BACHELOR'S THESIS

Base-pair free-energy differences estimated from frameshifting efficiencies for SARS coronavirus

Freie Energie Unterschiede von Basenpaaren bestimmt durch Frameshifting Effizienzen für den SARS Coronavirus

prepared by

Lisa-Marie Heß

in Göttingen at the department of

THEORETICAL AND COMPUTATIONAL BIOPHYSICS

Submission date: 30.09.2022

Supervisors: Dr. Lars Bock and Sara Gabrielli

First Referee: Prof. Dr. Helmut Grubmüller

Second Referee: Prof. Dr. Stefan Klumpp

Acknowledgement

First and foremost, I want to thank Professor Helmut Grubmüller for the opportunity to write my bachelor thesis in his group. I also want to thank Professor Stefan Klumpp for being my second referee.

Furthermore, I want to thank my supervisors Dr. Lars Bock and Sara Gabrielli for their guidance, patience and all the helpful discussions throughout this project. I am very grateful that your door was always open for questions and for the detail with which you answered them.

Finally, I want to thank the whole group and especially Annke de Maeyer for the wonderful working environment.

Contents

1. Introduction

A virus is a small infectious agent affecting cells. It uses the infected host cell to replicate itself and is, in some cases, the cause for the emergence of diseases. To disrupt the processes essential for a virus survival, it helps to know how they are triggered and controlled. Viral proteins are important for both structural composition and function of a virus, rendering their synthesis a crucial process. Every information needed for protein synthesis is encoded in a sequence of nucleotides in the virus, which takes up space and resources. However, a virus is small (by a factor of 100 to 1000 smaller than human cells). Therefore, using as little storage space as necessary is vital [1]. A way in which a virus reduces the storage space is a process called programmed ribosomal frameshifting (PRF) [2]. This enables the storage of the information required for the synthesis of two different (poly-)proteins in the same sequence. How often each of the two proteins is produced is described by the frameshifting efficiency. In the host cell proteins are synthesized by macromolecular complexes called ribosomes. Also the synthesis of virus proteins and, thus, PRF take place on the ribosomes of the host cells.

A thermodynamic approach on how frameshifting is controlled has been published by Bock et al. [3]. Here -1 programmed ribosomal frameshifting efficiencies have been successfully explained by taking into account base-pair free-energy differences. Base-pairs form between nucleotides to decode the information for protein synthesis. As frameshifting results in the production of two different (poly-)proteins, the given synthesis information has to be decoded differently for each protein. Therefore, different base-pairs form. The difference in free-energy between two base-pairs is called the base-pair free-energy difference. To determine the free-energy differences, Bayes' theorem was applied. In this previously published study, the model has been tested on 85 sequence variants (64 sequences *in vitro*, 21 sequences *in vivo*) of the dnaX frameshifting element. As dnaX is part of the *Escherichia coli* mRNA, the ribosomes involved in the frameshifting process where bacterial ribosomes.

A large data set of frameshifting efficiencies was recently published by Mikl et al. [4], and includes more than 13000 sequences from different viruses and human mRNA. This data set enables testing the thermodynamic frameshifting model under different conditions. The new set of frameshifting efficiencies was obtained in the *in vivo* environment of human cells and, therefore, the ribosomes involved are human ribosomes. An *in vivo* environment is closer to the physiological conditions of virus replication, but also more complex than an *in vitro* environment due to the presence of additional, and possibly interacting, elements. Thus it is of interest to test if the model of Bock et al., mostly tested *in vitro*, still applies to sufficiently predict frameshifting efficiencies *in vivo* [5]. Additionally, even though their cores are similar, bacterial and human ribosomes display significant differences (for example in size). These differences motivate to test the method for human ribosomes as well [6]. The third difference with respect to Bock et al.'s work is the usage of mRNA sequences from severe acute respiratory

syndrome coronavirus (SARS-CoV) to test the model.
SARS-CoV is a human coronavirus which can lead to severe symptoms similar to pneumonia. During a SARS-CoV outbreak in 2002 about 8000 people were infected by the virus in half a year and, of those, 800 died [7]. Furthermore, the portion of the SARS-CoV sequence surrounding the site of frameshifting is very similar to that of SARS-CoV-2. The outbreak of this variant in 2019 caused a global pandemic with over 3.5 million deaths (outbreak until march 2022)[8–10]. In SARS-CoV, frameshifting results in the production of polyproteins, which are essential to mediate transcription and replication of the virus (pp1a, pp1ab) [11, 12]. Therefore, understanding the mechanism of PRF in SARS-CoV might help to develop medical interventions that inhibit viral replication.
In this bachelor thesis, I tested if the thermodynamic model by Bock et al. is applicable to determine and predict frameshifting efficiencies for variations in the SARS-CoV slippery sequence. For this, base-pair free-energy differences are determined based on the data made available by Mikl et al. [4] from measurements in an *in vivo* environment with human ribosomes. For one of the sequence variants, the frameshifting efficiency is predicted based on the determined free-energy differences.

2. Biological Background

Starting from an introduction on the relevant polynucleotide structures and the central dogma, the following sections will then focus on ribosomal translation and frameshifting. Finally, an overview on the SARS-CoV and its properties will be provided.

2.1. Genetic information storage and the central dogma

Genetic information in cells is stored in DNA (deoxyribonucleic acid). DNA has a double helix structure with high stability and is made of a sequence of nucleotides (nt). A nucleotide always consists of a five-carbon sugar bound to a phosphate group and an organic nitrogenous base (nucleobase). According to the number of rings in their structure, the nucleobases are classified into purines (two rings) or pyrimidines (one ring). Among the canonical nucleobases present in DNA, adenine (A) and guanine (G) are purines, while cytosine (C) and thymine (T) are pyrimidines. The sugar in DNA is deoxyribose. [13]. From now on the nucleotides are labeled by their organic base.
In order to protect the genetic information, and thus the DNA in the nucleus, the information flow in cells follows the central dogma of molecular biology. The direction of information flow in cells occurs from DNA to mRNA (messenger ribonucleic acid), to proteins made of chains of amino acids. mRNA is the polynucleotide molecule where the genetic information is transferred to during transcription. mRNA molecules are single stranded nucleotide sequences. As

an ribunucleic acid, mRNA contains the pyrimidine uracil (U) instead of thymine [13].
Translation is the step in which an amino-acid chain is formed based on the sequence of nucleotides in a mRNA molecule, and a protein is assembled [13]. For RNA-viruses the genetic material stored in RNA, therefore, once they have infected the host cell, the central dogma has two additional steps: the RNA replication and the reverse transcription [6] (figure 1).
In ribosomes, mRNA molecules are translated into proteins where the sequence of nucleotides encodes the sequence of amino acids in the protein. This genetic information is generally read in sequences of three nucleotides (codons). Each of the codons encodes a specific amino acid. The codons are decoded by sequences of three nucleotides called anticodons. Due to the amount of different components in the code (four different bases) 64 different codons and anticodons are possible [13].
Like DNA and mRNA, every polynucleotide has a chemical orientation. The end containing a phosphate or hydroxyl group at 5' carbon of the terminal sugar is the 5' end. The end having a hydroxyl group at the 3' carbon is called 3' end. Polynucleotides are read from the 5' to the 3' end [13].
Polynucleotides functioning like keys to decode the sequence are called tRNAs (transfer ribonucleic acids). Different tRNAs share a similar composition [14], consisting of a single strand of RNA (70-80 nt) folded into a 3D L-shape [13]. In 2D it is visible that tRNAs have four main arms: one of those contains the 3' and 5' ends, while the others are loops. The loop in the center is called anticodon arm. It includes the anticodon sequence, which forms base pairs with the corresponding codon in the mRNA and therefore decodes the mRNA sequence [6]. All tRNAs contain a CCA nucleotide sequence at the 3' end to which a specific amino acid can bind. This arm is thus called amino acid arm [6]. The amino acids are linked to the corresponding tRNA by aminoacyl tRNA synthetases [14]. The resulting complex is called aminoacyl-tRNA [13]. Although there are 64 possible codons, there are only 20 amino acids. As a result in most cases different codons encode the same amino-acid [13].
If codons and anticodons match during translation, base-pair interactions are formed between the bases in the codon of the mRNA and the bases in the anticodon of the tRNA. The most common base-pairs are A-U and G-C, and are called Watson-Crick base pairs. The first and second base pair of the codon anticodon bond are usually of the Watson-Crick type. For the third base pair nonstandard pairing (Wobble base pair) is possible. Examples of Wobble pairs are G-U or I-U, I-A and I-C, where the I stands for inosine, which is derived from adenine [13].

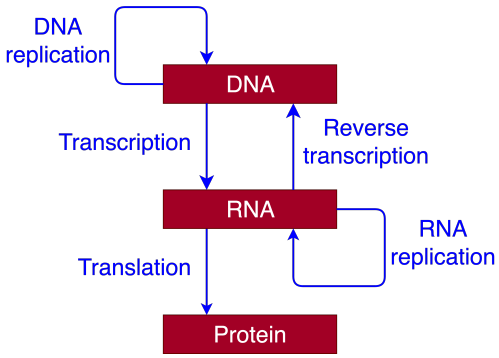

Figure 1: The central dogma for a RNA-virus [6].

2.2. Translation in ribosomes

During translation the mRNA is read in order to form polypeptides, i.e. the proteins [13]. This synthesis process takes place in the ribosome, which is a biomolecular complex made of rRNA (ribosomal ribonucleic acid) and proteins [14] (figure 2). The ribosome consists of two subunits: the large ribosomal subunit (LSU) and the small ribosomal subunit (SSU). Generally the LSU is important for the coordination of the translation process and the formation of the peptide bond between amino acids to form a polypeptide. Decoding, i.e. the pairing between codons and anticodons takes place in the SSU [15]. In eukaryotic cells, the LSU has a size of 60 S and the SSU of 40 S [14]. The unit S stands for Svedberg, which gives information on the size and is based on a sedimentation rate [13]. Furthermore, the ribosome contains three binding sites for the tRNAs: A site, P site and E site (aminoacyl, peptidyl and exit site). Before translation starts, the two subunits of the ribosome are separate [15].

The protein synthesis itself is subdivided into three steps: initiation, elongation and termination [14]. During the initiation step a methionyl aminoacyl-tRNA binds to the AUG-start codon of the mRNA and to the P site of the SSU [13, 14]. In eukaryotic cells this is controlled by eukaryotic initiation factors. Afterwards, the LSU binds to the complex [14].

At this stage elongation [13] begins with the help of proteins called elongation factors. The binding of the aminoacyl-tRNA to the matching codon in the A site leads to a conformational change of the ribosome [14]. Afterward, a peptide bond between the amino acid in the P site and the one in the A site is formed resulting in a peptidyl-chain connected to the tRNA in the A site and an uncharged tRNA in the P site [14]. This process is catalysed by the large subunit. During translocation, the ribosome moves along the mRNA by three nts in 5'-3' direction, resulting in the bound tRNAs moving from P and A site to E and P site [14]. The tRNA in the E site is then released and a new matching aminoacyl tRNA can bind to the A site.

The cycle is repeated until a stop codon is reached. Thanks to the release factors (proteins), the ribosome recognizes this codon and terminates the protein synthesis [14].

The sequence of codons translated throughout this process from start to stop codon is the read-

ing frame [13].

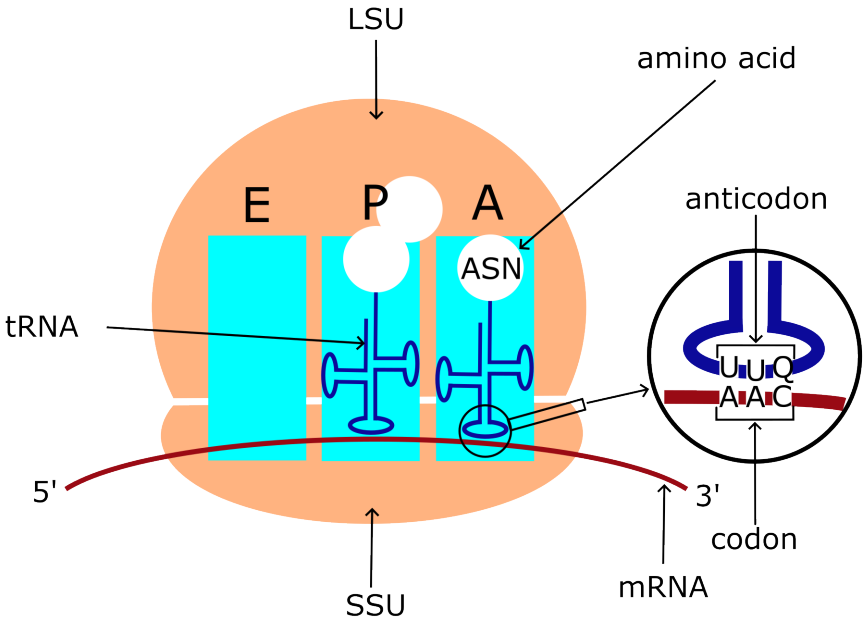

Figure 2: Schematic 2D graphic of translation in a ribosome for an AAC example codon in the A site.

2.3. Ribosomal frameshifting

Ribosomal frameshifting is an event during translation during which the reading frame is altered due to a shift of the tRNAs in the ribosome relative to the mRNA by a number of nt different from 3 [2]. Spontaneous frameshifting which results in erroneous proteins, only happens with a very low probability due to control mechanisms. In contrast, programmed frameshifting has evolved as a mechanism to change the reading frame at a specific position with a high probability (frameshifting efficiency) [16]. During -1 programmed frameshifting, the reading frame is shifted by one nt towards the 5' end of the mRNA [17].

Viruses use frameshifting to maximize the amount of information which can be stored in a sequence and to control the relative ratio of the two products. As they have little space available, frameshifting is extremely beneficial to the viruses [2]. However it is crucial for the virus that programmed frameshifting occurs with the correct probability, so that the required amount of each protein is produced [17].

Frameshifting is regulated and promoted by specific structures in the mRNA (cis-regulatory elements) as well as other factors (trans-acting factors). The main cis-regulatory elements are the slippery sequence, the secondary structure of the downstream sequence and sometimes the

upstream structure [16, 18].

The slippery sequence is a nucleotide sequence often of the structure X XXY YYZ, where X,Y, and Z stand for any of the nucleotides. With the use of Wobble base pairs this sequence enables base-pairing in 0 frame (before -1 programmed frameshift) as well as -1 frame (after -1 programmed frameshift) [16].

The slippage is generally enabled when the secondary structure (for example a stem-loop or pseudoknot), located a few nt downstream of the slippery site, stalls the translation process [16, 19].

It is likely that the frameshift occurs while the tRNAs move from the A and P sites to the P and E sites during translocation [3, 16].

2.4. Frameshifting in SARS-CoV

As previously mentioned, frameshifting is important for viruses to expand and fine-tune the storage and expression of proteins. This also applies to SARS-CoV [9], although it is an unusually large RNA virus (the genetic code consists out of approximately 30000 nt [12]). In general coronaviruses have the ability to rapidly adapt due to the unique open reading frames (orf) toward the 3' end of the sequences [7]. Therefore, they are slightly different from one another. However, a similarity they share is the importance of frameshifting for the formation of a replication/transcription complex (RTC) [9, 11].

The repilicase-transcriptase proteins are part of polyproteins translated from the two open reading frames orf1a (0 frame) and orf1ab (-1 frame) [9]. The distribution of the two polyproteins is controlled by the frameshifting efficiency, which has been shown in previous studies to be 18-40 % for WT coronavirus sequences [20].

This efficiency is, in turn, controlled by specific cis-acting factors. These factors are mainly the slippery sequence U_UUA_AAC [9] (figure 4), and a mRNA pseudoknot consisting of three stem loops as downstream secondary structure (figure 3). The pseudoknot of SARS-CoV follows a spacer region positioned after the slippery sequence [8, 20].This downstream secondary structure generates a back-pull which leads to stalling of the translation process, and thus enables slip-

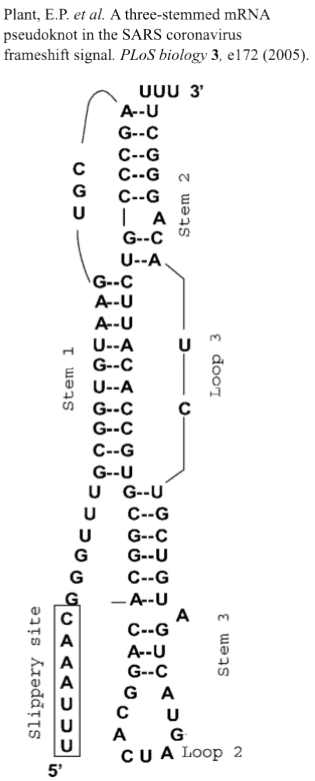

Figure 3: Secondary structure representation of the SARS-CoV cis-acting factors. Figure adapted from [8].

page on the slippery sequence [9]. As an evidence of the crucial role played by frameshifting, it has been shown that even single mutations in the range of the slippery sequence and pseudoknot can prevent viral replication [9].

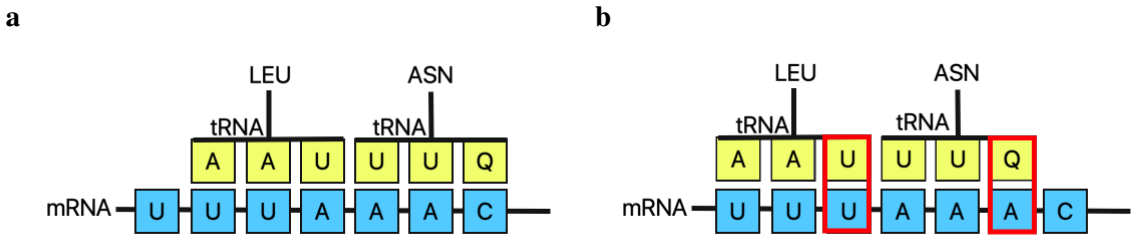

Figure 4: a SARS-CoV WT slippery sequence in the 0 frame. **b** SARS-CoV WT slippery sequence in the -1 frame.

3. Thermodynamic Background

3.1. Free energy on a macromolecular scale

The thermodynamic background and the equations of this section are based on Nelson et al. [21]. Free energy is defined as the quantity that is minimized when a system a is in equilibrium. Depending on whether the volume V or the pressure p of the system is fixed, this quantity is either Helmholtz free energy F_a (fixed volume)

$$F_a = E_a - TS_a, \tag{1}$$

or Gibbs free energy G_a (fixed pressure)

$$G_a = E_a + pV_a - TS_a, \tag{2}$$

where S_a is the entropy, E_a is the total energy, T is the temperature and $E_a + pV_a$ is the enthalpy. On a molecular scale, the free energy is defined as

$$G_a = \langle E_a \rangle - TS_a, \tag{3}$$

where the energy E_a is averaged for all possible states $\langle E_a \rangle = \sum_j P_j E_j$ and the entropy is $S_a = -k_B \sum_j P_j \ln(P_j)$.

Consider a complex macromolecular system that can occupy two possible states, indicated by the index $l \in 1, 2$. If each state consists of N_l substates, the total number of substates is

$N = N_1 + N_2$. The probability for a system to be in one of the substates j follows the Boltzmann distribution

$$P_j = \frac{1}{Z} e^{-E_j/k_\mathrm{B}T}, \tag{4}$$

where the normalising factor Z is the partition function

$$Z = \sum_{j=1}^{N} e^{-E_j/k_\mathrm{B}T}. \tag{5}$$

The probability for a system a to be in state $l \in 1, 2$ consisting of N_l substates therefore is

$$P_l = \frac{1}{Z} \sum_{j=1}^{N_l} e^{-\frac{E_j}{k_\mathrm{B}T}}. \tag{6}$$

The free energy of a system a in state $l \in 1, 2$ can, based on equation (3), be deducted to

$$\begin{aligned} G_{a,l} &= \langle E_a \rangle_l - T S_{a,l} \\ &= \sum_{j}^{N_l} P_{j,l} E_j + k_B T \sum_{j}^{N_l} P_{j,l} \ln(P_{j,l}) \\ &= -k_\mathrm{B} T \ln Z_l, \end{aligned} \tag{7}$$

where $Z_l = \sum_{j=1}^{N_l} e^{-E_j/k_\mathrm{B}T}$ and $P_{j,l}$ is the probability for the system to be in the substate j, when it is in state l. For a more detailed deduction see equation (54) in the appendix. Based on (7), the free energy difference between the two states is $\Delta G = G_{a,1} - G_{a,2} = -k_\mathrm{B}T \ln(Z_1/Z_2)$. Using equation (6) it can be shown that the ratio of probabilities equals the ratios of partition functions $P_1/P_2 = Z_1/Z_2$. This results in

$$\frac{P_1}{P_2} = e^{-\Delta G/k_\mathrm{B}T}. \tag{8}$$

Solving for P_1 gives the probability for a system to be in state 1

$$P_1 = \frac{e^{-\Delta G/k_\mathrm{B}T}}{e^{-\Delta G/k_\mathrm{B}T} + 1}. \tag{9}$$

3.2. Relation between free-energy differences and frameshifting efficiencies

The frameshifting efficiency is defined as the probability for a sequence to be translated in the -1 reading-frame. The model proposed by Bock et al. assumes that, if the translation process

occurs slowly enough, as in the case of stalling induced by a downstream mRNA secondary structure element, the system can be approximated as being in equilibrium. Therefore, the free-energy difference between the -1 and 0 frame is the only parameter determining the frameshifting efficiency [3]. The free-energy difference results from different interactions between mRNA and tRNA due to a different codon-anticodon match when the frameshifting takes place. This assumption enables to classify this complex problem into two overall states. One state corresponds to the 0 frame and the other to the -1 frame. As described in section 3.1, this results in the following frameshifting efficiency (FS) for a given sequence:

$$\mathrm{FS} = \frac{e^{-\Delta G/k_\mathrm{B}T}}{e^{-\Delta G/k_\mathrm{B}T} + 1}, \tag{10}$$

where the difference in free energy is $\Delta G = G_{-1} - G_0$. Accordingly, a negative free-energy difference results in a frameshifting efficiency above 50% [3]. The relation between the frameshifting efficiency and the free-energy difference is visualised in figure 5.

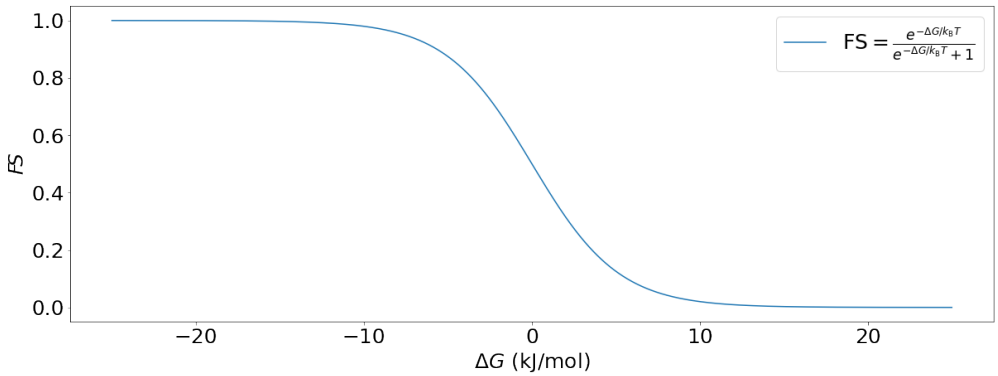

Figure 5: Frameshifting efficiency as a function of the free-energy difference between 0 and -1 frame, according to (10) for T=310.15 K.

The different interactions between mRNA and tRNA due to a different codon-anticodon match can be further split down into different interaction between bases. It is assumed that these tRNA-mRNA base-pair free-energy differences are additive and result in the overall free-energy difference between 0 and -1 frame. This assumption leads to

$$\Delta G_j = \sum_{r=1}^{m} \Delta G_{\mathrm{bp},r}, \tag{11}$$

where m is the overall number of base-pair changes occurring in a frameshifting event with sequence j and $\Delta G_{\mathrm{bp},r}$ are the free-energy differences corresponding to specific base-pair changes for the sequence j[3]. The base-pair free-energy difference depends on the position of the bases in the codons and ribosome. As a result, changes between the same base pairs but at different

positions are counted as different base-pair changes [3].

4. Bayesian statistics

In Bayesian statistics a statistical model is created based on the Bayes' theorem to analyse experimental data. This model can be fitted to determine unknown parameters by using a Markov Chain Monte Carlo algorithm, like the Metropolis-Algorithm [22].

4.1. Bayes' Theorem

A conditional probability is defined as

$$P(C_i|T) = \frac{P(T \cap C_i)}{P(T)}, \tag{12}$$

where C_i, with $i \in I$, and T are two events and the probability of C_i, $P(C_i|T)$, depends on the occurrence of T [23]. I is the ensemble of all possible outcomes for C. According to Bayes' theorem this can also be written as [24]

$$\mathrm{P}(C_i|T) = \frac{\mathrm{P}(T|C_i) \cdot \mathrm{P}(C_i)}{\sum_{\gamma \in I} \mathrm{P}(T|C_\gamma) \cdot \mathrm{P}(C_\gamma)}, \tag{13}$$

where $\mathrm{P}(T|C_i)$ is called the likelihood, $\mathrm{P}(C_i)$ is the prior, and the evidence is $\mathrm{P}(T)$ [25]. Here the evidence is defined as

$$\mathrm{P}(T) = \sum_{\gamma \in I} \mathrm{P}(T|C_\gamma) \cdot \mathrm{P}(C_\gamma). \tag{14}$$

In the Bayesian setting, the conditional probability $\mathrm{P}(C_i|T)$ is called the posterior probability. When this theorem is applied in Bayesian statistics, C_i is a parameter set and T the experimental data. Therefore, the probability to have a set of parameters when given the experimental data $\mathrm{P}(C_i|T)$ is described [22].

4.2. The Metropolis-Algorithm

In the framework of Bayesian statistics, to determine the parameters of ones statistical model, one samples the distribution function of the posterior and thereby obtains the parameter distributions. This can be done using the Metropolis-Algorithm [22].
In a Markov chain, the probability to move from one state to another is independent of all previous states. The Metropolis-Algorithm is a Markov chain Monte Carlo algorithm, which generates a Markov chain that converges to a given distribution. The Metropolis-Algorithm

compares probabilities to decide on the chain links [26].
In the beginning of the algorithm a start value C_0 is given, as well as a density function f and a symmetric proposal distribution $q(Y|C)$. The latter describes how likely it is to change states from C to Y without knowledge of any previous state. First, a value Y is randomly selected given the proposal distribution $q(Y|C_0)$. With this value the acceptance probability α is calculated as [26]

$$\alpha(C, Y) = \min\left\{\frac{f(Y)}{f(C)}, 1\right\}. \tag{15}$$

Then a random value u is generated using the uniform distribution in range [0,1] [27]. The next element of the chain C_{i+1} given the previous chain element C_j is determined with [26]

$$C_{i+1} = \begin{cases} Y, & u \leq \alpha(C_i, Y) \\ C_i, & \text{otherwise} \end{cases}. \tag{16}$$

This algorithm converges, however the right choice of proposal density is crucial for efficiency [27].

5. Frameshifting efficiencies from high-throughput experiments

In a previous publication, Mikl et al. provided measurements related to frameshifting potentials for more than 13 000 different mRNA sequences associated with PRF. The frameshifting potential was tested for sections of different viral and human mRNA sequences which are involved in PRF. The experiments were run on both wild type and variants of the sequences. The measurements were performed by monitoring the expression of GFP (green fluorescent protein) in the *in vivo* environment of human cells. The GFP-encoding mRNA sequence was introduced downstream of the tested sequence sections. In particular, it was positioned in the -1 frame so that GFP would be produced only if a frameshifting event occurred. The tested sections were 162 nt long and consisted of the slippery sequence plus its immediate upstream and downstream region. The data were collected for both wild type and variants of each sequence [4].

5.1. GFP expression measurement

The experimental method used by Mikl et al. to obtain the GFP expression percentages is fluorescence activated cell sorting (FACS)[4].
FACS is a high-throughput measuring technique to sort single cells according to their fluores-

cence [4, 28]. During a FACS experiment, droplets containing single cells pass through a laser beam. Depending on their fluorescence intensity, the droplets are charged and subsequently sorted into different bins by an electric field [28].
The measured fluorescence intensity is related to frameshifting because of the inclusion of a GFP encoding sequence in the -1 reading frame of the tested sequences and because GFP emits green fluorescence. When frameshifting occurs, the sequence encoding the GFP protein is in frame and, hence, the protein is synthesized. A higher frameshifting efficiency results in a higher production of GFP proteins and, thus, in a brighter fluorescence signal [4].
In the experiment performed by Mikl et al., the cells were sorted, according to their fluorescence intensity, into 16 bins. The green fluorescence of each bin was calculated as the median of the log2 green fluorescence intensity of all the cells sorted into the bin. If the resulting distribution was not considered likely, this values were smoothed . For example, bins were set to 0 if their neighbouring values were zero. The overall green fluorescence reported for a given sequence was obtained as the weighted average of the intensities of the 16 bins [4].
The measured green fluorescence values were then scaled in such a way that the smallest observed green fluorescence value of the whole data set is interpreted as 0 % GFP expression and the largest observed value as 100 % GFP expression. As a consequence, the given percentages of green fluorescence do not necessarily match the actual percentages of occurring PRF events [4].

5.2. Background fluorescence measurement

The natural fluorescence of the cells introduces background noise in the measurements of the GFP expression. Mikl et al. measured this background fluorescence by including an early stop codon upstream of the slippery sequence, thereby preventing any expression of GFP, and measuring the resulting green fluorescence. The background noise was measured for all of the tested sequences. Based on these results, a threshold of 1.3 % GFP expression was introduced. Smaller GFP expressions were attributed to background fluorescence [4].

6. Methods

6.1. Selection of GFP expression measurement points

As mentioned in section 5, the green fluorescence values used as input data for this work were selected from the more extensive data set published by Mikl et al.[4]. In this section, the criteria followed in the selection of the input data points are reported.
One of the reasons why not all of the measurements from Mikl et al.[4] could be used is that

some of the reported GFP expression were calculated by the authors from multimodal fluorescence distributions. Multimodal distributions indicate problems in the measurements or the interference of other unaccounted factors in the cell, as it is not to be expected that there are no or several possible frameshifting efficiencies to be assigned to a sequence. Hence, in this analysis, only data points that were obtained from an unimodal distribution were considered.
In some cases where multiple data points were provided, the range of provided GFP expressions is very wide, for example 91.36 % for the West Nile Virus. As the presence of outliers misrepresents the actual distribution of the measurements, a criteria to identify them was introduced by defining the interquartile range I. Here, I is defined as the difference between the upper quartile Q_3 and the lower quartile Q_1 of all GFP expressions values given for one sequence [29]. Thus, for data points obtained from the same sequence, all values outside the interval $[Q_1 - 1.5 \cdot I, Q_3 + 1.5 \cdot I]$ were excluded. However, if for a sequence less than 4 data points were provided, none of them was excluded due to the lack of data to determine reasonable quartiles.
This resulted in a subset of the original database containing data points for 15 different virus and human mRNA WT sequences, 8 sequences of HIV HXB2 variants (including the WT) and 53 sequences of SARS-CoV variants (including the WT).
For the 53 slippery sequences of SARS-CoV variants, the anticondons had to be determined in order to later determine base-pair changes. To this end, the human tRNAs that are linked to the each aminoacid were, as a first step, identified using the database GtRNAdb [30]. Afterwards, the database tRNAdb 2009 [31] was used to search for the presence of any modified nts in the anticodon of the selected tRNAs. From the resulting set of tRNAs, the anticodons involved in frameshifting were selected based on the codons of the slippery sequence in the 0 frame. When doing so, the most likely base-pairing possible was assumed, i.e. Watson-Crick base-pairs were chosen over Wobble base-pairs. The codon-anticodon pairs that resulted relevant for the 53 analysed SARS-CoV slippery sequences are printed in the appendix in table 1 .
When frameshifting events take place, the 53 SARS-CoV variants result in the 40 different base-pair changes listed in table 2. The base-pair changes are labeled as for example P3: AU $\rightarrow$ GU (figure 6). Here, P is the site in the ribosome where the codon is located, while 3 is the position of the base in the codon, counted along the 5'-3' direction of the mRNA. A is the base of the mRNA at the position P3 in the 0 frame, G is the mRNA base that occupies the same position in -1 frame. U is the base on the tRNA at the position P3.

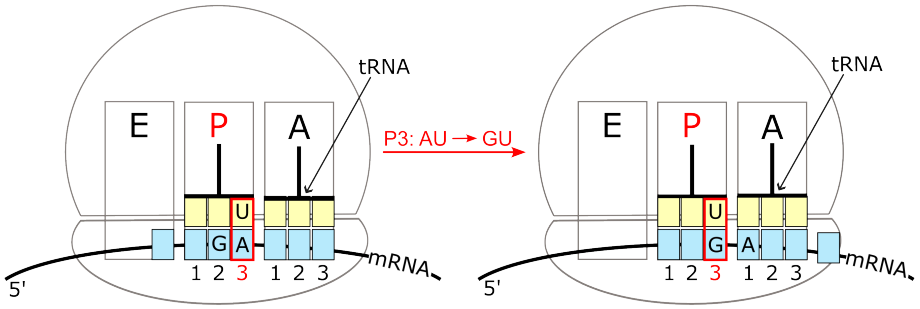

Figure 6: Visualisation of the base-pair change P3: AU $\rightarrow$ GU.

6.2. Metropolis-Algorithm sampling parameters based on measured GFP expressions

The first goal of this project was to sample distributions for parameters, based on the probability of a parameter combination to cause the experimentally provided GFP expression measurements. This was achieved by using the Metropolis-Algorithm (section 4.2) to compare the probability of different parameter sets of generating the experimental data when introduced in the tested model. For this, a set of experimentally determined GFP expression measurements T_{all}, a set of starting values for the parameters C_0, proposal distributions q_k for each parameter C^k in the set, and a density function f are given.

Since a set of parameters was to be optimised instead of only a single parameter, the parameters were varied, as described in section 4.2, iteratively in each iteration step. Therefore, in each iteration step, a parameter specific proposal function q_k was used to propose a new value for the parameter. Depending on the acceptance probability, the proposed value was either rejected or accepted. Gaussian distributions were used as proposal functions. Their mean was equal to the current value of the parameter value and their standard deviation was parameter specific (see appendix tables 3, 5). The density function used to calculate the acceptance probability was derived from the posterior probability $P(C_i|T_{\text{all}})$, which was calculated using the Bayes' theorem introduced in section 4.1.

For the whole data set, T_{all} contains the experimental measurements of the **T**otal amount of green fluorescence (background noise + signal)

$$\mathrm{P}(C_i|T_{\text{all}}) = \frac{\mathrm{P}(T_{\text{all}}|C_i) \cdot \mathrm{P}(C_i)}{\mathrm{P}(T_{\text{all}})}. \tag{17}$$

Here C_i with $i \in I$ is a vector of all varied parameters. I is the index-set over all varied parameter combinations. As the parameters are continuous rather than discrete, from now on probability densities will be considered instead of probabilities. Because $\mathrm{P}(T_{\text{all}})$ is independent from the parameters C_i, it is also possible to compare $\mathrm{P}(T_{\text{all}}|C_i) \cdot \mathrm{P}(C_i)$ of two steps in the

algorithm instead of $\mathrm{P}(C_i|T_{\mathrm{all}})$.

$$\mathrm{P}(C_i|T_{\mathrm{all}}) \propto \mathrm{P}(T_{\mathrm{all}}|C_i) \cdot \mathrm{P}(C_i). \tag{18}$$

The prior $\mathrm{P}(C_i)$ is defined based on the previous knowledge of the varied parameters. It is specific to the parameter set and, as a consequence, introduced when the parameter set is defined (sections 6.4, 6.5, 6.6).
The likelihood $\mathrm{P}(T_{\mathrm{all}}|C_i)$ to obtain all GFP expression measurements T_{all} given a parameter set C_i, is characterized by the density function $f_{T_j}(t_j)$ described below. This function describes the probability density for one specific sequence j to measure a GFP expression value t_j.
As previously stated, the ***T***otal fluorescence in the experimental data is a combination of ***B***ackground fluorescence B and GFP fluorescence ***S***ignal S (section 5.2). Therefore, for one sequence j, the total fluorescence T_j can be written as $T_j = B + S_j$, where we can assume that B and S_j are independent variables.
For two independent variables B, S_j the following statement on the combined probability density function is true: [24]

$$f_{(B,S_j)}(b, s) = f_B(b) \cdot f_{S_j}(s). \tag{19}$$

The probability density function for T_j is given by [24]

$$f_{T_j}(t_j) = \int_{\mathbb{R}} f_B(b) f_{S_j}(t_j - b)\, \mathrm{d}b. \tag{20}$$

For both fluorescences a Gaussian distribution is assumed. The resulting integral, given the standard deviations σ_B and σ_{S_j}, the mean of the background noise μ_B, and the mean of the GFP signal μ_{S_j} is

$$f_{T_j}(t_j) = \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_B\sigma_{S_j}} \exp\left(-\frac{(b-\mu_b)^2}{2\sigma_B^2}\right) \exp\left(-\frac{(t_j - b - \mu_{S_j})^2}{2\sigma_{S_j}^2}\right) \mathrm{d}b. \tag{21}$$

The integral is solved using the following relation, which is based on the result of a Gaussian integral [32] and the completion of the sum in the exponent

$$\int_{-\infty}^{\infty} e^{-ax^2+bx+c}\, \mathrm{d}x = \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{4a} + c\right). \tag{22}$$

Therefore, the probability density function for the overall fluorescence T_j of one sequence is

$$f_{T_j}(t_j) = \frac{1}{\sqrt{2\pi \cdot (\sigma_B^2 + \sigma_{S_j}^2)}} \exp\left(-\frac{(t_j - \mu_B - \mu_{S_j})^2}{2(\sigma_B^2 + \sigma_{S_j}^2)}\right). \tag{23}$$

All N sequences, with $L(j)$ measured values for sequence j, are combined in T_{all} and have to be included in the likelihood for a parameter set. As the measurements were performed independently, the final likelihood is a product of the density distributions for all considered sequences evaluated for the measured data points $t_{j,l}$ with $j \in \{1, \ldots, N\}$, $l \in \{1, \ldots, L(j)\}$ (24).

$$\mathrm{P}(T_{\text{all}}|C_i) = \prod_{j=1}^{N} \prod_{l=1}^{L(j)} \mathrm{f}_{T_{j,l}}(t_{j,l}) = \prod_{j=1}^{N} \prod_{l=1}^{L(j)} \frac{1}{\sqrt{2\pi \cdot (\sigma_B^2 + \sigma_{S_j}^2)}} \exp\left(-\frac{(t_{j,l} - \mu_B - \mu_{S_j})^2}{2(\sigma_B^2 + \sigma_{S_j}^2)}\right) \tag{24}$$

Extremely large and small numbers for the likelihood may lead to computational instabilities, which is why the logarithm of equation (24) was used in the calculations.

6.3. Determining mean and standard deviation of the background noise μ_{B} and σ_{B}

As mentioned in section 5.2, early stop codon experiments provided, for each sequence, measurements of the background fluorescence. The distribution f_B of all of these measurements was obtained by plotting them in a histogram. This distribution, for the most part, resembles a Gaussian distribution. However, in addition, it has a few values that are spread out on the right-hand side of the distribution. (section 7.1). As the spread-out values were too few to draw any conclusion on their contribution to the distribution of the background signal, values higher than 99 % of the measured GFP expressions were cut off. The mean μ_B and the standard deviation σ_B of the Gaussian function fitting the remaining data were, then, calculated.

6.4. Determining the standard deviation of the GFP signal σ_{S_j}

As shown in equation 24, the likelihood depends on the standard deviation of the GFP signal σ_{S_j}. Deviations might be caused by coincidental frameshifting events, a wrongly constructed sequence or as a consequence of the measurement technique. However, from the database, it was not possible to determine σ_{S_j} for all the considered sequences, as, for some of them, not enough individual measurements were gathered to determine a standard deviation. As a consequence, σ_{S_j} had to be considered in the Metropolis-algorithm either as an independent parameter, or as a function of the GFP signal mean μ_{S_j}. In order to decide how to proceed, I tested if and how the standard deviation in the GFP signal σ_{S_j} was dependent on the mean of the

% GFP μ_{S_j}. For this purpose, a subset containing only WT sequences of different viruses and human mRNA was analysed. The reason for choosing this subset is that only for WT sequences enough measurements (8 or more) were reported to estimate the standard deviations. For every sequence j, the set of measured data points is $\mathbf{t}_\mathrm{j}$.
As a first step, the Bayesian algorithm introduced in section 6.2 was applied on data sets belonging to each WT sequence separately to maximize the probability $\mathrm{P}(\mathbf{t}_\mathrm{j}|\sigma_{S_j},\mu_{S_j})\cdot\mathrm{P}(\sigma_{S_j},\mu_{S_j})$ and obtain, for each sequence j, distributions of the parameters σ_{S_j} and μ_{S_j}. The likelihood for each sequence j, $\mathrm{P}(\mathbf{t}_\mathrm{j}|\sigma_{S_j},\mu_{S_j})$, was defined by equation (24) as

$$\mathrm{P}(\mathbf{t}_\mathrm{j}|\sigma_{S_j},\mu_{S_j}) = \prod_{l=1}^{L(j)} \frac{1}{\sqrt{2\pi\cdot(\sigma_B^2+\sigma_{S_j}^2)}} \exp\left(-\frac{(t_{j,l}-\mu_B-\mu_{S_j})^2}{2(\sigma_B^2+\sigma_{S_j}^2)}\right). \tag{25}$$

σ_B and μ_B were replaced by the values computed in section 6.3. The prior $\mathrm{P}(\sigma_{S_j},\mu_{S_j})$ was chosen to prevent negative values for standard deviations or expression percentages as

$$P(\sigma_{S_j},\mu_{S_j}) = \begin{cases} 0, & \sigma_{S_j} < 0\,\% \text{ or } \mu_{S_j} < 0\,\% \\ 1, & \text{otherwise} \end{cases}. \tag{26}$$

The starting parameters C_0 and the standard deviations of the proposal distributions are displayed in table 3. For each WT sequence 100000 iteration steps were executed. The algorithm was run twice to check for convergence and the parameters converged in less than 50000 steps. The mean and the standard deviation of σ_{S_j} and μ_{S_j} were calculated from the last 50000 steps. Noticeably, the obtained values for σ_{S_j} and μ_{S_j} were correlated, showing a trend of increased σ_{S_j} with increased μ_{S_j}. From here on, as an approximation, a linear relation between the two parameters is assumed (section 7.2):

$$\sigma_{S_j} = m_\sigma \cdot \mu_{S_j}. \tag{27}$$

A value for m_σ was first determined, as a reference for section 6.5, with an orthogonal distance regression based on the signal deviations and means for the WT-sequences. This was done with the python package scipy.odr. However, when running the Metropolis-algorithm over the complete data set, m_σ was introduced as a parameter to obtain the most likely distribution of its values (see section 6.6).

6.5. The relation between the measured GFP expression and the frameshifting efficiency

Based on the data provided by Mikl et al. [4] on the GFP expression measurement technique, it is not clear how the GFP expressions are related to the frameshifting efficiency. One of the uncertainties stems from the fact that the GFP expression percentages were determined based on the log2 of the measured green fluorescence. Additionally, the provided percentages of GFP expression were scaled by the maximal and minimal fluorescence intensities measured in the experiments (section 5). This does not necessarily mean that a GFP expression percentage of 100% corresponds to a frameshifting efficiency of 1. Furthermore, the measurements were performed *in vivo* and previous research has shown differences between frameshifting efficiencies *in vivo* and *in vitro* [5]. Based on the provided data, it is not possible to distinguish the contribution of each one of these factors to the total difference between measured GFP expressions and frameshifting efficiencies. As a consequence, I made the approximation that all factors result in the same relation between GFP expression and frameshifting efficiency. All factors were, therefore, summarised in one relation.

To determine the nature of the relation and its parameters, the GFP expression percentages reported by Mikl et al. were compared to other published i*n vitro* measured values of programmed ribosomal frameshifting efficiency FS. This was done for eight variants of the HIV HXB2 WT sequence [33–35] (table 4). For every sequence j the set of measured data points is $\mathbf{t}_\mathrm{j}$.

To enable comparison, the mean GFP signal μ_{S_j} was determined, for each sequence, based on the data by Mikl et al. using the algorithm introduced in section 6.2. Not for all sequences more than one measurement point of GFP expression was provided, therefore, only one parameter could be varied. As a consequence σ_{S_j} is not varied, but calculated with equation (27) with the slope $m_\sigma = 0.24$ given by the orthogonal distance regression from section 7.2. Therefore, the only parameter that is varied is μ_{S_j}, starting from an initial value of 2 % and using a proposal function with a standard deviation of 0.7 %. For each selected sequence j the following likelihood was used:

$$\mathrm{P}(\mathbf{t}_\mathrm{j}|\mu_{S_j}) = \prod_{l=1}^{L(j)} \frac{1}{\sqrt{2\pi \cdot (\sigma_B^2 + (0.16 \cdot \mu_{S_j})^2)}} \exp\left(-\frac{(t_{j,l} - \mu_B - \mu_{S_j})^2}{2(\sigma_B^2 + (0.24 \cdot \mu_{S_j})^2)}\right). \tag{28}$$

σ_B and μ_B were replaced by the values computed in section 6.3. The prior for μ_{S_j} is chosen to prevent unreasonable probabilities as

$$P(\mu_{S_j}) = \begin{cases} 0, & \mu_{S_j} < 0\,\% \text{ or } \mu_{S_j} > 100\,\% \\ 1, & \text{otherwise} \end{cases}. \tag{29}$$

For all sequences, the mean of the GFP signal mean μ_{S_j} was calculated after 100000 iteration steps over the last 50000 steps.
As stated in the beginning of this section the nature of the relation between GFP expression and frameshifting efficiency had to be analysed, as those do not necessarily match. Based on the relation observed between the determined μ_{S_j} and the frameshifting efficiency previously known, a linear relation was approximated (section 7.3). Based on this the parameter m_{eff} is introduced into the density function likelihood $\mathrm{P}(T_{\text{all}}|C_i)$ as

$$\mu_{S_j} = m_{\text{eff}} \cdot FS_j. \tag{30}$$

m_{eff} is a parameter varied in the Metropolis-Algorithm to determine base-pair free-energy differences (section 6.6). To enable comparison with the m_{eff} distribution resulting from the Metropolis-Algorithm , m_{eff} was calculated also via orthogonal distance regressions with a linear approach $\mu_{S_j} = m_{\text{eff}} \cdot FS_j + b_{\text{eff}}$. This was done with the python package scipy.odr.

6.6. Determining base-pair free-energy differences

To determine base-pair free-energy differences, the algorithm described in section 6.2 varied 50 parameters. Of those, 33 are base-pair free-energy differences
$\mathbf{\Delta G}_{\text{bp}} = (\Delta G_{\text{bp},1}, \Delta G_{\text{bp},2}, ..., \Delta G_{\text{bp},33})$ (parameters $C^0 - C^{32}$) and 15 parameters are free-energies differences of WT sequences $\mathbf{\Delta G}_{\text{WT}} = (\Delta G_{\text{WT},1}, \Delta G_{\text{WT},2}, ..., \Delta G_{\text{WT},13})$ (parameters $C^{33} - C^{47}$) (table 5). These 15 free-energy differences are not base-pair free-energy differences, but rather the overall free-energy differences between 0 and -1 frame for the WT sequences. They are varied not to quantify their values, but to better approximate m_σ. This is possible, because there are more data points for the WT sequences, and therefore, deviations can be determined more accurately. The two remaining parameters are m_σ (parameter C^{48}) and m_{eff} (parameter C^{49}).
The total data set, composed of N data points, was divided into three subsets containing N_{SARS}, N_{WT} and N_{HIV} sequences. The likelihood was derived from equation (24) and calculated as the product of three likelihoods, each one computed over the corresponding subset. In all of the likelihoods, the relation introduced in section 6.4 was taken into account for the standard deviation of the GFP signal σ_S for a sequence j:

$$\sigma_{S_j} = m_\sigma \cdot \mu_{S_j}. \tag{31}$$

The first N_{SARS} sequences were used to determine $\mathbf{\Delta G}_{\text{bp}}$, these are the 53 SARS-CoV sequences selected in section 6.1. For these sequences the set of $\mathbf{t}_{\text{SARS}}$ measurement points is provided with $L(j)$ measurements per sequence j, where $t_{\text{SARS}_{j,l}} \in \mathbf{t}_{\text{SARS}}$, $j < 53$, $l < L(j)$.

The GFP signal μ_S was modeled as a function of the free-energy differences for a sequence between 0 and -1 frame according to the thermodynamic approach in section 3.2. Additionally, the slope parameter m_{eff} was included as defined in section 6.5, resulting in

$$\mu_S(\Delta G, m_{\text{eff}}) = m_{\text{eff}} \cdot \frac{e^{-\Delta G/k_{\text{B}}T}}{e^{-\Delta G/k_{\text{B}}T} + 1}. \tag{32}$$

The temperature was set to 310.15 K [4], while k_{B} is the Boltzmann constant ($1.38 \cdot 10^{-23}$ J/K [21]). The free-energy differences for the different SARS-CoV sequences can be written as $\mathbf{\Delta G}_{\text{SARS}}$=($\Delta G_{\text{SARS},1}, \Delta G_{\text{SARS},2}, ..., \Delta G_{\text{SARS},53}$). For each sequence j, $\Delta G_{\text{SARS},j}$ was calculated based on the sum of base-pair free-energy differences relevant for the sequence.

$$\Delta G_{\text{SARS},j} = \sum_{r=1}^{m} \Delta G_{\text{bp},r} \cdot g(r, j) \tag{33}$$

Here $m = 33$ is the amount of considered base-pair changes for all the sequences. The function $g(r, j)$ enables to only sum up free-energy differences for base-pair changes, which occur during a frameshifting event in SARS-CoV sequence j:

$$g(j, r) = \begin{cases} 1, & \text{base-pair}r \text{ changes during PRF for sequence } j \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

The resulting part of the likelihood function for the N_{SARS} sequences is

$$\begin{aligned} \mathrm{P}(\mathbf{t}_{\text{SARS}} | \mathbf{\Delta G}_{\text{bp}}, m_\sigma, m_{\text{eff}}) = \prod_{j=1}^{N_{\text{SARS}}} \prod_{l=1}^{L(j)} & \frac{1}{\sqrt{2\pi \cdot \left(\sigma_B^2 + \left(m_\sigma \cdot \mu_S(\Delta G_{\text{SARS}_j}, m_{\text{eff}})\right)^2\right)}} \cdot \\ & \exp\left(-\frac{\left(t_{\text{SARS}_{j,l}} - \mu_B - \mu_S(\Delta G_{\text{SARS}_j}, m_{\text{eff}})\right)^2}{2 \cdot \left(\sigma_B^2 + \left(m_\sigma \cdot \mu_S(\Delta G_{\text{SARS}_j}, m_{\text{eff}})\right)^2\right)}\right). \end{aligned} \tag{35}$$

The next N_{WT} sequences were used in the first place to determine m_σ more accurately. As a side-effect, they also contributed to the determination of 15 ΔG_{WT} distributions. The sequences are the 15 WT sequences chosen in section 6.4, which are suitable to this purpose due to their high number of provided data points. For these sequences the set of $\mathbf{t}_{\text{WT}}$ measurement points is provided with $L(j)$ measurements per sequence j, where $t_{\text{WT}_{j,l}} \in \mathbf{t}_{\text{WT}}$, $j < 15$, $l < L(j)$. Here the GFP signal μ_S was also modeled according to the thermodynamic approach in section 3.2 after including the slope parameter m_{eff}

$$\mu_S(\Delta G, m_{\text{eff}}) = m_{\text{eff}} \cdot \frac{e^{-\Delta G/k_{\text{B}}T}}{e^{-\Delta G/k_{\text{B}}T} + 1}. \tag{36}$$

The temperature was set to 310.15 K [4], while k_B is the Boltzmann constant ($1.38 \cdot 10^{-23}$ J/K [21]). The free-energy differences for the different WT sequences can be written as $\mathbf{\Delta G}_\mathrm{WT}$=$(\Delta G_{\mathrm{WT},1}, \Delta G_{\mathrm{WT},2}, ..., \Delta G_{\mathrm{WT},15})$. The free-energy differences for the WT sequences are not calculated by base-pair free-energy differences, but are varied parameters. The resulting part of the likelihood function for the N_WT sequences is:

$$\mathrm{P}(\mathbf{t}_\mathrm{WT}|\mathbf{\Delta G}_\mathrm{WT}, m_\sigma, m_\mathrm{eff}) = \prod_{j=1}^{N_\mathrm{WT}} \prod_{l=1}^{L(j)} \frac{1}{\sqrt{2\pi \cdot \left(\sigma_B^2 + \left(m_\sigma \cdot \mu_S(\Delta G_{\mathrm{WT}_j}, m_\mathrm{eff})\right)^2\right)}} \cdot \exp\left(-\frac{\left(t_{\mathrm{WT}_{j,l}} - \mu_B - \mu_S(\Delta G_{\mathrm{WT}_j}, m_\mathrm{eff})\right)^2}{2 \cdot \left(\sigma_B^2 + \left(m_\sigma \cdot \mu_S(\Delta G_{\mathrm{WT}_j}, m_\mathrm{eff})\right)^2\right)}\right). \tag{37}$$

(38)

The last $N_\mathrm{HIV} = 8$ sequences are HIV sequence variants used to determine m_eff (section 6.1). For them, the set of $\mathbf{t}_\mathrm{HIV}$ measurement points is provided with $L(j)$ measurements per sequence j, where $t_{\mathrm{HIV}_{j,l}} \in \mathbf{t}_\mathrm{HIV},\ j < 8,\ l < L(j)$. For these sequences the frameshifting efficiencies published by different authors [33–35] $\mathbf{FS}_\mathrm{HIV} = (FS_{\mathrm{HIV},1}, FS_{\mathrm{HIV},2}, ...FS_{\mathrm{HIV},8})$, were included in the likelihood using their relation to the GFP signal μ_S:

$$\mu_S(FS, m_\mathrm{eff}) = m_\mathrm{eff} \cdot FS. \tag{39}$$

The resulting likelihood function for these N_HIV sequences is

$$\mathrm{P}(\mathbf{t}_\mathrm{HIV}, \mathbf{FS}_\mathrm{HIV}|m_\sigma, m_\mathrm{eff}) = \prod_{j=1}^{N_\mathrm{HIV}} \prod_{l=1}^{L(j)} \frac{1}{\sqrt{2\pi \cdot \left(\sigma_B^2 + \left(m_\sigma \cdot \mu_S(\Delta FS_{\mathrm{HIV}_j}, m_\mathrm{eff})\right)^2\right)}} \cdot \exp\left(-\frac{\left(t_{\mathrm{HIV}_{j,l}} - \mu_B - \mu_S(\Delta FS_{\mathrm{HIV}_j}, m_\mathrm{eff})\right)^2}{2 \cdot \left(\sigma_B^2 + \left(m_\sigma \cdot \mu_S(\Delta FS_{\mathrm{HIV}_j}, m_\mathrm{eff})\right)^2\right)}\right). \tag{40}$$

(41)

Combined, this leads to the overall likelihood $\mathrm{P}(\mathbf{T}_\mathrm{all}|C_i)$

$$\mathrm{P}(\mathbf{T}_\mathrm{all}|C_i) = \mathrm{P}(\mathbf{t}_\mathrm{SARS}|\mathbf{\Delta G}_\mathrm{bp}, m_\sigma, m_\mathrm{eff}) \cdot \mathrm{P}(\mathbf{t}_\mathrm{WT}|\mathbf{\Delta G}_\mathrm{WT}, m_\sigma, m_\mathrm{eff}) \cdot \mathrm{P}(\mathbf{t}_\mathrm{HIV}, \mathbf{FS}_\mathrm{HIV}|m_\sigma, m_\mathrm{eff}) \tag{42}$$

σ_B and μ_B were replaced by the values computed in section 6.3. For all parameters, Gaussian proposal distributions with different deviations (table 5) were used. The prior was chosen based

on the free-energy differences determined by Bock et al. [3] in order to set a reasonable range for the free-energy differences as

$$P(\mathbf{\Delta G}_{\mathrm{bp}}, \mathbf{\Delta G}_{\mathrm{WT}}) = \begin{cases} 0, & \min(\mathbf{\Delta G}_{\mathrm{bp}}, \mathbf{\Delta G}_{\mathrm{WT}}) < -25\ \frac{\mathrm{kJ}}{\mathrm{mol}} \text{ or } \max(\mathbf{\Delta G}_{\mathrm{bp}}, \mathbf{\Delta G}_{\mathrm{WT}}) > 25\ \frac{\mathrm{kJ}}{\mathrm{mol}} \\ 1, & \text{otherwise} \end{cases} . \tag{43}$$

The algorithm ran for 100000 steps and was executed twice to check for convergence. Distributions of the different parameters and 95 % confidence intervals were computed over the last 50000 iteration steps.

6.7. Correlation analysis

The presence of correlated parameters gives rise to convergence issues of the Metropolis-Hastings algorithm. As a consequence, correlations between the parameters used for the determination of base-pair free-energy differences were identified with a correlation matrix. This matrix displays the Pearson correlation coefficient r for each pair of varied parameters. r is defined as [36]

$$r = \frac{\sum_{i=1}^{N} z_{x,i} z_{y,i}}{N}. \tag{44}$$

Here x, y are the possibly correlated parameters, N is the amount of data points and z is the standard score. Given a mean value μ and a standard deviation σ, the standard score of an observed value x is [36]

$$z = \frac{x - \mu}{\sigma}. \tag{45}$$

A positive r shows a linear positive correlation and a negative r a linear anti-correlation. The closer $|r|$ is to 1, the stronger the values are correlated/anti-correlated [36]. The correlation matrix was generated from the last 50000 steps of the algorithm determining base-pair free-energy differences.

Just as one needs at least as many equations as variables to solve a system of equations, one needs enough sequences here to determine base-pair free-energy differences. Base-pair free-energy differences that cannot be determined unambiguously for this reason were varied from the beginning only as the sum of the correlated differences. As an example, the base-pair changes P1: AU $\rightarrow$ UU and P2: UA $\rightarrow$ AA only occur for the slippery sequence (UAUAAAC) of one of the SARS-CoV variants for which data points are available. To determine the two individual base-pair free-energy differences based on the frameshifting efficiency where both

change at the same time is not possible. However, the sum of the two free-energy differences can be determined and is used as a parameter in the model.

6.8. Prediction of a frameshifting efficiency based on base-pair free-energy differences

In order to perform cross-validation and test the predictive power of the model introduced in section 6.6, the algorithm was run after having excluded one of the GFP expression measurements from the input data set. The value of the excluded measurement was then predicted from the free-energy differences estimated based on the remaining data points. The excluded data point corresponds to the SARS-CoV mRNA variant containing the slippery sequence A_AAU_UUA. The base-pair changes occurring when the sequence slips from the 0 frame to the -1 frame are P3: UA $\rightarrow$ AA and A3: AU $\rightarrow$ UU.
In order to predict the GFP expression for this sequence, the corresponding frameshifting free-energy difference was first calculated as the sum of the base-pair free-energy differences obtained for P3: UA $\rightarrow$ AA and A3: AU $\rightarrow$ UU. A distribution of the frameshifting free-energy difference was obtained by considering the last 50000 iteration steps and summing each free-energy difference obtained for P3: UA $\rightarrow$ AA to each free-energy difference obtained for A3: AU $\rightarrow$ UU . Afterwards, the GFP signal without the background fluorescence was calculated using equation 32 for every iteration step. The mean of the background fluorescence expression was added to this signal to compare the resulting GFP expression to the one provided by Mikl. et al.

7. Results and Discussion

In the beginning of this section the background fluorescence signal is analysed and the dependency between the mean μ_{S_j} and standard deviation σ_{S_j} of the GFP signal is studied. Based on the results of these analyses, the parameters and relations to include in the likelihood $\mathrm{P}(T_{\mathrm{all}}|C_i)$ are determined. The distributions of the parameters' values obtained from the Bayesian approach, described in section 6, are then reported. In particular, section 7.4 focuses on the results for the base-pair free-energy differences involved in the frameshifting of the SARS-CoV variants. Next, after having discussed correlation and convergence issues, the computed base-pair free-energy differences are compared to the ones determined for dnaX variations by Bock et al. [3]. Finally, for cross-validation, section 7.7 reports on the prediction of one GFP expression for one of the sequence based on the base-pair free-energy differences obtained from the other sequences in the data set.

7.1. Determining mean and standard deviation of the background noise μ_{B} and σ_{B}

A portion of the green fluorescence signal measured to quantify GFP expression is caused by background fluorescence. With the data available from the early stop codon measurements, it can be observed that the background fluorescence distribution resembles a Gaussian distribution for $\%\,\mathrm{GFP} < 2\%$ (figure 7). However, it seems like there are multiple reasons for background fluorescence, as there are also higher GFP expressions in the distribution, which do match the Gaussian distribution. For these, there are not enough data points to asses their distribution. If one uses all data points to calculate a mean and a standard deviation for a Gaussian approximation of the distribution, the resulting distribution is quite different from the measured one (orange curve in figure 7). This is because one would mix different reasons of background fluorescence into one distribution. For a better approximation, the GFP expressions, which did not fit the Gaussian, were cut of before calculating the mean and the standard deviation 6.3. This results in

$$\mu_B = 0.73\,\%, \tag{46}$$

$$\sigma_B = 0.29\,\%. \tag{47}$$

The Gaussian distribution described by μ_B and σ_B is shown in figure 7. This distribution overlaps nicely with the actual distribution of the measured GFP expression for $\%\,\mathrm{GFP} < 2\%$. However, it fails to capture background noise in the measured GFP expressions bigger than $2\,\%$. $95\,\%$ of the GFP expressions caused by background fluorescence are smaller than $1.3\,\%$. This is in agreement the threshold determined by Mikl at al. [4].

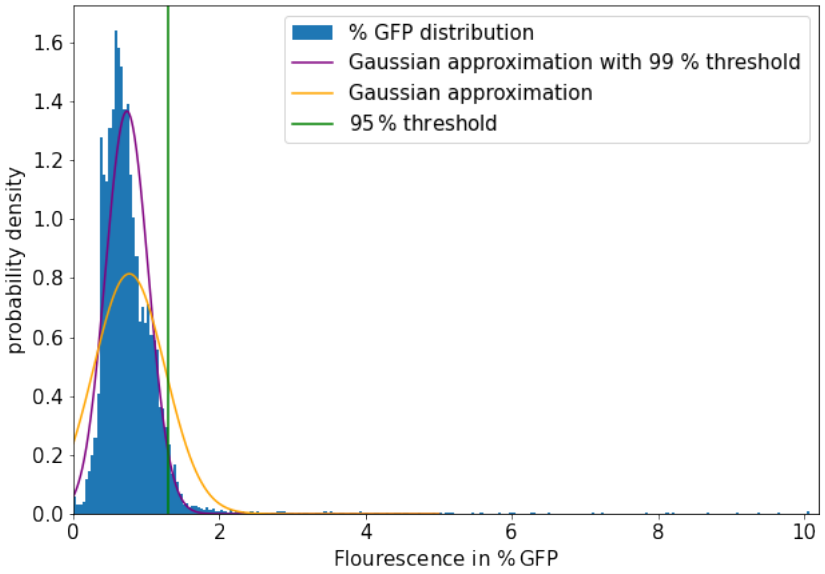

Figure 7: Normalised distribution of the background signal in the reported GFP expressions. In blue: the distribution of the measurements. In orange: Gaussian approximation using all data points. In purple: Gaussian approximation using only data points below a 99 % threshold. The green vertical line displays the 95% threshold of the measurement points.

7.2. Determining the standard deviation of the GFP signal σ_{S_j}

As described in section 6.4, not enough data points were provided for every sequence j to determine the standard deviation of the GFP signal σ_{S_j} directly from the measurements. Therefore σ_{S_j} needed to be estimated. This could be achieved by introducing separate μ_{S_j} and σ_{S_j} parameters for each sequence j in the likelihood, by setting σ_{S_j} as a constant for all sequences, or by relating the standard deviation of the GFP signal σ_{S_j} to the GFP signal mean μ_{S_j}. If μ_{S_j} and σ_{S_j} are highly correlated, the relation can be expressed by a function which can then be included in the likelihood. Additionally, an advantage of using a relation between σ_{S_j} and μ_{S_j}, is that it results in a reduction of the number of varied parameters. In order to test if a relation between σ_{S_j} and μ_{S_j} was present, both parameters were first determined for different virus and human mRNA WT sequences using a Bayesian Metropolis-Algorithm. For each sequence the probability densities were calculated independently resulting in 15 μ_{S_j} and σ_{S_j} values. An example of the resulting parameter chains obtained from the algorithm is provided for the μ_{S_j} and σ_{S_j} corresponding to the PLRV WT sequence (figure 8).

When comparing the resulting σ_{S_j} for different WT sequences, it can be observed that higher values of σ_{S_j} correspond to higher values of μ_{S_j}. However, after having plotted σ_{S_j} vs. μ_{S_j} (figure 9), I observed that the exact relation between the two parameters is not clear. In a first approximation, a linear dependency is the simplest relation to describe the distribution of the

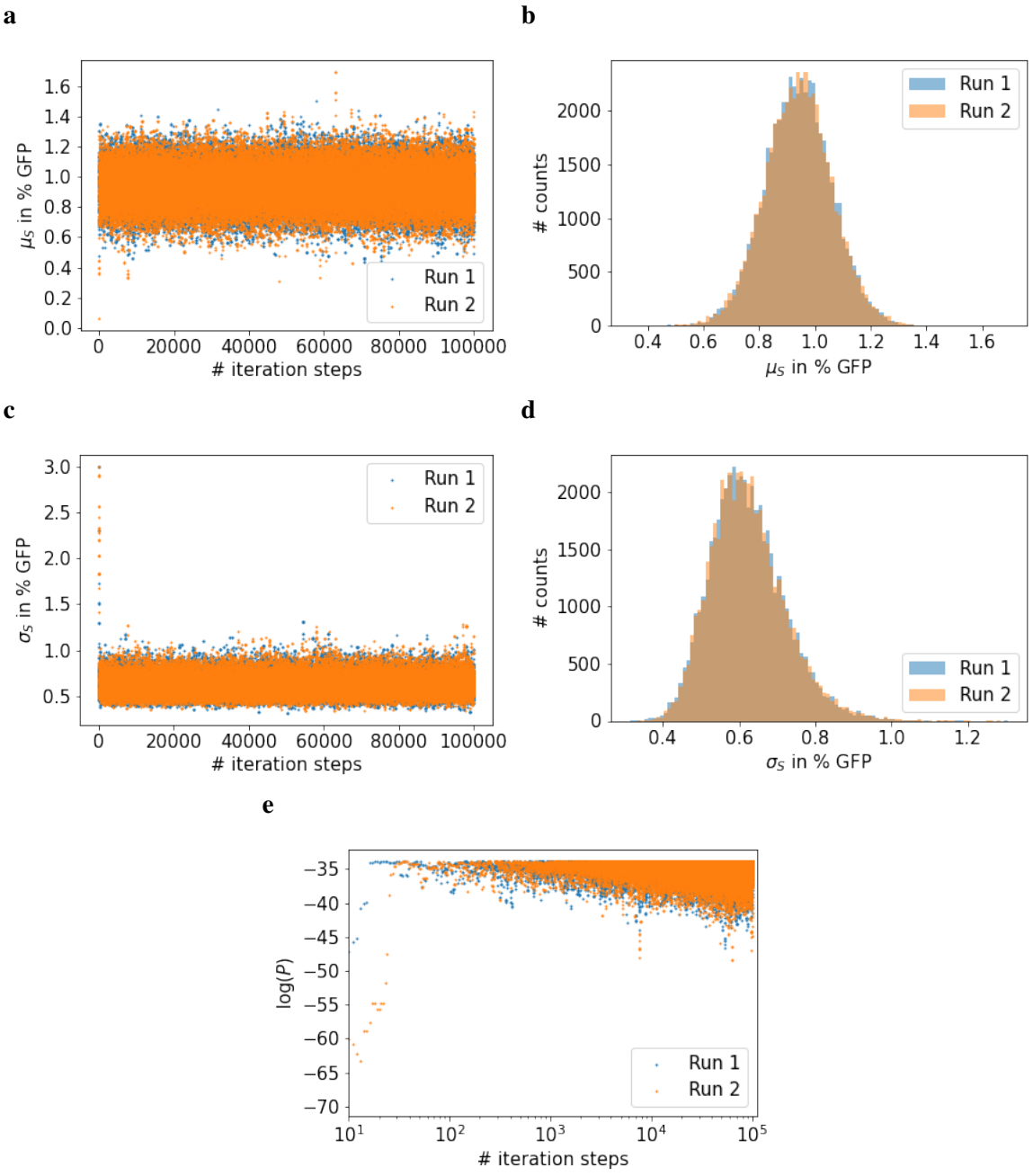

Figure 8: Results of the Metropolis-Algorithm determining μ_{S_j} and σ_{S_j} for PLRV. The algorithm was executed twice, the first run is visualised in blue and the second in orange. **a** μ_{S_j}(PLRV) convergences for 100000 iteration steps. **b** The histogram of μ_{S_j} for PLRV over the last 50000 iteration steps. **c** σ_{S_j}(PLRV) convergence for 100000 iteration steps. **d** The histogram of σ_{S_j} for PLRV over the last 50000 iteration steps. **e** The logarithm of the product of likelihood and prior as function of iteration steps.

data. I thus tested this dependency by introducing into the model a slope parameter m_σ

$$\sigma_{S_j} = m_\sigma \cdot \mu_{S_j}. \tag{48}$$

A first estimate for m_σ was obtained by an orthogonal distance regression (ORD). The line resulting from the regression is shown in figure 9. The obtained m_σ value of 0.24 ± 0.07 was only used for the Metropolis-Algorithm analysing the relation between the measured GFP expression and the frameshifting efficiency in the next section. In the final calculation, the distribution of m_σ values that is most likely to generate the experimental data was, however, determined by including this slope parameter in the likelihood, together with the free-energy parameters as described in section (6.6). The value of the parameter converged as shown in figure 10 (section 6.6). The 95 % confidence interval for m_σ is displayed in figure 9. The range in which values for $\sigma_{S_j}(\mu_{S_j})$ should lie based on this 95 % confidence interval (grey area in figure 9) can be compared to the sampled values of σ_{S_j} for the different WTs. One can observe that the standard deviations of the signal are almost evenly spread below and above this range. Additionally, the general tendency that greater GFP signal means have greater standard deviations applies. Based on these observations, I decided, as a first approximation, to use a linear dependency in our model as a relation between σ_{S_j} and μ_{S_j}.

It is interesting to observe that slope determined by ORD, is not in the 95 % interval of the slope parameter determined with the Metropolis-Algorithm. A possible reason for this, is that when m_σ is optimized in the Metropolis-algorithm, its most likely distribution is affected by the values assumed by the other parameters at each iteration (sections 6.4,6.6). A further difference between the two algorithms is that, while the Metropolis-algorithm provides the distribution of m_σ, and hence the confidence interval, as an output, the ORD requires to include *a priori* the standard deviations for the determined WT μ_{S_j} and σ_{S_j} values.

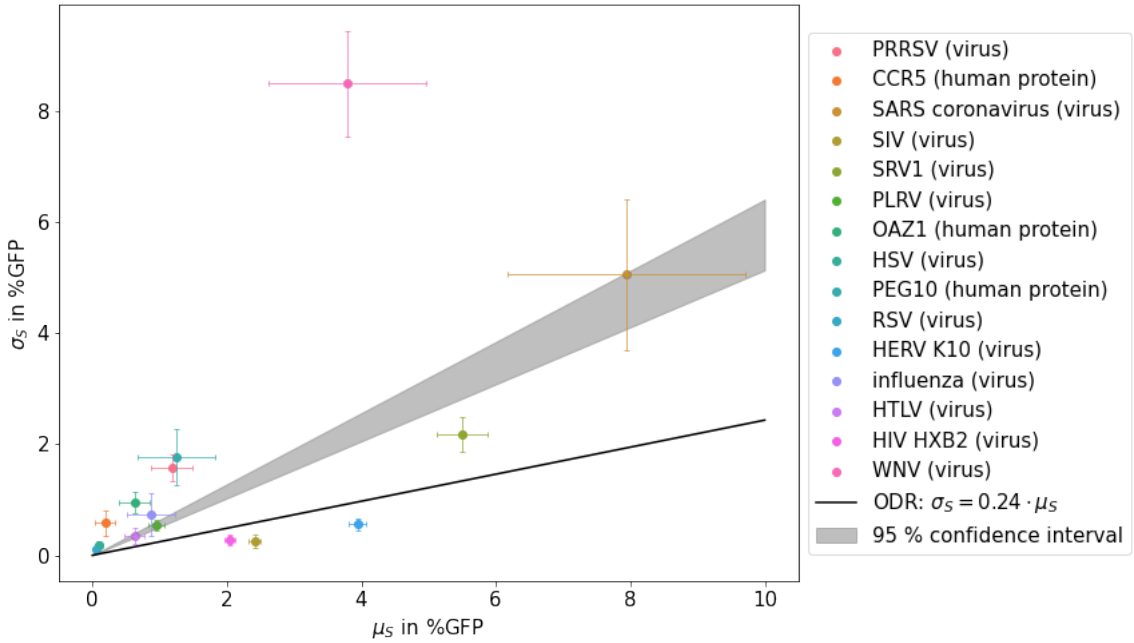

Figure 9: σ_{S_j} and μ_{S_j} values for different WT sequences. A different colour is used for each sequence and error bars are obtained by the standard deviation of the sampled distributions. The black line shows the result of the linear orthogonal distance regression performed to fit the data taking into account the standard deviations on both σ_S and μ_S. In grey: the $95\,\%$ confidence interval for the slope parameter m_σ resulting from the the Metropolis-Algorithm determining base-pair free-energies (section 6.6).

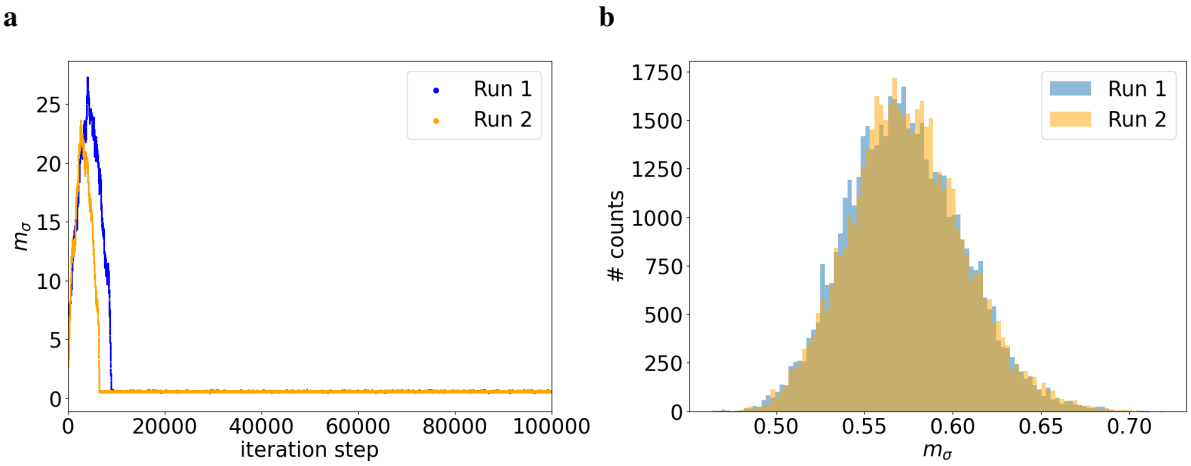

Figure 10: The slope parameter m_σ is one out of 50 varied parameters of the Metropolis-Algorithm determining base-pair free-energies (section 6.6). The algorithm was executed twice over 100000 steps for a convergence check. The resulting runs are labeled in orange and blue. **a** m_σ in dependence of the iteration step. **b** The histogram of m_σ over the last 50000 iteration steps.

7.3. The relation between the measured GFP expression and the frameshifting efficiency

As previously stated, the GFP expression for a sequence provided by Mikl et al. is the sum of the background fluorescence and the GFP signal for this sequence. The distribution of the background fluorescence has already been approximated in section 7.1. The GFP signal is caused by frameshifting events, but, as introduced in section 6.5, it is not necessarily equal to the frameshifting efficiencies. However, as the thermodynamic model introduced in section 3.2 relates the base-pair free-energy differences to the frameshifting efficiency, it is necessary to understand how the provided GFP expressions are related to the frameshifting efficiencies. To better understand this relation, previously published frameshifting efficiencies for HIV sequence variations ([33], [34], [35]) were compared to GFP signals sampled based on the data provided by Mikl et al. [4]. For this, the GFP signals μ_{S_j} were sampled with a Metropolis-Algorithm using the GFP expressions of the HIV sequences as input data and varying μ_{S_j} (section 6.5). For each sequence, the GFP signal resulting from the Metropolis-sampling is plotted against the corresponding frameshifting efficiency obtained *in vitro* (figure 11). As a linear relationship between GFP expression and frameshifting efficiency seems to be a reasonable approximation, it was introduced in the Metropolis-algorithm (see section 6.6) as:

$$\mu_{S_j} = m_{\text{eff}} \cdot FS_j. \tag{49}$$

The slope-parameter m_{eff} was varied in the algorithm together with the free-energy parameters and m_σ. It was also tested, if the y-axis intercept b_{eff} should also be included in the model. For this the final Metropolis-Algorithm, determining the base-pair free-energy differences, was executed including also b_{eff} as varied parameter. However, b_{eff} was heavily correlated to the slope parameter m_{eff}, which resulted in convergence problems. Additionally, including the b_{eff} parameter reduced the value of the likelihood to which the algorithm converged. To further evaluate the necessity to include b_{eff} in the model, linear ODRs were performed and resulted in the following relations with the *in vitro* efficiencies (figure 11)

$$\text{Dulude et al.} : \mu_{S_j} = (0.34 \pm 0) \cdot FS - (1.08 \pm 0)\,\%, \tag{50}$$

$$\text{Leger et al.} : \mu_{S_j} = (0.22 \pm 0.04) \cdot FS + (0.2 \pm 0.3)\,\%, \tag{51}$$

$$\text{Biswas et al.} : \mu_{S_j} = (0.41 \pm 0.11) \cdot FS + (0 \pm 0.2)\,\%, \tag{52}$$

$$\text{All papers} : \mu_{S_j} = (0.25 \pm 0.05) \cdot FS + (0 \pm 0.3)\,\%. \tag{53}$$

For the paper by Dulude et al. [33] only 2 sequences were suitable for comparison and, as a re-

sult, no error on the ODR result could be estimated. The offsets calculated in (50)-(53) support the choice, not to include b_{eff} into the model, as they are close to 0 % [33–35].
Figure 12 shows the distribution of the slope-parameter m_{eff} obtained with the algorithm determining the base-pair free-energy differences (section 6.6). The 95 % confidence interval resulting for this m_{eff} is included in figure 11. The ODR calculated when considering all the data sets together is within the range given by the interval.

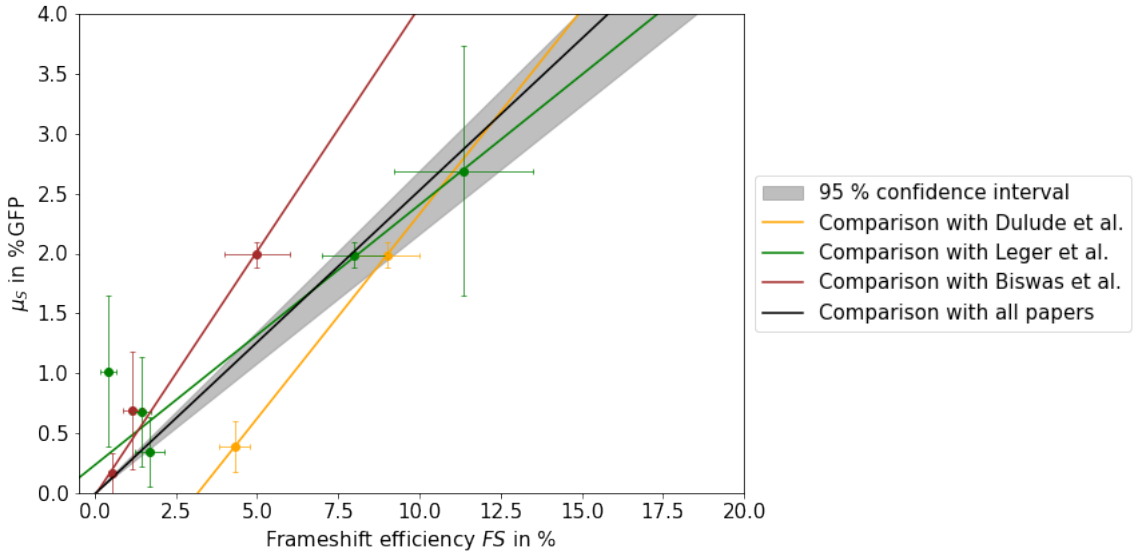

Figure 11: Comparison between the GFP signals estimated with a Metropolis-algorithm from Mikl et al. data (*in vivo*) and the frameshifting efficiencies obtained *in vitro* for HIV sequence variations. Different colours are used for different data sets. The black line displays the linear relation obtained by a ODR (12). The 95 % confidence interval of m_{eff} calculated by varying m_{eff} in the Metropolis-Algorithm determining base-pair free-energy differences (section 6.6) is shown in grey.

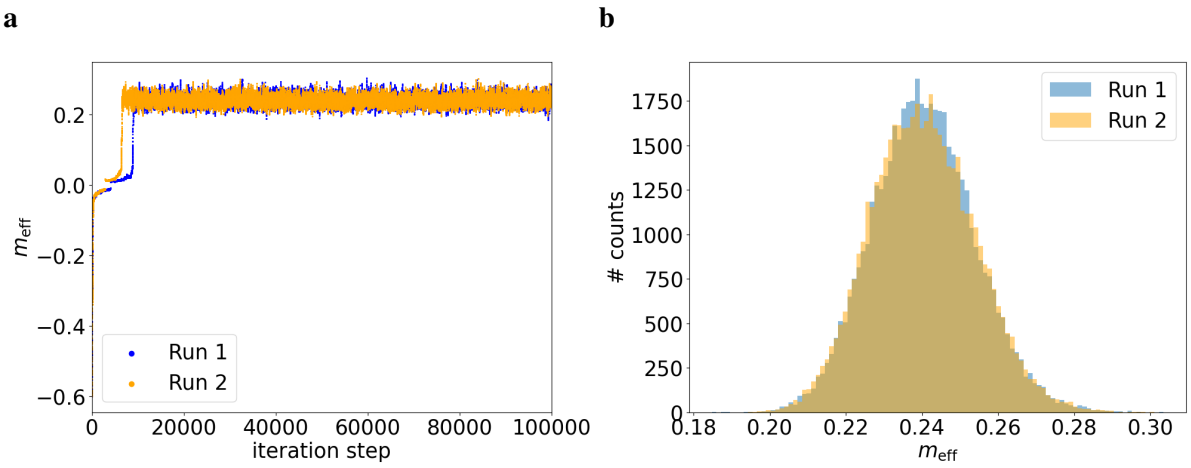

Figure 12: The slope parameter m_{eff} is one out of 50 varied parameters of the Metropolis-Algorithm determining base-pair free-energies (section 6.6). The algorithm was executed twice over 100000 steps for a convergence check. The resulting runs are labeled in orange and blue. **a** m_{eff} in dependence of the iteration step. **b** The histograms of m_{eff} over the last 50000 iteration steps.

7.4. Determining base-pair free-energy differences

As a result of the last three sections, base-pair free-energy differences for SARS-CoV can now be sampled using Bayesian statistics, based on the provided GFP expressions [4].
This was done with the Metropolis-Algorithm introduced in section 6.6, where, in addition to the 33 base-pair free-energy differences ΔG_{bp}, also m_{eff}, m_σ and 15 ΔG_{WT} were varied. The algorithm was stopped once the product of likelihood and prior was maximised. Based on the results shown in figure 13, I considered the calculations to be converged after 50000 iteration steps. The algorithm ran for an additional 50000 steps to sample the parameters.

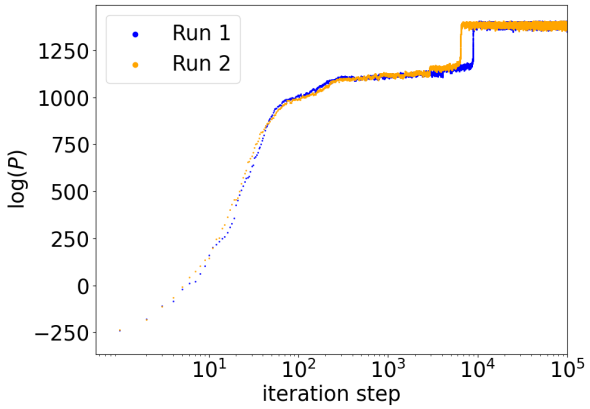

Figure 13: The logarithm of the product of likelihood and prior calculated with the Metropolis-Algorithm determining base-pair free-energy differences in dependence of the iteration step. The two colours represent two independent runs of the algorithm.

The resulting distributions of the parameters will be presented and discussed in this section. The convergence of the slope parameters m_σ and m_{eff} was presented in the previous sections 7.3 and 7.2.

The 15 free-energy differences between 0 and -1 frame of WT sequences ($\Delta G_{\text{WT,j}}$) were varied to determine m_σ. The distributions obtained for 12 of them resembled Gaussian distributions. For the other three, the distributions only have a lower boundary and their upper boundary is determined by the boundary of the prior at 25 kJ/mol (figure 20). As this convergence behavior was observed also in the distributions of some of the base-pair free-energy differences, it will be further explained later in this section.

For the base-pair free-energy differences ($\Delta G_{\text{bp,r}}$) three different convergence behaviours can be observed. Most commonly, the probability densities of some base-pair free-energy differences converge to localized densities around a certain value in the range $[-25\,\text{kJ/mol}, 25\,\text{kJ/mol}]$. An example for this is the free-energy difference of the base-pair change A3: UA $\rightarrow$ CA shown in figures 14a,14b.

For other base-pair changes the free-energy difference distribution only had a lower boundary. This can happen if, for the sequences in which the base-pair change occurs, the measured GFP expressions are very small. Indeed, it can be observed that many of these GFP expressions are below the background noise threshold assumed by Mikl et al. of $1.3\,\%$. As a consequence, the frameshifting efficiency would be close to $0\,\%$. For such small efficiencies the slope of the the frameshift efficiency with respect to the free-energy difference (10) is also very low. Therefore, the range of possible free-energy differences resulting in an efficiency close to the measured one is very large. For better visualisation note figure 5 in section 6.4. Additionally, as the frameshifting efficiency approaches 0 %, the free-energy difference tends to infinity. This can explain why the free-energy distribution does not have an upper boundary. An example base-pair change, where the free-energy difference does not have an upper boundary is A3: GC$\rightarrow$ CC. The convergence behavior is shown in figures 14c,14d.

Surprisingly some of the base-pair changes, for example P1:UA$\rightarrow$ CA (figures 14e,14f), have only an upper boundary and no lower boundary. For all free-energy differences smaller than 0 kJ/mol the base pair would be more likely in -1 frame than in 0 frame. This is not intuitive as the affected base-pair changes are Watson-Crick base pairs in 0 frame and non standard pairs in -1 frame. Noticeably, these distributions still have a Gaussian shaped maxima in the range $\Delta G_{\text{bp}} > 0\,\text{kJ/mol}$. This distribution shape is probably based on the limited amount of data resulting in a large uncertainty. For example ΔG_{bp}(P1:UA$\rightarrow$ CA) is sampled based only on one measurement point. This is problematic for a statistical analysis.

An overview over all estimated base-pair free-energy differences is shown in figure 15. If a bar is labeled with two or three base-pair changes together, it means that the variation of the sum of their base-pair free energy differences was considered in the algorithm. This was necessary for the reason explained in section 6.7. In particular, in the sequences of the SARS-C data set

where one of those base-pair changes took place, this change always occurred together with the other base-pair changes included in the sum and never in combination with any other base-pair change.

As expected, most of the probability densities for the ΔG_{bp} values have a maximum for a ΔG_{bp} value larger than 0 kJ/mol, which means that the 0 frame is more probable. This is in agreement with the fact that the base pairs in the 0 frame are Watson-Crick pairs.

Noticeably, the distributions of the free-energy difference for ΔG_{bp}(P3: UA $\rightarrow$ AA), ΔG_{bp}(A3: AU $\rightarrow$ UU) and ΔG_{bp}(P3: AU $\rightarrow$ UU) are among those closest to 0 kJ/mol. This labels them, among all the converged base-pair changes, as the most likely to happen. This is surprising as, for those changes, the -1 frame base-pairs are purine-purine or pyrimidine-pyrimidine pairs and thus, even in wobble position, expected to be unlikely pairs. However, as the conformations that the base pairs assume in the ribosome in these cases was not provided, no further conclusions can be drawn.

Reasonably ΔG_{bp} (P1: UA $\rightarrow$ AA) is higher than other free-energy differences, as non-standard paring is less likely in the P1 position (not Wobble-position).

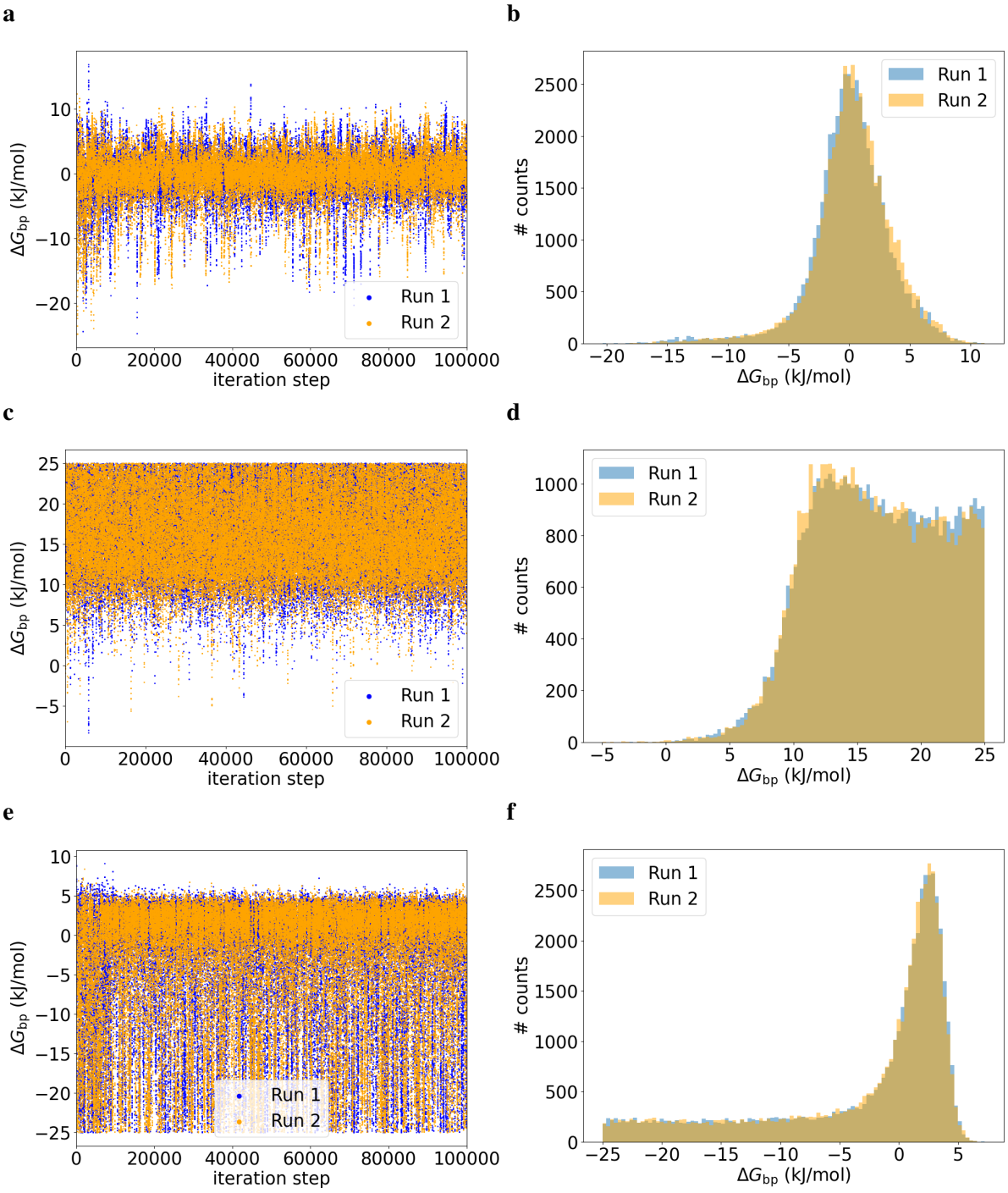

Figure 14: Examples of convergence behaviors for base-pair free-energy differences determined with the Metropolis-Algorithm. The algorithm was executed twice, the first run is labeled in blue and the second one in orange **a** The base-pair free-energy difference ΔG_{bp}(A3: UA $\rightarrow$ CA) as a function of the iteration step. **b** A histogram of ΔG_{bp}(A3: UA $\rightarrow$ CA) over the last 50000 iteration steps. **c** The base-pair free-energy difference ΔG_{bp}(A3: GC$\rightarrow$ CC) as a function of the iteration step. **d** A histogram of ΔG_{bp}(A3: GC$\rightarrow$ CC) over the last 50000 iteration steps. **e** The base-pair free-energy difference ΔG_{bp}(P1: UA $\rightarrow$ CA) as a function of the iteration step. **f** A histogram of ΔG_{bp}(P1: UA $\rightarrow$ CA) over the last 50000 iteration steps.

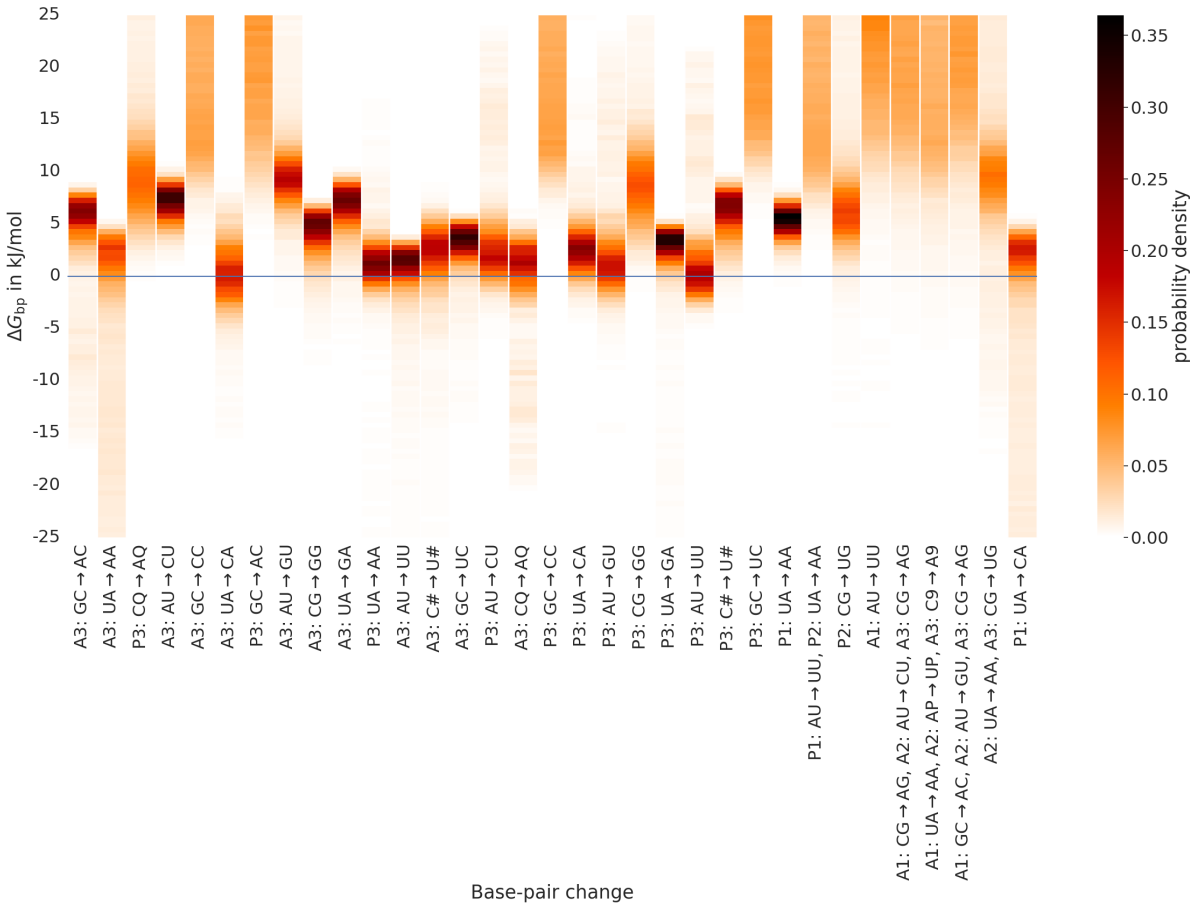

Figure 15: Probability distributions of all base-pair free-energy differences obtained from histograms over the last 50000 iteration steps of the Metropolis-Algorithm. For bars labeled with multiple base-pair changes, the sum of their base-pair free energy differences was considered as a parameter in the likelihood.

7.5. Correlation analysis and convergence issues

In the Metropolis-Algorithm, correlation between varied parameters can generate convergence problems and introduce limitations in the determination of some parameters (section 6.7). For example, if two parameters are correlated, only their sum or their difference might be determinable. To check for the presence of correlations, a correlation matrix showing the linear correlation of the parameters was generated (figure 16) (section 6.7). From the matrix, several groups of correlated parameters can be observed. Possible reasons for the correlations are presented below. An assessment that can be used in case the presence of correlation is problematic follows.

Noticeably, the slope parameters m_σ and m_{eff} are correlated (Pearson factor $|r| = 0.35$). The correlation of these two parameters does not markedly affect the determination of the base-pair free-energy differences as the correlations between the $\mathbf{\Delta G}_{\text{bp}}$ parameters and either m_{eff} or m_σ are small (Pearson factor $|r| < 0.13$).

The correlation of the WT free-energy differences is a consequence of their correlation to the slope parameter m_{eff}. Given a constant GFP signal, a smaller m_{eff} results in larger frameshifting efficiencies and, thus, in smaller free-energy differences. The free-energy differences for WT sequences, which are not correlated with m_{eff} are the ones, which only have a lower boundary (see section 7.4). For these WT sequences the GFP signal is very small. Therefore, if m_{eff} changes, the difference in the resulting frameshifting efficiency is also small $FS = \mu_S / m_{\text{eff}}$. As a consequence the frameshifting efficiency is still in the range of curve 10 where the slope is small (see section 3.2). Hence, the probability density of the free-energy difference can only be determined with a lower bound and the parameter sampling is unaffected by the slope change. As the WT parameters are not strongly correlated with the base-pair free-energy differences, the correlation between the WT free-energy differences is not problematic to determine base-pair free-energy differences. Surprisingly, the same correlation observed between WT free-energy differences and m_{eff} cannot be observed between the base-pair free-energy differences and m_{eff}. The range of the base-pair free-energy differences is very similar to that of the WT free-energy differences and the same reasoning could be applied to explain a correlation.

In theory, enough different combinations of base-pair changes occurred in the considered sequences, for all 33 base-pair free-energy differences to be clearly sampled. This means that every base-pair change occurs either in: (a) at least one sequence where it is the only base-pair change involved in frameshifting, or (b) at least one sequence where the other base-pair changes can be sampled also in other sequences (either as the only base-pair changes involved in frameshifting, or in combination with changes differ from sequence to sequence). Nevertheless, for the base-pair free-energy differences, the matrix in figure 16 clearly shows sets of correlated parameters are present.

An example of these sets includes the following base-pair changes: P3: AU $\rightarrow$ CU, A3: CQ $\rightarrow$ AQ, P3: AU $\rightarrow$ GU, P3: AU $\rightarrow$ UU, A3: GC $\rightarrow$ AC and A3: UA $\rightarrow$ AA. If one has a closer look, it stands out that the sequences necessary to determine those base-pair free-energy differences are rarely used to determine base-pair free-energy differences that do not belong to the set. This explains the presence of correlation in particular sets of parameters.

Now it will be explained how parameters within one set are correlated. Among the sequences in the set mentioned above, only the base-pair changes P3: AU $\rightarrow$ CU and A3: GC$\rightarrow$ AC occur as in case (a). Based on the GFP expressions provided for their corresponding sequence, the free-energy difference of these two base-pair changes can be clearly sampled. The other base-pair free-energy differences of the set can, then, be sampled based on the GFP expression of the sequences where they occur. According to the model in section 3.2, the total frameshifting free-energy difference is the sum over the base-pair free-energy differences. Therefore, within the same frameshifting event, if one base-pair free-energy difference gets smaller, the other has to become bigger. As a result, it is reasonable that the base-pair free-energy differences involved in the same frameshifting events are anti-correlated. Despite these correlations, the base-pair

free-energy differences can still be sampled with clear distributions. An additional challenge for convergence in this specific set is that the base-par changes, relevant for the SARS-CoV WT sequence, P3: AU $\rightarrow$ UU and A3: CQ $\rightarrow$ AQ, are part of this set. The GFP expressions given for the SARS-CoV WT sequence are distributed over a very large range (11 %), even after excluding out outlier values. This makes convergence difficult to reach as the base-pair free-energy differences involved in frameshifting for the WT SARS-CoV cannot fit all the widely-spread data points.

In general, uncertainties and errors in the measurements contribute to the presence of convergence problems. Indeed, based on the analysis of the measured values for the WT sequences, it is known that there are outlier values. These values increased the range of measured GFP expression for one sequence to up to $91\,\%$. Outlier values are most likely present also for sequences with less data points, but in this case they cannot be systematically filtered out with interquartile ranges. When outlier values are included in the algorithm, they affect the accuracy of the result, or, if they do not agree with the measured values of other sequences, make convergence more difficult.

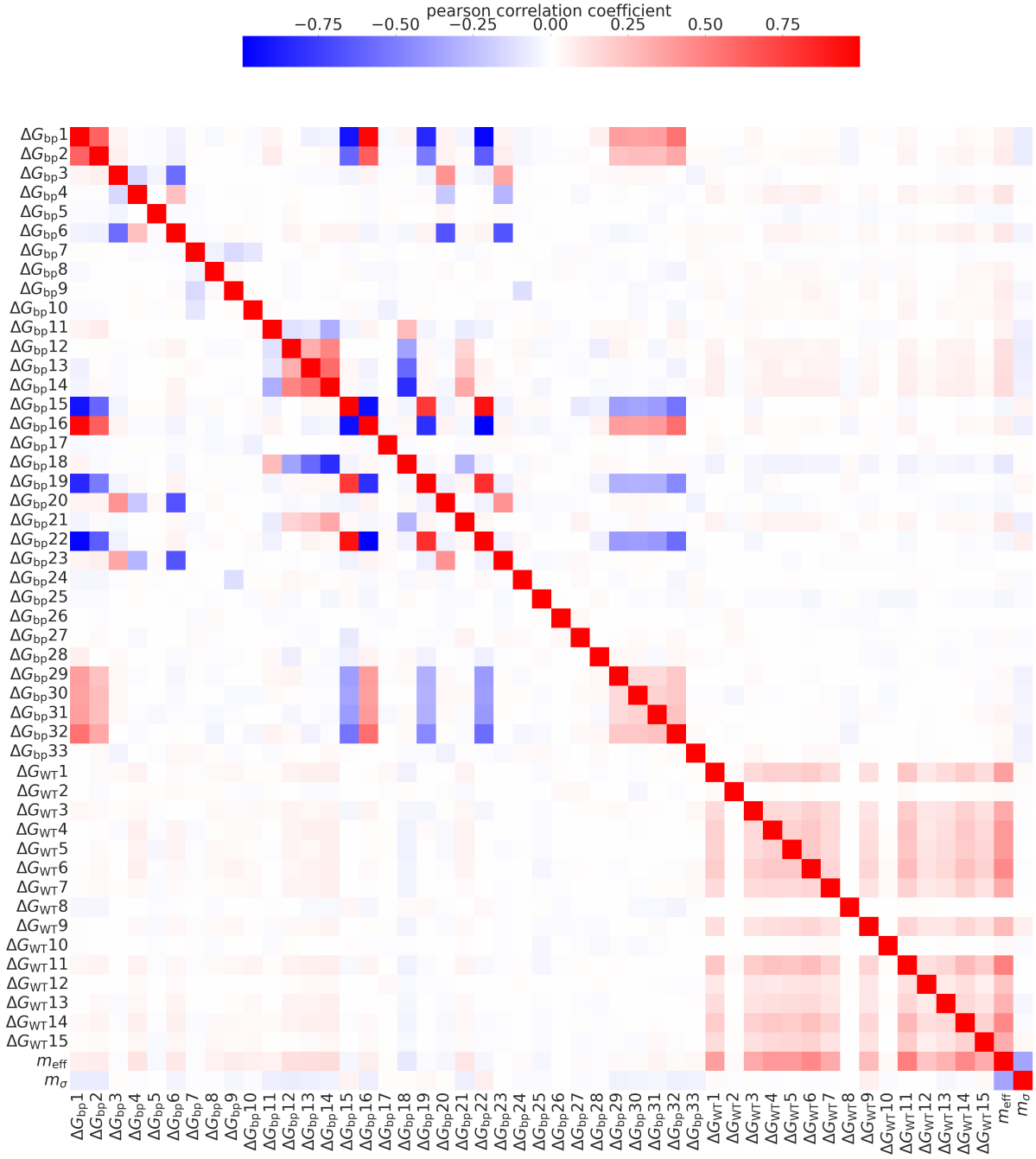

Figure 16: The correlation matrix for all parameters varied in the final Metropolis-algorithm using Pearson correlation, specification of the parameters can be found in table 6.

7.6. Comparison of base-pair free-energies

The basic assumption of the model used to explain frameshifting efficiencies is that the ribosome is slowed down by the mRNA secondary structure, resulting in the two frames being in equilibrium. Thus, the frameshifting efficiency depends only on the free-energy differences of the base-pair changes within the slippery sequence between 0 and -1 frame. Based on these

assumptions, the base-pair free-energy differences should be wild-type independent, as long as a suitable secondary structure is present.

To test if the model is applicable for SARS-CoV, the base-pair free-energy difference distributions determined here based on Mikl et al. [4] are compared to those determined by Bock et al. for dnaX [3]. The base-pair changes that can be used for this purpose, since they are included both here and in the work published by Bock et al. [3], are P1: UA $\rightarrow$ CA, P1: UA $\rightarrow$ AA and A1: AU $\rightarrow$ UU. The results from this thesis and from [3] are compared in figure 17, where the normalized distributions of the free-energy differences associated to these base-pair changes are shown.

First, it is noticeable that the distributions determined for the SARS-CoV base-pair free-energy differences are, by far, wider than the ones obtained for dnaX. This probably stems from the FACS measurement technique, as the range of GFP expression values obtained by repeating the measurements on the same sequence are also very wide, even after eliminating outlier values (SARS WT 11%). Additionally, large standard deviations might result from the way Mikl et al. calculated the GFP expressions based on the measured green fluorescence described in section 5.

The maximum height of the free-energy difference distribution of the base-pair change P1: UA $\rightarrow$ CA determined for SARS-CoV is at approximately 2 kJ/mol. This lies within the 5σ-interval of the free-energy difference determined for dnaX (3.4 kJ/mol $\pm$ 0.3 kJ/mol). Additionally, both distributions largely overlap as can be observed in figure 17a. This result does not disagree with the assumption that the SARS-CoV sequence is in fact stalled long enough, so that the frameshifting efficiency depends solely on the base-pair free-energy differences.

The 95 % confidence interval of free-energy difference distribution resulting from the base-pair change P1: UA $\rightarrow$ AA determined for dnaX, however, does not overlap with the 95 % confidence interval of the distribution determined for the SARS-CoV. This suggests that either the different environments, *in vivo* vs. *in vitro* and a different type of ribosome, result in this difference. Or the assumptions necessary to apply the model cannot be made. Therefore, for SARS-CoV frameshifting efficiencies might not depend only on the base-pair free-energy differences. The model needs to be adapted to further analyse the origin of the offset.

For the last base-pair change, where data is provided for both SARS-CoV and dnaX, A1: UA $\rightarrow$ AA, the free-energy difference determined from SARS-CoV only converged with a lower bound. The absence of an upper bound makes a comparison unfeasible. However, it is noticeable that the lower bound is within the range of the free-energy difference distribution based on dnaX.

Overall, not enough base-pair changes are available for a thorough comparison and it is unclear how measurement problems and possible stray values may effect the comparison. Nevertheless, as the free-energy difference distributions do not overlap for all comparable base-pair changes, the base-pair free-energy differences might not be the only relevant factors determin-

ing frameshifting efficiencies for SARS-CoV.

A possible systematic difference in the base-pair free energy differences for dnaX and SARS-CoV caused by the different environments (*in vivo* versus *in vitro*) was counteracted by including m_{eff} in the Metropolis-Algorithm. Nonetheless the linear approximation might lack in complexity to convert the values measured *in vivo* into values measured *in vitro*.

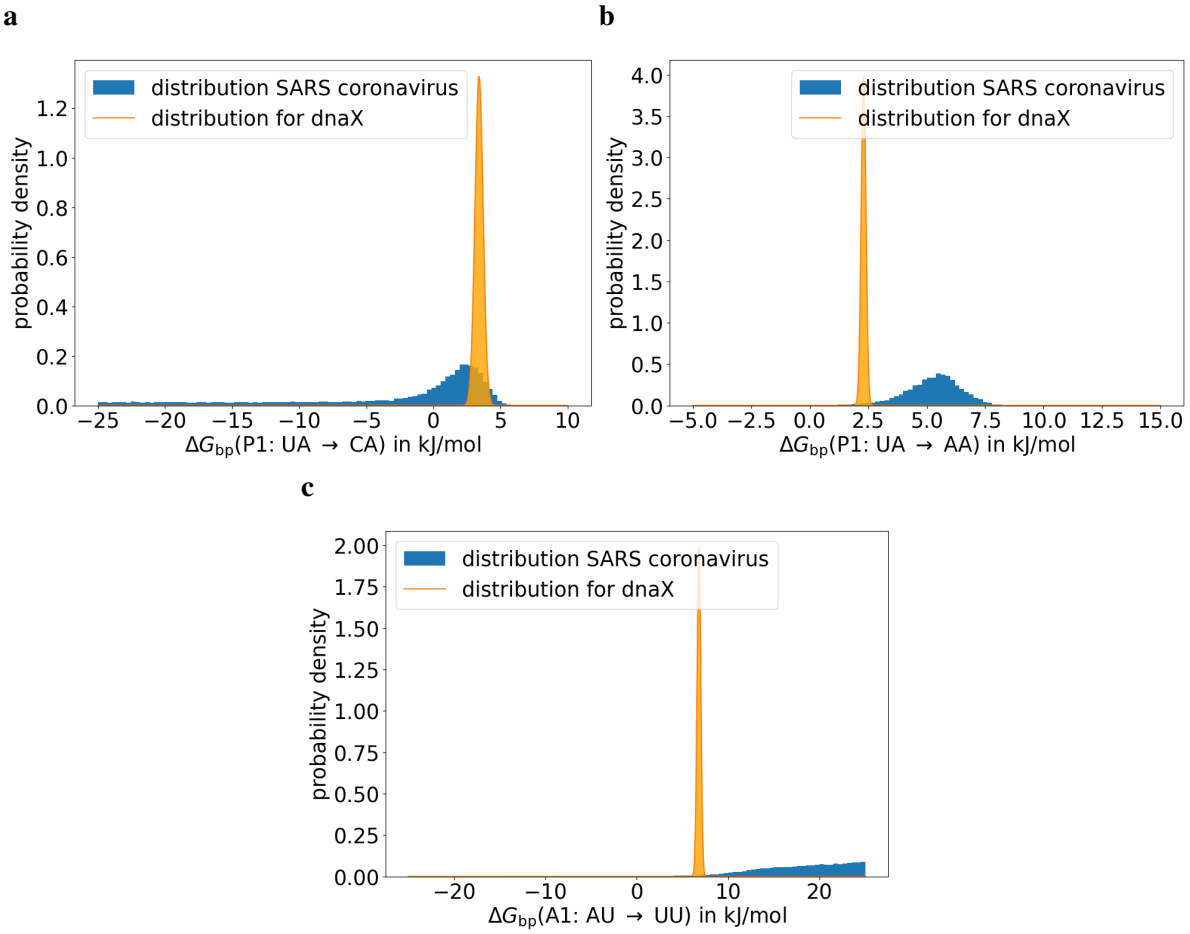

Figure 17: Comparison of base-pair free-energy differences for SARS-CoV (in blue) and dnaX (in orange). **a** Probability density distribution of the base-pair free-energy difference of P1: UA → CA for SARS-CoV and dnaX. **b** Probability density distribution of the base-pair free-energy difference of P1: UA → AA for SARS-CoV and dnaX. **c** Probability density distribution of the base-pair free-energy difference of A1: AU → UU for SARS-CoV and dnaX.

7.7. Prediction of a frameshifting efficiency based on base-pair free-energy differences

In this section, the results for two of the the determined base-pair free-energy difference distributions were cross-validated. For this purpose it was tested, if base-pair free-energy differences can predict a GFP expression for SARS-CoV with the slippery sequence A_AAU_UUA. In order to do so, the relevant base-pair changes in this sequence, P3: UA → AA and A3: AU

→ UU, were determined using the Metropolis-Algorithm (section 6.8) without using any data points for the SARS-CoV sequence with the slippery site A_AAU_UUA.
The results for the free-energy differences determined with and without passing the GFP expression for the tested sequence are in similar ranges (figures 21, 15). The probability density of the free-energy differences of the changes P3: UA → AA and A3: AU → UU, however, has a smaller deviation when including GFP expressions for the slippery sequence A_AAU_UUA. Based on the base-pair free-energy differences the GFP expression for the SARS-CoV sequence with slippery side A_AAU_UUA is predicted (section 6.8) (figure 18).

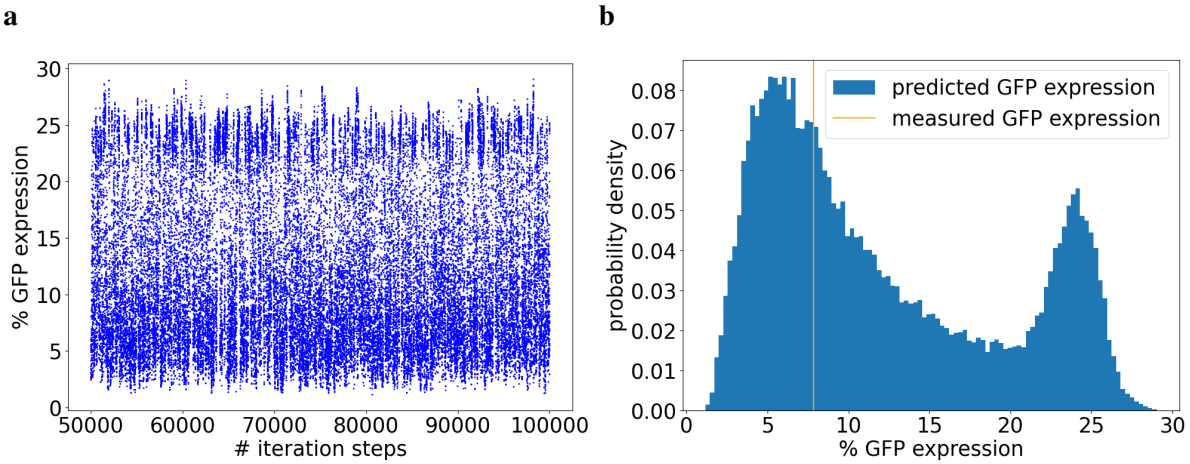

Figure 18: The predicted GFP expression for SARS-CoV with slippery site A_AAU_UUA **a** The predicted GFP expression for the slippery sequence A_AAU_UUA in dependence of the iteration step. **b** In blue: the probability density distribution of the predicted GFP expression for the slippery sequence A_AAU_UUA over the last 50000 iteration steps. In orange: GFP expression measured by Mikl et al. [4] for the same sequence.

The resulting distribution over the predicted GFP expression (figure 18b) has two peaks. The actual measured GFP expression by Mikl et al. of 7.83 % is in the range of the peak on the left hand side. However, this peak has a wide range of about 15 %.
To understand the second peak, the parameters used for the calculation of the GFP expression are plotted against the number of iteration steps (figure 19). All parameters, which result in a GFP expression greater than 20%, and therefore are part of the second peak, are marked in orange. The second peak is, consequentially, a result of the base-pair free-energy values, which are not in a reasonable range around the peak of the distribution of the free-energy differences.
For the slippery sequence A_AAU_UUA a GFP expression range was predictable based on the determined base-pair free-energy differences. However, this range is wide (15 %) and does not give information about an exact percentage. In addition, a convergence of the base-pair free-energy differences to a Gaussian distribution is important for a prediction.

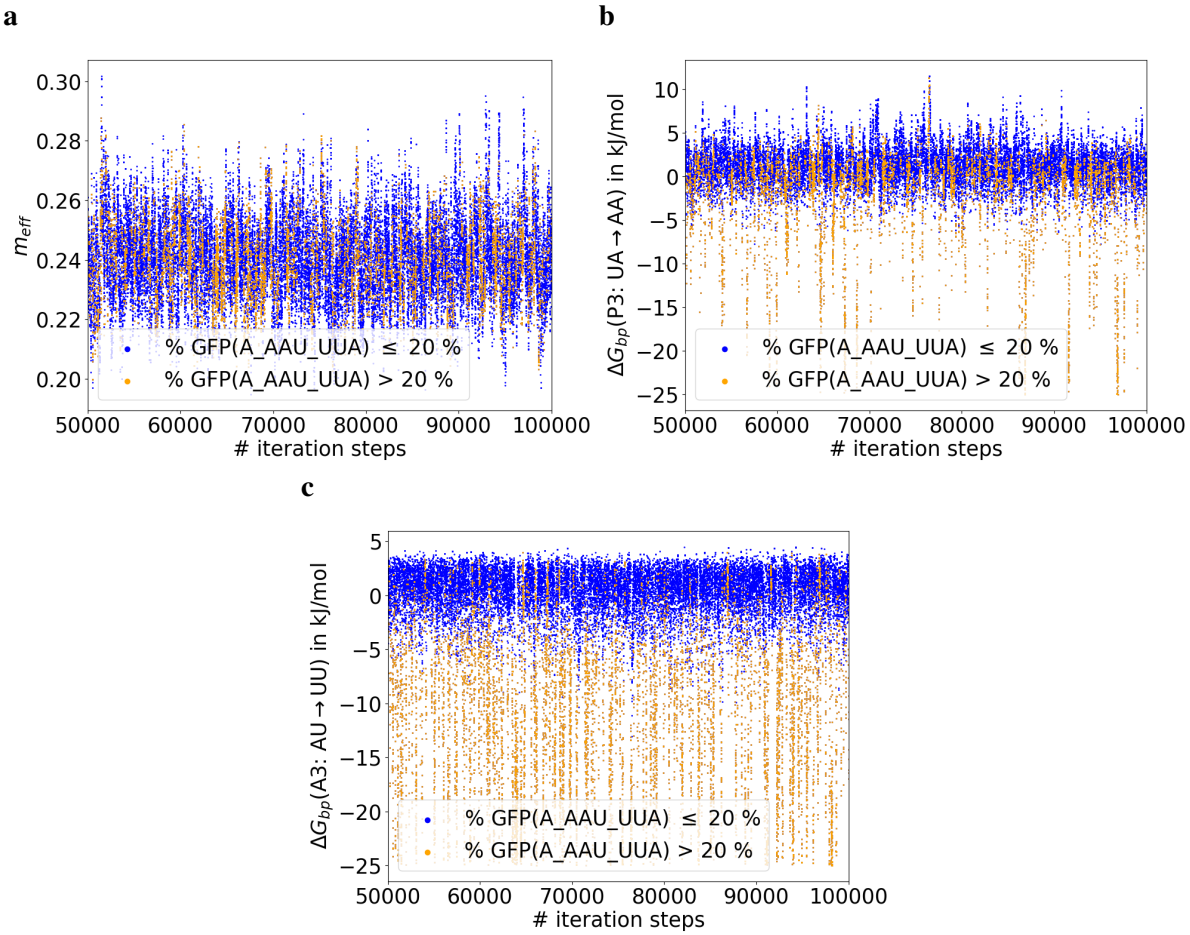

Figure 19: **a** The parameter m_{eff} in dependence of the iteration step. **b** The parameter ΔG_{bp}(P3: UA $\rightarrow$ AA) in dependence of the iteration step. **c** The parameter ΔG_{bp}(A3: AU $\rightarrow$ UU) in dependence of the iteration step. The values generating the second peak in the GFP expression distribution for SARS-CoV with slippery sequence A_AAU_UUA are highlighted in orange.

8. Conclusion

The goal of this bachelor thesis was to determine base-pair free-energy differences for SARS-CoV based on frameshifting efficiencies and to check if the simple free-energy model is applicable to SARS-CoV frameshifting in the human ribosome and under *in-vivo* conditions. To that aim, parameters of the free-energy model and parameters to describe the experiment were obtained using a Bayesian statistics approach. Based on the results, I assessed if the frameshifting efficiencies measured for SARS-CoV *in vivo* can be explained only by base-pair free-energy differences.

The free-energy model applied to the data set of frameshifting efficiencies for 53 different SARS-CoV sequences variations resulted in localized probability distributions of the free-energy differences for 16 out of the total 40 base-pair changes. Additionally, it was possible

to cross-validate two of the base-pair free-energy differences. For this purpose the GFP expression of a SARS sequence variation was predicted with an accuracy of $15\,\%$ based on the two base-pair free-energy differences. The measured GFP expression of the variant lies within this $15\,\%$ range. The model relies on the assumption that frameshifting in SARS-CoV occurs in equilibrium and the frameshifting efficiencies for a sequence only depends on base-pair free-energy differences.

Due to large uncertainties in the measured frameshifting efficiencies arising from large measurement errors and few data points, it was not possible to estimate more frameshifting efficiencies based on the determined base-pair free-energy differences in a cross-validation manner. In addition, for many base-pair free-energy differences only either a lower or an upper boundary could be provided. Furthermore, the determined base-pair free-energy differences do not always match the free-energy differences determined for the same base-pair change in dnaX sequences by Bock et al. [3]. Reasons for not obtaining the same free-energy differences could be that assumptions underlying the model are not valid, i.e. the equilibrium condition or the additivity of base-pair free-energy differences. Further, the free energies could be affected differently by the different chemical environments of the base pairs in the decoding centers of bacterial ribosome (work by Bock et al. [3]) and human ribosomes (this work). The data set by Mikl et al. causes multiple problems in the determination of the resulting base-pair free-energy differences. First of all, the spread of GFP expressions provided for some sequences is large (SARS WT 11 %). This explains why a GFP expression cannot be predicted more precisely. Additionally, many of the GFP expressions lie within the range of background noise. Moreover, for the majority of SARS-CoV sequences there is only one data point of GFP expression resulting in large uncertainties. Therefore, for many base-pair changes the determined free-energy difference is based on a very small amount of data. More data would enable to sort out outlier values as well as limit the amount of base-pair free-energy distributions, where only one boundary can be estimated. This can be assumed based on the difference it makes in section 7.7 to leave one sequence out of the algorithm.

Including the slope parameters m_σ and m_{eff} in the model might have also been a problematic choice. Both of the corresponding relations are simple models, which might lack necessary detail. For example it is known from previous research that *in vivo* frameshifting efficiencies are smaller than those measured *in vitro* [5]. A simple slope parameter m_{eff} might not capture the real relation. Additionally, choosing a linear relation between the frameshifting efficiency and the GFP signal requires the log2(green fluorescence) to be proportional to the frameshifting efficiency. This is not a trivial assumption.

Another way to improve the evaluation would be, to test a more accurate approximated distribution for the background noise. This approximation should than include the GFP expressions measured, which are larger than 2 %.

Most important is to question if the model can be applied for SARS-CoV assuming frameshift-

ing occurs in equilibrium and no other factors but the base-pair free-energy differences determine the frameshifting efficiency. Thus, it should be further analysed, if translation is sufficiently stalled during the frameshifting event by the SARS-CoV pseudoknot. This can be done by including a kinetic contribution into the model. A model like this has been published by Bock et al. to analyse the effects of the stem-loop on the kinetics of frameshifting for dnaX [3]. If kinetics play a role, the base-pair free-energy differences determined in this work would not be the actual free-energy differences but would include other factors.

Additionally, other research has shown that the position of the stop codon is relevant for the frameshifting efficiency of the SARS-CoV [9] and variations of the upstream sequence also influence the frameshifting efficiency [37]. Both qualities do not match the assumption that the frameshifting efficiency for the SARS-CoV can be predicted purely based on the base-pair free-energy differences.

All in all, with a limit in accuracy, a frameshifting efficiency for a variation of SARS-CoV can be predicted based on provided efficiencies for other SARS-CoV sequence variations. For further analysis on the role of base-pair free-energies for PRF in SARS-CoV, a more precise and extensive data set of SARS-CoV frameshifting efficiencies is needed. It would be of great interest to extend the model to take the role kinetics into account as well as other factors that might contribute to frameshifting of the SARS-CoV RNA.

References

1. Louten, J. Virus Structure and Classification. *Essential Human Virology,* 19–29 (2016).
2. Atkins, J. F., Loughran, G., Bhatt, P. R., Firth, A. E. & Baranov, P. V. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Research* **44,** 7007–7078 (2016).
3. Bock, L. V., Caliskan, N., Korniy, N., Peske, F., Rodnina, M. V. & Grubmüller, H. Thermodynamic control of -1 programmed ribosomal frameshifting. *Nature Communications* **10,** 4598 (2019).
4. Mikl, M., Pilpel, Y. & Segal, E. High-throughput interrogation of programmed ribosomal frameshifting in human cells. *Nature Communications* **11,** 3061 (2020).
5. Kim, Y. G., Su, L., Maas, S., O'Neill, A. & Rich, A. Specific mutations in a viral RNA pseudoknot drastically change ribosomal frameshifting efficiency. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 14234–14239 (1999).
6. Lehninger, A. L., Nelson, D. L. & Cox, M. M. *Lehninger principles of biochemistry* 4th ed. ISBN: 0716743396 (W.H. Freeman, New York, 2005).
7. Bolles, M., Donaldson, E. & Baric, R. SARS-CoV and emergent coronaviruses: viral determinants of interspecies transmission. *Current Opinion in Virology* **1,** 624–634 (2011).
8. Plant, E. P., Pérez-Alvarado, G. C., Jacobs, J. L., Mukhopadhyay, B., Hennig, M. & Dinman, J. D. A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biology* **3,** e172 (2005).
9. Bhatt, P. R., Scaiola, A., Loughran, G., Leibundgut, M., Kratzel, A., Meurs, R., Dreos, R., O'Connor, K. M., McMillan, A., Bode, J. W., Thiel, V., Gatfield, D., Atkins, J. F. & Ban, N. Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome. *Science (New York, N.Y.)* **372,** 1306–1313 (2021).
10. Lamers, M. M. & Haagmans, B. L. SARS-CoV-2 pathogenesis. *Nature Reviews Microbiology* **20,** 270–284 (2022).
11. van Hemert, M. J., van den Worm, S. H. E., Knoops, K., Mommaas, A. M., Gorbalenya, A. E. & Snijder, E. J. SARS-coronavirus replication/transcription complexes are membrane-protected and need a host factor for activity in vitro. *PLoS Pathogens* **4,** e1000054 (2008).
12. Sawicki, S. G., Sawicki, D. L. & Siddell, S. G. A contemporary view of coronavirus transcription. *Journal of Virology* **81,** 20–29 (2007).
13. Lodish, H. *Molecular Cell Biology* 9th ed. ISBN: 9781319383602 (Macmillan Learning, New York, 2021).

14. Cooper, G. M. *The cell: A molecular approach* Eighth edition. ISBN: 9781605357072 (Sinauer Associates Oxford University Press, New York and Oxford, 2019).

15. Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K. & Walter, P. *Molecular biology of the cell* Sixth edition. ISBN: 978-0815345244 (Garland Science, Taylor and Francis Group, New York, NY, 2015).

16. Caliskan, N., Peske, F. & Rodnina, M. V. Changed in translation: mRNA recoding by -1 programmed ribosomal frameshifting. *Trends in Biochemical Sciences* **40,** 265–274 (2015).

17. Liao, P.-Y., Choi, Y. S., Dinman, J. D. & Lee, K. H. The many paths to frameshifting: kinetic modelling and analysis of the effects of different elongation steps on programmed -1 ribosomal frameshifting. *Nucleic Acids Research* **39,** 300–312 (2011).

18. Brierley, I., Gilbert, R. J. & Pennell, S. in *Recoding: Expansion of Decoding Rules Enriches Gene Expression* (eds Atkins, J. F. & Gesteland, R. F.) 149–174 (Springer New York, New York, NY, 2010). ISBN: 978-0-387-89382-2.

19. Chen, J., Petrov, A., Johansson, M., Tsai, A., O'Leary, S. E. & Puglisi, J. D. Dynamic pathways of -1 translational frameshifting. *Nature* **512,** 328–332 (2014).

20. Dos Ramos, F., Carrasco, M., Doyle, T. & Brierley, I. Programmed -1 ribosomal frameshifting in the SARS coronavirus. *Biochemical Society Transactions* **32,** 1081–1083 (2004).

21. Nelson, P. *Biological physics: Energy, information, life* Student edition. ISBN: 9780578687025 (Chiliagon Science, Philadelphia, PA, 2020).

22. Van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., *et al.* Bayesian statistics and modelling. *Nature Reviews Methods Primers* **1,** 1–26 (2021).

23. Bolstad, W. M. & Curran, J. *Introduction to Bayesian statistics* Third edition. ISBN: 1118593162 (Wiley, Hoboken, New Jersey, 2017).

24. Tappe, S. *Einführung in die Wahrscheinlichkeitstheorie* ISBN: 9783642375446 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).

25. McElreath, R. *Statistical rethinking: A Bayesian course with examples in R and Stan* ISBN: 9781482253443 (CRC Press, Boca Raton, London, and New York, 2016).

26. Kroese, D. P., Taimre, T. & Botev, Z. I. *Handbook of Monte Carlo methods* Online-Ausg. ISBN: 9780470177938 (Wiley, Hoboken, NJ, 2011).

27. Koch, K.-R. Monte Carlo methods. *GEM - International Journal on Geomathematics* **9,** 117–143 (2018).

28. Herzenberg, L. A., Sweet, R. G. & Herzenberg, L. A. Fluorescence-activated cell sorting. *Scientific American* **234,** 108–118 (1976).

29. Henze, N. *Stochastik für Einsteiger: Eine Einführung in die faszinierende Welt des Zufalls* 12., verbesserte und erweiterte Auflage. ISBN: 9783658220440 (Springer Spektrum, Wiesbaden, 2018).

30. Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research* **37,** D93–D97 (2009).

31. Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F. & Pütz, J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research* **37,** D159–D162 (2009).

32. Hieber, M. *Analysis I* 1. Aufl. 2018. ISBN: 9783662575383 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2018).

33. Dulude, D., Berchiche, Y. A., Gendron, K., Brakier-Gingras, L. & Heveker, N. Decreasing the frameshift efficiency translates into an equivalent reduction of the replication of the human immunodeficiency virus type 1. *Virology* **345,** 127–136 (2006).

34. Léger, M., Dulude, D., Steinberg, S. V. & Brakier-Gingras, L. The three transfer RNAs occupying the A, P and E sites on the ribosome are involved in viral programmed -1 ribosomal frameshift. *Nucleic Acids Research* **35,** 5581–5592 (2007).

35. Biswas, P., Jiang, X., Pacchia, A. L., Dougherty, J. P. & Peltz, S. W. The human immunodeficiency virus type 1 ribosomal frameshifting site is an invariant sequence determinant and an important target for antiviral therapy. *Journal of Virology* **78,** 2082–2087 (2004).

36. Hinton, P. R. *Statistics explained* 3. ed. ISBN: 9781848723115 (Routledge, London, 2014).

37. Su, M.-C., Chang, C.-T., Chu, C.-H., Tsai, C.-H. & Chang, K.-Y. An atypical RNA pseudoknot stimulator and an upstream attenuation signal for -1 ribosomal frameshifting of SARS coronavirus. *Nucleic Acids Research* **33,** 4265–4275 (2005).

A. Derivation of free energy $G_{a,l}$

$$
\begin{aligned}
G_{a,l} &= \langle E_a \rangle_l - TS_{a,l} \qquad (54)\\
&= \sum_j^{N_l} P_{j,l} E_j + Tk_B \sum_j^{N_l} P_{j,l} \ln(P_{j,l}) \\
&= \sum_j^{N_l} P_{j,l} E_j + Tk_B \sum_j^{N_l} P_{j,l} \ln\left(\frac{P_j}{P_l}\right) \\
&= \sum_j^{N_l} P_{j,l} E_j + Tk_B \sum_j^{N_l} P_{j,l} \ln\left(\frac{\exp\left(\frac{-E_j}{k_B T}\right)}{Z P_l}\right) \\
&= \sum_j^{N_l} P_{j,l} E_j + Tk_B \sum_j^{N_l} P_{j,l} \left(\ln\left(\exp\left(\frac{-E_j}{k_B T}\right)\right) - \ln\left(Z P_l\right)\right) \\
&= \sum_j^{N_l} P_{j,l} E_j - \frac{Tk_B}{Tk_B} \sum_j^{N_l} P_{j,l} E_j - Tk_B \sum_j^{N_l} P_{j,l} \ln\left(Z P_l\right) \\
&= -Tk_B \sum_j^{N_l} P_{j,l} \ln\left(Z P_l\right) \\
&= -Tk_B \ln\left(Z_l\right) \sum_j^{N_l} \frac{P_j}{P_l} \\
&= -Tk_B \ln\left(Z_l\right)
\end{aligned}
$$

B. Additional tables and figures

Table 1: Codon - anticodon pairings of the codons in 0 frame of the analysed SARS-CoV variants read from 5' to 3' of the mRNA.

Codon	Anticodon
AAA	UUt
AAG	UUC
AAU	UUA
AAC	UUQ
CCA	GGU
CCG	GGC
CCU	GGA
GGA	CCU
GGC	CCG
GGU	CCA
UUA	AAU
UUC	AA#
UUG	AAC
CCC	GGG
GGG	CCC
UUU	AAA
AUA	UAU
UCA	AGU
CAC	GUG
UAC	AP9
GAC	CUG
AUC	UAG

Table 2: Slippery sequences in SARS-CoV variants and the base-pair changes in case of a frameshifting event.

Slippery site	Base-pair changes
AAAAAAG	A3: GC $\rightarrow$ AC
AAAAAAU	A3: UA $\rightarrow$ AA
AAACCCA	P3: CQ $\rightarrow$ AQ, A3: AU $\rightarrow$ CU
AAACCCG	P3: CQ $\rightarrow$ AQ, A3: GC $\rightarrow$ CC
AAACCCU	P3: CQ $\rightarrow$ AQ, A3: UA $\rightarrow$ CA
AAAGGGA	P3: GC $\rightarrow$ AC, A3: AU $\rightarrow$ GU
AAAGGGC	P3: GC $\rightarrow$ AC, A3: CG $\rightarrow$ GG
AAAGGGU	P3: GC $\rightarrow$ AC, A3: UA $\rightarrow$ GA

Continued on next page

Table 2 – continued from previous page

Slippery site	Base-pair changes
AAAUUUA	P3: UA → AA, A3: AU → UU
AAAUUUC	P3: UA → AA, A3: C# → U#
AAAUUUG	P3: UA → AA, A3: GC → UC
CCCAAAA	P3: AU → CU
CCCAAAC	P3: AU → CU, A3: CQ → AQ
CCCAAAG	P3: AU → CU, A3: GC → AC
CCCCCCA	A3: AU → CU
CCCCCCG	A3: GC → CC
CCCCCCU	A3: UA → CA
CCCGGGA	P3: GC → CC, A3: AU → GU
CCCGGGG	P3: GC → CC
CCCGGGU	P3: GC → CC, A3: UA → GA
CCCUUUA	P3: UA → CA, A3: AU → UU
CCCUUUC	P3: UA → CA, A3: C# → U#
CCCUUUG	P3: UA → CA, A3: GC → UC
GGGAAAC	P3: AU → GU, A3: CQ → AQ
GGGAAAG	P3: AU → GU, A3: GC → AC
GGGCCCA	P3: CG → GG, A3: AU → CU
GGGCCCG	P3: CG → GG, A3: GC → CC
GGGCCCU	P3: CG → GG, A3: UA → CA
GGGGGGA	A3: AU → GU
GGGGGGC	A3: CG → GG
GGGGGGU	A3: UA → GA
GGGUUUG	P3: UA → GA, A3: GC → UC
GGGUUUU	P3: UA → GA
UUUAAAG	P3: AU → UU, A3: GC → AC
UUUAAAU	P3: AU → UU, A3: UA → AA
UUUCCCA	P3: C# → U#, A3: AU → CU
UUUCCCC	P3: C# → U#
UUUCCCU	P3: C# → U#, A3: UA → CA
UUUGGGC	P3: GC → UC, A3: CG → GG
UUUGGGG	P3: GC → UC
UUUGGGU	P3: GC → UC, A3: UA → GA
UUUUUUA	A3: AU → UU

Continued on next page

Table 2 – continued from previous page

Slippery site	Base-pair changes
UUUUUUG	A3: GC → UC
AUUAAAC	P1: UA → AA, P3: AU → UU, A3: CQ → AQ
UAUAAAC	P1: AU → UU, P2: UA → AA, P3: AU → UU, A3: CQ → AQ
UUCAAAC	P2: CG → UG, P3: AU → CU, A3: CQ → AQ
UUUAAAC	P3: AU → UU, A3: CQ → AQ
UUUUAAC	A1: AU → UU, A3: CQ → AQ
UUUACAC	P3: AU → UU, A1: CG → AG, A2: AU → CU, A3: CG → AG
UUUAUAC	P3: AU → UU, A1: UA → AA, A2: AP → UP, A3: C9 → A9
UUUAGAC	P3: AU → UU, A1: GC → AC, A2: AU → GU, A3: CG → AG
UUUAAUC	P3: AU → UU, A2: UA → AA, A3: CG → UG
CUUAAAC	P1: UA → CA, P3: AU → UU, A3: CQ → AQ

Table 3: Parameter-set for the Metropolis-Algorithm for analysis of the GFP signal standard deviation.

Parameter	Meaning	Startvalue C_0	Proposal distribution - standard deviation
C^0	σ_{S_j}	3 %	0.3 %
C^1	μ_{S_j}	1 %	0.3 %

Table 4: Sequences of HIV HXB2 variants used for comparison of GFP signal and frameshifting efficiency.

Paper	HIV HXB2 sequence 3 nt prior of slipperey-site and slippery site
Dulude et al. [33]	UAAUUUUUUA
	UAAUUUUUUU
Leger et al. [34]	UAAUUUUUUA
	UAAAUUUUUA
	UAUUUUUUUA
	UAACCCUUUA
	UAAUUUAAAC
Biswas et al. [35]	UAAUUUUUUA
	UAAUUUAAAC
	UAAGGGUUUA
	UAAAAAUUUA

Table 5: Parameter-set for the Metropolis-Algorithm determining the base-pair free-energy differences.

Parameter	Meaning	Startvalue C_0	Proposal distribution - standard deviation
C^0	ΔG_{bp}(A3: GC $\rightarrow$ AC)	1 kJ/mol	3 kJ/mol
C^1	ΔG_{bp}(A3: UA $\rightarrow$ AA)	1 kJ/mol	3 kJ/mol
C^2	ΔG_{bp}(P3: CQ $\rightarrow$ AQ)	1 kJ/mol	3 kJ/mol
C^3	ΔG_{bp}(A3: AU $\rightarrow$ CU)	1 kJ/mol	3 kJ/mol
C^4	ΔG_{bp}(A3: GC $\rightarrow$ CC)	1 kJ/mol	3 kJ/mol
C^5	ΔG_{bp}(A3: UA $\rightarrow$ CA)	1 kJ/mol	3 kJ/mol
C^6	ΔG_{bp}(P3: GC $\rightarrow$ AC)	1 kJ/mol	3 kJ/mol
C^7	ΔG_{bp}(A3: AU $\rightarrow$ GU)	1 kJ/mol	3 kJ/mol
C^8	ΔG_{bp}(A3: CG $\rightarrow$ GG)	1 kJ/mol	3 kJ/mol
C^9	ΔG_{bp}(A3: UA $\rightarrow$ GA)	1 kJ/mol	3 kJ/mol
C^{10}	ΔG_{bp}(P3: UA $\rightarrow$ AA)	1 kJ/mol	3 kJ/mol
C^{11}	ΔG_{bp}(A3: AU $\rightarrow$ UU)	1 kJ/mol	3 kJ/mol
C^{12}	ΔG_{bp}(A3: C# $\rightarrow$ U#)	1 kJ/mol	3 kJ/mol
C^{13}	ΔG_{bp}(A3: GC $\rightarrow$ UC)	1 kJ/mol	3 kJ/mol
C^{14}	ΔG_{bp}(P3: AU $\rightarrow$ CU)	1 kJ/mol	3 kJ/mol
C^{15}	ΔG_{bp}(A3: CQ $\rightarrow$ AQ)	1 kJ/mol	3 kJ/mol
C^{16}	ΔG_{bp}(P3: GC $\rightarrow$ CC)	1 kJ/mol	3 kJ/mol
C^{17}	ΔG_{bp}(P3: UA $\rightarrow$ CA)	1 kJ/mol	3 kJ/mol
C^{18}	ΔG_{bp}(P3: AU $\rightarrow$ GU)	1 kJ/mol	3 kJ/mol
C^{19}	ΔG_{bp}(P3: CG $\rightarrow$ GG)	1 kJ/mol	3 kJ/mol
C^{20}	ΔG_{bp}(P3: UA $\rightarrow$ GA)	1 kJ/mol	3 kJ/mol
C^{21}	ΔG_{bp}(P3: AU $\rightarrow$ UU)	1 kJ/mol	3 kJ/mol
C^{22}	ΔG_{bp}(P3: C# $\rightarrow$ U#)	1 kJ/mol	3 kJ/mol
C^{23}	ΔG_{bp}(P3: GC $\rightarrow$ UC)	1 kJ/mol	3 kJ/mol
C^{24}	ΔG_{bp}(P1: UA $\rightarrow$ AA)	1 kJ/mol	3 kJ/mol
C^{25}	ΔG_{bp}(P1: AU $\rightarrow$ UU, P2: UA $\rightarrow$ AA)	1 kJ/mol	3 kJ/mol
C^{26}	ΔG_{bp}(P2: CG $\rightarrow$ UG)	1 kJ/mol	3 kJ/mol
C^{27}	ΔG_{bp}(A1: AU $\rightarrow$ UU)	1 kJ/mol	3 kJ/mol
C^{28}	ΔG_{bp}(A1: CG $\rightarrow$ AG, A2: AU $\rightarrow$ CU, A3: CG $\rightarrow$ AG)	1 kJ/mol	3 kJ/mol

Continued on next page

Table 5 – continued from previous page

Parameter	Meaning	Startvalue C_0	Proposal distribution - standard deviation
C^{29}	ΔG_{bp}(A1: UA $\rightarrow$ AA, A2: AP $\rightarrow$ UP, A3: C9 $\rightarrow$ A9)	1 kJ/mol	3 kJ/mol
C^{30}	ΔG_{bp}(A1: GC $\rightarrow$ AC, A2: AU $\rightarrow$ GU, A3: CG $\rightarrow$ AG)	1 kJ/mol	3 kJ/mol
C^{31}	ΔG_{bp}(A2: UA $\rightarrow$ AA, A3: CG $\rightarrow$ UG)	1 kJ/mol	3 kJ/mol
C^{32}	ΔG_{bp}(P1: UA $\rightarrow$ CA)	1 kJ/mol	3 kJ/mol
C^{33}	ΔG_{WT}(PRRSV) (virus)	0.5 kJ/mol	1 kJ/mol
C^{34}	ΔG_{WT}(CCR5) (human protein)	2 kJ/mol	1 kJ/mol
C^{35}	ΔG_{WT}(SARS-CoV) (virus)	0.5 kJ/mol	1 kJ/mol
C^{36}	ΔG_{WT}(SIV) (virus)	0.5 kJ/mol	1 kJ/mol
C^{37}	ΔG_{WT}(SRV1) (virus)	0.5 kJ/mol	1 kJ/mol
C^{38}	ΔG_{WT}(PLRV) (virus)	0.5 kJ/mol	1 kJ/mol
C^{39}	ΔG_{WT}(OAZ1) (human protein)	0.5 kJ/mol	1 kJ/mol
C^{40}	ΔG_{WT}(HSV) (virus)	2 kJ/mol	1 kJ/mol
C^{41}	ΔG_{WT}(PEG10) (human protein)	0.5 kJ/mol	1 kJ/mol
C^{42}	ΔG_{WT}(RSV) (virus)	2 kJ/mol	1 kJ/mol
C^{43}	ΔG_{WT}(HERV K10) (virus)	0.5 kJ/mol	1 kJ/mol
C^{44}	ΔG_{WT}(influenza) (virus)	0.5 kJ/mol	1 kJ/mol
C^{45}	ΔG_{WT}(HTLV) (virus)	0.5 kJ/mol	1 kJ/mol
C^{46}	ΔG_{WT}(HIV HXB2) (virus)	0.5 kJ/mol	1 kJ/mol
C^{47}	ΔG_{WT}(WNV) (virus)	0.5 kJ/mol	1 kJ/mol
C^{48}	m_{eff}	-0.6	0.08
C^{49}	m_{σ}	4.5	0.007

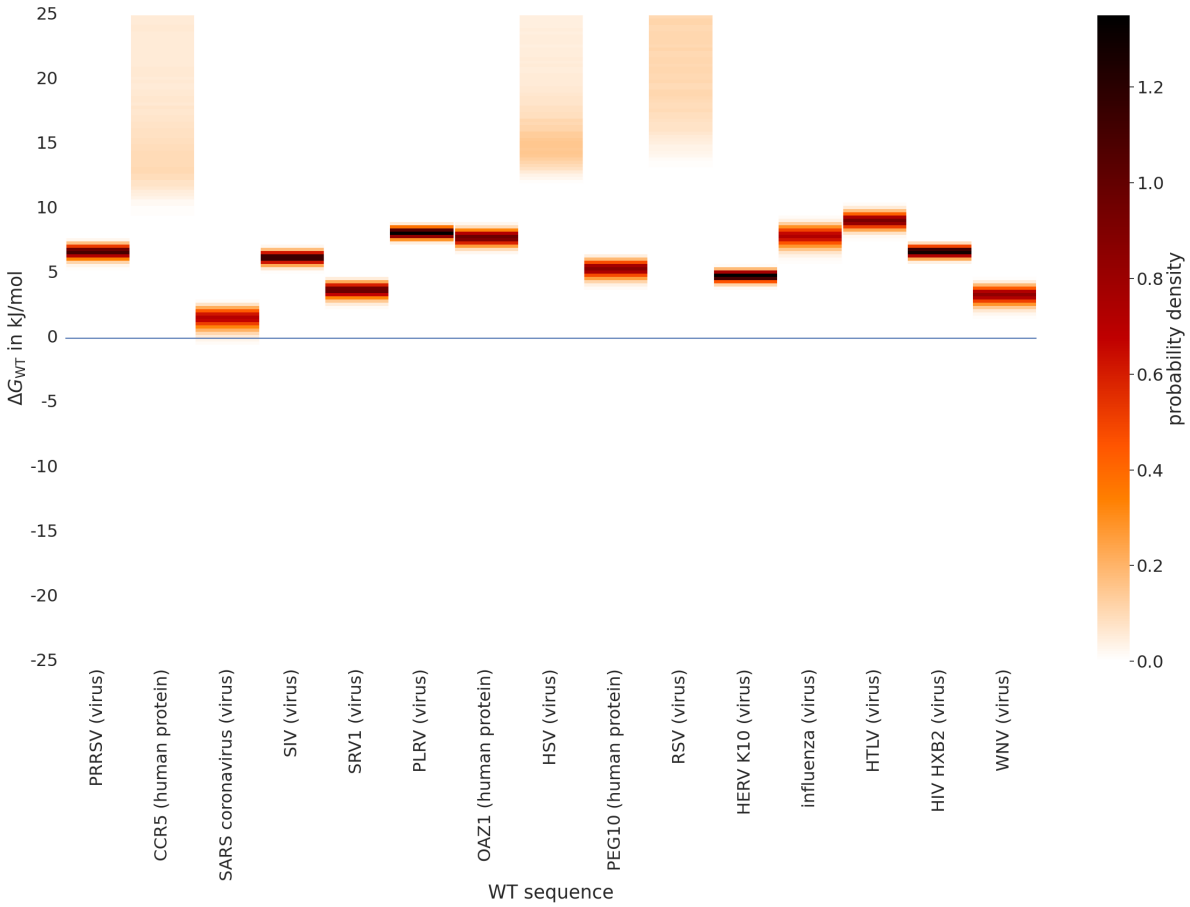

Figure 20: Probability distributions of all WT free-energy differences obtained from histograms over the last 50000 iteration steps of the Metropolis-Algorithm.

Table 6: Legend of the abbreviations for the correlation matrix.

Label	Meaning
$\Delta G_{\mathrm{bp}}1$	ΔG_{bp}(A3: GC $\rightarrow$ AC)
$\Delta G_{\mathrm{bp}}2$	ΔG_{bp}(A3: UA $\rightarrow$ AA)
$\Delta G_{\mathrm{bp}}3$	ΔG_{bp}(P3: CQ $\rightarrow$ AQ)
$\Delta G_{\mathrm{bp}}4$	ΔG_{bp}(A3: AU $\rightarrow$ CU)
$\Delta G_{\mathrm{bp}}5$	ΔG_{bp}(A3: GC $\rightarrow$ CC)
$\Delta G_{\mathrm{bp}}6$	ΔG_{bp}(A3: UA $\rightarrow$ CA)
$\Delta G_{\mathrm{bp}}7$	ΔG_{bp}(P3: GC $\rightarrow$ AC)
$\Delta G_{\mathrm{bp}}8$	ΔG_{bp}(A3: AU $\rightarrow$ GU)
$\Delta G_{\mathrm{bp}}9$	ΔG_{bp}(A3: CG $\rightarrow$ GG)
$\Delta G_{\mathrm{bp}}10$	ΔG_{bp}(A3: UA $\rightarrow$ GA)
$\Delta G_{\mathrm{bp}}11$	ΔG_{bp}(P3: UA $\rightarrow$ AA)

Continued on next page

Table 6 – continued from previous page

Label	Meaning
$\Delta G_{\mathrm{bp}}12$	ΔG_{bp}(A3: AU $\rightarrow$ UU)
$\Delta G_{\mathrm{bp}}13$	ΔG_{bp}(A3: C# $\rightarrow$ U#)
$\Delta G_{\mathrm{bp}}14$	ΔG_{bp}(A3: GC $\rightarrow$ UC)
$\Delta G_{\mathrm{bp}}15$	ΔG_{bp}(P3: AU $\rightarrow$ CU)
$\Delta G_{\mathrm{bp}}16$	ΔG_{bp}(A3: CQ $\rightarrow$ AQ)
$\Delta G_{\mathrm{bp}}17$	ΔG_{bp}(P3: GC $\rightarrow$ CC)
$\Delta G_{\mathrm{bp}}18$	ΔG_{bp}(P3: UA $\rightarrow$ CA)
$\Delta G_{\mathrm{bp}}19$	ΔG_{bp}(P3: AU $\rightarrow$ GU)
$\Delta G_{\mathrm{bp}}20$	ΔG_{bp}(P3: CG $\rightarrow$ GG)
$\Delta G_{\mathrm{bp}}21$	ΔG_{bp}(P3: UA $\rightarrow$ GA)
$\Delta G_{\mathrm{bp}}22$	ΔG_{bp}(P3: AU $\rightarrow$ UU)
$\Delta G_{\mathrm{bp}}23$	ΔG_{bp}(P3: C# $\rightarrow$ U#)
$\Delta G_{\mathrm{bp}}24$	ΔG_{bp}(P3: GC $\rightarrow$ UC)
$\Delta G_{\mathrm{bp}}25$	ΔG_{bp}(P1: UA $\rightarrow$ AA)
$\Delta G_{\mathrm{bp}}26$	ΔG_{bp}(P1: AU $\rightarrow$ UU, P2: UA $\rightarrow$ AA)
$\Delta G_{\mathrm{bp}}27$	ΔG_{bp}(P2: CG $\rightarrow$ UG)
$\Delta G_{\mathrm{bp}}28$	ΔG_{bp}(A1: AU $\rightarrow$ UU)
$\Delta G_{\mathrm{bp}}29$	ΔG_{bp}(A1: CG $\rightarrow$ AG, A2: AU $\rightarrow$ CU, A3: CG $\rightarrow$ AG)
$\Delta G_{\mathrm{bp}}30$	ΔG_{bp}(A1: UA $\rightarrow$ AA, A2: AP $\rightarrow$ UP, A3: C9 $\rightarrow$ A9)
$\Delta G_{\mathrm{bp}}31$	ΔG_{bp}(A1: GC $\rightarrow$ AC, A2: AU $\rightarrow$ GU, A3: CG $\rightarrow$ AG)
$\Delta G_{\mathrm{bp}}32$	ΔG_{bp}(A2: UA $\rightarrow$ AA, A3: CG $\rightarrow$ UG)
$\Delta G_{\mathrm{bp}}33$	ΔG_{bp}(P1: UA $\rightarrow$ CA)
$\Delta G_{\mathrm{WT}}1$	ΔG_{WT}(PRRSV) (virus)
$\Delta G_{\mathrm{WT}}2$	ΔG_{WT}(CCR5) (human protein)
$\Delta G_{\mathrm{WT}}3$	ΔG_{WT}(SARS-CoV) (virus)
$\Delta G_{\mathrm{WT}}4$	ΔG_{WT}(SIV) (virus)
$\Delta G_{\mathrm{WT}}5$	ΔG_{WT}(SRV1) (virus)
$\Delta G_{\mathrm{WT}}6$	ΔG_{WT}(PLRV) (virus)
$\Delta G_{\mathrm{WT}}7$	ΔG_{WT}(OAZ1) (human protein)
$\Delta G_{\mathrm{WT}}8$	ΔG_{WT}(HSV) (virus)
$\Delta G_{\mathrm{WT}}9$	ΔG_{WT}(PEG10) (human protein)
$\Delta G_{\mathrm{WT}}10$	ΔG_{WT}(RSV) (virus)
$\Delta G_{\mathrm{WT}}11$	ΔG_{WT}(HERV K10) (virus)
$\Delta G_{\mathrm{WT}}12$	ΔG_{WT}(influenza) (virus)

Continued on next page

Table 6 – continued from previous page

Label	Meaning
$\Delta G_{\mathrm{WT}}13$	ΔG_{WT}(HTLV) (virus)
$\Delta G_{\mathrm{WT}}14$	ΔG_{WT}(HIV HXB2) (virus)
$\Delta G_{\mathrm{WT}}15$	ΔG_{WT}(WNV) (virus)

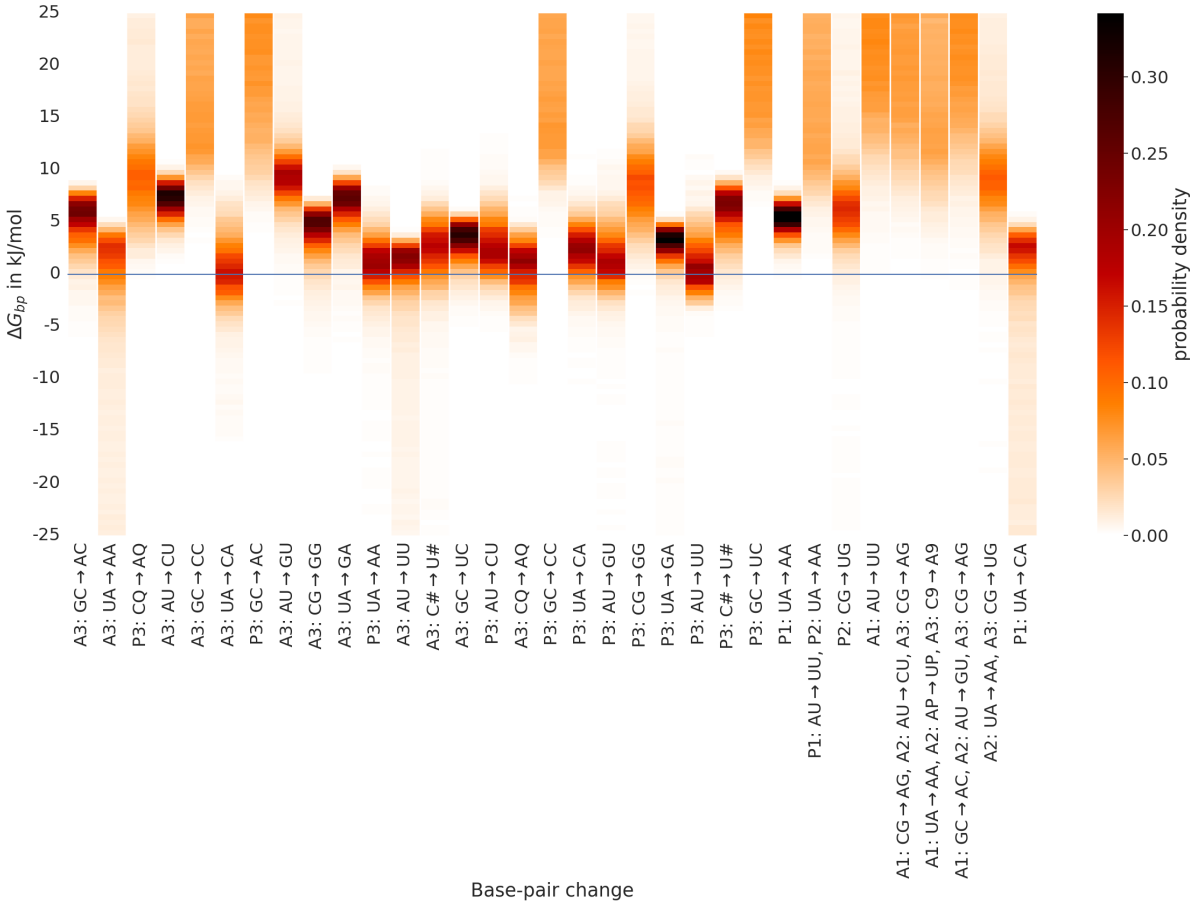

Figure 21: Probability distributions of all base-pair free-energy differences obtained from histograms over the last 50000 iteration steps of the Metropolis-Algorithm (without slippery sequence A_AAU_UUA). For bars labeled with multiple base-pair changes, the sum of their base-pair free energy differences was considered as a parameter in the likelihood.