RSC
Chemical Biology

ROYAL SOCIETY
OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Cite this: *RSC Chem. Biol.*, 2023, 4, 271

A primer to directed evolution: current methodologies and future directions

Lara Sellés Vidal, †*[ab] Mark Isalan, [ac] John T. Heap [acd] and Rodrigo Ledesma-Amaro *[ab]

Received 21st November 2022,
Accepted 18th January 2023

DOI: 10.1039/d2cb00231k

rsc.li/rsc-chembio

Directed evolution is one of the most powerful tools for protein engineering and functions by harnessing natural evolution, but on a shorter timescale. It enables the rapid selection of variants of biomolecules with properties that make them more suitable for specific applications. Since the first *in vitro* evolution experiments performed by Sol Spiegelman in 1967, a wide range of techniques have been developed to tackle the main two steps of directed evolution: genetic diversification (library generation), and isolation of the variants of interest. This review covers the main modern methodologies, discussing the advantages and drawbacks of each, and hence the considerations for designing directed evolution experiments. Furthermore, the most recent developments are discussed, showing how advances in the handling of ever larger library sizes are enabling new research questions to be tackled.

Introduction

Enzymes have attracted increasing interest from industry as more efficient and less costly alternatives to other synthetic chemistry tools, such as transition metal catalysts or organocatalysts.[1] However, while the repertoire in nature provides a vast variety of biocatalysts, it is frequently the case that natural enzymes need to be tailored by protein engineering in order to maximise their performance for specific applications. The same applies to biomolecules performing other functions, such as binding partners (including antibodies), fluorescent or bioluminescent macromolecules and biocatalysts acting on other macromolecules.

Two main approaches can be taken to carry out protein engineering: rational design and directed evolution. Rational design involves performing chosen point mutations, insertions or deletions in the coding sequence, and mutation choice is typically based on structural and functional information about the target biomolecule. Nonetheless, the sequence–structure–function relationship is often difficult to predict accurately, particularly at the single residue level. Additionally, reliable structural information is frequently not available for the protein of interest and, while progress is being made in methods for protein structure prediction thanks to artificial intelligence,[2] these still remain rather limited, especially for larger proteins and macromolecular complexes.[3] Often therefore, rationally designed mutations do not have the desired effect.

Directed evolution, on the other hand, bypasses the need to determine specific mutations *a priori* by mimicking the process of natural evolution in the laboratory.[4,5] In nature, mutations which are beneficial for individuals are iteratively selected through numerous generations. In directed evolution, this process takes place on a much shorter timescale, and generates biomolecules that suit human-defined applications.

The first *in vitro* evolution experiments can be traced back to the 1960s. In a pioneering Darwinian experiment, Sol Spiegelman *et al.* iteratively selected RNA molecules based on their ability to be replicated by Q bacteriophage RNA polymerase.[6] Over the next two decades, such *in vitro* evolution experiments shifted towards more application-driven approaches, which is exemplified by the development of phage display.[7] In phage display, an exogenous sequence is fused to a gene encoding a minor coat protein of a filamentous phage, leading the assembled viral particles to display the extra amino acids. A set of phages with different fused peptides could then be subjected to affinity purification, against desired binding partners, to obtain variants with high affinity towards them.

During the last 30 years, directed evolution approaches have diversified and shifted their focus towards more complex and varied properties and biomolecules (Fig. 1). This can be easily visualized by analyzing the frequency of keywords associated with directed evolution papers (Fig. 2). Our analysis suggests that from the early days of directed evolution and until the mid 2000s, directed evolution focused mostly on altering binding

[a] *Imperial College Centre for Synthetic Biology, Imperial College London, London, SW7 2AZ, UK. E-mail: r.ledesma-amaro@imperial.ac.uk*
[b] *Department of Bioengineering, Imperial College London, London, SW7 2AZ, UK*
[c] *Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK*
[d] *School of Life Sciences, The University of Nottingham, University Park, Nottingham, NG7 2RD, UK*
† Current address: Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan. E-mail: lara.selles@oist.jp

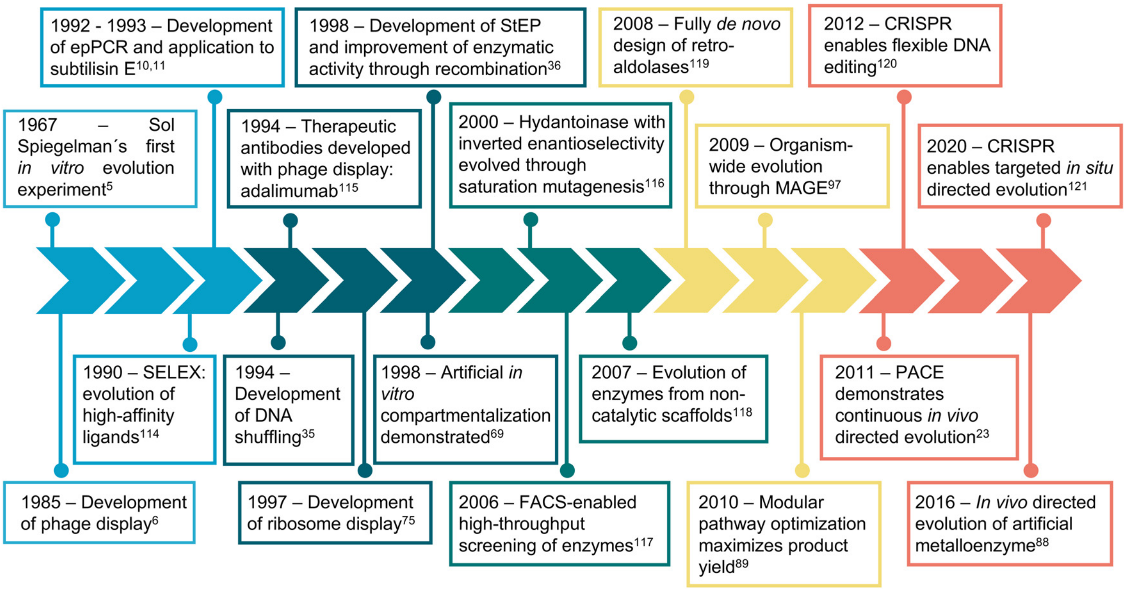
1967 – Sol Spiegelman´s first *in vitro* evolution experiment[5]
1985 – Development of phage display[6]
1990 – SELEX: evolution of high-affinity ligands[114]
1992 - 1993 – Development of epPCR and application to subtilisin E[10,11]
1994 – Therapeutic antibodies developed with phage display: adalimumab[115]
1994 – Development of DNA shuffling[35]
1997 – Development of ribosome display[75]
1998 – Development of StEP and improvement of enzymatic activity through recombination[36]
1998 – Artificial *in vitro* compartmentalization demonstrated[69]
2000 – Hydantoinase with inverted enantioselectivity evolved through saturation mutagenesis[116]
2006 – FACS-enabled high-throughput screening of enzymes[117]
2007 – Evolution of enzymes from non-catalytic scaffolds[118]
2008 – Fully *de novo* design of retro-aldolases[119]
2009 – Organism-wide evolution through MAGE[97]
2010 – Modular pathway optimization maximizes product yield[89]
2011 – PACE demonstrates continuous *in vivo* directed evolution[23]
2012 – CRISPR enables flexible DNA editing[120]
2016 – *In vivo* directed evolution of artificial metalloenzyme[88]
2020 – CRISPR enables targeted *in situ* directed evolution[121]

Fig. 1 Timeline of directed evolution major developments. Representative examples of some of the most relevant developments and achievements in directed evolution are shown.

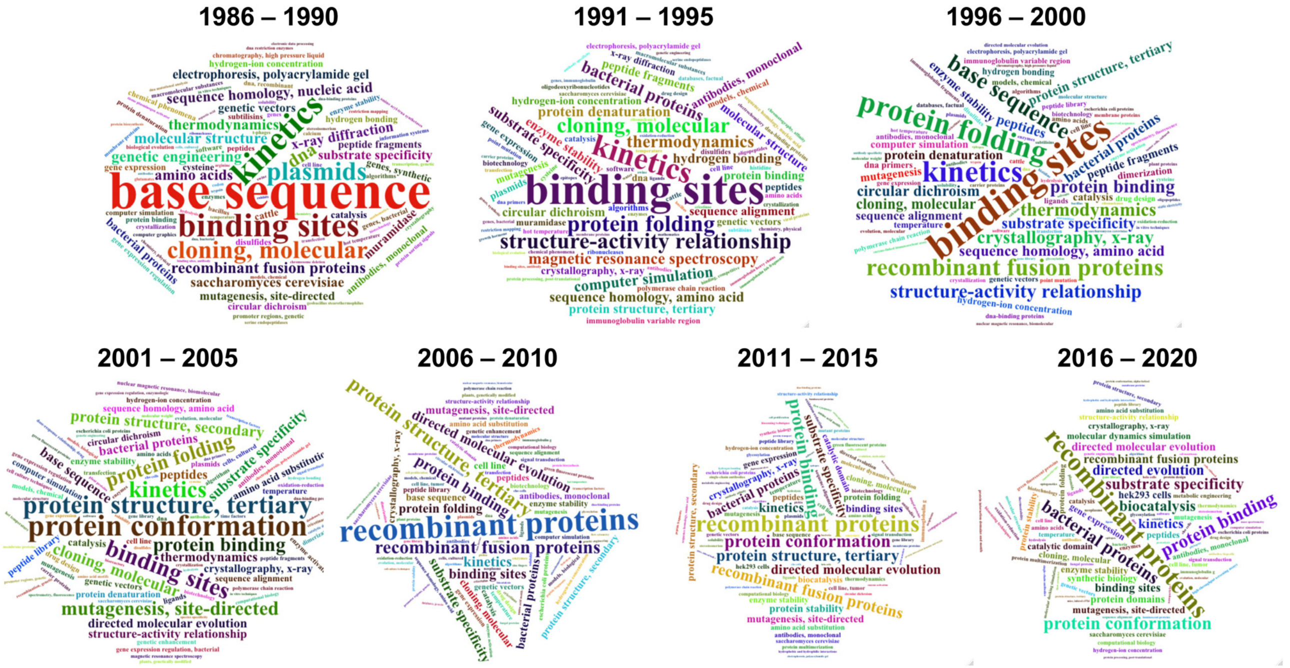

Fig. 2 Evolution of keywords associated to directed evolution articles. Wordclouds of the keywords associated to directed evolution articles since 1985 are shown. Articles were grouped in blocks of five years. The wordclouds reveal a change over time of the focus and scope of directed evolution, with the most recent decade showing a larger variety of targeted properties and biomolecules. Data was retrieved, analysed and visualised with R.

sites and improving enzyme kinetic parameters, with a special emphasis on the analysis of nucleotide and amino acid sequences when choosing the approach to take. At the beginning of the 21st century, protein structure and conformation became a major point of interest in directed evolution studies, probably due to the increasing availability of macromolecular structures as a consequence of the improvement and accessibility of structural biology techniques. Indeed, in the period 1995–2000, ‘‘structure–activity relationship’’ was one of the top keywords found in directed evolution papers. Such keywords ranked considerably lower from 2005, probably due to the realization that the mechanisms through which structure

Table 1 Summary of techniques frequently applied in directed evolution

	Technique	Purpose	Advantages	Disadvantages	Application examples
Mutagenesis	Error-prone PCR and error-prone RCA	Insertion of point mutations across whole sequence	• Easy to perform • Does not require prior knowledge about key positions	• Reduced sampling of mutagenesis space • Mutagenesis bias	Subtilisin E[12] Glycolyl-CoA carboxylase[124]
	RAISE	Insertion of random short insertions and deletions	• Enables random indels across sequence	• Indels limited to few nucleotides • Frameshifts introduced	β-Lactamase[17]
	TRINS	Insertion of random tandem repeats	• Mimics duplications that occur in natural evolution	• Frameshifts introduced	β-Lactamase[18]
	Mini-mu based techniques	Random insertion or deletion of one or multiple codons	• Conservation of reading frame • Modification of transposon enables customization of mutations	• Additional steps of DNA manipulation required	Arylesterase[18] GFP[125]
	Mutator strains	*In vivo* random mutagenesis	• Simple system	• Biased and uncontrolled mutagenesis spectrum • Mutagenesis not restricted to target	Vitamin K epoxide reductase[126] Cells resistant to DMF[127]
	Orthogonal systems based on DNA Pol I, pGLK1/2, Ty1, T7RNAP and CRISPR	*In vivo* random mutagenesis	• Mutagenesis restricted to target sequence • Could be coupled to *in vivo* selection	• Mutation frequency relatively low • Limitations on size of target sequence	β-Lactamase[28] Dihydrofolate reductase[128] Orotidine-5′-phosphate decarboxylase[31]
	DNA shuffling	Random sequence recombination	• Recombination advantages	• High homology between parental sequences required	Thymidine kinase[129] Non-canonical esterase[130]
	StEP	Random sequence recombination	• Recombination advantages • Easy to perform	• High homology between parental sequences required	Photoswitchable fluorescent protein[131] Synthetic antibodies[39]
	RACHITT	Random sequence recombination	• Increased crossover frequency • Parental sequences removed from library	• High homology between parental sequences required	Monooxygenase[40] DNA polymerase[132]
Mutagenesis	ITCHY and SCRATCHY	Random recombination of any two sequences	• No homology between sequences required	• Gene length and reading frame not preserved • Recombination at sites not structurally related • Single crossover per variant (solved in SCRATCHY)	Alcohol dehydrogenase[41] Deoxyribonucleoside kinase[133]
	SHIPREC	Random recombination of any two sequences	• No homology between sequences required • Crossovers at structurally-related sites	• Single crossover per variant • Reading frame not preserved	Cytochrome P450[42]
	SISDC	Recombination of a few sequences at specific sites	• No homology between sequences required • Crossover points can be chosen	• Limited number of potential crossover points • Additional steps of DNA manipulation required	β-Lactamase[44]
	MORPHING	*In vivo* generation of recombination libraries	• Can be coupled to *in vivo* selection techniques	• Crossovers only occur at overlapping regions	Peroxidase[45] Aryl alcohol oxidase[134]
	Site-saturation mutagenesis	Focused mutagenesis of specific positions	• In-depth exploration of mutagenesis at chosen positions • Possibility to incorporate previous information for efficient mutagenesis	• Only a few positions mutated • Libraries can easily become very large	Widely applied to enzyme evolution[56,57]

Table 1 (*continued*)

	Technique	Purpose	Advantages	Disadvantages	Application examples
			• Iterative cycles and smart libraries can reduce library sizes		
	StLois	Sequential extension of loops	• Insertions performed at sites less likely to result in non-functional variants • Reasonable library sizes due to sequential extension of loops	• Limited number of insertions in each extension cycle	Cumene dioxygenase[64]
Identification of variants	Colorimetric/fluorimetric analysis of colonies/cultures	Screening of variants	• Fast and easy to perform	• Limited to biomolecules exhibiting appropriate spectral properties	Fluorescent proteins[57,58]
	Plate-based automated enzymatic assays	Screening of variants	• Automation has increased throughput • Surrogate substrates expand scope • Coupling to GC/HPLC enables analysis of enantiomers	• Throughput remains limited compared to other methods, especially if substrate or product do not have characteristic spectral or fluorescent properties • Results with surrogate substrates do not always replicate with original ones	Lipase[135] Laccase[136]
	FACS-based methods	Screening of variants	• High throughput • Product entrapment expands application scope • Similar techniques can be applied with *in vitro* compartmentalization	• Evolved property must be linked to a change in fluorescence	Sortase[81] Cre recombinase[82] β-galactosidase[83]
	MS-based methods	Screening of variants	• High throughput • Does not rely on specific properties of substrates	• Less widely-available equipment required • For MALDI-based methods, requirement of immobilization on matrix	Fatty acid synthase[77] Cytochrome P411[78] Cyclodipeptide synthase[79]
	Display techniques	Selection of variants	• High throughput	• Limited to selection of biomolecules with specific binding properties	Antibodies[87] Fbs1 glycan-binding protein[90] RNA-binding peptides[137] Random sequence ATP-binding proteins[138]
	QUEST	Selection of variants	• High throughput	• Limited scope due to substrate/ligand constraints	Scytalone dehydratase[101] Arabinose isomerase[139]
	Cofactor regeneration coupling	Selection of variants	• High throughput • Applicable to wide range of small molecule biocatalysts and properties	• An indirect link to NAD-related activities must be established	Alcohol dehydrogenase[109] Imine reductase[109] Nitrorreductase[109] Isopropanol pathway[109]
	In vitro compartmentalized self-replication	Selection of variants	• High throughput • Bypasses library transformation	• Limited to activities that can be linked to replication or transcription of its coding sequence	DNA polymerase[104]

determines function in biological macromolecules cannot be easily unraveled. During the last decade, the variety of targeted properties has quickly increased (including a larger proportion of studies aiming to alter substrate specificity). The range of organisms employed in such studies has also been expanded, as demonstrated by the presence of keywords such as ''HEK293 cells'' or ''cell line''. Interestingly, a persistent interest in recombinant proteins seems to have been sustained during the same period.

The large expansion of the scope of directed evolution is tightly linked to the ample variety of developed techniques that allow researchers to tackle more efficiently the two main steps of the process of artificial evolution of biomolecules (Table 1). The first step consists in generating enough genetic diversity in a given parental sequence to cover the sequence-function space to be sampled. The resulting set of sequences, or library, often includes a majority of variants without the desired property or improvement. In a second step, the individual genotypes must be linked to the individual phenotypes[8] to allow the variants of interest to be identified and isolated from the library. In this review, we aim to provide an overview of the different available methodologies, presenting the underlying principles, advantages and disadvantages of each one in order to facilitate readers to make an informed choice when determining the most appropriate techniques for a particular application of directed evolution.

Generation of a library of variants

The rate of natural mutation is usually insufficient for generating the genetic diversity required for laboratory directed evolution. For example, the mutation rate of wild-type *E. coli* is approximately 1×10^{-3} mutations per genome per generation, or 2.2×10^{-10} mutations per base pair per generation.[9] With such a low mutation rate, over 100 000 generations would be required on average to obtain a single point mutation in a target gene of 1000 base pairs.[10] Therefore, it is necessary to artificially enhance genetic diversification to increase the sampling of mutations. A plethora of techniques to achieve this are now available, each having their own advantages and drawbacks. They can be broadly classified into random and rational mutagenesis, although it is often found that a given technique combines aspects from both types of approaches.

Random mutagenesis

In random mutagenesis approaches, no specific sequence positions are targeted. Such techniques are particularly useful for directed evolution of proteins for which there is not enough structure–function information available to determine which residues to diversify, or when the property to be evolved cannot be easily attributed to a few specific positions. For instance, enhanced stability under extreme conditions such as high temperature or organic solvents.

***In vitro* techniques.** Some of the earliest mutagenesis techniques were based on the use of chemical or physical mutagens to introduce random mutations in the sequence of interest[11] (Fig. 3a). However, these have been superseded by other methods for directed evolution due to their strong bias towards certain mutations, and the need to perform further manipulations to obtain double-stranded mutagenised DNA.

Error-prone PCR (epPCR)[12] is one of the most widely used techniques for random mutagenesis (Fig. 3b). In epPCR, a PCR is performed with low-fidelity polymerases (such as Taq polymerase) and altered reaction conditions (including Mn^{2+} and increased Mg^{2+} ions, and using unequal concentrations of the different deoxyribonucleotides) to increase the error frequency. In a landmark study, Chen and Arnold demonstrated the potential of the technique by alternating rounds of epPCR and screening to obtain a variant of subtilisin E enzyme, with increased activity in the presence of 60% dimethylformamide.[13]

Approaches based on epPCR require fine-tuning of the mutation rate, which must be neither too high nor too low (this varies by application, but a rule of thumb is 1 mutation per kb per generation). Achieving an optimal mutation rate is crucial to obtain a library with a good density of functional variants, since an excessive amount of mutations can easily lead to a prevalence of deleterious effects, depleting the library from variants of interest. Approaches to alleviate the load of non-functional variants have been devised, such as the construction of neutral drift libraries.[14] In such methodology, variants are pre-screened to identify those that maintain wild-type activity, which are then pooled to create a library of variants without deleterious effects, used as the starting point for additional rounds of selection for the actual target property. In addition, transformation into a host strain is often the limiting factor in library size, leading to the loss of some variants. It can therefore be advantageous to amplify the library prior to transformation, for example using plasmid rolling circle amplification (RCA)[15] (Fig. 3c). Thus, in 2004, a simple protocol for error-prone RCA (epRCA) was developed, where the mutagenesis rate is enhanced through the addition of $MnCl_2$ to the reaction, and by reducing the concentration of the DNA template.[16]

Both epPCR and epRCA only introduce substitutions in the sequences subjected to mutagenesis. However, insertions and deletions (indels) also play an important role in natural genetic diversification, since they can alter the backbone of the encoded proteins in a way not achievable simply by point mutation.[17,18] Several methodologies have been developed for introducing indels into gene libraries. Initial efforts focused on the introduction of random insertions and deletions, such as the RAISE method, where 3′-terminal deoxynucleotidyl transferase introduces random extensions[19] (Fig. 3d).

Insertions by duplication are another of the most frequent naturally occurring indels, with some estimates placing their frequency as high as two thirds of all insertions observed in natural genomes.[20] TRINS was developed as a method to mimic such insertions *in vitro*, resulting in the formation of tandem repeats of segments of the sequence of interest through assembly PCR with random linear and circularised fragments[20] (Fig. 3e).

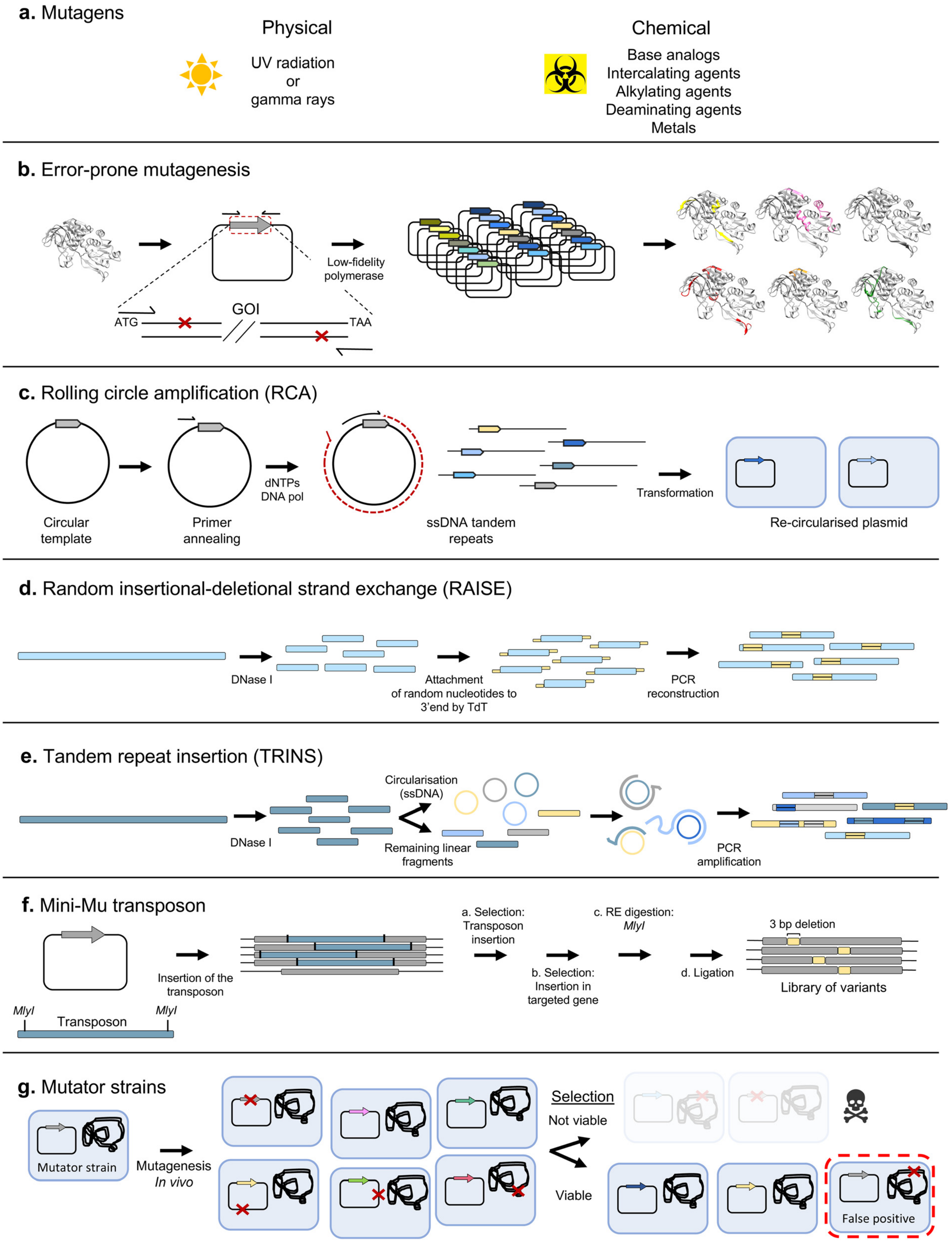

Fig. 3 Random mutagenesis techniques. Some of the most widely applied random mutagenesis approaches are represented. (a) Chemical and physical mutagens were the basis of many genome-wide screening experiments, but have not been extensively used in directed evolution. (b) In error-prone PCR, random mutations are introduced by a low fidelity polymerase, resulting in linear DNA fragments with point mutations, which generally require further manipulation to be inserted into an appropriate vector. (c) RCA based methods bypass the need for further manipulation, since the products can be automatically recircularized by the host cells. (d) RAISE was one of the first methodologies designed to introduce random insertions in the sequence to be mutated. Random small extensions are attached to digested fragments of the parental sequence, and full-length genes are reconstructed through PCR.

(e) TRINS aims to generate repeats of random short fragments of the parental sequence. First, the parental sequence is digested with DNase I. Part of the obtained fragments are circularized and mixed with the remaining linear fragments. An assembly PCR reaction is then performed. When a linear fragment anneals with a circular fragment, a reaction similar to RCA takes place, leading to the replication of multiple copies of the region corresponding to the circularized fragment. (f) Several mutagenesis techniques based on the mini-Mu transposon have been devised. Such techniques allow the generation of random insertions or even deletions while preserving the appropriate reading frame. (g) Mutator strains were the first tool enabling *in vivo* random mutagenesis. However, the increased mutagenesis rate applies to the whole genetic material, and not only to the sequence of interest. More sophisticated approaches where only the gene of interest is targeted have been developed. GOI: gene of interest.

Nevertheless, such techniques have the drawback of generating variants with frameshifts, which usually lead to loss of function. More refined indels can be achieved with modified versions of the mini-Mu transposon, a transposable element which can be easily inserted at random points of a DNA sequence by treatment with the MuA transposase.[21] Jones modified the mini-Mu transposon to allow the removal of single codons at insertion sites after treatment with MlyI and ligation[22] (Fig. 3f).

TRIAD uses a modified Mu protocol to insert or delete up to 3 codons.[18] In it, an engineered transposon is inserted at random points, and then removed through treatment with MlyI. Multiple cassettes can then be inserted through ligation at the generated restriction sites. The deletion cassettes are designed such that their removal with a final treatment with the appropriate restriction enzyme also leads to the deletion of a fixed number of adjacent nucleotides. On the other hand, insertion cassettes include degenerate NNN codons, which remain in the sequence of interest even after removal of the rest of the cassette. More recently, a DNA assembly platform to introduce customizable insertions has been devised,[23] based on cycles of endonuclease treatment and ligation of DNA sequences with the desired inserts.

***In vivo* techniques.** While *in vitro* mutagenesis techniques enable a relatively good control over the mutation rate and spectrum, they also suffer from inherent bottlenecks, such as the limitation on the library size imposed by the transformation efficiency and the requirement of manipulating the genetic material. *In vivo* approaches try to bypass these limitations by taking advantage of the cell machinery to directly perform mutagenesis within the host cells. This also allows the coupling of mutation and screening or selection cycles, enabling more efficient automated workflows, including modern continuous directed evolution[24] (Fig. 8).

The first *in vivo* mutagenesis systems were mutator strains, such as the XL1-Red strain,[25] where inactivation of DNA repair pathways results in increased mutation rates (Fig. 3g). While such hypermutator strains provide a simple *in vivo* mutagenesis system, they present major drawbacks, including a relatively low and hard-to-control mutagenesis rate, biased mutagenesis spectrum, and the indiscriminate mutagenesis of the genome and other sequences outside the gene of interest. This causes a series of undesirable side effects, such as slower growth, reduced transformation efficiencies, and loss of successful variants. More recent random mutagenesis approaches have improved this by having inducible mutator plasmids of different strengths[26] and have overcome genome background mutagenesis by using evolving phage that continuously infect virgin cells containing mutator plasmids.[27]

Despite these advances, the 'Holy Grail' in directed evolution is to perform targeted mutagenesis of the gene of interest. An early example is the two-plasmid system based on error-prone DNA polymerase I (Pol I), where Pol I introduces mutations in sequences of up to 3000 bp.[28] A similar system for *in vivo* mutagenesis in yeast has also been devised, termed OrthoRep, based on the pGKL1/2 plasmid system of *Kluveromyces lactis*, where mutations are introduced by error-prone terminal-protein DNA polymerase 1.[29,30] Other *in vivo* systems have expanded the idea to use enzymes other than DNA-dependent DNA polymerases as the main source of mutagenesis. An example is the Ty1 retrotransposon-based *in vivo* mutagenesis system for yeasts, based on the high error propensity of its encoded retrotranscriptase,[31] or the MutaT7 system, based on chimeric protein fusing T7 RNA polymerase and a nucleobase deaminase, responsible for the introduction of mutations.[32]

More recently, there have been advances in targeted mutation based on CRISPR technology, such a EvolvR, which allows targeted mutation rates that are up to ∼8-million-fold greater than rates seen in wild-type cells, and editing windows with lengths of up to 350 nucleotides.[33] These techniques promise a step-change for *in vivo* directed evolution. Nevertheless, despite these recent advances, it should be noted that *in vivo* mutagenesis is in practice rarely applied to the evolution of biocatalysts acting on small molecule substrates. This is due to the fact that there is not yet a generalized method to assess the fitness of the generated variants *in vivo*, although instances of successful directed evolution of novel enzymes acting on small-molecule substrates through *in vivo* mutagenesis coupled to selection exist, as later discussed. The lack of such a general method that enables the coupling of the different properties of a wide range of substrates or products to an increased survival rate or an easily screenable indicator, makes *in vitro* and rational mutagenesis the current preferred choice for the evolution of enzymes acting on small molecules.

Recombination techniques

Recombination is one of the most powerful mechanisms in nature to generate the genetic diversification required for evolution.[34,35] Correspondingly, a large variety of recombination-based techniques have been developed for directed evolution. They differ greatly from random mutagenesis techniques in the fact that they produce libraries of combinatorial variants where segments from several functional biomolecules are combined. In principle, libraries generated through recombination methods possess a higher percentage of functional variants, avoid the

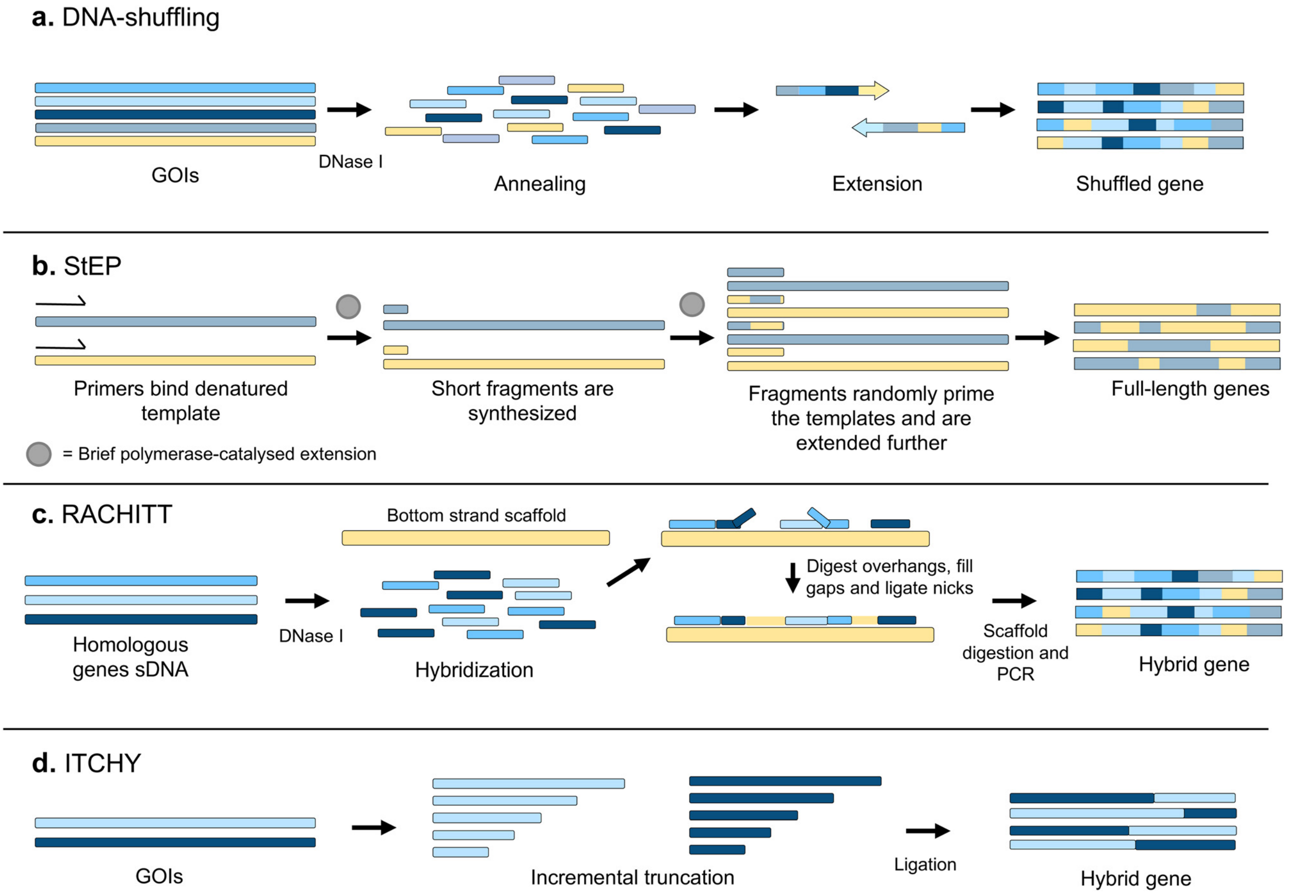
a. DNA-shuffling
GOIs
DNase I
Annealing
Extension
Shuffled gene
b. StEP
Primers bind denatured template
Short fragments are synthesized
Fragments randomly prime the templates and are extended further
Full-length genes
● = Brief polymerase-catalysed extension
c. RACHITT
Homologous genes sDNA
DNase I
Bottom strand scaffold
Hybridization
Digest overhangs, fill gaps and ligate nicks
Scaffold digestion and PCR
Hybrid gene
d. ITCHY
GOIs
Incremental truncation
Ligation
Hybrid gene

Fig. 4 Recombination-based mutagenesis techniques. A set of some of the main mutagenesis techniques based on recombination is displayed. (a) DNA shuffling was the first recombination-based *in vitro* mutagenesis technique to be developed. A set of homologous sequences is treated with DNase I, and the resulting mix of fragments is used to reassemble full-length sequences through self-priming PCR. (b) In StEP, a set of homologous sequences is used as templates for a series of cycles of annealing with primers and extension of the primers by a DNA polymerase. In each cycle, the growing primers can anneal with a different template, resulting in chimeric full-length sequences. (c) RACHITT requires less homology between parental sequences than DNA shuffling and StEP. A set of fragments obtained by treatment with DNase I of the complementary strands of the sequences to be recombined is hybridized to a single-stranded copy of one of the parental sequences. After digesting overhangs, filling the gaps and ligating the nicks, the scaffold strands are digested, and double-stranded chimeric sequences are obtained by means of PCR. (d) ITCHY decreases even further the sequence homology requirements, but it is limited to a single crossover per variant. Exonuclease III is used to incrementally truncate one of the parental genes from its 3′ end, and the other one from its 5′ end. Then, random-length fragments of each gene are ligated. GOI: gene of interest.

introduction of stop codons, and can lead to the discovery of epistatic mutations.[36]

The first recombination-based technique for library generation was DNA shuffling.[37] A pool of closely related sequences is fragmented by DNase I treatment, and full-length sequences are reassembled through self-priming PCR, combining different parental sequences (Fig. 4a). Alternatively, StEP (Staggered Expansion Process) provides a technically simpler approach where the need to fragment the parental sequences with DNase I is bypassed.[38] StEP consists of repeated cycles of denaturation of the templates and very short annealing and extension by DNA polymerase steps, performed until combined full-length products are obtained (Fig. 4b). Typically, StEP is applied to a pool of sequences generated by another mutagenesis technique, to further increase genetic variability and has proven to be useful for evolving synthetic antibodies with improved affinities.[39]

However, both DNA shuffling and StEP require a high degree of homology between the different parental sequences to be recombined. This makes them unsuitable for the recombination of distantly related sequences, which would be very useful for recombining homologous enzymes, for example. Several strategies have therefore been developed to reduce the required degree of homology.

RACHITT (Random Chimeragenesis on Transient Templates) employs a single strand of one of the sequences to be recombined as a scaffolding template, to which fragments of the other sequences are hybridised and ligated[40] (Fig. 4c). Others went even further, seeking to develop techniques that completely removed any requirement for homology between the sequences. The first methodologies allowed only two genes to be recombined, such as ITCHY (Incremental Truncation for the Creation of Hybrid enzymes, Fig. 4d)[41] and SHIPREC (Sequence Homology-Independent Protein Recombination),[42] both of which combine 3′ and 5′ fragments of two genes. Further variation can be introduced with the SCRATCHY protocol,[43] where variants from an ITCHY library are further recombined by DNA shuffling. Other methods, such as SISDC (Sequence Independent Site Directed Chimeragenesis),[44] bypass the need for sequence homology by pre-establishing the crossover points in the parental sequences.

More recently, the focus of novel recombination methods has shifted towards the *in vivo* generation of libraries, allowing a direct coupling to selection techniques. One of the first such approaches was MORPHING (Mutagenic Organized Recombination Process by Homologous *in vivo* Grouping), which takes advantage of the naturally high recombination frequency of *Saccharomyces cerevisiae* to assemble full-length variants of a gene from sets of overlapping fragments.[45]

Rational mutagenesis

In contrast to random mutagenesis approaches, rational mutagenesis focuses on mutating only a limited number of positions in the target sequence, which must be determined based on prior knowledge. The latter includes structures, multiple sequence alignments, biochemical data and computer-based predictions.

Site-saturation mutagenesis (SSM) enables fast and efficient generation of libraries where all possible substitutions of the chosen positions are included. Several different strategies have been developed to obtain libraries with saturated positions, but most rely on the use of mutagenic primers that randomise chosen positions in PCR (typically using the base N in oligonucleotides, where N is an equimolar mixture of A, C, G and T). One of the first developed protocols was overlap extension PCR (OE-PCR),[46] where two sequential steps of amplification are performed. In the first step, overlapping primers, degenerate on the position(s) to be mutated, result in two overlapping segments of sequence containing the mutation(s). The full-length sequence is reconstituted in a second amplification step where primers annealing with both ends of the sequence are used (Fig. 5).

While this protocol can be easily used to saturate one or a few proximal codons (close enough to be included in one or two primers), more specialised strategies are required in order to generate libraries where several distant sites are randomised. One possibility is to use several pairs of mutagenic degenerate primers covering the full sequence, such that both the reverse primer for a given segment and the forward primer for the next segment overlap partially and cover the same degenerate codon. The resulting products are separated by agarose (MOE-PCR) or polyacrylamide gels (POEP), and full-length sequences are assembled by means of a second PCR with flanking primers.[47,48] It is also possible to anneal all mutagenic primers simultaneously to the parental sequence, extend them with T4 DNA polymerase and perform a PCR to obtain the library. However, this can be difficult to achieve if the thermodynamic parameters of each set of primers differ. OmniChange was developed as an alternative to saturate up to five independent codons in a sequence-independent manner by using primers with phosphorothiodiester bonds.[49]

In addition to the mutagenesis protocol, another key parameter that needs to be decided before undertaking SSM is the degeneracy scheme. It is possible to vary the target codons to NNN (where N represents A, G, T or C) to obtain libraries where all possible 64 triplets are present. However, since some amino acids are encoded by more possible codons than others, not all substitutions become equally frequent at the protein level. Furthermore, premature stop codons are introduced, leading to a background of variants with loss of function. Moreover, if a large number of codons is targeted, the library size can easily become excessive.

An NNK scheme (where K represents either G or T) reduces the library size by half and includes only one possible stop codon instead of three, while still allowing for all 20 amino acids to be encoded in the resulting triplets.[50] More restrictive degeneration strategies can be implemented to completely remove stop codons and balance better the representation of amino acids with different chemical properties, albeit at the cost of not encoding all 20 amino acids[50,51] (Table 1). A potential workaround is the usage of mixtures of primers with restricted degeneration schemes that encode, as an ensemble, all possible substitutions with only one possible codon per amino acid. This allows the generation of "small-intelligent" libraries, where stop codons and rare codons can be completely removed. DC-Analyzer was developed as a software tool to assist in the design of such libraries, with the possibility of generating libraries including only polar or hydrophobic residues.[52]

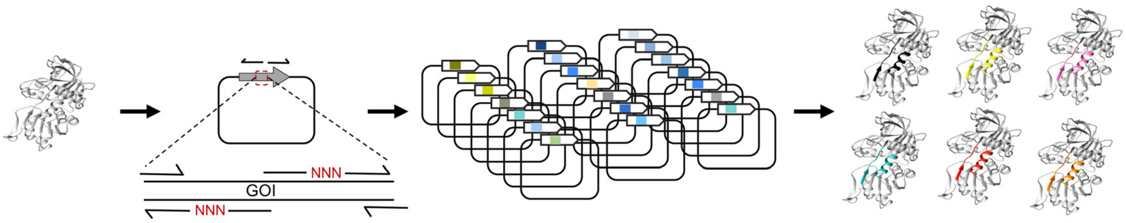

Fig. 5 Site-saturation and site-directed mutagenesis. Site-saturation mutagenesis allows the introduction of a large range of point-mutations at specific sites, while site-directed mutagenesis introduces a specific set of mutations. In both cases, mutagenic primers, which contain mismatches with the parental sequence in the positions to be mutated, are used for PCR reactions with the parental sequence as the template. The amplification products are the mutated variants. In the case of site-directed mutagenesis, primers carrying specific mutations are used. For site-saturation mutagenesis, degenerated primers containing a range of possible mutations are employed. GOI: gene of interest.

Nevertheless, the approach cannot be easily applied to cases where more than two sites must be randomised. Library size can be further reduced by randomizing only small groups of spatially-close residues located, for example, at the active site, as in the combinatorial active-site saturation test (CAST).[53] ISM[54] extended this methodology by employing the best resulting variants as templates for iterative cycles of randomization of additional clusters of residues, proving to be particularly successful for evolving enzymes with altered substrate specificity. Based on the same concept, FRISM[55] was devised to iterate over multiple sites predicted to be key for the evolved property, but introducing only a few rationally determined mutations in a given position at each cycle, instead of generating libraries. Such approaches have achieved the highest success rate when evolving enzymes with altered substrate or product specificity (especially when performed iteratively), since they allow for mutagenesis to be focused around the active site and substrate binding pocket.[56,57] Nevertheless, it is necessary to keep in mind that the assumptions that must often be made in order to reduce the library size to reasonable numbers, together with the fact that frequently not all possible combinations of mutations are assessed (which prevents the discovery of synergistic effects), risk removing optimal variants from the sequence space, making the evolution of enzymes with substrate specificity altered at will a non-trivial process (Fig. 8).

Some alternative attempts at improving the standard SSM protocols tackle generating optimal degenerate libraries from the point of view of the mutagenic primer synthesis. A near optimal solution was provided by the use of mixtures of 20 trinucleotide phosphoramidites, each encoding a different amino acid.[58,59] Although this procedure allows for the elimination of codon bias and termination codons, the high cost of the technique has prevented its widespread application. Gaytán *et al.* developed a system to produce degenerate mutagenic primers in a cost-effective way, where stop and redundant codons are eliminated, named TrimerDimer.[60] The method is based on the use of modified di and trinucleotides that are sequentially combined during solid phase synthesis to yield 20 random codons per position encoding all 20 amino acids and no stop codons. While TrimerDimer provides an attractive way of obtaining well-balanced libraries without variants with premature stop codons, its technical and equipment requirements do not make it readily available to all standard molecular biology laboratories. Such libraries can also be synthesized through commercially available services, at increasingly lower costs, which is becoming a more prominent route to obtain libraries. However, it should be noted that it is still not possible to synthesize some particularly complex library designs, such as those having multiple clusters of degenerate codons spread along the sequence or highly repeated sequences. Protocols such as Combinatorial Codon Mutagenesis (CCM) provide an interesting alternative for these cases.[61] In short, iterative rounds of fragment PCR using mutagenic primers enabling codon degeneration at different target points, and joining PCR to generate full-length assemblies are performed. This approach, an example of a combination of rational and random mutagenesis, does not require specialized reagents, and therefore offers a cost-effective way to obtain libraries with a high number of saturated sites, including non-contiguous codons.

Computational tools can also provide valuable information to guide library generation. For example, CSR-SALAD employs structural information to predict a set of residues likely to have a key role in determining the nicotinamide cofactor preference of NAD(P)-dependent oxidoreductases.[62] However, it relies on the availability of an accurate model of the structure of the target protein. While experimentally solved structures are not always available, the recent developments of accurate machine learning-based techniques, such as AlphaFold 2,[2] could largely facilitate the choice of key residues to be targeted during mutagenesis, expanding the number of macromolecules that can be subjected to rational directed evolution (Fig. 8). The availability of a large number of accurate structural predictions also facilitates the adoption of mixed rational-random mutagenesis approaches. For example, the Stepwise Loop Insertion Strategy (StLois) aims to perform multiple rounds of successive insertions of random pairs of residues coupled to screening to identify the best variants. Insertions are introduced at loops expected to tolerate variability in proximity to active sites and are identified through structural alignment of multiple related proteins. Performing stepwise insertions of one or two residues only allows to keep reasonable library sizes at each round, and has been shown to be an effective tool for engineering biocatalysts.[63,64]

Machine learning (ML) has also been applied to guide library generation by modelling the fitness landscape incorporating multiple data sources of tested variants, achieving success at even evolving novel enantiospecific enzymes.[65,66] One of the main difficulties found when applying ML to directed evolution is the scarcity of labeled data, *i.e.*, biological sequences with an associated measurement of the target property.[67] One currently active research line aims to employ unlabeled sequence data to capture a set of underlying rules assumed to be followed by any functional protein, which can then be employed to generate a sort of compressed numerical representation of protein sequences (known as ''embeddings''). Embeddings can then be provided to other models trained on the available labeled data, to expand the sequence space for which prediction of function is performed.[68,69] Alternatively, training models on unlabeled data has been proven to be useful in predicting weather a given sequence is likely to be functional or not through zero-shot prediction, which allows to eliminate non-functional sequences from subsequent experimental efforts.[70] It is also possible to use the learnt underlying sequence distribution to generate candidate sequences for novel functional variants with generative models such as generative adversarial networks.[71,72] This approach has been shown to be successful at identifying a functional malate dehydrogenase differing as much as 106 point mutations from the original sequence, which would have been impossible to identify either experimentally or with traditional predictive models due to the large size of the involved sequence space.[72]

Identification of the variants of interest

The main bottleneck of directed evolution is often the identification of the variants of interest. Typical libraries generated in modern directed evolution experiments can include many millions of variants, with the vast majority of sequences lacking the desired properties. Two main strategies have been developed to isolate successful variants: screening and selection.

Screening approaches aim to evaluate the target property across the library, leading to the identification of the best-performing variants. On the other hand, selection approaches link an improvement in the evolved property to a physical recovery of the corresponding coding sequence, or an increased survival rate.

Screening techniques

The most basic screening techniques rely on spatially separating each variant and then assessing their activities individually. This can be achieved by expressing the library of variants in a model organism, such as *E. coli*, and plating on solid media to isolate colonies corresponding to clones containing a single variant. In some cases, it is possible to perform colorimetric or fluorometric measurements directly on the colonies obtained, thanks to automated digital imaging techniques (Fig. 6a). It is also possible to transfer the colonies to multi-well liquid culture devices to perform further analysis with the liquid cultures themselves (or cell lysates) on microtiter plates. Robotic systems allow tracking thousands of such assays simultaneously (Fig. 6b).

Some biomolecules possess spectral properties that can be directly screened, such as fluorescent proteins. There are indeed multiple cases of successful identification of variants of GFP with more intense fluorescence,[73] or altered absorption or emission spectra.[74] However, most frequently it is the consumption of an enzyme substrate, or product, that is measured. When either substrate or product has a distinctive

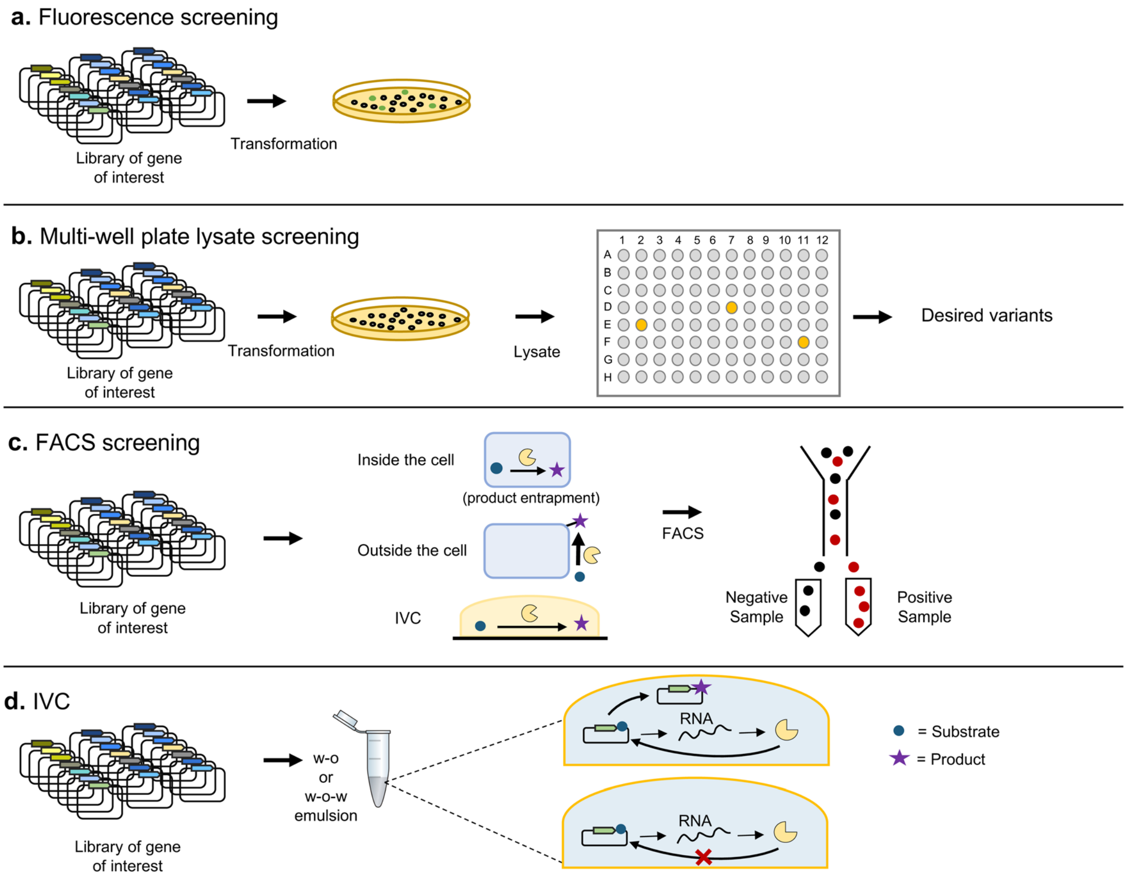

Fig. 6 Screening techniques. Some of the most frequent screening techniques are depicted. (a) Variants of proteins that confer fluorescence can be screened by analysing with digital imaging techniques cultures in solid media. (b) For proteins whose activity can be linked to a colorimetric assay or to the generation of fluorescence, it is possible to automatically transfer individual colonies to liquid cultures by means of automated multi-well liquid culture devices. The liquid cultures or their lysates can then be screened by colorimetric or fluorescent-based assays. (c) FACS enables the physical separation of individual cells based on their fluorescence properties, allowing for a higher throughput and reduced material and physical requirements. However, it is limited to biomolecules whose activity can be linked to a change in fluorescence. (d) IVC techniques replace the compartmentalization provided by cells with artificial compartmentalization, most frequently provided by emulsions of water and oil. This allows to bypass the limitation imposed by transformation efficiency, but incompatibilities between the conditions required for transcription and translation and those required for the activity of the biomolecule of interest reduce its scope of application. IVC: *in vitro* compartmentalization; w–o: water-in-oil; w-o-w: water-in-oil-in-water.

absorbance or fluorescence peak, it is possible to use the signal to evaluate enzymatic activity. Alternatively, variants can be assessed with a surrogate substrate exhibiting the desired spectral or fluorescent properties. However, the identified candidates of interest must then be reassessed with the original substrate to confirm activity. This is of special relevance when the target property is altered or improved enantiospecificity, since different substituents on the substrate can greatly affect such properties.[75] Under such cases, automated chiral GC and HPLC offer a better solution for screening positive hits.

When a convenient property of the product of interest that can be easily assessed cannot be identified, methods based on mass spectrometry (MS) provide a powerful alternative. MS offers multiple advantages over other screening methods, such as its high sensitivity and specificity, as well as the absence of a requirement for specific spectroscopic properties in the target molecule (which also eliminates the need to create surrogate substrates). Furthermore, the main limitation of the technique, the requirement for a time-consuming separation step (typically performed by HPLC), has been considerably overcome with recent developments in automated and fast autosamplers, such as Agilent's RapidFire platform, which have greatly reduced the required time down to only a few seconds per sample while employing electrospray ionization (ESI) for sample preparation.[76] Matrix-assisted laser desorption/ionization (MALDI) can also provide a very high throughput but requires immobilization of the targets onto a matrix. Therefore, this requires careful consideration of the chemistry of the target molecules to develop an appropriate immobilization method and matrix, although the effectiveness of the method for screening of novel biocatalysts has been demonstrated for a variety of cases.[77–79]

Despite automation, screening by physical separation in wells limits the number of variants that can be tested. This can be overcome by directly analyzing the different variants in bulk without previous separation, for example by fluorescence-activated cell sorting (FACS)[80] (Fig. 6c). FACS allows the separation of individual cells by means of flow cytometry, based on fluorescent signals, allowing screening of up to 10^8 variants in a few hours,[81] as long as their activities can be linked to a change in fluorescence. For example, Santoro and Schultz developed a reporter where GFPuv would only be expressed upon recombination of two loxP sites. With such an arrangement they were able to select Cre recombinase variants with increased activity towards modified loxP sites.[82]

It is also possible to use a substrate that is converted by an enzyme of interest to a fluorescent product unable to leave the cell. After washing off the permeable substrate, cells containing an active variant can be identified through the fluorescence of the product. This methodology, known as product entrapment, has been applied to a variety of cases, including glycosyltransferases, glutathione transferase and β-galactosidase.[83] FACS has also been applied to screen protein–protein interactions by fusing one of the interactors to the yeast Aga2 surface receptor.[81]

There have also been attempts to replace the natural compartmentalization provided by cells with artificial *in vitro* compartmentalization (IVC) (Fig. 6d). IVC offers the advantage of bypassing the transformation of the library, which can limit its effective size. Artificial compartmentalization is most frequently achieved through water-in-oil and water-in-oil-in-water emulsions. The latter, combined with microfluidics, allow the processing of up to millions of droplets per second, providing the basis for high-throughput screening.

IVC was first applied to identify genes encoding the HaeIII DNA methyltransferase amongst a hundred-fold excess of genes encoding other enzymes, by exploiting the resistance to restriction enzyme cleavage introduced upon methylation.[84] IVC can also be coupled to FACS, with the requirement that water-oil-water must be used in order for the flow cytometer to handle the emulsion. Tawfik and collaborators employed this approach to obtain variants of paraoxonase 1 with increased activity towards sarin-like nerve agents.[85]

More recently, microfluidic systems have also been combined with direct detection through ESI-MS to achieve a throughput of nearly 1 sample per s.[86] In principle, such a system enables sorting of nanodroplets without the limitations imposed by fluorescence-based methods, and without the requirement to immobilize the samples onto a MALDI matrix.

Selection techniques

Selection methods automatically discard all undesired variants and thus tend to allow a higher throughput than screening, but are usually less generally applicable to a broad range of biomolecules. Two main categories of approaches can be distinguished, depending on whether the activity of a variant of interest leads to a physical segregation of its encoding sequence, or to an increased survival rate of the host organism.

Display techniques. Display techniques rely on a physical connection between a nucleic acid sequence and the product it encodes, and are most frequently applied to obtain variants with improved binding to desired targets, with only a few exceptions of application to the evolution of catalysts available.[81] The most common method employs the previously-described phage display technique to expose variants in the surface of phages which contain the gene encoding the corresponding variant (Fig. 7a). This has become one of the most powerful techniques to select peptides and antibody fragments with specific binding properties, leading to the development of numerous antibody-based pharmaceuticals which are in late-stage clinical trials or have already been approved.[87] In addition to filamentous phages, other viruses can be used, such as the T7 phage for cytoplasmic proteins[88] or retroviruses and other eukaryotic viruses for eukaryotic proteins.[89]

Another, conceptually simpler, approach is plasmid display (Fig. 7b), where each encoded variant is fused to a DNA-binding protein and is placed into a vector containing the target DNA sequence. After expressing the library of variants in an appropriate organism, the cells are lysed and the resulting variant-plasmid complexes can be subjected to selection by exploiting a

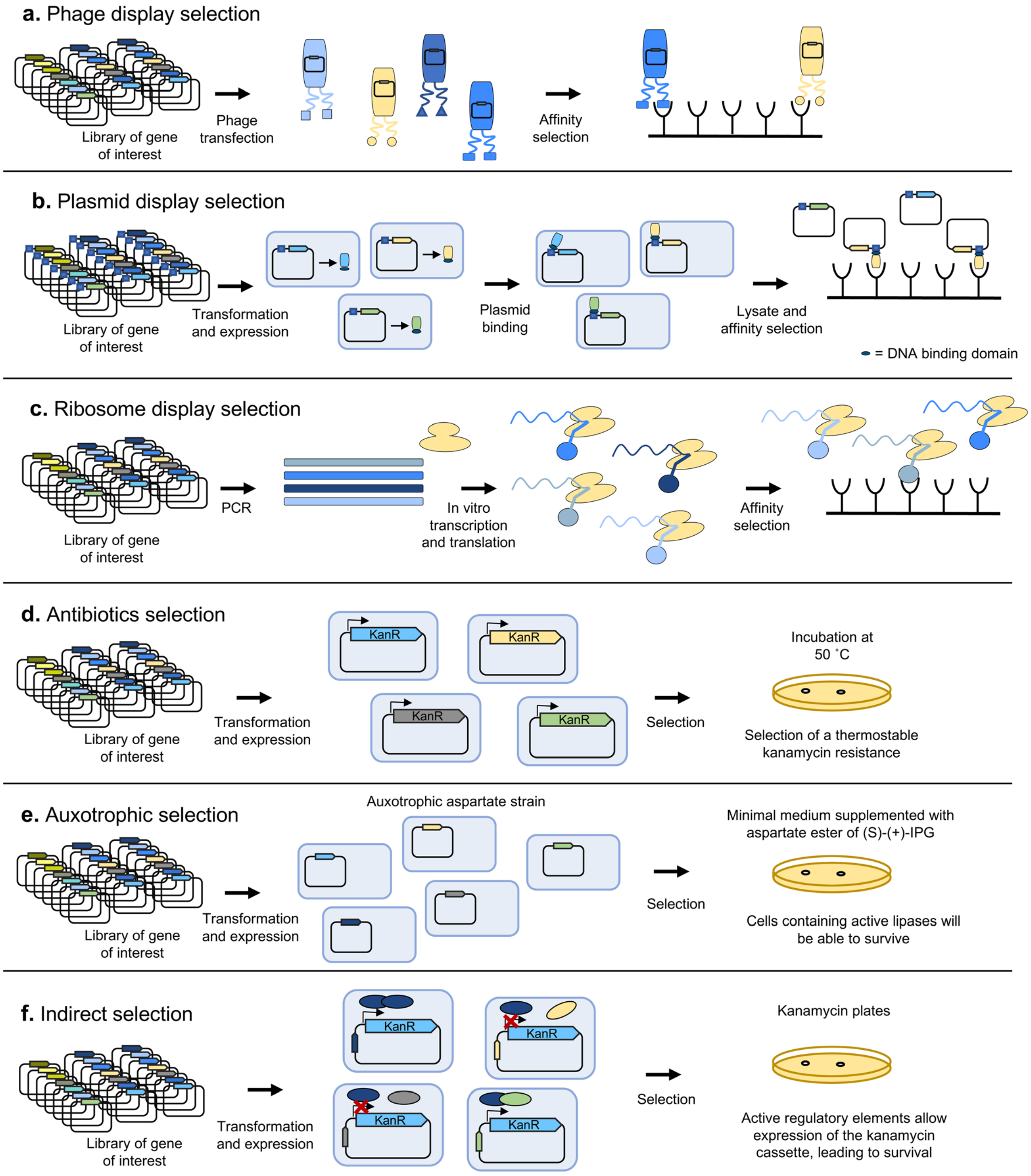
a. Phage display selection
Library of gene of interest
Phage transfection
Affinity selection
b. Plasmid display selection
Library of gene of interest
Transformation and expression
Plasmid binding
Lysate and affinity selection
= DNA binding domain
c. Ribosome display selection
Library of gene of interest
PCR
In vitro transcription and translation
Affinity selection
d. Antibiotics selection
Library of gene of interest
Transformation and expression
KanR
KanR
KanR
KanR
Selection
Incubation at 50 °C
Selection of a thermostable kanamycin resistance
e. Auxotrophic selection
Auxotrophic aspartate strain
Library of gene of interest
Transformation and expression
Selection
Minimal medium supplemented with aspartate ester of (S)-(+)-IPG
Cells containing active lipases will be able to survive
f. Indirect selection
Library of gene of interest
Transformation and expression
KanR
KanR
KanR
KanR
Selection
Kanamycin plates
Active regulatory elements allow expression of the kanamycin cassette, leading to survival

Fig. 7 Selection techniques. Some of the most frequent selection techniques are represented. (a) In phage display, protein or peptide variants are exposed on the surface of phages, and selected based on their binding affinity to a target binding partner. (b) In plasmid display, a DNA-binding protein is fused to each variant. The encoding plasmid contains the target sequence for the DNA-binding protein. After lysing cells, variants can be selected based on their binding affinity to specific target interactors. Variant sequence can then be determined from the associated plasmid thanks to the linkage provided by the DNA-binding protein. (c) Ribosome display can be used to link protein variants to their corresponding mRNA. Translation is stopped by cooling on ice, and protein–ribosome–mRNA complexes are stabilized through the addition of magnesium, enabling affinity selection and amplification of the sequence of selected variants by treatment with reverse transcriptase, followed by PCR. (d) Antibiotic resistance selection is one of the most basic types of growth complementation selection techniques. Cells are transformed with a library of variants and grown in selective medium supplemented with a certain antibiotic. Only cells expressing a variant able to confer resistance for the added antibiotic and functional under the selection conditions (such as high temperature) will survive. (e) In auxotrophy-based selections, cells auxotrophic for a certain metabolite are grown in minimal medium without said metabolite but with a precursor that can yield the required compound upon transformation by a certain enzymatic activity. Cells are transformed with a library of variants, such that only those carrying a variant able to catalyse the conversion of the precursor will be able to survive. (f) In indirect growth-complementation based selection techniques, the activity of the gene of interest is not directly responsible for an increased survival rate. Instead, its activity (such as activation of transcription) leads to an increased expression or activity of the biomolecule directly responsible for it, such as an antibiotic resistance.

functional property of the variant proteins, such as binding affinity for a specific target. For example, Samuelson and colleagues generated a library of the *N*-glycan binding protein Fbs1 fused to NF-κB1, and used beads with fetuin (a serum glycoprotein containing complex sialylated *N*-glycans) to select Fbs1 variants with improved binding to a broad range of *N*-glycans.[90]

To make even larger libraries, is also possible to link the protein variants to their corresponding mRNA or DNA, *in vitro*. One of the first methodologies to achieve this linking was ribosome display (Fig. 7c), developed in 1997 by Hanes and Plückthun[91] based on the previously devised polysome display.[92] These techniques were mostly applied to the selection of proteins with enhanced binding properties, such as antibodies, and used stop codons to stall ribosomes on mRNA, and ice/magnesium to stabilise phenotype–genotype linkage. Alternatively, protein and mRNA can be directly linked with mRNA display.[93] Modern equivalents such as CIS-display[94,95] use *cis*-acting DNA-binding proteins (repA) to stably link each displayed protein to its coding gene. Such systems can handle some of the largest library sizes ever reported: $>10^{12}$ variants, because no cells or transformation steps are required.

Growth coupling techniques. In growth coupling approaches, cell fitness is linked to the target property in the biomolecule of interest. Growth coupling techniques allow competition between variants and thus true evolution, when combined with *in vivo* mutagenesis. However, it is not always possible to couple a desired biological function to growth, and function of the evolved biomolecule can lower cell fitness. Nonetheless, proteins conferring antibiotic resistance are easily linked to increased survival rate by simply adding the antibiotic to the culture medium (Fig. 7d). Several examples exist where variants of natural antibiotic resistances have been evolved to confer enhanced or novel protection against a particular antibiotic, including proteins such as efflux pumps and β-lactamases.[96–99] Similarly, enzymes involved in essential metabolic reactions can be selected by using auxotrophic mutants (Fig. 7e). For example, Boersma *et al.* used an aspartate-auxotrophic *E. coli* strain to select an enantioselective variant of *Bacillus subtilis* lipase A with preference towards the (S) isomer of 1,2-*O*-isopropylidene-*sn*-glycerol, a precursor for the synthesis of β-adrenoceptor antagonists.[100]

It is also possible to indirectly couple the activity of the biomolecule of interest to increased survival (Fig. 7f). For example, in the QUEST method, the selection marker is placed under a synthetic transcriptional activator comprising the AraC DNA-binding domain and a domain able to bind both a molecule inducing dimerization, and the substrate of the enzyme of interest, which compete with each other. The selection marker is only expressed upon transcription factor dimerization and conversion of the substrate to a product by the enzyme of interest.[101] The scope of applicability of transcription factors as biosensors leading to the expression of a gene linked to survival (or even a fluorescent product for screening-based methods) is dependent on the availability of a transcription factor able to respond with high enough specificity to the presence of a desired product.[102] While the repertoire of naturally available transcription factors is not enough to cover a wide range of target products, novel allosteric transcription factors have been engineered through directed evolution, most typically with FACS.[103]

There have also been successful attempts at mimicking the growth coupling principle by using *in vitro* compartmentalization. This is better suited to select enzymes which can promote the replication or transcription of their coding sequences. For example, Holliger and collaborators compartmentalised coding sequences for DNA polymerases in a water-in-oil emulsion and provided flanking primers and dNTPs within each aqueous compartment.[104] They then performed PCR cycles, where a higher yield of DNA was obtained in compartments with the most active polymerases, resulting in enrichment of the sequences encoding the ’fittest’ variants. By performing several rounds of selection at increasingly higher temperatures or heparin concentrations, they managed to obtain Taq polymerase variants with enhanced thermal stability and reduced heparin inhibition.

Emerging novel paradigms in directed evolution

Recently, techniques coupling mutagenesis and selection under constant flow have been devised, enabling continuous directed evolution. Advantageously, this is >100 times faster than rounds of conventional batch selection because each cell generation can potentially provide a round of selection. One of the most famous examples is phage-assisted continuous evolution (PACE), developed by the laboratory of David Liu.[24] In the PACE system, the evolving gene of interest restores the activity of a missing gene (gene III of the M13 bacteriophage), essential gene for phage replication. Evolving phages infect a constant flow of host cells carrying an inducible mutagenesis plasmid[26] and an accessory plasmid. The latter responds to evolving gene variants to induce the essential gene III, so completing the phage life cycle. Thus, the system allows for the continuous evolution of the gene of interest as long as its activity can be linked to the expression of gIII. A recent variant technique (PACEmid) adapted the method to phagemids, allowing large starting-library generation.[105] The method was illustrated by evolving the smallest dual transcription factor reported to date: a 63-amino acid peptide.[27]

Continuous directed evolution is best suited to applications where it is straightforward to link desired property to increased survival of cells. While continuous *in vivo* evolution techniques have been devised as previously discussed, the identification and isolation of the variants of interest remains a central bottleneck.[106] This has limited its practical application, specially to small molecule biocatalysts. Indeed, only a few examples of application of PACE-like protocols to this type of cases have been described.[107,108] Therefore, novel selection methods allowing for such linking to be established for a wider range of properties and biomolecules must be developed, to expand the

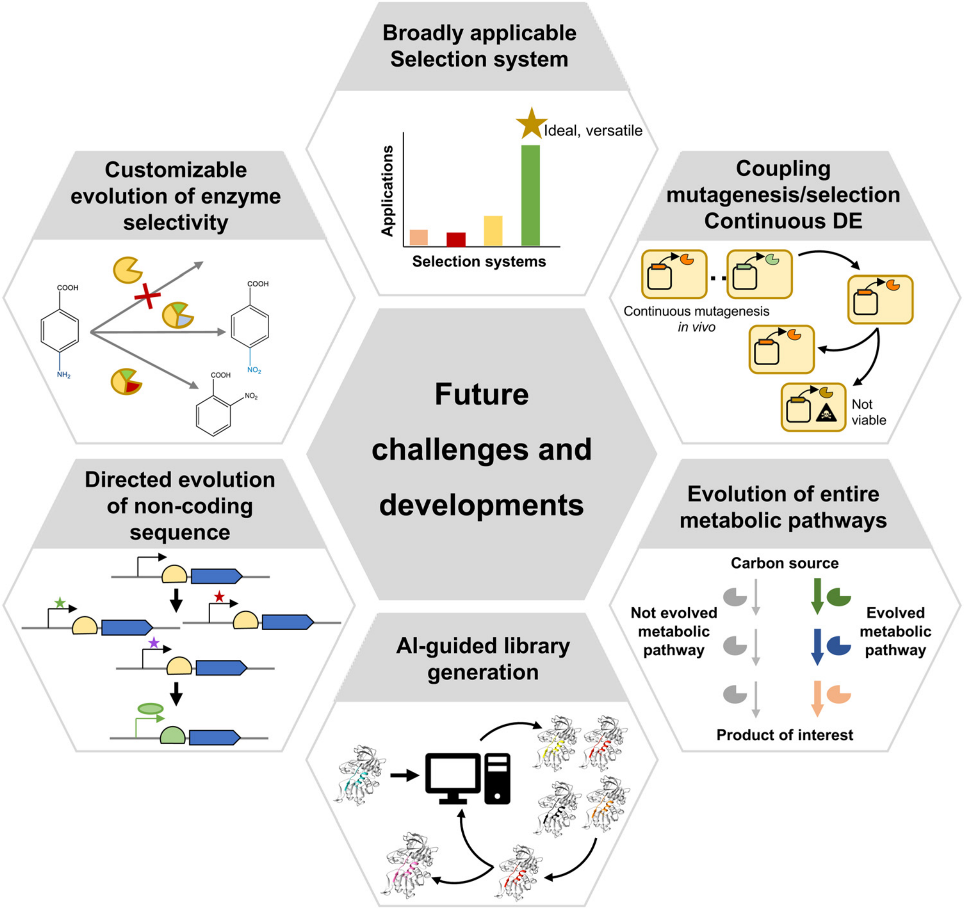
Broadly applicable Selection system
Ideal, versatile
Applications
Selection systems
Customizable evolution of enzyme selectivity
COOH
NH₂
COOH
NO₂
COOH
NO₂
Coupling mutagenesis/selection Continuous DE
Continuous mutagenesis in vivo
Not viable
Future challenges and developments
Directed evolution of non-coding sequence
Evolution of entire metabolic pathways
Carbon source
Not evolved metabolic pathway
Evolved metabolic pathway
Product of interest
AI-guided library generation

Fig. 8 Summary of main challenges and future developments of directed evolution.

scope of continuous evolution approaches. Novel selection methods recently developed based on direct or indirect regeneration of essential cofactors as a consequence of enzymatic activities might provide an useful tool in this direction.[109]

Directed evolution is moving towards the evolution of ever more complex biomolecules and systems (Fig. 8). Metalloenzymes are examples, which offer an ample range of catalysed reactions thanks to their metallic cofactors. Artificial metalloenzymes have been developed either using natural metalloenzymes as the starting point, or by engineering protein backbones to incorporate metal cofactors.[110] In both cases, evolving novel metalloenzymes poses additional challenges, due to the requirement that the cofactor must be efficiently produced or incorporated by the expression system, and the toxicity or propensity to react with other metabolites (such as glutathione) of some metal cofactors, which can inactivate them. Such issues can be partially overcome by secreting candidate variants to the periplasmic space through the fusion of a signal peptide, as demonstrated by the evolution of a ruthenium-binding olefin "metathase".[110] It is also possible to perform bioconjugation of the required cofactor after expressing the apo protein and lysing cells, as demonstrated for the case of evolution of an improved cyclopropanase dependent on a dirhodium cofactor with random mutagenesis not necessarily focused on the active site.[111] Alternatively, cells expressing transporters that import an exogenously provided cofactor provide another viable strategy, as long as cofactor toxicity is not too high. This strategy was successfully applied for the evolution of a novel cytochrome P450 variant with an iridium-containing heme-like cofactor able to produce chiral amines enantioselectively.[112]

Of increasing interest is also the evolution of entire metabolic pathways, where a set of enzymes work together towards the generation of a desired product (Fig. 8). A possible approach is to individually evolve enzymes as required for a novel, rationally-designed synthetic pathway to perform *in vitro* synthesis through a biocatalytic cascade, as shown for the manufacturing of islatravir.[113] However, complete pathway optimization requires not only evolving the individual enzymes in the pathway, but also their arrangement and regulation. For example, Ajikumar *et al.* split a taxadiene production pathway into two modules and screened for different combinations of promoters and replication origins to find the optimal expression level and copy number of each module, thus maximizing taxadiene yield.[114] Similar modular optimization approaches have since been applied to a variety of metabolic pathways.[115] Simultaneously, more specialised mutagenesis methods aimed at assembling and optimising combinatorial expression libraries have been developed, such as Start-Stop Assembly[116] or the Yeast Toolkit for Modular Assembly.[117] However, the application of such methodologies to directed evolution has, so far, remain relatively limited. This is partly due to the fact that the effect of optimization of combinations of regulatory, non-coding elements can only be discerned *in vivo* (Fig. 8), and therefore the generated libraries must be subjected to selection. Given that, as previously mentioned, no selection methodology widely applicable to a large range of pathways is available, the extent to which such combinatorial libraries can be employed for directed evolution is constrained.

An extreme case is the application of directed evolution to whole genomes. One of the most popular methods for performing genome-scale directed evolution is Adaptive Laboratory Evolution (ALE), where microorganisms are cultured under controlled, specific conditions that enable the selection of phenotypes associated with improved growth under the chosen environment. A typical application example is the selection of *E. coli* strains with increased resistance to high temperatures by growing them at 42.2 °C during multiple generations.[118] This has been successfully attempted up to 50 °C with libraries of transcription network rewirings as the source of variation.[119]

More recently, mutagenesis methods for genome-scale directed evolution have been devised, including several techniques based on Multiplex Automated Genome Engineering (MAGE) which employ combinations of multiple oligonucleotides to target up to thousands of genomic locations simultaneously.[120,121] The potential of such approaches to develop new variant organisms serving as optimised whole-cell catalysts was demonstrated by Wang *et al.*, who managed to obtain an *E. coli* strain overproducing lycopene.[122] The latest approaches employ a combination of RNA interference and CRISPR/Cas9 systems, to achieve both knockdown and activation of genes within a single cell, showing promising results in the selection of yeast strains with multiple phenotypes, such as cellulase expression and isobutanol production.[123]

Conclusion

Over the past few decades, directed evolution has clearly demonstrated its potential to develop variants of biomolecules with novel or enhanced properties. Many cases of successful application of directed evolution have been possible thanks to the development of myriad techniques for library generation and variant identification. While efficient genetic diversification can be achieved relatively easily with modern molecular biology and chemical synthesis techniques, as well as AI-assisted computational tools, variant identification often remains as a major bottleneck. A tool that is current lacking is a selection system broadly applicable to a wider range of biomolecules, especially enzymes acting on small molecules, and even more complex systems such as cellular pathways and non-coding sequences. Such a system could be coupled to *in vivo* mutagenesis to enable real-time automated enzyme evolution. Ultimately, this would provide a solution to two of the current major challenges of directed evolution: widening applicability to multiple types of biomolecules and properties while minimising human intervention.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by a JSPS Postdoctoral Fellowship for Research in Japan (Standard) Grant Number P20810 to LSV (Overseas researcher under Postdoctoral Fellowship of Japan Society for the Promotion of Science). R. L.-A. received funding from BBSRC (BB/R01602X/1, BB/T013176/1, BB/T011408/1 – 19-ERACoBioTech-33 SyCoLim), British Council 527429894, Newton Advanced Fellowship (NAF\R1\201187), Yeast4Bio Cost Action 18229, European Research Council (ERC) (DEUSBIO – 949080) and the Bio-based Industries Joint (PERFECOAT-101022370) under the European Union's Horizon 2020 research and innovation programme.

References

1 T. Johannes, M. R. Simurdiak and H. Zhao, Biocatalysis, in *Encyclopedia of Chemical Processing*, CRC Press, 2005, pp. 101–110.

2 E. Callaway, It will change everything': DeepMind's AI makes gigantic leap in solving protein structures, *Nature*, 2020, **588**(7837), 203–204.

3 T. Schwede, Protein modeling: what happened to the 'protein structure gap'?, *Structure*, 2013, **21**(9), 1531–1540.

4 F. H. Arnold, Innovation by evolution: bringing new chemistry to life (Nobel lecture), *Angew. Chem., Int. Ed.*, 2019, **58**(41), 14420–14426.

5 Y. Wang, P. Xue, M. Cao, T. Yu, S. T. Lane and H. Zhao, Directed evolution: methodologies and applications, *Chem. Rev.*, 2021, **121**(20), 12384–12444.

6 D. R. Mills, R. L. Peterson and S. Spiegelman, An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule, *Proc. Natl. Acad. Sci. U. S. A.*, 1967, **58**(1), 217–224.

7 G. P. Smith, Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface, *Science*, 1985, **228**(4705), 1315–1317.

8 H. Leemhuis, V. Stein, A. D. Griffiths and F. Hollfelder, New genotype-phenotype linkages for directed evolution of functional proteins, *Curr. Opin. Struct. Biol.*, 2005, **15**(4), 472–478.

9 H. Lee, E. Popodi, H. Tang and P. L. Foster, Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**(41), E2774–E2783.

10 A. Greener, M. Callahan and B. Jerpseth, An Efficient Random Mutagenesis Technique Using an E. coli Mutator Strain, in *In Vitro Mutagenesis Protocols*, ed. M. K. Trower, Humana Press, Totowa, NJ, 1996, pp. 375–85.

11 R. M. Myers, L. S. Lerman and T. Maniatis, A general method for saturation mutagenesis of cloned DNA fragments, *Science*, 1985, **229**(4710), 242–247.

12 R. C. Cadwell and G. F. Joyce, Randomization of genes by PCR mutagenesis, *PCR Methods Appl.*, 1992, **2**(1), 28–33.

13 K. Chen and F. H. Arnold, Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide, *Proc. Natl. Acad. Sci. U. S. A.*, 1993, **90**(12), 5618–5622.

14 M. Kaltenbach and N. Tokuriki, Generation of effective libraries by neutral drift, *Methods Mol. Biol.*, 2014, **1179**, 69–81.

15 A. Fire and S. Q. Xu, Rolling replication of short DNA circles, *Proc. Natl. Acad. Sci. U. S. A.*, 1995, **92**(10), 4641–4645.

16 R. Fujii, M. Kitaoka and K. Hayashi, One-step random mutagenesis by error-prone rolling circle amplification, *Nucleic Acids Res.*, 2004, **32**(19), e145.

17 D. Shortle and J. Sondek, The emerging role of insertions and deletions in protein engineering, *Curr. Opin. Biotechnol*, 1995, **6**(4), 387–393.

18 S. Emond, M. Petek, E. J. Kay, B. Heames, S. R. A. Devenish and N. Tokuriki, *et al.*, Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis, *Nat. Commun.*, 2020, **11**(1), 3469.

19 R. Fujii, M. Kitaoka and K. Hayashi, RAISE: a simple and novel method of generating random insertion and deletion mutations, *Nucleic Acids Res.*, 2006, **34**(4), e30.

20 Y. Kipnis, E. Dellus-Gur and D. S. Tawfik, TRINS: a method for gene modification by randomized tandem repeat insertions, *Protein Eng. Des. Sel.*, 2012, **25**(9), 437–444.

21 S. Haapa, S. Suomalainen, S. Eerikäinen, M. Airaksinen, L. Paulin and H. Savilahti, An efficient DNA sequencing strategy based on the bacteriophage mu in vitro DNA transposition reaction, *Genome Res.*, 1999, **9**(3), 308–315.

22 D. D. Jones, Triplet nucleotide removal at random positions in a target gene: the tolerance of TEM-1 beta-lactamase to an amino acid deletion, *Nucleic Acids Res.*, 2005, **33**(9), e80.

23 P. A. G. Tizei, E. Harris, S. Withanage, M. Renders and V. B. Pinheiro, A novel framework for engineering protein loops exploring length and compositional variation, *Sci. Rep.*, 2021, **11**(1), 9134.

24 K. M. Esvelt, J. C. Carlson and D. R. Liu, A system for the continuous directed evolution of biomolecules, *Nature*, 2011, **472**(7344), 499–503.

25 A. Greener and M. Callahan, XL1-Red: a highly efficient random mutagenesis strain, *Strategies*, 1994, **7**, 32–34.

26 A. H. Badran and D. R. Liu, Development of potent in vivo mutagenesis plasmids with broad mutational spectra, *Nat. Commun.*, 2015, **6**(1), 8425.

27 A. K. Brödel, R. Rodrigues, A. Jaramillo and M. Isalan, Accelerated evolution of a minimal 63-amino acid dual transcription factor, *Sci. Adv.*, 2020, **6**(24), eaba2728.

28 M. Camps, J. Naukkarinen, B. P. Johnson and L. A. Loeb, Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**(17), 9727–9732.

29 A. Ravikumar, A. Arrieta and C. C. Liu, An orthogonal DNA replication system in yeast, *Nat. Chem. Biol.*, 2014, **10**(3), 175–177.

30 A. Ravikumar, G. A. Arzumanyan, M. K. A. Obadi, A. A. Javanpour and C. C. Liu, Scalable, continuous evolution of genes at mutation rates above genomic error thresholds, *Cell*, 2018, **175**(7), 1946–1957.e13.

31 N. Crook, J. Abatemarco, J. Sun, J. M. Wagner, A. Schmitz and H. S. Alper, In vivo continuous evolution of genes and pathways in yeast, *Nat. Commun.*, 2016, **7**, 13051.

32 C. L. Moore, L. J. Papa III and M. D. Shoulders, A processive protein chimera introduces mutations across defined DNA regions in vivo, *J. Am. Chem. Soc.*, 2018, **140**(37), 11560–11564.

33 S. O. Halperin, C. J. Tou, E. B. Wong, C. Modavi, D. V. Schaffer and J. E. Dueber, CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window, *Nature*, 2018, **560**(7717), 248–252.

34 X. Didelot and M. C. J. Maiden, Impact of recombination on bacterial evolution, *Trends Microbiol.*, 2010, **18**(7), 315–322.

35 J. Stapley, P. G. D. Feulner, S. E. Johnston, A. W. Santure and C. M. Smadja, Recombination: the good, the bad and the variable, *Philos. Trans. R. Soc. London, Ser. B*, 2017, **372**(1736), 20170279.

36 A. J. Ruff, A. Dennig and U. Schwaneberg, To get what we aim for--progress in diversity generation methods, *FEBS J.*, 2013, **280**(13), 2961–2978.

37 W. P. Stemmer, Rapid evolution of a protein in vitro by DNA shuffling, *Nature*, 1994, **370**(6488), 389–391.

38 H. Zhao, L. Giver, Z. Shao, J. A. Affholter and F. H. Arnold, Molecular evolution by staggered extension process (StEP) in vitro recombination, *Nat. Biotechnol.*, 1998, **16**(3), 258–261.

39 K. Skamaki, S. Emond, M. Chodorge, J. Andrews, D. G. Rees and D. Cannon, *et al.*, In vitro evolution of antibody affinity via insertional scanning mutagenesis of an entire antibody variable region, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**(44), 27307–27318.

40 W. M. Coco, W. E. Levinson, M. J. Crist, H. J. Hektor, A. Darzins and P. T. Pienkos, *et al.*, DNA shuffling method for generating highly recombined genes and evolved enzymes, *Nat. Biotechnol.*, 2001, **19**(4), 354–359.

41 M. Ostermeier, J. H. Shim and S. J. Benkovic, A combinatorial approach to hybrid enzymes independent of DNA homology, *Nat. Biotechnol.*, 1999, **17**(12), 1205–1209.

42 V. Sieber, C. A. Martinez and F. H. Arnold, Libraries of hybrid proteins from distantly related sequences, *Nat. Biotechnol.*, 2001, **19**(5), 456–460.

43 S. Lutz, M. Ostermeier, G. L. Moore, C. D. Maranas and S. J. Benkovic, Creating multiple-crossover DNA libraries independent of sequence identity, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**(20), 11248–11253.

44 K. Hiraga and F. H. Arnold, General method for sequence-independent site-directed chimeragenesis, *J. Mol. Biol.*, 2003, **330**(2), 287–296.

45 D. Gonzalez-Perez, P. Molina-Espeja, E. Garcia-Ruiz and M. Alcalde, Mutagenic Organized Recombination Process by Homologous IN vivo Grouping (MORPHING) for directed enzyme evolution, *PLoS One*, 2014, **9**(3), e90919.

46 R. Higuchi, B. Krummel and R. K. Saiki, A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions, *Nucleic Acids Res.*, 1988, **16**(15), 7351–7367.

47 Y. An, J. Ji, W. Wu, A. Lv, R. Huang and Y. Wei, A rapid and efficient method for multiple-site mutagenesis with a modified overlap extension PCR, *Appl. Microbiol. Biotechnol.*, 2005, **68**(6), 774–778.

48 R.-H. Peng, A.-S. Xiong and Q.-H. Yao, A direct and efficient PAGE-mediated overlap extension PCR method for gene multiple-site mutagenesis, *Appl. Microbiol. Biotechnol.*, 2006, **73**(1), 234–240.

49 A. Dennig, A. V. Shivange, J. Marienhagen and U. Schwaneberg, OmniChange: the sequence independent method for simultaneous site-saturation of five codons, *PLoS One*, 2011, **6**(10), e26222.

50 C. Neylon, Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution, *Nucleic Acids Res.*, 2004, **32**(4), 1448–1459.

51 M. Isalan, Construction of semi-randomized gene libraries with weighted oligonucleotide synthesis and PCR, *Nat. Protoc.*, 2006, **1**(1), 468–475.

52 L. Tang, H. Gao, X. Zhu, X. Wang, M. Zhou and R. Jiang, Construction of 'small-intelligent' focused mutagenesis libraries using well-designed combinatorial degenerate primers, *Biotechniques*, 2012, **52**(3), 149–158.

53 M. T. Reetz, M. Bocola, J. D. Carballeira, D. Zha and A. Vogel, Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test, *Angew. Chem., Int. Ed.*, 2005, **44**(27), 4192–4196.

54 M. T. Reetz, L.-W. Wang and M. Bocola, Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space, *Angew. Chem., Int. Ed.*, 2006, **45**(8), 1236–1241.

55 D. Li, Q. Wu and M. T. Reetz, Focused rational iterative site-specific mutagenesis (FRISM), *Methods Enzymol.*, 2020, **643**, 225–242.

56 M. T. Reetz, *Directed evolution of selective enzymes*, Wiley-VCH Verlag, Weinheim, Germany, 2016, p. 320.

57 L. P. Parra, R. Agudo and M. T. Reetz, Directed evolution by using iterative saturation mutagenesis based on multiresidue sites, *ChemBioChem*, 2013, **14**(17), 2301–2309.

58 B. Virnekäs, L. Ge, A. Plückthun, K. C. Schneider, G. Wellnhofer and S. E. Moroney, Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis, *Nucleic Acids Res.*, 1994, **22**(25), 5600–5607.

59 A. Kayushin, M. Korosteleva and A. Miroshnikov, Large-scale solid-phase preparation of 3'-unprotected trinucleotide phosphotriesters-precursors for synthesis of trinucleotide phosphoramidites, *Nucleosides, Nucleotides Nucleic Acids*, 2000, **19**(10–12), 1967–1976.

60 P. Gaytán, C. Contreras-Zambrano, M. Ortiz-Alvarado, A. Morales-Pablos and J. Yáñez, TrimerDimer: an oligonucleotide-based saturation mutagenesis approach that removes redundant and stop codons, *Nucleic Acids Res.*, 2009, **37**(18), e125.

61 K. D. Belsare, M. C. Andorfer, F. S. Cardenas, J. R. Chael, H. J. Park and J. C. Lewis, A simple combinatorial Codon Mutagenesis method for targeted protein engineering, *ACS Synth. Biol.*, 2017, **6**(3), 416–420.

62 J. K. B. Cahn, C. A. Werlang, A. Baumschlager, S. Brinkmann-Chen, S. L. Mayo and F. H. Arnold, A General Tool for Engineering the NAD/NADP Cofactor Preference of Oxidoreductases, *ACS Synth. Biol.*, 2016, **6**(2), 326–333.

63 M. A. Hoque, Y. Zhang, L. Chen, G. Yang, M. A. Khatun and H. Chen, *et al.*, Stepwise loop insertion strategy for active site remodeling to generate novel enzyme functions, *ACS Chem. Biol.*, 2017, **12**(5), 1188–1193.

64 P. M. Heinemann, D. Armbruster and B. Hauer, Active-site loop variations adjust activity and selectivity of the cumene dioxygenase, *Nat. Commun.*, 2021, **12**(1), 1095.

65 Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann and F. H. Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**(18), 8852–8858.

66 G. Li, Y. Dong and M. T. Reetz, Can machine learning revolutionize directed evolution of selective enzymes?, *Adv. Synth. Catal.*, 2019, **361**(11), 2377–2386.

67 B. J. Wittmann, K. E. Johnston, Z. Wu and F. H. Arnold, Advances in machine learning for directed evolution, *Curr. Opin. Struct. Biol.*, 2021, **69**, 11–18.

68 T. Bepler and B. Berger, Learning protein sequence embeddings using information from structure, arXiv [cs.LG], 2019, arXiv:1902.08661, DOI: **10.48550/arXiv.1902.08661**.

69 S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt and G. M. Church, Low-N protein engineering with data-efficient deep learning, *Nat. Methods*, 2021, **18**(4), 389–396.

70 A. J. Riesselman, J. B. Ingraham and D. S. Marks, Deep generative models of genetic variation capture the effects of mutations, *Nat. Methods*, 2018, **15**(10), 816–822.

71 A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand and R. R. Eguchi, *et al.*, ProGen: Language modeling for protein generation, arXiv, 2020, arXiv:2004.03497, DOI: **10.48550/arXiv.2004.03497**.

72 D. Repecka, V. Jauniskis, L. Karpus, E. Rembeza, I. Rokaitis and J. Zrimec, *et al.*, Expanding functional protein sequence spaces using generative adversarial networks, *Nat. Mach. Intell.*, 2021, **3**(4), 324–333.

73 A. Crameri, E. A. Whitehorn, E. Tate and W. P. Stemmer, Improved green fluorescent protein by molecular evolution using DNA shuffling, *Nat. Biotechnol.*, 1996, **14**(3), 315–319.

74 R. Heim, D. C. Prasher and R. Y. Tsien, Wavelength mutations and posttranslational autoxidation of green

fluorescent protein, *Proc. Natl. Acad. Sci. U. S. A.*, 1994, **91**(26), 12501–12504.

75 L. G. Otten, F. Hollmann and I. W. C. E. Arends, Enzyme engineering for enantioselectivity: from trial-and-error to rational design?, *Trends Biotechnol.*, 2010, **28**(1), 46–54.

76 T. Bretschneider, C. Ozbal, M. Holstein, M. Winter, F. H. Buettner and S. Thamm, *et al.*, RapidFire BLAZE-mode is boosting ESI-MS toward high-throughput-screening, *SLAS Technol.*, 2019, **24**(4), 386–393.

77 P. Xue, T. Si, S. Mishra, L. Zhang, K. Choe and J. V. Sweedler, *et al.*, A mass spectrometry-based high-throughput screening method for engineering fatty acid synthases with improved production of medium-chain fatty acids, *Biotechnol. Bioeng.*, 2020, **117**(7), 2131–2138.

78 A. J. Pluchinsky, D. J. Wackelin, X. Huang, F. H. Arnold and M. Mrksich, High throughput screening with SAMDI mass spectrometry for directed evolution, *J. Am. Chem. Soc.*, 2020, **142**(47), 19804–19808.

79 S. Zhang, J. Zhu, S. Fan, W. Xie, Z. Yang and T. Si, Directed evolution of a cyclodipeptide synthase with new activities via label-free mass spectrometric screening, *Chem. Sci.*, 2022, **13**(25), 7581–7586.

80 W. Zeng, L. Guo, S. Xu, J. Chen and J. Zhou, High-throughput screening technology in industrial biotechnology, *Trends Biotechnol.*, 2020, **38**(8), 888–906.

81 I. Chen, B. M. Dorr and D. R. Liu, A general strategy for the evolution of bond-forming enzymes using yeast display, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**(28), 11399–11404.

82 S. W. Santoro and P. G. Schultz, Directed evolution of the site specificity of Cre recombinase, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**(7), 4185–4190.

83 G. Yang and S. G. Withers, Ultrahigh-throughput FACS-based screening for directed enzyme evolution, *ChemBioChem*, 2009, **10**(17), 2704–2715.

84 D. S. Tawfik and A. D. Griffiths, Man-made cell-like compartments for molecular evolution, *Nat. Biotechnol.*, 1998, **16**(7), 652–656.

85 R. D. Gupta, M. Goldsmith, Y. Ashani, Y. Simo, G. Mullokandov and H. Bar, *et al.*, Directed evolution of hydrolases for prevention of G-type nerve agent intoxication, *Nat. Chem. Biol.*, 2011, **7**(2), 120–125.

86 D. A. Holland-Moritz, M. K. Wismer, B. F. Mann, I. Farasat, P. Devine and E. D. Guetschow, *et al.*, Mass activated droplet sorting (MADS) enables high-throughput screening of enzymatic reactions at nanoliter scale, *Angew. Chem., Int. Ed.*, 2020, **59**(11), 4470–4477.

87 A. E. Nixon, D. J. Sexton and R. C. Ladner, Drugs derived from phage display: from candidate identification to clinical practice, *MAbs.*, 2014, **6**(1), 73–85.

88 H. K. Binz, P. Amstutz, A. Kohl, M. T. Stumpp, C. Briand and P. Forrer, *et al.*, High-affinity binders selected from designed ankyrin repeat protein libraries, *Nat. Biotechnol.*, 2004, **22**(5), 575–582.

89 J. H. Urban and C. A. Merten, Retroviral display in gene therapy, protein engineering, and vaccine development, *ACS Chem. Biol.*, 2011, **6**(1), 61–74.

90 M. Chen, X. Shi, R. M. Duke, C. I. Ruse, N. Dai and C. H. Taron, *et al.*, An engineered high affinity Fbs1 carbohydrate binding protein for selective capture of *N*-glycans and *N*-glycopeptides, *Nat. Commun.*, 2017, **8**, 15487.

91 J. Hanes and A. Plückthun, In vitro selection and evolution of functional proteins by using ribosome display, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**(10), 4937–4942.

92 L. C. Mattheakis, R. R. Bhatt and W. J. Dower, An in vitro polysome display system for identifying ligands from very large peptide libraries, *Proc. Natl. Acad. Sci. U. S. A.*, 1994, **91**(19), 9022–9026.

93 D. S. Wilson, A. D. Keefe and J. W. Szostak, The use of mRNA display to select high-affinity protein-binding peptides, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**(7), 3750–3755.

94 R. Odegrip, D. Coomber, B. Eldridge, R. Hederer, P. A. Kuhlman and C. Ullman, *et al.*, CIS display: In vitro selection of peptides from libraries of protein-DNA complexes, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**(9), 2806–2810.

95 S. Patel, P. Mathonet, A. M. Jaulent and C. G. Ullman, Selection of a high-affinity WW domain against the extracellular region of VEGF receptor isoform-2 from a combinatorial library using CIS display, *Protein Eng., Des. Sel.*, 2013, **26**(4), 307–315.

96 Á. Nyerges, B. Csörgő, G. Draskovits, B. Kintses, P. Szili and G. Ferenc, *et al.*, Directed evolution of multiple genomic loci allows the prediction of antibiotic resistance, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115**(25), E5726–E5735.

97 C. Feiler, A. C. Fisher, J. T. Boock, M. J. Marrichi, L. Wright and P. A. M. Schmidpeter, *et al.*, Directed evolution of Mycobacterium tuberculosis β-lactamase reveals gatekeeper residue that regulates antibiotic resistance and catalytic efficiency, *PLoS One*, 2013, **8**(9), e73123.

98 S. Sun, W. Zhang, B. Mannervik and D. I. Andersson, Evolution of broad spectrum β-lactam resistance in an engineered metallo-β-lactamase, *J. Biol. Chem.*, 2013, **288**(4), 2314–2324.

99 E. Bokma, E. Koronakis, S. Lobedanz, C. Hughes and V. Koronakis, Directed evolution of a bacterial efflux pump: adaptation of the E. coli TolC exit duct to the Pseudomonas MexAB translocase, *FEBS Lett.*, 2006, **580**(22), 5339–5343.

100 Y. L. Boersma, M. J. Dröge, A. M. van der Sloot, T. Pijning, R. H. Cool and B. W. Dijkstra, *et al.*, A novel genetic selection system for improved enantioselectivity of Bacillus subtilis lipase A, *ChemBioChem*, 2008, **9**(7), 1110–1115.

101 S. M. Firestine, F. Salinas, A. E. Nixon, S. J. Baker and S. J. Benkovic, Using an AraC-based three-hybrid system to detect biocatalysts in vivo, *Nat. Biotechnol.*, 2000, **18**(5), 544–547.

102 J. A. Dietrich, D. L. Shis, A. Alikhani and J. D. Keasling, Transcription factor-based screens and synthetic selections for microbial small-molecule biosynthesis, *ACS Synth. Biol.*, 2013, **2**(1), 47–58.

103 L. F. M. Machado and N. Dixon, Directed evolution of transcription factor-based biosensors for altered effector specificity, *Methods Mol. Biol.*, 2022, **2461**, 175–193.

104 F. J. Ghadessy, J. L. Ong and P. Holliger, Directed evolution of polymerase function by compartmentalized self-replication, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**(8), 4552–4557.

105 A. K. Brödel, A. Jaramillo and M. Isalan, Engineering orthogonal dual transcription factors for multi-input synthetic promoters, *Nat. Commun.*, 2016, **7**(1), 13858.

106 R. S. Molina, G. Rix, A. A. Mengiste, B. Álvarez, D. Seo and H. Chen, *et al.*, In vivo hypermutation and continuous evolution, *Nat. Rev. Methods Primers*, 2022, **2**(1), 1–22.

107 T. B. Roth, B. M. Woolston, G. Stephanopoulos and D. R. Liu, Phage-Assisted Evolution of Bacillus methanolicus Methanol Dehydrogenase 2, *ACS Synth. Biol.*, 2019, **8**(4), 796–806.

108 K. A. Jones, H. M. Snodgrass, K. Belsare, B. C. Dickinson and J. C. Lewis, Phage-assisted continuous evolution and selection of enzymes for chemical synthesis, *ACS Cent. Sci.*, 2021, **7**(9), 1581–1590.

109 L. Sellés Vidal, J. W. Murray and J. T. Heap, Versatile selective evolutionary pressure using synthetic defect in universal metabolism, *Nat. Commun.*, 2021, **12**(1), 6859.

110 M. Jeschek, R. Reuter, T. Heinisch, C. Trindler, J. Klehr and S. Panke, *et al.*, Directed evolution of artificial metalloenzymes for in vivo metathesis, *Nature*, 2016, **537**(7622), 661–665.

111 H. Yang, A. M. Swartz, H. J. Park, P. Srivastava, K. Ellis-Guardiola and D. M. Upp, *et al.*, Evolving artificial metalloenzymes via random mutagenesis, *Nat. Chem.*, 2018, **10**(3), 318–324.

112 Y. Gu, B. J. Bloomer, Z. Liu, R. Chen, D. S. Clark and J. F. Hartwig, Directed evolution of artificial metalloenzymes in whole cells, *Angew. Chem., Int. Ed.*, 2022, **61**(5), e202110519.

113 M. A. Huffman, A. Fryszkowska, O. Alvizo, M. Borra-Garske, K. R. Campos and K. A. Canada, *et al.*, Design of an in vitro biocatalytic cascade for the manufacture of islatravir, *Science*, 2019, **366**(6470), 1255–1259.

114 P. K. Ajikumar, W.-H. Xiao, K. E. J. Tyo, Y. Wang, F. Simeon and E. Leonard, *et al.*, Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli, *Science*, 2010, **330**(6000), 70–74.

115 M. Jeschek, D. Gerngross and S. Panke, Combinatorial pathway optimization for streamlined metabolic engineering, *Curr. Opin. Biotechnol*, 2017, **47**, 142–151.

116 G. M. Taylor, P. M. Mordaka and J. T. Heap, Start-Stop Assembly: a functionally scarless DNA assembly system optimized for metabolic engineering, *Nucleic Acids Res.*, 2019, **47**(3), e17.

117 M. E. Lee, W. C. DeLoache, B. Cervantes and J. E. Dueber, A highly characterized yeast toolkit for modular, multipart assembly, *ACS Synth. Biol.*, 2015, **4**(9), 975–986.

118 O. Tenaillon, A. Rodríguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett and A. D. Long, *et al.*, The molecular diversity of adaptive convergence, *Science*, 2012, **335**(6067), 457–461.

119 M. Isalan, C. Lemerle, K. Michalodimitrakis, C. Horn, P. Beltrao and E. Raineri, *et al.*, Evolvability and hierarchy in rewired bacterial gene networks, *Nature*, 2008, **452**(7189), 840–845.

120 M. T. Bonde, S. Kosuri, H. J. Genee, K. Sarup-Lytzen, G. M. Church and M. O. A. Sommer, *et al.*, Direct mutagenesis of thousands of genomic targets using microarray-derived oligonucleotides, *ACS Synth. Biol.*, 2015, **4**(1), 17–22.

121 E. M. Barbieri, P. Muir, B. O. Akhuetie-Oni, C. M. Yellman and F. J. Isaacs, Precise editing at DNA replication forks enables multiplex genome engineering in eukaryotes, *Cell*, 2017, **171**(6), 1453–1467.e13.

122 H. H. Wang, F. J. Isaacs, P. A. Carr, Z. Z. Sun, G. Xu and C. R. Forest, *et al.*, Programming cells by multiplex genome engineering and accelerated evolution, *Nature*, 2009, **460**(7257), 894–898.

123 T. Si, R. Chao, Y. Min, Y. Wu, W. Ren and H. Zhao, Automated multiplex genome-scale engineering in yeast, *Nat. Commun.*, 2017, **8**(1), 15187.

124 M. Scheffen, D. G. Marchal, T. Beneyton, S. K. Schuller, M. Klose and C. Diehl, *et al.*, A new-to-nature carboxylation module to improve natural and synthetic CO_2 fixation, *Nat. Catal.*, 2021, **4**(2), 105–115.

125 S.-S. Liu, X. Wei, Q. Ji, X. Xin, B. Jiang and J. Liu, A facile and efficient transposon mutagenesis method for generation of multi-codon deletions in protein sequences, *J. Biotechnol.*, 2016, **227**, 27–34.

126 F. Hatahet, J. L. Blazyk, E. Martineau, E. Mandela, Y. Zhao and R. E. Campbell, *et al.*, Altered Escherichia coli membrane protein assembly machinery allows proper membrane assembly of eukaryotic protein vitamin K epoxide reductase, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**(49), 15184–15189.

127 O. Selifonova, F. Valle and V. Schellenberger, Rapid evolution of novel traits in microorganisms, *Appl. Environ. Microbiol.*, 2001, **67**(8), 3645–3649.

128 Z. Zhong, B. G. Wong, A. Ravikumar, G. A. Arzumanyan, A. S. Khalil and C. C. Liu, Automated continuous evolution of proteins in vivo, *ACS Synth. Biol.*, 2020, **9**(6), 1270–1276.

129 F. C. Christians, L. Scapozza, A. Crameri, G. Folkers and W. P. Stemmer, Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling, *Nat. Biotechnol.*, 1999, **17**(3), 259–264.

130 A. J. Burke, S. L. Lovelock, A. Frese, R. Crawshaw, M. Ortmayer and M. Dunstan, *et al.*, Design and evolution of an enzyme with a non-canonical organocatalytic mechanism, *Nature*, 2019, **570**(7760), 219–223.

131 D. K. Tiwari, Y. Arai, M. Yamanaka, T. Matsuda, M. Agetsuma and M. Nakano, *et al.*, A fast- and positively photoswitchable fluorescent protein for ultralow-laser-power RESOLFT nanoscopy, *Nat. Methods*, 2015, **12**(6), 515–518.

132 G. Raghunathan and A. Marx, Identification of Thermus aquaticus DNA polymerase variants with increased mismatch discrimination and reverse transcriptase activity from a smart enzyme mutant library, *Sci. Rep.*, 2019, **9**(1), 590.

133 M. L. Gerth and S. Lutz, Non-homologous recombination of deoxyribonucleoside kinases from human and Drosophila melanogaster yields human-like enzymes with novel activities, *J. Mol. Biol.*, 2007, **370**(4), 742–751.

134 J. Viña-Gonzalez, D. Jimenez-Lalana, F. Sancho, A. Serrano, A. T. Martinez and V. Guallar, *et al.*, Structure-guided evolution of aryl alcohol oxidase from Pleurotus eryngii for the selective oxidation of secondary benzyl alcohols, *Adv. Synth. Catal.*, 2019, **361**(11), 2514–2525.

135 Y. Wikmark, M. Svedendahl Humble and J.-E. Bäckvall, Combinatorial library based engineering of Candida antarctica lipase A for enantioselective transacylation of sec-alcohols in organic solvent, *Angew. Chem., Int. Ed.*, 2015, **54**(14), 4284–4288.

136 V. Brissos, M. Ferreira, G. Grass and L. O. Martins, Turning a hyperthermostable metallo-oxidase into a laccase by directed evolution, *ACS Catal.*, 2015, **5**(8), 4932–4941.

137 J. E. Barrick and R. W. Roberts, Sequence analysis of an artificial family of RNA-binding peptides, *Protein Sci.*, 2002, **11**(11), 2688–2696.

138 A. D. Keefe and J. W. Szostak, Functional proteins from a random-sequence library, *Nature*, 2001, **410**(6829), 715–718.

139 T. van Rossum, A. Muras, M. J. J. Baur, S. C. A. Creutzburg, J. van der Oost and S. W. M. Kengen, A growth- and bioluminescence-based bioreporter for thein vivo-detection of novel biocatalysts, *Microb. Biotechnol.*, 2017, **10**(3), 625–641.