Hot Spots-Making Directed Evolution Easier

Haoran Yu[ab,*], Shuang Ma[ab], Yiwen Li[c], Paul A. Dalby[c,*]

[a] Institute of Bioengineering, College of Chemical and Biological Engineering, Zhejiang University, Hangzhou 310027, China

[b] Hangzhou Global Scientific and Technological Innovation Centre, Zhejiang University, Hangzhou 311200, China

[c] Advanced Centre for Biochemical Engineering, Department of Biochemical Engineering, University College London, Torrington Place, London, WC1E 7JE, UK

[*]Corresponding authors

Haoran Yu email address: yuhaoran@zju.edu.cn

Paul Dalby email address: p.dalby@ucl.ac.uk

Abstract

Directed evolution has emerged as a powerful strategy to engineer various properties of proteins. Traditional methods to construct libraries such as error-prone PCR and DNA shuffling commonly produce large, relatively inefficient libraries. In the absence of a high-throughput screening method, searching such libraries is time-consuming, laborious and costly. On the other hand, targeted mutagenesis guided by structure or sequence information has become a popular way to produce so-called smart libraries. With an increased ratio of advantageous to deleterious mutations, smart libraries increase the efficiency of directed evolution, provided that target site prediction is reliable. Mutation target site or hot spot prediction is critical to the quality of libraries and the performance of directed evolution. Appropriate selection of hot spots enables the generation of proteins with desired properties efficiently and rationally. Here, we give an overview of seven kinds of hot spots that are divided into two categories: sequence-based hot spots including CbD (conserved but different) sites and coevolving residues, and then 3D structure-based hot spots including active-site residues, access tunnel sites, flexible sites, distal sites coupled to active center, and interface sites. This review also covers the latest advances in computational tools for identifying these hot spots and many successful cases using them for enzyme engineering.

Keywords： directed evolution; enzyme engineering; hot spots; target mutagenesis

1. Introduction

Natural evolution has generated a large number of proteins which can be harnessed for various applications in biotechnology and pharmaceutical science (Lane and Seelig, 2014). Protein engineering technology is frequently used to alter and improve proteins for specific applications. Directed evolution has emerged as a powerful strategy to engineer known properties and design or optimize proteins with functions not encountered in nature (Chen and Arnold, 2020). Traditional directed evolution consists of iterative cycles of library construction using random mutagenesis and high-throughput screening for specific features (Dalby, 2011). There are some potential drawbacks to this approach. In randomly mutated libraries, 60–70% of mutations are deleterious, 30–40% are neutral, and less than 5% of mutations give functional gains (Goldsmith and Tawfik, 2013). Traditional library construction procedures, such as error-prone PCR and DNA shuffling produce large, relatively inefficient libraries. In the absence of a high-throughput screening method, searching such libraries is time-consuming, laborious and costly (Yang et al., 2019).

Over the last few decades, the number of protein structures has been gradually increasing. The overall number of PDB entries has grown from under 400 at the beginning of 1990 to over 182,418 presently. AlphaFold2 greatly improved the accuracy of protein structure prediction, and the AlphaFold protein structure database is expanding with presently over 365,198 predicted structures from more than 20 key organisms (Jumper et al., 2021, Tunyasuvunakool et al., 2021). However, our detailed understanding of the relationships between structure and function of proteins is still incomplete. We cannot even predict the effect of a single mutation in a single protein with confidence. Fortunately, our rudimentary understanding of structure-function relationship of proteins has helped to make directed evolution easier.

Using protein structure or sequence information to guide targeted mutagenesis has become a popular method to produce so-called smart libraries (Sebestova et al., 2014, Jochens and Bornscheuer, 2010). This strategy is less likely to disrupt the global protein fold and increases the probability of obtaining active mutants. Focusing on specific amino-acid locations minimizes the size of created libraries and hence improves directed evolution efficiency, assuming that the target site prediction is accurate (Sinha and Shukla, 2019). Such libraries can yield results that are comparable to several rounds of conventional directed evolution, while the strategies for constructing smart libraries in one enzyme can often be applied to other enzymes.

A prerequisite for designing smart libraries is to select appropriate mutagenesis targets or hot spots. This review aims to discuss the commonly used techniques for defining hot spots, and how they are used for engineering enzymes more rationally. Based on the required structure information, we divide seven kinds of hot spots into two types: sequence based and structure based (Table 1). Guided by different hot spots, different enzyme properties can be engineered, and corresponding examples are provided.

Table 1 Hot spots for constructing smart libraries.

	Hot spots	**Properties**	**Comments**	**Selected references**
Sequence-based hot spots	CbD sites [a]	Stability	Most commonly used rational design strategy when 3D-structures are not available.	(Gómez et al., 2020, Sternke et al., 2019)
		Activity	Assumes that homologous proteins with better activity exist.	(Wu et al., 2014, Motoyama et al., 2020)
		Enantioselectivity	Assumes that homologous proteins with better enantioselectivity exist.	(Godinho et al., 2012, Wang et al., 2020a)

	Coevolving residues	Stability	Combinational coevolving-site saturation mutagenesis (CCSA) approach has been developed.	(Liu et al., 2021, Chang et al., 2021)
		Activity	Has the potential to modulate activity over a very large range.	(Wang et al., 2020b)
Structure-based hot spots	Active-site residues	Stability	Rigidifying flexible active-site residues has been proven useful in enhancing kinetic stability.	(Xie et al., 2014)
		Activity and substrate specificity	Well-explored strategy which has been proven useful in many cases.	(Ranoux and Hanefeld, 2013, Hailes et al., 2013)
		Enantioselectivity	Well-explored strategy. Importance of computation design has been showed recently in this area.	(Wijma et al., 2015, Zheng et al., 2021)
	Access tunnel sites	Stability	Mutants with enhanced stability showed preference to appear at access tunnel sites.	(Gihaz et al., 2018, Stimple et al., 2020)
		Activity	Useful for engineering enzymes with access tunnels.	(Meng et al., 2021, Bata et al., 2021)
	Flexible sites	Stability	Recently established strategy, consisting of two steps: identifying flexible sites then rigidifying them.	(Zhu et al., 2021, Liu et al., 2018)
		Activity	Still under exploration. Few successful cases have been reported.	(Kazuyo et al., 2014, Saavedra et al., 2018, Reetz and Carballeira, 2007)
	Distal sites coupled to active center	Activity	A strategy indicating the importance of dynamic correlations in enzyme engineering	(Yu and Dalby, 2018a, Yu and Dalby, 2018b)
	Interface sites	Stability	A useful approach to engineer multimeric enzymes.	(Bosshart et al., 2013, Basu and Sen, 2013)

[a] CbD sites: Positions that are conserved in the multiple sequence alignments but different in the sequence of the target protein.

2. Sequence-based hot spots

2.1 CbD (Conserved but Different) sites

The advent of efficient low-cost sequencing technologies has resulted in a massive growth in the number of sequences available in key protein sequence databases like UniProt, which currently contains over 177 million sequences (Gligorijević et al., 2021). Based on the assumption that sequence similarity implies functional similarity, exploiting statistical amino acid frequencies from multiple sequence alignments (MSAs) has been widely used in protein engineering, especially in cases where the structural data is of limited accuracy, or where the mechanism of substrate recognition is not fully understood. One of the purposes carrying out MSAs is to identify the positions that are conserved in the pool of sequences but different in the target protein sequence. To better discuss the MSAs method used for protein engineering, we name these positions as CbD (conserved but different) sites (Figure 1). The WW domain is one of the smallest protein modules, found in a number of unrelated signalling and structural proteins, which mediates specific protein-protein interactions. If we consider the WW domain of zygote-specific protein 3 (ZYS3) from *Chlamydomonas reinhardtii* as the target, its CbD sites are the amino acids different from those in the consensus sequence, such as those at positions 1 to 6 shown (Figure 1) (Porebski and Buckle, 2016). Once CbD sites are identified, desired properties can be obtained through mutating original residues at CbD sites to the equivalent consensus residues. Two things normally need to be considered before applying MSAs to identify CbD sites for protein engineering: (i) whether homologous structures with desired properties exist for the target enzyme; (ii) whether the regions

determining those desired properties have been identified in the target protein. Comparing the sequence of the target enzyme with homologous structures with desired properties is a straightforward way to identify key sites for mutagenesis. When this condition is not met, and the local region controlling the desired properties is known, CbD sites can be easily identified by comparing sequences of that region among homologous structures. For example, the catalytic properties of enzymes are significantly controlled by the active-site region. Thus, when homologous enzymes with superior activity exist, hot spots for engineering activity could be identified by comparing the catalytic regions of two enzymes (Wu et al., 2014). Of course, MSAs are still very useful when the two key conditions are not met. With CbD sites as hot spots, numerous enzymes have been engineered to have improved properties, including stability, activity, and enantioselectivity (Sternke et al., 2019, Motoyama et al., 2020, Wang et al., 2020a, Sumbalova et al., 2018).

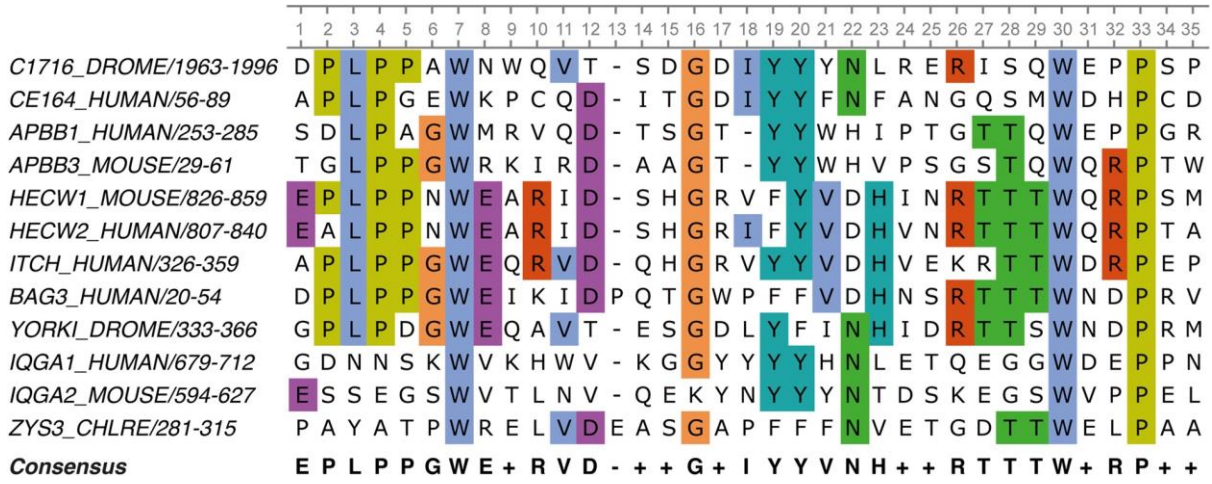

Figure 1 Sequence alignment of 12 WW domains across several species and parent proteins. In the consensus, a '–' is a gap, whilst a '+' is an ambiguous position with no consensus. The most conserved residues are highlighted (Porebski and Buckle, 2016). Reprinted with permission, copyright 2016 Oxford University Press.

2.1.1 Stability

When limited structure data were available, MSAs have been a useful method to guide the engineering of stability into enzymes. Amino acids that appear most frequently at a specific position among homologous structures tend to contribute more to the stability than other residues at the same position. Based on this assumption, the “consensus design” approach was first used by Steipe *et al*. to stabilize an antibody domain through substituting ten residues with consensus amino acids (Steipe et al., 1994). Following this study, many researchers have applied the “back to the consensus mutations” approach to engineer the stability of various enzymes such as glucose dehydrogenase (Vazquez-Figueroa et al., 2007), endoglucanase (Anbar et al., 2012), amylase (Ranjani et al., 2014), xylanase (Han et al., 2017), laccase (Gómez et al., 2020) and so forth. Consensus sequence design has been proved as a general strategy to create stable and biologically active proteins (Sternke et al., 2019). After designing and characterizing consensus sequences for six unrelated protein families, Sternke *et al*. found that consensus design showed high success rates in creating well-folded, hyperstable proteins and retaining their biological activities. More importantly, these consensus proteins showed higher stability than the naturally occurring sequences of their respective protein families, highlighting the utility of consensus sequence design.

In addition, ancestral sequence reconstruction (ASR) has emerged as a useful methodology for engineering enzymes with enhanced stability, heterologous expression, activity, or unique activity profiles (Spence et al., 2021). Ancestral sequences are reconstructed by inferring a phylogenetic relationship between homologous sequences and applying a statistical model of amino acid substitution to calculate sequences at internal nodes of the phylogenetic tree (Hall, 2006). ASR is different from other enzyme engineering methods as the new sequences are generated based upon probabilistic searches of non-conserved functional space, giving each generated sequence a high likelihood of being functional as long as an accurate multiple

sequence alignment input is provided (Thomas et al., 2019). Recently, to address the poor catalytic performance of the existing β-1,3-xylanase, its ancestor sequence named AncXyl09 was reconstructed using an optimized ancestral sequence reconstruction strategy and the generated sequence showed an excellent thermostability with a half-life of 65.08 h at 50 ^{o}C (Zeng et al., 2021).

2.1.2 Catalytic activity

CbD sites have been used as hot spots to engineer the catalytic activity of enzymes as well. *Paenibacillus barcinonensis* esterase (EstA) converts tertiary alcohol esters with limited activity and enantioselectivity. However, its homologous enzymes, two large groups of esterases and lipases can convert tertiary alcohols efficiently. MSAs of 1343 sequences revealed that in the oxyanion hole, all these enzymes contain a highly conserved motif with the sequence of GGG(A), in addition to EstA whose third position is a serine. Thus, the mutant EstA-GGG was constructed. As expected, this mutant showed 26-fold faster conversion of tertiary alcohols than the wild type (Bassegoda et al., 2010). Additionally, *Escherichia coli* phytase (EcAppA) was engineered to have improved activity and thermostability guided by sequence alignment (Wu et al., 2014). EcAppA has two homologous structures, *Citrobacter braakii* (CbAppA) and *Citrobacter amalonaticus* (CaAppA), which have 60.6% and 57.1% protein sequence identity with EcAppA and show superior specific activity to EcAppA. The alignment of these three sequences revealed four EcAppA-unique residues around the phytate-binding pocket and a variable loop region. Mutating these hot spots to the consensus residues led to a variant showing a 17.5% increase in the specific activity. Threonine dehydrogenase (TDH) has also been engineered using the consensus design method to have unique enzymatic properties. Five artificial TDHs have been designed by full consensus

protein design (FCD) utilizing sequences selected from databases, and four of them were successfully expressed in soluble form, but with various catalytic properties (Motoyama et al., 2020).

2.1.3 Enantioselectivity

Comparison of homologous sequences can also be applied as a guide for engineering the enantioselectivity of enzymes. In some circumstances, homologous enzymes showed different properties, and sequence comparison could be a method to engineer the enzymes to have a desired target property. Two carboxylesterases CesA and CesB from *B. subtilis 168* shared 60% sequence identity and showed very different enantioselectivity towards substrate 1,2-*O*-isopropylideneglycerol (IPG). Sequence alignment was used to identify sites that might lead to the difference in enantioselectivity of these two enzymes, and it was found that most active site residues are conserved in both CesA and CesB, with the exception of positions 166 and 182. These two residues in CesA were then identified as a target for site-directed or saturation mutagenesis to enhance its enantioselectivity. A CesA double mutant F166V/F182C was generated, which showed a 13-fold increased enantioselectivity and without significant activity loss compared to wild type (Godinho et al., 2012). An epoxide hydrolase from *Phaseolus vulgaris* (*Pv*EH2) was also engineered to significantly increase enantioselectivity. A cap-loop of *Pv*EH2 was speculated to be relevant to EH's catalytic properties based on previous studies and was carried out for sequence alignment with four EHs including *St*EH, *Pv*EH1, *Vr*EH1 and *Vr*EH2 from *Solanum tuberosum*, *Phaseolus vulgaris* and *Vigna radiata,* respectively. As a result, the differences of their cap-loops mainly focused on their non-conserved middle segments (^{190}EGMGSNLNTSMP^{201} in *Pv*EH2), in which the residue composition and chain length clearly varied. By replacing this variable cap loop in PvEH2 with

the corresponding fragments in the homologous enzymes, four PvEH2 variants including Pv2St, Pv2Pv1, Pv2Vr1 and Pv2Vr2 were designed respectively. Utilizing rac-1,2-epoxyhexane as the target substrate, hybrid enzyme *Pv2St* showed the highest E value with 11.5-fold improvement compared to *Pv*EH2 (Wang et al., 2020a).

2.2 Coevolving residues

Multiple sequence alignments (MSAs) allow for identifying residues that are completely conserved, partially conserved, or non-conserved. Completely conserved residues are functionally important as they did not change through random meanderings of evolutionary change. Some other positions are conserved only within subfamilies, but are different between them, which play important roles in functional diversity of homologs. These positions are so-called the specificity determining positions/residues/sites (SDPs/SDRs/SDSs) or subfamily/family-specific positions (SSPs/FSPs) (Figure 2) (Suplatov et al., 2020, de Juan et al., 2013, Chagoyen et al., 2016). Such positions are helpful to understand how enzymes perform their natural functions, and can also be selected as hotspots for protein engineering experiments. In addition, in evolutionary processes, pairs of residues that are mutually proximate in the tertiary structure often coevolve to maintain their structure. For example, when one becomes larger, the other becomes smaller. And, when one becomes a positively charged residue, the other becomes a negatively charged residue. Such evolutionary couplings provide accurate information about residue pair contacts, important to protein three-dimensional structure prediction (Ovchinnikov et al., 2017, Anishchenko et al., 2017, Yang et al., 2020). In addition, these evolutionary couplings could also control the adaptation of proteins to natural environments through maintaining overall structural-functional integrity while fine-tuning the function of proteins, including catalytic activity, substrate

selectivity, and tolerance to unusual environmental conditions. These residues are called coevolving residues (Figure 2) (de Juan et al., 2013). Coevolving residues can be much more important to the stability and folding of proteins than other residues. Choosing coevolving residues as hot spots for protein engineering has the potential to introduce novel functions into proteins, or to create radical changes in stability. In recent years, much attention has been paid to develop tools for predicting coevolving residues (de Juan et al., 2013, Dickson and Gloor, 2014, Sumbalova et al., 2018) and several successful engineering examples taking coevolution into account have been reported.

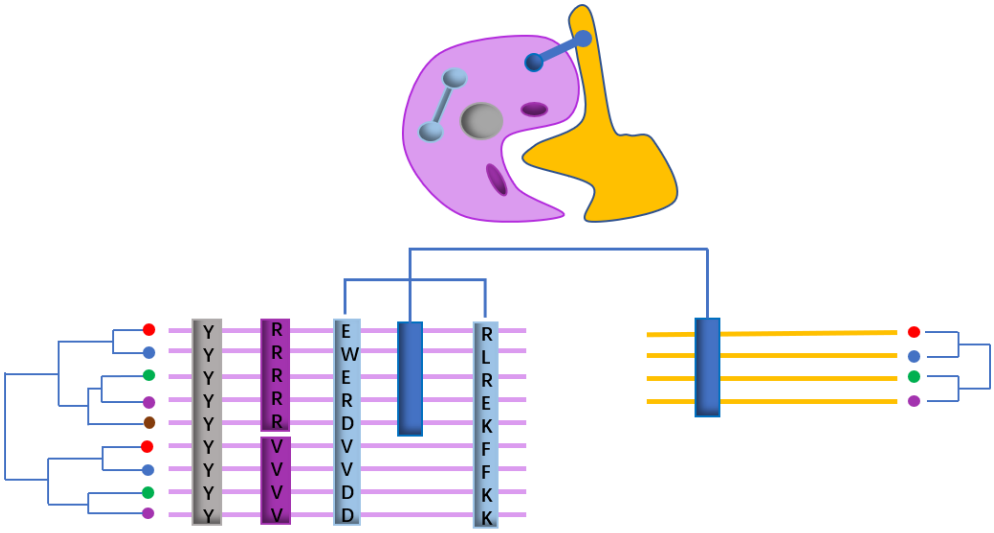

Figure 2 Coevolutionary features extracted from protein multiple sequence alignments. The three-dimensional structures of two interacting proteins (purple and yellow) are schematized as well as their MSAs and phylogenetic trees based on orthologous sequences from a number of organisms. Circles of different colours represent different species from which the protein sequences are derived. Intra-protein coevolving residues (light blue) are related to residue spatial proximity, whereas inter-protein coevolving residues (dark blue) reflect in many cases proximity between residues in different protein chains. Fully conserved positions (grey) tend to form a part of the protein core and are also in functional regions (such as protein interaction sites and catalytic sites). Specificity-determining positions (SDPs; purple) tend to be in functional sites conferring specificity. Reprinted with permission, copyright 2000 Springer Nature BV.

2.2.1 Activity

Coevolving residues provided an efficient strategy for engineering the catalytic activity of enzymes. The pullulanase from *Bacillus naganoensis* has been engineered to have improved catalytic activity by an evolutionary coupling saturation mutagenesis (ECSM) strategy. This strategy identified residues for saturation mutagenesis by calculating the covariance of residue pairs. Seven residue pairs were selected as evolutionary mutation hotspots, and none of these sites were located at or closed to the active sites. The best-performing quadruple mutant K631V/Q597K/D541I/D473E obtained after mutagenesis and screening showed a 3.0-fold increase of k_{cat} and 6.3-fold increase of k_{cat}/K_m relative to the WT enzyme, demonstrating that coevolutionary analysis can identify distal sites that affect enzyme activity (Wang et al., 2020b). It has also been reported that enzyme activity could be modulated over a 100-fold range by mutating coevolving residues (McMurrough et al., 2014). McMurrough *et al.* identified a coevolving network consisting of two catalytic metal-binding residues (Asp and Glu) and two adjacent noncatalytic residues (Ala and Gly) in LAGLIDADG homing endonucleases (LHEs). Saturation mutagenesis was used to mutate two non-catalytic residues, while the metal-binding residues were held at either Glu or Asp. Variants with the highest activity showed 3-fold decreased K_m while variants with the lowest activity revealed a 65-fold decreased k_{cat}. Additionally, they concluded that variants with low activity could be rescued by compensatory mutations in relative coevolving network residues, and optimization of the coevolving network could be an important consideration in engineering catalytic activity of enzymes. In another study, coevolutionary sites were proven crucial in improving the efficiency and specificity in genome editing with the CRISPR/Cas9 system (Li et al., 2019).

2.2.2 Thermostability

Coevolving residues can also be used as hot spots to engineer the thermostability of enzymes (Strafford et al 2012). Wang *et al*. have improved the thermostability of alpha-amylase by a strategy called CCSA (combinational coevolving site-saturation mutagenesis) (Wang et al., 2012). Six coevolving sites and 10 pairs of coevolutionary interactions were identified in α-amylase firstly. Each pair of residues involved in evolutionary interactions was randomly mutated to construct a library. 10,010 clones from ten libraries were screened for improved thermostability. The best mutant showed 8 ^{o}C enhanced thermostability. Additionally, they found that deleterious effects caused by disadvantageous mutation could be compensated by the covariation at the other coevolving site. The thermostability of amine transaminase was also improved by evolutionary coupling saturation mutagenesis. The Mutual Information Server to Infer Coevolution (MISTIC) (Simonetti et al., 2013) was used to predict eight residues with strong interactions in the coevolution network as the mutation targets, and subsequent alanine screening and saturation mutagenesis identified several improved variants with the best mutant F115L/L118T showing 9.55-fold improvement in half-life compared to wild type (Zhu et al., 2019, Liu et al., 2021).

Molecular dynamics (MD) simulations showed that these mutations reduced the overall flexibility of AT-ATA and this could have a stabilizing effect on the double mutant F115L/L118T (Liu et al., 2021). Using the adenylate kinase (ADK) family as a model system, Chang *et al.* improved the thermal stability of ADK by coevolution and sequence divergence analysis. The method identified a series of amino acid sites that were closely related to thermal stability. Single and double site mutants showed improved thermostability and better enzymatic activity at higher temperatures (Chang et al., 2021) .

3. Structure-based hot spots

3.1 Active-site residues

The active-site is the reaction centre which binds the substrates, eases the formation of the transition state, and then releases the products (Toscano et al., 2007). The active-site occupies nearly 10–20% of the volume of an enzyme and consists of amino acid residues that bind substrate and residues that catalyse a reaction of that substrate. For example, aminoglycoside-3'-phosphotransferase-IIa catalyses ATP-mediated phosphate transfer to chemically modify and inactivate aminoglycoside antibiotics such as kanamycin (Nurizzo et al., 2003). The residues including D159, E160, R211, D227, E230, E262, F264 involve binding substrate kanamycin and the catalytic residue D190 plays a role in the deprotonation of 3’ hydroxyl of kanamycin to allow for efficient attack on the γ-phosphate (Figure 3) (Nurizzo et al., 2003). To engineer novel enzymes, active-site residues are good starting points. Many enzymes have homologous structures, and these homologous enzymes use similar mechanisms to catalyse different reactions. It has been observed that a small change in the composition of active-site residues of homologous enzymes can contribute to their different activities. Such observations demonstrate that an inherent structural plasticity of the active site has the potential to give enzymes new functions. Chemical space in the active site of extant enzymes has not been fully explored by nature, which allows us to evolve active site residues to create novel enzymes. In recent years, numerous researchers have revealed that mutation of active-site residues often dramatically changes the properties of enzymes. Of course, most of the changes lead to deactivated mutants. However, sometimes, variants have improved or new catalytic activity (Chen and Arnold, 2020), broadened substrate scope (Hibbert et al., 2007), altered stereospecificity (Goldsmith et al., 2012), regioselectivity (Balke et al., 2017, Wang et al., 2017), enantioselectivity (Gao et al., 2018, Sandstrom et al., 2012), and even thermostability (Xie et al., 2014).

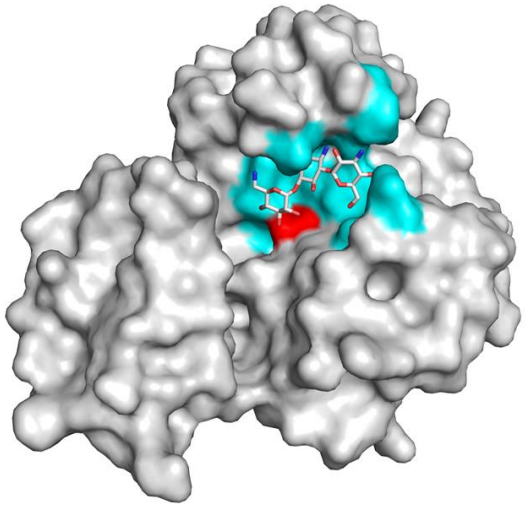

Figure 3 Surface representation of aminoglycoside-3'-phosphotransferase-IIa in complex with kanamycin (PDB ID 1ND4) (Nurizzo et al., 2003). Residues including D159, E160, R211, D227, E230, E262, F264 involved in binding substrate kanamycin (sticks) are shown in cyan. The catalytic residue D190 is shown in red, playing a role in removing a proton from O3' hydroxyl of substrate prior to phosphoryl transfer.

3.1.1 Activity and substrate specificity

A major problem in applying enzymes as catalysts in organic chemistry is that only limited substrates can be accepted. Here, the process of engineering transketolase (TK) for accepting various substrates is used to illustrate how to apply active-site residues as hot spots in directed evolution. The wild type TK catalyses the reversible transfer of a C2-ketol unit from D-xylulose-5-phosphate to either D-ribose-5-phosphate or D-erythrose-4-phosphate, linking glycolysis to the pentose phosphate pathway (Figure 4A) (Sprenger et al., 1995). Using active site residues as the hotspots, a semi-rational directed evolutionary strategy was applied to design *E. coli* TK to accept non-phosphorylated substrates. In 2007, TK was first designed to accept glycolaldehyde (GA) as a substrate by saturation mutagenesis of active site residues (Hibbert et al., 2007), which were selected in two ways: structurally defined sites and phylogenetic defined sites. Structurally defined sites were those within 4 Å of the docked substrate erythrose-4-phosphate. Natural phylogenetic diversity has also turned out to be useful in guiding directed evolution. Through phylogenetic analysis of 52 TK sequences from

bacteria, fungi, plants and trypanosomes, 10 different residues within 10 Å of the cofactor TPP were identified as phylogenetically diverse sites (Figure 4B). Finally, variants with improved activity against GA were obtained by screening saturation libraries.

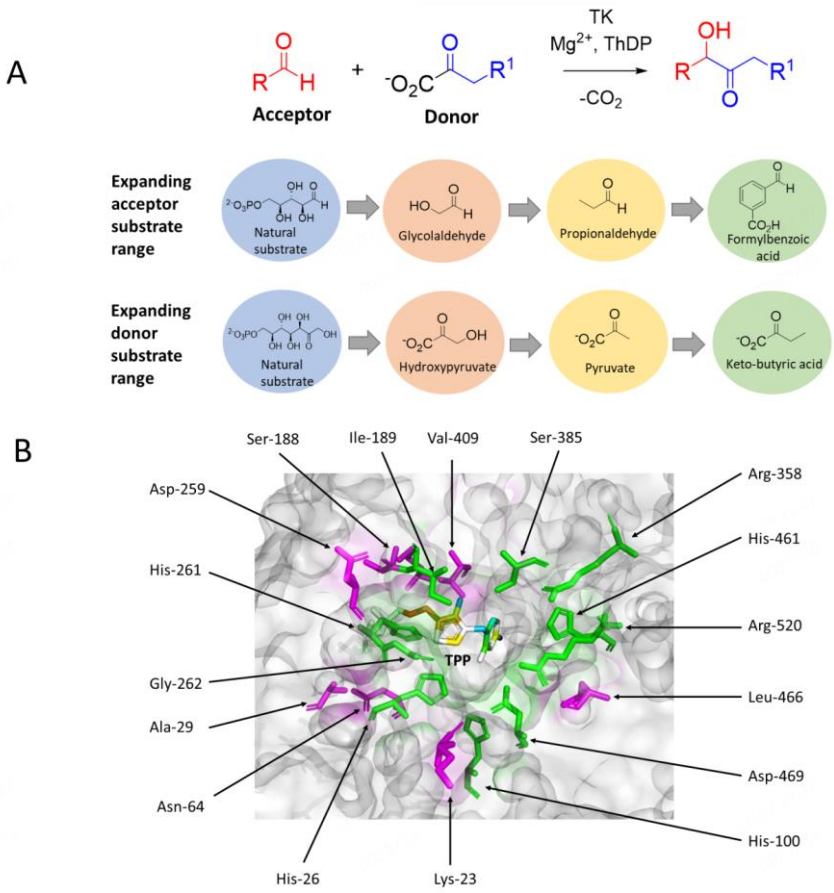

Figure 4 Engineering transketolase to expand substrate range with active sites as the mutation targets. A, Range of acceptor and donor substrates accepted with engineered *E. coli* transketolase. B, Phylogenetically defined sites (in magenta) and structurally defined sites (in green) in *E. coli* transketolase with cofactor TPP (PDB ID: 1QGD). Structurally defined sites were those within 4 Å of the docked substrate erythrose-4-phosphate. Phylogenetically diverse sites were identified within 10 Å of the cofactor TPP through phylogenetic analysis of 52 TK sequences from bacteria, fungi, plants and trypanosomes. These defined sites were applied as the mutation targets to expand the range of substrates for the transketolase.

To expand the range of acceptor substrates for TK, libraries constructed at the above active sites were further screened to identify TK mutants with better activity against the non-hydroxylated aldehyde substrate, propionaldehyde (PA) (Hibbert et al., 2008). Thirteen

mutants with enhanced activity were eventually identified out of a randomised mutagenesis library targeted phylogenetic defined sites, with D469T showing the greatest improvement of 4.9-fold relative to the wild type. By further modification of the active sites described above, TK variants were successfully screened for high activity against other non-natural substrates, including long aliphatic (Cazares et al. 2010), cyclic (Cazares et al. 2010), aromatic (Galman et al. 2010, Payongsri et al. 2012), heteroaromatic aldehyde substrates (Galman et al. 2010), polar aromatic aldehyde substrates (Panwajee Payongsria, 2015) and even novel donor substrates including pyruvate and keto-butyric acid (Figure 4A) (Yu et al., 2020). Recently, the activity of TK has been modified by the incorporation of unnatural amino acids (UAAs). With the variant S385Y/D469T/R520Q showing high activity towards unnatural substrate 3-FBA as the template, Y385 was further replaced with a series of phenylalanine derivatives to reduce aromatic ring electron density, including *p*-aminophenylalanine (*p*AMF), *p*-cyanophenylalanine (*p*CNF) and *p*-nitrophenylalanine (*p*NTF) (Wilkinson and Dalby, 2021). The results showed that the *p*CNF variant simultaneously increased the activity and stability of TK against 3-hydroxybenzaldehyde (3-HBA).

3.1.2 Enantioselectivity

Enantioselectivity is a property that allows certain enzymes to be utilised to produce enantiomerically pure chemicals for industrial applications including for agrochemicals and pharmaceuticals (Yu et al., 2021). Since most enzymes do not have perfect enantioselectivity when transforming non-natural substrates, protein engineering is often applied to adjust the enantioselectivity (Otten et al., 2010). With active-site residues as hot spots, the enantioselectivity of various enzymes has been engineered, including for esterases, lipases, cytochrome P450s and so forth, through producing small high-quality libraries (Jochens and

Bornscheuer, 2010, Bartsch et al., 2008, Reetz, 2011, Wijma et al., 2015). Recently, the so-called combinatorial active-site saturation test/iterative saturation mutagenesis (CAST/ISM) strategy has been used to engineer L-threonine aldolase (LTA) for improved C_β stereoselectivity (Zheng et al., 2021) . LTA, a 5'-phosphate (PLP)-dependent enzyme, was used to catalyse the formation of β-hydroxy-α-amino acids with two chiral centres (Figure 5). This enzyme is strictly selective for C_α of β-hydroxy-α-amino acids but moderately selective for C_β, limiting its wide use in stereospecific carbon-carbon bond synthesis. In this study, the CAST/ISM strategy was applied to build a small and smart library with the active sites as the mutation targets. At last, the RS1 variants containing the Y8H, Y31H, I143R and N305R mutations showed significant improvements in the diastereoselectivity of many other aromatic aldehydes and has the potential to be used industrially for the synthesis of high value β-hydroxy-α-amino acids. Compared to the traditional directed evolution method, this approach dramatically decreased the size of the library and achieved a similar outcome in terms of enzyme performance.

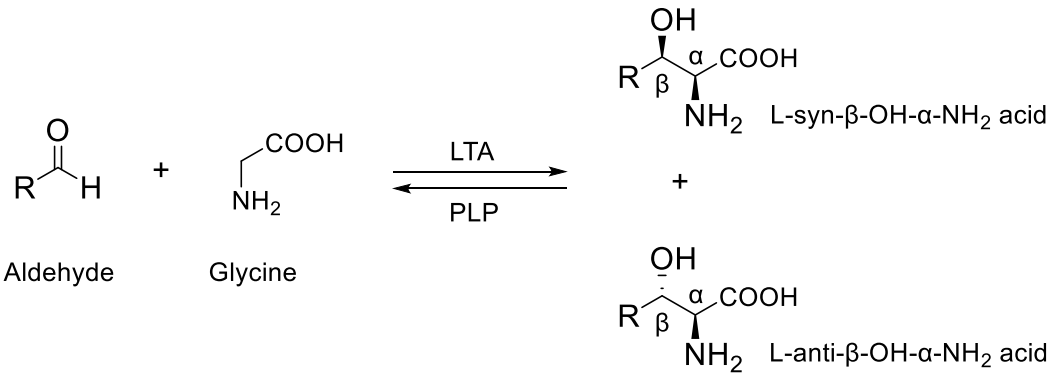

Figure 5 The reversible aldol reaction catalyzed by LTA. The β-hydroxy-α-amino acid with two chiral centers (C_α and C_β) is produced, and LTA is more selective for C_α than C_β. The CAST/ISM strategy was applied to improve its diastereoselectivity. The active sites were selected as the mutation targets to build a small and smart library, and mutations were obtained with significant improvements in the diastereoselectivity towards many aromatic aldehydes.

3.2 Access tunnel sites

As we discussed in section 3.1, the active-site residues are critical for enzyme engineering. In some enzymes, such as chymotrypsin, active sites are relatively surface exposed, while in others, such as dehydrogenases, the active sites are often deeply buried in the core of the protein (Gora et al., 2013). For enzymes with buried active sites, potential substrates must pass through the body of the protein to access the active sites. Active sites inside the protein core are often connected with the protein surface by one or more access tunnels (Kingsley and Lill, 2015). Structurally, buried active sites accessed by tunnels enhance the complexity of the ligand binding process. To address how tunnels influence the enzyme activity, reaction mechanism, specificity, and stereoselectivity, the "keyhole-lock-key" model, different from the traditional "lock and key" model, has been proposed (Figure 6A) (Kokkonen et al., 2019). In this model, an active site is represented by a lock, an access tunnel is represented by a keyhole and a substrate is represented by key. For enzymes with buried active sites, the recognition of substrate is seen as a two-step process (i) migration of substrate through access tunnel and subsequently (ii) substrate binding in the active site. By deconstructing substrate recognition in this way, it is known that before complementarity between substrate and active site, there must be a complementarity between the substrates and the access tunnel. The tunnel itself is hence also responsible for the substrate specificity, and this has been observed in many enzymes such as haloalkane dehalogenases (Kokkonen et al., 2021), aldehyde-deformylating oxygenase (Bao et al., 2016), cytochrome P450s (Cojocaru et al., 2007), and fructosyl peptide oxidase (Rigoldi et al., 2020). Theoretically, it is possible to modify enzyme properties by altering the substrate access tunnels. Substitutions in these access tunnel sites do not disrupt the active-site architecture and have the potential to generate high yields of functional variants.

In many studies, the replacement of residues located at access tunnels has provided impressive improvements in valuable enzyme properties including reaction mechanism (Biedermannova et al., 2012), resistance to substrate inhibition (Kokkonen et al., 2021) , activity (Brouk et al., 2010, Panizza et al., 2015, Jung et al., 2018, Marques et al., 2017b, Kong et al., 2014, Luan et al., 2015), substrate specificity (Bao et al., 2016, Kaushik et al., 2018), stability (Koudelakova et al., 2013) and enantioselectivity (Liskova et al., 2017). Enzyme tunnels could also be the potential targets for designing new biocatalysts, materials or drugs (Jurcik et al., 2018, Marques et al., 2017a) .

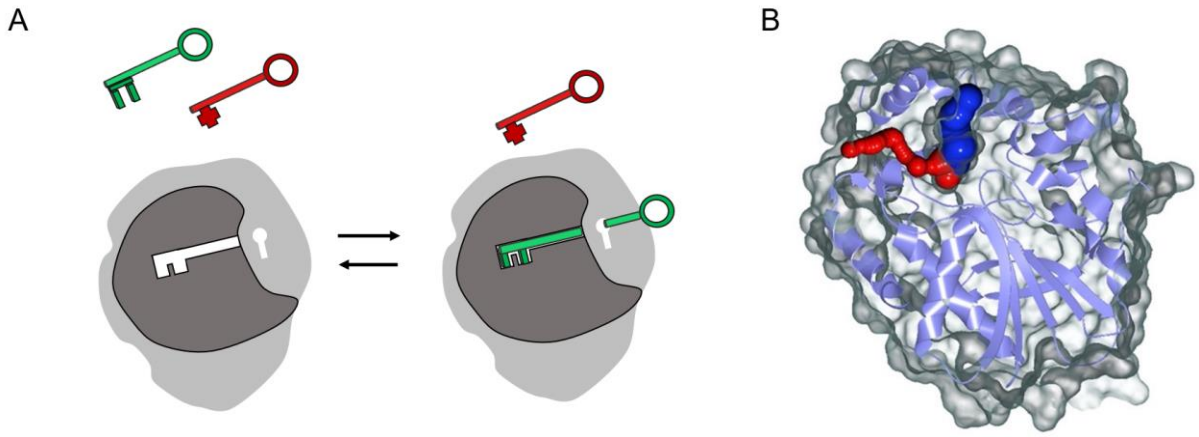

Figure 6 Access tunnel sites of enzymes. A, Keyhole-lock-key model for enzymatic catalysis. Two-step process composed of a passage of substrate (key) via the tunnel (keyhole) and molecular recognition in the active site (lock). B, The two substrate tunnels of haloalkane dehalogenase (Kokkonen et al., 2019). Reprinted with permission, copyright 2019 Elsevier.

3.2.1 Activity

The most pronounced example of improving activity by modifying enzyme tunnels has been the process of engineering the haloalkane dehalogenase DhaA (Figure 6B). DhaA cleaves carbon-halogen bonds by a hydrolytic mechanism to yield the corresponding alcohol, a proton, and a halide, and has potential applications in various fields. DhaA was firstly evolved to have improved activity to convert a toxic, non-natural compound 1,2,3-trichloropropane

(TCP) to the less toxic 2,3-dichloropropan-1-ol (DCP) (Pavlova et al., 2009). In the study, three residues located in the access tunnel were chosen as hot spots for saturation mutagenesis. After screening around 5000 clones, 51 positive clones and 25 unique sequences with enhanced activity were obtained. The best DhaA mutant showed 26-fold higher catalytic activity than wild type. After this study, the same group found that introducing bulky amino acids into the substrate channel can also increase enzyme activity (Marques et al., 2017b). Pavlova et al. obtained a variant containing five mutations (I135F/C176Y/V245F/L246I/Y273F) through targeted evolution of a molecular channel of haloalkane dehalogenase, which increased the activity against the novel substrate TCP by 32-fold. Tertiary structural analysis indicated that the large volume of amino acids in substrate channel were fundamental for positioning of substrate TCP in the reactive conformations and increasing the productive binding of substrate in the enzyme.

Recently, a study showed that the substrate preference of cytochrome $P450_{Bs\beta}$ from *Bacillus subtilis* could be modulated by tunnel engineering strategies. $P450_{Bs\beta}$ showed low decarboxylase activity towards long-chain fatty acids, and enlarging the access tunnel in the variants F79A and F173V gave a 15.2-fold and a 3.9-fold increase in conversion of palmitic acid and pentadecanoic acid, respectively (Meng et al., 2021). In addition, by engineering the substrate access tunnel, an (*R*)-aminomutase was converted to a highly selective (*S*)-ammonia lyase (Bata et al., 2021). Furthermore, it has been shown that the geometry of the tunnels may have different effects on the binding and catalysis of different ligands. The design of enzyme tunnels must take into account not only geometry, kinetics and physicochemical properties but also how mutations may affect key steps in the catalytic cycle (Kaushik et al., 2018).

3.2.2 Stability

The DhaA was also engineered to have improved stability by modifying the access tunnel sites (Koudelakova et al., 2013). The stability of DhaA was engineered through error-prone PCR, and it was found that the structural stabilization of highly stable variants essentially came from four mutations, T148L, G171Q, A172V, and C176F in the access tunnel and their effects were additive. The variant DhaA80 containing only these four mutations exhibited 4000-fold higher kinetic stability in 40 vol.% DMSO and 17 ^{o}C enhanced melting temperature relative to wild type. However, a stability-activity trade-off was observed with the activity towards 1,2-dibromoethane having been reduced by two orders of magnitude in the DhaA80 variant. Recently, this issue has been addressed by fine-tuning access tunnel sites (Liskova et al., 2015). In this study, libraries were constructed with two access tunnel sites V172, F176 as targets. After screening 236 colonies in these two libraries, the hit with the greatest improvement in activity relative to the template was obtained, F176G. The catalytic activity of this mutant towards 1,2-dibromoethane was 32-times higher than that of DhaA80, and its melting temperature was only 4 ^{o}C lower. Access tunnels have shown a huge potential as hot spots to improve stability and even balance the activity and stability of enzymes (Stimple et al., 2020). Access tunnel sites have also been applied as mutation targets to engineer thermostability or resistance to organic solvent for enzymes including for lipase (Gihaz et al., 2018), and esterase (Singh et al., 2017). A detailed discussion about the enzyme properties modified by engineering access tunnels can be found in the review article from Damborsky et al (Kokkonen et al., 2019).

3.3 Flexible sites

Flexibility is an important structural property determining protein functions such as protein-protein interactions, ligand binding, allosteric regulation and signal transduction. Conformational changes are also frequently observed as part of enzyme mechanisms, and protein motions are hence critical for enzymatic function (Nestl and Hauer, 2014). On the other hand, highly flexible regions are often located on the surface loops as showed by the flexibility analysis of Calf-1 domain of integrin $\alpha_{IIb}\beta_3$ which mediates platelet aggregation and thrombus formation (Figure 7). The flexible regions have a relatively low number of contacts with other amino acids and large thermal fluctuation of flexible regions might expose the hydrophobic core of protein to water penetration, triggering protein unfolding. The assumption that rigidity is the prerequisite for high thermostability has been supported by studies that compare flexibility in mesophilic and thermophilic proteins (Paredes et al., 2011, Mamonova et al., 2013, Reetz et al., 2006). Although the relationship between flexibility, activity and stability is complex, flexible sites can be chosen as hot spots for guiding protein engineering of different properties.

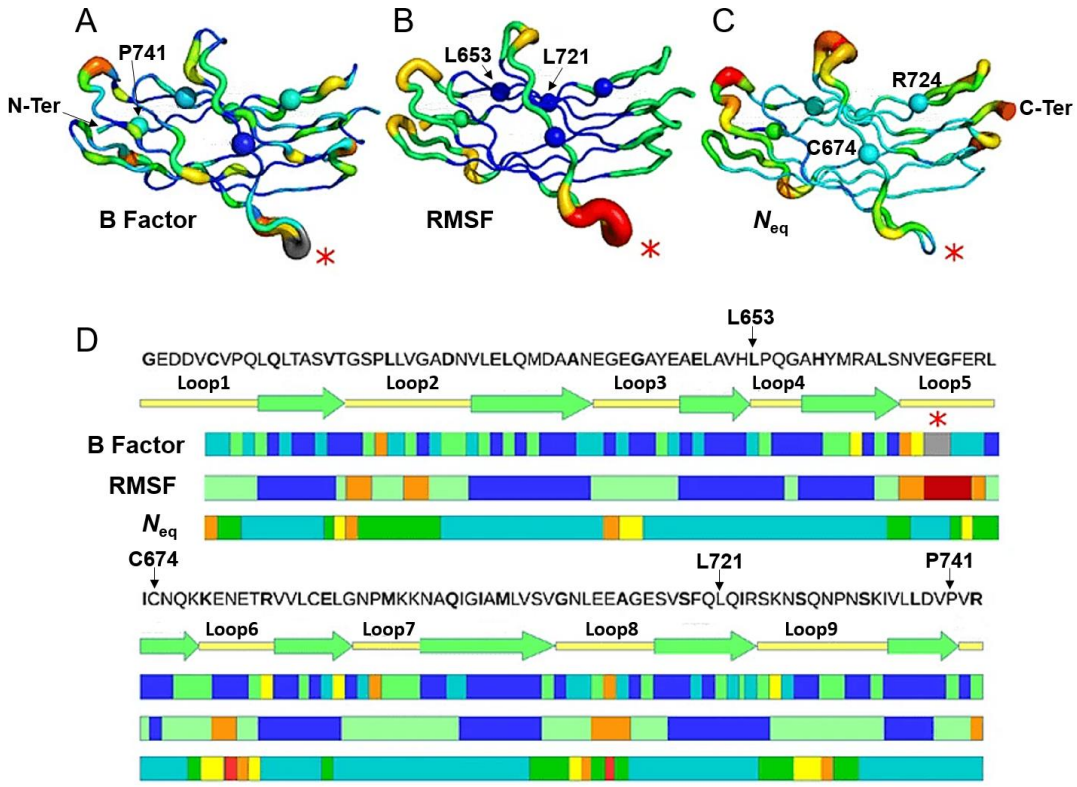

Figure 7 Comparison of the protein flexibility of Calf-1 through different metrics. 3D structures of Calf-1 domain represented through (A) B-factor values, (B) RMSF values, and (C) N_{eq} values. Local structure is ranked from rigid (thin blue line, a value of 0.0) to flexible (thick red line, a value of 4.0). Residues with completed missing atoms are in grey in the B-factor cartoon (A). D, The Calf-1 amino acid sequence is placed in regard to its secondary structures assignment and to protein flexibility according to the B-factor, the RMSF or the N_{eq} values. Blue, green, yellow, orange and red colours scale the structure from rigid to flexible (Goguet et al., 2017). Reprinted with permission, copyright 2017 Springer Nature.

3.3.1 Stability

B-factor is commonly used to represent the flexibility and the iterative saturation mutagenesis on the basis of B-factor (B-FIT approach) has been a greatly useful strategy to improve the thermal stability of proteins(Blum et al., 2012, Reetz et al., 2006). In addition to the use of B-factors, root mean square fluctuations (RMSF) and equivalent number of protein blocks (N_{eq}) calculated in MD simulations could be used to represent protein flexibility (Figure 7). The detailed discussion about computational tools to predict protein flexibility is provided

in the section 4.5. Hydrogen–deuterium exchange mass spectrometry can also be used to experimentally investigate the flexibility.

There are also many other methods available for rigidifying flexible sites such as structure-guided consensus mutagenesis, the introduction of prolines or disulfide bridges, or the addition of salt bridges, which have been reviewed previously (Yu and Huang, 2014). Two parallel strategies have been applied to identify mutation candidates within the flexible loops of *Escherichia coli* transketolase (Yu et al., 2017). The first was a 'back to consensus mutations' approach, and the second was computational design based on ΔΔ*G* calculations in Rosetta. After an experimental characterization, three single-mutant variants I189H, A282P, D143K were found to be more thermostable than wild-type TK. Additionally, thermostable enzymes usually have deletions of exposed loop regions found in their mesophilic homologs, as it reduces the flexibility and, therefore, inherent entropy in the protein structure. Thus, deleting or shortening dynamic loops could be one way to enhance the thermostability of mesophilic proteins. Residues 78-90 in porcine trypsin were predicted by molecular dynamic simulations, FlexPred and FoldUnfold, to be a highly flexible region. By truncating this region, the variant D9 exhibited higher thermal stability, with a 5 °C increase in T_{opt}, a 5.8 °C increase in T^{10}_{50} and a 4.5 °C increase in T_m compared to the wild type (Liu et al., 2018). Fang *et al*. have improved the thermostability of bacterial laccase Lac 15 by deleting the residues step by step (Fang et al., 2014). From the crystal structure of Lac15-His_6, they noticed that a few regions, including the C-terminal His-tag, are invisible in the electron density map, indicating high flexibility of these areas. When residues (323-332) were deleted from Lac 15, a variant Lac15D was obtained which exhibited significantly improved thermostability and extraordinary solubility. The strategy of rigidifying flexible sites has also been applied to many other enzymes including

chondroitinase ABC (Kheirollahi et al., 2017), esterase (Zhu et al., 2021), artificial metalloenzymes (Obrecht et al., 2021) and so forth.

3.3.2 Function

Although flexible regions have shown a promising role in guiding the evolution of enzyme stability, it is still difficult to engineer functional properties taking flexibility into account. Due to the so-called activity–stability trade-off, modifying them risks a negative correlation between enzyme stability and activity. Several studies attempted to explore the possibility of applying flexibility modulation as a means to enhance enzyme activity. Young Je Yoo's group attempted to enhance the activity of xylanase from *Bacillus circulans* (Kazuyo et al., 2014). Hinge regions between moving subunits were firstly identified using the computational tool PiSQRD. Then, in the hinge regions, four target residues including Val131, Arg132, Asn141, and Ala142 were mutated to several amino acids with different flexibility. In the results, it was observed that mutants with increased rigidity showed increased catalytic activity while mutants with decreased rigidity showed decreased catalytic activity (Kazuyo et al., 2014, Hong et al., 2014).

The investigation of the relationship between structural dynamics and enzyme catalysis is of increasing interest. In another study, MSA and MD simulations were performed for cellulase Cel5A from *Bacillus agaradherans*. Three specific positions were selected based on the analysis of hydrogen bond patterns between residues within the active site. After site-directed mutagenesis, Cel5A variants showed a concomitant increase in the catalytic activity at low temperatures and a decrease in activation energy and activation enthalpy, similar to cold-active enzymes, indicating that disrupting a hydrogen bond network in the vicinity of the active site increases local flexibility (Saavedra et al., 2018).

The impact of enzyme flexibility on catalytic efficiency has also been shown in cytochrome P450. A cytochrome P450 variant M.aqRLT showed strongly improved substrate binding and catalytic efficiency, and the MD simulations revealed the tunnel modifications caused greatly reduced flexibility of the two loop regions (Rapp et al., 2021). Recently, both the insertion-deletion mutagenesis and anisotropic network model highlighted the importance of the conformational flexibility of a loop-helix fragment of *Renilla* luciferases RLuc8 for ligand binding. And, transplanting this dynamic fragment from RLuc8 to $Anc^{HLD\text{-}Rluc}$, a thermostable ancestral protein catalysing both dehalogenase and luciferase reactions, to yield an enzyme AncFT with 7000-fold improved catalytic efficiency (Schenkmayerova et al., 2021). Hence, what becomes very clear is that protein flexibility is very crucial to enzyme catalysis, promiscuity and evolution but understanding how to tune flexible sites to change enzyme functions remains challenging (Pabis et al., 2018).

3.4 Distal sites coupled to active center

Allostery describes the binding affinity of a ligand or substrate that is changed by binding another ligand far from the active site. Nature uses the principle of minimum perturbation and maximum response to change the dynamics of functional key sites through allostery, rather than obtaining new and large effect mutations. An increasing number of studies indicated that small perturbations at the distal coupling site can lead to changes in a series of functional active sites (Figure 8A). Enzymes obtained through directed evolution have also produced many cases where distal mutations in regions previously thought not to affect function are actually functionally relevant. Some mutations have improved thermal stability and protein expression, and others have increased catalytic efficiency by regulating the conformational space or affecting the dynamics of the active site (Modi and Ozkan, 2018,

Khersonsky et al., 2010, Yang and Lai, 2016, Taylor et al., 2015). Hence, the distal sites coupled to an active center could be potential hot spots to engineer functional properties of enzymes (Figure 8A).

The critical allosteric interactions in various systems can be identified by DCI (dynamic coupling index) scores (Modi and Ozkan, 2018). TEM-1 is an evolved enzyme of β-lactamase which inactivates antibiotics by hydrolyzing β-lactams. It was shown that most mutations of TEM-1 leading to resistance were located distal to the catalytic site, and their DCI scores indicated a higher coupling to the active sites (Figure 8B)(Modi and Ozkan, 2018). The distal antibiotic-resistant mutations also remotely altered the flexibility of the active site. Due to their strong dynamic coupling to the active sites, these mutations created a series of changes in the interaction network, resulting in changes in the flexibility profile of regions that play a key role in function.

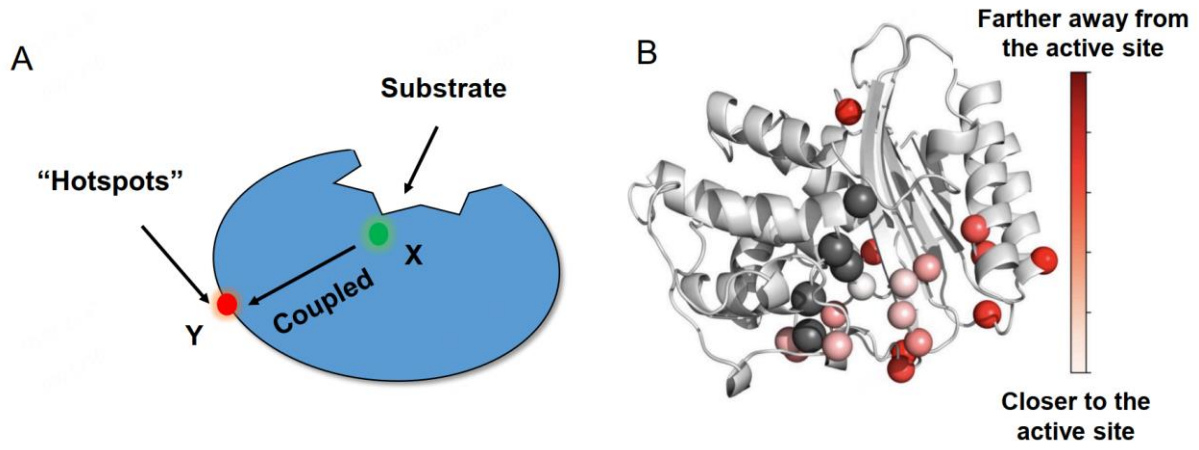

Figure 8 The distal sites important to enzymes functions. A, distal hot spots coupled to active center. B, Functional mutations distal from the TEM-1 active sites. Reprinted with permission, copyright 2018 Molecular Diversity Preservation International.

Recently, we have counteracted the stability-activity trade-off observed in the *Escherichia coli* transketolase 3M variant, also a common problem in directed evolution, by making mutations

targeted to the distal sites with strong dynamics correlations with the active center (Yu and Dalby, 2018b). A clear activity-stability trade-off was found in the 3M variants, and the MD simulations revealed increased flexibility in several interconnected active-site regions that also form part of the dimer interface. Mutating the newly flexible active-site residues to regain stability risked losing the new activity. In earlier work, it was shown that mutations influenced the dynamics of their local environment, but also in some cases the dynamics of regions distant in the structure (Yu and Dalby, 2018a). Hence, six variants were constructed in the regions outside of the active sites, whose dynamics were correlated with the newly flexible active sites. The best variant had a 10.8-fold improved half-life at 55 °C, and increased the T_m and T_{agg} by 3 °C and 4.3 °C, respectively. The variants even increased the activity, by up to threefold. This study highlights how protein engineering strategies could be potentially improved by considering long-range dynamics.

3.5 Interface sites

Interface sites of a multimeric enzyme have been chosen as hot spots for engineering the thermostability of enzymes (Bosshart et al., 2013). For a multimeric enzyme, denaturation typically starts with a loss of integrity of the quaternary structure and is followed by an irreversible denaturation step (Rogers and Bommarius, 2010, Peterson et al., 2007). It was reasoned that mutagenesis targeting non-conserved residues of the interface could strengthen the inter-subunit interaction and protect proteins from disintegration.

Bosshart *et al.* tested this hypothesis using D-tagatose 3-epimerase of *P. cichorii* (PcDTE) as an example (Bosshart et al., 2013). The software PDBePISA (Krissinel and Henrick, 2007) was applied to identify the residues involved in interface formation in the crystal structure of PcDTE (Figure 9). They discarded the high consensus residues in the interface and randomly

mutated 31 remaining sites using site-saturation mutagenesis. At least one improved variant from each of nine of the 31 libraries was achieved. ISM was subsequently applied to accumulate beneficial mutations. Finally, after a limited screening (<4000 clones), a mutant was produced which showed 21.4 °C enhanced thermostability, and comparable substrate specificity and selectivity relative to wild type.

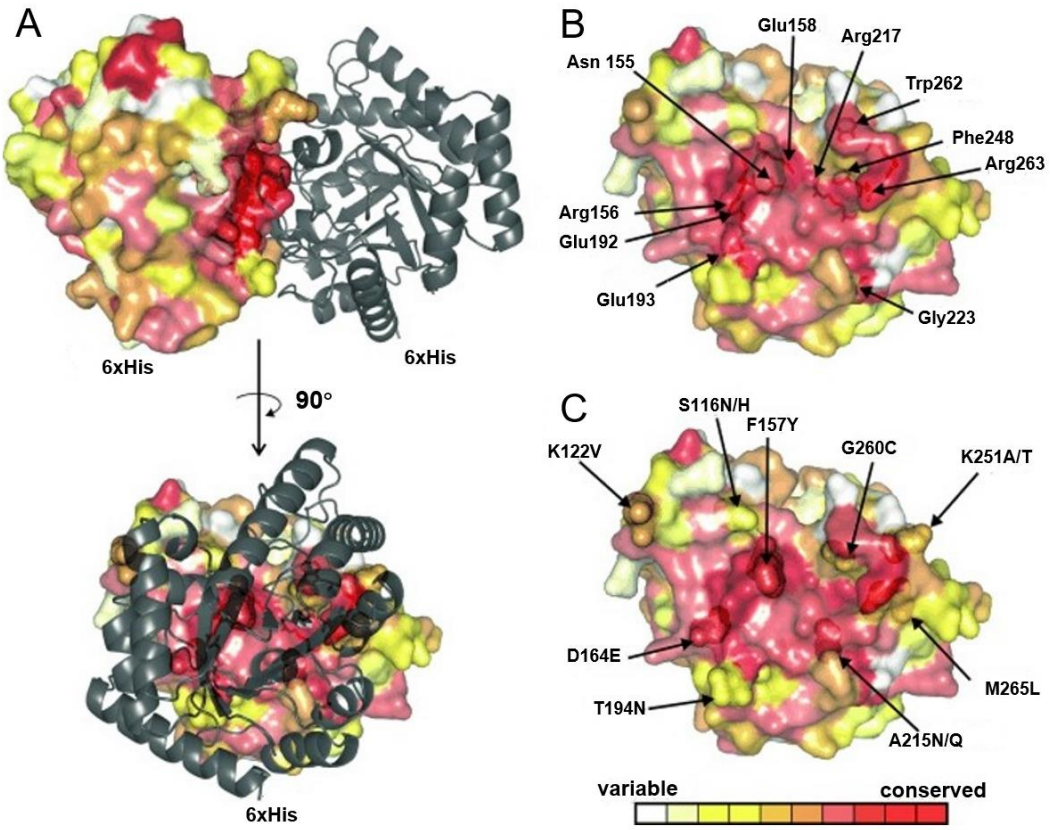

Figure 9 Localization of strictly conserved interface amino acid residues and those affording more thermostable PcDTE variants. A, PcDTE dimer with chain A in surface representation and chain B (dark gray) shown in cartoon representation, the C-terminal 6xHis-tag is marked. B, Chains A/B with all 10 strictly conserved interface residues shown as sticks. C, Chains A/B with the nine interface sites that afforded an improved mutant during the initial stability screening, highlighted as spheres. The coloring of each residue corresponds to its degree of conservation in 10 % increments. Reprinted with permission, copyright 2013 WILEY - VCH VERLAG GMBH & CO. KGAA.

In addition to random mutagenesis, site-directed mutagenesis could also be used to strengthen interactions between subunits. Numerous studies have shown the importance of salt bridges at interfaces in dominating stability of multimeric proteins (De Jesus et al., 2014, Letai and Fuchs, 1995). Based on this understanding, Basu *et al*. designed proteins with enhanced thermostability by introducing ion pairs at interface sites (Basu and Sen, 2013). They firstly identified several polar or charged residues on the protein surface which have weak interactions with other residues, then replaced the side-chains of suitable interface residues to introduce electronic interactions between monomers. Applying this strategy, they successfully improved the thermostability of both a homo-dimeric protein and a hetero-dimeric protein (Basu and Sen, 2013). Similarly to salt bridges, disulfide bonds are also important interactions for maintaining protein structure. With interface sites as hot spots, Zhao *et al.* stabilized a single-chain fragment variable by adding an interdomain disulfide bond (Zhao et al., 2010, Zhang et al., 2018, Hong et al., 2017, Meng et al., 2020)

4. Computational tools to identify hot spots

Recent computational technological advances have greatly facilitated the identification of the hot spots mentioned above. Our goal here is not to review all the methods used to predict hot spots, and instead only those commonly used in the successful enzyme engineering cases will be introduced (Table 2). Some of the tools have been reviewed before (Marques et al., 2021, Planas-Iglesias et al., 2021). In addition to the computational technology, some more general methods might be used for identification of hot spots. For example, if no structure or homologous sequence information is available and the first step in protein design can be random mutagenesis and screening to experimentally identify hot spots, which can subsequently be investigated by saturation mutagenesis.

Table 2 Computational tools to identify hot spots

Hot spots	Name	Web	References
CbD sites	Muscle	https://www.ebi.ac.uk/Tools/msa/muscle/	(Edgar, 2004)
	ClustalW	https://www.genome.jp/tools-bin/clustalw	(Larkin et al., 2007)
	T-coffee	http://www.tcoffee.org/	(Notredame et al., 2000)
	EVcoupling	https://evcouplings.org/	(Hopf et al., 2019)
	Hot-Spot Wizard 3.0	http://loschmidt.chemi.muni.cz/hotspotwizard.	(Sumbalova et al., 2018)
Coevolving residues	CCMpred	https://github.com/soedinglab/ccmpred	(Seemayer et al., 2014)
	InterMap3D	https://services.healthtech.dtu.dk/service.php?InterMap3D-1.3	(Gouveia-Oliveira et al., 2009)
	BIS2Analyzer	http://www.lcqb.upmc.fr/BIS2Analyzer/	(Oteri et al., 2017)
	SDPpred	http://bioinf.fbb.msu.ru/SDPpred/	(Kalinina et al., 2004)
	SDPsite	http://bioinf.fbb.msu.ru/SDPsite/	(Kalinina et al., 2009)
Active-site residues	CRpred	http://biomine.cs.vcu.edu/datasets/CRpred/CRpred.html	(Zhang et al., 2008)
	ConSurf	https://consurf.tau.ac.il/	(Ashkenazy et al., 2010)
	DISCERN	http://phylogenomics.berkeley.edu/software	(Sankararaman et al., 2010)
Access tunnel sites	CAVER	https://loschmidt.chemi.muni.cz/caverweb/	(Stourac et al., 2019
	MOLEonline	https://mole.upol.cz/	(Pravda et al., 2018)
Flexible sites	B-FITTER	https://www.kofo.mpg.de/en/research/biocatalysis	(Blum et al., 2012)
	WHAT IF	https://swift.cmbi.umcn.nl/servers/html/index.html	(Vriend, 1990)
Distal sites coupled to active center	AlloPred	https://github.com/jgreener64/allopred	(Greener and Sternberg, 2015)
	PARS	http://bioinf.uab.cat/pars.	(Panjkovich and Daura, 2014)
	DynOmics ENM	http://enm.pitt.edu/	(Li et al., 2017)
	CORRSITE	http://www.pkumdl.cn:8000/cavityplus/index.php	(Xu et al., 2018)

	AllosMod	https://modbase.compbio.ucsf.edu/allosmod/	(Weinkam et al., 2012)
	PASSer	https://passer.smu.edu/	(Tian et al., 2021)
	Allosite	http://mdl.shsmu.edu.cn/AST	(Huang et al., 2013)
Interface sites	PredUs 2.0	https://honiglab.c2b2.columbia.edu/PredUs/index_omega.html	(Hwang et al., 2016)
	InterProSurf	http://curie.utmb.edu/	(Negi et al., 2007)
	iFraG	http://sbi.imim.es/web/index.php/research/servers/iFrag	(Garcia-Garcia et al., 2017)
	PDBePisa	https://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver	(Krissinel and Henrick, 2007)
	MolSurfer	https://molsurfer.h-its.org/index.html	(Gabdoulline et al., 2003)
	SPPIDER	http://sppider.cchmc.org/	(Porollo and Meller, 2007)

4.1 CdD sites

Obtaining the consensus sequence is the key step in identifying the CbD sites. To prepare a starting MSA for determining the consensus sequence, a sequence set of a query protein need to be collected firstly. The easiest way to obtain the sequence set is from databases including Pfam (El-Gebali et al., 2019), InterPro (Mitchell et al., 2019), SMART (Letunic and Bork, 2018) which contain MSAs for a vast number of protein families. Another option is to build an MSA using homolog search and alignment tools such as HMMER (Finn et al., 2011) and PSI-BLAST (Altschul et al., 1997). The second step is to curate the sequence set to create an alignment of diverse yet nonredundant sequences by removing sequences that share high identity or are too long or too short. The MSA programs including MAFFT (Katoh et al., 2019), ClustalW (Larkin et al., 2007), MUSCLE (Edgar, 2004) and T-Coffee (Notredame et al., 2000) are then used to generate MSA before calculating the residue frequencies at each position to generate consensus sequence. Python scripts for cleaning and curating MSAs, calculating residue

frequencies from MSAs and determining the consensus sequences can be found on Github at github.com/msternke/protein-consensus-sequence (Sternke et al., 2020). The above methods have been used to generate consensus sequences of six protein families including N-terminal domain of ribosomal protein L9, the SH3 domain, the SH2 domain, dihydrofolate reductase, adenylate kinase and phosphoglycerate kinase. All six consensus proteins adopt cooperatively folded structures in solution, and four of them showed increased thermodynamic stability over naturally occurring homologs (Sternke et al., 2019). In addition, tools including EVcoupling, Gremlin and Hot-Spot Wizard 3.0 are able to take a single sequence as input, from which they automatically find a set of homologous sequences, construct a multiple alignment, generate the consensus sequence (Hopf et al., 2019, Sumbalova et al., 2018, Kamisetty et al., 2013). However, a manually curated dataset will most often be of better quality than the automatically generated one, thus improving the quality of the predictions.

4.2 Coevolving residues

Coevolving residues are identified also by utilizing MSA to compute couplings between pairs of positions in a protein sequence. Since statistical dependency between amino acid positions may arise either from direct or indirect correlates residues, these methods are commonly classified into two categories: methods that consider all covarying interactions as independent between each other, and direct coupling methods that deconvolute the covariation signal in order to infer only direct interactions (Colell et al., 2018). Directed coupling analysis (DCA) (http://dca.rice.edu/portal/dca) method was developed to estimate direct inter-residue contacts, which is widely applied to predict structural proximity such as both AlphaFold and RaptorX relying on the inter-residue contacts predicted by CCMpred, a

DCA-based approach (Morcos et al., 2011, Ju et al., 2021). By contrast, statistical coupling anslysis (SCA) introduced by Lockless and Ranganathan in 1999 is a way to infer energetic interactions within a protein from a statistical analysis of MSA (Lockless and Ranganathan, 1999). DCA and SCA produced different results by analysis of same MSAs, which was due to differences in the algorithmic approaches: SCA uses clustering to identify larger groups of coevolving sites (sectors), whereas DCA uses maximum-entropy modeling to extract pairs of directly coupled residues (Morcos et al., 2011). SCA has been previously used to identify evolutionarily correlated networks of residues that included the active-site residues for transketolase. Screening of libraries targeted to one of these networks led to the R520Q mutation that stabilized the transketolase variant D469T sufficiently to restore soluble expression(Strafford et al., 2012). In addition, the methods such as InterMap3D are based on mutual information (MI), a statistical measure of the codependency between two random variables, for detecting coevolving residues (Gouveia-Oliveira et al., 2009). The MI and SCA might not always correspond to residue proximity, but is useful in allosteric pathway prediction, which has been applied to engineer the thermostability of an apha-amylase (Wang et al., 2012). MISTIC2 is a server that allows to calculate coevolving residues with both DCA and MI methods, specifically three DCA approaches including mean field DCA, pseudo-likehood maximization DCA, multivariate gaussian modelling DCA and a corrected mutual information approach (Colell et al., 2018). MI can also be applied to the prediction of SDPs by calculating MSA into subgroups. Tools like SDPpred and SDPsite combine MI with subsequent statistical significance, enabling simple and effective prediction of SDPs (Kalinina et al., 2004, Kalinina et al., 2009). In addition, with the input of manually curated MSAs or MSAs from databases such as Pfam (El-Gebali et al., 2019) and HMMER (Finn et al., 2011), several tools including CoeViz 2.0 (Baker and Porollo, 2016), BIS2Analyzer (Oteri et al., 2017) and CAPS 2.0

(Fares and Travers, 2006) can detect groups of amino acids that evolve together. And, in addition to the consensus sequence, the webservers including EVcoupling (Hopf et al., 2019), Gremlin (Kamisetty et al., 2013) and Hot-Spot Wizard 3.0 (Sumbalova et al., 2018) are able to automatically identify the coevolving residues with the input of a single sequence. EVcoupling server can also be used to predict mutation effect and predict protein structure, which has recently been used to develop a machine learning-assisted directed evolution method and engineer activity of a *Bacillus naganoensis* pullulanase (Wittmann et al., 2021, Wang et al., 2020b).

4.3 Active-site residues

The enzyme 3D structure with a substrate bound provides accurate identification of active-site residues. However, experimental studies to predict active sites are cumbersome and time-consuming. In the past decade, many sequence or structure-based methods have been developed to predict enzyme active-site residues. Purely sequence-based approaches use phylogenetic information, relying on the idea that functional sites are conserved during evolution, and hence the computational tools to predict consensus sequence are useful to identify active-site residues (Aubailly and Piazza, 2015). CRpred (http://biomine.cs.vcu.edu/datasets/CRpred/CRpred.html) is a widely used sequence-based method that uses several sequence features including residue type, hydrophobicity, and PSI-BLAST profiles in a support vector machine (SVM) to predict residues to be catalytic residues or not (Zhang et al., 2008). Another method, ConSurf (https://consurf.tau.ac.il/) identifies functionally important regions in proteins by estimating the degree of conservation of the amino acid sites among their close sequence homologues (Ashkenazy et al., 2010). The ConSurf webserver has been used to guide the activity engineering of prodigiosin ligase PigC

to exclude those residues from mutagenesis that showed high conservation scores as they are likely to be essential for enzymatic function (Brands et al., 2021). With the availability of tertiary structures, methods were developed by using structure similarity searches with pre-calculated active site structural template library, such as CATSID which enables rapid searches for structural matches to a user-specified catalytic site among all PDB structures (Kirshner et al., 2013). Many other methods combine sequence and structural features to improve prediction accuracy. For example, DISCERN (http://phylogenomics.berkeley.edu/software) uses statistical models based on phylogenomic conservation score of sequence and several structural features including B-factors and solvent accessibility to predict catalytic residues (Sankararaman et al., 2010). PREvaIL is an integrative approach for inferring catalytic residues using sequence, structural, and network features in a random forest machine-learning framework (Song et al., 2018). In addition, to pinpoint the specific amino acids modulating binding of substrates, molecular docking tools such as AutoDock (Morris et al., 2009), AutoDock Vina (Trott and Olson, 2010), Glide (Friesner et al., 2004), and Gold (Verdonk et al., 2003) could be used. Since their performance is largely influenced by the type of ligand being docked and the target system, it is worthwhile to select a tool that shows high success rates with molecules similar to those of interest (Ebert and Pelletier, 2017).

4.4 Access tunnel sites

Since the software tools for the determination of protein tunnels have been reviewed by Brezovsky et al. in 2018 (Brezovsky et al., 2018), Kingsley and Lill in 2015 (Kingsley and Lill, 2015), and Damborsky et al. in 2019 (Kokkonen et al., 2019), we give only a brief overview of the tools used in recent enzyme engineering cases. CAVER Analyst 2.0 (Jurcik et al., 2018) has been used for engineering xylanase activity (Lu et al., 2019), and CAVER 3.0 (Chovancova et

al., 2012) has been applied to engineer haloalkane dehalogenase activity and carotenoid cleavage dioxygenase substrate scope (Kaushik et al., 2018, Liang et al., 2021). CAVER 3.0 is a software tool widely used for the identification and characterization of transport pathways in both static macromolecular structures and large ensembles of protein conformations (Heinemann et al., 2021). It implements algorithms for the calculation and clustering of pathways and hence enables calculating detailed characteristics and statistics of the time evolution of individual pathways with a trajectory from a MD simulation as the input, which is critical to find occasionally closed tunnels. CAVER Analyst 2.0 (Jurcik et al., 2018) enables visualization of access tunnels computed by the CAVER 3.0 algorithms, which also provides an intuitive graphic user interface for setting up the calculation and interactive exploration of identified tunnels and their characteristics. CAVER software was also available for easy-to-use as a web server CABER Web 1.0 (Stourac et al., 2019). MOLEonline is another interactive, web-based application for the detection and characterization of access tunnels within protein structures (Pravda et al., 2018), which was recently used for engineering cytochrome P450 and phenylalanine ammonia-lyases substrate preference (Meng et al., 2021, Bata et al., 2021). Both CAVER and MOLE detect tunnels based on the Voronoi diagram representation of a protein structure alone and offer high-quality results in short calculation time.

Another time-demanding method for *in silico* analyses of ligand transport is based on use of MD simulations which simulate small ligands passing through channel and provide highly robust and accurate results. Since timescale of ligand (un)binding is very long, the MD simulations often employ various enhanced sampling approaches such as Random Accelerated Molecular Dynamics (Kokh et al., 2018), Steered Molecular Dynamics (Chen, 2015, Do et al., 2018, Skovstrup et al., 2012), Umbrella Sampling (Zhang and Voth, 2011), Adaptive

Sampling (Marques et al., 2018) and Metadynamics (Zhang and Voth, 2011, Furini and Domene, 2016).

4.5 Flexible sites

B-Factors and MD simulations are two commonly used methods to identify flexible sites (Yu and Huang, 2014, Sun et al., 2019). The average B-factors for a residue can be calculated by averaging the B-values of each atom in the amino acid with programs including B-FITTER (Blum et al., 2012) (https://www.kofo.mpg.de/en/research/biocatalysis) and WHAT IF Web server (Vriend, 1990) (https://swift.cmbi.umcn.nl/servers/html/index.html). However, it is important to note that the flexibility of proteins in solution may be qualitatively different from that in crystals. MD simulations is a more powerful technique to study the flexibility of proteins. It provides an accurate representation of protein flexibility under similar physiological environments. RMSF values in the MD simulations were used to represent flexibility of the protein, which measure mean amplitude of each atom motions relative to a mean reference position during MD trajectory. The equivalent number of protein blocks, N_{eq} value, is also an indicator of protein flexibility (Figure 7). Protein blocks (PBs) are a structural alphabet composed of 16 local prototypes representing β-strand N-caps, β-strand C-caps, coils, α-helix N-caps, α-helix C-caps and so forth (Joseph et al., 2010). Each residue can be assigned a PB type and PB assignments are done for each residue over every snapshot extracted from MD simulations. N_{eq} is a statistical measurement that represents the average number of PBs for a residue at a given position in the MD simulations (Goguet et al., 2017, de Brevern et al., 2000). MD simulations can be carried out in packages such as GROMACS (Van Der Spoel et al., 2005), AMBER (Case et al., 2005), NAMD (Phillips et al., 2005) or CHARMM (Brooks et al., 2009).

4.6 Distal sites coupled to active centers

The methods to predict distal sites coupled to active centers are related to investigating allostery. Several of such tools have been reviewed before (Guarnera and Berezovsky, 2016, Schueler-Furman and Wodak, 2016, Greener and Sternberg, 2018, Sheik Amamuddy et al., 2020). Generally, the methods could be classified as sequence-based and structure-based. Based on the assumption that allosteric communication paths are under evolutionary pressure, the computational tools such as InterMap3D and SCA for predicting coevolving residues based on MSAs could also be used to predict allostery signalling (Schueler-Furman and Wodak, 2016). The key characteristic of the allosteric site is its ability to couple to the intrinsic dynamics of protein, which, in turn, underlies communication with relevant functional sites through coherent collective motions. Hence, as the standard computational tool for dynamics analysis, MD simulations are frequently used for the search of allosteric binding sites.

The dynamical cross-correlation matrix (DCCM) measured by the analysis of simulation trajectories could be used to determine the potential allosteric sites showing dynamics correlations (Yu and Dalby, 2020). Perturbation response scanning (PRS) is also a popular MD method for allosteric prediction, which examines the response of the structure to random perturbations caused by systematically applying a series of uniformly distributed forces at specific positions (Atilgan and Atilgan, 2009). PRS has been successfully applied to predict the allosteric hotspot residues of the TEM-1 β-lactamase (Modi and Ozkan, 2018), chaperone Hsp70 (Penkler et al., 2017) and two PDZ domain proteins (Gerek and Ozkan, 2011). In addition, AllosMod (https://modbase.compbio.ucsf.edu/allosmod/) has been developed by combining MD simulations and energy landscape construction which can

sample the conformational transitions sufficiently well to accurately link microscopic motions to macroscopic allosteric phenomena (Weinkam et al., 2012). However, since conformational changes that cause allostery are often large enough to occur on timescales of microseconds or milliseconds, it is too computationally expensive to use MD simulations to predict allosteric sites.

Normal model analysis (NMA) methods based on assumption of harmonic motion around an energy minimum provide another faster tool for allostery analysis. AlloPred (Greener and Sternberg, 2015) (https://github.com/jgreener64/allopred) is available as web server which calculates the normal modes of a protein, then holds the springs in the region of a potential allosteric site rigid and measures the effect of this perturbation at the active site. The DynOmics ENM server (Li et al., 2017) (http://enm.pitt.edu/) finds hinge residues that control the two slowest normal modes of a protein, and hence is able to influence its dynamics. CORRSITE (Xu et al., 2018) (http://www.pkumdl.cn:8000/cavityplus/index.php) identifies potential allosteric sites based on motion correlation between ligand-binding sites and corresponding orthosteric sites. These methods are expected to reveal the perturbations to vibrations, but other factors contributing to the allostery such as local unfolding are not taken into account. Several machine learning methods such as PASSer (https://passer.smu.edu/) have been developed for prediction allosteric sites (Tian et al., 2021). Allosite (Huang et al., 2013)(http://mdl.shsmu.edu.cn/AST) is a method for predicting allosteric sites using support vector machine (SVM) based on topological and physiochemical pocket features. Chen et al. used random forest (RF) to construct a predictive model to classify protein cavities into three categories: allosteric, regular or orthosteric (Chen et al., 2016).

4.7 Interface sites

Many different approaches have been developed to predict protein interface sites. Methods might use intrinsic features of the sequence or the structure, evolutionary relationships or use an existing complex as a reference template. Sequence-based interface predictors such as iFraG (http://sbi.imim.es/web/index.php/research/servers/iFrag) use only sequence features to detect interfaces, useful for the proteins without structure information available, but typically have lower accuracies than methods incorporating evolutionary and structural information (Garcia-Garcia et al., 2017). Structure-based methods such as PDBePisa (Krissinel and Henrick, 2007) (https://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver), MolSurfer (Gabdoulline et al., 2003) (https://molsurfer.h-its.org/index.html), SPPIDER (Porollo and Meller, 2007) (http://sppider.cchmc.org/) make use of the different structural features including secondary structure, solvent-accessible surface area, geometric shape of the protein surface and crystallographic B-factor. As discussed in the section of 3.5, PDBePISA was successfully applied to predict residues involved in interface formation and provided the mutation targets for thermostability engineering (Bosshart et al., 2013). In addition, the InterProSurf (Negi et al., 2007) web server (http://curie.utmb.edu/) predicts a list of amino acid residues based on their accessible surface area and propensities most likely to be responsible for protein interaction, which has been applied to *Bacillus anthracis* toxin and measles virus hemagglutinin proteins to identify interface regions (Negi et al., 2007). The machine learning-based techniques such as partial least squares (PLS) regression, SVM and random forest have also been combined with sequence or structure information to act as the predictors of interface sites including PAIRpred (Minhas et al., 2014), Protein IntErface Recognition (PIER) (Kufareva et al., 2007), PredUs 2.0 (https://honiglab.c2b2.columbia.edu/PredUs/index_omega.html) (Zhang et al., 2011) and

the predictor based on 3D Zernike descriptors (Daberdaku and Ferrari, 2018). Since these predictors make use of many distinct quality measures, different training and testing data sets, it is hard to have a fair comparison between them and the detailed discussion has been reviewed before (Esmaielbeiki et al., 2016). Recently, the proteome-wide amino acid coevolution analysis and deep-learning–based structure modelling have been used to systematically build accurate models of core eukaryotic protein complexes (Humphreys et al., 2021). This enabled a large-scale screen of protein-protein interactions and the accurate identification of interface sites of many protein complexes.

Table 3 Guide for choosing suitable hot spots

	Activity	**Enantioselectivity**	**Stability**
Sequence	CbD sites[a] Coevolving residues	CbD sites	CbD sites Coevolving residues
3D structure	Active-site residues Access tunnel sites Distal sites coupled to active center	Active-site residues	Flexible sites Interface sites Access tunnel sites Distal sites coupled to active center

[a]CbD sites: Positions that are conserved in the multiple sequence alignments but different in the sequence of the target protein.

5. Conclusion

The dream of all protein engineers is to predict a single amino acid sequence that would work with desired functions. However, it is far from being reached due to our incomplete understanding of the relationship between the function and structure of proteins. In order to make extant enzymes applicable to wider fields, the evolutionary strategy of trial and error is an inevitable choice (Goldsmith and Tawfik, 2013). The current challenge is to decrease the exploration of sequence space from an impractical number of all possible sequence (20^n, n being the number of amino acids) to a controllable number of mutants. Targeted mutagenesis

is an effective approach to narrow the sequence space and increase the efficiency of directed evolution by constructing smart libraries. We have reviewed recent studies applying seven kinds of hot spots as mutagenesis targets for engineering various properties of enzymes. These spots can be divided into two types based on whether crystal structures are available or not: sequence-based hot spots and 3D structure-based hot spots. Choosing hot spots should be based on the desired property and available structure information (Table 3). If one wants to enhance thermostability and the 3D structure is available, flexible sites and interface sites are good choices since these sites are far away from catalytic sites. During the process of identifying hot spots, consensus information from MSAs is also very useful. On the one hand, conserved residues are not often suitable for mutation, whereas CbD sites are good targets for engineering thermostability. After identifying the hot spots, saturation mutagenesis was commonly applied to construct mutation libraries. However, some enzymes are not amenable to high-throughput screening. In this case, site-directed mutagenesis can be used to introduce salt bridges or disulfide bonds to stabilize flexible sites or interface sites.

It is important to note that these hot spots only provide potential mutation targets. Other criteria might be needed to shrink the scope of mutation sites in the real situation. For example, many enzymes have multiple tunnels connecting their active sites with the surrounding solvent and each tunnel has different number of amino acids (Kokkonen et al., 2019). To identify the specific mutation sites in a tunnel for modifying the substrate preference, other criteria are needed to be set, such as location in a bottleneck area or initial part of the tunnel, location in a loop area not completely conserved in the homologous families (Meng et al., 2021, Bata et al., 2021). Flexible sites are great hot spots for engineering thermostability. Except identifying the most flexible sites as the mutation target, the flexible

regions could be located initially, and then sequence statistics used to pinpoint the mutation positions and substitution types (Yu et al., 2015, Yu et al., 2017).

When analysing structural data to identify hot spots, literature is an important means to acquire information about the previous mutagenesis attempts of the target enzyme. Using information from previous studies, Bornscheuer *et al.* identified three hot spots of *Pseudomonas fluorescens* esterase (PFE I) and then generated a mutant with 15-fold improved enantioselectivity (Nobili et al., 2015). Additionally, with the recent breakthrough of deep learning in protein three-dimensional structure prediction (Jumper et al., 2021, Senior et al., 2020, Baek et al., 2021), our understanding of protein structure-function relationships will be greatly improved, and new hot spots for engineering enzymes are expected to be shown up in the near future.

Reference

ALTSCHUL, S. F., MADDEN, T. L., SCHäFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25**,** 3389-402.

ANBAR, M., GUL, O., LAMED, R., SEZERMAN, U. O. & BAYER, E. A. 2012. Improved thermostability of Clostridium thermocellum endoglucanase Cel8A by using consensus-guided mutagenesis. *Appl Environ Microbiol,* 78**,** 3458-64.

ANISHCHENKO, I., OVCHINNIKOV, S., KAMISETTY, H. & BAKER, D. 2017. Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci U S A,* 114**,** 9122-9127.

ASHKENAZY, H., EREZ, E., MARTZ, E., PUPKO, T. & BEN-TAL, N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res,* 38**,** W529-33.

ATILGAN, C. & ATILGAN, A. R. 2009. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput Biol,* 5**,** e1000544.

AUBAILLY, S. & PIAZZA, F. 2015. Cutoff lensing: predicting catalytic sites in enzymes. *Sci Rep,* 5**,** 14874.

BAEK, M., DIMAIO, F., ANISHCHENKO, I., DAUPARAS, J., OVCHINNIKOV, S., LEE, G. R., WANG, J., CONG, Q., KINCH, L. N., SCHAEFFER, R. D., MILLáN, C., PARK, H., ADAMS, C., GLASSMAN, C. R., DEGIOVANNI, A., PEREIRA, J. H., RODRIGUES, A. V., VAN DIJK, A. A., EBRECHT, A. C., OPPERMAN, D. J., SAGMEISTER, T., BUHLHELLER, C., PAVKOV-KELLER, T., RATHINASWAMY, M. K., DALWADI, U., YIP, C. K., BURKE, J. E., GARCIA, K. C., GRISHIN, N. V., ADAMS, P. D., READ, R. J. & BAKER, D. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*.

BAKER, F. N. & POROLLO, A. 2016. CoeViz: a web-based tool for coevolution analysis of protein residues. *BMC Bioinformatics,* 17**,** 119.

BALKE, K., BäUMGEN, M. & BORNSCHEUER, U. T. 2017. Controlling the Regioselectivity of Baeyer-Villiger Monooxygenases by Mutation of Active-Site Residues. *Chembiochem,* 18**,** 1627-1638.

BAO, L., LI, J. J., JIA, C., LI, M. & LU, X. 2016. Structure-oriented substrate specificity engineering of aldehyde-deformylating oxygenase towards aldehydes carbon chain length. *Biotechnol Biofuels,* 9**,** 185.

BARTSCH, S., KOURIST, R. & BORNSCHEUER, U. T. 2008. Complete inversion of enantioselectivity towards acetylated tertiary alcohols by a double mutant of a Bacillus subtilis esterase. *Angew Chem Int Ed Engl,* 47**,** 1508-11.

BASSEGODA, A., NGUYEN, G.-S., SCHMIDT, M., KOURIST, R., DIAZ, P. & BORNSCHEUER, U. T. 2010. Rational Protein Design of Paenibacillus barcinonensis Esterase EstA for Kinetic Resolution of Tertiary Alcohols. *ChemCatChem,* 2**,** 962-967.

BASU, S. & SEN, S. 2013. An in silico method for designing thermostable variant of a dimeric mesophilic protein based on its 3D structure. *J Mol Graph Model,* 42**,** 92-103.

BATA, Z., MOLNáR, Z., MADARAS, E., MOLNáR, B., SáNTA-BELL, E., VARGA, A., LEVELES, I., QIAN, R., HAMMERSCHMIDT, F., PAIZS, C., VéRTESSY, B. G. & POPPE, L. 2021. Substrate Tunnel Engineering Aided by X-ray Crystallography and Functional Dynamics Swaps the Function of MIO-Enzymes. *ACS Catal,* 11**,** 4538-4549.

BIEDERMANNOVA, L., PROKOP, Z., GORA, A., CHOVANCOVA, E., KOVACS, M., DAMBORSKY, J. & WADE, R. C. 2012. A single mutation in a tunnel to the active site changes the mechanism and kinetics of product release in haloalkane dehalogenase LinB. *J Biol Chem,* 287**,** 29062-74.

BLUM, J. K., RICKETTS, M. D. & BOMMARIUS, A. S. 2012. Improved thermostability of AEH by combining B-FIT analysis and structure-guided consensus method. *J Biotechnol,* 160**,** 214-21.

BOSSHART, A., PANKE, S. & BECHTOLD, M. 2013. Systematic optimization of interface interactions increases the thermostability of a multimeric enzyme. *Angew Chem Int Ed Engl,* 52**,** 9673-6.

BRANDS, S., SIKKENS, J. G., DAVARI, M. D., BRASS, H. U. C., KLEIN, A. S., PIETRUSZKA, J., RUFF, A. J. & SCHWANEBERG, U. 2021. Understanding substrate binding and the role of gatekeeping residues in PigC access tunnels. *Chem Commun (Camb),* 57**,** 2681-2684.

BREZOVSKY, J., KOZLIKOVA, B. & DAMBORSKY, J. 2018. Computational Analysis of Protein Tunnels and Channels. *Methods Mol Biol,* 1685**,** 25-42.

BROOKS, B. R., BROOKS, C. L., 3RD, MACKERELL, A. D., JR., NILSSON, L., PETRELLA, R. J., ROUX, B., WON, Y., ARCHONTIS, G., BARTELS, C., BORESCH, S., CAFLISCH, A., CAVES, L., CUI, Q., DINNER, A. R., FEIG, M., FISCHER, S., GAO, J., HODOSCEK, M., IM, W., KUCZERA, K., LAZARIDIS, T., MA, J., OVCHINNIKOV, V., PACI, E., PASTOR, R. W., POST, C. B., PU, J. Z., SCHAEFER, M., TIDOR, B., VENABLE, R. M., WOODCOCK, H. L., WU, X., YANG, W., YORK, D. M. & KARPLUS, M. 2009. CHARMM: the biomolecular simulation program. *J Comput Chem,* 30**,** 1545-614.

BROUK, M., DERRY, N.-L., SHAINSKY, J., ZELAS, Z. B.-B., BOYKO, Y., DABUSH, K. & FISHMAN, A. 2010. The influence of key residues in the tunnel entrance and the active site on activity and selectivity of toluene-4-monooxygenase. *J Mol Catal B: Enzym,* 66**,** 72-80.

CASE, D. A., CHEATHAM III, T. E., DARDEN, T., GOHLKE, H., LUO, R., MERZ JR, K. M., ONUFRIEV, A., SIMMERLING, C., WANG, B. & WOODS, R. J. 2005. The Amber biomolecular simulation programs. *J Comput Chem,* 26**,** 1668-1688.

CHAGOYEN, M., GARCíA-MARTíN, J. A. & PAZOS, F. 2016. Practical analysis of specificity-determining residues in protein families. *Brief Bioinform,* 17**,** 255-61.

CHANG, J., ZHANG, C., CHENG, H. & TAN, Y. W. 2021. Rational Design of Adenylate Kinase Thermostability through Coevolution and Sequence Divergence Analysis. *Int J Mol Sci,* 22.

CHEN, A. S., WESTWOOD, N. J., BREAR, P., ROGERS, G. W., MAVRIDIS, L. & MITCHELL, J. B. 2016. A Random Forest Model for Predicting Allosteric and Functional Sites on Proteins. *Mol Inform,* 35**,** 125-35.

CHEN, K. & ARNOLD, F. H. 2020. Engineering new catalytic activities in enzymes. *Nat Catal,* 3**,** 203-213.
CHEN, L. Y. 2015. Hybrid Steered Molecular Dynamics Approach to Computing Absolute Binding Free Energy of Ligand-Protein Complexes: A Brute Force Approach That Is Fast and Accurate. *J Chem Theory Comput,* 11**,** 1928-38.
CHOVANCOVA, E., PAVELKA, A., BENES, P., STRNAD, O., BREZOVSKY, J., KOZLIKOVA, B., GORA, A., SUSTR, V., KLVANA, M., MEDEK, P., BIEDERMANNOVA, L., SOCHOR, J. & DAMBORSKY, J. 2012. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol,* 8**,** e1002708.
COJOCARU, V., WINN, P. J. & WADE, R. C. 2007. The ins and outs of cytochrome P450s. *Biochim Biophys Acta,* 1770**,** 390-401.
COLELL, E. A., ISERTE, J. A., SIMONETTI, F. L. & MARINO-BUSLJE, C. 2018. MISTIC2: comprehensive server to study coevolution in protein families. *Nucleic Acids Res,* 46**,** W323-w328.
DABERDAKU, S. & FERRARI, C. 2018. Exploring the potential of 3D Zernike descriptors and SVM for protein-protein interface prediction. *BMC Bioinform,* 19**,** 35.
DALBY, P. A. 2011. Strategy and success for the directed evolution of enzymes. *Curr Opin Struct Biol,* 21**,** 473-80.
DE BREVERN, A. G., ETCHEBEST, C. & HAZOUT, S. 2000. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins,* 41**,** 271-87.
DE JESUS, M. C., INGLE, B. L., BARAKAT, K. A., SHRESTHA, B., SLAVENS, K. D., CUNDARI, T. R. & ANDERSON, M. E. 2014. The role of strong electrostatic interactions at the dimer interface of human glutathione synthetase. *Protein J,* 33**,** 403-9.
DE JUAN, D., PAZOS, F. & VALENCIA, A. 2013. Emerging methods in protein co-evolution. *Nat Rev Genet,* 14**,** 249-61.
DICKSON, R. J. & GLOOR, G. B. 2014. Bioinformatics identification of coevolving residues. *Methods Mol Biol,* 1123**,** 223-43.
DO, P. C., LEE, E. H. & LE, L. 2018. Steered Molecular Dynamics Simulation in Rational Drug Design. *J Chem Inf Model,* 58**,** 1473-1482.
EBERT, M. C. & PELLETIER, J. N. 2017. Computational tools for enzyme improvement: why everyone can - and should - use them. *Curr Opin Chem Biol,* 37**,** 89-96.
EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.
EL-GEBALI, S., MISTRY, J., BATEMAN, A., EDDY, S. R., LUCIANI, A., POTTER, S. C., QURESHI, M., RICHARDSON, L. J., SALAZAR, G. A., SMART, A., SONNHAMMER, E. L. L., HIRSH, L., PALADIN, L., PIOVESAN, D., TOSATTO, S. C. E. & FINN, R. D. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res,* 47**,** D427-d432.
ESMAIELBEIKI, R., KRAWCZYK, K., KNAPP, B., NEBEL, J. C. & DEANE, C. M. 2016. Progress and challenges in predicting protein interfaces. *Brief Bioinform,* 17**,** 117-31.
FANG, Z., ZHOU, P., CHANG, F., YIN, Q., FANG, W., YUAN, J., ZHANG, X. & XIAO, Y. 2014. Structure-based rational design to enhance the solubility and thermostability of a bacterial laccase Lac15. *PLoS One,* 9**,** e102423.
FARES, M. A. & TRAVERS, S. A. 2006. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics,* 173**,** 9-23.
FINN, R. D., CLEMENTS, J. & EDDY, S. R. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res,* 39**,** W29-37.
FRIESNER, R. A., BANKS, J. L., MURPHY, R. B., HALGREN, T. A., KLICIC, J. J., MAINZ, D. T., REPASKY, M. P., KNOLL, E. H., SHELLEY, M., PERRY, J. K., SHAW, D. E., FRANCIS, P. & SHENKIN, P. S. 2004. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem,* 47**,** 1739-49.
FURINI, S. & DOMENE, C. 2016. Computational studies of transport in ion channels using metadynamics. *Biochim Biophys Acta,* 1858**,** 1733-40.

GABDOULLINE, R. R., WADE, R. C. & WALTHER, D. 2003. MolSurfer: A macromolecular interface navigator. *Nucleic Acids Res,* 31**,** 3349-51.
GAO, S., ZHU, S., HUANG, R., LI, H., WANG, H. & ZHENG, G. 2018. Engineering the Enantioselectivity and Thermostability of a (+)-γ-Lactamase from Microbacterium hydrocarbonoxydans for Kinetic Resolution of Vince Lactam (2-Azabicyclo[2.2.1]hept-5-en-3-one). *Appl Environ Microbiol,* 84.
GARCIA-GARCIA, J., VALLS-COMAMALA, V., GUNEY, E., ANDREU, D., MUñOZ, F. J., FERNANDEZ-FUENTES, N. & OLIVA, B. 2017. iFrag: A Protein-Protein Interface Prediction Server Based on Sequence Fragments. *J Mol Biol,* 429**,** 382-389.
GEREK, Z. N. & OZKAN, S. B. 2011. Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. *PLoS Comput Biol,* 7**,** e1002154.
GIHAZ, S., KANTEEV, M., PAZY, Y. & FISHMAN, A. 2018. Filling the Void: Introducing Aromatic Interactions into Solvent Tunnels To Enhance Lipase Stability in Methanol. *Appl Environ Microbiol,* 84.
GLIGORIJEVIĆ, V., RENFREW, P. D., KOSCIOLEK, T., LEMAN, J. K., BERENBERG, D., VATANEN, T., CHANDLER, C., TAYLOR, B. C., FISK, I. M., VLAMAKIS, H., XAVIER, R. J., KNIGHT, R., CHO, K. & BONNEAU, R. 2021. Structure-based protein function prediction using graph convolutional networks. *Nat Commun,* 12**,** 3168.
GODINHO, L. F., REIS, C. R., ROZEBOOM, H. J., DEKKER, F. J., DIJKSTRA, B. W., POELARENDS, G. J. & QUAX, W. J. 2012. Enhancement of the enantioselectivity of carboxylesterase A by structure-based mutagenesis. *J Biotechnol,* 158**,** 36-43.
GOGUET, M., NARWANI, T. J., PETERMANN, R., JALLU, V. & DE BREVERN, A. G. 2017. In silico analysis of Glanzmann variants of Calf-1 domain of αIIbβ3 integrin revealed dynamic allosteric effect. *Sci Rep,* 7**,** 8001.
GOLDSMITH, M., ASHANI, Y., SIMO, Y., BEN-DAVID, M., LEADER, H., SILMAN, I., SUSSMAN, J. L. & TAWFIK, D. S. 2012. Evolved stereoselective hydrolases for broad-spectrum G-type nerve agent detoxification. *Chem Biol,* 19**,** 456-66.
GOLDSMITH, M. & TAWFIK, D. S. 2013. Enzyme engineering by targeted libraries. *Methods Enzymol,* 523**,** 257-83.
GóMEZ, B., RISSO, V., SANCHEZ-RUIZ, J. & ALCALDE, M. 2020. Consensus Design of an Evolved High-Redox Potential Laccase. *Front Bioeng Biotechnol,* 8**,** 354.
GORA, A., BREZOVSKY, J. & DAMBORSKY, J. 2013. Gates of enzymes. *Chem Rev,* 113**,** 5871-923.
GOUVEIA-OLIVEIRA, R., ROQUE, F. S., WERNERSSON, R., SICHERITZ-PONTEN, T., SACKETT, P. W., MøLGAARD, A. & PEDERSEN, A. G. 2009. InterMap3D: predicting and visualizing co-evolving protein residues. *Bioinformatics,* 25**,** 1963-5.
GREENER, J. G. & STERNBERG, M. J. 2015. AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinform,* 16**,** 335.
GREENER, J. G. & STERNBERG, M. J. 2018. Structure-based prediction of protein allostery. *Curr Opin Struct Biol,* 50**,** 1-8.
GUARNERA, E. & BEREZOVSKY, I. N. 2016. Allosteric sites: remote control in regulation of protein activity. *Curr Opin Struct Biol,* 37**,** 1-8.
HAILES, H. C., ROTHER, D., MULLER, M., WESTPHAL, R., WARD, J. M., PLEISS, J., VOGEL, C. & POHL, M. 2013. Engineering stereoselectivity of ThDP-dependent enzymes. *FEBS J,* 280**,** 6374-94.
HALL, B. G. 2006. Simple and accurate estimation of ancestral protein sequences. *Proc Natl Acad Sci U S A,* 103**,** 5431-6.
HAN, N., MIAO, H., DING, J., LI, J., MU, Y., ZHOU, J. & HUANG, Z. 2017. Improving the thermostability of a fungal GH11 xylanase via site-directed mutagenesis guided by sequence and structural analysis. *Biotechnol Biofuels,* 10**,** 133.
HEINEMANN, P. M., ARMBRUSTER, D. & HAUER, B. 2021. Active-site loop variations adjust activity and selectivity of the cumene dioxygenase. *Nat Commun,* 12**,** 1095.

HIBBERT, E. G., SENUSSI, T., SMITH, M. E., COSTELLOE, S. J., WARD, J. M., HAILES, H. C. & DALBY, P. A. 2008. Directed evolution of transketolase substrate specificity towards an aliphatic aldehyde. *J Biotechnol,* 134**,** 240-5.

HONG, E. Y., LEE, S. G., PARK, B. J., LEE, J. M., YUN, H. & KIM, B. G. 2017. Simultaneously Enhancing the Stability and Catalytic Activity of Multimeric Lysine Decarboxylase CadA by Engineering Interface Regions for Enzymatic Production of Cadaverine at High Concentration of Lysine. *Biotechnol J,* 12.

HONG, S. Y., PARK, H. J. & YOO, Y. J. 2014. Flexibility analysis of activity-enhanced mutants of bacteriophage T4 lysozyme. *J Mol Catal B: Enzym,* 106**,** 95-99.

HOPF, T. A., GREEN, A. G., SCHUBERT, B., MERSMANN, S., SCHäRFE, C. P. I., INGRAHAM, J. B., TOTH-PETROCZY, A., BROCK, K., RIESSELMAN, A. J., PALMEDO, P., KANG, C., SHERIDAN, R., DRAIZEN, E. J., DALLAGO, C., SANDER, C. & MARKS, D. S. 2019. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics,* 35**,** 1582-1584.

HUANG, W., LU, S., HUANG, Z., LIU, X., MOU, L., LUO, Y., ZHAO, Y., LIU, Y., CHEN, Z., HOU, T. & ZHANG, J. 2013. Allosite: a method for predicting allosteric sites. *Bioinformatics,* 29**,** 2357-9.

HUMPHREYS, I. R., PEI, J., BAEK, M., KRISHNAKUMAR, A., ANISHCHENKO, I., OVCHINNIKOV, S., ZHANG, J., NESS, T. J., BANJADE, S., BAGDE, S. R., STANCHEVA, V. G., LI, X. H., LIU, K., ZHENG, Z., BARRERO, D. J., ROY, U., KUPER, J., FERNáNDEZ, I. S., SZAKAL, B., BRANZEI, D., RIZO, J., KISKER, C., GREENE, E. C., BIGGINS, S., KEENEY, S., MILLER, E. A., FROMME, J. C., HENDRICKSON, T. L., CONG, Q. & BAKER, D. 2021. Computed structures of core eukaryotic protein complexes. *Science,* 374**,** eabm4805.

HWANG, H., PETREY, D. & HONIG, B. 2016. A hybrid method for protein-protein interface prediction. *Protein Sci,* 25**,** 159-65.

JOCHENS, H. & BORNSCHEUER, U. T. 2010. Natural diversity to guide focused directed evolution. *Chembiochem,* 11**,** 1861-6.

JOSEPH, A. P., AGARWAL, G., MAHAJAN, S., GELLY, J. C., SWAPNA, L. S., OFFMANN, B., CADET, F., BORNOT, A., TYAGI, M., VALADIé, H., SCHNEIDER, B., ETCHEBEST, C., SRINIVASAN, N. & DE BREVERN, A. G. 2010. A short survey on protein blocks. *Biophys Rev,* 2**,** 137-147.

JU, F., ZHU, J., SHAO, B., KONG, L., LIU, T. Y., ZHENG, W. M. & BU, D. 2021. CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. *Nat Commun,* 12**,** 2535.

JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A., POTAPENKO, A., BRIDGLAND, A., MEYER, C., KOHL, S. A. A., BALLARD, A. J., COWIE, A., ROMERA-PAREDES, B., NIKOLOV, S., JAIN, R., ADLER, J., BACK, T., PETERSEN, S., REIMAN, D., CLANCY, E., ZIELINSKI, M., STEINEGGER, M., PACHOLSKA, M., BERGHAMMER, T., BODENSTEIN, S., SILVER, D., VINYALS, O., SENIOR, A. W., KAVUKCUOGLU, K., KOHLI, P. & HASSABIS, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*.

JUNG, E., PARK, B. G., YOO, H. W., KIM, J., CHOI, K. Y. & KIM, B. G. 2018. Semi-rational engineering of CYP153A35 to enhance ω-hydroxylation activity toward palmitic acid. *Appl Microbiol Biotechnol,* 102**,** 269-277.

JURCIK, A., BEDNAR, D., BYSKA, J., MARQUES, S. M., FURMANOVA, K., DANIEL, L., KOKKONEN, P., BREZOVSKY, J., STRNAD, O., STOURAC, J., PAVELKA, A., MANAK, M., DAMBORSKY, J. & KOZLIKOVA, B. 2018. CAVER Analyst 2.0: analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories. *Bioinformatics,* 34**,** 3586-3588.

KALININA, O. V., GELFAND, M. S. & RUSSELL, R. B. 2009. Combining specificity determining and conserved residues improves functional site prediction. *BMC Bioinform,* 10**,** 174.

KALININA, O. V., NOVICHKOV, P. S., MIRONOV, A. A., GELFAND, M. S. & RAKHMANINOVA, A. B. 2004. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res,* 32**,** W424-8.

KAMISETTY, H., OVCHINNIKOV, S. & BAKER, D. 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A,* 110**,** 15674-9.

KATOH, K., ROZEWICKI, J. & YAMADA, K. D. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform,* 20**,** 1160-1166.

KAUSHIK, S., MARQUES, S. M., KHIRSARIYA, P., PARUCH, K., LIBICHOVA, L., BREZOVSKY, J., PROKOP, Z., CHALOUPKOVA, R. & DAMBORSKY, J. 2018. Impact of the access tunnel engineering on catalysis is strictly ligand-specific. *Febs j,* 285**,** 1456-1476.

KAZUYO, F., HONG, S. Y., YEON, Y. J., JOO, J. C. & YOO, Y. J. 2014. Enhancing the activity of Bacillus circulans xylanase by modulating the flexibility of the hinge region. *J Ind Microbiol Biotechnol,* 41**,** 1181-90.

KHEIROLLAHI, A., KHAJEH, K. & GOLESTANI, A. 2017. Rigidifying flexible sites: An approach to improve stability of chondroitinase ABC I. *Int J Biol Macromol,* 97**,** 270-278.

KHERSONSKY, O., RöTHLISBERGER, D., DYM, O., ALBECK, S., JACKSON, C. J., BAKER, D. & TAWFIK, D. S. 2010. Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series. *J Mol Biol,* 396**,** 1025-42.

KINGSLEY, L. J. & LILL, M. A. 2015. Substrate tunnels in enzymes: structure-function relationships and computational methodology. *Proteins,* 83**,** 599-611.

KIRSHNER, D. A., NILMEIER, J. P. & LIGHTSTONE, F. C. 2013. Catalytic site identification--a web server to identify catalytic site structural matches throughout PDB. *Nucleic Acids Res,* 41**,** W256-65.

KOKH, D. B., AMARAL, M., BOMKE, J., GRäDLER, U., MUSIL, D., BUCHSTALLER, H. P., DREYER, M. K., FRECH, M., LOWINSKI, M., VALLEE, F., BIANCIOTTO, M., RAK, A. & WADE, R. C. 2018. Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *J Chem Theory Comput,* 14**,** 3859-3869.

KOKKONEN, P., BEDNAR, D., PINTO, G., PROKOP, Z. & DAMBORSKY, J. 2019. Engineering enzyme access tunnels. *Biotechnol Adv,* 37**,** 107386.

KOKKONEN, P., BEIER, A., MAZURENKO, S., DAMBORSKY, J., BEDNAR, D. & PROKOP, Z. 2021. Substrate inhibition by the blockage of product release and its control by tunnel engineering. *RSC Chemical Biology,* 2**,** 645-655.

KONG, X. D., YUAN, S., LI, L., CHEN, S., XU, J. H. & ZHOU, J. 2014. Engineering of an epoxide hydrolase for efficient bioresolution of bulky pharmaco substrates. *Proc Natl Acad Sci U S A,* 111**,** 15717-22.

KOUDELAKOVA, T., CHALOUPKOVA, R., BREZOVSKY, J., PROKOP, Z., SEBESTOVA, E., HESSELER, M., KHABIRI, M., PLEVAKA, M., KULIK, D., KUTA SMATANOVA, I., REZACOVA, P., ETTRICH, R., BORNSCHEUER, U. T. & DAMBORSKY, J. 2013. Engineering enzyme stability and resistance to an organic cosolvent by modification of residues in the access tunnel. *Angew Chem Int Ed Engl,* 52**,** 1959-63.

KRISSINEL, E. & HENRICK, K. 2007. Inference of macromolecular assemblies from crystalline state. *J Mol Biol,* 372**,** 774-97.

KUFAREVA, I., BUDAGYAN, L., RAUSH, E., TOTROV, M. & ABAGYAN, R. 2007. PIER: protein interface recognition for structural proteomics. *Proteins,* 67**,** 400-17.

LANE, M. D. & SEELIG, B. 2014. Advances in the directed evolution of proteins. *Curr Opin Chem Biol,* 22C**,** 129-136.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-8.

LETAI, A. & FUCHS, E. 1995. The importance of intramolecular ion pairing in intermediate filaments. *Proc Natl Acad Sci U S A,* 92**,** 92-6.

LETUNIC, I. & BORK, P. 2018. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res,* 46**,** D493-d496.

LI, H., CHANG, Y. Y., LEE, J. Y., BAHAR, I. & YANG, L. W. 2017. DynOmics: dynamics of structural proteome and beyond. *Nucleic Acids Res,* 45**,** W374-w380.

LI, Y., DE LA PAZ, J. A., JIANG, X., LIU, R., POKKULANDRA, A. P., BLERIS, L. & MORCOS, F. 2019. Coevolutionary Couplings Unravel PAM-Proximal Constraints of CRISPR-SpCas9. *Biophys J,* 117**,** 1684-1691.

LIANG, N., YAO, M. D., WANG, Y., LIU, J., FENG, L., WANG, Z. M., LI, X. Y., XIAO, W. H. & YUAN, Y. J. 2021. CsCCD2 Access Tunnel Design for a Broader Substrate Profile in Crocetin Production. *J Agric Food Chem,* 69**,** 11626-11636.

LISKOVA, V., BEDNAR, D., PRUDNIKOVA, T., REZACOVA, P., KOUDELAKOVA, T., SEBESTOVA, E., SMATANOVA, I. K., BREZOVSKY, J., CHALOUPKOVA, R. & DAMBORSKY, J. 2015. Balancing the Stability-Activity Trade-Off by Fine-Tuning Dehalogenase Access Tunnels. *ChemCatChem,* 7**,** 648-659.

LISKOVA, V., STEPANKOVA, V., BEDNAR, D., BREZOVSKY, J., PROKOP, Z., CHALOUPKOVA, R. & DAMBORSKY, J. 2017. Different Structural Origins of the Enantioselectivity of Haloalkane Dehalogenases toward Linear β-Haloalkanes: Open-Solvated versus Occluded-Desolvated Active Sites. *Angew Chem Int Ed Engl,* 56**,** 4719-4723.

LIU, C.-Y., SEVERIN LUPALA, C., LYU, C.-J., ZHU, W.-L., WANG, H.-P., JIANG, C.-J., MEI, L.-H., LIU, H.-G. & HUANG, J. 2021. Improving thermostability of (R)-selective amine transaminase from Aspergillus terreus by evolutionary coupling saturation mutagenesis. *Biochemical Engineering Journal,* 167**,** 107926.

LIU, L., YU, H., DU, K., WANG, Z., GAN, Y. & HUANG, H. 2018. Enhanced trypsin thermostability in Pichia pastoris through truncating the flexible region. *Microb Cell Fact,* 17**,** 165.

LOCKLESS, S. W. & RANGANATHAN, R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science,* 286**,** 295-9.

LU, Z., LI, X., ZHANG, R., YI, L., MA, Y. & ZHANG, G. 2019. Tunnel engineering to accelerate product release for better biomass-degrading abilities in lignocellulolytic enzymes. *Biotechnol Biofuels,* 12**,** 275.

LUAN, Z. J., LI, F. L., DOU, S., CHEN, Q. & XU, J. H. 2015. Substrate channel evolution of an esterase for the synthesis of Cilastatin. *Catal Sci Technol,* 5**,** 2622-2629.

MAMONOVA, T. B., GLYAKINA, A. V., GALZITSKAYA, O. V. & KURNIKOVA, M. G. 2013. Stability and rigidity/flexibility-two sides of the same coin? *Biochim Biophys Acta,* 1834**,** 854-66.

MARQUES, S. M., BEDNAR, D. & DAMBORSKY, J. 2018. Computational Study of Protein-Ligand Unbinding for Enzyme Engineering. *Front Chem,* 6**,** 650.

MARQUES, S. M., DANIEL, L., BURYSKA, T., PROKOP, Z., BREZOVSKY, J. & DAMBORSKY, J. 2017a. Enzyme Tunnels and Gates As Relevant Targets in Drug Design. *Med Res Rev,* 37**,** 1095-1139.

MARQUES, S. M., DUNAJOVA, Z., PROKOP, Z., CHALOUPKOVA, R., BREZOVSKY, J. & DAMBORSKY, J. 2017b. Catalytic Cycle of Haloalkane Dehalogenases Toward Unnatural Substrates Explored by Computational Modeling. *J Chem Inf Model,* 57**,** 1970-1989.

MARQUES, S. M., PLANAS-IGLESIAS, J. & DAMBORSKY, J. 2021. Web-based tools for computational enzyme design. *Curr Opin Struct Biol,* 69**,** 19-34.

MCMURROUGH, T. A., DICKSON, R. J., THIBERT, S. M., GLOOR, G. B. & EDGELL, D. R. 2014. Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. *Proc Natl Acad Sci U S A,* 111**,** E2376-83.

MENG, Q., CAPRA, N., PALACIO, C. M., LANFRANCHI, E., OTZEN, M., VAN SCHIE, L. Z., ROZEBOOM, H. J., THUNNISSEN, A. W. H., WIJMA, H. J. & JANSSEN, D. B. 2020. Robust ω-Transaminases by Computational Stabilization of the Subunit Interface. *ACS Catal,* 10**,** 2915-2928.

MENG, S., AN, R., LI, Z., SCHWANEBERG, U., JI, Y., DAVARI, M. D., WANG, F., WANG, M., QIN, M., NIE, K. & LIU, L. 2021. Tunnel engineering for modulating the substrate preference in cytochrome P450BsβHI. *Bioresour Bioprocess,* 8**,** 26.

MINHAS, F., GEISS, B. J. & BEN-HUR, A. 2014. PAIRpred: partner-specific prediction of interacting residues from sequence and structure. *Proteins,* 82**,** 1142-55.

MITCHELL, A. L., ATTWOOD, T. K., BABBITT, P. C., BLUM, M., BORK, P., BRIDGE, A., BROWN, S. D., CHANG, H. Y., EL-GEBALI, S., FRASER, M. I., GOUGH, J., HAFT, D. R., HUANG, H., LETUNIC, I., LOPEZ, R., LUCIANI, A., MADEIRA, F., MARCHLER-BAUER, A., MI, H., NATALE, D. A., NECCI, M., NUKA, G., ORENGO, C., PANDURANGAN, A. P., PAYSAN-LAFOSSE, T., PESSEAT, S., POTTER, S. C., QURESHI, M. A., RAWLINGS, N. D., REDASCHI, N., RICHARDSON, L. J., RIVOIRE, C., SALAZAR, G. A., SANGRADOR-VEGAS, A., SIGRIST, C. J. A., SILLITOE, I., SUTTON, G. G., THANKI, N., THOMAS, P. D., TOSATTO, S. C. E., YONG, S. Y. & FINN, R. D. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res,* 47**,** D351-d360.

MODI, T. & OZKAN, S. B. 2018. Mutations Utilize Dynamic Allostery to Confer Resistance in TEM-1 β-lactamase. *Int J Mol Sci,* 19.

MORCOS, F., PAGNANI, A., LUNT, B., BERTOLINO, A., MARKS, D. S., SANDER, C., ZECCHINA, R., ONUCHIC, J. N., HWA, T. & WEIGT, M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A,* 108**,** E1293-301.

MORRIS, G. M., HUEY, R., LINDSTROM, W., SANNER, M. F., BELEW, R. K., GOODSELL, D. S. & OLSON, A. J. 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem,* 30**,** 2785-91.

MOTOYAMA, T., HIRAMATSU, N., ASANO, Y., NAKANO, S. & ITO, S. 2020. Protein Sequence Selection Method That Enables Full Consensus Design of Artificial l-Threonine 3-Dehydrogenases with Unique Enzymatic Properties. *Biochemistry,* 59**,** 3823-3833.

NEGI, S. S., SCHEIN, C. H., OEZGUEN, N., POWER, T. D. & BRAUN, W. 2007. InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics,* 23**,** 3397-9.

NESTL, B. M. & HAUER, B. 2014. Engineering of Flexible Loops in Enzymes. *ACS Catal,* 4**,** 3201-3211.

NOBILI, A., TAO, Y., PAVLIDIS, I. V., VAN DEN BERGH, T., JOOSTEN, H. J., TAN, T. & BORNSCHEUER, U. T. 2015. Simultaneous use of in silico design and a correlated mutation network as a tool to efficiently guide enzyme engineering. *Chembiochem,* 16**,** 805-10.

NOTREDAME, C., HIGGINS, D. G. & HERINGA, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol,* 302**,** 205-17.

NURIZZO, D., SHEWRY, S. C., PERLIN, M. H., BROWN, S. A., DHOLAKIA, J. N., FUCHS, R. L., DEVA, T., BAKER, E. N. & SMITH, C. A. 2003. The crystal structure of aminoglycoside-3'-phosphotransferase-IIa, an enzyme responsible for antibiotic resistance. *J Mol Biol,* 327**,** 491-506.

OBRECHT, L., JOOSTEN, H.-J., LEE, M., ROZEBOOM, H., BRANIGAN, E., NAISMITH, J., JANSSEN, D., JARVIS, A. & KAMER, P. 2021. Engineering Thermostability in Artificial Metalloenzymes to Increase Catalytic Activity. *ACS Catal,* 11**,** 3620-3627.

OTERI, F., NADALIN, F., CHAMPEIMONT, R. & CARBONE, A. 2017. BIS2Analyzer: a server for co-evolution analysis of conserved protein families. *Nucleic Acids Res,* 45**,** W307-w314.

OTTEN, L. G., HOLLMANN, F. & ARENDS, I. W. 2010. Enzyme engineering for enantioselectivity: from trial-and-error to rational design? *Trends Biotechnol,* 28**,** 46-54.

OVCHINNIKOV, S., PARK, H., VARGHESE, N., HUANG, P. S., PAVLOPOULOS, G. A., KIM, D. E., KAMISETTY, H., KYRPIDES, N. C. & BAKER, D. 2017. Protein structure determination using metagenome sequence data. *Science,* 355**,** 294-298.

PANIZZA, P., CESARINI, S., DIAZ, P. & RODRIGUEZ GIORDANO, S. 2015. Saturation mutagenesis in selected amino acids to shift Pseudomonas sp. acidic lipase Lip I.3 substrate specificity and activity. *Chem Commun (Camb),* 51**,** 1330-3.

PANJKOVICH, A. & DAURA, X. 2014. PARS: a web server for the prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics,* 30**,** 1314-5.

PANWAJEE PAYONGSRIA, D. S., HELEN C. HAILESB, PAUL A. DALBY 2015. Second Generation Engineering of Transketolase for Polar Aromatic Aldehyde Substrates. *Enzyme Microb Technol,* 71, 45-52.

PAREDES, D. I., WATTERS, K., PITMAN, D. J., BYSTROFF, C. & DORDICK, J. S. 2011. Comparative void-volume analysis of psychrophilic and mesophilic enzymes: Structural bioinformatics of psychrophilic enzymes reveals sources of core flexibility. *BMC Struct Biol,* 11**,** 42.

PAVLOVA, M., KLVANA, M., PROKOP, Z., CHALOUPKOVA, R., BANAS, P., OTYEPKA, M., WADE, R. C., TSUDA, M., NAGATA, Y. & DAMBORSKY, J. 2009. Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat Chem Biol,* 5**,** 727-33.

PENKLER, D., SENSOY, Ö., ATILGAN, C. & TASTAN BISHOP, Ö. 2017. Perturbation-Response Scanning Reveals Key Residues for Allosteric Control in Hsp70. *J Chem Inf Model,* 57**,** 1359-1374.

PETERSON, M. E., DANIEL, R. M., DANSON, M. J. & EISENTHAL, R. 2007. The dependence of enzyme activity on temperature: determination and validation of parameters. *Biochem J,* 402**,** 331-7.

PHILLIPS, J. C., BRAUN, R., WANG, W., GUMBART, J., TAJKHORSHID, E., VILLA, E., CHIPOT, C., SKEEL, R. D., KALé, L. & SCHULTEN, K. 2005. Scalable molecular dynamics with NAMD. *J Comput Chem,* 26**,** 1781-802.

PLANAS-IGLESIAS, J., MARQUES, S. M., PINTO, G. P., MUSIL, M., STOURAC, J., DAMBORSKY, J. & BEDNAR, D. 2021. Computational design of enzymes for biotechnological applications. *Biotechnol Adv,* 47**,** 107696.

POREBSKI, B. T. & BUCKLE, A. M. 2016. Consensus protein design. *Protein Eng Des Sel,* 29**,** 245-51.

POROLLO, A. & MELLER, J. 2007. Prediction-based fingerprints of protein-protein interactions. *Proteins,* 66**,** 630-45.

PRAVDA, L., SEHNAL, D., TOUŠEK, D., NAVRáTILOVá, V., BAZGIER, V., BERKA, K., SVOBODOVá VAREKOVá, R., KOCA, J. & OTYEPKA, M. 2018. MOLEonline: a web-based tool for analyzing channels, tunnels and pores (2018 update). *Nucleic Acids Res,* 46**,** W368-w373.

RANJANI, V., JANECEK, S., CHAI, K. P., SHAHIR, S., ABDUL RAHMAN, R. N., CHAN, K. G. & GOH, K. M. 2014. Protein engineering of selected residues from conserved sequence regions of a novel Anoxybacillus alpha-amylase. *Sci Rep,* 4**,** 5850.

RANOUX, A. & HANEFELD, U. 2013. Improving Transketolase. *Topics in Catalysis,* 56**,** 750-764.

RAPP, L. R., MARQUES, S. M., ZUKIC, E., ROWLINSON, B., SHARMA, M., GROGAN, G., DAMBORSKY, J. & HAUER, B. 2021. Substrate Anchoring and Flexibility Reduction in CYP153AM.aq Leads to Highly Improved Efficiency toward Octanoic Acid. *ACS Catal,* 11**,** 3182-3189.

REETZ, M. T. 2011. Laboratory evolution of stereoselective enzymes: a prolific source of catalysts for asymmetric reactions. *Angew Chem Int Ed Engl,* 50**,** 138-74.

REETZ, M. T. & CARBALLEIRA, J. D. 2007. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat Protoc,* 2**,** 891-903.

REETZ, M. T., CARBALLEIRA, J. D. & VOGEL, A. 2006. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew Chem Int Ed Engl,* 45**,** 7745-51.

RIGOLDI, F., DONINI, S., TORRETTA, A., CARBONE, A., REDAELLI, A., BANDIERA, T., PARISINI, E. & GAUTIERI, A. 2020. Rational backbone redesign of a fructosyl peptide oxidase to widen its active site access tunnel. *Biotechnol Bioeng,* 117**,** 3688-3698.

ROGERS, T. A. & BOMMARIUS, A. S. 2010. Utilizing Simple Biochemical Measurements to Predict Lifetime Output of Biocatalysts in Continuous Isothermal Processes. *Chem Eng Sci,* 65**,** 2118-2124.

SAAVEDRA, J. M., AZóCAR, M. A., RODRíGUEZ, V., RAMíREZ-SARMIENTO, C. A., ANDREWS, B. A., ASENJO, J. A. & PARRA, L. P. 2018. Relevance of Local Flexibility Near the Active Site for Enzymatic Catalysis: Biochemical Characterization and Engineering of Cellulase Cel5A From Bacillus agaradherans. *Biotechnol J,* 13**,** e1700669.

SANDSTROM, A. G., WIKMARK, Y., ENGSTROM, K., NYHLEN, J. & BACKVALL, J. E. 2012. Combinatorial reshaping of the Candida antarctica lipase A substrate pocket for enantioselectivity using an extremely condensed library. *Proc Natl Acad Sci U S A,* 109**,** 78-83.

SANKARARAMAN, S., SHA, F., KIRSCH, J. F., JORDAN, M. I. & SJöLANDER, K. 2010. Active site prediction using evolutionary and structural information. *Bioinformatics,* 26**,** 617-24.

SCHENKMAYEROVA, A., PINTO, G. P., TOUL, M., MAREK, M., HERNYCHOVA, L., PLANAS-IGLESIAS, J., DANIEL LISKOVA, V., PLUSKAL, D., VASINA, M., EMOND, S., DöRR, M., CHALOUPKOVA, R., BEDNAR, D., PROKOP, Z., HOLLFELDER, F., BORNSCHEUER, U. T. & DAMBORSKY, J. 2021. Engineering the protein dynamics of an ancestral luciferase. *Nat Commun,* 12**,** 3616.

SCHUELER-FURMAN, O. & WODAK, S. J. 2016. Computational approaches to investigating allostery. *Curr Opin Struct Biol,* 41**,** 159-171.

SEBESTOVA, E., BENDL, J., BREZOVSKY, J. & DAMBORSKY, J. 2014. Computational tools for designing smart libraries. *Methods Mol Biol,* 1179**,** 291-314.

SEEMAYER, S., GRUBER, M. & SöDING, J. 2014. CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics,* 30**,** 3128-30.

SENIOR, A. W., EVANS, R., JUMPER, J., KIRKPATRICK, J., SIFRE, L., GREEN, T., QIN, C., ŽÍDEK, A., NELSON, A. W. R., BRIDGLAND, A., PENEDONES, H., PETERSEN, S., SIMONYAN, K., CROSSAN, S., KOHLI, P., JONES, D. T., SILVER, D., KAVUKCUOGLU, K. & HASSABIS, D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature,* 577**,** 706-710.

SHEIK AMAMUDDY, O., VELDMAN, W., MANYUMWA, C., KHAIRALLAH, A., AGAJANIAN, S., OLUYEMI, O., VERKHIVKER, G. & TASTAN BISHOP, O. 2020. Integrated Computational Approaches and Tools forAllosteric Drug Discovery. *Int J Mol Sci,* 21.

SIMONETTI, F. L., TEPPA, E., CHERNOMORETZ, A., NIELSEN, M. & MARINO BUSLJE, C. 2013. MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Res,* 41**,** W8-14.

SINGH, M. K., SHIVAKUMARASWAMY, S., GUMMADI, S. N. & MANOJ, N. 2017. Role of an N-terminal extension in stability and catalytic activity of a hyperthermostable α/β hydrolase fold esterase. *Protein Eng Des Sel,* 30**,** 559-570.

SINHA, R. & SHUKLA, P. 2019. Current Trends in Protein Engineering: Updates and Progress. *Curr Protein Pept Sci,* 20**,** 398-407.

SKOVSTRUP, S., DAVID, L., TABOUREAU, O. & JøRGENSEN, F. S. 2012. A steered molecular dynamics study of binding and translocation processes in the GABA transporter. *PLoS One,* 7**,** e39360.

SONG, J., LI, F., TAKEMOTO, K., HAFFARI, G., AKUTSU, T., CHOU, K. C. & WEBB, G. I. 2018. PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. *J Theor Biol,* 443**,** 125-137.

SPENCE, M. A., KACZMARSKI, J. A., SAUNDERS, J. W. & JACKSON, C. J. 2021. Ancestral sequence reconstruction for protein engineers. *Curr Opin Struct Biol,* 69**,** 131-141.

SPRENGER, G. A., SCHORKEN, U., SPRENGER, G. & SAHM, H. 1995. Transketolase A of Escherichia coli K12. Purification and properties of the enzyme from recombinant strains. *Eur J Biochem,* 230**,** 525-32.

STEIPE, B., SCHILLER, B., PLUCKTHUN, A. & STEINBACHER, S. 1994. Sequence statistics reliably predict stabilizing mutations in a protein domain. *J Mol Biol,* 240**,** 188-92.

STERNKE, M., TRIPP, K. W. & BARRICK, D. 2019. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc Natl Acad Sci U S A,* 116**,** 11275-11284.

STERNKE, M., TRIPP, K. W. & BARRICK, D. 2020. The use of consensus sequence information to engineer stability and activity in proteins. *Methods Enzymol,* 643**,** 149-179.

STIMPLE, S. D., SMITH, M. D. & TESSIER, P. M. 2020. Directed evolution methods for overcoming trade-offs between protein activity and stability. *AIChE J,* 66.

STOURAC, J., VAVRA, O., KOKKONEN, P., FILIPOVIC, J., PINTO, G., BREZOVSKY, J., DAMBORSKY, J. & BEDNAR, D. 2019. Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. *Nucleic Acids Res,* 47**,** W414-w422.

STRAFFORD, J., PAYONGSRI, P., HIBBERT, E. G., MORRIS, P., BATTH, S. S., STEADMAN, D., SMITH, M. E., WARD, J. M., HAILES, H. C. & DALBY, P. A. 2012. Directed evolution to re-adapt a co-evolved network within an enzyme. *J Biotechnol,* 157**,** 237-45.

SUMBALOVA, L., STOURAC, J., MARTINEK, T., BEDNAR, D. & DAMBORSKY, J. 2018. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res,* 46**,** W356-w362.

SUN, Z., LIU, Q., QU, G., FENG, Y. & REETZ, M. T. 2019. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem Rev,* 119**,** 1626-1665.

SUPLATOV, D., SHARAPOVA, Y., GERASEVA, E. & ŠVEDAS, V. 2020. Zebra2: advanced and easy-to-use web-server for bioinformatic analysis of subfamily-specific and conserved positions in diverse protein superfamilies. *Nucleic Acids Res,* 48**,** W65-w71.

TAYLOR, J. L., PRICE, J. E. & TONEY, M. D. 2015. Directed evolution of the substrate specificity of dialkylglycine decarboxylase. *Biochim Biophys Acta,* 1854**,** 146-55.

THOMAS, A., CUTLAN, R., FINNIGAN, W., VAN DER GIEZEN, M. & HARMER, N. 2019. Highly thermostable carboxylic acid reductases generated by ancestral sequence reconstruction. *Commun Biol,* 2**,** 429.

TIAN, H., JIANG, X. & TAO, P. 2021. PASSer: Prediction of Allosteric Sites Server. *Mach Learn Sci Technol,* 2.

TOSCANO, M. D., WOYCECHOWSKY, K. J. & HILVERT, D. 2007. Minimalist active-site redesign: teaching old enzymes new tricks. *Angew Chem Int Ed Engl,* 46**,** 3212-36.

TROTT, O. & OLSON, A. J. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem,* 31**,** 455-61.

TUNYASUVUNAKOOL, K., ADLER, J., WU, Z., GREEN, T., ZIELINSKI, M., ŽÍDEK, A., BRIDGLAND, A., COWIE, A., MEYER, C., LAYDON, A., VELANKAR, S., KLEYWEGT, G. J., BATEMAN, A., EVANS, R., PRITZEL, A., FIGURNOV, M., RONNEBERGER, O., BATES, R., KOHL, S. A. A., POTAPENKO, A., BALLARD, A. J., ROMERA-PAREDES, B., NIKOLOV, S., JAIN, R., CLANCY, E., REIMAN, D., PETERSEN, S., SENIOR, A. W., KAVUKCUOGLU, K., BIRNEY, E., KOHLI, P., JUMPER, J. & HASSABIS, D. 2021. Highly accurate protein structure prediction for the human proteome. *Nature,* 596**,** 590-596.

VAN DER SPOEL, D., LINDAHL, E., HESS, B., GROENHOF, G., MARK, A. E. & BERENDSEN, H. J. 2005. GROMACS: fast, flexible, and free. *J Comput Chem,* 26**,** 1701-18.

VAZQUEZ-FIGUEROA, E., CHAPARRO-RIGGERS, J. & BOMMARIUS, A. S. 2007. Development of a thermostable glucose dehydrogenase by a structure-guided consensus concept. *Chembiochem,* 8**,** 2295-301.

VERDONK, M. L., COLE, J. C., HARTSHORN, M. J., MURRAY, C. W. & TAYLOR, R. D. 2003. Improved protein-ligand docking using GOLD. *Proteins,* 52**,** 609-23.

VRIEND, G. 1990. WHAT IF: a molecular modeling and drug design program. *J Mol Graph,* 8**,** 52-6, 29.

WANG, C., HUANG, R., HE, B. & DU, Q. 2012. Improving the thermostability of alpha-amylase by combinatorial coevolving-site saturation mutagenesis. *BMC Bioinform,* 13**,** 263.

WANG, J. B., LI, G. & REETZ, M. T. 2017. Enzymatic site-selectivity enabled by structure-guided directed evolution. *Chem Commun (Camb),* 53**,** 3916-3928.

WANG, T., LIANG, C., HOU, Y., ZHENG, M., XU, H., AN, Y., XIAO, S., LIU, L. & LIAN, S. 2020a. Small design from big alignment: engineering proteins with multiple sequence alignment as the starting point. *Biotechnol Lett,* 42**,** 1305-1315.

WANG, X., JING, X., DENG, Y., NIE, Y., XU, F., XU, Y., ZHAO, Y. L., HUNT, J. F., MONTELIONE, G. T. & SZYPERSKI, T. 2020b. Evolutionary coupling saturation mutagenesis: Coevolution-guided identification of distant sites influencing Bacillus naganoensis pullulanase activity. *FEBS Lett,* 594**,** 799-812.

WEINKAM, P., PONS, J. & SALI, A. 2012. Structure-based model of allostery predicts coupling between distant sites. *Proc Natl Acad Sci U S A,* 109**,** 4875-80.

WIJMA, H. J., FLOOR, R. J., BJELIC, S., MARRINK, S. J., BAKER, D. & JANSSEN, D. B. 2015. Enantioselective Enzymes by Computational Design and In Silico Screening. *Angew Chem Int Ed Engl,* 54, 3726-30.

WILKINSON, H. C. & DALBY, P. A. 2021. Fine-tuning the activity and stability of an evolved enzyme active-site through noncanonical amino-acids. *Febs j,* 288**,** 1935-1955.

WITTMANN, B. J., YUE, Y. & ARNOLD, F. H. 2021. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst,* 12, 1026-45.
WU, T. H., CHEN, C. C., CHENG, Y. S., KO, T. P., LIN, C. Y., LAI, H. L., HUANG, T. Y., LIU, J. R. & GUO, R. T. 2014. Improving specific activity and thermostability of Escherichia coli phytase by structure-based rational design. *J Biotechnol,* 175**,** 1-6.
XIE, Y., AN, J., YANG, G., WU, G., ZHANG, Y., CUI, L. & FENG, Y. 2014. Enhanced enzyme kinetic stability by increasing rigidity within the active site. *J Biol Chem,* 289**,** 7994-8006.
XU, Y., WANG, S., HU, Q., GAO, S., MA, X., ZHANG, W., SHEN, Y., CHEN, F., LAI, L. & PEI, J. 2018. CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Res,* 46**,** W374-w379.
YANG, J., ANISHCHENKO, I., PARK, H., PENG, Z., OVCHINNIKOV, S. & BAKER, D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A,* 117**,** 1496-1503.
YANG, K. K., WU, Z. & ARNOLD, F. H. 2019. Machine-learning-guided directed evolution for protein engineering. *Nat Methods,* 16**,** 687-694.
YANG, W. & LAI, L.-H. 2016. Computational design of proteins with novel structure and functions. *Chinese Physics B,* 25, 018702.
YU, H. & DALBY, P. A. 2018a. Coupled molecular dynamics mediate long- and short-range epistasis between mutations that affect stability and aggregation kinetics. *Proc Natl Acad Sci U S A,* 115**,** E11043-E11052.
YU, H. & DALBY, P. A. 2018b. Exploiting correlated molecular-dynamics networks to counteract enzyme activity-stability trade-off. *Proc Natl Acad Sci U S A,* 115**,** E12192-e12200.
YU, H. & DALBY, P. A. 2020. A beginner's guide to molecular dynamics simulations and the identification of cross-correlation networks for enzyme engineering. *Methods Enzymol,* 643**,** 15-49.
YU, H., HERNáNDEZ LóPEZ, R. I., STEADMAN, D., MéNDEZ-SáNCHEZ, D., HIGSON, S., CáZARES-KöRNER, A., SHEPPARD, T. D., WARD, J. M., HAILES, H. C. & DALBY, P. A. 2020. Engineering transketolase to accept both unnatural donor and acceptor substrates and produce α-hydroxyketones. *Febs j,* 287**,** 1758-1776.
YU, H. & HUANG, H. 2014. Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnol Adv,* 32**,** 308-15.
YU, H., YAN, Y., ZHANG, C. & DALBY, P. A. 2017. Two strategies to engineer flexible loops for improved enzyme thermostability. *Sci Rep,* 7**,** 41212.
YU, H., ZHAO, Y., GUO, C., GAN, Y. & HUANG, H. 2015. The role of proline substitutions within flexible regions on thermostability of luciferase. *Biochim Biophys Acta,* 1854**,** 65-72.
YU, Z., YU, H., TANG, H., WANG, Z., WU, J., YANG, L. & XU, G. 2021. Site-specifically Incorporated Non-Canonical Amino Acids into Pseudomonas alcaligenes Lipase to Hydrolyze L-menthol Propionate among the Eight Isomers. *ChemCatChem,* 13**,** 2691-2701.
ZENG, B., ZHOU, Y., YI, Z., ZHOU, R., JIN, W. & ZHANG, G. 2021. Highly thermostable and promiscuous β-1,3-xylanasen designed by optimized ancestral sequence reconstruction. *Bioresour Technol,* 340**,** 125732.
ZHANG, Q. C., DENG, L., FISHER, M., GUAN, J., HONIG, B. & PETREY, D. 2011. PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res,* 39**,** W283-7.
ZHANG, T., ZHANG, H., CHEN, K., SHEN, S., RUAN, J. & KURGAN, L. 2008. Accurate sequence-based prediction of catalytic residues. *Bioinformatics,* 24**,** 2329-2338.
ZHANG, W., ZHANG, Y., HUANG, J., CHEN, Z., ZHANG, T., GUANG, C., MU, W. J. J. O. A. & CHEMISTRY, F. 2018. Thermostability Improvement of the d-Allulose 3-Epimerase from Dorea sp. CAG317 by Site-Directed Mutagenesis at the Interface Regions. *J Agric Food Chem,* 66, 5593-601.
ZHANG, Y. & VOTH, G. A. 2011. A Combined Metadynamics and Umbrella Sampling Method for the Calculation of Ion Permeation Free Energy Profiles. *J Chem Theory Comput,* 7**,** 2277-2283.

ZHAO, J. X., YANG, L., GU, Z. N., CHEN, H. Q., TIAN, F. W., CHEN, Y. Q., ZHANG, H. & CHEN, W. 2010. Stabilization of the single-chain fragment variable by an interdomain disulfide bond and its effect on antibody affinity. *Int J Mol Sci,* 12**,** 1-11.

ZHENG, W., YU, H., FANG, S., CHEN, K., WANG, Z., CHENG, X., XU, G., YANG, L. & WU, J. 2021. Directed Evolution of l-Threonine Aldolase for the Diastereoselective Synthesis of β-Hydroxy-α-amino Acids. *ACS Catal,* 11**,** 3198-3205.

ZHU, C., CHEN, Y., ISUPOV, M., LITTLECHILD, J., SUN, L., XIAODONG, L., WANG, Q., GONG, H., DONG, P., ZHANG, N. & WU, Y. 2021. Structural Insights into a Novel Esterase from the East Pacific Rise and Its Improved Thermostability by a Semirational Design. *J Agric Food Chem,* 69**,** 1079-1090.

ZHU, W. L., HU, S., LV, C. J., ZHAO, W. R., WANG, H. P., MEI, J. Q., MEI, L. H. & HUANG, J. 2019. A Single Mutation Increases the Thermostability and Activity of Aspergillus terreus Amine Transaminase. *Molecules,* 24, 1194.