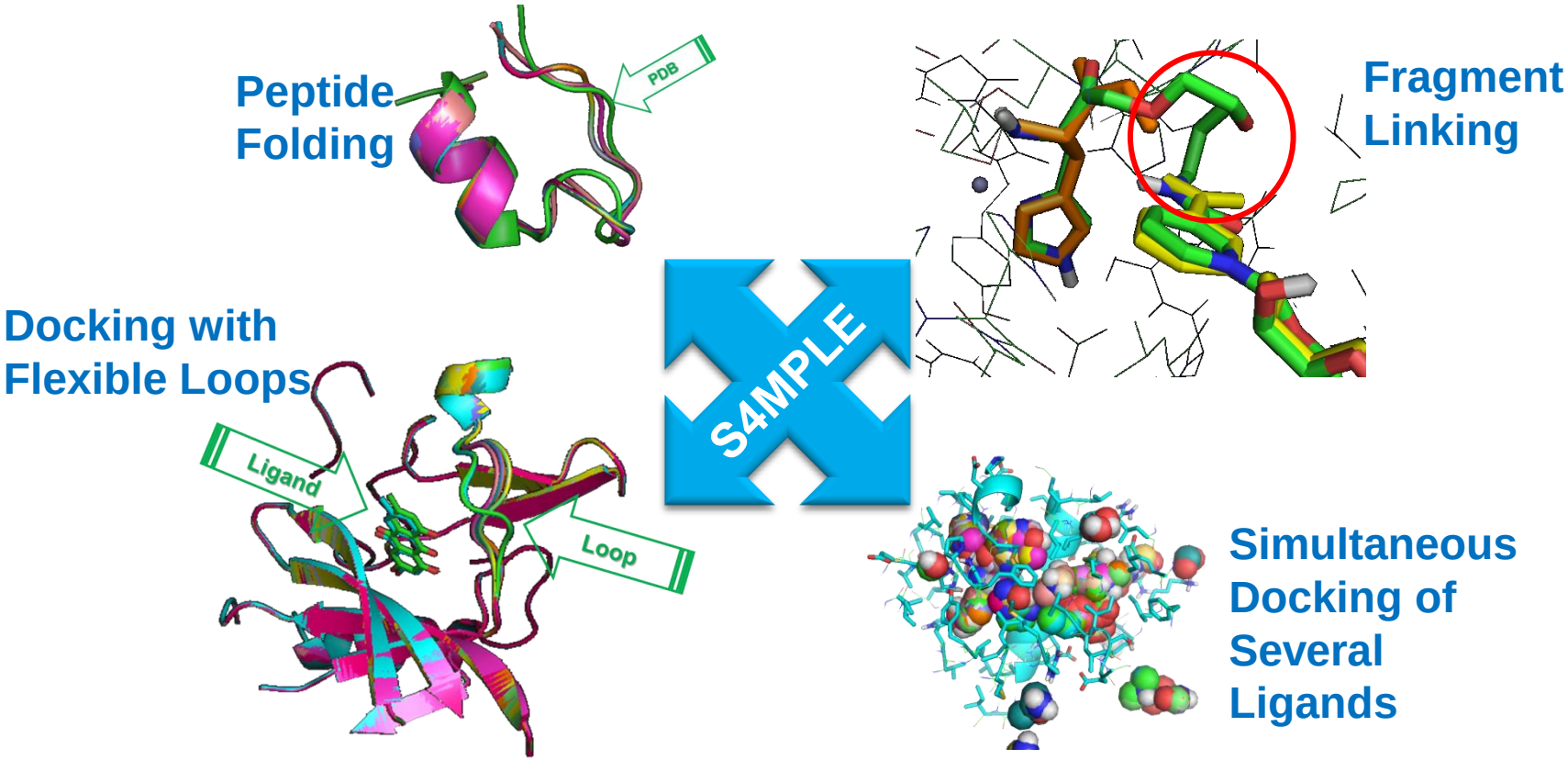
# S4MPLE – Sampler For Multiple Protein-Ligand Entities

Dragos Horvath, Laurent Hoffer, Gilles Marcou, Alexandre Varnek
Laboratoire de Chemoinformatique UMR 7140 CNRS/Université de Strasbourg
**dhorvath@unistra.fr**

# The Concept…

- Folding, Docking, Protein Loop Modeling, *etc.* – all these are particular instances of Conformational Sampling.
  - Address the GENERAL problem in order to make progress on each particular issue… or any combination thereof!



# The Issues…

- Finding the relevant minima of the energy landscape, function of molecular geometry, is a NP-hard problem.
  - Nature-inspired, computer grid-deployed sampling heuristics: Genetic Algorithm enhanced (hybridized) with molecular modeling know-how
  - Use Interaction Fingerprints to control diversity of the current population of solutions.
- Sampling the wrong energy landscape is a no go… but nothing better than fallible Force Field (FF)-based methods qualify for energy estimation.
  - Fast implicit solvent model add-on to classical vacuum force field (AMBER/GAFF)
  - Add-on terms need *fitting…*

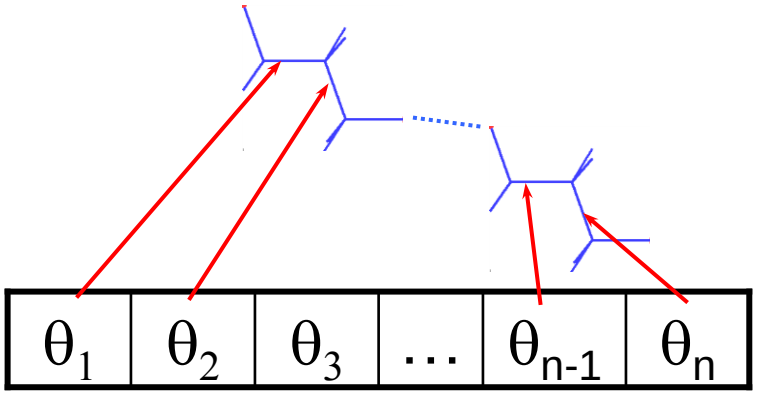
# This Presentation…

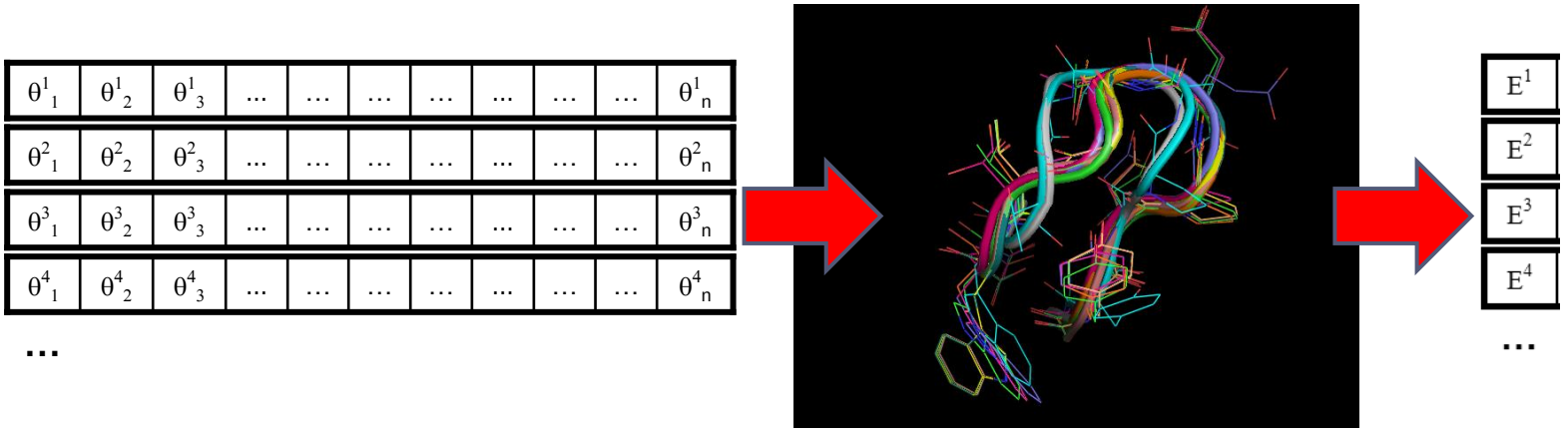
- How does S4MPLE work?
  - The underlying Hybrid Genetic Algorithm & GRID Deployment Scheme – Past, Present & Future.
- Force Field Issues.
  - Force field: classical and additional terms
- Benchmarks & Applications:
  - Treatment of large-scale protein site flexibility.
  - Redocking, Cross-docking
  - Key water-mediated interaction predcition
  - S4MPLE in fragment-based drug design (L. Hoffer, NovAlix).

# S4MPLE – a Genetic Algorithm in Torsional (Old) and Cartesian (New) Coordinate Space

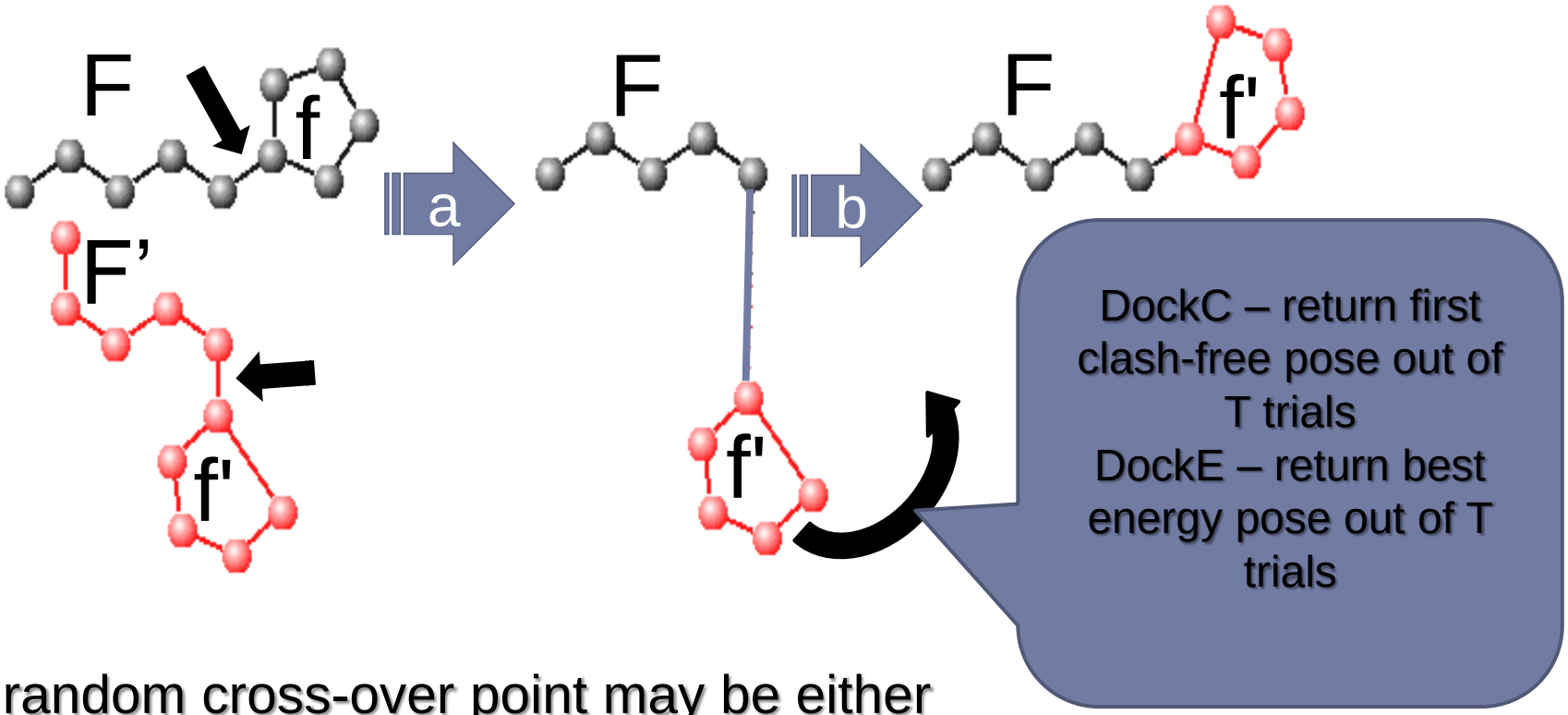
- Apply a Darwinian Evolution Scenario to a population of vectors ("chromosomes" ***C***), encoding points in the variable space of the "Fitness" function *f(**C**)* to maximize.

« individual »
or
« chromosome »
=
list of its torsional angles



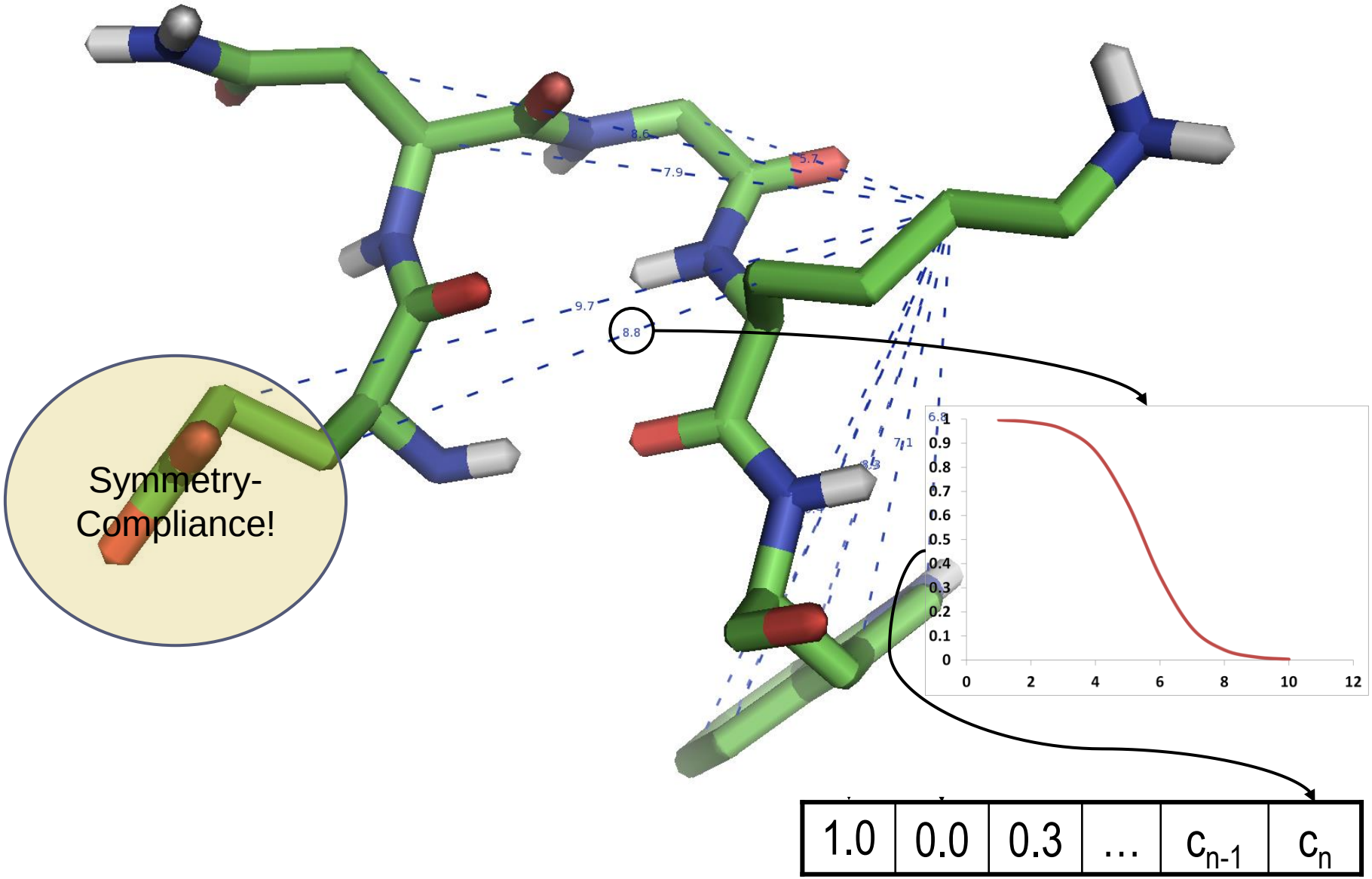



# The S4MPLE Core Trick: Connectivity-Independent Genetic Operators




- The random cross-over point may be either
  - a covalent bond, as shown above, or
  - a putatively favorable inter-molecular contact (H-bond, hydrophobic).
- S4MPLE makes no formal distinction between inter/intra-molecular degrees of freedom
  - … thus, it makes no distinction between sampling and (multi-species) docking

# Smooth, differentiable Interaction Fingerprints



# Diversity Control needed to avoid premature convergence

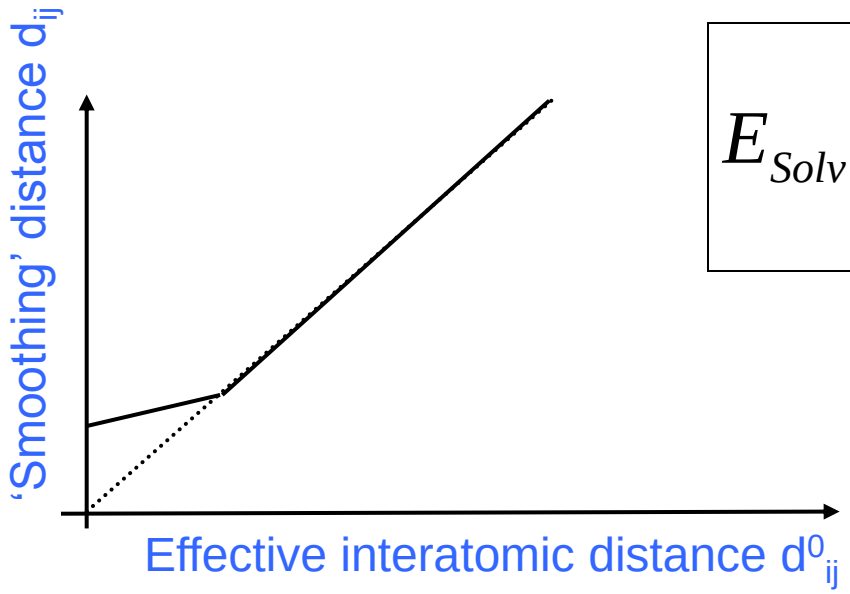
- Dissimilarity of two individuals I and J is given by the Hamming distance H between their contact fingerprint vectors $\mathbf{c}^I$ and $\mathbf{c}^J$. J is *redundant* if I is fitter and $H(\mathbf{c}^I,\mathbf{c}^J) < minfpdiff$.
- Two diversity control operators were defined
  - ***ReplaceMostRedundant*** – lets current individual enter the population instead of its most redundant member.
  - ***RandomizeMostRedundant*** – used once every generation, to avoid accumulation of redundant individuals (these are replaced by randomized individuals)
- Multi-Island Model
  - run many almost-independent simulations with smaller populations in parallel (allow occasional chromosome exchanges).

# This Presentation…

- How does S4MPLE work?
  - The underlying Hybrid Genetic Algorithm & GRID Deployment Scheme – Past, Present & Future.
- Force Field Issues.
  - Force field: classical and additional terms
- Benchmarks & Applications:
  - Treatment of large-scale protein site flexibility.
  - Redocking, Cross-docking
  - Key water-mediated interaction prediction
  - S4MPLE in fragment-based drug design (L. Hoffer, NovAlix).

# The Force Field Challenge

- AMBER for *proteins* and General Amber FF (GAFF) for *others*.
- … modified to avoid singularities.
- Simple Implicit Solvent Model added to classical vacuum/explic…need fittings...




$$E_{Solv} = k_{solv} \frac{Q_i^2 V_j + Q_j^2 V_i}{d_{i,j}^4} - k_h \sigma^{hphob}(i,j)$$

$$E_{coulomb} = \frac{Q_i Q_j}{4\pi\varepsilon_{dd} * d_{ij}^2}$$

# This Presentation…

- How does S4MPLE work?
  - The underlying Hybrid Genetic Algorithm & GRID Deployment Scheme – Past, Present & Future.
- Force Field Issues.
  - Force field: classical and additional terms
- Benchmarks & Applications:
  - Treatment of large-scale protein site flexibility.
  - Redocking, Cross-docking
  - Key water-mediated interaction prediction
  - S4MPLE in fragment-based drug design (L. Hoffer, NovAlix).

# Classical Benchmarking: Rigid Redocking of Astex Complexes.

| | Success Rate (%) Top ranked-pose | Success Rate (%) Saved poses |
|---|---|---|
| S4MPLE (Core FF) | 77 | 94 |
| S4MPLE (Fit FF) | 84 | 98 |
| GOLD (Goldscore) (1) | 75 | Unavailable |
| Plants (ChemPLP) (2) | 87 | 97 |
| Plants (PLP) (2) | 84 | |
| LGA “LargeAll” (3) | 63 | Unavailable |

- State-of-the-art results… *but, unlike the other programs, S4MPLE employs no fitted scoring function (other than the FF energy)*

*(1) Hartshorn, JMedChem, 50, 726-741 (2007)*
*(2) Korb et al, JChemInfModel, 49, 84–96 (2009)*
*(3) Fuhrmann et al, JComputChem, 31, 1911–1918 (2010)*

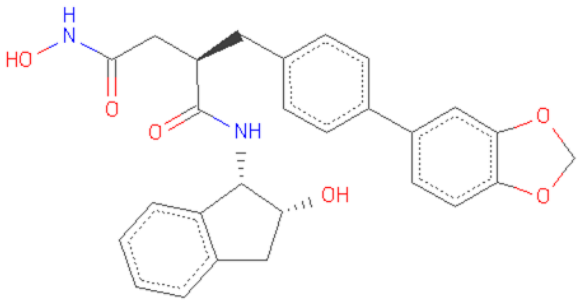
# Flexible Cross-Docking

- Human ADAMTS-5 (several ligand-protein PDB complexes available)
- Different conformations of the binding site (“induced fit” from the ligand : rearrangements of backbone and/or several side chains)

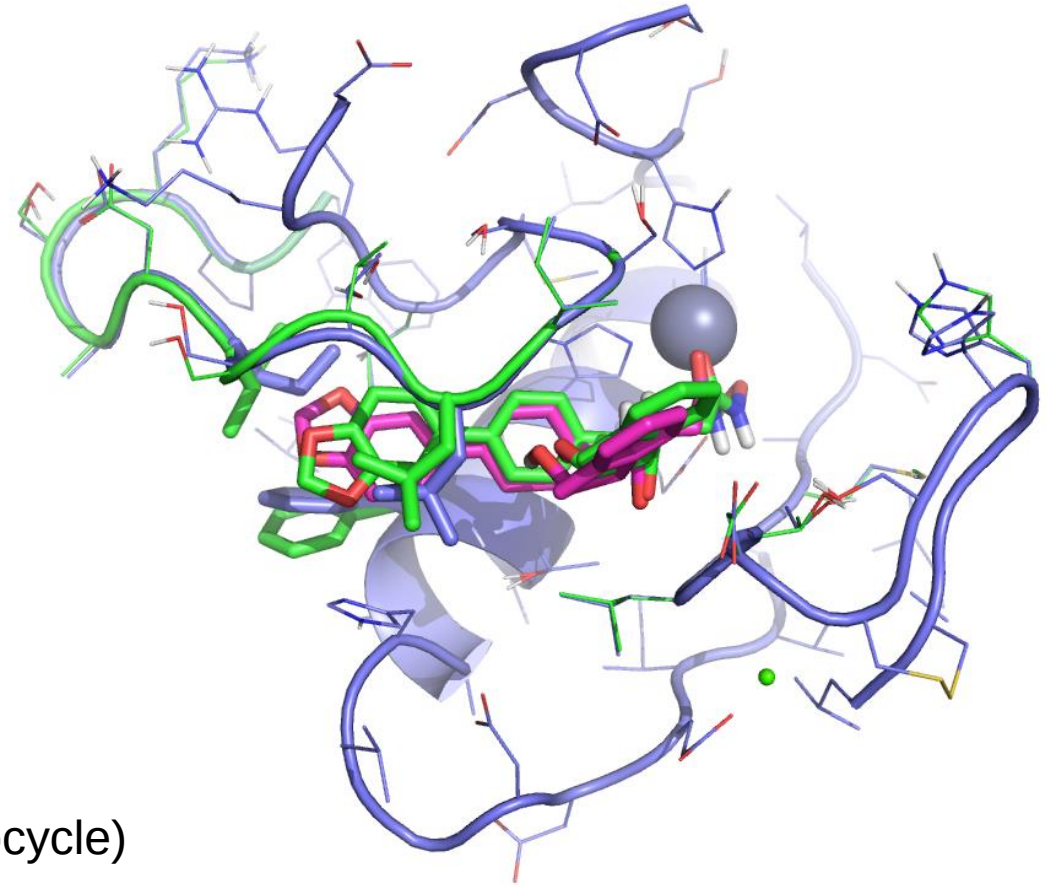
| Rigid Mode (RMSD) | | Ligands | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2RJQ | 3B8Z | 3HY7 | 3HY9 | 3HYG | 3LJT |
| Binding Site | 2RJQ | 1,7 | 4,9 | 0,9 | 1,4 | 1,4 | 7,0 |
| | 3B8Z | 2,1 | 0,3 | 2,1 | 2,3 | 2,4 | 2,7 |
| | 3HY7 | 1,6 | 2,7 | 0,6 | 1,1 | 1,4 | 3,6 |
| | 3HY9 | 1,8 | 3,3 | 1,9 | 1,2 | 1,3 | 4,2 |
| | 3HYG | 4,2 | 2,5 | 2,1 | 1,0 | 0,9 | 3,5 |
| | 3LJT | 1,3 | 1,3 | 1,9 | 1,7 | 1,4 | 0,5 |

| Flexible Mode (RMSD) | | Ligands | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2RJQ | 3B8Z | 3HY7 | 3HY9 | 3HYG | 3LJT |
| Binding Site | 2RJQ | 1,6 | 1,4 | 0,6 | 1,6 | 1,6 | 2,7 |
| | 3B8Z | 1,9 | 0,4 | 1,0 | 2,7 | 2,3 | 2,5 |
| | 3HY7 | 2,7 | 0,7 | 0,7 | 2,2 | 1,6 | 2,4 |
| | 3HY9 | 2,1 | 0,7 | 0,7 | 1,5 | 1,5 | 1,7 |
| | 3HYG | 1,9 | 0,8 | 0,7 | 1,0 | 1,5 | 1,5 |
| | 3LJT | 1,5 | 1,1 | 1,8 | 1,2 | 1,6 | 1,3 |

# Flexible Cross-Docking




Example: 3LJT ligand into the 3HYG binding site



Legend

Input binding site

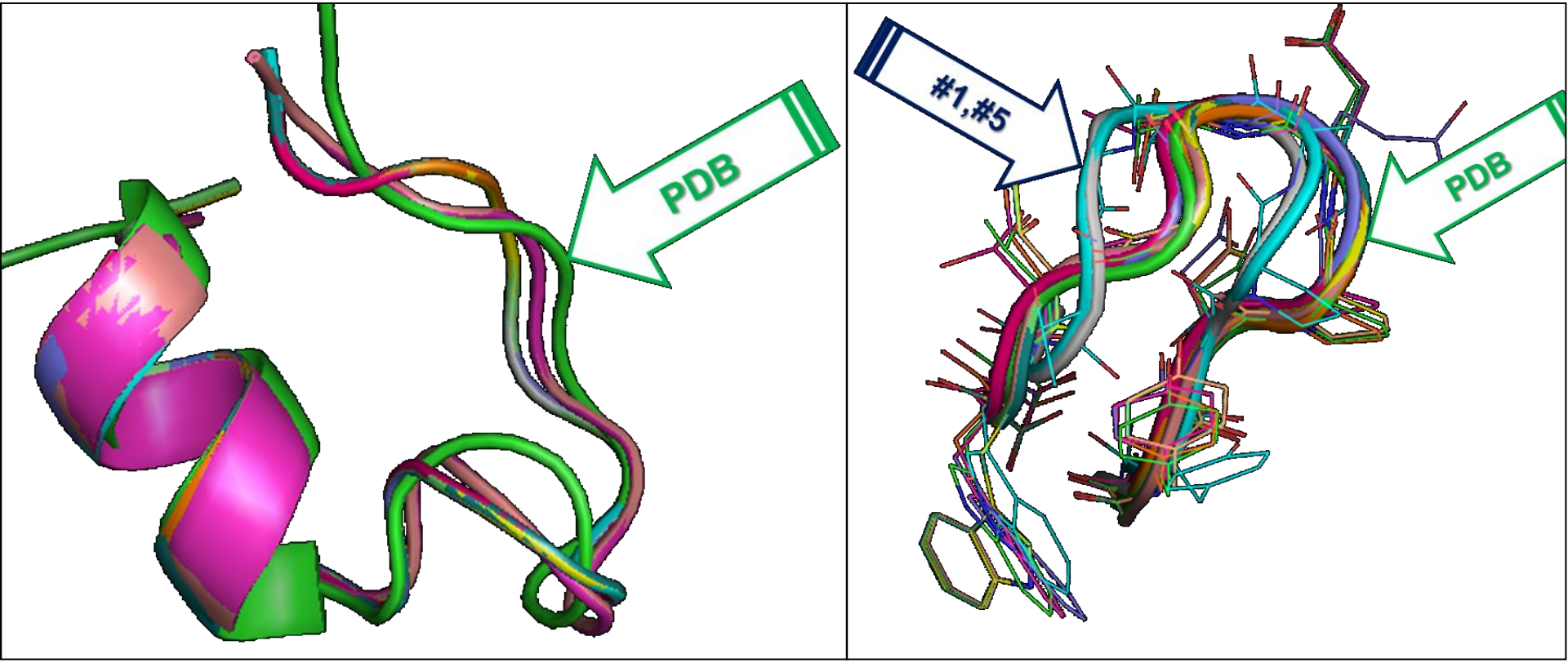
X-ray ligand

Docking output

**RMSD = 1.5 Å** (flip of heterocycle)

Flexible atoms : all around ligands (6.5 Å) ; except Zn-coordination residues

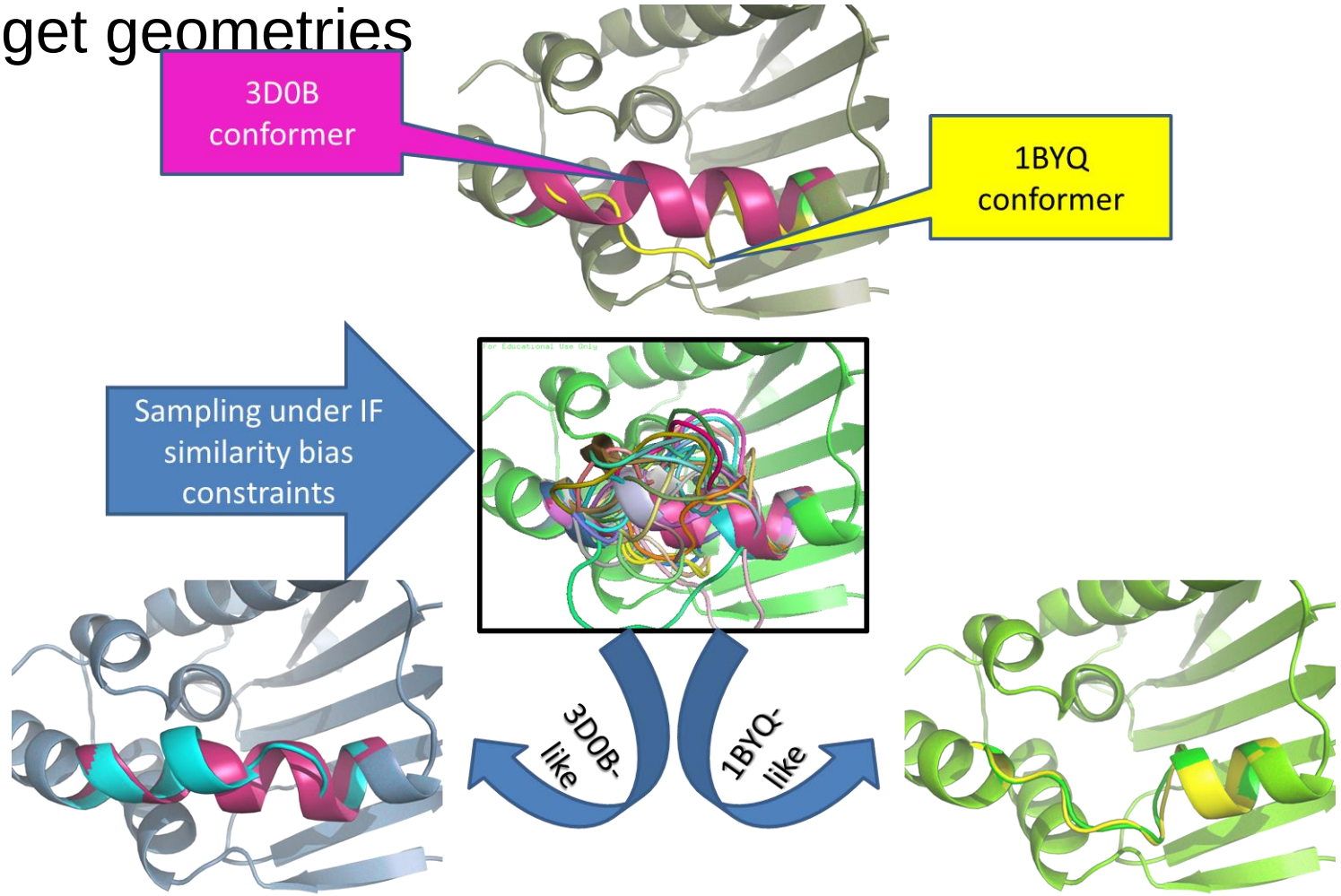
# Small Protein Folding (ongoing)




- .. more difficult than docking, because problem space is less constrained (many near-optimal unfolded geometries are possible)

# Difficult Problems Made Feasible (on a Workstation) under Constraints

- HSP90 ligand-induced restructuring – fully flexible S4MPLE version, using IF-based 'attractors' towards target geometries



# Treating crystallographic waters as additional ligands

- A few explicit waters are allowed to move freely, being modeled as additional ligands (but do not appear or disappear).
- Like the site and the organic ligand, they are embedded in the same continuum solvent reaction field…
- Will they be positioned at ligand-site bridging interaction spots?
- Four protocols were studied for these simulations:
  A. Default S4MPLE settings (2000 generations)
  B. Default S4MPLE setting, but longer runs (5000 generations)
  C. Protocol A, with a bias favoring ligand-water contacts
  D. Protocol C, followed by an explicit water molecule position optimization, at fixed organic ligand pose

# Water challenge: Typical Examples

Examples obtained with protocol D:

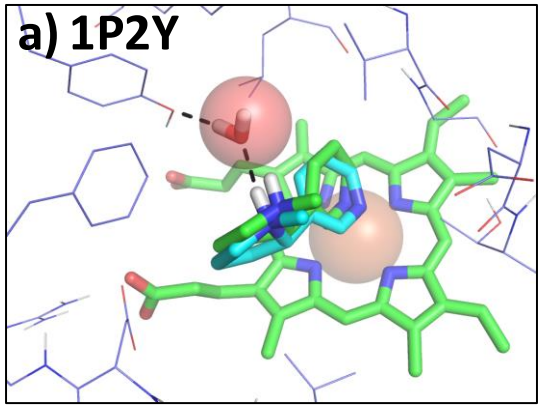
1P2Y: correct prediction of both ligand and waters in top pose

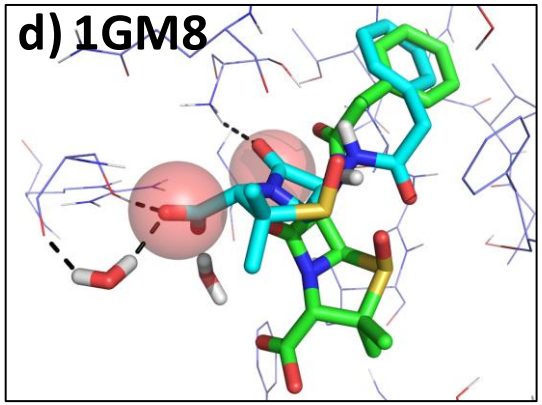
1LPZ: correct geometry found in suboptimal, saved poses.

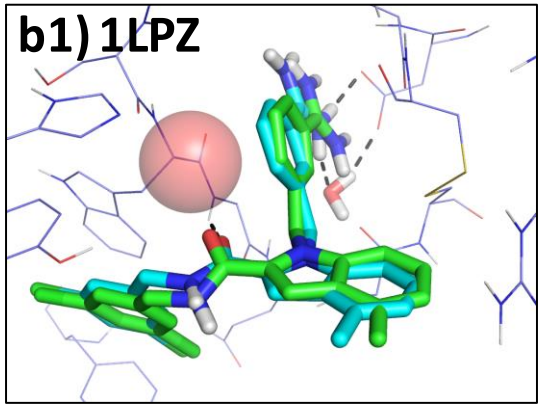
1G9V: partial success (correct positioning of ligand + at least one water)

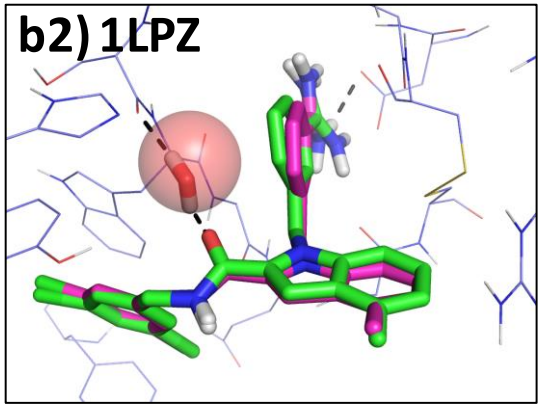
1GM8: no acceptable solution.

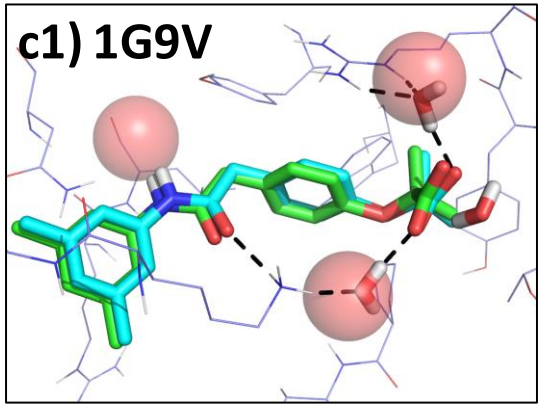
Experimental ligand conformer is green, spheres are expected water locations. Predictions are in blue (top-ranked) or purple (other rank), with waters as stick models.

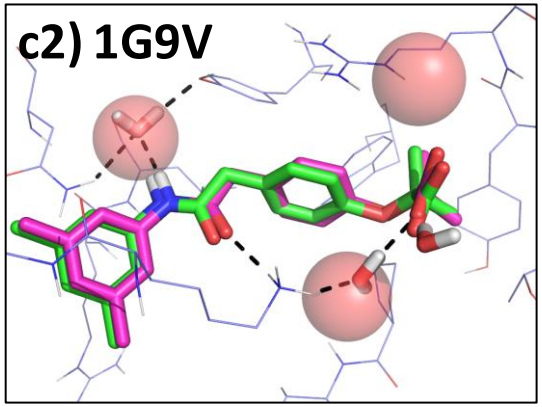


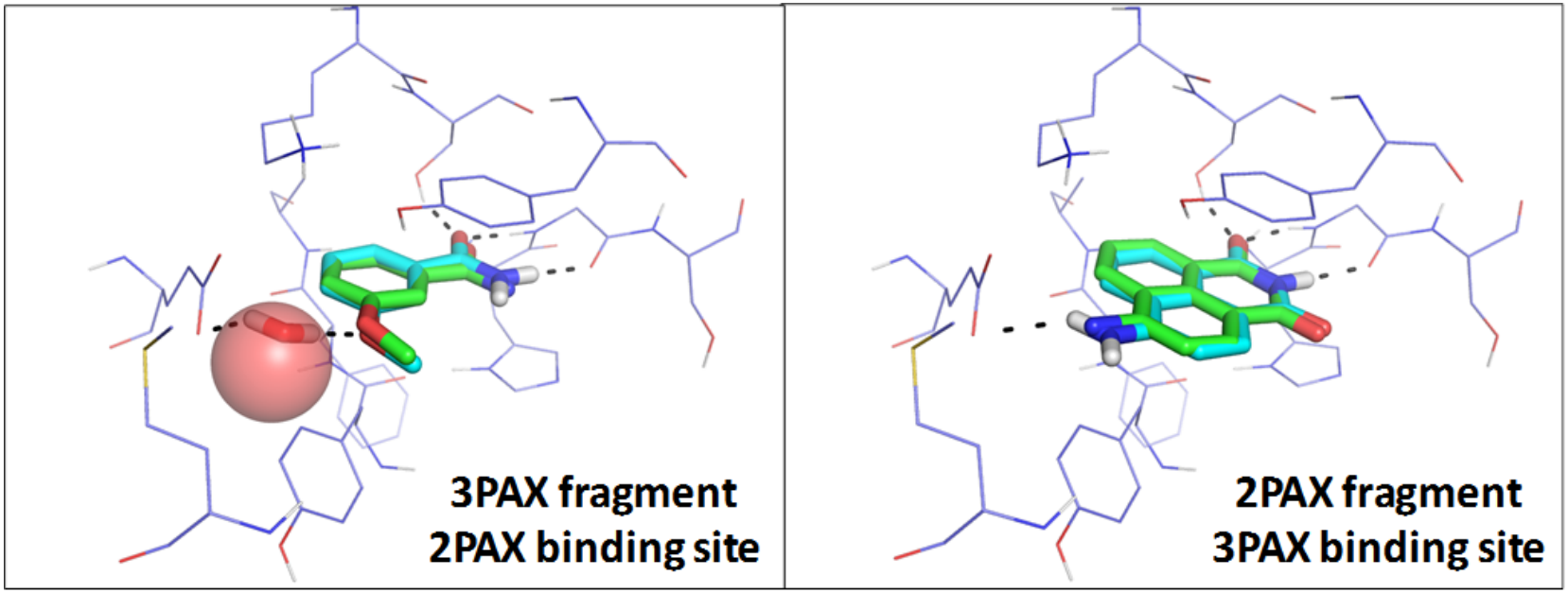



# Detecting water-mediated *vs.* direct ligand-site interactions…




Structures from cross-docking run on the PARP target. Two experimental X-ray structures are reported. In 3PAX, the bound fragment features a water-mediated interaction, while in 2PAX the larger ligand is designed such as to replace this water by its $–NH_2$ group.

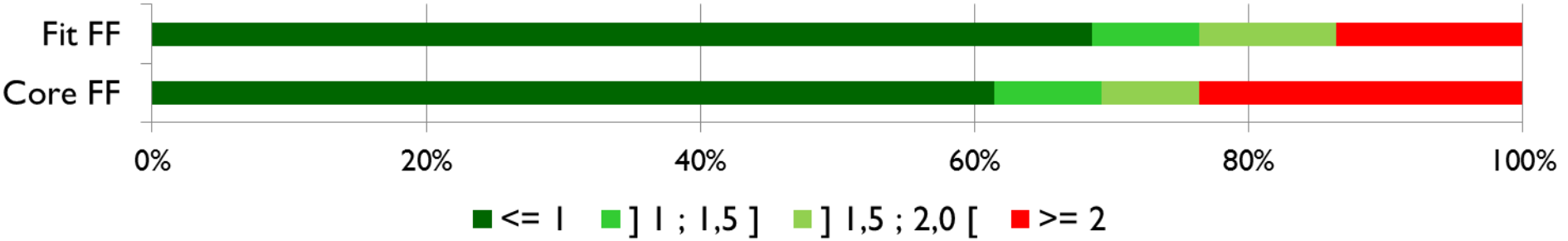
Each ligand was docked into the site of the other structure, in presence of an additional water. This water was successfully positioned at its ligand-site bridging spot, with the 3PAX fragment, but displaced by 2PAX.

# S4MPLE in Fragment-Based Drug Design.

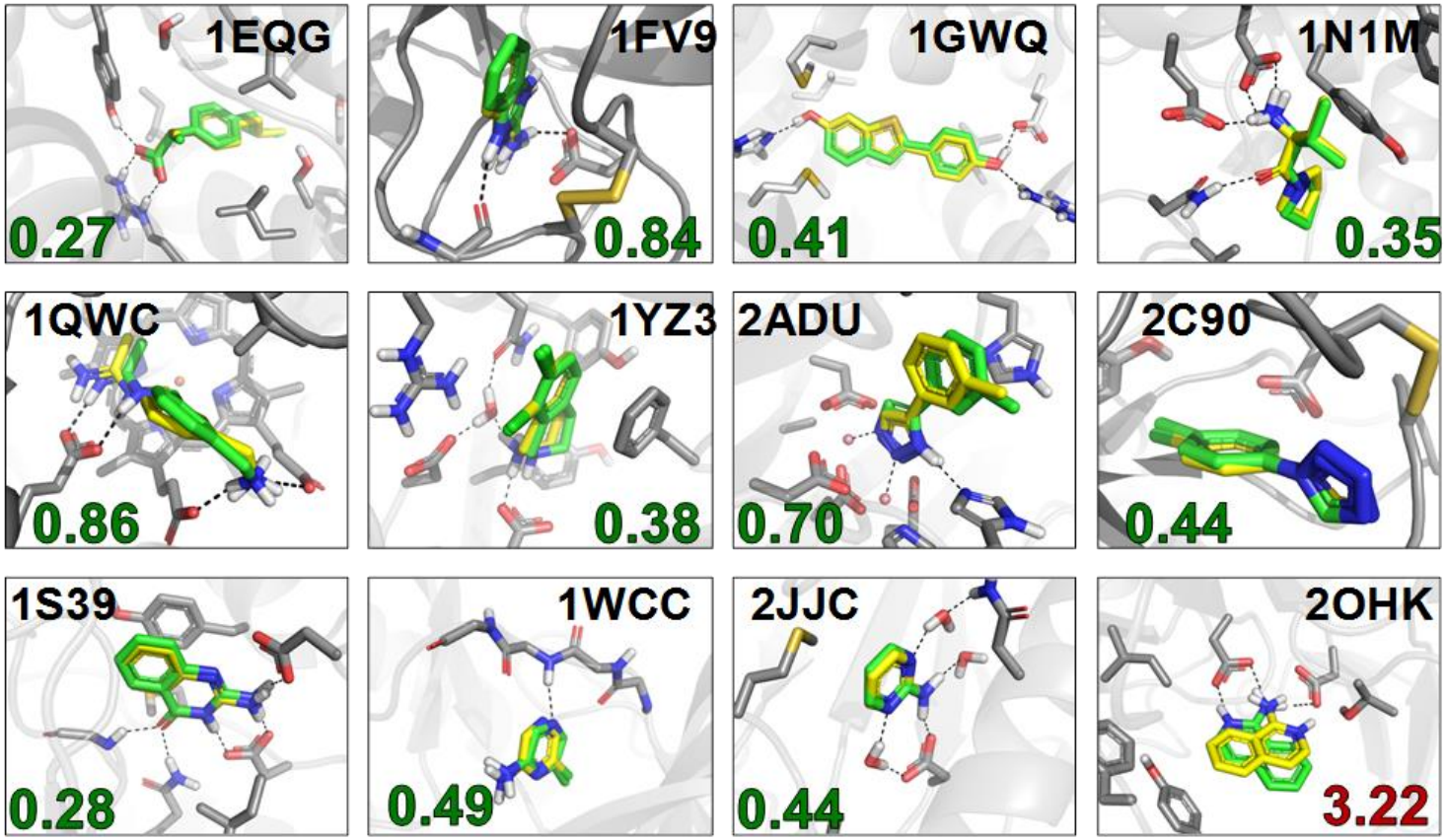
1. The Congreve Dataset – a small but well-documented set of fragment-site com

| Tool | Waters Kept Fixed in Site | Success Rate (%) Top ranked-pose |
|---|---|---|
| S4MPLE (Core FF) | 2JJC | 92 |
| S4MPLE (Fit FF) | 2JJC | 92 |
| GLIDE (1) | All around 5Å | 100 |
| MCSS (2) | 2JJC / 1YZ3 | 67 |
| MCSS (GBSA) (2) | 2JJC / 1YZ3 | 67 |
| GOLD (2) | 2JJC / 1YZ3 | 67 |
| GOLD (GBSA) (2) | 2JJC / 1YZ3 | 75 |

II. More robust – the collection of 140 fragment-like ligands in Astex/CCDC-clean

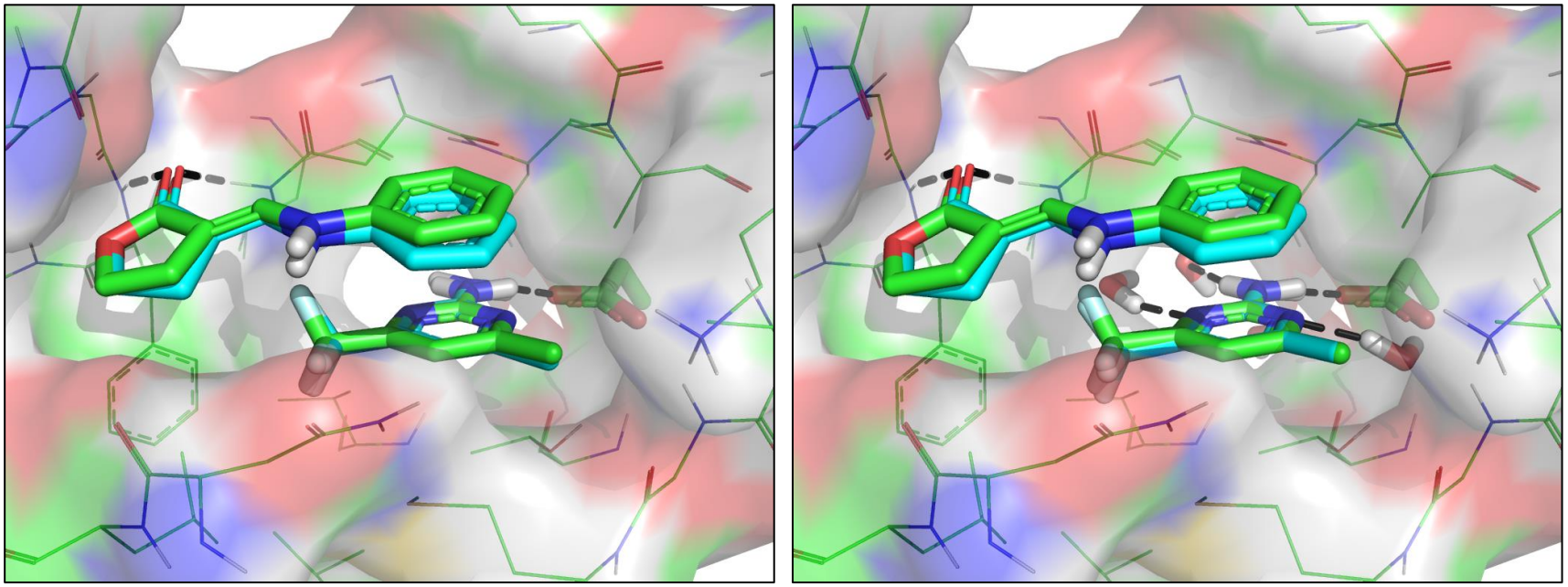



(1) Loving et al, *JCAMD,23, 541–554 (2009)*
(2) Haider et al., J. Chem. Inf. Model, in press (2011)

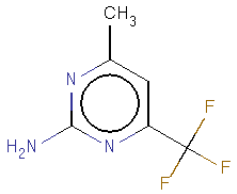
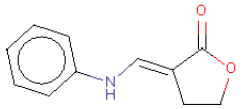
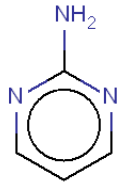
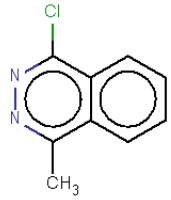
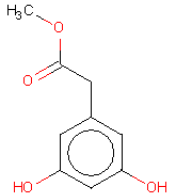
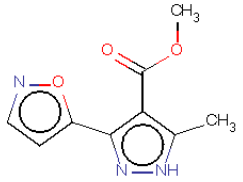
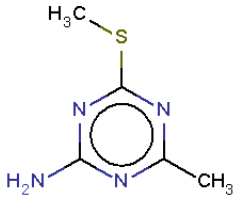
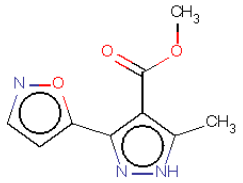
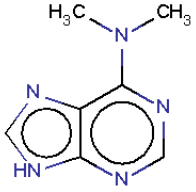
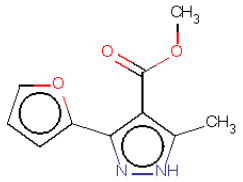


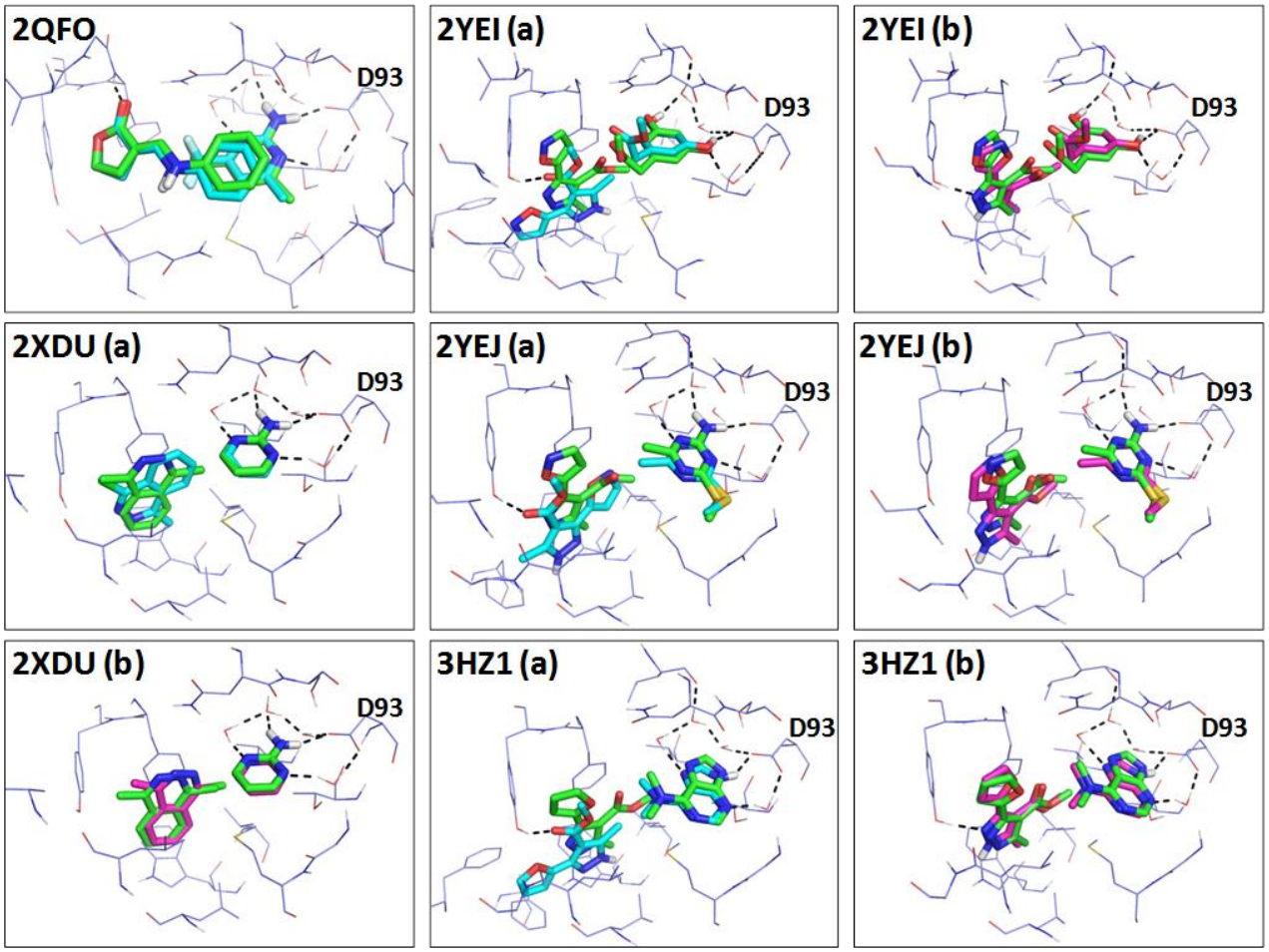

Top-ranked poses from redocking simulations of Congreve's dataset, using the Fit FF energy function (RMSD with respect to X-ray coordinates are provided for each complex). Experimental and docked locations are represented in green and yellow, respectively. Main polar interactions are depicted as dash lines.

# Fragment-Based Drug Design: Simultaneous Docking of Several Fragments



Successful redocking (Fit FF) of two fragments at the same time, in absence and respectively presence of crystallographic waters, into the binding site of HSP90 (PDB : 2QFO)

| PDB | Fragment 1 | Fragment 2 |
|---|---|---|
| 2QFO |  |  |
| 2XDU |  |  |
| 2YEI |  |  |
| 2YEJ |  |  |
| 3HZ1 |  |  |



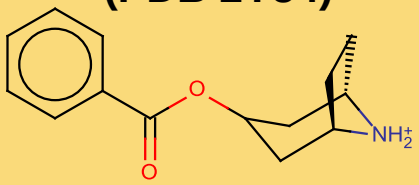

Study case: Heat Shock Potein ternary complexes.

Experimental locations are in green. Predicted poses are in blue (top-ranked) or purple (other rank). Main polar interactions are depicted as dash lines. The hot spot D93 is indicated.

# Fragment-to-Lead (F2L) Growing, accounting for protein flexibility

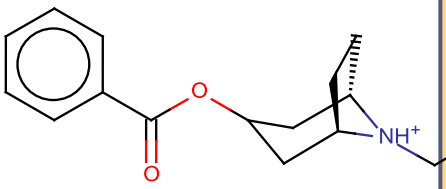
- AChBP Study Case : Experimental Data

**Anchor Fragment with $pK_i \approx 5$ (PDB 2Y54)**

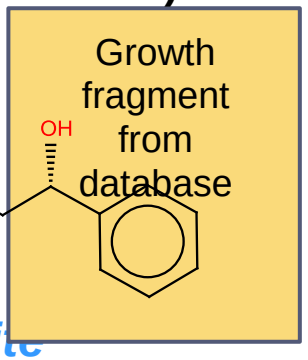


F2L

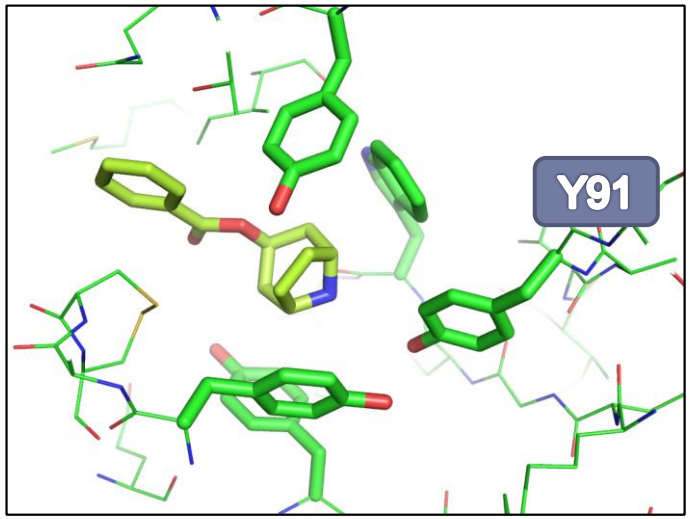
**Lead with $pK_i \approx 7$ (PDB 2Y56)**

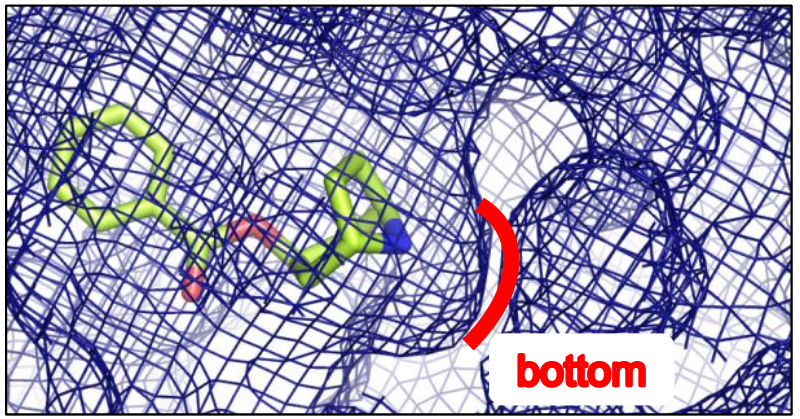


Growth fragment from database



❑ Growing-based optimization → ***conformational changes in binding site***



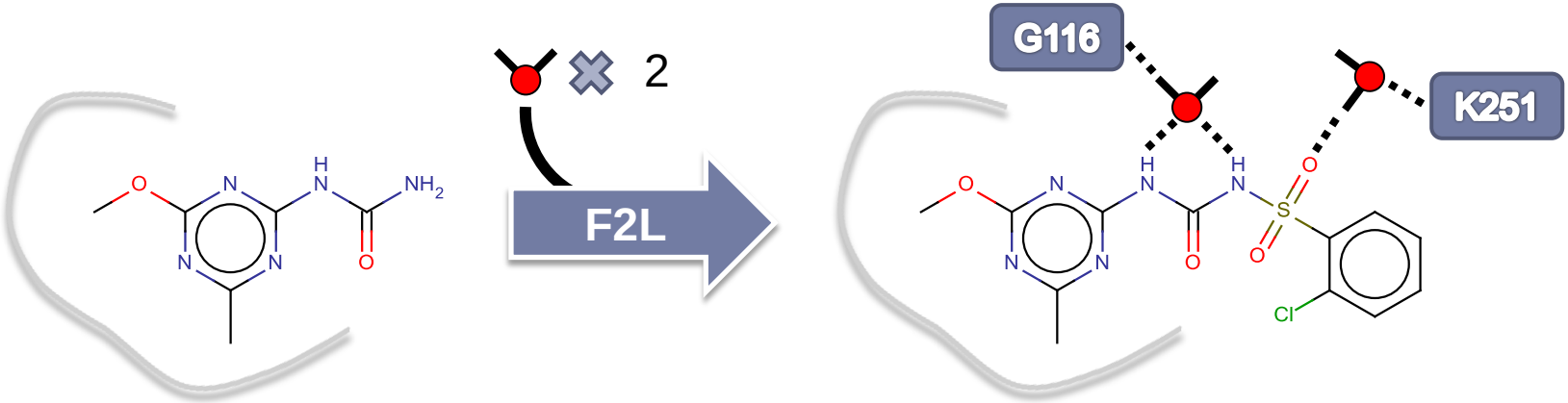

**2Y54 fragment** **2Y54 site**

*Edink et al, JACS (2011)*

# Fragment-to-Lead (F2L) Growing, accounting for water-mediated interactions

Goal : to challenge S4MPLE to ***prioritize growth partners involved in water-mediated hydrogen bonds*** with the binding site

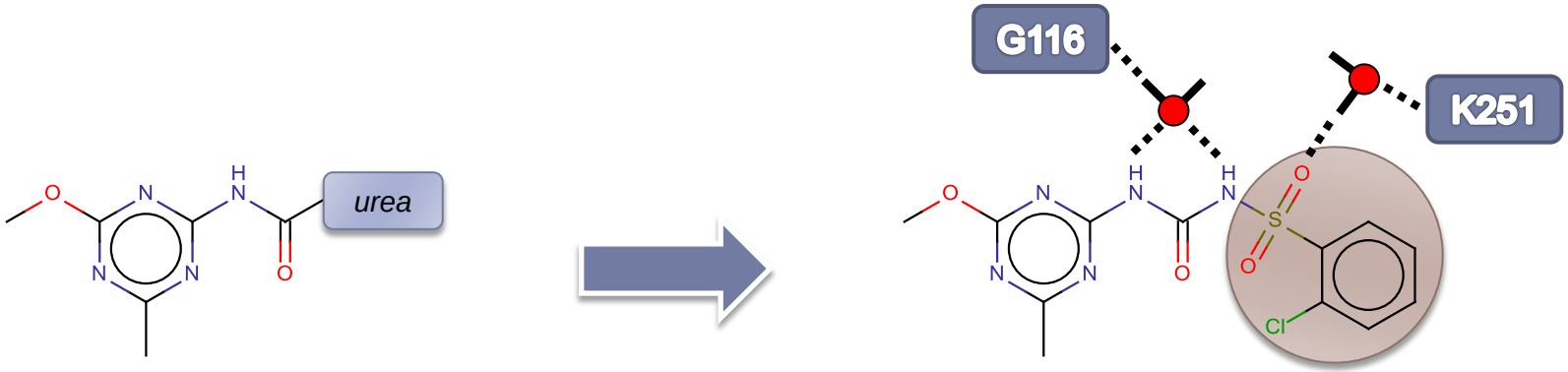
❑ ***Growing-based optimization including 2 free waters*** (PDB 1T9B)




❑ 2 different growth campaigns, (***water-free & water-included***) involving 1500+ compounds

❑ Monitored criteria of success

✓Reference ligand is automatically generated from reference fragment




✓All geometries (***ligand + waters***) are accurately reproduced

✓Fair rank in ***water-free*** simulations (***rank #62***)

✓Better rank in ***water-included*** simulations (***rank #6***)

***Mediated HB promoted expected moiety***

# Some Conclusions…

- General approach:
  - no compound-class specific hypotheses (proteins are molecules like any other!)
  - no problem-class specific hypotheses (docking is conformational sampling like any other – no *scoring functions*!)
- ☞ Vast Applicability Domain – from Folding to Docking to Multi-Entity Docking and Growing/Linking.
- However, the Force Field Fitting Battle is not won (yet?)
  - Is there an ideal, generally competent Force Field for docking & Folding? Things look fine for Docking…
  - Allowing for full flexibility makes Folding infinitely more complex…
  - Only the planned massive deployment of fully flexible S4MPLE onto GPU Grids may tell – open to international collaborations