frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 15 November 2023
DOI 10.3389/fpls.2023.1252166

OPEN ACCESS

EDITED BY
Ahmad Bazli Ramzi,
National University of Malaysia, Malaysia

REVIEWED BY
Diego Orzaez,
Polytechnic University of Valencia, Spain
Tsan-Yu Chiu,
Beijing Genomics Institute (BGI), China
Johannes Felix Buyel,
University of Natural Resources and Life Sciences, Austria

*CORRESPONDENCE
Ramalingam Sathishkumar
rsathish@buc.edu.in
Ashutosh Sharma
asharma@tec.mx

RECEIVED 03 July 2023
ACCEPTED 17 October 2023
PUBLISHED 15 November 2023

CITATION
Parthiban S, Vijeesh T, Gayathri T, Shanmugaraj B, Sharma A and Sathishkumar R (2023) Artificial intelligence-driven systems engineering for next-generation plant-derived biopharmaceuticals. *Front. Plant Sci.* 14:1252166. doi: 10.3389/fpls.2023.1252166

Artificial intelligence-driven systems engineering for next-generation plant-derived biopharmaceuticals

Subramanian Parthiban[1], Thandarvalli Vijeesh[1], Thashanamoorthi Gayathri[1], Balamurugan Shanmugaraj[1], Ashutosh Sharma[2]* and Ramalingam Sathishkumar[1]*

[1]Plant Genetic Engineering Laboratory, Department of Biotechnology, Bharathiar University, Coimbatore, India, [2]Tecnologico de Monterrey, School of Engineering and Sciences, Centre of Bioengineering, Queretaro, Mexico

Recombinant biopharmaceuticals including antigens, antibodies, hormones, cytokines, single-chain variable fragments, and peptides have been used as vaccines, diagnostics and therapeutics. Plant molecular pharming is a robust platform that uses plants as an expression system to produce simple and complex recombinant biopharmaceuticals on a large scale. Plant system has several advantages over other host systems such as humanized expression, glycosylation, scalability, reduced risk of human or animal pathogenic contaminants, rapid and cost-effective production. Despite many advantages, the expression of recombinant proteins in plant system is hindered by some factors such as non-human post-translational modifications, protein misfolding, conformation changes and instability. Artificial intelligence (AI) plays a vital role in various fields of biotechnology and in the aspect of plant molecular pharming, a significant increase in yield and stability can be achieved with the intervention of AI-based multi-approach to overcome the hindrance factors. Current limitations of plant-based recombinant biopharmaceutical production can be circumvented with the aid of synthetic biology tools and AI algorithms in plant-based glycan engineering for protein folding, stability, viability, catalytic activity and organelle targeting. The AI models, including but not limited to, neural network, support vector machines, linear regression, Gaussian process and regressor ensemble, work by predicting the training and experimental data sets to design and validate the protein structures thereby optimizing properties such as thermostability, catalytic activity, antibody affinity, and protein folding. This review focuses on, integrating systems engineering approaches and AI-based machine learning and deep learning algorithms in protein engineering and host engineering to augment protein production in plant systems to meet the ever-expanding therapeutics market.

KEYWORDS

artificial intelligence, molecular pharming, synthetic biology, deep learning, machine learning

1 Introduction

Plant molecular pharming refers to the recombinant expression of biologics including vaccines, hormones, therapeutics and diagnostic reagents in plant-based systems. The field is gaining attention since the biologics produced from plants are efficient and similar to products from other conventional systems with the advantage of eukaryotic host performing post-translational modifications. Some of these recombinant biologics produced in plant systems are SARS-CoV2 virus-like particle (VLPs), spike antigen, anti-SARS-CoV2 mAb H4 and B38, anti-EBV (Ebola virus) mAb 6D8, 4H2 IgG and IgM (against *Coccidioides*), antimicrobial peptide (AMP) LL-37 and human apolipoprotein A-I_{Milano} (Apo A-I_{Milano}) (Fulton et al., 2015; Holásková et al., 2018; Ali and Kim, 2019; Shanmugaraj et al., 2020; Jugler et al., 2022; Zhao et al., 2023). Various model plant systems have been used as stable or transient heterologous expression hosts for recombinant protein production that include, tobacco (*Nicotiana benthamiana* and *Nicotiana tabacum*), Arabidopsis, tomato, potato, rice, maize, soybean, etc. (Ghag et al., 2021; Lobato Gómez et al., 2021). The plant host systems are useful in many aspects such as cost-effectiveness, multimeric protein assembly, scale-up and safety (minimal/no risk of human pathogen contaminations). Even with the listed advantages, there are few limitations to use plants as expression systems such as lack of humanized *N*-glycosylation post-translational modification which is needed for antibody production and stability of plant-produced proteins are still a concern (Sethi et al., 2021). Recombinant biologics production is dependent on several factors such as vector construction, codon optimization, regulatory components, protein localization and glycosylation (Amack and Antunes, 2020; Jin et al., 2022; Mirzaee et al., 2022; Moon et al., 2022; Zhao et al., 2023).

Systems Engineering in biology can be defined as a holistic approach that analyzes, models, alters, optimizes, and regulates the complex processes of biological systems resulting in desired functions. Artificial Intelligence (AI) refers to the development of machines and systems that use algorithms and statistical models to analyze data, identify patterns and can perform/outperform tasks that demand human intelligence in learning, reasoning, planning, communicating, and problem-solving (Russell, 2010). Machine Learning (ML) is a subset of AI that enables the systems to learn by providing abundant training datasets and is classified into supervised, unsupervised and semi-supervised learning algorithms. Supervised algorithms are the most used of the three since they are developed using labelled datasets from databases with minimum data redundancy, feature extraction, analysis & selection of main traits, prediction methods, and performance evaluation. They provide an excellent prospect for biologists in identifying patterns of gene expression and relevant features, thereby governing the identification through deep understanding of different combinations of the responsible factors (Singh et al., 2016; Silva et al., 2019). Deep Learning (DL) is a network-based supervised learning method with multiple layers of simple modules pooled and arrayed for learning, computing, and mapping a big dataset through each layer. It takes advantage over other AI-based ML algorithms in exploring complex structures of high-dimensional data built from the simplest layers (Lecun et al., 2015). Industry 4.0 revolutionizes traditional practices of manufacturing in industrial settings with the integration of digital technologies, automation, and data exchange, which concourses physical and digital systems leading to increased efficiency, productivity and innovation. Intervention of automation, cyber-physical systems, internet of things (IoT) and big data analytics would prove to be efficient and robust in plant-based biologics production (Dubey et al., 2018; Chen et al., 2020).

AI has been used in recombinant biologics production in host systems such as mammalian cells (CHO and HEK293), yeast (*Saccharomyces cerevisiae* and *Pichia pastoris*) and bacterial (*Escherichia coli* and *Bacillus subtilis*) systems (Van Brempt et al., 2020; Smiatek et al., 2021; Feng et al., 2022a; Li et al., 2022a; Packiam et al., 2022). Application of AI or ML algorithms include protein engineering, protein-protein interaction, stability, localization, solubility, functional motif prediction and catalytic activity which increases the production and functionality of recombinant proteins (Han et al., 2019; Jiang et al., 2021; Feng et al., 2022a; LaFleur et al., 2022; Masson et al., 2022; Kalemati et al., 2023). Till date, AI finds very least or no intervention in plant molecular pharming. In this review, we discuss about the systems biology concepts with the introduction of AI, as shown in Figure 1, in different aspects of recombinant biologics production to increase the stability, functionality and applications of AI-based ML algorithms in engineering systems to overcome the challenges and to enhance the production of next generation plant-based biologics.

2 Advantages of plant expression system

The market size of plant-based biologics was valued at $116.1 million during the year 2021, and with the compound annual growth rate (CAGR) at 4.8%, it is being estimated to reach $182.9 million by the year 2031. Few of the major plant-based production firms include Leaf Expression Systems, Zea Biosciences, Plant Biotechnology Inc., InVitria, Mapp Biopharmaceutical and PlantForm (Allied Market Research, 2023). Very few plant-based recombinant therapeutics have been commercialized following development and many are under clinical trials (He et al., 2021; Lobato Gómez et al., 2021). Elelyso, taliglucerase alfa, produced in carrot cell culture by ProtalixBio Therapeutics was approved by FDA in 2012 to treat Gaucher disease and has been commercialized (Mor, 2015). ZMapp – an antibody cocktail produced in *N. benthamiana* by Leaf Biopharmaceutical (commercialization arm of Mapp Biopharmaceutical) was used to treat Ebola outbreak under emergency use authorization during 2014 in Africa (Qureshi, 2016). Recombinant growth factors were produced in the endosperm of barley grain by ORF Genetics and have been commercialized as skincare products (ORF Genetics, 2023). Covifenz, a plant-based SARS CoV2 VLP vaccine against COVID19, developed by Medicago was authorized by Health Canada during 2022 (Hager et al., 2022).

Protein-based pharmaceutical products are growing rapidly in recent years and most of them are produced in mammalian and

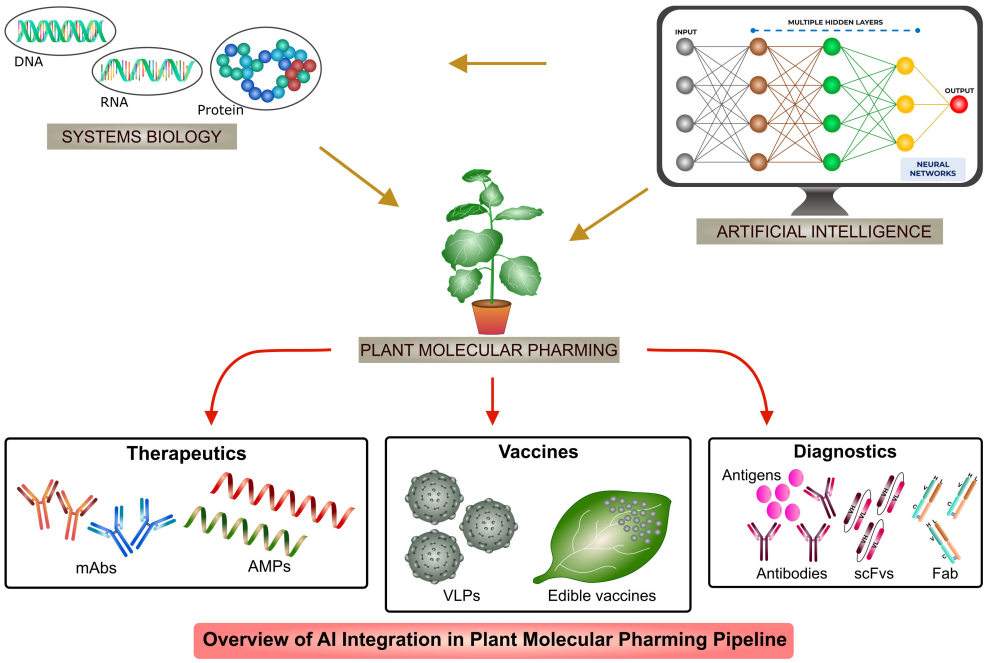

FIGURE 1
Overview of AI integration in plant molecular pharming pipeline.

microbial expression systems. Now-a-days, plant systems have emerged as an alternative platform for large scale production of recombinant proteins as they necessitate no capital-intensive infrastructure, bioreactors, or expensive culture media, but may be quickly scaled in low-cost greenhouses using simple reagents (Chen and Davis, 2016). When compared with prokaryotic and other host systems, plants offer an alternative bioreactor system for recombinant expression due to their glycan profile and cost-effective management system (Schillberg et al., 2019). Apart from the advantages mentioned above, plant systems are human pathogen free, sterile conditions are not required during production and scalable due to open-field cultivation (Buyel, 2019). For all these reasons plant expression system has been established as a prominent bioreactor for the production of therapeutic proteins such as vaccines, therapeutic proteins and growth hormones (Limkul et al., 2016; Moon et al., 2022).

Each expression host has its advantages and limitations. For instance, mammalian cell systems are capable of inherently producing recombinant biologics in humanized form, but it is difficult to maintain cell lines free from human pathogens and contaminants (Sethi et al., 2021). Plant system has many advantages over other systems including rapid (production of recombinant protein starts at day 2-3 post infiltration), cost-effective (produced at a cost of $0.27 for 3 mg dose of recombinant AMP), scale-up (increasing the plant biomass as required and thereby protein yield), purity (up to 99%), safety (production without any contaminant interference and functionally safe in humans) and post-translational modifications (*N*-glycosylation in engineered tobacco plants, which prokaryotic host system lacks). These advantages can be briefed with an example each using *N. benthamiana* transient expression host system. SARS-CoV2 RBD (Receptor binding domain) Fc fusion vaccine candidate was expressed in *N. benthamiana* and was extracted 4 days post infiltration which gave an yield of 25 µg/g FW (Siriwattananon et al., 2021). Alam et al. (2018) were able to produce antiviral compound Griffithsin at 99% purity from tobacco plant. Two mAb isotypes, 4H2 IgG and 4H2 IgM antibodies against *Coccidioides* CTS1 (Valley Fever) antigen were expressed in *N. benthamiana* plants showing homogenous N-glycosylation profile with a dominant GnGn/GnM structure, highly similar to mammals. Techno-economic analysis by McNulty et al. (2020) of *N. benthamiana*-based recombinant protein production reveals that the plant can produce up to 4 g of protein per kg FW (g/kg FW) with the yield up to 300 kg of recombinant protein per year through transient expression.

3 Systems engineering approaches to produce recombinant biopharmaceuticals in plants

Plant-based biologics have emerged as a promising alternative for therapeutics production due to their low-cost and scalable nature. This is critical for meeting the demand for immunizations during pandemics. Production of recombinant therapeutics in plants can be achieved by either stable or transient expression. Stable expression systems are developed by nuclear transformation or chloroplast transformation through *Agrobacterium*-mediated or biolistic gene transfer (Gelvin, 2003; Tien et al., 2019; Bolaños-Martínez et al., 2020; Heenatigala et al., 2020; Kumar and Ling, 2021). Meanwhile, transient expression systems are developed by plant virus-based vectors or agroinfiltration. Stable expression

systems possess advantages including scale-up, low storage costs, glycosylation patterns and reduced cross contamination of animal-borne agents; Transient expression systems are known for their rapid, cost-effective, increased protein accumulation and commercialization potential (Moon et al., 2019). Transient expression of recombinant biopharmaceuticals in plant system is the most preferred mode of production since the system accumulates large quantities of proteins quickly. Different immunogens and therapeutic agents have been produced through transient expression in leaves by agroinfiltration (Iyappan et al., 2018; Page et al., 2019; Rattanapisit et al., 2020).

Proteins reach functional state by proper folding, disulphide bond formation, subunit assembly and post-translational modifications. Prokaryotic host systems pose limitations such as lack of post-translational modifications (glycosylation and sialylation), signal peptide cleavage and pro-peptide processing (Gomord and Faye, 2004). Glycosylation is the most prevalent and diverse type of post-translational modification of proteins shared by all eukaryotic cells. A complex metabolic network and many glycosylation pathways are used during the enzymatic glycosylation of proteins to produce a wide variety of proteoforms (Schjoldager et al., 2020). For instance in humans, N-acetylglucosaminyl transferases IV and V present in Golgi functions in galactosylation, branch elongation and sialic acid capping, which is not found in plants (Strasser, 2022; Strasser, 2023). In order to produce therapeutic proteins of interest in plant with desired glycosylation pattern, β-1,4 galactosyl transferase co-expression and sub-cellular localization to Golgi is preferred (Navarre et al., 2017; Strasser, 2022). Recombinant glycoproteins produced in plants have residues of α1,3-fucose and β1,2-xylose linked to the same core N-glycan. These two sugar residues could be immunogenic since they are absent in human glycoproteins (Margolin et al., 2020a). In Arabidopsis, tobacco, and rice, multiplex CRISPR-Cas9 technology was used to knock out two glycosyl transferases, β1,2-xylosyltransferase and α1,3-fucosyltransferase, in order to humanize glycosylation patterns in plants and produced biopharmaceuticals. The results demonstrate that complete suppression of these two sugar residues was reported in Arabidopsis and tobacco, while the presence of Lewis structure in rice shows that the glycosylation pattern differs between dicots like Arabidopsis and tobacco and monocots like rice (Jansing et al., 2019; Jung et al., 2021). Many therapeutic proteins that are glycosylated need to be sialylated ultimately to fully activate their biological functions, however plants are not capable of N-glycan sialylation, in contrast to mammals. The ability to perform N-glycan sialylation is much sought after in the plant-based biopharmaceutical industry since sialic acids are a frequent terminal alteration on human N-glycans. Plants can be engineered across α2,6-sialylation or α2,3-sialylation pathways that showed active IgG with anti-inflammatory properties and increased pharmacokinetic activity of therapeutics produced in plants (Strasser, 2023). N-glycan sialylation is highly desirable due to its function in extended half-life, stability, solubility, and receptor binding (Bohlender et al., 2020; Chia et al., 2023). A whole mammalian biosynthetic pathway, including the coordinated expression of the genes for (i) biosynthesis, (ii) activation, (iii) transport, and (iv) transfer of Neu5Ac to terminal galactose, has been introduced into *N. benthamiana* in order to achieve *in planta* protein sialylation (Izadi et al., 2023).

Recombinant biologics expressed in plants are designed as fusion proteins to contain an N-terminal or C-terminal tag (His, FLAG, HA, CBM3 etc.) for easy purification and analysis. Immobilized metal-ion affinity chromatography is widely used for purification of hexahistidine tagged proteins (Vafaee and Alizadeh, 2018; Islam et al., 2019; Hanittinan et al., 2020; Islam et al., 2020; Marques et al., 2020; Soni et al., 2022). Other techniques such as one-step cation-exchange chromatography, Protein G-/A-based affinity chromatography, diafiltration (antibody purification) and polyelectrolyte precipitation (removal of plant proteins), hydrophobic interaction chromatography (HIC) followed by hydrophobic charge induction chromatography (HCIC) are employed in recombinant plant protein purification (Fulton et al., 2015; Park et al., 2015; Shi et al., 2019; Miura et al., 2020; Lim et al., 2022; Grandits et al., 2023).

4 AI-based ML algorithms in recombinant protein production

Gene designing and genetic engineering are key tools in molecular pharming, which enable the expression of protein of interest in host system, and development of genetically modified organisms with desirable traits. The design of gene and its expression cassette is the first step in getting desired protein in the plant system (Rozov and Deineko, 2019). Proper designing plays a major role in the production of biologics that includes selection of host system, codon optimization, regulatory components associated with foreign gene, host engineering, mode of expression, and purification of biopharmaceuticals (Webster et al., 2017; Peyret et al., 2019; Belcher et al., 2020; Sainsbury, 2020; Hassan et al., 2021; Vazquez-Vilar et al., 2023). AI-based ML algorithms are proven choice for cost-cutting and efficient designing of product manufacturing in different host systems. Few of the competent network models were built on Convolutional Neural Networks (CNNs), a DL architecture inspired from connectivity patterns of animal visual cortex to identify, locate and differentiate objects in any image (Barré et al., 2017). Different AI-based ML and DL algorithms have been developed to increase the recombinant biopharmaceutical production in the hosts by detecting, analyzing and optimizing the conditions such as screening and candidate selection, vector construction, codon optimization, protein modelling and design, growth condition optimization and protein solubilization and purification. A model architecture of CNN is shown in Figure 2.

4.1 AI in codon optimization

Introduction of native genes into alternate host system causes incompatibility in codon usage bias, sequence repeats, % of GC, negative cis-regulatory elements and Shine-Dalgarno sequence (Tuan-Anh et al., 2017; Constant et al., 2023; Jain et al., 2023).

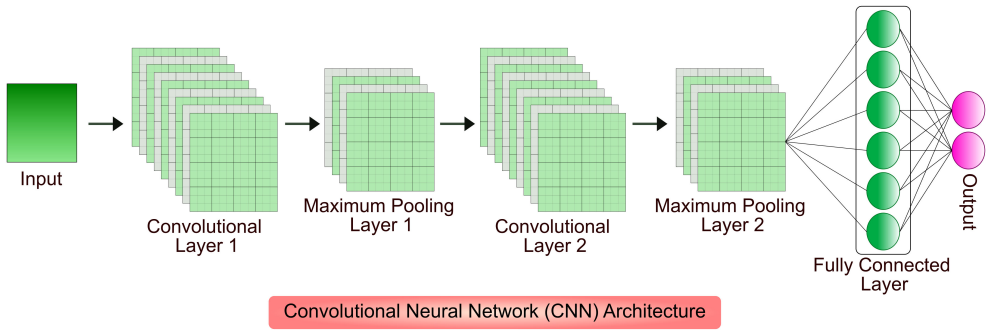

FIGURE 2
An illustration of input characteristics recognition using CNN. The input spatial features pass through multiple convolutional and pooling layers; processed data is received at a fully connected layer. The convolutional layer applies filters to extract features from input, pooling layer downsamples the features in order to reduce computation and fully connected layer makes the final prediction to result the output.

Codon bias affects the expression of transgene in the host plant which result in stopping at disfavored codons, truncation, misincorporation or frameshift. Site directed mutagenesis can resolve these problems by introducing silent mutations in coding region of the transgene and help the host species read transgene codon without any hindrance (Ma et al., 2003). Heterologous expression of recombinant proteins in different hosts needs optimization of coding sequences with synonymous codons as the host systems tend to remove heterologous proteins through proteolysis. Further, codon optimization renders the recombinant protein with structural and functional conformation at increased levels of expression in different host systems (Al-Hawash et al., 2017; Argentinian AntiCovid Consortium, 2020; Ding et al., 2022). The codon optimization percentage is proportional to the level of recombinant transgene expression. The amount of expression of the four variants of the *bar* gene with varying percentages of optimized codons was examined using experimental and *in silico* methods, and it was found that genes with 50–70% of optimized codons were expressed effectively in *N. tabacum* (Agarwal et al., 2019). Beta-defensin from chicken called chicken β Gallinacin-3 has demonstrated broad-spectrum antibacterial action against plant infections. Using DNAWORKS3.0 and the Genscript Rare Codon Analysis Tool, chicken β Gallinacin-3 gene sequences were codon optimized and tested. The results demonstrated constitutive expression in *Medicago sativa* and improved antibacterial activity against *E. coli*, *S. aureus*, and *Salmonella typhi* (Jin et al., 2022). Despite species difference, the codon optimizer program improved translation efficiency in tobacco and lettuce by using codon usage hierarchy of the *psbA* gene (Kwon et al., 2016). Adiponectin, an adipokine and a cell signaling protein, is produced as a secretory protein in *Withania somnifera* hairy root culture. Codon usage data, base composition and codon adaptive index (CAI) of *W. somnifera* were analyzed; the human adiponectin gene sequence was optimized and expressed as secretory product. Optimization of codons increased the expression levels of protein secretion (Dehdashti et al., 2020). The synthesis and expression of therapeutic proteins depend heavily on codon optimization. Effective methods are required to efficiently optimize codons for the generation of recombinant proteins in plants (Webster et al., 2017). Codon usage bias was utilized to optimize nucleotide sequences for host-specific expression in many systems including *E. coli*, Chinese Hamster Ovary (CHO) cells, HEK293, etc (Al-Hawash et al., 2017; Shayesteh et al., 2020; Lu et al., 2021). Till date, no AI tool has been designed to optimize codons for increasing the plant-based recombinant biologics production. The challenges posed by conventional methods include a vast possibility of codon combinations, irrational effects following transcription and translation, protein misfolding and loss of function (Constant et al., 2023).

Neural network (NN) models identify unexplored patterns in the native DNA sequences from the training set, predicts the most valid coding sequences using the test set and optimize DNA sequence for translation. The NN-optimization is found to be more efficient than conventional methods resulting in significantly higher yields of recombinant biologics (Goulet et al., 2023). Many sequence-based ML algorithms using deep neural networks (DNN) extract features from input codon data, predict and evaluate sequence data. Two major parameters that play a crucial role in codon optimization are 1) codon adaptation index (CAI) and 2) tRNA adaptation index (tAI). CAI is the frequency of codon usage in an organism's coding DNA sequence (CDS) and tAI is the measure of intracellular tRNA to translate into proteins and individual codon-anticodon pairing efficiency (Sabi et al., 2017; Tuan-Anh et al., 2017; Fu et al., 2020; Constant et al., 2023; Goulet et al., 2023). A Recurrent Neural Network (RNN) model trained sequence was tested for its efficiency by transient transfection of unoptimized and optimized sequences in CHO (ExpiCHO) cells. The titres of model protein, human programmed death ligand 1 (PD-L1) extracellular domain, were quantitated nine days after transfection. The RNN-optimized sequence was expressed largely (179.5 ± 12.4 μg/mL) than the native sequence (104.5 ± 5.7 μg/mL). The RNN model was used in optimization of mAb and stable integration of mAb CDS in CHO-K1-derived cells. The RNN-optimization of CDS yielded 2030 μg/mL and the unoptimized sequence resulted in an yield of 960 μg/mL (Goulet et al., 2023). Influence of AI in bacterial expression system is more than any other eukaryotic systems and so codon optimization was widely carried out through ML-based models. Tuan-Anh et al. (2017) used

neural network with CAI and GC content for optimizing codons expressing prochymosin, the chymosin-precursor in *E. coli* system. Codon optimization could preferably not just used for increasing heterologous recombinant expression, but also for increasing the protein solubility. MPEPE, a newly developed protein solubility prediction DNN model was built using convolution layers, pooling layers and long-short term memory (LSTM) layers. The architecture was built as embedded matrix, through 'one-hot encoding' technique using integers '1' and '0', to include synonymous codons of individual amino acids. Point mutation in sites was scrutinized through evolutionary analysis without interfering the protein function. The target nucleotides for expression studies were used as inputs in MPEPE for virtual screening and recombinant proteins were expressed in *E. coli* BL21 (DE3) cells with an increased level of soluble protein expression (Ding et al., 2022). Bidirectional LSTM Conditional Random Field (BiLSTM-CRF) model is a codon optimization model built for *E. coli* by H. Fu et al. (2020). The model converts codon optimization to sequence annotation and trains the data of *E. coli* gene set through word-embedding vector. The multivalent *Plasmodium falciparum* vaccine antigen FALVAC-1 and PTP4A3, a prognostic cancer biomarker optimized by BiLSTM-CRF were expressed in *E. coli* BL21 (DE3). The model efficiently optimized the low-expression candidate to higher expression levels, which proved the robustness of the model and the high expression candidate PTP4A3 was expressed in similar levels which proved the stability of algorithm. Jain et al. (2023) designed ICOR (Improving Codon Optimization with RNNs), a DL tool, built on BiLSTM architecture through 'one-hot encoding' method, with a large non-redundant dataset of *E. coli* genomes and upon correlation comparison with the mRNA expression in real-time based on a work by dos Reis et al. (2003), the improvement in expression observed was about 236%. The multilayer network model may be trained for other host systems including model plants (such as *N. benthamiana* or *N. tabacum*) as shown in Figure 3 with complete omics dataset through transfer learning approach to increase the yield. CO-BERTa, a deep contextual language model was trained with GFP (Green Fluorescent Protein) and anti-HER2 VHH CDSs on *Enterobacterales* dataset for functional protein measurement. The mCherry reporter protein which showed 28.7% pairwise identity to GFP and anti-SARS-CoV2 VHH which showed 73.7% pairwise identity to anti-HER2 VHH was chosen to test the model. These proteins differ in their length but share similar structural features, a major feature being β-barrel. ACE (Activity-specific Cell Enrichment) measurement of CO-BERTa codon optimized proteins in SoluPro^{TM} *E. coli* B strain showed highest expression levels than commercial algorithms (except Genewiz, $p<0.05$) (Constant et al., 2023). Further, genome analysis and codon usage patterns of plant host systems through artificial neural networks (ANNs) could significantly increase the expression of recombinant biologics (Doyle et al., 2016).

Quantum computers can be used to optimize codons for high expression of proteins. Quantum Annealing (QA) algorithm uses quantum computers to give high-dimensional combinatorial optimization of codons using Binary Quadratic Model (BQM) built on 'one-hot encoding' technique. mRNA codons of peptide fragments and full length proteins of SARS-CoV2 spike glycoprotein were optimized using Quantum Approximate Optimization Algorithm (QAOA) (Fox et al., 2021).

Currently, there are no ML-based algorithms available for codon optimization of recombinant proteins to express in plants. The algorithms available for other host systems could be adapted, remodelled and designed for plant-based expression hosts since many of the model plants' genome is available publicly.

4.2 AI in protein modelling and design

The recombinant proteins expressed in different systems are influenced majorly by factors including structure, solubility, catalytic activity, protein folding and stability. Vector and gene of interest is designed to overcome the challenges of recombinant protein expression. The components of protein modelling include host and expression vector selection, promoter, selectable marker, fusion tags. ML based algorithms enhance the expression and overcome the challenges in expression of recombinant biologics in multiple expression systems. These algorithms analyses and tests (either nucleotides – CDS/RNA-seq or amino acids) sequences and provides with the fitness of protein variants (Wittmann et al., 2021). Few ML models utilize structure along with sequences of amino acids for modelling of proteins. The RNNs and other neural network models are powerful than other ML models since these could learn from raw data directly without any sequence alignment and heuristic scoring (Deep RNN for Protein Function Prediction from Sequence). While molecular dynamics simulations for an antibody through supercomputers require hours and even days, neural networks such as CNN models take only seconds to get the work done in personal computers (Lai, 2022). Regulatory elements are one of the key components of recombinant protein production and synthetic promoters have been designed using ML models to increase the transcription efficiency. Highly functional Synthetic Promoters with Enhanced Cell-State Specificity (SPECS) were identified from a library of 6107 promoters using multiple ML regression algorithms, from which a generalized linear model with elastic net regularization (GLMNET) was chosen as the efficient model to predict highly active promoters. The spatiotemporal activity of each promoter was analyzed by expression of fluorescent protein in HEK-293T cells (Wu et al., 2019). In the work by Vo ngoc et al. (2020), human PolII core promoter was analyzed to create HARPE (high-throughput analysis of randomized promoter elements). The HARPE training dataset included 200,000 variants of promoter sequences and downstream core promoter region (DPR) models were generated by support vector regression (SVR) algorithm and tested *in vitro* and in HeLa cells. Designing protein includes predicting counterparts, which are involved in structural integrity and stability of proteins (Masson et al., 2022). These include epitope prediction, vaccine designing and remote homology detection, which utilize parts of the protein molecule to increase its activity (Mettu et al., 2016; Moss et al., 2019; Yang et al., 2021b; Koşaloğlu-Yalçın et al., 2022; Routray et al., 2022).

Using DeepLoc, a deep convolutional network Kraus et al. (2017) showed improved performance over traditional approaches

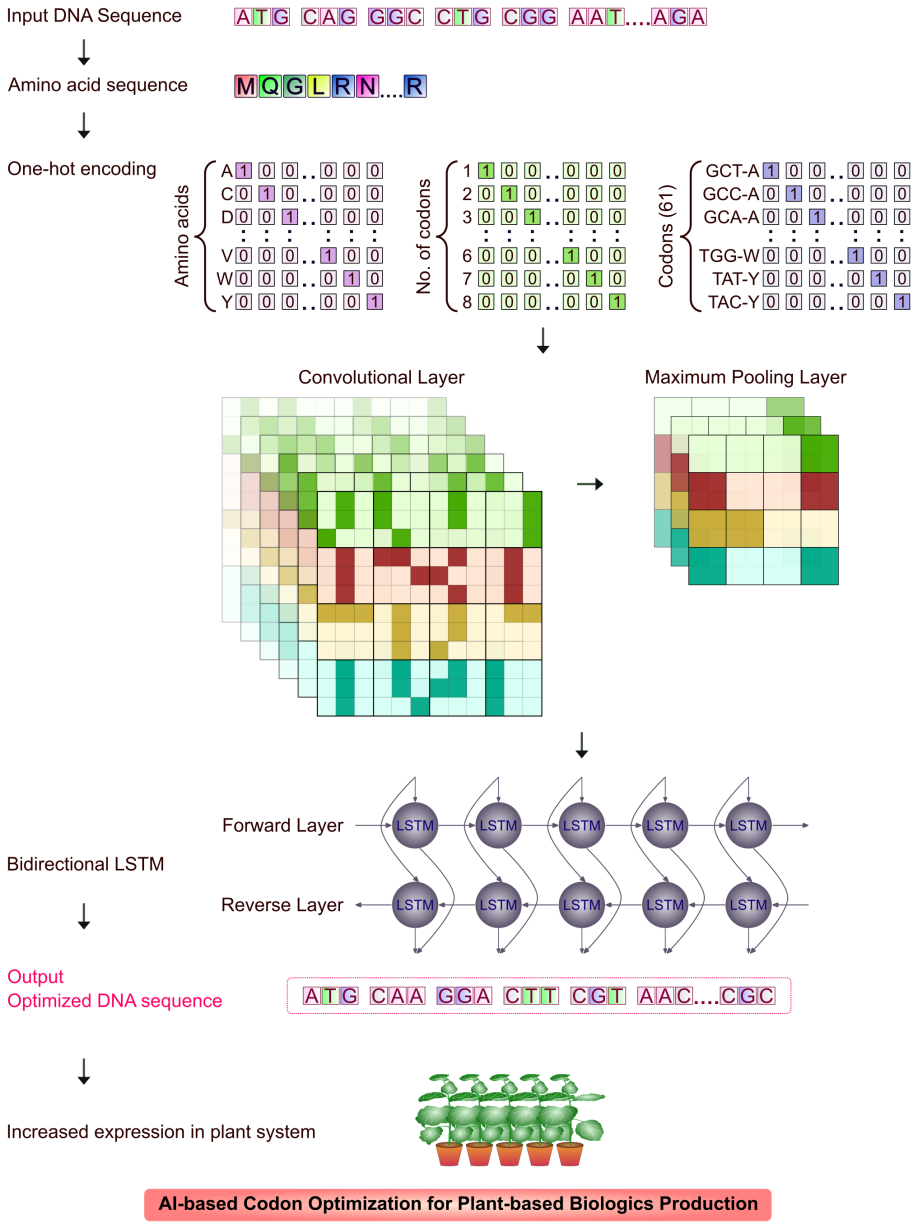

FIGURE 3
An overview of CNN and BiLSTM in codon optimization for plant expression system. Input DNA sequence is translated to amino acid sequence and reverse translated to DNA sequence. The sequence is further tested using CNN for codon optimization metrics prediction for plant expression system and BiLSTM tests with synonymous codons for optimal codon usage for high expression in plant system.

in the automated classification of protein subcellular localization in yeast cells. Organelle targeting and sub-cellular localization increases the recombinant therapeutic protein expression in plants to higher levels. Localization of recombinant proteins in cytosol and different plant organelles such as nucleus, chloroplast, mitochondria and endoplasmic reticulum (ER) of plant tissues such as seeds and leaves are useful in increased accumulation and stability of expressed proteins (Vafaee and Alizadeh, 2018; Arcalis et al., 2019; Bidarigh fard et al., 2019; Islam et al., 2019; Shi et al., 2019; Hanittinan et al., 2020; Islam et al., 2020; Li et al., 2022b; Lim et al., 2022). Signal sequences are added to N-terminus or C-terminus of the biologics to increase the yield and a C-terminal ER retention signal is the most widely used strategy to accumulate higher amount of proteins in recombinant expression. Sahu et al. (2021) developed a tool, Plant-mSubP, based on integrated ML approaches with SVM as the model to predict localization of proteins to single and dual organelle targets.

Analysis of the enriched bococizumab yeast cell libraries along with similar library for antibody affinity was done using an ML model, which enabled the identification of rare variants with co-optimized levels of low self-association and high affinity (Makowski et al., 2022). Similarly, mAbs can be screened and optimized for production in specific host systems that could include plants as well (Feng et al., 2022a; Lai, 2022). Proteins such as toxins which are

difficult to produce in certain hosts can be expressed easily using deep-learning based CNN algorithms (Pan et al., 2020). A wide range of ML algorithms used in various eukaryotic and prokaryotic systems for modelling different proteins is shown in Table 1.

4.3 ML models in engineering strains for recombinant protein production

A large repertoire of omics data is obtained from the host system at different levels of replication (genome), transcription (transcriptome), translation (proteome), and regulation (metabolome). These data can be used to engineer host cells to improve recombinant protein yield (Ramzi et al., 2020; Samoudi et al., 2021). ML algorithms can be implemented in understanding the genome-scale metabolic models (GEMs), which encompasses hundreds of metabolic pathways and thousands of metabolic reactions. ML can be a stand-alone or a complementary approach, in learning regulatory levels of complex pathways in plants such as transcriptional, translational and allosteric regulation. These ML algorithms are shown to exhibit more robustness than statistical tools (Radivojević et al., 2020; Zhang et al., 2020; Strain et al., 2023).

Multilayer Perceptron (MLP), an NN model was used to analyse the human RNA-seq data from ARCHS4 database based on secretory index (SI) and extrapolated to engineer CHO cells (Zaragoza, 2022). In order to predict yeast cell growth Culley et al. (2020) proposed ML–based data integration techniques, combining gene expression profiles that rigorously assess and compare with computationally generated metabolic flux. A total

TABLE 1 AI in protein modelling and design.

Component	Name of the program	Type of ML algorithm	Architecture	Function/ Parameter	Model system/ training dataset	References
mRNA	APARENT (APA REgressionNeT)	CNN	One-hot encoded matrix system with two convolutional layers	• mRNA isoform prediction and polyadenylation within +10 to +35 nt downstream of 6-base central sequence element (CSE) • cleavage site prediction across polyA signal	HEK293	Bogard et al. (2019)
	6-mer Logistic Regression Baseline	Linear logistic regression	One-hot encoded matrix system with 6-mer counts	• mRNA isoform prediction and polyadenylation • cleavage site prediction		
mRNA, gene enhancers and protein	DEN (Deep Exploration Network)	Deep Convolutional Generative Adversarial Networks (DC-GANs)	One-hot encoded matrix Latent Seed Sequence Tensor	• polyadenylation signals conformed to mRNA isoforms and 3' cleavage sites • differential splicing • maximum transcriptional activation of gene enhancers • functional variants of GFP (Green Fluorescent Protein)	HEK293 HeLa MCF7 CHO	Linder et al. (2020)
	APARENT	CNN				
	-	GP regression				
	APA VAE (Variational Autoencoder)	Residual Neural Network (ResNet)				
	KL-bounded DEN	CNN				
Gene interaction and expression	scCapsNet	DNN	Capsule Neural Network	Discovery of gene interactions; closely related in function but presenting differential gene expression pattern in single cell types (based on transcriptome analysis)	scRNA-seq dataset including mouse retinal bipolar (mRBC) cells and human peripheral blood mononuclear cells (hPBMC)	Wang et al. (2020)
Transcription factor	Independent Component Analysis (ICA)	Unsupervised ML	-	Gene expression and transcriptional regulation in *E. coli* through transcriptome analysis	*E. coli* K12 RNA-seq expression profiles	Sastry et al. (2019)
Transcription factor binding	FactorNet	Convolutional RNN	One hot encoded 4-row bit matrix, LSTM	Transcription Factor (TF) cell type specific binding site prediction. (Eg.TF E2F1	DNase-seq, ChIp-seq and RNA-seq data of chromosomes X and	Quang and Xie (2019)

(Continued)

TABLE 1 Continued

Component	Name of the program	Type of ML algorithm	Architecture	Function/ Parameter	Model system/ training dataset	References
				binding to GM12878 and HeLa-S3)	1-22 from ENCODE-DREAM challenge	
Promoters	Hybrid biophysical-ML approach	Ridge regression model	-	• Synthetic promoter designing • Identification of -35 and -10 motifs and optimal spacer length	*E. coli*	LaFleur et al. (2022)
Synthetic promoter	DL model	Deep CNN	Transformer model with BiLSTM	Design regulatory sequences including orthologous promoters	RNA-seq data from *S. cerevisiae* and 10 other Ascomycota species	Vaishnav et al. (2022)
Protein	DeepRHD	DNN	CNN based bidirectional GRU (Gated Recurrent Units)	Remote homology prediction of protein sequences using physico chemical properties and evolutionary information	SCOP1.67 dataset	Routray et al. (2022)
Protein	ProtT5	pLM (protein language models) Logistic Regression	Attention based deep dilated residual networks consisting of convolution layers (ResNet CNN)	Protein (transmembrane beta barrel proteins – OmpX and variants) structure prediction from sequences	High resolution protein 3D structure dataset from ProteinNet12	Weissenow et al. (2022)
Protein	ML model	Linear regression models including glmnet, partial least squares, averaged neural network, SVM with radial basis function kernel, stochastic gradient boosting, boosted generalized linear model, random forest, cubist and naïve Bayes models	Caret package in R	Factors influencing recombinant protein stability including Molecular weight, cysteine residues and N-linked glycosylation	CHO cells expressing human secretome	Masson et al. (2022)
Protein	ASPIRER	DL model	XGBoost and N-terminal sequence-based CNN	Prediction of Non-classical secreted proteins (NCSPs)	Gram positive bacteria NCSPs dataset from UniProt	Wang et al. (2022)
Protein	eUniRep	DL NN	UniRep multiplicative LSTM	Protein, avGFP and TEM-1 β-lactamase, engineering (Low-*N* engineering) using small number of functional variants	*E. coli* DH5α	Biswas et al. (2021)
Protein	UniRep	SVM	RNN	Prediction of recombinant gene expression and protein solubility	*B. subtilis*	Martiny et al. (2021)
		LR				
		Random Forest (RF)				
		ANN				
Protein	ECNet	RNN	BiLSTM, Transformer architecture with TAPE integration	Protein fitness prediction based on evolutionary context, engineered TEM-1 β-lactamase variants showing enhanced ampicillin resistance	*E. coli* DH5α Diverse large-scale deep mutational scanning (DMS) datasets and random mutagenesis datasets	Luo et al. (2021)
Protein	EPSOL	Keras based DL model	Multidimensional Embedding, multi-convolutional-pooling	Protein solubility prediction	Heterologous expressed *E. coli* soluble and insoluble	Wu and Yu (2021)

(Continued)

TABLE 1 Continued

Component	Name of the program	Type of ML algorithm	Architecture	Function/ Parameter	Model system/ training dataset	References
			module and a Multi-layer Perceptron (MLP)		protein dataset compiled by Smialowski et al. (2012)	
Protein	DEEPred	Multi-layered perceptrons (MLPs)	Feed-forward multitask DNN	Sequence/Gene Ontology (GO) based functional definition prediction of proteins	*Pseudomonas aeruginosa* strain reference genome and UniProtKB/Swiss-Prot dataset	Sureyya Rifaioglu et al. (2019)
Protein	ML models	GANs	Generator Neural Network and Discriminator Neural Network	Prediction of Protein solubility	eSol database dataset	Han et al. (2019)
		Logistic regression				
		Decision Tree				
		SVM				
		Naïve Bayes				
		Cforest				
		XGboost				
		ANNs				
Protein	DeepSol	DL model	CNN, non-linear high-dimensional k-mer vector spaces, deep feed-forward neural network (FFNN)	Protein solubility prediction	Heterologous expressed *E. coli* soluble and insoluble protein dataset compiled by Smialowski et al. (2012)	Khurana et al. (2018)
Protein	ML	RNN	BiLSTM, One-hot encoded matrix	Identification and function prediction of protein homologs including iron sequestering proteins, cytochrome P450, serine and cysteine proteases and G-Protein coupled receptors, detection through fluorescence (GFP)	*E. coli*	Liu (2017)
Protein	SPIDER2	Deep learning neural network	Stacked sparse autoencoder	Protein secondary structure, solvent accessible surface area, main chain torsion angle prediction	Non-redundant high resolution protein structures dataset	Yang et al. (2017)
Amyloidogenic proteins	AbsoluRATE	SVM	Sequence-based regression	Aggregation kinetics prediction of amyloidogenic proteins	CPAD 2.0 database dataset	Rawat et al. (2021)
Antibody	DeepAb	Deep residual convolutional network (Deep RCN) with Rosetta-based protocol	RNN, BiLSTM, LSTM	Antibody Fv structure prediction from sequence	Observed Antibody Space (OAS) database, SAbDab database	Ruffolo et al. (2022)
Antibody	DeepH3	Deep residual network	One dimensional and two dimensional convolutions	Prediction of *de novo* CDR H3 loop structures	Rosetta and SAbDab dataset	Ruffolo et al. (2020)
mAbs	solPredict	ESM1b-based Multilayer perceptron (MLP2Layer)	Pretrained protein language model EMS1b embedding	• Rapid, large-scale high throughput screening of mAb sequences (IgG1, IgG2 and IgG4) and quantitative solubility prediction	HEK293/CHO	Feng et al. (2022a)

(Continued)

TABLE 1 Continued

Component	Name of the program	Type of ML algorithm	Architecture	Function/ Parameter	Model system/ training dataset	References
		transfer learning model		eliminating precipitation in Histidine pH 6.0 (H6) buffer system • Eliminates the need for 3D modelling		
mAbs/IgG1	DeepSCM	Scikit-learn	CNN architecture	Molecular dynamics simulation to screen high concentration antibody viscosity prediction	SAbDab and AbYsis database dataset	Lai (2022)
	Keras v2.7.0	-				
Multiepitope vaccine	DeepVacPred	DNN-V	Multi-layer CNN and a 4-layer linear neural network	Designing vaccine subunit containing both T- and B-cell epitopes of Spike glycoprotein against SARS-CoV2	*E. coli* K12	Yang et al. (2021b)
T-cell Epitope	Antigen eXpression based Epitope Likelihood-Function (AXEL-F)/ NetMHCpan 4.1 combination	-	Neural networks	• Expression of source antigen; T cell epitope prediction and peptide presentation to MHC Class I molecule • SARS-CoV2 epitope prediction	IEDB HLA class I ligands dataset; RNA-Seq data of HeLa cells; SARS-CoV2 expression dataset from Finkel et al. (2021)	Koşaloğlu-Yalçın et al. (2022)
T-cell Epitope	-	Epitope likelihood	Aggregate z-score, structure-based processing likelihood	*P. aeruginosa* endotoxin domain III (PE-III) epitope prediction	*P. aeruginosa*	Moss et al. (2019)
T-cell Epitope	-	Epitope likelihood	Aggregate z-score	CD4+ T-cell epitope prediction in bacterial and viral antigens without genotype information through antigen processing constraint modelling	Sequence data from different studies in C57BL/6 mice, HLA-DR4-transgenic mice and humans	Mettu et al. (2016)
Protein localization	MULocDeep	Bayesian optimization & Attention visualization	LSTM	Protein localization in organelles such as nucleus, mitochondria, plastid and thylakoid and extracellular matrix	Mitochondrial proteome data of *A. thaliana* cell cultures, *Solanum tuberosum* tubers, *Vicia faba* roots	Jiang et al. (2021)
Protein localization	Plant-mSubP	SVM	OvR (One-vs.-Rest)	Single- and dual- organelle targeting/subcellular localization of proteins in plants	Plant protein sequence dataset from Uniprot Database	Sahu et al. (2021)
Cytokines and peptides	ProtConv	Transfer learning CNN	LSTM, ResNet and Transformer with TAPE embedding LeNet-5 architecture	Function prediction of proinflammatory cytokines and anticancer peptides	IEDB and CancerPPD database dataset	Sara et al. (2021)
Peptide	FBGAN (Feedback GAN)	GANs	RNN and Feedback loop training architecture	• Generation of synthetic AMPs and α-helical peptide coding genes • Optimization of secondary structure	Uniprot database dataset	Gupta and Zou (2018)
Peptide-MHC Class I binding	CapsNet-MHC	CNN	Capsule Neural Network	Prediction of interaction between allelic variants of MHC and peptides with rare sequence lengths	IEDB dataset	Kalemati et al., (2023)
Peptide-HLA binding	DeepSeqPanII	Pan-specific DNN with attention mechanism	LSTM	Prediction of Peptide-HLA Class II binding	IEDB datasets BD2013 and BD2016	Liu et al. (2022b)

(Continued)

TABLE 1 Continued

Component	Name of the program	Type of ML algorithm	Architecture	Function/ Parameter	Model system/ training dataset	References
MHC Class II Antigen Presentation	NNAlign_MAC	ANN	NNAlign_MA ML framework	• CD4 T cell epitope prediction • MHC class II antigen presentation prediction • Prediction of protein-drug immunogenicity	Single allele and Multiple allele dataset & IEDB dataset	Barra et al. (2020)
Signal Peptide	XGBoost	Regression model	-	Increasing the protein translocation rates to ER by optimizing synthetic signal peptide-protein (mAb/ScFv) complex formation	CHO-K1 cells	O'Neill et al. (2023)
Signal peptide	Sequence-to-sequence model	Attention-based neural network	Transformer model	Signal peptide prediction from Amylase, lipase, protease and xylanase enzymes	*B. subtilis*	Wu et al. (2020)
Signal peptide	SignalP 5.0	DL model	Non-linear PSSMs (position specific scoring matrix), BiLSTM and a conditional random field	Peptide identification (three classes including Sec/SPI, Sec/SPII, Tat/SPI) in prokaryotes	Reference proteomes of *E. coli* K12 and *S. cerevisiae*	Almagro Armenteros et al. (2019)
Toxic motifs	ToxDL	Deep CNN	Bidirectional GRU, one-hot encoded matrix	Toxicity assessment of genetically engineered organisms by highlighting toxic motifs and alteration of toxicity	Toxic/venom protein dataset from Animal Toxin Annotation Project in UniProt	Pan et al. (2020)
	Domain2Vec		Skip-gram model			
NSAID	Ensemble Decision Tree (DT)	Extremely Random Tree (ET)	Multiple base trees with bagging strategy	Non-steroidal anti-inflammatory drug, Oxaprozin, solubility in supercritical CO_2 fluid	Oxaprozin solubility dataset from Khoshmaram et al. (2021)	Alshehri et al. (2022)
		Random Forest (RF)				
		Gradient Boosting	Sequence of base predictors			

of 1,143 *S. cerevisiae* mutants were tested and 27 machine learning methods were analyzed.

ART (Automated Recommendation Tool) and EVOLVE algorithm are ML-based Bayesian ensemble optimization tools used in increasing the production of tryptophan in yeast, *S. cerevisiae*. These ML algorithms were used to design 30 different promoter combinations from the transcriptome dataset, which were used to predict engineered strains to show increased productivity. The engineered strain SP606 was found to possess higher synthesis rate of proxy GFP than other strains designed using ML and library preparation. Also, the engineered yeast strain SP606 was identified to have an increased titre and productivity of tryptophan (Zhang et al., 2020). ART was also trained with concentration dataset of proteins/enzymes involved in heterologous pathway for the production of limonene. New strain design sets of *E. coli* for enhanced production of limonene were provided by ART (Radivojević et al., 2020).

Similarly, supervised learning algorithms have predicted pathway dynamics with the use of multiomics data (proteome and metabolome data) in *E. coli* for enhancing limonene production (Costello and Martin, 2018). In contrast, an unsupervised ML approach termed as HybridFBA, was proposed by Ramos et al. (2022) that combined GEM and metabolic flux balance analysis (FBA) using principle component analysis (PCA) in CHO cells (Strain et al., 2023). Machine Learning Predictions Having Amplified Secretion (MaLPHAS) by Eden Bio Ltd is an ML algorithm that predicted knock out of five genes, out of which Component of Oligomeric Golgi Complex (*cog6*) knockout strain resulted in doubled secretion of recombinant protein in the host *Komagataella phaffii* (*P. pastoris*) compared with the *bgs7* supersecretor strain (Markova et al., 2022).

DCell is a virtual eukaryotic cell composed of 2,526 subsystems embedded as VNNs (visible neural networks), a deep ANN, in hierarchy. The model was built using the hierarchical architecture of subsystems of *S. cerevisiae*. Being trained on several million genotypes, during simulation, DCell generates patterns of molecular activities based on genotype to phenotype relationship (Ma et al., 2018). DCell can identify gene deletions/knockouts using Gene Ontology (GO), which will result in phenotype change (Ma et al., 2018; Kim et al., 2020).

The ML algorithms and tools can be used to introduce or remove genes from a pathway to direct the increased production of humanized recombinant biologics in plant system. Knock-out approach of removing plant-specific glycans [β(1,2)-Xyl and

α(1,3)-Fuc] or knock-in strategy to express human [β(1,4)-Gal]and addition of sialic acid residues in specific host plants result in humanized protein expression. Such mechanisms could be explored and analyzed through ML tools such as ART (Sethi et al., 2021). Also, metabolic flux of host plant systems can be studied to generate stable lines with optimized metabolic pathways for desired post translational modifications of recombinant biologics.

4.4 Automation and AI in plant growth monitoring and biomass production

One of the big attributes of plant molecular pharming for recombinant biologics production, next to host selection and engineering is plant growth and maintenance. Plants are efficient biofactories for the manufacture of recombinant proteins and growth monitoring is a vital aspect when it comes to both laboratory scale and commercial production. Several automation technologies including affordable sensors built on Raspberry Pi, robotics and high-definition cameras work based on image acquisition (Jahnke et al., 2016; Jolles, 2021; Banerjee et al., 2022; Wan et al., 2022). The camera sensors have been deployed to analyze the plant growth patterns, phenotypes such as plant morphology, height, canopy, temperature, leaf biomass, leaf area index, greenness, age and different stresses. Similarly, seed count, shape, size and color, parameters for plant growth such as temperature, photoperiod, grow light color, etc. were studied by robot-assisted systems. A large training dataset of raw images captured in the camera sensors are analyzed through DNN modules and processed for color correction and segmentation for analysis (Jahnke et al., 2016; Ubbens and Stavness, 2017; Tovar et al., 2018; Zheng et al., 2019; Tausen et al., 2020; Bose and Hautop Lund, 2022). The efficient analysis of images are carried out by models based on CNNs that include U-Net, R-CNN and ResNet (Ubbens and Stavness, 2017; Lin et al., 2019; Zheng et al., 2019; Tausen et al., 2020; Bose and Hautop Lund, 2022). The IoT based sensors and programs are not limited to phenotyping the growth and morphology of plants but could detect plant nutrient deficiencies, diseases and soil parameters, thereby reduce the labor intensive maintenance and increase the sustainability (Dhivya et al., 2021; Monteiro et al., 2021; Bose and Hautop Lund, 2022). Plant monitoring and phenotyping using integrated automation and ML approaches is illustrated in Figure 4.

With the wider and large-scale biologics production environment, a large number of sensors in plant monitoring are needed and it becomes highly difficult to build the architecture for plant maintenance. Hence remote sensing using unmanned aerial vehicles (UAVs) is used in place at low altitudes to acquire high-resolution multispectral images of plants grown in agricultural field and greenhouses. The UAV high-throughput phenotyping

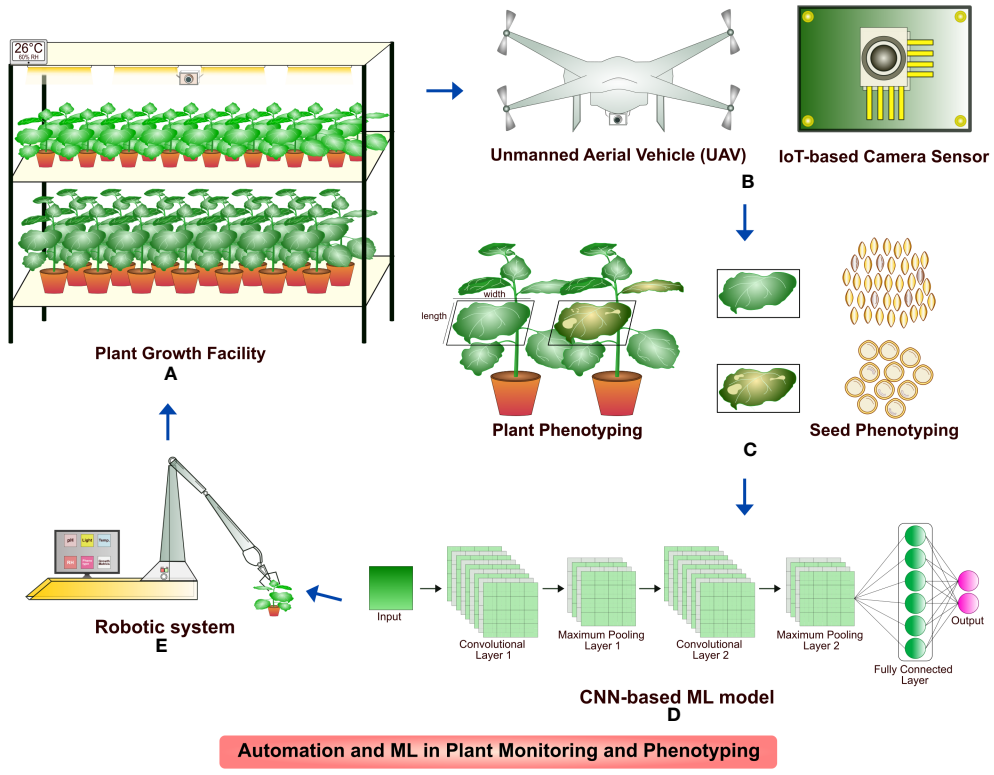

FIGURE 4
An illustration of plant monitoring and phenotyping with the integration of automation and ML approaches. **(A)** and **(B)** *N. benthamiana* plants grown in containment facility monitored continuously by image sensors; UAV deployed for plant monitoring in greenhouse; **(C)** Image acquisition of plant phenotypes such as height, width, leaf greenness, disease identification, seed count, morphology and segregation; **(D)** Phenotyping of plants by CNN-based ML model; **(E)** Robotic system for optimization and maintenance of environmental requirements for plant growth trained by CNN model.

platform, working on support vector machine (SVM) and SVM-derived models, processes the spectral information of optical images for the identification of plant growth, biomass, stress and disease stages (Maimaitijiang et al., 2020; Fu et al., 2021; Yang et al., 2021a; Aslan et al., 2022; Jiang et al., 2022a; Bai et al., 2023a). Several plants used as hosts in production of recombinant biopharmaceuticals such as *Glycine max* (L.) Merr. (soybean), *Triticum aestivum* (wheat), *Hordeum vulgare* (barley), *Oryza sativa* (rice), *Zea mays* (maize), *Arachis hypogaea* L. (peanut), *Arabidopsis thaliana* (Arabidopsis), *Brassica napus* (rapeseed), *Lycopersicon esculentum Mill.* (Tomtato), *Cucumis Linn.* (cucumber), *L. sativa Linn.* (lettuce), *Brassica oleracea linn.* (cabbage), *Raphanus sativus linn.* (turnip), *Apium graveliens Linn.* (celery) and *Spinacia oleracea Linn.* (spinach) and *N. tabacum* (tobacco) can be monitored using the sensors for high product yield (Minervini et al., 2015; Jahnke et al., 2016; Minervini et al., 2017; Ubbens and Stavness, 2017; Zheng et al., 2019; Maimaitijiang et al., 2020; Fu et al., 2021; Sangjan et al., 2021; Sarkar et al., 2021; Yang et al., 2021a; Banerjee et al., 2022; Bai et al., 2023a; Bai et al., 2023b; Sun et al., 2023). A detailed list of automation and AI-based tools used in plant monitoring is listed in Table 2. These technologies are not limited to monitoring the mentioned plants but can be extended to all the plant host systems used in expression of recombinant biologics.

4.5 ML approaches in cell suspension cultures and bioreactors

Plant cell suspension cultures offer a unique platform for the production of recombinant proteins due to their ability to perform post-translational modifications similar to mammalian cells (Gutierrez-valdes et al., 2020). Plant cell suspension cultures are usually prepared from callus tissue in shaker flasks or fermenters to form single cells and small aggregates and growing plant cells in a liquid medium in a controlled environment, such as bioreactor, where various factors like temperature, pH, and ratio of nutrient are to be optimized for cell growth and protein production (Cardon et al., 2019). Several proteins have been produced in bioreactor using cell suspension cultures including ORF8, an accessory protein of SARS-CoV2 in suspension cultured tobacco BY-2 cells (Imamura et al., 2021), rrBChE, rice recombinant butyrylcholinesterase in rice cell suspension culture (Macharoen et al., 2021), LBT-Syn protein in carrot cell suspension culture (Carreño-Campos et al., 2022), taliglucerase (ELELYSO), a recombinant version of human glucocerebrosidase in carrot cell cultures (Mor, 2015) etc.

Large scale production of plant-expressed recombinant proteins can be achieved by growing the transformed plant cell in different bioreactor shapes, however, there are diverse problems to be addressed such as pH of media, minerals, growth regulators, cell density, gaseous atmosphere, agitation system and sterilization conditions (Ruffoni et al., 2010).

Now-a-days AI techniques are increasingly being applied to bioreactors, which are essential tools in bioprocessing for the production of various biological products such as recombinant proteins, vaccines, and biofuels. ML models can identify the optimal operating conditions, such as temperature, pH, dissolved oxygen, and nutrient concentrations, to maximize product yield and quality. By integrating with sensors, data acquisition systems and control algorithms, AI models can analyze data in real time and automatically adjust process parameter accordingly. AI can adapt and adjust process parameters for optimal performance, reducing the need for manual intervention.

Optimizing plant tissue culture media is a complicated and time-consuming process, which is influenced by genotype, mineral nutrients, plant growth regulators, vitamins and other factors. ML approaches such as multilayer perceptron neural network (MLPNN), k-nearest neighbors (KNN) and gene expression programming (GEP) were used for developing prediction models in optimizing plant tissue culture media composition (Hosseini et al., 2022). In another work, three ANN models: CIPnet, CWnet and DCnet were developed to predict the best media composition for callus weight (CW), callus induction percentage (CIP) and days to callus initiation (DC). The performance was satisfactory and showed the R^2 values of 0.95, 0.95 and 0.88 for CIPnet, CW, and DCnet respectively (Munasinghe et al., 2020). The formation of foam in bioreactor is another major issue in pharmaceutical industry and creates operational issues. To address the issue in bioreactor, a CNN-based model was developed for the real-time identification of foam formation (Austerjost et al., 2021). Cell proliferation could be monitored through ML based algorithms. An ML model was trained for monitoring insect cell proliferation and viability percentage upon baculovirus infection in the bioreactor (Altenburg et al., 2023).

ANN based ML algorithm was used to control the micro-aerobic conditions to achieve a satisfactory product yield. Metabolic flux-based control strategy technique (SUPERSYS_MCU) was used to address the issue. To generate a surrogate model in the form of an ANN, the control strategy used simulations of a genome-scale metabolic model. The meta-model provided setpoints to the controller, allowing adjustment of the inlet airflow to control oxygen uptake rate (Zangirolami et al., 2021). Application of ANN models in predicting the system performance of osmotic membrane bioreactors (OMBRs) was investigated and such models developed showed good performance for the prediction of water flux and membrane fouling simulations (Viet and Jang, 2021).

Deep learning techniques in a hybrid semi metric modelling contest, such as deep feed forward neural network with varying depths, the rectified linear unit (ReLU) activation function, dropout regularization of network weights, and stochastic training with the ADAM method were explored (Mestre et al., 2022). Performance of ML algorithms was analyzed to predict n-caproate and n-caprylate productivities in bacteria using 16S rRNA amplicons in a bioreactor. The bioreactor performance was analyzed quantitatively and accurately from the dataset generated from different bioreactors. ML models were trained independently and tested with 16S rRNA amplicon sequencing data to predict n-caproate and n-caprylate productivities. The tests concluded that random forest was the best algorithm producing more consistent results with low error rate and more than 90% accuracy in the prediction of n-caproate and n-caprylate (Liu et al., 2022a). To predict the accuracy of real-time liquid level four ML algorithms, multiple linear regression (MLR), artificial neural network (ANN),

random forest (RF), and support vector machine (SVM) with radial basis kernel were analyzed and found that ANN and RF models performed well (Yu et al., 2022).

4.6 AI in downstream processing

The market demand of biopharmaceutical products is constantly increasing every year and there is an increasing pressure on price reduction for global access to biological drugs. In order to meet the market demand, significant improvement has been carried out in upstream processes, however the productivity in downstream has not increased accordingly (Ötes et al., 2017). The most challenging phase of therapeutic protein production in industries is the downstream processing (DSP) and DSP is accounting for a large portion of the total production costs. The growing demand and developments in upstream processing of therapeutics have burdened the downstream purification processes, due to high cost and insufficient processing capacity (Li et al., 2019). DSP of recombinant therapeutic proteins involves a series of operation such as filtration, followed by capture, purification, and polishing steps mainly done by chromatography (Gaughan, 2016). Chromatography is considered as the workhorse of DSP because it can selectively enrich the target proteins while eliminating impurities and this is achieved by exploiting differences in molecular properties, such as size, charge and hydrophobicity (Bernau et al., 2022). The development of product specific chromatography-based purification techniques is time consuming and expensive because target proteins make up a small portion of the total protein in the initial plant extract. To address this issue, Buyel and Fischer (2014) created a general downstream procedure for the purification of recombinant proteins produced in plants with diverse features. This was done by concentrating on the resin's ability to bind tobacco host cell proteins (HCPs) under various conditions such as pH and conductivity.

TABLE 2 Automation and AI Tools in plant monitoring.

Platform	Automation Technology	Imaging Device	Phenotype/ Parameter	Plant Species	References
UAV remote sensing	Multirotor UAV with CNN architecture	XIMEAMQ022MG-CM Camerawith CMOS sensor and 16 mm lens and Sony NEX-7 Camera	Disease severity at 25m altitude	*O. sativa* (rice)	Bai et al. (2023a)
High throughput UAV remote sensing	DJI Phantom 4 Advanced quadcopter	Drone RGB camera	Accurate plant count, location and size determination to distinguish in paddy field at 7m altitude	*O. sativa*(rice)	Bai et al. (2023b)
RiceNet	Deep Learning Network				
Edge-computing based network monitoring	IoT monitoring with deep learning algorithm-based Edge Image Processing Architecture	Raspberry Pi Camera with 5MP sensor	• Plant growth • Environment and Water quality	-	Wan et al. (2022)
GrowBot	Robotic system with U-Net: CNN	OV5647 CMOS image sensor with Raspberry Pi4	Plant growth based on nutrient deficiency and temperature stress	*Ocimum basilicum* (basil)	Bose and Hautop Lund (2022)
AscTec Navigator 3.4.5	UAV with built-in GPS	AscTec Falcon 8 octocopter (Ascending technologies, Germany) Sony α6000 24.3 MP camera with 20mm f/2.8 lens	• Leaf Area Index at 20m altitude • Leaf/biomass growth • Vegetation indices • Chlorophyll index	*A. hypogaea* L. (peanut)	Sarkar et al. (2021)
WEKA (Waikato Environment for Knowledge Analysis) software v3.8.4	ANN				
WOFOST	UAV imaging integration	-	Leaf area index (LAI), biomass, yield	*T. aestivum* (winter wheat)	Yang et al. (2021a)
Hyperspectral Reflectance	MLP, SVM and RF with remote sensing	UniSpec-DC Spectral Analysis System (PP Systems International Inc., USA)	• Biomass yield • Plant growth and development stages	*G. max* (soybean)	Yoosefzadeh-Najafabadi et al. (2021)
Greenotyper	U-Net: CNNs	RPi3 Model B with RPi Camera module v2.1	• Plant area • Greenness • Overlapping growth patterns	*Trifolium repens* (white clover)	Tausen et al. (2020)
Keras	U-Net based CNN segmentation model	2592 x 1944 x 3 resolution camera (5 MP)	Powdery mildew disease detection	*Cucumis sativus* (cucumber)	Lin et al. (2019)

(Continued)

TABLE 2 Continued

Platform	Automation Technology	Imaging Device	Phenotype/ Parameter	Plant Species	References
CropDeep	RetNet with ResNet50 CNN	IoT cameras, Autonomous Spray robots, Autonomous Picking Robots, Mobicamera and Smartphone camera	• Precision farming • Plant identification, growth and location • Different plant variety monitoring • Fruit and vegetable health status	25 plant varieties including *L. sativa Linn.* (lettuce), *A. graveliens Linn.* (celery), *Cucumis Linn.* (cucumber), *B. oleracea Linn.* (cabbage), *S. oleracea Linn.* (spinach), *L. esculentum Mill.* (tomato), *R. sativus Linn.* (turnip)	Zheng et al. (2019)
Alexnet	CNN-Long-Short Term Memories (LSTM) architecture	Canon EOS 650D	Plant growth pattern of different genotypes	*A. thaliana*	Taghavi Namin et al. (2018)
Persistent Homology based topological methods	DIRT (Digital Imaging of Root Traits) Gaussian kernel density estimator Elliptical Fourier descriptors	-	• Leaf shape, serrations and root architecture • Discrimination between genotypes	*Solanum pennellii* (wild tomato)	Li et al. (2018)
PlantCV	U-Net based CNN	Raspberry Pi Camera	Plant convex hull, width and length	*A. thaliana*	Tovar et al. (2018)
		Nikon COOLPIX L830 Camera	Seed size, shape, count and color	*Chenopodium quinoa* Willd. (Quinoa)	
LeafNet	Caffe framework based Deep Learning CNN	*LeafSnap*, *Flavia* and *Foliage* dataset images using Mobile cameras (iPhones mostly)	Species identification through leaf features like edges and venations	*LeafSnap*, *Flavia* and *Foliage* dataset	Barré et al. (2017)
Deep Plant Phenomics (DPP)	Deep CNN with PlantCV module	Canon PowerShot SD1000 7 MP camera, Model B with Raspberry Pi 5 MP camera module	Leaf size, shape and leaf count	*A. thaliana* *N. tabacum* (tobacco)	Ubbens and Stavness (2017) Minervini et al. (2015)
phenoSeeder	KR 10 scara R600-Z300 robot (KUKA Roboter GmbH, Germany)	Oscar F-810C Camera (Allied-Vision Technologies, GmbH, Germany)	Seed projected area, length, width and color	*B. napus* (rapeseed), *H. vulgare* (barley) and *A. thaliana*	Jahnke et al. (2016)
		Grasshopper GRAS-50S5M-C Camera (Point Grey, Canada) with 35mm lens	Seed volume		
UAV remote sensing SAMPLINGTSPN	UAV and GPML (Gausian Processes for Machine Learning) Toolbox	MikroKopter, Hexa XL with Multispectral Tetracam Camera	Nitrogen level prediction at 30m altitude	*Z. mays* (maize)	Tokekar et al. (2016)
DIRT (Digital Imaging of Root Traits)	-	-	Root angles (top and bottom), stem diameter, width of root system	*Z. mays* (maize)	Das et al. (2015)
GARNICS	Robotic system with ML-based algorithms	Robot head with 4 x Point Grey Grasshopper, 3.45 µm pixels Camera and Schneider KreuznachXenoplan 1.4/ 17-0903 lenses Canon PowerShot SD1000 7 MP camera, Model B with Raspberry Pi 5 MP camera module	• Plant detection and localization • Plant and leaf segmentation • Leaf shade, appearance and difference detection • Leaf counting • Leaf growth tracking • Classification based on mutant and treatment recognition and age regression	*A. thaliana* *N. tabacum* (tobacco)	Minervini et al. (2015)

Recent developments in ML and DL based programs can be utilized to overcome the challenges in downstream processing (Bernau et al., 2022). ML has been applied to chromatography system to monitor real time processing, process optimization, retention time prediction and peak monitoring. In order to predict the chromatographic conditions (i.e., solvents and solvent ratio), three vectorization types such as learned embedding, extended-connectivity fingerprints (ECFP), ECFP encoder+FFNN and three machine learning approaches (FFNN, LSTM and CNN), DNN architectures and a set of hyperparameter values were investigated. The best results were achieved for the prediction of solvents and solvent ratio with ECFP LSTM auto-encoder with FFNN as the supervised machine-learning method with an accuracy of 0.95 for first task and 0.982 for second task respectively (Vaškevičius et al., 2021). Several ML models have been developed so far to address some of the challenges in downstream processing such as XGboost for the prediction of column performance (Jiang et al., 2022b), PeakBot for chromatographic peak prediction (Bueschl et al., 2022), DeepRT for peptide retention time prediction (Ma et al., 2017) and an algorithm to predict the HCPs elution behavior (Buyel et al., 2013).

5 Challenges and current limitations

Plant-based expression systems have several advantages for producing proteins, however, also come with limitations and challenges. Here are few limitations and challenges in plant-based expression systems such as low productivity, post-translational modification, protein stability, biosafety concerns, high costs of downstream processing, regulatory approval, and slow translation to applications (Schillberg et al., 2019; Schillberg and Finnern, 2021; Sethi et al., 2021). Even though the plant expression system is cheaper and more scalable than conventional expression systems, expression yields and appropriate post-translational modifications along the plant secretory pathway remain a challenge for many proteins. For instance, fusion viral glycoproteins often expressed in plants give low yield and may not be properly processed in some cases (Margolin et al., 2020b). In comparison to mammalian systems, plant-based expression systems introduce different glycosylation patterns which could have an effect on the immunogenicity and functionality of proteins. Although difficult, methods for achieving human-like glycosylation patterns in plants are being explored by engineering host systems using CRISPR/Cas9-based technologies. The intellectual property (IP) and regulatory body approval is one of the main hurdles in the adoption of molecular farming compared to commercial microbial and mammalian cell expression systems which have a proven track record, particularly in the field of biopharmaceutical manufacture. As a result, the industry continues to view molecular farming as risky and chooses to depend on its tried-and-true systems in most circumstances (Schillberg and Finnern, 2021). The possible hazards posed by genetically modified (GM) plants or animals, including the effect on biodiversity, ecological interactions, and possibility of unforeseen effects, must be carefully evaluated. There is a risk that the transgenes may unintentionally spread to other organisms through gene flow, such as cross-pollination or horizontal gene transfer. For molecular pharming processes and products to be safe, it is crucial to implement effective containment strategies, risk assessment and mitigation measures. Techniques such as chloroplast expression and transient expression in closed culture systems could circumvent the environmental risk of transgene transmission through pollen (Moon et al., 2019; Feng et al., 2022b).

AI-based tools have been developed and deployed for various microbial expression systems such as *E. coli*, *P. pastoris*, *S. cerevisiae* and mammalian cell expression systems including CHO, HEK293, HeLa and MCF7 (Linder et al., 2020; Van Brempt et al., 2020; Smiatek et al., 2021; Feng et al., 2022a; Li et al., 2022a; Packiam et al., 2022). Plant host system remains an unexplored arena for AI incorporation. Creation and maintenance of AI-based training models is mainly hindered by lack of abundant experimental dataset that include but not limited to genome, transcriptome and metabolome sequences; plant cell culture, plant growth and bioreactor conditions; protein extraction and optimization, purification strategies and relative parameters such as protein localization, structure, stability, catalytic activity and solubility. Such limited training dataset renders the ML approaches overfitting (Feng et al., 2020; van Dijk et al., 2021). Intervention of automation and AI models discussed in Tables 1, 2 to predict the conditions and maintenance for the large-scale production in plants is yet to be established as illustrated in Figure 4. Data integration of multiple parameters discussed in Table 1 is needed for optimal protein expression. Further the generation of training dataset for plant cell culture condition optimization necessitates a large collection of data (van Dijk et al., 2021); and *in vitro* testing of enormous experimental procedures in different test conditions for an individual recombinant protein production in real-time is laborious; time-consuming; requires well-equipped research facility and investment for growth optimization, plant maintenance and downstream processing (Schillberg et al., 2019; Hesami et al., 2020; Sarker, 2021; van Dijk et al., 2021; Packiam et al., 2022). Even with the available omics data of model plants used in recombinant biologics production, expression training datasets are insufficient for AI-based host engineering and host selection, vector and gene designing, protein modelling, solubility and stability prediction as they are not integrated yet (van Dijk et al., 2021). A large number of data for each parameter (more than 10,000 data points if required) is needed to perform as an effective training dataset (Barré et al., 2017; Hesami et al., 2020; LaFleur et al., 2022; Yang et al., 2023). The illustration in Figure 5 highlights the requirement of training datasets available globally that could build a web of AI-based prediction and optimization tools to tackle the challenges and increase the production of highly active next generation biologics. Several algorithms have been under-utilized or unutilized to increase the recombinant protein yield. ML algorithm could predict the signal peptides and increase the ER translocation rates in CHO cells (O'Neill et al., 2023), and yet not used in exploring recombinant biologics production in plants. CNN-based prediction models have been used effectively for increased protein expression in microbial systems (Zrimec et al., 2020) and so far no tool has been adapted for plant-based expression systems.

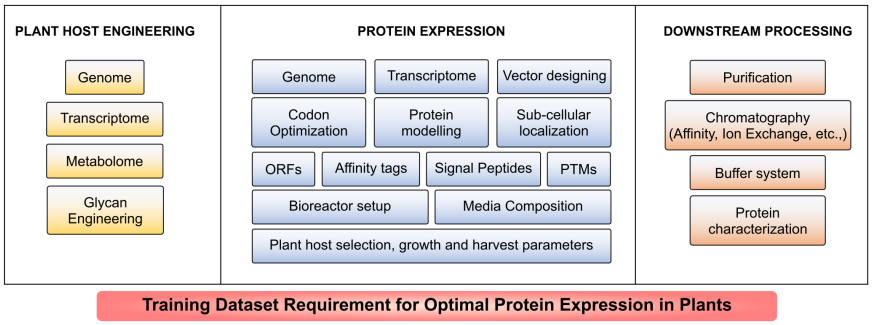

FIGURE 5
Training dataset requirement for optimal protein expression in plants. A large volume of data is required for prediction of optimum conditions at each stage including host engineering, expression and downstream processing for a specific protein to be expressed large-scale in plants.

6 Conclusion and future directions

Plant molecular pharming offers efficient alternate host systems for expression of recombinant biologics. Moreover, the system is robust and cost-effective compared to other hosts. In this review, the concepts of AI in systems engineering for improved production of recombinant biologics were discussed. Several prediction and optimization parameters are known to increase the yield in different expression hosts and integration of machine learning algorithms is new to the plant molecular pharming field. Such plant-based expression parameters include host engineering, growth and maintenance, protein model designing, glycosylation, sialylation, epitope prediction, antibody identification& optimization, regulatory element prediction & optimization and protein stability and activity. Neural network-based ML models when integrated with systems engineering approaches could be advantageous during the manufacture of humanized forms of biologics at various stages of production including seed selection, germination, plant growth parameter optimization, monitoring, recombinant protein modelling, expression, extraction, purification and downstream processing. GEMs and other omics data availability favor the process of designing and optimization of protein production yet more omics (genomics, proteomics, transcriptomics and metabolomics) based studies are needed for complete utilization of ML tools. Transcriptome and metabolome profiles of specific plant hosts in the form of large training data sets need to be fed into neural networks, which then can be used to test the desired function (such as gene knock-out or knock-in). Similarly, parameters of protein production solely based on plant system are to be created as codes using language models and integrated as hierarchical architectures using neural networks. Datasets trained with the discussed parameters using ML models for protein expression in plants could aid in an effective modelling of recombinant biologics and prediction of accurate conditions for protein expression in different plant hosts including but not limited to *N. benthamiana*, *N. tabacum*, *L. sativa* and *O. sativa*. Such ML-based techniques will reduce the time frame and cost of reagents in all the levels of plant-based biologics production rendering functional and active products.

Author contributions

RS proposed the idea of application of AI in plant molecular pharming; SP designed the review manuscript. TG drafted systems biology; SP and TV drafted AI integration concepts and improvised systems biology concepts. RS and BS revised and corrected the manuscript. AS gave expert comments on the technical aspects. All authors contributed to the article and approved the submitted version.

Funding

The authors would like to acknowledge the funding support of University Grants Commission-UK-India Research Initiative (UGC-UKIERI), No.F 184-9/2018(IC), and RashtriyaUchchatar Shiksha Abhiyan (RUSA) 2.0, No. BU/RUSA2.0/BCTRC/2020/BCTRC-CD06, Bharathiar University, India.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Agarwal, P., Gautam, T., Singh, A. K., and Burma, P. K. (2019). Evaluating the effect of codon optimization on expression of bar gene in transgenic tobacco plants. *J. Plant Biochem. Biotechnol.* 28, 189–202. doi: 10.1007/s13562-019-00506-2

Alam, A., Jiang, L., Kittleson, G. A., Steadman, K. D., Nandi, S., Fuqua, J. L., et al. (2018). Technoeconomic modeling of plant-based griffithsin manufacturing. *Front. Bioeng. Biotechnol.* 6. doi: 10.3389/fbioe.2018.00102

Al-Hawash, A. B., Zhang, X., and Ma, F. (2017). Strategies of codon optimization for high-level heterologous protein expression in microbial expression systems. *Gene Rep.* 9, 46–53. doi: 10.1016/j.genrep.2017.08.006

Ali, S., and Kim, W. C. (2019). A fruitful decade using synthetic promoters in the improvement of transgenic plants. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01433

Allied Market Research. (2023). *Plant-Based Biologics Market by Product Type (Leaf-based, Seed-Based, Fruit-based, Others), by Source (Carrot, Tobacco, Rice, Duckweed, Others), by Target Disease (Gaucher Disease, Fabry Disease, Others): Global Opportunity Analysis and Industry Forecast*. Available at: https://www.alliedmarketresearch.com/plant-based-biologics-market-A74549#:~:text.

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Alshehri, S., Alqarni, M., Namazi, N. I., Naguib, I. A., Venkatesan, K., Mosaad, Y. O., et al. (2022). Design of predictive model to optimize the solubility of Oxaprozin as nonsteroidal anti-inflammatory drug. *Sci. Rep.* 12, 1–10. doi: 10.1038/s41598-022-17350-5

Altenburg, J. J., Klaverdijk, M., Cabosart, D., Desmecht, L., Brunekreeft-Terlouw, S. S., Both, J., et al. (2023). Real-time online monitoring of insect cell proliferation and baculovirus infection using digital differential holographic microscopy and machine learning. *Biotechnol. Prog.* 39, e3318. doi: 10.1002/btpr.3318

Amack, S. C., and Antunes, M. S. (2020). CaMV35S promoter – A plant biology and biotechnology workhorse in the era of synthetic biology. *Curr. Plant Biol.* 24, 100179. doi: 10.1016/j.cpb.2020.100179

Arcalis, E., Ibl, V., Hilscher, J., Rademacher, T., Avesani, L., Morandini, F., et al. (2019). Russell-like bodies in plant seeds share common features with prolamin bodies and occur upon recombinant protein production. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00777

Argentinian AntiCovid Consortium. (2020). Structural and functional comparison of SARS-CoV-2-spike receptor binding domain produced in Pichia pastoris and mammalian cells. *Sci. Rep.* 10, 21779. doi: 10.1038/s41598-020-78711-6

Aslan, M. F., Durdu, A., Sabanci, K., and Ropelewska, E. (2022). A comprehensive survey of the recent studies with UAV for precision agriculture in open fields and greenhouses. *Appl. Sci.* 12, 1047. doi: 10.3390/app12031047

Austerjost, J., Söldner, R., Edlund, C., Trygg, J., Pollard, D., and Sjögren, R. (2021). A machine vision approach for bioreactor foam sensing. *SLAS Technol.* 26 (4), 408–414. doi: 10.1177/24726303211008861

Bai, X., Fang, H., He, Y., Zhang, J., Tao, M., Wu, Q., et al. (2023a). Dynamic UAV phenotyping for rice disease resistance analysis based on multisource data. *Plant Phenomics* 5, 1–13. doi: 10.34133/plantphenomics.0019

Bai, X., Liu, P., Cao, Z., Lu, H., Xiong, H., Yang, A., et al. (2023b). Rice plant counting, locating, and sizing method based on high-throughput UAV RGB images. *Plant Phenomics* 5, 1–16. doi: 10.34133/plantphenomics.0020

Banerjee, B. P., Spangenberg, G., and Kant, S. (2022). CBM: an ioT enabled liDAR sensor for in-field crop height and biomass measurements. *Biosensors* 12, 16. doi: 10.3390/bios12010016

Barra, C., Ackaert, C., Reynisson, B., Schockaert, J., Jessen, L. E., Watson, M., et al. (2020). Immunopeptidomic data integration to artificial neural networks enhances protein-drug immunogenicity prediction. *Front. Immunol.* 11. doi: 10.3389/fimmu.2020.01304

Barré, P., Stöver, B. C., Müller, K. F., and Steinhage, V. (2017). LeafNet: A computer vision system for automatic plant species identification. *Ecol. Inform.* 40, 50–56. doi: 10.1016/j.ecoinf.2017.05.005

Belcher, M. S., Vuu, K. M., Zhou, A., Mansoori, N., Agosto Ramos, A., Thompson, M. G., et al. (2020). Design of orthogonal regulatory systems for modulating gene expression in plants. *Nat. Chem. Biol.* 16, 857–865. doi: 10.1038/s41589-020-0547-4

Bernau, C. R., Knödler, M., Emonts, J., Jäpel, R. C., and Buyel, J. F. (2022). The use of predictive models to develop chromatography-based purification processes. *Front. Bioeng. Biotechnol.* 10. doi: 10.3389/fbioe.2022.1009102

Bidarigh fard, A., Dehghan Nayeri, F., and Habibi Anbuhi, M. (2019). Transient expression of etanercept therapeutic protein in tobacco (Nicotiana tabacum L.). *Int. J. Biol. Macromol.* 130, 483–490. doi: 10.1016/j.ijbiomac.2019.02.153

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. (2021). Low-N protein engineering with data-efficient deep learning. *Nat. Methods* 18, 389–396. doi: 10.1038/s41592-021-01100-y

Bogard, N., Linder, J., Rosenberg, A. B., and Seelig, G. (2019). A deep neural network for predicting and engineering alternative polyadenylation. *Cell* 178, 91–106.e23. doi: 10.1016/j.cell.2019.04.046

Bohlender, L. L., Parsons, J., Hoernstein, S. N. W., Rempfer, C., Ruiz-Molina, N., Lorenz, T., et al. (2020). Stable protein sialylation in physcomitrella. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.610032

Bolaños-Martínez, O. C., Govea-Alonso, D. O., Cervantes-Torres, J., Hernández, M., Fragoso, G., Sciutto-Conde, E., et al. (2020). Expression of immunogenic poliovirus Sabin type 1 VP proteins in transgenic tobacco. *J. Biotechnol.* 322, 10–20. doi: 10.1016/j.jbiotec.2020.07.007

Bose, R., and Hautop Lund, H. (2022). Convolutional neural network for studying plant nutrient deficiencies. *Proc. Int. Conf. Artif. Life Robot.* 27, 25–29. doi: 10.5954/icarob.2022.is2-2

Bueschl, C., Doppler, M., Varga, E., Seidl, B., Flasch, M., Warth, B., et al. (2022). PeakBot: Machine-learning-based chromatographic peak picking. *Bioinformatics* 38, 3422–3428. doi: 10.1093/bioinformatics/btac344

Buyel, J. F. (2019). Plant molecular farming – Integration and exploitation of side streams to achieve sustainable biomanufacturing. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01893

Buyel, J. F., and Fischer, R. (2014). Generic chromatography-based purification strategies accelerate the development of downstream processes for biopharmaceutical proteins produced in plants. *Biotechnol. J.* 9, 566–577. doi: 10.1002/biot.201300548

Buyel, J. F., Woo, J. A., Cramer, S. M., and Fischer, R. (2013). The use of quantitative structure-activity relationship models to develop optimized processes for the removal of tobacco host cell proteins during biopharmaceutical production. *J. Chromatogr. A* 1322, 18–28. doi: 10.1016/j.chroma.2013.10.076

Cardon, F., Pallisse, R., Bardor, M., Caron, A., Vanier, J., Pierre, J., et al. (2019). Brassica rapa hairy root based expression system leads to the production of highly homogenous and reproducible profiles of recombinant human alpha-L-iduronidase. *Plant Biotechnol. J.* 17 (2), 505–516. doi: 10.1111/pbi.12994

Carreño-Campos, C., Arevalo-Villalobos, J. I., Villarreal, M. L., Ortiz-Caltempa, A., and Rosales-Mendoza, S. (2022). Establishment of the carrot-made LTB-syn antigen cell line in shake flask and airlift bioreactor cultures. *Planta Med.* 88, 1060–1068. doi: 10.1055/a-1677-4135

Chen, Q., and Davis, K. R. (2016). The potential of plants as a system for the development and production of human biologics [version 1; referees: 3 approved]. *F1000Research* 5, 1–8. doi: 10.12688/F1000RESEARCH.8010.1

Chen, Y., Yang, O., Sampat, C., Bhalode, P., Ramachandran, R., and Ierapetritou, M. (2020). Digital twins in pharmaceutical and biopharmaceutical manufacturing. *Processes* 8, 1–33. doi: 10.3390/pr8091088

Chia, S., Tay, S. J., Song, Z., Yang, Y., Walsh, I., and Pang, K. T. (2023). Enhancing pharmacokinetic and pharmacodynamic properties of recombinant therapeutic proteins by manipulation of sialic acid content. *Biomed. Pharmacother.* 163, 114757. doi: 10.1016/j.biopha.2023.114757

Constant, D. A., Gutierrez, J. M., Sastry, A. V., Viazzo, R., Smith, N. R., Hossain, J., et al. (2023). Deep learning-based codon optimization with large-scale synonymous variant datasets enables generalized tunable protein expression. *bioRxiv* 2023, 2.11.528149. doi: 10.1101/2023.02.11.528149

Costello, Z., and Martin, H. G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ Syst. Biol. Appl.* 4, 1–14. doi: 10.1038/s41540-018-0054-3

Culley, C., Vijayakumar, S., Zampieri, G., and Angione, C. (2020). A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proc. Natl. Acad. Sci.* 117, 18869–18879. doi: 10.1073/pnas.2002959117

Das, A., Schneider, H., Burridge, J., Ascanio, A. K. M., Wojciechowski, T., Topp, C. N., et al. (2015). Digital imaging of root traits (DIRT): A high-throughput computing and collaboration platform for field-based root phenomics. *Plant Methods* 11, 1–12. doi: 10.1186/s13007-015-0093-3

Dehdashti, S. M., Acharjee, S., Nomani, A., and Deka, M. (2020). Production of pharmaceutical active recombinant globular adiponectin as a secretory protein in Withania Somnifera hairy root culture. *J. Biotechnol.* 323, 302–312. doi: 10.1016/j.jbiotec.2020.07.012

Dhivya, S., Priya, S. H., and Sathishkumar, R. (2021). "Opportunities in Agriculture, Biomedicine, and Healthcare," in *Artificial Intelligence Theory, Models, and Applications*. Eds. P. Kaliraj and T. Devi (Boca Raton, FL, Oxon, OX: CRC Press), 121.

Ding, Z., Guan, F., Xu, G., Wang, Y., Yan, Y., Zhang, W., et al. (2022). MPEPE, a predictive approach to improve protein expression in E. coli based on deep learning. *Comput. Struct. Biotechnol. J.* 20, 1142–1153. doi: 10.1016/j.csbj.2022.02.030

dos Reis, M., Wernisch, L., and Savva, R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res.* 31, 6976–6985. doi: 10.1093/nar/gkg897

Doyle, F., Leonardi, A., Endres, L., Tenenbaum, S. A., Dedon, P. C., and Begley, T. J. (2016). Gene- and genome-based analysis of significant codon patterns in yeast, rat and mice genomes with the CUT Codon UTilization tool. *Methods* 107, 98–109. doi: 10.1016/j.ymeth.2016.05.010

Dubey, K. K., Luke, G. A., Knox, C., Kumar, P., Pletschke, B. I., Singh, P. K., et al. (2018). Vaccine and antibody production in plants: Developments and computational tools. *Brief. Funct. Genomics* 17, 295–307. doi: 10.1093/bfgp/ely020

Feng, J., Jiang, M., Shih, J., and Chai, Q. (2022a). Antibody apparent solubility prediction from sequence by transfer learning. *iScience* 25, 105173. doi: 10.1016/j.isci.2022.105173

Feng, Z., Li, X., Fan, B., Zhu, C., and Chen, Z. (2022b). Maximizing the production of recombinant proteins in plants: from transcription to protein stability. *Int. J. Mol. Sci.* 23, 13516. doi: 10.3390/ijms232113516

Feng, L., Zhang, Z., Ma, Y., Du, Q., Williams, P., Drewry, J., et al. (2020). Alfalfa yield prediction using UAV-based hyperspectral imagery and ensemble learning. *Remote Sens.* 12, 2028. doi: 10.3390/rs12122028

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., et al. (2021). The coding capacity of SARS-CoV-2. *Nature* 589, 125–130. doi: 10.1038/s41586-020-2739-1

Fox, D. M., Branson, K. M., and Walker, R. C. (2021). mRNA codon optimization with quantum computers. *PloS One* 16, 1–16. doi: 10.1371/journal.pone.0259101

Fu, H., Liang, Y., Zhong, X., Pan, Z., Huang, L., Zhang, H., et al. (2020). Codon optimization with deep learning to enhance protein expression. *Sci. Rep.* 10, 17617. doi: 10.1038/s41598-020-74091-z

Fu, Y., Yang, G., Song, X., Li, Z., Xu, X., Feng, H., et al. (2021). Improved estimation of winter wheat aboveground biomass using multiscale textures extracted from UAV-based digital images and hyperspectral feature analysis. *Remote Sens.* 13, 1–22. doi: 10.3390/rs13040581

Fulton, A., Lai, H., Chen, Q., and Zhang, C. (2015). Purification of monoclonal antibody against Ebola GP1 protein expressed in Nicotiana benthamiana. *J. Chromatogr. A* 1389, 128–132. doi: 10.1016/j.chroma.2015.02.013

Gaughan, C. L. (2016). The present state of the art in expression, production and characterization of monoclonal antibodies. *Mol. Divers.* 20, 255–270. doi: 10.1007/s11030-015-9625-z

Gelvin, S. B. (2003). Agrobacterium-mediated plant transformation: the biology behind the "gene-jockeying" tool. *Microbiol. Mol. Biol. Rev.* 67, 16–37. doi: 10.1128/MMBR.67.1.16-37.2003

Ghag, S. B., Adki, V. S., Ganapathi, T. R., and Bapat, V. A. (2021). Plant platforms for efficient heterologous protein production. *Biotechnol. Bioprocess Eng.* 26, 546–567. doi: 10.1007/s12257-020-0374-1

Gomord, V., and Faye, L. (2004). Posttranslational modification of therapeutic proteins in plants. *Curr. Opin. Plant Biol.* 7, 171–181. doi: 10.1016/j.pbi.2004.01.015

Goulet, D. R., Yan, Y., Agrawal, P., Waight, A. B., Mak, A. N. S., and Zhu, Y. (2023). Codon optimization using a recurrent neural network. *J. Comput. Biol.* 30, 70–81. doi: 10.1089/cmb.2021.0458

Grandits, M., Grünwald-Gruber, C., Gastine, S., Standing, J. F., Reljic, R., Teh, A. Y.-H., et al. (2023). Improving the efficacy of plant-made anti-HIV monoclonal antibodies for clinical use. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1126470

Gupta, A., and Zou, J. (2018). Feedback GAN (FBGAN) for DNA: a novel feedback-loop architecture for optimizing protein functions. *arXiv Prepr. arXiv* 1804, 1694. doi: 10.48550/arXiv.1804.01694

Gutierrez-valdes, N., Häkkinen, S. T., Lemasson, C., Guillet, M., Ritala, A., and Cardon, F. (2020). Hairy root cultures — A versatile tool with multiple applications. *Front. Plant. Sci.* 11, 1–11. doi: 10.3389/fpls.2020.00033

Hager, K. J., Pérez Marc, G., Gobeil, P., Diaz, R. S., Heizer, G., Llapur, C., et al. (2022). Efficacy and safety of a recombinant plant-based adjuvanted covid-19 vaccine. *N. Engl. J. Med.* 386, 2084–2096. doi: 10.1056/nejmoa2201300

Han, X., Wang, X., and Zhou, K. (2019). Develop machine learning-based regression predictive models for engineering protein solubility. *Bioinformatics* 35, 4640–4646. doi: 10.1093/bioinformatics/btz294

Hanittinan, O., Oo, Y., Chaotham, C., Rattanapisit, K., Shanmugaraj, B., and Phoolcharoen, W. (2020). Expression optimization, purification and in *vitro* characterization of human epidermal growth factor produced in Nicotiana benthamiana. *Biotechnol. Rep.* 28, e00524. doi: 10.1016/j.btre.2020.e00524

Hassan, M. M., Zhang, Y., Yuan, G., De, K., Chen, J. G., Muchero, W., et al. (2021). Construct design for CRISPR/Cas-based genome editing in plants. *Trends Plant Sci.* 26, 1133–1152. doi: 10.1016/j.tplants.2021.06.015

He, W., Baysal, C., Lobato Gómez, M., Huang, X., Alvarez, D., Zhu, C., et al. (2021). Contributions of the international plant science community to the fight against infectious diseases in humans—part 2: Affordable drugs in edible plants for endemic and re-emerging diseases. *Plant Biotechnol. J.* 19, 1921–1936. doi: 10.1111/pbi.13658

Heenatigala, P. P. M., Sun, Z., Yang, J., Zhao, X., and Hou, H. (2020). Expression of lamB vaccine antigen in wolffia globosa (Duck weed) against fish vibriosis. *Front. Immunol.* 11. doi: 10.3389/fimmu.2020.01857

Hesami, M., Naderi, R., Tohidfar, M., and Yoosefzadeh-Najafabadi, M. (2020). Development of support vector machine-based model and comparative analysis with artificial neural network for modeling the plant tissue culture procedures: Effect of plant growth regulators on somatic embryogenesis of chrysanthemum, as a case study. *Plant Methods* 16, 1–15. doi: 10.1186/s13007-020-00655-9

Holásková, E., Galuszka, P., Mičúchová, A., Šebela, M., Öz, M. T., and Frébort, I. (2018). Molecular farming in barley: development of a novel production platform to produce human antimicrobial peptide LL-37. *Biotechnol. J.* 13, 1700628. doi: 10.1002/biot.201700628

Hosseini, M. S., Arab, M. M., Soltani, M., and Eftekhari, M. (2022). Predictive modeling of Persian walnut (Juglans regia L.) in *vitro* proliferation media using machine learning approaches : a comparative study of ANN, KNN and GEP models. *Plant Methods* 18 (1), 1–24. doi: 10.1186/s13007-022-00871-5

Imamura, T., Isozumi, N., Higashimura, Y., Ohki, S., and Mori, M. (2021). Production of ORF8 protein from SARS – CoV – 2 using an inducible virus – mediated expression system in suspension – cultured tobacco BY – 2 cells. *Plant Cell Rep.* 40, 433–436. doi: 10.1007/s00299-020-02654-5

Islam, M. R., Choi, S., Muthamilselvan, T., Shin, K., and Hwang, I. (2020). *In vivo* removal of N-terminal fusion domains from recombinant target proteins produced in nicotiana benthamiana. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00440

Islam, M. R., Kwak, J. W., Lee, J.-s., Hong, S. W., Khan, M. R. I., Lee, Y., et al. (2019). Cost-effective production of tag-less recombinant protein in Nicotiana benthamiana. *Plant Biotechnol. J.* 17, 1094–1105. doi: 10.1111/pbi.13040

Iyappan, G., Shanmugaraj, B. M., Inchakalody, V., Ma, J. K.-C., and Ramalingam, S. (2018). Potential of plant biologics to tackle the epidemic like situations - case studies involving viral and bacterial candidates. *Int. J. Infect. Dis.* 73, 363. doi: 10.1016/j.ijid.2018.04.4236

Izadi, S., Kunnummel, V., Steinkellner, H., Werner, S., and Castilho, A. (2023). Assessment of transient expression strategies to sialylate recombinant proteins in N. benthamiana. *J. Biotechnol.* 365, 48–53. doi: 10.1016/j.jbiotec.2023.02.004

Jahnke, S., Roussel, J., Hombach, T., Kochs, J., Fischbach, A., Huber, G., et al. (2016). phenoSeeder - A robot system for automated handling and phenotyping of individual seeds. *Plant Physiol.* 172, 1358–1370. doi: 10.1104/pp.16.01122

Jain, R., Jain, A., Mauro, E., LeShane, K., and Densmore, D. (2023). ICOR: improving codon optimization with recurrent neural networks. *BMC Bioinf.* 24, 132. doi: 10.1186/s12859-023-05246-8

Jansing, J., Sack, M., Augustine, S. M., Fischer, R., and Bortesi, L. (2019). CRISPR/Cas9-mediated knockout of six glycosyltransferase genes in Nicotiana benthamiana for the production of recombinant proteins lacking β-1,2-xylose and core α-1,3-fucose. *Plant Biotechnol. J.* 17, 350–361. doi: 10.1111/pbi.12981

Jiang, J., Johansen, K., Stanschewski, C. S., Wellman, G., Mousa, M. A. A., Fiene, G. M., et al. (2022a). Phenotyping a diversity panel of quinoa using UAV-retrieved leaf area index, SPAD-based chlorophyll and a random forest approach. *Precis. Agric.* 23, 961–983. doi: 10.1007/s11119-021-09870-3

Jiang, Q., Seth, S., Scharl, T., Schroeder, T., Jungbauer, A., and Dimartino, S. (2022b). Prediction of the performance of pre-packed purification columns through machine learning. *J. Sep. Sci.* 45, 1445–1457. doi: 10.1002/jssc.202100864

Jiang, Y., Wang, D., Yao, Y., Eubel, H., Künzler, P., Møller, I. M., et al. (2021). MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. *Comput. Struct. Biotechnol. J.* 19, 4825–4839. doi: 10.1016/j.csbj.2021.08.027

Jin, L., Wang, Y., Liu, X., Peng, R., Lin, S., Sun, D., et al. (2022). Codon optimization of chicken β Gallinacin-3 gene results in constitutive expression and enhanced antimicrobial activity in transgenic Medicago sativa L. *Gene* 835, 146656. doi: 10.1016/j.gene.2022.146656

Jolles, J. W. (2021). Broad-scale applications of the Raspberry Pi: A review and guide for biologists. *Methods Ecol. Evol.* 12, 1562–1579. doi: 10.1111/2041-210X.13652

Jugler, C., Grill, F. J., Eidenberger, L., Karr, T. L., Grys, T. E., Steinkellner, H., et al. (2022). Humanization and expression of IgG and IgM antibodies in plants as potential diagnostic reagents for Valley Fever. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.925008

Jung, J. W., Shin, J. H., Lee, W. K., Begum, H., Min, C. H., Jang, M. H., et al. (2021). Inactivation of the β (1, 2)-xylosyltransferase and the α (1, 3)-fucosyltransferase gene in rice (Oryza sativa) by multiplex CRISPR/Cas9 strategy. *Plant Cell Rep.* 40, 1025–1035. doi: 10.1007/s00299-021-02667-8

Kalemati, M., Darvishi, S., and Koohi, S. (2023). CapsNet-MHC predicts peptide-MHC class I binding based on capsule neural networks. *Commun. Biol.* 6, 492. doi: 10.1038/s42003-023-04867-2

Khoshmaram, A., Zabihi, S., Pelalak, R., Pishnamazi, M., Marjani, A., and Shirazian, S. (2021). Supercritical process for preparation of nanomedicine: oxaprozin case study. *Chem. Eng. Technol.* 44, 208–212. doi: 10.1002/ceat.202000411

Khurana, S., Rawi, R., Kunji, K., Chuang, G. Y., Bensmail, H., and Mall, R. (2018). DeepSol: A deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* 34, 2605–2613. doi: 10.1093/bioinformatics/bty166

Kim, G. B., Kim, W. J., Kim, H. U., and Lee, S. Y. (2020). Machine learning applications in systems metabolic engineering. *Curr. Opin. Biotechnol.* 64, 1–9. doi: 10.1016/j.copbio.2019.08.010

Koşaloğlu-Yalçın, Z., Lee, J., Greenbaum, J., Schoenberger, S. P., Miller, A., Kim, Y. J., et al. (2022). Combined assessment of MHC binding and antigen abundance improves T cell epitope predictions. *iScience* 25, 103850. doi: 10.1016/j.isci.2022.103850

Kraus, O. Z., Grys, B. T., Ba, J., Chong, Y., Frey, B. J., Boone, C., et al. (2017). Automated analysis of high-content microscopy data with deep learning. *Mol. Syst. Biol.* 13, 1–15. doi: 10.15252/msb.20177551

Kumar, A. U., and Ling, A. P. K. (2021). Gene introduction approaches in chloroplast transformation and its applications. *J. Genet. Eng. Biotechnol.* 19 (1), 1–10. doi: 10.1186/s43141-021-00255-7

Kwon, K. C., Chan, H. T., León, I. R., Williams-Carrier, R., Barkan, A., and Daniell, H. (2016). Codon optimization to enhance expression yields insights into chloroplast translation. *Plant Physiol.* 172, 62–77. doi: 10.1104/pp.16.00981

LaFleur, T. L., Hossain, A., and Salis, H. M. (2022). Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat. Commun.* 13, 5159. doi: 10.1038/s41467-022-32829-5

Lai, P. K. (2022). DeepSCM: An efficient convolutional neural network surrogate model for the screening of therapeutic antibody viscosity. *Comput. Struct. Biotechnol. J.* 20, 2143–2152. doi: 10.1016/j.csbj.2022.04.035

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, F., Chen, Y., Qi, Q., Wang, Y., Yuan, L., Huang, M., et al. (2022a). Improving recombinant protein production by yeast through genome-scale modeling using proteome constraints. *Nat. Commun.* 13, 1–13. doi: 10.1038/s41467-022-30689-7

Li, M., Frank, M. H., Coneva, V., Mio, W., Chitwood, D. H., and Topp, C. N. (2018). The persistent homology mathematical framework provides enhanced genotype-to-phenotype associations for plant morphology. *Plant Physiol.* 177, 1382–1395. doi: 10.1104/pp.18.00104

Li, X., Li, X., Fan, B., Zhu, C., and Chen, Z. (2022b). Specialized endoplasmic reticulum-derived vesicles in plants: Functional diversity, evolution, and biotechnological exploitation. *J. Integr. Plant Biol.* 64, 821–835. doi: 10.1111/jipb.13233

Li, Y., Stern, D., Lin, L., Mills, J., Ou, S., Morrow, M., et al. (2019). Emerging biomaterials for downstream manufacturing of therapeutic proteins. *Acta Biomaterialia* 95, 73–90. doi: 10.1016/j.actbio.2019.03.015

Lim, C. Y., Kim, D. S., Kang, Y., Lee, Y. R., Kim, K., Kim, D. S., et al. (2022). Immune responses to plant-derived recombinant colorectal cancer glycoprotein epCAM-fcK fusion protein in mice. *Biomol. Ther.* 30, 546–552. doi: 10.4062/biomolther.2022.103

Limkul, J., Iizuka, S., Sato, Y., Misaki, R., Ohashi, T., Ohashi, T., et al. (2016). The production of human glucocerebrosidase in glyco-engineered Nicotiana benthamiana plants. *Plant Biotechnol. J.* 14, 1682–1694. doi: 10.1111/pbi.12529

Lin, K., Gong, L., Huang, Y., Liu, C., and Pan, J. (2019). Deep learning-based segmentation and quantification of cucumber powdery mildew using convolutional neural network. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00155

Linder, J., Bogard, N., Rosenberg, A. B., and Seelig, G. (2020). A generative neural network for maximizing fitness and diversity of synthetic DNA and protein sequences. *Cell Syst.* 11, 49–62.e16. doi: 10.1016/j.cels.2020.05.007

Liu, X. (2017). Deep recurrent neural network for protein function prediction from sequence. *arXiv Prepr*. doi: 10.48550/arXiv.1701.08318

Liu, Z., Jin, J., Cui, Y., Xiong, Z., Nasiri, A., Zhao, Y., et al. (2022b). DeepSeqPanII: an interpretable recurrent neural network model with attention mechanism for peptide-HLA class II binding prediction. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 19, 2188–2196. doi: 10.1109/TCBB.2021.3074927

Liu, B., Sträuber, H., Saraiva, J., Harms, H., Silva, S. G., and Kasmanas, J. C. (2022a). Machine learning-assisted identification of bioindicators predicts medium-chain carboxylate production performance of an anaerobic mixed culture. *Microbiome* 10, 1–21. doi: 10.1186/s40168-021-01219-2

Lobato Gómez, M., Huang, X., Alvarez, D., He, W., Baysal, C., Zhu, C., et al. (2021). Contributions of the international plant science community to the fight against human infectious diseases – part 1: epidemic and pandemic diseases. *Plant Biotechnol. J.* 19, 1901–1920. doi: 10.1111/pbi.13657

Lu, C., Liu, C., Sun, X., Wan, P., Ni, J., Wang, L., et al. (2021). Bioinformatics analysis, codon optimization and expression of ovine pregnancy associated Glycoprotein-7 in HEK293 cells. *Theriogenology* 172, 27–35. doi: 10.1016/j.theriogenology.2021.05.027

Luo, Y., Jiang, G., Yu, T., Liu, Y., Vo, L., Ding, H., et al. (2021). ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat. Commun.* 12, 5743. doi: 10.1038/s41467-021-25976-8

Ma, J. K. C., Drake, P. M. W., and Christou, P. (2003). The production of recombinant pharmaceutical proteins in plants. *Nat. Rev. Genet.* 4, 794–805. doi: 10.1038/nrg1177

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., et al. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* 15, 290–298. doi: 10.1038/nmeth.4627

Ma, C., Zhu, Z., Ye, J., Yang, J., Pei, J., Xu, S., et al. (2017). DeepRT: deep learning for peptide retention time prediction in proteomics. *arXiv Prepr*. doi: 10.48550/arXiv.1705.05368

Macharoen, K., Du, M., Jung, S., McDonald, K. A., and Nandi, S. (2021). Production of recombinant butyrylcholinesterase from transgenic rice cell suspension cultures in a pilot-scale bioreactor. *Biotechnol. Bioeng.* 118, 1431–1443. doi: 10.1002/bit.27638

Maimaitijiang, M., Sagan, V., Sidike, P., Daloye, A. M., Erkbol, H., and Fritschi, F. B. (2020). Crop monitoring using satellite/UAV data fusion and machine learning. *Remote Sens.* 12, 1357. doi: 10.3390/RS12091357

Makowski, E. K., Chen, H., Lambert, M., Bennett, E. M., Eschmann, N. S., Zhang, Y., et al. (2022). Reduction of therapeutic antibody self-association using yeast-display selections and machine learning. *MAbs* 14, 2146629. doi: 10.1080/19420862.2022.2146629

Margolin, E., Oh, Y. J., Verbeek, M., Naude, J., Ponndorf, D., Meshcheriakova, Y. A., et al. (2020b). Co-expression of human calreticulin significantly improves the production of HIV gp140 and other viral glycoproteins in plants. *Plant Biotechnol. J.* 18, 2109–2117. doi: 10.1111/pbi.13369

Margolin, E. A., Strasser, R., Chapman, R., Williamson, A.-L., Rybicki, E. P., and Meyers, A. E. (2020a). Engineering the plant secretory pathway for the production of next-generation pharmaceuticals. *Trends Biotechnol.* 38, 1034–1044. doi: 10.1016/j.tibtech.2020.03.004

Markova, E. A., Shaw, R. E., and Reynolds, C. R. (2022). Prediction of strain engineerings that amplify recombinant protein secretion through the machine learning approach MaLPHAS. *Eng. Biol.* 6, 82–90. doi: 10.1049/enb2.12025

Marques, L. É. C., Silva, B. B., Dutra, R. F., Florean, E. O. P. T., Menassa, R., and Guedes, M. I. F. (2020). Transient expression of dengue virus NS1 antigen in nicotiana benthamiana for use as a diagnostic antigen. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01674

Martiny, H. M., Armenteros, J. J. A., Johansen, A. R., Salomon, J., and Nielsen, H. (2021). Deep protein representations enable recombinant protein expression prediction. *Comput. Biol. Chem.* 95, 107596. doi: 10.1016/j.compbiolchem.2021.107596

Masson, H. O., Kuo, C.-C., Malm, M., Lundqvist, M., Sievertsson, Å., Berling, A., et al. (2022). Deciphering the determinants of recombinant protein yield across the human secretome. *bioRxiv* 2022, 12.12.520152. doi: 10.1101/2022.12.12.520152

McNulty, M. J., Gleba, Y., Tusé, D., Hahn-Löbmann, S., Giritch, A., Nandi, S., et al. (2020). Techno-economic analysis of a plant-based platform for manufacturing antimicrobial proteins for food safety. *Biotechnol. Prog.* 36, e2896. doi: 10.1002/btpr.2896

Mestre, M., Ramos, J., Costa, R. S., Striedner, G., and Oliveira, R. (2022). *A general deep hybrid model for bioreactor systems : Combining first principles with deep neural networks*. Amsterdam: Elsevier. Vol. 165. doi: 10.1016/j.compchemeng.2022.107952

Mettu, R. R., Charles, T., and Landry, S. J. (2016). CD4+ T-cell epitope prediction using antigen processing constraints. *J. Immunol. Methods* 432, 72–81. doi: 10.1016/j.jim.2016.02.013

Minervini, M., Fischbach, A., Scharr, H., and Tsaftaris, S. A. (2015). Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognit. Lett.* 81, 80–89. doi: 10.1016/j.patrec.2015.10.013

Minervini, M., Giuffrida, M. V., Perata, P., and Tsaftaris, S. A. (2017). Phenotiki: an open software and hardware platform for affordable and easy image-based phenotyping of rosette-shaped plants. *Plant J.* 90, 204–216. doi: 10.1111/tpj.13472

Mirzaee, M., Osmani, Z., Frébortová, J., and Frébort, I. (2022). Recent advances in molecular farming using monocot plants. *Biotechnol. Adv.* 58, 107913. doi: 10.1016/j.bioteChadv.2022.107913

Miura, K., Yoshida, H., Nosaki, S., Kaneko, M. K., and Kato, Y. (2020). RAP tag and PMab-2 antibody: A tagging system for detecting and purifying proteins in plant cells. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.510444

Monteiro, A., Santos, S., and Gonçalves, P. (2021). Precision agriculture for crop and livestock farming—Brief review. *Animals* 11, 1–18. doi: 10.3390/ani11082345

Moon, K.-B., Jeon, J.-H., Choi, H., Park, J.-S., Park, S.-J., Lee, H.-J., et al. (2022). Construction of SARS-CoV-2 virus-like particles in plant. *Sci. Rep.* 12, 1005. doi: 10.1038/s41598-022-04883-y

Moon, K., Park, J., Park, Y., Song, I., Lee, H., Cho, H. S., et al. (2019). Development of systems for the production of plant-derived biopharmaceuticals. *Plants* 9, 30. doi: 10.3390/plants9010030

Mor, T. S. (2015). Molecular pharming's foot in the FDA's door: Protalix's trailblazing story. *Biotechnol. Lett.* 37, 2147–2150. doi: 10.1007/s10529-015-1908-z

Moss, D. L., Park, H. W., Mettu, R. R., and Landry, S. J. (2019). Deimmunizing substitutions in Pseudomonas exotoxin domain III perturb antigen processing without eliminating T-cell epitopes. *J. Biol. Chem.* 294, 4667–4681. doi: 10.1074/jbc.RA118.006704

Munasinghe, S. P., Somaratne, S., and Weerakoon, S. R. (2020). Prediction of chemical composition for callus production in Gyrinops walla Gaetner through machine learning. *Inf. Process. Agric.* 7, 511–522. doi: 10.1016/j.inpa.2019.12.001

Navarre, C., Smargiasso, N., Duvivier, L., Nader, J., Far, J., De Pauw, E., et al. (2017). N-Glycosylation of an IgG antibody secreted by Nicotiana tabacum BY-2 cells can be modulated through co-expression of human β-1,4-galactosyltransferase. *Transgenic Res.* 26, 375–384. doi: 10.1007/s11248-017-0013-6

O'Neill, P., Mistry, R. K., Brown, A. J., and James, D. C. (2023). Protein-specific signal peptides for mammalian vector engineering. *bioRxiv*, 532380. doi: 10.1101/2023.03.14.532380

ORF Genetics. (2023). Available at: https://www.orfgenetics.com/.

Ötes, O., Flato, H., Winderl, J., Hubbuch, J., and Capito, F. (2017). Feasibility of using continuous chromatography in downstream processing : Comparison of costs and product quality for a hybrid process vs. a conventional batch process. *J. Biotechnol.* 259, 213–220. doi: 10.1016/j.jbiotec.2017.07.001

Packiam, K. A. R., Ooi, C. W., Li, F., Mei, S., Tey, B. T., Ong, H. F., et al. (2022). PERISCOPE-Opt: Machine learning-based prediction of optimal fermentation conditions and yields of recombinant periplasmic protein expressed in Escherichia coli. *Comput. Struct. Biotechnol. J.* 20, 2909–2920. doi: 10.1016/j.csbj.2022.06.006

Page, M. T., Parry, M. A. J., and Carmo-Silva, E. (2019). A high-throughput transient expression system for rice. *Plant Cell Environ.* 42, 2057–2064. doi: 10.1111/pce.13542

Pan, X., Zuallaert, J., Wang, X., Shen, H. B., Campos, E. P., Marushchak, D. O., et al. (2020). ToxDL: Deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics* 36, 5159–5168. doi: 10.1093/bioinformatics/btaa656

Park, S. R., Lim, C. Y., Kim, D. S., and Ko, K. (2015). Optimization of ammonium sulfate concentration for purification of colorectal cancer vaccine candidate recombinant protein GA733-Fck isolated from plants. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.01040

Peyret, H., Brown, J. K. M., and Lomonossoff, G. P. (2019). Improving plant transient expression through the rational design of synthetic 5′ and 3′ untranslated regions. *Plant Methods* 15, 1–13. doi: 10.1186/s13007-019-0494-9

Quang, D., and Xie, X. (2019). FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* 166, 40–47. doi: 10.1016/j.ymeth.2019.03.020

Qureshi, A. I. (2016). "Chapter 11 - Treatment of Ebola Virus Disease: Therapeutic Agents," in *Ebola Virus Disease: From Origin to Outbreak*. Ed. A. Qureshi (London, San Diego, Cambridge, Oxford: Academic Press), 159–166. doi: 10.1016/B978-0-12-804230-4.00011-X

Radivojević, T., Costello, Z., Workman, K., and Garcia Martin, H. (2020). A machine learning Automated Recommendation Tool for synthetic biology. *Nat. Commun.* 11, 1–14. doi: 10.1038/s41467-020-18008-4

Ramos, J. R. C., Oliveira, G. P., Dumas, P., and Oliveira, R. (2022). Genome-scale modeling of Chinese hamster ovary cells by hybrid semi-parametric flux balance analysis. *Bioprocess Biosyst. Eng.* 45, 1889–1904. doi: 10.1007/s00449-022-02795-9

Ramzi, A. B., Baharum, S. N., Bunawan, H., and Scrutton, N. S. (2020). Streamlining natural products biomanufacturing with omics and machine learning driven microbial engineering. *Front. Bioeng. Biotechnol.* 8. doi: 10.3389/fbioe.2020.608918

Rattanapisit, K., Shanmugaraj, B., Manopwisedjaroen, S., Purwono, P. B., Siriwattananon, K., Khorattanakulchai, N., et al. (2020). Rapid production of SARS-CoV-2 receptor binding domain (RBD) and spike specific monoclonal antibody CR3022 in Nicotiana benthamiana. *Sci. Rep.* 10, 17698. doi: 10.1038/s41598-020-74904-1

Rawat, P., Prabakaran, R., Kumar, S., and Gromiha, M. M. (2021). AbsoluRATE: An in-silico method to predict the aggregation kinetics of native proteins. *Biochim. Biophys. Acta - Proteins Proteomics* 1869, 140682. doi: 10.1016/j.bbapap.2021.140682

Routray, M., Vipsita, S., Sundaray, A., and Kulkarni, S. (2022). DeepRHD: An efficient hybrid feature extraction technique for protein remote homology detection using deep learning strategies. *Comput. Biol. Chem.* 100, 107749. doi: 10.1016/j.compbiolchem.2022.107749

Rozov, S. M., and Deineko, E. V. (2019). Strategies for optimizing recombinant protein synthesis in plant cells: classical approaches and new directions. *Mol. Biol.* 53, 157–175. doi: 10.1134/S0026893319020146

Ruffolo, J. A., Guerra, C., Mahajan, S. P., Sulam, J., and Gray, J. J. (2020). Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinformatics* 36, I268–I275. doi: 10.1093/BIOINFORMATICS/BTAA457

Ruffolo, J. A., Sulam, J., and Gray, J. J. (2022). Antibody structure prediction using interpretable deep learning. *Patterns* 3, 100406. doi: 10.1016/j.patter.2021.100406

Ruffoni, B., Pistelli, L., Bertoli, A., and Pistelli, L. (2010). Plant cell cultures: Bioreactors for industrial production. *Adv. Exp. Med. Biol.* 698, 203–221. doi: 10.1007/978-1-4419-7347-4_15

Russell, S. J. (2010). *Artificial intelligence a modern approach* (New Jersey: Pearson Education, Inc).

Sabi, R., Daniel, R. V., and Tuller, T. (2017). StAIcalc: tRNA adaptation index calculator based on species-specific weights. *Bioinformatics* 33, 589–591. doi: 10.1093/bioinformatics/btw647

Sahu, S. S., Loaiza, C. D., and Kaundal, R. (2021). Plant-mSubP: A computational framework for the prediction of single- And multi-target protein subcellular localization using integrated machine-learning approaches. *AoB Plants* 12, 1–10. doi: 10.1093/AOBPLA/PLZ068

Sainsbury, F. (2020). Innovation in plant-based transient protein expression for infectious disease prevention and preparedness. *Curr. Opin. Biotechnol.* 61, 110–115. doi: 10.1016/j.copbio.2019.11.002

Samoudi, M., Masson, H. O., Kuo, C. C., Robinson, C. M., and Lewis, N. E. (2021). From omics to cellular mechanisms in mammalian cell factory development. *Curr. Opin. Chem. Eng.* 32, 100688. doi: 10.1016/j.coche.2021.100688

Sangjan, W., Carter, A. H., Pumphrey, M. O., Jitkov, V., and Sankaran, S. (2021). Development of a raspberry pi-based sensor system for automated in-field monitoring to support crop breeding programs. *Inventions* 6, 42. doi: 10.3390/inventions6020042

Sara, S. T., Hasan, M. M., Ahmad, A., and Shatabda, S. (2021). Convolutional neural networks with image representation of amino acid sequences for protein function prediction. *Comput. Biol. Chem.* 92, 107494. doi: 10.1016/j.compbiolchem.2021.107494

Sarkar, S., Cazenave, A. B., Oakes, J., McCall, D., Thomason, W., Abbott, L., et al. (2021). Aerial high-throughput phenotyping of peanut leaf area index and lateral growth. *Sci. Rep.* 11, 1–17. doi: 10.1038/s41598-021-00936-w

Sarker, I. H. (2021). Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* 2, 1–21. doi: 10.1007/s42979-021-00592-x

Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., et al. (2019). The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nat. Commun.* 10, 1–14. doi: 10.1038/s41467-019-13483-w

Schillberg, S., and Finnern, R. (2021). Plant molecular farming for the production of valuable proteins - Critical evaluation of achievements and future challenges. *J. Plant Physiol.* 258–259, 153359. doi: 10.1016/j.jplph.2020.153359

Schillberg, S., Raven, N., Spiegel, H., Rasche, S., and Buntru, M. (2019). Critical analysis of the commercial potential of plants for the production of recombinant proteins. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00720

Schjoldager, K. T., Narimatsu, Y., Joshi, H. J., and Clausen, H. (2020). Global view of human protein glycosylation pathways and functions. *Nat. Rev. Mol. Cell Biol.* 21, 729–749. doi: 10.1038/s41580-020-00294-x

Sethi, L., Kumari, K., and Dey, N. (2021). Engineering of plants for efficient production of therapeutics. *Mol. Biotechnol.* 63, 1125–1137. doi: 10.1007/s12033-021-00381-0

Shanmugaraj, B., Rattanapisit, K., Manopwisedjaroen, S., Thitithanyanont, A., and Phoolcharoen, W. (2020). Monoclonal Antibodies B38 and H4 Produced in Nicotiana benthamiana Neutralize SARS-CoV-2 in *vitro*. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.589995

Shayesteh, M., Ghasemi, F., Tabandeh, F., Yakhchali, B., and Shakibaie, M. (2020). Design, construction, and expression of recombinant human interferon beta gene in CHO-s cell line using EBV-based expression system. *Res. Pharm. Sci.* 15, 144–153. doi: 10.4103/1735-5362.283814

Shi, X., Cordero, T., Garrigues, S., Marcos, J. F., Daròs, J. A., and Coca, M. (2019). Efficient production of antifungal proteins in plants using a new transient expression vector derived from tobacco mosaic virus. *Plant Biotechnol. J.* 17, 1069–1080. doi: 10.1111/pbi.13038

Silva, J. C. F., Teixeira, R. M., Silva, F. F., Brommonschenkel, S. H., and Fontes, E. P. B. (2019). Machine learning approaches and their current application in plant molecular biology: A systematic review. *Plant Sci.* 284, 37–47. doi: 10.1016/j.plantsci.2019.03.020

Singh, A., Ganapathysubramanian, B., Singh, A. K., and Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci.* 21, 110–124. doi: 10.1016/j.tplants.2015.10.015

Siriwattananon, K., Manopwisedjaroen, S., Shanmugaraj, B., Rattanapisit, K., Phumiamorn, S., Sapsutthipas, S., et al. (2021). Plant-produced receptor-binding domain of SARS-coV-2 elicits potent neutralizing responses in mice and non-human primates. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.682953

Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., and Frishman, D. (2012). PROSO II - A new method for protein solubility prediction. *FEBS J.* 279, 2192–2200. doi: 10.1111/j.1742-4658.2012.08603.x

Smiatek, J., Clemens, C., Herrera, L. M., Arnold, S., Knapp, B., Presser, B., et al. (2021). Generic and specific recurrent neural network models: Applications for large and small scale biopharmaceutical upstream processes. *Biotechnol. Rep.* 31, e00640. doi: 10.1016/j.btre.2021.e00640

Soni, A. P., Lee, J., Shin, K., Koiwa, H., and Hwang, I. (2022). Production of recombinant active human TGFβ1 in nicotiana benthamiana. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.922694

Strain, B., Morrissey, J., Antonakoudis, A., and Kontoravdi, C. (2023). Genome-scale models as a vehicle for knowledge transfer from microbial to mammalian cell systems. *Comput. Struct. Biotechnol. J.* 21, 1543–1549. doi: 10.1016/j.csbj.2023.02.011

Strasser, R. (2022). Recent developments in deciphering the biological role of plant complex N-glycans. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.897549

Strasser, R. (2023). Plant glycoengineering for designing next-generation vaccines and therapeutic proteins. *Biotechnol. Adv.* 67, 108197. doi: 10.1016/j.bioteChadv.2023.108197

Sun, X., Yang, Z., Su, P., Wei, K., Wang, Z., Yang, C., et al. (2023). Non-destructive monitoring of maize LAI by fusing UAV spectral and textural features. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1158837

Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R., and Atalay, V. (2019). DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. *Sci. Rep.* 9, 1–16. doi: 10.1038/s41598-019-43708-3

Taghavi Namin, S., Esmaeilzadeh, M., Najafi, M., Brown, T. B., and Borevitz, J. O. (2018). Deep phenotyping: Deep learning for temporal phenotype/genotype classification. *Plant Methods* 14, 1–14. doi: 10.1186/s13007-018-0333-4

Tausen, M., Clausen, M., Moeskjær, S., Shihavuddin, A. S. M., Dahl, A. B., Janss, L., et al. (2020). Greenotyper: image-based plant phenotyping using distributed computing and deep learning. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.01181

Tien, N. Q. D., Huy, N. X., and Kim, M. Y. (2019). Improved expression of porcine epidemic diarrhea antigen by fusion with cholera toxin B subunit and chloroplast transformation in Nicotiana tabacum. *Plant Cell. Tissue Organ Cult.* 137, 213–223. doi: 10.1007/s11240-019-01562-1

Tokekar, P., Vander Hook, J., Mulla, D., and Isler, V. (2016). Sensor planning for a symbiotic UAV and UGV system for precision agriculture. *IEEE Trans. Robot.* 32, 1498–1511. doi: 10.1109/TRO.2016.2603528

Tovar, J. C., Hoyer, J. S., Lin, A., Tielking, A., Callen, S. T., Elizabeth Castillo, S., et al. (2018). Raspberry Pi–powered imaging for plant phenotyping. *Appl. Plant Sci.* 6, 1–12. doi: 10.1002/aps3.1031

Tuan-Anh, T., Ly, L. T., Viet, N. Q., and Bao, P. T. (2017). Novel methods to optimize gene and statistic test for evaluation - an application for Escherichia coli. *BMC Bioinf.* 18, 1–10. doi: 10.1186/s12859-017-1517-z

Ubbens, J. R., and Stavness, I. (2017). Deep plant phenomics: A deep learning platform for complex plant phenotyping tasks. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01190

Vafaee, Y., and Alizadeh, H. (2018). Heterologous production of recombinant anti-HIV microbicide griffithsin in transgenic lettuce and tobacco lines. *Plant Cell. Tissue Organ Cult.* 135, 85–97. doi: 10.1007/s11240-018-1445-2

Vaishnav, E. D., de Boer, C. G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., et al. (2022). The evolution, evolvability and engineering of gene regulatory DNA. *Nature* 603, 455–463. doi: 10.1038/s41586-022-04506-6

Van Brempt, M., Clauwaert, J., Mey, F., Stock, M., Maertens, J., Waegeman, W., et al. (2020). Predictive design of sigma factor-specific promoters. *Nat. Commun.* 11, 1–13. doi: 10.1038/s41467-020-19446-w

van Dijk, A. D. J., Kootstra, G., Kruijer, W., and de Ridder, D. (2021). Machine learning in plant science and plant breeding. *iScience* 24, 101890. doi: 10.1016/j.isci.2020.101890

Vaškevičius, M., Kapočiūtė-Dzikienė, J., and Šlepikas, L. (2021). Prediction of chromatography conditions for purification in organic synthesis using deep learning. *Molecules* 26, 2474. doi: 10.3390/molecules26092474

Vazquez-Vilar, M., Selma, S., and Orzaez, D. (2023). The design of synthetic gene circuits in plants: new components, old challenges. *J. Exp. Bot.* 74, 3791–3805. doi: 10.1093/jxb/erad167

Viet, N. D., and Jang, A. (2021). Journal of Environmental Chemical Engineering Development of artificial intelligence-based models for the prediction of filtration performance and membrane fouling in an osmotic membrane bioreactor. *J. Environ. Chem. Eng.* 9, 105337. doi: 10.1016/j.jece.2021.105337

Vo ngoc, L., Huang, C. Y., Cassidy, C. J., Medrano, C., and Kadonaga, J. T. (2020). Identification of the human DPR core promoter element using machine learning. *Nature* 585, 459–463. doi: 10.1038/s41586-020-2689-7

Wan, S., Zhao, K., Lu, Z., Li, J., Lu, T., and Wang, H. (2022). A modularized ioT monitoring system with edge-computing for aquaponics. *Sensors* 22, 9260. doi: 10.3390/s22239260

Wang, X., Li, F., Xu, J., Rong, J., Webb, G. I., Ge, Z., et al. (2022). ASPIRER: A new computational approach for identifying non-classical secreted proteins based on deep learning. *Brief. Bioinform.* 23, 1–12. doi: 10.1093/bib/bbac031

Wang, L., Nie, R., Yu, Z., Xin, R., Zheng, C., Zhang, Z., et al. (2020). An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data. *Nat. Mach. Intell.* 2, 693–703. doi: 10.1038/s42256-020-00244-4

Webster, G. R., Teh, A. Y. H., and Ma, J. K. C. (2017). Synthetic gene design—The rationale for codon optimization and implications for molecular pharming in plants. *Biotechnol. Bioeng.* 114, 492–502. doi: 10.1002/bit.26183

Weissenow, K., Heinzinger, M., and Rost, B. (2022). Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure* 30, 1169–1177.e4. doi: 10.1016/j.str.2022.05.001

Wittmann, B. J., Johnston, K. E., Wu, Z., and Arnold, F. H. (2021). Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* 69, 11–18. doi: 10.1016/j.sbi.2021.01.008

Wu, M. R., Nissim, L., Stupp, D., Pery, E., Binder-Nissim, A., Weisinger, K., et al. (2019). A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). *Nat. Commun.* 10, 1–10. doi: 10.1038/s41467-019-10912-8

Wu, Z., Yang, K. K., Liszka, M. J., Lee, A., Batzilla, A., Wernick, D., et al. (2020). Signal peptides generated by attention-based neural networks. *ACS Synth. Biol.* 9, 2154–2161. doi: 10.1021/acssynbio.0c00219

Wu, X., and Yu, L. (2021). EPSOL: sequence-based protein solubility prediction using multidimensional embedding. *Bioinformatics* 37, 4314–4320. doi: 10.1093/bioinformatics/btab463

Yang, Z., Bogdan, P., and Nazarian, S. (2021b). An in silico deep learning approach to multi-epitope vaccine design: a SARS-CoV-2 case study. *Sci. Rep.* 11, 1–21. doi: 10.1038/s41598-021-81749-9

Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., et al. (2017). Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. *Methods Mol. Biol.* 1484, 55–63. doi: 10.1007/978-1-4939-6406-2_6

Yang, H. S., Rhoads, D. D., Sepulveda, J., Zang, C., Chadburn, A., and Wang, F. (2023). Challenges and considerations of developing and implementing machine learning tools for clinical laboratory medicine practice. *Arch. Pathol. Lab. Med.* 147, 826–836. doi: 10.5858/arpa.2021-0635-RA

Yang, T., Zhang, W., Zhou, T., Wu, W., Liu, T., and Sun, C. (2021a). Plant phenomics & precision agriculture simulation of winter wheat growth by the assimilation of unmanned aerial vehicle imagery into the WOFOST model. *PloS One* 16, 1–9. doi: 10.1371/journal.pone.0246874

Yoosefzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J., and Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.624273

Yu, S. I., Rhee, C., Cho, K. H., and Shin, S. G. (2022). Comparison of different machine learning algorithms to estimate liquid level for bioreactor management. *Environ. Eng. Res.* 28, 220037–220030. doi: 10.4491/eer.2022.037

Zangirolami, T. C., Campani, G., Horta, A. C. L., and Giordano, R. C. (2021). Machine learning applied for metabolic flux - based control of micro - aerated fermentations in bioreactors.. *Biotechnol. Bioeng.* 118, 2076–2091. doi: 10.1002/bit.27721

Zaragoza, J. M. C. (2022). *Data-Driven Cell Engineering of Chinese Hamster Ovary Cells through Machine Learning*. Denmark: Technical University of Denmark.

Zhang, J., Petersen, S. D., Radivojevic, T., Ramirez, A., Pérez-Manríquez, A., Abeliuk, E., et al. (2020). Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat. Commun.* 11, 4880. doi: 10.1038/s41467-020-17910-1

Zhao, W., Zhou, L. Y., Kong, J., Huang, Z. H., Gao, Y., Zhang, Z. X., et al. (2023). Expression of recombinant human Apolipoprotein A-IMilano in Nicotiana tabacum. *Bioresour. Bioprocess.* 10 (1), 1–14. doi: 10.1186/s40643-023-00623-w

Zheng, Y. Y., Kong, J. L., Jin, X. B., Wang, X. Y., Su, T. L., and Zuo, M. (2019). Cropdeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors (Switzerland)* 19, 1058. doi: 10.3390/s19051058

Zrimec, J., Börlin, C. S., Buric, F., Muhammad, A. S., Chen, R., Siewers, V., et al. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* 11, 6141. doi: 10.1038/s41467-020-19921-4