From PDB to AlphaFold
Detailed timeline

This timeline is an ongoing collective history and reflection about the datasets (especially the PDB) and scientific advances that have given origin to AlphaFold.
Contributions are welcome, for example detailed accounts and verifications from the individuals involved, interviews with them, identifications of relevant documents and reflections about the future of AI in science that are inspired by this history.
The following have already contributed:
Jane Richardson (Duke), Helen Berman (USC and Rutgers, a former Director of PDB, see Interview) , Pierre Baldi (UC Irvine, see comment and Interview), Søren Brunak (University of Copenhagen, see comment and Interview), Adam Godzik (UC Riverside), Tomaso Poggio (MIT, see comment), Alyssa Cruz (Sanford Burnham Prebys), Martin Steinegger (Seoul National University, a co-author in the original AlphaFold2 paper, see comment), Johannes Söding (Max-Planck Institute for Multidisciplinary Sciences, see comment), Joel Sussman (Weizmann Institute, a former Director of PDB, see comment), Philip Campbell (former Editor in Chief of Nature, see comment), Alexander Wlodawer (NIH, see comment), Tom Cech (former HHMI President, see comment), Harold Varmus (former NIH Director, see comment), Jake Feala (cofounder at Lila Sciences, see comment) and Mohammed AlQuraishi (Columbia University, see comment).

The initial conclusions, about which the scientists listed above have been consulted, are the following:
- AlphaFold was built on several decades of contributions from scientists working in multiple fields, including protein structural biology, bioinformatics and deep learning AI.
- The PDB data were produced by an international scientific community. The PDB was one of the first large-scale, openly available, scientific data sharing resources. Full data sharing of protein structures was slowly accepted by scientists. This acceptance took 20-30 years and required a change in culture, which benefited from support from scientific institutions, including funders, journals and scientific associations. It was part of a broader trend towards open science.
- Key innovations also originated from a company, DeepMind, which was founded in 2010, when the majority of the AI scientific community was skeptical about the value of deep learning methods.
- The innovations introduced by AlphaFold have led to further academic contributions, as shown by the examples of RoseTTAFold and OpenFold.

Future applications of AI in science might benefit from synergies similar to those that were described in this history.

A brief timeline and a summary are available.

1958-1960 The first protein structures, myoglobin and hemoglobin, were determined by John Kendrew and Max Perutz at Cambridge, UK; both were solved using X-ray crystallography.

1962 John Kendrew and Max Perutz received the Nobel Prize in Chemistry for their discoveries.

1965 At MIT's Project Mac, Cyrus Levinthal and Bob Langridge used computer graphics for the first time to display a protein structure, *i.e.,* myoglobin. 3D visualization was achieved simply by rotating the structure on the screen. (Levinthal 1966).

1965 Margaret Oakley Dayhoff's Atlas of Protein Sequence and Structure.
Margaret O. Dayhoff pioneered the systematic collection and analysis of protein sequences with the publication of the Atlas of Protein Sequence and Structure. This work compiled all ~70 known protein sequences at the time. Dayhoff's efforts laid the groundwork for bioinformatics and the development of substitution matrices—such as the PAM (Point Accepted Mutation) matrices—which are used for sequence alignment scoring to date. (Dayhoff, 1965)

1969 Cyrus Levinthal described the paradox of protein folding: the folding process must be guided by specific interactions and not by a random search through all possible conformations, which would take an immensely long time (Levinthal, 1969).

1970–1982 Needleman–Wunsch, Smith-Waterman to Gotoh.
In 1970, Saul Needleman and Christian Wunsch introduced the Needleman–Wunsch algorithm, the first systematic method for global sequence alignment using dynamic programming (Needleman, 1970).
This method was followed by the Smith–Waterman algorithm in 1981, which provided a framework for local alignments to detect conserved regions (Smith 1981).
The local alignment approach was motivated by the discovery of exons and introns in 1977 (by Richard Roberts and Phillip Sharp). It made it easier to detect homologous protein-coding regions in the surrounding sea of less conserved non-coding sequence in introns.
In 1982, Osamu Gotoh refined these methods by devising an elegant approach to compute affine gap penalties, thereby enabling rapid and biologically accurate sequence alignments (Gotoh, 1982). Together these seminal works developed the algorithm that is now executed billions of times daily to compute pairwise protein alignments.

1970-1971 As described by Helen M. Berman (Berman, 2008):
"The establishment of the Protein Data Bank (PDB) began in the 1970's as a grassroots effort. A group of (then) young crystallographers, including Edgar Meyer, Gerson Cohen and myself, began discussing the idea of establishing a central repository for coordinate data at an American Crystallographic Association (ACA) meeting in Ottawa, Canada, in 1970. Those conversations were continued with a larger group at the ACA meeting in Columbia, South Carolina, USA, in 1971. At that meeting, a petition was written, and a proposal was submitted to the United States National Committee for Crystallography (USNCCr)."

In 1970 Meyer wrote to Helen Berman that he "initially thought about approaching the International Union of Crystallography (IUCr) but became discouraged when told he would run into the opposition of “certain blocking groups.” (Strasser, 2019)

1971 June. The meeting at Cold Spring Harbor.
As written by (Strasser, 2019):
"Any solution to this problem would require a broad international consensus. Fortunately, a unique opportunity soon arose to discuss the data bank project with the international crystallographic community. In June 1971, the Cold Spring Harbor Symposium on Quantitative Biology was devoted to the “Structure and Function of Proteins at the Three-Dimensional Level.” Organized by James Watson, the list of attendees of this select meeting read like a “who’s who” in protein crystallography, including (future) Nobel Prize winners Dorothy Crowfoot Hodgkin, Max Perutz, Aaron Klug, and William N. Lipscomb. Although the meeting was by invitation only, a few scientists who were too junior to be on the list decided to participate anyway and “kind of crashed the meeting.” Helen Berman and three friends, self-described “hippies” who valued communitarian ideals, drove from Philadelphia to Long Island to attend the meeting and present the idea for a crystallographic data bank." The online memoir of Helen Berman shows a picture of herself with several of the young scientists that participated in these discussions (Sung-Hou Kim, Joel L. Sussman, and Nadrian C. Seeman) at a time close to the meeting.
From Helen M. Berman (Berman, 2008):
"The discussions within the meeting room, on the lawn, and on the beach were exciting and intense. In an informal meeting convened by Max Perutz, protein crystallographers discussed how best to collect and distribute data.
During the CSH meeting, [Walter] Hamilton was approached with the idea that had been discussed within the ACA community – a public data bank of protein structures. At an ad hoc meeting of protein crystallographers attending the Symposium, it was proposed that there should be a repository with identical files in the United Kingdom and in the USA. Hamilton volunteered to set up the American data bank at Brookhaven.
When Max Perutz returned to England, he discussed this proposal with Olga Kennard, who was the founder of the Cambridge Crystallographic Data Centre (CCDC) and had wide experience in assembling and archiving crystallographic data. Walter Hamilton wrote to her with an offer of collaboration and proposed to meet and discuss some of the details of coordinating the activities. He visited England that summer and, by October 1971, the establishment of the Protein Data Bank archive, jointly operated by the CCDC and BNL, was announced in Nature New Biology"(1971, Nature New Biology)
At the time Walter Hamilton was Deputy Chairman of the Chemistry Department at Brookhaven National Laboratory. He was also a former President of the American Crystallographic Association.

1971 August. At the ACA Conference in Ames, Iowa, the first 3D molecular graphics film was shown in a lecture by Joel L. Sussman on very small RNA structure, UpA [https://www.youtube.com/watch?v=PraieqBi048] (Seeman et al, 1971).

1971-1976 The PDB was established at Brookhaven National Laboratory under the leadership of Walter Hamilton. After Hamilton's death, in 1973, Thomas Koetzle became the next leader. It originally contained 7 structures and initially it grew slowly. By 1976 a total of 13 structures were contained in the database.
The PDB was one of the first large-scale, openly available, scientific data sharing resources. An older example is the World Data Center (WDC) system, which was created for the International Geophysical Year (1957–58) to archive and openly exchange geophysical observations worldwide and was later continued. The Cambridge Structural Database (CSD), maintained by the CCDC, began in 1965 as a major crystallographic database, but it eventually became license restricted. Olga Kennard, John Desmond Bernal and the International Union of Crystallography were involved in the origin of the CSD (as explained by Olga Kennard in this interview, after minute 7:36: https://www.youtube.com/watch?v=Jz6Hk6N2sWA)

1972 Christian Anfinsen received the Nobel Prize in Chemistry for his work showing that all the information for a 3D protein structure is contained in the sequence of amino acids.

1982 The sequence databases at GenBank in the US (Jordan, 1982) and at EMBL in Europe (Hamm, 1986) were opened to the public (Strasser, 2008; Strasser, 2019). The databases at GenBank, EMBL and after 1986 at DDBJ in Japan are mirror organizations with the same content.

1986–2002 Swiss-Prot, TrEMBL, and UniProt.
In 1986, Amos Bairoch established Swiss-Prot, a database of curated proteins and later, to handle the exponential growth in protein sequence data, TrEMBL was introduced in 1996 as a complementary database containing computationally annotated entries.
In 2002, Swiss-Prot, TrEMBL and the Protein Information Resource (PIR) at Georgetown University in the US received support by NIH and merged to form the Universal Protein Resource (UniProt) (Butler, 2002). The extensive open-source protein sequence data in UniProt was indispensable for generating diverse multiple sequence alignments, critical for training AlphaFold2.

1988 First applications of neural networks to predict the secondary structure of proteins from the sequence (Quian and Sejnowski, 1988; Bohr et al, 1988). Applications of this approach achieved great improvements in accuracy over the following years (Baldi, 2018). Søren Brunak (see comment) trained some of the earliest neural networks predicting protein structure. He was part of a Danish group that was first to predict distance matrices for proteins by neural networks (Bohr, 1990), a step that is a key element in the original AlphaFold methodology. As pointed out by Poggio, even before deep learning was fully developed several other machine learning methods started to have successful applications to scientific and engineering problems in the 1990s.

1989 An article by Marcia Barinaga in Science about "The Missing Crystallography Data" provides a very informative snapshot of the ongoing discussions (Barinaga, 1989). The article mentions a letter initiated by Frederic Richards (Yale) in 1987 and co-signed by 173 colleagues, encouraging the sharing of protein structure data. Among the leading petition signatories in addition to Richards were Jane and David Richardson. A second letter by Richard Dickerson (UCLA) in 1989 made the point again and presented data showing that less than half of published DNA structures disclosed the coordinates. Dickerson stated that "we are on our way to developing a miniature scandal." There was a difference of opinion among Editors of the major scientific journals, including Science and Nature, regarding the necessity of sharing all the details of the structures at the time of publication. Industry opinions also differed. The deputy director of NIGMS, Marvin Cassman, encouraged public deposition of data and hoped that the scientific community would come to an agreement about this.

The International Union of Crystallography published guidelines (IUCr, 1989) recommending deposition of data, but as a compromise among different viewpoints the release of the coordinates could be delayed for up to 1 year after publication, and the release of the experimental data (structure factors) could be delayed for up to 4 years.

An editorial by John Maddox, the Editor of Nature, was published in September 1989, soon after the Barinaga article, and defended the policy of not requesting a database deposition as a condition of publication of structural biology and DNA sequencing papers (Maddox, 1989). In November 1989 letters to Nature by scientists condemning this policy followed. Richard J. Roberts (at CSHL, later awarded the 1993 Nobel Prize in Physiology or Medicine) stated that he was "appalled by the comments of John Maddox" (Roberts, 1989). Thomas Koetzle, at the time Director of PDB, more diplomatically encouraged Nature to "reconsider its policy of not requiring deposition of data in the appropriate databases" (Koetzle, 1989).

In 1989 Renato Dulbecco (Nobel Medicine 1975) published remarks about the world of science moving away from open communication and sharing, which he confirmed in his interview shown on this site, as did many other sources. What was eventually achieved by the PDB is even more remarkable because it worked against this broader trend.

1990 BLAST (Basic Local Alignment Search Tool).

Altschul et al. introduced BLAST, a tool that revolutionized sequence searches by enabling rapid detection of sequence similarities (Altschul, 1990). This innovation dramatically improved researchers' ability to search through ever-growing protein databases through a seed-and-extend-based alignment scheme. Additionally, the introduction of the E-value—derived from the Karlin-Altschul statistical framework—provided a robust measure for assessing the likelihood of a match occurring by chance, thereby grounding sequence alignment in solid statistical principles (Karlin, 1990).

1994 CASP (Critical Assessment of protein Structure Prediction) was co-founded by John Moult and Krzysztof Fidelis, as a blind and independent test of software for the prediction of protein structure from sequence.

Rosetta (Rohl, 2004), contributed by the lab of David Baker, was one of the most successful methods in the initial phase.

The results improved until 2002, but after that date they were essentially flat. The next major improvements were in 2018 and 2020 with Alphafold and Alphafold2.

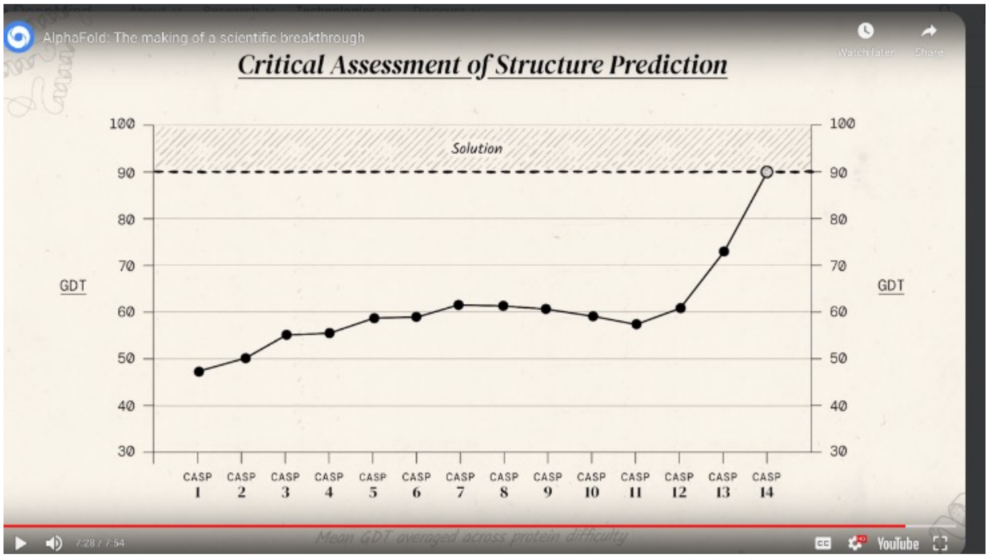


Figure 1: The scores of the winner of the CASP competition, that took place every 2 years, starting in 1994. Note the long period of stasis. The last two time points are AlphaFold in 2018 and AlphaFold2 in 2020. GDT (Global Distance Test) is the main metric used to evaluate predictions submitted to CASP. The figure is from of a DeepMind video about the making of AlphaFold (minute 7:27).

1995 The first PDB Browser was released by the PDB-BNL. It dramatically enhanced the PDB's printed index listings and various ad hoc search protocols developed to find PDB entries. Selected proteins in the PDB could be easily downloaded, and their molecular structures visualized on lab computers (Stampf et al, 1995 & Sussman et al, 2001) via RasM (Sayle, 1995) and other 3D visualization tools. The following year, the browser was significantly improved, becoming the “3DB Browser” (Prilusky et al, 1996). See comment from Joel Sussman with more details.

1996: The PDB release of AutoDep, the first web-based tool for macromolecular structure deposition and validation. It was developed at the PDB-BNL, but was also given to the PDBe, who used it as the first remote site for deposition in the PDB. Within 3 months of its release, over 50% of all new submissions were deposited via AutoDep. AutoDep predated any web submission of papers to journals (Lin, 2000).

1996 November. A conference celebrating the 25th Anniversary of the PDB and the 10th Anniversary of Swiss-Prot was held in Jerusalem. [http://www.weizmann.ac.il/csb/faculty_pages/Sussman/pdb25sp10]. This was one of the first meetings at which 3D structural data and sequence information were analyzed synergistically.

1997 PSI-BLAST (Position-Specific Iterated BLAST).
Building upon the BLAST framework, PSI-BLAST constructs sequence profiles from initial

alignments and iteratively searches through the database (Altschul, 1997). This allowed researchers to detect even more remotely homologous relationships efficiently.

1998 HMMER: fast profile Hidden Markov Models to sequence alignments.
Sean Eddy developed HMMER, an efficient suite of methods that applies hidden Markov models to sequence search and alignment (Eddy, 1998; Eddy 2011). By incorporating probabilities for insertions and deletions into the profile scoring, HMMER significantly improved the sensitivity of sequence comparisons, establishing itself as a critical tool for the large-scale annotation of protein families and domains.

1998 Nature, Science and PNAS reversed their long-standing policy of not requiring the immediate release of high-resolution structural coordinate data upon publication.
Nature stated in their 9 July 1998 issue that "It is clear that there is a significant majority opinion in the community against permitting a one-year hold. Accordingly, Nature, simultaneously with Science, is changing its policy. Any paper containing new structural data received on or after 1 October 1998 will not be accepted without an accession number from the Brookhaven Protein Data Bank (PDB) accompanied by an assurance that unrestricted ("layer-1") release will occur at or before the time of publication."
Floyd Bloom, of Science, published in their 10-Jul-1998 issue a very similar editorial: https://www.science.org/doi/full/10.1126/science.281.5374.175c. The Editor-in-Chief of PNAS, Nick Cozzarelli, also published a similar editorial policy in the 28-April 1998 issue. See the comment of Joel Sussman and that of Philip Campbell, at the time Editor in Chief of Nature, for an account of the discussions leading to this decision. Joel Sussman mentions the numerous meeting and letters to journals that took place during the 1990s, while he was PDB Director. Philip Campbell describes the close coordination with Floyd Bloom at Science and makes it clear that they were following the desires of the scientific community. A letter was published in Science in early 1998 urging the adoption of the immediate release policy (Wlodawer, 1998) and was accompanied by a note of Floyd Bloom encouraging readers to respond to a survey to make their preference know about this change. See also comment from Alex Wlodawer about the role played by funders like HHMI and NIH in promoting data sharing. The NIH changed its policy on January 29, 1999, requiring NIH grant recipients to deposit atomic coordinates for immediate release upon publication. Tom Cech, a former HHMI President, has provided insights about the discussions ongoing at the time about the new policy. Harold Varmus, NIH Director at that time, has also shared the reasons behind this policy change.

1999 The Research Collaboratory of Structural Bioinformatics (RCSB) became the new manager of the PDB (Berman, 2000). The three member institutions of the RSCB were: Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology (NIST).
The new Director was Helen Berman of Rutgers University. John Westbrook from Rutgers, Peter Arzberger (and then Phil Bourne) from SDSC at UCSD, and Gary Gilliland from NIST became the co-Directors.

A new data representation format, eventually called PDBx/mmCIF, started to be adopted by the RCSB, even if the complete buy in by the community and by the crystallographic software developers took several more years. This new data representation format was fully machine readable and facilitated quality control.

2000 Larry Page, who co-founded Google in 1998, predicts in an interview (minute 3:59) the importance of AI for providing answers to search inquiries and for the future Google.

2000 The US National Institute of General Medical Sciences (NIGMS) at NIH supported the Protein Structure Initiative (PSI) for 15 years, from 2000 to 2015, providing grants for around $1 billion in total. The PSI was created to solve novel protein structures in a high-throughput manner. During the initial years the aim was to explore the protein structure space in a systematic way. More than 6,500 protein structures were solved (Editorial Nature Methods, 2014). The goal of the PSI was "to make the three-dimensional, atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences".
There was much debate during the period of NIH funding of the PSI about the relative merits of investigator-initiated, hypothesis-motivated science versus more systematic discovery science (Banci et al, 2007). This debate is not new, see the reflections of Medicine Nobel winner Barry Blumberg. It is a question that can be seen from a different perspective now that the importance of large, high-quality datasets for AI based science has become clear. The PSI was the largest of the structural genomics consortia, which were based not only in the US but also in Europe, Canada and Japan. These efforts (including the PSI) solved more than 13,500 protein structures by 2015, for a combined expense of around $2 billion (Grabowski et al, 2016).

2003 The worldwide PDB is announced (Berman, 2003) as a collaboration of three organizations: The RCSB, the Macromolecular Structure Database at the European Bioinformatics Institute (EBI) and the Protein Data Bank Japan (PDBj) at the Institute for Protein Research in Osaka University. The goal is "maintaining a single archive of macromolecular structural data that is freely and publicly available to the global community".

2005–2012 HH-suite - fast HMM-HMM alignment.
Johannes Söding and colleagues developed HHsearch and HHblits, which compare hidden Markov models against hidden Markov models (HMM–HMM) (Söding, 2005; Remmert, 2012). This method greatly enhances sensitivity for detecting remote homology, enabling the discovery of extremely distant relationships that might be missed by traditional approaches.

2008 The first significant use of GPUs (graphics processing unit) in machine learning applications. An account was presented by Rajat Raina and Andrew Ng at a NIPS workshop. GPUs were initially developed for digital image processing and used by the videogame industry but were later found to be able to considerably speed up calculations needed in AI applications.

2008 PDB depositions required not only the coordinates but also the structure factors (the experimental data). The experimental data allowed a more comprehensive validation.

2009 Initial publication of ImageNet, a very large and systematic dataset of labelled images built by a group coordinated by Fei-Fei Li and designed to support AI vision research.
According to Fei-Fei Li "One thing ImageNet changed in the field of AI is suddenly people realized the thankless work of making a dataset was at the core of AI research. People really recognize the importance the dataset is front and center in the research as much as algorithms."

2010 14-15 August. Demis Hassabis presented at a conference in San Francisco, the Singularity Summit, a series of yearly conferences about artificial intelligence, initially supported by Peter Thiel. The singularity is the moment when artificial intelligence becomes more capable than human intelligence. The title of Hassabis talk was "A Systems Neuroscience Approach to Building AGI" and the first slide already showed the DeepMind logo. He suggested that machine learning and knowledge of neuroscience could be combined to design artificial general intelligence. Several slides showed the work of Tomaso Poggio, a leading AI scientist. Hassabis had been a visiting scientist in the lab of Poggio at MIT.
Shane Legg also presented and spoke about "Universal measures of intelligence"
In November DeepMind was officially founded by Hassabis, Shane Legg and Mustafa Suleyman.

2010 to early 2011 Funding of DeepMind by Venture Capital groups, led by Peter Thiel and his Founders Fund. Tomaso Poggio was also a minor investor (DeepMind 2011 Annual return). According to a 2024 interview with Shane Legg (minute 7:06) in 2010 people thought that AI was a failed area, and nobody wanted to fund it. Especially in the case of DeepMind, because they did not just propose to do some machine learning, they wanted to build artificial general intelligence. Peter Thiel funded them because he is a well-known contrarian investor. He obtained opinions from other people who probably told him that investing in this company was a bad idea. One of the first breakthroughs was an algorithm that could play many different Atari videogames. This was the first general algorithm.

2012 AlexNet, a convolutional neural network (CNN), designed by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton from the University of Toronto, won the ImageNet Large Scale Visual Recognition Challenge. It was the first model based on neural networks to win the competition, achieving a large improvement compared to previous methods. The three authors were hired by Google in 2013.
It has been widely commented that the ImageNet 2012 competition triggered the most recent explosion of interest in AI (Sejnowski, 2018).

2014 January. Google bought DeepMind for around $600m but DeepMind remained as a separate entity. DeepMind obtained access to a large computational infrastructure and capital for expanding and acquiring top talent for their team.
In 2024 DeepMind merged with Google Brain and Hassabis was put in charge of the entire Google AI effort, possibly as a response to the success of ChatGPT.

2014 The total number of structures deposited in PDB surpasses 100,000.

2015 Emails from this period between Elon Musk, Sam Altman and others were recently published as part of a court case.
They show how a concern about Google and DeepMind dominating AI was one of the motivations for starting OpenAI, the developer of ChatGPT. Musk and Altman wrote that "OpenAI is a non-profit artificial intelligence research company with the goal of advancing digital intelligence in the way that is most likely to benefit humanity as a whole, unencumbered by an obligation to generate financial returns." An email by Altman stated "OpenAI's mission is to ensure that artificial general intelligence (AGI) - by which we mean highly autonomous systems that outperform humans at most economically valuable creative work - benefits all of humanity."
We now know that for-profit motives have later become more central at OpenAI, and a reflection about the appropriate governance structure for non-profit AI might be needed.

2016 The AlphaGo match was another proof of principle for DeepMind. After the AlphaGo match Hassabis remembered playing Foldit, a game designed by David Baker and others to allow the general public to participate in protein folding efforts, and other discussions about this problem going back to his college days. Foldit showed the potential of crowdsourcing in science. (Cooper, 2010)
DeepMind started a serious effort on the protein folding problem.

2016 MMseqs2: Fast iterative profile searches for building MSAs.
The exploitation of the huge metagenomics sequence sets for iterative sequence searching to build MSAs required a fast sequence profile search tool that can handle datasets of billions of sequences. MMseqs2 filled that gap, with a search speed two to three orders of magnitude faster than PSI-BLAST or HMMER yet similar sensitivity (Steinegger & Söding, 2017). It would later enable the fast generation of MSAs for AlphaFold2 and Colabfold.

2017: Metagenomic Data Integration for structure prediction.
The integration of metagenomic sequencing data dramatically expanded the pool of available protein sequences by adding billions of sequences from diverse microbial communities (Ovchinnikov, 2017). This expansion—driven by a global effort to sequence and deposit experiments—has vastly improved the breadth and accuracy of multiple sequence alignments used in protein structure prediction and other analyses.

2017-2018: Linear-time sequence clustering enabled the exploitation of huge metagenomic sequence corpora. In as much as large language models have profited from ever increasing sizes of their training corpus, the deep-learning revolution in protein biology, including AlphaFold, relies critically on training protein language models with huge numbers of non-redundant sets of protein sequences. AlphaFold2, for instance, was trained on a collection of representative sequences obtained by clustering 4 billion sequences from metagenomic and genomic sources (BFD database) and 1.6 billion sequences from MGnify v18. Generating such huge reference sets only became possible with Linclust, the first algorithm whose runtime scaled linearly instead of quadratically with the size of the input sequence set (Steinegger & Söding, 2018). Before Linclust, the practical limit for sequence clustering was at around 100 million sequences. AlphaFold2 profits in another way from the huge and diverse databases such as MGnify and BFD clustered with Linclust. The model quality depends on a sufficient diversity of the MSA built from the query sequence, and that diversity may depend crucially on the diversity of the sequence databases in which it searches for homologous sequences. Removing both MGnify and BFD for the MSA generation reduced AlphaFold2's mean GDT score by 6.1. Additionally, the Uniclust resource was established to provide deeply clustered and annotated protein sequence databases based on UniProt data utilized for AlphaFold2 training (Mirdita, 2017).

2017 Publication of the "Attention Is All You Need" paper about transformers by a group from Google (Vaswani, 2017).

2017 Recently launched validation efforts at PDB are described (Gore, 2017). PDB produces a validation report that is required for review by an increasing number of journals. The report provides metrics to evaluate the quality of the experimental data, the structural model, and the fit between them.

2018 AlphaFold from DeepMind wins CASP13 (Senior, 2020).

2020 AlphaFold2 wins CASP14 and it is considered by many to have essentially solved the protein folding problem (Jumper, 2021). The authors stated that "This bioinformatics approach has benefited greatly from the steady growth of experimental protein structures deposited in the Protein Data Bank (PDB), the explosion of genomic sequencing and the rapid development of deep learning techniques to interpret these correlations." (Jumper, 2021)
Various parts of the model used copies of PDB obtained at different times from 2019 to 2020. In 2019 PDB contained 158,794 structures and in 2020 contained 172,779 structures.
2,423,213,294 protein sequences were obtained from UniProt and other open resources and were used for multiple sequence alignments providing evolutionary information. The majority of the proteomes in UniProt are based on the translation of genome sequences from GenBank and the other mirror sites in Europe and Japan. (UniProt Consortium, 2021)

Alphafold2 is based on a modified transformer architecture. It uses comparative evolutionary information in the Evoformer, and then passes information to another transformer called the structure module. The information cycles between the two modules. The DeepMind team working on AlphaFold was led by John Jumper and supervised by Demis Hassabis.
After the AlphaFold2 success, a debate started about the reasons why this solution to the problem of protein folding was not found by an academic group. Insights can also be obtained from a comparative analysis with Genentech, another start-up company that contributed to important scientific advances.

2021 RoseTTAFold (Baek, 2021), developed by a team led by David Baker, incorporated ideas from AlphaFold2 and achieved accuracies approaching it.

2022: Making AlphaFold2 accessible to all through ColabFold.
ColabFold made AlphaFold2 predictions widely accessible to researchers and practitioners without access to large-scale computing infrastructure by providing high-quality, rapid and free-of-charge multiple-sequence alignment (MSA) generation through a publicly accessible MMseqs2 search server and a user-friendly Google Colab-based notebook interface (Mirdita, 2022).

2022 DeepMind releases structure predictions for 218 million proteins, nearly all known proteins.

2023 - 2024 AlphaMissense is another AI tool developed by DeepMind. It was published in Science in September 2023 and predicts the pathogenicity of all possible human single amino acid substitutions (Cheng, 2023). All the components of the AlphaFold and AlphaFold2 AI models were shared openly, but in the case of AlphaMissense the trained weights, a set of parameters essential for running the model, were not shared.
In 2024 DeepMind released AlphaFold3, which adds a diffusion-based method to predict binding structures and interactions of proteins with other molecules.
When AlphaFold 3 was published in Nature the code was not provided (Abramson, 2024). A server was offered for non-commercial use, but the number and types of queries allowed was limited.
A petition signed by more than one thousand scientists expressed disappointment with the lack of disclosure of the AlpaFold3 code at the time of publication in Nature. Not even reviewers were given access to the code during review.
Six months after publication the code of AlphaFold3 was released for academic use.
Opinions about transparency and AI from several journal editors (FEBS Journal, JBC and Science) are also shared on this site.

2024 The lab of David Baker releases RoseTTAFold All-Atom, which predicts 3D structures of assemblies of proteins and other small molecules. (Krishna, 2024)

OpenFold (Ahdritz, 2024), an open-source implementation of AlphaFold2 including the code and data required to train new models, is produced by a large academic collaboration and yields insights into its learning mechanisms and capacity for generalization.
One of the leaders of the OpenFold Consortium, Mohammed AlQuraishi, shared a comment highlighting contributions to the scientific development of AlphaFold and related efforts.

2024 Oct. Nobel in Chemistry for David Baker, Demis Hassabis and John Jumper "for computational protein design and protein structure prediction".
The Nobel Prize in Physics was awarded to John J. Hopfield and Geoffrey Hinton “for foundational discoveries and inventions that enable machine learning with artificial neural networks”. The scientific background to the Nobel Prize in Physics stated: “So far, the most spectacular scientific breakthrough using deep learning ANN [artificial neural networks] methods is the AlphaFold tool for prediction of three-dimensional protein structures, given their amino acid sequences.”

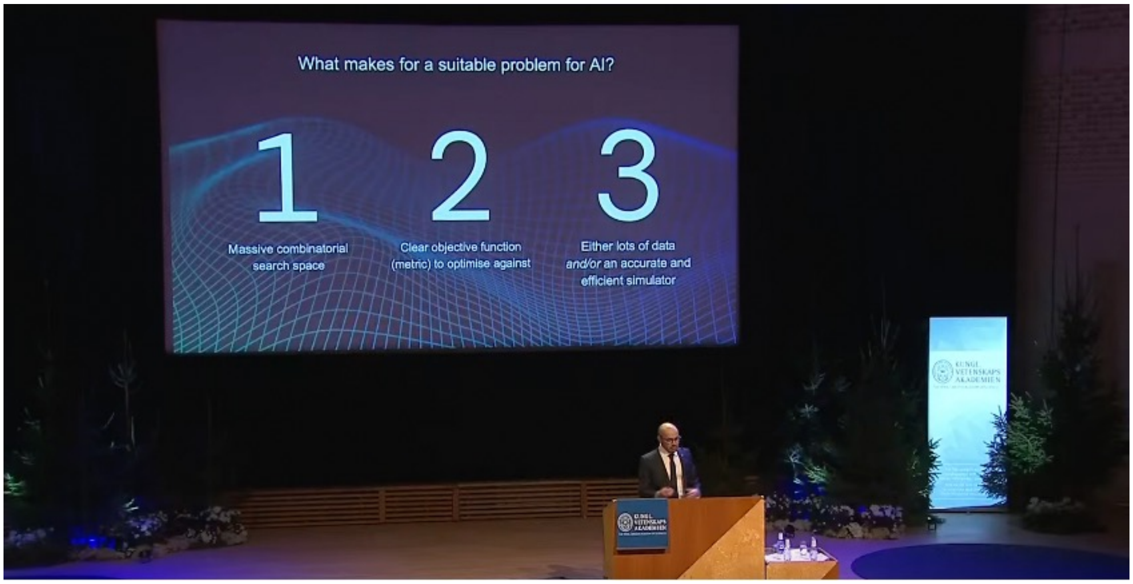
2024 Nov 18. AI for Science Forum, co-hosted by Google DeepMind and the Royal Society. In one of the sessions, Janet Thornton (minute 19:25), Director Emeritus, European Molecular Biology Laboratory - European Bioinformatics Institute, which was closely involved in the PDB, said that it took 20 years for every scientist to come around to the idea of sharing the data. A key step was a statement from the International Union of Crystallographers saying that unless people deposited their data, they would not be able to publish in various journals. Many in the community were already on board, with some outstanding exceptions. Some of the most famous scientists did not initially share. A change of culture was needed. In academic research the data are obtained with public funds, so the case for sharing is even stronger.
Siddhartha Mukherjee (Columbia University) (minute 18:28) pointed out that patients might freely share their data to benefit the public good but might be less inclined to agree to do it for the benefit of a company. The same might be said of scientists.
Anna Greka (Core Institute Member, Broad Institute of MIT and Harvard) (minute 32:17) suggested that a dataset that could play the same role as the PDB for future AI models of the cell could be obtained by systematically perturbing human cells. It would need to be a well-controlled and clean dataset including single cell data, transcriptomics and imaging measurements.

Most Interviews on this site, especially the recent ones, also mention datasets that could support AI models, as the PDB did. For example, Aviv Regev and Sarah Teichmann described the Human Cell Atlas, a consortium that aims to create a comprehensive reference map of all human cells; Gene Yeo mentioned the potential importance of collecting data about all possible RNA modifications; Jack Gilbert described the complexities and opportunities arising from microbiome data; Gary Siuzdak and Bruno Conti spoke about metabolomics and metabolomics databases.

Andrew McCulloch mentioned the importance of AI models at multiple scales, from atomic resolution to populations. In the Interviews there were several mentions of virtual models at the cell or higher level that would need integrations of many datasets.

In another session of the AI for Science Forum, Paul Nurse (Director of the Francis Crick Institute and 2001 Nobel Prize in Physiology or Medicine) made several thought-provoking remarks (minutes 1:13 and 21:34). He said:
Science has increased in complexity and silos have been created, quite often self-referential silos. We must begin by seeing how we can break down those silos, how we can actually get the different parts of the scientific community talking to each other, not just collaborating, but interacting and talking one with another.
That's particularly the case with respect to artificial intelligence, because we are all being influenced by it. We must increase the permeability to it, so it doesn't become a sort of new priesthood that is somehow separate from the rest of the scientific endeavor. There are social science aspects to this, of actually getting a better working community, working across disciplines, working across different types of organizations, from universities through to industrial and commercial organizations. And that requires, not only the will, including the political will, but actually us thinking carefully about how we talk to each other, how we think about it, how people are trained in different ways in different scientific fields. We would benefit by using social scientists to help us. We need advice and help from them.

2024 Dec 8. Nobel lectures by Baker, Hassabis and Jumper. All the speakers said that the PDB data had been essential for their work.



What makes for a suitable problem for AI?

1 Massive combinatorial search space

2 Clear objective function (metric) to optimise against

3 Either lots of data *and/or* an accurate and efficient simulator

Figure 2: A list of criteria suggested by Demis Hassabis to determine if a scientific problem is suitable for an AI solution.
1- Massive combinatorial search space
2- Clear objective function (metric) to optimize against

3- Either lots of data and/or an accurate and efficient simulator
The slide was presented during his Nobel lecture.

2025 In an Interview published on January 23 Demis Hassabis stated that most of the old AlphaFold team at Google DeepMind is now working on the Virtual Cell (minute 39:21), building an AI simulation of a working cell. They expect to solve this problem within the next five years.
Other approaches to AI in science are also being explored, as shown by the comments on this website from industry and academic scientists, including Jake Feala, Aviv Regev and Sarah Teichmann, Gene Yeo, Jack Gilbert, Gary Siuzdak and Bruno Conti, Andrea Califano, Pierre Baldi, Soren Brunak, Talmo Pereira, Andrew McCulloch and many others, included in the sections on Interviews , Roundtable and Surveys.

REFERENCES

(see also links within the text)

1971. Crystallography: Protein Data Bank. Nature New Biology, 233(42), pp.223-223.

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J. and Bodenstein, S.W., 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 630(8016), p.493

Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T.J., Berenberg, D., Fisk, I., Zanichelli, N. and Zhang, B. et al, 2024. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Nature Methods, pp.1-11.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D. and Millán, C. et al , 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 373(6557), pp.871-876.

Baldi, P., 2018. Deep learning in biomedical data science. *Annual review of biomedical data science*, *1*(1), pp.181-205.

Banci, L., Baumeister, W., Heinemann, U., Schneider, G., Silman, I., Stuart, D.I. and Sussman, J.L., 2007. An idea whose time has come. *Genome Biology*, *8*, pp.1-3.

Barinaga, M., 1989. The missing crystallography data: some disgruntled researchers are mounting a campaign to force crystallographers to make available key data when they publish the structure of complex molecules. Science, 245(4923), pp.1179-1181.

Berman, H.M., 2008. The protein data bank: a historical perspective. Acta Crystallographica Section A: Foundations of Crystallography, 64(1), pp.88-95.

Berman, H., Henrick, K. and Nakamura, H., 2003. Announcing the worldwide protein data bank. Nature structural & molecular biology, 10(12), pp.980-980.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., 2000. The protein data bank. Nucleic acids research, 28(1), pp.235-242.

Bohr H, Bohr J, Brunak S, Cotterill RM, Fredholm H, Lautrup B, Petersen SB. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. FEBS Lett. 1990 Feb 12;261(1):43-6.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M., Lautrup, B., Nørskov, L., Olsen, O.H. and Petersen, S.B., 1988. Protein secondary structure and homology by neural networks The α-helices in rhodopsin. *FEBS letters*, *241*(1-2), pp.223-228.

Butler D: NIH pledges cash for global protein database. Nature 2002, 419:101.

Campbell, P., 1998. New policy for structure data. Nature, 394(6689).

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T. and Schneider, R.G., 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science, 381(6664), p.eadg7492.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D. Popović, Z., & Foldit players, 2010. Predicting protein structures with a multiplayer online game. Nature, 466(7307), pp.756-760.

Dayhoff, M.O. et al. (1965). Atlas of Protein Sequence and Structure, National Biomedical Research Foundation.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics, 14(9), 755–763.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Computational Biology

Editorial, 2014. A bittersweet celebration of crystallography. Nat Methods. Jun;11(6):593. doi: 10.1038/nmeth.2995.

Gore, S., García, E.S., Hendrickx, P.M., Gutmanas, A., Westbrook, J.D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J.M., Hudson, B.P., Ikegawa, Y. et al, 2017. Validation of structures in the Protein Data Bank. Structure, 25(12), pp.1916-1927.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. Journal of Molecular Biology

Grabowski, M., Niedzialkowska, E., Zimmerman, M.D. and Minor, W., 2016. The impact of structural genomics: the first quindecennial. *Journal of structural and functional genomics*, *17*, pp.1-16.

Hamm, G.H. and Cameron, G.N., 1986. The EMBL data library. Nucleic acids research, 14(1), pp.5-9.

IUCr Commission on Biological Macromolecules (1989). Acta Cryst. A45, 658. https://journals.iucr.org/a/issues/1989/09/00/es0121/es0121.pdf

Jordan, E. and Carrico, C., 1982. DNA database. Science, 218(4568), pp.108-108.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., et al, 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), pp.583-589.

Karlin, S., & Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences

Koetzle, T.F., 1989. Benefits of databases. Nature, 342(6246), pp.114-114.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G.R., Morey-Burrows, F.S., Anishchenko, I., Humphreys, I.R. and McHugh, R. et al, 2024. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science, 384(6693), p.eadl2528.

Levinthal, C. 1966. Molecular model-building by computer, Scientific American, 214(6). pp. 42-52.

Levinthal, C. (1969) How to fold graciously, in: J.T.P. DeBrunner, E. Munck (Eds.) Mossbauer Spectroscopy in Biological Systems Proceedings, Univ of Illinois Press, Illinois, 67(41), pp. 22-24.

Lin, D., Manning, N.O., Jiang, J., Abola, E.E., Stampf, D., Prilusky, J., and Sussman, J.L. 2000. AutoDep©: A web based system for deposition and validation of macromolecular structural information, Acta Crystallographica D Biological Crystallograrphy, 56(7) pp. 828-841.

Maddox, J., 1989. Making good databanks better. Nature, 341(6240), pp.277-277.

Mirdita, M., Von Den Driesch, L., Galiez, C., Martin, M. J., Söding, J., & Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, *45*(D1), D170-D176.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature methods*, *19*(6), 679-682.

Needleman, S.B., & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C. and Baker, D., 2017. Protein structure determination using metagenome sequence data. *Science*, *355*(6322), pp.294-298.

Prilusky, J., Sussman, J.L. and Abola, E.E. (1996) Three dimensional database of biomacromolecules structure (3DB): a 'multi-tool' based browser as a solution for complex data and complex queries. in Abstract of 10 Anniversay of the Swiss-Prot Database and the 25th anniversary of the Protein Data Bank [https://www.weizmann.ac.il/csb/faculty_pages/Sussman/pdb25sp10/abstracts/Abola.html]

Qian, N. and Sejnowski, T.J., 1988. Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology*, *202*(4), pp.865-884.

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, *9*(2), 173-175.

Roberts, R.J., 1989. Benefits of databases. Nature, 342(6246), pp.114-114.

Rohl, C.A., Strauss, C.E., Misura, K.M. and Baker, D., 2004. Protein structure prediction using Rosetta. In Methods in enzymology (Vol. 383, pp. 66-93). Academic Press.

Sayle R.A. and Milner-White, E.J. (1995). RASMOL: biomolecular graphics for all. TIBS, 20(9) pp. 374-376

Seeman, N.C., Sussman, J.L., Berman, H.M., & Kim, S.-H. (1971). Nucleic acid conformation: crystal structure of a naturally occurring dinucleoside phosphate (UpA). Nature New Biology, 233(37). pp. 90-92.

Sejnowski, T.J., 2018. The deep learning revolution. MIT Press.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A. and Penedones, H., et al, 2020. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), pp.706-710.

Smith, T.F., & Waterman, M.S. (1981). Identification of common molecular subsequences. Journal of Molecular Biology

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. Bioinformatics

Stampf, D.R., Felder, C.E., & Sussman, J..L. (1995). PDBbrowse--a graphics interface to the Brookhaven Protein Data Bank, Nature, 374(6522) pp. 572-574.

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology

Steinegger M. & Söding J. (2018). Clustering huge protein sequence sets in linear time. Nature communications 9(1) 2542.

Strasser, B.J., 2008. GenBank-Natural History in the 21st Century?. Science, 322(5901), pp.537-538.

Strasser, B.J., 2019. Collecting experiments: Making big data biology. University of Chicago Press.

Sussman, J.L, Lin, D., Jiang, J., Manning, N.O., Prilusky, J. & Abola, E.E. 2001. The protein data bank at Brookhaven, in: M.G. Rossmann, E. Arnold (Eds.) International Tables for Crystallography, Volume F. Crystallography of Biological Macromolecules, Kluwer Academic Publishers, Dordrecht, pp. 649-656.

UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49(D1), p.D480.

Vaswani, A. et al, 2017. Attention is all you need. Advances in Neural Information Processing Systems.
arxiv.org/abs/1706.03762

Wlodawer A, Davies D, Petsko G, Rossmann M, Olson A, Sussman JL. Immediate release of crystallographic data: a proposal. Science. 1998 Jan 16;279(5349):306-7.

https://www.science.org/doi/10.1126/science.279.5349.302e