

CELPP

Continuous Evaluation of Ligand Pose Prediction



Jeff Wagner

D3R Workshop, February 2018

CELPP

“Weekly blinded challenges for pose prediction”



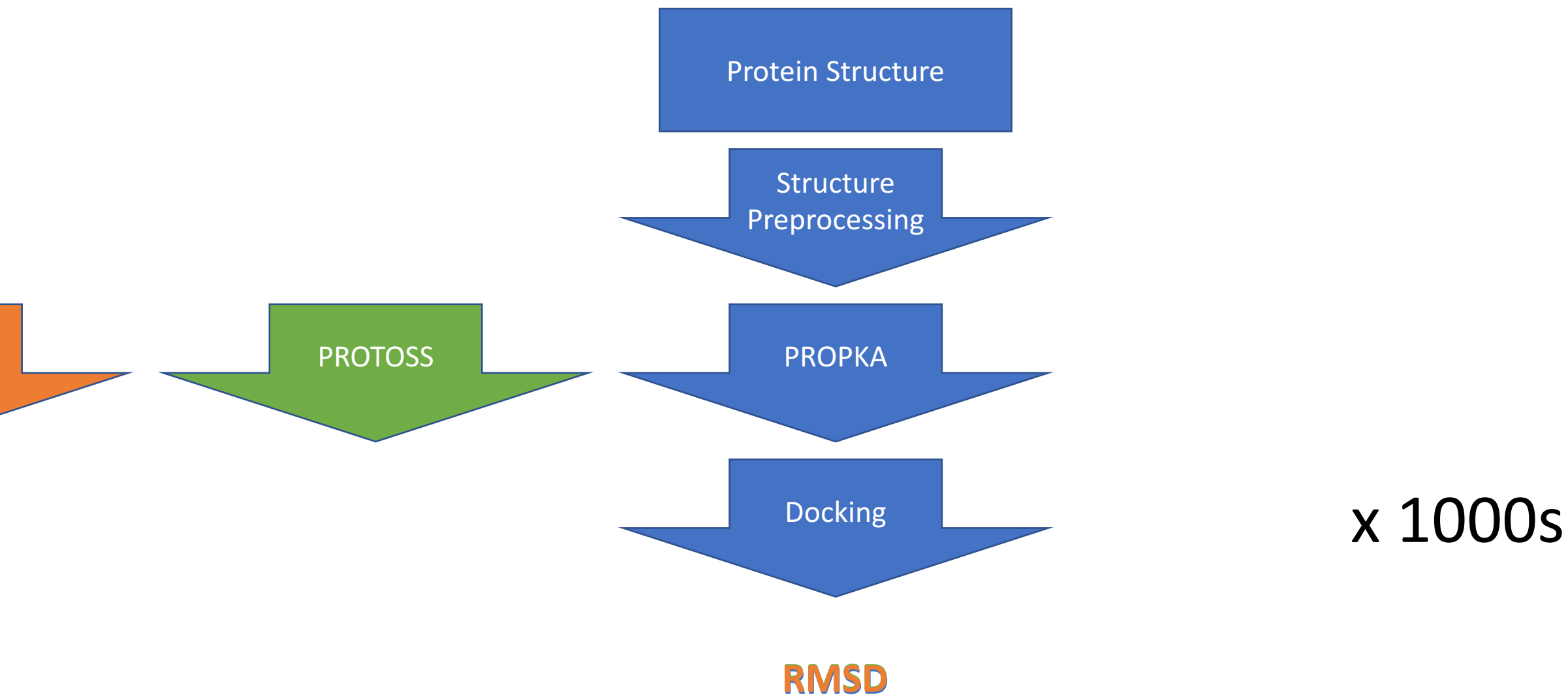
Jeff Wagner

D3R Workshop, February 2018

How can we achieve statistical significance?

- In Grand Challenge format, the amount of data curation work scales linearly with number of targets
- Grand Challenge participants often use human input, which is frequently non-reproducible
- Best-performing workflows in GCs are often fundamentally different, making it difficult to understand the individual contribution of each step

The Pipe Dream



Where can we get thousands of test cases for docking?

RCSB PDB Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Download ▾ Learn ▾ More ▾

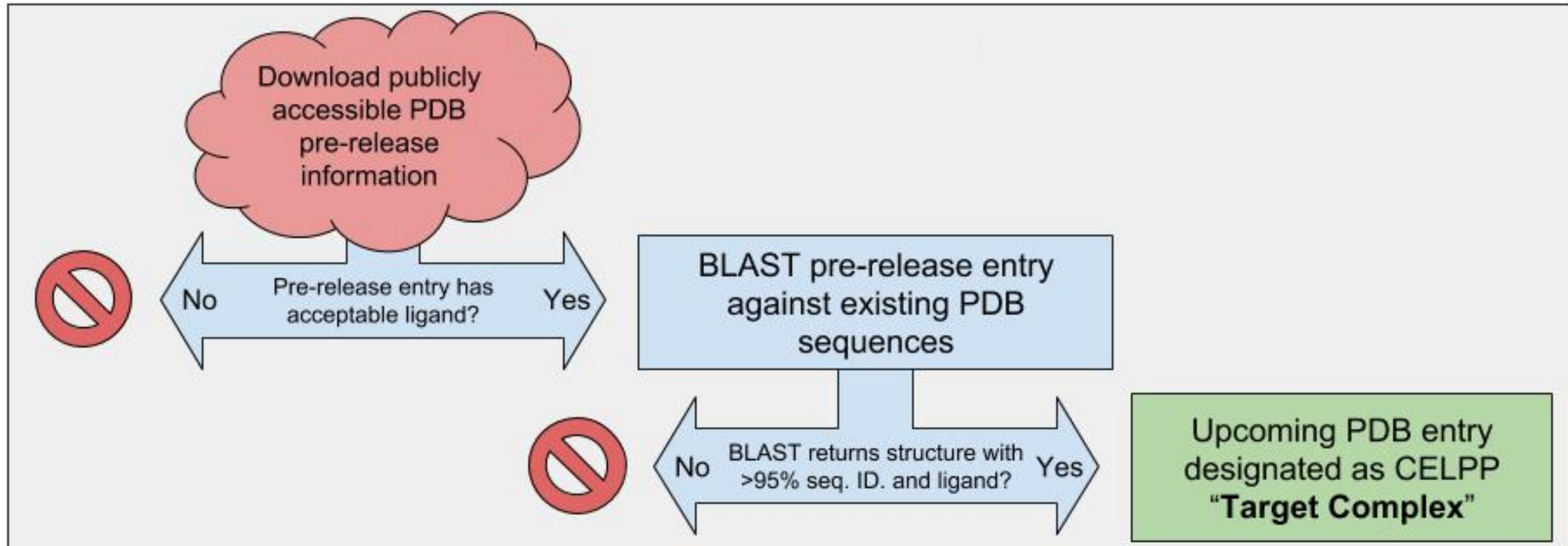
List pre-release sequences in FASTA format

Get a list of pre-release sequence in FASTA format.

- Example: Access the information for a subset of PDB IDs: </pdb/rest/getStatusSequence?structureId=3MU6,3QV1,3SSF>
- Example: Access the information for a subset of PDB IDs, entity type, and range of deposition date: </pdb/rest/getStatusSequence?entityType=RNA&depositionDateMin=2011-07-01&depositionDateMax=2011-07-30> (The entity type can be **RNA**, **DNA**, and **Polypeptide**.)
- Example: Access the information for all unreleased PDB IDs: </pdb/rest/getStatusSequence>

Identifying suitable PDB pre-release structures for the challenge

5UGB	1	SGEAPNQALLRIL...TEFKKIKVLGSGAFGTVYK
5UGB	8BM	InChI=1S/C25H35N9O/c1-17(2)34-16-26...,30)/t20-/m0/s1



Protein Structures for Docking Come From Existing PDB Entries

- D3R does a BLAST search of the PDB and provides crystal structures of homologous proteins
- Most weeks require predicting 20 - 100 upcoming PDB complexes
 - Up to 500 docking jobs

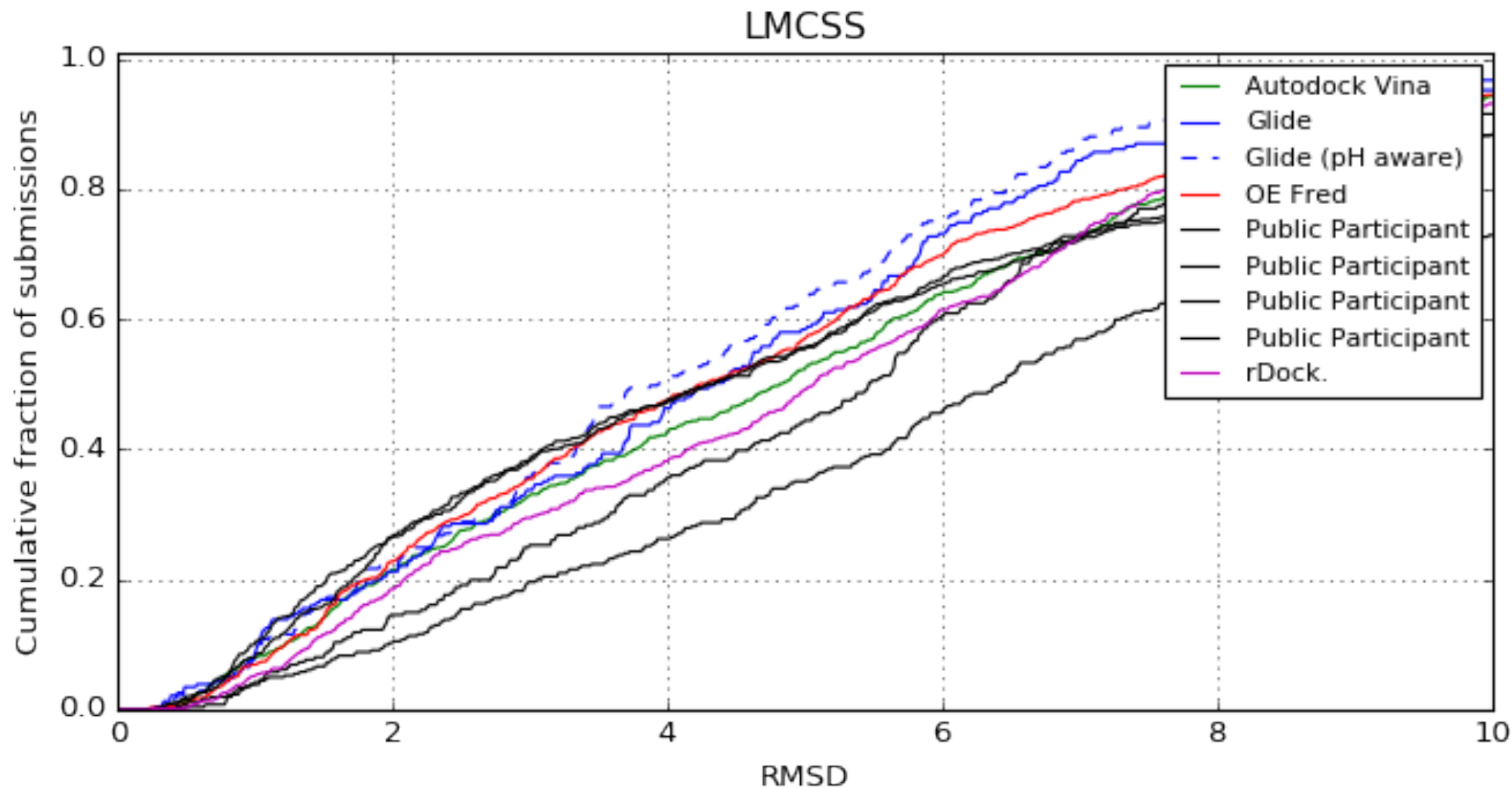
40 Weeks Later...

- D3R has been running baseline “participants”
- Infrastructure for selecting targets, distributing weekly challenge packages, and evaluating results is automated
- External groups/beta testers volunteered to participate and have automated workflows running each week

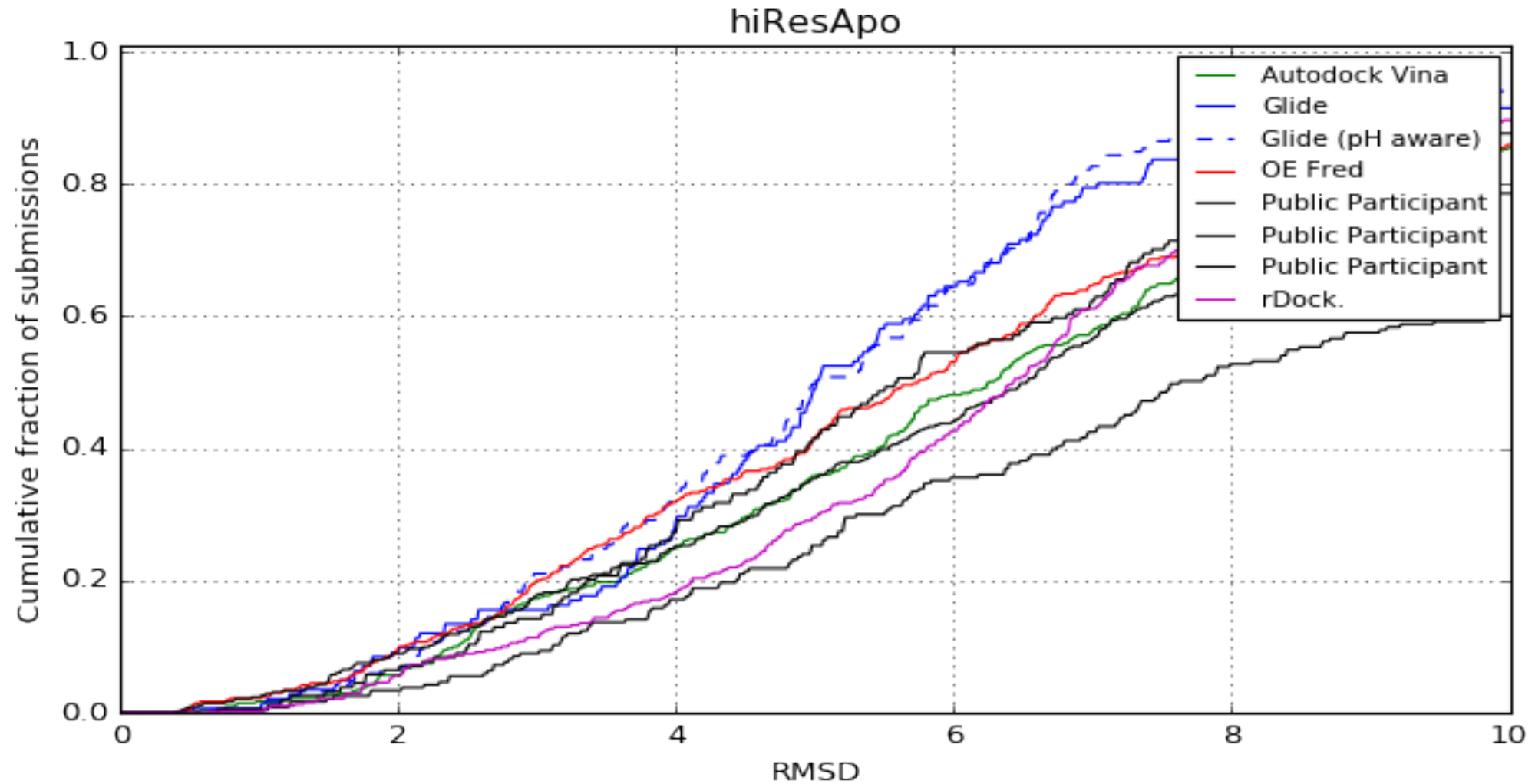
Different measures of performance

Method	% under 2Å	Median RMSD	Mean RMSD	# submissions
Public Participant	26%	4.29	5.75	589
Public Participant	26%	4.33	5.39	654
OE Fred [D3R]	23%	4.24	4.66	637
Glide (pH aware) [D3R]	22%	3.94	4.29	277
Glide [D3R]	21%	4.35	4.53	231
Autodock Vina [D3R]	21%	4.82	5.01	595
rDock [D3R]	18%	5.07	5.19	584
Public Participant	15%	5.47	5.48	261
Public Participant	10%	6.17	7.62	124
Public Participant	10%	6.42	8.44	362
Public Participant	9%	5.75	5.90	249

Similar Ligand Crystallized in Binding Site

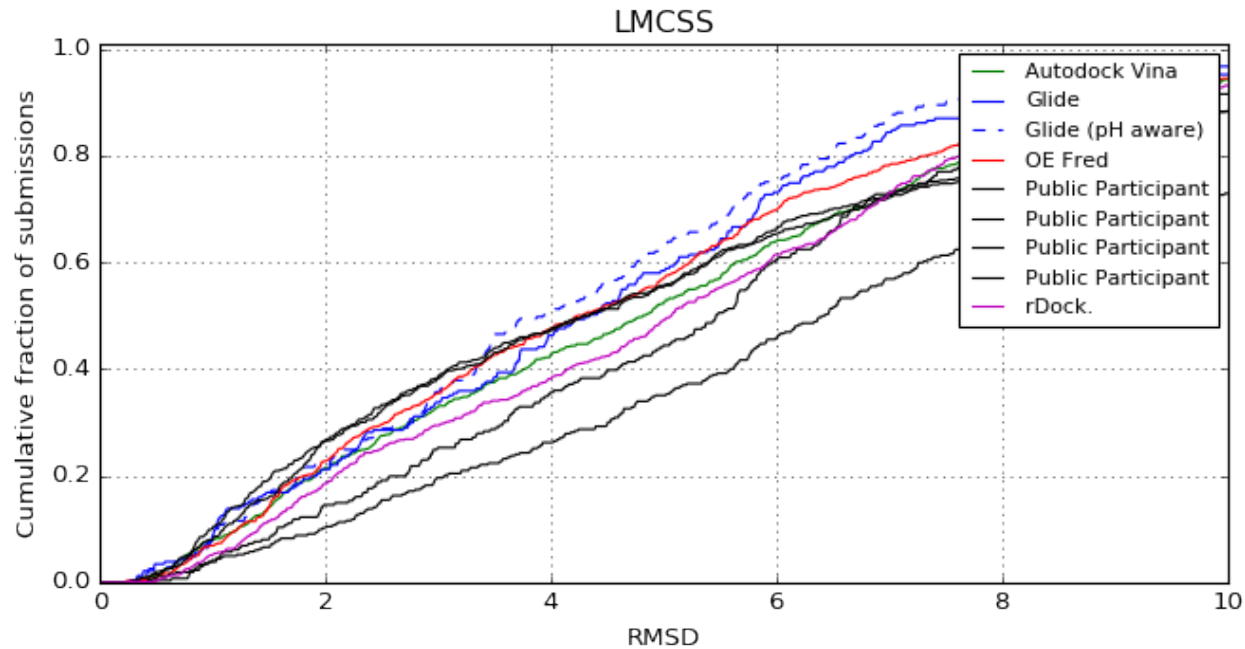


No Ligand Crystallized in Binding Site

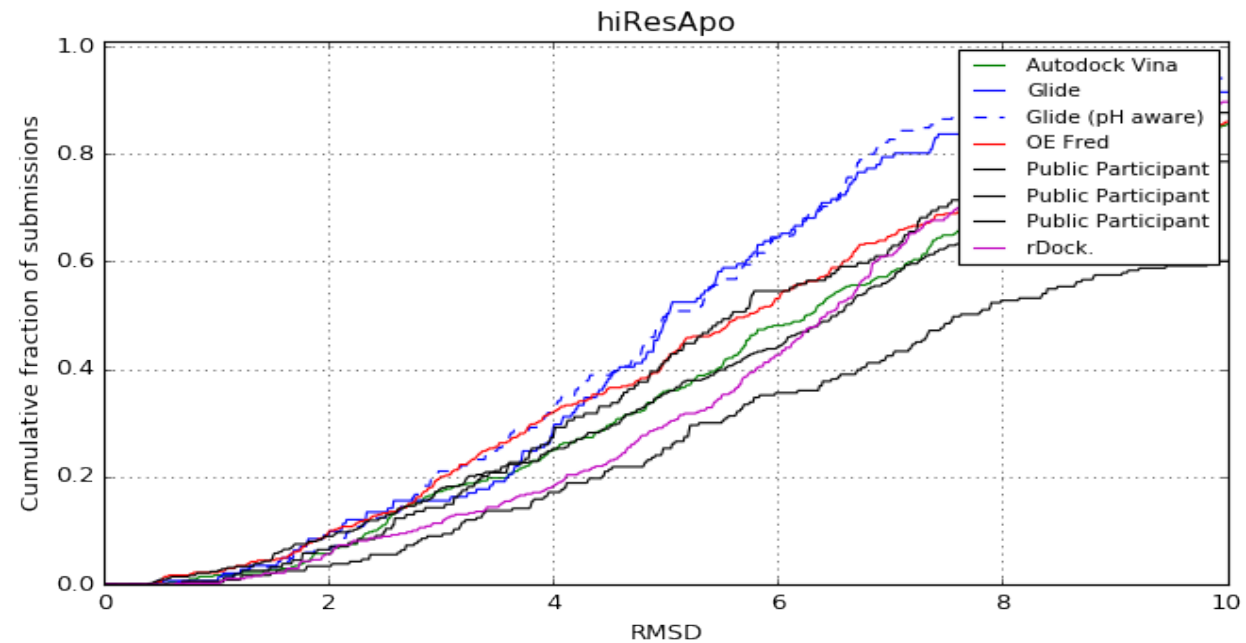


The Value of a Well-Arranged Binding Site

Similar Ligand in Binding Site



No Ligand in Binding Site



About a threefold improvement (10% → 30%) in the odds of getting <2 Å RMSD

HOW TO WIN CELPP AND INFLUENCE PEOPLE

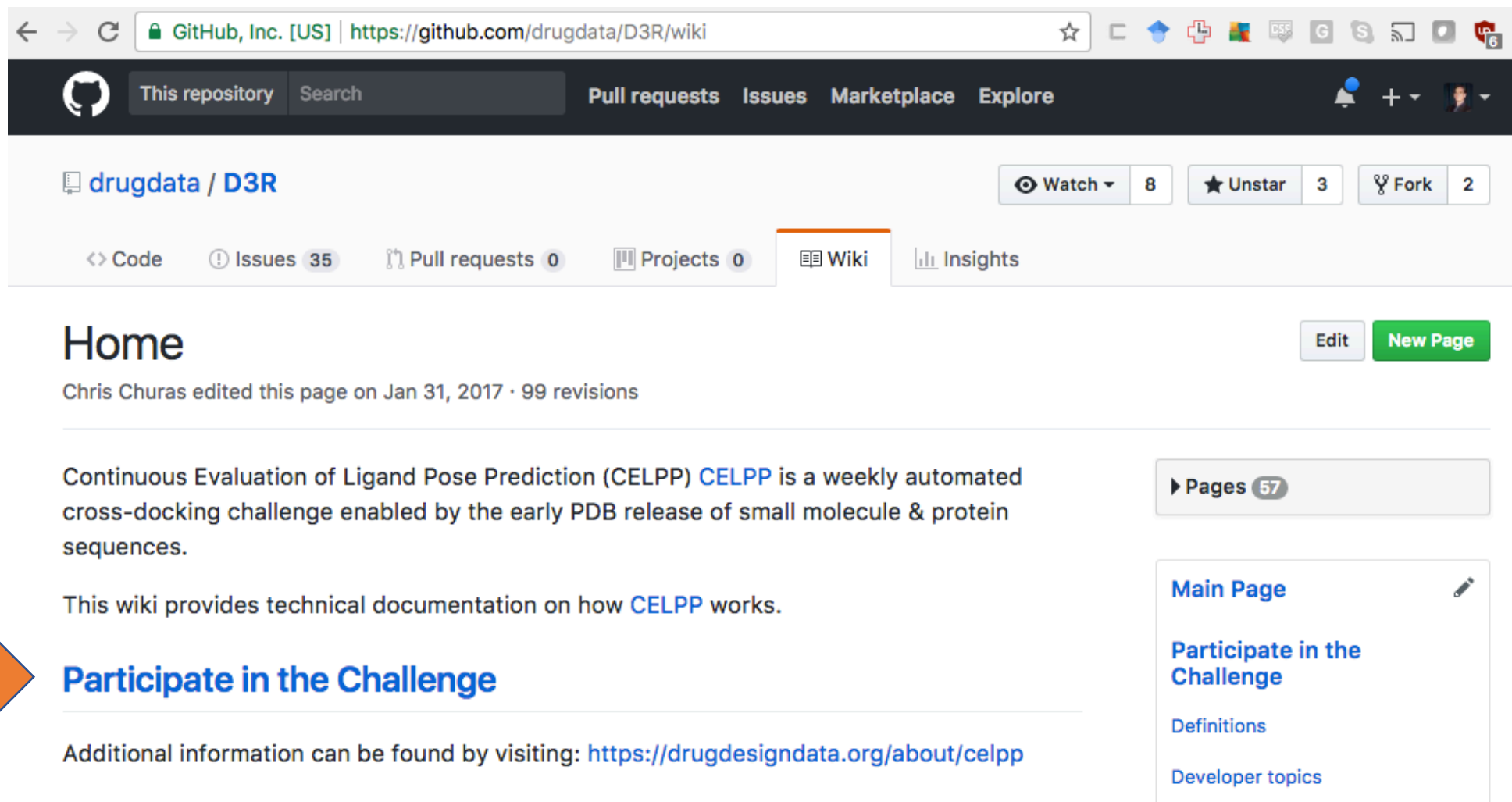


THIS IS COPY N° 3,288,873 OF
THE MOST POPULAR WORK OF NON-FICTION OF OUR TIME

1. What are the six ways of making people like you? *See pages 59-104.*
2. What are the twelve ways of winning people to your way of thinking?
See pages 105-169.
3. What are the nine ways to change people without giving offense or
arousing resentment? *See pages 170-194.*

DALE CARNEGIE

github.com/drugdata/D3R/wiki



GitHub, Inc. [US] | <https://github.com/drugdata/D3R/wiki>

drugdata / D3R

Watch 8 Unstar 3 Fork 2

Code Issues 35 Pull requests 0 Projects 0 Wiki Insights

Home

Chris Churas edited this page on Jan 31, 2017 · 99 revisions

Continuous Evaluation of Ligand Pose Prediction (CELPP) [CELPP](#) is a weekly automated cross-docking challenge enabled by the early PDB release of small molecule & protein sequences.

This wiki provides technical documentation on how [CELPP](#) works.

Participate in the Challenge

Additional information can be found by visiting: <https://drugdesigndata.org/about/celpp>

Pages 57

Main Page

Participate in the Challenge


Definitions


Developer topics

Getting Started

CELPPade Tutorial Video





Written CELPPade Tutorial


 Search





An Introduction to CELPPade

51 views

 0  0  SHARE 

 **Drug Design Data Resource**
Published on Nov 4, 2016


 

This video shows the basics of implementing a CELPP contestant server using the CELPPade template code. The code from this video is fully functional and competes CELPP weekly runs.

- The code from this video is hosted at https://github.com/drugdata/tutorial_...
- CELPP software/structure hub: <https://github.com/drugdata/D3R/wiki>
- Overview of D3R CELPP: <https://drugdesigndata.org/about/celpp>
- Be sure to follow the D3R twitter for updates! <https://twitter.com/drugdesigndata>

CELPPade Tutorial ☆

File Edit View Insert Format Tools Table Add-ons Help Last edit was on September 15, 2017



Purpose	1
To run the finished product immediately	2
Overview	3
"Scientific" vs. "Technical" molecule preparation	3
My simple AutoDock Vina workflow	5
Helper scripts	6
Ligand scientific preparation	6
Protein scientific preparation	6
Docking	7
Getting the template code and implementing the workflow	7
ftp_config	8
internal_autodockvina_contestant_protein_prep.py	9
internal_autodockvina_contestant_ligand_prep.py	11
internal_autodockvina_contestant_dock.py	13
Running your workflow	17
Installation	17
Running locally on test data	18
Running the weekly challenge	19
Setting up automatic weekly runs	19
Uploading the package to GitHub	19

Purpose

This document walks through the steps that created the [internal_autodockvina_contestant](#) package using the [CELPPade](#) template. D3R runs this package each week to simulate a participant using an AutoDock Vina-based pose prediction workflow. It is intended for:

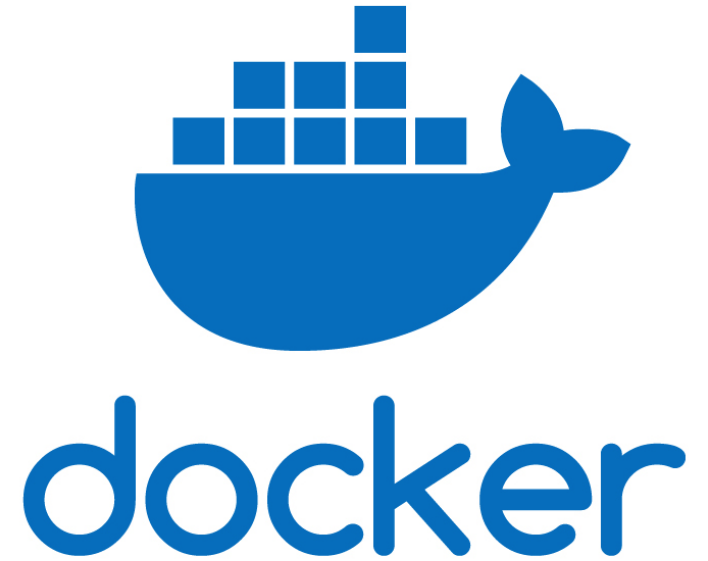
- Users interested in running this package on their own machine
- CELPP participants who want to understand the CELPPade template
- CELPP participants who want to begin workflow development with an already-functional package

Handling Different Computer Environments

- Machine images
- Cloud Implement-able



Cloud Credits
for Research



Moving Forward

- Engage community
 - Add participants
 - Share results
 - Enable sharing of methods (containerization)
 - Reach out to MolSSI and BioSimSpace
- Enrich the results dataset
 - Which methods work best for flexible molecules? Deep vs. shallow pockets?
 - What are the best methods to evaluate predictions?
- On pace to reach ~1000 targets per year
- Toward affinity predictions

Acknowledgements

- The CELPP Team
 - Chris Churas
 - Mike Chiu
 - Victoria Feher
 - Shuai Liu
 - Rob Swift
 - Rommie Amaro
 - Michael Gilson
- D3R and the early CELPP participants
- The Protein Data Bank team
- NIH U01GM111528

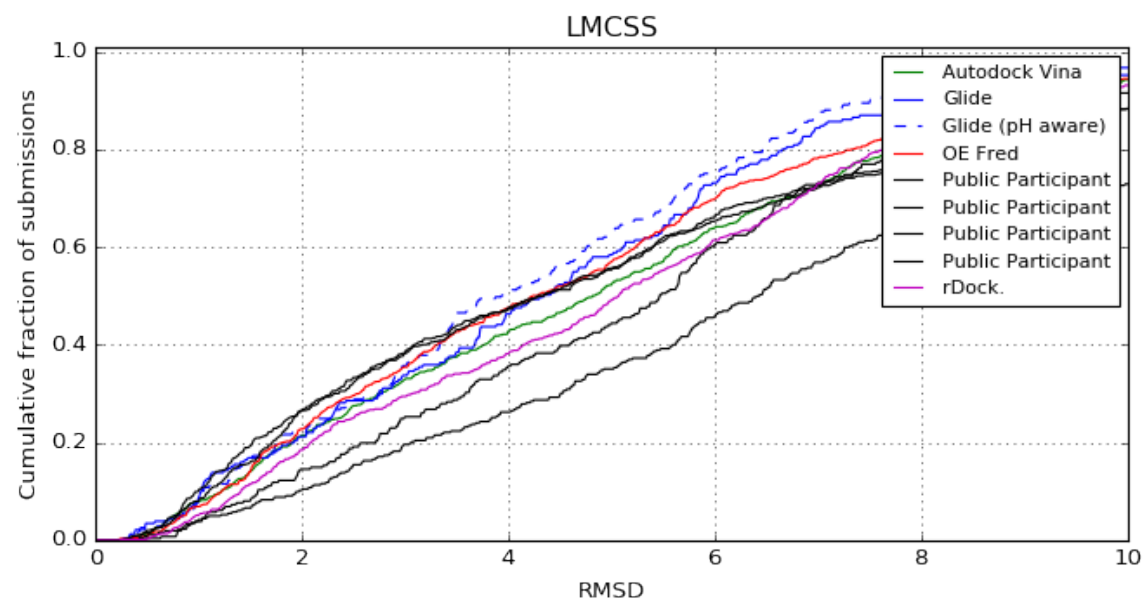
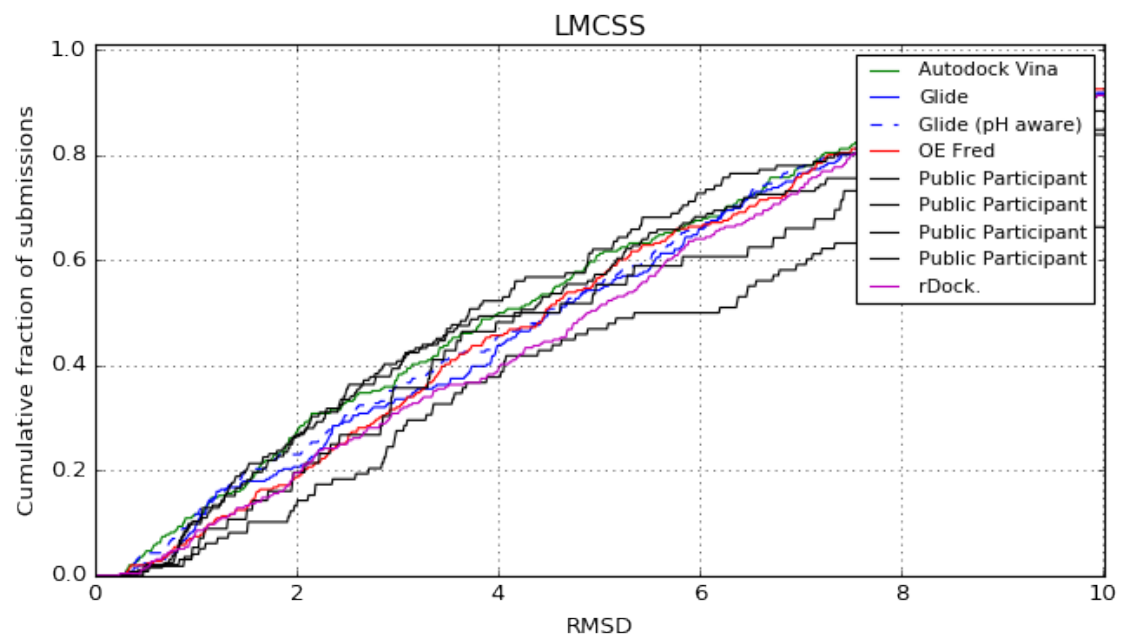
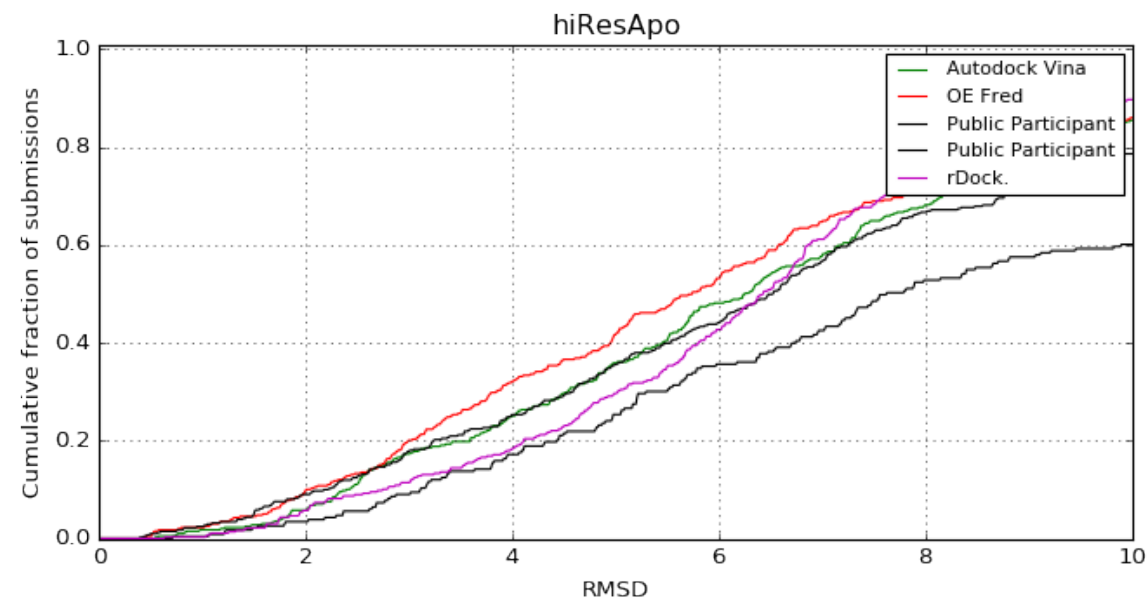
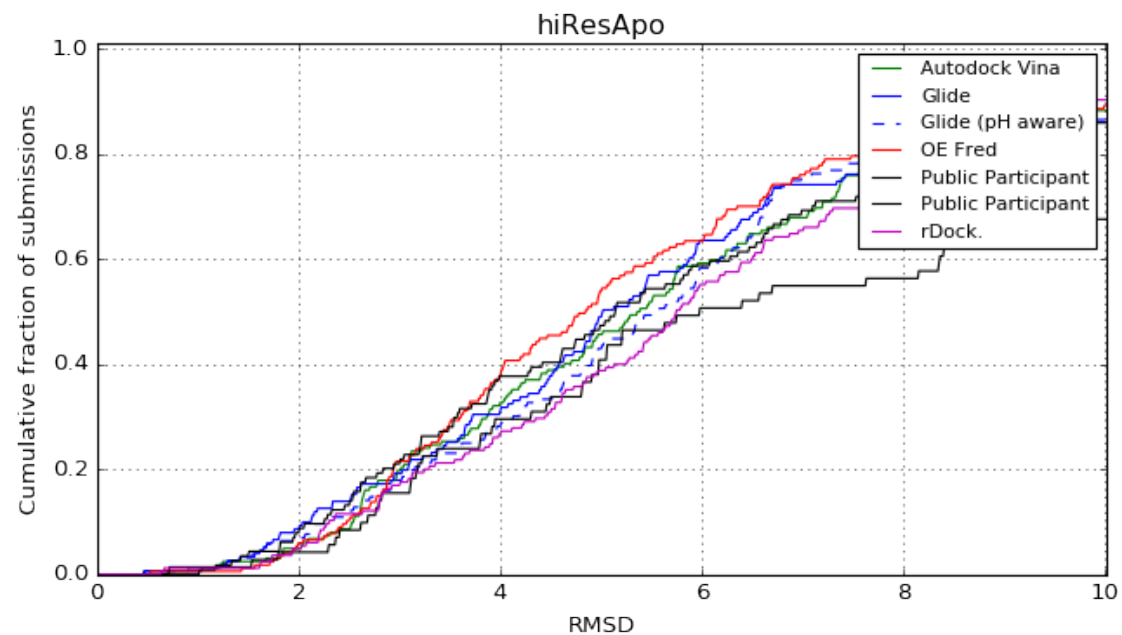


Evaluation and Results – Coming Soon

- Website visualization
- Pandas dataframes
 - Enrich with fields like candidate ligand overlap size, protein classification, ligand charge/description, # of torsions
- Anonymity handling
- Evaluation algorithm

```
RMSD_list = []
for crystal_chain in crystal_structure:
    predicted_complex = merge(predicted_chain, predicted_ligand)
    Align predicted_complex to crystal_chain
    aligned_predicted_protein, aligned_predicted_ligand = split(predicted_complex)
    for heavy_atom_mapping in atom_symmetries(crystal_ligand, predicted_ligand):
        RMSD_list.append(this_mapping_RMSD)
Return min(RMSD_list)
```

- Search for other sources of error (eg. Incorrect pocket predictions)
- Greater throughput will allow hyperparameter optimization for docking

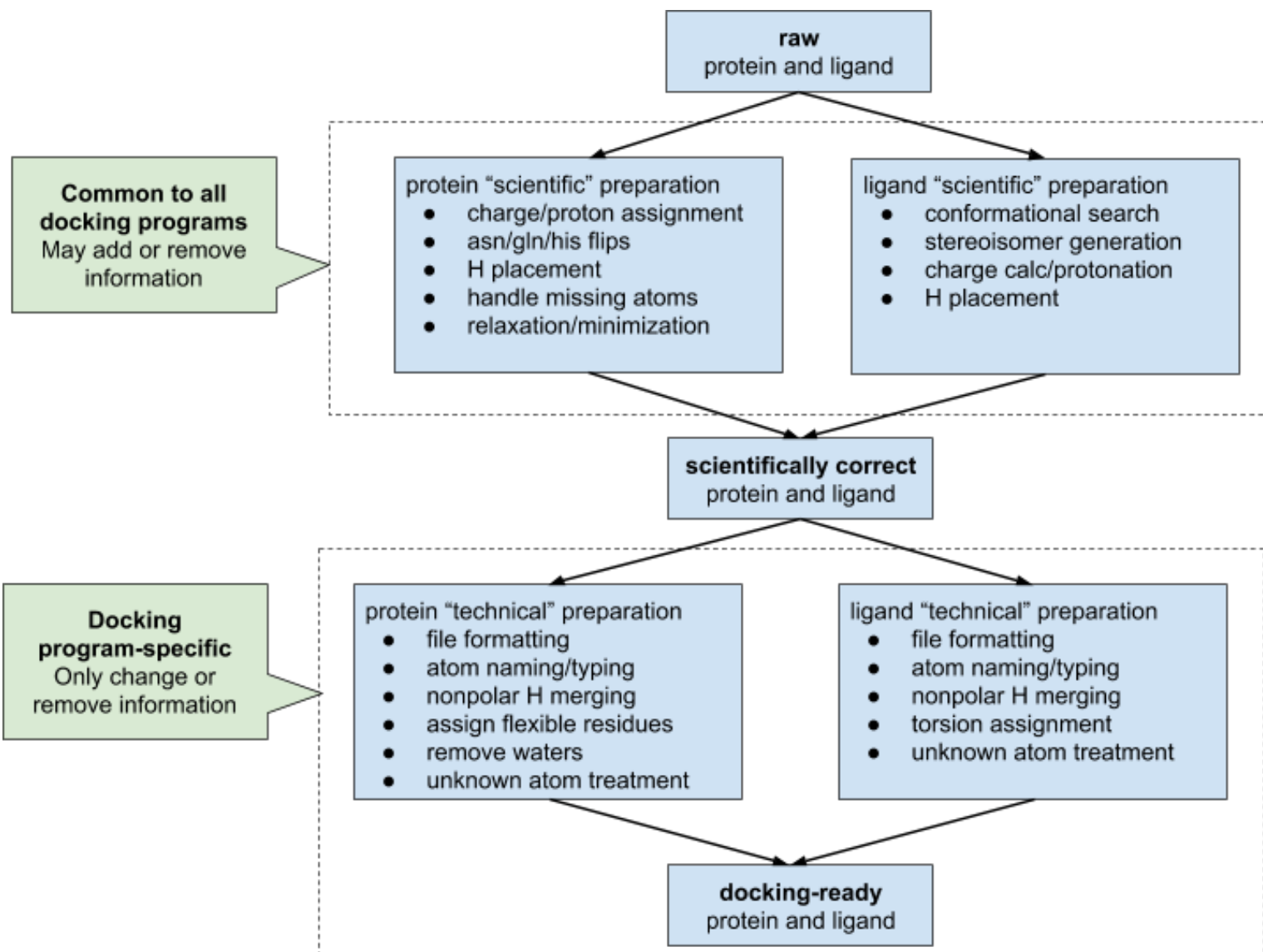


What does a docking workflow need to do?

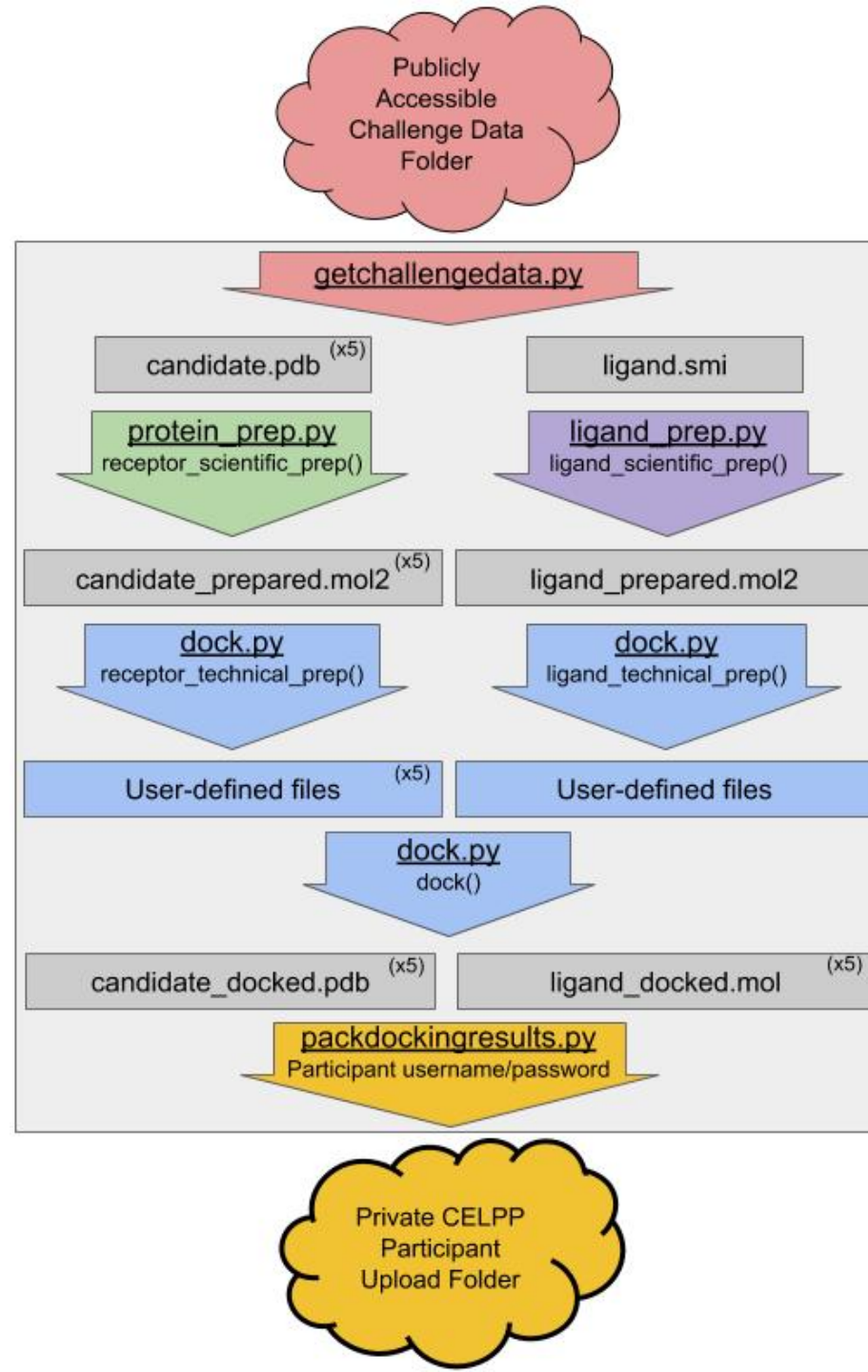
D3R provides existing PDB structure, ligand SMILES/InChI, and pocket coordinates

- Protein
 - Missing atom replacement
 - Proton assignments
 - Correct file formatting
 - Cofactor parameterization
 - Decisions about keeping water
 - Assign flexible residues
- Ligand
 - 3D conformer generation
 - Correct file formatting
 - Atom typing/parameterizing
 - Charge
 - Protonation
- Running docking

Molecule Handling in a Modular Docking Workflow



CELPP-ade



Download/upload format

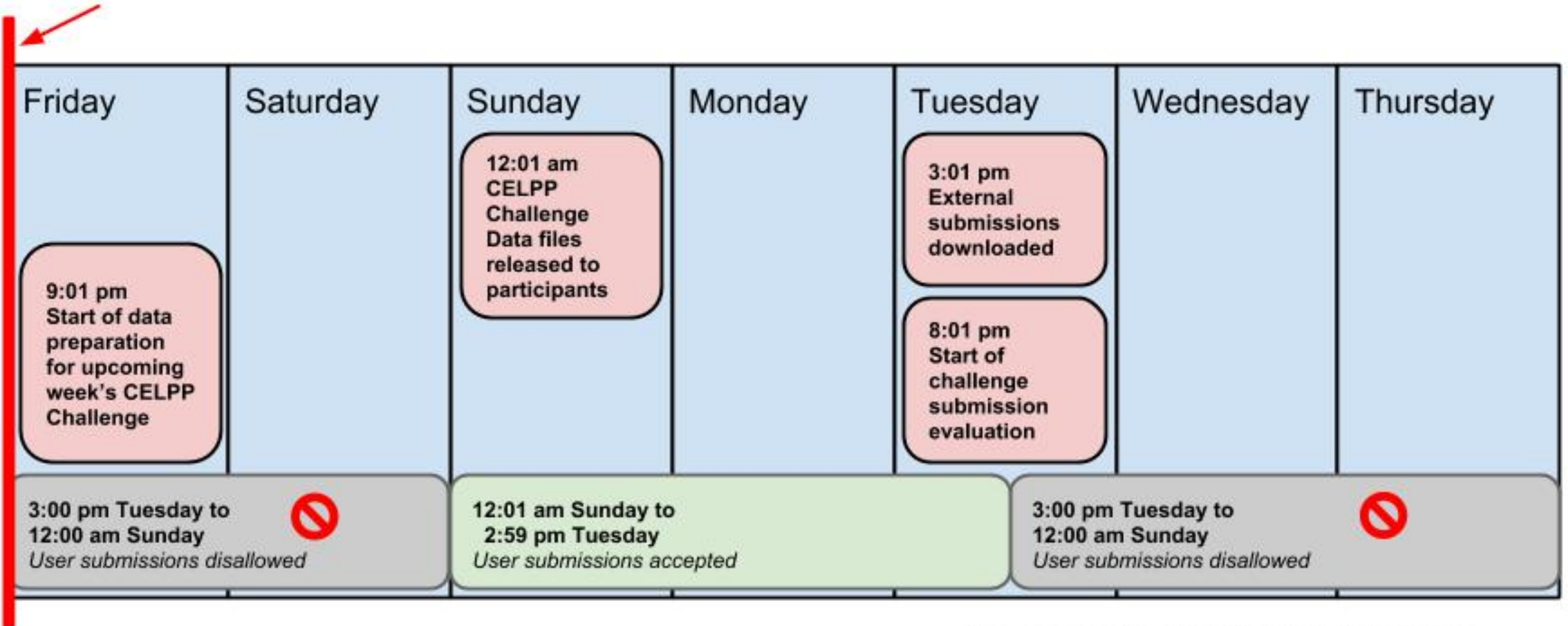
- We're very strict about this!

What CELPPade provides

- `getchallengedata.py`
- `packdockingresults.py`
- Scientific prep, technical prep, and docking functions
 - Think of these as a “for” loop over all the molecules
 - Each function knows some details of the target (pH, ligand structure, etc)
- These are entirely a convenience and you don't have to use them

The CELPP Week

CELPP week start



Times are in Pacific Daylight Time

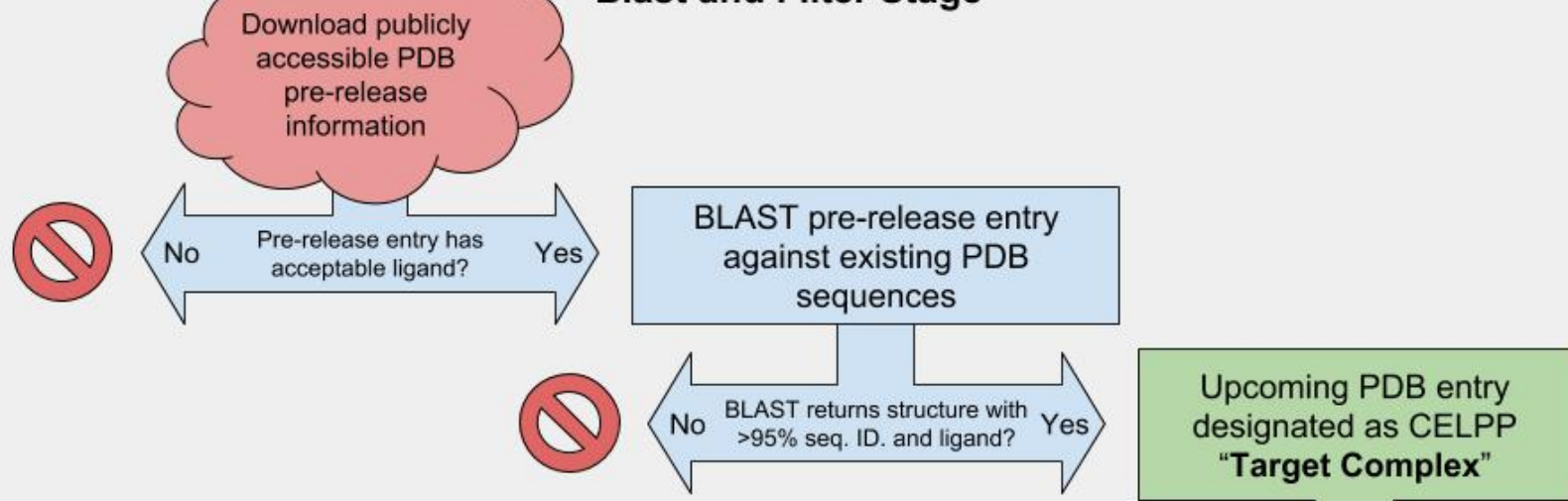
Pandas

```
In [12]: valid = [target for target in PTR.index.levels[1] if (PTR[:,target,'LMCSS_ori_distance']<5).any()]
PTR.loc[idx['33567_oefred-smallbox',valid,'LMCSS']]
```

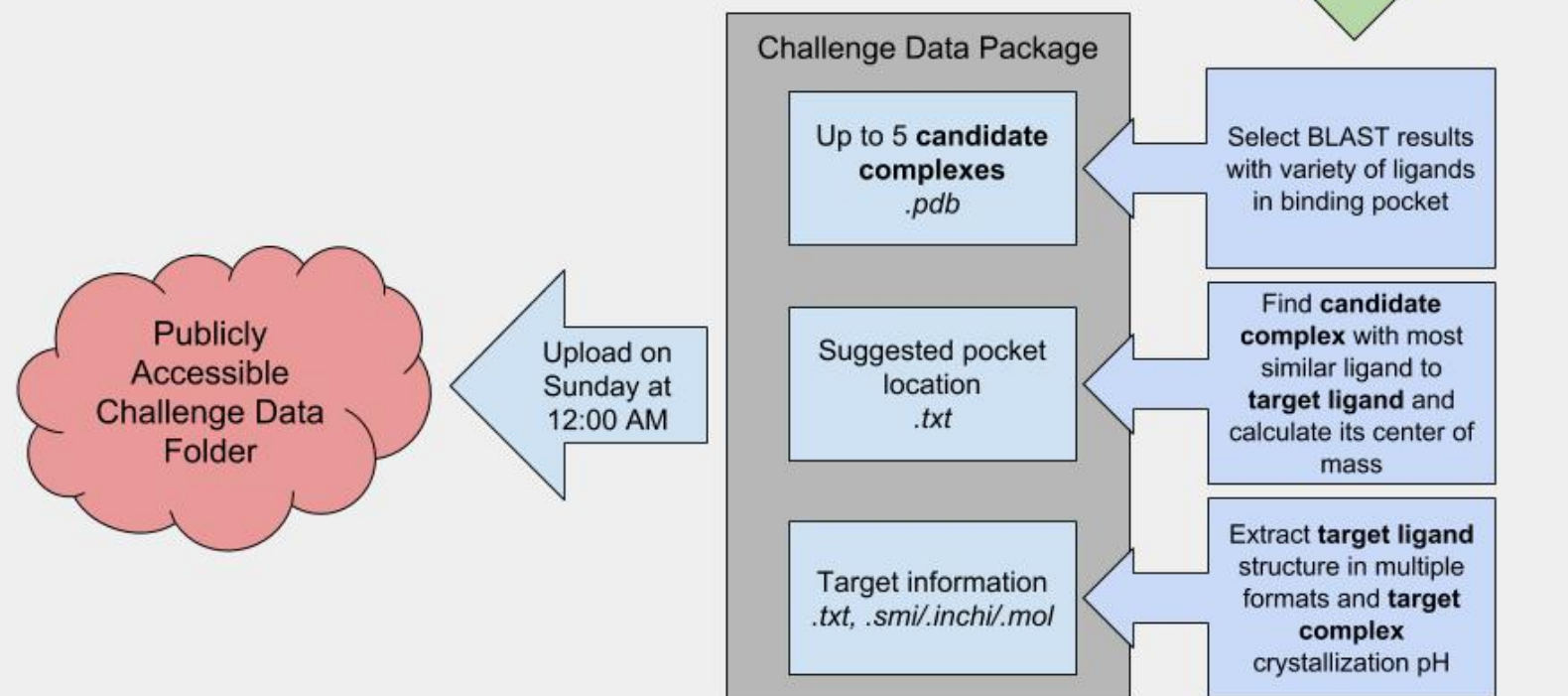
```
Out[12]:
```

participant	target	category	
33567_oefred-smallbox	1fcz	LMCSS	0.774
	5g2b	LMCSS	1.319
	5g48	LMCSS	NaN
	5g4q	LMCSS	NaN
	5g57	LMCSS	7.751
	5g5j	LMCSS	6.114
	5giy	LMCSS	5.423
	5gkw	LMCSS	1.491
	5gmm	LMCSS	3.930
	5gmn	LMCSS	4.473
	5gn7	LMCSS	4.321
	5gn9	LMCSS	4.611
	5grw	LMCSS	9.562
	5gso	LMCSS	NaN
	5gsw	LMCSS	4.306
	5gud	LMCSS	4.332
	5gzw	LMCSS	2.315
	5h0b	LMCSS	3.167
	5h21	LMCSS	2.169
	5h22	LMCSS	1.908
	5h4j	LMCSS	6.192
	5h5g	LMCSS	1.524
	5h5h	LMCSS	1.602
	5h5i	LMCSS	2.086
	5h5o	LMCSS	6.823
	5isp	LMCSS	NaN
	5isq	LMCSS	NaN
	5j20	LMCSS	1.472
	5j27	LMCSS	5.207
	5j2x	LMCSS	0.564
	...		
	6bku	LMCSS	1.032
	6ble	LMCSS	0.765
	6bmb	LMCSS	3.096
	6bmj	LMCSS	6.026
	6bnj	LMCSS	6.545
	6bql	LMCSS	8.740

Blast and Filter Stage



Challenge Data Generation Stage



Challenge Data Generation Stage

