Computing in the Life Sciences: From Early Algorithms to Modern AI

This manuscript (permalink) was automatically generated from In-Vivo-Group/generative-biology@547185c on June 17, 2024.

Authors

- **Samuel A. Donkor**
  · samadon1
  In Vivo Group

- **Matthew E. Walsh**
  0000-0003-1514-7761 · mwalsh52
  U.S. National Security Commission on Emerging Biotechnology; Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health

- **Alexander J. Titus** ✉
  0000-0002-0145-9564 · alexandertitus
  In Vivo Group; U.S. National Security Commission on Emerging Biotechnology; Information Sciences Institute & Iovine and Young Academy, University of Southern California

✉ — Correspondence possible via GitHub Issues or email to Alexander J. Titus <publications@theinvivogroup.com>.

Abstract

Computing in the life sciences has undergone a transformative evolution, from early computational models in the 1950s to the applications of artificial intelligence (AI) and machine learning (ML) seen today. This paper highlights key milestones and technological advancements through the historical development of computing in the life sciences. The discussion includes the inception of computational models for biological processes, the advent of bioinformatics tools, and the integration of AI/ML in modern life sciences research. Attention is given to AI-enabled tools used in the life sciences, such as scientific large language models and bio-AI tools, examining their capabilities, limitations, and impact to biological risk. This paper seeks to clarify and establish essential terminology and concepts to ensure informed decision-making and effective communication across disciplines.

The views and opinions expressed within this manuscript are those of the authors and do not necessarily reflect the views and opinions of any organization the authors are affiliated with.

Introduction

Computing technologies have become indispensable to life scientists, changing how research is conducted and expanding the scope of scientific discovery. The history of computing in the life sciences is marked by significant milestones that have advanced research, including early algorithmic approaches to the application of artificial intelligence (AI) and machine learning (ML). Early uses of computers in the 1950s for population genetics calculations and the pioneering work of Alan Turing in biological morphogenesis set the stage for subsequent developments. Over the following decades,

computational biology evolved from basic protein structure analysis to complex genomic studies, driven by advancements in DNA sequencing and computing.

Today, the terms AI, ML, deep learning, and large language models (LLMs) are often used interchangeably in the life sciences. Although these terms are related, they each have distinct meanings (Figure 2). AI broadly refers to machines designed to mimic human intelligence. ML is a subset of AI focused on algorithms that improve through experience. Deep learning is a subset of ML involving neural networks with many layers that can learn from vast amounts of data, and LLMs, such as GPT (Generative Pre-trained Transformer) models like ChatGPT, are a specific type of deep learning that excel in understanding and generating human-like text. By processing and analyzing biological data at unprecedented scale and speeds, these technologies have advanced fields such as bioinformatics, structural biology, and genomics. Understanding distinctions among AI-related nomenclature is crucial as technology development accelerates. Decisions about funding, regulation, new product development, and the implementation of new technologies rely on an accurate understanding of what these technologies can and cannot do. A nuanced understanding of the capabilities and limitations of AI, ML, LLMs and other computational tools can help to correctly estimate their potential and effectively utilize valuable resources.

This paper provides an overview of historical context, current applications, and future directions of computing in the life sciences. By explaining key terms, concepts, and timelines, we aim to bridge the knowledge gap between practitioners and stakeholders, fostering an environment for progress that supports scientific innovation and public benefit outcomes.

Computers, Algorithms and the Internet

1950s and 1960s: Early computers and algorithms

Computers were used in the early 1950s for population genetics calculations [1]. The inception of computational modeling in biology coincides with the origins of computer science itself. British mathematician and logician Alan Turing, often referred to as “the father of computing”, used primitive computers to implement a model of biological morphogenesis (the emergence of pattern and shape in living organisms) in 1952 [2]. At about the same time, a computer called MANIAC was used for measuring speculative genetic codes; it was originally built for weaponry research at the Los Alamos National Laboratory in New Mexico [3].

Computers were used for the study of protein structure by the 1960s, and other increasingly diverse analyses. These developments marked the rise of the computational biology field, stemming from research focused on protein crystallography, in which scientists found computers indispensable for carrying out laborious Fourier analyses to determine the three-dimensional structure of proteins [4,5].

In addition to advances in determination of protein structures through crystallography, the first sequence of protein, insulin, was published [6,7]. More efficient protein sequencing methods, such as the Edman degradation technique [8], enabled sequencing 15 different proteins over a decade [9]. COMPROTEIN, one of the first bioinformatics softwares developed in the early 1960s, was designed to overcome the limitations of Edman sequencing [10]. In an effort to simplify the handling of protein sequence data for the COMPROTEIN software, a one-letter amino acid code was developed [11]. This one-letter code was first used in the Atlas of Protein Sequence and Structure [12], the first biological sequence database, laying the groundwork for paleogenetic studies.

Development of methods to compare protein sequences followed. The Needleman-Wunsch algorithm [13], the first dynamic programming algorithm developed for pairwise protein sequence alignments,

was introduced in the 1970s. Multiple sequence alignment (MSA) algorithms followed in the early 1980s. Progressive sequence alignment was introduced by Feng and Doolittle in 1987 [14]. The MSA software CLUSTAL, a simplification of the Feng-Doolittle algorithm [15] was developed in 1988. It is still used and maintained to this day [16].

1970s: From protein to DNA analysis

The deciphering of all 64 triplet codons of the genetic code in 196817 fueled a desire to efficiently determine the sequence of DNA that existed into the 1970s. This desire led to the development of cost-efficient DNA sequencing methods, such as the Maxam-Gilbert and Sanger sequencing techniques in the mid-1970s [6,7,17]. With this new ability to generate DNA sequence data, a paradigm shift from protein analysis to DNA analysis occurred in the late 1970s. Concurrently, concerns over recombinant DNA research led to safety protocols established during the 1975 Asilomar conference [18].

New DNA sequencing techniques resulted in significantly more data to be analyzed, a task at which computation could help. The first software dedicated to analyzing Sanger sequencing reads was published in 1979 [19]. DNA sequences began to be utilized in phylogenetic inference with pioneering methods like maximum likelihood for inferring phylogenetic trees from DNA sequences [20]. Several bioinformatics tools and statistical methods were developed following this work. The adoption of Bayesian statistics in molecular phylogeny in the 1990s was inspired by this [21] and is still commonly used in biology today [22]. Yet, numerous computational limitations needed to be overcome during the latter half of the 1970s to expand the utilization of computing in the life sciences, especially in DNA analysis. The subsequent decade proved instrumental in addressing these challenges.

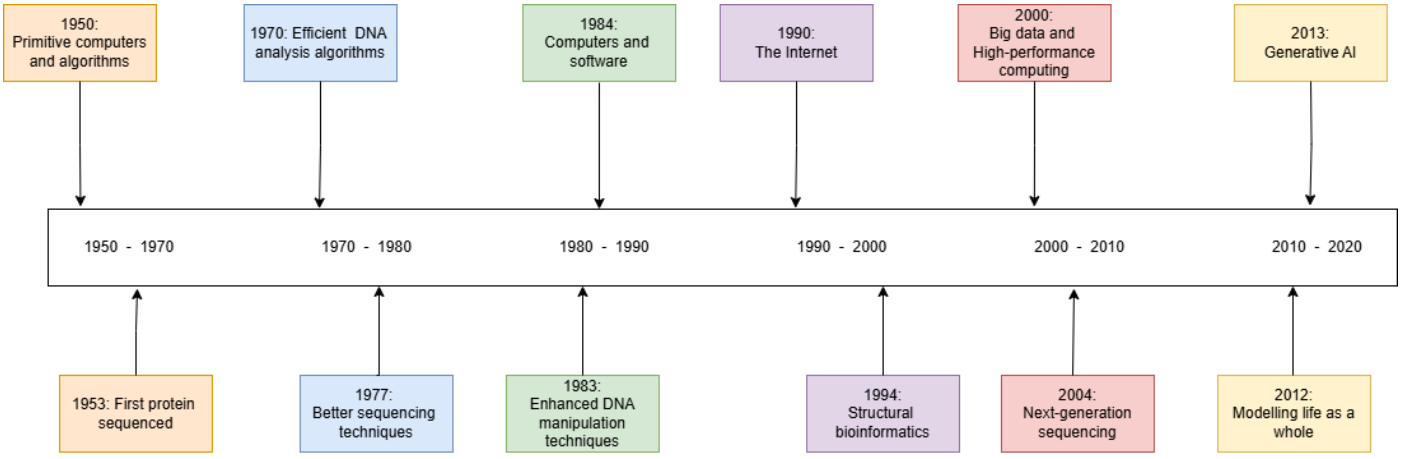

Figure 1: The history of parallel advancements in computing and the life sciences: A timeline of major milestones.

1980s: Simultaneous advances in computing and biology

Parallel advancements in biology and computing propelled bioinformatics forward during the 1980s and 1990s. Molecular techniques like gene targeting and amplification, using enzymes like restriction endonucleases and DNA ligases, laid the groundwork for genetic engineering [18]. The polymerase chain reaction (PCR) transformed gene amplification, while innovations like Taq polymerase and thermal cyclers optimized the process [23].

Computing accessibility surged with microcomputers like the Commodore PET, Apple II, and Tandy TRS-80, along with bioinformatics software like the GCG software suite [24] and DNASTAR [25], another sequence manipulation suite capable of assembling and analyzing Sanger sequencing data. Other sequence manipulation suites were developed to run on CP/M, Apple II, and Macintosh

computers [26] in the years 1984 and 1985. Free code copies of this software were offered on demand by some developers. This propelled an upcoming software-sharing movement in the programming world [27,28].

The free software movement, led by the GNU project, promoted open-source bioinformatics tools. Major sequence databases (EMBL, GenBank, DDBJ) standardized data formatting and enabled global sharing. Bioinformatics journals, like CABIOS, which is now known as Bioinformatics (Oxford, England) accentuated computational methods' importance. Desktop workstations with Unix-like systems and scripting languages aided bioinformatics analyses, and scripting languages simplified tool development.

1990s: The genomics era and web-based bioinformatics

The genomics era began in the mid-1990s with the complete sequencing of the Haemophilus influenzae genome [29], initiating genome-scale analyses. This milestone was followed by the publication of the human genome at the beginning of the 21st century, which served as the definitive catalyst for the genomic era [30]. This transformative event spurred the design and development of several specialized Perl-based software to assemble whole-genome sequencing reads: PHRAP [31], Celera Assembler [32] among others.

Tim Berners-Lee's pioneering work at CERN in the early 1990s resulted in the World Wide Web, transforming global communication and ushering in an era of unprecedented access to information. With the advent of the internet, researchers gained a powerful platform to share and access vast amounts of biological data efficiently. This facilitated collaborative efforts in biology and genomics, leading to the establishment of foundational databases such as the EMBL Nucleotide Sequence Data Library [33] and the GenBank database became the responsibility of the NCBI [34] in 1992. Also, the famous NCBI website came online in 1994, featuring the efficient pairwise alignment tool BLAST [35]. After that, the world saw the birth of major databases we still rely on today: Genomes (1995), PubMed (1997), and Human Genome (1999) [36,37,38].

The proliferation of web-based resources transformed access to bioinformatics tools, democratizing their availability and usability for researchers worldwide. Through the development of web platforms, bioinformatics tools became more user-friendly and accessible. This shift enabled researchers to interact with sophisticated analytical tools without needing extensive computational expertise or access to specialized hardware. Consequently, the widespread adoption of web-based bioinformatics resources facilitated broader participation in genomic and molecular research, accelerating scientific discovery and collaboration on a global scale. Graphical web servers emerged as a convenient alternative to traditional UNIX-based systems, simplifying data analysis without the need for complex installations. The continued relevance of servers for scientific purposes is exemplified by the AlphaFold Server which uses the latest AlphaFold 3 model [39], released in 2024, to provide highly accurate biomolecular structure predictions in a unified platform.

The internet facilitated the dissemination of scientific research through online publications, challenging traditional print-based methods. Early initiatives like BLEND [40] paved the way for internet-based scientific publishing by shedding insights into the potentials and obstacles associated with using the internet for scientific publications. This study paved the way for leveraging the Internet for both data set storage and dissemination, leading up to the establishment of preprint servers like arXiv (est. 1991) [41] and bioRxiv (est. 2013) [42] which changed the way scientific findings are shared and accessed. These platforms democratized access to scientific knowledge by enabling researchers to share their work rapidly and openly, facilitating interdisciplinary collaborations and the cross-pollination of ideas.

The experimental determination of the first three-dimensional structure of a protein, specifically, myoglobin, occurred in 1958 via X-ray diffraction [4]. However, earlier groundwork by Pauling and Corey with the publication of two articles in 1951 that reported the prediction of α-helices and β-sheets [43] laid the foundation for predicting protein structures. Similar to advances in other biological sciences, the utilization of computers has made it feasible to conduct calculations aimed at predicting the secondary and tertiary structure of proteins, with varying levels of confidence. This capability has been notably enhanced by the development of fold recognition algorithms, also known as threading algorithms [44,45]. However, proteins are dynamic entities, requiring advanced biophysical models to describe their interactions and movements accurately. Force fields have been formulated to describe the interactions among atoms, enabling the introduction of tools for modeling the molecular dynamics of proteins during the 1990s [46]. Used to study the behavior and interactions of atoms and molecules over time, molecular dynamics simulations calculate the positions and velocities of atoms based on physical principles. Despite the theoretical advancements and availability of tools, executing molecular dynamics simulations remained challenging in practice due to the substantial computational resources they demanded.

Graphical processing Units (GPUs) have made molecular dynamics more accessible [47], with applications extending to other bioinformatics fields requiring intensive computation. However, the internet's role in data dissemination, coupled with increasing computational power, has led to the proliferation of 'Big Data' in bioinformatics.

2000s: High-throughput sequencing and big data

Second-generation sequencing technologies democratized high-throughput bioinformatics. For example '454' pyrosequencing, a high-throughput DNA sequencing technique played a significant role in advancing genomics research by enabling rapid and cost-effective sequencing of DNA samples, particularly for applications such as whole-genome sequencing [48], but computational challenges arose with increased data volumes. Decreasing sequencing costs resulted in more data being generated, emphasizing data organization and accessibility. Specialized repositories and standardization efforts were needed to ensure data interoperability. High-performance computing adaptation became vital to address the increased amounts of data within bioinformatics projects. The surge in bioinformatics projects, accompanied by a vast influx of data, prompted adjustments from funding bodies to accommodate the demand for high-performance computing resources and collaborative initiatives.

While basic computer setups suffice for some projects, others demand complex infrastructures and substantial expertise. Government-sponsored entities like Compute Canada, New York State's High-Performance Computing Program, The European Technology Platform for High-Performance Computing, and National Center for High-Performance Computing served researchers' computational needs. Companies like Amazon, Microsoft, and Google, among many others, offer bioinformatics and life sciences services, emphasizing the field's importance.

Table 1. Organizations providing High-Performance Computing Resources for Bioinformatics and Life Sciences

Organization	Computing Resources
Compute Canada	Provides high-performance computing resources and support services to researchers and innovators across Canada. They offer supercomputers, cloud platforms, data storage, and training programs to advance scientific research and innovation in various fields.

Organization	Computing Resources
New York State's High-Performance Computing Program	Provides researchers, businesses, and educational institutions with access to high-performance computing (HPC) resources and expertise to support their computational research and development efforts.
The European Technology Platform for High-Performance Computing	Fosters collaboration among industry, research, and academic stakeholders to advance high-performance computing (HPC) technology in Europe.
National Center for High-Performance Computing	Facility for high-performance computing (HPC) resources including large-scale computational science and engineering, cluster and grid computing, middleware development, visualization and virtual reality, data storage, networking, and HPC-related training.
National Center for Supercomputing Applications	Offers high-performance computing resources such as the Blue Waters supercomputer, provides advanced data storage solutions, data analysis, and visualization tools, and supports interdisciplinary research in fields such as astrophysics, climate modeling, and genomics.
Oak Ridge Leadership Computing Facility	Provides supercomputing resources, such as the Summit supercomputer, for scientific research, offers support services including software development, data storage, and visualization, and facilitates research in various fields including climate science, biology, and materials science.
Swiss National Supercomputing Centre	Provides high-performance computing systems including the Piz Daint supercomputer, offers cloud computing services, data management, and user support, and facilitates scientific research in areas such as climate modeling, physics, and life sciences.
Barcelona Supercomputing Center	Provides access to MareNostrum, one of the most powerful supercomputers in Europe, offers resources for high-performance computing, data storage, and computational sciences, and supports research in fields including bioinformatics, computational biology, and engineering.
Japan's RIKEN Center for Computational Science	Houses the Fugaku supercomputer, one of the world's fastest supercomputers, provides resources for computational science, data processing, and artificial intelligence, and supports research in fields such as life sciences, materials science, and disaster prevention.
National Supercomputing Centre Singapore	Provides high-performance computing resources and support services, offers data storage, cloud computing, and software development services, and supports research in fields including bioinformatics, environmental modeling, and smart cities.

Community computing platforms democratized participation and expanded bioinformatics research's reach. Platforms like BOINC enabled broad participation in bioinformatics. Experts can submit computing tasks to BOINC, while non-experts and science enthusiasts can volunteer their computer resources to process these tasks. Several life sciences projects are available through BOINC, including protein-ligand docking, malaria simulations, and protein folding [49].

2010+: The present and future

The integration of computers into biology has ushered in a new era of research possibilities, allowing for increasingly complex studies. While before, the focus was on individual genes or proteins, advancements today enable the analysis of entire genomes or proteomes [50]. This shift toward a holistic approach in biology is evident in disciplines like genomics, proteomics, and glycomics, which have limited interconnection between them.

The next leap at the intersection of computing and the life sciences lies in modeling entire living organisms and their environments simultaneously, integrating all molecular categories. This has already been achieved in a whole cell model of Mycoplasma genitalium, in which all its genes, products and their known metabolic interactions have been reconstructed [51]. Driven by advancements in measurement techniques, improved computational performance and artificial intelligence (AI) techniques, whole-cell modeling is increasingly becoming realistic and feasible. In contrast to traditional bottom-up approaches relying on molecular interaction networks, a predictive model has been developed for genome-wide phenotypes of budding yeast using deep learning [52]. The main applications of whole-cell modeling have been in producing useful substances and discovering drugs, such as antimicrobials [53,54,55,56] since whole-cell modeling was first directed toward unicellular organisms. Meanwhile, models of cultured human cells have also been developed, which have found applications in cell differentiation and medical research [57]. The possibility of modeling entire multicellular organisms may not be far off, considering the rapid pace of technological and computational advancements like artificial intelligence (AI) .

Artificial Intelligence (AI)

Artificial intelligence (AI) refers to a set of tools, techniques and paradigms that enable computers to mimic human behavior and either replicate the decision-making process typically performed by humans or exceed human performance in solving complex tasks independently or with minimal human intervention [58]. AI is concerned with a variety of central problems, including knowledge representation, reasoning, learning, planning, perception, and communication. It also refers to a variety of tools and methods, including case-based reasoning, rule-based systems, genetic algorithms, fuzzy models, and multi-agent systems [59]. Early AI research focused primarily on hard-coded statements in formal languages, which a computer can then automatically reason about based on logical inference rules. These computer systems known as expert systems, excelled in specific domains but lacked adaptability. Over time, AI has evolved to include a variety of approaches, each with its own strengths and weaknesses. For instance, expert systems are highly accurate within narrow fields but struggle with tasks outside their programmed knowledge. In contrast, machine learning algorithms can generalize from data and adapt to new situations, though they require large datasets and extensive training. Other AI techniques, such as deep learning, neural networks, and natural language processing also offer their own unique advantages and challenges.

Expert systems

Expert systems are a type of artificial intelligence (AI) that aims to replicate the decision-making capabilities of human experts in specific domains. They are made of a knowledge base containing domain-specific facts, rules, and heuristics, and an inference engine that applies logical reasoning to this knowledge to draw conclusions or make decisions [60]. Users are typically able to input queries and receive advice or recommendations through a simplified user interface. The primary user action, which involves pointing and clicking, is known as selecting [61].

An expert system for chemical analysis was developed in 1965 by AI researcher Edward Feigenbaum and geneticist Joshua Lederberg. This system was originally known as Heuristic DENDRAL and later as DENDRAL [62]. DENDRAL was developed to analyze molecular structures, particularly those containing elements like carbon, hydrogen, and nitrogen, based on spectrographic data. It proposed molecular structures for the compounds, with accuracy comparable to that of expert chemists.

Edward Shortliffe's work on MYCIN [63] began in 1972 at Stanford University. MYCIN, an expert system, was designed to assist physicians in diagnosing and selecting therapies for patients with bacterial infections, particularly patients with meningitis. It used a rule-based system that analyzed

patient symptoms and medical history to suggest appropriate antibiotic treatments. MYCIN exhibited proficiency equivalent to infectious disease doctors.

However, despite their capabilities, the paradigm faces several limitations as humans generally struggle to explicitly articulate all their tacit knowledge that is required to perform complex tasks [64], leading to challenges such as difficulty in extrapolation, handling out-of-distribution data, managing uncertainty, and addressing biases. These limitations arise because expert systems heavily rely on predefined rules and knowledge encoded by humans. Consequently, the involvement of humans in specifying these parameters is essential but can also introduce limitations due to human cognitive constraints and biases. In contrast, machine learning algorithms overcome some of these limitations by learning from data, and making them more adaptable without relying heavily on explicit human guidance.

Machine learning and Deep learning

Machine learning (ML) is a subset of AI that focuses on the development of algorithms and statistical models that enable computers to perform tasks without being explicitly programmed to do so [65]. It involves the use of data and algorithms to imitate the way humans learn, gradually improving the system's performance on a specific task over time through iterative learning processes. Machine learning is effective for tasks such as classification, regression, and clustering, particularly when they involve high-dimensional data. These algorithms analyze data, identify patterns, and make predictions or decisions without being explicitly programmed for each task.

Based on the given problem and the available data, there are many potential model and training paradigms, three of the most prominent types of ML being: supervised learning [66], unsupervised learning [67,68], and reinforcement learning [69]. The goal of machine learning is to develop an output model that can make predictions or decisions based on input data. In supervised learning, the model is trained on a labeled dataset, where each training example is paired with an output label. A label is the desired output or result for a given piece of data. For example, in an image recognition task, labels could be the names of objects in the images (e.g., "cat," "dog," "car"). In a spam detection task, emails could be labeled as "spam" or "not spam.". The goal is to learn a mapping from inputs to outputs. Unsupervised learning involves training a model on data without labeled responses. The goal is to uncover patterns or structures within the data. In reinforcement learning, an agent learns to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties based on its actions and learns to maximize cumulative rewards over time.

Depending on the learning task, the field offers various classes of ML algorithms, each of them coming in multiple specifications and variants, including regression models, instance-based algorithms, decision trees, Bayesian methods, and artificial neural networks, among others.

Artificial neural networks (ANNs) span all three major types of machine learning. ANNs are inspired by biological systems and consist of interconnected processing units called neurons, with connections akin to synapses in the human brain. Signals are processed based on thresholds set by activation functions, and organized into layers for input, hidden, and output layers. Shallow machine learning encompasses simpler ANNs and other algorithms, often being more interpretable than deep neural networks. Deep neural networks, which have multiple hidden layers, perform complex calculations to automatically discover patterns in data. This ability is known as deep learning64. Deep learning excels with large, high-dimensional data like text, images, and videos, while shallow learning may outperform with low-dimensional data or limited training data. Time series, image, and text data present various application domains.

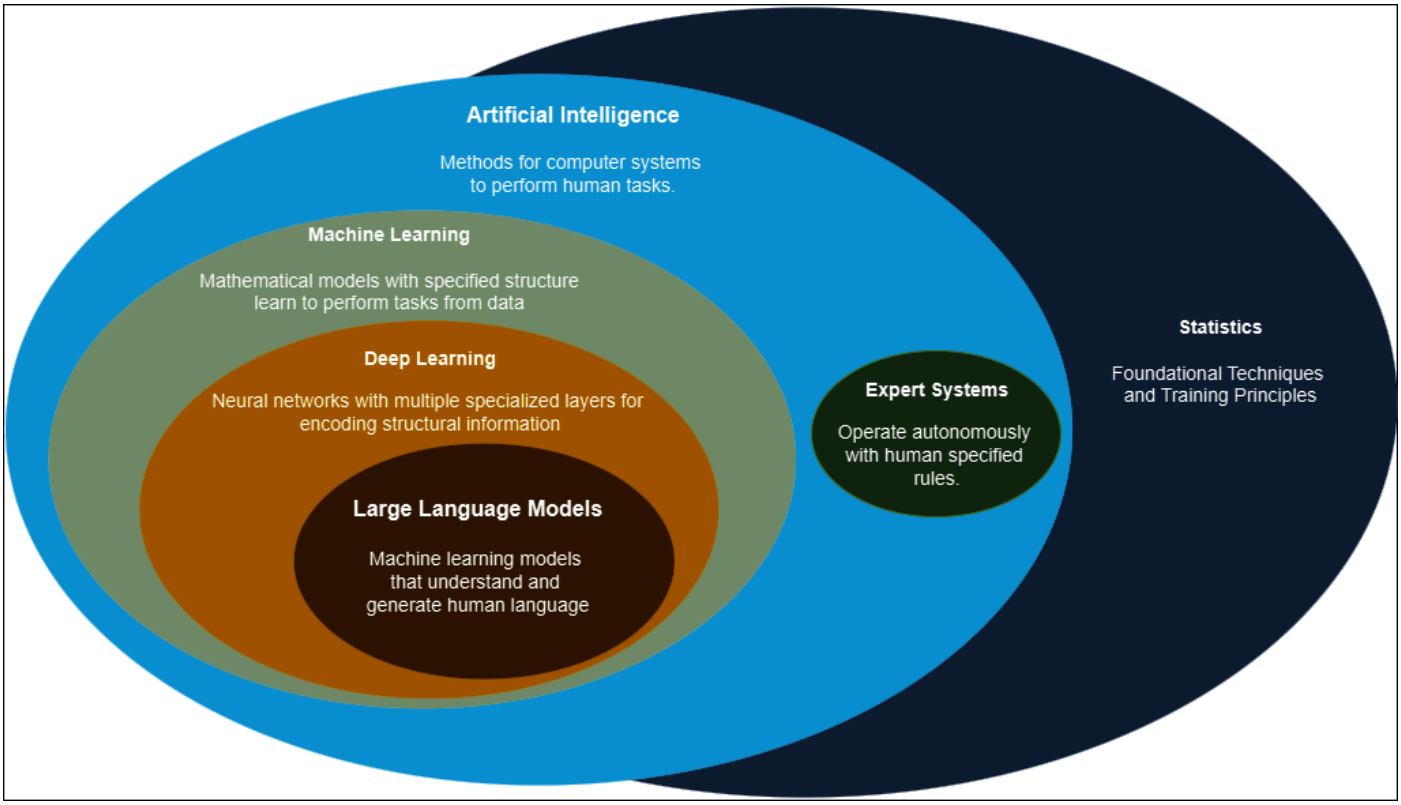

Figure 2: Relationship between statistics, artificial intelligence, expert systems, machine learning, deep learning and large language models.

Automated model building in machine learning involves using input data for pattern identification relevant to the learning task. Shallow machine learning relies on predefined features such as pixel values in images or word frequencies in text. For example, in image classification, shallow learning might rely on handcrafted features like color histograms or edge detectors. In contrast, deep learning can operate directly on high-dimensional raw input data, such as the raw pixel values of an image or the sequence of words in text. It automatically learns features at multiple levels of abstraction, allowing it to capture patterns in the data without the need for manual feature engineering. For instance, in image classification with deep learning, the model learns to detect edges, shapes, and textures from raw pixel data, resulting in improved accuracy [70].

Deep learning architectures often combine both aspects into end-to-end systems or extract features for use in other learning subsystems. Various deep learning architectures have emerged, including convolutional neural networks (CNNs) [71], recurrent neural networks (RNNs) [72], distributed representations [73], autoencoders [74], generative adversarial neural networks (GANs) [75], among others. CNNs excel in computer vision and speech recognition tasks, learning hierarchical features essential for image recognition. RNNs specialize in sequential data structures like time-series data and natural language processing (NLP), addressing the challenges of vanishing gradients through advanced mechanisms like long short-term memory (LSTM) networks [76]. Distributed representations, such as word embeddings, play a crucial role in NLP tasks by projecting language entities into numerical representations, preserving semantic relationships between words. Autoencoders provide dense feature representations and are applied for unsupervised feature learning, dimensionality reduction, and anomaly detection. GANs, belonging to generative models, learn probability distributions over training data to generate new data samples, using a generator-discriminator framework in a non-cooperative game setting.

Generative AI and Transformers

Generative AI (GenAI) analyzes vast amounts of data, looking for patterns and relationships, then uses these insights to create fresh, new content that mimics the original data [77]. It does this by leveraging machine learning models, especially unsupervised and semi-supervised algorithms. There are three popular techniques for implementing Generative AI: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Transformers.

Variational Autoencoders (VAEs) [78] first introduced by Diederik P. Kingma et al. in 2013 are generative models in unsupervised machine learning that generate new data similar to the input data. They consist of an encoder that compresses the input data into a lower-dimensional latent space by producing parameters for a probability distribution (mean and variance). The decoder reconstructs the data from this latent representation. The loss function, which combines reconstruction loss and regularization loss (KL Divergence), ensures the output data is both accurate and diverse. VAEs are used in applications like image generation, data imputation, anomaly detection, offering a flexible framework for generating and understanding data despite some challenges in balancing the loss components and achieving high-quality outputs [79,80,81].

In 2014, GANs [75] were proposed by researchers at the University of Montreal. GANs use two models that work in tandem: One learns to generate a target output (like an image) and the other learns to discriminate true data from the generator's output. The generator tries to fool the discriminator, and in the process learns to make more realistic outputs. The image generator StyleGAN [82] is based on these types of models.

Diffusion models [83] were introduced a year later by researchers at Stanford University and the University of California at Berkeley. By iteratively refining their output, these models learn to generate new data samples that resemble samples in a training dataset and have been used to create realistic-looking images. A diffusion model is at the heart of the text-to-image generation system Stable Diffusion [84].

Recurrent neural networks (RNNs) and their variants like long short-term memory (LSTM) networks are commonly used for sequential data processing tasks. However, these models suffer from limitations such as vanishing gradients and inefficiency in parallelization. Transformers revolutionized the field with the ability to capture long-range dependencies in sequential data efficiently and was first reported in the seminal 2017 paper, "Attention is All You Need" [85]. The introduction of transformers, with their superior performance and scalability, initiated a departure from RNNs. Transformers were used to train the large language models (LLMs) that power ChatGPT [86].

The transformer architecture consists of an encoder and a decoder, each with multiple layers of self-attention and feedforward neural networks. The self-attention mechanism enables the model to assess the significance of a piece of data, such as a word in a sentence, based on that word's relations with other words in the sentence. To preserve the ordering of the words and the meaning of the sentence, the transformer incorporates positional bias to maintain the relative positions of words within a sentence.

The transformer encoder-decoder architecture performs well at tasks like language translation. In a language translation task, the model transforms a sentence by encoding inputs from one language and then decoding outputs in another. The encoder processes the input sentence and creates a fixed-size vector representation, which the decoder then uses to generate the output sentence. The encoder-decoder employs both self-attention and cross-attention mechanisms, where self-attention is applied to the decoder's inputs, and cross-attention focuses on the encoder's output.

A prominent example of the transformer encoder-decoder architecture is Google's T5 (Text-to-Text Transfer Transformer) [87], introduced in 2019. T5 can be fine-tuned for various NLP tasks, including language translation, question answering, and summarization.Real-world applications of the

transformer encoder-decoder architecture include Google Translate, which utilizes the T5 model for translating text between languages, and Facebook's M2M-10080, a multilingual machine translation model capable of translating among 100 different languages.

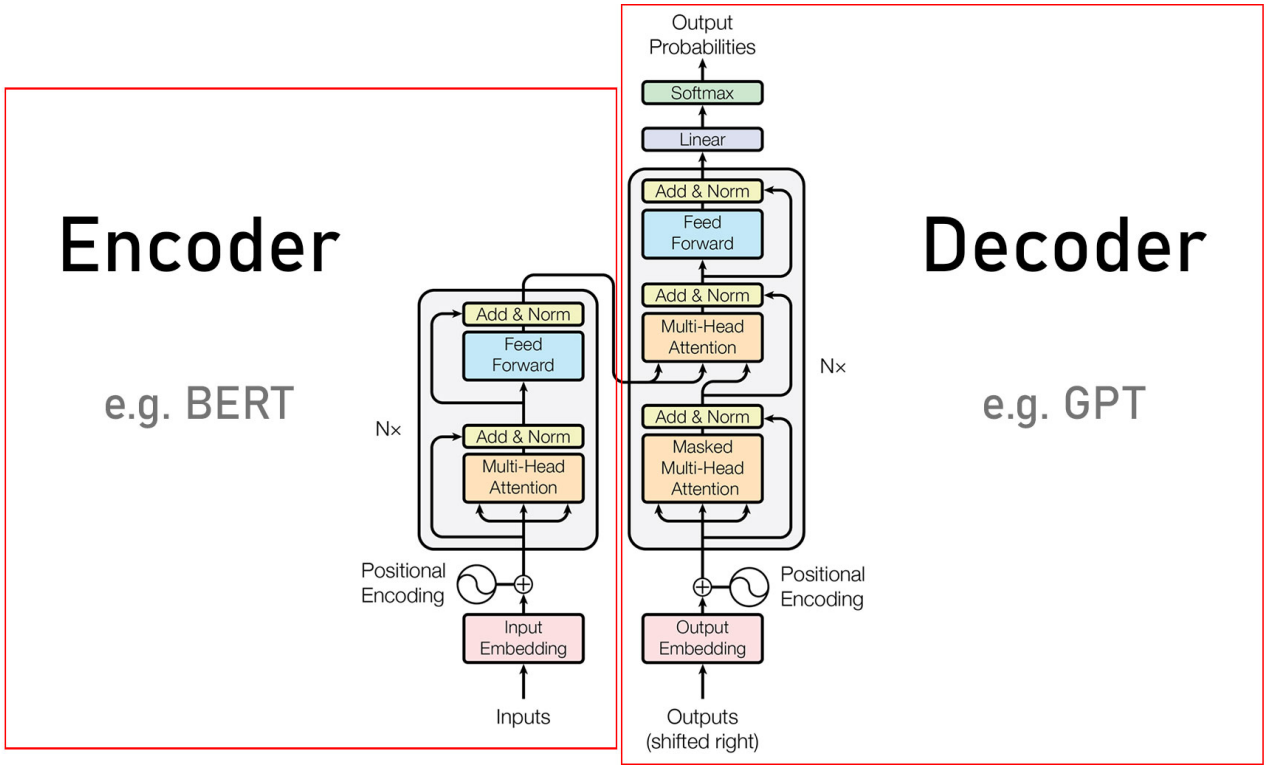

Figure 3: The encoder-decoder structure of the Transformer architecture. Adapted from "Attention Is All You Need" **Encoder-only models**: Ideal for tasks requiring a deep understanding of the input, such as sentence classification and named entity recognition. **Decoder-only models**: Suited for generative tasks like text generation. **Encoder-decoder models (or sequence-to-sequence models)**: Best for generative tasks that depend on an input, such as translation or summarization.

Transformer Encoder

The transformer encoder architecture is used for tasks such as text classification, where the goal is to categorize a piece of text into predefined categories. Text classification tasks include determining the sentiment of a piece of text, determining the topic and detecting if the text is spam. The encoder processes a sequence of tokens and produces a fixed-size vector representation of the entire sequence, which is then used for classification. The most notable transformer encoder model is BERT (Bidirectional Encoder Representations from Transformers) [88], introduced by Google in 2018. BERT is pre-trained on large text datasets and can be fine-tuned for a wide range of NLP tasks.

Unlike the encoder-decoder architecture, the transformer encoder focuses solely on the input sequence without generating an output sequence and instead the output is a classification task. It uses the self-attention mechanism to identify the most relevant parts of the input for the given task. Real-world applications of the transformer encoder architecture include sentiment analysis, where models classify reviews as positive or negative, and email spam detection, where models classify emails as spam or not.

Transformer Decoder

The transformer decoder architecture is tailored for tasks like language generation, where the model creates a sequence of words based on an input prompt or context. The decoder takes a fixed-size vector representation of the context and generates a sequence of words one at a time, with each word depending on the previously generated words. A well-known transformer decoder model is GPT-

3 (Generative Pre-trained Transformer 3) [89], introduced by OpenAI in 2020. GPT-3 is a large language model capable of generating human-like text across various styles and genres. ChatGPT, which is based on the GPT-3 model, was officially launched by OpenAI in November 2020. It was a significant milestone in the development of large language models (LLMs), characterized by its ability to generate human-like text across various styles and genres. Real-world applications of the transformer decoder architecture include text generation, where models generate stories or articles based on a given prompt, and chatbots, where models create natural and engaging responses to user inputs.

Large Language Models (LLMs)

Large language models are machine learning models that can comprehend and generate human language text. In the life sciences, LLMs such as GPT (Generative Pre-trained Transformer) and BERT, have revolutionized natural language processing, enabling researchers to extract insights from vast repositories of biomedical literature, accelerate drug discovery, and personalize patient care [90].

Large language models use transformer models and are trained using massive datasets — hence, large. This enables them to recognize, translate, predict, or generate text or other content. They are composed of multiple neural network layers – recurrent layers, feedforward layers, embedding layers, and attention layers work in tandem to process the input text and generate output content.

There are three main kinds of large language models:

- **Generic or raw language models** predict the next word based on the language in the training data. These language models perform information retrieval tasks.
- **Instruction-tuned language models** are trained to predict responses to the instructions given in the input. This allows them to perform sentiment analysis, or to generate text or code.
- **Dialog-tuned language models** are trained to have a dialog by predicting the next response. Think of chatbots or conversational AI.

Before functioning, LLMs undergo two crucial processes: training and fine-tuning. They are pre-trained on massive textual datasets from sources like Wikipedia and GitHub, comprising trillions of words to form a foundation model or a pre-trained model. This unsupervised learning stage allows the model to understand word meanings, relationships, and contextual distinctions, such as discerning whether “right” means “correct” or the opposite of “left.”. To perform specific tasks, pretrained models undergo fine-tuning, which tailors them to particular activities like translation. This process optimizes task-specific performance. A related method, prompt-tuning, trains the model using few-shot or zero-shot prompting. Few-shot prompting provides examples to teach the model how to respond, while zero-shot prompting directly instructs the model on the task without examples.

LLMs serve various purposes:

- **Information retrieval**: Used by search engines like Google and Bing to produce and communicate answers conversationally.
- **Sentiment analysis**: Used to evaluate the sentiment of textual data.
- **Text generation**: Powers generative AI, such as ChatGPT, to create text based on prompts.
- **Code generation**: Similar to text generation, LLMs can generate code by recognizing patterns.
- **Chatbots and conversational AI**: Facilitate customer service interactions by interpreting and responding to customer queries.

AI in the Life Sciences

The intersection of AI and the life sciences (AIxBio) has given rise to new capabilities where advanced computational techniques are applied to understand the complexities of biological systems and engineer novel solutions to pressing challenges in medicine and biotechnology [91]. The two primary modern AI categories used in the life sciences are large language models (LLMs) and bio-AI tools.

LLM-based chatbots like ChatGPT are designed to process human language inputs and generate output in human-like fashion. In the life sciences, ChatGPT can assist researchers by drafting and editing scientific manuscripts, generating hypotheses, summarizing datasets, and retrieving information from the scientific literature. LLM-based chatbots can also streamline literature reviews and facilitate the comprehension of complex biological concepts.

As a general-purpose LLM, ChatGPT and its equivalents are trained on a broad range of text from the internet. This results in models that function across topics and contexts. However, the generalist nature comes at the cost of precision and depth required for highly specialized tasks. For example, LLM-based chatbots can provide outputs with information with unfounded details, aiming to fill knowledge gaps. This behavior is known as “Confabulation”, and it can limit the utility of the tool. Furthermore, ethical concerns related to biased outputs are often attributed to biases within the training data.

Additionally, training and using general-purpose language models can be computationally expensive, time-consuming, and resource and energy intensive. Given the cost of training general purpose LLMs and their limitations, evaluations are essential for understanding their performance. Evaluations help developers identify strengths and weaknesses of the model, and often measure generalizability of models to real-world applications. This process can also identify biased or misleading model outputs. Typically, models undergo evaluation on standardized benchmarks such as GLUE (General Language Understanding Evaluation) [92], SuperGLUE [93], HellaSwag [94], TruthfulQA [95], and MMLU (Massive Multitask Language Understanding) [96] using established metrics, as shown in Table 2.

Table 2. Common Benchmarks for LLMs

Benchmark	Description	Format of Task
MMLU	MMLU (Massive Multitask Language Understanding) evaluates how well the LLM can multitask	Multiple-choice
TruthfulQA	Measures truthfulness of model responses	Generation, Multiple-choice
HellaSwag	Evaluates how well an LLM can complete a sentence	Sentence completion
SuperGLUE Benchmark	Compares more challenging and diverse tasks with GLUE, with comprehensive human baselines	Sentence- and sentence-pair classification (main task), coreference resolution and question answering
GLUE Benchmark	GLUE (General Language Understanding Evaluation) benchmark provides a standardized set of diverse NLP tasks to evaluate the effectiveness of different language models	Classification and prediction

The behavior of LLMs can be modified through model alignment, domain-specific pre-training, and supervised fine-tuning. These methods can be used to address limitations of generic LLMs, tailor behavior to meet specific requirements, and infuse general knowledge into the LLMs. Domain-specific language models, trained or fine-tuned on specific datasets relevant to particular domains, offer more

contextually accurate responses for specific domains. Evaluating domain-specific or fine-tuned models typically involves comparing their performance against a ground truth dataset if available. This process is crucial because it ensures that the model performs as expected and generates the desired outputs.

In the life sciences, specialized models can interpret complex biological data, provide detailed insights, thereby enhancing both the accuracy and reliability of the information provided. These models are known as scientific large language models (Sci-LLMs) [97].

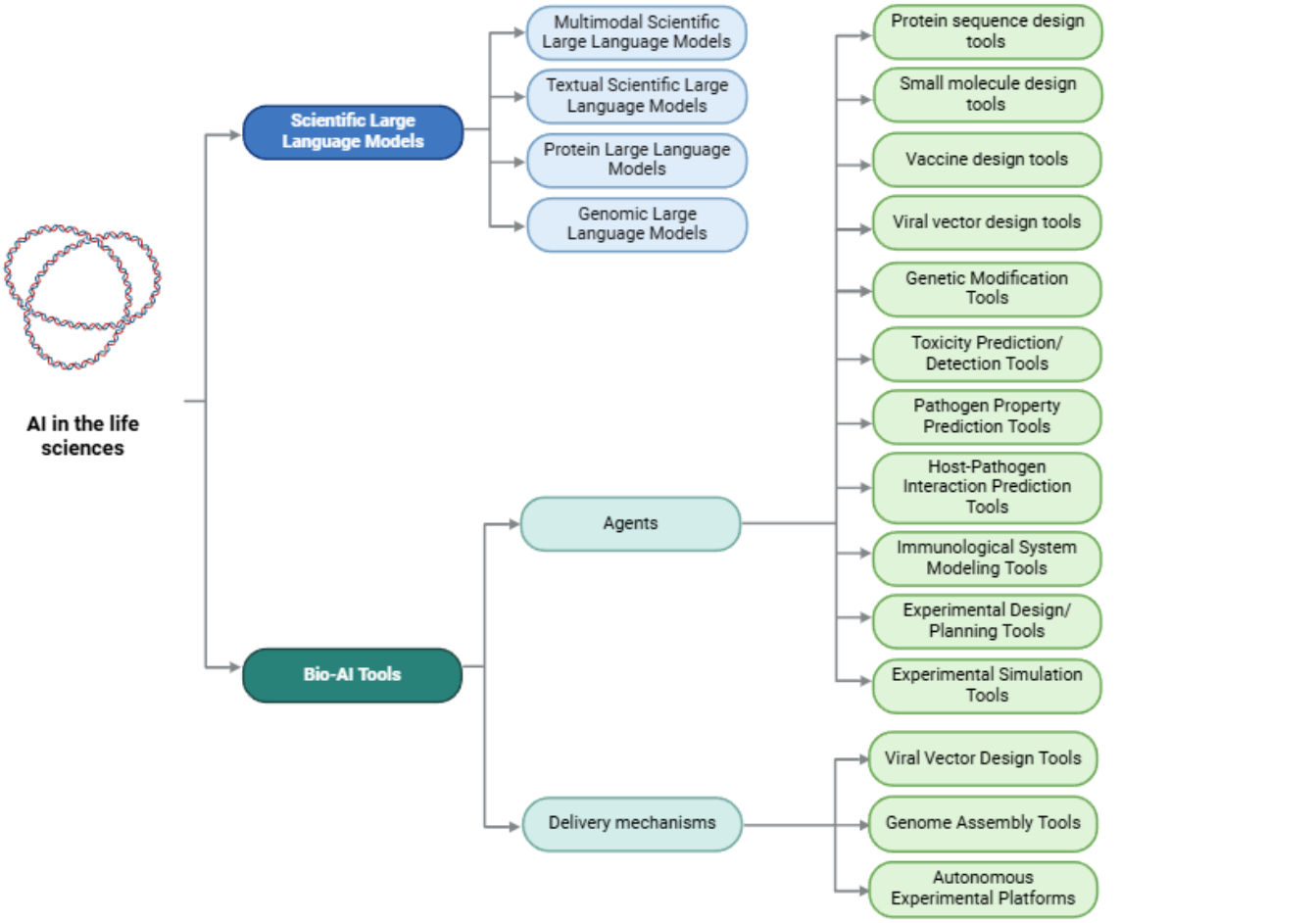

Figure 4: AI-enabled tools used in the biological sciences; Large Language Models (LLMs) and Bio-AI.

Scientific Large Language Models (Sci-LLMs)

LLMs in the life sciences have been trained on natural language, molecular, protein, and genomic sequence data. These LLMs are collectively known as Scientific Large Language Models (Sci-LLMs). Sci-LLMs are specialized models designed to process and understand various types of scientific data. They extend the capabilities of general LLMs to handle domain-specific tasks in biology, chemistry, and other scientific fields. Sci-LLMs in the biological field include Textual Scientific Large Language Models (Text-Sci-LLMs), Protein Large Language Models (Pro-LLMs), and Genomic Large Language Models (Gene-LLMs) [97].

Textual Scientific Large Language Models (Text-Sci-LLMs)

Text-Sci-LLMs are trained on vast amounts of scientific textual data, such as scientific publications. Text-Sci-LLMs excel at understanding, generating, and interacting with written human language from scientific domains. LLMs trained on vast, diverse datasets, such as BERT [88] and its variations which

have been fine-tuned specifically on biological corpora with the encoder-only architecture, have demonstrated significant potential in natural language processing (NLP) tasks within biology. Models initially trained on broad corpora such as Wikipedia and textbooks and then fine-tuned on specific biological NLP tasks, show substantial improvements in various downstream tasks including biological terminology understanding, named entity recognition, text similarity, and relation extraction [98,99,100,101,102].

GPT and its variants [89,103,104], with decoder-only architectures, have become dominant in the field of biological NLP because they can generate textual information as an output. BioGPT [105], an extension of GPT-2 [104], has been extensively fine-tuned on biomedical literature, showcasing remarkable performance in biomedical relation extraction and question answering. It also generates coherent and fluent descriptions within the biomedical context. BioMedGPT-LM [106], incrementally pre-trained on LLaMA2 [107], enables a comprehensive understanding of various biological modalities and aligns them with natural language. BioGPT and BioMedGPT-LM are both specialized language models designed for biomedical applications; however, BioGPT focuses on generating and understanding biomedical literature, while BioMedGPT-LM integrates a broader range of tasks including text generation, question answering, and classification within the biomedical domain.

Capabilities Evaluation

The evaluation of LLMs often uses Bloom's taxonomy [108,109], which includes six cognitive levels:

Table 3. Bloom's Taxonomy

Cognitive Level	Description	Examples of Activities/Tasks
Remember	Recall facts and basic concepts	List, define, identify, memorize, repeat, state
Understand	Explain ideas or concepts	Describe, explain, interpret, summarize, paraphrase, discuss
Apply	Use information or existing knowledge in new contexts	Use, demonstrate, solve, implement, execute, carry out
Analyze	Explore connections, causes, and relationships among ideas	Differentiate, organize, relate, compare, contrast, examine
Evaluate	Justify a decision or course of action based on sound analysis	Judge, critique, recommend, justify, assess, appraise
Create	Produce new or original work using existing information	Design, assemble, construct, develop, formulate, author

SciEval [110] has recently introduced a framework for evaluating scientific LLMs across four dimensions: basic knowledge, knowledge application, scientific calculation, and research ability. These dimensions are based on the cognitive domains in Bloom's taxonomy. KnowEval [97] assesses the depth of knowledge LLMs can grasp, aiming for human-level comprehension. KnowEval categorizes Text-Sci-LLMs into Pre-college, College, and Post-college levels based on the complexity of scientific knowledge.

Table 4. Categories for KnowEval

Category	Description

Category	Description
Pre-college Level	This level covers fundamental concepts and principles, aligning with the Remember and Understand stages of Bloom's taxonomy and the basic knowledge dimension of SciEval. Evaluations focus on basic knowledge comprehension, using benchmarks like MMLU [96] and C-Eval [111]
College Level	At this level, knowledge becomes more specialized and abstract, requiring logical reasoning and proof. It corresponds to the Apply and Analyze stages of Bloom's taxonomy and the knowledge application and scientific calculation dimensions of SciEval. Evaluations like PubMedQA [112] and SciQ [113] focus on this advanced understanding.
Post-college Level	This level involves mastering current knowledge and generating innovative ideas, aligning with the Evaluate and Create stages of Bloom's taxonomy and the research ability dimension of SciEval. It requires capabilities beyond standard question-answering, including summarizing advancements and designing novel experiments. Few benchmarks, such as a subset in the SciEval dataset [110], assess these high-level capabilities.

Benchmarks for Text-Sci-LLMs

Table 5. Summary of Benchmarks for Text-Sci-LLMs

Benchmark	Description	Type
MMLU	Offers a detailed and challenging benchmark that tests the comprehension and problem-solving capabilities of LLMs across a wide spectrum of tasks and subjects.	Multiple choice
C-Eval	Consists of 13,948 multi-choice questions spanning 52 diverse disciplines and four difficulty levels.	Multiple choice
AGIEval	Evaluates the general abilities of foundation models in tasks pertinent to human cognition and problem-solving.	Multiple choice
ScienceQA	A dataset designed for question answering in the scientific domain, covering various scientific topics and requiring reasoning over structured and unstructured information.	Multiple choice / Question answering (QA)
SciEval	A benchmark dataset for evaluating language models in the scientific domain, covering a range of tasks related to scientific text understanding and generation.	Multiple choice / Question answering (QA)

Benchmark	Description	Type
Bioinfo-Bench-QA	A benchmark dataset focused on question answering in the field of bioinformatics, covering topics related to biological information processing and analysis.	Multiple choice
SciQ	A dataset designed for evaluating language models in scientific question answering tasks, covering various scientific disciplines and requiring both factual and reasoning-based answers.	Multiple choice
ARC	A dataset that challenges models with questions that require a mix of comprehension and reasoning skills across a wide range of topics, including science.	Multiple choice
BLURB	A comprehensive set of datasets and tasks designed to evaluate the performance of natural language processing (NLP) models specifically in the biomedical domain.	Multiple NLP tasks
PubMedQA	A dataset designed for question answering based on biomedical literature available on PubMed, aiming to evaluate models' ability to comprehend and extract information from scientific articles.	True or False

Protein Large Language Models (Prot-LLMs).

Protein Large Language Models (Prot-LLMs) are trained on protein-related sequence data, including amino acid sequences, protein folding patterns, and other biological information. As a result, they can accurately predict protein structures, functions, and interactions. Prot-LLMs can be categorized into three main types based on their architectures: encoder-only, decoder-only, and encoder-decoder models, each suited for various protein research applications. For instance, encoder-only models are primarily used for predicting protein functions or properties, while decoder-only models are mainly employed for protein generation tasks.

Encoder-only models: Encoder-only models are a specialized form of the transformer architecture, dedicated solely to understanding and encoding input sequences.The essence of an encoder-only model revolves around extracting significant context from input sequences. These models encode protein sequences into fixed-length vectors for tasks like pattern recognition and prediction. Techniques like the Pairwise Masked Language Model (PMLM) [114] and mixed-chunk attention aim to capture co-evolutionary information and reduce complexity. Non-parametric models like ProteinNPT [115] handle sparse labels and multitask learning. Some models, like ESM-GearNet [116] and LM-GVP [117], integrate 3D structure information for better performance.

Decoder-only models: Utilizing the GPT [89] architecture, these models, such as ProGen [118] and ProGen2 [119], are essential for controllable protein generation. They explore unseen regions of the protein space while designing proteins with nature-like properties. Similar capabilities are exemplified by models like RITA [120], PoET [121], and LM-Design [122].

Encoder-decoder models: Used for sequence-to-sequence tasks, these models, including ProstT5 [123] and pAbT5 [124] are adept at tasks where an input sequence is transformed into an output sequence. A common example of a sequence-to-sequence task is machine translation, where a model translates a sentence from one language to another. In the context of Prot-LLMs, sequence-to-sequence tasks could involve tasks such as translating between protein sequences and structures. They can incorporate Multiple Sequence Alignment (MSA) modules to improve sequence generation and utilize reinforcement learning for structure-based design, as seen in Fold2Seq [125].

Capabilities Evaluation

Prot-LLMs are evaluated in three key areas: protein structure prediction, protein function prediction, and protein sequence generation.

Protein Structure Prediction: Prot-LLMs can predict the 3D structure of proteins from their sequences, which aids in understanding protein function, drug design, and biomedical research. Based on the 3D structure of known proteins, prot-LLMs can predict the three-dimensional structure of proteins based on an input sequence, which includes determining the atomic coordinates and the spatial relationships between atoms. Encoder-based Prot-LLMs are used to extract sequence information from the training data and predict tertiary and quaternary structures.

Protein Function Prediction: Prot-LLMs can predict the biological function of proteins and their interactions with other biomolecules. These tasks can be grouped into several categories. Firstly, protein classification involves categorizing proteins based on their structure, function, or sequence similarity. Prediction of protein-protein interactions focuses on identifying and forecasting interactions crucial for various biological processes. Localization and homology detection tasks include predicting a protein's subcellular location and identifying distant relationships between protein sequences. Spectral characteristics and stability prediction involve forecasting fluorescence properties and stability under specific conditions, respectively. Furthermore, specific tasks such as β-Lactamase activity prediction, solubility prediction, and mutation effect prediction focus on understanding specific protein functions, compound solubility, and the effects of genetic mutations on protein function, respectively. These tasks collectively contribute to explaining the complex functions and behaviors of proteins in biological systems. Biological systems are inherently complex and multifaceted, often requiring the simultaneous optimization of multiple properties. Unlike single-objective optimization, which focuses on one specific goal, multi-objective optimization allows researchers to consider and balance several objectives at once. This is particularly important in protein function prediction, where factors such as stability, activity, solubility, and interaction with other molecules need to be optimized concurrently. By providing a more comprehensive optimization framework and utilizing techniques such as Pareto optimization, researchers can identify solutions that offer the best trade-offs among different objectives, rather than a single optimal solution for one objective. multi-objective optimization can enhance the practical applicability of Prot-LLMs, leading to more effective and efficient solutions in understanding and manipulating protein functions.

Protein Sequence Generation: Prot-LLMs can propose amino acid sequences not found in nature and with a predicted function, useful in drug design and enzyme engineering. It includes:

- De novo protein design: Proposing protein sequences with a desired property that are not based on existing proteins with some or all of the desired property. Autoregressive generative models, such as the ProGen series, are commonly utilized for tasks involving the generation of protein sequences.
- Protein sequence optimization: proposing modification to an existing protein sequence to alter (i.e., optimize) its function or characteristic in an intended manner.

Benchmarks for Prot-LLMs

Table 6. Summary of Benchmarks for Prot-LLMs

Benchmark	Description
CASP	CASP (Critical Assessment of Structure Prediction) evaluates different methods and algorithms for protein structure prediction, providing a standard assessment for progress in the field.
EC	EC (Enzyme Commission) dataset is used to classify enzymes based on the chemical reactions they catalyze. This system is used to evaluate the functional prediction of proteins, specifically enzymes.
GO	GO (Gene Ontology) provides a framework for the representation of gene and gene product attributes across species. GO terms are used to annotate proteins with their associated biological processes, cellular components, and molecular functions.
CATH	CATH (Class, Architecture, Topology, Homologous superfamily) is a protein structure classification database that organizes protein domains into a hierarchical structure based on their folding patterns. It is used to classify protein domains into these categories: Class, Architecture, Topology, Homologous superfamily.
SCOP	SCOP (Structural Classification of Proteins) classifies proteins based on their structural and evolutionary relationships. SCOP benchmarks evaluate the ability of computational methods to classify protein structures into appropriate categories: Class, Fold, Superfamily, and Family.
ProteinGym	ProteinGym is a benchmark suite designed for evaluating the generalization capabilities of machine learning models in protein sequence prediction tasks. It includes various datasets and metrics to assess the performance of models in predicting protein sequences and related properties under different conditions.
TAPE	TAPE (Task Assessing Protein Embeddings) is a benchmark suite designed to evaluate the performance of protein sequence embeddings learned by machine learning models. It includes a variety of tasks, such as secondary structure prediction, contact prediction, and remote homology detection, to assess how well these embeddings capture the underlying biological properties of proteins.

Genomic Large Language Models (Gene-LLMs)

Gene-LLMs, specialized in genomic data, are trained to comprehend and predict genetic and genomic aspects of biology. They analyze DNA sequences, interpret genetic variations, and aid in genetic research, like identifying disease-related genetic markers or exploring evolutionary biology. Built on the Transformer architecture, genomic LLMs effectively model nucleic acid sequence data, capturing long-range dependencies for prediction and generation tasks. Through self-supervised learning on genomic sequences, Gene-LLMs gradually grasp genome understanding. Once fine-tuned or contextually learned, they prove valuable for downstream tasks, enhancing accuracy and reducing manual intervention.

Encoder-only models: With an encoder-only architecture for genomics, numerous significant models utilize the Transformer encoder to process gene sequences and extract meaningful patterns. Models like SpliceBERT, DNABERT, DNABERT-2, iEnhancer-BERT [126,127,128,129], and others employ mask training mechanisms to predict and complete masked gene sequences, achieving improved performance in tasks such as promoter prediction and transcription factor binding site prediction.

For instance, MoDNA [130] adopts a BERT-like encoder with a unique stacked Generator-Discriminator training paradigm, facilitating motif-oriented learning. GENA-LM [131] introduces encoder-based foundational DNA language models capable of handling sequences up to 36,000 base pairs. The Nucleotide-Transformer model [132], pre-trained on diverse human and species genomes, enhances the prediction of molecular phenotypes from DNA sequences. EpiGePT [133] predicts genome-wide epigenomics signals, offering insights into gene regulation. Uni-RNA [134] predicts RNA structures and functions, useful in RNA research and drug development. Models like Enformer [135] and LOGO [136] address the quadratic time complexity of attention mechanisms in handling long sequences, while BioSeq-BLM [137] integrates traditional analysis methods with language models, marking advancements in pre-training and fine-tuning.

Decoder-only models: Decoder-only models, like GenSLMs [138] and DNAGPT [139], demonstrate generative capabilities, capturing the evolutionary dynamics of viruses and enabling species identification and regulatory factor prediction. HyenaDNA [140] stands out for its exceptional ability to efficiently handle ultra-long DNA sequences while preserving single-nucleotide resolution. This unique combination of features enables researchers to analyze and manipulate genetic data at an unprecedented level of detail. Its capability to handle long sequences while maintaining single-nucleotide resolution greatly enhances its utility in various genomic applications, representing a significant advancement in computational genomics.

Encoder-decoder models: Encoder-decoder models in genomics, such as ENBED [141], combine the strengths of both components to compress and encode genomic data into meaningful representations. These representations are then used by the decoder to generate sequences or make predictions, enhancing bioinformatics research capabilities.

Capabilities Evaluation

Gene-LLMs undergo evaluation across four key domains: function prediction, structure prediction, sequence generation, and sequence variation and evolution analysis.

Protein Function Prediction: Traditionally, gene function prediction relied on models trained on specific sequences. With the advent of LLMs, pre-training on extensive genomic data followed by task-specific fine-tuning has enhanced accuracy and contextual understanding. Key subtasks include promoter prediction, enhancer prediction, and binding site prediction, tackled by models like DNABERT [127] and EpiGePT [133]

Structure Prediction: Leverages computational tools to identify and model biologically significant nucleic acid structures, aiding in the design of novel molecular architectures for nanotechnology and synthetic biology. Recent advancements include predicting RNA three-dimensional structures directly from sequences and designing sequences for predefined DNA and RNA nanostructures, demonstrating that nucleic acid structure can be both predictable and controllable. Subtasks include chromatin profile prediction and DNA/RNA-protein interaction prediction, addressed by models like HyenaDNA [140] and TFBert [142].

Sequence Generation: Proposing artificial sequences resembling real biological ones is crucial for bioinformatics, particularly for creating artificial human genomes serving as tools to safeguard genetic privacy and reduce costs linked with genetic sample collection [143,144]. The generated data strives to

retain the utility of the source data by replicating most of its characteristics. Consequently, they could serve as viable alternatives for many genomic databases that are either not publicly available or face accessibility barriers. DNAGPT [139] excels in this task, generating artificial genomes covering regions of single nucleotide polymorphisms (SNPs).

Sequence Variation and Evolution Analysis: Understanding biological sequence variation and evolution is vital for uncovering the genetic basis of traits, disease, and evolutionary patterns. Models like GenSLMs [138] and GPN-MSA [145] analyze the evolutionary landscape of genomes, focusing on species-specific and whole-genome sequence alignments.

Benchmarks for Gene-LLMs

Table 7. Summary of Benchmarks for Gene-LLMs

Benchmark	Description
CAGI5 Challenge Benchmark	The Critical Assessment of Genome Interpretation (CAGI) is a benchmark designed to rigorously assess computational methods in predicting a wide array of genetic and genomic outcomes.
Protein-RNA Interaction Prediction Benchmark (Protein-RNA)	A set of 37 machine learning (primarily deep learning) methods for in vivo RNA-binding proteins RBP–RNA interaction prediction. This benchmark systematically evaluates a subset of 11 representative methods across hundreds of CLIP-seq datasets and RBPs.
Nucleotide Transformer Benchmark (NT-Bench)	A comprehensive evaluation framework designed to assess the performance of genomics foundational models. This benchmark pits the Nucleotide Transformer models against other prominent genomics models, such as DNABERT, HyenaDNA (with both 1kb and 32kb context lengths), and Enformer.

Multimodal Scientific Large Language Models (MM-Sci-LLMs)

Multimodal scientific large language models (MM-Sci-LLMs) possess the ability to process and combine various types of scientific data, including text, molecules, and proteins, making them indispensable for interdisciplinary research requiring insights from multiple domains. An emerging research area, MM-Sci-LLMs utilize LLMs as their core to handle diverse data types effectively. These models exhibit remarkable adaptability in incorporating text, images, audio, and other forms of information, enabling comprehensive problem-solving across scientific domains, particularly in biological sciences encompassing protein, molecular, and genomic studies.

Categorized into four distinct groups based on the specific modality they focus on, MM-Sci-LLMs demonstrate specialized capabilities.

Table 8. Summary of MM-Sci-LLMs

Category	Description	Encoder-only models	Encoder-Decoder models	Decoder-only models

Category	Description	Encoder-only models	Encoder-Decoder models	Decoder-only models
Molecule-to-text	Leverage various techniques like multimodal embedding and cross-modal learning to associate chemical structures with textual descriptions, enhancing tasks such as cross-modal retrieval and molecular property prediction.	Text2Mol, KV-PLM, MoMu	DrugChat, MolReGPT, Text+Chem, ChatMol, GIT-Mol	MolET5, MolFM, GPT-Mol
Protein-to-text models	Utilize textual data for protein function prediction and multimodal representation learning, enriching protein annotation and design by integrating natural language descriptions with protein data.	ProTranslator, ProtST-ProtBert	InstructionProtein	ProteinDT, Prot2Text, ProtST-ESM-1B, ProtST-ESM-2
Protein-to-molecule models	Focus on linking protein sequences with molecular information, improving drug discovery through techniques like adversarial networks and contrastive learning.	DrugCLIP	DrugGPT	ChemBERTaLM, DeepTarget
Comprehensive models	Integrate multiple scientific modalities to excel in diverse tasks like biological data analysis, and material prediction, leveraging advanced multimodal learning techniques to support fundamental science research.	BioTranslator	Galactica, ChatDrug, DARWIN-MDP, BioMedGPT-10B, Mol-Instructions	BioT5

Capabilities Evaluation

MM-Sci-LLMs undergo evaluation focusing on three pivotal areas: cross-modal prediction, retrieval, and generation.

Cross-Modal Prediction: This involves using multimodal models to predict the functionality of biological entities like molecules, proteins, and genomes based on textual instructions. Models like MoleculeSTM [146] and Mol-Instructions [147] integrate molecular structures and text data for function prediction, which is crucial for bioinformatics and drug discovery.

Cross-Modal Retrieval: Involves retrieving information from one modality based on a query from another modality. Key models like KV-PLM [148] and ProtST-ESM-1b [149] enable retrieving molecules, proteins, or genes based on textual descriptions, aiding drug discovery and biological mechanism understanding.

Cross-Modal Generation: Aims to create data in one modality based on data from another. Models like Text2Mol [150] and ProteinDT [151] generate molecular information from text descriptions, while models like Prot2Text [152] and ChemBERTaLM [153] convert protein sequences into detailed text descriptions. This capability facilitates cohesive multi-modal data creation, bridging the gap between different modalities in scientific research.

Benchmarks for MM-Sci-LLMs

Table 9. Summary of Benchmarks for MM-Sci-LLMs

Benchmark	Description
MoleculeNet	MoleculeNet is a large-scale benchmark for molecular machine learning. It curates multiple public datasets, establishes metrics for evaluation, and offers high-quality open-source implementations of multiple previously proposed molecular featurization and learning algorithms.
MARCEL	MARCEL (MoleculAR Conformer Ensemble Learning) provides a comprehensive platform for evaluating learning from molecular conformer ensembles. It focuses on diverse molecular conformer structures, marking a significant shift in molecular representation learning.
GuacaMol	GuacaMol is an evaluation framework designed for de novo molecular design. It aims to generate molecules with specific property profiles through virtual design-make-test cycles.

Our technical exploration is primarily confined to Transformer-based languages, excluding alternative neural architectures like graph neural networks and diffusion models, despite their widespread applications in protein folding. However, the concepts discussed in biological languages can be extended to other scientific languages, such as molecular and mathematical languages.

Molecular large language models (Mol-LLMs) are specialized LLMs trained on molecular data, enabling them to understand and predict the chemical properties and behaviors of molecules. This specialized knowledge makes them invaluable tools in drug discovery, materials science, and the study of complex chemical interactions.

Encoder-only Mol-LLMs, like SMILES-BERT [154], focus on understanding and interpreting input molecules, making them ideal for tasks requiring a deep comprehension of molecular structures and properties. SMILES-BERT, for instance, leverages the BERT architecture to interpret SMILES representations of molecules.

Decoder-only Mol-LLMs, such as MolGPT [155] and SMILESGPT [156], use SMILES strings as input to navigate the vast chemical space. These models are crucial in drug discovery and materials science, enabling the synthesis of molecules with specific properties. MolGPT, which utilizes GPT for molecular generation with conditional training for property optimization, excels in molecular modeling and drug discovery by demonstrating strong control over multiple properties for accurate generation.

In encoder-decoder Mol-LLMs, encoders convert raw molecules into latent vectors, which decoders then reconstruct into functional chemical structures. Most Transformer-based encoder-decoder models use SMILES or SELFIES as inputs for the encoder, with outputs varying by task. For example, in chemical reaction prediction, the decoder generates the anticipated outcomes for reactants. The Molecular Transformer [157], a Transformer-based model for reaction prediction, effectively handles complex, long-range sequence interactions.

Biological data with graph structures can be modeled in two primary ways: molecular structure-based modeling and biological network-based modeling. In molecular structure-based modeling, atoms or valid chemical substructures are used as nodes, and bonds serve as edges to construct the molecular graph. Molecular graphs are extensively used for predicting molecular properties and designing new molecules.

In biological network-based modeling, nodes represent various entities such as genes, diseases, or RNAs, with edges indicating known associations between pairs of entities, such as miRNA–disease interactions. This creates a relational network. Graph Neural Networks (GNNs) excel at extracting information from graph structures, making them suitable for processing omics data in fields such as genomics, proteomics, RNomics, and radiomics. By applying GNNs to these omics data using the aforementioned modeling methods, a variety of tasks can be performed, including molecular property prediction, de novo molecular design, link prediction, and node classification in biological networks.

Bio-AI Tools (BDTs)

Bio-AI tools, commonly referred to as biological design tools (BDTs) are computational tools that help design proteins, viral vectors, or other biological agents. Traditional methods molecular biology like site-directed mutagenesis (SDM) involve the deliberate alteration of specific nucleotide sequences in DNA to create desired changes in the resulting protein. This process typically requires designing and synthesizing specific DNA primers, followed by PCR amplification and cloning steps to introduce the mutated DNA into a host organism. While SDM allows for precise modifications at predetermined sites, it can be time-consuming and labor-intensive, especially when multiple iterations are required to achieve the desired outcome. Additionally, the success rate of SDM experiments can vary depending on factors such as the efficiency of DNA synthesis and the stability of the resulting mutant proteins.

Random mutagenesis, another traditional method, involves introducing random mutations throughout the genome of an organism using techniques such as chemical mutagenesis or UV irradiation. This approach generates a pool of mutants with diverse genetic variations, which are then screened to identify individuals with desired phenotypic traits. While random mutagenesis can uncover novel genetic variants and phenotypes, it lacks the precision and control offered by targeted mutagenesis techniques like SDM. A related concept that enhances the utility of random mutagenesis is directed evolution. Directed evolution is an iterative process where organisms undergo random mutations, are tested against a screening process, and the best performers are selected for subsequent rounds of mutation. This cycle of mutating, screening, and selecting can be analogized to the training process of deep learning models. In deep learning, a model makes predictions based on input data, receives feedback on the accuracy of these predictions, and then adjusts its parameters through a process known as backpropagation.

In directed evolution, the organism's genetic material is repeatedly altered and tested, much like a model's parameters are iteratively refined to improve performance. Each cycle of directed evolution involves creating genetic diversity through random mutations, screening the resultant mutants for desirable traits, and then selecting the top performers for the next round of mutations. This method has been instrumental in fields such as enzyme engineering, where it has led to the development of

proteins with enhanced or novel functions. However, It is resource-intensive, requiring significant time and high-throughput screening capabilities.

In contrast, BDTs can accelerate experimentation by suggesting optimized properties of biological agents upfront, thereby potentially reducing the number of tests required to achieve desired outcomes. While the speed of individual experiments may not change, the efficiency of the overall experimentation process is enhanced, as researchers may need to conduct fewer experiments to reach the same or improved results [158]. Examples of BDTs include RFDiffusion [159], Protein MPNN [160], and protein language models like ProGen2 [119] and Ankh [161]. These models can be considered both Prot-LLMs and specific instances within the broader category of BDTs due to their training and output characteristics.

A crucial difference between LLMs and BDTs is both the training data — as LLMs are trained on natural language while BDTs are trained on biological data — and the output — LLMs typically produce outputs in natural language while BDTs produce outputs in the form of biological sequences, structures, and predictions. Although BDTs currently focus on creating sequences by optimizing for a single function, they may eventually evolve to design complex proteins and enzymes with multiple functions and properties. BDTs may eventually develop the capability to engineer whole organisms optimized for various functions and characteristics, addressing a comprehensive range of biological properties [162].

Of all the categories of AI-enabled BDTs, protein structural prediction tools have the highest relative maturity. Protein structure prediction tools, commonly referred to as 'folding tools,' contribute to the field by predicting a protein's 3D structure, including its secondary, tertiary and quaternary structures from its amino acid sequences. This prediction aids in understanding protein function and interactions. Determining the precise structure of proteins, vital for their functions, has historically posed significant challenges in experimental biology [163], often requiring years of dedicated effort. However, the landscape has shifted with the advent of AI, tailored to predict protein structures directly from their amino acid sequences.

Notably, pioneering AI systems like AlphaFold [164] and RoseTTAFold [165] have emerged, revolutionizing the field by drastically reducing structure determination times from months to mere hours. While AlphaFold provides measured structures based on experimental data and computational predictions, RoseTTAFold predicts structures solely through computational methods, sometimes eliminating the need for experimental measurements. AlphaFold 2, released in 2021, marked a significant breakthrough for deep learning in biology by unveiling a vast array of previously unknown protein structures. It quickly became a valuable tool for researchers working to understand everything from cellular structures [166] to tuberculosis [167]. It also inspired the development of other biological deep learning tools. Most notably, the biochemist David Baker and his team at the University of Washington developed a competing algorithm in 2021 called RoseTTAFold, which, like AlphaFold2, predicts protein structures from sequence data. Both systems have since been enhanced with new features. RoseTTAFold Diffusion is designed to create new proteins that do not exist in nature, while AlphaFold Multimer focuses on the interaction of multiple proteins. These advancements have propelled the development of numerous complementary tools that contextualize biochemical data, screen for protein interactions, and aid in experimental structure elucidation. Furthermore, the predictions from these tools have been integrated into publicly accessible databases, fostering widespread access and collaboration.

Proteins, intricate molecular machines honed by evolution, are built from a repertoire of 20 canonical amino acids, intricately arranged to yield diverse structures crucial for biological functions. Understanding a protein's 3D structure is paramount, as it dictates its functional properties; for instance, an enzyme's precise folding enables effective catalysis. Thus, deciphering protein structures not only determines their biological roles but also sheds light on disease-related mutations and their

impacts. A longstanding aspiration in structural biology has been the computational prediction of protein structures, circumventing the laborious and expensive experimental methods. Milestones such as the Critical Assessment of Structure Prediction (CASP) [168] have gauged progress in this domain. AlphaFold's breakthrough at the 13th CASP competition, and subsequent advancements like AlphaFold2 and RoseTTAFold at 14th CASP competition, harnessed the pattern recognition prowess of machine-learning algorithms, trained on vast structural data repositories like the Protein Data Bank (PDB) [169]. These algorithms, unencumbered by prior exposure to certain proteins, demonstrated remarkable accuracy in structure prediction.

Following the 14th CASP competition, a proliferation of AI-enabled structure predictors has emerged. These predictors employ diverse strategies but share a common goal of understanding spatial proximity among amino acids by tracing evolutionary relationships. Multiple sequence alignment structure predictors (MSA-SPs), exemplified by AlphaFold 2 and RoseTTAFold, analyze co-evolutionary signals gleaned from input sequences to predict structures. In contrast, protein language model structure predictors (pLM-SPs), exemplified by ESMFold [170] and OmegaFold [171], embed evolutionary insights directly into their algorithms, eliminating the need for explicit MSA generation.

AlphaFold 3 [172], a successor to previous AlphaFold models, was released in 2024 by Google DeepMind. This new version extends its capabilities by predicting the structures of nearly all biological molecules and modeling their interactions. While researchers have previously developed specialized computational methods for modeling interactions between specific types of biological molecules, AlphaFold 3 is the first system capable of predicting interactions between almost all molecular types with state-of-the-art performance. The properties and functions of molecules in biological systems typically depend on their interactions with other molecules. Experimental methods to understand these interactions can take years and be prohibitively expensive. However, if these interactions can be accurately estimated computationally, biological research can be significantly accelerated. For instance, researchers looking for a promising drug candidate that binds a specific protein site can use computational systems like AlphaFold 3 to test potential drug molecules efficiently.

Other subcategories of BDTs include:

Table 10. Other subcategories of BDTs

Category	Description	Examples
Protein sequence design tools	Also known as 'inverse folding tools,' predict the sequence of a protein with a user-specified structure and/or functional property, such as binding to a target. These tools play a crucial role in designing proteins tailored to specific requirements.	Rosetta, RoseTTAFold, RF Diffusion
Small molecule design tools	Designed to predict molecular structures with specific profiles, such as generating drugs that provoke desired biological responses while maintaining acceptable pharmacokinetic properties. These tools are essential in drug discovery and development processes.	REINVENT 4, Chemistry42

Category	Description	Examples
Vaccine design tools	Pivotal in predicting protective antigens or vaccine subunits from the protein or proteome of target pathogens. By identifying vaccine candidates, these tools contribute significantly to the development of effective vaccines against infectious diseases.	LinearDesign, VSeq-Toolkit
Viral vector design tools	Focus on predicting the amino acid sequences of virus capsids with the aim of optimizing them as delivery vectors. These vectors are crucial in gene therapy and vaccine development, enabling the efficient delivery of therapeutic genes or vaccine antigens into target cells.	VSeq-Toolkit
Genetic modification tools	Analyze genetic sequences to identify sequence features or optimize them for specific purposes. These tools aid in genetic engineering applications by facilitating the modification of DNA sequences to achieve desired outcomes.	OpenCRISPR-1, ZFDesign
Genome assembly tools	Play a vital role in assembling genomes from multiple short reads generated by DNA sequencing technologies. These tools contribute to genome sequencing projects by reconstructing complete genome sequences from fragmented data.	DeepConsensus
Toxicity prediction/detection tools	Designed to predict or detect the molecular toxicity of given molecules or metabolites. These tools are valuable in drug safety assessment and environmental toxicology, aiding in the identification of potentially harmful substances.	TOXSCAPE, GENESCAPE
Pathogen property prediction tools	Predict or detect features of pathogens, such as propensity for zoonotic spillover or virulence. These tools are crucial in infectious disease surveillance and control, providing insights into the behavior and potential risks associated with pathogens.	MP4
Host-pathogen interaction prediction tools	Focus on predicting protein-protein interactions between hosts and pathogenic agents. By elucidating the mechanisms of host-pathogen interactions, these tools contribute to understanding disease pathogenesis and identifying potential therapeutic targets.	HPIPred, deepHPI

Category	Description	Examples
Immunological system modeling tools	Replicate components of the human immune system to predict immune responses, such as T-cell receptor epitope recognition. These tools aid in vaccine design and immunotherapy development by simulating immune responses to pathogens or therapeutic agents.	SIMMUNE
Experimental design/planning tools	Generate designs for experiments based on predefined objectives, optimizing experimental variables and methods to achieve desired outcomes. These tools streamline the experimental process, improving efficiency and data quality.	The Experimental Design Assistant (EDA)
Experimental simulation tools	Simulate and predict experimental outcomes in silico, aiding in the design and interpretation of experiments. By providing insights into potential experimental outcomes, these tools inform experimental planning and hypothesis testing.	PhET, BioSimulators
Autonomous experimental platforms	Conduct experiments without human intervention, utilizing laboratory automation equipment to perform physical tests, modeling, or data mining. These platforms enhance experimental throughput and reproducibility, accelerating scientific research and discovery.	BO algorithm with expected Improvement based (EI-based) policy

Risks, Limitations and Future Directions

While recent advancements in AI have enabled rapid progress in the life sciences, it also has several limitations and presents potential risks.

Risks and Limitations

Inaccurate outputs from AI models

The effectiveness of AI tools relies heavily on the quality of their algorithms and the data they are trained on. When these algorithms contain errors or the datasets are biased or incomplete, the AI models can produce inaccurate outputs. If the models and logic underlying an AI algorithm are incorrect, the AI's predictions or recommendations will also be incorrect. This can occur due to coding errors, incorrect assumptions in the model design, or inadequate tuning of the model parameters. AI models learn from the data they are trained on. If the training data is biased (e.g., over-represents certain conditions or populations) or incomplete (e.g., missing critical variables or having insufficient diversity), the model's outputs will reflect these shortcomings. This means the AI could give incorrect advice or predictions, which in biological experiments can lead to wasted time and resources as researchers follow flawed directions. The inaccuracies can misguide researchers, causing them to

conduct experiments based on false premises. This not only wastes valuable resources like time, money, and materials but can also delay scientific progress.

Development of harmful biological agents

AI models have the potential to assist in the creation and distribution of harmful biological agents. They could, for example, enable an actor to design a biological agent with favorable properties [173] and modify the agent's delivery mechanism in a manner that optimizes infectious doses and ensures environmental survival [174].This possibility raises significant biosecurity concerns. Amateur users are unlikely to utilize BDTs, but experts with malicious intent could leverage their scientific training and specific AI models to design new pathogens, develop synthetic DNA strands that evade screening measures, or enhance the efficiency of bioweapon production [175]. As with any AI system, BDTs depend on the quality of their training data, which can sometimes be limited by incompleteness or unintentional biases. While BDTs have been used to digitally generate potentially risky genetic sequences, research has yet to show if the synthesized sequences could be used to create harmful biological agents. Establishing empirical baselines metrics is essential for conducting risk assessments and tracking changes in risk over time [**Titus2023?**]. In AI applications within the life sciences, these metrics and baselines are not yet defined. To assess this risk, we need to systematically evaluate current AI systems' abilities to generate new sequences versus enhancing existing ones.

Ethics in AI for Life Sciences

The integration of artificial intelligence (AI) in life sciences presents significant ethical challenges that must be addressed to ensure responsible and beneficial use. Key ethical considerations include data privacy, informed consent, and the potential for bias in AI algorithms. Ensuring data privacy is paramount, as AI systems often require access to vast amounts of sensitive biological and medical data. This necessitates robust data protection measures and compliance with legal standards to prevent misuse and unauthorized access [176]. Informed consent is another critical issue, as individuals must be fully aware of how their data will be used and the potential implications of AI-driven analyses [177]. Additionally, AI algorithms can inadvertently perpetuate or exacerbate existing biases if the training data is not representative of diverse populations, leading to inequitable outcomes in healthcare and research [176]. Addressing these ethical concerns requires a multi-faceted approach, including rigorous testing of AI systems, transparency in AI operations, and the establishment of ethical guidelines and governance frameworks to guide the development and deployment of AI in life sciences [178]. By prioritizing these ethical considerations, we can harness the transformative potential of AI while safeguarding human rights and promoting equitable access to its benefits.

Future Directions

Introduction of new benchmarks

Recent studies have highlighted the shortcomings of existing benchmarks in evaluating LLMs for clinical applications [179,180]. Traditional benchmarks, which focus mainly on accuracy in medical question-answering, fail to capture the full range of clinical skills necessary for LLMs [181]. Critics argue that using human-centric standardized medical exams to evaluate LLMs is insufficient, as passing these tests does not reflect the nuanced expertise required in real-world clinical settings [181].

There is a growing consensus on the need for more comprehensive benchmarks. These new benchmarks should assess capabilities such as sourcing information from authoritative medical

references, adapting to the evolving medical knowledge landscape, and clearly communicating uncertainties [181,182]. To further enhance their relevance, benchmarks should include scenarios that test an LLM's performance in real-world applications and its ability to adapt to feedback from clinicians while maintaining robustness. Given the sensitive nature of healthcare, these benchmarks should evaluate factors like fairness, ethics, and equity, which are crucial yet challenging to quantify [181]. By expanding benchmarks to encompass scientific domains, especially the biological domain, we can ensure that LLMs are rigorously evaluated across a broad spectrum of applications, thereby promoting their responsible and effective use in advancing scientific and medical knowledge.

Red, blue and violet teaming

Due to increasing concerns about the safety, security, and trustworthiness of Generative AI models, both practitioners and regulators emphasize the importance of AI red-teaming [183]. Originally from cybersecurity, red-teaming involves adopting an adversary's perspective to find vulnerabilities. In AI, this means simulating attacks on AI applications to identify weaknesses and develop preventive measures [184]. For example, red teams can simulate backdoor attacks or data poisoning to test the AI model's defenses. Prompt injection, a common attack on generative AI models like LLMs, tricks the model into producing harmful content. Red teams can also prompt AI systems to extract sensitive information from training data.

Blue teaming, which focuses on defending against these simulated attacks, and purple teaming, which combines both red and blue teams for a comprehensive security assessment [185]. However, as AI systems continuously evolve, these strategies might be insufficient, especially in critical sectors like the life sciences [186].

Violet teaming goes further by pairing red and blue teams to build resilient systems that intend to simultaneously minimize harm and maximize benefit using the very technology that poses potential security risks [187]. In the life sciences, this might involve using AI models to screen for harmful sequences generated by the models themselves, preventing them from being produced and shared with the end user.

Additionally, Machine Learning Security Operations (MLSecOps) could play a crucial role in ensuring the safety of AI models in the life sciences by employing machine learning (ML) techniques to protect against cyber threats and secure AI/ML models [188]. MLSecOps focuses on encrypting sensitive genome data, detecting ransomware and Trojan attacks, and ensuring the integrity of ML algorithms used in critical applications. It also addresses vulnerabilities in software and IoT devices within biotechnology labs, enhances supply chain security, and mitigates biases in healthcare ML systems.

Conclusion

The integration of computing technologies into the life sciences has profoundly transformed the field, enabling unprecedented advancements in biological research and applications. From the early days of population genetics calculations in the 1950s to the sophisticated AI-driven models of today, the evolution of computational tools has paralleled and propelled the growth of life sciences.

Historical Milestones and Technological Advancements

The journey began with the use of primitive computers for biological modeling, such as Alan Turing's work on morphogenesis and the MANIAC computer's genetic code measurements. The 1960s and 1970s saw the rise of computational biology, driven by protein crystallography and the development

of bioinformatics software like COMPROTEIN. The advent of dynamic programming algorithms for sequence alignment and the shift from protein to DNA analysis marked significant milestones.

The Genomic Era and Beyond

The 1980s and 1990s were pivotal, with the development of gene targeting techniques, the polymerase chain reaction (PCR), and the emergence of bioinformatics software suites. The completion of the Haemophilus influenzae genome and the human genome project ushered in the genomic era, leading to the creation of specialized software for whole-genome sequencing.

Artificial Intelligence and Machine Learning

The recent decades have witnessed the integration of artificial intelligence (AI) and machine learning (ML) into life sciences, revolutionizing data analysis, drug discovery, and personalized medicine. AI models, from expert systems like DENDRAL and MYCIN to modern deep learning architectures, have enhanced our ability to predict protein structures, analyze genomic data, and design novel biological entities.

Emerging Technologies and Future Directions

Emerging technologies such as cloud computing, big data analytics, and the Internet of Things (IoT) are further enhancing the capabilities of life sciences research. Cloud-based high-performance computing enables complex data analysis and reduces research cycles, while IoT facilitates real-time data collection and patient monitoring.

Challenges and Ethical Considerations

Despite these advancements, challenges remain. The accuracy of AI models depends on the quality of training data, and there are significant ethical concerns regarding data privacy and the potential misuse of AI in creating harmful biological agents. Addressing these challenges requires robust ethical frameworks, continuous monitoring, and the development of explainable AI systems.

Altogether, the integration of computing in the life sciences has not only accelerated research but also opened new frontiers in understanding and manipulating biological systems. As we move forward, the synergy between computational technologies and life sciences will continue to drive innovation, offering new solutions to complex biological problems and improving human health. The future promises even greater advancements as we harness the power of AI, cloud computing, and other emerging technologies to explore the intricacies of life at unprecedented scales.

Glossary

- **Algorithm:** A step-by-step procedure or formula for solving a problem, often used in computer programming and computational biology.
- **AlphaFold:** An AI program developed by DeepMind that predicts protein structures with high accuracy.
- **Autoencoders:** A type of artificial neural network used to learn efficient codings of unlabeled data.

- **Bioinformatics:** The application of computer technology to the management and analysis of biological data.
- **Computational Biology:** A field that uses mathematical models, algorithms, and computational techniques to understand and analyze biological systems.
- **CRISPR:** A technology used for editing genomes, allowing researchers to alter DNA sequences and modify gene function.
- **Deep Learning:** A subset of machine learning involving neural networks with many layers.
- **DNA Sequencing:** The process of determining the nucleic acid sequence – the order of nucleotides in DNA.
- **Fourier Analysis:** A mathematical technique used to transform signals between time (or spatial) domain and frequency domain, applied in protein crystallography to determine structures.
- **Fixed-size vector representation:** A numerical representation of a fixed length that encapsulates the information or features extracted from a variable-length input. In machine learning and natural language processing (NLP), fixed-size vector representations are commonly used to represent textual or sequential data.
- **Genome:** The complete set of genes or genetic material present in a cell or organism.
- **Genomics:** The study of genomes, the complete set of DNA within an organism, including its structure, function, evolution, and mapping.
- **Generative Adversarial Networks (GANs):** A class of machine learning systems where two neural networks contest with each other in a game.
- **Machine Learning (ML):** A subset of artificial intelligence (AI) that involves the development of algorithms that allow computers to learn from and make predictions based on data.
- **Maximum Likelihood Methods:** Statistical methods for estimating the parameters of a model, used in phylogenetic inference to determine the most likely tree structure.
- **Metagenomics:** The study of genetic material recovered directly from environmental samples.
- **Multiple Sequence Alignment (MSA):** A method used to align three or more biological sequences to identify regions of similarity that may indicate functional, structural, or evolutionary relationships.
- **Natural Language Processing (NLP):** The ability of a computer program to understand human language as it is spoken.
- **Needleman-Wunsch Algorithm:** An algorithm used for pairwise sequence alignment that employs dynamic programming to find the optimal alignment between two sequences.
- **Neural Networks:** A series of algorithms that mimic the operations of a human brain to recognize relationships between vast amounts of data.
- **Next-Generation Sequencing (NGS):** High-throughput sequencing technologies that allow for rapid sequencing of DNA or RNA samples.

- **PCR (Polymerase Chain Reaction):** A method widely used in molecular biology to make several copies of a specific DNA segment.
- **Phylogenetics:** The study of the evolutionary history and relationships among individuals or groups of organisms.
- **Proteomics:** The large-scale study of proteins, particularly their structures and functions.
- **Reinforcement Learning:** A type of machine learning where an agent learns to behave in an environment by performing actions and seeing the results.
- **Systems Biology:** An approach in biomedical research to understanding the larger picture by putting its pieces together (holism instead of reductionism).
- **Transcriptomics:** The study of the complete set of RNA transcripts produced by the genome.
- **Unsupervised Learning:** A type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels.

References

1. **Computers in the study of evolution**
   JL Crosby
   Science Progress (1967) https://pubmed.ncbi.nlm.nih.gov/4859964

2. **The chemical basis of morphogenesis**
   Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences (1952-08-14) https://royalsocietypublishing.org/doi/10.1098/rstb.1952.0012
   DOI: 10.1098/rstb.1952.0012

3. **Scientific uses of the MANIAC**
   HL Anderson
   Journal of Statistical Physics (1986-06-01) https://doi.org/10.1007/BF02628301
   DOI: 10.1007/bf02628301

4. **A three-dimensional model of the myoglobin molecule obtained by x-ray analysis**
   JC Kendrew, G Bodo, HM Dintzis, RG Parrish, H Wyckoff, DC Phillips
   Nature (1958-03-08) https://pubmed.ncbi.nlm.nih.gov/13517261
   DOI: 10.1038/181662a0

5. **Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution**
   JC Kendrew, RE Dickerson, BE Strandberg, RG Hart, DR Davies, DC Phillips, VC Shore
   Nature (1960-02-13) https://pubmed.ncbi.nlm.nih.gov/18990802
   DOI: 10.1038/185422a0

6. **The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates**
   F Sanger, EOP Thompson
   Biochemical Journal (1953-02) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198157/

7. **The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates**
   F Sanger, EOP Thompson
   Biochemical Journal (1953-02) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198158/

8. **A method for the determination of amino acid sequence in peptides**
   P Edman
   Archives of Biochemistry (1949-07) https://pubmed.ncbi.nlm.nih.gov/18134557

9. **The origins of bioinformatics**
   JB Hagen
   Nature Reviews. Genetics (2000-12) https://pubmed.ncbi.nlm.nih.gov/11252753
   DOI: 10.1038/35042090

10. **Comprotein: a computer program to aid primary protein structure determination**
    Margaret Oakley Dayhoff, Robert S Ledley
    Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall) (1962) http://portal.acm.org/citation.cfm?doid=1461518.1461546
    DOI: 10.1145/1461518.1461546

11. https://febs.onlinelibrary.wiley.com/toc/14321033/5/2

12. **Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965**
Bruno J Strasser
Journal of the History of Biology (2010) https://pubmed.ncbi.nlm.nih.gov/20665074
DOI: 10.1007/s10739-009-9221-0

13. **A general method applicable to the search for similarities in the amino acid sequence of two proteins**
SB Needleman, CD Wunsch
Journal of Molecular Biology (1970-03) https://pubmed.ncbi.nlm.nih.gov/5420325
DOI: 10.1016/0022-2836(70)90057-4

14. **Progressive sequence alignment as a prerequisite to correct phylogenetic trees**
DF Feng, RF Doolittle
Journal of Molecular Evolution (1987) https://pubmed.ncbi.nlm.nih.gov/3118049
DOI: 10.1007/bf02603120

15. **CLUSTAL: a package for performing multiple sequence alignment on a microcomputer**
DG Higgins, PM Sharp
Gene (1988-12-15) https://pubmed.ncbi.nlm.nih.gov/3243435
DOI: 10.1016/0378-1119(88)90330-7

16. **Clustal Omega, accurate alignment of very large numbers of sequences**
Fabian Sievers, Desmond G Higgins
Methods in Molecular Biology (Clifton, N.J.) (2014) https://pubmed.ncbi.nlm.nih.gov/24170397
DOI: 10.1007/978-1-62703-646-7_6

17. **A new method for sequencing DNA**
AM Maxam, W Gilbert
Proceedings of the National Academy of Sciences of the United States of America (1977-02)
https://pubmed.ncbi.nlm.nih.gov/265521
DOI: 10.1073/pnas.74.2.560

18. **Summary statement of the Asilomar conference on recombinant DNA molecules.**
P Berg, D Baltimore, S Brenner, RO Roblin, MF Singer
Proceedings of the National Academy of Sciences of the United States of America (1975-06)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC432675/

19. **A strategy of DNA sequencing employing computer programs.**
R Staden
Nucleic Acids Research (1979-06-11) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC327874/

20. **Evolutionary trees from DNA sequences: a maximum likelihood approach**
J Felsenstein
Journal of Molecular Evolution (1981) https://pubmed.ncbi.nlm.nih.gov/7288891
DOI: 10.1007/bf01734359

21. **Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference**
B Rannala, Z Yang
Journal of Molecular Evolution (1996-09) https://pubmed.ncbi.nlm.nih.gov/8703097
DOI: 10.1007/bf02338839

22. **A biologist's guide to Bayesian phylogenetic analysis**
Fabrícia F Nascimento, Mario Dos Reis, Ziheng Yang
Nature Ecology & Evolution (2017-10) https://pubmed.ncbi.nlm.nih.gov/28983516

DOI: 10.1038/s41559-017-0280-x

23. **The Nobel Prize in Chemistry 1993**
NobelPrize.org
https://www.nobelprize.org/prizes/chemistry/1993/mullis/lecture/

24. **A comprehensive set of sequence analysis programs for the VAX**
J Devereux, P Haeberli, O Smithies
Nucleic Acids Research (1984-01-11) https://pubmed.ncbi.nlm.nih.gov/6546423
DOI: 10.1093/nar/12.1part1.387

25. **DNASTAR's Lasergene Sequence Analysis Software**
Timothy G Burland
Bioinformatics Methods and Protocols (1999) https://doi.org/10.1385/1-59259-192-2:71
ISBN: 9781592591923

26. **Apple II PASCAL programs for molecular biologists**
B Malthiery, B Bellon, D Giorgi, B Jacq
Nucleic Acids Research (1984-01-11) https://pubmed.ncbi.nlm.nih.gov/6320099
DOI: 10.1093/nar/12.1part2.569

27. **The GNU Manifesto - GNU Project - Free Software Foundation**
https://www.gnu.org/gnu/manifesto.en.html

28. https://www.researchgate.net/publication/221307757_The_Free_Software_Movement_and_the_GNULinux_Operating_System

29. **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd**
RD Fleischmann, MD Adams, O White, RA Clayton, EF Kirkness, AR Kerlavage, CJ Bult, JF Tomb, BA Dougherty, JM Merrick
Science (New York, N.Y.) (1995-07-28) https://pubmed.ncbi.nlm.nih.gov/7542800
DOI: 10.1126/science.7542800

30. **The sequence of the human genome**
JC Venter, MD Adams, EW Myers, PW Li, RJ Mural, GG Sutton, HO Smith, M Yandell, CA Evans, RA Holt, ... X Zhu
Science (New York, N.Y.) (2001-02-16) https://pubmed.ncbi.nlm.nih.gov/11181995
DOI: 10.1126/science.1058040

31. **Consed: a graphical tool for sequence finishing**
D Gordon, C Abajian, P Green
Genome Research (1998-03) https://pubmed.ncbi.nlm.nih.gov/9521923
DOI: 10.1101/gr.8.3.195

32. **A whole-genome assembly of Drosophila**
EW Myers, GG Sutton, AL Delcher, IM Dew, DP Fasulo, MJ Flanigan, SA Kravitz, CM Mobarry, KH Reinert, KA Remington, ... JC Venter
Science (New York, N.Y.) (2000-03-24) https://pubmed.ncbi.nlm.nih.gov/10731133
DOI: 10.1126/science.287.5461.2196

33. **The EMBL data library**
CM Rice, R Fuchs, DG Higgins, PJ Stoehr, GN Cameron
Nucleic Acids Research (1993-07-01) https://pubmed.ncbi.nlm.nih.gov/8332519
DOI: 10.1093/nar/21.13.2967

34. **GenBank**

Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, Eric W Sayers
Nucleic Acids Research (2013-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531190/
DOI: 10.1093/nar/gks1195

35. **BLAST: at the core of a powerful and diverse set of sequence analysis tools**
Scott McGinnis, Thomas L Madden
Nucleic Acids Research (2004-07-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC441573/
DOI: 10.1093/nar/gkh435

36. **Genomes OnLine Database (GOLD): a monitor of genome projects world-wide**
Axel Bernal, Uy Ear, Nikos Kyrpides
Nucleic Acids Research (2001-01-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC29859/

37. **PubMed**
PubMed
https://pubmed.ncbi.nlm.nih.gov/

38. https://academic.oup.com/nar/article/26/1/94/2379498

39. **Accurate structure prediction of biomolecular interactions with AlphaFold 3**
Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, … John M Jumper
Nature (2024-06) https://pubmed.ncbi.nlm.nih.gov/38718835
DOI: 10.1038/s41586-024-07487-w

40. **The BLEND system Programme for the study of some 'electronic journals'∗**
B Shackel
Ergonomics (1982-04) http://www.tandfonline.com/doi/abs/10.1080/00140138208924954
DOI: 10.1080/00140138208924954

41. **arXiv.org e-Print archive** https://arxiv.org/

42. **bioRxiv.org - the preprint server for Biology** https://www.biorxiv.org/

43. **Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets**
L Pauling, RB Corey
Proceedings of the National Academy of Sciences of the United States of America (1951-11)
https://pubmed.ncbi.nlm.nih.gov/16578412
DOI: 10.1073/pnas.37.11.729

44. **Sixty-five years of the long march in protein secondary structure prediction: the final stretch?**
Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, Yaoqi Zhou
Briefings in Bioinformatics (2018-05-01) https://pubmed.ncbi.nlm.nih.gov/28040746
DOI: 10.1093/bib/bbw129

45. **Computational methods for protein structure prediction and modeling**
New York, N.Y. : Springer
(2007) http://archive.org/details/computationalmet0000unse_u4q5
ISBN: 9780387333212

46. **Molecular dynamics simulations: advances and applications**
Adam Hospital, Josep Ramon Goñi, Modesto Orozco, Josep L Gelpí

Advances and Applications in Bioinformatics and Chemistry : AABC (2015-11-19)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4655909/
DOI: 10.2147/aabc.s70333

47. **To Milliseconds and Beyond: Challenges in the Simulation of Protein Folding**
Thomas J Lane, Diwakar Shukla, Kyle A Beauchamp, Vijay S Pande
Current opinion in structural biology (2013-02)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3673555/
DOI: 10.1016/j.sbi.2012.11.002

48. **Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality**
Nidhi Gupta, Vijay K Verma
Microbial Technology for the Welfare of Society (2019-09-13)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7122948/
DOI: 10.1007/978-981-13-8844-6_15

49. **BOINC: A Platform for Volunteer Computing**
David P Anderson
Journal of Grid Computing (2020-03-01) https://doi.org/10.1007/s10723-019-09497-9
DOI: 10.1007/s10723-019-09497-9

50. **Structural proteomics by NMR spectroscopy**
Joon Shin, Woonghee Lee, Weontae Lee
Expert Review of Proteomics (2008-08) https://pubmed.ncbi.nlm.nih.gov/18761469
DOI: 10.1586/14789450.5.4.589

51. **A whole-cell computational model predicts phenotype from genotype**
Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, Markus W Covert
Cell (2012-07-20) https://pubmed.ncbi.nlm.nih.gov/22817898
DOI: 10.1016/j.cell.2012.05.044

52. **Using deep learning to model the hierarchical structure and function of a cell**
Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, Trey Ideker
Nature Methods (2018-04) https://www.nature.com/articles/nmeth.4627
DOI: 10.1038/nmeth.4627

53. **Why Build Whole-Cell Models?**
Javier Carrera, Markus W Covert
Trends in Cell Biology (2015-12) https://pubmed.ncbi.nlm.nih.gov/26471224
DOI: 10.1016/j.tcb.2015.09.004

54. **The future of whole-cell modeling**
Derek N Macklin, Nicholas A Ruggero, Markus W Covert
Current Opinion in Biotechnology (2014-08) https://pubmed.ncbi.nlm.nih.gov/24556244
DOI: 10.1016/j.copbio.2014.01.012

55. **Computer-Aided Whole-Cell Design: Taking a Holistic Approach by Integrating Synthetic With Systems Biology**
Lucia Marucci, Matteo Barberis, Jonathan Karr, Oliver Ray, Paul R Race, Miguel de Souza Andrade, Claire Grierson, Stefan Andreas Hoffmann, Sophie Landon, Elibio Rech, … Christopher Woods

Frontiers in Bioengineering and Biotechnology (2020-08-07)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7426639/
DOI: 10.3389/fbioe.2020.00942

56. **Accelerated discovery via a whole-cell model**
Jayodita C Sanghvi, Sergi Regot, Silvia Carrasco, Jonathan R Karr, Miriam V Gutschow, Benjamin Bolival, Markus W Covert
Nature methods (2013-12) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3856890/
DOI: 10.1038/nmeth.2724

57. **A blueprint for human whole-cell modeling**
Balázs Szigeti, Yosef D Roth, John AP Sekar, Arthur P Goldberg, Saahith C Pochiraju, Jonathan R Karr
Current opinion in systems biology (2018-02)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5966287/
DOI: 10.1016/j.coisb.2017.10.005

58. https://www.researchgate.net/publication/30874496_Artificial_Intelligence_A_Modern_Approach

59. https://www.researchgate.net/publication/223020409_Chen_CM_Intelligent_Web-based_Learning_System_with_Personalized_Learning_Path_Guidance_Computers_Education_512_787-814

60. **Expert systems: An overview | IEEE Journals & Magazine | IEEE Xplore**
https://ieeexplore.ieee.org/document/1145205

61. **On Interface Requirements for Expert Systems**
R Wexelblat
The AI Magazine (1989) https://www.semanticscholar.org/paper/On-Interface-Requirements-for-Expert-Systems-Wexelblat/291bffa7fec4fafff62462d015dd86c466273d4c

62. https://stacks.stanford.edu/file/druid:pj337tr4694/pj337tr4694.pdf

63. **MYCIN: a knowledge-based consultation program for infectious disease diagnosis**
William van Melle
International Journal of Man-Machine Studies (1978-05-01)
https://www.sciencedirect.com/science/article/pii/S0020737378800492
DOI: 10.1016/s0020-7373(78)80049-2

64. https://www.researchgate.net/publication/235028224_The_Applicability_and_Limitations_of_Expert_System_Shells

65. **Artificial intelligence, machine learning and deep learning: Potential resources for the infection clinician**
Anastasia A Theodosiou, Robert C Read
The Journal of Infection (2023-10) https://pubmed.ncbi.nlm.nih.gov/37468046
DOI: 10.1016/j.jinf.2023.07.006

66. **A survey on semi-supervised learning**
Jesper E van Engelen, Holger H Hoos
Machine Learning (2020-02-01) https://doi.org/10.1007/s10994-019-05855-6
DOI: 10.1007/s10994-019-05855-6

67. **Unsupervised Learning and Pattern Recognition of Biological Data Structures with Density Functional Theory and Machine Learning**
Chien-Chang Chen, Hung-Hui Juan, Meng-Yuan Tsai, Henry Horng-Shing Lu

Scientific Reports (2018-01-11) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5765025/
DOI: 10.1038/s41598-017-18931-5

68. **Unsupervised K-Means Clustering Algorithm | IEEE Journals & Magazine | IEEE Xplore**
https://ieeexplore.ieee.org/document/9072123

69. **Efficient Training Management for Mobile Crowd-Machine Learning: A Deep Reinforcement Learning Approach | IEEE Journals & Magazine | IEEE Xplore**
https://ieeexplore.ieee.org/document/8716527

70. **Machine learning in construction: From shallow to deep learning**
Yayin Xu, Ying Zhou, Przemyslaw Sekula, Lieyun Ding
Developments in the Built Environment (2021-05-01)
https://www.sciencedirect.com/science/article/pii/S2666165921000041
DOI: 10.1016/j.dibe.2021.100045

71. **An Introduction to Convolutional Neural Networks**
Keiron O'Shea, Ryan Nash
arXiv (2015-12-02) http://arxiv.org/abs/1511.08458
DOI: 10.48550/arxiv.1511.08458

72. **Recurrent Neural Networks (RNNs): A gentle Introduction and Overview**
Robin M Schmidt
arXiv (2019-11-23) http://arxiv.org/abs/1912.05911
DOI: 10.48550/arxiv.1912.05911

73. **Distributed representations, simple recurrent networks, and grammatical structure**
Jeffrey L Elman
Machine Learning (1991-09-01) https://doi.org/10.1007/BF00114844
DOI: 10.1007/bf00114844

74. **Autoencoders**
Dor Bank, Noam Koenigstein, Raja Giryes
arXiv (2021-04-03) http://arxiv.org/abs/2003.05991
DOI: 10.48550/arxiv.2003.05991

75. **Generative Adversarial Networks**
Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio
arXiv (2014-06-10) http://arxiv.org/abs/1406.2661
DOI: 10.48550/arxiv.1406.2661

76. **Natural language processing: state of the art, current trends and challenges**
Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh
Multimedia Tools and Applications (2023) https://pubmed.ncbi.nlm.nih.gov/35855771
DOI: 10.1007/s11042-022-13428-4

77. **Generative AI: A systematic review using topic modelling techniques**
Priyanka Gupta, Bosheng Ding, Chong Guan, Ding Ding
Data and Information Management (2024-06-01)
https://www.sciencedirect.com/science/article/pii/S2543925124000020
DOI: 10.1016/j.dim.2024.100066

78. **Auto-Encoding Variational Bayes**
Diederik P Kingma, Max Welling
arXiv (2022-12-10) http://arxiv.org/abs/1312.6114

DOI: 10.48550/arxiv.1312.6114

79. https://www.researchgate.net/publication/377955158_Autoencoders_and_their_applications_in_machine_learning_a_survey

80. **Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction**
Alexander J Titus, Owen M Wilkins, Carly A Bobak, Brock C Christensen
bioRxiv (2018-11-07) https://www.biorxiv.org/content/10.1101/433763v5
DOI: 10.1101/433763

81. https://www.researchgate.net/publication/322870935_A_New_Dimension_of_Breast_Cancer_Epigenetics_-_Applications_of_Variational_Autoencoders_with_DNA_Methylation

82. **A Style-Based Generator Architecture for Generative Adversarial Networks**
Tero Karras, Samuli Laine, Timo Aila
arXiv (2019-03-29) http://arxiv.org/abs/1812.04948
DOI: 10.48550/arxiv.1812.04948

83. **Denoising Diffusion Probabilistic Models**
Jonathan Ho, Ajay Jain, Pieter Abbeel
arXiv (2020-12-16) http://arxiv.org/abs/2006.11239
DOI: 10.48550/arxiv.2006.11239

84. **High-Resolution Image Synthesis with Latent Diffusion Models**
Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer
arXiv (2022-04-13) http://arxiv.org/abs/2112.10752
DOI: 10.48550/arxiv.2112.10752

85. **Attention Is All You Need**
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin
arXiv (2023-08-01) http://arxiv.org/abs/1706.03762
DOI: 10.48550/arxiv.1706.03762

86. https://openai.com/index/chatgpt

87. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**
Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu
arXiv (2023-09-19) http://arxiv.org/abs/1910.10683
DOI: 10.48550/arxiv.1910.10683

88. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
arXiv (2019-05-24) http://arxiv.org/abs/1810.04805
DOI: 10.48550/arxiv.1810.04805

89. **Language Models are Few-Shot Learners**
Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, … Dario Amodei
arXiv (2020-07-22) http://arxiv.org/abs/2005.14165
DOI: 10.48550/arxiv.2005.14165

90. **A Survey of Large Language Models in Medicine: Progress, Application, and Challenge**

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, ... David A Clifton
arXiv (2024-05-15) http://arxiv.org/abs/2311.05112
DOI: 10.48550/arxiv.2311.05112

91. **Artificial intelligence in medicine: current trends and future possibilities**
Varun H Buch, Irfan Ahmed, Mahiben Maruthappu
The British Journal of General Practice (2018-03)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5819974/
DOI: 10.3399/bjgp18x695213

92. **GLUE Benchmark** https://gluebenchmark.com/

93. **SuperGLUE Benchmark**
SuperGLUE Benchmark
https://super.gluebenchmark.com/

94. **HellaSwag: Can a Machine Really Finish Your Sentence?**
Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, Yejin Choi
arXiv (2019-05-19) http://arxiv.org/abs/1905.07830
DOI: 10.48550/arxiv.1905.07830

95. **TruthfulQA: Measuring How Models Mimic Human Falsehoods**
Stephanie Lin, Jacob Hilton, Owain Evans
arXiv (2022-05-07) http://arxiv.org/abs/2109.07958
DOI: 10.48550/arxiv.2109.07958

96. **Measuring Massive Multitask Language Understanding**
Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt
arXiv (2021-01-12) http://arxiv.org/abs/2009.03300
DOI: 10.48550/arxiv.2009.03300

97. **Scientific Large Language Models: A Survey on Biological & Chemical Domains**
Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, ... Huajun Chen
arXiv (2024-01-26) http://arxiv.org/abs/2401.14656
DOI: 10.48550/arxiv.2401.14656

98. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**
Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang
Bioinformatics (2020-02-15) http://arxiv.org/abs/1901.08746
DOI: 10.1093/bioinformatics/btz682

99. **BioMegatron: Larger Biomedical Domain Language Model**
Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, Raghav Mani
arXiv (2020-10-13) http://arxiv.org/abs/2010.06060
DOI: 10.48550/arxiv.2010.06060

100. **Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing**
Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, Hoifung Poon

ACM Transactions on Computing for Healthcare (2022-01-31) http://arxiv.org/abs/2007.15779
DOI: 10.1145/3458754

101. **BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA**
Sultan Alrowili, Vijay Shanker
Proceedings of the 20th Workshop on Biomedical Language Processing (2021-06)
https://aclanthology.org/2021.bionlp-1.24
DOI: 10.18653/v1/2021.bionlp-1.24

102. **LinkBERT: Pretraining Language Models with Document Links**
Michihiro Yasunaga, Jure Leskovec, Percy Liang
arXiv (2022-03-29) http://arxiv.org/abs/2203.15827
DOI: 10.48550/arxiv.2203.15827

103. **Improving Language Understanding by Generative Pre-Training**
Alec Radford, Karthik Narasimhan
(2018) https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035

104. **Language Models are Unsupervised Multitask Learners**
Alec Radford, Jeff Wu, R Child, D Luan, Dario Amodei, I Sutskever
(2019) https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

105. **BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining**
Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, Tie-Yan Liu
Briefings in Bioinformatics (2022-11-19) http://arxiv.org/abs/2210.10341
DOI: 10.1093/bib/bbac409

106. **BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine**
Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, Zaiqing Nie
arXiv (2023-08-21) http://arxiv.org/abs/2308.09442
DOI: 10.48550/arxiv.2308.09442

107. **LLaMA: Open and Efficient Foundation Language Models**
Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, ... Guillaume Lample
arXiv (2023-02-27) http://arxiv.org/abs/2302.13971
DOI: 10.48550/arxiv.2302.13971

108. **Article Citations - References - Scientific Research Publishing**
https://www.scirp.org/reference/referencespapers?referenceid

109. **A Revision of Bloom's Taxonomy: An Overview**
David R Krathwohl
Theory Into Practice (2002-11-01)
https://www.tandfonline.com/doi/full/10.1207/s15430421tip4104_2
DOI: 10.1207/s15430421tip4104_2

110. **SciEval: A Multi-Level Large Language Model Evaluation Benchmark for Scientific Research**
Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, Kai Yu
arXiv (2023-08-24) http://arxiv.org/abs/2308.13149
DOI: 10.48550/arxiv.2308.13149

111. **C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models**
Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, ... Junxian He
arXiv (2023-11-06) http://arxiv.org/abs/2305.08322
DOI: 10.48550/arxiv.2305.08322

112. **PubMedQA: A Dataset for Biomedical Research Question Answering**
Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, Xinghua Lu
arXiv (2019-09-13) http://arxiv.org/abs/1909.06146
DOI: 10.48550/arxiv.1909.06146

113. **Crowdsourcing Multiple Choice Science Questions**
Johannes Welbl, Nelson F Liu, Matt Gardner
arXiv (2017-07-19) http://arxiv.org/abs/1707.06209
DOI: 10.48550/arxiv.1707.06209

114. **Pre-training Co-evolutionary Protein Representation via A Pairwise Masked Language Model**
Liang He, Shizhuo Zhang, Lijun Wu, Huanhuan Xia, Fusong Ju, He Zhang, Siyuan Liu, Yingce Xia, Jianwei Zhu, Pan Deng, ... Tie-Yan Liu
arXiv (2021-10-29) http://arxiv.org/abs/2110.15527
DOI: 10.48550/arxiv.2110.15527

115. **ProteinNPT: Improving Protein Property Prediction and Design with Non-Parametric Transformers**
Pascal Notin, Ruben Weitzman, Debora S Marks, Yarin Gal
bioRxiv (2023-12-07) https://www.biorxiv.org/content/10.1101/2023.12.06.570473v1
DOI: 10.1101/2023.12.06.570473

116. https://openreview.net/forum?id

117. **LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction**
Zichen Wang, Steven A Combs, Ryan Brand, Miguel Romero Calvo, Panpan Xu, George Price, Nataliya Golovach, Emmanuel O Salawu, Colby J Wise, Sri Priya Ponnapalli, Peter M Clark
Scientific Reports (2022-04-27) https://pubmed.ncbi.nlm.nih.gov/35477726
DOI: 10.1038/s41598-022-10775-y

118. **ProGen: Language Modeling for Protein Generation**
Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, Richard Socher
arXiv (2020-03-07) http://arxiv.org/abs/2004.03497
DOI: 10.48550/arxiv.2004.03497

119. **ProGen2: Exploring the Boundaries of Protein Language Models**
Erik Nijkamp, Jeffrey Ruffolo, Eli N Weinstein, Nikhil Naik, Ali Madani
arXiv (2022-06-27) http://arxiv.org/abs/2206.13517
DOI: 10.48550/arxiv.2206.13517

120. **RITA: a Study on Scaling Up Generative Protein Sequence Models**
Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, Debora Marks
arXiv (2022-07-14) http://arxiv.org/abs/2205.05789
DOI: 10.48550/arxiv.2205.05789

121. **PoET: A generative model of protein families as sequences-of-sequences**
Timothy F Truong Jr, Tristan Bepler

arXiv (2023-11-01) http://arxiv.org/abs/2306.06156
DOI: 10.48550/arxiv.2306.06156

122. **Structure-informed Language Models Are Protein Designers**
Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei YE, Quanquan Gu
arXiv (2023-02-09) http://arxiv.org/abs/2302.01649
DOI: 10.48550/arxiv.2302.01649

123. **Bilingual Language Model for Protein Sequence and Structure**
Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, Burkhard Rost
bioRxiv (2024-03-24) https://www.biorxiv.org/content/10.1101/2023.07.23.550085v2
DOI: 10.1101/2023.07.23.550085

124. **Generative Antibody Design for Complementary Chain Pairing Sequences through Encoder-Decoder Language Model**
Simon KS Chu, Kathy Y Wei
arXiv (2023-11-20) http://arxiv.org/abs/2301.02748
DOI: 10.48550/arxiv.2301.02748

125. **Fold2Seq: A Joint Sequence(1D)-Fold(3D) Embedding-based Generative Model for Protein Design**
Yue Cao, Payel Das, Vijil Chenthamarakshan, Pin-Yu Chen, Igor Melnyk, Yang Shen
arXiv (2021-06-24) http://arxiv.org/abs/2106.13058
DOI: 10.48550/arxiv.2106.13058

126. **Predicting Retrosynthetic Reaction using Self-Corrected Transformer Neural Networks**
Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, Yuedong Yang
arXiv (2019-07-02) http://arxiv.org/abs/1907.01356
DOI: 10.48550/arxiv.1907.01356

127. **DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome**
Yanrong Ji, Zhihan Zhou, Han Liu, Ramana V Davuluri
Bioinformatics (Oxford, England) (2021-08-09) https://pubmed.ncbi.nlm.nih.gov/33538820
DOI: 10.1093/bioinformatics/btab083

128. **DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome**
Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, Han Liu
arXiv (2024-03-18) http://arxiv.org/abs/2306.15006
DOI: 10.48550/arxiv.2306.15006

129. **iEnhancer-BERT: A Novel Transfer Learning Architecture Based on DNA-Language Model for Identifying Enhancers and Their Strength**
springerprofessional.de
https://www.springerprofessional.de/en/ienhancer-bert-a-novel-transfer-learning-architecture-based-on-d/23365796

130. https://www.researchgate.net/publication/362540943_MoDNA_motif-oriented_pre-training_for_DNA_language_model

131. **GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences**
Veniamin Fishman, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, Mikhail Burtsev
bioRxiv (2023-06-13) https://www.biorxiv.org/content/10.1101/2023.06.12.544594v1
DOI: 10.1101/2023.06.12.544594

132. **The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics**
Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, … Thomas Pierrot
bioRxiv (2023-01-15) https://www.biorxiv.org/content/10.1101/2023.01.11.523679v1
DOI: 10.1101/2023.01.11.523679

133. **EpiGePT: a Pretrained Transformer model for epigenomics**
Zijing Gao, Qiao Liu, Wanwen Zeng, Rui Jiang, Wing Hung Wong
bioRxiv (2024-02-03) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10370089/
DOI: 10.1101/2023.07.15.549134

134. **Uni-Rna: Universal Pre-Trained Models Revolutionize Rna Research**
Xi Wang, Ruichu Gu, Zhiyuan Chen, Yongge Li, Xiaohong Ji, Guolin Ke, Han Wen
bioRxiv (2023-07-12) https://www.biorxiv.org/content/10.1101/2023.07.11.548588v1
DOI: 10.1101/2023.07.11.548588

135. **Effective gene expression prediction from sequence by integrating long-range interactions**
Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, David R Kelley
Nature Methods (2021-10) https://pubmed.ncbi.nlm.nih.gov/34608324
DOI: 10.1038/s41592-021-01252-x

136. https://www.researchgate.net/publication/354105080_LOGO_a_contextualized_pre-trained_language_model_of_human_genome_flexibly_adapts_to_various_downstream_tasks_by_fine-tuning

137. **BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models**
Hong-Liang Li, Yi-He Pang, Bin Liu
Nucleic Acids Research (2021-12-16) https://pubmed.ncbi.nlm.nih.gov/34581805
DOI: 10.1093/nar/gkab829

138. **GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics**
Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, … Arvind Ramanathan
bioRxiv (2022-10-11) https://www.biorxiv.org/content/10.1101/2022.10.10.511571v1
DOI: 10.1101/2022.10.10.511571

139. **DNAGPT: A Generalized Pretrained Tool for Multiple DNA Sequence Analysis Tasks**
Daoan Zhang, Weitong Zhang, Bing He, Jianguo Zhang, Chenchen Qin, Jianhua Yao
bioRxiv (2023-07-12) https://www.biorxiv.org/content/10.1101/2023.07.11.548628v1
DOI: 10.1101/2023.07.11.548628

140. **HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution**
Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, … Chris Ré
arXiv (2023-11-14) http://arxiv.org/abs/2306.15794
DOI: 10.48550/arxiv.2306.15794

141. **Understanding the Natural Language of DNA using Encoder-Decoder Foundation Models with Byte-level Precision**
Aditya Malusare, Harish Kothandaraman, Dipesh Tamboli, Nadia A Lanman, Vaneet Aggarwal

arXiv (2024-02-13) http://arxiv.org/abs/2311.02333
DOI: 10.48550/arxiv.2311.02333

142. **Improving language model of human genome for DNA-protein binding prediction based on task-specific pre-training**
Hanyu Luo, Wenyu Shan, Cheng Chen, Pingjian Ding, Lingyun Luo
Interdisciplinary Sciences, Computational Life Sciences (2023-03)
https://pubmed.ncbi.nlm.nih.gov/36136096
DOI: 10.1007/s12539-022-00537-9

143. **AnomiGAN: Generative Adversarial Networks for Anonymizing Private Medical Data**
Ho Bae, Dahuin Jung, Hyun-Soo Choi, Sungroh Yoon
Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (2020)
https://pubmed.ncbi.nlm.nih.gov/31797628

144. **Creating artificial human genomes using generative neural networks**
Burak Yelmen, Aurélien Decelle, Linda Ongaro, Davide Marnetto, Corentin Tallec, Francesco Montinaro, Cyril Furtlehner, Luca Pagani, Flora Jay
PLoS genetics (2021-02) https://pubmed.ncbi.nlm.nih.gov/33539374
DOI: 10.1371/journal.pgen.1009303

145. **GPN-MSA: an alignment-based DNA language model for genome-wide variant effect prediction**
Gonzalo Benegas, Carlos Albors, Alan J Aw, Chengzhong Ye, Yun S Song
bioRxiv (2024-04-06) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10592768/
DOI: 10.1101/2023.10.10.561776

146. **Multi-modal Molecule Structure-text Model for Text-based Retrieval and Editing**
Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, Anima Anandkumar
arXiv (2024-01-29) http://arxiv.org/abs/2212.10789
DOI: 10.48550/arxiv.2212.10789

147. **Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models**
Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, Huajun Chen
arXiv (2024-03-04) http://arxiv.org/abs/2306.08018
DOI: 10.48550/arxiv.2306.08018

148. **A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals**
Zheni Zeng, Yuan Yao, Zhiyuan Liu, Maosong Sun
Nature Communications (2022-02-14) https://pubmed.ncbi.nlm.nih.gov/35165275
DOI: 10.1038/s41467-022-28494-3

149. **ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts**
Minghao Xu, Xinyu Yuan, Santiago Miret, Jian Tang
arXiv (2023-07-04) http://arxiv.org/abs/2301.12040
DOI: 10.48550/arxiv.2301.12040

150. **Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries**
Carl Edwards, ChengXiang Zhai, Heng Ji
Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (2021-11) https://aclanthology.org/2021.emnlp-main.47

DOI: 10.18653/v1/2021.emnlp-main.47

151. **A Text-guided Protein Design Framework**
Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, … Anima Anandkumar
arXiv (2023-12-03) http://arxiv.org/abs/2302.04611
DOI: 10.48550/arxiv.2302.04611

152. **Prot2Text: Multimodal Protein's Function Generation with GNNs and Transformers**
Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, Michalis Vazirgiannis
Proceedings of the AAAI Conference on Artificial Intelligence (2024-03-24)
http://arxiv.org/abs/2307.14367
DOI: 10.1609/aaai.v38i10.28948

153. **Exploiting pretrained biochemical language models for targeted drug design**
Gökçe Uludoğan, Elif Ozkirimli, Kutlu O Ulgen, Nilgün Karalı, Arzucan Özgür
Bioinformatics (Oxford, England) (2022-09-16) https://pubmed.ncbi.nlm.nih.gov/36124801
DOI: 10.1093/bioinformatics/btac482

154. **SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction**
Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, Junzhou Huang
Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (2019-09-04) https://dl.acm.org/doi/10.1145/3307339.3342186
DOI: 10.1145/3307339.3342186

155. **MolGPT: Molecular Generation Using a Transformer-Decoder Model**
Viraj Bagal, Rishal Aggarwal, PK Vinod, UDeva Priyakumar
Journal of Chemical Information and Modeling (2022-05-09)
https://pubs.acs.org/doi/10.1021/acs.jcim.1c00600
DOI: 10.1021/acs.jcim.1c00600

156. **Generative Pre-Training from Molecules**
Sanjar Adilov
ChemRxiv (2021-09-16) https://chemrxiv.org/engage/chemrxiv/article-details/6142f60742198e8c31782e9e
DOI: 10.26434/chemrxiv-2021-5fwjd

157. **Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction**
Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, Alpha A Lee
ACS Central Science (2019-09-25) https://pubs.acs.org/doi/10.1021/acscentsci.9b00576
DOI: 10.1021/acscentsci.9b00576

158. **AI-based screening method could boost speed of new drug discovery**
ScienceDaily
https://www.sciencedaily.com/releases/2022/09/220923090832.htm

159. **Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models**
Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, … David Baker
bioRxiv (2022-12-10) https://www.biorxiv.org/content/10.1101/2022.12.09.519842v1
DOI: 10.1101/2022.12.09.519842

160. **Robust deep learning-based protein sequence design using ProteinMPNN**

J Dauparas, I Anishchenko, N Bennett, H Bai, RJ Ragotte, LF Milles, BIM Wicky, A Courbet, RJ de Haas, N Bethel, … D Baker
Science (New York, N.Y.) (2022-10-07) https://pubmed.ncbi.nlm.nih.gov/36108050
DOI: 10.1126/science.add2187

161. **Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling**
Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, Burkhard Rost
arXiv (2023-01-16) http://arxiv.org/abs/2301.06568
DOI: 10.48550/arxiv.2301.06568

162. **DNAI - The Artificial Intelligence / Artificial Life convergence**
Jim Thomas
https://www.scanthehorizon.org/p/dnai-the-artificial-intelligence

163. **Challenges in protein folding simulations: Timescale, representation, and analysis**
Peter L Freddolino, Christopher B Harrison, Yanxin Liu, Klaus Schulten
Nature physics (2010-10-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3032381/
DOI: 10.1038/nphys1713

164. **Highly accurate protein structure prediction with AlphaFold**
John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, … Demis Hassabis
Nature (2021-08) https://www.nature.com/articles/s41586-021-03819-2
DOI: 10.1038/s41586-021-03819-2

165. **Accurate prediction of protein structures and interactions using a three-track neural network**
Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, RDustin Schaeffer, … David Baker
Science (New York, N.Y.) (2021-08-20) https://pubmed.ncbi.nlm.nih.gov/34282049
DOI: 10.1126/science.abj8754

166. **Building the nuclear pore complex**
Di Jiang
Science (2022-06-10) https://www.science.org/doi/10.1126/science.add2210
DOI: 10.1126/science.add2210

167. **Structure of an endogenous mycobacterial MCE lipid transporter**
James Chen, Alice Fruhauf, Catherine Fan, Jackeline Ponce, Beatrix Ueberheide, Gira Bhabha, Damian C Ekiert
Nature (2023-08) https://pubmed.ncbi.nlm.nih.gov/37495693
DOI: 10.1038/s41586-023-06366-0

168. **Critical assessment of methods of protein structure prediction (CASP)-Round XIII**
Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, John Moult
Proteins (2019-12) https://pubmed.ncbi.nlm.nih.gov/31589781
DOI: 10.1002/prot.25823

169. **The Protein Data Bank**
HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, PE Bourne
Nucleic Acids Research (2000-01-01) https://pubmed.ncbi.nlm.nih.gov/10592235
DOI: 10.1093/nar/28.1.235

170. **Evolutionary-scale prediction of atomic-level protein structure with a language model**

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, ... Alexander Rives
Science (New York, N.Y.) (2023-03-17) https://pubmed.ncbi.nlm.nih.gov/36927031
DOI: 10.1126/science.ade2574

171. **High-resolution de novo structure prediction from primary sequence**
Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, ... Jian Peng
bioRxiv (2022-07-22) https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1
DOI: 10.1101/2022.07.21.500999

172. **Accurate structure prediction of biomolecular interactions with AlphaFold 3**
Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, ... John M Jumper
Nature (2024-06) https://www.nature.com/articles/s41586-024-07487-w
DOI: 10.1038/s41586-024-07487-w

173. **AI-enhanced protein design makes proteins that have never existed**
Michael Eisenstein
Nature Biotechnology (2023-03-01) https://www.nature.com/articles/s41587-023-01705-y
DOI: 10.1038/s41587-023-01705-y

174. **Artificial Intelligence's Impact on Drug Discovery and Development From Bench to Bedside**
KS Vidhya, Ayesha Sultana, Naveen Kumar M, Harish Rangareddy
Cureus (2023-10) https://pubmed.ncbi.nlm.nih.gov/37881323
DOI: 10.7759/cureus.47486

175. **Artificial intelligence challenges in the face of biological threats: emerging catastrophic risks for public health**
Renan Chaves de Lima, Lucas Sinclair, Ricardo Megger, Magno Alessandro Guedes Maciel, Pedro Fernando da Costa Vasconcelos, Juarez Antônio Simões Quaresma
Frontiers in Artificial Intelligence (2024)
https://www.frontiersin.org/articles/10.3389/frai.2024.1382356

176. **Navigating the legal and ethical challenges of AI in healthcare - KPMG UK**
Caroline Rivett Simpson Isabel
KPMG (2024-03-01) https://kpmg.com/uk/en/home/insights/2024/03/navigating-the-legal-and-ethical-challenges-of-ai-in-healthcare.html

177. **Ethical Issues of Artificial Intelligence in Medicine and Healthcare**
Dariush D Farhud, Shaghayegh Zokaei
Iranian Journal of Public Health (2021-11)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8826344/
DOI: 10.18502/ijph.v50i11.7600

178. **Ethical AI in Life Sciences: Impact & Guidelines** https://www.aciinfotech.com/blogs/ethical-ai-in-life-sciences-impact-guidelines

179. **An extensive benchmark study on biomedical text generation and mining with ChatGPT**
Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, Zhangmin Niu
Bioinformatics (Oxford, England) (2023-09-02) https://pubmed.ncbi.nlm.nih.gov/37682111
DOI: 10.1093/bioinformatics/btad557

180. **Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations**
Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, Hua Xu
arXiv (2024-01-20) http://arxiv.org/abs/2305.16326
DOI: 10.48550/arxiv.2305.16326

181. **Large Language Models Encode Clinical Knowledge**
Karan Singhal, Shekoofeh Azizi, Tao Tu, SSara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, … Vivek Natarajan
arXiv (2022-12-26) http://arxiv.org/abs/2212.13138
DOI: 10.48550/arxiv.2212.13138

182. **Large language models in medicine**
Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, Daniel Shu Wei Ting
Nature Medicine (2023-08) https://pubmed.ncbi.nlm.nih.gov/37460753
DOI: 10.1038/s41591-023-02448-8

183. **Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence**
NIST
(2023-10-10) https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence

184. **What Does AI Red-Teaming Actually Mean?**
Tessa Baker
Center for Security and Emerging Technology (2023-10-24)
https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/

185. **Purple Teaming: A comprehensive and collaborative approach to cyber security**
Erik Van Buggenhout
Cyber Security: A Peer-Reviewed Journal (2024)
https://ideas.repec.org//a/aza/csj000/y2024v7i3p207-216.html

186. https://www.researchgate.net/publication/372592054_Violet_Teaming_AI_in_the_Life_Sciences_A_Preprint

187. **The Promise and Peril of Artificial Intelligence -- Violet Teaming Offers a Balanced Path Forward**
Alexander J Titus, Adam H Russell
arXiv (2023-08-27) http://arxiv.org/abs/2308.14253
DOI: 10.48550/arxiv.2308.14253

188. **Integrating MLSecOps in the Biotechnology Industry 5.0**
Naseela Pervez, Alexander J Titus
IntechOpen (2024-05-10) https://www.intechopen.com/online-first/89417
ISBN: 9780850144840