Future-proofing the DSI multilateral mechanism: possible implications of artificial intelligence & other upcoming technologies

Executive Summary

Artificial intelligence (AI) is transforming the use of digital sequence information (DSI) in the life sciences. AI applications have major implications for benefit sharing from DSI on genetic resources under the Convention on Biological Diversity (CBD) as well as other UN fora that address DSI. This report aims to support policy makers by facilitating a better understanding of AI applications on DSI in order to "future-proof" the design of the multilateral mechanism (MLM) by ensuring that the AI-based DSI benefits will be captured.

At the CBD's 15th Conference of the Parties (COP15), member states agreed to establish a MLM for benefit sharing from the use of DSI on genetic resources that includes a global fund. However, the current rapid evolution of the technologies associated with DSI - particularly the rise of AI applications on DSI - generates unique challenges in ensuring a robust, future-proofed mechanism. AI technologies are capable of using, analyzing, interpreting, and even generating new DSI on an unprecedented scale, offering both opportunities and challenges for benefit-sharing frameworks globally.

AI on DSI has a transformative impact in several fields and will lead to innovations in genomics, molecular biology, medicine, and beyond. The ability of AI to manage, analyze, and map information associated with huge and diverse biological datasets, enables it to generate novel contextual and predictive DSI information and even design entirely new DSI sequences and biological structures.

A noteworthy example amongst artificial intelligence-driven models is the new-found ability to rapidly predict protein folding. DeepMind's AlphaFold has revolutionized the prediction of 3D protein structures which enables researchers to better understand protein functions and accelerate the development of DSI-based applications and products. Generative AI applied to DSI

is another powerful tool. This application can create new DSI sequences that do not exist in nature. For example, generative AI supports the design of proteins or DNA or RNA sequences with specific or optimized functions, opening the door to breakthroughs in synthetic biology and biotechnology.

The report also discusses the "black box" nature of AI models. AI models are trained on vast DSI datasets, mixing information obtained using different types of biological data from multiple databases. The complexity of most AI models makes it virtually impossible to map the contribution of individual sequences to the final AI result.

To ensure that the MLM can adapt to the evolving role of AI in DSI research and of future technologies, the report offers the following recommendations:

Capture the collective benefits derived by the use of DSI: A future-proof MLM must be designed to recognize the aggregate value of DSI. Individual DSI are becoming irrelevant and the collective whole is the valuable commodity. The monetary benefit-sharing triggers should anticipate and account for this scientific reality.

A broad definition for DSI: A broad definition of DSI that includes DNA, RNA, proteins, metabolites, and other biologically active molecules will best ensure that future innovations derived from the application of AI to DSI are covered by the benefit-sharing framework. By adopting a broad definition, the MLM would be better positioned to accommodate future technological advances beyond current AI applications.

Revenue-based triggers: Triggers for benefit sharing should be based on the profit generated by AI applications that use DSI commercially. A trigger based on products or services could miss out on future AI applications and thus a trigger based on aggregate financial triggers such as turnover, sale, or profit seems more future-proof.

Horizon Scanning for future technologies: Policymakers must remain vigilant about new AI applications and technologies that could impact DSI benefit sharing. Regular horizon scanning and expert consultations can help ensure that MLM evolves in step with technological advances.

Background

At COP15, Parties to the Convention on Biological Diversity (CBD) agreed to establish a multilateral mechanism (MLM) including a global fund for benefit-sharing from the use of digital sequence information (DSI) on genetic resources[1]. At the first DSI Open-Ended Working Group meeting in 2023, during discussions to determine possible elements for the mechanism, Parties highlighted the need to ensure the mechanism is future-proof and captures, inter alia, the results of artificial intelligence (AI) applied to DSI on genetic resources[2] while recognizing the difficulties in achieving this goal as the future is impossible to predict[3].

The objective of this report is to identify 1) how AI is used in DSI-related research in the life sciences and 2) whether those uses have implications for benefit-sharing to the DSI MLM and mobilization of resources at scale. This report will consider the most relevant AI applications that could have implications for futureproofing the multilateral DSI benefit-sharing system for DSI under the CBD (the potential scope of DSI is discussed below in "What is DSI?").

Given time and resource constraints, the primary focus of this report is on the use of "generative AI", that is AI that can generate brand-new DSI or novel predictions. In the interest of brevity and given the focus on benefit-sharing, the report does not provide a comprehensive analysis of AI in the life sciences.

Important advancements, for example, the use of AI to mine scientific publications or large text-based repositories or other emerging applications are not covered here. Similarly, the report does not address important societal, and ethical concerns (Messeri and Crockett 2024), as our mandate was to focus on the intersections of DSI, AI, and benefit-sharing in a compact manner.

Why is future-proofing important for designing the DSI MLM?

Negotiated in 2010 and entering into force in 2014, the Nagoya Protocol (NP) is a supplementary agreement to the CBD. Its main objective is to create a legal framework that provides clarity and predictability for both providers and users of genetic resources (GRs) which was, ultimately, intended to lead to greater benefit-sharing. The NP requires the fair and equitable sharing of monetary (e.g., royalties) and nonmonetary (e.g., scientific training) benefits resulting from utilization of GRs[4]. During the decade before the negotiation of the NP, the issue of whether genetic data should be included in the agreement was raised and debated, but, ultimately, it was left out. High-throughput DNA sequencing was still a relatively nascent technology and a cutting-edge field led by a few universities and research institutes. For example, the massive and very expensive Human Genome Project was first completed in 2009 (Dolgin 2009) and was a front-page breakthrough.

Since the NP was negotiated, there has been an explosion in DNA sequencing (Cantelli et al. 2022) (Figure 1) and significant advances in related technologies that generate and use

1. https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-09-en.pdf
2. Paragraph 70, https://www.cbd.int/doc/c/b3c5/e301/e4cdc9663fb0001e5196ef8e/wgdsi-01-l-02-en.pdf
3. https://www.cbd.int/doc/c/50f2/3f82/e1db68327616c51aae0cd29f/wgdsi-02-02-add1-en.pdf
4. https://www.cbd.int/abs

DSI. The use and study of DSI, such as DNA and RNA sequences, is now indispensable for understanding the biological mechanisms underlying GR, monitoring biodiversity, developing new genetic varieties, and developing new therapies and commercial products. DNA sequencing, DNA synthesis, and high-throughput, cloud-based bioinformatics have become standardized, cheap, and mainstream practice across many scientific fields. Hundreds of millions of DNA sequences are now available in thousands of public and private databases which are used daily by millions of researchers across the world.

The parties negotiating the NP did not and perhaps simply could not anticipate the genomics revolution that would unfold in the decade ahead; therefore, the NP did not address the complexities of a world in which genetic data can be readily digitized, shared, and analyzed globally at high speed and scale. No law or policy can ever be completely "future-proof", so policymakers often avoid going into details and use ambiguity to allow for flexibility in the system for future eventualities. This ambiguity has led to a decade of negotiations under the CBD and other UN fora to address the gap in the NP regarding sequence data and its implications for benefit sharing.

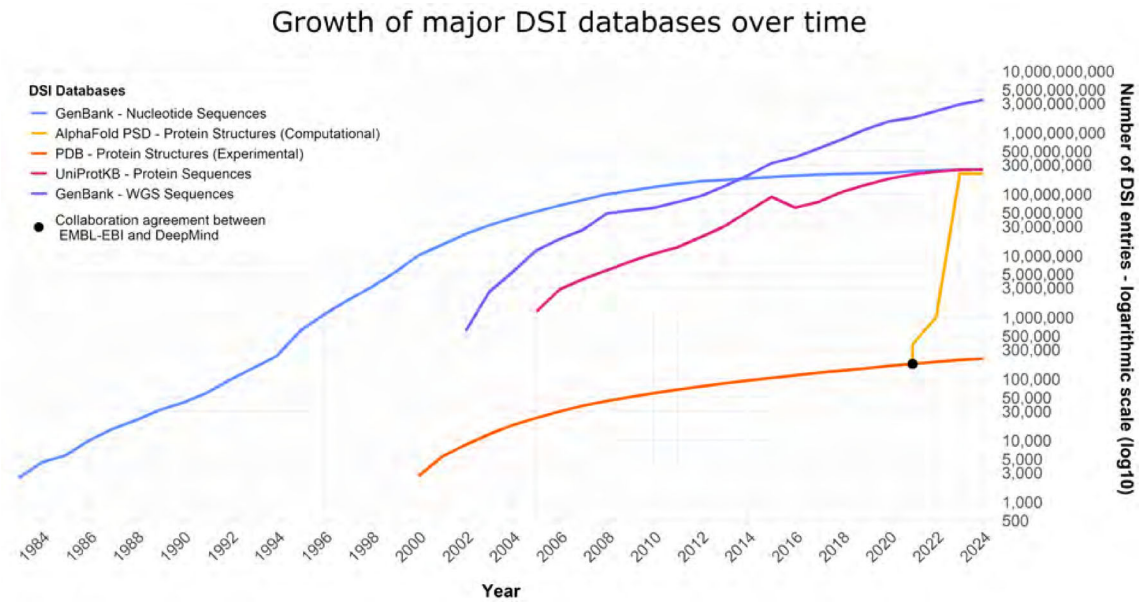


Figure 1: DSI databases have seen a steady and progressive increase in DSI entries over the last decades, with the GenBank database (light blue line) seeing a doubling of its nucleotide sequences deposited approximately every 18 months including growth in Whole Genome Shotgun sequences (purple line). UniProtKB, the database that stores non-redundant protein sequences (magenta line), exceeded 245,000,000 protein DSI in 2024 (The UniProt Consortium et al. 2023). The dramatic drop in the number of DSI proteins in the UniProtKB that occurred between 2015 and 2016 is due to the removal of redundant sequences that resulted in the elimination of about 50,000,000 sequences, almost half of the entire database in 2015 (Bursteinas et al. 2016). Protein Data Bank (orange line), a repository of three-dimensional protein structures, is continuously enriched with three-dimensional structures generated painstakingly by "wet-lab" crystallography experiments, reaching 213,045 structures in 2024. A breakthrough in the study of three-dimensional DSI structures of proteins occurred in 2021, when a collaboration agreement was signed between EMBL-EBI and DeepMind, the company that developed AlphaFold (black dot in the graph). By training AlphaFold's models on the protein structures available from PDB and the sequence data from the INSDC and other databases, AlphaFold was able to computationally increase the number of three-dimensional protein structures available to researchers from 174,395 to 365,000 in 2021 alone. In 2022, AlphaFold' generated structures reached 995,000 units, and by 2023, due to AlphaFold 2, the number of structures available in the AlphaFold Protein Structure Database, yellow line, jumped by three orders of magnitude in a single year reaching 214,000,000 (Varadi et al. 2024).

In the current negotiations on DSI, policymakers are focused on single sequences, individual DSI databases and DSI used in individual products. Yet these are the concerns of today not of tomorrow. AI technologies can analyze, interpret, and generate massive amounts of digital genetic and other molecular biological data much faster and more accurately than traditional methods. AI-driven breakthroughs on and with DSI have the potential to improve the way sequences and other molecular data are generated and used, resulting in improved global scientific outputs and foreseeable rapid advances in fields such as genomics, molecular biology, plant breeding and agriculture, industrial biotechnology, clean energy, vaccine and drug design, and personalized medicine to name a few.

The following analysis is intended to harness the expertise of the DSI Scientific Network[5] to provide a new perspective for the current DSI negotiations. What is the cutting edge of DSI use now? What should one consider in designing the policy framework of the DSI MLM to ensure that the access and benefit-sharing (ABS) community does not end up 10 years from now in a post-AI, post-DSI MLM benefit-sharing debate? What are the benefit-sharing implications of AI applied to large DSI datasets? As the examples below demonstrate, new opportunities and horizons for DSI use through AI technologies abound and are changing the life sciences. If a DSI MLM does not capture and deliver the varied types of benefits from AI applied to DSI, this will result in a smaller global fund with fewer benefits. That result will lead to frustration and mistrust from all stakeholders and Parties.

Different types of AI can be applied to DSI

AI is a field of computer science dedicated to the development of systems and programs that resemble human intelligence in their operations and their ability to generate creative or optimized outputs. These technologies use complex algorithms and mathematical models that enable computers to learn from data, adapt to new information, and improve their performance over time without the need for task-specific programming. AI is routinely applied in data analysis and is changing many fields that rely on digital applications.

While there are considerations about the development of AI models with superhuman capabilities and corresponding ethical and socioeconomic repercussions in human societies, these analyses are beyond the scope of this report. AI applied to DSI is characterized by being mainly task specific, optimized to the solution of well-defined biological problems, and which is reliant predominantly on machine learning and, more specifically, deep learning (a subset of machine learning).

AI can be developed to be able to handle different types of data, such as text, images, audio, numerical data, as well as DSI which (depending on the definition) can take the form of text strings (e.g. DNA, RNA, or protein sequences), 3-D structure (e.g. for proteins), matrices of interaction partners (for protein-molecule docking) amongst other forms. Most DNA sequence data, for example, the large datasets available in the International Nucleotide Sequence Database Collaboration (INSDC), provide an excellent type of clean, structured, pre-processed input that can be analyzed efficiently by AI algorithms.

Several common types of AI can be applied to the study and research of DSI. The integration of **predictive AI** to biological data promises to revolutionize fields such as structural genomics, proteomics, and metabolomics. Through the use of sophisticated machine learning models, AI systems are trained to interpret the complex interactions between DNA, RNA, proteins and small molecules. During the training phase, the model begins to understand annotations associated with DSI (i.e. the metadata on what the raw data means). Once trained, the model is able to generate novel annotation information to unknown DSI, and is thus making a prediction based on all data it has ever encountered.

5. https://dsiscientificnetwork.org/

The application of predictive models to entire databases makes it possible to computationally characterize millions of DSI sequences in a very short time, bringing a revolutionary advancement in our understanding of biological data that would otherwise require decades of work with experimental methods, thus accelerating research and development processes.

Generative artificial intelligence goes a step further, allowing the creation of completely new DSI sequences that do not exist in nature, based on the researcher's specific requirements (Winnifrith, Outeiral, and Hie 2024). For example, when asked to generate a protein that can bind to a particular substrate, the model will create and explore new DSI sequences that meet the desired criteria. During the generative process, the model creates new DSI and explores computationally their function or three-dimensional conformations optimizing those that meet the desired criteria, such as genes with specific functions or proteins that have the ability to bind a certain substrate. Eventually, the model will converge on a few final candidate DSI that will be made available to the researcher. This approach paves the way for innovations in synthetic biology, drug design and protein development with biotechnological applications, enabling the design of DSI with functions never observed in nature.

Large Language Models (LLMs) have also found powerful applications in biology. Just as they are used to process and "understand" human languages, LLMs can analyze vast sets of genetic information, such as DNA, RNA and protein sequences. By recognizing the patterns and structures underlying these sequences, LLMs can begin to "speak" the language of genomics, predicting the functions and regulation of genes through a process known as annotation (Benegas, Batra, and Song 2023; Sanabria et al. 2024). These AI models, trained on huge data sets of genetic sequences and related functional annotations, can infer the biological rules that link specific nucleotide or amino acid sequences to given functions.

As LLMs develop a contextual understanding of these sequences – analogous to how it learns language patterns in human languages – they can make predictions even for completely unknown genes (i.e. without known homologs) or nucleotide sequences that belong to unknown regions of the genome (i.e. so-called junk DNA that we now know likely has novel and interesting functions). AI can thus help discover entirely new genomic regions, expanding the boundaries of our knowledge in molecular biology. Moreover, because LLMs can be continuously updated with new genomic data, their predictive power will grow stronger over time, constantly evolving along with the expansion of the data we will make available to it.

Discriminative models, a subset of LLMs, have more targeted applications. These models are designed to recognize specific features within DSI, using supervised training to categorize data based on well-labeled datasets (He et al. 2019). Supervised training is a training process in which the AI model is fed labeled data, meaning that each input is matched with a corresponding correct output or annotation.

Trained on extensive databases of genetic sequences, discriminative models are able to identify species specific sequences and to locate genetic variants associate to DSI segments. This can be used, for example, to better understand the evolutionary relationships between DSI in support of taxonomic identification (Łysko et al. 2022), illegal wildlife trade, and invasive species monitoring. Furthermore, discriminative AI can be employed to monitor and conserve biodiver-

sity by analyzing the genomic sequences of natural populations and identifying endangered species or discovering new taxa. Thus, aiding the discovery of new genetic resources, the implementation of targeted conservation interventions and the protection of habitats and species that are critical to a particular ecological system.

Together, these AI-driven models – predictive, generative, linguistic, and discriminative – not only improve our understanding of biological data, but also pave the way for revolutionary advances in synthetic biology, conservation, and biotechnology.

What is DSI actually?

Before diving into the application of AI to DSI, we faced the problem that it is not clear what DSI actually is. To date, the Parties of the CBD have agreed to continue to use the expression "digital sequence information" as a placeholder term[6]. However, the 2020 Ad Hoc Technical Expert Group (AHTEG) on DSI received a study on the potential definition of DSI and formulated a range of options of how DSI could be understood (Figure 2)[6].

In its most narrow definition (group 1), DSI would be defined as nucleic acid sequences, which include both DNA and RNA and constitute the core genetic information of an organism. The intermediate definition (group 2) would include all DSI types in group 1 and add amino acid (protein) sequences that, when folded inside a living cell, take on structure and become the "machines" of the cell carrying out the daily work of making, breaking, transporting, and recognizing other molecules. The definition of DSI could be expanded further (group 3) to include all DSI types in group 2 and add metabolites and biologically-active molecules as well. The AHTEG recommended that subsidiary information should not be included in the definition.

For the purposes of this report, we have collected examples of AI applications for groups 1,2, and 3 (Figure 2) with the aim to provide a broad overview of how AI can be used on all major molecular entities obtained from genetic resources and stored digitally in databases.

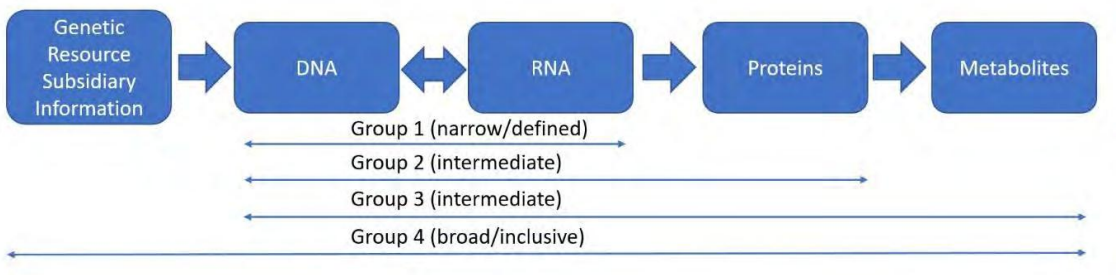


Figure 2: Grouping proposed for digital data of molecular information derived from genetic resources (adapted from[6]).

6. CBD/DSI/AHTEG/2020/1/3 Digital Sequence Information on Genetic Resources: Concept, Scope and Current Use https://www.cbd.int/doc/c/fef9/2f90/70f037ccc5da885dfb293e88/dsi-ahteg-2020-01-03-en.pdf

AI can be applied to all types of molecular biological data

In the following section we present examples of AI applied to DSI following the "central dogma" of molecular biology[7] – the uni-directional flow of genetic information from DNA to RNA to protein and, then moving onward, to metabolites (Figure 3), which mirrors the potential broad definition of DSI.

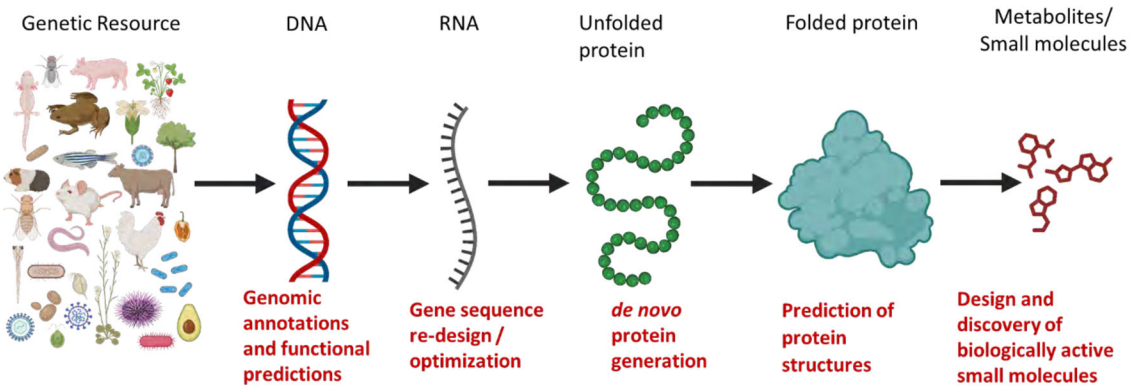


Figure 3: AI applications of DSI (in red) mapped according to the type of biological molecule (DSI type) they are targeted towards.

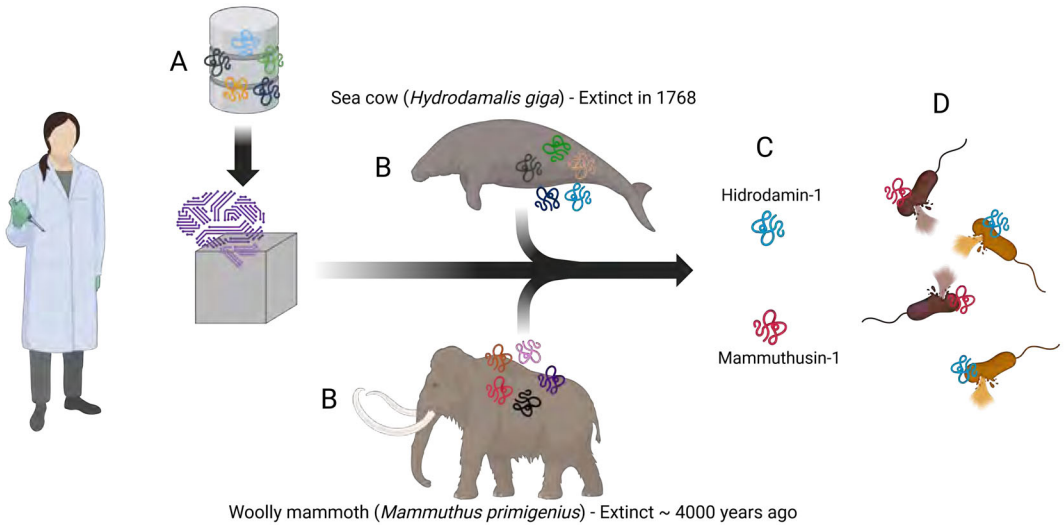
These examples are based on literature research to identify the main areas of AI application to DSI, discussions with DSI Scientific Network members, and expert interviews to identify, evaluate, and select representative case studies. A more exhaustive list of examples of AI on DSI can be found in Annex 1. Here a brief overview:

1. **AI applied to DNA and proteins: Genomic annotations and functional predictions**. AI can be applied to protein or DNA sequences generated by genome sequencing projects to produce improved functional annotations and predictions of gene function, thereby increasing information content and filling in previously large gaps in information.
2. **AI applied to RNA: Gene Sequence Optimization.** AI models can suggest how to optimize the expression of genes via non-coding **RNA** (which are like on-off switches that activate the expression of a gene) to make them more efficient in their biological tasks compared to the original sequences.
3. **AI applied to protein folding: Prediction of protein structures**. AI can predict **protein structures** which helps scientists understand how proteins work and interact with each other.
4. **AI applied to protein sequence: *de novo* protein design.** AI can generate novel **protein** sequences that do not occur in nature, which gives researchers new tools to invent creative biological solutions for different biological functions and applications.
5. **AI applied to metabolites.** Finally, AI can support the design and characterization of **metabolites**, which facilitates the development of active ingredients and pharmaceutical

7. https://www.genome.gov/genetics-glossary/Central-Dogma

compounds and interactions with proteins and other parts of the cell.

1. AI applied to DNA and proteins: Ggenomic annotations and functional predictions



Hien is a senior researcher working on antibiotic-resistant infections, which cause about 1.27 million deaths annually (Murray et al. 2022). Her research focuses on identifying short amino acid sequences (polypeptides) with antimicrobial activity that can be found scattered throughout the protein assemblies (the proteome) of many living organisms.
To efficiently expand her research and identify hitherto previously unknown polypeptides, Hien used an artificial intelligence model trained with many antimicrobial peptides obtained from databases (A). The AI trained by Hien will be able to identify these types of peptides in the proteome of extinct organisms such as the woolly mammoth and the sea cow (B). Partial proteomes of extinct genetic resources are available in the INSDC databases.
The AI model enables her to identify two peptides belonging to the two extinct animals (C), successively Hien “de-extinguishes” the two macromolecules in her laboratory, one from the mammoth and one from the sea cow, and tests them for antimicrobial activity (D) with positive results.
Upon further investigation, the two peptides may contribute to the development of new medicine to overcome antibiotic resistance.

Background

AI can be trained to recognize DNA’s and proteins’ “hidden” structural and regulatory patterns. This allows genomic and amino acid sequences to be better and more rapidly annotated by the model, identifying protein and genomic regions that may have key functions. AI can detect specific patterns like motifs conserved between different GRs that indicate functionally important regions of proteins, of the genome, and to discover evolutionary relationships between different

organisms.

Summary

Molecular de-extinction aims to resurrect molecules from extinctic GR to address current biological and biomedical problems, such as antibiotic resistance. This study demonstrates that deep learning can be used to explore the proteomes of existing and extinct organisms in search of antibiotic peptides. The authors trained deep learning models to functionally predict the antimicrobial activity of 10,311,899 peptides, identifying 37,176 sequences with putative broad-spectrum antibiotic activity, of which 11,035 are not found in current organisms. Of these, 69 peptides were synthesized and successfully tested against bacterial pathogens. Key compounds showed efficacy against infections in mouse models, suggesting that molecular de-extinction supported by deep learning may accelerate the discovery of new therapeutic molecules (Wan et al. 2024; Maasch et al. 2023).

Biological objective

- Identification of peptides with antibiotic action at scale, including DSI obtained from extinct genetic resources.

Data input into the AI model

- 11,581 peptides from public and private databases.
- In-house dataset of 14,738 antimicrobial activity data values obtained from 34 bacterial strains.

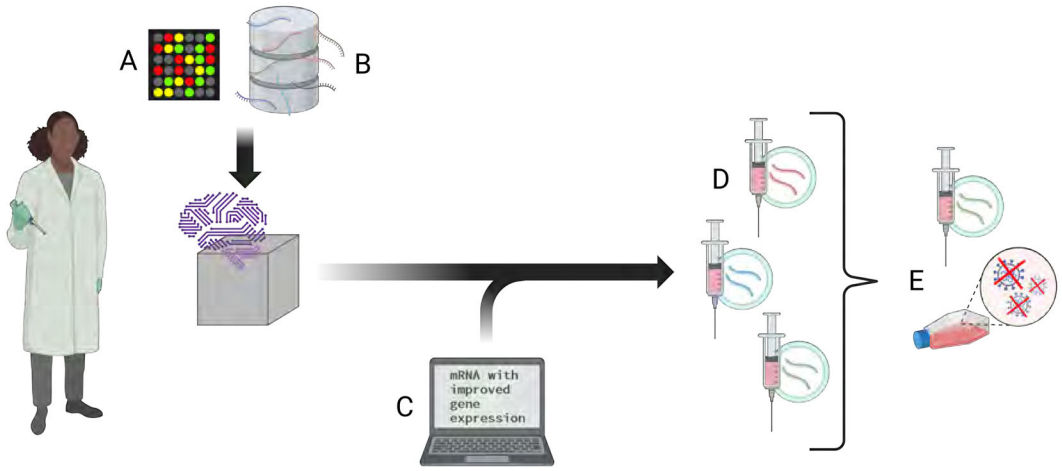
AI Model outputs

- Identification of new molecular traits.

Innovations produced by the model

- **Scalability:** The model can be trained on larger datasets.
- **Integration of multimodal data:** The model combines multimodal data, such as peptide sequences and antimicrobial activity data. Moreover, in principle it can be expanded to include three-dimensional sequences to obtain more accurate predictions.

2. AI applied to RNA: Gene Sequence Optimization



Ayanna is a researcher working on mRNA vaccines who wants to develop more efficient vaccines that produce more vaccine with less mRNA to improve tolerability of vaccines. Ayanna wants to maximize the potential of the information available in the scientific literature about mRNA expression performance, a task that would be impossible to achieve in a sensible timeframe if done by humans.
She develops a multimodal artificial intelligence model (i.e., a model that can integrate information from different types of data) into which she inputs experimental results from millions of sequences (A) and the associated sequences obtained from databases (B).
Ayanna prompts the AI model to generate make messenger RNAs (mRNAs) with better gene expression performance (C).
The AI model suggests some candidates to her (D), which Ayanna will test experimentally by eventually isolating those that perform better (E) and can be further characterized in preclinical studies.

Background

DNA and RNA can be modified by advanced computational and biotechnological methods to improve specific biological attributes like stability, folding, and expression capacity. AI can suggest how to alter and optimize genetic sequences to enhance the efficiency of gene expression which can boost, for example, protein production. By analyzing vast amounts of genetic data, AI algorithms identify patterns in the DNA to suggest modifications that will increase the effectiveness of gene transcription (i.e. reading DNA to make mRNA) and/or the protein translation processes (i.e. reading mRNA and producing proteins). This enables the design of synthetic genes with more efficient properties (optimized codons (i.e., optimizing the ability to produce protein), regulatory elements, and structural motifs) that maximize yield or biological functionality. AI-driven optimization can also reduce the occurrence of natural errors and inefficiencies in gene expression and improve the stability and solubility of proteins.

Summary

The regulatory region at the beginning of mRNA (5' UTR), is critical for the regulation of translation and protein expression. A new language model, called UTR-LM, was developed and pre-trained on naturally-occurring 5' UTRs from different species. Structural and energetic information was also added to the model. UTR-LM was able to generate optimized 5' UTRs sequences that allow high protein production compared with their nonoptimized counterparts. Optimized 5'UTRs can be used for the production of more efficient mRNA vaccines, which produce more antigens (i.e. immune system stimulation) with less vaccine mRNA, or for the development of gene therapies (Chu et al. 2024).

Biological objectives

- Improvement of the expression of DSI of interest.
- Discovery of mutations that increase the stability or functions of mRNA or protein DSI.
- Identification of modification to DSI that can be included in vaccine or therapeutics to modulate immune responses.

Data input into the model

- Non-coding DNA: 214,349 5' untranslated region (5' UTR) nucleotide sequences from the Ensembl database.
- Experimental information (microarray data) on the gene expression levels of ~2,315,000 untranslated region (5' UTR) nucleotide sequences.
- 5' UTR annotations from DSI databases.

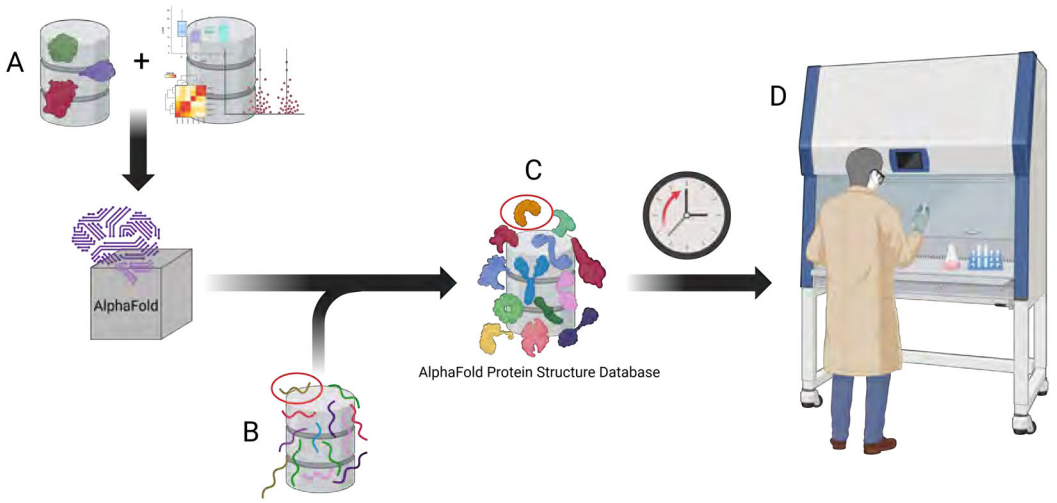
Model outputs

- A set of 20 new DSI 5'UTR nucleotide sequences with improved gene expression profiles.

Innovations produced by the model

- **Efficiency in experimentation:** AI model massively reduces the number of laboratory experiments needed to find optimal gene expression sequences, saving time and resources.
- **Scalability:** Models can analyze and optimize large DSI datasets, accelerating the discovery of sequences useful for different applications.

3. AI applied to protein folding: Prediction of protein structures



Darwin is a PostDoc working on optimizing an enzyme with detergent properties that he has characterized and deposited in the INSDC database as part of his PhD thesis. In his PostDoc project, he aims to decrease the temperature at which the enzyme is active. Making the enzyme more efficient at low temperatures would allow it to be used in household detergents that would work at lower temperatures thus saving energy during washing.

The enzyme is very difficult to crystallize in the laboratory, and during his PhD Darwin has been unable to accurately determine its three-dimensional structure. Only by knowing the three-dimensional structure could Darwin identify the parts of the enzyme that contribute to the detergent activity so that they could be improved for it to be active at lower temperature.

AlphaFold's AI was trained by integrating information available on the limited amount of three-dimensional structures of proteins available in databases with experimental data from the scientific community (A).

AlphaFold was then able to generalize this information and predict the three-dimensional structure of all "orphan" sequences that had not previously been characterized structurally (B to C, and yellow line in Figure 1).

Among these sequences is the enzyme Darwin is working on, which now has a three-dimensional structure predicted by AlphaFold deposited in the database (red circled protein sequence in B and protein structure in C). Thanks to AlphaFold's prediction, Darwin was able to determine the precise location of the five amino acids responsible for the enzyme's catalytic activity and immediately set to work on optimizing them (D).

Alphafold allowed Darwin to obtain information about the three-dimensional structure of the enzyme of his interest in a fraction of the time compared to his unsuccessful attempts during his PhD, enabling him to continue his PostDoc work seamlessly. Not having access to the three- dimensional structure of that enzyme, would otherwise have been precluded him to pursue his research.

Background

One of the most impactful examples of AI applied to DSI is the Alpha Fold model, which was able to expand the database of predicted 3-D protein sequences by 200-fold, from 1,000,000 to 200,000,000, dramatically increasing in the blink of an eye the information content associated with millions of DSI available in the databases[8].

Summary

DeepMind's AlphaFold 2 and AlphaFold 3 have demonstrated an unprecedented ability to predict the three-dimensional structures of proteins, and the molecules with which they interact, based only on their amino acid sequences. This ability opens the path to a deeper understanding of protein function, the discovery of targeted drugs, and the development of bioactive peptides and antibodies. (Jumper et al. 2021; Abramson et al. 2024).

Biological objectives

- Determination of the three-dimensional structure of a protein from its amino acid sequence.

Data input into the model

- Big Fantastic Database (BFD) covering 2,204,359,010 protein sequences, custom-made by joining the entirety of UniProt database of natural protein sequences of natural and synthetic proteins, a soil reference protein catalogue and the marine eukaryotic reference catalogue.
- Three-dimensional structure information of natural proteins obtained by X-ray crystallography, NMR spectroscopy or electron cryomicroscopy from the Protein Data Bank (PDB).
- Experimental data on the properties and function of proteins, such as catalytic activity, substrate specificity, thermal stability, etc.

Model output

- 3-D model of an unknown protein(s) showing the functional 3-D structure of the polypeptide chain (protein).
- Interactions between the predicted protein and other molecules or proteins such as metabolites, lipids, or small molecules (drugs).

Model innovation

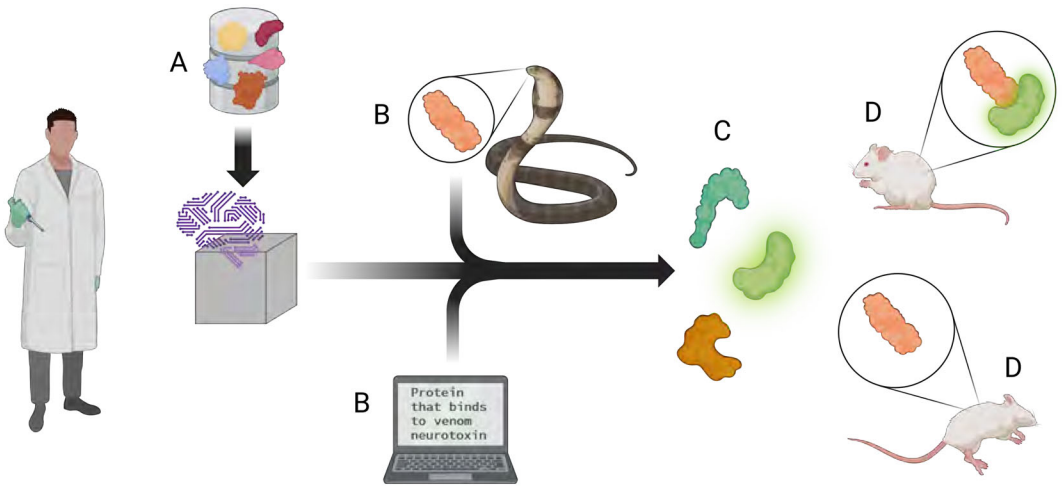
- Unprecedented accuracy: Models such as AlphaFold have achieved levels of accuracy comparable to experimental methods, revolutionizing the field of structural biology by de

8. https://deepmind.google/discover/blog/alphafold-reveals-the-structure-of-the-protein-universe/

facto solving the 50-year challenge on developing the best computational model for predicting the 3-D structure of proteins (Moult et al., 1995).

- Acceleration of scientific discovery: Allows structures of protein DSI that would take years of experimental work to be obtained rapidly in minutes, accelerating research in biology, medicine and biotechnology.
- Access to experimentally inaccessible structures: Allows structural information to be obtained for proteins that are difficult to study by traditional methods, such as membrane proteins or large protein complexes.

4. AI applied to protein sequences: *de novo* protein design



Thiago is a junior researcher working on venomous snake bites, a neglected tropical disease (Chippaux 2017).
In his lab, he wants to identify a possible antidote to the venom of elapid snakes; elapid venom includes a neurotoxin that can cause severe tissue damage, and neurotoxicity that can be life threatening. Thiago wants to leverage artificial intelligence to develop an antiserum by exploring any type of protein that could block the neurotoxin, while not necessarily limiting himself to naturally occurring proteins found in the genetic resources so far characterized.
He therefore uses an AI model that is trained on 100,000 experimentally verified three-dimensional structures of natural proteins and 1,000,000 three-dimensional structures generated by Alphafold 2, obtained from databases (A).
He further interrogates the AI model to obtain proteins that binds the active part of the neurotoxin, inhibiting its toxicity (B). The AI model proposes several synthetic protein models that do not exist in nature (compare A with C). Thiago identifies through laboratory experiments the most promising candidate that protects mouse models from intoxication due to the venom neurotoxin (D). He will continue to analyze in preclinical studies the suggested best candidate protein to characterize it in detail.

Background

AI models can also be trained on the shapes of hundreds of thousands of proteins. The model processes detailed information about the three-dimensional structures of proteins and utilize the physical architecture of the 3-D models to generate entirely new protein sequences that do not exist in nature. The use of this application can accelerate the discovery of therapeutic proteins, industrial enzymes, and advanced biomaterials.

Summary

Snakebites are a neglected tropical disease (Chippaux 2017) that causes more than 100,000 deaths a year and severe disabilities. The highly toxic and sometimes lethal compound found in the venom of elapid snakes is a three-finger toxin (3FTx). Using the deep learning model RFdiffusion, this study developed stable synthetic proteins not found in nature that can neutralize these toxins with high affinity. These proteins, which have demonstrated efficacy in vitro and in animal models, could lead to antivenom treatments that are easier to develop and cheaper than traditional antibody therapies, and therefore accessible even in resource-limited settings (Baker et al. 2024; Watson et al. 2023).

Biological objectives

- DSI design of proteins with specific functions not found in nature, e.g., enzymes with enhanced or novel catalytic activity or therapeutic proteins with increased stability and affinity for predetermined targets.

Data input into the model

- The RFdiffusion model was trained on ~100,000 experimentally-determined three-dimensional structures of natural proteins retrieved from the Protein Data Bank (PDB) ~1,000,000 3-D structures of natural proteins computationally generated by AlphaFold 2 sourced from the AlphaFold DB (https://alphafold.com/).

Model output

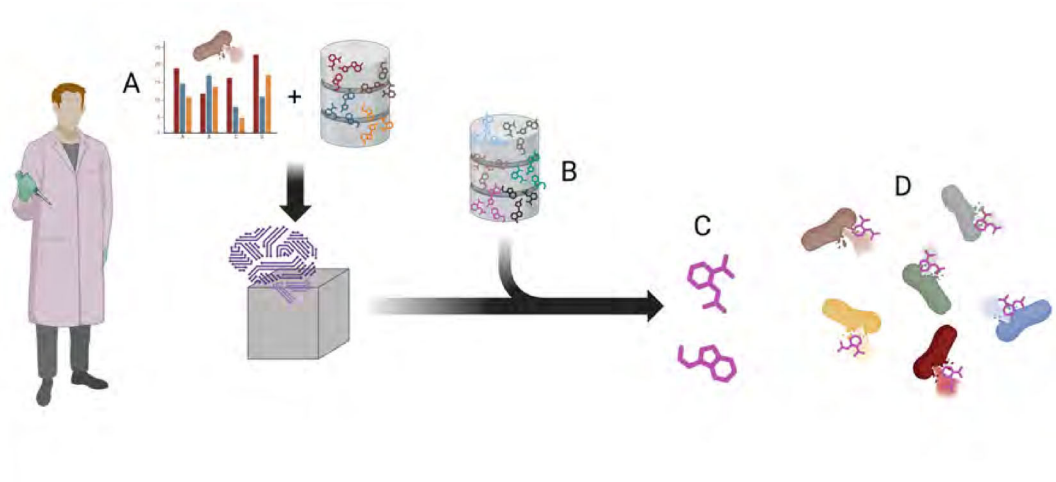
- Short protein sequences with custom biochemical and biophysical properties that have a high affinity and neutralizing ability toward snake venom toxins.

Model innovation

- **Generation of new functionalities:** Design of proteins not found in nature with optimized custom functionalities.
- **Enhanced efficiency in protein design:** Dramatic reduction of the time and cost associated with the design and synthesis of new proteins compared to traditional trial-and-error experimental methods.

- **Advanced protein optimization:** Simultaneous optimization of multiple protein properties that would be difficult to achieve with traditional approaches.

5. AI applied to metabolites: Design and discovery of biologically active small molecules



Hassan is a researcher working on antibiotics and for his master thesis is looking for small, broad-spectrum antibiotic molecules that are active against different bacterial strains.
He therefore decides to train an AI model using as training dataset the conformation information of 2,335 molecules and their lytic activity against the bacterium Escherichia coli (A).
He feeds into the AI model an even larger database of molecules with unknown properties against E. coli (B) and the model presents him with molecular hits that might have broad-spectrum antibacterial activity (C). If he had to do a manual screening of the molecules in the database, he would have had to invest years of laboratory trials having to bear extremely onerous expenses.
After laboratory tests on the few hits suggested by the AI model, Hassan succeeds in identifying among the candidates a molecule with broad-spectrum antibacterial activity (D).
Hassan will focus his research on this molecule to improve its pharmacological properties in preclinical studies.

Background

AI can also be trained on extensive chemical and biological datasets to uncover the fundamental principles of the structure of bioactive molecules, such as how molecules adopt different shapes and configurations and achieve their biological activity. After identifying these key attributes, AI facilitates the optimization of existing molecules and the ex novo generation of new bioactive molecules. This innovative technology has potential applications in precision

medicine, the development of antibiotics, the generation of new phytopharmaceuticals, and the development of biologically active compounds with novel properties.

Summary

The search for new antibiotics is a constant arms race to find new active compounds to bypass the emergence of resistances. To address this challenge, a deep neural network-based AI model was generated to predict molecules with antibacterial activity. The model was initially trained on the empirical efficacy of 2,335 compounds against E. coli. Once trained, it was then run on more than 107 million active chemical compounds obtained from the Drug Repurposing Hub and ZINC15 libraries and led to the discovery of Halicin, a molecule structurally different from conventional antibiotics, and to the identification of eight structurally distinct antibacterial compounds. (Stokes et al. 2020).

Biological objectives

- Identification of chemical compounds with specific biological activity with the intention to develop of more effective or less toxic drugs.
- Optimization of physicochemical and pharmacokinetic properties of already known compounds to improve their efficacy and safety.

Data input into the model

- Empirical data quantifying the E. coli growth inhibition of an FDA-approved Drug Library supplemented by a modest library of natural products, totaling 2,335 molecules.
- Information on the two- or three-dimensional structures of 107,349,233 DSI of existing bioactive molecules, available from the ZINC15 database[9].

Model outputs

- List of chemical compound candidates that can be used for empirical validation experiments.
- Information on the chemical and physical properties and biological activities of the candidate molecules.

Model innovation

- **New bioactive compounds:** Identification or repurposing of bioactive molecules for particular biological activities (such as antibiotics).
- **Accelerated molecular design:** Significantly reduces the time and cost associated with new drug discovery and optimization compared to traditional methods. AI-mediated screen-

9. https://cartblanche22.docking.org/

ing of 107,349,233 DSI of bioactive compounds proposed in this study is two orders of magnitude larger than what standards empirical study by traditional methods can allow.

- **Drug personalization:** Facilitates the design of compounds specific to individual targets, improving the precision of personalized medicine.

Implications of AI on DSI for the Multilateral Mechanism for benefit sharing from the use of DSI

What do the above examples mean for DSI benefit-sharing? What do they teach us about the possible "future of the life sciences" and how might policymakers anticipate the nature of future technological outcomes and thus benefits that comes from DSI? There are five overarching implications that should be ideally be reflected in the design of the multilateral mechanism for DSI benefit-sharing.

1. The AI "black box" means that a traceable direct connection between individual sequences and their quantitative contribution to AI outcomes is not possible

As the above examples illustrate, AI models are trained on millions of data points which themselves are integrated together from many databases as well as experimental and other data sources. Compared to more traditional computational analysis, AI models offer high throughput performance and scalability to extraordinary volumes of data.

The "black box" nature of AI applications is used to convey the observation that concrete input and output from the AI model are observable, but the internal process of transformation and decision making is opaque and difficult and sometimes impossible to understand or explain (Figure 4). The AI model itself – why it gives the answers it does come from a "black box". The opaqueness of AI applications is common in deep learning systems and complex neural networks, where the complexity and nonlinearity of internal operations make it difficult to interpret how decisions are made. **Therefore, the contribution of individual DSI used to train the most common AI models used in the life sciences, or to generate the outputs from those models is neither quantifiable nor traceable.** Outputs of AI models are disconnected from a single genetic resource, DSI, or protein structure. Tracking the country of origin of DSI and resulting AI outputs is simply not possible.

Some have argued that this "black box" poses a risk to society and raised alarm bells about AI applied to DSI[10]. There is concern that black-box outcomes, if wrong, could produce dangerous products. However, DSI-based AI output is only a prediction. The outputs of AI models applied to DSI must be tested experimentally to substantiate their validity and pass through the same regulatory pathways and review processes and procedures of any other life science outcomes – commercial and non-commercial. This critical experimental step in which real-life biology proves or disproves the AI prediction, provides a strict quality control step to determine the

10. 'Black Box' Biotechnology - Integration of artificial intelligence with synthetic biology https://acbio.org.za/gm-biosafety/black-box-biotechnology-integration-of-artificial-intelligence-with-synthetic-biology/

accuracy and efficacy of the outputs. This is a feature and a requirement that many applications of AI do not have in other societal applications. For example, in “deep fake” videos, it becomes hard to tell truth from fiction and once a video has been released who is the ultimate judge? However, in biology, the experiment in the lab or in the field and the actual biological outcomes will necessarily need to be verified. AI outcomes in the life sciences are thus not a final outcome but an important middle step to accelerate R&D. The “black box” nature of AI outcomes is relevant for benefit-sharing design but should not itself be a reason for concern.

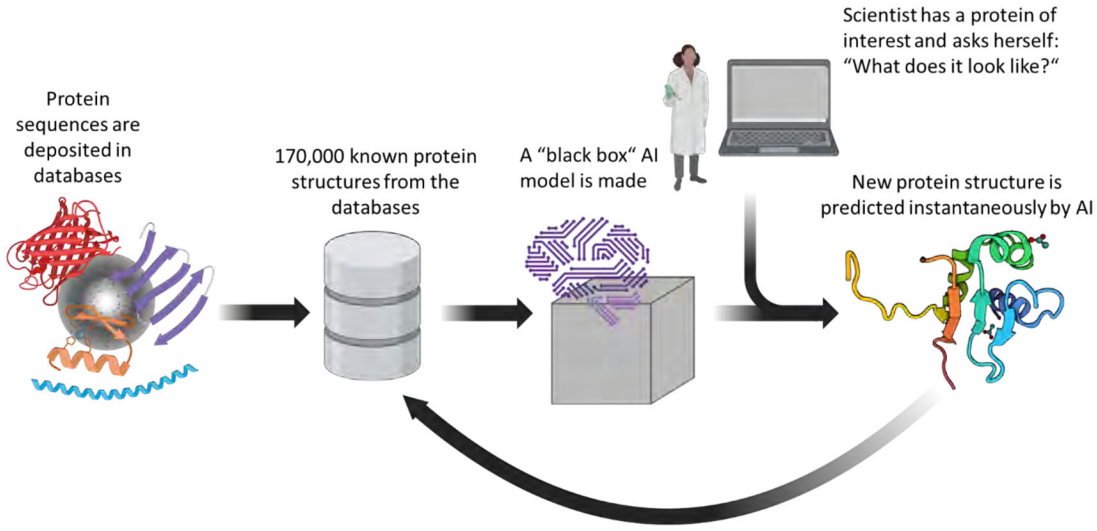


Figure 4: Schematic representation of a black box training model.

2. XAI, explainable artificial intelligence

Explainable AI (XAI) is an emerging field that aims to make AI models more transparent and understandable to humans. In order to address the ”black box” characteristic of many AI systems, whose decision-making processes are often opaque, XAI develops methods to at least partially disclose and explain how a model arrives at a certain conclusion. XAI could be a technology that could help to overcome the black-box problem and its implications for benefit-sharing presented above.

Furthermore, XAI may facilitate the identification and correction of bias in models, improving the fairness and reliability of AI applications. It also provides a better understanding of AI decision-making patterns and could foster greater understanding of DSI and GR being studied by researchers.

Particularly in the medical field, where AI applications have the potential to deliver some outstanding benefits in patient care, AI model decisions can take into account a patient’s medical history and directly contribute to diagnosis and treatment plans. It is therefore crucial for medical practitioners to have an understanding of how the model interprets the data, its possible biases, and the privacy of the data used (Brahma and Vimal 2024). These AI models used in medicine often rely on clinical images or clinical parameters associated with the constitution of

the patient, such as body mass, physiological data, etc. (Fujihara et al. 2023).

One promising approach to make AI models less opaque is given by SHAP (SHapley Additive exPlanations)[11], a game theory-based approach that aims to make any machine learning-based model explainable (Lundberg and Lee 2017). These developments hold great promise for allowing humans to be able to observe the internal decision-making process of AI models, although the authors of SHAP themselves remind us that we need to be cautious from considering the results of such approaches as causal explanations when, depending on the type of data analyzed, they actually provide us with correlational information[12].

However, in the current state of XAI research and based on the major AI models in use for DSI analysis, it is not possible to observe and understand the operation of the AI models as one might do by opening the hood of a car. Thus, the conclusions above remain the same but with the caveat that XAI could change or improve the penetrability of the black-box model over time. At present, XAI offers the possibility to better understand the operation of the AI model through external manipulation and approximations (Jiménez-Luna, Grisoni, and Schneider 2020) but cannot fully explain and attribute causal relationships to AI outcomes.

3. Identifying and monitoring biodiversity and synthetic constructs

One interesting LLM-based application of AI is the quick and accurate identification of the original GR for naturally-occurring DNA sequences. This identification is done by mapping specific molecular markers, like marker sequences often referred to as DNA barcoding (Riza et al. 2023), and associating them with a specific species. This application can be useful to identify unique biodiversity or endemic species. It can also help biodiversity management, protect genetic resources, and contribute to the identification of synthetic constructs.

AI can improve the accuracy and speed of taxonomic and phylogenetic analyses, which are essential for studying the diversity of genetic resources and characterizing their genetic traits. Machine learning algorithms can efficiently analyze large datasets of nucleotide DSI to extract information about genetic diversity and map illegal use of genetic resources.

South Africa recently announced the development and use of the BioInnovation Monitoring Tool (BioMoT), an AI model that can identify use of endemic South African GR in scientific publications and patent applications[13]. BioMoT leverages artificial intelligence to gather data from three major online databases, global patent information, published scientific papers, and commercial listings to detect trends in research, patents, and products. Its outputs facilitate the effective monitoring of South Africa's biological and genetic resources, including DSI, and supports the creation and implementation of policies that ensure these resources are used in accordance with the Nagoya Protocol. BioMoT benefits key regional stakeholders—such as industry, government, and academia—by aiding national biodiversity strategies and by providing early warnings of market trends that might jeopardize the sustainable and equitable use

11. https://shap.readthedocs.io/en/latest/index.html
12. https://towardsdatascience.com/be-careful-when-interpreting-predictive-models-in-search-of-causal-insights-e68626e664b6
13. https://www.abs-biotrade.info/news-1/analysing-the-use-of-south-africas-biological-and-genetic-resources-through-artificial-intelligence-ai/

of indigenous biological resources. Additionally, it helps combat illegal trade of South African flora and allows industry to identify new market opportunities through the tracking of trends in patents and scientific research.

Despite their significant capabilities, these methods are inherently limited in one crucial aspect: they cannot trace the origin of novel, AI-generated DSI of organisms that have geographical distribution outside of a single country. In these cases, they can only tell the user "this genetic data is a tiger" but not "this tiger came from a zoo or from Bangladesh". For synthetic DSI, these methods have no ability to determine origin or location. AI-generated (i.e. synthetic) DSI does not stem from a unique, tangible DSI or genetic resource collected from the natural world. Unlike natural DSI, which can be linked back to a specific species, AI-generated sequences are synthesized by AI model that draw on vast datasets, mixing and matching elements in ways that do not correspond directly to any existing organism and that cannot be tracked back to any exact genetic resources.

4. The legal lines between UN fora, between human DSI and other living organisms that form part of the world's biodiversity, and between synthetic and natural DSI are completely blurred

These new AI innovations are only possible because of the collective contributions from various types of DSI datasets and raw experimental data available in multiple databases. Model training is only possible because of globally available DSI that scientists have shared, made public, and inter-connected over decades. From genomic annotations to design of novel proteins, AI works because it harnesses the power of the aggregate, of the collective whole of DSI databases and experimental results. These are scientific outputs and direct result of international collaboration of researchers from around the world.

Similarly, the data itself cover literally all of biology – from humans to viruses and they are sourced from every single environment on the Earth (and even beyond). The policy implications are quite complex. **Should the benefits that arise from use of the global corpus of DSI, such as AI applications, go to the CBD? To the High Seas? To the Food and Agriculture Organization? Or to the World Health Organization? Or none of the above?** If millions of DSI data points from many databases are used and merged together in a model, who should rightly benefit from the scientific results of that model? Who should receive the benefits?

Similarly, although the CBD has historically excluded human genetic resources from any benefit-sharing requirements, that now becomes virtually impossible. The results of AI on global DSI datasets will, of course, inadvertently, also include data on human genetic resources because, simply put, humans are part of biodiversity. Relatedly, synthetic sequences, although not directly connected to genetic resources, that are a combination of nature and human innovation, would not be possible without the knowledge and information gained from global DSI datasets. Thus, it also makes intellectual sense, that benefits arising from synthetic DSI should also require benefit-sharing.

Many of these complex questions about what is "in" or "out" of scope have been noted prior to the widespread use of AI on DSI. The genomics and bioinformatics revolutions and related

biotechnology applications have been “mixing up” the biological input and output of scientific research for many years now.

This is a desired outcome necessary to solve societal challenges for health, food, conservation, but can create confusion if new legal and administrative measures require the implementing authority to determine which benefits get shared with whom under what conditions and on what legal basis. AI increases the complexity of the policy challenges to find a fair and equitable solution.

5. A definition for DSI is needed sooner rather than later...

The AI examples above demonstrate that there are a wide range of applications of AI on various types of biological molecules and data. If DSI were to be defined as broadly as possible, it would include DNA, RNA, amino acid (protein), metabolites and biologically-active molecules as well: now in group 1-3 (Figure 2). This definition essentially includes all of the bio-molecules and their related data types inside the cell and so a rather clear concept could be established for scientific users. This definition would likely be future-proof as the contents of the cell have remained relatively stable over the past three billion years and are unlikely to change in the coming decades.

If a DSI definition is unclear, then users might interpret a rather narrow definition such as only DNA and RNA data. This could lead to attempt to avoid the multilateral mechanism by targeting R&D activities focused on other biological data types, such as amino acid (protein), metabolites, which would have negative consequences for the global fund. Users (especially those that would be required to pay monetary benefits) might focus R&D efforts on non-DNA/RNA research and focus on other types of biological data outside of the scope of DSI thereby avoiding payment into the global fund.

If an actually definition is deemed to be too time-consuming or difficult to negotiate, policymakers might reference the 2020 DSI AHTEG[14] report and specify which definition of DSI (based on the groups described in the report) they intend without necessarily negotiating a definition.

Policy Recommendations for benefit-sharing from AI-based DSI use

How does all of the above translate into the operationalization of the multilateral mechanism for benefit-sharing from DSI? There are three areas where the lessons above point towards options that could support future-proofing the mechanisms and ensuring that benefits arising from DSI that has been processed through AI are part of the multilateral mechanisms if Parties so wish.

14. CBD/DSI/AHTEG/2020/1/3 Digital Sequence Information on Genetic Resources: Concept, Scope and Current Use https://www.cbd.int/doc/c/fef9/2f90/70f037ccc5da885dfb293e88/dsi-ahteg-2020-01-03-en.pdf

1. A broad scope of DSI

A broad definition of DSI that includes DNA, RNA, amino acid (protein), metabolites and biologically-active molecules could be a clear and useful definition for DSI. Alternatively, policymakers could reference the 2020 study and the AHTEG report[15]. Either way, a definition of DSI that includes DNA, RNA, proteins, metabolites and other celluar small molecules seems to prevent avoidance of the multilateral mechanism. This would thus facilitate the capturing of a wider variety of benefits emerging from the application of AI to DSI. A broader definition would capture a wider spectrum of applications and technologies by allowing to be open to potential innovations not yet foreseeable.

2. Benefit sharing from the collective rather than individual DSI

Benefit-sharing triggers that focus specifically on individual DSI or subsets of DSI or that aim to meticulously track each DSI individually across the value chain are not feasible with AI. These benefit-sharing approaches may overlook the long-term and far-reaching research results from AI applications on DSI. AI has the potential to generate significant research benefits over time leveraging the aggregate use of DSI, which may not be fully captured if benefit-sharing mechanisms focus too narrowly on individual DSI transactions.

3. Overarching triggers rather than product-focused triggers

Over the course of our discussions, we did not find concrete examples of commercial outcomes or consumer products where AI has been successfully applied to DSI at scale. At this point, most concrete commercial activity using AI on DSI is on data analysis and interim data processing steps and predictions.

The design for the multilateral DSI benefit-sharing mechanism should provide for global and aggregate use of the DSI from the start; this would improve the mechanism's ability to achieve its resource mobilization goal. Considering the extensive use of DSI during the design phase of the MLM would ensure the ability to leverage the full range of benefits generated by the global application of IA on DSI. Tech companies that develop and commercialize AI models or software or provide DSI-based services should trigger benefit sharing and pay monetary benefits. As such, trigger points based on revenues from commercial activity of use of DSI seem more likely to be able to capture these types of DSI use activities that do not directly relate to commercial products.

Although AI is expected to have a profound impact in many fields, it is also worth considering that its potential is potentially also hyped. Expectations about AI must be balanced with a realistic understanding of its capabilities and limitations. The intersection of DSI and AI is a relatively new frontier. This emerging field requires further exploration and development before its commercial feasibility and practical benefits can be fully realized. Thus, the DSI multilateral mechanism is well-poised to build in these considerations and, ideally, evolve the mechanism, as the field changes over time. Without question, if AI issues are overlooked in the design of the DSI multilateral mechanism or other UN fora's own DSI mechanisms, they run a high risk

of not delivering as expected.

Outlook and Open Questions

Technological breakthroughs on and with DSI, such as the AI applications referenced in this report, have the potential to improve the way sequences are generated and used, resulting in improved global scientific outputs and foreseeable advances in fields such as genomics, molecular biology, plant breeding and agriculture, industrial biotechnology, clean energy, vaccine and drug design, and personalized medicine, to name a few. The present report only scratches the surface of questions that should be explored further within the context of AI applications that apply to DSI.

- **How AI can help benefit sharing?**

Initiatives that promote principles that can guide the responsible development of AI technologies in the field of protein design call for more equitable participation in the research itself and its benefits[15][16]. How these initiatives can help enable building capacity to address the DSI gap and contribute Non-monetary benefits is an important question that should be further explored.

- **How the field of "explainability" will improve the "black box" problem?**

There is an ongoing development of tools that are dedicated to explain the output of "black box" AI such as SHAP, Lime, Mean Decrease in Impurity or GINI (Saarela and Jauhiainen 2021). A horizon scanning exercise to monitor how these tools will advance and the impacts they may have on current applications is important for the implementation of the multilateral mechanism.

- **Horizon scanning for future AI applications:**

In the interviews carried out for the development of this report, experts signaled the development of new AI applications that use DSI and would be relevant for the multilateral mechanism. An exploration of these innovations to understand their trajectory and their possible impacts on the multilateral mechanism could be helpful to Parties.

The DSI Scientific Network can build on the results of this report, and work on these questions, developing a more in-depth analysis on whether AI will have implications for benefit-sharing to the DSI multilateral mechanism and the mobilization of resources at scale per the COP16 decision. The activity will also serve to counter-balance the more alarmist perspectives on AI coming from the Third World Network [17].

Methodology for figure 1

GenBank datasets (Nucleotide sequences and WGS sequences) were downloaded on August 22, 2024 from this link: https://www.ncbi.nlm.nih.gov/genbank/statistics/. AlphaFold data were extracted from (Varadi et al. 2024). PDB data were collected from https://www.wwpdb.

15. https://www.ipd.uw.edu/responsible-ai/
16. https://responsiblebiodesign.ai/
17. 'Black Box' Biotechnology – Integration of artificial intelligence with synthetic biology https://acbio.org.za/gm-biosafety/black-box-biotechnology-integration-of-artificial-intelligence-with-synthetic-biology/

org/stats/deposition, while UniProtKB data are those pertaining to release number 1 of each year obtained from here:https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/, https://www.uniprot.org/help/synchronization (The UniProt Consortium et al. 2023). The datasets were compiled into a .csv file and analyzed with R.

Acknowledgments

The main text of the report was drafted by Davide Faggionato, Pablo Orozco and Amber H. Scholz. We would like to express our sincere gratitude to the DSI Network members and in particular to the members of the Working group on AI of the DSI Scientific Network for their valuable contributions to the preparation of this document: Andrew Lee Hufton, Hiroko Muraki Gottlieb, Débora Raposo, Aylin Haas, Ann Mc Cartney, Masanori Arita, Mathieu Rouard, Carolina dos Santos Ribeiro, Alejandra Sierra, Guilherme Oliveira, Irma Klünker, Hanzhi Yu, Gabriele Rinck as well as the broader Network membership for their support and discussion. The expertise and input offered by the members greatly improved and enriched the document.

We are very grateful for the valuable input provided by the AI experts who participated in the expert interviews: Anna Poetsch and Melissa Sanabria, Technical University Dresden; Marina Camacho Sanz and Cristian Izquierdo Morcillo, University of Barcelona; Ian Haydon, University of Washington; Campbell Watson, IBM Research.

The DSI Scientific Network (https://www.dsiscientificnetwork.org/) developed this report as a knowledge product with funding from NORAD to support Parties in the inter-sessional period and beyond.

All figures except Figure 1 and 2 were generated with Biorender.com. The header banner on the start page was generated using the artificial intelligence of DALL-E.

Disclaimer regarding the examples

The examples and people represented for the five applications are fictional and are intended to explain in simplistic terms the scope of the five applications by relying loosely on the case studies. They are therefore not intended as specific examples of the case studies. For more detailed descriptions of the case studies please refer to the texts below the boxes with the examples and the primary literature from which the case studies are drawn.

ANNEX: Other technologies with implications for benefit-sharing

1. Crowdsourcing DSI analysis

Through crowdsourcing, small analysis packages solved by humans can be integrated into more complex analytical frameworks. AI has the potential to revolutionize the field of crowdsourcing by contributing greater efficiency, accuracy and scalability to data collection and analysis.

Scientific projects that aim to investigate large, computationally intensive DSI datasets, can fragment DSI data analysis transforming it in smaller tasks, which can then be solved by humans in the form of a small assignments included, for example, in video game or a website plug in.

AI can increase the performance of output collection from distributed sources such as video games, social media, or user feedback platforms. Machine learning algorithms can preprocess data for human analysis (Sarrazin-Gendron et al. 2024) or can filter and refine the collected results, eliminating irrelevant or duplicate information and integrating the packages of information with each other to achieve the result of the DSI analysis. This procedure makes data collection faster and more efficient, while ensuring that the data are restructured uniformly for formal analysis.

This type of decentralized approach to DSI processing, allows hundreds if not thousands of users to contribute to the final result. A single user, as anticipated by traditional benefit-sharing, is non-existent. In addition, because users interface with DSI through a third-party platform (a commercial video game, a social medium, a web page), they can contribute to DSI use from widely disparate geographic locations and be subject to dramatically different regimes of privacy and anonymity.

2. Synthetic communities of microorganisms (SynCom)

Microbiomes are increasingly recognized as critical contributors of ecological services that unite human health, animals and the environment . Synthetic communities of microorganisms, also known as SynCom, are communities of microbial species that are assembled artificially to mimic the functions of natural microbiomes or to generate microbiomes with determined characteristics. They are composed of a definite number of microbial species that interact with each other according to well-characterized ecological relationships. SynCom lend themselves to the exploitation of interactions and synergies existing between microbe species, between microbes and their hosts or between microbes and the environment. The applications of synthetic communities of microorganisms are many and span a variety of fields, including agriculture, medicine, the environment, and industry (D'Hondt et al. 2021).

In agriculture, synthetic microbial communities can be used to promote plant growth by improving nutrient uptake and protecting plants from pathogens or environmental stresses; these applications can reduce dependence on chemical pesticides and fertilizers, making agriculture more sustainable (Shayanthan, Ordoñez, and Oresnik 2022).

In medicine and human health, SynComs can be used to study and modulate the human gut microbiome, therefore improving digestion and metabolic health, dysbiosis (microbiome imbalance) and inflammatory bowel disease. Similarly, synthetic communities can be used to improve the microbiome of animals thus reducing the need for antibiotics while improving productivity and welfare on livestock farms (Leeuwen et al. 2023).

Synthetic microbial communities have many applications in industry and biotechnology, they can be used to produce biofuels through biomass fermentation, as these communities are gen-

erally more efficient and resilient than individual microbial strains, allowing for longer production cycles. In addition, SynComs can be engineered to efficiently produce chemicals such as antibiotics, vitamins, enzymes, and other pharmaceutical or industrial products through optimized fermentation processes that require different chemical reactions. These chemical reactions for bioactive molecules are often difficult to reproduce step by step in the laboratory but can be reproduced as metabolic chains by microbial communities (Oleskowicz-Popiel 2018).

In bioremediation and environmental protection, synthetic communities have the potential to be engineered ad hoc to degrade pollutants such as hydrocarbons, plastics, or pesticides. Another advantage of synthetic microbial communities in research and development is that they can be used to study complex microbiomes such as gut or soil microbiomes under laboratory and controlled conditions, reducing the complexity of their natural counterparts, without losing key ecological and functional dynamics (Gianetto-Hill et al. 2023).

With more recent advances revealing the fundamental role of microbiomes in multiple areas of global ecosystems, synthetic microbial communities offer enormous potential for a variety of applications. With microbiome engineering, rational design and the support of AI, the future possibilities of these systems are almost limitless.

3. Molecular de-extinction

Compared with the de-extinction of entire species, which with modern technologies is outside our reach (Lin et al. 2022), molecular de-extinction targeting proteins and other functional parts of extinct genomes is within reach, as we presented a case study in application number one of this report (Wan et al. 2024). Even without the aid of artificial intelligence, molecular de-extinction remains a research field with great potential and significant repercussions in the fields of biology, medicine and biotechnology.

This targeted de-extinction approach, does not aim to bring entire species back to life, but to resurrect functional parts of ancient genomes (paleogenomes) such as proteins, enzymes or genes. Studying the structure and function of extinct proteins and genomes and comparing them with modern proteins already facilitates the understanding of biological functions and adaptations present in organisms now living on planet earth, an iconic example being the comparison of the genome of modern Homo sapiens with ancient hominins species. With the help of such analysis it has been possible to trace the origin of specific genetic characteristics of some human populations, such as the ability to withstand high altitudes of Tibetan populations, plausibly inherited from the Denisovans, an extinct species of hominins (Zeberg, Jakobsson, and Pääbo 2024).

In general, de-extinction of genes and proteins can be achieved in two ways: by cloning genes from paleogenomes (Wan et al. 2024), or by statistical reconstruction of the evolutionary process from existing molecular structures to go “back in time” and statistically reconstruct ancestral forms of genes and proteins (also called genetic resurrection) (Thornton 2004). In the latter case, these are statistical reconstructions that do not 100% reflect the original genomic sequences but are able to bring “back to life” functional genes and proteins with extinct functions and characteristics that no longer occur in nature and that first emerged on earth dozen

of millions of years ago (Thornton, Need, and Crews 2003; Yokoyama et al. 2015).

In both cases, either by cloning directly from the genome of extinct beings or by reconstruction through statistical inference, it is possible to trace functional forms of genes and proteins that can be tested and characterized in the laboratory and that cannot be associated with genetic resources currently existing on planet earth. In the face of the innovations and potential for the discovery of new molecular functions, there is a need, while the field of research is still young, to weigh its implications for benefit sharing (Torrance and De La Fuente-Nunez 2024).

4. Bacterial and viral therapies to fight cancer

The use of therapeutic bacteria and viruses to fight cancer is an emerging field of research with promising results (Yarahmadi et al. 2024).

In bacterial cancer therapy, modified strains of bacteria are exploited to attack tumors directly. These bacteria, in some cases genetically engineered, can preferentially proliferate in tumor microenvironments, where they find ideal conditions such as low oxygen concentration (hypoxia) and specific nutrients that promote their growth. Depending on their use, the bacteria can be used to kill the tumor through the release of toxins or molecules that interfere with tumor growth or to initiate inflammatory processes that activate a host immune response that can clarify the tumor.

Cancer virotherapy uses modified viruses, known as oncolytic viruses, to selectively infect and destroy cancer cells. Oncolytic viruses replicate inside cancer cells, leading to cell lysis and, at may the same time, stimulate an immune response against the tumor.

There are many bacterial and viral strains that are candidates for these applications (Bifidobacteria, Clostridium, Listeria monocytogenes, Salmonella typhimurium, Bacillus, Vaccinia viruses, Adenoviruses, Reoviruses, Herpesviruses, and Coxsackieviruses) and ongoing research tests these treatments in preclinical and clinical trials (Harimoto et al. 2022; Toso et al. 2002).

References

Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, et al. 2024. "Accurate structure prediction of biomolecular interactions with AlphaFold 3." *Nature* 630, no. 8016 (June 13, 2024): 493–500. ISSN: 0028-0836, 1476-4687, accessed August 16, 2024. https://doi.org/10.1038/s41586-024-07487-w. https://www.nature.com/articles/s41586-024-07487-w.

Baker, David, Susana Vazquez Torres, Melisa Benard Valle, Stephen Mackessy, Stefanie Menzies, Nicholas Casewell, Shirin Ahmadi, et al. 2024. *De novo designed proteins neutralize lethal snake venom toxins,* May 17, 2024. Accessed October 9, 2024. https://doi.org/10.21203/rs.3.rs-4402792/v1. https://www.researchsquare.com/article/rs-4402792/v1.

Benegas, Gonzalo, Sanjit Singh Batra, and Yun S. Song. 2023. "DNA language models are powerful predictors of genome-wide variant effects." *Proceedings of the National Academy of Sciences* 120, no. 44 (October 31, 2023): e2311219120. ISSN: 0027-8424, 1091-6490, accessed August 15, 2024. https://doi.org/10.1073/pnas.2311219120. https://pnas.org/doi/10.1073/pnas.2311219120.

Brahma, Neha, and S. Vimal. 2024. "Artificial intelligence in neuroimaging: Opportunities and ethical challenges." *Brain and Spine* 4:102919. ISSN: 27725294, accessed October 7, 2024. https://doi.org/10.1016/j.bas.2024.102919. https://linkinghub.elsevier.com/retrieve/pii/S2772529424001759.

Bursteinas, Borisas, Ramona Britto, Benoit Bely, Andrea Auchincloss, Catherine Rivoire, Nicole Redaschi, Claire O'Donovan, and Maria Jesus Martin. 2016. "Minimizing proteome redundancy in the UniProt Knowledgebase." *Database* 2016:baw139. ISSN: 1758-0463, accessed September 6, 2024. https://doi.org/10.1093/database/baw139. https://academic.oup.com/database/article-lookup/doi/10.1093/database/baw139.

Cantelli, Gaia, Alex Bateman, Cath Brooksbank, Anton I Petrov, Rahuman S Malik-Sheriff, Michele Ide-Smith, Henning Hermjakob, et al. 2022. "The European Bioinformatics Institute (EMBL-EBI) in 2021." *Nucleic Acids Research* 50 (D1 2022): D11–D19. ISSN: 0305-1048, 1362-4962, accessed August 12, 2024. https://doi.org/10.1093/nar/gkab1127. https://academic.oup.com/nar/article/50/D1/D11/6439668.

Chippaux, Jean-Philippe. 2017. "Snakebite envenomation turns again into a neglected tropical disease!" *Journal of Venomous Animals and Toxins including Tropical Diseases* 23, no. 1 (December): 38. ISSN: 1678-9199, accessed September 6, 2024. https://doi.org/10.1186/s40409-017-0127-6. http://jvat.biomedcentral.com/articles/10.1186/s40409-017-0127-6.

Chu, Yanyi, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. 2024. "A 5' UTR language model for decoding untranslated regions of mRNA and function predictions." *Nature Machine Intelligence* 6, no. 4 (April 5, 2024): 449–460. ISSN: 2522-5839, accessed August 16, 2024. https://doi.org/10.1038/s42256-024-00823-

9. https://www.nature.com/articles/s42256-024-00823-9.

D'Hondt, Kathleen, Tanja Kostic, Richard McDowell, Francois Eudes, Brajesh K. Singh, Sara Sarkar, Marios Markakis, Bettina Schelkle, Emmanuelle Maguin, and Angela Sessitsch. 2021. "Microbiome innovations for a sustainable future." *Nature Microbiology* 6, no. 2 (January 28, 2021): 138–142. ISSN: 2058-5276, accessed October 3, 2024. https://doi.org/10.1038/s41564-020-00857-w. https://www.nature.com/articles/s41564-020-00857-w.

Dolgin, Elie. 2009. "Human genomics: The genome finishers." *Nature* 462, no. 7275 (December 17, 2009): 843–845. ISSN: 0028-0836, 1476-4687, accessed August 12, 2024. https://doi.org/10.1038/462843a. https://www.nature.com/articles/462843a.

Fujihara, Kazuya, Mayuko Yamada Harada, Chika Horikawa, Midori Iwanaga, Hirofumi Tanaka, Hitoshi Nomura, Yasuharu Sui, et al. 2023. "Machine learning approach to predict body weight in adults." *Frontiers in Public Health* 11 (June 15, 2023): 1090146. ISSN: 2296-2565, accessed October 7, 2024. https://doi.org/10.3389/fpubh.2023.1090146. https://www.frontiersin.org/articles/10.3389/fpubh.2023.1090146/full.

Gianetto-Hill, Connor M., Sarah J. Vancuren, Brendan Daisley, Simone Renwick, Jacob Wilde, Kathleen Schroeter, Michelle C. Daigneault, and Emma Allen-Vercoe. 2023. "The Robogut: A Bioreactor Model of the Human Colon for Evaluation of Gut Microbial Community Ecology and Function." *Current Protocols* 3, no. 4 (April): e737. ISSN: 2691-1299, 2691-1299, accessed October 4, 2024. https://doi.org/10.1002/cpz1.737. https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cpz1.737.

Harimoto, Tetsuhiro, Jaeseung Hahn, Yu-Yu Chen, Jongwon Im, Joanna Zhang, Nicholas Hou, Fangda Li, et al. 2022. "A programmable encapsulation system improves delivery of therapeutic bacteria in mice." *Nature Biotechnology* 40, no. 8 (August): 1259–1269. ISSN: 1087-0156, 1546-1696, accessed October 4, 2024. https://doi.org/10.1038/s41587-022-01244-y. https://www.nature.com/articles/s41587-022-01244-y.

He, Tuo, Lichao Jiao, Alex C. Wiedenhoeft, and Yafang Yin. 2019. "Machine learning approaches outperform distance- and tree-based methods for DNA barcoding of Pterocarpus wood." *Planta* 249, no. 5 (May): 1617–1625. ISSN: 0032-0935, 1432-2048, accessed October 8, 2024. https://doi.org/10.1007/s00425-019-03116-3. http://link.springer.com/10.1007/s00425-019-03116-3.

Jiménez-Luna, José, Francesca Grisoni, and Gisbert Schneider. 2020. "Drug discovery with explainable artificial intelligence." *Nature Machine Intelligence* 2, no. 10 (October 13, 2020): 573–584. ISSN: 2522-5839, accessed August 19, 2024. https://doi.org/10.1038/s42256-020-00236-4. https://www.nature.com/articles/s42256-020-00236-4.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596, no. 7873 (August 26, 2021): 583–589. ISSN: 0028-0836, 1476-4687, accessed October 9, 2024. https://doi.org/10.1038/s41586-021-03819-2. https://www.nature.com/articles/s41586-021-03819-2.

Leeuwen, Pim T van, Stanley Brul, Jianbo Zhang, and Meike T Wortel. 2023. "Synthetic microbial communities (SynComs) of the human gut: design, assembly, and applications." *FEMS Microbiology Reviews* 47, no. 2 (March 10, 2023): fuad012. ISSN: 1574-6976, accessed October 4, 2024. https://doi.org/10.1093/femsre/fuad012. https://academic.oup.com/femsre/article/doi/10.1093/femsre/fuad012/7080139.

Lin, Jianqing, David Duchêne, Christian Carøe, Oliver Smith, Marta Maria Ciucani, Jonas Niemann, Douglas Richmond, et al. 2022. "Probing the genomic limits of de-extinction in the Christmas Island rat." *Current Biology* 32, no. 7 (April): 1650–1656.e3. ISSN: 09609822, accessed October 4, 2024. https://doi.org/10.1016/j.cub.2022.02.027. https://linkinghub.elsevier.com/retrieve/pii/S0960982222002494.

Lundberg, Scott M., and Su-In Lee. 2017. "A unified approach to interpreting model predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems,* 4768–4777. NIPS'17. Long Beach, California, USA: Curran Associates Inc. ISBN: 9781510860964.

Łysko, Andrzej, Agnieszka Popiela, Paweł Forczmański, Attila Molnár V., Balázs András Lukács, Zoltán Barta, Witold Maćków, and Grzegorz J. Wolski. 2022. "Comparison of discriminant methods and deep learning analysis in plant taxonomy: a case study of Elatine." *Scientific Reports* 12, no. 1 (November 28, 2022): 20450. ISSN: 2045-2322, accessed October 8, 2024. https://doi.org/10.1038/s41598-022-24660-1. https://www.nature.com/articles/s41598-022-24660-1.

Maasch, Jacqueline R.M.A., Marcelo D.T. Torres, Marcelo C.R. Melo, and Cesar De La Fuente-Nunez. 2023. "Molecular de-extinction of ancient antimicrobial peptides enabled by machine learning." *Cell Host & Microbe* 31, no. 8 (August): 1260–1274.e6. ISSN: 19313128, accessed October 9, 2024. https://doi.org/10.1016/j.chom.2023.07.001. https://linkinghub.elsevier.com/retrieve/pii/S1931312823002962.

Messeri, Lisa, and M. J. Crockett. 2024. "Artificial intelligence and illusions of understanding in scientific research." *Nature* 627, no. 8002 (March 7, 2024): 49–58. ISSN: 0028-0836, 1476-4687, accessed October 8, 2024. https://doi.org/10.1038/s41586-024-07146-0. https://www.nature.com/articles/s41586-024-07146-0.

Murray, Christopher J L, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, et al. 2022. "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis." *The Lancet* 399, no. 10325 (February): 629–655. ISSN: 01406736, accessed October 2, 2024. https://doi.org/10.1016/S0140-6736(21)

02724-0. https://linkinghub.elsevier.com/retrieve/pii/S0140673621027240.

Oleskowicz-Popiel, Piotr. 2018. "Designing Reactor Microbiomes for Chemical Production from Organic Waste." *Trends in Biotechnology* 36, no. 8 (August): 747–750. ISSN: 01677799, accessed October 4, 2024. https://doi.org/10.1016/j.tibtech.2018.01.002. https://linkinghub.elsevier.com/retrieve/pii/S0167779918300234.

Riza, Lala Septem, Muhammad Iqbal Zain, Ahmad Izzuddin, Yudi Prasetyo, Topik Hidayat, and Khyrina Airin Fariza Abu Samah. 2023. "Implementation of machine learning in DNA barcoding for determining the plant family taxonomy." *Heliyon* 9, no. 10 (October): e20161. ISSN: 24058440, accessed August 19, 2024. https://doi.org/10.1016/j.heliyon.2023.e20161. https://linkinghub.elsevier.com/retrieve/pii/S2405844023073693.

Saarela, Mirka, and Susanne Jauhiainen. 2021. "Comparison of feature importance measures as explanations for classification models." *SN Applied Sciences* 3, no. 2 (February): 272. ISSN: 2523-3963, 2523-3971, accessed October 8, 2024. https://doi.org/10.1007/s42452-021-04148-9. http://link.springer.com/10.1007/s42452-021-04148-9.

Sanabria, Melissa, Jonas Hirsch, Pierre M. Joubert, and Anna R. Poetsch. 2024. "DNA language model GROVER learns sequence context in the human genome." *Nature Machine Intelligence* 6, no. 8 (July 23, 2024): 911–923. ISSN: 2522-5839, accessed September 6, 2024. https://doi.org/10.1038/s42256-024-00872-0. https://www.nature.com/articles/s42256-024-00872-0.

Sarrazin-Gendron, Roman, Parham Ghasemloo Gheidari, Alexander Butyaev, Timothy Keding, Eddie Cai, Jiayue Zheng, Renata Mutalova, et al. 2024. "Improving microbial phylogeny with citizen science within a mass-market video game." *Nature Biotechnology* (April 15, 2024). ISSN: 1087-0156, 1546-1696, accessed August 19, 2024. https://doi.org/10.1038/s41587-024-02175-6. https://www.nature.com/articles/s41587-024-02175-6.

Shayanthan, Ambihai, Patricia Ann C. Ordoñez, and Ivan John Oresnik. 2022. "The Role of Synthetic Microbial Communities (SynCom) in Sustainable Agriculture." *Frontiers in Agronomy* 4 (June 30, 2022): 896307. ISSN: 2673-3218, accessed October 4, 2024. https://doi.org/10.3389/fagro.2022.896307. https://www.frontiersin.org/articles/10.3389/fagro.2022.896307/full.

Stokes, Jonathan M., Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, et al. 2020. "A Deep Learning Approach to Antibiotic Discovery." *Cell* 180, no. 4 (February): 688–702.e13. ISSN: 00928674, accessed August 19, 2024. https://doi.org/10.1016/j.cell.2020.01.021. https://linkinghub.elsevier.com/retrieve/pii/S0092867420301021.

The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Shadab Ahmad, Emanuele Alpi, et al. 2023. "UniProt: the Universal Protein Knowledgebase in 2023." *Nucleic Acids Research* 51 (D1 2023): D523–D531. ISSN: 0305-1048, 1362-4962, accessed August 22, 2024. https://doi.org/10.1093/nar/gkac1052.

https://academic.oup.com/nar/article/51/D1/D523/6835362.

Thornton, Joseph W. 2004. "Resurrecting ancient genes: experimental analysis of extinct molecules." *Nature Reviews Genetics* 5, no. 5 (May): 366–375. ISSN: 1471-0056, 1471-0064, accessed October 4, 2024. https://doi.org/10.1038/nrg1324. https://www.nature.com/articles/nrg1324.

Thornton, Joseph W., Eleanor Need, and David Crews. 2003. "Resurrecting the Ancestral Steroid Receptor: Ancient Origin of Estrogen Signaling." *Science* 301, no. 5640 (September 19, 2003): 1714–1717. ISSN: 0036-8075, 1095-9203, accessed October 4, 2024. https://doi.org/10.1126/science.1086185. https://www.science.org/doi/10.1126/science.1086185.

Torrance, Andrew W., and Cesar De La Fuente-Nunez. 2024. "The patentability and bioethics of molecular de-extinction." *Nature Biotechnology* 42, no. 8 (August): 1179–1180. ISSN: 1087-0156, 1546-1696, accessed October 4, 2024. https://doi.org/10.1038/s41587-024-02332-x. https://www.nature.com/articles/s41587-024-02332-x.

Toso, John F., Vee J. Gill, Patrick Hwu, Francesco M. Marincola, Nicholas P. Restifo, Douglas J. Schwartzentruber, Richard M. Sherry, et al. 2002. "Phase I Study of the Intravenous Administration of Attenuated *Salmonella typhimurium* to Patients With Metastatic Melanoma." *Journal of Clinical Oncology* 20, no. 1 (January 1, 2002): 142–152. ISSN: 0732-183X, 1527-7755, accessed October 4, 2024. https://doi.org/10.1200/JCO.2002.20.1.142. https://ascopubs.org/doi/10.1200/JCO.2002.20.1.142.

Varadi, Mihaly, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, et al. 2024. "AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences." *Nucleic Acids Research* 52 (D1 2024): D368–D375. ISSN: 0305-1048, 1362-4962, accessed August 22, 2024. https://doi.org/10.1093/nar/gkad1011. https://academic.oup.com/nar/article/52/D1/D368/7337620.

Wan, Fangping, Marcelo D. T. Torres, Jacqueline Peng, and Cesar De La Fuente-Nunez. 2024. "Deep-learning-enabled antibiotic discovery through molecular de-extinction." *Nature Biomedical Engineering* 8, no. 7 (June 11, 2024): 854–871. ISSN: 2157-846X, accessed October 4, 2024. https://doi.org/10.1038/s41551-024-01201-x. https://www.nature.com/articles/s41551-024-01201-x.

Watson, Joseph L., David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, et al. 2023. "De novo design of protein structure and function with RFdiffusion." *Nature* 620, no. 7976 (August 31, 2023): 1089–1100. ISSN: 0028-0836, 1476-4687, accessed August 16, 2024. https://doi.org/10.1038/s41586-023-06415-8. https://www.nature.com/articles/s41586-023-06415-8.

Winnifrith, Adam, Carlos Outeiral, and Brian L. Hie. 2024. "Generative artificial intelligence for de novo protein design." *Current Opinion in Structural Biology* 86 (June): 102794. ISSN: 0959440X, accessed October 8, 2024. https://doi.org/10.1016/j.sbi.2024.102794. https://

linkinghub . elsevier . com / retrieve / pii / S0959440X24000216.

Yarahmadi, Aref, Mitra Zare, Masoomeh Aghayari, Hamed Afkhami, and Gholam Ali Jafari. 2024. "Therapeutic bacteria and viruses to combat cancer: double-edged sword in cancer therapy: new insights for future." *Cell Communication and Signaling* 22, no. 1 (April 24, 2024): 239. ISSN: 1478-811X, accessed October 4, 2024. https : / / doi . org / 10 . 1186 / s12964 - 024 - 01622 - w. https : / / biosignaling . biomedcentral . com / articles / 10 . 1186 / s12964-024-01622-w.

Yokoyama, Shozo, Ahmet Altun, Huiyong Jia, Hui Yang, Takashi Koyama, Davide Faggionato, Yang Liu, and William T. Starmer. 2015. "Adaptive evolutionary paths from UV reception to sensing violet light by epistatic interactions." *Science Advances* 1, no. 8 (September 4, 2015): e1500162. ISSN: 2375-2548, accessed October 4, 2024. https://doi.org/10.1126/sciadv.1500162. https://www.science.org/doi/10.1126/sciadv.1500162.

Zeberg, Hugo, Mattias Jakobsson, and Svante Pääbo. 2024. "The genetic changes that shaped Neandertals, Denisovans, and modern humans." *Cell* 187, no. 5 (February): 1047–1058. ISSN: 00928674, accessed October 4, 2024. https://doi.org/10.1016/j.cell.2023.12.029. https://linkinghub.elsevier.com/retrieve/pii/S0092867423014034.