

D3R Synergies with the Protein Data Bank

March 11th 2016

Stephen K. Burley, M.D., D.Phil.
Distinguished Professor, Chemistry and Chemical Biology
Director, RCSB Protein Data Bank
Director, Center for Integrative Proteomics Research
Member, Cancer Institute of New Jersey
Founding Director, Institute for Quantitative Biomedicine



Outline

- Introducing the Protein Data Bank
- Improving the Quality of Co-Crystal Structures in the PDB
- Enabling Industry PDB Depositions with D3R
- Electron Density Scoring of Predicted Ligand Poses

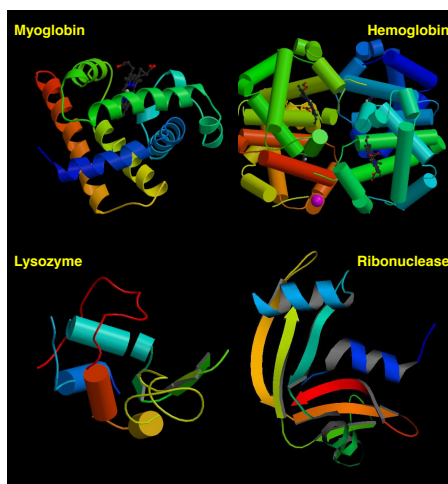


Outline

- Introducing the Protein Data Bank
- Improving the Quality of Co-Crystal Structures in the PDB
- Enabling Industry PDB Depositions with D3R
- Electron Density Scoring of Predicted Ligand Poses

Protein Data Bank

- PDB: 1st Open Access digital resource in biology (est. in 1971 with 7 entries)
- Primary Data Archive
- Managed jointly by wwPDB partners in US (RCSB PDB), UK (PDBe), and Japan (PDBJ)
- Today, single global PDB macromolecular structure archive (>116,000 entries)



Some of the very first structures in the PDB

Protein Xtallography NOT High-Accuracy!

- Precision of non-H Atomic (x,y,z) $\sim 0.1 \times$ Resolution Limit

3Å Resolution: $\sim 0.3-0.4\text{\AA}$

2Å Resolution: $\sim 0.2\text{\AA}$

Non-bonded Interaction Distances: $\sim \pm 0.3\text{\AA}$

- Ligand Quality depends critically on

Accuracy of Chemical Structure Description
 Correct Ideal Geometry for X-ray Refinement
 Correct Interpretation of Electron Density
 Chemical Knowledge of Crystallographer

- Refereeing of Co-crystal Structure Papers “Uneven”

Protein Data Bank is an Archival Resource!

- Global wwPDB Deposition/Annotation/Validation

- Annotated by professional Biocurators employed at RCSB PDB (Americas/Oceania), PDBj (Asia), and PDBe (Europe/Africa)

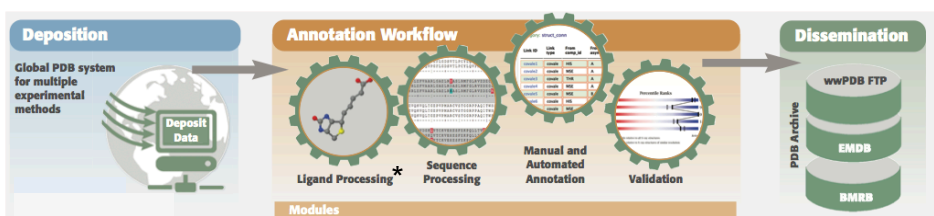
- Validated using tools developed with help from wwPDB Validation Task Forces (X-ray, NMR, EM)

- PDB does NOT reject archival depositions (Depositor is Ultimately Responsible for Quality!)

Outline

- Introducing the Protein Data Bank
- Improving the Quality of Co-Crystal Structures in the PDB
- Enabling Industry PDB Depositions with D3R
- Electron Density Scoring of Predicted Ligand Poses

wwPDB Deposition/Annotation/Validation



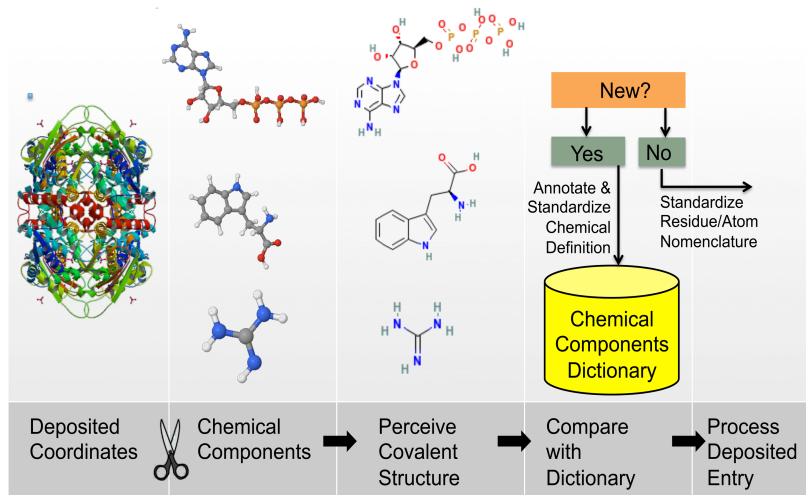
*Full description in poster

- Regional Workload Balancing and Increased Productivity
- Better QA/QC of Polymer Sequences and Ligand Chemistry
- PDBx/mmCIF is now the Master File Format
- Validation based on Recommendations from wwPDB Task Forces
- **Version 2.0 released Jan. 2016 (X-ray/NMR/EM)!**
- Federation with other Data Resources (e.g., EMDB, SASBDB, ...)

Chemical Component Dictionary (CCD)

- Complete descriptions of constituent small molecules in experimentally-determined 3D macromolecular structures in the PDB
- Data items include
 - Atom Nomenclature
 - Connectivity/Chirality
 - Chemical Formula, InChI/SMILES, etc.
 - Molecular Names
 - Idealized 3D Structure
 - 3D Structure Exemplar from PDB Archive

Chemical Component Deposition



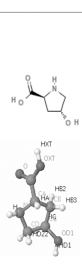
Chemical Component Dictionary Entry

a)

```

data_HYP
#
_chem_comp_id          HYP
_chem_comp_name        L-HYDROXYPROLINE
_chem_comp_type        STAMP
_chem_comp_type_symbol "CS H9 N O3"
_chem_comp_formula     C5H9NO3
_chem_comp_mon_nstd_parent_comp_id
_chem_comp_pdbx_synonym HYDROXYPROLINE
_chem_comp_pdbx_formal_charge 0
_chem_comp_pdbx_initial_date 1999-07-08
_chem_comp_pdbx_modified_date 2008-04-29
_chem_comp_pdbx_release_status REL
_chem_comp_pdbx_replaced_by 7
_chem_comp_pdbx_replaces 7
_chem_comp_formula_weight 131.130
_chem_comp_chem_letter_code HYP
_chem_comp_three_letter_code HYP
_chem_comp_pdbx_model_coordinates_details
_chem_comp_pdbx_model_coordinates_missing_flag 0
_chem_comp_pdbx_ideal_coordinates_details
_chem_comp_pdbx_ideal_coordinates_missing_flag 0
_chem_comp_pdbx_model_coordinates_db_code 1086
_chem_comp_pdbx_processing_site RCSB
#

```



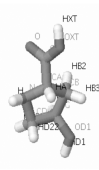
b)

```

loop
_chem_comp_atom_comp_id
_chem_comp_atom_atom_id
_chem_comp_atom_atm_name
_chem_comp_atom_type_symbol
_chem_comp_atom_charge
_chem_comp_atom_pdbx_align
_chem_comp_atom_pdbx_aromatic_flag
_chem_comp_atom_pdbx_leaving_atom_flag
_chem_comp_atom_model_Cartn_x
_chem_comp_atom_model_Cartn_y
_chem_comp_atom_model_Cartn_z
_chem_comp_atom_pdbx_model_Cartn_x_ideal
_chem_comp_atom_pdbx_model_Cartn_y_ideal
_chem_comp_atom_pdbx_model_Cartn_z_ideal
_chem_comp_atom_pdbx_orbital
HYP H1  H  0  1  W  W  -3.266 16.348 44.188  0.168  1.360 -0.282 1
HYP H2  H  0  1  W  W  -2.950 15.768 42.044 -0.284 -0.003 -0.493 2
HYP C1  C  0  1  W  W  -1.447 15.609 42.030 -1.411 -0.072 -0.012 3
HYP C2  C  0  1  W  W  -0.722 15.484 42.503 -2.233 -0.764  0.759 4
HYP CB  C  0  1  W  W  -3.488 16.578 41.829  0.515 -0.924  0.359 5
HYP CD  C  0  1  W  W  -4.427 17.482 42.330  1.647 -0.159  0.305 6
HYP OD  O  0  1  W  W  -6.493 16.810 42.294  2.817 -0.911 -0.071 8
HYP OZ  O  0  1  W  W  -0.978 16.502 42.469 -2.614 -1.062 -0.812 9
HYP H3  H  0  1  W  W  -3.980 16.047 44.765 -0.107  1.961 -1.628 10
HYP H4  H  0  1  W  W  -3.285 16.750 42.068 -0.323 -0.278 -1.546 11
HYP H22  H  0  1  W  W  -2.587 17.141 41.398  0.046 -1.092  1.337 12
HYP H23  H  0  1  W  W  -2.780 16.926 41.026  0.978 -0.375 -0.153 13
HYP H2  H  0  1  W  W  -4.588 16.399 41.726  2.092  0.648  1.555 14
HYP H222  H  0  1  W  W  -6.498 16.005 44.370  2.818 -1.085 -1.289 15
HYP H232  H  0  1  W  W  -4.497 16.712 43.848  2.132  1.965  0.242 16
HYP H21  H  0  1  W  W  -5.990 16.464 43.181  3.780 -0.479 -0.309 17
HYP H21  H  0  1  W  W  -6.027 16.531 42.699 -1.749 -1.066 -0.898 18
#

```

Atom names
Stereochemistry & aromaticity
Model coordinates
Ideal coordinates



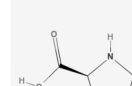
c)

```

loop
_chem_comp_bond_comp_id
_chem_comp_bond_atom_id_1
_chem_comp_bond_atom_id_2
_chem_comp_bond_type_symbol
_chem_comp_bond_pdbx_aromatic_flag
_chem_comp_bond_pdbx_stereo_config
_chem_comp_bond_pdbx_orbital
HYP H1  CA  SING  W  1
HYP H1  H  SING  W  2
HYP H1  H  SING  W  3
HYP CA  C  SING  W  4
HYP CA  CB  SING  W  5
HYP CA  HA  SING  W  6
HYP C  O  DOUB  W  7
HYP C  OZ  SING  W  8
HYP CB  C2  SING  W  9
HYP CB  H22  SING  W  10
HYP CB  H23  SING  W  11
HYP CB  OZ  SING  W  11
#

```

Connected atoms
Bond type
Stereochemistry & aromaticity



d)

```

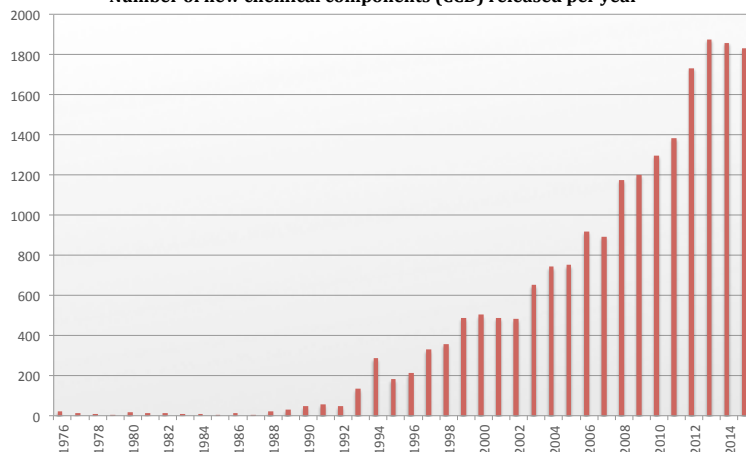
#
_pdbx_chem_comp_descriptor_comp_id
_pdbx_chem_comp_descriptor_type
_pdbx_chem_comp_descriptor_program
_pdbx_chem_comp_descriptor_program_version
_pdbx_chem_comp_descriptor
HYP SMILES          ACLabels  12.01 O=C(O)CNC(C)O(C)
HYP SMILES_CARCICAL  CACTV5     3.385 O(C)[CH](C)(C)C(O)=O
HYP SMILES           CACTV5     3.385 O(C)[CH](C)(C)C(O)=O
HYP SMILES_CARCICAL  "openType @Toolkits" 1.7.5 C1(C)[CH](C)(C)C(O)=O
HYP SMILES           "openType @Toolkits" 1.7.5 C1(C)[CH](C)(C)C(O)=O
HYP InChI            InChI      1.03  "InChI=1S/C5H9NO3/c7-3-1-(4)(8)(9)6-2-3/3-4,6-7H,1,..."
HYP InChIKey         InChI      1.03  PNHRYVYMAAGQ-OTRCVYIQLA-N
#
_pdbx_chem_comp_identifier_comp_id
_pdbx_chem_comp_identifier_type
_pdbx_chem_comp_identifier_program

```

SMILES
InChI

Growth of Chemical Components in PDB

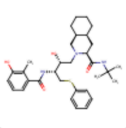
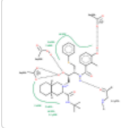
Number of new chemical components (CCD) released per year



Growth trends new Chemical Components in the PDB

Linking to Binding Data Resources

Ligands 1 Unique

ID	Chains	Name / Formula / InChI Key	2D Diagram & Interactions
1UN Query on 1UN Download SDF File Download CCD File	A	2-[2-HYDROXY-3-(3-HYDROXY-2-METHYL-BENZOYLAMINO)- 4-PHENYL SULFANYL-BUTYL]-DECAHYDRO-ISOQUINOLINE- 3-CARBOXYLIC ACID TERT-BUTYLAMIDE NELFINAVIR MESYLATE AG1343 (<i>Synonym</i>) C ₃₂ H ₄₅ N ₃ O ₄ S QAGYKUNXZHKKMR-HKWSIXNMSA-N	 

External Ligand Annotations

ID	Binding Affinity (Sequence Identity %)
1UN	Ki: 0.53 - 2 nM (88 - 98) BindingDB Ki: 2 nM BindingMOAD Ki: 2 nM PDBbind

PCSB PKOT 12

Tabular Report: HSP90 with Binding Data

Custom Report

Click on column headers to sort up/down. Click again to reverse order. Download options: [EXCEL](#) | [EXCEL 2007 or later](#) | [CSV](#)

Page 1 of 4

	PDB ID	Chain I	Ligand SMILES	HET ID	Ki (nM)	Kd (nM)
	<input type="checkbox"/> x	<input type="checkbox"/> x	<input type="checkbox"/> x	<input type="checkbox"/> x	<input type="checkbox"/> x	<input type="checkbox"/> x
4	2QF6	D	<chem>c1ccc2cc(c(cc2c1)c3nc(nc(n3)N)N)Br</chem>	A56	<ul style="list-style-type: none"> 320 (PDBbind) 320 (BDB) 	<ul style="list-style-type: none"> <1.0e+4 (BDB)
5	2QFO	A	<chem>c1ccc(cc1)NC=C1/CCOC2=O</chem>	A51		<ul style="list-style-type: none"> 150000 (BMOAD) 150000 (BDB)
6	2QFO	B				
7	2QFO	A	<chem>Cc1cc(nc(n1)N)C(F)(F)F</chem>	A13	<ul style="list-style-type: none"> 18000 (BMOAD) 18000 (BDB) 	<ul style="list-style-type: none"> 20000 (PDBbind) 20000 (BDB)
8	2QG0	A	<chem>Cc1cc(nc(n1)N)CNS(=O)(=O)c2ccccc2)N/C=C</chem>	A94	<ul style="list-style-type: none"> 1900 (PDBbind) 1900 (BMOAD) 	
9	2QG0	B	<chem>Cc1cc(nc(n1)N)CNS(=O)(=O)c2ccccc2)N/C=C</chem>	A94	<ul style="list-style-type: none"> 1900 (PDBbind) 1900 (BMOAD) 	

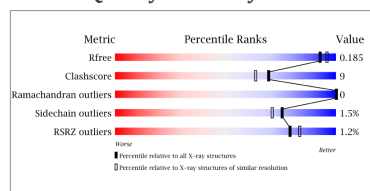
Improved Validation

- wwPDB Validation Task Forces for X-ray, NMR, SAS
- wwPDB/EMDataBank VTF for EM
- Recommendations about validating new and existing structures
 - Implemented in software pipeline
 - Produces summary report (PDF) and XML file with detailed statistics
- Validation at different stages
 - While determining/depositing the structure
 - After annotation (official; should be sent to journals)
 - Upon release (publicly available; updated annually)

X-ray Validation Reporting

- Model Quality
 - Bond lengths and angles (outlier info, RMS-Z)
 - Chirality, planarity
 - Close contacts (including worst clashes, MolProbity clash score)
 - Torsion angles (Ramachandran statistics, protein rotamers)
 - Ligand geometry (Mogul analysis)
- Residue Plots
 - Residues with model-quality outliers (0, 1, 2, >2)
 - Residues with RSR-Z > 2 are highlighted
 - Residues not observed

Overall Quality Summary

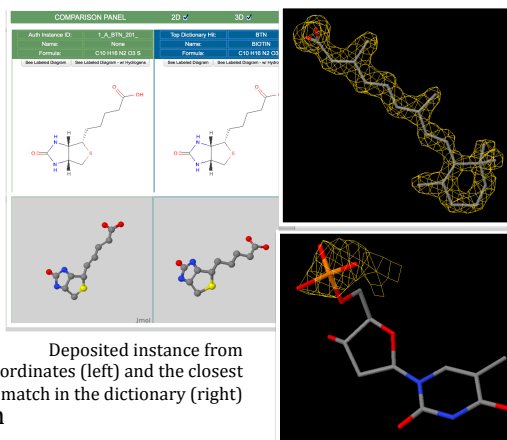


Residue Plots



Improved Ligand Annotation at Deposition

- Batch search against Chemical Component Dictionary with automated CCD ID assignment
- Captures and displays author-provided chemical information
- Comparison panel
 - 2D and 3D views of ligand for Biocurator Review
 - ID assignment
- Display of local ligand electron density fit
- **Soon to be provided to Depositors**



Local ligand density display (1.5 sigma omit map)
 Top: REA in entry 1CBS with LLDF=1.31 (RSR=0.10, CC=0.95)
 Bottom: TMP in entry 3HW4 with LLDF=6.77 (RSR=0.41, CC=0.70)

wwPDB/CCDC/D3R Ligand Validation Workshop

Objectives: Bring together co-crystal structure determination experts from Academe and Industry with Crystallography and Computational Chemistry Software Developers to discuss, develop, and recommend Best Practices for

- PDB Deposition/Validation of Co-crystal Structures
- Editorial/Refereeing/Publication of Co-crystal Structures
- Improvement of Ligand Representations across PDB



LVW Meeting Metrics

- Demographics
 - 41 Organizations represented by
 - 57 Registered Participants
- Participant breakdown
 - Large Pharmaceutical Companies: 10
 - Small Biotechnology Companies: 3
 - Universities: 12
 - National Laboratories: 8
 - Technology/Computational Chemistry Companies: 8
 - wwPDB: 13; CCDC: 2; D3R: 1

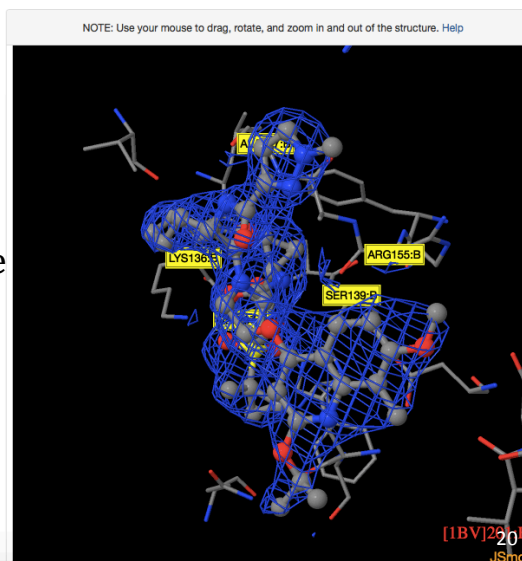
Workshop White Paper Process

- White Paper describing recommendations re deposition/validation and editorial/refereeing/publication standards now in press at *Structure*
- Recommendations were approved *en bloc* by the wwPDB X-ray Validation Task Force (Nov. 2015)
- Recommendations to be implemented in 2016
- Journals have been asked to make the wwPDB Validation Report part of their Refereeing Process

Ligand Electron Density in 3D @ rcsb.org

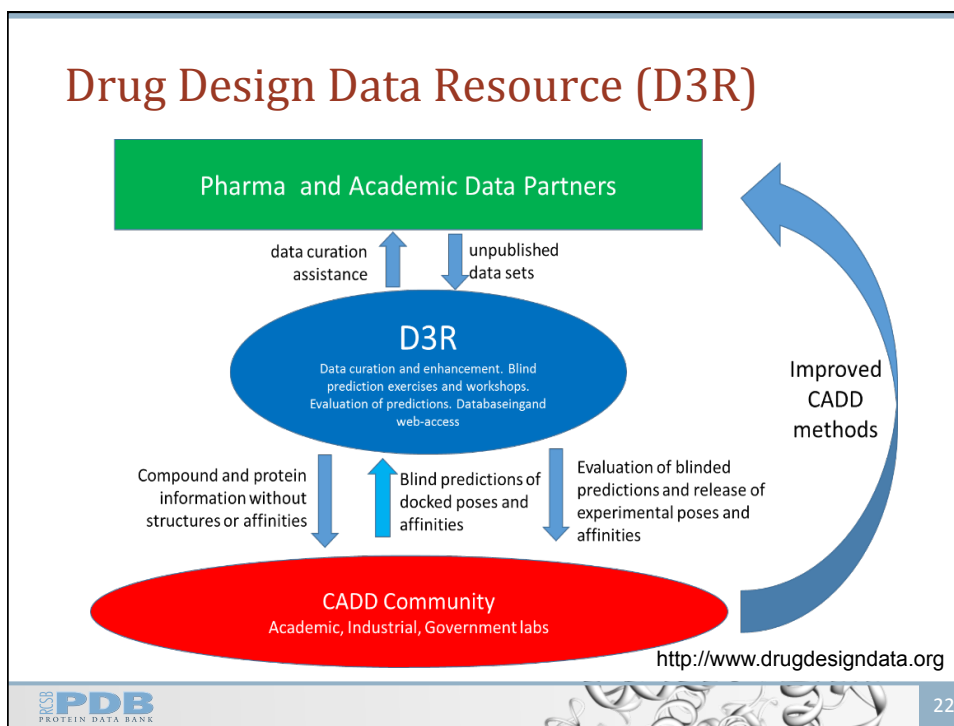
Crystal structure of HCV NS3/4A D168V protease complexed with compound 4

- rcsb.org now offers pre-computed electron density views of all ligands (with exptl. |Fobs|)
- HCV NS3/4A Protease
PDB: 4I32; CCD: 1BV
RSCC=0.9
RSR=0.11
RSZD=0.1
 $R_{\text{free}} - R_{\text{work}} = 5.1\%$



Outline

- Introducing the Protein Data Bank
- Improving the Quality of Co-Crystal Structures in the PDB
- Enabling Industry PDB Depositions with D3R
- Electron Density Scoring of Predicted Ligand Poses



wwPDB Data Release Policy Change

- Effective April 10th 2015
Two-Stage Weekly PDB Archive Data Release
- Every Saturday 03:00 UTC Release of ~200 Entries
 - Polymer Sequence(s)
 - InChI String(s)
 - Crystallization pH value(s)
 - **Need to add RSR, RSCC, RSZD, ($R_{\text{free}}-R_{\text{work}}$), etc.**
- Following Wednesday 0:00 UTC All Data Release
- Weekly Blinded CASP-like Competitions
 - Protein Structure Modeling (CAMEO)
 - Ligand Docking (CELPP)

Automated Batch Deposition to PDB

Boundary Condition:

- Input coordinate files must meet deposition requirements established for the wwPDB Global D&A system

Plan:

- Provide a template mmCIF file and specifications that fulfills D&A deposition requirements for D3R partners to prepare their deposition files
- Continue improving pdb_extract to produce a deposition file that meets the specifications in the template file
- Create new RCSB PDB website to host upload/download operations for automated batch deposition
- Incorporate batch deposition and annotation functionality into wwPDB D&A system



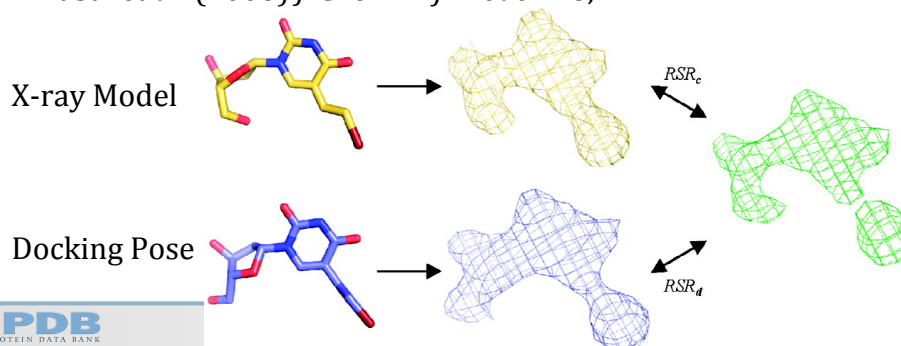
Outline

- Introducing the Protein Data Bank
- Improving the Quality of Co-Crystal Structures in the PDB
- Enabling Industry PDB Depositions with D3R
- Electron Density Scoring of Predicted Ligand Poses

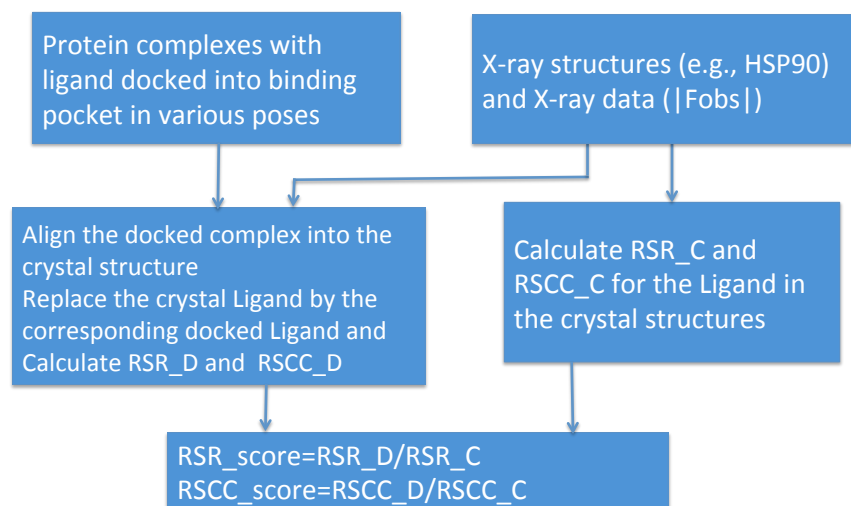


Electron Density (ED) Scoring of Poses

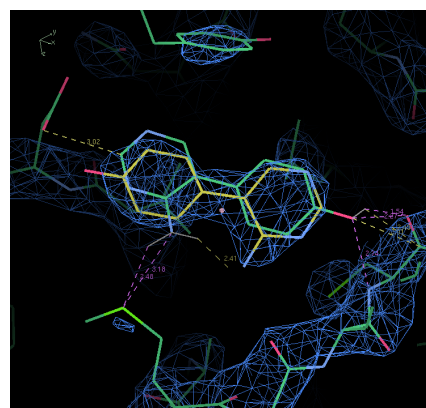
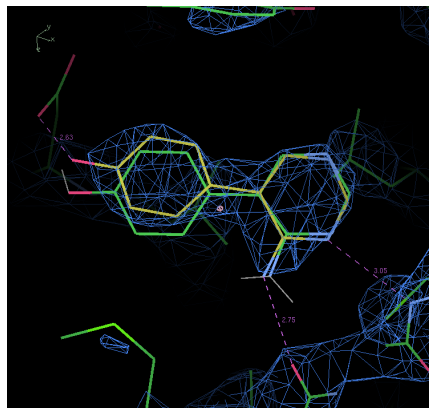
- Predicted ED for Ligand X-ray Model (yellow)
- Predicted ED for Ligand Docking Pose (blue)
- Measured ED Ligand X-ray Experiment (green)
- Compute Real Space R-factors (RSR_c versus RSR_D)
- Yusuf *et al.* (2008) *J. Chem. Inf. Model.* 48, 1411-1422



D3R RSR/RSCC Scoring Pipeline



Multiple Metrics Required!



5670edaea4508 21 1 (4OBQ) RMSD=0.93 RSR=1.51 RSCC=0.88 (left)
 5670edaea4508 21 3 (4ZK5) RMSD=5.39 RSR=1.32 RSCC=0.88 (right)

The yellow is the crystal structure, green is the docked.
 Both poses show good RSCC, but pose 1 give good RMSD

Acknowledgements (H. Yang, J. Young)

