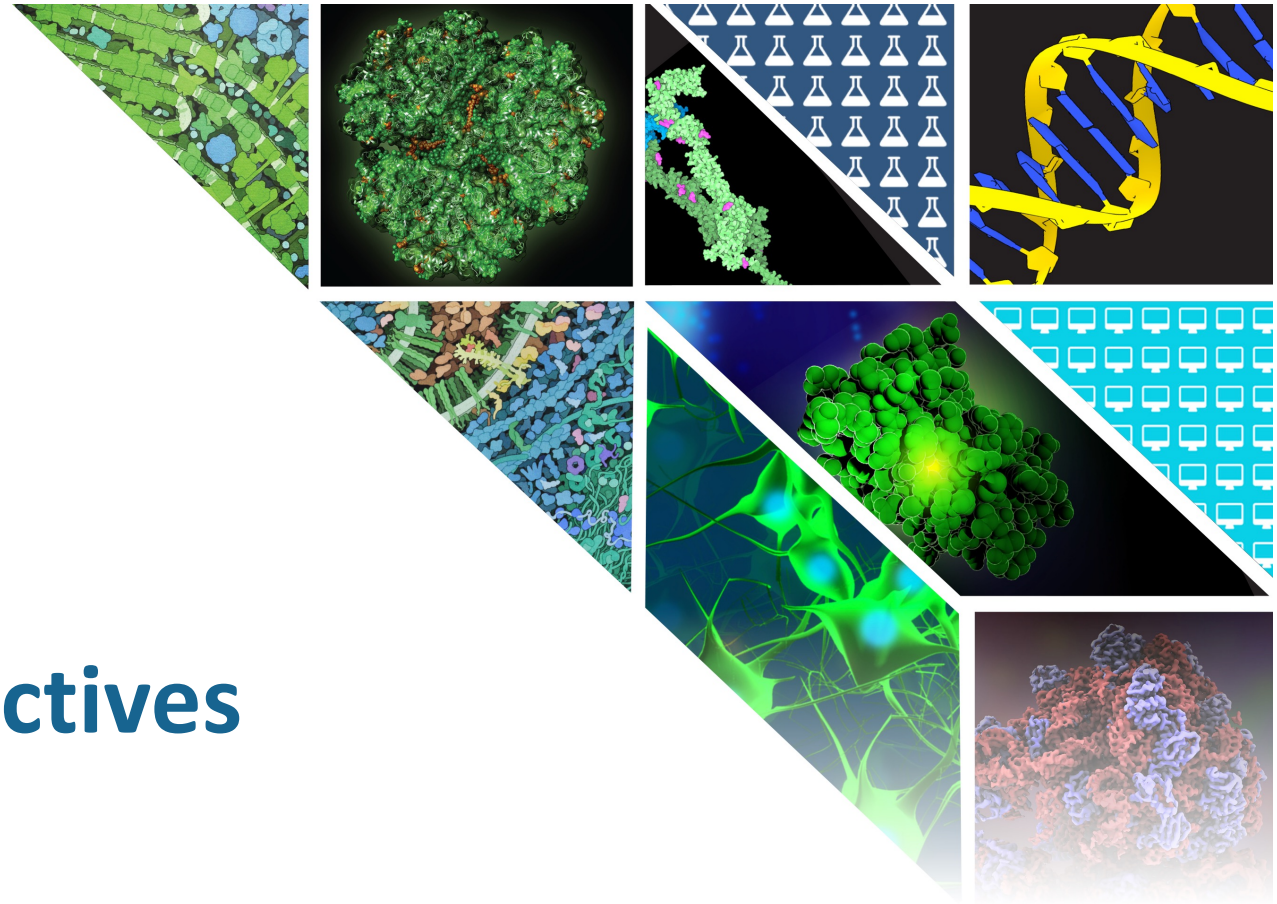
RCSB PDB
PROTEIN DATA BANK

rcsb.org



HDRMX: Perspectives from the PDB

Ezra Peisach 2025-03-25

History….

- PDB has been around since 1971
- Meta-data present in legacy PDB file format as structured remarks
- Remark 200 – mirrored the “table 1” information from a publication
  - Data source, Rmerge, resolution limits, # reflections, etc.
- Not until 2008 was “structure factor” data mandatory
  - Amplitude or intensity data – not really “structure factors”
- Prior to then, only minimal checking on agreement between data and coordinate model

Loose coupling between model & SF files

- PDB has treated coordinate file as authoritative for validation
  - Unit cell, space group, wavelength
  - Data collection statistics
  - PDB attempts to “reproduce” R-factors
- The only connection between coordinate file and SF is _diffrn.id
  - Assumed to be “1”
  - Does not handle connection of multiple datasets well

What is PDB doing currently

- Minimal cross checking experimental data
  - Test set present if used for refinement
  - Unit cell similar
  - Space group matching
  - Wavelength mismatch
  - First data block only!
- Validation software attempts to reproduce refined R/Rfree values
  - Makes various assumptions when multiple columns present in data
    - I vs F vs anomalous signal

Unmerged and unscaled data

- Gold standard would be to archive all raw images
- Not practical, but there is value to collecting unscaled/unmerged data
  - After the fact analysis of crystal decay
  - Removal of sweeps of data with poor quality
  - Possible reprocessing
- Unmerged data have been accepted by PDB but
  - Assume single panel detector that is perfectly aligned
- Work in conjunction with PDBx/mmCIF working group
  - Parameterize multi-panel detectors using imgCIF
  - More meta-data collected about location of diffraction image

In ideal world...

- PDB should treat the structure factor file as authoritative for experimental statistics
- Multi-datablock SF files should
  - Contain a machine readable description of each block
  - Data collection and processing should control each block and meta-data
    - Collection source, parameters, processing information, etc.
    - Anything that the community feels would be useful
  - Refinement software get their own block
    - Indicate which data in SF used (I vs F, etc)
    - Statistics