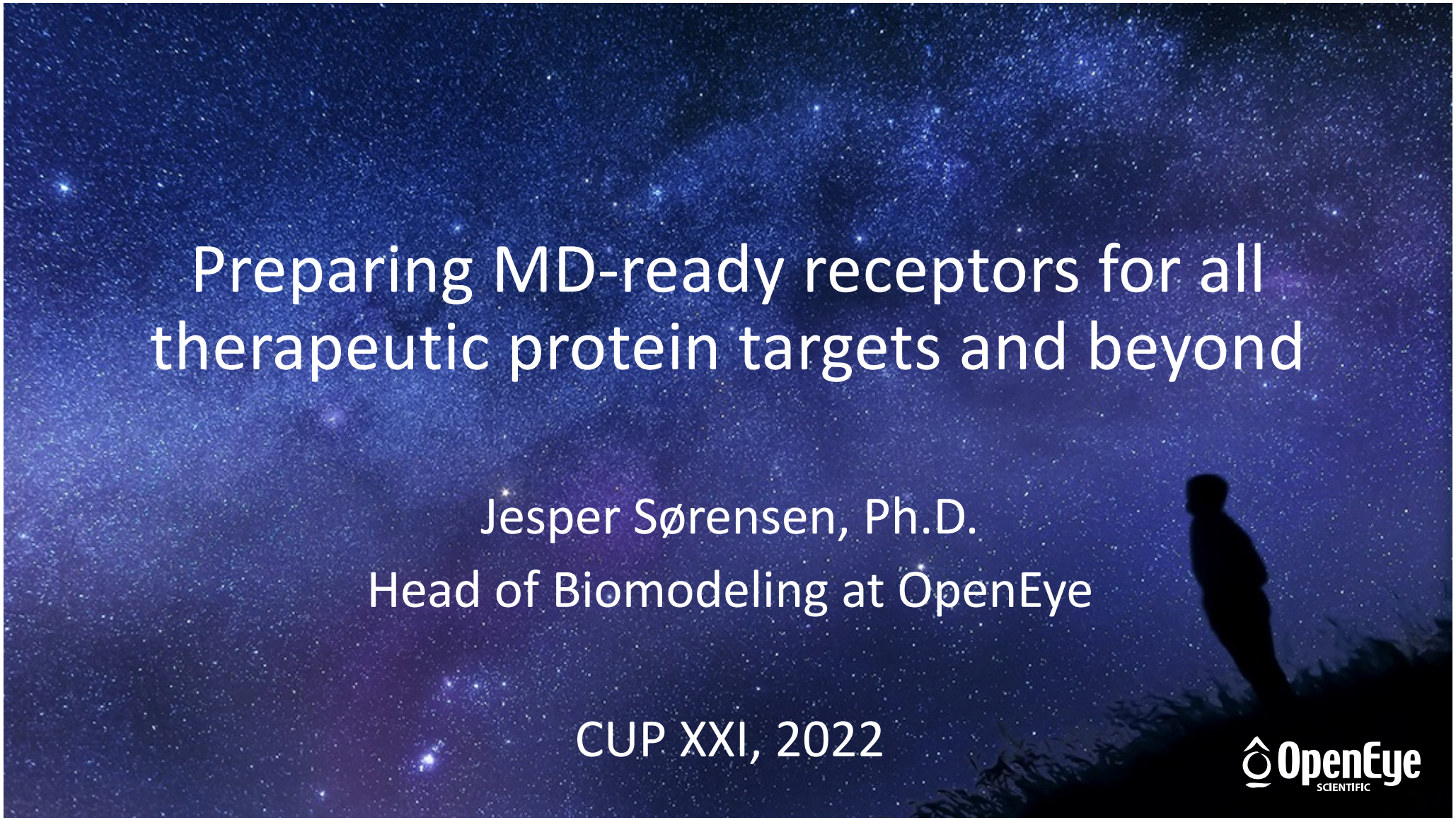
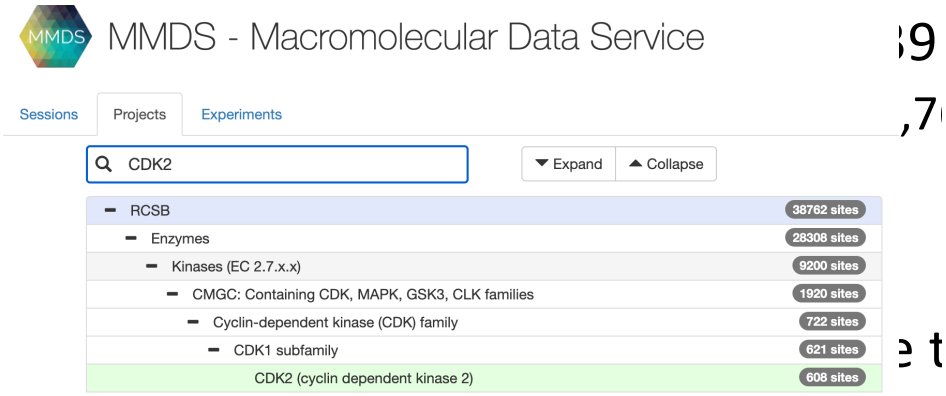

Preparing MD-ready receptors for all therapeutic protein targets and beyond
Jesper Sørensen, Ph.D.
Head of Biomodeling at OpenEye
CUP XXI, 2022
OpenEye
SCIENTIFIC

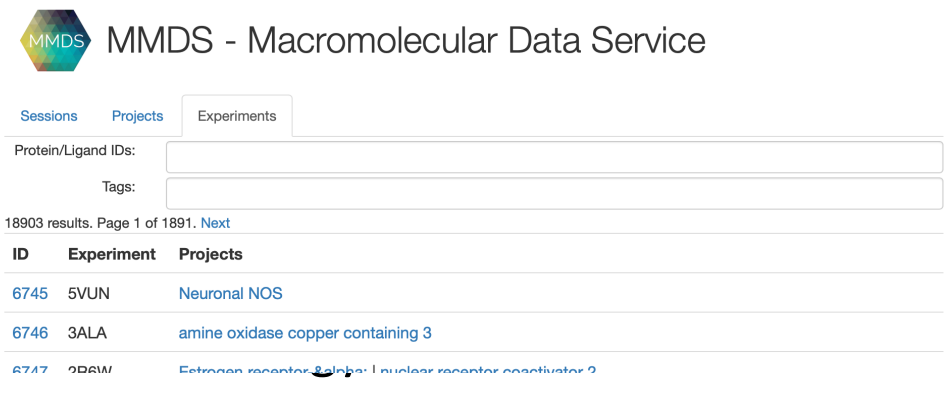
The Goal

- Prepare the entire PDB
  - PDB: 187,844 experiments
- Prepare AlphaFold2
  - AF2v2-human: 23,391 models
- Organize the data
  - Target and family classification:
    - So many choices…
- Mine the data
  - So many things to ask

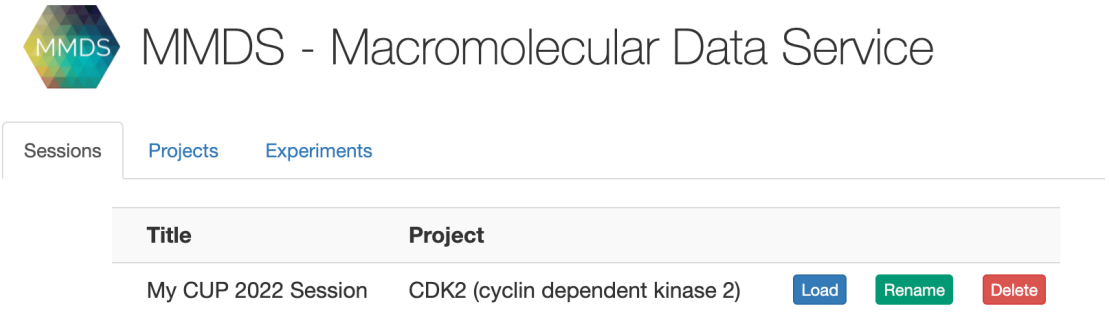
All data sources as of 3/4/2022

The first iteration – focus on pharma targets



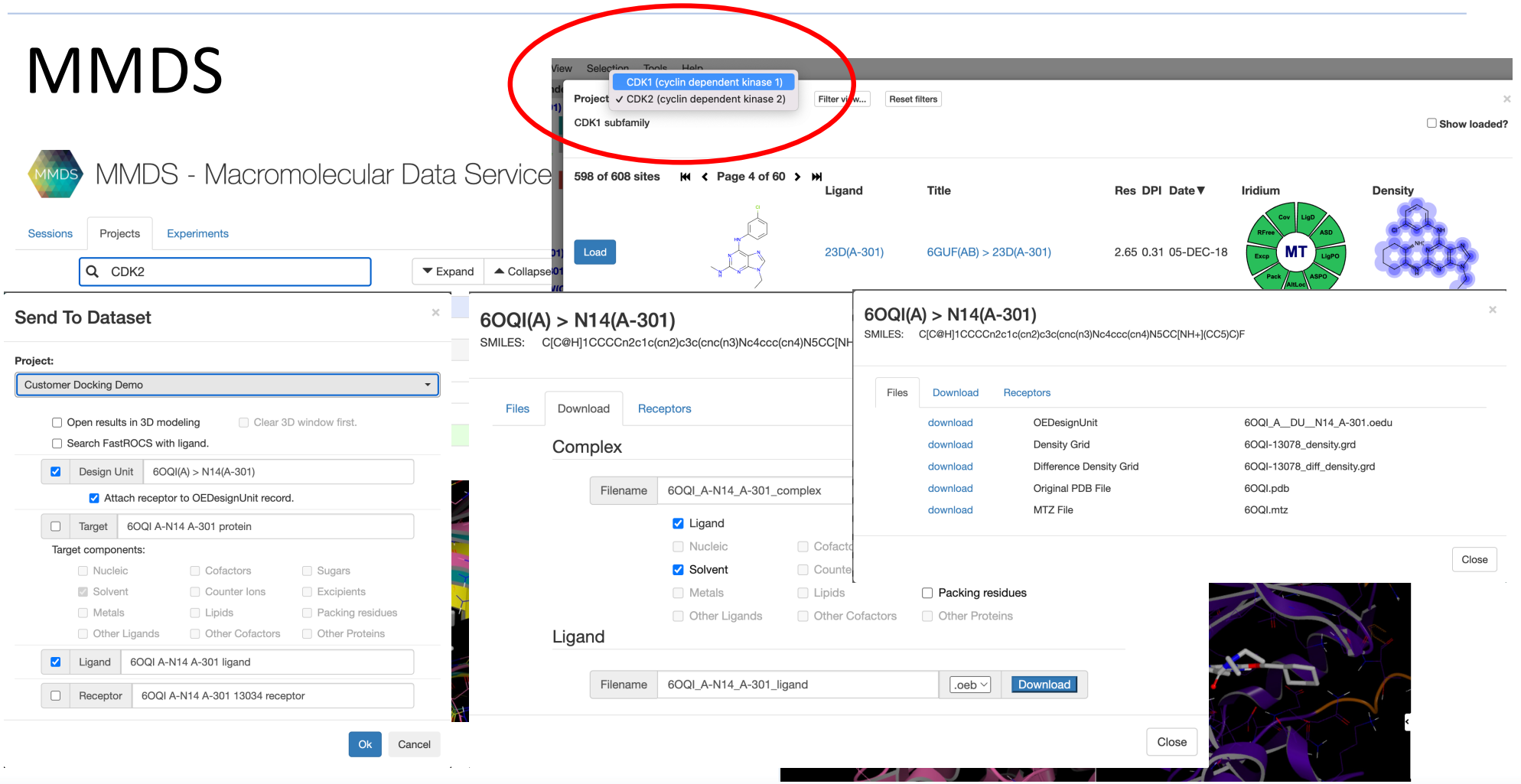


- 2,673 Targets in the family tree



OpenEye SCIENTIFIC

MMDS



OpenEye SCIENTIFIC

Acknowledgements

- Zhe Mei
- Addison Smith

- Bob Tolbert
- David LeBard

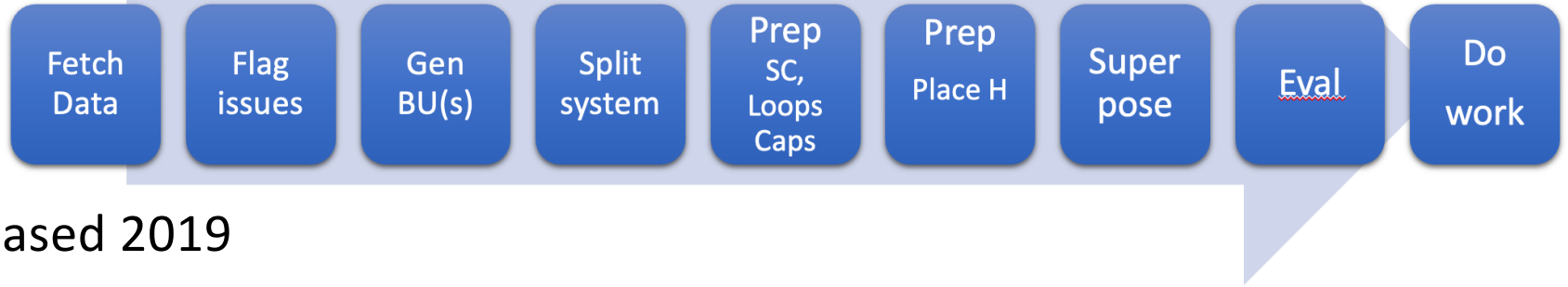
- Mike Word
- Stan Wlodek
- Krisztina Boda
- Greg Warren
- Geoff Skillman
- Ant Nicholls
- Christopher Bayly
- Gaetano Calabro
- Hyesu Jang
- She Zhang
- Orion/Cloud Teams
- Customer Success Group

OpenEye SCIENTIFIC

Connecting data sources, compute, and analysis

OpenEye SCIENTIFIC

SPRUCE – OpenEye's Protein Preparation Tool

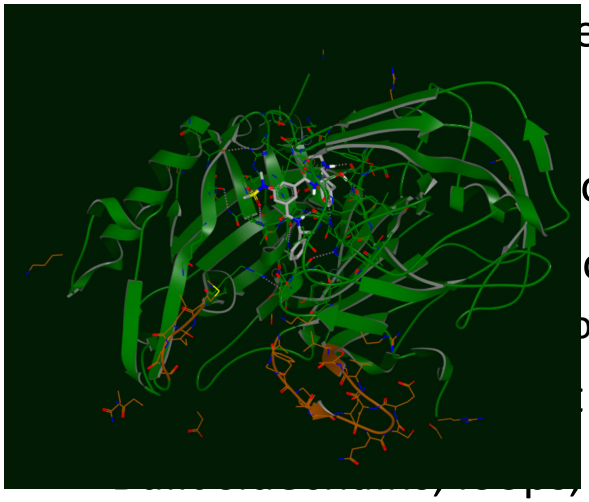


- 1.0 Released 2019

- Reads a variety of inputs as PDB/MMCIF files + MTZ Files
  - Experiments: X-ray, NMR, cryoEM
  - Models: Homology models, AI models, one-off models

- Forms the basis for a lot of our tools
  - OEDOCKING, SZMAP, SZYBKI, ZAP, STMD, NES, Enhanced Sampling

Target Preparation Goals



ed

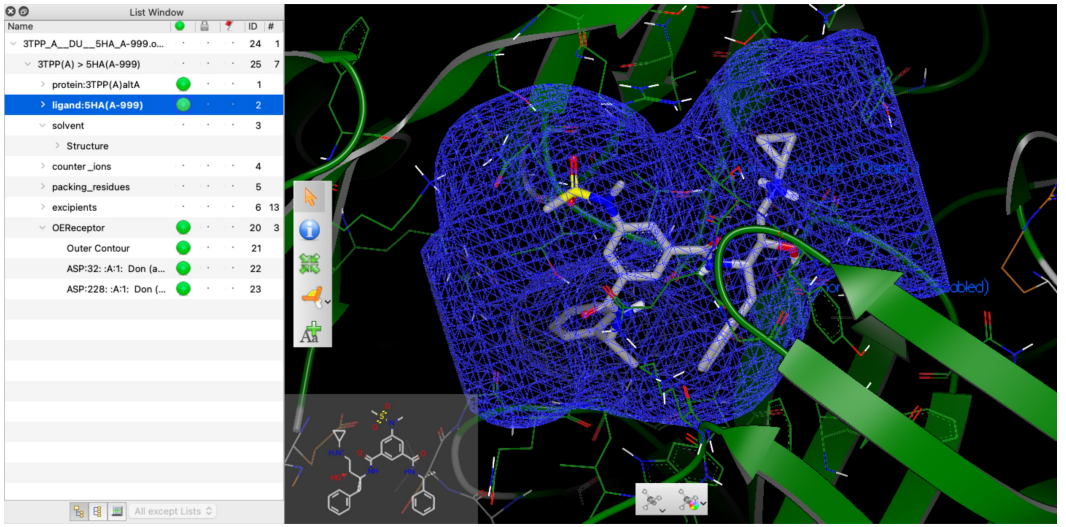
orized

optimized

omer states

(site residues)

tails, caps if needed



bound ligand	Active Site Depiction	B-Factor Depiction	du_title	Iridium Depiction	Ligand Density Depiction	pdb_code
Ligand from 3T...	Legend		3TPP(A) > 5HA(...	Cov, LigD, ASD, LigPO, ASPO, AltLoc, Pack, Excp, RFree, HT		3tpp

OpenEye SCIENTIFIC

Iridium Categorization

Active site:
Residues within 5A from the ligand
Co-factors within 5A from the ligand

Density coverage:

	HT	MT	NT
Ligand	> 0.90	> 0.5	< 0.5
Active Site	> 0.95	> 0.5	< 0.5

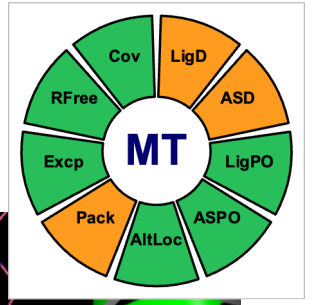
- HT → MT
  - DPI > 0.50
  - Alternate conformations in the ligand or active site residues
  - Any ligand heavy atom with an occupancy < 0.90
  - Any active site heavy atom with occupancy < 0.5
  - Packing residues: Interactions with ligand
  - Excipients: Interactions with ligand
  - Covalently bound ligand, using covalent definition from OEInteractionHints

* → NT

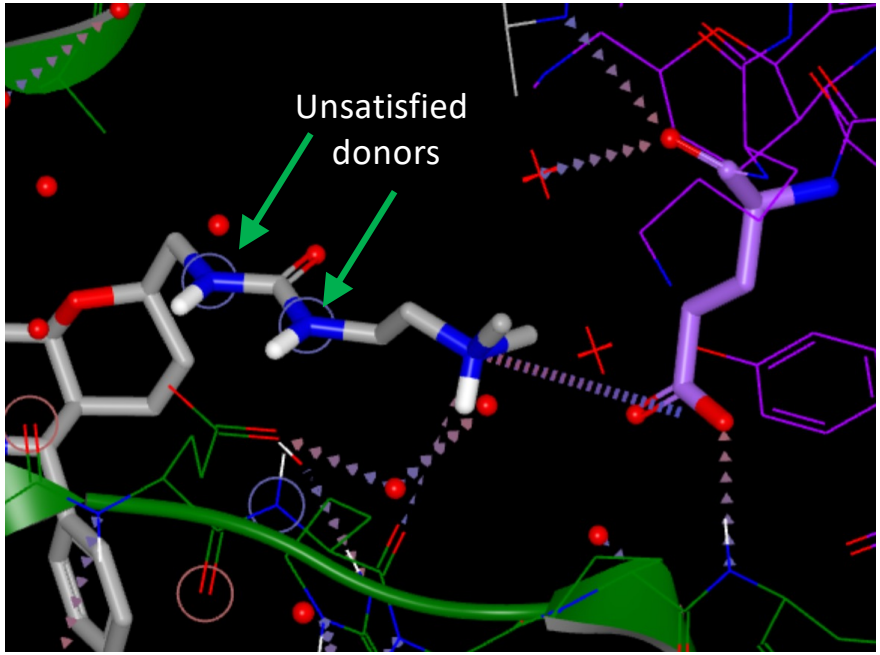
Rfree > 0.45 and Resolution < 3.5

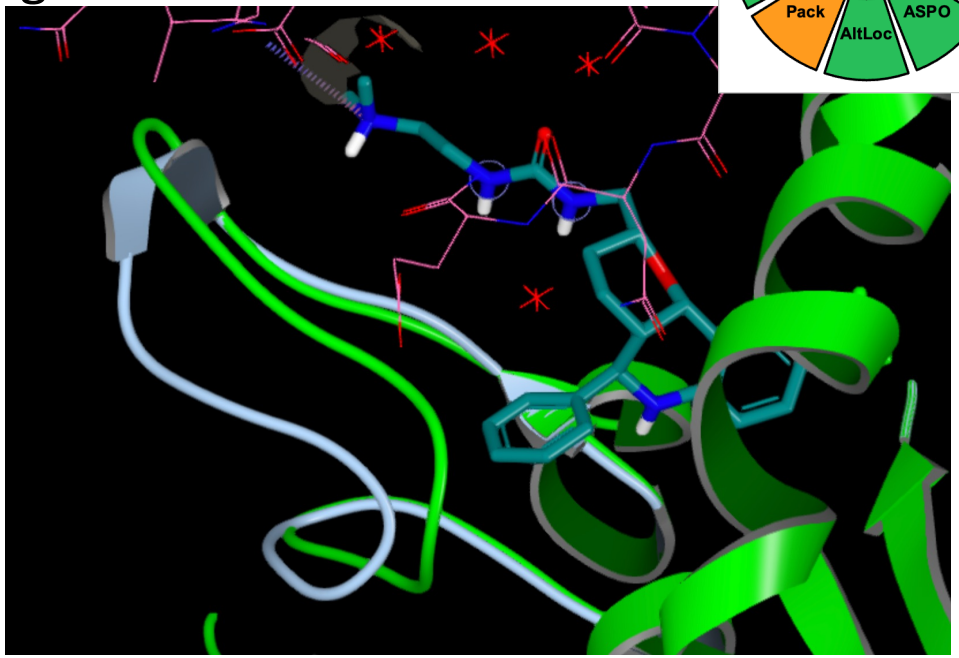
Iridium criteria: Warren et al., *Drug Disc. Today*, **17**, 1270 (2012).

Packing residues are important



EG5 - 3L9H – Packing residue interaction



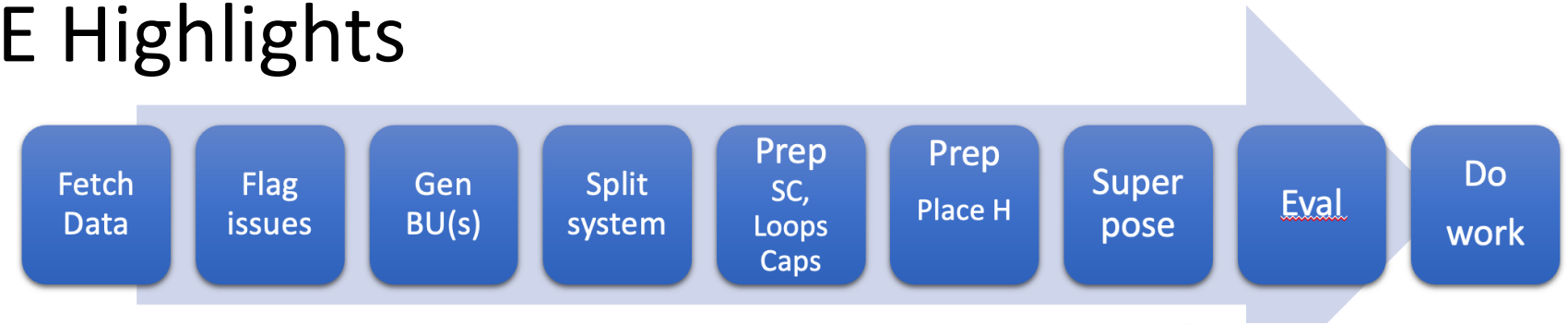


Conformation of ligand from chain A is strongly affected by residue from chain B.

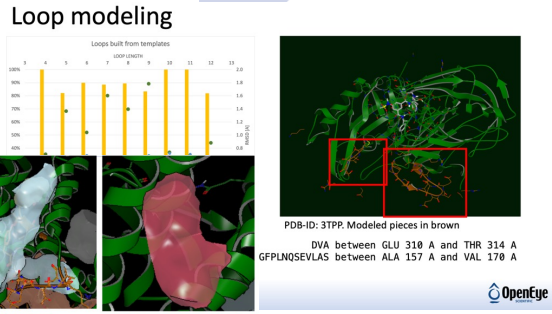
Original PDB coordinates
Re-modeled in Merck kga dataset

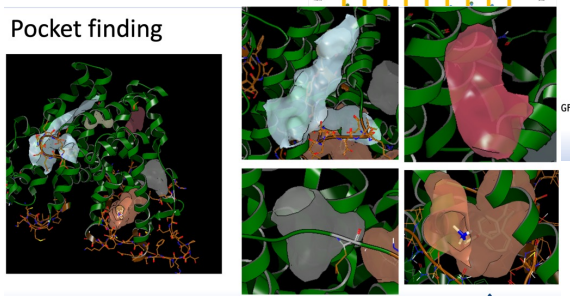
OpenEye SCIENTIFIC

SPRUCE Highlights



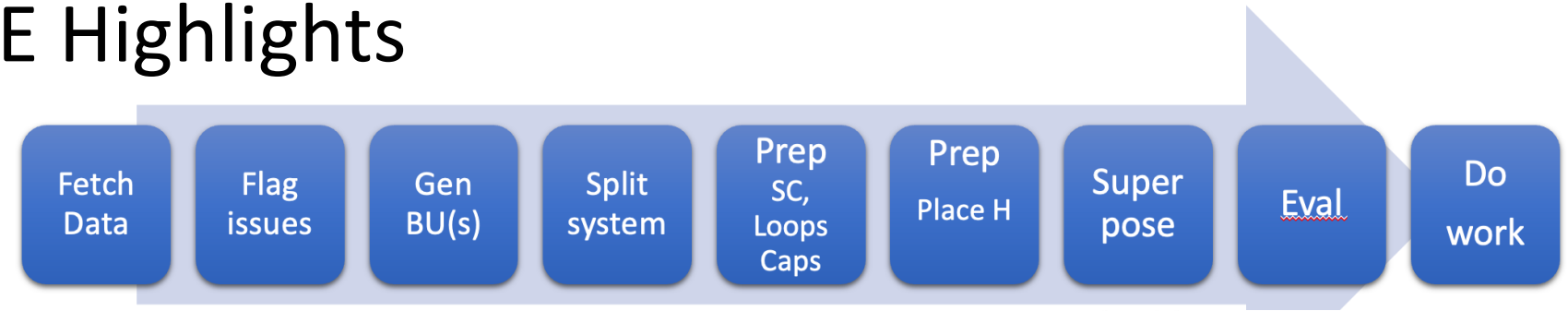
- Spring 2020
  - Added Loop Modeling (database available, build your own)
- Fall 2020
  - Pocket search with OEPocket
- Spring 2021
  - Superposition API was finalized
  - Protein residue perception improvements with capping group support





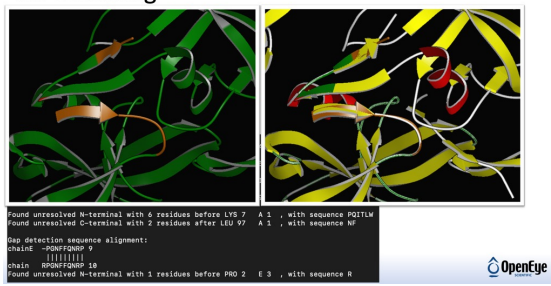
OpenEye SCIENTIFIC

SPRUCE Highlights



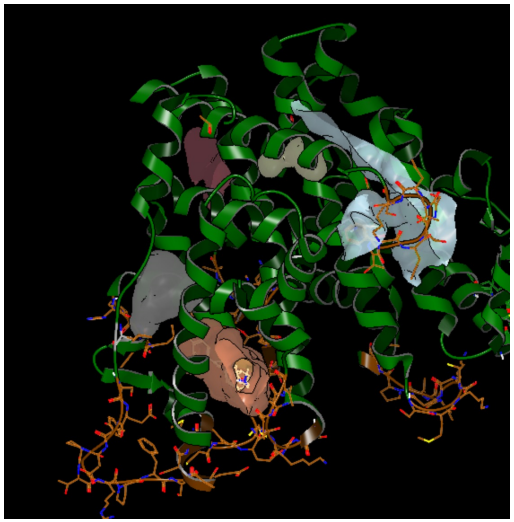
- Fall 2021
  - Improved Hydrogen placement for Histidine
  - Improved Tautomer selection for ligands and co-factors
  - Added Tail modeling (same database approach)
  - SiteHopper Toolkit released
  - SiteHopper as a superposition method (joined Seq, Site Seq, DDM, SSE)
- Spring 2022 upcoming
  - Biounit extraction performance
  - Logging/messaging prep issues to customers (pre-filter and post prep)

Tail Modeling

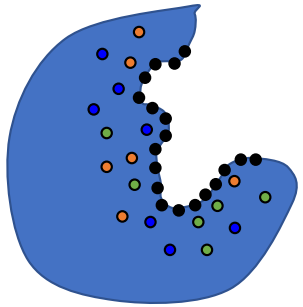


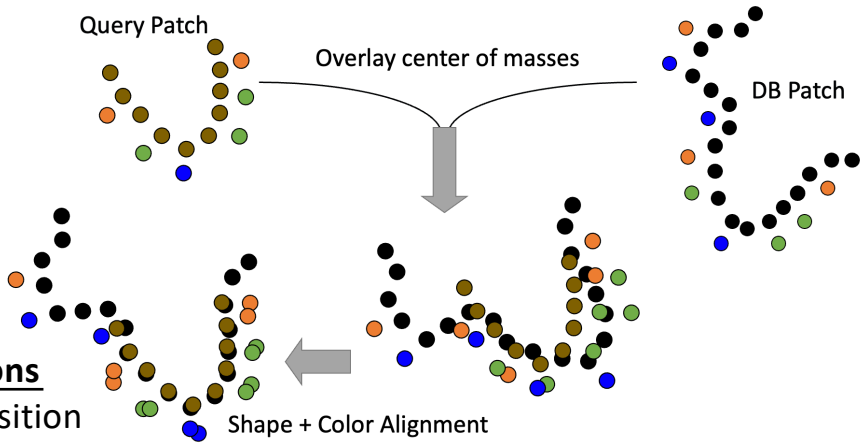
OpenEye SCIENTIFIC

SPRUCE Highlights



Pocket Finding with OEPocket
Liganded and prospective Sites





SiteHopper Applications
SAR Analysis/Superposition
Homology Modeling
Selectivity
Drug Repurposing
Off-target effects
Target Fishing

Organization Data / OpenEye Data / SiteHopper Collections ☆

SiteHopper Guide to Pharmacology Alternate Sites Patch DB Collection OEv1.0	316 GB
SiteHopper Guide to Pharmacology Patch DB Collection OEv1.0	25 GB
AlphaFold2v1 Human 20211019 Alternate Sites Patch DB Collection OEv1.0	67 GB

OpenEye SCIENTIFIC

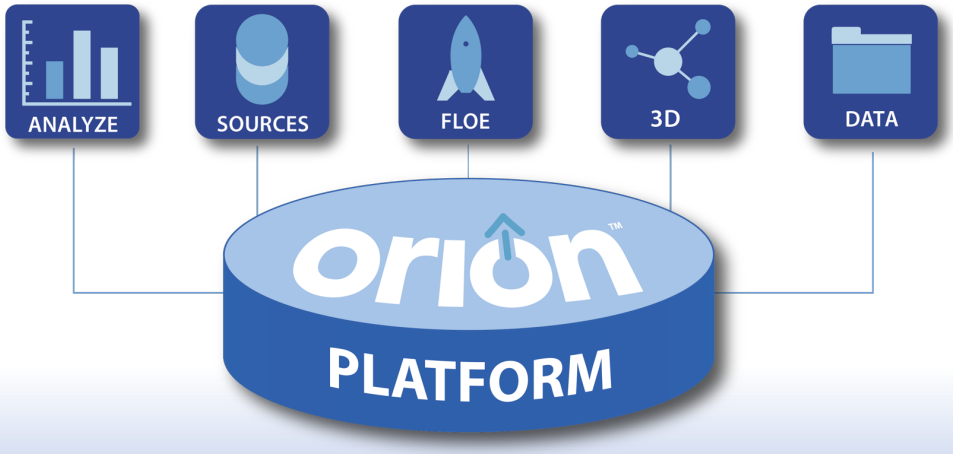
The Goal

- Prepare the entire PDB
  - PDB: 187,844 experiments
- Prepare AlphaFold2
  - AF2v2-human: 23,391 models
- Organize the data
  - Target and family classification:
    - Guide to Pharmacology
    - …
- Mine the data
  - So many things to ask

All data sources as of 3/4/2022

Good Data Science Practices

- Versioned workflows to access and prepare data
- Versioned datasets
- Consistent generation & update of datasets
- Reproducible science, reproducible validation
- Democratization of good science



OpenEye SCIENTIFIC

The Floes

and data organization

MMDS 1. Make RCSB PDB Collection

Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

MMDS 2. Generate Guide to Pharmacology Datasets

Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

MMDS 3. Target Reference Picker

Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

MMDS 4. Family reference picker

Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

MMDS 5. Structure Prep

Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

MMDS 6. Add family data to MMDS

Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

MMDS 7. Add context data to MMDS

Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

MMDS 8. Add structures to MMDS

Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

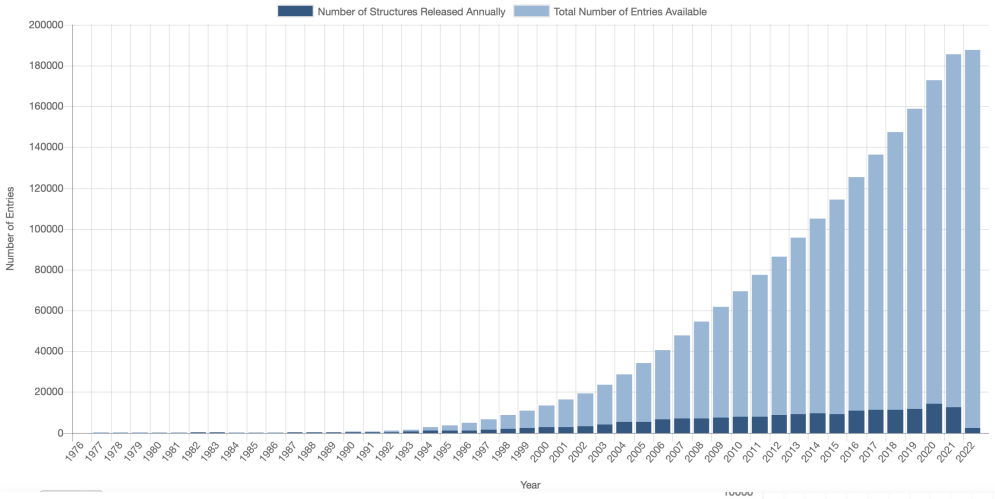
MMDS 9. Add receptors to MMDS

Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

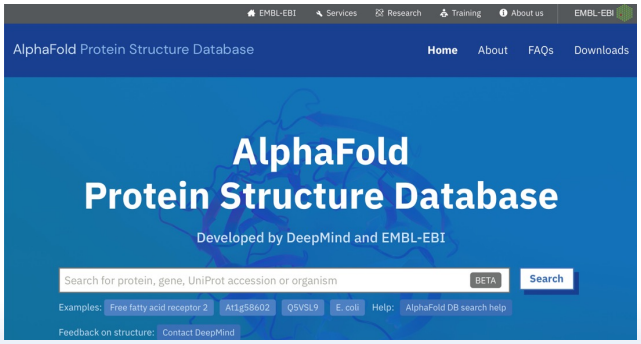
MMDS 10. Add sequence alignments to MMDS

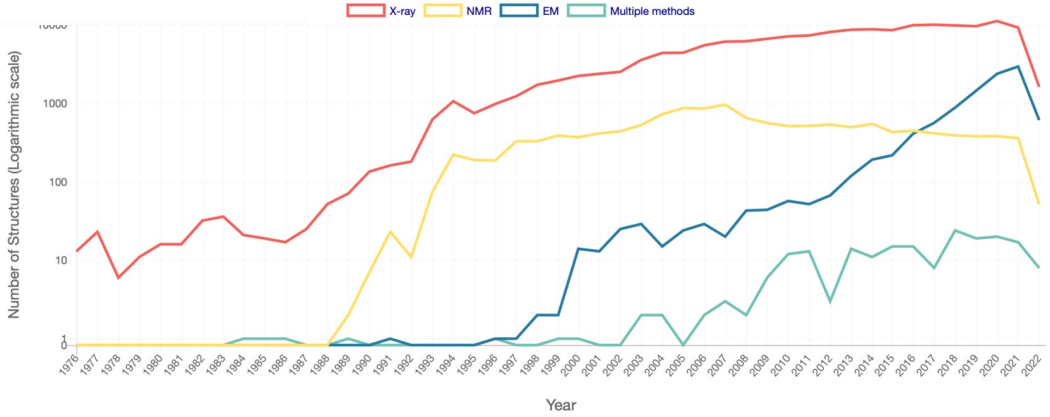
Package: OpenEye Biomodeler Floes v0.1.1b16
Show description

The Input Data



- Public data
  - ~188K structures
  - growth ~10-15k per year
- Customer data
  - Large pharma – 20k+ structures,
    Growth 1-2k per year
  - Small pharma – 10-500 structures
    Growth 1-100 per year

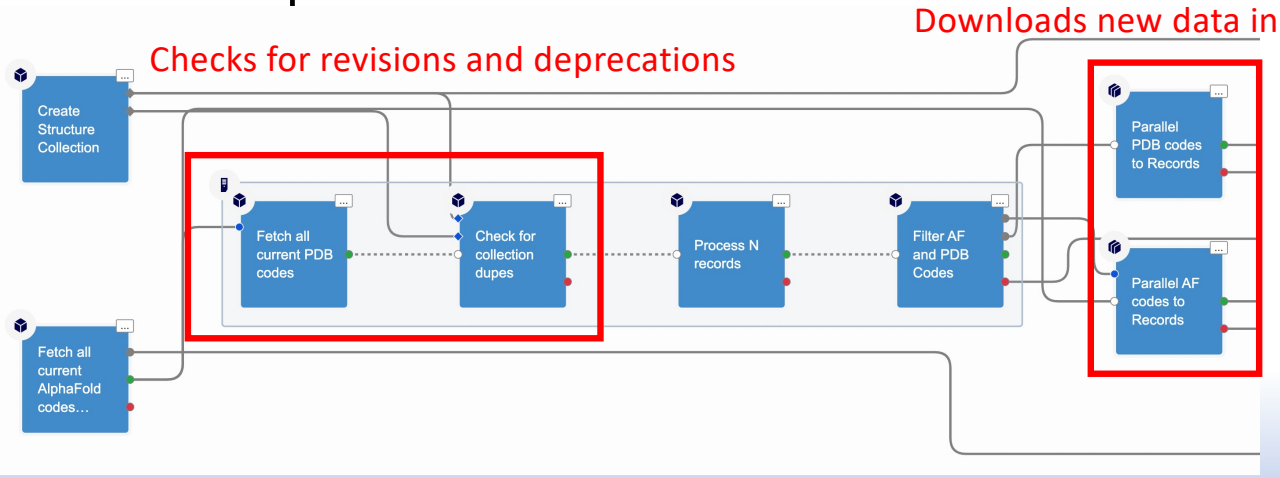




OpenEye SCIENTIFIC

Updateable Source Collection

- Orion storage of each PDB/MMCIF and MTZ file
- The workfloe makes the data updatable
  - New revisions of PDBs, Removal of deprecated codes
  - New versions of AlphaFold models
- Currently **764 GB of data**
- EM Maps inclusion is planned



OpenEye SCIENTIFIC

Data organization

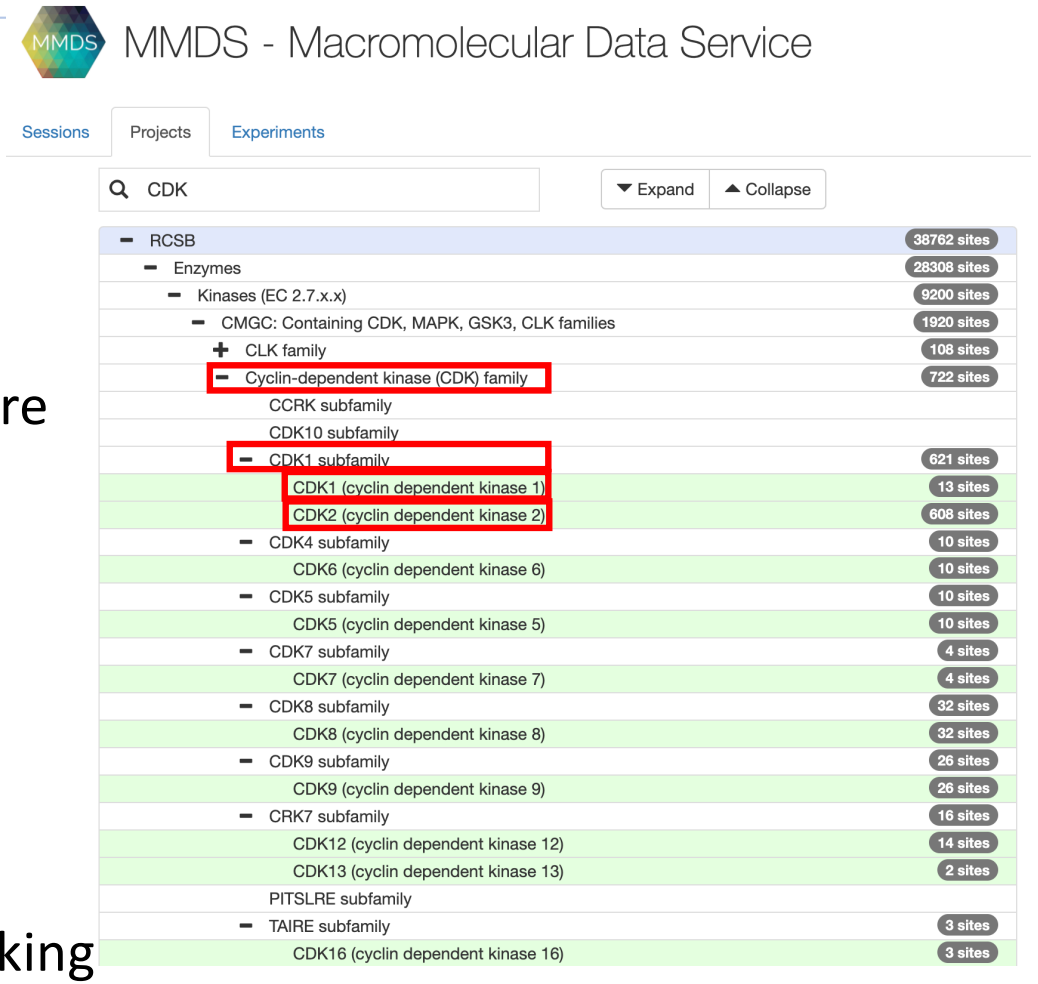
- No convention on “organized” n
- A target can be organized by
  - Class (e.g. enzymes, receptors, tra
  - Evolutionary relationship (i.e. sequ
  - Therapeutic area (e.g. cancer, neu
  - … your scheme here …

- We have some great public databases
  - PDB at RCSB, Alphafold and UniprotKB at EBI
  - Lots of cross-links and annotations per structure
  - They are indexed, but **not “organized”**

OpenEye SCIENTIFIC

Target definition

- Targets are defined in a hierarchy
- Each target has a reference structure
  - Defines the biological unit/form
  - Defines the frame of reference
- Each family (parent) can have a ref
  - If each child below can superpose
- Automated reference structure picking
  - Based on consensus of structures for the target



Problems in first iteration…

- Targets not in GtoP
  - Viral targets*
  - Bacterial targets
  - e.g. Haemoglobin

Gag Polyprotein: P04585

Molecule processing

Feature key	Position(s)	Description	Actions	Graphical view	Length
Initiator methionine		Removed; by host (By similarity)			
Chain (PRO_0000261261)	2 – 1447	Gag-Pol polyprotein	Add BLAST		1446
Chain (PRO_0000042285)	2 – 132	Matrix protein p17 (By similarity)	Add BLAST		131
Chain (PRO_0000042286)	133 – 363	Capsid protein p24 (By similarity)	Add BLAST		231
Peptide (PRO_0000042287)	364 – 377	Spacer peptide 1 (By similarity)	Add BLAST		14
Chain (PRO_0000042288)	378 – 432	Nucleocapsid protein p7 (By similarity)	Add BLAST		55
Peptide (PRO_0000246710)	433 – 440	Transframe peptide (Sequence analysis)			8
Chain (PRO_0000042289)	441 – 500	p6-pol (Sequence analysis)	Add BLAST		60
Chain (PRO_0000038647)	501 – 599	Protease (By similarity)	Add BLAST		99
Chain (PRO_0000042290)	600 – 1159	Reverse transcriptase/ribonuclease H (By similarity)	Add BLAST		560
Chain (PRO_0000042291)	600 – 1039	p51 RT (By similarity)	Add BLAST		440
Chain (PRO_0000042292)	1040 – 1159	p15	Add BLAST		120
Chain (PRO_0000042293)	1160 – 1447	Integrase (By similarity)	Add BLAST		288

- Problems mapping using UniprotKB IDs
  - **Viral targets***

Entry	Entry name	Protein names	Gene names	Organism	Length
P04585	POL_HV1H2	**Gag-Pol polyprotein**	**gag-pol**	Human immunodeficiency virus type 1 group M subtype B (isolate HXB2) (HIV-1)	1,435
P03369	POL_HV1A2	**Gag-Pol polyprotein**	**gag-pol**	Human immunodeficiency virus type 1 group M subtype B (isolate ARV2/SF2) (HIV-1)	1,437
P05961	POL_HV1MN	**Gag-Pol polyprotein**	**gag-pol**	Human immunodeficiency virus type 1 group M subtype B (isolate MN) (HIV-1)	1,441
Q9QSR3	POL_HV1VI	**Gag-Pol polyprotein**	**gag-pol**	Human immunodeficiency virus type 1 group M subtype F1 (isolate VI850) (HIV-1)	1,430
Q9Q720	POL_HV1V9	**Gag-Pol polyprotein**	**gag-pol**	Human immunodeficiency virus type 1 group M subtype H (isolate VI991) (HIV-1)	1,436
Q75002	POL_HV1ET	**Gag-Pol polyprotein**	**gag-pol**	Human immunodeficiency virus type 1 group M subtype C (isolate ETH2220) (HIV-1)	1,439
Q89290	POL_HV193	**Gag-Pol polyprotein**	**gag-pol**	Human immunodeficiency virus type 1 group M subtype F1 (isolate 93BR020) (HIV-1)	1,430

OpenEye SCIENTIFIC

UniProt Features with Sequence Splitting

Sequence splitting occurs when a protein is split into non-overlapping features. This is easily seen in the HIV target Gag Polyprotein (P04591).

UniProtKB - P04591 (GAG_HV1H2)

Feature key	Position(s)	Description	Actions	Graphical view	Length
Initiator methionine		Removed; by host By similarity			
Chain (PRO_0000261216)	2 – 500	Gag polyprotein	Add BLAST		499
Chain (PRO_0000038593)	2 – 132	Matrix protein p17 By similarity	Add BLAST		131
Chain (PRO_0000038594)	133 – 363	Capsid protein p24 By similarity	Add BLAST		231
Peptide (PRO_0000038595)	364 – 377	Spacer peptide 1 By similarity	Add BLAST		14
Chain (PRO_0000038596)	378 – 432	Nucleocapsid protein p7 By similarity	Add BLAST		
Peptide (PRO_0000038597)	433 – 448	Spacer peptide 2 By similarity	Add BLAST		
Chain (PRO_0000038598)	449 – 500	p6-gag By similarity	Add BLAST		

- Uncategorized
  - Gag polyprotein
    - Capsid protein p27, alternate cleaved 2
    - Phosphorylated protein pp24
    - Capsid protein p24
    - Gag protein
    - Matrix protein p10
    - Matrix protein p15
    - Matrix protein p17
    - Protease p15

UniprotKB “target” is broken into smaller feature-targets that represent the druggable targets in HIV.

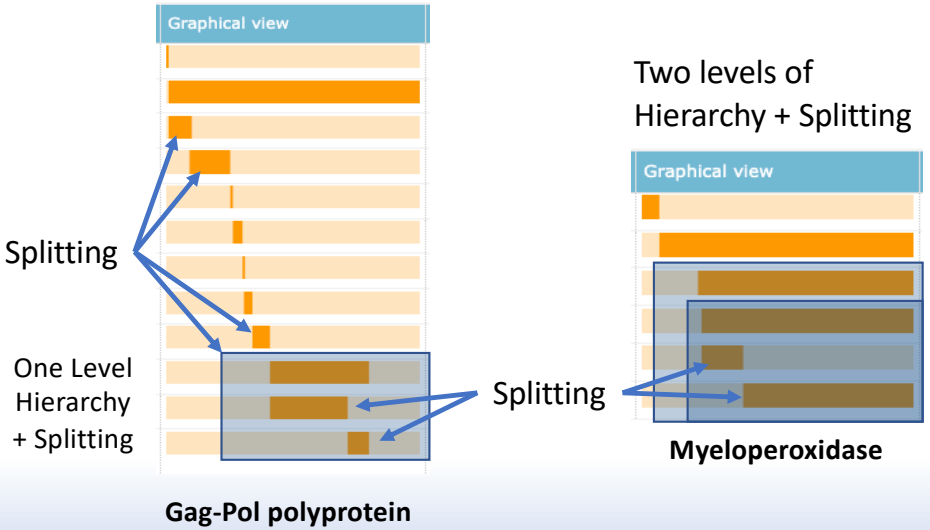
Updates the family and target tree in MMDS

OpenEye SCIENTIFIC

Complex Target Features and Inconsistencies

Combining both sequence splitting and sequence hierarchy can give much greater protein detail. Inconsistencies in UniProt submissions inappropriately convolutes.

The Gag-Pol polyprotein and Myeloperoxidase targets are complicated examples where sequence splitting occurs inside sequence hierarchies of variable depths.



Description	Actions	Graphical view
Ubiquitin-related	Add BLAST	
Ubiquitin	Add BLAST	
Ubiquitin	Add BLAST	
Ubiquitin	Add BLAST	
Ubiquitin	Add BLAST	
Ubiquitin	Add BLAST	
Ubiquitin	Add BLAST	
Ubiquitin	Add BLAST	
Ubiquitin	Add BLAST	

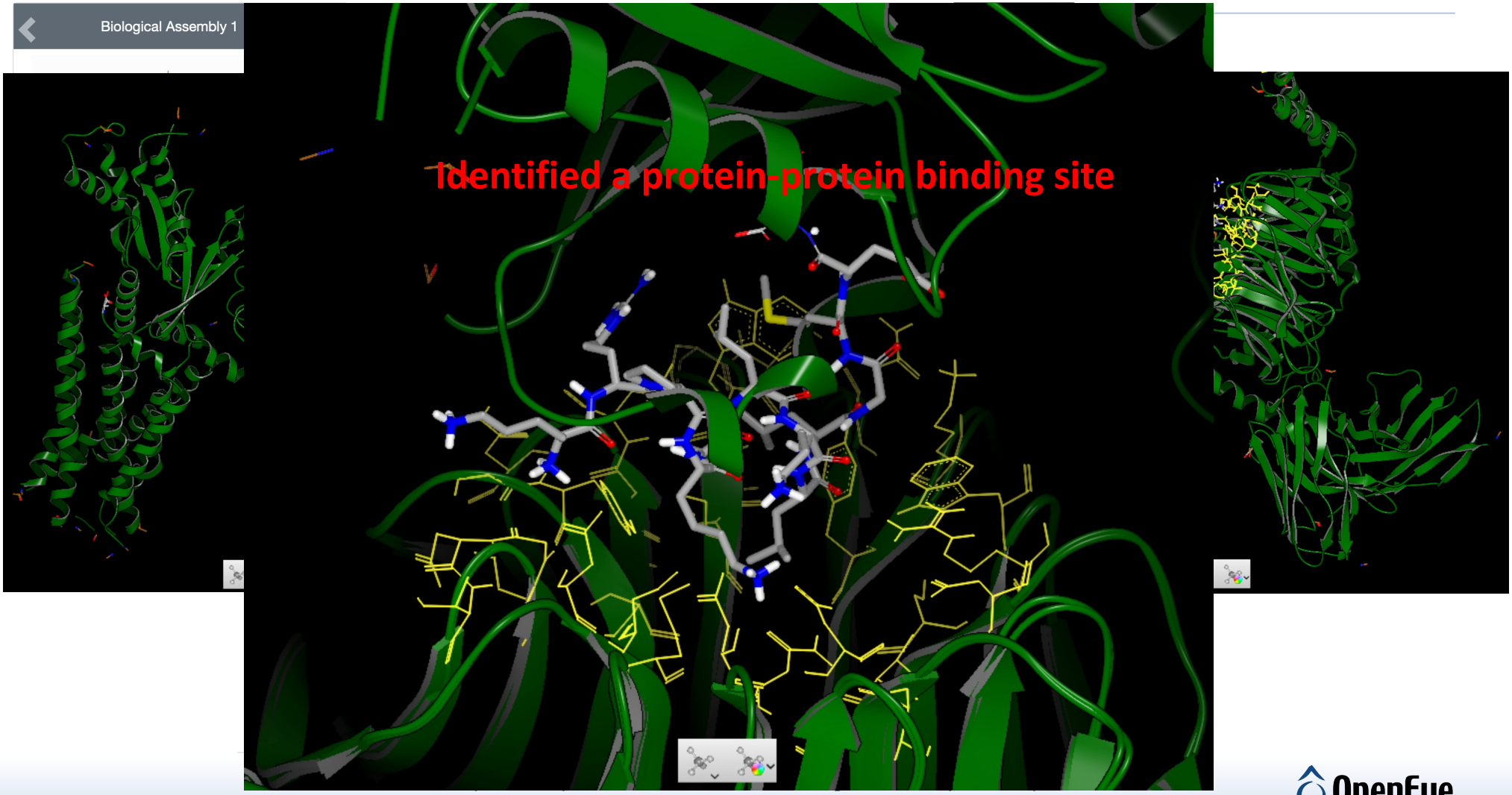
In the case of Ubiquitin, variations in UniProt submission produce splits that do not actually contribute to uniqueness, and generate redundancies that complicate its utility.

OpenEye SCIENTIFIC

Problems in first iteration…

- Targets not in GtoP
  - Viral targets*
  - Bacterial targets
  - e.g. Haemoglobin

- Problems mapping using UniprotKB IDs
  - Viral targets*
  - **UniprotKB IDs map to multiple structures, e.g. macromolecular assemblies**



OpenEye SCIENTIFIC

Problems in first iteration…

- Targets not in GtoP
  - Viral targets*
  - Bacterial targets
  - e.g. Haemoglobin

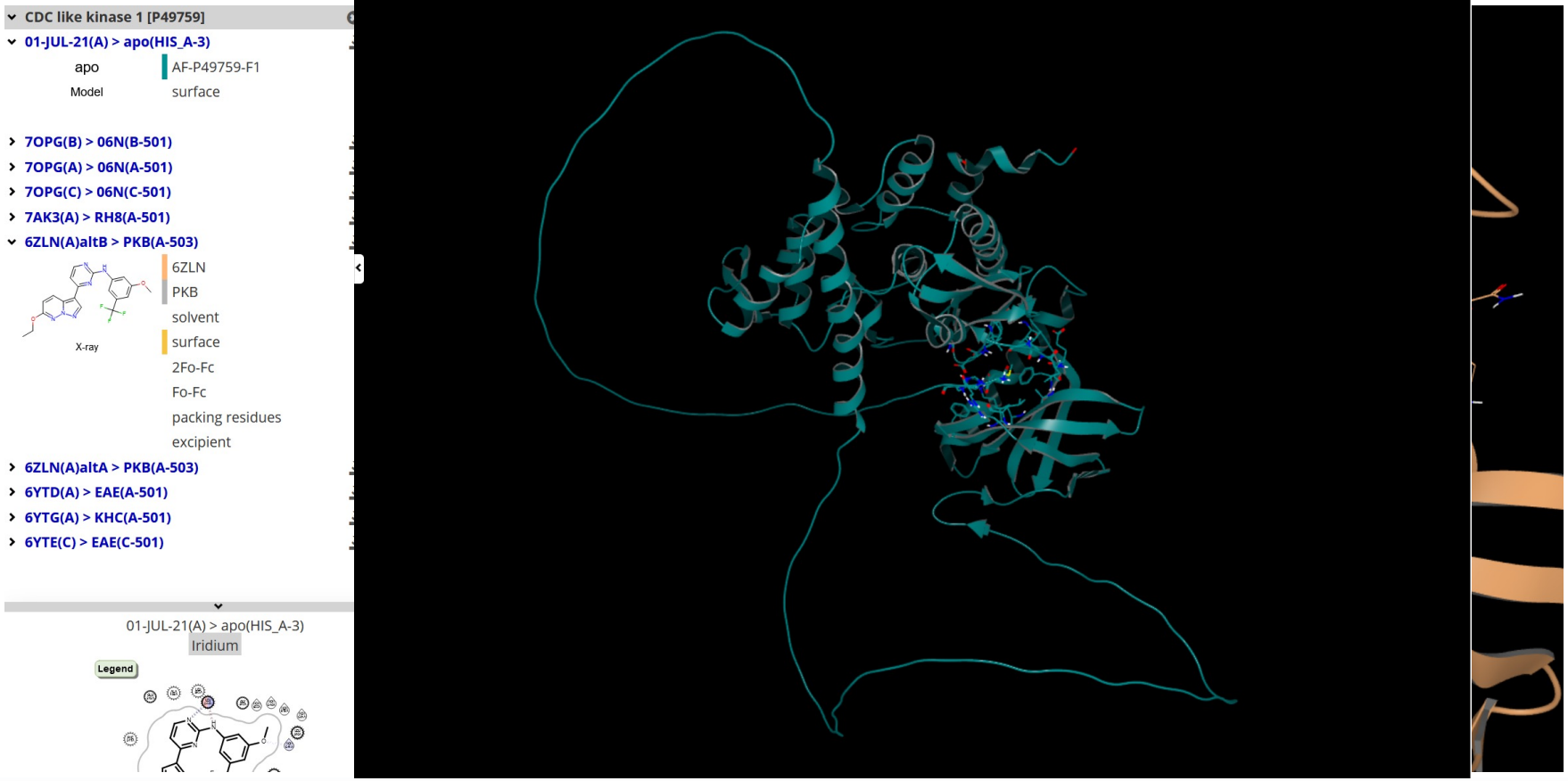
- Problems mapping using UniprotKB IDs
  - Viral targets*
  - UniprotKB IDs map to multiple structures, e.g. macromolecular assemblies

- Target reference issues
  - **No liganded structures in a target** – which “site” is of interest? Big problem for AlphaFold models

Current Status

- Prepare the entire PDB
  - 105,779 experiments were prepared
    - ~200,000+ design units
- Prepare AlphaFold2
  - 1,325 models were prepared
    - 1,325 design units
- Organize the data
  - Organized in MMDS
    - Guide to pharmacology tree
    - Uncategorized tree
    - 12,728 (**4x improvement**)
- Mine the data
  - Let’s turn to this next…

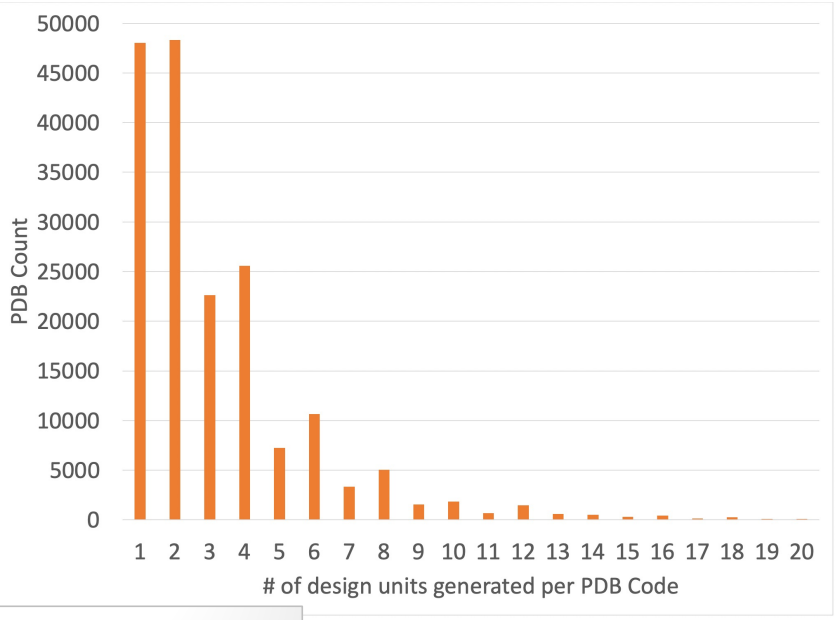
AlphaFold

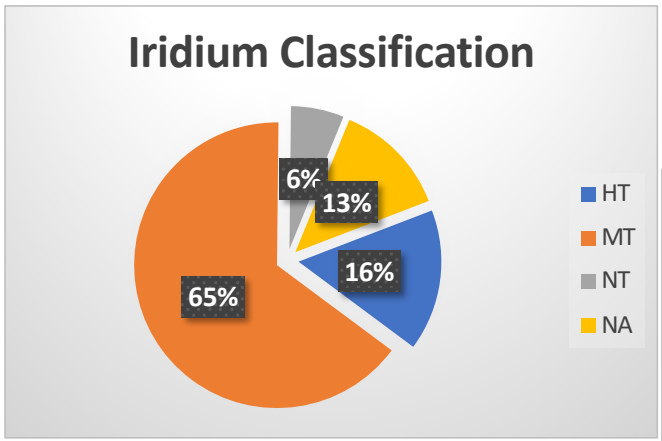


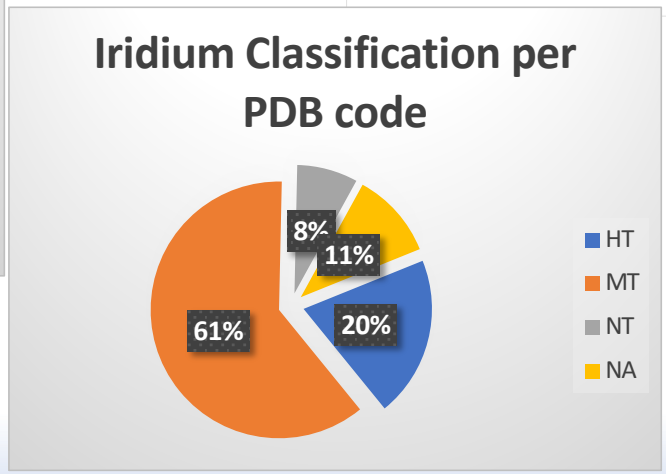
OpenEye SCIENTIFIC

Key bits of early data

- Success rate: 97%
- Stats from 239,298 liganded DUs

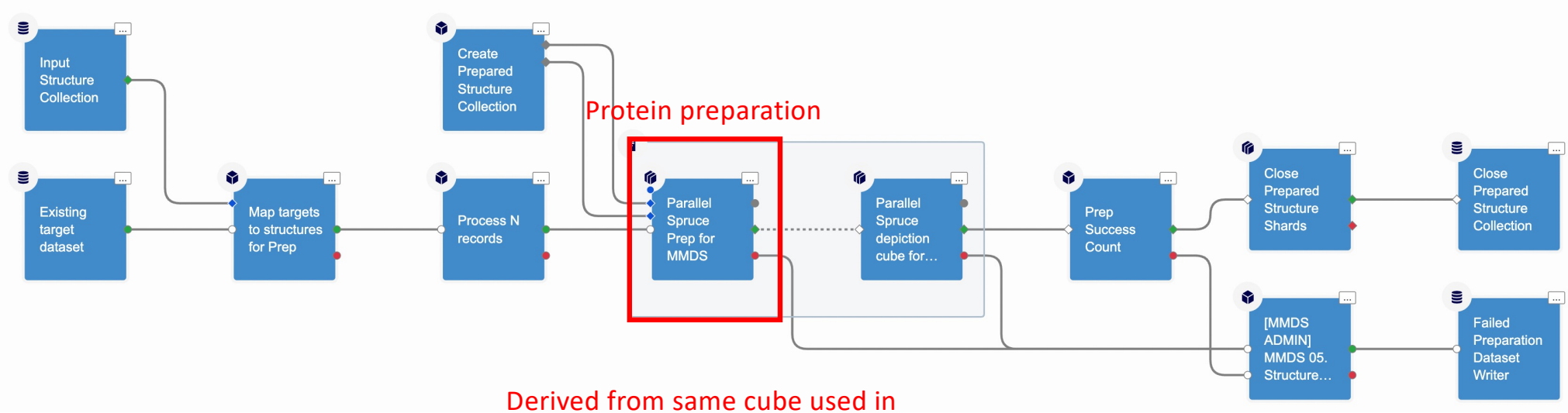






OpenEye SCIENTIFIC

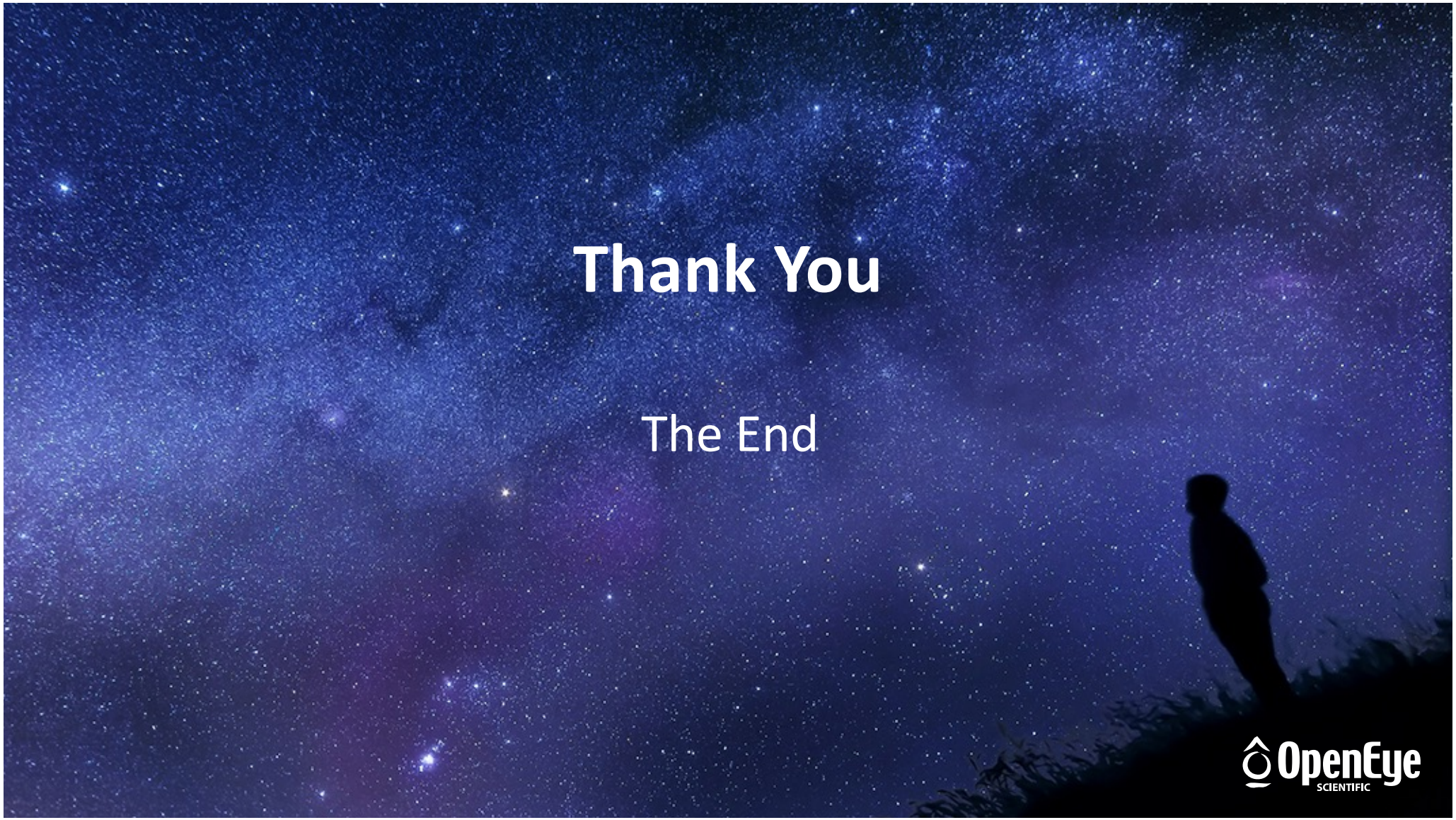
Structure Prep at Scale



Derived from same cube used in
Classic Floes Protein Preparation Floes

Summary

- Closing in on goal of preparing entire PDB + AlphaFold using Orion
- Organizing the prepared data in MMDS
- Solved some interesting problems with linking databases of experiments to a definition of “target”
- Data mining and learning from this data has just begun


Thank You
The End
OpenEye
SCIENTIFIC

For more information, please contact:

sales@eyesopen.com | **info@eyesopen.com**

www.openeye.inc

+1-505-473-7385