Creating New Enzymes with Evolution and Intuition

Patrick J. Almhjell and Frances H. Arnold, California Institute of Technology

Introduction

Enzymes are nature's catalysts, performing the chemical reactions that living organisms use to extract materials and energy from their environments and create new life. Consider how sunlight and CO_2 are converted into sugars by plants, and how those sugars are later metabolized by the organisms that rely on plants for food. These reactions are all catalyzed by enzymes. Even the conversion of solubilized CO_2 in the blood (in the form of carbonate) to gaseous CO_2 that is exhaled and the metabolism of different pharmaceuticals to remove them from a human circulatory system are performed by these remarkable biological machines. Guided by evolution, enzymes have been optimized for their particular reactions and can exhibit extraordinary rate enhancements and selectivities compared to uncatalyzed reactions and other catalysts.

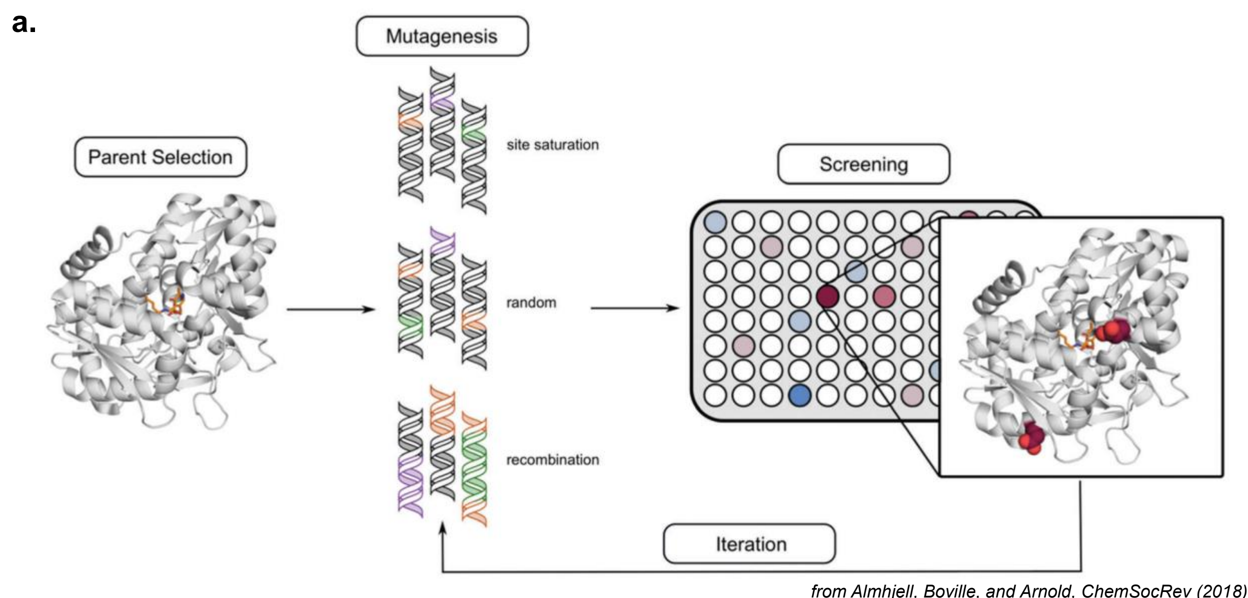
Enzymes are made of linear arrangements of the 20 canonical amino acids that fold into a complex three-dimensional structure. They often bind different cofactors to facilitate their reactions and are sometimes further modified by post-translational modifications which alter their structures and functions. Optimization of enzymes via evolution occurs due to the action of two main events: mutations occur in the DNA encoding the enzyme, followed by selection on its function, or *phenotype*. Mutations that improve function can become enriched in the population. This process enables organisms to co-opt and consume the resources around them; those that do this better have a better chance at passing their genes to the next generation. This process has also given rise to the great diversity of life that has adapted to occupy nearly every niche of the world.

The powerful design algorithm of evolution has been brought into the laboratory in the form of directed evolution, which can be used to engineer enzymes and other proteins to have properties useful for addressing human needs. Analogous to natural evolution, directed evolution uses mutagenesis, a method of creating genetic diversity, followed by some form of selection or screen to identify variant enzymes with improved properties. This cycle can be repeated until the property is sufficiently improved (**Fig 1a**). For the most part, directed evolution is conceptually and technically straightforward: once an enzyme is identified that displays a low level of activity for a desired function, mutagenesis and screening for improved function will often provide enhancements. However, that first step, the identification of enzymes with some initial activity, can be far from straightforward. Both in nature and in the laboratory, we often find ourselves wondering how exactly *new* functions arose over the course of natural evolution.

Similar to the tenet of cellular biology that "*omnis cellula e cellula*"—or "all cells come from cells"—enzymes are derived from earlier enzymes and other proteins, acquiring new functions over the course of evolution. What circumstances allow an enzyme that performs one function to evolve into another with a different function? This question is not completely answered, but this process has generally been observed to occur incrementally in nature, through small changes and chance—the right change or the right set of new conditions—over a long time (**Fig 1b**). We enzyme engineers, however, wish to create enzymes that address current, time-sensitive problems and thus do not have the luxury of waiting "evolutionary timescales" before a solution appears and is optimized. Furthermore, we might want to go in a different direction and create enzymes that serve us rather than support the organism that makes it. How, then, can we make faster, more directed jumps toward new catalytic activities? As we will see in this chapter, it is often about using one's chemical intuition, the *how* and *why* a reaction might happen, and making the right changes under the right set of conditions.

Evolution is not just a thing of the past—it's happening right now

First, it is important to note that not all natural evolution requires millennia. Given a sufficiently strong selection pressure or advantage, enzymes and even entire organisms can adapt to a new environment in a matter of years, or even months. Examples of rapid evolution can be attributed to changes in single enzymes within an organism, such as the appearance of resistance to an antibiotic or the emergence of bacterial strains which can consume plastic, a consequence of the plastic waste permeating our ecosystems.



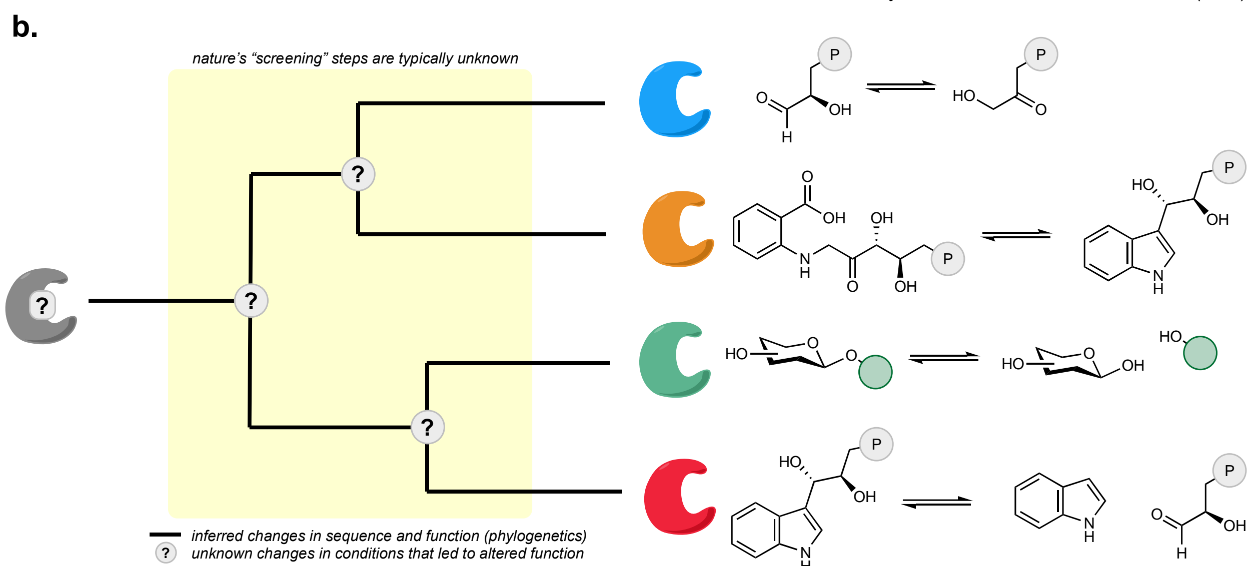
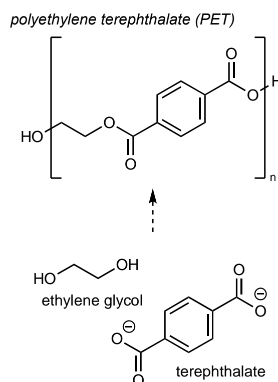


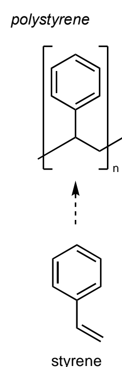
Fig 1. Evolution in the laboratory and in nature. a. The process of directed enzyme evolution starts with finding a "parent" enzyme with some initial activity for the desired reaction. Mutations are made to the gene encoding this enzyme via mutagenesis techniques, and then these enzyme variants are screeded for improved activity. An improved variant or variants can then be subjected to additional rounds of mutagenesis and screening until a sufficiently active enzyme is created. **b.** Natural evolution is often visualized with phylogenetic trees that map the branching changes in protein sequence from a progenitor. While excellent at representing speciation and diversification events, it does not tell us about the specific conditions that led to the observed diversification of enzyme functions.

Plastics, such as polyethylene terephthalate (PET), are organic polymers that were nonexistent in nature until recently (**Fig 2a**). They often contain very stable bonds that are uncommon—or at least unusually decorated—compared to those found in biology, which makes them difficult for enzymes to access and break down. This prevents organisms from converting these polymers into digestible monomeric units that can be used as a carbon source. Furthermore, some of these polymers assemble into materials with high crystallinity at all but extremely high temperatures, which reduces their accessibility to enzymes in the first place. These qualities mean that plastics accumulate in the environment. Nonetheless, other materials that *are* biodegradable share some of these qualities (**Fig 2b**). Wood, comprised of crystalline cellulose and highly cross-linked lignin polymers, is an example of a material that is difficult to break down into monomers, yet it is obviously biodegradable. Only a select few organisms, however, can efficiently perform this function, such as fungi and microbes in the guts of termites. Is plastic, a product of human engineering, fundamentally different from biological materials like cellulose? Or have organisms simply not had sufficient time to adapt to these new carbon sources?

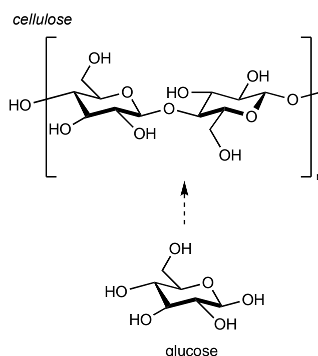
Recently, researchers identified a strain of bacteria named *Ideonella sakaiensis*, isolated from a plastic bottle recycling facility, that is capable of hydrolyzing PET and metabolizing it as its primary carbon source to sustain growth (**Fig 2c**). There are two enzymes involved in this process, one that breaks down the PET polymer (a PETase) into its mono-(2-hydroxyethyl)terephthalate (MHET) monomers and one that then breaks down MHET (a MHETase) into two pieces. These pieces—ethylene glycol and terephthalate—can then enter a cell's metabolic cycle as carbon sources. While these new enzymes perform their functions well enough to sustain the growth of *I. sakaiensis,* they are slow compared to enzymes that perform other, similar hydrolysis reactions. Therefore, there is likely still room for improvement via natural or laboratory evolution, presenting an exciting opportunity for laboratory evolution to "compete" with natural evolution.

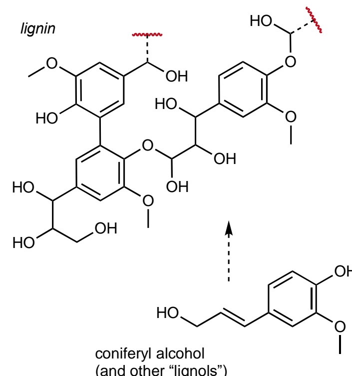
a. Examples of plastics





b. Examples of wood polymers





c. Biodegradation of PET by *Ideonella sakaiensis*

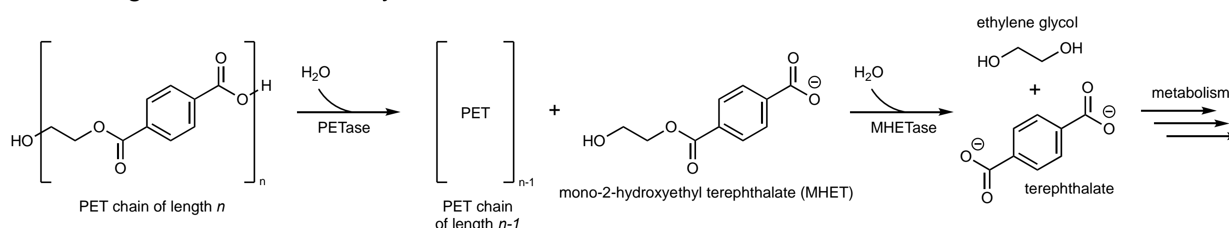


Fig 2. Organic polymers are potentially rich carbon sources if they can be broken down. **a.** Representative structures and monomeric units of plastics. **b.** Representative structures and monomeric units of the polymers found in wood. **c.** The bacterium *Ideonella sakaiensis*, recently isolated from a plastic recycling plant, can degrade PET into pieces that it can use as carbon sources.

Directed evolution readily re-optimizes enzymes

Some of the most successful examples of directed laboratory evolution have involved asking an enzyme to perform its native function in a non-native way, such as under harsher conditions or on a substrate that is not its natural substrate. The enzyme carbonic anhydrase rapidly catalyzes the interconversion of CO_2, a poorly soluble gas, and bicarbonate, a highly soluble salt (**Fig 3a**). In fact, this reaction happens so fast—up to 1 million times per second—that it is limited only by the rate of diffusion of the substrate to the enzyme, which has afforded carbonic anhydrase the title of a "perfect" enzyme. Engineering the enzyme to perform the same transformation but for longer periods in hot and highly alkaline environments has enabled its application within carbon capture systems. These systems typically use a basic solvent, such as a solution of an amine, that helps absorb the CO_2 and retain it as bicarbonate within water. The solvent is then cycled to a different chamber at a higher temperature to facilitate the desorption of bicarbonate from the solvent, releasing a stream of pure CO_2 that can then be captured and stored. Unfortunately, needing to use the same amine for absorption and desorption presents a problem: amines that are faster at absorbing CO_2 are typically slower at releasing it, requiring even higher temperatures and more energy input to make these systems viable for carbon capture. With the use of a prodigious catalyst like carbonic anhydrase, however, to speed up the interconversion of CO_2 and bicarbonate, the kinetics of amine absorption become less important. Amines with lower desorption temperatures could thus be coupled with a catalyst to create a better system (**Fig 3b**). After engineering carbonic anhydrase to withstand the conditions of a carbon capture system—a pH of 10.0 with cycling between 25 °C for absorption and 87 °C for desorption—the catalyzed system captured roughly two-thirds of all CO_2 released from a power plant flue, up to 25-fold more CO_2 than without the enzyme, with no noticeable loss in enzyme activity over 60 hours.

A striking example of engineering an enzyme to perform on a substrate analog is found in the directed evolution of a transaminase to synthesize sitagliptin, an anti-diabetes drug that contains a chiral amine (**Fig 3c**). Transaminases are a broad class of enzymes that catalyze the interconversion of ketones and amines. While ketones lack chirality, the conversion of a ketone to an amine sets a new chiral center, which enzymes can perform with exquisite selectivity to afford only a single enantiomer. However, no transaminase had been known to work on a molecule as large and complex as prositagliptin, the achiral precursor that could be transaminated to sitagliptin. This was overcome by a clever strategy called substrate walking in which an enzyme is iteratively adapted to substrate analogs progressively more similar to the final target. Protein engineers were first able to increase the activity of a transaminase toward a fragment of prositagliptin that was closer in structure to the native substrate of the enzyme (**Fig 3d**). Once activity was improved toward this substrate, the enzyme then also displayed some activity toward the full prositagliptin molecule. This is called substrate promiscuity, where an enzyme can react with multiple substrates rather than just the one for which it was evolved, usually to a lower degree (**Fig 3d, purple region**). Once it was possible to reliably screen for activity on the desired prositagliptin substrate, the transaminase could be evolved to yield a nearly perfectly enantioselective and highly efficient catalyst for the synthesis of sitagliptin at industrial scales.

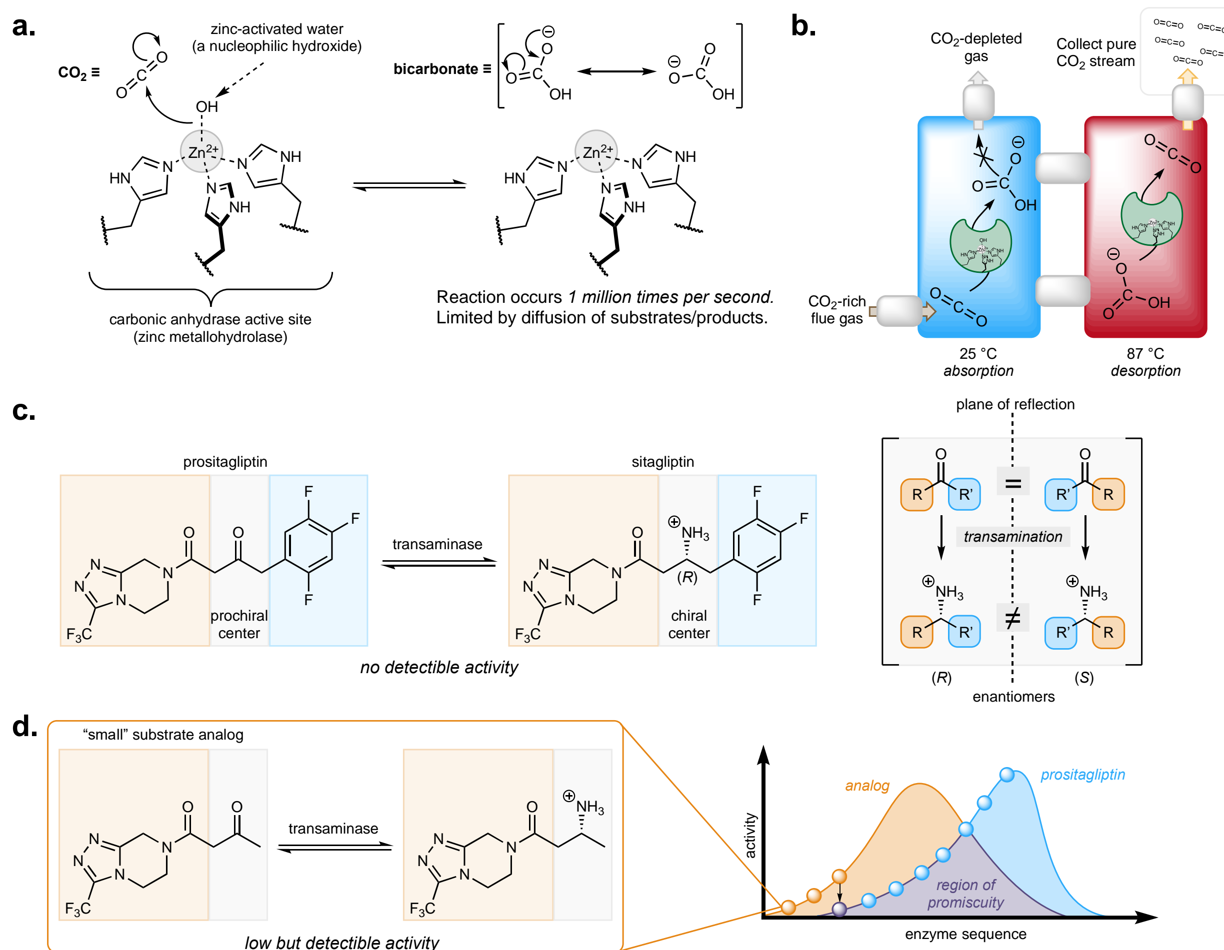


Fig 3. Re-optimizing enzymes to new conditions and substrate analogs. a. The reaction of carbonic anhydrase, a zinc metallohydrolase, which interconverts CO_2 and bicarbonate. **b.** Schematic of a carbonic anhydrase-catalyzed carbon capture system. **c.** Sitagliptin synthesis requires the installation of a chiral amine, which can theoretically be accomplished by a transaminase in a highly selective fashion, but no transaminase was identified to act on the precursor prositagliptin. **d.** A smaller prositagliptin analog was accepted by a transaminase (initial orange sphere), which could be evolved until an enzyme variant was identified with activity toward the full prositagliptin molecule (purple sphere). This could then be evolved into an industrially useful catalyst (blue spheres). This process is called a "substrate walk", and relies on substrate promiscuity—the ability of an enzyme to perform its reaction with a different substrate—and the selection of appropriate intermediate substrates.

But how can we evolve enzymes to perform *new* functions?

What qualifies as a "new" catalytic function is somewhat arbitrary. The "function" of an enzyme is often equated to the reaction it performs. However, reactions are distinguished not just by the specific products that are made, but also by the character of the transition states and intermediates, referred to as the catalytic "mechanism". Two reactions that make the same product but proceed through very different mechanisms might be considered to be more different than two reactions that create different products but do so through an identical mechanism. We will consider the two primary variables of an enzyme's function to be the specific mechanism and the nature of the bonds being broken and/or formed (**Fig 4a**). In the previous examples, the evolved enzymes still used their native mechanisms: the carbonic anhydrase was adapted to a new environment whereas the transaminase was adapted toward a more decorated substrate with the same reactive moiety—the same bonds being broken and formed. Here we will consider an enzyme to have a *new* catalytic function if it is working on a different substrate class or reactive

moiety—even if the reaction is mechanistically similar—or if it is reacting through a completely new mechanism. Through chemical intuition and an in-depth understanding of a reaction, we can often visualize how an enzyme could in principle perform a whole new reaction.

As with an enzyme catalyzing the same reaction on different substrates, we say an enzyme is catalytically promiscuous when it can catalyze multiple different reactions (**Fig 4b**). This concept is often used to explain an observed reaction. Take, for example, the fact that carbonic anhydrase has promiscuous esterase activity, the ability to hydrolyze an ester into an alcohol and carboxylic acid (**Fig 4c**). We can intuit how this reaction occurs in a similar way to the native mechanism, but with a new bond-breaking step from the carbonate-like intermediate (compare **Fig 3a** to **Fig 4c**). However, when looking to create a new enzyme for a reaction that is not known to be catalyzed by an enzyme, we instead must hypothesize how a *new* reaction might be catalyzed by

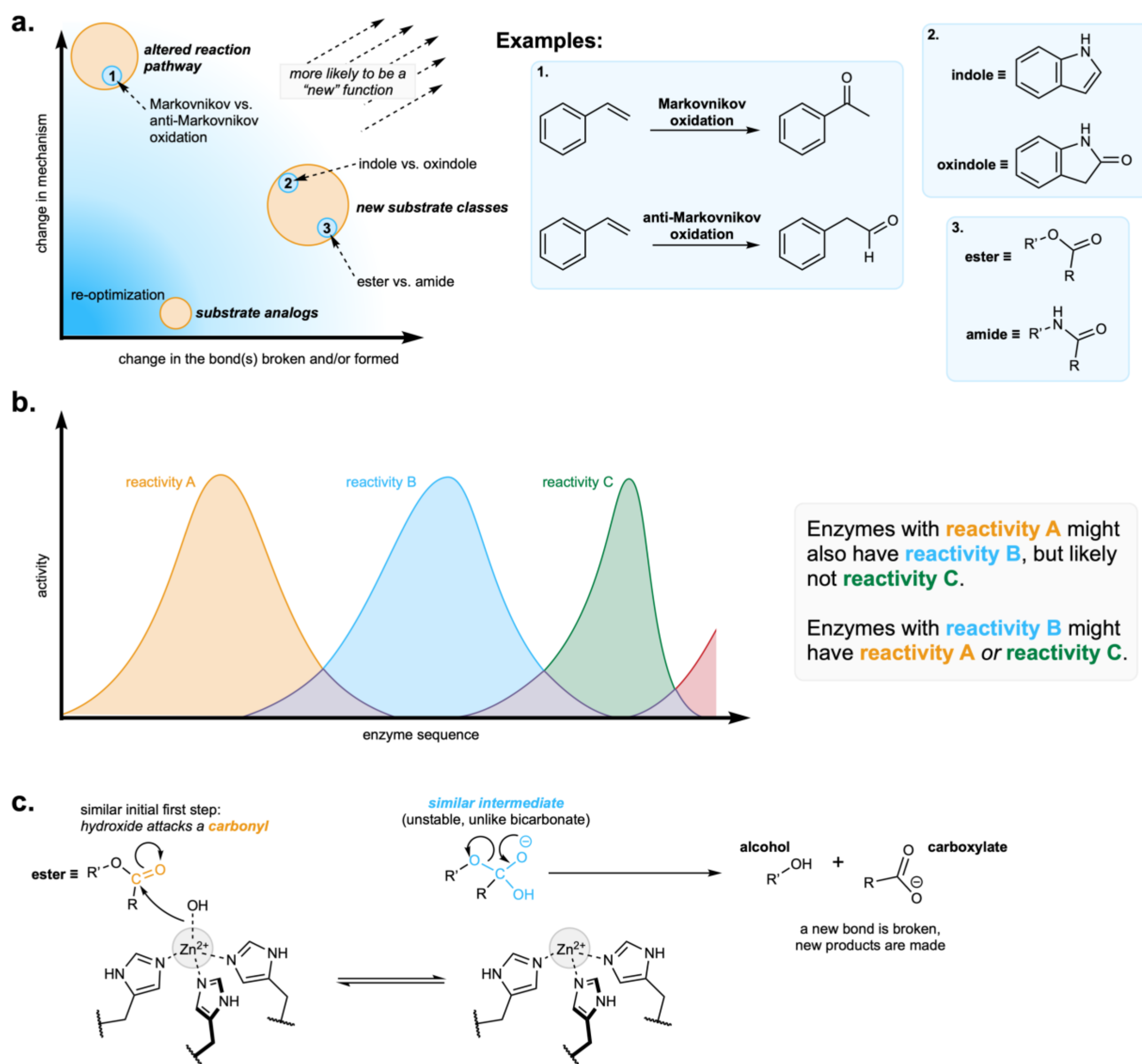


Fig 4. The appearance of new enzyme functions. a. A two-dimensional representation of enzyme functional differences, with the two primary axes being the character of the bonds that are broken and formed and the nature of the reaction mechanism. Specific examples are highlighted. **b.** A cartoon representation of catalytic promiscuity, the ability of an enzyme with a given reactivity to have (or not have) other reactivities. Such reactivities can be enhanced and tuned by directed evolution. **c.** Likely mechanisms of the promiscuous esterase activity of carbonic anhydrase, analogous to (but quite different from) its native reactivity.

an *existing* enzyme *a priori*, and provide the appropriate substrates (and sometimes even cofactors). This process can guide us to identify the initial activity required to begin directed evolution of a new enzyme function. The final section will discuss how this approach has been used in the laboratory evolution of tryptophan synthase.

Tryptophan synthase: a powerful biocatalyst for amino acid synthesis
Tryptophan synthase (TrpS) is an ancient enzyme responsible for the final two steps in the biosynthesis of L-tryptophan (Trp), one of the twenty amino acids used in protein synthesis. It is present in all organisms with the exception of animals, which instead must obtain it in some other way (it is *essential*), such as through diet or with the help of gut microbes. TrpS is commonly found as a heterodimeric complex of two subunits, the α-subunit (TrpA) and the β-subunit (TrpB), which are in some cases tethered together in a single protein chain. This close proximity is important. TrpA converts indole glycerol phosphate (IGP) into glyceraldehyde 3-phosphate (a three-carbon sugar that re-enters metabolism) and indole, a small and somewhat toxic molecule that can readily escape the cell or cause a stress response within (**Fig 5a**; TrpA). Escaping the cell would result in loss of the resources that went into making IGP in the first place, and a stress response is obviously undesirable. Before indole can do either of these things, however, it is quickly shuttled into the adjacent active site of TrpB where it is immediately converted to Trp.

Two key features contribute to the efficiency of this process. The first is that TrpB uses its pyridoxal 5'-phosphate (PLP) cofactor to convert L-serine (Ser) into a highly electrophilic intermediate—the amino-acrylate—that is poised to react with the nucleophilic indole substrate as soon as they meet, proceeding via electrophilic aromatic substitution (**Fig 5a**; TrpB). This intermediate can be degraded through a different reaction pathway, but this side activity is reduced by the second key feature: allostery. Allostery is a process of communication between the two subunits, such that their functions are coordinated through different conformational states. Upon binding IGP, TrpA signals TrpB to convert Ser to the amino-acrylate; once done, TrpB signals TrpA to produce indole. The amino-acrylate intermediate can, theoretically, react with any other nucleophile that it might encounter, but it is sequestered in time and space to only appear when and where indole is available. Thus, the highly reactive intermediate and allosterically controlled structural changes work together to meet the biological demands of synthesizing Trp quickly and with minimal loss of resources.

Can we change the function of TrpB to make very different amino acids?
Outside of the demands of biology, however, one can envision TrpB to be capable of using its catalytic machinery for other reactions. The amino-acrylate is a potent electrophile, able to react with numerous different nucleophilic species. However, it is only given the chance to encounter indole in its native environment. When placed in an *in vitro* setting and engineered to favor the formation of the amino-acrylate without relying on allosteric signals from TrpA, TrpB became a highly general amino acid synthase that can take on a broad array of new catalytic activities (**Fig 5b**). Provided a competent nucleophile, even those significantly different from indole that undergo C–C bond formation via a different mechanism, variants of this enzyme can produce numerous other *noncanonical* amino acids (ncAAs) that are not part of the twenty used in protein synthesis but are important pharmaceuticals, building blocks, and biological probes.

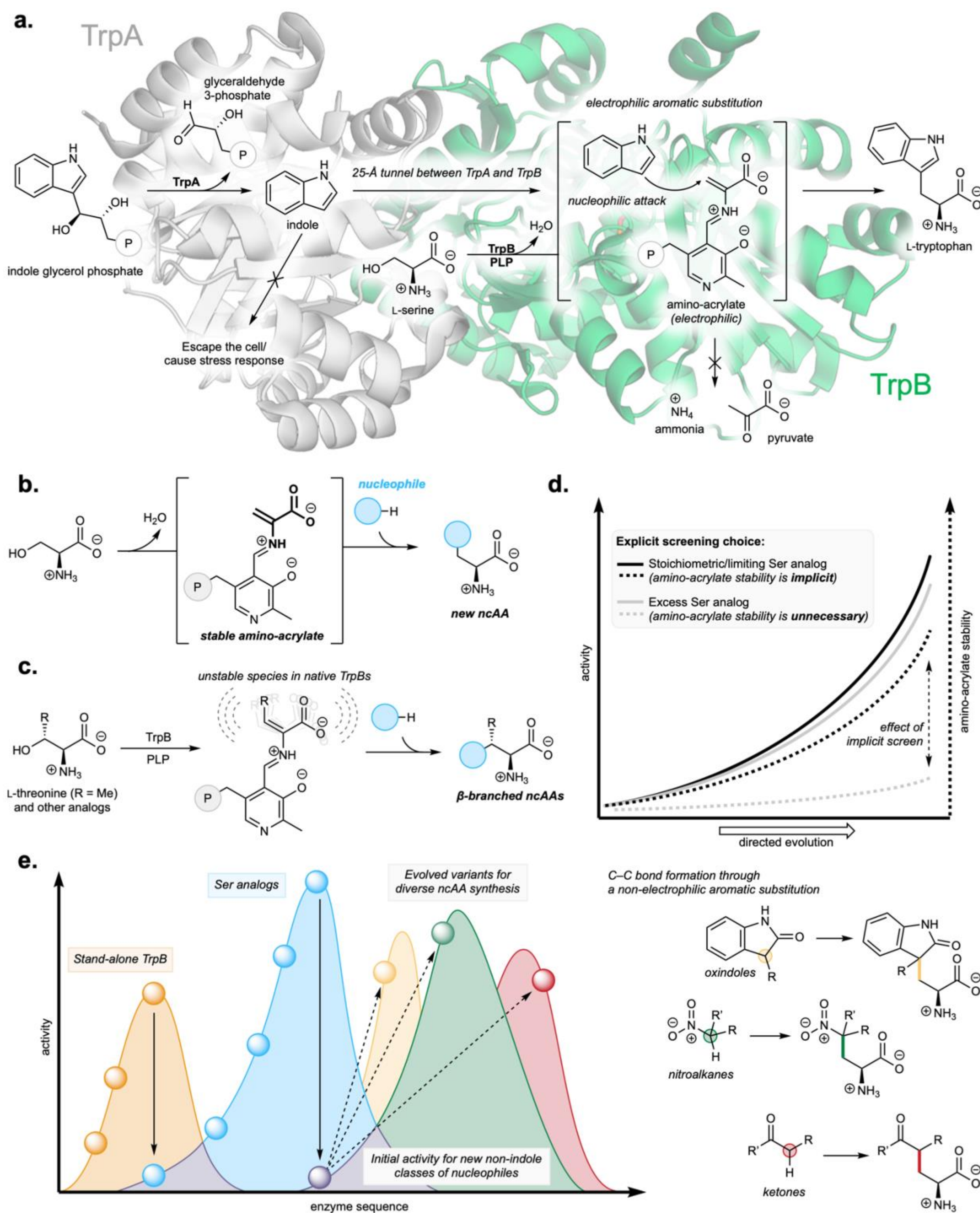


Fig 5. Directed evolution of tryptophan synthase. a. Tryptophan synthase (TrpS) converts indole glycerol phosphate (IGP) to indole and glyceraldehyde 3-phosphate within its TrpA subunit, then couples indole and L-serine (Ser) to make L-tryptophan in its TrpB subunit, avoiding non-productive outcomes. **b.** TrpB can work as a general noncanonical amino acid (ncAA) synthase, coupling a nucleophile with the electrophilic amino-acrylate intermediate. **c.** Ser analogs like L-threonine can also be used to generate an amino-acrylate analog, but directed evolution was required to stabilize this intermediate. **d.** The amino-acrylate could be implicitly stabilized by screening under conditions where the Ser analog was provided in stoichiometric (or limiting) quantities. The hypothetical curves demonstrate how using excess Ser might still improve activity with a given nucleophile to the same extent as limiting Ser, but would not stabilize the amino-acrylate as much. **e.** Fitness landscapes for various activities, showing how improving amino-acrylate stability might then unlock new activities with less reactive nucleophiles that proceed through a different, non-native bond-forming mechanism.

Many of these new catalytic activities—particularly those providing non-tryptophan ncAAs—were not detectible, however, with the native enzyme. They needed to be coaxed out by directed evolution. Protein engineers applied the two concepts described above: they optimized the TrpB subunit to new conditions and adapted it to new substrates. Directed evolution to improve native Trp synthesis under new *in vitro* conditions meant that the reaction took place in the absence of TrpA, with indole already in the reaction mixture, and with TrpB no longer having to discriminate against other nucleophiles as indole was the only one. The enzyme also did not have to discriminate against Ser analogs and could use L-threonine (Thr) and other Ser analogs to generate amino-acrylate analogs (**Fig 5c**). However, this raised another potential issue: the amino-acrylate could still be degraded, and its analogs were even less stable.

How does one then increase the stability of these species? An apt adage in directed evolution is "you get what you screen for". If you *want* a stable amino-acrylate you need to *screen for* a stable amino-acrylate, either explicitly or implicitly. One way to accomplish this implicitly is to use stoichiometric amounts of substrates in the screening reaction. Therefore, the only way to achieve 100% yield is to stabilize the amino-acrylate so that it is not degraded over the course of the reaction (**Fig 5d**). Indeed, when TrpB was evolved using stoichiometric amounts of substrates, greater amino-acrylate stability came right along with increased yield. Additional engineering using poorly reactive indole analogs (which required an even more stable amino-acrylate intermediate) yielded a remarkably efficient Trp-analog synthase.

At this point, the enzyme had been through very little *functional* change. Under the conditions used for directed evolution, the enzyme merely favored certain intermediates and disfavored non-productive pathways. Its "native" chemistry—synthesizing Trp—was relatively unchanged, it just did this in the absence of TrpA and without needing to discriminate against other nucleophiles and Ser analogs. However, by stabilizing the reactive intermediate, these new enzyme variants were capable of reacting with completely new substrate classes, those previously inaccessible with the native TrpB enzymes. With some chemical intuition, new nucleophilic species were identified—oxindoles, nitroalkanes, ketones, and more—that underwent a carbon–carbon (C–C) bond-forming reaction with the amino-acrylate, despite looking and behaving quite differently from the native aromatic indole substrate (**Fig 5e**). While the initial activities were typically low, they provided starting points for directed evolution that were absent from the native enzymes and could be further evolved into productive biocatalysts.

Who cares about this?

Just as life adapts to new challenges and opportunities through the evolution of enzymes, directed enzyme evolution has provided a reliable and powerful approach to address human needs. As we saw in this chapter, enzymes can provide solutions to problems in fields as disparate as pharmaceutical manufacturing to industrial carbon capture. New technologies have certainly improved our ability to optimize enzymes, such as improvements in DNA synthesis and sequencing, analytical instruments and techniques for screening, and computational methods that can efficiently learn from collected data. (In fact, next-generation DNA sequencing uses enzymes that themselves have been the subject of directed evolution!) Our ability to identify enzymes with *new* functions—which we can then throw into the process of directed evolution—can start with chemical intuition and in some instances requires the connection of multiple steps to form a path from a known enzyme function to a new one. By appreciating the fundamentals of natural and laboratory enzyme evolution, we can improve our chances of quickly creating new enzyme functions when they are needed.

Key questions that remain to be answered:
What if there simply is no enzyme that performs a desired reaction? We are currently limited by the enzymes that already exist, either through natural or laboratory evolution, which is an infinitesimal fraction of the possible protein sequences. Is there a way we can begin to create enzymes that can catalyze a reaction for which there is no good enzyme starting point? Advances in computational tools, such as machine learning and structure-based protein design, have begun to provide a glimmer of hope, but they also highlight the difficulties of this challenge. As stated above, a catalytic mechanism describes the transition states and intermediates that provide an energetically feasible path from substrates to products. To *design* an enzyme one needs to know the nature of the transition states and intermediates—this is quite difficult—and then to compose a protein sequence that will fold into a structure that stabilizes the transition states and accommodates any necessary intermediates—this is *extraordinarily* difficult. Nonetheless, computational advances have shown promise for relatively simple and well-characterized reactions, generating enzymes that could be further optimized through directed evolution. While it is clear that there is still much work to be done in this area, our ability to go from a hypothetical enzyme function to an observed one will only continue to improve with our understanding of enzymes and computational prowess. We predict that engineered enzymes will play increasingly important roles in future technology, thanks to the power of evolution.

Further reading
For more information about directed evolution and TrpB, visit the Arnold lab website.

Read more about TrpB evolution and biocatalytic ncAA synthesis:
Almhjell PJ, Boville CE, and **Arnold FH**. (2018) Engineering enzymes for noncanonical amino acid synthesis. *Chem. Soc. Rev.* **47**, 8980–8997.
Rix G, Watkins-Dulaney E, **Almhjell PJ**, Boville CE, **Arnold FH**, and Liu CC. (2020) Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities. *Nat. Commun.* **11**:5644.

About the authors
Patrick Almhjell is a graduate student in the Biochemistry and Molecular Biophysics option at the California Institute of Technology, where he works in the lab of Frances Arnold to engineer TrpB for noncanonical amino acid synthesis.
Frances Arnold is the Linus Pauling Professor of Chemical Engineering, Bioengineering, and Biochemistry at the California Institute of Technology. She was awarded the 2018 Nobel Prize in Chemistry for pioneering directed enzyme evolution.