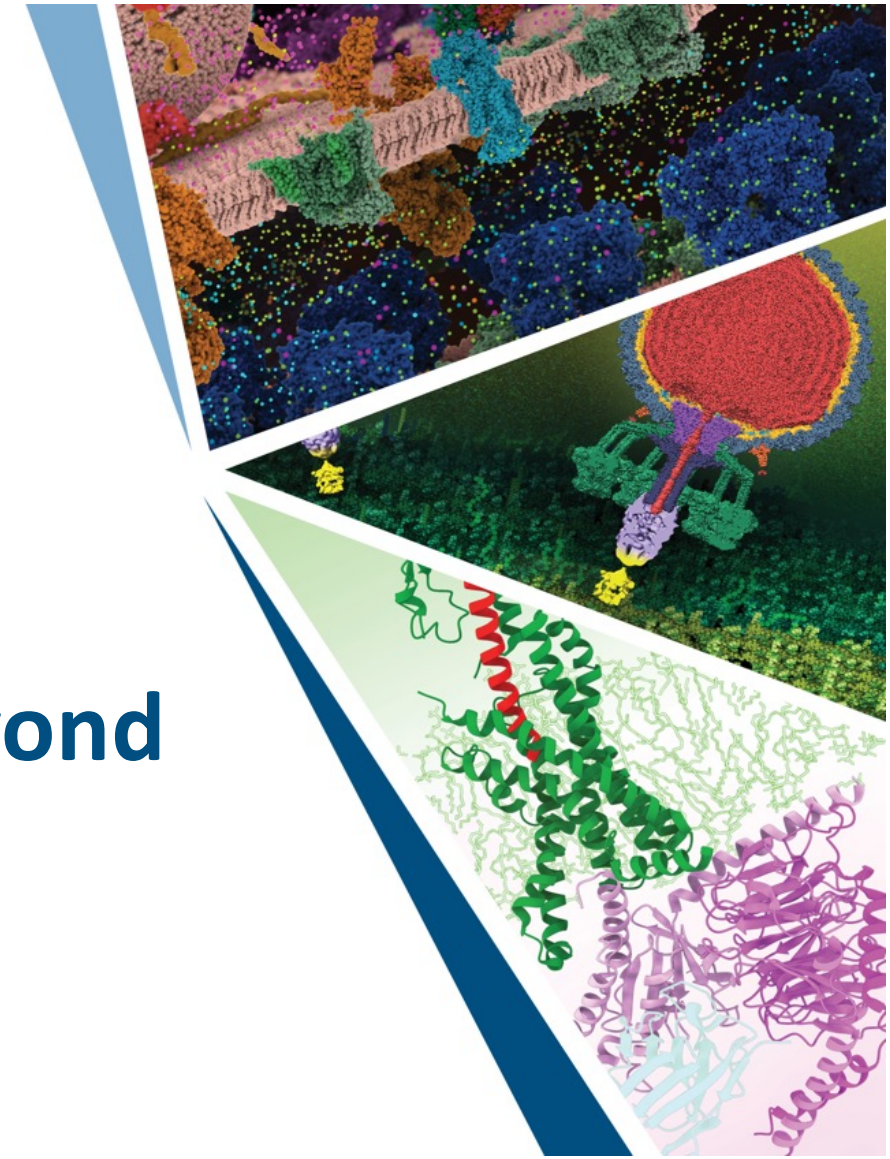
RCSB PDB
PROTEIN DATA BANK

RCSB.org
info@rcsb.org

Data Science and Artificial Intelligence for Biomedicine and Beyond



Stephen K. Burley, M.D., D.Phil.
University Professor and Henry Rutgers Chair
Director, RCSB Protein Data Bank
Interim Director, Rutgers Artificial Intelligence and Data Science (RAD) Collaboratory
Founding Director, Institute for Quantitative Biomedicine
Tenured Member, Department of Chemistry & Chemical Biology
Cancer Pharmacology Research Program Co-Lead, Rutgers Cancer Institute
Rutgers, The State University of New Jersey

Rutgers American Medical Students Association 17th Annual Pre-Health Conference February 9th 2025

Disclosures

- Member, Harrington Discovery Institute Investment Advisory Board
- Member, *Cell Press Structure* Editorial Advisory Board
- Member, *Nature Scientific Data* Editorial Board
- Member, *Nature Oncogene* Editorial Advisory Board
- Advisor, Ligo Analytics, Inc.
- Member, Vincere Biosciences, Inc., Scientific Advisory Board
- Consultant, HanAll Biopharma

Outline

- My Journey: 4 Careers and Counting
- Data Science: Protein Data Bank (a.k.a. Rutgers' Best Kept Secret!)
- Artificial Intelligence/Machine Learning: *De Novo* Protein Structure Prediction
- Impact of PDB Data and Computed Structure Models on Basic and Applied Research in Biology and Medicine
- What do we mean by the term Artificial Intelligence?
- Rutgers Artificial Intelligence and Data Science (RAD) Collaboratory
- Artificial Intelligence/Machine Learning Applications in Biomedicine
- Prepare for a Medical Career in "The Era of Data Science and Artificial Intelligence/Machine Learning"
- Acknowledgments

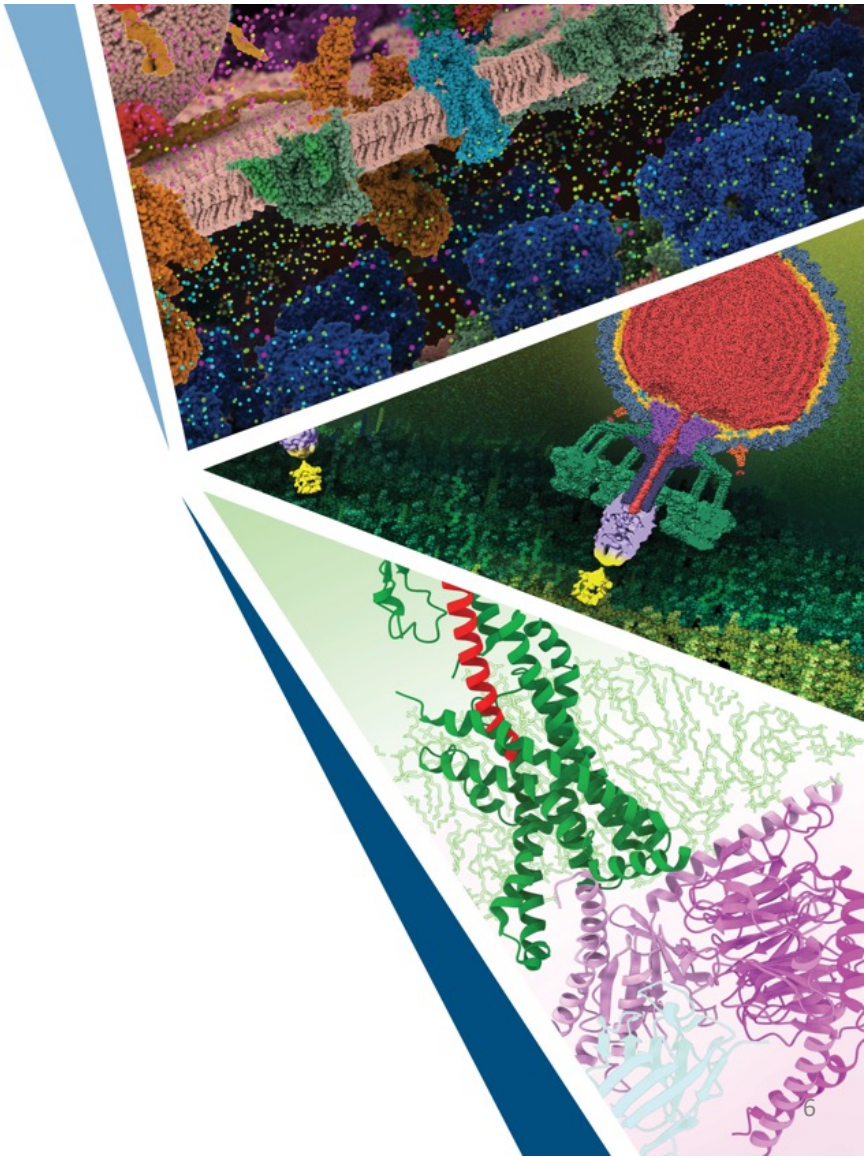
My Journey: 4 Careers and Counting – Part I

- Brought up in United Kingdom, Australia, and Canada
- B.Sc. In Physics and Mathematics, University of Western Ontario, Canada — 1976-1980
- D.Phil. In Structural Biology, Oxford University, United Kingdom — 1980-1983
- M.D., Harvard-MIT Health Sciences and Technology Program — 1983-1987
- Internship/Residency Medicine, Brigham and Women's Hospital — 1987-1990
- Post-Doctoral Fellow, Harvard Chemistry — 1987-1990
- Chaired Professor, The Rockefeller University and Investigator, Howard Hughes Medical Institute — 1990-2002
- Chief Scientific Officer, SGX Pharmaceuticals, Inc. — 2002-2008
- NASDAQ Initial Public Offering (SGXP) — 2006
- Distinguished Lilly Scholar, Eli Lilly and Company — 2008-2012
- **Retirement No. 1 — 2012 FAILED!**

My Journey: 4 Careers and Counting – Part II

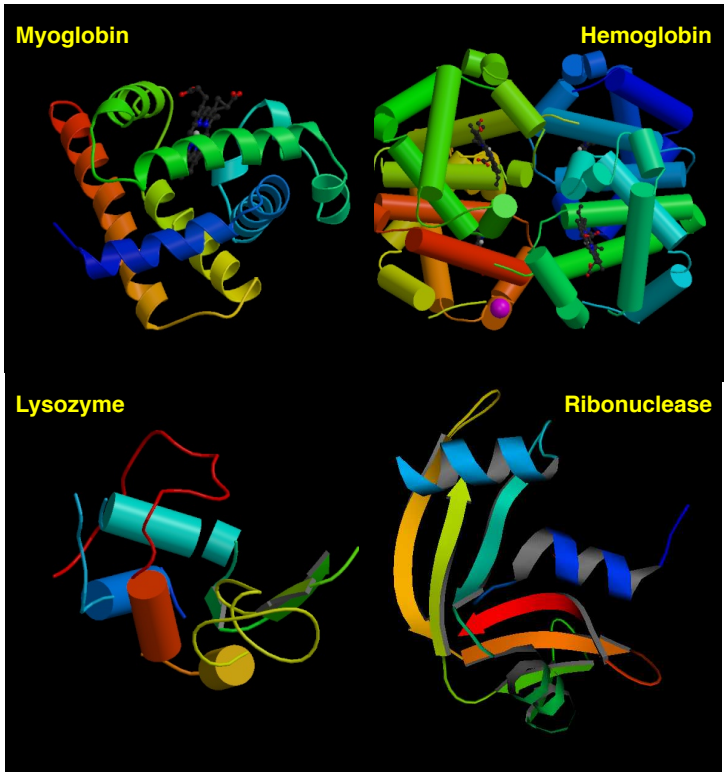
- **Saved by Rutgers, The State University of New Jersey — 2013**
- Distinguished Professor, Department of Chemistry & Chemical Biology — 2013-2017
- Director, Center for Integrative Proteomics Research — 2013-2018
- Member and Cancer Pharmacology Co-Lead, Rutgers Cancer Institute — 2013-Present
- Director, RCSB Protein Data Bank — 2014-Present
- Founding Director, Institute for Quantitative Biomedicine — 2015-Present
- University Professor & Henry Rutgers Chair — 2017-Present
- Interim Director, Rutgers AI and Data (RAD) Collaboratory — 2024-Present
  (International Search underway for my Successor)
- Return to Full-Time Research — ASAP-203?
- **Retirement No. 2 — 203? Hopefully, more successful than No. 1**

RCSB Protein Data Bank



Protein Data Bank (Established 1971)

- PDB 1st online Open Access digital data resource in all of biology
- Founded 1971 with 7 protein structures
- Single global **archive** for protein and DNA/RNA experimental structures
- **Open Access >230,000 structures!**
- wwPDB Partnership founded in 2003
- Members: RCSB PDB (US), PDBe (EMBL-EBI), PDBj (Japan), and PDBc (China); plus EMDB (3DEM) and BMRB (NMR)

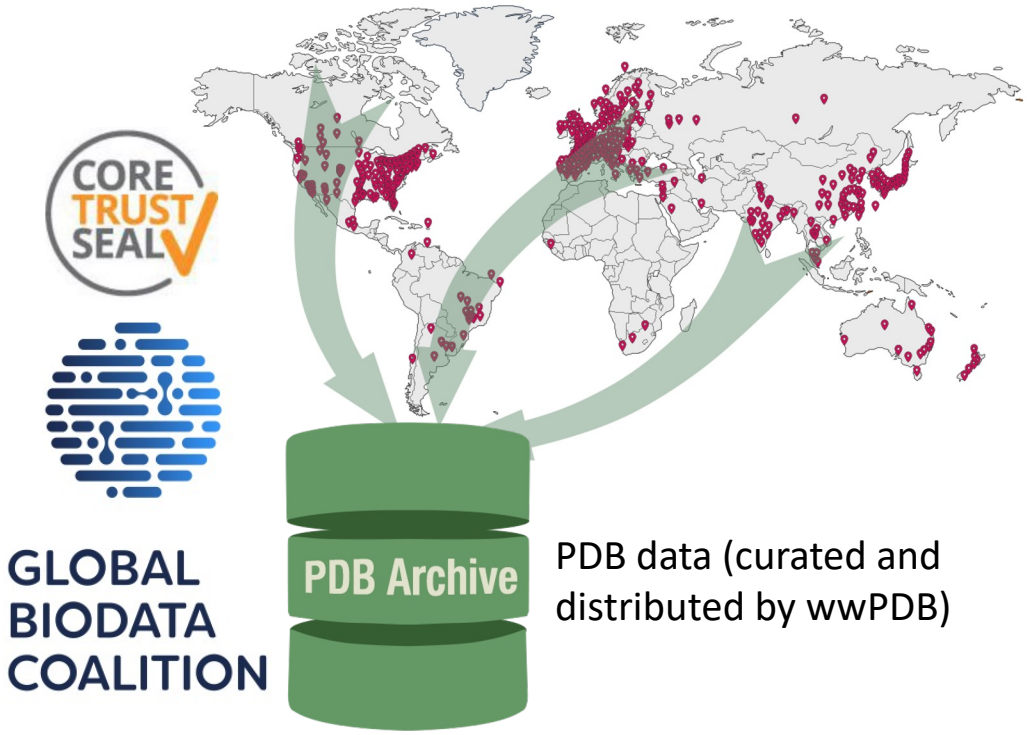


Structures that Inspired Launch of the PDB

Protein Data Bank (1971) *Nature New Biology 233,* 223.
Worldwide Protein Data Bank (2019) *Nucleic Acids Research 47*, D520–D528.

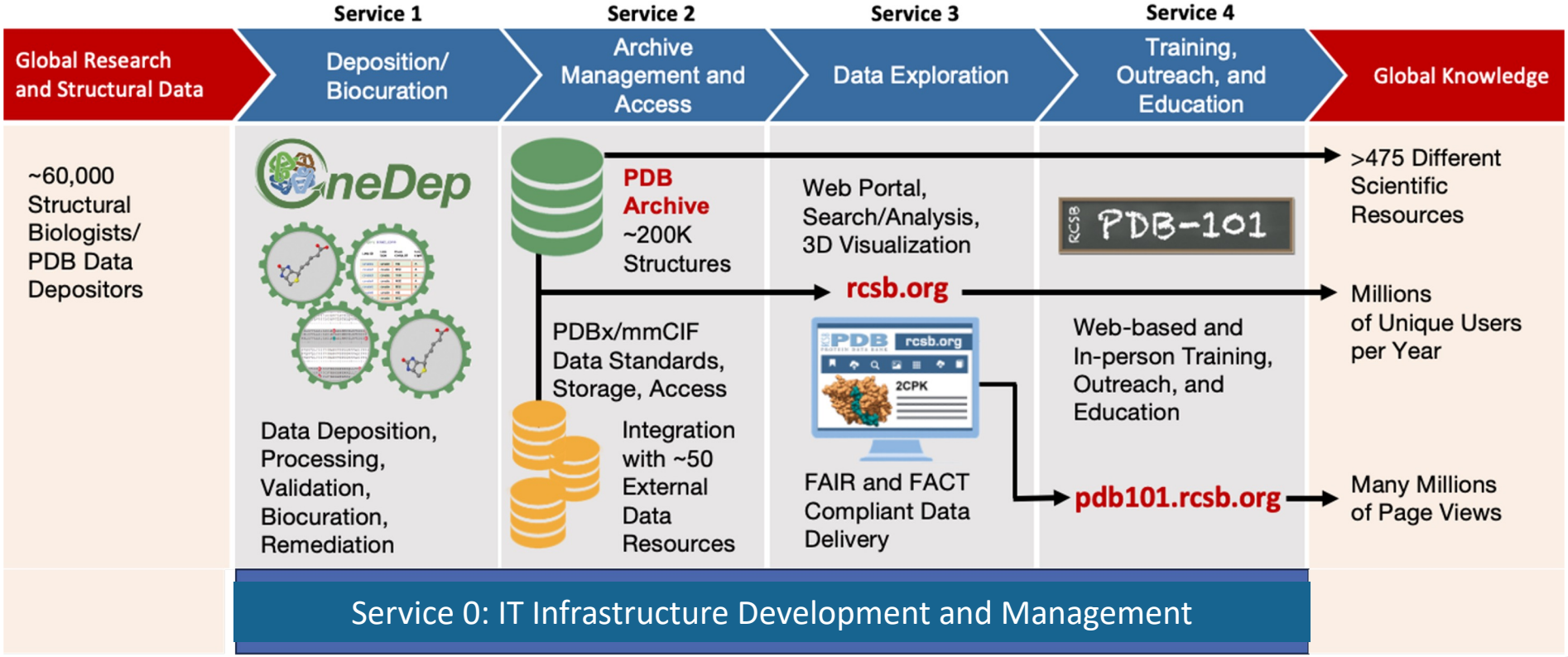
Protein Data Bank is "Crucial for Sustaining the Broader Biodata Infrastructure"

- Single FAIR/FACT compliant global archive providing Open Access to public domain experimental 3D biostructures
- PDB data distributed under the Creative Commons CC0 License
- RCSB Protein Data Bank is the US wwPDB Data Center jointly supported by NSF, NIH, and DOE



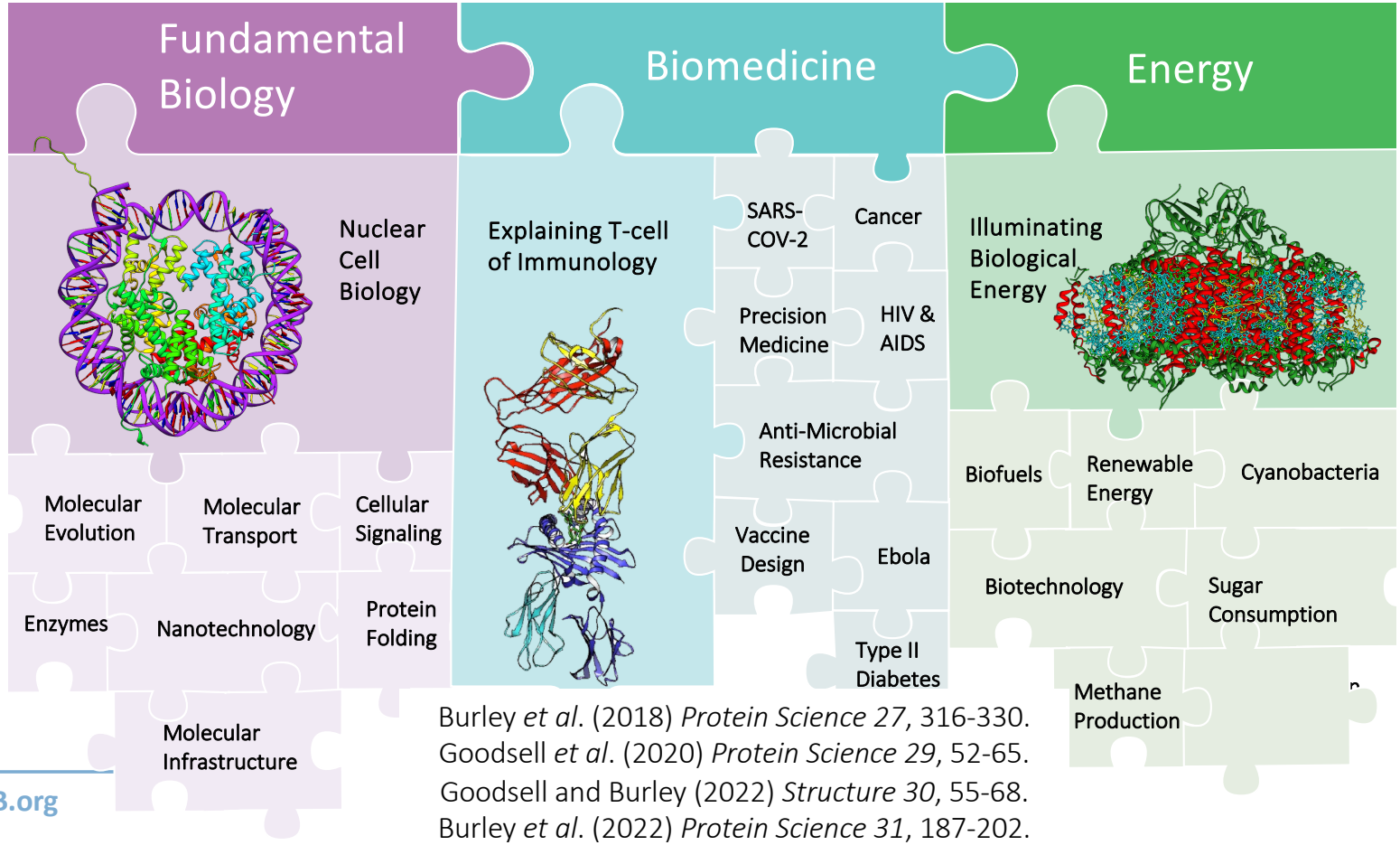
Protein Data Bank (1971) *Nature New Biology 233*, 223.
Worldwide Protein Data Bank (2019) *Nucleic Acids Research 47*, D520–D528.

RCSB Protein Data Bank (RCSB PDB) Services Convert Global Data into Global Knowledge



Burley *et al*. (2022) *Protein Science 31*, E4482.
Burley *et al*. (2025) *Nucleic Acids Research 53*, D564–D574.

RCSB PDB Impact Across the Biosciences



Burley *et al*. (2018) *Protein Science 27*, 316-330.
Goodsell *et al*. (2020) *Protein Science 29*, 52-65.
Goodsell and Burley (2022) *Structure 30*, 55-68.
Burley *et al*. (2022) *Protein Science 31*, 187-202.

Impact: All Therapeutic Areas 2010-2016

- All Therapeutic Areas 2010-2016
  Structure 27, 211-217
- Anti-neoplastic Agents 2010-2018
  Drug Discovery Today 25, 837-850
- Anti-neoplastic Agents 2019-2023
  Nature Oncogene 43, 2229-2243
- Review Article 2021
  Journal of Biological Chemistry 296, 100559

Structure
Perspective

CellPress

How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals

John D. Westbrook[1,*] and Stephen K. Burley[1,2,3,*]
[1]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[2]Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ 08903, USA
[3]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA
*Correspondence: john.westbrook@rcsb.org (J.D.W.), stephen.burley@rcsb.org (S.K.B.)
https://doi.org/10.1016/j.str.2018.11.007

Discovery and development of 210 new molecular entities (NMEs; new drugs) approved by the US Food and Drug Administration 2010–2016 was facilitated by 3D structural information generated by structural biologists worldwide and distributed on an open-access basis by the PDB. The molecular targets for 94% of these NMEs are known. The PDB archive contains 5,914 structures containing one of the known targets and/or a new drug, providing structural coverage for 88% of the recently approved NMEs across all therapeutic areas. More than half of the 5,914 structures were published and made available by the PDB at no charge, with no restrictions on usage >10 years before drug approval. Citation analyses revealed that these 5,914 PDB structures significantly affected the very large body of publicly funded research reported in publications on the NME targets that motivated biopharmaceutical company investment in discovery and development programs that produced the NMEs.

Introduction

The PDB (Berman et al., 2003) is an enormously valuable gold standard, reference data resource for education/training and research (both basic and applied) across the biological and biomedical sciences. It was established in 1971 as the first open-access, digital data resource in biology with just seven protein structures (Protein Data Bank, 1971). Forty-seven years later, the PDB continues to serve as the single global repository for 3D structural data, making >146,000 experimentally determined structures (of proteins, DNA, RNA, and their complexes with drugs and/or other small molecules) freely available without restrictions on usage. Since 2003, the PDB has been managed jointly by the Worldwide PDB (wwPDB) partnership (Berman et al., 2003), including US Research Collaboratory for Structural Bioinformatics (RCSB) PDB (Berman et al., 2000), PDBe in Europe (Velankar et al., 2010), PDB Japan (Kinjo et al., 2017), and BioMagResBank (Ulrich et al., 2008). The wwPDB OneDep global system for deposition-validation-biocuration of incoming structures supports all PDB data depositors, helping to ensure that every structure archived in the PDB is well validated and expertly biocurated (Young et al., 2017, 2018; Gore et al., 2017). wwPDB partners are committed to ensuring adherence to the FAIR (findability, accessibility, interoperability, reusability) (Wilkinson et al., 2016). Publication of new macromolecular structures in most scientific journals is contingent on mandatory deposition to the PDB of the 3D atomic coordinates comprising the structure, together with experimental data and metadata. Many governmental/non-governmental research funders also require PDB deposition of macromolecular structural data.

Over the past two decades, structural biology and structure-guided drug discovery have become well established within the biopharmaceutical industry (Blundell, 2017; Klebe, 2013). 3D structures frequently provide information about how individual small-molecule ligands bind to their target proteins (e.g., imatinib in chronic myeloid leukemia; Capdeville et al., 2002; Nagar et al., 2002). Structural data have also proved useful in overcoming some of the myriad challenges inherent in turning biochemically active compounds into potent drug-like molecules suitable for safety and efficacy testing in animals and humans (Stoll et al., 2011). In the realm of biologics (~20% of approved drugs over the past decade; Mullard, 2016), 3D structural information is now routinely being used to drive engineering of monoclonal antibodies (Gillliland et al., 2012).

Previously published case studies (e.g., Hu et al., 2018) and largely anecdotal reports presented at scientific meetings leave no doubt as to the importance of individual contributions to drug discovery made by macromolecular crystallographers working in industry, but they do not address the impact on the biopharmaceutical industry by public-sector structural biologists and the data they contribute to the PDB archive. We, therefore, undertook a quantitative assessment of the impact of structural biologists and the PDB on the discovery and development of 210 New Medical Entities (NMEs or drugs) approved by the US Food and Drug Administration (FDA) between 2010 and 2016.

Results and Discussion

Overview

We analyzed PDB archival holdings and identified 3D structures that include the 210 NMEs and/or their 150 known molecular targets (Table 1). The 210 NMEs break down into three classes: small chemicals of molecular weight <1,000 (low-molecular-weight [LMW]-NMEs: 81.4%, 171/210), proteins or peptide

Westbrook & Burley (2019) *Structure 27*, 211-217

Impact: All Therapeutic Areas 2010-2016[1]

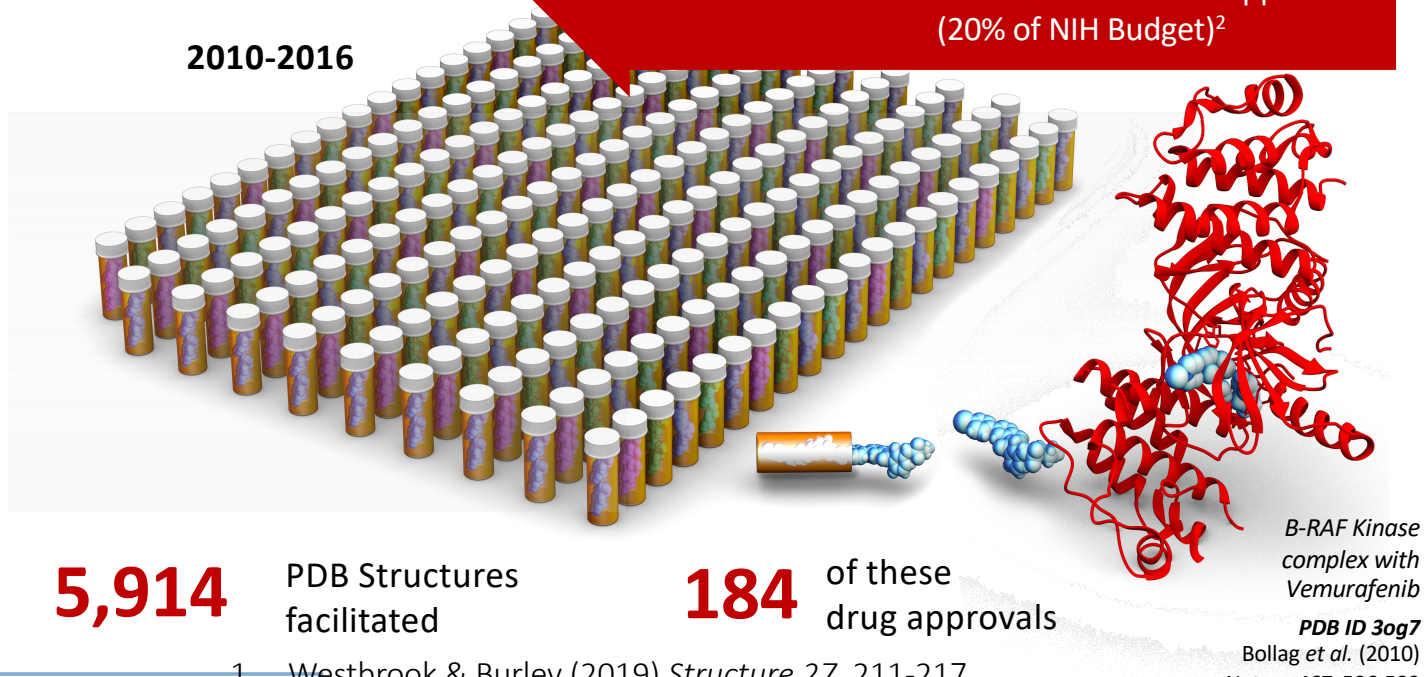
210 **NEW DRUGS**

approved

2010-2016

2000-2016

~$100 BILLION **of NIH funding**
contributed to these approvals
(20% of NIH Budget)[2]



B-RAF Kinase complex with Vemurafenib

PDB ID 3og7
Bollag *et al.* (2010)
Nature 467, 596-599

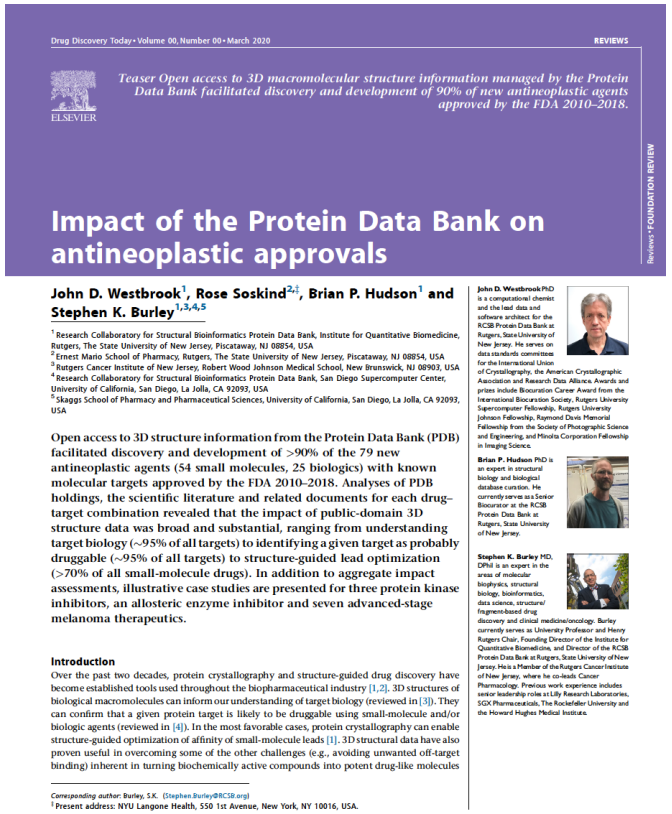
5,914 PDB Structures facilitated

184 of these drug approvals

1. Westbrook & Burley (2019) *Structure 27*, 211-217
2. Galkina Cleary *et al.* (2018) *PNAS 115,* 2329-2334; Value in 2016 US$

Impact: Oncology Indications 2010-2018

- All Therapeutic Areas 2010-2016
  Structure 27, 211-217
- Anti-neoplastic Agents 2010-2018
  Drug Discovery Today 25, 837-850
- Anti-neoplastic Agents 2019-2023
  Nature Oncogene 43, 2229-2243
- Review Article 2021
  Journal of Biological Chemistry 296, 100559



Drug Discovery Today • Volume 00, Number 00 • March 2020 REVIEWS

Teaser Open access to 3D macromolecular structure information managed by the Protein Data Bank facilitated discovery and development of 90% of new antineoplastic agents approved by the FDA 2010–2018.

Impact of the Protein Data Bank on antineoplastic approvals

John D. Westbrook[1], Rose Soskind[2,†], Brian P. Hudson[1] and Stephen K. Burley[1,3,4,5]

[1] Research Collaboratory for Structural Bioinformatics Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[2] Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[3] Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ 08903, USA
[4] Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA
[5] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA

Open access to 3D structure information from the Protein Data Bank (PDB) facilitated discovery and development of >90% of the 79 new antineoplastic agents (54 small molecules, 25 biologics) with known molecular targets approved by the FDA 2010–2018. Analyses of PDB holdings, the scientific literature and related documents for each drug–target combination revealed that the impact of public-domain 3D structure data was broad and substantial, ranging from understanding target biology (~95% of all targets) to identifying a given target as probably druggable (~95% of all targets) to structure-guided lead optimization (>70% of all small-molecule drugs). In addition to aggregate impact assessments, illustrative case studies are presented for three protein kinase inhibitors, an allosteric enzyme inhibitor and seven advanced-stage melanoma therapeutics.

Introduction

Over the past two decades, protein crystallography and structure-guided drug discovery have become established tools used throughout the biopharmaceutical industry [1,2]. 3D structures of biological macromolecules can inform our understanding of target biology (reviewed in [3]). They can confirm that a given protein target is likely to be druggable using small-molecule and/or biologic agents (reviewed in [4]). In the most favorable cases, protein crystallography can enable structure-guided optimization of affinity of small-molecule leads [1]. 3D structural data have also proven useful in overcoming some of the other challenges (e.g., avoiding unwanted off-target binding) inherent in turning biochemically active compounds into potent drug-like molecules

Corresponding author: Burley, S.K. (Stephen.Burley@RCSB.org)
† Present address: NYU Langone Health, 550 1st Avenue, New York, NY 10016, USA.

Westbrook et al. (2020) *Drug Discovery Today 25*, 837-850

Impact: Oncology Indications 2010-2018[1]

79 **NEW ANTI-CANCER DRUGS** Approved 2010-2018

74 **of these new drugs**

had a
total of **2412**

unique structures in the PDB
explaining target biology and
facilitating discovery/development

Structure-guided drug discovery → >70% of small-molecule drugs

1. Westbrook *et al*. (2020) *Drug Discovery Today 25*, 837-850.

Impact: Oncology Indications 2019-2023

- All Therapeutic Areas 2010-2016
  Structure 27, 211-217
- Anti-neoplastic Agents 2010-2018
  Drug Discovery Today 25, 837-850
- Anti-neoplastic Agents 2019-2023
  Nature Oncogene 43, 2229-2243
- Review Article 2021
  Journal of Biological Chemistry 296, 100559

Oncogene www.nature.com/onc

REVIEW ARTICLE OPEN

Impact of structural biology and the protein data bank on us fda new drug approvals of low molecular weight antineoplastic agents 2019–2023

Stephen K. Burley[1,2,3,4✉], Amy Wu-Wu[1], Shuchismita Dutta[1,2], Shridar Ganesan[2] and Steven X. F. Zheng[2]

Open access to three-dimensional atomic-level biostructure information from the Protein Data Bank (PDB) facilitated discovery/development of 100% of the 34 new low molecular weight, protein-targeted, antineoplastic agents approved by the US FDA 2019–2023. Analyses of PDB holdings, the scientific literature, and related documents for each drug-target combination revealed that the impact of structural biologists and public-domain 3D biostructure data was broad and substantial, ranging from understanding target biology (100% of all drug targets), to identifying a given target as likely druggable (100% of all targets), to structure-guided drug discovery (>80% of all new small-molecule drugs, made up of 50% confirmed and >30% probable cases). In addition to aggregate impact assessments, illustrative case studies are presented for six first-in-class small-molecule anti-cancer drugs, including a selective inhibitor of nuclear export targeting Exportin 1 (selinexor, Xpovio), an ATP-competitive CSF-1R receptor tyrosine kinase inhibitor (pexidartinib,Turalio), a non-ATP-competitive inhibitor of the BCR-Abl fusion protein targeting the myristoyl binding pocket within the kinase catalytic domain of Abl (asciminib, Scemblix), a covalently-acting G12C KRAS inhibitor (sotorasib, Lumakras or Lumykras), an EZH2 methyltransferase inhibitor (tazemostat, Tazverik), and an agent targeting the basic-Helix-Loop-Helix transcription factor HIF-2α (belzutifan, Welireg).

Oncogene (2024) 43:2229–2243; https://doi.org/10.1038/s41388-024-03077-2

INTRODUCTION

X-ray protein crystallography and structure-guided approaches have been mainstays for drug discovery for more than two decades [1, 2]. Atomic-level, three-dimensional (3D) structures of biological macromolecules inform our understanding of target biology (reviewed in [3]), and provide important insights into target druggability for both small-molecule and/or biologic agents (reviewed in [4]). Today, macromolecular crystallography (MX) and 3D electron microscopy (3DEM) are routinely used in most large and many small biopharmaceutical companies for structure-guided optimization of affinity of small-molecule screening hits and lead compounds [1]. 3D biostructure data can also aid in surmounting some of the myriad challenges (e.g., avoiding unwanted off-target binding) inherent in turning biochemically active compounds into potent, drug-like molecules suitable for safety and efficacy testing in animals and humans [5]. Finally, starting points for medicinal chemistry campaigns (i.e., selectively binding chemical scaffolds) can be identified via fragment screening using nuclear magnetic resonance spectroscopy or NMR [6], MX [7], and 3DEM [8].

Public-domain 3D biostructure information generated using MX, 3DEM, or NMR is distributed on an open-access basis by a singular global data resource, known as the Protein Data Bank (PDB [9]). When the PDB was established in 1971 as the first open-access digital data resource in biology, it housed only seven protein structures [9]. Today, the PDB is regarded as a global public good vital to basic and applied research and education/training across the biological and biomedical sciences. In the spring of 2024, the PDB housed >220,000 experimentally determined, atomic-level 3D structures of biological macromolecules (i.e., proteins, DNA, and RNA), many of which have been visualized in the act of binding one or more small-molecule ligands, including United States Food and Drug Administration (US FDA) approved drugs. Since 2003, the PDB has been managed jointly according to the FAIR Principles of Findability-Accessibility-Interoperability-Reusability [10] by the Worldwide Protein Data Bank (wwPDB) partnership [11, 12], including the US Research Collaboratory for Structural Bioinformatics Protein Data Bank or RCSB PDB [13–15], Protein Data Bank in Europe [16], Protein Data Bank Japan [17], Protein Data Bank China [18], Biological Magnetic Resonance Data Bank [19], and the Electron Microscopy Data Bank [20].

The RCSB PDB (RCSB.org) headquartered at Rutgers, The State University of New Jersey (with additional performance sites at the University of California San Diego and the University of California San Francisco) serves as the US wwPDB data center and as the wwPDB-designated Archive Keeper for the PDB. On two previous occasions, we have reviewed the impact of structural biologists and PDB

[1]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. [2]Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ 08903, USA. [3]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA. [4]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. ✉email: Stephen.Burley@RCSB.org

Received: 28 March 2024 Revised: 4 June 2024 Accepted: 5 June 2024
Published online: 17 June 2024

SPRINGER NATURE

Burley et al. (2024) *Nature Oncogene 43*, 2229-2243

Impact: LMW*-NMEs Oncology Targets 2010-2023

Area of Impact	Fraction of NMEs Impacted
• Target Biology	• 100% (88/88)
• Target Druggability	• 100% (88/88)
• Structure-Guided Drug Discovery	• 75% (66/88)
• Structure of NME+Target in PDB	• >70% (63/88)
• ADME^/Safety Optimization	• Cytochrome P450s, P-glycoprotein, Herg Channel, *etc*.

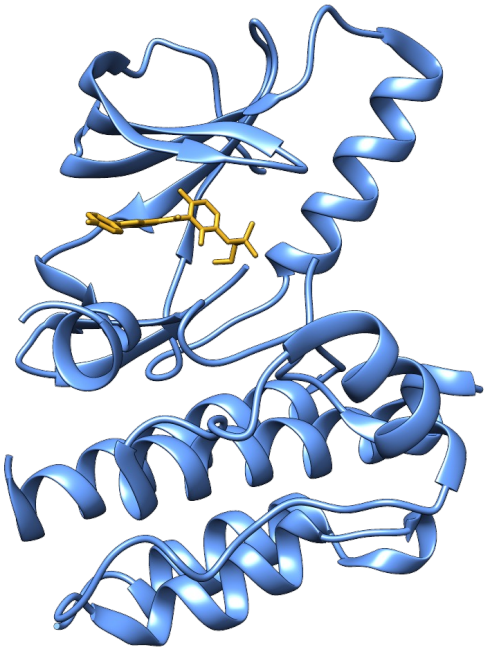
 * LMW=Low Molecular Weight (<1000 Da); ^ Absorption-Distribution-Metabolism-Excretion

BRAF Structure-Guided Drug Discovery

BRAF History Synopsis

- 2004: PDB ID 1uwh (Academic)
  1st BRAF structure in PDB
  Cancer causing mutations in catalytic domain
- 2007: V600E BRAF occurs in ~50% late-stage melanomas
- 2010: PDB ID 3og7 (Plexxikon)
  V600E BRAF-Vemurafenib co-crystal structure
- 2011: Vemurafenib approved by US FDA
- Acquired resistance to vemurafenib limited it utility as a single agent treatment for late-stage melanoma

US FDA New Drug Approval 2011



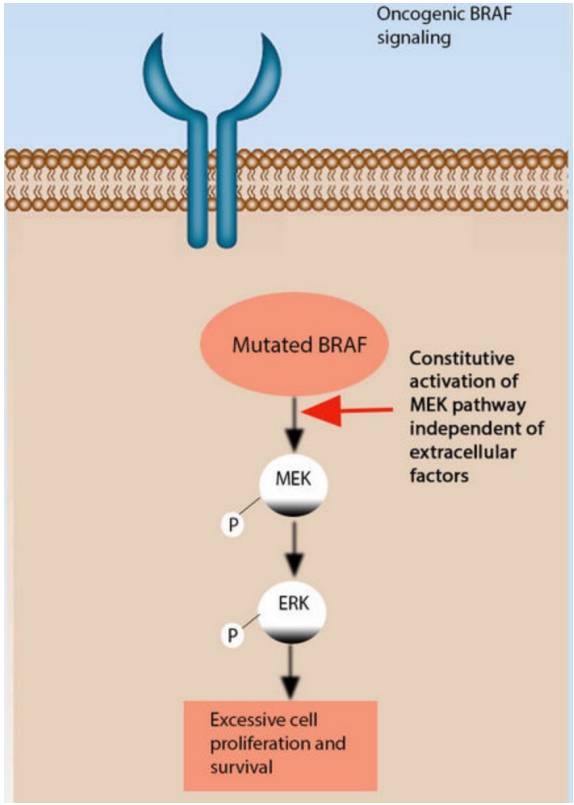
PDB ID 3og7

Bollag *et al*. (2010) *Nature 467*, 596-599.

BRAF/MEK Combinations for Late-stage Melanoma

- BRAF and MEK occur in the same signaling pathway
- V600E BRAF targeted by Plexxikon*, GSK*, and Novartis*
- BRAF inhibition can be overcome by MEK upregulation
- MEK targeted by Japan Tobacco, Array Biopharma*, and Exelixis*
- BRAF and MEK inhibitors are now used in combination

* Structure-Guided Drug Discovery

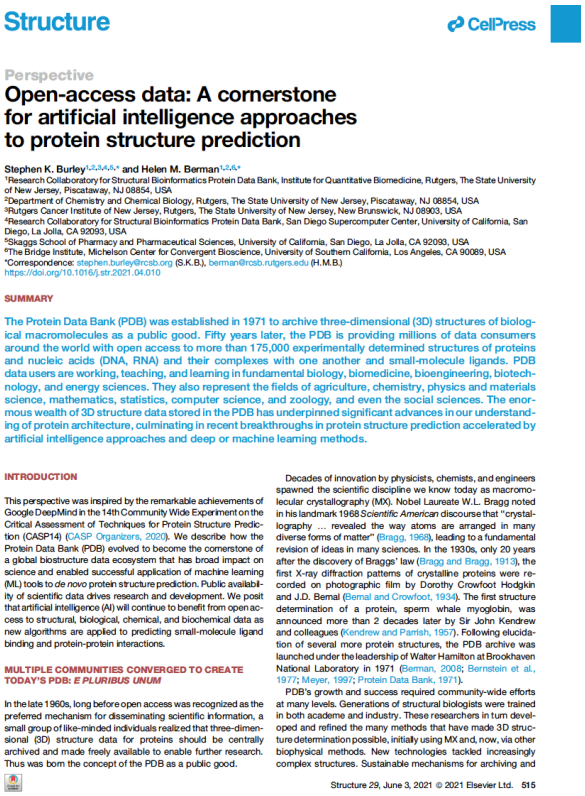


Maraka & Janku (2018) *Discov Med 26*, 51-60.

AI/ML Computed Structure Modeling

- Open access PDB enabled
  - Homology modeling using "templates" from the PDB
  - *De novo* protein structure prediction (template free)
- Machine Learning approaches use PDB plus genomic sequence information
  - AlphaFold2 (Google DeepMind)*
  - RoseTTAFold2 (Baker, UWashington)*
  - OpenFold (AlQuraishi, Columbia U)
- Reasonable accuracy computed structure models (CSMs) now accessible for nearly every protein sequence in UniProt

* **2024 Nobel Prize in Chemistry**



Structure

CellPress

Perspective

Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction

Stephen K. Burley[1,2,3,4,5,*] and Helen M. Berman[1,2,6,*]
[1]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[2]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[3]Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, NJ 08903, USA
[4]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA
[5]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA
[6]The Bridge Institute, Michelson Center for Convergent Bioscience, University of Southern California, Los Angeles, CA 90089, USA
*Correspondence: stephen.burley@rcsb.org (S.K.B.), berman@rcsb.rutgers.edu (H.M.B.)
https://doi.org/10.1016/j.str.2021.04.010

SUMMARY

The Protein Data Bank (PDB) was established in 1971 to archive three-dimensional (3D) structures of biological macromolecules as a public good. Fifty years later, the PDB is providing millions of data consumers around the world with open access to more than 175,000 experimentally determined structures of proteins and nucleic acids (DNA, RNA) and their complexes with one another and small-molecule ligands. PDB data users are working, teaching, and learning in fundamental biology, biomedicine, bioengineering, biotechnology, and energy sciences. They also represent the fields of agriculture, chemistry, physics and materials science, mathematics, statistics, computer science, and zoology, and even the social sciences. The enormous wealth of 3D structure data stored in the PDB has underpinned significant advances in our understanding of protein architecture, culminating in recent breakthroughs in protein structure prediction accelerated by artificial intelligence approaches and deep or machine learning methods.

INTRODUCTION

This perspective was inspired by the remarkable achievements of Google DeepMind in the 14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP14) (CASP Organizers, 2020). We describe how the Protein Data Bank (PDB) evolved to become the cornerstone of a global biostructure data ecosystem that has broad impact on science and enabled successful application of machine learning (ML) tools to *de novo* protein structure prediction. Public availability of scientific data drives research and development. We posit that artificial intelligence (AI) will continue to benefit from open access to structural, biological, chemical, and biochemical data as new algorithms are applied to predicting small-molecule ligand binding and protein-protein interactions.

MULTIPLE COMMUNITIES CONVERGED TO CREATE TODAY'S PDB: *E PLURIBUS UNUM*

In the late 1960s, long before open access was recognized as the preferred mechanism for disseminating scientific information, a small group of like-minded individuals realized that three-dimensional (3D) structure data for proteins should be centrally archived and made freely available to enable further research. Thus was born the concept of the PDB as a public good.

Decades of innovation by physicists, chemists, and engineers spawned the scientific discipline we know today as macromolecular crystallography (MX). Nobel Laureate W.L. Bragg noted in his landmark 1968 *Scientific American* discourse that "crystallography … revealed the way atoms are arranged in many diverse forms of matter" (Bragg, 1968), leading to a fundamental revision of ideas in many sciences. In the 1930s, only 20 years after the discovery of Braggs' law (Bragg and Bragg, 1913), the first X-ray diffraction patterns of crystalline proteins were recorded on photographic film by Dorothy Crowfoot Hodgkin and J.D. Bernal (Bernal and Crowfoot, 1934). The first structure determination of a protein, sperm whale myoglobin, was announced more than 2 decades later by Sir John Kendrew and colleagues (Kendrew and Parrish, 1957). Following elucidation of several more protein structures, the PDB archive was launched under the leadership of Walter Hamilton at Brookhaven National Laboratory in 1971 (Berman, 2008; Bernstein et al., 1977; Meyer, 1997; Protein Data Bank, 1971).

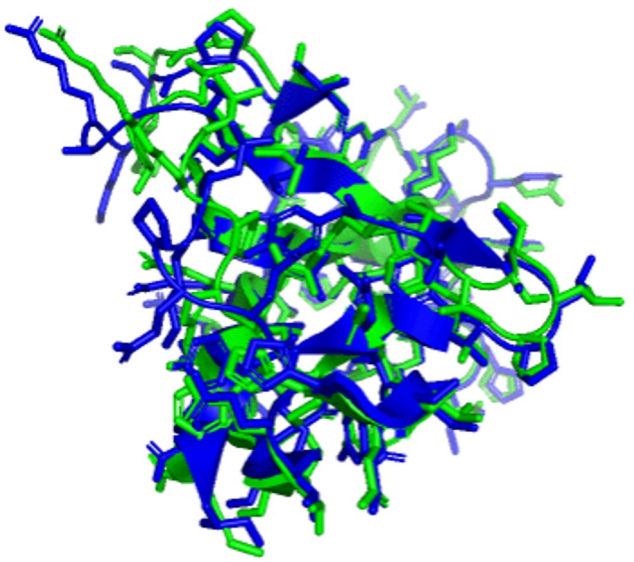
PDB's growth and success required community-wide efforts at many levels. Generations of structural biologists were trained in both academe and industry. These researchers in turn developed and refined the many methods that have made 3D structure determination possible, initially using MX and, now, via other biophysical methods. New technologies tackled increasingly complex structures. Sustainable mechanisms for archiving and

Burley & Berman (2021) *Structure 29*, 515-519.

AI/ML Computed Structure Modeling

- Open access PDB enabled
  - Homology modeling using "templates" from the PDB
  - *De novo* protein structure prediction (template free)
- Machine Learning approaches use PDB plus genomic sequence information
  - AlphaFold2 (Google DeepMind)*
  - RoseTTAFold2 (Baker, UWashington)*
  - OpenFold (AlQuraishi, Columbia U)
- Reasonable accuracy computed structure models (CSMs) now accessible for nearly every protein sequence in UniProt

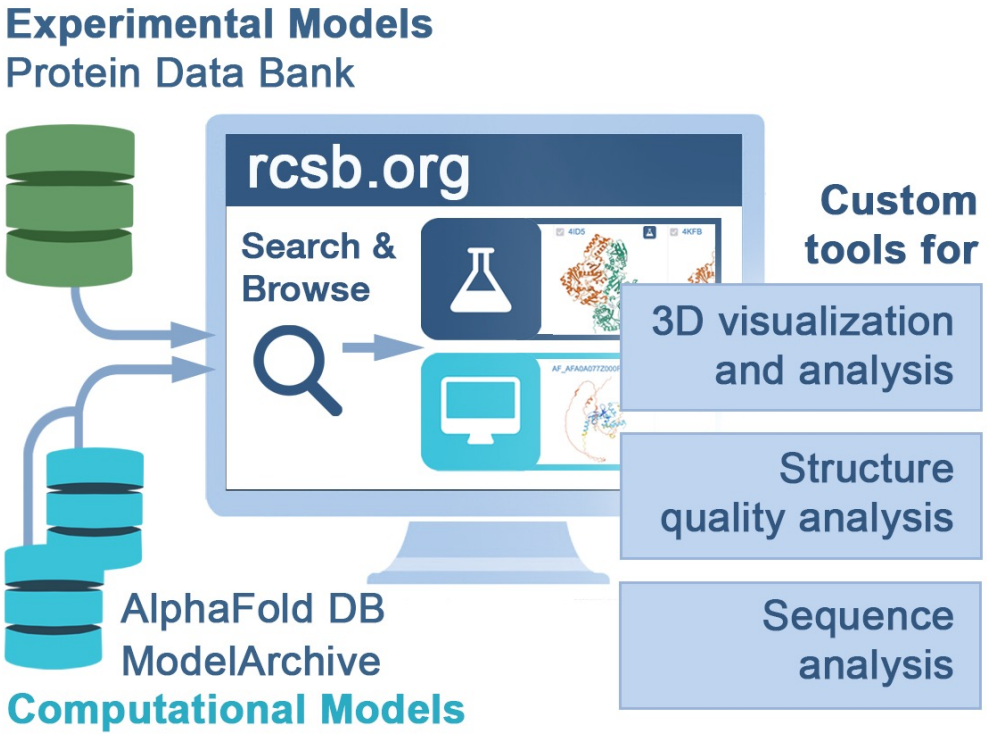
* **2024 Nobel Prize in Chemistry**



Experiment
Prediction

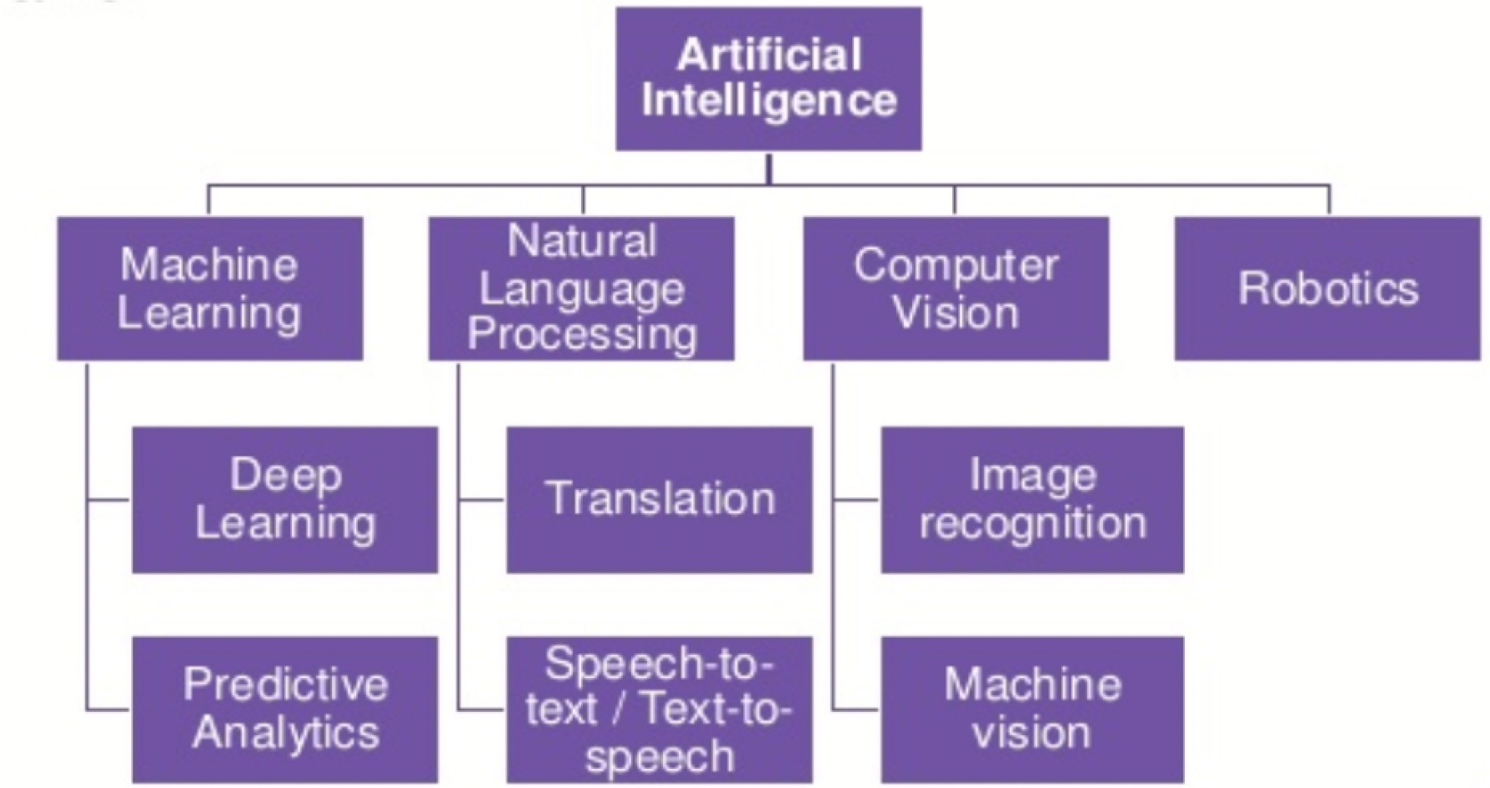
RCSB.org Research-focused Web Portal: One-Stop-Shop for Public 3D Biostructure Data

- RCSB.org delivers
  - >220,000 PDB structures
  - >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



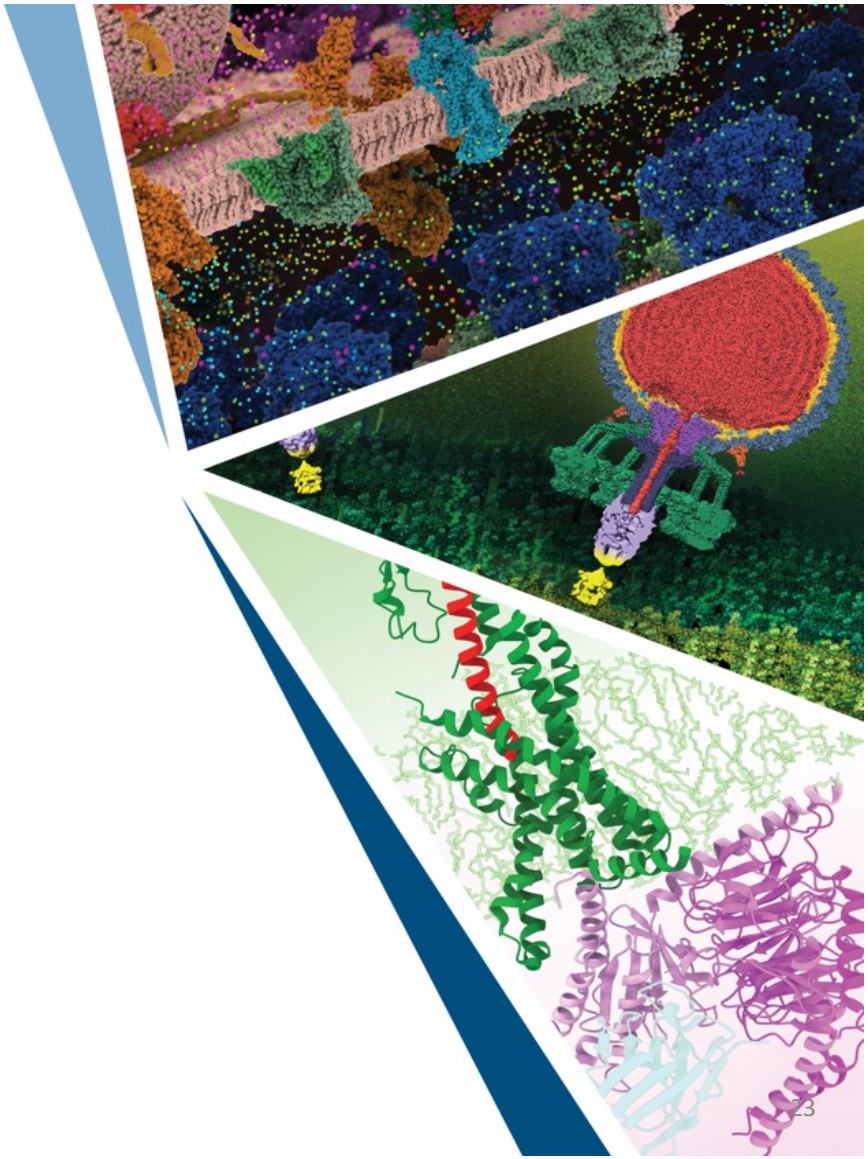
Burley *et al*. (2022) *Protein Science 31*, E4482.
Burley *et al*. (2025) *Nucleic Acids Research 53*, D564–D574.

What do we mean by Artificial Intelligence?



RADCollaboratory.Rutgers.edu

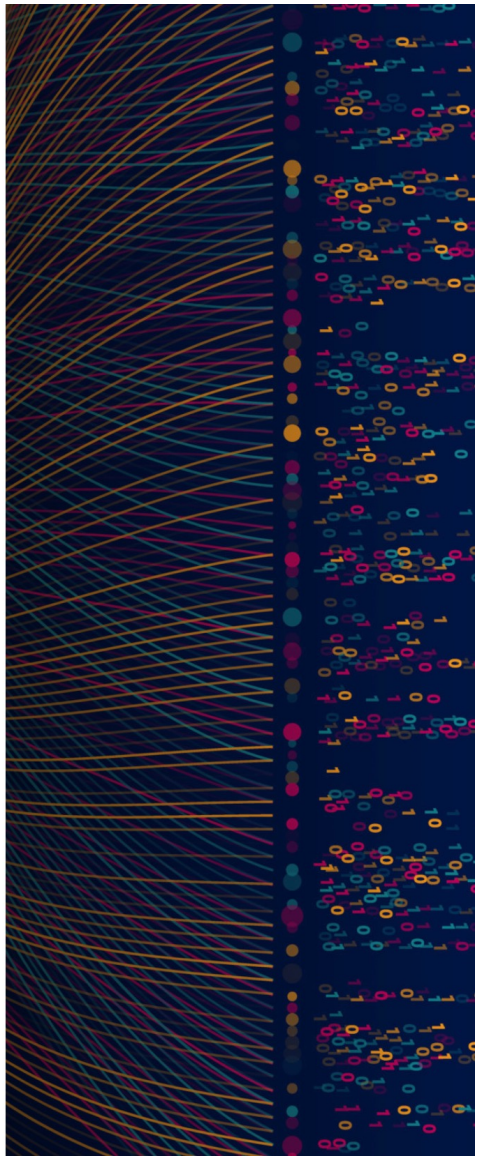
RAD Collaboratory



RUTGERS–NEW BRUNSWICK
Artificial Intelligence
and Data Science Collaboratory

RADCollaboratory.Rutgers.edu

RAD Collaboratory Inaugural Networking Event January 22nd 2025



Stephen K. Burley, M.D., D.Phil.
University Professor and Henry Rutgers Chair
Interim Director, RAD Collaboratory
Director, RCSB Protein Data Bank
Founding Director, Institute for Quantitative Biomedicine
Department of Chemistry & Chemical Biology
Rutgers Cancer Institute
Rutgers, The State University of New Jersey

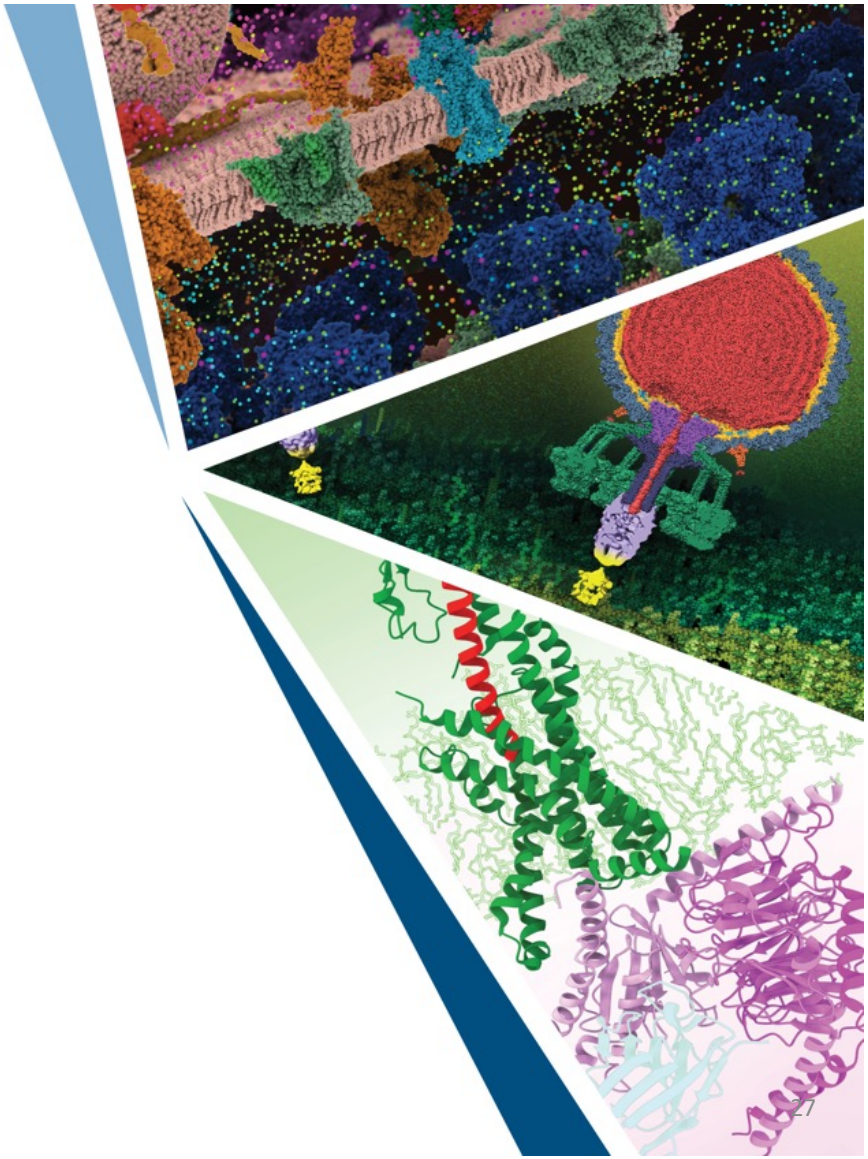
What is the RAD Collaboratory?

The RAD Collaboratory was established as a New Brunswick Chancellor-reporting Signature Initiative to serve as a hub that fosters Data Science and Artificial Intelligence/Machine Learning (AI/ML) research collaborations, student and postdoctoral programming, and community engagement in synergistic partnerships with related initiatives across the University.

What will the RAD do in Spring 2025?

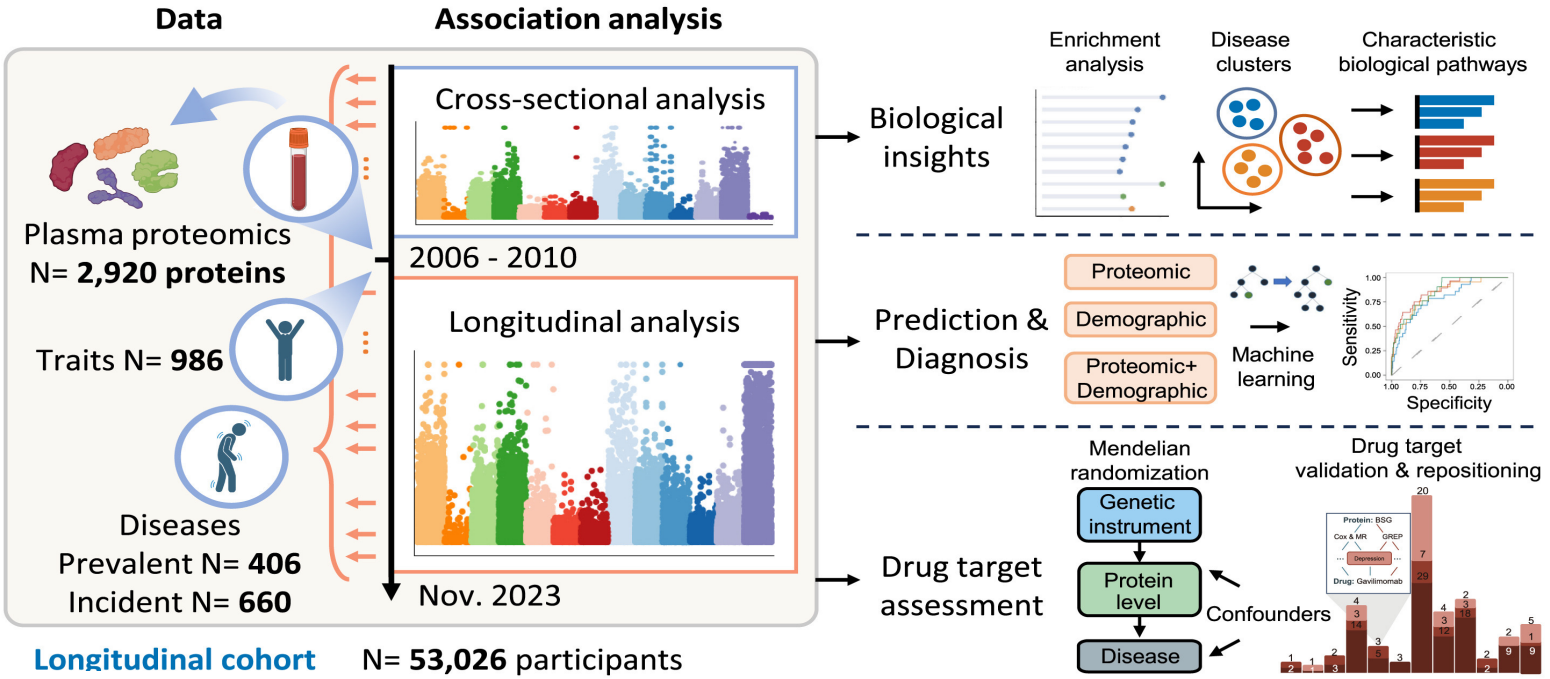
- Host Hands-on Training in AI/DL Tutorials: Save-the-Date in your packet
- Recruit a RAD Research Programmer: Rutgers Posting No. 25FA0062
  Supporting RAD Research Teams with cyberinfrastructure and AI/ML expertise to develop competitive interdisciplinary research programs
- RAD submission of large Multi-PI Federal Grant Applications
  Indirect cost revenues will help support RAD initiatives and PIs
- Award Five RAD Postdoctoral Fellowships: Rutgers Posting No. 25FA00624
- **Establish a RAD Undergraduate Research Program: Applications Early March**
- **Host additional Networking Events : March 5th, April 8th (1-3pm Eastern)**

AI in Biomedicine



AI: ML for Basic Biomedical Research

Atlas of the plasma human proteome in health and disease in 53,026 adults

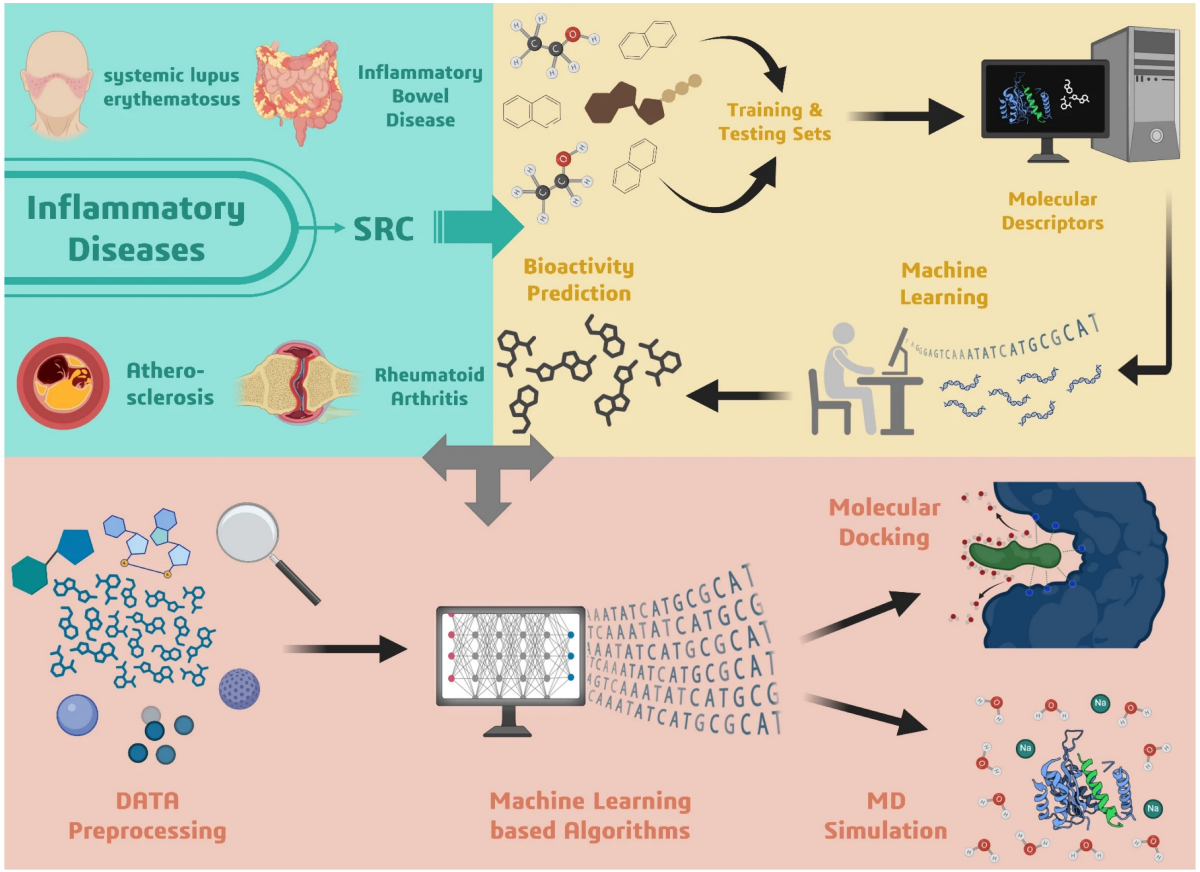


RADCollaboratory.Rutgers.edu

Deng *et al*. (2025) *Cell 188*, 253-372.

AI: ML for Applied Biomedical Research

Integrating machine learning and structure-based approaches for repurposing Src protein tyrosine kinase inhibitors for inflammation

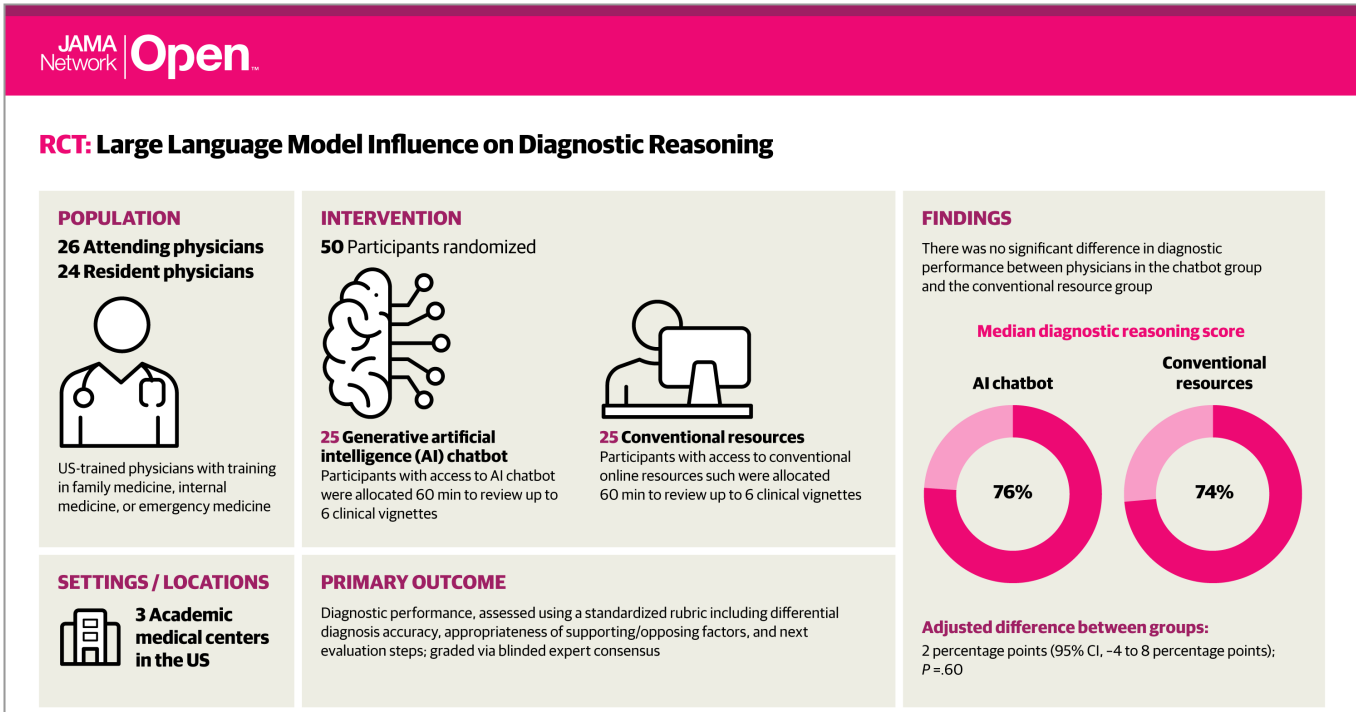


Iqbal *et al*. (2025) *Scientific Reports 15*, 1836.

AI: Natural Language Processing in Clinical Practice (50 Internists were not ready in 2024!)

CHAT GPT 4.0 assistance gave no meaningful improvement in diagnostic reasoning score for 50 MDs (76% *versus* 74%)

But CHAT GPT 4.0 alone scored 90%!



JAMA Network Open

RCT: Large Language Model Influence on Diagnostic Reasoning

POPULATION

26 Attending physicians
24 Resident physicians

US-trained physicians with training in family medicine, internal medicine, or emergency medicine

INTERVENTION

50 Participants randomized

25 Generative artificial intelligence (AI) chatbot
Participants with access to AI chatbot were allocated 60 min to review up to 6 clinical vignettes

25 Conventional resources
Participants with access to conventional online resources such were allocated 60 min to review up to 6 clinical vignettes

FINDINGS

There was no significant difference in diagnostic performance between physicians in the chatbot group and the conventional resource group

Median diagnostic reasoning score

Adjusted difference between groups:
2 percentage points (95% CI, -4 to 8 percentage points); *P* =.60

SETTINGS / LOCATIONS

3 Academic medical centers in the US

PRIMARY OUTCOME

Diagnostic performance, assessed using a standardized rubric including differential diagnosis accuracy, appropriateness of supporting/opposing factors, and next evaluation steps; graded via blinded expert consensus

Goh *et al*. (2024) *JAMA Network Open 7*, e2440969.

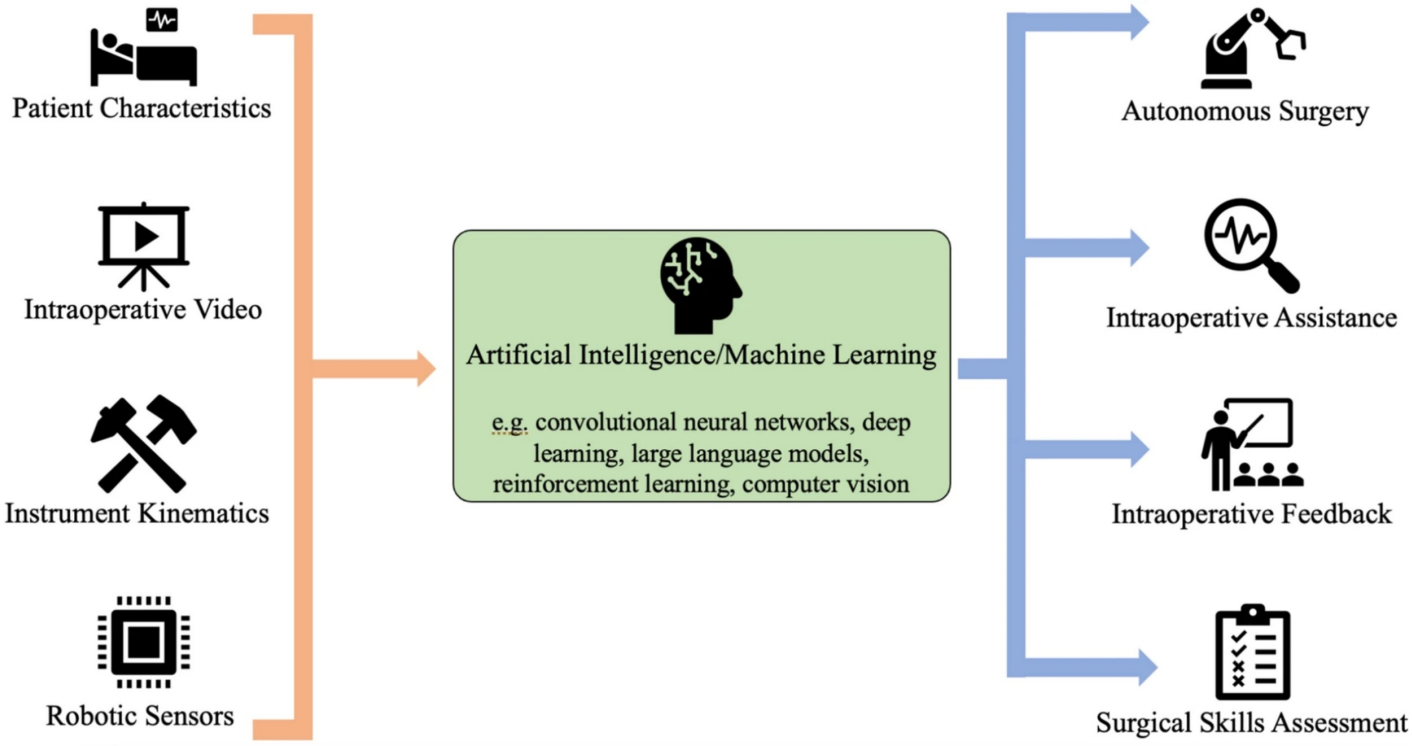
AI: Computer Vision/ML in Clinical Practice (Radiologists are ready, willing, and able!)

Computer Vision combined with Machine Learning can improve early detection of lung and breast cancer

Intellias.com

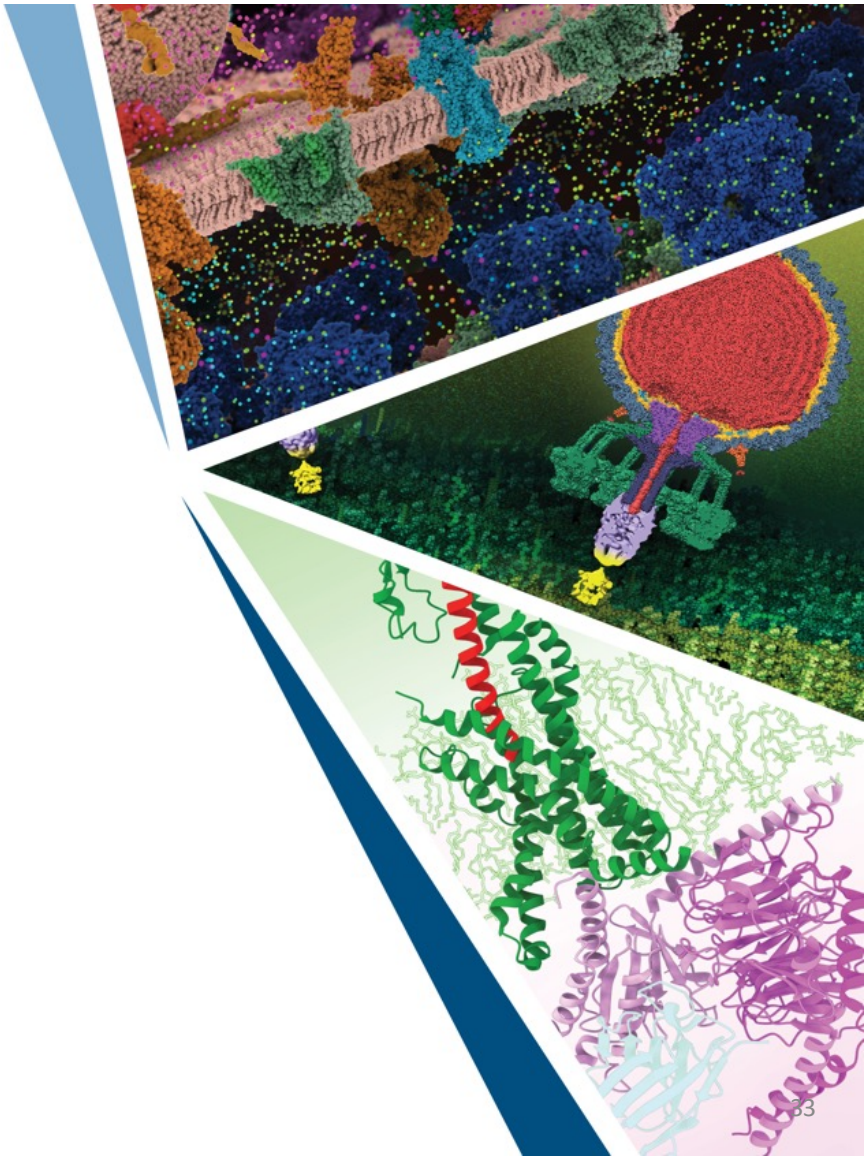
AI: Robotics/ML in Clinical Practice (Surgeons are ready, willing, and able!)

Surgeons are using AI to monitor interoperative metrics (force, etc.) and enhance detection of +ve surgical margins



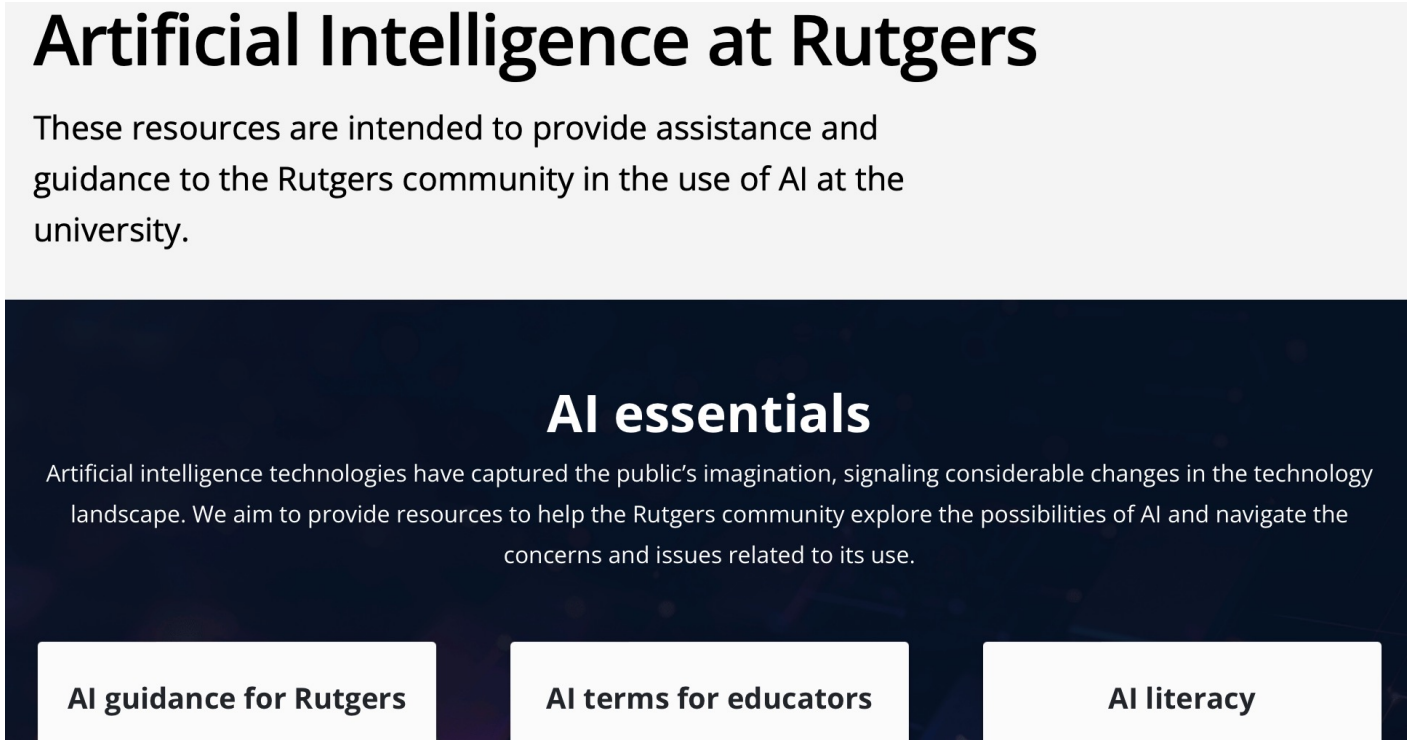
Knudsen et al. (2024) *Journal of Robotic Surgery 18*, 102.

Rutgers IT AI Resources

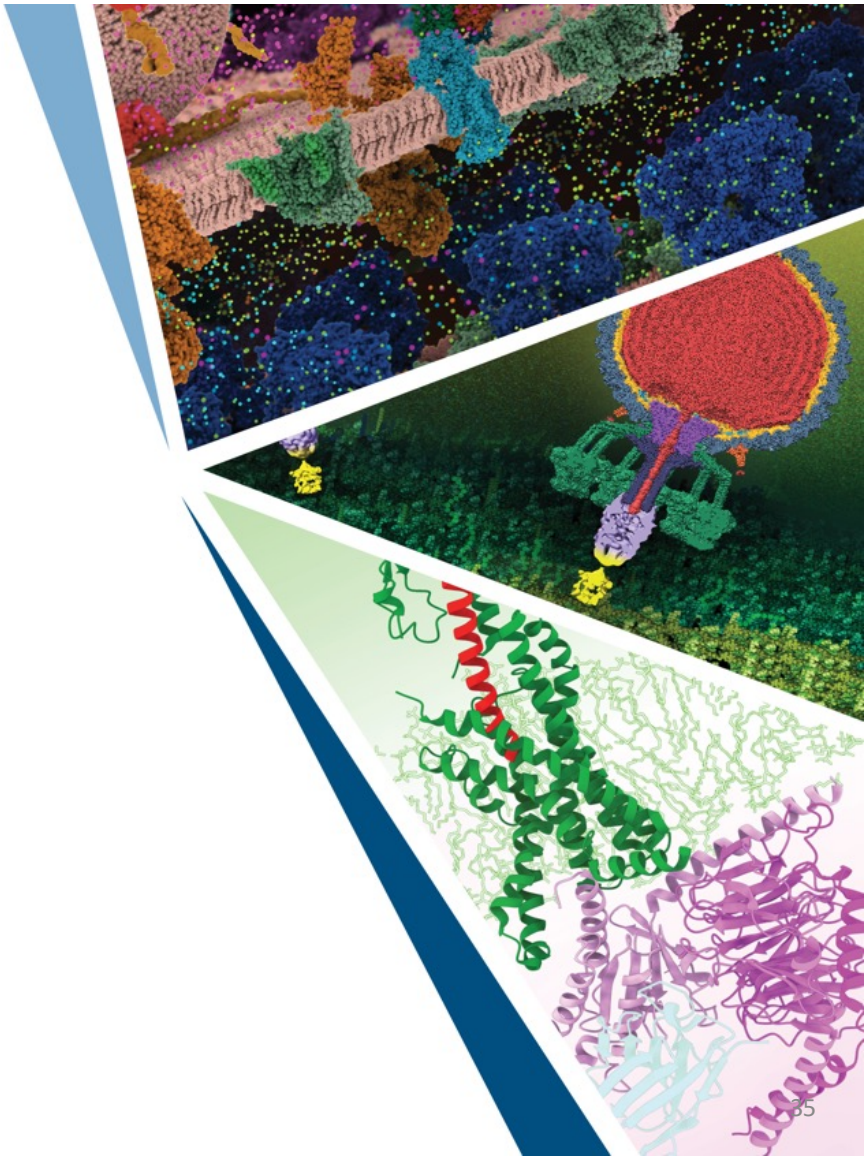


Rutgers IT AI Resources Web Page

URL: https://it.rutgers.edu/ai/



Skills You Need



Skills Needed to Succeed in the DS/AI Era

- Excel in your chosen major (subject matter expertise matters!)
- Learn to think in 3D at the atomic level (use RCSB.org in your bio courses)
- Bio or Chem of Physics Majors need to be well-grounded in
  - Computer Programming (coursework or Coursera)
  - Statistics (Statistics Department coursework)
  - Data Science (Data Science Program coursework)
  - Artificial Intelligence (Data Science Program coursework)

- Undergraduate research in a federally funded lab using AI/ML tools
- Co-author a peer-reviewed paper that uses AI/ML tools

Take Home Messages I

- Open access to PDB structure data informs research across fundamental biology, biomedicine, and bioenergy
- Open access to PDB structure data has contributed substantially to US FDA Drug Approvals
- Open access to PDB structure data was central to recent advances in computational structure prediction using Artificial Intelligence methods
- Combining PDB structure data with Computed Structure Models will accelerate basic and applied research
- RCSB Protein Data Bank RCSB.org research-focused web portal is a one-stop-shop for using 3D biostructure data in your education and your research

Take Home Messages II

- Rutgers AI and Data Science (RAD) Collaboratory is here to serve you as hub for related research, training, and community-building activities
- Rising Jrs/Srs consider applying for the RAD Summer Undergraduate Research Program through the ARESTY website starting late Feb/early Mar
- Learn more about the RAD Collaboratory at our March 8th (Data Science) and April 5th (AI/ML) networking events
- Appreciate the myriad ways that AI and Data Science are changing the face of basic and applied biomedical research and clinical practice
- Think carefully about the AI/DS-related skills you will need to succeed when you move on to medical school, *etc*. (Do Not Get Left Behind!)

RCSB PDB Team

RCSB.ORG
info@rcsb.org

Core Operations Funding

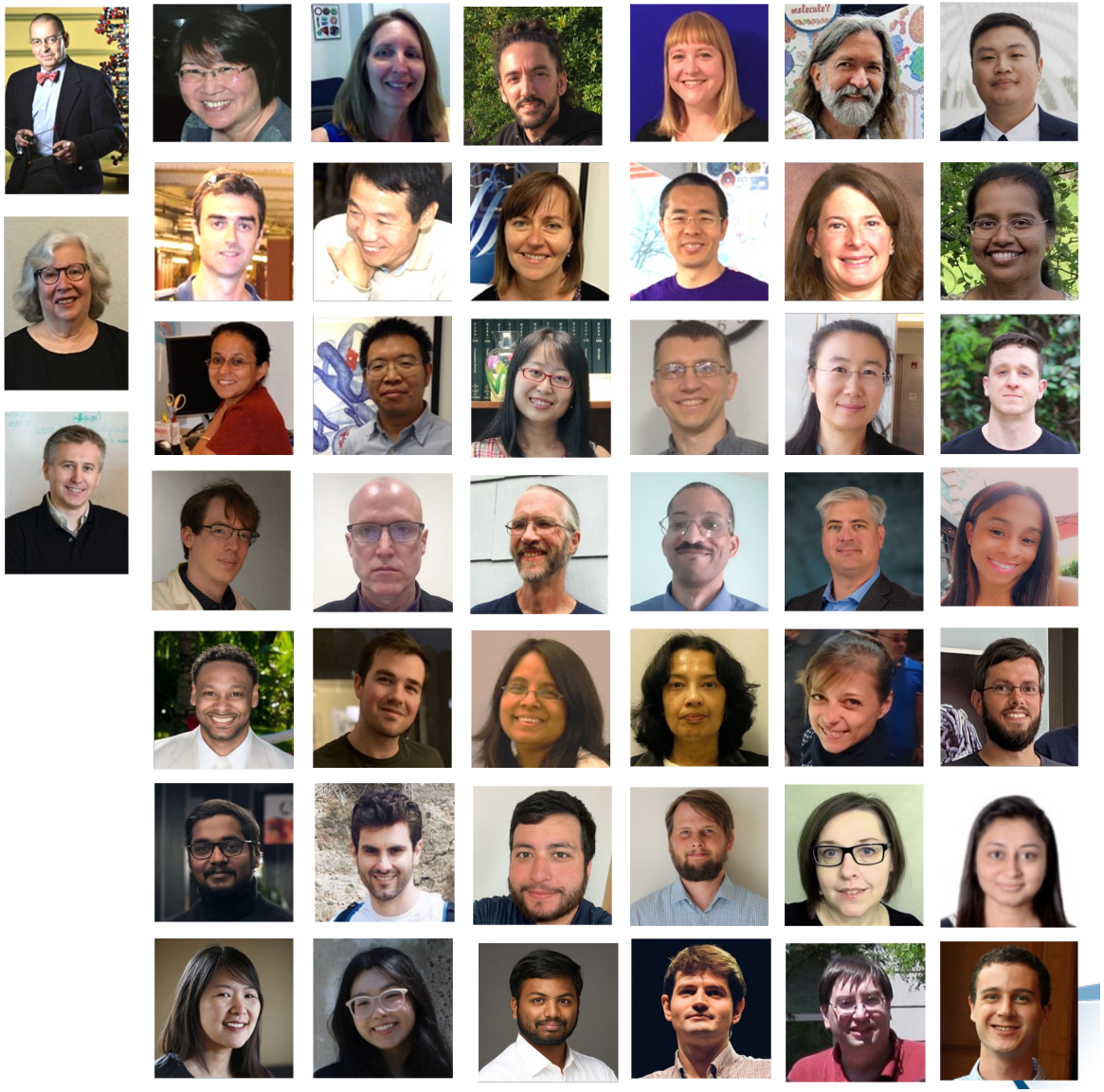
US National Science Foundation (DBI-2321666), National Institute of General Medical Sciences, National Institute of Allergy and Infectious Disease, and National Cancer Institute (NIH R01GM133198), and the US Department of Energy (DE-SC0019749)

Management

Member of the Worldwide Protein Data Bank (wwPDB; **wwpdb.org**)

Follow us

RCSB.org



Training Resources on PDB-101

pdb101.rcsb.org > *Train*

Materials to help effectively use **RCSB.org** tools for searching, visualizing, and analyzing 3D biostructure data

- Guide to Understanding PDB Data
- Training Courses
- Education Corner
- PDB & Data Archiving Curriculum

Sign up for Training Event Announcements