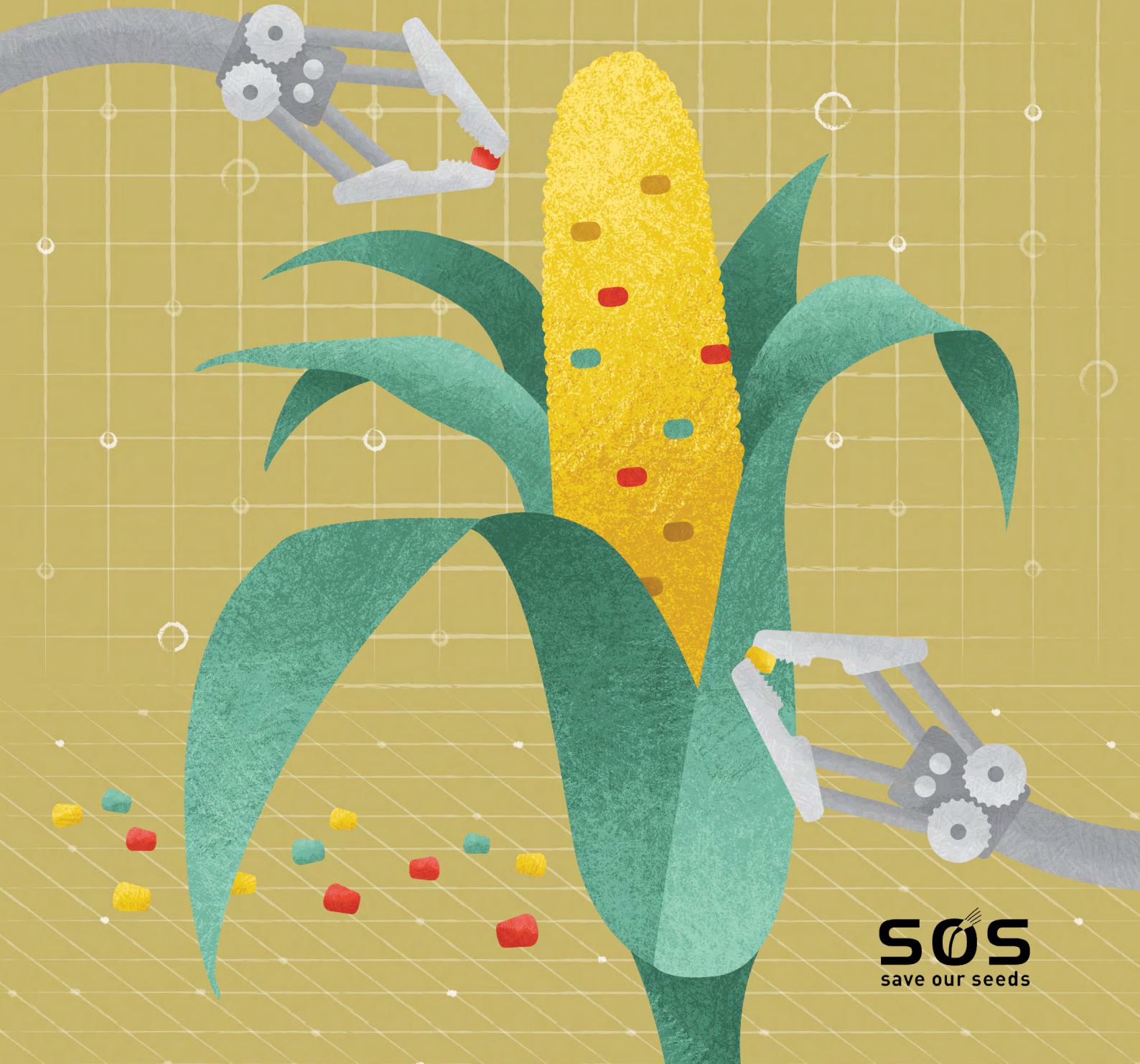
When chatbots breed new plant varieties

Generative Artificial Intelligence and New Genetic Engineering Techniques



SOS
save our seeds

Publisher:
Save Our Seeds / Foundation on Future Farming
Marienstr. 19-20
10117 Berlin
Phone: +49 30-28482326
Email: info@saveourseeds.org

Author:
Benno Vogel
www.bennovogel.eu

Cover and Design:
Beatriz Francisco
www.linkedin.com/in/beatriz-francisco/

Publication:
January 2025

SOS
save our seeds

Contents

Contents

List of abbreviations

AMP	Antimicrobial proteins
CRE	Cis-regulatory element
EU	European Union
GEIGS	Gene Editing Induced Gene Silencing
AI	Artificial intelligence
SME	Small and medium-sized enterprises
miRNA	microRNA
NGT	New genomic techniques
RNAi	RNA interference
scRNA-Seq	Single-cell RNA sequencing
siRNA	small interfering RNA
uORF	Upstream Open Reading Frame

1. Introduction

Increasingly, university laboratories, start-ups and tech giants such as Meta, Google and Microsoft are creating generative artificial intelligence (AI) tools for biotechnology and genetic engineering. They take the AI architectures of the diffusion and large language models used in chatbots like ChatGPT or image generators like DALL-E and train them in the 'languages' of biology – with protein and genome sequences. This results in tools that are radically changing the way genetic engineering is used to intervene in the genetic material of organisms. Equipped with improved descriptive capabilities, the new AI models make it possible to simulate the effects of genetic engineering on the computer. Thanks to their generative capabilities, the AI models can even design functional DNA and RNA sequences, as well as proteins, that evolution has not yet produced and that are, in technical jargon, 'new-to-nature'.

While generative AI is finding its way into genetically-engineered plant breeding, the EU is in the process of relaxing the regulation of genetically modified plants produced using newer methods of genetic engineering – methods known as 'new genomic techniques' (NGT). The European Commission presented a draft NGT law in July 2023. The draft divides those plants produced using genome editing that do not contain any genetic material from outside their breeding gene pool into two categories: genome-edited plants that contain up to 20 targeted changes in their genome form category 1 (NGT1 plants). Genome-edited plants with more than 20 targeted changes form category 2 (NGT2 plants). As the European Commission assumes that the risk profiles of NGT1 plants and conventionally bred plants are comparable, it proposes that NGT1 plants be exempt from the requirements of GMO legislation and be subject to the applicable legal provisions for conventionally bred plants. NGT2 plants, on the other hand, are proposed to remain within the regulatory area of GMO legislation.

To enable the EU Parliament and EU Council of Ministers to have a well-informed discussion on the proposed legislation covering the regulation of NGT plants, the EU Commission has provided policymakers with a range of documents: an impact assessment, case studies by the Joint Research Centre (JRC), work by the European Food Safety Authority (EFSA) and the results of a stakeholder consultation. However, what these documents and the ongoing political debate on the regulation of NGT plants do not take

into account is the convergence of genome editing and generative AI that is currently taking place in molecular plant breeding laboratories. What does this mean for the future-proof regulation of NGT plants? The question is all the more pressing given that the removal of precautionary measures such as risk assessment and traceability for NGT1 plants is under discussion.

The fact that the convergence of NGT and generative AI has so far played virtually no role in the discussion on the planned deregulation of NGT1 plants led the Save Our Seeds initiative to commission this paper. This paper aims to provide insight into the development of protein- and genome-based diffusion and large language models that could be used in plant genome editing, and to present the regulatory issues raised by the convergence of genome editing and generative AI in the production of NGT plants.

The information, gathered through a literature and internet search, is presented in the following order: firstly, Chapter 2 briefly presents the biological data that are available for training generative AI models. Chapter 3 describes the development of generative AI models trained on proteins and genomes that are being considered for use in NGT-based plant breeding. Subsequent chapters describe how research (Chapter 4), corporations (Chapter 5) and small and medium-sized enterprises (Chapter 6) use AI models when they modify the genetic makeup of plants. Finally, Chapter 7 highlights the questions and challenges that arise in the regulation of NGT1 plants.

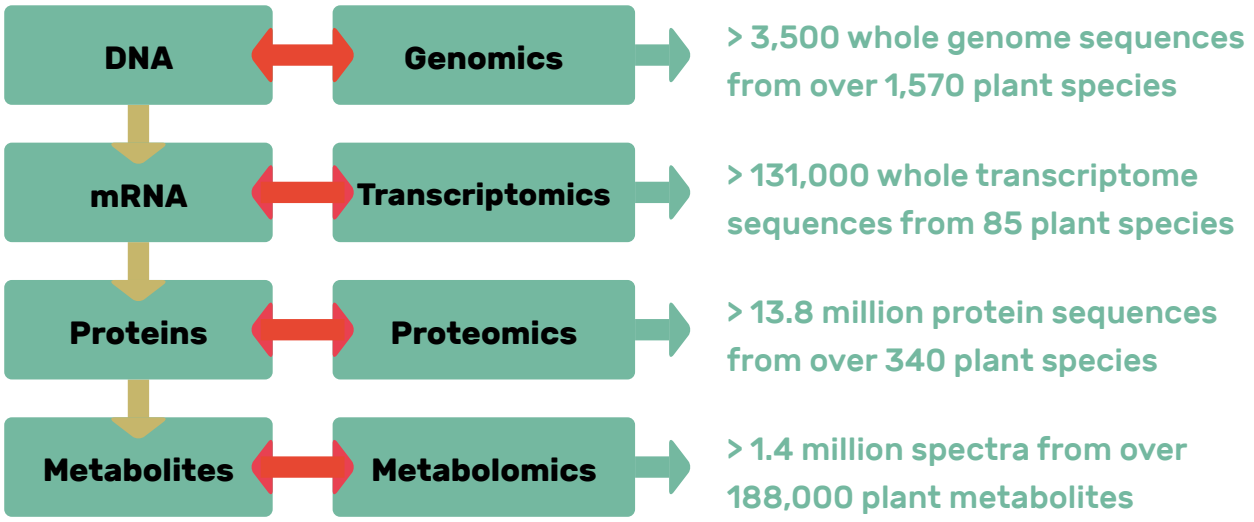
2. Big Data – the raw material for generative AI

The fields of genomics, transcriptomics, proteomics and metabolomics have uncovered an immense wealth of data on DNA and mRNA sequences, proteins and metabolites of plants in recent years (Figure 1).[1] For example, 13.82 million protein sequences from 342 plant species are included in the PlantMWpIDB database.[2] The PlantExp platform contains 131,400 whole-transcriptome sequences consisting of 572.4 terabases from 85 different plant species.[3] Meanwhile, the plant metabolome hub PMhub chemically describes 188,837 different plant metabolites.[4] These and other data form the raw material that allows the development of generative AI for NGT-based plant breeding to be

possible in the first place. Without big data, there would be no generative AI: modern algorithms need to be trained with huge data sets, and are generally more powerful the larger the data sets are.

Though the amount of data on plants is already extremely large today, both the quality of the techniques of data acquisition and the sheer amount of data gathered are set to develop exponentially in the coming years.

Figure 1: Omics techniques, their relationship to DNA, RNA, proteins and metabolites, and the data gathered from them.



2.1 Genomes, pangenomes and super-pangenomes

For the successful convergence of AI and NGT, plant genome data is essential in two ways: firstly, genomic data is indispensable for the use of CRISPR, and secondly, the data is the raw material for training AI tools.

Since the publication of the first genome sequence of a plant – *Arabidopsis thaliana* – in 2000, technical advances have simultaneously drastically increased the speed of sequencing and significantly reduced the costs associated with it. The first sequencing of the model plant *Arabidopsis* genome took ten years and 100 million US dollars to complete. The plant's genome can now be determined within a week for less than 1,000 US dollars.[5] Since 2000, the sequencing of entire genomes has become a relatively routine practice, and not only for model plants. Reflecting this, the amount of genomic data has increased dramatically in recent years: between 2021 and 2023 alone, twice as many

plant genomes were sequenced as in the previous twenty years combined.[6]

In June 2024, the N3 database at Zhejiang University in China contains 3,505 genomes sequenced from 1,575 plant species.[7] At the same point in time, the database of the US National Center for Biotechnology Information contains 4,604 plant genomes from 1,482 species.[8] With a number of projects, such as Ten Thousand Plant Genomes, African Orphan Crops, Genomics for Australian Plants and Darwin Tree of Life, currently underway in the sequencing of further plant genomes, these numbers will continue to grow.[9] The most ambitious goal, however, is being pursued by the Earth BioGenome Project, which aims to create a reference genome for all known animal, plant and fungal species in the world by 2030.[10]

With the advent of high-performance sequencing, the number of plant species for which a Pangenome is available has also increased in recent years. Aiming to represent the broad spectrum of genetic diversity within a species, Pangenomes are based on the genome sequences of several organisms within the same species, rather than just being based on the genome of one organism of that species.[11] They are intended to help in the identification of agronomically interesting or desirable alleles that can then be transferred to existing elite crop varieties using genome editing. So far, more than a thousand plant genomes are thought to have been compiled together for the construction of pangenomes. Published pangenomes are already available for several important crop plants such as rice, maize, wheat, soy, barley and potato.[12]

While pangenomes aim to include the entire set of genes within a species, the so-called 'super pangenome' extends this concept further by also including the genomes of closely-related species.[13] Initial projects are underway to compile super-pangenomes of rice,[14] maize,[15] tomato[16] and chickpea,[17] among others. Super-pangenomes are hoped to be used to transfer valuable traits from wild plants to elite varieties using genome editing techniques.

2.2 Omics techniques now available for single cells

Until recently, so-called 'omics' techniques could only be used at the level of cell clusters or entire plants. Although the data obtained using these techniques have significantly expanded our understanding of plant biology, the functions of rare cell types and low-concentration molecules remained largely obscured due to the 'dilution effect'. Now, new methods are making it

possible to obtain omics data at the level of a single plant cell. The addition of rare cell types and molecules that this offers adds a new depth to the data which was previously lost during the mass measurement of cell material.

The results obtained from single-cell omics expand the range of big data available for training AI tools.[18] Currently, the single-cell omics method used most commonly in plant biotechnology research utilises so-called 'scRNA-Seq', short for single-cell RNA sequencing techniques.[19] These techniques enable the analysis of RNA molecules of individual cells using high throughput sequencing, opening up in particular new possibilities for understanding gene expression.[20] Researchers at Nanjing University in China recently combed through existing scRNA-Seq studies on 17 plant species and created a database comprising of data from around 2.5 million cells.[21]

2.3 Google Maps for plants

Single-cell omics techniques also form an important foundation for the Plant Cell Atlas project.[22] The project, running since 2019, aims to generate extensive data on the structure and organisation of plant cells.[23] Experts from various disciplines, including genetics, cell biology, bioinformatics and imaging technology have been researching which types of plant cells exist and where and when certain molecules are present within them. The goal is to draw up a 'molecular map' which contains high-resolution temporo-spatial information about the DNA, RNA, proteins and metabolites found in plant cells – akin to a Google Maps for plants. The results of this endeavour are viewed as an important resource for plant and breeding research. The first Plant Cell Atlas symposium in 2021 was attended by nearly 500 leading experts from academia, industry and government agencies, including employees of BASF, Bayer, Syngenta and Google.[24]

The large amounts of scRNA-Seq and other single-cell omics data generated by the Plant Cell Atlas project will in turn be used to train AI tools for NGT-based breeding.[25]

3. Generative AI for NGT

Deep learning, artificial neural networks, language and diffusion models are currently undergoing rapid technological developments. These developments ensure that both the generative and descriptive performance of AI is constantly improving. As in many other areas, these advances are expected to bring profound changes both to the life sciences in general and specifically to plant breeding.

Deep learning is a general term for machine learning algorithms that consist of deep neural networks. Neural networks are computer programmes modelled on the way the human brain works, able to analyse huge amounts of unstructured data. Large language models and diffusion models, on the other hand, are variants of artificial neural networks. They form the AI architecture used in chatbots such as ChatGPT or image generators such as DALL-E and have been causing a furore worldwide since 2022.

The fact that diffusion and large language models are also causing a stir in the life sciences and NGT-based plant breeding can be attributed to two reasons: firstly, large language models can be trained with data from the scientific literature, functioning similarly to a novel form of research assistant. Secondly – and most importantly – rather than being trained with language texts, diffusion and large language models can also be trained using the vast amounts of DNA, RNA and protein data that have been collected through omics techniques in recent years. On one hand, the resulting AI tools are descriptive: like conventional deep learning algorithms, they can work with the 'languages' of biology and make predictions from them. On the other hand, they are also generative, able to generate functional DNA, RNA and amino acid sequences, including those that are new to nature.

3.1 Large language models: research assistants for NGT plant breeding

Google Scholar returns 498,000 results for a search with the keyword 'genome editing', 657,000 with 'synthetic biology', 1.9 million with 'plant breeding' and 2.2 million with 'genetic engineering'. AI tools that sift through these vast amounts of data and analyse them according to the wishes of researchers are intended to further facilitate and drive forward NGT-based breeding projects.

At least four such tools already exist: Open AI, the company behind ChatGPT, has developed a DNA programming tool that can help in the design of CRISPR projects and write programming code for DNA-related applications.[26] The company also offers the so-called 'Plant Breeding Optimiser'.[27] This is a chatbot designed to improve breeding programmes and – according to the company itself – can also help to predict breeding results in NGT-based projects. At the end of April 2024, Google presented CRISPR-GPT[28] – an AI assistant developed in collaboration with US universities to facilitate and automate the planning and running of CRISPR-based experiments. PLLaMa[29] has also been available since 2024. The tool is a joint product of researchers at universities in China, Sweden and the USA, who have trained Meta's LLaMa model with more than 1.5 million articles from the field of plant sciences. An international committee of agricultural engineers, plant scientists and breeders is currently testing how well PLLaMa can answer questions.

In the future, an increasing number of text-based models will be developed to analyse the constantly growing body of scientific literature and research results, making the conclusions and developments more easily accessible to plant breeders.[30]

3.2 Generative AI trained on proteins

Proteins control important biological processes and determine what happens in plants at a molecular level. AI tools that can analyse proteins, simulate their interactions or redesign their functions, constitute powerful tools for synthetic biology and genetic engineering in plants. There is extremely high interest in generative AI. Although the practice of training diffusion and large language models with protein data only arose in the early 2020's, the number of tools that have emerged is already becoming overwhelming. In addition to academic laboratories and numerous start-ups, tech companies have also become involved in their development. The two software giants Microsoft and Salesforce, the chip manufacturer NVIDIA, the internet companies Google and Meta as well as ByteDance, the company behind TikTok – all offer AI tools that, depending on the tool, can either understand protein sequences, generate them, or do both (Table 1).

The most famous protein tool is Google's Alphafold. Within a year, it mapped out the 3D structures of over 200 million proteins[31] – a feat that, without

powerful algorithms, would have taken researchers millions of years of work. According to the head of Google's AI department, Demis Hassabis, the database of protein structures created with Alphafold has already been visited by over one million researchers from 190 countries.[32] In June 2024, Google Scholar already contains over 23,000 publications citing the original article[33] on Alphafold published three years earlier in the scientific journal *Nature*.

Another protein tool named ESM-2, developed by Meta, has also been attracting a lot of attention. One factor drawing interest is its speed, which is said to be 60 times faster than Alphafold. Another reason is the Metagenomic Atlas, a database created by Meta's ESM-2 containing the structure of over 600 million microbial proteins.[34] Published in the journal *Science* in 2023, one year later in 2024 ESM-2 already has 1,200 citations in Google Scholar.

In addition to Alphafold and ESM-2, tools primarily used to model the structure of proteins, more and more generative AI tools have recently emerged that can be used in the design of proteins.[35,36,37] These include models developed by private AI laboratories such as Chroma[38] from Generate Biomedicines, EvoDiff[39] from Microsoft or ProGEN2[40] from Profluent and Salesforce, as well as university-developed models such as RFdiffusion,[41] ProtGPT2[42] and ForceGen.[43] The diversity and rapid growth of models produced have led researchers to speak of an 'explosion of possibilities.'[44] The tools offer not only new ways of redesigning natural proteins into versions with optimised or entirely novel functions - they also enable the de novo design of proteins previously unknown in nature.

Many of the design tools are so novel that the necessary experimental data is not yet available to evaluate the performance of their algorithms. However, it is already apparent that they open up a new design space that goes beyond natural limits. The number of mathematically possible protein variants is close to 10^{1300}. This unimaginably large number, exceeding the number of atoms in the universe many times over, clearly also includes an extremely large proportion of functionless amino acid sequences. However, it is also conceivable that this huge design space harbours functioning proteins that do not exist in nature. Such 'new-to-nature' constructs are of particular interest for researchers in the field of molecular plant breeding.

Table 1: AI tools (co)developed by tech companies for structural analysis and/or protein design.

AI Tool	Tech company	Training data	Year
Alphafold-2[45]	Google	> 170,000 protein structures	2021
Alphafold-3[46]	Google	*Not published*	2024
ESM-2[47]	Meta	65m protein sequences	2023
EvoDiff[48]	Microsoft	45m protein sequences	2023
LM-Design[49]	ByteDance	45m protein sequences	2023
OpenFold[50]	Microsoft/NVIDIA	> 170,000 protein structures	2024
ProGen[51]	Salesforce	280m protein sequences	2023
ProtTrans[52]	Google/NVIDIA	390b amino acids	2021
ProT-VAE[53]	NVIDIA	46m protein sequences	2023

3.3 Generative AI trained on genomes

The first large language models trained with huge amounts of DNA sequences and therefore able to simulate the 'language' of genomes came into existence in 2021. DNABERT, Nucleotide Transformer, GenSLM, megaDNA and EVO - these are the peculiar names of a selection of the two dozen or so models that currently exist (Table 2). While protein models are concerned with the coding sequences in the genome, genomic models also include the non-coding sequences and therefore also allow insight into the regulation of genes. This opens the door to entirely new possibilities for researchers.

To date, most genomic language models have been based on DNA sequences from humans and animals. However, as employees of Instadeep and BioNTech recently demonstrated, such models can also be used to analyse plant genomes.[54]

The existence of the first language models for RNA sequences should also not go without mention. Examples of

these are CodonBERT[55] from Sanofi or ERNIE-RNA[56] and scGPT[57] from Microsoft, all three of which have been trained with human RNA sequences. In the future, it is likely that large language models based on RNA sequences from plants will be developed. Models based on scRNA-seq data, such as scGPT, are seen as being particularly promising for the field of plant science and breeding.[58]

3.3.1 GPN, FloraBERT and AgroNT – the first language models for plant genomes

In June 2024, there are four models on Google Scholar that were specifically trained using plant DNA sequences. One of these is the Genomic Pre-trained Network (GPN). It originates from the AI laboratories of the University of California, where it was equipped with DNA data from *Arabidopsis* and seven other cruciferous plant species.[59] GPN can be used to predict how individual mutations in regulatory sequences will affect the plant.

FloraBERT is also specialised in regulatory sequences. The 2022 model is a product of Inari, a start-up that creates genome-edited plants (see 6.1).[60] The model's training data are promoter sequences stemming from the genetic material of 93 plant species and 25 different maize varieties. FloraBERT is designed to predict, in several different maize tissues, how changes in the promoter sequences affect gene activity.

The most powerful language model for plant genomes to date is the Agronomic Nucleotide Transformer, or AgroNT for short. It was developed as a collaboration between the AI forges of Google and Instadeep. The internet company and the AI company joined forces in 2022 to develop a computer model for the genome editing of plants that would enable the simulation and evaluation of individual changes in a desired region of the genome. The model, published at the end of 2023, was trained using 10 million genome sequences from 48 plant species.[61] To demonstrate its performance, the model was used to simulate more than 10 million mutations in the cassava genome and predict how each one would affect gene activity in the plant. As the developers highlighted, modelling the effects of so many mutations would be almost unachievable through experiments on plants and would be essentially impossible in nature.

The fourth language model for plant genomes is PlantCaduceus, a model trained on genetic data from 16 plant species, presented in the USA in July 2024 - shortly before the completion of this report.[62]

3.3.2 Will AI-designed genomes soon be possible?

It is currently difficult to predict the extent to which genomic language models will influence synthetic biology and genome editing. A number of articles on these models are currently only available on preprint servers such as bioRxiv and arXiv and are therefore not yet peer-reviewed. In addition, experimental data that could be used to assess the actual performance of the algorithms is often lacking.

However, regarding their use as tools for the functional annotation of genomes and as predictive models, it is already clear that the large language models are likely to outperform previous AI tools.[63] According to Instadeep, genomic models may also be suitable for modelling proteins and therefore be a good starting point for the construction of multimodal models for biology (see below).[64]

Large language models trained with microbial DNA sequences also demonstrate the potential that genomic AI tools could have. At the end of 2023, a researcher at Harvard University presented megaDNA - a model based on DNA data from bacteriophages.[65] As megaDNA can be used to generate new sequences up to 96 kilobases in length with the functional structure of phages, the model paves the way for the de novo design of entire phage genomes. The AI company Together AI and the privately funded Arc Institute are also discussing the design of new genomes. Together with university institutes, the two have developed EVO, a model based on 300 billion DNA bases from 80,000 bacterial genomes and millions of phage and plasmid DNA sequences.[66] EVO is not only capable of generating sequences for small molecules such as non-coding RNA, but can also code DNA sequences up to 650 kilobases in length. According to the developers, EVO has the potential to generate sequences on the scale of entire microbial genomes.

Table 2: Examples of generative AI tools trained on genomes.

AI tool	Company/University	Training data	Year
AgroNT[67]	Instadeep & Google	10 million sequences from geno-mes from 48 plant species	2023
DNABERT[68] (multi-species version)	Northwestern University	> 32 billion bases from genomes from 135 species (animals, fungi and bacteria)	2023

AI tool	Company/University	Training data	Year
EVO[69]	Together AI & Arc Institute	300 billion bases from over 80,000 bacteria and phage genomes	2023
FloraBERT[70]	Inari	Promoter sequences from 93 plant species and 25 maize varieties	2022
GenSLM[71]	NVIDIA & several unis	110 million prokaryotic gene sequences and 1.5 million SARS-CoV-2 genomes	2023
GPN[72]	University of California	Genome sequences from 8 plant species	2023
Nucleotide Transformer[73]	Instadeep & NVIDIA	Sequences from over 3000 human genomes and 850 genomes from animals, fungi and bacteria	2023
megaDNA[74]	Harvard University	> 99,000 phage genome sequences	2023
PlantCaduceus[75]	Cornell University	Genome sequences from 16 plant species	2024

3.4 Multimodal tools - on the way to supermodels

As powerful as text, protein and genome-based language models are, it is already becoming clear that they could soon be replaced by even more powerful models. The keyword here is multimodality. While previous language and diffusion models are still restricted to working with a single type of data, AI companies are now working on models that can process multiple types of data.

At the beginning of May 2024, Instadeep and BioNTech presented ChatNT, a multimodal AI tool aiming to bridge the gap for the first time between a Conversational Agent trained with text data and a model trained with biological data.[76] Although the newly developed chatbot is primarily aimed at medical research, applications in the plant biology sector are possible. According to Instadeep, ChatNT "signals a potential shift towards the creation of a truly universal, multimodal AI system for genomics".[77]

At the end of June, shortly before this paper was completed, Instadeep and BioNTech presented the first multimodal AI architecture for connecting DNA, RNA and protein data.[78]

3.4.1 CropGPT for Breeding 5.0

AI tools that can process text from scientific literature as well as genomics, proteomics, transcriptomics and metabolomics data could be especially interesting in plant breeding.[79] They are being presented as pioneers of a new type of molecular plant breeding, which researchers call 'Breeding 5.0' or 'data-driven genomic design breeding'.[80]

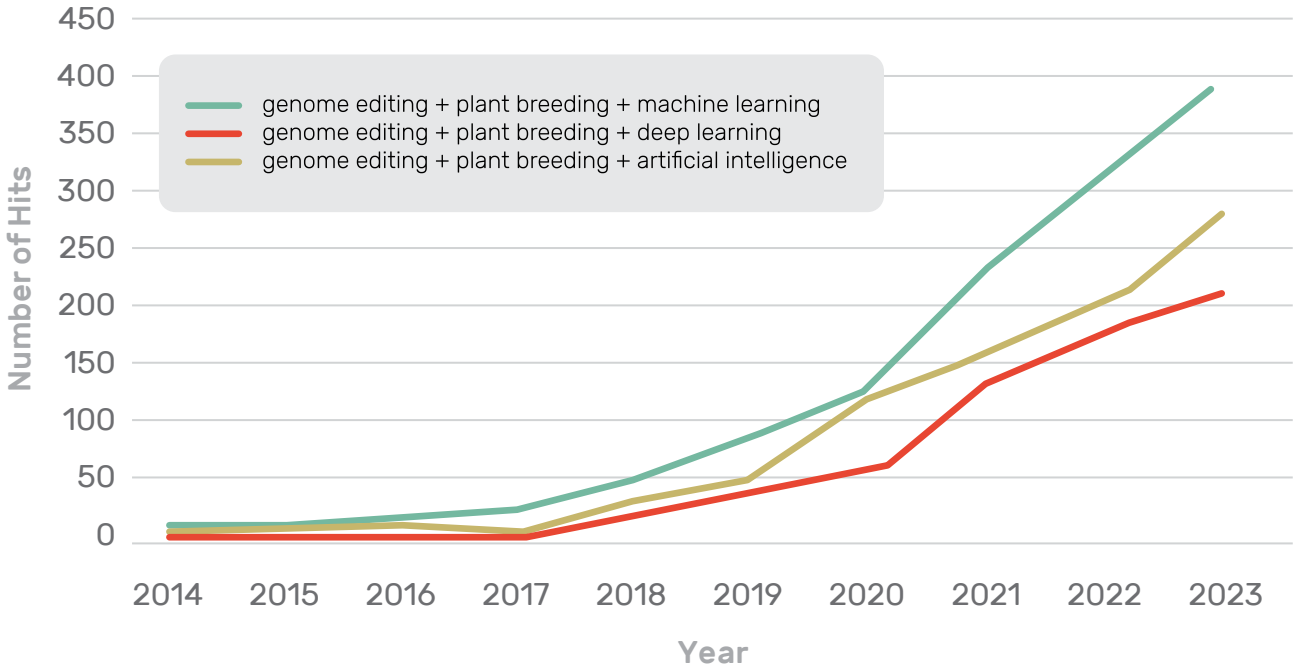
A universal, multimodal AI system for NGT-based plant breeding could be coming soon. In early 2024, researchers from several Chinese universities envisioned a global CropGPT project, published in the journal *Molecular Plant*.[81] In the publication they called for breeders, biologists, computer scientists and mathematicians worldwide to work together with biotech companies and breeding companies to develop a multimodal tool based on diverse omics data, with the aim of the project being to accelerate the development of AI-driven design breeding.

4. AI applications in NGT-based breeding research

AI is nothing new in NGT-based breeding research. If you type the keywords 'genome editing', 'plant breeding' and 'deep learning' into Google Scholar, you will find the first hits as early as 2017. Since then, interest in AI has been growing significantly. Between 2018 and 2023, the number of hits with the above keywords increased thirteenfold (see Figure 2). This trend is likely to continue, as generative AI tools are considered ideal for convergence with NGT.

As a brief review of the literature shows, AI models in NGT-oriented breeding research have so far mainly been used to analyse genomic data, identify regulatory elements in the genome and make genome editing more precise and efficient. There are still hardly any publications in the literature that report on specific applications of the modern AI tools described in Chapter 3. Most of these tools have only become available after 2022 and are therefore too new to have contributed results published in breeding research. An exception are AI tools such as Alphafold from Google, which can be used to predict protein structures. As in other areas of biological research, they are increasingly becoming part of the 'infrastructure' in NGT-based breeding research, and several projects for AI-supported redesign of traits for crops can be found in the literature.

Figure 2: Annual hits on Google Scholar with selected AI and NGT search terms between 2014 and 2023.



4.1 AI tools for efficient and precise genome editing

NGT-based plant research and development today relies primarily on the CRISPR-Cas method. Anyone who uses CRISPR to edit plant genomes must not only know the sequence of the target region beforehand, but also know which changes are to be made, and at which position, to produce the desired trait. In addition, in order to ensure the efficiency and precision of the experiments, an optimal sequence for the guide RNA that determines the location of the double-break must be selected. AI tools are now available for all these tasks:[82,83] algorithms help researchers to identify optimal targets for editing by analysing the genomic context, functional annotations and potential off-target sites. Other algorithms suggest optimal sequences for the guide RNA and which of the various CRISPR cutting enzymes might be the most suitable. These tools make genome editing with CRISPR more precise, efficient and successful.

Some examples of such tools are listed in Table 3.

Table 3: Examples of AI tools that increase efficiency and precision in the genetic modification of plants using CRISPR.

Tool	Year	Scope of application	Citations*
Plant-specific tool			
CRISPR-P[84]	2014	49 plant species	701
CRISPR-P 2.0[85]	2017	49 plant species	585
CRISPR-GE[86]	2017	> 40 plant species	326
CRISPR-Plant v2[87]	2019	7 plant species	74
Tools also suitable for plants			
CRISPOR[88]	2018	> 100 species	1413
CHOPCHOP[89]	2014	> 100 species	1292

* Number of citations on Google Scholar on 29.6.2024

4.2 Regulation rather than knockout: generating quantitative trait variations

By far the most common form of intervention in the genome of plants using NGT involves switching off genes. The result is hundreds of plants with loss-of-function mutations. However, such loss-of-function mutations are often of limited value in breeding. Especially when it comes to generating quantitative traits that are influenced by several genes, so-called 'knockouts' reach their limits.[90]

So far, researchers have lacked tools to generate variation in quantitative traits. This could now change. This new trend is called quantitative trait engineering:[91] researchers no longer knock out genes, but instead control their expression by specifically modifying regulatory network sequences. Controlling gene expression should make it possible to influence complex quantitative traits.[92,93,94]

Quantitative trait engineering is to be achieved with CRISPR-based base and prime editors. These tools can be used to precisely generate mutations at regulatory elements in the genome that lead to the desired level of gene

expression. Researchers intend to target cis-regulatory elements (CRE) such as promoters, enhancers or silencers that control transcription, as well as upstream open reading frames (uORF) that regulate translation.

Quantitative trait engineering is made possible largely thanks to AI. One of the tools available is iCREPCP, a deep learning-based platform developed at Huazhong Agricultural University in China.[95] It is designed to find promoter sequences in plant genomes and make them accessible for genome editing. A second example, CAPE, developed by researchers from several Chinese universities, combines multiplex genome editing with an algorithm that predicts how edits in promoter sequences will affect gene activity.[96] There are also a number of tools used for identifying and editing uORF that can be used in plants, such as uORFSCAN, uORFlight and PsORF.[97]

AI tools are also being developed that can be used to generate CRE and uORF that are new-to-nature.[98,99] Outside the field of plant biotechnology, a number of such AI tools already exist.[100,101,102,103] In June 2024 the PhytoExpr model was presented, a model created to design CRE for plants. Researchers at the National Maize Improvement Center in China have developed two algorithms for PhytoExpr: one for redesigning natural CREs for genome editing and one for designing artificial CREs for synthetic biology in plants.[104]

4.3 NGT breeding with protein redesign

AI tools for predicting protein structures, such as Alphafold from Google or ESM-2 from Meta, are considered particularly promising in NGT-based breeding research because they expand the possibilities for the development of 'designer' plants.[105,106] Although the tools have only recently become available, a number of publications can already be found in the literature in which researchers report how they intend to use the tools to create NGT plants.

In a recent study, for example, researchers used AlphaFold to simulate how the protease Pip1 in tomatoes interacts with the protease-inhibiting protein EpiC2B of the pathogenic fungus *Phytophthora infestans*.[107] They found that two amino acids in Pip1 need to be changed to make the protease resistant to inhibition by EpiC2B. The aim is now to use CRISPR to edit the Pip1 gene accordingly and increase the resistance of the tomato to disease.

In another study, researchers recently used Alphafold to redesign patatin,[108] a protein occurring naturally in potatoes. A new version of the protein was generated that, according to the AI, should improve the viscosity and nutritional properties of dough made from potato flour. Using CRISPR-based prime editors, researchers aim to create the AI-generated version of patatin in the genome of potatoes.

Further examples can be found in the literature: in maize, researchers are planning to use AI-guided protein design and genome editing to modify the architecture of plants so that they can grow closer together in the field.[109] in wheat, the baking quality is to be optimised by modelling the structure of storage proteins.[110] The development of low-allergen plants is also a goal: by analysing the structure and function of allergenic proteins using AI, they want to use computers to determine which changes can reduce allergenicity while maintaining the nutritional value of the plants.[111] Also a focus of research are protein kinases and phosphatases, enzymes that significantly influence plant growth and their interactions with the environment and pathogens. Redesigning them with the help of AI could enable the development of edited plants that produce higher yields and are more resistant to pathogens.[112] Proteins that transport sugars are also considered possible candidates for optimisation, using AI tools such as Alphafold and genome editing to increase disease resistance.[113] In addition, thanks to newly designed proteins, plants are being designed that have higher levels of photosynthetic activity[114,115,116] or fix more carbon from the soils.[117] Finally, NLR proteins and thus the immune system of plants are also targets for research.[118,119] NLR proteins function like sentinels, with different protein variants capable of recognising different pathogens and triggering an alarm when they are attacked, causing plants to activate their defence systems. Using tools such as Alphafold or ESM-2, researchers aim to create new NLR variants that enable the plant to recognise pathogens that it previously overlooked. To do this, they plan to use the computer-based tools to determine the necessary changes in amino acid sequences required to expand the specificity of an NLR protein, before recreating them in the plant's genome using genome editing.

4.4 SynEpi and epigenome editing

The agronomic traits of crops are often not only regulated genetically, but also epigenetically. Variability in the epigenome has so far hardly played a role in molecular breeding. However, with the use of AI and CRISPR-based tools, researchers hope to change this. In recent years, several algorithms have been developed that can be used to identify epialleles and predict changes in plant epigenomes.[120,121,122] Alongside this, researchers have also used CRISPR to develop novel tools that can be used to specifically generate epialleles in plant genomes.[123,124] So far, epigenome editing has been largely limited to model plants. However, researchers at the Jiangsu Co-Innovation Center in China believe that, with the help of AI, epigenome editing could in the future develop into a widely applicable and effective method of plant breeding.[125]

In 2022, researchers from the Chinese Academy of Agricultural Sciences presented a new breeding strategy which relies on AI-generated predictions and epigenomic editing tools. The strategy is called 'Synthetic Epigenetics', or SynEpi for short. It follows engineering principles and aims to alter or completely redesign the epigenetic systems of plants. The aim is to develop varieties that react in specific and predetermined ways when exposed to exogenous or endogenous triggers.[126]

4.5 Automation

AI is viewed as being a driving force behind the automation of biotechnology and genetic engineering, and can also ensure that processes in NGT-based plant breeding run more autonomously. While the first autonomous laboratories for research on genetically-modified microorganisms already exist, the first plant biotechnology automation projects are also beginning. At the end of May 2024, researchers at the University of Illinois presented FAST-PB – a fast, automated and scalable high-throughput pipeline for plant bioengineering.[127] It is now possible to automate the cloning of genes and genome editing of protoplasts and callus cells in the laboratory. The NGT company Cibus has also automated its workflows and now aims to use this to industrialise the genome editing of oilseed rape.[128] Syngenta, in turn, has automated genome editing and transgene expression studies in maize and soybeans in order to accelerate the development of new varieties.[129,130]

5. AI, NGT and corporations

5.1 AI applications in seed companies

The large seed companies Bayer, BASF, Corteva and Syngenta have been collecting large amounts of omics data for many years. Using these data, they train algorithms for the selecting of genetic combinations in plants from conventional breeding techniques. The companies usually keep the specific tools with these algorithms under lock and key.[131] Therefore, little is known about how the companies use AI models for their NGT-based breeding programmes. What is almost certain however, is that they are using such tools.[132]

The company Corteva is known to have its own generative AI tool for NGT-based breeding. The company used Google's BigBird to develop it, a language model that can process DNA data. To use BigBird in its breeding programmes, Corteva supplied the AI with DNA data from 14 crop species, including canola, rice, corn, soy, wheat and barley. The result is a computer-based prediction tool that can determine how individual mutations in regulatory DNA sequences will affect gene activity.[133]

To supplement their AI expertise, these companies are also cooperating with other firms. For example, in June 2024 Syngenta announced that it uses the generative tool AgroNT, developed by Instadeep and Google (see 3.3.1). Together with Instadeep, the agricultural company now wants to develop AI-developed traits for corn and soy.[134] Syngenta is also working with Biographica, a startup founded in 2024 that develops state-of-the-art AI techniques to identify "high-value targets for gene editing" in crops.[135]

BASF and Corteva have both started collaborations with Tropic Biosciences, which has proprietary AI, working to develop genome-edited disease-resistant plants (see 6.2).[136]

Similarly, Bayer is using Evogene's AI platform to identify sequences in the genome of corn that could be edited to engineer disease-resistance in plants.[137] In addition, through its impact investment unit Leaps by Bayer, Bayer is supporting the startups Ukko and Amfora, both of which use the combined power of AI and CRISPR to develop new plant varieties.[138]

5.2 AI for NGT offered by tech companies

As providers of generative AI models, tech companies also play a role in R&D projects on NGT plants. Although the tools for protein analysis and design offered by Meta, NVIDIA, Google, Microsoft, Salesforce and ByteDance are not designed and manufactured specifically for genetic engineering-based plant breeding, they can also be used for this purpose.

Using Alphafold, Google is not only active in the field of protein design, but also develops tools specifically for plant genetic engineering. These include the large language model AgroNT, developed jointly with Instadeep, which is used by Syngenta and other companies (see 3.3.1 and 5.1). As early as 2021, Google's Moonshot Factory X already filed a patent for an AI model designed to help discover interesting genes in plants and make recommendations on which genome edits will produce a desired trait.[139]

6. AI applications in small and medium-sized enterprises

There are a number of small and medium-sized enterprises (SMEs) worldwide whose business models are based entirely or partially on the convergence of AI and NGT (Table 4). The strategies for utilising this convergence vary. There are companies such as Traitseq, Evogene, Instadeep, McClintock, Biographica and Computomics, which develop AI tools for NGT-based plant breeding and offer them to third parties. Companies such as Arzeda or Gingko Bioworks create traits for plant breeding companies using their proprietary AI. Ohalo, Amfora, Finally Foods, Plastomics and Hudson River Biotechnology, on the other hand, are SMEs using third-party AI tools to develop their NGT plants. And finally, there are the companies that have proprietary AI tools, usually developed with their own specific uses in mind, to produce their plant varieties. These companies include Inari, NeoCrop, genXtraits, Phytoform, Plantae Bioscience, TreeCo and Tropic Biosciences.

Companies in the latter group use their tools primarily for predictive modelling and hope to be able to develop NGT plants faster and more cost-effectively. According to the scientific journal *Nature Biotechnology,* they have the potential to break the dominance of agricultural companies

in the commercialisation of genetically modified plants.[140] The fact that the companies could have a chance of competing with the seed giants in NGT plants is reflected in the money they receive from investment companies.

According to data from the company databases Tracxn,[141] Crunchbase[142] and PitchBook,[143] more than 900 million euros in venture capital has been channelled into SMEs combining AI with NGT since 2016.

6.1 CRISPR AI start-ups for gene regulation

So far, Inari has received the most investment money. Since its foundation in 2016, the US company has received around 530 million euros.[144] Today, the 'SEEDesign Company' not only has a proprietary base editor and a licence for multiplex genome editing of promoters, but also has FloraBERT (see 3.3.1), a generative AI tool that can be used to predict how mutations in promoters will affect the characteristics of a plant. Equipped with these tools, Inari aims to influence the gene activity in maize, soya and wheat in a way that leads to the plants producing a 10 to 20 per cent higher yield. Inari's first plants - soya[145] and maize[146] with increased yield potential as well as a shorter variety of maize[147] - have already been given the green light for cultivation by the relevant authorities in the USA. In 2025, the company plans to carry out field trials with edited, high-yield wheat at several locations in Australia.[148] Inari is also carrying out field tests of a short-growing maize in Belgium.[149]

Phytoform is also focussing on promoters and the regulation of gene activity. The start-up, based between London and Boston, also has developed its own AI tool, CRE.AI.TIVE.[150] According to its own advertising, its algorithm is designed to determine the minimal changes in promoter sequences that can be used to achieve maximum effects in crops, thus "enabling unprecedented control over gene expression".[151]

Table 4: A selection of small and medium-sized companies offering AI tools for NGT-based breeding and/or using AI tools in NGT-based breeding.

Company	Country	Year*	Use of artificial intelligence
Amfora	US	2016	Uses McClintock's algorithm to create ultra-high protein peas and soybeans using NGT.
Arzeda	US	2008	Develops new traits for plants using AI protein design.
BellaGen	CN	2020	Uses the DNA-cutting enzyme Cas-SF01, created using AI-driven protein design, for genome editing.
Benson Hill	US	2012	Uses proprietary AI system *CropOS* to identify gene sequences conferring interesting traits.
Biographica	UK	2024	Offers AI tools for NGT-based plant breeding.
Computomics	DE	2012	Offers AccelATrait for identifying editing targets.
Evogene	IL	2002	Offers GeneRator for identifying candidate genes.
Finally Foods	IL	2024	Uses Evogene's GeneRator for molecular farming plants.
genXtraits	US	2022	Uses proprietary algorithm to identify DNA segments for editing that act as 'dimmer switches'.
Ginkgo Bioworks	US	2008	Designs new proteins for breeding with proprietary tool Owl.
Inari	US	2016	Uses generative AI FloraBERT for genome editing.
Instadeep	UK	2014	Offers the generative model AgroNT (developed with Google) for genome editing.
Hudson River Biotechnology	NL	2015	Uses AccelATrait from Computomics to identify gene loci for genome editing.
McClintock	US	2022	Offers AI tools for NGT-based plant breeding.
NeoCrop	CL	2020	Uses proprietary AI prediction model for genome editing.
Ohalo	US	2019	Uses AI from Google in its NGT-based breeding work.
NRGene	IL	2010	Offers the AI tool GO-GENOME for genome editing.

Company	Country	Year*	Use of artificial intelligence
Phytoform Labs	US	2017	Uses proprietary AI tool CRE.AI.TIVE for genome editing.
Plantae Bioscience	IL	2020	Uses AI-driven protein design for NGT-based breeding.
Plastomics	US	2017	Transforms soybean chloroplast genome with genes discovered using Evogene's GeneRator.
Qi Biodesign	CN	2021	Created a base editor using Google's Alphafold.
Traitseq	US	2023	Offers AI prediction models for genome editing.
TreeCo	US	2019	Uses a prediction model for genome editing of trees.
Tropic Bioscience	UK	2016	Uses proprietary GEiGS bio-compute tool to discover and mutate non-coding RNA genes.
Ukko	US	2016	Uses proprietary AI platform to create novel gluten tolerated in coeliac disease for wheat breeding.
Viridian Seeds	IE	2021	Uses AI for genome editing of legumes.
Wild Bioscience	UK	2021	Uses proprietary AI for gene identification in wild plants.

*Founding year of company

GenXtraits is another company focussing on the fine-tuning of gene activity. Founded in the USA in 2022, it claims to have a portfolio of intellectual property focussing on key regulatory elements in plant genomes.[152] Unlike Inari and Phytoform, however, genXtraits does not work with promoters but with uORFs. In order to identify these elements, which genXtraits calls 'dimmer switches', in the genetic material of plants, the company has developed a specialised AI tool.[153] It intends to alter the activity of the uORFs discovered with this tool using NGT.

6.2 Plant breeding with AI & RNAi & CRISPR

GEIGS, which stands for 'Gene Editing Induced Gene Silencing', is the name of an unconventional method in NGT-based plant breeding. The patent for GEIGS belongs to Tropic Biosciences.[154] The British start-up combines genome editing with RNA interference (RNAi): with the GEIGS platform, genes coding for siRNA or miRNA can be edited in such a way that the silencing functions of the RNAs are directed towards new targets - such as genes from insects and fungi or even the plant's own genes. For this purpose, Tropic uses the company's own AI, GEIGS-BioCompute. The tool analyses the genomic data of a plant and uses it to predict where in the genome the smallest changes need to be made in RNAi genes to achieve the desired trait. The ability to redirect silencing functions using AI-driven genome editing has been met with great interest. Tropic has joined forces with BASF and Corteva as cooperation partners (see 5.1) and has received over 70 million euros in venture capital since its foundation in 2016.[155]

6.3 AI from Google and 'boosted breeding'

Ohalo is similarly popular with investment companies. The US start-up, founded in 2019, is said to have already raised around 100 million euros towards the breeding of new plant varieties using NGT and AI from Google.[156,157] It is not known exactly which AI tools Ohalo uses: the company talks about predictive models used to find out which crosses, out of hundreds of thousands or millions of possible combinations, lead to plant varieties with the desired characteristics. Ohalo has already received the green light for the cultivation of two varieties of potato plant in the USA: one called RedVin, which can be cold-stored without it sweetening,[158] and the other for a potato with more beta-carotene in the tuber.[159] At the end of May 2024, the company presented its 'boosted breeding' technology, which aims to "accelerate evolution to unlock nature's potential". In the patent-pending technology,[160] Ohalo uses CRISPR-based ribonucleoproteins to cause the germ cells of plants to retain their genome in its entirety rather than halving it as usually happens. This results in so-called 'clonal germ cells', which have a double set of chromosomes. When two of these germ cells fuse together, the result is offspring that have 100 per cent of the genes of each their parent

plants instead of the normal 50 per cent from each. According to Ohalo, the production of such polyploid plants offers the possibility of "combining characteristics in boosted plants that, through conventional breeding, would either take thousands of years to combine, or would not combine at all".

6.4 Simulation of over 69,000 editing strategies

The results from TreeCo show why AI-based prediction models are of so much interest to companies. The US company is working on genome-edited poplar trees that produce less lignin and should therefore make paper production easier. To this end, it has developed an AI tool based on decades of forestry biotechnology studies. The tool can be used to predict how changes in the 21 genes involved in lignin synthesis will affect the wood composition, growth rate and other traits of the trees.[161,162] As the tool shows, more than 69,000 editing strategies could be considered for editing the 21 genes. Using practical experiments to determine the best strategies would therefore be too time-consuming - the AI tool, on the other hand, searches for the best strategies using a computer simulation. The company decided to use multiplex genome editing to experimentally generate only the seven most promising combinations of gene edits in the poplar genome.[163] In the greenhouse, some of the poplars genome-edited in this way contained up to 49 per cent less lignin.

6.5 Searching through the genomes of wild plants

Using AI to tap into the genetic diversity of wild plants for NGT-based plant breeding - that is the goal of Wild Bioscience. Founded in England in 2021, the company has an AI tool that it uses to search sequenced genomes of wild plants for variations in genes that it considers interesting for breeding climate-robust varieties. Wild Bioscience then transfers these variants by making "small changes to the genetic makeup of crop plants". By widening the search to the genomes of wild plants, the start-up has access to a new range of genetic source material that is much wider than previously available.[164]

6.6 Protein design for plant breeding

In addition to the companies mentioned above that use genome-based AI tools, there are also a number of companies in the field of NGT plant breeding working with protein design tools.

One of these companies is Arzeda. It calls itself 'The Protein Design Company' and uses generative AI to develop proteins for everything from enzymes for industry to traits for plant breeding. The company's projects include, for example, designing an improved version of Rubisco,[165] the protein that plants use to fix C02 from air and thus to draw carbon into the food chain. Arzeda also has its sights set on a plant enzyme that can break down a widely used herbicide.[166] According to Forbes, the manufacturer tests 10,000 designer proteins per week,[167] through which it claims to be able to "go beyond what nature has given us".[168]

Another company claiming to be able to exceed the limits of natural evolution is Gingko Bioworks. Like Arzeda, the Synbio company offers the AI-controlled development of proteins that can be used as traits for crop plants. Its generative tool for this is called Owl.[169]Gingko aims to use it to customise the activity and specificity of proteins.

"We accelerate nature's evolution beyond what was imaginable before, dramatically shortening the path to healthier, more sustainable plant-centric food", writes Plantae Bioscience, a company founded in Israel in 2020, on its website.[170] To achieve this acceleration, it uses AI-supported protein design and genome editing: first, it uses Google's Alphafold and the AI tool FuncLib to design new variants of existing plant proteins on the computer. It then engineers the amino acid sequences determined in this way into the genetic material of the plants through genome editing. With this AI-CRISPR combination, Plantae Bioscience aims to develop plants suitable for vertical farming that are smaller, grow faster, flower synchronously and thrive despite low light conditions.

Another company working with protein design tools and CRISPR is Ukko.[171] Supported by funding from Leaps by Bayer, the company is pursuing the goal of redesigning plant proteins that can trigger food intolerances.[172] Ukko's AI-supported platform aims to make it possible to modify pathogenic proteins so that they become tolerable without losing their other properties. Ukko plans to use genome editing to implement the changes in protein sequence deemed necessary by the platform into the genetic material of the plants.

6.7 Start-ups with self-made NGT tools

Companies are not only working with AI-driven protein engineering on new traits in plants, but also on new tools for the genome editing of plants. BellaGen, for example, claims to be the first company in China to carry out genome editing in plants on an industrial scale. The company recently started using the two tools hfCas12Max** and Cas-SF0, which it developed itself with the help of AI. Both tools are variants of Cas12 - a class of nucleases that are interesting due to their small size but are less efficient in editing than other classes of Cas nucleases. Thanks to Google's Alphafold and structure-guided protein engineering, BellaGen has turned Cas12 enzymes into efficient tools.[173,174] A first application took place in soya: Cas-SF01 was used to edit soya plants so that they develop larger seeds[175] and to engineer varieties to be resistant to the herbicide flucarbazone.[176]

Qi Biodesign is another Chinese start-up that has developed a new NGT tool using AI. Together with several university research institutes, the company searched the InterPro protein database for proteins that have similar sequences to deaminases. These are enzymes that can be fused with Cas nucleases to form a base editor. Qi Biodesign then used Alphafold to select the proteins considered to have the most promising structures from the proteins found. The result was a base editor with which, for the first time, C-G base pairs can be converted into T-A base pairs in the genetic material of soya.[177] Previously, soya was one of the plant species in which this edit, for unknown reasons, was not possible.

6.8 First start-ups with robot-compatible plants

It should not go without mentioning here that advances in AI can also influence the goals of genetic plant breeding. For example, to simplify the digitalisation and automation of agriculture, there are proposals to genetically modify plants so that they send out signals that can be interpreted by AI-controlled agricultural machinery such as robots or drones.[178,179]

Innerplant is one of the companies already developing such 'robot-ready crops'.[180] The company equips soya, maize, tomatoes and cotton with genes coding for fluorescent substances, with which the plants can signal when they are under attack from pests. In partnership with John Deere and Syngenta, Innerplant aims to develop a system that combines plants, equipment and

resources as follows: satellites recognise the stress signals from Innerplant's plants and guide tractors, equipped with fluorescence detectors by John Deere, to the affected fields, where they spray Syngenta pesticides in a targeted manner.[181] In 2023, Innerplant was given the green light for three plants in the USA: for Innersoy,[182] which emits signals in the event of pathogen infestation, and for a soya[183] and a tomato,[184] which produce fluorescent substances needed to calibrate the remote sensing devices. The US company Insignum AgTech is also developing signalling plants. Unlike Innerplant, however, it does not use genes from other species. Instead, it regroups genes that are already present in the genome, so that plants react to an external trigger with colour changes. For example, Insignum has developed a maize that produces purple anthocyanin at the site of infection when attacked by pathogens - a dye that can be recognised by AI-controlled agricultural machinery. The robot-ready maize was released for cultivation in the USA at the end of 2023.[185]

7. Generative AI and regulation of NGT1 plants

In July 2023, the European Commission submitted a proposal to the European Parliament and the Council of the European Union to deregulate genetically modified plants that are produced by targeted mutagenesis, cisgenesis, or a combination of the two techniques and do not contain any genetic material from outside the breed's gene pool.[186] The proposal distinguishes between two categories of genetically modified plants (GMPs), depending on the extent of the genetic modifications: NGT1 plants, which contain up to 20 targeted changes in their genome (see section 7.3), and NGT2 plants, which have more than 20 changes. The EU Commission assumes that the risk profiles of NGT1 plants and conventionally bred plants are comparable and proposes that NGT1 plants be exempted from the requirements of GMO legislation and be subject to the same regulations that cover conventionally bred plants. NGT2 plants, on the other hand, should remain within the regulatory scope of GMO legislation.

To enable the EU Parliament and Council to have a well-informed discussion, the EU Commission has published an impact assessment, case studies from the Joint Research Centre (JRC), work by the European Food Safety Authority (EFSA) and the results of a stakeholder survey. However, not explored in these documents, and therefore also not

considered in the ongoing political debate on the regulation of NGT plants, is the convergence of NGT and generative AI, as it is currently taking place in the NGT laboratories of the world. The convergence of these technologies only gained momentum after most of the EU Commission's NGT activities had been completed or were nearing completion.

As this convergence is thought to have great potential to change NGT-based breeding, it is all the more pressing to proactively discuss and clarify which regulatory issues and safety concerns it brings with it, before a new law is passed. This is especially important due to the proposal that precautionary measures such as risk assessment and traceability be waived for NGT1 plants.

In the following sections, we will first list some general aspects that should be considered in a proactive regulatory discussion. We will then use a scenario to present and explore regulatory questions arising from the convergence of NGT and generative AI in the production of NGT plants. After that, we will focus on NGT1 plants: first, we will present the new design space that would be available for the AI-controlled production of NGT1 plants under the proposed legislation. After that, exploring the aspects of risk assessment, traceability and labelling, we will present why the convergence of NGT and generative AI should be considered in the regulatory debate.

7.1 General regulatory aspects

The following are some general aspects to be considered when planning the governance of the convergence of NGT and generative AI. They largely reflect aspects found in the literature on the possible risks of and concerns over the use of generative AI in science and technology.[187,188,189,190,191]

7.1.1 Generative AI lowers the skill threshold

So far, modifying plants with NGT has been reserved for highly trained professionals who are extensively trained in molecular biological techniques. Generative AI is likely to change this. Its models are becoming increasingly more sophisticated and are acquiring the expertise and decision-making skills that were previously only available to experienced researchers.

In the near future, chatbots could provide instructions and support for laypeople, thus making NGT plant breeding accessible to students, computer scientists, entrepreneurs or DIY biologists. These people will have neither the experience in dealing with GMPs nor a sufficient awareness of biosafety issues. The loss of specialist skill in this area, in conjunction with the black box (7.1.4), hallucination (7.1.5) and possible data errors (7.1.6), raises concerns that NGT1 plants with undesirable or inappropriate properties may be created and released into the environment.

7.1.2 Generative AI increases productivity

Automation, AI-based research assistants, powerful computer simulations and design tools are increasingly turning NGT-based plant breeding into a data-driven process that produces more and more NGT plants at an ever-faster rate, which are then tested as possible candidates for new plant varieties in the natural environment. From the perspective of health and environmental protection, the acceleration and the associated productivity boost also give rise to concerns. This is because the increased pace and the sheer volume of possible candidates for the development of new varieties are likely to make it more difficult to identify and select those new plant varieties unexpectedly displaying properties undesirable for human, animal and environmental health during development.

7.1.3 Generative AI provides new tools

Dozens of simple gene scissors and sophisticated base, prime and epigenome editors are already available for NGT-based plant breeding. Generative AI will greatly expand this toolbox. Startups such as BellaGen and Qi Biodesign are examples of how structural analysis tools like Alphafold can be used to create new CRISPR-based gene scissors and base editors (see 6.7). Powerful protein design tools will further increase the possibilities. In April 2024, NGT tools developed using large protein language models were presented for the first time: a base editor from Westlake University,[192] and OpenCRISPR-1 from the startup Profluent.[193] OpenCRISPR-1 is particularly noteworthy. Firstly, the protein comes from a pool of millions

of new, computer-designed CRISPR protein sequences that Profluent designed with ProGen. Secondly, OpenCRISPR-1 is a very novel protein: it differs from any natural CRISPR protein by at least 182 mutations and from the widely used SPCas9 gene scissors by as many as 403 mutations.

Genomic language models are also soon likely to contribute to the development of new forms of NGT tools. For example, EVO, a model trained with genetic data from bacteria, can generate sequences that are new-to-nature but nevertheless are designed to function like Cas9 gene scissors.[194,195]

Whether nucleases, deaminases, recombinases, transposases or methyltransferases – today, the toolbox of NGT plant breeding is still largely restricted to components that come from natural sources. Generative AI will not only help to redesign these 'natural' NGT tools to make them even more powerful, but it could also give rise to a range of novel NGTs that facilitate multiplex genome editing, gene stacking, single sequence rearrangements, and chromosome remodelling.[196] Thanks to generative AI, research and industry will have tools to manipulate plant genomes on an even larger scale than today.

7.1.4 Black Box

Generative AI models often work as a 'black box':[197,198,199] they make predictions or recommendations without humans being able to understand exactly how and why the models came up with them. While this lack of transparency does not hinder the technological use of the AI models, it does limit the ability to evaluate them in terms of reliability or safety.

In sensitive areas such as NGT-based plant breeding, where the products can affect the health of many people and the environment, the lack of traceability and reproducibility of the results undermines trust in generative AI models. Therefore, ways must be sought to make future AI models transparent and comprehensible to the interested public and, in particular, to regulatory authorities. In addition, solutions must be found to ensure that human intelligence, control and governance remain integrated at critical points in the AI-driven production of NGT plants.

7.1.5 Hallucinations

Alongside the black box, 'hallucinations' are also a cause for concern: generative AI models can produce results that often appear reasonable but are actually factually wrong or irrelevant.[200] How often and in what contexts AI models 'hallucinate' and how this can be prevented or reduced has yet to be determined. What is clear, however, is that the unquestioned production of false and irrelevant results is to be expected. The combination of black box and hallucination is particularly problematic where generative AI models make suggestions for extensive interventions in the genetic makeup of plants, and the modified plants are then released into the environment.

7.1.6 Data distortion and lack of logical understanding

The outputs and predictions of generative AI models always reflect the data used to train the models. If the training data contains errors or distortions stemming either from the underlying biological systems or from the human curators, they can be transferred to the model's results. Furthermore, AI models lack an understanding of causality. They can correctly identify patterns and relationships in the data, but they cannot grasp what the immediate causes or mechanistic explanations for the identified relationships are. This lack of causal understanding ultimately limits the ability to anticipate possible side effects or malfunctions that may arise when implementing AI predictions into real-world applications.[201]

7.1.7 Speed and future-proofing

Technological advances in both NGT and generative AI are currently taking place at a breathtaking pace. Managing this rapid technological change poses a challenge for the governance of the convergence of NGT and generative AI. In areas where NGT and generative AI are used together, authorities and legislators will need to constantly assess whether existing regulations can keep pace with rapidly changing technological possibilities.

7.1.8 Corporate power

Tech companies have immense power when it comes to developing generative AI. They have the necessary infrastructure, highly qualified personnel, access to powerful computers and huge cloud capacities, and the financial resources needed for the very costly production of generative models.

Since SMEs and public academic institutions are rarely able to afford the high development costs, a large proportion of AI breakthroughs are in the hands of private corporations. A few tech giants can determine market trends, set standards, decide whether they disclose the codes of their models and dictate who, and under which conditions, can have access to the tools. What is more, they also have the power to influence ethical and regulatory discussions and thus also political decisions.[202]

With tech companies such as Meta, Google, NVIDIA, Salesforce and Microsoft now also developing generative AI models for life sciences, synthetic biology and NGT-based plant breeding, the question arises of whether and how corporate power influences the process. Do the goals of a company influence how their protein and genome-trained models work? How transparent, reproducible and comprehensible are the tech giants' tools? What are the consequences if AI models for NGT-based breeding become increasingly large and only a few companies can develop the best and most powerful tools? What forms and possibilities of state control are needed? And what resources, expertise and powers of intervention should be given to national or international institutions to carry out these controls? A broad public debate on these issues is needed. So far, the discussion has only taken place behind closed doors.[203,204]

7.1.9 'Open-washing'

Many of the AI models from private companies mentioned in this paper are public. Google/Instadeep's AgroNT, for example, is available on Hugging Face[205] and Inari's FloraBERT is on Github.[206] However, it remains to be seen what exactly 'public' means for each of the individual models.

A recent study by the Dutch Radboud University shows that the 'open source' label in the field of generative AI does

not always deliver what it promises. Two researchers there looked at how open, transparent and accessible chatbots and image generators from private companies actually are. The result: tech giants such as Google, Meta and Microsoft often describe their AI models as open source, but only disclose a few key pieces of information such as code or training data. In short, tech companies are doing 'open-washing'.[207,208]

In May 2024, Google received a lot of criticism from the scientific community when it presented Alphafold 3, the latest version of its revolutionary AI for predicting protein structures, in the journal *Nature*.[209,210] Although Alphafold 3 is available on a public web server, its use is subject to a license limited to non-commercial use. What's more, Google also refrained for the first time from making the computer code describing the progress of the model public. In addition to an outcry on social media, more than 1,000 researchers criticised the journal *Nature* in an open letter to the editors for accepting Google's article without computer code, thus deviating from the standards of the research community.[211]

7.2 'Google Crops' scenario

The development of generative AI models for NGT-based plant breeding is still in its infancy. It is not yet possible to say where it will lead. Stuart Smyth of the University of Saskatchewan recently predicted 'Google Crops' – yield-optimised varieties designed with AI and created with CRISPR.[212] The following scenario is based on this. It is intended to illustrate the regulatory issues that arise with the advent of generative AI models.

2027: Google has developed AgroNT into a multimodal model that not only understands the language of proteins and genomes, but also the legal conditions for breeding and growing NGT plants. Google offers its tool – let's call it 'The AI-Breeder' – to breeding companies, which can fine-tune it with their own data. Syngenta has been working with AgroNT since 2024 and is now using The AI-Breeder to produce NGT1 oilseed rape for the European market. To do this, the company enters the genome sequences of its elite crop varieties, as well as data on the climate of the growing areas and the soil quality of the fields, into the tool and receives information on how to edit

these varieties using genome editing to simultaneously achieve high yields and remain within the boundaries of NGT1. An automated genome editing workflow implements the suggestions of The AI-Breeder. Without having to check for possible environmental or health effects beforehand, Syngenta then releases the edited oilseed rape varieties into the environment and tests their yields in several locations on a trial basis. Since Syngenta does not have to take any measures to limit its trials in terms of time and space, edited rapeseed escapes from the trial areas via seeds and also passes on its Google-designed genes to other rapeseed plants and related wild species via pollen.

The scenario raises questions that it could be sensible to discuss at the regulatory and political level before the planned deregulation of NGT plants: is it conceivable that generative AI tools are prone to error and make unwanted suggestions, the implementation of which could lead to edited varieties with undesirable effects on humans, animals or the environment? If so, should it be the companies' own responsibility to decide whether or not to use reliable and safe tools? Are binding quality standards required for this? Should the companies themselves check that errors are detected and that no NGT1 plants with undesirable effects leave the laboratories? Should companies be able to choose for themselves how much decision-making they hand over to an AI and at which points in their AI-controlled breeding process they rely on human intelligence, control and decision-making? In short, is the personal responsibility and self-control of SMEs and corporations sufficient, or is state intervention needed to make sure the tools and plants are safe using appropriate regulation?

Another important question is how to determine whether an AI tool is reliable and makes safe suggestions. Does this require a step-by-step approach in which data are first collected on screen, then in the laboratory, in greenhouses and in controlled release experiments, and then submitted to authorities for evaluation? Or should companies and corporations figure it out on their own in the process of breeding, as would be the case under the proposed deregulation?

The following question is particularly important for this political and regulatory discussion: can a generative AI, with the design space legally available to it for designing an NGT1 genome, design plants whose risk profile differs from conventionally bred plants?

7.3 The design space for NGT1 plants

The design space available for the production of NGT1 plants under the European Commission's proposal is defined in Annex 1 of the NGT draft regulation. It sets out the criteria for the equivalence of NGT1 plants with conventionally bred plants (Figure 3). If the criteria are met, equivalence applies even if the properties of the NGT1 plants are novel and do not occur in conventionally bred varieties of the same species.

The following lines demonstrate the possibilities that the planned design space theoretically offers for the AI-controlled production of NGT1 plants: a generative AI can, for example, suggest the introduction of 18 new nucleotides for each of 20 different coding sites in the genome. Since this corresponds to six amino acids per site, it is possible to redesign several proteins. The AI can also suggest edits at 20 sites in the genome that act as CRE or uORF. This in turn allows it to design the regulatory network of a plant. The AI also has access to all DNA sequences from the gene pool of a plant species for its design proposals. It can, for example, select up to 20 genes from the super pangenome of a species and use them to suggest the formation of a new metabolic pathway. The design space for an AI becomes very large when it also exploits the possibilities of crosses. The EU Commission's proposal envisages that crosses between two different NGT1 plants will in turn lead to more NGT1 plants, even if the offspring then have more than 20 genetically engineered modifications.[213,214]

Figure 3: Criteria proposed by the EU Commission for the equivalence of NGT1 plants with conventional plants

A NGT plant is considered equivalent to conventional plants when it differs from the recipient/parental plant by no more than 20 genetic modifications of the types referred to in points 1 to 5, in any DNA sequence sharing sequence similarity with the targeted site that can be predicted by bioinformatic tools.

1. **Substitution or insertion of no more than 20 nucleotides;**
2. **Deletion of any number of nucleotides;**
3. **On the condition that the genetic modification does not interrupt an endogenous gene:**
   - a. **Targeted insertion of a contiguous DNA sequence existing in the breeder's gene pool;**
   - b. **Targeted substitution of an endogenous DNA sequence with a contiguous DNA sequence existing in the breeder's gene pool;**
4. **Targeted inversion of a sequence of any number of nucleotides;**
5. **Any other targeted modification of any size, on the condition that the resulting DNA sequences already occur (possibly with modifications as accepted under points (1) and/or (2)) in a species from the breeders' gene pool.**

7.4 Risk assessment of NGT1 plants

Under EU law, anyone who wants to release a genetically modified plant for experimental purposes or market it in the EU must first carry out a risk assessment. This requirement is designed to avoid any adverse effects of GMOs on humans, animals, the environment and biodiversity in advance. The EU Commission assumes that the risk profiles of NGT1 plants are the same as those of conventionally bred plants and therefore proposes that the requirement for a risk assessment of NGT1 plants under GMO legislation be waived. According to the EU Commission's plans, there would only be a risk assessment for human health for NGT1 foods that are classified as novel foods and fall under Regulation 2015/2283.

The question of whether the use of protein- and genome-based generative AI models can lead to NGT1 plants that differ in their risk profile from conventionally bred plants has not yet played a role in the regulatory debate. However, in view of the potential attributed to these AI models, their inclusion in the discussion is advisable. To allow an informed decision about the pros and cons of a potential mandatory risk assessment to be reached, it should first be clarified what risk profiles NGT1 plants might have, whose genome modifications have been proposed by an AI model.

Protein-based generative AI models are characterised by several abilities: they can predict protein structures, infer protein functions and predict both protein-protein interactions and interactions between proteins and small molecules. The AI models thus significantly improve the possibilities for the genetic engineering of natural plant proteins. For the regulatory debate, it would therefore be useful to take stock of the protein-based generative AI models and to assess the current and future redesign potential of AI models in the NGT1 design space (7.3). In particular, the question should be answered as to whether redesigned proteins with novel functions that give NGT1 plants characteristics with an increased risk profile are currently feasible or will in the future be feasible.

Genome-based generative AI models improve our understanding of genomes and provide insight into the way DNA elements interact at different levels to enable complex functions. They can help predict the effects of genome modifications and design functional DNA sequences. As with the protein-based models, the genome-based generative AI models should also clarify what

design potential currently exists within the NGT1 design space and what might be expected in the future as AI models progress. In turn, this should answer the question of whether the AI models also enable the production of NGT1 plants whose risk profile is increased compared to conventionally bred plants.

A scenario that describes possible NGT1 plants with an increased risk profile: the plant's own microRNAs can be modified using genome editing in such a way that they prevent the formation of essential proteins by RNAi in harmful insects. With a genome-based generative AI model, it might be possible to search the genome of a plant variety for sequences that code for microRNAs. In this scenario, the model finds several such sequences and suggests for three of them the modifications necessary, in less than 20 nucleotides in each of the three, in order to have a toxic effect on two different insect pests via RNA interference. The suggestions are implemented using genome editing, resulting in an NGT1 plant that produces six insect-toxic substances and could not be developed in a practical time frame using conventional breeding methods. According to the EU Commission's plans, no tests would be carried out on the NGT1 plant to determine the effect of the newly formed microRNAs on non-target insects before it was placed on the market, even though such undesirable effects are conceivable.[215] If, on the other hand, the newly formed microRNAs were to be sprayed on fields as a plant protection product, they would have to undergo a risk assessment under current EU plant protection product legislation.

Another scenario: a breeding company uses a genome-based generative AI model to determine how to edit the regulatory elements of the zmm28 gene in maize to increase the expression of the gene. The zmm28 gene encodes for a transcription factor that regulates the activity of genes involved in processes such as photosynthesis, nitrogen assimilation and growth-regulating hormone signalling. The breeding company implements the AI model's suggestions and generates an NGT1 maize with a higher grain yield. It is unlikely that a conventional breeding programme could produce exactly the same genome edits as were identified by the AI model. Overexpression of the ZMM28 transcription factor raises safety concerns:[216] food and feed from NGT1 maize could produce more auxins, indolylacetic acid, indolylbutyric acid or nitrate than usual. These safety concerns would remain unresolved if there were no requirement for risk assessment for NGT1 plants.

A third scenario: a start-up commissions a genome-based AI model to search

a super pangenome, consisting of all publicly available genome sequences of oilseed rape and five of its related wild relatives, for sequences that could potentially code for antimicrobial peptides (AMPs). AMPs are part of the plant defence system against fungi, viruses and bacteria, as well as to some extent against insects and nematodes. Depending on their sequence and structure, AMPs are categorised into different types such as thionins, defensins, knottins, snakin, cyclotins or hevein-like peptides. The AI model provides the startup with two dozen sequences. The startup then has the corresponding genes synthetically produced, uses them to create several different cisgenic variants of oilseed rape and then tests the variants in the environment. One variant, which contains six potential AMP genes from four related wild species, proves to be particularly robust against harmful fungi, and the startup commercialises it as a variety for biodiesel production. Using conventional breeding methods, the variety could not be produced within a practical time frame. Questions that arise for the regulatory debate in this scenario: what information about the new variety should the startup be required to provide to the relevant authorities in the process of verifying its status as an NGT1 plant? Is it sufficient to state that the six inserted cisgenes are AMP genes according to the AI model? Or would the start-up have to clarify experimentally in advance whether the six proteins proposed by the AI model are actually AMPs? And is it justifiable for the start-up to release the cisgenic oilseed rape variant into the environment without a risk assessment, even though it is known that certain AMPs can also have a toxic effect on animals and humans?

7.5 Labelling of NGT1 plants

Under current EU law, food that consists of or is produced from genetically modified plants must be labelled as GMO. This labelling requirement ensures freedom of choice at the retail and consumer level. The decisive factor for the labelling requirement is not the presence of foreign DNA in plant-based food. Rather, the decisive factor is whether genetic engineering techniques have been used to produce the plant. The EU Commission now wants to abolish this process labelling for NGT1 plants. Even though classical genetic engineering methods are generally used today in the production of NGT1 plants, the EU Commission no longer wants to make this process transparent and proposes in its draft legislation to exempt NGT1 plants from labelling

requirements. The absence of foreign DNA and the assumed equivalence to conventionally bred plants are now to become the criteria that determine freedom of choice in the case of NGT1 plants.

When the pros and cons of a labelling requirement for NGT1 plants were previously debated politically, the possible use of AI in the plant production process did not play a role. Now however, protein- and genome-based generative AI models cast a new light on the removal of process labelling and make it necessary to include AI in the deregulation debate.

Can the information as to whether generative AI was used in the production of NGT1 plants be essential for consumers to decide whether or not to buy the NGT1 product? Should consumers be informed if AI was used to design the genome of a tomato they want to buy? These are two of the questions that the possible use of generative AI models brings up for debate and that need to be brought into consideration politically. Important aspects here are the extent of AI design, the possible waiving of risk assessments for NGT1 plants (7.4), the black box (7.1.4) and the artificiality of the change created (new-to-nature) by the use of AI models.

7.6 Traceability of NGT1 plants

Traceability systems are a legal requirement for the commercial handling of GMPs in the EU. They ensure that GMPs and the products derived from them can be traced seamlessly throughout the entire manufacturing and distribution chain and in nature. The traceability requirement was introduced in order to enable the rapid recall of any defective products. This risk-prevention measure, a precaution stepping into play after the GMP comes onto the market, the EU Commission now wants to abolish with regards to NGT1 plants.

How the removal of the traceability requirement for NGT1 plants, whose modifications were proposed by an AI model, is to be evaluated has not yet been discussed by the relevant authorities, political institutions or the interested public, but should be included in the debate on the regulation of NGT plants. For example, it should be discussed whether black box (7.1.4), hallucinations (7.1.5) and data distortions (7.1.6) could not also lead to unsafe or defective NGT1 products and whether a recall option would be useful if AI-driven suggestions lead directly to modified genomes and the resulting plants can come onto the market without risk assessment and government oversight.

Glossary

(created with the help of ChatGPT)

Antimicrobial proteins

Antimicrobial proteins (AMPs) are small proteins produced by plants to defend themselves against various pathogens such as bacteria, fungi and viruses. AMPs function against microorganisms by destroying their cell walls or membranes or by interfering with their metabolic processes. AMPs play an important role in the immune system of plants and can also be used in plant breeding to develop disease-resistant varieties.

Base editor

A base editor is a genome editing tool based on the CRISPR system. It allows the targeted modification of individual DNA bases in a genome without creating the double-strand breaks that are typical of the conventional CRISPR-Cas9 system. For example, a base editor can specifically convert a single base, such as a C-G base pair, into a T-A base pair. A base editor consists of a modified Cas protein that mediates DNA binding, but not DNA cutting, and a → deaminase component that causes the chemical conversion of one base into another.

Cis-regulatory element

A cis-regulatory element – CRE for short – is a DNA sequence that controls the activity of a gene by enabling or preventing the binding of → transcription factors and other regulatory proteins. These elements are usually located near the gene they regulate. CREs play a key role in gene expression by determining when, where and how strongly a gene is expressed. Examples of cis-regulatory elements are → promoters, → enhancers and → silencers.

Deep learning

Deep learning is a subfield of machine learning based on artificial neural networks. It is a method by which a computer learns to recognise and understand complex patterns and relationships within large amounts of data. These neural networks consist of several layers (hence ‘deep’), through which data is processed and transformed step by step to identify patterns, features or decisions.

Deaminase

A deaminase is an enzyme that catalyses a chemical reaction called deamination. This reaction removes an amino group (-NH_2) from a molecule. Deaminases play important roles in amino acid and nucleotide metabolism by enabling the removal of amino groups, which is essential for energy production and the breakdown of excess nitrogen compounds. Deaminases play an important role in genome editing because they can be used to create → base editors.

Descriptive artificial intelligence

Descriptive artificial intelligence refers to the use of → artificial intelligence (AI) to analyse and describe existing data. It can identify patterns in large data sets that are often difficult for humans to see. Descriptive AI helps to better understand existing data.

Diffusion model

A diffusion model is a type of generative model in → artificial intelligence designed to learn complex data patterns by gradually adding noise to a data structure and then reversing the process to restore that structure. These models learn how to reconstruct data from a noisy state, which allows them to generate new, realistic-looking data.

In a biological context, diffusion models could be used to capture and reconstruct patterns in genetic sequences, such as DNA or RNA, or to aid in the simulation of molecular processes. For example, they could be used to model the folding of proteins or to predict genetic mutations by learning the transition from a disordered to an ordered state and vice versa.

Enhancer

An enhancer is a segment of DNA in the genome that can boost the expression of one or more genes. Enhancers are a type of → cis-regulatory element. They work by binding to specific transcription factors or promoting the binding of activator proteins. Enhancers can be located far from the gene they regulate and still influence its expression. An enhancer is the counterpart to a → silencer.

Single-cell omics

Single-cell omics refers to a collection of techniques and methods that make it possible to obtain and analyse biological information at the level of individual cells. In contrast to conventional 'bulk' analyses, which collect data from a mixed population of cells and only provide average values, single-cell omics allows a detailed examination of individual cells. This can reveal differences between individual cells that would remain hidden in a bulk analysis.

Epiallele

Epialleles are genes or alleles that match in their DNA sequence but have different epigenetic modifications (e.g. methylation) and are therefore usually expressed differently.

Epigenome

The epigenome refers to the entirety of all epigenetic modifications to the DNA of an organism that regulate gene activity and gene expression without changing the DNA sequence itself. The epigenetic modifications include, among others, DNA methylation, histone modifications and the organisation of the chromatin structure. In contrast to the genome, which is relatively stable, the epigenome can be dynamic and change more extensively over the course of a lifetime.

Epigenome editing

Epigenome editing is a technique that aims to make targeted changes in the epigenome to control gene expression without altering the underlying DNA sequence. While traditional genome editing methods such as CRISPR/Cas9 directly modify the DNA sequence, epigenome editing focuses on the modification of epigenetic markers such as DNA methylation and histone modifications.

One of the techniques for epigenome editing relies on the use of dCas9, an inactive Cas9 enzyme that does not cut DNA. Enzymes that act as epigenetic modulators can be coupled to dCas9. For example, dCas9 can be fused with a DNA methyltransferase to specifically alter DNA methylations.

Functional annotation

In genomics, functional annotation refers to the process of assigning certain biological functions to the identified sequences of a genome. A sequenced genome initially exists in the form of a long sequence of DNA bases (A, T, C, G). Functional annotation helps to interpret this sequence and to find out which sections of the DNA play which role in the organism.

Generative artificial intelligence

Generative artificial intelligence refers to the use of → artificial intelligence (AI) that not only analyses data but also creates (generates) new data. Examples of generative AI are models like ChatGPT, which can generate human language, or DALL-E, which can create images.

Large language model

A large language model (LLM) is a type of artificial intelligence designed to analyse, understand and generate complex sequences of symbols, characters or data. It is based on deep neural networks. The models learn patterns, relationships and structures within large amounts of sequential data, be it text or biological sequences such as those found in DNA, RNA or proteins. By training on extensive data sets, large language models can perform various tasks, such as predicting sequences, generating new sequences or classifying data. For example, ChatGPT relies on a large language model that can understand and generate human language.

Artificial intelligence

Artificial intelligence (AI) refers to the development of computer systems capable of performing tasks that normally require human intelligence. This includes pattern recognition, natural language understanding, decision-making and learning from experience. In life sciences, AI encompasses machine learning techniques and other intelligent algorithms that can analyse complex biological data and are implemented in areas including genomics, proteomics and image analysis.

Machine learning

Machine learning is a branch of → Artificial Intelligence that involves developing algorithms and statistical models that allow computers to perform tasks without

explicit instructions. In machine learning, an algorithm learns from patterns in a large set of labelled data. Once trained, predictions or decisions can be made based on that learning in response to new and unseen data.

Metabolomics

Metabolomics refers to the analysis and identification of all metabolites in plant samples to better understand the biochemical processes and metabolic pathways in plants. The method can be used to gain insights into plant physiology, responses to environmental stress, the biosynthesis of phytochemicals, and also the effects of genetic engineering.

Methyltransferases

Methyltransferases are enzymes that transfer a methyl group ($-CH_3$) to a substrate. These substrates can be DNA, RNA, proteins or other molecules. Methylation by methyltransferases is an important biochemical process that can regulate the function of genes and proteins. For example, methylation of DNA segments by DNA methyltransferases can inactivate genes, which is a form of epigenetic gene regulation. In combination with the CRISPR/Cas system, methyltransferases can be used for targeted → epigenome editing

MicroRNA

Micro-RNA – or miRNA for short – is a short, non-coding and single-stranded RNA molecule that is an important component of RNA interference (RNAi). miRNAs are about 21-25 nucleotides long and involved in a variety of biological processes, such as cell growth and differentiation, apoptosis (programmed cell death) and stress reactions.

Multiplex genome editing

Multiplex genome editing refers to the simultaneous editing of several target sites in the genome of a single cell. Multiplexing is mainly achieved with the CRISPR system: by introducing several different guide RNAs (gRNAs) simultaneously with the Cas cutting enzyme into cells, changes are also made at several different sites in the genome. The method makes it possible to edit or switch off several genes at once.

Neural network

A neural network is a model or programme for machine learning inspired by the way the human brain works. It consists of a large number of interconnected nodes – so-called artificial neurons. The nodes are organised into layers: an input layer, one or more hidden layers, and an output layer. Neural networks rely on training data to learn and improve their accuracy over time. They can classify, cluster and generate data at high speed.

Nucleases

Nucleases are enzymes that can cleave nucleic acids, i.e. DNA or RNA. They can break specific bonds between nucleotides, causing the DNA or RNA strands to be cut. In genome editing, nucleases are crucial because they can specifically cut at specific sites in the genome, paving the way for various genetic engineering manipulations. The Cas9 enzyme of the CRISPR system is an example of a nuclease.

Omics techniques

Omics techniques are a group of techniques and approaches that aim to study the entirety (the 'om') of certain classes of molecules in cells, tissues or organisms. They enable the acquisition of comprehensive information about the structure, function and dynamics of biological systems. Omics techniques include, but are not limited to, → genomics, → proteomics, → transcriptomics and → metabolomics.

Pangenome

The pangenome of a plant species encompasses, as far as possible, the entire set of genes that occur within that species. It consists of the core genome (the genes that are present in all individuals of the species) and the variable genome (the genes that are present only in some, but not all, individuals of the species).

Peptide

A peptide is a molecule consisting of a short chain of amino acids linked together by peptide bonds. Peptides can consist of only two amino acids (dipeptides) or of longer chains of up to about 100 amino acids (polypeptides). Peptides play important roles in biological processes such as signal transduction or defence reactions.

Polyploid plant

A polyploid plant has more than two sets of chromosomes in its cells. Unlike diploid plants, which have two sets of chromosomes (one from each parent), polyploid plants often have three (triploid), four (tetraploid) or even more sets of chromosomes. Polyploidy occurs naturally and can arise through evolutionary processes, such as errors in cell division. Polyploid plants often exhibit increased robustness, larger cells and fruits, and higher genetic diversity, making them particularly valuable for breeding and agriculture.

Prime editor

A prime editor is a genome editing tool based on the CRISPR system that is used for precise editing of DNA sequences. Unlike the conventional CRISPR-Cas9 system, which creates double-strand breaks in DNA, the prime editor combines Cas enzymes that introduce single-strand breaks with a reverse transcriptase that can transcribe RNA into DNA. This combination enables the insertion, deletion or exchange of specific DNA sequences without double-strand breaks. The prime editor system uses a so-called prime editing guide RNA (pegRNA): it not only guides the enzymes to the target site in the genome, but also contains the information for the desired change. The reverse transcriptase copies this information into the target site in the genome.

Promoter

A promoter is a → cis-regulatory element. It is located directly in front of the → transcription start site of a gene and serves as a binding site for RNA polymerase and other → transcription factors to initiate transcription.

Proteomics

Proteomics refers to the analysis and characterisation of all proteins that are produced in a plant cell or tissue at a given time. The method can be used to determine the range of proteins, including protein modifications and interactions, which in turn provides insights into the functional biology of the plant, its reactions to environmental conditions or the effects of genetic engineering.

Quantitative Traits

Quantitative traits are plant characteristics influenced by many genes (polygenic), which not only occur in distinct categories but show continuous variation. In contrast to qualitative traits, which are determined by few genes and show discrete, clearly identifiable classes (e.g. flower colour), quantitative traits are characterised by a broad scale of expressions.

Recombinase

A recombinase is an enzyme that mediates genetic recombination by recognising and cutting DNA strands at specific sites and rejoining them. Recombinases play a central role in the rearrangement of DNA sequences that occur in nature, for example, in DNA repair, in the exchange of genetic material between chromosomes or in the integration of viral DNA into the host genome. In genome editing, recombinases are used to specifically modify DNA sequences.

RNA interference

RNA interference – or RNAi for short – is a natural cellular process that regulates gene expression by degrading specific mRNA molecules or preventing their translation into proteins. RNAi plays a central role in gene regulation and serves as a defence mechanism against viruses.

scRNA-Seq data

scRNA-Seq data (Single-Cell RNA Sequencing data) come from a technology that allows measuring gene expression in single cells. Unlike traditional RNA-sequencing methods, which measure average gene expression across many cells, scRNA-Seq provides a detailed view of gene expression at the level of individual cells.

Silencer

A silencer is a DNA segment in the genome that can repress or reduce the expression of one or more genes. Silencers are one of the → cis-regulatory elements. Silencers can be located far from the gene they regulate and still influence its expression. A silencer is the counterpart of the → enhancer.

Small interfering RNA

Small interfering RNA (siRNA) is a short, non-coding, double-stranded RNA molecule that is an important component of RNA interference (RNAi). siRNAs are about 20-25 nucleotides long and are used by plants to regulate the expression of specific genes. Plants use siRNA to defend themselves against viruses.

Structure-guided protein engineering

Structure-guided protein engineering is a biotechnology approach that uses the three-dimensional structure of a protein to make specific changes to the amino acid sequence of the protein. The aim is to improve or modify the function, stability, binding affinity or other properties of the protein.

Super pangenome

A super pangenome of a plant species encompasses, as far as possible, the entire set of genes that occur within this species and its relatives. Super pangenomes usually correspond to → pangenomes at the genus level. They provide insights into the evolutionary history, domestication processes and genetic relationships within a genus.

Trait

Trait is a term used in genetic plant breeding. It refers to a specific property or ability of a plant, either specifically introduced by genetic engineering or naturally occurring. A trait can be, for example, resistance to certain pests, higher tolerance to herbicides, or improved nutrient composition. By inserting one or more genes that are responsible for the desired trait, researchers can introduce certain properties into plants in a targeted manner.

Transcription

Transcription refers to the process by which the genetic information in DNA is rewritten into a complementary RNA sequence. During this process, a specific section of DNA containing a gene is read by an enzyme called RNA polymerase and converted into mRNA (messenger RNA). This mRNA later serves as a template for translation, during which the information encoded in the RNA is translated into a protein. Transcription is the first step in gene expression.

Transcriptomics

Transcriptomics is the study of all RNA molecules (in particular mRNA) in a plant cell or tissue. It is used to determine the pattern of gene expression under specific conditions, which makes it possible to study gene activity at different stages of development or in response to environmental factors.

Translation

Translation is the name given to the process by which the genetic information encoded in mRNA (messenger RNA) is translated into an amino acid sequence to form a protein. This process occurs in the ribosomes. During translation, the ribosomes read the mRNA sequence in groups of three nucleotides, called codons, adding the appropriate amino acids to a growing polypeptide chain. This chain then folds into a functional protein.

Transposase

A transposase is an enzyme responsible for the mobility of transposons (jumping genes or mobile genetic elements). Transposons are DNA sequences that can be moved from one location to another within a genome. The transposase recognises specific DNA sequences at the ends of a transposon, cuts them out of their original position and integrates them into a new location in the genome. In genome editing, transposases can be used to insert or remove genes from a genome in a targeted manner.

Upstream Open Reading Frame

An upstream open reading frame – uORF for short – is a short open reading frame (ORF) located upstream of the main ORF of a gene. A uORF can potentially code for a small peptide. → ORFs are important elements that help to fine-tune protein production by regulating → translation in response to cellular conditions and signals.

References

1 Chao, H., Zhang, S., Hu, Y., Ni, Q., Xin, S., Zhao, L., ... & Chen, M. (2024). Integrating omics databases for enhanced crop breeding. Journal of Integrative Bioinformatics 20(4): 20230012.

2 Mohanta, T. K., Kamran, M. S., Omar, M., Anwar, W., & Choi, G. S. (2022). PlantMW pI DB: a database for the molecular weight and isoelectric points of the plant proteomes. Scientific Reports 12(1): 7421.

3 Liu, J., Zhang, Y., Zheng, Y., Zhu, Y., Shi, Y., Guan, Z., ... & Dou, D. (2023). PlantExp: a platform for exploration of gene expression and alternative splicing based on public plant RNA-seq samples. Nucleic Acids Research 51(D1): D1483-D1491.

4 Tian, Z., Hu, X., Xu, Y., Liu, M., Liu, H., Li, D., ... & Chen, W. (2024). PMhub 1.0: a comprehensive plant metabolome database. Nucleic Acids Research 52(D1): D1579-D1587.

5 Li, F. W., & Harkess, A. (2018). A guide to sequence your favorite plant genomes. Applications in Plant Sciences, 6(3), e1030.

6 Xie, L., Gong, X., Yang, K., Huang, Y., Zhang, S., Shen, L., ... & Fan, L. (2024). Technology-enabled great leap in deciphering plant genomes. Nature Plants 10(4): 551-566.

7 http://ibi.zju.edu.cn/N3database/index.php

8 https://www.ncbi.nlm.nih.gov/datasets/genome/

9 Bernal-Gallardo, J. J., & de Folter, S. (2024). Plant genome information facilitates plant functional genomics. Planta 259(5): 117.

10 Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M., ... & Zhang, G. (2022). The earth BioGenome project 2020: Starting the clock. Proceedings of the National Academy of Sciences 119(4): e2115635118.

11 Schreiber, M., Jayakodi, M., Stein, N., & Mascher, M. (2024). Plant pangenomes for crop improvement, biodiversity and evolution. Nature Reviews Genetics *in press*

12 Hu, H., Li, R., Zhao, J., Batley, J., & Edwards, D. (2024). Technological development and advances for constructing and analyzing plant pangenomes. Genome Biology and Evolution 16(4): evae081.

13 Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D., & Varshney, R. K. (2020). Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends in Plant Science 25(2): 148-158.

14 Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., ... & Qian, Q. (2022). A super pan-genomic landscape of rice. Cell Research 32(10): 878-896.

15 Gui, S., Wei, W., Jiang, C., Luo, J., Chen, L., Wu, S., ... & Yan, J. (2022). A pan-Zea genome map for enhancing maize improvement. Genome Biology 23(1): 178.

16 Li, N., He, Q., Wang, J., Wang, B., Zhao, J., Huang, S., ... & Yu, Q. (2023). Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nature Genetics 55(5): 852-860.

17 Khan, A. W., Garg, V., Sun, S., Gupta, S., Dudchenko, O., Roorkiwal, M., ... & Varshney, R. K. (2024). Cicer super-pangenome provides insights into species evolution and agronomic trait loci for crop improvement in chickpea. Nature Genetics 56: 1225-1234.

18 Lam, H. Y. I., Ong, X. E., & Mutwil, M. (2024). Large language models in plant biology. Trends in Plant Science *in press*

19 Islam, M. T., Liu, Y., Hassan, M. M., Abraham, P. E., Merlet, J., Townsend, A., ... & Yang, X. (2024). Advances in the application of single-cell

transcriptomics in plant systems and synthetic biology. BioDesign Research 6: ID0029.

20 Kaur, H., Jha, P., Ochatt, S. J., & Kumar, V. (2024). Single-cell transcriptomics is revolutionizing the improvement of plant biotechnology research: recent advances and future opportunities. Critical Reviews in Biotechnology 44(2): 202-217.

21 He, Z., Luo, Y., Zhou, X., Zhu, T., Lan, Y., & Chen, D. (2024). scPlantDB: a comprehensive database for exploring cell types and markers of plant cell atlases. Nucleic Acids Research 52(D1): D1629-D1638.

22 https://www.plantcellatlas.org

23 Rhee, S. Y., Birnbaum, K. D., & Ehrhardt, D. W. (2019). Towards building a plant cell atlas. Trends in Plant Science 24(4): 303-310.

24 https://www.plantcellatlas.org/2021-pca-symposium---dec-2021.html

25 Zheng, D., Xu, J., Lu, Y., Chen, H., Chu, Q., & Fan, L. (2023). Recent progresses in plant single-cell transcript-omics. Crop Design 2: 100041.

26 https://chatgpt.com/g/g-00Xk9QlqJ-crispr-gpt

27 https://chatgpt.com/g/g-20ZVLapH9-plant-breeding-optimizer

28 Huang, K., Qu, Y., Cousins, H., Johnson, W. A., Yin, D., Shah, M., ... & Cong, L. (2024). Crispr-GPT: An LLM agent for automated design of gene-editing experiments. arXiv:2404.18021.

29 Yang, X., Gao, J., Xue, W., & Alexandersson, E. (2024). Pllama: An open-source large language model for plant science. arXiv:2401.01600.

30 Fang, J. (2024). Breeding 5.0: AI-driven revolution in designed plant breeding. Molecular Plant Breeding 15

31 Callaway, E. (2022). The entire protein universe': AI predicts shape of nearly every known protein. Nature 608(7921): 15-16.

32 https://www.theatlantic.com/sponsored/google-2023/unlocking-lifes-building-blocks-demis-hassabis/3867/

33 Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596(7873). 583-589.

34 https://esmatlas.com

35 Kortemme, T. (2024). De novo protein design – From new structures to programmable functions. Cell 187(3): 526-544.

36 Winnifrith, A., Outeiral, C., & Hie, B. L. (2024). Generative artificial intelligence for de novo protein design. Current Opinion in Structural Biology 86: 102794.

37 Notin, P., Rollins, N., Gal, Y., Sander, C., & Marks, D. (2024). Machine learning for functional protein design. Nature Biotechnology 42(2): 216-228

38 Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., ... & Grigoryan, G. (2023). Illuminating protein space with a programmable generative model. Nature 623(7989): 1070-1078.

39 Alamdari, S., Thakkar, N., van den Berg, R., Lu, A. X., Fusi, N., Amini, A. P., & Yang, K. K. (2023). Protein generation with evolutionary diffusion: sequence is all you need. bioRxiv, 2023-09.

40 Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., ... & Naik, N. (2023). Large language models generate functional protein sequences across diverse families. Nature Biotechnology, 41(8), 1099-1106.

41 Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., ... & Baker,

D. (2023). De novo design of protein structure and function with RFdiffusion. Nature 620(7976): 1089-1100.

42 Ferruz, N., Schmidt, S., & Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. Nature Communications 13(1): 4348.

43 Ni, B., Kaplan, D. L., & Buehler, M. J. (2024). ForceGen: End-to-end de novo protein generation based on nonlinear mechanical unfolding responses using a language diffusion model. Science Advances 10(6): eadl4000.

44 Callaway, E. (2023). AI tools are designing entirely new proteins that could transform medicine. Nature 619 (7969): 236-238.

45 Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596(7873). 583-589.

46 Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature in press

47 Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379(6637: 1123-1130.

48 Alamdari, S., Thakkar, N., van den Berg, R., Lu, A. X., Fusi, N., Amini, A. P., & Yang, K. K. (2023). Protein generation with evolutionary diffusion: sequence is all you need. bioRxiv, 2023-09.

49 Zheng, Z., Deng, Y., Xue, D., Zhou, Y., Ye, F., & Gu, Q. (2023, July). Structure-informed language models are protein designers. In: International Conference on Machine Learning, pp. 42317-42338. PMLR.

50 Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., ... & AlQuraishi, M. (2024). OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Nature Methods, 1-11.

51 Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., ... & Naik, N. (2023). Large language models generate functional protein sequences across diverse families. Nature Biotechnology 41(8): 1099-1106.

52 Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., ... & Rost, B. (2021). Prottrans: Toward understanding the language of life through self-supervised learning. IEEE transactions on pattern analysis and machine intelligence 44(10): 7112-7127.

53 Sevgen, E., Moller, J., Lange, A., Parker, J., Quigley, S., Mayer, J., ... & Ferguson, A. L. (2023). ProT-VAE: protein transformer variational autoencoder for functional protein design. bioRxiv, 2023-01.

54 de Almeida, B. P., Dalla-Torre, H., Richard, G., Blum, C., Hexemer, L., Gélard, M., ... & Pierrot, T. (2024). SegmentNT: annotating the genome at single-nucleotide resolution with DNA foundation models. bioRxiv, 2024-03.

55 Li, S., Moayedpour, S., Li, R., Bailey, M., Riahi, S., Kogler-Anele, L., ... & Jager, S. (2023). CodonBERT: Large Language Models for mRNA design and optimization. bioRxiv, 2023-09.

56 Yin, W., Zhang, Z., He, L., Jiang, R., Zhang, S., Liu, G., ... & Xie, Z. (2024). ERNIE-RNA: An RNA Language Model with Structure-enhanced Representations. bioRxiv, 2024-03.

57 Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., & Wang, B. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods *in press*

58 Lam, H. Y. I., Ong, X. E., & Mutwil, M. (2024). Large language models in plant biology. Trends in Plant Science in press

59 Benegas, G., Batra, S. S., & Song, Y. S. (2023). DNA language models are powerful predictors of genome-wide variant effects. Proceedings of the National Academy of Sciences, 120(44), e2311219120.

60 Levy, B., Xu, Z., Zhao, L., Kremling, K., Altman, R., Wong, P., & Tanner, C. (2022). FloraBERT: cross-species transfer learning with attention-based neural networks for gene expression prediction. Research Square

61 Mendoza-Revilla, J., Trop, E., Gonzalez, L., Roller, M., Dalla-Torre, H., de Almeida, B. P., ... & Lopez, M. (2023). A Foundational Large Language Model for Edible Plant Genomes. bioRxiv, 2023-10.

62 Zhai, J., Gokaslan, A., Schiff, Y., Berthel, A., Liu, Z. Y., Miller, Z. R., ... & Kuleshov, V. (2024). Cross-species plant genomes modeling at single nucleotide resolution using a pre-trained DNA language model. bioRxiv, 2024-06.

63 Lam, H. Y. I., Ong, X. E., & Mutwil, M. (2024). Large language models in plant biology. Trends in Plant Science *in press*

64 Boshar, S., Trop, E., de Almeida, B. P., Copoiu, L., & Pierrot, T. (2024). Are genomic language models all you need? exploring genomic language models on protein downstream tasks. bioRxiv, 2024-05.

65 Shao, B. (2023). A long-context language model for deciphering and generating bacteriophage genomes bioRxiv, 2023-12.

66 Nguyen, E., Poli, M., Durrant, M. G., Thomas, A. W., Kang, B., Sullivan, J., ... & Hie, B. L. (2024). Sequence modeling and design from molecular to genome scale with Evo. bioRxiv, 2024-02.

67 Mendoza-Revilla, J., Trop, E., Gonzalez, L., Roller, M., Dalla-Torre, H., de Almeida, B. P., ... & Lopez, M. (2024). A foundational large language model for edible plant genomes. Communications Biology 7(1): 835.

68 Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 37(15): 2112-2120.

69 Nguyen, E., Poli, M., Durrant, M. G., Thomas, A. W., Kang, B., Sullivan, J., ... & Hie, B. L. (2024). Sequence modeling and design from molecular to genome scale with Evo. bioRxiv, 2024-02.

70 Levy, B., Xu, Z., Zhao, L., Kremling, K., Altman, R., Wong, P., & Tanner, C. (2022). FloraBERT: cross-species transfer learning with attention-based neural networks for gene expression prediction. Research Square

71 Zvyagin, M., Brace, A., Hippe, K., Deng, Y., Zhang, B., Bohorquez, C. O., ... & Ramanathan, A. (2023). GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. The International Journal of High Performance Computing Applications 37(6): 683-705.

72 Benegas, G., Batra, S. S., & Song, Y. S. (2023). DNA language models are powerful predictors of genome-wide variant effects. Proceedings of the National Academy of Sciences, 120(44), e2311219120.

73 Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L.,... & Pierrot, T. (2023). The nucleotide transformer: Building and evaluating robust foundation models for human genomics. BioRxiv, 2023-01.

74 Shao, B. (2023). A long-context language model for deciphering and generating bacteriophage genomes bioRxiv, 2023-12.

75 Zhai, J., Gokaslan, A., Schiff, Y., Berthel, A., Liu, Z. Y., Miller, Z. R., ... & Kuleshov, V. (2024). Cross-species plant genomes modeling at single nucleotide resolution using a pre-trained DNA language model. bioRxiv, 2024-06.

76 Richard, G., de Almeida, B. P., Dalla-Torre, H., Blum, C., Hexemer, L., Pandey, P., ... & Pierrot,

T. (2024). ChatNT: A Multimodal Conversational Agent for DNA, RNA and Protein Tasks. bioRxiv, 2024-04.

77 https://www.instadeep.com/2024/04/building-the-next-generation-of-ai-models-to-decipher-human-biology/

78 Garau-Luis, J. J., Bordes, P., Gonzalez, L., Roller, M., de Almeida, B. P., Hexemer, L., ... & Richard, G. (2024). Multi-modal transfer learning between biological foundation models. arXiv preprint arXiv:2406.14150.

79 Li, J., Xu, M., Xiang, L., Chen, D., Zhuang, W., Yin, X., & Li, Z. (2024). Foundation models in smart agriculture: Basics, opportunities, and challenges. Computers and Electronics in Agriculture 222: 109032.

80 Yan, J., & Wang, X. (2023). Machine learning bridges omics sciences and plant breeding. Trends in Plant Science 28(2): 199-210.

81 Zhu, W., Han, R., Shang, X., Zhou, T., Liang, C., Qin, X., ... & Li, L. (2024). The CropGPT project: Call for a global, coordinated effort in precision design breeding driven by AI using biological big data. Molecular Plant 17(2): 215-218.

82 Mattiello, L., Rütgers, M., Sua-Rojas, M. F., Tavares, R., Soares, J. S., Begcy, K., & Menossi, M. (2022). Molecular and computational strategies to increase the efficiency of CRISPR-based techniques. Frontiers in Plant Science 13: 868027.

83 Chen, L., Liu, G., & Zhang, T. (2024). Integrating machine learning and genome editing for crop improve-ment. aBIOTECH 1-16.

84 Lei, Y., Lu, L., Liu, H. Y., Li, S., Xing, F., & Chen, L. L. (2014). CRISPR-P: a web tool for synthetic single-guide RNA design of CRISPR-system in plants. Molecular Plant 7(9): 1494-1496.

85 Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K., & Chen, L. L. (2017). CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants. Molecular Plant 10(3): 530-532.

86 Xie, X., Ma, X., Zhu, Q., Zeng, D., Li, G., & Liu, Y. G. (2017). CRISPR-GE: a convenient software toolkit for CRISPR-based genome editing. Molecular Plant 10(9): 1246-1249.

87 Minkenberg, B., Zhang, J., Xie, K., & Yang, Y. (2019). CRISPR-PLANT v2: An online resource for highly specific guide RNA spacers based on improved off-target analysis. Plant Biotechnology Journal 17(1): 5.

88 Concordet, J. P., & Haeussler, M. (2018). CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Research 46(W1): W242-W245.

89 Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M., & Valen, E. (2014). CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic Acids Research 42(W1): W401-W407.

90 Khaipho-Burch, M., Cooper, M., Crossa, J., de Leon, N., Holland, J., Lewis, R., ... & Buckler, E. S. (2023). Genetic modification can improve crop yields – but stop overselling it. Nature 621(7979): 470-473.

91 Patel-Tupper, D., Kelikian, A., Leipertz, A., Maryn, N., Tjahjadi, M., Karavolias, N. G., ... & Niyogi, K. K. (2024). Multiplexed CRISPR-Cas9 mutagenesis of rice PSBS1 noncoding sequences for transgene-free overexpression. Science Advances 10(23): eadm7452.

92 Luo, G., & Palmgren, M. (2023). Fine-tuning of quantitative traits. Science China Life Sciences 66(6): 1456-1458.

93 Tang, X., & Zhang, Y. (2023). Beyond knockouts: fine-tuning regulation of gene expression in plants with CRISPR-Cas-based promoter editing. New Phytologist 239(3): 868-874.

94 Li, Y., & Wei, P. Editing of upstream

regulatory elements advances plant gene silencing. New Phytologist *in press*

95 Deng, K., Zhang, Q., Hong, Y., Yan, J., & Hu, X. (2023). iCREPCP: A deep learning-based web server for iden-tifying base-resolution cis-regulatory elements within plant core promoters. Plant Communications 4(1): 100455.

96 Zhou, J., Liu, G., Zhao, Y., Zhang, R., Tang, X., Li, L., ... & Zhang, Y. (2023). An efficient CRISPR–Cas12a promoter editing system for crop improvement. Nature Plants 9(4): 588-604.

97 Xue, C., Qiu, F., Wang, Y., Li, B., Zhao, K. T., Chen, K., & Gao, C. (2023). Tuning plant phenotypes by precise, graded downregulation of gene expression. Nature Biotechnology 41(12): 1758-1764.

98 Yasmeen, E., Wang, J., Riaz, M., Zhang, L., & Zuo, K. (2023). Designing artificial synthetic promoters for accurate, smart, and versatile gene expression in plants. Plant Communications 4: 100558.

99 Hu, X., Fernie, A. R., & Yan, J. (2023). Deep learning in regulatory genomics: from identification to design. Current Opinion in Biotechnology 79: 102887.

100 DaSilva, L. F., Senan, S., Patel, Z. M., Reddy, A. J., Gabbita, S., Nussbaum, Z., ... & Pinello, L. (2024). DNA-Diffusion: Leveraging generative models for controlling chromatin accessibility and gene expression via synthetic regulatory elements. bioRxiv.

101 Xia, Y., Du, X., Liu, B., Guo, S., & Huo, Y. X. (2024). Species-specific design of artificial promoters by transfer-learning based generative deep-learning model. Nucleic Acids Research gkae429.

102 Gosai, S. J., Castro, R. I., Fuentes, N., Butts, J. C., Kales, S., Noche, R. R., ... & Tewhey, R. (2023). Machine-guided design of synthetic cell type-specific cis-regulatory elements. bioRxiv.

103 Lal, A., Garfield, D., Biancalani, T., & Eraslan, G. (2024, April). regLM: Designing realistic regulatory DNA with autoregressive language models. In: International Conference on Research in Computational Molecular Biology (pp. 332-335). Cham: Springer Nature Switzerland.

104 Li, T., Xu, H., Teng, S., Suo, M., Bahitwa, R., Xu, M., ... & Wang, H. (2024). Modeling 0.6 million genes for the rational design of functional cis-regulatory variants and de novo design of cis-regulatory sequences. Proceedings of the National Academy of Sciences 121(26): e2319811121.

105 Van der Oost, J., & Patinios, C. (2023). The genome editing revolution. Trends in Biotechnology 41(3): 396-409.

106 McClelland, A. J., & Ma, W. (2024). Zig, Zag, and'Zyme: leveraging structural biology to engineer disease resistance. aBIOTECH 1-5.

107 Schuster, M., Eisele, S., Armas-Egas, L., Kessenbrock, T., Kourelis, J., Kaiser, M., & van der Hoorn, R. A. (2024). Enhanced late blight resistance by engineering an EpiC2B-insensitive immune protease. Plant Biotechno¬logy Journal 22(2): 284.

108 Luo, X., Cao, L., Yu, L., Gao, M., Ai, J., Gao, D., ... & Shang, Y. (2024). Deep learning-based characterization and redesign of major potato tuber storage protein. Food Chemistry 443: 138556.

109 Jafari, F., Wang, B., Wang, H., & Zou, J. (2023). Breeding maize of ideal plant architecture for high-density planting tolerance through modulating shade avoidance response and beyond. Journal of Integrative Plant Biology 66(5): 849-864.

110 Feldman, M., & Levy, A. A. (2023). Future prospects. In: Wheat Evolution and Domestication (pp. 665-673). Cham: Springer International Publishing.

111 Lokya, V., Parmar, S., Pandey, A. K., Sudini, H. K., Huai, D., Ozias-Akins, P., ... & Pandey, M. K. (2023). Prospects for developing allergen-depleted food crops. The Plant Genome 16(4): e20375.

112 Sojka, J., Šamajová, O., & Šamaj, J. (2024). Gene-edited protein kinases and phosphatases in molecular plant breeding. Trends in Plant Science 29(6): 694-710.

113 Chen, Y., Miller, A. J., Qiu, B., Huang, Y., Zhang, K., Fan, G., & Liu, X. (2024). The role of sugar transporters in the battle for carbon between plants and pathogens. Plant Biotechnology Journal *in press*

114 https://experiment.com/projects/bnioorxwrtxozfglnwqr

115 Rühle, T., Leister, D., & Pasch, V. (2024). Chloroplast ATP synthase: From structure to engineering. The Plant Cell koae081.

116 Roze, L. V., Antoniak, A., Sarkar, D., Liepman, A. H., Tejera-Nieves, M., Vermaas, J. V., & Walker, B. J. (2024). Advancing thermostability of the key photorespiratory enzyme glycerate 3-kinase by structure-based recombination. bioRxiv, 2024-05.

117 https://experiment.com/projects/jvqjplzolhyncosrbpob

118 Outram, M. A., Figueroa, M., Sperschneider, J., Williams, S. J., & Dodds, P. N. (2022). Seeing is believing: Exploiting advances in structural biology to understand and engineer plant immunity. Current Opinion in Plant Biology 67: 102210.

119 Joshi, A., Song, H. G., Yang, S. Y., & Lee, J. H. (2023). Integrated molecular and bioinformatics approaches for disease-related genes in plants. Plants 12(13): 2454.

120 Zhang, P., Wang, Y., Chachar, S., Tian, J., & Gu, X. (2020). eRice: a refined epigenomic platform for japonica and indica rice. Plant Biotechnology Journal 18(8): 1642.

121 Wang, Y., Zhang, P., Guo, W., Liu, H., Li, X., Zhang, Q., ... & Gu, X. (2021). A deep learning approach to auto¬mate whole-genome prediction of diverse epigenomic modifications in plants. New Phytologist 232(2): 880-897.

122 Sinha, D., Dasmandal, T., Paul, K., Yeasin, M., Bhattacharjee, S., Murmu, S., ... & Archak, S. (2023). MethSemble-6mA: an ensemble-based 6mA prediction server and its application on promoter region of LBD gene family in Poaceae. Frontiers in Plant Science 14: 1256186.

123 Cheng, Y., Zhou, Y., & Wang, M. (2024). Targeted gene regulation through epigenome editing in plants. Current Opinion in Plant Biology 80: 102552.

124 Subramanian, A. T., Roy, P., Aravind, B., Kumar, A. P., & Mohannath, G. (2024). Epigenome editing strategies for plants: a mini review. The Nucleus 67: 75-87.

125 Chen, L., Liu, G., & Zhang, T. (2024). Integrating machine learning and genome editing for crop improvement. aBIOTECH, 1-16.

126 Yang, L., Zhang, P., Wang, Y., Hu, G., Guo, W., Gu, X., & Pu, L. (2022). Plant synthetic epigenomic engineering for crop improvement. Science China Life Sciences 65(11): 2191-2204.

127 Dong, J., Croslow, S., Lane, S., Castro, D., Blanford, J., Zhou, S., ... & Hudson, M. (2024). Enhancing lipid production in plant cells through high-throughput genome editing and phenotyping via a scalable automated pipeline. bioRxiv, 2024-05.

128 Walker, A., Narváez-Vásquez, J., Mozoruk, J., Niu, Z., Luginbühl, P., Sanders, S., ... & Beetham, P. (2023). Industrial Scale Gene Editing in Brassica napus. International Journal of Plant Biology 14(4): 1064-1077.

129 Rigoulot, S. B., Park, J., Fabish, J., Seaberry, E. M., Parrish, A., Meier, K. A., ... & Dong, S. (2024). Enabling high-throughput transgene

expression studies using automated liquid handling for etiolated maize leaf protoplasts. Journal of Visualized Experiments 204: e65989.

130 Rigoulot, S. B., Barco, B., Zhang, Y., Zhang, C., Meier, K. A., Moore, M., ... & Que, Q. (2023). Automated, high-throughput protoplast transfection for gene editing and transgene expression studies. In: Plant Genome Engineering: Methods and Protocols (pp. 129-149). New York, NY: Springer US.

131 Waltz, E. (2017). Digital farming attracts cash to agtech startups. Nature Biotechnology 35(5): 397-398.

132 Waltz, E. (2019). With CRISPR and machine learning, startups fast-track crops to consume less, produce more. Nature Biotechnology 37(11): 1251-1253.

133 https://www.lens.org/lens/patent/147-221-689-821-247/frontpage

134 https://www.syngenta.com/en/company/media/syngenta-news/year/2024/syngenta-and-instadeep-collaborate-accelerate-crops-seeds

135 https://graphica.bio

136 https://tropic.bio

137 https://www.prnewswire.com/news-releases/evogene-amends-its-collaboration-agreement-with-bayer-to-include-genome-editing-targets-300885511.html

138 https://leaps.bayer.com/companies/agriculture

139 https://www.lens.org/lens/patent/031-797-944-942-034/frontpage

140 Waltz, E. (2019). With CRISPR and machine learning, startups fast-track crops to consume less, produce more. Nature Biotechnology 37(11): 1251-1253.

141 https://tracxn.com

142 https://www.crunchbase.com

143 https://pitchbook.com

144 https://tracxn.com/d/companies/inari/__GeBti0I5FOhQvboXpyDz1lWejC1W32h0_edfpL-ySkI

145 USDA (2024). RE: Regulatory Status Review of soybean developed using genetic engineering for enhanced yield, and changes to plant architecture and development. https://www.aphis.usda.gov/sites/default/files/23-132-01rsr-response.pdf

146 USDA (2024). RE: Regulatory Status Review of corn developed using genetic engineering for enhanced yield traits. https://www.aphis.usda.gov/sites/default/files/23-040-01rsr-response.pdf

147 USDA (2023). RE: Regulatory Status Review of maize developed using genetic engineering for altered plant height. https://www.aphis.usda.gov/sites/default/files/23-101-01rsr-review-response.pdf

148 https://www.reuters.com/markets/commodities/australian-trial-gene-edited-wheat-aims-10-bigger-yields-2024-05-23/

149 European Comission (2024). Information on the notifications submitted under Directive 2001/18/EC. Part B – GM plants: Notification B/Be/23/V4. https://webgate.ec.europa.eu/fip/GMO_Registers/GMO_Part_B_Plants.php

150 https://www.lens.org/lens/patent/147-494-399-043-305/fulltext?l=en

151 https://www.phytoformlabs.com/technology

152 https://genxtraits.com

153 https://www.lens.org/lens/patent/173-231-633-687-511/frontpage

154 https://www.lens.org/lens/patent/019-427-725-925-397/frontpage?l=en

155 https://tracxn.com

156 https://agfundernews.com/armed-with-100m-in-funding-dave-friedberg-unveils-boosted-breeding-tech-at-ohalo-in-holy-shit-moment-for-crop-breeders

157 https://cloud.google.com/blog/topics/startups/ai-startups-at-next24?hl=en

158 USDA (2023). RE: Regulatory Status Review of potato developed using genetic engineering for reduced glucose and fructose content in tubers. https://www.aphis.usda.gov/sites/default/files/23-081-01rsr-review-response.pdf

159 USDA (2023). RE: Regulatory Status Review of potato developed using genetic engineering for increased beta-carotene in tubers. https://www.aphis.usda.gov/sites/default/files/22-224-01rsr-review-response.pdf

160 https://www.lens.org/lens/patent/000-280-791-396-901/frontpage

161 https://www.science.org/content/article/genetically-edited-wood-could-make-paper-more-sustainable

162 Oliveira, D. M., & Cesarino, I. (2023). Genome editing of wood for sustainable pulping. Trends in Plant Science 29(2): 111-113.

163 Sulis, D. B., Jiang, X., Yang, C., Marques, B. M., Matthews, M. L., Miller, Z., ... & Wang, J. P. (2023). Multiplex CRISPR editing of wood for sustainable fiber production. Science 381(6654): 216-221.

164 https://innovation.ox.ac.uk/news/wild-bioscience-transforming-agriculture-through-innovation/

165 https://medium.com/future-literacy/thinking-outside-of-the-evolutionary-box-how-arzeda-is-re-imagining-proteins-the-building-blocks-79a1301c06dc

166 Eisenstein, M. (2023). AI-enhanced protein design makes proteins that have never existed. Nature Biotechnology 41(3): 303.

167 https://www.forbes.com/sites/johncumbers/2019/11/26/molecule-maker-arzeda-wants-to-grow-phone-screens-that-wont-scratch/?sh=295987046785

168 https://www.wsj.com/articles/ai-accelerates-ability-to-program-biology-like-software-9962a975

169 https://www.ginkgobioworks.com/2022/10/18/ag-biologics-division-bayer-joyn/

170 https://www.plantae.net

171 https://www.ukko.us

172 https://www.bayer.com/media/losung-fur-lebensmittelallergien-ukko-erhalt-40-millionen-us-dollar-in-series-b-finanzierungsrunde-mit-leaps-by-bayer-als-leadinvestor/

173 Duan, Z., Liang, Y., Sun, J., Zheng, H., Lin, T., Luo, P., ... & Zhu, J. K. (2024). An engineered Cas12i nuclease that is an efficient genome editing tool in animals and plants. The Innovation 5(2): 100564

174 Han, X., Chen, Y., Liu, R., Zhu, J. K., & Duan, Z. (2024). Engineering hfCas12Max for improved gene editing efficiency. The Innovation Life 2(2): 100068.

175 Xie, H., Su, F., Niu, Q., Geng, L., Cao, X., Song, M., ... & Zhu, J. (2024). Knockout of miR396 genes increases seed size and yield in soybean. Journal of Integrative Plant Biology in press

176 Niu, Q., Xie, H., Cao, X., Song, M., Wang, X., Li, S., ... & Zhu, J. (2024). Engineering soybean with high levels of herbicide resistance with a Cas12-SF01-based cytosine base editor. Plant Biotechnology Journal in press

177 Huang, J., Lin, Q., Fei, H., He, Z., Xu, H., Li, Y., ... & Gao, C. (2023). Discovery of deaminase functions by structure-based protein clustering. Cell, 186(15), 3182-3195.

178 Dixon, T. A., Williams, T. C., & Pretorius, I. S. (2021). Sensing the future of bio-informational engineering. Nature Communications 12(1): 388.

179 Correia, P. M., Najafi, J., & Palmgren, M. (2024). De novo domestication: what about the weeds?. Trends in Plant Science *in press*

180 https://www.scanthehorizon.org/p/dnai-the-artificial-intelligence

181 Wilke, C. (2023). Remote sensing for crops spots pests and pathogens. ACS Central Science 9: 339-342.

182 USDA (2023). RE: Regulatory Status Review of soybean developed using genetic engineering for inducible expression of a fluorescent protein and an antibiotic marker gene. https://www.aphis.usda.gov/sites/default/files/22-235-01rsr-review-response.pdf

183 USDA (2023). RE: Regulatory Status Review of soybean developed using genetic engineering for expression of a fluorescent protein and an antibiotic marker gene. https://www.aphis.usda.gov/sites/default/files/22-276-01rsr-review-response.pdf

184 USDA (2023). RE: Regulatory Status Review of tomato developed using genetic engineering for expression of a fluorescent protein and an antibiotic marker gene. https://www.aphis.usda.gov/sites/default/files/22-276-02rsr-review-response.pdf

185 USDA (2023). RE: Regulatory Status Review of corn developed using genetic engineering to produce anthocyanins in response to pathogen infection, and to have a disrupted pathogen-responsive gene. https://www.aphis.usda.gov/sites/default/files/23-087-01rsr-review-response.pdf

186 https://food.ec.europa.eu/plants/genetically-modified-organisms/new-techniques-biotechnology_en#commission-proposal-on-plants-obtained-by-certain-new-genomic-techniques

187 Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. Nature, 627(8002), 49-58.

188 Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9).

189 Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. Nature Reviews Physics, 5(5), 277-280.

190 He, J., Feng, W., Min, Y., Yi, J., Tang, K., Li, S., ... & Zheng, S. (2023). Control risk for potential misuse of artificial intelligence in science. arXiv preprint arXiv:2312.06632.

191 Undheim, T. A. (2024). The whack-a-mole governance challenge for AI-enabled synthetic biology: literature review and emerging frameworks. Frontiers in Bioengineering and Biotechnology, 12, 1359768.

192 He, Y., Zhou, X., Chang, C., Chen, G., Liu, W., Li, G., ... & Chang, X. (2024). Protein language models-assisted optimization of an uracil-N-glycosylase variant enables programmable T-to-G and T-to-C base editing. Molecular Cell 84(7): 1257-1270.

193 Ruffolo, J. A., Nayfach, S., Gallagher, J., Bhatnagar, A., Beazer, J., Hussain, R., ... & Madani, A. (2024). De-sign of highly functional genome editors by modeling the universe of CRISPR-Cas sequences. bioRxiv, 2024-04.

194 Callaway, E. (2024). 'ChatGPT for CRISPR'creates new gene-editing tools. Nature, 629(8011), 272-272.

195 https://www.together.ai/blog/evo

196 Li, B., Sun, C., Li, J., & Gao, C. (2024). Targeted genome-modification tools and their advanced applications in crop breeding. Nature Reviews Genetics *in press*

197 Pei, Q., Wu, L., Gao, K., Zhu, J., Wang, Y., Wang, Z., ... & Yan, R. (2024). Leveraging Biomolecule and Natural Language through Multi-Modal Learning: A Survey. arXiv preprint arXiv:2403.01528.

198 Lam, H. Y. I., Ong, X. E., & Mutwil, M. (2024). Large language models in plant biology. Trends in Plant Science in press

199 Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., ... & Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. Nature, 620(7972), 47-60.

200 Verspoor, K. (2024). 'Fighting fire with fire' – using LLMs to combat LLM hallucinations. Nature 630(8017): 569-570.

201 Vindman, C., Trump, B., Cummings, C., Smith, M., Titus, A. J., Oye, K., ... & Linkov, I. (2024). The convergence of AI and Synthetic Biology: The looming deluge. arXiv preprint arXiv:2404.18973.

202 https://www.lobbycontrol.de/wp-content/uploads/Study_en_LobbyNetwork_31.8.2021.pdf

203 Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. Nature 627(8002): 49-58.

204 https://www.cbd.int/synbio/current_activities/open-ended_online_forum/november_2023?threadid=3038

205 https://huggingface.co/InstaDeepAI/agro-nucleotide-transformer-1b

206 https://github.com/benlevyx/florabert

207 Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: open washing and the EU AI Act. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1774-1787).

208 Gibney, E. (2024). Not all 'open source' AI models are actually open. Nature News 19 June 2024. https://www.nature.com/articles/d41586-024-02012-5

209 Lin, F. (2024). AlphaFold 3 angst: Limited accessibility stirs outcry from researchers. GEN Biotechnology 3(3): 103-106.

210 Callaway, E. (2024). Who will make AlphaFold3 open source? Scientists race to crack AI model. Nature 630(8015): 14-15.

211 https://zenodo.org/records/11206103

212 https://saifood.ca/google-crops/

213 Winter, G. (2024). The European Union's deregulation of plants obtained from new genomic techniques: a critique and an alternative option. Environmental Sciences Europe 36(1): 47.

214 Zentrale Kommission für die Biologische Sicherheit (2023). Statement of the ZKBS on the proposal of the European Commission to re-regulate plants bred with «New Genomic Techniques (NGT)». https://www.zkbs-online.de/ZKBS/EN/Commentaries/03_Kommissionsentwurf%20Neuregulierung%20NGT/Kommissionsentwurf%20Neuregulierung%20NGT_node.html

215 Bohle, F., Schneider, R., Mundorf, J., Zühl, L., Simon, S., & Engelhard, M. (2024). Where does the EU-path on new genomic techniques lead us?. Frontiers in Genome Editing, 6, 1377117.

216 EFSA Panel on Genetically Modified Organisms (GMO), Mullins, E., Bresson, J. L., Dalmay, T., Dewhurst, I. C., Epstein, M. M., ... & Raffaello, T. (2024). Assessment of genetically modified maize DP202216 for food and feed uses, under Regulation (EC) No 1829/2003 (application EFSA-GMO-NL-2019-159). EFSA Journal, 22(3), e8655.


SOS
save our seeds