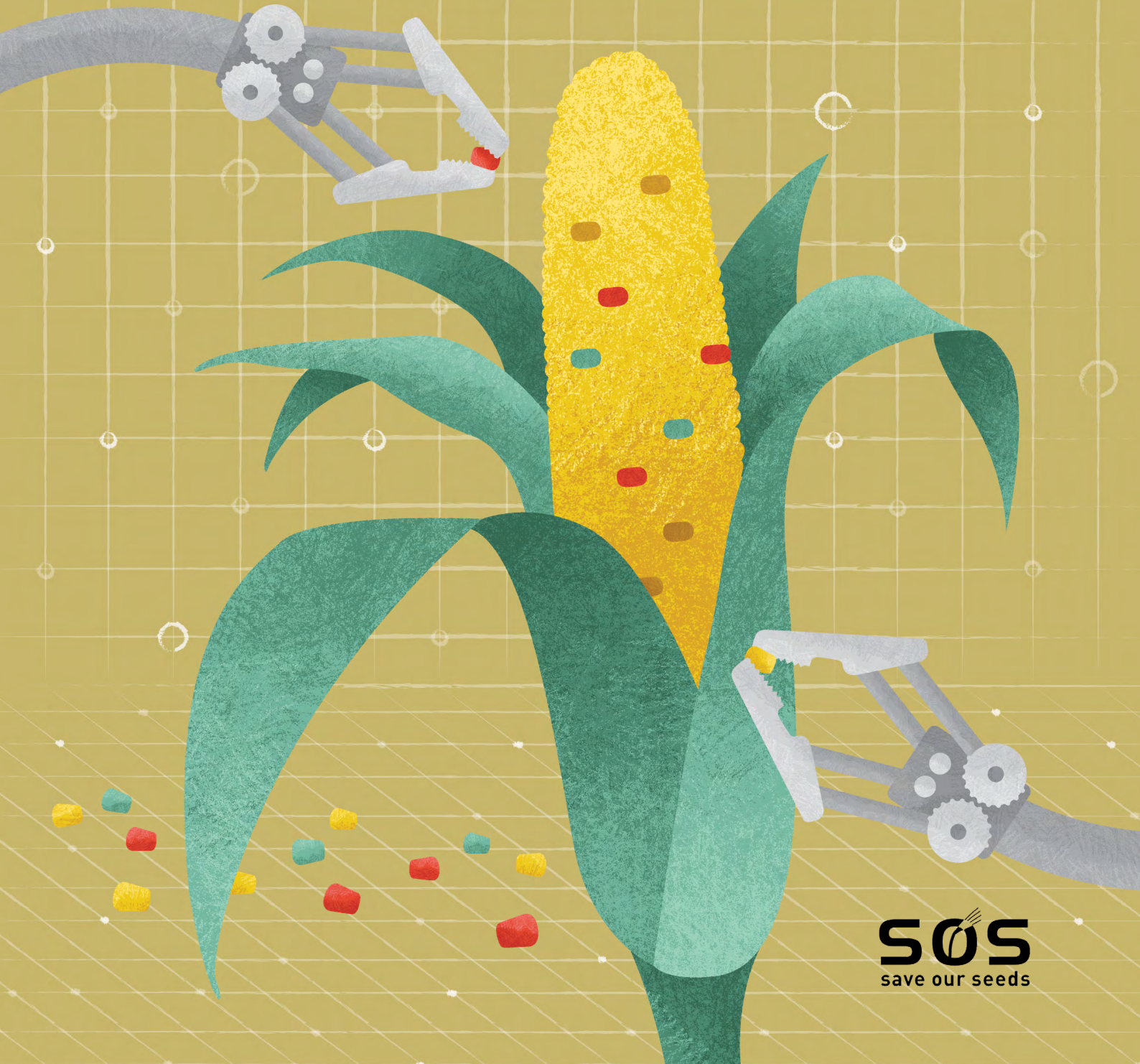
Wenn Chatbots neue Sorten züchten

Generative Künstliche Intelligenz und neue Gentechniken



SOS
save our seeds

Herausgeber:
Save Our Seeds / Zukunftsstiftung Landwirtschaft
Marienstr. 19-20
10117 Berlin
Telefon: +49 30-28482326
E-Mail: info@saveourseeds.org

Autor:
Benno Vogel
www.bennovogel.eu

Layout und Gestaltung:
Beatriz Francisco
www.linkedin.com/in/beatriz-francisco/

Veröffentlichung:
Januar 2025

SOS
save our seeds

Inhaltsverzeichnis

Inhaltsverzeichnis

Generative KI und Gentechnik

Zusammenfassung von Save Our Seeds

Der Einsatz generativer künstlicher Intelligenz (KI) hat eine tiefgreifende Transformation der Biotechnologie eingeleitet und verändert auch grundlegend den Einsatz von Gentechnik an Pflanzen. Einerseits kann KI die Präzision und Effizienz der CRISPR-basierten Gentechnik steigern und über die bisher üblichen Gen-Knockouts hinaus deutlich erweitern. Andererseits ist die KI-gesteuerte Gentechnik anfällig für die bekannten Risiken der KI, wie etwa den Black-Box-Effekt, Halluzinationen und Datenverzerrungen. Dadurch entstehen neue Möglichkeiten, gentechnisch veränderte Organismen mit unerwünschten Eigenschaften zu schaffen und in die Natur freizusetzen. Der Bericht **Wenn Chatbots neue Sorten züchten** *fasst den Stand der Technik in Bezug auf die Pflanzenzüchtung zusammen. Wie sollten Wissenschaft und Gesetzgebung in der EU mit den sich abzeichnenden neuen Herausforderungen umgehen?*

KI-Modelle, die in den ‚Sprachen' der Biologie geschult sind

Entwickler:innen passen die KI-Architekturen von Diffusionsmodellen und großen Sprachmodellen, wie sie in Chatbots wie ChatGPT oder Bildgeneratoren wie DALL-E verwendet werden, an die „Sprachen" der Biologie an und trainieren sie mit gigantischen Datenmengen von Protein- und Genomsequenzen.

Diese Entwicklung wurde durch die enorme Menge an Daten zu DNA- und RNA-Sequenzen, Proteinen und Metaboliten möglich, die in den letzten Jahren verfügbar wurde. Diese Daten bilden nun das Rohmaterial, das die Entwicklung generativer KI für die Gentechnik ermöglicht.

Die resultierenden KI-Werkzeuge sind sowohl deskriptiv als auch generativ. Wie herkömmliche Deep-Learning-Algorithmen können sie biologische Daten analysieren und Vorhersagen treffen. Darüber hinaus ermöglichen sie das Design funktionaler DNA-, RNA- und Proteinsequenzen, einschließlich „new-to-nature"-Sequenzen, die in der Natur so noch nicht vorkommen.

Je nach der Art der „Sprache“, werden verschiedene Modelle unterschieden:

- **Proteinmodelle:** Obwohl der Einsatz von KI-Modellen, die mit Proteindaten trainiert wurden, eine relativ neue Entwicklung ist, wächst die Zahl solcher Tools rasant. Diese Modelle können Proteine analysieren, ihre Interaktionen simulieren und ihre Funktionen neu gestalten. Das berühmteste Protein-Tool ist *Alphafold* von Google. Demis Hassabis, der Leiter der KI-Abteilung von Google, wurde für die Entwicklung dieses Modells gemeinsam mit zwei Kollegen mit dem Nobelpreis für Chemie 2024 ausgezeichnet.

- **DNA-Modelle:** Seit 2021 werden große Sprachmodelle mit riesigen Mengen an DNA-Sequenzen trainiert, um dadurch die „Sprache“ der Genome zu simulieren. Hervorzuheben sind vier Modelle, die spezifisch mit DNA-Se¬quenzen von Pflanzen trainiert wurden. Das bisher leistungsstärkste Sprachmodell für Pflanzengenome, *AgroNT* von Google und Instadeep, wurde Ende 2023 veröffentlicht und mit 10 Millionen Erbgutsequenzen von 48 Pflan¬zenarten trai-niert.

- **RNA-Modelle:** KI-Modelle, die mit menschlichen RNA-Sequenzen trainiert wurden, sind bereits im Einsatz. Es ist zu erwarten, dass es bald auch große Sprachmodelle geben wird, die auf RNA-Sequenzen von Pflanzen beruhen. Als vielversprechend für die Pflanzenwissenschaften und -züchtung gelten vor allem Modelle wie scGPT, die auf Einzelzell-RNA-Sequenzierungsdaten (scRNA-seq) basieren.

- **Multimodale Modelle:** Während bisherige Sprach- und Diffusionsmodelle noch auf einzelne Datentypen beschränkt sind, arbeiten KI-Firmen jetzt an Modellen, die mehrere Arten von Daten verarbeiten. Im Jahr 2024 stellten Instadeep und BioNTech die erste multimodale KI-Architektur zur Verbindung von DNA-, RNA- und Protein-Daten vor.

Einsatz von KI in der Gentechnik an Pflanzen

Die sogenannte Genom-Editierung stützt sich heute hauptsächlich auf die CRISPR-Cas-Methode. Spezifische KI-Tools sind verfügbar, um diesen Prozess zu verbessern. Sie unterstützen Forschende dabei, optimale Zielorte zu finden, die effektivsten Sequenzen für die Leit-RNA zu identifizieren und die am besten geeigneten CRISPR-Schneideenzyme auszuwählen. Der

Einsatz dieser Tools kann die Genom-Editierung mit CRISPR präziser und effizienter machen.

Außerdem haben KI-Tools dazu beigetragen, die Fähigkeiten von CRISPR über herkömmliche Anwendungen hinaus zu erweitern. For¬schende schalten Gene nicht mehr einfach aus (Knockout), sondern steuern nun deren Expression, indem sie gezielt Sequenzen des regulatorischen Netzwerkes verändern (quantitative Trait-Engineering). Die gezielte Steuerung der Gen¬expression soll die Beeinflussung komplexer quantitativer Merkmale ermöglichen.

Hier sind einige Beispiele, wie KI-Modelle die gentechnische Manipulation von Pflanzen verändern:

- Das US-Unternehmen TreeCo möchte **Pappeln** so verändern, dass sie weniger Lignin bilden, und dadurch die Papierherstellung erleichtert wird. Das Unternehmen hat ein eigenes KI-Tool entwickelt, das vorhersagt, wie sich Veränderungen in den 21 Genen, die an der Ligninsynthese beteiligt sind, auf die Holzzusammensetzung, die Wachstumsrate und andere Merkmale der Bäume auswirken. Für diese 21 Gene hat das Tool über 69.000 potenzielle Editierungsstrategien gefunden und mit einer Computersimulation nach den besten Strategien gesucht. Die Firma hat schließlich die sieben vielversprechendsten Kombinationen von Genveränderungen im Pappelerbgut erzeugt.

- Ein weiteres US-Unternehmen, Inari, entwickelt **Maispflanzen** mit reduzierter Höhe und erhöhter Blattbiomasse. Das Unternehmen nutzt ein generatives KI-Tool, das voraussagen soll, wie sich Mutationen in Promotoren auf die Eigen-schaften einer Pflanze auswirken. In Belgien testet Inari einen kurzwachsenden Mais im Freiland.

- Forschende haben das Alphafold-Proteinmodell genutzt, um Patatin – ein Protein, das natürlicherweise in **Kartoffeln** vorkommt – neu zu gestalten. Am Computer ist eine Version entstanden, die laut KI die Viskosität und die ernährungsphysiologischen Eigenschaften von Kartoffelmehlteig verbessern soll. Die Forschenden wollen die KI-generierte Patatin-Version nun im Erbgut von Kartoffeln erzeugen.

Große Saatgutfirmen wie Corteva, Bayer, BASF und Syngenta setzen zunehmend KI-Tools in ihren Gentechnikprogrammen ein. Dabei gehen sie häufig Partner-schaften mit spezialisierten KI-Firmen ein. So haben BASF und Corteva jeweils Kooperationen mit der Firma Tropic Biosciences gestartet, die über eine proprietäre KI-Technologie verfügt. Syngenta hat sich mit Instadeep

und Biographica zusammengetan, während Bayer Startups wie Ukko und Amfora unterstützt, die beide für die Entwicklung neuer Sorten auf den kombinierten Einsatz von KI und CRISPR setzen.

Was kommt als Nächstes?

Die Entwicklung generativer KI-Modelle für die Genom-Editierung steckt noch in den Kinderschuhen. Viele der derzeit verfügbaren Design-Tools sind so neu, dass noch nicht die nötigen experimentellen Daten vorliegen, um die Leistung ihrer Algorithmen zu bewerten. Es ist jedoch bereits jetzt erkennbar, dass diese Tools neue Designmöglichkeiten schaffen, die auch über natürliche Grenzen hinausgehen.

In den kommenden Jahren wird erwartet, dass sich die Qualität der Datenerhebungstechniken, der Umfang der gesammelten Daten und die Rechenleistung zu deren Verarbeitung exponentiell erhöhen werden. Die deskriptiven und generativen Fähigkeiten der KI verbessern sich ständig. Erfahrungen mit großen Sprachmodellen, die mit mikrobiellem DNA-Material trainiert wurden, zeigen das Potenzial, das genomische KI-Tools haben könnten. Ein solches Modell, EVO, hat laut seinen Entwicklern das Potential, Sequenzen in der Größenordnung ganzer mikro¬bieller Genome zu erzeugen.

Wie in vielen anderen Bereichen steht zu erwarten, dass diese Fortschritte die Life Sciences insgesamt und die Pflanzenzüchtung tiefgreifend verändern werden.

Was könnte schiefgehen?

Die Integration von KI in die Gentechnik wirft eine Reihe von Bedenken auf. Viele Aspekte werden auch in anderen Bereichen diskutiert, in denen die generative KI zum Einsatz kommt. Dazu gehören unter anderem:

- **Niedrigere Qualifikationsschwelle.** Bislang ist die gentechnische Veränderung von Pflanzen hochqualifizierten Fachleuten vorbehalten, die umfassend in molekularbiologischen Techniken

geschult sind. Mit dem Aufkommen von KI-Tools könnte die Gentechnik zunehmend auch für Studierende, Informatiker:innen, Unternehmer:innen oder sogar Hobby-Biolog:innen zugänglich werden.

- **Black Box.** Generative KI-Modelle liefern Vorhersagen oder machen Empfehlungen, ohne dass nachvollziehbar wäre, wie und warum sie zu diesen Ergebnissen kommen. In sensiblen Bereichen wie der Pflanzen-Gentechnik, deren Produkte sich fortpflanzen, in der Natur interagieren und die Gesundheit vieler Menschen und der Umwelt tangieren, ist der Mangel an Nachvollziehbarkeit und Reproduzierbarkeit besonders problematisch.

- **Halluzinationen.** Generative KI-Modelle können Ergebnisse liefern, die plausibel erscheinen, aber sachlich falsch oder irrelevant sind. Wie oft und in welchen Zusammenhängen KI-Modelle „halluzinieren" und wie man dem entgegenwirken kann, muss noch ermittelt werden.

- **Datenverzerrungen.** Die Outputs und Vorhersagen von generativen KI-Modellen spiegeln immer die Daten wider, mit denen die Modelle trainiert wurden. Wenn diese Daten Fehler oder Verzerrungen enthalten, die von den zugrundeliegenden biologischen Systemen oder von den menschlichen Kuratoren stammen, können sich diese Verzerrungen auf die Ergebnisse des Modells übertragen.

Der Mangel an spezialisiertem Fachwissen, in Verbindung mit der Black-Box-Problematik, Halluzinationen und möglichen Datenfehlern lässt befürchten, dass Pflanzen mit unerwünschten Eigenschaften entwickelt und in die Umwelt freigesetzt werden könnten. Daher gilt es, mit Vorsicht voranzugehen und strenge Aufsichtsmechanismen zu entwickeln.

EU plant, die Regulierung von KI-designten Pflanzen zu lockern

In diesem kritischen Moment will die EU nun die regulatorischen Anforderungen an die Kommerzialisierung von gentechnisch veränderten Pflanzen erheblich lockern. In einem Vorschlag vom Juli 2023 schlägt die Europäische Kommission vor, Pflanzen, die mit Verfahren wie CRISPR gentechnisch verändert wurden, ähnlich zu behandeln wie konventionell gezüchtete Pflanzen. Konkret würden Pflanzen mit bis zu 20 gentechnischen

Veränderungen ihres Erbguts vom EU-Gentechnikrecht ausgenommen. Der Kommission zufolge könnten diese Pflanzen ohne Risikoprüfung, Nachweismethode, Rückverfolgbarkeit oder Kennzeichnungspflicht auf den Markt gebracht werden.

Zahlreiche Wissenschaftler:innen, Behörden und NGOs haben den Vorschlag der Kommission kritisiert. Das Bundesamt für Naturschutz (BfN) in Deutschland wies darauf hin, dass die Mehrzahl der genomeditierten Pflanzen ohne Risikobewertung in die Umwelt freigesetzt würden und warnte davor, dass auch geringfügige Veränderungen des Genoms hohe Risiken bergen können. Die französische Lebensmittelbehörde ANSES argumentierte, dass der Schwellenwert von 20 Nukleotiden nicht geeignet sei, um die Äquivalenz zu konventionell gezüchteten Pflanzen nachzuweisen. Die Europäische Lebensmittelbehörde (EFSA) verteidigt jedoch den Vorschlag der Kommission.

Die Konvergenz von KI und Gentechnik könnte die bestehende Problematik erheblich verschärfen. Der Einsatz generativer KI-Modelle könnte es Entwickler:innen ermöglichen, den „Designraum" von 20 genetischen Veränderungen vollständig auszuschöpfen. Dies könnte zu der – absichtlichen oder unbeabsichtigten – Schaffung von Pflanzen führen, die für den Menschen und die Umwelt gefährlich sind. So könnten Forschende zum Beispiel eine Pflanze entwickeln, die eine Vielzahl von Insektengiften produziert. Nach dem Vorschlag der Europäischen Kommission wären jedoch keine Tests erforderlich, um die Auswirkungen der Pflanze auf Nicht-Zielarten zu prüfen.

Wie weiter?

Anstatt die GVO-Vorschriften zu lockern, sollte die EU die grundlegenden Anforderungen ihrer Gentechnik-Gesetze auch für Pflanzen aufrechterhalten, die mit den neuesten Technologien entwickelt wurden. Die Risikoprüfung sollte so angepasst werden, dass sie die spezifischen Merkmale dieser neuen Verfahren und Technologien berücksichtigt.

Darüber hinaus sollte die EU Schritte unternehmen, um die KI-gestützte Gentechnik wirksam zu regulieren. Klare Vorschriften sollten dafür sorgen, dass die verwendeten KI-Modelle zuverlässig und in der Lage sind, sichere Empfehlungen abzugeben, während Verständnis, Aufsicht und Entscheidungsfindung durch den Menschen in kritischen Phasen des Gentechnikprozesses gewahrt bleiben.

Die Kontrolle von KI-Tools und -Technologien, die in der Gentechnik eingesetzt werden, sowie von KI-generierten künstlichen Organismen, ist für die Sicherheit der Forschung und Entwicklung in diesem Bereich entscheidend. Risikobewertung, Monitoring, Rückverfolgbarkeit und Rückholbarkeit sollten Mindestanforderungen sein, bevor solche Organismen in die Umwelt freigesetzt werden.

Eine internationale Aufsicht sollte die Schaffung neuer Organismen oder genetischer Materialien verhindern, die pathogen sind oder andere schwerwiegende Bedrohungen darstellen. Der Zugang zu Hochrisikotechnologien, -tools und genetischen Daten, die anfällig für Missbrauch sein könnten, muss streng kontrolliert werden.

Die biologische Sicherheit muss ein integraler Bestandteil aller Forschungsaktivitäten sein, unabhängig davon, ob die Projekte von privaten Unternehmen oder öffentlichen Forschungseinrichtungen durchgeführt werden. Bei Unsicherheiten bezüglich potenziell hoher Risiken sollten Alternativen mit geringeren Risiken vorgezogen werden.

Schließlich sollte die unabhängige Forschung zu den Risiken der Gentechnik mit öffentlichen Mitteln unterstützt werden, mit einem besonderen Augenmerk auf die KI-gesteuerte Gentechnik. Besondere Aufmerksamkeit sollte den systemischen und langfristigen Auswirkungen gewidmet werden, die über den Rahmen einzelner Projekte hinausgehen (Technikfolgenabschätzung).

Abkürzungsverzeichnis

AMP	Antimikrobielle Proteine
CRE	Cis-regulatorisches Element
EU	Europäische Union
GEIGS	Gene Editing Induced Gene Silencing
KI	Künstliche Intelligenz
KMU	Kleinere und mittlere Unternehmen
miRNA	mikroRNA
NGT	Neue genomische Techniken
RNAi	RNA Interferenz
scRNA-Seq	Single-cell RNA sequencing
siRNA	small interfering RNA
uORF	Upstream Open Reading Frame

1. Einleitung

Immer öfter kreieren universitäre Labore, Startups und Tech-Giganten wie Meta, Google und Microsoft Werkzeuge der generativen Künstlichen Intelligenz (KI) für die Bio- und Gentechnologie. Sie nehmen dazu die KI-Architekturen der Diffusions- und großen Sprachmodelle, die in Chatbots wie ChatGPT oder Bildergeneratoren wie DALL-E stecken, und trainieren sie in den „Sprachen" der Biologie – mit Protein- und Genomsequenzen. Dabei entstehen Tools, die die Art und Weise, wie mit Gentechnik ins Erbgut von Organismen eingegriffen wird, stark verändern. Ausgestattet mit verbesserten deskriptiven Fähigkeiten ermöglichen es die neuen KI-Modelle, die Auswirkungen gentechnischer Eingriffe am Computer zu simulieren. Dank ihrer generativen Fähigkeiten können die KI- Modelle sogar funktionale DNA- und RNA-Sequenzen sowie Proteine entwerfen, die die Evolution noch nicht hervorgebracht hat und die – so der Fachjargon – „new-to-nature" sind.

Während die generative KI auch in der gentechnischen Pflanzenzüchtung Einzug hält, ist die EU im Begriff, die Regulierung von gentechnisch veränderten Pflanzen zu lockern, die mit neueren Methoden der Gentechnik – so genannten neuen genomischen Techniken (NGT) – hergestellt werden. Seit Juli 2023 liegt ein Gesetzesentwurf der EU-Kommission vor. Er unterteilt Pflanzen, die mit Genomeditierung erzeugt werden und kein genetisches Material von außerhalb ihres züchterischen Genpools enthalten, in zwei Kategorien: Genomeditierte Pflanzen, die bis zu 20 gezielte Veränderungen in ihrem Erbgut enthalten, bilden die Kategorie 1 (NGT1-Pflanzen). Genomeditierte Pflanzen mit mehr als 20 gezielten Veränderungen bilden die Kategorie 2 (NGT2-Pflanzen). Da die EU-Kommission davon ausgeht, dass die Risikoprofile von NGT1-Pflanzen und herkömmlich gezüchteten Pflanzen vergleichbar sind, schlägt sie vor, NGT1-Pflanzen von den Anforderungen der GVO-Rechtsvorschriften auszunehmen und den geltenden Bestimmungen für herkömmlich gezüchteten Pflanzen zu unterstellen. NGT2-Pflanzen hingegen sollen im Regulierungsbereich des Gentechnikrechts bleiben.

Damit EU-Parlament und EU-Ministerrat den Gesetzesentwurf zur Regulierung von NGT-Pflanzen fundiert diskutieren können, hat die EU-Kommission den politischen Entscheidungsträger:innen eine Reihe von Dokumenten vorgelegt: eine Folgenabschätzung, Fallstudien der Gemeinsamen Forschungsstelle (JRC), Arbeiten der Europäische Behörde für Lebensmittelsicherheit (EFSA) sowie die Resultate einer Stakeholderbefragung. Was in diesen Dokumenten und in

der laufenden politischen Debatte zur Regulierung von NGT-Pflanzen jedoch unberücksichtigt bleibt, ist die Konvergenz von Genomeditierung und generativer KI, wie sie derzeit in den Laboren der molekularen Pflanzenzucht stattfindet. Was bedeutet dies für eine zukunftssichere Regulierung von NGT-Pflanzen? Die Frage stellt sich umso dringlicher, als bei NGT1-Pflanzen zur Diskussion steht, vorsorgliche Maßnahmen wie Risikoprüfung und Rückverfolgbarkeit aufzuheben.

Weil die Konvergenz von NGT und generativer KI in der Diskussion zur geplanten Deregulierung von NGT1-Pflanzen bisher praktische keine Rolle spielte, hat die Initiative Save Our Seeds die vorliegende Arbeit in Auftrag gegeben. Sie soll einen Einblick in die Entwicklung von Protein- und Genom-basierten Diffusions- und großen Sprachmodellen geben, die bei der Genomeditierung von Pflanzen zum Einsatz kommen könnten, und die regulatorischen Fragen darstellen, die die Konvergenz von Genomeditierung und generativer KI bei der Herstellung von NGT-Pflanzen mit sich bringt.

Die mit einer Literatur- und Internetrecherche dazu gesammelten Informationen sind in folgender Reihenfolge dargestellt: Zuerst werden in Kapitel 2 kurz die biologischen Daten dargestellt, die für das Training der Modelle der generativen KI zur Verfügung stehen. Kapitel 3 beschreibt die Entwicklung bei generativen KI-Modellen, die an Proteinen und Genomen geschult sind und für den Einsatz in der NGT-basierten Pflanzenzüchtung in Frage kommen. Anschließend wird beschrieben, wie Forschung (Kapitel 4), Konzerne (Kapitel 5) sowie kleinere und mittlere Unternehmen (Kapitel 6) KI-Modelle verwenden, wenn sie das Erbgut von Pflanzen verändern. Kapitel 7 beleuchtet schließlich, welche Fragen und Herausforderungen sich bei der Regulierung von NGT1-Pflanzen stellen.

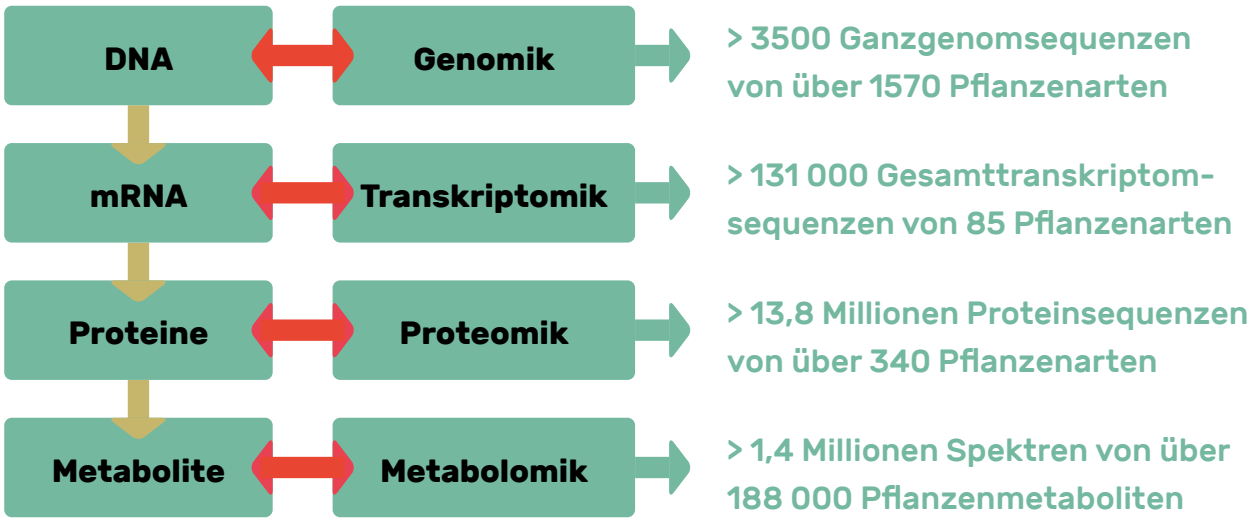
2. Big Data – die Rohstoffe für die generative KI

Genomik, Transkriptomik, Proteomik und Metabolomik – sie haben in den letzten Jahren zu einem immensen Schatz an Daten zu DNA- und mRNA-Sequenzen, Proteinen und Metaboliten von Pflanzen geführt (Abbildung 1).[1] 13,82 Millionen Proteinsequenzen von 342 Pflanzenarten sind beispielsweise in der Datenbank PlantMWpIDB enthalten.[2] Auf der Plattform PlantExp finden sich 131.400 Gesamttranskriptomsequenzen mit 572,4 Tera-Basen von 85 Pflanzenarten.[3] Im Pflanzenmetabolom-Hub PMhub wiederum sind 188.837

Stoffwechselprodukte von Pflanzen chemisch beschrieben.[4] Diese und weitere Daten sind heute der Rohstoff, der die Entwicklung generativer KI für die NGT-basierte Pflanzenzüchtung erst möglich macht. Denn ohne Big Data gibt es auch keine generative KI: Die modernen Algorithmen müssen mit riesigen Datensätzen trainiert werden und sind in der Regel umso leistungsfähiger je größer die Datensätze sind.

So riesig die Menge an Daten über Pflanzen heute bereits ist, so gewiss ist es, dass ihrer Gewinnung in den kommenden Jahren eine exponentielle quantitative aber auch qualitative Entwicklung bevorsteht.

Abbildung 1: Omik-Techniken, ihre Beziehung zu DNA, RNA, Proteinen und Metaboliten sowie die mit ihnen gewonnenen Daten.



2.1 Genome, Pangenome und Super-Pangenome

Daten zu Genomen von Pflanzen sind in zweifacher Hinsicht unerlässlich für die Konvergenz von KI und NGT. Erstens sind Genomdaten unentbehrlich für den Einsatz von CRISPR, zweitens sind sie der Rohstoff für das Training von KI-Werkzeugen.

Seit der Veröffentlichung der ersten Genomsequenz einer Pflanze – von Arabidopsis thaliana – im Jahr 2000 haben technische Fortschritte das Tempo der Sequenzierung drastisch erhöht und die Kosten massiv gesenkt. Brauchte es noch zehn Jahre und 100 Millionen US-Dollar um das erste Arabidopsis-Genom zu sequenzieren, ist die Erbgutsequenz der Pflanze heute innerhalb einer Woche für weniger als 1000 US-Dollar ermittelt.[5] Seit 2000

ist die Sequenzierung ganzer Genome nicht nur von Modellpflanzen längst zur Routine geworden. In jüngster Zeit hat die Menge an Genomdaten stark zugenommen. Allein zwischen 2021 und 2023 sind doppelt so viele Pflanzengenome sequenziert worden wie in den zwanzig Jahren zuvor.[6]

Im Juni 2024 liegen in der Datenbank N3 der Zhejiang University in China 3505 sequenzierte Genome von 1575 Pflanzenarten vor.[7] Die Datenbank des Nationalen Zentrums für Biotechnologieinformation der USA wiederum führt zu diesem Zeitpunkt 4604 Pflanzengenome von 1482 Arten.[8] Die Zahlen werden weiter steigen, laufen doch derzeit eine Reihe von Initiativen zur Sequenzierung weiterer Pflanzengenome wie Ten Thousand Plant Genome, African Orphan Crops, Genomics for Australian Plants oder Darwin Tree of Life.[9] Das ehrgeizigste Ziel verfolgt das Earth BioGenome-Projekt. Es will bis 2030 für alle bekannten Tier-, Pflanzen- und Pilz-Arten der Welt ein Referenzgenom erstellen.[10]

Mit den Hochleistungs-Sequenzierverfahren ist in den letzten Jahren auch die Zahl der Pflanzenarten gestiegen, für die ein Pangenom verfügbar ist. Pangenome beruhen auf den Erbgutsequenzen mehrerer Individuen einer Art,[11] die ein breites Spektrum der genetischen Variation innerhalb einer Art repräsentieren. Sie sollen dabei helfen, agronomisch interessante Allele zu finden, die sich dann mit Genomeditierung in Elitesorten übertragen lassen. Bisher sollen über Tausend Pflanzengenome für die Konstruktion von Pangenomen zusammengestellt worden sein. Veröffentlichte Pangenome gibt es bei wichtigen Kulturpflanzenarten wie Reis, Mais, Weizen, Soja, Gerste und Kartoffel.[12]

Während Pangenome möglichst das gesamte Set von Genen innerhalb einer Art umfassen wollen, erweitert das sogenannte Super-Pangenom dieses Konzept um die Genome naher Verwandter.[13] Erste Projekte dazu gibt es unter anderem bei Reis,[14] Mais,[15] Tomate[16] und Kichererbse.[17] Super-Pangenome sollen dazu genutzt werden, wertvolle Eigenschaften aus Wildpflanzen mittels Genomeditierung auf Elitesorten zu übertragen.

2.2 Omik-Techniken jetzt auch für die Einzelzelle

Bis vor kurzem kamen sogenannte Omik-Techniken erst auf Ebene von Zellverbänden oder ganzen Pflanzen zum Einsatz. Die damit gewonnenen Daten haben zwar das Verständnis der Pflanzenbiologie stark erweitert, aber die Funktionen von seltenen Zelltypen oder gering konzentrierten Molekülen blieben wegen des „Verdünnungseffekts" weitgehend verschleiert. Jetzt machen neue Verfahren es möglich, auch auf Ebene einer einzelnen Pflanzenzelle Omik-Daten zu gewinnen. Die Datentiefe, die sich dadurch bietet, ist neu, werden doch erstmals auch seltene Zelltypen und Moleküle erfasst, die bisher in der Massenmessung untergingen.

Mit Einzel-Zell-Omik gewonnene Resultate erweitern die Big Data, die für das Training von KI-Tools zur Verfügung stehen.[18] Derzeit werden in der pflanzenbiotechnologischen Forschung vor allem mit scRNA-Seq genannten Methoden viele solcher neuen Trainings-Daten erhoben.[19] scRNA-Seq ist das Kürzel für Single-Cell RNA-Sequenzierungstechniken. Mit ihnen lassen sich die RNA-Moleküle einzelner Zellen mit hohem Durchsatz sequenzieren, woraus sich vor allem neue Möglichkeiten für das Verständnis der Genexpression ergeben.[20] Forschende der Nanjing University in China haben kürzlich für 17 Pflanzenarten die vorhandenen scRNA-Seq-Studien durchforstet und eine Datenbank erstellt, die Daten aus rund 2,5 Millionen Zellen umfasst.[21]

2.3 Google Maps für Pflanzen

Omik-Techniken für Einzelzellen sind auch eine wichtige Grundlage für das Projekt „Plant Cell Atlas".[22] Es startete 2019 und will umfangreiche Daten über die Struktur und Organisation von Pflanzenzellen generieren.[23] Fachleute aus verschiedenen Disziplinen wie Genetik, Zellbiologie, Bioinformatik und Bildgebungstechnologie erforschen seither, welche Typen von Pflanzenzellen es gibt und wo und wann bestimmte Moleküle innerhalb der Zellen vorhanden sind. Das Ziel ist eine „molekulare Landkarte", die hochauflösende räumlich-zeitliche Informationen über DNA- und RNA-Moleküle, Proteine und Metabolite in Pflanzenzellen enthält, eine Art Google Maps für Pflanzen. Das Unterfangen wird als eine wichtige Ressource für die Pflanzen- und Züchtungsforschung angesehen. Am ersten Pflanzenzellatlas-Symposium

nahmen 2021 fast 500 führende Fachleute aus Wissenschaft, Industrie und Behörden teil – darunter auch Mitarbeitende von BASF, Bayer, Syngenta und Google.[24]

Die großen Mengen an scRNA-Seq- und anderen Einzelzell-Omik-Daten, die im Rahmen des „Plant Cell Atlas"-Projekts entstehen, werden wiederum Rohstoffe für das Training von KI-Tools für die NGT-basierte Züchtung sein.[25]

3. Generative KI für NGT

Deep Learning, künstliche neuronale Netze, Sprach- und Diffusionsmodelle erleben derzeit rasante technologische Fortentwicklungen. Sie sorgen dafür, dass die KI sowohl bei den deskriptiven als auch den generativen Leistungen immer besser wird. Wie in vielen anderen Bereichen sollen diese Fortschritte auch in den Biowissenschaften und der Pflanzenzüchtung tiefgreifende Veränderungen bringen

Deep Learning ist ein allgemeiner Begriff für Algorithmen des maschinellen Lernens, die aus tiefen neuronalen Netzen bestehen. Neuronale Netze sind Computerprogramme, die der Funktionsweise des menschlichen Gehirns nachempfunden sind. Sie sind in der Lage riesige Mengen an unstrukturierten Daten auszuwerten. Große Sprachmodelle und Diffusionsmodelle wiederum sind Varianten künstlicher neuronaler Netze. Sie bilden die KI-Architekturen, die in Chatbots wie ChatGPT oder Bildergeneratoren wie DALL-E stecken und seit 2022 weltweit für Furore sorgen.

Dass Diffusions- und große Sprachmodelle auch in den Biowissenschaften und der NGT-basierten Pflanzenzüchtung für Aufregung sorgen, hat zwei Gründe. Erstens lassen sich große Sprachmodelle mit Daten aus der wissenschaftlichen Literatur trainieren und in neuartige Forschungsassistenten verwandeln. Zweitens – und das ist der wichtigere Grund – lassen sich Diffusions- und große Sprachmodelle auch statt mit Texten der menschlichen Sprache mit den Unmengen an DNA-, RNA- und Proteindaten trainieren, die in den letzten Jahren mit den Omik-Techniken erschlossen wurden. Die KI-Tools, die dadurch entstehen, sind zum einen deskriptiv und können wie herkömmliche Deep Learning-Algorithmen mit den „Sprachen" der Biologie umgehen und daraus Vorhersagen treffen. Zum anderen sind sie aber auch generativ: Sie können funktionale DNA-, RNA- und Aminosäurensequenzen generieren, unter anderem auch solche, die neu für die Natur sind.

3.1 Große Sprachmodelle: Forschungsassistenten für NGT-Pflanzenzüchtung

498.000 Ergebnisse liefert Google Scholar bei einer Suche mit dem Stichwort „genome editing", 657.000 mit „synthetic biology", 1,9 Millionen mit „plant breeding" und 2,2 Millionen mit „genetic engineering". KI-Tools, die sich durch diese Unmengen an Daten wühlen und sie nach den Wünschen von Forschenden analysieren, sollen NGT-basierte Züchtungsprojekte weiter erleichtern und vorantreiben.

Mindestens vier solcher Forschungshilfen gibt es schon: Open AI, die Firma hinter ChatGPT, hat einen DNA-Programmierassistenten entwickelt, der bei der Gestaltung von CRISPR-Projekten helfen und Programmiercode für DNA-bezogene Anwendungen schreiben kann.[26] Zudem bietet die Firma den *Plant Breeding Optimizer*[27] an. Das ist ein Chatbot, der Züchtungsprogramme verbessern soll und – nach eigenen Angaben – auch bei NGT-basierten Projekten hilft, Züchtungsergebnisse vorherzusagen. Google hat Ende April 2024 *CRISPR-GPT*[28] vorgestellt – einen gemeinsam mit US-Universitäten entwickelten KI-Assistenten, der die Planung und Durchführung von CRISPR-basierten Experimenten erleichtern und automatisieren soll. Ebenfalls seit 2024 gibt es *PLLaMa*[29]. Das Tool ist ein gemeinsames Produkt von Forschenden an Universitäten in China, Schweden und den USA. Sie haben das Modell *LLaMa* von Meta mit mehr als 1,5 Millionen Artikeln aus dem Bereich Pflanzenwissenschaften trainiert. Ein internationales Gremium von Agraringenieuren, Pflanzenwissenschaftlerinnen und Züchterinnen prüft derzeit, wie gut *PLLaMa* Fragen beantworten kann.

In Zukunft werden weitere Text-basierte Modelle hinzukommen, die die stetig wachsende Menge an wissenschaftlicher Literatur und Forschungsergebnissen analysieren und ihr Wissen und ihre Informationen Züchtenden zur Verfügung stellen können.[30]

3.2 An Proteinen geschulte generative KI

Proteine lenken wichtige biologische Prozesse und bestimmen das Geschehen in Pflanzen auf molekularer Ebene. Wer über KI-Werkzeuge verfügt, mit denen sich Proteine analysieren, ihre Interaktionen simulieren oder ihre Funktionen neugestalten lassen, verfügt über mächtige Werkzeuge für die Synthetische Biologie und die Gentechnik bei Pflanzen. Das Interesse an generativer KI ist riesig. Obwohl erst seit den frühen 2020er Jahren klar ist, dass sich Diffusions- und große Sprachmodelle mit Proteindaten trainieren lassen, ist die Anzahl der entstandenen Tools bereits unübersichtlich geworden. Neben akademischen Laboren und zahlreichen Startups mischen bei ihrer Entwicklung auch Tech-Konzerne mit. Die beiden Software-Giganten Microsoft und Salesforce, der Chip-Hersteller NVIDIA, die Internetkonzerne Google und Meta sowie ByteDance, der Konzern hinter TikTok – sie alle bieten KI-Werkzeuge an, die Proteinsequenzen verstehen und/ oder generieren können (Tabelle 1).

Das berühmteste Protein-Tool ist *Alphafold* von Google. Innerhalb eines Jahres hat es die 3D-Strukturen von über 200 Millionen Proteinen ermittelt[31] – eine Aufgabe, für die Forschende ohne leistungsstarke Algorithmen Millionen von Arbeitsjahren benötigt hätten. Laut Demis Hassabis, dem Kopf von Googles KI-Abteilung, ist die mit *Alphafold* erstellte Strukturdatenbank bereits von über einer Million Forschenden aus 190 Ländern besucht worden.[32] Im Juni 2024 finden sich auf Google Scholar bereits über 23.000 Publikationen, die den drei Jahre zuvor in der Fachzeitschrift Nature veröffentlichten Original-Artikel[33] zu *Alphafold* zitieren.

Viel Aufmerksamkeit erhält auch das von Meta entwickelte Protein-Tool *ESM-2*. Ein Grund dafür ist seine Geschwindigkeit: Es soll 60mal schneller sein als *Alphafold*. Ein anderer Grund ist der Metagenomic Atlas, eine Datenbank, die Meta mit *ESM-2* gebildet hat.[34] Sie enthält die Struktur von über 600 Millionen Proteinen, die in Mikroben vorkommen. 2023 im Fachjournal Science veröffentlicht, hat *ESM-2* ein Jahr später bereits 1200 Zitierungen in Google Scholar.

Neben Tools wie *Alphafold* und *ESM-2*, mit denen sich vor allem die Struktur von Proteinen modellieren lässt, entstehen in jüngster Zeit immer mehr generative KI-Werkzeuge, mit denen sich Proteine designen lassen.[35,36,37] Dazu gehören Modelle aus privaten KI-Laboren wie *Chroma*[38] von Generate Biomedicines,

EvoDiff[39] von Microsoft oder *ProGEN2*[40] von Profluent und Salesforce sowie Modelle, die aus Universitäten kommen, wie *RFdiffusion,*[41] *ProtGPT2*[42] und *ForceGen.*[43] Forschende sprechen von einer „Explosion der Möglichkeiten".[44] So bieten die Tools nicht nur neue Wege für das Redesign – also die Umgestaltung natürlicher Proteine in Versionen mit optimierten oder neuartigen Funktionen. Sie ermöglich auch ein De-novo-Design von Proteinen, die in der Natur bisher nicht bekannt sind.

Viele der Design-Tools sind so neu, dass noch nicht die nötigen experimentellen Daten vorliegen, um die Leistung ihrer Algorithmen zu bewerten. Doch es zeichnet sich ab, dass sie einen Designraum eröffnen, der natürliche Grenzen überschreitet. Die Zahl der mathematisch möglichen Proteinvarianten liegt in der Nähe von 10^{1300}. Dass in dieser unvorstellbar großen Menge, die die Zahl der Atome im Universum um ein Vielfaches übersteigt, auch unvorstellbar viele funktionslose Aminosäurensequenzen sind, ist klar. Denkbar ist aber auch, dass dieser riesige Designraum funktionierende Proteine birgt, die es in der Natur nicht gibt. Für die Forschung im Bereich der molekularen Pflanzenzüchtung haben derartige „new-to-nature" Konstrukte eine besondere Faszination.

Tabelle 1: Von Tech-Konzernen (mit)entwickelte KI-Tools für die Strukturanalyse und/oder das Design von Proteinen.

KI-Tool	Tech-Konzern	Trainingsdaten	Jahr
Alphafold-2[45]	Google	>170 000 Proteinstrukturen	2021
Alphafold-3[46]	Google	*nicht veröffentlicht*	2024
ESM-2[47]	Meta	65 Mio. Proteinsequenzen	2023
EvoDiff[48]	Microsoft	45 Mio. Proteinsequenzen	2023
LM-Design[49]	ByteDance	45 Mio. Proteinsequenzen	2023
OpenFold[50]	Microsoft/NVIDIA	>170 000 Proteinstrukturen	2024
ProGen[51]	Salesforce	280 Mio. Proteinsequenzen	2023
ProtTrans[52]	Google/NVIDIA	390 Mia. Aminosäuren	2021
ProT-VAE[53]	NVIDIA	46 Mio. Proteinsequenzen	2023

3.3 An Genomen geschulte generative KI

Seit 2021 gibt es auch erste große Sprachmodelle, die mit riesigen Mengen an DNA-Sequenzen trainiert sind und dadurch die „Sprache" der Genome simulieren können. *DNABERT*, *Nucleotide Transformer*, *GenSLM*, *megaDNA* und *EVO* – das sind die eigentümlich wirkenden Namen einer Auswahl der rund zwei Dutzend Modelle, die es derzeit gibt (Tabelle 2). Während Protein-Modelle die kodierenden Sequenzen im Erbgut abdecken, umfassen genomische Modelle zusätzlich die nicht-kodierenden Sequenzen und erlauben deshalb auch Einblicke in die Genregulation. Für Forschende bieten sich dadurch neue Möglichkeiten.

Bisher beruhen die meisten genomischen Sprachmodelle auf DNA-Sequenzen von Menschen und Tieren. Wie Mitarbeitende von Instadeep und BioNTech kürzlich zeigten, können solche Modelle jedoch auch für die Analyse von Pflanzengenomen herangezogen werden.[54]

Nicht unerwähnt bleiben soll hier, dass es auch erste Sprachmodelle für RNA-Sequenzen gibt. So zum Beispiel *CodonBERT*[55] von Sanofi oder *ERNIE-RNA*[56] und *scGPT*[57] von Microsoft, die alle drei mit RNA-Sequenzen des Menschen trainiert worden sind. In Zukunft dürfte es auch große Sprachmodelle geben, die auf RNA-Sequenzen von Pflanzen beruhen. Als vielversprechend für die Pflanzenwissenschaften und -züchtung gelten vor allem Modelle, die wie *scGPT* auf scRNA-seq-Daten beruhen.[58]

3.3.1 GPN, FloraBERT und AgroNT – erste Sprachmodelle für Pflanzengenome

Im Juni 2024 finden sich auf Google Scholar vier Modelle, die spezifisch mit DNA-Sequenzen von Pflanzen trainiert wurden. Eines davon ist das *Genomic Pre-trained Network* oder kurz *GPN*. Es stammt aus den KI-Laboren der Universität von Kalifornien, wo es mit DNA-Daten von Arabidopsis und sieben weiteren Kreuzblütler-Arten ausgestattet wurde.[59] Mit *GPN* lässt sich vorhersagen, wie sich einzelne Mutationen in regulatorischen Sequenzen auf die Pflanze auswirken.

Auch *FloraBERT* ist auf regulatorische Sequenzen spezialisiert. Das 2022 vorgestellte Modell ist ein Produkt von Inari, einem Startup, das

genomeditierte Pflanzen herstellt (siehe 6.1).[60] Die Trainingsdaten des Modells sind Promotorsequenzen aus dem Erbgut von 93 Pflanzenarten und 25 verschiedenen Maissorten. *FloraBERT* soll für mehrere Maisgewebe vorhersagen, wie sich Veränderungen in den Promotorsequenzen auf die Genaktivität auswirken.

Das bisher leistungsstärkste Sprachmodell für Pflanzengenome ist der *Agronomic Nucleotide Transformer*, kurz *AgroNT* genannt. Entstanden ist er in den KI-Schmieden von Google und Instadeep. Der Internetkonzern und das KI-Unternehmen haben sich 2022 zusammengetan, um ein Computermodell für die Genomeditierung von Pflanzen zu entwickeln, das die Simulation und Bewertung einzelner Veränderungen in einer Genomregion ermöglichen soll. Ende 2023 wurde das mit 10 Millionen Erbgutsequenzen von 48 Pflanzenarten trainierte Modell veröffentlicht.[61] Um seine Leistungsfähigkeit zu demonstrieren wurden mit dem Modell im Maniok-Erbgut mehr als 10 Millionen Mutationen simuliert und für jede einzelne prognostiziert, wie sie sich auf die Genaktivität in der Pflanze auswirkt. Wie die Entwickler schreiben, wären Ergebnisse in dieser Größenordnung mit Experimenten an den Pflanzen kaum zu erreichen und in der Natur nahezu unmöglich.

Das vierte Sprachmodell für Pflanzengenome ist *PlantCaduceus*. Es beruht auf Erbgutdaten von 16 Pflanzenarten und ist im Juli 2024 in den USA vorgestellt worden.[62]

3.3.2 Bald ganze Genome aus KI-Design?

Wie groß der Einfluss genomischer Sprachmodelle auf die Synthetische Biologie und die Genomeditierung sein wird, ist derzeit noch kaum abschätzbar. Etliche Artikel über diese Modelle sind erst auf Preprint-Servern wie bioRxiv und arXiv erhältlich und noch ohne Peer-Review. Oft fehlen auch noch die experimentellen Daten, anhand derer sich die tatsächliche Leistung der Algorithmen bewerten ließe.

Was sich aber bereits andeutet: Als Werkzeuge für die funktionelle Annotation von Genomen und als Vorhersagemodelle dürften die großen Sprachmodelle die Leistung bisheriger KI-Tools übertreffen.[63] Nach Ansicht von Instadeep können sich genomische Modelle zudem auch für die Modellierung von Proteinen eignen und deshalb ein guter Ausgangspunkt für den Bau einheitlicher, multimodaler

Grundmodelle für die Biologie sein (siehe unten).[64]

Mit mikrobiellen DNA-Sequenzen gefütterte große Sprachmodelle weisen zudem darauf hin, welche Potentiale genomische KI-Tools haben könnten. Ende 2023 hat ein Forschender der Harvard University *megaDNA* vorgestellt – ein Modell, das auf DNA-Daten von Bakteriophagen beruht.[65] Da sich mit *megaDNA* neue Sequenzen bis zu einer Länge von 96 Kilobasen generieren lassen, die die funktionelle Struktur von Phagen haben, weist das Modell den Weg zum De-novo-Entwurf ganzer Phagengenome. Auch das KI-Unternehmen Together AI und das mit privaten Geldern arbeitende Arc Institute sprechen von Entwürfen neuer Genome. Die beiden haben gemeinsam mit universitären Instituten *EVO* entwickelt, ein Modell, das auf 300 Milliarden DNA-Buchstaben aus 80.000 Bakteriengenomen und Millionen von Phagen- und Plasmidsequenzen beruhen soll.[66] *Evo* kann nicht nur Sequenzen für kleine Moleküle wie nicht-kodierende RNA generieren, sondern auch kodierende DNA-Sequenzen bis zu einer Länge von 650 Kilobasen. Laut den Entwicklern soll EVO das Potential haben, Sequenzen in der Größenordnung ganzer mikrobieller Genome zu erzeugen.

Tabelle 2: Beispiele von generativen KI-Tools, die an Genomen geschult sind.

KI-Tool	Tech-Konzern	Trainingsdaten	Jahr
AgroNT[67]	Instadeep & Google	10 Millionen Sequenzen aus Genomen von 48 Pflanzenarten	2023
DNABERT[68] (Mehrarten-Version)	Northwestern University	> 32 Milliarden Basen aus den Genomen von 135 Arten (Tiere, Pilze und Bakterien)	2023
EVO[69]	Together AI & Arc Institute	300 Milliarden Basen aus über 80.000 Bakterien- und Phagengenomen	2023
FloraBERT[70]	Inari	Promotorsequenzen von 93 Pflanzenarten und 25 Maissorten	2022
GenSLM[71]	NVIDIA & mehrere Unis	110 Millionen prokaryotische Gensequenzen und 1,5 Millionen SARS-CoV-2 Genome	2023

KI-Tool	Tech-Konzern	Trainingsdaten	Jahr
GPN[72]	Universität von Kalifornien	Genomsequenzen von 8 Pflanzenarten	2023
Nukleotid Transformer[73]	Instadeep & NVIDIA	Sequenzen aus über 3000 Humangenomen und 850 Genomen von Tieren, Pilzen und Bakterien	2023
megaDNA[74]	Harvard University	> 99.000 Phagengenomsequenzen	2023
PlantCaduceus[75]	Cornell University	Genomsequenzen von 16 Pflanzenarten	2024

3.4 Multimodale Tools – Auf dem Weg zu den Supermodellen

So leistungsstark Text-, Protein- und Genom-basierte Sprachmodelle auch sind, zeichnet sich bereits ab, dass sie schon bald von noch leistungsstärkeren Modellen abgelöst werden könnten. Das Stichwort dazu lautet Multimodalität. Während bisherige Sprach- und Diffusionsmodelle noch auf einen einzigen Datentyp beschränkt sind, arbeiten KI-Firmen jetzt an Modellen, die mehrere Arten von Daten verarbeiten können.

Anfang Mai 2024 stellten Instadeep und BioNTech mit *ChatNT* ein multimodales KI-Tool vor, das erstmals die Lücke zwischen einem mit Textdaten trainierten Gesprächsagenten und einem mit biologischen Daten trainierten Modell schließen soll.[76] Der neu entwickelte Chatbot ist zwar primär auf die medizinische Forschung ausgerichtet, Anwendungen im Pflanzenbereich sind aber möglich. Laut Instadeep signalisiert *ChatNT* „einen potenziellen Wandel hin zur Schaffung eines wirklich universellen, multimodalen KI-Systems für die Genomik".[77]

Ende Juni, kurz vor Abschluss dieser Arbeit, stellten Instadeep und BioNTech die erste multimodale KI-Architektur zur Verbindung von DNA-, RNA- und Protein-Daten vor.[78]

3.4.1 CropGPT für Breeding 5.0

KI-Tools, die Texte aus der wissenschaftlichen Literatur genauso verarbeiten können wie Daten aus der Genomik, Proteomik, Transkriptomik und Metabolomik, könnten für die Pflanzenzüchtung besonders interessant sein.[79] Sie werden als Wegbereiter einer neuen Art molekularer Pflanzenzüchtung präsentiert, die Forschende „Breeding 5.0" oder auch datengesteuerte genomische Design-Züchtung[80] nennen.

Bald könnte es ein derart universelles, multimodales KI-System für die NGT-basierte Pflanzenzüchtung geben. Anfang 2024 riefen Forschende mehrerer chinesischer Universitäten in der Zeitschrift Molecular Plant zu einem globalen *CropGPT*-Projekt auf:[81] Weltweit sollen Züchter, Biologinnen, Informatiker und Mathematikerinnen gemeinsam mit Biotechfirmen und Züchtungsunternehmen ein multimodales, auf diversen Omik-Daten beruhendes Tool entwickeln, um die Entwicklung der KI-gesteuerten Design-Züchtung zu beschleunigen.

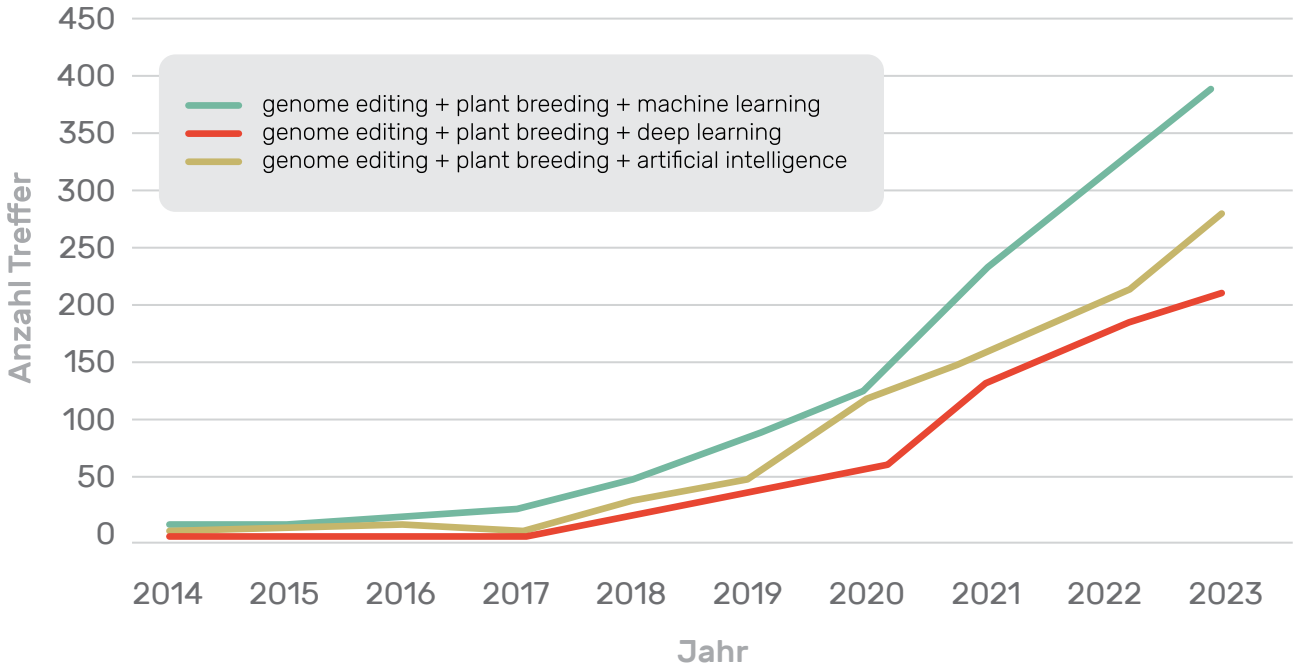
4. KI-Anwendungen in der NGT-basierten Züchtungsforschung

KI ist in der NGT-basierten Züchtungsforschung nichts neues. Tippt man bei Google Scholar die Suchworte ‚Genomeditierung', ‚Pflanzenzüchtung' und ‚Deep Learning' ein, finden sich bereits im Jahr 2017 erste Treffer. Seither steigt das Interesse an KI an. Zwischen 2018 und 2023 hat sich die Zahl der Treffer mit den genannten Stichworten verdreizehnfacht (siehe Abbildung 2). Der Trend dürfte sich weiter fortsetzen, gelten doch gerade die Tools der generativen KI als ideal für die Konvergenz mit NGT.

Wie eine kurze Sichtung der Literatur zeigt, dienten KI-Modelle in der NGT-orientierten Züchtungsforschung bisher vor allem dazu, Genom-Daten zu analysieren, regulatorische Elemente im Erbgut zu identifizieren und die Genomeditierung präziser und effizienter zu machen. Noch finden sich in der Literatur kaum Publikationen, die von konkreten Anwendungen der modernen, in Abschnitt 3 beschriebenen KI-Tools berichten. Die meisten dieser Werkzeuge sind erst nach 2022 verfügbar geworden und deshalb noch zu jung, um in der Züchtungsforschung bereits zu publizierbaren Resultaten geführt zu haben. Eine Ausnahme sind KI-Tools wie *Alphafold* von Google, mit denen sich Proteinstrukturen

vorhersagen lassen. Wie in anderen Bereichen der biologischen Forschung gehören sie auch in der NGT-basierten Züchtungsforschung zunehmend zur „Infrastruktur“ und in der Literatur finden sich mehrere Projekte zum KI-gestützten Redesign von Traits für Nutzpflanzen.

Abbildung 2: Anzahl der jährlichen Treffer in Google Scholar mit ausgewählten Suchworten zu KI und NGT für den Zeitraum 2014 bis 2023.



4.1 KI-Tools für Effizienz und Präzision der Genomeditierung

NGT-basierte Pflanzenforschung und -entwicklung beruht heute vor allem auf der CRISPR-Cas Methode. Wer mit CRISPR ins Erbgut von Pflanzen eingreift, muss vorher nicht nur die Sequenz der Zielregion kennen, sondern auch wissen, an welcher Stelle dort welche Änderungen vorzunehmen sind, um die erwünschte Eigenschaft zu erzeugen. Zudem ist eine optimale Sequenz für die Leit-RNA zu wählen, die den Ort des Doppelbruchs bestimmt, um Effizienz und Präzision der Experimente zu gewährleisten. Für alle diese Aufgaben gibt es heute KI-Tools:[82,83] Algorithmen helfen Forschenden dabei, optimale Zielorte für die Editierung zu identifizieren, indem sie den genomischen Kontext, funktionelle Annotationen und potenzielle

Off-Target-Standorte analysieren. Andere Algorithmen schlagen vor, welche Sequenzen für die Leit-RNA optimal und welche der verschiedenen CRISPR-Schneideenzyme die geeignetsten sein könnten. Diese Tools machen Erbguteingriffe mit CRISPR präziser, effizienter und erfolgreicher.

Einige Beispiele solcher Tools sind in Tabelle 3 aufgeführt.

Tabelle 3: Beispiele von KI-Tools zur Erhöhung der Effizienz und Präzision bei der gentechnischen Veränderung von Pflanzen mit CRISPR.

Werkzeug	Jahr	Basis	Zitierungen*
Pflanzenspezifische Werkzeuge			
CRISPR-P[84]	2014	49 Pflanzenarten	701
CRISPR-P 2.0[85]	2017	49 Pflanzenarten	585
CRISPR-GE[86]	2017	> 40 Pflanzenarten	326
CRISPR-Plant v2[87]	2019	7 Pflanzenarten	74
Werkzeuge, die auch für Pflanzen geeignet sind			
CRISPOR[88]	2018	> 100 Arten	1413
CHOPCHOP[89]	2014	> 100 Arten	1292

* Anzahl Zitierungen in Google Scholar am 29.6.2024

4.2 Steuerung statt Knockout: Erzeugung quantitativer Merkmalsvariationen

Die mit Abstand häufigste Form mit NGT ins Erbgut von Pflanzen einzugreifen besteht darin, Gene auszuschalten. Das Resultat sind Hunderte von Pflanzen mit Loss-of-Function-Mutationen. Züchterisch sind solche Funktionsverluste jedoch oft von beschränktem Wert. Vor allem wenn es um die Erzeugung quantitativer Merkmale geht, die durch mehrere Gene beeinflusst sind, stoßen Knockouts an ihre Grenzen.[90]

Was Forschenden bisher fehlte, waren Werkzeuge, um quantitative

Merkmalsvariationen zu erzeugen. Das könnte sich jetzt ändern. Dieser neue Trend heißt quantitative Trait-Engineering:[91] Forschende knocken Gene nicht mehr aus, sondern steuern nun deren Expression, indem sie gezielt Sequenzen des regulatorischen Netzwerkes verändern. Das Steuern der Genexpression soll die Beeinflussung komplexer quantitativer Merkmale ermöglichen.[92,93,94]

Erreicht werden soll das quantitative Trait-Engineering mit CRISPR-basierten Basen- und Prime-Editoren. Mit diesen Werkzeugen lassen sich an regulatorischen Elementen im Erbgut erstmals die Mutationen punktgenau erzeugen, die zur gewünschten Stärke der Genexpression führen. Im Visier haben Forschende dabei vor allem Cis-regulatorische Elemente (CRE) wie Promotoren, Enhancer oder Silencer, die die Transkription steuern, sowie so genannte Upstream Open Reading Frames (uORF), die die Translation regulieren.

Möglich wird das quantitative Trait-Engineering vor allem auch dank KI. Eines der vorhandenen Werkzeuge ist *iCREPCP*, eine auf Deep Learning basierende Plattform, die an der Huazhong Agricultural University in China entwickelt wurde.[95] Sie soll in Pflanzengenomen Promotorsequenzen finden und für die Genomeditierung erschließen. Ein zweites Beispiel ist *CAPE*. Das von Forschenden mehrerer chinesischer Universitäten entwickelte System kombiniert Multiplex-Genomeditierung mit einem Algorithmus, der vorhersagt, wie sich Editierungen in Promotorsequenzen auf die Genaktivität auswirken.[96] Auch für die Suche und das Editieren von uORF gibt es eine Reihe von Tools, die bei Pflanzen einsetzbar sind, wie zum Beispiel *uORFSCAN*, *uORFlight* und *PsORF.*[97]

Entwickelt werden zudem KI-Werkzeuge, mit denen sich CRE und auch uORF generieren lassen, die new-to-nature sind.[98,99] Außerhalb der Pflanzenbiotechnologie gibt es bereits eine Reihe derartiger KI-Tools.[100,101,102,103] Im Juni 2024 wurde mit *PhytoExpr* ein Modell vorgestellt, das CRE für Pflanzen designen soll. Forschende des National Maize Improvement Center in China haben für *PhytoExpr* zwei Algorithmen entwickelt: einen für das Redesign von natürlichen CRE für die Genomeditierung und einen für das Design von künstlichen CRE für die Synthetische Biologie bei Pflanzen.[104]

4.3 NGT-Züchtung mit Protein-Redesign

KI-Tools für die Vorhersage von Proteinstrukturen, wie zum Beispiel *Alphafold* von Google oder *ESM-2* von Meta, gelten in der NGT-basierten Züchtungsforschung als interessant, weil sie die Möglichkeiten für die Entwicklung von, „Designer"-Pflanzen erweitern.[105,106] Obwohl die Tools erst seit kurzem verfügbar sind, finden sich in der Literatur bereits eine Reihe von Publikationen, in denen Forschende berichten, wie sie die Werkzeuge für die Herstellung von NGT-Pflanzen nutzen wollen.

In einer kürzlich durchgeführten Studie haben Forschende mit *AlphaFold* zum Beispiel simuliert, wie die Protease Pip1 von Tomaten mit dem Protease-inhibierenden Protein EpiC2B des pathogenen Pilzes *Phytophthora infestans* interagiert.[107] Dabei stellten sie fest, dass zwei Aminosäuren in Pip1 zu ändern sind, um die Protease unempfindlich gegen die Hemmung durch EpiC2B zu machen. Mit CRISPR soll das Pip1-Gen nun entsprechend ediert und die Krankheitsresistenz der Tomate erhöht werden.

In einer anderen Studie haben Forschende jüngst mit Hilfe von *Alphafold* Patatin redesignt.[108] Patatin ist ein Protein, das natürlicherweise in Kartoffeln vorkommt. Am Computer ist eine Version entstanden, die laut KI die Viskosität und die ernährungsphysiologischen Eigenschaften von Kartoffelmehlteig verbessern soll. Mit CRISPR-basierten Prime-Editoren wollen die Forschenden die KI-generierte Patatin-Version im Erbgut von Kartoffeln erzeugen.

Weitere Beispiele, die sich in der Literatur finden lassen: Bei Mais planen Forschende, mit KI-geleitetem Proteindesign und Genomeditierung die Architektur der Pflanzen so zu verändern, dass sie auf dem Feld dichter nebeneinander wachsen können.[109] Bei Weizen soll durch die Modellierung der Struktur von Speicherproteinen die Backqualität optimiert werden.[110] Auch die Entwicklung allergenarmer Pflanzen ist ein Ziel: Indem Forschende mit KI die Struktur und Funktion allergener Proteine analysieren, wollen sie am Computer diejenigen Veränderungen ermitteln, mit denen sich die Allergenität verringern und gleichzeitig der Nährwert der Pflanzen erhalten lässt.[111] Im Visier der Forschung sind auch Proteinkinasen und Phosphatasen – zwei Enzyme, die das Wachstum von Pflanzen und ihre Interaktionen mit der Umwelt und Krankheitserregern wesentlich beeinflussen. Durch ihr KI-gesteuertes Redesign sollen editierte Pflanzen entwickelt werden, die höhere Erträge liefern und resistent gegen Krankheitserreger sind.[112] Proteine,

die Zucker transportieren, gelten als mögliche Kandidaten, die sich mit KI-Tools wie *Alphafold* und Genomeditierung für die Krankheitsresistenz optimieren lassen.[113] In Planung sind zudem Pflanzen, die dank neu designter Proteine stärker photosynthetisieren[114,115,116] oder die Kohlenstoffbindung von Böden erhöhen[117] sollen. Im Visier der Forschung sind schließlich auch NLR-Proteine und somit das Immunsystem von Pflanzen:[118,119] NLR-Proteine sind eine Art Wächter. Sie können jeweils bestimmte Pathogene erkennen und lösen bei Befall Alarm aus, worauf Pflanzen ihr Abwehrsystem aktivieren. Mit Tools wie *Alphafold* oder *ESM-2* wollen Forschende nun zuerst am Computer die Änderungen in der Aminosäurensequenz ermitteln, die für die Erweiterung der Spezifität eines NLR-Proteins notwendig sind, um sie dann mittels Genomeditierung im Erbgut der Pflanze zu erzeugen. Mit der neuen NLR-Variante soll die Pflanze Pathogene erkennen, die sie zuvor verfehlte.

4.4 SynEpi und Epigenomeditierung

Agronomische Merkmale von Nutzpflanzen sind oft nicht nur genetisch sondern auch epigenetisch gesteuert. Diese Variabilität im Epigenom ist für die molekulare Züchtung bisher kaum erschlossen. Doch Forschende wollen das mittels KI und CRISPR-basierten Werkzeugen ändern. So sind in den letzten Jahren nicht nur mehrere Algorithmen entstanden, die für die Identifikation von Epiallelen und für die Vorhersage von Veränderungen in Pflanzenepigenomen einsetzbar sind.[120,121,122] Forschende haben mit CRISPR auch neuartige Werkzeuge entwickelt, mit denen sich im Erbgut von Pflanzen gezielt Epiallele erzeugen lassen.[123,124] Bisher ist die Epigenomeditierung zwar noch weitgehend auf Modellpflanzen beschränkt, aber Forschende des Jiangsu Co-Innovation Center in China gehen davon aus, dass sie in Zukunft mit Hilfe von KI zu einer weithin anwendbaren und effektiven Methode der Pflanzenzüchtung werden könnte.[125]

2022 haben Forschende der Chinesischen Akademie der Agrarwissenschaften eine neue Züchtungsstrategie vorgestellt, die auf den Vorhersagen von KI und Werkzeugen für die Epigenomeditierung beruht. Der Name der Strategie ist „Synthetic Epigenetics“ oder kurz SynEpi. Sie folgt ingenieurwissenschaftlichen Prinzipien und will die epigenetischen Systeme von Pflanzen umgestalten oder komplett neu entwerfen. Das Ziel sind Sorten, die auf gewünschte, vorbestimmte Art auf exogene oder endogene Trigger reagieren.[126]

4.5 Automatisierung

KI gilt als Treiber für die Automatisierung der Bio- und Gentechnologie und kann auch in der NGT-basierten Pflanzenzüchtung dafür sorgen, dass Prozesse vermehrt autonom ablaufen. Während in der Forschung mit gentechnisch veränderten Mikroorganismen bereits erste autonome Labore existieren, beginnen auch in der Pflanzenbiotechnologie erste Projekte zur Automatisierung. Ende Mai 2024 haben Forschende der University of Illinois FAST-PB vorgestellt – eine schnelle, automatisierte und skalierbare Hochdurchsatz-Pipeline für das Bioengineering von Pflanzen.[127] Was im Labor nun automatisiert ablaufen kann, ist das Klonieren von Genen und die Genomeditierung von Protoplasten und Kalluszellen. Auch die NGT-Firma Cibus hat ihre Workflows automatisiert und will damit nun die Genomeditierung von Raps industrialisieren.[128] Syngenta wiederum hat Genomeditierungs- und Transgenexpressionsstudien bei Mais und Soja automatisiert und will damit die Entwicklung neuer Sorten beschleunigen.[129,130]

5. KI, NGT und Konzerne

5.1 KI-Anwendungen in Saatgutkonzernen

Die großen Saatgutkonzerne Bayer, BASF, Corteva und Syngenta sammeln seit Jahren riesige Mengen an Omik-Daten und trainieren damit Algorithmen, die sie in der herkömmlichen Züchtung zur Auswahl der genetischen Kombinationen von Pflanzen einsetzen. Die Tools mit diesen Algorithmen halten die Konzerne meist unter Verschluss.[131] Wenig ist denn auch darüber bekannt, wie die Konzerne KI-Modelle für ihre NGT-basierten Zuchtprogramme einsetzen. Dass sie solche Tools verwenden, davon ist mit ziemlicher Sicherheit auszugehen.[132]

Zumindest Corteva verfügt über ein eigenes generatives KI-Tool für NGT-basierte Züchtung. Der Konzern hat für dessen Entwicklung *BigBird* von Google verwendet, ein Sprachmodell, das DNA-Daten verarbeiten kann. Um *BigBird* für seine Züchtungsprogramme nutzen zu können, fütterte Corteva die KI mit DNA-Daten von 14 Kulturpflanzenarten – darunter Raps, Reis, Mais, Soja, Weizen und Gerste. Entstanden ist so ein Vorhersagewerkzeug um am Computer zu ermitteln, wie sich einzelne Mutationen in regulatorischen DNA-Sequenzen auf die Genaktivität auswirken.[133]

Um ihr KI-Knowhow zu ergänzen, kooperieren die Konzerne auch mit anderen Firmen. Syngenta zum Beispiel hat im Juni 2024 bekannt gegeben, das von Instadeep und Google entwickelte generative Tool *AgroNT* zu nutzen (siehe 3.3.1). Gemeinsam mit Instadeep will der Agrarkonzern nun KI-vermittelte Traits für Mais und Soja entwickeln.[134] Zudem arbeitet Syngenta mit Biographica zusammen – einem 2024 gegründeten Startup, das modernste KI-Techniken entwickelt, um in Nutzpflanzen „hochwertige Ziele für die Geneditierung" zu identifizieren.[135]

BASF und Corteva haben jeweils beide Kooperationen mit der Firma Tropic Biosciences gestartet, die eine proprietäre KI hat, um genomeditierte krankheitsresistente Pflanzen zu entwickeln (siehe 6.2).[136]

Bayer wiederum nutzt die KI-Plattform von Evogene, um im Erbgut von Mais Sequenzen zu entdecken, deren Editierung die Pflanzen krankheitsresistent machen soll.[137] Zudem unterstützt Bayer via seine Impact Investment-Einheit Leaps by Bayer die Startups Ukko und Amfora, die beide für die Entwicklung neuer Sorten auf den kombinierten Einsatz von KI und CRISPR setzen.[138]

5.2 KI für NGT im Angebot von Tech-Konzernen

Als Anbieter generativer KI-Modelle spielen bei F&E-Projekten mit NGT-Pflanzen auch Tech-Konzerne eine Rolle. Die Tools für die Analyse und das Design von Proteinen, wie sie Meta, NVIDIA, Google, Microsoft, Salesforce und ByteDance im Angebot haben, sind zwar nicht extra für die Gentechnik-basierte Pflanzenzüchtung konzipiert und hergestellt, sie können dort aber auch zur Anwendung kommen.

Google ist mit *Alphafold* nicht nur im Proteindesign aktiv, sondern entwickelt auch Tools speziell für gentechnische Pflanzenzucht. Dazu gehört das gemeinsam mit Instadeep entwickelte große Sprachmodell *AgroNT*, das unter anderem von Syngenta genutzt wird (siehe 3.3.1 und 5.1). Googles Moonshot Factory X hat bereits 2021 ein KI-Modell zum Patent angemeldet, das bei der Entdeckung interessanter Gene helfen soll und Empfehlungen dazu abgibt, welche Genomeditierungen in einer Pflanze die gewünschte Eigenschaft erzeugen.[139]

6. KI-Anwendungen bei kleineren und mittleren Unternehmen

Weltweit gibt es eine Reihe kleinerer und mittlerer Unternehmen (KMU), deren Geschäftsmodelle ganz oder zumindest teilweise auf der Konvergenz von KI und NGT beruhen (Tabelle 4). Die Strategien zur Nutzung der Konvergenz sind dabei verschieden. Da sind Firmen wie Traitseq, Evogene, Instadeep, McClintock, Biographica und Computomics, die KI-Tools für die NGT-basierte Pflanzenzüchtung entwickeln und sie Dritten anbieten. Firmen wie Arzeda oder Gingko Bioworks, kreieren mittels firmeneigener KI Traits für Pflanzenzuchtunternehmen. Ohalo, Amfora, Finally Foods, Plastomics und Hudson River Biotechnology wiederum sind KMU, die für die Entwicklung ihrer NGT-Pflanzen KI-Tools Dritter nutzen. Und schließlich sind da noch die Firmen, die für die Herstellung ihrer Sorten über proprietäre, meist auf ganze bestimmte Zwecke ausgerichtete KI-Tools verfügen. Zu diesen Firmen gehören Inari, NeoCrop, genXtraits, Phytoform, Plantae Bioscience, TreeCo und Tropic Biosciences.

Firmen der letzten Gruppe setzen ihre Tools vor allem zur Vorhersagemodellierung ein und hoffen, damit NGT-Pflanzen schneller und kostengünstiger entwickeln zu können. Ihnen wird von der Fachzeitschrift Nature Biotechnology das Potenzial zugesprochen, beim Inverkehrbringen von gentechnisch veränderten Pflanzen die Dominanz der Agrarkonzerne zu brechen.[140] Dass die Firmen bei NGT-Pflanzen eine Chance haben könnten, mit den Saatgutgiganten zu konkurrieren, zeigt sich auch an den Geldern, die sie von Investmentfirmen erhalten. Laut Daten aus den Unternehmensdatenbanken Tracxn,[141] Crunchbase[142] und PitchBook[143] sind seit 2016 mehr als 900 Millionen Euro an Wagniskapital in KMU geflossen, die KI mit NGT kombinieren.

6.1 CRISPR-KI-Startups zur Genregulierung

Die weitaus meisten Investmentgelder sind bisher an Inari gegangen. Seit der Gründung im Jahr 2016 hat die US-Firma rund 530 Millionen Euro erhalten.[144] Heute verfügt die „SEEDesign Company“ nicht nur über einen firmeneigenen Baseneditor und eine Lizenz für die Multiplex-Genomeditierung von Promotoren, sondern mit *FloraBERT* (siehe 3.3.1)

auch über ein generatives KI-Tool, mit dem sich voraussagen lässt, wie sich Mutationen in Promotoren auf die Eigenschaften einer Pflanze auswirken. Ausgerüstet mit diesen Werkzeugen will Inari bei Mais, Soja und Weizen so auf die Aktivität von Genen einwirken, dass die Pflanzen 10 bis 20 Prozent mehr Ertrag bringen. Erste Pflanzen von Inari – Soja[145] und Mais[146] mit erhöhtem Ertragspotenzial sowie ein kurzwachsender Mais[147] – haben in den USA von den zuständigen Behörden bereits grünes Licht für den Anbau erhalten. 2025 will die Firma in Australien an mehreren Orten Freilandversuche mit editiertem, ertragreichem Weizen durchführen.[148] In Belgien wiederum testet Inari einen kurzwachsenden Mais im Freiland.[149]

Auch Phytoform legt seinen Fokus auf Promotoren und die Regulierung von Genaktivitäten. Das in London und Boston ansässige Startup verfügt dabei mit *CRE.AI.TIVE* ebenfalls über ein firmeneigenes KI-Tool.[150] Sein Algorithmus soll laut Eigenwerbung ermitteln, mit welchen minimalen Änderungen in Promotorsequenzen sich maximale Wirkungen in Nutzpflanzen erzielen lassen, und so „eine noch nie dagewesene Kontrolle über die Genexpression ermöglichen".[151]

Tabelle 4: Auswahl kleinerer und mittlerer Unternehmen, die KI-Tools für NGT-basierte Züchtung anbieten und/oder KI-Tools bei der NGT-basierten Züchtung einsetzen.

Firma	Land	Jahr*	Nutzung von Künstlicher Intelligenz
Amfora	US	2016	Benutzt Algorithmus der Firma McClintock, um mit NGT Erbsen und Sojabohnen mit ultrahohem Proteingehalt zu erzeugen.
Arzeda	US	2008	Entwickelt mit KI-Proteindesign neue Traits für Pflanzen.
BellaGen	CN	2020	Nutzt das mit KI-gesteuertem Proteindesign hergestellte DNA-Schneideenzym Cas-SF01 für die Genomeditierung.
Benson Hill	US	2012	Verwendet firmeneigenes KI-System *CropOS*, um Gensequenzen zu identifizieren, die interessante Eigenschaften verleihen.
Biographica	UK	2024	Bietet KI-Werkzeuge für NGT-basierte Pflanzenzüchtung an.
Computomics	DE	2012	Bietet *AccelATrait* für die Identifizierung von Editing-Zielen an.

Firma	Land	Jahr*	Nutzung von Künstlicher Intelligenz
Evogene	IL	2002	Bietet *GeneRator* für die Identifizierung von Kandidatengenen an.
Finally Foods	IL	2024	Nutzt *GeneRator* von Evogene für Molecular Farming-Pflanzen.
genXtraits	US	2022	Verwendet firmeneigenen Algorithmus, um DNA-Abschnitte für die Editierung zu identifizieren, die als „Dimmer-Schalter" fungieren.
Ginkgo Bioworks	US	2008	Entwirft mit proprietärem Tool *Owl* neue Proteine für die Züchtung.
Inari	US	2016	Nutzt die generative KI *FloraBERT* für die Genomeditierung.
Instadeep	UK	2014	Bietet das gemeinsam mit Google entwickelte generative Modell *AgroNT* für die Genomeditierung an.
Hudson River Biotechnology	NL	2015	Setzt *AccelATrait* von Computomics ein, um Genorte für die Genomeditierung zu identifizieren.
McClintock	US	2022	Bietet KI-Werkzeuge für NGT-basierte Pflanzenzüchtung an.
NeoCrop	CL	2020	Nutzt für die Genomeditierung firmeneigenes KI-Vorhersagemodell.
Ohalo	US	2019	Nutzt bei seinen NGT-basierten Züchtungsarbeiten KI von Google.
NRGene	IL	2010	Bietet für Genomeditierung das KI-Tool GO-GENOME an.
Phytoform Labs	US	2017	Verwendet für Genomeditierung proprietäres KI-Tool *CRE.AI.TIVE.*
Plantae Bioscience	IL	2020	Nutzt KI-gesteuertes Proteindesign für NGT-basierte Züchtung.
Plastomics	US	2017	Transformiert das Erbgut von Soja-Chloroplasten mit Genen, die mit *GeneRator* von Evogene entdeckt worden sind.
Qi Biodesign	CN	2021	Hat mit Hilfe von Googles *Alphafold* einen Baseneditor hergestellt.
Traitseq	US	2023	Bietet KI-Vorhersagemodelle für Genomeditierung an.

Firma	Land	Jahr*	Nutzung von Künstlicher Intelligenz
TreeCo	US	2019	Nutzt für Genomeditierung von Bäumen ein Vorhersagemodell.
Tropic Bioscience	UK	2016	Setzt firmeneigenes *GEiGS-BioCompute*-Tool ein, um Gene für nicht-kodierende RNA zu entdecken und zu mutieren.
Ukko	US	2016	Nutzt eine eigene KI-Plattform, um für die Weizenzüchtung neuartiges Gluten zu kreieren, das Menschen mit Zöliakie vertragen.
Viridian Seeds	IE	2021	Nutzt KI für die Genomeditierung von Hülsenfrüchten.
Wild Bioscience	UK	2021	Setzt firmeneigene KI für Genidentifikation in Wildpflanzen ein.

*Gründungsjahr

genXtraits ist eine weitere Firma, die die Feinabstimmung der Aktivität von Genen im Visier hat. Sie ist 2022 in den USA gegründet worden und hat nach eigenen Angaben ein Portfolio an geistigem Eigentum, das sich auf wichtige regulatorische Elemente in Pflanzengenomen konzentriert.[152] Anders als Inari und Phytoform arbeitet genXtraits jedoch nicht mit Promotoren sondern mit uORFs. Um diese von genXtraits „Dimmerschalter" genannten Elemente im Erbgut von Pflanzen identifizieren zu können, hat die Firma ein spezialisiertes KI-Tool entwickelt.[153] Die Aktivität der damit entdeckten uORFs will sie mit NGT steuern.

6.2 Pflanzenzucht mit KI & RNAi & CRISPR

GEIGS – das Kürzel steht für „*Gene Editing* Induced Gene Silencing" und ist der Name für eine außergewöhnliche Methode in der NGT-basierten Pflanzenzucht. Das Patent für GEIGS gehört Tropic Biosciences.[154] Das britische Startup kombiniert Genom-editierung mit RNA-Interferenz (RNAi): Mit der GEIGS-Plattform lassen sich Gene, die für siRNA oder miRNA kodieren, so editieren, dass sich die Stilllegungs-Funktionen der RNAs auf neue Ziele richten – etwa auf Gene von Insekten und Pilzen oder auch auf

pflanzeneigene Gene. Dafür nutzt Tropic die firmeneigene KI *GEIGS-BioCompute*. Das Tool analysiert die Genomdaten einer Pflanze und sagt dann voraus, wo im Erbgut die geringsten Änderungen in RNAi-Genen vorzunehmen sind, um die gewünschte Eigenschaft zu erzielen. Die Fähigkeit, Stilllegungs-Funktionen mittels KI-gesteuerter Genomeditierung umzuleiten, stößt auf großes Interesse. Tropic hat BASF und Corteva als Kooperationspartner gewonnen (siehe 5.1) und seit der Gründung im Jahr 2016 über 70 Millionen Euro an Risikokapital erhalten.[155]

6.3 KI von Google und „Boosted Breeding“

Ebenfalls beliebt bei Investmentfirmen ist Ohalo. Rund 100 Millionen Euro soll das 2019 gegründete US-Startup bereits eingeholt haben, um mit NGT und KI von Google neue Pflanzensorten zu züchten.[156,157] Welche KI-Tools Ohalo dabei genau nutzt, ist unbekannt. Die Firma spricht von Vorhersagemodellen, mit denen sie herausfinden will, welche Kreuzungen aus Hunderttausenden oder Millionen möglicher Kreuzungen zu Sorten mit den gewünschten Eigenschaften führen. Für zwei Pflanzen hat Ohalo in den USA bereits grünes Licht für den Anbau erhalten: Für eine *RedVin* genannte Kartoffel, die ohne zu versüßen kalt lagerbar ist,[158] und für eine Kartoffel mit mehr Betacarotin in den Knollen.[159] Ende Mai 2024 hat die Firma ihre „Boosted Breeding“-Technik vorgestellt. Sie soll die „Evolution beschleunigen, um das Potenzial der Natur zu erschließen“. Bei der zum Patent angemeldeten Technik[160] bringt Ohalo mit CRISPR-basierten Ribonukleoproteinen Keimzellen von Pflanzen dazu, ihr Genom nicht wie üblich zu halbieren, sondern vollständig zu erhalten. Was entsteht, sind sogenannte klonale Keimzellen, die über einen doppelten Chromosomensatz verfügen. Verschmelzen zwei dieser Keimzellen miteinander, entstehen Nachkommen, die statt 50 Prozent der Gene ihrer Elternpflanzen 100 Prozent der Gene erhalten. Mit der Herstellung solch polyploider Pflanzen bietet sich laut Ohalo die Möglichkeit, „Merkmale in Boosted-Pflanzen zu kombinieren, die mit herkömmlichen Methoden erst in Tausenden von Jahren oder gar nicht kombiniert werden könnten“.

6.4 Simulation von über 69.000 Editierungsstrategien

Weshalb KI-basierte Vorhersagemodelle für Firmen so interessant sind, zeigen Zahlen von TreeCo. Die US-Firma arbeitet an genomeditierten Pappeln, die weniger Lignin bilden und dadurch die Papierherstellung erleichtern sollen. Sie hat dazu ein KI-Tool entwickelt, das auf jahrzehntelangen forstbiotechnologischen Studien basiert. Mit ihm lässt sich vorhersagen, wie sich Veränderungen in den 21 Genen, die an der Ligninsynthese beteiligt sind, auf die Holzzusammensetzung, die Wachstumsrate und andere Faktoren der Bäume auswirken.[161,162] Wie das Tool zeigt, kommen mehr als 69.000 Editierungsstrategien für die Veränderung der 21 Gene in Frage. Mit praktischen Experimenten daraus die besten Strategien zu ermitteln, würde am Aufwand scheitern. Das KI-Tool hingegen sucht nach den besten Strategien mit einer Computersimulation. So hat die Firma schließlich mit Multiplex-Genomeditierung bloß die sieben vielversprechendsten Kombinationen von Genveränderungen im Pappelerbgut experimentell erzeugt.[163] Im Gewächshaus enthielten einige der so genomeditierten Pappeln bis zu 49 Prozent weniger Lignin.

6.5 Genome von Wildpflanzen durchsuchen

Mit KI die genetische Vielfalt von Wildpflanzen für die NGT-basierte Pflanzenzüchtung erschließen – das ist das Ziel von Wild Bioscience. Die 2021 in England gegründete Firma verfügt über ein KI-Tool, mit dem sie in sequenzierten Genomen von Wildpflanzen nach Genvarianten sucht, die sie für die Züchtung Klima-robuster Sorten als interessant hält. Wild Bioscience überträgt dann diese Varianten, indem sie „kleine Änderungen am genetischen Makeup von Nutzpflanzen“ vornimmt. Mit der Durchforstung der Genome von Wildpflanzen öffnet das Startup einen genetischen Suchraum, der wesentlich größer ist als der bisher zur Verfügung stehende.[164]

6.6 Protein-Design für die Pflanzenzucht

Neben den oben erwähnten Firmen, die genombasierte KI-Tools einsetzen, gibt es im Bereich der NGT-Pflanzenzucht auch eine Reihe von Firmen, die mit Proteindesign-Tools arbeiten.

Eine dieser Firmen ist Arzeda. Sie nennt sich selbst „The Protein Design Company" und nutzt generative KI, um Proteine für alles Mögliche zu entwickeln, von Enzymen für die Industrie bis hin zu Traits für die Pflanzenzüchtung. Zu den Projekten des Unternehmens gehört zum Beispiel, eine verbesserte Version von Rubisco zu designen.[165] Das ist das Protein, mit dem Pflanzen CO_2 aus der Luft fixieren und so Kohlenstoff in die Nahrungskette einschleusen. Auch ein Pflanzenenzym, das ein weit verbreitetes Herbizid abbauen kann, hat Arzeda im Visier.[166] Laut Forbes testet die Manufaktur pro Woche 10.000 Designer-Proteine.[167] Dabei gelingt es ihr, nach eigener Aussage, „über das hinauszugehen, was uns die Natur gegeben hat."[168]

Die Grenzen der natürlichen Evolution zu überschreiten – das gelingt nach eigenen Angaben auch Gingko Bioworks. Wie Arzeda bietet die Synbio-Firma die KI-gesteuerte Entwicklung von Proteinen an, die als Traits für Nutzpflanzen in Frage kommen. Ihr generatives Tool dafür heißt *Owl*.[169] Mit ihm will Gingko die Aktivität und Spezifität von Proteinen den Kundenwünschen anpassen.

„Wir beschleunigen die Evolution der Natur über das bisher Vorstellbare hinaus und verkürzen den Weg zu gesünderen, nachhaltigeren pflanzenbasierten Lebensmitteln dramatisch," schreibt Plantae Bioscience, eine 2020 in Israel gegründete Firma, auf ihrer Webseite.[170] Für diese Beschleunigung nutzt sie KI-gestütztes Proteindesign und Genomeditierung: Zuerst entwirft sie mit Hilfe von Googles *Alphafold* und dem KI-Werkzeug *FuncLib* am Computer neue Varianten existierender Pflanzenproteine. Dann setzt sie die so ermittelten Aminosäurensequenzen durch Genomeditierung im Erbgut der Pflanzen um. Mit dieser KI-CRISPR-Kombination will Plantae Bioscience für die vertikale Landwirtschaft Pflanzen entwickeln, die kleiner sind, schneller wachsen, synchron blühen und selbst unter schwachem Licht gedeihen.

Eine weitere Firma, die mit Proteindesign-Tools und CRISPR arbeitet, ist Ukko.[171] Unterstützt mit Geldern von Leaps by Bayer verfolgt sie das Ziel, Pflanzenproteine neu zu gestalten, die Lebensmittelunverträglichkeiten

auslösen können.[172] Ukkos KI-gestützte Plattform soll es dabei möglich machen, krankmachende Proteine so zu verändern, dass sie verträglich werden, ohne ihre sonstigen Eigenschaften zu verlieren. Die im Computer ermittelten notwendigen Änderungen in der Aminosäurensequenz der Proteine will Ukko mit Genomeditierung im Erbgut der Pflanzen umsetzen.

6.7 Startups mit selbstgemachten NGT-Werkzeugen

Firmen arbeiten mit KI-gesteuertem Protein-Engineering nicht nur an neuen Eigenschaften, sondern auch an neuen Werkzeugen für die Genomeditierung von Pflanzen. So zum Beispiel BellaGen: Sie ist nach eigenen Angaben die erste Firma in China, die Genomeditierung bei Pflanzen im industriellen Maßstab betreibt. Seit kurzem nutzt sie dafür die beiden Werkzeuge hfCas12Max** und Cas-SF01, die sie mit Hilfe von KI selber entwickelt hat. Beide Tools sind Varianten von Cas12 – einer Klasse von Nukleasen, die wegen ihrer geringen Größe interessant sind, aber eine schlechtere Editierungseffizienz als andere Cas-Klassen haben. Mit Googles *Alphafold* und strukturgeleitetem Protein-Engineering hat BellaGen Cas12-Enzyme in effiziente Tools verwandelt.[173,174] Ein erster Einsatz erfolgte bei Soja: Mit Cas-SF01 wurden Pflanzen so editiert, dass sie größere Samen bilden[175] bzw. resistent gegen das Herbizid Flucarbazon sind.[176]

Qi Biodesign ist ein weiteres chinesisches Startup, das mit KI ein neues NGT-Werkzeug entwickelt hat. Gemeinsam mit mehreren universitären Forschungsinstituten durchforstete die Firma die Proteindatenbank InterPro nach Proteinen, die Sequenzähnlichkeit mit Deaminasen haben. Das sind Enzyme, die sich mit Cas-Nukleasen zu einem Baseneditor fusionieren lassen. Aus den gefundenen Proteinen wählte Qi Biodesign dann mit Hilfe von *Alphafold* diejenigen mit der vielversprechendsten Struktur aus. So entstand schließlich ein Baseneditor, mit dem sich erstmals auch im Erbgut von Soja C-G-Basenpaare in T-A-Basenpaare umwandeln lassen.[177] Zuvor gehörte Soja zu den Pflanzenarten, bei denen diese Umwandlung aus unbekannten Gründen nicht gelang.

6.8 Erste Startups mit Roboter-tauglichen Pflanzen

Nicht unerwähnt bleiben soll hier, dass Fortschritte in der KI auch die Ziele der gentechnischen Pflanzenzüchtung beeinflussen können. Zur Vereinfachung der Digitalisierung und Automatisierung der Landwirtschaft gibt es beispielsweise Vorschläge, Pflanzen gentechnisch so zu verändern, dass sie Signale aussenden, die KI-gesteuerte Landmaschinen wie Roboter oder Drohnen interpretieren können.[178,179]

Innerplant ist eine der Firmen, die bereits solche *„robot-ready crops"*[180] entwickelt. Sie stattet Soja, Mais, Tomate und Baumwolle mit Genen für Fluoreszenzstoffe aus, mit denen die Pflanzen signalisieren können, wenn sie von Schädlingen befallen werden. Gemeinsam mit John Deere und Syngenta will Innerplant ein System entwickeln, das Pflanzen, Geräte und Betriebsmittel etwa so vereint: Satelliten erkennen die Stresssignale von Innerplants Pflanzen und lenken Traktoren, die John Deere mit Fluoreszenzdetektoren bestückt hat, zu den betroffenen Feldern, wo sie gezielt Pestizide von Syngenta versprühen.[181] 2023 hat Innerplant in den USA grünes Licht für drei Pflanzen erhalten: Für Innersoy,[182] die bei Pathogenbefall Signale aussendet, sowie für eine Soja[183] und Tomate,[184] die zum Kalibrieren der Fernerkundungsgeräte notwendig sind.

Auch die US-Firma Insignum AgTech entwickelt Signalpflanzen. Anders als Innerplant nutzt sie dafür jedoch keine artfremden Gene. Sie gruppiert vielmehr Gene, die bereits im Erbgut vorhanden sind, so um, dass Pflanzen auf einen externen Auslöser mit Farbveränderungen reagieren. So hat Insignum einen Mais entwickelt, der bei Pathogenbefall an der Infektionsstelle lilafarbenes Anthocyanin bildet – einen Farbstoff, der von KI-gesteuerten Landmaschinen erkannt werden kann. Der robot-ready-Mais ist Ende 2023 in den USA für den Anbau frei gegeben worden.[185]

7. Generative KI und Regulierung von NGT1-Pflanzen

Im Juli 2023 hat die EU-Kommission dem EU-Parlament und dem EU-Ministerrat einen Vorschlag zur Deregulierung von gentechnisch veränderten Pflanzen vorgelegt, die durch gezielte Mutagenese, Cisgenese oder einer Kombination der beiden Techniken hergestellt werden und kein genetisches Material von außerhalb ihres züchterischen Genpools enthalten.[186]

Der Vorschlag unterscheidet dabei – je nach Ausmaß der gentechnischen Veränderungen – zwei Kategorien von gentechnisch veränderten Pflanzen (GVP): NGT1-Pflanzen, die bis zu 20 gezielte Veränderungen in ihrem Erbgut enthalten (siehe Abschnitt 7.3), und NGT2-Pflanzen, die mehr als 20 Veränderungen aufweisen. Die EU-Kommission geht davon aus, dass die Risikoprofile von NGT1-Pflanzen und herkömmlich gezüchteten Pflanzen vergleichbar sind und schlägt vor, NGT1-Pflanzen von den Anforderungen der GVO-Rechtsvorschriften auszunehmen und den geltenden Bestimmungen für herkömmlich gezüchteten Pflanzen zu unterstellen. NGT2-Pflanzen hingegen sollen im Regulierungsbereich des Gentechnikrechts bleiben.

Um EU-Parlament und -Ministerrat eine fundierte Diskussion zu ermöglichen, hat die EU-Kommission dem Gesetzesentwurf eine Folgenabschätzung, Fallstudien der Gemeinsamen Forschungsstelle (JRC), Arbeiten der Europäische Behörde für Lebensmittelsicherheit (EFSA) sowie die Resultate einer Stakeholderbefragung beigelegt. Was in diesen Dokumenten und somit auch in der laufenden politischen Debatte zur Regulierung von NGT-Pflanzen unberücksichtigt bleibt, ist die Konvergenz von NGT und generativer KI, wie sie derzeit in den NGT-Laboren der Welt stattfindet. Sie hat erst an Fahrt aufgenommen, als die meisten NGT-Aktivitäten der EU-Kommission bereits abgeschlossen waren oder kurz vor Abschluss standen.

Da dieser Konvergenz ein großes Potenzial zugesprochen wird, die NGT-basierte Züchtung zu verändern, drängt sich vielmehr auf, vor der Verabschiedung eines neuen Gesetzes proaktiv zu diskutieren und zu klären, welche regulatorischen Fragen und Sicherheitsbedenken mit der Konvergenz einhergehen. Dies gilt umso mehr, als für NGT1-Pflanzen vorgeschlagen wird, vorsorgliche Maßnahmen wie Risikoprüfung und Rückverfolgbarkeit aufzuheben.

Im Folgenden werden zuerst einige allgemeine Aspekte aufgeführt, die bei einer proaktiven regulatorischen Diskussion eine Rolle spielen sollten. Anhand eines Szenarios werden anschließend regulatorischen Fragen dargestellt, die die Konvergenz von NGT und generativer KI bei der Herstellung von NGT-Pflanzen mit sich bringt. Danach liegt der Fokus auf NGT1-Pflanzen: Zuerst wird der Designraum vorgestellt, der gemäß des Gesetzesvorschlags für die KI-gesteuerte Herstellung von NGT1-Pflanzen zur Verfügung stehen würde. Danach wird für die Aspekte Risikoprüfung, Rückverfolgbarkeit und Kennzeichnungspflicht jeweils dargestellt, weshalb es notwendig ist, die Konvergenz von NGT und generativer KI in der Regulierungsdebatte zu berücksichtigen.

7.1 Allgemeine regulatorische Aspekte

Im Folgenden sind einige allgemeine Aspekte aufgeführt, die bei der Gestaltung der Governanz der Konvergenz von NGT und generativer KI zu berücksichtigen sind. Sie widerspiegeln weitgehend Aspekte, die in der Literatur über mögliche Risiken und Bedenken des Einsatzes von generativer KI in Wissenschaft und Technik zu finden sind.[187,188,189,190,191]

7.1.1 Generative KI senkt Qualifikationsschwelle

Bislang ist die Veränderung von Pflanzen mit NGT hochqualifizierten Fachleuten vorbehalten, die umfassend in molekularbiologischen Techniken geschult sind. Generative KI dürfte das ändern. Ihre Modelle sind immer ausgefeilter und verfügen zunehmend über Fachkenntnisse und Entscheidungsfähigkeiten, die bisher nur erfahrene Forscher und Forscherinnen hatten. In Naher Zukunft könnten Chatbots Anleitungen und Unterstützungen für Laien und Laiinnen bieten und damit auch Studenten, Informatikerinnen, Unternehmern oder DIY-Biologinnen Zugang zur NGT-Pflanzenzüchtung ermöglichen. Diese Menschen werden weder Erfahrungen im Umgang mit gentechnisch veränderten Pflanzen mitbringen noch sich der Biosicherheitsfragen ausreichend gewahr sein. Diese Dequalifizierung wirft deshalb in Verbindung mit der Black Box (7.1.4), dem Halluzinieren (7.1.5) und möglichen Datenverzerrungen (7.1.6) die Bedenken auf, dass NGT1-Pflanzen mit unerwünschten oder unangemessenen Eigenschaften geschaffen und in die Umwelt entlassen werden.

7.1.2 Generative KI bringt Produktivitätsschub

Automatisierung, KI-basierte Forschungsassistenten, leistungsstarke Computersimulationen und Designtools – sie machen die NGT-basierte Pflanzenzüchtung zunehmend zu einem Daten-gesteuerten Prozess, aus dem immer schneller immer mehr NGT-Pflanzen hervorgehen, die als mögliche Kandidaten für die Sortenentwicklung in der Umwelt getestet werden. Aus Sicht des Gesundheits- und Umweltschutzes geben die Beschleunigung und der damit einhergehende Produktivitätsschub auch Anlass für

Bedenken. Denn das höhere Tempo und die Vielzahl unterschiedlicher Sortenkandidaten dürften es schwieriger machen, während der Sortenentwicklung die Pflanzen zu erkennen und auszusortieren, die unerwartet Eigenschaften aufweisen, die für die Gesundheit von Mensch, Tier und Umwelt unerwünscht sind.

7.1.3 Generative KI bringt neue Werkzeuge

Dutzende einfache Genscheren und hochentwickelte Basen-, Prime- und Epigenom-Editoren stehen heute bereits für die NGT-basierte Pflanzenzüchtung zu Verfügung. Die generative KI wird diesen Werkzeugkasten stark erweitern. Startups wie BellaGen und Qi Biodesign zeigen beispielhaft, dass sich nur schon mit Strukturanalyse-Tools wie *Alphafold*, neue CRISPR-basierte Genscheren und Basen-Editoren kreieren lassen (siehe 6.7). Leistungsstarke Proteindesign-Tools werden die Möglichkeiten noch weiter erhöhen. Im April 2024 sind erstmals NGT-Werkzeuge vorgestellt worden, die mit Hilfe von großen Protein-Sprachmodellen entstanden: ein Baseneditor der Westlake University[192] sowie OpenCRISPR-1 des Startups Profluent.[193] OpenCRISPR-1 ist besonders bemerkenswert. Ersten kommt das Protein aus einem Pool von Millionen neuer CRISPR-Proteinsequenzen, die Profluent mit *ProGen* am Computer entworfen hat. Und zweitens ist OpenCRISPR-1 ein sehr neuartiges Protein: Es unterschiedet sich durch mindestens 182 Mutationen von jedem natürlichen CRISPR-Protein; zur weitverbreiteten Genschere SPCas9 sind es sogar 403 Mutationen.

Auch genomische Sprachmodelle dürften bald zur Entwicklung neuer Formen von NGT-Werkzeugen beitragen. *EVO* zum Beispiel, ein mit Erbgutdaten von Bakterien trainiertes Modell, kann Sequenzen generieren, die new-to-nature sind und dennoch wie Cas9-Genscheren funktionieren sollen.[194,195]

Ob Nukleasen, Deaminasen, Rekombinasen, Transposasen oder Methyltransferasen – heute ist der Werkzeugkasten der NGT-Pflanzenzüchtung noch weitgehend auf Komponenten beschränkt, die aus natürlicher Quelle kommen. Generative KI wird nicht nur dazu beitragen, dass diese „natürlichen" NGT-Tools durch Redesign noch leistungsstärker werden. Sie könnte auch eine Reihe neuartiger NGT hervorbringen, die die Multiplex-Genomeditierung, das Stapeln von Genen, die Umlagerungen einzelner Sequenzen und den Umbau von

Chromosomen erleichtern.[196] Forschung und Industrie werden dank generativer KI über Werkzeuge verfügen, mit denen sie das Erbgut von Pflanzen in noch größerem Maßstab manipulieren können als heute.

7.1.4 Black Box

Generative KI-Modelle arbeiten oft als „Black Box“:[197,198,199] Sie liefern Vorhersagen oder machen Empfehlungen, ohne dass es für Menschen nachvollziehbar ist, wie und weshalb die Modelle genau diese Vorhersagen und Empfehlungen machen. Diese Undurchsichtigkeit behindert zwar nicht den technologischen Nutzen der KI-Modelle, sie schränkt aber die Evaluierbarkeit in Bezug auf Zuverlässigkeit oder Sicherheit ein.

In sensiblen Bereichen wie der NGT-basierten Pflanzenzüchtung, deren Produkte die Gesundheit vieler Menschen und die Umwelt tangieren können, unterminiert der Mangel an Nachvollziehbarkeit und Reproduzierbarkeit ihrer Ergebnisse das Vertrauen in generative KI-Modelle. Hier ist deshalb nach Wegen zu suchen, wie zukünftige KI-Modelle für die interessierte Öffentlichkeit und insbesondere für Regulierungsbehörden transparent und nachvollziehbar gemacht werden können. Zudem sind Lösungen zu finden, die bei der KI-gesteuerten Herstellung von NGT-Pflanzen gewährleisten, dass menschliche Intelligenz, Kontrolle und Steuerung an kritischen Stellen integriert bleiben.

7.1.5 Halluzinationen

Neben der Black Box gibt auch das „Halluzinieren“ Anlass für Bedenken: Generative KI-Modelle können Ergebnisse liefern, die vernünftig erscheinen, aber sachlich falsch oder irrelevant sind.[200] Wie oft und in welchen Zusammenhängen KI-Modelle „halluzinieren“ und wie dies verhindert oder vermindert werden kann, muss noch ermittelt werden. Klar ist aber, dass mit der unhinterfragten Produktion von falschen und irrelevanten Ergebnissen zu rechnen ist. Die Kombination von Black Box und Halluzination ist vor allem da höchst problematisch, wo generative KI-Modelle Vorschläge für umfangreiche Eingriffe ins Erbgut von Pflanzen machen und die veränderten Pflanzen dann in die Umwelt freigesetzt werden.

7.1.6 Datenverzerrungen und Mangel an logischem Verständnis

Die Outputs und Vorhersagen von generativen KI-Modellen spiegeln immer die Daten wider, mit denen die Modelle trainiert wurden. Enthalten die Trainingsdaten Verzerrungen, die von den zugrundeliegenden biologischen Systemen oder von den menschlichen Kuratoren stammen, können sich diese Verzerrungen auf die Ergebnisse des Modells übertragen. Zudem fehlt den KI-Modellen auch ein Verständnis für Kausalitäten. Sie können zwar Muster und Beziehungen in den Daten korrekt identifiziert, jedoch nicht erfassen, was die unmittelbaren Ursachen oder mechanistischen Erklärungen für die identifizierten Zusammenhänge sind. Dieser Mangel an kausalem Verständnis schränkt letztendlich die Fähigkeit ein, mögliche Nebenwirkungen oder Fehlfunktionen zu antizipieren, die bei der Umsetzung von KI-Vorhersagen in reale Anwendungen auftreten können.[201]

7.1.7 Geschwindigkeit und Zukunftssicherheit

Sowohl bei den NGT als auch bei der generativen KI finden derzeit technologische Fortschritte in einem atemberaubenden Tempo statt. Die Bewältigung dieses raschen technologischen Wandels stellt für die Governanz der Konvergenz von NGT und generativer KI eine Herausforderung dar. In den Bereichen, in denen NGT und generative KI gemeinsam zum Einsatz kommen, haben Behörden und Gesetzgeber laufend zu prüfen, ob die geltenden Vorschriften noch mit den sich schnell verändernden technologischen Möglichkeiten Schritt halten können.

7.1.8 Konzernmacht

Die Macht von Tech-Konzernen bei der Entwicklung von generativer KI ist immens. Sie verfügen über die nötige Infrastruktur und hochqualifiziertes Fachpersonal, haben Zugang zu leistungsstarken Rechnern und riesigen Cloudkapazitäten und besitzen die finanziellen Mittel, die es für die sehr kostspielige Herstellung generativer Modelle braucht.

Da KMU und öffentliche akademische Einrichtungen die hohen Entwicklungskosten selten aufbringen

können, liegt ein Großteil der KI-Durchbrüche in den Händen privater Konzerne. Einige wenige Tech-Giganten können Markttrends bestimmen, Standards setzen und darüber entscheiden, ob sie die Codes ihrer Modelle offenlegen und wem sie unter welchen Bedingungen Zugang zu den Werkzeugen geben. Mehr noch: Sie haben auch die Macht, ethische und regulatorische Diskussionen und somit auch politische Entscheidungen zu beeinflussen.[202]

Wenn Tech-Konzerne wie Meta, Google, NVIDIA, Salesforce und Microsoft jetzt auch generative KI-Modelle für Biowissenschaften, Synthetische Biologie und NGT-basierten Pflanzenzüchtung entwickeln, stellt sich die Frage, ob und wie sich die Konzernmacht hier auswirkt. Beeinflussen die Ziele der Konzerne, wie ihre an Proteinen und Genomen geschulten Modelle funktionieren? Wie transparent, reproduzier- und nachvollziehbar sind die Tools der Tech-Giganten? Was sind die Folgen, wenn KI-Modelle für die NGT-basierte Züchtung immer größer werden und nur noch wenige Unternehmen die besten und leistungsstärksten Tools entwickeln können? Welche Formen und Möglichkeiten staatlicher Kontrolle braucht es? Und mit welchen Ressourcen, Kompetenzen und Interventionsmöglichkeiten sind nationale oder auch internationale Institutionen auszustatten, die für die Kontrollen zuständig sind? Eine breite, öffentliche Debatte dieser Fragen tut not. Bisher findet die Diskussion erst in kleinen Kreisen statt.[203,204]

7.1.9 „Open-Washing“

Viele der in dieser Arbeit erwähnten KI-Modelle privater Unternehmen sind öffentlich. *AgroNT* von Google/ Instadeep zum Beispiel ist bei Hugging Face[205] und *FloraBERT* von Inari bei Github[206] erhältlich. Was „öffentlich“ bei den einzelnen Modellen jeweils genau bedeutet, bleibt jedoch zu prüfen.

Dass das Label „Open Source” im Bereich generativer KI nicht immer hält, was es verspricht, zeigt eine aktuelle Untersuchung der niederländischen Radboud University. Dort haben zwei Forschende geschaut, wie offen, transparent und zugänglich Chatbots und Bildgeneratoren privater Firmen tatsächlich sind. Das Resultat: Tech-Giganten wie Google, Meta und Microsoft bezeichnen ihre KI-Modelle zwar oft als open source, geben aber nur wenige Schlüsselinformationen wie

Code oder Trainingsdaten preis. Kurzum: Tech-Konzerne betreiben „Open-Washing".[207,208]

Viel Kritik aus der Wissenschaftsgemeinde gab es im Mai 2024, als Google in der Zeitschrift Nature *Alphafold 3* vorstellte, die neueste Version seiner revolutionären KI zur Vorhersage von Proteinstrukturen.[209,210] *Alphafold 3* ist zwar auf einem öffentlich Webserver abrufbar, aber anders als bei früheren Versionen unterliegt die Nutzung einer Lizenz, die auf nichtkommerzielle Nutzung beschränkt ist. Noch mehr: Google verzichtete erstmals auch darauf, den Computercode öffentlich zu machen, der den Fortschritt des Modells beschreibt. Neben einem Aufschrei in den sozialen Medien kritisierten mehr als 1000 Forschende in einem offenen Brief an die Redaktion von Nature, dass die Zeitschrift Googles Artikel ohne Computercode akzeptierte und damit von den Standards der Forschungsgemeinschaft abgewichen ist.[211]

7.2 Szenario „Google Crops"

Noch steht die Entwicklung generativer KI-Modelle für die NGT-basierte Pflanzenzüchtung ganz in den Anfängen. Wohin sie führen wird, ist offen. Stuart Smyth von der Universität von Saskatchewan hat kürzlich die „Google Crops" prophezeit – ertragsoptimierte, mit KI designte und mit CRISPR erzeugte Sorten.[212] Daran anlehnend ist im Folgenden ein Szenario skizziert. Es soll veranschaulichen, welche regulatorischen Fragen sich mit dem Einzug generativer KI-Modelle stellen.

2027: Google hat *AgroNT* zu einem multimodalen Modell weiterentwickelt, das nicht nur die Sprache der Proteine und Genome versteht, sondern auch die rechtlichen Bedingungen für Züchtung und Anbau von NGT-Pflanzen kennt. Google bietet sein Tool – nennen wir es „The KI-Breeder" – Züchtungskonzernen an, die es mit eigenen Daten feinabstimmen können. Syngenta arbeitet bereits seit 2024 mit *AgroNT* und nutzt nun „The KI-Breeder", um NGT1-Raps für den europäischen Markt herzustellen. Der Konzern gibt dazu die Genomsequenzen seiner Elitesorten sowie Daten zum Klima der Anbaugebiete und zur Bodenqualität der Felder in das Tool ein und erhält die Information, wie er seine Elitesorten genomeditieren muss, um

gleichzeitig hohe Erträge zu erzielen und im NGT1-Geltungsbereich zu bleiben. Ein automatisierter Genomeditierungs-Workflow setzt die Vorschläge von „The KI-Breeder“ um. Ohne vorher mögliche Umwelt- oder Gesundheitswirkungen prüfen zu müssen, bringt Syngenta dann die editierten Rapsvarianten in die Umwelt aus und testet ihre Erträge an mehreren Orten versuchsweise. Da Syngenta keine Maßnahmen treffen muss, um seine Versuche zeitlich und räumlich zu begrenzen, entweicht editierter Raps via Samen aus den Versuchsflächen und gibt zudem seine von Google designten Gene via Pollen an andere Rapspflanzen sowie verwandte Wildarten weiter.

Das Szenario wirft Fragen auf, deren regulatorische und politische Diskussion vor der geplanten Deregulierung von NGT-Pflanzen durchaus sinnvoll sein könnte: Ist es denkbar, dass generative KI-Tools fehleranfällig sind und ungewollt Vorschläge machen, deren Umsetzung zu editierten Sorten mit unerwünschten Wirkungen auf Mensch, Tier oder Umwelt führen kann? Falls ja: Soll es in der Eigenverantwortung der Firmen liegen, ob sie zuverlässige und sichere Tools verwenden oder nicht? Sind hierfür verbindliche Qualitätsstandards erforderlich? Sollen die Firmen selbst kontrollieren, dass Fehler erkannt werden und keine NGT1-Pflanzen mit unerwünschten Wirkungen die Labore verlassen? Sollen die Firmen selbst wählen können, wie viel Entscheidungen sie an eine KI abgeben und an welchen Stellen ihres KI-gesteuerten Züchtungsprozesses sie menschliche Intelligenz, Kontrolle und Entscheidung einsetzen? Kurzum: Reichen Eigenverantwortung und Selbstkontrolle der KMU und Konzerne aus oder braucht es den Staat, der mit geeigneter Regulierung für sichere Tools und Pflanzen sorgt?

Wichtig ist auch die Frage, wer denn wie eruiert, ob ein KI-Tool zuverlässig ist und sichere Vorschläge macht? Erfordert dies ein schrittweises Vorgehen, bei dem Daten erst am Bildschirm, dann im Labor, in Gewächshäusern und in kontrollierten Freisetzungsversuchen gesammelt und dann Behörden zur Bewertung vorgelegt werden? Oder sollen Firmen und Konzerne das in Eigenverantwortung in der Züchtungspraxis herausfinden, so wie es bei der geplanten Deregulierung der Fall wäre?

Wichtig für diese politische und regulatorische Diskussion ist insbesondere die Frage: Kann eine generative KI mit dem Raum, der ihr rechtlich für das Design eines NGT1-Genoms zur Verfügung steht, Pflanzen entwerfen, deren Risikoprofil sich von herkömmlich gezüchteten Pflanzen unterscheidet?

7.3 Der Designraum für NGT1-Pflanzen

Der Designraum, der nach dem Vorschlag der EU-Kommission für die Herstellung von NGT1-Pflanzen zur Verfügung steht, ist in Anhang 1 des NGT-Verordnungsentwurfs definiert. Er benennt die Kriterien für die Gleichwertigkeit von NGT1-Pflanzen mit herkömmlich gezüchteten Pflanzen (Abbildung 3). Sind die Kriterien erfüllt, gilt die Gleichwertigkeit auch dann, wenn die Eigenschaften der NGT1-Pflanzen neuartig sind und in herkömmlich gezüchteten Sorten der gleichen Art nicht vorkommen.

Welche Möglichkeiten der geplante Designraum für die KI-gesteuerte Herstellung von NGT1-Pflanzen theoretisch bietet, zeigen folgende Ausführungen: Eine generative KI kann beispielsweise für 20 kodierende Stellen des Genoms jeweils die Einführung 18 neuer Nukleotide vorschlagen. Da dies pro Stelle sechs Aminosäuren entspricht, ist ein Neudesign mehrerer Proteine möglich. Die KI kann auch Editierungen an 20 Stellen des Genoms vorschlagen, die als CRE oder uORF wirken. Dadurch kann sie wiederum das regulatorische Netzwerk einer Pflanze gestalten. Der KI stehen für ihre Designvorschläge zudem alle DNA-Sequenzen aus dem züchterischen Genpool einer Pflanzenart zur Verfügung. Sie kann aus dem Super-Pangenom einer Art bis zu 20 Gene auswählen und damit zum Beispiel die Bildung eines neuen Stoffwechselweges vorschlagen. Sehr groß wird der Designraum für eine KI, wenn sie zudem die Möglichkeiten von Kreuzungen ausschöpft. Denn der Vorschlag der EU-Kommission sieht vor, dass Kreuzungen zweier unterschiedlicher NGT1-Pflanzen wiederum zu NGT1-Pflanzen führen, auch wenn die Nachkommen dann mehr als 20 gentechnisch erzeugte Veränderungen aufweisen.[213,214]

Abbildung 3: Von der EU-Kommission vorgeschlagene Kriterien für die Gleichwertigkeit von NGT1-Pflanzen mit herkömmlich Pflanzen

Eine NGT-Pflanze gilt als gleichwertig mit herkömmlichen Pflanzen, wenn sie sich von der Empfänger-/Elternpflanze durch nicht mehr als 20 genetische Veränderungen der unter den Nummern 1 bis 5 genannten Arten in einer DNA-Sequenz unterscheidet, die eine Sequenzähnlichkeit mit der Zielstelle aufweist, die durch bioinformatische Werkzeuge vorhergesagt werden kann.

1. **Ersatz oder Einführung von höchstens 20 Nukleotiden;**
2. **Streichung einer beliebigen Anzahl von Nukleotiden;**
3. **sofern die genetische Veränderung ein endogenes Gen nicht unterbricht:**
   - a. **gezielte Einführung einer zusammenhängenden DNA-Sequenz in den Genpool des Züchters;**
   - b. **gezielter Ersatz einer endogenen DNA-Sequenz durch eine im Genpool des Züchters vorhandene zusammenhängende DNA-Sequenz;**
4. **gezielte Umkehrung einer Abfolge beliebiger Nukleotide;**
5. **jede andere gezielte Veränderung jeglicher Größe unter der Bedingung, dass die resultierenden DNA-Sequenzen bereits (möglicherweise mit Veränderungen gemäß den Nummern 1 und/oder 2) in einer Art aus dem Genpool der Züchter auftreten.**

7.4 Risikoprüfung von NGT1-Pflanzen

Wer in der EU eine gentechnisch veränderte Pflanze versuchsweise freisetzen oder Inverkehrbringen will, muss nach geltendem Recht vorab eine Risikoprüfung durchführen. Diese Pflicht dient dazu, etwaige unerwünschte Auswirkungen von GVP auf Mensch, Tier, Umwelt und Biodiversität schon im Vorfeld zu vermeiden. Die EU-Kommission geht davon aus, dass die Risikoprofile von NGT1-Pflanzen gleich sind wie die Risikoprofile von herkömmlich gezüchteten Pflanzen, und schlägt deshalb vor, die Pflicht zur gentechnikrechtliche Risikoprüfung von NGT1-Pflanzen aufzuheben. Nach den Plänen der EU-Kommission würde es lediglich für NGT1-Lebensmittel eine Risikobewertung für die menschliche Gesundheit geben, die als neuartige Lebensmittel einzustufen sind und unter die Verordnung 2015/2283 fallen.

Die Frage, ob der Einsatz von Protein- und Genom-basierten generativen KI-Modellen zu NGT1-Pflanzen führen kann, die sich im Risikoprofil von herkömmlich gezüchteten Pflanzen unterscheiden, spielte in der Regulierungsdebatte bisher keine Rolle. Angesichts des Potenzials, das diesen KI-Modellen zugesprochen wird, mahnt sich ein Einbezug jedoch an. Um über das Für und Wider einer Risikoprüfungspflicht sachkundig entscheiden zu können, sollte vorab geklärt sein, welche Risikoprofile NGT1-Pflanzen haben könnten, deren Genomveränderungen ein KI-Modell vorgeschlagen hat.

Protein-basierte generative KI-Modelle zeichnen sich durch mehrere Fähigkeiten aus: sie können Proteinstrukturen voraussagen, Proteinfunktionen erschließen und sowohl Protein-Protein-Interaktionen wie auch Interaktionen zwischen Proteinen und kleinen Molekülen prognostizieren. Die KI-Modelle verbessern damit die Möglichkeiten für gentechnische Veränderungen natürlicher Pflanzenproteine deutlich. Für die Regulierungsdebatte wäre es daher sinnvoll, eine Bestandsaufnahme der Protein-basierten generativen KI-Modelle zu machen und zu beurteilen, welches Redesign-Potenzial die KI-Modelle im NGT1-Designraum (7.3) derzeit und künftig haben. Dabei sollte vor allem die Frage beantwortet werden, ob auch redesignte Proteine mit neuartigen Funktionen machbar sind/werden, die NGT1-Pflanzen Eigenschaften mit erhöhtem Risikoprofil verleihen.

Genom-basierte generative KI-Modelle verbessern das Verständnis von Genomen und geben einen Einblick in die Art und Weise, wie DNA-

Elemente auf verschiedenen Ebenen zusammenwirken, um komplexe Funktionen zu ermöglichen. Sie können helfen, Auswirkungen von Genomveränderungen vorherzusagen und funktionale DNA-Sequenzen zu entwerfen. Wie bei den Protein-basierten sollte auch bei den Genom-basierten generativen KI-Modellen geklärt werden, welches Designpotenzial innerhalb des NGT1-Designraums derzeit besteht und bei weiteren Fortschritten der KI-Modelle in Zukunft zu erwarten sein könnte. Dabei sollte wiederum die Frage beantwortet werden, ob die KI-Modelle auch die Herstellung von NGT1-Pflanzen ermöglichen, deren Risikoprofil sich gegenüber herkömmlich gezüchteten Pflanzen erhöht.

Ein Szenario, das mögliche NGT1-Pflanzen mit erhöhtem Risikoprofil schildert: Pflanzeneigene mikroRNAs können mit Genomeditierung so verändert werden, dass sie in Schadinsekten die Bildung essenzieller Proteine durch RNAi unterbinden. Mit einem Genom-basierten generativen KI-Modell könnte es möglich sein, das Erbgut einer Sorte nach Sequenzen zu durchsuchen, die für mikroRNAs kodieren. Das Modell findet mehrere solcher Sequenzen und schlägt für je drei davon vor, wie sie mit jeweils in weniger als 20 Nukleotiden geändert werden müssten, um gegen zwei verschiedene Schadinsekten via RNA-Interferenz giftig zu wirken. Die Vorschläge werden mit Genomeditierung umgesetzt und es entsteht eine NGT1-Pflanze, die sechs insektentoxische Substanzen bildet und mit herkömmlichen Züchtungsmethoden in praktikablen Zeiträumen nicht realisierbar wäre. Nach den Plänen der EU-Kommission wäre vor dem Inverkehrbringen der NGT1-Pflanze nicht zu prüfen, wie die neu gebildeten mikroRNAs auf Nicht-Zielinsekten wirken, obwohl solche unerwünschten Wirkungen denkbar sind.[215] Würden die neu gebildeten mikroRNAs hingegen als Pflanzenschutzmittel auf die Felder gesprüht werden, müssten sie nach geltendem EU-Pflanzenschutzmittelrecht einer Risikoprüfung unterzogen werden.

Ein weiteres Szenario: Eine Züchtungsfirma ermittelt mit einem Genom-basierten generativen KI-Modell, wie die regulatorischen Elemente des zmm28-Gens von Mais zu editieren sind, damit sich die Expression des Gens erhöht. Das zmm28-Gen codiert für einen Transkriptionsfaktor, der wiederum die Aktivität von Genen reguliert, die an Prozessen wie Photosynthese, Stickstoffassimilation und wachstumsregulierenden Hormonsignalen beteiligt sind.
Die Züchtungsfirma setzt die Vorschläge des KI-Modells um und generiert einen NGT1-Mais mit einem

höheren Kornertrag. Dass in einem herkömmlichen Züchtungsprogramm genau die vom KI-Modell ermittelten Genomeditierungen erzeugt werden können, ist wenig wahrscheinlich. Eine Überexpression des ZMM28-Transkriptionsfaktors wirft Sicherheitsbedenken auf:[216] Lebens- und Futtermittel aus NGT1-Mais könnten mehr Auxine, Indolylessigsäure, Indolylbuttersäure oder Nitrat als üblich bilden. Diese Sicherheitsbedenken blieben ungeklärt, wenn für NGT1-Pflanzen keine Pflicht zur Risikoprüfung bestünde.

Ein drittes Szenario: Ein Startup gibt einem Genom-basierten KI-Modell den Auftrag, in einem Super-Pangenom, das aus allen öffentlich erhältlichen Genomsequenzen von Raps und fünf seiner verwandten Wildarten besteht, nach Sequenzen zu suchen, die potentiell für Antimikrobielle Peptide (AMP) kodieren könnten. AMP sind Teil des pflanzlichen Abwehrsystems gegen Pilze, Viren und Bakterien sowie teilweise auch gegen Insekten und Nematoden. Je nach Sequenz und Struktur werden AMP unterschiedlichen Typen zugeordnet wie etwa Thioninen, Defensinen, Knottinen, Snakinen, Cyclotiden oder heveinartigen Peptiden. Das KI-Modell liefert dem Startup zwei Dutzend Sequenzen. Das Startup lässt die entsprechenden Gene synthetisch herstellen, kreiert damit mehrere unterschiedliche cisgene Varianten von Raps und testet dann die Varianten in der Umwelt. Eine Variante, die sechs potenzielle AMP-Gene aus vier verwandten Wildarten enthält, erweist sich als besonders robust gegenüber Schadpilzen, und das Startup bringt sie als Sorte für die Herstellung von Biodiesel in Verkehr. Mit herkömmlichen Züchtungsmethoden wäre die Sorte innerhalb praktikabler Zeiträume nicht herstellbar. Fragen, die sich bei diesem Szenario für die Regulierungsdebatte stellen: Welche Informationen über die neue Sorte sollte das Startup im Verfahren zur Überprüfung des Status als NGT1-Pflanze den zuständigen Behörden liefern müssen? Reicht die Information, dass die sechs eingefügten Cisgene laut dem KI-Modell AMP-Gene sind? Oder müsste das Startup vorab experimentell abklären, ob die sechs vom KI-Modell vorgeschlagenen Proteine tatsächlich AMP sind? Und ist es vertretbar, dass das Startup die cisgene Rapsvariante ohne Risikoprüfung in die Umwelt bringen darf, obwohl bekannt ist, dass gewisse AMP auch toxisch auf Tier und Mensch wirken können?

7.5 Kennzeichnung von NGT1-Pflanzen

Nach geltendem EU-Recht sind Lebensmittel, die aus gentechnisch veränderten Pflanzen bestehen oder daraus hergestellt werden, als GVO zu kennzeichnen. Diese Kennzeichnungspflicht gewährt heute auf Einzelhandels- und Verbraucherebene die Wahlfreiheit. Ausschlaggebend für die Kennzeichnungspflicht ist nicht das Vorhandensein von artfremder DNA in den pflanzlichen Lebensmitteln. Entscheidend ist vielmehr, ob bei der Herstellung der Pflanze gentechnische Verfahren angewendet worden sind. Diese Prozesskennzeichnung will die EU-Kommission für NGT1-Pflanzen nun abschaffen. Auch wenn bei der Herstellung von NGT1-Pflanzen heute in der Regel klassische gentechnische Verfahren eingesetzt werden, will die EU-Kommission diesen Prozess nicht mehr transparent machen und schlägt in ihrem Gesetzesentwurf vor, NGT1-Pflanzen von der Kennzeichnungspflicht auszunehmen. Die Abwesenheit von artfremder DNA sowie die angenommene Gleichwertigkeit mit herkömmlich gezüchteten Pflanzen sollen nun die Kriterien werden, die über Wahlfreiheit bei NGT1-Pflanzen entscheiden.

Wenn bisher das Für und Wider einer Kennzeichnungspflicht für NGT1-Pflanzen zur politischen Debatte stand, spielte der mögliche Einsatz von KI im Herstellungsprozess der Pflanzen keine Rolle. Doch jetzt werfen die Protein- und Genom-basierten generativen KI-Modelle auf die Aufhebung der Prozesskennzeichnung ein neues Licht und machen einen Einbezug der KI in die Deregulierungsdebatte notwendig.

Kann die Information, ob bei der Herstellung von NGT1-Pflanzen generative KI zum Einsatz kam, für Verbraucherinnen und Verbraucher wesentlich sein, sich für oder gegen einen Kauf des NGT1-Produktes zu entscheiden? Muss Verbraucherinnen und Verbrauchern etwa kommuniziert werden, ob KI das Erbgut einer Tomate entworfen hat, die sie kaufen wollen? Das sind zwei der Fragen, die der mögliche Einsatz generativer KI-Modelle neu in die Debatte bringt und die politisch zu beantworten sind. Dabei sind das Ausmaß des KI-Designs, der mögliche Verzicht auf Risikoprüfungen für NGT1-Pflanzen (7.4), die Black Box (7.1.4) sowie die Künstlichkeit der erzeugten Veränderung (new-to-nature) durch den Einsatz der KI-Modelle wichtige Aspekte.

7.6 Rückverfolgbarkeit von NGT1-Pflanzen

Rückverfolgbarkeitssysteme sind beim kommerziellen Umgang mit GVP in der EU gesetzliche Pflicht. Sie sorgen dafür, dass GVP und daraus gewonnene Produkte über die ganze Herstellungs- und Vertriebskette hinweg lückenlos rückverfolgbar und in der Natur . Die Plicht zur Rückverfolgbarkeit hatte der Gesetzgeber unter anderem eingeführt, um einen schnellen Rückruf etwaiger fehlerhafter Produkte zu ermöglichen. Diese dem Inverkehrbringen nachgelagerte Risikovorsorge will die EU-Kommission nun für NGT1-Pflanzen aufheben.

Wie die Aufhebung der Rückverfolgbarkeitspflicht bei NGT1-Pflanzen zu bewerten ist, deren Veränderungen ein KI-Modell vorgeschlagen hat, ist bisher weder von zuständigen Behörden noch von politischen Gremien und der interessierten Öffentlichkeit diskutiert worden, sollte aber in die Debatte um die Regulierung von NGT-Pflanzen miteinbezogen werden. Zu diskutieren ist etwa, ob Black Box (7.1.4), Halluzinationen (7.1.5) und Datenverzerrungen (7.1.6) nicht auch zu unsicheren oder fehlerhaften NGT1-Produkten führen könnten und eine Rückrufmöglichkeit sinnvoll wäre, wenn KI-gesteuerte Vorschläge direkt in veränderte Genome münden und die resultierenden Pflanzen ohne Risikoprüfung und staatliche Aufsicht auf den Markt kommen können.

Glossar

(erstellt mit Unterstützung von ChatGPT)

Antimikrobielle Proteine

Antimikrobielle Proteine (AMP) sind kleine Proteine, die von Pflanzen produziert werden, um sich gegen verschiedene Krankheitserreger wie Bakterien, Pilze und Viren zu verteidigen. AMP wirken gegen die Mikroorganismen, indem sie deren Zellwände oder Membranen zerstören oder deren Stoffwechselprozesse behindern. AMP spielen eine wichtige Rolle im Immunsystem der Pflanzen und können auch in der Pflanzenzüchtung genutzt werden, um krankheitsresistente Sorten zu entwickeln.

Basen-Editor

Ein Basen-Editor ist ein Werkzeug für die Genomeditierung, das auf dem CRISPR-System basiert. Es ermöglicht, einzelne DNA-Basen in einem Genom gezielt zu verändern, ohne dabei die DNA-Doppelstrangbrüche zu erzeugen, die typisch für das herkömmliche CRISPR-Cas9-System sind. Ein Basen-Editor kann zum Beispiel gezielt eine einzelne Base, wie etwa eine C-G-Paarung, in eine T-A-Paarung umwandeln. Ein Basen-Editor besteht aus einem modifizierten Cas-Protein, das die DNA-Bindung, aber nicht den DNA-Schnitt vermittelt, und einer → Deaminase-Komponente, die die chemische Umwandlung einer Base in eine andere bewirkt.

Cis-regulatorisches Element

Ein cis-regulatorisches Element – kurz CRE – ist eine DNA-Sequenz, die die Aktivität eines Gens steuert, indem sie die Bindung von → Transkriptionsfaktoren und anderen regulatorischen Proteinen ermöglicht oder verhindert.
Diese Elemente befinden sich in der Regel in der Nähe des Gens, das sie regulieren. CRE spielen eine Schlüsselrolle in der Genexpression, indem sie bestimmen, wann, wo und wie stark ein Gen exprimiert wird. Beispiele für cis-regulatorische Elemente sind → Promotoren, → Enhancer und → Silencer.

Deep Learning

Deep Learning ist ein Teilbereich des maschinellen Lernens, der auf künstlichen neuronalen Netzen basiert. Es handelt sich um eine Methode, bei der ein Computer lernt, komplexe Muster und Beziehungen in großen Datenmengen zu erkennen und zu

verstehen. Diese neuronalen Netze bestehen aus mehreren Schichten (daher „deep" für tief), durch die Daten schrittweise verarbeitet und transformiert werden, um Muster, Merkmale oder Entscheidungen zu identifizieren.

Deaminase

Eine Deaminase ist ein Enzym, das eine chemische Reaktion namens Deaminierung katalysiert. Bei dieser Reaktion wird eine Aminogruppe (-NH_2) aus einem Molekül entfernt. Deaminasen spielen eine wichtige Rolle im Stoffwechsel von Aminosäuren und Nukleotiden, indem sie die Abspaltung von Aminogruppen ermöglichen, was für die Energiegewinnung und den Abbau von überschüssigen Stickstoffverbindungen essenziell ist. Deaminasen spielen eine wichtige Rolle in der Genomeditierung, da sich mit ihnen → Basen-Editoren kreieren lassen.

Deskriptive Künstliche Intelligenz

Deskriptive Künstliche Intelligenz bezieht sich auf den Einsatz von → Künstlicher Intelligenz (KI), um bestehende Daten zu analysieren und zu beschreiben. Sie kann in großen Datensätzen Muster identifizieren, die oft für Menschen schwer erkennbar sind. Deskriptive KI hilft, vorhandene Daten besser verstehen zu können.

Diffusionsmodell

Ein Diffusionsmodell ist ein Typ von generativem Modell in der → Künstlichen Intelligenz, das darauf ausgelegt ist, komplexe Datenmuster zu lernen, indem es schrittweise Rauschen zu einer Datenstruktur hinzufügt und dann den Prozess umkehrt, um diese Struktur wiederherzustellen. Diese Modelle lernen, wie man Daten aus einem verrauschten Zustand rekonstruiert, was ihnen ermöglicht, neue, realistisch wirkende Daten zu erzeugen.

In einem biologischen Kontext könnten Diffusionsmodelle verwendet werden, um Muster in genetischen Sequenzen, wie DNA oder RNA, zu erfassen und zu rekonstruieren, oder um bei der Simulation von molekularen Prozessen zu helfen. Zum Beispiel könnten sie bei der Modellierung der Faltung von Proteinen oder der Vorhersage von genetischen Mutationen genutzt werden, indem sie den Übergang von einem ungeordneten zu einem geordneten Zustand lernen und umgekehrt.

Enhancer

Ein Enhancer ist ein DNA-Abschnitt im Genom, der die Expression eines oder mehrerer Gene steigern kann. Enhancer gehören zu den → cis-regulatorischer Elementen. Sie wirken unter anderem, indem sie spezifische Transkriptionsfaktoren binden oder die Bindung von Aktivatorproteinen fördern. Enhancer können weit entfernt von dem Gen liegen, das sie regulieren, und dennoch Einfluss auf dessen Expression nehmen. Ein Enhancer ist das Gegenstück zum → Silencer.

Einzel-Zell-Omik

Einzel-Zell-Omik (engl. Single-Cell-Omics) bezeichnet eine Sammlung von Techniken und Methoden, die es ermöglichen, biologische Informationen auf der Ebene einzelner Zellen zu gewinnen und zu analysieren. Im Gegensatz zu herkömmlichen „Bulk"-Analysen, die Daten aus einer gemischten Population von Zellen sammeln und nur Durchschnittswerte liefern, erlaubt Einzel-Zell-Omik eine detaillierte Untersuchung individueller Zellen. Dadurch können Unterschiede zwischen einzelnen Zellen aufgedeckt werden, die in der Gesamtanalyse verborgen bleiben würden.

Epiallel

Epiallele sind Gene oder Allele, die zwar in ihrer DNA-Sequenz übereinstimmen, aber unterschiedliche epigenetische Modifikationen (z.B. Methylierung) aufweisen und deshalb in der Regel unterschiedlich stark exprimiert werden.

Epigenom

Das Epigenom bezeichnet die Gesamtheit aller epigenetischen Modifikationen an der DNA eines Organismus, die die Genaktivität und Genexpression regulieren, ohne die DNA-Sequenz selbst zu verändern. Die epigenetischen Modifikationen umfassen unter anderem DNA-Methylierung, Histon-Modifikationen und die Organisation der Chromatinstruktur. Im Gegensatz zum Genom, das relativ stabil ist, kann das Epigenom dynamisch sein und sich im Laufe des Lebens stärker ändern.

Epigenomeditierung

Epigenomeditierung ist eine Technik, die darauf abzielt, gezielt Änderungen im Epigenom vorzunehmen, um die Genexpression zu steuern, ohne die zugrunde liegende DNA-Sequenz zu verändern. Während traditionelle

Genomeditierungsmethoden wie CRISPR/Cas9 direkt die DNA-Sequenz verändern, konzentriert sich die Epigenomeditierung auf die Modifikation von epigenetischen Markierungen wie zum Beispiel DNA-Methylierungen und Histonmodifikationen.

Eine der Techniken für die Epigenomeditierung beruht auf der Verwendung von dCas9, einem inaktiven Cas9-Enzym, das die DNA nicht schneidet. An dCas9 können Enzyme gekoppelt werden, die als epigenetische Modulatoren wirken. Zum Beispiel kann dCas9 mit einer DNA-Methyltransferase fusioniert werden, um gezielt DNA-Methylierungen zu ändern.

Funktionelle Annotation

Funktionelle Annotation bezeichnet in der Genomik den Prozess, bei dem den identifizierten Sequenzen eines Genoms bestimmte biologische Funktionen zugeordnet werden. Ein sequenziertes Genom liegt zunächst in Form einer langen Abfolge von DNA-Basen (A, T, C, G) vor. Die funktionelle Annotation hilft dabei, diese Abfolge zu interpretieren und herauszufinden, welche Abschnitte der DNA welche Rolle im Organismus spielen.

Generative Künstliche Intelligenz

Generative Künstliche Intelligenz bezieht sich auf den Einsatz von → Künstlicher Intelligenz (KI), die Daten nicht nur analysiert, sondern auch neue Daten erstellt (generiert). Beispiele für generative KI sind Modelle wie ChatGPT, die menschliche Sprache erzeugen können, oder wie DALL-E, die Bilder fertigen können.

Großes Sprachmodell

Ein Großes Sprachmodell (Large Language Model, LLM) ist eine Art von künstlicher Intelligenz, die entwickelt wurde, um komplexe Sequenzen von Symbolen, Zeichen oder Daten zu analysieren, zu verstehen und zu generieren. Es basiert auf tiefen neuronalen Netzen. Die Modelle lernen Muster, Beziehungen und Strukturen innerhalb großer Mengen von sequentiellen Daten, sei es Text oder biologische Sequenzen wie sie in DNA, RNA oder Proteinen vorkommen. Durch ihr Training auf umfangreichen Datensätzen können Große Sprachmodelle verschiedene Aufgaben erfüllen, wie das Vorhersagen von Sequenzen, das Generieren neuer Sequenzen oder das Klassifizieren von Daten. ChatGPT beruht beispielsweise auf einem großen Sprachmodell, das menschliche Sprache verstehen und generieren kann.

Künstliche Intelligenz

Künstliche Intelligenz (KI) bezeichnet die Entwicklung von Computersystemen, die in der Lage sind, Aufgaben zu erledigen, die normalerweise menschliche Intelligenz erfordern. Dazu gehören unter anderem das Erkennen von Mustern, das Verstehen natürlicher Sprache, das Treffen von Entscheidungen und das Lernen aus Erfahrungen. In den Biowissenschaften umfasst KI maschinelle Lernverfahren und andere intelligente Algorithmen, die komplexe biologische Daten analysieren können und unter anderem in der Genomik, Proteomik oder bei der Bildanalyse zum Einsatz kommen.

Maschinelles Lernen

Maschinelles Lernen ist ein Bereich der → Künstlichen Intelligenz, in dem Algorithmen und statistische Modelle entwickelt werden, die es Computern ermöglichen, Aufgaben ohne ausdrückliche Anweisungen auszuführen. Beim maschinellen Lernen lernt ein Algorithmus aus Mustern in einer großen Menge markierter Daten. Sobald er trainiert ist, können Vorhersagen oder Entscheidungen auf der Grundlage dieses Lernens als Reaktion auf neue und ungesehene Daten getroffen werden.

Metabolomik

Metabolomik bezieht sich auf die Analyse und Identifizierung aller Metaboliten in pflanzlichen Proben, um die biochemischen Prozesse und Stoffwechselwege in Pflanzen besser zu verstehen. Mit der Methode lassen sich Einblicke in die Pflanzenphysiologie, die Reaktion auf Umweltstress, die Biosynthese von sekundären Pflanzenstoffen und auch die Auswirkungen gentechnischer Eingriffe gewinnen.

Methyltransferasen

Methyltransferasen sind Enzyme, die eine Methylgruppe ($-CH_3$) auf ein Substrat übertragen. Diese Substrate können DNA, RNA, Proteine oder andere Moleküle sein. Die Methylierung durch Methyltransferasen ist ein wichtiger biochemischer Prozess, der die Funktion von Genen und Proteinen regulieren kann. Beispielsweise kann die Methylierung von DNA-Abschnitten durch DNA-Methyltransferasen Gene inaktivieren, was eine Form der epigenetischen Genregulation darstellt. In Kombination mit dem CRISPR/Cas-System können Methyltransferasen für die gezielte → Epigenomeditierung eingesetzt werden.

Mikro-RNA

Mikro-RNA – kurz miRNA – ist ein kurzes, nicht-kodierendes und einzelsträngiges RNA-Molekül, das ein wichtiger Bestanteil der RNA-Interferenz (RNAi) ist. miRNAs sind etwa 21-25 Nukleotide lang und an einer Vielzahl von biologischen Prozessen beteiligt; so etwa bei Zellwachstum und -differenzierung, Apoptose (programmierter Zelltod) und Stressreaktionen.

Multiplex-Genomeditierung

Multiplex-Genomeditierung bezeichnet die gleichzeitige Bearbeitung mehrerer Zielstellen im Genom einer einzelnen Zelle. Das Multiplexing gelingt vor allem mit dem CRISPR-System: Indem mehrere verschiedene guide RNAs (gRNAs) gleichzeitig mit dem Cas-Schneideenzym in Zellen eingeführt werden, kommt es auch an mehreren verschiedenen Stellen des Erbguts zu Veränderungen. Die Methode ermöglicht es, mehrere Gene auf einmal zu editieren oder auszuschalten.

Neuronales Netz

Ein neuronales Netz ist ein Modell oder Programm für maschinelles Lernen, das von der Funktionsweise des menschlichen Gehirns inspiriert ist. Es besteht aus einer Vielzahl von miteinander verbundenen Knoten – sogenannten künstlichen Neuronen. Die Knoten sind in Schichten organisiert: einer Eingabeschicht, einer oder mehrerer verborgenen Schichten und einer Ausgabeschicht. Neuronale Netze stützen sich auf Trainingsdaten, um zu lernen und ihre Genauigkeit im Laufe der Zeit zu verbessern. Sie können Daten mit hoher Geschwindigkeit klassifizieren, clustern und generieren.

Nukleasen

Nukleasen sind Enzyme, die Nukleinsäuren, also DNA oder RNA, spalten können. Sie können spezifische Bindungen zwischen Nukleotiden aufbrechen, was dazu führt, dass die DNA- oder RNA-Stränge geschnitten werden. In der Genomeditierung sind Nukleasen entscheidend, da sie gezielt an spezifischen Stellen des Genoms schneiden können, was den Weg für verschiedene gentechnische Manipulationen ebnet. Das Cas9-Enzyme des CRISPR-Systems ist ein Beispiel für eine Nuklease.

Omik-Techniken

Omik-Techniken sind eine Gruppe von Techniken und Ansätzen, die darauf abzielen, die Gesamtheit (das „Om“) bestimmter Molekülklassen in Zellen, Geweben oder Organismen zu untersuchen. Sie ermöglichen es,
umfassende Informationen über die Struktur, Funktion und Dynamik biologischer Systeme zu erhalten. Zu den Omik-Techniken gehören unter anderem → Genomik, → Proteomik, → Transkriptomik und → Metabolomik

Pangenom

Ein Pangenom einer Pflanzenart umfasst möglichst das ganze Set von Genen, das innerhalb dieser Art vorkommt. Es setzt sich aus dem Kern-Genom (den Genen, die in allen Individuen der Art vorhanden sind) und dem variablen Genom (den Genen, die nur in einigen, aber nicht allen Individuen der Art vorkommen) zusammen.

Peptid

Ein Peptid ist ein Molekül, das aus einer kurzen Kette von Aminosäuren besteht, die durch Peptidbindungen miteinander verknüpft sind. Peptide können aus nur zwei Aminosäuren bestehen (Dipeptide) oder aus längeren Ketten mit bis zu etwa 100 Aminosäuren (Polypeptide). Peptide spielen eine wichtige Rolle in biologischen Prozessen wie Signalübertragungen oder Abwehrreaktionen.

Polyploide Pflanze

Eine polyploide Pflanze besitzt mehr als zwei Sätze von Chromosomen in ihren Zellen. Im Gegensatz zu diploiden Pflanzen, die zwei Chromosomensätze (einen von jedem Elternteil) haben, haben polyploide Pflanzen oft drei (triploid), vier (tetraploid) oder noch mehr Chromosomensätze. Polyploidie tritt natürlich auf und kann durch evolutionäre Prozesse, wie z.B. Fehler in der Zellteilung, entstehen. Polyploide Pflanzen weisen häufig erhöhte Robustheit, größere Zellen und Früchte sowie eine höhere genetische Vielfalt auf, was sie für die Züchtung und Landwirtschaft besonders wertvoll macht.

Prime-Editor

Ein Prime-Editor ist ein Werkzeug für die Genomeditierung, das auf dem CRISPR-System basiert und zur präzisen Bearbeitung von DNA-Sequenzen verwendet wird. Im Gegensatz zum herkömmlichen CRISPR-Cas9-System, das DNA-Doppelstrangbrüche erzeugt, kombiniert der Prime-Editor Cas-Enzyme, die Einzelstrangbrüche einfügen, mit einer reversen Transkriptase, die RNA in DNA umschreiben kann. Diese Kombination ermöglicht das Einfügen, Löschen oder Austauschen spezifischer DNA-Sequenzen ohne Doppelstrangbrüche. Das Prime-Editor-System verwendet eine sogenannte Prime-Editing-Guide-RNA (pegRNA): Sie steuert die Enzyme nicht nur an den Zielort im Erbgut, sondern enthält auch die Information für die gewünschte Änderung. Die reverse Transkriptase kopiert diese Information in den Zielort im Genom.

Promotor

Ein Promotor ist ein → cis-regulatorischer Element. Er befindet sich direkt vor dem Startpunkt der → Transkription eines Gens und dient als Bindungsstelle für die RNA-Polymerase und andere → Transkriptionsfaktoren, um die Transkription zu initiieren.

Proteomik

Proteomik bezeichnet die Analyse und Charakterisierung aller Proteine, die in einer pflanzlichen Zelle oder einem Gewebe zu einem bestimmten Zeitpunkt gebildet werden. Mit der Methode lässt sich das Proteinspektrum, einschließlich Proteinmodifikationen und Interaktionen, ermitteln, womit wiederum Einblicke in die funktionelle Biologie der Pflanze ihre Reaktionen auf Umweltbedingungen oder die Auswirkungen gentechnischer Eingriffe gewinnen lassen.

Quantitative Merkmale

Quantitative Merkmale sind Eigenschaften von Pflanzen, die durch viele Gene (polygen) beeinflusst werden und nicht nur in klar abgrenzbaren Kategorien vorkommen, sondern eine kontinuierliche Variation zeigen.
Im Gegensatz zu qualitativen Merkmalen, die von wenigen Genen bestimmt werden und diskrete, eindeutig identifizierbare Klassen aufweisen (z. B. Blütenfarbe), sind quantitative Merkmale durch eine breite Skala von Ausprägungen gekennzeichnet.

Rekombinase

Eine Rekombinase ist ein Enzym, das genetische Rekombination vermittelt, indem es DNA-Stränge erkennt und an spezifischen Stellen schneidet und wieder zusammenfügt. Rekombinasen spielen eine zentrale Rolle bei der Umlagerung von DNA-Sequenzen, die in der Natur beispielsweise bei der DNA-Reparatur, beim Austausch von genetischem Material zwischen Chromosomen oder bei der Integration von Virus-DNA in das Wirtsgenom vorkommen. In der Genomeditierung werden Rekombinasen eingesetzt, um gezielt DNA-Sequenzen zu verändern.

RNA-Interferenz

RNA-Interferenz – kurz RNAi – ist ein natürlicher zellulärer Prozess, der die Genexpression reguliert, indem er spezifische mRNA-Moleküle abbaut oder deren Übersetzung in Proteine verhindert. RNAi spielt eine zentrale Rolle in der Genregulation und dient als Abwehrmechanismus gegen Viren.

scRNA-Seq-Daten

scRNA-Seq-Daten (Single-Cell RNA Sequencing-Daten) stammen aus einer Technologie, die es ermöglicht, die Genexpression in einzelnen Zellen zu messen. Im Gegensatz zu traditionellen RNA-Sequenzierungsmethoden, die die durchschnittliche Genexpression über viele Zellen hinweg messen, bietet scRNA-Seq eine detaillierte Ansicht der Genexpression auf der Ebene einzelner Zellen.

Silencer

Ein Silencer ist ein DNA-Abschnitt im Genom, der die Expression eines oder mehrerer Gene unterdrücken oder verringern kann. Silencer gehören zu den → cis-regulatorischer Elementen. Silencer können weit entfernt von dem Gen liegen, das sie regulieren, und dennoch Einfluss auf dessen Expression nehmen. Ein Silencer ist das Gegenstück zum → Enhancer.

Small interfering RNA

Small interfering RNA (siRNA) ist ein kurzes, nicht-kodierendes und doppelsträngiges RNA-Molekül, das ein wichtiger Bestanteil der RNA-Interferenz (RNAi) ist. siRNAs sind etwa 20-25 Nukleotide lang und dienen Pflanzen dazu, die Expression von

spezifischen Genen zu regulieren. Pflanzen setzen siRNA insbesondere bei der Abwehr von Viren ein.

Strukturgeleitetes Protein-Engineering

Strukturgeleitetes Protein-Engineering ist ein Ansatz in der Biotechnologie, bei dem die dreidimensionale Struktur eines Proteins genutzt wird, um gezielt Änderungen an der Aminosäuresequenz des Proteins vorzunehmen. Das Ziel ist es, die Funktion, Stabilität, Bindungsaffinität oder andere Eigenschaften des Proteins zu verbessern oder zu verändern.

Super-Pangenom

Ein Super-Pangenom einer Pflanzenart umfasst möglichst das ganze Set von Genen, das innerhalb dieser Art und ihrer Verwandten vorkommt. Super-Pangenom sind meist → Pangenome auf Gattungsebene. Sie geben Aufschluss über die Evolutionsgeschichte, Domestikationsprozesse und genetische Beziehungen innerhalb einer Gattung geben.

Trait

Trait ist ein Begriff aus der gentechnischen Pflanzenzüchtung. Er bezeichnet eine spezifische, durch den gentechnischen Eingriff eingeführte Eigenschaft oder Fähigkeit einer Pflanze. Ein Trait kann zum Beispiel die Resistenz gegenüber bestimmten Schädlingen, eine höhere Toleranz gegenüber Herbiziden, oder eine verbesserte Nährstoffzusammensetzung sein. Durch das Einfügen eines oder mehrerer Gene, die für den gewünschten Trait verantwortlich sind, können Forschende bestimmte Eigenschaften gezielt in Pflanzen einführen.

Transkription

Transkription bezeichnet den Prozess, bei dem die genetische Information in der DNA in eine komplementäre RNA-Sequenz umgeschrieben wird. Während dieses Prozesses wird ein spezifischer Abschnitt der DNA, der ein Gen enthält, von einem Enzym namens RNA-Polymerase abgelesen und in eine mRNA (messenger RNA) umgewandelt. Diese mRNA dient später als Vorlage für die → Translation, bei der die in der RNA kodierte Information in ein Protein übersetzt wird. Transkription ist der erste Schritt der Genexpression.

Transkriptomik

Transkriptomik bezeichnet die Analyse und Quantifizierung aller RNA-Moleküle (insbesondere mRNA) in pflanzlichen Zellen oder Geweben. Mit der Methode wird das Muster der Genexpression unter bestimmten Bedingungen ermittelt, wodurch es möglich wird, die Genaktivität in verschiedenen Entwicklungsstadien oder bei Reaktionen auf Umweltfaktoren zu erforschen.

Translation

Translation ist der Name für den Prozess, bei dem die genetische Information, die in der mRNA (messenger RNA) kodiert ist, in eine Aminosäuresequenz übersetzt wird, um ein Protein zu bilden. Dieser Prozess findet in den Ribosomen statt. Während der Translation lesen die Ribosomen die mRNA-Sequenz in Dreiergruppen von Nukleotiden, die als Codons bezeichnet werden, und fügen die entsprechenden Aminosäuren zu einer wachsenden Polypeptidkette hinzu. Diese Kette faltet sich dann zu einem funktionalen Protein.

Transposase

Eine Transposase ist ein Enzym, das für die Mobilität von Transposons (springenden Genen oder mobilen genetischen Elementen) verantwortlich ist. Transposons sind DNA-Sequenzen, die innerhalb eines Genoms von einer Position zu einer anderen verschoben werden können. Die Transposase erkennt spezifische DNA-Sequenzen an den Enden eines Transposons, schneidet diese aus ihrer ursprünglichen Position heraus und integriert sie in eine neue Stelle im Genom. In der Genomeditierung können Transposasen eingesetzt werden, um Gene gezielt in ein Genom einzufügen oder sie daraus zu entfernen.

Upstream Open Reading Frame

Ein upstream Open Reading Frame – kurz uORF – ist ein kurzer offener Leseraster (ORF), der sich stromaufwärts (upstream) des Haupt-ORFs eines Gens befindet. Ein uORF kann potenziell ein kleines → Peptid kodieren. uORFs sind wichtige Elemente, die zur Feinabstimmung der Proteinproduktion beitragen, indem sie die → Translation in Abhängigkeit von zellulären Bedingungen und Signalen regulieren.

Quellenverzeichnis

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 transcriptomics. 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 identifying 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 Biotechnology 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 automate 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/__GeBti0I5F0hQvboXpyDz1lWejC1W32h0_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 Lam, H. Y. I., Ong, X. E., & Mutwil, M. (2024). Large language models in plant biology. Trends in Plant Science in press

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