The Continuous Evaluation of Ligand Pose Prediction (CELPP) Challenge

Jeffrey Wagner[1], Shuai Liu[1], Christopher Churas[1], Victoria Feher[1], Stephen Burley[2,3], Robert Swift[1], Michael Chiu[1], Michael Gilson[1]*, Rommie Amaro[1]*

1. Drug Design Data Resource (D3R), University of California San Diego, La Jolla, California
2. RCSB Protein Data Bank, Center for Integrative Proteomics Research, Institute for Quantitative Biomedicine, Rutgers Cancer Institute of New Jersey, Rutgers University, The State University of New Jersey, Piscataway, New Jersey
3. RCSB Protein Databank, San Diego Supercomputer Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California

Introduction

The Drug Design Data Resource (D3R; www.drugdesigndata.org) is an NIH funded resource aimed at providing benchmark datasets and blinded challenges to assist in the evaluation and improvement of computational algorithms such as small molecule protein docking.

Ligand pose prediction algorithms have enabled researchers to make great strides in the discovery of novel therapeutics. However, the leading participants in pose prediction challenges often report using different algorithms and diverse strategies. In the interest of developing pose prediction best practices, D3R has sought a high-volume source of blinded protein-ligand structure data.

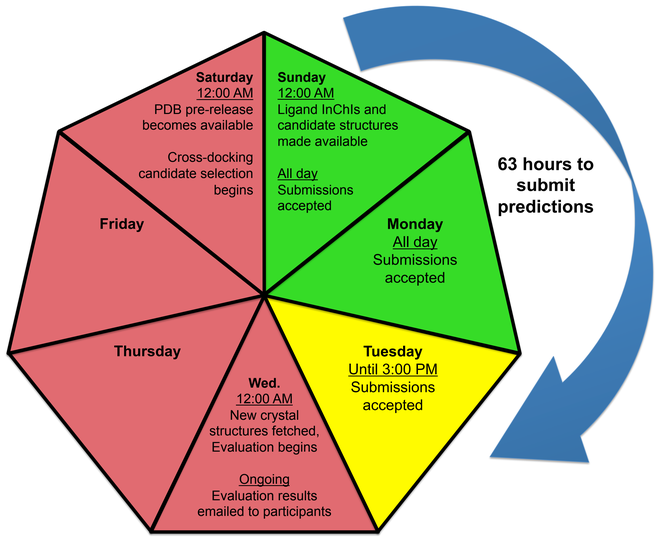
One novel source of benchmarking data is the ongoing release of new protein-ligand complexes in the Protein Data Bank (PDB). In the current release process, new PDB entries are announced five days prior to their 3D coordinates becoming available. Therefore, D3R has established the Continuous Evaluation of Ligand Pose Prediction (CELPP) challenge to evaluate whether the drug design community is able to predict protein-ligand crystal structures before their coordinates are released.

A D3R server hosts these weekly challenges by: 1) selecting appropriate pre-release docking targets, 2) making these targets available to CELPP participant prediction servers, and 3) evaluating the accuracy of the participant predictions.

CELPP challenge participants implement their pose prediction workflows on their own servers. To lower the barrier to participation, D3R provides CELPPade, a Python framework for contestant servers that downloads the weekly challenge package, applies the contestant's prediction workflow, and uploads the predictions back to D3R for evaluation.

This presentation discusses the CELPP challenge, the CELPPade framework, and pathways for community participation.

The CELPP Week



CELPP Overview

```
5UGB  1     SGEAPNQALLRIL…TEFKKIKVLGSGAFGTVYK

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

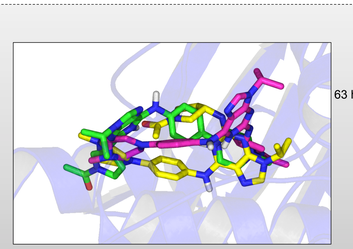
Protein Data Bank

Structure pre-release notifications

```
query, 5ugb
ph, N/A
center  -52.035,   -0.147,  -23.507
ligand, 8BM
inchi, InChI=1S/C25H35N9O/c1-17(2)34-16-26…,30)/t20-/m0/s1
size, 36
rotatable_bond, 5
LMCSS, 4r5s, FI3, chain: A, (size: 48, mcss_size: 25, resolution:  3.0)
SMCSS, 4wkq, CSX, chain: A, (size: 9, mcss_size: 5, resolution:
hiResHolo, 5hg8, 634, chain: A, (resolution: 1.42)
hiResApo, 2eb2
hiTanimoto, 2jiu, AEE, chain: A, (tanimoto_similarity: 0.48, resolution: 3.0
hiTanimoto, 2j6m, AEE, chain: A, (tanimoto_similarity: 0.48, resolution:  3.1)
…
hiTanimoto, 2ity, IRE, chain: A, (tanimoto_similarity: 0.45, resolution: 3.42)
```

D3R Server

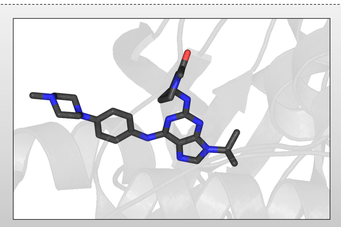
"Challenge data" package



63 hours for predictions

CELPP Participant 1 | CELPP Participant 2 | CELPP Participant 3

Ligand pose predictions



Protein Data Bank → Revealed ligand pose → D3R Server

Evaluation emails

Target_PDB_RMSD:	LMCSS	SMCSS	hiResApo	hiResHolo	hiTanimoto
5ugb	10.163	9.770	8.371	8.351	9.225

Target_PDB_RMSD:	LMCSS	SMCSS	hiResApo	hiResHolo	hiTanimoto
5ugb	4.595	8.658	10.063	10.032	7.638

Target_PDB_RMSD:	LMCSS	SMCSS	hiResApo	hiResHolo	hiTanimoto
5ugb	8.649	7.642	10.690	9.317	3.811

CELPP Resources

CELPP Wiki | Challenge data packages | CELPPade | D3R

https://github.com/drugdata/D3R/wiki, all code available on PyPI

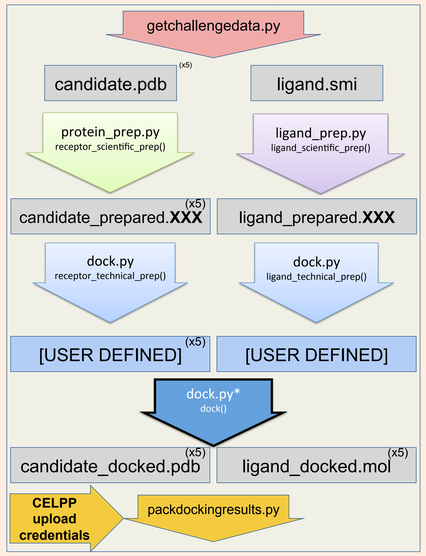
Pose Prediction Workflow Template – "CELPPade"

- **Modular** - Separates prep and docking code to encourage recombination
- **Shareable** - Template code is Github-ready
- **Minimal** - Only one function must be written
- **Simple** - Provides challenge data handlers
- **Supported** - Imports core D3R CELPP code

```
protein_prep.py
  • receptor_scientific_prep()
ligand_prep.py
  • ligand_scientific_prep()
dock.py
  • receptor_technical_prep()
  • ligand_technical_prep()
  • *dock()
packdockingresults.py
  • CELPP upload credentials
```

* dock() is a required function

CELPPade flow diagram



Receptor "Scientific" Preparation
Missing loop replacement
Protonation
Charge calculation
Solvent/HETATM deletion
Resolve multiple-occupancy residues

Ligand "Scientific" Preparation
SMILES/InChI → 3D structure
Protonation
Charge calculation
Conformer generation

Receptor "Technical" Preparation
File format conversion
Docking grid generation

Ligand "Technical" Preparation
File format conversion

Separating preparation stages into different files enables modular recombination and fast improvement of existing workflows.

Future Work

- Additional/improved scoring metrics
- Scaling on Amazon cloud
- Result website development
- Identification of best-performing participants by structure class
- "Recombination" of top prediction workflows
- Development of transferrable workflow machine images
- Unrestricted protein structure selection

Acknowledgements

We would like to acknowledge support from NIH U01GM111528 as well as valuable discussions with the RCSB PDB and the early CELPP participants.