Drug Design
Data Resource

UCSD

Continuous Evaluation of Ligand Pose Predictions (CELPP)

An Open, Rolling Blinded Prediction Challenge

ACS San Francisco
April 5th 2017

Shuai Liu
Postdoc
D3R at UCSD

Outlines

- 1, Introduction of D3R
- 2, Introduction of CELPP
- 3, CELPP workflows
- 4, How to join CELPP
- 5, Acknowledgement

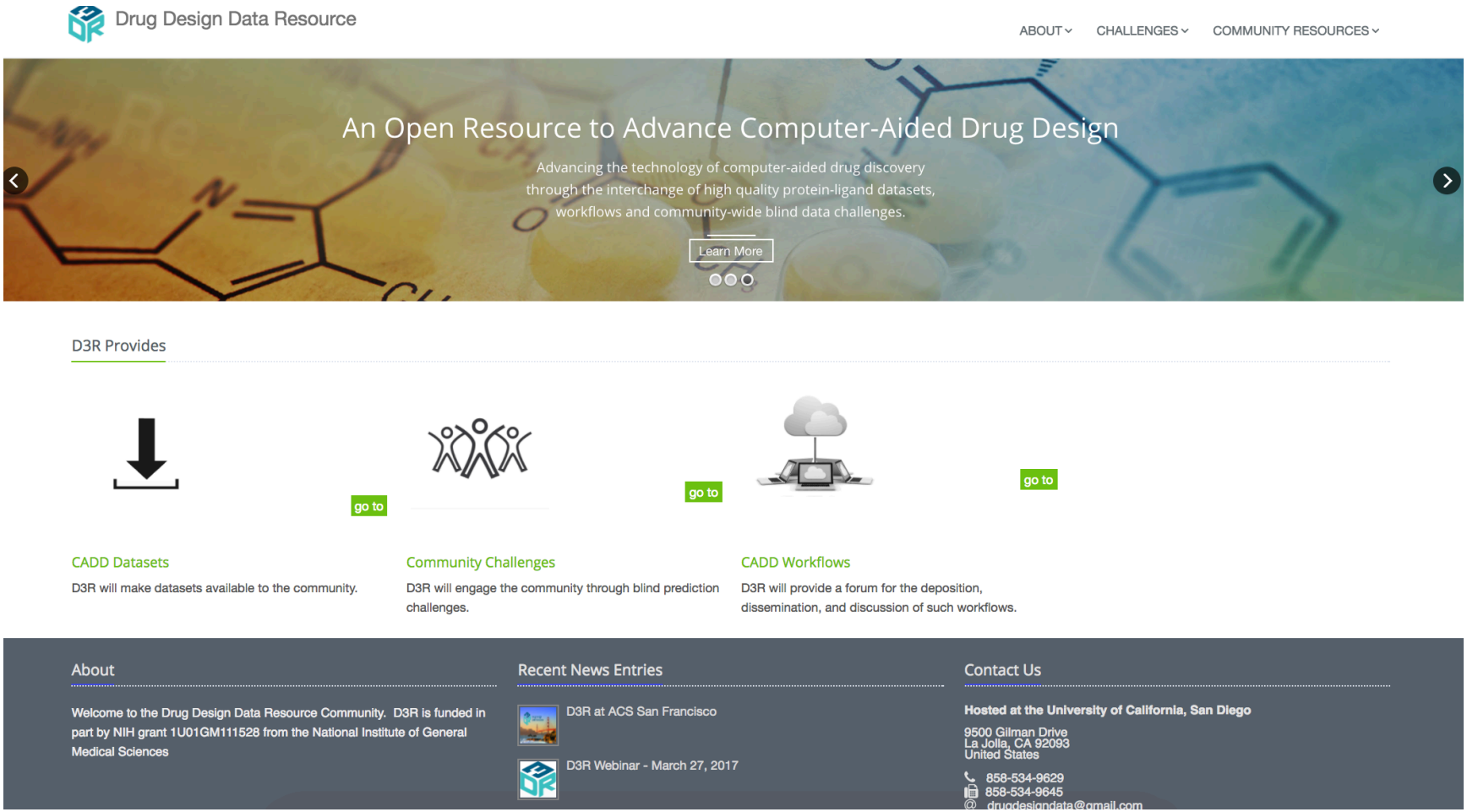
Drug Design Data Resource--D3R

- 1, D3R grand challenge: We collect data from industry and academic groups and run blind challenges for pose, binding affinity, free energy predictions.

- 2, Continuous Evaluation of Ligand Pose Prediction (CELPP) : weekly cross-docking challenge.

Drug Design Data Resource--D3R

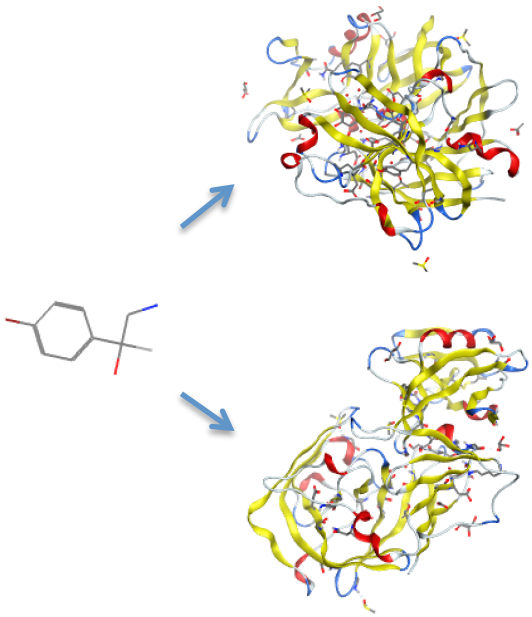
drugdesigndata.org/



CELPP Motivation

Friday night

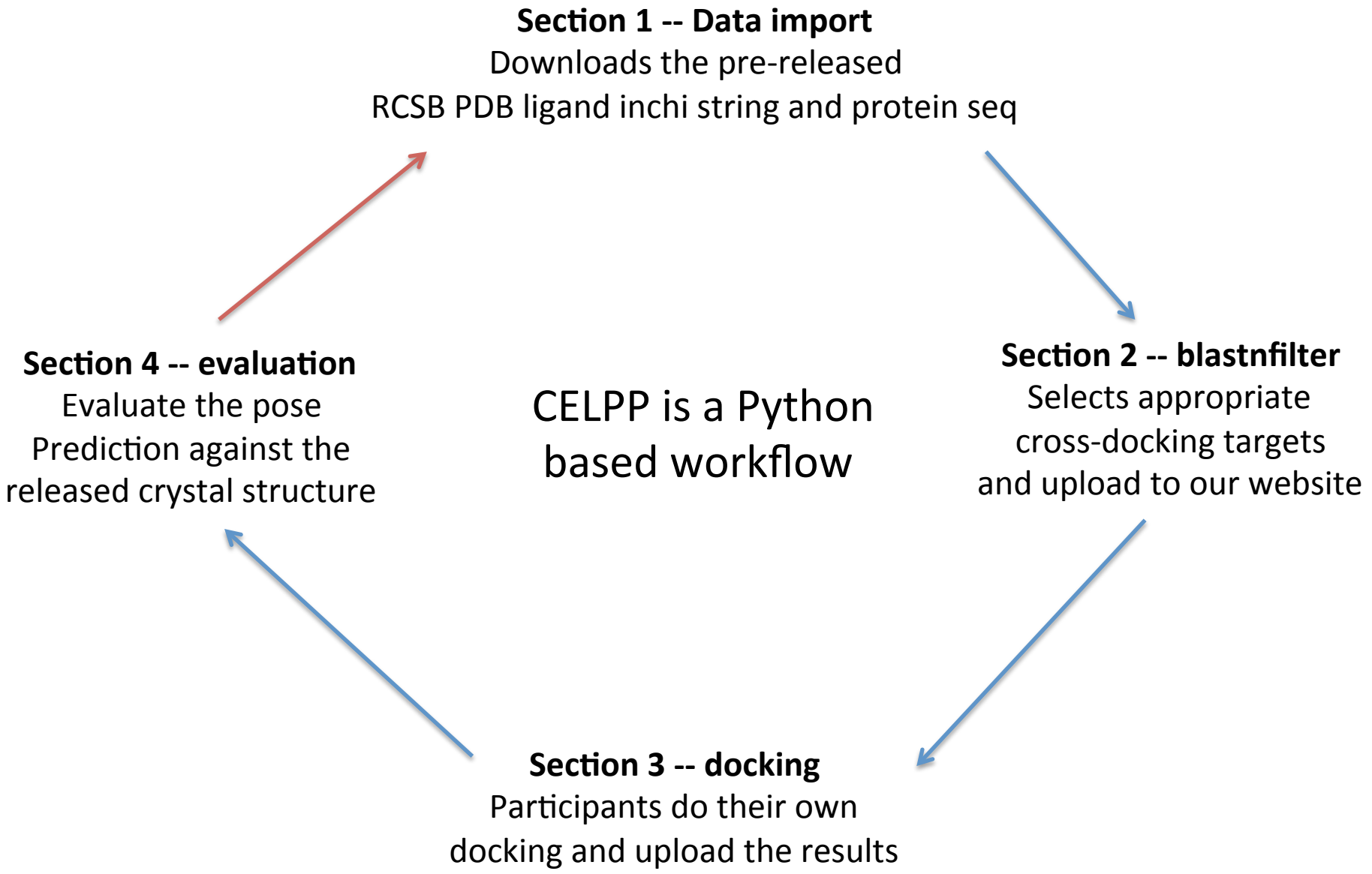
compound inchi string and protein seq



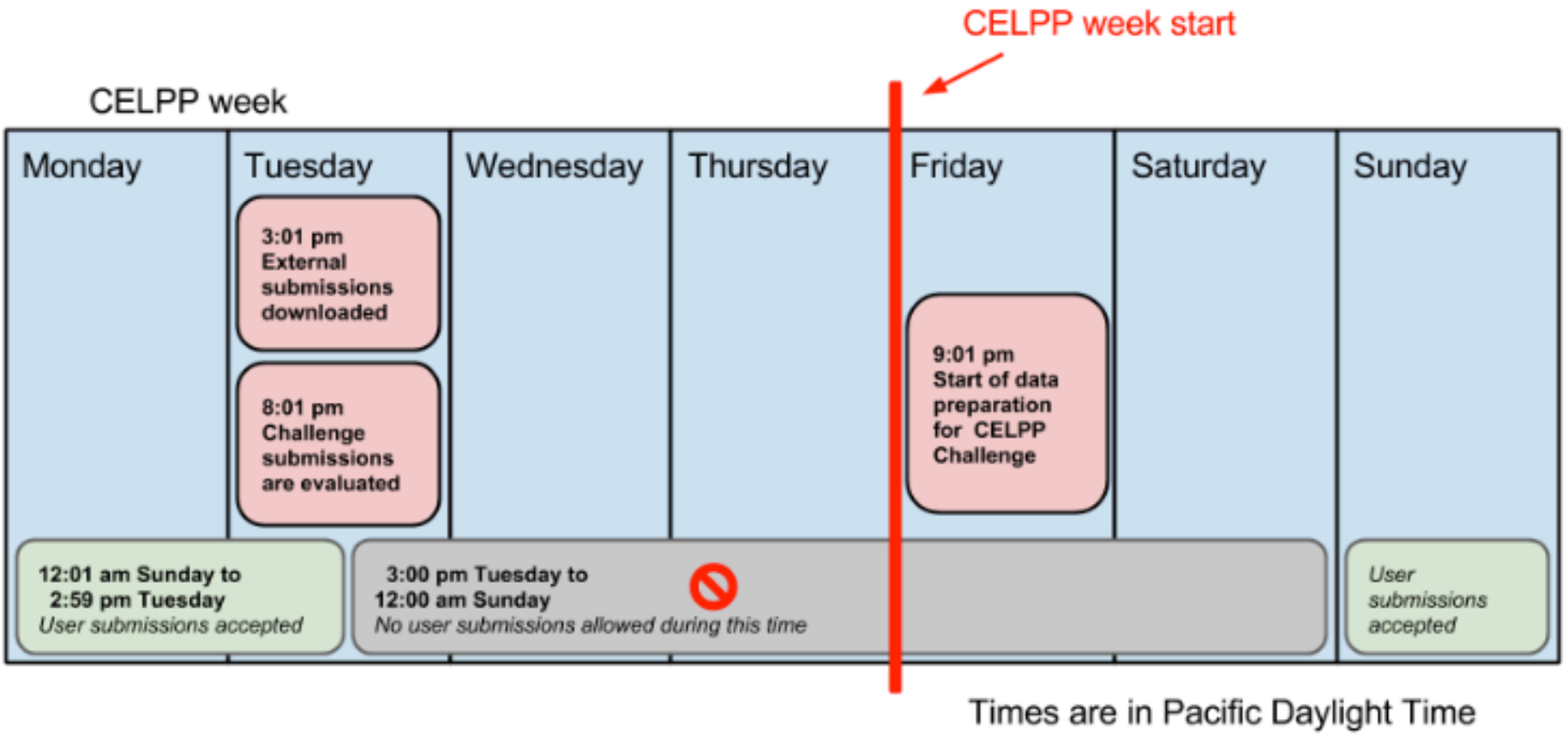
Tuesday night

3D coordinates

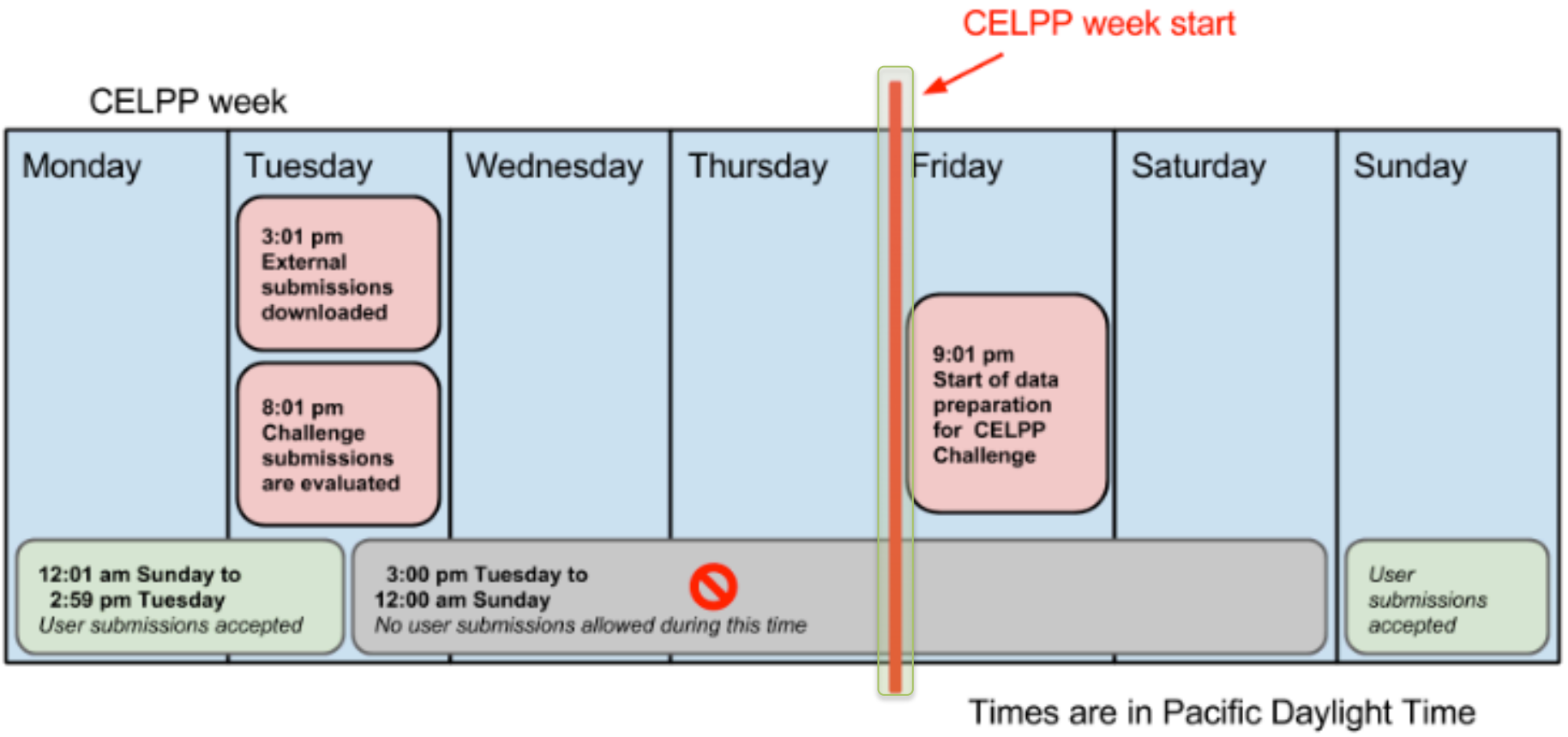
RCSB PDB provides compound INCHI strings and the protein polymer sequence four days prior to the release of their 3D coordinates, which gives an opportunity to predict small molecule protein docked poses each week.


Section 1 -- Data import
Downloads the pre-released
RCSB PDB ligand inchi string and protein seq
Section 2 -- blastnfilter
Selects appropriate
cross-docking targets
and upload to our website
Section 3 -- docking
Participants do their own
docking and upload the results
Section 4 -- evaluation
Evaluate the pose
Prediction against the
released crystal structure
CELPP is a Python
based workflow

CELPP week


CELPP week start
CELPP week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
3:01 pm
External
submissions
downloaded
8:01 pm
Challenge
submissions
are evaluated
9:01 pm
Start of data
preparation
for CELPP
Challenge
12:01 am Sunday to
2:59 pm Tuesday
User submissions accepted
3:00 pm Tuesday to
12:00 am Sunday
No user submissions allowed during this time
User
submissions
accepted
Times are in Pacific Daylight Time

Section 1: Data import



Section 1: Data import

- CELPP download the RCSB PDB pre-released

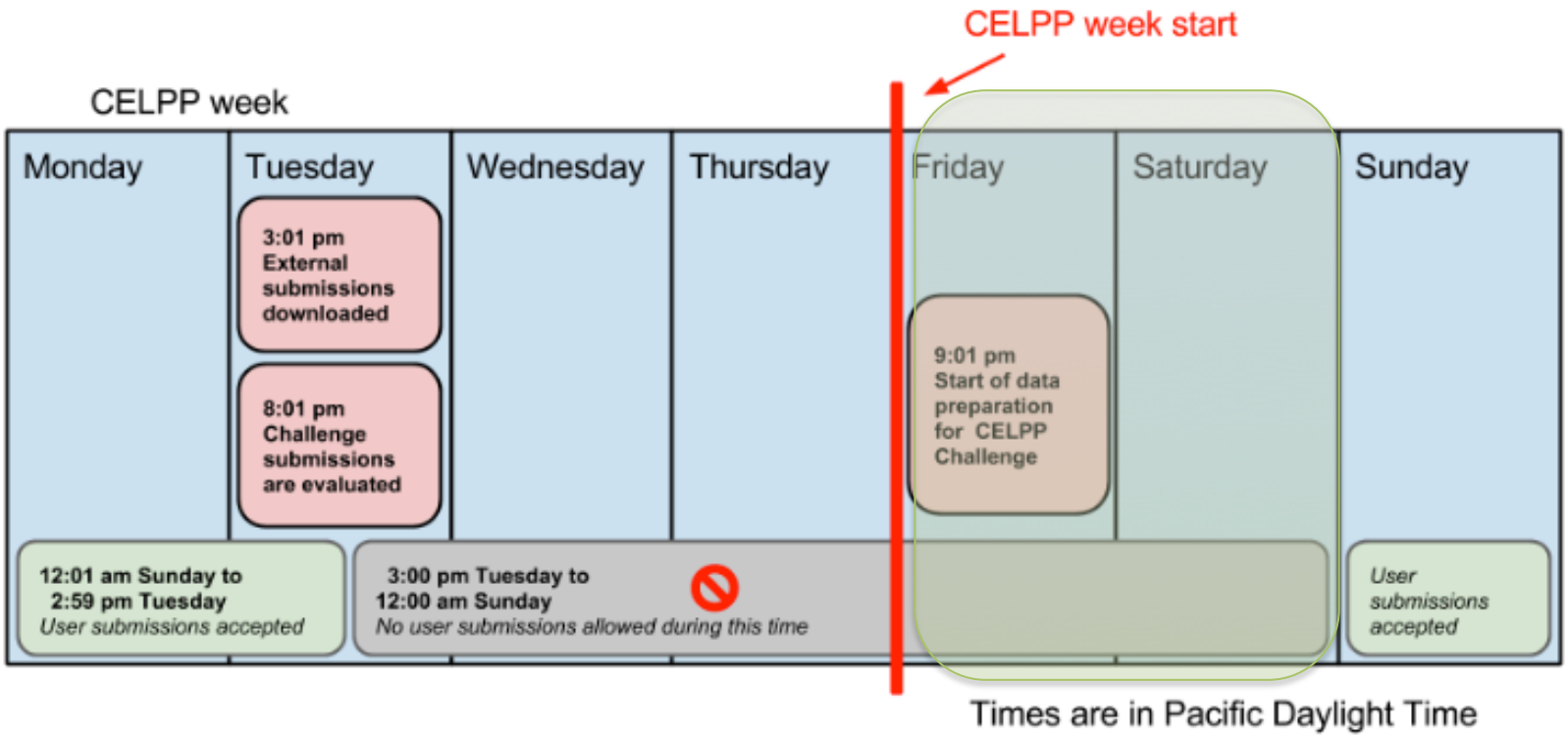
  1, compound International Chemical Identifiers (INCHI) strings (query ligand)

PDBID	Ligand ID	Inchi string
5IS4	6LY	InChI=1S/C8H10BrNO/c9-7-3-1-6(2-4-7)8(11)5-10/h1-4,8,11H,5,10H2/t8-/m1/s1

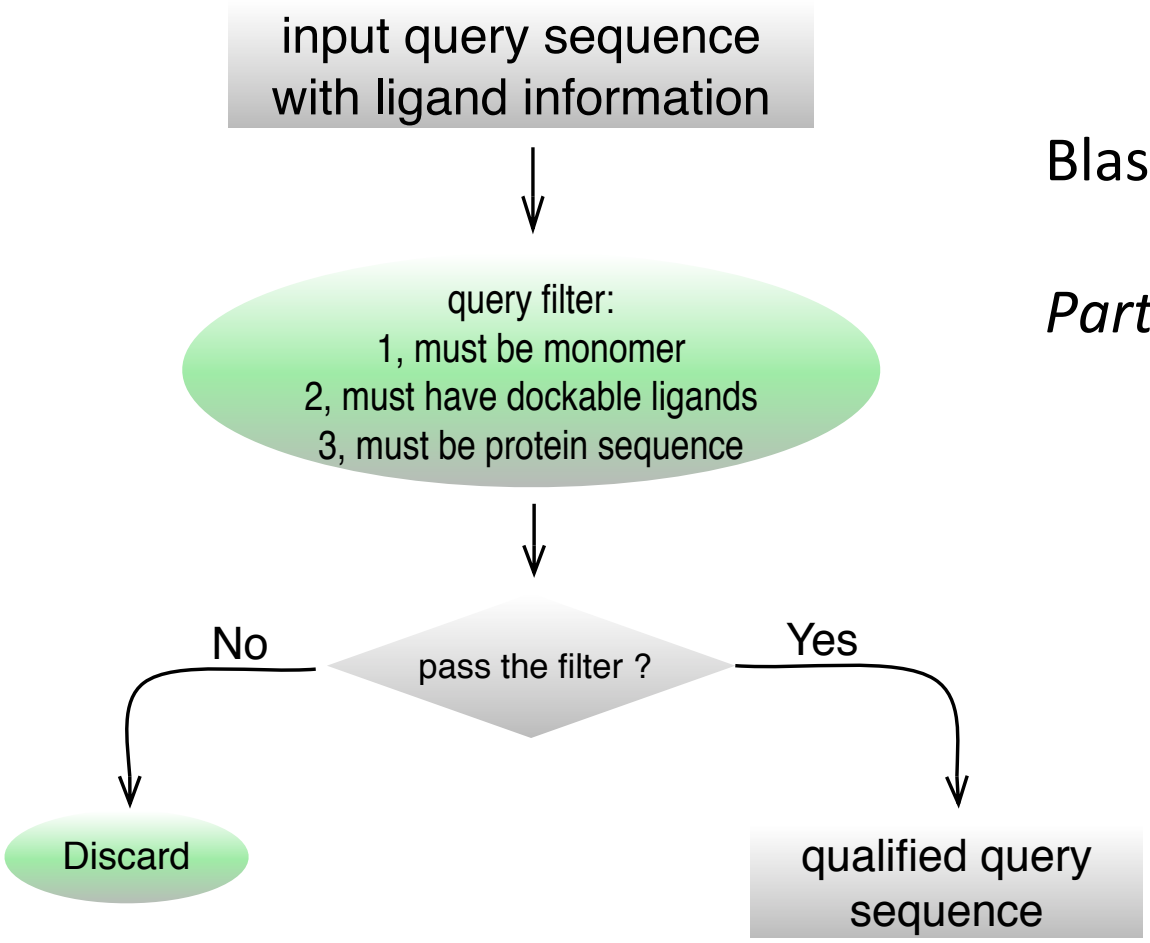
  2, the protein polymer sequence (query protein)

PDBID	Sequence number	Sequence
5IS4	1	STGSATTTPIDSLDDAYITPVQIGTPAQTLNLDFDTGSSDLWVFSSETTASEVDGQTIYTPSKSTTAKLLSGATWSISYGDGSSSSGDVYTDTVSVGGLTVTGQAVESAKKVSSSFTEDSTIDGLLGLAFSTLNTVSPTQQKTFFDNAKASLDSPVFTADLGYHAPGTYNFGFIDTTAYTGSITYTAVSTKQGFWEWTSTGYAVGSGTFKSTSIDGIADTGTTLLYLPATVVSAYWAQVSGAKSSSSVGGYVFPCSATLPSFTFGVGSARIVIPGDYIDFGPISTGSSSCFGGIQSSAGIGINIFGDVALKAAFVVFNGATTPTLGFASK

Section 2 – blastnfilter

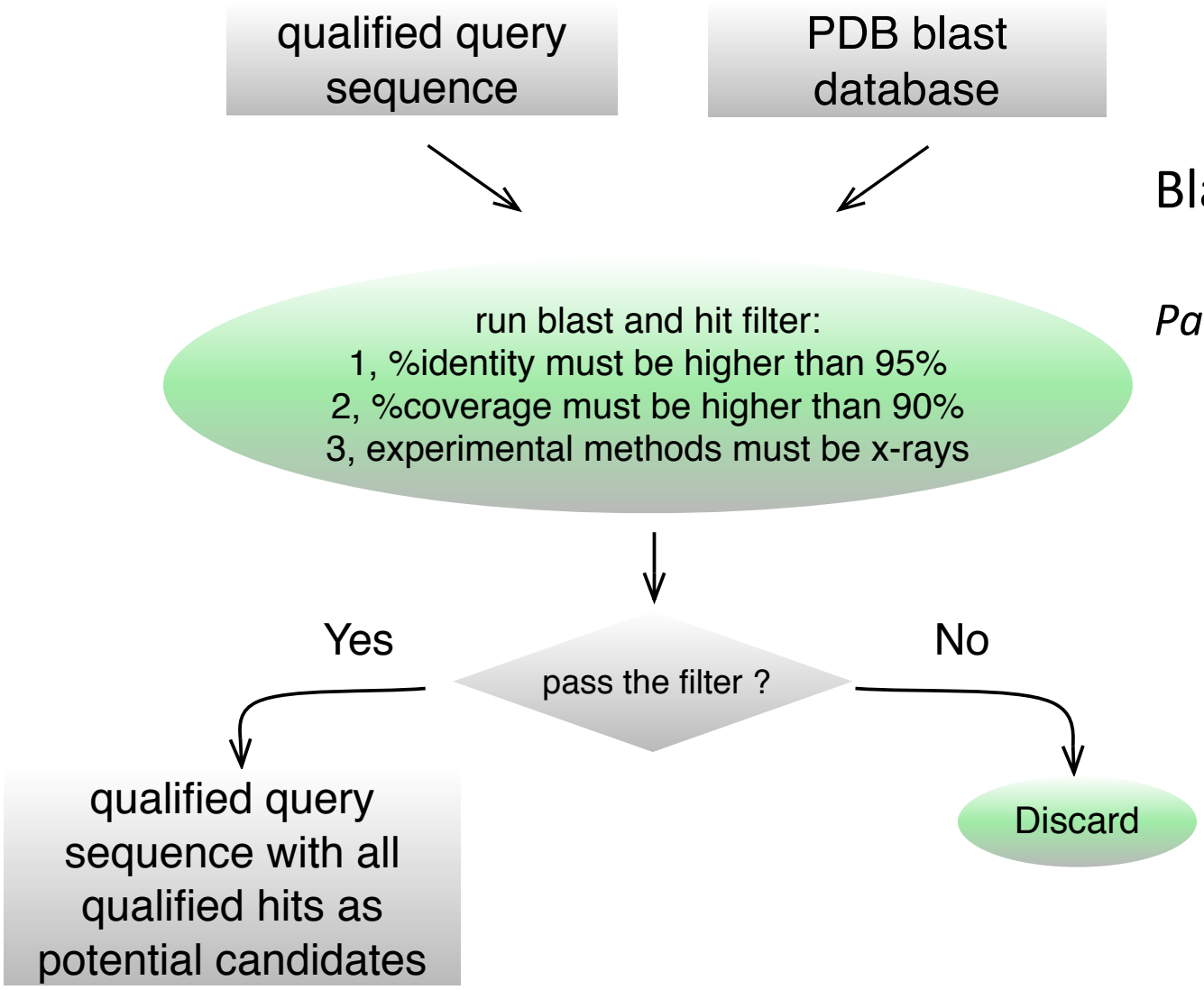


CELPP blasts the query sequence against the PDB database and finds similar proteins as hits, and then we filter them using few rules to get the final docking candidates.



Blastnfilter workflows:

Part 1: Query filter



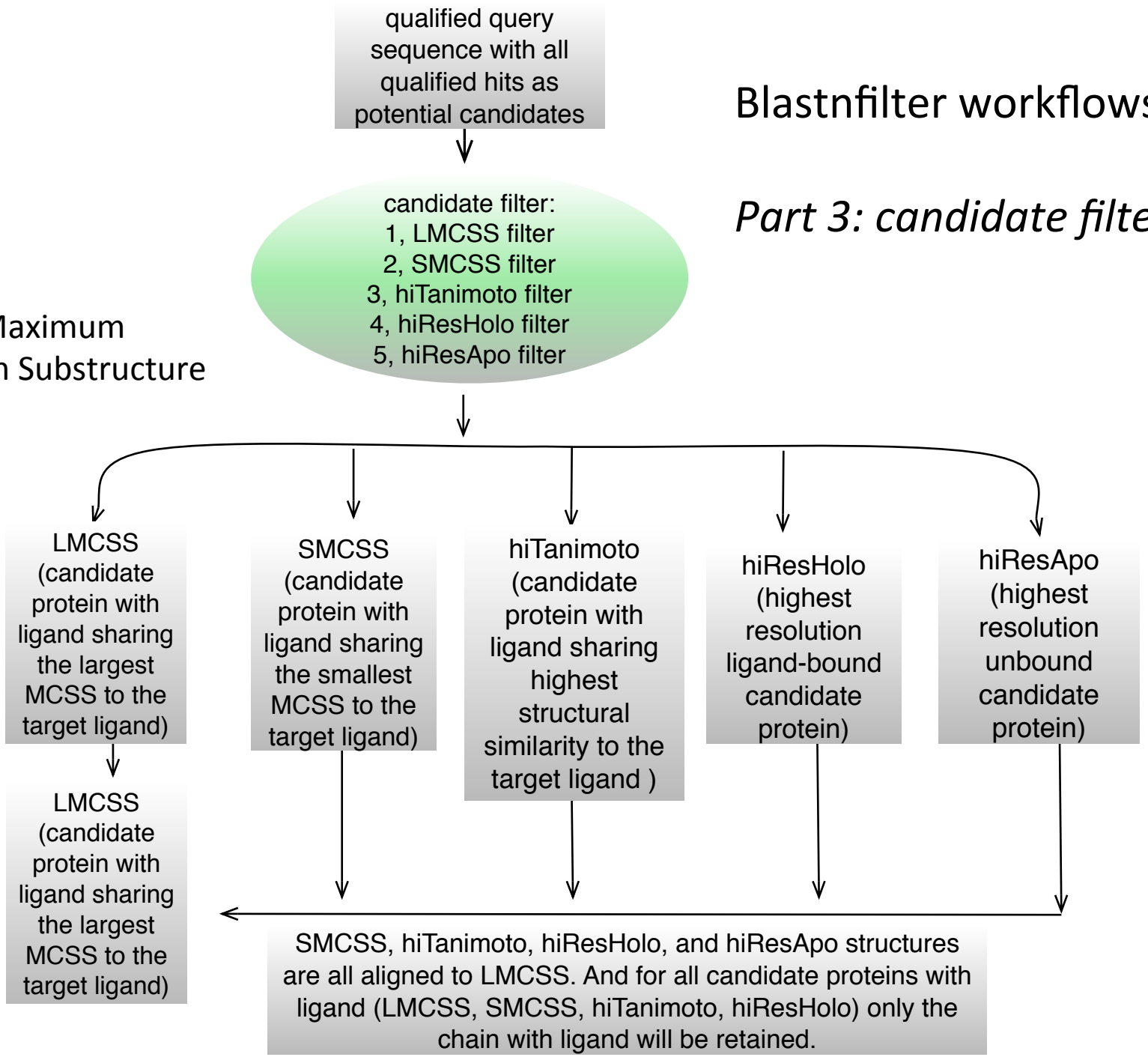
Blastnfilter workflows

Part 2: Blast and hit filter

Blastnfilter workflows

Part 3: candidate filter

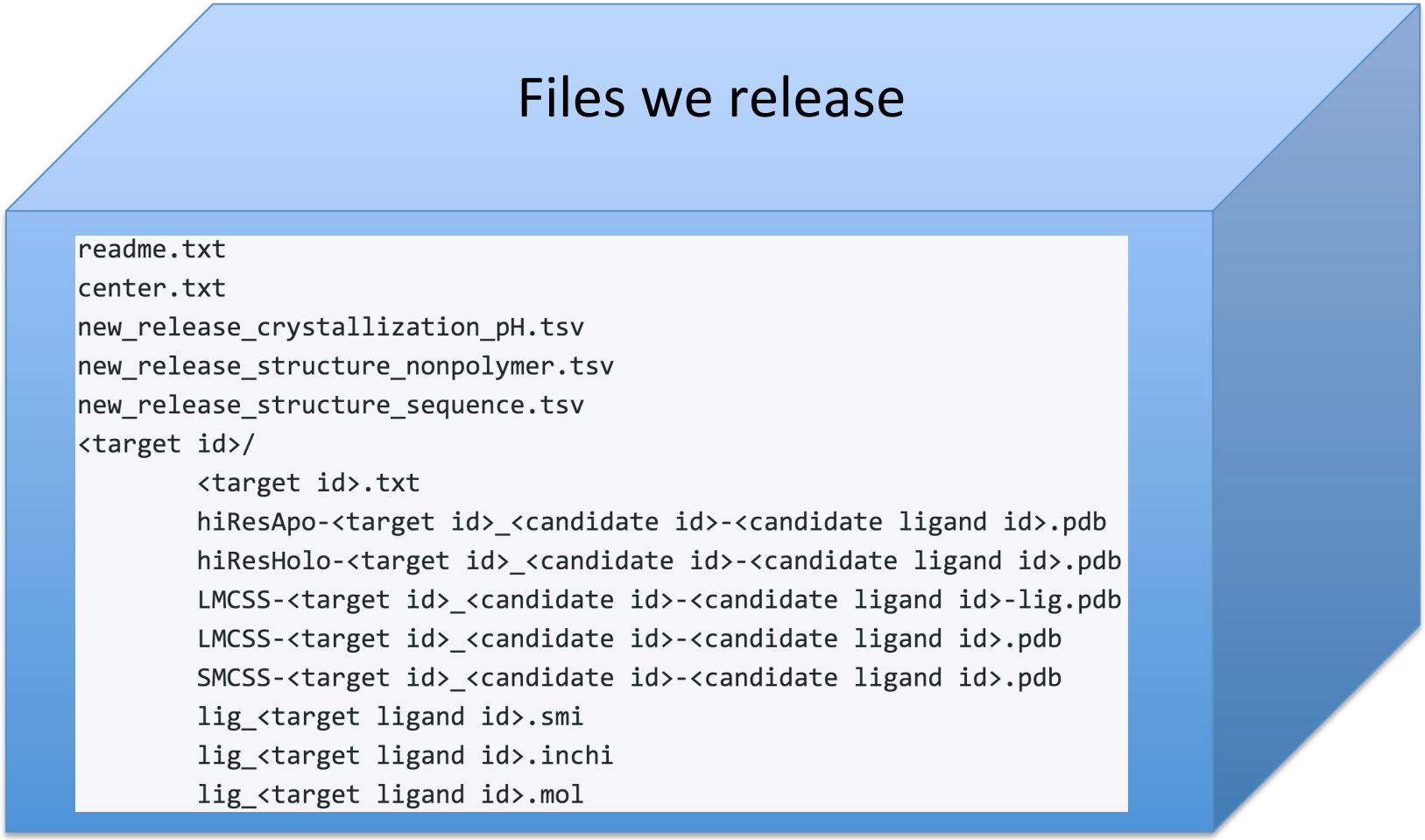
MCSS: Maximum Common Substructure



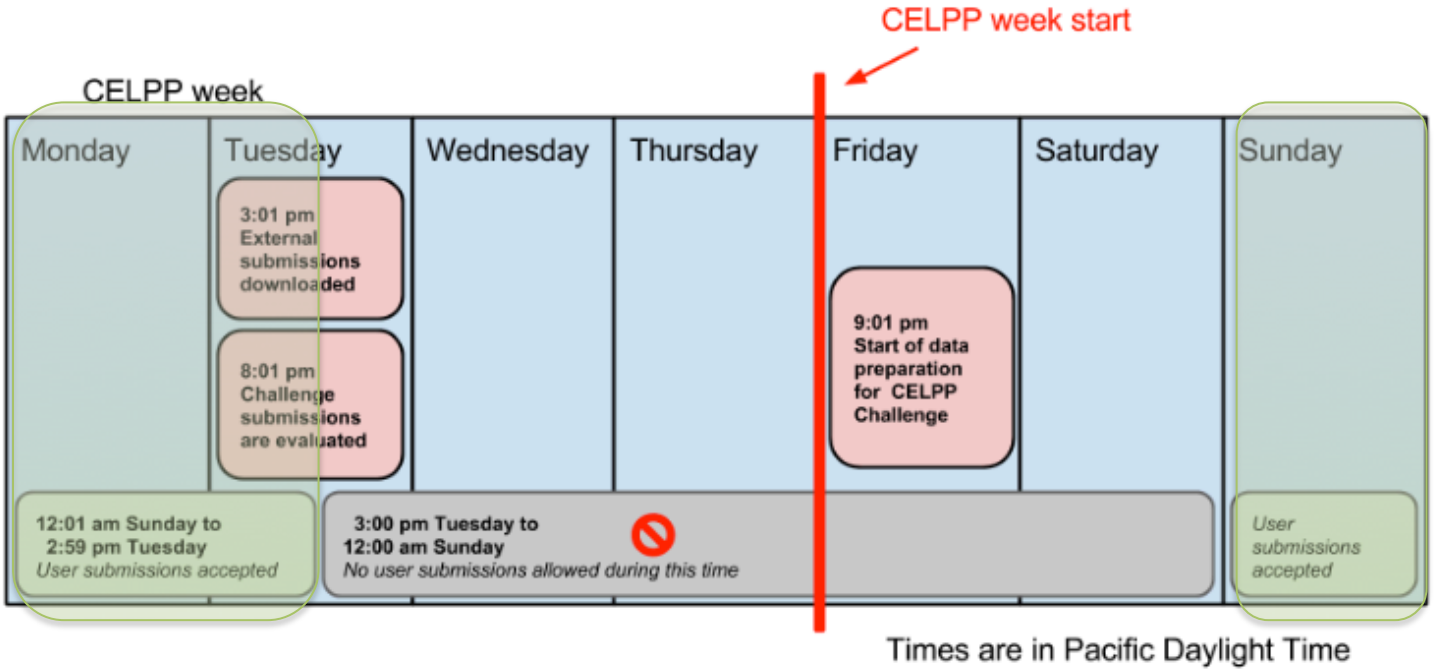
Blastnfilter reports

Query PDBID	query, 5is4
	ph, 4.6
Query LigandID	ligand, 6LY
	inchi, InChI=1S/C8H10BrNO/c9-7-3-1-6(2-4-7)8(11)5-10/h1-4,8,11H,5,10H2/t8-/m1/s1
Heavy atom num	size, 12
	rotatable_bond, 2
Candidates info Type, PDBID, LigandID …	LMCSS, 4y4j, LNR, chain: A, (size: 13, mcss_size: 11, resolution: 1.03)
	LMCSS, 4y5k, 489, chain: A, (size: 12, mcss_size: 11, resolution: 1.44)
	SMCSS, 4y4d, CFF, chain: A, (size: 14, mcss_size: 3, resolution: 1.27)
	hiResHolo, 1oew, SUI, chain: A, (resolution: 0.9)
	hiResApo, 4y5l
	hiTanimoto, 4y5k, 489, chain: A, (tanimoto_similarity: 0.70, resolution: 1.44)

After blastnfilter, we release the data package through *box.com*



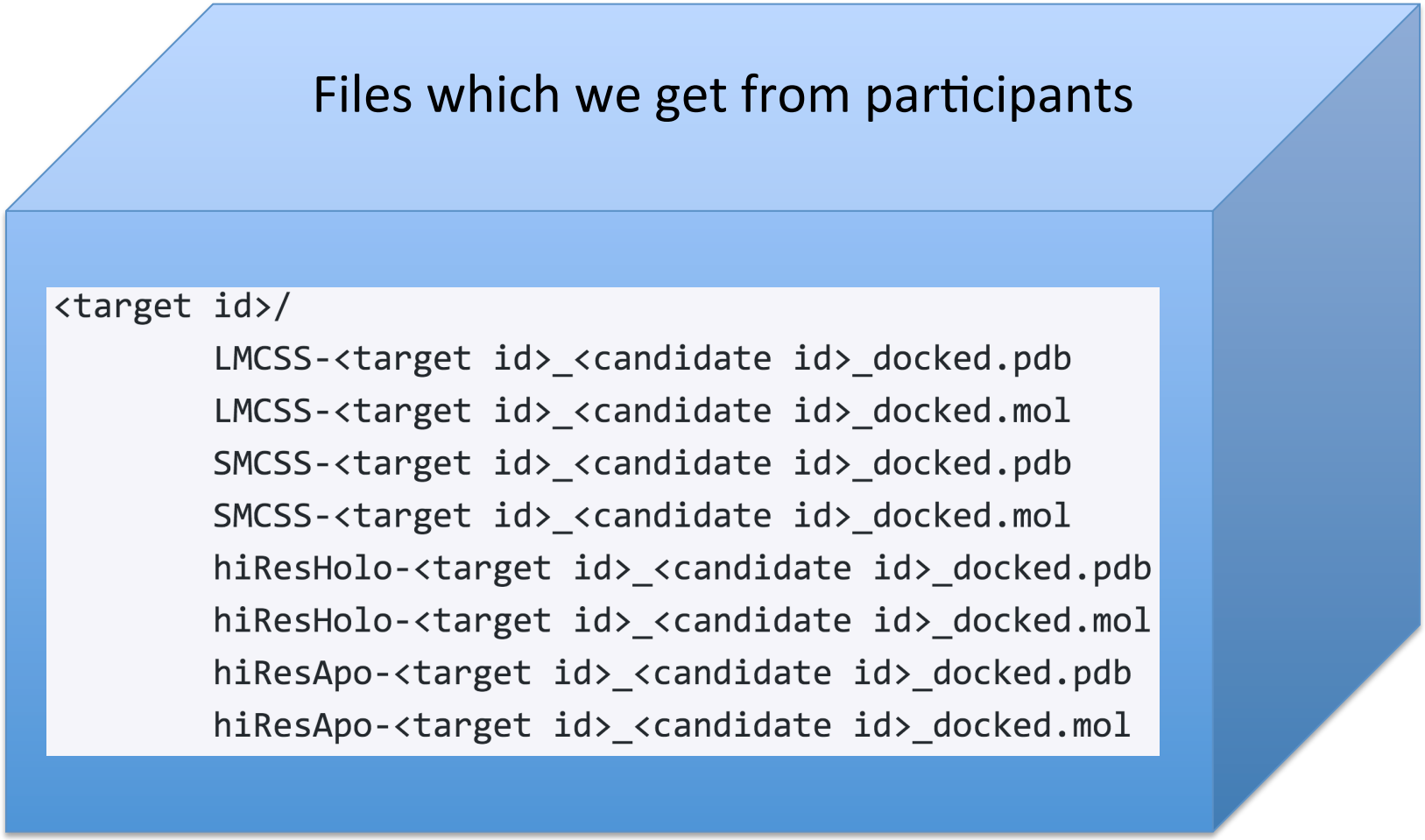
Section 3 – docking



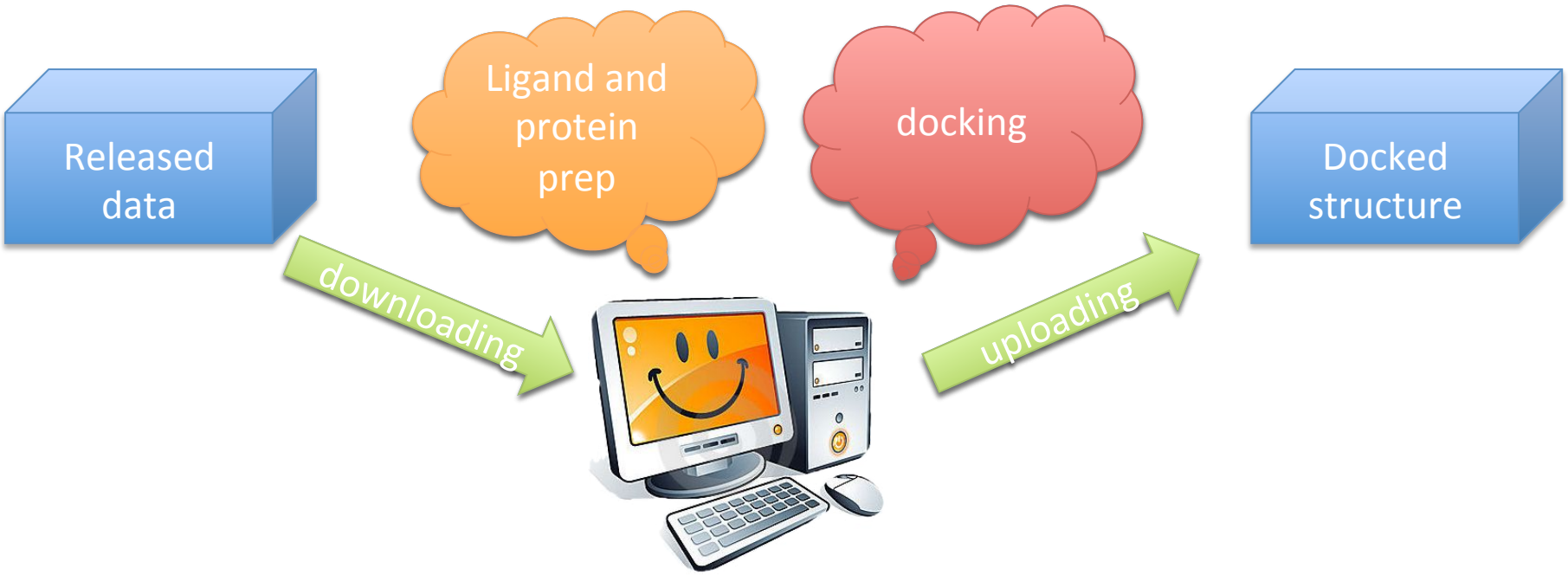
- 1, Ligand and protein preparation
- 2, Grid generation and docking

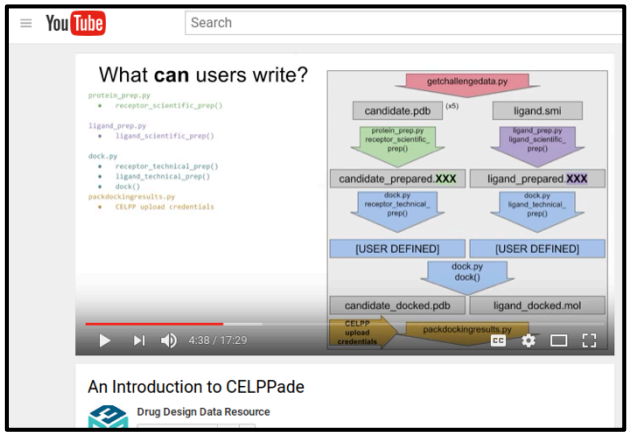
Participants could join in the challenge in this stage with customized preparation and docking methods and send us the docked poses for evaluation.

After docking, participants will send the docked structures to *box.com*

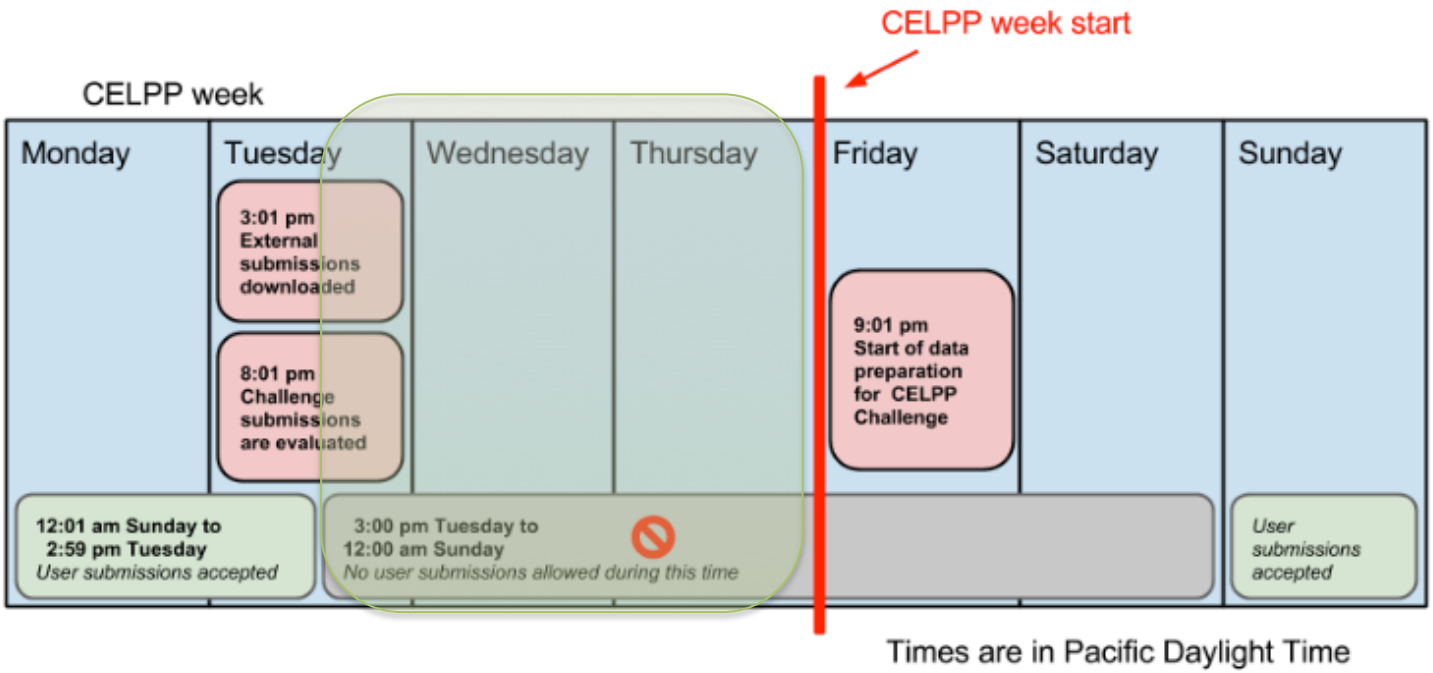


CELPPade will help handle the data downloading and uploading, and it provides a template for ligand, protein preparation, and docking.





Section 4 – Evaluation



- 1, Evaluate the pose prediction using multiple scoring matrixes like RMSD, RSCC and potentially ROCS, protein ligand interaction fingerprint etc.

- 2, We will then send the overall statistics to the participants.

Below is the RMSD result of an internal submission using Glide

Target_PDBID	LMCSS	SMCSS	hiResApo	hiResHolo	hiTanimoto
Number_of_cases	22	22	11	22	22
Average	4.654	5.887	5.667	4.309	5.722
Minimum	0.347	0.387	1.156	0.664	0.347
Maximum	12.184	21.465	12.241	14.828	12.184
5ka3	0.846	6.635		1.142	0.846
5uud	1.324	2.287	2.071	1.327	9.255
5uua	3.999	3.521	4.004	3.229	3.999
5uub	4.007	3.589	4.036	3.25	4.007
5jsl	10.608	9.737		5.931	10.608
5l13	10.201	11.272		9.432	11.352
5v0u	0.66	5.082	4.821	5.082	5.082
5ur3	6.909	6.005	12.241	6.005	6.909
5ka7	0.867	6.61		1.181	0.867
5i96	12.184	6.781		6.781	12.184
1fcz	0.347	0.387		0.664	0.347
5kab	0.836	6.718		1.227	0.836
5ggz	0.519	0.948	1.156	1.111	7.322
5uu9	4.006	3.53	4.019	3.247	4.006
5uu8	4.045	3.572	4.042	3.242	4.045
5lwo	3.351	3.284		3.585	3.351

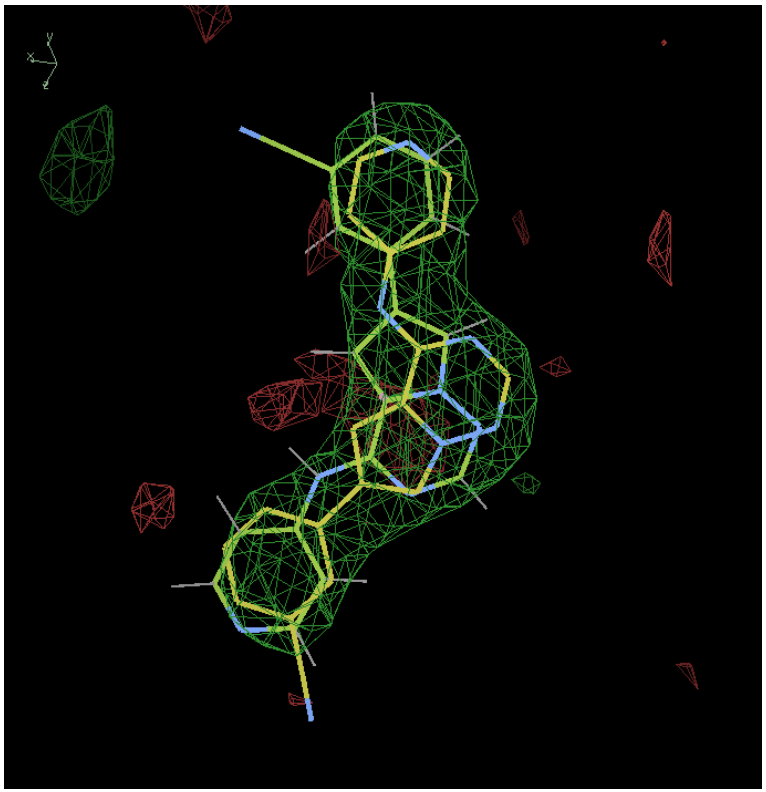
Using other metrics to evaluate the pose prediction?

Motivation: Some cases found in the D3R grand challenge evaluations suggested that beside RMSD, other pose prediction evaluation matrixes like RSCC maybe also useful.

RSCC: real-space correlation coefficient is a measure of the similarity between an electron-density map calculated directly from a structural model and one calculated from experimental data.

RMSD vs RSCC – Case 1

In most of the cases, RSCC results agreed with the RMSD, while there are several outliers



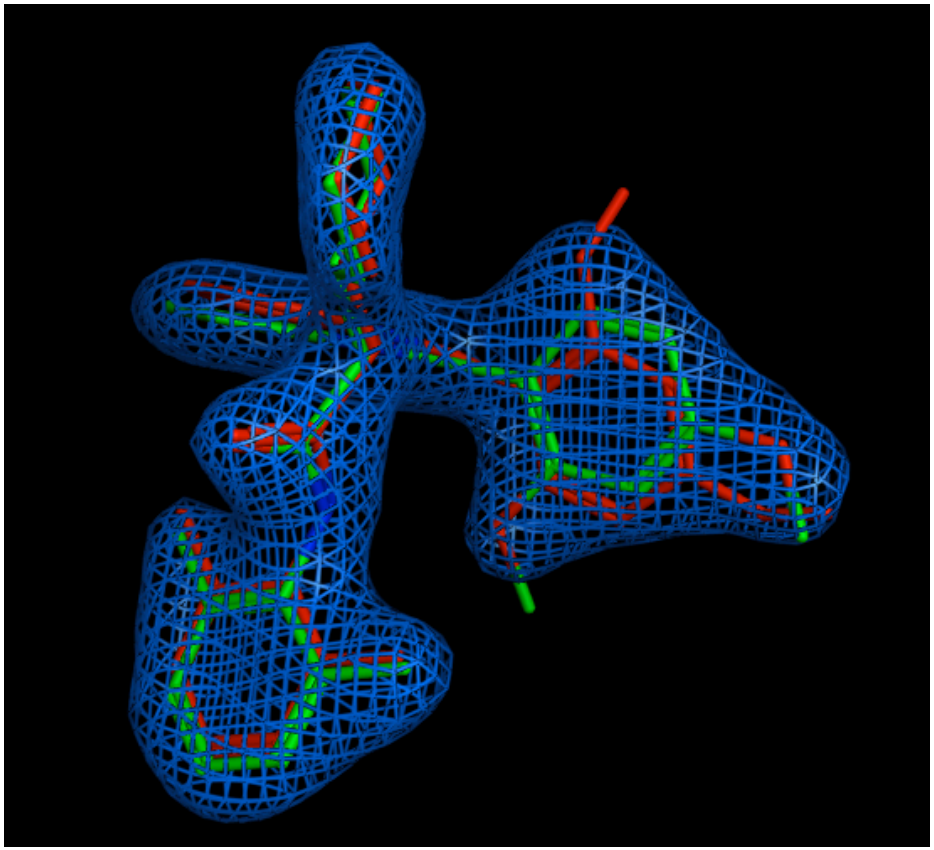
Yellow: crystal ligand
Green: docked ligand

Case 1:

rscc_d/rscc_c=0.86 RMSD=8.4

The rscc score is good while the RMSD is not

RMSD vs RSCC– Case 2



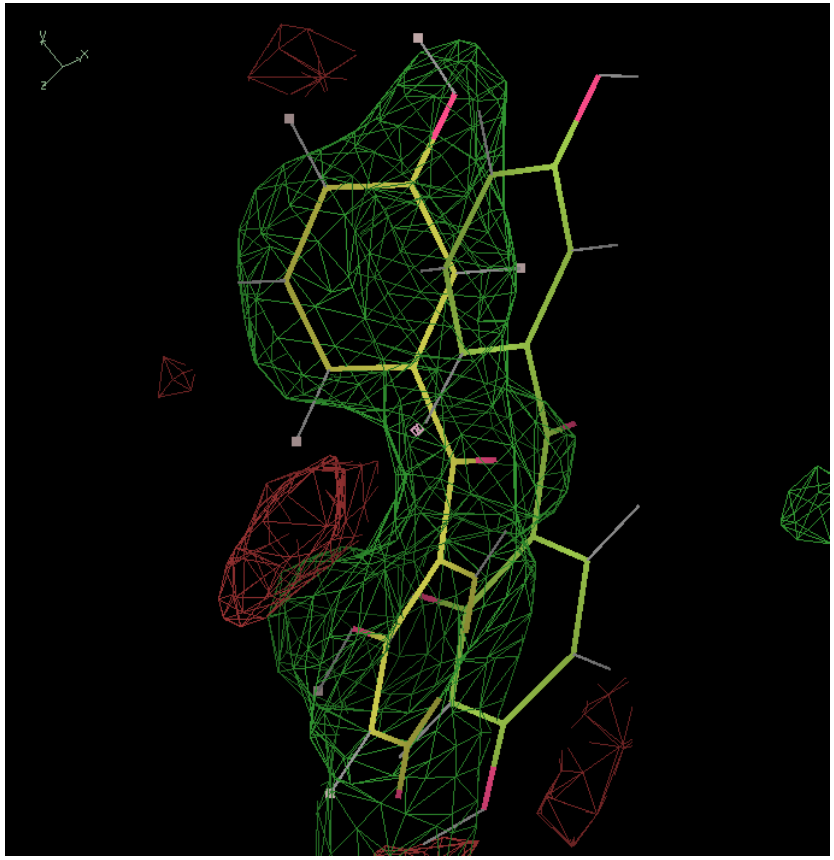
Green: crystal ligand
Red: docked ligand

Case 2:

rscc_d/rscc_c=0.96 RMSD=1.6

The RSCC is perfect and the RMSD is also good while could we make the RMSD better?

RMSD vs RSCC – Case 3



Yellow: crystal ligand
Green: docked ligand

Case 3:

rscc_d/rscc_c=0.16 RMSD=1.7

The RMSD is good while the RSCC is not

Lessons learned from RMSD vs RSCC

- 1, RMSD has advantage if the ligand just shifts a little bit – case 3
- 2, RSCC suggests that we may need different crystal models – case 1, 2
- 3, We may need other metrics like protein ligand fingerprint, ROCS score etc.

How to participate?

- CELPP overview: drugdesigndata.org/about/celpp
- Google group: groups.google.com/forum/#!forum/celpp-developers
- CELPP GitHub Wiki:github.com/drugdata/d3r/wiki

Contact us directly at drugdesigndata@gmail.com

Acknowledgements

- UCSD: Jeff Wagner, Chris Churas, Mike Gilson, Rommie Amaro, Mike Chiu, Jeff Grethe, Vicki Feher, Rob Swift
- Rutgers: Stephen Burley, Huanwang Yang

Thank you!