Protein Engineering

Romas Kazlauskas

2026-07-20

Table of contents

 # Preface **1**

1 Setting Protein Engineering Goals **3**
1.1 Why proteins need to be engineered 4
1.2 Four molecular-level protein engineering goals 8
1.3 Application areas of engineered proteins 11
1.3.1 Engineering proteins for medicine 12
1.3.2 Engineering proteins for agriculture 17
1.3.3 Engineering proteins for industry 19
1.4 A look ahead: two main strategies of protein engineering 22
Glossary 23
References 24
Problems 45
Answers 47

2 Making Protein Variants **49**
2.1 Protein expression (DNA $\rightarrow$ mRNA $\rightarrow$ protein) 50
2.2 Protein overexpression 53
2.2.1 A better RNA polymerase 53
2.2.2 Multiple copies of the target protein gene 55
2.2.3 Ribosome binding site 56
2.3 Turning protein overexpression on or off 56
2.4 pET plasmids 62
2.5 Site-directed mutagenesis 64
2.6 Template-independent synthesis of DNA 68
Glossary 68

References 70
Problems 91
Answers 97

3 The logic of protein engineering 101
3.1 Gibbs energy diagrams 102
3.2 The relative Gibbs energies of states determine protein properties 106
3.3 Protein engineering manipulates the Gibbs energies of states 109
3.4 Non-covalent interactions & entropy 113
3.5 Logic of protein engineering 125
3.6 Amino acid structures 127
Glossary 132
References 133
Problems 154
Answers 156

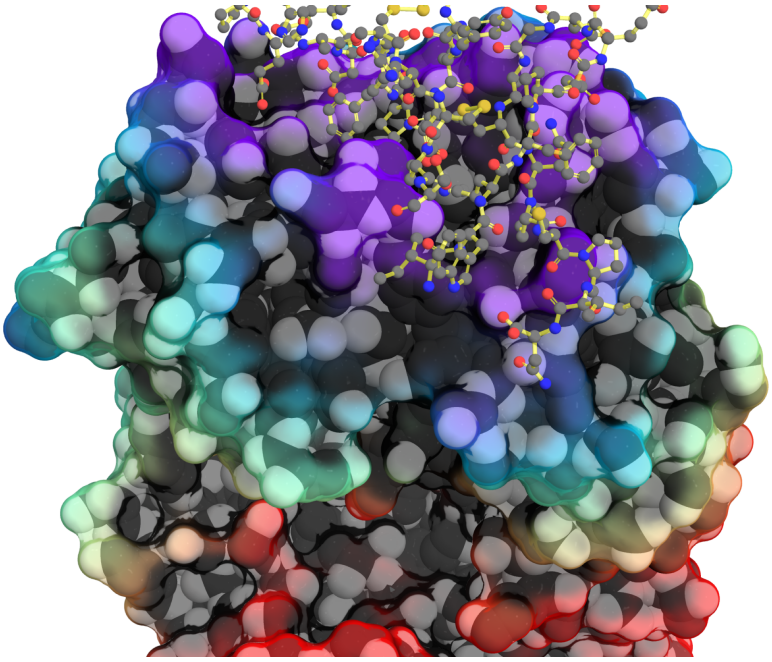
4 Protein Structure and Molecular Modeling 159
4.1 Protein structure 160
4.1.1 Determining protein structures 160
4.1.2 Structure data files 162
4.1.3 Visualizing proteins 166
4.1.4 Intuitive protein engineering 168
4.2 Computer modeling of proteins 168
4.2.1 Calculating the energy of a structure 170
4.2.2 Optimizing the geometry of a structure 183
4.2.3 Searching for conformations 189
4.2.4 Predicting protein properties 194
4.3 Conclusions 198
Glossary 198
References 200
Problems 221
Answers 226

5 Engineering More Stable Proteins **229**
5.1 Introduction . . . 230
5.2 Native and denatured conformations equilibrate . . . 233
5.3 Measuring the folding-unfolding equilibrium . . . 237
5.3.1 Denaturation with urea, ΔG_{H_2O} . . . 239
5.3.2 Heat denaturation, $\Delta G^{unfold}(T)$ . . . 243
5.4 Stabilization to cooperative unfolding . . . 247
5.4.1 Restore conserved amino acids . . . 250
5.4.2 Destabilize the unfolded ensemble . . . 255
5.4.3 Stabilize the folded form . . . 266
5.5 Stabilization to stepwise unfolding . . . 275
5.5.1 Multiple domain proteins . . . 276
5.5.2 Oligomeric proteins . . . 279
5.6 Weaknesses other than unfolding . . . 280
5.6.1 Protein aggregation . . . 280
5.6.2 Chemical modification . . . 284
5.6.3 Serum half-life . . . 286
5.7 Concluding remarks . . . 289
References . . . 290
Supporting Information . . . 311
Problems . . . 315
Answers . . . 319

6 Engineering tighter binding **321**
6.1 Introduction . . . 322
6.2 Measuring binding . . . 327
6.3 Engineering tighter binding . . . 333
6.4 Computational design of tighter binding . . . 347
6.5 Selective binding . . . 349
Glossary . . . 351
References . . . 351
Supporting Information . . . 372
Problems . . . 373

Answers 377

Preface



Welcome to protein engineering, the art and science of improving proteins. This text will teach you how to better frame your protein engineering problem and to choose approaches to reach your goal. Both online and pdf versions of this text are free to view and download here (https://betterenzyme.com).

The cover image shows part of the antibody drug trastuzumab (Her-

ceptin) in a space-filling representation bound to its target receptor HER2, shown in a ball-and-stick representation. Protein engineering transformed a mouse antibody into a humanized therapeutic while preserving its ability to recognize HER2, making trastuzumab one of the first successful antibody drugs.

1 Setting Protein Engineering Goals

 Last revised: March 2026.

Summary. Protein engineering is the modification of proteins to improve their properties, usually by replacing amino acid residues with other amino acids. Proteins evolved in nature for their biological role, but for other applications, they usually require improvements. The four molecular goals are improvements in stability, binding, reactivity and selectivity. Biotechnology uses engineered proteins in three main application areas: medicine, agriculture, and industry. Protein engineering in medicine has tuned the bioavailability of insulin and increased the potency and plasma half-life of monoclonal antibodies. Protein engineering in agriculture has increased the insect resistance of crops and stabilized enzymes for use as feed additives. Protein engineering for industrial applications include stabilizing detergent enzymes and engineering enzymes for pharmaceutical synthesis. Replacing chemical methods for pharmaceutical manufacture with biocatalysis usually yields a greener, more environmentally friendly process.

Learning goals

- While evolution in nature improves proteins for their natural function, application of proteins to non-natural functions requires protein engineering.
- The four molecular protein engineering goals are improvements in stability, binding, reactivity and selectivity. Identifying the goal is a critical first step of any protein engineering project.

- Achieving these molecular goals leads to a broad range of user benefits including lower costs, safer drugs, greener chemistry, improved diagnostics, or accessible healthcare.
- Engineered proteins have applications in medicine, agriculture and industry. Your home likely contains engineered proteins.

1.1 Why proteins need to be engineered

Over millennia humans have domesticated crops and animals by breeding to select for desired traits. Current wheat produces more grain than wild wheat; dogs are better companions than wolves. The molecular understanding of biology in recent decades created more powerful approaches to improve biology for human applications. These applications of biology to make useful products in medicine, agriculture and industry are collectively called biotechnology. The worldwide sales of biotechnology were ~US$160 billion in 2017.[1] The most significant application area is medicine, sometimes called red biotechnology, Figure 1.1. The next largest is agriculture, called green biotechnology and the smallest is industry, called white biotechnology.

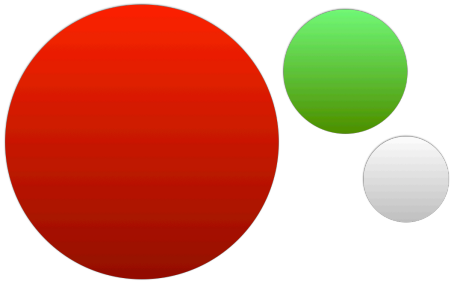


Figure 1.1. Biotechnology includes applications in medicine (red), agriculture (green) and industry (white). The areas of the circles approximate the relative worldwide sales in each area.

The two primary technologies within biotechnology are genetic engineering and protein engineering. Genetic engineering changes the genetic code of an organism. The most common genetic engineering is adding a gene from one organism to another organism so that it makes a protein that it did not make previously. Such recombinant organisms gain new abilities. For example, adding the gene encoding human insulin to *E. coli* bacteria allowed manufacture of this essential therapeutic protein. Adding genes for proteins toxic to insects into crop plants made them resistant to insects. Combining genes to degrade different hydrocarbons into one bacteria enabled it to degrade oil spills more effectively.

Protein engineering is the modification of proteins to improve them for specific applications. Altering the gene that encodes the natural protein creates a new gene that encodes an improved variant protein. Changing the protein can improve its stability, binding, catalytic activity or even give it new functions.[2] The modifications are typically substitution of one amino acid residue with another one, but sometimes include adding or removing amino acid residues, even entire domains of a protein. For example, engineering of subtilisin, a detergent protease, replaced an easily oxidized amino acid with one less easily oxidized, thereby increasing its stability to bleach during washing,[3] Figure 1.2.

Applications that differ from the natural function usually require protein engineering. Proteins have evolved in nature for specific natural functions. When proteins are used for another function or in a different way, they will likely be not good enough. Natural selection chooses proteins that improve fitness in nature, not for various applications that humans identify. Natural proteins may be accidentally suitable for these applications, but won't be optimized for them. For example, nature has not evolved proteins tolerant to organics solvents since no organisms live in organic solvents, but some proteins evolved to tolerate high temperatures often also tolerate organic solvents. In some cases, the different requirements are obvious. For example, an enzyme-catalyzed manufacture of a pharmaceutical intermediate may involve an unnatural substrate, organic cosolvents, elevated temperatures and high concentrations of sub-

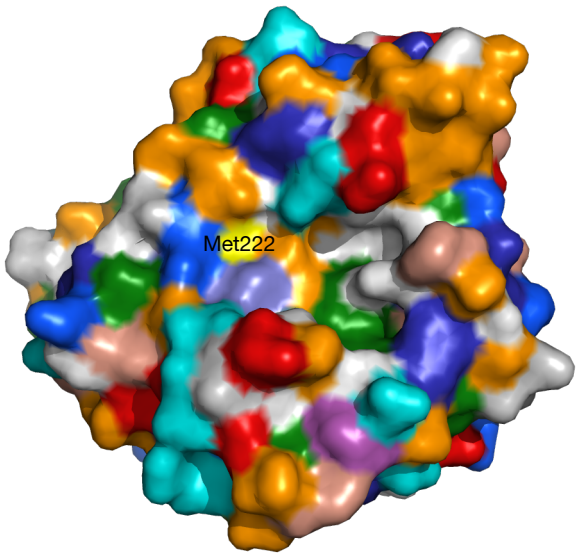


Figure 1.2. Surface representation of the protease subtilisin where the twenty amino acids are colored in one of twelve different colors according to traditional amino acid properties. Protein engineering increased the stability of this protease to bleach by replacing methionine 222 (labelled yellow patch near the center) with alanine.

strates and products. In these cases, protein engineering can adjust the active site to fit the new substrate to increase reaction rate and stabilize the protein to organic solvents, elevated temperatures and high substrate and product concentrations.

In other cases, the differences between the natural protein function and application may be subtle. For example, human insulin regulates the glucose levels in the blood. Patients with type I diabetes cannot make insulin and require injections of human insulin. Although the goal is the same - to maintain glucose levels, the dosing method differs. Patients inject insulin several times per day in contrast to the pancreas continuously secreting varying amounts of insulin. This difference in dosing required engineering fast-acting variants of human insulin for injection before meals and slow-acting variants to maintain glucose levels overnight.

Evolution does not maximize the properties of a protein, but only maintains the minimum necessary. Natural selection cannot distinguish between a protein that is improved just enough so that another protein limits function and a protein that is much better. Selection favors both equally since the much better property does not contribute to fitness. A hammer made from extra-hard steel does not drive nails any better than a hammer with steel just hard enough. Since most random mutations are detrimental, a much-better protein will eventually lose its extra abilities and return to just good enough.

An example of 'just enough' is that most proteins are just stable enough for their natural function.[4,5] Proteins from mesophiles (organisms that grow at moderate temperatures) are not stable at high temperatures. In contrast, homologous proteins from thermophiles (organisms that grow at high temperatures) are stable a high temperatures. This stability shows that proteins from mesophiles could be more stable if needed.[6] The reason for the marginal stability is that evolution cannot select for more than the minimum needed stability.[7] An extra-stable protein in a mesophile does not give it any selective advantage, so this extra stability is slowly lost through genetic drift. One can also generalize this

'just enough' quality to other protein properties. Enzymes are just fast enough for their function, they bind their substrates just tightly enough and discriminate between substrates just as much as needed.

1.2 Four molecular-level protein engineering goals

> To the person who does not know where he wants to go there is no favorable wind. — Seneca

The application improvements enabled by engineered proteins vary widely because they depend on the function of the protein being engineered. As described below, some improvements make vaccines more effective, others improve the synthesis of pharmaceutical, still others enable plants to grow faster under heat stress. Despite this wide range of applications, the improvements themselves stem from only four types of protein property changes: stability, binding, reactivity or selectivity, Table 1.1. Later tables in this chapter list a wide range of applications, but also classify each application according to one of these four protein property changes.

Table 1.1. Four molecular-level goals in protein engineering.

Goal	Examples of changes
stability	tolerate heat, organic solvents, extremes of pH or other harsh conditions; tolerate storage or use for a long time at normal conditions; tolerate destabilizing substitutions; longer plasma lifetime; resistant to digestive proteases
binding	bind target ligand more tightly or less tightly; avoid binding to antibodies to avoid allergic reaction

Goal	Examples of changes
catalytic activity	faster catalysis of existing substrates; expand catalysis to new substrates; inactivate undesired catalytic activity; enable catalysis of a new chemical reaction
selectivity	favor binding one ligand over another in the same solution; favor reaction of one of several competing substrates; favor formation of one of several possible products

Stability refers to the ability of a protein to maintain its function, usually binding to a target or catalyzing a reaction. Increasing protein stability allows it to tolerate harsh conditions such as high temperatures or the presence of bleach as in the detergent protease example above. It may make the protein last longer under normal application or storage conditions. For biopharmaceuticals, stability may also refer to extending the plasma lifetime of the protein. Most proteins are rapidly removed from the blood thus limiting their therapeutic benefit. Increasing the plasma lifetime enhances the therapeutic benefit.

Binding refers to the affinity of the protein for a target molecule. Increases in the affinity of a protein for its target can lower the effective dose of a biopharmaceutical or lower the detection limit of a diagnostic test. For biopharmaceuticals, *decreases* in binding can also be important. Biopharmaceuticals should not bind to human antibodies to avoid causing an allergic reaction. Humanization of proteins engineers them so that they do not bind to human antibodies.

Reactivity refers to the ability of an enzyme to catalyze chemical reactions. Increases in reaction rates lowers the amount of enzyme needed for the chemical transformation. Increases in reactivity may improve reactions for existing substrates or may create new catalytic abilities by expanding reactivity to new substrates. Improvements in reactivity may

enable an enzyme to catalyze new reaction types of chemical reaction, even those that enzymes in nature do not catalyze.

Selectivity refers to the relative binding or relative reactivity of two or more competing molecules in the solution. Selectivity may refer to the ability of a protein to bind one of several ligands or to catalyze the reaction of one of several substrates. It can also refer to the selective formation of one of several possible products from a single substrate. A common type of selectivity exhibited by enzymes is enantioselectivity where one enantiomer reacts faster than the other.

Identifying the protein engineering goal is a critical first step in any protein engineering project because it defines the approaches that one can use. Typically the target protein already has some desirable properties, but needs improvement on one of them. One seeks to improve this one property (stability, binding, reactivity or selectivity) while not degrading the already desirable properties. To fully define that goal one should also state the amount of change needed (e.g., a two-fold increase in stability) because it reveals how difficult the engineering will be.

Sometimes protein engineering changes several protein properties so that it can be hard to identify which property is the true protein engineering goal. The true goal is the one property that *must* change to improve the protein for the application. The other changes are side effects of engineering to achieve this goal.

For example, researchers wanted to engineer a protease that could degrade the protein gluten in the human digestive track so they could add it to the diet of gluten-intolerant patients.[8] The researchers started with the protease KumaWT, which favored hydrolysis after a ProArg or ProLys sequence and wanted to change it to favor hydrolysis after a ProGln sequence since gluten contains many ProGln amino acid pairs in its sequence. The engineered protein showed changes in three properties: KumaWT was more reactive toward ProGln-containing peptides, more selective for ProGln versus ProArg sequences and bound the ProGln sequence more tightly to the protease binding site.

Which of these properties was the true goal? Imagine if the catalytic activity toward the ProGln sequence increased, but there was no change in the selectivity (the ProLys sequence also reacted) and no change in the binding (the increase in reactivity was not associated with better binding). The engineered protein would still be improved because it would degrade the ProGln containing peptides. Increasing reactivity is the true goal. In contrast, imagine if selectivity improved without changing the catalytic activity. (A decrease in the activity toward the ProLys sequence could increase the selectivity for the ProGln sequence without increasing the activity toward the ProGln sequence.) No more ProGln would be cleaved than with the starting enzyme. Thus, selectivity is not the true goal and the increase is selectivity is a side effect of increased reactivity. Finally, consider the case where the binding improved, but not the selectivity or reactivity. Again, without an increase in reactivity, no more ProGln sequences would be cleaved. The authors hypothesized that the reason for the poor reactivity was that the ProGln sequence binds poorly. Thus, increasing binding was an approach to increase reactivity. If they achieved increased binding without a concomitant increase in reactivity, they would have failed and tried another approach to increase reactivity. Increased binding alone would not have yielded a better enzyme. Thus, protein engineering can simultaneously change several protein properties. The true protein engineering goal is the property that *must* change in order to see an improvement in the proposed application.

1.3 Application areas of engineered proteins

Examples from different application areas show how changes in protein stability, binding, reactivity and selectivity can improve proteins in widely different applications.

1.3.1 Engineering proteins for medicine

Biopharmaceuticals are biomolecules used as drugs. They can be proteins such as antibodies or enzymes; they can also be nucleic acids such as micro RNAs and even complete microbes such as attenuated live vaccines. This text considers only protein biopharmaceuticals.

The first protein biopharmaceuticals were recombinant equivalents of the natural protein where the biotechnology advantage was the ability to produce large amounts of protein. For example, blood factors that were isolated from donated blood could now be produced in cell cultures due to genetic engineering. The subsequent generations of blood factors were improved engineered variants. For example, engineering increased the potency of the blood coagulation protein, Factor VII used to treat hemophilia, so that each dose required less protein,[9] Table 1.2. The action of this blood coagulation protein depends on its binding to a membrane surface containing acidic phospholipids, which is characteristic of damaged vascular cells or of platelets that adhere to the damage as the clot forms. Mutagenesis of the membrane-binding domain of Factor VII yielded variants that bound more tightly to the membrane and therefore were more potent. The best variant contained five substitutions and was 149-296 fold better than wild-type Factor VII. Other researchers increased its potency with three amino acid substitutions that increased its catalytic activity 30-fold.[10] Proteolytic activity of Factor VII activates the next factor in the coagulation cascade.

Table 1.2. Examples of protein engineering to improve biopharmaceuticals.

Category	Example of improvement	Property change
blood factors	more potent coagulation factor VII	increase binding to damaged cells; increased reactivity (proteolysis)
thrombolytics & anticoagulants	more potent and longer plasma half-life of tissue plasminogen activator	increase binding to fibrin; reduce binding to receptor that removes proteins from blood
hormones	fast-acting or long-lasting insulins	alter binding with other monomers
monoclonal antibodies	reduced immunogenicity of mouse-derived antibody	reduce binding to immune response proteins
growth factors	longer serum half-life of granulocyte colony-stimulating factor	increase binding to receptors that recycle proteins to blood stream
vaccines	less toxic pertussis (whooping cough) vaccine	inactivate reactivity that causes toxicity
other	less immunogenic asparaginase	reduce binding to immune response proteins

Insulin variants engineered for altered bioavailability are even better than human insulin at controlling blood sugar levels. The engineering relied on the fact that monomeric insulin is the active form, while multimeric forms are inactive. Insulin lispro is a fast-acting insulin analog for injections before meals because it favors the monomeric

form.[11] The amino acid sequence of lispro differs only in the reversal of the penultimate lysine and proline residues on the C-terminal end of the B-chain. (Insulin consists of 51 amino acids arranged as a dimer of an A-chain and a B-chain linked by disulfide bonds.) This residue change blocks the formation of insulin dimers and hexamers and increases the amount of monomeric insulin (the active form), thereby creating a faster-acting variant. In contrast, substitutions to create a slow-acting insulin analog, glargine, promote the association of the monomers into an insoluble precipitate, which then slowly releases monomeric insulin to maintain basal levels of insulin for an extended time.

An essential class of therapeutic proteins is monoclonal antibodies and monoclonal antibody drug conjugates. Common goals of antibody engineering are increasing binding to the target and minimizing immunogenicity.[12] One example of a therapeutic monoclonal antibody is trastuzumab (Herceptin®) used to treat breast cancer. This antibody interferes with the HER2 receptor (human ErbB2 receptor tyrosine kinase), which signals cell proliferation. Cancer cells, usually breast cancer, overexpress the HER2 receptor causing the cancer cells to grow uncontrollably. Interfering with the receptor stops uncontrolled growth.

Trastuzumab is a humanized version of a mouse antibody, Figure 1.3. Humanization prevents an allergic reaction to a mouse antibody. This humanization involved first, the transplantation of the complementarity-determining regions from the mouse antibody into a human IgG antibody and second, an optimization of the surrounding regions, called framework regions.[13] The transplantation of the complementarity-determining regions changed 24 amino acids in the light chain and 32 in the heavy chain. This substitution did not yield a useful antibody. The engineered version bound less tightly than the mouse antibody and did not block cancer cell proliferation upon binding. Next, an additional seven substitutions in the surrounding (framework) regions increased the binding to the target 100-fold making it three-fold tighter than the mouse antibody. This tighter binding restored the ability of the antibody

to block cell proliferation. Trastuzumab was approved in the USA in 1998 to treat metastatic breast cancer.

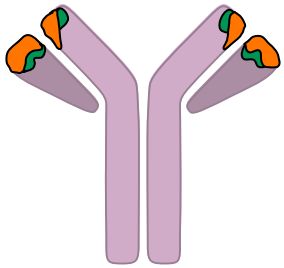


Figure 1.3. To engineer trastuzumab, researchers combined the complementarity-determining regions from a mouse antibody (orange; 56 amino acids) with a human IgG antibody (lavender). Full activity required seven additional substitutions (green) outside the binding region.

Growth factors and cytokines, such as vascular endothelial growth factor, epidermal growth factor and granulocyte colony-stimulating factor, stimulate cell recruitment, proliferation, morphogenesis, and differentiation. Inhibitors of these growth factors act as anticancer drugs, while enhanced growth factors promote wound repair. Natural growth factors lack stability and specificity; they are prone to degradation in serum and have multiple activities. In one example of protein engineering of a cytokine, Sarkar and coworkers increased the half-life and potency of granulocyte colony-stimulating factor by modifications that increased its recycling back to the bloodstream.[14]

Cantor and coworkers engineered asparaginase, an enzyme for leukemia treatment, to be less immunogenic.[15] Treatment of childhood acute lymphoblastic leukemia involves injections of asparaginase, which cleaves asparagine to aspartate, thereby depriving the cancerous cells of asparagine. Healthy cells can make asparagine, but these cancer cells cannot. Humans do not have an asparaginase enzyme, so the treatment uses enzyme from the bacteria *Escherichia coli*. One limitation is that

some patients' immune systems recognize this enzyme as a non-human protein and inhibit it by binding it with an antibody or even inducing a severe allergic reaction. Cantor and coworkers identified amino acid sequences on the surface of *E. coli* asparaginase that are likely to be recognized by human T-cells and replaced them with non-recognized amino acid sequences. The replacements maintained the catalytic activity of the asparaginase.

Vaccines against pertussis (whooping cough) contain engineered proteins. The bacteria that causes pertussis, Bordetella pertussis, secrete pertussis toxin protein, which consists of five subunits. Immunization with the pertussis toxin or just its S1 subunit protects against disease. However, the S1 subunit is an enzyme that catalyzes the ADP-ribosylation of GTP-binding proteins. This catalytic activity is the origin of the toxicity of the pertussis toxin. Two amino acid substitutions in the S1 subunit - Glu129Gly, Arg9Lys - deactivate its catalytic activity. This inactivated version of the S1 subunit is the vaccine against pertussis.[16]

The mRNA vaccines against the SARS-CoV-2 virus encode the spike protein, which sits on the virus surface and binds to cell-surface proteins during infection. Upon binding to the target cell, the spike protein changes conformation. Researchers reasoned that the spike protein in the pre-binding conformation would make a better vaccine than the spike protein in the post-binding conformation. Binding antibodies to the spike protein before it binds to the cell could prevent infection. To stabilize the pre-binding conformation, researchers replaced six residues in the spike protein with proline.[17] Proline limits the flexibility at those sites due to its ring structure and keeps the spike protein in the pre-binding conformation. This stabilization of the protein in the pre-binding conformation improved the vaccine.

1.3.2 Engineering proteins for agriculture

Agricultural biotechnology aims mainly to improve productivity of crops and animals. Some products, like bovine somatotrophin (a growth hormone), are copies of the natural protein produced in recombinant *E. coli* bacteria. The pituitary gland produces small amounts of somatotrophin, but production in bacteria increases the availability and lowers the cost. Treating cows with somatotrophin increases milk production. In other cases, natural proteins are improved by engineering, Table 1.3.

Table 1.3. Examples of protein engineering to improve agriculture.

Category	Example of improvement	Property change
herbicide-resistant crops	glyphosate-resistant soybeans	add new reactivity (glyphosate oxidase)
insect-resistant crops	cotton resistant to bollworm	increase binding of *Bt* toxins to target proteins in insect gut
enhanced photosynthesis	faster growth of *Arabidopsis* at higher temperatures	increase stability of rubisco activase
modified products	algae producing an increased fraction of medium-chain fatty acids	increased reactivity toward medium-chain fatty acid precursors
feed additives	animal feed with higher phosphorus availability	increase stability of phytase

Glyphosate is a broad-spectrum, systemic herbicide that acts by inhibiting the enzyme EPSP synthase in aromatic amino acid biosynthesis.

Farmers can control weeds by spraying glyphosate on fields, but only if the crops are insensitive to glyphosate. Adding a glyphosate-insensitive variant of EPSP synthase from *Agrobacterium* sp. created the first glyphosate-resistant soybean plants.[18] The next generation of glyphosate-resistant crops contain enzymes such as oxidases or acetyl transferases, which inactivate glyphosate. Protein engineering to increase the catalytic activity of these enzymes is an important goal.[19]

Genetic modification of crops such as cotton, corn, and rice to express insecticidal crystal proteins produced by *Bacillus thuringiensis* protects the crops against insect pests. The toxicity of these proteins varies toward different insects, and some insects develop resistance to these toxins. Enhancing the toxicity of these proteins by increasing the their binding to target proteins in the insect gut is an essential goal in agricultural biotechnology.[20]

Heat stress hinders plant growth and lowers crop yields mainly due to reduced photosynthesis. This reduction is mainly due to an inactivation of a thermolabile ATPase, Rubisco activase, whose role is to maintain ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) in its active state. Kurek and coworkers engineered a thermostable rubisco activase, which improved the growth of *Arabidopsis* plants at higher temperatures.[21] The leaf area increased approximately 20% during moderate heat stress as compared to wild-type plants.

Engineering the selectivity of fatty acid synthesis enzymes changes the composition of the lipids that accummulate in plants. Whittle and Shanklin altered the chain length selectivity of a plant fatty acid desaturase to accept a 16-carbon instead of an 18-carbon substrate.[22] Lin and Lee engineered algae to produce more medium chain length fatty acids relative to long-chain-length fatty acids for use in biodiesel production.[23]

Engineering a more stable phytase, a feed additive enzyme, increased the nutritional value of feed and reduced pollution. Phosphorus in grains and oilseeds often occurs as phytic acid (*myo*-inositol hexakisphosphate).

Non-ruminants such as pigs and chicken digest phytic acid incompletely leading to high phosphorus in wastewater. The enzyme phytase in animal feed aids digestion of phytic acid by catalyzing its hydrolysis. Forming animal feed pellets requires pressing at high temperatures, so the phytases have been engineered for higher thermal stability to withstand this step.[24]

1.3.3 Engineering proteins for industry

Industry uses enzymes for manufacturing and similar non-natural applications, Table 1.4. The three main application areas are detergents, food and beverage, and other, which includes enzymes for the synthesis of biofuels, fine chemicals and pharmaceuticals. The advantages of using enzymes over chemical reagents or catalysts are that they are faster, greener and more selective. The primary engineering goals are stability and faster catalysis, both of which lower the cost of the enzyme.

Table 1.4. Examples of protein engineering to improve industrial enzymes.

Category	Example of improvement	Property change
detergents	bleach-tolerant protease	stabilize subtilisin to oxidation
food and beverage	simplify starch hydrolysis to maltodextrins and glucose	stabilize α-amylase to low pH
other (e.g., biofuels, pharmaceuticals)	engineer transaminase for manufacture of diabetes drug	increase reactivity toward target substrate; increase stability to reaction conditions

Detergent enzymes such as proteases speed removal of food, blood and other stains on clothing. Subtilisin was the first industrial enzyme to be engineered.[3] Subtilisin tolerated hot water and surfactants, but bleach rapidly inactivated it, which limited its usefulness. Replacement of an oxidation-sensitive methionine 222 in the active site (Figure 1.2) by alanine stabilized subtilisin while maintaining high activity. The cost of these industrial proteins are typically $100/kg. In contrast, proteins for medical applications can cost 10^5-fold more, $10,000/g, because their manufacture and regulatory approval are more complex.

The largest volume application in food and beverages is the conversion of cornstarch to glucose catalyzed by α-amylase and glucoamylase. Glucose isomerase can isomerize the resulting glucose to high fructose corn syrup, which tastes sweeter than glucose. This process uses high temperatures to prevent microbial growth and high concentrations to minimize reactor size. Both α-amylase[25] and glucose isomerase[26] have been engineered to tolerate high temperatures and high glucose concentration. Glucose isomerase has a half-life = 50-100 d at the operating temperature of 55 °C. More than 500 tons of glucose isomerase produce millions of tons of high fructose corn syrup each year, corresponding to productivities of >10,000 kg corn syrup per kg enzyme.

The use of enzymes for chemical synthesis is known as biocatalysis.[27] In pharmaceutical manufacture, the most significant application is the penicillin G amidase-catalyzed manufacture of penicillin and cephalosporin antibiotics. This process yields more than 10,000 tons of antibiotics annually with a productivity of >600 kgs product/kg enzyme.[28] These antibiotics are natural products, so they are the natural substrates for this enzymes. However, many other pharmaceuticals are not natural products, so the ability of enzymes to act on them is accidental. Engineering of enzymes to improve their ability to act on unnatural substrates and tolerate the harsh conditions of a chemical reactor is a common goal.

Engineered enzymes can also create new biochemical pathways within cells, and the whole cells can be used for synthesis. For example,

Ran and Frost expanded the substrate range of an aldolase to create a new metabolic pathway to make shikimic acid for an influenza drug synthesis.[29] Keasling and coworkers combined enzymes from different organisms and biochemical pathways to create a new biochemical pathway for the synthesis of artemisinin, an anti-malarial.[30]

Green chemistry

One reason to use enzymes for manufacture of pharmaceuticals is to decrease the environmental impact of the process. The risk associated with a chemical depends both on how dangerous it is (hazard) and on one's contact with it (exposure), Equation 1.1.

$$risk = hazard \cdot exposure \tag{1.1}$$

In the past, governments and industry focused on reducing risk by minimizing exposure. Lab coats, safety glasses, and other chemical handling rules limit the exposure of workers to hazardous chemicals. Green chemistry instead focuses on the hazard and tries to minimize it instead of exposure.[31]. Preventing problems is inevitably easier and less expensive than contending with difficulties after they occur. Green chemistry is the design of chemical products and processes that reduce or eliminate hazardous substances. Replacing chemical manufacturing steps with enzyme-catalyzed steps is an essential tool in green chemistry. Enzyme-catalyzed steps often replace hazardous solvents with water and hazardous catalysts with biodegradable enzymes. One example of greener pharmaceutical manufacture is the improved synthesis of sitagliptin, the active ingredient in an oral type 2 diabetes drug. The difficult step is the addition of the amino group with the correct orientation, Figure 1.4.

Researchers at Merck & Co. had already dramatically improved the original synthesis of sitagliptin by using a catalytic asymmetric hydrogenation

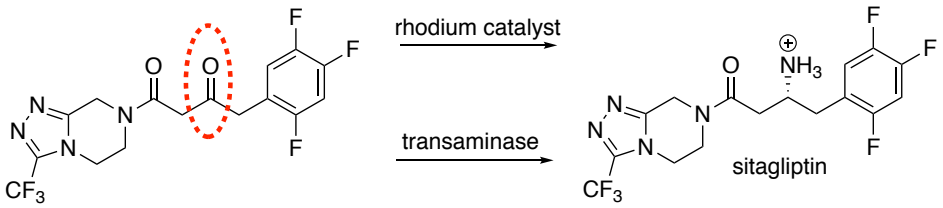


Figure 1.4. Two improved routes to sitagliptin showing the difficult step of inserting the amino group with the correct configuration. Improvements in the chemocatalytic process (rhodium catalyst) won a Green Chemistry award in 2006. Further improvements by switching to a biocatalytic process (transaminase) won a second Green Chemistry award in 2010.

with a rhodium catalyst. This improved synthesis won a Green Chemistry award in 2006. Biocatalysis further improved this synthesis. Researchers at Codexis and Merck & Co. used a transaminase-catalyzed formation of an amine, which eliminated four steps including one that required a high pressure reactor.[32] The biocatalysis approach resulted in a 10-13% higher overall yield and 19% less total waste. This improved synthesis won a second Green Chemistry award in 2010. Enabling this synthesis required extensive protein engineering of the transaminase. The starting transaminase did not react at all with the required substrate and was unstable under the reaction conditions. The engineering replaced 27 amino acids creating a highly active, highly selective and stable enzyme.

1.4 A look ahead: two main strategies of protein engineering

Protein engineering typically begins with proteins that already possess some, but not all, of the properties required for a particular application.

These starting points are often natural proteins, but they may also be proteins proposed through computational design, although computational methods do not yet reliably generate proteins with all the properties needed for most applications.

The next chapter introduces the experimental methods used in protein engineering. Genetic modification of microorganisms enables the production of desired proteins in large quantities. Molecular biology techniques can then be used to alter the DNA encoding these proteins and create variants with improved properties.

Chapter 3 introduces the logic of protein engineering: protein states, defined as particular structural and functional conditions of proteins, determine protein properties, and changes in the relative Gibbs energies of these states alter those properties. Chapter 4 introduces protein structure and the molecular modeling of protein states.

The remaining chapters focus on the two principal strategies of protein engineering: rational design and directed evolution. Rational design seeks beneficial changes on the basis of protein structure, molecular mechanism, and computational modeling. Chapters 5–8 examine the rational design of protein stability, binding, reactivity, and selectivity, explaining both how these properties are measured and how they can be improved. Chapters 9–11 cover directed evolution, in which protein variants are generated, screened for improved properties, and subjected to repeated cycles of optimization.

Glossary

Binding is the ability of a protein to associate with a target molecule.

Protein engineering goal is the protein property that *must* change in order to reach the desired application improvement. Most protein

engineering goals are one of these four: stability, binding, reactivity or selectivity. The specific application improvement depends on the function of the protein being improved.

Reactivity is the ability of an enzyme to catalyze a chemical transformation.

Selectivity is the ability of a protein to discriminate between competing molecules in solution. This discrimination can involve selectivity in binding the competing molecules or selectivity in reacting with the competing molecules. Selectivity can also refer to the favored formation of one of several competing products from a single substrate.

Stability is the ability of a protein to remain folded and functional. The function can be binding to a target or catalyzing a reaction. Stability can also refer to the ability of a biopharmaceutical to persist in plasma.

References

1. Morrison, C., & Lähteenmäki, R. (2018). Public biotech in 2017-the numbers. *Nat. Biotechnol.*, *36*(7), 576–584. https://doi.org/10.1038/nbt.4175
2. Brannigan, J. A., & Wilkinson, A. J. (2002). Protein engineering 20 years on. *Nat. Rev. Mol. Cell Biol.*, *3*(12), 964–970. https://doi.org/10.1038/nrm975
3. Estell, D. A., Graycar, T. P., & Wells, J. A. (1985). Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.*, *260*(11), 6518–6521. https://doi.org/10.1016/S0021-9258(18)88811-1
4. Pace, C. N., & Hermans, J. (1975). The stability of globular proteins. *CRC Crit. Rev. Biochem.*, *3*(1), 1–43. https://doi.org/10.3109/10409237509102551

5. Taverna, D. M., & Goldstein, R. A. (2002). Why are proteins marginally stable? *Proteins*, *46*(1), 105–109. https://doi.org/10.1002/prot.10016
6. Fersht, A. R., & Serrano, L. (1993). Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.*, *3*(1), 75–83. https://doi.org/10.1016/0959-440X(93)90205-Y
7. Bloom, J. D., Lu, Z., Chen, D., Raval, A., Venturelli, O. S., & Arnold, F. H. (2007). Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.*, *5*(1), 97–21. https://doi.org/10.1186/1741-7007-5-29
8. Gordon, S. R., Stanley, E. J., Wolf, S., Toland, A., Wu, S. J., Hadidi, D., Mills, J. H., Baker, D., Pultz, I. S., & Siegel, J. B. (2012). Computational design of an α-gliadin peptidase. *J. Am. Chem. Soc.*, *134*(50), 20513–20520. https://doi.org/10.1021/ja3094795
9. Harvey, S. B., Stone, M. D., Martinez, M. B., & Nelsestuen, G. L. (2003). Mutagenesis of the γ-carboxyglutamic acid domain of human factor VII to generate maximum enhancement of the membrane contact site. *J. Biol. Chem.*, *278*(10), 8363–8369. https://doi.org/10.1074/jbc.M211629200
10. Persson, E., Olsen, O. H., Bjørn, S. E., & Ezban, M. (2012). Vatreptacog alfa from conception to clinical proof of concept. *Semin. Thromb. Hemost.*, *38*(3), 274–281. https://doi.org/10.1055/s-0032-1302442
11. Bolli, G. B., Di Marchi, R. D., Park, G. D., Pramming, S., & Koivisto, V. A. (1999). Insulin analogues and their potential in the management of diabetes mellitus. *Diabetologia*, *42*(10), 1151–1167. https://doi.org/10.1007/s001250051286
12. Ryu, J. K., Kim, H. S., & Nam, D. H. (2012). Current status and perspectives of biopharmaceutical drugs. *Biotechnol. Bioprocess Eng.*, *17*(5), 900–911. https://doi.org/10.1007/s12257-012-0095-1

13. Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B. B., Henner, D., Wong, W. L. T., Rowland, A. M., Kotts, C., Carver, M. E., & Shepard, H. M. (1992). Humanization of an anti-p185^{HER2} antibody for human cancer therapy. *Proc. Natl. Acad. Sci. U. S. A.*, *89*(10), 4285–4289. https://doi.org/10.1073/pnas.89.10.4285
14. Sarkar, C. A., Lowenhaupt, K., Horan, T., Boone, T. C., Tidor, B., & Lauffenburger, D. A. (2002). Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nat. Biotechnol.*, *20*(9), 908–913. https://doi.org/10.1038/nbt725
15. Cantor, J. R., Yoo, T. H., Dixit, A., Iverson, B. L., Forsthuber, T. G., & Georgiou, G. (2011). Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(4), 1272–1277. https://doi.org/10.1073/pnas.1014739108
16. Loosmore, S. M., Zealey, G. R., Boux, H. A., Cockle, S. A., Radika, K., Fahim, R. E., Zobrist, G. J., Yacoob, R. K., Chong, P. C., & Yao, F. L. (1990). Engineering of genetically detoxified pertussis toxin analogs for development of a recombinant whooping cough vaccine. *Infect. Immun.*, *58*(11), 3653–3662. https://doi.org/10.1128/iai.58.11.3653-3662.1990
17. Hsieh, C.-L., Goldsmith, J. A., Schaub, J. M., DiVenere, A. M., Kuo, H.-C., Javanmardi, K., Le, K. C., Wrapp, D., Lee, A. G., Liu, Y., Chou, C.-W., Byrne, P. O., Hjorth, C. K., Johnson, N. V., Ludes-Meyers, J., Nguyen, A. W., Park, J., Wang, N., Amengor, D., … McLellan, J. S. (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science*, *369*(6510), 1501–1505. https://doi.org/10.1126/science.abd0826

18. Padgette, S. R., Kolacz, K. H., Delannay, X., Re, D. B., Tinius, C. N., Rhodes, W. K., Otero, Y. I., Barry, G. F., Eichholtz, D. A., Peschke, V. M., Nida, D. L., Taylor, N. B., & Kishore, G. M. (1995). Development, identification, and characterization of a glyphosate-tolerant soybean line. *Crop Sci.*, *35*(5), 1451–1461. https://doi.org/10.2135/cropsci1995.0011183x003500050032x
19. Castle, L. A., Siehl, D. L., Gorton, R., Patten, P. A., Chen, Y. H., Bertain, S., Cho, H.-J., Duck, N., Wong, J., Liu, D., & Lassner, M. W. (2004). Discovery and directed evolution of a glyphosate tolerance gene. *Science*, *304*(5674), 1151–1154. https://doi.org/10.1126/science.1096770
20. Bravo, A., & Soberon, M. (2008). How to cope with insect resistance to Bt toxins? *Trends Biotechnol.*, *26*(10), 573–579. https://doi.org/10.1016/j.tibtech.2008.06.005
21. Kurek, I., Chang, T. K., Bertain, S. M., Madrigal, A., Liu, L., Lassner, M. W., & Zhu, G. (2007). Enhanced thermostability of *Arabidopsis* rubisco activase improves photosynthesis and growth rates under moderate heat stress. *Plant Cell*, *19*(10), 3230–3241. https://doi.org/10.1105/tpc.107.054171
22. Whittle, E., & Shanklin, J. (2001). Engineering Δ^9-16:0-acyl carrier protein (ACP) desaturase specificity based on combinatorial saturation mutagenesis and logical redesign of the castor Δ^9-18:0-ACP desaturase. *J. Biol. Chem.*, *276*(24), 21500–21505. https://doi.org/10.1074/jbc.M102129200
23. Lin, H., & Lee, Y. K. (2017). Genetic engineering of medium-chain-length fatty acid synthesis in *Dunaliella tertiolecta* for improved biodiesel production. *J. Appl. Phycol.*, *29*(6), 2811–2819. https://doi.org/10.1007/s10811-017-1210-7
24. Yao, M. Z., Zhang, Y. H., Lu, W. L., Hu, M. Q., Wang, W., & Liang, A. H. (2012). Phytases: Crystal structures, protein engineering and potential biotechnological applications. *J. Appl. Microbiol.*, *112*(1), 1–14. https://doi.org/10.1111/j.1365-2672.2011.05181.x

25. Richardson, T. H., Tan, X., Frey, G., Callen, W., Cabell, M., Lam, D., Macomber, J., Short, J. M., Robertson, D. E., & Miller, C. (2002). A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable α-amylase. *J. Biol. Chem.*, *277*(29), 26501–26507. https://doi.org/10.1074/jbc.M203183200
26. Quax, W. J., Mrabet, N. T., Luiten, R. G. M., Schuurhuizen, P. W., Stanssens, P., & Lasters, I. (1991). Enhancing the thermostability of glucose-isomerase by protein engineering. *Bio/Technology*, *9*(8), 738–742. https://doi.org/10.1038/nbt0891-738
27. Bell, E. L., Finnigan, W., France, S. P., Green, A. P., Hayes, M. A., Hepworth, L. J., Lovelock, S. L., Niikura, H., Osuna, S., Romero, E., Ryan, K. S., Turner, N. J., & Flitsch, S. L. (2021). Biocatalysis. *Nat. Rev. Methods Primers*, *1*(1), 46. https://doi.org/10.1038/s43586-021-00044-z
28. van der Wielen, L. A. M., Ottens, M., & Straathof, A. J. J. (2001). Process Technology and Process Integration in the Preparation of Penicillins. In A. Bruggink (Ed.), *Synthesis of β-Lactam antibiotics: Chemistry, biocatalysis & process integration* (pp. 150–205). Springer Netherlands.
29. Ran, N., & Frost, J. W. (2007). Directed evolution of 2-keto-3-deoxy-6-phosphogalactonate aldolase to replace 3-deoxy-D-arabino-heptulosonic acid 7-phosphate synthase. *J. Am. Chem. Soc.*, *129*(19), 6130–6139. https://doi.org/10.1021/ja067330p
30. Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., ... Newman, J. D. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, *496*(7446), 528–532. https://doi.org/10.1038/nature12051
31. Anastas, P. T., & Warner, J. C. (2000). *Green chemistry : Theory and practice*. Oxford University Press.

32. Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J., Brands, J., Devine, P. N., Huisman, G. W., & Hughes, G. J. (2010). Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, *329*(5989), 305–309. https://doi.org/10.1126/science.1188934
33. Stone, E. M., Paley, O., Hu, J., Ekerdt, B., Cheung, N.-K., & Georgiou, G. (2012). De novo engineering of a human cystathionine-γ-lyase for systemic L-methionine depletion cancer therapy. *ACS Chem. Biol.*, *7*(11), 1822–1829. https://doi.org/10.1021/cb300335j
34. Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, *27*(10), 946–950. https://doi.org/10.1038/nbt.1568
35. Jobe, A., & Bourgeois, S. (1972). *lac* Repressor-operator interaction. VI. The natural inducer of the lac operon. *J. Mol. Biol.*, *69*(3), 397–408. https://doi.org/10.1016/0022-2836(72)90253-7
36. Egel, R. (1988). The '*lac*' operon - an irrelevant paradox? *Trends Genet.*, *4*(2), 31–31. https://doi.org/10.1016/0168-9525(88)90062-5
37. Valdez-Cruz, N. A., Caspeta, L., Pérez, N. O., Ramírez, O. T., & Trujillo-Roldán, M. A. (2010). Production of recombinant proteins in *E. Coli* by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microb Cell Fact*, *9*(1), 18. https://doi.org/10.1186/1475-2859-9-18
38. Austin, H. P., Allen, M. D., Donohoe, B. S., Rorrer, N. A., Kearns, F. L., Silveira, R. L., Pollard, B. C., Dominick, G., Duman, R., El Omari, K., Mykhaylyk, V., Wagner, A., Michener, W. E., Amore, A., Skaf, M. S., Crowley, M. F., Thorne, A. W., Johnson, C. W., Woodcock, H. L., ... Beckham, G. T. (2018). Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(19), E4350–E4357. https://doi.org/10.1073/pnas.1718804115

39. Hutchison 3rd, C. A., Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., & Smith, M. (1978). Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem.*, *253*(18), 6551–6560. https://doi.org/10.1016/S0021-9258(19)46967-6
40. Silva, D., Santos, G., Barroca, M., & Collins, T. (2017). Inverse PCR for point mutation introduction. In L. Domingues (Ed.), *PCR: Methods and Protocols* (pp. 87–100). Springer New York. https://doi.org/10.1007/978-1-4939-7060-5_5
41. Roy, S., & Caruthers, M. (2013). Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. *Molecules*, *18*(11), 14268–14284. https://doi.org/10.3390/molecules181114268
42. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S., & Ryadnov, M. G. (2023). DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.*, *7*(3), 144–161. https://doi.org/10.1038/s41570-022-00456-9
43. Seydel, C. (2023). DNA writing technologies moving toward synthetic genomes. *Nat Biotechnol.*, *41*(11), 1504–1509. https://doi.org/10.1038/s41587-023-02006-0
44. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison 3rd, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318
45. Green, M. R., & Sambrook, J. (2012). *Molecular Cloning: A Laboratory Manual* (4th ed.). Cold Spring Harbor Press.
46. Ball, P. (2018). A compassionate substance. How science and myth collide in water. *Lapham's Quart.*, *11*(Summer).
47. Pace, C. N., Scholtz, J. M., & Grimsley, G. R. (2014). Forces stabilizing proteins. *FEBS Lett.*, *588*(14), 2177–2184. https://doi.org/10.1016/j.febslet.2014.05.006

48. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427
49. 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., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
50. Mura, C., McCrimmon, C. M., Vertrees, J., & Sawaya, M. R. (2010). An introduction to biomolecular graphics. *PLoS Comput. Biol.*, *6*(8), e1000918. https://doi.org/10.1371/journal.pcbi.1000918
51. Leach, A. R. (2001). *Molecular modelling: Principles and applications*. Prentice Hall.
52. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz Jr, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, *117*(19), 5179–5197. https://doi.org/10.1021/ja00124a002
53. Allinger, N. L., Zhou, X., & Bergsma, J. (1994). Molecular mechanics parameters. *J. Mol. Struct. THEOCHEM*, *312*(1), 69–83. https://doi.org/10.1016/S0166-1280(09)80008-0

54. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem\@. Theory Comput.*, *13*(6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125
55. Homeyer, N., & Gohlke, H. (2012). Free energy calculations by the molecular mechanics Poisson-Boltzmann surface area method. *Mol. Inf.*, *31*(2), 114–122. https://doi.org/10.1002/minf.201100135
56. Bae, E., & Phillips Jr, G. N. (2005). Identifying and engineering ion pairs in adenylate kinases. *J. Biol. Chem.*, *280*(35), 30943–30948. https://doi.org/10.1074/jbc.M504216200
57. Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J., & Osuna, S. (2019). Hidden conformations in *Aspergillus niger* monoamine oxidase are key for catalytic efficiency. *Angew. Chem. Int. Ed.*, *58*(10), 3097–3101. https://doi.org/10.1002/anie.201812532
58. Thomas, L. L., Christakis, T. J., & Jorgensen, W. L. (2006). Conformation of alkanes in the gas phase and pure liquids. *J. Phys. Chem. B*, *110*(42), 21198–21204. https://doi.org/10.1021/jp064811m
59. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N., & Hilvert, D. (2013). Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.*, *9*(8), 494–498. https://doi.org/10.1038/nchembio.1276
60. Crean, R. M., Gardner, J. M., & Kamerlin, S. C. L. (2020). Harnessing conformational plasticity to generate designer enzymes. *J. Am. Chem. Soc.*, *142*(26), 11324–11342. https://doi.org/10.1021/jacs.0c04924
61. Robic, S., Guzman-Casado, M., Sanchez-Ruiz, J. M., & Marqusee, S. (2003). Role of residual structure in the unfolded state of a thermophilic protein. *Proc. Natl. Acad. Sci. U. S. A.*, *100*(20), 11345–11349. https://doi.org/10.1073/pnas.1635051100

62. Rosenthal, L., Rabolt, J. F., & Hummel, J. (1982). An investigation of the conformational equilibrium of *n*-butane in a solvent using Raman spectroscopy. *J. Chem. Phys.*, *76*(2), 817–820. https://doi.org/10.1063/1.443052
63. Vázquez-Figueroa, E., Yeh, V., Broering, J. M., Chaparro-Riggers, J. F., & Bommarius, A. S. (2008). Thermostable variants constructed via the structure-guided consensus method also show increased stability in salts solutions and homogeneous aqueous-organic media. *Protein Eng. Des. Sel.*, *21*(11), 673–680. https://doi.org/10.1093/protein/gzn048
64. Mayer, S., Rüdiger, S., Ang, H. C., Joerger, A. C., & Fersht, A. R. (2007). Correlation of levels of folded recombinant P53 in *Escherichia coli* with thermodynamic stability in vitro. *J. Mol. Biol.*, *372*(1), 268–276. https://doi.org/10.1016/j.jmb.2007.06.044
65. Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, *103*(15), 5869–5874. https://doi.org/10.1073/pnas.0510098103
66. Adams, M. W., & Kelly, R. M. (1998). Finding and using hyperthermophilic enzymes. *Trends Biotechnol.*, *16*(8), 329–332. https://doi.org/10.1016/s0167-7799(98)01193-7
67. Merz, A., Yee, M.-C., Szadkowski, H., Pappenberger, G., Crameri, A., Stemmer, W. P. C., Yanofsky, C., & Kirschner, K. (2000). Improving the catalytic activity of a thermophilic enzyme at low temperatures. *Biochemistry*, *39*(5), 880–889. https://doi.org/10.1007/pl00003701
68. Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.*, *17*(4), 263–270. https://doi.org/10.1002/humu.22
69. Joerger, A. C., Ang, H. C., & Fersht, A. R. (2006). Structural basis for understanding oncogenic P53 mutations and designing rescue drugs. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(41), 15056–15061. https://doi.org/10.1073/pnas.0607286103

70. Park, S. H., & Bao, G. (2021). CRISPR/Cas9 gene editing for curing sickle cell disease. *Transfusion and Apheresis Science*, *60*(1), 103060. https://doi.org/10.1016/j.transci.2021.103060
71. Pace, C. N., & Shaw, K. L. (2000). Linear extrapolation method of analyzing solvent denaturation curves. *Proteins Struct. Funct. Bioinforma.*, *Suppl 4*, 1–7. https://doi.org/10.1002/1097-0134(2000)41:4+%3C1::aid-prot10%3E3.3.co;2-u
72. Baxa, M. C., Haddadian, E. J., Jumper, J. M., Freed, K. F., & Sosnick, T. R. (2014). Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(43), 15396–15401. https://doi.org/10.1073/pnas.1407768111
73. Horovitz, A., & Fersht, A. R. (1992). Co-operative interactions during protein folding. *J. Mol. Biol.*, *224*(3), 733–740. https://doi.org/10.1016/0022-2836(92)90557-Z
74. Scholtz, J. M., Grimsley, G. R., & Pace, C. N. (2009). Solvent denaturation of proteins and interpretations of the m value. *Methods Enzymol.*, *466*, 549–565. https://doi.org/10.1016/S0076-6879(09)66023-7
75. Fersht, A. (1999). Chapter 17, Protein Stability. In *Structure and Mechanism in Protein Science*. Freeman.
76. Becktel, W. J., & Schellman, J. A. (1987). Protein stability curves. *Biopolymers*, *26*(11), 1859–1878. https://doi.org/10.1002/bip.360261104
77. Robertson, A. D., & Murphy, K. P. (1997). Protein structure and the energetics of protein stability. *Chem. Rev.*, *97*(5), 1251–1267.
78. Pucci, F., Kwasigroch, J. M., & Rooman, M. (2017). SCooP: An accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*, *33*(21), 3415–3422.

79. Pastore, A., Martin, S. R., Politou, A., Kondapalli, K. C., Stemmler, T., & Temussi, P. A. (2007). Unbiased cold denaturation: Low- and high-temperature unfolding of yeast frataxin under physiological conditions. *J. Am. Chem. Soc.*, *129*(17), 5374–5375. https://doi.org/10.1021/ja0714538
80. Razvi, A., & Scholtz, J. M. (2006). Lessons in stability from thermophilic proteins. *Protein Sci.*, *15*(7), 1569–1578. https://doi.org/10.1110/ps.062130306
81. Eijsink, V. G. H., Bjørk, A., G\a aseidnes, S., Sirev\a ag, R., Synstad, B., van den Burg, B., & Vriend, G. (2004). Rational engineering of enzyme stability. *J. Biotechnol.*, *113*(1-3), 105–120. https://doi.org/10.1016/j.jbiotec.2004.03.026
82. Dumon, C., Varvak, A., Wall, M. A., Flint, J. E., Lewis, R. J., Lakey, J. H., Morland, C., Luginbühl, P., Healey, S., Todaro, T., DeSantis, G., Sun, M., Parra-Gessert, L., Tan, X., Weiner, D. P., & Gilbert, H. J. (2008). Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure. *J. Biol. Chem.*, *283*(33), 22557–22564. https://doi.org/10.1074/jbc.M800936200
83. Shoichet, B. K., Baase, W. A., Kuroki, R., & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.*, *92*(2), 452–456. https://doi.org/10.1073/pnas.92.2.452
84. Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L., van Loon, A. P. G. M., & Wyss, M. (2002). The consensus concept for thermostability engineering of proteins: Further proof of concept. *Protein Eng.*, *15*(5), 403–411. https://doi.org/10.1093/protein/15.5.403
85. Magliery, T. J. (2015). Protein stability: Computation, sequence statistics, and new experimental methods. *Current Opinion in Structural Biology*, *33*, 161–168. https://doi.org/10.1016/j.sbi.2015.09.002

86. Jones, B. J., Kan, C. N. E., Luo, C., & Kazlauskas, R. J. (2020). Consensus Finder web tool to predict stabilizing substitutions in proteins. In D. S. Tawfik (Ed.), *Enzyme Engineering and Evolution: General Methods* (Vol. 643, pp. 129–148). Academic Press. https://doi.org/10.1016/bs.mie.2020.07.010
87. Musil, M., Stourac, J., Bendl, J., Brezovsky, J., Prokop, Z., Zendulka, J., Martinek, T., Bednar, D., & Damborsky, J. (2017). FireProt: Web server for automated design of thermostable proteins. *Nucleic Acids Res.*, *45*(W1), W393–W399. https://doi.org/10.1093/nar/gkx285
88. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L., & Bryan, P. N. (1989). Large increases in general stability for subtilisin BPN' through incremental changes in the free-energy of unfolding. *Biochemistry*, *28*(18), 7205–7213. https://doi.org/10.1021/bi00444a012
89. Schwehm, J. M., Fitch, C. A., Dang, B. N., García-Moreno E., B., & Stites, W. E. (2003). Changes in stability upon charge reversal and neutralization substitution in staphylococcal nuclease are dominated by favorable electrostatic effects. *Biochemistry*, *42*(4), 1118–1128. https://doi.org/10.1021/bi0266434
90. Trudeau, D. L., Kaltenbach, M., & Tawfik, D. S. (2016). On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.*, *33*(10), 2633–2641. https://doi.org/10.1093/molbev/msw138
91. Clarke, J., Henrick, K., & Fersht, A. R. (1995). Disulfide mutants of barnase I: Changes in stability and structure assessed by biophysical methods and x-ray crystallography. *J. Mol. Biol.*, *253*(3), 493–504. https://doi.org/10.1006/jmbi.1995.0568
92. Pace, C. N., Grimsley, G. R., Thomson, J. A., & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.*, *263*(24), 11820–11825.

93. Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.*, *151*(2), 261–287. https://doi.org/10.1016/0022-2836(81)90515-5
94. Perry, L. J., & Wetzel, R. (1984). Disulfide bond engineered into T4 lysozyme: Stabilization of the protein toward thermal inactivation. *Science*, *226*(4674), 555–557. https://www.jstor.org/stable/1693872
95. Matsumura, M., Becktel, W. J., Levitt, M., & Matthews, B. W. (1989). Stabilization of phage T4 Iysozyme by engineered disulfide bonds. *Proc. Natl. Acad. Sci. U. S. A.*, *86*, 6562–6566. https://doi.org/10.1073/pnas.86.17.6562
96. Roesler, K. R., & Rao, A. G. (2000). A single disulfide bond restores thermodynamic and proteolytic stability to an extensively mutated protein. *Protein Science*, *9*(9), 1642–1650. https://doi.org/10.1110/ps.9.9.1642
97. Kanaya, S., Katsuda, C., Kimura, S., Nakai, T., Kitakuni, E., Nakamura, H., Katayanagi, K., Morikawa, K., & Ikehara, M. (1991). Stabilization of *Escherichia coli* ribonuclease H by introduction of an artificial disulfide bond. *Journal of Biological Chemistry*, *266*(10), 6038–6044. https://doi.org/10.1016/S0021-9258(18)38080-3
98. Hazes, B., & Dijkstra, B. W. (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng. Des. Sel.*, *2*(2), 119–125. https://doi.org/10.1093/protein/2.2.119
99. Dombkowski, A. A. (2003). Disulfide by Design^{TM}: A computational method for the rational design of disulfide bonds in proteins. *Bioinformatics*, *19*(14), 1852–1853. https://doi.org/10.1093/bioinformatics/btg231
100. Suplatov, D., Timonina, D., Sharapova, Y., & Švedas, V. (2019). Yosshi: A web-server for disulfide engineering by bioinformatic analysis of diverse protein families. *Nucleic Acids Res.*, *47*(W1), W308–W314. https://doi.org/10.1093/nar/gkz385

101. Pikkemaat, M. G., Linssen, A. B. M., Berendsen, H. J. C., & Janssen, D. B. (2002). Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng.*, *15*(3), 185–192. https://doi.org/10.1093/protein/15.3.185
102. Stewart, E. J., Åslund, F., & Beckwith, J. (1998). Disulfide bond formation in the *Escherichia coli* cytoplasm: An *in Vivo* role reversal for the thioredoxins. *EMBO J.*, *17*(19), 5543–5550. https://doi.org/10.1093/emboj/17.19.5543
103. Zhou, H.-X. (2004). Loops, linkages, rings, catenanes, cages, and crowders: Entropy-based strategies for stabilizing proteins. *Acc. Chem. Res.*, *37*(2), 123–130. https://doi.org/10.1021/ar0302282
104. Watanabe, K., & Suzuki, Y. (1998). Protein thermostabilization by proline substitutions. *J. Mol. Catal. B Enzym.*, *4*(4), 167–180. https://doi.org/10.1016/s1381-1177(97)00031-3
105. Fu, H., Grimsley, G. R., Razvi, A., Scholtz, J. M., & Pace, C. N. (2009). Increasing protein stability by improving beta-turns. *Proteins*, *77*(3), 491–498. https://doi.org/10.1002/prot.22509
106. Matthews, B. W., Nicholson, H., & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U. S. A.*, *84*(19), 6663–6667. https://doi.org/10.1073/pnas.84.19.6663
107. Scott, K. A., Alonso, D. O. V., Sato, S., Fersht, A. R., & Daggett, V. (2007). Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(8), 2661–2666. https://doi.org/10.1073/pnas.0611182104
108. Liu, Z., Lemmonds, S., Huang, J., Tyagi, M., Hong, L., & Jain, N. (2018). Entropic contribution to enhanced thermal stability in the thermostable P450 CYP119. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(43), E10049–E10058. https://doi.org/10.1073/pnas.1807473115
109. 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*(46), 7745–7751. https://doi.org/10.1002/anie.200602795

110. Floor, R. J., Wijma, H. J., Colpa, D. I., Ramos-Silva, A., Jekel, P. A., Szymanski, W., Feringa, B. L., Marrink, S. J., & Janssen, D. B. (2014). Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem*, *15*(11), 1660–1672. https://doi.org/10.1002/cbic.201402128

111. Xiao, Z., Bergeron, H., Grosse, S., Beauchemin, M., Garron, M.-L., Shaya, D., Sulea, T., Cygler, M., & Lau, P. C. K. (2008). Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-temperature-guided sequence alignment. *Appl. Environ. Microbiol.*, *74*(4), 1183–1189. https://doi.org/10.1128/AEM.02220-07

112. Petsko, G. A. (2001). Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat." *Methods Enzymol.*, *334*, 469–478. https://doi.org/10.1016/S0076-6879(01)34486-5

113. Gribenko, A. V., Patel, M., Liu, J., McCallum, S., Wang, C., & Makhatadze, G. I. (2009). Rational stabilization of enzymes by computational redesign of surface charge–charge interactions. *Proc. Natl. Acad. Sci. U. S. A.*, *106*(8), 2601–2606. https://doi.org/10.1073/pnas.0808220106

114. Contessoto, V. G., de Oliveira, V. M., Fernandes, B. R., Slade, G. G., & Leite, V. B. P. (2018). TKSA-MC: A web server for rational mutation through the optimization of protein charge interactions. *Proteins*, *86*(11), 1184–1188. https://doi.org/10.1002/prot.25599

115. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987–2998. https://doi.org/10.1021/bi902153g

116. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Res.*, *33*(Web Server issue), W382–8. https://doi.org/10.1093/nar/gki387

117. Korkegian, A., Black, M. E., Baker, D., & Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. *Sci. Wash.*, *308*(5723), 857–860. https://doi.org/10.1126/science.1107387
118. Arabnejad, H., Dal Lago, M., Jekel, P. A., Floor, R. J., Thunnissen, A.-M. W. H., Terwisscha van Scheltinga, A. C., Wijma, H. J., & Janssen, D. B. (2017). A robust cosolvent-compatible halohydrin dehalogenase by computational library design. *Protein Eng. Des. Sel.*, *30*(3), 173–187. https://doi.org/10.1093/protein/gzw068
119. Jin, Y., Johannissen, L. O., & Hay, S. (2021). Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins*, *89*(8), 915–921. https://doi.org/10.1002/prot.26068
120. Weinstein, J. J., Goldenzweig, A., Hoch, S., & Fleishman, S. J. (2021). PROSS 2: A new server for the design of stable and highly expressed protein variants. *Bioinformatics*, *37*(1), 123–125. https://doi.org/10.1093/bioinformatics/btaa1071
121. Gao, D., Narasimhan, D. L., Macdonald, J., Brim, R., Ko, M.-C., Landry, D. W., Woods, J. H., Sunahara, R. K., & Zhan, C.-G. (2009). Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, *75*(2), 318–323. https://doi.org/10.1124/mol.108.049486
122. van den Burg, B., Vriend, G., Veltman, O. R., & Eijsink, V. G. H. (1998). Engineering an enzyme to resist boiling. *Proc. Natl. Acad. Sci. U. S. A.*, *95*(5), 2056–2060. https://doi.org/10.1073/pnas.95.5.2056
123. Bjørk, A., Dalhus, B., Mantzilas, D., Sirev\a ag, R., & Eijsink, V. G. H. (2004). Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface. *J. Mol. Biol.*, *341*(5), 1215–1226. https://doi.org/10.1016/j.jmb.2004.06.079

124. Kabashima, T., Li, Y., Kanada, N., Ito, K., & Yoshimoto, T. (2001). Enhancement of the thermal stability of pyroglutamyl peptidase I by introduction of an intersubunit disulfide bond. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1547*(2), 214–220. https://doi.org/10.1016/s0167-4838(01)00185-6
125. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., & Trout, B. L. (2010). Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B*, *114*(19), 6614–6624. https://doi.org/10.1021/jp911706q
126. Kuriata, A., Iglesias, V., Pujols, J., Kurcinski, M., Kmiecik, S., & Ventura, S. (2019). Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. *Nucleic Acids Res.*, *47*(W1), W300–W307. https://doi.org/10.1093/nar/gkz321
127. Van Durme, J., De Baets, G., Van Der Kant, R., Ramakers, M., Ganesan, A., Wilkinson, H., Gallardo, R., Rousseau, F., & Schymkowitz, J. (2016). Solubis: A webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel.*, *29*(8), 285–289. https://doi.org/10.1093/protein/gzw019
128. Suen, W.-C., Zhang, N., Xiao, L., Madison, V., & Zaks, A. (2004). Improved activity and thermostability of *Candida antarctica* lipase B by DNA family shuffling. *Protein Eng. Des. Sel.*, *17*(2), 133–140. https://doi.org/10.1093/protein/gzh017
129. Lawrence, M. S., Phillips, K. J., & Liu, D. R. (2007). Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.*, *129*(33), 10110–10112. https://doi.org/10.1021/ja071641y
130. Robinson, N. E. (2002). Protein deamidation. *Proc. Natl. Acad. Sci. U. S. A.*, *99*(8), 5283–5288. https://doi.org/10.1073/pnas.082102799
131. Mark, D. F., Lu, S. D., Creasey, A. A., Yamamoto, R., & Lin, L. S. (1984). Site-specific mutagenesis of the human fibroblast interferon gene. *Proc. Natl. Acad. Sci. U. S. A.*, *81*(18), 5662–5666.

132. Kontermann, R. E. (2009). Strategies to extend plasma half-lives of recombinant antibodies. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *23*(2), 93–109. https://doi.org/10.2165/00063030-200923020-00003
133. Strohl, W. R. (2015). Fusion proteins for half-life extension of biologics as a strategy to make biobetters. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *29*(4), 215–239. https://doi.org/10.1007/s40259-015-0133-6
134. Cheang, J. Y., & Moyle, P. M. (2018). Glucagon-like peptide-1 (GLP-1)-based therapeutics: Current status and future opportunities beyond type 2 diabetes. *ChemMedChem*, *13*(7), 662–671. https://doi.org/10.1002/cmdc.201700781
135. Feldmann, M. (2002). Development of anti-TNF therapy for rheumatoid arthritis. *Nat. Rev. Immunol.*, *2*(5), 364–371. https://doi.org/10.1038/nri802
136. Nielsen, J. E., & Borchert, T. V. (2000). Protein engineering of bacterial α-amylases. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1543*(2), 253–274. https://doi.org/10.1016/s0167-4838(00)00240-5
137. Jones, B. J., Lim, H. Y., Huang, J., & Kazlauskas, R. J. (2017). Comparison of five protein engineering strategies for stabilizing an α/β-hydrolase. *Biochemistry*, *56*(50), 6521–6532. https://doi.org/10.1021/acs.biochem.7b00571
138. Thieker, D. F., Maguire, J. B., Kudlacek, S. T., Leaver-Fay, A., Lyskov, S., & Kuhlman, B. (2022). Stabilizing proteins, simplified: A Rosetta-based webtool for predicting favorable mutations. *Protein Sci.*, *31*(10). https://doi.org/10.1002/pro.4428
139. Banta, S., Dooley, K., & Shur, O. (2013). Replacing antibodies: Engineering new binding proteins. *Annu. Rev. Biomed. Eng.*, *15*(1), 93–113. https://doi.org/10.1146/annurev-bioeng-071812-152412
140. Rossi, A. M., & Taylor, C. W. (2011). Analysis of protein-ligand interactions by fluorescence polarization. *Nat Protoc*, *6*(3), 365–387. https://doi.org/10.1038/nprot.2011.305

141. Pollard, T. D. (2010). MBOC technical perspective: A guide to simple and informative binding assays. *Mol. Biol. Cell*, *21*(23), 4061–4067. https://doi.org/10.1091/mbc.E10-08-0683
142. Jeffrey, P. D., Strong, R. K., Sieker, L. C., Chang, C. Y. Y., Campbell, R. L., Petsko, G. A., Haber, E., Margolies, M. N., & Sheriff, S. (1993). 26-10 Fab-digoxin complex: Affinity and specificity due to surface complementarity. *Proc. Natl. Acad. Sci. U. S. A.*, *90*(21), 10310–10314. https://doi.org/10.1073/pnas.90.21.10310
143. Tian, W., Chen, C., Lei, X., Zhao, J., & Liang, J. (2018). CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.*, *46*(W1), W363–W367. https://doi.org/10.1093/nar/gky473
144. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, *255*(5041), 178–183. https://doi.org/10.1126/science.1553543
145. Jones, B. J., Evans, R. L., Mylrea, N. J., Chaudhury, D., Luo, C., Guan, B., Pierce, C. T., Gordon, W. R., Wilmot, C. M., & Kazlauskas, R. J. (2020). Larger active site in an ancestral hydroxynitrile lyase increases catalytically promiscuous esterase activity. *PLoS ONE*, *15*, e0235341. https://doi.org/10.1371/journal.pone.0235341
146. DeLorbe, J. E., Clements, J. H., Teresk, M. G., Benfield, A. P., Plake, H. R., Millspaugh, L. E., & Martin, S. F. (2009). Thermodynamic and structural effects of conformational constraints in protein-ligand interactions. Entropic paradoxy associated with ligand preorganization. *J. Am. Chem. Soc.*, *131*(46), 16758–16770. https://doi.org/10.1021/ja904698q
147. Barros, E. P., Schiffer, J. M., Vorobieva, A., Dou, J., Baker, D., & Amaro, R. E. (2019). Improving the efficiency of ligand-binding protein design with molecular dynamics simulations. *J. Chem. Theory Comput.*, *15*(10), 5703–5715. https://doi.org/10.1021/acs.jctc.9b00483

148. Thakkar, S., Nanaware-Kharade, N., Celikel, R., Peterson, E. C., & Varughese, K. I. (2014). Affinity improvement of a therapeutic antibody to methamphetamine and amphetamine through structure-based antibody engineering. *Sci. Rep.*, *4*, 3673. https://doi.org/10.1038/srep03673
149. Zhao, Q., Ahmed, M., Guo, H., Cheung, I. Y., & Cheung, N.-K. V. (2015). Alteration of electrostatic surface potential enhances affinity and tumor killing properties of anti-ganglioside GD2 monoclonal antibody hu3F8. *J. Biol. Chem.*, *290*(21), 13017–13027. https://doi.org/10.1074/jbc.M115.650903
150. Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, *25*(10), 1171–1176. https://doi.org/10.1038/nbt1336
151. Selzer, T., Albeck, S., & Schreiber, G. (2000). Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.*, *7*(7), 537–541. https://doi.org/10.1038/76744
152. Bugnon, M., Röhrig, U. F., Goullieux, M., Perez, M. A. S., Daina, A., Michielin, O., & Zoete, V. (2024). SwissDock 2024: Major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. *Nucleic Acids Res*, *52*(W1), W324–W332. https://doi.org/10.1093/nar/gkae300
153. Zoete, V., Schuepbach, T., Bovigny, C., Chaskar, P., Daina, A., Röhrig, U. F., & Michielin, O. (2016). Attracting cavities for docking. Replacing the rough energy landscape of the protein by a smooth attracting landscape. *J Comput Chem*, *37*(4), 437–447. https://doi.org/10.1002/jcc.24249
154. Röhrig, U. F., Goullieux, M., Bugnon, M., & Zoete, V. (2023). Attracting Cavities 2.0: Improving the flexibility and robustness for small-molecule docking. *J. Chem. Inf. Model.*, *63*(12), 3925–3940. https://doi.org/10.1021/acs.jcim.3c00054

155. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., & Baker, D. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, *501*(7466), 212–216. https://doi.org/10.1038/nature12443

156. Morin, A., Kaufmann, K. W., Fortenberry, C., Harp, J. M., Mizoue, L. S., & Meiler, J. (2011). Computational design of an endo-1,4-beta-xylanase ligand binding site. *Protein Eng. Des. Sel.*, *24*(6), 503–516. https://doi.org/10.1093/protein/gzr006

Problems

1. *The application benefits depend on the function of the target protein.* Three examples of improvements achieved by engineering a more stable protein mentioned in this chapter were rubisco activase, Covid-19 vaccine and subtilisin. For each of these three target proteins: a) identify the function of this protein in the application, b) describe the type of stability that was improved by the engineering, and c) how this increased stability improved the function of the target protein in the application.

2. *Proteins in nature are just good enough for the natural function.* Proteins acquire point mutations through mistakes in DNA replication. Natural selection selects against deleterious mutations - those that hinder the function of a protein. Although mutations can also yield improvements in protein properties, these improvements are lost if they do not contribute to the fitness of the organism. For example, imagine a metabolic enzyme that acquires a substitution that makes it ten-fold faster. The organism does not grow ten-fold faster, but only two-fold faster because a second enzyme in the pathway now limits the rate. What will happen as these two protein accumulate additional substitutions? Keep in mind that beneficial substitutions are rare, while detrimental substitutions are common.

3. *Distinguish between binding as a goal and as a hypothesis for how to increase reactivity.* Tumors derived from the nervous system such as glioblastomas are more sensitive than normal tissues to L-methionine starvation. Injections of bacterial methionine-γ-lyase can deplete methionine and inhibit tumor growth. Researchers sought to avoid using a bacterial enzyme by engineering human cystathionine-γ-lyase (hCGL) to accept a new substrate, L-methionine,[33] Figure 1.5. hCGL showed no detectable activity with L-methionine. The researchers hypothesized that L-methionine does not react because it does not bind to the active site of hCGL. Three substitutions in the active site of hCGL (E59N, R119L, E339V) created variant hCGL-NLV with a more hydrophobic binding site. This variant protein improved both binding and catalysis; it could both bind L-methionine and catalyze its cleavage. Although the protein engineering changed two properties only one of these was the true goal. Identify the true goal by considering the possibility that the researchers changed only one of these properties: improved binding with no change in catalysis or improved catalysis with no change in binding. Which of these hypothetical improvements would yield an enzyme that could be useful as a cancer treatment?

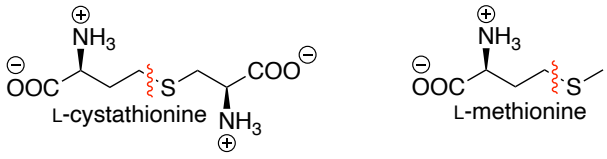


Figure 1.5. Cystathionine-γ-lyase catalyzes the cleavage of cystathionine at the bond marked with a wavy red line. L-Methionine is smaller and more hydrophobic than cystathionine. Methionine-γ-lyase catalyzes the cleavage of L-methionine at the bond marked with a wavy red line.

4. *Industrial enzymes in your kitchen* Automatic dishwasher detergents contain mixtures of enzymes. Suggest three enzymes that catalyze different reactions that may be useful to include in dishwasher detergents and ex-

plain their potential role. Suggest improvements in two different types of stability that may be required for each enzyme to be useful in dishwasher detergents.

Answers

1. Rubisco activase: a) Function: An enzyme that remodels Rubisco active sites to maintain CO_2 fixation in photosynthesis. b) Stability improved: Thermal stability — resistance to inactivation at elevated leaf temperatures. c) Benefit: Plants engineered with heat-stable rubisco activase maintain photosynthetic efficiency under heat stress, improving crop productivity in warmer climates. COVID-19 vaccine antigen (spike protein) a) Function: Viral surface glycoprotein used as the immunogen to elicit protective antibodies. b) Stability improved: Conformational stability — spike engineered into a prefusion-stabilized form. c) Benefit: The stabilized antigen better mimics the native viral surface structure, increasing neutralizing antibody responses and vaccine effectiveness. Subtilisin a) Function: A protease used in industrial applications, especially as an active ingredient in laundry detergents. b) Stability improved: Chemical stability — resistance to autolysis, oxidation, and inactivation by surfactants or high pH. The requirements are broader than just oxidative stability. c) Benefit: The stabilized enzyme retains activity longer in harsh detergent conditions, improving cleaning performance and extending product shelf life.

2. As mutations accumulate, most substitutions will be neutral or deleterious rather than beneficial. The initially improved enzyme will gradually lose its advantage because destabilizing or activity-reducing mutations are much more frequent than further rare beneficial ones. Since the organism's growth is constrained by other steps in the pathway, selection pressure to maintain the extra efficiency of the faster enzyme is weak. Thus, the first enzyme will tend to drift back toward "good enough"

performance, while the newly rate-limiting enzyme may now experience stronger selection for stabilizing or activity-enhancing mutations. Overall, proteins converge on the minimum performance sufficient for organismal fitness, rather than maximizing their possible catalytic or stability potential.

3. The true goal was improved catalysis, not just binding. Binding alone, without catalytic turnover, would not deplete methionine and thus would be useless as a cancer treatment. In contrast, if catalysis improved while binding affinity remained unchanged, the enzyme could still process L-methionine effectively at physiological concentrations. Therefore, binding was a hypothesis for how to enable catalysis, but the actual engineering goal was to create an enzyme that catalyzes methionine cleavage.

4. Protease (e.g., subtilisin): Cleaves peptide bonds in food residues such as egg or meat; Amylase (e.g., α-amylase): Hydrolyzes starch from foods like potatoes or pasta into soluble sugars; Lipase: Breaks down triglycerides and fat residues into glycerol and fatty acids; Cellulase: Hydrolyzes cellulose and microfibrils from plant-based food soils (e.g., vegetable fibers). Also helps prevent redeposition of small particles on dishes; Mannanase (β-mannanase): Degrades galactomannans found in thickeners like guar gum, common in sauces and ice cream residues. All enzymes require alkaline stability (stable and active at detergent pH of 9–11), oxidative stability (some contain bleach or other oxidants to remove stains), thermal stability (hot wash cycles), surfactant stability (resistance to denaturation by detergent surfactants), shelf-life stability (resistance to long storage in formulated detergent powders/liquids).

2 Making Protein Variants

Summary. Protein expression converts DNA into protein. It involves two steps: the transcription of the coding region of the DNA into mRNA by RNA polymerase followed by the translation of the mRNA into protein by the ribosome. Proteins are manufactured by modification of the natural protein synthesis pathways in microbes like *Escherichia coli*. The modifications specify the protein to make and increase the amount of protein made. The DNA encoding the target proteins is carried on plasmids which are circular, independently replicating genetic elements. Protein synthesis also requires DNA regions outside the coding sequence to start and stop transcription of the DNA to mRNA and translation of the mRNA to protein. Experimental techniques for construction of plasmids include template-independent DNA synthesis, cutting of DNA with restrictions enzymes and assembly of DNA fragments using the polymerase chain reaction.

Key learning goals

- Protein overexpression plasmids encode the target proteins, the control elements needed to make large amounts of protein, and other regions required for plasmids to replicate in microbes.
- The main control of protein expression is at transcription of DNA to mRNA step. The *lac* operator (a region of DNA) and lac repressor (a protein that binds to the *lac* operator) serve as on/off switches for

transcription of DNA to mRNA. Addition of a non-hydrolyzable lactose analog, IPTG, turns on transcription. Other control points are the plasmid copy number and the strength of the ribosome binding site.

- The pET21 plasmid contains the gene encoding the target protein, T7 RNA polymerase promoter & terminator, *lac* operator, *lac*I gene and an ampicillin resistance gene. The pET plasmids uses T7 RNA polymerase, which is 5-10-fold more efficient than endogenous *E. coli* RNA polymerase.
- Site-directed mutagenesis uses the polymerase chain reaction with chemically or enzymatically synthesized DNA fragments (primers) to alter the DNA sequence to encode protein variants.

2.1 Protein expression (DNA $\rightarrow$ mRNA $\rightarrow$ protein)

Protein expression is the conversion of a DNA sequence into the corresponding protein. It consists of two steps. First, the DNA encoding the target protein is transcribed into messenger RNA (mRNA) by RNA polymerase. Second, this mRNA is translated into protein by the ribosome.

In biotechnology, proteins are manufactured by microbes, typically bacteria or yeast, that have been modified to make large amounts of the target protein. In the laboratory, researchers grow these microbes in 2-liter shake flasks containing approximately 400 mL of culture media (a mixture of amino acids, sugars, and other nutrients) in a warm incubator. Industrial manufacture uses large fermenters similar to those in a brewery. The details below refer to protein expression in prokaryotes like *Escherichia coli*.

Protein expression requires control elements outside the region that encodes the target protein. One set of control elements bind and release the RNA polymerase that transcribes the DNA, while another set of control elements bind and release the ribosome that translates the genetic

code into protein. Thus, the DNA surrounding the gene that encodes the target protein also contains four additional regions (two start signals, two stop signals) outside the DNA sequence that encodes the target protein.

Transcription starts by binding the RNA polymerase to the double-stranded DNA upstream of the coding sequence, Figure 2.1. The recognition site, called the promoter, includes of two regions each 6 bases long. One occurs approximately 10 base pairs upstream of the start codon and the other occurs 35 base pairs upstream of the start codon. The RNA polymerase is larger than this region so that it covers a larger region when it binds - about 40 base pairs upstream and 20 base pairs downstream of the start codon.

Upon binding, the RNA polymerase unwinds approximately 16 bases of DNA around the start codon to expose single-stranded DNA as a template for transcription. The RNA polymerase starts the synthesis of an RNA strand that is complementary to the DNA. As the synthesis proceeds the polymerase unwinds the template DNA ahead of it and rewinds the DNA behind it, maintaining an unwound region of about 16 base pairs in the region of transcription.

RNA synthesis continues until the polymerase encounters a termination signal. Here the transcription stops, the RNA is released from the polymerase, and the polymerase dissociates from its DNA template. The termination signal in *E. coli* consists of a symmetrical inverted repeat of a GC-rich sequence followed by four or more A residues. Transcription of the GC-rich inverted repeat results in the formation of a segment of RNA that can form a stable stem-loop structure by complementary base pairing. This structure in the RNA disrupts its base pairing with the DNA template, which causes the termination. Termination occurs *after* the terminator region because it is the RNA, not the DNA that forms the stem-loop structure. The DNA must first be transcribed into RNA to form the stem-loop structure that causes termination.

Translation starts when the ribosome binds to the mRNA. The ribosome binding site (Shine-Dalgarno sequence) is approximately 8 bases up-

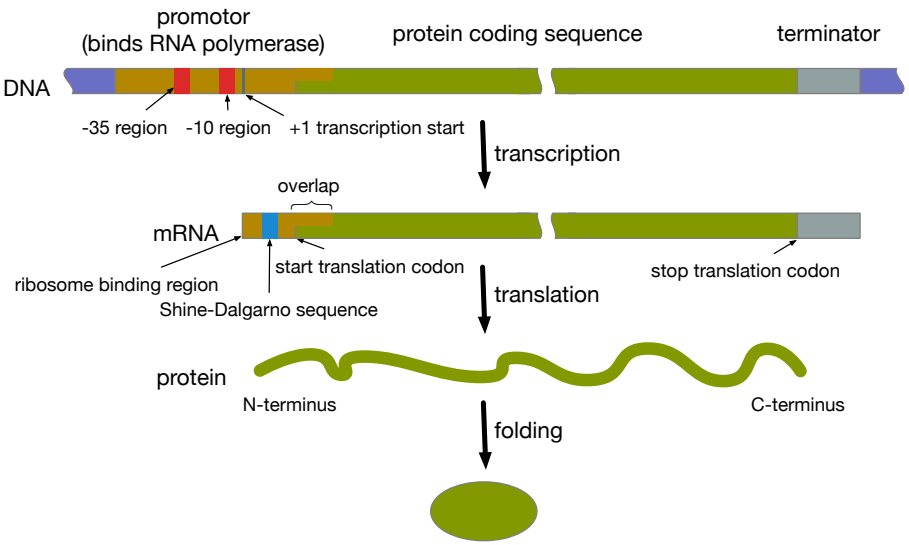


Figure 2.1. Protein expression requires control elements outside the protein coding region. The transcription of DNA into RNA requires an upstream region to bind the RNA polymerase (promoter region) and a downstream region to release the RNA polymerase (terminator region). The promoter region includes two six-base regions at 10 bp upstream and 35 bp upstream that recognize the RNA polymerase. The translation of RNA to protein requires a ribosome binding site, a start codon and a stop codon. Both the RNA polymerase and the ribosome are large molecular machines. Although they recognize and bind to the regions marked, they also cover a much larger region upon binding including part of the protein coding region marked overlap.

stream of the start codon (methionine codon), but the ribosome covers a much larger region including the start codon and the beginning of the coding region. Protein synthesis occurs as the ribosome moves along the RNA sequence in groups of three nucleotides and attaches the next amino acid to the growing protein chain. Termination occurs when the ribosome encounters a stop codon and detaches from the RNA and releases the completed protein. Thus, translation requires a ribosome binding site before the coding sequence and a stop codon after the coding region. The start codon, also required, is within the coding region.

The DNA sequence encoding the control elements for expression of a β-lactamase protein, Figure 2.2, shows a specific example of these elements.

2.2 Protein overexpression

Cells typically make small amounts of each protein, but to manufacture proteins for biotechnology applications one wants to make as much protein as possible. This overproduction of the target protein is called protein overexpression. The two common strategies to increase protein expression are to use a more efficient RNA polymerase and to include multiple copies of the DNA encoding the target protein. These approaches are very effective; up to 50% of the protein in an *E. coli* culture can be the target protein.

2.2.1 A better RNA polymerase

E. coli bacteria contain an endogenous RNA polymerase, but RNA polymerases from bacteriophage are often faster. Bacteriophage RNA polymerases hijack the cellular machinery of the host bacteria to make phage proteins instead of bacterial proteins. The most common replacement

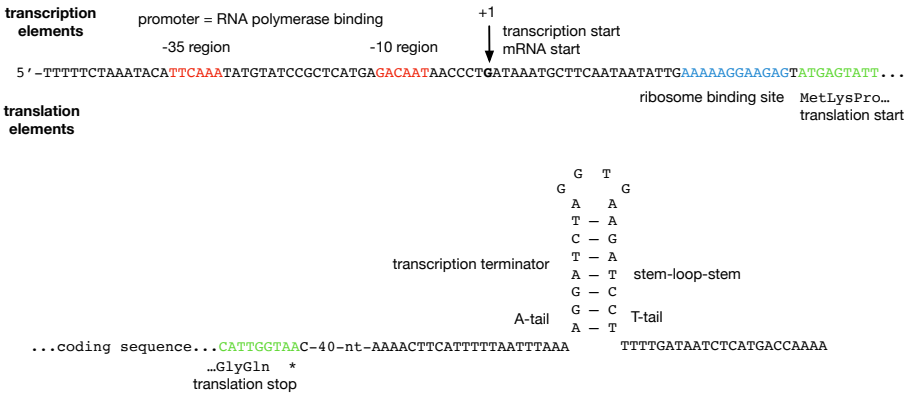


Figure 2.2. DNA sequence for the expression of a β-lactamase protein showing the transcription and translation control elements. The -35 and -10 regions (red text) in the promoter bind the endogenous RNA polymerase, but the polymerase extends beyond these regions. The ribosome binds to the ribosome binding site (blue text), but also extends beyond this region. During transcription, the terminator region of the RNA forms the stem-loop-stem structure shown, which releases the mRNA from the polymerase and stops transcription. The DNA does not for the stem-loop-stem structure.

polymerase for increased protein expression is the T7 RNA polymerase, which originates from bacteriophage T7. The T7 RNA polymerase catalyzes RNA synthesis 5-10-fold faster than the *E. coli* RNA polymerase, so more than the normal amount of mRNA will be produced, which will lead to more protein.

Overexpressing the target protein with T7 RNA polymerase requires a gene that encodes and expresses the T7 RNA polymerase protein. Some strains of *E. coli* (such as BL21(DE3) and BL21(DE3)pLysS) have been engineered to contain the T7 RNA polymerase gene in their genomic DNA. These strains should be used for protein overexpression that requires the T7 RNA polymerase.

Overexpressing the target protein with T7 RNA polymerase also requires a different promoter upstream of the target gene so that it binds the T7, not the *E. coli*, RNA polymerase. The DNA sequence of the T7 promoter differs from the endogenous promoters in *E. coli*.

2.2.2 Multiple copies of the target protein gene

A second way to increase the amount of protein is to have many copies of the DNA that encodes the target protein and its control elements. Protein overexpression uses plasmids, which are small (5-10 kb), independently replicating, circular DNA molecules. The pET plasmid, which will be discussed later in this chapter, makes 15-20 copies of itself in *E. coli*. These multiple copies of the target protein DNA increase the amount of protein that is produced. pET plasmids are classified as low copy number plasmids; some high copy number plasmids make 500-700 copies of themselves in each cell.

2.2.3 Ribosome binding site

The slow step of translating mRNA into protein is the initiation of protein synthesis, which depends on binding of the ribosome to the ribosome binding site (Shine-Dalgarno sequence). Optimization of the initiation step can include optimizing the distance between the Shine-Dalgarno sequence and the start codon, eliminating any secondary structure in the mRNA that could hinder binding of the ribosome, and adding additional binding elements upstream of the Shine-Dalgarno sequence.[34]

2.3 Turning protein overexpression on or off

Strong overexpression of the target protein hinders the growth of microbes. Diverting most of the cell's energy and resources to synthesis of the target protein can even prevent growth entirely. The solution is to initially turn off the overepression of the target protein to allow the bacteria to grow, then turn it on to produce the protein. During overexpression phase the bacteria no longer multiply, but instead make the target protein. The most commonly used on/off switch is the *lac* operator.

The *lac* operator originates from the lactose operon or *lac* operon. An operon is a set of genes whose expression is regulated together. The *lac* operon consists of three structural genes and a *lac* operator, Figure 2.3. The *lac* operator is a section of DNA that controls the expression of the three genes by reversibly binding the *lac* repressor protein, which is encoded by the *lac*I gene elsewhere in the genome. The *lac* operator is in the promoter region where the RNA polymerase must bind, so when the *lac* repressor protein binds to the *lac* operator, it blocks the RNA polymerase from binding. By default, the expression of the three genes in the lactose operon is off. Binding of allolactose to the *lac* repressor protein changes its shape so that it dissociates from the *lac* operator. Expression is now turned on because the RNA polymerase can bind to the promoter.

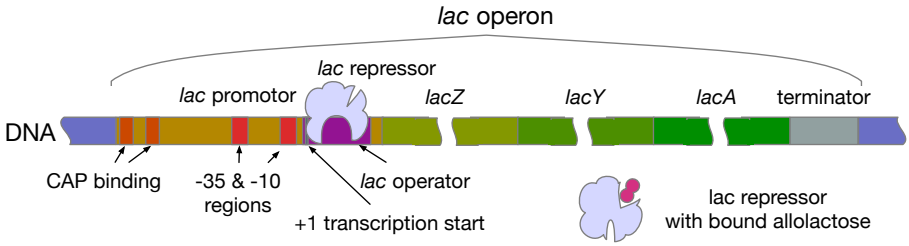


Figure 2.3. The *lac* operon consists of three structural genes and control elements that turn on expression of these genes only when lactose is present in the cell. By default the *lac* repressor protein binds to the *lac* operator and prevents the RNA polymerase from binding in this region. When allolactose is present, it binds to the *lac* repressor protein and changes it shape so that it no longer binds to the *lac* operator. Upon release of the *lac* repressor protein, the RNA polymerase can bind and start transcription of the three structural genes.

The off state of the lactose operator is not perfect. The *lac* repressor protein sometimes dissociates from the *lac* operator even with no lactose present leading to leakiness, that is, constant low-level expression. Given the strong expression from the T7 promoter, the imperfect off state can lead to difficulties in cell growth.

To improve the off state, researchers use a low copy number plasmid and two copies of the *lac* operator. The low copy number of the plasmid reduces the probability that the *lac* repressor protein spontaneously dissociates from the *lac* operator to turn on protein synthesis.

The two *lac* operator switches are used to turn off expression of both the target gene and the T7 RNA polymerase gene. Overexpression of the target protein first requires expression of the T7 RNA polymerase. Placing the *lac* operator in the promoter region blocks this first step. (Expression of the T7 RNA polymerase uses the endogenous *E. coli* promoter.) The second location for the *lac* operator is in the T7 promoter region upstream of

the target gene. In the no lactose state, the *lac* repressor protein should prevent expression of the T7 RNA polymerase and of the target protein thus decreasing the leakiness of the off state.

To further ensure that protein overexpression is turned off, some researcher add a third switch - an inhibitor of the T7 RNA polymerase. This inhibitor is a protein, T7 lysozyme. It requires an additional plasmid, pLysS, that encodes synthesis of the T7 lysozyme protein.

In the presence of lactose or an analog, the switches are on. Lactose binds to the lactose repressor protein, which changes its shape in a way that releases it from the two *lac* operators. Upon release of the lactose repressor protein, the endogenous *E. coli* polymerase binds and starts to transcribe the gene encoding the T7 RNA polymerase. After it is expressed, the T7 RNA polymerase can bind to the promoter upstream of the target gene to initiate protein expression.

Lactose induces the *lac* promoter indirectly. Lactose can be isomerized into allolactose, which induces the *lac* promoter.[35] The natural inducer of the *lac* promoter is likely galactosyl glycerol found in plant galactolipids and common in the diet of ruminants and herbivores,[36] Figure 2.4.

For biotechnology applications, researchers use a non-hydrolyzable analog of galactosyl glycerol to induce the *lac* promoter. Isopropyl β-D-thiogalactopyranoside, IPTG, contains a sulfur atom at the glycosidic link, which prevent hydrolysis. Induction starts upon addition of IPTG and continues indefinitely since the compound does not hydrolyze.

A less expensive alternative to the *lac* promoter often used in industry is the λ promoter, which is turned on by heating the culture to 42 °C for 30 min instead of adding IPTG.[37] Initially the culture is grown at 30 °C and a thermolabile variant of the λ cI857 repressor protein binds to the operator site of the λ promoter and blocks transcription. Upon heating the repressor irreversibly denatures and no longer binds to the operator allowing protein expression to proceed.

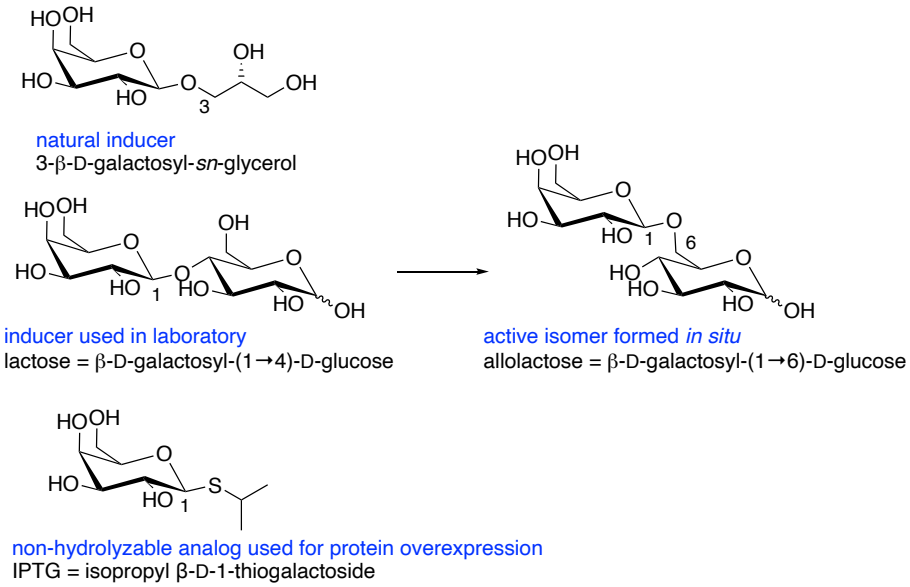


Figure 2.4. Inducers of the *lac* promoter. The natural inducer is likely galactosyl glycerol found in plant galactolipids. Lactose itself is not an inducer. IPTG is an non-hydrolyzable analog of galactosyl glycerol and is used to induce the *lac* promoter in biotechnology.

In summary, there are two or three off switches for protein expression using pET plasmids. Two of the off switches are lac operators in the promoter region: one in the T7lac promoter upstream of the coding region on the plasmid and one in the endogenous promoter upstream of the gene for the T7 RNA polymerase on chromosome. Both of these switches control the synthesis of mRNA. The addition of IPTG turns these switches on. The final switch is the low level expression of T7 lysozyme by the pLysS plasmid to inhibit the T7 RNA polymerase. The BL21(DES)pLysS strain of *Escherichia coli* includes this plasmid.

Figure 2.5 shows an example gene and its associated control elements for expression in *Escherichia coli*. This coding region encodes a poly(ethyleneterephthalate)-degrading enzyme, a PETase.[38] The gene originates from the bacterium *Ideonella sakaiensis*, but the DNA sequence has been optimized for expression in *Escherichia coli*. This optimization replaces codons that are rarely used in *E. coli* with a synonymous, more commonly used codon. The start and stop elements for the RNA polymerase are the T7 promoter and the T7 terminator. The T7 promoter contains the *lac* operator, which serves as an on/off switch. The start elements for translation are the ribosome binding site (T7_trans_en_RBS) and the start codon (ATG) and the stop element for translation is the stop codon (TGA).

The PETase gene contains eight additional amino acid residues at the C-terminus: a linker (not labelled) and a His-tag. The linker (CTCGAG) encodes the dipeptide Leu-Glu, while the His-tag (CACCACCATCACCACCAC) encodes six histidine residues. The addition of six histidine residues to the protein simplifies purification. After protein expression, the mixture of proteins is added to a column containing bound nickel or cobalt ions, which have a high affinity for the imidazole ring of histidine. Most proteins bind poorly to the column and elute quickly, but the His-tagged protein binds tightly. After washing away the unwanted proteins, eluting with a buffer containing imidazole releases the His-tagged protein.

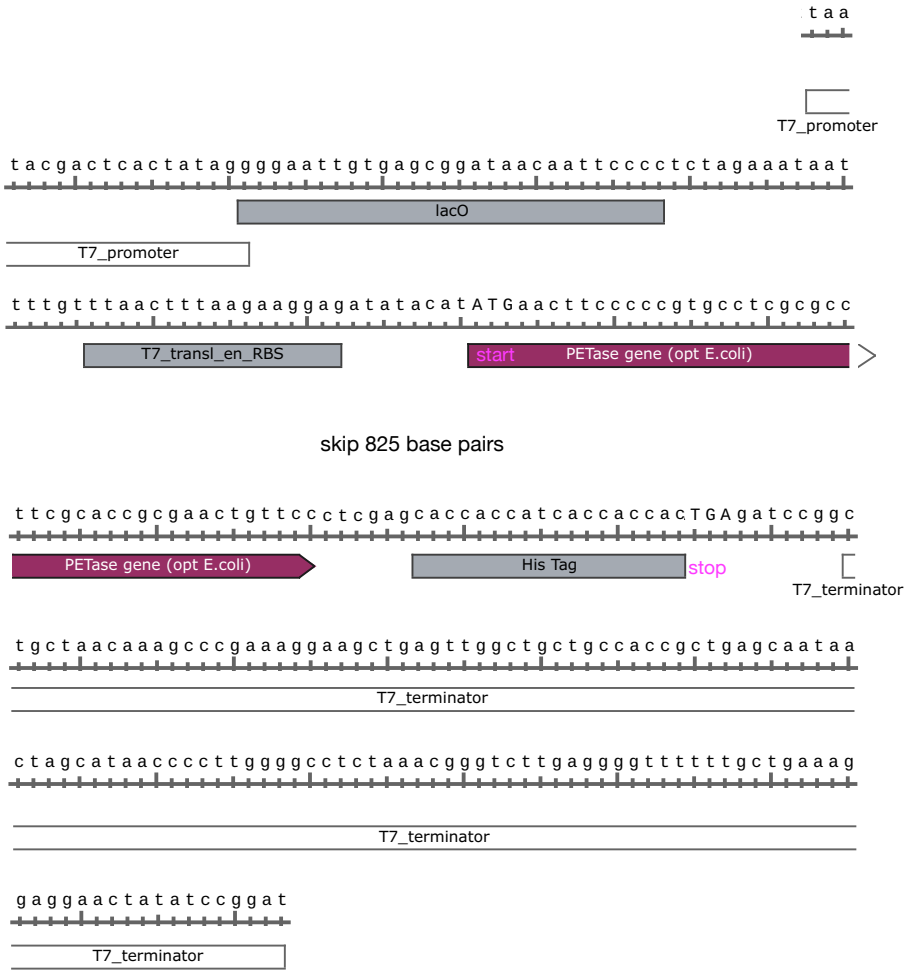


Figure 2.5. Control elements before and after the DNA encoding a PETase gene (maroon fill). The T7 promoter region binds the T7 RNA polymerase. The *lac* operator (lacO) is an on/off switch to control expression. The ribosome binding site is labelled T7_trans_en_RBS. The PETase gene is 870 bp long not including the added His-tag; the central 825 bp are not shown to save space. The His-tag encodes a six-histidine C-terminal tag added to the PETase. The T7_terminator forms secondary structures in the RNA that stop the RNA polymerase.

2.4 pET plasmids

The pET family of plasmids are often used for protein overexpression. The pET family of plasmids all use the T7 RNA polymerase (pET stands for plasmid for Expression by T7 RNA polymerase), but differ in the some details. The expression plasmid used to make the PETase protein mentioned above is derived from the pET21 plasmid, Figure 2.6. The pET21 plasmid contains an origin of replication, which determines its copy number, and genes to express the lactose repressor protein (*lac*I), and an ampicillin resistance protein (AmpR). The lactose repressor protein binds to the *lac* operator to turn off protein expression. The expression of *lac*I and AmpR are constitutive (always on) at a normal level (the promoters bind the *E. coli* RNA polymerase).

The ampicillin resistance protein or β-lactamase is a selectable marker that ensures that only bacteria containing the pET21 plasmid grow in the culture. Researchers add ampicillin to the growth media, which prevents bacterial growth. Bacteria containing the pET21 plasmid express the β-lactamase, so they grow in the presence of ampicillin. This expression of the AmpR gene is constitutive and at a normal level. Figure 2.2 above shows the sequences of control elements for the AmpR gene on a pET21 plasmid.

The different pET plasmids (about fifty) differ in mainly in the purification tag. The His-tag may be at the N- or C-terminus; it may be removable with a protease if the linker contains a protease-sensitive site. Some plasmids encode a glutathione-S-transferase (GST) purification tag. GST is a 26 kDa protein. It has a high affinity for glutathione and yields higher purity protein than the His-tag. Another family of plasmids used for protein expression are the pBAD plasmids, which use an arabinose promoter, Table 2.1. Protein expression from these plasmids is more tightly controlled than for pET plasmids and can be turned off more completely or tuned to different levels. Protein synthesis is turned on by the addition of different amounts of arabinose.

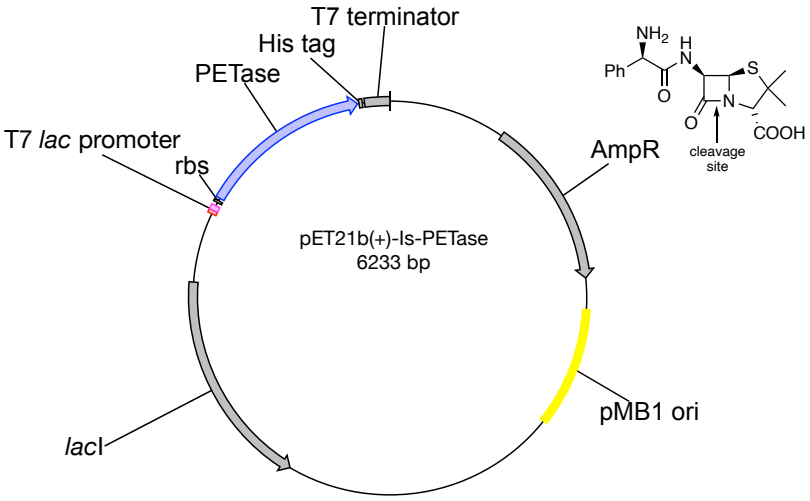


Figure 2.6. A plasmid derived from pET21 to overexpress the PETase protein in *Escherichia coli*. This plasmid contains an origin of replication (yellow section) that controls the copy number of the plasmid. The plasmid also expresses the lactose repressor protein (*lac*I) and the β-lactamase (AmpR) at normal levels. The structure of the antibiotic ampicillin is at the upper right.

Table 2.1. Plasmid families used for protein overexpression.

name	antibiotic resistance	origin of replication	promoter, inducer
pET	ampicillin or kanamycin	pMB1 ori, 15-20 copies per cell	strong, T7 or T7*lac*, IPTG
pBAD	ampicillin or kanamycin	p15A ori, 10-12 copies per cell	tunable, arabinose

2.5 Site-directed mutagenesis

The most common method of protein engineering is site-directed mutagenesis to replace an existing amino acid residue in a protein with a different amino acid that is expected to improve the protein. Site-directed mutagenesis changes the DNA sequence encoding the target protein to encode a different amino acid at the selected location. Although amino acid replacements are the most common change made to improve proteins, one can also insert additional amino acids or delete amino acids using the same methods.

As an example of site-directed mutagenesis, consider the replacement of methionine 222 with alanine to stabilize subtilisin for use in laundry detergent,[3] Fig Figure 2.7. This replacement required replacing the ATG codon for methionine with GCG to encode alanine. In contrast to methionine, alanine lacks a sulfur atom and is unaffected by bleach. Alanine is also smaller than methionine so it does not hinder access of the substrate to the active site.

Smith and coworkers first reported site-directed mutagenesis in 1978,[39] but advances in molecular biology methods have yielded improved procedures. The section below describes using an inverse polymerase chain

```
... 218 219 220 221 222 223 224 225 ... amino acid numbering
... AAC GGT ACG TCA ATG GCA TCT CCG ... original DNA sequence (forward strand)
... Asn Gly Thr Ser Met Ala Ser Pro ... original protein encoded

... AAC GGT ACG TCA gcG GCA TCT CCG ... modified DNA sequence
... Asn Gly Thr Ser Ala Ala Ser Pro ... variant protein encoded
```

Figure 2.7. Site-directed mutagenesis to replace methionine at position 222 with alanine involves replacement of two nucleotides in the codon. The original DNA sequence encoding subtilisin (top) specifies methionine at position 222 using the ATG codon. Replacement of this codon with GCG encodes a variant of subtilisin where position 222 is an alanine (bottom). The two substituted nucleotides are shown in lower case.

reaction (inverse PCR) to replace methionine 222 in subtilisin with alanine. In the original work the researchers at Genentech and Genencor used another method since their experiments were completed before the invention of the polymerase chain reaction.[3]

This site-directed mutagenesis method uses an inverse polymerase chain reaction (inverse PCR) to copy the whole plasmid using back-to-back primers,[40] Figure 2.8. The primers are designed to bind back-to-back with the DNA synthesis proceeding outward. Since the plasmid is circular, the result is a linear double stranded copy of the plasmid. Copying the whole plasmid avoids the cut and paste steps that complicated older site-directed mutagenesis methods.

The DNA primers are single-stranded 20-30 nucleotide long oligomers. The location of the primers should be such that desired mutation (typically 1-3 base changes) lies approximately in the center of one of the primers, which is the mutagenic primer. The mutagenic primer encodes the new DNA sequence; thus several nucleotides are not complementary to the plasmid DNA. One primer is complementary to one DNA strand; the other primer is complementary to the other DNA strand. The primers should have similar melting temperatures and not form dimers with themselves or each other. Figure 2.9 shows an example of primers

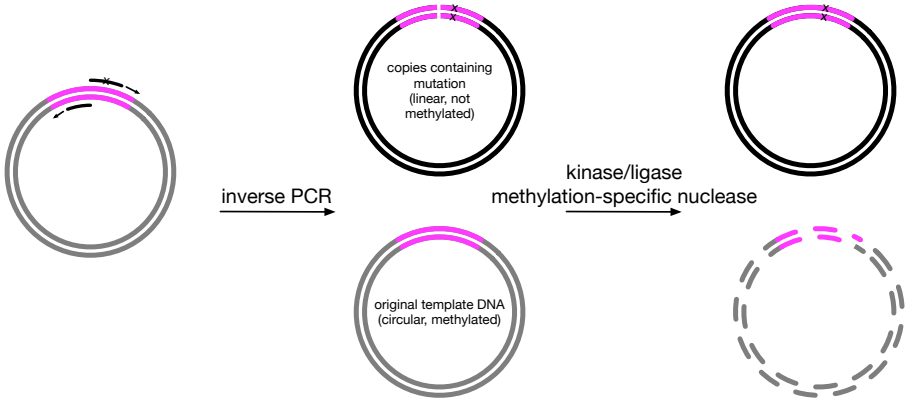


Figure 2.8. Site-directed mutagenesis involves copying the entire plasmid using mutagenic primers followed by a clean-up step. The magenta section represents the target gene for mutagenesis. Extension of two back-to-back primers (black) using the inverse polymerase chain reaction in the direction shown by the arrows yields a linear copy of the plasmid. If one of the primers contains a mismatch (x), then the copies contain this altered DNA sequence. Clean up involves circularization of the linear copies with kinase and ligase enzymes and fragmentation of the original wild type plasmid with a nuclease that only cleaves DNA containing methylation. The wild-type plasmid, prepared by growth in *E. coli*, contains methylation, but the mutated plasmid, created by the polymerase chain reaction, does not contain methylation, so the nucleases does not fragment it.

designed to replace methionine 222 in subtilisin with alanine.

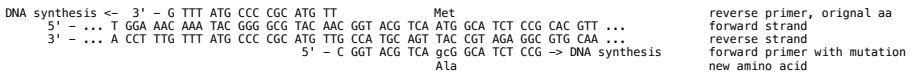


Figure 2.9. Mutagenic back-to-back primers to replace methionine 222 in subtilisin with alanine and a section of the circular plasmid DNA sequence. Top: The reverse primer (18 nt, 50% GC, Tm = 63 °C) binds to the forward strand and contains no mutations. DNA polymerase extends this primer to the left generating the complement to the forward strand. In subsequent PCR cycles, this primer extension will also create complements to the mutated forward strand. Bottom: The forward primer (22 nt, 68% GC, Tm = 61 °C) binds to the reverse strand and contains a two-nucleotide mismatch. DNA polymerase extends this primer to the right generating a complement to the reverse strand. Primers were predicted by NEBaseChanger (https://nebasechanger.neb.com).

An additional incubation with three enzymes circularizes this linear DNA fragment for insertion and replication in bacteria, Figure 2.8. Simultaneous treatment with a kinase to add a phosphoryl group to the 5'-ends of the PCR products and a ligase to join the two ends yield a circular double stranded DNA. Also in the same reaction, methylation-specific nuclease fragments the original template, which does not contain the mutation. The template DNA is the circular wild-type plasmid isolated from *E. coli*. When plasmids replicate in *E. coli*, methyl groups are added to the DNA. The nuclease recognizes and cleaves DNA containing these methyl groups. In contrast, the primers and the DNA synthesized using the polymerase chain reaction do not contain methyl groups and are ignored by this nuclease. The result is a circular copy of the plasmid only containing the altered DNA sequence. Transfer of this plasmid into *E. coli* allows researchers to grow cells, which then make the variant protein encoded.

2.6 Template-independent synthesis of DNA

The site-directed mutagenesis experiment above relies on the ability to synthesize DNA fragments (primers) with a defined sequence. DNA polymerase can synthesize DNA, but only by copying an existing template. To make primers and even entire genes, one needs a template-independent synthesis of DNA.

The first approach to template-independent DNA synthesis was chemical synthesis.[41] This synthesis is relatively complex and involves reactive intermediates and toxic solvents. Researchers normally order the desired sequence from a commercial service. The synthesis starts with a solid silica support. Reagents such as protected nucleoside phosphoramidites are washed over the silica to carry out the multiple steps needed to add each base. The harsh reaction conditions slowly damage the DNA, which caps the length of the DNA to about 200 bases. More recently, researchers have developed enzymatic approaches for the template-independent synthesis of DNA, Figure 2.10. This enzymatic approach relies on an engineered terminal deoxynucleotidyl transferase (Tdt) for the coupling step. The enzymatic approach generates less waste, can make longer oligonucleotides (>1000 bases) and can make repetitive sequences that the chemical methods cannot.[42],[43]

Entire genes and longer fragments can be synthesized by chemical or enzymatic synthesis of 300 bp DNA fragments, followed by joining them to create longer oligonucleotides.[44] The experimental procedures for the methods described in this chapter be found online or in molecular biology handbooks.[45]

Glossary

Operator is a section of DNA upstream of the a gene that binds a transcription regulator. For example, the *lac* operator binds the *lac*

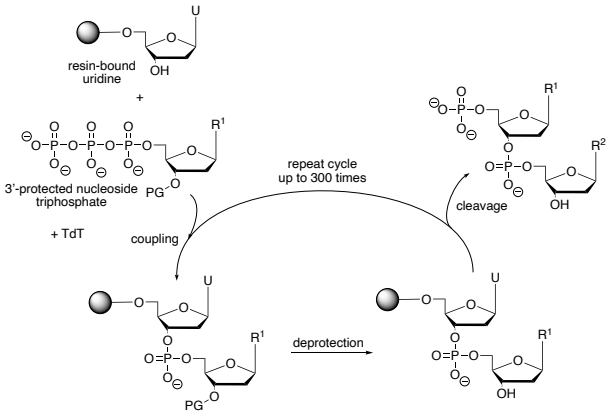


Figure 2.10. Enzymatic synthesis of DNA using 3'-protected nucleoside 5'-triphosphates. Synthesis starts at a uridine attached to a resin bead, which provides a place for terminal deoxynucleotidyl transferase (TdT) to bind and a cleavage site once synthesis is complete. TdT ligates the 3'-protected nucleoside 5'-triphosphate to the 3'-terminus of the growing oligonucleotide chain. Washing removes excess reagents and the pyrophosphate by-product of ligation. Deprotection of the 3'-PG exposed the 3'-hydroxyl for the next synthesis cycle. On completion, the oligonucleotide is cleaved from the resin bead by uracil DNA glycosylase.

repressor protein, which inhibits transcription of the downstream gene.

Promoter a region of DNA upstream of a region that encodes one or more proteins. The promoter region binds the RNA polymerase that will transcribe the DNA into mRNA.

Plasmid is a circular, independently replicating, DNA within a microbe separate from the chromosomal DNA. For example, pET plasmids are inserted into *Escherichia coli* to overexpress proteins under the control of a T7 promoter.

Primer is a short (10-50 nucleotides), single-stranded DNA fragment. When this fragment bind to a complementary region on a longer DNA fragment, then DNA polymerase can extend the primer making a complementary copy of the longer DNA fragment. If the primer contains a mutation, then the longer copy will also contain that mutation.

T7 promoter is a promoter that binds the T7 RNA polymerase, which is more efficient than the endogenous RNA polymerase in *Escherichia coli*.

References

1. Morrison, C., & Lähteenmäki, R. (2018). Public biotech in 2017-the numbers. *Nat. Biotechnol.*, *36*(7), 576–584. https://doi.org/10.1038/nbt.4175
2. Brannigan, J. A., & Wilkinson, A. J. (2002). Protein engineering 20 years on. *Nat. Rev. Mol. Cell Biol.*, *3*(12), 964–970. https://doi.org/10.1038/nrm975
3. Estell, D. A., Graycar, T. P., & Wells, J. A. (1985). Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.*, *260*(11), 6518–6521. https://doi.org/10.1016/S0021-9258(18)88811-1

4. Pace, C. N., & Hermans, J. (1975). The stability of globular proteins. *CRC Crit. Rev. Biochem.*, *3*(1), 1–43. https://doi.org/10.3109/10409237509102551
5. Taverna, D. M., & Goldstein, R. A. (2002). Why are proteins marginally stable? *Proteins*, *46*(1), 105–109. https://doi.org/10.1002/prot.10016
6. Fersht, A. R., & Serrano, L. (1993). Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.*, *3*(1), 75–83. https://doi.org/10.1016/0959-440X(93)90205-Y
7. Bloom, J. D., Lu, Z., Chen, D., Raval, A., Venturelli, O. S., & Arnold, F. H. (2007). Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.*, *5*(1), 97–21. https://doi.org/10.1186/1741-7007-5-29
8. Gordon, S. R., Stanley, E. J., Wolf, S., Toland, A., Wu, S. J., Hadidi, D., Mills, J. H., Baker, D., Pultz, I. S., & Siegel, J. B. (2012). Computational design of an α-gliadin peptidase. *J. Am. Chem. Soc.*, *134*(50), 20513–20520. https://doi.org/10.1021/ja3094795
9. Harvey, S. B., Stone, M. D., Martinez, M. B., & Nelsestuen, G. L. (2003). Mutagenesis of the γ-carboxyglutamic acid domain of human factor VII to generate maximum enhancement of the membrane contact site. *J. Biol. Chem.*, *278*(10), 8363–8369. https://doi.org/10.1074/jbc.M211629200
10. Persson, E., Olsen, O. H., Bjørn, S. E., & Ezban, M. (2012). Vatreptacog alfa from conception to clinical proof of concept. *Semin. Thromb. Hemost.*, *38*(3), 274–281. https://doi.org/10.1055/s-0032-1302442
11. Bolli, G. B., Di Marchi, R. D., Park, G. D., Pramming, S., & Koivisto, V. A. (1999). Insulin analogues and their potential in the management of diabetes mellitus. *Diabetologia*, *42*(10), 1151–1167. https://doi.org/10.1007/s001250051286
12. Ryu, J. K., Kim, H. S., & Nam, D. H. (2012). Current status and perspectives of biopharmaceutical drugs. *Biotechnol. Bioprocess Eng.*, *17*(5), 900–911. https://doi.org/10.1007/s12257-012-0095-1

13. Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B. B., Henner, D., Wong, W. L. T., Rowland, A. M., Kotts, C., Carver, M. E., & Shepard, H. M. (1992). Humanization of an anti-p185^{HER2} antibody for human cancer therapy. *Proc. Natl. Acad. Sci. U. S. A.*, *89*(10), 4285–4289. https://doi.org/10.1073/pnas.89.10.4285
14. Sarkar, C. A., Lowenhaupt, K., Horan, T., Boone, T. C., Tidor, B., & Lauffenburger, D. A. (2002). Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nat. Biotechnol.*, *20*(9), 908–913. https://doi.org/10.1038/nbt725
15. Cantor, J. R., Yoo, T. H., Dixit, A., Iverson, B. L., Forsthuber, T. G., & Georgiou, G. (2011). Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(4), 1272–1277. https://doi.org/10.1073/pnas.1014739108
16. Loosmore, S. M., Zealey, G. R., Boux, H. A., Cockle, S. A., Radika, K., Fahim, R. E., Zobrist, G. J., Yacoob, R. K., Chong, P. C., & Yao, F. L. (1990). Engineering of genetically detoxified pertussis toxin analogs for development of a recombinant whooping cough vaccine. *Infect. Immun.*, *58*(11), 3653–3662. https://doi.org/10.1128/iai.58.11.3653-3662.1990
17. Hsieh, C.-L., Goldsmith, J. A., Schaub, J. M., DiVenere, A. M., Kuo, H.-C., Javanmardi, K., Le, K. C., Wrapp, D., Lee, A. G., Liu, Y., Chou, C.-W., Byrne, P. O., Hjorth, C. K., Johnson, N. V., Ludes-Meyers, J., Nguyen, A. W., Park, J., Wang, N., Amengor, D., … McLellan, J. S. (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science*, *369*(6510), 1501–1505. https://doi.org/10.1126/science.abd0826

18. Padgette, S. R., Kolacz, K. H., Delannay, X., Re, D. B., Tinius, C. N., Rhodes, W. K., Otero, Y. I., Barry, G. F., Eichholtz, D. A., Peschke, V. M., Nida, D. L., Taylor, N. B., & Kishore, G. M. (1995). Development, identification, and characterization of a glyphosate-tolerant soybean line. *Crop Sci.*, *35*(5), 1451–1461. https://doi.org/10.2135/cropsci1995.0011183x003500050032x
19. Castle, L. A., Siehl, D. L., Gorton, R., Patten, P. A., Chen, Y. H., Bertain, S., Cho, H.-J., Duck, N., Wong, J., Liu, D., & Lassner, M. W. (2004). Discovery and directed evolution of a glyphosate tolerance gene. *Science*, *304*(5674), 1151–1154. https://doi.org/10.1126/science.1096770
20. Bravo, A., & Soberon, M. (2008). How to cope with insect resistance to Bt toxins? *Trends Biotechnol.*, *26*(10), 573–579. https://doi.org/10.1016/j.tibtech.2008.06.005
21. Kurek, I., Chang, T. K., Bertain, S. M., Madrigal, A., Liu, L., Lassner, M. W., & Zhu, G. (2007). Enhanced thermostability of *Arabidopsis* rubisco activase improves photosynthesis and growth rates under moderate heat stress. *Plant Cell*, *19*(10), 3230–3241. https://doi.org/10.1105/tpc.107.054171
22. Whittle, E., & Shanklin, J. (2001). Engineering Δ^9-16:0-acyl carrier protein (ACP) desaturase specificity based on combinatorial saturation mutagenesis and logical redesign of the castor Δ^9-18:0-ACP desaturase. *J. Biol. Chem.*, *276*(24), 21500–21505. https://doi.org/10.1074/jbc.M102129200
23. Lin, H., & Lee, Y. K. (2017). Genetic engineering of medium-chain-length fatty acid synthesis in *Dunaliella tertiolecta* for improved biodiesel production. *J. Appl. Phycol.*, *29*(6), 2811–2819. https://doi.org/10.1007/s10811-017-1210-7
24. Yao, M. Z., Zhang, Y. H., Lu, W. L., Hu, M. Q., Wang, W., & Liang, A. H. (2012). Phytases: Crystal structures, protein engineering and potential biotechnological applications. *J. Appl. Microbiol.*, *112*(1), 1–14. https://doi.org/10.1111/j.1365-2672.2011.05181.x

25. Richardson, T. H., Tan, X., Frey, G., Callen, W., Cabell, M., Lam, D., Macomber, J., Short, J. M., Robertson, D. E., & Miller, C. (2002). A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable α-amylase. *J. Biol. Chem.*, *277*(29), 26501–26507. https://doi.org/10.1074/jbc.M203183200
26. Quax, W. J., Mrabet, N. T., Luiten, R. G. M., Schuurhuizen, P. W., Stanssens, P., & Lasters, I. (1991). Enhancing the thermostability of glucose-isomerase by protein engineering. *Bio/Technology*, *9*(8), 738–742. https://doi.org/10.1038/nbt0891-738
27. Bell, E. L., Finnigan, W., France, S. P., Green, A. P., Hayes, M. A., Hepworth, L. J., Lovelock, S. L., Niikura, H., Osuna, S., Romero, E., Ryan, K. S., Turner, N. J., & Flitsch, S. L. (2021). Biocatalysis. *Nat. Rev. Methods Primers*, *1*(1), 46. https://doi.org/10.1038/s43586-021-00044-z
28. van der Wielen, L. A. M., Ottens, M., & Straathof, A. J. J. (2001). Process Technology and Process Integration in the Preparation of Penicillins. In A. Bruggink (Ed.), *Synthesis of β-Lactam antibiotics: Chemistry, biocatalysis & process integration* (pp. 150–205). Springer Netherlands.
29. Ran, N., & Frost, J. W. (2007). Directed evolution of 2-keto-3-deoxy-6-phosphogalactonate aldolase to replace 3-deoxy-D-arabino-heptulosonic acid 7-phosphate synthase. *J. Am. Chem. Soc.*, *129*(19), 6130–6139. https://doi.org/10.1021/ja067330p
30. Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., ... Newman, J. D. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, *496*(7446), 528–532. https://doi.org/10.1038/nature12051
31. Anastas, P. T., & Warner, J. C. (2000). *Green chemistry : Theory and practice*. Oxford University Press.

32. Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J., Brands, J., Devine, P. N., Huisman, G. W., & Hughes, G. J. (2010). Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, *329*(5989), 305–309. https://doi.org/10.1126/science.1188934
33. Stone, E. M., Paley, O., Hu, J., Ekerdt, B., Cheung, N.-K., & Georgiou, G. (2012). De novo engineering of a human cystathionine-γ-lyase for systemic L-methionine depletion cancer therapy. *ACS Chem. Biol.*, *7*(11), 1822–1829. https://doi.org/10.1021/cb300335j
34. Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, *27*(10), 946–950. https://doi.org/10.1038/nbt.1568
35. Jobe, A., & Bourgeois, S. (1972). *lac* Repressor-operator interaction. VI. The natural inducer of the lac operon. *J. Mol. Biol.*, *69*(3), 397–408. https://doi.org/10.1016/0022-2836(72)90253-7
36. Egel, R. (1988). The '*lac*' operon - an irrelevant paradox? *Trends Genet.*, *4*(2), 31–31. https://doi.org/10.1016/0168-9525(88)90062-5
37. Valdez-Cruz, N. A., Caspeta, L., Pérez, N. O., Ramírez, O. T., & Trujillo-Roldán, M. A. (2010). Production of recombinant proteins in *E. Coli* by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microb Cell Fact*, *9*(1), 18. https://doi.org/10.1186/1475-2859-9-18
38. Austin, H. P., Allen, M. D., Donohoe, B. S., Rorrer, N. A., Kearns, F. L., Silveira, R. L., Pollard, B. C., Dominick, G., Duman, R., El Omari, K., Mykhaylyk, V., Wagner, A., Michener, W. E., Amore, A., Skaf, M. S., Crowley, M. F., Thorne, A. W., Johnson, C. W., Woodcock, H. L., ... Beckham, G. T. (2018). Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(19), E4350–E4357. https://doi.org/10.1073/pnas.1718804115

39. Hutchison 3rd, C. A., Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., & Smith, M. (1978). Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem.*, *253*(18), 6551–6560. https://doi.org/10.1016/S0021-9258(19)46967-6
40. Silva, D., Santos, G., Barroca, M., & Collins, T. (2017). Inverse PCR for point mutation introduction. In L. Domingues (Ed.), *PCR: Methods and Protocols* (pp. 87–100). Springer New York. https://doi.org/10.1007/978-1-4939-7060-5_5
41. Roy, S., & Caruthers, M. (2013). Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. *Molecules*, *18*(11), 14268–14284. https://doi.org/10.3390/molecules181114268
42. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S., & Ryadnov, M. G. (2023). DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.*, *7*(3), 144–161. https://doi.org/10.1038/s41570-022-00456-9
43. Seydel, C. (2023). DNA writing technologies moving toward synthetic genomes. *Nat Biotechnol.*, *41*(11), 1504–1509. https://doi.org/10.1038/s41587-023-02006-0
44. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison 3rd, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318
45. Green, M. R., & Sambrook, J. (2012). *Molecular Cloning: A Laboratory Manual* (4th ed.). Cold Spring Harbor Press.
46. Ball, P. (2018). A compassionate substance. How science and myth collide in water. *Lapham's Quart.*, *11*(Summer).
47. Pace, C. N., Scholtz, J. M., & Grimsley, G. R. (2014). Forces stabilizing proteins. *FEBS Lett.*, *588*(14), 2177–2184. https://doi.org/10.1016/j.febslet.2014.05.006

48. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427
49. 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., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
50. Mura, C., McCrimmon, C. M., Vertrees, J., & Sawaya, M. R. (2010). An introduction to biomolecular graphics. *PLoS Comput. Biol.*, *6*(8), e1000918. https://doi.org/10.1371/journal.pcbi.1000918
51. Leach, A. R. (2001). *Molecular modelling: Principles and applications*. Prentice Hall.
52. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz Jr, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, *117*(19), 5179–5197. https://doi.org/10.1021/ja00124a002
53. Allinger, N. L., Zhou, X., & Bergsma, J. (1994). Molecular mechanics parameters. *J. Mol. Struct. THEOCHEM*, *312*(1), 69–83. https://doi.org/10.1016/S0166-1280(09)80008-0

54. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem\@. Theory Comput.*, *13*(6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125

55. Homeyer, N., & Gohlke, H. (2012). Free energy calculations by the molecular mechanics Poisson-Boltzmann surface area method. *Mol. Inf.*, *31*(2), 114–122. https://doi.org/10.1002/minf.201100135

56. Bae, E., & Phillips Jr, G. N. (2005). Identifying and engineering ion pairs in adenylate kinases. *J. Biol. Chem.*, *280*(35), 30943–30948. https://doi.org/10.1074/jbc.M504216200

57. Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J., & Osuna, S. (2019). Hidden conformations in *Aspergillus niger* monoamine oxidase are key for catalytic efficiency. *Angew. Chem. Int. Ed.*, *58*(10), 3097–3101. https://doi.org/10.1002/anie.201812532

58. Thomas, L. L., Christakis, T. J., & Jorgensen, W. L. (2006). Conformation of alkanes in the gas phase and pure liquids. *J. Phys. Chem. B*, *110*(42), 21198–21204. https://doi.org/10.1021/jp064811m

59. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N., & Hilvert, D. (2013). Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.*, *9*(8), 494–498. https://doi.org/10.1038/nchembio.1276

60. Crean, R. M., Gardner, J. M., & Kamerlin, S. C. L. (2020). Harnessing conformational plasticity to generate designer enzymes. *J. Am. Chem. Soc.*, *142*(26), 11324–11342. https://doi.org/10.1021/jacs.0c04924

61. Robic, S., Guzman-Casado, M., Sanchez-Ruiz, J. M., & Marqusee, S. (2003). Role of residual structure in the unfolded state of a thermophilic protein. *Proc. Natl. Acad. Sci. U. S. A.*, *100*(20), 11345–11349. https://doi.org/10.1073/pnas.1635051100

62. Rosenthal, L., Rabolt, J. F., & Hummel, J. (1982). An investigation of the conformational equilibrium of *n*-butane in a solvent using Raman spectroscopy. *J. Chem. Phys.*, *76*(2), 817–820. https://doi.org/10.1063/1.443052
63. Vázquez-Figueroa, E., Yeh, V., Broering, J. M., Chaparro-Riggers, J. F., & Bommarius, A. S. (2008). Thermostable variants constructed via the structure-guided consensus method also show increased stability in salts solutions and homogeneous aqueous-organic media. *Protein Eng. Des. Sel.*, *21*(11), 673–680. https://doi.org/10.1093/protein/gzn048
64. Mayer, S., Rüdiger, S., Ang, H. C., Joerger, A. C., & Fersht, A. R. (2007). Correlation of levels of folded recombinant P53 in *Escherichia coli* with thermodynamic stability in vitro. *J. Mol. Biol.*, *372*(1), 268–276. https://doi.org/10.1016/j.jmb.2007.06.044
65. Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, *103*(15), 5869–5874. https://doi.org/10.1073/pnas.0510098103
66. Adams, M. W., & Kelly, R. M. (1998). Finding and using hyperthermophilic enzymes. *Trends Biotechnol.*, *16*(8), 329–332. https://doi.org/10.1016/s0167-7799(98)01193-7
67. Merz, A., Yee, M.-C., Szadkowski, H., Pappenberger, G., Crameri, A., Stemmer, W. P. C., Yanofsky, C., & Kirschner, K. (2000). Improving the catalytic activity of a thermophilic enzyme at low temperatures. *Biochemistry*, *39*(5), 880–889. https://doi.org/10.1007/pl00003701
68. Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.*, *17*(4), 263–270. https://doi.org/10.1002/humu.22
69. Joerger, A. C., Ang, H. C., & Fersht, A. R. (2006). Structural basis for understanding oncogenic P53 mutations and designing rescue drugs. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(41), 15056–15061. https://doi.org/10.1073/pnas.0607286103

70. Park, S. H., & Bao, G. (2021). CRISPR/Cas9 gene editing for curing sickle cell disease. *Transfusion and Apheresis Science*, *60*(1), 103060. https://doi.org/10.1016/j.transci.2021.103060
71. Pace, C. N., & Shaw, K. L. (2000). Linear extrapolation method of analyzing solvent denaturation curves. *Proteins Struct. Funct. Bioinforma.*, *Suppl 4*, 1–7. https://doi.org/10.1002/1097-0134(2000)41:4+%3C1::aid-prot10%3E3.3.co;2-u
72. Baxa, M. C., Haddadian, E. J., Jumper, J. M., Freed, K. F., & Sosnick, T. R. (2014). Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(43), 15396–15401. https://doi.org/10.1073/pnas.1407768111
73. Horovitz, A., & Fersht, A. R. (1992). Co-operative interactions during protein folding. *J. Mol. Biol.*, *224*(3), 733–740. https://doi.org/10.1016/0022-2836(92)90557-Z
74. Scholtz, J. M., Grimsley, G. R., & Pace, C. N. (2009). Solvent denaturation of proteins and interpretations of the m value. *Methods Enzymol.*, *466*, 549–565. https://doi.org/10.1016/S0076-6879(09)66023-7
75. Fersht, A. (1999). Chapter 17, Protein Stability. In *Structure and Mechanism in Protein Science*. Freeman.
76. Becktel, W. J., & Schellman, J. A. (1987). Protein stability curves. *Biopolymers*, *26*(11), 1859–1878. https://doi.org/10.1002/bip.360261104
77. Robertson, A. D., & Murphy, K. P. (1997). Protein structure and the energetics of protein stability. *Chem. Rev.*, *97*(5), 1251–1267.
78. Pucci, F., Kwasigroch, J. M., & Rooman, M. (2017). SCooP: An accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*, *33*(21), 3415–3422.

79. Pastore, A., Martin, S. R., Politou, A., Kondapalli, K. C., Stemmler, T., & Temussi, P. A. (2007). Unbiased cold denaturation: Low- and high-temperature unfolding of yeast frataxin under physiological conditions. *J. Am. Chem. Soc.*, *129*(17), 5374–5375. https://doi.org/10.1021/ja0714538
80. Razvi, A., & Scholtz, J. M. (2006). Lessons in stability from thermophilic proteins. *Protein Sci.*, *15*(7), 1569–1578. https://doi.org/10.1110/ps.062130306
81. Eijsink, V. G. H., Bjørk, A., G\a aseidnes, S., Sirev\a ag, R., Synstad, B., van den Burg, B., & Vriend, G. (2004). Rational engineering of enzyme stability. *J. Biotechnol.*, *113*(1-3), 105–120. https://doi.org/10.1016/j.jbiotec.2004.03.026
82. Dumon, C., Varvak, A., Wall, M. A., Flint, J. E., Lewis, R. J., Lakey, J. H., Morland, C., Luginbühl, P., Healey, S., Todaro, T., DeSantis, G., Sun, M., Parra-Gessert, L., Tan, X., Weiner, D. P., & Gilbert, H. J. (2008). Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure. *J. Biol. Chem.*, *283*(33), 22557–22564. https://doi.org/10.1074/jbc.M800936200
83. Shoichet, B. K., Baase, W. A., Kuroki, R., & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.*, *92*(2), 452–456. https://doi.org/10.1073/pnas.92.2.452
84. Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L., van Loon, A. P. G. M., & Wyss, M. (2002). The consensus concept for thermostability engineering of proteins: Further proof of concept. *Protein Eng.*, *15*(5), 403–411. https://doi.org/10.1093/protein/15.5.403
85. Magliery, T. J. (2015). Protein stability: Computation, sequence statistics, and new experimental methods. *Current Opinion in Structural Biology*, *33*, 161–168. https://doi.org/10.1016/j.sbi.2015.09.002

86. Jones, B. J., Kan, C. N. E., Luo, C., & Kazlauskas, R. J. (2020). Consensus Finder web tool to predict stabilizing substitutions in proteins. In D. S. Tawfik (Ed.), *Enzyme Engineering and Evolution: General Methods* (Vol. 643, pp. 129–148). Academic Press. https://doi.org/10.1016/bs.mie.2020.07.010
87. Musil, M., Stourac, J., Bendl, J., Brezovsky, J., Prokop, Z., Zendulka, J., Martinek, T., Bednar, D., & Damborsky, J. (2017). FireProt: Web server for automated design of thermostable proteins. *Nucleic Acids Res.*, *45*(W1), W393–W399. https://doi.org/10.1093/nar/gkx285
88. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L., & Bryan, P. N. (1989). Large increases in general stability for subtilisin BPN' through incremental changes in the free-energy of unfolding. *Biochemistry*, *28*(18), 7205–7213. https://doi.org/10.1021/bi00444a012
89. Schwehm, J. M., Fitch, C. A., Dang, B. N., García-Moreno E., B., & Stites, W. E. (2003). Changes in stability upon charge reversal and neutralization substitution in staphylococcal nuclease are dominated by favorable electrostatic effects. *Biochemistry*, *42*(4), 1118–1128. https://doi.org/10.1021/bi0266434
90. Trudeau, D. L., Kaltenbach, M., & Tawfik, D. S. (2016). On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.*, *33*(10), 2633–2641. https://doi.org/10.1093/molbev/msw138
91. Clarke, J., Henrick, K., & Fersht, A. R. (1995). Disulfide mutants of barnase I: Changes in stability and structure assessed by biophysical methods and x-ray crystallography. *J. Mol. Biol.*, *253*(3), 493–504. https://doi.org/10.1006/jmbi.1995.0568
92. Pace, C. N., Grimsley, G. R., Thomson, J. A., & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.*, *263*(24), 11820–11825.

93. Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.*, *151*(2), 261–287. https://doi.org/10.1016/0022-2836(81)90515-5
94. Perry, L. J., & Wetzel, R. (1984). Disulfide bond engineered into T4 lysozyme: Stabilization of the protein toward thermal inactivation. *Science*, *226*(4674), 555–557. https://www.jstor.org/stable/1693872
95. Matsumura, M., Becktel, W. J., Levitt, M., & Matthews, B. W. (1989). Stabilization of phage T4 Iysozyme by engineered disulfide bonds. *Proc. Natl. Acad. Sci. U. S. A.*, *86*, 6562–6566. https://doi.org/10.1073/pnas.86.17.6562
96. Roesler, K. R., & Rao, A. G. (2000). A single disulfide bond restores thermodynamic and proteolytic stability to an extensively mutated protein. *Protein Science*, *9*(9), 1642–1650. https://doi.org/10.1110/ps.9.9.1642
97. Kanaya, S., Katsuda, C., Kimura, S., Nakai, T., Kitakuni, E., Nakamura, H., Katayanagi, K., Morikawa, K., & Ikehara, M. (1991). Stabilization of *Escherichia coli* ribonuclease H by introduction of an artificial disulfide bond. *Journal of Biological Chemistry*, *266*(10), 6038–6044. https://doi.org/10.1016/S0021-9258(18)38080-3
98. Hazes, B., & Dijkstra, B. W. (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng. Des. Sel.*, *2*(2), 119–125. https://doi.org/10.1093/protein/2.2.119
99. Dombkowski, A. A. (2003). Disulfide by Design^{TM}: A computational method for the rational design of disulfide bonds in proteins. *Bioinformatics*, *19*(14), 1852–1853. https://doi.org/10.1093/bioinformatics/btg231
100. Suplatov, D., Timonina, D., Sharapova, Y., & Švedas, V. (2019). Yosshi: A web-server for disulfide engineering by bioinformatic analysis of diverse protein families. *Nucleic Acids Res.*, *47*(W1), W308–W314. https://doi.org/10.1093/nar/gkz385

101. Pikkemaat, M. G., Linssen, A. B. M., Berendsen, H. J. C., & Janssen, D. B. (2002). Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng.*, *15*(3), 185–192. https://doi.org/10.1093/protein/15.3.185
102. Stewart, E. J., Åslund, F., & Beckwith, J. (1998). Disulfide bond formation in the *Escherichia coli* cytoplasm: An *in Vivo* role reversal for the thioredoxins. *EMBO J.*, *17*(19), 5543–5550. https://doi.org/10.1093/emboj/17.19.5543
103. Zhou, H.-X. (2004). Loops, linkages, rings, catenanes, cages, and crowders: Entropy-based strategies for stabilizing proteins. *Acc. Chem. Res.*, *37*(2), 123–130. https://doi.org/10.1021/ar0302282
104. Watanabe, K., & Suzuki, Y. (1998). Protein thermostabilization by proline substitutions. *J. Mol. Catal. B Enzym.*, *4*(4), 167–180. https://doi.org/10.1016/s1381-1177(97)00031-3
105. Fu, H., Grimsley, G. R., Razvi, A., Scholtz, J. M., & Pace, C. N. (2009). Increasing protein stability by improving beta-turns. *Proteins*, *77*(3), 491–498. https://doi.org/10.1002/prot.22509
106. Matthews, B. W., Nicholson, H., & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U. S. A.*, *84*(19), 6663–6667. https://doi.org/10.1073/pnas.84.19.6663
107. Scott, K. A., Alonso, D. O. V., Sato, S., Fersht, A. R., & Daggett, V. (2007). Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(8), 2661–2666. https://doi.org/10.1073/pnas.0611182104
108. Liu, Z., Lemmonds, S., Huang, J., Tyagi, M., Hong, L., & Jain, N. (2018). Entropic contribution to enhanced thermal stability in the thermostable P450 CYP119. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(43), E10049–E10058. https://doi.org/10.1073/pnas.1807473115
109. 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*(46), 7745–7751. https://doi.org/10.1002/anie.200602795

110. Floor, R. J., Wijma, H. J., Colpa, D. I., Ramos-Silva, A., Jekel, P. A., Szymanski, W., Feringa, B. L., Marrink, S. J., & Janssen, D. B. (2014). Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem*, *15*(11), 1660–1672. https://doi.org/10.1002/cbic.201402128

111. Xiao, Z., Bergeron, H., Grosse, S., Beauchemin, M., Garron, M.-L., Shaya, D., Sulea, T., Cygler, M., & Lau, P. C. K. (2008). Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-temperature-guided sequence alignment. *Appl. Environ. Microbiol.*, *74*(4), 1183–1189. https://doi.org/10.1128/AEM.02220-07

112. Petsko, G. A. (2001). Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat." *Methods Enzymol.*, *334*, 469–478. https://doi.org/10.1016/S0076-6879(01)34486-5

113. Gribenko, A. V., Patel, M., Liu, J., McCallum, S., Wang, C., & Makhatadze, G. I. (2009). Rational stabilization of enzymes by computational redesign of surface charge–charge interactions. *Proc. Natl. Acad. Sci. U. S. A.*, *106*(8), 2601–2606. https://doi.org/10.1073/pnas.0808220106

114. Contessoto, V. G., de Oliveira, V. M., Fernandes, B. R., Slade, G. G., & Leite, V. B. P. (2018). TKSA-MC: A web server for rational mutation through the optimization of protein charge interactions. *Proteins*, *86*(11), 1184–1188. https://doi.org/10.1002/prot.25599

115. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987–2998. https://doi.org/10.1021/bi902153g

116. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Res.*, *33*(Web Server issue), W382–8. https://doi.org/10.1093/nar/gki387

117. Korkegian, A., Black, M. E., Baker, D., & Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. *Sci. Wash.*, *308*(5723), 857–860. https://doi.org/10.1126/science.1107387
118. Arabnejad, H., Dal Lago, M., Jekel, P. A., Floor, R. J., Thunnissen, A.-M. W. H., Terwisscha van Scheltinga, A. C., Wijma, H. J., & Janssen, D. B. (2017). A robust cosolvent-compatible halohydrin dehalogenase by computational library design. *Protein Eng. Des. Sel.*, *30*(3), 173–187. https://doi.org/10.1093/protein/gzw068
119. Jin, Y., Johannissen, L. O., & Hay, S. (2021). Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins*, *89*(8), 915–921. https://doi.org/10.1002/prot.26068
120. Weinstein, J. J., Goldenzweig, A., Hoch, S., & Fleishman, S. J. (2021). PROSS 2: A new server for the design of stable and highly expressed protein variants. *Bioinformatics*, *37*(1), 123–125. https://doi.org/10.1093/bioinformatics/btaa1071
121. Gao, D., Narasimhan, D. L., Macdonald, J., Brim, R., Ko, M.-C., Landry, D. W., Woods, J. H., Sunahara, R. K., & Zhan, C.-G. (2009). Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, *75*(2), 318–323. https://doi.org/10.1124/mol.108.049486
122. van den Burg, B., Vriend, G., Veltman, O. R., & Eijsink, V. G. H. (1998). Engineering an enzyme to resist boiling. *Proc. Natl. Acad. Sci. U. S. A.*, *95*(5), 2056–2060. https://doi.org/10.1073/pnas.95.5.2056
123. Bjørk, A., Dalhus, B., Mantzilas, D., Sirev\a ag, R., & Eijsink, V. G. H. (2004). Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface. *J. Mol. Biol.*, *341*(5), 1215–1226. https://doi.org/10.1016/j.jmb.2004.06.079

124. Kabashima, T., Li, Y., Kanada, N., Ito, K., & Yoshimoto, T. (2001). Enhancement of the thermal stability of pyroglutamyl peptidase I by introduction of an intersubunit disulfide bond. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1547*(2), 214–220. https://doi.org/10.1016/s0167-4838(01)00185-6
125. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., & Trout, B. L. (2010). Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B*, *114*(19), 6614–6624. https://doi.org/10.1021/jp911706q
126. Kuriata, A., Iglesias, V., Pujols, J., Kurcinski, M., Kmiecik, S., & Ventura, S. (2019). Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. *Nucleic Acids Res.*, *47*(W1), W300–W307. https://doi.org/10.1093/nar/gkz321
127. Van Durme, J., De Baets, G., Van Der Kant, R., Ramakers, M., Ganesan, A., Wilkinson, H., Gallardo, R., Rousseau, F., & Schymkowitz, J. (2016). Solubis: A webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel.*, *29*(8), 285–289. https://doi.org/10.1093/protein/gzw019
128. Suen, W.-C., Zhang, N., Xiao, L., Madison, V., & Zaks, A. (2004). Improved activity and thermostability of *Candida antarctica* lipase B by DNA family shuffling. *Protein Eng. Des. Sel.*, *17*(2), 133–140. https://doi.org/10.1093/protein/gzh017
129. Lawrence, M. S., Phillips, K. J., & Liu, D. R. (2007). Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.*, *129*(33), 10110–10112. https://doi.org/10.1021/ja071641y
130. Robinson, N. E. (2002). Protein deamidation. *Proc. Natl. Acad. Sci. U. S. A.*, *99*(8), 5283–5288. https://doi.org/10.1073/pnas.082102799
131. Mark, D. F., Lu, S. D., Creasey, A. A., Yamamoto, R., & Lin, L. S. (1984). Site-specific mutagenesis of the human fibroblast interferon gene. *Proc. Natl. Acad. Sci. U. S. A.*, *81*(18), 5662–5666.

132. Kontermann, R. E. (2009). Strategies to extend plasma half-lives of recombinant antibodies. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *23*(2), 93–109. https://doi.org/10.2165/00063030-200923020-00003
133. Strohl, W. R. (2015). Fusion proteins for half-life extension of biologics as a strategy to make biobetters. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *29*(4), 215–239. https://doi.org/10.1007/s40259-015-0133-6
134. Cheang, J. Y., & Moyle, P. M. (2018). Glucagon-like peptide-1 (GLP-1)-based therapeutics: Current status and future opportunities beyond type 2 diabetes. *ChemMedChem*, *13*(7), 662–671. https://doi.org/10.1002/cmdc.201700781
135. Feldmann, M. (2002). Development of anti-TNF therapy for rheumatoid arthritis. *Nat. Rev. Immunol.*, *2*(5), 364–371. https://doi.org/10.1038/nri802
136. Nielsen, J. E., & Borchert, T. V. (2000). Protein engineering of bacterial α-amylases. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1543*(2), 253–274. https://doi.org/10.1016/s0167-4838(00)00240-5
137. Jones, B. J., Lim, H. Y., Huang, J., & Kazlauskas, R. J. (2017). Comparison of five protein engineering strategies for stabilizing an α/β-hydrolase. *Biochemistry*, *56*(50), 6521–6532. https://doi.org/10.1021/acs.biochem.7b00571
138. Thieker, D. F., Maguire, J. B., Kudlacek, S. T., Leaver-Fay, A., Lyskov, S., & Kuhlman, B. (2022). Stabilizing proteins, simplified: A Rosetta-based webtool for predicting favorable mutations. *Protein Sci.*, *31*(10). https://doi.org/10.1002/pro.4428
139. Banta, S., Dooley, K., & Shur, O. (2013). Replacing antibodies: Engineering new binding proteins. *Annu. Rev. Biomed. Eng.*, *15*(1), 93–113. https://doi.org/10.1146/annurev-bioeng-071812-152412
140. Rossi, A. M., & Taylor, C. W. (2011). Analysis of protein-ligand interactions by fluorescence polarization. *Nat Protoc*, *6*(3), 365–387. https://doi.org/10.1038/nprot.2011.305

141. Pollard, T. D. (2010). MBOC technical perspective: A guide to simple and informative binding assays. *Mol. Biol. Cell*, *21*(23), 4061–4067. https://doi.org/10.1091/mbc.E10-08-0683
142. Jeffrey, P. D., Strong, R. K., Sieker, L. C., Chang, C. Y. Y., Campbell, R. L., Petsko, G. A., Haber, E., Margolies, M. N., & Sheriff, S. (1993). 26-10 Fab-digoxin complex: Affinity and specificity due to surface complementarity. *Proc. Natl. Acad. Sci. U. S. A.*, *90*(21), 10310–10314. https://doi.org/10.1073/pnas.90.21.10310
143. Tian, W., Chen, C., Lei, X., Zhao, J., & Liang, J. (2018). CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.*, *46*(W1), W363–W367. https://doi.org/10.1093/nar/gky473
144. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, *255*(5041), 178–183. https://doi.org/10.1126/science.1553543
145. Jones, B. J., Evans, R. L., Mylrea, N. J., Chaudhury, D., Luo, C., Guan, B., Pierce, C. T., Gordon, W. R., Wilmot, C. M., & Kazlauskas, R. J. (2020). Larger active site in an ancestral hydroxynitrile lyase increases catalytically promiscuous esterase activity. *PLoS ONE*, *15*, e0235341. https://doi.org/10.1371/journal.pone.0235341
146. DeLorbe, J. E., Clements, J. H., Teresk, M. G., Benfield, A. P., Plake, H. R., Millspaugh, L. E., & Martin, S. F. (2009). Thermodynamic and structural effects of conformational constraints in protein-ligand interactions. Entropic paradoxy associated with ligand preorganization. *J. Am. Chem. Soc.*, *131*(46), 16758–16770. https://doi.org/10.1021/ja904698q
147. Barros, E. P., Schiffer, J. M., Vorobieva, A., Dou, J., Baker, D., & Amaro, R. E. (2019). Improving the efficiency of ligand-binding protein design with molecular dynamics simulations. *J. Chem. Theory Comput.*, *15*(10), 5703–5715. https://doi.org/10.1021/acs.jctc.9b00483

148. Thakkar, S., Nanaware-Kharade, N., Celikel, R., Peterson, E. C., & Varughese, K. I. (2014). Affinity improvement of a therapeutic antibody to methamphetamine and amphetamine through structure-based antibody engineering. *Sci. Rep.*, *4*, 3673. https://doi.org/10.1038/srep03673
149. Zhao, Q., Ahmed, M., Guo, H., Cheung, I. Y., & Cheung, N.-K. V. (2015). Alteration of electrostatic surface potential enhances affinity and tumor killing properties of anti-ganglioside GD2 monoclonal antibody hu3F8. *J. Biol. Chem.*, *290*(21), 13017–13027. https://doi.org/10.1074/jbc.M115.650903
150. Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, *25*(10), 1171–1176. https://doi.org/10.1038/nbt1336
151. Selzer, T., Albeck, S., & Schreiber, G. (2000). Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.*, *7*(7), 537–541. https://doi.org/10.1038/76744
152. Bugnon, M., Röhrig, U. F., Goullieux, M., Perez, M. A. S., Daina, A., Michielin, O., & Zoete, V. (2024). SwissDock 2024: Major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. *Nucleic Acids Res*, *52*(W1), W324–W332. https://doi.org/10.1093/nar/gkae300
153. Zoete, V., Schuepbach, T., Bovigny, C., Chaskar, P., Daina, A., Röhrig, U. F., & Michielin, O. (2016). Attracting cavities for docking. Replacing the rough energy landscape of the protein by a smooth attracting landscape. *J Comput Chem*, *37*(4), 437–447. https://doi.org/10.1002/jcc.24249
154. Röhrig, U. F., Goullieux, M., Bugnon, M., & Zoete, V. (2023). Attracting Cavities 2.0: Improving the flexibility and robustness for small-molecule docking. *J. Chem. Inf. Model.*, *63*(12), 3925–3940. https://doi.org/10.1021/acs.jcim.3c00054

155. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., & Baker, D. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, *501*(7466), 212–216. https://doi.org/10.1038/nature12443

156. Morin, A., Kaufmann, K. W., Fortenberry, C., Harp, J. M., Mizoue, L. S., & Meiler, J. (2011). Computational design of an endo-1,4-beta-xylanase ligand binding site. *Protein Eng. Des. Sel.*, *24*(6), 503–516. https://doi.org/10.1093/protein/gzr006

Problems

1. Identify whether the items below are proteins, oligonucleotides or small organic molecules. Explain their role in protein overexpression.

*lac*I gene product, T7 RNA polymerase, *lac* operator, ampicillin, IPTG, ribosome binding site, stop codon, T7 terminator, His-tag, mutagenic PCR primer, transcription start site

2. a) Why does DNA transcription stop after the terminator, but mRNA translation stop at the stop codon? b) What problems could arise when using a T7–lac promoter on a high–copy number plasmid? c) Why do researchers put the gene encoding T7 RNA polymerase on the genome instead of on the expression plasmid?

3. *Plain text files*. In computing, plain text files are files that contain only text characters including white spaces and punctuation. For example, a plain text file could contain the following FASTA-formatted amino acid sequence for subtilisin:

```
>tr|M5B284|M5B284_BACIU Subtilisin OS=Bacillus subtilis OX=1423 GN=aprE
MRNKKLWISLLFALTLIFTMAFSNMSAQAAGKSSTEKKYIVGFKQTMSAMSSAKKKDVIS
EKGGKVQKQFKYVNAATATLDEKAVKELKQDPSVAYVEEDHIAHEYAQSVPYGISQIKAP
```

```
ALHSQGYTGSNVKVAVIDSGIDSSHPDLNVRGGASFVPSETNPYQDGSSHGTHVAGTIAA
LNNSIGVLGVAPSASLYAVKVLDSTGSGQYSWIINGIEWAISNNMDVINMSLGGPSGSTA
LKTVVDKAVSSGIVVAAAAGNEGSSGSSSTVGYPAKYPSTIAVGAVNSSNQRASFSSAGS
ELDVMAPGVSIQSTLPGGTYGAYNGTSMATPHVAGAAALILSKHPTWTNAQVRDRLESTA
TYLGSSFYYGKGLINVEAAAQ
```

In contrast, formatted text files include additional characters within the text to indicate font, bold, line spacing, and other information. For example, and indicate bold in the example below. The program that you use to open this file must know that these formatting characters are not part of the text, but indicate formatting commands. If not, it may misinterpret these characters as data. The same subtilisin sequence with HTML bold tags added to the first line:

```
>tr|M5B284|M5B284_BACIU Subtilisin OS=Bacillus subtilis OX=1423 GN=aprE
<b>MRNKKLWISLLFALTLIFTMAFSNMSAQAAGKSSTEKKYIVGFKQTMSAMSSAKKKDVIS</b>
EKGGKVQKQFKYVNAATATLDEKAVKELKQDPSVAYVEEDHIAHEYAQSVPYGISQIKAP
ALHSQGYTGSNVKVAVIDSGIDSSHPDLNVRGGASFVPSETNPYQDGSSHGTHVAGTIAA
LNNSIGVLGVAPSASLYAVKVLDSTGSGQYSWIINGIEWAISNNMDVINMSLGGPSGSTA
LKTVVDKAVSSGIVVAAAAGNEGSSGSSSTVGYPAKYPSTIAVGAVNSSNQRASFSSAGS
ELDVMAPGVSIQSTLPGGTYGAYNGTSMATPHVAGAAALILSKHPTWTNAQVRDRLESTA
TYLGSSFYYGKGLINVEAAAQ
```

Bioinformatics and protein engineering often use plain text files to store information and commands. You can open plain text files using word processing programs like Word or Pages, but to keep them as plain text files, that is, not add formatting text to them, be sure to use a plain text format when you save them (usually 'txt' extension) and not a word processing format (e.g., 'pages' or 'docx' extension). More commonly researchers use text editor programs, which are programs intended to manipulate

plain text files: Mac: TextEdit,* BBEdit; Windows: Notepad, Textpad; Linux: pico, vim, nano). The next question requires you to work with a plain text file similar to the example above, which is a FASTA format file. FASTA format files are plain text files used to store DNA or amino acid sequences.

Practice: The `<b>` and `</b>` indicate bold in web pages, which use 'html' format files. Copy the second amino acid sequence text above into a text editor, save it with the extension 'html,' then open that file with a web browser. The `<b>` and `</b>` characters should be hidden and the characters in between should be displayed as bold. Next, change the extension of this file from 'html' to 'txt' and open with the web browser again. The browser now displays all the characters in the file, but does not show anything in bold. A web browser can read both plain text files and web pages, which have the html format. The file extension tells the browser whether to interpret the characters as text or as formatting information. Most bioinformatics and computational chemistry programs expect text files as input. If the file contains formatting or other codes, they will change the meaning of the file and the program will not accept them.

4. PCR exponentially amplifies DNA in a sample through repeated cycles of denaturation, primer annealing, and extension. In site-directed mutagenesis using PCR with a mutagenic primer, almost all newly synthesized DNA strands contain the mutation because they either incorporate the mutagenic primer sequence or copy a DNA strand that already contains the mutation (Figure 2.11). In the first cycle, extension from the mutagenic primer yields a daughter strand containing the mutation, while extension from the complementary primer produces a strand without the mutation. In subsequent cycles, the mutagenic primer can anneal to and extend from both wild-type and already-mutated template strands.

*Unfortunately, the default settings use rich text formatting which includes formatting characters. Change the default by selecting Format → Make Plain Text option in the menu bar. This change will also set the extension of files saved in the future to '.txt' from the default '.rtf'.

Non-mutated strands increase only by copying the original wild-type template using the non-mutagenic primer. Each cycle increases the proportion of mutated strands. After approximately 25–35 cycles, nearly every amplified DNA molecule incorporates the mutagenic primer sequence and therefore contains the mutation. The original wild-type template remains as a small fraction of the total (it is not amplified, only diluted by the exponentially increasing mutated products).

As an exercise, show that the 32 strands after the next cycle (including both template and synthesized strands) would consist of 26 strands with the mutation and 6 strands without the mutation. Further show that 11 of the resulting 16 double-stranded DNA molecules would contain mutations on both strands.

5. *Site-directed mutagenesis.* Gluten, a protein found in wheat, rye, and barley, contains a high proportion of proline and glutamine. Digestive proteases cleave gluten to oligopeptides enriched in the dipeptide sequence proline-glutamine because peptides containing this sequence are resistant to digestive proteases. In celiac disease patients, these oligopeptides are believed to trigger an autoimmune response, which causes diarrhea and other gastrointestinal problems. One potential solution is adding a protease that can degrade Pro-Gln containing peptide to the meals of celiac patients. Gordon and coworkers describe protein engineering to create such a protease, which they called KumaMAX.[†] You do not need to read that paper to answer this question. Some of the authors of that paper formed the company PvP Biologics to further improve KumaMAX. Recently this improved protease passed Phase I clinical trials and in 2020 Takeda Pharmaceutical Company acquired PvP Biologics. In the future (after 2025), it may be a treatment for celiac disease.

KumaMAX contains seven amino acid substitutions as compared to KumaWT. This question focuses on the Gly319Ser substitution, which con-

[†]Non-linear least squares fitting of eq. Equation 6.5 to the data starts with an initial guess for the dissociation constrant followed by iteration to find the best value. The supporting information includes a Python function for this fit.

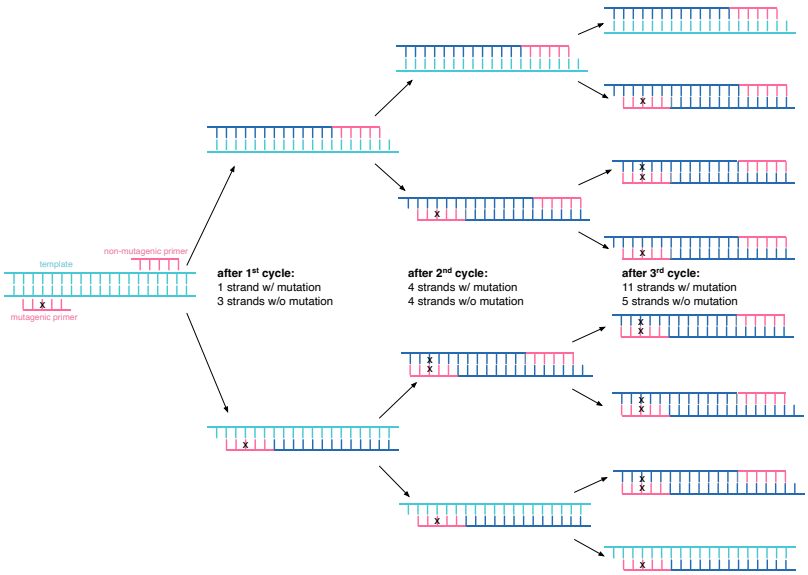


Figure 2.11. Site-directed mutagenesis with the polymerase chain reaction using a mutagenic primer. The mutation is marked by an 'x.' Among the PCR-synthesized DNA strands (dark blue), mutation-containing strands increase exponentially (from 1 to 4 to 11), while non-mutated strands increase linearly (from 1 to 2 to 3). The exponential increase in mutated strands results from extending the mutagenic primer and copying existing mutated strands by extensions with the non-mutagenic primer. Non-mutated strands increase linearly because they arise only from copying one of the template strands. The exponential accumulation of mutation-containing strands dilutes the non-mutated strands to insignificance after the typical 30 cycles of PCR.

tributes to the switch in selectivity from Arg/Lys to Gln by decreasing the amount of space available and by adding a hydrogen bond partner for Gln.

a) *Find the DNA sequence for KumaWT in GenBank.* GenBank is found at https://www.ncbi.nlm.nih.gov/nuccore/. Search for kumamolisin; the sequence should be the first hit. The enzyme is synthesized as a 552 amino acid residue precursor, which corresponds to $552 \cdot 3 = 1656$ nucleotides. The DNA sequence contains 5841 nucleotides, so we must find the section that codes for the protein. In the search results, click on the text 'GenBank' below the name of the gene. This page contains the DNA sequence and information about the DNA sequence. In the 'FEATURES' section, it mentions that the coding region (CDS) corresponds to nucleotides 2940 to 4598. How many nucleotides does that correspond to? (Your answer should be a number.) Why does that number not match $552 \cdot 3$? *Hint:* The length of an amino acid or nucleotide sequence is last_number – first_number + 1. The '+ 1' is needed to include the starting amino acid or nucleotide. For example, a sequence of five nucleotides numbered 1 through 5 has a length of 5 - 1 + 1 = 5.

b) *Save the coding region as a FASTA-formatted file on your computer.* Click on the text 'CDS' in the features section to show the DNA sequence with the coding region highlighted. On this page, click on 'FASTA' at the bottom right (near the pop-up with the translation to amino acids) to view only the coding region in FASTA format. FASTA-formatted files are plain text files, often named using .txt or .fasta as the extension, but this extension is not required. The FASTA format requires at least two lines. The first line starts with '>' (greater than symbol) and contains a description of the sequence. This line is followed by one or more lines containing single-letter codes for the amino acid or nucleotide sequence. Copy and paste the FASTA file into an empty text file on your computer. Save the file with the name 'KumaWT.fasta'. (Do not copy the entire

page; only the FASTA part.) Your file should contain about 1660 nucleotides, not about 5800. It must be a text file.

c) Design primers for site directed mutagenesis of Gly319 to Ser in KumaWT using the web tool NEBaseChanger (https://nebasechanger.neb.com). Upload the KumaWT.fasta file in the 'enter your sequence' section and G319S in the 'select input mode and set mutations' section. Keep the other entries at default values, then click 'build primers'. What is the DNA sequence that codes for Gly318Gly319Pro320? What will be the new sequence that encodes Gly318Ser319Pro320?

d) Sketch a diagram of how the primers will bind to a circular plasmid containing this gene; include the sequence of the primers in your diagram oriented correctly and indicate the direction of DNA synthesis.

Answers

1. oligonucleotides: ribosome binding site (RNA), stop codon (RNA), *lac* operator (DNA), T7 terminator (DNA), transcription start site (DNA)

proteins: *lac*I gene product, T7 RNA polymerase, His-tag (part of a protein)

small molecules: ampicillin, IPTG

2. a) In bacteria, transcription termination requires that the RNA polymerase transcribe through the terminator sequence to produce the RNA hairpin and poly-U tract. This RNA secondary structure destabilizes the RNA–DNA hybrid and causes polymerase dissociation. Because the RNA structure only forms after the terminator has been transcribed, transcription stops after the terminator. By contrast, in translation, a stop codon is not decoded into an amino acid but instead recruits release factors that

trigger termination. Thus translation ends at the stop codon. b) The lac operator is leaky, so some transcription occurs even without IPTG induction. On a high–copy plasmid, many T7 promoters combined with leaky expression can produce enough toxic or burdensome protein to overwhelm host metabolism and inhibit growth. c) Placing the T7 RNA polymerase gene in the chromosome limits its copy number to one (two during replication). If it were carried on a plasmid, the multiple copies would amplify leaky expression from the T7–lac promoter, leading to uncontrolled polymerase production, runaway target gene expression, and severe metabolic stress.

5. a) 4598 - 2940 + 1 = 1659. The extra three nucleotides in the DNA sequence encode the stop codon.

b) text file containing the text below

```
>AB070740.2:2940-4598 Bacillus sp. MN-32 kscp gene for kuma ...
ATGAGCGACATGGAGAAGCCTTGGAAGGAAGAGGAGAAGCGCGAGGTCCTCGCGGGACACGCTCGCAGGC
AGGCGCCGCAGGCTGTCGATAAGGGACCGGTGACCGGGGACCAGCGGATTTCCGTTACGGTCGTGCTGCG
CCGCCAACGAGGCGATGAACTCGAGGCGCATGTGGAACGCCAGGCCGCGCTCGCACCTCACGCGCGAGTG
CATCTGGAGCGAGAAGCGTTTGCCGCTTCGCACGGCGCTTCGCTCGACGACTTTGCGGAGATTCGAAAGT
TCGCGGAAGCGCACGGGCTCACGCTCGATCGCGCCCACGTGGCTGCGGGCACCGCGGTGCTGAGCGGCCC
GGTGGACGCCGTTAACCAAGCGTTTGGGGTCGAGTTGCGCCATTTCGATCATCCAGACGGATCCTATCGA
AGCTACGTCGGCGACGTGCGTGTGCCGGCTTCCATCGCGCCTCTGATTGAAGCGGTGTTTGGCCTGGACA
CGCGCCCGGTGGCGCGGCCCCACTTTCGGCTGCGGAGGCGCGCCGAGGGCGAGTTTGAGGCGAGATCGCA
GTCCGCGGCGCCGACTGCGTACACGCCGCTCGACGTCGCGCAGGCGTACCAATTTCCCGAGGGGCTCGAC
```

plus 15 more lines

c)

```
GGTGGCCCT and GGTagcCCT
GlyGlyPro      GlySerPro
```

d) answer should also show a circular diagram.

```
<-- cggcaggtagcacaggta
5'- GCCGTCCATCGTGTCCATCAGCTGGGGTGGCCCTGAGGACA
3'- CGGCAGGTAGCACAGGTAGTCGACCCCACCGGGACTCCTGT
                     cagctggggtagccctgaggaca -->
```

3 The logic of protein engineering

 Last revised: March 2026.

Summary. Gibbs energy diagrams are the conceptual tool that connects the goals of protein engineering to the changes in amino acid sequence. Gibbs energy diagrams show the states available to a protein, the relative energies of these states and barriers between them. Protein states (macrostates) are different protein forms, for example, folded and unfolded protein states. Protein states differ in their flexibility and in their non-covalent interactions within the protein and between protein, solvent and any ligands. These differences in molecular interactions in different protein states create Gibbs energy differences between them. The Gibbs energy differences between protein states determine the properties of proteins, including those usually targeted for protein engineering: stability, binding, reactivity and selectivity. Protein engineering works by selecting amino acid replacements that alter the relative Gibbs energy of protein states.

Key learning goals

- Gibbs energy diagrams show the options available to a protein including the available states, their relative energies and the barriers between the states.
- Protein states (macrostates) are protein forms that have macroscopic or bulk properties that one can measure. For example, the folded and unfolded protein states differ in their fluorescence properties.

- Protein states consist of countless numbers of microstates, which are individual conformations.
- Protein states differ in Gibbs energy due to differences in non-covalent interactions (electrostatic and van der Waals) and in entropy. Amino acids substitutions alter these interactions.
- The Gibbs energy difference between various states determine a protein's properties. The Gibbs energy difference between folded and unfolded protein, ΔG_{unfold}, determines protein stability. The Gibbs energy difference between bound protein and target and free protein and target, ΔG_{diss}, determine binding strength. The Gibbs energy difference between the substrate state and the transition state for the reaction, $\Delta G^{\ddagger}$, determine the reaction rate.
- Replacing amino acid residues in a protein changes the relative energies of the states and therefore the properties of the protein. This change in Gibbs energy of the protein states is the basis of protein engineering.

3.1 Gibbs energy diagrams

Protein states. Proteins exist in different forms called macrostates or more commonly states. Protein states differ in their non-covalent interactions and flexibility, but not in their covalent structure. For example, the folded state of an enzyme is the native, catalytically active form and the unfolded state is the denatured, catalytically inactive form. States refer to protein forms that have macroscopic or bulk properties that one can measure. For example, the folded and unfolded states of a protein have different fluorescence properties and migrate differently in a gel electrophoresis experiment.

Protein states can interconvert with each other. For example, a solution may contain interconverting folded and unfolded states, Equation 3.1.

The relative amounts of each state corresponds to the equilibrium constant for the interconversion. This balance between the folded and unfolded states determines protein stability.

$$folded \overset{K_{unfold}}{\rightleftharpoons} unfolded \quad K_{unfold} = \frac{[unfolded]}{[folded]} \tag{3.1}$$

Gibbs energy diagrams. Gibbs energy diagrams show the options available to a protein. The diagram shows the available states, their relative energies and the barriers between the states, Figure 3.1. The relative Gibbs energies of the states determine the relative amounts of protein in each state; in other words, the relative Gibbs energies determine the equilibrium constant between the two states, Equation 5.1. The difference in Gibbs energy, ΔG, between two states is proportional to the natural logarithm of the equilibrium constant between the two states. The constant R is the gas constant, which corresponds to 1.987 cal/mol·K when ΔG has the units of calories and to 8.314 J/mol·K when ΔG has the units of joules. The constant R connects the temperature scale to the units of energy used in physics (calories or joules). R is a molar quantity suitable for chemistry calculations, while Boltzmann's constant is used when working with particles. T represents the temperature in degrees Kelvin.

$$\Delta G = -RT \ln K_{eq} \tag{3.2}$$

If the reaction is favorable, then the equilibrium constant is >1, the natural logarithm of the equilibrium constant is positive, and the Gibbs energy change is negative. For example, in Figure 3.1, state 1 could represent the denatured state and state 2 could represent the native state. The native state is more stable, so most of the protein exists in that form. The equilbrium constant for unfolding is unfavorable (<1) and the Gibbs energy change for unfolding is positive. At room temperature (298 K) an equilibrium constant of 10 corresponds to a 1.36 kcal/mol Gibbs energy difference. An equilibrium constant of 100 corresponds to 2.73 kcal/mol,

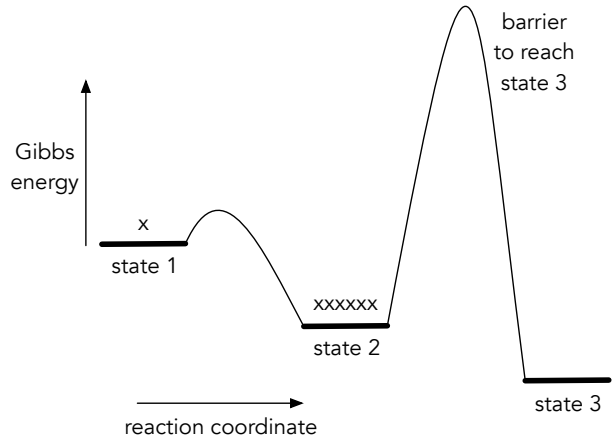


Figure 3.1. Gibbs energy diagrams for protein states show the available states, their relative energies and the barriers between them. The proteins (x) equilibrate between states 1 and 2 with more proteins in state 2 because it is lower in energy. The barrier between states 1 and 2 is low so that the proteins equilibrate between them on the time scale of the experiment. State 3 is even lower in energy than states 1 and 2, but a high barrier prevents proteins from reaching that state. The x-axis on these diagrams, reaction coordinate, refers to changes in protein conformations.

which is an additional 1.36 kcal/mol. Each additional factor of 10 in the equilibrium constant corresponds to an additional 1.36 kcal/mol of Gibbs energy difference.

The barriers between protein states indicate how fast the proteins interconvert between the states. The barrier between states 1 and 2 is small, so those two states interconvert readily on the time scale of the experiment. The barrier to reach state 3 is high so that state, although lower in energy, is not populated because there has not been enough time for the proteins to equilibrate. Since protein states differ in their conformation and many conformational changes are fast, many barriers between states are low. When the conformational change requires collective movements of many atoms it can be slow. For example, large scale movement of a loop to open an active site can be slow because it requires two hinge movements and release of existing interactions between amino acids. State 3 could represent a very slow-to-form protein conformation such as one with several knots.

Chemical reactions, such as the conversion of a substrate to product, are also described by Gibbs energy diagrams. The relative Gibbs energies of the substrate and product correspond to the equilibrium constant between them. Enzymes cannot change this equilibrium; it depends only on the properties of substrate and product. The height of the barrier corresponds to rate at which they reach this equilibrium. Enzymes can change this rate by stabilizing the transition state for the reaction. This text will use Gibbs energy diagrams for both protein states, where only conformational changes occur and for chemical reactions where covalent bonds are broken and formed. For most chemical reactions, this text will use the standard Gibbs energy, indicated by the superscript °. This standard free energy refers to the standard state of 1 M concentration for both [S] and [P] and is used for simplicity. In reality, the Gibbs energies differ as the concentrations of [S] and [P] change. At the beginning of a reaction (high [S], low [P]) the Gibbs energy of the product is lower than the Gibbs energy of the substrate. As the reaction proceeds, the concentrations of substrate and product change, the Gibbs energy levels change

such that at equilibrium (low [S], high [P]) the Gibbs energy levels are equal.

3.2 The relative Gibbs energies of states determine protein properties

Proteins properties depend on the relative energies of different protein or chemical states, Figure 3.2. Protein stability depends on the difference between the native state and denatured state, Equation 3.1 above. The denatured state is less stable (ΔG_{unfold} is positive) so most of the protein exists in the native state.

The binding strength of an antibody, Ab, for an antigen, Ag, depends on the energy difference between that bound and unbound states, Equation 3.3. A positive ΔG_{diss} corresponds to unfavorable dissociation, thus favorable binding, which means the Gibbs energy of free Ag + Ab is higher than the Gibbs energy of Ag·Ab.

$$\mathrm{Ag \cdot Ab} \overset{\Delta G_{diss}}{\rightleftharpoons} \mathrm{Ag + Ab} \quad \Delta G_{diss} = G_{\mathrm{Ag+Ab}} - G_{\mathrm{Ag \cdot Ab}} \tag{3.3}$$

Enzyme catalysis involve a chemical reaction, Equation 3.4, which has a certain ΔG_{rxn} that reflects the energy difference between starting material and product. Enzymes cannot change this ΔG_{rxn}; it is determined by the starting material and product. Enzymes only speed up the approach to equilibrium.

$$starting\ material \overset{K_{eq,rxn}}{\rightleftharpoons} product \quad \Delta G_{rxn} = \frac{[product]}{[starting\ material]} \tag{3.4}$$

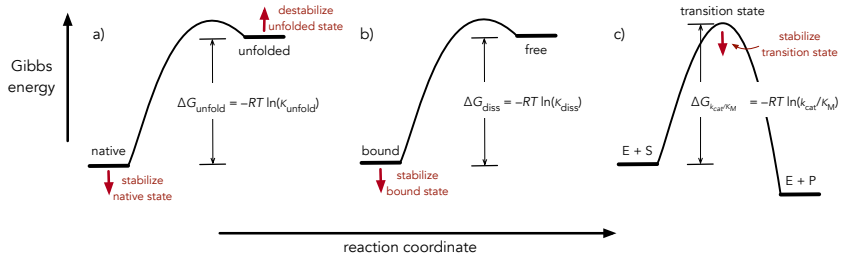


Figure 3.2. Gibbs energy differences determine protein properties. a) The Gibbs energy difference between the native and unfolded states, ΔG_{unfold}, sets the equilibrium constant between the native and unfolded states, K_{unfold}, to determine the stability of a protein. Increasing ΔG_{unfold} by either stabilizing the native state or destabilizing the unfolded state yields a more stable protein because unfolding becomes less favorable. b) The Gibbs energy difference, ΔG_{diss}, sets the equilibrium constant between the bound and free states to determine the binding affinity of a protein. Increasing ΔG_{diss} by stabilizing the bound state yields tighter binding because dissociation becomes less favorable. c) The Gibbs energy difference, $\Delta G_{k_{cat}/K_M}$, between enzyme and starting material, E + S, and the transition state sets the rate constant, k_{cat}/K_M, that determines how fast the reaction proceeds. Decreasing $\Delta G_{k_{cat}/K_M}$ by stabilizing the transition state speeds up the reaction because it lowers the barrier separating enzyme and starting material from enzyme and product, E + P.

Transition state theory proposes that the rate of a chemical reaction depends on the energy difference between starting materials and the transition state, $\Delta G^{\ddagger}$, Equation 3.5. The transition state lies at higher energy than the starting material, so $\Delta G^{\ddagger}$ is positive. This equilibrium is a pseudo-equilibrium because some of the transition state species are continuing on to product.

$$\begin{gathered} starting\,material \overset{K_{eq,ts}}{\rightleftharpoons} transition\,state \\ \Delta G^{\ddagger}_{uncatalyzed} = \frac{[transition\,state]}{[starting\,material]} \end{gathered} \tag{3.5}$$

Enzymes catalyze reactions by stabilizing the transition state. Enzymes bind the transition state and stabilize the conformation and charge distribution needed for the bond-breaking and bond-making, Equation 3.6. This stabilization lowers the barrier between starting material and product so the reaction proceeds faster. The two states that determine enzyme catalysis are the $starting\ material + enzyme$ state and the $transition\ state \cdot enzyme$ state.

$$\begin{gathered} starting\,material + enzyme \overset{K_{eq,ts,cat}}{\rightleftharpoons} transition\,state \cdot enzyme \\ \Delta G^{\ddagger}_{catalyzed} = \frac{[transition\,state \cdot enzyme]}{[starting\,material][enzyme]} \end{gathered} \tag{3.6}$$

3.3 Protein engineering manipulates the Gibbs energies of states

Improving protein properties requires changing the relative Gibbs energies of different protein states, Table 3.1. Increasing the value of ΔG_{unfold} stabilizes proteins, increasing the value of ΔG_{diss} strengthens binding, and decreasing $\Delta G_{k_{cat}/K_M}$ speeds up catalysis. Changing the relative Gibbs energies is called differential stabilization. The structural changes introduced by protein engineering must stabilize some states more than others so that the difference between them changes, which results in a change in the protein properties.

Table 3.1. Changes in Gibbs energy needed to improve protein function.

Protein Function	Gibbs Energy Change	Comparison States	Engineering Approaches for Improvement
stability	ΔG_{unfold}	folded protein vs. unfolded ensemble	increase ΔG_{unfold} by stabilizing folded | protein or destabilizing unfolded ensemble
binding	ΔG_{diss}	protein bound to ligand vs. solvated protein & ligand	increase ΔG_{diss} by stabilizing the bound | complex
catalysis	$\Delta G_{k_{cat}/K_M}$	enzyme and starting materials vs. transition state	decrease $\Delta G_{k_{cat}/K_M}$ by stabilizing | the transition state

For example, altering the relative Gibbs energies of the two states

involved in binding, the unbound state (free Ag + Ab) and the bound state (Ag·Ab), alters binding strength, Equation 3.3 above. To create a differential stabilization, the substitution must alter the Gibbs energies of the two states by different amounts. The substitution of an alanine residue with phenylalanine in the antigen-binding region of the antibody might strengthen binding if the antigen interacts with the added phenyl group. Thus, the bound state would be stabilized. In the free antibody, exposing an additional hydrophobic phenyl group to water is unfavorable, thus the free state would be destabilized. The net result is a larger Gibbs energy of dissociation, which corresponds to stronger binding, Equation 3.7 and Equation 3.8.

$$\text{Ag} \cdot \text{Ab}_{Ala} \rightleftharpoons \text{Ag} + \text{Ab}_{Ala} \quad \Delta G_{diss,Ala} = \text{original} \tag{3.7}$$

$$\text{Ag} \cdot \text{Ab}_{Phe} \rightleftharpoons \text{Ag} + \text{Ab}_{Phe} \quad \Delta G_{diss,Phe} = \text{larger than original} \tag{3.8}$$

In contrast, the substitution of an alanine residue with phenylalanine in the core of the antibody, remote from the antigen-binding site, would alter both states similarly so the difference in Gibbs energy for Equation 3.3 would remain the same. The binding strength would not improve.

The improvement factor of a protein engineering experiment is the factor by which the stability, binding, reaction rate or selectivity has improved. This improvement corresponds to a ratio of equilibrium constants or rate contants for the property being improved. The value for the variant protein is divided by the value for the original protein, Equation 3.9.

$$\text{improvement factor} = \frac{\text{variant } K_{eq} \text{or rate constant}}{\text{original } K_{eq} \text{or rate constant}} \tag{3.9}$$

The improvement factor should be >1 if the protein has improved. If the improvement factor is less than one when it seems that it should be

>1, consider whether the equilibrium or rate constants being compared should increase or decrease to improve the protein. For example, protein stability depends on the unfolding equilibrium constant. An increase in the unfolding equilibrium constant corresponds to more unfolding and decreased stability. An increase in the unfolding equilibrium constant is not an improvement, but a degradation of stability. The equilibrium constant that should be compared to identify improvement is the inverse, or $1/K_{unfold}$. The ratio of $1/K_{unfold}$ for the variant divided by that for the orginal corresponds to the improvement factor, Equation 3.10. If the variant unfolds less readily, then $1/K_{unfold}$ for the variant will be larger and the improvement will be >1.

$$\text{stability improvement factor} = \frac{1/K_{unfold,\,variant}}{1/K_{unfold,\,original}} \tag{3.10}$$

One can also assign a Gibbs energy to the improvement. Protein properties in Figure 3.2 correspond to Gibbs energy difference between two states, a ΔG. The Gibbs energy for an improvement contains two Δ's because it is the difference between ΔG for the variant and the ΔG for the original, Equation 3.11.

$$\Delta\Delta G_{improvement} = \Delta G_{variant} - \Delta G_{original} \tag{3.11}$$

An improvement can be a positive or a negative $\Delta\Delta G$ depending on how the property is defined. An increase in protein stability corresponds to a positive value for $\Delta\Delta G_{unfold}$, Equation 3.12, because making unfolding less favorable corresponds to a more stable protein.

$$\Delta\Delta G_{unfold} = \text{larger than original} - \text{original} = \text{positive value} \tag{3.12}$$

An increase in antibody binding affinity corresponds to a positive value for $\Delta\Delta G_{diss}$, Equation 3.13, because making dissociation less favorable corresponds to tighter binding.

$$\Delta\Delta G_{diss} = \text{larger than original} - \text{original} = \text{positive value} \tag{3.13}$$

In contrast, an increase in reaction rate corresponds to a negative value for$\Delta\Delta G^{\ddagger}$, Equation 3.14, because lowering the energy of the transition state speeds up the reaction.

$$\Delta\Delta G^{\ddagger} = \text{smaller than original} - \text{original} = \text{negative value} \tag{3.14}$$

Selectivity is another protein property that is often a target for protein engineering. Selectivity refers to selective binding or selective catalysis. Binding selectivity compares the binding of a protein to two targets, so it is the difference in Gibbs energies for binding of the two targets, Equation 3.15.

$$\begin{aligned} \text{binding selectivity} &= \Delta G_{diss,target\ 1} - \Delta G_{diss,target\ 2} \\ &= \Delta\Delta G_{diss,target\ 1\ vs.\ target\ 2} \end{aligned} \tag{3.15}$$

There are two Δ's in the Gibbs energy for binding selectivity. Similarly, selectivity for reaction compares the reactivity of two substrates or the formation of two different products from an enzyme. Since the selectivity of binding or reaction compare two reactions, it already contains two Δ's: $\Delta\Delta G$. The Gibbs energy change for the improvement in selectivity will have three Δ's: $\Delta\Delta G$, Equation 3.16.

improvement in binding selectivity

$$= \Delta\Delta G_{diss,\ variant,\ target\ 1\ vs.\ target\ 2} - \Delta\Delta G_{diss,\ original,\ target\ 1\ vs.\ target\ 2}$$
$$= \Delta\Delta\Delta G_{diss,\ original \rightarrow variant,\ target\ 1\ vs.\ target\ 2} \tag{3.16}$$

One can write a similar equation for reaction selectivity.

In Chapter 1 we learned to define protein engineering goal in terms of protein properties. Now we learned that relative Gibbs energies of specific protein states must change to improve these protein properties. The next sections describes the non-covalent interactions and entropy contributions that we can modify to engineer changes in these Gibbs energies.

3.4 Non-covalent interactions & entropy

To engineer changes in the Gibbs energies, one needs to alter the molecular features that contribute to differences in the Gibbs energies of different protein states. The two contributors to differences in Gibbs energy are enthalpy, ΔH, and entropy, ΔS, Equation 3.17.

$$\Delta G = \Delta H – T\Delta S \tag{3.17}$$

Protein states differ not in their covalent bonds, but in their conformations and associations with solvent and other molecules. These non-covalent interactions differ between states so manipulating these interactions can change the properties of a protein. For protein states, the enthalpic contributions come from favorable or unfavorable non-covalent interactions within a protein, between protein and solvent and between protein and target molecules. Stabilizing non-covalent interactions lowers the Gibbs energies of the corresponding states,

while destabilizing interactions raise these energies. The entropic contributions come from differences in molecular flexibility and solvent interactions between the states. Entropy is the degree of disorder in a system so states with higher disorder are more favorable and have lower Gibbs energy.

Non-covalent interactions. The two fundamental non-covalent interactions are electrostatic interactions and van der Waals interactions. Charged or partially charged atoms attract or repel each other. Interactions between oppositely charged atoms are attractive (negative energies), while interactions between like-charged atoms are repulsive (positive energies). These energies vary inversely with the distance between them, r, according to Coulomb's law, Equation 3.18, where q_1 and q_2 are the charges on the atoms and ϵ_0 is the dielectric constant of the medium separating them.

$$\text{electrostatic energy contribution} = \frac{q_1 \cdot q_2}{4\pi \cdot \epsilon_0 \cdot r} \tag{3.18}$$

Since the energy varies inversely with r, they persist over long distances, Figure 3.3. Hydrogen bonds are a special type of electrostatic interaction that includes contributions from covalent bonding.

All atoms interact via van der Waals interactions. The interaction is zero at long distances, attractive at intermediate distances and repulsive at short distances. The Lennard-Jones potential, Equation 3.19, describes the energetics of van der Waals interactions where r is the distance between the atoms and r_o is the distance where their interaction energy is lowest (optimal distance or van der Waals distance). Here the constant, ϵ, is *not* the dielectric constant, but the maximum interaction energy (depth of the attractive well). When $r_o = r$, the van der Waals energy is $-\epsilon$. Different values of ϵ and r_o account for different interactions between different pairs of atoms. The 12^{th}- and 6^{th}-power inverse dependence on r indicates that both favorable and unfavorable interactions act only at short distances. Molecular modeling programs typically include van der

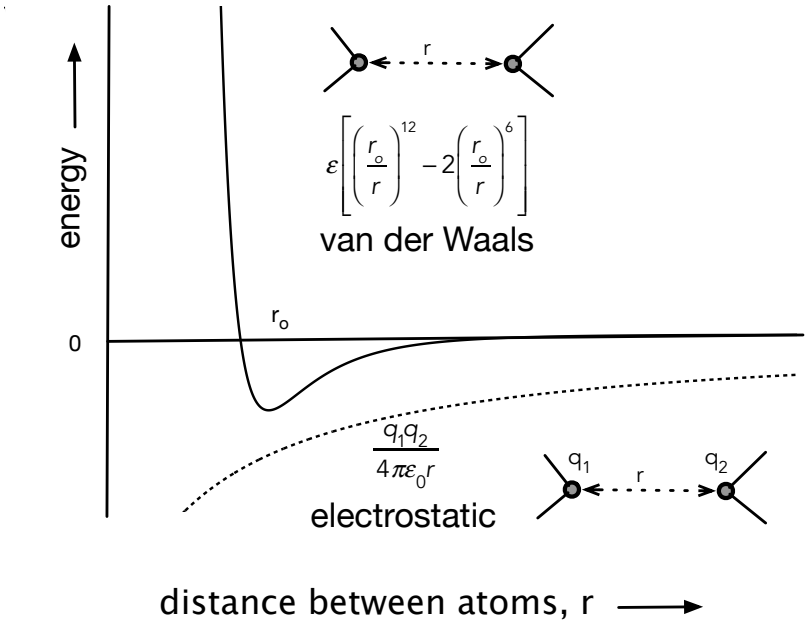


Figure 3.3. The two fundamental non-covalent interactions are electrostatic interactions (dashed line) and van der Waals interactions (solid line). The attractive electrostatic interactions between oppositely charged atoms become increasing favorable (lower energy) as the two atoms approach each other. The energy decreases according to $1/r$, which allows electrostatic interactions to persist over long distances. A similar unfavorable interaction occurs between two like charges (not shown). The van der Waals interactions are unfavorable at short distances due to steric bumping, slightly favorable at intermediate distances due to the interaction of permanent and induced dipoles. The favorable interaction decreases according to $1/r^6$, so it dissipates to zero more quickly than electrostatic interactions.

Waals interactions between atoms separated by less than 10 Å. This approach saves computing interactions that are zero or very close to zero.

$$\text{van der Waals energy contribution} = \epsilon \left[\left(\frac{r_o}{r} \right)^{12} - 2 \left(\frac{r_o}{r} \right)^{6} \right] \quad (3.19)$$

The physical basis for the repulsion of neutral atoms at short distances is the Pauli exclusion principle. This principle states that electrons cannot have the same quantum numbers. This principle prevents atoms from collapsing since outer electrons of an atom cannot move closer to the nucleus into shells already occupied by other electrons. One can imagine this interaction as atoms bumping into one another.

The physical basis of the attractive interaction can be orientation, dispersion or induction depending on whether the molecules are non-polar or polar, Figure 3.4. If the atoms are in a polar bond, then they create a permanent dipole. Orientations between the bonds that cancel the dipoles create an attractive interaction, while orientations that reinforce the dipoles create repulsive interactions. If the atoms are non-polar (lack a permanent dipole), then fluctuation of their electron clouds creates instantaneous, temporary dipoles. Favorable alignment of these dipoles creates a weak net attraction, called London dispersion. For large, non-polar atoms, this attraction between instantaneous, temporary dipoles is the main origin of van der Waals attraction. The final possibility for van der Waals forces is between a non-polar atom and a polar bond. The permanent dipole in the polar bond induces a dipole in the non-polar atom, which creates an attractive interaction, sometimes called a Debye force.

Atoms approach each other until attractive interactions between them (electrostatic and van der Waals attractions) balance the repulsive interactions between them (electrostatic repulsion and bumping).

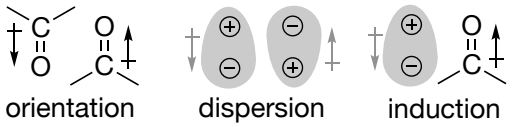


Figure 3.4. Three contributions to the attractive van der Waals interaction. *Orientation:* Polar molecules contain dipoles (black arrows), which create favorable electrostatic interactions when they orient to cancel each other. *Dispersion:* Fluctuating polarization of electron clouds (gray) in nonpolar atoms creates temporary dipoles (gray arrows) that attract each other. *Induction:* A polar bond induces a temporary dipole in the electron cloud of a nearby non-polar atom to create a attractive interaction.

Non-covalent interations in water. Moving from vacuum to water changes the interaction between atoms in two ways. First, the net strength of the interactions between atoms decreases because in water the atoms always interact with something - water or each other - instead of nothing or each other in vacuum. Second, the hydrophobic effect strengthens the association between non-polar atoms to minimize the unfavorable association of water with non-polar atoms.

The equations above describe interactions between atoms in vacuum where the atoms have a choice between interacting with nothing (vacuum) or another atom. In water, the choice changes. Atoms interact with either water or other atoms; interacting with nothing is no longer an option. For polar and charged atoms electrostatic interactions, including hydrogen bonds, are weaker in water than they would be in vacuum, Table 3.2, because the atoms also make favorable electrostatic interactions and hydrogen bonds to water when they are not interacting with each other. For example, hydrogen bonds between solutes in water are only 0-2 kcal/mol, not the 3-5 kcal/mol in vacuum. Hydrogen bonds in water are simply traded to different partners, Equation 3.20, so the advantage is any net gain in the new set of hydrogen bonds. Two solutes start with

hydrogen bonds to water. If the solutes trade those hydrogen bonds for hydrogen bond to each other, they break the solute-water hydrogen bond and create water-water hydrogen bonds from the released water and solute-solute hydrogen bonds. The net gain depends on the relative strength of all these changing hydrogen bonds.

$$D{-}H(water) + A(water) \rightleftharpoons water{-}water + D{-}H \cdots A \qquad (3.20)$$

Here A is the hydrogen bond acceptor atom and $D{-}H$ is the hydrogen bond donor. In a similar manner van der Waals interactions are weaker in water than in vacuum because van der Waals interactions between water and solutes offsets the advantage of van der Waals interactions between solutes.

Table 3.2. Noncovalent interactions between atoms in vacuum and in water.[a]

Interaction	Type of Atom	Origin	Strength in Vacuum	Strength in Water
electrostatic interactions	charged & partially charged atoms	attraction between opposite charges, repulsion between like charges	3-7 kcal/mol	1-5 kcal/mol
hydrogen bonds	H + O, N, halogen	sharing a hydrogen atom between two electronegative atoms, an electrostatic interaction with a covalent contribution	3-5 kcal/mol	0-2 kcal/mol
van der Waals repulsion (bumping)	all	unfavorable bumping between atoms	0-large	0-large
van der Waals attraction	non-polar, polar bonds	orientation, dispersion, induction	≈1 kcal/mol for Me-Me	≤1 kcal/mol for Me-Me

Interaction	Type of Atom	Origin	Strength in Vacuum	Strength in Water
hydrophobic effect	non-polar, only occurs in water	burying of hydrophobic surfaces minimizes unfavorable water orientations & creates van der Waals attraction	0	≈2 kcal/mol for Me-Me

[a] Estimates from Lodish, H. F., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000). Molecular cell biology (4th ed). W.H. Freeman. Section 2.2

The unique structure of water creates the hydrophobic effect, which is the tendency of non-polar solutes to cluster in water. The hydrophobic effect is not a bond or interaction between non-polar atoms; instead, it is a consequence of the behavior of water in the presence of a hydrophobic solute. The quote below describes the dynamic structure of water.

> Water isn't unique in forming hydrogen bonds, but it is the only common substance that can be joined by these gentle, frangible molecular handclasps into a three-dimensional network. Most liquids are little more than a disorderly scrum of jostling molecules. But water is delicately poised between order and disorder, constantly adopting a defective version of the framework structure that, in ice, immobilizes the water molecules into crystalline regularity. - Philip Ball[46]

Adding a hydrophobic solute disrupts this dynamic structure of water. The balance between order and disorder tips toward an ordered structure surrounding the hydrophobic solute. Water molecules at the non-polar surface cannot form hydrogen bonds with the non-polar surface, so they make fewer total hydrogen bonds, Figure 3.5. These hydrogen bonds are stronger than those in bulk water, immobilizing the water molecules at the surface of a non-polar solute into an ice-like cage structure. The restricted mobility of the water molecules in this cage decreases their entropy, making this arrangement less stable than bulk water. Clustering of non-polar solutes reduces the contact area between water and non-polar surface, releases water molecules from the cage-like structure and lowers their energy. The clustering of hydrophobic solutes occurs not due to an attractions between the non-polar atoms, but because of the change in water structure when they cluster. Weak van der Waals interactions between the non-polar atoms contribute a small amount to the hydrophobic effect, but the main origin is the behavior of the water molecules. The spontaneous separation of oil and water in a mixture is a macroscopic demonstration of the hydrophobic effect.

Entropy. Besides the non-covalent interactions between atoms, entropy

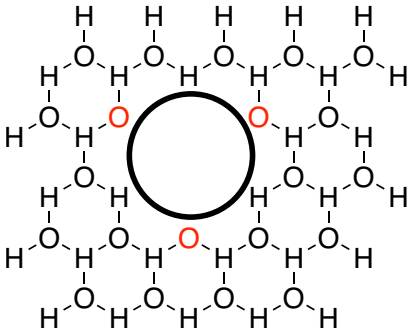


Figure 3.5. Water surrounds hydrophobic solutes with an ice-like structure. In this 2-D diagram of a hydrophobic circle in water, three oxygens near the circle (red color) have only two hydrogen partners, while the rest of the oxygen atoms have three hydrogen partners. Fewer partners strengthen the remaining bonds leading to a rigid, ice-like structure at the hydrophobic interface. This 2-D diagram simplifies the true 3-D structure, where each oxygen atom has four hydrogen partners in the bulk water and two or three hydrogen partners in the water at the hydrophobic interface.

also contributes to the Gibbs energy of the different protein states. Entropy is the degree of disorder in a system. For molecules, entropy is related to the number of microstates that the molecule can adopt. Assuming equal probability for each microstates, the entropy of a molecular state is proportional to the natural logarithm of the number of microstates within that state, Equation 3.21.

$$S = R\ln(\#\text{ of microstates}) \tag{3.21}$$

A typical example of microstates is different conformations along a single bond, such as the three different conformations of the side chain of serine. Another typical case is two enantiomeric possibilities, such as the gauche+ and gauche– conformations of butane.

The microstate of a protein refers to the a specific conformation or arrangement of atoms and bonds at a given moment. Microstates can differ in the position of each amino acid main chain, the orientation of individual side chains and the hydrogen bonds that form between them. Both the amino acid sequence of the protein and environmental factors such as temperature, pH, and ionic strength determine which microstates are accessible to a protein.

Proteins exist in countless numbers of possible microstates; fortunately, we do not need to count them because we are concerned only with *differences* in entropy and therefore only the relative numbers of microstates. For example to compare the entropy of two states, such as folded and unfolded protein, one needs the relative number of available microstates in each state, Equation 3.22.

$$\Delta S = R\ln\left(\frac{\#\text{ of microstates}_{unfolded\ protein}}{\#\text{ of microstates}_{folded\ protein}}\right) \tag{3.22}$$

A configuration that has more microstates will have a lower Gibbs energy than one with fewer microstates, assuming all other things are equal. For

example, if the product state allows free rotation of the side-chain atoms, while the starting material state does not, then this difference lowers the Gibbs energy of the product state. The contribution of entropy to the Gibbs energy is $-T\Delta S$, Equation 3.23.

$$\begin{aligned}&\text{entropy contribution to } \Delta G = -T\Delta S \\ &= -TR\ln\left(\frac{\text{\# of microstates}_{product}}{\text{\# of microstates}_{starting\ material}}\right)\end{aligned} \tag{3.23}$$

If the flexible side chain in the product has three rotatable bonds, each of which can adopt three stable conformations, then there are nine possible microstates in the product state. The fixed side chain in the starting material can adopt only one microstate. At 25 °C,

$$\text{entropy contribution to } \Delta G = -298\ \text{K}\cdot 1.987\ \text{cal/mol}\cdot\text{K}\cdot ln(9/1) = -1.3\ \text{kcal/mol}$$

Thus, amino acid substitutions that increase the disorder or flexibility of a state stabilize that state. While protein structures show stabilizing noncovalent interactions like hydrogen bonds, static structures do not show the flexibility of a state. One must remember to consider differences in flexibility in addition to structural differences when estimating Gibbs energy differences between states.

Protein folding. Folded structures are the low-energy state for most proteins. This low energy comes mainly from the hydrophobic effect. Burying a –CH_2 group contributes 1.1 ± 0.5 kcal/mol to protein stability. The hydrophobic effect provides ~60% of the driving force to collapse the amino acid chain into a compact structure.[47] The remaining ~40% comes from attractive interactions between the amino acids. The chains orient to maximize the hydrophobic effect and the attractive electrostatic and van der Waals interactions including hydrogen bonds. This favorable contact between non-polar atoms tightly packs amino acids within the hydrophobic core of proteins. The only empty space is small cavities where the packing is imperfect.

The size, hydrophobicity, charge, and hydrogen bonding abilities of the twenty proteinogenic amino acids differ from one another. Polar amino acids favor the outside of the structure to interact with the surrounding polar water molecules, while non-polar amino acids favor the interior of the structure due to the hydrophobic effect. The shape, charge, and hydrogen bonding determine the specificity of the protein folding; that is, which sections fold into helices and strands and subsequently how these secondary structures associate into the tertiary structures that form protein domains.

3.5 Logic of protein engineering

Protein engineering will involve substituting one, or more likely, many amino acids in a protein. Amino acid substitutions change non-covalent interactions and entropy, which in turn change the Gibbs energy of states, which changes protein properties. Gibbs energy differences connect protein properties to these amino acid substitutions, Figure 3.6. Protein properties like stability, binding, and catalysis depend on Gibbs free energies differences between different states. Amino acid substitutions change the molecular interactions (flexibility and non-covalent interactions) that determine these Gibbs energies. Thus, replacing amino acids changes interactions, which changes Gibbs energies, which changes protein properties.

Protein engineering involves changing the structure of the protein to change interactions within each state such that protein properties improve. In most cases, the goal is to improve some protein properties while keeping others unchanged. Amino acid replacements must both improve the targeted properties and not ruin the existing favorable properties. The next chapter describes working with protein structures and computer modeling approaches to predict how amino acids substitutions affect the Gibbs energies of various states.

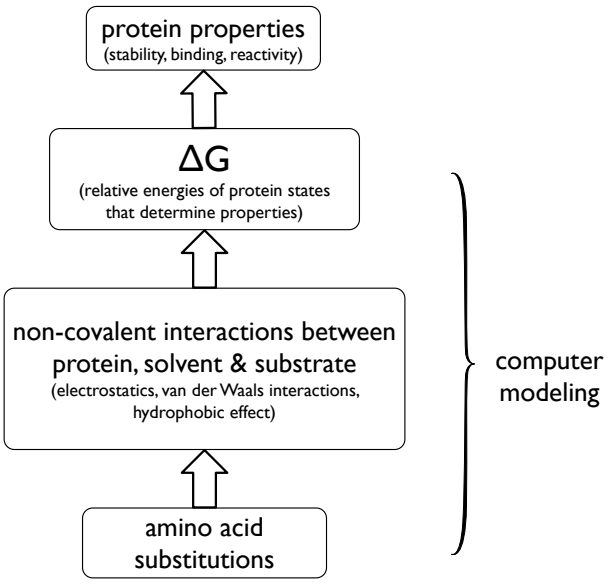


Figure 3.6. The logic of protein engineering. Differences in Gibbs energies of various states determine protein properties like stability, binding, and catalysis. For example, the relative Gibbs energies of the folded and unfolded protein states determine a protein's stability. The states differ in energy because the unfolded state of a protein is flexible and the amino acid residues interact mainly with solvent, while the folded state is less flexible and amino acid residues interact mostly with each other. Protein engineering involves changing the amino acid sequence of the protein to change the flexibility and non-covalent interactions to change the relative Gibbs energies of the states that determine protein properties. Computer modeling keeps track of the complex interactions between amino acids to predict the Gibbs energy of the different states.

3.6 Amino acid structures

Most protein engineering will involve substituting one amino acid for another with the expectation that the replacement will improve the desired property. The basis for this expectation is the differences in the side chains, R, of the different α-amino acids. Substitutions change the size and shape of the side chain and may remove and/or introduce new functional groups.

The structure below shows the general structure of an L-α-amino acid, Figure 3.7. Biochemists use the Fischer nomenclature (D or L) instead of the Cahn-Ingold-Prelog nomenclature (*R* or *S*) for amino acids and sugars. In the Fischer nomenclature all proteinogenic amino acids have the L-configuration, while the Cahn-Ingold-Prelog nomenclature assigns *R* to cysteine and *S* to the others making it less convenient. The number in red indicates the pK_a values for the protonated amino and carboxylic acid groups. At a pH above this value, these groups are deprotonated. At pH 7, α-amino acids exist as zwitterions: a positive charge at the α-amino group and a negative charge at the carboxylate. The carboxylic acid deprotonates to the carboxylate at pH 7 because this pH is above the pK_a of the carboxylic acid group, but the α-amino group does not deprotonate because pH 7 is below the pK_a of the protonated amino group.

Planning amino acid replacements requires knowing the structures and properties of the amino acid side chains. Figure 3.8 shows the side chains of the twenty standard amino acids grouped according to their classification as hydrophobic, polar, or charged. The figures also include mnemonics for the one-letter codes for the amino acids.

The nine amino acids with hydrophobic side chains are typically buried inside the protein core when the protein folds. The six amino acids with polar side chains form hydrogen bonds as proton donors or acceptors. The number indicates the pK_a values for these residues in typical unfolded peptides. At high pH or in unusual protein environments, cysteine and tyrosine may deprotonate and become negatively charged.

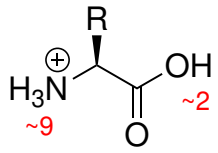


Figure 3.7. The general structure of an L-α-amino acid in its fully protonated form. The amino group is attached at the α-position of the carboxylic acid. All proteinogenic amino acids have the absolute configuration shown, which is called L in the Fischer nomenclature. The fully protonated form shown exists below pH 2; at pH 7, the carboxylic acid is deprotonated to the carboxylate and the net charge on the amino acid is zero.

The five amino acids with charged side chains often form salt bridges. Figure 3.8 shows these five in their protonated forms. At pH 7, aspartate and glutamate would be negatively charged, histidine would be mostly neutral, and lysine and arginine would be positively charged. Histidine often occurs in enzyme active sites because the pK_a of the side chain near-neutral pH allows it to both accept and donate protons readily.

Figure 3.8 also shows the atom naming for the side chains of several amino acids. The first character of the atom name is the chemical symbol for the atom type. For example, all names for carbon atoms begin with C. The next character indicates the remoteness from the carboxylic acid according to the Greek alphabet: α = alpha, β = beta, γ = gamma, δ = delta, ε = epsilon, ζ = zeta, η = eta. For example, the side chain atoms in serine are Cβ and Oγ. The last character of the atom name is a number to indicate the branch if required. For example, Oδ1 and Oδ2 indicate the carboxyl oxygens in aspartate, which are equivalent due to resonance but correspond to different branches. Greek letters are inconvenient in protein structure files, so they are transliterated with Roman letters according to α = A, β = B, γ = G, δ = D, ε = E, ζ = Z, η = H. For example, the side chain atoms in serine are named CB and OG in protein structure files. The main chain atoms in the protein structure files are

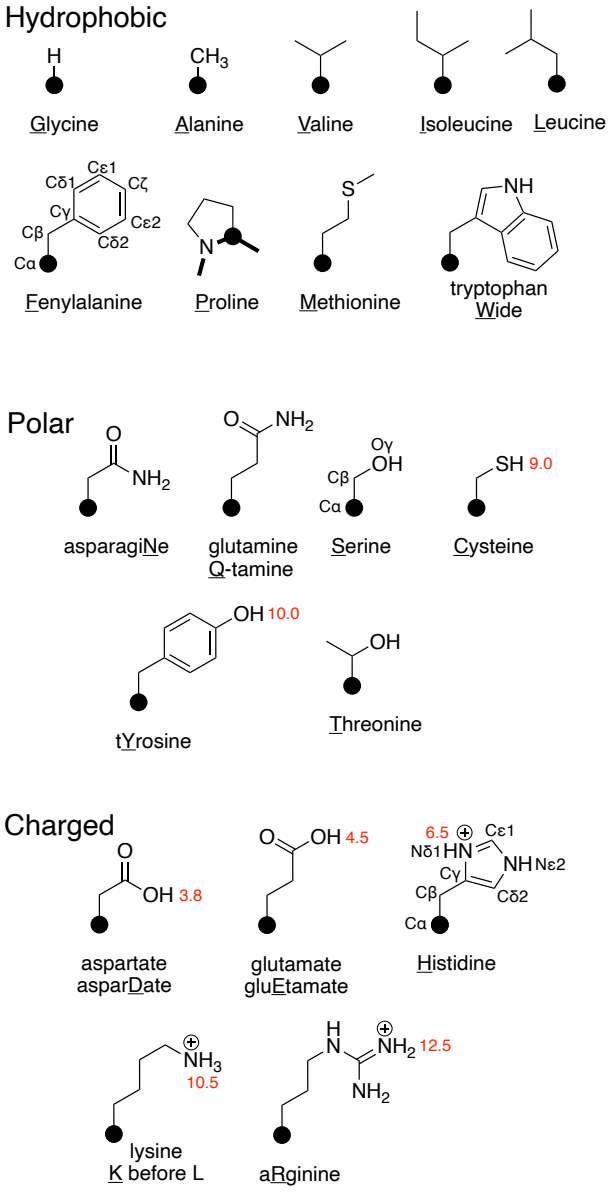


Figure 3.8. The twenty standard amino acids. The large black dot marks the alpha carbon. The drawing of proline includes the backbone nitrogen and backbone bonds shown in bold. The amino acids names (or corresponding mnemonics) contain underlined letters signifying their single-letter abbreviations. The drawings of phenylalanine, serine and histidine include the atom names of the carbons. The Greek letters in the atom names indicate the remoteness from the carboxylic acid carbon atom: α = alpha, β = beta, γ = gamma, δ = delta, ε = epsilon, ζ = zeta. The number in the atom name indicates the branch. For example, Cδ1 and Cδ2 indicate the equivalent carbon atoms along two different branches. The red numbers indicate the pK_a of this group in a typical unfolded peptide.

named N, CA, C, and O corresponding to the amino nitrogen, Cα carbon, carbonyl carbon, and carbonyl oxygen, respectively.

Acid/base behavior, resonance structures and tautomers of imidazole. The versatile nature of the imidazole ring of histidine make it important for catalysis, but the versatility also creates complexity. Below is a review of the acid/base behavior, resonance structures and tautomers for the imidazole ring.

The neutral, uncharged imidazole can act both as an acid and as a base, but only the basic behavior is relevant to most biochemistry, Figure 3.9. Imidazole often acts as a base because it accepts a proton at pH 6.5, which is a common pH in biochemistry. The protonated form is also called an imidazolium cation; its pK_a is 6.5. Because histidine contains an the imidazole ring in its side chain, it can act as a base or, in the imidazolium form, as an acid near neutral pH. This ability allows it to catalyze proton transfers near neutral pH so histidine often occurs in the active site of enzymes. Note that neutral imidazole rarely acts as an acid because it loses its proton only at pH 15, which is an extreme pH for biochemistry. The pK_a of imidazole is 15.

The imidazolium cation is a single structure that is best represented by two resonance structures, Figure 3.10. These structures show that two nitrogen atoms are equivalent because the positive charge of the imidazolium cation is equally distributed between both nitrogen atoms. One Lewis structure is inadequate to accurately represent the imidazolium cation. The resonance structures are two Lewis structures that contribute to the true structure. An alternative to resonance structures is to draw dotted lines to represent delocalized electrons and partial charge indicators. Note that resonance is *not* a rapid equilibrium between the two structures; instead, the true structure is a weighted average of the two resonance structures.

In contrast to the imidazolium cation, the two nitrogens in neutral imidazole differ from one another - one has an added proton, while the other does not. Neutral imidazole exists as two equivalent isomers where the

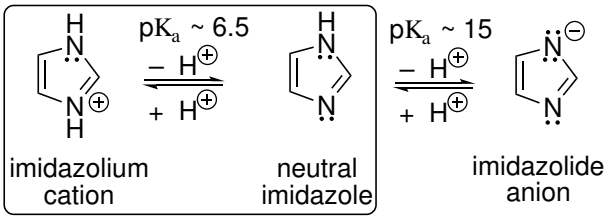


Figure 3.9. Protons transfer readily between neutral imidazole and the imidazolium cation near neutral pH because the pK_a of the imidazolium cation is 6.5. Imidazole is the base (proton acceptor) while the imidazolium cation is the acid (proton donor). Imidazole can also donate a proton, thereby acting as an acid, but this behavior is rare in biochemistry. The pK_a of imidazole is 15, so harsh base is required for imidazole to act as an acid. Expect to see only the neutral imidazole and the imidazolium cation (boxed) in biochemical mechanisms.

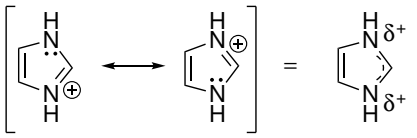


Figure 3.10. Resonance structures of the immidazolium cation show that the positive charge is delocalized on both nitrogen atoms and both C-N bonds have partial double bond character. The two nitrogen atoms are equivalent. Resonance structures are represented by double-headed arrows and represent a single molecular species, not equilibrium between separate species. Atom locations are identical between resonance structures, but the electron locations differ. (Other resonance structures of the immidazolium cation, not shown, make minor contributions to the structure because they are much higher in energy.)

proton is on different nitrogen atoms. The two isomers exchange rapidly by transferring a proton. These two equivalent isomers of imidazole are tautomers, which are rapidly equilibrating isomers, usually by a proton transfer. In histidine, the imidazole ring contains a substituent making the two two tautomers different, Figure 3.11. The tautomers of histidine differ by which nitrogen is protonated. Catalysis may require one of the two tautomers of histidine in the active site.

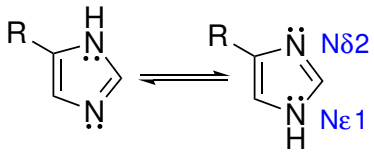


Figure 3.11. Histidine with a neutral imidazole ring exists as a pair of tautomers, which differ by which nitrogen is bonded to a proton. In one tautomer, the proton sits on Nδ2; in the other it sits on Nε1. The tautomers equilibrate by proton transfer via an immidazolium cation (not shown).

Glossary

Conformation see microstate

Differential stabilization of states alters the Gibbs energy difference between states and therefore the equilibrium constant between the states. Protein engineering seeks to differentially stabilize the states that determine protein stability, binding, reactivity, or selectivity in order to improve protein properties.

Entropy is the degree of disorder quantified by the number of accessible microstates. Entropy is stabilizing contribution to the Gibbs energy.

Gas constant (R) is the conversion factor used in chemistry to connects energy to the temperature of the system for a mole of particles. The units are calories (or joules)/°K·mole. Physicists use the Boltzmann

constant, which makes the same connection, but per particle. Its units are calories (or joules)/°K.

Hydrophobic effect is the tendency of non-polar solutes to cluster in water. The origin of this clustering is the unfavorable ordering of water molecules at a non-polar interface. Minimizing non-polar interface by clustering releases the ordered water molecules.

Microstates are specific conformations or arrangements of atoms and bonds at a given moment. Protein microstates can differ in the position of each amino acid main, the orientation of individual side chains and the hydrogen bonds that form between them. The relative number of microstates between two protein states determines the entropy difference between them.

States are different forms of a protein that differ in their conformation and/or association with other molecules, but not in covalent structure. States, sometimes called macrostates, consist of countless microstates. States refer to protein forms that have macroscopic or bulk properties that one can measure. The folded and unfolded states of a protein have different fluorescence properties.

Tautomers are rapidly equilibrating isomers, most often by transfer of a proton to a different atom. The keto and enol forms of carbonyl compounds are tautomers. The imidazole ring of histidine exists in two tautomeric forms.

References

1. Morrison, C., & Lähteenmäki, R. (2018). Public biotech in 2017-the numbers. *Nat. Biotechnol.*, *36*(7), 576–584. https://doi.org/10.1038/nbt.4175
2. Brannigan, J. A., & Wilkinson, A. J. (2002). Protein engineering 20 years on. *Nat. Rev. Mol. Cell Biol.*, *3*(12), 964–970. https://doi.org/10.1038/nrm975

3. Estell, D. A., Graycar, T. P., & Wells, J. A. (1985). Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.*, *260*(11), 6518–6521. https://doi.org/10.1016/S0021-9258(18)88811-1
4. Pace, C. N., & Hermans, J. (1975). The stability of globular proteins. *CRC Crit. Rev. Biochem.*, *3*(1), 1–43. https://doi.org/10.3109/10409237509102551
5. Taverna, D. M., & Goldstein, R. A. (2002). Why are proteins marginally stable? *Proteins*, *46*(1), 105–109. https://doi.org/10.1002/prot.10016
6. Fersht, A. R., & Serrano, L. (1993). Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.*, *3*(1), 75–83. https://doi.org/10.1016/0959-440X(93)90205-Y
7. Bloom, J. D., Lu, Z., Chen, D., Raval, A., Venturelli, O. S., & Arnold, F. H. (2007). Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.*, *5*(1), 97–21. https://doi.org/10.1186/1741-7007-5-29
8. Gordon, S. R., Stanley, E. J., Wolf, S., Toland, A., Wu, S. J., Hadidi, D., Mills, J. H., Baker, D., Pultz, I. S., & Siegel, J. B. (2012). Computational design of an α-gliadin peptidase. *J. Am. Chem. Soc.*, *134*(50), 20513–20520. https://doi.org/10.1021/ja3094795
9. Harvey, S. B., Stone, M. D., Martinez, M. B., & Nelsestuen, G. L. (2003). Mutagenesis of the γ-carboxyglutamic acid domain of human factor VII to generate maximum enhancement of the membrane contact site. *J. Biol. Chem.*, *278*(10), 8363–8369. https://doi.org/10.1074/jbc.M211629200
10. Persson, E., Olsen, O. H., Bjørn, S. E., & Ezban, M. (2012). Vatreptacog alfa from conception to clinical proof of concept. *Semin. Thromb. Hemost.*, *38*(3), 274–281. https://doi.org/10.1055/s-0032-1302442

11. Bolli, G. B., Di Marchi, R. D., Park, G. D., Pramming, S., & Koivisto, V. A. (1999). Insulin analogues and their potential in the management of diabetes mellitus. *Diabetologia*, *42*(10), 1151–1167. https://doi.org/10.1007/s001250051286
12. Ryu, J. K., Kim, H. S., & Nam, D. H. (2012). Current status and perspectives of biopharmaceutical drugs. *Biotechnol. Bioprocess Eng.*, *17*(5), 900–911. https://doi.org/10.1007/s12257-012-0095-1
13. Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B. B., Henner, D., Wong, W. L. T., Rowland, A. M., Kotts, C., Carver, M. E., & Shepard, H. M. (1992). Humanization of an anti-p185^{HER2} antibody for human cancer therapy. *Proc. Natl. Acad. Sci. U. S. A.*, *89*(10), 4285–4289. https://doi.org/10.1073/pnas.89.10.4285
14. Sarkar, C. A., Lowenhaupt, K., Horan, T., Boone, T. C., Tidor, B., & Lauffenburger, D. A. (2002). Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nat. Biotechnol.*, *20*(9), 908–913. https://doi.org/10.1038/nbt725
15. Cantor, J. R., Yoo, T. H., Dixit, A., Iverson, B. L., Forsthuber, T. G., & Georgiou, G. (2011). Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(4), 1272–1277. https://doi.org/10.1073/pnas.1014739108
16. Loosmore, S. M., Zealey, G. R., Boux, H. A., Cockle, S. A., Radika, K., Fahim, R. E., Zobrist, G. J., Yacoob, R. K., Chong, P. C., & Yao, F. L. (1990). Engineering of genetically detoxified pertussis toxin analogs for development of a recombinant whooping cough vaccine. *Infect. Immun.*, *58*(11), 3653–3662. https://doi.org/10.1128/iai.58.11.3653-3662.1990

17. Hsieh, C.-L., Goldsmith, J. A., Schaub, J. M., DiVenere, A. M., Kuo, H.-C., Javanmardi, K., Le, K. C., Wrapp, D., Lee, A. G., Liu, Y., Chou, C.-W., Byrne, P. O., Hjorth, C. K., Johnson, N. V., Ludes-Meyers, J., Nguyen, A. W., Park, J., Wang, N., Amengor, D., … McLellan, J. S. (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science*, *369*(6510), 1501–1505. https://doi.org/10.1126/science.abd0826
18. Padgette, S. R., Kolacz, K. H., Delannay, X., Re, D. B., Tinius, C. N., Rhodes, W. K., Otero, Y. I., Barry, G. F., Eichholtz, D. A., Peschke, V. M., Nida, D. L., Taylor, N. B., & Kishore, G. M. (1995). Development, identification, and characterization of a glyphosate-tolerant soybean line. *Crop Sci.*, *35*(5), 1451–1461. https://doi.org/10.2135/cropsci1995.0011183x003500050032x
19. Castle, L. A., Siehl, D. L., Gorton, R., Patten, P. A., Chen, Y. H., Bertain, S., Cho, H.-J., Duck, N., Wong, J., Liu, D., & Lassner, M. W. (2004). Discovery and directed evolution of a glyphosate tolerance gene. *Science*, *304*(5674), 1151–1154. https://doi.org/10.1126/science.1096770
20. Bravo, A., & Soberon, M. (2008). How to cope with insect resistance to Bt toxins? *Trends Biotechnol.*, *26*(10), 573–579. https://doi.org/10.1016/j.tibtech.2008.06.005
21. Kurek, I., Chang, T. K., Bertain, S. M., Madrigal, A., Liu, L., Lassner, M. W., & Zhu, G. (2007). Enhanced thermostability of *Arabidopsis* rubisco activase improves photosynthesis and growth rates under moderate heat stress. *Plant Cell*, *19*(10), 3230–3241. https://doi.org/10.1105/tpc.107.054171
22. Whittle, E., & Shanklin, J. (2001). Engineering Δ^9-16:0-acyl carrier protein (ACP) desaturase specificity based on combinatorial saturation mutagenesis and logical redesign of the castor Δ^9-18:0-ACP desaturase. *J. Biol. Chem.*, *276*(24), 21500–21505. https://doi.org/10.1074/jbc.M102129200

23. Lin, H., & Lee, Y. K. (2017). Genetic engineering of medium-chain-length fatty acid synthesis in *Dunaliella tertiolecta* for improved biodiesel production. *J. Appl. Phycol.*, *29*(6), 2811–2819. https://doi.org/10.1007/s10811-017-1210-7
24. Yao, M. Z., Zhang, Y. H., Lu, W. L., Hu, M. Q., Wang, W., & Liang, A. H. (2012). Phytases: Crystal structures, protein engineering and potential biotechnological applications. *J. Appl. Microbiol.*, *112*(1), 1–14. https://doi.org/10.1111/j.1365-2672.2011.05181.x
25. Richardson, T. H., Tan, X., Frey, G., Callen, W., Cabell, M., Lam, D., Macomber, J., Short, J. M., Robertson, D. E., & Miller, C. (2002). A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable α-amylase. *J. Biol. Chem.*, *277*(29), 26501–26507. https://doi.org/10.1074/jbc.M203183200
26. Quax, W. J., Mrabet, N. T., Luiten, R. G. M., Schuurhuizen, P. W., Stanssens, P., & Lasters, I. (1991). Enhancing the thermostability of glucose-isomerase by protein engineering. *Bio/Technology*, *9*(8), 738–742. https://doi.org/10.1038/nbt0891-738
27. Bell, E. L., Finnigan, W., France, S. P., Green, A. P., Hayes, M. A., Hepworth, L. J., Lovelock, S. L., Niikura, H., Osuna, S., Romero, E., Ryan, K. S., Turner, N. J., & Flitsch, S. L. (2021). Biocatalysis. *Nat. Rev. Methods Primers*, *1*(1), 46. https://doi.org/10.1038/s43586-021-00044-z
28. van der Wielen, L. A. M., Ottens, M., & Straathof, A. J. J. (2001). Process Technology and Process Integration in the Preparation of Penicillins. In A. Bruggink (Ed.), *Synthesis of β-Lactam antibiotics: Chemistry, biocatalysis & process integration* (pp. 150–205). Springer Netherlands.
29. Ran, N., & Frost, J. W. (2007). Directed evolution of 2-keto-3-deoxy-6-phosphogalactonate aldolase to replace 3-deoxy-D-arabino-heptulosonic acid 7-phosphate synthase. *J. Am. Chem. Soc.*, *129*(19), 6130–6139. https://doi.org/10.1021/ja067330p

30. Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., ... Newman, J. D. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, *496*(7446), 528–532. https://doi.org/10.1038/nature12051
31. Anastas, P. T., & Warner, J. C. (2000). *Green chemistry : Theory and practice*. Oxford University Press.
32. Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J., Brands, J., Devine, P. N., Huisman, G. W., & Hughes, G. J. (2010). Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, *329*(5989), 305–309. https://doi.org/10.1126/science.1188934
33. Stone, E. M., Paley, O., Hu, J., Ekerdt, B., Cheung, N.-K., & Georgiou, G. (2012). De novo engineering of a human cystathionine-γ-lyase for systemic L-methionine depletion cancer therapy. *ACS Chem. Biol.*, *7*(11), 1822–1829. https://doi.org/10.1021/cb300335j
34. Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, *27*(10), 946–950. https://doi.org/10.1038/nbt.1568
35. Jobe, A., & Bourgeois, S. (1972). *lac* Repressor-operator interaction. VI. The natural inducer of the lac operon. *J. Mol. Biol.*, *69*(3), 397–408. https://doi.org/10.1016/0022-2836(72)90253-7
36. Egel, R. (1988). The '*lac*' operon - an irrelevant paradox? *Trends Genet.*, *4*(2), 31–31. https://doi.org/10.1016/0168-9525(88)90062-5
37. Valdez-Cruz, N. A., Caspeta, L., Pérez, N. O., Ramírez, O. T., & Trujillo-Roldán, M. A. (2010). Production of recombinant proteins in *E. Coli* by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microb Cell Fact*, *9*(1), 18. https://doi.org/10.1186/1475-2859-9-18

38. Austin, H. P., Allen, M. D., Donohoe, B. S., Rorrer, N. A., Kearns, F. L., Silveira, R. L., Pollard, B. C., Dominick, G., Duman, R., El Omari, K., Mykhaylyk, V., Wagner, A., Michener, W. E., Amore, A., Skaf, M. S., Crowley, M. F., Thorne, A. W., Johnson, C. W., Woodcock, H. L., ... Beckham, G. T. (2018). Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(19), E4350–E4357. https://doi.org/10.1073/pnas.1718804115
39. Hutchison 3rd, C. A., Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., & Smith, M. (1978). Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem.*, *253*(18), 6551–6560. https://doi.org/10.1016/S0021-9258(19)46967-6
40. Silva, D., Santos, G., Barroca, M., & Collins, T. (2017). Inverse PCR for point mutation introduction. In L. Domingues (Ed.), *PCR: Methods and Protocols* (pp. 87–100). Springer New York. https://doi.org/10.1007/978-1-4939-7060-5_5
41. Roy, S., & Caruthers, M. (2013). Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. *Molecules*, *18*(11), 14268–14284. https://doi.org/10.3390/molecules181114268
42. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S., & Ryadnov, M. G. (2023). DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.*, *7*(3), 144–161. https://doi.org/10.1038/s41570-022-00456-9
43. Seydel, C. (2023). DNA writing technologies moving toward synthetic genomes. *Nat Biotechnol.*, *41*(11), 1504–1509. https://doi.org/10.1038/s41587-023-02006-0
44. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison 3rd, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318
45. Green, M. R., & Sambrook, J. (2012). *Molecular Cloning: A Laboratory Manual* (4th ed.). Cold Spring Harbor Press.

46. Ball, P. (2018). A compassionate substance. How science and myth collide in water. *Lapham's Quart.*, *11*(Summer).
47. Pace, C. N., Scholtz, J. M., & Grimsley, G. R. (2014). Forces stabilizing proteins. *FEBS Lett.*, *588*(14), 2177–2184. https://doi.org/10.1016/j.febslet.2014.05.006
48. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427
49. 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., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
50. Mura, C., McCrimmon, C. M., Vertrees, J., & Sawaya, M. R. (2010). An introduction to biomolecular graphics. *PLoS Comput. Biol.*, *6*(8), e1000918. https://doi.org/10.1371/journal.pcbi.1000918
51. Leach, A. R. (2001). *Molecular modelling: Principles and applications*. Prentice Hall.
52. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz Jr, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, *117*(19), 5179–5197. https://doi.org/10.1021/ja00124a002
53. Allinger, N. L., Zhou, X., & Bergsma, J. (1994). Molecular mechanics parameters. *J. Mol. Struct. THEOCHEM*, *312*(1), 69–83. https://doi.org/10.1016/S0166-1280(09)80008-0

54. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem\@. Theory Comput.*, *13*(6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125
55. Homeyer, N., & Gohlke, H. (2012). Free energy calculations by the molecular mechanics Poisson-Boltzmann surface area method. *Mol. Inf.*, *31*(2), 114–122. https://doi.org/10.1002/minf.201100135
56. Bae, E., & Phillips Jr, G. N. (2005). Identifying and engineering ion pairs in adenylate kinases. *J. Biol. Chem.*, *280*(35), 30943–30948. https://doi.org/10.1074/jbc.M504216200
57. Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J., & Osuna, S. (2019). Hidden conformations in *Aspergillus niger* monoamine oxidase are key for catalytic efficiency. *Angew. Chem. Int. Ed.*, *58*(10), 3097–3101. https://doi.org/10.1002/anie.201812532
58. Thomas, L. L., Christakis, T. J., & Jorgensen, W. L. (2006). Conformation of alkanes in the gas phase and pure liquids. *J. Phys. Chem. B*, *110*(42), 21198–21204. https://doi.org/10.1021/jp064811m
59. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N., & Hilvert, D. (2013). Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.*, *9*(8), 494–498. https://doi.org/10.1038/nchembio.1276
60. Crean, R. M., Gardner, J. M., & Kamerlin, S. C. L. (2020). Harnessing conformational plasticity to generate designer enzymes. *J. Am. Chem. Soc.*, *142*(26), 11324–11342. https://doi.org/10.1021/jacs.0c04924
61. Robic, S., Guzman-Casado, M., Sanchez-Ruiz, J. M., & Marqusee, S. (2003). Role of residual structure in the unfolded state of a thermophilic protein. *Proc. Natl. Acad. Sci. U. S. A.*, *100*(20), 11345–11349. https://doi.org/10.1073/pnas.1635051100

62. Rosenthal, L., Rabolt, J. F., & Hummel, J. (1982). An investigation of the conformational equilibrium of *n*-butane in a solvent using Raman spectroscopy. *J. Chem. Phys.*, *76*(2), 817–820. https://doi.org/10.1063/1.443052
63. Vázquez-Figueroa, E., Yeh, V., Broering, J. M., Chaparro-Riggers, J. F., & Bommarius, A. S. (2008). Thermostable variants constructed via the structure-guided consensus method also show increased stability in salts solutions and homogeneous aqueous-organic media. *Protein Eng. Des. Sel.*, *21*(11), 673–680. https://doi.org/10.1093/protein/gzn048
64. Mayer, S., Rüdiger, S., Ang, H. C., Joerger, A. C., & Fersht, A. R. (2007). Correlation of levels of folded recombinant P53 in *Escherichia coli* with thermodynamic stability in vitro. *J. Mol. Biol.*, *372*(1), 268–276. https://doi.org/10.1016/j.jmb.2007.06.044
65. Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, *103*(15), 5869–5874. https://doi.org/10.1073/pnas.0510098103
66. Adams, M. W., & Kelly, R. M. (1998). Finding and using hyperthermophilic enzymes. *Trends Biotechnol.*, *16*(8), 329–332. https://doi.org/10.1016/s0167-7799(98)01193-7
67. Merz, A., Yee, M.-C., Szadkowski, H., Pappenberger, G., Crameri, A., Stemmer, W. P. C., Yanofsky, C., & Kirschner, K. (2000). Improving the catalytic activity of a thermophilic enzyme at low temperatures. *Biochemistry*, *39*(5), 880–889. https://doi.org/10.1007/pl00003701
68. Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.*, *17*(4), 263–270. https://doi.org/10.1002/humu.22
69. Joerger, A. C., Ang, H. C., & Fersht, A. R. (2006). Structural basis for understanding oncogenic P53 mutations and designing rescue drugs. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(41), 15056–15061. https://doi.org/10.1073/pnas.0607286103

70. Park, S. H., & Bao, G. (2021). CRISPR/Cas9 gene editing for curing sickle cell disease. *Transfusion and Apheresis Science*, *60*(1), 103060. https://doi.org/10.1016/j.transci.2021.103060
71. Pace, C. N., & Shaw, K. L. (2000). Linear extrapolation method of analyzing solvent denaturation curves. *Proteins Struct. Funct. Bioinforma.*, *Suppl 4*, 1–7. https://doi.org/10.1002/1097-0134(2000)41:4+%3C1::aid-prot10%3E3.3.co;2-u
72. Baxa, M. C., Haddadian, E. J., Jumper, J. M., Freed, K. F., & Sosnick, T. R. (2014). Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(43), 15396–15401. https://doi.org/10.1073/pnas.1407768111
73. Horovitz, A., & Fersht, A. R. (1992). Co-operative interactions during protein folding. *J. Mol. Biol.*, *224*(3), 733–740. https://doi.org/10.1016/0022-2836(92)90557-Z
74. Scholtz, J. M., Grimsley, G. R., & Pace, C. N. (2009). Solvent denaturation of proteins and interpretations of the m value. *Methods Enzymol.*, *466*, 549–565. https://doi.org/10.1016/S0076-6879(09)66023-7
75. Fersht, A. (1999). Chapter 17, Protein Stability. In *Structure and Mechanism in Protein Science*. Freeman.
76. Becktel, W. J., & Schellman, J. A. (1987). Protein stability curves. *Biopolymers*, *26*(11), 1859–1878. https://doi.org/10.1002/bip.360261104
77. Robertson, A. D., & Murphy, K. P. (1997). Protein structure and the energetics of protein stability. *Chem. Rev.*, *97*(5), 1251–1267.
78. Pucci, F., Kwasigroch, J. M., & Rooman, M. (2017). SCooP: An accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*, *33*(21), 3415–3422.

79. Pastore, A., Martin, S. R., Politou, A., Kondapalli, K. C., Stemmler, T., & Temussi, P. A. (2007). Unbiased cold denaturation: Low- and high-temperature unfolding of yeast frataxin under physiological conditions. *J. Am. Chem. Soc.*, *129*(17), 5374–5375. https://doi.org/10.1021/ja0714538
80. Razvi, A., & Scholtz, J. M. (2006). Lessons in stability from thermophilic proteins. *Protein Sci.*, *15*(7), 1569–1578. https://doi.org/10.1110/ps.062130306
81. Eijsink, V. G. H., Bjørk, A., G\a aseidnes, S., Sirev\a ag, R., Synstad, B., van den Burg, B., & Vriend, G. (2004). Rational engineering of enzyme stability. *J. Biotechnol.*, *113*(1-3), 105–120. https://doi.org/10.1016/j.jbiotec.2004.03.026
82. Dumon, C., Varvak, A., Wall, M. A., Flint, J. E., Lewis, R. J., Lakey, J. H., Morland, C., Luginbühl, P., Healey, S., Todaro, T., DeSantis, G., Sun, M., Parra-Gessert, L., Tan, X., Weiner, D. P., & Gilbert, H. J. (2008). Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure. *J. Biol. Chem.*, *283*(33), 22557–22564. https://doi.org/10.1074/jbc.M800936200
83. Shoichet, B. K., Baase, W. A., Kuroki, R., & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.*, *92*(2), 452–456. https://doi.org/10.1073/pnas.92.2.452
84. Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L., van Loon, A. P. G. M., & Wyss, M. (2002). The consensus concept for thermostability engineering of proteins: Further proof of concept. *Protein Eng.*, *15*(5), 403–411. https://doi.org/10.1093/protein/15.5.403
85. Magliery, T. J. (2015). Protein stability: Computation, sequence statistics, and new experimental methods. *Current Opinion in Structural Biology*, *33*, 161–168. https://doi.org/10.1016/j.sbi.2015.09.002

86. Jones, B. J., Kan, C. N. E., Luo, C., & Kazlauskas, R. J. (2020). Consensus Finder web tool to predict stabilizing substitutions in proteins. In D. S. Tawfik (Ed.), *Enzyme Engineering and Evolution: General Methods* (Vol. 643, pp. 129–148). Academic Press. https://doi.org/10.1016/bs.mie.2020.07.010
87. Musil, M., Stourac, J., Bendl, J., Brezovsky, J., Prokop, Z., Zendulka, J., Martinek, T., Bednar, D., & Damborsky, J. (2017). FireProt: Web server for automated design of thermostable proteins. *Nucleic Acids Res.*, *45*(W1), W393–W399. https://doi.org/10.1093/nar/gkx285
88. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L., & Bryan, P. N. (1989). Large increases in general stability for subtilisin BPN' through incremental changes in the free-energy of unfolding. *Biochemistry*, *28*(18), 7205–7213. https://doi.org/10.1021/bi00444a012
89. Schwehm, J. M., Fitch, C. A., Dang, B. N., García-Moreno E., B., & Stites, W. E. (2003). Changes in stability upon charge reversal and neutralization substitution in staphylococcal nuclease are dominated by favorable electrostatic effects. *Biochemistry*, *42*(4), 1118–1128. https://doi.org/10.1021/bi0266434
90. Trudeau, D. L., Kaltenbach, M., & Tawfik, D. S. (2016). On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.*, *33*(10), 2633–2641. https://doi.org/10.1093/molbev/msw138
91. Clarke, J., Henrick, K., & Fersht, A. R. (1995). Disulfide mutants of barnase I: Changes in stability and structure assessed by biophysical methods and x-ray crystallography. *J. Mol. Biol.*, *253*(3), 493–504. https://doi.org/10.1006/jmbi.1995.0568
92. Pace, C. N., Grimsley, G. R., Thomson, J. A., & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.*, *263*(24), 11820–11825.

93. Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.*, *151*(2), 261–287. https://doi.org/10.1016/0022-2836(81)90515-5
94. Perry, L. J., & Wetzel, R. (1984). Disulfide bond engineered into T4 lysozyme: Stabilization of the protein toward thermal inactivation. *Science*, *226*(4674), 555–557. https://www.jstor.org/stable/1693872
95. Matsumura, M., Becktel, W. J., Levitt, M., & Matthews, B. W. (1989). Stabilization of phage T4 lysozyme by engineered disulfide bonds. *Proc. Natl. Acad. Sci. U. S. A.*, *86*, 6562–6566. https://doi.org/10.1073/pnas.86.17.6562
96. Roesler, K. R., & Rao, A. G. (2000). A single disulfide bond restores thermodynamic and proteolytic stability to an extensively mutated protein. *Protein Science*, *9*(9), 1642–1650. https://doi.org/10.1110/ps.9.9.1642
97. Kanaya, S., Katsuda, C., Kimura, S., Nakai, T., Kitakuni, E., Nakamura, H., Katayanagi, K., Morikawa, K., & Ikehara, M. (1991). Stabilization of *Escherichia coli* ribonuclease H by introduction of an artificial disulfide bond. *Journal of Biological Chemistry*, *266*(10), 6038–6044. https://doi.org/10.1016/S0021-9258(18)38080-3
98. Hazes, B., & Dijkstra, B. W. (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng. Des. Sel.*, *2*(2), 119–125. https://doi.org/10.1093/protein/2.2.119
99. Dombkowski, A. A. (2003). Disulfide by Design^{TM}: A computational method for the rational design of disulfide bonds in proteins. *Bioinformatics*, *19*(14), 1852–1853. https://doi.org/10.1093/bioinformatics/btg231
100. Suplatov, D., Timonina, D., Sharapova, Y., & Švedas, V. (2019). Yosshi: A web-server for disulfide engineering by bioinformatic analysis of diverse protein families. *Nucleic Acids Res.*, *47*(W1), W308–W314. https://doi.org/10.1093/nar/gkz385

101. Pikkemaat, M. G., Linssen, A. B. M., Berendsen, H. J. C., & Janssen, D. B. (2002). Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng.*, *15*(3), 185–192. https://doi.org/10.1093/protein/15.3.185
102. Stewart, E. J., Åslund, F., & Beckwith, J. (1998). Disulfide bond formation in the *Escherichia coli* cytoplasm: An *in Vivo* role reversal for the thioredoxins. *EMBO J.*, *17*(19), 5543–5550. https://doi.org/10.1093/emboj/17.19.5543
103. Zhou, H.-X. (2004). Loops, linkages, rings, catenanes, cages, and crowders: Entropy-based strategies for stabilizing proteins. *Acc. Chem. Res.*, *37*(2), 123–130. https://doi.org/10.1021/ar0302282
104. Watanabe, K., & Suzuki, Y. (1998). Protein thermostabilization by proline substitutions. *J. Mol. Catal. B Enzym.*, *4*(4), 167–180. https://doi.org/10.1016/s1381-1177(97)00031-3
105. Fu, H., Grimsley, G. R., Razvi, A., Scholtz, J. M., & Pace, C. N. (2009). Increasing protein stability by improving beta-turns. *Proteins*, *77*(3), 491–498. https://doi.org/10.1002/prot.22509
106. Matthews, B. W., Nicholson, H., & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U. S. A.*, *84*(19), 6663–6667. https://doi.org/10.1073/pnas.84.19.6663
107. Scott, K. A., Alonso, D. O. V., Sato, S., Fersht, A. R., & Daggett, V. (2007). Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(8), 2661–2666. https://doi.org/10.1073/pnas.0611182104
108. Liu, Z., Lemmonds, S., Huang, J., Tyagi, M., Hong, L., & Jain, N. (2018). Entropic contribution to enhanced thermal stability in the thermostable P450 CYP119. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(43), E10049–E10058. https://doi.org/10.1073/pnas.1807473115
109. 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*(46), 7745–7751. https://doi.org/10.1002/anie.200602795

110. Floor, R. J., Wijma, H. J., Colpa, D. I., Ramos-Silva, A., Jekel, P. A., Szymanski, W., Feringa, B. L., Marrink, S. J., & Janssen, D. B. (2014). Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem*, *15*(11), 1660–1672. https://doi.org/10.1002/cbic.201402128
111. Xiao, Z., Bergeron, H., Grosse, S., Beauchemin, M., Garron, M.-L., Shaya, D., Sulea, T., Cygler, M., & Lau, P. C. K. (2008). Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-temperature-guided sequence alignment. *Appl. Environ. Microbiol.*, *74*(4), 1183–1189. https://doi.org/10.1128/AEM.02220-07
112. Petsko, G. A. (2001). Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat." *Methods Enzymol.*, *334*, 469–478. https://doi.org/10.1016/S0076-6879(01)34486-5
113. Gribenko, A. V., Patel, M., Liu, J., McCallum, S., Wang, C., & Makhatadze, G. I. (2009). Rational stabilization of enzymes by computational redesign of surface charge–charge interactions. *Proc. Natl. Acad. Sci. U. S. A.*, *106*(8), 2601–2606. https://doi.org/10.1073/pnas.0808220106
114. Contessoto, V. G., de Oliveira, V. M., Fernandes, B. R., Slade, G. G., & Leite, V. B. P. (2018). TKSA-MC: A web server for rational mutation through the optimization of protein charge interactions. *Proteins*, *86*(11), 1184–1188. https://doi.org/10.1002/prot.25599
115. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987–2998. https://doi.org/10.1021/bi902153g
116. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Res.*, *33*(Web Server issue), W382–8. https://doi.org/10.1093/nar/gki387

117. Korkegian, A., Black, M. E., Baker, D., & Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. *Sci. Wash.*, *308*(5723), 857–860. https://doi.org/10.1126/science.1107387
118. Arabnejad, H., Dal Lago, M., Jekel, P. A., Floor, R. J., Thunnissen, A.-M. W. H., Terwisscha van Scheltinga, A. C., Wijma, H. J., & Janssen, D. B. (2017). A robust cosolvent-compatible halohydrin dehalogenase by computational library design. *Protein Eng. Des. Sel.*, *30*(3), 173–187. https://doi.org/10.1093/protein/gzw068
119. Jin, Y., Johannissen, L. O., & Hay, S. (2021). Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins*, *89*(8), 915–921. https://doi.org/10.1002/prot.26068
120. Weinstein, J. J., Goldenzweig, A., Hoch, S., & Fleishman, S. J. (2021). PROSS 2: A new server for the design of stable and highly expressed protein variants. *Bioinformatics*, *37*(1), 123–125. https://doi.org/10.1093/bioinformatics/btaa1071
121. Gao, D., Narasimhan, D. L., Macdonald, J., Brim, R., Ko, M.-C., Landry, D. W., Woods, J. H., Sunahara, R. K., & Zhan, C.-G. (2009). Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, *75*(2), 318–323. https://doi.org/10.1124/mol.108.049486
122. van den Burg, B., Vriend, G., Veltman, O. R., & Eijsink, V. G. H. (1998). Engineering an enzyme to resist boiling. *Proc. Natl. Acad. Sci. U. S. A.*, *95*(5), 2056–2060. https://doi.org/10.1073/pnas.95.5.2056
123. Bjørk, A., Dalhus, B., Mantzilas, D., Sirev\a ag, R., & Eijsink, V. G. H. (2004). Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface. *J. Mol. Biol.*, *341*(5), 1215–1226. https://doi.org/10.1016/j.jmb.2004.06.079

124. Kabashima, T., Li, Y., Kanada, N., Ito, K., & Yoshimoto, T. (2001). Enhancement of the thermal stability of pyroglutamyl peptidase I by introduction of an intersubunit disulfide bond. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1547*(2), 214–220. https://doi.org/10.1016/s0167-4838(01)00185-6
125. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., & Trout, B. L. (2010). Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B*, *114*(19), 6614–6624. https://doi.org/10.1021/jp911706q
126. Kuriata, A., Iglesias, V., Pujols, J., Kurcinski, M., Kmiecik, S., & Ventura, S. (2019). Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. *Nucleic Acids Res.*, *47*(W1), W300–W307. https://doi.org/10.1093/nar/gkz321
127. Van Durme, J., De Baets, G., Van Der Kant, R., Ramakers, M., Ganesan, A., Wilkinson, H., Gallardo, R., Rousseau, F., & Schymkowitz, J. (2016). Solubis: A webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel.*, *29*(8), 285–289. https://doi.org/10.1093/protein/gzw019
128. Suen, W.-C., Zhang, N., Xiao, L., Madison, V., & Zaks, A. (2004). Improved activity and thermostability of *Candida antarctica* lipase B by DNA family shuffling. *Protein Eng. Des. Sel.*, *17*(2), 133–140. https://doi.org/10.1093/protein/gzh017
129. Lawrence, M. S., Phillips, K. J., & Liu, D. R. (2007). Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.*, *129*(33), 10110–10112. https://doi.org/10.1021/ja071641y
130. Robinson, N. E. (2002). Protein deamidation. *Proc. Natl. Acad. Sci. U. S. A.*, *99*(8), 5283–5288. https://doi.org/10.1073/pnas.082102799
131. Mark, D. F., Lu, S. D., Creasey, A. A., Yamamoto, R., & Lin, L. S. (1984). Site-specific mutagenesis of the human fibroblast interferon gene. *Proc. Natl. Acad. Sci. U. S. A.*, *81*(18), 5662–5666.

132. Kontermann, R. E. (2009). Strategies to extend plasma half-lives of recombinant antibodies. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *23*(2), 93–109. https://doi.org/10.2165/00063030-200923020-00003
133. Strohl, W. R. (2015). Fusion proteins for half-life extension of biologics as a strategy to make biobetters. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *29*(4), 215–239. https://doi.org/10.1007/s40259-015-0133-6
134. Cheang, J. Y., & Moyle, P. M. (2018). Glucagon-like peptide-1 (GLP-1)-based therapeutics: Current status and future opportunities beyond type 2 diabetes. *ChemMedChem*, *13*(7), 662–671. https://doi.org/10.1002/cmdc.201700781
135. Feldmann, M. (2002). Development of anti-TNF therapy for rheumatoid arthritis. *Nat. Rev. Immunol.*, *2*(5), 364–371. https://doi.org/10.1038/nri802
136. Nielsen, J. E., & Borchert, T. V. (2000). Protein engineering of bacterial α-amylases. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1543*(2), 253–274. https://doi.org/10.1016/s0167-4838(00)00240-5
137. Jones, B. J., Lim, H. Y., Huang, J., & Kazlauskas, R. J. (2017). Comparison of five protein engineering strategies for stabilizing an α/β-hydrolase. *Biochemistry*, *56*(50), 6521–6532. https://doi.org/10.1021/acs.biochem.7b00571
138. Thieker, D. F., Maguire, J. B., Kudlacek, S. T., Leaver-Fay, A., Lyskov, S., & Kuhlman, B. (2022). Stabilizing proteins, simplified: A Rosetta-based webtool for predicting favorable mutations. *Protein Sci.*, *31*(10). https://doi.org/10.1002/pro.4428
139. Banta, S., Dooley, K., & Shur, O. (2013). Replacing antibodies: Engineering new binding proteins. *Annu. Rev. Biomed. Eng.*, *15*(1), 93–113. https://doi.org/10.1146/annurev-bioeng-071812-152412
140. Rossi, A. M., & Taylor, C. W. (2011). Analysis of protein-ligand interactions by fluorescence polarization. *Nat Protoc*, *6*(3), 365–387. https://doi.org/10.1038/nprot.2011.305

141. Pollard, T. D. (2010). MBOC technical perspective: A guide to simple and informative binding assays. *Mol. Biol. Cell*, *21*(23), 4061–4067. https://doi.org/10.1091/mbc.E10-08-0683
142. Jeffrey, P. D., Strong, R. K., Sieker, L. C., Chang, C. Y. Y., Campbell, R. L., Petsko, G. A., Haber, E., Margolies, M. N., & Sheriff, S. (1993). 26-10 Fab-digoxin complex: Affinity and specificity due to surface complementarity. *Proc. Natl. Acad. Sci. U. S. A.*, *90*(21), 10310–10314. https://doi.org/10.1073/pnas.90.21.10310
143. Tian, W., Chen, C., Lei, X., Zhao, J., & Liang, J. (2018). CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.*, *46*(W1), W363–W367. https://doi.org/10.1093/nar/gky473
144. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, *255*(5041), 178–183. https://doi.org/10.1126/science.1553543
145. Jones, B. J., Evans, R. L., Mylrea, N. J., Chaudhury, D., Luo, C., Guan, B., Pierce, C. T., Gordon, W. R., Wilmot, C. M., & Kazlauskas, R. J. (2020). Larger active site in an ancestral hydroxynitrile lyase increases catalytically promiscuous esterase activity. *PLoS ONE*, *15*, e0235341. https://doi.org/10.1371/journal.pone.0235341
146. DeLorbe, J. E., Clements, J. H., Teresk, M. G., Benfield, A. P., Plake, H. R., Millspaugh, L. E., & Martin, S. F. (2009). Thermodynamic and structural effects of conformational constraints in protein-ligand interactions. Entropic paradoxy associated with ligand preorganization. *J. Am. Chem. Soc.*, *131*(46), 16758–16770. https://doi.org/10.1021/ja904698q
147. Barros, E. P., Schiffer, J. M., Vorobieva, A., Dou, J., Baker, D., & Amaro, R. E. (2019). Improving the efficiency of ligand-binding protein design with molecular dynamics simulations. *J. Chem. Theory Comput.*, *15*(10), 5703–5715. https://doi.org/10.1021/acs.jctc.9b00483

148. Thakkar, S., Nanaware-Kharade, N., Celikel, R., Peterson, E. C., & Varughese, K. I. (2014). Affinity improvement of a therapeutic antibody to methamphetamine and amphetamine through structure-based antibody engineering. *Sci. Rep.*, *4*, 3673. https://doi.org/10.1038/srep03673
149. Zhao, Q., Ahmed, M., Guo, H., Cheung, I. Y., & Cheung, N.-K. V. (2015). Alteration of electrostatic surface potential enhances affinity and tumor killing properties of anti-ganglioside GD2 monoclonal antibody hu3F8. *J. Biol. Chem.*, *290*(21), 13017–13027. https://doi.org/10.1074/jbc.M115.650903
150. Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, *25*(10), 1171–1176. https://doi.org/10.1038/nbt1336
151. Selzer, T., Albeck, S., & Schreiber, G. (2000). Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.*, *7*(7), 537–541. https://doi.org/10.1038/76744
152. Bugnon, M., Röhrig, U. F., Goullieux, M., Perez, M. A. S., Daina, A., Michielin, O., & Zoete, V. (2024). SwissDock 2024: Major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. *Nucleic Acids Res*, *52*(W1), W324–W332. https://doi.org/10.1093/nar/gkae300
153. Zoete, V., Schuepbach, T., Bovigny, C., Chaskar, P., Daina, A., Röhrig, U. F., & Michielin, O. (2016). Attracting cavities for docking. Replacing the rough energy landscape of the protein by a smooth attracting landscape. *J Comput Chem*, *37*(4), 437–447. https://doi.org/10.1002/jcc.24249
154. Röhrig, U. F., Goullieux, M., Bugnon, M., & Zoete, V. (2023). Attracting Cavities 2.0: Improving the flexibility and robustness for small-molecule docking. *J. Chem. Inf. Model.*, *63*(12), 3925–3940. https://doi.org/10.1021/acs.jcim.3c00054

155. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., & Baker, D. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, *501*(7466), 212–216. https://doi.org/10.1038/nature12443

156. Morin, A., Kaufmann, K. W., Fortenberry, C., Harp, J. M., Mizoue, L. S., & Meiler, J. (2011). Computational design of an endo-1,4-beta-xylanase ligand binding site. *Protein Eng. Des. Sel.*, *24*(6), 503–516. https://doi.org/10.1093/protein/gzr006

Problems

1. To increase the binding of an antibody to its target, do you need to increase or decrease the value of ΔG_{diss}? How large must the change be in kcal/mol for the binding to increase by a factor of one hundred? Explain whether the change is a ΔG_{diss}, a $\Delta\Delta G_{diss}$ or a $\Delta\Delta\Delta G_{diss}$.

2. Two stable conformations of butane are the anti and gauche conformation. The gauche conformation exists as a pair of enantiomers. Explain how entropy favors the gauche conformation over the anti conformation. How large, in kcal/mol, is this entropy contribution?

3. Draw resonance structures for the deprotonated form of imidazole (the imidazolide anion). Draw an arrow-pushing reaction mechanism for the tautomerization of histidine.

4. Draw the structure of amino acid His showing the correct absolute configuration and the correct protonation state at pH 7 for any acidic or basic functional groups. The pK_a of a carboxylic acids like acetic acid is ~5, but the pK_a of the carboxylic acid in free amino acids is ~2. Rationalize this difference in pK_a.

5. *Optional.* Use the Python script below to calculate the change in Gibbs energy associated with an increase in reaction rate for an enzyme variant

as compared to the wild-type enzyme. To use the script, copy the text below and save it as a text file named GibbsEnergyChange.py on your computer Desktop, then follow the instructions below.

```
#!/usr/bin/env python3
'''
This script calculates the Gibbs energy change in kcal/mol associated
with the improvement of the reaction rate of an enzyme.

To use the script:
1. place the script on your Desktop
2. open a terminal window
3. change the working directory to the Desktop by typing at the prompt:
   cd ~/Desktop
4. in the terminal type: python3 GibbsEnergyChange.py
5. enter the rates for the wild type and variant protein when prompted.

To test the script enter 1 for the rate of wt enzyme and 10 for the
rate of the variant enzyme; the output should be -1.4 kcal/mol.
The negative value indicates that the transition state is lower in
the variant, so the reaction is faster.

'''
#load math function needed to calculate logarithms
import math

# define R in units of cal/mol*K
# to change the units of R to J/mol*K, replace this value
# with 8.314 and also change the units in the print
# statement below to kJ/mol
R = 1.987

# temperature of the comparison in units of degrees Kelvin
T = 298
```

```
# asks for user input
rate_wt = float(input("What is the rate of the wt enzyme?"))
rate_variant = float(input("What is the rate of the mutant enzyme?"))

# calculates Gibbs energy change
def GibbsEnergyChange(rate_wt,rate_variant):
    return (- R*T*math.log(rate_variant / rate_wt))/1000.0

# print result
print('The mutation changed the free energy of the transition '
      'state by %.1f kcal/mol.' % (GibbsEnergyChange(rate_wt,
      rate_variant)))
```

Answers

1. For a 100-fold increase in binding, the K_d of the new complex should be 100-fold lower (less favorable) than the starting complex. The value of ΔG_{diss} should be higher. The relationship between free energy change and equilibrium constant for bound/unbound is $\Delta G_{diss} = -RT \ln(K_{diss})$

To compare two complexes: $\Delta\Delta G_{diss}(newvsold) = -RT \ln(K_{diss}(new)/K_{diss}(old))$

For a 100-fold improvement in binding:

$\Delta\Delta G_{diss}(newvsold) = -RT \ln((x/100)/x)$ where x is $K_{diss}(old)$

$\Delta\Delta G_{diss}(newvsold) = -RT \ln(1/100) = +RT \ln(100)$

At 298K: $\Delta\Delta G_{diss} = +(1.987 \times 10^{-3}\ \text{kcal/mol/K})(298\ \text{K})(4.605) = +2.73$ kcal/mol

2. Entropy favors the gauche conformation because it has more accessible microstates (two enantiomeric conformations (gauche+ and gauche-))

than the anti conformation (one conformation). The contribution of microstates to entropy is given by $S = R\ln(\Omega)$ where Ω is the number of accessible microstates. For the reaction $anti \rightleftharpoons gauche$, the difference in entropy is,

$$\Delta S_{(\text{g vs. a})} = R\ln(\Omega_g/\Omega_a) = R\ln(2/1) = R\ln(2)$$

The free energy contribution from this entropy term is: $\Delta G_{(\text{g vs. a})} = -T\Delta S = -RT\ln(2)$

At room temperature (298K): $\Delta G_{(\text{g vs. a})} = -(1.987\times10^{-3}\ \text{kcal/mol/K})(298\text{K})(0.693) = -0.41$ kcal/mol

So entropy favors the gauche conformation by approximately 0.41 kcal/mol due to the availability of two equivalent conformations versus one for the anti form. This entropic advantage is why the gauche form, despite potential steric interactions, is significantly populated at room temperature.

3.

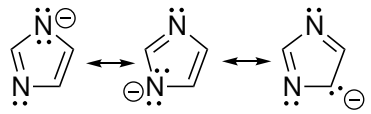


Figure 3.12. Resonance structures of the imidazolide anion.

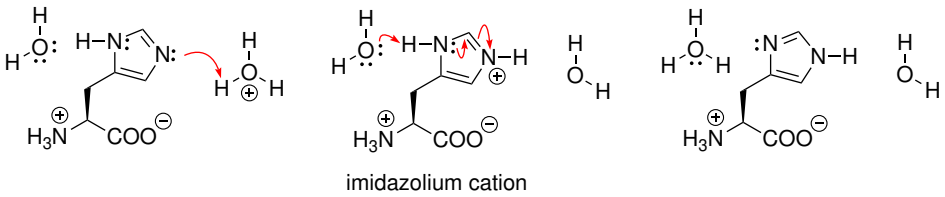


Figure 3.13. The tautomerization of histidine via the immidazolium cation is the most likely mechanism near neutral pH. The alternative mechanism via the imidazolide anion is unfavorable because the anion is unlikely to form near neutral pH.

4. The 3-unit difference in pK_a (5 → 2) is a 1000-fold increase in acidity of the histidine carboxylate as compared to that in acetic acid. This increase is due to the neighboring ammonium cation, which stabilizes the carboxylate form by electrostatics and to a lesser extent by inductive effects. At pH 7, the predominant form of histidine is COO^- (deprotonated), NH_3^+ (protonated), imidazole ring: neutral (its pK_a is 6.5); net charge = 0, it has the (S) absolute configuration at the α-carbon.

4 Protein Structure and Molecular Modeling

Summary. Protein structures show the fold of the protein, which amino acid residues interact with each other and the shape of cavities and tunnels. Protein structure files specify the locations (x, y, z coordinates) of the atoms in a protein as well as any bound molecules and solvent. Protein visualization programs such as PyMOL create images from these files. Different representations highlight different aspects of these intricate structures including shape, electrostatic charge and flexibility. Computer modeling programs such as AMBER or Rosetta predict protein structures using simplified mechanical models of molecules. Accurately modeling protein properties remains difficult because it is difficult to model protein states. Protein states are collections of interconverting conformations that are accessible to a particular protein form. Molecular dynamics simulations reveal how proteins move on short time scales, but finding a representative sample of all the conformations that contribute to a state is difficult. In practice, researchers use approximations to model protein states, which lowers the accuracy of the prediction of protein properties.

Key learning goals

- Protein structure files are text files containing the x,y,z coordinates of the protein atoms as well as other structural information.

- Protein visualization software displays the data in protein structure files. This software simplifies the complexity of protein structures by using different representations for different parts of the protein. This software can also measure distances between atoms within a protein and predict structures of variants with amino acid substitutions.
- Computer modeling of proteins uses simplified mechanical models of molecules called force fields to calculate the energy of a structure. Most predictions of protein structures from their amino acid sequence are extrapolations from known structures. These extrapolations include machine learning models and correctly predict most protein folds and are increasingly accurate in predicting side chain orientations.
- Protein move constantly and all of the accessible protein conformations contribute to the protein states. Gibbs energy differences between protein states determine protein properties.
- Protein states are difficult to model because important conformations may be difficult to find. For example, an open conformation important for catalysis may occur rarely or not at all in molecular dynamics simulations because the conformation requires simultaneous orientation of multiple bonds.
- Most predictions rely on approximations to model protein states. Their accuracy varies according to the validity of these approximations.

4.1 Protein structure

4.1.1 Determining protein structures

X-ray crystallography determines the structure of proteins by measuring the electron density distribution within protein crystals, Figure 4.1. Fit-

ting a model of the known amino acid sequence of the protein to this electron density data yields the structure.

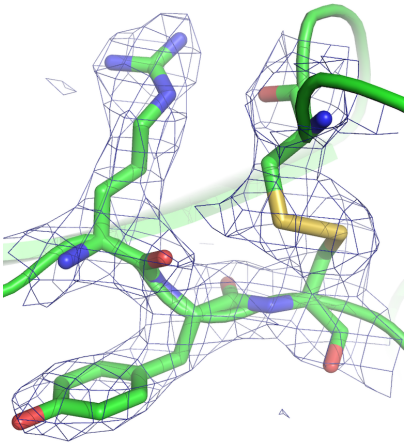


Figure 4.1. Measuring electron density reveals a protein's structure. The mesh diagram represents the observed electron density in an x-ray crystallography experiment for four amino acids in a protein. The mesh is contoured at 1σ; that is, it encloses the regions that contain more than one standard deviation above the mean electron density. The line diagram is the best fit of the atoms in the known amino acid sequence (arginine, tyrosine, and two cysteines linked by a disulfide bond) to the observed electron density. Image from: https://en.wikipedia.org/wiki/file:example_of_electron_density_map.png; CC BY-SA 3.0 license.

One can also determine the structures of proteins by NMR (nuclear magnetic resonance). NMR measures the distances between various nuclei in the protein. Fitting a model of the known amino acid sequence to these distances yields a set of matching structures with differing conformations. While x-ray structures typically yield a single structure, NMR structures yield a collection of similar structures that represent confor-

mations explored by the protein.

Homology modeling is a computational approach to predict protein structures. Extrapolation from a known protein structure (the template) yields predicted structures for homologous proteins. If the two proteins share >80% identical amino acids, then the model is typically within 1-2 Å root mean square deviation of the correct structure. For more distant homologs, the reliability depends on the degree of sequence identity, the choice of the best template, and the alignment of the template and target protein sequences. Web tools such as SwissModel[48] and the machine learning model AlphaFold[49] automate this extrapolation.

4.1.2 Structure data files

One way to describe the structure of a molecule is to specify the location of each atom using Cartesian coordinates (x,y,z). The example structure file for aspirin lists these coordinates as well as some additional information, Figure 4.2. The first line states the total number of atoms. Lines 2-22 list the information for each atom: the element, serial number, (x,y,z) coordinates, atom type number (see molecular mechanics in Section 4.2.1). The final columns list the other atoms bonded to the atom. For example, line 2 describes carbon atom numbered 1 located at x = –1.573, y = 0.146 and z = –0.7046. The units are Ångstroms. It is a carbon atom type 2 (sp^2-carbon, alkene) and it makes bonds to the atoms numbered 2, 6, and 12. Upon rotation of the molecule, these (x, y, z) coordinates change.

An alternative file format is internal coordinates or a Z-matrix (not shown). The position of the first atom is arbitrary. Subsequent atom positions are indicated relative to this first atom by specifying measurements (distance, bond angle, dihedral angle) relative to previous atoms. The advantage of this format is that the coordinates do not change when the molecule rotates.

```
21
  C    1   -1.573105    0.146286   -0.704605    2    2    6   12
  C    2   -2.248077   -1.036957   -1.043121    2    1    3    7
  C    3   -1.549103   -2.223541   -1.264496    2    2    4    8
  C    4   -0.161728   -2.223724   -1.153229    2    3    5    9
  C    5    0.504456   -1.045914   -0.804138    2    4    6   10
  C    6   -0.167023    0.158997   -0.554840    2    1    5   11
  H    7   -3.343964   -1.055359   -1.158142    5    2
  H    8   -2.085190   -3.146332   -1.540329    5    3
  H    9    0.406052   -3.148911   -1.342590    5    4
  H   10    1.604492   -1.086975   -0.736557    5    5
  O   11    0.508545    1.292389   -0.190369    6    6   16
  C   12   -2.393173    1.382111   -0.517700    3    1   13   14
  O   13   -3.599030    1.262833   -0.317047    7   12
  O   14   -1.863129    2.626800   -0.546539    6   12   15
  H   15   -0.953934    2.553543   -0.871445   24   14
  C   16    1.617218    1.154007    0.589493    3   11   17   18
  C   17    2.717529    2.093964    0.139771    1   16   19   20   21
  O   18    1.691452    0.407181    1.540344    7   16
  H   19    3.599045    2.013824    0.815094    5   17
  H   20    2.360992    3.148911    0.162857    5   17
  H   21    3.047653    1.844300   -0.893768    5   17
```

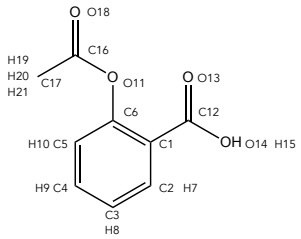


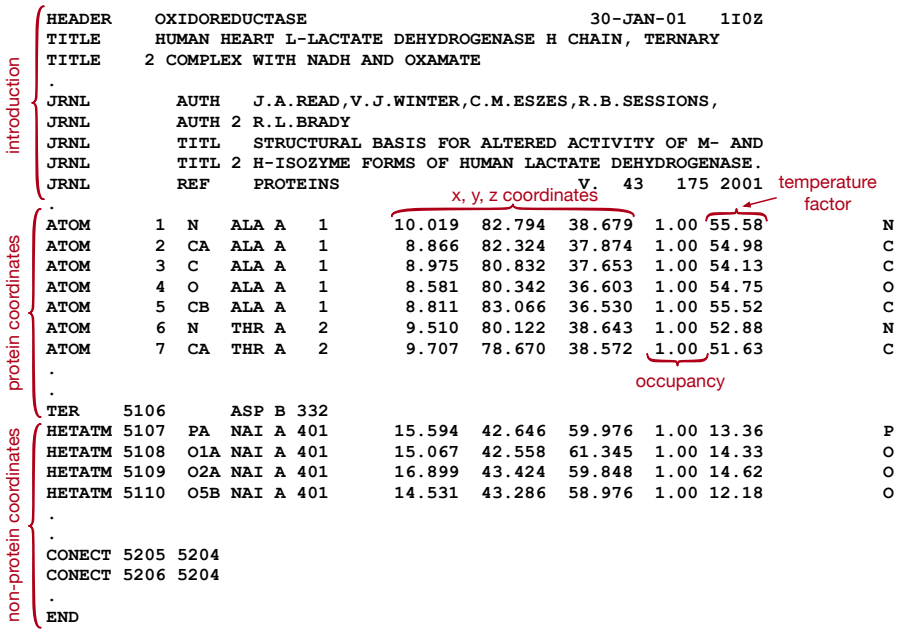
Figure 4.2. Cartesian coordinate file for aspirin and, at the right, a line diagram of the structure showing the atom numbering. The first line gives the total number of atoms. The subsequent lines specify the element, a serial number for each atom (1-21), the three x, y, and z coordinates of its location, the atom type number (1, 2, 3, 5, 6, 7 or 24; these numbers are used by the molecular modeling program) and the last four columns list the serial numbers of the atoms bonded to it.

Protein structure files contain thousands of atoms, use the protein data bank (pdb) file format, and are freely available at https://www.rcsb.org. These pdb format files specify the Cartesian coordinates (x, y, z) of each atom and include additional information. Protein structure files are text files, which can be opened by text editors or word processing programs, Figure 4.3. The files are large (5,524 atom locations, 115 pages for the example in Figure 4.3). The information in these files can be grouped into introduction, protein atom (and nucleic acid atom, if present) coordinates, and non-protein atom coordinates.

The file partly shown in Figure 4.3 contains 6,391 lines; 5,524 lines contain x, y, z coordinates of atoms and the remaining lines contain additional information. The introductory information describes the structure and the experimental methods. It includes the title of the structure, journal reference, remarks, amino acid sequence, list of heteroatoms in the structure, list of disulfide bonds and crystallographic information.

The protein coordinates section lists the atoms numbered sequentially. The atom naming follows the convention described in the previous chapter and includes the amino acid name, amino acid number, and the protein chain indicated by a letter. For example, atom 1 in Figure 4.3 is the amino nitrogen of serine, which is the first amino acid in chain A. The next numbers are the Cartesian (x, y, z) coordinates for this atom. Most x-ray crystal structure files do not include hydrogen atoms because most protein crystals do not diffract well enough to allow hydrogen atoms to be resolved. Molecular modeling programs will add hydrogen atoms by inferring their positions from the positions of other atoms. In contrast, protein structure files from NMR experiments do include hydrogen atom positions because the NMR experiments measure locations of hydrogens.

The (x, y, z) coordinates are followed by the atom occupancy, which is 1.00 in most cases, indicating that the atom is located at one position. For flexible regions, the atoms may lie in multiple positions. In these cases, the structure file lists coordinates for each position and assign each one an occupancy. If an atom occupies two positions equally, then the pbb file




```
HEADER    OXIDOREDUCTASE                          30-JAN-01   1I0Z
TITLE     HUMAN HEART L-LACTATE DEHYDROGENASE H CHAIN, TERNARY
TITLE    2 COMPLEX WITH NADH AND OXAMATE
.
JRNL        AUTH   J.A.READ,V.J.WINTER,C.M.ESZES,R.B.SESSIONS,
JRNL        AUTH 2 R.L.BRADY
JRNL        TITL   STRUCTURAL BASIS FOR ALTERED ACTIVITY OF M- AND
JRNL        TITL 2 H-ISOZYME FORMS OF HUMAN LACTATE DEHYDROGENASE.
JRNL        REF    PROTEINS                       V.  43   175 2001
.
ATOM      1  N   ALA A   1      10.019  82.794  38.679  1.00 55.58           N
ATOM      2  CA  ALA A   1       8.866  82.324  37.874  1.00 54.98           C
ATOM      3  C   ALA A   1       8.975  80.832  37.653  1.00 54.13           C
ATOM      4  O   ALA A   1       8.581  80.342  36.603  1.00 54.75           O
ATOM      5  CB  ALA A   1       8.811  83.066  36.530  1.00 55.52           C
ATOM      6  N   THR A   2       9.510  80.122  38.643  1.00 52.88           N
ATOM      7  CA  THR A   2       9.707  78.670  38.572  1.00 51.63           C
.
.
TER    5106      ASP B 332
HETATM 5107  PA  NAI A 401      15.594  42.646  59.976  1.00 13.36           P
HETATM 5108  O1A NAI A 401      15.067  42.558  61.345  1.00 14.33           O
HETATM 5109  O2A NAI A 401      16.899  43.424  59.848  1.00 14.62           O
HETATM 5110  O5B NAI A 401      14.531  43.286  58.976  1.00 12.18           O
.
.
CONECT 5205 5204
CONECT 5206 5204
.
END
```

Figure 4.3. Three sections from the x-ray structure file for human heart lactate dehydrogenase from the protein data bank (pdb id = 1i0z). The first section, introduction, contains the name of the molecule, a reference to the peer-reviewed article describing this structure and experimental details about the structure determination. The second section lists the coordinates of protein atoms. The third section lists coordinates for non-protein atoms (solvent and other co-crystallized molecules). It also contains 'CONECT' lines to specify the connectivity of non-protein molecules such as NADH (named NAI) in this structure. The END record marks the end of the file.

contains separate lines for each position and each position has an atom occupancy of 0.50.

The numbers in the column after occupancy are the temperature factors, or B-factors, for each atom. The B-factor has units of $Å^2$ and describes the displacement of the atomic positions from a mean value. Typical values are 15-30 $Å^2$ in the protein core, but more flexible or disordered regions (such as the N-terminal serine in the structure above) have higher values (57 $Å^2$).

The last section lists the coordinates and connectivity for non-protein atoms within the structure, which are called heteroatoms (HETATM). These include water, other solvent molecules (e.g., glycerol) and any inhibitors or substrates co-crystallized with the protein. The solvent water molecules are those that remain in fixed positions within the crystal so that they diffract some of the x-rays. In reality, the protein is completely surrounded by water molecules. The records starting with CONECT specify the connectivity for the heteroatoms. In contrast, the protein coordinates section assumes standard atom connectivity for amino acid residues and nucleotides, so it does not contain CONECT records. The end of the pdb structure file contains a record for bookkeeping (MASTER) and an END record.

4.1.3 Visualizing proteins

Opening a structure file with molecular visualization software instead of a text editor reveals a three-dimensional image that can be rotated and zoomed on the screen. PyMOL is popular molecular visualization software for proteins, which is available as an open-source version, a free educational version, and also paid versions. Tutorials are available online and a publication gives an overview of the different types of visualizations that are possible.[50] PyMOL includes a Python interpreter so that Python commands can extend and automate PyMOL. Later chapters

include PyMOL scripts that automate highlighting features like surface residues, holes in the protein core or ion pairs.

Molecules look like the blobs of electron density shown in Figure 4.1, but molecular visualization software represents molecules in ways that simplify and emphasize different aspects of their structure. For example, the ball and stick representation of aspirin resembles mechanical molecular models used for conformational analysis, Figure 4.4. The wireframe view focuses on the connectivity of the atoms, while the space-filling representation consists of overlapping van der Waals spheres to show overall size and shape of the molecule. The model may be colored by element, as shown in Figure 4.4, but colors can also indicate the electrostatic charge, hydrophobicity, or other atomic property.

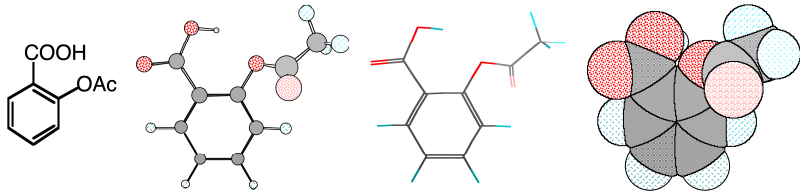


Figure 4.4. Line drawing of aspirin (left) and three representations of its three-dimensional structure: ball and stick, wireframe, and space-filling.

Since proteins contain so many more atoms than small organic molecules, researchers use additional simplifying representations that hide atoms. The default settings in PyMOL show the protein as a ribbon that traces the main chain of a protein without showing any individual protein atoms, Figure 4.5. This representation, called a cartoon, uses flat arrows to indicate β-sheet secondary structures, helices to indicate α-helices, and thin tubes to indicate loops. This representation emphasizes the protein fold while ignoring the atomic details of the structure. The center structure in Figure 4.5 adds gray and white colors to the ribbon to indicate different domains as well as stick represention of the bound cofactor NADH and inhibitor. The right-most representation in

Figure 4.5 shows a protein property of the same protein. The tubes are colored and scaled to indicate the degree of flexibility.

4.1.4 Intuitive protein engineering

In some cases, protein engineers use only inspection of the protein structure combined with chemical reasoning to choose substitutions. Looking at the three-dimensional structures of proteins using using a visualization program such as PyMOL reveals the overall fold of the protein, the location of the active site, which residues line the active site, and many other details. Molecular visualization software can measure distances and angles in a structure and even create models of where amino acids have been replaced. For example, inspection of the structure identifies substitutions that could expand the active site to accommodate a larger substrate.

This intuitive approach is a good first step, but one should not be disappointed if it fails. First, this approach typically examines only one structure while protein properties depend on differences between two protein states. The intuitive approach ignores one state completely and also assumes that the single structure examined is a good representation of the folded state. In the example above, substitutions to expand the active site may cause the surrounding residues to readjust such that the active site contracts instead. Finally, some protein features such as relative energies, solvation, protein movements, and pK_a of catalytic groups are not evident from simple inspection of the structure and require calculations.

4.2 Computer modeling of proteins

All models are wrong, but some are useful.

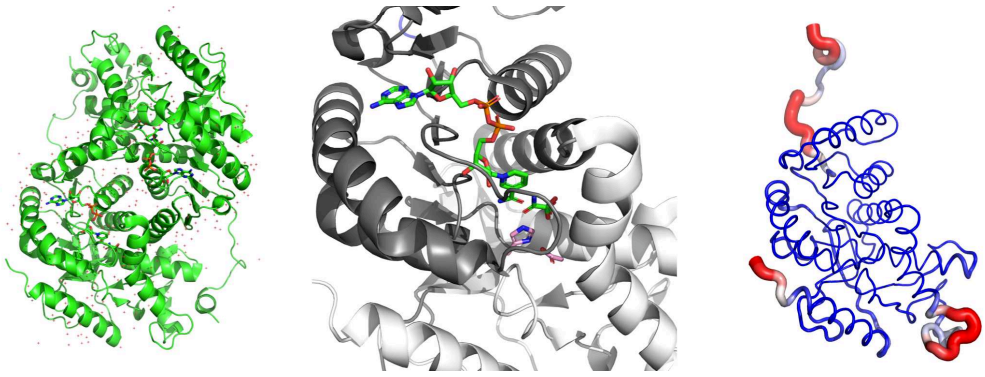


Figure 4.5. Three images of lactate dehydrogenase (pdb id = 1i0z) as displayed in PyMOL. Left: Default settings show the dimer observed in the crystal. The green ribbon represents the protein atoms. Helical flat ribbons indicate α-helices, flat arrows indicate β-sheets, and thin lines indicate turns. The red crosses represent solvent water molecules, and the bound NADH and inhibitor oxamate are shown in stick representation with the carbon atoms colored green. Center: A view of the active site region with the catalytic domain in white ribbons and the Rossmann fold domain in gray ribbons. The side chains of the catalytic residue (His193 and Asp166) are shown as sticks with pink carbons. Right: A ribbon representation where the colors and thickness of the tubes indicate the temperature factors of the atoms. The central part of the protein (blue, thin tubes) has well-defined locations, but the N- and C-terminii and residues 213-228 (red and white, thick tubes) have poorly defined positions suggesting flexibility. The text version of this pdb file was shown in Figure 4.3.

— George Box, statistician

The goal of computer modeling in protein engineering[51] is to predict which substitutions will improve the target properties of the protein. This prediction requires models of the protein states that contribute to the target properties. The two components of modeling a protein state are first to estimate the Gibbs energy of individual protein structures in each protein state and second to collect a representative sample of conformations that contribute to that state. Estimating the Gibbs energy of protein structures is relatively reliable, but collecting a representative sample of all contributing conformation remains difficult. Some important conformations occur infrequently and are difficult to find with computational methods.

4.2.1 Calculating the energy of a structure

The theoretically correct approach to model molecules is quantum mechanics where the electron distributions in molecules are calculated using approximations to solve the Schrödinger equation. Quantum mechanics is a first-principles method without experimentally derived parameters, but is far too complicated and slow to model entire proteins. Quantum mechanics is required to model the bond-making and bond-breaking steps of a reaction. These steps involve electron redistribution that cannot be modeled with simpler approaches. In these cases, quantum mechanics is used to model the active-site region (<100 atoms) while simpler methods are used for the rest of the protein.

Most computer modeling of proteins uses molecular mechanics, which simplifies molecules by treating them like macroscopic mechanical objects. Molecular mechanics uses classical physics to model molecules where spherical balls represent atoms and distance-dependent attractions between them represent bonding interactions. This theoretical model of molecules is incorrect because bonds between atoms arise

from delocalization of electrons and interaction of electrons with nuclei. These electron interactions do not behave according to classical physics. Nevertheless, molecular mechanics yields accurate structures of molecules. Empirical adjustment of the equations and parameters of molecular mechanics to experimental structures created predictions that match the known structures of molecules.

Different molecular mechanics software uses different force fields to calculate the energy of a molecular structure. The force field is the mechanical model used to represent molecules in molecular mechanics calculations. It consists of 1) an equation with multiple terms that yields the energy of a structure, 2) atom types to describe elements with different bonding arrangements, and 3) parameters (constants) adjusted so that the equations yield the correct energies and structures. Force fields differ in the complexity of the terms in the energy equation, the number and definition of atom types, and the number and value of the parameters. Each force field is best suited for specific types of molecules. Force fields like MM3 (Molecular Mechanics 3) are best to predict the structures of small organic molecules, while force fields like AMBER (Assisted Model Building with Energy Refinement) or Rosetta are better suited for biomolecules.

Energy equation for physics-based force fields. The energy equation of a force field consists of multiple terms, some of which describe interactions between bonded atoms and others describe the interactions between non-bonded atoms, Equation 4.1. Bonded interaction terms describe how energy varies with bond length, bond angle, and torsion angle along single bonds. Non-bonded interaction terms describe repulsive and attractive van der Waals and electrostatic interactions. The total energy, υ, varies with the structure of the molecule as defined by the distances between the atoms, r^N.

$$\upsilon(r^N) = \sum_{\text{bonded atoms}} \text{bonded interactions} + \sum_{\text{all atom pairs}} \text{non-bonded interactions} \tag{4.1}$$

The force field energy equation includes at least three terms to describe bonded interactions: bond stretching, angle bending and torsion along single bonds, Figure 4.6. Each term defines an ideal value and the energy cost for deviations from this ideal. The bond stretch term defines an optimum bond length, l_0 between two bonded atoms. When $l = l_0$, this term is zero indicating no energy penalty. Bond lengths shorter or longer than l_0 increase the energy of the molecule by the square of the deviation multiplied by k_b, which represents the stiffness of the bond. The angle bend term applies to a group of three bonded atoms and similarly has an optimum angle, θ_0, and angle stiffness, k_θ. The torsion angle term defines the orientation along a single bond for a group of four bonded atoms. There are usually several torsion angle minima. For example, the H-C-C-H torsion angle in ethane has three minima, which correspond to the three staggered conformations. The constant V_n defines the amplitude of the curve, the n defines its periodicity (number of minima), γ shifts the entire curve along the torsion angle (ω) axis. Multiple substituents at the ends of a single bond require several torsional angle terms to include the different interactions created as the torsion angle varies.

The non-bonded interactions depend on the distance between atoms, r, as shown in the bottom two terms in Figure 4.6. These terms were explained in the previous chapter. Pairs of atoms connected by chemical bonds are normally excluded from computation of non-bonded interactions because bonded energy terms replace non-bonded interactions. In biomolecular force fields all pairs of connected atoms separated by up to 2 bonds (1-2 and 1-3 pairs) are excluded from non-bonded interactions.

The force field calculates the sum of all the possible interactions to estimate the energy of a particular conformation of a molecule, Equation 4.2.

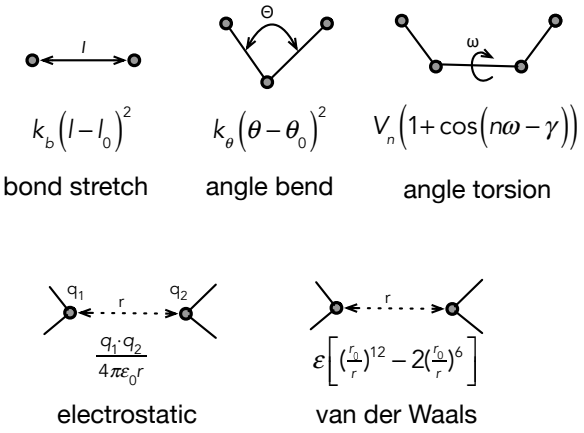


Figure 4.6. Simple force field terms that represent bonded (top three) and non-bonded interactions (bottom two). Bonded interactions describe distances, angle, and torsions between two, three, and four atoms connected by bonds. Non-bonded interactions describe the pairwise interactions between atoms that are nearby but not bonded covalently. All force fields model these five interactions but may modify these terms or add additional terms to describe them more accurately.

The sum includes terms for interaction between bonded atoms (all bonds, all angles, all torsions) and terms for interactions of non-bonded atoms (all possible pairwise interactions between atoms).

$$E_{total} = \sum_{\text{bonds}} k_{b,i}(l_i - l_{i,0})^2 + \sum_{\text{angles}} k_{\theta,i}(\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} V_n(1 - \cos(n\omega - \gamma)) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\epsilon_{ij} \left[\left(\frac{r_{i,j,0}}{r_{i,j}} \right)^{12} - 2 \left(\frac{r_{i,j,0}}{r_{i,j}} \right)^{6} \right] + \frac{q_i \cdot q_j}{4\pi \cdot \epsilon_0 \cdot r_{i,j}} \right) \tag{4.2}$$

For example, calculating the energy of n-butane would contain bond stretch terms for each C-C and C-H bond in the molecule, angle bending terms for each H-C-H, H-C-C, and C-C-C bond, and torsion angle terms for each H-C-C-H, H-C-C-C, and C-C-C-C bond. Non-optimal bond lengths or angles would raise the energy of the structure. The van der Waals interaction term would calculate bumping or weak attractive interactions between all atom pairs. The calculation of the non-bonded interactions may ignore electrostatic interactions for this non-polar molecule.

The number of bonded interactions increases linearly with the number of atoms in the molecule, N. For example, if each atom is bonded to an average of three other atoms, then the number of bonded interactions is $3N$. In contrast, the number of non-bonded interaction increase with the square of number of atoms. Each atom interacts with every other atom so the number of non-bonded interactions is $N(N-1)/2$. Division by two eliminates double counting of the interaction between atoms 1 and 2 and the interaction between atom 2 and 1 as separate interactions.* This different scaling with the number of atoms means that the number of

*The precise number of non-bonded interaction is $2N$ less than this value because connected atoms separated by up to two bonds are excluded from the calculation, but this small correction does not change the conclusion that there are many more non-bonded than bonded interactions.

non-bonded interactions is much larger than the number of bonded interactions for large molecules like proteins. For example, the dehydrogenase in Figure 4.5 above contains 374 amino acids per monomer. Each monomer contains approximately 2800 atoms and the software will add 2900 hydrogen atoms for a total of 5700 atoms. There are $5700 \cdot 3 = 17,100$ bonded interactions to calculate and $5700 \cdot 5699/2 = 16$ million non-bonded interactions. The number of non-bonded interaction is ~1000-fold more than the number of bonded interactions. The number of non-bonded interactions will further increase if the calculation includes water molecules. To save computation time, programs usually ignore non-bonded interactions between atoms that are more than 6 Å from one another.

Atom types. Atom types are the molecular subunits in a molecular mechanics calculation and include both element type and its bonding arrangement. While elements are the authentic subunits of molecules, molecular mechanics uses atom types to distinguish different bonding arrangements of the same element. This approach compensates for the fact that molecular mechanics does not explicitly include electrons in the calculations. For example, in the aspirin example in Figure 4.2 above, C1-C6 are all atom-type 2, which in the MM2 force field is an sp^2-hybridized carbon in an alkene (a non polar bond). C12 and C16 are atom-type 3, which is an sp^2-hybridized carbon in a carbonyl or imine (a polar bond), while C17 is atom type 1, which is an sp^3-hybridized carbon. Even though these atoms are all carbon atoms, molecular mechanics requires different parameters to describe their different bonding arrangements. In contrast, a quantum mechanics approach would assign all these atoms as carbon atoms and the calculated differences in the electron distribution density of each carbon would lead to different bonding arrangements.

Using atom types instead of elements as the subunits of molecules leads to some unusual atom types. For example, the MM2 force field models lone pairs of electrons as a special atom types. These pseudo atoms account for the non-spherical nature of atoms containing lone

pairs, such as oxygen and nitrogen. These lone-pair pseudo atoms improve the model of van der Waals interactions, torsional potentials and hydrogen bonding for these atoms. Other force fields (e.g., MM3, MM4) do not use lone-pair pseudo atoms and instead model lone pairs with a more complex electrostatic treatment and adjusted torsional potentials. These unusual atoms types are a reminder that molecular mechanics oversimplifies chemical bonding.

Different force fields use different definitions for atoms types. For example, AMBER, a force field optimized for biopolymers, includes united atoms types which include hydrogen atoms implicitly. For example, the AMBER-assigned atom types for tyrosine include CD (sp^2 aromatic carbon in 6-membered ring with 1 hydrogen), C2 (sp^3 carbon with 2 hydrogens), and CH (sp^3 carbon with 1 hydrogen), Figure 4.7. The rationale for this approximation is that large structures adjust to avoid small distortions in C-H subunits, so they are all similar in biopolymers. Force fields for small organic molecules never use united atom types for a methyl group. Strained organic molecules may distort methyl groups, and these distortions contribute to the overall energy of a molecule.

A force field for protein modeling such as AMBER[52] differs from a force field for modeling small organic molecules such as MM3,[53] Table 4.1. MM3 includes 153 atom types to model a wide range of elements and bonding arrangement in organic molecules. In contrast, AMBER includes only 41 atom types, which is enough to describe proteins, nucleic acids and carbohydrates. MM3 includes elaborate functions for the bond stretch, bond angle and cross terms to accurately model the structures of a wide range of organic molecules, including strained structures. In contrast, AMBER uses a simplified treatment of bond stretch, bond angle and van der Waals interactions to be fast enough to calculate large structures. Distortions of bonds, angles and torsions are uncommon in proteins since the proteins can relieve strain by moving the backbone. MM3 uses a simplified treatment of electrostatic interactions using only formal charges and bond dipoles, since most organic molecules are non-polar. MM3 adds a point charge term for charged molecules, but

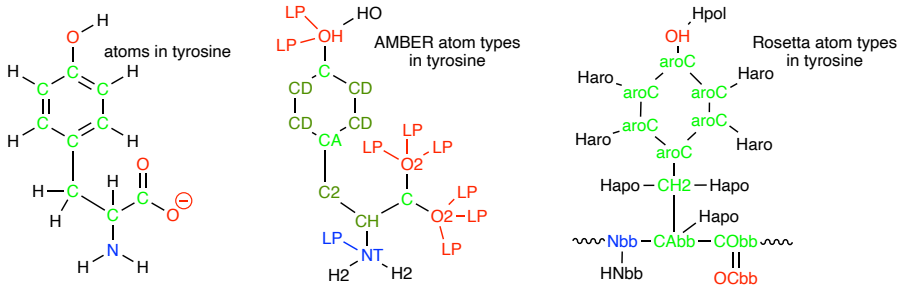


Figure 4.7. The AMBER and Rosetta force fields assign the atoms in tyrosine differently (middle and right structure) from the actual elements of the atoms in in tyrosine (left structure). The different atom types for the same element to indicate different bonding or polarity. The AMBER hydrogen atoms include HO (hydrogen on oxygen) and H2 (amino hydrogen in NH_2); the oxygen atoms include OH (alcohol oxygen) and O2 (carboxyl or phosphate nonbonded oxygen); the carbon atoms include C (sp^2 carbonyl carbon and aromatic carbon with hydroxyl substituent in tyrosine) and CA (sp^2 aromatic carbon in 6-membered ring with 1 substituent). Tyrosine contains only one nitrogen, which is assigned AMBER atom type NT (sp^3 nitrogen with 3 substituents). In other cases, several atoms are combined into a single unit called a united atom type: CD (sp^2 aromatic carbon in 6-membered ring with 1 hydrogen), C2 (sp^3 carbon with 2 hydrogens), and CH (sp^3 carbon with 1 hydrogen). AMBER assigns lone pairs, which are not atoms, to the atom type LP. Rosetta hydrogen atoms include Hapo (apolar hydrogen), Haro (hydrogen on an aromatic ring), Hpol (polar hydrogen), and HNbb (backbone amino hydrogen). Rosetta carbon atoms include aroC (carbons in an aromatic ring), CH2 (methylene carbons), CAbb (backbone C alpha), and CObb (backbone carbonyl carbon). Rosetta oxygen atoms are OH (hydroxyl oxygen) and OCbb (backbone carbonyl oxygen). The backbone amide nitrogen is assigned atom type Nbb.

errors can occur in highly polar molecules. AMBER uses partial point charges for all atoms to more accurately model electrostatic interactions in biomolecules, which are highly polar.

Table 4.1. Comparison of a force field for small organic molecules (MM3) with one for proteins and nucleic acids (AMBER).

Force field	Number of atom types	Application	Stretch	Bend	vdW	Electrostatic
MM3	153	general, organic molecules	r^2, r^3, r^4	6 terms	e^r, r^2	dipole or charge
AMBER	41	biopolymers	r^2	2 terms	r^{12}, r^6, r^{12}, r^{10}	charge	charge
Rosetta	59	biopolymers	statistical potential	statistical potential	r^{12}, r^6	charge with modifications	with modifications

Rosetta, a statistics-based force field. The MM3 and AMBER force fields rely on rules that describe physical interactions and are known as physics-based force fields. Another class of force fields, called statistics- or knowledge-based, replaces some of the rules with statistical data from databases. Statistics-based force fields replace rules of how to build structures with examples of structures. The best known statistics-based force field is Rosetta developed by the Baker laboratory at the University of Washington. Rosetta replaces thef physics-based equations for bonded interactions with statistically derived energies from structures in the protein data bank.[54] For example, instead of defining optimum bond distances and bond angles, Rosetta assigns

backbone conformation according to how frequently they occur for similar amino acids. Frequently-occurring structures are scored as more stable than rarely-occurring structures. Similarly, Rosetta includes the propensity of amino acids to be buried in protein structures, which approximates solvation effects, and the propensity of amino acids to occur next to each other to weight the overall energy score of a protein structure. This approach is much faster than a physics-based force field. Rosetta uses similar physics based energy terms to calculate non-bonded interactions. One disadvantage of knowledge-based force fields as compared to physics-based force fields is the loss of insight as to why some interactions are favored over others.

Modeling bond stretching. The best description of how energy varies with bond length is the Morse potential, Equation 4.3, where l is the bond length and l_0 is the ideal bond length.

$$v(l)_{\text{bond}} = D\left(1 - e^{-a(l-l_0)}\right)^2 - D \tag{4.3}$$

D is a constant that describe the depth of the energy well (bond strength), while a is a constant that describes the width of the well (bond stiffness). At large bond lengths, the expression inside the parentheses approaches one and the energy approaches zero, Figure 4.8. This behavior indicates no interaction at large distances. When $l = l_0$, the expression inside the parentheses is zero and the energy is $-D$. This behavior indicates that as the two atoms approach each other the energy decreases due to the formation of the bond reaching a minimum at l_0, which is the optimal bond length for this atom pair. If the distance between the atoms decreases further, the energy increases sharply. The computer evaluation of the exponential term in the Morse potential equation requires expansion to a series and evaluation of multiple terms, which is computationally costly.

To speed up calculation, force fields replace the Morse potential with polynomials, which are faster to calculate. AMBER uses a single

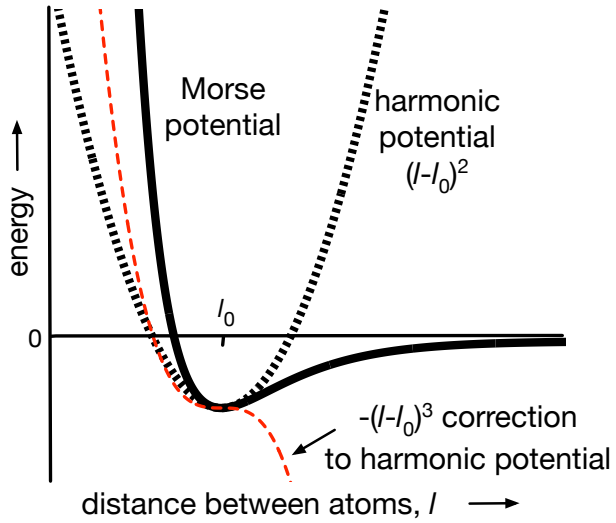


Figure 4.8. Equations to model the bond stretch in molecular mechanics. The most accurate equation is the Morse potential (solid line), which predicts no interaction at long distances, stabilization at an intermediate distances near l0 due to formation of the bond, and a sharply increasing repulsive interaction a short distances. A harmonic potential (parabola, short dashed line) matches the Morse potential near l0, but deviates at shorter and longer distances. A cubic correction to the harmonic potential (red dash line) decreases the energy at long distances and increases the energy at short distance to match the Morse potential over a wider range of bonds lengths.

quadratic term, eq, Equation 4.4. This parabola equation describes a harmonic oscillator or spring where the energy increases when the distance deviates from the optimal distance. This simple harmonic matches the Morse potential closely near the optimal distance, l_0, but deviates when the bond is distorted. The simple harmonic increases in energy symmetrically when the bond is shortened or lengthened, so it underestimates the energy when the distance is significantly shortened and overestimates the energy when the distance is significantly lengthened. AMBER tolerates these deviation because it is intended to model proteins where the bond distances are always near their optimal distances.

$$v(l)_{\text{bond}} = k_b(l - l_0)^2 \tag{4.4}$$

In contrast, the bonding and interatomic interactions within small organic molecules often forces bond lengths to adopt shorter or longer distances than the optimum. Accurate modeling of organic structures requires accurate modeling of a wider range of bond lengths. MM3 adds cubic and quartic polynomial terms to match the Morse potential over a wider range of interatomic distances, Equation 4.5. The cubic polynomial term increases the energy at distances shorter than l_0 and decreases the energy at distances longer than l_0 to extend the range of bond lengths that match the Morse potential. The positive quartic term offsets the large negative values of the cubic term at long bond distances. The constants CS (cubic stretch) and QS (quartic stretch) adjust the magnitude of these additional terms. Including these corrections allows MM3 to yield accurate bond distances for even strained organic molecules.

$$v(l)_{\text{bond}} = k_b(l - l_0)^2 - CS(l - l_0)^3 + QS(l - l_0)^4 \tag{4.5}$$

MM3 also includes cross terms that account for interactions between the terms. For example, bond angles can affect the energy of the bond stretch. As a bond angle (e.g, H–O–H) narrows, the two H atoms start to bump

each other. Opening the angle relieves the strain, but lengthening the O–H bond also relieves the strain. The stretch-bend cross term accounts for this interaction between angle bending and bond stretching for more accurate energy calculations. AMBER does not use cross terms.

Rosetta does not model bond stretching at all, but uses statistical potentials to define the conformation of backbone and the side chain orientations.

Modeling electrostatic interactions. In contrast to the bond stretching term where MM3 adds additional terms, AMBER uses more complex calculations for electrostatic interactions. Proteins and nucleic acids are charged molecules so electrostatic interactions are an important contributions to their structure. Most small organic molecules are not charged, so electrostatic interactions are less important. AMBER assigns point charges to all atoms and calculates the interactions between all atom pairs using Coulomb's law plus an additional term for hydrogen bond interactions, Equation 4.6. Hydrogen bonds are partially covalent and have preferred bonds angles. If hydrogen bonds were modeled using a purely electrostatic interaction, then this partial covalent character would be lost. AMBER calculate pairwise electrostatic interactions for all atoms within 30-40 Å of each other. The diameter of a spherical 30 kDa protein is ~40 Å, so charges on the protein surface weakly interact with charges on the opposite surface.

$$\upsilon(r)_{\text{electrostatics}} = \frac{q_1 q_2}{4\pi\epsilon_0 r} + \left(\frac{A}{r^{12}} - \frac{C}{r^{10}}\right) \tag{4.6}$$

Most organic molecules are uncharged, so MM3 ignores electrostatic interactions between the partial atomic charges. Instead, MM3 assigns bond dipoles to polar bonds and calculates the interactions between them, Equation 4.7. This interaction decreases with the third power of the distance between them so pairwise interactions beyond 18 Å are ignored. In cases where an organic molecule is charged, MM3 adds a

Coulombic interaction term for the charged atoms like the first term in Equation 4.6 above.

$$v(r)_{\text{electrostatics}} = \frac{\mu_1 \mu_2}{\epsilon_0 r^3} \left(\cos\chi - \cos\alpha_1 \cdot \cos\alpha_2\right) + \text{if molecule is charged} \frac{q_1 q_2}{4\pi\epsilon_0 r} \tag{4.7}$$

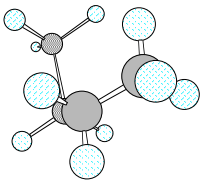
Rosetta also uses Coulomb's law with partial charges to model pairwise electrostatic interactions, but adds modifications to improve accuracy.[54] Rosetta modifies the dielectric constant to vary between the protein core (a low dielectric of 6) and the solvent-exposed surface (a high dielectric of 80). Rosetta also modifies the electrostatic interactions at short and long distances and when the atoms are within four bonds of each other. These empirical modifications improve its ability to match known protein structures.

4.2.2 Optimizing the geometry of a structure

The calculated energy of a structure is a score, and the geometry optimization is the strategy to get the best score, which correspond to a minimum energy structure. Predicting molecular structures using molecular mechanics involves cycles of energy calculations followed by adjustments of the structure (bond lengths, angles, etc.) and energy calculation of the new structure. This calculation finds structures that lie at the bottom of an energy well and is called a geometry optimization or an energy minimization. The convergence criterium is an energy difference between successive calculation, which determines when the cycles of calculation and structure adjustment should stop. When the energy difference drops below the convergence criterium, the calculation stops. A lower convergence criteria finds a structure closer to the center of the energy minimum and requires a longer calculation.

For example, a geometry optimization of *n*-butane yielded the structure in Figure 4.9. This calculated structure matches the experimental structure of the gauche conformation of butane. The constants in the molecular mechanics energy equation have been adjusted so that the energy minima matches the experimental structures.

Note that gauche conformation of butane is not the most stable conformation of *n*-butane. The geometry optimization yielded the gauche conformation because the initial structure (not shown) was closest to the gauche conformation. Geometry optimization yields a minimum energy structure that lies closest to the starting structure. The algorithm only moves down the energy slope to find the closest energy minimum. This energy minimum may not be the global minimum, which is the lowest energy conformation among all possible conformations. The gauche conformation is not the global minimum because it has a C-C-C-C torsion angle of ~ –60°, which retains some bumping interaction between the two methyl groups.



```
Note: All parameters used are finalized (Quality = 4).
  Iteration   51: Minimization terminated normally because
the gradient norm is less than the minimum gradient norm
  Stretch:          0.1675
  Bend:             0.6051
  Stretch-Bend:     0.0727
  Torsion:          0.4535
  Non-1,4 VDW:     -0.3814
  1,4 VDW:          2.1239
Total:              3.0412
```

Figure 4.9. Geometry optimization of *n*-butane using the MM2 molecular mechanics force field in the software Chem3D. The calculation, called an geometry optimization or energy minimization, involved calculating the energy of the starting *n*-butane structure, adjusting the geometry slightly to lower energy and repeating until the decrease in energy was insignificant (51 iterations). The list shows the energy components of the final structure show as a ball and stick model. This software refers to the convergence criterium as the 'minimum gradient norm.' The units for energy are kcal/mol.

The global minimum energy conformation of *n*-butane is the anti conformation, where the two methyl group lie across from each other with a torsion angle of 180°, Figure 4.10. A second geometry optimization of *n*-butane starting from a structure similar to the anti conformation yielded the anti conformation, Figure 4.11. It has a lower energy showing that the first optimized structure was not the global minimum, but a local minimum. In both cases the geometry optimization yield the *closest* energy minimum, not necessarily the lowest energy minimum.

The energy from a molecular mechanics calculation, called a steric energy, is the energy of the molecule relative to a hypothetical strain-less molecule where all bond lengths, angles, torsions, and non-bonded interactions are at their optimum values, Figure 4.12. The absolute values of steric energy have no physical significance, but differences in steric energy correspond to enthalpy differences. Thus, the steric energy value of 3.04 kcal/mol calculated for gauche butane above is not significant by itself, but the difference in steric energies between gauche and anti butane (3.04 - 2.17 = 0.87 kcal/mol) predicts the enthalpy difference between these two conformations. Differences in steric energy are only valid for different conformations or configurations of the same system. The steric energies of butane and pentane cannot be compared since they differ in molecular formula. One should not over interpret the individual energy contributions to the steric energy shown in Figures Figure 4.9 and Figure 4.11. The parameters are adjusted so the total energies and the structures match experimental values, but there are no experimental values for the individual energy contributions, so they should not be used for analysis.

Molecular properties depend on Gibbs energy differences, but the molecular mechanics calculations in Figure 4.10 yield enthalpies, but not entropies, for anti and gauche butane. The torsion diagram also provides the information needed to estimate the entropy difference between the anti and gauche conformations of butane. Butane can adopt one anti conformation, but two enantiomeric gauche conformations. Entropy favors the gauche conformation because there are two possibilities. Problem

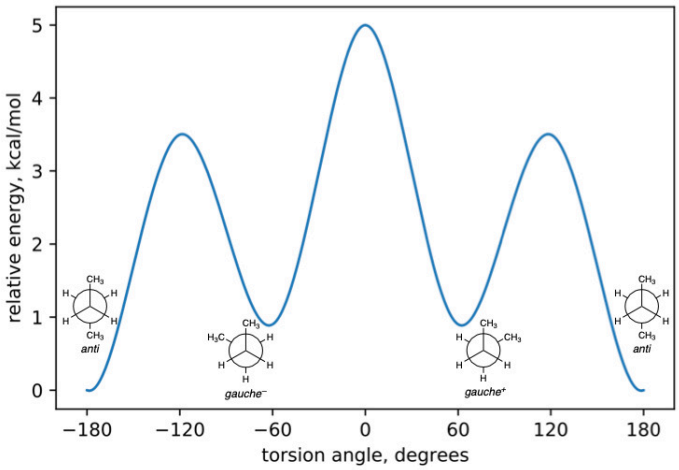
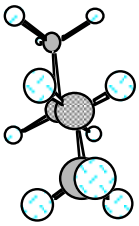


Figure 4.10. The relative energy of *n*-butane varies with the C–C–C–C torsion angle. The maxima correspond to eclipsed conformations along the central carbon-carbon bond, while the minima correspond to staggered conformations. The lowest energy minima is the anti conformation where the torsion angle between the two methyl groups is 180°. The two minima with energies of +0.9 kcal/mol relative to the anti conformation are enantiomeric gauche conformations where the torsion angle is ±60°. The two gauche conformations differ by an either clockwise or counter clockwise orientation of the methyl groups.



```
Note: All parameters used are finalized (Quality = 4).
  Iteration   14: Minimization terminated normally because
the gradient norm is less than the minimum gradient norm

  Stretch:           0.1577
  Bend:              0.2943
  Stretch-Bend:      0.0547
  Torsion:           0.0075
  Non-1,4 VDW:      -0.4054
  1,4 VDW:           2.0653

Total:               2.1742
```

Figure 4.11. The geometry-optimized structure of the anti conformation of *n*-butane. The energy of the anti conformation is 0.87 kcal/mol lower than for the gauche conformation in Figure 4.9 above.

1 at the end of the chapter asks the reader to estimate the Gibbs energy difference between anti and gauche butane and compare to the experimental equilibrium constant. The Gibbs energy estimate comes from the steric energy, which corresponds to enthalpy, and from an estimate of the entropy difference.

Rosetta differs from MM3 and AMBER in yielding Gibbs energies instead of enthalpies as the result of an energy calculation. The assumption of statistics-based potentials like Rosetta is that the data in the protein structure database follows a Boltzmann distribution. The Boltzmann distribution is a probability distribution developed for gases that varies exponentially with Gibbs energy of the state. States with lower energy are more likely to be occupied. If the database contains a random distribution of structures and the interactions within them act independently, then one expects that the more stable interactions will occur more frequently than less favorable interactions. The statistical frequency of an interaction in a protein structure database reveals its Gibbs energy relative to other, similar interactions. Interactions that occur frequently in the database are assigned a low energy, while interactions that occur rarely are assigned a high energy. The probability of finding a specific conformation, p_i, within an equilibrium mixture of conformations decreases exponentially as the Gibbs energy of the conformation, G_i, increases. k is Boltzmann's

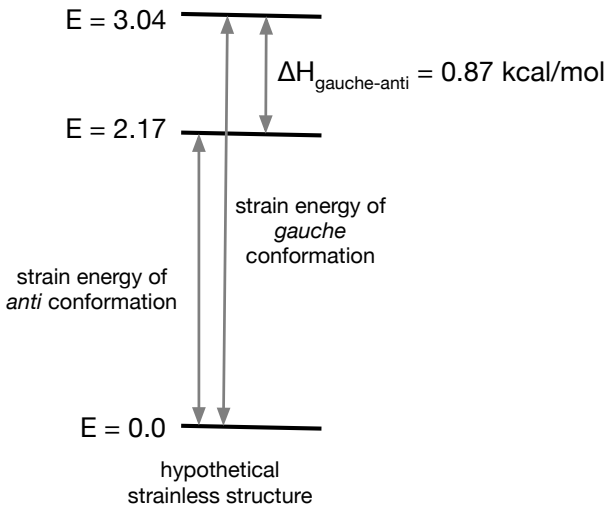


Figure 4.12. Using molecular mechanics to calculate enthalpy differences between conformations. Geometry optimization of the gauche conformation of butane yielded a strain energy of 3.04 kcal/mol. This strain energy compares the energy of this gauche conformation to a hypothetical strainless structure of butane. A similar geometry optimization yielded a lower strain energy for the anti-conformation of butane, 2.17 kcal/mol. While the absolute values of these strain energies have no physical significance, the difference between them, 0.87 kcal/mol, corresponds to the enthalpy difference between the two conformations. Rosetta uses a different approach where energy differences between conformations correspond to Gibbs energy differences.

constant and T is the temperature, Equation 4.8.

$$p_i \propto e^{-G_i/kT} \tag{4.8}$$

The RosettaDesign tool is useful for protein engineering and a web interface to this software is available at http://rosettadesign.med.unc.edu.

4.2.3 Searching for conformations

The most important reason to search for conformations (microstates) is to create a model of the protein state, which is needed to predict protein properties. Finding all conformation is impossible; the goal is to find a representative sample of conformations for that state. If one seeks to estimate protein stability, then finding all the low-energy conformations is important. If one seeks to estimate reactivity, then one needs to find the conformations that contribute to reactivity, even if they are high energy conformations. Other reasons to search for conformation is to find the global energy minimum calculation or to estimate entropy when using AMBER for modeling.[55] Modeling with Rosetta does not require a separate estimate for entropy.

Three approaches to conformational search are systematic search, random search, and molecular dynamics. The systematic search tests all possible orientations along rotatable bonds. A systematic search identifies all of the minimum conformations, including the global minimum. For butane, a systematic adjustment of the C-C-C-C torsion angle by 30° followed by geometry optimization of each of the six structures to the closest local minimum identified the three minima including the global minimum, Figure 4.10.

The number of possible conformations in a systematic search increases exponentially as the number of rotatable bonds increases. In general, the

number of possibilities is given by the number of choices at each location raised to the power of the number of locations, Equation 4.9.

$$\text{possibilities} = \text{choices at each location}^{\text{locations}} \tag{4.9}$$

For the number of possible conformations:

$$\text{possible conformations} = \text{positions along each bond}^{(\text{bonds})} \tag{4.10}$$

Even with a small molecule like a five-amino-acid peptide, a systematic search for the global minimum is difficult. If an amino acid has an average of five rotatable bonds (two in the main chain, three in the side chain), then a five-amino-acid peptide has 25 rotatable bonds. If there are six conformational possibilities at each bond, then the number of possible conformations is $(6)^{25}$ or ~10^{19}. This large number of possible conformations limits the application of systematic searches to small molecules.

The random search approach, called Monte Carlo search to suggest gambling, generates random structures by a torsion angle rotation followed by geometry optimization to yield a local energy minimum conformation. The Monte Carlo search prioritizes low energy structures because these will be most abundant. If after a random torsion angle rotation and geometry optimization, the energy of the structure is lower than the previous structure, then the current structure and the starting point for another random torsion angle rotation. If the current structure is not lower, then the previous structure is used again as a starting point. Occasionally, higher energy structures are accepted to avoid getting stuck in a local minimum. The search stops when structures with lower energy minima can no longer be found. The Monte Carlo search is more efficient than the systematic search because it focuses on the low energy structures instead of all the structures.

For compact structures like proteins, random search methods like Monte Carlo are inefficient. Most random changes of a torsion angle in a protein create unfavorable bumping interactions and are not realistic possibilities. To limit the search to more plausible possibilities, researchers limit the changes to side chain orientations. This approach generates plausible structures, but makes it difficult to escape local minima because alternative conformations of proteins often require cooperative adjustments in several amino acid side chains combined with main chain adjustments.

Molecular dynamics is the most common conformational search method for proteins. Molecular dynamics models molecular motions by allowing the atoms to move according to the force field. Newton's equation of motion, $a = F/m$, describes how a force, F, on an object with mass, m, accelerates the object in certain direction and by a certain amount, a. The force field describes the force on each atom due to distortions of the bonding interactions from their ideal values and due to non-bonded interactions with other atoms. The step size for the calculation of atom movements must be shorter than molecular vibrations (femtoseconds) to remain realistic. Calculating one nanosecond of protein movement may take several hours of computer time. Combining these steps creates a movie of molecular motion.

A typical molecular dynamic calculation for protein is a simulation. It seeks to model the folded protein state to show the conformations that a protein explores in solution. Its goal is not find all conformations, but to find a representative sample of conformations that the protein explores. Molecular dynamics simulations do not have an end point where the simulation is finished. Instead the researchers continue the simulation for an arbitrary time. For example, the simulation of a thermostable adenylate kinase below modeled 1 nanosecond of protein movement, but required hours of computation time. This simulation modeled the folded protein state in solution to identify which of the four ion pairs on its surface contributed to its stability.[56] Close contacts in the x-ray structure suggested that all four ion pairs contribute to stability, Table 4.2. The au-

thors reasoned that these distances could be misleading. First, the packing of the protein within the crystal creates additional electrostatic interactions that are not present in solution. Second, ion pairs on the surface are often flexible due to side-chain motion leading to uncertain positions. The molecular dynamics simulation of this protein modeled how the protein moves in solution. One ion pair had the shortest average distance because close contact persisted throughout the simulation. This behavior identified it as a stabilizing interaction. Two ions pairs had longer average distances because they maintained close contact through only parts of the simulation. The authors classified these two pairs as slightly stabilizing. The last ion pair had the longest distance because the pair stayed apart throughout the simulation. The authors classified this pair as destabilizing. Transferring these ion pairs into a less stable homolog confirmed these assignments. This molecular dynamics simulation revealed how the folded protein behaved in solution, which was different from that suggested by the single structure shown by x-ray.

Table 4.2. Molecular dynamic simulation of ion pairs on the surface of adenylate kinase identified the Arg116-Glu198 pair as the strongest.

Amino acid residues	Close N-O contacts in x-ray structure	Average N-O distance in MD simulation	Close contact maintained during simulation
Lys19-Glu202	2.8 Å, 3.2 Å	3.2 Å	partly
Arg116-Glu198	3.0 Å	2.7 Å (strongest)	yes
Arg131-Glu156	3.0 Å, 3.2 Å	3.8 Å	partly
Lys180-Asp114	4.0 Å, 4.8 Å	10.9 Å (weakest)	no

Typical molecular dynamics simulations model 1-10 nanoseconds of protein movement; in unusual cases simulations extend to one microsecond. These simulation times are short as compared to the times required a protein state. A protein state consists of all the conformations (microstates) that contribute to a protein property. For example, the turnover time for an enzyme catalyzed reaction may be a millisecond, so simulations shorter than a millisecond may omit conformations that occur rarely, but nevertheless contribute to catalysis. An example of a puzzle solved by long-timescale molecular dynamics simulations is the effect of mutations on the substrate scope and catalytic activity of monoamine oxidase.[57] The x-ray structure showed a closed conformation which could not explain the effects of mutations. The molecular dynamics simulations identified partially open and fully open conformations that could account for the changes in catalysis.

Since Gibbs energy consists of enthalpy and entropy, there are two ways that conformations can be rare or unfavorable. The first is high enthalpies or internal energies. Conformations with distorted bond torsions, angles, or lengths or with unfavorable bumping or electrostatic interactions have high enthalpies. The second type of unfavorable conformation are those that require several things to occur simultaneously. For example, moving a buried side chain requires not only rotating along one of the single bonds in the side chain, but simultaneously rotating along single bonds in adjacent side chains to create space for buried side chain to move. The buried side chain only has an opportunity to move when the adjacent side chain has moved and therefore occurs more rarely than conformations that don't require simultaneous movements. This requirement for simultaneous movements is an entropy cost since it imposes a requirement for order - simultaneous movement instead of movement anytime.

Some computational approaches to find rare conformations during a molecular dynamics simulation are 1) to extend the simulation for a longer time, 2) to repeat short simulations from several different starting conformations, and 3) to temporarily raise the temperature

unrealistically high (500 °C) to speed up slower conformational changes, then return the modeling to normal temperatures. While molecular dynamics is the best way to search for protein conformations, researchers recognize that the search may miss rarely-occurring conformations.

Molecular dynamics simulations typically include hundreds of explicit water molecules to more accurately model the solvation of the protein. The models used for the water molecule differ in their complexity. A typical model is called TIP3P (transferable intermolecular potential with 3 points) and represents water with three fixed sites corresponding to the three atoms. Each site has a charge, and the oxygen site also has van der Waals interactions. Explicit solvation models can find specific interactions between water and the protein, but adding hundreds of molecules slows down the simulation. The alternative to explicit solvation is implicit solvation, which models the solvent and counter ions as a continuous medium with a particular dielectric constant. This approach accounts for the average behavior of the solvent but does not find specific interactions with solvent molecules.

4.2.4 Predicting protein properties

The observable properties of a molecule depend not only on the global minimum conformation, but on the relative abundance of all the conformations (microstates) that contribute to the state. For example, accurate prediction of the density of liquid butane, which depends on its molecular volume, must include both the anti and gauche conformations.[58] The state of liquid butane is an unequal mixture of these two conformations and both contribute to the measured property of density.

The properties of a protein are determined not by a single structure, but by states, which are collections of interconverting conformations. Computer modeling must find these conformations, which is more difficult than calculation the energy of individual structures. There are too many

conformations of a protein to find all of them, so modeling uses simulations of protein movements to locate a representative sample of these conformations. If this sample of conformations is not representative of the state, then the prediction will be incorrect. For example, the design of a retroaldolase failed when protein adopted not the calculated structure, but a subtly different conformation that was not considered in the calculations.[59] Finding all the possibilities to consider is still an unsolved and active area of research.[60]

Because of the difficulties in accurately modeling protein states, researchers rely on approximations to model protein properties, Table 4.3. For example, one often assumes that a single x-ray structure is representative of the conformations available to a folded protein. In other example, the unfolded protein's structure is unknown, so researchers assume it is an unfolded random coil, but when this assumption fails, the predictions are incorrect.[61] These approximations reduce the accuracy of computational predictions making protein engineering challenging. Most computer modeling predictions of substitutions to improve proteins are incorrect, but the fraction of correct predictions, typically 1-10%, is vastly better than random guessing, so computer modeling is a useful tool. Computer modeling narrows the choice of substitutions from astronomical numbers to smaller numbers that can be tested experimentally.

Table 4.3. Computer modeling approaches to common protein engineering problems.

Goal	Approach	Comments
What is the structure of a variant protein?	homology modeling (extrapolation from known structures) SwissModel, AlphaFold	generally reliable single structure
Which conformations contribute to a protein state?	molecular dynamics simulations	challenging to find rare conformations
What is the structure of the enzyme-substrate complex or the protein-target complex?	docking calculations identify binding sites followed by molecular dynamics	challenging to find adjustments upon binding |
What is the structure of the transition state?	quantum mechanical modeling to define charges and shape of an energy maximum with partially broken and formed bonds	challenging for multi-step reactions | |
Which protein variant is more stable?	compare energies calculated by FoldX or Rosetta	low reliability, unfolded protein state is unknown, use single structure for folded protein

Goal	Approach	Comments
How to increase the binding of my protein to its target?	modeling to predict changes that increase | challenging, but several successful designs | favorable interactions between protein and | using Rosetta have been reported | target and remove unfavorable interactions |	
How to engineer a faster or more selective enzyme?	identify and compare the protein conformations that contribute to each property	currently impossible or difficult to predict

4.3 Conclusions

Examination of protein structures reveal many details about proteins and are the critical first step in protein engineering. However, protein structures alone are not enough to accurately predict protein properties like stability, binding, reactivity and selectivity. These properties depend on differences between two protein states, whose structure and energy may be unknown. One rarely has information about the unfolded protein state, the binding orientation of a target molecule, or the rate-limiting step of a reaction. In addition, protein states consist of numerous conformations, some of which may be critical for the protein property, but difficult to find. Researchers often assume that most features of protein states will cancel out and focus on a few differences between them. The accuracy of the predictions varies with the validity of these assumptions.

Computer models of proteins only approximate reality and in that sense they are all incorrect. Models differ in complexity because they make different assumptions. A good computer model will capture enough of the essential features so that it predicts the behavior in question. The most common assumption is that a single structure is representative of the protein state. A good approach to computer modeling is to first test the calculation on cases where the answer is known. If the predictions from this control calculation match experiments, then the predictions for cases where the result is unknown have a better chance of matching as well.

Glossary

Atom types are the molecular subunits used in molecular mechanics calculations. An atom type includes both an element and a bonding arrangement such as an sp^2-hybridized carbon in a polar double

bond. Some atom types, called united atom types, include several atoms such as a CH_2 group.

Deep neural network is a multi-layered machine learning model trained on large datasets of protein sequences, structures, and experimental data to learn complex, hierarchical patterns in protein sequence-structure-function relationships for predicting and designing protein properties. These networks excel at tasks like protein fold prediction due to their hierarchical architecture matching protein organization, but often struggle with more precise protein engineering predictions that require understanding subtle sequence-function relationships.

Force field is a mathematical description of how atoms interact with one another. A force field consists of a set of equations that describe how bonded and non-bonded atoms interact, a set of atom types that represent elements in different bonding arrangements and a set of parameters that scale the magnitude of the interactions to fit experimental values. Physics-based force fields, such as MM3 and AMBER, attempt to model the physical interactions between atoms, while knowledge-based force fields, such as Rosetta, include statistical potential that scores geometries that occur frequently in known protein structures as more stable.

Molecular dynamics is a computational method to simulate the motion of interacting atoms. The computation uses Newton's equations of motion with a force field that describes how atoms interact. Molecular dynamics simulations generate a collection of conformations, which can be played as a movie, or, more commonly, analyzed to extract average geometry relevant to a protein property like catalysis.

Molecular mechanics is a computational modeling approach to predict the structure, energetics, and dynamics of molecular. Molecular mechanics treats atoms as balls that interact with other atoms via classical physics interactions like balls on a spring. Although this model of molecules is incorrect because it does not include electrons explicitly in the model, molecular mechanics yields accurate

values because the equations and parameters have been adjusted to fit experimental values.

Quantum mechanics is a computational modeling approach that include electrons explicitly in the calculations. Quantum mechanics are needed to model bond-making and bond-breaking steps in a reaction. Quantum mechanics calculations are much more complex than molecular mechanics, so quantum mechanics is limited to 10-50 atoms in a calculation.

References

1. Morrison, C., & Lähteenmäki, R. (2018). Public biotech in 2017-the numbers. *Nat. Biotechnol.*, *36*(7), 576–584. https://doi.org/10.1038/nbt.4175
2. Brannigan, J. A., & Wilkinson, A. J. (2002). Protein engineering 20 years on. *Nat. Rev. Mol. Cell Biol.*, *3*(12), 964–970. https://doi.org/10.1038/nrm975
3. Estell, D. A., Graycar, T. P., & Wells, J. A. (1985). Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.*, *260*(11), 6518–6521. https://doi.org/10.1016/S0021-9258(18)88811-1
4. Pace, C. N., & Hermans, J. (1975). The stability of globular proteins. *CRC Crit. Rev. Biochem.*, *3*(1), 1–43. https://doi.org/10.3109/10409237509102551
5. Taverna, D. M., & Goldstein, R. A. (2002). Why are proteins marginally stable? *Proteins*, *46*(1), 105–109. https://doi.org/10.1002/prot.10016
6. Fersht, A. R., & Serrano, L. (1993). Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.*, *3*(1), 75–83. https://doi.org/10.1016/0959-440X(93)90205-Y

7. Bloom, J. D., Lu, Z., Chen, D., Raval, A., Venturelli, O. S., & Arnold, F. H. (2007). Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.*, *5*(1), 97–21. https://doi.org/10.1186/1741-7007-5-29
8. Gordon, S. R., Stanley, E. J., Wolf, S., Toland, A., Wu, S. J., Hadidi, D., Mills, J. H., Baker, D., Pultz, I. S., & Siegel, J. B. (2012). Computational design of an α-gliadin peptidase. *J. Am. Chem. Soc.*, *134*(50), 20513–20520. https://doi.org/10.1021/ja3094795
9. Harvey, S. B., Stone, M. D., Martinez, M. B., & Nelsestuen, G. L. (2003). Mutagenesis of the γ-carboxyglutamic acid domain of human factor VII to generate maximum enhancement of the membrane contact site. *J. Biol. Chem.*, *278*(10), 8363–8369. https://doi.org/10.1074/jbc.M211629200
10. Persson, E., Olsen, O. H., Bjørn, S. E., & Ezban, M. (2012). Vatreptacog alfa from conception to clinical proof of concept. *Semin. Thromb. Hemost.*, *38*(3), 274–281. https://doi.org/10.1055/s-0032-1302442
11. Bolli, G. B., Di Marchi, R. D., Park, G. D., Pramming, S., & Koivisto, V. A. (1999). Insulin analogues and their potential in the management of diabetes mellitus. *Diabetologia*, *42*(10), 1151–1167. https://doi.org/10.1007/s001250051286
12. Ryu, J. K., Kim, H. S., & Nam, D. H. (2012). Current status and perspectives of biopharmaceutical drugs. *Biotechnol. Bioprocess Eng.*, *17*(5), 900–911. https://doi.org/10.1007/s12257-012-0095-1
13. Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B. B., Henner, D., Wong, W. L. T., Rowland, A. M., Kotts, C., Carver, M. E., & Shepard, H. M. (1992). Humanization of an anti-p185$^{\mathrm{HER2}}$ antibody for human cancer therapy. *Proc. Natl. Acad. Sci. U. S. A.*, *89*(10), 4285–4289. https://doi.org/10.1073/pnas.89.10.4285

14. Sarkar, C. A., Lowenhaupt, K., Horan, T., Boone, T. C., Tidor, B., & Lauffenburger, D. A. (2002). Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nat. Biotechnol.*, *20*(9), 908–913. https://doi.org/10.1038/nbt725
15. Cantor, J. R., Yoo, T. H., Dixit, A., Iverson, B. L., Forsthuber, T. G., & Georgiou, G. (2011). Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(4), 1272–1277. https://doi.org/10.1073/pnas.1014739108
16. Loosmore, S. M., Zealey, G. R., Boux, H. A., Cockle, S. A., Radika, K., Fahim, R. E., Zobrist, G. J., Yacoob, R. K., Chong, P. C., & Yao, F. L. (1990). Engineering of genetically detoxified pertussis toxin analogs for development of a recombinant whooping cough vaccine. *Infect. Immun.*, *58*(11), 3653–3662. https://doi.org/10.1128/iai.58.11.3653-3662.1990
17. Hsieh, C.-L., Goldsmith, J. A., Schaub, J. M., DiVenere, A. M., Kuo, H.-C., Javanmardi, K., Le, K. C., Wrapp, D., Lee, A. G., Liu, Y., Chou, C.-W., Byrne, P. O., Hjorth, C. K., Johnson, N. V., Ludes-Meyers, J., Nguyen, A. W., Park, J., Wang, N., Amengor, D., … McLellan, J. S. (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science*, *369*(6510), 1501–1505. https://doi.org/10.1126/science.abd0826
18. Padgette, S. R., Kolacz, K. H., Delannay, X., Re, D. B., Tinius, C. N., Rhodes, W. K., Otero, Y. I., Barry, G. F., Eichholtz, D. A., Peschke, V. M., Nida, D. L., Taylor, N. B., & Kishore, G. M. (1995). Development, identification, and characterization of a glyphosate-tolerant soybean line. *Crop Sci.*, *35*(5), 1451–1461. https://doi.org/10.2135/cropsci1995.0011183x003500050032x

19. Castle, L. A., Siehl, D. L., Gorton, R., Patten, P. A., Chen, Y. H., Bertain, S., Cho, H.-J., Duck, N., Wong, J., Liu, D., & Lassner, M. W. (2004). Discovery and directed evolution of a glyphosate tolerance gene. *Science*, *304*(5674), 1151–1154. https://doi.org/10.1126/science.1096770
20. Bravo, A., & Soberon, M. (2008). How to cope with insect resistance to Bt toxins? *Trends Biotechnol.*, *26*(10), 573–579. https://doi.org/10.1016/j.tibtech.2008.06.005
21. Kurek, I., Chang, T. K., Bertain, S. M., Madrigal, A., Liu, L., Lassner, M. W., & Zhu, G. (2007). Enhanced thermostability of *Arabidopsis* rubisco activase improves photosynthesis and growth rates under moderate heat stress. *Plant Cell*, *19*(10), 3230–3241. https://doi.org/10.1105/tpc.107.054171
22. Whittle, E., & Shanklin, J. (2001). Engineering Δ^9-16:0-acyl carrier protein (ACP) desaturase specificity based on combinatorial saturation mutagenesis and logical redesign of the castor Δ^9-18:0-ACP desaturase. *J. Biol. Chem.*, *276*(24), 21500–21505. https://doi.org/10.1074/jbc.M102129200
23. Lin, H., & Lee, Y. K. (2017). Genetic engineering of medium-chain-length fatty acid synthesis in *Dunaliella tertiolecta* for improved biodiesel production. *J. Appl. Phycol.*, *29*(6), 2811–2819. https://doi.org/10.1007/s10811-017-1210-7
24. Yao, M. Z., Zhang, Y. H., Lu, W. L., Hu, M. Q., Wang, W., & Liang, A. H. (2012). Phytases: Crystal structures, protein engineering and potential biotechnological applications. *J. Appl. Microbiol.*, *112*(1), 1–14. https://doi.org/10.1111/j.1365-2672.2011.05181.x
25. Richardson, T. H., Tan, X., Frey, G., Callen, W., Cabell, M., Lam, D., Macomber, J., Short, J. M., Robertson, D. E., & Miller, C. (2002). A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable α-amylase. *J. Biol. Chem.*, *277*(29), 26501–26507. https://doi.org/10.1074/jbc.M203183200

26. Quax, W. J., Mrabet, N. T., Luiten, R. G. M., Schuurhuizen, P. W., Stanssens, P., & Lasters, I. (1991). Enhancing the thermostability of glucose-isomerase by protein engineering. *Bio/Technology*, *9*(8), 738–742. https://doi.org/10.1038/nbt0891-738
27. Bell, E. L., Finnigan, W., France, S. P., Green, A. P., Hayes, M. A., Hepworth, L. J., Lovelock, S. L., Niikura, H., Osuna, S., Romero, E., Ryan, K. S., Turner, N. J., & Flitsch, S. L. (2021). Biocatalysis. *Nat. Rev. Methods Primers*, *1*(1), 46. https://doi.org/10.1038/s43586-021-00044-z
28. van der Wielen, L. A. M., Ottens, M., & Straathof, A. J. J. (2001). Process Technology and Process Integration in the Preparation of Penicillins. In A. Bruggink (Ed.), *Synthesis of β-Lactam antibiotics: Chemistry, biocatalysis & process integration* (pp. 150–205). Springer Netherlands.
29. Ran, N., & Frost, J. W. (2007). Directed evolution of 2-keto-3-deoxy-6-phosphogalactonate aldolase to replace 3-deoxy-D-arabino-heptulosonic acid 7-phosphate synthase. *J. Am. Chem. Soc.*, *129*(19), 6130–6139. https://doi.org/10.1021/ja067330p
30. Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., ... Newman, J. D. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, *496*(7446), 528–532. https://doi.org/10.1038/nature12051
31. Anastas, P. T., & Warner, J. C. (2000). *Green chemistry : Theory and practice*. Oxford University Press.
32. Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J., Brands, J., Devine, P. N., Huisman, G. W., & Hughes, G. J. (2010). Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, *329*(5989), 305–309. https://doi.org/10.1126/science.1188934

33. Stone, E. M., Paley, O., Hu, J., Ekerdt, B., Cheung, N.-K., & Georgiou, G. (2012). De novo engineering of a human cystathionine-γ-lyase for systemic L-methionine depletion cancer therapy. *ACS Chem. Biol.*, *7*(11), 1822–1829. https://doi.org/10.1021/cb300335j
34. Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, *27*(10), 946–950. https://doi.org/10.1038/nbt.1568
35. Jobe, A., & Bourgeois, S. (1972). *lac* Repressor-operator interaction. VI. The natural inducer of the lac operon. *J. Mol. Biol.*, *69*(3), 397–408. https://doi.org/10.1016/0022-2836(72)90253-7
36. Egel, R. (1988). The '*lac*' operon - an irrelevant paradox? *Trends Genet.*, *4*(2), 31–31. https://doi.org/10.1016/0168-9525(88)90062-5
37. Valdez-Cruz, N. A., Caspeta, L., Pérez, N. O., Ramírez, O. T., & Trujillo-Roldán, M. A. (2010). Production of recombinant proteins in *E. Coli* by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microb Cell Fact*, *9*(1), 18. https://doi.org/10.1186/1475-2859-9-18
38. Austin, H. P., Allen, M. D., Donohoe, B. S., Rorrer, N. A., Kearns, F. L., Silveira, R. L., Pollard, B. C., Dominick, G., Duman, R., El Omari, K., Mykhaylyk, V., Wagner, A., Michener, W. E., Amore, A., Skaf, M. S., Crowley, M. F., Thorne, A. W., Johnson, C. W., Woodcock, H. L., ... Beckham, G. T. (2018). Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(19), E4350–E4357. https://doi.org/10.1073/pnas.1718804115
39. Hutchison 3rd, C. A., Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., & Smith, M. (1978). Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem.*, *253*(18), 6551–6560. https://doi.org/10.1016/S0021-9258(19)46967-6
40. Silva, D., Santos, G., Barroca, M., & Collins, T. (2017). Inverse PCR for point mutation introduction. In L. Domingues (Ed.), *PCR: Methods and Protocols* (pp. 87–100). Springer New York. https://doi.org/10.1007/978-1-4939-7060-5_5

41. Roy, S., & Caruthers, M. (2013). Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. *Molecules*, *18*(11), 14268–14284. https://doi.org/10.3390/molecules181114268
42. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S., & Ryadnov, M. G. (2023). DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.*, *7*(3), 144–161. https://doi.org/10.1038/s41570-022-00456-9
43. Seydel, C. (2023). DNA writing technologies moving toward synthetic genomes. *Nat Biotechnol.*, *41*(11), 1504–1509. https://doi.org/10.1038/s41587-023-02006-0
44. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison 3rd, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318
45. Green, M. R., & Sambrook, J. (2012). *Molecular Cloning: A Laboratory Manual* (4th ed.). Cold Spring Harbor Press.
46. Ball, P. (2018). A compassionate substance. How science and myth collide in water. *Lapham's Quart.*, *11*(Summer).
47. Pace, C. N., Scholtz, J. M., & Grimsley, G. R. (2014). Forces stabilizing proteins. *FEBS Lett.*, *588*(14), 2177–2184. https://doi.org/10.1016/j.febslet.2014.05.006
48. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427

49. 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., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
50. Mura, C., McCrimmon, C. M., Vertrees, J., & Sawaya, M. R. (2010). An introduction to biomolecular graphics. *PLoS Comput. Biol.*, *6*(8), e1000918. https://doi.org/10.1371/journal.pcbi.1000918
51. Leach, A. R. (2001). *Molecular modelling: Principles and applications*. Prentice Hall.
52. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz Jr, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, *117*(19), 5179–5197. https://doi.org/10.1021/ja00124a002
53. Allinger, N. L., Zhou, X., & Bergsma, J. (1994). Molecular mechanics parameters. *J. Mol. Struct. THEOCHEM*, *312*(1), 69–83. https://doi.org/10.1016/S0166-1280(09)80008-0
54. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem\@. Theory Comput.*, *13*(6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125
55. Homeyer, N., & Gohlke, H. (2012). Free energy calculations by the molecular mechanics Poisson-Boltzmann surface area method. *Mol. Inf.*, *31*(2), 114–122. https://doi.org/10.1002/minf.201100135

56. Bae, E., & Phillips Jr, G. N. (2005). Identifying and engineering ion pairs in adenylate kinases. *J. Biol. Chem.*, *280*(35), 30943–30948. https://doi.org/10.1074/jbc.M504216200
57. Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J., & Osuna, S. (2019). Hidden conformations in *Aspergillus niger* monoamine oxidase are key for catalytic efficiency. *Angew. Chem. Int. Ed.*, *58*(10), 3097–3101. https://doi.org/10.1002/anie.201812532
58. Thomas, L. L., Christakis, T. J., & Jorgensen, W. L. (2006). Conformation of alkanes in the gas phase and pure liquids. *J. Phys. Chem. B*, *110*(42), 21198–21204. https://doi.org/10.1021/jp064811m
59. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N., & Hilvert, D. (2013). Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.*, *9*(8), 494–498. https://doi.org/10.1038/nchembio.1276
60. Crean, R. M., Gardner, J. M., & Kamerlin, S. C. L. (2020). Harnessing conformational plasticity to generate designer enzymes. *J. Am. Chem. Soc.*, *142*(26), 11324–11342. https://doi.org/10.1021/jacs.0c04924
61. Robic, S., Guzman-Casado, M., Sanchez-Ruiz, J. M., & Marqusee, S. (2003). Role of residual structure in the unfolded state of a thermophilic protein. *Proc. Natl. Acad. Sci. U. S. A.*, *100*(20), 11345–11349. https://doi.org/10.1073/pnas.1635051100
62. Rosenthal, L., Rabolt, J. F., & Hummel, J. (1982). An investigation of the conformational equilibrium of *n*-butane in a solvent using Raman spectroscopy. *J. Chem. Phys.*, *76*(2), 817–820. https://doi.org/10.1063/1.443052
63. Vázquez-Figueroa, E., Yeh, V., Broering, J. M., Chaparro-Riggers, J. F., & Bommarius, A. S. (2008). Thermostable variants constructed via the structure-guided consensus method also show increased stability in salts solutions and homogeneous aqueous-organic media. *Protein Eng. Des. Sel.*, *21*(11), 673–680. https://doi.org/10.1093/protein/gzn048

64. Mayer, S., Rüdiger, S., Ang, H. C., Joerger, A. C., & Fersht, A. R. (2007). Correlation of levels of folded recombinant P53 in *Escherichia coli* with thermodynamic stability in vitro. *J. Mol. Biol.*, *372*(1), 268–276. https://doi.org/10.1016/j.jmb.2007.06.044
65. Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, *103*(15), 5869–5874. https://doi.org/10.1073/pnas.0510098103
66. Adams, M. W., & Kelly, R. M. (1998). Finding and using hyperthermophilic enzymes. *Trends Biotechnol.*, *16*(8), 329–332. https://doi.org/10.1016/s0167-7799(98)01193-7
67. Merz, A., Yee, M.-C., Szadkowski, H., Pappenberger, G., Crameri, A., Stemmer, W. P. C., Yanofsky, C., & Kirschner, K. (2000). Improving the catalytic activity of a thermophilic enzyme at low temperatures. *Biochemistry*, *39*(5), 880–889. https://doi.org/10.1007/pl00003701
68. Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.*, *17*(4), 263–270. https://doi.org/10.1002/humu.22
69. Joerger, A. C., Ang, H. C., & Fersht, A. R. (2006). Structural basis for understanding oncogenic P53 mutations and designing rescue drugs. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(41), 15056–15061. https://doi.org/10.1073/pnas.0607286103
70. Park, S. H., & Bao, G. (2021). CRISPR/Cas9 gene editing for curing sickle cell disease. *Transfusion and Apheresis Science*, *60*(1), 103060. https://doi.org/10.1016/j.transci.2021.103060
71. Pace, C. N., & Shaw, K. L. (2000). Linear extrapolation method of analyzing solvent denaturation curves. *Proteins Struct. Funct. Bioinforma.*, *Suppl 4*, 1–7. https://doi.org/10.1002/1097-0134(2000)41:4+%3C1::aid-prot10%3E3.3.co;2-u
72. Baxa, M. C., Haddadian, E. J., Jumper, J. M., Freed, K. F., & Sosnick, T. R. (2014). Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(43), 15396–15401. https://doi.org/10.1073/pnas.1407768111

73. Horovitz, A., & Fersht, A. R. (1992). Co-operative interactions during protein folding. *J. Mol. Biol.*, *224*(3), 733–740. https://doi.org/10.1016/0022-2836(92)90557-Z
74. Scholtz, J. M., Grimsley, G. R., & Pace, C. N. (2009). Solvent denaturation of proteins and interpretations of the m value. *Methods Enzymol.*, *466*, 549–565. https://doi.org/10.1016/S0076-6879(09)66023-7
75. Fersht, A. (1999). Chapter 17, Protein Stability. In *Structure and Mechanism in Protein Science*. Freeman.
76. Becktel, W. J., & Schellman, J. A. (1987). Protein stability curves. *Biopolymers*, *26*(11), 1859–1878. https://doi.org/10.1002/bip.360261104
77. Robertson, A. D., & Murphy, K. P. (1997). Protein structure and the energetics of protein stability. *Chem. Rev.*, *97*(5), 1251–1267.
78. Pucci, F., Kwasigroch, J. M., & Rooman, M. (2017). SCooP: An accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*, *33*(21), 3415–3422.
79. Pastore, A., Martin, S. R., Politou, A., Kondapalli, K. C., Stemmler, T., & Temussi, P. A. (2007). Unbiased cold denaturation: Low- and high-temperature unfolding of yeast frataxin under physiological conditions. *J. Am. Chem. Soc.*, *129*(17), 5374–5375. https://doi.org/10.1021/ja0714538
80. Razvi, A., & Scholtz, J. M. (2006). Lessons in stability from thermophilic proteins. *Protein Sci.*, *15*(7), 1569–1578. https://doi.org/10.1110/ps.062130306
81. Eijsink, V. G. H., Bjørk, A., G\a aseidnes, S., Sirev\a ag, R., Synstad, B., van den Burg, B., & Vriend, G. (2004). Rational engineering of enzyme stability. *J. Biotechnol.*, *113*(1-3), 105–120. https://doi.org/10.1016/j.jbiotec.2004.03.026

82. Dumon, C., Varvak, A., Wall, M. A., Flint, J. E., Lewis, R. J., Lakey, J. H., Morland, C., Luginbühl, P., Healey, S., Todaro, T., DeSantis, G., Sun, M., Parra-Gessert, L., Tan, X., Weiner, D. P., & Gilbert, H. J. (2008). Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure. *J. Biol. Chem.*, *283*(33), 22557–22564. https://doi.org/10.1074/jbc.M800936200
83. Shoichet, B. K., Baase, W. A., Kuroki, R., & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.*, *92*(2), 452–456. https://doi.org/10.1073/pnas.92.2.452
84. Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L., van Loon, A. P. G. M., & Wyss, M. (2002). The consensus concept for thermostability engineering of proteins: Further proof of concept. *Protein Eng.*, *15*(5), 403–411. https://doi.org/10.1093/protein/15.5.403
85. Magliery, T. J. (2015). Protein stability: Computation, sequence statistics, and new experimental methods. *Current Opinion in Structural Biology*, *33*, 161–168. https://doi.org/10.1016/j.sbi.2015.09.002
86. Jones, B. J., Kan, C. N. E., Luo, C., & Kazlauskas, R. J. (2020). Consensus Finder web tool to predict stabilizing substitutions in proteins. In D. S. Tawfik (Ed.), *Enzyme Engineering and Evolution: General Methods* (Vol. 643, pp. 129–148). Academic Press. https://doi.org/10.1016/bs.mie.2020.07.010
87. Musil, M., Stourac, J., Bendl, J., Brezovsky, J., Prokop, Z., Zendulka, J., Martinek, T., Bednar, D., & Damborsky, J. (2017). FireProt: Web server for automated design of thermostable proteins. *Nucleic Acids Res.*, *45*(W1), W393–W399. https://doi.org/10.1093/nar/gkx285

88. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L., & Bryan, P. N. (1989). Large increases in general stability for subtilisin BPN' through incremental changes in the free-energy of unfolding. *Biochemistry*, *28*(18), 7205–7213. https://doi.org/10.1021/bi00444a012
89. Schwehm, J. M., Fitch, C. A., Dang, B. N., García-Moreno E., B., & Stites, W. E. (2003). Changes in stability upon charge reversal and neutralization substitution in staphylococcal nuclease are dominated by favorable electrostatic effects. *Biochemistry*, *42*(4), 1118–1128. https://doi.org/10.1021/bi0266434
90. Trudeau, D. L., Kaltenbach, M., & Tawfik, D. S. (2016). On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.*, *33*(10), 2633–2641. https://doi.org/10.1093/molbev/msw138
91. Clarke, J., Henrick, K., & Fersht, A. R. (1995). Disulfide mutants of barnase I: Changes in stability and structure assessed by biophysical methods and x-ray crystallography. *J. Mol. Biol.*, *253*(3), 493–504. https://doi.org/10.1006/jmbi.1995.0568
92. Pace, C. N., Grimsley, G. R., Thomson, J. A., & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.*, *263*(24), 11820–11825.
93. Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.*, *151*(2), 261–287. https://doi.org/10.1016/0022-2836(81)90515-5
94. Perry, L. J., & Wetzel, R. (1984). Disulfide bond engineered into T4 lysozyme: Stabilization of the protein toward thermal inactivation. *Science*, *226*(4674), 555–557. https://www.jstor.org/stable/1693872
95. Matsumura, M., Becktel, W. J., Levitt, M., & Matthews, B. W. (1989). Stabilization of phage T4 Iysozyme by engineered disulfide bonds. *Proc. Natl. Acad. Sci. U. S. A.*, *86*, 6562–6566. https://doi.org/10.1073/pnas.86.17.6562

96. Roesler, K. R., & Rao, A. G. (2000). A single disulfide bond restores thermodynamic and proteolytic stability to an extensively mutated protein. *Protein Science*, *9*(9), 1642–1650. https://doi.org/10.1110/ps.9.9.1642
97. Kanaya, S., Katsuda, C., Kimura, S., Nakai, T., Kitakuni, E., Nakamura, H., Katayanagi, K., Morikawa, K., & Ikehara, M. (1991). Stabilization of *Escherichia coli* ribonuclease H by introduction of an artificial disulfide bond. *Journal of Biological Chemistry*, *266*(10), 6038–6044. https://doi.org/10.1016/S0021-9258(18)38080-3
98. Hazes, B., & Dijkstra, B. W. (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng. Des. Sel.*, *2*(2), 119–125. https://doi.org/10.1093/protein/2.2.119
99. Dombkowski, A. A. (2003). Disulfide by Design^{TM}: A computational method for the rational design of disulfide bonds in proteins. *Bioinformatics*, *19*(14), 1852–1853. https://doi.org/10.1093/bioinformatics/btg231
100. Suplatov, D., Timonina, D., Sharapova, Y., & Švedas, V. (2019). Yosshi: A web-server for disulfide engineering by bioinformatic analysis of diverse protein families. *Nucleic Acids Res.*, *47*(W1), W308–W314. https://doi.org/10.1093/nar/gkz385
101. Pikkemaat, M. G., Linssen, A. B. M., Berendsen, H. J. C., & Janssen, D. B. (2002). Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng.*, *15*(3), 185–192. https://doi.org/10.1093/protein/15.3.185
102. Stewart, E. J., Åslund, F., & Beckwith, J. (1998). Disulfide bond formation in the *Escherichia coli* cytoplasm: An *in Vivo* role reversal for the thioredoxins. *EMBO J.*, *17*(19), 5543–5550. https://doi.org/10.1093/emboj/17.19.5543
103. Zhou, H.-X. (2004). Loops, linkages, rings, catenanes, cages, and crowders: Entropy-based strategies for stabilizing proteins. *Acc. Chem. Res.*, *37*(2), 123–130. https://doi.org/10.1021/ar0302282

104. Watanabe, K., & Suzuki, Y. (1998). Protein thermostabilization by proline substitutions. *J. Mol. Catal. B Enzym.*, *4*(4), 167–180. https://doi.org/10.1016/s1381-1177(97)00031-3
105. Fu, H., Grimsley, G. R., Razvi, A., Scholtz, J. M., & Pace, C. N. (2009). Increasing protein stability by improving beta-turns. *Proteins*, *77*(3), 491–498. https://doi.org/10.1002/prot.22509
106. Matthews, B. W., Nicholson, H., & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U. S. A.*, *84*(19), 6663–6667. https://doi.org/10.1073/pnas.84.19.6663
107. Scott, K. A., Alonso, D. O. V., Sato, S., Fersht, A. R., & Daggett, V. (2007). Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(8), 2661–2666. https://doi.org/10.1073/pnas.0611182104
108. Liu, Z., Lemmonds, S., Huang, J., Tyagi, M., Hong, L., & Jain, N. (2018). Entropic contribution to enhanced thermal stability in the thermostable P450 CYP119. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(43), E10049–E10058. https://doi.org/10.1073/pnas.1807473115
109. 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*(46), 7745–7751. https://doi.org/10.1002/anie.200602795
110. Floor, R. J., Wijma, H. J., Colpa, D. I., Ramos-Silva, A., Jekel, P. A., Szymanski, W., Feringa, B. L., Marrink, S. J., & Janssen, D. B. (2014). Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem*, *15*(11), 1660–1672. https://doi.org/10.1002/cbic.201402128
111. Xiao, Z., Bergeron, H., Grosse, S., Beauchemin, M., Garron, M.-L., Shaya, D., Sulea, T., Cygler, M., & Lau, P. C. K. (2008). Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-temperature-guided sequence alignment. *Appl. Environ. Microbiol.*, *74*(4), 1183–1189. https://doi.org/10.1128/AEM.02220-07

112. Petsko, G. A. (2001). Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat." *Methods Enzymol.*, *334*, 469–478. https://doi.org/10.1016/S0076-6879(01)34486-5

113. Gribenko, A. V., Patel, M., Liu, J., McCallum, S., Wang, C., & Makhatadze, G. I. (2009). Rational stabilization of enzymes by computational redesign of surface charge–charge interactions. *Proc. Natl. Acad. Sci. U. S. A.*, *106*(8), 2601–2606. https://doi.org/10.1073/pnas.0808220106

114. Contessoto, V. G., de Oliveira, V. M., Fernandes, B. R., Slade, G. G., & Leite, V. B. P. (2018). TKSA-MC: A web server for rational mutation through the optimization of protein charge interactions. *Proteins*, *86*(11), 1184–1188. https://doi.org/10.1002/prot.25599

115. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987–2998. https://doi.org/10.1021/bi902153g

116. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Res.*, *33*(Web Server issue), W382–8. https://doi.org/10.1093/nar/gki387

117. Korkegian, A., Black, M. E., Baker, D., & Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. *Sci. Wash.*, *308*(5723), 857–860. https://doi.org/10.1126/science.1107387

118. Arabnejad, H., Dal Lago, M., Jekel, P. A., Floor, R. J., Thunnissen, A.-M. W. H., Terwisscha van Scheltinga, A. C., Wijma, H. J., & Janssen, D. B. (2017). A robust cosolvent-compatible halohydrin dehalogenase by computational library design. *Protein Eng. Des. Sel.*, *30*(3), 173–187. https://doi.org/10.1093/protein/gzw068

119. Jin, Y., Johannissen, L. O., & Hay, S. (2021). Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins*, *89*(8), 915–921. https://doi.org/10.1002/prot.26068

120. Weinstein, J. J., Goldenzweig, A., Hoch, S., & Fleishman, S. J. (2021). PROSS 2: A new server for the design of stable and highly expressed protein variants. *Bioinformatics*, *37*(1), 123–125. https://doi.org/10.1093/bioinformatics/btaa1071
121. Gao, D., Narasimhan, D. L., Macdonald, J., Brim, R., Ko, M.-C., Landry, D. W., Woods, J. H., Sunahara, R. K., & Zhan, C.-G. (2009). Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, *75*(2), 318–323. https://doi.org/10.1124/mol.108.049486
122. van den Burg, B., Vriend, G., Veltman, O. R., & Eijsink, V. G. H. (1998). Engineering an enzyme to resist boiling. *Proc. Natl. Acad. Sci. U. S. A.*, *95*(5), 2056–2060. https://doi.org/10.1073/pnas.95.5.2056
123. Bjørk, A., Dalhus, B., Mantzilas, D., Sirev\a ag, R., & Eijsink, V. G. H. (2004). Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface. *J. Mol. Biol.*, *341*(5), 1215–1226. https://doi.org/10.1016/j.jmb.2004.06.079
124. Kabashima, T., Li, Y., Kanada, N., Ito, K., & Yoshimoto, T. (2001). Enhancement of the thermal stability of pyroglutamyl peptidase I by introduction of an intersubunit disulfide bond. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1547*(2), 214–220. https://doi.org/10.1016/s0167-4838(01)00185-6
125. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., & Trout, B. L. (2010). Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B*, *114*(19), 6614–6624. https://doi.org/10.1021/jp911706q
126. Kuriata, A., Iglesias, V., Pujols, J., Kurcinski, M., Kmiecik, S., & Ventura, S. (2019). Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. *Nucleic Acids Res.*, *47*(W1), W300–W307. https://doi.org/10.1093/nar/gkz321

127. Van Durme, J., De Baets, G., Van Der Kant, R., Ramakers, M., Ganesan, A., Wilkinson, H., Gallardo, R., Rousseau, F., & Schymkowitz, J. (2016). Solubis: A webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel.*, *29*(8), 285–289. https://doi.org/10.1093/protein/gzw019
128. Suen, W.-C., Zhang, N., Xiao, L., Madison, V., & Zaks, A. (2004). Improved activity and thermostability of *Candida antarctica* lipase B by DNA family shuffling. *Protein Eng. Des. Sel.*, *17*(2), 133–140. https://doi.org/10.1093/protein/gzh017
129. Lawrence, M. S., Phillips, K. J., & Liu, D. R. (2007). Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.*, *129*(33), 10110–10112. https://doi.org/10.1021/ja071641y
130. Robinson, N. E. (2002). Protein deamidation. *Proc. Natl. Acad. Sci. U. S. A.*, *99*(8), 5283–5288. https://doi.org/10.1073/pnas.082102799
131. Mark, D. F., Lu, S. D., Creasey, A. A., Yamamoto, R., & Lin, L. S. (1984). Site-specific mutagenesis of the human fibroblast interferon gene. *Proc. Natl. Acad. Sci. U. S. A.*, *81*(18), 5662–5666.
132. Kontermann, R. E. (2009). Strategies to extend plasma half-lives of recombinant antibodies. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *23*(2), 93–109. https://doi.org/10.2165/00063030-200923020-00003
133. Strohl, W. R. (2015). Fusion proteins for half-life extension of biologics as a strategy to make biobetters. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *29*(4), 215–239. https://doi.org/10.1007/s40259-015-0133-6
134. Cheang, J. Y., & Moyle, P. M. (2018). Glucagon-like peptide-1 (GLP-1)-based therapeutics: Current status and future opportunities beyond type 2 diabetes. *ChemMedChem*, *13*(7), 662–671. https://doi.org/10.1002/cmdc.201700781
135. Feldmann, M. (2002). Development of anti-TNF therapy for rheumatoid arthritis. *Nat. Rev. Immunol.*, *2*(5), 364–371. https://doi.org/10.1038/nri802

136. Nielsen, J. E., & Borchert, T. V. (2000). Protein engineering of bacterial α-amylases. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1543*(2), 253–274. https://doi.org/10.1016/s0167-4838(00)00240-5
137. Jones, B. J., Lim, H. Y., Huang, J., & Kazlauskas, R. J. (2017). Comparison of five protein engineering strategies for stabilizing an α/β-hydrolase. *Biochemistry*, *56*(50), 6521–6532. https://doi.org/10.1021/acs.biochem.7b00571
138. Thieker, D. F., Maguire, J. B., Kudlacek, S. T., Leaver-Fay, A., Lyskov, S., & Kuhlman, B. (2022). Stabilizing proteins, simplified: A Rosetta-based webtool for predicting favorable mutations. *Protein Sci.*, *31*(10). https://doi.org/10.1002/pro.4428
139. Banta, S., Dooley, K., & Shur, O. (2013). Replacing antibodies: Engineering new binding proteins. *Annu. Rev. Biomed. Eng.*, *15*(1), 93–113. https://doi.org/10.1146/annurev-bioeng-071812-152412
140. Rossi, A. M., & Taylor, C. W. (2011). Analysis of protein-ligand interactions by fluorescence polarization. *Nat Protoc*, *6*(3), 365–387. https://doi.org/10.1038/nprot.2011.305
141. Pollard, T. D. (2010). MBOC technical perspective: A guide to simple and informative binding assays. *Mol. Biol. Cell*, *21*(23), 4061–4067. https://doi.org/10.1091/mbc.E10-08-0683
142. Jeffrey, P. D., Strong, R. K., Sieker, L. C., Chang, C. Y. Y., Campbell, R. L., Petsko, G. A., Haber, E., Margolies, M. N., & Sheriff, S. (1993). 26-10 Fab-digoxin complex: Affinity and specificity due to surface complementarity. *Proc. Natl. Acad. Sci. U. S. A.*, *90*(21), 10310–10314. https://doi.org/10.1073/pnas.90.21.10310
143. Tian, W., Chen, C., Lei, X., Zhao, J., & Liang, J. (2018). CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.*, *46*(W1), W363–W367. https://doi.org/10.1093/nar/gky473

144. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, *255*(5041), 178–183. https://doi.org/10.1126/science.1553543

145. Jones, B. J., Evans, R. L., Mylrea, N. J., Chaudhury, D., Luo, C., Guan, B., Pierce, C. T., Gordon, W. R., Wilmot, C. M., & Kazlauskas, R. J. (2020). Larger active site in an ancestral hydroxynitrile lyase increases catalytically promiscuous esterase activity. *PLoS ONE*, *15*, e0235341. https://doi.org/10.1371/journal.pone.0235341

146. DeLorbe, J. E., Clements, J. H., Teresk, M. G., Benfield, A. P., Plake, H. R., Millspaugh, L. E., & Martin, S. F. (2009). Thermodynamic and structural effects of conformational constraints in protein-ligand interactions. Entropic paradoxy associated with ligand preorganization. *J. Am. Chem. Soc.*, *131*(46), 16758–16770. https://doi.org/10.1021/ja904698q

147. Barros, E. P., Schiffer, J. M., Vorobieva, A., Dou, J., Baker, D., & Amaro, R. E. (2019). Improving the efficiency of ligand-binding protein design with molecular dynamics simulations. *J. Chem. Theory Comput.*, *15*(10), 5703–5715. https://doi.org/10.1021/acs.jctc.9b00483

148. Thakkar, S., Nanaware-Kharade, N., Celikel, R., Peterson, E. C., & Varughese, K. I. (2014). Affinity improvement of a therapeutic antibody to methamphetamine and amphetamine through structure-based antibody engineering. *Sci. Rep.*, *4*, 3673. https://doi.org/10.1038/srep03673

149. Zhao, Q., Ahmed, M., Guo, H., Cheung, I. Y., & Cheung, N.-K. V. (2015). Alteration of electrostatic surface potential enhances affinity and tumor killing properties of anti-ganglioside GD2 monoclonal antibody hu3F8. *J. Biol. Chem.*, *290*(21), 13017–13027. https://doi.org/10.1074/jbc.M115.650903

150. Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, *25*(10), 1171–1176. https://doi.org/10.1038/nbt1336
151. Selzer, T., Albeck, S., & Schreiber, G. (2000). Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.*, *7*(7), 537–541. https://doi.org/10.1038/76744
152. Bugnon, M., Röhrig, U. F., Goullieux, M., Perez, M. A. S., Daina, A., Michielin, O., & Zoete, V. (2024). SwissDock 2024: Major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. *Nucleic Acids Res*, *52*(W1), W324–W332. https://doi.org/10.1093/nar/gkae300
153. Zoete, V., Schuepbach, T., Bovigny, C., Chaskar, P., Daina, A., Röhrig, U. F., & Michielin, O. (2016). Attracting cavities for docking. Replacing the rough energy landscape of the protein by a smooth attracting landscape. *J Comput Chem*, *37*(4), 437–447. https://doi.org/10.1002/jcc.24249
154. Röhrig, U. F., Goullieux, M., Bugnon, M., & Zoete, V. (2023). Attracting Cavities 2.0: Improving the flexibility and robustness for small-molecule docking. *J. Chem. Inf. Model.*, *63*(12), 3925–3940. https://doi.org/10.1021/acs.jcim.3c00054
155. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., & Baker, D. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, *501*(7466), 212–216. https://doi.org/10.1038/nature12443
156. Morin, A., Kaufmann, K. W., Fortenberry, C., Harp, J. M., Mizoue, L. S., & Meiler, J. (2011). Computational design of an endo-1,4-beta-xylanase ligand binding site. *Protein Eng. Des. Sel.*, *24*(6), 503–516. https://doi.org/10.1093/protein/gzr006

Problems

1. *Gibbs energy and entropy estimate*. The goal of this question to compare an experimental equilibrium constant to one calculated using molecular mechanics. First, calculate the experimental equilibrium constant between the anti and gauche conformations for the conversion of gauche to anti. At equilibrium in the gas phase experimental measurements show that *n*-butane consists of a mixture of 70% anti and 30% gauche conformations.[62] Assume this refers to 300 °K. Next, calculate the equilibrium constant from the molecular mechanics calculations. The enthalpy comes from the molecular mechanics calculation (use data from Figure 4.12). To estimate the entropy difference between the anti and gauche conformations recall that there is only one anti conformation, but two enantiomeric gauche conformations. Combine the enthalpy and entropy values to calculate the Gibbs energy difference and convert this energy into an equilibrium constant. Compare the experimental and calculated values and comment on any differences.

2. *Protein Data Bank*. The PDB is a repository for the 3D structural data of proteins and nucleic acids. The data, typically obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world, are freely accessible on the internet.

a) *Searching for a structure file*. Go to the PDB website at http://www.rcsb.org. Each structure has a unique 4-digit pdb id. You can search for structures by name, by pdb id, by author, and by other criteria. The structure we are looking for is subtilisin from *Bacillus amyloliquefaciens* or subtilisin BPN'. How many structures of subtilisin BPN' are contained in the Protein Data Bank? Searching for 'subtilisin BPN'' in the search box at the top yields >120 structures, but specifying that 'subtilisin BPN'' must be in the UniProt molecule name (in the drop-down menu or the Advanced Search panel) returns ~60 structures. For example, the first search results include 3SSI, but

the second does not. Is file 1THM a structure of subtilisin BPN'? Explain your answer.

b) *Viewing the text of the structure file.* Find the entry with pdb id 1S01. Download the file (in PDB Format) to your computer. The file should be named 1s01.pdb; if it is named 1s01.cif.gz, look for a different download link where a menu pops up to allow you to chose PDB Format. Open the file with a text editor or a word processing program. How many protein atoms does the structure contain? How many water molecules (HOH)? What are the x, y, z coordinates for the Cα atom of Ser221? Note that the structure data does not include hydrogens.

3. *Introduction to PyMOL*. Download and install the most recent version (3.1) of PyMOL to your computer from http://pymol.org/ep

a) *Viewing the structure file with PyMOL*. Open the file 1S01.pdb with PyMOL. (Alternatively, you can directly download it from within PyMOL using the sequence: File $\rightarrow$ Get PDB... $\rightarrow$ enter 1S01 and click Download.) Use the trackpad or mouse to rotate and zoom. Try the different combinations of keyboard and mouse. Click on an atom, note that PyMOL tells you which atom it is. The protein is shown as a ribbon representation, but there are also water molecules, two isopropyl alcohol solvent molecules, and one calcium ion. Select Display $\rightarrow$ Sequence to show the amino acid sequence. Click on the calcium ion (large sphere). A small red dot indicates that it is selected. Selection also created an object called (sele); see gray rectangle in panel to the right of the structure. What is its residue number? Note that it is also highlighted in the sequence.

b) *Viewing the active site residues and saving your work.* Type the commands below, including spaces, but not the comment lines starting with '#'.

```
# create a new object named 'triad' containing
```

```
# residues numbered 32, 64 and 221
> select triad, resi 32 + resi 64 + resi 221
# show object triad in sticks representation
> show sticks, triad
# (or click on the 'S(how)' near triad (make sure
# triad is selected and select sticks)
# zoom in on the triad with:
> zoom triad
# hide water molecules
> hide everything, resn hoh
# or click on the H(ide) near 1S01 and choose waters
```

4. *Save your work two different ways.* Saving your session allows you to return to this point with all changes that you have made. Saving an image creates a static picture for use in a presentation or report. It loses all molecule information. Since it is an image file, you can adjust it (e.g., add a text label) with an image editing program.The difference between the commands is the extension chosen for the file name: pse versus png. PyMOL matches the information saved to the extension. The files are saved in your current directory. You many not know what your current directory is, so you may have to hunt for them. You can set your working directory under with File $\rightarrow$ Working Directory... $\rightarrow$. Alternatively, you can add a path to the file name like: `~/Desktop/file_name.pse`.

```
> save file_name.pse
# (or File -> Save Session, then type file_name.pse)
> save file_name.png
# (or File -> Export Image As -> PNG... ->
# Save PNG image as ... -> file_name.png)
```

5. *Measuring distances and specifying histidine tautomers in PyMOL.* Open your 1s01.pse file that you saved in the previous question. It should show

the protein as a cartoon, hidden water molecules, and the catalytic triad as sticks.

a) *Tautomers of histidine.* Neutral histidine can exist as one of two tautomers, sometimes called HID (hydrogen on Nδ2) and HIE (hydrogen on Nϵ1). Figure 4.13 shows catalytic triads showing these two tautomers. Draw potential hydrogen bonding patterns in both and explain why only one of these tautomers creates a catalytic configuration. (Review catalytic triads of serine proteases, if needed.)

b) *Identifying hydrogen bonds.* PyMOL can display hydrogen bonds. Select the triad (select triad) A(ction) $\rightarrow$ find $\rightarrow$ polar contacts $\rightarrow$ within selection. Click on the black space to deselect triad. The geometric criteria for the existence of a hydrogen bond between N and O are 1) an N-O distance or 2.9 Å and 2) a nearly linear N-H-O angle ($\geq$ 120°). Measure the distance and, if possible, the angle (you will need to add hydrogens: A(ction) $\rightarrow$ hydrogens $\rightarrow$ add polar). To measure the angle select Wizard $\rightarrow$ Measurement. Change the default measurement from "Distances" to "Angles" by clicking on "Distances" and choosing "Angles" from the pop-up menu. Then click on the three atoms that define the angle.

c) *Structure shows the incorrect tautomer of histidine*. PyMOL does not identify a hydrogen bond between OD1 or OD2 of Asp32 and ND of His 64. Why not? Are the distances between O and N appropriate for forming a hydrogen bond? Does the 1s01 structure contain the catalytic or the non-catalytic tautomers of histidine? Do you think the structures of the two tautomers differ enough that an x-ray experiment could distinguish the two?

d) *Specifying the correct tautomer of histidine.* To correct this error, you can specify which tautomer of histidine that PyMOL draws. The choices are: HIS = nothing specified, HID = neutral tautomer with hydrogen on ND, HIE = neutral tautomer with hydrogen on NE, HIP = hydrogens on both nitrogens; positively charged. Open pdb file

1S01.pdb and replace the names of all of the atoms in histidine 64 from 'HIS' to 'HID' as shown in Figs. Figure 4.14 and Figure 4.15. Save the file and reopen it with PyMOL. Remeasure the hydrogen bond distances between the serine, histidine, and aspartate and show that the structure now shows a catalytically productive hydrogen bond arrangement. However, one detail is still incorrect. The angle for the hydrogen bond between His NE2 and SerOG is 83°, which suggests a poor hydrogen bond. Suggest a reason why this conclusion is likely incorrect.

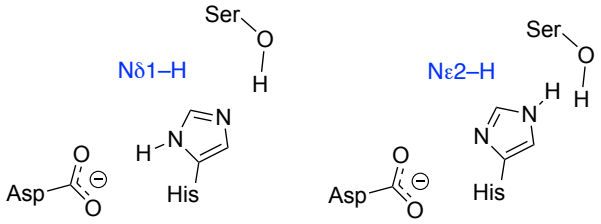


Figure 4.13. Only one tautomer of histidine creates a viable catalytic triad. One tautomer has the hydrogen on Nδ1, while the other has the hydrogen on Nϵ2.

```
ATOM    455  N   HIS A  64      -1.852  38.475  29.575  1.00  6.26           N
ATOM    456  CA  HIS A  64      -0.431  38.903  29.664  1.00  5.63           C
ATOM    457  C   HIS A  64      -0.018  39.823  28.532  1.00  4.99           C
ATOM    458  O   HIS A  64       0.508  40.917  28.774  1.00  4.27           O
ATOM    459  CB  HIS A  64       0.462  37.666  29.791  1.00  7.00           C
ATOM    460  CG  HIS A  64       1.882  37.897  30.144  1.00  5.95           C
ATOM    461  ND1 HIS A  64       2.883  38.035  29.195  1.00  6.43           N
ATOM    462  CD2 HIS A  64       2.491  37.967  31.341  1.00  6.26           C
ATOM    463  CE1 HIS A  64       4.038  38.217  29.803  1.00  6.84           C
ATOM    464  NE2 HIS A  64       3.834  38.154  31.126  1.00  6.08           N
```

Figure 4.14. Old listing of coordinates for His64 in pdb file 1s01 does not specify which tautomer should be drawn.

6. *Site-directed mutagenesis of subtilisin BPN'*.The Met222Ala substitution in subtilisin BPN' stabilized it to inactivation by bleach.

a) Select Met222 aand show it as sticks using a different color for the carbon atoms than the one used for the catalytic triad.

```
ATOM    455  N   HID A  64      -1.852  38.475  29.575  1.00  6.26           N
ATOM    456  CA  HID A  64      -0.431  38.903  29.664  1.00  5.63           C
ATOM    457  C   HID A  64      -0.018  39.823  28.532  1.00  4.99           C
ATOM    458  O   HID A  64       0.508  40.917  28.774  1.00  4.27           O
ATOM    459  CB  HID A  64       0.462  37.666  29.791  1.00  7.00           C
ATOM    460  CG  HID A  64       1.882  37.897  30.144  1.00  5.95           C
ATOM    461  ND1 HID A  64       2.883  38.035  29.195  1.00  6.43           N
ATOM    462  CD2 HID A  64       2.491  37.967  31.341  1.00  6.26           C
ATOM    463  CE1 HID A  64       4.038  38.217  29.803  1.00  6.84           C
ATOM    464  NE2 HID A  64       3.834  38.154  31.126  1.00  6.08           N
```

Figure 4.15. New listing of coordinates for His64 in pdb file 1s01 specifies the neutral tautomer with hydrogen on ND1.

b) To switch from viewing your protein to editing it, click on '3-button viewing' in the Mouse control legend on the bottom right. This click will toggle the mouse mode to '3-button editing' so that you can modify your protein. Mutate Met222 to alanine (Wizard $\rightarrow$ Mutagenesis $\rightarrow$ Protein $\rightarrow$) Then follow the prompts to select Met222 by clicking on it and choose a new amino acid by clicking on the light purple box labelled 'No Mutation'.

Answers

1. experimental = 70% anti in gas phase; K_{eq} = 2.3. A value greater than 1 indicates that anti if favored for the gauche $\leftrightharpoons$ anti equilibrium.

Now we will predict the same equilibrium constant using the molecular mechanics calculation of ΔH and an estimate of ΔS from microstates

$$\Delta G = \Delta H - T\Delta S$$

$$\Delta G = -900 \text{ cal/mol} - 300 \text{ K}(1.987 \text{ cal/mol} \cdot \text{K} \cdot \ln(1/2))$$

$$\Delta G = -900 + 413 = -500 \text{ cal/mol}$$

$$\Delta G = -RT \ln K_{eq}$$

$$-500/1.987 \times 300 = \ln K_{eq}$$

$-0.83 = \ln K_{eq}$; $K_{eq} = 2.3$!

2. a) 1THM is not a subtilisin, but a subtilisin-type serine proteinase. A search with the keyword 'subtilisin' anywhere in the record will return this protein, but not if the search is limited to the molecule name. b) 1938 protein atoms; note that 1939 is not an atom is has no coordinates; 215 water molecules; `ATOM 1529 CA SER A 221 x=7.506 y=40.969 z=32.045`

3. a) The calcium ion is residue 301. b) You should have a zoomed-in view of the catalytic triad.

5. a) Catalysis requires the histidine to act as a base and deprotonate the serine Oγ in the first step; this requires the Nδ 1 tautomer. b) 3.0 Å between H64 NE2 and S221 OG; 134° for the hydrogen attached to H64 NE2, which is a hydrogen bond. The hydrogen on S221 OG does not form a hydrogen bond because the angle is too acute, 83°. This H-O-C-C rotamer is likely not the orientation during catalysis. c) no, PyMOL has not added a hydrogen to N64 ND1, so no hydrogen bond is possible. The distance between H64 ND1 and D32 OD1 is 2.6 Å, which is short enough for a hydrogen bond; similarly the distance between H64 ND1 and D32 OD2 is 3.2 Å, which is also short enough. d) non catalytic; PyMOL has drawn the incorrect tautomer of H64; the pdb file does not specify tautomers because tautomers differ in the location of hydrogens, but the pdb file does not contain hydrogens.

5 Engineering More Stable Proteins

Summary. Protein function requires the folded protein form, but this form is unstable mainly because it readily unfolds into a flexible, unstructured form. Protein folding is favored by burying of hydrophobic side chains and hydrogen bonding between the amino acids. Protein unfolding is favored by the increase in flexibility of the main chain of amino acids upon unfolding. Protein stability is usually measured by the reversible unfolding of the protein with either heat or chemical additives like urea. Protein stabilization involves making substitutions that shift the folding-unfolding balance toward the folded form. Stabilizing substitutions can either stabilize the folded conformation or destabilize the unfolded ensemble. This tutorial emphasizes web-based tools to identify substitutions that stabilize proteins. Besides unfolding, other sources of protein instability are chemical modifications like oxidations or cleavage by proteases and aggregation of partly unfolded proteins into insoluble particles.

Key learning goals

- Proteins are dynamic, equilibrating collections of structures. Most of the structures are folded, but a small fraction are unfolded. This protein unfolding is the primary source protein instability. Stabilization shifts the folding-unfolding balance toward folding, but does not prevent unfolding.

- To measure protein stability, one measures the ratio of folded to unfolded protein. Under normal conditions, the amount of unfolded protein is too small to detect. Adding denaturants such as urea or heating the sample increases the amount of unfolded protein to measurable amounts. Extrapolation back to normal conditions reveals the stability of the protein.
- One way to stabilize proteins is to restore amino acid residues that are conserved in homologs. The rationale for this approach is evolution conserves residues that provide some benefit; one benefit is a contribution to stability.
- Another way to stabilize proteins is to destabilize the unfolded form. The main driving force for unfolding is the increased flexibility of the unfolded form. Decreasing the flexibility of unfolded form by adding disulfide crosslinks or introducing proline residues often stabilizes proteins.
- The most direct way to stabilize proteins is to create or strengthen attractive interactions between amino acids in the folded conformation. Although proteins will remain dynamic and still unfold, the stronger interactions either slow down unfolding or speed up refolding.

5.1 Introduction

Protein stability usually refers to resistance to unfolding. Stresses like high temperatures, organic cosolvents, high substrate or product concentrations, extremes of pH or high ionic strength can all cause unfolding. In many cases, stabilizing a protein to one stress also stabilizes it to other stresses. For example, proteins that tolerate high temperatures often also tolerate organic cosolvents.[63]

One advantage of more stable proteins is the ability to use them as biocatalysts in artificial environments where naturally-occurring proteins

are unstable. For example, using biocatalysis at higher than physiological temperatures yields faster reaction rates, higher substrate solubility, and lower solution viscosities. Reactions at high temperature can also avoid microbial contamination since most contaminating microbes cannot grow at high temperature. Biocatalysts that tolerates organic cosolvents and high concentrations of substrates and products simplify the scale-up of industrial processes by requiring less solvent, smaller equipment, and less complicated product isolation.

The second advantage of more stable proteins is extended lifetime or storage at ordinary temperatures. For example, manufacture of β-lactam antibiotics with penicillin G acylase recycles the immobilized enzyme many times to lower the overall cost. More stable therapeutic proteins last longer during storage and are more resistant to proteases in serum.

A third advantage of more stable proteins is higher yields of soluble protein during manufacture, especially for small, single-domain proteins. Over-expression of recombinant proteins in microbes creates high concentrations of protein. The unfolded forms can aggregate into insoluble particles called inclusion bodies. More stable single-domain proteins aggregate less and yield higher amounts of soluble protein.[64]

A fourth advantage is that more stable proteins are also more 'evolvable' or able to acquire beneficial traits. Substitutions that change protein function may destabilize proteins. If the destabilized protein fails to fold, then the potential improvement is lost. More stable proteins can tolerate greater numbers of destabilizing mutations and are thus more evolvable than their marginally stable variants.[65] One source of stable proteins is thermophiles and other extremophiles.[66] For example, the Pfu DNA polymerase used for the polymerase chain reaction (PCR) comes from the hyperthermophile Pyrococcus furiosus, Figure Figure 5.1. This polymerase catalyzes DNA synthesis at 72 °C and tolerates the high temperatures (95 °C) needed to dissociate complementary DNA strands.

One disadvantage of enzymes from thermophiles is the typically low catalytic activity of these enzymes at room temperatures.[67] Since ther-

Figure 5.1. The archaeon Pyrococcus furiosus is a hyperthermophile that grows best at 100 °C. The heat-stable Pfu DNA polymerase used in PCR comes from this microbe. This painting simulates a scanning electron micrograph. Image by Fulvio314 from https://en.wikipedia.org/wiki/Pyrococcus_furiosus (CC BY-SA 3.0).

mophiles do not live at room temperature, there is no selective pressure for high activity at room temperature. If the application requires activity at room temperature, enzymes from thermophiles may not be suitable. Pfu DNA polymerase has negligible activity at room temperature. Other limitation of enzymes from thermophiles is the lack of other needed protein characteristics. For example, applications may require unusual substrate specificity unavailable in thermophiles or human proteins to minimize immune response.

Many human disease-causing mutations are associated with amino acid substitutions that decrease protein stability.[68] For example, mutations that decrease the stability of fructose bisphosphate aldolase cause hereditary fructose intolerance and mutations of the tumor suppressor protein p53 cause some cancers.[69] The ability to predict substitutions that alter protein stability may help identify mutations that cause genetic diseases and treat them with genetic editing technologies. For example, sickle cell disease, caused by the lower stability of the Glu6Val variant of β-globin, has been treated by expressing the unmutated fetal β-globin gene.[70]

5.2 Native and denatured conformations equilibrate

The native protein state, N, is the folded, functional form. It is compact with the hydrophobic side chains mainly buried and the polar side chains mainly exposed to solvent. It has a specific structure (or similar set of structures), typically with α-helices, β-sheets, and turns folded in a particular arrangement, Figure Figure 5.2. While the structure is dynamic and moves, it has an overall stable structure. Most of the amino acids interact with each other; only amino acids on the protein surface interact with solvent water. In contrast, the denatured protein state, D, is not functional and is not a single state, but a collection or ensemble of more or less folded states. The main chain makes large, random motions and the amino acids interact mainly with solvent water, not with each other.

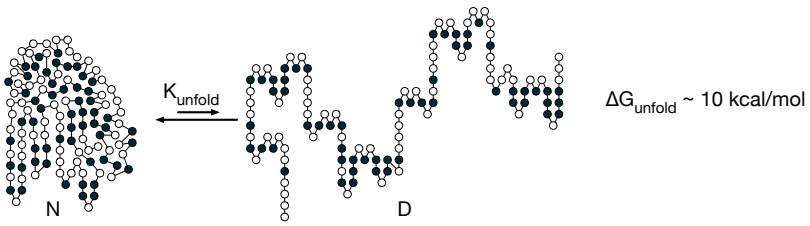


Figure 5.2. Schematic of protein unfolding showing the equilibrium between a compact folded native structure (N) and an ensemble of flexible unfolded states (D). Folding creates a specific structure that mostly buries hydrophobic amino acids (filled circles) in the interior of the folded structure while mostly exposing hydrophilic amino acids (open circles). Unfolding exposes more the amino acids to solvent. For clarity, the figure shows only one example unfolded structure, but in reality it is an ensemble of rapidly interchanging unfolded structures.

The native form of a typical protein is more stable than the denatured state by $\sim$ 10 kcal/mol, Figure Figure 5.3. A Gibbs energy of unfold-

ing, ΔG^{unfold}, of +10 kcal/mol corresponds to an equilibrium constant $e^{-(10{,}000/RT)}$ or 4.6×10^{-8} at room temperature indicating that unfolding is rare, eq. Equation 5.1. Folding and unfolding are fast for single domain proteins. Native forms continuously unfold to denatured states and these continuously refold to native forms.

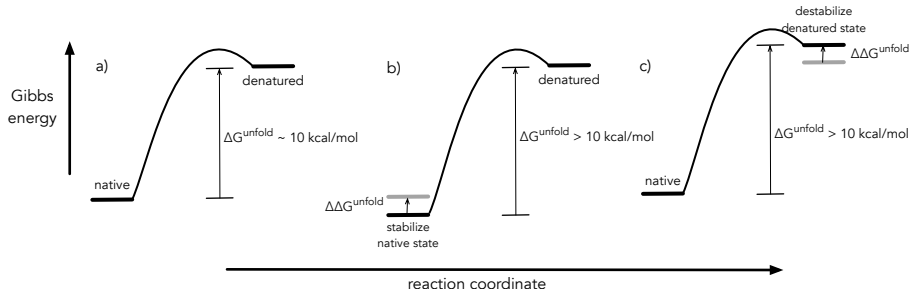


Figure 5.3. The Gibbs energy difference between the folded, native state of a protein and the denatured ensemble of states determines protein stability. The barrier to unfolding for single domain proteins is low. a) The denatured state for the original protein lies ~ 10 kcal/mol above the native state. b) Stabilization of the native form stabilizes the protein because it increases the energy difference between the native and denatured forms. c) Destabilization of the denatured form also stabilizes the protein because it increases the energy difference between the native and denatured forms.

$$\Delta G^{unfold} = -RTln\left(K^{unfold}\right) \text{ or } K^{unfold} = e^{-\Delta G^{unfold}/RT} \quad (5.1)$$

Protein solutions contain unfolded protein molecules even for stably folded proteins. In the case where the native form is ~ 10 kcal/mol more stable than the unfolded form, one in twenty-two million protein molecules unfolds at any given time in solution at room temperature, eq. Equation 5.2. This tiny fraction is too small to detect by spectroscopic

methods, but it is nevertheless a large number of molecules. A solution containing 1 mg of a 30 kDa protein contains 4×10^{16} molecules; of these, nearly a billion (10^9) are in the denatured form at any time.

$$K^{unfold} = \frac{[D]}{[N]} = 4.6 \times 10^{-8} \text{ or } [N] = 22 \times 10^6 \cdot [D] \qquad (5.2)$$

The main driving force for protein folding is the hydrophobic effect, which is the tendency of non-polar solutes to cluster in water. Burying a –CH_2 group contributes 1.1 ± 0.5 kcal/mol to protein stability. The hydrophobic effect provides $\sim$ 60% of the driving force to collapse the amino acid chain into a compact structure.[71] The folded form of a protein is its lowest energy conformation in dilute solutions. (In concentrated protein solutions, oligomeric or aggregated states may be lower in energy.) The chains orient to maximize the hydrophobic effect and also to make attractive interactions between amino acids: hydrogen bonds and other electrostatic interactions.

Hydrogen bonds contribute the remaining 40% of the driving forces for protein folding. The hydrogen bonds between protein atoms each contribute on average 1.1 ± 0.8 kcal/mol each to protein stability. The unfolded protein makes hydrogen bonds between protein and water, so this energy contribution is the net gain in hydrogen bond strength. Hydrogen bonds also define how the protein will fold into helices and sheets.

The dominant driving force for unfolding of the main protein chain is the increase in flexibility. The denatured ensemble is a collection of highly flexible forms. This flexibility creates many conformations or micro-states and the existence of these states is the main driving force for unfolding. The micro-states create a favorable entropy contribution, $-T\Delta S$, to the Gibbs energy. To find the Gibbs energy contribution of entropy, one multiplies the entropy change by temperature and a negative sign since an increase in entropy lowers Gibbs energy, eq. Equation 5.3,

where W_{folded} and W_2 are the different numbers of micro-states in the states being compared.

$$\Delta G_{unfolded-folded} = -T\Delta S_{2-1} = -RTln\left(\frac{W_2}{W_1}\right) \tag{5.3}$$

To find the effect of a molecule's conformational flexibility on Gibbs energy, one compares the number of conformations in the flexible state to the non-flexible state. For example, one can estimate the difference in Gibbs energy between the folded and unfolded states for one amino acid in a protein at 300 °K only due to differences in backbone flexibility while ignoring any differences in side-chain flexibility. Assume that the backbone has a single conformation in the folded state, N, but can adopt three staggered conformations along each of the two rotatable bonds (ψ, ϕ) in the unfolded state, D, Figure Figure 5.4.

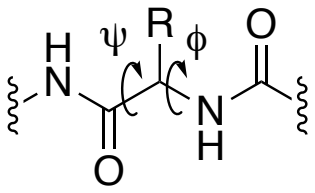


Figure 5.4. Each amino acid residue contains two rotatable bonds (ψ, ϕ) along the main chain. The C–N bond in the amide link does not readily rotate because it has partial double bond character.

The approach is to estimate the difference in the number of conformations available to the folded and unfolded states and then convert this difference to Gibbs energy using eq. Equation 5.3 above. The amino acid in the folded state has one available conformation while in the unfolded state, an amino acid can adopt $3 \cdot 3 = 9$ possible conformations. Converting this difference in possible conformations into Gibbs energy yields:

$$\Delta G_{N-D} = -RTln\left(\frac{W_2}{W_1}\right) = 1.987 \text{ cal/mol}\cdot^{\circ}\text{K}\cdot 300^{\circ}\text{K}\cdot ln\left(\frac{1}{9}\right) = +1.3 \text{ kcal/mol}$$

Thus, entropy due to differences in the backbone flexibility favors the unfolded state by 1.3 kcal/mol for each amino acid residue. More accurate estimates that account for the unequal likelihood of the nine conformations and differences in side chain flexibility yield a similar number: 1.4 kcal/mol per amino acid residue.[72] For a typical protein of 300 amino acids, this entropy effect of backbone flexibility contributes 400 kcal/mol, which is large as compared to the balance of ~ 10 kcal/mol in favor of the folded state. Hydrophobic interactions and hydrogen bonds in the native state counterbalance this flexibility advantage of the denatured state with a slightly larger Gibbs energy contribution. Protein stability is a balance between one set of large forces favoring the folded state and another set favoring the unfolded state.

5.3 Measuring the folding-unfolding equilibrium

Measuring the equilibrium constant requires measuring the ratio of folded and unfolded protein at equilibrium. Under normal conditions, the equilibrium amount of unfolded protein is too small to measure. Instead, researchers use destabilizing conditions where easily detectable amounts of both the folded and unfolded protein exist, Figure Figure 5.5. Typical destabilizing conditions are additives like urea or guanidinium chloride, changes in the solution pH, or increases in temperature. After measuring the equilibrium constant at these destabilizing conditions, researchers extrapolate back to normal conditions to get the desired equilibrium constant under normal conditions.

Most of this book treats protein folding and unfolding as a cooperative, two-state process (Figure Figure 5.2 above). The entire protein is either

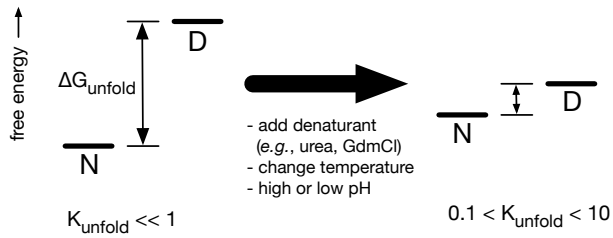


Figure 5.5. Changing the environment shifts the equilibrium between native and denatured states. The equilibrium constant can be measured when the relative amounts of the two forms are in the range of 0.1: 1 to 10:1, that is, when ΔG^{unfold} is within 1.4 kcal/mol of zero. GdmCl = guanidinium chloride.

folded or denatured, and the process is entirely reversible. There are no partially folded proteins or folding intermediates. The cooperative folding might start with one amino acid adopting an α-helix conformation, quickly followed by neighboring amino acids adopting the α-helix structure to create intrachain hydrogen bonds. This assumption of a simple, reversible two-state process is reasonable for single domain monomeric proteins. In these cases, adding stabilizing substitutions anywhere in the protein can contribute to stabilization since the entire protein unfolds simultaneously.

Changing the environment does not gradually change protein structure from folded to unfolded. Proteins remain fully folded or fully unfolded, but their ratio changes with the environment. Unfolding is cooperative because the interactions that stabilize the folded protein are stronger in combination than their individual contributions. The protein remains folded even when some stabilizing interactions break, but breaking too many destabilizes the others leading to complete unfolding of the protein.[73]

Section Section 5.5.1 below will briefly consider multiple domain and oligomeric proteins. Multiple domain proteins usually fold and unfold

stepwise via intermediates. The same stabilization strategies apply, but the substitutions must be in unfolded regions, not anywhere in the protein.

5.3.1 Denaturation with urea, ΔG_{H_2O}

Urea unfolds proteins because it 1) solvates exposed hydrophobic groups to reduce the hydrophobic effect and 2) forms hydrogen bonds to the backbone to disrupt secondary structures. Typical proteins unfold at 3-5 M urea; the solubility limit of urea is 18 M at room temperature.

The urea-induced denaturation experiment involves preparing solutions of protein in various concentrations of urea, waiting until the folding-unfolding reaches equilibrium, and measuring the amounts of native and denatured protein. The most common method to detect the folded and unfolded proteins is measuring the intensity of protein absorbance or fluorescence, Figure Figure 5.6, but any method that distinguishes between the native and denatured states is suitable. Some examples are measuring shifts in the wavelength of the fluorescence maximum, changes in the circular dichroism spectra, changes in the NMR spectra, decreases in catalytic activity, or changes in the dye fluorescence upon binding to hydrophobic regions of the unfolded protein.

The reasoning above predicts that the fluorescence emission wavelength increases for the denatured protein, but it is not easy to predict the relative intensity of this emission. The intensity change also depends wavelength chosen to monitor the fluorescence. For the example in Figure Figure 5.6a, the fluorescence intensity at 320 nm (near the emission maximum of the native protein) decreases as the protein unfolds, but the fluorescence intensity at 360 nm (near the emission maximum of the denatured protein) increases (not shown).

To extract the constants for unfolding equilibrium at different urea concentrations, $N \rightleftharpoons D$ from the fluorescence changes, one needs to know

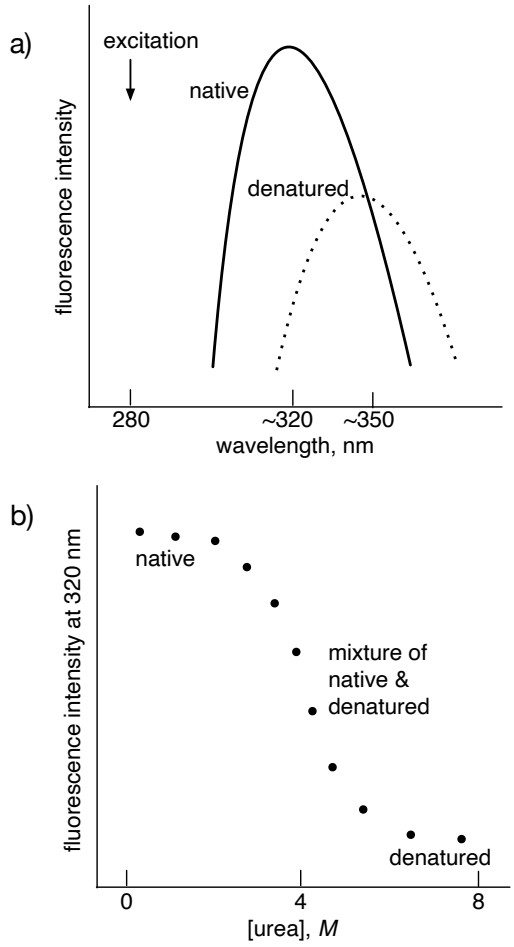


Figure 5.6. Unfolding of a protein in urea as monitored by tryptophan fluorescence. a) The indole ring of tryptophan absorbs light (absorbance maximum 280 nm; not shown), then emits light at lower energy (higher wavelengths). The emission maximum of a folded protein is typically 320 nm, which is similar to the emission maximum of indole in hexane. Upon unfolding the emission maximum shifts to 350 nm, which is similar to the emission maximum of tryptophan in water. b) Hypothetical urea unfolding data showing a decrease in fluorescence intensity at 320 nm as the protein unfolds in increasing concentrations of urea.

the fractions of native, F_N, and denatured protein, F_D. The ratio of these fractions is the equilibrium constant, eq. Equation 5.4. One assigns the fluorescence in low or no denaturant, Y_N, to the native protein and the fluorescence at high denaturant concentrations, Y_D, to the denatured protein. The supporting information includes a derivation of this equation. Intermediate values of fluorescence, Y_{obs}, correspond to a mixture of native and denatured states. Only data in the transition region of the denaturation experiment contribute to the analysis because when the protein is nearly completely folded or completely unfolded, the data do not precisely define the ratio of the two forms.

$$K^{unfold} = \frac{F_N}{F_D} = \frac{Y_{obs} - Y_N}{Y_D - Y_{obs}} \tag{5.4}$$

Applying the a linear extrapolation approach[74] to these measured equilibrium constants reveals the Gibbs energy of unfolding in water. This linear extrapolation approach starts by converting these equilibrium constants to Gibbs energy values according to eq. Equation 5.1 above and plotting them versus urea concentration, Figure Figure 5.7. The plot reveals a straight line with a negative slope since adding more urea favors protein unfolding. The line crosses the x-axis at $\Delta G^{unfold} = 0$. The goal of this experiment is to measure ΔG^{unfold} without any urea, that is, $[\text{urea}] = 0$. Extrapolation of the line to y-axis intercept reveals the Gibbs energy of unfolding in pure water, $\Delta G^{unfold}_{H_2O}$.

The equation for this line has a slope of -m and and a y-intercept at $\Delta G^{unfold}_{H_2O}$. The m-value is the absolute value of the slope and indicates the sensitivity of the protein to denaturant, which depends on the solvent-accessible surface area exposed upon unfolding.[74] A protein that unfolds completely has a higher m-value than a similar protein that unfolds only partially. Similar proteins with different m-values indicate different degrees of unfolding in their denatured states. Typical m-values for proteins undergoing urea-induced unfolding are ~ 1000

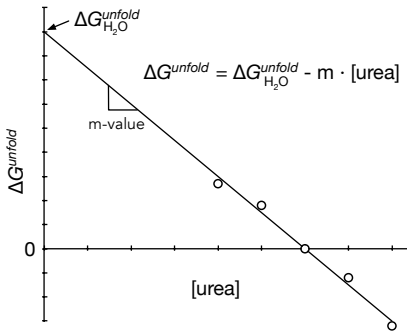


Figure 5.7. Linear extrapolation of the Gibbs energies calculated from the data in Figure Figure 5.6b reveals the Gibbs energy of unfolding in water. The data where the protein is nearly completely folded or completely unfolded are not included because the folded to unfolded ratio is not precisely defined by the data.

cal/(mol·M).This linear extrapolation approach is a convenient way to measure protein stability and the supporting information includes an example calculation.

$$\Delta G^{unfold} = \Delta G^{unfold}_{H_2O} - m \cdot [\text{urea}] \tag{5.5}$$

Positive Gibbs energies of unfolding in pure water indicate that unfolding of the protein is unfavorable. A higher Gibbs energy of unfolding indicates a more stable folded protein. The molecular basis for the linear relationship between unfolding Gibbs energy and urea concentration may be a binding interaction between urea and protein. The binding of urea destabilizes the protein and the fraction of urea bound increases linearly with concentration.

5.3.2 Heat denaturation, $\Delta G^{unfold}(T)$

Heat also unfolds proteins. Increases in temperature increase the entropy contribution to Gibbs energy of unfolding, eq. Equation 5.6 where ΔH^{unfold} and ΔS^{unfold} are, respectively, the changes in enthalpy and entropy upon unfolding. The entropy contribution due to the increase in flexibility of the unfolded protein increases at higher temperature and eventually leads to unfolding.

$$\Delta G^{unfold} = \Delta H^{unfold} - T\Delta S^{unfold} \tag{5.6}$$

Here we consider reversible heat-induced unfolding. Section Section 5.6.1 below considers irreversible unfolding caused by aggregation of the unfolded protein. As with urea-induced denaturation, this unfolding can be monitored spectroscopically, Figure Figure 5.8. For reversible unfolding, the equilibrium constant yields the Gibbs energy of unfolding at these high temperatures. The melting temperature, T_m, is the temperature where the amounts of folded and unfolded protein are equal; $\Delta G^{unfold} = 0$.

These measured Gibbs energies of unfolding at high temperatures can be extrapolated to room temperature, but this extrapolation is more complex than extrapolating the Gibbs energy of unfolding to pure water from a urea denaturation curve.[75] One might expect a plot of ΔGunfold versus temperature to yield a straight line according to eq. 3.7 above, where ΔSunfold is the slope and ΔHunfold is the y-intercept. However, the Gibbs energy of protein unfolding varies with temperature according to a more complex equation, eq. Equation 5.7.[76] Extrapolating the melting curve data to lower temperatures also requires measuring ΔC_p which is the change in heat capacity upon unfolding, Figure Figure 5.8.

$$\Delta G^{unfold} = \Delta H_0^{unfold} - T\Delta S_0^{unfold} + \Delta C_p\left[(T - T_0) - T\ln(T/T_0)\right] \tag{5.7}$$

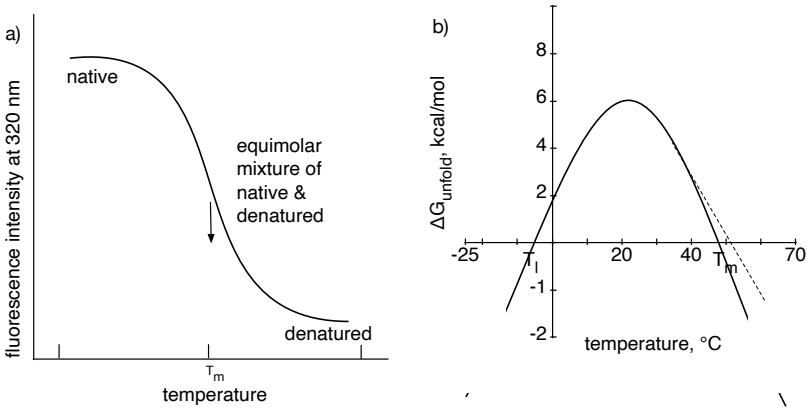


Figure 5.8. Heat induced unfolding of proteins. a) Hypothetical melting curve for a protein measured by the decrease in fluorescence intensity at 320 nm as the protein unfolds. The melting temperature, Tm, is the temperature where the amounts of folded and unfolded protein are equal. b) The Gibbs energy of unfolding of a protein varies non-linearly with temperature according to eq. Equation 5.6. The stability of a protein also decreases at lower temperatures; T_l marks the low temperature unfolding temperature. This plot uses values typical for a 300 aa protein: $T_0 = 300$ °K, $\Delta H_0 = 30$ kcal/mol, ΔS_0 = 80 cal/mol·deg, ΔC_p = 4.2 kcal/mol·deg. The dotted line is an approximation sometimes used near the melting temperature, see eq. Equation 5.8.

Here T_0 is the reference temperature and ΔH_0^{unfold} and ΔS_0^{unfold} are the enthalpy and entropy of unfolding at that temperature. For reactions involving small molecules, the heat capacity of starting materials and products are similar, and differences may be ignored. Eq. Equation 5.7 simplifies to eq. Equation 5.6 when $\Delta C_p = 0$.

However, the heat capacities of the folded and unfolded proteins differ significantly. Another way to state the reason for the more complex equation is that for protein unfolding ΔH^{unfold} and ΔS^{unfold} vary with temperature while eq. Equation 5.6 assumes they are constant. The extra terms in eq. Equation 5.7 account for the variation of ΔH^{unfold} and ΔS^{unfold} with temperature.

The unfolded protein has a higher heat capacity than the folded protein because unfolding exposes hydrophobic groups to water. Water surrounds these hydrophobic groups with an ice-like structure, which has slightly higher energy than bulk water. The fluctuation of water between bulk and ice-like structure increases the heat capacity.

The change in heat capacity for protein unfolding, ΔC_p, is always positive with a typical value of 14 cal/mol·deg per residue or 4.2 kcal/mol·deg for a 300 aa protein.[77] Differential scanning calorimetry can measure this change in heat capacity. Differential scanning calorimetry measures the heat required to increase the temperature of a sample. The baseline value before unfolding reveals the heat capacity of the folded protein, while the baseline value after unfolding reveals the heat capacity of the unfolded protein. The difference is the change in heat capacity. In the absence of experimental data, the web tool SCooP[78] predicts the parabolic curve for temperature-dependent unfolding of a protein based on its structure.

The parabolic curve in Figure Figure 5.8b predicts that proteins will also denature at low temperatures. For most proteins the predicted cold denaturation temperature, Tl, lies well below the freezing point of water as

shown in the sketch, but for some proteins, the cold denaturation temperature lies in the liquid range of water. For example, the yeast protein frataxin denatures at both 7 °C and at 30 °C.[79] At high temperatures the increase in conformational entropy (flexibility) of the protein atoms drives unfolding, while at low temperatures a decrease of the hydrophobic effect drives unfolding. At lower temperatures, the cost of ordering water molecules around a hydrophobic group decreases. As the tendency of hydrophobic groups to aggregate in water decreases, the folded form loses a stabilizing force.

Increases in melting temperature usually correspond to increases in protein stability. Comparing the unfolding or melting temperatures of protein variants is a common, approximate measure of their stability. One assumes that the unfolding process is similar for protein variants so the parabolic curve of Gibbs energy of unfolding versus temperature (Figure Figure 5.8b above) remains similar. One heats a protein sample slowly while monitoring the protein tryptophan fluorescence (Figure Figure 5.8a). The temperature corresponding to the midpoint of the change in fluorescence indicates the melting temperature. Usually one is on the right side of the parabola describing protein heat stability so an increase in melting temperature corresponds to a increase in protein stability.

A comparison of pairs of homologous proteins from thermophiles and mesophiles showed that the melting temperatures, T_m, were 31.5 °C higher and the Gibbs energies of unfolding 8.7 kcal/mol higher for the thermophiles.[80] From this comparison, we can derive a rule of thumb that stabilizing a protein by 1 kcal/mol will increase the melting temperature by 3.6 °C, eq. Equation 5.8, see dotted line in Figure Figure 5.8b.

$$\Delta\Delta G^{unfold}(\text{kcal/mol}) \sim \Delta T_m(^{\circ}\text{C})/3.6\ ^{\circ}\text{C/kcal/mol} \qquad (5.8)$$

In many cases, the Gibbs energy of heat-induced unfolding,$\Delta G^{unfold}(T)$,

is comparable to that obtained by urea unfolding. When these values differ, it is likely because the unfolded states differ in the two experiments.

5.4 Stabilization to cooperative unfolding

Stabilizing a dynamic object like a protein differs from stabilizing a rigid object like a chair. To stabilize a chair one can add a supporting brace, which keeps the chair in the desired structure. In contrast, a protein folds and unfolds continuously; it is dynamic. Adding a stabilizing interaction to a protein does not prevent unfolding. Instead, the stabilizing interaction shifts the folding-unfolding equilibrium toward folding so the protein spends more time in the folded form. Since the equilibrium constant is the ratio of the forward and reverse rate constants, eq. Equation 5.9, shifting the equilibrium constant also changes the rates of folding and unfolding. A stabilizing interaction may slow down unfolding, speed up folding, or both; but, unlike stabilizing a chair, stabilizing a protein does not *prevent* unfolding.

$$N \underset{k_{fold}}{\overset{k_{unfold}}{\rightleftharpoons}} D \tag{5.9}$$

Stabilizing a protein also differs from stabilizing a rigid object because one can stabilize a protein by destabilizing the unfolded form.[81] Everyday objects like chairs are not dynamic and cannot be stabilized this way. This book uses the expression 'stabilize a protein' or 'increase protein stability' to mean either stabilizing the folded form or destabilizing the unfolded form.

One reason that it is hard to predict which substitutions will stabilize a protein is that the structure of the denatured state (an ensemble of rapidly equilibrating structures) is unknown. Stabilizing substitution must affect

the folded and unfolded forms differently. A substitution that equally stabilizes (or destabilizes) the folded and unfolded forms does not change stability of the protein. X-ray crystal structures reveal changes to the folded form, but changes to the denatured state are invisible because its structure is unknown. Stabilization can occur without strengthening the folded structure. For example, seven amino acid substitutions dramatically stabilized xylanase (T_m increased by 25 °C), but the structures of the wild-type and stabilized proteins showed similar interactions between amino acids in both structures.[82] In this case, the stabilizing substitutions likely destabilized the unfolded state and changed its structure, but nature of these changes is unknown because the structure of the unfolded state is unknown.

A second reason that it is hard to predict stabilizing substitutions is that the substitutions must not hinder catalytic or binding ability of the target proteins. Substitutions that increase stability often also decrease activity and vice versa.[83] One reason for this trade-off is that catalysis and binding rely on interactions of the substrate or transition state with unsatisfied hydrogen bonds, exposed hydrophobic groups and unpaired charges in the protein. Satisfying these interactions with amino acid substitutions will stabilize the protein, but also disrupt binding and catalysis. Researchers typically avoid substitutions near the active site when searching for stabilizing substitutions to prevent disrupting binding and catalysis.

Some protein stabilization strategies seek to stabilize the native protein, others destabilize the denatured conformation, and, for a few, the molecular mechanism is unknown, Table Table 5.1. The design strategies that seek to stabilize the native protein require a protein structure or at least a homology model to start the design. A homology model is an extrapolated 3D-structure of a protein based on the known structure of a similar protein. Swiss Model and AlphaFold are web tools to generate homology models from protein sequences. Strategies that seek to destabilize the unfolded ensemble also require a protein structure to ensure that the changes do not distort the folded form. Strategies with unknown mecha-

nism do not require a protein structure. For example, restoring residues that are conserved in homologs (consensus sequence approach, see below) requires only a protein sequence. Random mutagenesis followed by screening to find stabilized variants also does not require a protein structure. Chapters 9 and 10 discuss random mutagenesis methods.

Table 5.1. Design strategies to stabilize proteins and web tools to implement these strategies.

Strategy	Rationale	Example
restore residues conserved in homlogs	not specified	Consensus Finder identifies conserved amino acids in homologs that are missing from target protein
add disulfide crosslinks	destabilize unfolded protein	Disulfide by Design identifies suitable locations, additional analysis needed to narrow choices
add Pro residues	destabilize unfolded protein	Analysis of structure to identify locations suitable for proline WHAT-IF
substitutions in or near flexible regions	stabilize folded protein	Random mutagenesis in or near flexible regions
random mutagenesis and screening	not specified	Random mutagenesis anywhere followed by screening for stabilized variants

Strategy	Rationale	Example
optimize electrostatic interactions	stabilize folded protein	TKSA-MC identifies destabilizing electrostatic interactions
optimize hydrophobic and other interactions	stabilize folded protein	Modeling with Rosetta and FoldX to identify stabilizing substitutions

To engineer a change, one needs to choose both the location of the substitution and the replacement amino acid. The 'restore residues conserved in homologs' approach predicts both, so it is easy to implement. The 'substitutions in or near flexible regions' approach defines the location, but not the replacements. Additional modeling or experiments are needed to identify the replacements. The 'add disulfide links' approach specifies that cysteine is the replacement amino acid, but additional modeling is needed to find suitable locations. Design strategies like 'improve hydrophobic and other interactions' are the least specific and require computer modeling to predict both the location and the replacement amino acid. Each of the approaches in Table Table 5.1 will be described in more detail below. The effect of each stabilizing mutation is typically small (0.5 kcal/mol), so substantial stabilization of the protein (2-4 kcal/mol) requires multiple substitutions and often more than one design strategy.

5.4.1 Restore conserved amino acids

Evolution conserves amino acids that contribute to protein function. This contribution may be to structure, catalysis, stability or other protein property. The consensus approach hypothesizes that conserved amino acids residues are more likely to contribute to protein stability than a

non-conserved amino acid. To use the consensus approach to stabilize a protein, one identifies amino acid residues conserved within homologous proteins, but which are missing in the target protein. Restoring these conserved amino acids is more likely to stabilize the protein than random substitutions.

Consensus design works because natural proteins rarely follow the consensus sequence at all structurally important positions. Natural proteins only have to be stable enough to fulfill their biological function; proteins with stabilities above a certain threshold will have no further selection advantage. There are many stabilizing residues, but individual proteins do not need all of them to reach the threshold of being stable enough. Thus, replacing a residue with the corresponding consensus amino acid may improve the stability or folding efficiency of a protein of interest. The same reasoning leads to the conclusion that the stability of a protein designed by consensus can be higher than that of the proteins in the alignment.

The consensus sequences approach does not hypothesize any molecular basis for the stabilization. It relies only on amino acid sequence comparisons from bioinformatics and does not require any structural information.[84,85] It is easy to implement, and several web-based tool to generate a consensus sequence for a protein are available: Consensus Finder[86] or FireProt.[87] Typically researchers restore one or a few conserved amino acids. Usually at least half of these predicted substitutions prove to be stabilizing. Pantoliano and coworkers first suggested that substitutions to the consensus amino acid may stabilize proteins and showed that the consensus-type mutation Met50Phe increased the unfolding temperature of subtilisin BPN' by 1.8 °C[88] and many others have used this approach. For example, a sequence alignment of 351 homologs of *Staphylococcus* nuclease A identified the His124Glu substitutions as potentially stabilizing, Figure Figure 5.9.

One can also restore all the conserved amino acids simultaneously. Lehmann and coworkers predicted the consensus sequence of fungal

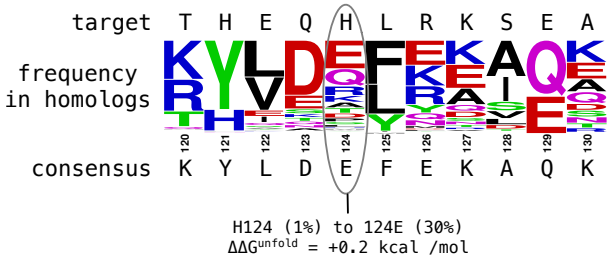


Figure 5.9. The most frequently-occurring amino acid in a multiple sequence alignment defines a consensus sequence. The target sequence (*Staphylococcus* nuclease A) was aligned with 351 homologs. The height of the letters in the logo plot reflects the frequency of the amino acid in the multiple sequence alignment. The sequence of the most common amino acid at each position is the consensus sequence. This consensus sequence predicts that an H124E substitution would stabilize this protein since 30% proteins in the alignment have E, but only 1% have H. Unfolding experiments showed that this substitution slightly stabilized Staphylococcus nuclease.[89] The WebLogo tool generated the logo plot of amino acid frequencies.

phytases based on the sequence alignment of thirteen sequences, Figure Figure 5.10.[84] The consensus phytase differed by at least 80 amino acids from the parent sequences. A synthetic gene encoded the consensus phytase containing all these changes. The consensus protein was 15-26°C more thermostable than any of its parents.

source	weight	sequence
Aspergillus terreus 9A-1	0.50	54- QVLARHGARS PTHSKEKAYA
Aspergillus terreus cbs	0.50	QVLARHGARS PTDSKTKAYA
Aspergillus niger var. *awamori*	0.33	QVLSRHGARY PTESKGKKYS
Aspergillus niger T213	0.33	QVLSRHGARY PTESKGKKYS
Aspergillus niger NRRL3135	0.33	QVLSRHGARY PTDSKGKKYS
Aspergillus fumigatus 13073	0.20	QVLSRHGARY PTSSKSKKYK
Aspergillus fumigatus 32722	0.20	QVLSRHGARY PTSSKSKKYK
Aspergillus fumigatus 58128	0.20	QVLSRHGARY PTSSKSKKYK
Aspergillus fumigatus 26906	0.20	QVLSRHGARY PTSSKSKKYK
Aspergillus fumigatus 32239	0.20	QVLSRHGARY PTASKSKKYK
Emericella nidulans	1.0	QVLSRHGARY PTESKSKAYS
Talaromyces thermophilus	1.0	QLLSRHGARY PTSSKTELYS
Myceliophthora thermophilia	1.0	QVLSRHGARA PTLKRAASYV
Consensus		**QVLSRHGARY PTSSK-KAYS**

Figure 5.10. Part of the sequence alignment (amino acid positions 54–73) of fungal phytases and the derived consensus sequence.[84] Amino acid sequences from the same species are weighted so that each species contributes equally to the consensus sequence. The '-' in the consensus sequence indicates that the consensus amino acid is ambiguous at this position.

At some positions, counting the numbers of occurrences identifies the consensus amino acid. For example, all thirteen sequences start with QVL, and most have S in the next position (boxed), so the consensus sequence begins with QVLS, Figure Figure 5.10. At other positions, identifying the consensus amino acid requires care to ensure unbiased sampling of sequences. Identifying the most common residues requires a uniform sampling of amino acid sequences throughout the evolutionary tree,

but the available amino acid sequences may not be evenly distributed. For the phytase example above, researchers reduced the bias of multiple sequences from different strains of the same species by weighting the amino acid sequences so that each *Aspergillus* species had equal weight. For example, the list included five sequences from *Aspergillus fumigatus* strains, Figure Figure 5.10. Each of these sequences was weighted by 0.2 in deriving the consensus to avoid bias toward this species. For example, consider the last amino acid in Figure Figure 5.10 (position 73, boxed). The first two sequences (*A. terreus*) have A at this position, the next three (*A. niger*) and two others (*Emericella* and *Talaromyces*) have S, the five *A. fumigatus* sequences have K, and the last sequence (*Myceliophthora*) has V. Although S and K both occur five times, K occurs only in the five *A. fumigatus* sequences and receives a total weight of 1.0, while S occurs in three different species and gets a total weight of 3.0. The consensus amino acid at this position is S. The consensus sequence is not unique, but depends on the set of comparison sequences. In later work, as more phytase sequences became available, researchers tested a revised consensus sequence, which further increased stability.

In other cases, a database search may yield thousands of similar sequences and it is impractical to assign different weights to the sequences. To minimize phylogenetic bias in these cases, one clusters similar sequences (typically >90% identity) together and uses only one sequence from each cluster for the consensus calculation. The web-based tools to find consensus sequences mentioned above use this clustering to minimize phylogenetic bias.

One disadvantage of the consensus sequence approach is that its conservative nature predicts small numbers of substitutions, often less than a handful or even none. Although experimental screening reveals that many stabilizing substitutions exist, the consensus sequence approach, by focusing on those conserved among homologs, misses most of them.

Ancestral proteins are extinct proteins that correspond to branch points

in a phylogenetic tree. Gene synthesis and expression of these proteins in microbial hosts resurrects these proteins. In many cases, ancestral proteins are more stable than modern proteins. Some ancestral proteins may be more stable because the earth was warmer in prehistoric times. However, ancestral proteins from times when the earth's temperature was only slightly warmer are also more stable, so other effects may contribute.[90] One contribution is that reconstructed ancestral proteins are often similar to the consensus sequence for that group since many descendants retain the ancestral residues. However, ancestral proteins differ from consensus proteins is two ways. First, ancestral proteins identify amino acids conserved within a related group, while consensus proteins include data from all homologous proteins, including those outside a group. Second, the consensus approach weights all sequences equally, but ancestral sequence reconstruction takes branch lengths into account, so more recent changes count less in an ancestral sequence prediction.

5.4.2 Destabilize the unfolded ensemble

The main property favoring the unfolded ensemble is its flexibility. Reducing this flexibility destabilizes the unfolded ensemble, thus shifting the folding-unfolding equilibrium toward folding. Two ways to reduce this flexibility are to add disulfide cross-links and to replace amino acid residues with proline. In both cases the replacement amino acid is defined; the challenge is to find suitable locations for the replacement. The ideal site in the folded structure would fit the replacement perfectly, so the only effect of the substitution is to reduce the flexibility of the unfolded ensemble.

Add disulfide cross links. Replacing a nearby pair of amino acid residues with cysteines followed by spontaneous oxidation creates a disulfide cross link, Figure Figure 5.11. Such disulfide links often stabilize proteins to unfolding. For example, the replacement of Ala43 and Ser80 in

Bacillus RNAse with cysteines followed by spontaneous oxidation yielded a more stable protein that unfolded in 5.77 M urea as compared to 4.58 M urea for the wild type.[91] This engineering involved both amino acid replacements and the formation of a cross link, both of which could contribute to changes in stability. To measure the effect of the cross link alone, the researchers compared the stabilities of the oxidized disulfide form with the reduced dithiol form of the protein. These two forms differ only in the presence or absence of a cross link. For the example above, the Gibbs energy of unfolding of the oxidized form was 3.2 kcal/mol higher than for the reduced dithiol form demonstrating that the cross link stabilized the protein.

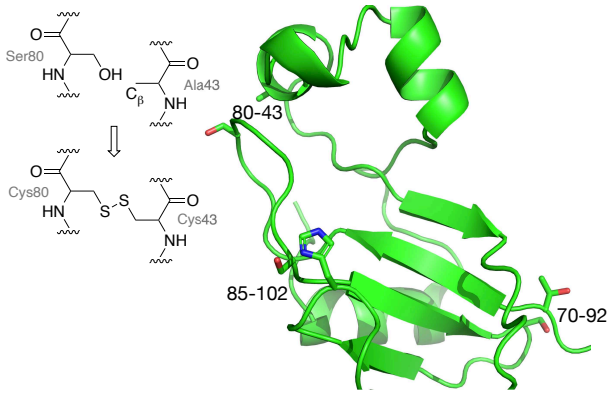


Figure 5.11. Replacement of Ser80 and Ala43 in *Bacillus* RNAse with cysteines followed by spontaneous oxidation created a disulfide cross link. This cross link stabilized the protein by 3.2 kcal/mol. Similar cross links introduced at two other locations had variable effects.

The reason that disulfide cross links stabilize proteins is that cross links reduce the flexibility of the unfolded protein. Upon unfolding, the cross link remains and limits the movement of the unfolded protein. Flexibility creates disorder (entropy), which is stabilizing. Reducing flexibility of the unfolded protein destabilizes it. The cross link constrains the dis-

tance between the linked cysteines. In the folded form, the cysteines are already close, so this constraint has little effect. In contrast, this constraint dramatically limits the movement and flexibility of the unfolded protein. Since increased flexibility and the associated increased entropy is the main reason that proteins unfold, reducing the flexibility of the unfolded form shifts the folding-unfolding equilibrium toward folding. This shift stabilizes the protein to unfolding.

A common misconception is that disulfide cross links prevent unfolding; they do not. Proteins remain flexible, dynamic structures even after adding disulfide cross links. They still unfold in urea solutions and upon heating. The RNAse protein in the example above still unfolded in urea solutions, although unfolding required higher urea concentrations that for wild-type. A similar misconception is that protein cross links strengthen and stabilize the folded protein state. They do not. The locations chosen for a disulfide cross link are already nearby in the folded protein. The introduction of the cysteines and the cross link may slightly change the folded protein stability, but this change is most often a destabilization, see below.

The chain entropy model predicts the amount of expected stabilization from a disulfide cross link. This model calculates the change in entropy due to reduced flexibility of the unfolded protein state caused by the disulfide cross link. This model assumes that the introduction of cysteines and formation of the disulfide cross link has no effect on the stability of the folded protein state. The amount of protein stabilization expected from disulfide cross link depends on the size of the ring created by the cross-link. Larger rings limit the flexibility of more amino acids and are therefore more stabilizing. Equation eq. Equation 5.10 below estimates the chain entropy effect of disulfide cross link on the unfolded state.[92]

$$\Delta\Delta S^{unfold} = -2.1 \text{cal/mol} \cdot {}^{\circ}K - \frac{3}{2} R \ln n \tag{5.10}$$

The $\Delta\Delta S^{unfold}$ refers to change due to the crosslink for the entropy change associated with unfolding; the n is the number of residues in the ring created by the crosslink and R is the gas constant. A link between amino acid residues 43 and 80 creates a ring of 38 residues; one more than the difference between 80 and 43. This equation comes from from polymer theory estimates of the probability that the two ends of the chain will coincide in the same volume element. If linked, the two ends must remain in the same volume element, while if unlinked they may move apart and allow greater numbers of conformations. To estimate the Gibbs energy change of the disulfide cross link using the chain entropy model, one multiplies by $-T$ where T is the temperature (in °K) yields the anticipated Gibbs energy contribution, eq. Equation 5.11.

$$\Delta\Delta G^{unfold} \text{ (cal/mol)} = -T \cdot \Delta\Delta S^{unfold} = T\left[2.1 + \frac{3}{2} \cdot 1.987 \cdot \ln n\right] \tag{5.11}$$

For example, the chain entropy model predicts that at 25 °C a cross link between amino acids 43 and 80 will stabilize a protein by 3.9 kcal/mol, which is slightly higher than the measured value of 3.2 kcal/mol in *Bacillus* RNase, Figure Figure 5.11. The stabilization predicted by eq. Equation 5.11 increases non-linearly with the number of amino acid residues in the ring, Figure Figure 5.12. The estimated stabilization for the smallest possible ring (two amino acids) is 1.2 kcal/mol. The stabilization increases to 3.0 kcal/mol at n = 15, which is the average separation of disulfide links in natural proteins.[93] With larger numbers of amino acid residues in the ring, the increase slows because the cross link restricts a smaller proportion of the motions in larger rings as compared to smaller rings.

The measured values for stabilization often do not match that predicted by the chain entropy model, likely because the chain entropy model ignores any effect on the folded state. The x-ray structure of *Bacillus*

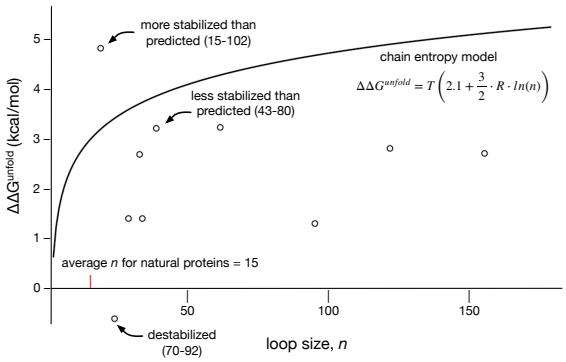


Figure 5.12. The predicted and measured effects of a cross link on protein stability. The open circles are the measured differences in stability between the oxidized protein (the two cysteines form a cross link) and the reduced protein (contains the two cysteines, but no cross link). The line is the predicted difference at 24 °C according to the chain entropy model, eq. Equation 5.11. Loop size, n, is the number of amino acids connected by the disulfide. In all but one case the measured effect is lower than predicted, which may be due to disruption of the folded protein by the cross link. Five data points are for T4 lysozyme,[94,95] three (marked by arrows) are for *Bacillus* RNase,[91] and one each for chymotrypsin inhibitor 2[96] and ribonuclease H.[97]

RNase 43-80 disulfide link mentioned above shows a small reorganization of backbone structure that may account for the slightly lower than predicted stabilization (3.2 vs. 3.9 kcal/mol).[91] Most of the examples in Figure Figure 5.12 seem to fit in this category. Two other disulfide links introduced in *Bacillus* RNase also are unusual cases. The disulfide at positions 85-102 stabilized the protein more than predicted by the chain entropy model, while the one at 70-92 was so much lower that it destabilizes the protein. The x-ray structures of the more-stable-than-expected protein showed no structural changes as compared to wild type, while the destabilized cross-linked protein showed substantial reorganization that apparently destabilized the folded protein.

To stabilize proteins by introducing a disulfide link, one must identify locations where the disulfide cross link fits in the folded protein without disrupting it. This step requires a structure of the target protein, although a homology model may also be used. Most prediction programs use geometry-only calculations where the distances and angles required for a disulfide crosslink are fit to the current fixed locations of possible amino acid pairs. Three geometry-only programs available as web tools are SSBOND (currently offline),[98] Disulfide by Design,[99] and YOSSHI.[100] The web interface requires only a pdb file of the protein structure for the calculation. YOSSHI also identifies disulfide cross links that occur in homologs. Since these cross links have been tested by natural selection, they may be prioritized for testing. Figure Figure 5.13 shows part of the results of an SSBOND prediction.

Typically the programs first identify amino acid pairs with suitable Cβ-Cβ distances (2.9-4.6 Å) and second, estimate the energy cost of distorting a disulfide bond to fit. Links with high predicted strain energies should be avoided. For example, conformation 1 in Figure Figure 5.13 between Ala4 and Asp89 is predicted to be destabilizing. The expected strain energy (total energy, TOTNRG column) is 6.2 kcal/mol, which is larger than the predicted stabilization energy for this link using eq. Equation 5.11 (4.6 kcal/mol). Since disulfide cross links that create larger rings are predicted

C_β–C_β distance test

NR	RES1	--	RES2	NAME1	NAME2	CB DIST	CA DIST
1	4	--	89	ALA	ASP	4.561	6.400
2	15	--	18	LEU	TYR	3.902	5.848
3	16	--	201	ASP	ALA	3.945	6.239

C_β–S_γ, S_γ–S_γ distance & angle test

	SGDIST	X1	X2	X3	X2'	X1'	CHINRG	TAUNRG	DISNRG	TOTNRG
1	CONFORMATION BETWEEN :			4 -	89	ALA ASP				
1	2.030	173.2	-134.8	-132.4	-62.1	-156.9	5.83	.24	.10	6.18
2	CONFORMATION BETWEEN :			15 -	18	LEU TYR				
1	2.030	69.2	156.5	-78.3	-88.6	120.1	4.40	.53	.42	5.35
2	2.030	-156.7	-98.7	-66.0	153.3	-26.6	5.48	.65	.21	6.35
3	CONFORMATION BETWEEN :			16 -	201	ASP ALA				
1	2.030	34.5	171.5	109.6	84.9	97.0	4.45	.15	.04	4.64
2	2.030	-90.4	117.7	85.9	-149.2	-159.6	4.32	.39	.02	4.73
3	2.030	1.2	-122.6	-103.9	-75.2	-165.5	5.25	.11	.02	5.37
4	2.030	-99.2	174.3	-114.1	137.2	82.4	5.41	.37	.87	6.65

Figure 5.13. The program SSBOND predicts potential locations for a disulfide link in a protein. A possible pair of amino acids in a protein: Gly and Asp. The program first identifies amino acid pairs with Cβ-Cβ distances in the range of 2.9–4.6 Å. In the case of glycine, which does not have a Cβ atom, SSBOND predicts its possible location from the location of C, N, and Cα. An example SSBOND calculation for a haloalkane dehalogenase protein (pdb ID = 1ede) with default settings found 81 pairs that match the Cβ-Cβ distances. The first three are shown here. In the second stage, SSBOND checks for suitable Cβ-Sγ distances, Sγ-Sγ distances, and proper angles; then calculates the energy cost for deviation from ideal geometry (TOTNRG = total energy). SSBOND ignores any potential interactions of the disulfide link with rest of the protein. This second stage eliminated two pairs leaving 79 possible pairs. The third pair (Asp16-Ala201) has predicted strain energy of at least 4.6 kcal/mol, while equation 3 above estimates 5.3 kcal/mol of stabilization at 25 °C. The prediction is stabilization by 0.7 kcal/mol. An experimental test revealed an increase in the melting transition of 5.2 °C[101] which corresponds to 1.5 kcal/mol of stabilization according to ΔG^{unfold}(kcal/mol) $\sim$ ΔT_m(°C)/3.6.[80]

to be more stabilizing by the chain entropy model, one normal favors larger rings over smaller rings.

Since geometry-only calculations ignore interactions between amino acids, one must consider these before choosing sites for mutagenesis. First, most researchers also avoid cross links near the active site to prevent possible disruption of catalysis. Second, geometry-only calculations do not consider possible loss of favorable interactions due to removal of existing amino acids, nor do they consider potential bumping interactions between the cysteines and adjacent amino acids. Examination of the protein structure can identify these potential problems. Some researchers also include a molecular dynamics simulations to generate alternative protein conformations.[101] These simulations could identify additional locations where a disulfide cross link can fit or identify locations that should be avoided since the cross link causes large shifts in the backbone that may destabilize the folded protein state.

Disulfide bonds usually form spontaneously upon air oxidation of the cysteines. If the application requires the protein to remain in the cytoplasm, which is a reducing environment, the disulfide bonds may not form spontaneously making this stabilization approach unsuitable. Mutant strains of *E. coli* where the enzyme thioredoxin reductase is inactivated can form disulfide cross links within the cytoplasm.[102]

While disulfides are the most common way to cross link the amino acid chain for protein stabilization, there are other possibilities.[103] For example, connecting the N- and C-termini into a ring also restricts the main chain motion of the unfolded protein and stabilizes proteins to unfolding. Formation of this ring requires special methods as well as a protein fold where the N- and C-termini are near one another.

Introduce proline residues. One can also reduce the backbone flexibility of the unfolded ensemble without creating crosslinks by selective amino acid substitutions. Proline is the least flexible amino acid because its ring limits its backbone conformations. Similarly, glycine is the most flexible amino acid since no side chain limits the possible backbone confor-

mations. Replacing any amino acid with proline is expected to reduce the flexibility of the denatured ensemble and stabilize the folded protein. Replacing glycine with any other amino acid should have a similar effect. To achieve a net stabilization, these substitutions should not otherwise stabilize the denatured ensemble and must not destabilize the folded form with unfavorable interactions. As in other cases, one should avoid changes in the active site of an enzyme to prevent disrupting catalysis.

flexibility: Gly > Ala & 17 other aa > Pro

Stabilization starts by finding a location where introducing a proline does not disrupt the native structure or interactions. Substitution of an amino acid with proline at such a location should cause no change in the Gibbs energy of the native state, Figure Figure 5.14. The Gibbs energy of the denatured state for the proline variant will be higher than for the wild type due to reduced flexibility of the denatured state.

Introducing proline is a reliable approach to stabilizing proteins. Replacing a typical amino acid with proline reduces the number of conformation in the denatured ensemble by an estimated factor of 7.5 or a ΔS of -4 cal/mol· deg,[104] eq. Equation 5.12, which corresponds to a Gibbs energy contribution of 1.2 kcal/mol at 298 °K.

$$\Delta S = R \ln \left[\frac{\text{\# of conformations for typical aa}}{\text{\# of conformations for proline}} \right] = R \ln (7.5) = 4 \text{ cal/mol·deg} \tag{5.12}$$

This approach requires identifying a suitable location for the proline. The main chain angles for proline fit well in three places: near the start of an α-helix, the i+1 position in a type I or II β-turn[104] or the i position of a type II β-turn.[105] These choices of proline location consider only main chain angles; the success of any substitution also assumes minimal changes to side chain interactions. The WHAT-IF web tool at can identify

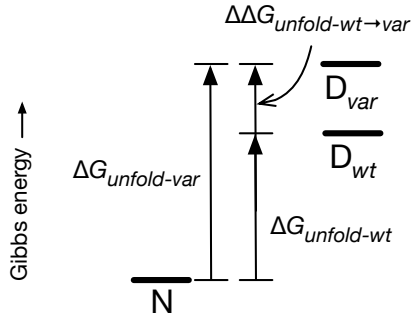


Figure 5.14. Hypothetical free energy changes upon stabilizing a protein with an alanine to proline substitution. Unfolding the native (N) wild-type protein to the denatured state (D) has a Gibbs energy of unfolding of $\Delta G_{unfold-wt}$. The location for the alanine to proline substitution is chosen so that the proline fits without altering the stability of the native state. The proline substitution limits the flexibility of the denatured state, thus the Gibbs energy of the denatured variant state is higher than the wild type and the Gibbs energy of unfolding the variant is higher. The increase in stability of the variant, $\Delta\Delta G_{unfold-wt\rightarrow var}$, is positive indicating that it is less favorable to unfold the variant as compared to wild type.

these locations in a protein structure (Click on 'mutation prediction', then 'Suggest proline mutations'). The first example of protein stabilization by proline substitution was an Ala82Pro substitution in T4 lysozyme, which occurs near the start of an α-helix and stabilized the protein by 0.8 kcal/mol, Table Table 5.2.[106] The restricted backbone angles of proline fit well near the start of an α-helix and the rest of the structure also fit a proline residue. The first four residues of a helix lack a partner to which the main chain N-H can donate a hydrogen bond, so the lack of an N-H in proline is not a disadvantage at the start of a helix. The typical success rate for stabilization upon proline substitution is $\sim 50\%$.

Table 5.2. Stabilization of lysozyme from phage T4 by substitutions that reduce the flexibility of the denatured state.

Protein	Location of substitution	T_m, °C	$\Delta\Delta G^{unfold}$ (kcal/mol)
wild type	none	64.7	0
Ala82Pro	near start of α-helix	66.8	0.8
Gly77Ala	near start of α-helix	65.6	0.4

Although replacing glycine with alanine also reduces the flexibility of the denatured state, other considerations limit the effectiveness of this substitution. Replacing glycine residues with alanine reduces the number of conformations available to the unfolded ensemble by approximately a factor of three, which corresponds to an Gibbs energy contribution of 0.4 kcal/mol at 298 °K, which is smaller than that of proline substitution, 1.2 kcal/mol. As with the addition of proline, one must ensure that the glycine to alanine substitution does not strain the folded form. Since glycine can adopt conformations that are not accessible to alanine (e.g., a left-handed helix conformation which occurs in some β-turns), re-

placements at these locations would destabilize proteins. Second, alanine can create stabilizing interactions within the denatured state that counterbalance the reduced flexibility so the net effect may be zero.[107] The Gly77Ala example in Table Table 5.2 above stabilized the protein, but the x-ray structure suggests that it stabilized the folded form due to the burial of the hydrophobic methyl group and not due to changes in flexibility of the denatured state.

5.4.3 Stabilize the folded form

Strengthening interactions between amino acids in the folded protein is the most obvious approach to stabilizing proteins. Starting from the three-dimensional structure of the protein, one designs various improved interactions. One difficulty with this approach is that proteins already create stabilizing interactions, so this design searches for additional interactions. One must predict both the location for the changes and the replacement amino acids.

Flexible regions identify weak spots in proteins. The effect of flexible regions in proteins can be confusing. Flexibility is a weakly stabilizing feature because it increases entropy.[108] This flexibility may contribute a factor of two, or 0.4 kcal/mol, to stability. While this flexibility is stabilizing, it also indicates that interactions with other amino acids are weak. Creating strong interactions between amino acids such as hydrogen bonds and hydrophobic interactions (several kcal/mol) in place of the weak stabilization of flexibility can be a net gain for stability, Figure Figure 5.15. Thus, it is not removing flexibility that stabilizes the protein, but the addition of new interactions between the amino acids. Loops on the surface and the N- and C-termini of the protein are often the most flexible.

There are two steps to this stabilization approach: first to identify the flexible regions of the protein and second, to identify stabilizing substitutions. One way to identify flexible areas is molecular dynamics simulation, which directly models protein motion.[101] Another way to identify

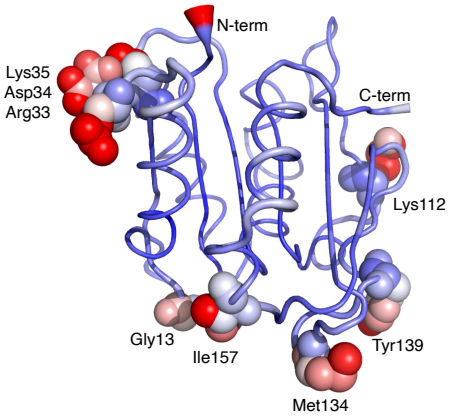


Figure 5.15. X-ray structure of a lipase (pdb code 1isp) colored to show the B-factor from low (blue) to high (red). The most flexible regions are the N– and C–termini (on the left) and several loop regions at the top. Reetz and coworkers29 replaced the residues in eight regions with the highest flexibility (B-factor), excluding the N- and C-termini, using random mutagenesis. Stabilizing substitutions (2-4 °C increase in T_m) occurred in the six regions where the residues shown as spheres.

flexible regions is the high B-factors in the x-ray crystal structure.[109] B-factors describe the spreading of electron density assigned to that atom. This spread may be due to movement of the atom during the x-ray analysis (temperature-dependent atomic vibrations), or may be due to the atom occupying several fixed positions in the structure (static disorder), which suggests motion in solution. Errors in model-building can also increase B-factors. The B-FITTER program identifies the twenty amino acids in a protein that have the highest B-factors. B-FITTER calculates the average of the B-factors for all atoms (excluding hydrogen) for each amino acid and displays a list of the twenty amino acid residues with the highest B-factors.

The second step is to identify stabilizing substitutions. One can make changes within the flexible region or next to flexible region. Interactions between amino acids require partners so either approach can yield improvements, but one comparison found larger stabilization with the replacements in the neighboring areas.[110] Some researchers targeted the flexible regions with random mutations,[109] while others used modeling to predict stabilizing substitutions.[110]

Copy features from thermotolerant homolog. Proteins from microorganisms that live in extreme environments (such as high temperatures >80 °C; extremes of pH, high salt, high pressure) tolerate their extreme environment better than homologous proteins from mesophiles (organisms that grow best at moderate temperatures). One protein stabilization strategy is to transfer the amino acids responsible for this extra stability in proteins from extremophiles to the corresponding proteins from mesophiles. This difficulty of this strategy is identifying which amino acids to transfer since some sequence differences impart increased stability, but most of the differences reflect random genetic drift. One approach relies on amino acid sequence comparison within the extremophiles. The expectation is that stabilizing amino acids are conserved in the proteins from extremophiles, but missing from the target protein from mesophiles. This approach differs from the consensus sequence approach, identifies residues conserved among all homologs,

including both those from extremophiles and mesophiles. One pitfall of this approach is that residues may be conserved within extremophiles because they are closely related, not because they contribute to stability. For example, comparison of a pectate lyase with four homologs from thermophiles identified nine substitutions common to the enzymes from thermophiles, but absent in the target enzyme.[111] However, only one of these nine substitutions (Arg236Phe) significantly increased the thermostability of the target lyase (12-fold), likely by improving hydrophobic packing.

A second approach relies on the structures of proteins from extremophiles and the identification of stabilizing interactions within them. The stabilizing interactions are not apparent from the structure, so identifying them requires modeling. For example, researchers identified which ion pairs in adenylate kinase from a thermophile contribute to stability using molecular dynamics simulations.[56] During the simulated movement of the protein some ion pairs broke apart, while three of them remained tightly paired. Transfer of these tightly-paired ion pairs into a less stable homolog increased its melting temperature by 0.8 to 3.3 °C.

Stabilizing interactions are not apparent from a comparison of structures because no single factor dominates.[112] Each protein uses a unique mixture of interactions to increase protein stability. For example, many proteins from thermophiles contain more extensive hydrogen bond networks to strengthen electrostatic interactions, but some proteins from thermophiles do not include more extensive hydrogen bond networks but are nevertheless stable at high temperature. Similarly, some, but not all, proteins from thermophiles contain increased atomic packing to strengthen hydrophobic interactions, increased numbers of ion pairs, shortened loops to minimize interactions with the solvent, increased occurrence of Ala in helices and increased oligomerization to enhance interactions between amino acids in the folded state. Other stabilizing features are not apparent in the folded structure because their effect is to

destabilize the denatured state. This wide range of features strengthens the argument that there are multiple paths to protein stabilization.

Remove destabilizing electrostatic interactions on the protein surface. Charged residues create favorable electrostatic interactions between residues with opposite charge and unfavorable interactions between residues with the same charge. Optimizing these interactions in the folded protein, especially by removing destabilizing interactions on the surface of the protein, stabilizes the protein. For example, five substitutions on the surface of an acylphosphatase increased the Gibbs energy of unfolding by 2.2 kcal/mol without affecting catalytic activity.[113] Three of these substitutions reversed the charge of the residues and two substitutions introduced new charges.

The two reasons to choose substitutions on the protein surface are that they are likely to be far from the active site and that they are likely to maintain good solvation. Avoiding mutations near the active site increases the probability that the modified protein will maintain catalytic activity (or binding in the case of a binding protein). Choosing residues with good solvation for substitution makes the prediction of stabilizing or destabilizing more accurate. The effect of a charged amino acid residue on protein stability depends on the differences in both electrostatics and solvation in the folded versus unfolded states. Since all residues are assumed to be well-solvated in the unfolded state, choosing residues that are well solvated in the folded state allows the prediction to ignore solvation in the comparison. One reason to avoid substitutions on the protein surface is that they may have a smaller effect on stability than substitutions in the interior of the protein. The environment of residues on the protein surface changes less drastically upon unfolding than the environment of residues in the interior of the protein. Nevertheless, residues on the surface move apart when the protein unfolds, thereby changing the electrostatic interactions.

TKSA-MC is a web-tool to identify residues for replacement due to unfavorable electrostatic interactions.[114] It calculates the electrostatic contri-

bution of all the charged residues in the protein and suggests replacing them if they make an unfavorable electrostatic contribution and lie on the protein surface. It does not suggest what the replacement amino acid should be, but uncharged polar residues or oppositely charged residues are the obvious choices. The electrostatic calculation requires a protein structure and considers the pH of the solution, the distance between the charges, different dielectric constants for the protein and solvent, and the fraction of the residue exposed to solvent. For example, TKSA-MC predicts that replacement of residues Asp36, Asp40, and Glu42 in Staphyloccocal binding protein would stabilize it, Figure Figure 5.16.

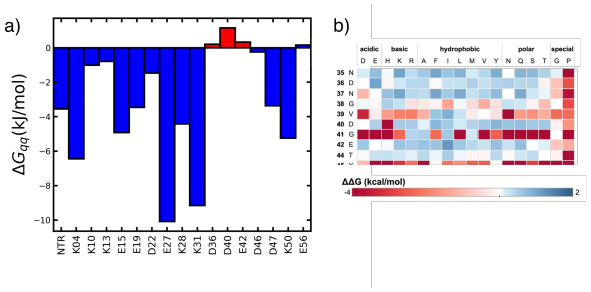


Figure 5.16. Stabilization of Staphyloccocal binding protein by replacement of destabilizing residues on the protein surface. a) Prediction of electrostatic interactions at pH 7 by web tool TKSA-MC based on the x-ray crystal structure (PDB id: 1pga). Positive values indicate destabilizing interactions. Replacement of the destabilizing residues filled in red (Asp36, Asp40, and Glu42), which indicates those on the surface, is predicted to stabilize this protein. b) Blue squares indicate replacements of amino acids in Staphyloccocal binding protein that stabilize the protein. In agreement with predictions, most replacements of Asp36, Asp40, and Glu42 lead to a more stable protein.

Most electrostatic interactions are stabilizing and only a few are destabilizing, so this approach identifies only a few locations for mutagenesis.

The ability to predict new stabilizing interactions would yield additional locations for mutagenesis, but this approach has not been tested with this web tool.

Modeling to identify stabilizing substitutions. Rosetta[115] and FoldX[116] are modeling programs widely used to predict protein stability. Both are available online. The programs model the physical interactions of the atoms in the folded form, but Rosetta adds statistical analysis of different properties extracted from protein databases.

FoldX is force field specifically for predicting protein stability. FoldX requires an experimental structure of a protein as the starting point. It does not do geometry optimization or conformational searching; it only calculates unfolding Gibbs energy for the protein and variants of the protein. For protein engineering, it predicts substitutions that stabilize or destabilize a protein. The force field equation contains terms for steric clash and electrostatic interaction like the physics-based force fields, but it also contains terms for solvation and entropy, Equation 5.13.

$$\begin{aligned}\Delta G = {} & a \cdot \Delta G_{vdw} + b \cdot \Delta G_{solvH} + c \cdot \Delta G_{solvP} + d \cdot \Delta G_{wb} \Rightarrow \text{solvation terms} \\ & + e \cdot \Delta G_{hbond} + f \cdot \Delta G_{el} + g \cdot \Delta G_{kon} \qquad \Rightarrow \text{electrostatic terms} \\ & + h \cdot T\Delta S_{mc} + k \cdot T\Delta S_{sc} \qquad \Rightarrow \text{entropy terms} \\ & + l \cdot \Delta G_{clash} \qquad \Rightarrow \text{steric clash term}\end{aligned} \tag{5.13}$$

These added terms estimate the Gibbs energies of various interactions of the amino acids with each other in the folded form as compared to interaction with solvent, which mimics the unfolded form. For example, the four solvation terms estimate the attractive van der Waals interaction between water and the protein (from the energy of removal of a protein atom from water to vapor phase), the desolvation of hydrophobic and polar groups upon folding (from the energy of removal of a protein atom from water to an organic solvent), and specific interactions with water

when the water molecules that make more than two hydrogen bonds with the protein. In these cases the water molecules are included as atoms in the calculation. The terms in the FoldX force field were scaled with respect to each other to match the predicted Gibbs energy to experimental measurements of >1000 stabilizing and destabilizing amino acid substitutions. A FoldX calculation is part of the HotSpot Wizard web tool (Sumbalova et al., 2018) at: http://loschmidt.chemi.muni.cz/hotspotwizard.

Rosetta's equation for energy is a hybrid of simple terms that model physical interactions combined with knowledge-based terms to improve accuracy. These knowledge-based terms come from statistical analysis of known protein structures. For example, while FoldX estimates electrostatic interactions using Coulomb's law and hydrogen bonding terms, Rosetta also weights the estimate by the probability that the two charged atoms occur nearby in PDB structures. Molecular modeling methods typically identify substitutions to increase hydrophobic interactions and various electrostatic interactions. For example, Rosetta predicted improved hydrophobic interactions in cytosine deaminase with three substitutions (A23L, I140L, V108I), which increased the apparent melting temperature by 10 °C.[117] The supporting information contains a detailed example of using Rosetta to identify a stabilizing substitution.

The energies from a molecular mechanics calculation like Rosetta are energy relative to a hypothetical unstrained molecule. A geometry optimization gives you a stable, reasonable structure, but the energy value needs a comparison value to make sense. For example, if you calculate the energy of the wt and a mutant, you can conclude which one of those folded structures is more stable. (The folded state is a collection of many conformations, so you are assuming that this single geometry-optimized structure is a good representative all folded structures that contribute to the folded state.) If you further assume that the energies of the unfolded states are identical, then the difference between the energies of the wt and mutant is the difference in unfolding energies. The units in Rosetta are 'Rosetta energy units' which are similar to kcal/mol.

The disadvantage of computational predictions of stabilizing substitutions is their low accuracy, often below 10%. Computer predictions are often incorrect either because the calculated energy of the molecular structure is incorrect or because the structure itself is incorrect and the protein favors a different conformation.

The first type of error is due to approximations used to speed up calculations (e.g., fixing the protein backbone, omitting explicit water molecules, ignoring entropic effects) and due to training of computers using small proteins; in short, the force field is inaccurate. To minimize this first type of errot, Floor and coworkers[110] suggested checking for the following in computer-predicted substitutions: 1) steric clashes, 2) internal cavities, 3) solvent-exposed aromatic residues, 4) solvent-exposed methionine, 5) a hydrogen bond donor or acceptor that has fewer hydrogen bonding interactions than in wild type, 6) a large hydrophobic patch on the protein surface, 7) uncompensated removal of a salt bridge, and 8) destabilization of an α-helix by removal of α-helix capping. The supporting information contains PyMOL scripts to check for some of the problems. Even with these checks the success rate for predicting substitutions to stabilize a haloalkane dehalogenase was only 9.2%.

The second type of error, an incorrect structure, is an error of conformational sampling. Computer modeling usually adjusts the residues near the substitution, but in reality a substitution may also affect more distant regions or cause more dramatic rearrangements in the main chain positions. Simulating alternative conformations using molecular dynamics with explicit water molecules adds these interactions and improves the success rate to 13-19%.[101,118] Finding these changes is computationally expensive since it requires long molecular dynamics simulations of protein motions. Machine learning approaches may be a faster alternative by using statistical patterns to accelerate conformational searching.[119]

Even with a perfect structure of the folded protein, errors in predicting

the effect of a substitution remain because the computations ignore the effect of the substitution on the unfolded protein state. Protein stability depends on the difference between the folded and unfolded states; ignoring the unfolded state is a major approximation.

Another approach to increasing success is to average the predictions of numerous methods. The assumption is that errors arising from different simplifications will cancel out. For example, the web server PROSS (Protein Repair One-Stop Shop[120]) combines the homology search used by the consensus sequence approach with computational design. First, PROSS searches for homologs to identify which substitutions are allowed at each position. The rationale for this evolution-based constraint is to favor variants that maintain their original molecular function. While the consensus sequence approach predicts stabilizing substitutions from the frequency of occurrence, PROSS adds an energy calculation to predict which substitutions are best. This calculation requires a three-dimensional structure of target protein. The rationale for adding energy calculations is that individual proteins differ so that an amino acid residue that is suitable in most homologs may not be suitable in the target protein. For example, stabilization of acetylcholine esterase involved a replacement for glycine at position 416. There was no clear choice for a replacement based on the frequency of occurrence because nine amino acids, including glycine, appeared at this position with similar frequency. Computational modeling of the replacements using Rosetta eliminated the most commonly-occurring histidine because it fit poorly and suggested third-most-frequently occurring glutamine because it formed a hydrogen bond with a nearby tyrosine.

5.5 Stabilization to stepwise unfolding

So far the discussion assumed that the protein is a single domain protein that unfolds cooperatively, reversibly, and in a single step. Multiple

domain proteins likely unfold stepwise with one domain unfolding first followed by other domains. For these proteins, stabilizing substitution must stabilize the domain that unfolds first. Proteins may also associate into oligomers. Oligomerization of folded proteins creates stabilizing interactions between amino acids at the protein-protein interface. (If the interactions were destabilizing, the proteins would not associate.) Creating additional interactions at the protein-protein interface is a new stabilization strategy available for oligomeric proteins.

5.5.1 Multiple domain proteins

Multiple domain proteins may unfold in a single step like single domain proteins, but more often they unfold stepwise with each domain unfolding separately. Stepwise unfolding creates an metastable intermediate with part of the protein folded and part unfolded. The stabilization must focus on the domain that unfolds first.

Bacterial cocaine esterase consists of three domains: an α/β-hydrolase domain, a cap domain, and a jelly roll domain. Molecular dynamics simulations indicated that the cap domain unfolded first. Two substitutions in the cap domain (T172R/G173Q) increased the half-life from ~ 0.2 h to 6 h at 37 °C,[121] which corresponds to stabilization of 2.1 kcal/mol. Modeling suggested that the T172R substitution strengthened interactions within the cap domain, while the G173Q substitution strengthened interactions between the cap and α/β-hydrolase domains by creating a new hydrogen bond.

Eijsink and coworkers dramatically stabilized a neutral protease from *Bacillus*, NprT, to self-degradation by stabilizing it to unfolding, Table Table 5.3.[122]. Proteolytic degradation proceeds from the unfolded form and does not depend strongly on amino acid sequence for this broad specificity protease, so stabilization to unfolding slows the proteolytic degradation. The substitutions increased the apparent melting temperature of the protein. Upon melting, the protease self-degrades rapidly.

The stabilizing substitutions were either copied from a heat tolerant homolog or rationally designed to stabilize the protein. The half-life at 70 °C increased dramatically from <0.5 to 170 min, which corresponds to 4.0 kcal/mol. The researchers named this stabilized enzyme 'boilysin' because of its ability to remain active at 100 °C.

Table 5.3. Substitutions that stabilize neutral protease NprT to unfolding and self-proteolysis at >100 °C.

ΔT_m, °C	Substitution	How identified	Molecular basis for stabilization
+12.3	Ala69Pro Thr63Phe	heat tolerant homolog heat tolerant homolog	less flexible unfolded domain hydrophobic interaction
+7.6	Gly58Ala Thr56Ala Ala4Thr	heat tolerant homolog heat tolerant homolog heat tolerant homolog	less flexible unfolded domain not determined not determined
+6.9	Ser65Pro 8-60Cys link	rational design rational design	less flexible unfolded domain less flexible unfolded domain

The stabilizing mutations all cluster in one region of the protein, Figure Figure 5.17. This multiple domain protein unfolds stepwise, and this region of the protein is the part that unfolds first and is then cleaved by

other protease molecules. Stabilizing this region to unfolding was the key to preventing proteolytic degradation.

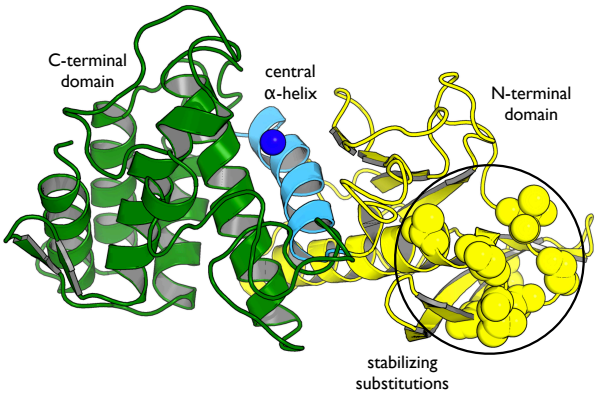


Figure 5.17. Cartoon representation of a model of thermolysin-like protease. The active site zinc ion (blue sphere) lies near the central a-helix (light blue) between the N-terminal domain (yellow) and C-terminal domain (green). The protease's stability depends on partial unfolding, so stabilizing substitutions (spheres within circle) cluster in that region: the N-terminal domain of the protein, in particular in the 55–69 surface loop. This model of protease NprT (P06874) created by Swiss-Model using the structure of thermolysin as the template, which has 87% identical amino acids.

5.5.2 Oligomeric proteins

Dimerization or oligomerization of monomeric native state proteins creates new protein-protein interactions while decreasing protein-water interactions and therefore stabilizes the folded native form, Figure Figure 5.18. In contrast, if unfolded or partially unfolded forms dimerize and oligomerize, then these interactions stabilize the unfolded structures, which eventually leads to aggregation.

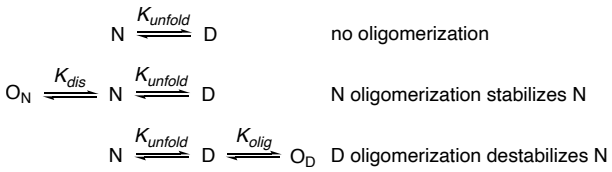


Figure 5.18. Dimerization or oligomerization of the native state (N) stabilizes the native state by shifting the overall distribution of conformations away from the denatured state (D). Strengthening interactions between native monomers stabilizes proteins to denaturation. In contrast, dimerization or oligomerization of the denatured state shifts the distribution of conformations away from the native state and eventually leads to aggregation. K_{dis} is the equilibrium constant for dissociation oligomers of native form, O_N, to monomers; K_{olig} is the equilibrium constant for association of denatured form into oligomers, O_D.

Substitutions at the oligomer interface that strengthen the interaction between monomers will stabilize proteins. In dramatic example, a replacement to remove a negative charge at the dimer interface of malate dehydrogenase (Glu165Gln or Glu165Lys) increased the melting temperature by 24 °C.[123] Another approach is to covalently link the monomers into dimers by introducing a disulfide link. Comparison of the structures of a peptidase and a thermostable homolog identified a cross-link between monomers in the thermostable homolog. Adding the this link to the peptidase increased its denaturation temperature by 30 °C.[124]

5.6 Weaknesses other than unfolding

While unfolding is the weak point of most proteins, in some cases, other weaknesses dominate protein instability. Cooking irreversibly aggregates egg white proteins so that they do not refold and restore a liquid

upon cooling. Other examples are chemical modification of proteins such oxidation or degradation by proteases. These other weaknesses may be related to unfolding. Proteins unfold at least partially before they aggregate or are degraded by proteases. Chemical modification can promote unfolding. A different type of protein instability is short serum half-life, which depends on biochemical processes that clear proteins from the bloodstream.

5.6.1 Protein aggregation

Protein aggregation is the irreversible association of partially or fully unfolded proteins into insoluble particles. Aggregation-prone regions assemble by intermolecular β-structured interactions to form the core of the aggregate. Aggregation can occur when unfolded proteins encounter one another. For example, cooking an egg first unfolds the egg white proteins to expose hydrophobic regions to the solvent. These regions associate with similar exposed hydrophobic regions of other egg white protein molecules leading to oligomerization and eventually insoluble aggregates, eq. Equation 5.14, Figure Figure 5.19. Here k_{agg} is the aggregation rate constant.

$$N \overset{K^{unfold}}{\rightleftharpoons} D \overset{k_{agg}}{\longrightarrow} \text{aggregate} \tag{5.14}$$

Aggregation can also occur under mild conditions. For example, overexpressing protein in bacteria creates high concentrations of unfolded protein. If protein folding is slow, the unfolded proteins can aggregate into insoluble particles called inclusion bodies. Biotherapeutic proteins can also aggregate during storage, which creates a danger of an immune response to the aggregates.

Measuring a loss of function after heating identifies aggregation as a contribution to protein instability. For example, one measures the enzyme

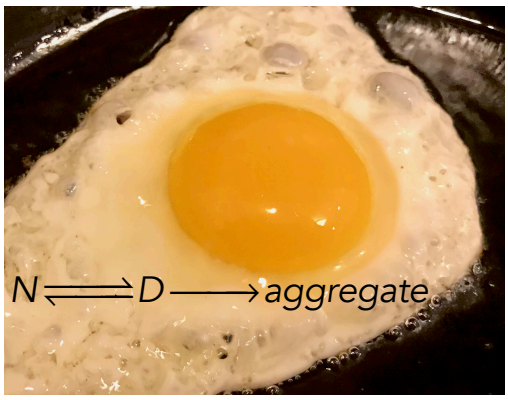


Figure 5.19. Cooking an egg involves heat-induced unfolding of the egg white proteins, followed by their aggregation into an insoluble gel. Egg white contains 10 wt% protein in water; this high concentration of protein promotes aggregation after unfolding.

activity at room temperature, incubates the sample at an elevated temperature, then cools it to room temperature and measures enzyme activity again. The decrease in activity reveals the fraction of enzyme irreversibly unfolded due to heating. The values reported are typically the half-life at a specific temperature. For non-catalytic proteins, the change in intrinsic fluorescence can measure the amount of natively folded protein remaining. The activity loss depends on both how much of the protein unfolds upon heating (its inherent stability) and on its ability to refold upon cooling (which may be prevented by aggregation), eq. 3.x above.

To convert the measured loss of activity to half-life, researchers assume the inactivation follows first-order kinetics, eq. Equation 5.15, where A is the activity at time t, A_0 is the initial enzyme activity and k is the inactivation rate constant. The natural logarithm of enzyme activity ($\ln A$) decreases linearly with time. A plot of several measurements of the natural logarithm of the remaining activity at different times yields a straight line, with a slope of $-k$ and a y-intercept of $\ln A_0$.

$$\ln A = -k \cdot t + \ln A_0 \tag{5.15}$$

By measuring the inactivation rate constants for both the wild-type enzyme and the engineered variant, eq. Equation 5.16, yields the change in Gibbs energy of activation, $\Delta\Delta G^{\ddagger}$, for the rate of enzyme inactivation. If the mutant is more stable, then $\frac{k_{mut}}{k_{wt}}$ is less than one, so the value of $\Delta\Delta G^{\ddagger}$ is positive, indicating an increase in the activation energy to unfold the protein.

$$\Delta\Delta G^{\ddagger} = -RT \ln \left[\frac{k_{mut}}{k_{wt}} \right] \tag{5.16}$$

In most cases, the rate determining step in enzyme inactivation is the unimolecular unfolding step, so the assumption of first-order kinetics is justified. However, refolding of the enzyme may also depend on aggregation of the protein into an insoluble precipitate, so the rate determining step may involve several enzyme molecules. In these cases, the inactivation will not fit eq. Equation 5.15, which assumes first order kinetics. Another way to measure irreversible denaturation by heat is to measure an apparent melting temperature, $T_{m,app}$. It is not a true melting temperature, only an apparent one, because it is an irreversible process.

Two strategies to reduce aggregation are 1) to reduce unfolding (K_{eq} in eq. Equation 5.14) or 2) slow down aggregation (k_{agg} in eq. Equation 5.14). Minimizing unfolding of the protein will use strategies above in Section Section 5.4. For example, a combination of three substitutions designed to stabilize the native cytosine deaminase slowed its aggregation 30-fold.[117]

Since aggregation occurs mainly through hydrophobic interactions, one method to slow aggregation is to minimize the hydrophobicity of solvent-exposed regions. To reduce the aggregation of antibodies, Chennamsetty

and coworkers[125] first identified hydrophobic areas on the surface including hydrophobic patches that become exposed as the protein moves. Replacing hydrophobic residues with hydrophilic ones within these patches reduced aggregation. Several web tools predict substitutions needed to reduce protein aggregation: Aggrescan3D[126] and Solubis.[127] After predicting the aggregation-prone regions (hydrophobic, β-sheet forming stretches) the web tools predict substitutions to 1) reduce their unfolding and exposure to solvent by strengthening interactions between the aggregation-prone area and the rest of the protein or 2) eliminate the aggregation propensity of the region. These substitutions are charged residues (R, K, D, or E) which reduce hydrophobic interactions needed for aggregation or proline, which hinders the formation of β-sheet structures in aggregates. The tools use FoldX to choose the substitution that is expected to best stabilize the folded form.

Hindering aggregation can even overcome a decrease in stability. Substitution A281E in Candida antarctica lipase B decreased its melting temperature from 58 to 51 °C indicating a decline in stability. However, this substitution increased the half-life of this enzyme at 70 °C 22-fold demonstrating a reduced propensity to aggregate.[128] This substitution makes a hydrophobic region of the enzyme less hydrophobic and less likely to aggregate.

A more drastic approach to reducing the association between protein molecules is supercharging. Supercharging is extensive substitutions on the protein surface to increase the net charge as high as +48, so the protein molecules repel one another and cannot aggregate.[129] Introducing large numbers of charge residues on the protein surface also reduce its hydrophobicity as charged amino acids replace uncharged amino acids. For example, green fluorescent protein (GFP, net charge = -7) loses its fluorescence at 100 °C due to unfolding, Equation 5.17. Cooling to room temperature does not restore fluorescence because the unfolded protein has aggregated and precipitated. In contrast, green fluorescent protein variants engineered to have a +36 or -30 net charge also lost their fluorescence upon heating to 100 °C due to unfolding, but upon

cooling, they did not aggregate and regained 62 and 28%, respectively, of their original fluorescence. This supercharging did not make the GFP variants more stable; in contrast, the GFP variants unfolded more readily than wild-type in urea, but supercharging did reduce their propensity to aggregate in the unfolded state.

$$\begin{aligned} GFP^{7-}_{folded} &\overset{\text{heat}}{\rightleftharpoons} GFP^{7-}_{unfolded} \xrightarrow{k_{agg}} \text{aggregate} \\ GFP^{30-}_{folded} &\overset{\text{heat}}{\rightleftharpoons} GFP^{30-}_{unfolded} \end{aligned} \tag{5.17}$$

Two disadvantages of supercharging are the potential to destabilize the protein and to disrupt its function. Residues with the same charge create unfavorable electrostatic interactions, which destabilize proteins. A change in electrostatic environment can also shift the pK_a of residues in the active site, which may affect binding or catalysis. For example, catalysis (k_{cat}) by a supercharged glutathione-S-transferase (net charge -40 for the dimer) was three-fold slower than wild-type (net charge +2 for the dimer).

5.6.2 Chemical modification

The side chains of three amino acids readily undergo spontaneous chemical modification: 1) asparagine deamidate to aspartate, 2) methionine oxidizes to a sulfoxide and 3) cysteine oxidizes to disulfide protein oligomers as well as sulfur oxides such as sulfenic acids (RS-OH). Replacement of problematic residues with a non-reactive residue eliminates the possibility of modification, but replacement of every asparagine, methionine, and cysteine in a protein is rarely needed. Many of the residues may not undergo modification, and even when they do, some modified residues will have little effect on protein properties. Spontaneous deamidation of asparagine to aspartate is a common chemical modification of proteins.

The most reactive are Asn-Gly sequences in sterically unhindered regions which have half-lives as short as 6 d in physiological conditions.[130] The least reactive asparagines are 10^5-fold more stable. Deamidation converts the neutral Asn residue to a negatively charged Asp residue. This change can impair function, destabilize the protein or have no effect. The web tool at www.deamidation.org predicts asparagine deamidation rates based on the 3-D structures of proteins. Glutamine can also deaminate to glutamate, but the rate is about a thousand-fold slower, except for the case of an N-terminal Gln, which deaminates at a rate similar to Asn. Oxidation of sulfur atoms in methionine and cysteine alters a proteins properties and may destabilize or inactivate it. The first protein engineering of an industrial enzyme was the removal of an oxidation sensitive methionine from the detergent protease subtilisin.[3] The existing subtilisin could tolerate most of the harsh conditions of laundry (heat, high pH, and detergent), but it could not tolerate bleach. Bleach oxidized a methionine near the active site to the methionine sulfoxide (R-S(O)-CH3), which hindered binding of the substrate proteins to inactivate the protease. Replacement of this problematic methionine with alanine created a bleach-tolerant protease. A cysteine-to-serine replacement stabilizes several cytokine drugs. This change does not affect biological activity, but avoids oligomerization by the formation of non-native intermolecular disulfide links between proteins. The specific activity of interferon-β, when expressed in *E. coli*, was about 10-fold lower than the native protein. The researchers hypothesized that oligomerization through intermolecular disulfide bonds caused the lower activity.[131] Replacement of Cys17 with Ser eliminated oligomerization and restored the specific activity to that of the native protein. The commercial drug, Betaseron®, contains this substitution. A similar Cys125Ser substitution stabilizes human interleukin 2, marketed as Proleukin®.

Some applications require proteins to work under conditions where other chemical modifications are possible. For example, glucose, an aldehyde, can react with lysine to form an imine. The enzymatic conversion of corn syrup (glucose) to high fructose corn syrup to increase its sweetness re-

quires the enzyme xylose isomerase to tolerate high concentrations of glucose. Stabilization involved replacing a surface lysine that reacted with the glucose.[26] At very high or low pH, irreversible chemical reactions can degrade the protein. For example, base-catalyzed β-elimination at pH > 8 alters cystine (disulfide) residues' side chains. Peptide backbone links next to aspartic acid residues can hydrolyze at pH < 4.

5.6.3 Serum half-life

Extending the serum half-life of a therapeutic protein enhances its efficacy because it remains active for a longer time. It also lowers cost because less therapeutic protein is required and improves delivery because injections are less frequent.[132,133] Three biochemical processes limit the serum half lives of proteins: (i) degradation by proteolytic enzymes, (ii) rapid filtration by the kidneys and (iii) receptor-mediated endocytosis. Small proteins (<50 kDa) typically have short serum half lives (5–50 min).

To stabilize proteins to proteolysis researchers identify the cleavage sites and modify them to slow hydrolysis. For example, human glucagon-like peptide-1 (GLP-1) has a serum half-life of only 2-3 minutes, in part due to rapid proteolytic cleavage by dipeptidyl peptidase 4 between the 8Ala-9Glu linkage, Figure Figure 5.20a. To slow down proteolysis researchers replaced the alanine residue with glycine (removal of the methyl group at Cα) or with 2-aminoisobutyric acid, Aib (addition of another methyl group at Cα).[134] Researchers replaced the alanine, not the glutamate because proteases are more selective for the amino acid on the carbonyl side of the amide bond rather than the amino acid on the amine side of the amide bond. Other protease that occur in serum besides dipeptidyl peptidase 4 are neutral endopeptidase, plasma kallikrein, and plasmin.

To slow filtration by the kidneys, researchers increase the molecular weight of the therapeutic protein to above ~50 kDa. In the case of GLP-1, researchers increased the molecular weight by adding an octadecanoic

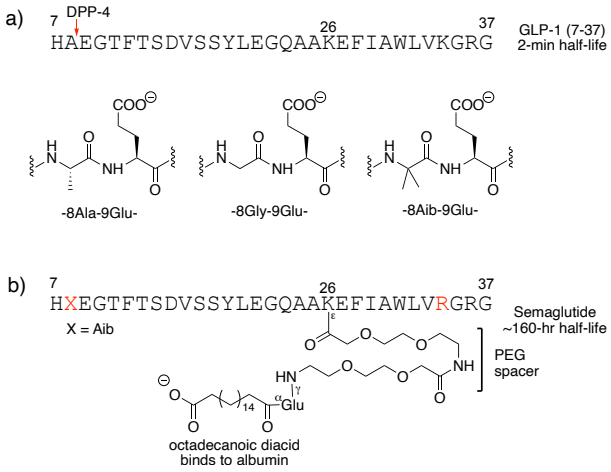


Figure 5.20. Structural modification of GLP-1 to increase its serum half-life. a) To prevent hydrolysis by the serum protease dipeptidyl peptidase 4, which favors hydrolysis at 8Ala-9Pro (marked by red arrow), researchers replaced 8Ala with Gly or the unnatural amino acid Aib. b) To prevent filtration by kidneys and to miminize receptor-mediated endocytosis, researchers attached a hydrophobic C18 group that binds to serum albumin. The red letters indicate changes in the peptide sequence from GLP-1.

diacid group, which binds to serum albumin, which has a molecular weight of 66.5 kDa, Figure Figure 5.20b above. The engineering of additional glycosylation sites or the chemical addition of poly(ethylene glycol) chain are other ways to increase the molecular weight of a therapeutic protein.

The choice of endogenous albumin as a binding protein also minimizes the third process that lowers serum half-life, receptor-mediated endocytosis. Endothelial cells lining the bloodstream internalize proteins via receptor-mediated endocytosis, which transfers the proteins from the blood to adjacent cells. However, some proteins, such as serum albumin and immunoglobulin G (IgG) are recycled back to the bloodstream. Upon endocytosis proteins transfer to the endosome, then to the lysosome for degradation. However, the endosome is acidic and contains FcRn receptor proteins, which tightly bind albumin and the Fc region of IgG proteins. Instead of degradation, this complex moves back to the cell surface where the higher pH of the blood weakens the binding between albumin or IgG and FcRn, causing release of the albumin or IgG back to the bloodstream. This recycling mechanism accounts for the long half-life of albumin and IgG in the blood. The modified form of GLP-1 in Figure Figure 5.20b above has a half-life of ~160 hours (3,200-fold longer than GLP-1) is an treatment for type 2 diabetes (Ozempic®) and obesity (Wegovy®).

Etanercept is another example where receptor-mediated recycling extends the serum half-life, but in this case using the Fc fragment of IgG, Figure Figure 5.21.[135]. The active protein is the extracellular domain of p75 tumor necrosis factor receptor, which binds the tumor necrosis factor cytokine to reduce inflammation in rheumatoid arthritis. Two copies of the active protein increase its affinity 100-4000-fold over monomeric counterparts, likely by better mimicking the dimeric structure of the receptor. These two active proteins are fused to the Fc fragment of human IgG, which increases serum half-life by enabling recycling of filtered protein back to the bloodstream.

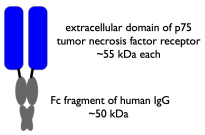


Figure 5.21. Etanercept is a fusion protein of two copies of the extracellular domain of tumor necrosis factor receptor (blue) with a single copy of the Fc fragment of IgG (gray). It is used to reduce inflammation of rheumatoid arthritis.

Another approach to increasing the serum half-life of therapeutic proteins is to create insoluble depot of the protein that slowly releases into the bloodstream. One example is the noncovalent entrapment of a GLP-1 homolog in poly(D, L-lactide-co-glycolide) to form 60 μm microspheres. Diffusion and hydrolysis of the polymer release the GLP-1 homolog with a half-life of 5-6 days. Another example is a long-acting insulin called insulin glargine. The protein sequence has been modified to alter the isoelectric point so that it precipitates as microcrystals at physiological pH and then slowly dissolves over twenty-four hours. The modified insulin is formulated in pH 4 solution where it is soluble, but then precipitates upon injection in the muscle.

5.7 Concluding remarks

Most single substitutions increase the stability of a protein by ≤1 kcal/mol (3–4 °C increase in melting temperature), so large stabilizations require multiple stabilizing mutations. In many cases, especially when the substitutions are far from one another, their stabilization effects will be approximately additive. For example, a heat-stabilized α-amylase contains at least ten amino acid substitutions, Table Table 5.4.[136] This enzyme catalyzes the hydrolysis of starch to glucose oligomers in the manufacture of corn syrup. High temperatures (90 °C) increase the solubility of

the starch and speed up hydrolysis, but require a heat-tolerant α-amylase. Stabilizing substitutions include replacement of amino acid residues that can undergo chemical modification, substitutions to stabilize the native form, substitutions to destabilize the unfolded form and substitutions discovered by random mutagenesis where the stabilization mechanism is unknown.

Table 5.4. Heat-stabilizing substitutions in α-amylase from *Bacillus licheniformis*.

Approach	Specifics
prevent chemical modification	remove oxidation (M197) and deamidation sites (N188, Q284)
stabilize N	bury Ca^{2+} ion (A181T), minimize electrostatic destabilization (H156Y)
destabilize D	introduce proline (R124P)
random mutagenesis	M15T, H133I, N199S, A209V

In another example, twelve substitutions in a halohydrin dehalogenase increased its apparent melting temperature by 28 °C and increased its ability to tolerate organic solvents.[118] The researchers first identified stabilizing single substitutions and then combined the best twelve into a single variant. The increased stability was mainly due to redistributed surface charges and improved interactions between subunits in this homotetrameric enzyme.

References

1. Morrison, C., & Lähteenmäki, R. (2018). Public biotech in 2017-the numbers. *Nat. Biotechnol.*, *36*(7), 576–584. https://doi.org/10.1038/nbt.4175
2. Brannigan, J. A., & Wilkinson, A. J. (2002). Protein engineering 20 years on. *Nat. Rev. Mol. Cell Biol.*, *3*(12), 964–970. https://doi.org/10.1038/nrm975
3. Estell, D. A., Graycar, T. P., & Wells, J. A. (1985). Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.*, *260*(11), 6518–6521. https://doi.org/10.1016/S0021-9258(18)88811-1
4. Pace, C. N., & Hermans, J. (1975). The stability of globular proteins. *CRC Crit. Rev. Biochem.*, *3*(1), 1–43. https://doi.org/10.3109/10409237509102551
5. Taverna, D. M., & Goldstein, R. A. (2002). Why are proteins marginally stable? *Proteins*, *46*(1), 105–109. https://doi.org/10.1002/prot.10016
6. Fersht, A. R., & Serrano, L. (1993). Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.*, *3*(1), 75–83. https://doi.org/10.1016/0959-440X(93)90205-Y
7. Bloom, J. D., Lu, Z., Chen, D., Raval, A., Venturelli, O. S., & Arnold, F. H. (2007). Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.*, *5*(1), 97–21. https://doi.org/10.1186/1741-7007-5-29
8. Gordon, S. R., Stanley, E. J., Wolf, S., Toland, A., Wu, S. J., Hadidi, D., Mills, J. H., Baker, D., Pultz, I. S., & Siegel, J. B. (2012). Computational design of an α-gliadin peptidase. *J. Am. Chem. Soc.*, *134*(50), 20513–20520. https://doi.org/10.1021/ja3094795

9. Harvey, S. B., Stone, M. D., Martinez, M. B., & Nelsestuen, G. L. (2003). Mutagenesis of the γ-carboxyglutamic acid domain of human factor VII to generate maximum enhancement of the membrane contact site. *J. Biol. Chem.*, *278*(10), 8363–8369. https://doi.org/10.1074/jbc.M211629200
10. Persson, E., Olsen, O. H., Bjørn, S. E., & Ezban, M. (2012). Vatreptacog alfa from conception to clinical proof of concept. *Semin. Thromb. Hemost.*, *38*(3), 274–281. https://doi.org/10.1055/s-0032-1302442
11. Bolli, G. B., Di Marchi, R. D., Park, G. D., Pramming, S., & Koivisto, V. A. (1999). Insulin analogues and their potential in the management of diabetes mellitus. *Diabetologia*, *42*(10), 1151–1167. https://doi.org/10.1007/s001250051286
12. Ryu, J. K., Kim, H. S., & Nam, D. H. (2012). Current status and perspectives of biopharmaceutical drugs. *Biotechnol. Bioprocess Eng.*, *17*(5), 900–911. https://doi.org/10.1007/s12257-012-0095-1
13. Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B. B., Henner, D., Wong, W. L. T., Rowland, A. M., Kotts, C., Carver, M. E., & Shepard, H. M. (1992). Humanization of an anti-p185^{HER2} antibody for human cancer therapy. *Proc. Natl. Acad. Sci. U. S. A.*, *89*(10), 4285–4289. https://doi.org/10.1073/pnas.89.10.4285
14. Sarkar, C. A., Lowenhaupt, K., Horan, T., Boone, T. C., Tidor, B., & Lauffenburger, D. A. (2002). Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nat. Biotechnol.*, *20*(9), 908–913. https://doi.org/10.1038/nbt725
15. Cantor, J. R., Yoo, T. H., Dixit, A., Iverson, B. L., Forsthuber, T. G., & Georgiou, G. (2011). Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(4), 1272–1277. https://doi.org/10.1073/pnas.1014739108

16. Loosmore, S. M., Zealey, G. R., Boux, H. A., Cockle, S. A., Radika, K., Fahim, R. E., Zobrist, G. J., Yacoob, R. K., Chong, P. C., & Yao, F. L. (1990). Engineering of genetically detoxified pertussis toxin analogs for development of a recombinant whooping cough vaccine. *Infect. Immun.*, *58*(11), 3653–3662. https://doi.org/10.1128/iai.58.11.3653-3662.1990
17. Hsieh, C.-L., Goldsmith, J. A., Schaub, J. M., DiVenere, A. M., Kuo, H.-C., Javanmardi, K., Le, K. C., Wrapp, D., Lee, A. G., Liu, Y., Chou, C.-W., Byrne, P. O., Hjorth, C. K., Johnson, N. V., Ludes-Meyers, J., Nguyen, A. W., Park, J., Wang, N., Amengor, D., … McLellan, J. S. (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science*, *369*(6510), 1501–1505. https://doi.org/10.1126/science.abd0826
18. Padgette, S. R., Kolacz, K. H., Delannay, X., Re, D. B., Tinius, C. N., Rhodes, W. K., Otero, Y. I., Barry, G. F., Eichholtz, D. A., Peschke, V. M., Nida, D. L., Taylor, N. B., & Kishore, G. M. (1995). Development, identification, and characterization of a glyphosate-tolerant soybean line. *Crop Sci.*, *35*(5), 1451–1461. https://doi.org/10.2135/cropsci1995.0011183x003500050032x
19. Castle, L. A., Siehl, D. L., Gorton, R., Patten, P. A., Chen, Y. H., Bertain, S., Cho, H.-J., Duck, N., Wong, J., Liu, D., & Lassner, M. W. (2004). Discovery and directed evolution of a glyphosate tolerance gene. *Science*, *304*(5674), 1151–1154. https://doi.org/10.1126/science.1096770
20. Bravo, A., & Soberon, M. (2008). How to cope with insect resistance to Bt toxins? *Trends Biotechnol.*, *26*(10), 573–579. https://doi.org/10.1016/j.tibtech.2008.06.005
21. Kurek, I., Chang, T. K., Bertain, S. M., Madrigal, A., Liu, L., Lassner, M. W., & Zhu, G. (2007). Enhanced thermostability of *Arabidopsis* rubisco activase improves photosynthesis and growth rates under moderate heat stress. *Plant Cell*, *19*(10), 3230–3241. https://doi.org/10.1105/tpc.107.054171

22. Whittle, E., & Shanklin, J. (2001). Engineering Δ^9-16:0-acyl carrier protein (ACP) desaturase specificity based on combinatorial saturation mutagenesis and logical redesign of the castor Δ^9-18:0-ACP desaturase. *J. Biol. Chem.*, *276*(24), 21500–21505. https://doi.org/10.1074/jbc.M102129200
23. Lin, H., & Lee, Y. K. (2017). Genetic engineering of medium-chain-length fatty acid synthesis in *Dunaliella tertiolecta* for improved biodiesel production. *J. Appl. Phycol.*, *29*(6), 2811–2819. https://doi.org/10.1007/s10811-017-1210-7
24. Yao, M. Z., Zhang, Y. H., Lu, W. L., Hu, M. Q., Wang, W., & Liang, A. H. (2012). Phytases: Crystal structures, protein engineering and potential biotechnological applications. *J. Appl. Microbiol.*, *112*(1), 1–14. https://doi.org/10.1111/j.1365-2672.2011.05181.x
25. Richardson, T. H., Tan, X., Frey, G., Callen, W., Cabell, M., Lam, D., Macomber, J., Short, J. M., Robertson, D. E., & Miller, C. (2002). A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable α-amylase. *J. Biol. Chem.*, *277*(29), 26501–26507. https://doi.org/10.1074/jbc.M203183200
26. Quax, W. J., Mrabet, N. T., Luiten, R. G. M., Schuurhuizen, P. W., Stanssens, P., & Lasters, I. (1991). Enhancing the thermostability of glucose-isomerase by protein engineering. *Bio/Technology*, *9*(8), 738–742. https://doi.org/10.1038/nbt0891-738
27. Bell, E. L., Finnigan, W., France, S. P., Green, A. P., Hayes, M. A., Hepworth, L. J., Lovelock, S. L., Niikura, H., Osuna, S., Romero, E., Ryan, K. S., Turner, N. J., & Flitsch, S. L. (2021). Biocatalysis. *Nat. Rev. Methods Primers*, *1*(1), 46. https://doi.org/10.1038/s43586-021-00044-z
28. van der Wielen, L. A. M., Ottens, M., & Straathof, A. J. J. (2001). Process Technology and Process Integration in the Preparation of Penicillins. In A. Bruggink (Ed.), *Synthesis of β-Lactam antibiotics: Chemistry, biocatalysis & process integration* (pp. 150–205). Springer Netherlands.

29. Ran, N., & Frost, J. W. (2007). Directed evolution of 2-keto-3-deoxy-6-phosphogalactonate aldolase to replace 3-deoxy-D-arabino-heptulosonic acid 7-phosphate synthase. *J. Am. Chem. Soc.*, *129*(19), 6130–6139. https://doi.org/10.1021/ja067330p
30. Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., ... Newman, J. D. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, *496*(7446), 528–532. https://doi.org/10.1038/nature12051
31. Anastas, P. T., & Warner, J. C. (2000). *Green chemistry : Theory and practice*. Oxford University Press.
32. Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J., Brands, J., Devine, P. N., Huisman, G. W., & Hughes, G. J. (2010). Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, *329*(5989), 305–309. https://doi.org/10.1126/science.1188934
33. Stone, E. M., Paley, O., Hu, J., Ekerdt, B., Cheung, N.-K., & Georgiou, G. (2012). De novo engineering of a human cystathionine-γ-lyase for systemic L-methionine depletion cancer therapy. *ACS Chem. Biol.*, *7*(11), 1822–1829. https://doi.org/10.1021/cb300335j
34. Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, *27*(10), 946–950. https://doi.org/10.1038/nbt.1568
35. Jobe, A., & Bourgeois, S. (1972). *lac* Repressor-operator interaction. VI. The natural inducer of the lac operon. *J. Mol. Biol.*, *69*(3), 397–408. https://doi.org/10.1016/0022-2836(72)90253-7
36. Egel, R. (1988). The '*lac*' operon - an irrelevant paradox? *Trends Genet.*, *4*(2), 31–31. https://doi.org/10.1016/0168-9525(88)90062-5

37. Valdez-Cruz, N. A., Caspeta, L., Pérez, N. O., Ramírez, O. T., & Trujillo-Roldán, M. A. (2010). Production of recombinant proteins in *E. Coli* by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microb Cell Fact*, *9*(1), 18. https://doi.org/10.1186/1475-2859-9-18
38. Austin, H. P., Allen, M. D., Donohoe, B. S., Rorrer, N. A., Kearns, F. L., Silveira, R. L., Pollard, B. C., Dominick, G., Duman, R., El Omari, K., Mykhaylyk, V., Wagner, A., Michener, W. E., Amore, A., Skaf, M. S., Crowley, M. F., Thorne, A. W., Johnson, C. W., Woodcock, H. L., ... Beckham, G. T. (2018). Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(19), E4350–E4357. https://doi.org/10.1073/pnas.1718804115
39. Hutchison 3rd, C. A., Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., & Smith, M. (1978). Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem.*, *253*(18), 6551–6560. https://doi.org/10.1016/S0021-9258(19)46967-6
40. Silva, D., Santos, G., Barroca, M., & Collins, T. (2017). Inverse PCR for point mutation introduction. In L. Domingues (Ed.), *PCR: Methods and Protocols* (pp. 87–100). Springer New York. https://doi.org/10.1007/978-1-4939-7060-5_5
41. Roy, S., & Caruthers, M. (2013). Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. *Molecules*, *18*(11), 14268–14284. https://doi.org/10.3390/molecules181114268
42. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S., & Ryadnov, M. G. (2023). DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.*, *7*(3), 144–161. https://doi.org/10.1038/s41570-022-00456-9
43. Seydel, C. (2023). DNA writing technologies moving toward synthetic genomes. *Nat Biotechnol.*, *41*(11), 1504–1509. https://doi.org/10.1038/s41587-023-02006-0

44. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison 3rd, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318
45. Green, M. R., & Sambrook, J. (2012). *Molecular Cloning: A Laboratory Manual* (4th ed.). Cold Spring Harbor Press.
46. Ball, P. (2018). A compassionate substance. How science and myth collide in water. *Lapham's Quart.*, *11*(Summer).
47. Pace, C. N., Scholtz, J. M., & Grimsley, G. R. (2014). Forces stabilizing proteins. *FEBS Lett.*, *588*(14), 2177–2184. https://doi.org/10.1016/j.febslet.2014.05.006
48. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427
49. 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., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
50. Mura, C., McCrimmon, C. M., Vertrees, J., & Sawaya, M. R. (2010). An introduction to biomolecular graphics. *PLoS Comput. Biol.*, *6*(8), e1000918. https://doi.org/10.1371/journal.pcbi.1000918
51. Leach, A. R. (2001). *Molecular modelling: Principles and applications*. Prentice Hall.

52. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz Jr, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, *117*(19), 5179–5197. https://doi.org/10.1021/ja00124a002
53. Allinger, N. L., Zhou, X., & Bergsma, J. (1994). Molecular mechanics parameters. *J. Mol. Struct. THEOCHEM*, *312*(1), 69–83. https://doi.org/10.1016/S0166-1280(09)80008-0
54. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem\@. Theory Comput.*, *13*(6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125
55. Homeyer, N., & Gohlke, H. (2012). Free energy calculations by the molecular mechanics Poisson-Boltzmann surface area method. *Mol. Inf.*, *31*(2), 114–122. https://doi.org/10.1002/minf.201100135
56. Bae, E., & Phillips Jr, G. N. (2005). Identifying and engineering ion pairs in adenylate kinases. *J. Biol. Chem.*, *280*(35), 30943–30948. https://doi.org/10.1074/jbc.M504216200
57. Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J., & Osuna, S. (2019). Hidden conformations in *Aspergillus niger* monoamine oxidase are key for catalytic efficiency. *Angew. Chem. Int. Ed.*, *58*(10), 3097–3101. https://doi.org/10.1002/anie.201812532
58. Thomas, L. L., Christakis, T. J., & Jorgensen, W. L. (2006). Conformation of alkanes in the gas phase and pure liquids. *J. Phys. Chem. B*, *110*(42), 21198–21204. https://doi.org/10.1021/jp064811m

59. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N., & Hilvert, D. (2013). Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.*, *9*(8), 494–498. https://doi.org/10.1038/nchembio.1276
60. Crean, R. M., Gardner, J. M., & Kamerlin, S. C. L. (2020). Harnessing conformational plasticity to generate designer enzymes. *J. Am. Chem. Soc.*, *142*(26), 11324–11342. https://doi.org/10.1021/jacs.0c04924
61. Robic, S., Guzman-Casado, M., Sanchez-Ruiz, J. M., & Marqusee, S. (2003). Role of residual structure in the unfolded state of a thermophilic protein. *Proc. Natl. Acad. Sci. U. S. A.*, *100*(20), 11345–11349. https://doi.org/10.1073/pnas.1635051100
62. Rosenthal, L., Rabolt, J. F., & Hummel, J. (1982). An investigation of the conformational equilibrium of *n*-butane in a solvent using Raman spectroscopy. *J. Chem. Phys.*, *76*(2), 817–820. https://doi.org/10.1063/1.443052
63. Vázquez-Figueroa, E., Yeh, V., Broering, J. M., Chaparro-Riggers, J. F., & Bommarius, A. S. (2008). Thermostable variants constructed via the structure-guided consensus method also show increased stability in salts solutions and homogeneous aqueous-organic media. *Protein Eng. Des. Sel.*, *21*(11), 673–680. https://doi.org/10.1093/protein/gzn048
64. Mayer, S., Rüdiger, S., Ang, H. C., Joerger, A. C., & Fersht, A. R. (2007). Correlation of levels of folded recombinant P53 in *Escherichia coli* with thermodynamic stability in vitro. *J. Mol. Biol.*, *372*(1), 268–276. https://doi.org/10.1016/j.jmb.2007.06.044
65. Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, *103*(15), 5869–5874. https://doi.org/10.1073/pnas.0510098103
66. Adams, M. W., & Kelly, R. M. (1998). Finding and using hyperthermophilic enzymes. *Trends Biotechnol.*, *16*(8), 329–332. https://doi.org/10.1016/s0167-7799(98)01193-7

67. Merz, A., Yee, M.-C., Szadkowski, H., Pappenberger, G., Crameri, A., Stemmer, W. P. C., Yanofsky, C., & Kirschner, K. (2000). Improving the catalytic activity of a thermophilic enzyme at low temperatures. *Biochemistry*, *39*(5), 880–889. https://doi.org/10.1007/pl00003701
68. Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.*, *17*(4), 263–270. https://doi.org/10.1002/humu.22
69. Joerger, A. C., Ang, H. C., & Fersht, A. R. (2006). Structural basis for understanding oncogenic P53 mutations and designing rescue drugs. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(41), 15056–15061. https://doi.org/10.1073/pnas.0607286103
70. Park, S. H., & Bao, G. (2021). CRISPR/Cas9 gene editing for curing sickle cell disease. *Transfusion and Apheresis Science*, *60*(1), 103060. https://doi.org/10.1016/j.transci.2021.103060
71. Pace, C. N., & Shaw, K. L. (2000). Linear extrapolation method of analyzing solvent denaturation curves. *Proteins Struct. Funct. Bioinforma.*, *Suppl 4*, 1–7. https://doi.org/10.1002/1097-0134(2000)41:4+%3C1::aid-prot10%3E3.3.co;2-u
72. Baxa, M. C., Haddadian, E. J., Jumper, J. M., Freed, K. F., & Sosnick, T. R. (2014). Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(43), 15396–15401. https://doi.org/10.1073/pnas.1407768111
73. Horovitz, A., & Fersht, A. R. (1992). Co-operative interactions during protein folding. *J. Mol. Biol.*, *224*(3), 733–740. https://doi.org/10.1016/0022-2836(92)90557-Z
74. Scholtz, J. M., Grimsley, G. R., & Pace, C. N. (2009). Solvent denaturation of proteins and interpretations of the m value. *Methods Enzymol.*, *466*, 549–565. https://doi.org/10.1016/S0076-6879(09)66023-7
75. Fersht, A. (1999). Chapter 17, Protein Stability. In *Structure and Mechanism in Protein Science*. Freeman.

76. Becktel, W. J., & Schellman, J. A. (1987). Protein stability curves. *Biopolymers*, *26*(11), 1859–1878. https://doi.org/10.1002/bip.360261104
77. Robertson, A. D., & Murphy, K. P. (1997). Protein structure and the energetics of protein stability. *Chem. Rev.*, *97*(5), 1251–1267.
78. Pucci, F., Kwasigroch, J. M., & Rooman, M. (2017). SCooP: An accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*, *33*(21), 3415–3422.
79. Pastore, A., Martin, S. R., Politou, A., Kondapalli, K. C., Stemmler, T., & Temussi, P. A. (2007). Unbiased cold denaturation: Low- and high-temperature unfolding of yeast frataxin under physiological conditions. *J. Am. Chem. Soc.*, *129*(17), 5374–5375. https://doi.org/10.1021/ja0714538
80. Razvi, A., & Scholtz, J. M. (2006). Lessons in stability from thermophilic proteins. *Protein Sci.*, *15*(7), 1569–1578. https://doi.org/10.1110/ps.062130306
81. Eijsink, V. G. H., Bjørk, A., G\a aseidnes, S., Sirev\a ag, R., Synstad, B., van den Burg, B., & Vriend, G. (2004). Rational engineering of enzyme stability. *J. Biotechnol.*, *113*(1-3), 105–120. https://doi.org/10.1016/j.jbiotec.2004.03.026
82. Dumon, C., Varvak, A., Wall, M. A., Flint, J. E., Lewis, R. J., Lakey, J. H., Morland, C., Luginbühl, P., Healey, S., Todaro, T., DeSantis, G., Sun, M., Parra-Gessert, L., Tan, X., Weiner, D. P., & Gilbert, H. J. (2008). Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure. *J. Biol. Chem.*, *283*(33), 22557–22564. https://doi.org/10.1074/jbc.M800936200
83. Shoichet, B. K., Baase, W. A., Kuroki, R., & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.*, *92*(2), 452–456. https://doi.org/10.1073/pnas.92.2.452

84. Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L., van Loon, A. P. G. M., & Wyss, M. (2002). The consensus concept for thermostability engineering of proteins: Further proof of concept. *Protein Eng.*, *15*(5), 403–411. https://doi.org/10.1093/protein/15.5.403
85. Magliery, T. J. (2015). Protein stability: Computation, sequence statistics, and new experimental methods. *Current Opinion in Structural Biology*, *33*, 161–168. https://doi.org/10.1016/j.sbi.2015.09.002
86. Jones, B. J., Kan, C. N. E., Luo, C., & Kazlauskas, R. J. (2020). Consensus Finder web tool to predict stabilizing substitutions in proteins. In D. S. Tawfik (Ed.), *Enzyme Engineering and Evolution: General Methods* (Vol. 643, pp. 129–148). Academic Press. https://doi.org/10.1016/bs.mie.2020.07.010
87. Musil, M., Stourac, J., Bendl, J., Brezovsky, J., Prokop, Z., Zendulka, J., Martinek, T., Bednar, D., & Damborsky, J. (2017). FireProt: Web server for automated design of thermostable proteins. *Nucleic Acids Res.*, *45*(W1), W393–W399. https://doi.org/10.1093/nar/gkx285
88. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L., & Bryan, P. N. (1989). Large increases in general stability for subtilisin BPN' through incremental changes in the free-energy of unfolding. *Biochemistry*, *28*(18), 7205–7213. https://doi.org/10.1021/bi00444a012
89. Schwehm, J. M., Fitch, C. A., Dang, B. N., García-Moreno E., B., & Stites, W. E. (2003). Changes in stability upon charge reversal and neutralization substitution in staphylococcal nuclease are dominated by favorable electrostatic effects. *Biochemistry*, *42*(4), 1118–1128. https://doi.org/10.1021/bi0266434
90. Trudeau, D. L., Kaltenbach, M., & Tawfik, D. S. (2016). On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.*, *33*(10), 2633–2641. https://doi.org/10.1093/molbev/msw138

91. Clarke, J., Henrick, K., & Fersht, A. R. (1995). Disulfide mutants of barnase I: Changes in stability and structure assessed by biophysical methods and x-ray crystallography. *J. Mol. Biol.*, *253*(3), 493–504. https://doi.org/10.1006/jmbi.1995.0568
92. Pace, C. N., Grimsley, G. R., Thomson, J. A., & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.*, *263*(24), 11820–11825.
93. Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.*, *151*(2), 261–287. https://doi.org/10.1016/0022-2836(81)90515-5
94. Perry, L. J., & Wetzel, R. (1984). Disulfide bond engineered into T4 lysozyme: Stabilization of the protein toward thermal inactivation. *Science*, *226*(4674), 555–557. https://www.jstor.org/stable/1693872
95. Matsumura, M., Becktel, W. J., Levitt, M., & Matthews, B. W. (1989). Stabilization of phage T4 Iysozyme by engineered disulfide bonds. *Proc. Natl. Acad. Sci. U. S. A.*, *86*, 6562–6566. https://doi.org/10.1073/pnas.86.17.6562
96. Roesler, K. R., & Rao, A. G. (2000). A single disulfide bond restores thermodynamic and proteolytic stability to an extensively mutated protein. *Protein Science*, *9*(9), 1642–1650. https://doi.org/10.1110/ps.9.9.1642
97. Kanaya, S., Katsuda, C., Kimura, S., Nakai, T., Kitakuni, E., Nakamura, H., Katayanagi, K., Morikawa, K., & Ikehara, M. (1991). Stabilization of *Escherichia coli* ribonuclease H by introduction of an artificial disulfide bond. *Journal of Biological Chemistry*, *266*(10), 6038–6044. https://doi.org/10.1016/S0021-9258(18)38080-3
98. Hazes, B., & Dijkstra, B. W. (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng. Des. Sel.*, *2*(2), 119–125. https://doi.org/10.1093/protein/2.2.119

99. Dombkowski, A. A. (2003). Disulfide by Design™: A computational method for the rational design of disulfide bonds in proteins. *Bioinformatics*, *19*(14), 1852–1853. https://doi.org/10.1093/bioinformatics/btg231
100. Suplatov, D., Timonina, D., Sharapova, Y., & Švedas, V. (2019). Yosshi: A web-server for disulfide engineering by bioinformatic analysis of diverse protein families. *Nucleic Acids Res.*, *47*(W1), W308–W314. https://doi.org/10.1093/nar/gkz385
101. Pikkemaat, M. G., Linssen, A. B. M., Berendsen, H. J. C., & Janssen, D. B. (2002). Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng.*, *15*(3), 185–192. https://doi.org/10.1093/protein/15.3.185
102. Stewart, E. J., Åslund, F., & Beckwith, J. (1998). Disulfide bond formation in the *Escherichia coli* cytoplasm: An *in Vivo* role reversal for the thioredoxins. *EMBO J.*, *17*(19), 5543–5550. https://doi.org/10.1093/emboj/17.19.5543
103. Zhou, H.-X. (2004). Loops, linkages, rings, catenanes, cages, and crowders: Entropy-based strategies for stabilizing proteins. *Acc. Chem. Res.*, *37*(2), 123–130. https://doi.org/10.1021/ar0302282
104. Watanabe, K., & Suzuki, Y. (1998). Protein thermostabilization by proline substitutions. *J. Mol. Catal. B Enzym.*, *4*(4), 167–180. https://doi.org/10.1016/s1381-1177(97)00031-3
105. Fu, H., Grimsley, G. R., Razvi, A., Scholtz, J. M., & Pace, C. N. (2009). Increasing protein stability by improving beta-turns. *Proteins*, *77*(3), 491–498. https://doi.org/10.1002/prot.22509
106. Matthews, B. W., Nicholson, H., & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U. S. A.*, *84*(19), 6663–6667. https://doi.org/10.1073/pnas.84.19.6663
107. Scott, K. A., Alonso, D. O. V., Sato, S., Fersht, A. R., & Daggett, V. (2007). Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(8), 2661–2666. https://doi.org/10.1073/pnas.0611182104

108. Liu, Z., Lemmonds, S., Huang, J., Tyagi, M., Hong, L., & Jain, N. (2018). Entropic contribution to enhanced thermal stability in the thermostable P450 CYP119. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(43), E10049–E10058. https://doi.org/10.1073/pnas.1807473115

109. 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*(46), 7745–7751. https://doi.org/10.1002/anie.200602795

110. Floor, R. J., Wijma, H. J., Colpa, D. I., Ramos-Silva, A., Jekel, P. A., Szymanski, W., Feringa, B. L., Marrink, S. J., & Janssen, D. B. (2014). Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem*, *15*(11), 1660–1672. https://doi.org/10.1002/cbic.201402128

111. Xiao, Z., Bergeron, H., Grosse, S., Beauchemin, M., Garron, M.-L., Shaya, D., Sulea, T., Cygler, M., & Lau, P. C. K. (2008). Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-temperature-guided sequence alignment. *Appl. Environ. Microbiol.*, *74*(4), 1183–1189. https://doi.org/10.1128/AEM.02220-07

112. Petsko, G. A. (2001). Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat." *Methods Enzymol.*, *334*, 469–478. https://doi.org/10.1016/S0076-6879(01)34486-5

113. Gribenko, A. V., Patel, M., Liu, J., McCallum, S., Wang, C., & Makhatadze, G. I. (2009). Rational stabilization of enzymes by computational redesign of surface charge–charge interactions. *Proc. Natl. Acad. Sci. U. S. A.*, *106*(8), 2601–2606. https://doi.org/10.1073/pnas.0808220106

114. Contessoto, V. G., de Oliveira, V. M., Fernandes, B. R., Slade, G. G., & Leite, V. B. P. (2018). TKSA-MC: A web server for rational mutation through the optimization of protein charge interactions. *Proteins*, *86*(11), 1184–1188. https://doi.org/10.1002/prot.25599

115. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987–2998. https://doi.org/10.1021/bi902153g
116. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Res.*, *33*(Web Server issue), W382–8. https://doi.org/10.1093/nar/gki387
117. Korkegian, A., Black, M. E., Baker, D., & Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. *Sci. Wash.*, *308*(5723), 857–860. https://doi.org/10.1126/science.1107387
118. Arabnejad, H., Dal Lago, M., Jekel, P. A., Floor, R. J., Thunnissen, A.-M. W. H., Terwisscha van Scheltinga, A. C., Wijma, H. J., & Janssen, D. B. (2017). A robust cosolvent-compatible halohydrin dehalogenase by computational library design. *Protein Eng. Des. Sel.*, *30*(3), 173–187. https://doi.org/10.1093/protein/gzw068
119. Jin, Y., Johannissen, L. O., & Hay, S. (2021). Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins*, *89*(8), 915–921. https://doi.org/10.1002/prot.26068
120. Weinstein, J. J., Goldenzweig, A., Hoch, S., & Fleishman, S. J. (2021). PROSS 2: A new server for the design of stable and highly expressed protein variants. *Bioinformatics*, *37*(1), 123–125. https://doi.org/10.1093/bioinformatics/btaa1071
121. Gao, D., Narasimhan, D. L., Macdonald, J., Brim, R., Ko, M.-C., Landry, D. W., Woods, J. H., Sunahara, R. K., & Zhan, C.-G. (2009). Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, *75*(2), 318–323. https://doi.org/10.1124/mol.108.049486
122. van den Burg, B., Vriend, G., Veltman, O. R., & Eijsink, V. G. H. (1998). Engineering an enzyme to resist boiling. *Proc. Natl. Acad. Sci. U. S. A.*, *95*(5), 2056–2060. https://doi.org/10.1073/pnas.95.5.2056

123. Bjørk, A., Dalhus, B., Mantzilas, D., Sirev\a ag, R., & Eijsink, V. G. H. (2004). Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface. *J. Mol. Biol.*, *341*(5), 1215–1226. https://doi.org/10.1016/j.jmb.2004.06.079
124. Kabashima, T., Li, Y., Kanada, N., Ito, K., & Yoshimoto, T. (2001). Enhancement of the thermal stability of pyroglutamyl peptidase I by introduction of an intersubunit disulfide bond. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1547*(2), 214–220. https://doi.org/10.1016/s0167-4838(01)00185-6
125. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., & Trout, B. L. (2010). Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B*, *114*(19), 6614–6624. https://doi.org/10.1021/jp911706q
126. Kuriata, A., Iglesias, V., Pujols, J., Kurcinski, M., Kmiecik, S., & Ventura, S. (2019). Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. *Nucleic Acids Res.*, *47*(W1), W300–W307. https://doi.org/10.1093/nar/gkz321
127. Van Durme, J., De Baets, G., Van Der Kant, R., Ramakers, M., Ganesan, A., Wilkinson, H., Gallardo, R., Rousseau, F., & Schymkowitz, J. (2016). Solubis: A webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel.*, *29*(8), 285–289. https://doi.org/10.1093/protein/gzw019
128. Suen, W.-C., Zhang, N., Xiao, L., Madison, V., & Zaks, A. (2004). Improved activity and thermostability of *Candida antarctica* lipase B by DNA family shuffling. *Protein Eng. Des. Sel.*, *17*(2), 133–140. https://doi.org/10.1093/protein/gzh017
129. Lawrence, M. S., Phillips, K. J., & Liu, D. R. (2007). Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.*, *129*(33), 10110–10112. https://doi.org/10.1021/ja071641y
130. Robinson, N. E. (2002). Protein deamidation. *Proc. Natl. Acad. Sci. U. S. A.*, *99*(8), 5283–5288. https://doi.org/10.1073/pnas.082102799

131. Mark, D. F., Lu, S. D., Creasey, A. A., Yamamoto, R., & Lin, L. S. (1984). Site-specific mutagenesis of the human fibroblast interferon gene. *Proc. Natl. Acad. Sci. U. S. A.*, *81*(18), 5662–5666.
132. Kontermann, R. E. (2009). Strategies to extend plasma half-lives of recombinant antibodies. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *23*(2), 93–109. https://doi.org/10.2165/00063030-200923020-00003
133. Strohl, W. R. (2015). Fusion proteins for half-life extension of biologics as a strategy to make biobetters. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *29*(4), 215–239. https://doi.org/10.1007/s40259-015-0133-6
134. Cheang, J. Y., & Moyle, P. M. (2018). Glucagon-like peptide-1 (GLP-1)-based therapeutics: Current status and future opportunities beyond type 2 diabetes. *ChemMedChem*, *13*(7), 662–671. https://doi.org/10.1002/cmdc.201700781
135. Feldmann, M. (2002). Development of anti-TNF therapy for rheumatoid arthritis. *Nat. Rev. Immunol.*, *2*(5), 364–371. https://doi.org/10.1038/nri802
136. Nielsen, J. E., & Borchert, T. V. (2000). Protein engineering of bacterial α-amylases. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1543*(2), 253–274. https://doi.org/10.1016/s0167-4838(00)00240-5
137. Jones, B. J., Lim, H. Y., Huang, J., & Kazlauskas, R. J. (2017). Comparison of five protein engineering strategies for stabilizing an α/β-hydrolase. *Biochemistry*, *56*(50), 6521–6532. https://doi.org/10.1021/acs.biochem.7b00571
138. Thieker, D. F., Maguire, J. B., Kudlacek, S. T., Leaver-Fay, A., Lyskov, S., & Kuhlman, B. (2022). Stabilizing proteins, simplified: A Rosetta-based webtool for predicting favorable mutations. *Protein Sci.*, *31*(10). https://doi.org/10.1002/pro.4428
139. Banta, S., Dooley, K., & Shur, O. (2013). Replacing antibodies: Engineering new binding proteins. *Annu. Rev. Biomed. Eng.*, *15*(1), 93–113. https://doi.org/10.1146/annurev-bioeng-071812-152412

140. Rossi, A. M., & Taylor, C. W. (2011). Analysis of protein-ligand interactions by fluorescence polarization. *Nat Protoc*, *6*(3), 365–387. https://doi.org/10.1038/nprot.2011.305
141. Pollard, T. D. (2010). MBOC technical perspective: A guide to simple and informative binding assays. *Mol. Biol. Cell*, *21*(23), 4061–4067. https://doi.org/10.1091/mbc.E10-08-0683
142. Jeffrey, P. D., Strong, R. K., Sieker, L. C., Chang, C. Y. Y., Campbell, R. L., Petsko, G. A., Haber, E., Margolies, M. N., & Sheriff, S. (1993). 26-10 Fab-digoxin complex: Affinity and specificity due to surface complementarity. *Proc. Natl. Acad. Sci. U. S. A.*, *90*(21), 10310–10314. https://doi.org/10.1073/pnas.90.21.10310
143. Tian, W., Chen, C., Lei, X., Zhao, J., & Liang, J. (2018). CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.*, *46*(W1), W363–W367. https://doi.org/10.1093/nar/gky473
144. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, *255*(5041), 178–183. https://doi.org/10.1126/science.1553543
145. Jones, B. J., Evans, R. L., Mylrea, N. J., Chaudhury, D., Luo, C., Guan, B., Pierce, C. T., Gordon, W. R., Wilmot, C. M., & Kazlauskas, R. J. (2020). Larger active site in an ancestral hydroxynitrile lyase increases catalytically promiscuous esterase activity. *PLoS ONE*, *15*, e0235341. https://doi.org/10.1371/journal.pone.0235341
146. DeLorbe, J. E., Clements, J. H., Teresk, M. G., Benfield, A. P., Plake, H. R., Millspaugh, L. E., & Martin, S. F. (2009). Thermodynamic and structural effects of conformational constraints in protein-ligand interactions. Entropic paradoxy associated with ligand preorganization. *J. Am. Chem. Soc.*, *131*(46), 16758–16770. https://doi.org/10.1021/ja904698q

147. Barros, E. P., Schiffer, J. M., Vorobieva, A., Dou, J., Baker, D., & Amaro, R. E. (2019). Improving the efficiency of ligand-binding protein design with molecular dynamics simulations. *J. Chem. Theory Comput.*, *15*(10), 5703–5715. https://doi.org/10.1021/acs.jctc.9b00483
148. Thakkar, S., Nanaware-Kharade, N., Celikel, R., Peterson, E. C., & Varughese, K. I. (2014). Affinity improvement of a therapeutic antibody to methamphetamine and amphetamine through structure-based antibody engineering. *Sci. Rep.*, *4*, 3673. https://doi.org/10.1038/srep03673
149. Zhao, Q., Ahmed, M., Guo, H., Cheung, I. Y., & Cheung, N.-K. V. (2015). Alteration of electrostatic surface potential enhances affinity and tumor killing properties of anti-ganglioside GD2 monoclonal antibody hu3F8. *J. Biol. Chem.*, *290*(21), 13017–13027. https://doi.org/10.1074/jbc.M115.650903
150. Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, *25*(10), 1171–1176. https://doi.org/10.1038/nbt1336
151. Selzer, T., Albeck, S., & Schreiber, G. (2000). Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.*, *7*(7), 537–541. https://doi.org/10.1038/76744
152. Bugnon, M., Röhrig, U. F., Goullieux, M., Perez, M. A. S., Daina, A., Michielin, O., & Zoete, V. (2024). SwissDock 2024: Major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. *Nucleic Acids Res*, *52*(W1), W324–W332. https://doi.org/10.1093/nar/gkae300
153. Zoete, V., Schuepbach, T., Bovigny, C., Chaskar, P., Daina, A., Röhrig, U. F., & Michielin, O. (2016). Attracting cavities for docking. Replacing the rough energy landscape of the protein by a smooth attracting landscape. *J Comput Chem*, *37*(4), 437–447. https://doi.org/10.1002/jcc.24249

154. Röhrig, U. F., Goullieux, M., Bugnon, M., & Zoete, V. (2023). Attracting Cavities 2.0: Improving the flexibility and robustness for small-molecule docking. *J. Chem. Inf. Model.*, *63*(12), 3925–3940. https://doi.org/10.1021/acs.jcim.3c00054
155. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., & Baker, D. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, *501*(7466), 212–216. https://doi.org/10.1038/nature12443
156. Morin, A., Kaufmann, K. W., Fortenberry, C., Harp, J. M., Mizoue, L. S., & Meiler, J. (2011). Computational design of an endo-1,4-beta-xylanase ligand binding site. *Protein Eng. Des. Sel.*, *24*(6), 503–516. https://doi.org/10.1093/protein/gzr006

Supporting Information

SI 5.1. Derivation of equation relating spectral changes and equilibrium constant, eq. Equation 5.4.

Deriving the relationship between spectral changes and equilibrium constant is straightforward. The protein is either in the native state or in the denatured state, so the sum of the two fractions is one: $F_N + F_D = 1$. The observed fluorescence at each urea concentration, Y_{obs}, is the sum of the fluorescence contributions from the native state, $Y_N \cdot F_N$, where Y_N is the fluorescence intensity from the pure native state, and the denatured state, $Y_D \cdot F_D$, where Y_D is the fluorescence intensity from the pure denatured state.

$$Y_{obs} = Y_N \cdot F_N + Y_D \cdot F_D$$

Substituting $F_N = 1 - F_D$ yields:

$$F_D = (Y_{obs} - Y_N)/(Y_D - Y_N)$$

Similarly substituting $F_D = 1 - F_N$ yields:

$$F_N = (Y_D - Y_{obs})/(Y_D - Y_N)$$

Dividing these two equations yields the equation below, which is the same as eq. Equation 5.4.

$$K^{unfold} = \frac{F_N}{F_D} = \frac{Y_{obs} - Y_N}{Y_D - Y_{obs}}$$

SI 5.2. Example of linear extrapolation of protein unfolding data in urea Protein A was dissolved in solutions of different concentrations of urea and allowed to reach equilibrium. The fluorescence spectra of these solutions showed an increase in fluorescence at 2-5 M urea, Figure Figure 5.22. Using this fluorescence data, calculate the Gibbs energy of unfolding in pure water for protein A.

To solve this problem, first estimate the fluorescence of the native and denatured states. The fluorescence of the native state is the flat part of the curve before unfolding begins. The average of the values at 0 and 1 M urea yields Y_N = 71. The fluorescence of the denatured state is the flat part of the curve after unfolding is complete. The average of the values at 6-9 M urea yields Y_D = 155.

Next, estimate the equilibrium constant between the native and denatured forms at each urea concentration in the unfolding region. For this example: $K^{unfold} = (Y_{obs} - 71)/(155 - Y_{obs})$, so one can calculate K_{unfold} at different urea concentrations by substituting the experimental values of Y_{obs}. This procedure yields the values in Table Table 5.5. Convert the equilibrium constants to Gibbs energy using R = 1.987 cal/mol °K and T = 298 °K, which yields the ΔG^{unfold} values in the table.

[urea]	Y_{obs}
0	72
1	70
2	81
3	95
4	140
5	150
6	155
7	152
8	156
9	156

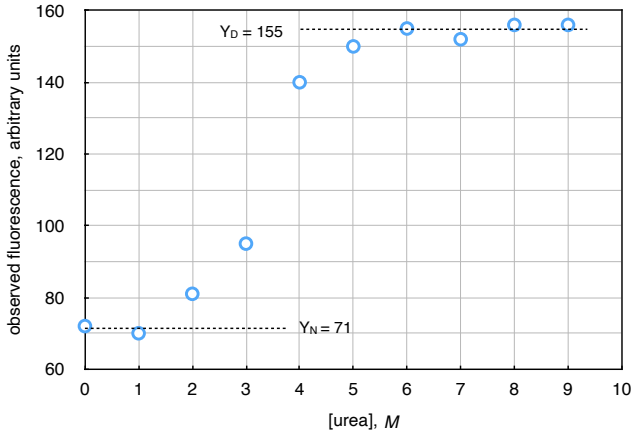


Figure 5.22. Typical experimental data from a urea-unfolding experiment to measure the Gibbs energy of unfolding of a protein. Depending on the protein and the wavelength used to monitor the fluorescence, the intensity of fluorescence can either increase or decrease upon protein unfolding; in this example, it increases.

Table 5.5. Calculated values of K_{unfold} and ΔG^{unfold}

[urea] (M)	K^{unfold}	ΔG^{unfold} (kcal/mol)
2.0	0.14	1.2
3.0	0.40	0.54
4.0	4.6	-0.90
5.0	16	-1.6

Finally, plot the ΔG^{unfold} values as a function of urea concentration, fit the data to a straight line and extrapolate to pure water ($[\text{urea}] = 0$), Figure Figure 5.23.

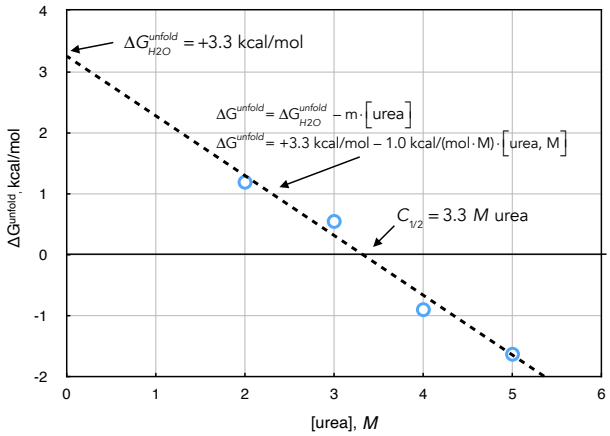


Figure 5.23. Linear extrapolation plot of the data from Table Table 5.5 above. The best fit line to the data is $\Delta G^{unfold} = 3.3 - 1.0 \cdot [\text{urea}]$, which indicates that $\Delta G^{unfold}_{H_2O}$ is 3.3 kcal/mol.

The extrapolation yields a y-intercept of +3.3 kcal/mol, which is the Gibbs energy of unfolding of this protein in pure water. This protein is less sta-

ble than typical proteins, but even for this one, the equilibrium amount of denatured protein in pure water is only 0.38%, which would be difficult to measure directly.

Problems

5.1. The following are examples of interactions that can stabilize proteins: a) improved core packing, b) higher backbone rigidity, c) increased surface polarity, d) increased salt bridges, e) increased hydrogen bonds between amino acids, f) replace residues with proline, g) increase number of charged amino acids. For each interaction, account for the stabilizing effect by describing the effect that it has on the folded protein and on the unfolded protein.

5.2. a) Floor and coworkers* used the 'Dynamic Disulfide Discovery algorithm', which they developed previously,† to predict disulfide cross link locations. That algorithm yielded seven residues pairs. (You do not need to look up the details of this algorithm.) The Dynamic Disulfide Discovery algorithm excludes residue pairs that are within 15 positions of each other. Explain why this is a good idea. If you apply this criterion to the results of SSBOND, how many residues pairs are still suitable for engineering?

b) Besides excluding residues pairs that are within 15 positions of each other, suggest another criterion that you could use to exclude a suggested pair. Provide a rationale for your criterion.

c) The apparent melting temperatures for the wild type and two of the variants are listed below. In which variant did the disulfide bond

*Floor, R. J., Wijma, H. J., Colpa, D. I., Ramos-Silva, A., Jekel, P. A., Szymanski, W. *et al.* (2014) Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem* 15, 1660-72.

†Wijma *et al.* (2014) Computationally designed libraries for rapid enzyme stabilization. *Prot. Eng. Des. Sel.* 27, 49–58; http://doi.org/10.1093/protein/gzt061

stabilize the protein? Explain how you came to your conclusion. For the variant that did not stabilize the protein, suggest a reason why it did not work based on the data in the table. Explain your reasoning. *Hint*: Consider the structure of amino acids R and D.

Substitutions	T_m(oxidized)	T_m(reduced)
none (wt)	51 °C	49 °C
A5C/A185C	56 °C	49 °C
R20C/D70C	50 °C	45.5 °C

d) Compare a prediction to a measured value. For the prediction, use equations 5.10 and 5.11 in the text to predict the expected Gibbs energy of stabilization for the two variants in part d above. For the experimental values, estimate the measured Gibbs energy of stabilization from the change in T_m using equation 5.8 (page 104) of the text. Comment on the match or mismatch between predictions and measured values. For the temperature, use the melting temperature for the variant that you are comparing.

5.3. *Predicting stabilizing single substitutions with Rosetta.* Jones and coworkers identified several substitutions that stabilize the esterase SABP2.[137] Table Table 5.7 lists four of the best substitutions. The increase in thermostability was measured two ways. The first method measured the time for irreversible inactivation at 60 °C. The sample was incubated at 60 °C for 15 min, then cooled to room temperature and the remaining catalytic activity was measured. The loss in catalytic activity was assumed to be first order in enzyme concentration and the expected half-life was calculated. Wild-type SABP2 had a half-life of 3.5 min. The single substitution variants in Table Table 5.7 all showed a longer half-life. The second two variants showed a larger increase in the half-life than the first two variants. The second measure of thermostability involved unfolding of the protein in urea and compared

the concentration of urea where the protein was half unfolded, $[\text{urea}]_{1/2}$. Wild-type SABP2 was half-unfolded at 2.23 M urea. The variants unfolded at higher concentrations of urea. The increase in urea concentration needed to half-unfold the protein was higher for the first two variants than for the second two.

Table 5.7. Two different measures of which single substitutions increase the thermostability of esterase SABP2.

Substitution	Increase in half-life at 60 °C	Increase in $[\text{urea}]_{1/2}$
L65F	1.2-fold	+1.5 M
E215P	1.5-fold	+1.0 M
Q221M	6.6-fold	+0.4 M
A230V	3.9-fold	+0.4 M

This question tests how well Rosetta predicts these four stabilizing substitutions. A webtool for Rosetta is available at ROSIE2 (Rosetta Online Server that Includes Everyone). In particular, we will use the tool stabilize-PM (point mutations), which predicts stabilizing single substitutions.[138] Using the tool is simple; one specifies the pdb code for the protein (1y7i for SABP2) and which positions you would like to mutate (A[65,215,221,230]), which means the four positions listed in chain A). By default, Rosetta tests 19 amino acids (all except cysteine) as replacements. During the calculations, Rosetta allows the two amino acids on either side of the mutated postion to move freely, all amino acids within 10 Å of the mutated position to move under weak constraints and all remaining amino acids to move only slightly (strong constraints). The results of the calculations (heat map in Figure Figure 5.24) show only one predicted stabilizing substitution, Leu65Phe, which is one of the experimentally identified stabilizing substitutions.

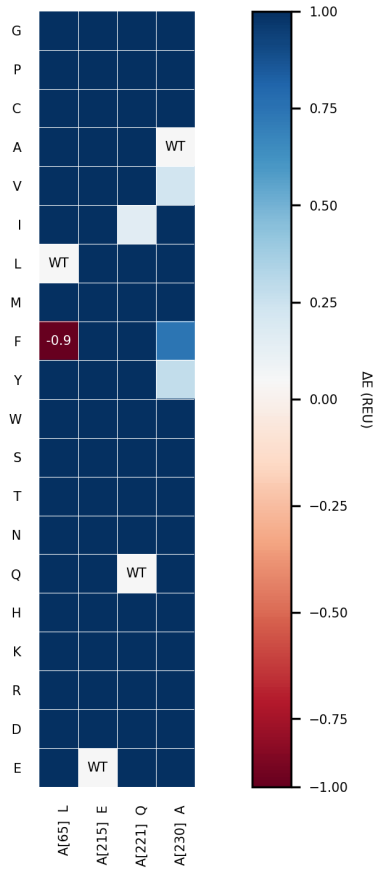


Figure 5.24. Heatmap shows the difference in score between the wild-type residue and each substitution (y-axis) for a given position (x-axis). Numerical values for the stabilizing scores (negative values) are shown, but only varying shades of blue are shown for the destabilizing scores (positive values). Scores greater than +1 are assigned the same navy color. This calculation predicted only one stabilizing substitution: Leu65Phe.

a) Thermostability was measured two different ways, which gave differing results about which substitution was the most stabilizing. Which of these experimental values should be compared to the Rosetta calculations? Explain your choice.
b) Rosetta failed to predict the Glu215Pro stabilizing substitution. Explain what was omitted from the Rosetta calculation that could have caused this error.

You may explore the Rosetta web tools for your own protein engineering project. The webtool requires login with a Github account, but is otherwise simple to use.

Answers

5.2. a) The stabilization increases with the size of the ring formed. If the residues are within 15 positions, the expected stabilization is only 3.1 kcal/mol (using eq 3.13 in the text, n = 16, T = 300 °K) and the actual stabilization is often less, so focusing on the larger rings increases the likelihood of success.

b) Avoid pairs near the active site. Avoid pairs like 2 where you replace an ion pair (Arg, Asp) which stabilizes the folded state. The replacement would destabilize both the folded and unfolded states leading to no net stabilization.

c) The A5C/A185C cross link stabilizes the protein because the melting temperature of the oxidized form increases by 5 °C as compared to wt. The R20C/D70C cross link destabilizes the protein because the melting temperature of the oxidized form decreases by 1 °C as compared to wt. Replacing the R/D pair with a cystine cross link removes an ion pair from the folded protein. The replacement destabilized the folded form (mp of reduced form is 3.5 °C lower than for wt).

d) A5C/A185C, n = 181, T = 327.5 °K, predicted $\Delta\Delta G = 5.8$ kcal/mol; observed 5/3.6 = 1.4 kcal/mol; also poor agreement. The cysteines seem to fit because the mp of the reduced form is unchanged from wt. The crosslink may not restrict the flexibility as much as expected.

R20C/D70C, n = 51, T = 323.5 °K, predicted $\Delta\Delta G = 4.5$ kcal/mol; observed -1/3.6 = -0.28 kcal/mol. The poor agreement may be due to removal of a ion pair from the folded form. Removal destabilizes folded form, but has no effect on the unfolded form because unfolding breaks the ion pair.

6 Engineering tighter binding

 Last revised: March 2026.

Summary. Tighter binding of proteins like antibodies to their target can increase the therapeutic effectiveness of biopharmaceuticals and lower the detection limits of diagnostic tests. Engineering a tighter-binding protein involves increasing the complementarity between the protein binding site and target. Both the shapes of the interacting surfaces and non-covalent interactions between them should match.

Key learning goals

- The relative stabilities of the bound state, where the protein and ligand interact with each other, and unbound state, where the protein and ligand interact with water, determine the strength of binding. Non-covalent interactions between the binding site and ligand favor the bound state, while solvation of the binding site and ligand favors the unbound state.
- Equilibrium dissociation constants, K_d, which have units of molarity, reveal the binding strength. K_d corresponds to the target concentration where half of the protein is bound to a ligand. Lower dissociation constants indicate tighter binding. Equilibrium dialysis is one technique to measure equilibrium dissociation constants.
- Increasing the affinity of a protein for a ligand requires increasing the shape and interaction complementarity between the two partners. Close contact between the surfaces creates non-covalent interactions including van der Waals interactions, hydrophobic inter-

actions, electrostatic interactions, and hydrogen bonds. The high cost of desolvating polar groups often makes burying them at the binding-site-ligand interface unfavorable even when they make favorable interactions like hydrogen bonds.

- Binding selectivity is the ability of a protein, often an antibody or receptor, to distinguish between two ligands. Binding selectivity depends on the ratio of the equilibrium association constants for the two binding possibilities.

I am always doing what I can't do yet in order to learn how to do it. — Vincent van Gogh

6.1 Introduction

Tighter binding of a protein to its ligand is a common goal of protein engineering.[139] A ligand is any species (usually a small organic molecule, an ion, a protein or an oligonucleotide) that binds selectively, stoichiometrically and reversibly to a protein. Increased binding of a diagnostic antibody to its antigen can lower detection limits (e.g., a pregnancy test could detect pregnancy earlier), while increased binding of a therapeutic antibody can decrease drug dosage or increase drug efficacy. Traditional methods to improve the binding of antibodies rely on the natural processes that generate antibodies. Mice are immunized with the antigen followed by screening for antibodies with higher affinity. These in vivo methods, called affinity maturation, reach a limit where affinity stops increasing due to biological limits of the innate immune system. A common protein engineering goal is improving the affinity of an antibody from good binding to great binding. Figure Figure 6.1 defines the different parts of an antibody.

The strength of binding a ligand to a protein depends on the equilibrium between the bound and unbound states, Figure Figure 6.2. In the bound

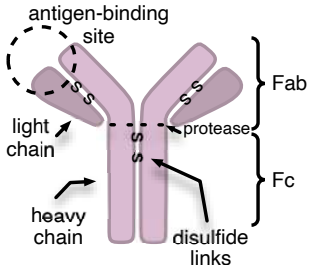


Figure 6.1. Antibodies consist of four protein chains: two identical heavy chains ($\sim$ 440 aa each) and two identical light chains ($\sim$ 220 aa each). Disulfide bonds connect the two heavy chains to each other and the light chains to their neighboring heavy chain. Antibodies contain two antigen-binding sites, each composed of the tips of the light and heavy chains. One of the two antigen binding sites is circled. Treatment with proteases cleaves both heavy chains as marked into three fragments: two Fab fragments (Fragment antigen-binding) and one Fc portion (Fragment crystallizable). The two protein chains that make up the Fab and Fc fragment remain linked to each other via disulfide cross links.

state the protein and ligand interact with each other, while in the unbound state they interact with solvent. Stronger non-covalent interactions with each other as compared to solvent favors binding.

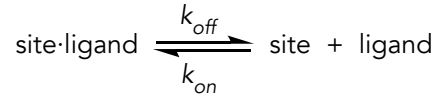


Figure 6.2. The bound state (ligand bound to the binding site) equilibrates with the unbound state (solvated ligand and site).

Two features influence the interactions between protein and ligand. First, the size and shape of the binding site in the protein must match the size and shape of the ligand. This match creates strong van der Waals interactions between them. Second, the chemical properties of regions within the binding site match those of ligand. Hydrophobic regions match and electrostatic interactions in polar regions are favorable. X-ray structures of antibody-antigen complexes reveal that 15-20 amino acids from the complementarity determining region of the antibody contact the antigen. Most of the contacts are hydrophobic and both the antibody and antigen adjust their conformations slightly upon complex formation to maximize the strength of the interactions. In the perfect case, no adjustments would be necessary since the most stable conformations of both protein and antibody match each another exactly.

The two comparison states that determine ligand-protein binding strength are the unbound state where water solvates the independently diffusing ligand and protein and the bound state where the ligand and protein interact with each other, Figure Figure 6.3. The solvation of polar and charged atoms at the surface stabilize the unbound state. In addition, the ability of the ligand and protein to diffuse independently in solution favors the unbound state. In contrast, the ordering of water at exposed hydrophobic regions of the protein and ligand destabilizes the unbound state as do unfavorable electrostatic interactions in high net charge regions.

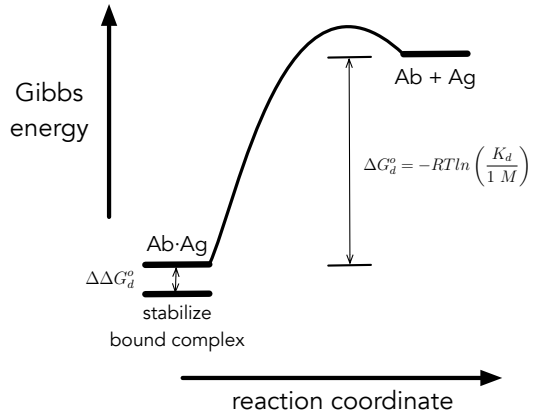


Figure 6.3. Gibbs energy changes associated with binding an antigen (Ag) to an antibody (Ab). The complex, Ab·Ag, is lower in energy than the free species, Ab + Ag, indicating that binding is favorable. Stabilizing the Ab·Ag complex increases the binding strength by an amount $\Delta\Delta G_d^o$.

In contrast, the bound state is stabilized by van der Waals attraction between touching atoms of the ligand and protein due to shape complementarity, the release of ordered water from hydrophobic regions of ligand and protein, electrostatic interactions between oppositely-charged atoms, and hydrogen bond formation between protein and ligand. The bound state is destabilized by desolvation of buried polar and charged atoms, electrostatic interactions between like-charged atoms, and the loss of independent diffusion for ligand.

Increasing binding requires increasing the energy difference between the bound and unbound states by stabilizing the bound state, destabilizing the unbound state or both. For example, one could destabilize the unbound state by adding exposed hydrophobic surface or high net charge regions within ligand-binding region. Ligand binding can alleviate this destabilization by binding a complementary hydrophobic surface or opposite charges. This alleviation of the destabilization when the ligand binds is critical to the success of this approach. If the destabilization persists in the bound state, then both states will have been destabilized and the binding strength will remain unchanged. Thus, the destabilizing changes must influence the ligand-binding region so that binding can alleviate the destabilization. One can stabilize the bound state relative to the unbound state by increasing the shape and charge complementarity between the protein and ligand stabilizes the bound state, while having little effect on the unbound state.

In all cases one must consider the effect of change to both the bound and unbound states. For example, engineering a hydrogen bond between the ligand and protein comes at the cost of reduced solvation of the polar atoms involved in the hydrogen bond. If the hydrogen bond is solvent accessible, then it remains solvated in the bound state and is often stabilizing. In contrast, buried hydrogen bonds are usually destabilizing because the cost of desolvating the polar atoms exceeds the gain from the hydrogen bond.

6.2 Measuring binding

Defining binding strength. The strength of binding is measured by the equilibrium constant for the reverse reaction, the ligand dissociation equilibrium constant, K_d, Figure Figure 6.2 and eq. Equation 6.1. For tighter binding, the dissociation equilibrium constant should be more unfavorable, so lower values of K_d indicate tighter binding. K_d has the units of molarity.

$$K_d = \frac{k_{off}}{k_{on}} = \frac{[site] \cdot [ligand]}{[site \cdot ligand]} \tag{6.1}$$

The value of K_d corresponds to the free ligand concentration when half of the sites contain a bound ligand and the other half are empty, eq Equation 6.2. Starting from eq. Equation 6.1, when concentration of empty sites, [site], equals the concentration of filled sites, [site·ligand], these two terms cancel in eq. Equation 6.1 yielding eq. Equation 6.2.

$$K_d = [ligand] \text{ when } [site] = [site \cdot ligand] \tag{6.2}$$

Antibody-antigen complexes are common cases of protein-ligand binding. For example, the mouse monoclonal antibody called 26-10 binds digoxin with high affinity. Digoxin, a hydrophobic steroid with an attached carbohydrate, is used to treat congestive heart failure. Anti-digoxin antibodies have been used to measure serum levels of digoxin and also to neutralize digoxin in cases of an overdose. Since antibody 26-10 contains two independently-acting binding sites for digoxin, the concentration of sites is twice the concentration of antibody. Antibody 26-10 binds digoxin very tightly ($K_d = 10^{-10}M$) and is half-saturated with digoxin when the concentration of free digoxin is $10^{-10}M$ (only 100 picomolar).

To convert the equilibrium constant in eq. Equation 6.1 into a Gibbs energy, one introduces the notion of a reference state. The purpose of the reference state is to cancel out the molarity unit on the equilibrium constant since one can only apply the logarithm function to dimensionless quantities. The standard reference state is 1 M and the value of K_d is divided by this reference yielding a dimensionless quantity, eq. Equation 6.3.

$$\Delta G_d^o = -RT \ln(K_d/1M) \tag{6.3}$$

The values of K_d for stable protein-ligand binding interactions are always less than 1 M, so the ratio $K_d/1M$ is always less than one. The logarithm of a value less than one is negative, so the Gibbs energy for dissociation, ΔG_d^o, eq. Equation 6.3, is positive indicating that dissociation is unfavorable. The 'o' superscript on the ΔG_d^o indicates that the K_d is compared to a reference state of 1 M. The free energy of dissociation of the digoxin-antibody 26-10 complex, ΔG_d^o, is +14.2 kcal/mol indicating that dissociation is highly unfavorable.

The Gibbs energy diagram shows that dissociation of an antibody-antigen complex is unfavorable, Figure Figure 6.3 above. To improve binding one seeks to make dissociation even more unfavorable. This approach is analogous to stabilizing a protein by makeing the unfavorable unfolding reaction even more unfavorable in the previous chapter.

Measuring K_d. One method to measure the strength of binding between a protein like an antibody, and a ligand (or antigen) is equilibrium dialysis. The antigen equilibrates between two compartments separated by a dialysis membrane, Figure Figure 6.4. The small pore diameter of the dialysis membrane allows the antigen to pass through, but not the antibody. At equilibrium, both compartments contain equal amounts of free antigen, but the antibody-containing compartment contains additional antigen as an antibody-antigen complex.

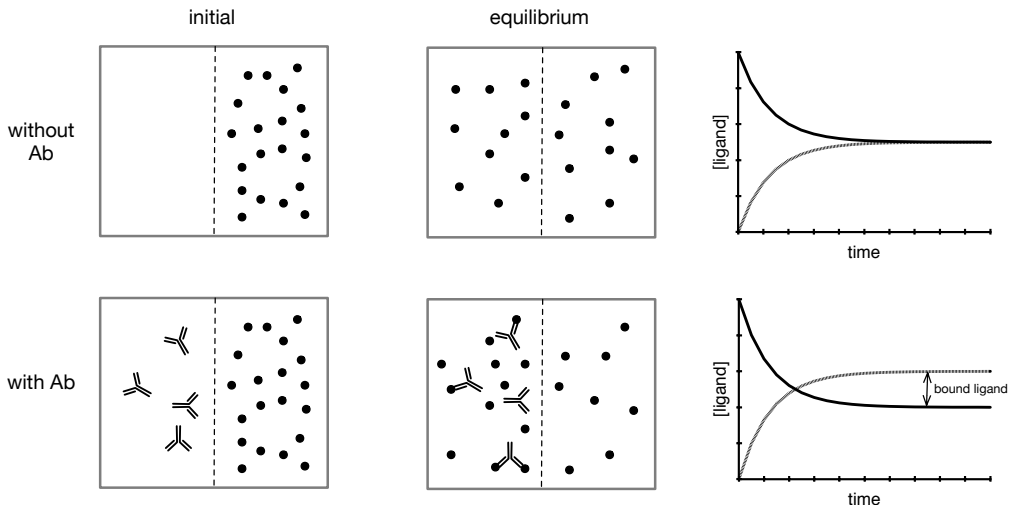


Figure 6.4. Measuring ligand (antigen) binding to an antibody by equilibrium dialysis. A dialysis membrane (dotted line) separates two compartments. Top row shows a control experiment without antibody. The ligand (twenty black circles) that is initially added to the right compartment distributes equally between both compartments (ten in each) because it passes freely through the membrane. Bottom row shows an experiment with antibody (Y-shapes) in left compartment. The dialysis membrane retains the antibody in its compartment. At equilibrium, both compartments contain the same concentration of free ligand (eight in each), but the antibody-containing compartment additionally contains antibody-bound ligand (four bound to the Y-shapes). The difference in the number of ligand molecules in the two compartments is the number of ligand molecules bound to antibody.

Measuring the equilibrium dissociation constant involves a series of equilibration experiments with different antigen concentrations. At lower concentrations of Ag, less Ab·Ag complex will form. The molar amount of antigen is always in excess of antibody. Since the equilibrium dissociation constants are very low, these experiments may require detecting very low antigen concentrations. For example, measuring the affinity of antibody 26-10 for digoxin requires measuring digoxin concentrations in the range of 100 picomolar.

Each equilibration experiment yields a ratio, Y, of sites complexed with ligand compared to the total number of sites. The concentration of sites complexed with ligand is the difference in the ligand concentrations between the antibody-containing and the antibody-free compartments, eq Equation 6.4. The total concentration of sites is known from the amount of antibody added to the experiment. For the example in Figure Figure 6.4 above, four of the binding sites are occupied by ligand and the total number of sites is eight, so Y = 4/8 = 0.5.

$$Y = \frac{\text{sites complexed with ligand}}{\text{total sites}} = \frac{[site \cdot ligand]}{[site] + [site \cdot ligand]} \tag{6.4}$$

Plotting the measured fraction saturation, Y at different concentrations of ligand as a function of free ligand on the x-axis yields a saturation binding curve, Figure Figure 6.5. The sites become saturated with ligand as the concentration of free ligand increases.

Equation Equation 6.5 defines the relationship between Y, the free ligand concentration and the equilibrium dissociation constant, K_d. At half-saturation, Y = 0.5, the free ligand concentration equals the equilibrium dissociation constant, 0.075 mM for this example.

$$Y = \frac{[ligand]}{K_d + [ligand]} \tag{6.5}$$

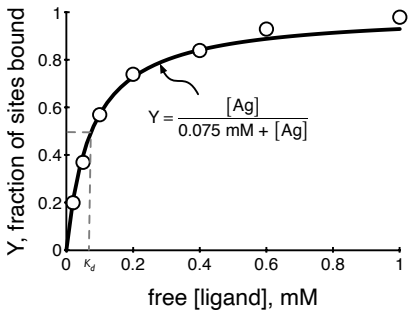


Figure 6.5. Fraction of antibody, Ab, bound to antigen, Ag, increases with increasing concentration of Ag until all of the antibody is saturated. Fitting equation Equation 6.5 to this data yields the equilibrium dissociation constant, K_d = 0.075 mM. Inspection of the data table reveals that a bound antigen fraction of 0.5 occurs between 0.05 mM and 0.10 mM of free antigen, so the value of K_d lies within this range.

This equation is derived by rearranging eq Equation 6.1 to define the concentration of free sites, eq Equation 6.6, and substituting it into eq Equation 6.4 above.

$$[site] = \frac{K_d \cdot [site \cdot ligand]}{[ligand]} \tag{6.6}$$

Finding K_d from the data in Figure Figure 6.5 involves adjusting the value of K_d so that the eq Equation 6.5 best fits the experimental data.*

Another way to measure the binding of a ligand to a protein is fluorescence polarization.[140] This method requires a fluorescent ligand, but

*Non-linear least squares fitting of eq. Equation 6.5 to the data starts with an initial guess for the dissociation constrant followed by iteration to find the best value. The supporting information includes a Python function for this fit.

does not require the two-compartment apparatus needed for equilibrium dialysis. The solution of ligand plus protein is irradiated with plane-polarized light to excite the fluorescent ligand. The emitted light (fluorescence) from the ligand is measured both parallel and perpendicular to the plane of the irradiated light. If the ligand does not turn during the few nanoseconds between the absorbtion and emission of light, then most of the light will be emitted parallel to the plane of the irradiated light. However, it the ligand turns during the few nanoseconds between the absorption and emission of light, then less of the light will be emitted parallel to the plane of the irradiating light. A ligand free in solution tumbles rapidly, but a ligand bound to a macromolecule tumbles more slowly. Fluoresence polarization detects whether a ligand is bound or not by measuring how much of the emitted light is parallel to the plane of irradiated light.

One can extend this method to also measure the binding of non-fluorescent ligands by adding the fluorescent ligand and a competing non-fluorescent ligand simultaneously. The non-fluorescent ligand competes with the fluorescent ligand for the protein binding site. Measuring the decrease in binding of the fluorescent ligand a varying concentrations of the non-fluorescent ligand reveals the relative strength of the two ligands. Since the binding strength of the fluorescent ligand is known, the binding strength of the non-fluorescent ligand can be calculated.

Practical details on binding measurements are available.[141] Other ways to measure binding include surface plasmon resonance and isothermal titration calorimetry.

Increasing binding strength requires increasing the difference in energy between the bound and unbound states. Engineering the antibody for tighter binding lowers the Gibbs energy of the Ag·Ab complex, raises the Gibbs energy of free Ag and Ab or both. The value of ΔG_d^o measures improvement of binding, Figure Figure 6.3. This value is the difference

in the Gibbs energy change for the two dissociation reactions, eq Equation 6.7.

$$\begin{aligned}\Delta\Delta G^o_d &= \Delta G^o_{d-new} - \Delta G^o_{d-original} \\ &= -RT \ln\left(\frac{K_{d-new}}{K_{d-original}}\right) = -RT \ln(improvement factor)\end{aligned} \tag{6.7}$$

Tighter binding corresponds to a larger positive Gibbs energy of dissociation for the improved variant and thus a positive value of ΔG^o_d. The value of K_{d-new} is a smaller (dissociation less favorable) than $K_{d-original}$, so $\left(\frac{K_{d-new}}{K_{d-original}}\right)$ is a value less than 1. For example, equilibrium dissociation constant may decrease from $10^{-6}M$ in the original to $10^{-8}M$ in the engineered case. The improvement factor is 10^{-2} and ln $\left(\frac{K_{d-new}}{K_{d-original}}\right)$ is –4.61. Then $-RT \ln\left(\frac{K_{d-new}}{K_{d-original}}\right)$ is +2.73 kcal/mol at 298 °K.

6.3 Engineering tighter binding

Engineering tighter binding involves modification of the antibody binding site to increase the complementarity between the antibody and antigen, Table Table 6.1. The first requirement of a binding site is that its size and shape must match the size and shape of the target ligand. This matching excludes solvent water by placing ligand atoms near binding-site atoms to create the opportunity for favorable interactions. The second criteria is that the binding site be preorganized for binding, that is, that the lowest energy conformation be the one that binds the ligand. The remaining criteria refer to optimizing different types of non-covalent interactions: hydrophobic interactions, hydrogen bonds and electrostatic interactions.

Table 6.1. Strategies to increase non-covalent interactions between protein and ligand.

Strategy	Reasoning
size and shape matching	creates favorable van der Waals interactionsminimizes bumpingreleases bound waterallows direct interactions
preorganization	binding conformation favored even before bindingno rearrangement needed for binding
hydrophobic matching	increases hydrophobic effect
optimize H bonds	compensates for loss of H bonds to solventavoids buried H bonds due to poor solvation
optimize electrostatics	minimizes interactions with like chargesmaximizes interactions with unlike charges

Size and shape matching. Matching the size and shape of the protein binding site to the target ligand maximizes their contact surface in order to maximize favorable non-covalent interactions, Figure Figure 6.6. Size matching ensures that all of the ligand interacts with the binding site. If the cavity is too large or too small, then part of the ligand remains exposed to solvent and the interaction between ligand and binding site will be less than the maximum possible. Shape matching refers to the surface contours of the ligand and binding site. Surface complementarity means the surfaces touch to create favorable van der Waals interactions. The close contact also makes additional non-covalent interactions between the ligand and binding site possible.

The intricate shapes of molecules prevents most of the atoms in two

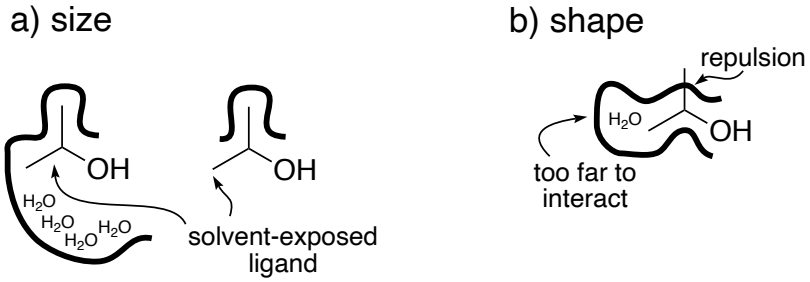


Figure 6.6. To maximize interactions with the target ligand (isopropyl side chain) the binding site must be the same size as the ligand and the contours of the binding site must match the shape of the ligand. a) A binding site that is larger or smaller than the ligand leaves part of the ligand exposed to solvent, which misses opportunities for interactions between ligand and binding site. b) Shape mismatches between the target ligand and contours of the binding site create repulsive interactions in regions where the binding site is too small and an inability to make favorable interactions in regions where the binding site is to large to contact the ligand.

contacting molecules to directly contact one another. The sphere representation of a molecule shows its van der Waals surface. The radius of each sphere is the van der Waals radius for the atom and the overlapping spheres show the van der Waals surface of the molecule, Fig Figure 6.7. This surface is intricate and includes narrow crevices too small to fit a water molecule. A water molecule cannot touch the regions within these crevices and a larger molecule will be excluded from even more regions. The *solvent contact surface* is a hypothetical smoother surface that eliminates crevices that are inaccessible to water. The solvent contact surface is the surface mapped by the edge of sphere with a radius of 1.4 Å as it rolls on the van der Waals surface of a molecule. The radius of 1.4 Å corresponds to the van der Waals radius of a water molecule. This surface shows smooth areas where the water molecule rides over the van der Waals surface as well as regions of protruding atoms where the water can directly touch the molecule.[†]

The majority of the surface atoms, even for highly complementary molecules, will not contact one another. The intricacy of molecular structures make complete direct contact geometrically impossible. Good complementary shapes will exclude solvent water from the interface and make many direct contacts.

An example of high surface complementarity is the interaction between antibody fragment 26-10 complexed with digoxin,[142] Figure Figure 6.8. The structure of 26-10 shows a deep pocket that surrounds the hydrophobic steroid portion. The hydrophilic carbohydrate groups of digoxin remain exposed to solvent and do not contribute to binding. The comple-

[†]Another surface, the solvent accessible surface, or Connolly surface, lies 1.4 Å outside the solvent contact surface. This surface represents the center, not the edge, of a water molecule on the surface. The solvent accessible surface is more convenient in some cases because positions on this surface will match the x,y,z coordinates of the rolling water molecule.

[‡]Charles Cullen, a nurse and serial killer convicted in 2003, used digitalis (and other drugs) to kill possibly forty hospital patients. This story and that of Amy Loughren, the ICU nurse who helped convict Cullen, is told in the book *The Good Nurse* (2013), and a 2022 movie by the same title currently on Netflix.

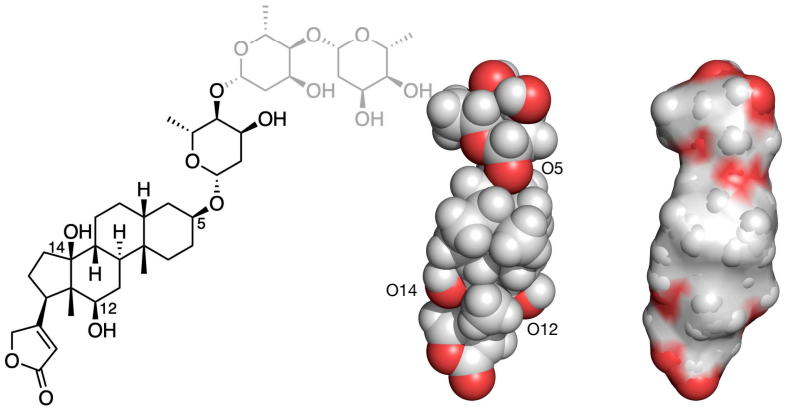


Figure 6.7. Structure of digoxin (digitalis), a cardiac glycoside.[‡] Left: The chemical structure of digoxin consists of a steroid (the aglycone) linked to a trisaccharide. The two sugar moieties in gray are not included in the two structures to the right. Middle: The overlapping van der Waals spheres of each atom create the van der Waals surface of digoxin. It contains numerous crevices that are too small to fit a water molecule. A water molecule would be slightly larger than one of the oxygen atoms shown as red spheres. Right: The solvent contact surface of digoxin is a hypothetical surface that shows the closest approach of a water molecule. Water can directly contact only a small fraction of the van der Waals surface.

mentarity between the solvent-accessible surface of the aglycone (compound without the glycoside moiety) of digoxin and 26-10 is imperfect, even in regions that are in contact. The shape complementary is close enough that there are no buried cavities in the complex that are large enough to contain a water molecule. This exclusion of water ensures a strong contribution of the hydrophobic effect to the binding. There are some regions exposed to the bulk water that are large enough to contain water molecules. The conformations of antibody and steroid remain the same in both unbound and bound states indicating that the low energy unbound shapes match each other.

Fragment 26-10 binds digoxin tightly, $K_d = 0.1$ nM. The binding is entirely due to the hydrophobic effect and van der Waals interactions. No hydrogen bonds or salt bridges form between antibody fragment 26-10 and digoxin. One consequence of the lack of interactions with the hydroxyl groups at C12 and C14 is a lack of selectivity for analogs without these hydroxyl groups. For example, 26-10 binds digitoxigenin (which lacks the hydroxyl group at C12) with equal affinity to digoxigenin.[§]

Sculpting a complementary surface will likely require multiple mutations and involve readjustment of the main chain and side chains. The smallest change that one can introduce by amino acid substitutions is the addition or removal of a methyl or hydroxyl group. This is a large change - about the size of a water molecule - and likely too large to precisely sculpt a surface. To make smaller changes one needs to make multiple substitutions including substitutions outside the surface being engineered so that a succession of adjusted positions make more subtle changes to the complementary surface. Precisely predicting the adjustments caused by multiple substitutions is more difficult making such sculpting difficult.

Substitutions often introduce readjustments of the main chain and side chains. A larger side chain causes readjustments when it bumps into nearby atoms. A smaller side chain may also cause readjustments because enlarging a hydrophobic pocket is unfavorable. Replacing a large

[§]The -genin suffix on these two compounds indicates the aglycone only.

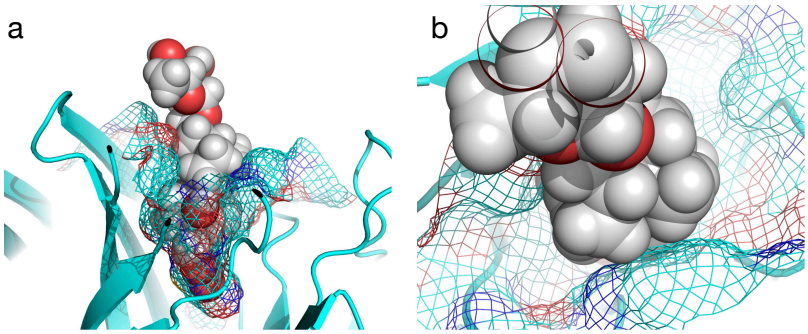


Figure 6.8. The binding pocket of antibody fragment 26-10 (mesh surface, backbone shown as a cartoon) binds the steroid digoxin (spheres) by creating a complementary surface. a) The two terminal saccharides extend into solvent and were not resolved in this structure. b) A top view shows close contact between the 26-10 and the aglycone in several parts, but much of the 26-10 surface does not directly contact the aglyclone. The figure was created using PyMOL from the x-ray crystal structure of the complex (pdb id = 1igj[142]). The surface of the antibody pocket was calculated using the web tool CASTp[143]

buried hydrophobic side chain with a smaller one destabilizes the protein because the smaller hydrophobic surface 1) reduces the contribution of the hydrophobic effect to protein folding and 2) reduces the favorable van der Waals contacts between the side chain and the rest of the protein. Structural readjustments often decrease the size of the created space to minimize these changes. The estimated decrease in stability due to removing one buried CH_2 group is 1.1 ± 0.5 kcal/mol.[47] For example, a Leu to Ala substitution removes three carbon atoms from the protein, so this substitution decreases protein stability by an estimated 3.3 ± 1.5 kcal/mol. If the protein structure readjusts, the destabilization will be at the smaller end of the estimate range, while if the structure does not readjust, it will be at the larger end of the estimated range. For example, a Leu to Ala mutation in T4 lysozyme destabilized the protein by 5.0 kcal/mol at position 99, but by only 2.7 kcal/mol at position 46. The x-ray structures revealed that at position 99 the structure did not readjust upon substitution leading to a larger cavity and a larger loss in van der Waals interactions, while at position 46 the structure relaxed leading to a smaller increase in the cavity size and a smaller loss in the van der Waals interactions.[144]

In a few cases, readjustments of the nearby residues reverse the intended change in size. For example, an ancestral hydroxynitrile lyase contains phenylalanines at positions 121 and 178 in the active site, while the corresponding modern enzyme from the rubber tree contains smaller leucines at these positions. One would expect that the substrate-binding site of the rubber tree enzyme would be larger, but the x-ray structures showed the opposite. In the rubber tree enzyme the side chain of Trp128 readjusts into the active site creating a smaller substrate binding site,[145] which is the opposite of what might be expected from comparison of the sizes of phenylalanine and leucine.

Preorganization. Preorganization of a dynamic protein means creating a global energy minimum conformation where the atoms in the apo protein are already oriented to interact with the partner atoms in the ligand. In the ideal case, this conformation has the correct structure with a minimum of flexibility. It must have the correct structure to avoid the energy

cost of moving the protein atoms from their most stable postions into the less stable interacting positions. This energy cost lowers the overall favorability of the binding. Flexibility of the apo protein also creates an energy cost. If protein flexibility decreases in the protein-ligand complex, then there is an entropy cost to binding, which also lowers the favorability of binding. Thus, the preorganized apo protein should be no more flexible than the protein-ligand structure.

Binding sites exist as an equilibrium mixture of non-binding and binding conformations, eqs. Equation 6.8 and Equation 6.9. The equilibrium constant (or the Gibbs energy) between the binding and not-binding conformations measures the propensity of the binding site to be preorganized for binding. It is convenient to express this equilibrium as the probability that the binding site is in a binding conformation, eq. Equation 6.10, which is a value between 0 and 1. For example, if the equilibrium constant is 10 in favor of the binding conformations, then the probability that the binding site exists in a binding conformation is 10/11 = 91%. This probability is sometimes called the Boltzmann factor or Boltzmann weight.

$$\text{non-binding conf.} \overset{K_{eq}}{\rightleftharpoons} \text{binding conf.} \tag{6.8}$$

$$K_{eq} = \frac{[\text{binding conf.}]}{[\text{non-binding conf.}]} \tag{6.9}$$

$$\text{probability} = \frac{[\text{binding conf.}]}{[\text{non-binding conf.}] + [\text{binding conf.}]} = \frac{K_{eq}}{1 + K_{eq}} \tag{6.10}$$

Preorganization excludes stable non-binding conformations of the binding site, Figure Figure 6.9. These conformations hinder binding because

they trap the binding site into a non-binding conformation. The binding site must first escape the trap, then bind the ligand, which lowers the binding interaction by the amount needed to escape the trap. The non-binding, low energy conformation stabilizes the unbound state by $\Delta\Delta G_{preorg}$, thereby reducing the Gibbs energy of dissociation for the complex.

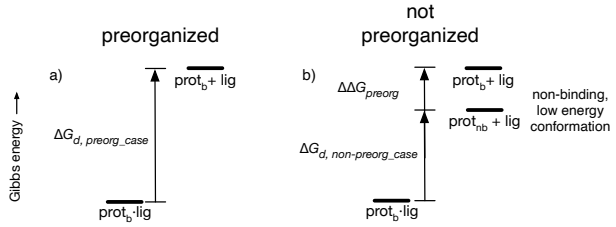
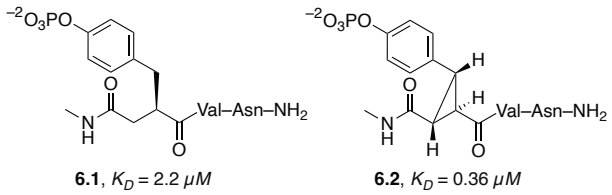


Figure 6.9. Preorganization excludes stable, non-binding conformations of the protein. a) When a protein is preorganized for binding, the binding site conformation in the ligand-free state has the same conformation as in the ligand-bound state. b) When the protein is not preorganized for binding, the protein exists in lower energy conformations that cannot bind the ligand. Binding requires the protein to first shift to a higher energy binding conformation. The cost of shifting to the binding conformation lowers the overall binding energy by $\Delta\Delta G_{preorg}$.

Constraining a flexible ligand into the three-dimensional shape it adopts when bound to a receptor often strengthens the binding interaction. For example, the phosphotyrosine-containing pseudopeptide 6.1, Figure Figure 6.10, bound tightly to the SH2 domain of a growth receptor binding protein, $K_D = 2.2\ \mu M$. Introducing a cyclopropyl ring to create compound 6.2 restricted the flexibility of the ligand so that its structure matches the bound conformation. The binding interaction tightened to $K_D = 0.36\ \mu M$.[146] Similar improvements are expected by restricting the flexibility of the binding site.



6.1, K_D = 2.2 μM **6.2**, K_D = 0.36 μM

Figure 6.10. Phosphotyrosine containing pseudopeptides bind tightly to the SH2 domain of a growth receptor binding protein. The cyclopropane ring in compound 6.2 restricted its flexibility and increased its affinity as compared to compound 6.1.

The preorganized protein structure for binding refers to both side chain and backbone orientations. The side chains must point in the correct direction to interact with the ligand, while the backbone must position these side chains at the correct distances. Incorrect backbone positions create binding sites that are too large or too small for the ligand. For example, a molecular dynamics simulation of digoxigenin-binding proteins found that favored conformation of the non-binding proteins was too large to interact with the entire digoxigenin surface due to incorrect backbone positions.[147]

Hydrophobic matching. Besides matching the shape of the ligand and binding site to maximize van der Waals interactions, matching additional non-covalent interactions further strengthens the interaction. For example, matching hydrophobic regions of the ligand and binding site avoids unsatisfied polar interactions. The structure of antibody fragment bound to methamphetamine showed a bound water molecule facing a hydrophobic region of the methamphetamine antigen.[148] Researchers hypothesized that releasing this water molecule would increase the hydrophobic character of the binding pocket and increase binding. Replacement of SerH93, which formed a hydrogen bond to the water molecule, with the slightly larger residue threonine increased the binding affinity 3.1-fold from K_d = 0.79 to 0.25 nM. A crystal structure of the improved variant

confirmed that this substitution forced out the water molecule.

Optimize hydrogen bonds. Hydrogen bonds between the binding site and ligand come at the cost losing hydrogen bonds to solvent and desolvation of the hydrogen bonding partners. Buried hydrogen bonds are rarely favorable, but solvent-exposed hydrogen maintain solvation so are more likely to be favorable. Hydrogen bonds often contribute to selectivity of binding.

Electrostatic interactions. Electrostatic interactions can contribute to the tighter binding of the protein and ligand in two ways. One way is to directly stabilize the protein-ligand complex; the other way is to speed up the formation of this complex, Figure Figure 6.11.

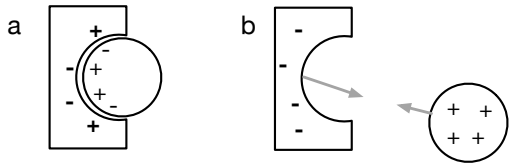


Figure 6.11. Two types of electrostatic interactions strengthen binding between partners. a) Partners with low or zero net charge associate with one another. Electrostatic interactions within the complex strengthen binding. b) Partners have opposite net charges. Favorable, long-range electrostatic interactions steer these oppositely charged molecules toward each other (gray arrows). Uncharged molecules of the same size would likely miss each other because their motions are random.

Direct stabilization of the protein-ligand complex refers to favorable interactions between oppositely charged atoms in the protein-ligand complex after it has formed. For example, improved electrostatic interactions increased the binding of monoclonal antibody hu3F8 (naxitamab) to the surface of cancer cells ~7-fold.[149] The antibody binds to two negatively charged sialic acid groups on glycolipid ganglioside

GD2. An Asp32His substitution replaced a negatively charged amino acid residue with a positively charged one just outside the binding site. Computer modeling predicted ~0.5 kcal/mol improvement in electrostatic interaction with the sialic acid groups in good agreement with the measured increase in binding. In another example, computer modeling predicted five substitutions that could strengthen electrostatic interactions between epidermal growth factor receptor and an antibody targeted to it (cetuximab, Erbitux®). Experiments confirmed that three of the five increased binding affinity. Substitutions Ser26Asp and Thr31Glu replaced polar residues with negatively charged residue, while Asn93Ala replaces a polar residue with an uncharged residue. All substitutions were in the antibody light chain. The combination of all three substitutions increased the binding 10-fold by decreasing K_d decreased from 490 pM to 52 pM.[150] The researchers noted that solvent-exposed electrostatic interactions and hydrogen bonds were more stabilizing than buried ones because they remain solvated. The loss of solvation energy made most buried electrostatic interactions and hydrogen bonds unfavorable.

Electrostatic interactions can also strengthen binding by long-range interactions that occur before binding occurs. Electrostatic steering increases binding strength by increasing the collision frequency between oppositely charged proteins and ligands, Figure {Figure 6.11}b above. Binding requires the protein and ligand to bump into one another as they randomly move through the solution. Electrostatic forces alter the collision rates between molecules because they act over long distances, that is, beyond van der Waals contact distances (see Chapter 2). This longer range of electrostatic interactions effectively makes charged molecules larger when they interact with other charged molecules. Nearby oppositely charged molecules can move toward each other to make van der Waals contact, while uncharged molecules must meet within their van der Waals radii. This increase in the number of collisions between oppositely charged molecules increases the rate of association of protein and ligand and thereby strengthens the binding interactions.

Increasing the electrostatic interaction between separated partners as compared to the complex by 1 kcal/mol increases the association rate by a factor of 2.8.[151] Charges that steer partners toward each other may also contribute to stronger binding after the complex forms. In contrast, partners with low net charge will not be steered even though they form favorable electrostatic interactions after binding.

To engineer such an increase, one increases the charge difference between the protein and ligand. For example, Selzer and coworkers[151] increased the affinity between β-lactamase (negative net charged) and its inhibitor (no net charge) 250-fold by giving the inhibitor a positive net charge. Four substitutions increased the positive charge by six units. A web tool predicts the increase in association rate upon mutagenesis: https://webhome.weizmann.ac.il/home/bcges/PARE.html.

The components of binding Gibbs energy, ΔH and −TΔS, do not map simply onto specific molecular events, making mechanistic interpretation challenging. Enthalpy (ΔH) reflects the balance of non-covalent interactions formed and broken during binding. Favorable contributions (−ΔH°) arise from hydrogen bonds, van der Waals contacts, and electrostatic interactions between protein and ligand. Unfavorable contributions (+ΔH°) result from desolvation penalties—breaking hydrogen bonds between the ligand and water, and between binding site residues and water. The observed ΔH° is the net sum of these competing effects.

Entropy (-TΔS°) reflects changes in the number of accessible microstates of the system. Unfavorable contributions (−ΔS°) include loss of translational and rotational freedom when the ligand binds, and loss of conformational flexibility in both protein (side chain immobilization, induced fit) and ligand (freezing of rotatable bonds). Favorable contributions (+ΔS°) primarily arise from release of ordered water molecules from the binding site and ligand surface into bulk solvent—the hydrophobic effect. Additionally, protein conformational changes may sometimes increase local disorder.

The observed Gibbs energy of binding represents the net balance of all these opposing entropic and enthalpic terms, complicating direct assignment to specific binding mechanisms.

6.4 Computational design of tighter binding

Docking models protein-ligand complexes. Designing improved binding interactions relies on computer modeling of the interactions. An accurate structure of the ligand complexed to the binding protein is the best starting point for computer modeling, but such structures are rarely available. Protein-ligand docking is a molecular modeling technique to create a model of a protein-ligand complex from a structure of the binding protein. SwissDock is a web tool for protein-ligand docking.[152] The most common application of docking is drug design, but it is also useful for protein design. Docking predicts where the ligand binds to the protein, the geometry of the bound ligand and the strength of the binding interaction.

As with all protein modeling, the two challenges are sampling and scoring. Sampling finds possible binding locations, orientations and conformations. Scoring evaluates and ranks these possibilities to identify the best ones. If the sampling does not generate the best-binding structure, then scoring cannot identify it.

For sampling, SwissDock uses an approach called Attracting Cavities.[153] It replaces the complex protein with a simplified set of points that represent attractive interactions with a protein cavity. This simplification eliminates bumping interactions from the protein atoms, which allows more efficient sampling of the conformations to find those that maximize the attractive interactions.

For scoring, SwissDock restores the protein structure and calculates the energies using the CHARMM force field, which is similar to AMBER and

intended for modeling biological molecules. Solvation calculations use an implicit solvent model, which lacks individual solvent molecules. The energy contributions to scoring are the ligand-protein interaction energy, the conformational energy of the ligand and protein and the desolvation energy of the protein and ligand. The calculated energies are estimates of the Gibbs energy of binding.

SwissDock accurately reproduced the experimental structures (rmsd ≤2 Å) of 285 protein-ligand complexes in 73% of the cases.[154] It performs best for ligands with limited numbers of possible conformations (≤10 rotatable bonds), and for binding orientations where most of the ligand surface (≥85%) interacts with protein. The conformations of large solvent-exposed regions of the ligand are difficult to predict accurately.

Computational design of binding proteins. Predicting substitutions that improve the affinity of a protein for a target ligand has shown some success, but remains challenging. Computational design using Rosetta predicted seventeen proteins that bind to the steroid portion of digoxin called digoxigenin.[155] Two of these proteins showed micromolar affinity for the target, which is relatively weak affinity, but the ability to predict such proteins is an impressive accomplishment. Further optimization did not use computation, but directed evolution methods (see Chapters 6 and 7) and yielded proteins with picomolar affinity for digoxigenin. The designed protein included hydrogen bond interactions with the hydroxyl groups of digoxigenin, so it showed selectivity against analogs lacking these hydroxyl groups. Other computational design were even less successful. Morin et al.[156] designed twelve possible binding sites in *endo*-1,4-β-xylanase for the antibiotic vancomycin, but none of the designs bound vancomycin. X-ray structure analysis of four designs showed the predicted conformation, but molecular dynamics showed that the designs were highly flexible. The authors suggest that the designed sites spent most of their time in conformations not suitable for binding.

6.5 Selective binding

Define selectivity. Selectivity, S, is the ability to distinguish between alternatives. Binding selectivity is the ability of a protein to bind one ligand more tightly than another one, Figure Figure 6.12. The selectivity value, S, is the ratio of the affinities of the protein for the preferred and non-preferred ligands. This value depends on comparison molecules, so they should always be included with the number. For many proteins, their ability of proteins to favor binding one ligand in the presence of competing ligands is what makes them valuable. For example, the cardiovascular antibody drug abciximab selectively binds to a receptor involved in platelet aggregation to inhibit formation of blood clots.

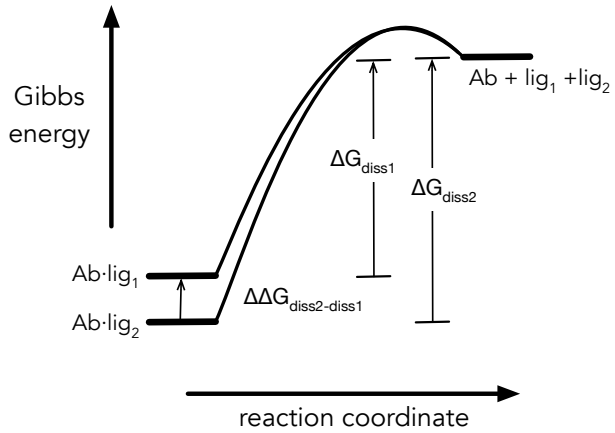


Figure 6.12. Gibbs energy diagram of selective binding of ligand 2 (lig_2) over ligand 1 (lig_1) to an antibody (Ab)

The affinity of a protein for a ligand is given by the equilibrium association constant, K_{as}, for the binding reaction, eq Equation 6.11. It is the inverse of the dissociation constant defined in Figure Figure 6.2 above.

$$\text{Ab + ligand} \rightleftharpoons \text{Ab·ligand} \quad K_{as} = \frac{\text{[Ab·ligand]}}{\text{[Ab]·[ligand]}} \tag{6.11}$$

Binding selectivity, S_{bind}, is the ratio of the affinities for the two protein-ligand complexes being compared, eq. Equation 6.12. By convention, the equilibrium association constant for the tighter-binding ligand, t, is in the numerator and that for the weaker-binding ligand, w, is in the denominator yielding binding selectivity values that are greater than one. The binding selectivity should also mention which ligands are being compared. A selectivity of ten corresponds to ten-fold tighter binding of one ligand as compared to the other ligand. A selectivity of one corresponds to no selectivity; that is, equal affinity for both.

$$S_{bind} = \frac{K_{as-t}}{K_{as-w}} = \frac{\frac{1}{K_{d-t}}}{\frac{1}{K_{d-w}}} = \frac{K_{d-w}}{K_{d-t}} \tag{6.12}$$

Binding selectivity can also be expressed in terms of the equilibrium constants for dissociation, which are the inverse of the equilibrium constants for association. The dissociation constant for the weaker-binding ligand is in the numerator while that for the tighter-binding ligand is in the denominator.The selectivity of an protein for two competing ligands can be measured by measuring each binding affinity separately, then dividing the two values.

The Gibbs energy change associated with binding selectivity, eq. Equation 6.13, is the difference between the Gibbs energy changes of the two association reactions. The Gibbs energy change is negative since association of the tighter-binding ligand is more favorable.

$$\Delta\Delta G_{S_{bind}} = \Delta G_{as-t} - \Delta G_{as-w} - RT \ln \left(\frac{K_{as-t}}{K_{as-w}} \right) = -RT \ln \left(S_{bind} \right) \tag{6.13}$$

To engineer a protein for increased selectivity between a preferred and a non-preferred ligand, one can either decrease the affinity of the protein for the non-preferred ligand (negative selection) or increase the affinity of the protein for the preferred ligand (positive selection). The best approach will depend on the details of the application.

Glossary

Boltzmann factor or weight is the probability that the binding site conformation when unbound is the same as that when it is bound. The Boltzmann factor is a way to measure the preorganization of a binding site for binding.

References

1. Morrison, C., & Lähteenmäki, R. (2018). Public biotech in 2017-the numbers. *Nat. Biotechnol.*, *36*(7), 576–584. https://doi.org/10.1038/nbt.4175
2. Brannigan, J. A., & Wilkinson, A. J. (2002). Protein engineering 20 years on. *Nat. Rev. Mol. Cell Biol.*, *3*(12), 964–970. https://doi.org/10.1038/nrm975
3. Estell, D. A., Graycar, T. P., & Wells, J. A. (1985). Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.*, *260*(11), 6518–6521. https://doi.org/10.1016/S0021-9258(18)88811-1
4. Pace, C. N., & Hermans, J. (1975). The stability of globular proteins. *CRC Crit. Rev. Biochem.*, *3*(1), 1–43. https://doi.org/10.3109/10409237509102551

5. Taverna, D. M., & Goldstein, R. A. (2002). Why are proteins marginally stable? *Proteins*, *46*(1), 105–109. https://doi.org/10.1002/prot.10016
6. Fersht, A. R., & Serrano, L. (1993). Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.*, *3*(1), 75–83. https://doi.org/10.1016/0959-440X(93)90205-Y
7. Bloom, J. D., Lu, Z., Chen, D., Raval, A., Venturelli, O. S., & Arnold, F. H. (2007). Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.*, *5*(1), 97–21. https://doi.org/10.1186/1741-7007-5-29
8. Gordon, S. R., Stanley, E. J., Wolf, S., Toland, A., Wu, S. J., Hadidi, D., Mills, J. H., Baker, D., Pultz, I. S., & Siegel, J. B. (2012). Computational design of an α-gliadin peptidase. *J. Am. Chem. Soc.*, *134*(50), 20513–20520. https://doi.org/10.1021/ja3094795
9. Harvey, S. B., Stone, M. D., Martinez, M. B., & Nelsestuen, G. L. (2003). Mutagenesis of the γ-carboxyglutamic acid domain of human factor VII to generate maximum enhancement of the membrane contact site. *J. Biol. Chem.*, *278*(10), 8363–8369. https://doi.org/10.1074/jbc.M211629200
10. Persson, E., Olsen, O. H., Bjørn, S. E., & Ezban, M. (2012). Vatreptacog alfa from conception to clinical proof of concept. *Semin. Thromb. Hemost.*, *38*(3), 274–281. https://doi.org/10.1055/s-0032-1302442
11. Bolli, G. B., Di Marchi, R. D., Park, G. D., Pramming, S., & Koivisto, V. A. (1999). Insulin analogues and their potential in the management of diabetes mellitus. *Diabetologia*, *42*(10), 1151–1167. https://doi.org/10.1007/s001250051286
12. Ryu, J. K., Kim, H. S., & Nam, D. H. (2012). Current status and perspectives of biopharmaceutical drugs. *Biotechnol. Bioprocess Eng.*, *17*(5), 900–911. https://doi.org/10.1007/s12257-012-0095-1

13. Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B. B., Henner, D., Wong, W. L. T., Rowland, A. M., Kotts, C., Carver, M. E., & Shepard, H. M. (1992). Humanization of an anti-p185^{HER2} antibody for human cancer therapy. *Proc. Natl. Acad. Sci. U. S. A.*, *89*(10), 4285–4289. https://doi.org/10.1073/pnas.89.10.4285
14. Sarkar, C. A., Lowenhaupt, K., Horan, T., Boone, T. C., Tidor, B., & Lauffenburger, D. A. (2002). Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nat. Biotechnol.*, *20*(9), 908–913. https://doi.org/10.1038/nbt725
15. Cantor, J. R., Yoo, T. H., Dixit, A., Iverson, B. L., Forsthuber, T. G., & Georgiou, G. (2011). Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(4), 1272–1277. https://doi.org/10.1073/pnas.1014739108
16. Loosmore, S. M., Zealey, G. R., Boux, H. A., Cockle, S. A., Radika, K., Fahim, R. E., Zobrist, G. J., Yacoob, R. K., Chong, P. C., & Yao, F. L. (1990). Engineering of genetically detoxified pertussis toxin analogs for development of a recombinant whooping cough vaccine. *Infect. Immun.*, *58*(11), 3653–3662. https://doi.org/10.1128/iai.58.11.3653-3662.1990
17. Hsieh, C.-L., Goldsmith, J. A., Schaub, J. M., DiVenere, A. M., Kuo, H.-C., Javanmardi, K., Le, K. C., Wrapp, D., Lee, A. G., Liu, Y., Chou, C.-W., Byrne, P. O., Hjorth, C. K., Johnson, N. V., Ludes-Meyers, J., Nguyen, A. W., Park, J., Wang, N., Amengor, D., ... McLellan, J. S. (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science*, *369*(6510), 1501–1505. https://doi.org/10.1126/science.abd0826

18. Padgette, S. R., Kolacz, K. H., Delannay, X., Re, D. B., Tinius, C. N., Rhodes, W. K., Otero, Y. I., Barry, G. F., Eichholtz, D. A., Peschke, V. M., Nida, D. L., Taylor, N. B., & Kishore, G. M. (1995). Development, identification, and characterization of a glyphosate-tolerant soybean line. *Crop Sci.*, *35*(5), 1451–1461. https://doi.org/10.2135/cropsci1995.0011183x003500050032x
19. Castle, L. A., Siehl, D. L., Gorton, R., Patten, P. A., Chen, Y. H., Bertain, S., Cho, H.-J., Duck, N., Wong, J., Liu, D., & Lassner, M. W. (2004). Discovery and directed evolution of a glyphosate tolerance gene. *Science*, *304*(5674), 1151–1154. https://doi.org/10.1126/science.1096770
20. Bravo, A., & Soberon, M. (2008). How to cope with insect resistance to Bt toxins? *Trends Biotechnol.*, *26*(10), 573–579. https://doi.org/10.1016/j.tibtech.2008.06.005
21. Kurek, I., Chang, T. K., Bertain, S. M., Madrigal, A., Liu, L., Lassner, M. W., & Zhu, G. (2007). Enhanced thermostability of *Arabidopsis* rubisco activase improves photosynthesis and growth rates under moderate heat stress. *Plant Cell*, *19*(10), 3230–3241. https://doi.org/10.1105/tpc.107.054171
22. Whittle, E., & Shanklin, J. (2001). Engineering Δ^9-16:0-acyl carrier protein (ACP) desaturase specificity based on combinatorial saturation mutagenesis and logical redesign of the castor Δ^9-18:0-ACP desaturase. *J. Biol. Chem.*, *276*(24), 21500–21505. https://doi.org/10.1074/jbc.M102129200
23. Lin, H., & Lee, Y. K. (2017). Genetic engineering of medium-chain-length fatty acid synthesis in *Dunaliella tertiolecta* for improved biodiesel production. *J. Appl. Phycol.*, *29*(6), 2811–2819. https://doi.org/10.1007/s10811-017-1210-7
24. Yao, M. Z., Zhang, Y. H., Lu, W. L., Hu, M. Q., Wang, W., & Liang, A. H. (2012). Phytases: Crystal structures, protein engineering and potential biotechnological applications. *J. Appl. Microbiol.*, *112*(1), 1–14. https://doi.org/10.1111/j.1365-2672.2011.05181.x

25. Richardson, T. H., Tan, X., Frey, G., Callen, W., Cabell, M., Lam, D., Macomber, J., Short, J. M., Robertson, D. E., & Miller, C. (2002). A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable α-amylase. *J. Biol. Chem.*, *277*(29), 26501–26507. https://doi.org/10.1074/jbc.M203183200
26. Quax, W. J., Mrabet, N. T., Luiten, R. G. M., Schuurhuizen, P. W., Stanssens, P., & Lasters, I. (1991). Enhancing the thermostability of glucose-isomerase by protein engineering. *Bio/Technology*, *9*(8), 738–742. https://doi.org/10.1038/nbt0891-738
27. Bell, E. L., Finnigan, W., France, S. P., Green, A. P., Hayes, M. A., Hepworth, L. J., Lovelock, S. L., Niikura, H., Osuna, S., Romero, E., Ryan, K. S., Turner, N. J., & Flitsch, S. L. (2021). Biocatalysis. *Nat. Rev. Methods Primers*, *1*(1), 46. https://doi.org/10.1038/s43586-021-00044-z
28. van der Wielen, L. A. M., Ottens, M., & Straathof, A. J. J. (2001). Process Technology and Process Integration in the Preparation of Penicillins. In A. Bruggink (Ed.), *Synthesis of β-Lactam antibiotics: Chemistry, biocatalysis & process integration* (pp. 150–205). Springer Netherlands.
29. Ran, N., & Frost, J. W. (2007). Directed evolution of 2-keto-3-deoxy-6-phosphogalactonate aldolase to replace 3-deoxy-D-arabino-heptulosonic acid 7-phosphate synthase. *J. Am. Chem. Soc.*, *129*(19), 6130–6139. https://doi.org/10.1021/ja067330p
30. Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., ... Newman, J. D. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, *496*(7446), 528–532. https://doi.org/10.1038/nature12051
31. Anastas, P. T., & Warner, J. C. (2000). *Green chemistry : Theory and practice*. Oxford University Press.

32. Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J., Brands, J., Devine, P. N., Huisman, G. W., & Hughes, G. J. (2010). Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, *329*(5989), 305–309. https://doi.org/10.1126/science.1188934

33. Stone, E. M., Paley, O., Hu, J., Ekerdt, B., Cheung, N.-K., & Georgiou, G. (2012). De novo engineering of a human cystathionine-γ-lyase for systemic L-methionine depletion cancer therapy. *ACS Chem. Biol.*, *7*(11), 1822–1829. https://doi.org/10.1021/cb300335j

34. Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, *27*(10), 946–950. https://doi.org/10.1038/nbt.1568

35. Jobe, A., & Bourgeois, S. (1972). *lac* Repressor-operator interaction. VI. The natural inducer of the lac operon. *J. Mol. Biol.*, *69*(3), 397–408. https://doi.org/10.1016/0022-2836(72)90253-7

36. Egel, R. (1988). The '*lac*' operon - an irrelevant paradox? *Trends Genet.*, *4*(2), 31–31. https://doi.org/10.1016/0168-9525(88)90062-5

37. Valdez-Cruz, N. A., Caspeta, L., Pérez, N. O., Ramírez, O. T., & Trujillo-Roldán, M. A. (2010). Production of recombinant proteins in *E. Coli* by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microb Cell Fact*, *9*(1), 18. https://doi.org/10.1186/1475-2859-9-18

38. Austin, H. P., Allen, M. D., Donohoe, B. S., Rorrer, N. A., Kearns, F. L., Silveira, R. L., Pollard, B. C., Dominick, G., Duman, R., El Omari, K., Mykhaylyk, V., Wagner, A., Michener, W. E., Amore, A., Skaf, M. S., Crowley, M. F., Thorne, A. W., Johnson, C. W., Woodcock, H. L., ... Beckham, G. T. (2018). Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(19), E4350–E4357. https://doi.org/10.1073/pnas.1718804115

39. Hutchison 3rd, C. A., Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., & Smith, M. (1978). Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem.*, *253*(18), 6551–6560. https://doi.org/10.1016/S0021-9258(19)46967-6
40. Silva, D., Santos, G., Barroca, M., & Collins, T. (2017). Inverse PCR for point mutation introduction. In L. Domingues (Ed.), *PCR: Methods and Protocols* (pp. 87–100). Springer New York. https://doi.org/10.1007/978-1-4939-7060-5_5
41. Roy, S., & Caruthers, M. (2013). Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. *Molecules*, *18*(11), 14268–14284. https://doi.org/10.3390/molecules181114268
42. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S., & Ryadnov, M. G. (2023). DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.*, *7*(3), 144–161. https://doi.org/10.1038/s41570-022-00456-9
43. Seydel, C. (2023). DNA writing technologies moving toward synthetic genomes. *Nat Biotechnol.*, *41*(11), 1504–1509. https://doi.org/10.1038/s41587-023-02006-0
44. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison 3rd, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318
45. Green, M. R., & Sambrook, J. (2012). *Molecular Cloning: A Laboratory Manual* (4th ed.). Cold Spring Harbor Press.
46. Ball, P. (2018). A compassionate substance. How science and myth collide in water. *Lapham's Quart.*, *11*(Summer).
47. Pace, C. N., Scholtz, J. M., & Grimsley, G. R. (2014). Forces stabilizing proteins. *FEBS Lett.*, *588*(14), 2177–2184. https://doi.org/10.1016/j.febslet.2014.05.006

48. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427
49. 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., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
50. Mura, C., McCrimmon, C. M., Vertrees, J., & Sawaya, M. R. (2010). An introduction to biomolecular graphics. *PLoS Comput. Biol.*, *6*(8), e1000918. https://doi.org/10.1371/journal.pcbi.1000918
51. Leach, A. R. (2001). *Molecular modelling: Principles and applications*. Prentice Hall.
52. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz Jr, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, *117*(19), 5179–5197. https://doi.org/10.1021/ja00124a002
53. Allinger, N. L., Zhou, X., & Bergsma, J. (1994). Molecular mechanics parameters. *J. Mol. Struct. THEOCHEM*, *312*(1), 69–83. https://doi.org/10.1016/S0166-1280(09)80008-0

54. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem\@. Theory Comput.*, *13*(6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125

55. Homeyer, N., & Gohlke, H. (2012). Free energy calculations by the molecular mechanics Poisson-Boltzmann surface area method. *Mol. Inf.*, *31*(2), 114–122. https://doi.org/10.1002/minf.201100135

56. Bae, E., & Phillips Jr, G. N. (2005). Identifying and engineering ion pairs in adenylate kinases. *J. Biol. Chem.*, *280*(35), 30943–30948. https://doi.org/10.1074/jbc.M504216200

57. Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J., & Osuna, S. (2019). Hidden conformations in *Aspergillus niger* monoamine oxidase are key for catalytic efficiency. *Angew. Chem. Int. Ed.*, *58*(10), 3097–3101. https://doi.org/10.1002/anie.201812532

58. Thomas, L. L., Christakis, T. J., & Jorgensen, W. L. (2006). Conformation of alkanes in the gas phase and pure liquids. *J. Phys. Chem. B*, *110*(42), 21198–21204. https://doi.org/10.1021/jp064811m

59. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N., & Hilvert, D. (2013). Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.*, *9*(8), 494–498. https://doi.org/10.1038/nchembio.1276

60. Crean, R. M., Gardner, J. M., & Kamerlin, S. C. L. (2020). Harnessing conformational plasticity to generate designer enzymes. *J. Am. Chem. Soc.*, *142*(26), 11324–11342. https://doi.org/10.1021/jacs.0c04924

61. Robic, S., Guzman-Casado, M., Sanchez-Ruiz, J. M., & Marqusee, S. (2003). Role of residual structure in the unfolded state of a thermophilic protein. *Proc. Natl. Acad. Sci. U. S. A.*, *100*(20), 11345–11349. https://doi.org/10.1073/pnas.1635051100

62. Rosenthal, L., Rabolt, J. F., & Hummel, J. (1982). An investigation of the conformational equilibrium of *n*-butane in a solvent using Raman spectroscopy. *J. Chem. Phys.*, *76*(2), 817–820. https://doi.org/10.1063/1.443052
63. Vázquez-Figueroa, E., Yeh, V., Broering, J. M., Chaparro-Riggers, J. F., & Bommarius, A. S. (2008). Thermostable variants constructed via the structure-guided consensus method also show increased stability in salts solutions and homogeneous aqueous-organic media. *Protein Eng. Des. Sel.*, *21*(11), 673–680. https://doi.org/10.1093/protein/gzn048
64. Mayer, S., Rüdiger, S., Ang, H. C., Joerger, A. C., & Fersht, A. R. (2007). Correlation of levels of folded recombinant P53 in *Escherichia coli* with thermodynamic stability in vitro. *J. Mol. Biol.*, *372*(1), 268–276. https://doi.org/10.1016/j.jmb.2007.06.044
65. Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, *103*(15), 5869–5874. https://doi.org/10.1073/pnas.0510098103
66. Adams, M. W., & Kelly, R. M. (1998). Finding and using hyperthermophilic enzymes. *Trends Biotechnol.*, *16*(8), 329–332. https://doi.org/10.1016/s0167-7799(98)01193-7
67. Merz, A., Yee, M.-C., Szadkowski, H., Pappenberger, G., Crameri, A., Stemmer, W. P. C., Yanofsky, C., & Kirschner, K. (2000). Improving the catalytic activity of a thermophilic enzyme at low temperatures. *Biochemistry*, *39*(5), 880–889. https://doi.org/10.1007/pl00003701
68. Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.*, *17*(4), 263–270. https://doi.org/10.1002/humu.22
69. Joerger, A. C., Ang, H. C., & Fersht, A. R. (2006). Structural basis for understanding oncogenic P53 mutations and designing rescue drugs. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(41), 15056–15061. https://doi.org/10.1073/pnas.0607286103

70. Park, S. H., & Bao, G. (2021). CRISPR/Cas9 gene editing for curing sickle cell disease. *Transfusion and Apheresis Science*, *60*(1), 103060. https://doi.org/10.1016/j.transci.2021.103060
71. Pace, C. N., & Shaw, K. L. (2000). Linear extrapolation method of analyzing solvent denaturation curves. *Proteins Struct. Funct. Bioinforma.*, *Suppl 4*, 1–7. https://doi.org/10.1002/1097-0134(2000)41:4+%3C1::aid-prot10%3E3.3.co;2-u
72. Baxa, M. C., Haddadian, E. J., Jumper, J. M., Freed, K. F., & Sosnick, T. R. (2014). Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(43), 15396–15401. https://doi.org/10.1073/pnas.1407768111
73. Horovitz, A., & Fersht, A. R. (1992). Co-operative interactions during protein folding. *J. Mol. Biol.*, *224*(3), 733–740. https://doi.org/10.1016/0022-2836(92)90557-Z
74. Scholtz, J. M., Grimsley, G. R., & Pace, C. N. (2009). Solvent denaturation of proteins and interpretations of the m value. *Methods Enzymol.*, *466*, 549–565. https://doi.org/10.1016/S0076-6879(09)66023-7
75. Fersht, A. (1999). Chapter 17, Protein Stability. In *Structure and Mechanism in Protein Science*. Freeman.
76. Becktel, W. J., & Schellman, J. A. (1987). Protein stability curves. *Biopolymers*, *26*(11), 1859–1878. https://doi.org/10.1002/bip.360261104
77. Robertson, A. D., & Murphy, K. P. (1997). Protein structure and the energetics of protein stability. *Chem. Rev.*, *97*(5), 1251–1267.
78. Pucci, F., Kwasigroch, J. M., & Rooman, M. (2017). SCooP: An accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*, *33*(21), 3415–3422.

79. Pastore, A., Martin, S. R., Politou, A., Kondapalli, K. C., Stemmler, T., & Temussi, P. A. (2007). Unbiased cold denaturation: Low- and high-temperature unfolding of yeast frataxin under physiological conditions. *J. Am. Chem. Soc.*, *129*(17), 5374–5375. https://doi.org/10.1021/ja0714538
80. Razvi, A., & Scholtz, J. M. (2006). Lessons in stability from thermophilic proteins. *Protein Sci.*, *15*(7), 1569–1578. https://doi.org/10.1110/ps.062130306
81. Eijsink, V. G. H., Bjørk, A., G\a aseidnes, S., Sirev\a ag, R., Synstad, B., van den Burg, B., & Vriend, G. (2004). Rational engineering of enzyme stability. *J. Biotechnol.*, *113*(1-3), 105–120. https://doi.org/10.1016/j.jbiotec.2004.03.026
82. Dumon, C., Varvak, A., Wall, M. A., Flint, J. E., Lewis, R. J., Lakey, J. H., Morland, C., Luginbühl, P., Healey, S., Todaro, T., DeSantis, G., Sun, M., Parra-Gessert, L., Tan, X., Weiner, D. P., & Gilbert, H. J. (2008). Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure. *J. Biol. Chem.*, *283*(33), 22557–22564. https://doi.org/10.1074/jbc.M800936200
83. Shoichet, B. K., Baase, W. A., Kuroki, R., & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.*, *92*(2), 452–456. https://doi.org/10.1073/pnas.92.2.452
84. Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L., van Loon, A. P. G. M., & Wyss, M. (2002). The consensus concept for thermostability engineering of proteins: Further proof of concept. *Protein Eng.*, *15*(5), 403–411. https://doi.org/10.1093/protein/15.5.403
85. Magliery, T. J. (2015). Protein stability: Computation, sequence statistics, and new experimental methods. *Current Opinion in Structural Biology*, *33*, 161–168. https://doi.org/10.1016/j.sbi.2015.09.002

86. Jones, B. J., Kan, C. N. E., Luo, C., & Kazlauskas, R. J. (2020). Consensus Finder web tool to predict stabilizing substitutions in proteins. In D. S. Tawfik (Ed.), *Enzyme Engineering and Evolution: General Methods* (Vol. 643, pp. 129–148). Academic Press. https://doi.org/10.1016/bs.mie.2020.07.010
87. Musil, M., Stourac, J., Bendl, J., Brezovsky, J., Prokop, Z., Zendulka, J., Martinek, T., Bednar, D., & Damborsky, J. (2017). FireProt: Web server for automated design of thermostable proteins. *Nucleic Acids Res.*, *45*(W1), W393–W399. https://doi.org/10.1093/nar/gkx285
88. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L., & Bryan, P. N. (1989). Large increases in general stability for subtilisin BPN' through incremental changes in the free-energy of unfolding. *Biochemistry*, *28*(18), 7205–7213. https://doi.org/10.1021/bi00444a012
89. Schwehm, J. M., Fitch, C. A., Dang, B. N., García-Moreno E., B., & Stites, W. E. (2003). Changes in stability upon charge reversal and neutralization substitution in staphylococcal nuclease are dominated by favorable electrostatic effects. *Biochemistry*, *42*(4), 1118–1128. https://doi.org/10.1021/bi0266434
90. Trudeau, D. L., Kaltenbach, M., & Tawfik, D. S. (2016). On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.*, *33*(10), 2633–2641. https://doi.org/10.1093/molbev/msw138
91. Clarke, J., Henrick, K., & Fersht, A. R. (1995). Disulfide mutants of barnase I: Changes in stability and structure assessed by biophysical methods and x-ray crystallography. *J. Mol. Biol.*, *253*(3), 493–504. https://doi.org/10.1006/jmbi.1995.0568
92. Pace, C. N., Grimsley, G. R., Thomson, J. A., & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.*, *263*(24), 11820–11825.

93. Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.*, *151*(2), 261–287. https://doi.org/10.1016/0022-2836(81)90515-5
94. Perry, L. J., & Wetzel, R. (1984). Disulfide bond engineered into T4 lysozyme: Stabilization of the protein toward thermal inactivation. *Science*, *226*(4674), 555–557. https://www.jstor.org/stable/1693872
95. Matsumura, M., Becktel, W. J., Levitt, M., & Matthews, B. W. (1989). Stabilization of phage T4 lysozyme by engineered disulfide bonds. *Proc. Natl. Acad. Sci. U. S. A.*, *86*, 6562–6566. https://doi.org/10.1073/pnas.86.17.6562
96. Roesler, K. R., & Rao, A. G. (2000). A single disulfide bond restores thermodynamic and proteolytic stability to an extensively mutated protein. *Protein Science*, *9*(9), 1642–1650. https://doi.org/10.1110/ps.9.9.1642
97. Kanaya, S., Katsuda, C., Kimura, S., Nakai, T., Kitakuni, E., Nakamura, H., Katayanagi, K., Morikawa, K., & Ikehara, M. (1991). Stabilization of *Escherichia coli* ribonuclease H by introduction of an artificial disulfide bond. *Journal of Biological Chemistry*, *266*(10), 6038–6044. https://doi.org/10.1016/S0021-9258(18)38080-3
98. Hazes, B., & Dijkstra, B. W. (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng. Des. Sel.*, *2*(2), 119–125. https://doi.org/10.1093/protein/2.2.119
99. Dombkowski, A. A. (2003). Disulfide by Design^{TM}: A computational method for the rational design of disulfide bonds in proteins. *Bioinformatics*, *19*(14), 1852–1853. https://doi.org/10.1093/bioinformatics/btg231
100. Suplatov, D., Timonina, D., Sharapova, Y., & Švedas, V. (2019). Yosshi: A web-server for disulfide engineering by bioinformatic analysis of diverse protein families. *Nucleic Acids Res.*, *47*(W1), W308–W314. https://doi.org/10.1093/nar/gkz385

101. Pikkemaat, M. G., Linssen, A. B. M., Berendsen, H. J. C., & Janssen, D. B. (2002). Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng.*, *15*(3), 185–192. https://doi.org/10.1093/protein/15.3.185
102. Stewart, E. J., Åslund, F., & Beckwith, J. (1998). Disulfide bond formation in the *Escherichia coli* cytoplasm: An *in Vivo* role reversal for the thioredoxins. *EMBO J.*, *17*(19), 5543–5550. https://doi.org/10.1093/emboj/17.19.5543
103. Zhou, H.-X. (2004). Loops, linkages, rings, catenanes, cages, and crowders: Entropy-based strategies for stabilizing proteins. *Acc. Chem. Res.*, *37*(2), 123–130. https://doi.org/10.1021/ar0302282
104. Watanabe, K., & Suzuki, Y. (1998). Protein thermostabilization by proline substitutions. *J. Mol. Catal. B Enzym.*, *4*(4), 167–180. https://doi.org/10.1016/s1381-1177(97)00031-3
105. Fu, H., Grimsley, G. R., Razvi, A., Scholtz, J. M., & Pace, C. N. (2009). Increasing protein stability by improving beta-turns. *Proteins*, *77*(3), 491–498. https://doi.org/10.1002/prot.22509
106. Matthews, B. W., Nicholson, H., & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U. S. A.*, *84*(19), 6663–6667. https://doi.org/10.1073/pnas.84.19.6663
107. Scott, K. A., Alonso, D. O. V., Sato, S., Fersht, A. R., & Daggett, V. (2007). Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(8), 2661–2666. https://doi.org/10.1073/pnas.0611182104
108. Liu, Z., Lemmonds, S., Huang, J., Tyagi, M., Hong, L., & Jain, N. (2018). Entropic contribution to enhanced thermal stability in the thermostable P450 CYP119. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(43), E10049–E10058. https://doi.org/10.1073/pnas.1807473115
109. 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*(46), 7745–7751. https://doi.org/10.1002/anie.200602795

110. Floor, R. J., Wijma, H. J., Colpa, D. I., Ramos-Silva, A., Jekel, P. A., Szymanski, W., Feringa, B. L., Marrink, S. J., & Janssen, D. B. (2014). Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem*, *15*(11), 1660–1672. https://doi.org/10.1002/cbic.201402128

111. Xiao, Z., Bergeron, H., Grosse, S., Beauchemin, M., Garron, M.-L., Shaya, D., Sulea, T., Cygler, M., & Lau, P. C. K. (2008). Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-temperature-guided sequence alignment. *Appl. Environ. Microbiol.*, *74*(4), 1183–1189. https://doi.org/10.1128/AEM.02220-07

112. Petsko, G. A. (2001). Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat." *Methods Enzymol.*, *334*, 469–478. https://doi.org/10.1016/S0076-6879(01)34486-5

113. Gribenko, A. V., Patel, M., Liu, J., McCallum, S., Wang, C., & Makhatadze, G. I. (2009). Rational stabilization of enzymes by computational redesign of surface charge–charge interactions. *Proc. Natl. Acad. Sci. U. S. A.*, *106*(8), 2601–2606. https://doi.org/10.1073/pnas.0808220106

114. Contessoto, V. G., de Oliveira, V. M., Fernandes, B. R., Slade, G. G., & Leite, V. B. P. (2018). TKSA-MC: A web server for rational mutation through the optimization of protein charge interactions. *Proteins*, *86*(11), 1184–1188. https://doi.org/10.1002/prot.25599

115. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987–2998. https://doi.org/10.1021/bi902153g

116. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Res.*, *33*(Web Server issue), W382–8. https://doi.org/10.1093/nar/gki387

117. Korkegian, A., Black, M. E., Baker, D., & Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. *Sci. Wash.*, *308*(5723), 857–860. https://doi.org/10.1126/science.1107387
118. Arabnejad, H., Dal Lago, M., Jekel, P. A., Floor, R. J., Thunnissen, A.-M. W. H., Terwisscha van Scheltinga, A. C., Wijma, H. J., & Janssen, D. B. (2017). A robust cosolvent-compatible halohydrin dehalogenase by computational library design. *Protein Eng. Des. Sel.*, *30*(3), 173–187. https://doi.org/10.1093/protein/gzw068
119. Jin, Y., Johannissen, L. O., & Hay, S. (2021). Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins*, *89*(8), 915–921. https://doi.org/10.1002/prot.26068
120. Weinstein, J. J., Goldenzweig, A., Hoch, S., & Fleishman, S. J. (2021). PROSS 2: A new server for the design of stable and highly expressed protein variants. *Bioinformatics*, *37*(1), 123–125. https://doi.org/10.1093/bioinformatics/btaa1071
121. Gao, D., Narasimhan, D. L., Macdonald, J., Brim, R., Ko, M.-C., Landry, D. W., Woods, J. H., Sunahara, R. K., & Zhan, C.-G. (2009). Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, *75*(2), 318–323. https://doi.org/10.1124/mol.108.049486
122. van den Burg, B., Vriend, G., Veltman, O. R., & Eijsink, V. G. H. (1998). Engineering an enzyme to resist boiling. *Proc. Natl. Acad. Sci. U. S. A.*, *95*(5), 2056–2060. https://doi.org/10.1073/pnas.95.5.2056
123. Bjørk, A., Dalhus, B., Mantzilas, D., Sirev\a ag, R., & Eijsink, V. G. H. (2004). Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface. *J. Mol. Biol.*, *341*(5), 1215–1226. https://doi.org/10.1016/j.jmb.2004.06.079

124. Kabashima, T., Li, Y., Kanada, N., Ito, K., & Yoshimoto, T. (2001). Enhancement of the thermal stability of pyroglutamyl peptidase I by introduction of an intersubunit disulfide bond. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1547*(2), 214–220. https://doi.org/10.1016/s0167-4838(01)00185-6
125. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., & Trout, B. L. (2010). Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B*, *114*(19), 6614–6624. https://doi.org/10.1021/jp911706q
126. Kuriata, A., Iglesias, V., Pujols, J., Kurcinski, M., Kmiecik, S., & Ventura, S. (2019). Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. *Nucleic Acids Res.*, *47*(W1), W300–W307. https://doi.org/10.1093/nar/gkz321
127. Van Durme, J., De Baets, G., Van Der Kant, R., Ramakers, M., Ganesan, A., Wilkinson, H., Gallardo, R., Rousseau, F., & Schymkowitz, J. (2016). Solubis: A webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel.*, *29*(8), 285–289. https://doi.org/10.1093/protein/gzw019
128. Suen, W.-C., Zhang, N., Xiao, L., Madison, V., & Zaks, A. (2004). Improved activity and thermostability of *Candida antarctica* lipase B by DNA family shuffling. *Protein Eng. Des. Sel.*, *17*(2), 133–140. https://doi.org/10.1093/protein/gzh017
129. Lawrence, M. S., Phillips, K. J., & Liu, D. R. (2007). Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.*, *129*(33), 10110–10112. https://doi.org/10.1021/ja071641y
130. Robinson, N. E. (2002). Protein deamidation. *Proc. Natl. Acad. Sci. U. S. A.*, *99*(8), 5283–5288. https://doi.org/10.1073/pnas.082102799
131. Mark, D. F., Lu, S. D., Creasey, A. A., Yamamoto, R., & Lin, L. S. (1984). Site-specific mutagenesis of the human fibroblast interferon gene. *Proc. Natl. Acad. Sci. U. S. A.*, *81*(18), 5662–5666.

132. Kontermann, R. E. (2009). Strategies to extend plasma half-lives of recombinant antibodies. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *23*(2), 93–109. https://doi.org/10.2165/00063030-200923020-00003
133. Strohl, W. R. (2015). Fusion proteins for half-life extension of biologics as a strategy to make biobetters. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.*, *29*(4), 215–239. https://doi.org/10.1007/s40259-015-0133-6
134. Cheang, J. Y., & Moyle, P. M. (2018). Glucagon-like peptide-1 (GLP-1)-based therapeutics: Current status and future opportunities beyond type 2 diabetes. *ChemMedChem*, *13*(7), 662–671. https://doi.org/10.1002/cmdc.201700781
135. Feldmann, M. (2002). Development of anti-TNF therapy for rheumatoid arthritis. *Nat. Rev. Immunol.*, *2*(5), 364–371. https://doi.org/10.1038/nri802
136. Nielsen, J. E., & Borchert, T. V. (2000). Protein engineering of bacterial α-amylases. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, *1543*(2), 253–274. https://doi.org/10.1016/s0167-4838(00)00240-5
137. Jones, B. J., Lim, H. Y., Huang, J., & Kazlauskas, R. J. (2017). Comparison of five protein engineering strategies for stabilizing an α/β-hydrolase. *Biochemistry*, *56*(50), 6521–6532. https://doi.org/10.1021/acs.biochem.7b00571
138. Thieker, D. F., Maguire, J. B., Kudlacek, S. T., Leaver-Fay, A., Lyskov, S., & Kuhlman, B. (2022). Stabilizing proteins, simplified: A Rosetta-based webtool for predicting favorable mutations. *Protein Sci.*, *31*(10). https://doi.org/10.1002/pro.4428
139. Banta, S., Dooley, K., & Shur, O. (2013). Replacing antibodies: Engineering new binding proteins. *Annu. Rev. Biomed. Eng.*, *15*(1), 93–113. https://doi.org/10.1146/annurev-bioeng-071812-152412
140. Rossi, A. M., & Taylor, C. W. (2011). Analysis of protein-ligand interactions by fluorescence polarization. *Nat Protoc*, *6*(3), 365–387. https://doi.org/10.1038/nprot.2011.305

141. Pollard, T. D. (2010). MBOC technical perspective: A guide to simple and informative binding assays. *Mol. Biol. Cell*, *21*(23), 4061–4067. https://doi.org/10.1091/mbc.E10-08-0683
142. Jeffrey, P. D., Strong, R. K., Sieker, L. C., Chang, C. Y. Y., Campbell, R. L., Petsko, G. A., Haber, E., Margolies, M. N., & Sheriff, S. (1993). 26-10 Fab-digoxin complex: Affinity and specificity due to surface complementarity. *Proc. Natl. Acad. Sci. U. S. A.*, *90*(21), 10310–10314. https://doi.org/10.1073/pnas.90.21.10310
143. Tian, W., Chen, C., Lei, X., Zhao, J., & Liang, J. (2018). CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.*, *46*(W1), W363–W367. https://doi.org/10.1093/nar/gky473
144. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, *255*(5041), 178–183. https://doi.org/10.1126/science.1553543
145. Jones, B. J., Evans, R. L., Mylrea, N. J., Chaudhury, D., Luo, C., Guan, B., Pierce, C. T., Gordon, W. R., Wilmot, C. M., & Kazlauskas, R. J. (2020). Larger active site in an ancestral hydroxynitrile lyase increases catalytically promiscuous esterase activity. *PLoS ONE*, *15*, e0235341. https://doi.org/10.1371/journal.pone.0235341
146. DeLorbe, J. E., Clements, J. H., Teresk, M. G., Benfield, A. P., Plake, H. R., Millspaugh, L. E., & Martin, S. F. (2009). Thermodynamic and structural effects of conformational constraints in protein-ligand interactions. Entropic paradoxy associated with ligand preorganization. *J. Am. Chem. Soc.*, *131*(46), 16758–16770. https://doi.org/10.1021/ja904698q
147. Barros, E. P., Schiffer, J. M., Vorobieva, A., Dou, J., Baker, D., & Amaro, R. E. (2019). Improving the efficiency of ligand-binding protein design with molecular dynamics simulations. *J. Chem. Theory Comput.*, *15*(10), 5703–5715. https://doi.org/10.1021/acs.jctc.9b00483

148. Thakkar, S., Nanaware-Kharade, N., Celikel, R., Peterson, E. C., & Varughese, K. I. (2014). Affinity improvement of a therapeutic antibody to methamphetamine and amphetamine through structure-based antibody engineering. *Sci. Rep.*, *4*, 3673. https://doi.org/10.1038/srep03673
149. Zhao, Q., Ahmed, M., Guo, H., Cheung, I. Y., & Cheung, N.-K. V. (2015). Alteration of electrostatic surface potential enhances affinity and tumor killing properties of anti-ganglioside GD2 monoclonal antibody hu3F8. *J. Biol. Chem.*, *290*(21), 13017–13027. https://doi.org/10.1074/jbc.M115.650903
150. Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, *25*(10), 1171–1176. https://doi.org/10.1038/nbt1336
151. Selzer, T., Albeck, S., & Schreiber, G. (2000). Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.*, *7*(7), 537–541. https://doi.org/10.1038/76744
152. Bugnon, M., Röhrig, U. F., Goullieux, M., Perez, M. A. S., Daina, A., Michielin, O., & Zoete, V. (2024). SwissDock 2024: Major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. *Nucleic Acids Res*, *52*(W1), W324–W332. https://doi.org/10.1093/nar/gkae300
153. Zoete, V., Schuepbach, T., Bovigny, C., Chaskar, P., Daina, A., Röhrig, U. F., & Michielin, O. (2016). Attracting cavities for docking. Replacing the rough energy landscape of the protein by a smooth attracting landscape. *J Comput Chem*, *37*(4), 437–447. https://doi.org/10.1002/jcc.24249
154. Röhrig, U. F., Goullieux, M., Bugnon, M., & Zoete, V. (2023). Attracting Cavities 2.0: Improving the flexibility and robustness for small-molecule docking. *J. Chem. Inf. Model.*, *63*(12), 3925–3940. https://doi.org/10.1021/acs.jcim.3c00054

155. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., & Baker, D. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, *501*(7466), 212–216. https://doi.org/10.1038/nature12443

156. Morin, A., Kaufmann, K. W., Fortenberry, C., Harp, J. M., Mizoue, L. S., & Meiler, J. (2011). Computational design of an endo-1,4-beta-xylanase ligand binding site. *Protein Eng. Des. Sel.*, *24*(6), 503–516. https://doi.org/10.1093/protein/gzr006

Supporting Information

Code Block S4.1. Python script to fit data for an equilibrium dialysis experiment (Fig Figure 6.5)

```
#! /usr/bin/env python
# coding: utf-8

# also run the command below in terminal
# sudo chmod +x file-name.py

import numpy as np                #import math functions to use arrays
from scipy import optimize        #import non-linear fit function
import matplotlib.pyplot as plt   #import plotting function

# enter your data here
freeAg = np.array([0.02, 0.05, 0.10, 0.2, 0.4, 0.6, 1])
Y = np.array([0.20, 0.37, 0.57, 0.74, 0.84, 0.93, 0.98])
Kd = 0.1 # initial guess for Kd

# define equation to be fit
def satCurve(freeAg, Kd):
```

```
    return freeAg/(freeAg + Kd)

# fit the data to the equation using non-linear least squares with
# default fit settings
popt, pcov = optimize.curve_fit(satCurve, freeAg, Y, Kd)
print("Kd =", "{0:.3f}".format(popt[0]))

# plot data and best fit curve
plt.scatter(freeAg, Y)
xfit = np.linspace(0,1)
plt.plot(xfit, satCurve(xfit, popt[0]), 'r-')
plt.xlim(0, 1)
plt.xlabel('free antigen concentration')
plt.ylabel('fraction of antibody containing bound antigen')
plt.show()
```

Problems

1. The equilibrium dissociation constant, K_d, of an antibody and its ligand (a peptide) is 5 μM at pH 5.0 and 25 °C.

    a) At what concentration of the ligand is half of the protein bound?

    b) What fraction of the protein is bound at ligand concentration of 1.25 μM?

    c) When the pH was raised to 6.5, the K_d increased to 20 μM. Is the binding tighter or weaker at this pH compared to pH 5.0? Explain why.

    d) What functional groups/residues are most likely responsible for this change in the binding affinity with pH?

The problems below use the experimental data from a paper reporting the design of a protein (DIG10.3) that tightly binds digoxigenin[155] to test the ability of docking calculations, SwissDock [¶] to predict the affinity of a protein for a ligand.

2. *Prepare the wt protein and Tyr34Phe variant for docking.* The x-ray structure of DIG10.3 (pdb id = 4J9A) contains 9 monomers in the unit cell. For simplicity, we will delete eight of these leaving only chain A. The structure also contains bound digoxigenin, which we will remove so that SwissDock can predict the binding position. Finally, we will save the modified file. Download this structure from the Protein Data Bank. Open the file with a text editor and read the notes at the beginning. In particular, note whether any amino acids are missing from the structure of chain A. Next, open the file using PyMOL. Enter the following commands on the input line (ignore comments after #).

```
remove (not chain A) # deletes all chains except chain A
remove (resn DOG) # deletes residue named 'DOG'
save DIG103_A.pdb # saves your modified structure as a pdb file
```

The file is saved in your working directory, usually your home directory; mine is '/Users/romas'. Search to find the file, if needed. Next, create the Tyr34Phe variant using the instructions in Homework 1 (Wizard => Mutagenesis => Protein ...). Use the suggested default rotamer of the Phe replacement. The following commands may be helpful.

```
show sticks, resi 34 # shows Tyr34 so that you can select it
save DIG103_A_Y34F.pdb # saves file once you have made the substitution.
```

[¶] Bugnon *et al.* (2024) SwissDock 2024: Major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. *Nucleic Acids Res.* 52, W324-W332. https://doi.org/10.1093/nar/gkae300

Additional troubleshooting: If SwissDock fails to verify your file of the mutated protein, check that your pdb file contains only the mutated protein. One student's file contained three proteins, not one. Open your file in a text editor. Check to see that a line that reads 'TER' occurs only once in your file and that it is the penultimate line. If you find 'TER' more than once, you must delete both the extra 'TER' and the atoms associated with that extra protein. Your file should contain 905 atoms and the last four lines of your file should read:

ATOM 904 O ARG A 123 0.634 -10.220 7.065 1.00 48.12 O

ATOM 905 CB ARG A 123 3.588 -8.274 5.874 1.00 50.26 C

TER

END

3. *Docking with default settings.* Submit the calculation to dock digoxigenin to the wild-type protein and to the Tyr34Phe variant at https://www.swissdock.ch/. Leave the default 'Docking with Attracting Cavities' selected. In section 1 - Submit a ligand, click on advanced search and enter 'digoxigenin', then click on the digoxigenin structure and click apply. This brings you back to the starting page. Click on 'Prepare ligand'. You should get a green check mark.

In section 2 - Submit a target, click on choose file, then select the DIG103_A.pdb file that you created above to upload it. This brings you back to the starting page. Click on 'Prepare target'. You should get a green check mark.

In sections 3 & 4, no changes are needed. Click on 'Check parameters'. You should get a green check mark and an estimate of how long the calculation will take, about 10 min. It usually takes a bit longer than this estimate; be patient.

In section 5 - Start docking, enter a docking name like DOG_DIG103_A, then click on 'START DOCKING'.

The result of the docking calculation is a set of ten clusters where each cluster contains a set of similar binding orientations. The tightest binding cluster is cluster 0 and usually contains eight binding orientations. We will compare the best binding orientations (most negative AC score in cluster 0) and assume that it is in kcal/mol.

a) *Predicted ΔG_d, $\Delta\Delta G_d$, and K_d.* Compare the AC score for the best binding orientation in cluster 0 of digoxigenin to the wt protein and to the Tyr34Phe variant. Assume the score corresponds the Gibbs energy of binding of the ligand to the target in kcal/mol. Calculate the predicted equilibrium dissociation constants, K_d. Calculate the predicted difference in binding affinity of the two proteins. Assume the temperature is 25 °C. Recall that favorable binding corresponds to a positive dissociation energy. Remember to compare *differences* in Gibbs energies, but *ratios* of dissociation constants. SwissDock yields negative numbers, which correspond to association energies. Reverse the sign on these values to convert them to dissociation energies.

b) *Compare predicted values to experimental values.* The experimental inhibition constants for DIG binding to DIG10.3 is 0.65±0.03 nM and to DIG10.3-Tyr34Phe is 59±6 nM (Figure 4, panels b & d)[155] The experiments involved displacement of DIG-PEG_3-Alexa488 from the protein with DIG, so they are called inhibition constants, not binding constants. Assume these inhibition constants are accurate measures of K_d. Calculate the same values as in 1a using these experimental values and comment on any differences.

c) Compare the predicted and experimental difference in the affinities of the two proteins for DIG to the expected contribution of one hydrogen bond to the binding. Discuss the lack of agreement between the calculated and experimental values. Suggest two reasons

that could contribute to the lack of agreement.

4. *Docking with increased sampling of conformations.* Repeat the calculation in question 1, but increase the number of random initial conditions (RIC) from the default of 1 to 3.

a) Compare the predicted values here with those in question 1. Rationalize any differences.

b) Compare the predicted values here with the experimental values. Rationalize any differences.

c) Discuss whether this docking web tool could be used in place of the Rosetta tool used by Tinberg *et al.*[155] to design a protein to that binds digoxigenin.

Answers

1. a) 5 μM b) using equation 6.5, Y = 1.25/6.25 = 20% c) weaker because the K_d has increased. d) A group with a pK_a in the range of 5-6.5; aspartate, glutamate or histidine side chains are possible. Since the pKa of functional groups changes when embedded into a protein environment, other groups may also be responsible and one could make a better estimate if the structure of the protein were available.
2. a) WT AC scores: -13.96097, -13.91753, -12.76379, -12.47792, -7.96029, -5.14183, 13.89088, 14.30629; AC scores: -20.42112, -16.18831, -14.83161, -14.22944, -14.00921, -13.65794, -13.34280, -12.30052 b) experimental: wt ($K_d = 0.653 * 10^{-9} M$); Tyr34Phe variant $≈100 - foldlesstightly(K_d = 59 * 10^{-9} M$). A factor of 100 corresponds to 2.8 kcal/mol at 298 K. A hydrogen bond is expected to contribute 0-2 kcal/mol to binding (p 25 of text), which corresponds to a factor of 1-30, so

a factor of 100 suggests that interactions beyond a hydrogen bond are involved. SwissDock calculations yield *association constants*, that is the inverse a dissociation constant, so use +14 kcal/mol to calculate $K_d = 5 * 10^{-11} M$, which is about 20-fold lower than the experimental value for wt. The value for Tyr34Phe, is $K_d = 10^{-15} M$, which is off by a factor of 50 million. SwissDock using default settings incorrectly predicts that the Tyr34Phe variant binds more 50,000-fold more tightly wt.

c) The lack of agreement may be a sampling failure due to insufficient search of conformations such as those that may require movement of the protein. The lack of agreement could be due to the inability of the force field used by SwissDock to accurately calculate the binding energy (a scoring failure).

4. a) increase the number of random initial conditions (RIC) from the default of 1 to 3 WT AC scores: -20.83291, -16.41948, -16.21564, -14.90450, -14.35758, -14.31006, -14.29569, -14.01725; Y34F AC scores: -16.303044, -14.243850, -14.050263, -13.886717, -13.784884, -13.460208, -12.810371, -12.024448

b) Now the ranking correctly predict that wt binds more tightly than Tyr34Phe, but numbers differ significantly from the experimental values. The SwissDock-predicted $K_d^{wt} = 5.5 * 10^{-16}$, which is a million-fold too high; SwissDock $K_d^{Y34F} = 1.1 * 10^{-12}$, which is 50,000-fold too high. The ratio of K_d's overestimates the experimental ratio by about a factor of 20. This difference could be due to sampling (SwissDock did not find a representative sample of the true binding orientations), perhaps because even more of the protein must move than the 5 Å permitted here or due to a scoring error, where the Gibbs energy of the interaction is not calculated correctly.

c) Rosetta predicted an interaction energy of $≈$15.5 REU for DIG10 and the ligand (Figure 1, panel b). If we assume that REU's (Rosetta energy units) correspond to kcal/mol, then this prediction corresponds to a K_d of $0.004 * 10^{-9}$, which is 3 million-fold stronger binding than the experimental value of $12,000 * 10^{-9}$. Thus, neither Rosetta, nor SwissDock accurately predict the binding constant. SwissDock did rank the wt and Tyr34Phe variant correctly; the Tinberg *et al.* paper did not calculate the relative binding strengths, they just measured them. Rosetta calculated two things that SwissDock did not: average H-bonding Boltzmann weight and the protein stability. Both of these are characteristics of the protein: is it preorganized for binding and is it stable. Both of these are important for design and missing from SwissDock. SwissDock assumes that the given structure is correct, while Rosetta predicts the structure. SwissDock cannot be used for protein design, but it may be helpful to predict small modifications.