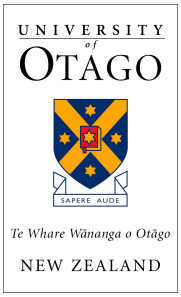
Computational tools for improving recombinant protein production



Bikash Kumar Bhandari

A thesis submitted for the degree of
Doctor of Philosophy
at the University of Otago, Dunedin,
New Zealand

3 May 2021

Blank page

Acknowledgements

I would like to thank my supervisor Associate Professor Paul P. Gardner and co-supervisor Dr Chun Shen Lim for their guidance, feedback, ideas, time, encouragement and many chats.

I am grateful to the committee members Associate Professor Chris Brown and Dr Alex Gavryushkin for providing many feedbacks along the way.

Special thanks to my family, Mammam and her kittens, and many others for their love and support.

Abstract

Recombinant protein production is a cornerstone of modern biotechnology and has been utilised to produce many proteins of scientific and commercial interest. The optimality of result is dependent on the balances among the involved intricate stochastic processes. In particular, two of the critical processes are protein expression and solubility. Collectively, the failures at these two steps drop down the success rate of protein production to around 25%. Furthermore, toxicity of recombinant proteins may also significantly reduce the amount of protein produced. Therefore, prediction and optimisation of expression, solubility and an early detection of these toxic proteins could save resources and assist in better planning of the experiment.

In this work, we show that mRNA accessibility, measured through the opening energy, and protein structural flexibility, measured by using the normalised B-factors, can describe protein expression and solubility respectively with a higher accuracy than other features. We also develop a new and more accurate protein solubility predicting metric called the Solubility-Weighted Index (SWI). Using these findings, we develop a gene expression prediction and optimisation tool: Translation Initiation coding region designer (TIsigner), available at `https://tisigner.com/tisigner` and protein solubility prediction and optimisation tool: Soluble Domain of Protein Expression (SoDoPE), available at `https://tisigner.com/sodope`. We also developed a third tool, Razor `https://tisigner.com/razor`, for the detection of toxins. To assist in maximising protein production, we also develop a pipeline for optimising protein expression, solubility and toxin detection by integrating these three tools.

Research articles associated with this thesis

1. **Bikash K Bhandari**[†], Chun Shen Lim[†], Daniela M. Remus, Augustine Chen, Craig van Dolleweerd, Paul P. Gardner, Analysis of 11,430 recombinant protein production experiments reveals that protein yield is tunable by synonymous codon changes of translation initiation sites, *PLoS Comput. Biol.*, 2021; doi: https://doi.org/10.1371/journal.pcbi.1009461.

2. **Bikash K Bhandari**[†], Chun Shen Lim[†], Paul P Gardner, TISIGNER.com: web services for improving recombinant protein production, *Nucleic Acids Research*, 2021, https://doi.org/10.1093/nar/gkab175.

3. **Bikash K Bhandari**, Paul P Gardner, Chun Shen Lim, Solubility-Weighted Index: fast and accurate prediction of protein solubility, *Bioinformatics*, 2020, https://doi.org/10
.1093/bioinformatics/btaa578.

4. **Bikash K Bhandari**, Paul P Gardner, Chun Shen Lim, Razor: annotation of signal peptides from toxins, *bioRxiv*, 2020; doi: https://doi.org/
10.1101/2020.11.30.405613.

5. **Bikash K Bhandari**, Paul P Gardner, Chun Shen Lim, Fast and accurate annotation of fungal signal peptides in metagenomic data, in preparation.

6. Chun Shen Lim[†], **Bikash K Bhandari**[†], Paul P Gardner, Capsicum: classification of capsid proteins from diverse viruses, in preparation.

[†] These authors contributed equally to this work.

Awards associated with this thesis

1. Student Paper Award, 2020. Paper of the year - Runner up. Department of Biochemistry, University of Otago [19].

Abbreviations

GFP	Green Fluorescent Protein
GRAVY	GRand AVerage of hydropathY
LB	Lysogeny Broth
MIDAS	Modular Idempotent DNA Assembly System
OD	Optical Density
PSI:Biology	Protein Structure Initiative:Biology
SoDoPE	Soluble Domain for Protein Expression
SP	Signal Peptide
SWI	Solubility-Weighted Index
Tisigner	Translation Initiation coding region desigNER

List of Tables

List of Figures

Table of Contents

Chapter 1

Introduction

After the introduction of recombinant protein production in 1977 [114], a large number of monoclonal antibodies, hormones, enzymes and other proteins of pharmaceutical, industrial and scientific importance are being synthesised using microbes such as bacteria and yeast, insect cells and mammalian cells. Consequently, recombinant proteins currently has a market value in billions of dollars, making it one of the highly valued technologies [255, 188].

There has been much research to improve the recombinant protein production technology. In particular, the development of the pET vector system in 1991 has revolutionised the use of *Escherichia coli* for protein production [65]. *E. coli* is often the host of choice because it is relatively inexpensive, and has a faster growth rate than other expression hosts [198, 58]. Now there are a multitude of optimised vectors such as pGEX, pMAL and pET as well as a number of engineered strains of *E. coli* such as BL21 and BL21(D3). Several guidelines and practices have also been proposed to maximise the chances of successful experiments [13, 198]. Furthermore, several high-throughput methods have made the process scalable [224, 31, 117]. These new advancements has made the process of recombinant protein production much easier and economical.

1.1 Recombinant protein production

The initial step in recombinant protein production is a successful protein expression. The expressed protein then needs to be soluble for use in many structural, functional

and pharmaceutical studies where concentrated protein samples are desired [132, 107]. Despite almost 40 years of refinements in protocols and technology, around half of the recombinant protein expression experiments fail at the expression stage and nearly half of expressed protein are insoluble [105] (Figure 1.1). This makes the protein production process more challenging. Predicting protein expression and solubility can help plan the experiment and save time and resources. Furthermore, using a highly optimised target gene can increase the success rate of recombinant protein production.

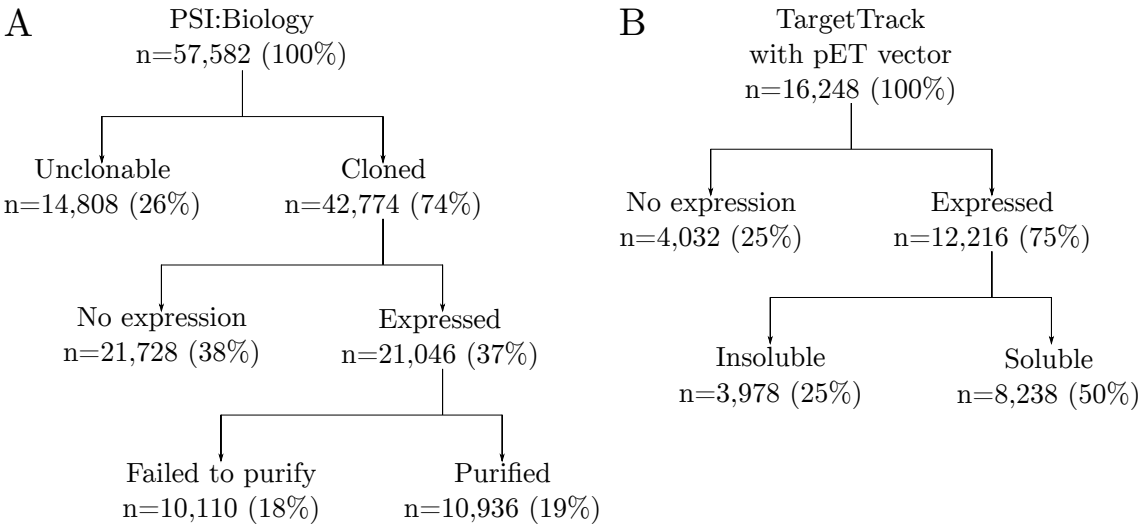


Figure 1.1: **The success rate of recombinant protein production is around a quarter. (A)** All experiments, using different vectors and hosts, preformed for deposition to the TargetTrack database shows around 19% of experiments are purified. Data taken from Protein Structural Initiative (PSI:Biology) metrics. **(B)** A subset of experiments from TargetTrack database using pET vector and *E. coli* as expression host, shows 50% of experiments produce soluble proteins.

In the following sub-sections, we will discuss these two steps—protein expression and solubility in details. Unless otherwise stated, these discussions will refer to prokaryotes, in particular, *E. coli* based systems.

1.1.1 Protein expression

Protein expression is the process by which protein is synthesised using the information in the messenger RNA (mRNA) (Figure 1.2).

Intuitively, we expect the protein yield to be predictable using mRNA levels, the cor-

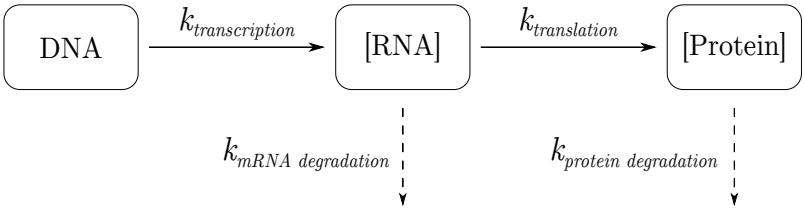


Figure 1.2: **Protein expression depends on the rates of RNA and protein synthesis and their degradation.** Solid arrow represents synthesis whereas dashed arrow represents degradation. [RNA], concentration of mRNA; [Protein], concentration of protein; $k_{transcription}$, rate of transcription; $k_{translation}$, rate of translation; $k_{mRNA\ degradation}$, rate of mRNA degradation; $k_{protein\ degradation}$, rate of protein degradation. Figure redrawn from Abreu *et al.* (2009).

relation is lower than expected [223, 231, 17]. This reflects the complexities of the underlying process. The amount of protein produced is determined by translation rate and protein degradation (Figure 1.2). This dynamic system can be mathematically described by a first order differential equation [223] whose solution at equilibrium gives the following relationship relationship between protein concentration (P_∞), mRNA concentration (R_∞), translation rate ($k_{translation}$) and protein degradation rate ($k_{protein\ degradation}$) :

$$\frac{P_\infty}{R_\infty} = \frac{k_{translation}}{k_{protein\ degradation}} \tag{1.1}$$

There are various mechanisms regulating translation and protein degradation, so the correlation between (P_∞) and (R_∞) is not perfect. Squared Pearson's correlation is around 0.4 for many organisms [223]. Assuming the protein degradation rate ($k_{protein\ degradation}$) to be a constant, the ratio of P/R depends upon the translation rate ($k_{translation}$) only. Hence, the amount of protein can be modulated by tuning $k_{translation}$.

1.1.2 Translation

Translation in prokaryotes is *initiated* when the small ribosome subunit $30S$ binds to the Shine-Dalgarno sequence and moves upto the start codon. The currently accepted model is that the initiator tRNA charged with N-formylmethionine, initiation factor (IF-2) and guanosine triphosphate (GTP) binds with this subunit followed by

the release of IF-3 [89]. This complex is called the 30S initiation complex. IF-1 and IF-2 are released and the large ribosomal subunit $50S$ now binds followed by the hydrolysis of GTP, to form a $70S$ initiation complex.

After the formation of initiation complex, the decoding of information in the mRNA begins. This is called *translation elongation*. The $70S$ complex consists of three active sites: peptidyl-tRNA site (P site), aminoacyl-tRNA site (A site) and exit site (E site). Initially, the start codon and the successive codon of mRNA is positioned at P site and A site respectively. The initiator tRNA charged with N-formylmethionine is coupled with start codon at the P site and the corresponding aminoacyl-tRNA is coupled to the next codon at the A site through codon-anticodon pairing (Figure 1.3). Peptide bond is formed between the amino acid carried by tRNA at P and A site to give a dipeptide. tRNA at P site is translocated to E site, tRNA at A site to P site and the ribosome moves along the mRNA. The deactylated tRNA at E site is released. Since the A site is now empty, it receives the next aminoacyl-tRNA and the process continues adding an amino acid to C terminal of the dipeptide.

Once the ribosome encounters a stop codon, release factor (RF1 or RF2 and RF3) bind to the ribosome resulting in a tranfer of polypeptide to a water molecule rather than aminoacyl-tRNA. The free polypeptide is released from the ribosome and $70S$ ribosome disassociates into $30S$ and $50S$ subunits. This step is called *translation termination*.

1.1.3 mRNA features and their roles in translation

Translation rate may depend on the rate formation of translation initiation complex and utilisation of available tRNA pool. Several features of a mRNA sequence are suggested to explain these two major dependencies of translation rate. However, many mRNA features are not independent, making it hard to distinguish the impacts of individual features [156]. The main features that have been considered to date, can be classified into three categories: codon preferences, mRNA folding (secondary structure) and mRNA:ncRNA avoidance. These three categories of features is the

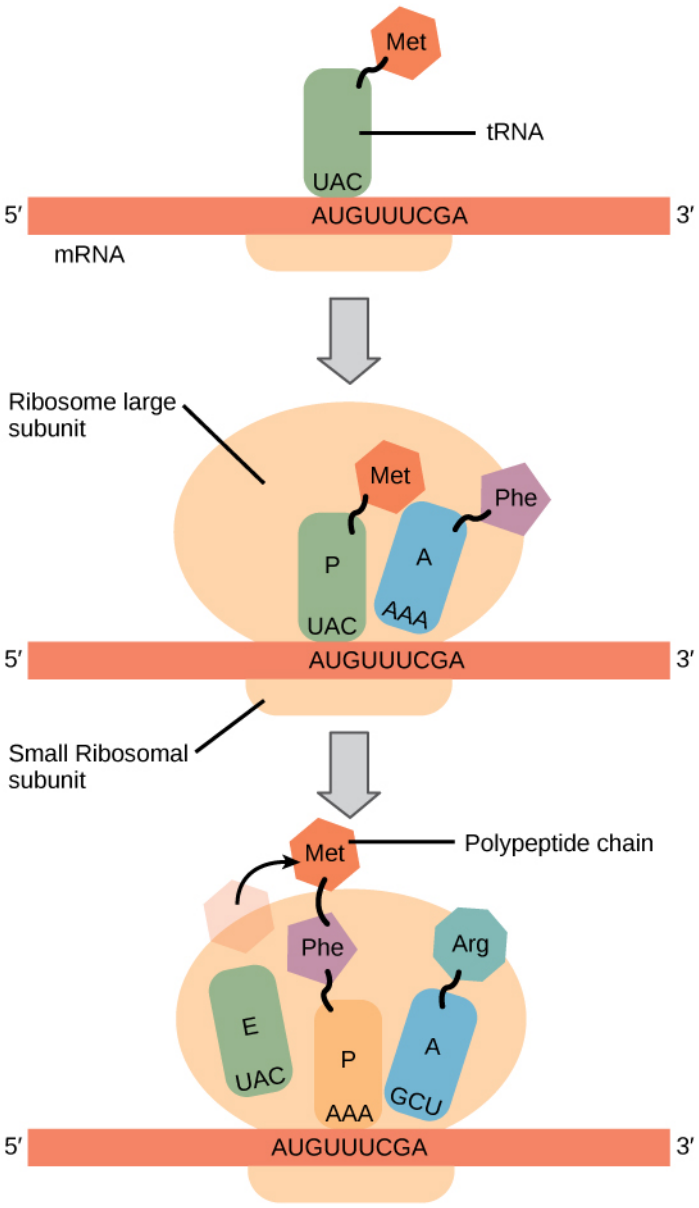


Figure 1.3: **Prokaryotic translation.** Initiator tRNA binds to the smaller subunit at the start codon. Larger subunit joins to form a translation initiation complex such that initiator tRNA is at P site and the next aminoacyl-tRNA is at A site. Initiator tRNA moves to the E site, dipeptide is formed at the P site and new aminoacyl-tRNA is received at the A site. Figure from OpenStax College, Concepts of Biology. OpenStax CNX (Creative Commons license: CC BY 4.0)

basis of our understanding and optimisation of protein production. Hence, we will describe them in more details below.

Features based on codon analysis

This category measures the bias in codon usage relative to endogenous mRNAs. Higher values of these indices is an indicator that the given mRNA sequence follows the codon usage pattern of the host. Features under this category are the codon adaptation index (CAI) [214], tRNA adaptation index (tAI) [196, 202] and related metrics such as codon pair usage [87]. For example: the codon adaptation index (CAI) for a given protein is the harmonic mean of the relative adaptiveness w [214] of the codons:

$$CAI_g = (\prod_{i=1}^{N} w_i)^{1/N} \tag{1.2}$$

where w_i is the relative adaptiveness of the i^{th} codon which is the ratio of observed frequency of the codon f_i upon consideration to the frequency of the most frequent synonymous codon.

$$w_i = \frac{f_i}{max(f_i)}$$

Based upon the idea of CAI, tAI was developed to measure the translational efficiency by taking into account of tRNA concentration and codon-anticodon coupling efficiency. We first define the absolute adaptiveness W_i of codon i as:

$$W_i = \sum_{j=1}^{n_i}(1 - s_{ij})tGCN_{ij} \tag{1.3}$$

where n_i is the number of anticodons pairing with codon i, $tGCN_{ij}$ is the copy number of the j^{th} tRNA that recognizes the i^{th} codon. $tGCN_{ij}$ is correlated with the tRNA concentration [123, 176]. s_{ij} is a constraint on the codon-anticodon pairing and has values between 0 (more efficient pairing) and 1 (less efficient pairing). The relative adaptiveness w_i of the i^{th} codon is W_i normalised by maximum of W_i among all codons. If W_i is zero, then the relative adaptiveness is the mean of all w_i. Once w_i are found, the tAI is the harmonic mean as in Equation 1.2.

Both CAI and tAI measures are equivalent to a zeroth order Markov model whereas codon pair usage or di-codon frequency is essentially a first order Markov model. It is thought that a higher value of these indices means that the sequence can utilise the available tRNA pool more efficiently which causes an increase in efficiency of translation [112, 87, 214, 196, 202, 34]. However, this proposition has been challenged and studies suggest that mRNA secondary structure might be more important in explaining translation efficiency. [133, 27, 35].

Secondary structure

If the region around a translation initiation site forms a strong secondary structure, this leads to disruption of the formation of initiation complex, which inhibits translation (Figure 1.4) [133, 69, 240]. Recent studies show that the RNA structure stability of this region explains variation in protein expression better then codon usage [133, 185, 35] indicating that translation initiation is a rate limiting step for translation. Furthermore, secondary structure has been shown to change the functional half-life of mRNA and thus further influence protein expression [156]. Minimum free energy (MFE) of mRNA is widely used to measure the strength and stability of secondary structure. A way to find the MFE is to enumerate all possible structures of a given mRNA and then find the minimum. However, this is impractical because combinatorial explosion occurs quickly as the length of mRNA increases. Clever dynamic programming algorithms has made this problem tractable which is described below.

Computation of minimum free energy (MFE) and suboptimal structures

Thermodynamically, an RNA structure with the lowest Gibb's free energy is the most stable. This energy is called minimum free energy. Thermodynamic ensembles usually follow the principle of locality, which says that interactions occur only between neighbouring particles. For example: In the Ising model of magnetism, spin interactions happen only between the nearest particles. Similarly, MFE is also calculated by using so called 'Turner parameters' [242] in a nearest neighbouring model and a set of recursive equations called Zuker's algorithm [280]. However, the accuracy of the computed MFE is only within $5 - 10\%$ and a large number of alternate RNA

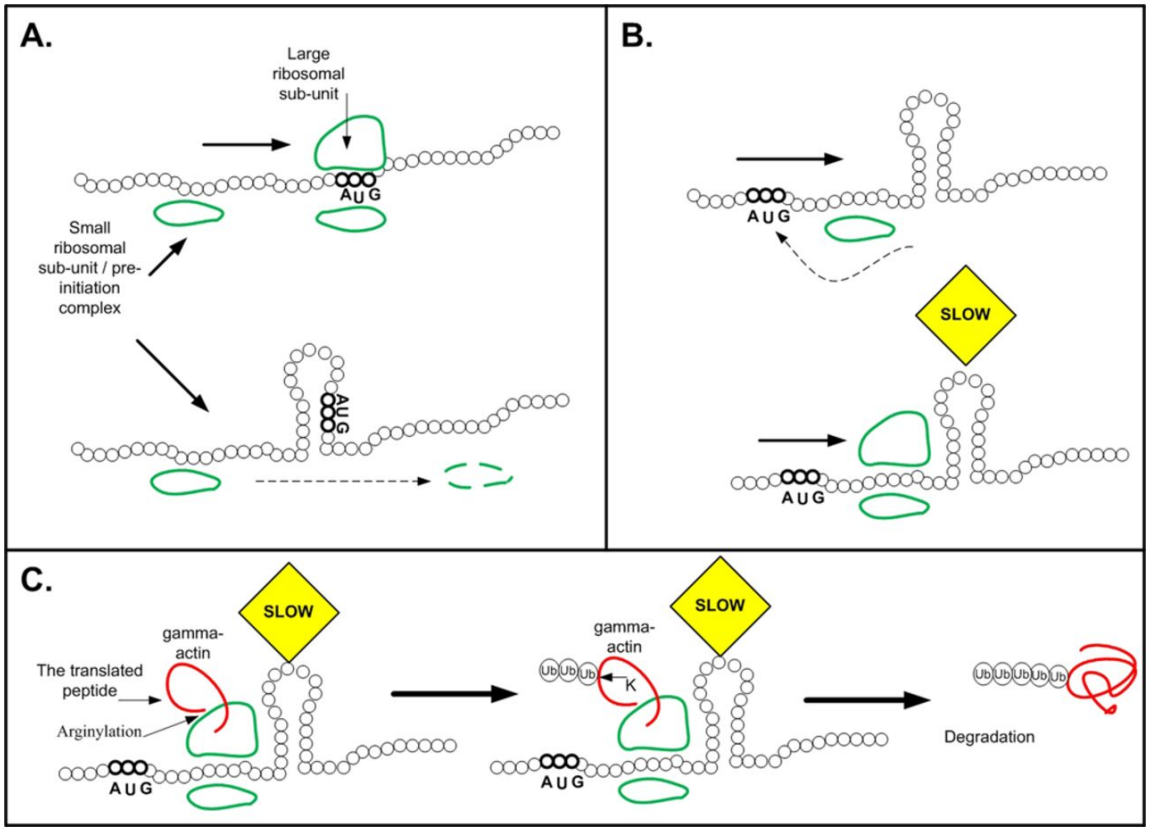


Figure 1.4: **Secondary structure at the translation initiation site inhibits translation. (A)** The start codon AUG is recognised by the pre-initiation complex if the secondary structure is weak (top) but fails to do so in presence of a strong structure (bottom). **(B)** Presence of strong structure downstream of translation initiation site prevents the movement of pre-initiation complex (top) which could improve translation efficiency by improving ribosomal allocation (bottom). **(C)** Such downstream structures could influence post-translation modification by slowing down the ribosome. For example: lysin residues in arginylated gamma actin are exposed due to slow translation and undergo ubiquitination. Figure from Tuller and Zur, (2011) (Creative Commons license: CC BY)

structures lie within $5 - 10\%$ of the predicted global minimum [67]. This prompts us to calculate some other free energies for different suboptimal conformations that RNA may achieve in a near thermodynamic equilibrium. Furthermore, depending on the criteria to pick an *optimal* sturcture from a Boltzmann's ensemble, suboptimal structures may not lie near the thermodynamic equilibrium at all. For example: Ding *et al.* [61] found that structures tend to form clusters in a Boltzmann's ensemble. Instead of free energy, if we use these clusters as a criteria for sampling, then structure that has a minimum distance from all clusters is the optimial structure [148]. This structure is also known as centroid structure and other structures around the centroid may as well be regarded as suboptimal.

There were some early attempts to compute suboptimal structures for example, by Zuker [279] and Waterman *et al.* [263]. However, the backtracing procedure in their algorithms was not efficient enough to compute energies for longer RNA molecules. This problem was solved by McCaskill [158] by proposing an efficient algorithm to compute energy through a partition function which has the same time complexity as Zuker's algorithm for MFE ($\mathcal{O}(n^3)$).

Partition function and base pairing probabilities Consider a structure s of an RNA molecule with free energy $E(s)$. In a thermodynamic ensemble of different structures at equilibrium, using the principle of maximum entropy, we see that the probability that the given RNA has the structure s follows a Boltzmann distribution:

$$p(s) \propto e^{-\beta E(s)} \tag{1.4}$$

where β is called 'thermodynamic beta' and equals $1/k_B T$, where k_B is the Boltzmann's constant and T is the absolute temperature (Subsection 1.3.1 Simulated annealing). Since sum of probabilities over set of all structures Ξ, must be equal to unity, we have:

$$\sum_s p(s) = \frac{1}{Z} \sum_{s \in \Xi} e^{-\beta E(s)} = 1$$
$$Z = \sum_{s \in \Xi} e^{-\beta E(s)} \tag{1.5}$$

The quantity Z, which plays a role of normalisation of probabilities, is called the *canonical partition function*. Many thermodynamic parameters of interest can be derived from Z, for example, free energy G of RNA in terms of Z is given by:

$$G = -\frac{1}{\beta} \ln(Z) \tag{1.6}$$

The efficient dynamic programming to enumerate Z was proposed by McCaskill [158] with time complexity of $\mathcal{O}(n^3)$. This method is essentially a recursive decomposition of Z similar to Zuker's algorithm only difference being the addition in Zuker's relation are now substituted by product because free energies are additive. If E is the total free energy and E_L are the energy contributions from various types of loops (hairpin, stacked pair, bulges, interior loops, multiloops) in a structure, then:

$$E = \sum_L E_L \tag{1.7}$$

If we suppose the term Q_{ij}^b accounts for all loops L enclosed by i, j, we see that additivity of free energy (Equation 1.7) implies a multiplicative contribution to the partition function (1.5) which gives the following recursive equation :

$$Q_{ij}^b = \sum_L e^{-\beta E_L} \prod_{i<h<k<j} Q_{hk}^b \tag{1.8}$$

Using this restricted partition function term for loop contributions, the total partition function between i^{th} and j^{th} nucleotides (Q_{ij}) can now be written as:

$$Q_{ij} = Q_{ij-1} + \sum_{i \leq k < j} Q_{ik-1} Q_{kj}^b \tag{1.9}$$

The full partition function of RNA with N nucleotides is given by $Z = Q_{1N}$. Equations 1.8 and 1.9 are McCaskill's recursions for partition function. Once Z is known, the probability of any structure s with free energy $E(s)$ is given by:

$$p(s) = \frac{1}{Z} e^{-\beta E(s)} \tag{1.10}$$

The computational approach outlined above is very generic and can be used for other specific cases. For example: if we want to know the probability that $[i^{th}, j^{th}]$ nucleotides are paired, then we can modify Equation 1.5, where partition function is found by simply summing Boltzmann's factor over all structures ζ where $[i^{th}, j^{th}]$ nucleotides are paired ($\zeta \subseteq \Xi$, where Ξ is the ensemble of structures):

$$Z_p = \sum_{s \in \zeta} e^{-\beta E(s)} \tag{1.11}$$

The base pairing probabilities are, given by equation 1.10 with an appropriately computed Z.

mRNA:ncRNA avoidance

Recently, Umu et al [243] found that in bacteria and the archaea, the strength of interactions between mRNAs and non-coding RNAs (ncRNAs) anti-correlate with protein levels. These signals are particularly obvious at the translation initiation site, suggesting that there is an avoidance of inappropriate interactions for highly expressed proteins. However, from a large pool of mRNA and ncRNA, a complete avoidance of interactions is unlikely and a trade off exists between interactions and protein expression. Further, it is suggested that compartmentalisation should minimise these cross talk interactions in eukaryotes. Compartmentalisation has been a topic of considerable research and is linked with noise filtering in gene expression and cellular feedback process [194, 227, 11, 63]. We now outline in brief, the necessary background to understand the computation of RNA interactions and mRNA:ncRNA avoidance.

Unpairing of bases and accessibility McCaskill's equations 1.8 and 1.9 for the partition function can also be 'inverted' to find the probability that $[i^{th}, j^{th}]$ nucleotides are unpaired in the given ensemble. The energy required to unpair the nucleotides is called accessibility or opening energy. If $\kappa \subseteq \Xi$ is the set of all structures s where $[i^{th}, j^{th}]$ nucleotides are unpaired, then the accessibility is given by :

$$\begin{aligned} E_{accessibility} &= E_{s\in\kappa} - E_{s\in\Xi} \\ E_{accessibility} &= -\frac{1}{\beta}\ln\frac{Z_{unpaired}}{Z} \end{aligned} \tag{1.12}$$

The term $\frac{Z_{unpaired}}{Z} = p_u$ is the probability that $[i^{th}, j^{th}]$ nucleotides are unpaired [15]. Since the pair i, j may or may not be enclosed by a base pair $k, l \ni \forall\ k, l$: $k < i < j < l$ (Figure 1.5), this probability can be computed by McCaskill approach [15]. Accessibility prediction forms the basis of RNA:RNA interactions.

RNA:RNA interactions For two RNA molecules to interact and pair, most computational tools assume that this is a two step process—unfolding of the RNA molecules at the target sites, followed by an actual interaction (hybridisation) [163]. Thus, the total binding energy ΔG is the sum of the accessibility of the target site of the longer RNA molecule $\Delta G_{unpaired}$ and the subsequent interaction between the unfolded region of the interacting molecules ΔG_{int}. $\Delta G_{unpaired}$ is computed by equation 1.8, where as ΔG_{int} is computed through equation 1.6 by replacing Z with Z_{int}. For an interaction between nucleotides $[i, j]$ and $[i^*, j*]$ the partition function Z_{int} is given by [163] (Figure 1.6):

$$\begin{aligned} Z_{int} &= p_u[i,j]\sum_{i^*>j*} Z^I[i,j,i^*,j^*] \\ &where, \\ Z^I[i,j,i^*,j^*] &= \sum_{\substack{i<k<j \\ j*<k^*<i^*}} Z^1[i,k,i^*,k^*]e^{E_I(i,k,i^*,k^*)} \end{aligned} \tag{1.13}$$

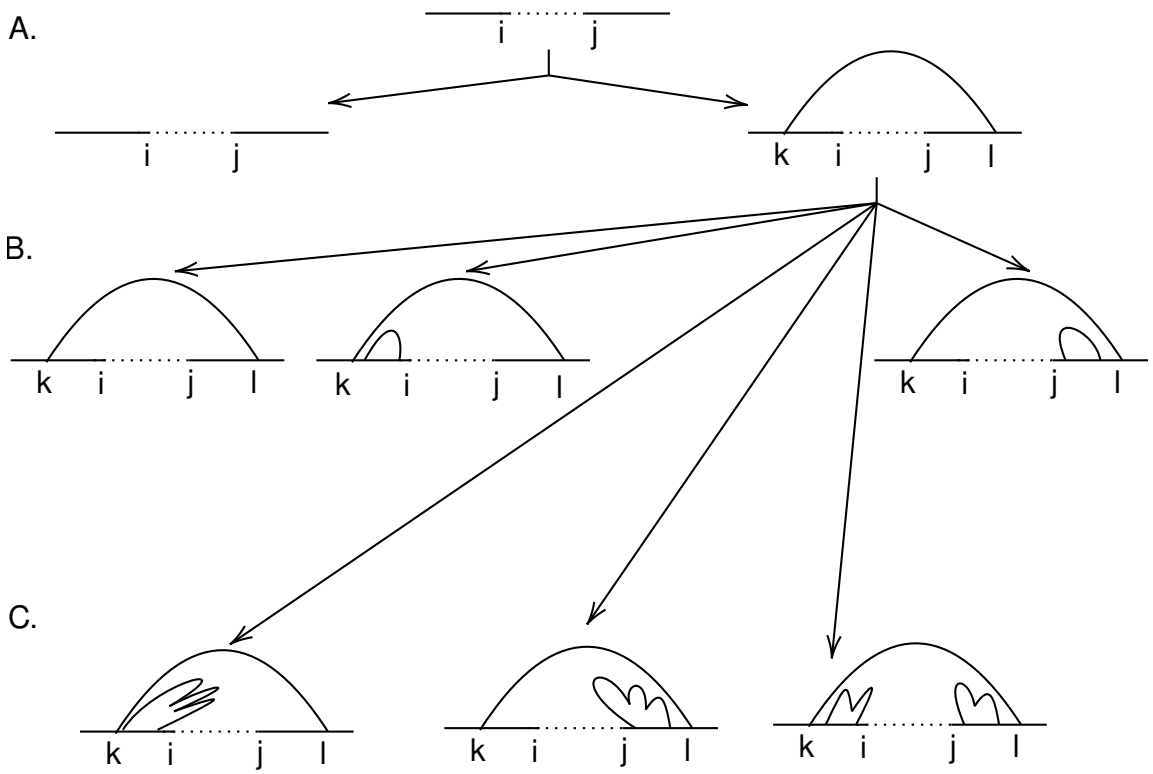


Figure 1.5: **Decomposition of partition function to calculate unpairing probability of the region $[i, j]$ in a nucleotide sequence. (A)** The interval [i,j] may either be enclosed by pair [k,l] (right) or may be open (left). **(B)** The enclosed pair may or may not contain an hairpin or **(C)** multiloops. The partition function $Z_{unpaired}$ is a sum of all these contributions. Figure adapted from Bernhart *et al.*, (2011)

with $E_I(i, k, i^*, k^*)$ as the free energy of the interior loop enclosed by (i, k) and (i^*, k^*) and $Z^1[i, k, i^*, k^*]$ contains at least one substructure between (i, k) and (i^*, k^*).

The RNA:RNA interaction prediction mechanisms are still being actively refined because many RNA:RNA interactions have important regulatory functions. For example, in eukaryotes, microRNAs can reduce the levels of mRNAs by interacting with the 3′UTRs of the mRNA targets [40, 246].

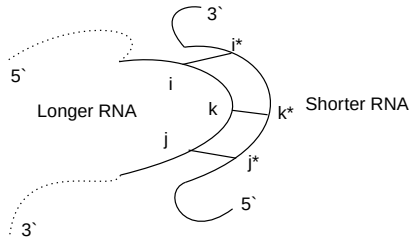


Figure 1.6: **RNA:RNA interaction and notations used for partition function.** Full shape of longer RNA is not shown. The nucleotides between $[i, j]$ and $[i^*, j^*]$ may contain mismatches. Figure adapted from Muckstein *et al.*, (2006).

Apart from these general mRNA features which play a role in protein expression, several specific features also exist. For example: Cis-regulatory elements such as promoters and enhancers, interactions of mRNA with sRNA and miRNA as well as introns in 5′ UTR in eukaryotes. However, our discussion will be based around the general features only.

A number of gene optimisation tools build a suitable cost or fitness function using a combination of these features. Typically, a genetic algorithm is then used to optimise the fitness. The synonymous mRNA sequence with the maximum fitness is regarded as the optimised mRNA sequence with optimal expression [251, 203, 191, 48, 236]. Despite being optimised on expression, the sequences may form aggregates, which cannot be used for further studies [86, 197]. This leads us to the discussion of optimising solubility.

1.1.4 Protein solubility

Solubility is defined as the proportion of the supernatant fraction, obtained after the centrifuging the the translation mixture, to the uncentrifuged total protein [174]. It ranges from 0% to 130% with solubility less than 30% categorised as aggregation-prone and greater than 70% are highly soluble. Several *intrinsic* features of the

protein itself such as molecular weight, flexibility, hydrophobicity, isoelectric point and structural propensities are also known to influence solubility [265, 46, 232, 60]. These intrinsic features can be modified by doing either mutagenesis or truncation, which might improve solubility. Several solubility enhancing tags are also available for example, thioredoxin (TRX), maltose binding protein (MBP), small ubiquitin-related modifier (SUMO) and glutathione S-transferase (GST). Although, the exact mechanisms of how these tags work is still unclear, it is proposed that they might act like a chaperone and assist in correct folding of the target protein or add charges which decreases the overall aggregation propensity [52].

Intrinsic properties of a protein

Intrinsic properties are based on the residues inside the poly-peptide chain. The commonly used intrinsic properties of proteins are described below.

Hydrophobicity Water soluble proteins fold such that the hydrophilic parts are exposed and can form hydrogen bonds with the water molecules whereas the hydrophobic parts are buried in the core. This hydrophobic effect is thought to be a driver of protein solubility [230]. Several scales have been proposed to measure the hydrophobicity of residues [136, 1, 116, 200]. However, none of these scales can fully capture the full range of behaviour of residues [42].

We will use Kyte-Doolittle's scale [136] for representative purposes. In this scale, residues are given a hydropathy score such that positive scores represent hydrophobicity and negative score represent hydrophilicity. The magnitude of score represents the strength. For example: isoleucine (I) is given 4.5 and is the most hydrophobic residue, where as arginine (R) is given -4.5 and is the most hydrophilic residue. Using these scores, a hydropathy plot for a given polypeptide can be drawn (Fig. 1.7). Hydropathy plot can be used to examine the hydrophobicity of protein region of interest. Using hydrophobic effect, we can then infer whether the residue is buried or located on the surface. Furthermore, the overall hydrophobicity of the protein can be determined by averaging the hydropathy scores across the polypeptide chain to obtain the GRand AVerages of hydropathY (GRAVY) score.

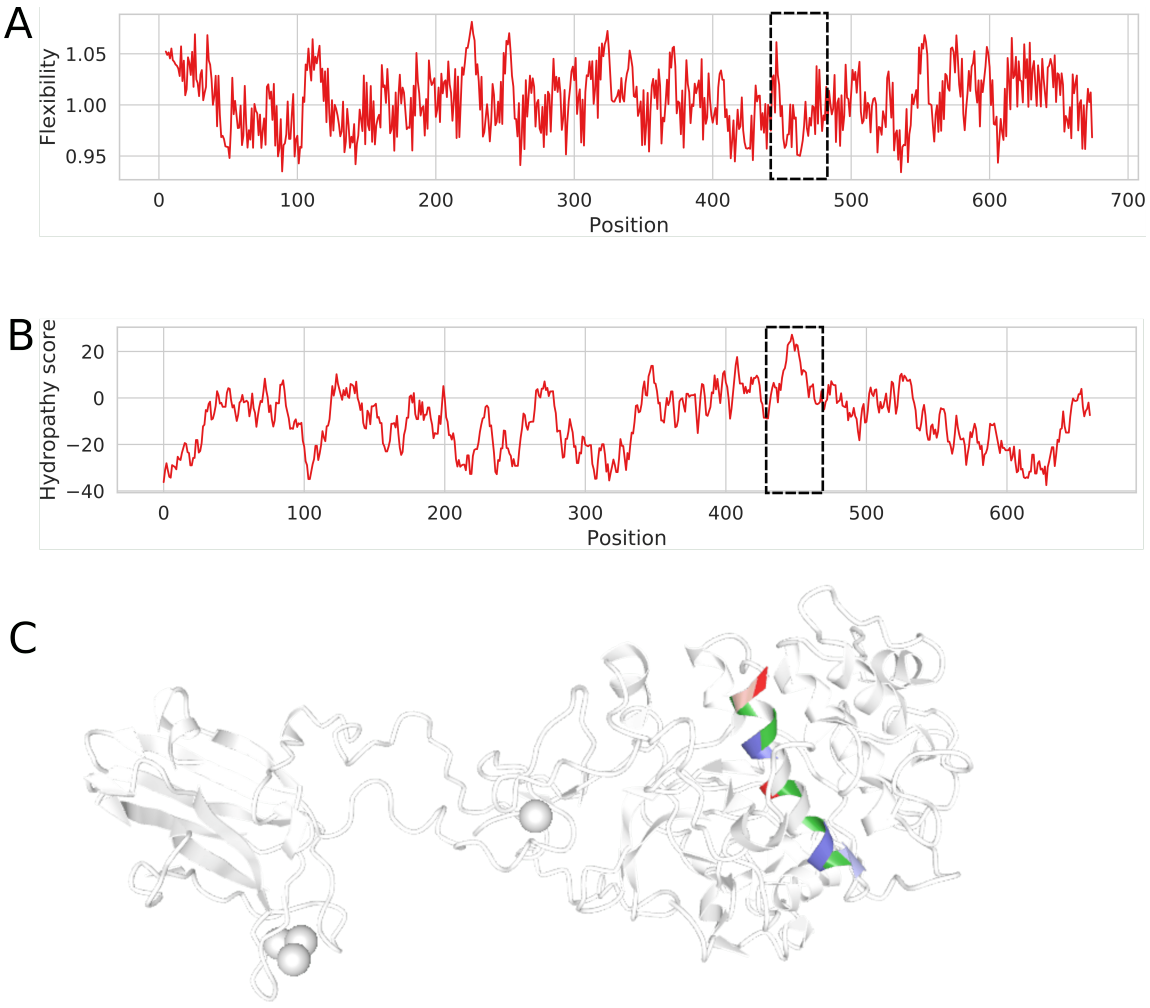


Figure 1.7: **Profiles of KPC1_DROME (UniProtKB P05130)**. **(A)** Flexibility plot generated using normalised B-factors from Vihinen et al. (1994), with a sliding window of 9 residues. For these normalised B-factors, values greater than one are regarded to be flexible and values less than one are rigid. **(B)** Hydropathy plot generated using Kyte-Doolittle's hydrophobicity scale, with a sliding window of 19 residues. For illustration, the hydropathy and flexibility of residues at around position 440 to 470 (shown by dotted box) are positive and rigid. This indicates the presence of an alpha helix which is supported by the actual 3D structure **(C)**. The coloured helix is the region 440 to 470.

Isoelectric point The net charge of a protein at a given pH depends on the acid dissociative constant (pKa) of ionisable groups such as amine and carboxyl group. At a certain pH, the amount of negative and positive charge are equal, resulting in a zero net charge. This pH is called the isoelectric point (pI). The net charge is positive at pH below pI and negative at pH above pI [215]. Since there are no net charges, the solubility of a protein is minimum at the isoelectric point.

Instability index Guruprasad et. al [85] found that the distribution of certain dipeptides on stable and unstable protein is different. Based on 12 unstable and 32 stable proteins, they assigned a weight called dipeptide instability weight value (DIWV) for all dipeptides. The instability index (II) is then given by equation 1.14.

$$II = \frac{10}{L} \sum_{i=1}^{L-1} DIWV(x_i y_{i+1}) \tag{1.14}$$

where L is the length of the sequence and $x_i y_{i+1}$ is a dipeptide.

Flexibility Protein molecules are dynamic and are inherently flexible due to the motion of the constituent atoms [250, 6, 234]. The structural flexibility of a protein can be inferred by using B-factors [250, 124, 220].

The B-factor (Equation 1.15) or temperature factor of the atoms in a crystalline structure is the measure of mean squared displacement vibration around their mean position ($u = \langle (x - x_0)^2 \rangle$), where x is the displacement of the atom from its mean position x_0. B-factor thus reflects the *orderedness* of the crystal lattice and subsequent uncertainty in X-ray scattering structure determination [209, 36, 30]. It has unit of $Å^2$.

$$B = 8\pi^2 u \tag{1.15}$$

To understand the effect of the B-factor, we define a quantity f called the atomic scattering factor as:

$$f = \frac{amplitude\ of\ wave\ scattered\ by\ an\ atom}{amplitude\ of\ wave\ scattered\ by\ one\ electron} \tag{1.16}$$

Atomic scattering factor after taking into account of the motion of atom becomes:

$$f_B = f \cdot e^{-B(sin\ \theta/\lambda)^2} \tag{1.17}$$

where θ is the Bragg's angle and λ is the wavelength of the wave. Thus, we see that B-factor attenuates the amplitude of wave scattered by an atom (Figure 1.8).

Experimental B-factors for different residues in a protein structure can be obtained the from Protein Data Bank (PDB). However, due to variation of structures, the B-factor of a given residue varies, even within the same polypeptide chain. For standardisation, the B-factor of residues within a chain are normalised using a z-score.

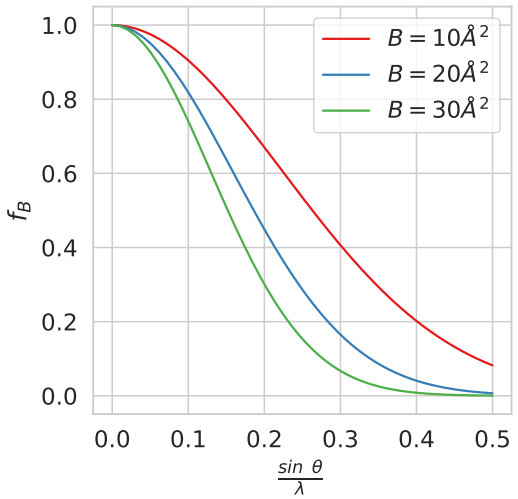


Figure 1.8: **Attenuation of the incident waves due to increasing B-factor.** As B-factor increases, atomic scattering factor (f) and consequently, the amplitude of wave scattered by an atom decreases rapidly.

$$B^i_{norm} = \frac{B^i - \langle B \rangle}{\sigma} \tag{1.18}$$

where, B^i_{norm} is the normalised B-factor of residue i, $\langle B \rangle$ is the mean and σ is the standard deviation of all B-factors across the chain.

A number of high resolution structures (for eg. 92 different proteins in Vihinen et al. [250], 292 in Smith et al. [220]) are sampled from a database (usually Protein Data Bank(PDB)) and B^i_{norm} is calculated for each residue on each protein. The mean of B^i_{norm} across the sampled structures is the final normalised B-factor [209, 220, 124, 250].

The structural flexibility of protein can either be determined by using these normal-

ised B-factors from experiments or by directly computing atomic displacements form molecular dynamics simulations (MDS) [64, 134]. Root mean square fluctuations and radius of gyration from MDS are useful in examining the flexibility. The profile plot (Fig. 1.7) can be used to visualise and infer the local flexibility and dynamics of the protein structure. Since structural flexibility is inherently related to the protein dynamics, it is thought to influence several properties such as conformal variations, functions, thermal stability, ligand binding and disordered regions [248, 233, 151, 276, 275, 7]. Although the relationship of flexibility with solubility has been noted previously [238], it has been overlooked.

Beside these features, amino acids also tend to have different structural propensities which is also thought to influence protein solubility [111, 109].

Many solubility prediction tools have been developed around these features using statistical models (e.g., linear and logistic regressions) and machine learning models (e.g., support vector machines and neural networks) [102, 88, 94, 221, 95, 270, 274]. Newer tools such as SOLart also employ 3D structural information for a precise estimation of solvent accessibility, which makes the prediction more accurate [108]. Despite a higher prediction accuracy, the generality of these structure based tools might be limited due to the lack of 3D structure information of many proteins of interest.

1.2 Signal Peptide

Secretory proteins such as hormones and toxins are some of the commercially important use cases of recombinant protein expression. These secretory proteins are often enriched with a short hydrophobic peptide at the N-terminal (Figure 1.10). This is called signal peptide (SP). Signal peptides are recognised by a protein complex called the signal recognition particle (SRP). SRP carries the signal peptide to the endoplasmic reticulum (ER) lumen, where post translation modification happens and the newly synthesised protein is secreted out. Despite having no consensus, SP have a tripartite structure as N-region, H-region and C-region [98] (Figure 1.9).

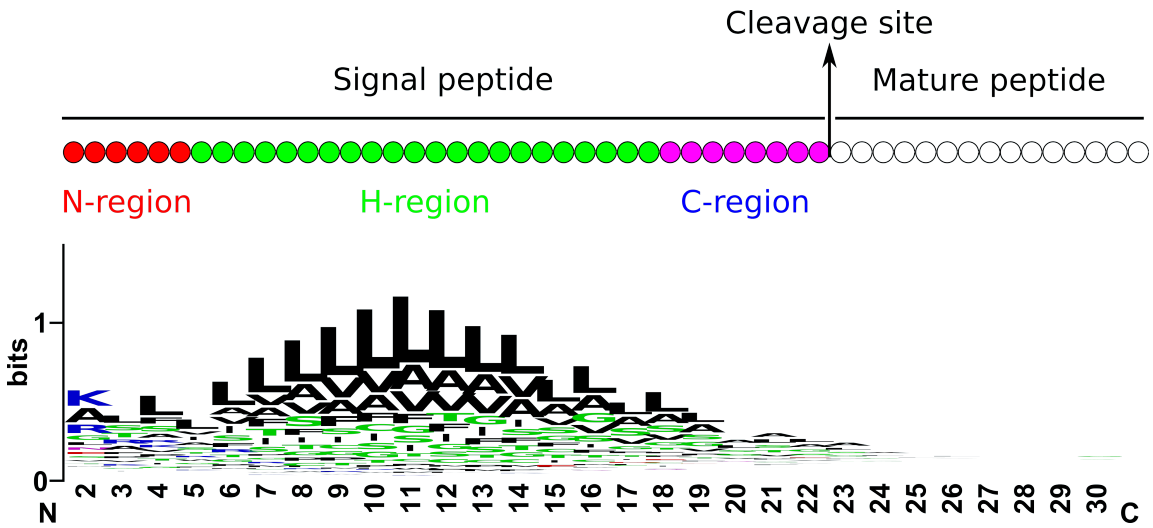


Figure 1.9: **Tripartite structure of a signal peptide.** N, H and C domains in a signal peptide. C domain also contains a cleavage site from which mature peptide is cleaved off after reaching the endoplasmic lumen. Sequence logo (bottom) shows an enrichment of Leucine (L) at the H-region.

The N-region usually consists of 2-5 charged residues, whereas the H-region is highly hydrophobic and forms an alpha helix. The C-region consists of small polar uncharged residues which often form a β-sheet structure. This topology is thought to help binding with signal peptidase and cleave off the signal peptide at the cleavage site.

The use of signal peptides in recombinant expression can lead to a high yield. As an added benefit, the proteins are often closer to the native activity because proper folding can happen inside the ER [78, 125].

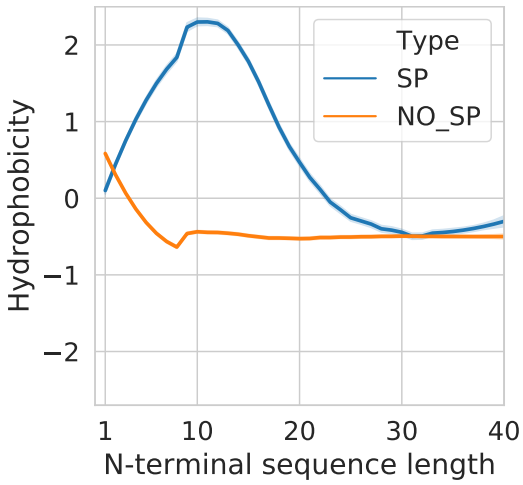


Figure 1.10: **Signal peptides are highly hydrophobic at the N-terminal.** Savitzky–Golay filtering is applied to the hydrophobicity. SP (N=2,609) and NO_SP (N=14,655) sequences from SignalP 5.0 dataset (Armenteros *et al.* (2019)). SP, Signal Peptide; NO_SP, Not a SP.

1.3 Mathematical optimisation

Optimisation is a way to find out the best solution to the problem. The starting point for optimisation is to define an objective function or the cost function. For an objective function $f : A \to \mathbb{R}$, the goal of optimisation is to find $x_0 \in A$ such that $f(x_0) \leq f(x) \; \forall \; x \in A$. If the objective function is differentiable, then we can use the gradient information to reach the optimal point. Non differentiable functions are usually optimised by heuristic optimisation methods. Heuristic methods often provide a near optimal solution even if the search space is large. The following optimisation methods are used in this study.

1.3.1 Simulated annealing

Simulated annealing is a heuristic optimisation technique inspired by the way metals cool and anneal. More precisely, it is based on the thermodynamics of a system undergoing a slow cooling so that the atoms have sufficient time to redistribute to form a crystalline structure—a state of minimum energy [131, 113, 127, 33, 187]. This algorithm is often used to solve combinatorial optimisation problems in bioinfomatics and other sciences. It has been used to align and predict non-coding RNAs from multiple sequences [144], to find consensus sequences [127] and optimise the ribosome binding sites [203] and mRNA folding using minimum free energy models [81].

Let p_i be the probability that a system is in a certain state i with energy ϵ_i. Then entropy of the system is defined as:

$$S = -k_B \sum_i p_i \ln(p_i) \tag{1.19}$$

where k_B is the Boltzmann's constant. For any system, the second law of thermodynamics states that the entropy is maximised as the system evolves towards a thermodynamic equilibrium. Hence, to know the behaviour of the system, Equation 1.19 needs to be maximised under these two constrains:

$$\begin{aligned} 1) \; \sum_j p_j &= 1 \\ 2) \; \sum_j p_j \epsilon_j &= E \end{aligned} \tag{1.20}$$

The first condition simply means that sum of all probabilities should be one while the second condition implies the total energy of the system is a constant E. Using Lagrange multipliers,

$$S = -k_B \sum_j p_j \ln(p_j) - \lambda[\sum_j p_j - 1] - \beta[\sum_j p_j \epsilon_j - E] \tag{1.21}$$

Setting the first derivative of Equation 1.21 to zero, we obtain:

$$p_i = e^{1-\lambda}.e^{-\beta\epsilon_i} \tag{1.22}$$

β, also known as thermodynamic beta, can be shown to be equal to $1/k_BT$, where T is the absolute temperature. λ is chosen to normalise the probability p_i in Equation 1.22. We thus arrive at the Boltzmann's probability distribution:

$$p_i = \frac{1}{Z} e^{-\beta\epsilon_i} \tag{1.23}$$

where Z is the partition function. For a system in thermal equilibrium at temperature T, Equation 1.23 gives us a set of probability mass functions for all different energy states ϵ_i. An interesting implication of the Boltzmann's distribution is that even at low temperature, there is a non zero probability of system being at high energy.

A mathematical way to reach the minima of the Boltzmann's distribution is by using a Markov chain sampling while simultaneously decreasing the temperature. The decreasing temperature is simulated by applying a cooling schedule, which is generally exponentially decreasing [131]. Markov chain sampling can be performed by using the Metropolis-Hastings algorithm or perhaps Gibb's sampling, although this is less commonly used [127]. For the Metropolis-Hastings algorithm, a *bad* move (uphill) E_2 from initial state E_1 such that $E_2 > E_1$, is accepted if $R(0,1) \geq p_2/p_1$, where $R(0,1)$ is a uniformly generated random number between 0 and 1 [93]. Unlike gradient descent, where only *good* moves (downhill) ($E_2 < E_1$) are accepted, this algorithm can move both uphill and downhill without getting trapped in any local minima. The probability of system moving uphill, however, decreases with temperature [187].

A use case of simulated annealing on a Rastrigin function (a test function commonly used to to demonstrate optimisation) (Figure 1.11) is shown in Figure 1.12.

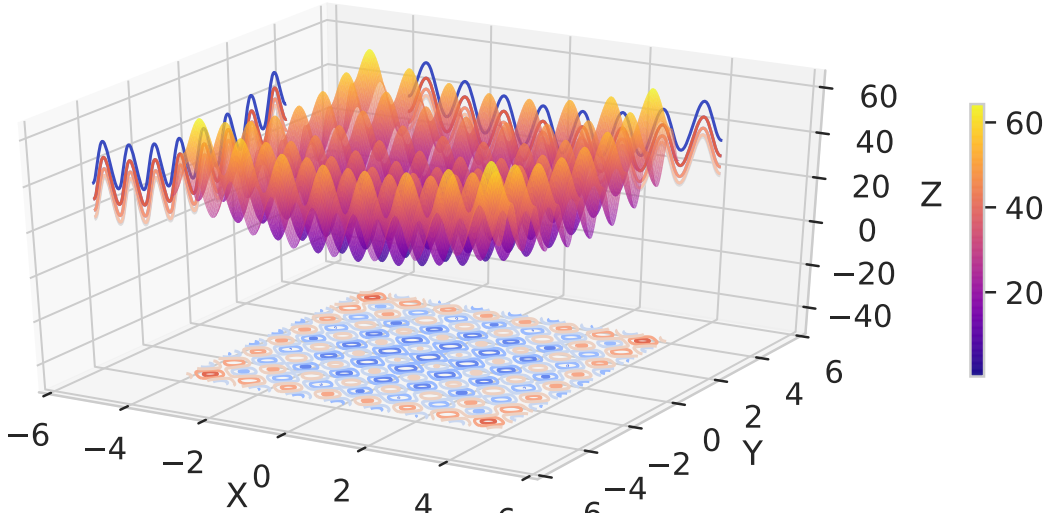


Figure 1.11: **Two dimensional Rastrigin function.** Rastrigin function is frequently used as a test function in optimisation problems because of a large number of local minima. The global minima is at $(0, 0)$ where the value of function is 0.

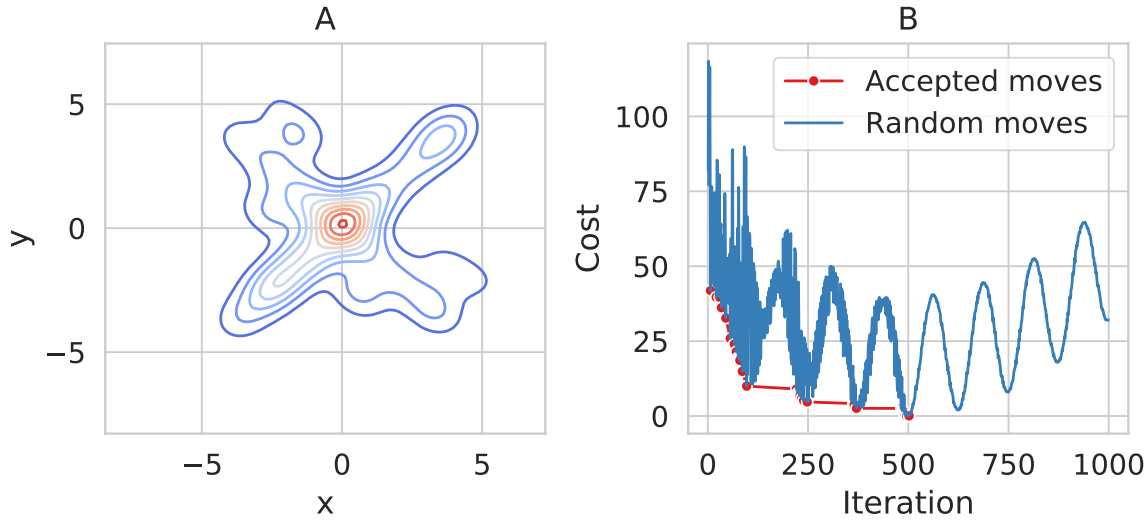


Figure 1.12: **Simulated annealing on a two dimensional Rastrigin function.** This simulation was run for $1,000$ iterations with an initial temperature of 1. The minimum found was at $(0.006, -0.013)$, where the value of function was 0.04. **(A)** Kernel density estimate of the moves during simulated annealing shows that they converge near the true minimum $(0, 0)$. **(B)** The accepted costs are usually immune to the traps of local minima. The algorithm converted at 502^{nd} iteration, where the accepted cost is 0.04. This is close to the global minima 0.

1.3.2 Nelder-Mead method

The Nelder-Mead method is another type of derivative-free, heuristic optimisation technique [168]. The method, however, requires the objective function to be evaluated at different points hence can also be categorised as a *direct search* method. Direct search methods usually employ a non-degenerate simplex at each step [138]. Simplex in n dimensions is a convex-hull of $n+1$ vertices and can be understood as a generalisation of triangles. By non-degenerate, we mean the volume of the simplex is non zero.

For an n dimensional function $f(x)$, initialisation is performed by choosing $n+1$ points $(x_i, 1 \leq i \leq n+1)$ which form a simplex. These points are ordered such that $f(x_1) \leq f(x_2) \leq ...f(x_{n+1})$. Since the goal is to minimise $f(x)$, in this case x_{n+1} is the worst point and x_1 is the best. The worst point is then reflected along the centroid of the simplex to give a new vertex x_r such that the volume is preserved and the non-degeneracy is maintained [187]. Three cases may happen:

- $f(x) \leq f(x_r) \leq f(x_{n+1})$

  In this case, the new move is neither good nor bad. x_r replaces some older point x_o where $f(x) \leq f(x_o) \leq f(x_{n+1})$.

- $f(x) \leq f(x_{n+1}) \leq f(x_r)$

  In this case, the new move is worse. The simplex is contracted by decreasing x_{n+1} to x_c. If $x_c < Min(f(x_{n+1}), f(x_r))$, x_{n+1} is replaced by x_c. Else, contraction is repeated.

- $f(x_r) \leq f(x_1) \leq f(x_{n+1})$

  In this case, the new move is good. x_r replaces the older best point x_1.

The ordering of points and reflection are repeated, until the change in the value of the function at the best point falls below a preset tolerance. This method tends to get stuck at local minima (Figure 1.14 A, B), because whenever it encounters one, the algorithm contracts the simplex rather than exploring the surrounding. However for functions with a few or no local minima, for example, Rosenbrock function (Fig.

1.13), if the starting point is good, this algorithm is very efficient in finding the minima (which is often global) (Figure 1.14 C, D).

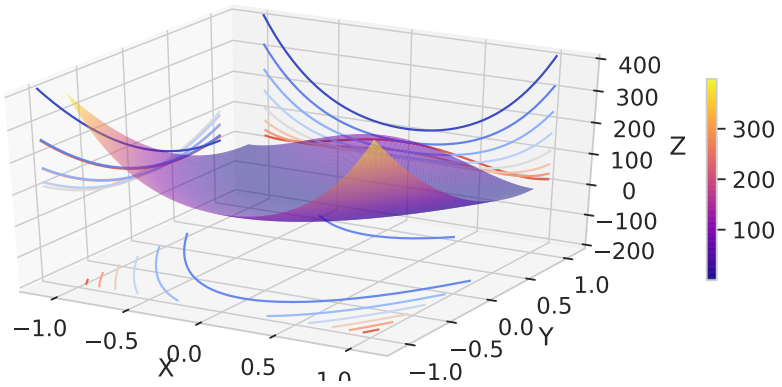


Figure 1.13: **Two dimensional Rosenbrock function.** Rosebrock function is a test function often used to demonstrate optimisation. It has a characteristic valley where the global minimum lies. The global minima is at $(1, 1)$ where the value of function is 0.

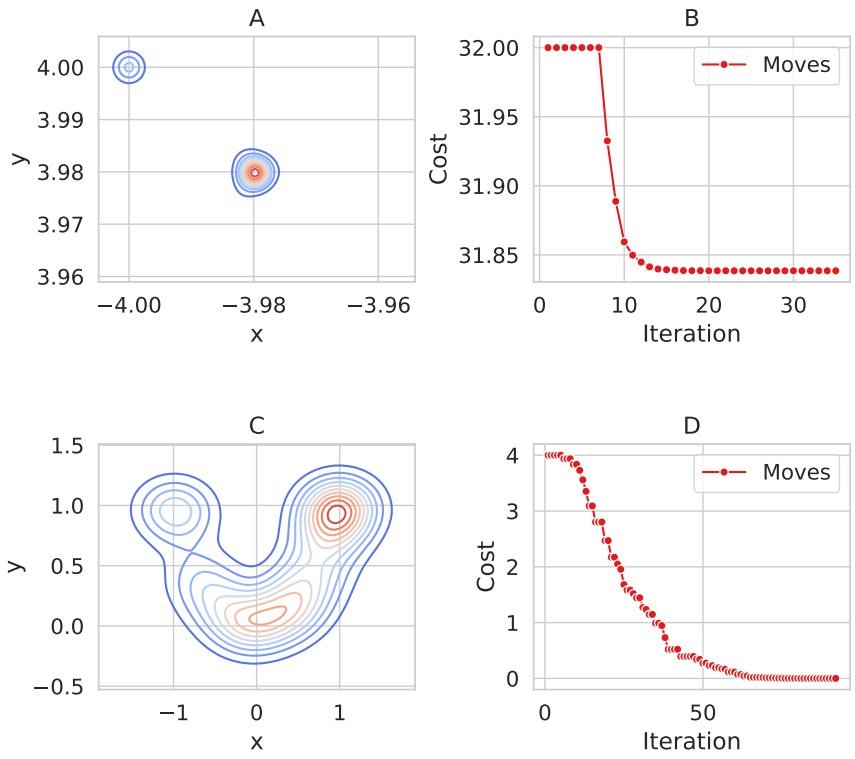


Figure 1.14: **Although the Nelder-Mead method tends to get stuck on local minima, a good initial point can result in a global optimum. (A)** Nelder-Mead method applied on a Rastrigin function. Rastrigin function is a test function often used to demonstrate optimisation. Kernel density estimate of the moves shows that the algorithm gets stuck at a local minima $(-3.98, 3.98)$. **(B)** The algorithm terminated after 35 iterations where the minimum was found to be 3.36, which is a local minimum. **(C)** Nelder-Mead method applied on a Rosenbrock function with $(-2, 2)$ as the starting point. Kernel density estimate of the moves shows that the algorithm moves around the valley and terminates at $(1.0001, 1.0002)$ close to the global minima. **(D)** The algorithm terminated after 92 iterations where the minimum was found to be 3.36×10^{-8}.

1.4 Classifiers

Classifiers are functions which map the input vector to some specific category. In the context of machine learning, we usually use a set of labelled data (known as training data) to fit a classifier. This fitted classifier, also known as model, can then be used to do predictions on unknown data. Many types of classifiers exist such as linear classifiers, support vector machines, decision trees and neural networks (Figure 1.15). One special classifier built from an ensemble of decision trees called the Random forest classifier is used in this work.

1.4.1 Random forest

A random forest is a classifier consisting of a collection of tree structured classifiers $\{h(x, \Theta_k), k = 1, ...\}$ where the $\{\Theta_k\}$ are independent, identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [32]. For a given data, a number of decision trees B are constructed using bootstrapping. Every time a new split is performed, a subset m of total vectors (features) f is used, such that $m < f$. If $m = f$, this procedure is called bagging. In practise, usually, $m \approx \sqrt{f}$. Surprisingly, the number of trees, B, is not critical and setting this to a very high value will not lead to overfitting, which can be shown to be a consequence of the strong law of large numbers [115, 32]. This also makes random forests more robust to noise and outliers than other classifiers.

Typically, each bagged tree uses around two-third of training points. Out-of-bag (OOB) estimates (errors) can be computed by preforming predictions on the remaining one-third points. OOB errors reflect the generalisability of the model. For each feature, we can also compute the total reduction of Gini index by that feature at each tree split, where Gini index is defined by:

$$G = \sum_{k=1}^{K} (p_{mk}(1 - p_{mk}))$$

and is a measure of variance across K classes for the proportion of training observations, p_{mk}, in the m^{th} region that are from the k^{th} class. This gives us the feature importance, sometimes called the Gini importance. However, it should be noted that m features used for each split are actually random and is not based on feature importance at all. This seemingly deceptive strategy forces all the subsequent trees to not use the same strongest predictor, thus forcing a decorrelation among the trees. This makes the outputs less variable and more reliable [115].

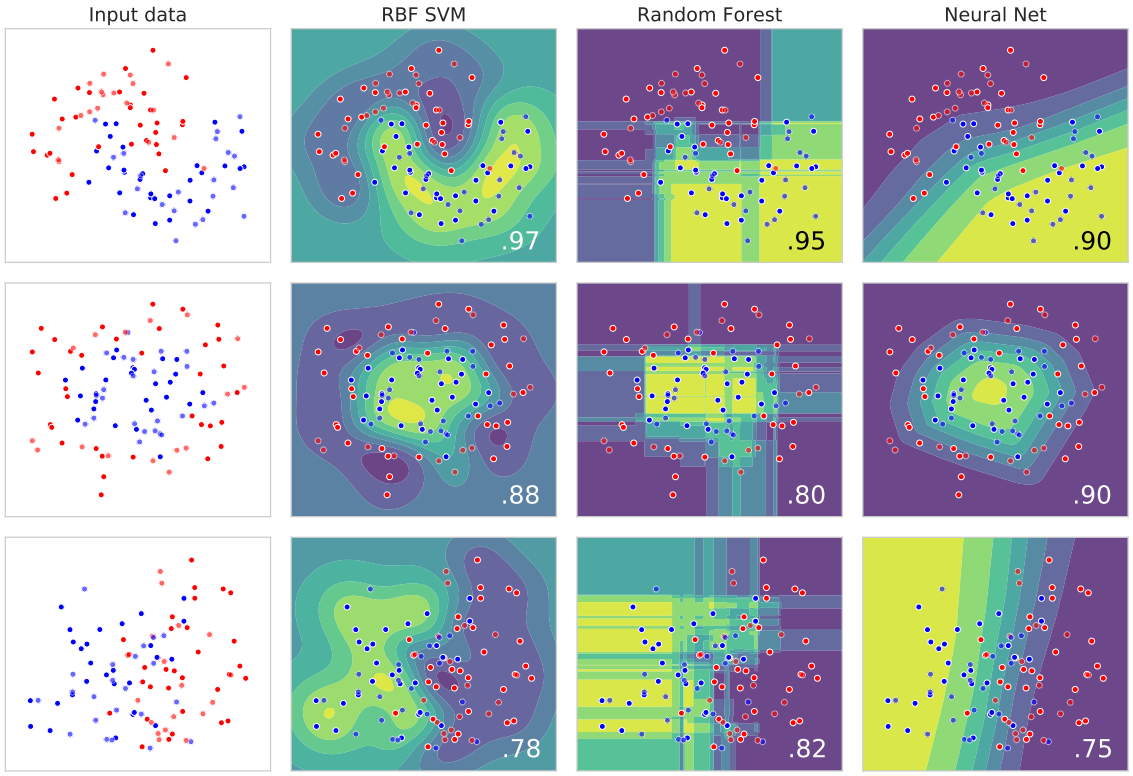


Figure 1.15: **Comparison between the some of the commonly used classifiers on synthetic datasets.** Three commonly used classifiers—Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, Random forest and a simple neural network (multi-layered perceptron) on three synthetic datasets as input. The input data is a binary data coloured as red and blue. The decision boundary is shown as contours and the accuracy of each classifier is shown on bottom right. Adapted from https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html.

Chapter 2

Protein yield is tunable by synonymous codon changes of translation initiation sites

Associate Professor Paul Gardner conceived the study. Dr Chun Shen Lim analysed multiple large-scale datasets and subsequently found that mRNA accessibility is potentially the best predictor of protein levels (Fig 2.1, 2.2, A.6, A.7) besides mRNA levels. In addition, he tested the translation elongation codon optimisation tool Iχnos (Fig A.8), fitted the logistic regression (Fig A.12) and did the densitometric analysis (Fig A.14). The GFP and *Renilla* luciferase plasmids were constructed by Dr Craig van Dolleweerd at the Callaghan Innovation Protein Science and Engineering, University of Canterbury (Fig A.1- A.5 and Table A.1 - A.3). The GFP reporter experiment was carried out by Dr Daniela M. Remus at the Callaghan Innovation Protein Science and Engineering. The *Renilla* luciferase reporter experiment was performed by Dr Augustine Chen at the University of Otago.

I conceived the idea of simulating the cellular system during overexpression and performed the coarse-grained simulations. Due to the lack of open source libraries specialised for sequence optimisation, I developed a simulated annealing (Metropolis-Hastings) based algorithm to maximise the accessibility at translation initiation sites using synonymous codon substitution. Using this algorithm, I developed TIsigner

(both the command line https://github.com/Gardner-BinfLab/TIsigner/tree/master/TIsigner_cmd and the web server https://tisigner.com/tisigner). I completed all the remaining analysis and figures, and drafted the manuscript [21]. Drs Gardner and Lim supervised the study.

2.1 Abstract

Recombinant protein production is a key process in generating proteins of interest in the pharmaceutical industry and biomedical research. However, about 50% of recombinant proteins fail to be expressed in a variety of host cells. To address this problem, we have modified up to the first nine codons of messenger RNAs with synonymous substitutions and showed that protein levels can be tuned. These modifications alter the 'accessibility' of translation initiation sites. We have also revealed the dynamics between accessibility, gene expression, and turnovers using a coarse-grained simulation.

2.2 Introduction

Recombinant protein expression has numerous applications in biotechnology and biomedical research. Despite extensive refinements in protocols over the past three decades, half of the experiments fail in the expression phase (http://targetdb.rcsb.org/metrics/). Notable problems are the low expression of 'difficult-to-express' proteins such as those found in, or associated with, membranes, and the poor growth of the expression hosts, which may relate to toxicity of heterologous proteins [129] (see [13, 198] for detailed reviews). Despite these issues, mRNA abundance can only explain up to 40% of the variation in protein abundance, due to the complexity of translation and turnover of biomolecules [223, 91, 143, 225, 210, 17, 231]. Furthermore, strong promoters used in expression vectors do not always lead to a desirable level of protein expression because of leaky expression [198].

For *Escherichia coli*, mainstream models that may explain the lower-than-expected correlation between mRNA and protein levels are codon-usage and mRNA structure.

Codon analysis is based on the frequency of codon usage in highly expressed proteins using codon adaptation index (CAI) [214] or tRNA adaptation index (tAI) [196, 202], whereas mRNA folding analysis predicts the stability of mRNA secondary structures. Codon usage bias is thought to correlate with tRNA abundance, translation efficiency and protein production [214, 87, 196, 202, 34, 178, 247] but its usefulness has been questioned [133, 185, 27, 35]. More recent studies show stronger support for models based on mRNA folding, in which the stability of RNA structures around the Shine-Dalgarno sequence and translation initiation sites inversely correlates with protein expression [219, 133, 185, 66, 240, 35]. We recently proposed a third model in which the avoidance of inappropriate interactions between mRNAs and non-coding RNAs (ncRNAs) has a strong effect on protein expression [243]. The roles of these models in protein expression is an active area of research.

The algorithms for gene optimisation sample synonymous protein-coding sequences using 'fitness' models based on CAI, tAI, mRNA folding, and/or G+C content (%) [251, 203, 191, 48, 236]. However, these 'fitness' models are usually based on some of the above findings that rely on either endogenous proteins, reporter proteins, or a few heterologous proteins with their synonymous variants. It is unclear whether these features are generalisable to explain the expression of all heterologous proteins. To address this question, we studied multiple large datasets across species in order to extract features that allow us to predict the outcomes of 11,430 experiments of recombinant protein expression in *E. coli*. With this information, we propose how such features can be exploited to fine-tune protein expression at a low cost.

2.3 Results

2.3.1 Accessibility of translation initiation sites strongly correlates with protein abundance

To identify a better energetic model for mRNA structure that explains protein expression, we examined an *E. coli* expression dataset of green fluorescent protein (GFP) fused in-frame with a library of 96-nt upstream sequences (N=244,000 vari-

ants) [35]. These 96-nt sequences were randomly generated to achieve a full factorial design by varying A+T content (%), CAI, codon ramp bottleneck position and strength, hydrophobicity of the encoded peptide, and MFEs. We removed the redundancy of these 96-nt upstream sequences by clustering on sequence similarity, giving rise to 14,425 representative sequences. We calculated the accessibility (also known as 'opening energy' based on unpairing probability) for all the corresponding sub-sequences (see Methods). We examined the correlation between the opening energies and GFP levels. We found that the opening energies of translation initiation sites, in particular from the nucleotide positions -30 to 18 ($-30:18$), shows the highest correlation with protein abundance (Fig 2.1A; Spearman's correlation, $R_s = -0.65$, $P < 2.20 \times 10^{-16}$). This is stronger than the highest correlation between the minimum free energy $-30:30$ and protein abundance, which was previously reported as the highest ranked feature (Fig 2.1A; $R_s = 0.51$, $P < 2.20 \times 10^{-16}$). To account for multiple-testing, the P-values were adjusted using Bonferroni's correction and reported to machine precision. The datasets used and results are summarised in Supplementary Table S4.

We repeated the analysis for a dataset of yellow fluorescent protein (YFP) expression in *Saccharomyces cerevisiae* [66]. This dataset corresponds to a library of $5'$ UTR variants, in which the 10-nt sequences preceding the YFP translation initiation site were randomly substituted (N=2,041 variants). In this case, the opening energy $-7:89$ showed a stronger correlation with protein abundance than that of the minimum free energy $-15:50$ reported previously (2.1B; $R_s = -0.55$ versus 0.46).

To examine the usefulness of accessibility in complex eukaryotes, we analysed a dataset of GFP expression in *Mus musculus* [175]. The reporter library was originally designed to measure the strength of translation initiation sequence context, in which all possible substitutions were made at the flanking regions of the GFP translation initiation site (6-nt upstream region and 2-nt downstream region of initiation codon; N=65,536 variants). Here the opening energy $-8:11$ showed a maximum correlation with expressed proteins, which again, is stronger than that of the minimum free energy $-30:30$ (2.1C; $R_s = -0.28$ versus 0.12).

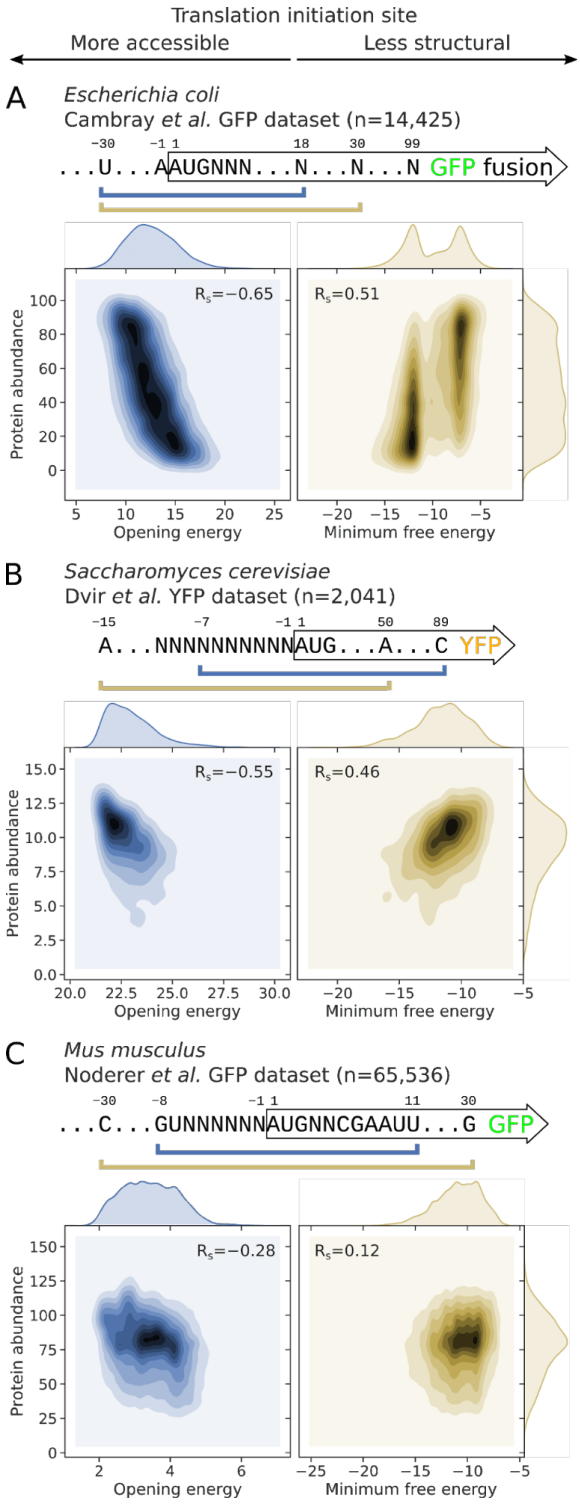


Figure 2.1: **Correlations between the opening energies of translation initiation sites and protein abundance are stronger than that of minimum free energy. (A)**. For *E. coli*, the opening energy at the region -30:18 shows the strongest correlation with protein abundance (see also 2.2B or Supplementary Fig S6A, sub-sequence l=48 at position i=18). For this analysis, we used a representative GFP expression dataset (N=14,425) from Cambray et al. (2018).The minimum free energy -30:30 shown was determined by Cambray et al. (right panel). **(B)** For *S. cerevisiae*, the opening energy -7:89 shows the strongest correlation with protein abundance (see also Supplementary Fig A.6B, sub-sequence l=96 at position i= 89). For this analysis, we used the YFP expression dataset (N=2,041) from Dvir et al. (2013). The minimum free energy -15:50 was previously shown to correlate the best with protein abundance (right panel). **(C)** For *M. musculus*, the opening energy -8:11 shows the strongest correlation with protein abundance (see also Supplementary Fig A.6C, sub-sequence l=19 at position i=11). For this analysis, we used the GFP expression dataset (N=65,536) from Noderer et al. (2014). The minimum free energy -30:30 was shown (right panel). See also Supplementary Table S4. R_s, Spearman's rho. Bonferroni adjusted P-values are below machine's underflow level for the correlations between opening energies and protein abundances shown in the left panels.

Taken together, our findings suggest that the accessibility of translation initiation sites strongly correlates with protein abundance across species. Interestingly, our findings also suggest that the Shine-Dalgarno sequence [217] at $-13:-8$ should be accessible to recruit ribosomes.

2.3.2 Accessibility predicts the outcome of recombinant protein expression

We investigated how accessibility performs in the real world in prediction of recombinant protein expression. For this purpose, we analysed 11,430 expression experiments in *E. coli* from the 'Protein Structure Initiative:Biology' (PSI:Biology) [43, 213, 2]. These PSI:Biology targets were expressed using the pET21_NESG expression vector that harbours the T7lac inducible promoter and a C-terminal His tag [2].

We split the experimental results of the PSI:Biology targets into protein expression 'success' and 'failure' groups that were previously curated by DNASU (N=8,780 and 2,650, respectively; see Supplementary Fig A.7). These PSI:Biology targets span more than 189 species and the failures are representative of various problems in heterologous protein expression. Only 1.6% of the targets were *E. coli* proteins, which is negligible (N=179; see Supplementary Fig A.7).

We calculated the opening energies for all possible sub-sequences of the PSI:Biology targets as above (2.2, positions relative to initiation codons). For each sub-sequence region, we used the opening energies to predict the expression outcomes and computed the prediction accuracy using the area under the receiver operating characteristic curve (AUC; see 2.2C). A closer look into the correlations between opening energies and expression outcomes, and AUC scores calculated for the sub-sequence regions reveals a strong accessibility signal of translation initiation sites (2.2B and C, Cambray's GFP and PSI:Biology datasets, respectively). We matched the correlations and AUC scores by sub-sequence regions and confirmed that sub-sequence regions that have strong correlations are likely to have high AUC scores (2.2D).

In contrast, the sub-sequence regions that have zero correlations are not useful for predicting the expression outcomes (AUC approximately 0.5).

We then asked how accessibility manifests in the endogenous mRNAs of *E. coli*, for which we studied a proteomics dataset of 3,725 proteins available from PaxDb [257]. As expected, we observed a similar accessibility signal, with the region -25:16 correlated the most with protein abundance (2.2E). However, the correlation was rather low ($R_s = -0.17$, $P < 2.2 \times 10^{-16}$), which may reflect the limitation of mass spectrometry to detect lower abundances [229, 173]. Furthermore, the endogenous promoters have variable strength, which gives rise to a broad range of mRNA and protein levels [59, 57]. Taken together, our results show that the accessibility signal of translation initiation sites is very consistent across various datasets analysed (Supplementary Fig A.6 and 2.2).

2.3.3 Accessibility outperforms other features in prediction of recombinant protein expression

To choose an accessibility region for subsequent analyses, we selected the top 200 regions from the above correlation analysis on Cambray's GFP dataset (2.2B) and used random forest to rank their Gini importance scores in prediction of the outcomes of the PSI:Biology targets. The region $-24 : 24$ was ranked first, which is nearly identical to the region $-23 : 24$ with the top AUC score (2.2C, AUC=0.70). We therefore used the opening energy at the region $-24 : 24$ in subsequent analyses.

We asked how the other features perform compared to accessibility in prediction of heterologous protein expression, for which we analysed the same PSI:Biology dataset. We first calculated the minimum free energy and avoidance at the regions -30:30 and 1:30, respectively. These are the local features associated with translation initiation rate. We also calculated CAI [214], tAI [239], codon context (CC) [8], G+C content, and Iχnos scores [241]. CC is similar to CAI except it takes codon-pairs into account, whereas the Iχnos scores are translation elongation rates predicted using a neural network model trained with ribosome profiling data (Supplementary Fig A.8). These are the global features associated with translation elongation rate. We

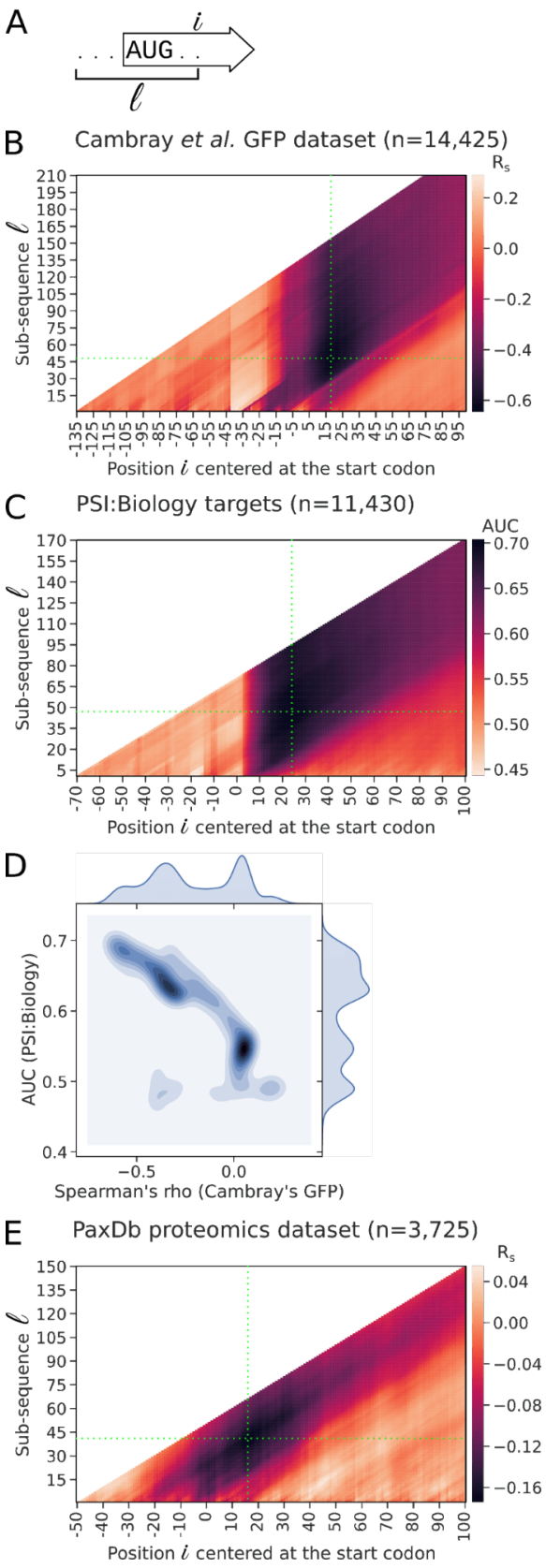


Figure 2.2: **Opening energies of regions surrounding the Shine-Dalgarno and start codons are predictive of protein expression in *E. coli*.** **(A)** Schematic representation of a transcript sub-sequence l at position i for the calculation of opening energy. **(B)** Correlation between the opening energies for the sub-sequences of GFP transcripts and protein abundance. The opening energy at the region -30 to 18 nt (green crosshair) shows the strongest correlation with protein abundance [$R_s = -0.65$; N=14,425, GFP expression dataset of Cambray et al. (2018)]. **(C)** Prediction accuracy of the expression outcomes of the PSI:Biology targets using opening energy (N=11,430). The opening energy at the region -23:24 (green crosshair) shows the highest prediction accuracy score (AUC=0.70). **(D)** Comparison between the correlations and AUC scores by sub-sequence region taken from the above analyses. Sub-sequences that have strong correlations are likely to have high AUC scores, whereas the sub-sequence regions that have no correlations are likely not useful in prediction of the expression outcomes. **(E)** Correlation between the opening energies for the sub-sequences of *E. coli* transcripts and protein abundance. The transcripts used for this analysis are protein-coding sequences concatenated with 50 and 10 nt located upstream and downstream, respectively. The opening energy at the region -25:16 (green crosshair) shows the strongest correlation with protein abundance ($R_s = -0.17$; N=3,725, PaxDb integrated proteomics dataset). See also Supplementary Table S4. R_s, Spearman's rho.

built a random forest model to rank the Gini importance scores of these local and global features. The local features ranked higher than the global features (2.3A). We then calculated and compared the prediction accuracy of these features. The AUC scores for the local features were 0.70, 0.67 and 0.62 for the opening energy, minimum free energy and avoidance, respectively, whereas the global features were 0.58, 0.57, 0.54, 0.54 and 0.51 for Iχnos, G+C content, CAI, CC and tAI, respectively (2.3B). The local features outperform the global features, suggesting that effects on translation initiation are a major predictor of the outcome of heterologous protein expression. We further examined the local G+C contents corresponding to the local features (Supplementary Fig A.9). The G+C contents in the regions $-24:24$ and $-30:30$ weakly correlate with opening energy and minimum free energy, respectively. The AUC scores for these local G+C contents are also lower than the corresponding local features, suggesting that these local G+C contents are not good proxies for the corresponding local features. Overall, our findings support previous reports that the effects on translation initiation are rate-limiting [133, 240] which, interestingly, correlate with the binary outcome of recombinant protein expression (2.3C). Importantly, accessibility outperformed all other features.

To identify a good opening energy threshold, we calculated positive likelihood ratios for different opening energy thresholds using the cumulative frequencies of true negative, false negative, true positive and false positive derived from the above receiver operating characteristic (ROC) analysis (Supplementary Fig A.12, top panel). Meanwhile, we calculated the 95% confidence intervals of these positive likelihood ratios using 10,000 bootstrap replicates. We reasoned that there is an upper and lower bound on translation initiation rate, therefore the relationship between translation initiation rate and accessibility is likely to follow a sigmoidal pattern. We fit the positive likelihood ratios into a four-parametric logistic regression model (Supplementary Fig A.12). As a result, we are 95% confident that an opening energy of 10 kcal/mol or below at the region $-24:24$ is about two times more likely to belong to the sequences which are successfully expressed than those that failed.

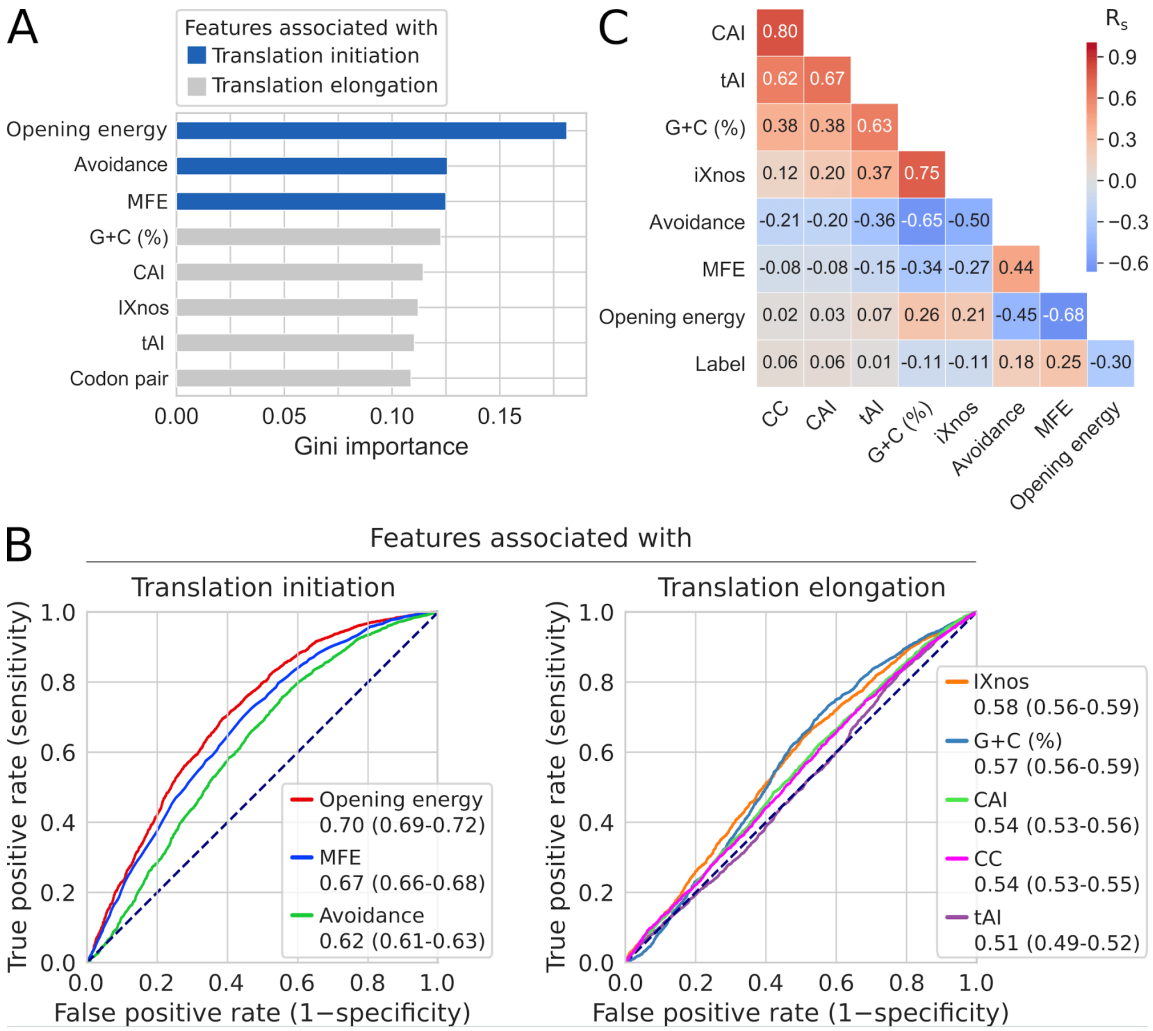


Figure 2.3: **Accessibility of translation initiation sites is the strongest predictor of heterologous protein expression in *E. coli*.** **(A)** mRNA features ranked by Gini importance for random forest classification of the expression outcomes of the PSI:Biology targets (N=8,780 and 2,650, 'success' and 'failure' groups, respectively). The features associated with translation initiation rate (purple; opening energy -24:24, minimum free energy -30:30, and mRNA:ncRNA avoidance 1:30) have higher scores than the feature associated with translation elongation rate [blue; tRNA adaptation index (tAI), codon context (CC), codon adaptation index (CAI), G+C content (%), and Iχnos]. The Iχnos scores are translation elongation rates predicted using a neural network model trained with ribosome profiling data (Supplementary Fig A.8). **(B)** ROC analysis shows that accessibility (opening energy -24:24) has the highest classification accuracy. The AUC scores with 95% confidence intervals are shown. See also Supplementary Table S4. **(C)** Accessibility (opening energy -24:24) is the best feature in explaining the expression outcomes. R_s, Spearman's rho.

2.3.4 Accessibility can be improved using a simulated annealing algorithm

The above results suggest that accessibility can, in part, explain the low expression problem of heterologous protein expression. Therefore, we sought to exploit this idea for optimising gene expression. We developed a simulated annealing algorithm to maximise the accessibility at the region $-24 : 24$ using synonymous codon substitution (see Methods). Previous studies have found that full-length synonymous codon-substituted transgenes may produce unexpected results, such as a reduction in mRNA abundance, RNA toxicity, and/or protein misfolding [12, 243, 241, 161]. Therefore, we sought to determine the minimum number of codons required for synonymous substitutions in order to achieve near-optimum accessibility. For this purpose, we used the PSI:Biology targets that failed to be expressed. We applied our simulated annealing algorithm such that synonymous substitutions can happen at any codon of the sequences except the start and stop codons, although the changes may not necessarily happen to all codons due to the stochastic nature of our optimisation algorithm (see Methods). Next, we constrained synonymous codon substitution to the first 14 codons and applied the same procedure (Supplementary Fig A.10A). Therefore, the changes may only occur at any or all of the first 14 codons. We repeated the same procedure for the first nine and also the first four codons. Thus a total of four series of codon-substituted sequences were generated. We then compared the distributions of opening energy -24:24 for these series using the Kolmogorov-Smirnov statistic (DKS; see Supplementary Fig A.10B). The distance between the distributions of the nine and full-length codon-substituted series was significantly different yet sufficiently close (DKS=0.087, P=3.3×10^{-8}), suggesting that optimisation of the first nine codons is sufficient in most cases to achieve an optimum accessibility of translation initiation sites. We named our software Translation Initiation coding region designer (TIsigner), which by default, allows synonymous substitutions in the first nine codons.

We asked to what extent the existing gene optimisation tools modify the accessibility of translation initiation sites. For this purpose, we first submitted the PSI:Biology

targets that failed to be expressed to the ExpOptimizer web server from NovoPro Bioscience (see Methods). We also optimised the PSI:Biology targets using the standalone version of Codon Optimisation OnLine (COOL) [48]. We found that both tools increase accessibility indirectly even though their algorithms are not specifically designed to do so. In fact, a purely random synonymous codon substitution on these PSI:Biology targets using our own script resulted in similar increases in accessibility (Supplementary Fig A.10C). These results may explain some indirect benefits from the existing gene optimisation tools (i.e. any change from suboptimal is likely to be an improvement, see below).

2.3.5 Low protein yields can be improved by synonymous codon changes in the vicinity of translation initiation sites

To demonstrate that heterologous protein expression is tunable with minimum effort, we designed and tested a series of GFP reporter gene constructs. We tested 29 plasmids harbouring GFP reporter genes with synonymous changes within the first nine codons (opening energies from 5.56 to 21.68 kcal/mol; Supplementary Table S5 and Supplementary Methods). GFP expression is controlled by an IPTG inducible T7lac promoter. In addition, all plasmids harbour a second reporter gene, i.e. mScarlet-I, which is controlled by the constitutive promoter from the nptII gene for aminoglycoside-3'-O-phosphotransferase of E. coli transposon Tn5 [25, 207]. mScarlet-I expression was measured to correct for plasmid copy number and as a proxy for bacterial growth [208].

Consistent with the above results, the GFP level significantly correlates with accessibility (i.e., anti-correlates with opening energy, $R_s = -0.53$, $P = 3.4 \times 10^{-3}$; 2.4A). This correlation was also the strongest compared to other features. Curiously, we observed a diminishing return with opening energies lower than that of the wild-type sequence (11.68 kcal/mol). To investigate this, we simulated a protein production experiment by modelling cell growth, transcription, translation, and turnovers (see Methods). We assumed that opening energies of 12 kcal/mol or below is favourable

in this model, based on our analysis of 8,780 PSI:Biology 'success' group (Supplementary Fig A.10). Interestingly, our in silico coarse-grained model shows a similar protein production trend as the actual experiment (2.4B).

We then tested this finding using the luciferase reporter of *Renilla reniformis* (RLuc). Similarly, we designed a series of RLuc variants, but with opening energies below that of the wild-type sequence (5.77 to 10.38 kcal/mol; Supplementary Fig S13 and Table S5). In addition, we tested commercially designed sequences, in which sequence optimisations were performed in full-length rather than the first 9 codons. However, RLuc is poorly soluble in the BL21Star(DE3) *E. coli* host (Supplementary Fig A.13B). We observed that TIsigner (9.9 kcal/mol) and commercially optimised luciferase reporter genes produced significantly higher luminescence than the wild-type. We also found that the levels of wild-type luciferase and many variants with lower opening energies (5-7 kcal/mol) were not significantly different, likely due to poor solubility.

As both wild-type GFP and RLuc genes are strongly expressed in *E. coli*, we asked whether poorly expressed proteins can be improved by increasing accessibility of translation initiation sites. We performed densitometric analysis of previously published Western blots, which include the results of a cell-free expression system using constructs harbouring a wild-type antibody fragment or archaebacterial dioxygenase and its synonymous variants (within the first six codons) [253]. Indeed, variants with opening energies lower than the wild-type sequences were expressed at higher levels (Supplementary Fig A.14).

2.4 Discussion

Our findings show that the accessibility of translation initiation sites is the strongest predictor of heterologous protein expression in *E. coli*. However, protein expression is inherently noisy due to the interplay of many cellular processes. At the transcript level, many mRNA features are not truly independent, which aggravates the problem of identifying the key features. As such, a careful design of experiments such as using factorial methods for generating mRNA sequences is crucial for a complete traversal

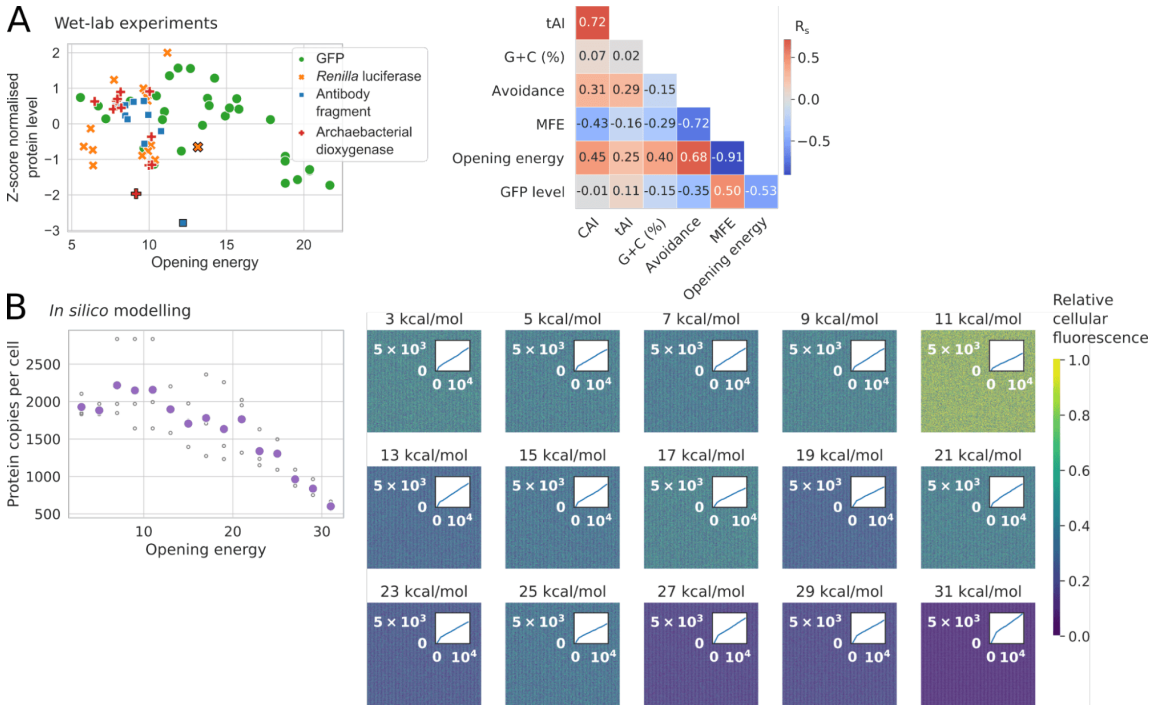


Figure 2.4: **The yields of heterologous protein productions are tunable by synonymous codon changes in the first nine codons. (A)** GFP level strongly correlates with accessibility, i.e., anti-correlates with opening energy ($R_s = -0.53$, $P = 3.4 \times 10^{-3}$; N=29). This correlation is the strongest compared to other features. The protein levels of GFP, *Renilla* luciferase (RLuc), an antibody fragment and an archaebacterial dioxygenase were transformed using z-score method. The GFP and RLuc levels were derived from the average values of at least two and three independent biological replicates, respectively. Black outlines denote wild-type sequences. See also Supplementary Fig A.13, A.14 and Table S5. CAI, codon adaptation index; CC, codon context; Rs, Spearman's rho; tAI, tRNA adaptation index. **(B)** Coarse-grained simulation of a protein production experiment by modelling cell growth, transcription, translation, and turnovers, given that translation initiation sites with opening energies less than or equal to 12 kcal/mol is optimum. The in silico model shows a similar trend of protein production as the wet-lab experimental results. Unfilled and filled (purple) circles denote the in silico replicates and their corresponding average values, respectively ($R_s = -0.75$, $P = 2.8 \times 10^{-9}$). Similar to an actual recombinant production experiment, the in silico model also shows that efficient protein production (higher relative cellular fluorescence) leads to slower cell growth and vice versa (right, see insets for the opening energies of 11 kcal/mol versus 31 kcal/mol).

of the feature landscape. Due to the large-scale nature of such designs, to-date few attempts have been made, e.g., 244,000 GFP and 86 firefly luciferase synonymous variants tested in *E. coli* and HeLa cells, respectively [35, 156]. These fluorescence reporter studies concluded that MFE was the best predictor (Fig 1, Spearman's correlations of 0.51 and 0.46 in *E. coli* and *S. cerevisiae*, respectively). These modest correlations reflect the noisiness of the system which further poses a problem for obtaining a better predictor. Furthermore, MFE estimation involves identifying the thermodynamically most probable structure from Boltzmann's ensemble, which is often inaccurate in a biological system where different constraints may prevent a mRNA from attaining the most probable conformation. With this in mind, we used opening energy, an accessibility-based approach that takes the full ensemble average energy into account. This includes all possible RNA structures, including suboptimal structures that are not reported by MFE models by default [163, 15]. Indeed, our approach gave us a better correlation from multiple datasets where MFE was previously concluded to be a better predictor. We have shown that accessibility is superior to MFE even for the datasets without factorial designs such as the PSI:Biology dataset (2.3), where the feature space is sampled irregularly, and the expression levels of recombinant proteins were categorised into 'Tested_Not_Found' and 'Protein_Confirmed' with SDS-PAGE analysis (Supplementary Fig S7, 11,430 proteins from over 189 diverse species). Moreover, the correlation between endogenous mRNA and protein levels is also limited in both bacteria and eukaryotes (0.4-0.7) [231], where theoretically mRNA levels should provide an upper-bound on correlation statistics of mRNA features. Besides mRNA level, accessibility is a sequence feature that explains most of the variation in protein abundance. Any further improvements in correlations are likely to be hindered by the noise and encountered diminishing returns.

Previous studies have largely used minimum free energy models to define the accessibility of a region of interest [203, 23, 171, 253, 182]. However, Terai and Asai (2020) and ourselves have independently discovered that the opening energy is a better choice for modelling accessibility [21, 235] (see 2.1A for example). Currently, the modelling of accessibility using opening energy is largely used for the prediction

of RNA-RNA intermolecular interactions, for example, as implemented in RNAup and IntaRNA [146, 153]. Our study has shown that this approach can be used to identify the key accessibility regions that are consistent across multiple large expression datasets. We have implemented our findings in TIsigner web server, which currently supports recombinant protein expression in *E. coli* and *S. cerevisiae* (optimisation regions $-24:24$ and $-7:89$, respectively; see 2.1). An independent yet similar implementation is available in XenoExpressO web server with the purpose of optimising protein expression for an *E. coli* cell-free system [278]. The authors showed that an increase in accessibility of a 30 bp region from the Shine-Dalgarno sequence enhances the expression level of human voltage dependent anion channel, which further supports our findings.

The strengths of our approaches are five-fold. Firstly, the likelihood of success or failure can be assessed prior to running an experiment. Users can compare the opening energies calculated for the input and optimised sequences and the distributions of the 'success' and 'failure' of the PSI:Biology targets. We also introduced a scoring scheme to score the input and optimised sequences based upon how likely they are to be expressed (Supplementary Fig A.12; see also Methods). Secondly, optimised sequences can have up to the first nine codons substituted (by default), meaning that gene optimisation using a standard PCR cloning method is feasible. For cloning, we propose a nested PCR approach, in which the final PCR reaction utilises a forward primer designed according to the optimised sequence [204] (Supplementary Fig A.10D). Thirdly, the cost of gene optimisation can be reduced dramatically as gene synthesis is replaced with PCR using our approach. This enables high-throughput protein expression screening using the optimised sequences, generated at a low cost. Fourthly, tunable expression is possible, i.e. high, intermediate or even low expression $5'$ codon sequences can be designed, allowing for more control over heterologous protein production, as demonstrated by our experiments (2.4). Finally, our fast, lightweight, coarse-grained simulation approach has opened up new avenues to study several aspects of gene expression, such as transcription, translation, cellular growth, and turnovers, which give good proxies to how cellular systems behave.

2.5 Material and methods

2.5.1 Plasmids

Plasmids were constructed using the MIDAS Golden Gate cloning system (see Supplementary Methods, Fig A.1, A.2, A.3, A.4, A.5, and Table A.1, A.2, A.3) [62].

2.5.2 Data

Datasets used in this study are listed in Supplementary Table S4. These include fluorescence reporter expression datasets previously generated using E. coli, Saccharomyces cerevisiae, and Mus musculus cultured cells (Supplementary Fig A.6), and recombinant protein production dataset from the Protein Structure Initiative: Biology (PSI:Biology; Supplementary Fig A.7). Two ribosome profiling libraries previously generated using *E. coli* were retrieved from the Sequence Read Archive (SRR7759806 and SRR7759807) [162].

2.5.3 Sequence features analysis

Representative sequences were chosen using CD-HIT-EST [142, 76]. Minimum free energies, opening energies and avoidance were calculated using RNAfold, RNAplfold and RNAup from ViennaRNA package (version 2.4.11), respectively [104, 163, 16, 28, 146, 15, 147]. RNAfold was run with default parameters. For RNAplfold, subsequences were generated from the input sequences to calculate opening energies (using the parameters -W 210 -u 210). For RNAup, we examined the stochastic interactions between the region 1:30 of each mRNA and 54 non-coding RNAs (using the parameters -b -o). RNAup reports the total interaction between two RNAs as the sum of energy required to open accessible sites in the interacting molecules Gu and the energy gained by subsequent hybridisation Gh[163]. For the interactions between each mRNA and 54 non-coding RNAs, we chose the most stable mRNA:ncRNA pair to report an inappropriate mRNA:ncRNA interaction, i.e. the pair with the strongest hybridisation energy, $(\Delta G_h)_{min}$.

CAI, tAI and Codon Context (CC) were calculated using the reference weights from

Sharp and Li [214], Tuller et al. [239] and Ang et al. [8], respectively. Translation elongation rate was predicted using Iχnos [241] that were trained using the *E. coli* ribosome profiling data (Supplementary Fig A.8). Local G+C contents were also examined (Supplementary Fig A.9).

2.5.4 Coarse-grained simulation

To understand the dynamics between accessibility and protein production, we performed a coarse grained simulation using constructs with increasing opening energy on a simulated cellular system. Despite being less precise than fine grained methods such as ab initio and molecular dynamics, coarse grained simulations often give similar results, with an added advantage of being scalable to very large systems.

To set the simulation, we binned the opening energies between 2 and 32 in intervals of two, with each bin representing a 'reporter plasmid construct' whose opening energy is the mean of the bin. For each construct, the 'technical replicates' were generated by allowing slight variations on the mean opening energy of the bin. This is to model variation between replicates, and the discrepancies between the estimated and the actual opening energies in vivo. For each round of transcription, mRNA copies were randomly generated from 30 to 60 plasmid DNA copies [100, 82, 198]. We chose an optimum opening energy of 12 kcal/mol or less for translation (Supplementary Fig A.10). However, this is probabilistic which occasionally allowed protein production from higher opening energy transcripts. We allowed mRNA to decay probabilistically when a mRNA molecule is translated for more than 10 rounds.

We also set a threshold of protein tolerance to be 1,000,000 copies where the copy numbers of endogenous proteins are usually less than 10,000 [231], beyond which there is a sporadic death of cells. However, in this simulation, the chances of staying viable and reproducing are higher than death, and cells grow steadily. This threshold also simulated random but low cell deaths in the experiment, without setting an extra variable.

To limit the computational complexity, our coarse-grained simulations used lower constants and iterations. Initialising with 100 cells, the algorithm was set to ter-

minate either after 10,000 iterations or when the total number of cells becomes zero. After termination, the total number of proteins and cells for each construct were taken from the endpoints. To imitate 'biological replicates', we repeated the above simulation three times with different random numbers, which provides slightly different initial conditions for each experiment.

2.5.5 Development of Translation Initiation coding region designer (TIsigner)

Finding a synonymous sequence with a maximum accessibility is a combinatorial problem that spans a vast search space. For example, for a protein-coding sequence of nine codons, assuming an average of 3 synonymous codons per amino acid, we can expect a total of 19,682 unique synonymous coding sequences. This number increases rapidly with increasing numbers of codons. Heuristic optimisation approaches are preferred in such situations because the search space can be explored more efficiently to obtain nearly optimal solutions.

To optimise the accessibility of a given sequence, TIsigner uses a simulated annealing algorithm [131, 113, 127, 33], a heuristic optimisation technique based on the thermodynamics of a system settling into a low energy state after cooling. Simulated annealing algorithms have been used to solve many combinatorial optimisation problems in bioinformatics. For example, we previously applied this algorithm to align and predict non-coding RNAs from multiple sequences [144]. Other studies use this algorithm to find consensus sequences [127], optimise ribosome binding sites [203] and predict mRNA foldings [81] using minimum free energy models.

In our implementation, each iteration consists of a move that may involve multiple synonymous codon substitutions. The algorithm begins at a high temperature where the first move is drastic, synonymous substitutions occur in all replaceable codons. At the end of the first iteration, a new sequence is accepted if the opening energy is smaller than that of the input sequence. However, if the opening energy of a new sequence is greater than that of the input sequence, acceptance depends on the Metropolis-Hastings criteria. The accepted sequence is used for the next iteration,

which repeats the above process. As the temperature cools, the moves get milder with fewer synonymous codon changes (Supplementary Fig A.10). Simulated annealing stops upon reaching a near-optimum solution.

For the web version of TIsigner, the default number of replaceable codons is restricted to the first nine codons. However, this default setting can be reset to range from the first four to nine codons, or the full length of the coding sequence. Since the accessibility of a fixed region is optimised, this process only takes O(1) time (Supplementary Fig A.11). Furthermore, TIsigner runs multiple simulated annealing instances, in parallel, to obtain multiple possible sequence solutions.

When users select T7lac promoter as the 5′ UTR, they can adjust 'Expression Score', that is calculated based on the PSI:Biology dataset (see below). This allows them to tune the expression level of a target gene. In contrast, when users input a custom 5′ UTR sequence, they only have the option to either maximise or minimise expression.

To implement 'Expression Score', the posterior probabilities of success for input and optimised sequences are evaluated using the following equations from Bayesian statistics:

$$positive\ posterior\ odds = prior\ odds \times fitted\ positive\ likelihood\ ratio$$

$$positive\ posterior\ probability = \frac{positive\ posterior\ odds}{1 + positive\ posterior\ odds}$$

The fitted positive likelihood ratios were obtained from the following 4-parametric logistic regression equation:

$$fitted\ positive\ likelihood\ ratio = d + \frac{(a-d)}{1 + (\frac{positive\ likelihood\ ratio}{c})^b}$$

with parameters a, b, c, and d. The prior probability was set to 0.49, which is the proportion of 'Expressed' (N=21,046) divided by 'Cloned' (N=42,774) of the PSI:Biology targets reported as of 28 June 2017 (http://targetdb.rcsb.org/metrics/). Posterior probabilities were scaled as percentages to score the input and optimised sequences (Supplementary Fig A.12).

The presence of terminator-like elements [44] in the protein-coding region may result in expression of truncated mRNAs due to early transcription termination. Therefore, we implemented an optional check for putative terminators in the input and optimised sequences by cmsearch (INFERNAL version 1.1.2) [167] using the covariance models of terminators from RMfam [80, 120]. We also allow users to filter the output sequences for the presence of restriction sites. Restriction modification sites (AarI, BsaI, and BsmBI) are avoided by default.

Besides *E. coli*, users can choose *S. cerevisiae*, *M. musculus* or 'Other' as the expression host. The regions for optimising accessibility are $-7:89$, $-8:11$ and $-24:89$ for *S. cerevisiae*, *M. musculus* and 'Other', respectively (2.1 and Supplementary Fig A.6). When users choose 'Custom' for expression host, the region for optimising accessibility becomes customisable.

2.5.6 Sequence optimisation

We submitted the PSI:Biology targets that failed to be expressed (N=2,650) to the ExpOptimizer web server from NovoPro Bioscience (https://www.novoprolabs.com/tools/codon-optimization). A total of 2,573 sequences were optimised. The target sequences were also optimised using a local version of COOL [48] and TIsigner using default settings. We also ran a random synonymous codon substitution as a control for these 2,573 sequences.

2.5.7 GFP assay

BL21(DE3)pLysS competent *E. coli* cells (Invitrogen) were transformed with plasmids and grown overnight on Luria-Bertani (LB) agar plates containing spectinomycin (50 $\mu g/ml$) and chloramphenicol (25 $\mu g/ml$). Single colonies were picked and inoculated into 3 ml LB broth containing the same antibiotics, and cultures were grown for 18 hours at 37°C, 200 rpm. Cultures were diluted with fresh media at 1:20 and grown at 37°C, 200 rpm, until reaching the mid-logarithmic growth phase (optical densities at 600 nm (OD_{600}) of 0.3). Of each culture, 20 μl was seeded into 96-well plates containing 180 μl LB broth supplemented with antibiotics and

isopropyl-β-D thiogalactopyranoside (IPTG) (1 mM final concentration) per well. Fluorescence intensities and ODs were measured in a black, flat, clear bottom 96-well plate with lid (CELLSTAR, Greiner) using a FLUOstar Omega plate reader (BMG Labtech) equipped with an excitation filter (band pass 485-12) and an emission filter (band pass 520) for GFP and excitation filter (band pass 484) and an emission filter (band pass 610-10) for mScarlet-I. The plate was incubated at 37°C with "meander corner well shaking" at 300 rpm for 7 hours measuring fluorescence and ODs every 10 minutes. Fluorescence was measured in a 2 mm circle recording the average of 8 measurements per well. Average values of technical replicates were calculated and normalised to the mScarlet-I second reporter, and then to the normalised value of the GFP variant with the highest opening energy (21.68 kcal/mol). Normalised fluorescence values were obtained from the average values of biological replicates (Supplementary Fig A.13 and Table S5).

2.5.8 Luciferase assay

BL21Star(DE3) competent cells (Invitrogen) were transformed with plasmids and grown overnight at 37°C on LB agar plates containing 50 $\mu g/ml$ spectinomycin. Single colonies were picked and inoculated into 5 ml LB broth (50 $\mu g/ml$ spectinomycin) and grown for 18 hours at 37°C, 200 rpm. Bacterial cultures were diluted with fresh media at 1:20 and grown at 37°C, 200 rpm, up to a mid-logarithmic phase (OD_{600} of 0.4). The cultures were split and induced with IPTG at a final concentration of 0.25 mM (or uninduced as controls), and seeded into a white, flat, clear bottom 96-well white plate with lid (Costar, Corning), 150 μl per well, in triplicates. Cells were incubated in a FLUOstar Omega Microplate Reader (BMG LABTECH) for 90 minutes at 25°C, 200 rpm, and OD_{600} was measured every 15 minutes (over 7 cycles). Cells were harvested by centrifugation at $3000 \times g$, for 10 minutes, at 20°C. Supernatants were removed. As the substrate can penetrate into cells, 50 μl of coelenterazine h (Promega) was added to living cells to minimise sample processing steps and variability [149, 77]. Luminescence was measured ($\lambda_{em} = 475nm$) in a Clariostar microplate reader (BMG LABTECH) at 25°C every 2 minutes (over 11 cycles). Average values of technical replicates were calculated and normalised to the

wild-type. Normalised luminescence values were obtained from the average values of biological replicates (Supplementary Fig A.14 and Supplementary Table S5).

2.5.9 Statistical analysis

AUC and Gini importance scores were calculated using scikit-learn (version 0.20.2) [181]. The 95% confidence intervals for AUC scores were calculated using DeLong's method [56]. Spearman's correlation coefficients and Kolmogorov-Smirnov statistics were calculated using Pandas (version 0.23.4) [159] and scipy (version 1.2.1) [177, 160], respectively. Positive likelihood ratios with 95% confidence intervals were calculated using the bootLR package [154, 190]. The P-values of multiple testing were adjusted using Bonferroni's correction and reported to machine precision. Plots were generated using Matplotlib (version 3.0.2) [110] and Seaborn (version 0.9.0) [261].

2.5.10 Code and data availability

Our code and data can be found in our GitHub repository (https://github.com/Gardner-BinfLab/TIsigner_paper_2019). These include the scripts and Jupyter notebooks to reproduce our results and figures. The source code of TIsigner is available at https://github.com/Gardner-BinfLab/TISIGNER-ReactJS. The public web version of this tool runs at https://tisigner.com/tisigner. The experimental data, analysis and results are available at https://github.com/bkb3/TIsignerExperiment/tree/master/Jupyter and an interactive version of results are available at https://bkb3.github.io/TIsignerExperiment/.

Chapter 3

Solubility-Weighted Index: fast and accurate prediction of protein solubility

This chapter is adapted from a paper published in *Bioinformatics* [19]. Dr Chun Shen Lim found that protein structural flexibility is potentially the best predictor of solubility in the PSI:Biology dataset and conceived the study. He then compared flexibility with 9,913 protein features calculated using the 'protr' R package, and benchmarked the existing solubility prediction tools [Fig B.2, 3.4, B.1, Table 3.1 B.2].

I derived the Solubility-Weighted Index from flexibility using Nelder-Mead algorithm. I did all the remaining analysis and figures, and developed SoDoPE (both the command line https://github.com/Gardner-BinfLab/SoDoPE_paper_2020/tree/master/SWI and the web server https://tisigner.com/sodope). I drafted the paper [19]. Dr Lim and Associate Professor Paul Gardner supervised the study.

3.1 Abstract

Motivation: Recombinant protein production is a widely used technique in the biotechnology and biomedical industries, yet only a quarter of target proteins are soluble and can therefore be purified.

Results: We have discovered that global structural flexibility, which can be modeled

by normalised B-factors, accurately predicts the solubility of 12,216 recombinant proteins expressed in *Escherichia coli*. We have optimised these B-factors, and derived a new set of values for solubility scoring that further improves prediction accuracy. We call this new predictor the 'Solubility-Weighted Index' (SWI). Importantly, SWI outperforms many existing protein solubility prediction tools. Furthermore, we have developed 'SoDoPE' (Soluble Domain for Protein Expression), a web interface that allows users to choose a protein region of interest for predicting and maximising both protein expression and solubility.

Availability: The SoDoPE web server and source code are freely available at https://tisigner.com/sodope and https://github.com/Gardner-BinfLab/TISIGNER-ReactJS, respectively. The code and data for reproducing our analysis can be found at https://github.com/Gardner-BinfLab/SoDoPE_paper_2020.

3.2 Introduction

High levels of protein expression and solubility are two major requirements of successful recombinant protein production [70]. However, recombinant protein production is a challenging process. Almost half of recombinant proteins fail to be expressed and half of the successfully expressed proteins are insoluble (http://targetdb.rcsb.org/metrics/). These failures hamper protein research, with particular implications for structural, functional and pharmaceutical studies that require soluble and concentrated protein solutions [132, 107]. Therefore, solubility prediction and protein engineering for enhanced solubility is an active area of research. Notable protein engineering approaches include mutagenesis, truncation (i.e., expression of partial protein sequences), or fusion with a solubility-enhancing tag [254, 70, 237, 41, 132, 52].

Protein solubility, in part, depends upon extrinsic factors such as ionic strength, temperature and pH, as well as intrinsic factors—the physicochemical properties of the protein sequence and structure, including molecular weight, amino acid composition, hydrophobicity, aromaticity, isoelectric point, structural propensities and

the polarity of surface residues [265, 46, 232, 60]. Many solubility prediction tools have been developed around these features using statistical models (e.g., linear and logistic regression) or other machine learning models (e.g., support vector machines and neural networks) [102, 88, 94, 221, 95, 270, 274].

In this study, we investigated the experimental outcomes of 12,216 recombinant proteins expressed in *Escherichia coli* from the 'Protein Structure Initiative:Biology' (PSI:Biology) [43, 2]. We showed that protein structural flexibility is more accurate than other protein sequence properties in solubility prediction [250, 53]. Flexibility is a standard feature appears to have been overlooked in previous solubility prediction attempts. On this basis, we derived a set of 20 values for the standard amino acid residues and used them to predict solubility. We call this new predictor the 'Solubility-Weighted Index' (SWI). SWI is a powerful predictor of solubility, and a good proxy for global structural flexibility. In addition, SWI outperforms many existing *de novo* protein solubility prediction tools.

3.3 Results

3.3.1 Global structural flexibility performs well at predicting protein solubility

We sought to understand what makes a protein soluble, and develop a fast and accurate approach for solubility prediction. To determine which protein sequence properties can accurately predict protein solubility, we analysed 12,216 target proteins from over 196 species that were expressed in *E. coli* [43, 2] (the PSI:Biology dataset; see Supplementary Fig B.1 and Table S1A). These proteins were expressed either with a C-terminal or N-terminal polyhistidine fusion tag (pET21_NESG and pET15_NESG expression vectors, N = 8,780 and 3,436, respectively). The protein entries were previously curated and classified as 'Protein_Soluble' or 'Tested_Not_Soluble' [213], based on the soluble analysis of cell lysate using SDS-PAGE [273]. Both the expression system and solubility analysis method are routinely used

in the labs [52]. This large collection of dataset captures a wide variety of protein solubility issues.

We evaluated nine standard and 9,920 miscellaneous protein sequence properties using the Biopython's ProtParam module and 'protr' R package, respectively [50, 272]. For example, the standard properties include the Grand Average of Hydropathy (GRAVY), secondary structure propensities, protein structural flexibility, etc., whereas miscellaneous properties include amino acid composition, autocorrelation, etc. Strikingly, protein structural flexibility outperformed other features in solubility prediction [Area Under the ROC Curve (AUC) = 0.67; Fig 3.1, Supplementary Fig B.2 and Table S2].

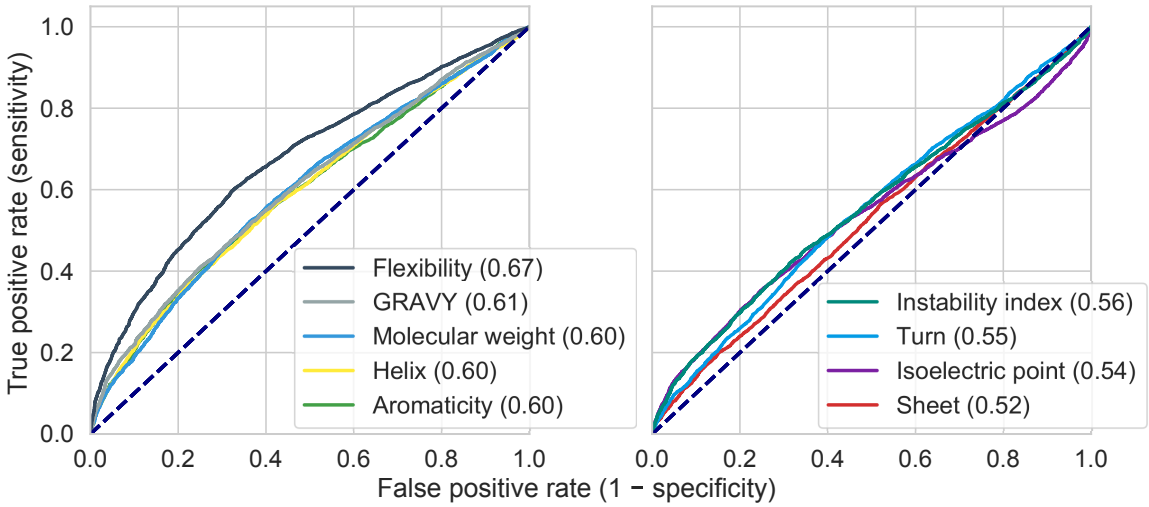


Figure 3.1: **Global structural flexibility outperforms other standard protein sequence properties in protein solubility prediction**. ROC analysis of the standard protein sequence features for predicting the solubility of 12,216 recombinant proteins expressed in *E. coli* (the PSI:Biology dataset). The ROC curves are shown in two separate panels for clarity. AUC scores (perfect = 1.00, random = 0.50) are shown in parentheses. Dashed lines denote the performance of random classifiers. See also Supplementary Fig B.2 and Table S2. AUC, Area Under the ROC Curve; GRAVY, Grand Average of Hydropathy; PSI:Biology, Protein Structure Initiative:Biology; ROC, Receiver Operating Characteristic.

3.3.2 The Solubility-Weighted Index (SWI) is an improved predictor of solubility

Protein structural flexibility, in particular, the flexibility of local regions, is often associated with function [53]. The local flexibility of an amino acid residue i can be written as:

$$\begin{aligned} f_i = \frac{1}{5.25} \times [B_i + 0.8125(B_{i-1} + B_{i+1}) + \\ 0.625(B_{i-2} + B_{i+2}) + \\ 0.4375(B_{i-3} + B_{i+3}) + \\ 0.25(B_{i-4} + B_{i+4})] \end{aligned} \tag{3.1}$$

where B_i denotes the normalised B-factor of amino acid residue i. These normalised B-factors were previously derived from the B-factors extracted from protein crystal structures [124, 193, 250, 220] (see also Supplementary Notes). These normalised B-factors can be applied to any protein sequences without crystallographic data for flexibility prediction, for example as implemented in Biopython.

To predict global protein structural flexibility F (as in Fig 3.1), F can be calculated as the sliding window average of normalised B-factors (i.e., the arithmetic mean of f_i) [250, 50].

$$F = \langle f_i \rangle \tag{3.2}$$

Therefore, we can simplify Equation 3.1 by setting $f'_i = B_i$ like a zeroth order Markov model. The simplified global flexibility F' is then the arithmetic mean of normalised B-factors (see Supplementary Notes B.1 for mathematical proof).

$$F' = \langle f'_i \rangle = \langle B_i \rangle \tag{3.3}$$

We found a strong correlation between F and F' for the PSI:Biology dataset (Spearman's rho = 0.98, P-value below machine's underflow level). Hence, the sliding window approach (Equation 3.1 and 3.2) is not necessary for this purpose.

We applied this arithmetic mean approach (i.e., sequence composition scoring) to the PSI:Biology dataset using four sets of previously published, normalised B-factors [22,

193, 250, 220]. Among these sets of B-factors, sequence composition scoring using the most recently published set of normalised B-factors produced the highest AUC score (AUC = 0.66; Supplementary Fig B.3).

To improve the prediction accuracy of solubility, we iteratively refined the weights of amino acid residues using the Nelder-Mead optimisation algorithm [168] (Fig 3.2). Smith *et al.*'s normalised B-factors were used as initial weights. To avoid testing and training on similar sequences, we generated 10 cross-validation sets with a maximised heterogeneity between these subsets (i.e. no similar sequences between subsets). We clustered all 12,216 PSI:Biology protein sequences by a 40% similarity threshold using USEARCH to produce 5,050 clusters with remote between-cluster similarity (see Methods and Supplementary Fig B.4). The clusters were grouped into 10 cross-validation sets of approximately 1,200 sequences each. As about 12% of clusters contain a mix of soluble and insoluble proteins, we avoided selecting a representative sequence for each cluster (Supplementary Fig B.4C). Furthermore, to avoid overfitting due to sequence similarity and imbalanced classes, we performed 1,000 bootstrap resamplings for each cross-validation step (Fig 3.2A) and Supplementary Fig B.5). We calculated the solubility scores using the optimised weights and the AUC scores for each cross-validation step as shown Fig 3.2A. Our training and test AUC scores were 0.72 ± 0.00 and 0.71 ± 0.01, respectively, showing a 7.5% improvement over flexibility in solubility prediction (mean $\pm$ standard deviation; Fig 3.2B and Supplementary Table B.3).

The final weights were derived from the arithmetic means of the weights for individual amino acid residues obtained from cross-validation (Fig 3.2 and Supplementary Table B.4). We observed over a 20% change on the weights for cysteine (C) and histidine (H) residues (Fig 3.2C and Supplementary Table B.4). These results are in agreement with the contributions of cysteine and histidine residues as shown in Supplementary Fig B.2B . We call the solubility score of a protein sequence calculated using the final weights the Solubility-Weighted Index (SWI):

$$SWI = \langle W_i \rangle \tag{3.4}$$

where W_i is the optimised weight of residue i.

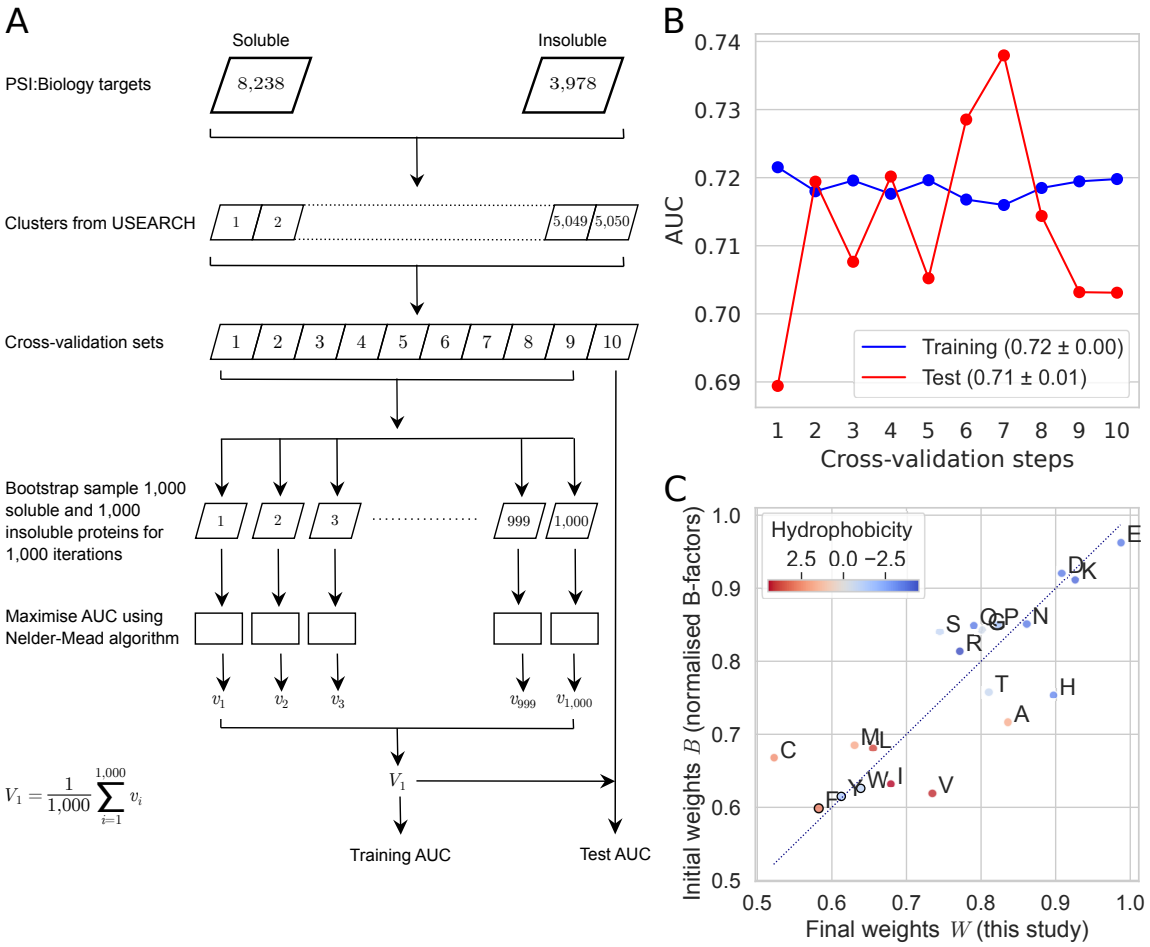


Figure 3.2: **Derivation of the Solubility-Weighted Index (SWI). (A)** Flow chart shows an iterative refinement of the weights of amino acid residues for solubility prediction. Each cross-validation step used separate sequence similarity clusters for training and testing. Furthermore, bootstrapping was used to resample each training set, avoiding training and testing on similar sequences. The solubility scores of protein sequences were calculated using a sequence composition scoring approach. These scores were used to compute the AUC scores for training and test datasets. **(B)** Training and test performance of solubility prediction using optimised weights for 20 amino acid residues in a 10-fold cross-validation (mean AUC ± standard deviation). Related data and figures are available as Supplementary Table B.3 and Supplementary Fig B.4 and B.5 . **(C)** Comparison between the 20 initial and final weights for amino acid residues. The final weights $W = \langle V_i \rangle$, $1 \leq i \leq 10$, were used to calculate the solubility score of a protein sequence (SWI) in the subsequent analyses. Filled circles, which represent amino acid residues, are colored by hydrophobicity [136]. Solid black circles denote aromatic residues phenylalanine (F), tyrosine (Y), tryptophan (W). Dotted diagonal line represents no change in weight. See also Supplementary Table B.4. AUC, Area Under the ROC Curve; ROC, Receiver Operating Characteristic.

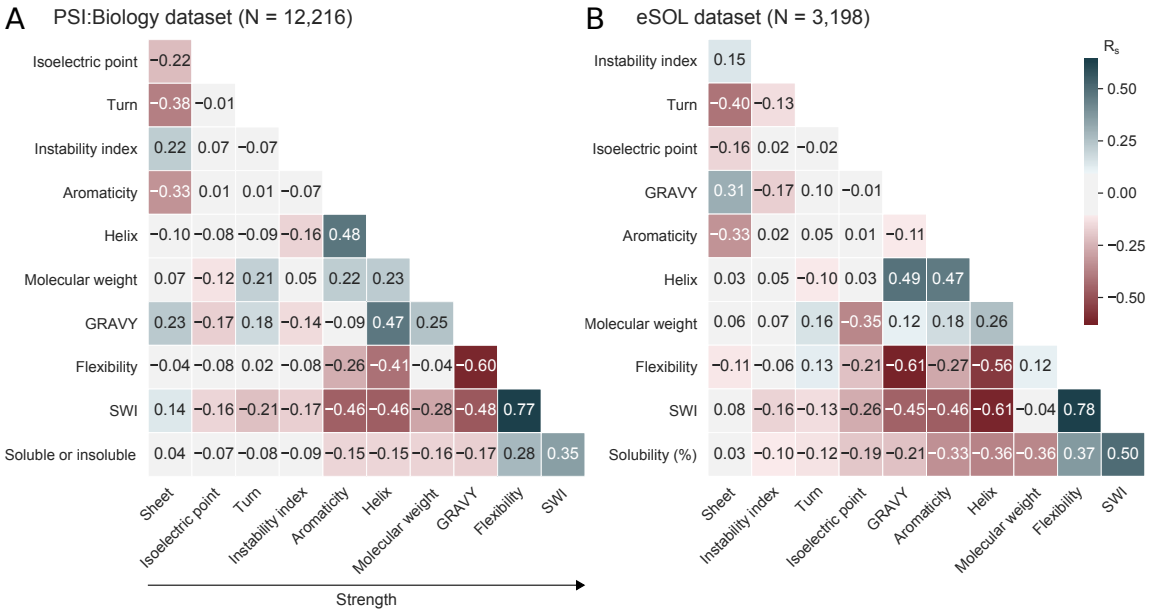


Figure 3.3: **SWI strongly correlates with protein solubility. (A)** Correlation matrix plot of the solubility of recombinant proteins expressed in *E. coli* and their standard protein sequence properties and SWI. These recombinant proteins are the PSI:Biology targets (N = 12, 216) with a binary solubility status of 'Protein_Soluble' or 'Tested_Not_Soluble'. Related data is available as Supplementary Table B.5. **(B)** Correlation matrix plot of the solubility percentages of *E. coli* proteins and their standard protein sequence properties and SWI. The solubility percentages were previously determined using an *E. coli* cell-free system (eSOL, N = 3, 198). Related data is available as Supplementary Table B.6. GRAVY, Grand Average of Hydropathy; PSI:Biology, Protein Structure Initiative:Biology; R_s, Spearman's rho; SWI, Solubility-Weighted Index.

To validate the cross-validation results, we used a dataset independent of the PSI:Biolgy known as eSOL [174] (Supplementary Table B.1). This dataset consists of the solubility percentages of *E. coli* proteins determined using an *E. coli* cell-free system (N = 3,198). Our solubility scoring using the final weights showed a significant improvement in correlation with *E. coli* protein solubility over the initial weights (Smith *et al.*'s normalised B-factors) [Spearman's rho of 0.50 ($P = 2.51 \times 10^{-205}$) versus 0.40 ($P = 4.57 \times 10^{-120}$)]. We repeated the correlation analysis by removing extra amino acid residues including His-tags from the eSOL sequences (MRGSHHHHHHT-DPALRA and GLCGR at the N- and C-termini, respectively). This artificial dataset was created based on the assumption that His-tags have little effect on solubility. We observed a slight decrease in correlation for this artificial dataset (Spearman's rho = 0.47, $P = 3.67 \times 10^{-176}$), which may be due to the effects of His-tags in solubility and/or the limitation(s) of our approach that may overfit to His-tag fusion proteins.

We performed Spearman's correlation analysis for both the PSI:Biology and eSOL datasets. SWI shows the strongest correlation with solubility compared to the standard and 9,920 miscellaneous sequence properties (Fig 3.3 and Supplementary Fig B.2 , respectively; see also Supplementary Table B.2, B.5, B.6). SWI strongly correlates with flexibility, suggesting that SWI is also a good proxy for global structural flexibility.

We asked whether protein solubility can be predicted by surface amino acid residues. To address this question, we examined a previously published dataset for the protein surface 'stickiness' of 397 *E. coli* proteins [140]. This dataset has the annotation for surface residues based on previously solved protein crystal structures. We observed little correlation between the protein surface 'stickiness' and the solubility data from eSOL (Spearman's rho = 0.05, P = 0.34, N = 348; Supplementary Fig B.6A). Next, we evaluated if amino acid composition scoring using surface residues is sufficient, in which optimising only the weights of surface residues should achieve similar or better results than SWI. As above, we iteratively refined the weights of surface residues using the Nelder-Mead optimisation algorithm. The method was initialised with Smith *et al.*'s normalised B-factors and a maximised correlation coefficient was the target. However, a low correlation was obtained upon convergence (Spearman's

rho = 0.18, P = 7.20×10^{-4}; Supplementary Fig B.6B). In contrast, the SWI of the full-length sequences has a much stronger correlation with solubility (Spearman's rho = 0.46, P = 2.97×10^{-19}; Supplementary Fig B.6C). These results show that the full-length of sequences contributes to protein solubility, not just surface residues, suggesting that solubility is modulated by cotranslational folding [55, 166].

To understand the properties of soluble and insoluble proteins, we determined the enrichment of amino acid residues in the PSI:Biology targets relative to the eSOL sequences (see Methods). We observed that the PSI:Biology targets are enriched in charged residues lysine (K), glutamate (E) and aspartate (D), and depleted in aromatic residues tryptophan (W), albeit to a lesser extend for insoluble proteins (Supplementary Fig B.7A). As expected, cysteine residues (C) are enriched in the PSI:Biology insoluble proteins, supporting previous findings that cysteine residues contribute to poor solubility in the *E. coli* expression system [265, 60].

In addition, we compared the distributions of the SWI scores of soluble and insoluble proteins in the PSI:Biology and eSOL datasets. We included an analysis of random sequences to confirm whether SWI can distinguish between biological and random sequences. In general, the SWI scores of soluble proteins are higher than those of insoluble proteins (Supplementary Fig B.7B), and the SWI scores of true biological sequences are higher than those of random sequences, addressing our concern about the potential flaw of this position independent, sequence composition scoring approach.

3.3.3 SWI outperforms many protein solubility prediction tools

To confirm the usefulness of SWI in solubility prediction, we compared SWI with the existing tools CamSol v2.1 [222, 221], ccSOL omics [3], DeepSol v0.3 [128], PaRSnIP [195], Protein-Sol [94], and the Wilkinson-Harrison model [265, 55, 92]. We did not include the specialised tools that model protein structural information such as surface geometry, surface charges and solvent accessibility because these tools require prior knowledge of protein tertiary structure. For example, Aggrescan3D and SOLart

Table 3.1: Comparison of protein solubility prediction methods and software.

	Approaches	Features	Wall time (s per sequence)[a]	PSI:Biology (AUC)[b]	eSOL [R_s (P-value)]
SWI	Arithmetic mean (this study). Sequence composition scoring using a set of 20 values for amino acid residues derived from Smith *et al.*'s normalised B-factors. Trained and tested using the PSI:Biology dataset curated by DNASU [213]. Available at https://tisigner.com/sodope and https://github.com/Gardner-BinfLab/SoDoPE_paper_2020	1	**0.00 ± 0.00**	**0.71 ± 0.01**	0.50 (2.51×10^{-205})
Protein-Sol	Linear model [94]. Trained and tested using eSOL dataset [174]. Available at https://protein-sol.manchester.ac.uk/	10	1.16 ± 0.75	0.68 ± 0.02	**0.54 (2.37×10^{-240})**
Flexibility	A sliding window of 9 amino acid residues using Vihinen *et al.*'s normalised B-factors. Available at https://github.com/biopython/biopython	1	0.38 ± 0.04	0.67 ± 0.02	0.37 (7.73×10^{-106})
DeepSol S2	Neural network models [128][c]. Trained and tested using a PSI:Biology dataset curated by ccSOL omics [3]. Available at https://github.com/sameerkhurana10/DSOL_rv0.2	57 (11 types)	2069.77 ± 1613.63	0.67 ± 0.02	0.23 (5.82×10^{-41})
DeepSol S3			2075.93 ± 1613.80	0.66 ± 0.02	0.35 (7.48×10^{-91})
DeepSol S1			2081.93 ± 1612.71	0.64 ± 0.03	0.39 (9.52×10^{-116})
CamSol intrinsic web server	Linear and logistic regression models (Sormanni *et al* 2015, 2017). Trained and tested using previously published datasets [72]. Available at http://www-vendruscolo.ch.cam.ac.uk/camsolmethod.html	4	NA	0.66 ± 0.01	0.44 (4.53×10^{-148})
PaRSnIP	Gradient boosting machine model [195]. Trained and tested using a PSI:Biology dataset curated by ccSOL omics [3]. Available at https://github.com/RedaRawi/PaRSnIP	8,477 (14 types)	2055.50 ± 1621.11	0.61 ± 0.02	0.29 (3.57×10^{-65})
Wilkinson- Harrison model	Linear model using charge average and turn-forming residue fraction [265, 55, 92]. Available at https://github.com/brunoV/bio-tools-solubility-wilkinson	2	0.09 ± 0.00	0.55 ± 0.03	−0.06 (1.16×10^{-4})
ccSOL omics web server	Support vector machine model [3]. Trained and tested using a PSI:Biology dataset curated in-house. Available at http://s.tartaglialab.com/new_submission/ccsol_omics_file	5	NA	0.51 ± 0.01	−0.02 (0.18)

Boldface values are the best results.
[a]The wall time was reported at the level of machine precision (mean seconds ± standard deviation). A total of 10 sequences were chosen from the PSI:Biology and eSOL datasets, related to Fig 3.4B and Supplementary Table B.7 (see Methods).
[b]For SWI, mean AUC ± standard deviation was calculated from a 10-fold cross-validation (see Methods). For other tools, no cross-validations were done as the AUC scores were calculated directly from the individual subsets used for cross-validation.
[c]DeepSol reports solubility prediction as probability and binary classes. The probability of solubility was used to calculate AUC and Spearman's correlation due to better results.
AUC, Area Under the ROC Curve; NA, not applicable; PDB, Protein Data Bank; PSI:Biology, Protein Structure Initiative:Biology; ROC, Receiver Operating Characteristic; R_s, Spearman's rho; SWI, Solubility-Weighted Index; s, seconds.

accept only PDB files that can be either downloaded from the Protein Data Bank or produced using a homology modeling program [108, 135].

SWI outperforms other tools except for Protein-Sol in predicting *E. coli* protein solubility (Fig 3.4 and Table 3.1). The test AUC scores of SWI were also less variable than most other tools, suggesting that SWI is less prone to overfitting (Fig 3.2 and 3.4A). Our SWI C program is also the fastest solubility prediction algorithm (Fig 3.4B, Table 3.1 and Supplementary Table B.6).

3.4 Discussion

The profile of normalised B-factors along a protein sequence can be used to infer the flexibility and dynamics of the protein structure [124, 250]. Protein structural flexibility has been associated with conformal variations, functions, thermal stability, ligand binding and disordered regions [248, 233, 192, 151, 209, 276, 275]. However, the use of flexibility in solubility prediction has been overlooked although their relationship has previously been noted [238]. In this study, we have shown that flexibility strongly correlates with solubility (Fig 3.3). Based on the normalised B-factors used to compute flexibility, we have derived a new position and length independent weights to score the solubility of a given protein sequence (i.e., sequence composition based score). We call this protein solubility score as SWI.

Upon further inspection, we observe some interesting properties in SWI. SWI anti-correlates with helix propensity, GRAVY, aromaticity and isoelectric point (Fig 3.2 and 3.3), suggesting that SWI incorporates the key propensities affecting solubility. Amino acid residues with a lower aromaticity or hydrophilic are known to improve protein solubility [265, 237, 174, 132, 259, 90]. Consistent with previous studies, the charged residues aspartate (D), glutamate (E) and lysine (K) are associated with high solubility, whereas the aromatic residues phenylalanine (F), tryptophan (W) and tyrosine (Y) are associated with low solubility (Fig 3.2C and Supplementary Fig B.7). Cysteine residue (C) has the lowest weight, probably because disulfide bonds couldn't be properly formed in the *E. coli* expression hosts [226, 9, 198, 117]. The weights are likely different if the solubility analysis was done using the reductase-deficient, *E. coli* Origami host strains, or eukaryotic hosts.

Higher helix propensity has been reported to increase solubility [111, 109]. However, our analysis has shown that helical and turn propensities anti-correlate with solubility, whereas sheet propensity lacks correlation with solubility, suggesting that disordered regions may tend to be more soluble (Fig 3.3). In accordance with these, SWI has stronger negative correlations with helix and turn propensities. Our findings also suggest that protein solubility can be largely explained by overall amino acid composition, not just the surface amino acid residues. This idea aligns with

our understanding that protein solubility and folding are closely linked, and folding occurs cotranslationally, a complex process that is driven various intrinsic and extrinsic factors [265, 55, 46, 232, 60, 166]. However, it is unclear why sheet propensity has little contribution to solubility as β-sheets have been shown to link closely with protein aggregation [111].

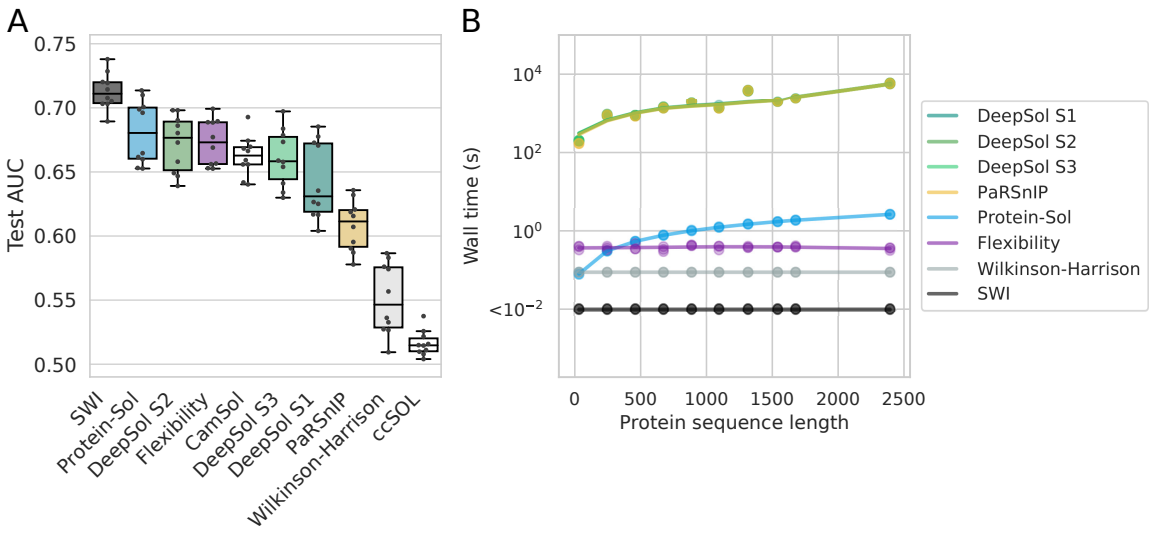


Figure 3.4: **SWI outperforms existing protein solubility prediction tools. (A)** Prediction accuracy of solubility prediction tools using the above cross-validation sets (Fig 3.2A). For SWI, the test AUC scores were calculated from a 10-fold cross-validation (i.e., a boxplot representation of Fig 3.2B). For other tools, no cross-validations were done as the AUC scores were calculated directly from the individual subsets used for cross-validation. CamSol and ccSOL omics are only available as web servers (no fill colors). **(B)** Wall time of protein solubility prediction tools per sequence (log scale). All command line tools were run three times using 10 sequences selected from the PSI:Biology and eSOL datasets. Related data is available as Supplementary Table B.7. AUC, Area Under the ROC Curve; PSI:Biology, Protein Structure Initiative:Biology; ROC, Receiver Operating Characteristic; SWI, Solubility-Weighted Index; s, seconds.

We conclude that SWI is a well-balanced index that is derived from a simple sequence composition scoring method. To demonstrate the usefulness of SWI, we developed a web server called SoDoPE (Soluble Domain for Protein Expression; https://tisigner.com/sodope). SoDoPE calculates the probability of solubility of a user-selected region based on SWI, which can either be a full-length or a partial sequence (see Methods and Supplementary Table B.8). This implementation is based on our observation that some protein domains tend to be more soluble than the others, and these soluble domains may enhance protein solubility as a whole. To

demonstrate this point, we used SoDoPE to analyse three commercial monoclonal antibodies and the proteomes of the severe acute respiratory syndrome coronaviruses (SARS-CoV and SARS-CoV-2) [155, 258, 267] (Supplementary Fig B.8 and B.9). SoDoPE also provides options for solubility prediction at the presence of solubility-enhancing tags. Similarly, these fusion tags may act as soluble 'protein domains' that can outweigh the aggregation propensity of insoluble proteins. However, some soluble fusion proteins may become insoluble after proteolytic cleavage of solubility tags [139]. In addition, SoDoPE is integrated with TIsigner, a web service for optimising protein expression [21]. This pipeline provides a holistic approach to improve the outcome of recombinant protein expression.

3.5 Methods

3.5.1 Data

We retrieved 12,216 PSI:Biology entries from the DNASU database [43, 2, 213]. These proteins were previously expressed in *E. coli* using pET21_NESG or pET15 _NESG expression vectors (N = 8,780 and 3,436, respectively). For validation, we used the solubility data of *E. coli* proteins from eSOL (N = 3,198; http://www.tanpaku.org/tp-esol/index.php?lang=en) [174]. See also Supplementary Fig B.1 and Table S1A.

In addition, we downloaded the 'stickiness' data of 397 *E. coli* proteins to examine the effects of surface amino acid residues (http://www.weizmann.ac.il/Structural_Biology/faculty_pages/ELevy/intDef/interface_def.html) [140].

3.5.2 Protein sequence properties

The standard protein sequence properties were calculated using the Bio.SeqUtils.ProtParam module of Biopython v1.73 [50]. All miscellaneous protein sequence properties were computed using the R package protr v1.6-2 [272].

3.5.3 Protein solubility prediction

We used the standard and miscellaneous protein sequence properties to predict the solubility of the PSI:Biology and eSOL targets. For method comparison, we chose the protein solubility prediction tools that are scalable (Table 3.1). Default configurations were used for running the command line tools.

To benchmark the wall time of solubility prediction tools, we selected 10 sequences that span a large range of lengths from the PSI:Biology and eSOL datasets (from 36 to 2,389 residues). All the tools were run and timed using a single process without using GPUs on a high performance computer [`/usr/bin/time -f '%E' <command>`]; CentOS Linux 7 (Core) operating system, 72 cores in 2 × Broadwell nodes (E5-2695v4, 2.1 GHz, dual socket 18 cores per socket), 528 GiB memory]. Single sequence fasta files were used as input files.

3.5.4 SWI

To improve protein solubility prediction, we optimised Smith *et al.*'s normalised B-factors using the PSI:Biology dataset (Fig 3.2). To avoid including homologous sequences in the test and training sets, we clustered the PSI:Biology targets using USEARCH v11.0.667, 32-bit [68]. His-tag sequences were removed from all sequences before clustering to avoid false cluster inclusions. We obtained 5,050 clusters using the parameters: `-cluster_fast <input_file> -id 0.4 -msaout <output_file> -threads 4`. These clusters were grouped into 10 subsets with approximately 1,200 sequences per subset manually. The subsequent steps were carried out using sequences with His-tags.

We iteratively refined the weights of amino acid residues for solubility scoring using a 10-fold cross-validation, in which a maximised AUC was the target (Fig 3.2A). Since AUC is non-differentiable, we used the Nelder-Mead optimisation method (implemented in SciPy v1.2.0), which is a derivative-free, heuristic, simplex-based optimisation [168, 177, 160]. For each step in cross-validation, we used bootstrap resamplings containing 1,000 soluble and 1,000 insoluble proteins. Optimisation was carried out

for each sample, giving 1,000 sets of weights. The arithmetic mean of these weights was used to determine the training and test AUC for the cross-validation step.

3.5.5 Bit score

To examine the enrichment of amino acid residues in soluble and insoluble proteins, we compute the bit scores for each residue in the PSI:Biology soluble and insoluble groups (Supplementary Fig B.7A). The count of each residue (x) in each group was normalised by the total number of residues in that group. We used the normalised count of amino acid residues using the eSOL *E. coli* sequences as the background. The bit score of residue x for soluble or insoluble group is then given by the following equation:

$$bit\ score(x_i) = log_2 \frac{f_i(x)}{f_{eSOL}(x)}, i = [soluble, insoluble] \tag{3.5}$$

where $f_i(x)$ is the normalised count of residue x in the PSI:Biology soluble or insoluble group and $f_{eSOL}(x)$ is the normalised count in the eSOL sequences.

For a control, random protein sequences were generated with incremental lengths, starting from a length of 50 residues to 6,000 residues with a step size of 50 residues. A hundred random sequences were generated for each length, giving a total of 12,000 unique random sequences.

3.5.6 The SoDoPE web server

To estimate the probability of solubility using SWI, we fitted the following logistic regression to the PSI:Biology dataset:

$$probability\ of\ solubility = \frac{1}{1 + exp(-(ax + b))} \tag{3.6}$$

where, x is the SWI of a given protein sequence, $a = 81.05812$ and $b = -62.7775$. The P-value of log-likelihood ratio test was below machine's underflow level. Equation 3.6 can be used to predict the solubility of a protein sequence given that the protein is successfully expressed in *E. coli* (Supplementary Table B.8).

On this basis, we developed a solubility prediction web service called SoDoPE (Sol-

uble Domain for Protein Expression). Our web server accepts either a nucleotide or amino acid sequence. Upon sequence submission, a query is sent to the HMMER web server to annotate protein domains (https://www.ebi.ac.uk/Tools/hmmer/) [186]. Once the protein domains are identified, users can choose a domain or any custom region (including full-length sequence) to examine the probability of solubility, flexibility and GRAVY. This functionality enables protein biochemists to plan their experiments and opt for the domains or regions with high probability of solubility. Furthermore, we implemented a simulated annealing algorithm that maximised the probability of solubility for a given region by generating a list of regions with extended boundaries. Users can also predict the improvement in solubility by selecting a commonly used solubility tag or a custom tag.

We linked SoDoPE with TIsigner, which is our existing web server for optimising the accessibility of translation initiation site [21]. This pipeline allows users to predict and optimise both protein expression and solubility for a gene of interest. The SoDoPE web server is freely available at https://tisigner.com/sodope.

3.5.7 Statistical analysis

Data analysis was done using Pandas v0.25.3 [159], scikit-learn v0.20.2 [181], numpy v1.16.2 [256] and statsmodel v0.10.1 [211]. Plots were generated using Matplotlib v3.0.2 [39] and Seaborn v0.9.0 [260].

3.5.8 Code and data availability

Jupyter notebook of our analysis can be found at https://github.com/Gardner-BinfLab/SoDoPE_paper_2020. The source code for our solubility prediction server (SoDoPE) can be found at https://github.com/Gardner-BinfLab/TISIGNER-ReactJS.

Chapter 4

Razor: annotation of signal peptides from toxins

Dr Chun Shen Lim conceived the study. I performed all the data analysis and developed Razor (both the command line https://github.com/Gardner-BinfLab/Razor and the web server https://tisigner.com/razor). I drafted the manuscript [18]. Dr Lim and Associate Professor Paul Gardner supervised the study.

4.1 Abstract

Signal peptides are responsible for protein transport and secretion and are ubiquitous to all forms of life. The annotation of signal peptides is important for understanding protein translocation and toxin secretion and evolution. Here we explore the features of these signal sequences from eukaryotic proteins. Strikingly, we find that the signal peptides from secretory toxins share universal features across kingdoms, supporting the idea of horizontal gene transfer or convergence of toxin genes across kingdoms as shown by previous studies. We leverage these features to build Razor, a simple yet powerful tool specialised in identifying signal peptides from toxins using the first 23 N-terminal residues. We demonstrate the usability of Razor by analysing all the sequences reviewed by UniProt. Indeed, Razor is able to identify toxins using their N-terminal sequences only. Interestingly, we also discover that many defensive proteins across kingdoms harbour a toxin-like signal peptide; some of these defensive proteins have been shown to emerge through convergent evolution, e.g. defensin and defensin-like protein families, and phospholipase families. In sum, Razor uses an

approach independent of homology search to identify novel and known toxin classes across species using N-terminal residues. Razor is available as a web application (https://tisigner.com/razor) and a command-line tool (https://github.com/Gardner-BinfLab/Razor).

4.2 Introduction

Secretory proteins are translocated in the secretory pathway with the assistance of a short peptide extension at the N-terminus. This special targeting peptide is known as the Signal Peptide (SP) [99]. Secretory pathways and their corresponding SPs have evolved across organisms to carry out different functions [96, 179]. Despite being ubiquitous across all domains of life, SPs do not share a consensus. Nevertheless, a SP usually consists of three regions: a positively charged domain (N-region), a hydrophobic core (H-region), followed by a polar but electrically neutral domain (C-region) containing a cleavage site [98, 99, 170]. Apart from translocating proteins, SPs are also responsible for several other roles, such as in regulatory functions, antigen presentation, and some human diseases [29, 54, 179].

An important group of secretory proteins is toxins, whose precursors almost always contain SPs [74]. Toxins have evolved in all domains of life primarily as a defense mechanism or for predation [37]. Furthermore, several organisms in the animal kingdom have evolved to create venoms, which consist of a complex mixture of different types of toxins, usually with a specialised apparatus to facilitate their delivery. Such adaptations may have evolved through convergence or duplication and neofunctionalisation [38]. However, a recent study found that at least five toxin gene families were horizontally transferred from bacteria and fungi to centipedes [244], suggesting common features exist in these gene families. Besides, the pharmacological actions of toxins on living cells are often employed to develop anti-toxins, novel drugs, and pathogen-resistant transgenic crops [130, 71, 24, 205, 141]. Hence, annotating SPs is essential in the functional and structural studies of proteins in fundamental research, commercial, and pharmaceutical industries. In addition, understanding the presence or absence of SPs in the genes of interest is critical for choosing the appro-

priate recombinant protein expression and purification systems, as the intracellular accumulation of secretory proteins and toxins may be toxic to the host cells. Indeed, the ability of SPs to translocate proteins has been utilised in recombinant protein expression systems for high quality and quantity results [78, 47, 125, 183].

Despite the immense use cases of toxins, there are very few tools to predict them, such as ClanTox, ToxinPred, TOXIFY, and ToxClassifier, some being specialised such as SpiderP for spider toxins [165, 84, 266, 79, 51]. Moreover, these methods are based on the properties of the mature peptides (or the propeptides), rather than the SPs. To address these issues, we first examined the features of SPs from eukaryotic proteins and toxins. We then exploited those features to build Razor, a new tool for annotating SPs. We have optimised the command-line version of Razor for high-throughput analysis and used it to predict new SPs by scanning all the sequences reviewed by UniProt [245]. We were able to predict novel toxins and defensive proteins using only the first 23 N-terminal residues, as evidenced by the protein family annotations.

4.3 Results

4.3.1 Toxin SPs have distinct sequence properties

We investigated the sequence composition of SPs by first aligning the sequences from the N-terminal residue or by centering at the cleavage sites, followed by computing bit scores for each residue (Fig 4.1). These approaches provide sufficient leverage to enumerate the tripartite domains of SPs (N-, H-, and C-domains). In general, hydrophobic residues are enriched towards the N-termini (H-region), which are characteristic features of SPs [99] (Supplementary Fig C.1). Strikingly, the SPs of toxins show a strong abundance of isoleucine (I) and lack leucine (L) and alanine (A) residues in contrast to other eukaryotic SPs (Fig 4.1). This is supported by an amino acid composition analysis of the N-terminal subsequences (Supplementary Fig C.2). We also analysed other features of these N-terminal subsequences, including GRAVY, structural flexibility, helix, sheet and turn propensities, instability index,

aromaticity, isoelectric point, and SWI. Interestingly, isoelectric point appears as a prominent feature of toxin SPs (Supplementary Fig C.3).

The cleavage sites mark the end of SPs and the beginning of the mature region (or the propeptide), which is a unique feature of SPs (Fig 4.1). By aligning the sequences at the cleavage sites, we observed a clear emergence of (-3,-1) rule preceding the cleavage sites, i.e. a distinctive presence of small and charged residues such as alanine (A) valine (V) and Glycine (G) [97].

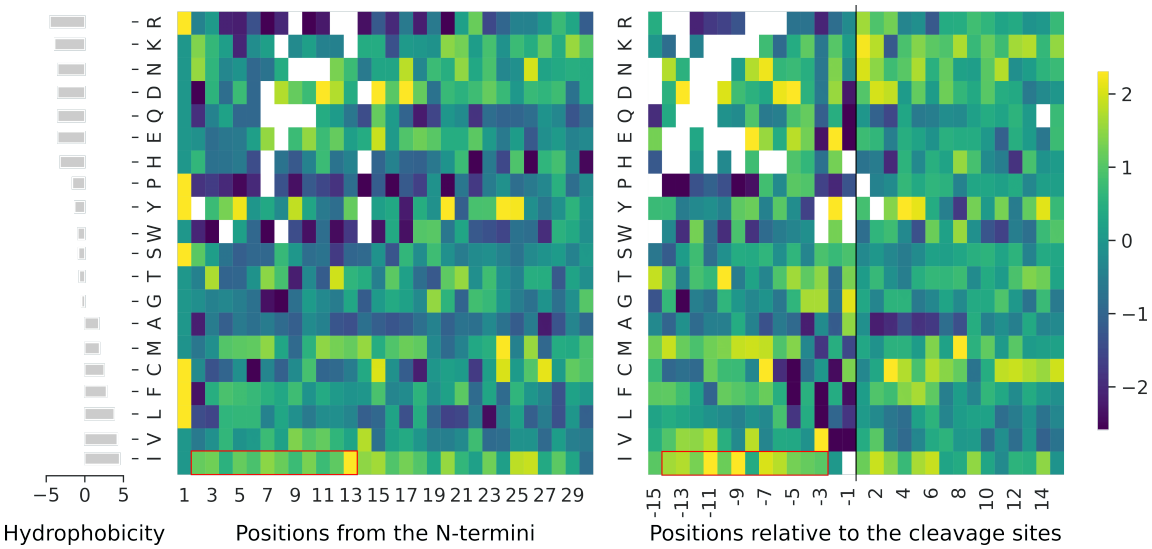


Figure 4.1: **The Signal Peptides (SPs) from toxins are enriched with isoleucine residues in contrast to other eukaryotic SPs.** The bar plot shows Kyte and Doolittle's hydrophobicity scale. The heatmaps show the enrichment of residues in bit scores by aligning SPs from the N-termini (left) and at the cleavage sites (right, black vertical line). The unfilled, red rectangles indicate the enrichment of isoleucine residues (I). The white spaces correspond to the absence of residues at certain positions due to limited sample size (261 toxin SPs and 1,738 non-toxin SPs that have been experimentally validated).

4.3.2 Razor accurately predicts toxin SPs

By taking these important features into account, we built SP classifiers to annotate eukaryotic and toxin SPs using random forest. Only SPs with experimental evidence were used for training. We compared these classifiers using an independent test set, where, the MCC, and the cleavage site precision and recall of Razor for eukaryotic SP prediction were 0.405, 0.136, 0.596, respectively (SPs vs non-SPs, see Supplementary Fig C.4, Table S3 and S4). More importantly, Razor outperforms state-of-the-art in

toxin SP prediction, achieving an MCC score of 0.611, and the cleavage site precision and recall of 0.355 and 0.831, respectively (toxin SPs vs non-SPs, see (Fig 4.2), and Supplementary Table C.5 and C.6).

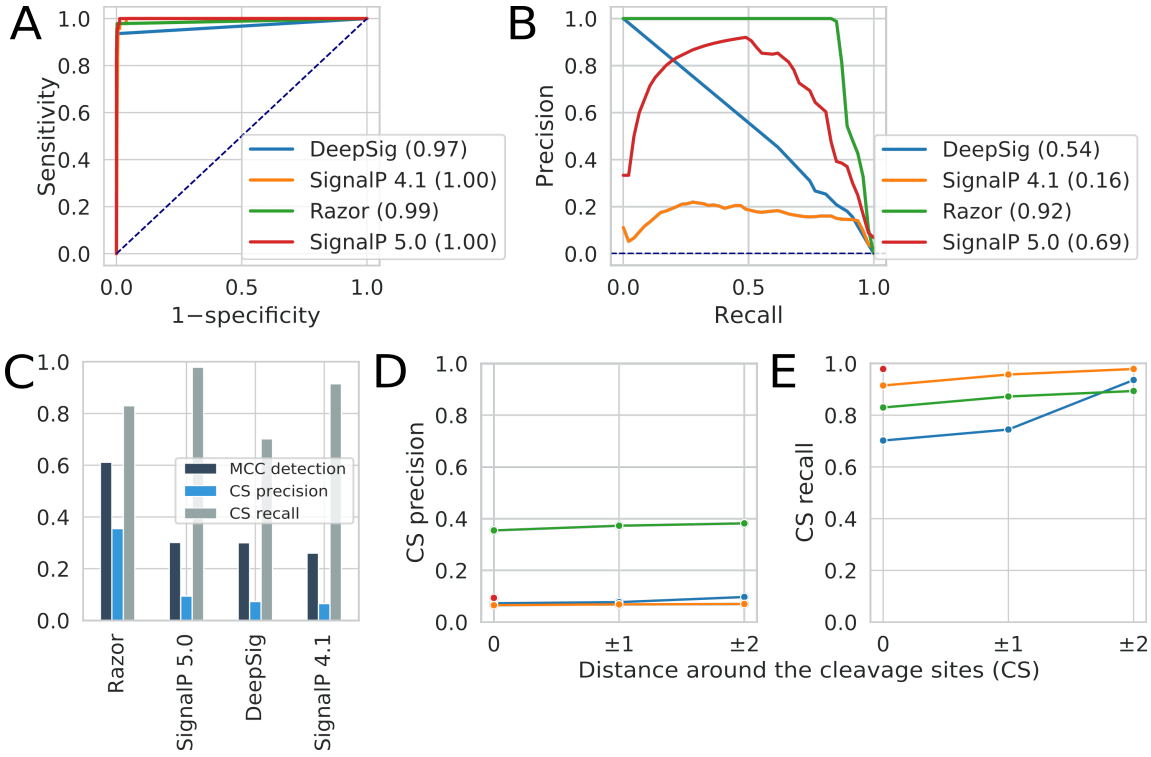


Figure 4.2: **Razor outperforms other tools in predicting toxin SPs.** Benchmarks were carried out using an independent test set (47 experimentally validated toxin SPs and 52,055 non-SPs). **(A)** Receiver operating characteristic curves **(B)** and precision recall curves **(C)** of the SP prediction tools. Areas under the curves are shown in parentheses. The dotted lines show the performance of a random classifier. **(C)** Matthew's Correlation Coefficients (MCC) of the SP prediction tools. The cleavage site (CS) precisions **(D)** and recalls **(E)** of windows surrounding the cleavage sites are shown. Data are available in Supplementary Tables C.5 and C.6.

4.3.3 Defensive proteins harbour a toxin-like SP

The training set for the toxin SP classifier was mainly composed of the SPs from animal toxins, e.g. snake three-finger toxins, scorpion toxins, and phospholipase A2, and plant toxins, e.g. ribosome-inactivating proteins (Fig 4.3A). To further assess our new toxin SP classifier, we scanned the reviewed sequences from UniProt (N=561,776). A total of 910 sequences were predicted positive from all SP detection models.

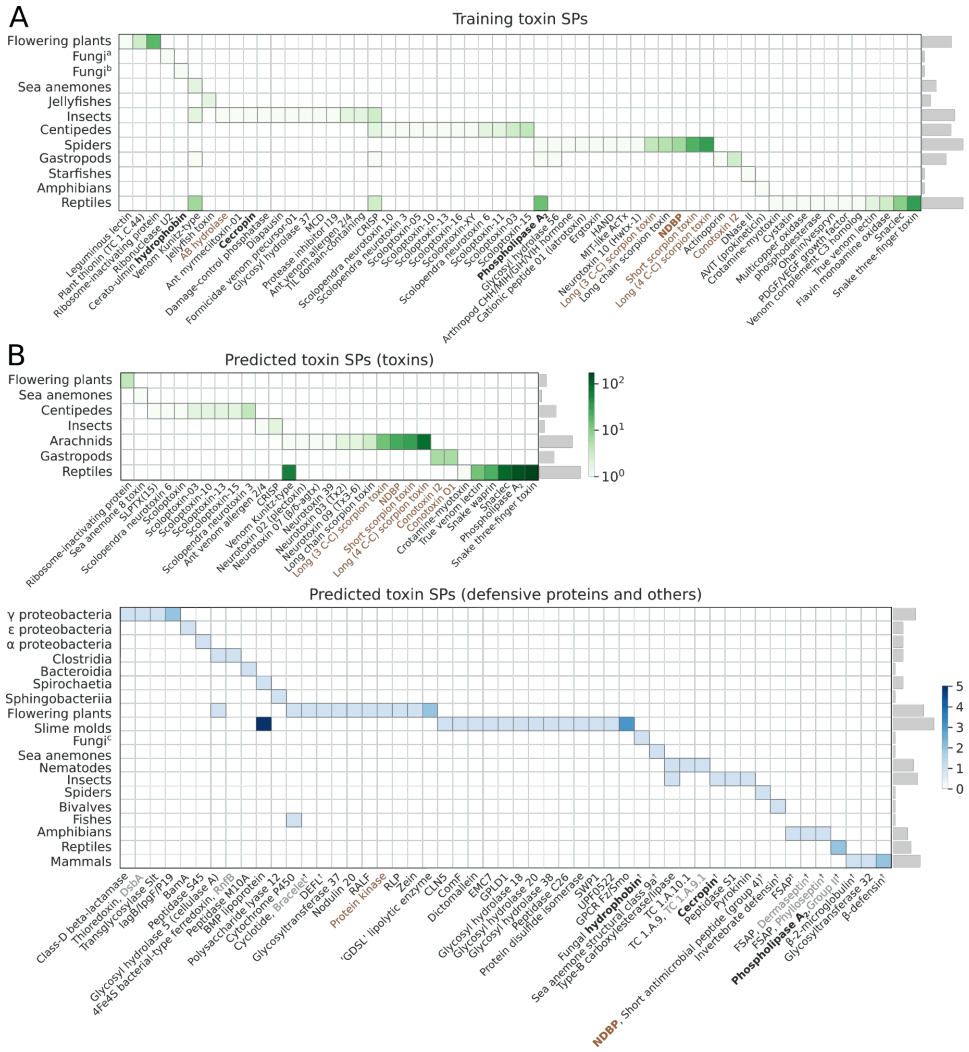


Figure 4.3: **Razor identifies SPs from toxins along with several classes of defensive proteins.** The reviewed sequences from UniProt were examined (N=561,776). **(A)** Heatmap shows the abundance of protein families in the training toxin sequences with SPs by taxa. A total of 237 of 261 training toxins had protein family annotations. **(B)** Heatmaps show the abundance of protein families in the sequences predicted to harbour toxin SPs. A total of 753 of 759 toxins predicted to harbour toxin SPs had protein family annotations (top). A total of 110 other types of proteins were predicted to harbour toxin SP, in which 76 of them had protein family annotations (bottom). The scale bars indicate the frequencies of protein families. Those protein families that have defensive properties are marked with † (bottom). Protein families that are in common between the training and predicted toxin SP sequences are bolded (bottom panel). Protein subfamily, family and superfamily are shown in grey, black and brown, respectively. Fungi[a], Eurotiomycetes; Fungi[b], Sordariomycetes; Fungi[c], Agaricomycetes; CLN5, Ceroid-Lipofuscinosis Neuronal protein; ComF, Competence protein F; CRISP, Cysteine RIch Secretory Protein; DEFL, DEFensin Like; EMC7, ER membrane protein complex subunit 7; FSAP, Frog Skin Antimicrobial Peptide; GPLD1, Glycosyl-phosphatidylinositol-specific phospholipase D; HAND, Helical Arthropod-Neuropeptide-Derived; RALF, Rapid ALkalinization Factor; RLP, Receptor Like Protein; SLPTX, Scoloptoxin; UPF, Uncharacterised Protein Family.

In (Fig 4.3), we excluded potential false positive hits, i.e. computationally annotated transmembrane proteins by UniProt (N=33). The remaining sequences were divided into two groups based on the presence or absence of toxin annotation. From these probable toxin SPs, 759 sequences had annotations for toxins. They included protein families such as scorpion toxin, phospholipase A2 and ribosome-inactivating protein (Fig 4.3B). The remaining 110 sequences had no annotations for toxins. These sequences were clustered at an identity threshold of 70%, which gave rise to 100 representative sequences. Interestingly, many of these proteins without toxin annotations have some defensive properties such as antibacterial peptides and cyclotides. Furthermore, other defensive proteins such as beta-defensin and defensin-like (DEFL) are the results of convergent evolution. For example, beta-defensin-like motifs are also found in toxins from lepidosauria (rattlesnakes and bearded dragons) and mammalia (platypus) [74, 75, 264]. This suggests why their SPs show some remote similarity with toxin SPs.

4.4 Discussion

We have studied the features of SPs from eukaryotic proteins. While SPs share a common hydrophobic nature, we have found several differences between toxin SPs and other eukaryotic SPs in their residue compositions and consequently the sequence properties. We have used these features to develop Razor for annotating eukaryotic SPs, which have specialised functionalities in annotating toxin SPs. Razor outperforms other sophisticated methods in predicting toxin SPs. Using Razor, we were able to predict several classes of probable toxins, which are yet to be annotated (Fig 4.3). Our predicted results consist of toxins and defensive proteins from diverse species, which gives us an overview of the source of toxins.

Since toxins and defensive proteins occur naturally in organisms to attack and neutralise foreign invaders, many of our predicted results include proteins involved in innate immune response and signalling. Some of the frequently observed biological processes of these proteins were 'killing of cells of other organism [GO:0031640]', 'defense response to fungus [GO:0050832]', 'defense response to bacterium [GO:0042742]' and

'innate immune response [GO:0045087]' (Supplementary Fig C.5 and C.6). Many toxins and defensive proteins are commercially important. For example, plant toxins such as defensin-like protein, animal toxins such as cecropin are used to develop disease-resistant transgenic crops [228, 137, 268, 26, 4]. Similarly, the cytotoxic activity of phospholipase A2 on cancer cells makes it a promising candidate for cancer therapy [271, 103, 145].

Taken together, Razor uses an approach independent of homology search to identify known and novel toxin classes across species. Razor was able to identify previously unannotated SPs and a spectrum of toxins and defensive proteins simply using the first 23 N-terminal residues. This also suggests a possible evolutionary constraint on SPs driven by the specialisation of the toxin secretory systems (or convergent evolution), and supports the idea of horizontal gene transfer of several toxin gene classes [244]. Therefore, accurate annotation of toxin SPs can enhance comparative genomics analysis and genome sequencing projects. Razor might also be useful in other research areas such as recombinant protein expression, toxicology, transgenics, and drug design.

4.5 Methods

4.5.1 Datasets

We retrieved the training dataset for the state-of-the-art SP prediction program SignalP 5.0, which is a curated set of the N-terminal sequences from all domains of life [5]. To get the full sequences and annotations of eukaryotic proteins, we used UniProt's ID mapping service [245] and obtained 17,264 fully annotated sequences, of which 2,609 sequences have been experimentally validated to harbour functional SPs. These sequences were used to build a generic, eukaryotic SP classifier. For feature analysis, we clustered these sequences (60 N-terminal residues) at an identity threshold of 70% using CD-HIT v4.8.1 [76]. A single representative sequence was retained for each cluster to reduce sequence redundancy (Supplementary Table C.1).

To build a classifier specialised for annotating toxin SPs, we manually curated a

separate positive set using the dataset from the animal toxin annotation project [118] and a subset from the above training set. Other SPs were assigned as a negative set. We then clustered the sequences as above and analysed the representative sequences (Supplementary Table C.1).

The SP classifiers were compared using an independent test set retrieved from UniProt on 16 February 2021. In particular, the eukaryotic SP classifier was evaluated using 241 SPs with experimental evidence and 52,055 non-SPs, whereas the toxin SP classifier was evaluated using a subset of this independent set (toxin SPs=47, non-SPs=52,055). We also scanned the reviewed sequences from UniProt (N=561,776, retrieved on 2 September 2020).

4.5.2 Bit score

The bit scores of the N-terminal residues were computed as:

$$bit\ score(residue) = log_2 \frac{Normalised\ count\ of\ residue\ in\ positive\ set}{Normalised\ count\ of\ residue\ in\ negative\ set} \quad (4.1)$$

For eukaryotic proteins, the positive set and the background set were SPs and non-SPs, respectively. For toxins, the positive set and the background set were toxin SPs and non-toxin SPs, respectively.

4.5.3 Protein sequence properties

The standard protein sequence properties, implemented in BioPython, were calculated using the Bio.SeqUtils.ProtParam module v1.73 [49]. These features include GRand AVerage of hydropathicitY (GRAVY), Flexibility, Helix, Sheet and Turn propensities, Instability Index, Aromaticity, and Isoelectric Point. An additional feature included is the Solubility-Weighted Index (SWI; [19].

4.5.4 SP classifiers

We built a random forest classifier based on several sequence features (GRAVY, flexibility, helix, and SWI), as well as the counts of residues (R, K, N, D, C, E, V, I, Y, F, W, L, Q, and P) of the first 30 N-terminal residues. The residues were chosen

such that they maximised Matthew's correlation coefficient (MCC) in five-fold cross-validations. After the cross-validation step, we generated five random forest models, which are used for scoring the N-terminal of a given sequence. The scores from these classifiers are comparable to the S-score of SignalP 4.0 except that our scores are non-position-specific [184].

For the prediction of the cleavage site, we took a total of 30 residues such that the cleavage site is aligned in between positions 15 and 16 in order to capture the major differences in residue distribution around the cleavage site. We built a 20×30 matrix and populated it with the hydrophobicity scale [136] as initial weights. We then used multi-objective simulated annealing [131] at each position such that the new weights maximised the AUC and precision-recall curve based on the training set. The scoring of the cleavage site (C-score) is done using the random forest classifier trained on the aligned set encoded using the optimised weight matrix. Small limitation of our approach is that we are unable to detect the correct cleavage site if it is located before the 15th position. Yet, based on training data, this is rarely observed (N=13).

After detecting the cleavage site, the final score for classification (Y-score) is the geometric mean $Y = \sqrt{S \times C}$, where S is the S-score and C is the max of C-scores along the sequence. For the final classifier, we chose a threshold of Y-score that maximised the MCC after five-fold cross-validations (MCC=0.914) on the training set.

We then built models specialised in annotating the toxin SPs based on hydrophobicity, SWI, flexibility, and turn. These features were selected such that they maximised the MCC using five-fold cross-validations on the training set. The N-terminal length of 23 was found to generate the maximum median MCC score for the toxin SP classifier (MCC=0.741, see also Supplementary Table C.2). Similar to the SP prediction models, the toxin SP classifiers consist of five models each.

4.5.5 Performance measures

We use MCC as a measure of performance to correctly identify eukaryotic SPs. We also use cleavage site precision ($CS_P = N_{corr}/N_P$) and recall ($CS_R = N_{corr}/N$),

where N_{corr} is the number of the correctly identified cleavage site, N_P is the number of predicted SPs and N is the number of SPs [5, 206].

4.5.6 Tool

We developed Razor for annotating SPs using the eukaryotic and toxin SP classifiers (Fig 4.4).

Razor accepts either a nucleotide sequence or a protein sequence. Sequences with a length of lower than 30 residues are padded with Serine (Ser, S), because it shows equal enrichment across all datasets, in particular after the H-region (Fig 4.1). Razor is available both as a command-line tool (https://github.com/Gardner-BinfLab/Razor) and a web application (https://tisigner.com/razor). For the web application, predictions from five models are displayed as stars. The final score is the median of scores from five models and is displayed along with the region for SP. A plot of C-scores along the sequence is also displayed along with the annotation for the cleavage site. In addition, we integrated the Razor web application with our protein expression and solubility optimisation tools, TIsigner and SoDoPE, respectively [19, 21]. Our web tools assist users in annotating SPs and protein domains, and making the decisions from gene cloning to protein expression and purification.

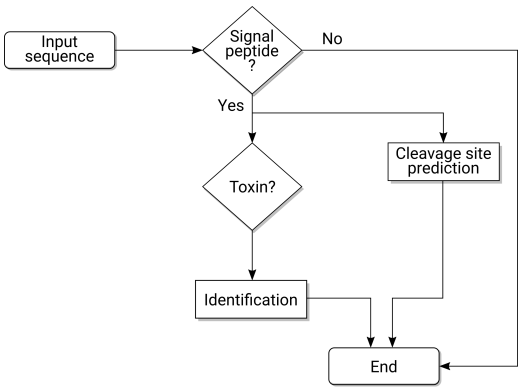


Figure 4.4: Flow chart of toxin SP classification using Razor.

4.5.7 Statistical analysis

Data analysis was performed using pandas v1.0.3 [159]. Hydrophobicity and SWI were smoothed for the classifier training using the Savitzky-Golay filter implemented in SciPy v1.4.1 [252]. Random forest classifier and MCC computation were done

using scikit-learn v0.23.1 [181]. Plots were generated using Matplotlib v3.1.3 and Seaborn v0.10.0 [110, 262].

4.5.8 Code and data availability

Jupyter notebooks for reproducing our analyses are available at https://github.com/Gardner-BinfLab/Razor_paper_2021. The source code for Razor, our SP annotation server can be found at https://github.com/Gardner-BinfLab/TISIGNER-ReactJS.

Chapter 5

TISIGNER.com: interactive web services for improving recombinant protein production

This chapter is adapted from a paper published in *Nucleic Acids Research*, the 2021 Web Server Issue [20]. Associate Professor Paul Gardner, Dr Chun Shen Lim and I conceived the study. Dr Chun Shen Lim and I designed the features of the web server. I developed https://tisigner.com and drafted the paper [20]. Drs Gardner and Lim supervised the study.

5.1 Abstract

Experiments that are planned using accurate prediction algorithms will mitigate failures in recombinant protein production. We have developed TISIGNER (https://tisigner.com) with the aim of addressing technical challenges to recombinant protein production. We offer three web services, TIsigner (Translation Initiation coding region designer), SoDoPE (Soluble Domain for Protein Expression) and Razor, which are specialised in synonymous optimisation of recombinant protein expression, solubility and signal peptide analysis, respectively. Importantly, TIsigner, SoDoPE and Razor are linked, which allows users to switch between the tools when optimising genes of interest.

5.2 Introduction

Recombinant protein production is a key process for life science research and the development of biotherapeutics. However, low protein expression and aggregation are the two major bottlenecks of recombinant protein production [13, 70, 107, 132, 157, 198, 249]. Since mRNA abundance alone is insufficient to explain protein abundance [17, 223, 143, 172, 231], several features of mRNA sequence have been proposed to affect protein expression. These features are mostly related to codon usage, such as the codon adaptation index and tRNA adaptation index [34, 196, 87, 202, 214], or measures of mRNA secondary structure, such as G+C content, minimum free energy (MFE) of RNA secondary structure, and mRNA:ncRNA interaction avoidance [219, 66, 133, 185, 240, 243]. Many of these features are not independent, making it challenging to distinguish the impacts of individual features [156]. This, in turn, hinders the development of accurate prediction/optimisation tools. Recent systematic studies suggest that MFE is the most important feature in protein expression [35, 156]. However, more recent work shows that the mRNA accessibility of translation initiation sites outperforms MFE in predicting relative protein levels from mRNA sequences [21, 235]. Accessibility is computed by considering all possible structures for a region, weighted by free energy, not just the single structure with the MFE [14].

In addition to high protein expression level, high solubility is preferable for the purification and long-term storage of recombinant proteins. However, almost half of the successfully expressed proteins are insoluble (http://targetdb.rcsb.org /metrics), which makes the recombinant protein production process challenging. A number of methods have been suggested to improve protein solubility, for example, truncation, mutagenesis, and the use of solubility-enhancing tags [41, 52, 70, 254]. Nevertheless, accurate solubility prediction could save resources and aid in designing soluble proteins before the experiments. With these in mind, we have recently formulated the Solubility-Weighted Index (SWI), which outperforms recent solubility prediction tools based on machine-learning algorithms [19].

Besides, many recombinant proteins of interest are secretory. The intracellular ac-

cumulation of heterologous secretory proteins may be toxic to the host cells. Therefore, the translocation efficiency of these proteins plays an important role in the yield quantity and quality. Secretory proteins usually have a short peptide at the N-terminus called Signal Peptide (SP) which is responsible for the translocation of secretory proteins via the Sec, Signal Recognition Particle (SRP) or Twin arginine transport (Tat) pathways [150, 180, 201, 99]. Detection of SPs or fusion of a suitable SP at the N-terminus is useful for optimising protein production [73, 126, 199, 277]. In addition, different pathways have different advantages, for example, the SRP dependent pathway can be used for rapidly folding proteins [179]. However, the Sec dependent pathway, which is common across all forms of life, has been widely used for recombinant protein expression because of higher protein production capacity and quality [152, 179]. In addition, the presence of SPs should almost always be checked when planning the expression experiments for uncharacterised proteins.

Existing web tools predict or optimise either protein expression or solubility alone [3, 45, 83, 189, 94, 106, 218, 222, 278]. There also exists several web tools for predicting SPs [5, 10, 101, 119, 206]. Only a very few tools can detect toxic proteins, for example, SpiderP, ClanTox, and ToxinPred [84, 164, 266]. These tools are either limited to predicting the venoms of certain organisms, such as spiders, or they are not designed to predict the signal peptides of toxins, rather to predict the toxicity of mature peptides. Moreover, these tools are offered through different independent services. We reasoned these functionalities should be integrated in order to assist not only in choosing appropriate expression systems, but also in optimising the expression and solubility levels of recombinant proteins. Here we present TISIGNER.com that integrates the optimisation tools TIsigner (Translation Initiation coding region designer), SoDoPE (Soluble Domain for Protein Expression) for protein expression and solubility, respectively, and Razor for detecting SPs [21, 19, 18]. Our web application provides easy, fast and interactive ways to assist users in planning and designing their experiments (Figure 5.1).

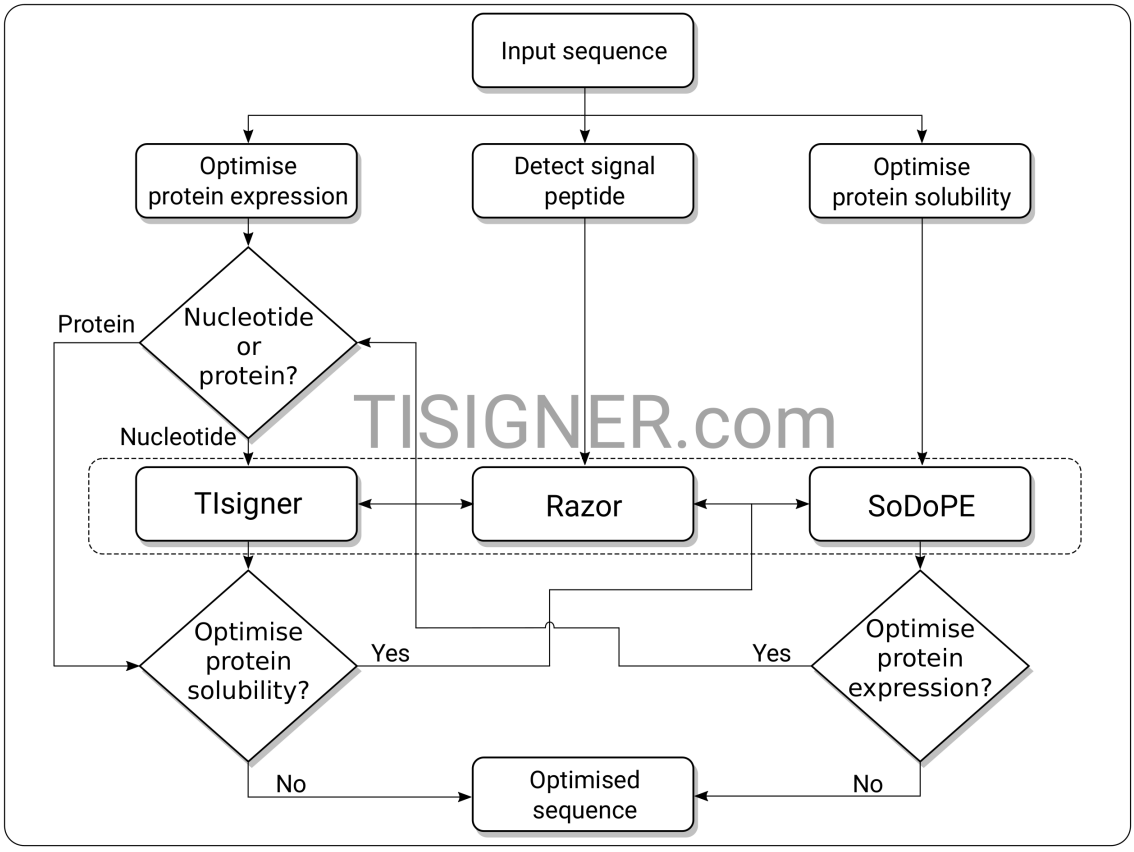


Figure 5.1: Flow chart for optimising recombinant protein production using the TISIGNER web application. TIsigner, SoDoPE and Razor are linked so that protein expression and solubility can be seamlessly optimised. TIsigner accepts a nucleotide sequence as input, whereas SoDoPE and Razor accept either a nucleotide or protein sequence. SoDoPE, Soluble Domain for Protein Expression; TIsigner, Translation Initiation coding region designer.

5.3 Web services

5.3.1 TIsigner

TIsigner offers tunable protein expression by optimising the mRNA accessibility of translation initiation sites [21]. The regions used to calculate accessibility (opening energy) are specific to the expression hosts, which is calculated using RNAplfold [14, 15, 146]. For *Escherichia coli*, *Saccharomyces cerevisiae*, and *Mus musculus* expression hosts, the optimal regions relative to the start codon for optimisation are $-24:24$, $-7:89$, $-8:11$, respectively. For other expression hosts, we provide an option 'Other', which optimises the accessibility of the region $-24:89$. Since *E. coli* is the most popular expression host, the default settings aim to optimise protein expression in *E. coli* with the T7 lac promoter system (see below). In this case, only the protein coding sequence is required for input (Figure 5.1). Otherwise, the 5′UTR (5′ untranslated region) sequence is also required.

The settings for TIsigner are grouped by complexity (i.e., general, extra, and advanced). The general settings include the options to modify the expression host, promoter and target expression score. The target expression score ranges from 0 to 100 (i.e., from the minimum to maximum predicted level), which is derived from a logistic regression of the opening energy distribution of $11,430$ expression experiments in *E. coli* from the 'Protein Structure Initiative:Biology' (PSI:Biology) [43, 213]. Hence, this scoring system is only applicable to the *E. coli* T7 lac promoter system. Since, there is a non-linear relationship between opening energy and expression score, an interactive plot is also displayed along with the slider to set the target expression score. For other expression hosts and promoters, the target expression level can be either maximised or minimised (i.e., binary). The extra settings have the options to optimise sequence within the translation initiation region or the full-length sequence. The AarI, BsaI, BsmBI restriction modification sites are filtered by default, whereas other sites can be manually supplied (e.g., a Shine-Dalgarno motif or terminator U-tract). The advanced settings allows users to tweak the random seed and sampling options (i.e., quick or deep, which uses different numbers of iterations

Figure 5.2: The results of TIsigner shows a protein expression optimised nucleotide sequence. The highlighted nucleotides show changes made to the input sequence. The opening energy of the input sequence before and after optimisation is annotated over the distributions of the opening energy for 8,780 'success' and 2,650 'failure' experiments from PSI:Biology. Further optimised sequences, if found, are also displayed. The results can be downloaded in either CSV or PDF format using the download icon on the bottom right. Each resulting sequence can be analysed for solubility or signal peptide.

and parallel processes). Here users can also customise the region for optimisation or disable the terminator checks.

Once the input sequence passes a sanity check, the optimisation task is rapid ($O(1)$ time using RNAplfold v2.4.11 (using parameters -W 210 -u 210)) with our simulated annealing algorithm. A list of optimised sequences are returned after checking for terminators using cmsearch (Infernal v1.1.2) [167] with RMfam models [80, 121]. If terminators are found, an option to use the full-length sequence for optimisation will be prompted to users. In a default case (*E. coli* T7 lac promoter system), the optimised sequence closest to the chosen expression level is selected as the first solution (Figure 5.2). For other expression hosts and/or promoters, the optimised sequence with the minimum changes in nucleotides is selected as the first solution. The altered nucleotides are highlighted (Figure 5.2). The accessibility of translation initiation sites for both the input and optimised sequences is shown as opening energy (kcal/mol). The results can be exported as a PDF or CSV file. When the default settings are used, the opening energy for each sequence is indicated on the distributions of the opening energy of $8,780$ 'success' and $2,650$ 'failure' groups of the PSI:Biology target genes. Furthermore, options for solubility and SP analyses using SoDoPE and Razor, respectively, are available for each sequence on the same results page (Figure 5.2).

5.3.2 SoDoPE

SoDoPE is our interactive solubility analysis and optimisation tool based on the Solubility-Weighted Index (SWI) [19]. SoDoPE accepts either a nucleotide or protein sequence (Figure 5.1). Upon submission, a query is sent to the HMMER web service for domain annotation [186]. Successful annotations are displayed as interactive graphics, in which the annotated domains are represented as discorectangles, above a grey band that represents the input protein sequence (Figure 5.3). Information about a protein domain is shown upon a mouse hover. The domains can be selected for solubility analysis. For a complete domain annotation report, a link to the HMMER results page is also provided.

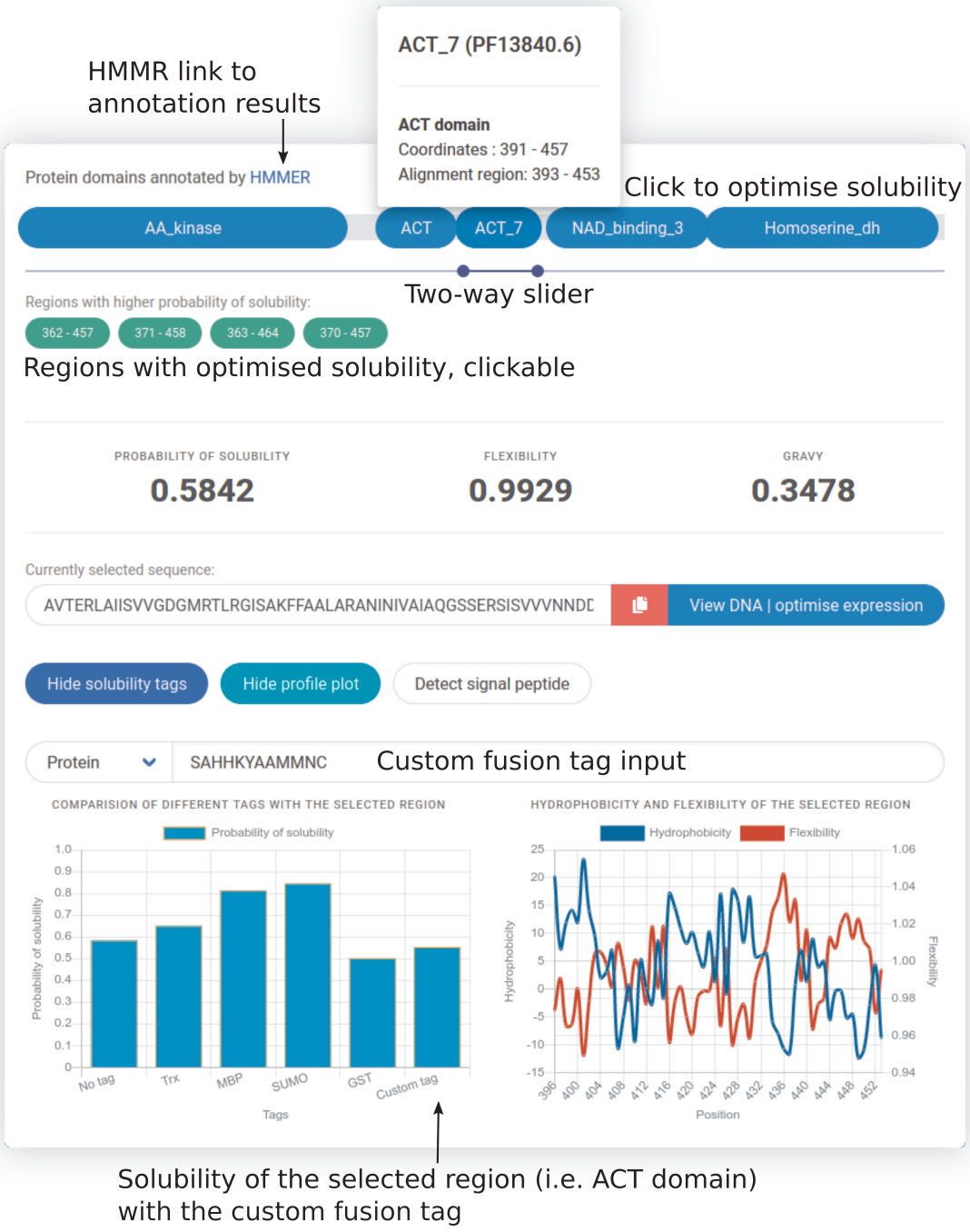


Figure 5.3: Exploring and optimising protein solubility using SoDoPE interactive graphics. Upon clicking a protein domain or selecting a region of interest, its solubility is optimised in real-time, and a list of regions with extended boundaries and higher probabilities of solubility is returned as green buttons (clickable). The probabilities of solubility of the selected region with and without fusion tags can be visualised in a barplot. The flexibility and hydrophobicity profile plots for the selected region can also be selectively viewed. The sequence can also be checked for the presence of a signal peptide or optimised for protein expression.

In addition, a two-way slider is available for navigation through any region of interest (Figure 5.3). The probability of solubility, flexibility and GRAVY (GRand AVerage of hydropathicitY) is shown in real-time according to the user-selected region. The selected region is optimised for higher solubility using simulated annealing. Only the regions with extended boundaries and also higher probability of solubility is returned. SP analysis can also be done using Razor (see below).

A profile plot of flexibility and/or hydrophilicity corresponding to the user selected region is generated (Figure 5.3). This allows an estimation of rigid/flexible regions and possible helices, that may be helpful for mutagenesis experiments. The sequence of the selected region is shown, with the option of sequence conversion between nucleotide and amino acid sequence format. In particular, the nucleotide sequence can be redirected to TIsigner for optimising protein expression (Figure 5.1 and 5.3, through the 'view DNA | optimise expression' button).

The contributions of several solubility-enhancing tags to user selected regions can be compared and shown in a bar plot, including thioredoxin (TRX), maltose binding protein (MBP), small ubiquitin-related modifier (SUMO) and glutathione S-transferase (GST) tags (Figure 5.3). Users can also input a fusion sequence of interest either in a nucleotide or protein sequence format.

5.3.3 Razor

Razor is our SP prediction tool which is based upon random forest models of protein features from the eukaryotic SP sequences of the SignalP 5.0 dataset and the animal toxin annotation project [5, 18, 118]. Razor accepts either a protein or a nucleotide sequence (Figure 5.1). After validation, the N-terminal region is checked for the presence of a SP using five random forest models. This gives five SP scores (S-scores) for a given sequence. For detecting the cleavage site, we use a sliding window of 30 residues and our optimised weight matrix for residues around the cleavage site. The scored subsequences are scored by additional five random forest models to give the cleavage site scores (C-scores) along the sequence, which is displayed as a step plot (Figure 5.4). The Y-score, which is the geometric mean of S-scores and the max

of C-scores, is used to infer whether the given sequence has a SP or not. The median of these five Y-scores is displayed as the final score. The cleavage site from the model with the median of max of C-scores is used to annotate the predicted region.

Figure 5.4: Detection of signal peptides using Razor. The dotted annotation in the step plot for the cleavage site scores (C-scores) shows the most likely position for proteolytic cleavage. The sequence can also be checked and optimised for protein solubility and expression.

If any of the models detect a SP in the input sequence, we further check whether the SP belongs to toxins, using five random forests trained on toxin-specific SPs. The final toxin score is the median of scores from those random forest models. Furthermore, since we noticed a lack of tools specialising in predicting SPs from fungi, any detected signal peptide is checked for such origin. Similarly, we use five random forests for detecting fungal SPs, with the final fungal score being the median

score of these models. This random forest is built using residue composition of the signal sequence. Since we have five random forest models in each step (SP, toxin- and fungal-specific SP detection steps), stars are displayed as an indication of the number of models agreeing on the sequence falling on either category (Figure 5.4).

Razor is linked with SoDoPE for checking and optimising protein solubility (Figure 5.4). If a nucleotide sequence was submitted, this sequence can also be optimised for protein expression using TIsigner (Figure 5.1).

5.4 Discussion

Low protein expression and solubility are the major hindrances to a successful recombinant protein production. Based on our comprehensive studies on these two problems, we have developed novel tools to optimise protein expression (TIsigner) and solubility (SoDoPE), and assessed their predictive performance using independent datasets (Table D.1). Our tools offer some unique features in an interactive way. TIsigner allows tuning of protein expression from low to high levels, whereas SoDoPE allows easy navigation of protein sequence/domains with real-time solubility prediction. Based on our assessment of similar tools, none of the publicly available tools provides these features.

Our third tool, Razor, is designed to check the presence of SPs. Compared to other related tools, Razor also predicts toxin- and fungal-specific SPs (Table D.2). These would be helpful for users in choosing the expression and purification systems that prevent the harmful intracellular accumulation of recombinant secretory proteins/toxins.

Our tools are interactive, fast, and accurate. Importantly, our tools are highly integrated, allowing a seamless transition between the optimisation tools. To make such transition intuitive, our web services limits one input sequence at a time and we aim to remove this input sequence limitation in the future. For optimising a large number of sequence, we provide the command-line version of each of our tools (see below).

5.5 General information

Demo input and results are available for new users to get started. A list of frequently asked questions is also available for each tool. The frontend is written in React and uses responsive web design principles. The backend is written in Flask and Python v3.6. The website is hosted on a virtual machine (Red Hat Enterprise Linux 8) running on Intel Xeon (8 × 2.60 GHz) with 4GiB RAM, by the Information Technology Services at the University of Otago.

5.6 Data availability

The web server is available at https://tisigner.com. This website is free and open to all users and there is no login required. All our tools, and the website are open-sourced (https://github.com/Gardner-BinfLab/TISIGNER-ReactJS; https://github.com/Gardner-BinfLab/TIsigner/tree/master/TIsigner_cmd; https://github.com/Gardner-BinfLab/SoDoPE_paper_2020/tree/master/SWI; https://github.com/Gardner-BinfLab/Razor) and privacy friendly (no user data stored).

Chapter 6

Discussion

Recombinant protein production is widely used in scientific research and industry. However, in general, the success rate of these experiments is around a 25%, the 75% failure rate is attributed to protein expression and solubility. In this work, we studied the major factors affecting these two steps and used the findings to develop methods to optimise the protein production. In addition, since many recombinant proteins of interest are secretory, translocation efficiency also plays an important role in the final yield. Since protein translocation is usually performed by signal peptides, fusing appropriate signal peptide to the protein of interest increases the yield. Therefore, we also developed tool to predict the presence of signal peptide in the sequence and identify the mature peptide.

6.1 Optimising protein expression using TIsigner (Translation Initiation coding region designer)

We have demonstrated that mRNA accessibility is a better predictor of protein expression across several datasets (Table D.1). Therefore, we used this feature to develop, TIsigner (https://tisigner.com/tisigner), a tool for optimising protein expression. TIsigner uses a simulated annealing algorithm to provide a novel mechanism for tuning the protein expression from low to high levels. Other unique features include an estimation of protein expression for mRNA sequences, which we've named as 'expression score' and synonymous changes limited to the first 10 codons of the input sequence. TIsigner's synonymous substitution algorithm is designed to make

a minimal number of changes. This approach is advantageous for two reasons—first, it is possible to do a PCR cloning using the optimised part as primers, which in turn reduces the cost of the experiment significantly as compared to the conventional full gene synthesis method. Second, it reduces the possibility of generating toxic mRNA sequences. mRNA toxicity is still a difficult problem to address and the mechanisms of toxicity are not fully understood yet [161]. However, TIsigner offers a possibility to do synonymous changes over all codons, which may be useful if terminators are found. The web-version of TIsigner automatically suggests doing a full length substitution if any terminators are found (Figure 6.1).

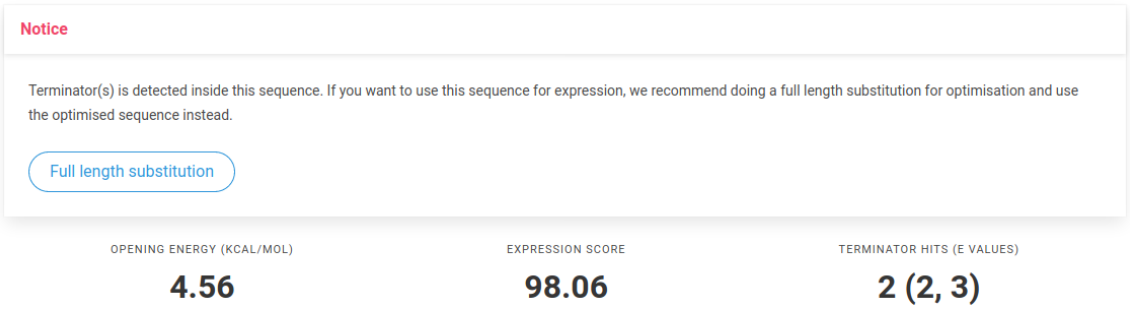


Figure 6.1: **TIsigner suggests to do a full length substitution if any transcription terminators are found in the sequence.** Mock data was used for demonstration purposes.

6.2 Optimising protein solubility using SoDoPE (Soluble Domains for Protein Expression)

Almost all use cases of protein require a soluble product. Hence, optimising just protein expression is not sufficient for a better protein production. Based on protein structural flexibility, which is the most accurate predictor of solubility among all conventional features, we have developed a new metric called the Solubility-Weighted Index (SWI). SWI is a very accurate predictor of solubility and outperforms other tools (Table: D.1). Using SWI, we developed a solubility prediction and optimisation tool, SoDoPE (https://tisigner.com/sodope). SoDoPE is an unparalleled tool with a distinctive interface for an easy navigation of protein sequence/domains with real-

time solubility prediction and optimisation. The effect of different solubility tags on solubility can also be compared to pick the best one for experiment.

6.3 Detection of signal peptides using Razor

The presence of signal peptides should almost always be checked when planning expression experiments for uncharacterised proteins. Based on the properties of signal sequences, we developed Razor (https://tisigner.com/razor) to detect the presence of N-terminal signal peptides. Compared to other related tools, Razor also predicts toxin and fungi—specific SPs. These would be helpful for users in choosing the expression and purification systems that prevent the harmful intracellular accumulation of recombinant secretory proteins/toxins. The performance summary of Razor is given in Table D.2.

6.4 Reception of tools by the community

TISIGNER (https://tisigner.otago.ac.nz, https://tisigner.com, https://tisigner.nz) is the web-suite for these tools. Initially, it consisted of TIsigner only, hence the name, but has been expanded to include both SoDoPE and Razor. Our tool has been online since February 2020. Despite the short period (March 2021 at the time of writing), our tools are gaining traction among researchers worldwide, with an exponential increase in the numbers of visitors from many countries (Figure 6.2).

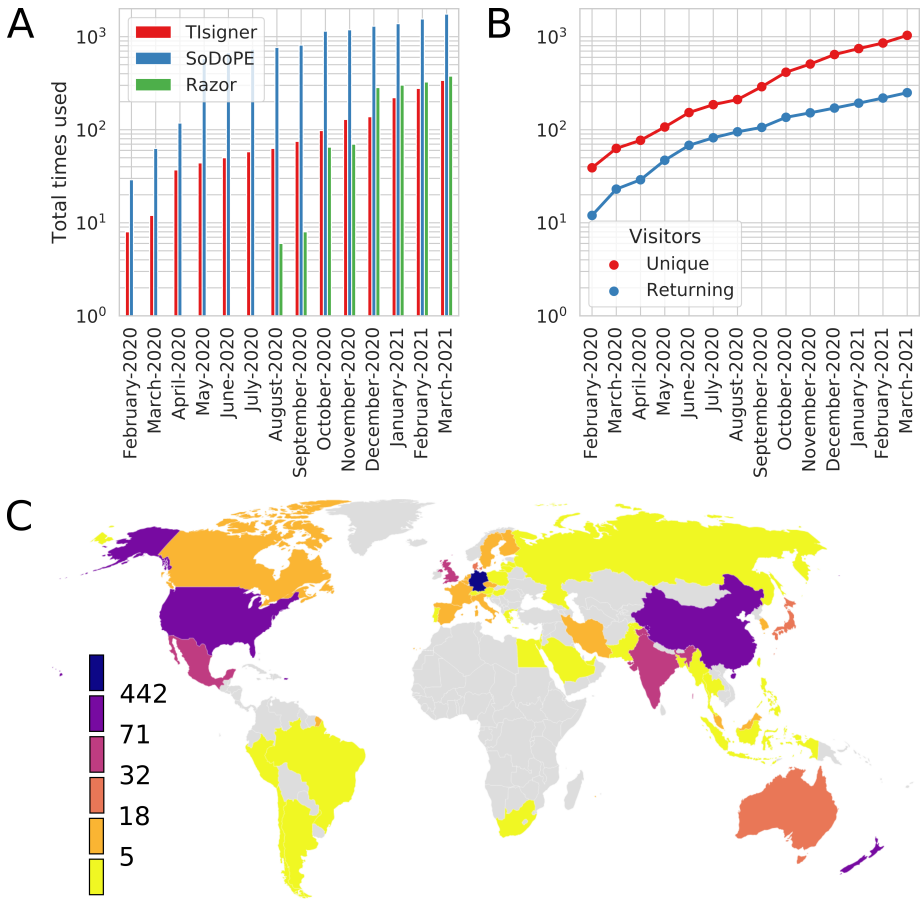


Figure 6.2: **TISIGNER.com web service is getting popular among researchers worldwide. (A)** Number of times each tool was used in a logarithmic scale. **(B)** The number of unique and returning visitors in a logarithmic scale. **(C)** Geographical location of users. Data is for the period of February 2020 to March 2021 from Google Analytics (https://analytics.google.com).

6.5 Outlook

The present work enumerates some of the cellular level processes responsible in protein synthesis. In this section, we will use a simplified version of higher level modelling by taking into account of the production system as a black-box (S) which has various levels of noise and uncertainties. In such a crowded environment, there might be some, possibly unknown, additional constrain and feedback between or within the cells, which could stabilise or destabilise the production system. Let us represent any and all of these uncertainties, noises and feedbacks collectively by a black-box (F) and simply refer to as feedbacks. A stable system is a result of interplay between the system and its feedback. These type of systems are called *control systems* (Figure

6.3A). The output behaviour of output of these systems when the input is switched from low to high is studied within the framework of step response.

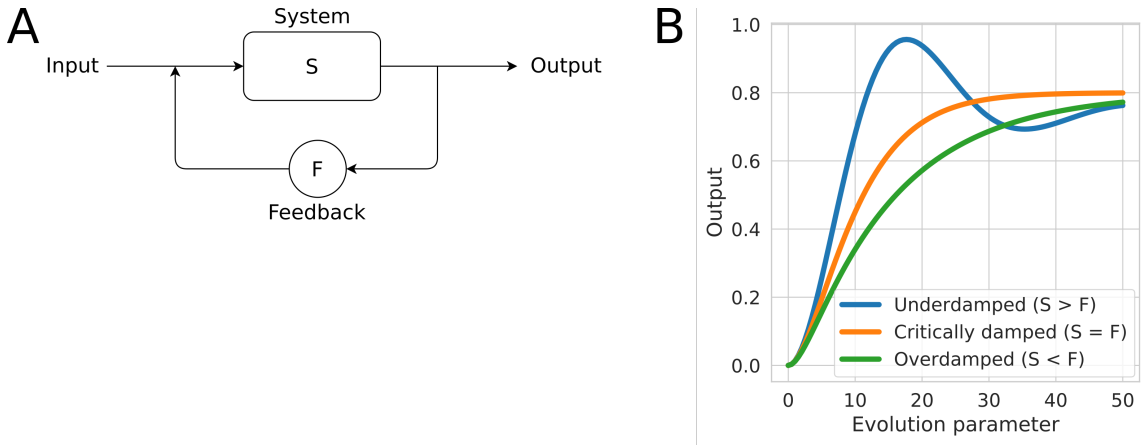


Figure 6.3: **Different possible step responses of a stable control system when input is switched from low to high. (A)** A control system (S) with a feedback loop (F). This system is assumed to be stable. **(B)** Output (step response) when the damping effects due to feedback (F) are low, equal and high compared to the system (S). Evolution parameter is an arbitrary parameter of the system.

In control theory, the evolution of higher order system with respect to some evolution parameter (often time, but could be any other variable) is approximated by a second order ordinary differential equation:

$$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = 0 \tag{6.1}$$

where ω_n is the eigen frequency, ζ is the damping ratio. The eigen frequency is the frequency at which the system oscillates when no external forces are present, whereas damping ratio is an indicative of the relative strength of feedback to the system (F/S). The most general solution of this equation can be written as a linear combination:

$$x = Ce^{-\omega_n(\zeta+i\sqrt{1-\zeta^2})} + De^{-\omega_n(\zeta-i\sqrt{1-\zeta^2})} \tag{6.2}$$

Hence for any system, multiple output possibilities exist which are shown in Figure 6.3B. For $0 \leq \zeta < 1$, the output is decaying exponential with oscillations and is called underdamped. For $\zeta > 1$, the system does not oscillate and is called overdamped whereas for $\zeta = 1$, the system is logistic in nature and is called critically damped.

Ideally, physical systems with $\zeta = 1$ (critically damped) are preferred because the maximum output can be reached easily. In contrast, biological systems are often noisy, which could make the output behave like either of the cases. As an illustration, the data from Cambray *et al.* [35], follows an overdamped trend ($\zeta > 1$) (Figure 6.4). However, the results of our GFP experiments using TIsigner, fits better to the underdamped system ($0 \leq \zeta < 1$) than a logistic regression (critically damped) (Figure 6.5). In this case, the output tends to oscillate when the input is increased beyond a certain limit, which we observed (Figure 6.5C).

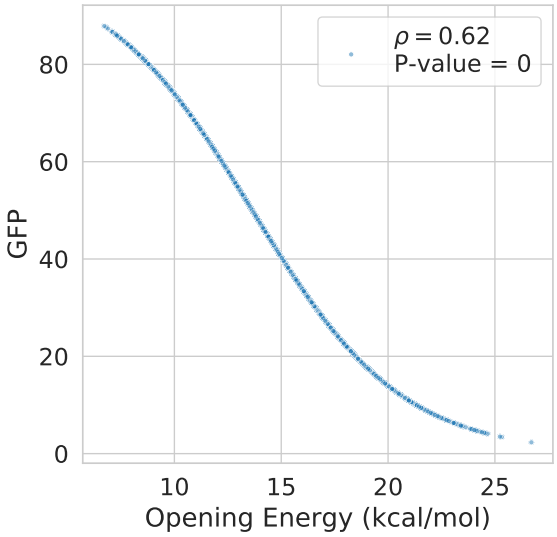


Figure 6.4: **Best fit shows an overdamped ($\zeta > 1$) trend in GFP data from Cambray *et al.* (2018)** Spearman's $\rho = 0.62$, P-value is less than machine's underflow. In these type of systems, the output increases steadily with an increase in input level. Reverse trend is because lower opening energy are more optimal than higher opening energy sequences.

These inherent and inevitable feedbacks and constrains, for example protein toxicity and solubility issues, may be different for different systems, protein of interest and protocols. In this work, these issues were treated separately. However, modelling the production system by taking into account of all these variables may be required to explain the outcomes of recombinant protein production systems with a greater accuracy. The simulated production system (Fig 2.4), which takes into account of only the protein toxicity and mRNA accessibility, was surprisingly close to the experimental results. Combined with a higher order modelling as outlined above, such simulations can be used to explore and study the effects of different features of mRNA and protein as well as stochasticity. Nevertheless, this work provides a sufficient background for such a complex modelling in the future.

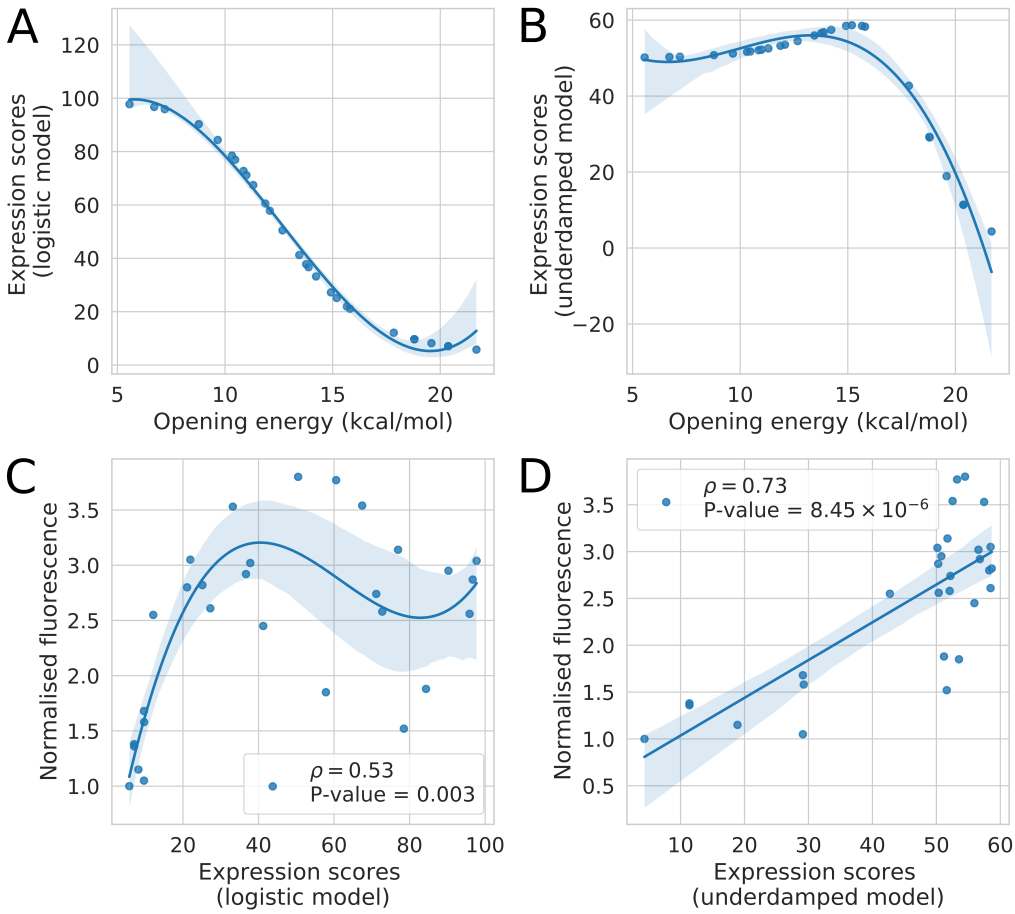


Figure 6.5: **An underdamped model fits better to the TIsigner experimental data (GFP) than a logistic (critically damped) model.** The conversion of opening energy to expression score using **(A)** logistic (critically damped) model and **(B)** underdamped model. Spearman'r ρ between normalised fluorescence and expression scores derived from **(C)** logistic model and **(D)** underdamped model. Spearman's ρ is stronger when we model the production as underdamped (0.73, P-value=8.45×10^{-6}) compared to the critically damped model (0.53, P-value=0.003). Lines represent the best fit regression.

Appendix A

Protein yield is tunable by synonymous codon changes of translation initiation sites

A.1 Supplementary notes

A.1.1 Cloning of TIsigner variants of GFP and Luciferase

The cloning of TIsigner sequence variants for E. coli expression was performed using the MIDAS Golden Gate cloning system [62]. As with other Golden Gate assembly (GGA) systems, MIDAS is a modular, hierarchical DNA assembly system that uses the Type IIS restriction enzymes AarI, BsaI and BsmBI to assemble genes, transcription units and other devices from basic parts, and subsequently enables multiple devices to be assembled together on a single plasmid. As per MIDAS, basic parts such as promoters, coding sequences and terminators were amplified by PCR or ordered as synthetic polynucleotide sequences from gene synthesis companies. The basic parts are listed in Table A.1. Protocols for the GGA reactions are as described in van Dolleweerd et al., 2018 [62].

N-terminal region of GFP (GFPN)

Overlapping oligonucleotide primer pairs corresponding to the first ten codons of each of the gfp sequence variants produced by the TIsigner algorithm were ordered

from Integrated DNA Technologies (IDT) (see Table A.2). Each overlapping pair of primers was annealed together and used as a template for amplification by Q5 polymerase (New England Biolabs) to create a double-stranded DNA product spanning the N-terminal region of GFP (GFPN; codons 1 to 10; see Fig A.1), and with MIDAS [CCAT] prefix and [GTTG] suffix nucleotides.

C-terminal region of GFP (GFPC)

The C-terminal region of GFP (designated GFPC), spanning codons 11 to 238 of the native gfp of *Aequoria victoria*, was synthesized (GeneArt) with flanking BsmBI recognition sites, and with the [GTTG] MIDAS prefix (compatible with the [GTTG] suffix on the GFPN part) and [GCTT] suffix nucleotides.

N-terminal region of luciferase (RLucN)

Overlapping oligonucleotide primer pairs corresponding to the first ten codons of each of the luciferase sequence variants generated by the TIsigner algorithm were ordered from Integrated DNA Technologies (IDT) (see Table A.3). Each overlapping pair of primers was annealed together and used as a template for amplification by Q5 polymerase (New England Biolabs) to create a double-stranded DNA product spanning the N-terminal region of luciferase (RLucN; codons 1 to 10), and with a MIDAS [CCAT] prefix and an [AGGA] suffix.

C-terminal region of luciferase (RLucC)

The C-terminal region of luciferase (designated RLucC), spanning codons 11 to 311 of the native luciferase of *Renilla reniformis*, was amplified from the full-length, native luciferase sequence (synthesized by GeneArt) using primers that add flanking BsmBI recognition sites, and a MIDAS [AGGA] prefix (compatible with the [AGGA] suffix on the RLucN part) and a [GCTT] suffix.

MIDAS Level-1 cloning of parts

PCR products, purified using commercially available column-based protocols (Macherey-Nagel), or parts produced by gene synthesis were cloned into the MIDAS pML1 vec-

tor by BsmBI-mediated Golden Gate assembly (BsmBI-GGA). As per the MIDAS design, BsmBI-GGA into the pML1 vector results in elimination of the BsmBI recognition sites and each part becomes flanked by BsaI recognition sites that cleave at the MIDAS prefix and suffix nucleotides.

In the case of the *Aequoria victoria* gfp TIsigner variants, each GFPN part cloned into the pML1 vector becomes flanked by BsaI recognition sites that are cleaved at the [CCAT] prefix and [GTTG] suffix (Fig A.2). Cloning of the GFPC part into the pML1 vector results in a GFPC module flanked by BsaI recognition sites that are cleaved at the [GTTG] prefix and at the [GCTT] suffix (Fig A.3).

In the case of the *Renilla reniformis* luciferase TIsigner variants, each RLucN part cloned into the pML1 vector becomes flanked by BsaI recognition sites that are cleaved at the [CCAT] prefix and [AGGA] suffix, while the C-terminal fragment, RLucC, becomes flanked by BsaI recognition sites that generate an [AGGA] prefix and a [GCTT] suffix upon cleavage.

All parts cloned into the pML1 vector were verified by sequencing.

MIDAS Level-2 assembly of devices

Devices were assembled from the cloned Level-1 modules described above, using BsaI-GGA, into the appropriate pML2 vector. As per the MIDAS design, multiple parts can be assembled together, with the position of each part in the assembled device dictated by the compatibility of the prefix and suffix nucleotides flanking each module:

- A lacI device was assembled in pML2(+)WR from the single lacI genetic element module.
- An mScarlet-I device was assembled in pML2(+)BR from nptII promoter, mScarlet-I CDS and lambda t0 transcription terminator modules.
- Full-length gfp devices for each TIsigner variant were assembled in pML2(+)WF from the following modules: T7lac promoter, GFPN, GFPC and T7 Tϕ transcription terminator. Since the prefix of the GFPC module, [GTTG] (see Fig

A.3), is identical to the suffix of each GFPN module (see Fig A.2) this allows the two modules to be genetically fused so that, together with the T7lac promoter and T7 Tϕ transcription terminator modules, full-length gfp devices are assembled for each variant. The mScarlet-I and gfp devices were assembled in pML2 vectors of opposite orientation (using the "Reverse" vector pML2(+)BR for mScarlet-I, and the "Forward" vector pML2(+)WF for each gfp device), so that they will be divergently transcribed once assembled into the expression vector (Level-3, see below).

- In a similar fashion, full-length luciferase devices for each TIsigner variant were assembled in pML2(+)BF from T7lac promoter, RLucN, RLucC and T7 Tϕ transcription terminator modules.

All cloned devices were verified by restriction mapping and sequencing.

MIDAS Level-3 assembly (construction of the expression plasmids)

E. coli gfp expression plasmids were constructed by sequentially loading the lacI, mScarlet-I and gfp devices, using alternating AarI- and BsmBI-GGA reactions, into the MIDAS Level-3 destination plasmid pML3.2, which has the medium copy replication origin from the pET series of vectors in place of the high copy pMB1 replicon of the pML3 destination vector originally described in van Dolleweerd et al, 2018 [62]. A representative map of an E. coli expression plasmid containing all three devices is shown in Fig A.4. For luciferase expression, the intermediate plasmid containing the lacI device (described above) was used for assembly of each of the luciferase devices (i.e., no mScarlet-I device was added), and a representative map of an E. coli plasmid for luciferase expression is shown in Fig A.5. The lacI and luciferase devices are divergently transcribed from the expression vector.

A.2 Supplementary figures

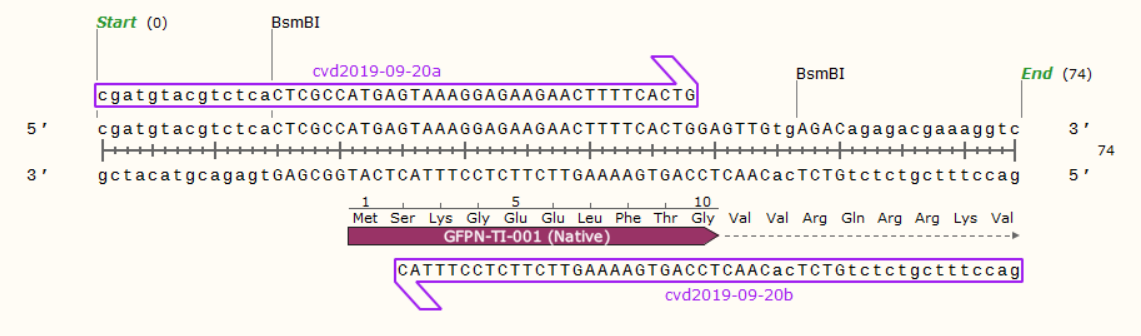


Figure A.1: **Strategy for producing a double stranded DNA corresponding to the first ten codons of each of the TIsigner variants of GFP.** The strategy employs a pair of primers that overlap at their 3' ends that, upon annealing together, can be used as a template for Q5 polymerase to generate a double stranded DNA spanning the N-terminal region of GFP (i.e. GFPN). Shown here is the sequence of variant GFPN-001 generated using the overlapping cvd2019-09-20a forward and cvd2019-09-20b reverse primer pair (see Table A.2). The resultant double stranded DNA can then be cloned into the MIDAS pML1 vector by digestion with the Type IIS restriction enzyme BsmBI (recognition site CGTCTC(1/5)). The same primer pair strategy was used for producing TIsigner variants of luciferase, albeit with a different [AGGA] suffix. This map was created with SnapGene.

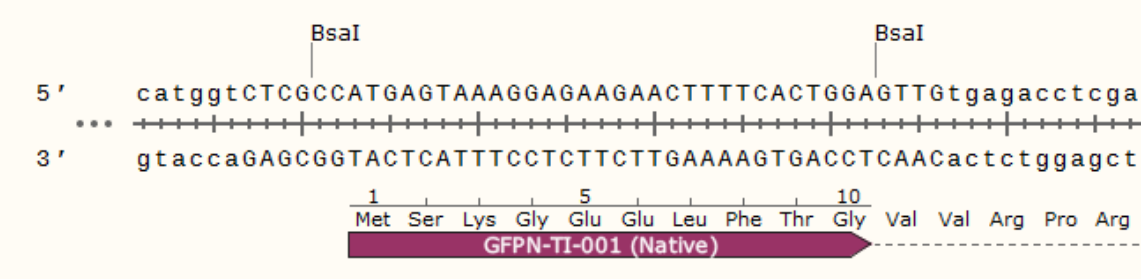


Figure A.2: **Structure of GFPN variants cloned into the pML1 vector.** Following BsmBI-mediated cloning into the pML1 vector, each GFPN variant becomes flanked by BsaI recognition sites (GGTCTC(1/5)). BsaI cleaves at the CCAT prefix upstream of the GFPN sequence and at the GTTG suffix (downstream of the GFPN module). For ease of depiction, only the sequences immediately surrounding the cloned GFPN fragment are shown (i.e., not the rest of the pML1 vector). The structure of luciferase RLucN variants is identical, except for having a different [AGGA] suffix sequence.

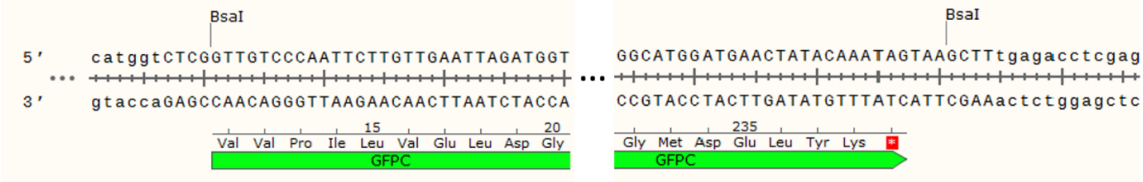


Figure A.3: **Structure of the GFPC fragment cloned into the pML1 vector.** Following BsmBI-mediated cloning into the pML1 vector, GFPC becomes flanked by BsaI recognition sites. BsaI cleaves at the [GTTG] prefix (upstream of the GFPC sequence) and at the downstream [GCTT] suffix. For ease of depiction, only sequences around the 5' and 3' ends of the GFPC fragment are shown (left- and right-hand sides, respectively). In the case of luciferase, the prefix sequence is [AGGA].

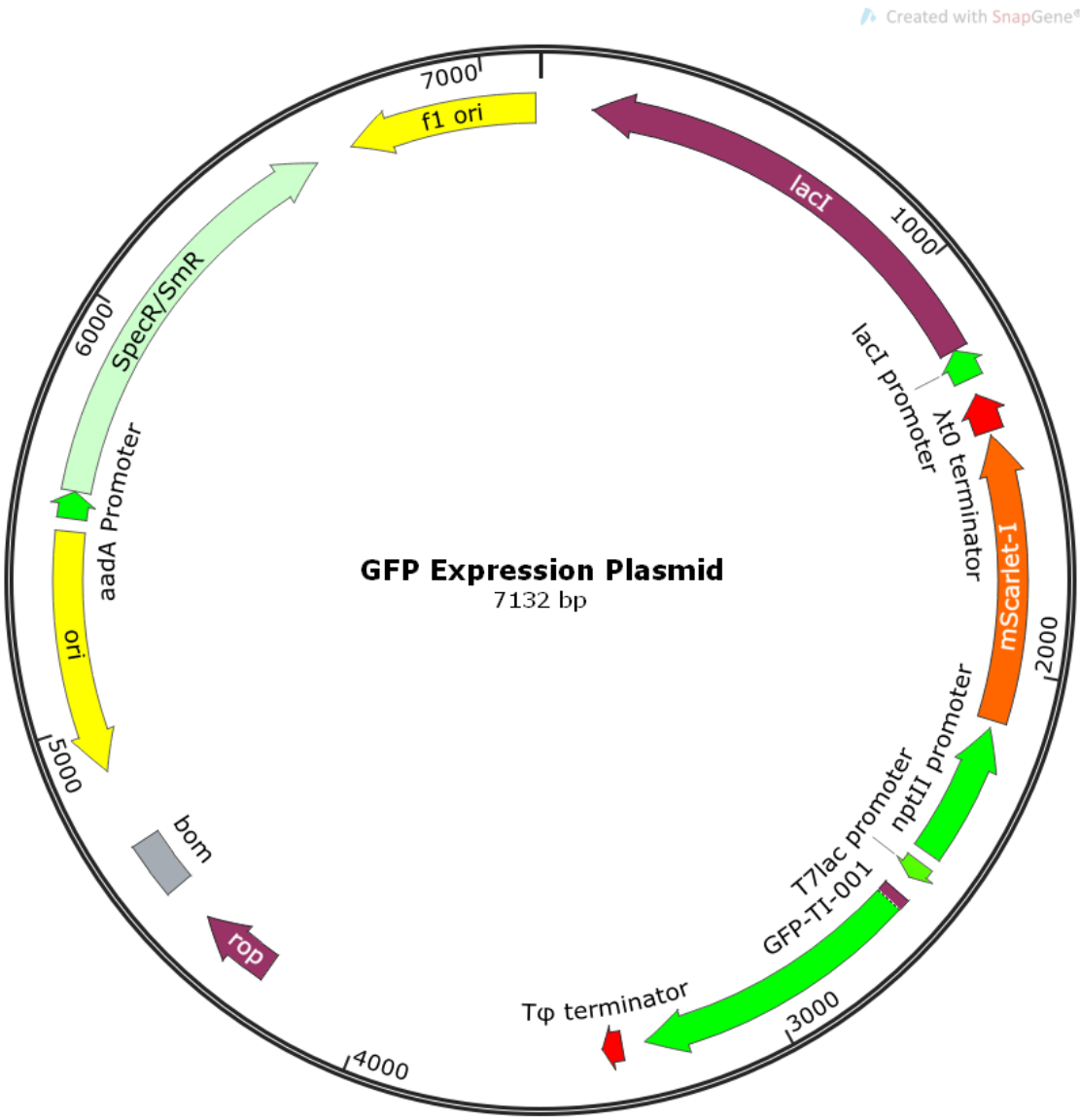


Figure A.4: **Structure of GFP expression plasmids.** Map view showing the architecture of MIDAS-assembled plasmids used for expression of GFP TIsigner variants. Expression of each of the gfp variants is controlled by the T7lac promoter and oriented such that they are divergently transcribed with respect to the mScarlet-I device, which is driven by the nptII promoter. The devices for lacI, mScarlet-I and gfp were loaded sequentially into plasmid pML3.2, which has the medium copy replication origin from the pET series of vectors (ori-bom-rop) in place of the high copy pMB1 replicon in the pML3 destination vector described in van Dolleweerd et al, 2018 [62], and a selectable marker conferring resistance to spectinomycin.

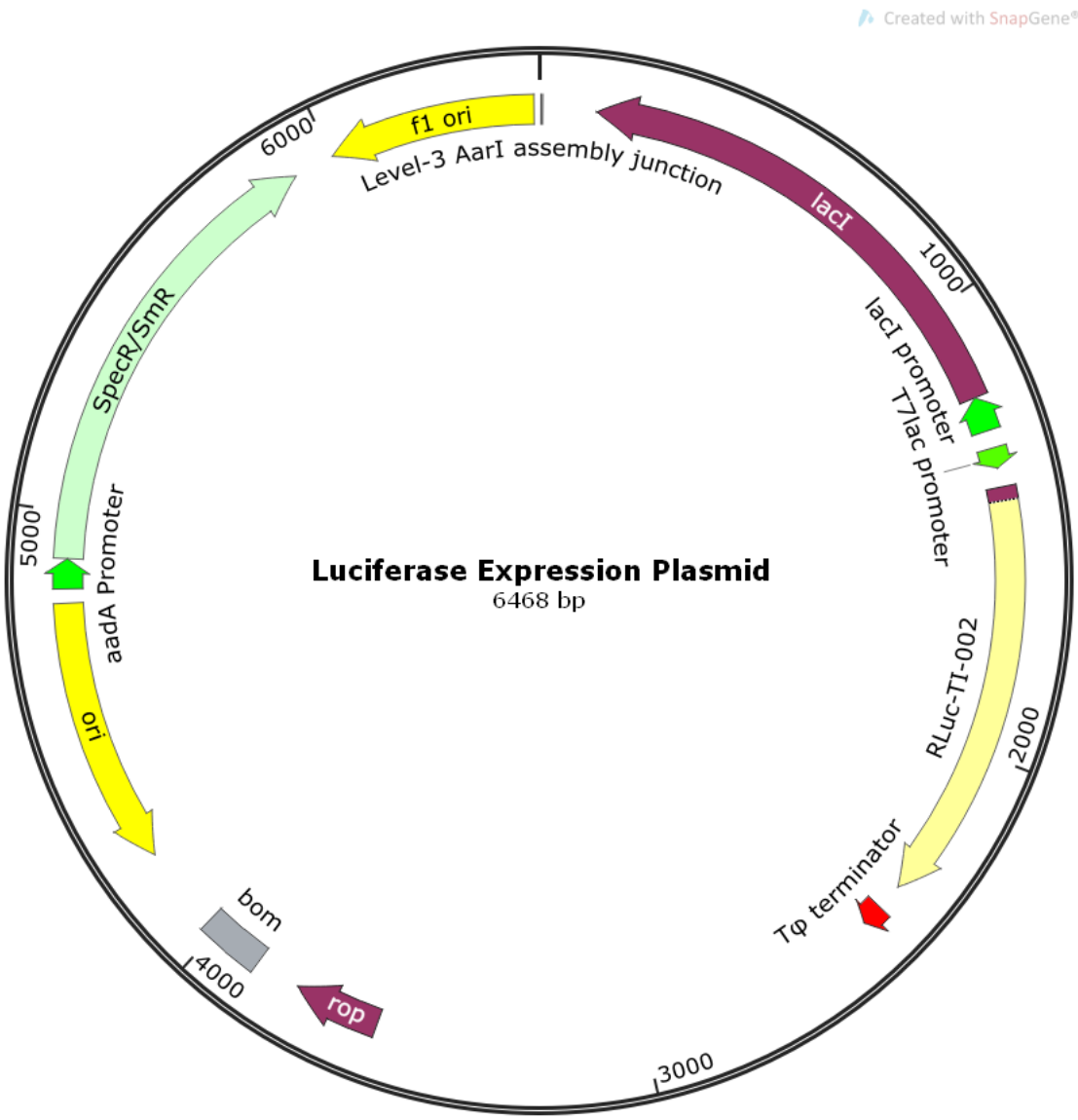


Figure A.5: **Structure of luciferase expression plasmids.** Map view showing the architecture of MIDAS-assembled plasmids used for expression of luciferase TIsigner variants. Expression of each of the luciferase variants is controlled by the T7lac promoter and oriented such that they are divergently transcribed with respect to the lacI device, which is driven by the lacI promoter. The devices for lacI and luciferase were loaded sequentially into plasmid pML3.2, which has the medium copy replication origin from the pET series of vectors (ori-bom-rop) in place of the high copy pMB1 replicon in the pML3 destination vector described in van Dolleweerd et al, 2018 [62], and a selectable marker conferring resistance to spectinomycin.

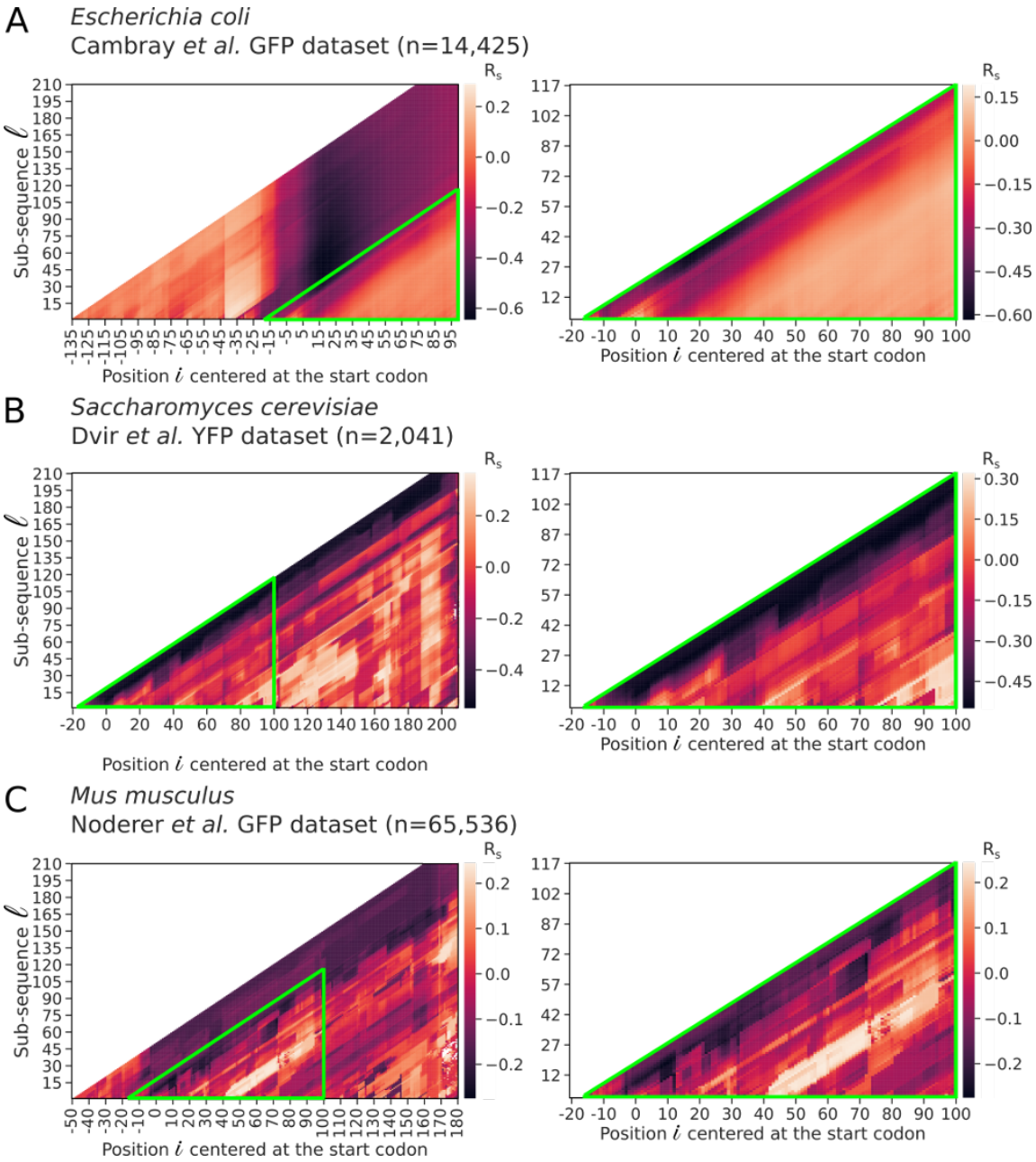


Figure A.6: **Heatmaps of correlations between opening energy and protein abundance for each of the sub-sequence regions (related to Fig 1).** Green unfilled triangles indicate the regions before and after scaling (left and right panels, respectively). (A) For E. coli, we used a representative GFP expression dataset from Cambray et al. (2018) [35]. The reporter library consists of GFP fused in-frame with a library of 96-nt upstream sequences (n=14,425). (B) For S. cerevisiae, we used a YFP expression dataset from Dvir et al. (2013) [66]. The YFP reporter library consists of 2,041 random decameric nucleotides inserted at the upstream of YFP start codon. (C) For M. musculus, we used the GFP expression dataset from Noderer et al. (2014) [175]. The GFP reporter library consists of 65,536 random hexameric and dimeric nucleotides inserted at the upstream and downstream of GFP start codon, respectively. R_s, Spearman's rho.

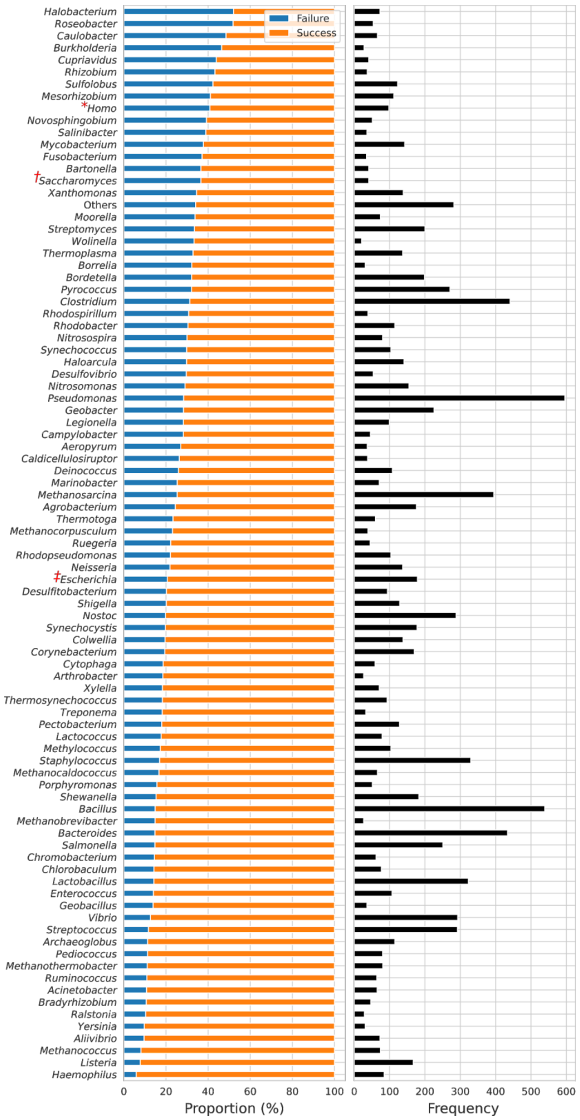


Figure A.7: **Expression outcomes of the PSI:Biology targets in *E. coli* (related to Fig 2.2C and 2.3).** A total of 11,430 PSI:Biology targets from over 189 species were analysed in this study (n=8,780 and 2,650, 'success' and 'failure' groups, respectively) [43, 212, 2]. Genera with at least 20 target genes are shown and the remaining as 'Others'. The top three PSI:Biology targets are from four Pseudomonas, five Bacillus and six Clostridium species. Red asterisk, obelisk and diesis indicate Homo sapiens, S. cerevisiae and E. coli, respectively. These target genes were inserted into the pET21_NESG expression vector, in which the promoter and fusion tag are T7lac and C-terminal His tag, respectively.

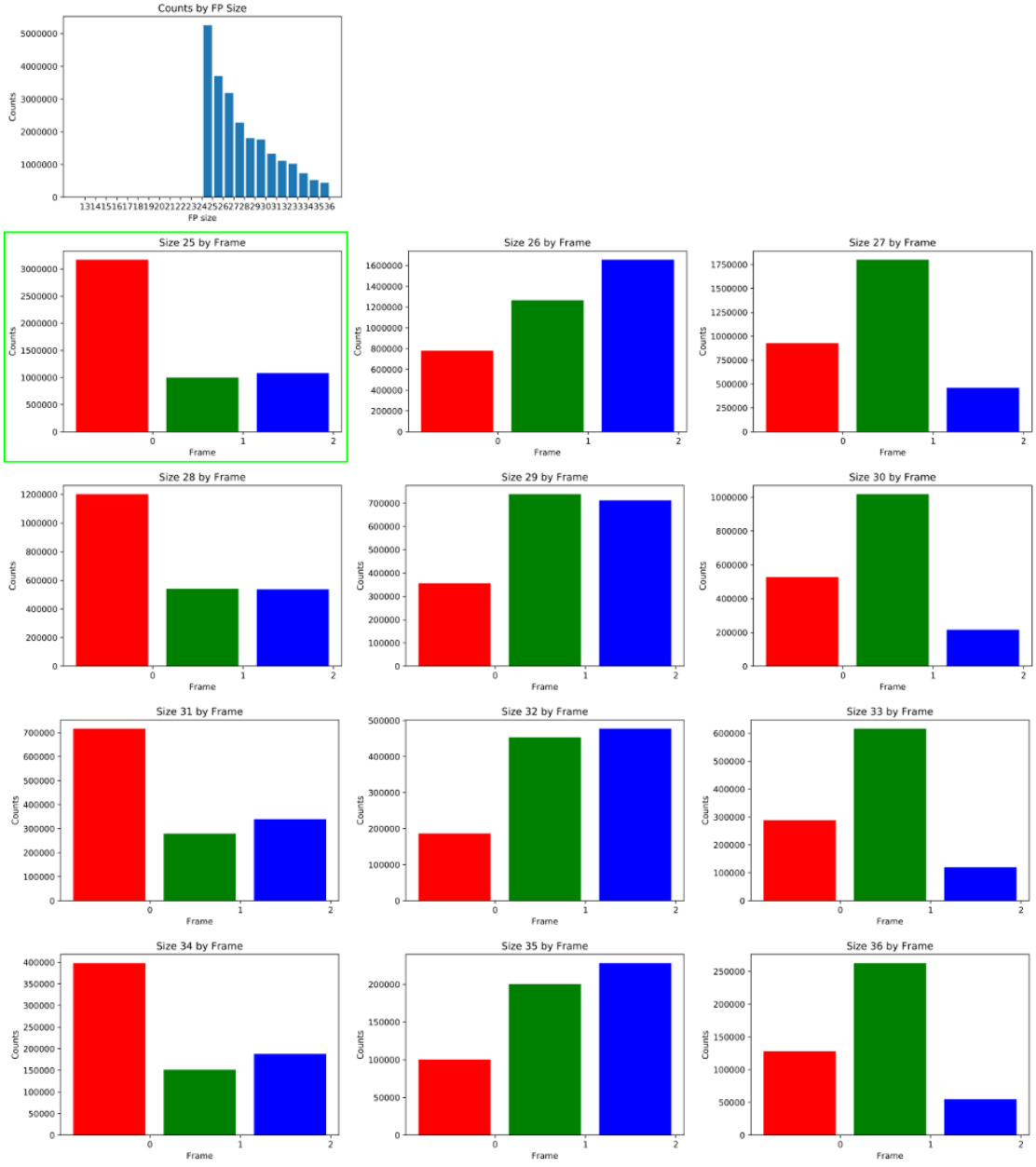


Figure A.8: **Ribosome footprints in 25-nt fragments show a strong triplet periodicity, indicating translation (related to Fig 2.3)** These 25-nt footprints (green unfilled rectangle) were used to train a neural network model [241] in order to predict the translation elongation rates of the PSI:Biology targets. Ribosome profiling data [SRR7759806 and SRR7759807 [162]] were first aligned to S. cerevisiae transcriptome. SAM alignment files were merged, and ribosome footprints which were mapped to each frame were enumerated. See https://github.com/Gardner-BinfLab/TIsigner_paper_2019. FP, footprints.

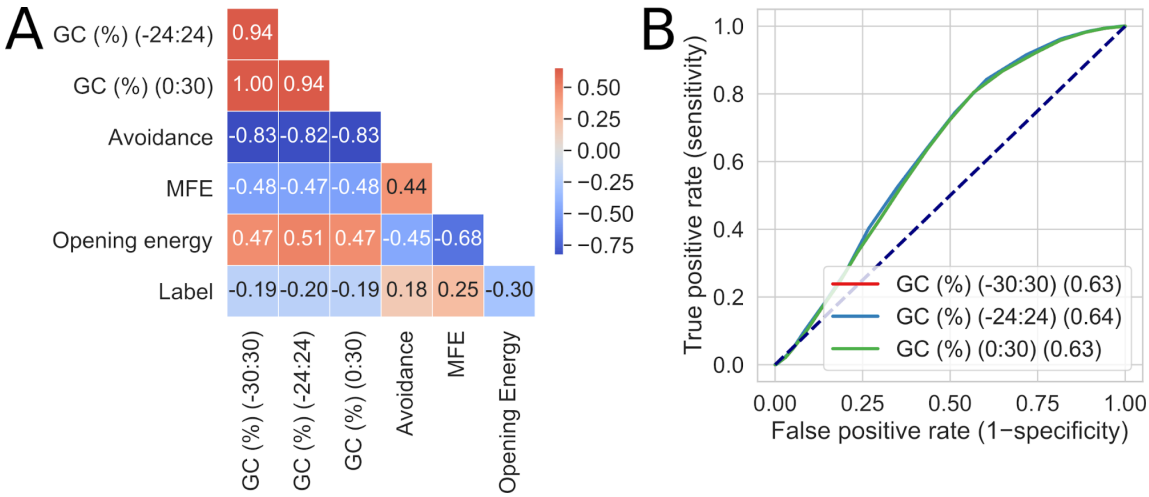


Figure A.9: **Analysis of the local G+C contents in the PSI:Biology target genes (related to Fig 2.3). (A)** The G+C contents in the regions -24:24 and -30:30 weakly correlate with opening energy and minimum free energy, respectively. Colourbar indicate Spearman's correlations (R_s) between the local G+C contents and the corresponding local features. **B** The local G+C contents show a similar prediction accuracy (AUC scores shown in parentheses).

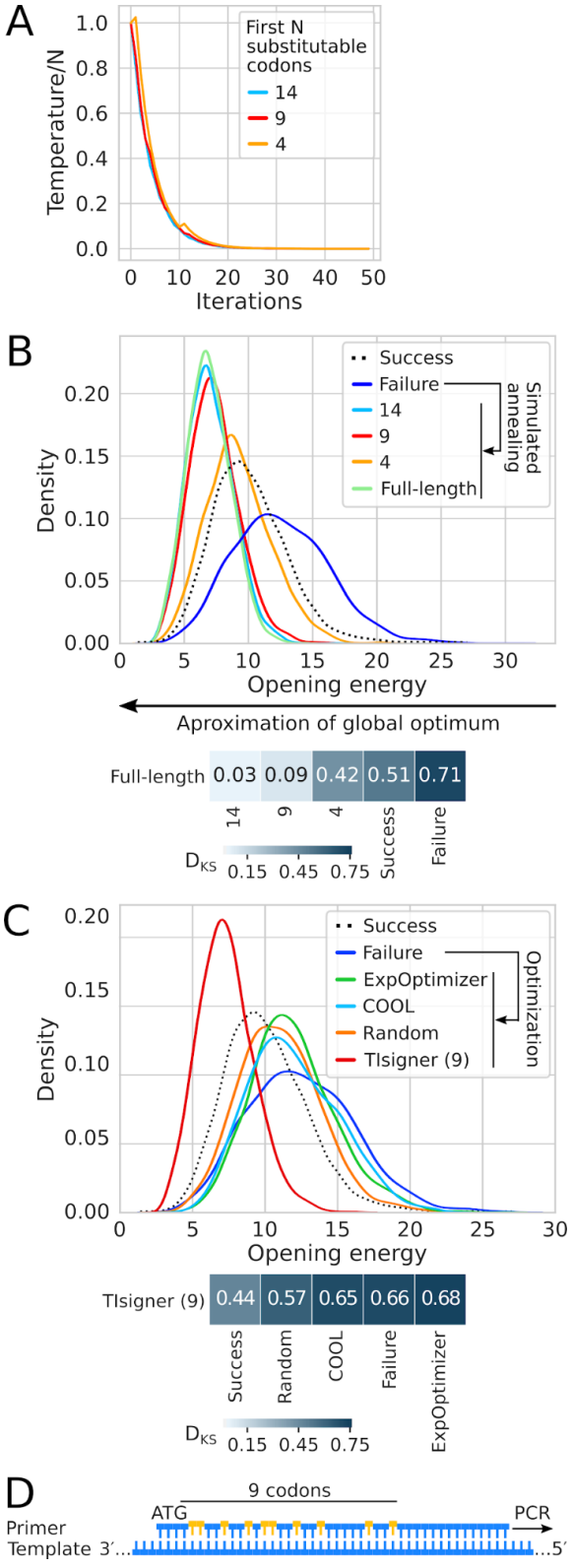


Figure A.10: **Accessibility of translation initiation sites can be increased by synonymous codon substitution within the first nine codons using simulated annealing.** **(A)** Schedules in simulated annealing. The ratio of temperature to the number of the first N substitutable codons decreases exponentially with increasing number of iterations. **(B)** Accessibility of translation initiation sites increases with increasing number of the first N replaceable codons. The PSI:Biology targets that failed to be expressed were optimised using simulated annealing (n=2,650). The Kolmogorov-Smirnov distance between the distributions of '9' and 'full-length' was significantly different but sufficiently close (DKS=0.09, P<10-7), indicating that optimisation of the first nine codons can achieve nearly optimum accessibility. For comparison, the distribution of the PSI:Biology targets that were successfully expressed are shown (n=8,780). See also Table S4. **(C)** Accessibility of translation initiation sites can be increased indirectly using the existing gene optimisation tools and random synonymous codon substitution. 'TIsigner (9)' refers to the default settings of our tool, which allows synonymous substitutions up to the first nine codons (as above). See also Table S4. **(D)** Accessibility of translation initiation sites can be optimised using PCR cloning. The forward primer should be designed according to TIsiger optimised sequences. For example, using a nested PCR approach, the optimised sequence can be produced using the forward primer designed with appropriate mismatches (gold bulges) to amplify the amplicon from the initial PCR reaction.

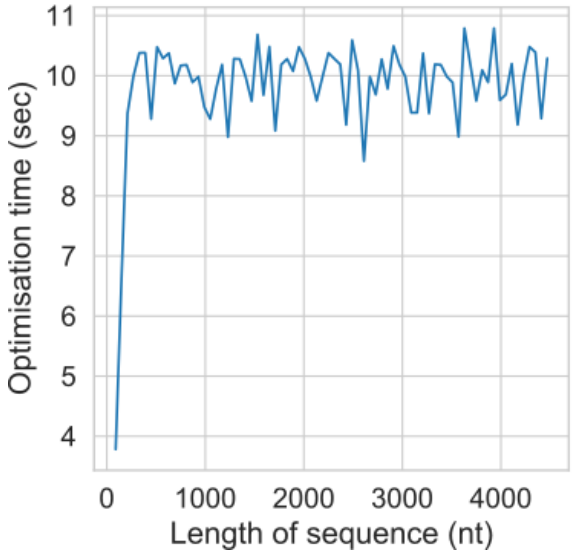


Figure A.11: **Sequence length does not affect software performance because only a fixed region is taken into account during optimisation ($\mathcal{O}(1)$ time).**

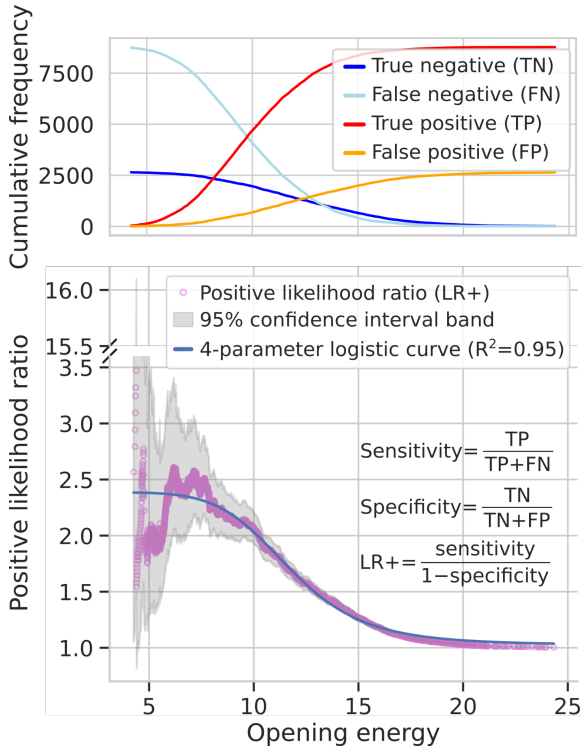


Figure A.12: **Opening energy of 10 or below at the region -24:24 is about two times more likely to come from the target genes that are successfully expressed than those that failed (related to Fig 2.3).** Cumulative frequency distributions of the true positive and false positive (less than type), and true negative and false negative (more than type) derived from the ROC analysis in Fig 2.3B (left panel, opening energy -24:24). These values were used to estimate positive likelihood ratios with 95% confidence intervals using 10,000 bootstrap replicates. The estimated ratios and/or confidence intervals are inaccurate at low numbers of true positives or true negatives. Therefore, a four-parameter logistic curve was fitted to the positive likelihood ratios. Fitted values are useful to estimate the posterior probability of protein expression.

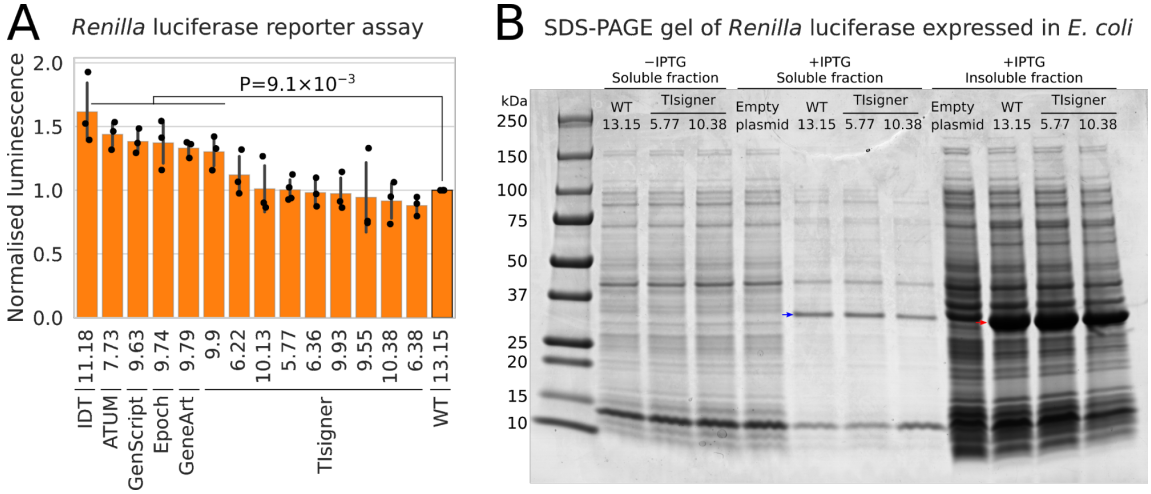


Figure A.13: **Luciferase reporter assay.** **(A)** The expression of RLuc can be improved, despite its poor solubility in E. coli. Opening energies are shown next to labels. The luciferase activities of commercially designed RLuc reporter genes (full-length sequence optimisation) and TIsigner (9.9 kcal/mol) are significantly higher than the wild-type luciferase (Mann-Whitney U tests, P=9.1×10^{-3}). **(B)** SDS-PAGE gel shows the protein bands of Renilla luciferase (RLuc) in the soluble and insoluble fractions of BL21Star(DE3) lysates. Selected bacterial clones were grown at 25° C, 200 RPM. The solubilities of wildtype (WT) RLuc and designed variants were compared after 4-hour IPTG induction. The blue and red arrows (36kDa) indicate that RLuc was poorly soluble. No RLuc protein bands were detected from the uninduced cultures and IPTG-induced negative control (empty vector control that lacks Rluc gene and T7lac promoter).

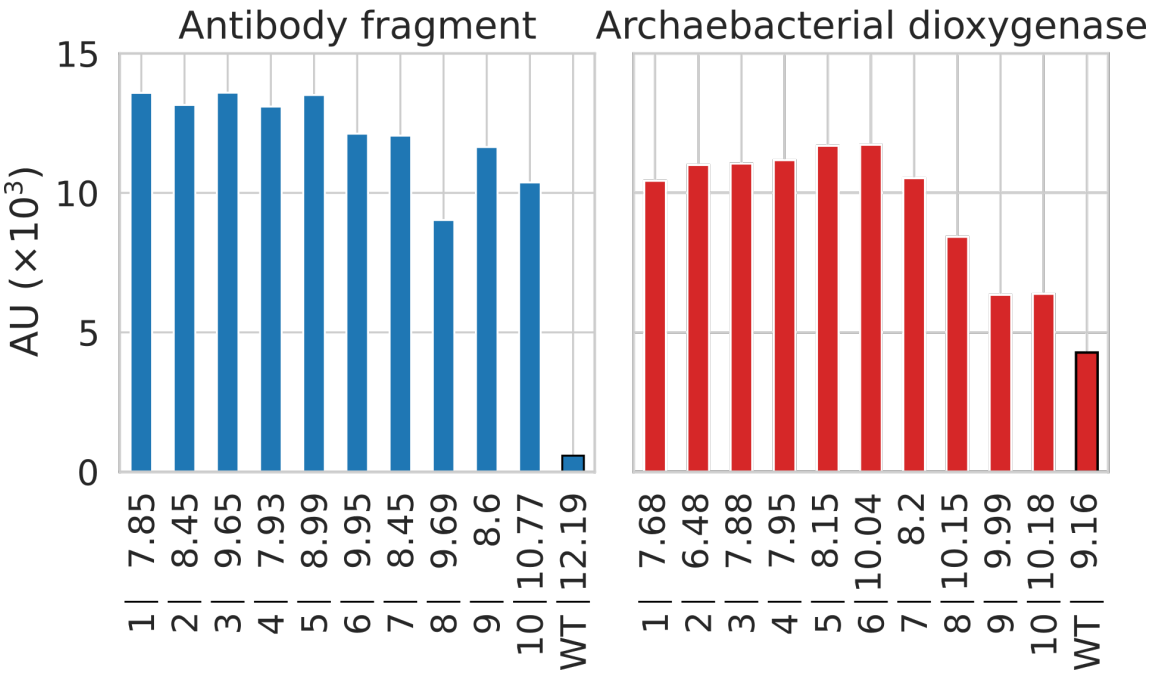


Figure A.14: **The yields of an antibody fragment and an archaebacterial dioxygenase can be improved by synonymous codon changes within the first six codons.** A RTS *E. coli* cell-free expression system was previously used to express these recombinant proteins (10). The expression levels are shown in arbitrary units (AU) based on the densitometric analysis of previously published Western blots (Supplementary Table S5). WT, wild-type.

A.3 Supplementary tables

Table A.1: MIDAS parts used in this work. The sequence of each part is shown, with BsmBI recognition sites used for cloning into the MIDAS pML1 vector. The T7lac promoter and T7 Tϕ transcription terminator (both sequences taken from pET-15b) were ordered as double-stranded DNA gBlocks from IDT. The lacI genetic element (with lacI promoter and lacI coding sequence) was from GeneArt. Other parts were amplified by polymerase chain reaction. Sequences of all parts were confirmed following cloning into the MIDAS pML1 vector.

Part	**Sequence**
T7lac promoter	cgatgtacgtctcaCTCGGGAGCGATCCCGCGAAATTAATACGACT CACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAA ATAATTTTGTTTAACTTTAAGAAGGAGATATACCATtgAGACag agacgaaaggtc
nptII promoter	cgatgtacgtctcaCTCGGGAGctcgcacgctgccgcaagcactcagggcg caagggctgctaaaggaagcggaacacgtagaaagccagtccgcagaaacgg tgctgaccccggatgaatgtcagctactgggctatctggacaagggaaaacgca agcgcaaagagaaagcaggtagcttgcagtgggcttacatggcgatagctaga ctgggcggtcttatggacagcaagcgaaccggaattgccagctggggcgccct ctggtaaggttgggaagccctgcaaagtaaactggatggctttcttgccgccaag gatctgatggcgcaggggatcaagatctgatcaagagacaggatactagtggag gaagaaaaAATGtgAGACagagacgaaaggtc

Table A.1 continued from previous page

Part	**Sequence**
lacI genetic element	cgatgtacgtctcaCTCGGGAGCGTCGAGATCCCGGACACCATCG AATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGG AAGAGAGTCAATTCAGGGTGGTGAATGTGAAACCAGTAACGT TATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGT TTCCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAA CGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTAC ATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAGTC GTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACG CGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGAT CAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAG CGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCG CGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATG ACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTTC CGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACA GTATTATTTTCTCCCATGAAGACGGTACGCGACTGGGCGTGG AGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAG CGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCT GGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAG

Table A.1 continued from previous page

Part	**Sequence**
lacI genetic element (cont..)	CGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAA CAAACCATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCG ATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCG CGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGATATCTC GGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTATATC CCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGG CAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCA GGCGGTGAAGGGCAATCAGCTGTTGCCCGTGTCACTGGTGAA AAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCC CCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGT TTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATG TAAGTTAGCTCACTCATTAGGCACCGGGATCTCGACCGATGCC CTTGAGAGCCTTCAACCCAGTCAGCTCCTTCCGGTGGGCGCG GGGCATGACTACGCTtgAGACagagacgaaaggtc

Table A.1 continued from previous page

Part	**Sequence**
mScarlet-I coding sequence (CDS)	cgatgtacgtctcaCTCGAATGgtgagcaagggcgaggcagtgatcaaggag ttcatgcggttcaaggtgcacatggagggctccatgaacggccacgagttcgagat cgagggcgagggcgagggccgcccctacgagggcacccagaccgccaagctga aggtgaccaagggtggcccctgcccttctcctgggacatcctgtcccctcagttca tgtacggctccagggccttcatcaagcaccccgccgacatccccgactactataag cagtccttccccgagggcttcaagtgggagcgcgtgatgaacttcgaggacggcg gcgccgtgaccgtgacccaggacacctccctggaggacggcaccctgatctacaa ggtgaagctccgcggcaccaacttccctcctgacggccccgtaatgcagaagaag acaatgggctgggaagcgtccaccgagcggttgtaccccgaggacggcgtgctg aagggcgacattaagatggccctgcgcctgaaggacggcggccgctacctggcg gacttcaagaccacctacaaggccaagaagcccgtgcagatgcccggcgcctaca acgtcgaccgcaagttggacatcacctcccacaacgaggactacaccgtggtgga acagtacgaacgctccgagggccgccactccaccggcggcatggacgagctgta caagtaaGCTTtgAGACagagacgaaaggtc

Table A.1 continued from previous page

Part	**Sequence**
Phage T7 Tϕ transcription terminator	cgatgtacgtctcaCTCGGCTTCAAAGCCCGAAAGGAAGCTGAGTT GGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGG GGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGG AACTATATCCGGATCGCTtgAGACagagacgaaaggtc
Lambda t0 transcription terminator	cgatgtacgtctcaCTCGGCTTggactcctgttgatagatccagtaatgacctc agaactccatctggatttgttcagaacgctcggttgccgccgggcgtttttttattggt gagaatccaagctagcttggCGCTtgAGACagagacgaaaggtc

Table A.2: Oligonucleotide primer pairs for constructing TIsigner variants of gfp. The sequences of each forward and reverse primer pair used for constructing each of the gfp TIsigner variants is shown. The start codon in each of the forward primers is shaded yellow. BsmBI recognition sites (used for Golden Gate assembly into the MIDAS pML1 vector) are underlined.

GFPN TIsigner ID	**Oligonucleotide Primer Pair**	**Primer Sequences (5' to 3')**
GFPN-001	cvd2019-09-20a	cgatgtacgtctcaCTCGCCATGAGTAAAGGAGAAGAACTTTTCACTG
	cvd2019-09-20b	gacctttcgtctctGTCTcaCAACTCCAGTGAAAAGTTCTTCTCCTTTAC
GFPN-002	cvd2019-09-21a	cgatgtacgtctcaCTCGCCATGTCGAAGGGTGAAGAACTCTTCAC
	cvd2019-09-21b	gacctttcgtctctGTCTcaCAACACCAGTGAAGAGTTCTTCACCCTTC
GFPN-003	cvd2019-09-21c	cgatgtacgtctcaCTCGCCATGAGTAAAGGGGAGGAACTCTTTAC
	cvd2019-09-21d	gacctttcgtctctGTCTcaCAACCCCGGTAAAGAGTTCCTCCCCTTTAC
GFPN-004	cvd2019-09-21e	cgatgtacgtctcaCTCGCCATGTCGAAGGGCGAAGAACTCTTC
	cvd2019-09-21f	gacctttcgtctctGTCTcaCAACACCAGTGAAGAGTTCTTCGCCCTTC
GFPN-005	cvd2019-09-21g	cgatgtacgtctcaCTCGCCATGTCTAAGGGTGAGGAGCTCTTC
	cvd2019-09-21h	gacctttcgtctctGTCTcaCAACTCCCGTGAAGAGCTCCTCACCCTTAG
GFPN-006	cvd2019-09-21i	cgatgtacgtctcaCTCGCCATGTCGAAAGGGGAAGAACTGTTCAC
	cvd2019-09-21j	gacctttcgtctctGTCTcaCAACGCCGGTGAACAGTTCTTCCCCTTTC
GFPN-007	cvd2019-09-21k	cgatgtacgtctcaCTCGCCATGTCTAAAGGAGAAGAGCTTTTCAC
	cvd2019-09-21l	gacctttcgtctctGTCTcaCAACCCCAGTGAAAAGCTCTTCTCCTTTAG
GFPN-008	cvd2019-09-21m	cgatgtacgtctcaCTCGCCATGAGTAAGGGTGAGGAATTATTCACG

Table A.2 continued from previous page

GFPN TIsigner ID	Oligonucleotide Primer Pair	Primer Sequences (5' to 3')
	cvd2019-09-21n	gacctttcgtctctGTCTcaCAACGCCCGTGAATAATTCCTCACCCTTAC
GFPN-009	cvd2019-09-22a	cgatgtacgtctcaCTCGCCATGAGTAAAGGGGAAGAACTGTTCAC
	cvd2019-09-22b	gacctttcgtctctGTCTcaCAACGCCAGTGAACAGTTCTTCCCCTTTAC
GFPN-010	cvd2019-09-22c	cgatgtacgtctcaCTCGCCATGTCTAAGGGTGAGGAGCTCTTC
	cvd2019-09-22d	gacctttcgtctctGTCTcaCAACTCCTGTGAAGAGCTCCTCACCCTTAG
GFPN-011	cvd2019-09-22e	cgatgtacgtctcaCTCGCCATGAGTAAAGGAGAAGAGTTATTTACTGG
	cvd2019-09-22f	gacctttcgtctctGTCTcaCAACTCCAGTAAATAACTCTTCTCCTTTAC
GFPN-012	cvd2019-09-22g	cgatgtacgtctcaCTCGCCATGAGTAAGGGAGAAGAGCTGTTC
	cvd2019-09-22h	gacctttcgtctctGTCTcaCAACTCCAGTGAACAGCTCTTCTCCCTTAC
GFPN-013	cvd2019-09-22i	cgatgtacgtctcaCTCGCCATGTCGAAAGGAGAAGAATTGTTCAC
	cvd2019-09-22j	gacctttcgtctctGTCTcaCAACGCCCGTGAACAATTCTTCTCCTTTCG
GFPN-014	cvd2019-09-22k	cgatgtacgtctcaCTCGCCATGAGCAAAGGAGAAGAATTATTTACTGG
	cvd2019-09-22l	gacctttcgtctctGTCTcaCAACTCCAGTAAATAATTCTTCTCCTTTGC
GFPN-015	cvd2019-09-22m	cgatgtacgtctcaCTCGCCATGAGCAAAGGAGAAGAATTATTTACGG
	cvd2019-09-22n	gacctttcgtctctGTCTcaCAACTCCCGTAAATAATTCTTCTCCTTTGC
GFPN-016	cvd2019-09-22o	cgatgtacgtctcaCTCGCCATGAGCAAAGGGGAAGAATTATTTACAG
	cvd2019-09-22p	gacctttcgtctctGTCTcaCAACACCTGTAAATAATTCTTCCCCTTTGC

Table A.2 continued from previous page

GFPN TIsigner ID	Oligonucleotide Primer Pair	Primer Sequences (5' to 3')
GFPN-017	cvd2020-03-05a	cgatgtacgtctcaCTCGCCATGAGTAAAGGGGAAGAACTCTTTACC
	cvd2020-03-05b	gacctttcgtctctGTCTcaCAACCCCGGTAAAGAGTTCTTCCCCTTTAC
GFPN-018	cvd2020-03-05c	cgatgtacgtctcaCTCGCCATGTCGAAAGGTGAGGAACTATTCACTG
	cvd2020-03-05d	gacctttcgtctctGTCTcaCAACACCAGTGAATAGTTCCTCACCTTTC
GFPN-019	cvd2020-03-05e	cgatgtacgtctcaCTCGCCATGTCGAAGGGTGAAGAACTGTTCACTG
	cvd2020-03-05f	gacctttcgtctctGTCTcaCAACACCAGTGAACAGTTCTTCACCCTTC
GFPN-020	cvd2020-03-05g	cgatgtacgtctcaCTCGCCATGTCGAAGGGTGAAGAACTTTTCACTG
	cvd2020-03-05h	gacctttcgtctctGTCTcaCAACCCCAGTGAAAAGTTCTTCACCCTTC
GFPN-021	cvd2020-03-05i	cgatgtacgtctcaCTCGCCATGTCCAAAGGGGAGGAACTCTTTACG
	cvd2020-03-05j	gacctttcgtctctGTCTcaCAACGCCCGTAAAGAGTTCCTCCCCTTTG
GFPN-022	cvd2020-03-05k	cgatgtacgtctcaCTCGCCATGTCCAAAGGTGAAGAGCTTTTCACC
	cvd2020-03-05l	gacctttcgtctctGTCTcaCAACCCCGGTGAAAAGCTCTTCACCTTTG
GFPN-023	cvd2020-03-05m	cgatgtacgtctcaCTCGCCATGTCGAAAGGTGAAGAGCTGTTCAC
	cvd2020-03-05n	gacctttcgtctctGTCTcaCAACACCGGTGAACAGCTCTTCACCTTTC
GFPN-024	cvd2020-03-05o	cgatgtacgtctcaCTCGCCATGTCGAAAGGTGAGGAACTGTTCAC
	cvd2020-03-05p	gacctttcgtctctGTCTcaCAACCCCAGTGAACAGTTCCTCACCTTTC
GFPN-025	cvd2020-03-05q	cgatgtacgtctcaCTCGCCATGAGTAAGGGGGAGGAGCTCTTCAC

Table A.2 continued from previous page

GFPN TIsigner ID	Oligonucleotide Primer Pair	Primer Sequences (5' to 3')
	cvd2020-03-05r	gacctttcgtctctGTCTcaCAACTCCGGTGAAGAGCTCCTCCCCCTTAC
GFPN-026	cvd2020-03-05s	cgatgtacgtctcaCTCGCCATGAGTAAAGGGGAAGAGCTTTTCAC
	cvd2020-03-05t	gacctttcgtctctGTCTcaCAACCCCGGTGAAAAGCTCTTCCCCTTTAC
GFPN-027	cvd2020-03-06a	cgatgtacgtctcaCTCGCCATGAGTAAAGGAGAAGAACTCTTTACCG
	cvd2020-03-06b	gacctttcgtctctGTCTcaCAACTCCGGTAAAGAGTTCTTCTCCTTTAC
GFPN-028	cvd2020-03-06c	cgatgtacgtctcaCTCGCCATGAGTAAAGGAGAAGAACTCTTCACC
	cvd2020-03-06d	gacctttcgtctctGTCTcaCAACACCGGTGAAGAGTTCTTCTCCTTTAC
GFPN-029	cvd2020-03-06e	cgatgtacgtctcaCTCGCCATGTCAAAGGGGGAAGAACTGTTCAC
	cvd2020-03-06f	gacctttcgtctctGTCTcaCAACGCCTGTGAACAGTTCTTCCCCCTTTG
GFPN-030	cvd2020-03-06g	cgatgtacgtctcaCTCGCCATGTCGAAAGGCGAGGAACTGTTCAC
	cvd2020-03-06h	gacctttcgtctctGTCTcaCAACTCCAGTGAACAGTTCCTCGCCTTTC
GFPN-031	cvd2020-03-06i	cgatgtacgtctcaCTCGCCATGAGCAAGGGTGAAGAGTTATTCACTG
	cvd2020-03-06j	gacctttcgtctctGTCTcaCAACTCCAGTGAATAACTCTTCACCCTTG
GFPN-032	cvd2020-03-06k	cgatgtacgtctcaCTCGCCATGTCTAAAGGTGAAGAACTATTCACAGG
	cvd2020-03-06l	gacctttcgtctctGTCTcaCAACCCCTGTGAATAGTTCTTCACCTTTAG
GFPN-033	cvd2020-03-06m	cgatgtacgtctcaCTCGCCATGTCTAAAGGTGAGGAGCTCTTCAC
	cvd2020-03-06n	gacctttcgtctctGTCTcaCAACTCCTGTGAAGAGCTCCTCACCTTTAG

Table A.2 continued from previous page

GFPN TIsigner ID	Oligonucleotide Primer Pair	Primer Sequences (5' to 3')
GFPN-034	cvd2020-03-06o	cgatgtacgtctcaCTCGCCATGAGTAAGGGAGAGGAACTGTTCAC
	cvd2020-03-06p	gacctttcgtctctGTCTcaCAACCCCTGTGAACAGTTCCTCTCCCTTAC
GFPN-035	cvd2020-03-06q	cgatgtacgtctcaCTCGCCATGTCGAAAGGGGAAGAATTGTTCAC
	cvd2020-03-06r	gacctttcgtctctGTCTcaCAACTCCAGTGAACAATTCTTCCCCTTTCG
GFPN-036	cvd2020-03-06s	cgatgtacgtctcaCTCGCCATGAGTAAGGGGGAGGAGCTGTTC
	cvd2020-03-06t	gacctttcgtctctGTCTcaCAACTCCTGTGAACAGCTCCTCCCCCTTAC
GFPN-037	cvd2020-03-07a	cgatgtacgtctcaCTCGCCATGAGTAAGGGAGAGGAATTGTTCAC
	cvd2020-03-07b	gacctttcgtctctGTCTcaCAACACCCGTGAACAATTCCTCTCCCTTAC
GFPN-038	cvd2020-03-07c	cgatgtacgtctcaCTCGCCATGAGTAAGGGAGAGGAACTTTTCAC
	cvd2020-03-07d	gacctttcgtctctGTCTcaCAACTCCCGTGAAAAGTTCCTCTCCCTTAC
GFPN-039	cvd2020-03-07e	cgatgtacgtctcaCTCGCCATGAGTAAAGGAGAGGAGCTTTTCACAG
	cvd2020-03-07f	gacctttcgtctctGTCTcaCAACTCCTGTGAAAAGCTCCTCTCCTTTAC
GFPN-040	cvd2020-03-07g	cgatgtacgtctcaCTCGCCATGAGCAAAGGAGAAGAGTTATTTACAGG
	cvd2020-03-07h	gacctttcgtctctGTCTcaCAACCCCTGTAAATAACTCTTCTCCTTTGC
GFPN-041	cvd2020-03-07i	cgatgtacgtctcaCTCGCCATGAGCAAAGGAGAGGAATTATTTACG
	cvd2020-03-07j	gacctttcgtctctGTCTcaCAACGCCCGTAAATAATTCCTCTCCTTTGC
GFPN-042	cvd2020-05-15a	cgatgtacgtctcaCTCGCCATGAGTAAAGGGGAGGAACTCTTTACTG

Table A.2 continued from previous page

GFPN TIsigner ID	Oligonucleotide Primer Pair	Primer Sequences (5' to 3')
	cvd2020-05-15b	gacctttcgtctctGTCTcaCAACACCAGTAAAGAGTTCCTCCCCTTTAC
GFPN-043	cvd2020-05-15c	cgatgtacgtctcaCTCGCCATGTCGAAAGGTGAAGAACTTTTCACTG
	cvd2020-05-15d	gacctttcgtctctGTCTcaCAACACCAGTGAAAAGTTCTTCACCTTTCG
GFPN-044	cvd2020-05-15e	cgatgtacgtctcaCTCGCCATGAGCAAGGGAGAAGAGCTGTTCACTG
	cvd2020-05-15f	gacctttcgtctctGTCTcaCAACGCCAGTGAACAGCTCTTCTCCC
GFPN-045	cvd2020-05-15g	cgatgtacgtctcaCTCGCCATGAGTAAGGGTGAGGAGTTATTCACG
	cvd2020-05-15h	gacctttcgtctctGTCTcaCAACGCCCGTGAATAACTCCTCACCCTTAC
GFPN-046	cvd2020-05-15i	cgatgtacgtctcaCTCGCCATGTCTAAAGGAGAAGAACTCTTCACAGG
	cvd2020-05-15j	gacctttcgtctctGTCTcaCAACCCCTGTGAAGAGTTCTTCTCCTTTAG
GFPN-047	cvd2020-05-15k	cgatgtacgtctcaCTCGCCATGTCCAAAGGAGAAGAACTATTCACC
	cvd2020-05-15l	gacctttcgtctctGTCTcaCAACTCCGGTGAATAGTTCTTCTCCTTTGG
GFPN-048	cvd2020-05-15m	cgatgtacgtctcaCTCGCCATGAGCAAAGGAGAAGAACTATTCACG
	cvd2020-05-15n	gacctttcgtctctGTCTcaCAACTCCCGTGAATAGTTCTTCTCCTTTGC
GFPN-049	cvd2020-05-15o	cgatgtacgtctcaCTCGCCATGTCGAAGGGAGAAGAATTATTTACGG
	cvd2020-05-15p	gacctttcgtctctGTCTcaCAACTCCCGTAAATAATTCTTCTCCCTTCG

Table A.3: Oligonucleotide primer pairs for constructing TIsigner variants of luciferase. The sequences of each forward and reverse primer pair used for constructing each of the luciferase TIsigner variants is shown. The start codon in each of the forward primers is shaded yellow. BsmBI recognition sites (used for Golden Gate assembly into the MIDAS pML1 vector) are underlined.

RLucN TIsigner ID	**Oligonucleotide Primer Pair**	**Primer Sequences (5' to 3')**
RLucN-TI-002	cvd2019-06-14c	cgatgtacgtctcaCTCGCCATGACATCAAAAGTATACGACCCAGAG
	cvd2019-06-14d	gacctttcgtctctGTCTcaTCCTCTGCTCTGGGTCGTATACTTTTGATG
RLucN-TI-003	cvd2019-06-14e	cgatgtacgtctcaCTCGCCATGACAAGTAAAGTTTATGACCCAGAGC
	cvd2019-06-14f	gacctttcgtctctGTCTcaTCCTCTGCTCTGGGTCATAAACTTTACTTG
RLucN-TI-004	cvd2019-06-14g	cgatgtacgtctcaCTCGCCATGACCAGCAAAGTTTATGACCCAGAG
	cvd2019-06-14h	gacctttcgtctctGTCTcaTCCTCTGCTCTGGGTCATAAACTTTGCTG
RLucN-TI-005	cvd2019-06-14i	cgatgtacgtctcaCTCGCCATGACAAGCAAAGTTTATGACCCAGAGC
	cvd2019-06-14j	gacctttcgtctctGTCTcaTCCTCTGCTCTGGGTCATAAACTTTGC
RLucN-TI-006	cvd2019-06-14k	cgatgtacgtctcaCTCGCCATGACTTCGAAAGTTTATGATCCAGAACAG
	cvd2019-06-14l	gacctttcgtctctGTCTcaTCCTCTGTTCTGGATCATAAACTTTCGAAG
RLucN-TI-007	cvd2019-06-14m	cgatgtacgtctcaCTCGCCATGACATCAAAAGTTTATGATCCAGAACAAAG
	cvd2019-06-14n	gacctttcgtctctGTCTcaTCCTTTGTTCTGGATCATAAACTTTTGATGTC
RLucN-TI-008	cvd2019-06-14o	cgatgtacgtctcaCTCGCCATGACGTCGAAAGTTTACGATCCAG
	cvd2019-06-14p	gacctttcgtctctGTCTcaTCCTTTGTTCTGGATCGTAAACTTTCGACG
RLucN-TI-009	cvd2019-06-14q	cgatgtacgtctcaCTCGCCATGACATCGAAAGTTTACGATCCAGAAC
	cvd2019-06-14r	gacctttcgtctctGTCTcaTCCTTTGTTCTGGATCGTAAACTTTCGATG
RLucN-TI-010	cvd2019-06-14s	cgatgtacgtctcaCTCGCCATGACCTCGAAAGTTTATGACCCAGAAC
	cvd2019-06-14t	gacctttcgtctctGTCTcaTCCTTTGTTCTGGGTCATAAACTTTCGAG

Appendix B

Solubility-Weighted Index: fast and accurate prediction of protein solubility

B.1 Supplementary notes

The B-factor or temperature factor of the atom in a crystalline structure is the measure of mean squared displacement $u = \langle (x - x_0)^2 \rangle$, where x is the displacement of atom from its mean position x_0. The B-factor thus reflects the *orderedness* of the crystal lattice and subsequent uncertainty in X-ray scattering structure determination [209, 36, 30]. It has unit of Å^2.

$$B = 8\pi^2 u$$

Since the distribution of B-factors varies with protein crystal structures, experimentally determined B-factors (for example from the Protein Data Bank) are not generalisable without appropriate normalisation. To address this issue, the B-factors of C_α atoms were extracted from a number of high-resolution protein crystal structures and normalised [209, 220, 124, 248]. The normalisation is often done by Z-scoring, for example, for a residue i, $B^i_{norm} = (B^i - \langle B \rangle)\sigma$, where σ is the standard deviation and $\langle B \rangle$ is the mean of B-factors within the polypeptide chain.

The profile of normalised B-factors along a protein sequence can be calculated using

a sliding window approach [e.g., 9 amino acid residues as implemented in Biopython [248, 50]]. The profile plot can be used to visualise and infer the local flexibility and dynamics of the protein structure [124, 248]. Previous studies that formulated flexibility also compared their computed values with the B-factors of previously solved protein structures using correlation tests [248, 250].

To calculate global structural flexibility, we reasoned that Vihinen *et al.*'s [250] sliding window method can be approximated by a more straightforward arithmetic mean. This sliding window method computes the local flexibility f_i of a given amino acid residue i as:

$$f_i = \frac{1}{5.25}[B_i + 0.8125(B_{i-1} + B_{i+1}) + 0.625(B_{i-2} + B_{i+2}) \\ +0.4375(B_{i-3} + B_{i+3}) + 0.25(B_{i-4} + B_{i+4})]$$

where, B_i is the normalised B-factor of the i^{th} C_α atom and so on. The arithmetic mean of these f_i can be approximately written as:

$$F = \langle f_i \rangle \approx \frac{1}{5.25(n-9)}(1 + 2 \times (0.8125 + 0.625 + 0.4375 + 0.25)) \sum_{i=5}^{n-4} B_i \\ = \frac{1}{n-9} \sum_{i=5}^{n-4} B_i$$

where, n is the number of residues in the protein. For sequence composition scoring, the arithmetic mean of B_i of a given full-length sequence is written as:

$$F' = \langle B \rangle = \frac{1}{n}(\sum_{i=1}^{n} B_i)$$

Approximating that the sums run at equal intervals, we can write:

$$\frac{F}{F'} = \frac{\langle f_i \rangle}{\langle B \rangle} \approx \frac{n}{n-9}$$

$n/(n-9)$ is monotonically decreasing for $n \geq 10$ and quickly approaches 1 with an increasing n. Thus, $\langle f_i \rangle$ is nearly equal to $\langle B \rangle$ and they are strongly correlated.

B.2 Supplementary figures

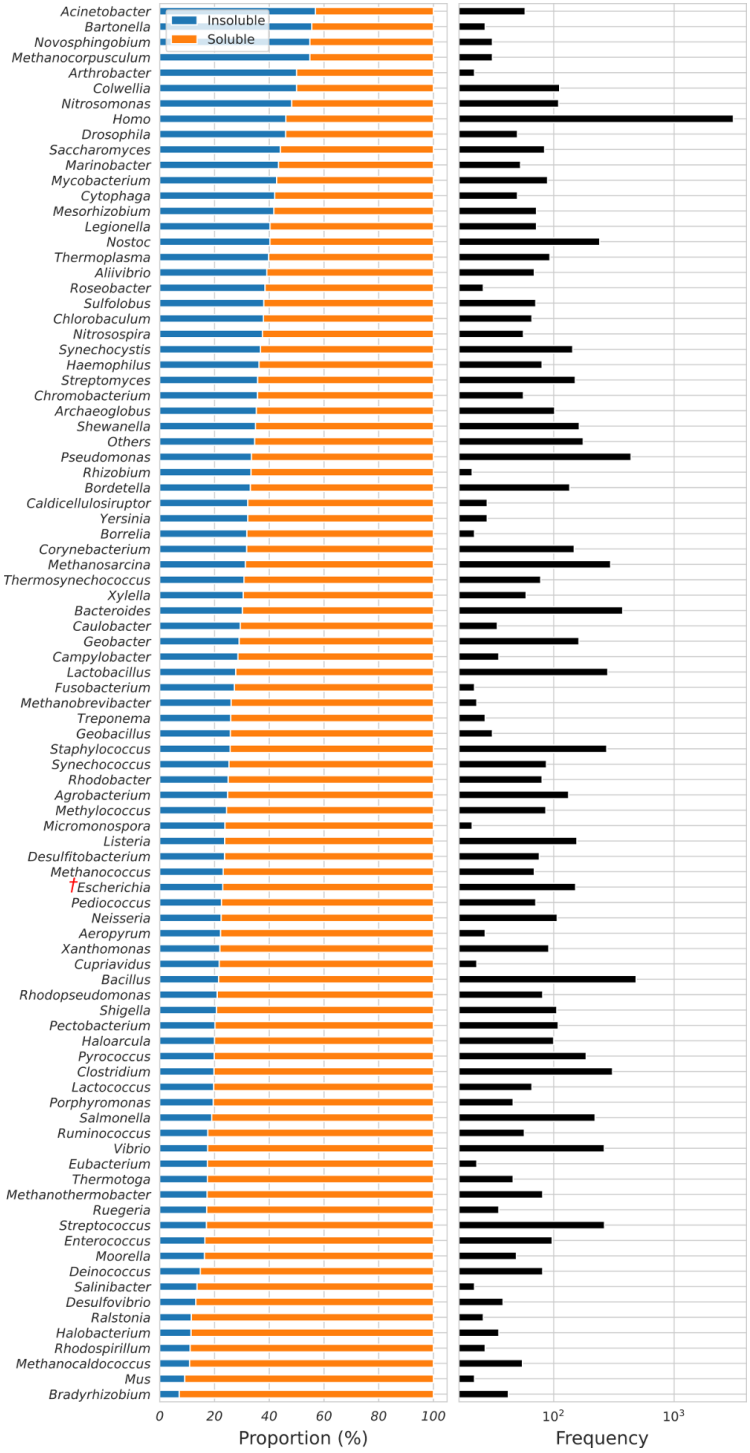


Figure B.1: **Solubility of the PSI:Biology targets grouped by source.** A total of 12,216 PSI:Biology targets from over 196 species were analysed in this study (8,238 soluble and 3,978 insoluble proteins). Genera with at least 20 target genes are shown and the remaining as 'Others'. Red obelisk indicates *E. coli*.

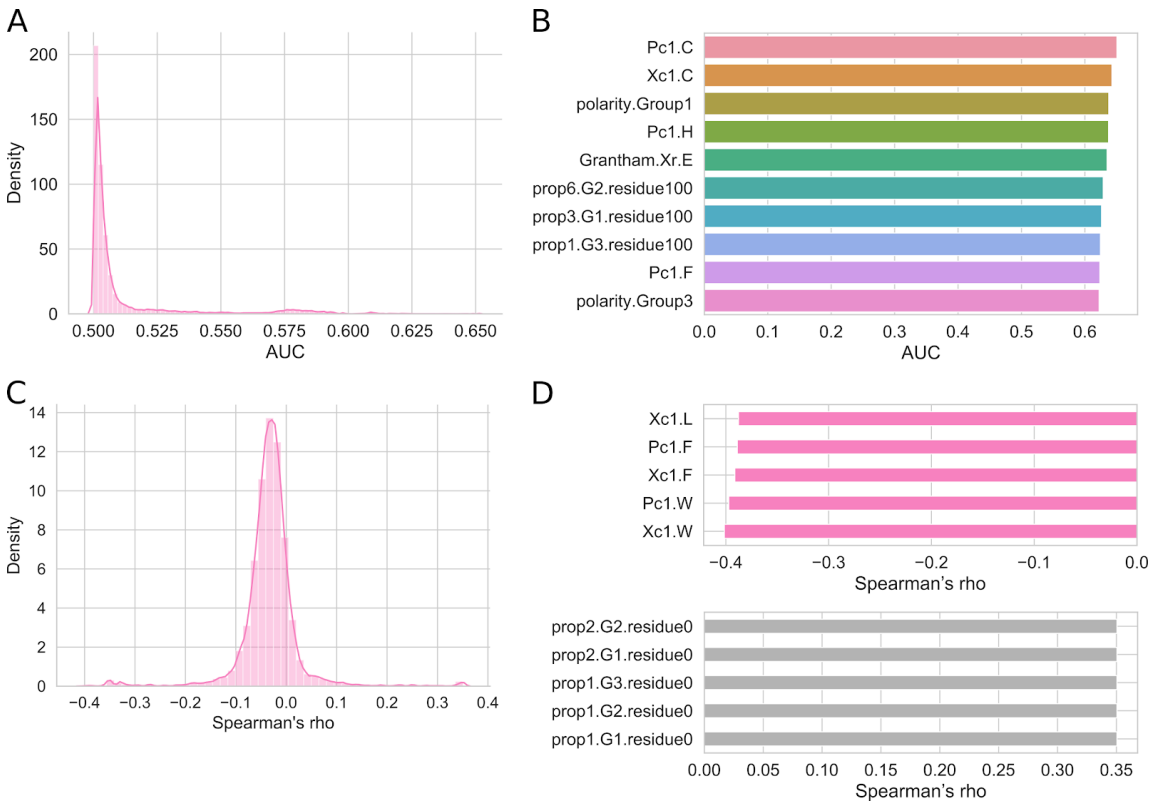


Figure B.2: **Prediction accuracy of 9,920 miscellaneous protein sequence properties.** Density distribution of AUC scores shows that relatively few features have high prediction accuracy (PSI:Biology dataset, N = 12,216). (B) Top-ranked features by AUC scores, which include the (amphiphilic) pseudo-amino acid compositions for cysteine residues (Pc1.C and Xc.1.C). (C) Density distribution of Spearman's rho shows that relatively few features have strong correlation coefficients with E. coli protein solubility (eSOL dataset, N = 3,198). (D) Top-ranked features by Spearman's correlation coefficients, which include the (amphiphilic) pseudo-amino acid compositions for aromatic amino acid residues (Xc1.W, Pc1.W, Xc1.F, and Pc1.F). The complete list of AUC scores and Spearman's correlation coefficients are available in Supplementary Table B.2. AUC, Area Under the ROC Curve; Pc1, amphiphilic pseudo-amino acid composition; polarity.Group1, one of the three groups of amino acid residues based on polarity (L, I, F, W, C, M, V, Y); polarity.Group3, one of the three groups of amino acid residues based on polarity (H, Q, R, K, N, E, D); prop$\{1-7\}$.G$\{1,2,3\}$.residue$\{0,25,50,100\%\}$, position percent for one of the three groups of amino acid residues by one of the seven properties listed in Table 1 of the protr vignettes, https://cran.r-project.org/web/packages/protr/vignettes/protr.html; PSI:Biology, Protein Structure Initiative:Biology; ROC, Grantham.Xr, Quasi-sequence-order based on Grantham's chemical distance matrix; Receiver Operating Characteristic; Xc1, pseudo-amino acid composition.

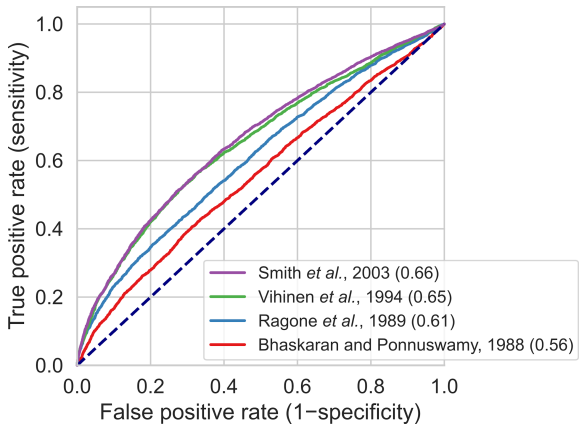


Figure B.3: **ROC analysis of sequence composition scores for solubility using previously published sets of normalised B-factors.**The PSI:Biology dataset (N = 12,216) was used for solubility prediction. AUC scores (perfect = 1.00, random = 0.50) are shown in parentheses. Dashed lines denote the performance of random classifiers. PSI:Biology, Protein Structure Initiative:Biology; ROC, Receiver Operating Characteristic.

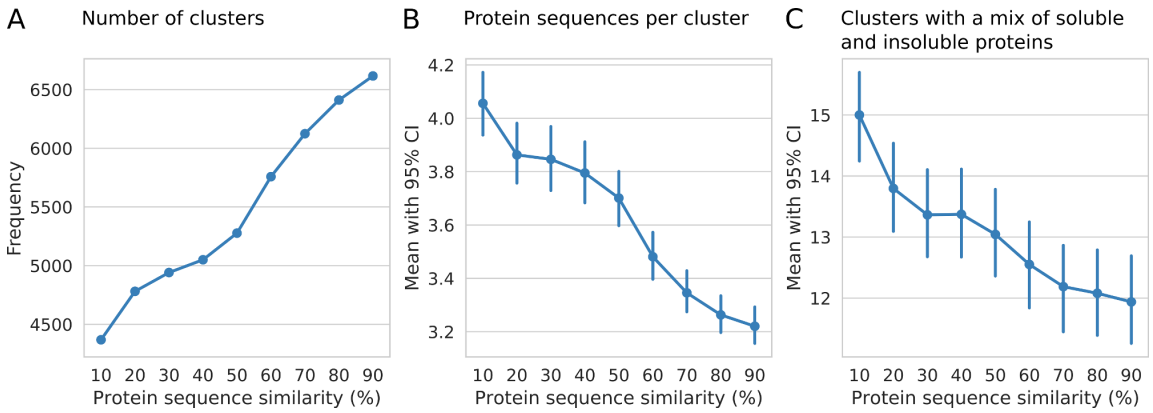


Figure B.4: **Relationship between protein solubility and sequence similarity, related to Fig 3.2** USEARCH was used to cluster the PSI:Biology targets (N = 12,216) at different percent similarity cutoffs (using the parameter -id 0.1 to 0.9; see https://drive5.com/usearch/manual/uclust_algo.html). **(A)** High numbers of clusters across different similarity cutoffs and **(B)** low numbers of sequences per cluster indicate that the PSI:Biology targets are highly diverse (Fig B.1). (C) Over about 12% of clusters contain a mix of soluble and insoluble proteins across different similarity cutoffs. CI, Confidence Intervals.

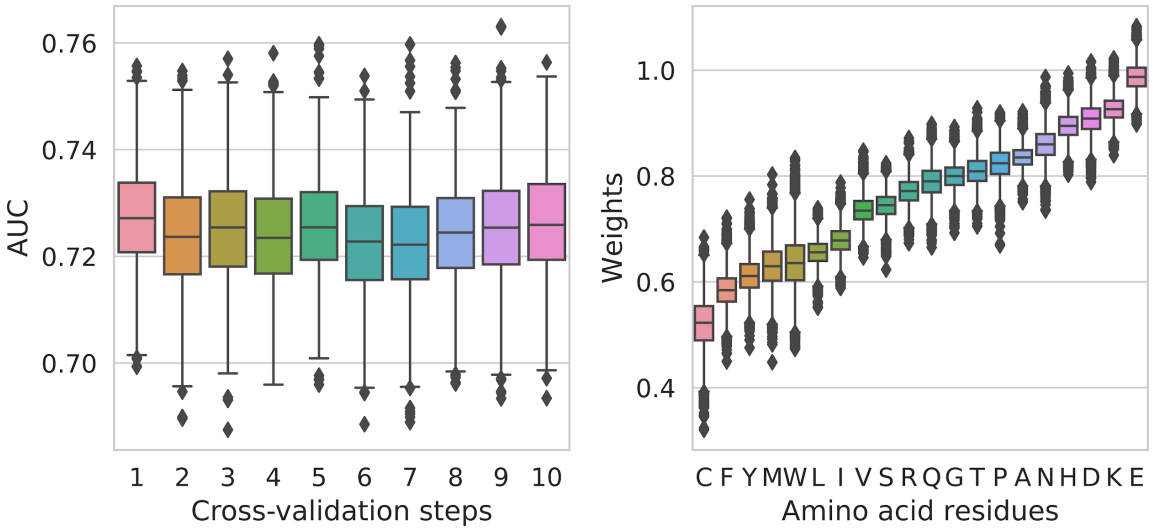


Figure B.5: **AUC scores and weights of amino acid residues obtained from individual bootstrap samples, related to Fig 3.2.** For each cross-validation step, 1,000 soluble and 1,000 insoluble proteins were resampled 1,000 times. For each bootstrap resampling, the weights of amino acid residues were optimised by maximising AUC using the Nelder-Mead algorithm. The optimised weights, i.e., the arithmetic means of the weights of individual amino acid residues in each cross-validation step, were used for sequence composition scoring. The training and test AUC scores were subsequently calculated (Fig 2B, 4A and Supplementary Table S3). AUC, Area Under the ROC Curve; ROC, Receiver Operating Characteristic.

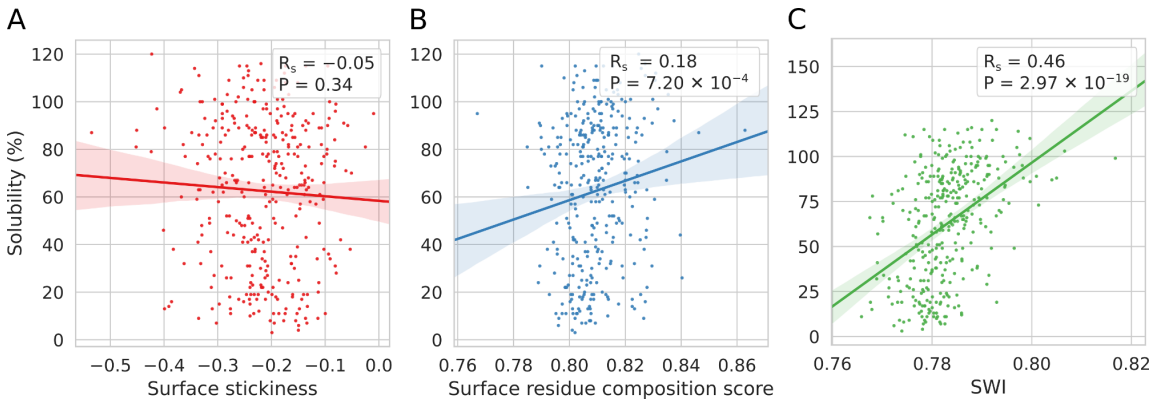


Figure B.6: **Relationship between protein solubility and surface amino acid residues.** The analyses were done using eSOL and the surface 'stickiness' of *E. coli* proteins (N = 348). **(A)** Protein solubility has a low correlation with surface 'stickiness'. **(B)** A low correlation was obtained after maximising the correlation between solubility and the surface residue composition scores using the Nelder-Mead algorithm. Smith et al.'s normalised B-factors were used as initial weights. **(C)** In contrast, protein solubility has a stronger correlation with SWI. Rs, Spearman's rho; SWI, Solubility-Weighted Index.

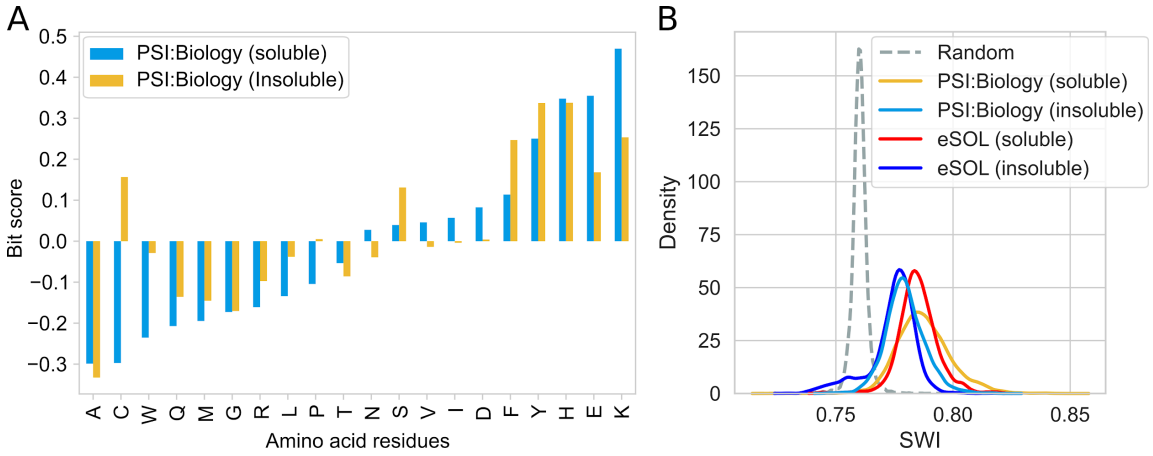


Figure B.7: **Properties of soluble and insoluble proteins.** **(A)**Enrichment of amino acid residues in the PSI:Biology targets relative to the eSOL sequences (N = 12,216 and 3,198, respectively). **(B)** Distribution of the SWI for soluble and insoluble proteins, and random sequences. The eSOL sequences were grouped into soluble and insoluble proteins, i.e, <30% and >70% solubility cutoffs, respectively (Supplementary Table S1B). Random sequences were generated from a length of 50 to 6,000 amino acid residues, with an increment of 50 residues. A total of 12,000 random sequences were generated, 100 sequences for each length. PSI:Biology, Protein Structure Initiative:Biology; SWI, Solubility-Weighted Index.

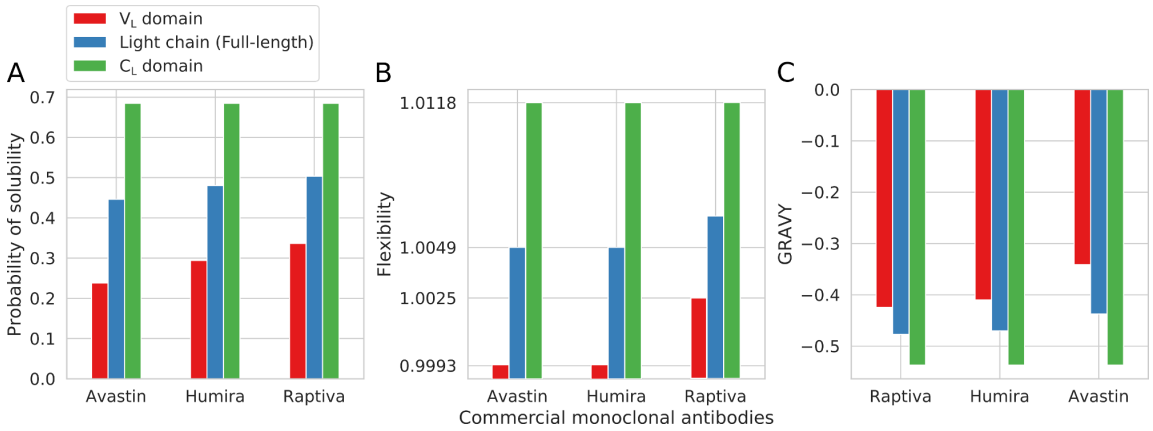


Figure B.8: **Solubility analysis of three commercial monoclonal antibodies.** The variable domains of immunoglobulin light chains (VL) have **(A)** lower probabilities of solubility, **(B)** lower structural flexibilities (log scale), and **(C)** higher GRAVY than the constant domains (CL). The sequences of Avastin (216974-75-3), Humira (331731-18-1), and Raptiva (214745-43-4) were retrieved from the Common Chemistry database. CAS registry numbers are shown in parentheses. GRAVY, Grand Average of Hydropathy.

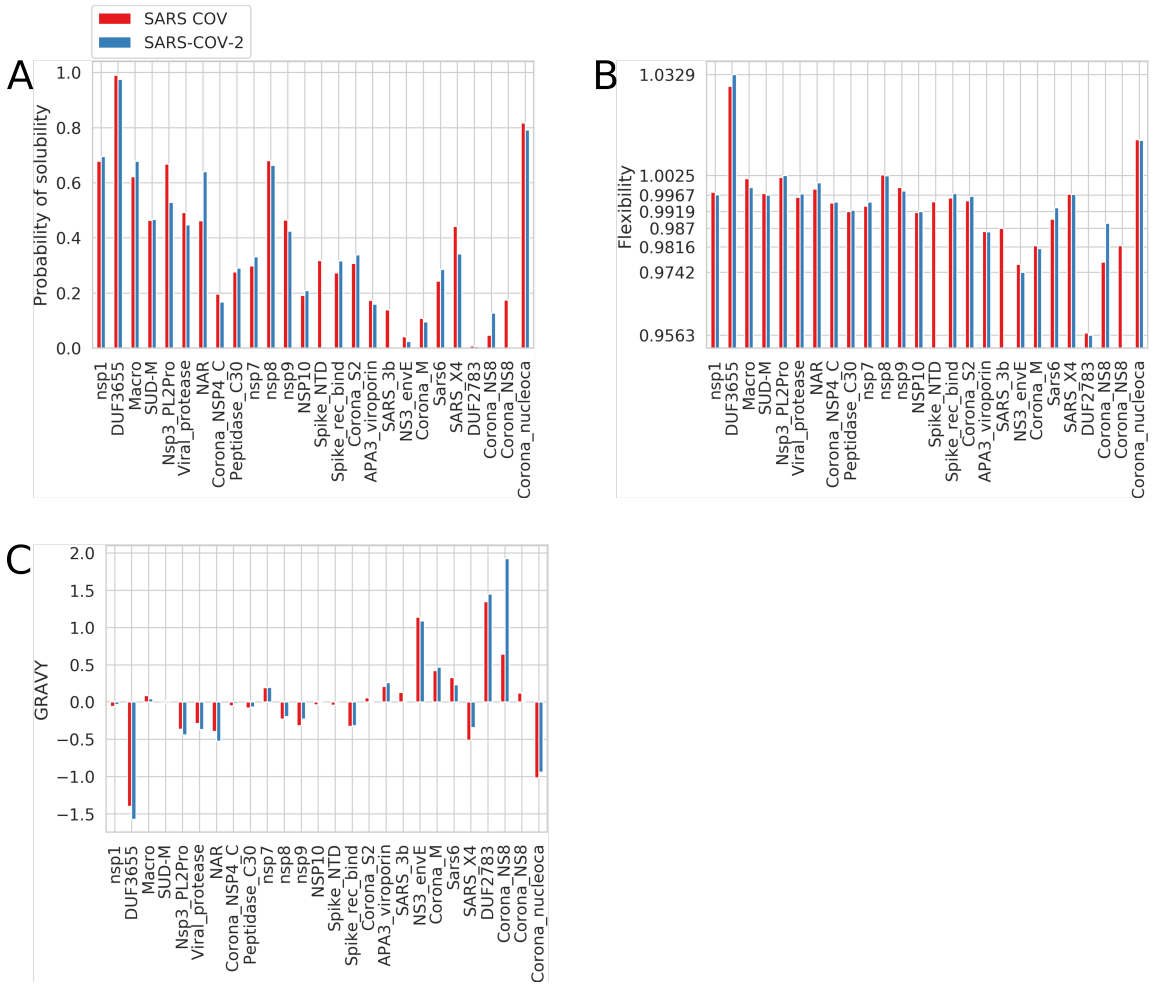


Figure B.9: **Solubility analysis of the SARS-CoV and SARS-CoV-2 proteomes.** The viral proteomes were retrieved from NCBI RefSeq on 23 March 2020 (NC_004718.3 and NC_045512.2). The polypeptides/domains were annotated by the HMMER web server using the Pfam database. No domains were annotated for ORF10. **(A)** The ORF2, 4, 5, and 8b proteins/domains have low probabilities of solubility, whereas the ORF9 protein have a high probability of solubility, which are consistent with previous protein expression studies (Wu *et al.*, 2004; Kam *et al.*, 2007; Neuman *et al.*, 2011; Shi *et al.*, 2019) [269, 122, 169, 216] . **(C)** The flexibility plot of each domain, shown in log scale. **(A)** GRAVY of each domain. GRAVY, Grand Average of Hydropathy; SARS-CoV, severe acute respiratory syndrome coronavirus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

B.3 Supplementary tables

Table B.1: Numbers of soluble and insoluble proteins examined in this study.

PSI:Biology dataset			
	pET21_NSEG	**pET15_NSEG**	**Total**
Soluble	6,342	1,896	8,238
Insoluble	2,438	1,540	3,978
Total	8,780	3,436	12,216
eSOL dataset			
Highly soluble (> 70% solubility)	1,029		
Partially soluble	905		
Aggregation prone (< 30% solubility)	1,264		
Total	3,198		

Table B.2: Analysis of miscellaneous protein sequence properties.

AUC scores for predicting the solubility of the PSI:Biology targets, related to Fig B.2		
Features	**AUC**	
Pc1.C	0.651	
Xc1.C	0.643	
polarity.Group1	0.638	
Pc1.H	0.637	
Grantham.Xr.E	0.635	
prop6.G2.residue100	0.629	
...	...	
Full table can be viewed at https://dx.doi.org/10.1093/bioinformatics/btaa578		
Spearman's correlation between 9,913 miscellaneous protein sequence features and E. coli protein solubility (eSOL dataset), related to Fig S2.		
Features	**Spearman's rho**	**P-value**
Xc1.W	-0.401	3.55E-124
Pc1.W	-0.397	2.47E-121
Xc1.F	-0.391	1.62E-117
Pc1.F	-0.389	5.77E-116
...	...	...
Full table can be viewed at https://dx.doi.org/10.1093/bioinformatics/btaa578		

Table B.3: Training and test AUC scores in a 10-fold cross-validation, related to Fig 3.2A and B.

Cross-validation step	AUC (train)	AUC (test)
1	0.721529	0.689413
2	0.718002	0.719417
3	0.71958	0.707645
4	0.717626	0.720168
5	0.719628	0.705219
6	0.716786	0.728537
7	0.715986	0.737957
8	0.718487	0.714374
9	0.719457	0.703184
10	0.719797	0.703105

Table B.4: Weights of amino acid residues for solubility scoring, related to Fig 3.2C.

Amino acid residues	Initial weights [normalised B-factors (Smith et al (2003)]	Final weights
A	0.717	0.835647
C	0.668	0.520809
E	0.963	0.987699
D	0.921	0.907904
G	0.843	0.799717
F	0.599	0.584979
I	0.632	0.678412
H	0.754	0.894791
K	0.912	0.92671
M	0.685	0.629662
L	0.681	0.655422
N	0.851	0.859743
Q	0.849	0.789435
P	0.85	0.823533
S	0.84	0.744091
R	0.814	0.771247
T	0.758	0.809692
W	0.626	0.637468
V	0.619	0.735784
Y	0.615	0.61128

Table B.5: Correlation test results, related to Fig 3.3A.

Spearman's correlation for the PSI:Biology dataset.											
	Sheet	**Isoelectric point**	**Turn**	**Instability index**	**Aromaticity**	**Helix**	**Molecular weight**	**GRAVY**	**Flexibility**	**SWI**	**Soluble or insoluble**
Sheet	1	-0.22206	-0.375104	0.220627	-0.328068	-0.097355	0.070636	0.232137	-0.041549	0.143317	0.040284
Isoelectric point	-0.22206	1	-0.010108	0.072233	0.009231	-0.082686	-0.124552	-0.171168	-0.083467	-0.158965	-0.065772
Turn	-0.375104	-0.010108	1	-0.07305	0.010417	-0.089198	0.210953	0.182721	0.024328	-0.211602	-0.080815
Instability index	0.220627	0.072233	-0.07305	1	-0.072715	-0.159392	0.04537	-0.143305	-0.07666	-0.172905	-0.090254
Aromaticity	-0.328068	0.009231	0.010417	-0.072715	1	0.476667	0.222074	-0.090969	-0.259083	-0.455687	-0.154726
Helix	-0.097355	-0.082686	-0.089198	-0.159392	0.476667	1	0.226637	0.470759	-0.409018	-0.460267	-0.154866
Molecular weight	0.070636	-0.124552	0.210953	0.04537	0.222074	0.226637	1	0.249705	-0.037168	-0.276888	-0.162451
GRAVY	0.232137	-0.171168	0.182721	-0.143305	-0.090969	0.470759	0.249705	1	-0.600141	-0.477966	-0.170855
Flexibility	-0.041549	-0.083467	0.024328	-0.07666	-0.259083	-0.409018	-0.037168	-0.600141	1	0.773422	0.280697
SWI	0.145739	-0.160047	-0.212443	-0.172425	-0.454848	-0.457766	-0.273635	-0.475792	0.774244	1	0.354619
Solubility	0.040284	-0.065772	-0.080815	-0.090254	-0.154726	-0.154866	-0.162451	-0.170855	0.280697	0.354597	1
Bonferroni corrected P-values for the correlation test.											
	Sheet	**Isoelectric point**	**Turn**	**Instability index**	**Aromaticity**	**Helix**	**Molecular weight**	**GRAVY**	**Flexibility**	**SWI**	**Solubile or Insoluble**
Sheet	0.00E+00	1.36E-134	0.00E+00	8.06E-133	1.16E-302	2.23E-25	3.00E-13	2.07E-147	2.39E-04	3.12E-57	4.64E-04
Isoelectric point	1.36E-134	0.00E+00	1.45E+01	7.23E-14	1.69E+01	3.02E-18	1.08E-41	3.14E-79	1.35E-18	3.72E-69	1.88E-11
Turn	0.00E+00	1.45E+01	0.00E+00	3.45E-14	1.37E+01	2.87E-21	3.47E-121	1.88E-90	3.94E-01	6.09E-123	2.03E-17
Instability index	8.06E-133	7.23E-14	3.45E-14	0.00E+00	4.68E-14	1.39E-68	2.89E-05	2.55E-55	1.19E-15	2.06E-80	8.84E-22
Aromaticity	1.16E-302	1.69E+01	1.37E+01	4.68E-14	0.00E+00	0.00E+00	1.31E-134	3.95E-22	7.94E-185	0.00E+00	1.38E-64
Helix	2.23E-25	3.02E-18	2.87E-21	1.39E-68	0.00E+00	0.00E+00	2.45E-140	0.00E+00	0.00E+00	0.00E+00	1.05E-64
Molecular weight	3.00E-13	1.08E-41	3.47E-121	2.89E-05	1.31E-134	2.45E-140	0.00E+00	2.80E-171	2.18E-03	5.55E-207	2.84E-71
GRAVY	2.07E-147	3.14E-79	1.88E-90	2.55E-55	3.95E-22	0.00E+00	2.80E-171	0.00E+00	0.00E+00	0.00E+00	6.17E-79
Flexibility	2.39E-04	1.35E-18	3.94E-01	1.19E-15	7.94E-185	0.00E+00	2.18E-03	0.00E+00	0.00E+00	0.00E+00	3.07E-218
SWI	3.12E-57	3.72E-69	6.09E-123	2.06E-80	0.00E+00	0.00E+00	5.55E-207	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Solubility	4.64E-04	1.88E-11	2.03E-17	8.84E-22	1.38E-64	1.05E-64	2.84E-71	6.17E-79	3.07E-218	0.00E+00	0.00E+00

Table B.6: Correlation test results, related to Fig 3.3B.

Spearman's correlation for the eSOL dataset.											
	Sheet	**Instability index**	**Turn**	**Isoelectric point**	**GRAVY**	**Aromaticity**	**Helix**	**Molecular weight**	**Flexibility**	**SWI**	**Solubility(%)**
Sheet	1	0.146456	-0.403058	-0.156209	0.309439	-0.326514	0.02755	0.059391	-0.112851	0.083016	0.027396
Instability index	0.146456	1	-0.134142	0.015775	-0.16924	0.015088	0.0546	0.069486	-0.057118	-0.158018	-0.099881
Turn	-0.403058	-0.134142	1	-0.018457	0.096262	0.047374	-0.103803	0.158678	0.126597	-0.134466	-0.122414
Isoelectric point	-0.156209	0.015775	-0.018457	1	-0.013594	0.005004	0.028062	-0.353713	-0.207243	-0.264909	-0.192862
GRAVY	0.309439	-0.16924	0.096262	-0.013594	1	-0.107626	0.487898	0.118364	-0.612059	-0.453011	-0.208084
Aromaticity	-0.326514	0.015088	0.047374	0.005004	-0.107626	1	0.468139	0.179587	-0.27243	-0.464074	-0.328931
Helix	0.02755	0.0546	-0.103803	0.028062	0.487898	0.468139	1	0.260863	-0.557496	-0.602545	-0.364232
Molecular weight	0.059391	0.069486	0.158678	-0.353713	0.118364	0.179587	0.260863	1	0.117507	-0.040323	-0.357656
Flexibility	-0.112851	-0.057118	0.126597	-0.207243	-0.612059	-0.27243	-0.557496	0.117507	1	0.780143	0.372535
SWI	0.083016	-0.158018	-0.134466	-0.264909	-0.453011	-0.464074	-0.602545	-0.040323	0.780143	1	0.503647
Solubility(%)	0.027396	-0.099881	-0.122414	-0.192862	-0.208084	-0.328931	-0.364232	-0.357656	0.372535	0.503647	1
Bonferroni corrected P-values for the correlation test.											
	Sheet	**Instability index**	**Turn**	**Isoelectric point**	**GRAVY**	**Aromaticity**	**Helix**	**Molecular weight**	**Flexibility**	**SWI**	**Solubility(%)**
Sheet	0.00E+00	4.68E-15	1.78E-123	3.52E-17	3.51E-70	1.37E-78	6.56E+00	4.28E-02	8.56E-09	1.42E-04	6.68E+00
Instability index	4.68E-15	0.00E+00	1.42E-12	2.05E+01	3.08E-20	2.17E+01	1.11E-01	4.62E-03	6.77E-02	1.37E-17	8.31E-07
Turn	1.78E-123	1.42E-12	0.00E+00	1.63E+01	2.71E-06	4.06E-01	2.21E-07	9.69E-18	3.69E-11	1.23E-12	2.07E-10
Isoelectric point	3.52E-17	2.05E+01	1.63E+01	0.00E+00	2.43E+01	4.27E+01	6.19E+00	3.80E-93	1.27E-30	9.31E-51	1.97E-26
GRAVY	3.51E-70	3.08E-20	2.71E-06	2.43E+01	0.00E+00	5.77E-08	3.35E-189	1.04E-09	0.00E+00	6.75E-160	7.07E-31
Aromaticity	1.37E-78	2.17E+01	4.06E-01	4.27E+01	5.77E-08	0.00E+00	3.46E-172	7.59E-23	8.59E-54	7.98E-169	7.99E-80
Helix	6.56E+00	1.11E-01	2.21E-07	6.19E+00	3.35E-189	3.46E-172	0.00E+00	3.64E-49	6.57E-259	0.00E+00	3.55E-99
Molecular weight	4.28E-02	4.62E-03	9.69E-18	3.80E-93	1.04E-09	7.59E-23	3.64E-49	0.00E+00	1.45E-09	1.24E+00	2.22E-95
Flexibility	8.56E-09	6.77E-02	3.69E-11	1.27E-30	0.00E+00	8.59E-54	6.57E-259	1.45E-09	0.00E+00	0.00E+00	4.25E-104
SWI	1.42E-04	1.37E-17	1.23E-12	9.31E-51	6.75E-160	7.98E-169	0.00E+00	1.24E+00	0.00E+00	0.00E+00	1.38E-203
Solubility(%)	6.68E+00	8.31E-07	2.07E-10	1.97E-26	7.07E-31	7.99E-80	3.55E-99	2.22E-95	4.25E-104	1.38E-203	0.00E+00

Table B.7: Runtime of protein solubility prediction tools per sequence, related to Fig 3.4B

Accession	**Length**	**Run**	**Wall time (hh:mm:ss)**	**Wall time (s)**	**Tool**
JW0031	1095	0	00:25:51	1551.001	DeepSol1
JW0031	1095	1	00:22:28	1348.001	DeepSol1
JW0031	1095	2	00:22:28	1348.001	DeepSol1
JW0560	249	0	00:14:22	862.001	DeepSol1
JW0560	249	1	00:15:23	923.001	DeepSol1
...	...	...	...	...	...
Full table can be viewed at https://dx.doi.org/10.1093/bioinformatics/btaa578					

Table B.8: Probability of solubility at selected SWI thresholds, related to Equation 3.5

SWI threshold	**True positive rate**	**False positive rate**	**Probability of solubility**
1.85	0	0	1
0.8	0.11	0.01	0.9
0.79	0.33	0.08	0.8
0.78	0.57	0.24	0.7
0.78	0.77	0.48	0.6
0.77	0.9	0.75	0.5
0.77	0.96	0.91	0.4
0.76	0.99	0.97	0.3
0.76	1	0.99	0.2
0.75	1	1	0.09
0.72	1	1	0.01

Appendix C

Razor: annotation of signal peptides from toxins

C.1 Supplementary figures

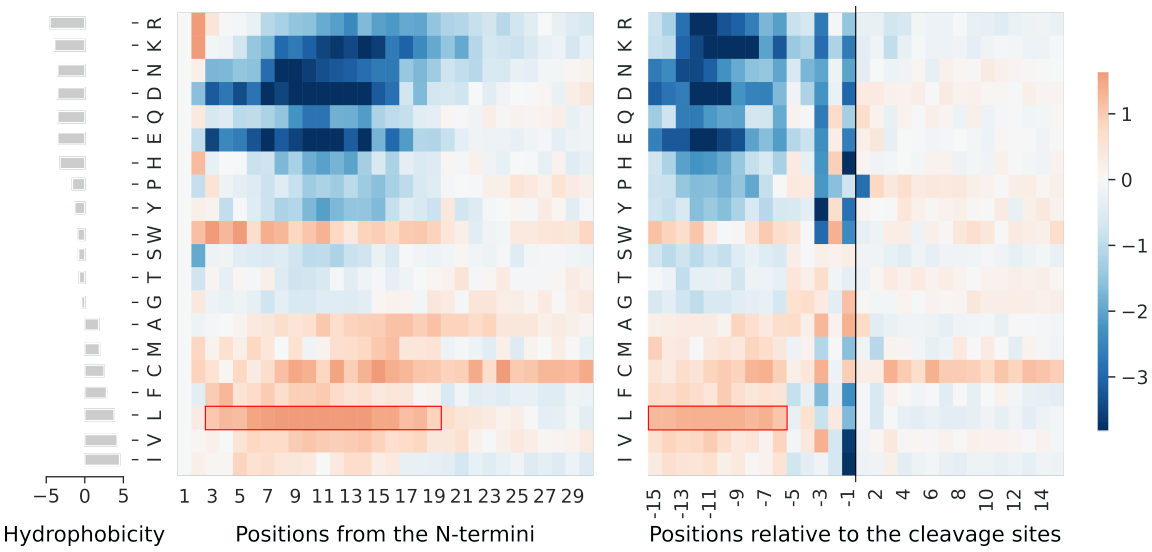


Figure C.1: **Signal peptides (SPs) show a strong hydrophobic property (1,964 experimentally validated SPs, 13,237 non-SPs).** The bar plot shows Kyte and Doolittle's hydrophobicity scale. The heatmaps show the enrichment of residues in bit scores by aligning SPs from the N-termini (left) and at the cleavage sites (right, black vertical line). The (-3, -1) rule for the cleavage site motif is shown (left). The unfilled, red rectangles indicate the enrichment of leucine residues (L).

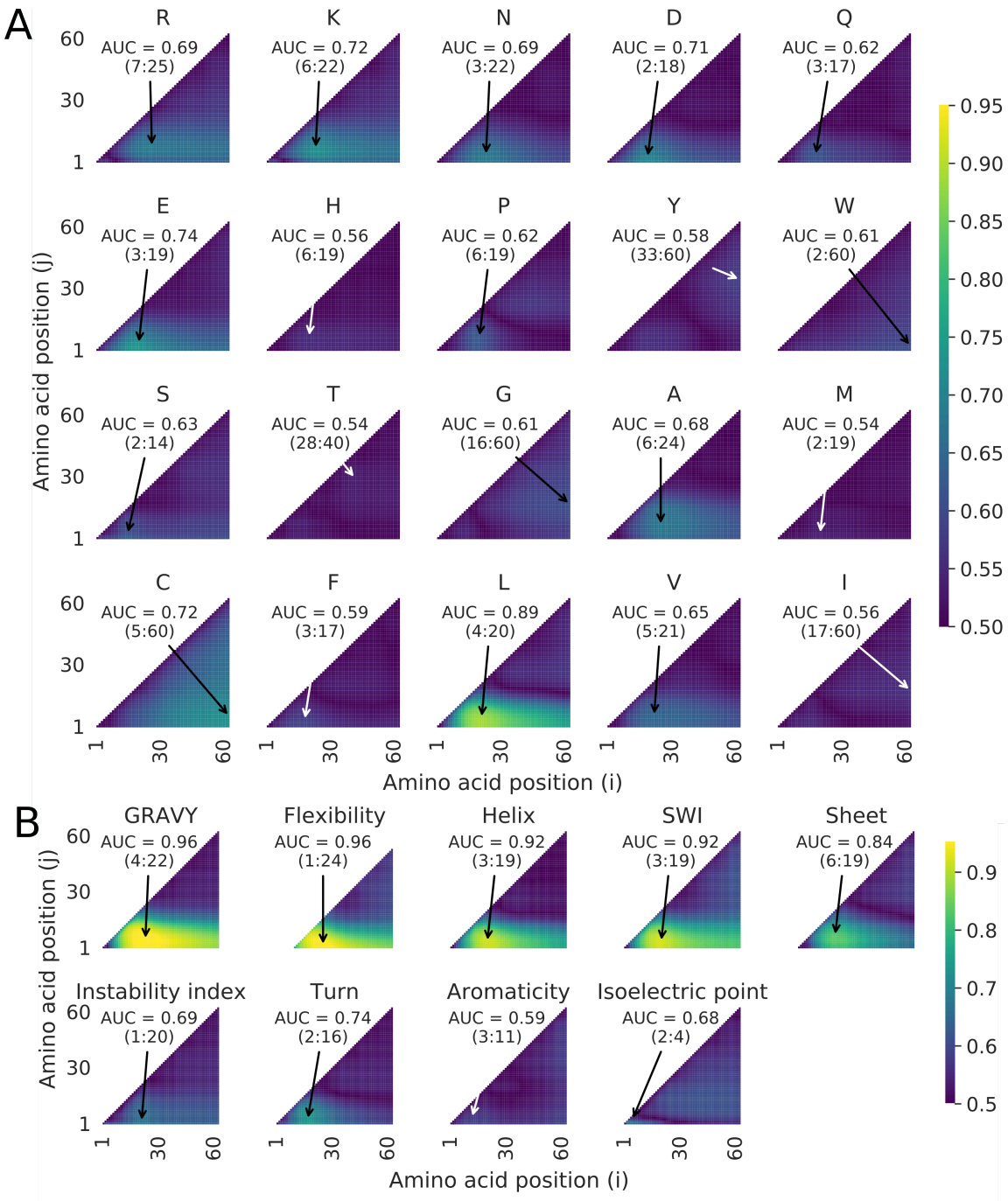


Figure C.2: **Leucine (L) composition within the N-terminal region 4:20 shows the highest AUC score in classifying the presence or absence of eukaryotic signal peptides (1,964 and 13,237, respectively).** Amino acid compositions were calculated from positions i to j. **(A)** AUC heatmaps for all residues. **(B)** GRAVY, Flexibility, Helix and SWI are the top four features ranked by AUC scores. AUC, Area Under the Curve; GRAVY, GRAnd average of hydropathicitY; SWI, Solubility-Weighted Index.

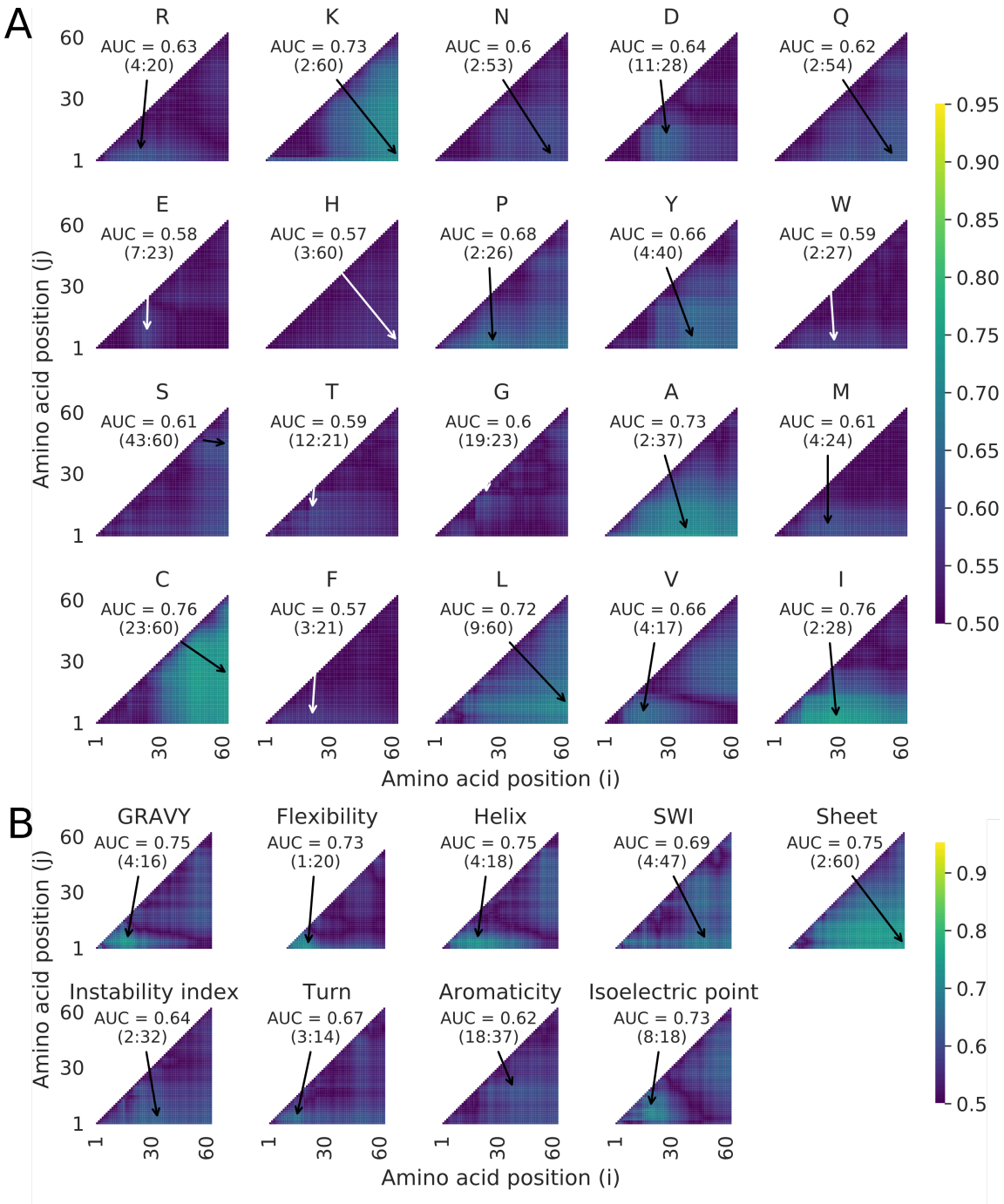


Figure C.3: **Isoleucine (I) composition within the N-terminal region 2:28 shows the highest AUC score in classifying toxin and non-toxin SPs (261 and 1,738, respectively).** Amino acid compositions were calculated from positions i to j. **(A)** AUC heatmaps for all residues. Cysteine shows a higher AUC score at the mature region (23:60) as many toxins are cysteine-rich. **(B)** GRAVY, Flexibility, Helix, SWI, Isoelectric point are the top features ranked by AUC scores. Although Sheet has a high AUC score, the region 2:60 extends beyond the normal SP length of around 30 residues. AUC, Area Under the Curve; GRAVY, GRAnd average of hydropathicitY; SWI, Solubility-Weighted Index.

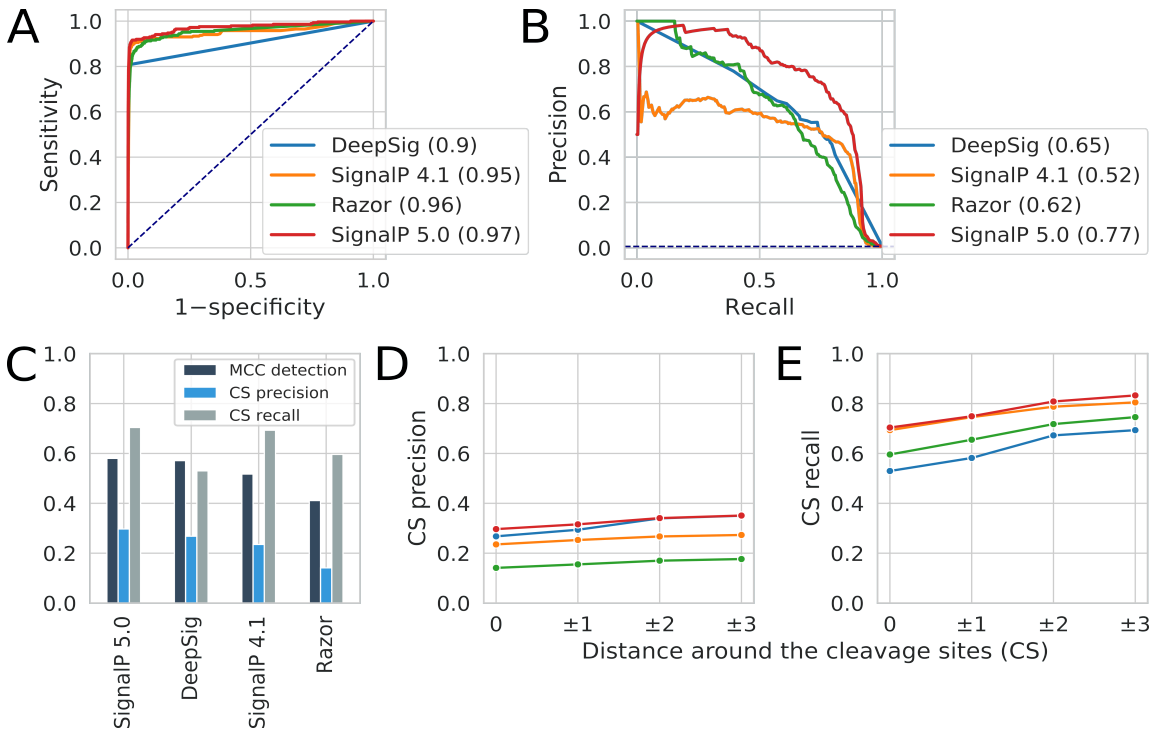


Figure C.4: **Performance of Razor and state-of-the-art in predicting eukaryotic SPs using an independent test set (SPs=241, non-SPs=52,055).** Receiver operating characteristic curves **(A)** and precision recall curves **(B)** of the SP prediction tools. Areas under the curves are shown in parentheses. The dotted lines show the performance of a random classifier. **(C)** Matthew's Correlation Coefficients (MCC) of the SP prediction tools. The cleavage site (CS) precisions **(D)** and recalls **(E)** of windows surrounding the cleavage sites are shown. Data are available in Supplementary Tables S3 and S4.

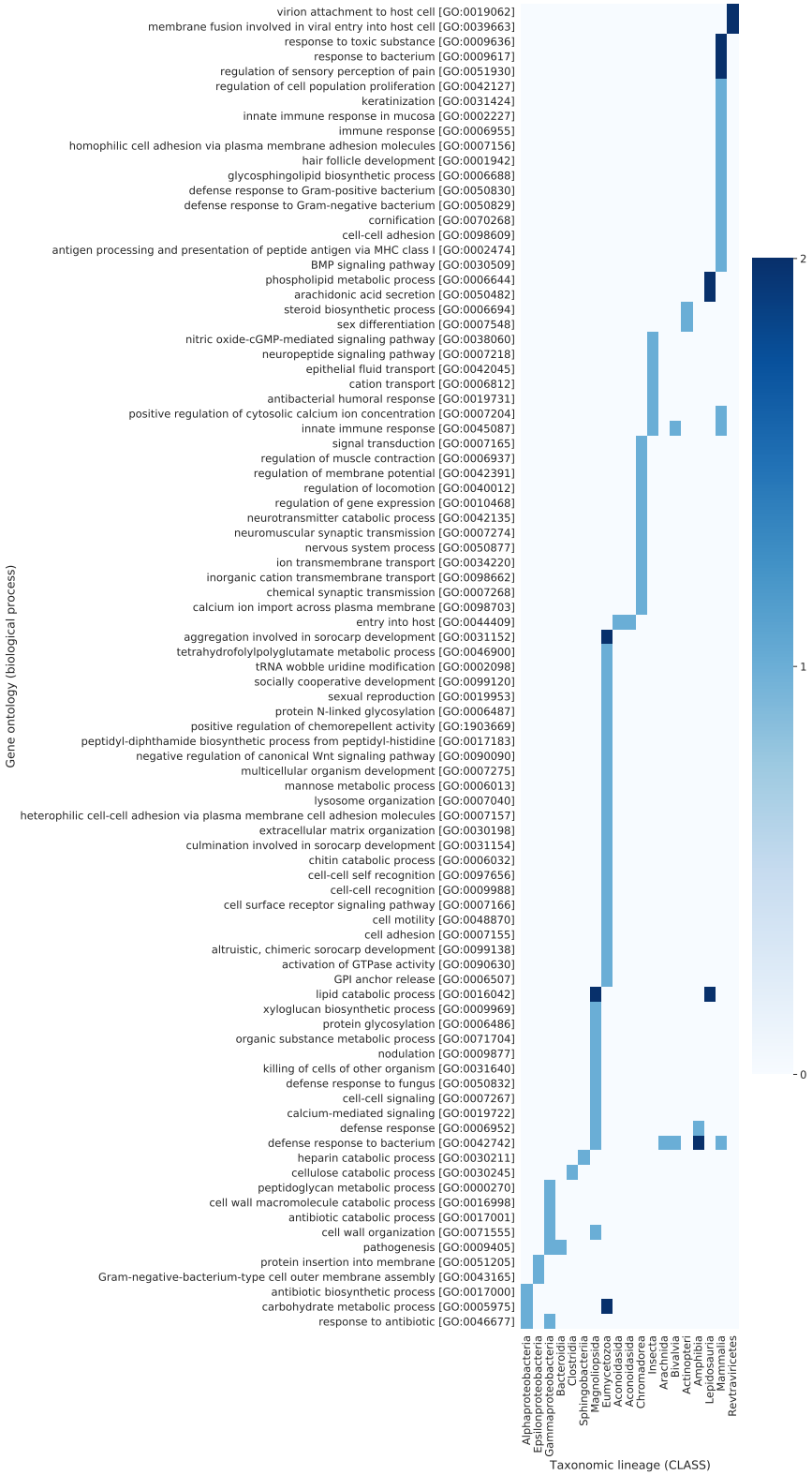


Figure C.5: **Gene ontology (GO) annotations (biological process) for the predicted toxin SPs.** A total of 54 out of 100 predicted sequences had GO terms. The scale bar indicates the frequencies of GO terms for the predicted sequences.

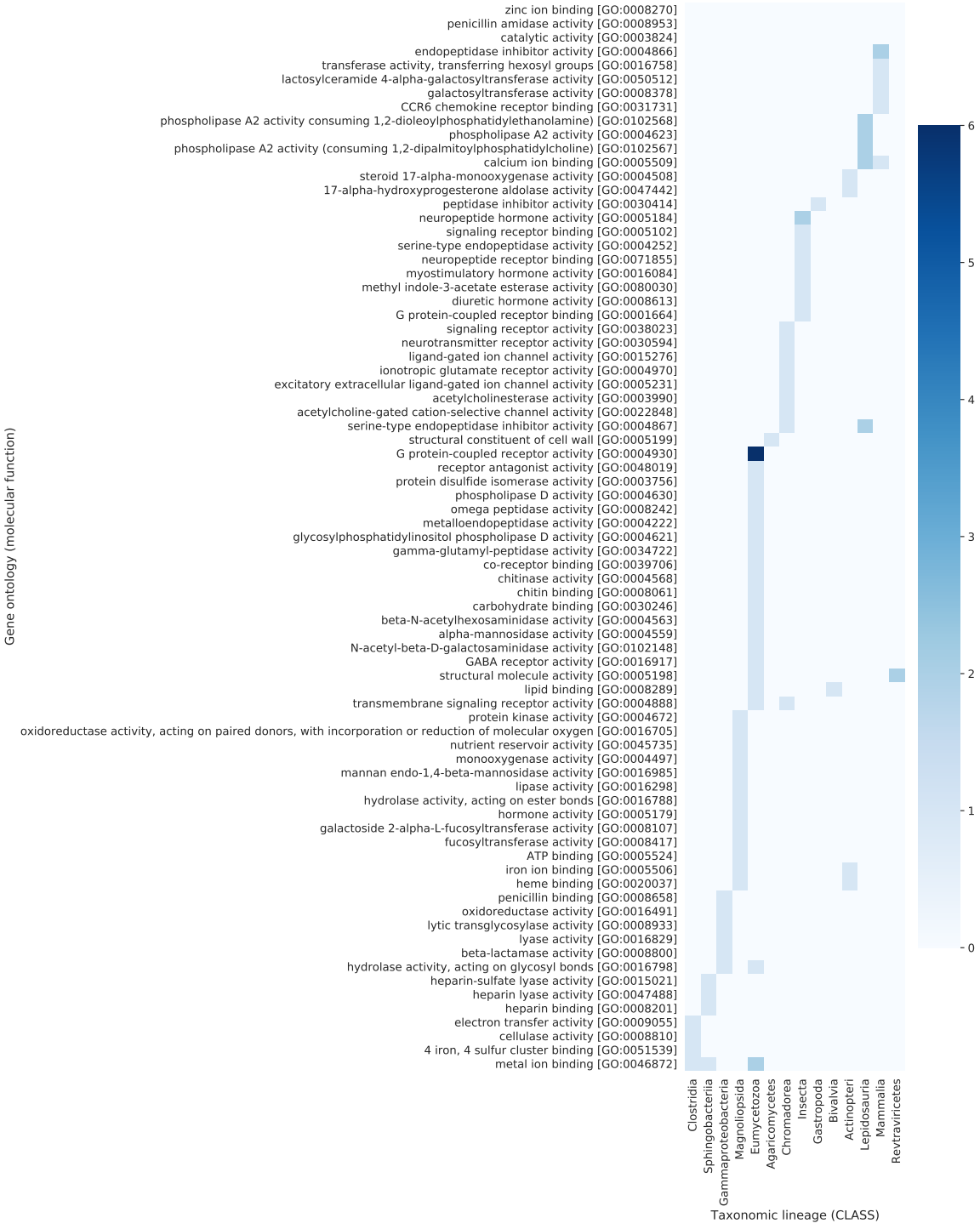


Figure C.6: **GO annotations (molecular function) for the predicted toxin SPs.** A total of 59 out of 100 predicted sequences had GO terms. The scale bar indicates the frequencies of GO terms for the predicted sequences.

C.2 Supplementary tables

Table C.1: Datasets used in this study.

	SPs	**Non-SPs**
Eukaryotic[a]	1,694	13,237
Toxin[a]	261	
Independent test set	214 (47 Toxin SPs, 194 Non-toxin SPs)	52,249

[a]Sequences were retrieved from the SignalP 5.0 training set and the animal toxin annotation program of UniProt and were clustered at 70% identity using CD-HIT.

Table C.2: Feature selection for building the toxin classifier using five-fold cross-validations. Boldface denotes the maximum MCC score.

N-terminal lengths	**MCC scores**	**Features**
15	0.727	Hydrophobicity, SWI, Flexibility, Helix, Turn
16	0.731	Hydrophobicity, SWI, Flexibility, Turn, Isoelectric point
17	0.716	Hydrophobicity, SWI, Helix, Turn
18	0.726	Hydrophobicity, SWI, Isoelectric Point
19	0.727	Hydrophobicity, SWI, Turn, Isoelectric Point
20	0.727	SWI, Flexibility, Helix, Turn
21	0.718	Hydrophobicity, SWI, Flexibility
22	0.717	Hydrophobicity, SWI, Isoelectric Point
23	**0.741**	**Hydrophobicity, SWI, Flexibility, Turn**
24	0.715	Hydrophobicity, SWI
25	0.718	Hydrophobicity, Flexibility, Turn
26	0.701	Hydrophobicity, SWI, Helix, Isoelectric Point
27	0.716	Hydrophobicity, SWI, Flexibility, Isoelectric Point
28	0.712	Hydrophobicity, SWI, Turn

Table C.3: Benchmarking of eukaryotic SP prediction using an independent test set (toxin SPs=287, Non-SPs=52,055). Boldface denotes the maximum MCC score.

	SignalP 5.0	**DeepSig**	**SignalP 4.1**	**Razor**
MCC	**0.571**	0.537	0.511	0.405

Table C.4: Benchmarking of the cleavage site prediction for eukaryotic SPs using an independent test set (SPs=287, Non-SPs=52,055). Boldface denotes the highest scores.

Tools	**Distance around the cleavage sites**			
	0	±1	±2	±3
Cleavage site precision				
SignalP 5.0	**0.287**	**0.306**	**0.33**	**0.34**
SignalP 4.1	0.229	0.247	0.26	0.266
Razor	0.136	0.15	0.164	0.171
DeepSig	0.237	0.261	0.301	0.31
Cleavage site recall				
SignalP 5.0	**0.704**	**0.749**	**0.808**	**0.833**
SignalP 4.1	0.693	0.746	0.787	0.805
DeepSig	0.53	0.582	0.672	0.693
Razor	0.596	0.655	0.718	0.746

Table C.5: Benchmarking of toxin SP prediction using an independent test set (toxin SPs=47, Non-toxin SPs=52,055). Boldface denotes the maximum MCC score.

	SignalP 5.0	**DeepSig**	**SignalP 4.1**	**Razor**
MCC	0.301	0.300	0.260	**0.611**

Table C.6: Benchmarking of the cleavage site prediction for toxin SPs using an independent test set (toxin SPs=47, Non-toxin SPs=52,055). Boldface denotes the highest scores.

Tools	**Distance around the cleavage sites**			
	0	±1	±2	±3
Cleavage site precision				
SignalP 5.0	0.094	N/A	N/A	0.34
SignalP 4.1	0.065	0.068	0.07	0.266
Razor	**0.355**	**0.373**	**0.382**	**0.171**
DeepSig	0.073	0.077	0.097	0.31
Cleavage site recall				
SignalP 5.0	**0.979**	**N/A**	**N/A**	**0.833**
SignalP 4.1	0.915	0.957	0.979	0.805
DeepSig	0.702	0.745	0.936	0.693
Razor	0.830	0.872	0.894	0.746

Appendix D

TISIGNER.com: interactive web services for improving recombinant protein production

D.1 Supplementary tables

Table D.1: Performance metrics of TIsigner and SoDoPE.

Tool (approach)	**Dataset and host (PubMed ID)**	**Type**	**Sample size**	**Spearman's correlation**	**AUROC**
TIsigner (mRNA accessibility)	GFP report, *E. coli* (PMID: 30247489)	Continuous	14,425	-0.65 ($P< 2.2 \times 10^{-16}$)	N/A
	YFP report, *S. cerevisiae* (PMID: 31209029)	Continuous	2,041	-0.55 ($P< 2.2 \times 10^{-16}$)	N/A
	GFP report, *M. musculus* (PMID: 25170020)	Continuous	65,536	-0.28 ($P< 2.2 \times 10^{-16}$)	N/A
	PSI:Biology, *E. coli* (PMID: 24225319)	Binary	8,780 (expressed), 2,650 (not expressed)	N/A	0.70
SoDoPE (SWI)	eSOL, *E. coli* (PMID: 19251648)	Continuous	3,198	0.50 ($P=9.46 \times 10^{-206}$)	N/A
	*PSI:Biology, *E. coli* (PMID: 24225319)	Binary	8,238 (soluble), 3,978 (insoluble)	N/A	0.71

*The metrics are derived from five-fold cross-validations for training purposes. AUROC, Area under the ROC curve; MCC, Matthew's correlation coefficient; SWI: Solubility-Weighted Index.

Table D.2: Performance metrics of Razor.

Classifier	**Dataset (references)**	**Sample size**	**MCC**	**AUROC**	**AUPRC**	**Cleavage site**	
						Precision	**Recall**
Eukaryotic SP	SignalP 5.0 benchmarking set, (PMID: 30778233)	211 SPs, 7,246 non-SPs	0.815	0.98	0.85	0.565	0.597
	Independent test set	287 SPs, 52,055 non-SPs	0.405	0.96	0.61	0.136	0.596
Toxin SP	Training set, (PMID: 22465017, 30395287)	261 toxin SPs, 1,738 non-toxin SPs	0.741	0.89	0.74	N/A	N/A
	Independent test set	47 toxin SPs, 194 non-toxin SPs	0.769	0.98	0.93	N/A	N/A
Fungal SP	Training set, (PMID: 30395287)	121 fungal SPs, 1,843 non-fungal SPs	0.506	0.87	0.48	N/A	N/A
	Independent test set	18 fungal SPs, 269 non-fungal SPs	0.6	0.94	0.75	N/A	N/A

References

[1] D. J. Abraham and A. J. Leo, 'Extension of the fragment method to calculate amino acid zwitterion and side chain partition coefficients', *Proteins*, 1987 (cit. on p. 15).

[2] T. B. Acton, K. C. Gunsalus, R. Xiao, L. C. Ma, J. Aramini, M. C. Baran, Y.-W. Chiang, T. Climent, B. Cooper, N. G. Denissova, S. M. Douglas, J. K. Everett, C. K. Ho, D. Macapagal, P. K. Rajan, R. Shastry, L.-Y. Shih, G. V. T. Swapna, M. Wilson, M. Wu, M. Gerstein, M. Inouye, J. F. Hunt and G. T. Montelione, 'Robotic cloning and protein production platform of the Northeast Structural Genomics Consortium', *Methods Enzymol.*, 2005 (cit. on pp. 34, 54, 65, 109).

[3] F. Agostini, D. Cirillo, C. M. Livi, R. Delli Ponti and G. G. Tartaglia, 'ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*', *Bioinformatics*, 2014 (cit. on pp. 61, 62, 83).

[4] S. Ali, B. A. Ganai, A. N. Kamili, A. A. Bhat, Z. A. Mir, J. A. Bhat, A. Tyagi, S. T. Islam, M. Mushtaq, P. Yadav, S. Rawat and A. Grover, 'Pathogenesis-related proteins and peptides as promising tools for engineering plants with multiple stress tolerance', *Microbiol. Res.*, 2018 (cit. on p. 76).

[5] J. J. Almagro Armenteros, K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne and H. Nielsen, 'SignalP 5.0 improves signal peptide predictions using deep neural networks', *Nat. Biotechnol.*, 2019 (cit. on pp. 76, 79, 83, 89).

[6] D. Alvarez-Garcia and X. Barril, 'Relationship between protein flexibility and binding: Lessons for structure-based drug design', *J. Chem. Theory Comput.*, 2014 (cit. on p. 17).

[7] M. Amaral, D. Kokh, J. Bomke, A. Wegener, H. Buchstaller, H. Eggenweiler, P. Matias, C. Sirrenberg, R. Wade and M. Frech, 'Protein conformational flexibility modulates kinetics and thermodynamics of drug binding', *Nat. Commun.*, 2017 (cit. on p. 19).

[8] K. S. Ang, S. Kyriakopoulos, W. Li and D.-Y. Lee, 'Multi-omics data driven analysis establishes reference codon biases for synthetic gene design in microbial and mammalian cells', *Methods*, 2016 (cit. on pp. 35, 46).

[9] F. Aslund and J. Beckwith, 'The thioredoxin superfamily: redundancy, specificity, and gray-area genomics', *J. Bacteriol.*, 1999 (cit. on p. 63).

[10] P. G. Bagos, K. D. Tsirigos, S. K. Plessas, T. D. Liakopoulos and S. J. Hamodrakas, 'Prediction of signal peptides in archaea', *Protein Eng. Des. Sel.*, 2009 (cit. on p. 83).

[11] S. F. Banani, H. O. Lee, A. A. Hyman and M. K. Rosen, 'Biomolecular condensates: organizers of cellular biochemistry', *Nat. Rev. Mol. Cell Biol.*, 2017 (cit. on p. 11).

[12] T. Ben-Yehezkel, S. Atar, H. Zur, A. Diament, E. Goz, T. Marx, R. Cohen, A. Dana, A. Feldman, E. Shapiro and T. Tuller, 'Rationally designed, heterologous *S. cerevisiae* transcripts expose novel expression determinants', *RNA Biol.*, 2015 (cit. on p. 39).

[13] A. Berlec and B. Strukelj, 'Current state and recent advances in biopharmaceutical production in *Escherichia coli*, yeasts and mammalian cells', *J. Ind. Microbiol. Biotechnol.*, 2013 (cit. on pp. 1, 30, 82).

[14] S. H. Bernhart, I. L. Hofacker and P. F. Stadler, 'Local RNA base pairing probabilities in large sequences', *Bioinformatics*, 2006 (cit. on pp. 82, 85).

[15] S. H. Bernhart, U. Mückstein and I. L. Hofacker, 'RNA Accessibility in cubic time', *Algorithms Mol. Biol.*, 2011 (cit. on pp. 12, 43, 45, 85).

[16] S. Bernhart, I. L. Hofacker and P. F. Stadler, 'Local Base Pairing Probabilities in Large RNAs', *Bioinformatics*, (cit. on p. 45).

[17] J. A. Bernstein, A. B. Khodursky, P.-H. Lin, S. Lin-Chao and S. N. Cohen, 'Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays', *Proc. Natl. Acad. Sci. U. S. A.*, 2002 (cit. on pp. 3, 30, 82).

[18] B. K. Bhandari, P. P. Gardner and C. S. Lim, 'Razor: annotation of signal peptides from toxins', *bioRxiv*, 2020 (cit. on pp. 69, 83, 89).

[19] ——, 'Solubility-Weighted Index: fast and accurate prediction of protein solubility', *Bioinformatics*, 2020 (cit. on pp. iv, 52, 77, 79, 82, 83, 87).

[20] B. K. Bhandari, C. S. Lim and P. P. Gardner, 'TISIGNER.com: web services for improving recombinant protein production', *Nucleic Acids Res.*, 2021, ISSN: 0305-1048 (cit. on p. 81).

[21] B. K. Bhandari, C. S. Lim, D. M. Remus, A. Chen, C. J. van Dolleweerd and P. P. Gardner, 'Protein yield is tunable by synonymous codon changes of translation initiation sites', *bioRxiv*, 2021 (cit. on pp. 30, 43, 65, 68, 79, 82, 83, 85).

[22] R. Bhaskaran and P. K. Ponnuswamy, 'Positional flexibilities of amino acid residues in globular proteins', *Int. J. Pept. Protein Res.*, 1988 (cit. on p. 56).

[23] S. Bhattacharyya, W. M. Jacobs, B. V. Adkar, J. Yan, W. Zhang and E. I. Shakhnovich, 'Accessibility of the Shine-Dalgarno Sequence Dictates N-Terminal Codon Bias in *E. coli*', *Mol. Cell*, 2018 (cit. on p. 43).

[24] L. F. Bidondo, L. Fernandez Bidondo, N. Almasia, A. Bazzini, R. Colombo, E. Hopp, C. Vazquez-Rovere and A. Godeas, 'The overexpression of antifungal genes enhances resistance to *Rhizoctonia solani* in transgenic potato plants without affecting arbuscular mycorrhizal symbiosis', *J. Crop Prot.*, 2019 (cit. on p. 70).

[25] D. S. Bindels, L. Haarbosch, L. van Weeren, M. Postma, K. E. Wiese *et al.*, 'mScarlet: a bright monomeric red fluorescent protein for cellular imaging', *Nat. Methods*, 2017 (cit. on p. 40).

[26] N. A. Boccardo, M. E. Segretin, I. Hernandez, F. G. Mirkin, O. Chacón, Y. Lopez, O. Borrás-Hidalgo and F. F. Bravo-Almonacid, 'Expression of pathogenesis-related proteins in transplastomic tobacco plants confers resistance to filamentous pathogens under field trials', *Sci. Rep.*, 2019 (cit. on p. 76).

[27] G. Boël, R. Letso, H. Neely, W. N. Price, K.-H. Wong, M. Su, J. D. Luff, M. Valecha, J. K. Everett, T. B. Acton *et al.*, 'Codon influence on protein expression in *E. coli* correlates with mRNA levels', *Nature*, 2016 (cit. on pp. 7, 31).

[28] A. F. Bompfünewerer, R. Backofen, S. H. Bernhart, J. Hertel, I. L. Hofacker, P. F. Stadler and S. Will, 'Variations on RNA folding and alignment: lessons from Benasque', *J. Math. Biol.*, 2008 (cit. on p. 45).

[29] F. Borrego, M. Ulbrecht, E. H. Weiss, J. E. Coligan and A. G. Brooks, 'Recognition of Human Histocompatibility Leukocyte Antigen (HLA)-E Complexed with HLA Class I Signal Sequence–derived Peptides by CD94/NKG2 Confers Protection from Natural Killer Cell–mediated Lysis', *J. Exp. Med.*, 1998 (cit. on p. 70).

[30] D. Bramer and G.-W. Wei, 'Blind prediction of protein B-factor and flexibility', *J. Chem. Phys.*, 2018 (cit. on pp. 17, 129).

[31] P. Braun and J. LaBaer, 'High throughput protein production for functional proteomics', *Trends Biotechnol.*, 2003 (cit. on p. 1).

[32] L. Breiman, 'Random forests', *Mach. Learn.*, 2001 (cit. on p. 27).

[33] J. Brownlee, *Clever Algorithms: Nature-inspired Programming Recipes*. Jason Brownlee, 2011 (cit. on pp. 21, 47).

[34] C. E. Brule and E. J. Grayhack, 'Synonymous Codons: Choose Wisely for Expression', *Trends Genet.*, 2017 (cit. on pp. 7, 31, 82).

[35] G. Cambray, J. C. Guimaraes and A. P. Arkin, 'Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*', *Nat. Biotechnol.*, 2018 (cit. on pp. 7, 31, 32, 43, 82, 98, 108).

[36] O. Carugo, 'How large B-factors can be in protein crystal structures', *BMC Bioinform.*, 2018 (cit. on pp. 17, 129).

[37] N. R. Casewell, W. Wüster, F. J. Vonk, R. A. Harrison and B. G. Fry, 'Complex cocktails: the evolutionary novelty of venoms', *Trends Ecol. Evol.*, 2013 (cit. on p. 70).

[38] N. Casewell, 'Solenodon genome reveals convergent evolution of venom in eulipotyphlan mammals (15 min)', *Toxicon*, 2020 (cit. on p. 70).

[39] T. A. Caswell, M. Droettboom, J. Hunter, E. Firing, A. Lee, D. Stansby, E. S. de Andrade, J. H. Nielsen, J. Klymak, N. Varoquaux, B. Root, P. Elson, D. Dale, R. May, J.-J. Lee, J. K. Seppänen, T. Hoffmann, D. McDougall, A. Straw, P. Hobson, cgohlke, T. S. Yu, E. Ma, A. F. Vincent, S. Silvester, C. Moad, J. Katins, N. Kniazev, F. Ariza and P. Würtz, *matplotlib/matplotlib v3.0.2*, 2018 (cit. on p. 68).

[40] C. Catalanotto, C. Cogoni and G. Zardo, 'MicroRNA in control of gene expression: an overview of nuclear functions', *Int. J. Mol. Sci.*, 2016 (cit. on p. 14).

[41] W.-C. Chan, P.-H. Liang, Y.-P. Shih, U.-C. Yang, W.-C. Lin and C.-N. Hsu, 'Learning to predict expression efficacy of vectors in recombinant protein production', *BMC Bioinform.*, 2010 (cit. on pp. 53, 82).

[42] M. Charton and B. I. Charton, 'The structural dependence of amino acid hydrophobicity parameters', *J. Theor. Biol.*, 1982 (cit. on p. 15).

[43] L. Chen, R. Oughtred, H. M. Berman and J. Westbrook, 'TargetDB: a target registration database for structural genomics projects', *Bioinformatics*, 2004 (cit. on pp. 34, 54, 65, 85, 109).

[44] Y.-J. Chen, P. Liu, A. A. K. Nielsen, J. A. N. Brophy, K. Clancy, T. Peterson and C. A. Voigt, 'Characterization of 582 natural and synthetic terminators and quantification of their design constraints', *Nat. Methods*, 2013 (cit. on p. 49).

[45] J. X. Chin, B. K.-S. Chung and D.-Y. Lee, 'Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design', *Bioinformatics*, 2014 (cit. on p. 83).

[46] F. Chiti, M. Stefani, N. Taddei, G. Ramponi and C. M. Dobson, 'Rationalization of the effects of mutations on peptide and protein aggregation rates', *Nature*, 2003 (cit. on pp. 15, 54, 64).

[47] H. J. Cho, B. M. Oh, J.-T. Kim, J. Lim, S. Y. Park, Y. S. Hwang, K. E. Baek, B.-Y. Kim, I. Choi and H. G. Lee, 'Efficient Interleukin-21 Production by Optimization of Codon and Signal Peptide in Chinese Hamster Ovarian Cells', *J. Microbiol. Biotechnol.*, 2019 (cit. on p. 71).

[48] B. K.-S. Chung and D.-Y. Lee, 'Computational codon optimization of synthetic gene for protein expression', *BMC Syst. Biol.*, 2012 (cit. on pp. 14, 31, 40, 49).

[49] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. L. de Hoon, 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*, 2009 (cit. on p. 77).

[50] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. L. de Hoon, 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*, 2009 (cit. on pp. 55, 56, 65, 130).

[51] T. J. Cole and M. S. Brewer, 'TOXIFY: a deep learning approach to classify animal venom proteins', *PeerJ*, 2019 (cit. on p. 71).

[52] S. Costa, A. Almeida, A. Castro and L. Domingues, 'Fusion tags for protein solubility, purification and immunogenicity in *Escherichia coli*: the novel Fh8 system', *Front. Microbiol.*, 2014 (cit. on pp. 15, 53, 55, 82).

[53] P. Craveur, A. P. Joseph, J. Esque, T. J. Narwani, F. Noël, N. Shinada, M. Goguet, S. Leonard, P. Poulain, O. Bertrand, G. Faure, J. Rebehmed, A. Ghozlane, L. S. Swapna, R. M. Bhaskara, J. Barnoud, S. Téletchéa, V. Jallu, J. Cerny, B. Schneider, C. Etchebest, N. Srinivasan, J.-C. Gelly and A. G. de Brevern, 'Protein flexibility in the light of structural alphabets', *Front. Mol. Biosci.*, 2015 (cit. on pp. 54, 56).

[54] R. Datta, A. Waheed, G. N. Shah and W. S. Sly, 'Signal sequence mutation in autosomal dominant form of hypoparathyroidism induces apoptosis that is corrected by a chemical chaperone', *Proc. Natl. Acad. Sci. U. S. A.*, 2007 (cit. on p. 70).

[55] G. D. Davis, C. Elisee, D. M. Newham and R. G. Harrison, 'New fusion protein systems designed to give soluble expression in Escherichia coli', *Biotechnol. Bioeng.*, 1999 (cit. on pp. 61, 62, 64).

[56] E. R. DeLong, D. M. DeLong and D. L. Clarke-Pearson, 'Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach', *Biometrics*, 1988 (cit. on p. 51).

[57] F. Delvigne, J. Baert, H. Sassi, P. Fickers, A. Grünberger and C. Dusny, 'Taking control over microbial populations: Current approaches for exploiting biological noise in bioprocesses', *Biotechnol. J.*, 2017 (cit. on p. 35).

[58] A. L. Demain and P. Vaishnav, 'Production of recombinant proteins by microbes and higher organisms', *Biotechnol. Adv.*, 2009 (cit. on p. 1).

[59] U. Deuschle, W. Kammerer, R. Gentz and H. Bujard, 'Promoters of *Escherichia coli*: a hierarchy of in vivo strength indicates alternate structures', *Embo J.*, 1986 (cit. on p. 35).

[60] A. A. Diaz, E. Tomba, R. Lennarson, R. Richard, M. J. Bagajewicz and R. G. Harrison, 'Prediction of protein solubility in *Escherichia coli* using logistic regression', *Biotechnol. Bioeng.*, 2010 (cit. on pp. 15, 54, 61, 64).

[61] Y. Ding, C. Y. Chan and C. E. Lawrence, 'RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble', *Rna*, 2005 (cit. on p. 9).

[62] C. J. van Dolleweerd, S. A. Kessans, K. C. Van de Bittner, L. Y. Bustamante, R. Bundela *et al.*, 'MIDAS: A Modular DNA Assembly System for Synthetic Biology', *ACS Synth. Biol.*, 2018 (cit. on pp. 45, 100, 103, 106, 107).

[63] P. Dong and Z. Liu, 'Shaping development by stochasticity and dynamics in gene regulation', *Open Biol.*, 2017 (cit. on p. 11).

[64] Y.-w. Dong, M.-l. Liao, X.-l. Meng and G. N. Somero, 'Structural flexibility and protein adaptation to temperature: Molecular dynamics analysis of malate dehydrogenases of marine molluscs', *Proc. Natl. Acad. Sci. U. S. A.*, 2018 (cit. on p. 18).

[65] J. W. Dubendorf and F. W. Studier, 'Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor', *J. Mol. Biol.*, 1991 (cit. on p. 1).

[66] S. Dvir, L. Velten, E. Sharon, D. Zeevi, L. B. Carey, A. Weinberger and E. Segal, 'Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast', *Proc. Natl. Acad. Sci. U. S. A.*, 2013 (cit. on pp. 31, 32, 82, 108).

[67] S. R. Eddy, 'How do RNA folding algorithms work?', *Nat. Biotechnol.*, 2004 (cit. on p. 9).

[68] R. C. Edgar, 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics*, 2010 (cit. on p. 66).

[69] A. Espah Borujeni, A. S. Channarasappa and H. M. Salis, 'Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites', *Nucleic Acids Res.*, 2014 (cit. on p. 7).

[70] D. Esposito and D. K. Chatterjee, 'Enhancement of soluble protein expression through the use of fusion tags', *Curr. Opin. Biotechnol.*, 2006 (cit. on pp. 53, 82).

[71] G. Estrada, E. Villegas and G. Corzo, 'Spider venoms: a rich source of acylpolyamines and peptides as new leads for CNS drugs', *Nat. Prod. Rep.*, 2007 (cit. on p. 70).

[72] C. Famılia, S. R. Dennison, A. Quintas and D. A. Phoenix, 'Prediction of Peptide and Protein Propensity for Amyloid Formation', *PLoS One*, 2015 (cit. on p. 62).

[73] R. Freudl, 'Signal peptides for recombinant protein secretion in bacterial expression systems', *Microb. Cell Fact.*, 2018 (cit. on p. 83).

[74] B. G. Fry, K. Roelants, D. E. Champagne, H. Scheib, J. D. A. Tyndall, G. F. King, T. J. Nevalainen, J. A. Norman, R. J. Lewis, R. S. Norton, C. Renjifo and R. C. R. de la Vega, 'The Toxicogenomic Multiverse: Convergent Recruitment of Proteins Into Animal Venoms', *Annu. Rev. Genomics Hum. Genet.*, 2009 (cit. on pp. 70, 75).

[75] B. G. Fry, K. Roelants, K. Winter, W. C. Hodgson, L. Griesman, H. F. Kwok, D. Scanlon, J. Karas, C. Shaw, L. Wong and J. A. Norman, 'Novel venom proteins produced by differential domain-expression strategies in beaded lizards and gila monsters (genus Heloderma)', *Mol. Biol. Evol.*, 2010 (cit. on p. 75).

[76] L. Fu, B. Niu, Z. Zhu, S. Wu and W. Li, 'CD-HIT: accelerated for clustering the next-generation sequencing data', *Bioinformatics*, 2012 (cit. on pp. 45, 76).

[77] M. Fuhrmann, A. Hausherr, L. Ferbitz, T. Schödl, M. Heitzer *et al.*, 'Monitoring dynamic expression of nuclear genes in *Chlamydomonas reinhardtii* by using a synthetic luciferase reporter gene', *Plant Mol. Biol.*, 2004 (cit. on p. 50).

[78] M. Futatsumori-Sugai and K. Tsumoto, 'Signal peptide design for improving recombinant protein secretion in the baculovirus expression vector system', *Biochem. Biophys. Res. Commun.*, 2010 (cit. on pp. 20, 71).

[79] R. Gacesa, D. J. Barlow and P. F. Long, 'Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions', *PeerJ Comput. Sci.*, 2016 (cit. on p. 71).

[80] P. P. Gardner and H. Eldai, 'Annotating RNA motifs in sequences and alignments', *Nucleic Acids Res.*, 2015 (cit. on pp. 49, 87).

[81] P. Gaspar, G. Moura, M. A. S. Santos and J. L. Oliveira, 'mRNA secondary structure optimization using a correlated stem-loop prediction', *Nucleic Acids Res.*, 2013 (cit. on pp. 21, 47).

[82] L. Gomes, G. Monteiro and F. Mergulhão, 'The Impact of IPTG Induction on Plasmid Stability and Heterologous Protein Expression by Biofilms', *Int. J. Mol. Sci.*, 2020 (cit. on p. 46).

[83] A. Grote, K. Hiller, M. Scheer, R. Münch, B. Nörtemann, D. C. Hempel and D. Jahn, 'JCat: a novel tool to adapt codon usage of a target gene to its potential expression host', *Nucleic Acids Res.*, 2005 (cit. on p. 83).

[84] S. Gupta, P. Kapoor, K. Chaudhary, A. Gautam, R. Kumar, Open Source Drug Discovery Consortium and G. P. S. Raghava, '*In silico* approach for predicting toxicity of peptides and proteins', *PLoS One*, 2013 (cit. on pp. 71, 83).

[85] K. Guruprasad, B. B. Reddy and M. W. Pandit, 'Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence', *Protein Eng. Des. Sel.*, 1990 (cit. on p. 17).

[86] C. Gustafsson, S. Govindarajan and J. Minshull, 'Codon bias and heterologous protein expression', *Trends Biotechnol.*, 2004 (cit. on p. 14).

[87] G. A. Gutman and G. W. Hatfield, 'Nonrandom utilization of codon pairs in *Escherichia coli*', *Proc. Natl. Acad. Sci. U. S. A.*, 1989 (cit. on pp. 6, 7, 31, 82).

[88] N. Habibi, S. Z. Mohd Hashim, A. Norouzi and M. R. Samian, 'A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in *Escherichia coli*', *BMC Bioinform.*, 2014 (cit. on pp. 19, 54).

[89] B. Hames, N. Hooper, B. Hames *et al.*, *Biochemistry/David Hames and Nigel Hooper.* 2005 (cit. on p. 4).

[90] X. Han, W. Ning, X. Ma, X. Wang and K. Zhou, 'Improve Protein Solubility and Activity based on Machine Learning Models', *bioRxiv*, 2019 (cit. on p. 63).

[91] G. Hanson and J. Coller, 'Codon optimality, bias and usage in translation and mRNA decay', *Nat. Rev. Mol. Cell Biol.*, 2018 (cit. on p. 30).

[92] R. G. Harrison, 'Expression of soluble heterologous proteins via fusion with NusA protein', *Innovations*, 2000 (cit. on pp. 61, 62).

[93] W. K. Hastings, 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika*, 1970 (cit. on p. 22).

[94] M. Hebditch, M. A. Carballo-Amador, S. Charonis, R. Curtis and J. Warwicker, 'Protein-Sol: a web tool for predicting protein solubility from sequence', *Bioinformatics*, 2017 (cit. on pp. 19, 54, 61, 62, 83).

[95] D. Heckmann, C. J. Lloyd, N. Mih, Y. Ha, D. C. Zielinski, Z. B. Haiman, A. A. Desouki, M. J. Lercher and B. O. Palsson, 'Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models', *Nat. Commun.*, 2018 (cit. on pp. 19, 54).

[96] R. S. Hegde and H. D. Bernstein, 'The surprising complexity of signal sequences', *Trends Biochem. Sci.*, 2006 (cit. on p. 70).

[97] G. von Heijne, 'Patterns of amino acids near signal-sequence cleavage sites', *Eur. J. Biochem.*, 1983 (cit. on p. 72).

[98] ——, 'Signal sequences', *J. Mol. Biol.*, 1985 (cit. on pp. 19, 70).

[99] ——, 'The signal peptide', *J. Membr. Biol.*, 1990 (cit. on pp. 70, 71, 83).

[100] D. Held, K. Yaeger and R. Novy, 'New coexpression vectors for expanded compatibilities in *E. coli*', Novagen, Tech. Rep., 2003 (cit. on p. 46).

[101] K. Hiller, A. Grote, M. Scheer, R. Münch and D. Jahn, 'PrediSi: prediction of signal peptides and their cleavage positions', *Nucleic Acids Res.*, 2004 (cit. on p. 83).

[102] S. Hirose and T. Noguchi, 'ESPRESSO: a system for estimating protein expression and solubility in protein expression systems', *Proteomics*, 2013 (cit. on pp. 19, 54).

[103] J. J. Hiu and M. K. K. Yap, 'Cytotoxicity of snake venom enzymatic toxins: phospholipase A2 and l-amino acid oxidase', *Biochem. Soc. Trans.*, 2020 (cit. on p. 76).

[104] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker and P. Schuster, 'Fast folding and comparison of RNA secondary structures', *Monatshefte für Chemie / Chemical Monthly*, 1994 (cit. on p. 45).

[105] *Home : Metrics*, `http://targetdb.rcsb.org/metrics/` (cit. on p. 2).

[106] J. Hon, M. Marusiak, T. Martinek, A. Kunka, J. Zendulka, D. Bednar and J. Damborsky, 'SoluProt: Prediction of Soluble Protein Expression in *Escherichia coli*', *Bioinformatics*, 2021 (cit. on p. 83).

[107] Q. Hou, R. Bourgeas, F. Pucci and M. Rooman, 'Computational analysis of the amino acid interactions that promote or decrease protein solubility', *Sci. Rep.*, 2018 (cit. on pp. 2, 53, 82).

[108] Q. Hou, J. M. Kwasigroch, M. Rooman and F. Pucci, 'SOLart: a structure-based method to predict protein solubility and aggregation', *Bioinformatics*, 2020 (cit. on pp. 19, 62).

[109] H.-L. Huang, P. Charoenkwan, T.-F. Kao, H.-C. Lee, F.-L. Chang, W.-L. Huang, S.-J. Ho, L.-S. Shu, W.-L. Chen and S.-Y. Ho, 'Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition', *BMC Bioinform.*, 2012 (cit. on pp. 19, 63).

[110] J. D. Hunter, 'Matplotlib: A 2D Graphics Environment', *Comput. Sci. Eng.*, 2007 (cit. on pp. 51, 80).

[111] S. Idicula-Thomas and P. V. Balaji, 'Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*', *Protein Sci.*, 2005 (cit. on pp. 19, 63, 64).

[112] T. Ikemura, 'Codon usage and tRNA content in unicellular and multicellular organisms', *Mol. Biol. Evol.*, 1985 (cit. on p. 7).

[113] L. Ingber, 'Adaptive simulated annealing (ASA): Lessons learned', 2000. eprint: `cs/0001018` (cit. on pp. 21, 47).

[114] K. Itakura, T. Hirose, R. Crea, A. D. Riggs, H. L. Heyneker, F. Bolivar and H. W. Boyer, 'Expression in *Escherichia coli* of a chemically synthesized gene for the hormone somatostatin', *Science*, 1977 (cit. on p. 1).

[115] G. James, D. Witten, T. Hastie and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013 (cit. on p. 27).

[116] J. Janin, 'Surface and inside volumes in globular proteins', *Nature*, 1979 (cit. on p. 15).

[117] B. Jia and C. O. Jeon, 'High-throughput recombinant protein expression in *Escherichia coli*: current status and future perspectives', *Open Biol.*, 2016 (cit. on pp. 1, 63).

[118] F. Jungo, L. Bougueleret, I. Xenarios and S. Poux, 'The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data', *Toxicon*, 2012 (cit. on pp. 77, 89).

[119] L. Käll, A. Krogh and E. L. L. Sonnhammer, 'A combined transmembrane topology and signal peptide prediction method', *J. Mol. Biol.*, 2004 (cit. on p. 83).

[120] I. Kalvari, J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, R. D. Finn and A. I. Petrov, 'Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families', *Nucleic Acids Res.*, 2018 (cit. on p. 49).

[121] I. Kalvari, E. P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weinberg, E. Rivas, S. R. Eddy, R. D. Finn, A. Bateman and A. I. Petrov, 'Rfam 14: expanded coverage of metagenomic, viral and microRNA families', *Nucleic Acids Res.*, 2021 (cit. on p. 87).

[122] Y. W. Kam, F. Kien, A. Roberts, Y. C. Cheung, E. W. Lamirande, L. Vogel, S. L. Chu, J. Tse, J. Guarner, S. R. Zaki, K. Subbarao, M. Peiris, B. Nal and R. Altmeyer, 'Antibodies against trimeric S glycoprotein protect hamsters against SARS-CoV challenge despite their capacity to mediate FcgammaRII-dependent entry into B cells in vitro', *Vaccine*, 2007 (cit. on p. 136).

[123] S. Kanaya, Y. Yamada, Y. Kudo and T. Ikemura, 'Studies of codon usage and trna genes of 18 unicellular organisms and quantification of bacillus subtilis tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis', *Gene*, 1999 (cit. on p. 6).

[124] P. A. Karplus and G. E. Schulz, 'Prediction of chain flexibility in proteins', *Naturwissenschaften*, 1985 (cit. on pp. 17, 18, 56, 63, 129, 130).

[125] A. Karyolaimos, H. Ampah-Korsah, T. Hillenaar, A. M. Borras, K. M. Dolata, S. Sievers, K. Riedel, R. Daniels and J.-W. de Gier, 'Enhancing Recombinant Protein Yields in the *E. coli* Periplasm by Combining Signal Peptide and Production Rate Screening', *Front. Microbiol.*, 2019 (cit. on pp. 20, 71).

[126] A. Karyolaimos, K. M. Dolata, M. Antelo-Varela, A. M. Borras, R. Elfageih, S. Sievers, D. Becher, K. Riedel and J.-W. de Gier, '*Escherichia coli* can adapt its protein translocation machinery for enhanced periplasmic recombinant protein production', *Front. Bioeng. Biotechnol.*, 2020 (cit. on p. 83).

[127] J. M. Keith, P. Adams, D. Bryant, D. P. Kroese, K. R. Mitchelson, D. A. E. Cochran and G. H. Lala, 'A simulated annealing algorithm for finding consensus sequences', *Bioinformatics*, 2002 (cit. on pp. 21, 22, 47).

[128] S. Khurana, R. Rawi, K. Kunji, G.-Y. Chuang, H. Bensmail and R. Mall, 'DeepSol: a deep learning framework for sequence-based protein solubility prediction', *Bioinformatics*, 2018 (cit. on pp. 61, 62).

[129] A. Kimelman, A. Levy, H. Sberro, S. Kidron, A. Leavitt, G. Amitai, D. R. Yoder-Himes, O. Wurtzel, Y. Zhu, E. M. Rubin and R. Sorek, 'A vast collection of microbial genes that are toxic to bacteria', *Genome Res.*, 2012 (cit. on p. 30).

[130] G. F. King, 'Venoms as a platform for human drugs: translating toxins into therapeutics', *Expert Opin. Biol. Ther.*, 2011 (cit. on p. 70).

[131] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, 'Optimization by Simulated Annealing', *Science*, 1983 (cit. on pp. 21, 22, 47, 78).

[132] R. M. Kramer, V. R. Shende, N. Motl, C. Nick Pace and J. Martin Scholtz, 'Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility', *Biophysical Journal*, 2012 (cit. on pp. 2, 53, 63, 82).

[133] G. Kudla, A. W. Murray, D. Tollervey and J. B. Plotkin, 'Coding-sequence determinants of gene expression in *Escherichia coli*', *Science*, 2009 (cit. on pp. 7, 31, 37, 82).

[134] I. Kufareva and R. Abagyan, 'Methods of protein structure comparison', in *Homology Modeling*, Springer, 2011 (cit. on p. 18).

[135] A. Kuriata, V. Iglesias, J. Pujols, M. Kurcinski, S. Kmiecik and S. Ventura, 'Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility', *Nucleic Acids Res.*, 2019 (cit. on p. 62).

[136] J. Kyte and R. F. Doolittle, 'A simple method for displaying the hydropathic character of a protein', *J. Mol. Biol.*, 1982 (cit. on pp. 15, 58, 78).

[137] A. F. Lacerda, E. A. R. Vasconcelos, P. B. Pelegrini and M. F. Grossi de Sa, 'Antifungal defensins and their role in plant defense', *Front. Microbiol.*, 2014 (cit. on p. 76).

[138] J. C. Lagarias, J. A. Reeds, M. H. Wright and P. E. Wright, 'Convergence properties of the Nelder–Mead simplex method in low dimensions', *SIAM J. Optim.*, 1998 (cit. on p. 24).

[139] M. Lebendiker and T. Danieli, 'Production of prone-to-aggregate proteins', *FEBS Lett.*, 2014 (cit. on p. 65).

[140] E. D. Levy, S. De and S. A. Teichmann, 'Cellular crowding imposes global constraints on the chemistry and evolution of proteomes', *Proc. Natl. Acad. Sci. U. S. A.*, 2012 (cit. on pp. 60, 65).

[141] L. Li, J. Huang and Y. Lin, 'Snake Venoms in Cancer Therapy: Past, Present and Future', *Toxins*, 2018 (cit. on p. 70).

[142] W. Li and A. Godzik, 'CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, 2006 (cit. on p. 45).

[143] C. S. Lim, S. J. T. Wardell, T. Kleffmann and C. M. Brown, 'The exon–intron gene structure upstream of the initiation codon predicts translation efficiency', *Nucleic Acids Res.*, 2018 (cit. on pp. 30, 82).

[144] S. Lindgreen, P. P. Gardner and A. Krogh, 'MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing', *Bioinformatics*, 2007 (cit. on pp. 21, 47).

[145] B. Lomonte and J. Rangel, 'Snake venom Lys49 myotoxins: From phospholipases A(2) to non-enzymatic membrane disruptors', *Toxicon*, 2012 (cit. on p. 76).

[146] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler and I. L. Hofacker, 'ViennaRNA Package 2.0', *Algorithms Mol. Biol.*, 2011 (cit. on pp. 44, 45, 85).

[147] R. Lorenz, I. L. Hofacker and P. F. Stadler, 'RNA folding with hard and soft constraints', *Algorithms Mol. Biol.*, 2016 (cit. on p. 45).

[148] R. Lorenz, M. T. Wolfinger, A. Tanzer and I. L. Hofacker, 'Predicting RNA secondary structures from sequence and probing data', *Methods*, 2016 (cit. on p. 9).

[149] W. W. Lorenz, M. J. Cormier, D. J. O'Kane, D. Hua, A. A. Escher *et al.*, 'Expression of the *Renilla reniformis* luciferase gene in mammalian cells', *J. Biolumin. Chemilumin.*, 1996 (cit. on p. 50).

[150] J. Luirink and B. Dobberstein, 'Mammalian and *Escherichia coli* signal recognition particles', *Mol. Microbiol.*, 1994 (cit. on p. 83).

[151] J. Ma, 'Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes', *Structure*, 2005 (cit. on pp. 19, 63).

[152] R. J. Ma, Y. H. Wang, L. Liu, L. L. Bai and R. Ban, 'Production enhancement of the extracellular lipase LipA in *Bacillus subtilis*: Effects of expression system and Sec pathway components', *Protein Expr. Purif*, 2018 (cit. on p. 83).

[153] M. Mann, P. R. Wright and R. Backofen, 'IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions', *Nucleic Acids Res.*, 2017 (cit. on p. 44).

[154] K. A. Marill, Y. Chang, K. F. Wong and A. B. Friedman, 'Estimating negative likelihood ratio confidence when test sensitivity is 100%: A bootstrapping approach', *Stat. Methods Med. Res.*, 2017 (cit. on p. 51).

[155] M. A. Marra, S. J. M. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. N. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, A. Cloutier, S. M. Coughlin, D. Freeman, N. Girn, O. L. Griffith, S. R. Leach, M. Mayo, H. McDonald, S. B. Montgomery, P. K. Pandoh, A. S. Petrescu, A. G. Robertson, J. E. Schein, A. Siddiqui, D. E. Smailus, J. M. Stott, G. S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T. F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G. A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R. C. Brunham, M. Krajden, M. Petric, D. M. Skowronski, C. Upton and R. L. Roper, 'The Genome sequence of the SARS-associated coronavirus', *Science*, 2003 (cit. on p. 65).

[156] D. M. Mauger, B. J. Cabral, V. Presnyak, S. V. Su, D. W. Reid, B. Goodman, K. Link, N. Khatwani, J. Reynders, M. J. Moore *et al.*, 'mRNA structure regulates protein expression through changes in functional half-life', *Proc. Natl. Acad. Sci. U. S. A.*, 2019 (cit. on pp. 4, 7, 43, 82).

[157] S. Mazurenko, 'Predicting protein stability and solubility changes upon mutations: data perspective', *ChemCatChem*, 2020 (cit. on p. 82).

[158] J. S. McCaskill, 'The equilibrium partition function and base pair binding probabilities for RNA secondary structure', *Biopolymers*, 1990 (cit. on pp. 9, 10).

[159] W. McKinney, 'Data Structures for Statistical Computing in Python', in *Proceedings of the 9th Python in Science Conference*, 2010 (cit. on pp. 51, 68, 79).

[160] K. J. Millman and M. Aivazis, 'Python for Scientists and Engineers', *Comput. Sci. Eng.*, 2011 (cit. on pp. 51, 66).

[161] P. Mittal, J. Brindle, J. Stephen, J. B. Plotkin and G. Kudla, 'Codon usage influences fitness through RNA toxicity', *Proc. Natl. Acad. Sci. U. S. A.*, 2018 (cit. on pp. 39, 94).

[162] F. Mohammad, R. Green and A. R. Buskirk, 'A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution', *Elife*, 2019 (cit. on pp. 45, 110).

[163] U. Mückstein, H. Tafer, J. Hackermüller, S. H. Bernhart, P. F. Stadler and I. L. Hofacker, 'Thermodynamics of RNA–RNA binding', *Bioinformatics*, 2006 (cit. on pp. 12, 43, 45).

[164] G. Naamati, M. Askenazi and M. Linial, 'ClanTox: a classifier of short animal toxins', *Nucleic Acids Res.*, 2009 (cit. on p. 83).

[165] ——, 'A predictor for toxin-like proteins exposes cell modulator candidates within viral genomes', *Bioinformatics*, 2010 (cit. on p. 71).

[166] E. Natan, T. Endoh, L. Haim-Vilmovsky, T. Flock, G. Chalancon, J. T. S. Hopper, B. Kintses, P. Horvath, L. Daruka, G. Fekete, C. Pál, B. Papp, E. Oszi, Z. Magyar, J. A. Marsh, A. H. Elcock, M. M. Babu, C. V. Robinson, N. Sugimoto and S. A. Teichmann, 'Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins', *Nat. Struct. Mol. Biol.*, 2018 (cit. on pp. 61, 64).

[167] E. P. Nawrocki and S. R. Eddy, 'Infernal 1.1: 100-fold faster RNA homology searches', *Bioinformatics*, 2013 (cit. on pp. 49, 87).

[168] J. A. Nelder and R. Mead, 'A Simplex Method for Function Minimization', *Comput. J.*, 1965 (cit. on pp. 23, 57, 66).

[169] B. W. Neuman, G. Kiss, A. H. Kunding, D. Bhella, M. F. Baksh, S. Connelly, B. Droese, J. P. Klaus, S. Makino, S. G. Sawicki, S. G. Siddell, D. G. Stamou, I. A. Wilson, P. Kuhn and M. J. Buchmeier, 'A structural analysis of M protein in coronavirus assembly and morphology', *J. Struct. Biol.*, 2011 (cit. on p. 136).

[170] H. Nielsen and A. Krogh, 'Prediction of signal peptides and signal anchors by a hidden Markov model', *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 1998 (cit. on p. 70).

[171] T. Nieuwkoop, N. J. Claassens and J. van der Oost, 'Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design', *Microb. Biotechnol.*, 2019 (cit. on p. 43).

[172] T. Nieuwkoop, M. Finger-Bou, J. van der Oost and N. J. Claassens, 'The Ongoing Quest to Crack the Genetic Code for Protein Production', *Mol. Cell*, 2020 (cit. on p. 82).

[173] T. Nilsson, M. Mann, R. Aebersold, J. R. Yates 3rd, A. Bairoch and J. J. M. Bergeron, 'Mass spectrometry in high-throughput proteomics: ready for the big time', *Nat. Methods*, 2010 (cit. on p. 35).

[174] T. Niwa, B.-W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda and H. Taguchi, 'Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins', *Proc. Natl. Acad. Sci. U. S. A.*, 2009 (cit. on pp. 14, 60, 62, 63, 65).

[175] W. L. Noderer, R. J. Flockhart, A. Bhaduri, A. J. Diaz de Arce, J. Zhang, P. A. Khavari and C. L. Wang, 'Quantitative analysis of mammalian translation initiation sites by FACS-seq', *Mol. Syst. Biol.*, 2014 (cit. on pp. 32, 108).

[176] E. M. Novoa, M. Pavon-Eternod, T. Pan and L. R. de Pouplana, 'A role for trna modifications in genome structure and codon usage', *Cell*, 2012 (cit. on p. 6).

[177] T. E. Oliphant, 'Python for Scientific Computing', *Comput. Sci. Eng.*, 2007 (cit. on pp. 51, 66).

[178] I. A. Osterman, Z. S. Chervontseva, S. A. Evfratov, A. V. Sorokina, V. A. Rodin, M. P. Rubtsova, E. S. Komarova, T. S. Zatsepin, M. R. Kabilov, A. A. Bogdanov, M. S. Gelfand, O. A. Dontsova and P. V. Sergiev, 'Translation at first sight: the influence of leading codons', *Nucleic Acids Res.*, 2020 (cit. on p. 31).

[179] H. Owji, N. Nezafat, M. Negahdaripour, A. Hajiebrahimi and Y. Ghasemi, 'A comprehensive review of signal peptides: Structure, roles, and applications', *Eur. J. Cell Biol.*, 2018 (cit. on pp. 70, 83).

[180] T. Palmer and B. C. Berks, 'The twin-arginine translocation (Tat) protein export pathway', *Nat. Rev. Microbiol.*, 2012 (cit. on p. 83).

[181] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, 'Scikit-learn: Machine Learning in Python', *J. Mach. Learn. Res.*, 2011 (cit. on pp. 51, 68, 80).

[182] J. Pelletier and N. Sonenberg, 'The involvement of mRNA secondary structure in protein synthesis', *Biochem. Cell Biol.*, 1987 (cit. on p. 43).

[183] C. Peng, C. Shi, X. Cao, Y. Li, F. Liu and F. Lu, 'Factors Influencing Recombinant Protein Secretion Efficiency in Gram-Positive Bacteria: Signal Peptide and Beyond', *Front. Bioeng. Biotechnol.*, 2019 (cit. on p. 71).

[184] T. N. Petersen, S. Brunak, G. von Heijne and H. Nielsen, 'SignalP 4.0: discriminating signal peptides from transmembrane regions', *Nature Methods*, 2011 (cit. on p. 78).

[185] J. B. Plotkin and G. Kudla, 'Synonymous but not the same: the causes and consequences of codon bias', *Nat. Rev. Genet.*, 2011 (cit. on pp. 7, 31, 82).

[186] S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez and R. D. Finn, 'HMMER web server: 2018 update', *Nucleic Acids Res.*, 2018 (cit. on pp. 68, 87).

[187] W. Presse, B. P. Flannery, S. A. Teukolsky and W. Wetterling, *Numerical recipes in C*. 1988 (cit. on pp. 21, 22, 24).

[188] J. Puetz and F. M. Wurm, 'Recombinant proteins for industrial versus pharmaceutical purposes: a review of process and pricing', *Processes*, 2019 (cit. on p. 1).

[189] P. Puigbò, E. Guzmán, A. Romeu and S. Garcia-Vallvé, 'OPTIMIZER: a web server for optimizing the codon usage of DNA sequences', *Nucleic Acids Res.*, 2007 (cit. on p. 83).

[190] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: `http://www.R-project.org/` (cit. on p. 51).

[191] D. Raab, M. Graf, F. Notka, T. Schödl and R. Wagner, 'The GeneOptimizer Algorithm: using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization', *Syst. Synth. Biol.*, 2010 (cit. on pp. 14, 31).

[192] P. Radivojac, 'Protein flexibility and intrinsic disorder', *Protein Sci.*, 2004 (cit. on p. 63).

[193] R. Ragone, F. Facchiano, A. Facchiano, A. M. Facchiano and G. Colonna, 'Flexibility plot of proteins', *Protein Eng. Des. Sel.*, 1989 (cit. on pp. 56, 57).

[194] C. V. Rao, D. M. Wolf and A. P. Arkin, 'Control, exploitation and tolerance of intracellular noise', *Nature*, 2002 (cit. on p. 11).

[195] R. Rawi, R. Mall, K. Kunji, C.-H. Shen, P. D. Kwong and G.-Y. Chuang, 'PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine', *Bioinformatics*, 2018 (cit. on pp. 61, 62).

[196] M. d. Reis and M. d. Reis, 'Solving the riddle of codon usage preferences: a test for translational selection', *Nucleic Acids Res.*, 2004 (cit. on pp. 6, 7, 31, 82).

[197] G. L. Rosano and E. A. Ceccarelli, 'Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted *Escherichia coli* strain', *Microb. Cell Factories* ., 2009 (cit. on p. 14).

[198] ——, 'Recombinant protein expression in *Escherichia coli*: advances and challenges', *Front. Microbiol.*, 2014 (cit. on pp. 1, 30, 46, 63, 82).

[199] G. L. Rosano, E. S. Morales and E. A. Ceccarelli, 'New tools for recombinant protein production in *Escherichia coli*: A 5-year update', *Protein Sci.*, 2019 (cit. on p. 83).

[200] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee and M. H. Zehfus, 'Hydrophobicity of amino acid residues in globular proteins', *Science*, 1985 (cit. on p. 15).

[201] S. L. Rusch and D. A. Kendall, 'Interactions That Drive Sec-Dependent Bacterial Protein Transport', *Biochemistry*, 2007 (cit. on p. 83).

[202] R. Sabi and T. Tuller, 'Modelling the Efficiency of Codon–tRNA Interactions Based on Codon Usage Bias', *DNA Research*, 2014 (cit. on pp. 6, 7, 31, 82).

[203] H. M. Salis, E. A. Mirsky and C. A. Voigt, 'Automated design of synthetic ribosome binding sites to control protein expression', *Nat. Biotechnol.*, 2009 (cit. on pp. 14, 21, 31, 43, 47).

[204] J. Sambrook and D. W. Russell, *Molecular cloning: a laboratory manual. Vol. 3*. CSHL Press, 2001 (cit. on p. 44).

[205] R. P. Samy, B. G. Stiles, O. L. Franco, G. Sethi and L. H. K. Lim, 'Animal venoms as antimicrobial agents', *Biochem. Pharmacol.*, 2017 (cit. on p. 70).

[206] C. Savojardo, P. L. Martelli, P. Fariselli and R. Casadio, 'DeepSig: deep learning improves signal peptide detection in proteins', *Bioinformatics*, 2017 (cit. on pp. 79, 83).

[207] R. O. Schlechter, H. Jun, M. Bernach, S. Oso, E. Boyd *et al.*, 'Chromatic Bacteria - A Broad Host-Range Plasmid and Chromosomal Insertion Toolbox for Fluorescent Protein Expression in Bacteria', *Front. Microbiol.*, 2018 (cit. on p. 40).

[208] R. O. Schlechter, D. M. Remus and M. N. P. Remus-Emsermann, 'Constitutively expressed fluorescent proteins allow to track bacterial growth and to determine relative fitness of bacteria in mixed cultures', 2020 (cit. on p. 40).

[209] A. Schlessinger and B. Rost, 'Protein flexibility and rigidity predicted from sequence', *Proteins*, 2005 (cit. on pp. 17, 18, 63, 129).

[210] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen and M. Selbach, 'Global quantification of mammalian gene expression control', *Nature*, 2011 (cit. on p. 30).

[211] S. Seabold and J. Perktold, 'Statsmodels: Econometric and statistical modeling with python', in *Proceedings of the 9th Python in Science Conference*, 2010 (cit. on p. 68).

[212] C. Y. Seiler, J. G. Park, A. Sharma, P. Hunter, P. Surapaneni *et al.*, 'DNASU plasmid and PSI:Biology-Materials repositories: resources to accelerate biological research', *Nucleic Acids Res.*, 2014 (cit. on p. 109).

[213] C. Y. Seiler, J. G. Park, A. Sharma, P. Hunter, P. Surapaneni, C. Sedillo, J. Field, R. Algar, A. Price, J. Steel, A. Throop, M. Fiacco and J. LaBaer, 'DNASU plasmid and PSI:Biology-Materials repositories: resources to accelerate biological research', *Nucleic Acids Res.*, 2014 (cit. on pp. 34, 54, 62, 65, 85).

[214] P. M. Sharp and W. H. Li, 'The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications', *Nucleic Acids Res.*, 1987 (cit. on pp. 6, 7, 31, 35, 46, 82).

[215] K. L. Shaw, G. R. Grimsley, G. I. Yakovlev, A. A. Makarov and C. N. Pace, 'The effect of net charge on the solubility, activity, and stability of ribonuclease Sa', *Protein Sci.*, 2001 (cit. on p. 17).

[216] C.-S. Shi, N. R. Nabar, N.-N. Huang and J. H. Kehrl, 'SARS-Coronavirus Open Reading Frame-8b triggers intracellular stress pathways and activates NLRP3 inflammasomes', *Cell Death Discov.*, 2019 (cit. on p. 136).

[217] J. Shine and L. Dalgarno, 'The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites', *Proc. Natl. Acad. Sci. U. S. A.*, 1974 (cit. on p. 34).

[218] P. Smialowski, G. Doose, P. Torkler, S. Kaufmann and D. Frishman, 'PROSO II–a new method for protein solubility prediction', *Febs J.*, 2012 (cit. on p. 83).

[219] M. H. de Smit and J. van Duin, 'Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis', *Proc. Natl. Acad. Sci. U. S. A.*, 1990 (cit. on pp. 31, 82).

[220] D. K. Smith, P. Radivojac, Z. Obradovic, A. K. Dunker and G. Zhu, 'Improved amino acid flexibility parameters', *Protein Sci.*, 2003 (cit. on pp. 17, 18, 56, 57, 129).

[221] P. Sormanni, L. Amery, S. Ekizoglou, M. Vendruscolo and B. Popovic, 'Rapid and accurate in silico solubility screening of a monoclonal antibody library', *Sci. Rep.*, 2017 (cit. on pp. 19, 54, 61).

[222] P. Sormanni, F. A. Aprile and M. Vendruscolo, 'The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility', *J. Mol. Biol.*, 2015 (cit. on pp. 61, 83).

[223] R. de Sousa Abreu, L. O. Penalva, E. M. Marcotte and C. Vogel, 'Global signatures of protein and mRNA expression levels', *Mol. Biosyst.*, 2009 (cit. on pp. 3, 30, 82).

[224] R. C. Stevens, 'Design of high-throughput methods of protein production for structural biology', *Structure*, 2000 (cit. on p. 1).

[225] S. G. Stevens and C. M. Brown, 'In silico estimation of translation efficiency in human cell lines: potential evidence for widespread translational control', *PLoS One*, 2013 (cit. on p. 30).

[226] E. J. Stewart, F. Aslund and J. Beckwith, 'Disulfide bond formation in the *Escherichia coli* cytoplasm: an in vivo role reversal for the thioredoxins', *Embo J.*, 1998 (cit. on p. 63).

[227] T. Stoeger, N. Battich and L. Pelkmans, 'Passive noise filtering by cellular compartmentalization', *Cell*, 2016 (cit. on p. 11).

[228] H. U. Stotz, J. G. Thomson and Y. Wang, 'Plant defensins: defense, development and application', *Plant Signal. Behav.*, 2009 (cit. on p. 76).

[229] D. L. Tabb, L. Vega-Montoto, P. A. Rudnick, A. M. Variyath, A.-J. L. Ham, D. M. Bunk, L. E. Kilpatrick, D. D. Billheimer, R. K. Blackman, H. L. Cardasis *et al.*, 'Repeatability and reproducibility in proteomic identifications by liquid chromatography- tandem mass spectrometry', *J. Proteome Res.*, 2009 (cit. on p. 35).

[230] C. Tanford, 'The hydrophobic effect and the organization of living matter', *Science*, 1978 (cit. on p. 15).

[231] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili and X. S. Xie, 'Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells', *Science*, 2010 (cit. on pp. 3, 30, 43, 46, 82).

[232] G. G. Tartaglia, A. Cavalli, R. Pellarin and A. Caflisch, 'The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates', *Protein Sci.*, 2004 (cit. on pp. 15, 54, 64).

[233] S. J. Teague, 'Implications of protein flexibility for drug discovery', *Nat. Rev. Drug Discov.*, 2003 (cit. on pp. 19, 63).

[234] K. Teilum, J. G. Olsen and B. B. Kragelund, 'Functional aspects of protein flexibility', *Cell. Mol. Life Sci.*, 2009 (cit. on p. 17).

[235] G. Terai and K. Asai, 'Improving the prediction accuracy of protein abundance in *Escherichia coli* using mRNA accessibility', *Nucleic Acids Res.*, 2020 (cit. on pp. 43, 82).

[236] G. Terai, S. Kamegai and K. Asai, 'CDSfold: an algorithm for designing a protein-coding sequence with the most stable secondary structure', *Bioinformatics*, 2016 (cit. on pp. 14, 31).

[237] S. R. Trevino, J. Martin Scholtz and C. Nick Pace, 'Amino Acid Contribution to Protein Solubility: Asp, Glu, and Ser Contribute more Favorably than the

other Hydrophilic Amino Acids in RNase Sa', *J. Mol. Biol.*, 2007 (cit. on pp. 53, 63).

[238] K. Tsumoto, D. Ejima, I. Kumagai and T. Arakawa, 'Practical considerations in refolding proteins from inclusion bodies', *Protein Expr. Purif.*, 2003 (cit. on pp. 19, 63).

[239] T. Tuller, Y. Y. Waldman, M. Kupiec and E. Ruppin, 'Translation efficiency is determined by both codon bias and folding energy', *Proc. Natl. Acad. Sci. U. S. A.*, 2010 (cit. on pp. 35, 46).

[240] T. Tuller and H. Zur, 'Multiple roles of the coding sequence 5' end in gene expression regulation', *Nucleic Acids Res.*, 2015 (cit. on pp. 7, 31, 37, 82).

[241] R. Tunney, N. J. McGlincy, M. E. Graham, N. Naddaf, L. Pachter and L. F. Lareau, 'Accurate design of translational output by a neural network model of ribosome distribution', *Nat. Struct. Mol. Biol.*, 2018 (cit. on pp. 35, 39, 46, 110).

[242] D. H. Turner and D. H. Mathews, 'NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure', *Nucleic Acids Res.*, 2009 (cit. on p. 7).

[243] S. U. Umu, A. M. Poole, R. C. Dobson and P. P. Gardner, 'Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea', *Elife*, 2016 (cit. on pp. 11, 31, 39, 82).

[244] E. A. B. Undheim and R. A. Jenner, 'Phylogenetic analyses suggest centipede venom arsenals were repeatedly stocked by horizontal gene transfer', *Nat. Commun.*, 2021 (cit. on pp. 70, 76).

[245] UniProt Consortium, 'UniProt: a worldwide hub of protein knowledge', *Nucleic Acids Res.*, 2019 (cit. on pp. 71, 76).

[246] M. A. Valencia-Sanchez, J. Liu, G. J. Hannon and R. Parker, 'Control of translation and mRNA degradation by miRNAs and siRNAs', *Genes & development*, 2006 (cit. on p. 14).

[247] M. Verma, J. Choi, K. A. Cottrell, Z. Lavagnino, E. N. Thomas *et al.*, 'A short translational ramp determines the efficiency of protein synthesis', *Nat. Commun.*, 2019 (cit. on p. 31).

[248] M. Vihinen, 'Relationship of protein flexibility to thermostability', *Protein Eng. Des. Sel.*, 1987 (cit. on pp. 19, 63, 129, 130).

[249] ——, 'Solubility of proteins', *ADMET and DMPK*, 2020 (cit. on p. 82).

[250] M. Vihinen, E. Torkkila and P. Riikonen, 'Accuracy of protein flexibility predictions', *Proteins*, 1994 (cit. on pp. 17, 18, 54, 56, 57, 63, 130).

[251] A. Villalobos, J. E. Ness, C. Gustafsson, J. Minshull and S. Govindarajan, 'Gene Designer: a synthetic biology tool for constructing artificial DNA segments', *BMC Bioinform.*, 2006 (cit. on pp. 14, 31).

[252] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nat. Methods*, 2020 (cit. on p. 79).

[253] D. Voges, M. Watzele, C. Nemetz, S. Wizemann and B. Buchberger, 'Analyzing and enhancing mRNA translational efficiency in an *Escherichia coli* in vitro expression system', *Biochem. Biophys. Res. Commun.*, 2004 (cit. on pp. 41, 43).

[254] G. S. Waldo, 'Genetic screens and directed evolution for protein solubility', *Curr. Opin. Chem. Biol.*, 2003 (cit. on pp. 53, 82).

[255] G. Walsh, 'Biopharmaceutical benchmarks 2014', *Nat. Biotechnol.*, 2014 (cit. on p. 1).

[256] S. van der Walt, S. C. Colbert and G. Varoquaux, 'The NumPy Array: A Structure for Efficient Numerical Computation', *Comput. Sci. Eng.*, 2011 (cit. on p. 68).

[257] M. Wang, C. J. Herrmann, M. Simonovic, D. Szklarczyk and C. von Mering, 'Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines', *Proteomics*, 2015 (cit. on p. 35).

[258] X. Wang, T. K. Das, S. K. Singh and S. Kumar, 'Potential aggregation prone regions in biotherapeutics: A survey of commercial monoclonal antibodies', *MAbs*, 2009 (cit. on p. 65).

[259] J. Warwicker, S. Charonis and R. A. Curtis, 'Lysine and arginine content of proteins: computational analysis suggests a new tool for solubility design', *Mol. Pharm.*, 2014 (cit. on p. 63).

[260] M. Waskom, O. Botvinnik, P. Hobson, J. B. Cole, Y. Halchenko, S. Hoyer, A. Miles, T. Augspurger, T. Yarkoni, T. Megies, L. P. Coelho, D. Wehner, cynddl, E. Ziegler, diego, Y. V. Zaytsev, T. Hoppe, S. Seabold, P. Cloud, M. Koskinen, K. Meyer, A. Qalieh and D. Allan, 'seaborn: v0.5.0 (November 2014)', 2014 (cit. on p. 68).

[261] M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, J. Ostblom, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Brunner,

T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, Brian and A. Qalieh, 'mwaskom/seaborn: v0.9.0 (July 2018)', 2018 (cit. on p. 51).

[262] M. Waskom, O. Botvinnik, J. Ostblom, S. Lukauskas, P. Hobson, MaozGelbart, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, C. Swain, A. Miles, T. Brunner, D. O'Kane, T. Yarkoni, M. L. Williams and C. Evans, *mwaskom/seaborn: v0.10.0 (January 2020)*, 2020 (cit. on p. 80).

[263] M. S. Waterman and T. H. Byers, 'A dynamic programming algorithm to find all solutions in a neighborhood of the optimum', *Math. Biosci.*, 1985 (cit. on p. 9).

[264] C. M. Whittington, A. T. Papenfuss, P. Bansal, A. M. Torres, E. S. W. Wong, J. E. Deakin, T. Graves, A. Alsop, K. Schatzkamer, C. Kremitzki, C. P. Ponting, P. Temple-Smith, W. C. Warren, P. W. Kuchel and K. Belov, 'Defensins and the convergent evolution of platypus and reptile venom genes', *Genome Res.*, 2008 (cit. on p. 75).

[265] D. L. Wilkinson and R. G. Harrison, 'Predicting the solubility of recombinant proteins in *Escherichia coli*', *Nat. Biotechnol.*, 1991 (cit. on pp. 15, 54, 61–64).

[266] E. S. W. Wong, M. C. Hardy, D. Wood, T. Bailey and G. F. King, 'SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula', *PLoS One*, 2013 (cit. on pp. 71, 83).

[267] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei *et al.*, 'A new coronavirus associated with human respiratory disease in China', *Nature*, 2020 (cit. on p. 65).

[268] J. Wu, S. G. Kim, K. Y. Kang, J.-G. Kim, S.-R. Park, R. Gupta, Y. H. Kim, Y. Wang and S. T. Kim, 'Overexpression of a Pathogenesis-Related Protein 10 Enhances Biotic and Abiotic Stress Tolerance in Rice', *Plant Pathol. J.*, 2016 (cit. on p. 76).

[269] X. D. Wu, B. Shang, R. F. Yang, H. Yu, Z. H. Ma, X. Shen, Y. Y. Ji, Y. Lin, Y. D. Wu, G. M. Lin, L. Tian, X. Q. Gan, S. Yang, W. H. Jiang, E. H. Dai, X. Y. Wang, H. L. Jiang, Y. H. Xie, X. L. Zhu, G. Pei, L. Li, J. R. Wu and B. Sun, 'The spike protein of severe acute respiratory syndrome (SARS) is cleaved in virus infected Vero-E6 cells', *Cell Res.*, 2004 (cit. on p. 136).

[270] Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann and F. H. Arnold, 'Machine learning-assisted directed protein evolution with combinatorial libraries', *Proc. Natl. Acad. Sci. U. S. A.*, 2019 (cit. on pp. 19, 54).

[271] H. Xiao, H. Pan, K. Liao, M. Yang and C. Huang, 'Snake Venom PLA, a Promising Target for Broad-Spectrum Antivenom Drug Development', *Biomed Res. Int.*, 2017 (cit. on p. 76).

[272] N. Xiao, D.-S. Cao, M.-F. Zhu and Q.-S. Xu, 'protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences', *Bioinformatics*, 2015 (cit. on pp. 55, 65).

[273] R. Xiao, S. Anderson, J. Aramini, R. Belote, W. A. Buchwald, C. Ciccosanti, K. Conover, J. K. Everett, K. Hamilton, Y. J. Huang, H. Janjua, M. Jiang, G. J. Kornhaber, D. Y. Lee, J. Y. Locke, L.-C. Ma, M. Maglaqui, L. Mao, S. Mitra, D. Patel, P. Rossi, S. Sahdev, S. Sharma, R. Shastry, G. V. T. Swapna, S. N. Tong, D. Wang, H. Wang, L. Zhao, G. T. Montelione and T. B. Acton, 'The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium', *J. Struct. Biol.*, 2010 (cit. on p. 54).

[274] K. K. Yang, Z. Wu and F. H. Arnold, 'Machine-learning-guided directed evolution for protein engineering', *Nat. Methods*, 2019 (cit. on pp. 19, 54).

[275] H. Yin, Y.-Z. Li and M.-L. Li, 'On the relation between residue flexibility and residue interactions in proteins', *Protein Pept. Lett.*, 2011 (cit. on pp. 19, 63).

[276] Z. Yuan, T. L. Bailey and R. D. Teasdale, 'Prediction of protein B-factor profiles', *Proteins*, 2005 (cit. on pp. 19, 63).

[277] M. Zamani, N. Nezafat, M. Negahdaripour, F. Dabbagh and Y. Ghasemi, '*In Silico* evaluation of different signal peptides for the secretory production of human growth hormone in *E. coli*', *Int. J. Pept. Res. Ther.*, 2015 (cit. on p. 83).

[278] S. Zayni, S. Damiati, S. Moreno-Flores, F. Amman, I. Hofacker and E.-K. Ehmoser, 'Enhancing the cell-free expression of native membrane proteins by in-silico optimization of the coding sequence – an experimental study of the human voltage-dependent anion channel', *bioRxiv*, 2018 (cit. on pp. 44, 83).

[279] M. Zuker, 'On finding all suboptimal foldings of an RNA molecule', *Science*, 1989 (cit. on p. 9).

[280] M. Zuker and P. Stiegler, 'Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information', *Nucleic Acids Res.*, 1981 (cit. on p. 7).