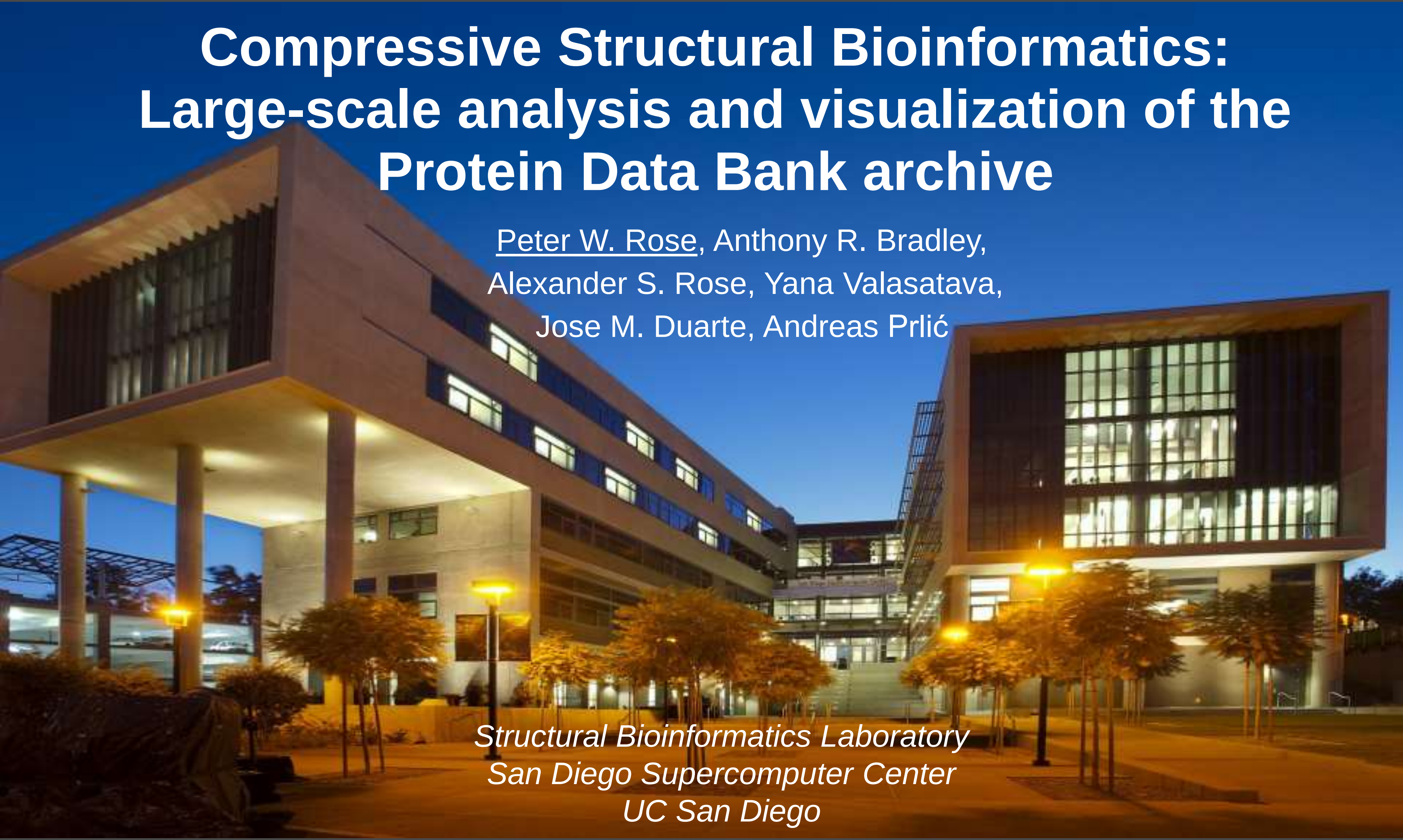
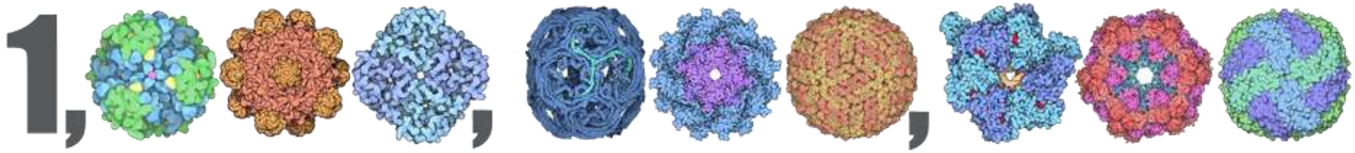
# Compressive Structural Bioinformatics: Large-scale analysis and visualization of the Protein Data Bank archive

Peter W. Rose, Anthony R. Bradley,
Alexander S. Rose, Yana Valasatava,
Jose M. Duarte, Andreas Prlić

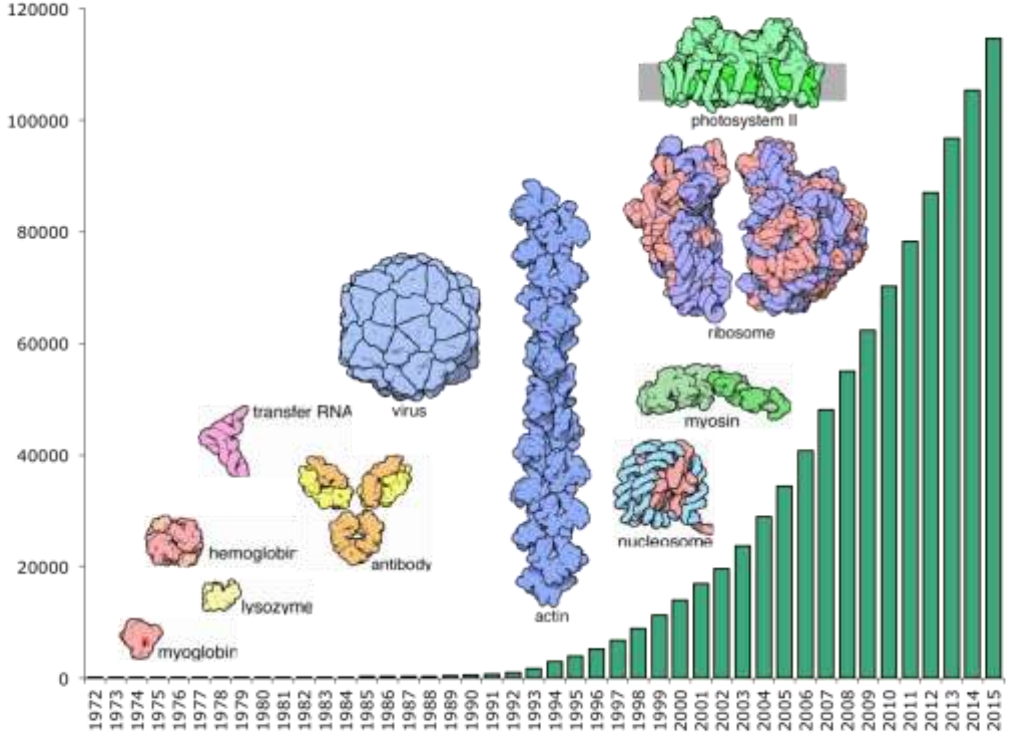
*Structural Bioinformatics Laboratory*
*San Diego Supercomputer Center*
*UC San Diego*



# PDB – A Billion Atom Archive

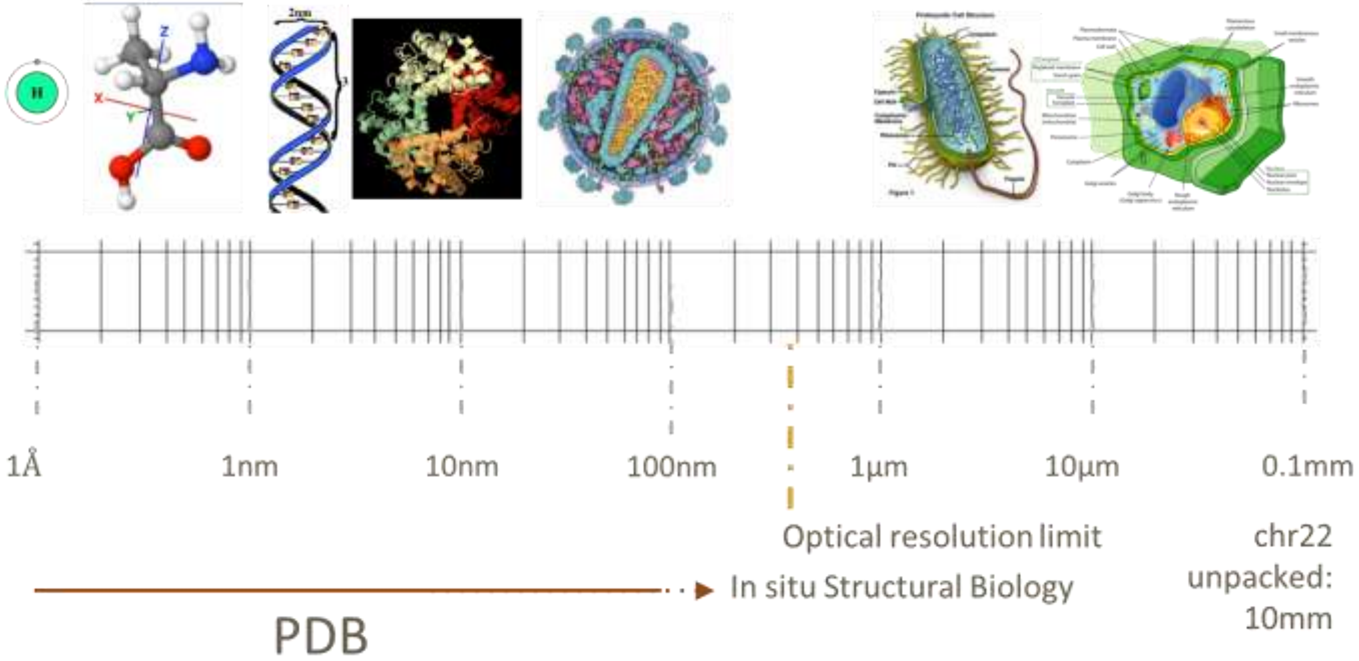


*> 1 billion atoms in the asymmetric units*



*120,000*
*structures*
*in June 2016*

# Growing Structure Size and Complexity



Optical resolution limit

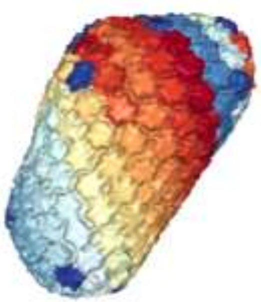
In situ Structural Biology

PDB

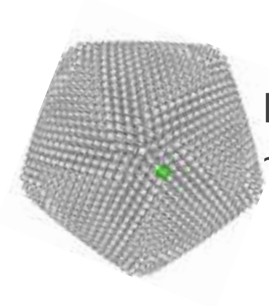
chr22
unpacked:
10mm

Largest asymmetric structure in PDB

Largest symmetric structure in PDB



HIV-1 capsid: PDB ID 3J3Q
~2.4M unique atoms

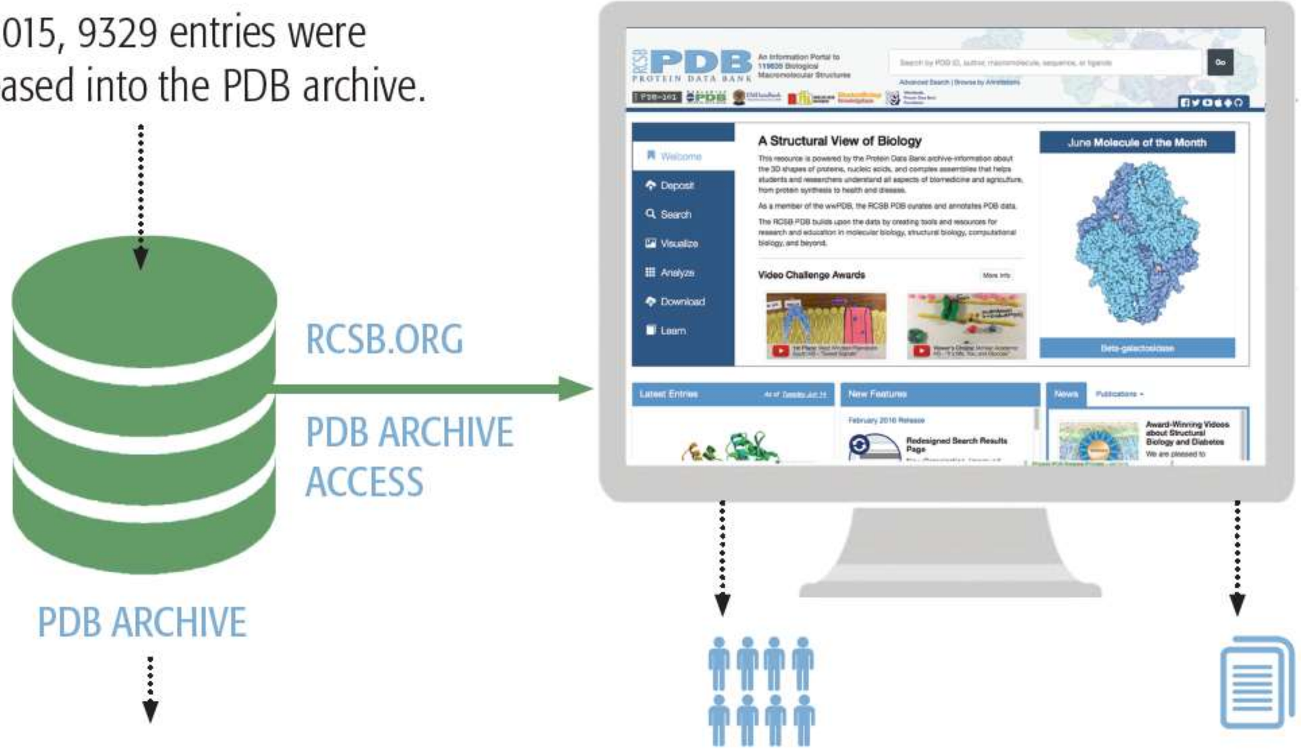


Faustovirus major capsid: PDB ID 5J7V
~40M overall atoms

# Growing User Base

## ACCESSING PDB AND RCSB PDB

In 2015, 9329 entries were released into the PDB archive.



Total PDB archive traffic from all wwPDB partners totaled 534,339,871 downloads

Each month in 2015, **rcsb.org** was visited 741,000 times on average by 315,000 unique visitors

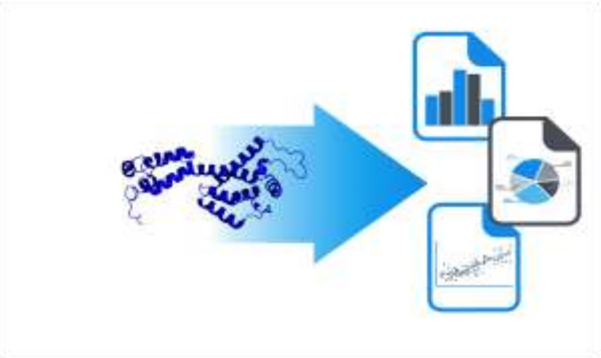
A total of 35,260 GB of data were accessed

# → Scalability Issues

- **Interactive visualization**
  - slow network transfer
  - slow parsing
  - slow rendering

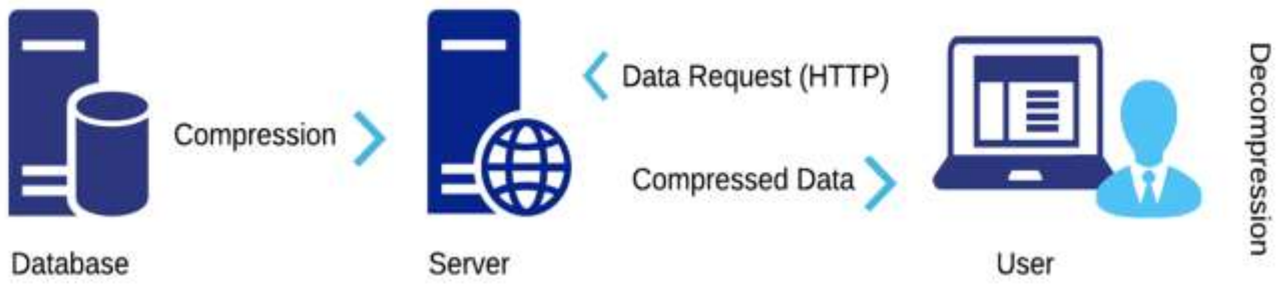
- **Mobile visualization**
  - limited bandwidth
  - limited memory

- **Large-scale structural analysis**
  - slow repeated I/O
  - slow repeated parsing

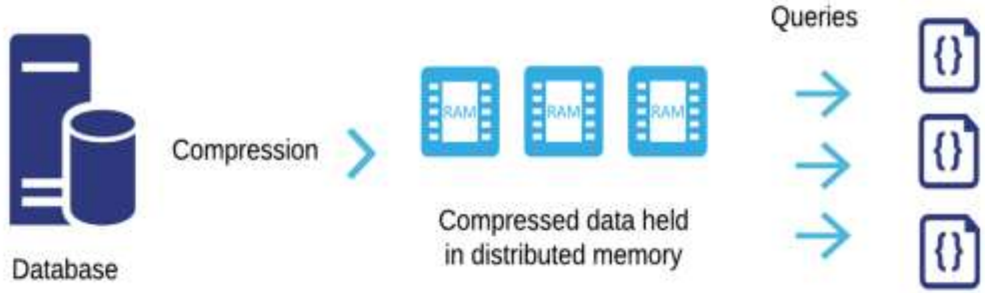


# Compressive Structural Bioinformatics

Efficiently store, transmit, and visualize 3D structures of biological macromolecules

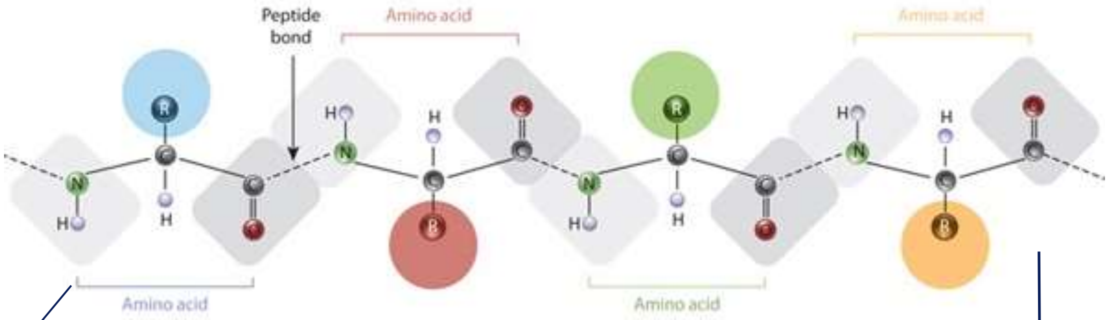


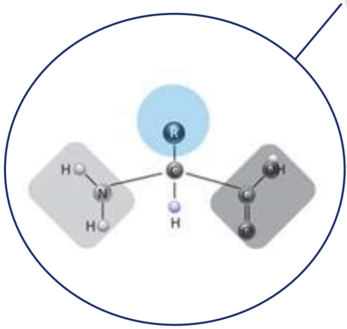
Perform large-scale structural calculations such as geometric queries or structural comparisons over the entire PDB archive held in memory



# Macromolecular 3D Structure

Biological macromolecules: proteins, nucleic acids

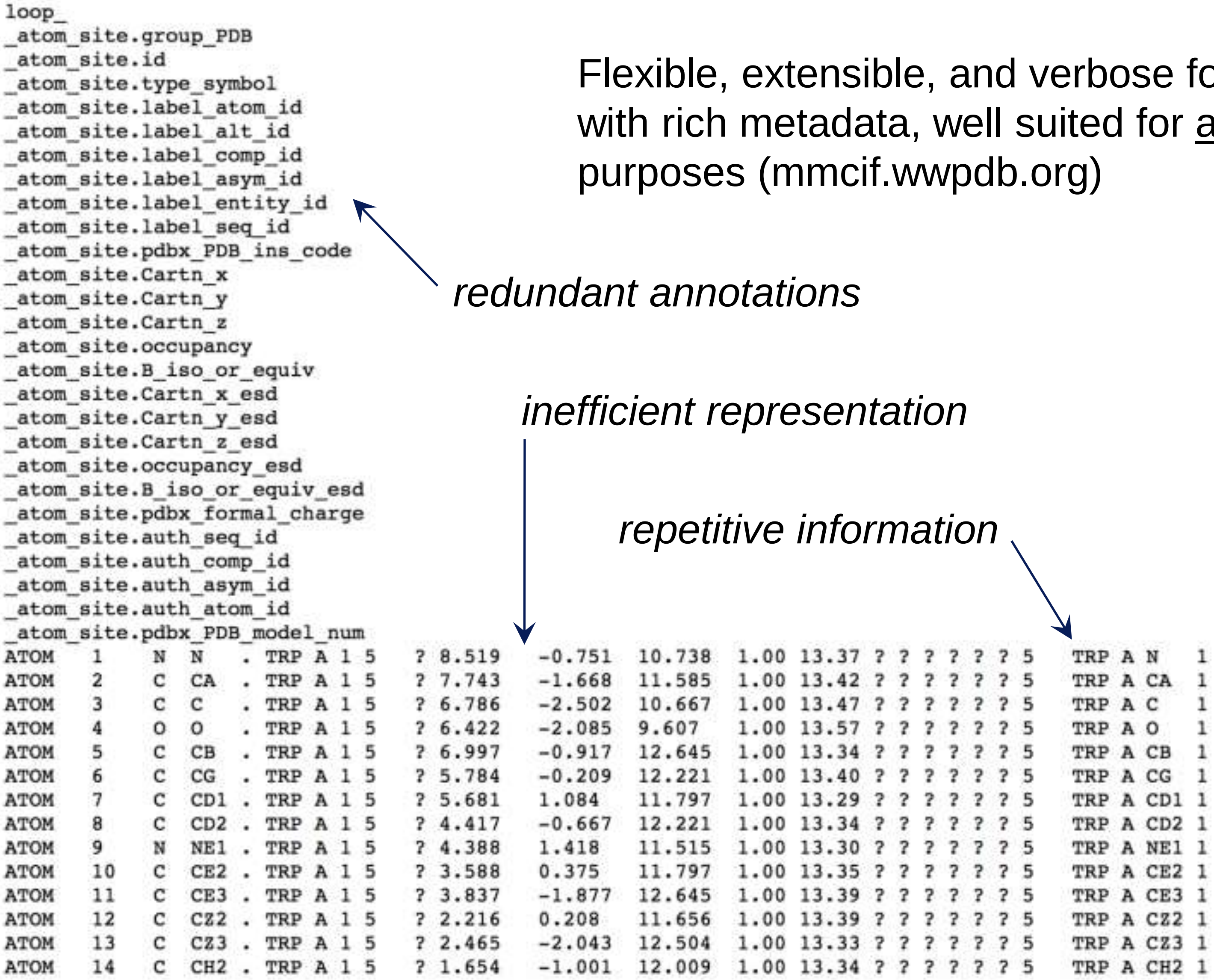




Biological macromolecules are polymers constructed by linking monomers by covalent bonds

# PDBx/mmCIF

Flexible, extensible, and verbose format with rich metadata, well suited for archival purposes (mmcif.wwpdb.org)



```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.Cartn_x_esd
_atom_site.Cartn_y_esd
_atom_site.Cartn_z_esd
_atom_site.occupancy_esd
_atom_site.B_iso_or_equiv_esd
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N N   . TRP A 1 5   ? 8.519   -0.751  10.738  1.00 13.37 ? ? ? ? ? ? 5   TRP A N   1
ATOM   2    C CA  . TRP A 1 5   ? 7.743   -1.668  11.585  1.00 13.42 ? ? ? ? ? ? 5   TRP A CA  1
ATOM   3    C C   . TRP A 1 5   ? 6.786   -2.502  10.667  1.00 13.47 ? ? ? ? ? ? 5   TRP A C   1
ATOM   4    O O   . TRP A 1 5   ? 6.422   -2.085  9.607   1.00 13.57 ? ? ? ? ? ? 5   TRP A O   1
ATOM   5    C CB  . TRP A 1 5   ? 6.997   -0.917  12.645  1.00 13.34 ? ? ? ? ? ? 5   TRP A CB  1
ATOM   6    C CG  . TRP A 1 5   ? 5.784   -0.209  12.221  1.00 13.40 ? ? ? ? ? ? 5   TRP A CG  1
ATOM   7    C CD1 . TRP A 1 5   ? 5.681   1.084   11.797  1.00 13.29 ? ? ? ? ? ? 5   TRP A CD1 1
ATOM   8    C CD2 . TRP A 1 5   ? 4.417   -0.667  12.221  1.00 13.34 ? ? ? ? ? ? 5   TRP A CD2 1
ATOM   9    N NE1 . TRP A 1 5   ? 4.388   1.418   11.515  1.00 13.30 ? ? ? ? ? ? 5   TRP A NE1 1
ATOM   10   C CE2 . TRP A 1 5   ? 3.588   0.375   11.797  1.00 13.35 ? ? ? ? ? ? 5   TRP A CE2 1
ATOM   11   C CE3 . TRP A 1 5   ? 3.837   -1.877  12.645  1.00 13.39 ? ? ? ? ? ? 5   TRP A CE3 1
ATOM   12   C CZ2 . TRP A 1 5   ? 2.216   0.208   11.656  1.00 13.39 ? ? ? ? ? ? 5   TRP A CZ2 1
ATOM   13   C CZ3 . TRP A 1 5   ? 2.465   -2.043  12.504  1.00 13.33 ? ? ? ? ? ? 5   TRP A CZ3 1
ATOM   14   C CH2 . TRP A 1 5   ? 1.654   -1.001  12.009  1.00 13.34 ? ? ? ? ? ? 5   TRP A CH2 1
```

*redundant annotations*

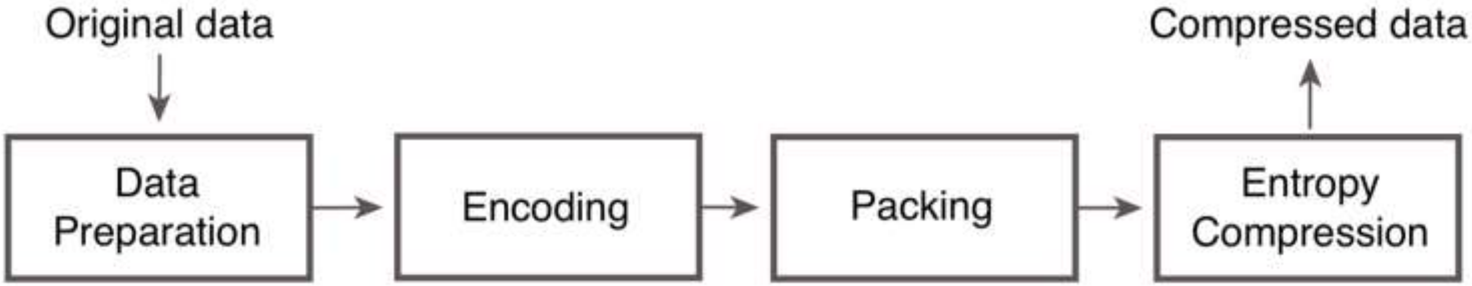
*inefficient representation*

*repetitive information*

MMTF

- **MacroMolecular Transmission Format (mmtf.rcsb.org)**
  - Compact
    - fast network transfer, less I/O
  - Fast to parse
    - binary, no string parsing
  - Contains information for structural analysis and visualization
    - covalent bonds and bond orders
    - consistently calculated secondary structure

# MMTF Compression Pipeline



Original data

Data Preparation → Encoding → Packing → Entropy Compression

Compressed data

*extract structural data*
*calculate bonds, SSE*

*integer encoding*
*dictionary encoding*
*run-length encoding*
*delta encoding*

*recursive indexing*

*GZIP*

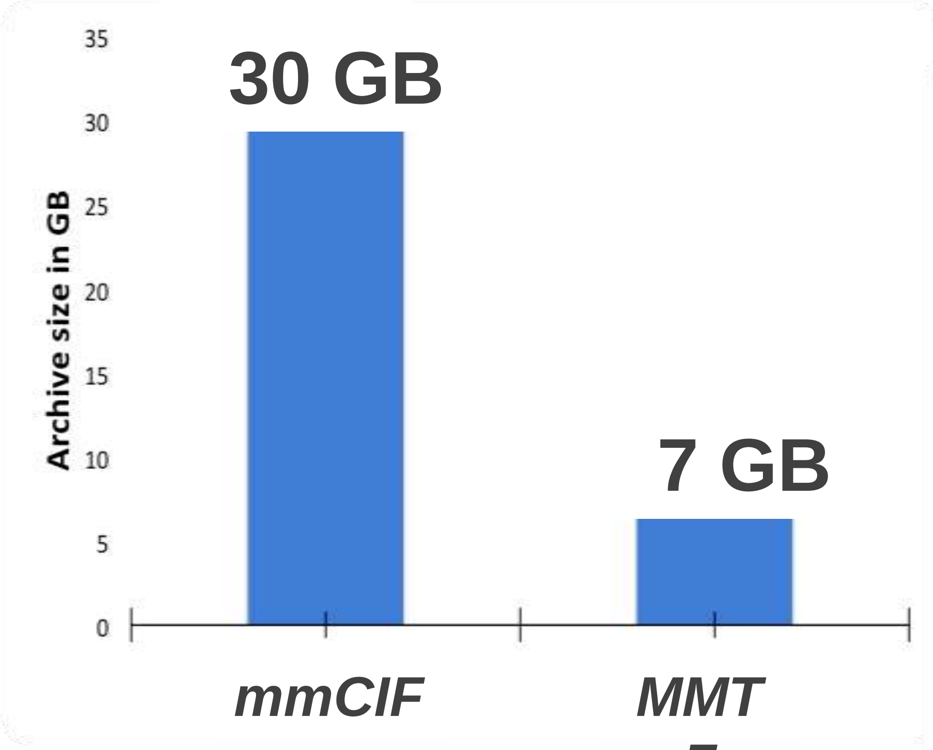
*Binary, extensible container format of MMTF*

MessagePack

*It's like JSON.*
*but fast and small.*

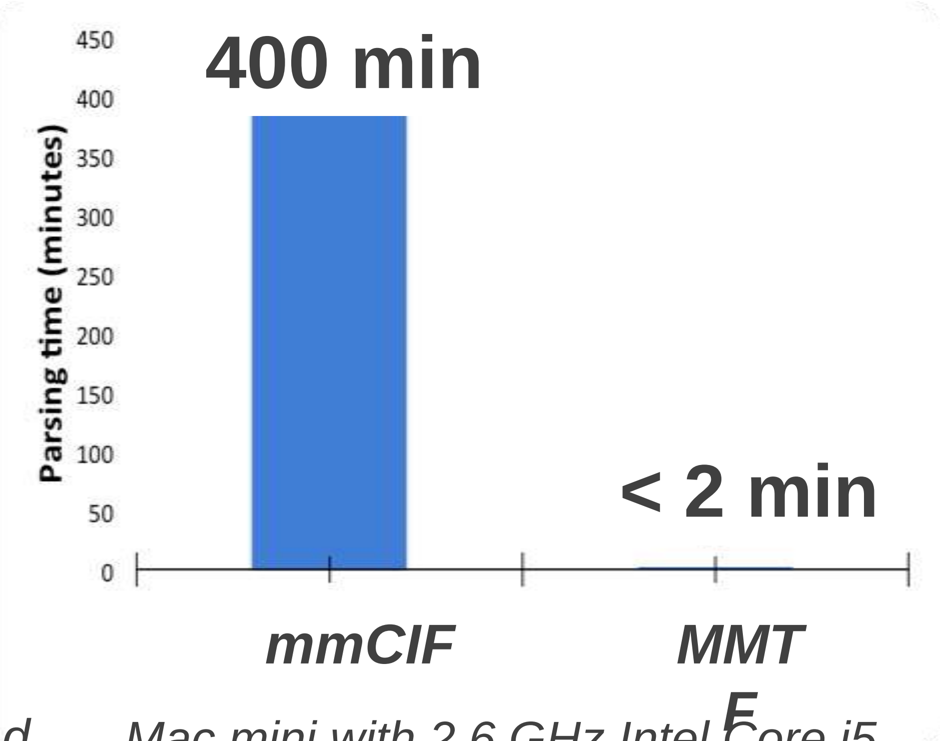
# Size and Parsing Speed
# mmCIF vs. MMTF for 120,000 Structures

## Small



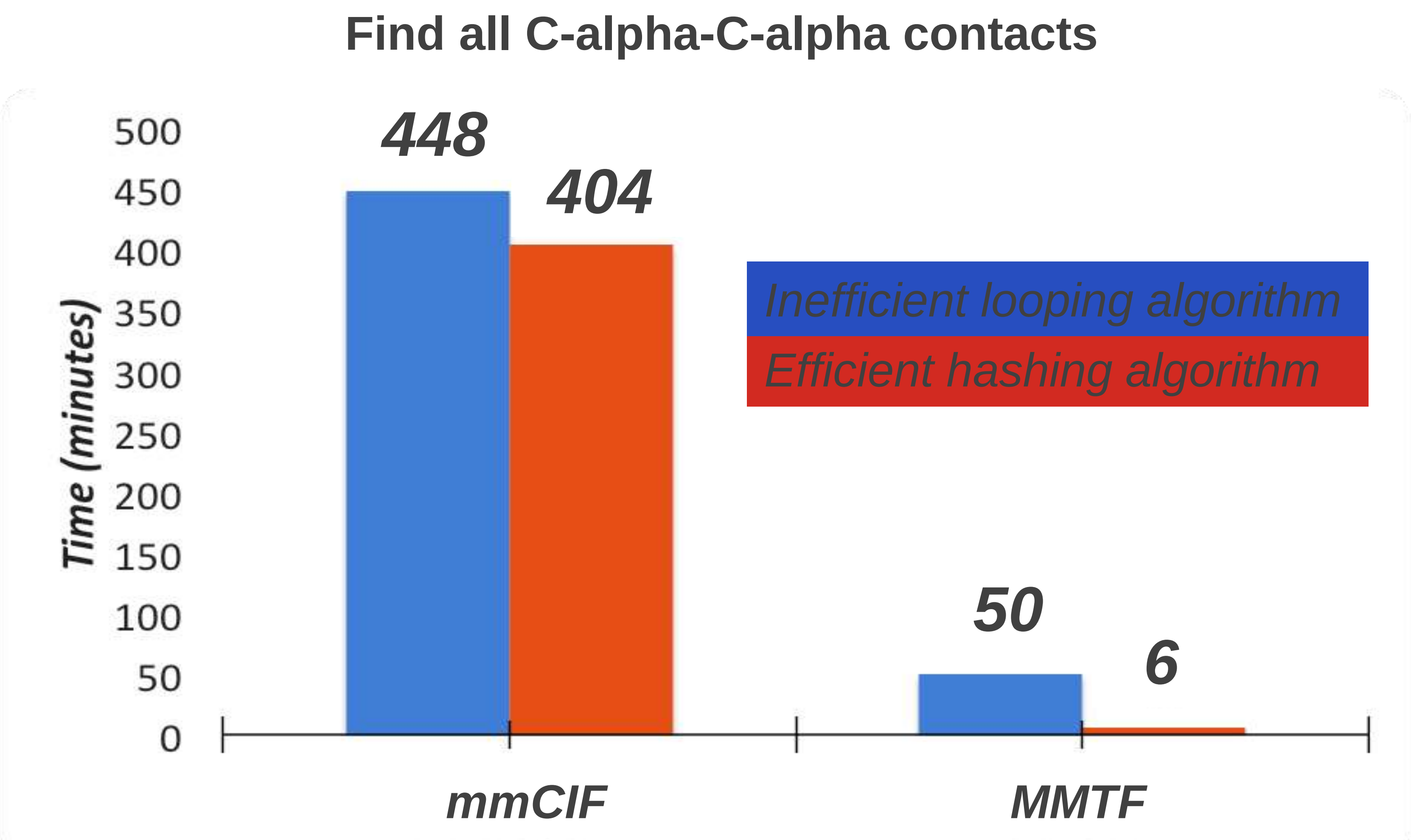
*Whole PDB archive GZIP compressed*
*(MMTF reduced/lossy: ~800 MB)*

## Fast



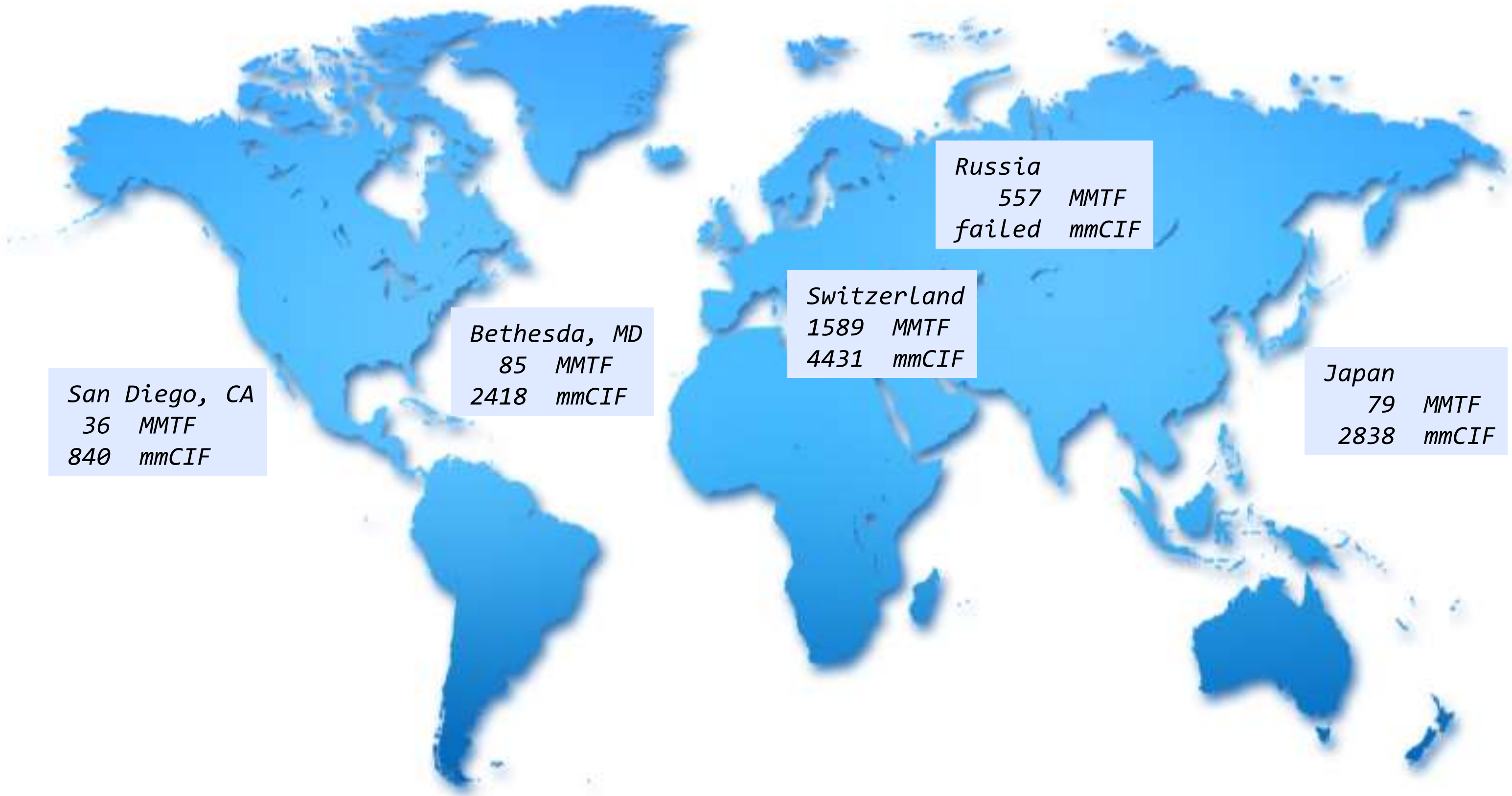
*Mac mini with 2.6 GHz Intel Core i5*
*(4 cores) and 16GB RAM using*

# Data Mining using Apache Spark mmCIF vs. MMTF



# Download + Parsing time MMTF vs. mmCIF

Time (seconds) to download* 100 large PDB structures from UCSD and parse with JavaScript decoder in Chrome browser



San Diego, CA
36 MMTF
840 mmCIF

Bethesda, MD
85 MMTF
2418 mmCIF

Switzerland
1589 MMTF
4431 mmCIF

Russia
557 MMTF
failed mmCIF

Japan
79 MMTF
2838 mmCIF

*Note: download times are highly variable and not representative

# Community Engagement

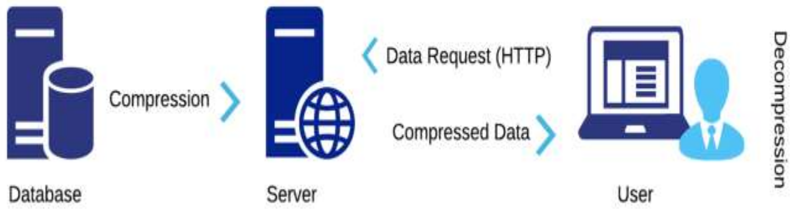
- **Open source specification**
- **Open source decoding libraries**
  - Java
  - JavaScript
  - Python
  - C/C++ (developed by community members)
- **Applications using MMTF**
  - 3Dmol.js, JSmol, iCn3D(NCBI), ICM Viewer, PyMol
  - BioJava, Biopython, MDAnalysis
  - RCSB PDB website

# Summary

- **MacroMolecular Transmission Format (MMTF, mmtf.rcsb.org)**
  - Compressed, binary, efficient representation of 3D structures
    - Lossless representation (~4x compression)
    - Lossy, reduced representation (~37x compression)

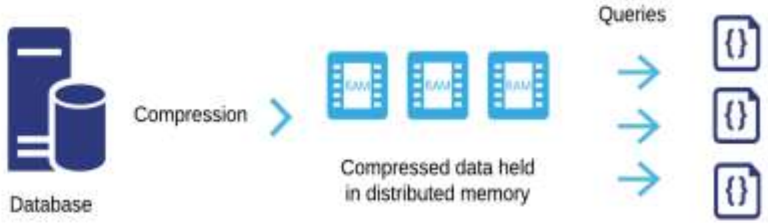
- **Compressive Structural Bioinformatics**
  - Algorithms, application, and workflows using MMTF
    - 10 to 100+ fold speedup

*Structure Visualization*



*Web-based molecular graphics for large complexes (2016)*
*Web 3D '16, 185-186, DOI: 10.1145/2945292.2945324*

*Large Scale PDB Mining*



# Acknowledgements

*Funding: NCI/NIH (U01 CA198942)*

NIH
Big Data to Knowledge(BD2K)

*MMTF Early Adopters*

molsoft
molecules *in silico*

BIOPYTHON

NCBI

SCHRÖDINGER.

BioJava

Jmol

3Dmol.js

RCSB PDB
PROTEIN DATA BANK