Biotechnology Notes 6 (2025) 126–132

Contents lists available at ScienceDirect

Biotechnology Notes

journal homepage: www.keaipublishing.com/en/journals/biotechnology-notes/

Cell-free protein synthesis platforms for accelerating drug discovery

Filippo Caschera

Independent Consultant, Rome, Italy

ARTICLE INFO

Keywords:
Cell-free protein synthesis
Platform
Streamlined
Open system
DBTL cycle
Machine learning

ABSTRACT

Cell-free protein synthesis is a platform for streamlined production of macromolecules. Recently, several proteins with pharmaceutical relevance were synthesised and characterised. Off-the-shelf reagents and parallelised experimentation have enabled the exploration of many different conditions for *in vitro* protein synthesis and engineering. Herein is described how machine learning algorithms were applied for protein yield maximisation as well as for protein engineering and *de novo* design. Cell-free protein synthesis provides the biotechnological platform to unlock the power and benefit of AI/ML for drug discovery and improve human health.

1. Introduction

The synthesis of proteins in the laboratory is an important process that need to be streamlined. This optimisation will improve human health through the discovery and prototyping of more effective biotherapeutics. Cell-free protein synthesis (CFPS) is the process that recapitulates the central dogma of molecular biology *in vitro* including applications in biomanufacturing, metabolic engineering[1–3] and synthetic cells.[4] A comprehensive review with all the historical landmarks and applications has been recently published[5] besides to an examination on pharmaceutical molecule production in cell-free systems.[6] CFPS is an integrated system with three modules combined: The lysate (proteins), the energy (small molecules) and the DNA modules are interlocked to generate an open system in which all the parameters necessary to synthesise the proteins *in vitro* can be varied independently.[7,8] Therefore, representing a perfect platform to unlock the combinatorial power of artificial intelligence (AI). CFPS is an easily accessible platform for protein production at small scales with several advantages compared to the expression *in vivo*, such as the decoupling of cell growth from protein synthesis.[9] Importantly, CFPS is enabled by linear DNA containing all the elements for transcription and translation.[10] This is advantageous because linear expression templates (LETs) can be synthesised in few days bypassing tedious molecular cloning procedures necessary for plasmids construction and propagation that require weeks. Furthermore, to detect proteins produced during or after the CFPS reaction complementation assays were developed with engineered split-proteins,[11,12] and a quantification method using a fluorescence dye and a minihelix was described.[13] Off-the-shelf reagents can be used for high-throughput drug discovery campaigns exploiting automatic workflows for protein synthesis and purification using liquid handling equipment[14,15] and microfluidic systems.[16] CFPS platforms have some key features that make this technology advantageous for accelerating proteins discovery and prototyping, such as customisable reagents[17,18] and accelerated timelines.[19] In addition, accessibility and quick response to drug discovery campaigns can be further streamlined with off-the-shelf reagents prepared lyophilized and stored at room temperature.[20] A streamlined workflow for protein production can shorten the time for synthesis and protein in hand from days to hours. As a result, protein variants can be quickly screened and selected for properties such as: solubility, purifiability, stability, binding capacity, etc. Overall, such features set the conditions for using CFPS as a time-efficient and cost-effective tool for drug discovery. The drug discovery workflows can benefit from CFPS integration to bypass laborious *in vivo* protein expression protocols and shorten the timeline of biopharmaceuticals design, synthesis and purification filtering for drug candidates that can be tested for efficacy in animal models. The open environment of CFPS systems enables multi-parameter optimisation with machine learning (ML) to find specific reaction formulations that maximise CFPS yileds, producing sufficient product quantities for testing in animal models. Furthermore, drug discovery pipelines can benefit from the amount of data generated through CFPS reaction miniaturisation to guide ML through iterative processes of peptide engineering exploring the sequence to function protein fitness landscape or to train models providing accurate predictions. In addition, the easy accessibility of CFPS systems can be useful to quickly validate the expression and activity of hundreds of *de novo* drug candidates designed with AI. The

Peer review under the responsibility of Editorial Board of Biotechnology Notes.
E-mail address: filippocaschera@gmail.com.

https://doi.org/10.1016/j.biotno.2025.02.001
Received 24 December 2024; Received in revised form 1 February 2025; Accepted 17 February 2025
Available online 19 February 2025

sections below highlight some milestones concerning pharmaceutical relevant proteins produced with CFPS and the use of the technology as platform for ML optimisation and protein design.

2. Milestone events of pharmaceutically relevant candidates produced with CFPS

Over the last 20 years CFPS has become a popular tool with many biotechnological applications and developments.[9] Some developments have included the implementation of large-scale production technologies following Good Manufacturing Practice (GMP) guidelines, which have resulted in several human clinical trials of CFPS-based products.[21] More recently, CFPS has been positioned as a tool for personalised medicine and diagnostics offering rapid and adaptable platforms for customisable therapeutics.[22] In this section are described some examples of CFPS platforms developed to produce pharmaceutically relevant proteins. An accelerated cell-free workflow for antibody discovery was implemented to determine the binding capacity for the SARS-CoV-2 spike protein of synthetically dimerized antigen-binding fragments (sdFab).[23] The platform leveraged high-throughput (HTP) liner DNA assembly *in vitro* and expression in an *E. coli* CFPS optimised for disulphide bonds formation. The output from B-cells was used to identify effective antibody sequences and the binding capacity of the cell-free synthesised fragments was evaluated with an AlphaLISA assay.[23] A novel workflow leveraging CFPS, called "deep screening" was demonstrated effective in discovering single-chain fragment variables (scFVs) with improved binding affinity against clinically relevant targets in a single 3-day experiment.[24] The "deep screening" pipeline implied the grouping and sequencing of antibody libraries, the conversion of DNA into mRNA on a chip and *in situ* translation into antibody fragments linked by ribosome display and screening through labelled antigens. Image analysis coupled to a ML algorithm (bidirectional encoder representations from transformers, BERT) enabled the design of complementary-determining regions (CDRs) and the discovery of antibodies with up to 5200-fold increased binding affinity. Notably, with a small library diversity (4×10^6 or 2×10^5) compared to batch selection methods such as, mRNA, ribosome, yeast and phage display (10^7 to 10^{12}).[24]

Important efforts were conceived to design CFPS systems for non-canonical amino acids (ncAAs) incorporation bypassing the drawbacks of cell viability. The designs relied on recoded *E. coli* strains leveraging the amber suppression mechanism.[25,26] A genomically recoded *E. coli* strain release factor 1-deficient, without the Sep-specific phosphatase gene *SerB*, and engineered with an orthogonal translation system (OTS) was utilised for site-specific incorporation of phosphoserine to the human MEK 1 kinase, which was successfully expressed in milligram quantities using CFPS(27). This result enabled an important post-translational modification (PTM) that was used to recapitulate the physiological signalling cascade *in vitro*.[27] The engineering of *E. coli* strains release factor 1-deficient was also relevant for site-specific conjugation via click chemistry to develop antibody-drug conjugates (ADCs) with payloads that enhance efficacy against tumour cells.[28] A cell-free workflow for the synthesis of ADCs was described based on the addition to the CFPS reaction of prefabricated light-chains (LCs) previously produced and purified from *E. coli*. The develop of a *in vivo/in vitro* protein synthesis platform based on efficient LCs manufacturing and CFPS incorporation of ncAAs into Immunoglobulin G (IgG) resulted in 50–100 % increase in monoclonal antibody (mAB) titers.[28] Resource allocation during CFPS was likely the reason for the increase in protein yield. The hybrid system was validated with five mABs at CFPS production scale from 0.1 to 250 ml. The precise incorporation of a ncAA into the IgG was utilised for azide-alkyne click chemistry to conjugate a linker (SC236) with the cytotoxic payload maytansine generating ADCs, which cytotoxicity was demonstrated with *in vitro* proliferation assays.[28] Another application of ncAAs incorporation with demonstrated efficacy regards the development of vaccines. Through antigen-adjuvant conjugation was demonstrated superior antigen presentation *in vitro* and major antigen specific $CD8^+$ T-cell production *in vivo*.[29] The toll-like-receptor (TLR) agonist was conjugated with copper free click chemistry following CFPS incorporation of the *p*-azidomethyl-L-phenylalanine (pAMF) into the amino acid sequence of the model antigen, ovalbumin (OVA). Overall, CFPS resulted in an efficient methodology and easy-accessible technique for the incorporation of ncAAs to produce large quantities of antigen-adjuvant vaccines.[29] CFPS was utilised as a platform to prototype virus-like particles (VLPs) and design a VLP-antigen vaccine with increased stability and lower immunogenicity.[30] An *E. coli* CFPS system was used to rapidly screen HBc protein variants to build an artificial disulphide bridge network stabilising the nanoparticle. Furthermore, the negative charge on its surface was reduced to improve conjugation efficiency via ncAAs incorporation and click chemistry of immuno-therapeutics and adjuvants. Importantly, low immunogenicity was conferred by transplanting a modified version of the spike from a natural mutant.[30] In addition, CFPS was also utilised to produce PD-O78 glycoconjugate vaccines leveraging an engineered *E. coli* strain harbouring the glycosylation machinery.[31] This study was important to demonstrate the production of vaccines that generate bactericidal antibodies using CFPS production in a low-cost lyophilized format and therefore exhibiting the potential of the technology for decentralised biomanufacturing augmenting non-refrigerated storage and transportation.[31]

CFPS is a practical platform for the expression and characterisation of G protein-coupled receptors (GPCRs). This family of membrane proteins is involved in many physiological processes and is the major target for drug development with 34 % of all drugs approved by FDA.[32] Several types of CFPS systems, prokaryotic and eukaryotic lysates, were applied for the synthesis of GPCRs. For instance, a human GPCR (hOR17-4) was synthesised using the wheat germ lysate in the presence of the detergent Digitonin maintaining the solubility of the membrane protein[33]. The protein was successfully purified with size exclusion chromatography (SEC) and α-helical structure characterised with circular dichroism (CD). The affinity constant K_D of the receptor for its ligand was derived from surface plasmon resonance (SPR) measurements.[33] In another example, six GPCRs (four human) were successfully synthesised in an *E. coli* lysate with the Brij78 detergent, purified and ligand binding investigated to determine the K_D with total internal reflection fluorescence spectroscopy (TIRFS).[34] An insect lysate (*Sf21*) containing endogenous microsomes, which are membrane compartments derived from the endoplasmic reticulum that offer a native lipid environment and active translocon mechanism for the correct insertion and folding of membrane proteins, in addition to PTMs, was utilised for the synthesis of a GPCR (E-TB).[35] Ligand binding was demonstrated without purification with a fluorescence assay and K_D was obtained from a radioligand binding assay. Interestingly, accessibility of the ligand for the receptor through the microsomes was enabled with the supplementation of 0.03 % Brij35 that perforated the membranes.[35] More recently, an *E. coli* CFPS system was used for the cryo-electron microscopy (Cryo-EM) structure determination of the human histamine 2 receptor (H_2R) in an active conformation with histamine and in complex with G_s heterotrimeric protein[36]. The complex was formed with full length H_2R, G_s and the G_s stabilising nanobody Nb35 in nanodiscs composed of DOPG lipids. In addition, this study was significant because the membrane protein was expressed in a detergent free environment, which may cause receptor incorrect folding.[36] Eukaryotic CFPS based on *Sf21* and CHO lysates were compared for efficient production of the GPCR glucagon-like peptide 1 receptor (GLP-1R).[37] Interestingly, the yield of the receptor was higher for the CHO CFPS system but the correct glycosylation, necessary for GPCR activity, was achieved only for the insect (*Sf21*) CFPS system. Importantly, the accessibility to the extracellular domain of the receptor integrated in immobilized microsomes was demonstrated with an ELISA assay and therefore a radioligand binding assay was applied successfully, proposing eukaryotic CFPS as a screening platform for drug development.[37]

Antimicrobial resistance (AMR) has recently raised concerns due to the overuse of antibiotics in the industry and the intrinsically antibiotic resistance of some pathogens. This concerning scenario has re-energised bacterio phage therapy as an efficient alternative to treat multidrug-resistant bacterial infections and CFPS platforms can accelerate phage production on demand.[38] A novel production cell-free workflow, "phactory", was presented to produce clinically relevant phages targeting *K. pneumoniae*, *Y. pestis* and ECEC with genomic DNA isolated from purified phage stocks at clinically relevant titers of 10^8 PFU/ml.[39] Taking advantage of CFPS as an open system through the pipeline was possible to reboot the capacity of phage production synthesising in the reaction the exogenous sigma factor of *B. subtilis*, and as a result extend the *in vitro* system to produce phages against a Gram-positive bacterium. Furthermore, the capsid of T7 was engineered via co-expression of the minor capsid protein with a polyhistidine-tag that has allowed purification with affinity chromatography of active phages at higher titers through a transient non-genomic modification.[39] A novel CFPS workflow for phage engineering named PHEIGES, phage engineering by in *vitro* gene expression and selection, was established for T7.[40] This represents a rapid workflow for genome engineering, synthesis and selection. Engineered T7 genomes were rapidly assembled *in vitro* and used to select mutants with increased host range without need of droplet compartimentalisation due to an emerging mechanism that guaranteed genotype-phenotype linkage in the CFPS reaction.[40] Further investigation will determine the generality of the genotype-phenotype linkage to other phages produced *in vitro*. Interestingly, PHEIGES was applied to the *Salmonella* phage FelixO1 which is the largest phage genome with a titer of 10^8 PFU/ml, and in addition was demonstrated the expression of *E.coli* phages VpaE1, T6 and the *Salmonella* phages FelixO1 and S16 from their purified genomes therefore validating the capability of CFPS to synthesise *E. coli* and non-*E. coli* phages with genomes ranging from 40 Kb to 170 Kb.[40]

Antibiotic resistance represents one of the greatest threats to global health and peptide-based antibiotics (PBAs) offer an alternative to traditional antibiotics with more effective mechanisms of action against resistant pathogens.[41] CFPS has emerged as a powerful platform for the discovery and synthesis of PBAs given its accessibility, flexibility and speed.[41] PBAs can be divided in three groups. The first group is the antimicrobial peptides (AMPs), which contain the 20 canonical amino acids and can be characterised by α-helical structure, proline-rich (PrAMPs), tryptophan-rich (WrAMPs), arginine-rich (RrAMPs) amino acids and cyclic structure formed by disulphide bonds. The AMPs have different mechanisms of action depending on the structure, such as disruption of the cell membranes or inhibition of protein synthesis after membrane penetration.[41] The second group is the ribosomally synthesised and post-translational modified peptides (RiPPs) containing more than the 20 canonical amino acids with structure past linear peptide and diversity arising from several PTMs including cyclisation, cyclodehydration, glycosylation, N-alkylation and more.[41] The RiPPs are a broad class of natural products with a vast structural diversity and functionalities that can go beyond antimicrobial activity.[42] A novel CFPS platform for the synthesis and activity evaluation of natural products such as RiPPs was developed in the context of AMR.[43] The unified biocatalyst (UniBioCat) recapitulates the biosynthetic pathway *in vitro* co-expressing the precursor peptide and modification enzymes to produce the mature RiPPs. The production efficiency of the lanthipeptide, salivaracin B, was increased using a lysate prepared from an engineered *E. coli* strain proteases deficient and with chaperones added to the CFPS reaction.[43] Overall, such modifications resulted in increased protein yields. The UniBioCat platform accelerated the engineering of salivaracin B with improved antimicrobial activity and the evaluation of ten uncharacterized lanthipeptides.[43] The first two groups of PBAs, AMPs and RiPPs, are synthesised by the translation apparatus while the third group includes non-ribosomally synthesised peptides produced by complex enzymatic processes mediated by the non-ribosomal peptide synthetases (NRPSs), which are large multi-modular enzyme complexes that catalyse the incorporation of diverse non-canonical substrates (D-amino acids, β-amino acids, hydroxyacids etc.) generating peptides with unique structures and efficacy.[41] Moreover, an innovative method to generate peptides with ncAAs is the Flexyzyme system that allows tRNA aminoacylation for the translation apparatus[44] consisting in artificial ribozymes generating mis-acylated tRNA enabling ncAAs incorporation during *in vitro* translation[45,46] leading, for instance, to the potential discovery of pseudo-natural macrocyclic peptides.[47]

A CFPS platform of the pathogenic bacterium *K. pnemoniae* was conceived to contrast the rise of AMR.[48] The platform increased the throughput and shortened the timelines to study antimicrobials and intracellular AMR factors. The development of the CFPS system from a World Health Organization significant pathogen improves accessibility for developing and screening new chemical and natural compounds in BSF-1 laboratories.[48] The CFPS system developed, was demonstrated effective to determine the minimum inhibitory concentration (MIC) and the half inhibitory concentration (IC_{50}) of 15 antibiotics with five compounds having the same dose-response curve between the CFPS system and an *in vivo* membrane-permeability assay. Therefore, demonstrating that CFPS from a pathogenic bacterium can be used to develop a useful bioactivity model to study antimicrobial activity within cells.[48]

Table 1 summarises cell-free synthesised products with pharmaceutical relevance described in this section.

3. Machine learning optimisation of CFPS for suitable drug efficacy validation

The CFPS integration into drug discovery workflows will shorten the timelines for biopharmaceuticals production and efficacy validation in animal models. The optimisation of CFPS should provide enough protein for efficacy tests *in vivo*. For instance, CFPS-conjugated vaccines were administered at 7.5–10 μg per mouse.[49] In another example, anti-PD-L1 nanobodies were tested in mouse models at a dose concentration range of 0.1–0.92 mg/kg.[50] Furthermore, nanobodies targeting GPCR were tested for efficacy in monkeys at a dose concentration range of 0.1–25 mg/kg.[51] These examples suggest that if a mouse weighs ~25 g, then 2.5–23 μg of protein would be necessary. Therefore, would be reasonable if 100 μl of CFPS reaction produced ~0.025–0.23 g/l. In contrast, if a monkey weighs ~5 kg, it could require a 10000 μl CFPS reaction producing 5 g/l or more. However, such yield has never been reported for a pharmaceutically relevant protein and is challenging to achieve. Therefore, the CFPS reaction would need to be scaled up to deliver enough protein for a study in large animal models. Assuming a 100 l CFPS reaction[52] producing ~0.5 g/l[53] this would be sufficient to guarantee 1000 doses at a 10 mg/kg dose in monkeys of 5 kg body weight. However, for a comprehensive study involving 100 monkeys would be sufficient a more accessible 10 l CFPS reaction format.[54] These observations suggest that CFPS should be optimised to guarantee enough protein for *in vivo* testing, considering that yields are protein dependent and cost-effective formulations should be devised to face the challenges of scaling-up the reaction. The most efficient method to simultaneously

Table 1
Cell-free synthesised pharmaceuticals.

Publication	Year	Pharmaceutical	Relevance
A. C. Hunt et al.[23]	2023	sdFab	SARS-CoV-2 spike protein
B. T. Porebski et al.[24]	2024	scFVs	HER 2, ERBB2, huIL-7
J. P. Oza et al.[27]	2015	MEK 1	Protein kinase
J. Hanson et al.[28]	2023	ADC	Cytotoxicity
A. M. Weiss et al.[29]	2021	OVA	Antigen-adjuvant conjugation
Y. Lu[30]	2015	VLP	VLP-antigen vaccine
K. F. Warfel[31]	2023	PD-O78	Glycoconjugate vaccine
Z. Kock[36]	2024	H_2R	Human GPCR receptor
L. Haueis[37]	2023	GLP-1R	Human GPCR receptor
W. Q. Liu[43]	2024	salivaracin B	Peptide based antibiotic

maximise CFPS reaction yields and reduce costs would be to apply ML algorithms incorporating for instance a multi-objective fitness function.

E. coli CFPS is the result of a modular design in which the lysate (proteins), the energy and the DNA modules are the three integrated subsystems (LED integration).[8] CFPS is an open system where small molecule mixtures (NTPs, amino acids, co-factors, salts, high-energy phosphate donor substrates, crowding agents etc.) fuel the transcription-translation representing the energy module.[7,55] The modules underline a complex experimental space where the parameters can be varied on different levels to optimise the CFPS system output. Large experimental spaces can be designed reaching for example 10^{12} possible combinations only varying the concentration levels of the small molecules fuelling the system.[56] As a result, it becomes impossible to explore the experimental space exhaustively and therefore ML algorithms can be used to allocate intelligently the experimental effort exploring only a small subset of the possible combinations and successfully optimise the fitness function of the system. The design-build-test-learn (DBTL) cycle depicted in Fig. 1 represents the iterative framework that leverages CFPS and ML to optimise protein synthesis and engineering. The DBTL cycle integrates CFPS and ML in an iterative process that leads to an increase of the fitness function accelerating drug discovery. The fitness function could be the yield of a target protein and/or the design of cost-effective formulations for large scale production. In addition, the DBTL cycle could be applied to engineer proteins with enhanced properties, such as antibodies with better affinity and efficacy or enzymes with higher catalytic efficiency.

A ML approach leveraging statistical modelling and intelligent stochastic exploration has made substantial improvement on the optimisation of CFPS.[57] During the DBTL iterative cycle, a neural network model predicted the best combinations of molecules for CFPS in a 16-dimesional input space. The parameters underling CFPS were divided in four groups comprising concentration and time variables yielding an experimental space of more than 10^6 possible combinations. In particular, the experimental design included different DNA sequences with diverse spacer regions between the regulatory parts in combinations with small molecule concentrations added at distinctive time intervals into a diluted *E. coli* lysate, therefore starting from a suboptimal condition. Through the ML algorithm developed it was possible to explore only a small percentage (0.014 %) of the total experimental space and improve protein expression ~ 3-fold with eight iterations.[57] The iSAT system (*in vitro* integrated synthesis, assembly and translation) for the bottom-up construction of ribosomes is a complex cell-free system and its optimisation using ML was presented in a 20-dimensional experimental space.[56] For the first time a computationally guided optimisation was proposed for the construction and optimisation of ribosomes *in vitro* utilising HTP cell-free ribosomes assembly and translation leveraging the DBTL cycle. The goal of the ML guided experimentation was to implement the iSAT system using a common accessible E. *coli* strain (suboptimal for mRNA stability) defining a cost-effective solution. An experimental space of ~10^{12} possible combinations was designed from a library of 20 molecules on four concentration levels. The ML algorithm was adapted to the liquid handling robotic constrains which allowed only the exploration of points located in proximity of different "pivot" experiments defined by the model. As a result, after seven iterations the iSAT system was improved ~10-fold with a 4-fold cost reduction omitting most of the components from the reaction. This report has demonstrated the power of ML in optimizing complex systems *in vitro* and postulated that the conditions discovered are a form of model overfitting of the specific protein expressed.[56] Maximisation of CFPS yields with ML was further demonstrated in a 11-dimensional experimental space varying the composition of the small molecules in the energy module on four concentrations levels.[58] With the iterative DBTL process an experimental space of ~4×10^6 possible combinations was explored using only 20 CFPS combinations to train the model during each iteration. As a result of the optimisation a ~ 4-fold increase in protein yield was achieved after ten iterations. Moreover, model prediction accuracy was demonstrated to define a large diversity of yields depending on cell-free reaction conditions.[58] In another study, ML has optimised the CFPS formulation for gene expression using a bacterial chromosome, an aspect important for synthetic cells research.[59] The exploration of a 11-dimensional experimental space with a ML algorithm defined the parameters for optimal protein synthesis from low DNA concentration in a diluted lysate. The algorithm has optimised the cell-free system through the DBTL cycle decreasing the exploration factor over ten iterations yielding an 8.6-fold increase in protein expressed. In addition, the data generated guided improvement in the upstream process of chromosomal DNA purification.[59] A novel scalable microfluidic platform named DropAI was presented as powerful tool for screening CFPS reactions with the potential to screen 10^8 combinatorial reactions per day through parallelisation.[60] DropAI is a droplet-based and AI-guided combinatorial screening platform established to optimise CFPS through a fluorescence-based coding de-coding system able to screen and identify optimal concentrations of small molecules in the energy module leveraging image analysis. The experimental results on protein synthesis were used to train a ML algorithm and predict high-yield combinations. Remarkably, after two screening rounds, 100000 droplets generated in ~ 4 h with ~25 μl of reagents, the yield of an *E. coli* CFPS system could be increased 1.9-fold with 2.1-fold cost reduction.[60] The result was further validated with the synthesis of 12 different proteins ranging from 27 to 370 kDa. Moreover, minimal experimental input was necessary in the optimisation of a *B. subtilis* CFPS system utilising transfer learning which predicted a 2-fold increase in protein yield.[60] Biofoundries aim to advance synthetic biology applications leveraging automation, HTP screening and computational methods, and CFPS has emerged as a technology to streamline biomanufacturing with automation.[61,62] Recently an automated workflow to optimise the CFPS of two antimicrobial peptides, i. e. colicin M and colicin E1, was presented leveraging large language models (LLMs), such as ChatGPT-4 for code generation and active learning.[63] Through the iterative DBTL cycle two energy modules underlining complex experimental spaces of 6^9 and 10^{12} possible combinations were explored using a ML algorithm coded by ChatGPT-4 that leveraged the cluster margin method to balance uncertainty with diversity and make the labelling

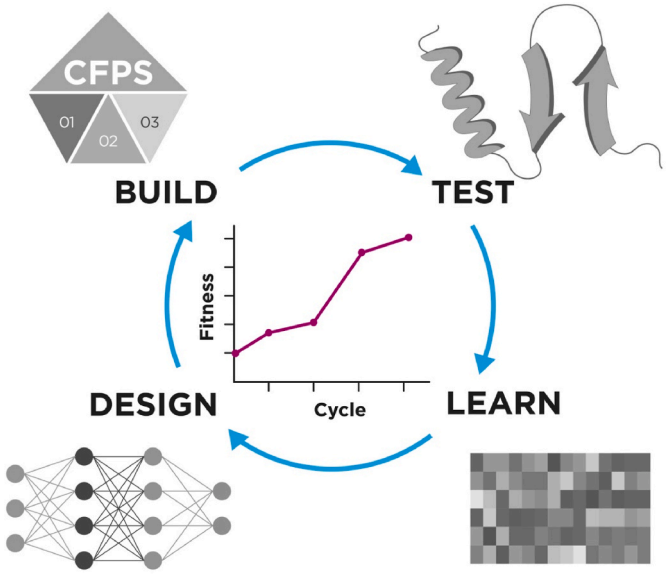

Fig. 1. Design-build-test-learn (DBTL) iterative cycle. The design step consists in a ML alghoritm that defines the experiments to be performed. In the build step, the CFPS reactions are prepared integrating the biochemical modules LED (01 Lysate, 02 Energy, 03 DNA). The test step is the expression of the protein of interest (POI) and the learn one is the data analysis. The four steps of the iterative cycle are repeated for a gradual increase of the fitness function.

process more efficient. Fewer data points were necessary to train the model an achieve 9-fold and 3-fold increase for colicin M and colicin E1 respectively after four iterations. Importantly, all the computational code involved in the automation of the DBTL cycle was generated by ChatGPT-4 reducing the time for programming from months to one week.[63]

4. Machine learning and CFPS for engineering and validation of proteins for drug discovery

CFPS is the biotechnology that assists AI/ML algorithms to engineer proteins and to validate *de novo* protein designs. One of the features of CFPS is reaction miniaturisation. This parallelisation enables the generation of many data points that can guide ML in the exploration of protein fitness landscapes or to make better sequence to function predictions. Moreover, CFPS is an easily accessible platform with short reaction time that offers a testbed for the validation of *de novo* designed proteins. Protein engineering and *de novo* design are important sides of the drug discovery process as they can generate candidates with better properties. For example, bispecific antibodies with superior binding affinity and efficacy could be engineered or designed to improve cancer therapies.[64,65] Moreover, enzymes with improved thermal stability and activity could be used to synthesise different classes of small molecules with pharmaceutical importance.[66] Furthermore, AMPs could be engineered or designed to have a broad-spectrum of antimicrobial activity.[67] The discovery of proteins with new and enhanced properties will be efficiently accelerated with CFPS and AI/ML.

A self-driving autonomous machine for protein landscape exploration (SAMPLE) platform based on Bayesian optimisation, liquid handling robotic and CFPS was presented to engineer glycoside hydrolase family 1 (GH1) enzymes from *Streptomyces* species with improved thermal stability.[68] The Bayesian optimisation deployed a multioutput Gaussian process (GP) trained on 331 cytochrome p450 inactive and 187 active sequences labelling for thermal stability. An upper confidence bound (UCB) algorithm with an active/inactive GP classifier was efficient in balancing exploration and exploitation and converge the optimisation towards functional sequences. The full combinatorial space consisted of 1352 unique GH1 sequences different in 116 mutations on average and by at least 16 mutations. The experimental space was composed of natural sequence fragments, Rosetta-designed fragments and evolution-designed fragments for a total of 34 fragments that were automatically assembled (3-parts one gene or data point) using Golden Gate cloning into a fully functional gene for CFPS using an *E. coli* system.[68] The whole automation process comprising, protein design, gene assembly, PCR, CFPS and thermostability assay required 9 h. The authors deployed four agents seeded with the same six natural GH1 sequences. The agents choose three fragments per round and ran for a total of 20 DBTL cycles. As a result, each agent discovered GH1 sequences 12 °C more stable than the six natural sequences exploring less than 2 % of the full combinatorial space. This result was achieved with minimal experimental effort, the agents converged on the same global fitness peak showing different learning behaviours during the exploration of the sequence to function fitness landscape.[68] A CFPS workflow leveraging expression from linear DNA and predictions from an evaluated machine learning algorithm has engineered amide synthetases to produce efficiently nine small molecule pharmaceuticals.[69] Importantly to gain insights in activity and sequence to function relationships the cell-free workflow has unlocked the preparation of >10000 unique CFPS reactions with results in 24 h. The ML algorithm was designed to prevent overfitting using supervised ridge regression, trained on data labelling with non-zero activity and expanded with zero-shot fitness predictors to foresee the top performing enzymes. Sixty-four residues in the active site of the amide synthetase were selected to engineer the enzyme according to the crystal structure available. With a hot spot screen (HSS) (64 X 19 = 1216 single mutants) four residues were down selected and approximately 80 single mutants were used to train the ML algorithm and to predict the top 25 converting enzymes from an experimental space of 20^4 possible combinations. As a result, three ML predictors had better conversion and six showed activities resulting in improved product yields ranging from 1.6 to 34-fold-increase. The ML model could predict mutations in the complex sequence to function fitness landscape more efficiently than an iterative site saturation mutagenesis campaign solving higher order epistatic interactions.[69] A straightforward CFPS workflow applying ML was devised to improve the efficiency of Con 1 protease.[70] First AplhaFold 3 was applied to model the binding of the substrate to the protease and then Robetta alanin scan was used to determine six residues influencing the substrate binding. With CFPS 48 random protease variants were screened and their activity scored with a FRET-based assay in 6 h. This initial round fed the ML algorithm leveraging active learning-assisted directed evolution (ALDE) that balances model uncertainty in the exploration of the protein sequence to function fitness landscape. The top 32 predicted variants were synthesised, screened and the best performing mutant has shown 4-fold fitness improvement compared to the wild-type protease.[70] Furthermore, four identified protease variants were expressed *in vivo*, purified and tested in six days to demonstrate activity correlation with the candidates obtained from the condensed 6 h CFPS workflow.[70]

A low cost CFPS platform was established for *de novo* design of AMPs using Generative AI.[70] An important group of AMPs contains linear peptides composed of 12–50 canonical amino acids (AAs), therefore assuming a 30 AAs AMP the possible space of combinations is 20^{30}, an immense number of possibilities. The generative AI model created 500 candidates that were screened and tested with a CFPS platform in 24 h. As a result, 30 were found functional and six showed broad-spectrum activity against multi-drug-resistant pathogens without developing bacterial resistance.[71] *De novo* drugs generation was achieved training unsupervised variational autoencoders (VAE) with 1.5 million peptide sequences from UniProt following a transfer learning to a dataset of ~5000 known AMPs to define a latent space for sequences generation. Afterwards, a regressor, based on convolutional neural networks (CNN) trained with ~5000 sequences with known minimum inhibitory concentration (MIC) prioritized 500 AMPs candidates for HTP CFPS and testing.[71] Remarkably, BLAST searching has shown that the *de novo* generated peptides have original sequences and AplhaFold analysis has unveiled structural similarities with training AMPs. Moreover, molecular dynamics (MD) simulations have shown that generally the generated peptides target bacterial membranes and not human plasma membrane.[71] CFPS was applied to validate *de novo* proteins generated by a LLM named ProGen.[72] The model is a 1.2 billion neural network based on Transformer architecture with a self-attention mechanism that weights the probability of each token (amino acid). Protein family labels play a crucial role in fine-tune the transformer-based model including structural information and specific patterns associated with different protein families. Labels (or tags) reduce the space complexity and improve the accuracy of the model. ProGen was trained on a large data base of 280 million naturally evolved proteins. A dataset of ~56000 sequences from five lysozyme families was collected to fine-tune the model and select 100 positive controls, without the ability to capture local sequences neighbourhood, therefore representing an additional challenge for the model. ProGen generated artificial proteins that spanned the landscape of natural proteins from five lysozyme families. One hundred artificial proteins were selected for CFPS validation along with the set of 100 natural proteins as positive control. Results have shown that the artificial proteins could be expressed and purified at the same level of the natural proteins (72 %) indicating that the artificial proteins were folded correctly. Additionally, AlphaFold2 has predicted artificial structures that roughly match structures found in nature including for low identity artificial sequences (<40 %). Importantly, the artificial proteins showed similar catalytic activities as natural lysozymes even for low sequence identity of 31.4 %.[72]

5. Conclusions and perspective

CFPS is a robust easily accessible platform to express several pharmaceuticals. Some examples described in this article include, sdFab,[23] scFVs,[24] ADCs[28] human kinase,[27] antigens,[29,31] VLPs,[30] GPCRs,[36] PBAs[43,71,72] and amide synthetases.[69] Moreover, CFPS platforms can be applied for phage synthesis, engineering[39,40] and as screening platform for antimicrobial compounds.[48] The system compared to traditional *in vivo* protein expression offers an easily accessible HTP platform with reduced timelines. Importantly, CFPS is an open system, therefore protein yields can be maximised while cost decreased optimising molecular compositions with advanced ML algorithms during iterative experimentation.[56,60] In addition, ML and CFPS can guide protein engineering through autonomous agents exploring the protein fitness landscape[68] or with predictions of the sequence to function relationships.[69] Drug discovery campaigns can be initiated with library sizes reduced several orders of magnitude applying deep screening[24] or even to a few hundred[73] or tens[70] with ALDE. CFPS offers a testbed to validate the expression and activity of hundreds *de novo* proteins (artificial proteins) designed with Generative AI[71] and LLM[72]. Looking forward, drug discovery campaigns can be accelerated using CFPS and ML algorithms programmed to cope with large experimental spaces using low data input and fine-tuned for optimal performance. CFPS will unlock the power and benefit of AI leveraging engineering and *de novo* design of proteins important for human health.

Declaration of competing interest

The author declares no conflicts of interest.

Acknowledgments

Javin P. Oza is acknowledged for helpful discussions about cell-free protein synthesis systems and biotechnology applications. This work has not received external funding.

References

1. Silverman AD, Karim AS, Jewett MC. Cell-free gene expression: an expanded repertoire of applications. *Nat Rev Genet*. 2020;21:151–170.
2. Garenne D, et al. Cell-free gene expression. *Nature Reviews Methods Primers*. 2021;1.
3. Gregorio NE, Levine MZ, Oza JP. A user's guide to cell-free protein synthesis. *Methods Protoc*. 2019;2.
4. Cao M, et al. Cell-free protein synthesis system for building synthetic cells. *J Vis Exp*. 2024;10:3791–66626.
5. Hunt AC, et al. Cell-free gene expression: methods and applications. *Chem Rev*. 2024. https://doi.org/10.1021/acs.chemrev.4c00116.
6. Casteleijn MG, et al. Beyond in vivo, pharmaceutical molecule production in cell-free systems and the use of noncanonical amino acids therein. *Chem Rev*. 2025. https://doi.org/10.1021/acs.chemrev.4c00126.
7. Lu Y. Cell-free synthetic biology: engineering in an open world. *Synth Syst Biotechnol*. 2017;2:23–27.
8. Caschera F. Bacterial cell-free expression technology to in vitro systems engineering and optimization. *Synth Syst Biotechnol*. 2017;2:97–104.
9. Brookwell A, Oza JP, Caschera F. Biotechnology applications of cell-free expression systems. *Life*. 2021;11.
10. Sun ZZ, Yeung E, Hayes CA, Noireaux V, Murray RM. Linear DNA for rapid prototyping of synthetic biological circuits in an Escherichia coli based TX-TL cell-free system. *ACS Synth Biol*. 2014;3:387–397.
11. Copeland CE, et al. Expanding the cell-free reporter protein toolbox by employing a split mNeonGreen system to reduce protein synthesis workload. *ACS Synth Biol*. 2024;13:1663–1668.
12. Levrier A, et al. Split reporters facilitate monitoring of gene expression and peptide production in linear cell-free transcription-translation systems. *ACS Synth Biol*. 2024;13:3119–3127.
13. Willi JA, Karim AS, Jewett MC. Cell-free translation quantification via a fluorescent minihelix. *ACS Synth Biol*. 2024;13:2253–2259.
14. Buss JA, Stoddard CD, Beck SV. *Scaling down to scale up – miniaturizing cell-free protein synthesis reactions with the Echo 525 Acoustic Liquid Handler*. New England BioLabs Inc.; 2023. Application note.
15. Buss JA, Lund S, Chen E, et al. *Automated cell-free protein expression and purification for high-throughput Screening using NEBExpress® Cell-free E. coli protein synthesis system and NEBExpress Ni-NTA magnetic beads*. New England BioLabs Inc.; 2024. Application note.
16. Saxena M, Tomas R. *Overcoming discovery challenges for highly disordered transcription factors using eProtein Discovery™*. Nuclera Ltd; 2024. Application note.
17. Guilliam T, Wall B, Shack S, Caschera F. *Customizing expression conditions with additives to attain soluble, active proteins within the eProtein Discovery™ System*. Nuclera Ltd; 2024. Application note.
18. Wall B, Guilliam T, Caschera F. *Screening the effects of solubility tags and their removal (detagging) with 3C protease on cell-free protein expression using the eProtein Discovery™ System*. Nuclera Ltd; 2024. Application note.
19. Wall B, Shack S, Khanna N, Tomas R. *Guiding in vivo protein expression with eProtein Discovery™ CFPS for disulfide-rich proteins*. Nuclera Ltd; 2024. Application note.
20. Gregorio NE, Kao WY, Williams LC, et al. Unlocking applications of cell-free biotechnology through enhanced shelf life and productivity of E. coli extracts. *ACS Synth Biol*. 2020;9:766–778.
21. Zawada JF, et al. Cell-free technologies for biopharmaceutical research and production. *Curr Opin Biotechnol*. 2022;76, 102719.
22. Kim KJL, Lee S-J, Kim D-M. Cell-free systems for a multi-pronged approach to next-generation therapeutics and diagnostics. *Biotechnol Bioproc E*. 2024;29:233–239.
23. Hunt AC, et al. A rapid cell-free expression and screening platform for antibody discovery. *Nat Commun*. 2023;14:3897.
24. Porebski BT, et al. Rapid discovery of high-affinity antibodies via massively parallel sequencing, ribosome display and affinity screening. *Nat Biomed Eng*. 2024;8:214–232.
25. Hong SH, et al. Cell-free protein synthesis from a release factor 1 deficient Escherichia coli activates efficient and multiple site-specific nonstandard amino acid incorporation. *ACS Synth Biol*. 2014;3:398–409.
26. Yin G, et al. RF1 attenuation enables efficient non-natural amino acid incorporation for production of homogeneous antibody drug conjugates. *Sci Rep*. 2017;7:3026.
27. Oza JP, et al. Robust production of recombinant phosphoproteins using cell-free protein synthesis. *Nat Commun*. 2015;6:8168.
28. Hanson J, et al. An integrated in vivo/in vitro protein production platform for site-specific antibody drug conjugates. *Bioengineering (Basel)*. 2023;10.
29. Weiss AM, et al. Site-specific antigen-adjuvant conjugation using cell-free protein synthesis enhances antigen presentation and CD8(+) T-cell response. *Sci Rep*. 2021;11:6267.
30. Lu Y, Chan W, Ko BY, VanLang CC, Swartz JR. Assessing sequence plasticity of a virus-like nanoparticle by evolution toward a versatile scaffold for vaccines and drug delivery. *Proc Natl Acad Sci U S A*. 2015;112:12360–12365.
31. Warfel KF, et al. A low-cost, thermostable, cell-free protein synthesis platform for on-demand production of conjugate vaccines. *ACS Synth Biol*. 2023;12:95–107.
32. Hauser AS, Attwood MM, Rask-Andersen M, Schioth HB, Gloriam DE. Trends in GPCR drug discovery: new agents, targets and indications. *Nat Rev Drug Discov*. 2017;16:829–842.
33. Kaiser L, et al. Efficient cell-free production of olfactory receptors: detergent optimization, structure, and ligand binding analyses. *Proc Natl Acad Sci U S A*. 2008;105:15726–15731.
34. Klammt C, et al. Cell-free production of G protein-coupled receptors for functional and structural studies. *J Struct Biol*. 2007;158:482–493.
35. Zemella A, et al. Qualifying a eukaryotic cell-free system for fluorescence based GPCR analyses. *Sci Rep*. 2017;7:3740.
36. Kock Z, et al. Cryo-EM structure of cell-free synthesized human histamine 2 receptor/G(s) complex in nanodisc environment. *Nat Commun*. 2024;15:1831.
37. Haueis L, et al. Rapid one-step capturing of native, cell-free synthesized and membrane-embedded GLP-1R. *Int J Mol Sci*. 2023;24.
38. Strathdee SA, Hatfull GF, Mutalik VK, Schooley RT. Phage therapy: from biological mechanisms to future directions. *Cell*. 2023;186:17–31.
39. Emslander Q, et al. Cell-free production of personalized therapeutic phages targeting multidrug-resistant bacteria. *Cell Chem Biol*. 2022;29:1434–1445. e1437.
40. Levrier A, et al. PHEIGES: all-cell-free phage synthesis and selection from engineered genomes. *Nat Commun*. 2024;15:2223.
41. Park H, Jin H, Kim D, Lee J. Cell-free systems: ideal platforms for accelerating the discovery and production of peptide-based antibiotics. *Int J Mol Sci*. 2024;25.
42. Arnison PG, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep*. 2013;30:108–160.
43. Liu WQ, et al. Cell-free biosynthesis and engineering of ribosomally synthesized lanthipeptides. *Nat Commun*. 2024;15:4336.
44. Ohuchi M, Murakami H, Suga H. The flexizyme system: a highly flexible tRNA aminoacylation tool for the translation apparatus. *Curr Opin Chem Biol*. 2007;11:537–542.
45. Lee J, et al. Expanding the limits of the second genetic code with ribozymes. *Nat Commun*. 2019;10:5097.
46. Lee J, et al. Ribosome-mediated biosynthesis of pyridazinone oligomers in vitro. *Nat Commun*. 2022;13:6322.
47. Goto Y, Suga H. The RaPID platform for the discovery of pseudo-natural macrocyclic peptides. *Acc Chem Res*. 2021;54:3604–3617.
48. Chengan K, et al. A cell-free strategy for host-specific profiling of intracellular antibiotic sensitivity and resistance. *npj Antimicrobials and Resistance*. 2023;1:16.
49. Stark JC, et al. On-demand biomanufacturing of protective conjugate vaccines. *Sci Adv*. 2021;7.
50. Zhang F, et al. Structural basis of a novel PD-L1 nanobody for immune checkpoint blockade. *Cell Discov*. 2017;3, 17004.
51. Jahnichen S, et al. CXCR4 nanobodies (VHH-based single variable domains) potently inhibit chemotaxis and HIV-1 replication and mobilize stem cells. *Proc Natl Acad Sci U S A*. 2010;107:20565–20570.

52. Zawada JF, et al. Microscale to manufacturing scale-up of cell-free cytokine production–a new approach for shortening protein production development timelines. *Biotechnol Bioeng*. 2011;108:1570–1578.
53. Ebbert LE, Free TJ, Soltani M, Bundy BC. The design and cell-free protein synthesis of a pembrolizumab single-chain variable fragment. *Drugs and Drug Candidates*. 2025;4:3.
54. Gupta MD, et al. Scaling eukaryotic cell-free protein synthesis achieved with the versatile and high-yielding tobacco BY-2 cell lysate. *Biotechnol Bioeng*. 2023;120: 2890–2906.
55. Dopp JL, Tamiev DD, Reuel NF. Cell-free supplement mixtures: elucidating the history and biochemical utility of additives used to support in vitro protein synthesis in E. coli extract. *Biotechnol Adv*. 2019;37:246–258.
56. Caschera F, et al. High-throughput optimization cycle of a cell-free ribosome assembly and protein synthesis system. *ACS Synth Biol*. 2018;7:2841–2853.
57. Caschera F, et al. Coping with complexity: machine learning optimization of cell-free protein synthesis. *Biotechnol Bioeng*. 2011;108:2218–2228.
58. Borkowski O, et al. Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat Commun*. 2020;11:1872.
59. Morini L, et al. Leveraging active learning to establish efficient in vitro transcription and translation from bacterial chromosomal DNA. *ACS Omega*. 2024;9: 19227–19235.
60. Zhu J, et al. *AI-Driven High-Throughput Droplet Screening of Cell-free Gene Expression*. 2024. bioRxiv 10.1101/2024.07.04.602084, 2024.2007.2004.602084.
61. Romantseva E, Piorino Macruz de Oliveira F, Sundberg C, Sittampalam GS, Strychalski E. *Advancing Cell-free Manufacturing: Challenges in Scale-Up and Automation Workshop Report*. Special Publication (NIST SP); 2024:1500–1526.
62. Hillson N, et al. Building a global alliance of biofoundries. *Nat Commun*. 2019;10: 2040.
63. Hérisson J, Hoang AN, El-Sawah A, Khalil MM, Faulon J-L. *Operate a Cell-free Biofoundry Using Large Language Models*. 2024. bioRxiv 10.1101/ 2024.10.28.619828, 2024.2010.2028.619828.
64. Klein C, Brinkmann U, Reichert JM, Kontermann RE. The present and future of bispecific antibodies for cancer therapy. *Nat Rev Drug Discov*. 2024;23:301–319.
65. Xu Y, et al. Production of bispecific antibodies in "knobs-into-holes" using a cell-free expression system. *mAbs*. 2015;7:231–242.
66. Moura M, et al. Evaluating enzymatic synthesis of small molecule drugs. *Metab Eng*. 2016;33:138–147.
67. Rathinakumar R, Wimley WC. High-throughput discovery of broad-spectrum peptide antibiotics. *FASEB J*. 2010;24:3232–3238.
68. Rapp JT, Bremer BJ, Romero PA. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat Chem Eng*. 2024;1:97–107.
69. Landwehr GM, et al. *Accelerated Enzyme Engineering by Machine-Learning Guided Cell-free Expression*. 2024. bioRxiv 10.1101/2024.07.30.605672, 2024.2007.2030.605672.
70. Thornton EL, Boyle JT, Laohakunakorn N, Regan L. *Cell-free Protein Synthesis as a Method to Rapidly Screen Machine Learning-Directed Protease Variants*. 2025. bioRxiv 10.1101/2025.01.24.634768, 2025.2001.2024.634768.
71. Pandi A, et al. Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides. *Nat Commun*. 2023;14:7197.
72. Madani A, et al. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol*. 2023;41:1099–1106.
73. Yang J, et al. *Active Learning-Assisted Directed Evolution*. 2024. bioRxiv 10.1101/ 2024.07.27.605457, 2024.2007.2027.605457.