# Intersection of the RCSB Protein Data Bank with AI/ML

Stephen K. Burley, M.D., D.Phil.

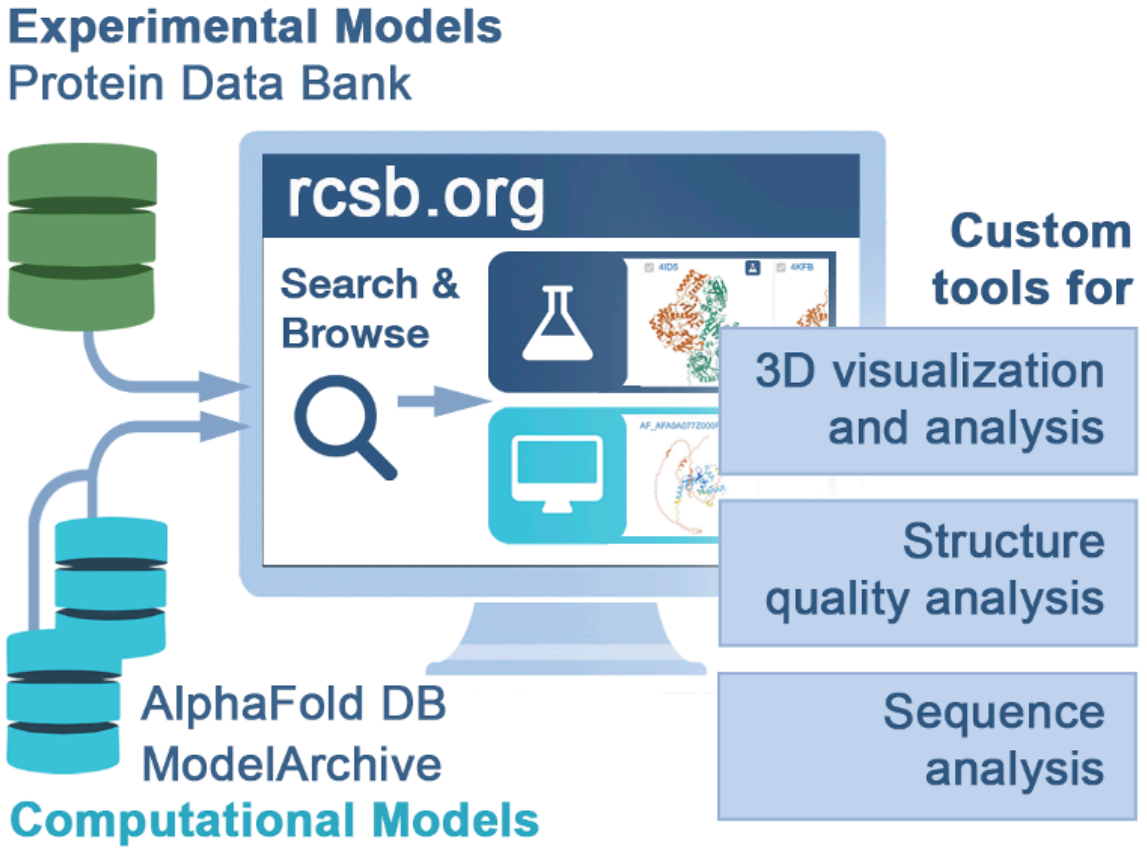
SAS-Chemistry and Chemical Biology/Institute for Quantitative Biomedicine

SKB104@IQB.Rutgers.Edu

RUTGERS–NEW BRUNSWICK
Artificial Intelligence
and Data Science Collaboratory

# RCSB.org Research-focused Web Portal
# One-Stop-Shop for Public 3D Biostructure Data

- RCSB.org delivers >230,000 PDB structures alongside >1 million Computed Structure Models (CSMs) from AlphaFold DB and the ModelArchive
- RCSB.org data exploration and visualization tools used by many millions of researchers, educators, and students worldwide
- Provenance/reliability of both data types are clearly identified




Burley *et al.* (2023) *Nucleic Acids Research 51*, D488–D508.
Burley *et al.* (2025) *Nucleic Acids Research 53*, D564–D574.

# AI-Embedded Structure Search (using NERSC)

- Deep Learning neural network trained to infer vector **embeddings** from structures
- **Vector database** used for searching domains, full-length proteins, assemblies
- Comparable in sensitivity to FoldSeek but much faster (>200M CSMs in seconds)
- PoC @ Embedding-Search.RCSB.org
- Manuscript at bioRxiv
- Full deployment at RCSB.org later this year