RCSB Protein Data Bank
Advisory Committee Meeting

October 31, 2011

RCSB PDB
PROTEIN DATA BANK

Overview

Helen Berman

Response to Major 2010 Recommendations

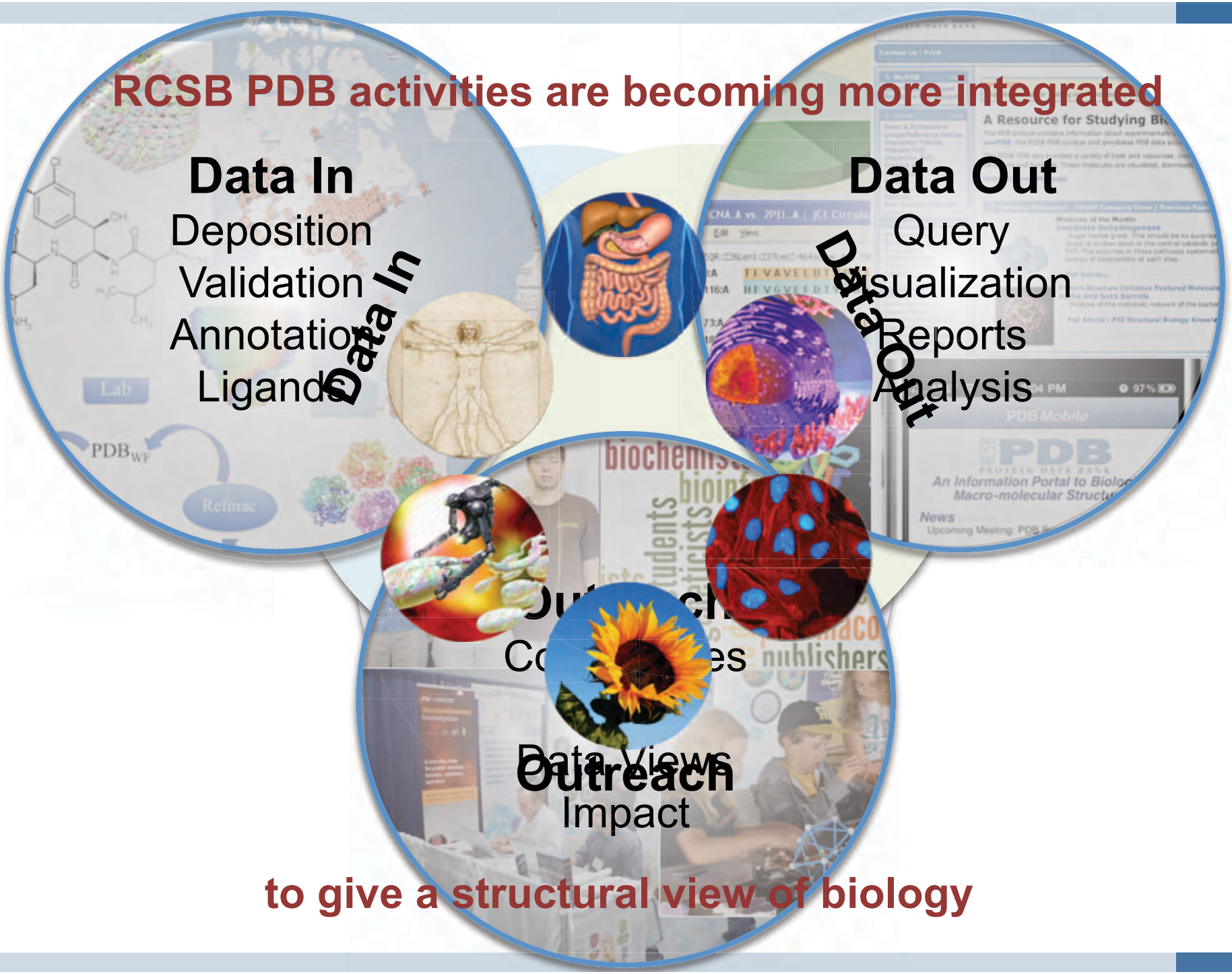
- Common Tool: Quantitative estimation of improvements in data processing
  - Processing time for ligands cut by up to 70% with new interface; benchmarking to continue
- Data Out: What are the benefits & costs of smart phone/iPad development?
  - Survey of PDB*Mobile* users
  - Improving website appearance for mobile devices
- Outreach: Add K12 Education Representative on AC
  - Jack G. Chirikjian, Ph.D., Georgetown University

Vision

To provide a global resource for the advancement of research and education in biology and medicine by curating, integrating, and disseminating biological macromolecular structural information in the context of function, biological processes, evolution, pathways and disease states.

We will implement standards, and anticipate and develop appropriate technologies to support evolving science.

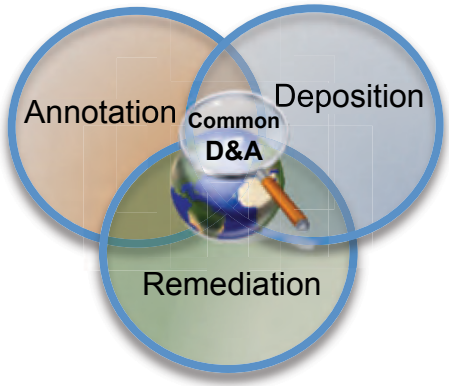
RCSB PDB activities are becoming more integrated



to give a structural view of biology

Data In

- Improved tools for deposition
- Improved data processing efficiency
- 2011 remediation release
- Common Tool partially in production
- Resolution of format issues
- wwPDB Validation Task Forces



Depositions

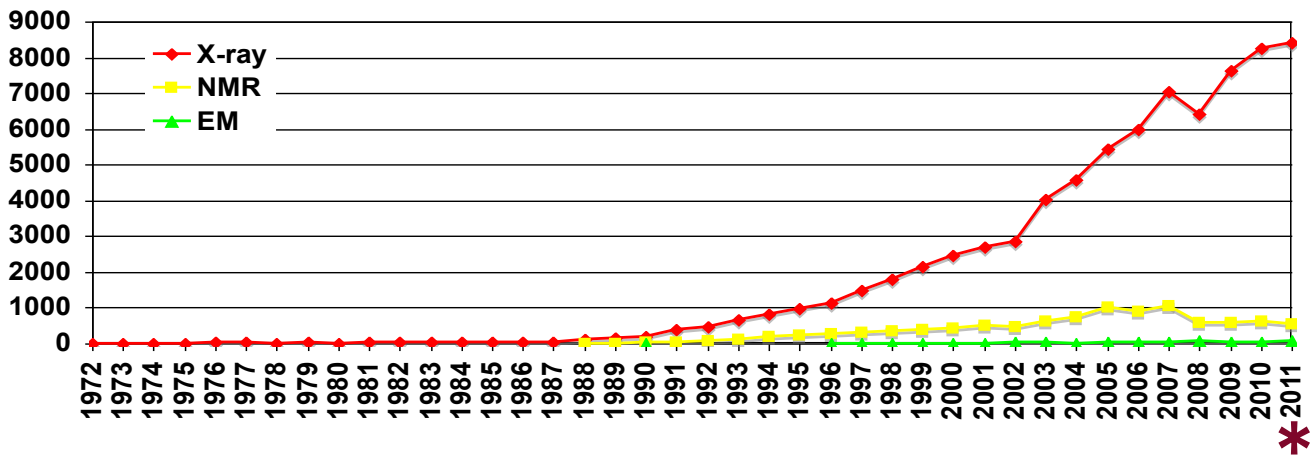
WORLDWIDE PDB
PROTEIN DATA BANK

By deposition and processing site

Last Updated: 5 Oct 2011

Year	Total Depositions	Deposited To			Processed By		
		RCSB PDB	PDBj	PDBe	RCSB PDB	PDBj	PDBe
2000	2983	2445	10	528	2297	158	528
2001	3287	2673	118	496	2408	383	496
2002	3565	2769	289	507	2401	657	507
2003	4830	3488	673	669	3135	1026	669
2004	5508	3796	900	812	3082	1614	812
2005	6678	4507	1166	1005	3563	2110	1005
2006	7282	5145	1052	1085	4252	1945	1085
2007	8130	5399	1603	1128	4703	2299	1128
2008	7073	5452	648	973	4106	1994	973
2009	8300	6715	527	1058	5069	2173	1058
2010	8878	6912	593	1373	5464	2041	1373
2011	6896(*9088)	5370	409	1117	4528	1251	1117
TOTAL	73410	54671	7988	10751	45008	17651	10751

By experimental type



*2011 projected

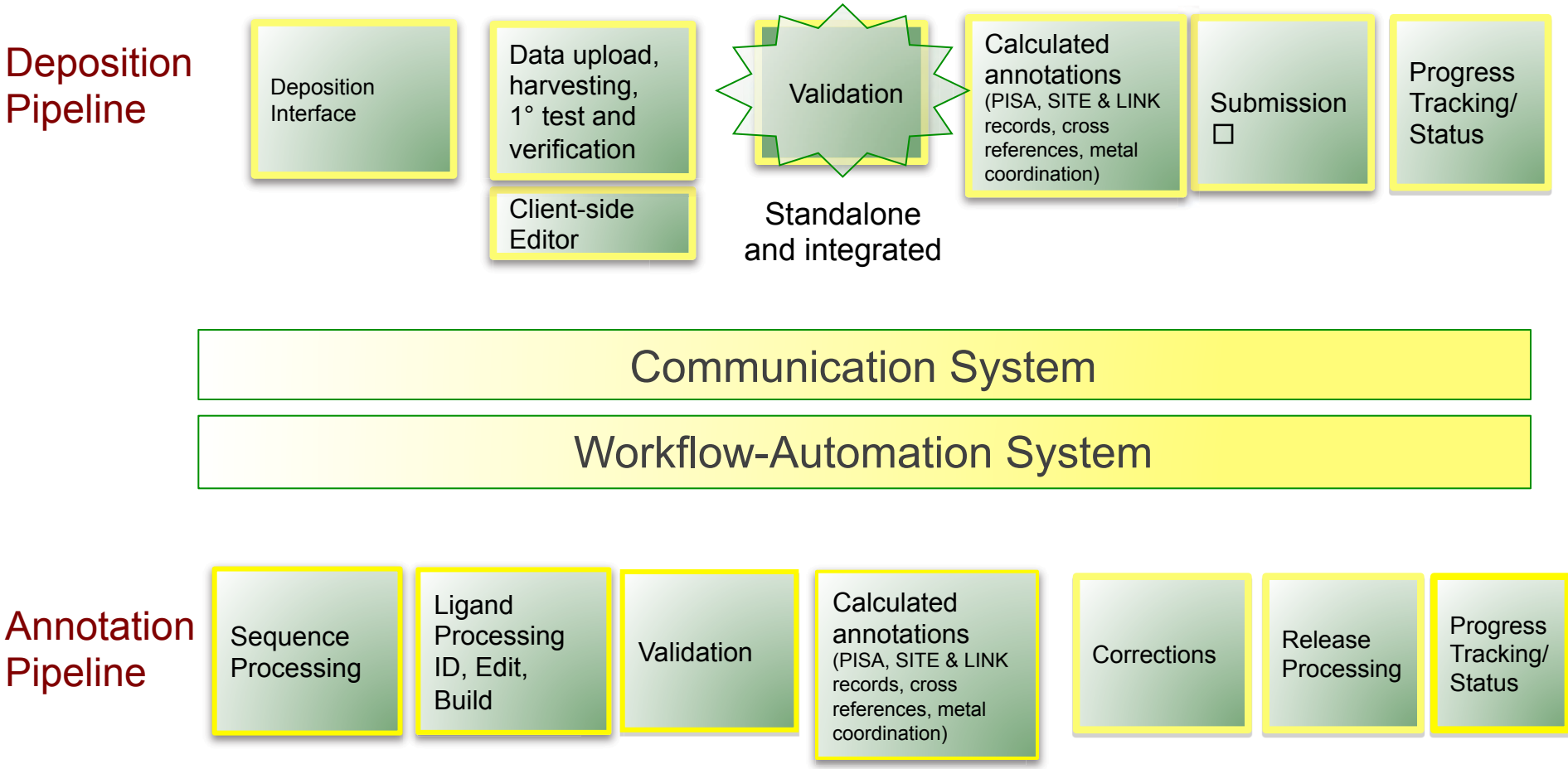
Common Deposition and Annotation Tool

The goal is to implement a set of common deposition and annotation processes and tools that will enable the wwPDB to deliver a resource of increasingly high quality and dependability over the next 10 years.

- addresses the increase in complexity and experimental variety of submissions and the increase in deposition throughput
- maximizes the efficiency and effectiveness of data handling and support for the scientific community

Common Deposition and Annotation Pipeline

WORLDWIDE PDB
PROTEIN DATA BANK



Task Forces

To collect recommendations and develop consensus on method-specific issues, including validation checks that should be performed and identification of validation software applications.

X-ray Validation

- 2008 Workshop
- 2011 *Structure* publication
- Chair: Randy J. Read (University of Cambridge)

NMR Validation

- Meetings held 2009, 2011
- Chairs: Gaetano Montelione (Rutgers), Michael Nilges (Institut Pasteur)
- Report in progress

3DEM Validation

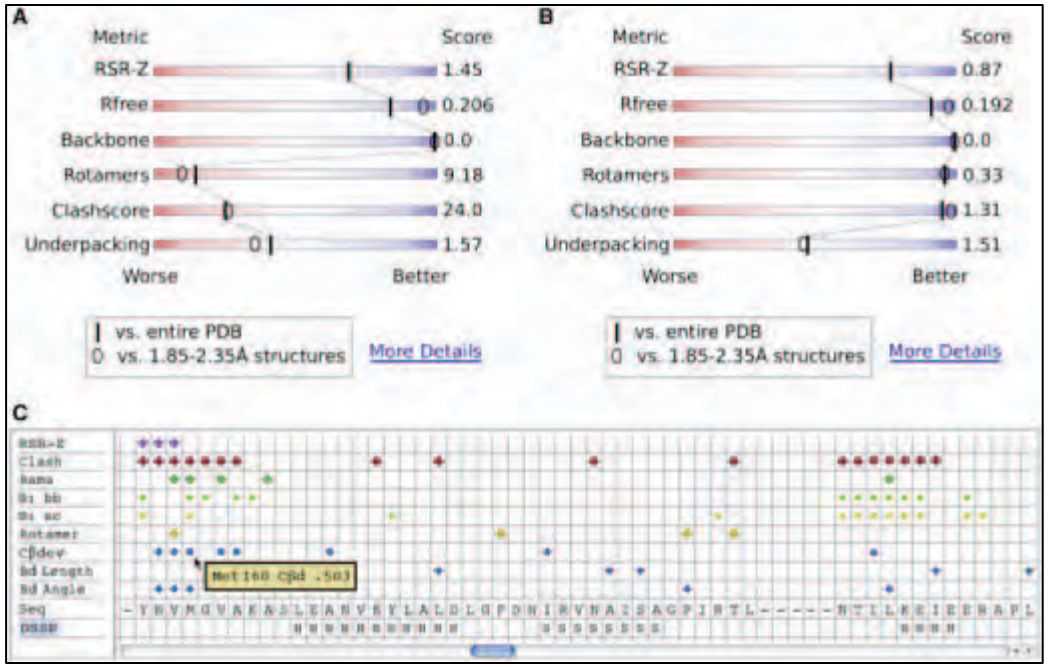
- 2010 Meeting
- Chairs: Richard Henderson (Maps, MRC-LMB), Andrej Sali (Models, UCSF)
- White paper in progress

Small-Angle Scattering

- Members: Jill Trewhella (Univ Sydney), Dmitri Svergun (EMBL Hamburg), Andrej Sali (UCSF), Mamoru Sato (Yokohama City Univ), John Tainer (Scripps)

X-ray VTF recommendations

- Integrated battery of quality checks
  - Pool together validation code from community software into a single pipeline
- Percentile scores
  - Put validation scores in context, helping a non-expert user in judging quality
- Access to validation analyses
  - Cater to all classes of users: depositors, reviewers, expert and non-expert end users, programmers



From *A new generation of crystallographic validation tools for the Protein Data Bank* Read *et al*. (2011) *Structure 19*, 1395-1412.

July 2011 Remediation

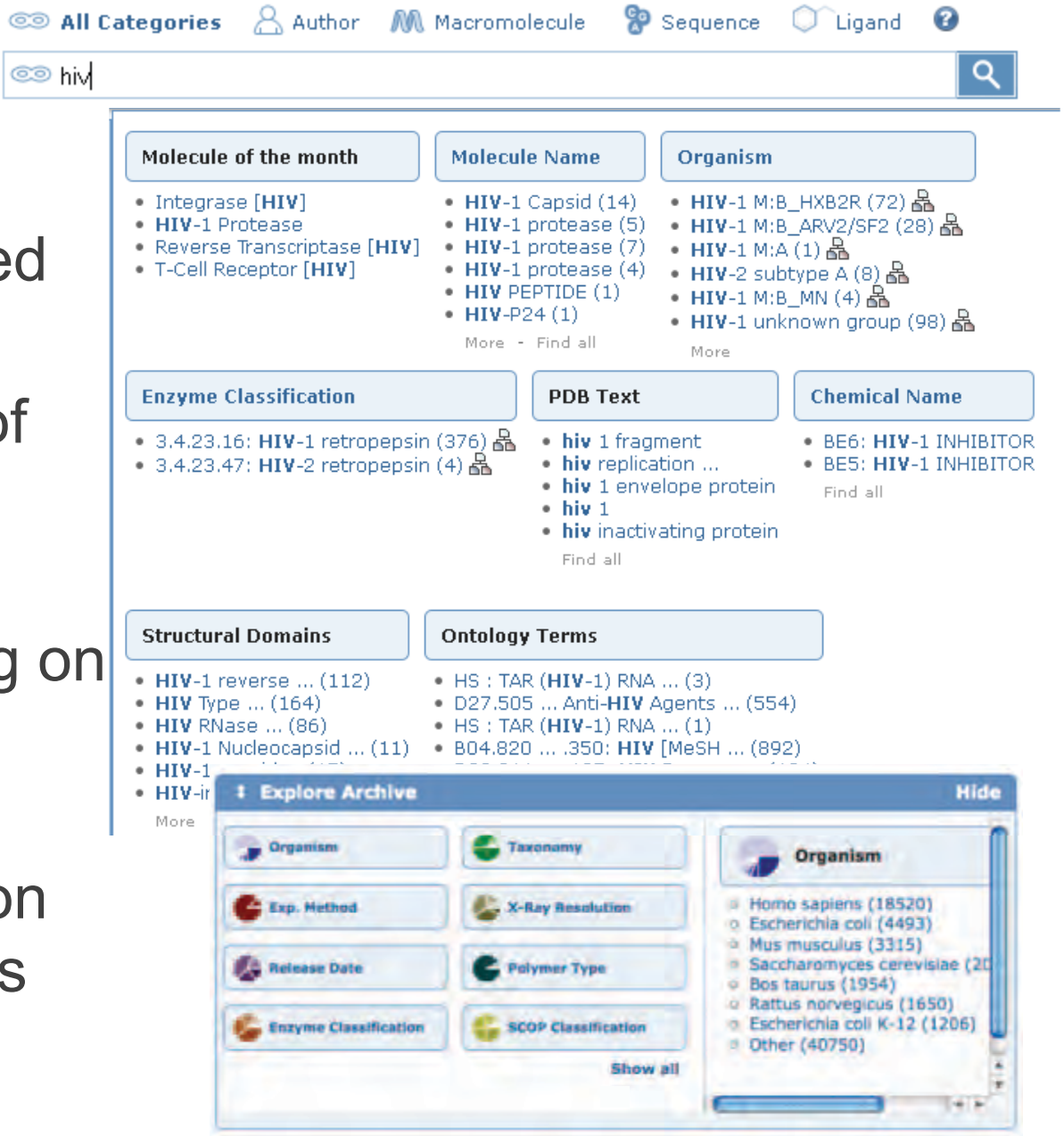
PDB File Format Version 3.3

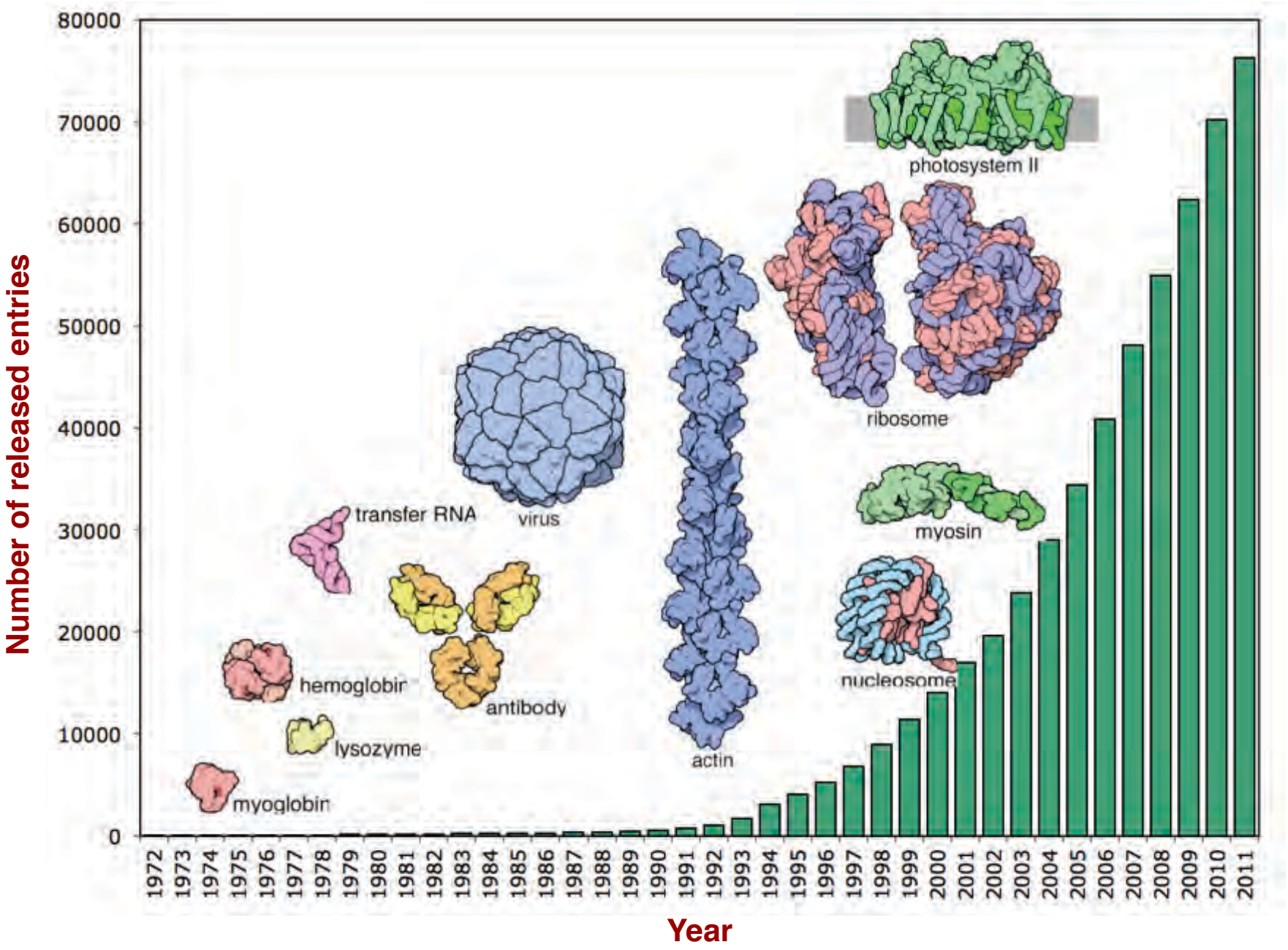
PDB Exchange Dictionary (PDBx) Version 4.0

- Entries containing residual B-factors labeled (7.3K entries)
- Antibiotics and peptide inhibitors standardized (1K entries)
- Entries in the nonstandard crystal frame labeled (148 entries)
- Biological assemblies corrected (5.8K entries)
- Added support for polymers containing nonstandard polymer linkages (58 entries)
- Added support for hybrid x-ray/neutron diffraction experiments (54 entries)
- Added new revision logging to PDBx/PDBML entries (all)

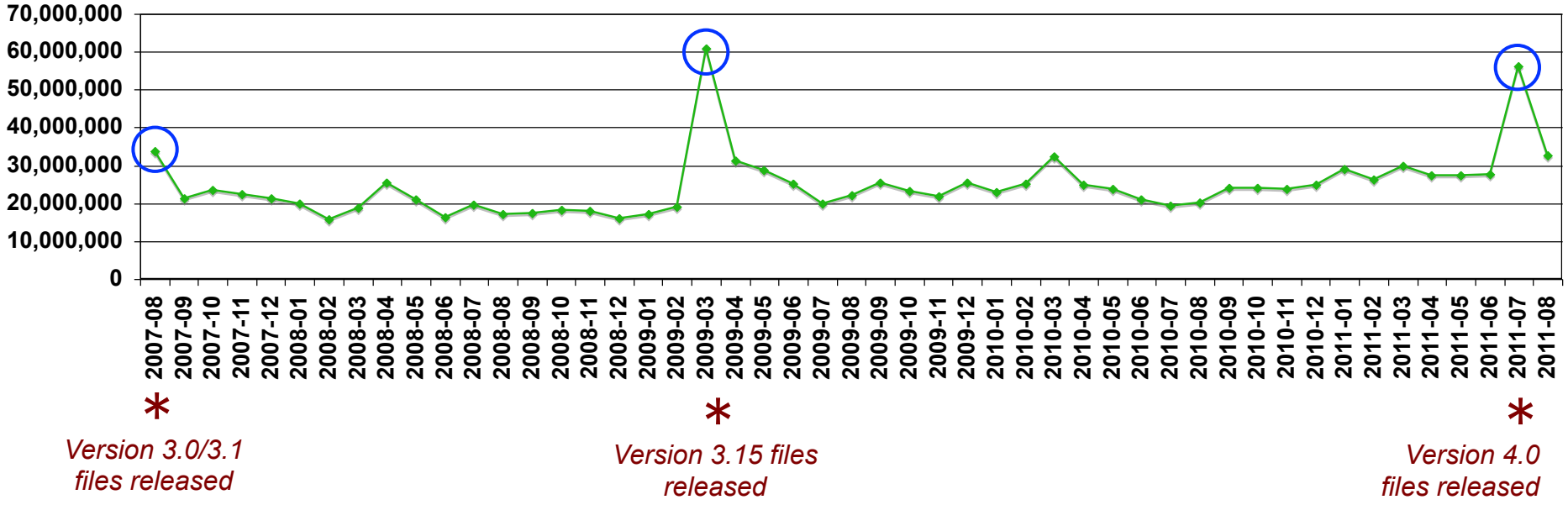
Data Out

- More user-friendly searching and sorted results
- Faceted browsing of PDB archive and search results
- Annotation mapping on to sequence and structure
- Increased integration with other resources




Number of released entries
80000
70000
60000
50000
40000
30000
20000
10000
0
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Year
photosystem II
ribosome
myosin
nucleosome
actin
virus
transfer RNA
antibody
hemoglobin
lysozyme
myoglobin

PDB FTP Downloads



2010 FTP Traffic

RCSB PDB
159 million
entry downloads

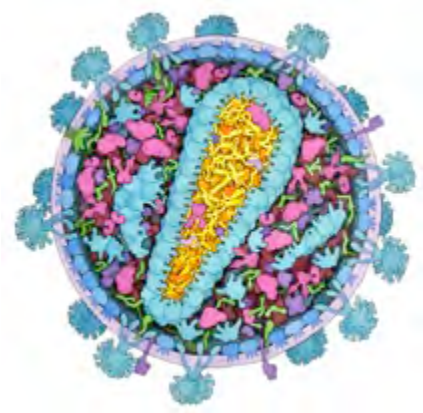
PDBe
34 million
entry downloads

PDBj
16 million
entry downloads

Outreach

Educational Communities

- PDB-101 packages together RCSB PDB resources of interest to teachers and students
- Meetings and events
- Molecular Anatomy Project



Structural Biology of HIV animation and poster

Research Communities

- Task Force Meetings
- Professional society meetings
- Publications
- Online resources
- PDB40

2011 AAAS meeting

Website and PDB*Mobile* Development

- Unified UI design for mobile and desktop display

Before After

Worldwide
Protein Data Bank
Foundation

- Established to support specific wwPDB activities
  - Advisory committee meetings
  - Outreach and education activities, including seminars and workshops
- 501(c)3 organization
  - American, tax-exempt association dedicated to scientific, literary, charitable, and educational purposes
- Fundraising on-going

WORLDWIDE PDB
PROTEIN DATA BANK

PDB40
Symposium

October 28 - 30, 2011
Cold Spring Harbor Laboratory

Come celebrate four decades of innovation in structural biology

- 230 registered
- 34 travel awards
- 95 posters

Confirmed Speakers

- **Cheryl Arrowsmith**, University of Toronto, Canada
- **David Baker**, University of Washington
- **Ad Bax**, NIH/DHHS/NIDDK/LCP
- **Axel Brunger**, Stanford University/HHMI
- **Stephen K. Burley**, Eli Lilly & Co.
- **Wah Chiu**, Baylor College of Medicine
- **Johann Deisenhofer,** UT Southwestern Medical Center
- **Angela Gronenborn**, University of Pittsburgh
- **Richard Henderson**, MRC Lab. of Molecular Biology
- **Wayne Hendrickson**, Columbia University
- **Mei Hong**, Iowa State University
- **So Iwata**, Imperial College London
- **Brian Matthews**, University of Oregon
- **Jane Richardson**, Duke University Medical Center
- **Michael Rossmann**, Purdue University
- **Andrej Sali**, University of California, San Francisco
- **David Searls**, Independent Consultant
- **Susan Taylor**, University of California, San Diego
- **Janet Thornton**, EMBL, Hinxton,
- **Soichi Wakatsuki**, IMMS-KEK
- **Kurt Wüthrich**, The Scripps Research Institute, ETH Zürich

meetings.cshl.edu/meetings/pdb40.shtml

Funding: Strategy for Sustainability?

- RCSB PDB competitive renewal funded by NSF
  - January 2009 - December 2013
  - Dates and procedures for renewal under discussion
- PDBe competitive grant from Wellcome Trust
  - January 2010 - December 2014
- PDBj competitive renewal funded by JST (Japan Science & Technology Agency)☐
  - April 2011 - March 2014
- BMRB competitive renewal funded from the National Library of Medicine
  - NLM will no longer fund BMRB after 2014

Agenda

Overview	**Helen M. Berman**
Data In	**Jasmine Young** **Marina Zhuravleva** **Martha Quesada** **John Westbrook**
Data Out	**Philip E. Bourne** **Peter W. Rose**
Education and Outreach	**Shuchismita Dutta** **Gregory B. Quinn**
Journal Interactions	**Christine Zardecki**

Data In

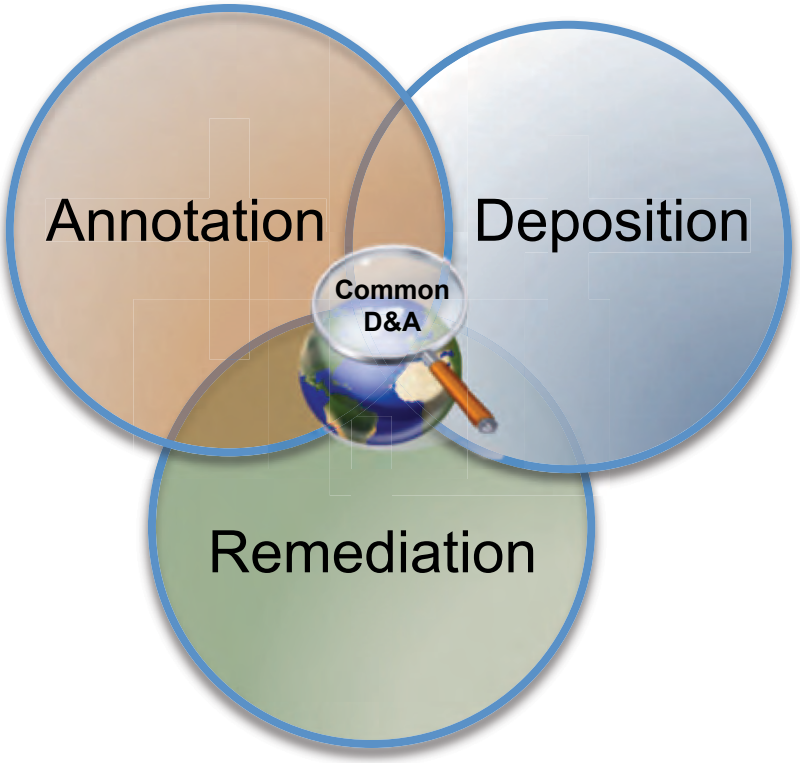
October 31, 2011

RCSB PDB
PROTEIN DATA BANK

Deposition, Annotation and Remediation

Jasmine Young

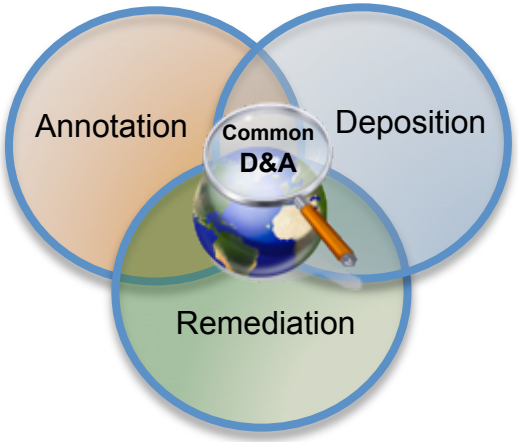
Interactivity of RCSB PDB Data In Activities



Each activity informs and contributes to the evolution of the others

Annotation Team

“Data In”



Goal: Enable research and discovery in the fields of structural biology and biomedical research

- Capture the experimental data defining the structure of macromolecules
- Maximize quality and completeness of data

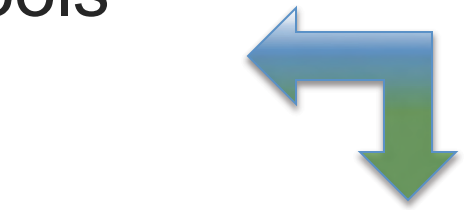
Data In Systems

Current RCSB PDB Deposition Tools

- ADIT: deposition tool
- PDB_extract: data harvesting tool
- Validation Server
- SF-Tool: converts and validates structure factor data
- Ligand Expo: search and create new ligands

Current RCSB PDB Annotation Tools

- Integrated Annotation Tool
- Chemical Component Tools
- Common D&A Ligand Module



Future:

wwPDB Common D&A System

Improved Existing Deposition Tools

Current RCSB PDB Validation Server

- Users can generate PDF reports anonymously
- Provides high-level geometric and experimental checking results

pdb_extract V3.11

- Support for hybrid methods
- Added support for new refinement and data processing programs/versions
- Better integration of extracted data into deposition pipeline - improved referential integrity of harvested data

SF-Tool

- Better handling of user defined data items in SF files
- Support for neutron and X-ray hybrid data

Both **pdb_extract** and **SF-Tool** will be incorporated into the D&A Deposition System

2011: RCSB PDB Annotated 66% of All Depositions

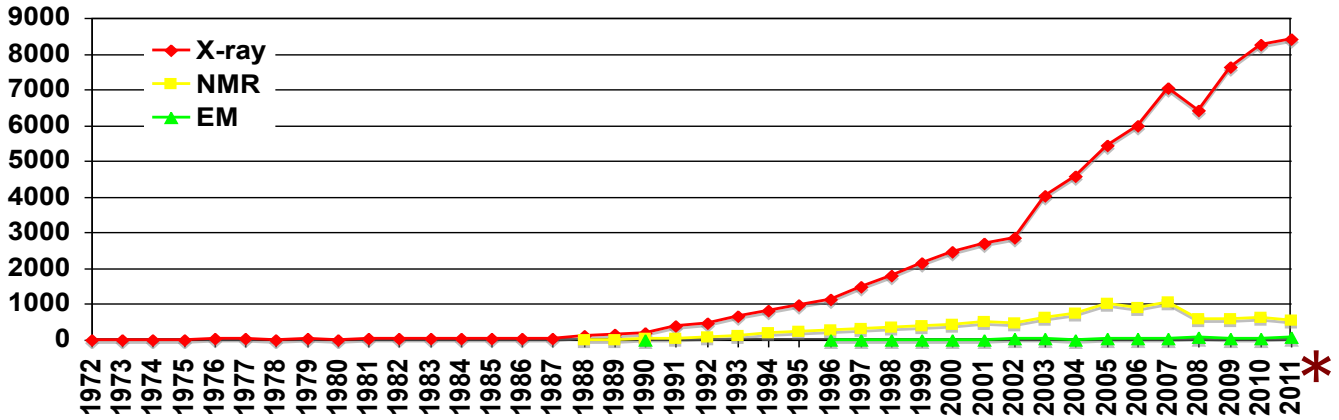
By deposition and processing Site

Last Updated: 5 Oct 2011

Year	Total Depositions	Deposited To			Processed By		
		RCSB PDB	PDBj	PDBe	RCSB PDB	PDBj	PDBe
2000	2983	2445	10	528	2297	158	528
2001	3287	2673	118	496	2408	383	496
2002	3565	2769	289	507	2401	657	507
2003	4830	3488	673	669	3135	1026	669
2004	5508	3796	900	812	3082	1614	812
2005	6678	4507	1166	1005	3563	2110	1005
2006	7282	5145	1052	1085	4252	1945	1085
2007	8130	5399	1603	1128	4703	2299	1128
2008	7073	5452	648	973	4106	1994	973
2009	8300	6715	527	1058	5069	2173	1058
2010	8878	6912	593	1373	5464	2041	1373
2011	6896 (*9088)	5370	409	1117	4528	1251	1117
TOTAL	73410	54671	7988	10751	45008	17651	10751

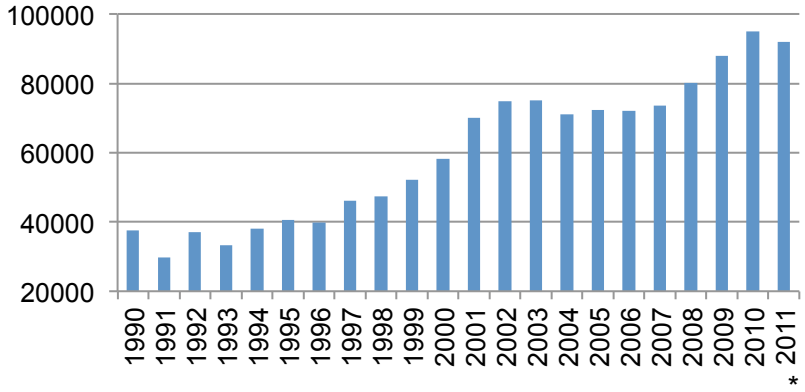
Experimental method distribution

(Updated 5 Oct 2011;
* projection for 2011)

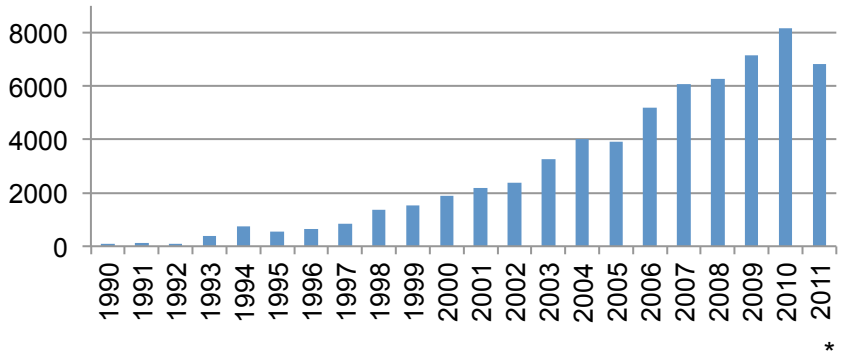


Evolving Complexity of PDB Entries

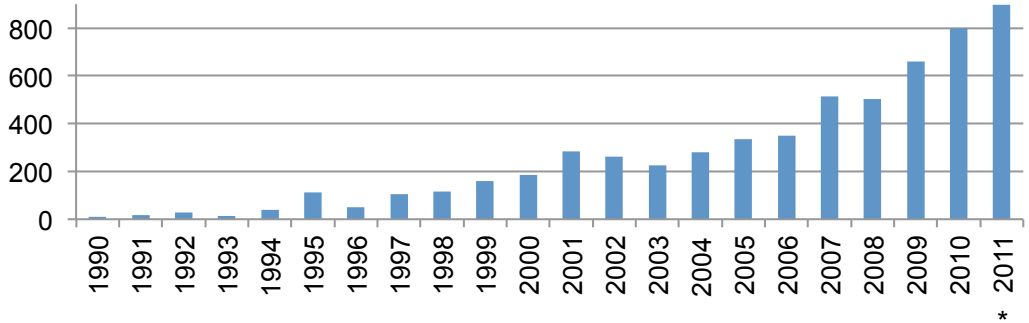
Polymer Molecular Weight



Number of Organic Ligands Deposited



Number of Entries with Inhibitors and Antibiotics



* 2011 - 9 months

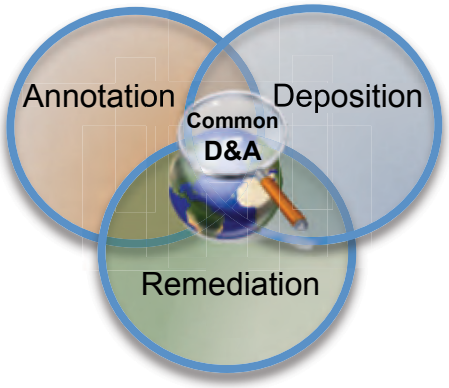
Improving Efficiency

Staff-Driven

- Domain experts hired
- Expert teams
  - Small and large ligands
  - Viruses
- “Jamborees”

Improvements to Production Pipeline

- Sequence processing
- Common D&A ligand module



Common D&A Ligand Module Incorporated into Production Pipeline

Instance: 1_A_B12_800_

TOP CANDIDATE RESULTS FOR: 1_A_B12_800_			
CANDIDATE ID	ASSIGN AS:	COMPOSITE SCORE	COMPARE
B12	○	100 / 86 / 86 / match / 85	☑
COB	○	98 / 92 / 92 / match / 85	☑
CNC	○	97 / 92 / 92 / match / 85	☐
COY	○	85 / 82 / 82 / match / 95	☐
OTHER CANDIDATES ADDED FOR COMPARISON			
HEM	Must Force Assign	n.a.	☑

Force Assign as: [ ]

Rerun Search

Edit / Create New Ligand

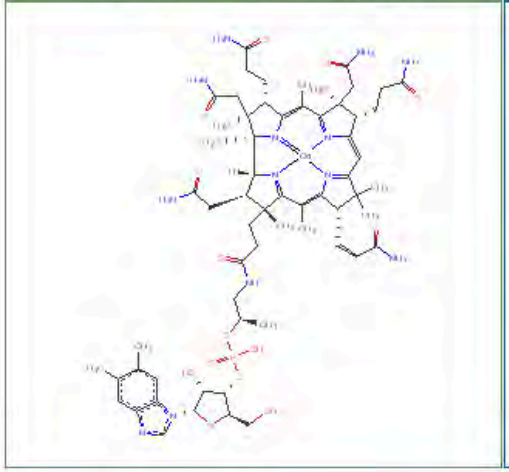
Chop Ligand

Enter chem component ID to add other candidate to Comparison Panel below: [ ] Add

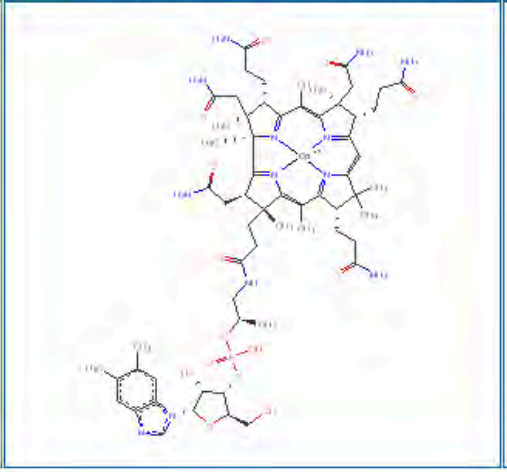
Assign

COMPARISON PANEL 2D ☑ 3D ☐ ATOM MAP ☐

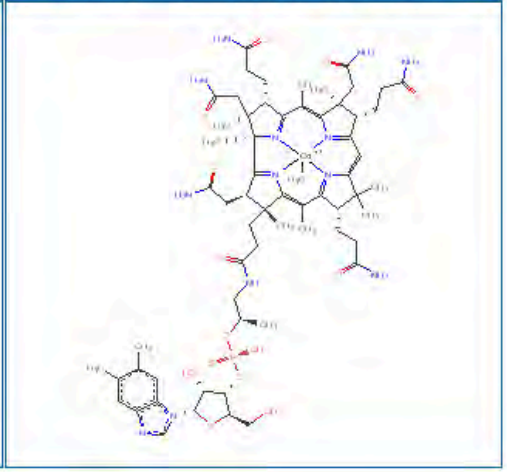
Auth Instance ID:	1_A_B12_800_
Name:	None
Formula:	C62 H87 Co N13 O14 P



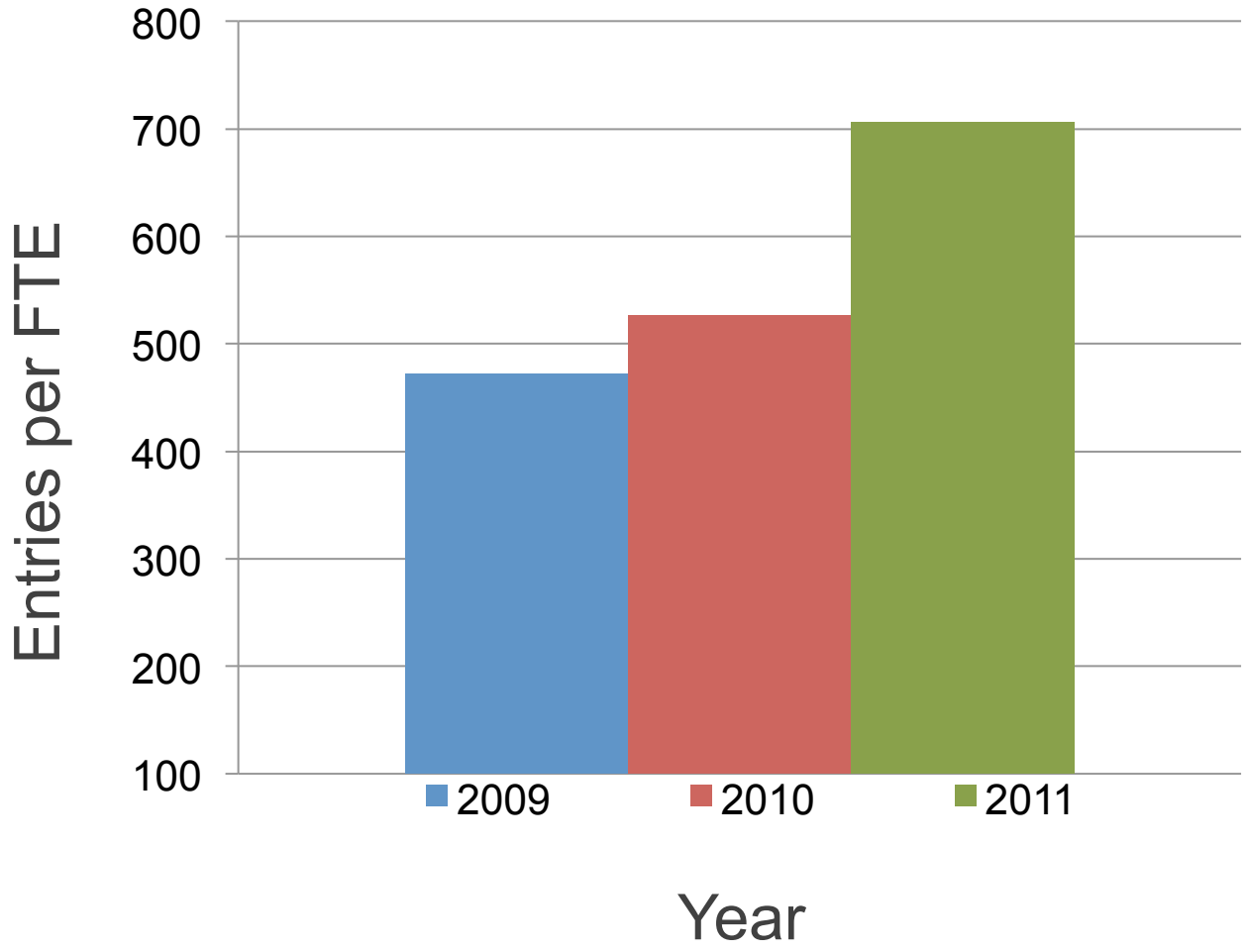
Top Dictionary Hit:	B12
Name:	COBALAMIN
Formula:	C62 H88 Co N13 O14 P



Dictionary ID:	COB
Name:	CO-METHYLCOBALAMIN
Formula:	C63 H91 Co N13 O14 P



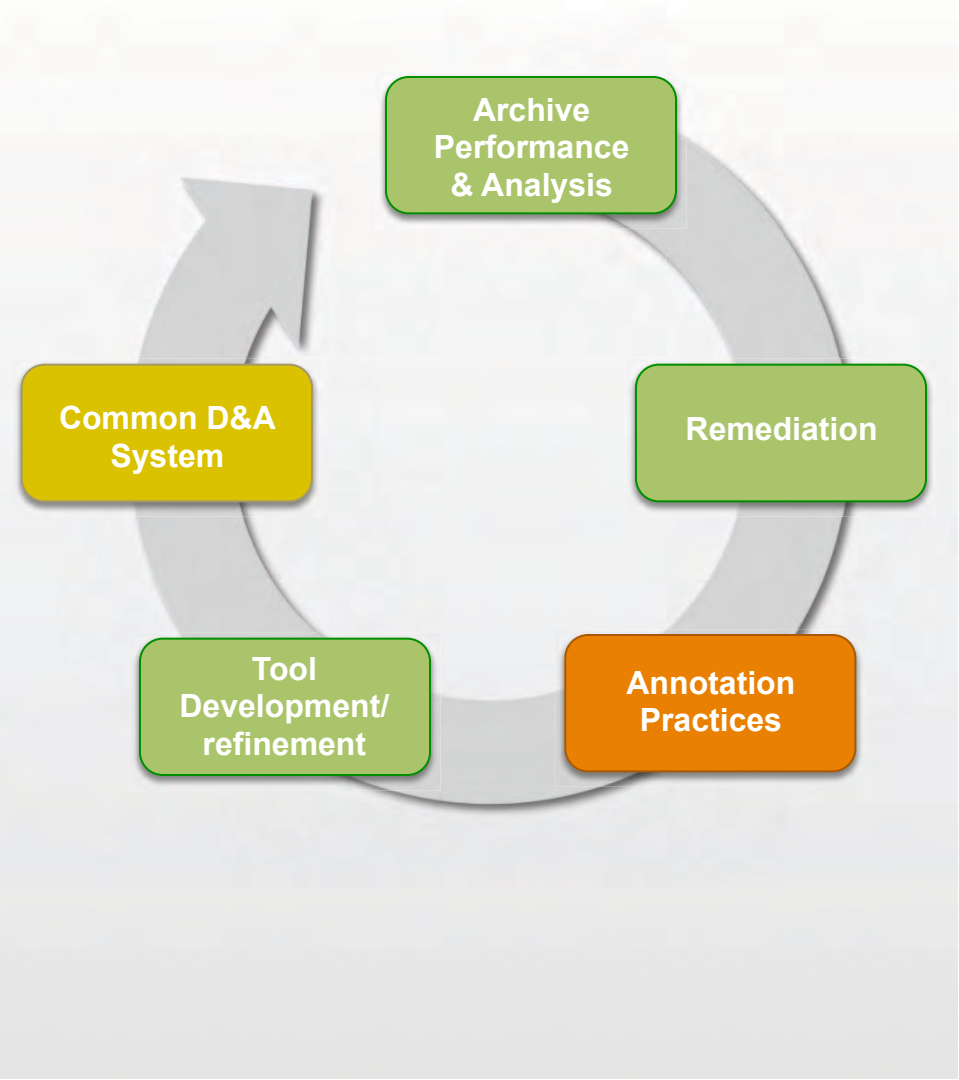
Improved Productivity



Remediation

- Informs all processes
- Improves consistency in file annotation
- Enhances chemistry representation

Better query capability



Remediation 2010-2011

1. Biological Assemblies

   Incomplete computational annotation in 6126 entries

   - Entries updated with curated PQS and PISA results
   - D&A implication: to be captured at deposition

2. B Factor

   Residual B factors deposited in place of full B factors

   - 7310 problematic entries identified and tagged
   - D&A implication: residual B factors to be identified and resolved at deposition

Remediation 2010-2011

3. Non-standard Crystal Frame

   Difficulty with transformation between Cartesian and fractional coordinates

   - 148 entries were identified and tagged

4. Hybrid X-ray/Neutron Diffraction Structures

   Relationships between method and data collection details to be better represented in the PDBx file

   - The PDBx exchange dictionary has been extended to handle hybrid X-ray/neutron diffraction methods

Remediation 2010-2011: Improved Chemistry Representation

5. Peptide Inhibitors/Antibiotics:
   Non-uniformity in representation
   - Chemistry representation corrected and standardized in 1029 entries, sequence and molecular views supported
   - Peptide Reference Dictionary created, tools developed.

6. Nonstandard Polymer Linkages:
   Use of standard amino acid residue names to represent nonstandard linking is misleading
   - Definitions for non-standard chemical components created and 58 entries corrected

Remediation 2010-2011: New Revision Log in PDBx/PDBML Files

- Any changes made to the data are recorded in the PDBX_VERSION data category

- A revision log created for this release is available at wwPDB website (XLS and CSV)

New Remediation in Progress

- Carbohydrates

Multiple representations in naming and linking

  - Archive analysis of carbohydrate-containing entries
  - Incorporate standard nomenclature
  - Incorporate standard representation for branched polymers

- Post Translational Modifications

Inconsistent annotation in archival files

  - PTMs will be identified and annotated

- Recalculation of B factors

Partial B-value cases labeled, but not replaced with full isotropic values

  - Full B-values will be recalculated and added to the data files.

Chemistry Quality Assessment

Marina Zhuravleva

Data Quality and Value Annotation

Annotation and validation specific to biopolymer entities (protein, DNA, etc.) have been worked out thoroughly

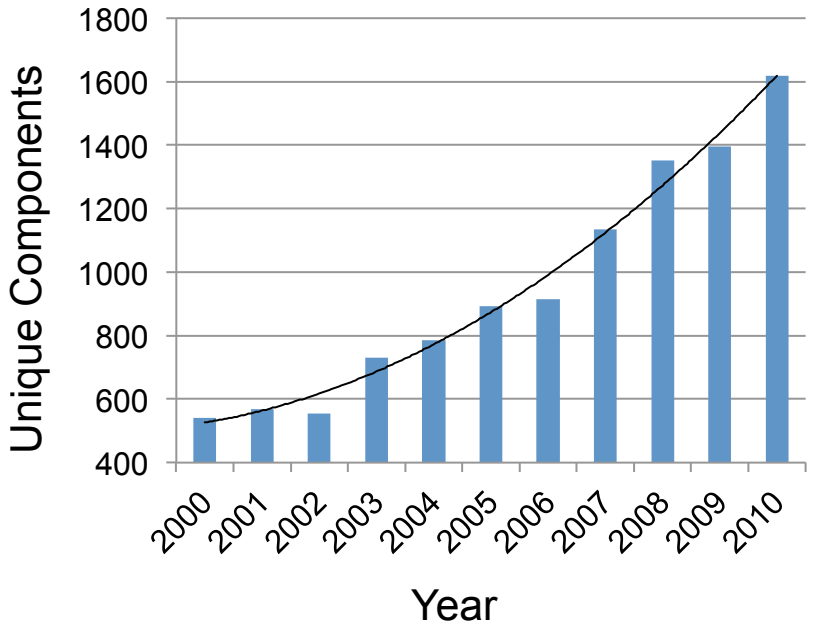
Annotation and validation procedures for small molecules are still lagging:

- Geometry Validation
- Molecule Presentation
- Sequence Reference
- Source Information
- Classification/Biological Function

Chemical Diversity at PDB

In addition to biopolymers, the PDB archive contains ~15,000 unique small molecule entities (components).

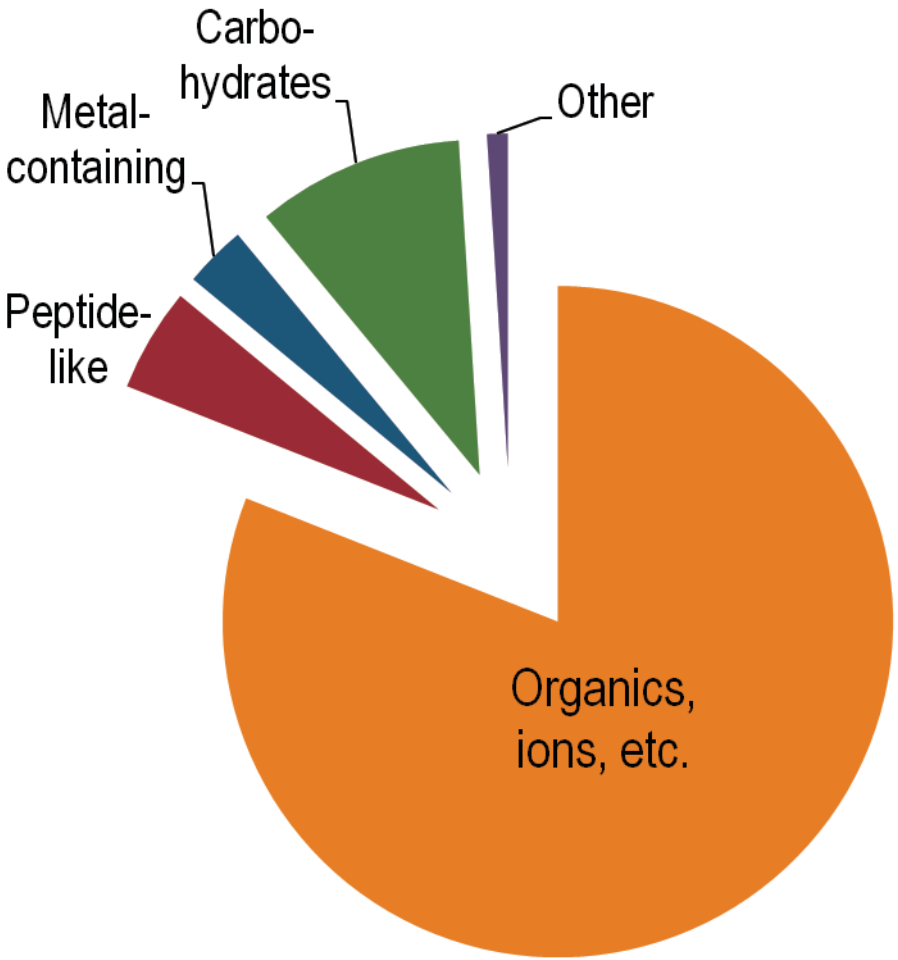
- Chemical Components Dictionary provides a systematic, standard and common point of reference for Components
- The diversity, complexity and number of components in PDB entries are constantly growing



Chemical Components Processing

- Chemistry assignment
  - Atomic positions and atom types come from deposited coordinates
  - Connectivities and bond types are derived from molecular geometry
- Geometry validation available only for modified amino acids and nucleic acids
  - Working with CCDC to use high resolution models as targets
- Search against Chemical Component Dictionary
  - Match to existing definition
  - Commit new component

Chemical Components Dictionary

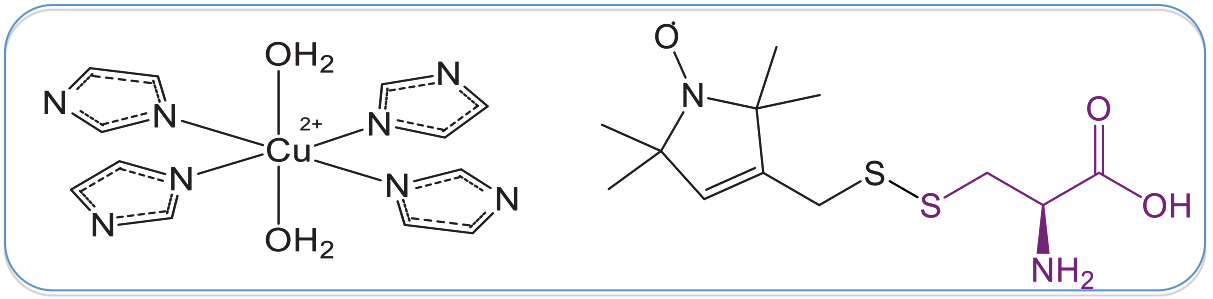


- Chemical components catalogued as chemically reasonable entities wherein atom types, bond orders, chiralities are defined and valences are satisfied
- Additional information is noted within the data files via flags and tokens
- Ideal coordinates computed

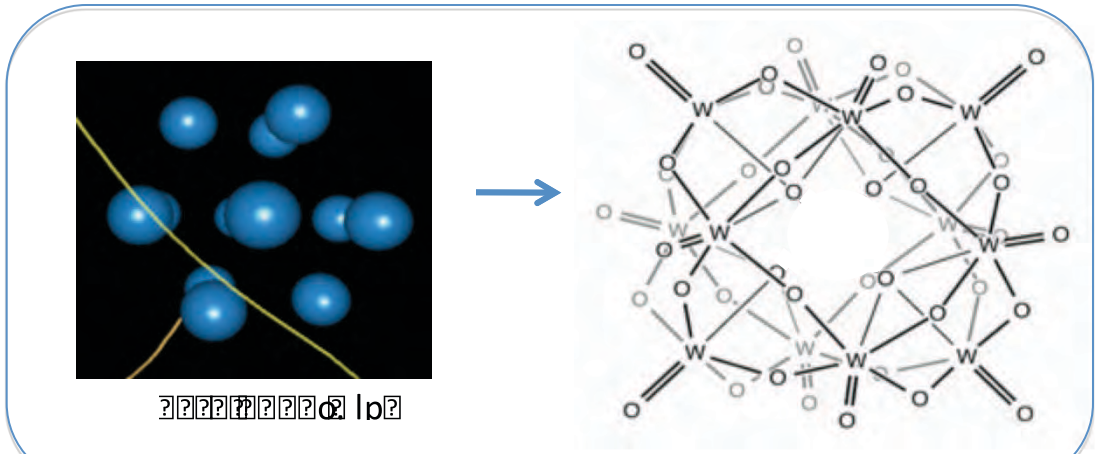
Challenges in chemical presentation:

- Component identity
- Incomplete models
- Dependent components
- Chemical reactions
- Oligomeric molecules

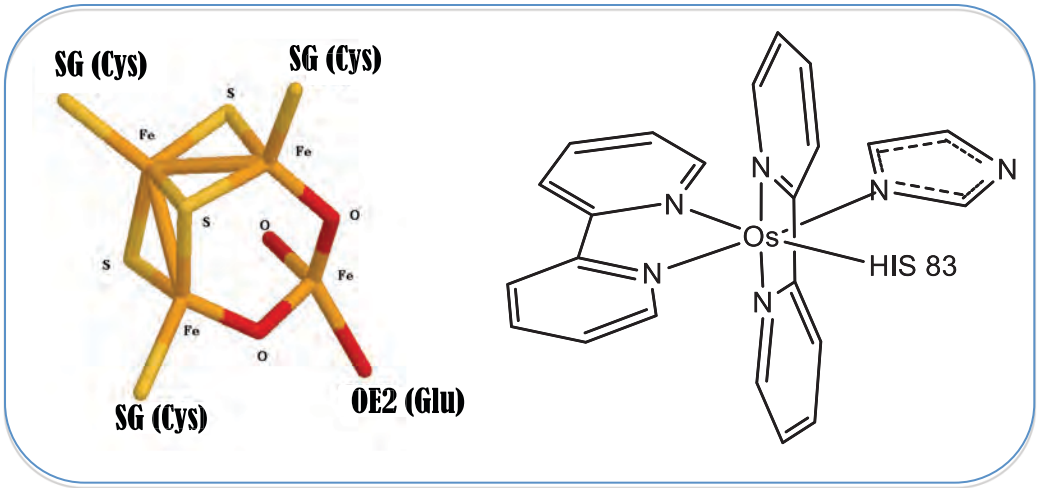
Component identity



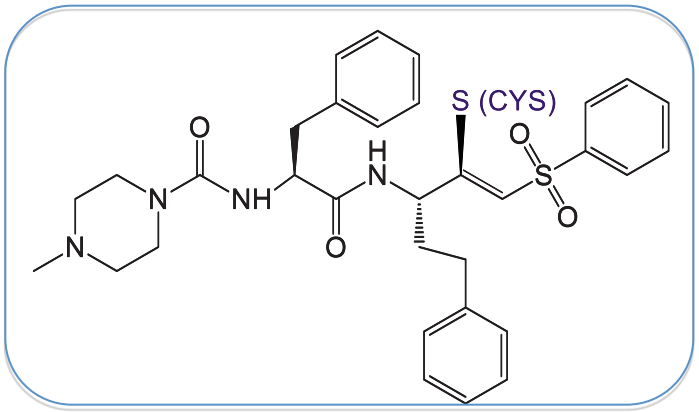
Incomplete models



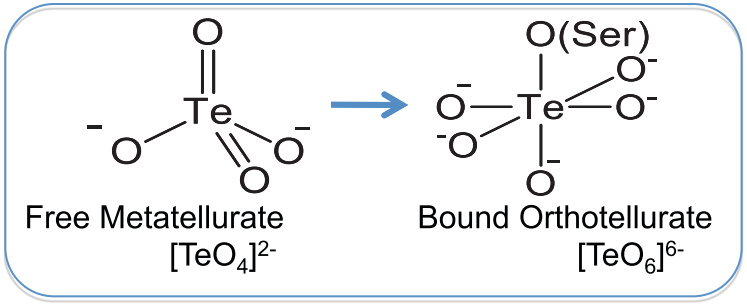
Dependent components



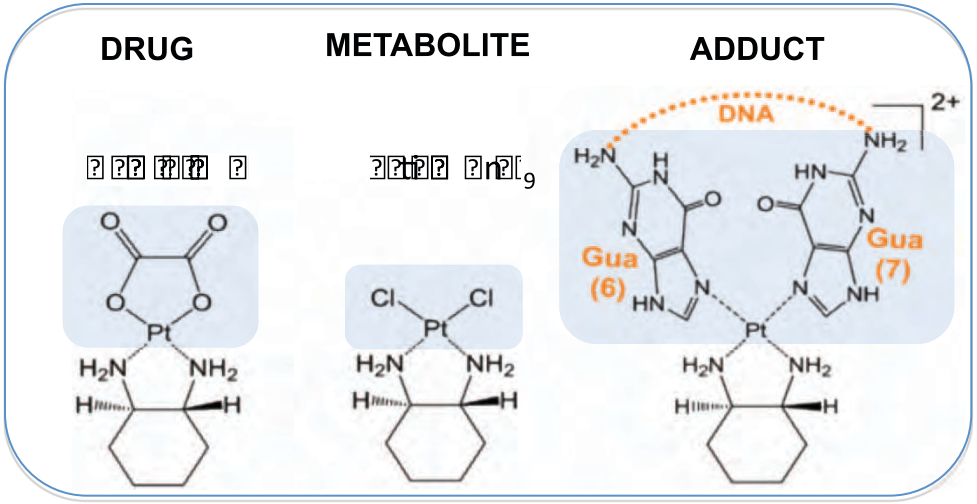
Reactions resulting in bond order change & introduction of stereocenter



Reactions resulting in geometry/oxidation state change

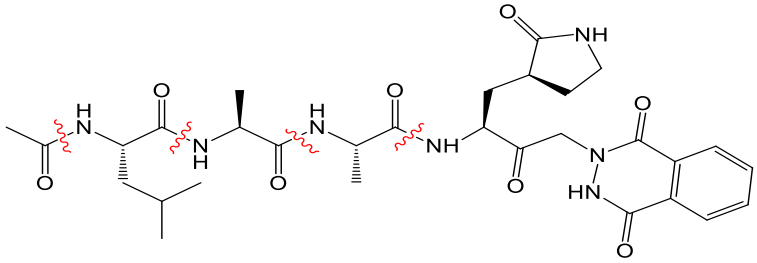


Reactions involving leaving atoms and groups



Oligomeric molecules

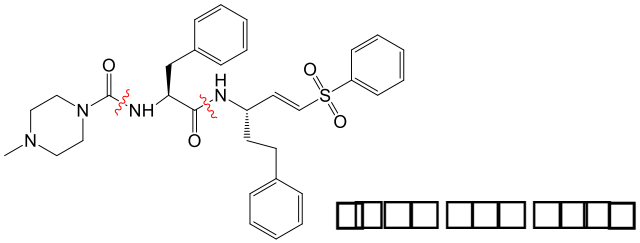
- Peptide-like molecules
- Carbohydrates
- PEGs and Jeffamines

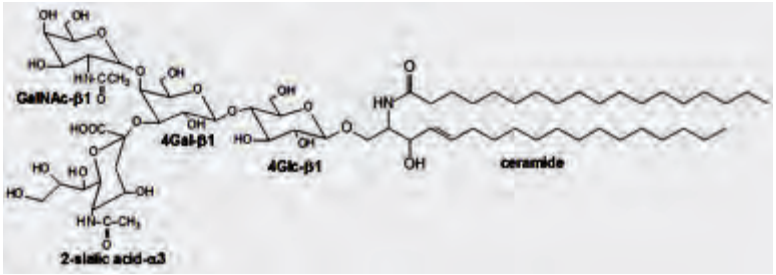


ACE LEU ALA ALA OCQ

Challenges

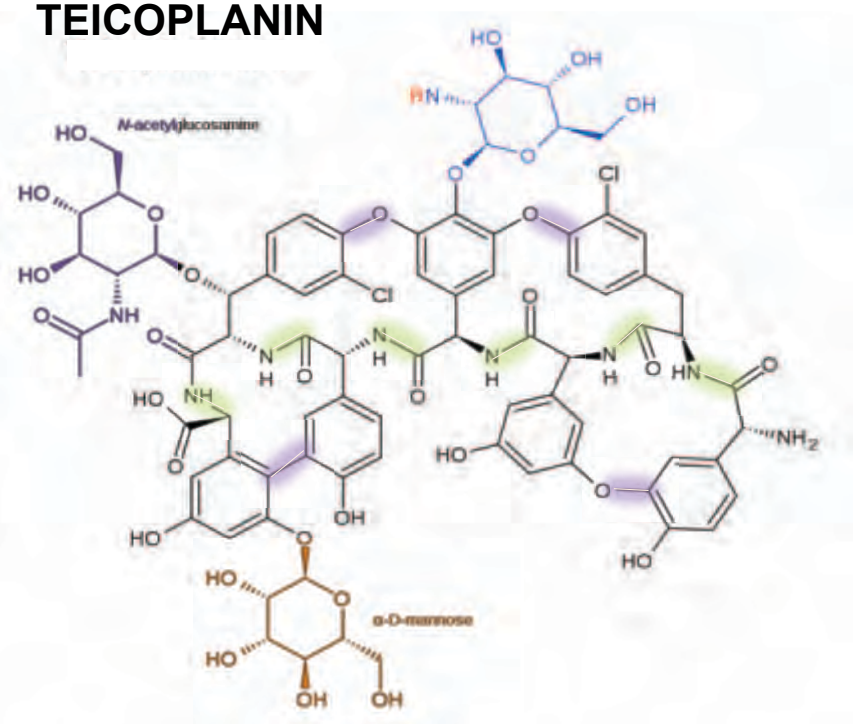
- Molecule presentation (polymeric sequence vs. single molecule)
- Decorations (sugars, lipids, etc.)
- Branching
- Linkages (glycosidic, iso-peptide)
- Classification





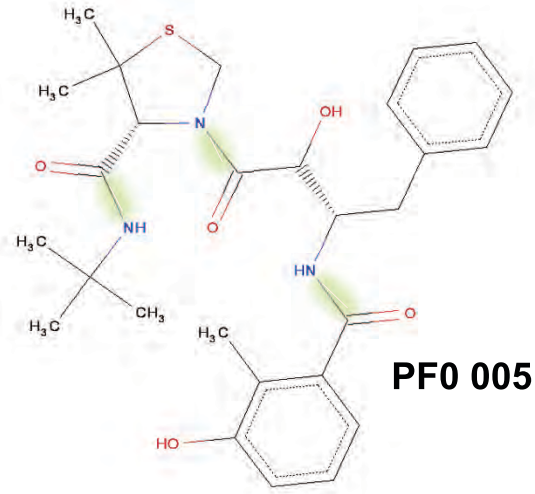
Oligomeric Molecules: Peptide-like 2011 Remediation

TEICOPLANIN



- Present as a polymer per biosynthetic pathway
- Provide group concept to describe decorations
- Provide sequence data base reference
- Provide host organism source information
- Capture evolutionarily-related families

KNI-577 HIV PROTEASE INHIBITOR



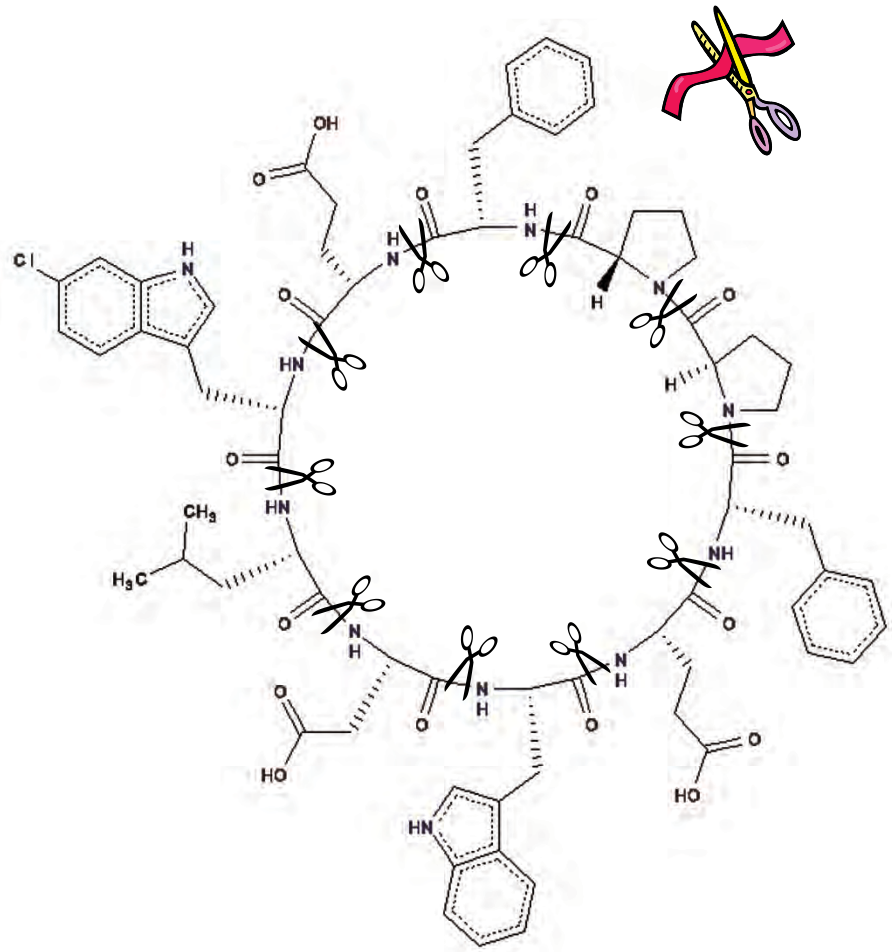
PF0 005 00B NTB

- Present as a single molecule
- Provide sequence information per chemical synthesis
- Capture functional and structural classes

Molecule Presentation and Sequence Remediation

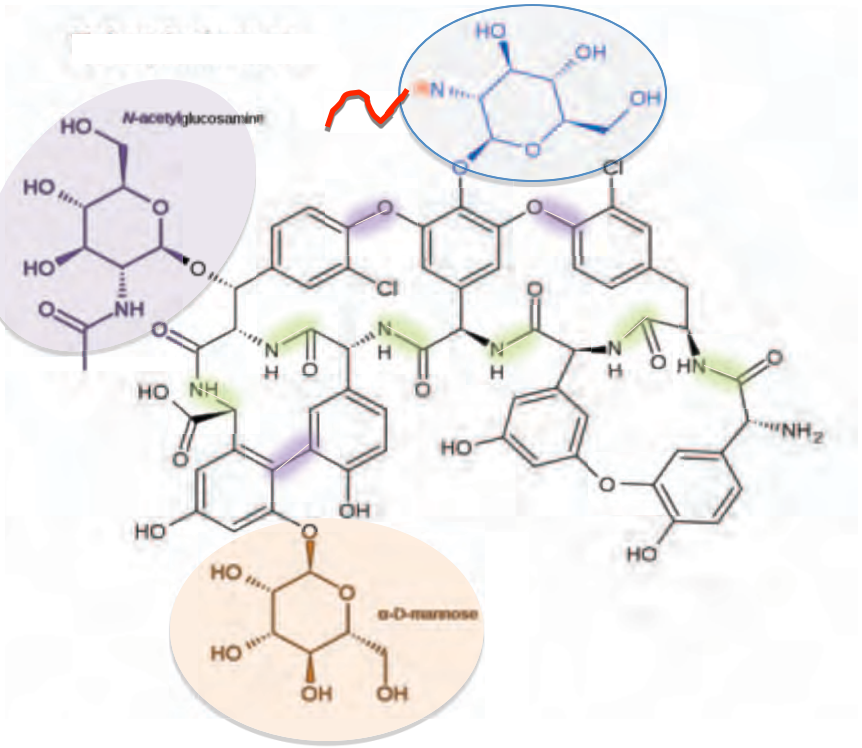
- SeqUeNce → SeQuEnCe
- Sequence + HET → SEQUENCE
- HET1 + HET2 + HET3 → SEQUENCE
- HET1 + HET2 → Component
- Component → SEQUENCE
- Component → ComPoNent
- ComPoNent → CoMpoNenT

Sequence Remediation: Peptide Molecule Chopper Tool



- Significantly improves efficiency and throughput
- Allows annotator to specify bond breaks
- Adds leaving atoms or groups
- Standardizes atom nomenclature
- Maps individual residue nomenclature to full molecule

Decorations: Group Concept



T55

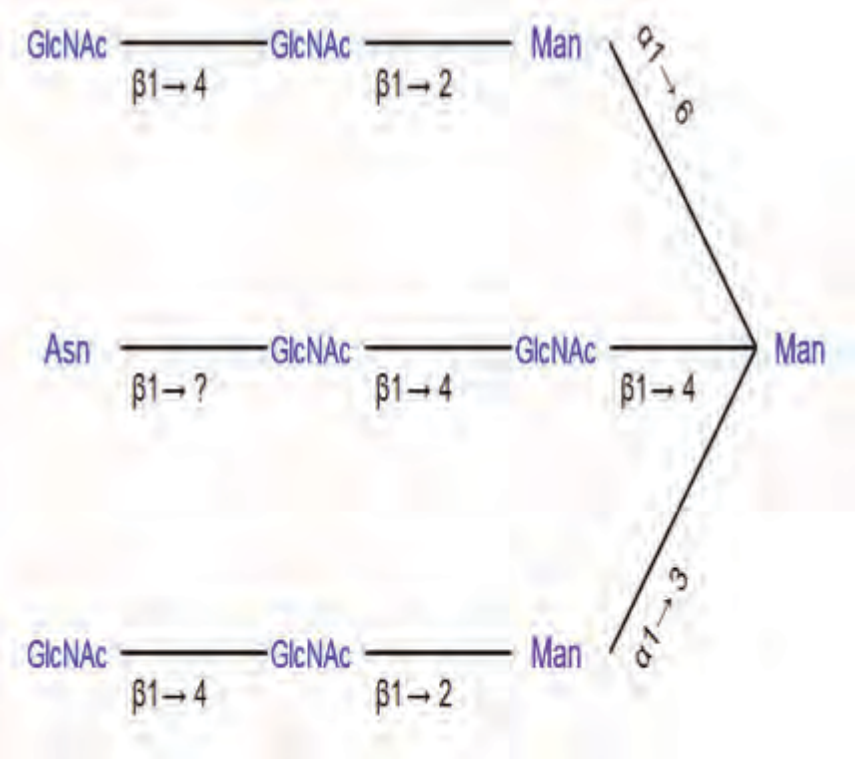
GSC NAG BMA

GHP 3MY 3FG GHP GHP OMY 3FG

REMARK 400
REMARK 400 GROUP: 1
REMARK 400 NAME: TEICOPLANIN
REMARK 400 CHAIN: E, F, G, H
REMARK 400 COMPONENT_1: PEPTIDE LIKE SEQUENCE RESIDUES 701 TO 707
REMARK 400 COMPONENT_2: SUGAR RESIDUES 708, 709 AND 710
REMARK 400 COMPONENT_3: FATTY ACID RESIDUE 711
REMARK 400
REMARK 400 DESCRIPTION: TEICOPLANIN IS A TETRACYCLIC HEPTAPEPTIDE
REMARK 400 GLYCOSYLATED BY THREE MONOSCCARIDES, RESIDUES 708, 709
REMARK 400 AND 710, ON RESIDUES 707, 706 AND 704, RESPECTIVELY. THE FATTY
REMARK 400 ACID IS LINKED TO THE BETA-D-GLUCOSAMINE (RESIDUE 710)

Oligomeric Molecules: Carbohydrates Future Remediation

- Need consistent molecule presentation
- Decorated (lipids, peptides, etc.)
- Branched (non-linear)
- Linkages (α1,3; β1,4 etc.)
- Attachment to protein (alpha vs beta)
- Classification

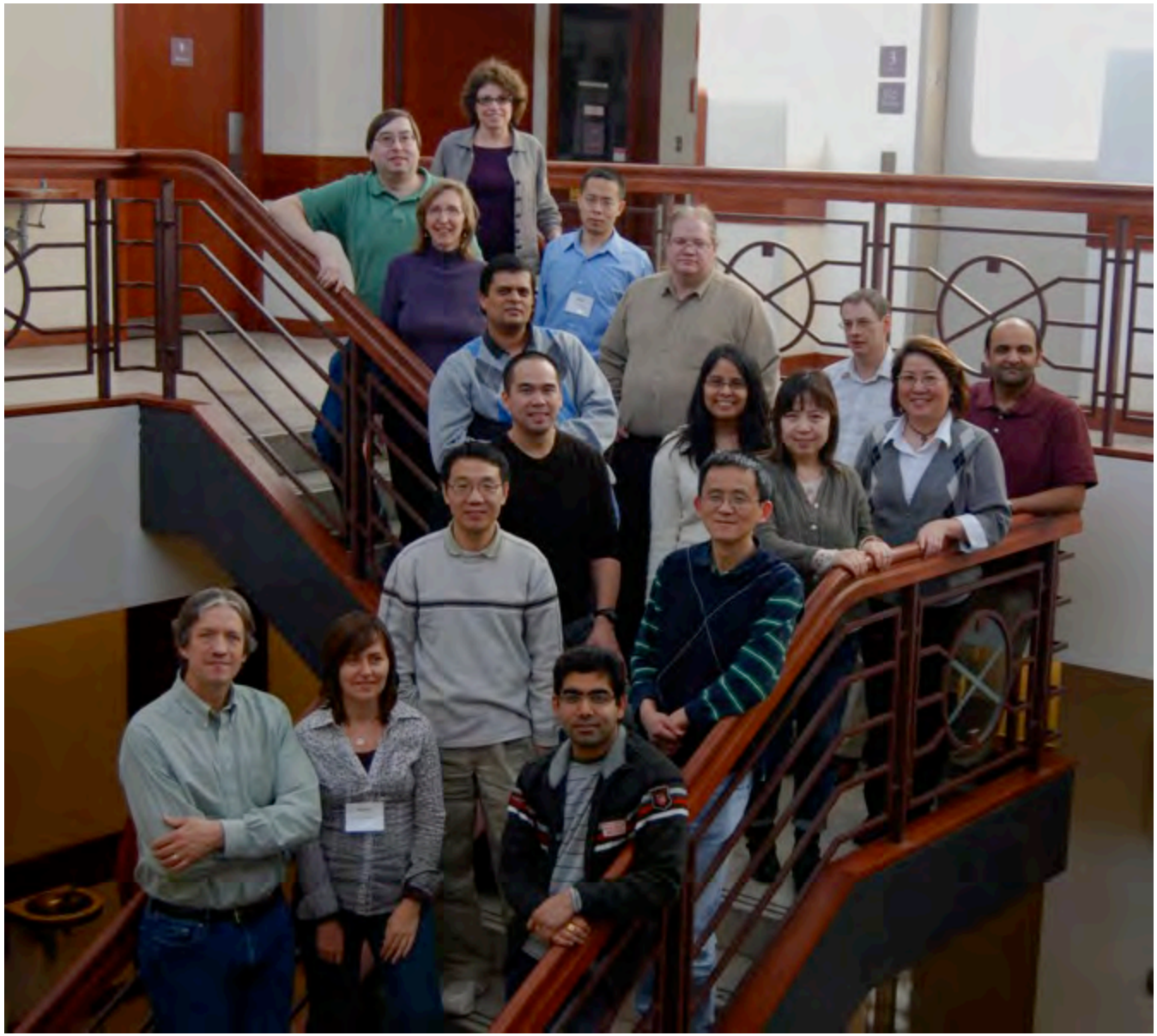


wwPDB Common Deposition and Annotation Tool

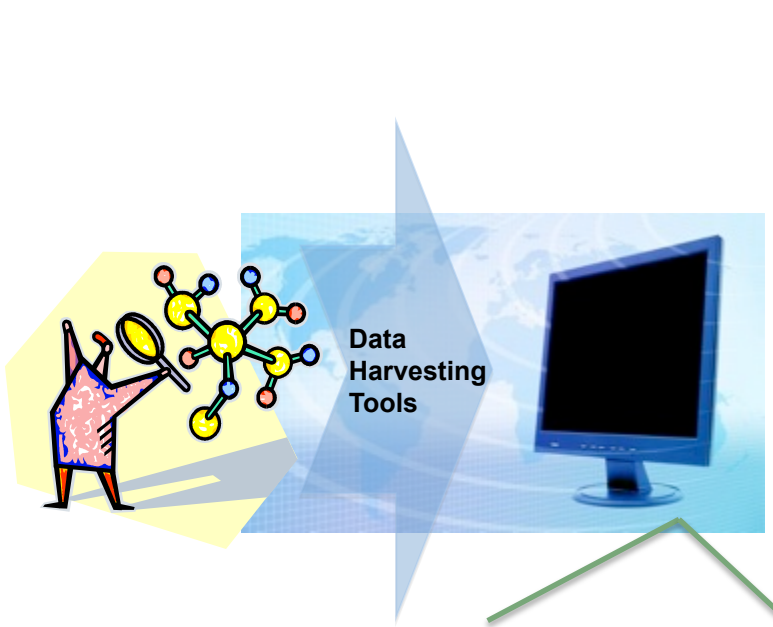
Martha Quesada

Common D&A Project Team March 2011

Experience, expertise and diverse skills representing the broad interests of wwPDB



The Vision



Common Deposition Interface
- Accession ID
- Validation Report
- Other

Author info, Citations

Restraints

Chemical shifts

NMR exp details

X-ray exp details

X-ray SF

EM exp details

EM maps

Integrated Data Capture

Coordinates and Models

BMRB Processing Pipeline

BMRB Entry

BMRB FTP

PDB Processing Pipeline

PDB Entry

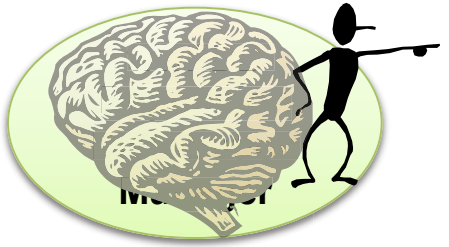
EM maps Processing Pipeline

EMDB Entry

wwPDB FTP

Workload Balance

- Depositions will be distributed taking into account:
  - Deposition session restart preference
  - Advisory and funding guidelines
  - Time zone: facilitate “help” and communication
  - Load balance: even distribution with respect to each site’s local capacity (e.g., taking into account local holidays)



- Single, wwPDB-branded, point of contact for all new depositions (e.g., http://wwpdb.org/deposit)

wwPDB Common Deposition and Annotation Pipeline

Deposition Pipeline

- Deposition Interface
- Data upload, harvesting, 1° test and verification
- Client side Editor
- Validation
  - Standalone and integrated
- Calculated annotations (PISA, SITE & LINK records, cross references, metal coordination)
- Submission
- Progress Tracking/ Status

Communication System

Workflow-Automation System

Annotation Pipeline

- Sequence Processing
- Ligand Processing ID, Edit, Build
- Validation
- Calculated annotations (PISA, SITE & LINK records, cross references, metal coordination)
- Corrections
- Release Processing
- Progress Tracking/ Status

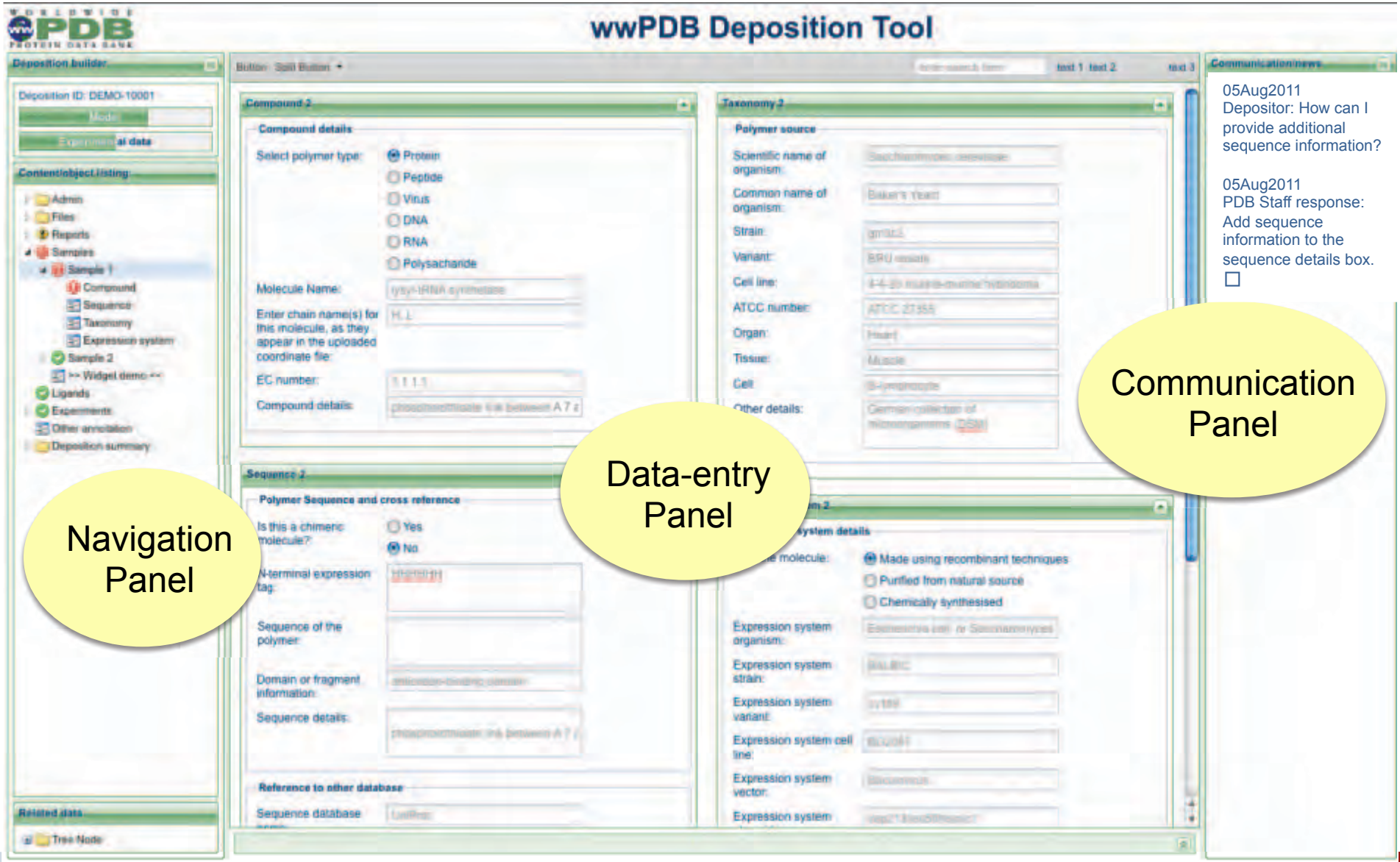
Deposition Interface Design and Community Input

- wwPDB partner groups: initial requirements and design
- Introduction to community at ACA 2010
- Deposition user interface: initial feedback at IUCr 2011
  - *Iterative evolution*
- Interface review by targeted external user representatives (November 2011)
  - *Iterative evolution*
- Broader review by community experts (January 2012)
  - *Iterative evolution*
- Community beta testing to begin Q3 2012

Interface Features for Depositors

- Automated batch data uploads
- Flexible manual data entry
- Restart deposition and re-upload data without loss of general information
- Build new submissions on previous depositions
- Easily view percentage complete
- Visually review data
- Structure validation reports

Interface Look and Feel: Deposition



EM Integration

- Functional requirements 90% completed
  - Dictionary for incorporation into D&A
  - Interface requirements underway
- Large data file requirements to be supported in V1.0 of the deposition module
- Additional visualization, data harvesting to be supported in V1.X
- Validation requirements from EM VTF to be supported in VN.0

NMR Integration

- Dictionary data items supporting NMR have been defined
- Data requirements defined for chemical shifts
- Integration of software for PDB atom nomenclature correspondence to NMR experimental data
- Implement Common D&A and ADIT-NMR data exchange

Technology and Development

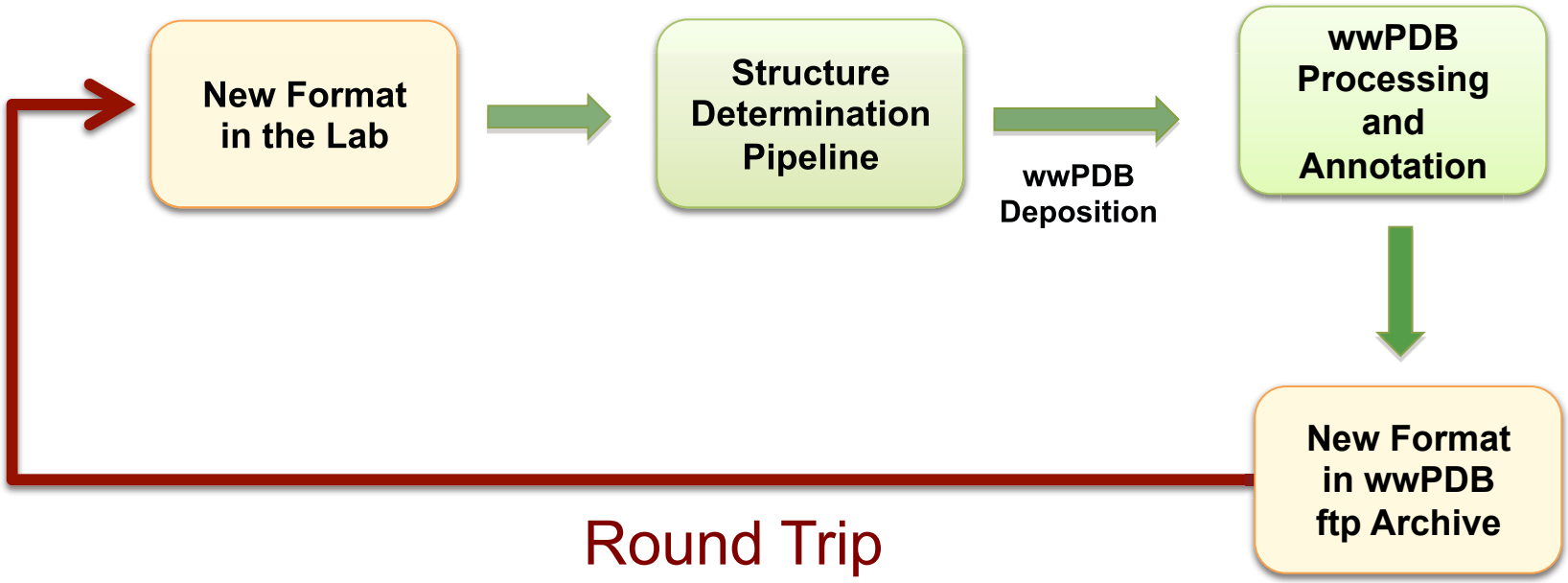
John Westbrook

Major Technical Challenges

- Multiple deposition data formats pose problems for data exchange and representation
- Different deposition and annotation systems currently in use at wwPDB sites

Deposition Data Format Options

- Current or compatible PDB format
- PDBx/mmCIF archival/exchange format
- A new or hybrid data format



PDB Format Issues

- PDB format is almost 40 years old and does not support today's science
  - Let alone tomorrow's science…
- Some key limitations include
  - Max 62 chains
    - and that's stretching it
  - Max 99,999 atoms
    - 5 ribosomes in ASU=10 PDB entries!
  - Very short chain, residue and atom names
    - 1, 3, 4 characters, respectively
  - No bond orders or chirality specified for ligands
  - No support for NMR, EM, hybrid methods, …
  - Meta-data specification cumbersome and inflexible

Other Format Options

PDBx/mmCIF

- Community/IUCr standard
- Well-supported within PDB and *technically sound*
- Not perceived as a popular option due to its complexity

Hybrid format solution

- Prototyped and circulated a hybrid format for review among key developers and users
- Conducted a workshop with participation from major structure determination systems

Format Workshop 26/27 Sept 2011

Format Workshop Outcomes

And the
"New PDB Deposition Format"
is …

PDBx

with a supporting report and
presentation format

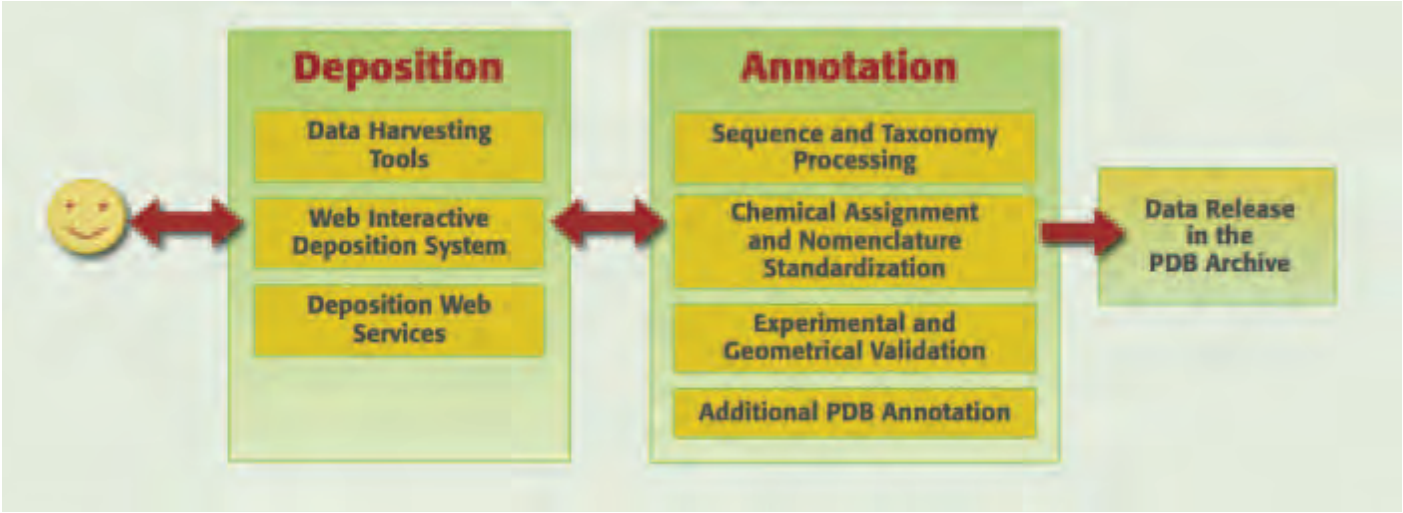
Data Format Plan

- PDBx as working and deposition format
  - Commitments from CCP4, Phenix and Global Phasing (*i.e.*, ~85% of all PDB depositions)
  - Agreement on managing development between these software providers and wwPDB
  - Projected completion – January 2013
- New, simplified, future-proof PDB report format
  - Think “wide-PDB”
  - Will be developed by wwPDB with input from stakeholder communities

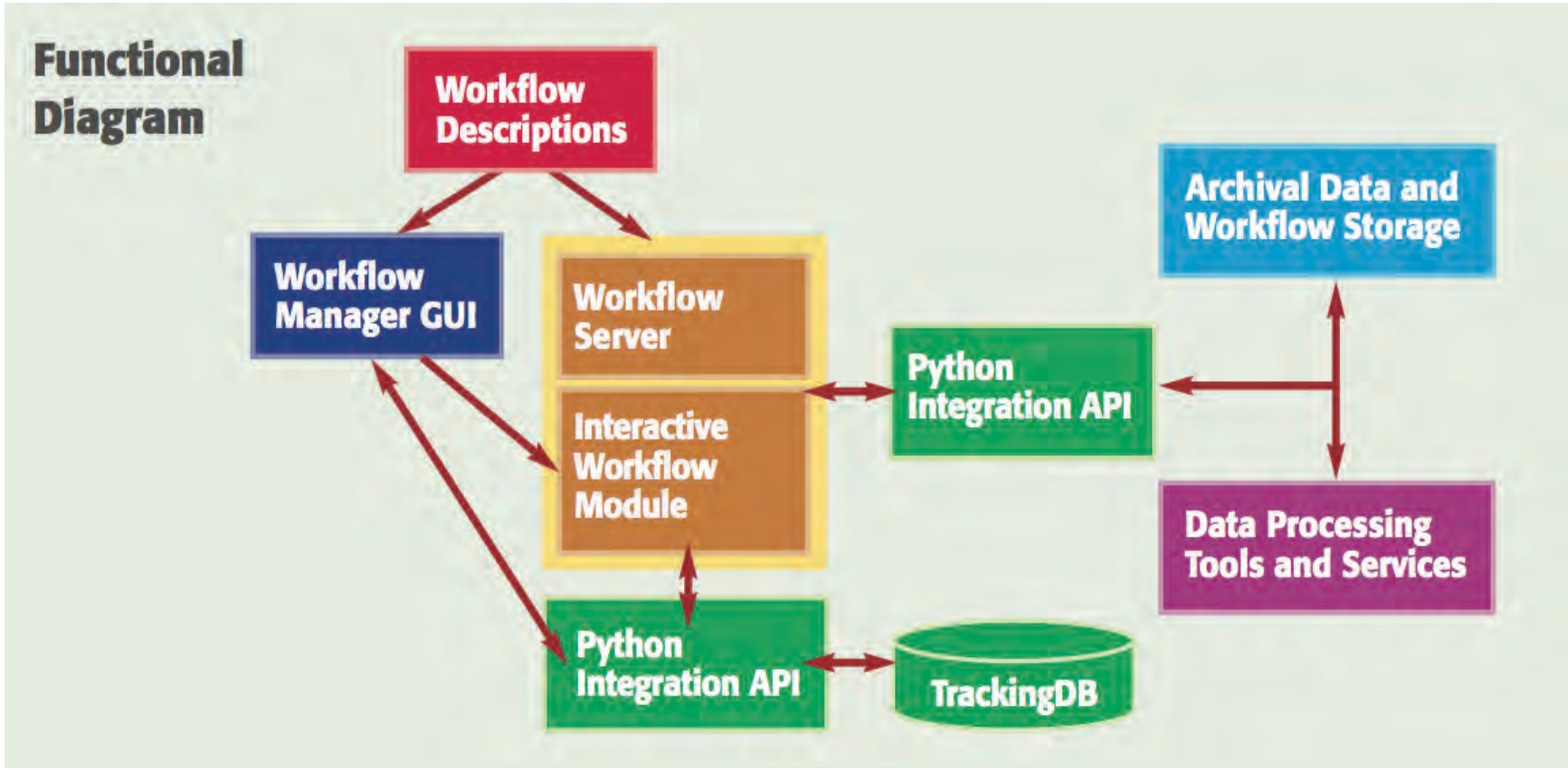
Building the New Common D&A System

Design Goals

- Greater automation and standardization of all deposition and annotation data processing tasks using a workflow system
- Tracking of important deposition and annotation steps
- Simplify communication with depositors and among annotators
- Easy deployment at all project sites
- Along with a long list of software engineering goals – portability, robustness, maintainability, extensibility, fault-tolerance



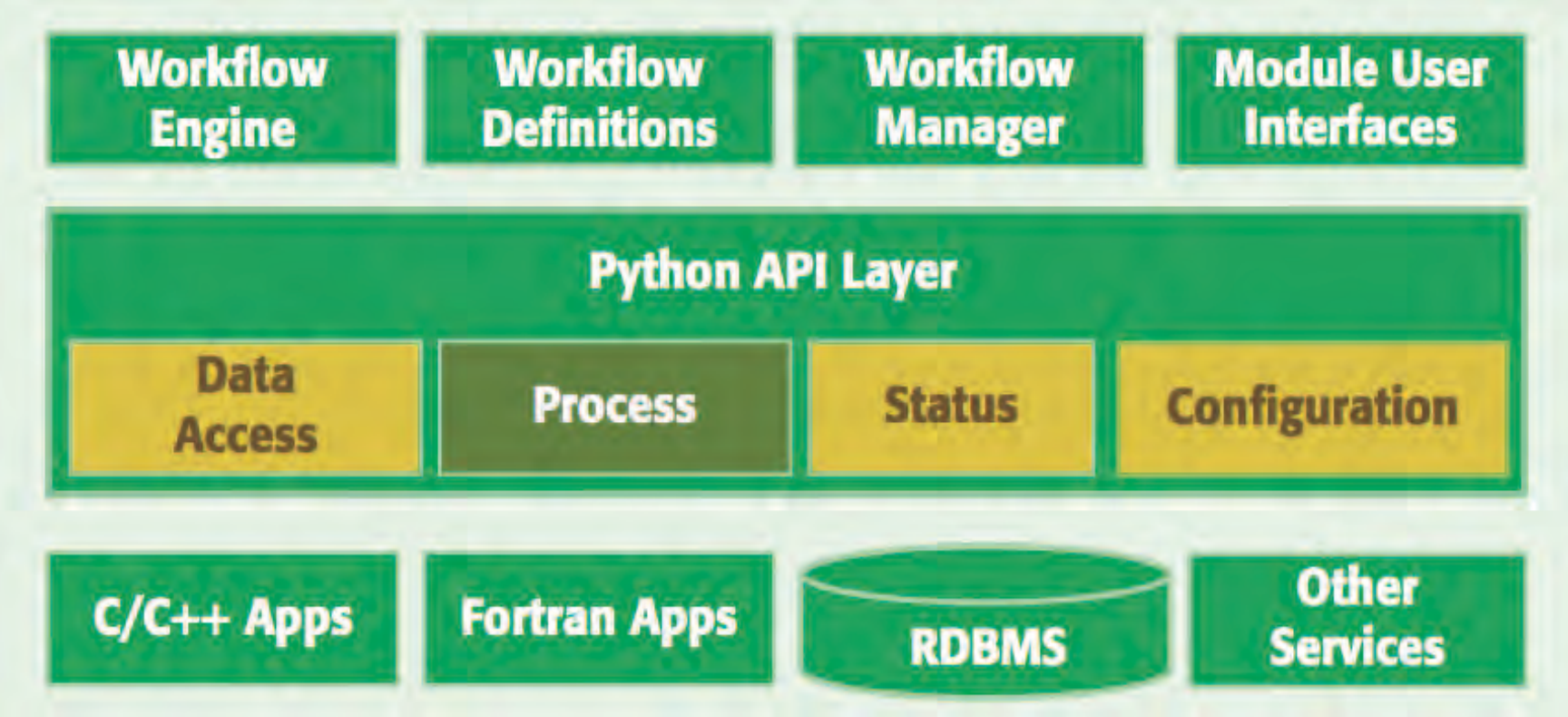
Workflow Architecture



Workflow System Evolution

- Workflow tracking performance impact shown to be manageable
- Workflow tasks aligned with process milestones
- Annotation workflows are being reused by deposition system
- Workflow supports remote execution of CPU intensive tasks

New System Software Architecture

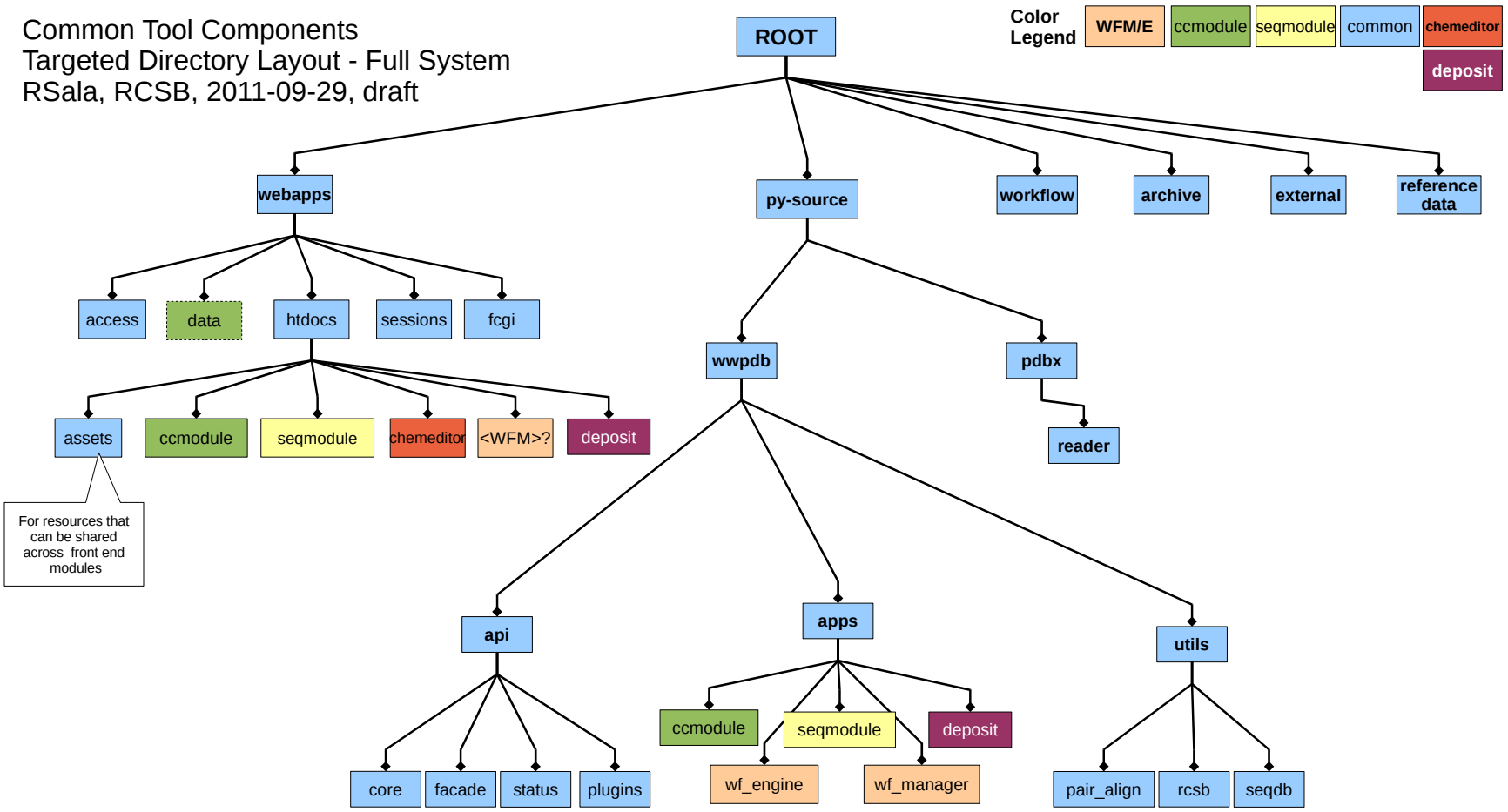


System Development and Deployment

- Rapid and flexible development
  - JavaScript/CSS for user interfaces
  - Python + wrappers for middleware
  - C++ for compute intensive applications
- Development has been □unit test driven□
- Created shared systems to host both development and integration testing across sites
- Software components are managed in SVN
- Site specific automated build and deployment tools are catching up with development

Enabling Deployment

Project tree contains all system components and dependencies.



Timeline

- Common Tool released for public use late 2012
- Full integration testing of the D&A pipeline modules to begin in Q2 2012
- All modules completed and integrated into the pipelines by end of Q1☐
- Deposition Interface – External user testing to begin early January 2012

Annotation Team

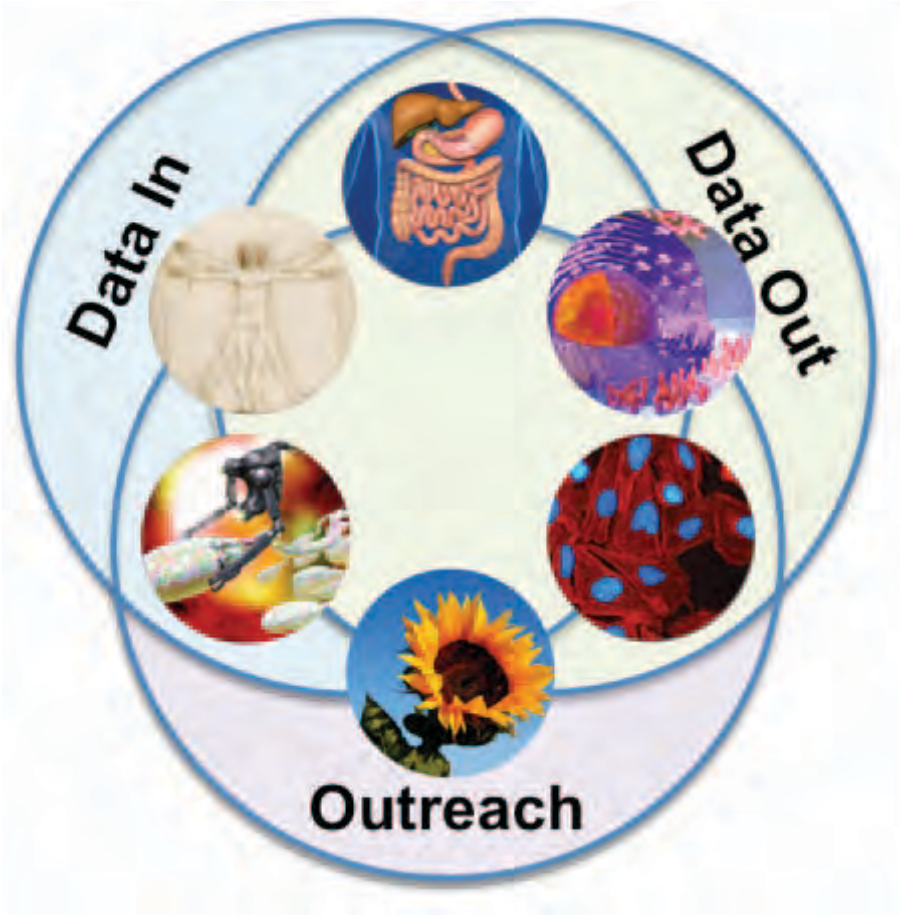
Data Out: Website & Impact

Peter Rose and Phil Bourne

RCSB PDB
PROTEIN DATA BANK


Pharmaceutical Sciences

Integration of RCSB PDB Activities



A New Query Interface

Why?

- Users wanted accurate results and not just on structures
- To fully utilize the remediated data

What did users get?

- More accurate results for
  - Structures
  - Ligands
  - Sequences
  - Biology
  - Educational resources

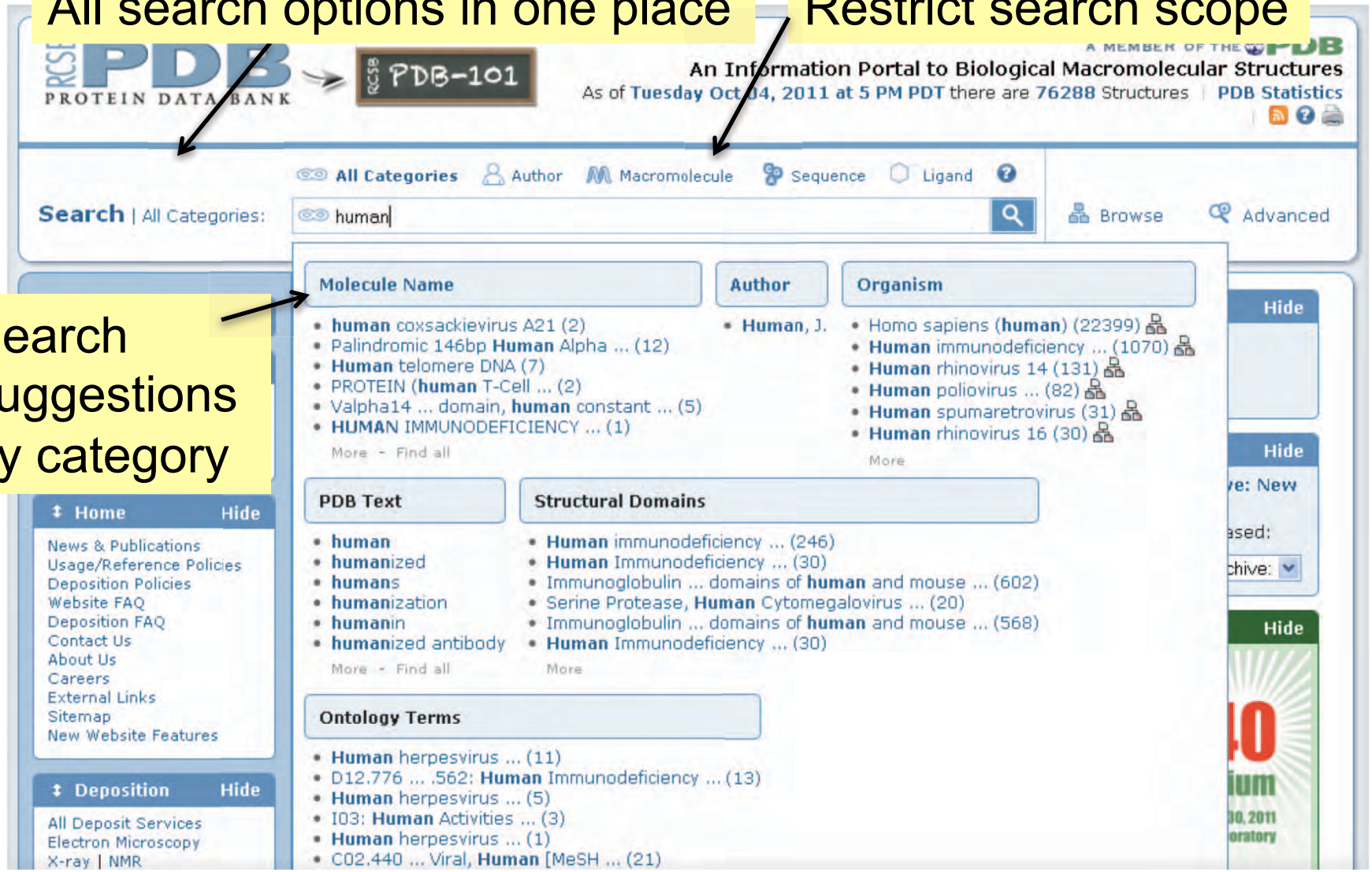
A New Search Interface

What we delivered – *productivity, efficiency, accuracy*

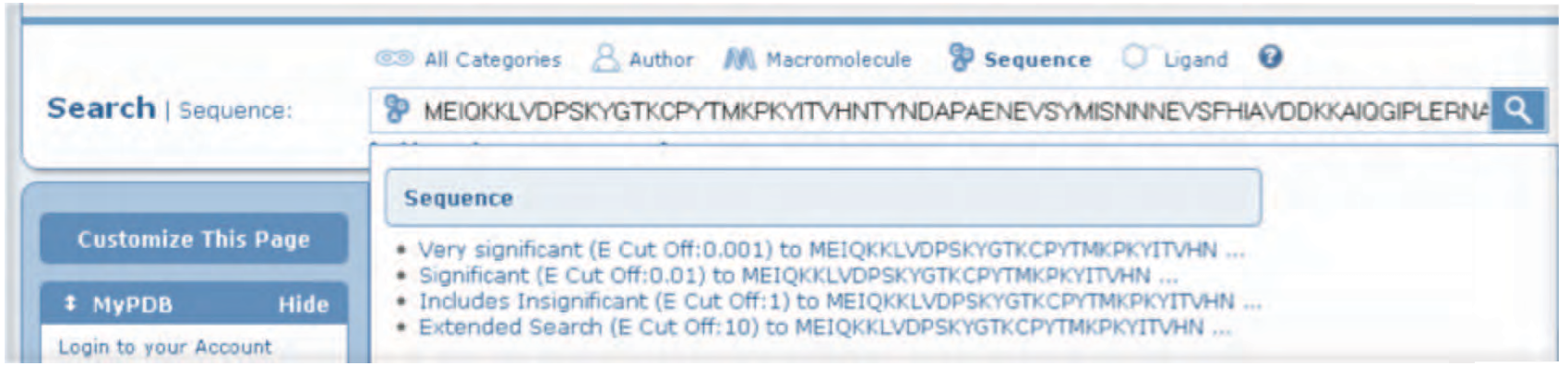
- Through autocomplete we suggest answers as the user types
- Through better ranking we return fewer but more relevant results
- We now return educational material, ligands, sequences *as well as* structures

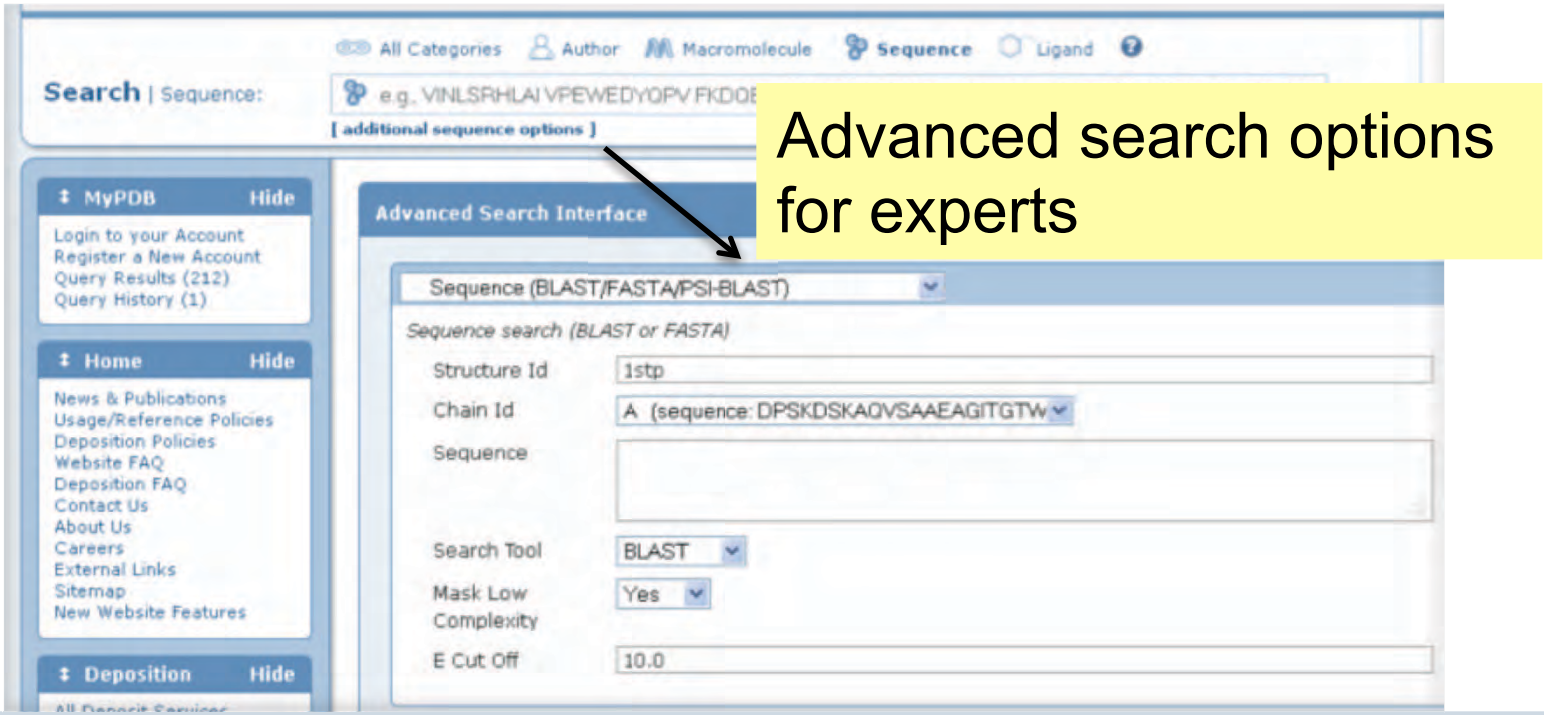
DEMO OF SEARCH FEATURES

New Layout and Autocomplete



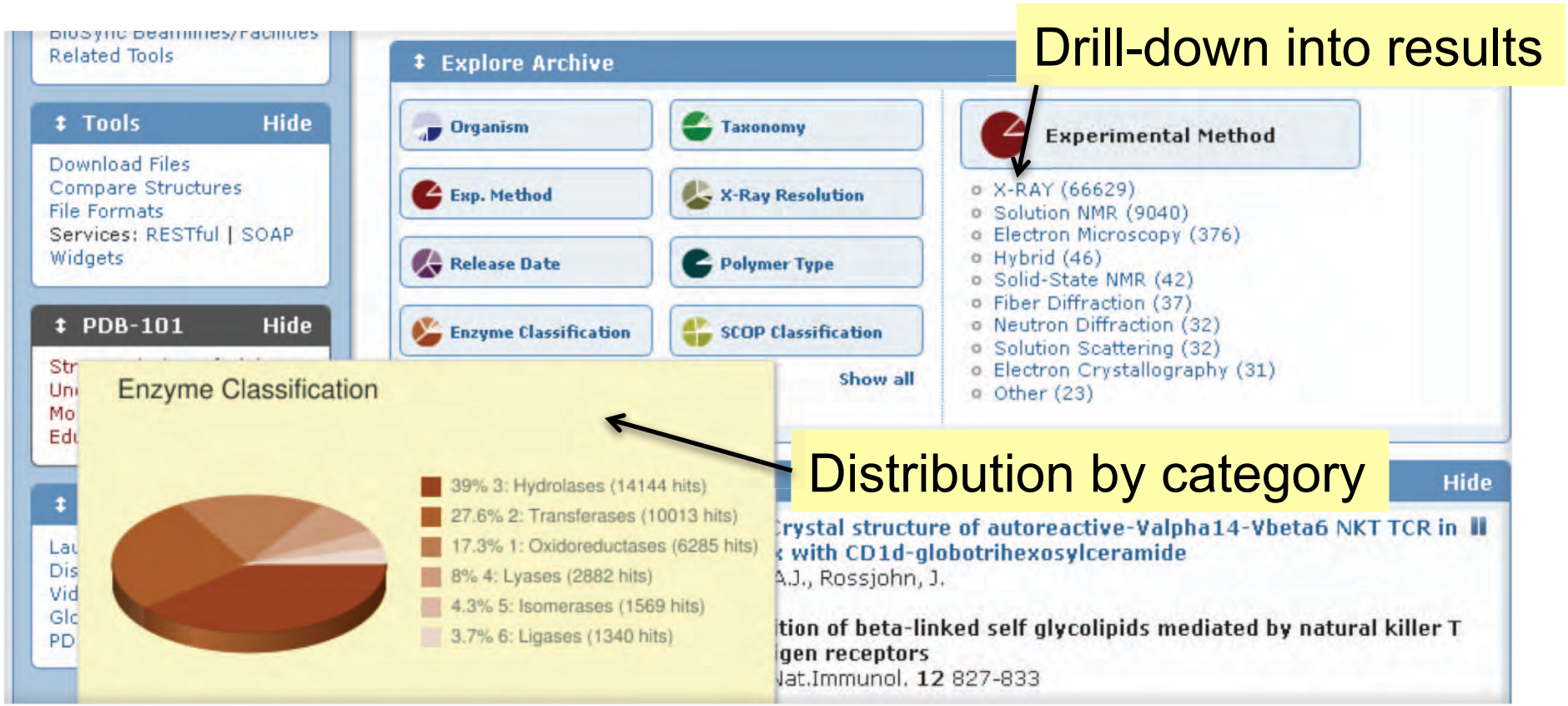
Sequence Search: Simple & Advanced





Exploring the Archive

- Some frequently asked questions can be answered by browsing – no search required



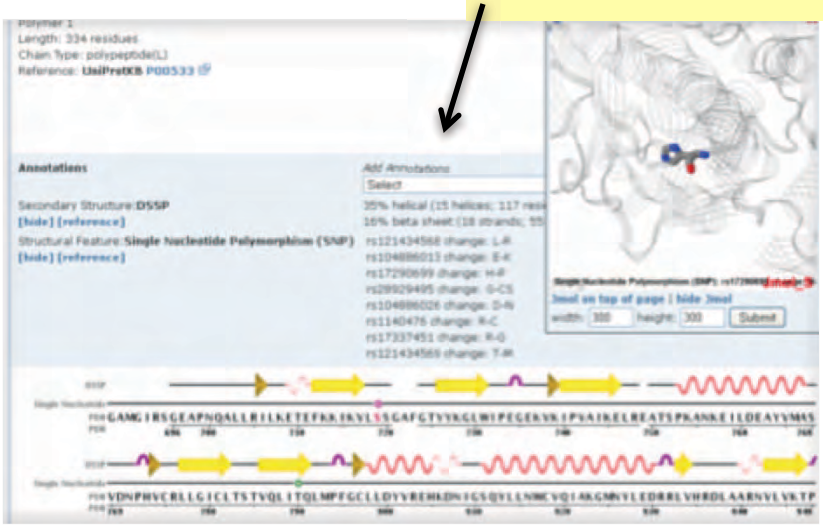
Visualization, Analysis & Data Reporting

Objectives

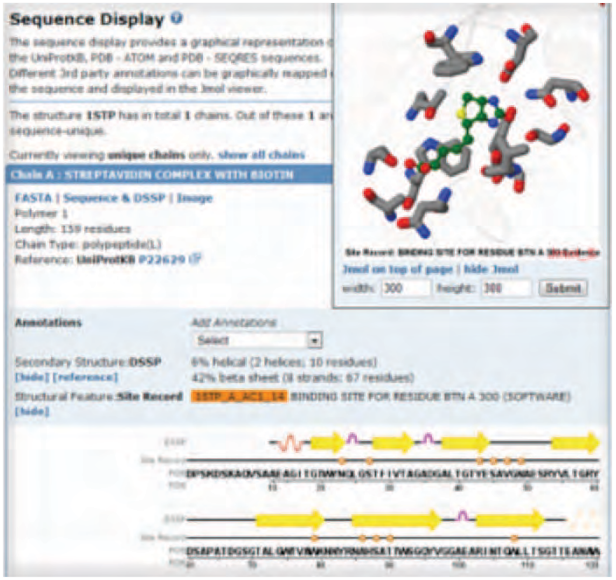
- Visualize sequence annotations mapped onto sequence and structure
- Provide simple visualization options including capabilities to generate publication ready images
- Extend structural alignment algorithms from the chain to the domain level
- Add further report capabilities to tabular reports and provide programmatic access

Mapping Annotations onto Sequence and Structure

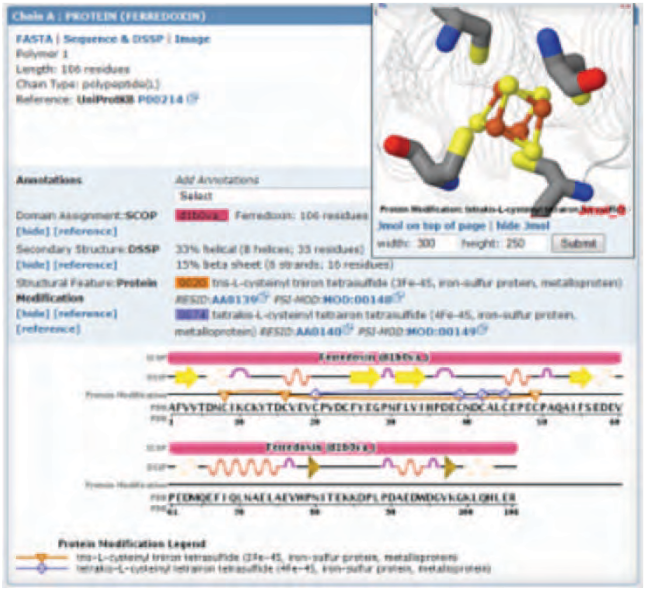
SNPs



Binding sites

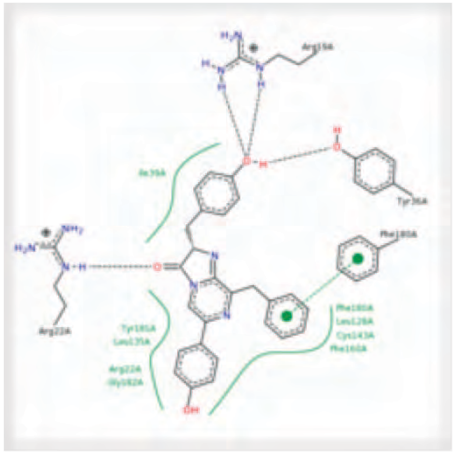


Protein modifi-cations

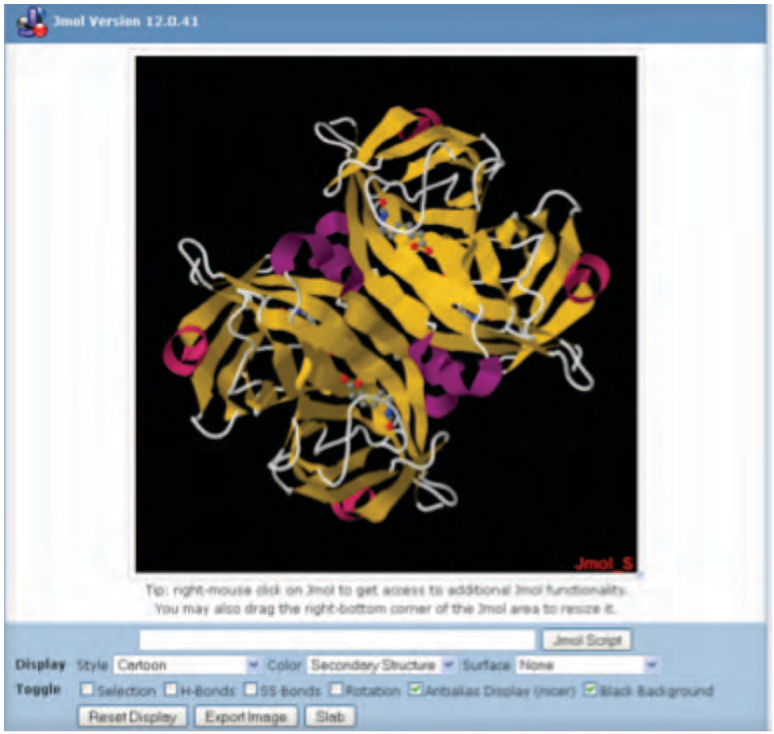


Structure Visualization

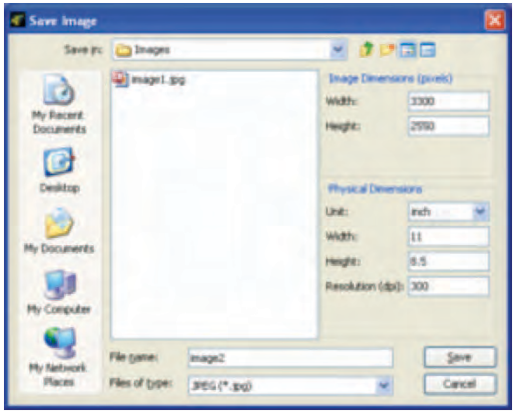
2D Macromolecule-ligand interaction diagrams*



Improved Jmol options panel



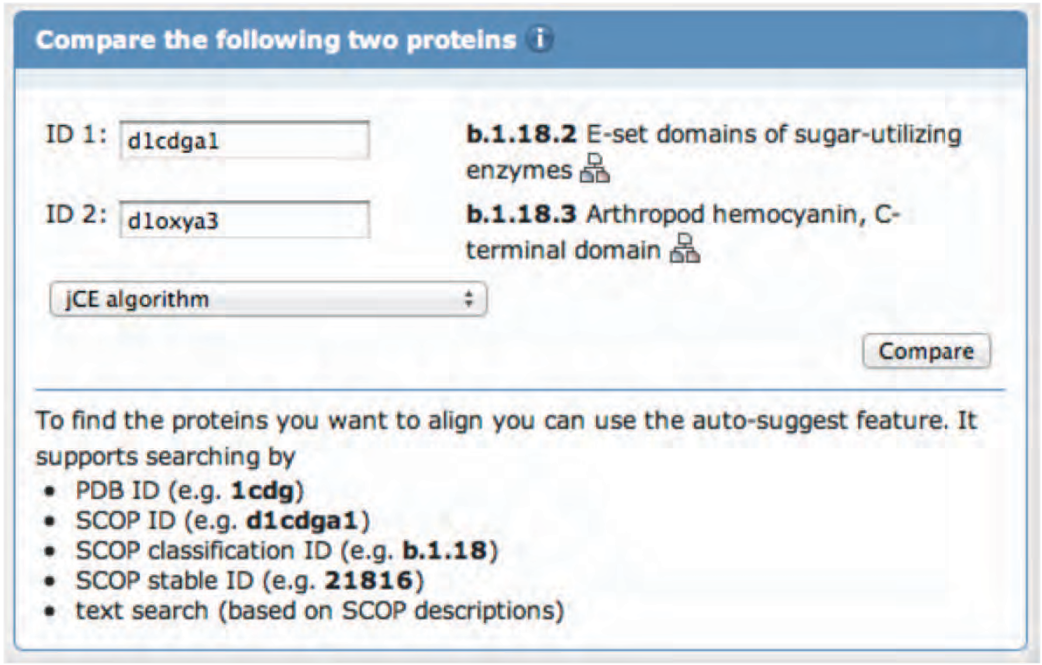
High-resolution image generation



* Stierand, K., Rarey, M. (2010) Drawing the PDB: Protein-ligand complexes in two dimensions. ACS Med. Chem. Lett., DOI: 10.1021/ml100164p.

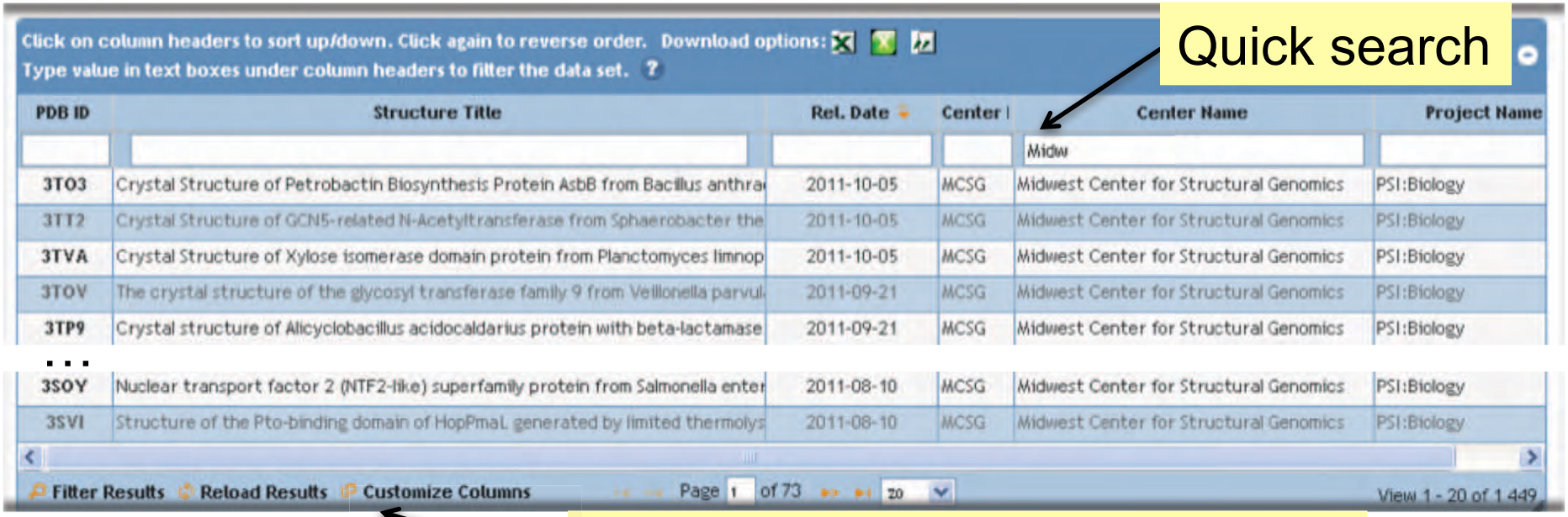
Structural Alignments

- Align two SCOP domains or a SCOP domain against a chain



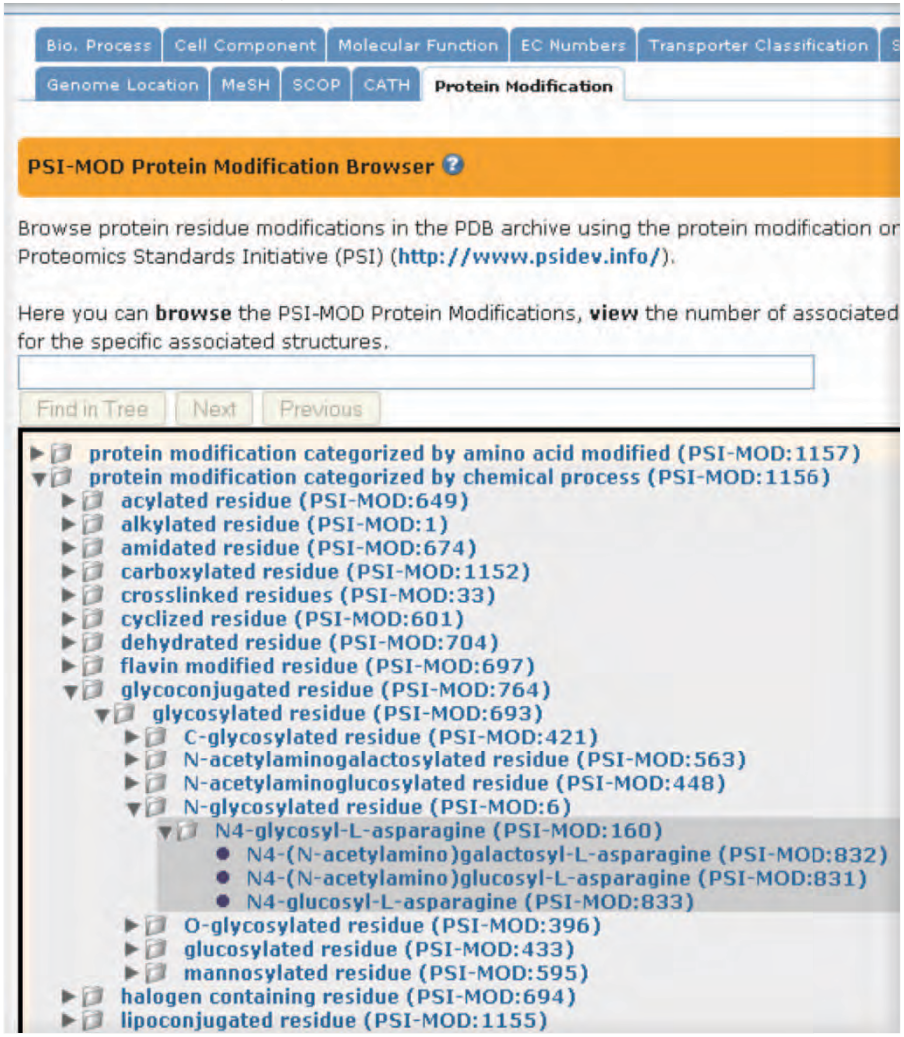
Tabular Reports – A Query Spreadsheet

- Accessible as web service
  - Programmatic download of custom tables
- New reports
  - Binding affinity
  - EM summary
  - Structural Genomics Centers



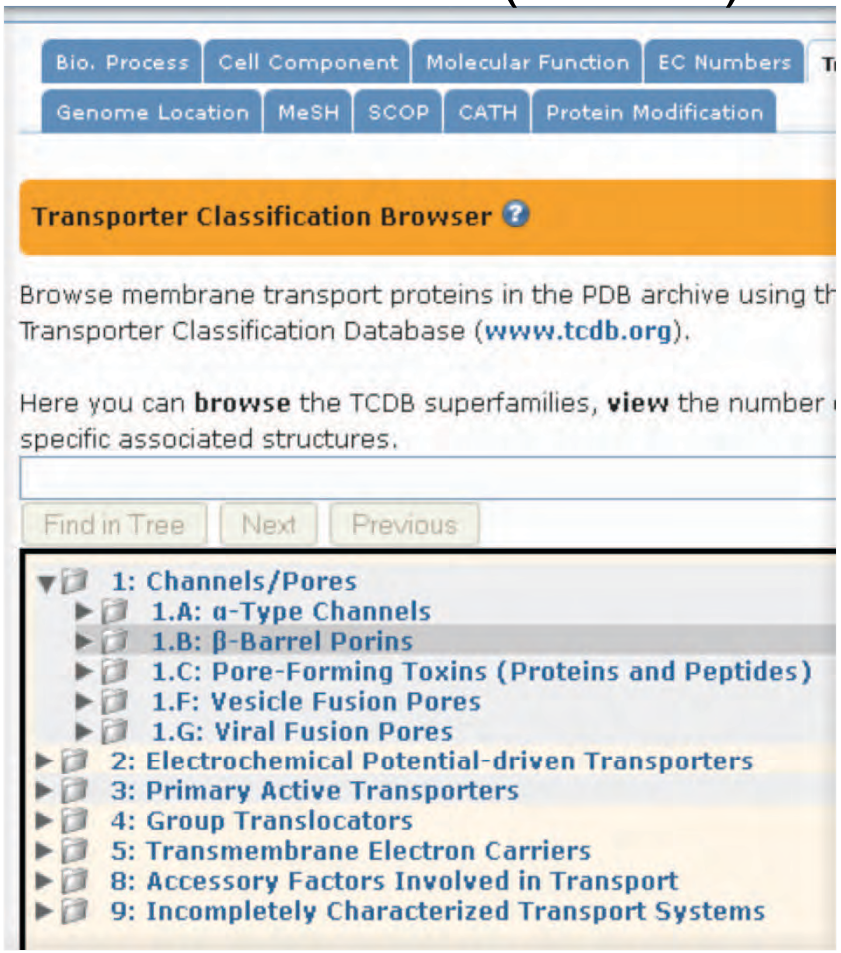
Integration with other Resources

Browse by Protein Modification



Montecchi-Palazzi, L., et al. (2008) The PSI-MOD community standard for representation of protein modification data. *Nature Biotechnology* 26, 864-6.

Browse by Transporter Classification (IUBMB)

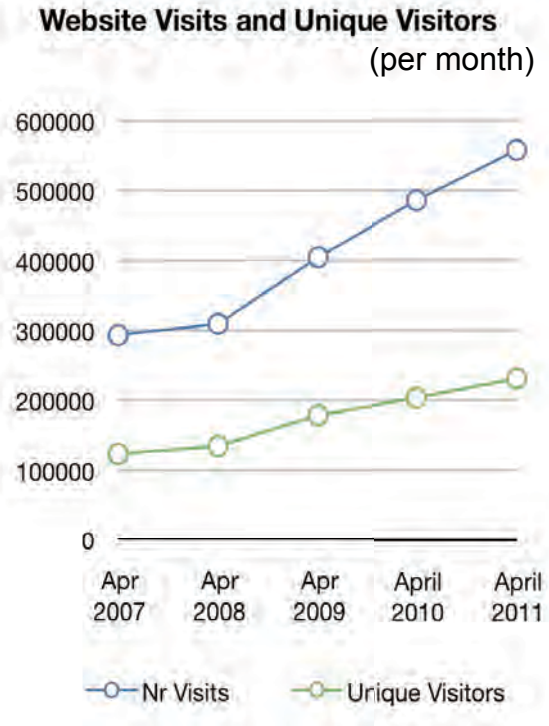


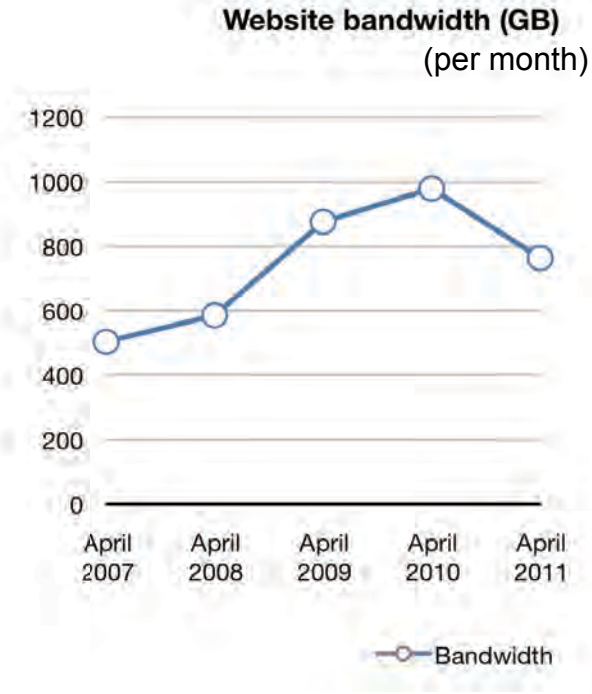
Saier M.H. Jr, et al. (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Res*. 37, D274-8.

RCSB PDB Website Usage

Number of visitors is growing linearly.

- Bandwidth optimization – delivering web pages more quickly

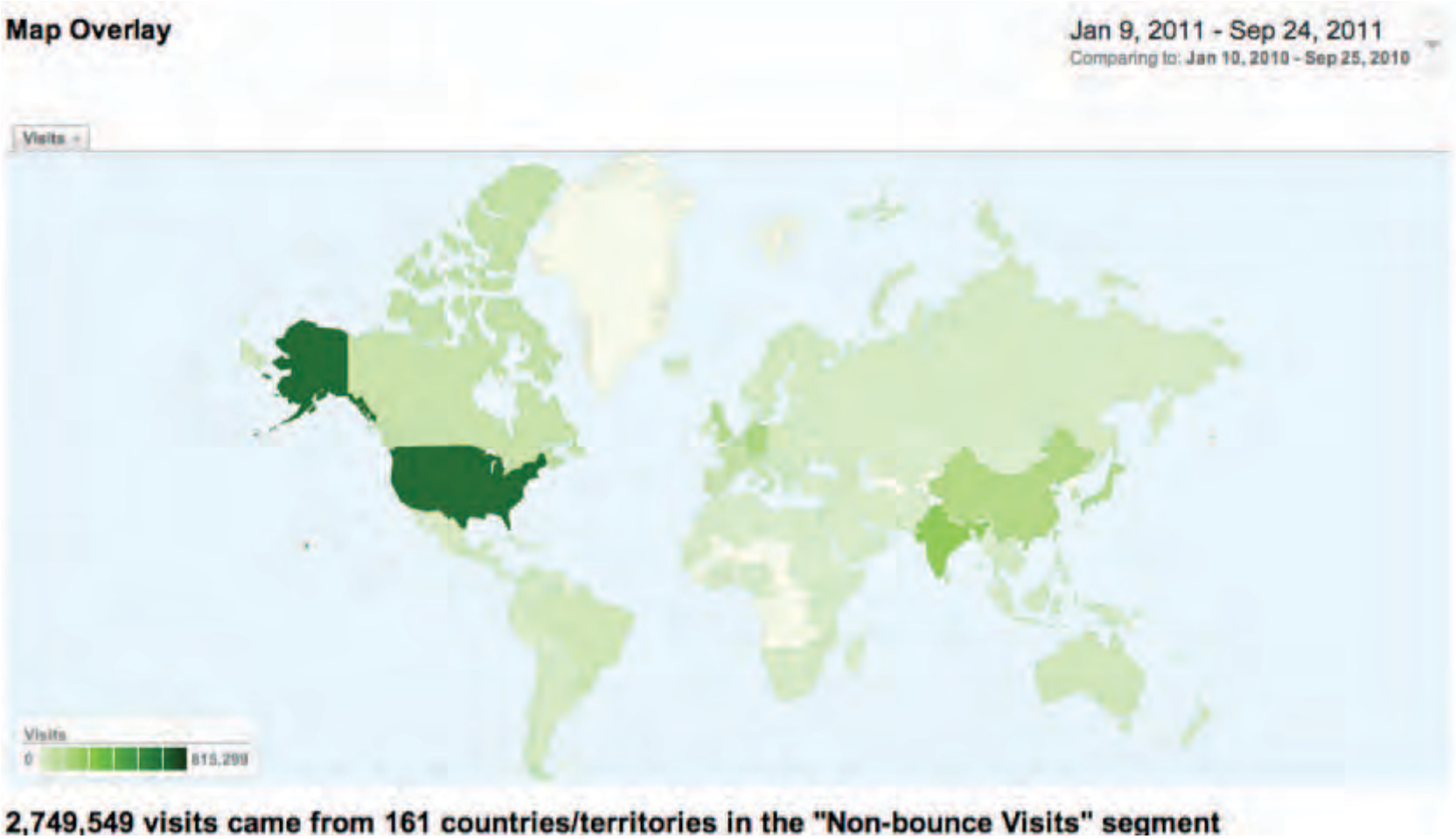




Statistics by AWStats

Website Usage by Country

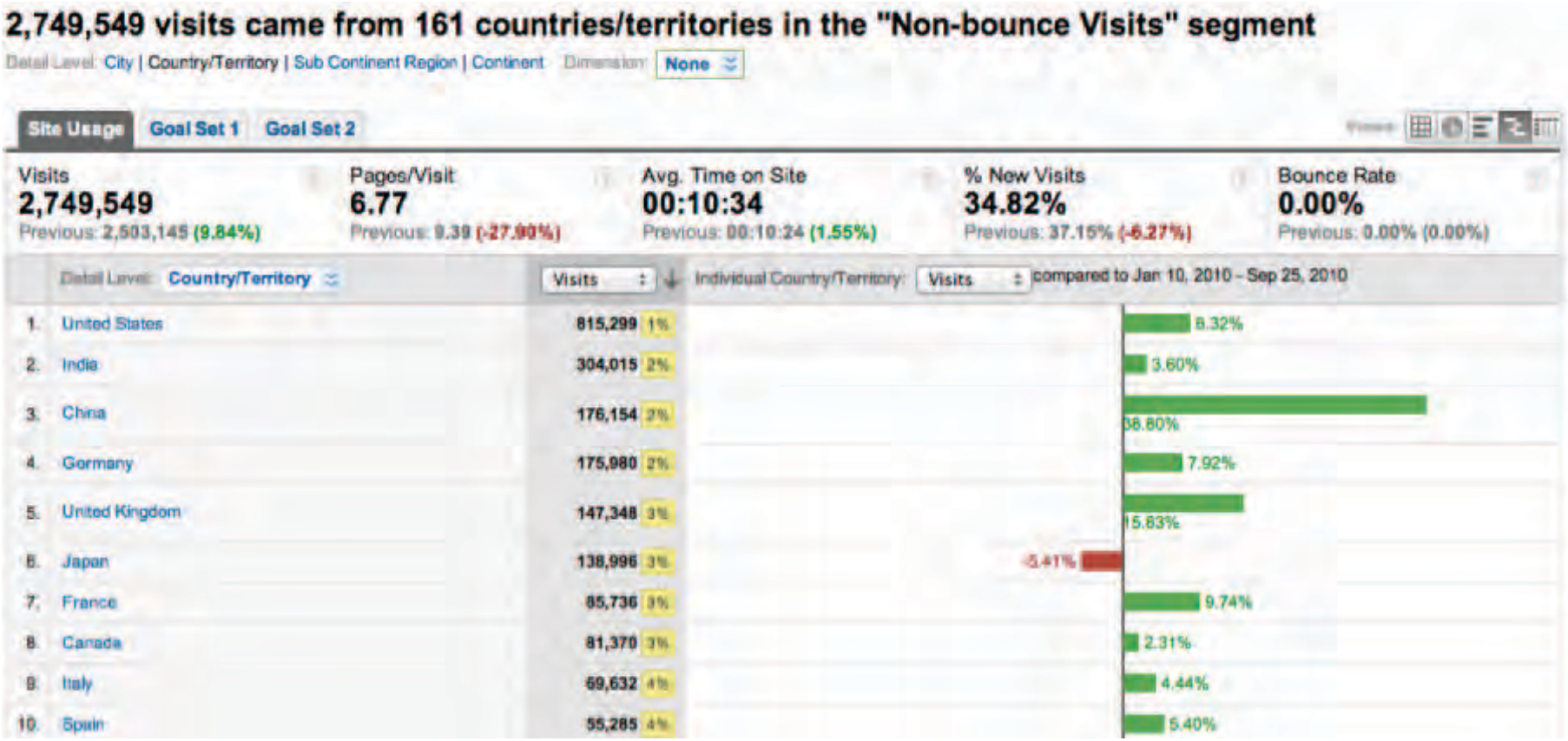
- US still number 1, followed by India, China



Statistics by Google Analytics

Where Did the Growth Come From?

- Visits increased by about 10% in last year
- China has highest growth rate

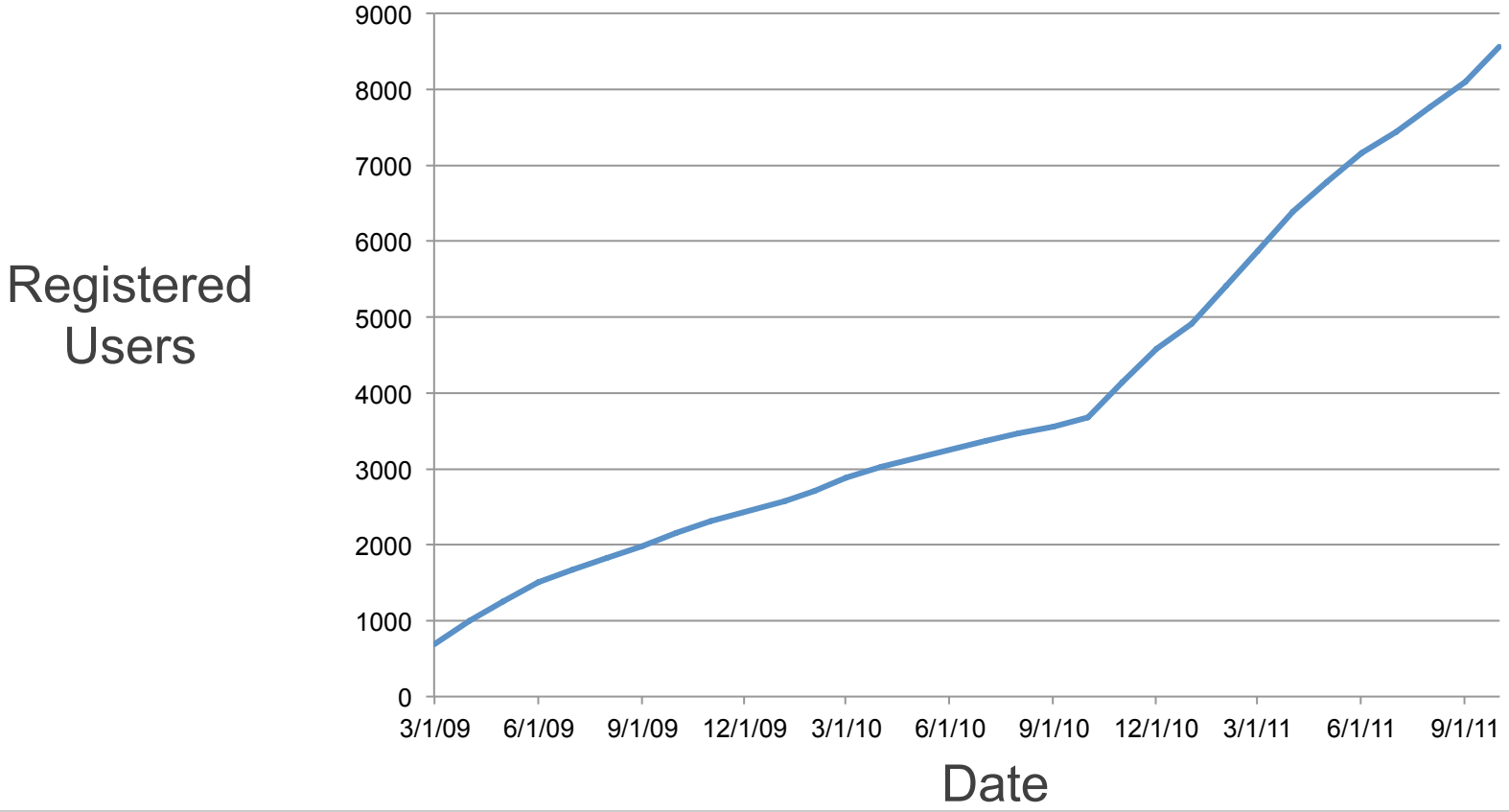


Statistics by Google Analytics

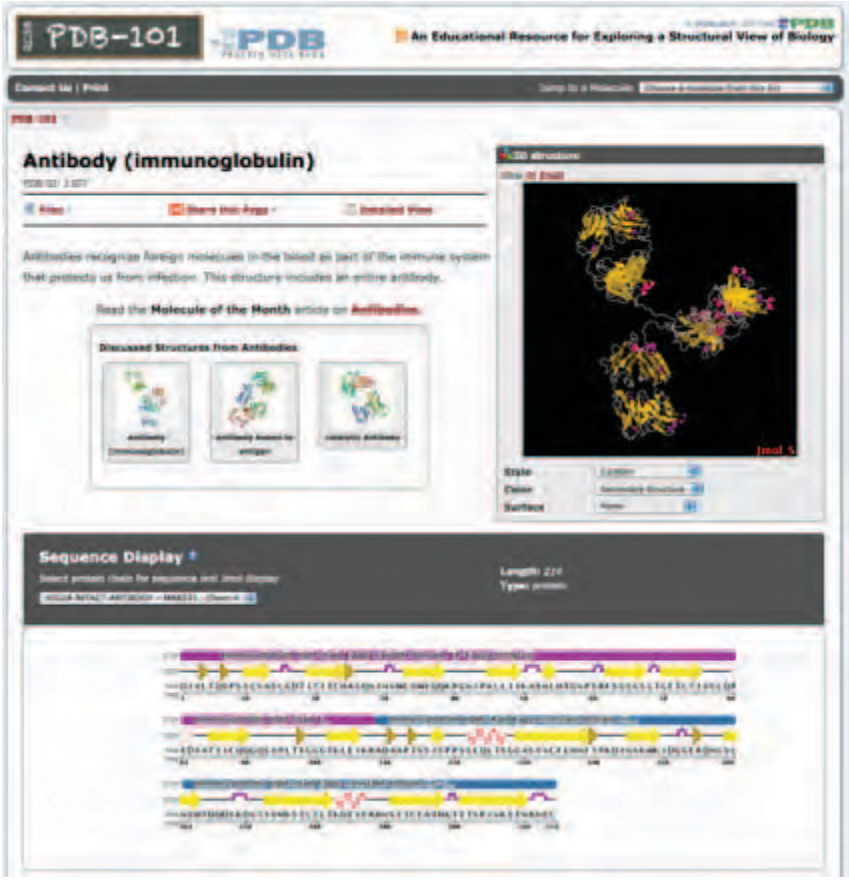
Growth of MyPDB

User base doubled over a year – Why?

- Useful to store queries and annotate results
- MyPDB located more prominently on site
- News items and flyers



PDB-101 Developed & Deployed

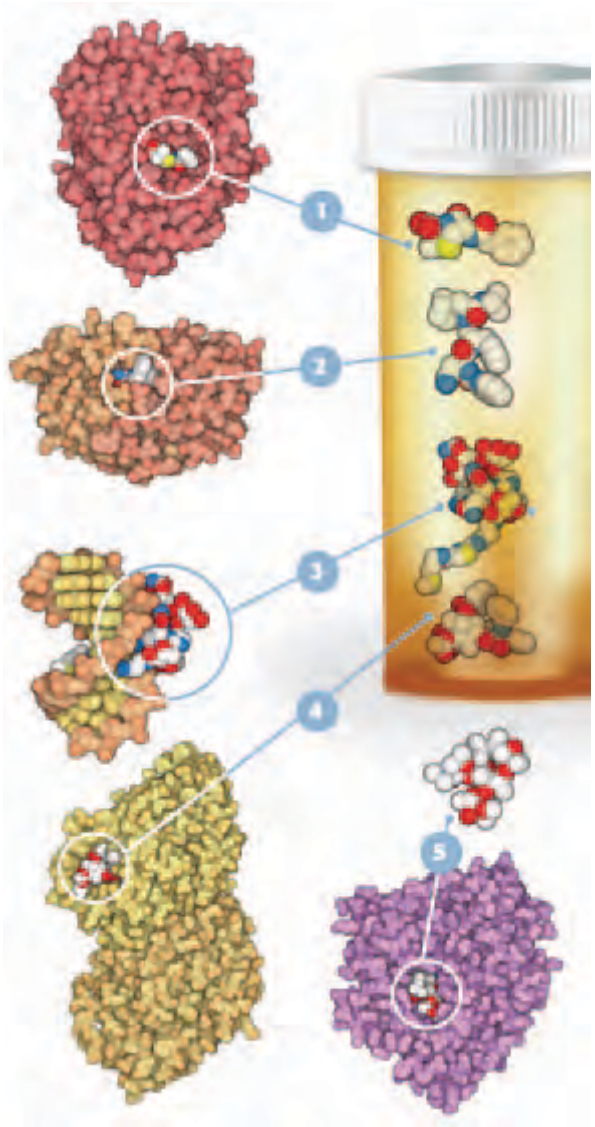


To be discussed in Outreach

Future Plans

- Drug view
- Adaption of parts of the website for mobile devices (initial focus PDB-101)
- Structural alignment database based on domains
- Further development of web services

Drug View: Scope



- Annotated set of drugs and drug targets including off-targets and metabolic enzymes
- Outreach materials about structural aspects of drug-receptor interactions and structure-based drug design

Drug View: Priorities

- Small molecule drugs, nutraceuticals, and their targets (DrugBank)
- Peptide derived compounds
- Human metabolites (HMDB), toxins and toxin targets (T3DB)
- Biotherapeutics, i.e., monoclonal antibodies
- Veterinary drugs (FDA Green Book)

Education & Outreach

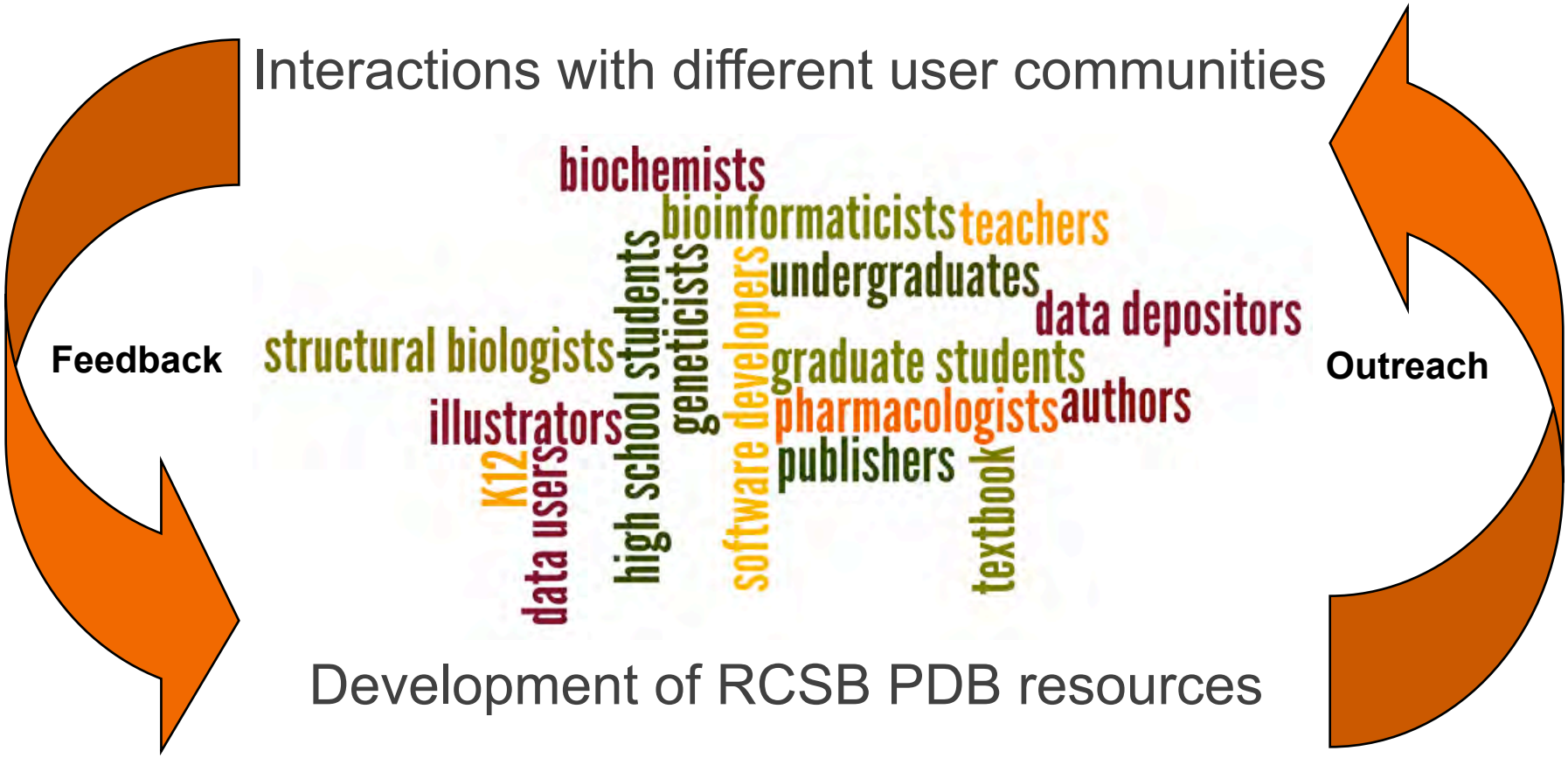
Shuchismita Dutta

RCSB PDB
PROTEIN DATA BANK

Goals

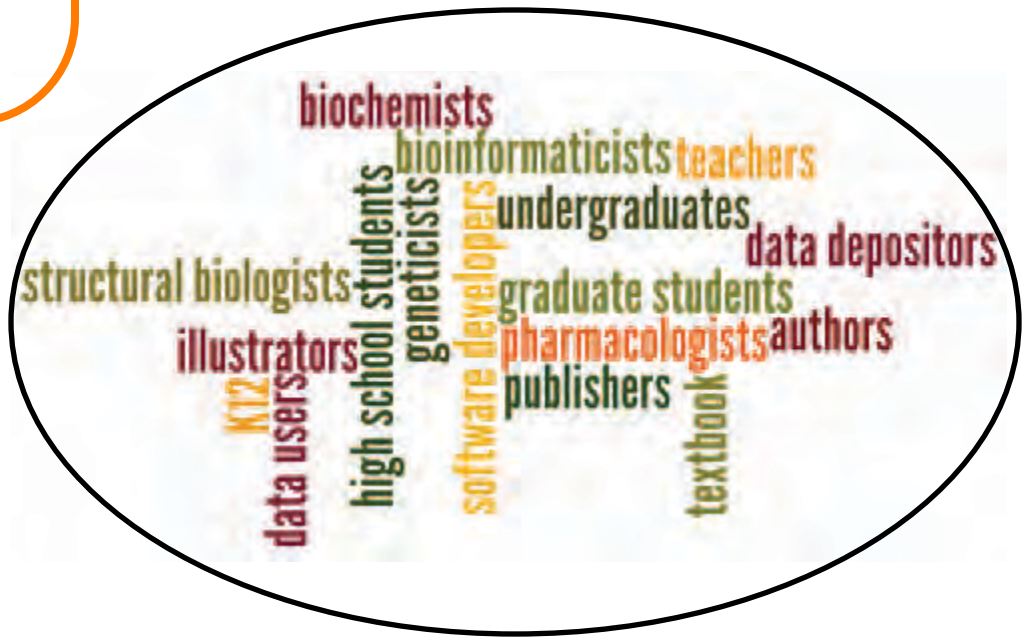
- RCSB PDB resource should meet its mission in the interest of science, medicine and education
- RCSB PDB is defined by, designed for, and owned by the communities it serves

The Outreach Cycle



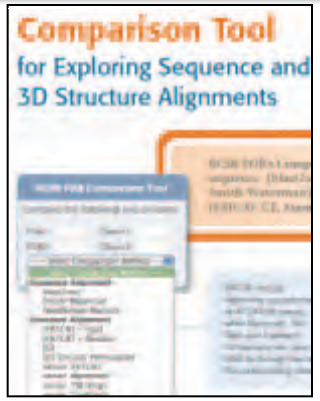
International User Communities

1. **Biologists**
   - Structural Biology
   - Biochemistry
   - Genetics
   - Pharmacology
2. **Other scientists**
   - Bioinformatics
   - Software developers
3. **Students** and **Educators**
   - K-12
   - Undergraduate
   - Graduate
4. **Media**
   - Writers
   - Illustrators
   - Textbook authors
5. **General public**



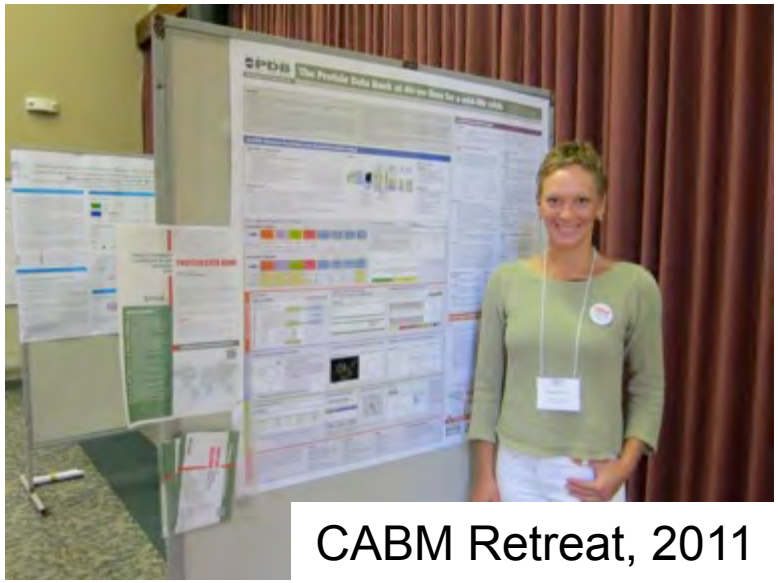
Tell them, tell them again

- International scientific meetings and workshops
- Electronic news, RSS feeds, support pages, tutorials, listserv
- Printed and online publications (annual report, newsletter, flyers, brochures)



Community Interactions: Feedback

- Electronic help desks, discussion groups
- Demonstrations/ presentations at professional meetings
- Personal interactions
- Exhibit booths
- Interactions with Journals
- Workshops, Posters
- Surveys



CABM Retreat, 2011

Experimental Biology, 2011

Community Interactions: Collaborations

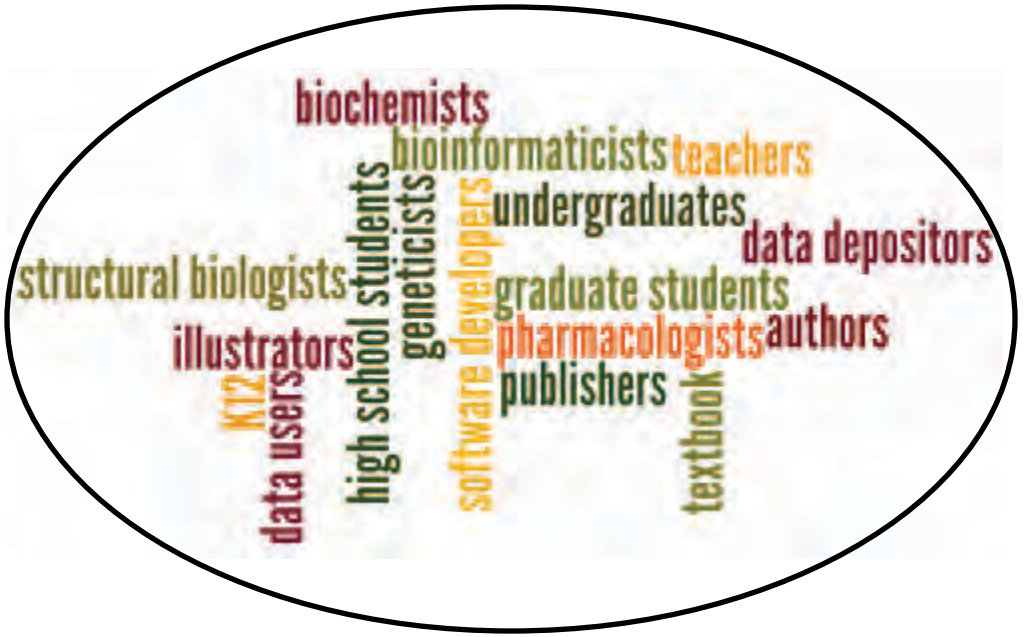
- Task forces to establish requirements and set standards
- Journal interactions
- New PDB Format meeting

Cambridge, Sep 2011

EM VTF, Sep 2010

International User Communities

1. **Biologists**
   - Structural Biology
   - Biochemistry
   - Genetics
   - Pharmacology
2. **Other scientists**
   - Bioinformatics
   - Software developers
3. **Students** and **Educators**
   - K-12
   - Undergraduate
   - Graduate
4. **Media**
   - Writers
   - Illustrators
   - Textbook authors
5. **General public**



Structural View of Biology: for All

- Presentations, demonstrations and booths at new society meetings
- Outreach events

AAAS, Feb 2011

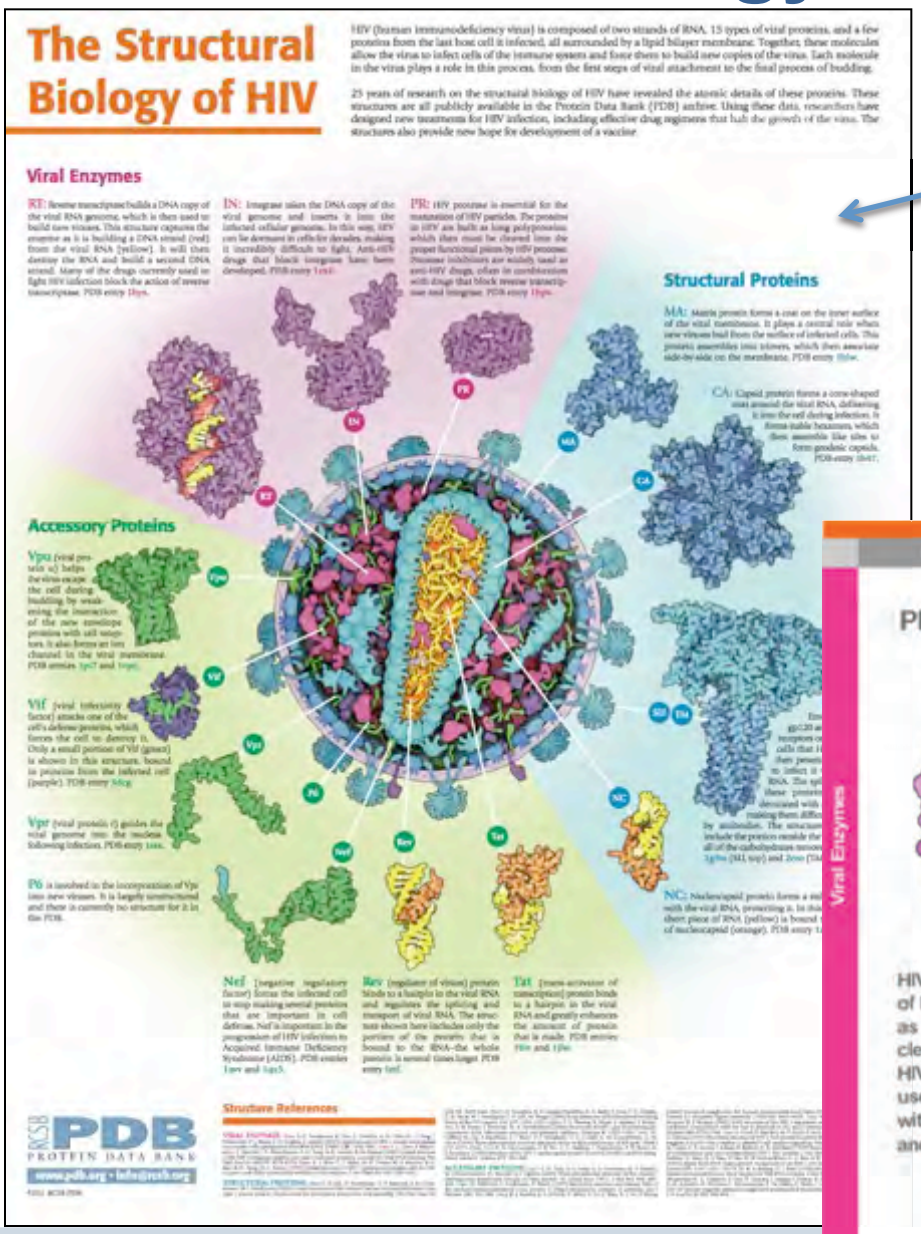
Apr 2011

RUTGERS DAY
Experience New Jersey's State University

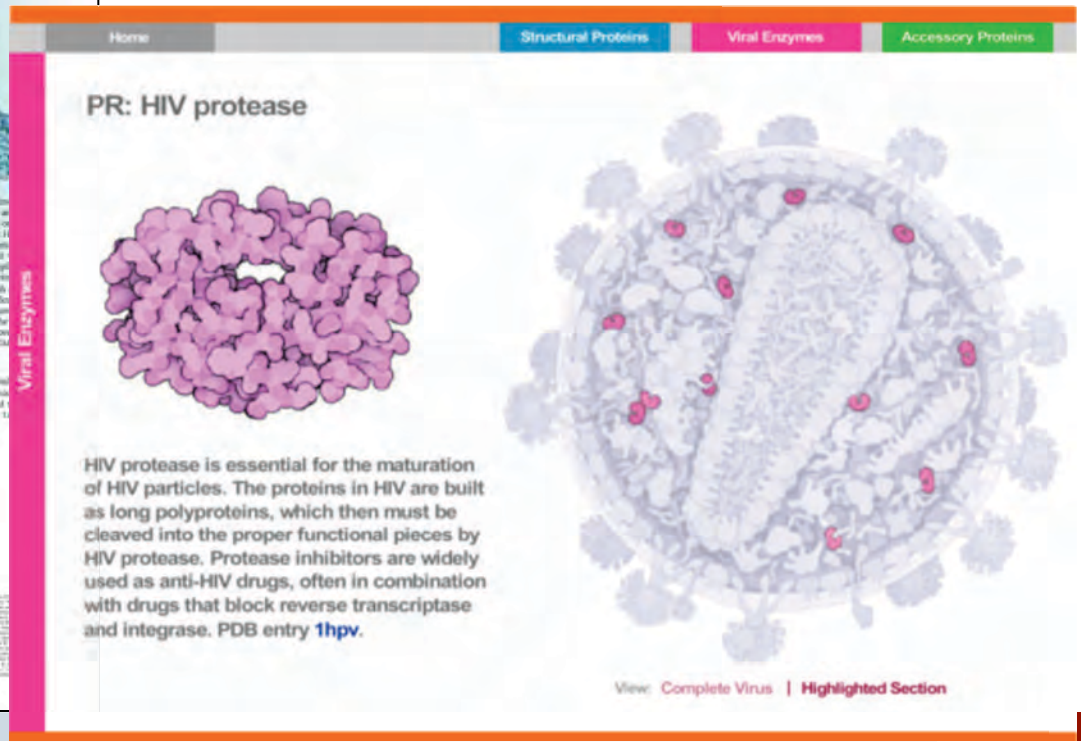
San Diego Festival of Science & Engineering
excite your mind

Mar 2011

Structural Biology of HIV

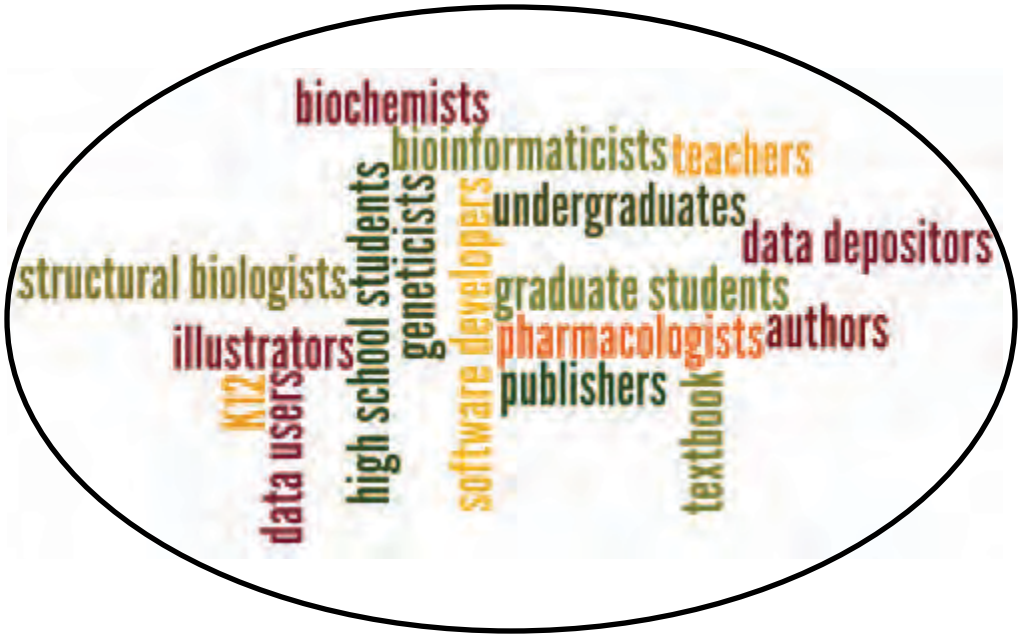


- Printed poster showing structures of all components of HIV
- Online interactive views



International User Communities

1. **Biologists**
   - Structural Biology
   - Biochemistry
   - Genetics
   - Pharmacology
2. **Other scientists**
   - Bioinformatics
   - Software developers
3. **Students** and **Educators**
   - K-12
   - Undergraduate
   - Graduate
4. **Media**
   - Writers
   - Illustrators
   - Textbook authors
5. **General public**



Teaching and Learning: A Structural View of Biology

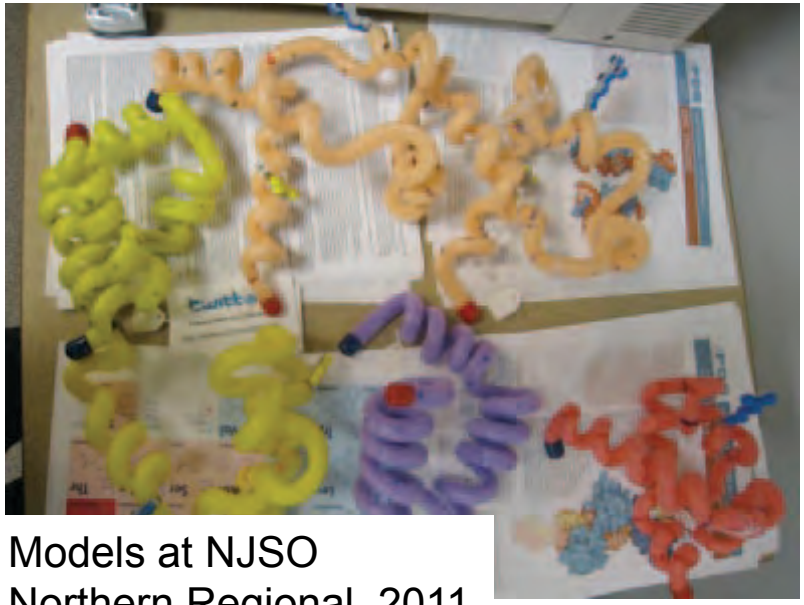
- Presentations, booths, demonstrations at Society meetings

- Online resources
  - *Molecule of the Month*
  - PDB-101
  - Education Corner
  - Lesson plans/Activities
  - MAP

NJSC, Oct 2011

Protein Modeling at the Science Olympiad

- HS students
- Annotators & software developers involved
- 2012: NJ and CA



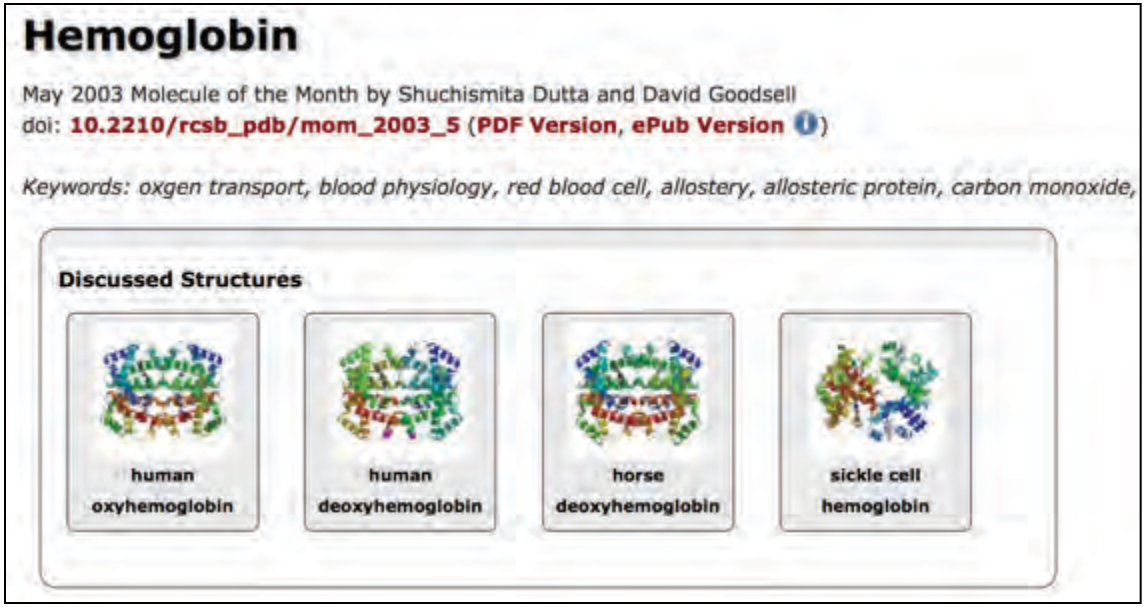
Models at NJSO Northern Regional, 2011

NJSO Central Regional, 2011

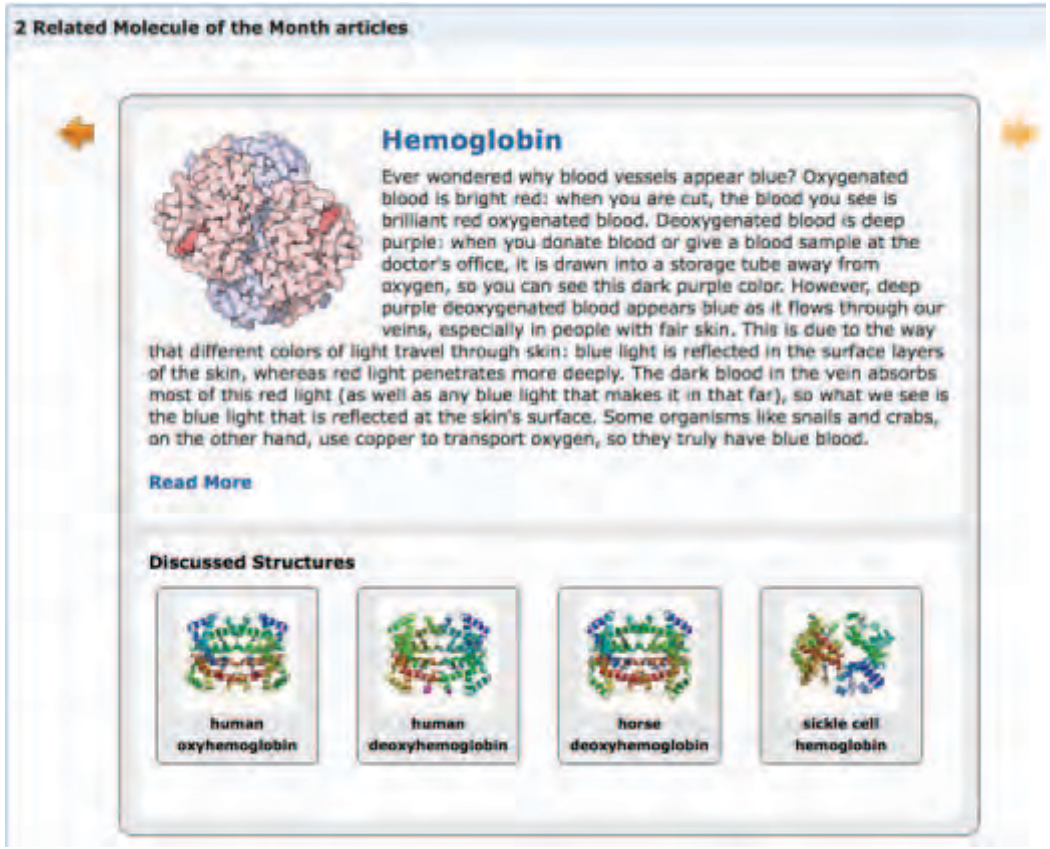
Molecule of the Month (MoM)

- Discussed structures and summaries added
- Keywords included
- Older, multipage features converted to single page
- Interactive Jmol displays for newer MoM features

- Topics for further explorations in newer MoM features



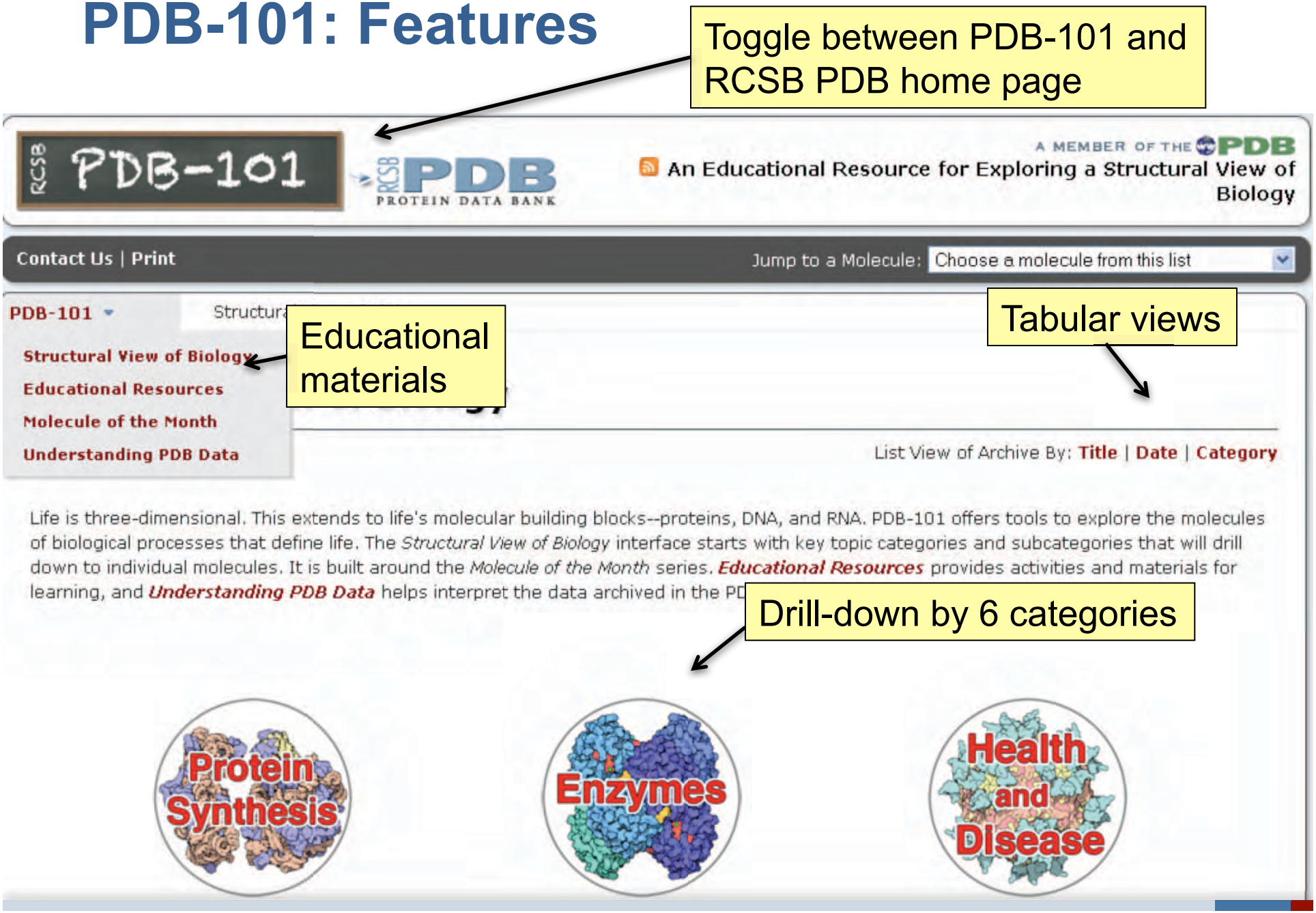
Molecule of the Month (MoM)



- If search results contain structures described in MoMs, a short description is displayed at top of results list
- An intuitive way to guide users to MoM content

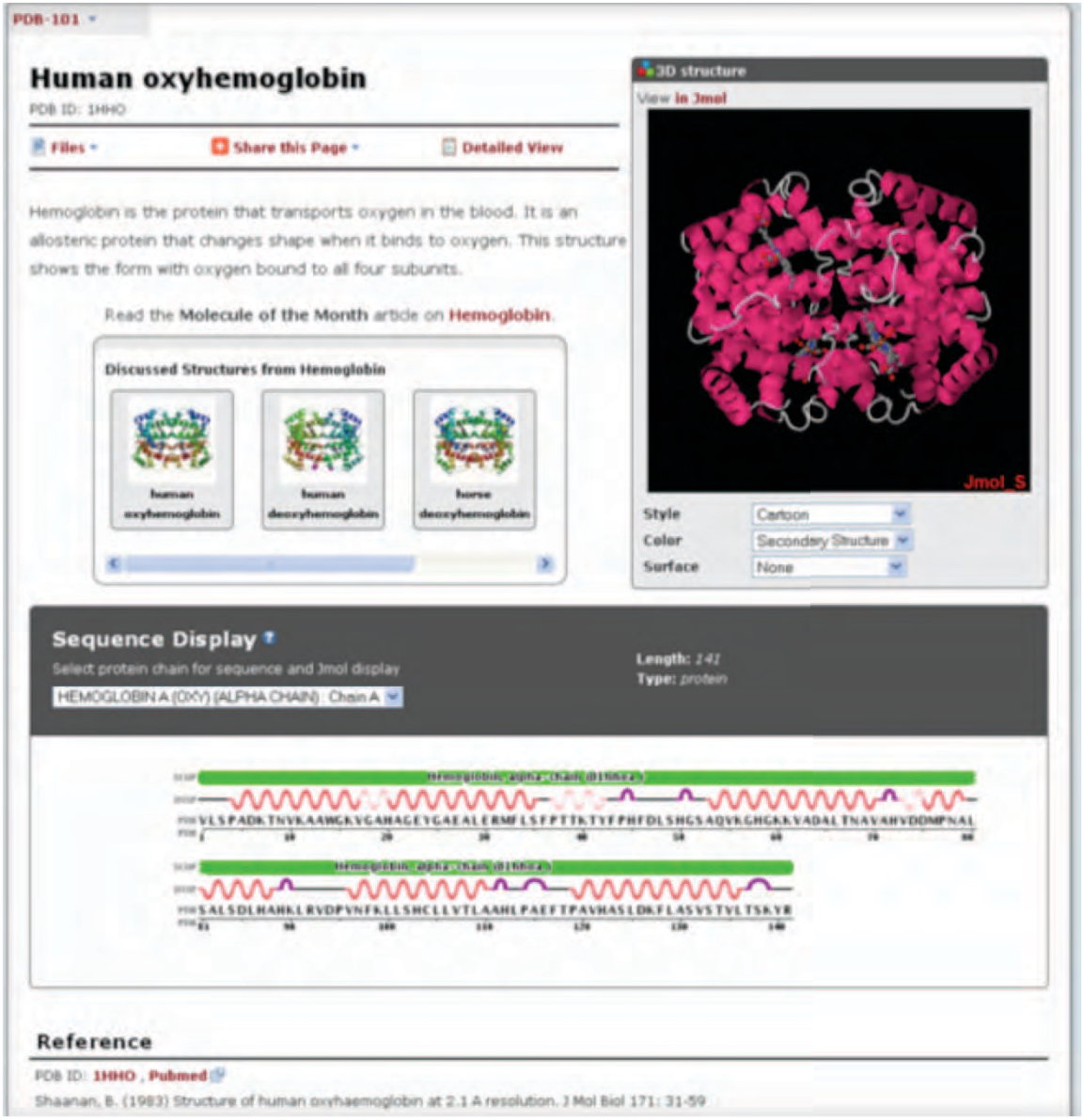
- Motivation
  - Structural View of Biology
  - Easy access to PDB data
- Process
  - Group MoM features into 6 broad categories
  - Include description of each category/sub-category
  - Identify and describe discussed structures
  - Simplified Structure Summary page
- Status: Implemented and deployed

PDB-101: Features



PDB-101: Summary View

- Simplified structure summary page
  - Sequence
  - Structure
  - Function
- Intended for educational audience
- Compatible with mobile devices, image sprites for structure visualization



PDB-101: Feedback

National Science Teachers Association Group on LinkedIn

J. Ladwig, Project-Based STEM Outreach Facilitator Seeking Opportunities to Reach Out...
Bloomington, Indiana Area

"Given that I've on occasion gotten the, "Sir, you know you're making this up, right..." look from a puzzled, yet well-meaning student, PDB-101 is an awesome educator help for student enrichment. In a time where many students need extra assistance with the basics, helpers like easy to navigate PDB-101 allow students ahead of the curve to self-engage and to "self-enrich"–skills that must be honed to succeed- while I work with others on understanding at grade level.

After years of discussing mitochondrial ATP Synthase activity in more general terms and with less than incredible imagery, these extreme images and excellent narratives concisely clarify the activities to the atomic and even the sub-atomic level in an excellent one-stop shopping experience. They allow amazing personal discovery moments for Secondary and post-Secondary Ed. students seeking a little extra enrichment and understanding well beyond the standards."

Molecular Anatomy Project (MAP)

MAP
Molecular Anatomy Project

Home | About Map | Help | Contact Us

MAP Resource
- Organ System
- Organ and Tissue
- Molecule Type
- Diseases

MAP Administrator
- View All Depositions

User login

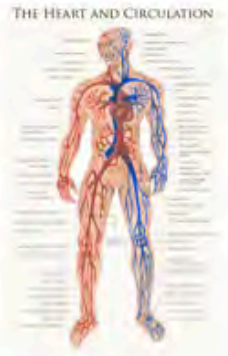
Username: *
admin

Password: *
••••••••••

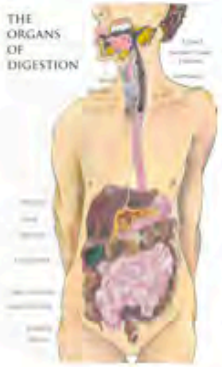
Organ System

Complex multicellular organisms like humans have specific groups of cells, tissues and organs to perform specific functions. These tissues and organs work together in organ systems to perform coordinated biological functions such as digestion or respiration. Browse through the different organ systems to learn about the structure and functions of molecules associated with them.

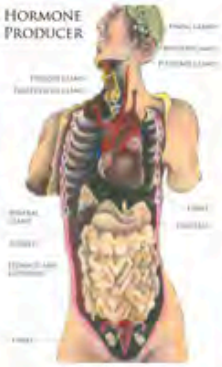
Circulatory system



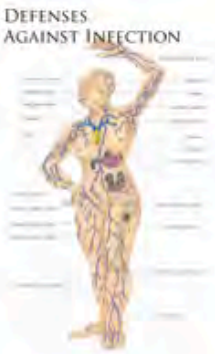
Digestive system



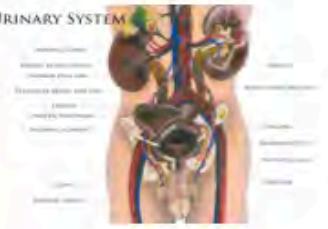
Endocrine system



Immune system



Excretory system



Integumentary system

Musculoskeletal system

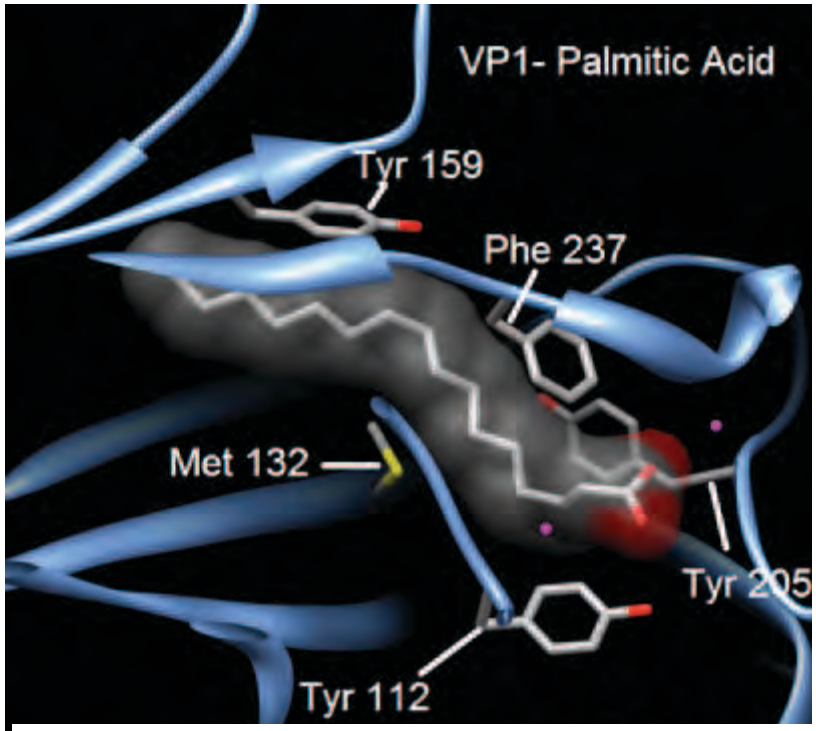
Nervous system

Respiratory system

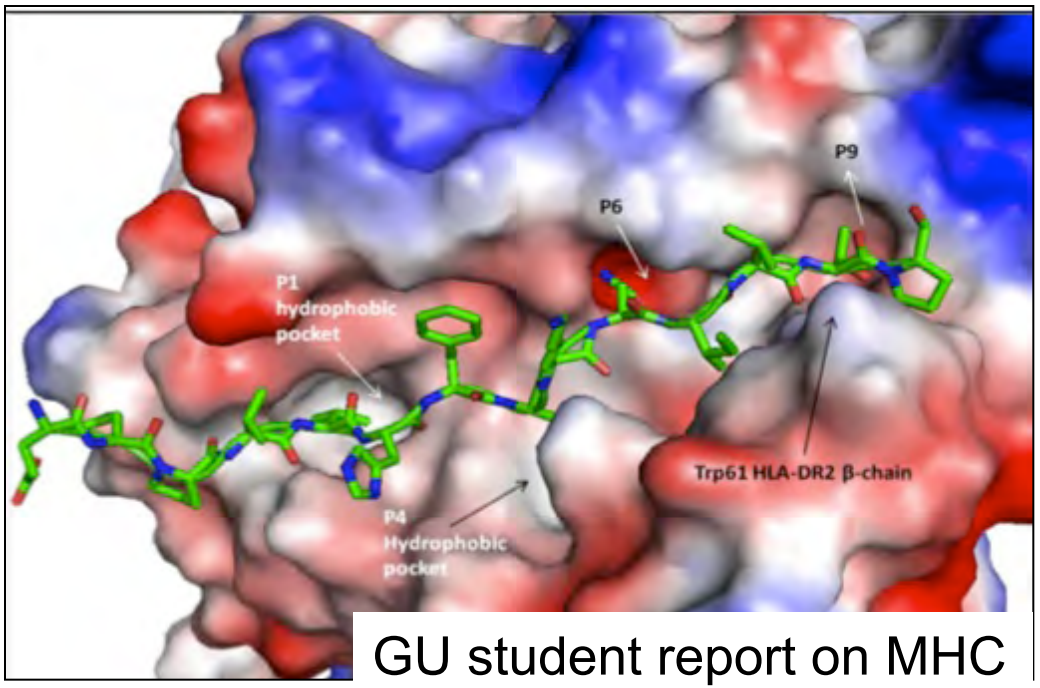
Reproductive system

MAP-based Courses: Spring 2011

- Rutgers University: Viral Infectious Diseases
- Georgetown University: Molecules in the Immune System

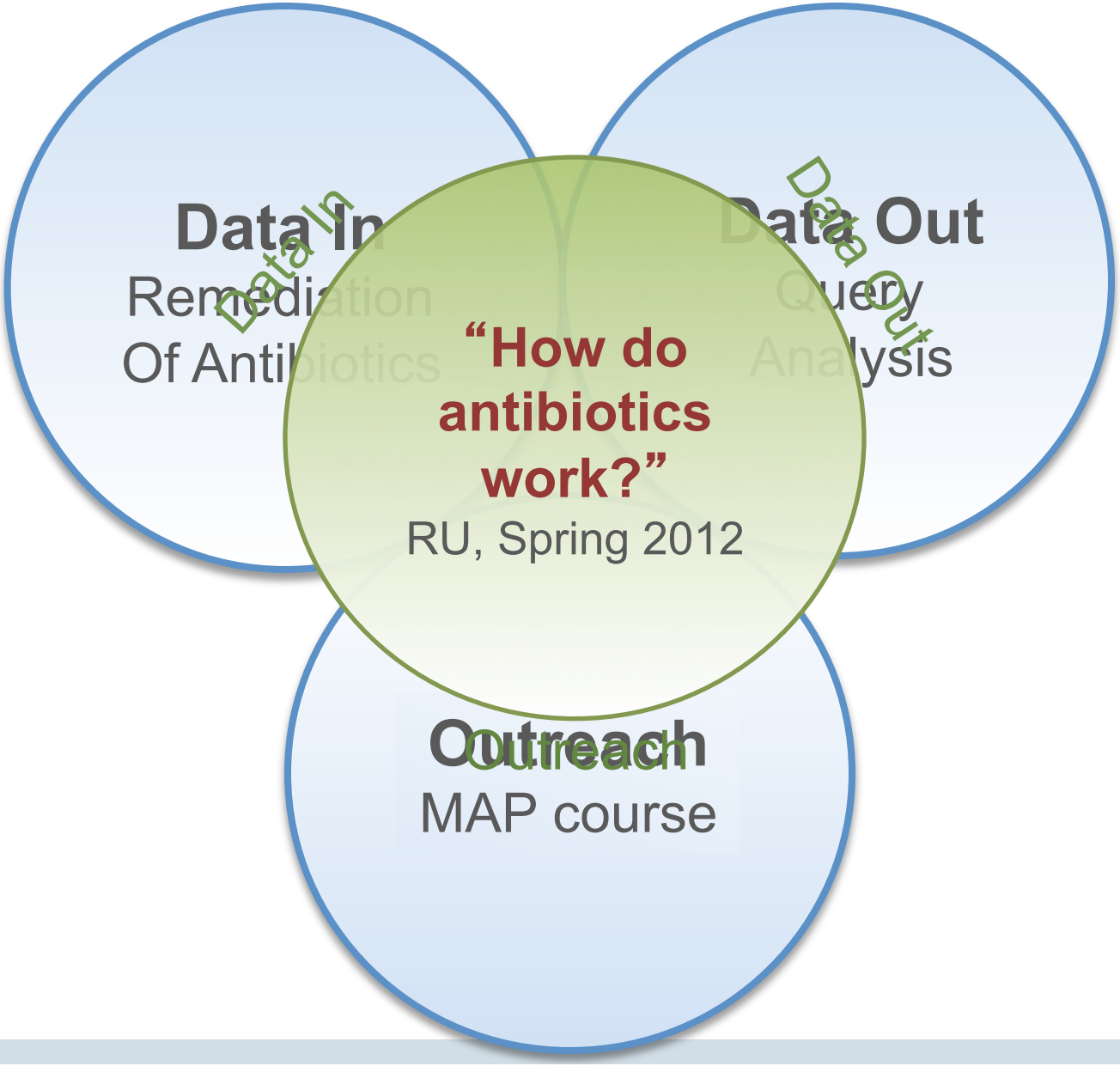


RU student report on Polio virus



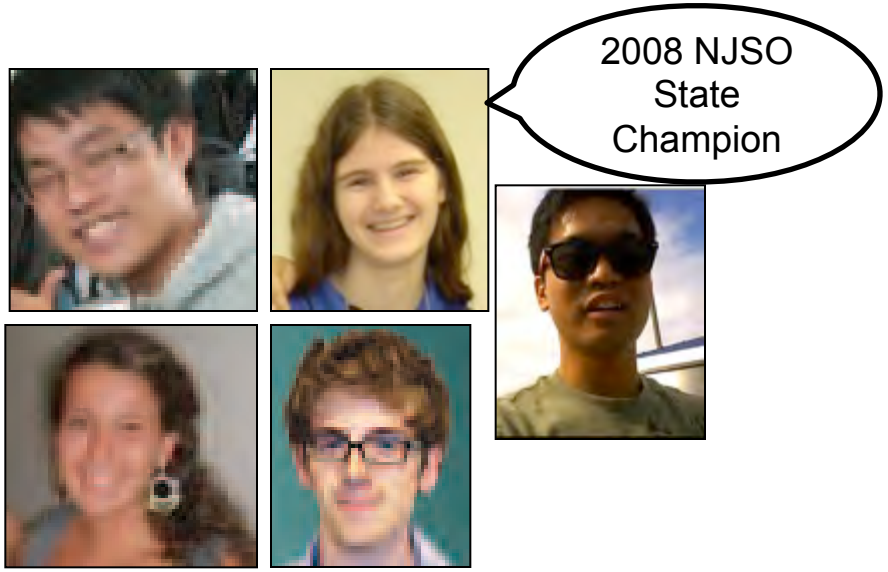
GU student report on MHC

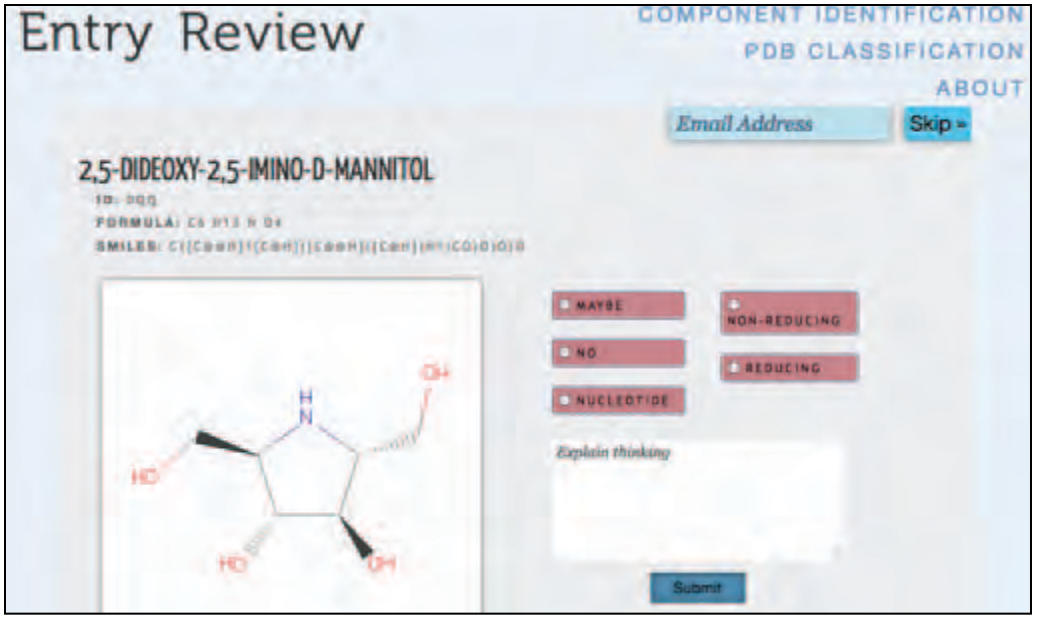
MAP Based Courses: Future



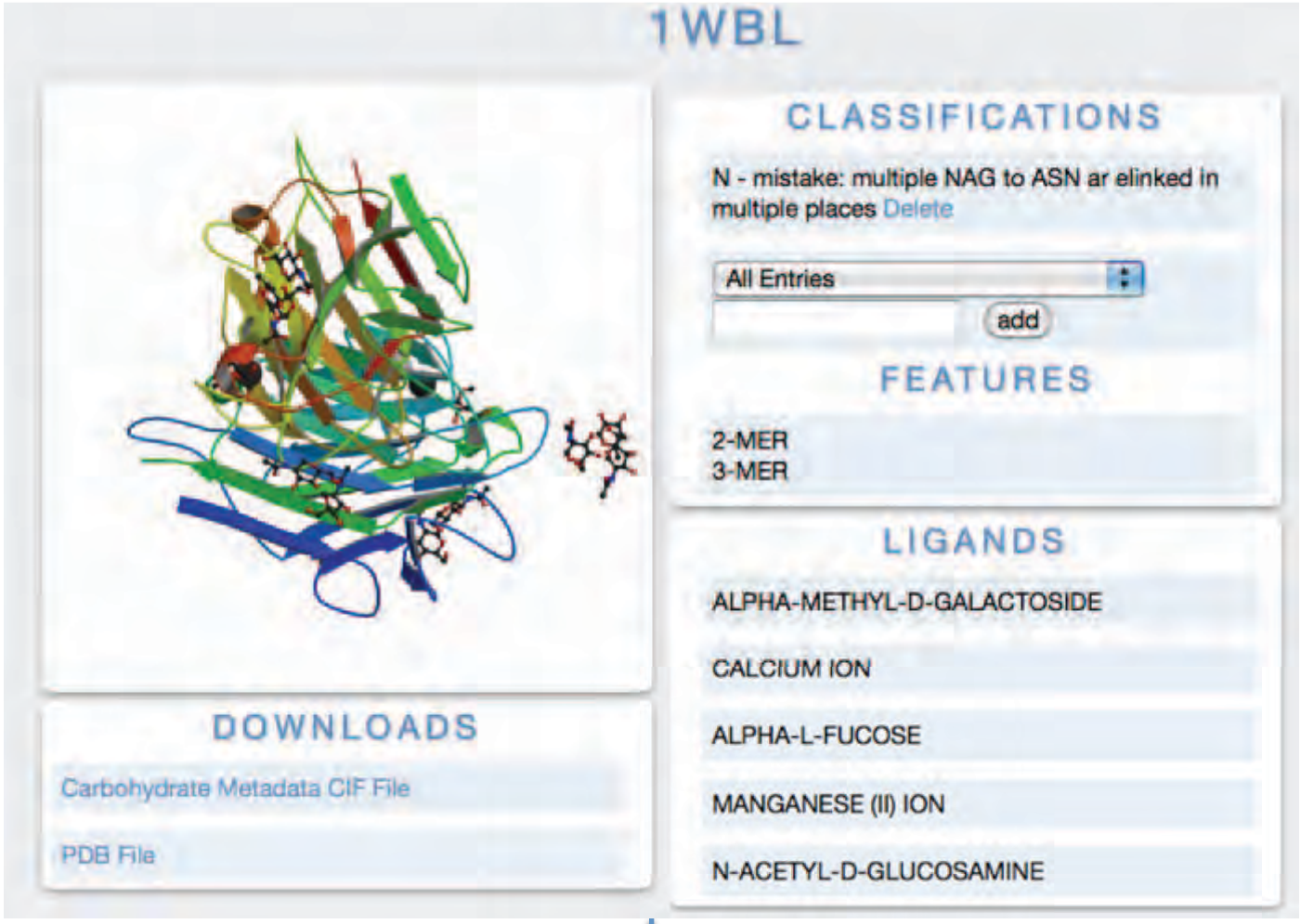
Summer Program 2011: Carbohydrates

- Identification, Analysis and Classification of Carbohydrates in the PDB
- 5 students (4 at RU, 1 at UCSD)
- Training by faculty from RU, UCSD, UK





Classification of Structures with Carbohydrates



Internal DB for Query, Analysis and Remediation

Proposed Collaborative Summer Programs

- Proposal submitted to NIH Sept. 2011
- Earliest date of offering Summer 2012
- Inspired by MAP based courses

RCSB PDB PROTEIN DATA BANK
RUTGERS

BioMaPS Institute
and
Graduate Program in
Computational Biology & Molecular Biophysics

UCSD
SDSC • SKAGGS SCHOOL of PHARMACY and PHARMACEUTICAL SCIENCES

Mobile Outreach Strategy

Greg Quinn

RCSB PDB
PROTEIN DATA BANK

Mobile Access Introduced in 2010-2011

- PDB*Mobile*

  HTML5-based mobile app released - more than 1500 downloads

- ePub

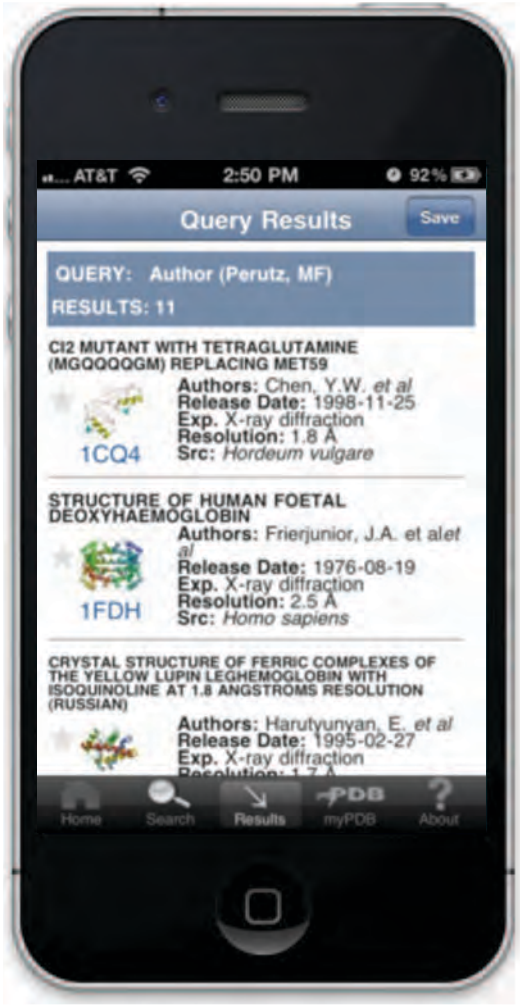
  Molecule of the Month articles can now be downloaded in portable ePub format

- Animations

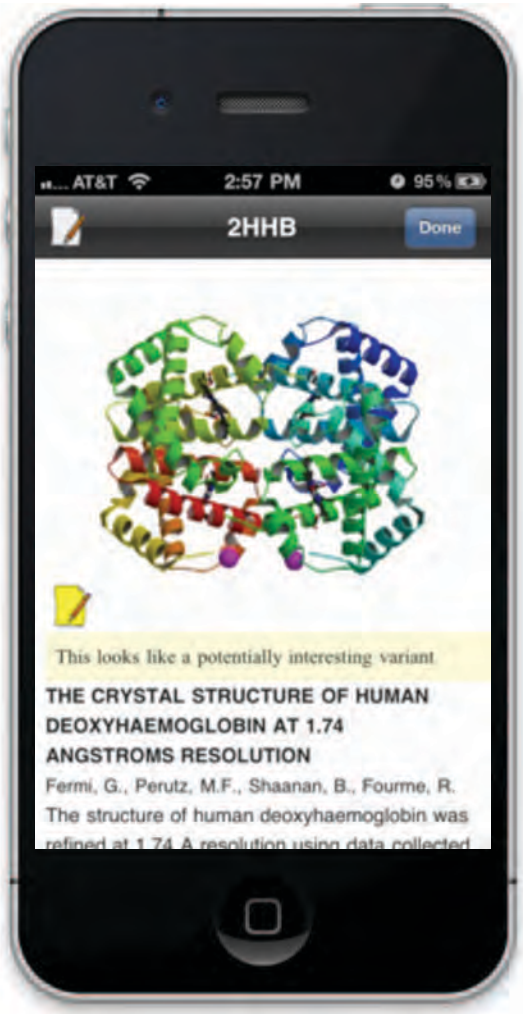
  Interactive molecular animations for PDB-101 discussed entries

PDB*Mobile*

Search

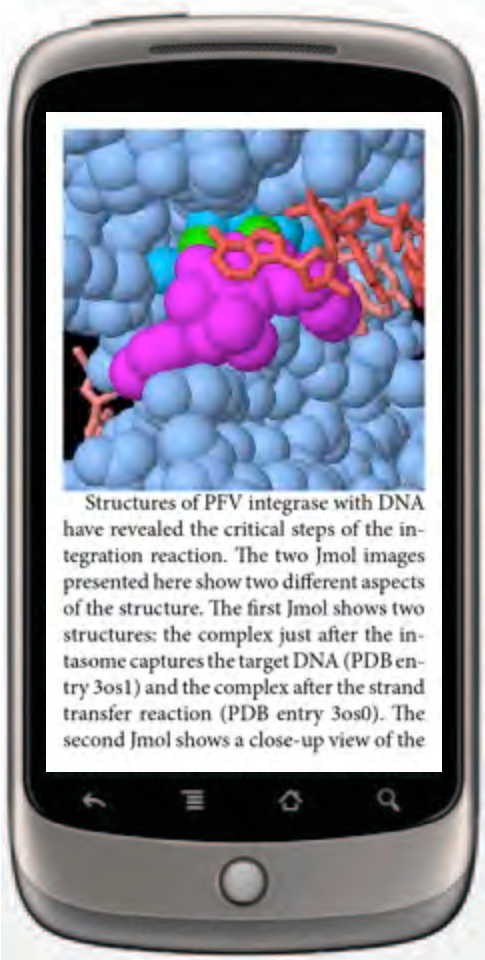


Browse

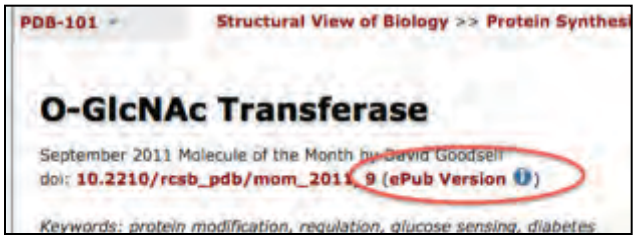


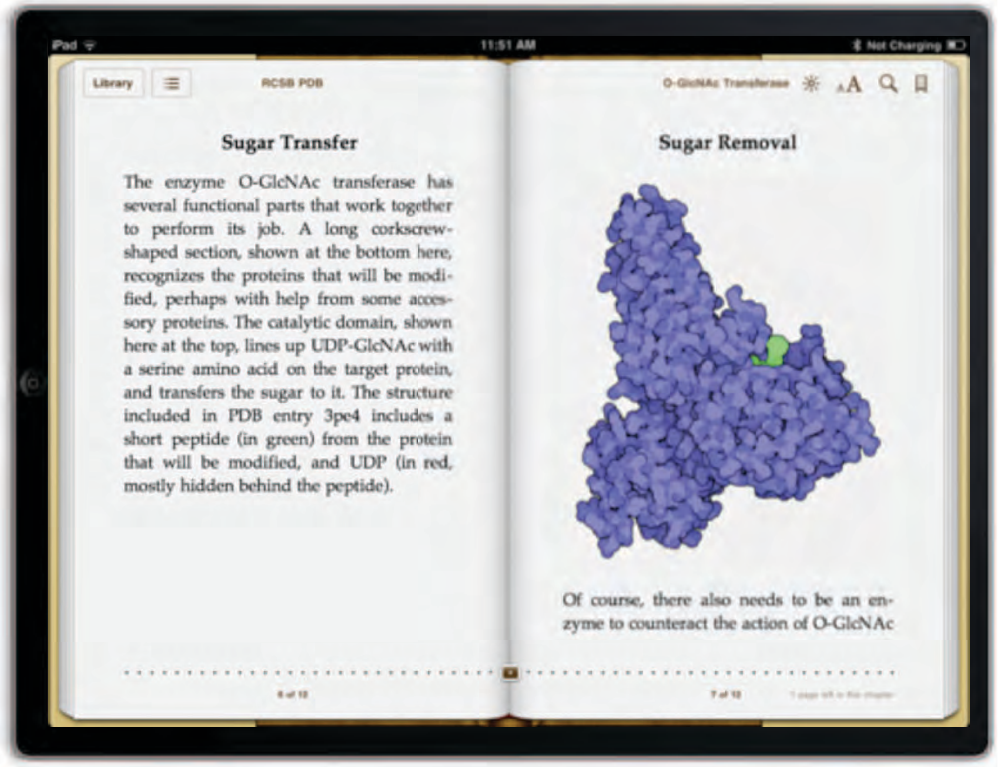
View and annotate

ePub Documents



Android

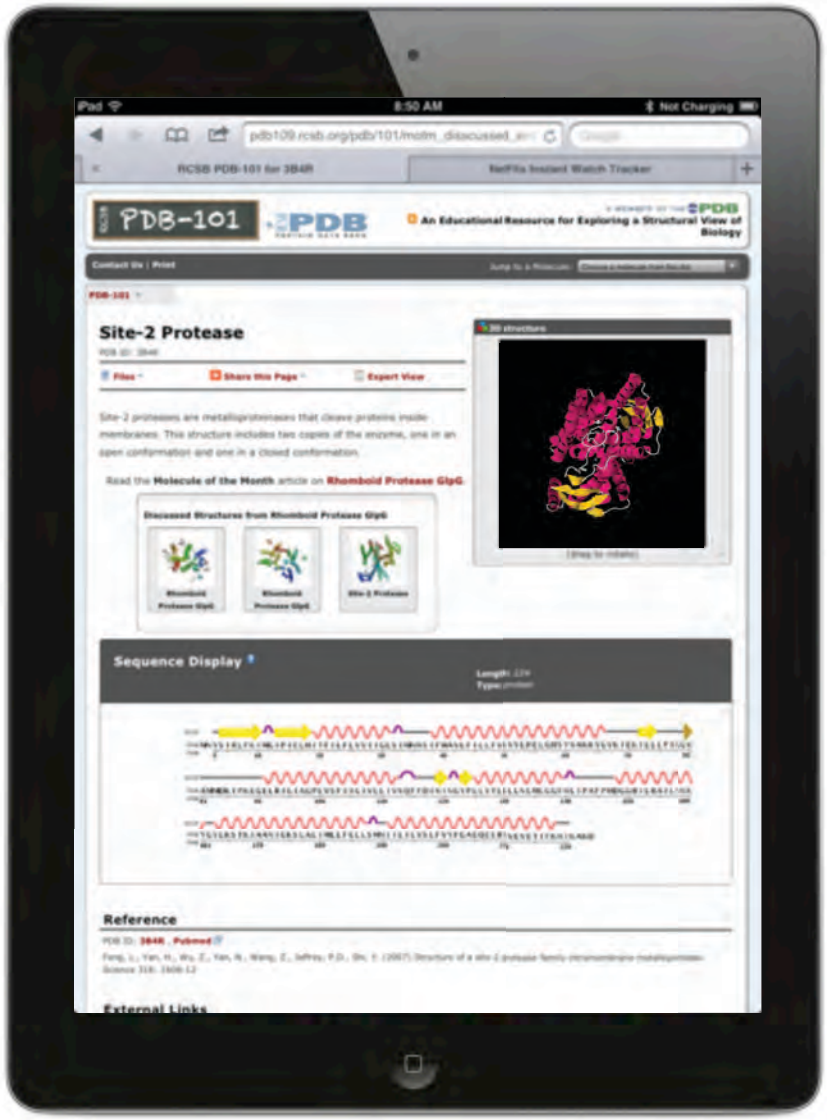




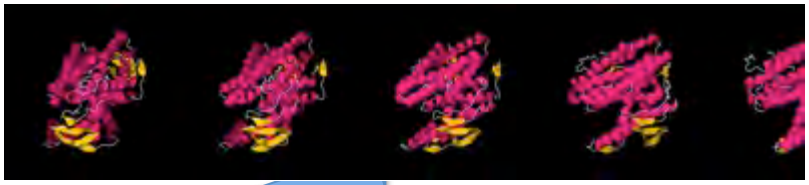
iOS (e.g. iPad, iPhone , iPod)

ePub Virtual Bookshelf of *Molecule of the Month* features

HTML5 Animations & Interactivity



Animation sprite strip



Animations and touch-based interactivity generated using HTML5 Canvas object and JavaScript

Looking Ahead: 2011-2012

Since the last AC meeting, the iPad has changed the landscape for web access

- Website needs to be far more mobile-friendly
- “apps” rule the day (even for the iPad!)

Many Different Platforms!

iOS 4

Windows phone

palm webOS™

RIM

BlackBerry

symbian OS

Many Different Mobile Devices!

Current Website Mobile View

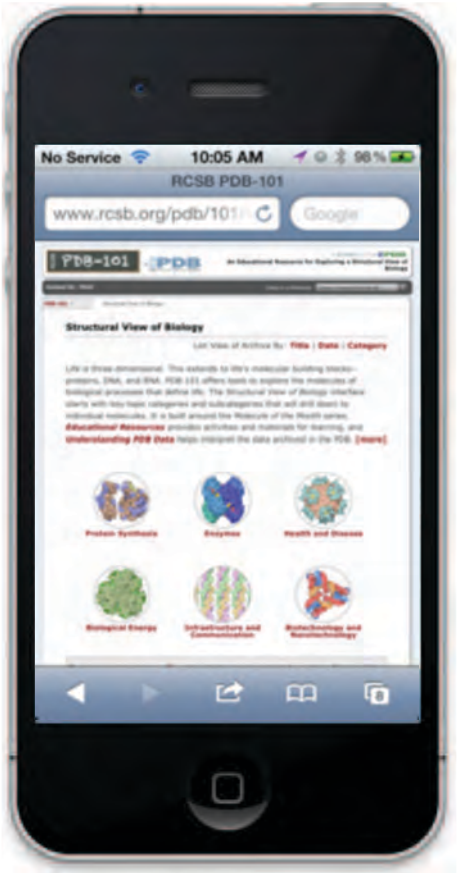
iPad

iPhone

Overall formatting is fine, but without pinch-to-zoom, the interface is close to impossible to read or navigate

- Unified UI design for both mobile and desktop display

Simplified UI



Current view

Simplified view

- Remove unnecessary user interface clutter
- Graphical element navigation
- Use icon prompts to access further textual content, instead of displaying by default

Content Adaptation

- Content-adaptation for mobile devices
  - Optimize layout
  - Optimize fonts
  - Optimize image sizes
  - Add navigation elements

CSS3 + HTML5

Query screen resolution

Query screen orientation

Normal view

With content adaptation

Website-Based Apps

- Platform-specific apps that are a portal to the website (i.e. a web browser)
- Support Android, iOS, any others that the open source Apache Callback SDK supports
- Push out through respective App Stores

Looking Ahead:
Gordon award to develop real-time animation solution for mobile devices

http request

animation

Jmol applet

Real time generation of animation sequences using ultra-fast I/O nodes on the Gordon compute resource

Gordon I/O nodes are solid state servers

Journal Interactions

Christine Zardecki

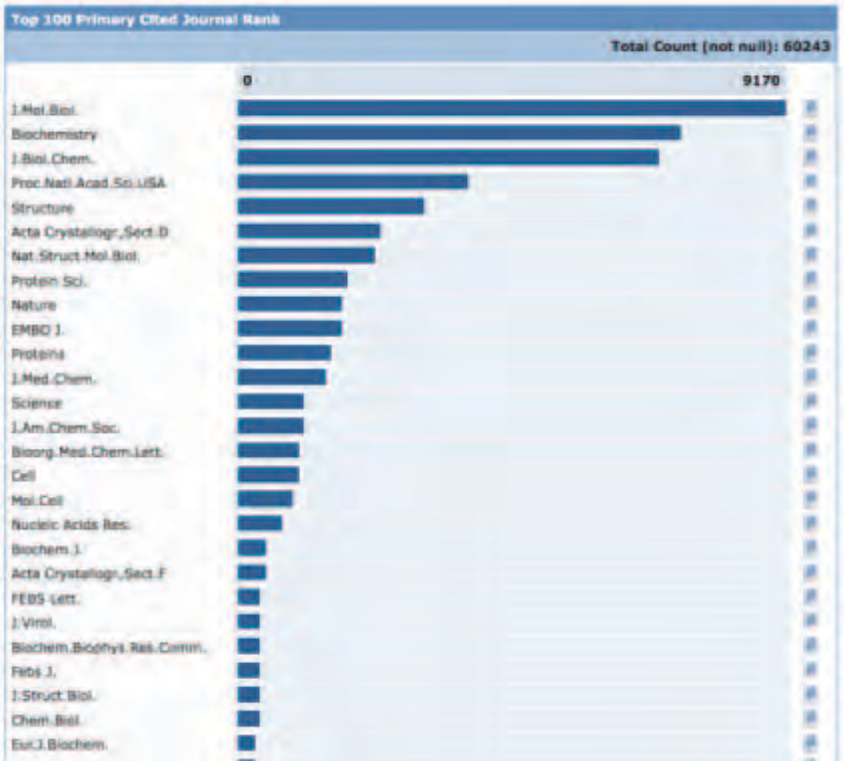
RCSB PDB
PROTEIN DATA BANK

Issues

- Better coordination of data release and publication
- Desire for improved data quality
  - How can we ensure that the claims made in a research paper are substantiated by the underlying data? (i.e., avoid retractions)
  - How can the PDB be an effective partner in the manuscript review process?

Top Journals Represented in the PDB

Top Overall

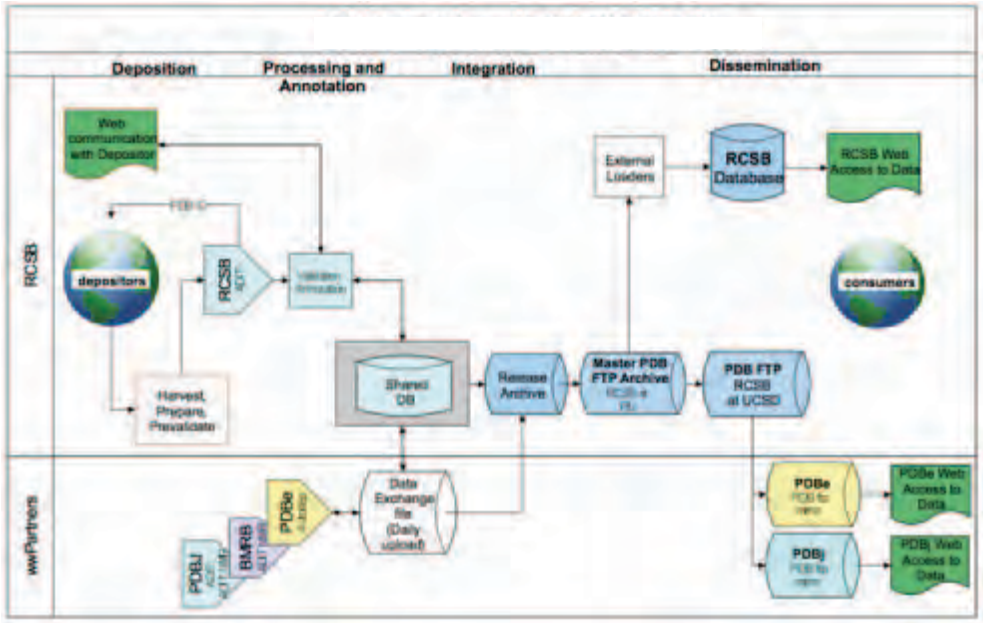


Top in 2010

1. *J.Biol.Chem.* 645
2. *J.Mol.Biol.* 592
3. *PNAS* 483
4. *Biochemistry* 425
5. *Nat.Struct.Mol.Biol.* 221
6. *J.Med.Chem.* 206
7. *J.Am.Chem.Soc.* 199
8. *Bioorg.Med.Chem.Lett.*192
9. *Nature* 181
10. *Structure* 176

Pipeline: Structure Deposition to Release

1. Author deposits data online
   - Release status set at this time
2. Entry validated and annotated
   - Validation Report and processed entry sent to author
   - Current processing status updated along the way
3. Once entry is approved by author, data can be released according to set status
4. Entry status can be checked by using the PDB ID in a search



Status codes:

HPUB: processing complete, entry on hold until publication;

HOLD: processing complete, entry on hold until a certain date;

PROC: to be processed;

WAIT: processing started, waiting for author input to continue processing;

AUTH: processed, waiting for author review and approval;

REPL: author sent new coordinates, entry to be reprocessed;

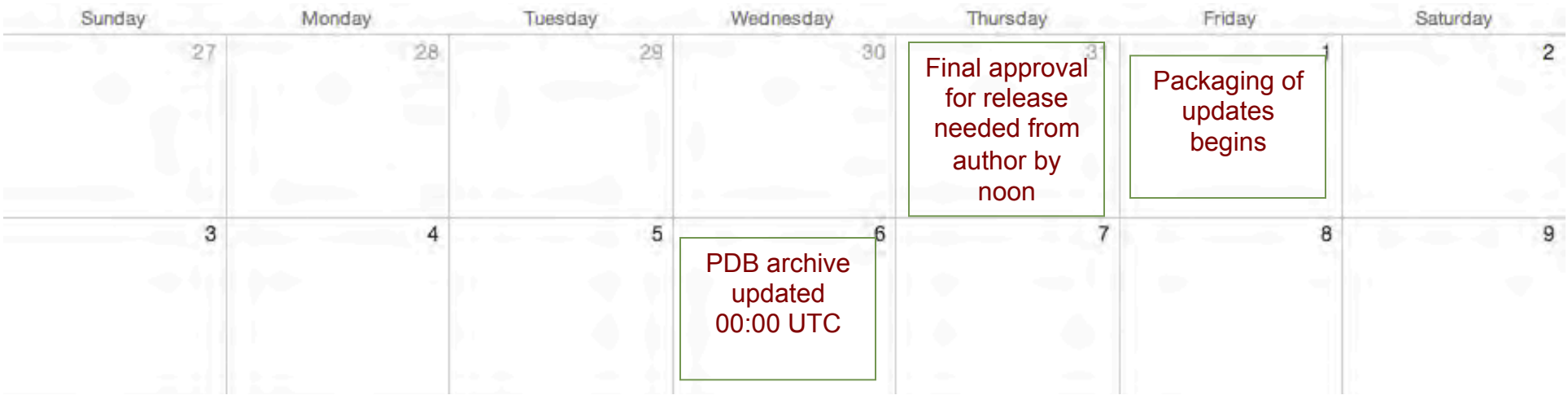
POLC: waiting for a policy decision;

REFI: re-refined entry, processing pending availability of primary publication;

WDRN: deposition withdrawn;

PDB Data Release Structure

- Final approval from author needed Thursdays by noon (local time)
  - Only entries approved by authors (implicit or explicit) are released
- Updates packaged on Fridays
- Archive is updated Wednesdays at 00:00 UTC (Coordinated Universal Time)



Request to Journals

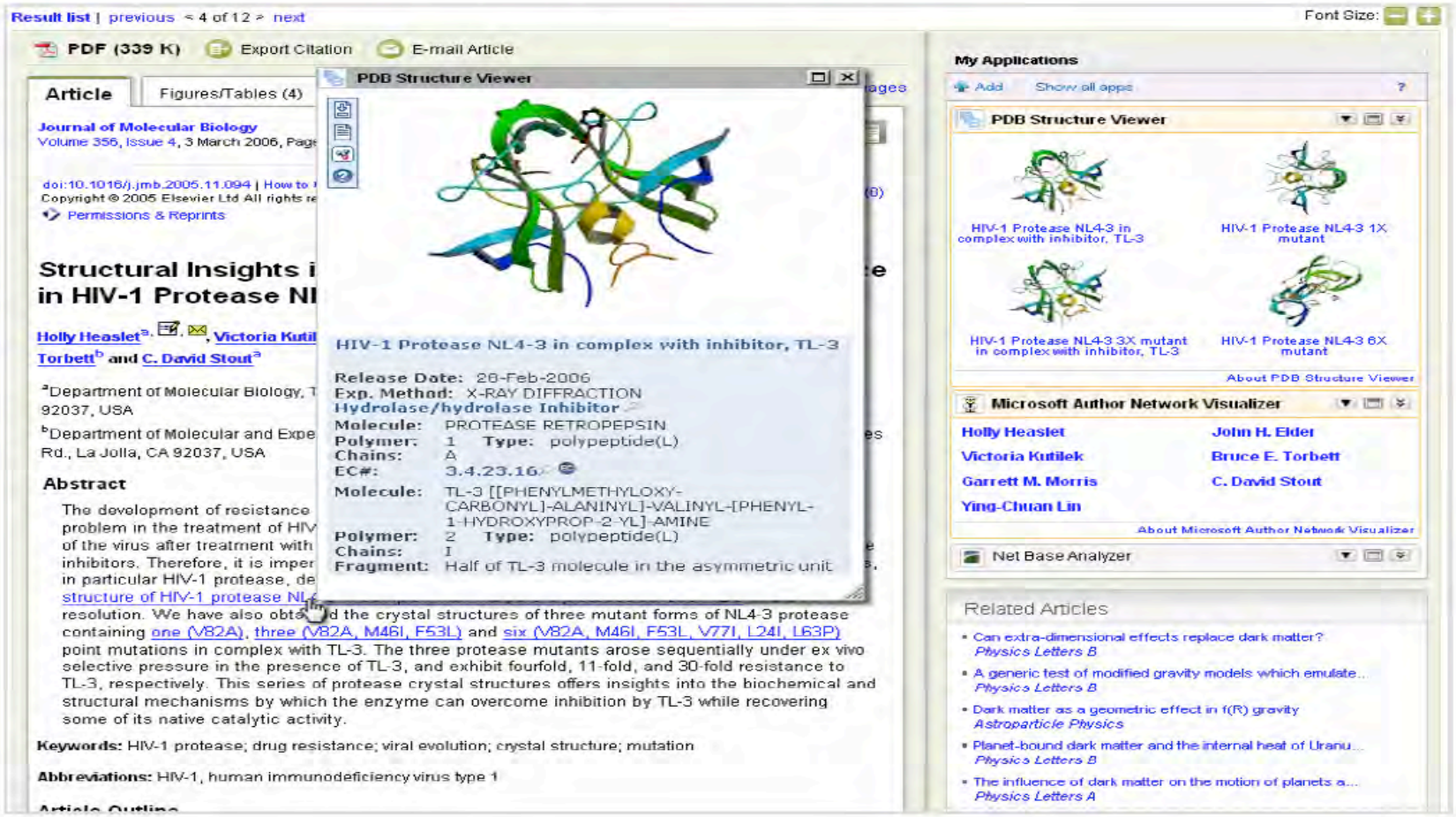
- To synchronize data release with online publication, provide in advance the information needed for release
  - Article title, author list, PDB ID(s), DOI, and publication date should be emailed to deposit@wwpdb.org 2 weeks prior to publication
  - Authors given opportunity for final verification
  - Data released as close to online publication as possible

Journals Notifications (ongoing or in discussion)

- Acta D&F (upon publication)
- FEBS
- Journal of Biological Chemistry
- Journal of Molecular Biology
- Nature, Nature Structural & Molecular Biology, Nature Chemical Biology
- Nucleic Acids Research
- Proteins
- PNAS

Journals Are Using the PDB Through Web Services

App Integration in Science Direct



From Anita de Waard, Elsevier

What Changes Occur After Release?

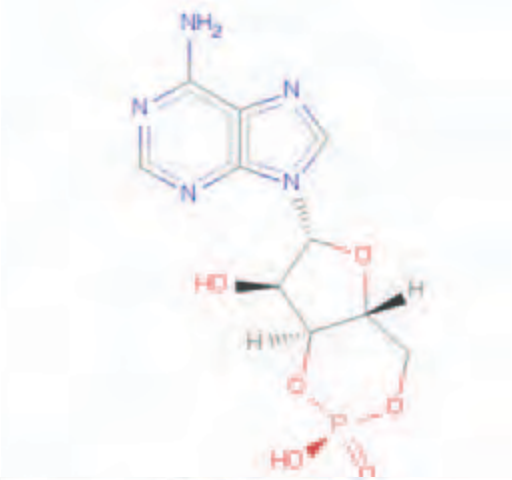
- Minor changes and updates (citation, numbering, etc.) are made regularly
- Major revisions to coordinates that change the geometry or chemical composition require the entry to be obsoleted (removed) and superceded (replaced by a new deposition)
  - Common procedure when authors have collected new data or have re-refined the entry
  - All obsolete entries remain available to the public through the PDB ftp archive

Cases of Obsolete Entries Without Replacement Entries

- Journal retracts the publication due to issues with the data; the retraction will be listed in the obsoleted file
- Author obsoletes an incorrect structure. The entry must contain a statement as to the reason for obsoleting the structure
- A third-party (such as the employer) requests that the entry is obsoleted (e.g., in case of malfeasance). The citation in the obsoleted entry must be a published explanation and retraction in a peer-reviewed journal

How Can We Prevent Retractions?

- PDB Validation Reports check
  - Geometry: Atom clashes, peptide linkage, covalent geometry
  - Sequence
  - Biological assembly
  - Ligand chemistry
  - Structure factor data
- High level summary provided as a PDF during deposition process



R-factors	
R-factor (Author reported)	0.150
R-factor (Calculated by SFCHECK, V7.02.4)	0.212
R-factor (Calculated by REFMAC, V5.5.0109)	0.1960
Free R-factor (Author reported)	0.188
Free R-factor (Calculated by SFCHECK, V7.02.4)	0.236
Free R-factor (Calculated by REFMAC, V5.5.0109)	0.2200

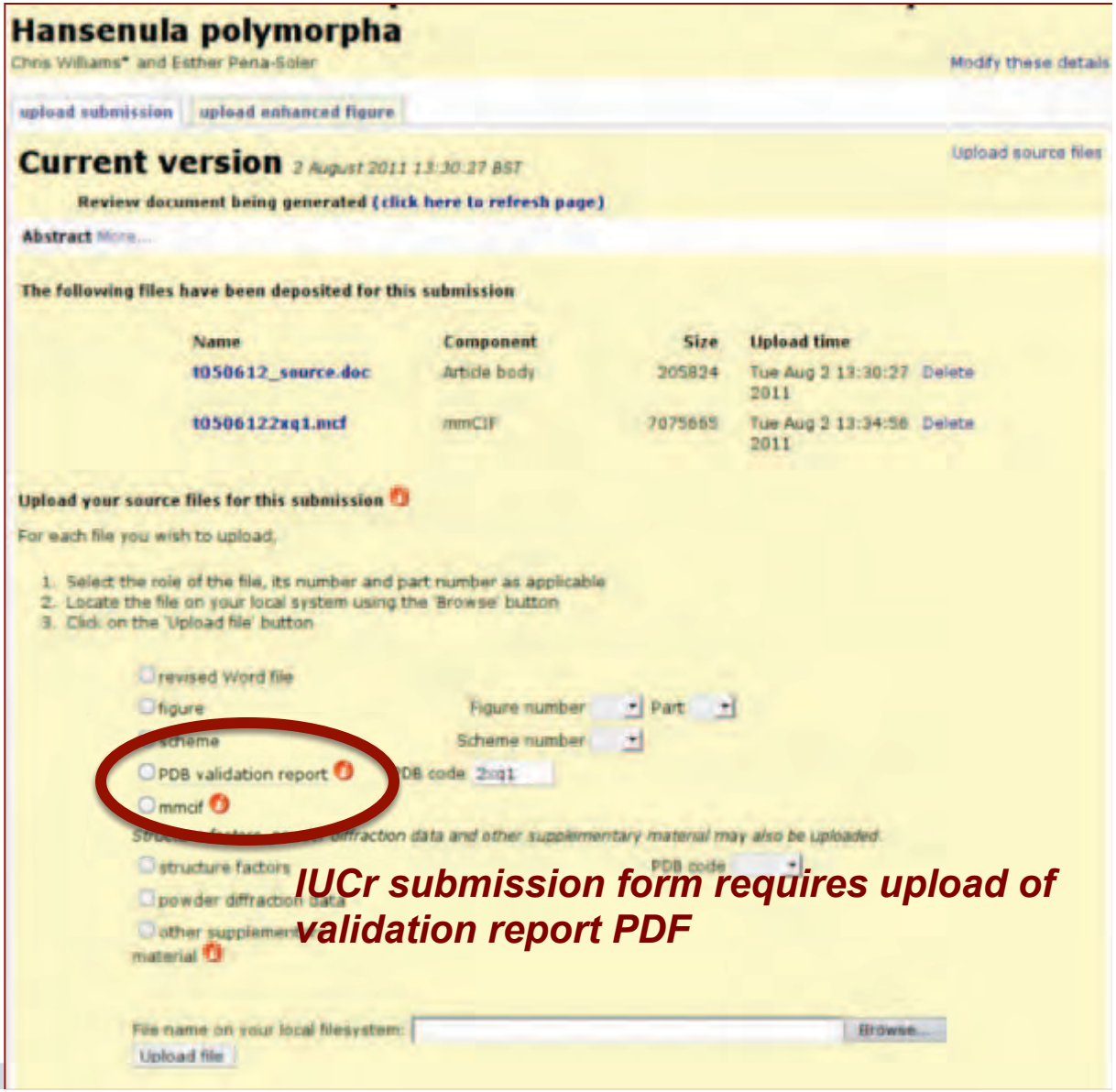
Structure quality	
Average Real space R-factor (Deviation) (Calculated by SFCHECK, V7.02.4)	0.0767
Average Real space R-factor (Deviation) (Calculated by MAPMAN, V7.8.5)	0.1007
Average Real-space correlation coefficient (Deviation) (Calculated by SFCHECK, V7.02.4)	0.9858
Average Real-space correlation coefficient (Deviation) (Calculated by MAPMAN, V7.8.5)	0.963
Average Occupancy-weighted avg temperature factor (Deviation)	35.25

Wilson statistics (PHENIX, V1.6-289)	
Wilson B-factor	31.65
Wilson Scale	0.12

Example from PDF Report

Current Validation PDF Report

- Provides quick assessment of structure quality without access to coordinates
- Identifies most reported issues that led to retractions
- Authors can provide validation PDF to journal reviewers
- Part of Acta Cryst pipeline



IUCr Experience so far

- Introduced at the end of 2010
- All authors seem able to submit validation reports
- No noticeable reduction in the number of articles submitted
- Welcomed by authors, editors and reviewers

Crystallography Journals Online

How Do We Get More Journal Involvement?

- Each journal has its own pipeline and issues
  - Needs one-on-one relationships (e.g., Acta D&F, Nature and Journal of Biological Chemistry)
- Community interactions