How to perform ensemble docking with GOLD

CCDC Virtual Workshop

Shalini John Lovis, Rupesh Chikhale, Francesca Stanzione, Ilaria Gimondi, Suzanna Ward

October 2022

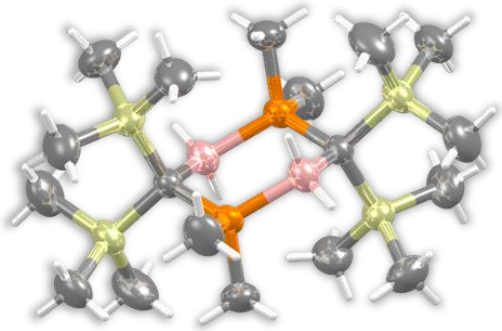
CCDC
advancing structural science

Learning outcomes for today

- Brief overview of ensemble docking.
- Basics of the Hermes interface, the CCDC's 3D visualizer for proteins.
- Step-by-step set-up of an ensemble docking simulation in GOLD, including:
  - Superimposing the protein structures;
  - Binding site definition;
  - Ligand selection and ligand flexibility;
  - Functional waters.
- Visualization and analysis of ensemble docking results.

CCDC

The Cambridge Structural Database



CSD Refcode ELOFUJ
The 1.1 millionth CSD structure
A *CSD Communication*
determined by Bill Clegg

- Every published structure
  - Inc. ASAP & early view
  - CSD Communications
  - Patents
  - University repositories
  - Thesis
- Every entry enriched and annotated by experts
- Discoverability of data and knowledge
- Sustainable for over 57 years
- A trusted CoreTrustSeal repository

Certified as Trustworthy by CoreTrustSeal

Inside the Cambridge Structural Database

The CSD is a database of all the published organic and metal-organic experimental crystal structures

Organic
43%

Metal-Organic
57%

At least one transition metal, lanthanide, actinide or any of Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po

Not Polymeric
89%

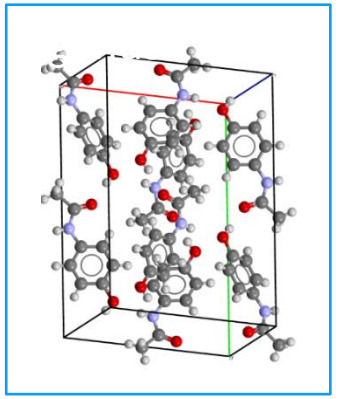
Polymeric: 11%

Single Component
56%

Multi Component
44%

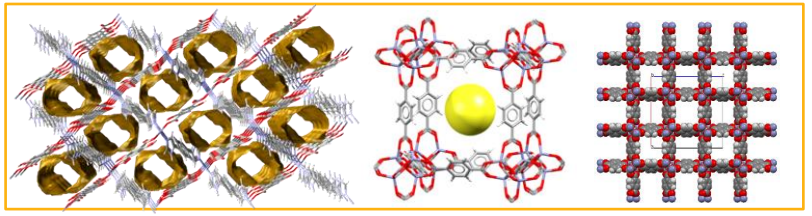
Organic

- Drugs
- Agrochemicals
- Pigments
- Explosives
- Protein ligands



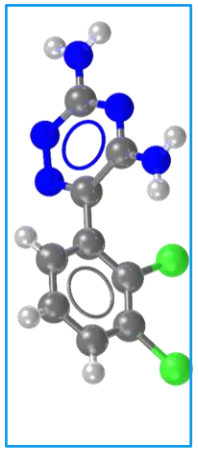
Metal-Organic

- Metal Organic Frameworks
- Models for new catalysts
- Porous frameworks for gas storage
- Fundamental chemical bonding



Additional data

- 11,525 polymorph families
- 171,683 melting points
- 909,992 crystal colours
- 778,663 crystal shapes
- 24,916 bioactivity details
- 11,387 natural source data
- > 250,000 oxidation states



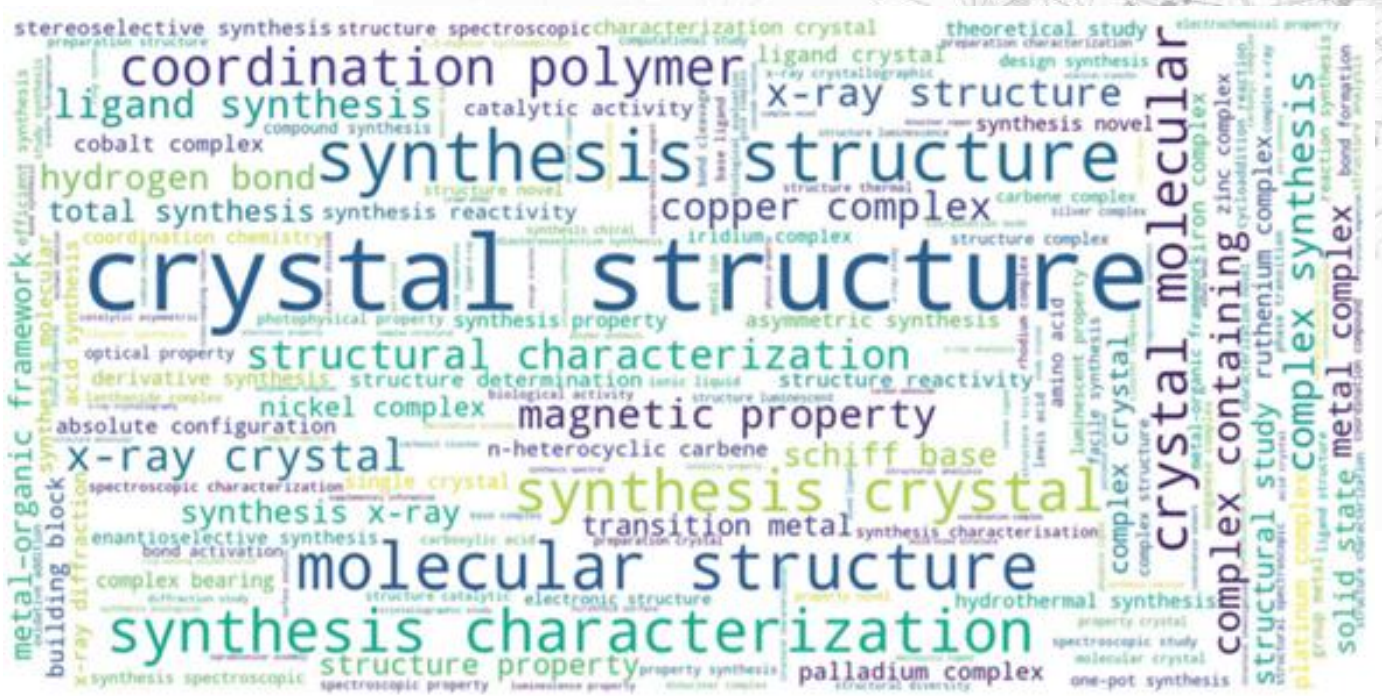
Links and subsets

- DrugBank
- Druglike
- MOFs
- PDB ligands
- PubChem
- ChemSpider
- Pesticide PDB

The CSD in research

The *Web of Science* subject categories that cite the 2006-2016 CCDC standard references most frequently

Subject category	Citations
Crystallography	5256
Chemistry Multidisciplinary	3275
Chemistry Inorganic Nuclear	1931
Chemistry Physical	1148
Materials Science Multidisciplinary	864
Chemistry Organic	568
Biochemistry Molecular Biology	307
Physics Atomic Molecular Chemical	203
Chemistry Medicinal	185
Spectroscopy	181



A word cloud of common bigrams in the titles of publications containing CSD-compliant crystal structures.

Willett P. et al, *CrystEngComm*, 2020,22, 7233-7241 DOI:10.1039/D0CE00045K

CCDC

Top 200 drugs

- By retail sale

Small molecule structure in the CSD

Biological structure in the PDB

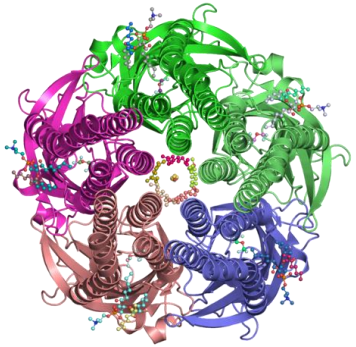
Small molecule structure CSD entry pending

- Adapted from poster compiled and produced by the Njardarson Group at the University of Arizona

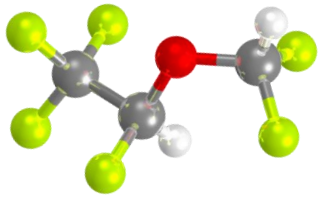
https://www.ccdc.cam.ac.uk/Community/blog/2020-FDA-Drug-Approvals/

Adapted from the poster compiled and produced by the Njardarson Group (The University of Arizona), J. Chem. Ed. 2010, 87, 1348

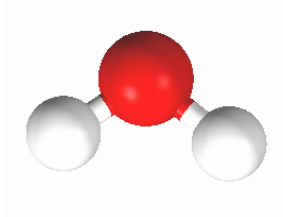
Proteins and ligands



Proteins are large biomolecules and macromolecules that comprise of one or more long chains of amino acids.



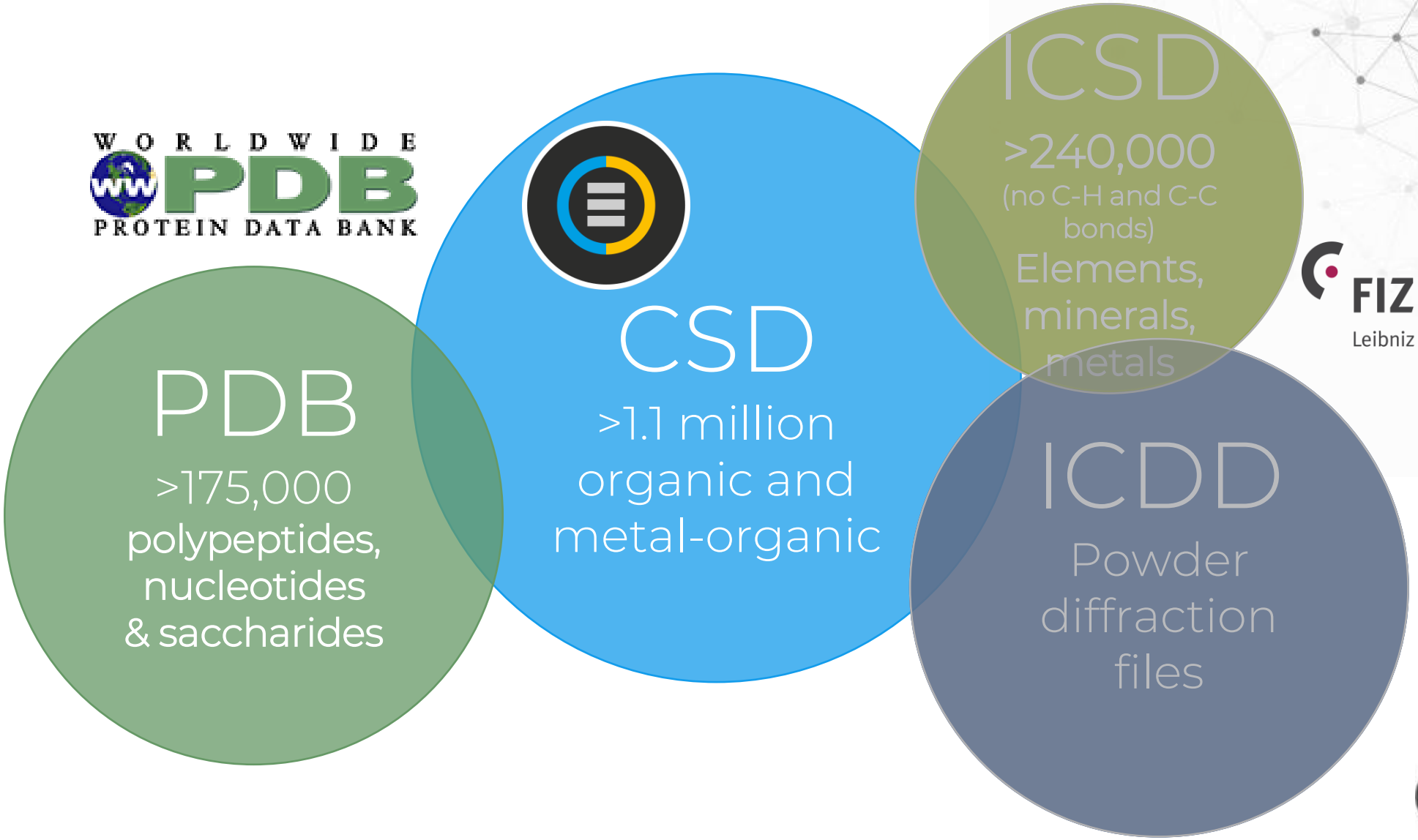
Ligands are small molecules that bind to the protein and can change the protein function.



Functional waters are found in the binding site of a protein and mediate the interactions between the ligand and the protein.

CCDC

Structural databases



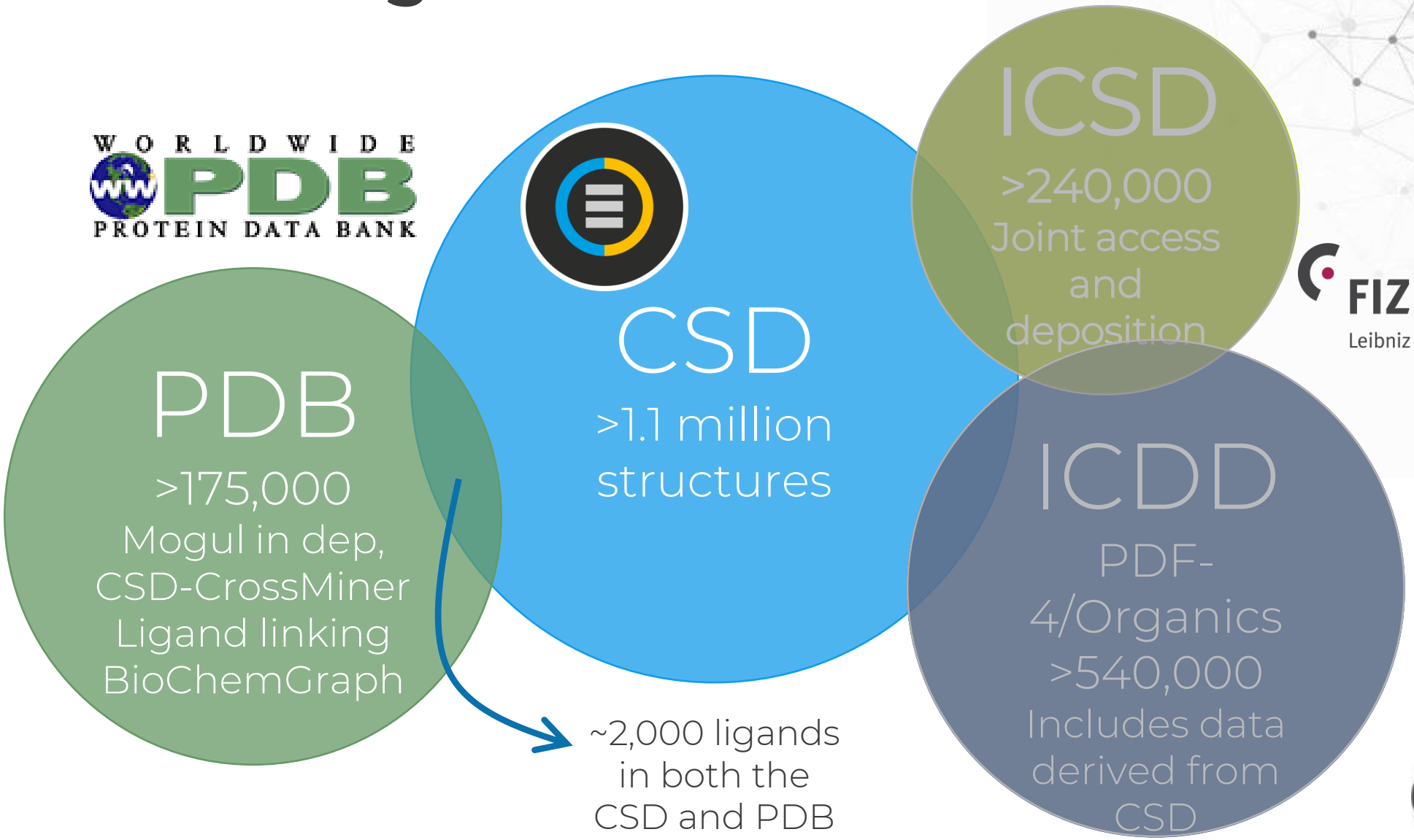
WORLDWIDE PDB
PROTEIN DATA BANK

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

ICDD®

CCDC

More integrated structural databases



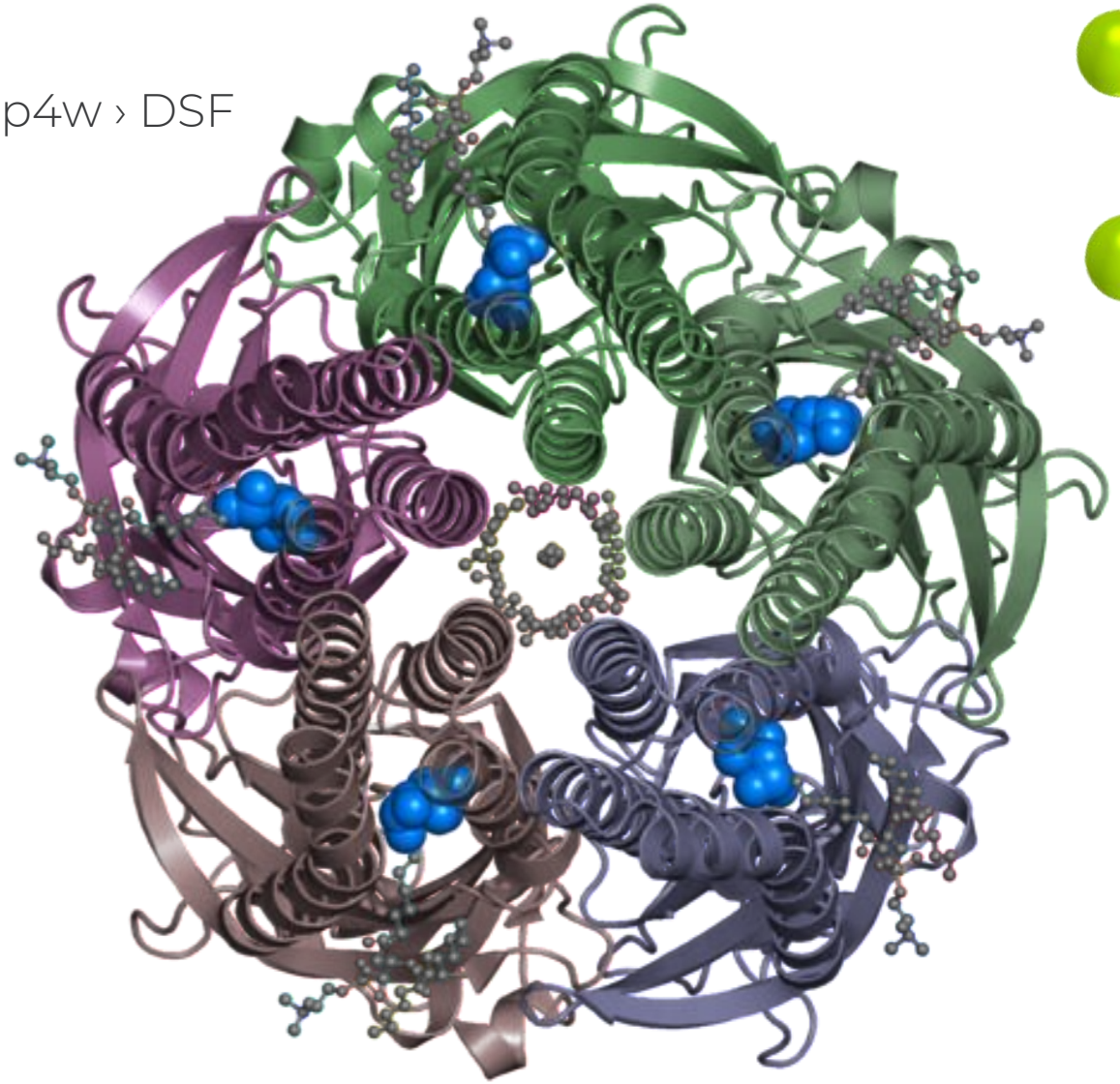
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

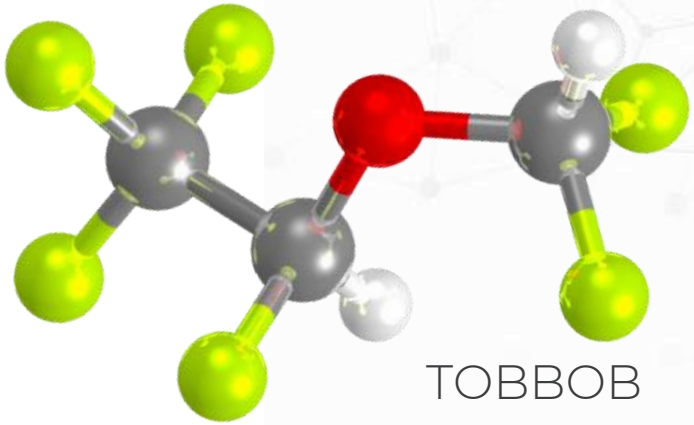
ICDD

CCDC

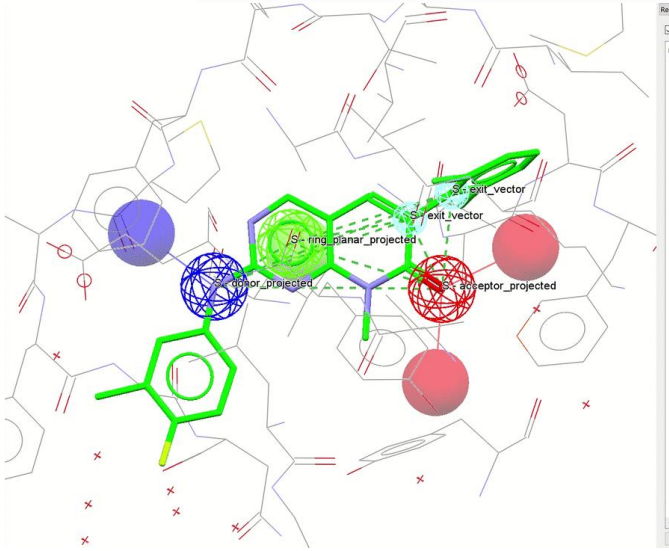
Connecting chemistry and biology

3p4w › DSF



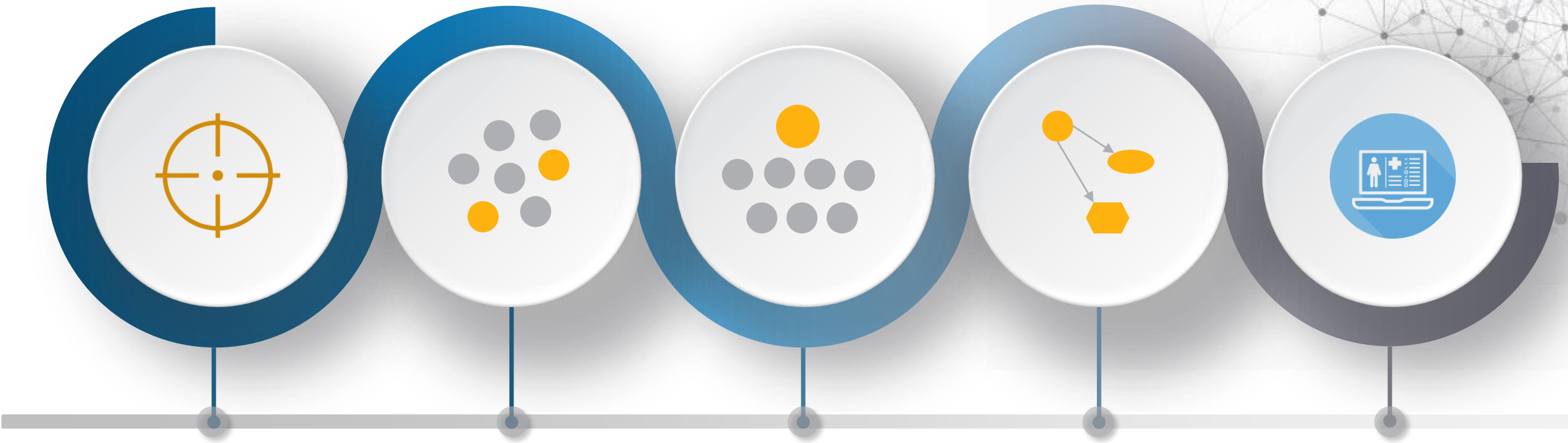


TOBBOB



CCDC

Drug Discovery Pipeline



TARGET SELECTION

HIT IDENTIFICATION

Structure- based virtual screening.

HIT TO LEAD

Assess how changes affect binding.
Optimize compound geometry.
Predict binding of small molecules to active pockets in proteins.

LEAD OPTIMISATION

Check the impact of changes with docking pose prediction.
Understand how changes affect conformations.

DRUG DEVELOPMENT

CCDC

Our knowledge-based software solutions

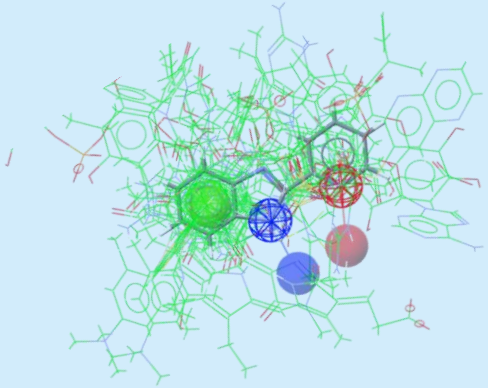
CSDCore.

Search, visualise, analyse and communicate structural data

Insights into molecular and crystal shape and interactions

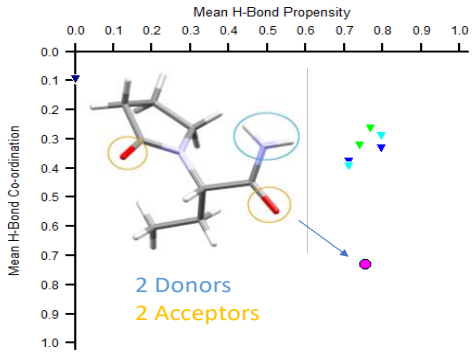
CSDDiscovery.

Design of new molecules



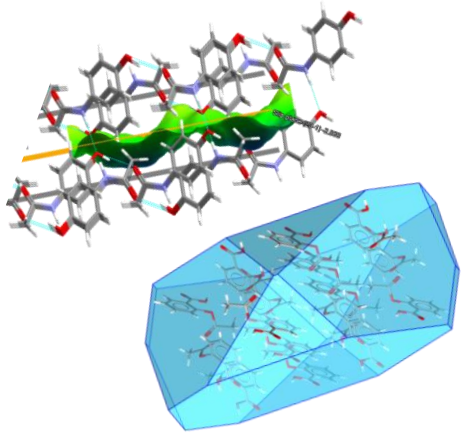
CSDMaterials.

Assessment of solid form stability and properties



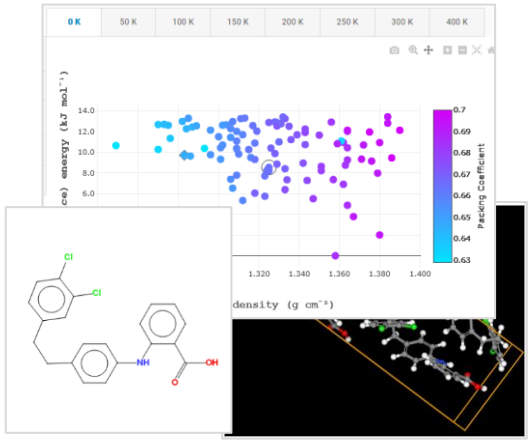
CSDParticle.

Anticipate particle properties and behaviour



CSDTheory.

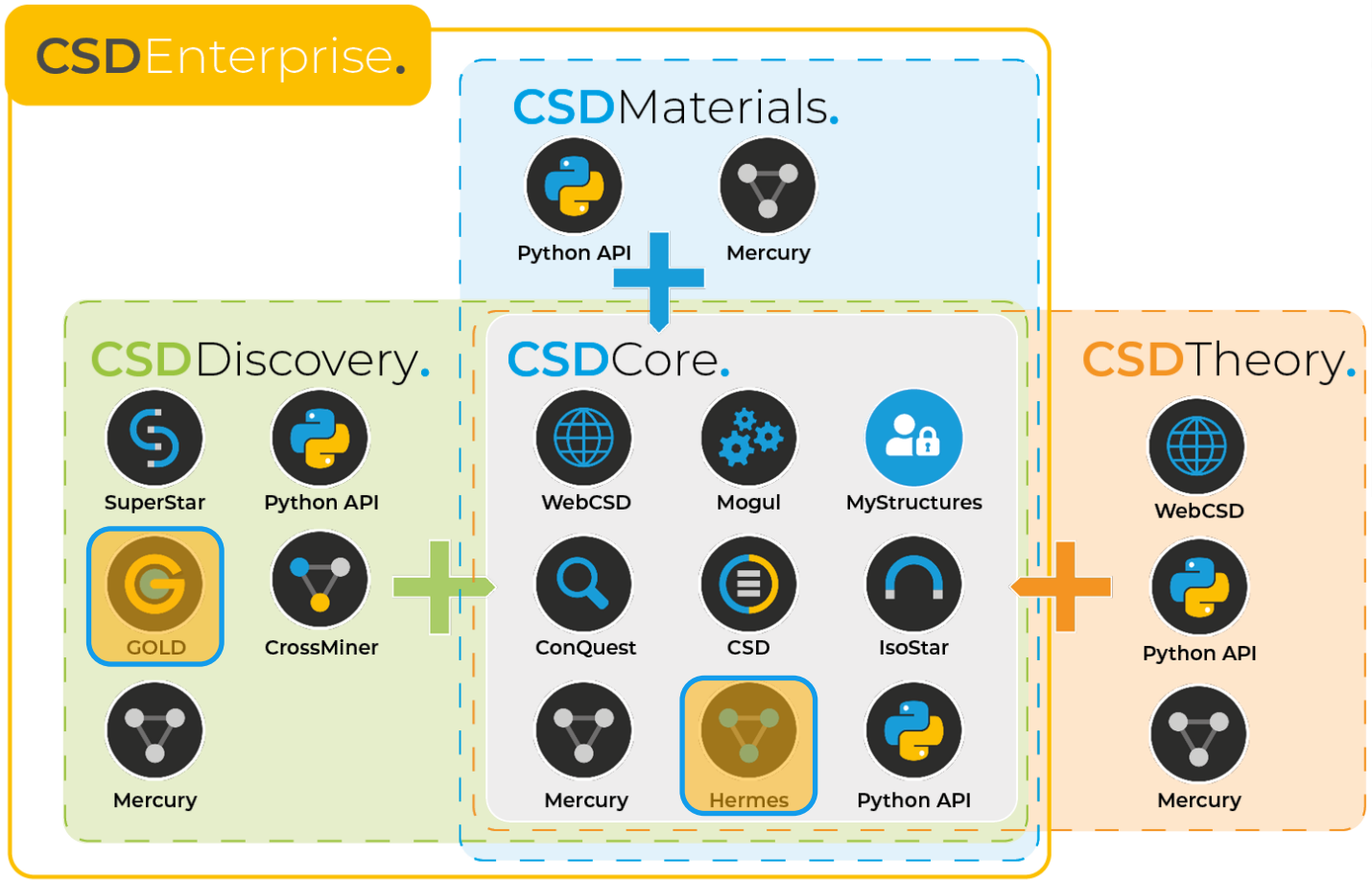
Insights from predicted structure landscapes



Medicinal & Computational Chemists ♦ Crystallographers & Structural Biologists ♦ Solid Form & Crystallisation Scientists

CCDC

Our products and services

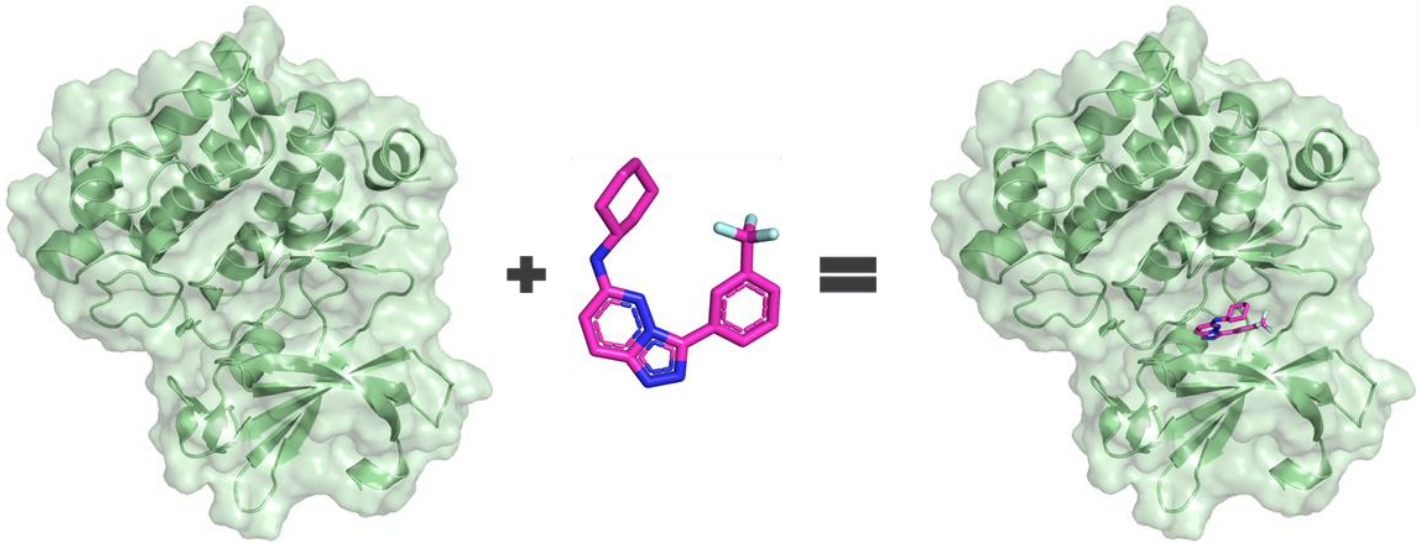


CCDC

Docking and its importance

- An efficient strategy to identify bioactive conformations.
- To identify possible compounds that bound to specific target of interest.
- Quick screen library of compounds.
- Many docking programs have been developed and applied in drug discovery.

An example of protein-ligand docking



Protein Ligand Protein-Ligand Complex

CCDC

GOLD: Protein-ligand docking software

- GOLD (Genetic Optimisation for Ligand Docking) is a genetic algorithm for docking flexible ligands into protein binding sites.

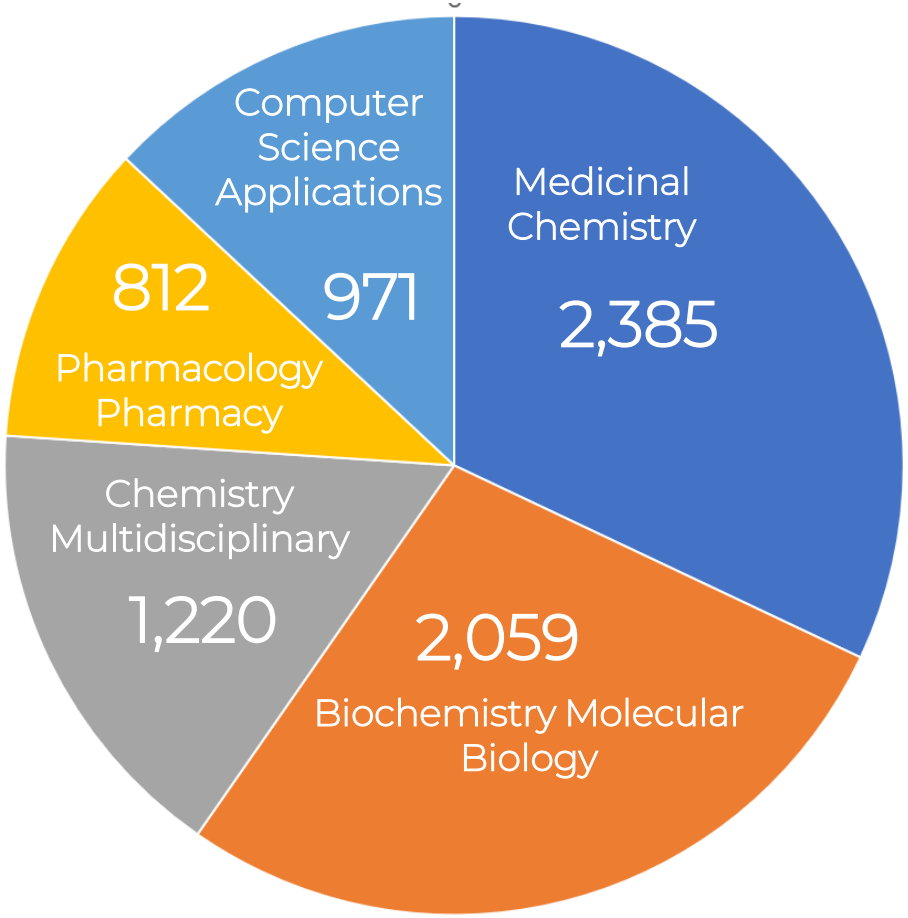
- GOLD has proven success in virtual screening, lead optimisation, and identifying the correct binding mode of active molecules.

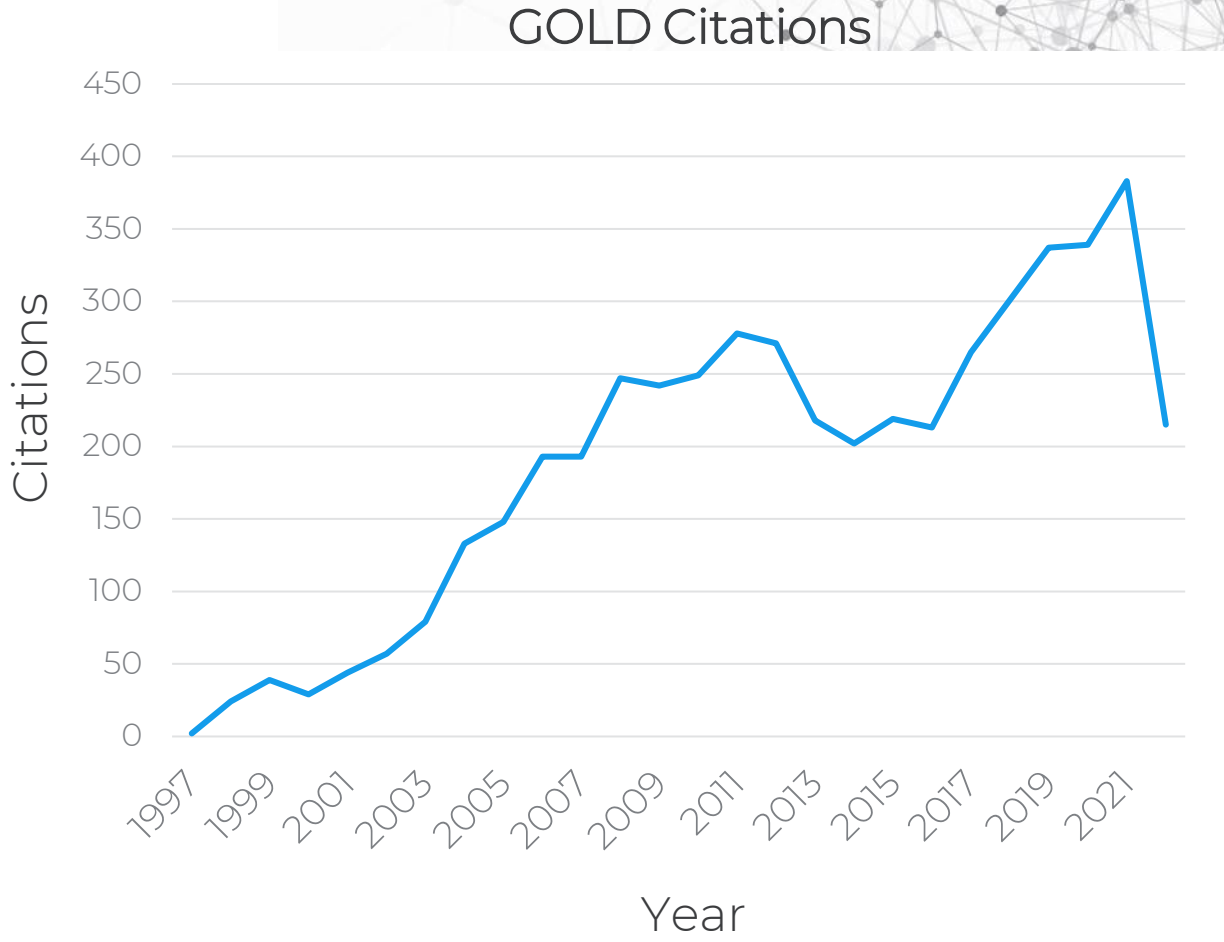
- Relied on by researchers in academia and industry worldwide.

- Reliable
- Flexible
- Configurable

CCDC

GOLD docking in research

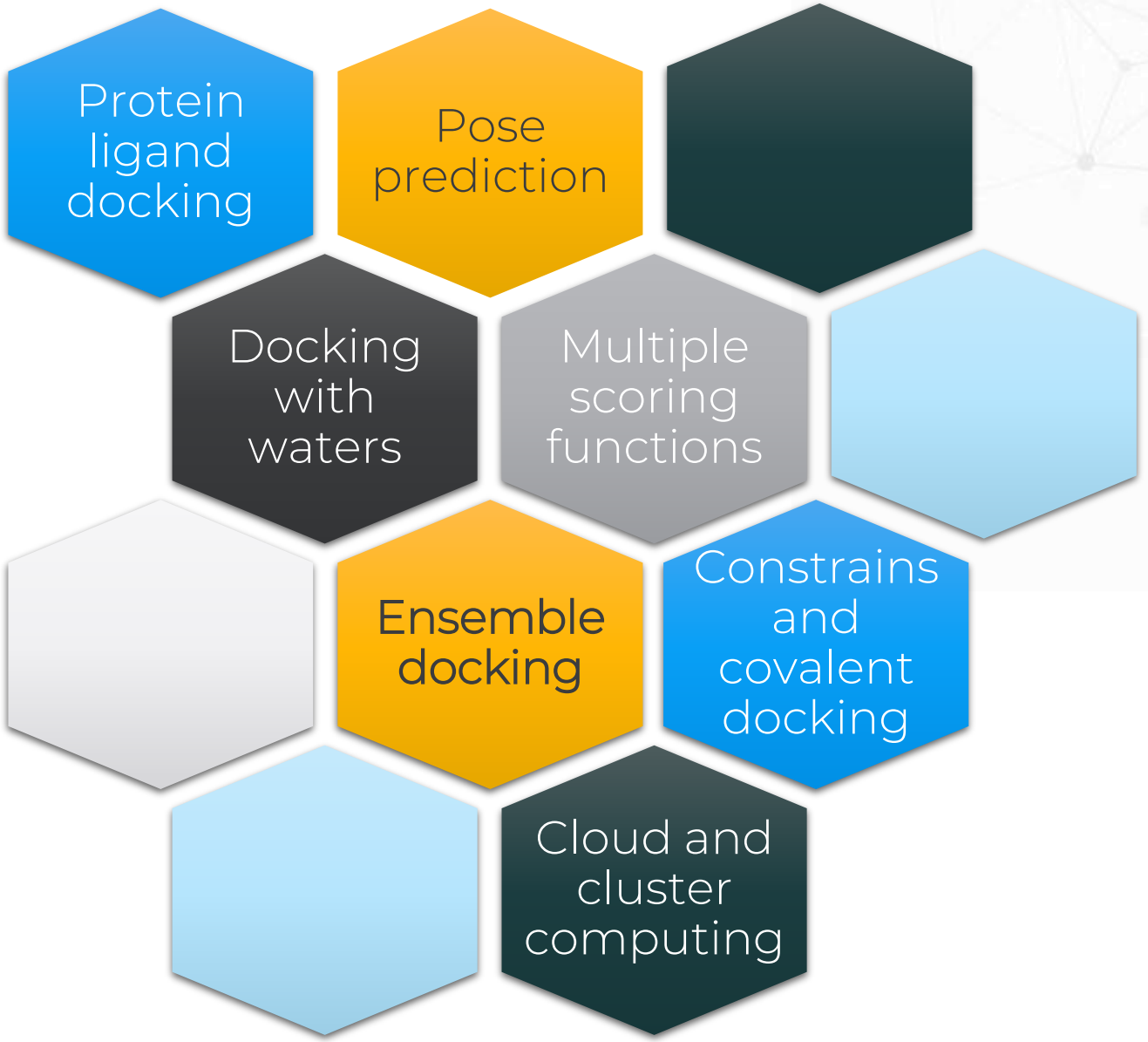




Development and validation of a genetic algorithm for flexible docking, *J. Mol. Biol.*, Vol. 267, pp. 727-748., 1997. DOI: https://doi.org/10.1006/jmbi.1996.0897

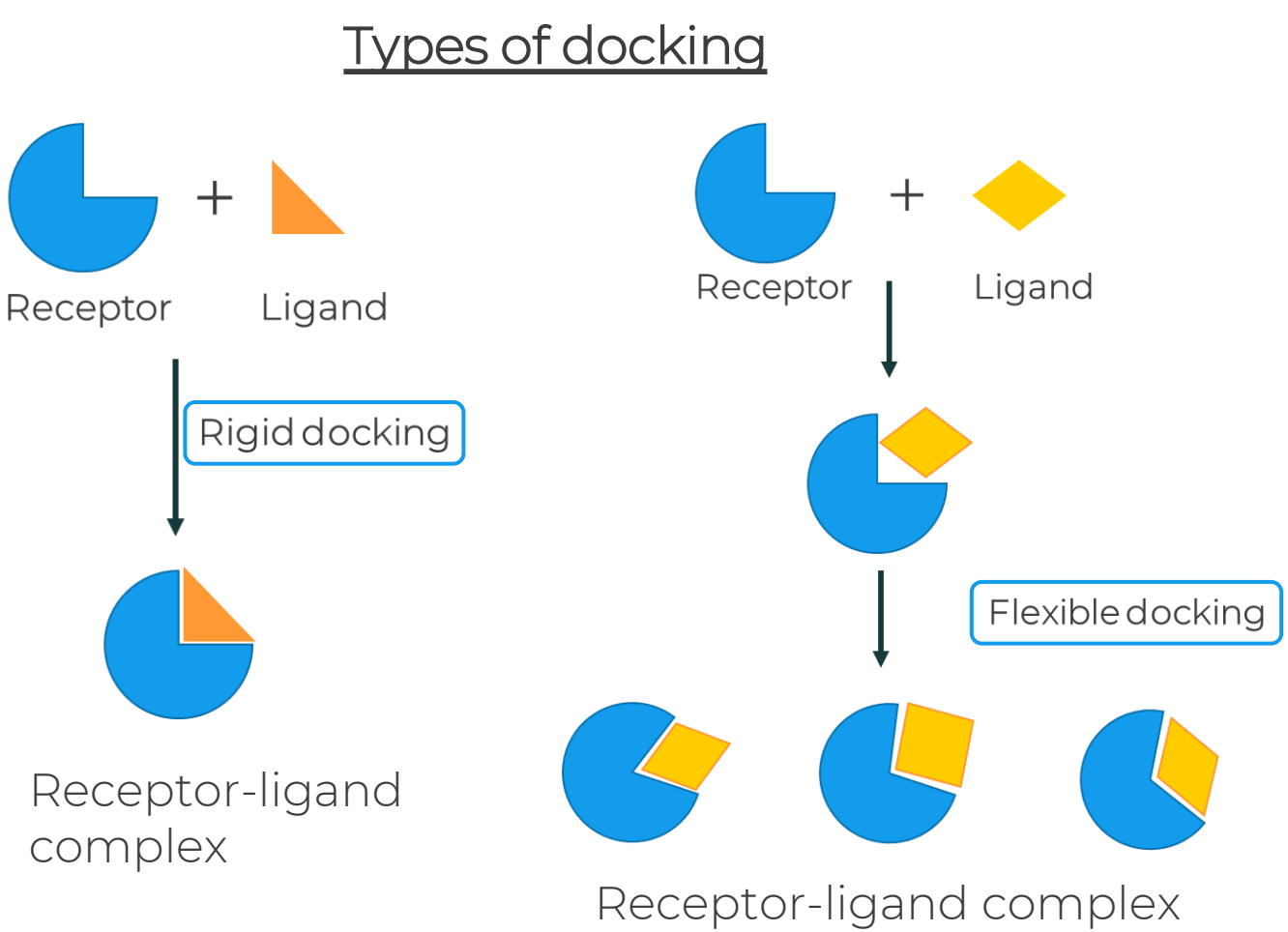
CCDC

GOLD: All in one molecular docking package



CCDC

Molecular docking: Challenges



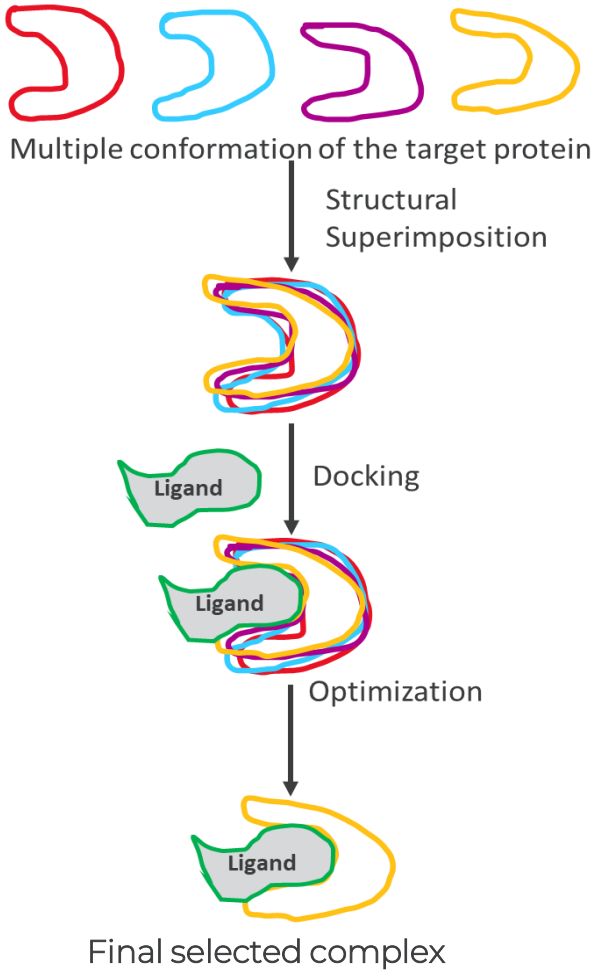
Challenge: Protein flexibility

- Slight conformational changes can affect ligand binding free energy.
- Methods to account for small protein changes exist.
- Structural rearrangements such as backbone movements are difficult to address.

Korb, O. *et al*. "Potential and Limitations of Ensemble Docking." *J. Chem. Inf. Model*. (2012). 52, 1261-1274
Chao, Y. *et al*. "Protein–Ligand Docking in the Machine-Learning Era." *Molecules* (2022). 27, 4568

CCDC

Ensemble docking: what is it and why it is important?



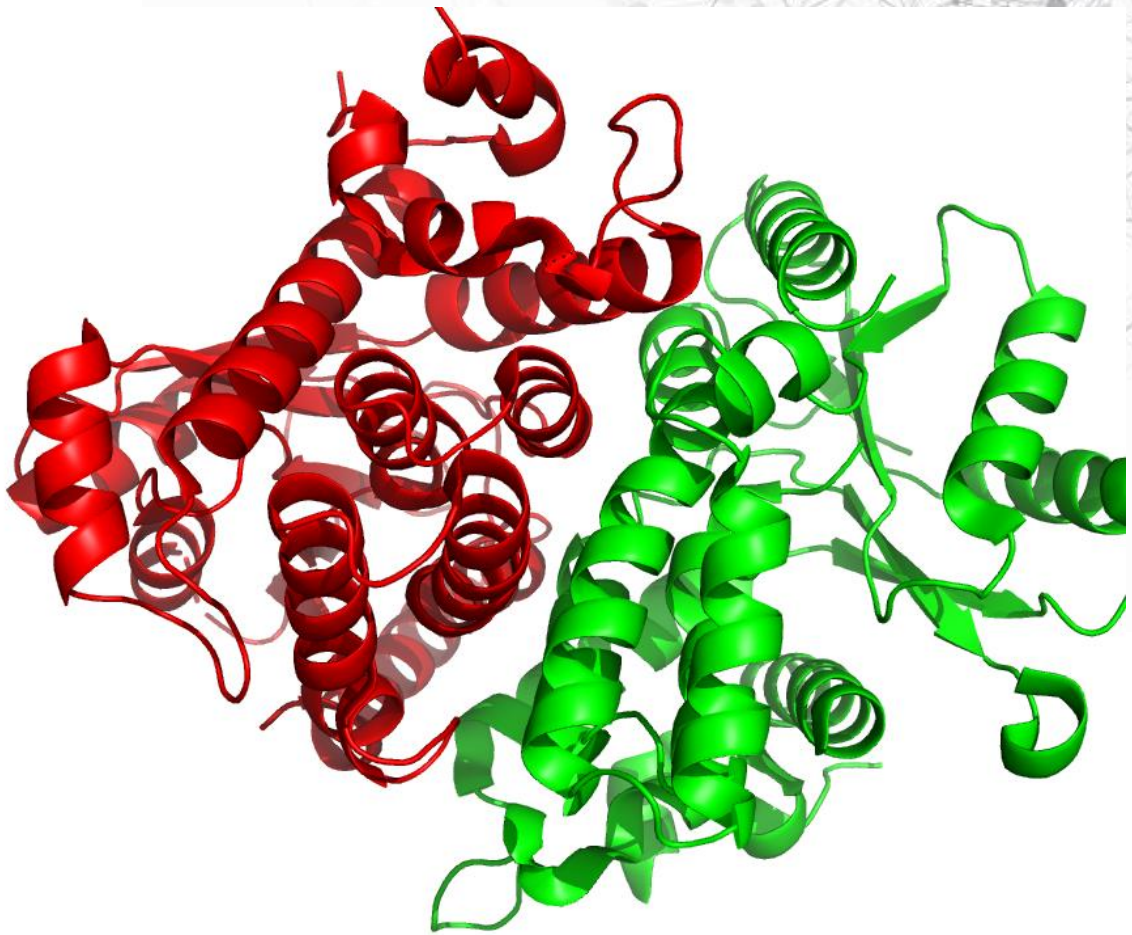
Solution: Ensemble docking

- It accounts for several discrete protein conformations for docking.
- Ensemble docking uses multiple conformations for a specific target.
- Resources: PDB is a good source to obtain multiple conformations of a protein.

Korb, O. *et al*. "Potential and Limitations of Ensemble Docking." *J. Chem. Inf. Model*. (2012). 52, 1261-1274

Ensemble docking : Case study

- Thymidine kinase (TK) catalyses the phosphorylation of thymidine (dT) to thymidine monophosphate (dTMP) in the presence of Mg2+ and ATP.
- Viral thymidine kinase from Herpes simplex virus type 1 (TK_{HSV1}) accepts a broad range of substrate
  - Enzyme-prodrug gene therapy of cancer.
- 2'-Exo-methanocarba-thymidine (MCT) is a potent antiviral compound characterised by high activity against HSV1 and HSV2.



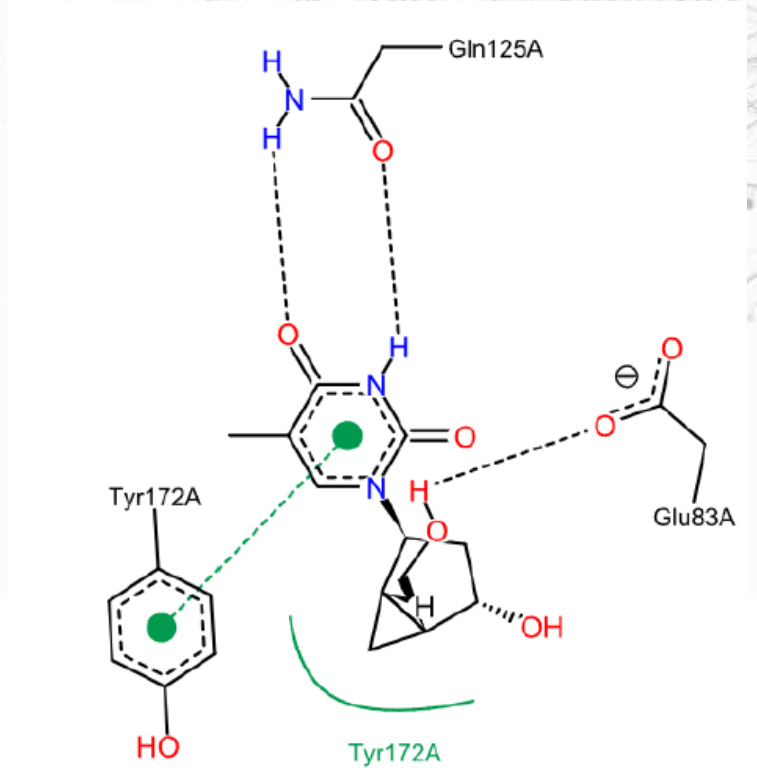
Crystal structure of TK_{HSV1}(apo) - PDB code 1e2h. The two asymmetric subunits A and B.

CCDC

Ensemble docking : Case study

- The thymidine ring of MCT is stacked between Met128 and Tyr172 and fixed by a complex hydrogen bonding network

- The hydrogen bonds between nucleobase, Gln125, Arg 176 and two water-mediated hydrogen bonds retains the nucleobase within the active site

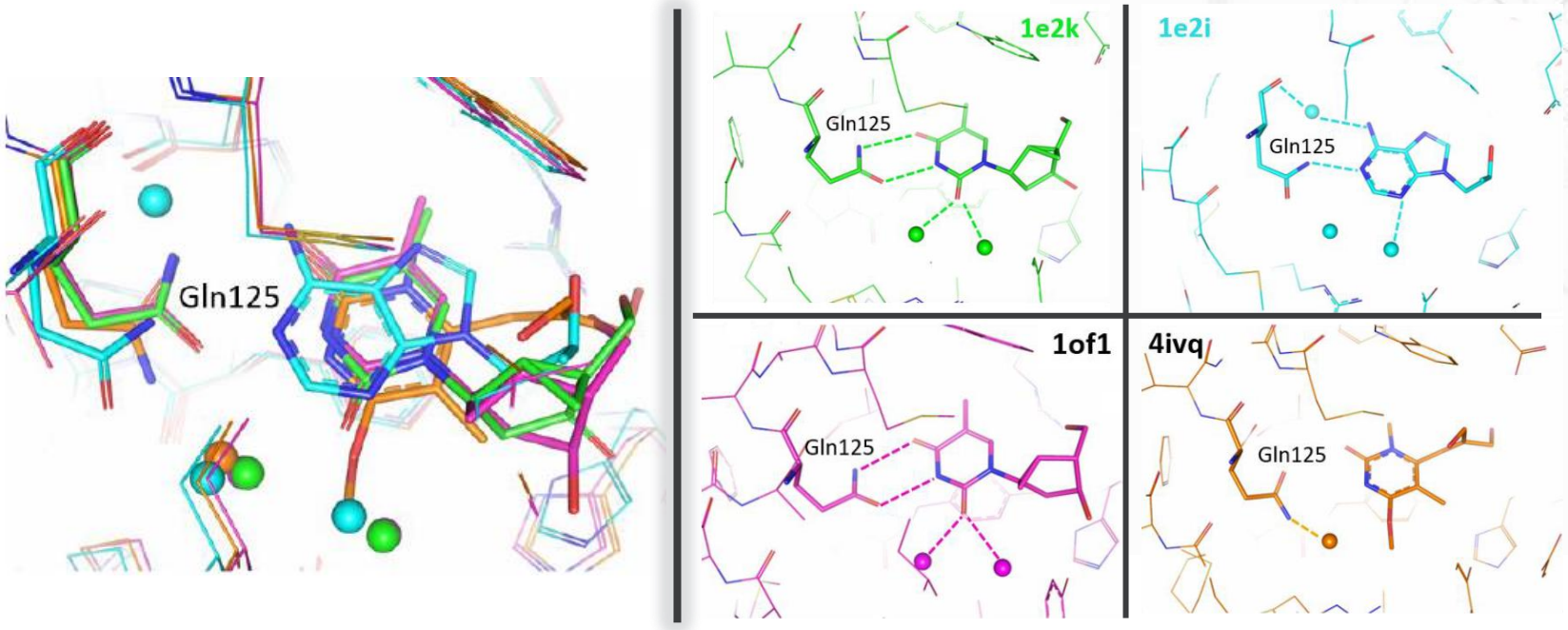


2D interactions diagram of MCT in complex with TK_{HCVI}, PDB code: 1e2k

CCDC

Ensemble docking : Case study

Virtual Screening example using TK



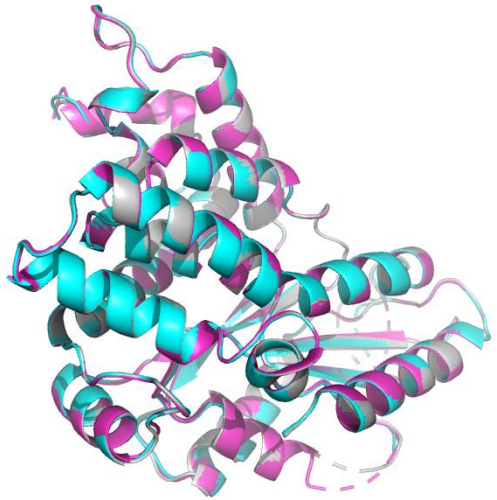
- Perform a non-native docking of MCT into an ensemble of four different protein conformers of TK_{HSV1}.
- Investigate how this inhibitor would fit into the protein target by accounting for its flexibility.

CCDC

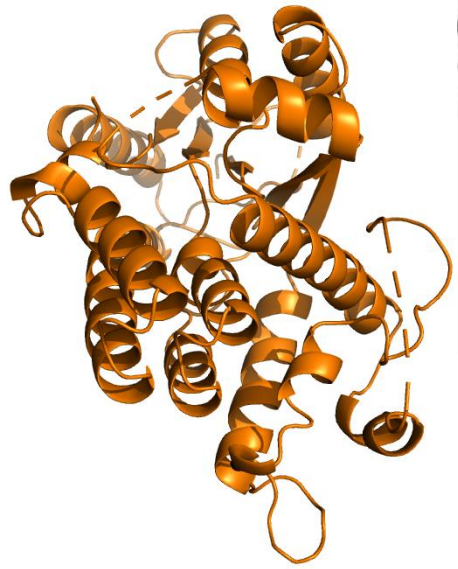
Workflow – Ensemble docking with GOLD

Target selection

- Co-crystal structure of protein-ligand complex
- PDB codes: **1E2K**, **1E2I**, **1FO1** and **4IVQ**
- Without missing residues and atoms



1E2K, 1E2I, 1FO1



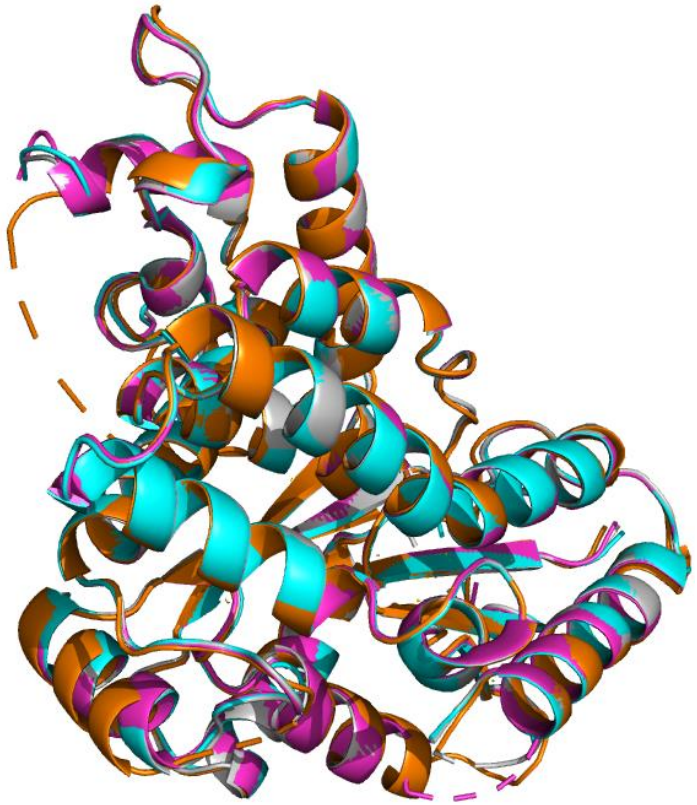
4IVQ

CCDC

Workflow – Ensemble docking with GOLD

Target selection → **Superimposition of protein**

1E2K
1E2I
1FO1
4IVQ

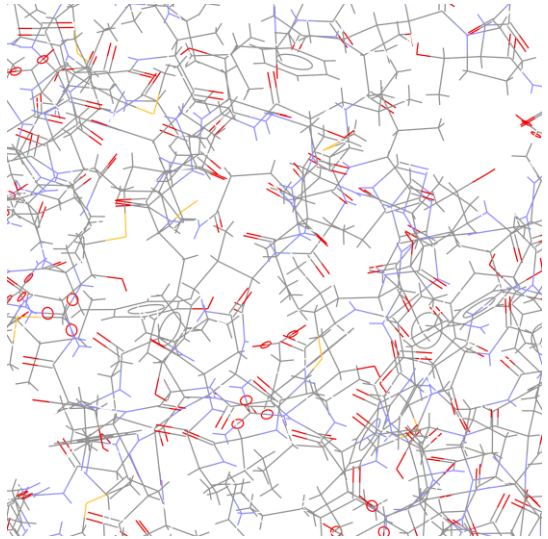


Overlay of 4 proteins

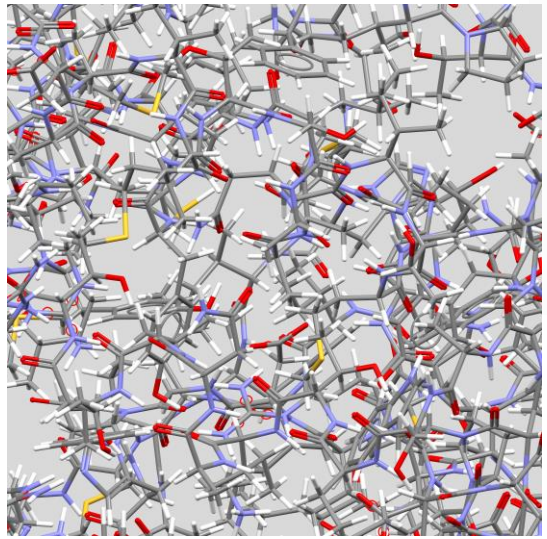
CCDC

Workflow – Ensemble docking with GOLD

Target selection → **Superimposition of protein** → **Protein set-up**



Add hydrogens →



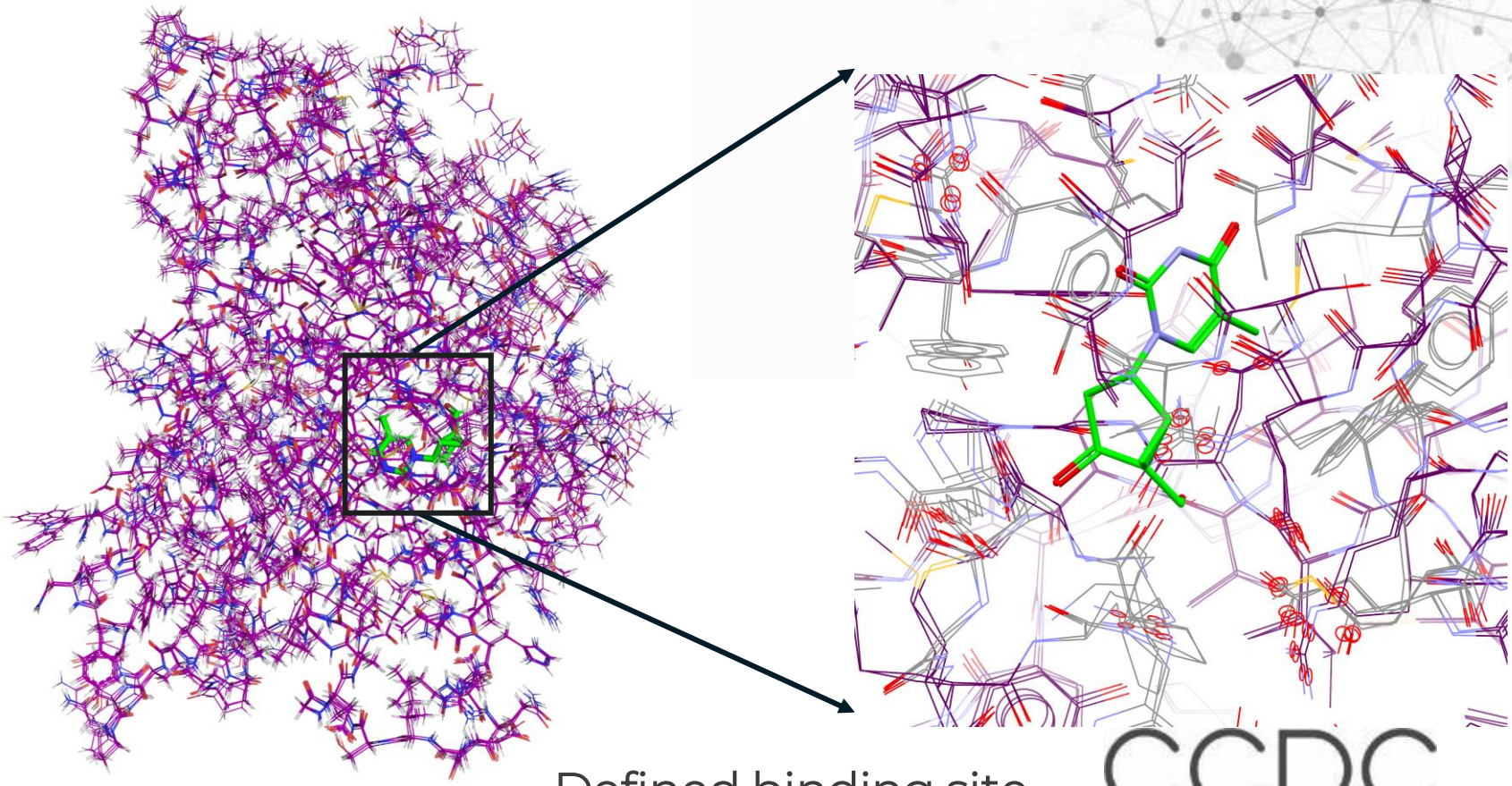
- Add hydrogens.
- Check protonation of His, Gln, Asn.
- Remove waters or retain required ones.
- Extract ligand and save.

CCDC

Workflow – Ensemble docking with GOLD

Target selection → Superimposition of protein → Protein set-up → Binding site definition

- For an ensemble it is not possible to define the binding site from an atom or a list of atoms or residues.
- It is only possible to define the binding site from a point in space or from a ligand.

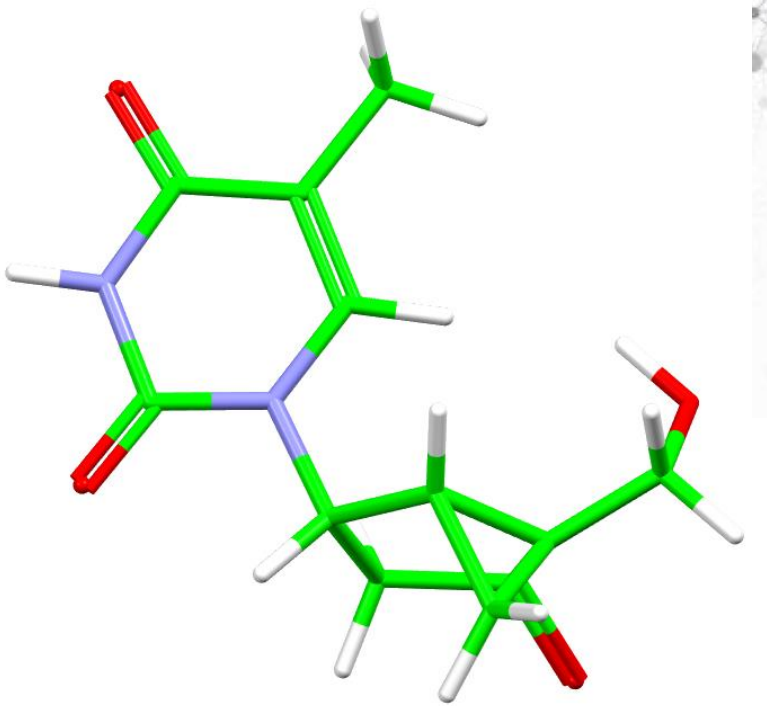


Defined binding site

CCDC

Workflow – Ensemble docking with GOLD

Target selection → Superimposition of protein → Protein set-up → Binding site definition → Selecting ligands

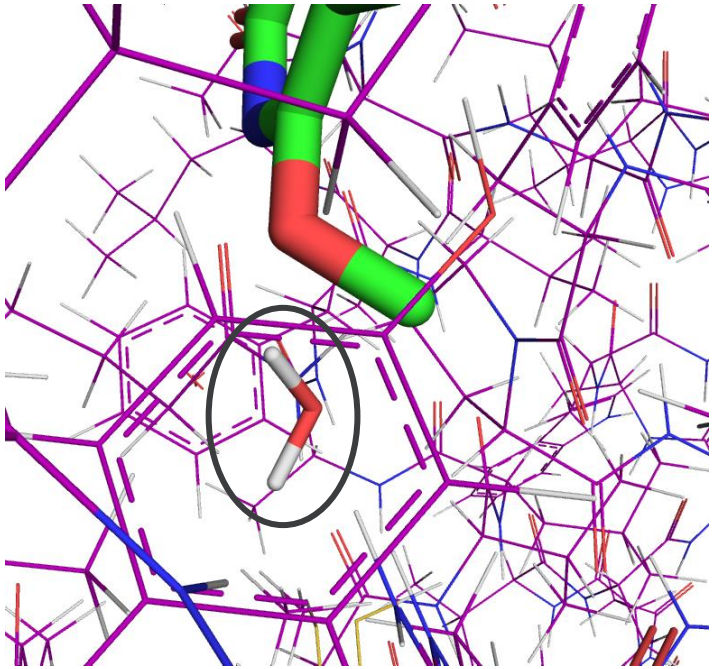


2’-exo-methanocarba-thymidine (MCT)

CCDC

Workflow – Ensemble docking with GOLD

Target selection → Superimposition of protein → Protein set-up → Binding site definition → Selecting ligands → Setting water in binding site

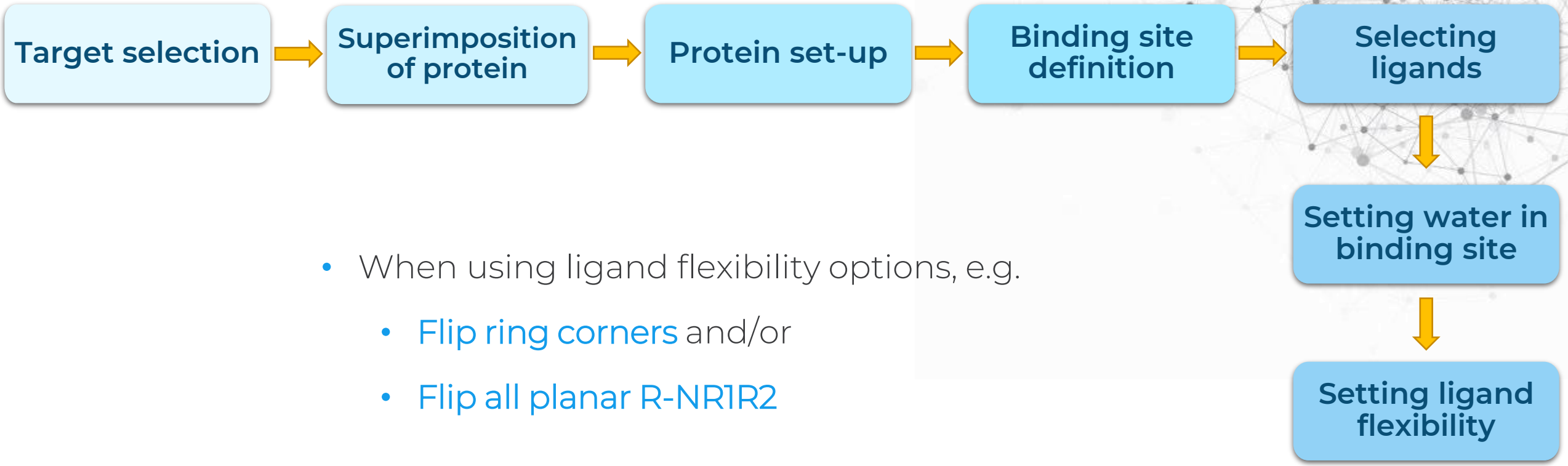


Water in binding site

- Must define the active waters to be considered during the ensemble docking.
- GOLD allows you to:
  - Switch on and off the water molecules.
  - Rotate water molecules around three principal axes during docking.

CCDC

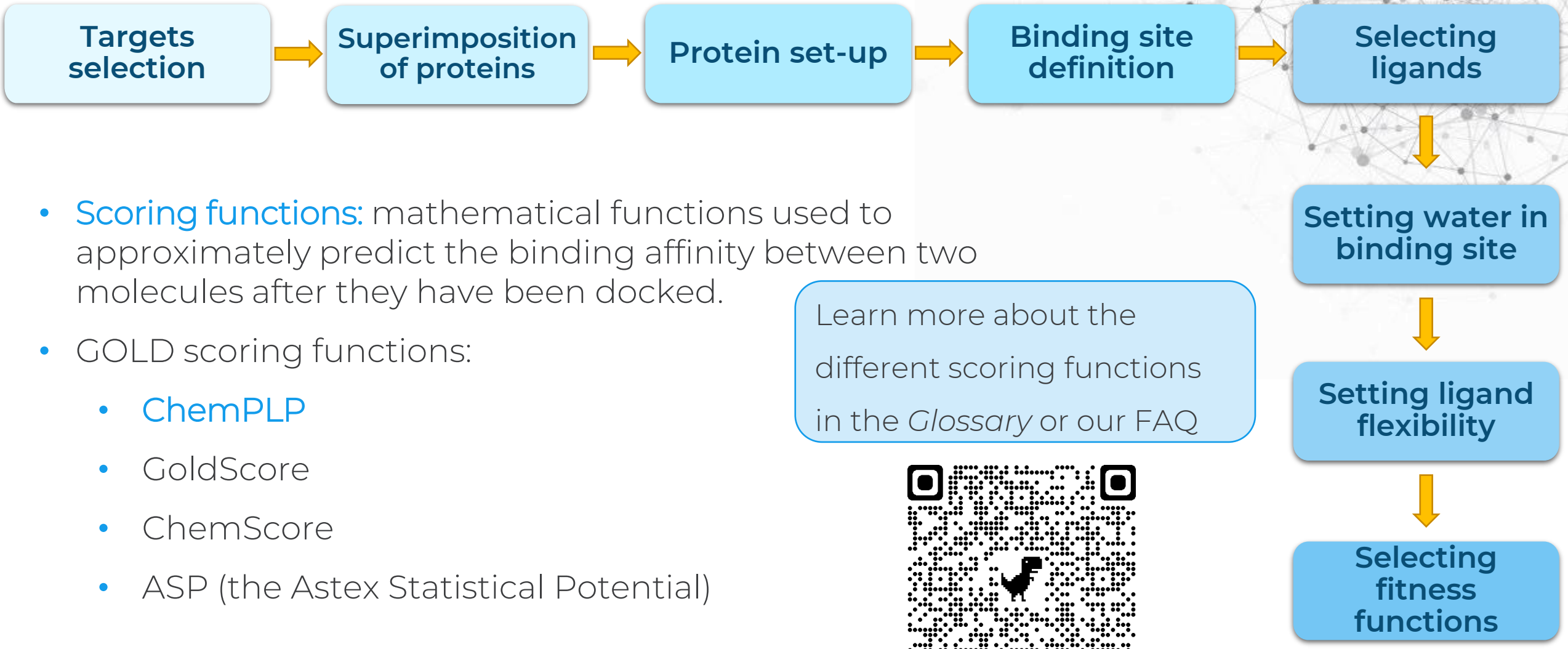
Workflow – Ensemble docking with GOLD



- When using ligand flexibility options, e.g.
  - Flip ring corners and/or
  - Flip all planar R-NR1R2

CCDC

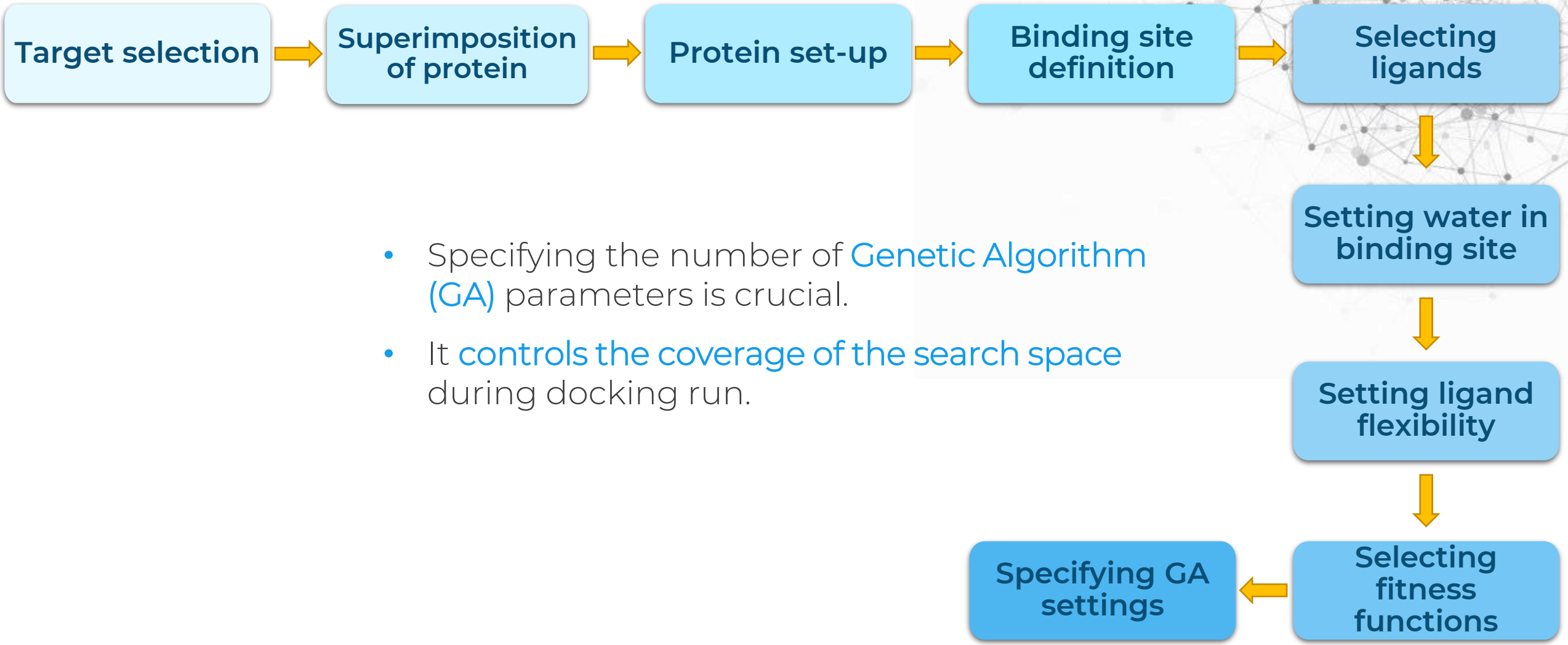
Workflow – Ensemble docking with GOLD



- Scoring functions: mathematical functions used to approximately predict the binding affinity between two molecules after they have been docked.
- GOLD scoring functions:
  - ChemPLP
  - GoldScore
  - ChemScore
  - ASP (the Astex Statistical Potential)

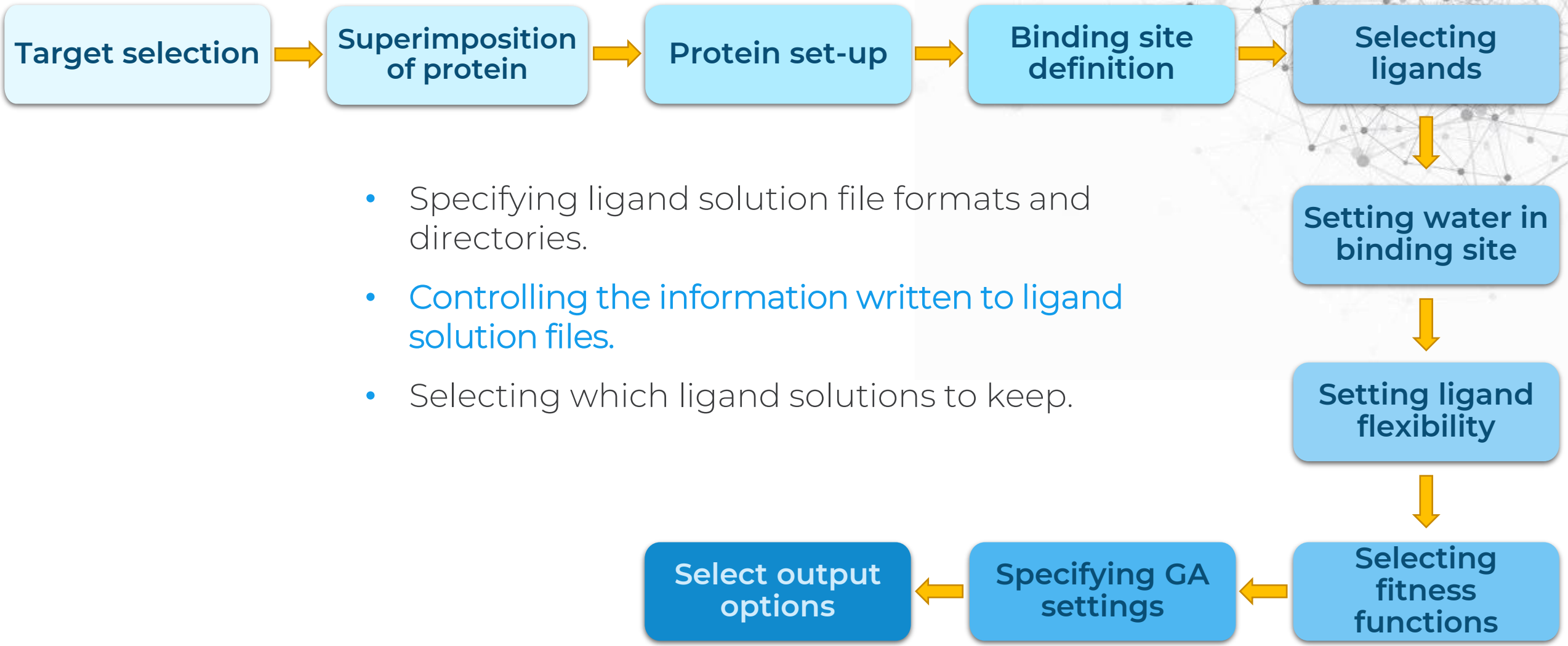
Learn more about the different scoring functions in the *Glossary* or our FAQ

Workflow – Ensemble docking with GOLD



- Specifying the number of Genetic Algorithm (GA) parameters is crucial.
- It controls the coverage of the search space during docking run.

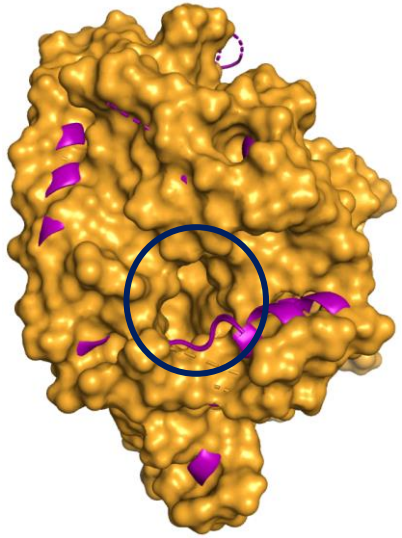
Workflow – Ensemble docking with GOLD



- Specifying ligand solution file formats and directories.
- Controlling the information written to ligand solution files.
- Selecting which ligand solutions to keep.

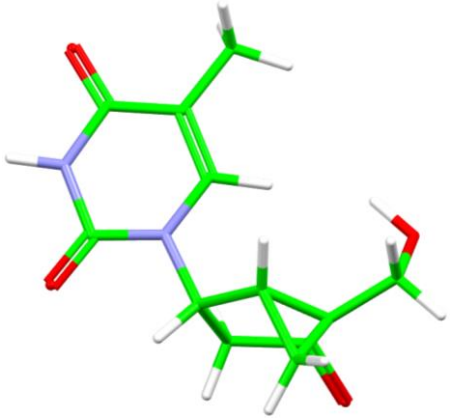
Workflow – Ensemble docking with GOLD

Target selection → Superimposition of protein → Protein set-up → Binding site definition → Selecting ligands → Setting water in binding site → Setting ligand flexibility → Selecting fitness functions → Specifying GA settings → Select output options → Run the ensemble docking



Ensembles of proteins

Docking

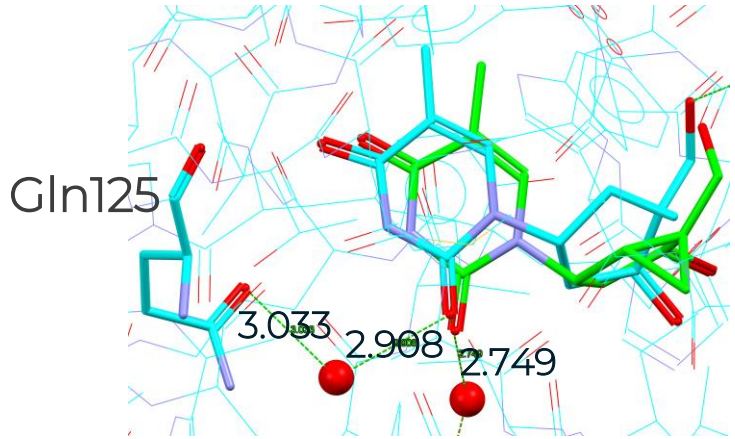


2’-exo-methanocarba-thymidine (MCT)

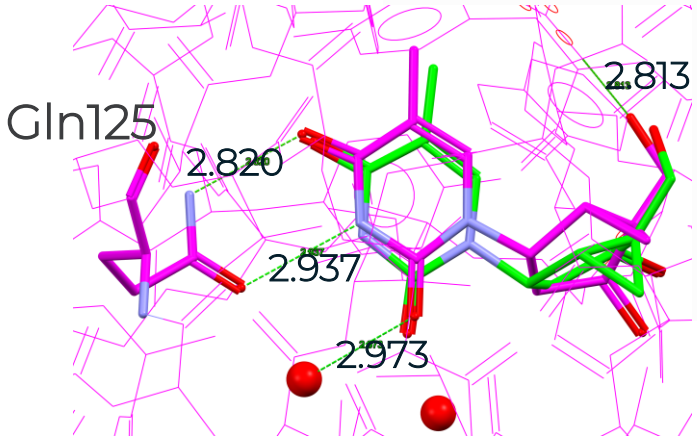
Workflow – Ensemble docking with GOLD

Target selection → Superimposition of protein → Protein set-up → Binding site definition → Selecting ligands → Setting water in binding site → Setting ligand flexibility → Selecting fitness functions → Specifying GA settings → Select output options → Run the ensemble docking → Analysing and visualizing docking results

Overlay of best hits on the reference molecule



1E2H

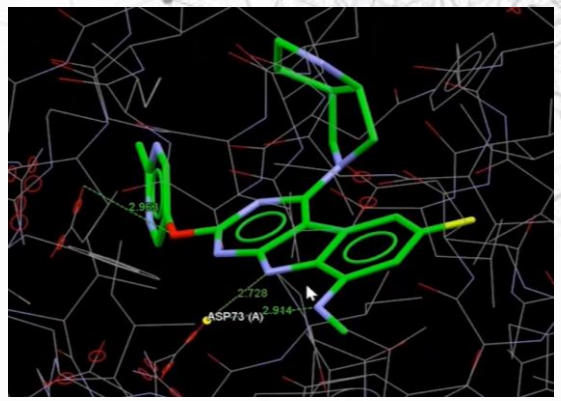


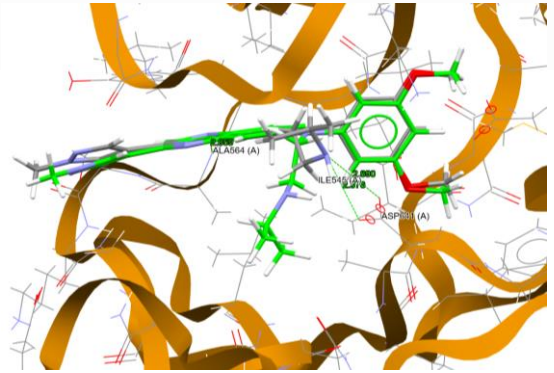
1FO1

Hermes – Structure visualisation

With Hermes you can:

- Visualise and edit macromolecules in 3D including proteins, nucleic acids, antibodies and ligands.
- Navigate around the structure and modify the appearance to display ribbons, etc.
- Interface to GOLD, Mogul, SuperStar, the CSD Ligand Overlay, and descriptors for GOLD docking poses.
- Generate publication quality images and share sessions for effective scientific communication.





Hermes is protein visualization software and is part of CSD-Core and CSD-Enterprise

CCDC

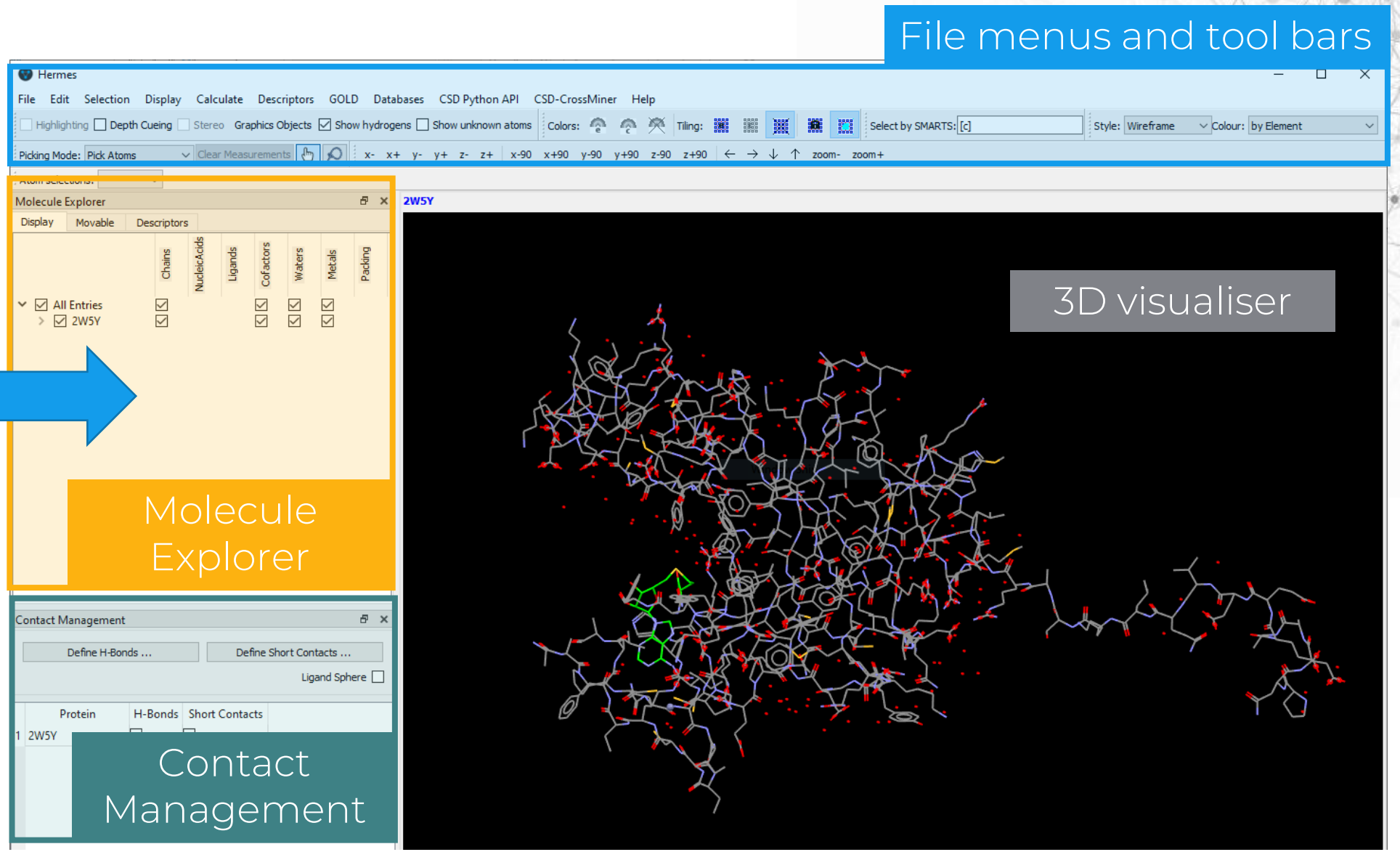
Hermes functionality

- 3D visualisation options:
  - Display styles, colours, labelling schemes, hide/unhide atoms, residues, ligands.
- Read in and prepare protein and ligand structures from external files:
  - Edit, prepare and add hydrogens, etc.
- The ability to:
  - Load and visualise contoured surfaces.
  - Overlay protein structures by least squares overlay or sequence alignment.
  - Measure and display distances, angles and torsion angles.
  - Find and display hydrogen bonds and nonbonded clashes, and customise how they are defined geometrically.
- Prepare and save publication quality displays

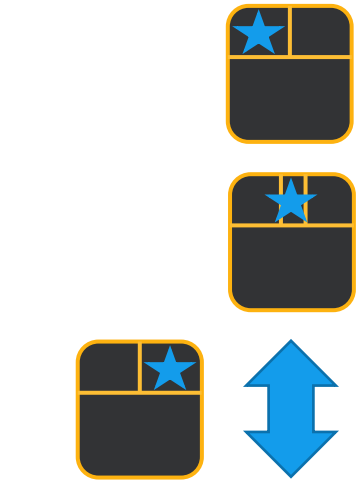
Hermes is protein visualization software and is part of CSD-Core and CSD-Enterprise

CCDC

Show One: Hermes Interface



Show One: Navigation basics



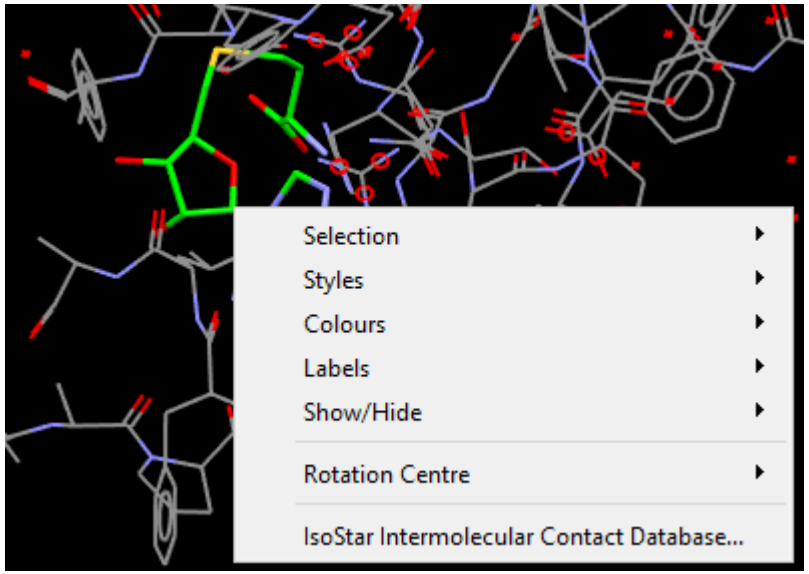
- Left mouse button and move – rotate molecules
- Middle Mouse wheel – move molecules up and down
- Right mouse button and move up and down – zoom in and out of molecules
- Shift + Left mouse button and move - rotate in the plane molecules
- Ctrl + Left mouse button and move - translate molecules

Did you know?
The basic navigation features in **Hermes** are the same in **Mercury** - our small molecule visualisation software. So... if you are an expert in one you are an expert in both!

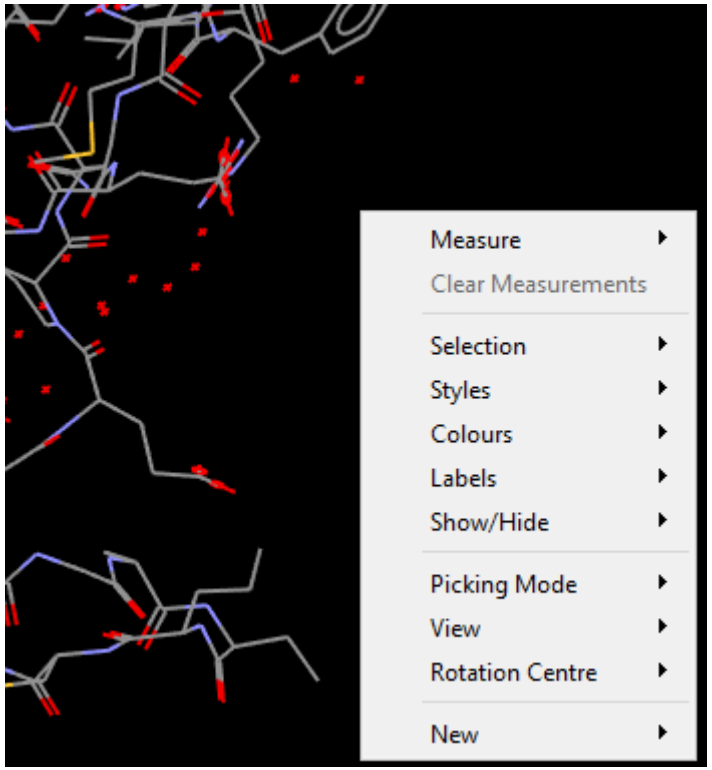
CCDC

Show One: Navigation basics – Right click

On a feature



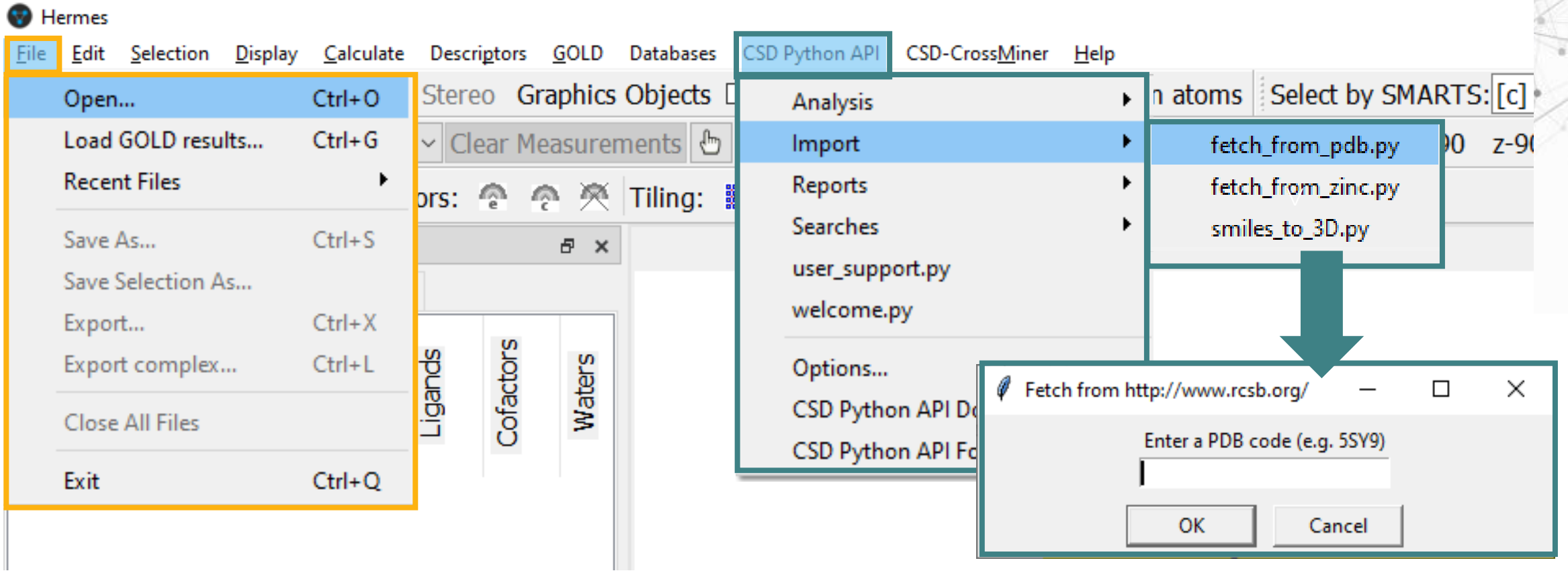
Away from a feature



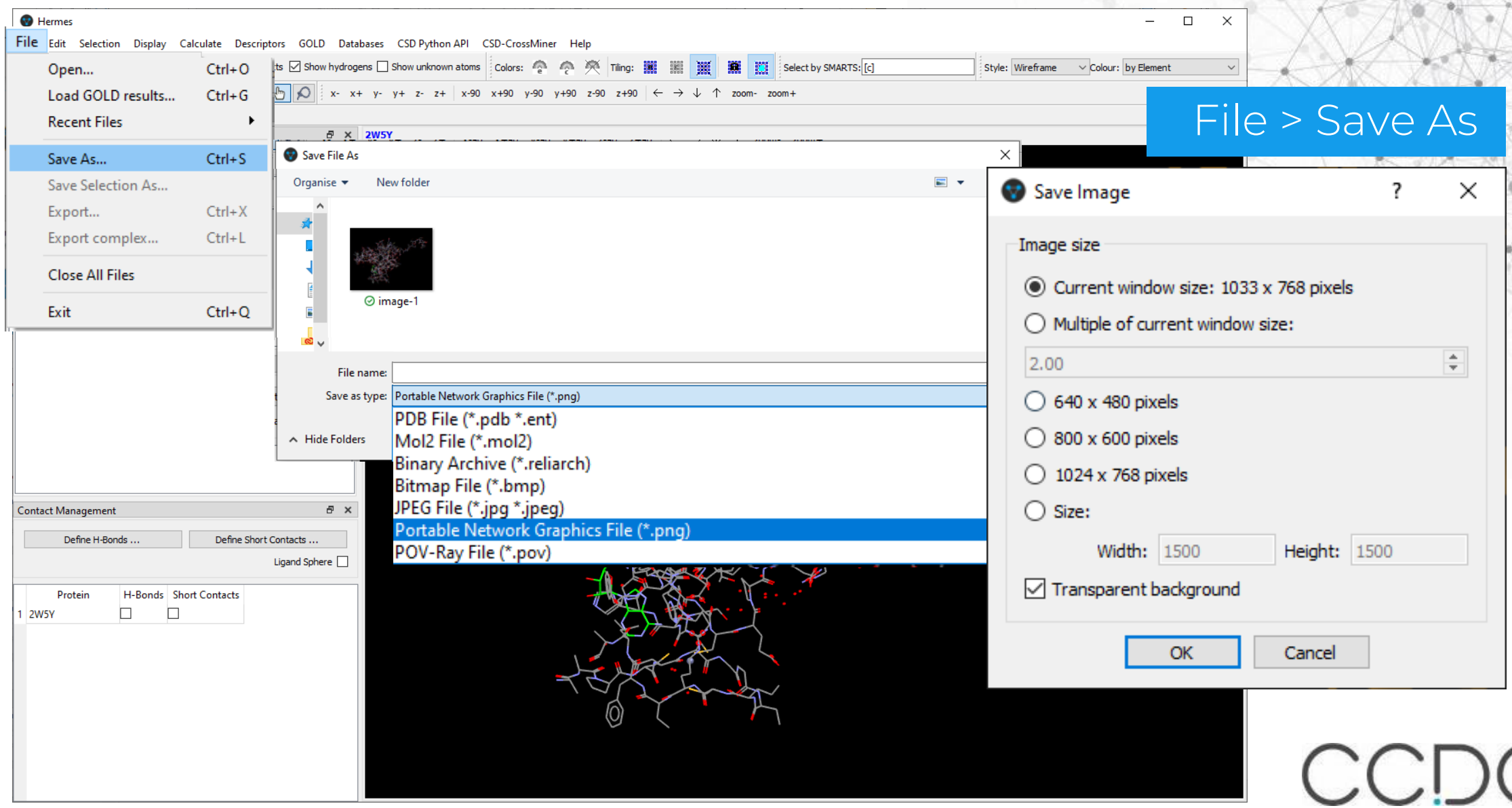
Show One: Loading structures into Hermes

From a file: File > Open

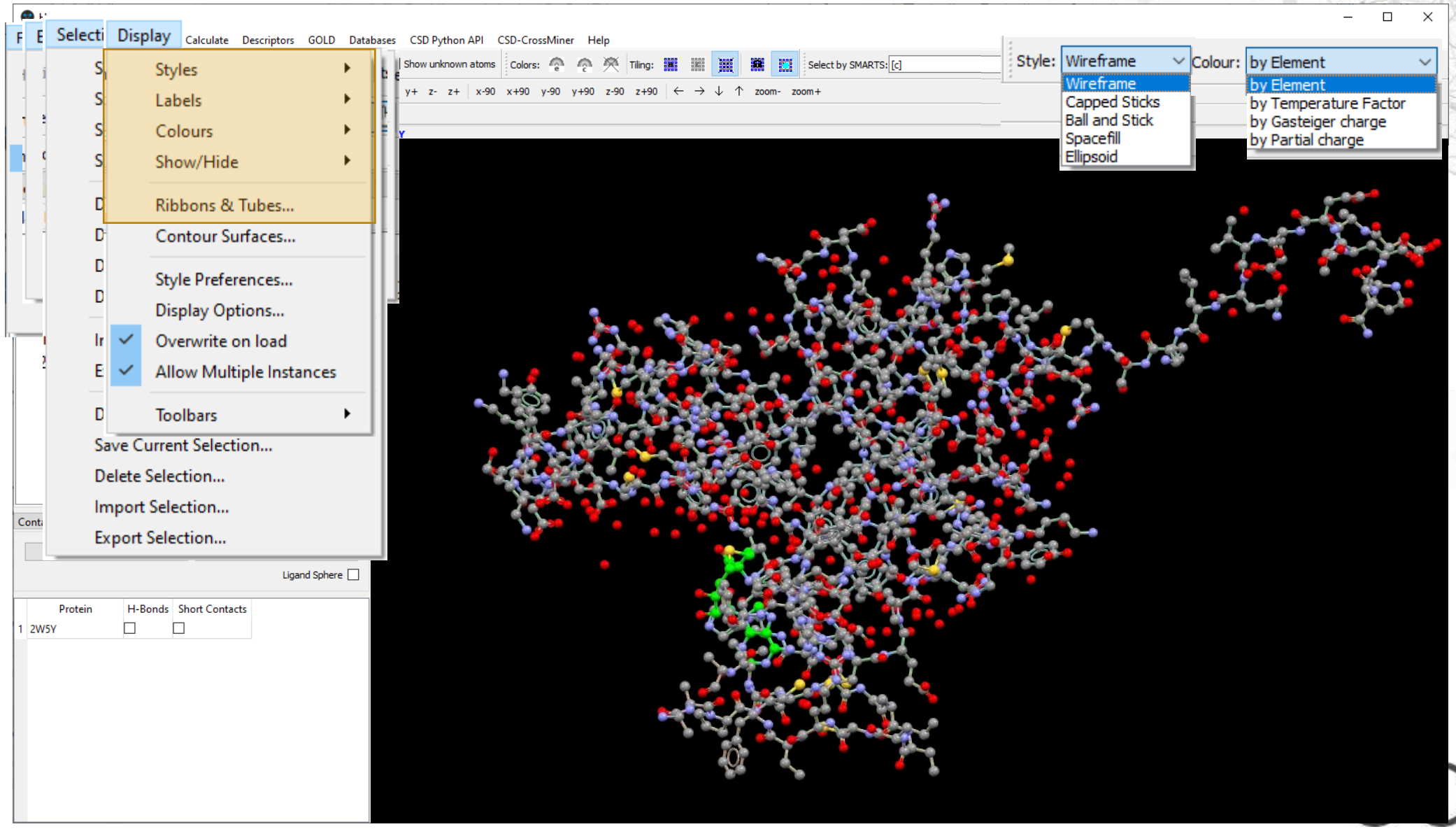
From a DB or SMILES: CSD Python API > Import



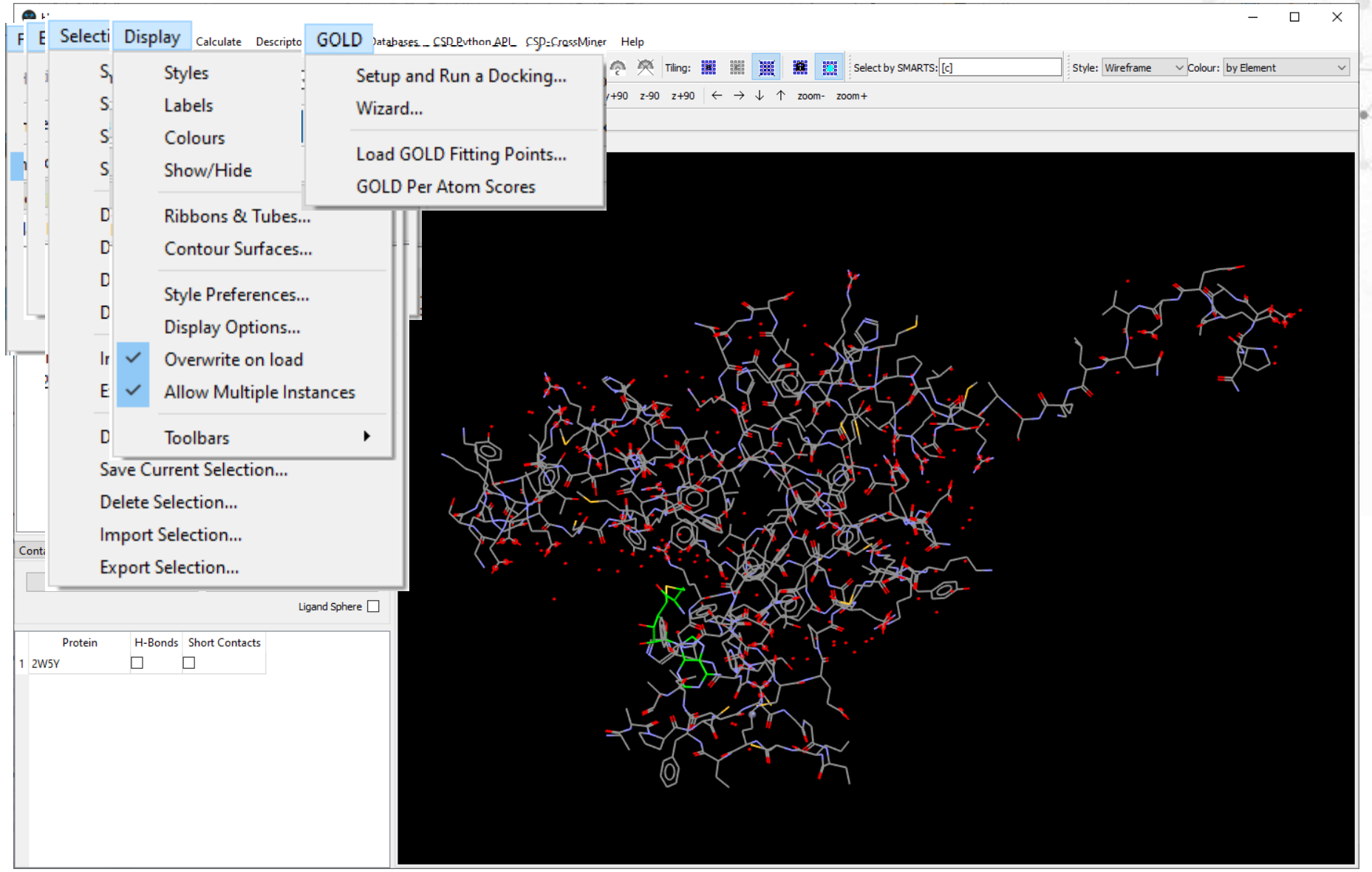
Show One: Hermes interface - Saving images



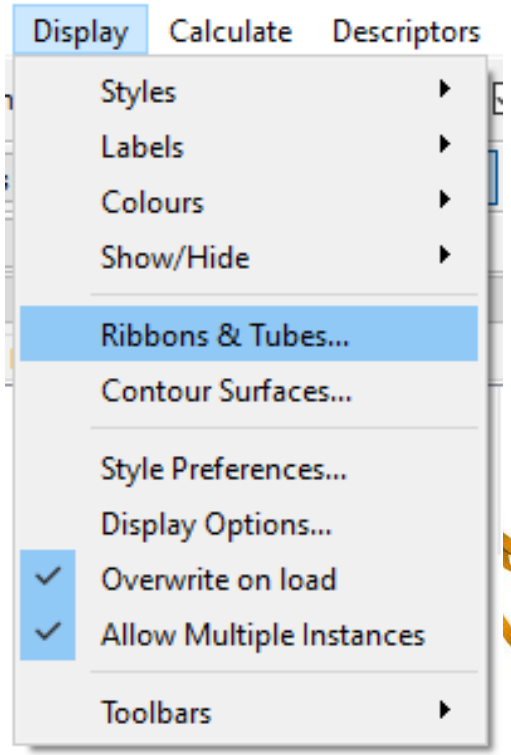
Show One: Hermes interface

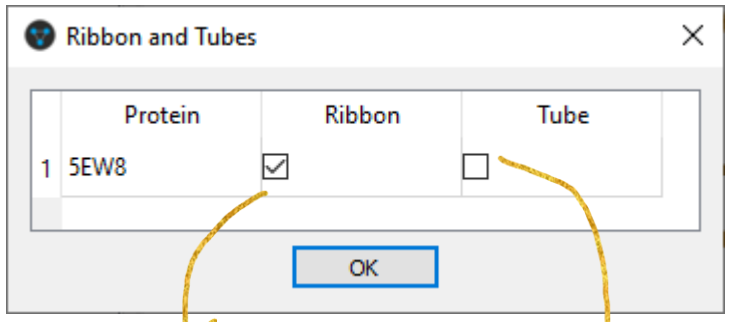


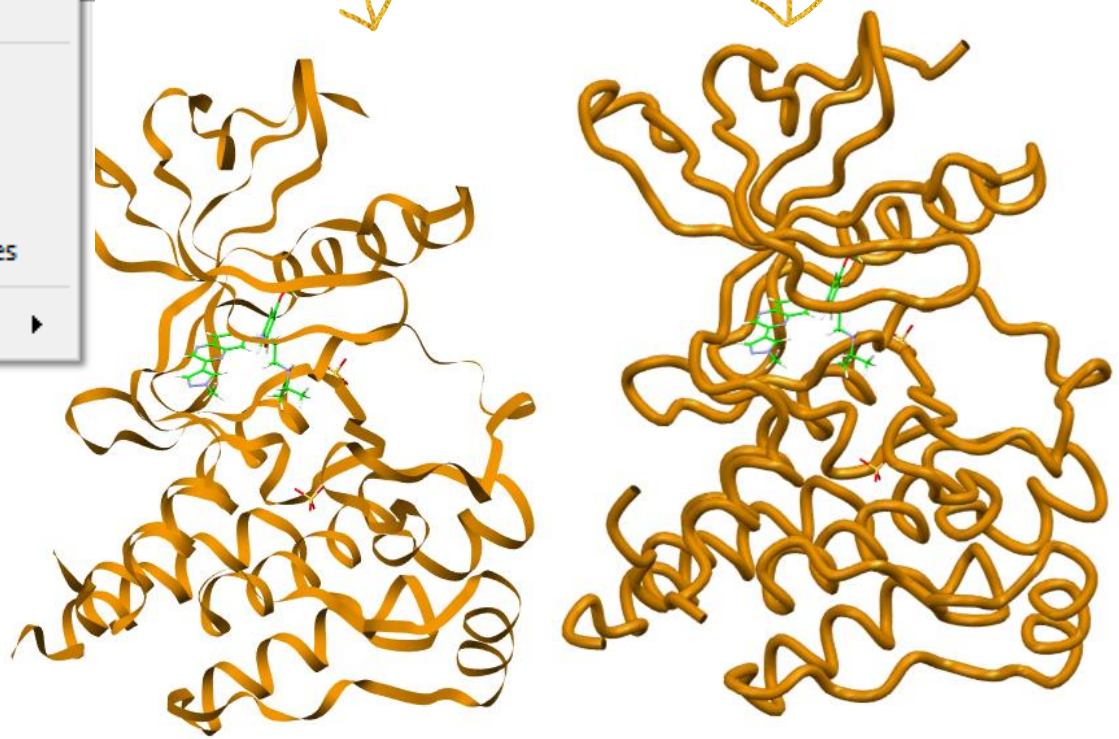
Show One: Hermes interface



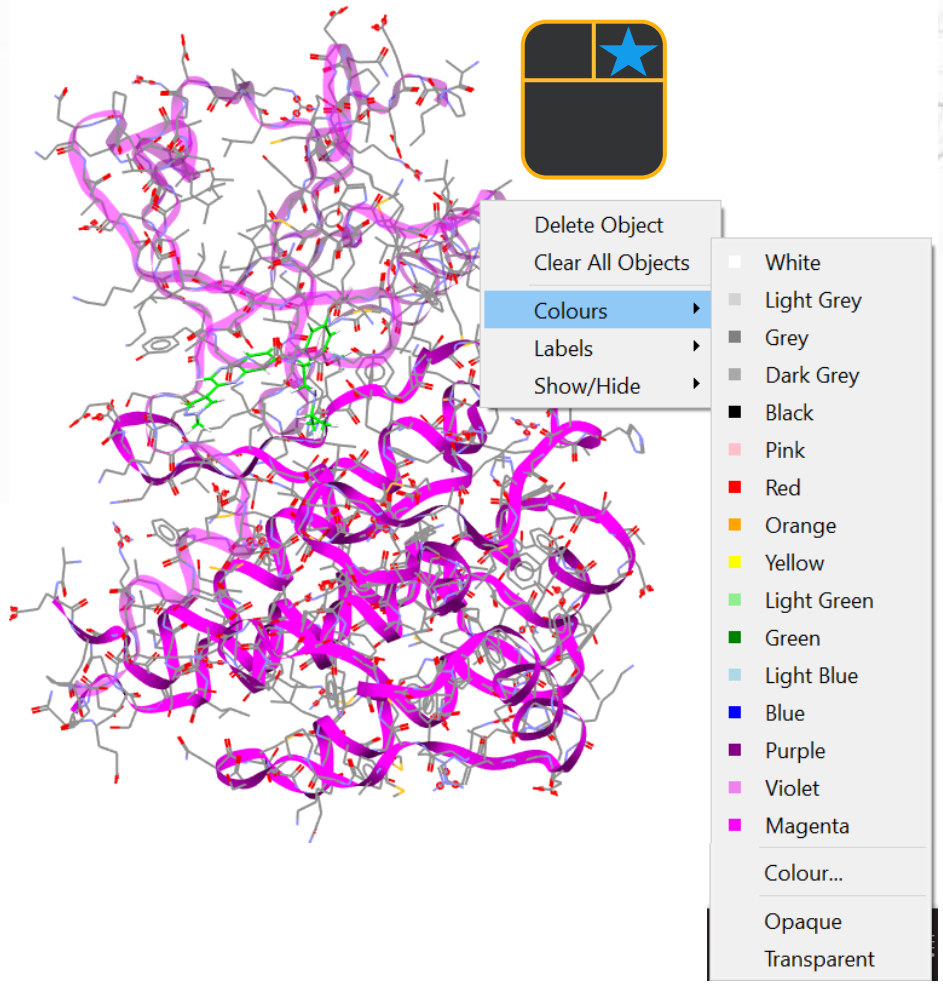
Show One: Ribbon & Tubes options







Colours & Transparency



GOLD: Complete molecular docking suite

Pose prediction

Validate your ligand docking results and optimise hits to leads

Highly configurable constraints

Use your existing knowledge of the system to bias results and focus on known features and behaviours

Multiple scoring functions

Score and rescore to build a full picture of your system or perform consensus scoring

Flexible docking

Perform ensemble docking or handle flexible side-chains with soft potentials

Water handling

Assess how structural waters affect binding, see if the ligand displaces waters or mediates the interaction during docking

CCDC

GOLD: Complete molecular docking suite

Virtual screening

Unlimited potential with virtual screening powered by cloud or cluster (HPC).

Python API access

Run dockings programmatically - for parameter optimisation and workflow incorporation.

KNIME component

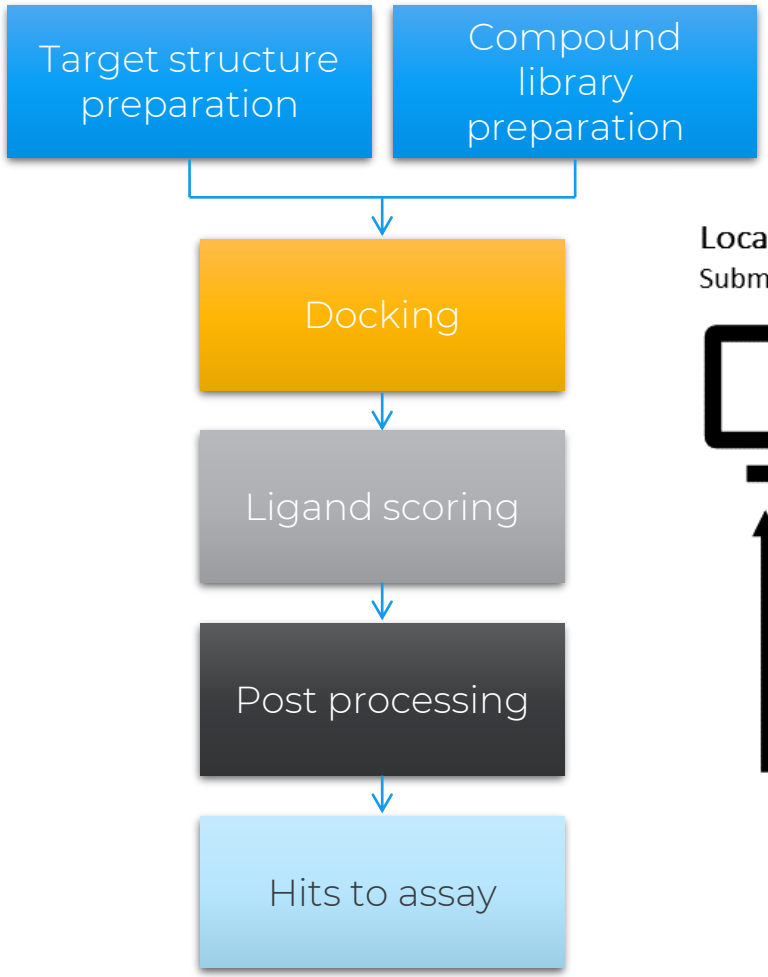
Perform protein-ligand docking in the KNIME interface to easily build into pipelines of work.

Covalent docking

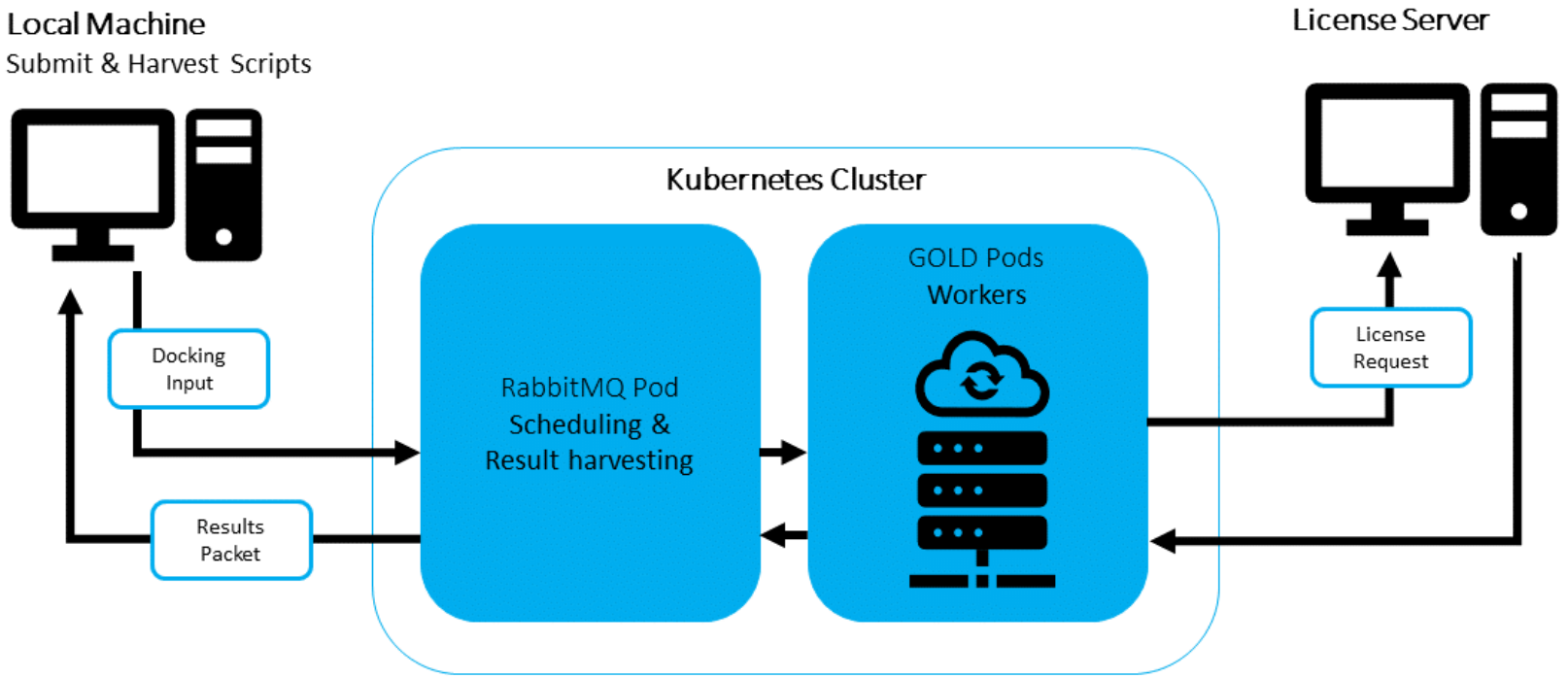
Understand irreversible binding with covalent docking to explore cancer, immunology and infectious disease targets.

CCDC

Docking virtual screening (VS) with GOLD



Typical workflow of a docking-based VS



Gold High Performance Computing

CCDC

CSD-Discovery overview

GOLD: Protein-ligand docking and virtual screening

CSD-CrossMiner: Interrogate the CSD and the PDB for common interaction patterns

Ligand-based **virtual screening** workflow to find new hits.

CSD-Conformer Generator: Generation of molecular conformations.

CSD-Ligand Overlay: Flexible alignment of ligands.

CSD Python API: Create CSD-driven analyses and workflows.

Script-based interfaces to the **Field-Based Ligand Screener**; protein **cavity and sub-pocket search** and comparison; **protein-ligand substructure & interaction pattern mining.**

SuperStar: Analyse, predict and understand protein and ligand interactions

CCDC

https://www.ccdc.cam.ac.uk/Community/educationalresources/

Want to explore more?

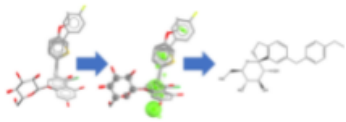
Training and Educational Resources

The wealth of information contained within the Cambridge Structural Database (CSD) extends far beyond a collection of crystal structures. Knowledge derived from these materials ch of chemistry, biochemistry, and biology. Chemical and structural concepts are often difficult to grasp without real world, interactive examples for students to explore.

and our colleagues continually produce educational materials for use in classroom and computer lab settings, or as independent study modules. Many of these materials f the Teaching Subset - a freely available set of over 750 structures that can be investigated with the free version of our Mercury visualisation and analysis program. Of database of over one million entries are available for free through our Access Structures portal.

n educator looking for supplementary teaching materials, find out more about the Teaching Database **here**. If you have developed your own modules using the CSD and o share them with the broader community, please contact us at education@ccdc.cam.ac.uk.

to date with the latest news from education and outreach at the CCDC, sign up for the Education and Outreach Newsletter **here**.

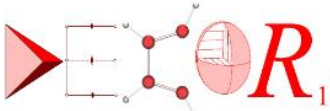


CSD-Discovery

For pharmaceutical and agrochemical researchers, tools for discovering new molecules and performing protein docking studies.

Information on the Teaching Subset

Access a series of teaching modules for use in the classroom



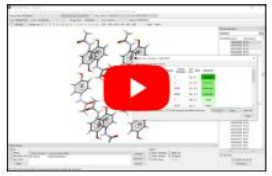
DECOR: Educational Resources for Teaching Crystallography

Register for E&O newsletter

On-demand modules with completion certificate

Self-guided workshops

Download a series of self-guided workshop materials for CCDC tools and features

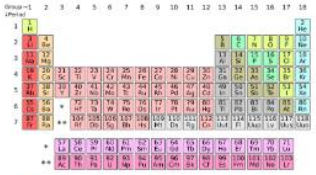


Watch software training and support videos

CSDU modules - Explore our on-demand training courses

YouTube and LabTube channels

Access fun science activities for kids through the CCDC Home learning page



Explore the Periodic Table through Crystal Structures

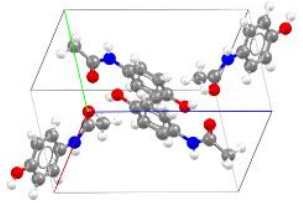
Bound! a Protein-Drug matching card game

A new docking card game

CSDU

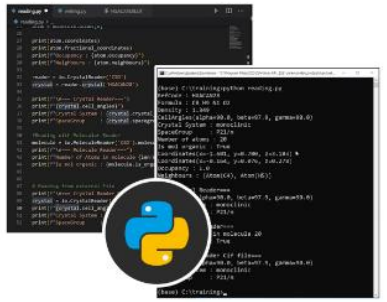
- On-demand modules to learn how to use the CSD Software at your own pace.
- Completion certificate after a final quiz!

Visualisation 101 - Visualising structural chemistry data with Mercury



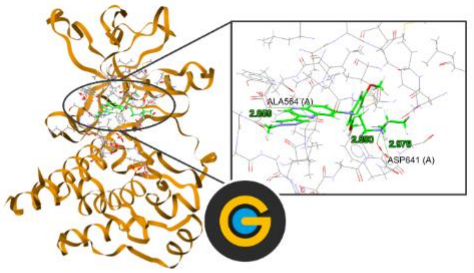
Begin module

Programmatic access to the CSD 101 – CSD Python API



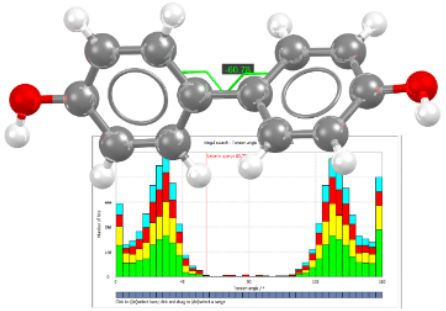
Begin module

Protein-ligand docking 101 - Running a simulation in GOLD



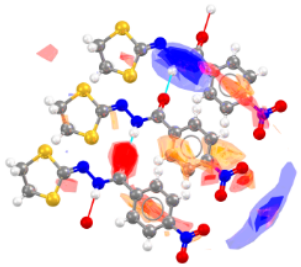
Begin module

Analysing molecular geometries 101 - basics of Mogul



Begin module

Analysing intermolecular interactions 101 - Full Interaction Maps



Begin module

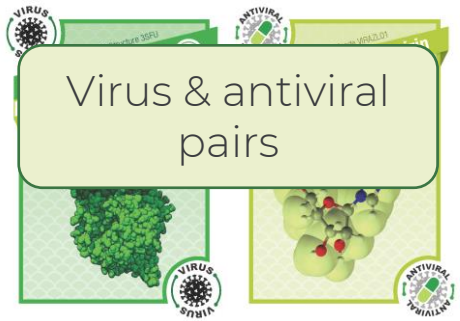
https://www.ccdc.cam.ac.uk/Community/educationalresources/CSDU/

CCDC

Bound! A Protein-Drug binding game

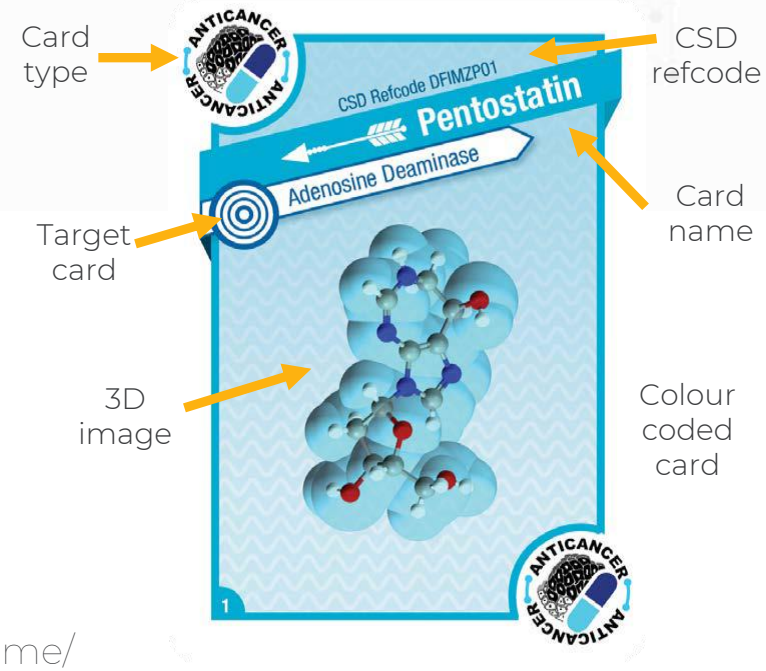
Leveraging structural data for educational purposes

For ages 12 and up



Rule cards:

- Introducing protein-drug binding and the cards.
- Game play instructions with bonus games.
- Linked to web resources.



https://www.ccdc.cam.ac.uk/Community/educationalresources/protein-ligand-card-game/

CCDC Engagement Grants

Engage, inspire, share!

- Do you wish to inspire others with your love of crystallography or structural science?
- CCDC Engagement Grants cover costs for you to produce resources or activities to increase schools and public engagement in crystallography and structural science.

https://www.ccdc.cam.ac.uk/Community/awardsandsponsorship/ccdc-engagement-grants/

CCDC