Inaugural Dissertation

for

Obtaining the Doctoral Degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht-Karls University of Heidelberg

Presented by

M.Sc. Federico Marotta

Born in Torino (Italy)

Oral Examination: 14th October 2025

Characterization, evolution, and dynamics of cryo-ET-derived macromolecular assemblies in *Mycoplasma pneumoniae*

Referees:
Prof. Dr. Robert Russell
Dr. Maria Zimmermann-Kogadeeva

Abstract

Mycoplasma pneumoniae, a genome-reduced pathogenic bacterium, is as an ideal model for studying the concept of a minimal cell. This organism benefits from a long history of analyses that encompass many biological processes and molecular types, providing a comprehensive foundation for computational studies. The advent of high-resolution cryogenic electron tomography (cryo-ET) has placed *M. pneumoniae* in a unique position to bridge *in-situ* imaging with bioinformatics and systems biology. This unprecedented direct visual access to protein complexes and other macromolecular assemblies presents novel opportunities and challenges. While some large and abundant molecular components can be readily identified in the tomograms, deciphering their functions requires sophisticated computational approaches that make use of the wealth of existing data. The recent introduction of AlphaFold2 has revolutionized the field of structural biology by providing high-confidence structural models for virtually every protein. In this thesis, I leverage this new paradigm to annotate the function of a newly identified protein complex observed exclusively in *M. pneumoniae* cryo-ET data. I show that access to protein structures dramatically improves the accuracy of functional annotations, particularly when structures are segmented into their constituent domains. One of the largest macromolecular assemblies in the cell is the ribosome, with its many interacting partners. Utilizing a novel cryo-ET dataset that captures the abundances of intermediate states of ribosomes engaged in protein synthesis, I build a kinetic model of the translation-elongation cycle. Furthermore, I develop and implement a generalizable method to calibrate the kinetic rates of biological processes by integrating cryo-ET data with know rates from a reference system. This thesis advances structural bioinformatics by designing innovative analytical frameworks downstream of cryo-ET. These frameworks enable a better annotation of the function of proteins from their structure, as well as a better understanding of the dynamics of molecular processes from static cryo-ET snapshots.

Zusammenfassung

Mycoplasma pneumoniae, ein genomreduziertes pathogenes Bakterium, ist ein ideales Modell zur Untersuchung des Konzepts einer minimalen Zelle. Dieses Organismus profitiert von einer langen Geschichte von Analysen, die viele biologische Prozesse und Molekültypen umfassen und eine umfassende Grundlage für computergestützte Studien bieten. Der Aufstieg der hochauflösenden cryo-ET hat *M. pneumoniae* in eine einzigartige Position gebracht, um *in-situ* Bildgebung mit Bioinformatik und Systembiologie zu verbinden. Dieser beispiellose direkte visuelle Zugang zu Proteinkomplexen und anderen makromolekularen Strukturen bietet neue Möglichkeiten und Herausforderungen. Während einige große und reichlich vorhandene molekulare Komponenten in den Tomogrammen leicht identifiziert werden können, erfordert die Entschlüsselung ihrer Funktionen anspruchsvolle rechnergestützte Ansätze, die den Reichtum an vorhandenen Daten nutzen. Die kürzliche Einführung von AlphaFold2 hat das Gebiet der Strukturbiologie revolutioniert, indem es hochzuverlässige Strukturmodelle für praktisch jedes Protein bereitstellt. In dieser Arbeit nutze ich dieses neue Paradigma, um die Funktion eines neu identifizierten Proteinkomplexes zu annotieren, der ausschließlich in *M. pneumoniae* cryo-ET-Daten beobachtet wird. Ich zeige, dass der Zugang zu Proteinstrukturen die Genauigkeit funktioneller Annotationen erheblich verbessert, insbesondere wenn Strukturen in ihre konstituierenden Domänen segmentiert werden. Eine der größten makromolekularen Strukturen in der Zelle ist das Ribosom mit seinen vielen interagierenden Partnern. Unter Verwendung eines neuartigen cryo-ET-Datensatzes, der die Häufigkeiten von Zwischenzuständen von Ribosomen erfasst, die an der Proteinsynthese beteiligt sind, erstelle ich ein kinetisches Modell des Translations-Elongationszyklus. Darüber hinaus entwickle und implementiere ich eine verallgemeinerbare Methode zur Kalibrierung der kinetischen Raten biologischer Prozesse, indem ich cryo-ET-Daten mit bekannten Raten aus einem Referenzsystem integriere. Diese Arbeit fördert die strukturelle Bioinformatik, indem sie innovative analytische Rahmenwerke im Anschluss an cryo-ET entwirft. Diese Rahmenwerke ermöglichen eine bessere Annotation der Funktion von Proteinen aus ihrer Struktur sowie ein besseres Verständnis der Dynamik molekularer Prozesse aus statischen cryo-ET-Schnappschüssen.

Contents

List of Figures

List of Tables

List of acronyms

AF AlphaFold

AI artificial intelligence

CASP Critical Assessment of Structure Prediction

ChIP-seq Chromatin immunoprecipitation followed by sequencing

E. coli *Escherichia coli*

cryo-EM cryogenic electron microscopy

cryo-ET cryogenic electron tomography

CTMC continuous-time Markov chain

DSSP Dictionary of Secondary Structure in Proteins

EMBL European Molecular Biology Laboratory

ENA European Nucleotide Archive

FDR false discovery rate

FEG field emission gun

FIB focused ion beam

GFF general feature format

HMM hidden Markov model

KG knowledge graph

LLM large language model

MAG metagenome-assembled genome

MDP major dome protein

MD molecular dynamics

MSA multiple sequence alignment

M. pneumoniae *Mycoplasma pneumoniae*

NMR nuclear magnetic resonance

OBO Open Biological and Biomedical

ODE ordinary differential equation

PAE predicted aligned error

PCA principal component analysis

PDB protein data bank

pLDDT predicted local distance difference test

RNAseq RNA shotgun sequencing

SBML Systems Biology Markup Language

SVA surrogate variable analysis

TAP tandem affinity purification

TED The Encyclopedia of Domains

TEM transmission electron microscopy

TPM transcripts per million

1 Introduction

1.1 *Mycoplasma pneumoniae* as a minimal cell model for systems biology

Mycoplasma pneumoniae is a pathogenic bacterium characterized by a small cell size and the lack of a cell wall. The first report concerting this bacterium came in 1941 from Eaton, Beck, and Pearson [1], who recovered it from the sputum of patients with a form of atypical pneumonia. Initially, the so-called Eaton agent was thought to be a virus due to the challenges associated with its cultivation [2]. Moreover, the lack of a cell wall also makes several classes of antibiotics ineffective. Eventually, in the 1960s, Chanock, Hayflick, and Barile [3] managed to grow it in cell-free medium. Its inoculation to volunteers confirmed it as the etiological agent for atypical pneumonia in humans. The first genome sequencing of *M. pneumoniae* was performed in 1996 in Heidelberg University [4]. It revealed a size of 816,394 base pairs with an average G+C content of 40.0%. The first modern annotation of the gene locations and functions came four years later from the Bork group at the European Molecular Biology Laboratory (EMBL) [5]. The reduced size of *M. pneumoniae*'s genome, which encodes for only about 700 proteins, together with its pathogenic lifestyle, make it relatively difficult to grow the bacterium in the lab. Indeed, it lacks many biosynthetic pathways, divides relatively slowly (every 6-20 hours), and depends on a rich culture medium for survival [6]. Another peculiarity is that it uses a non-standard genetic code, where UGA is not a stop codon, but is translated to tryptophan. Imaging studies revealed that the cells have a slightly elongated and asymmetrical shape, with a length of around 1 μm–2 μm and a thickness of about 0.1 μm–0.2 μm [7].

For the past almost 30 years, *M. pneumoniae* has been a constant presence at the EMBL. Its appeal stems mainly from the exceptionally small size of its genome and cell volume. Indeed, although its sister species, *Mycoplasma genitalium*, is the smallest known autonomously replicating organism in Nature [9], *pneumoniae* is already close to being a minimal cell, *i.e.* a cell that contains only the bare minimum to sustain autonomous life. As such, Mycoplasmas are ideal model organisms to answer fundamental biological questions. Their role in this niche is strengthened by studies that tried to further reduce their genome artificially and culminated with the top-down creation of a synthetic cell inspired by *Mycoplasma mycoides capri* [10]. JCVI-syn3, the name given to the artificial cell created at the J. Craig Venter Institute, contains only 473 genes (149 of which, almost one third of the total, have unknown function). Moreover, the relative simplicity of these organisms has stimulated the creation

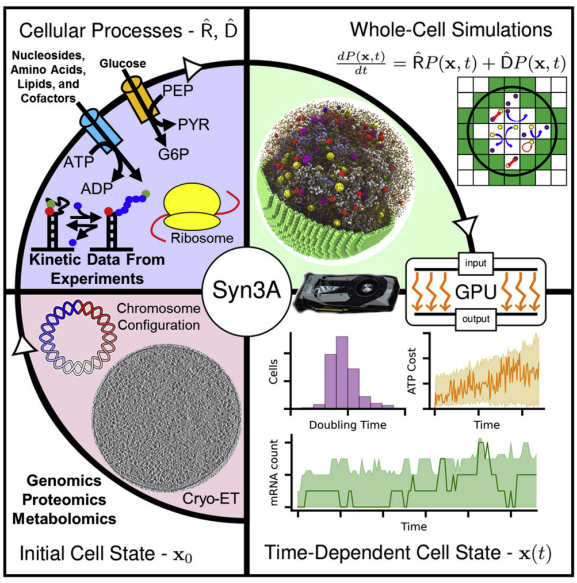


Figure 1.1: Figure from Thornburg et al. [8]. The authors developed a catalogue of all cellular process at the level of individual biochemical reactions. Using genomic, proteomic, and metabolomic data to determine an initial cell state, they used a reaction-diffusion equation system to simulate the evolution of the cell state over time.

of computational whole-cell models. Some of the issues and potential rewards of this approach were already discussed in 2007 by Betts and Russell [11]. Karr et al. [12] built a heterogeneous model of all biological processes in *Mycoplasma genitalium*, and Maritan et al. [13] complemented this model by adding structural information for the full proteome. Thornburg et al. [8] finally constructed a 3D dynamic spatial and temporal kinetic model of JCVI-syn3A which revealed connections between metabolism, genetic information, and cell growth (fig. 1.1).

At EMBL, *M. pneumoniae* has been chosen to answer important questions in various fields of biology. In 2000, its genome was reannotated with more modern approaches, adding valuable information [5]. In 2006, Seybert, Herrmann, and Frangakis [14] visualized *M. pneumoniae* cells for the first time at the high magnification achievable with cryo-ET. In 2009, a series of three landmark studies analyzed its metabolism, gene expression, and protein-protein interactions [15, 16, 17]. This was a collaboration among the groups of Peer Bork, Anne-Claude Gavin, Rob Russell, and Luis Serrano. The researchers involved in this collaborations went on to dissect key biological processes including post-translational modifications [18], regulation by transcription factors and environmental stimuli [19], and translation by ribosome profiling [20], to name a few examples. Recently, *M. pneumoniae* became an important model organism for *in cell* structural biology, thanks to work in the Mahamid group at EMBL, when the structure of the expressosome (a transient complex between the ribosome and RNA polymerase) and the structure of ten ribosome intermediates of the translation-elongation cycle were reconstructed to near-atomic resolution by cryo-ET [21, 22].

Traditionally, *M. pneumoniae* has been classified in the Mollicutes class [23], a name that captures the "soft skin" of this group of bacteria. Most bacteria in this group are pathogens of animals or plants, living either on or in the host's cells, and they have special nutritional

needs. Due to their marked small size and behavioural differences from other known bacteria, they have been classified in their own phylum, Mycoplasmatota (formerly known as Tenericutes) [24]. However, some researchers would place this clade under the Bacillati phylum [25, 26], and some argue that it is not even a monophyletic group [27, 28]. In 2018, Gupta et al. [28] proposed sweeping changes to the nomenclature of Mollicutes, including renaming *Mycoplasma pneumoniae* to *Mycoplasmoides pneumoniae*. These changes were initially rejected by the International Committee on Systematics of Prokaryotes [29], but after the original authors appealed again in 2020 [30], the new names are starting to gain traction and have been adopted in the NCBI and GTDB databases [24, 26]. Although their taxonomic classification is controversial, most studies point to the Bacillota phylum as the closest relative of Mollicutes. Thus, interestingly, although the lack of a cell wall makes them Gram-negative, their close relatives are Gram-positive. The current hypothesis is that mycoplasmas originated from bacilli, rapidly adapting to the pathogenic lifestyle and reducing their genome [31]. These evolutionary changes, including the adoption of the new genetic code, are quite dramatic and must have led to profound rearrangements and adaptations throughout the whole genome.

1.2 Cryogenic electron tomography (cryo-ET) as entry point for *in situ* biology

cryo-ET is an imaging method based on transmission electron microscopy (TEM). By acquiring multiple images of the same sample tilted at various angles, typically ranging from $-70°$ to $70°$, it is possible to reconstruct a 3D image, the tomogram, from the series of 2D projections. This method is particularly valuable for studying cells frozen in a near-native state[32, 33]. Indeed, the sample often consists of whole cells or even multicellular organisms such as *C. elegans*, as opposed to purified molecules [34]. This avoids the limitation of traditional structural methods like X-ray crystallography, which requires orthologous expression, purification, and crystallization, and can be challenging for large macromolecular complexes and membrane proteins. In recent years the field of cryo-EM has experienced a surge of interest, mainly due to technical innovations in the machines that generate the electrons (field emission gun (FEG)) and in the instruments that detect the electrons after they have crossed the sample, as well as software innovation. These evolutions have led to the so-called "resolution revolution" [35].

A key guiding principle of methods such as cryo-ET is the hope that understanding the structure can improve the understanding of the function of a biomolecular entity. For example, macromolecular X-ray crystallography and cryo-EM were instrumental in revealing the mechanism by which the ATP synthase works, since the structure of the complex clearly suggests how rotational movement can be coupled with chemical synthesis [37]. Moreover, capturing a dynamic process at different time points in different cells can give a

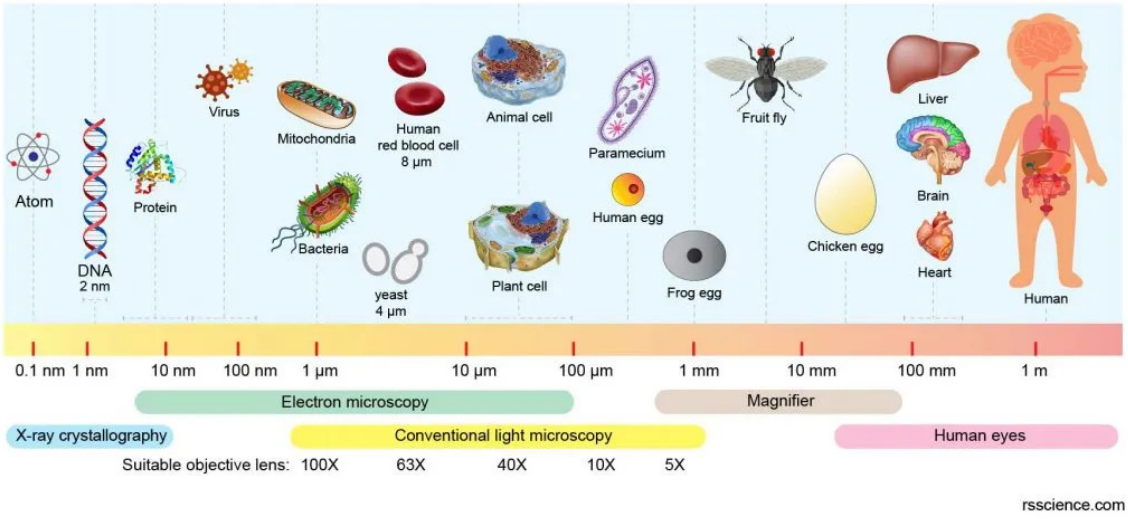


Figure 1.2: Size scales of biology (image from `rsscience.com`.) Room-temperature electron microscopy is limited because the samples have to be fixed and stained, but modern cryo-EM on purified and frozen samples can already achieve real atomic resolutions, with hallmark papers showing 1 Å [36].

glimpse into the dynamics of the biological process. What makes cryo-ET especially useful is that it allows biologists to probe life at different scales (fig. 1.2). For example, whole-cell tomograms let us look at big structures like organelles inside cells, giving a bird's eye view of how cells are spatially organized. But techniques such as subtomogram averaging can enhance the resolution of specific macromolecules up to near-atomic detail. Since the cells are kept intact, it is possible to not only examine the individual macromolecular complexes of interest, but also the broader context in which they are embedded, including other macromolecules or organelles with which they interact. Although I have never practiced cryo-ET myself, my work largely builds upon results obtained with this technique. This section is meant to provide an overview of the main principles in cryo-ET (see *Cryo-Electron Tomography: Structural Biology in situ* by Förster and Briegel [38] for a reference).

As TEM requires placing the sample in a vacuum, sample preparation is a critical step. Freezing is an obvious way of dealing with hydrated samples in a vacuum, but it comes with a cost: the ice crystals could damage the biological structures and/or hinder their visualization. Thus, rapid freezing is essential to preserve biological structures in a glass-like state, avoiding crystal growth. Improving the vitreous quality of ice was a breakthrough for which Jaques Dubochet, among others, was awarded the Nobel Prize in Chemistry in 2017. Currently, samples are typically plunge-frozen in liquid ethane, although high-pressure freezing is required for some of the thicker samples. Importantly, plunge-freezing and high-pressure freezing leave the cells almost in their native conditions, making sure that the molecules move as little as possible.

Once the samples are frozen, the tilt series can be collected and the tomograms reconstructed. Modern electron microscopes operate at energies of 200 keV to 300 keV. Higher energy electrons can penetrate thicker samples, but also lead to aberrations due to beam-induced motion of the atoms in the sample. Thus, sample thickness is one of the most important variables that influence the quality of the reconstruction. For optimal results, the sample shouldn't be thicker than 100 nm to 200 nm, especially because the thickness will increase dramatically at higher tilt angles. For some samples, it is therefore necessary to use a precisely targeted beam of gallium ions to make the sample thinner, a technique called focused ion beam (FIB) milling. The ion beam removes layers of biological material from the specimen, exposing the interesting structures underneath the surface. In the case of our model organism, *Mycoplasma pneumoniae* (*M. pneumoniae*), however, FIB milling is not necessary, since the thickness of the cells does not exceed 200 nm (see also section 1.1). This makes it possible to capture the whole cell in the microscope's field of view, which is not possible for larger cells.

After reconstructing the tomograms, they can be analyzed using a variety of methods. Here, we will focus on methods to extract and enhance the resolution of single macromolecules, which in this context are also referred to as "particles". Often, one of the first steps in a cryo-ET pipeline is particle picking, which consists in identifying the coordinates of the macromolecules of interest within the tomograms. This can be performed manually, which ensures high accuracy but is time-consuming, or automatically using template matching algorithms that detect particles based on their similarity with a user-provided known structure. These algorithms compute the cross-correlation at each voxel in the tomogram with a known 3D reference structure rotated in all possible orientation. Modern algorithms can also rely on deep learning to increase the accuracy of partickle picking [39]. In practice, all of these modalities are used iteratively in order to extract as many particles as possible from the tomograms. The particle picking process produces the bounding boxes for the particles of interest, which are then fed into the next step: subtomogram averaging.

Sutomogram averaging is designed to enhance the resolution of the structure reconstructed from the picked particles. By aligning and averaging multiple copies of the same macromolecule, the noise from the individual images tends to cancel out, and we are left with a clear structure. If the particle is known to have some symmetry, this fact can be exploited to achieve an even higher resolution. Often, the structure reconstruction achieves near-atomic resolution, *i.e.* below 4 Å. Such resolution allows structural biologists to build atomic models into the density.

But most macromolecules are not rigid and motionless, and cryo-ET can be used to capture conformational dynamics through a process called multibody-refinement [40]. In multibody refinement, the macromolecular structure is divided into two or more regions of interest, often separated by a relatively flexible interface. Then, the particles are aligned only on one region, leaving the others free. This reveals the multiple conformations and shapes taken by the particle, highlighting some of the possible movements of the structure. Multibody

refinement is particularly useful for studying flexible molecules such as molecular machines and ion channels.

One additional analysis, which is key to resolving multiple states of the same complex, is particle classification. It consists of clustering similar structures into groups which then are averaged and refined separately, under the supervision of the researcher. This method enables the identification of different conformations of a protein (*e.g.* the open and closed states of an ion channel), as well as the differentiation of distinct molecules bound to the same macromolecular assembly. For example, Xue et al. [22] classified ribosomes into 10 intermediate states of the translation elongation cycle, whereby in each state the ribosome interacts with a unique set of molecules or exhibits a different relative rotation of its subunits. Such classification will also form the foundation of the project described in chapter 4.

As any experimental method, cryo-ET comes with some limitations. One of the main issues in cryo-ET is the anisotropy in the reconstruction due to the missing wedge. Due to the thickness of the sample, the tilt series needs to stop at 60°–70°. Therefore, the imaging process leaves out a wedge-shaped volume, introducing artifacts and leading to a decreased resolution along the z axis (parallel to the beam path) compared to the x-y plane.

Moreover, although time-resolved approaches are starting to emerge [41], cryo-ET still mostly provides a static snapshot of the sample. Any dynamical processes are therefore challenging to investigate. chapter 4 addresses this limitation directly, making use of biology's other microscope: mathematics [42].

cryo-ET's ability to span multiple scales makes it complementary to other structural biology techniques. While single-particle methods such as cryo-EM and X-ray crystallography provide high-resolution reconstructions of purified molecules, cryo-ET excels at studying cellular structures within their native environment, allowing us to probe the interactions and broader context of a biological process. Any limitations in the resolution are more than compensated by the ability to investigate cells in their native state, without the need to purify the components or crystallize the molecules. This is particularly relevant for studying large complexes, integral membrane proteins, and macromolecular assemblies, all of which will be relevant in the next chapters of this thesis. In summary, cryo-ET offers a powerful multiscale framework for structural biology, bridging cellular and atomic levels of organization.

1.3 The AlphaFold2 revolution

Predicting the structure of a protein from its sequence alone has been a longstanding dream of structural and computational biologists. One of the initial assumptions was that proteins would acquire the conformation with the least free energy. Cyrus Levinthal famously estimated that the degrees of freedom in the bond angles and lengths between the atoms in a single protein easily lead to over 1×10^{300} possible conformations [43]. Although it is in

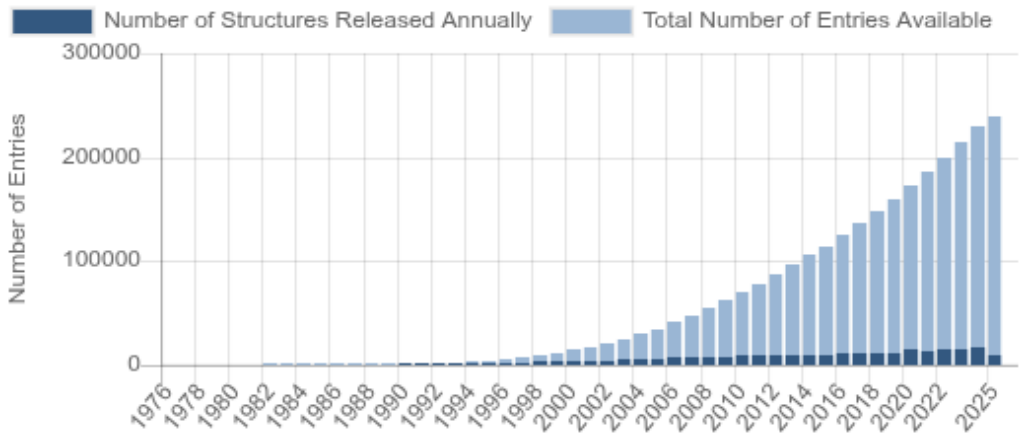


Figure 1.3: Number of structures deposited in the protein data bank (PDB) over the years [44].

principle possible to calculate the conformational energy of a structure from the position of its atoms, sequentially probing each conformation is clearly unfeasible for computational approaches as it would take an enormous amount of time. On the other hand, proteins in Nature fold into their native conformation in about one millisecond, a mismatch known as Levinthal's paradox. For decades, the field made incremental progress by recognizing the hierarchical and modular nature of protein folding. First, it is easier to predict the secondary structure than the tertiary structure directly. Second, proteins tend to organize in modular units called *domains*, and it is possible to recognize the same domain in different proteins, even if the proteins themselves are not directly related.

The computational methods for protein structure prediction lagged behind experimental techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and, more recently, cryo-EM, even though the experiments are labor intensive and time consuming. Although the number of protein structures deposited yearly in the protein data bank (PDB) has grown steadily (section 1.3), the number of experimentally resolved structures (250 thousand) is still tiny compared to the number of available protein sequences (250 million in UniProt [45]). Indeed, compared to resolving the structure of a protein, obtaining its sequence is much cheaper and easier. Methods like metagenomics have increased even more the boundaries of what is possible to obtain by genome sequencing, uncovering billion of distinct sequences [46, 47], whereas the number of resolved protein structures has been stuck orders of magnitudes below.

Most of the approaches to protein structure prediction are based on one of the unifying principles of biology: evolution. The key idea is that proteins which are neighbors in sequence space should also have a similar structure. Indeed, structure has proven to be even more evolutionary conserved than sequence [48, 49]. This means that once the structure of a protein is known, it can be propagated to proteins with a similar sequence [50]. The same

ideas inspired the classification of proteins into families and superfamilies, either based on their structure [51, 52] or their domains [53]. For years, the state-of-the-art method for predicting protein structure has been homology modelling [54]. Briefly, the process involves selecting a suitable template (a homologous protein of known structure), aligning sequences, and building a 3D model by copying structural elements from the template, using the sequence alignment as the reference. Homology modeling is most reliable when sequence identity exceeds 30%.

Everything changed in 2020, when the second version of AlphaFold was released. The Critical Assessment of Structure Prediction (CASP) programme was initiated by John Moult in 1994 as a biyearly challenge to track the progress of structure prediction methods [55]. For 13 editions, progress was incremental. In 2018 a neural-network based model, AlphaFold, outperformed homology-modelling approaches, but only by a small margin. In 2020, at CASP 14, AlphaFold2 achieved a ground-breaking median backbone accuracy of 0.96 Å rmsd_{95} (C_α root-mean-square deviation at 95% residue coverage), which is comparable to experimental methods [56].

Interestingly, this breakthrough came from DeepMind, a company that in 2014 was acquired by Google. DeepMind had already gained recognition for developing artificial intelligence (AI) models capable of mastering complex games such as Go and Starcraft, demonstrating the potential of artificial intelligence in solving complex problems [57, 58]. With AlphaFold, DeepMind ventured into scientific territory, revolutionizing the field of protein structure prediction. In 2024, Demis Hassabis (co-founder and CEO of DeepMind), John Jumper (scientist and lead author of AlphaFold2), and David Baker (biochemist and computational biologist) were awarded the Nobel Prize in Chemistry.

The deep-learning architecture employed by AlphaFold is completely novel in the field of structure prediction. A key innovation is the use of attention mechanisms, which allow the model to effectively capture relationships between distant residues in the sequence, leading to a better understanding of the protein folding constraints. Additionally, the model integrates evolutionary information by computing MSAs, which allow it to recognize conserved residues and structural motifs. Unlike previous approaches that rely on multiple intermediate steps, sometimes requiring manual intervention, AlphaFold is an end-to-end framework, where the network is trained to predict the final protein structure rather than an intermediate result that should then be refined. Moreover, the software was released in 2021 with an open source license for academic use, meaning that everyone in the scientific community could use the trained model to predict their protein structures of interest, as well as re-train or fine-tune the model. This has enabled the creation of a database of computationally predicted protein structures that, at last, rivals the size of sequence databases: the AlphaFold protein structure database, created in 2022 by a partnership between EMBL-EBI and DeepMind, now hosts predicted structures for 200 million proteins, covering most of UniProt [59, 60]. AlphaFold represented a paradigm-shift in fields such as drug discovery, protein function analysis [61], and protein design [62, 63].

Some of the limitations of this technology are as follows. First, AlphaFold2 was not trained to predict interactions with other proteins or small molecules, a drawback that was addressed with the release of AlphaFold-multimer first [64] and, more recently, AlphaFold3 [65]. Second, AlphaFold still largely relies on MSA, so, while it can predict novel folds that are not in the PDB [66], it still struggles to predict the structures of "orphan" proteins (*i.e.* sequences without many relatives).

An introduction on the field of protein structure prediction would not be complete without mentioning an alternative strategy to the AlphaFold approach. ESMFold [67] uses a language model based on the transformer architecture [68] trained on protein sequences to predict the structure without the need of a multiple sequence alignment. This greatly speeds up the inference and allowed researchers to predict the structure of more than 600 million proteins from metagenomes in MGnify [46].

Although not everything in biology is a protein, having access to high-quality protein structures in a matter of minutes rather than months or years has opened up many new research avenues. In this thesis, AlphaFold features prominently especially in chapter 2, where it was the starting point that enabled the project.

1.4 Modelling in biology

Jeremy Gunawardena defines models in biology as "accurate descriptions of our pathetic thinking" [69]. Although provocative, this definition captures the notion that models are not intended to be perfect representations of reality. Rather, they serve as a reflection of our current understanding and assumptions about a given system. Through models, we can leave out all the details that are irrelevant and focus on the minimal set of assumptions that we believe in, asking whether they are enough to describe the key features of the system. After establishing what is important, the modelling process entails following semi-automatic mathematical steps to get to the logical conclusions.

Models can also guide our intuition, often revealing surprising results. Analytic and quantitative approaches shed light on questions that would otherwise be much harder or impossible to tackle. For example, consider the problem of cell size control, discussed by Rhind [70]. There are three classes of models that can explain how cells know how big they are and maintain size homeostasis, referred to as timer, sizer, and adder. The timer model posits that cells grow for a fixed amount of time each cell cycle, then divide. In the sizer model, cells grow until they reach a certain size, which triggers the division. According to the adder model, cells accumulate a fixed amount of mass (one half of the cell's target size) before dividing. All three mechanisms could explain how cells maintain size homeostasis across many cell cycles, but only two of them are robust to noise. In the timer model, big cells grow faster and small cells grow slower, amplifying existing differences and leading

to heterogeneous populations.[1] Another example is provided by Rosenfeld and Alon [71], who show that increasing the degradation rate of a protein leads to faster response time when external stimuli occurs. This comes at the cost of increased protein production, which could be seen as a futile cycle or an evolutionary mis-adaptation. These examples show how quantitative analyses can provide compelling explanations for phenomena that might be confusing

Another benefit of adopting the mathematical language is that seemingly unrelated phenomena are actually expressed in the same way. Allostery provides a great example, as nicely written by Phillips [72],

> A wide variety of different biological phenomena are mediated by molecules that can exist in two different conformational states, one that we will dub the active state and the other the inactive state. A crucial feature of these molecules is that they can bind a ligand that has different binding affinities for the active and inactive states, thereby biasing the relative probabilities of these two states. By speaking the language of mathematics, it is possible to unite phenomena as diverse as the Bohr effect in hemoglobin, the accessibility of genomic DNA to DNA-binding proteins, the response of chemotaxis receptors to changes in chemoattractant concentration, the analysis of mutants in quorum sensing, and the induction of transcription factors. [...] all of these phenomena can be described by a single equation that parameterizes their activity as a function of ligand concentration, revealing a deep unity that is hidden from view when these problems are discussed verbally, although many theoretical challenges remain.

Models lead to simple narratives that are nonetheless accurate. In fact, simplification is a strict requirement, and its role in science has been profusely discussed, sometimes under the heading of "map-territory relation". Korzybski [73] wrote "a map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness". Another statement often heard is George Box's "all models are wrong (but some are useful)". Physics has a tradition of developing progressively more general models that describe natural phenomena in ever greater detail, but also becoming increasingly complex. Older and simpler models are not dismissed; rather, they are still taught for their pedagogical value and even used in practice whenever the simplification leads to negligible errors.

Finally, one of the most celebrated powers of models is prediction. When done correctly, modelling allows making quantitative predictions about experiments that were never made before. When these predictions are successful, it is a powerful indication of the correctness of the theory. Although less impressive, predictive power can also be achieved by fitting

[1] This assumes that cells grow exponentially, *i.e.* their growth rate is proportional to their size. The available evidence strongly leans towards exponential growth.

the parameters of a model to the observed data. This "fitting" approach is criticized by some [69], but it underlies the whole field of machine learning and has led to many success stories, including AlphaFold itself. The usefulness of such predictions is that they can spare researchers from performing complex or expensive experiments. It goes without saying that any prediction should undergo experimental validation before being trusted. The same is true for other types of analysis, including pure descriptive studies: evaluating claims from multiple perspective is always a good thing. Modelling is another point of view that sometimes can be useful. Predictive modelling serves as an additional perspective that can in many cases provide valuable insights.

In modelling, one of the primary objectives is the identification of the most important entities in a given system and the description of their interactions. In its essence, a model can be as simple as drawing a schematic diagram on a piece of paper. But the real power comes when the entities and interactions are quantified mathematically, which enables the application of quantitative reasoning to derive meaningful insights and predictions. As Phillips [72] argues, quantitative data demands quantitative reasoning, and data in biology, from high-throughput sequencing to imaging, is becoming increasingly quantitative. A quantitative mindset is just another tool in the scientist's belt, which can lead to better explainability and predictability of biological systems. In this thesis, I apply these principles to develop a model of the translation elongation cycle in chapter 4.

1.5 Aims of this thesis

In this thesis, I report the work I have done with *M. pneumoniae*, which leverages new data and computational tools that recently became available. This work is going to be multidisciplinary, reflecting my mixed background and interests. In chapter 2, in a tight collaboration with Rasmus K. Jensen (a post-doctoral researcher in the Mahamid group), I use structural bioinformatics to annotate the function of proteins forming a newly-discovered dome-shaped membrane complex that is prominently visible in the cryo-ET data. In chapter 3, drawing from lessons learned in the previous chapter for improving functional annotation, I aggregate and harmonize almost all known data about *M. pneumoniae*, design interactive visualizations to explore them, and develop a workflow based on structural homology at the domain level to greatly speed up the annotation of the still unknown part of the proteome. Finally, in chapter 4, I develop a kinetic model of the translation-elongation cycle from static snapshots obtained by cryo-ET and implement a method for the estimation of the unknown transition rates from cryo-ET steady-state distributions and biochemical data in a different organism.

2 Phylogenetic analysis of a family of hitherto uncharacterized membrane proteins assembling into a dome-shaped complex

2.1 Introduction

cryo-ET (section 1.2) has emerged as a powerful technique in structural biology, particularly for the visual exploration of life at different scales. Since this method preserves biological samples in their native state by embedding them in vitreous ice, it is particularly suitable for the visualization of large complexes and molecular interactions. In bioinformatics, some of the biggest breakthroughs were enabled by switching from targeted, hypothesis-driven methods to unbiased, shotgun approaches. For example, consider mRNA quantification. The targeted approach would be a traditional microarray, which contains a predefined number of probes, thereby limiting the analysis to only those sequences that are present in the array. On the other hand, RNA shotgun sequencing (RNAseq) is an unbiased approach, since all the RNA is sequenced. Another example is metagenomics. Here, the targeted approach would be the isolation of pure bacterial cultures consisting of one microorganism only, whereas the metagenomic approach is based on pooling all the DNA from a sample and trying to reconstruct its composition later on. cryo-ET stands in a similar position compared to traditional structural methods like X-ray crystallography: we don't just look at one molecule or complex in isolation, we look at everything that is in a cell, including therefore the native context. The challenge, then, is to infer the "molecular sociology" from what we can see in the tomograms [74].

As we have seen, this is especially true for *M. pneumoniae*, since the whole cell can in principle fit in a single 3D tomogram acquired at high resolution (original pixel size of 1 Å–2 Å) without the need for sample milling. However, one of the major difficulties is the identification of cellular components and macromolecular complexes from the images. The visual recognition of molecular complexes from the noisy tomograms is indeed a difficult skill to master. Even with the help of template matching [76] or deep-learning segmentation approaches [39], it is still impossible to identify all of the cell's molecules. First, for small or low-abundant molecules, the signal-to-noise ratio is still too low; second, even for some of the biggest complexes, there are no suitable templates. Nevertheless, the visual exploration of tomograms can still be a fruitful endeavor. The starting point for the project described

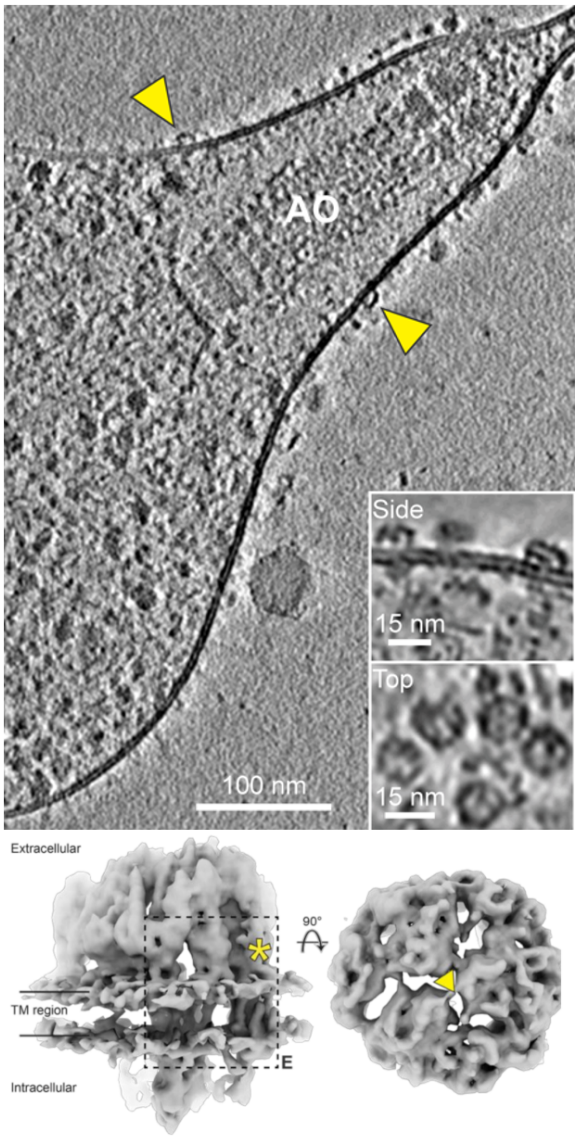


Figure 2.1: Top: slice of a tomogram of *M. pneumoniae* highlighting an interesting and uncharacterized membrane protein complex, often highly abundant near the attachment organelle (AO). This figure also illustrates that a single cell can almost fit in the field of view of the electron microscope. Moreover, it shows how difficult it is for the untrained eye to identify molecules in the tomograms. Although some of the bigger structures such as the membrane, the attachment organelle, and ribosomes are easily spotted, many macromolecular assemblies require sophisticated methods such as template matching or neural-networks. A comprehensive retrieval of the particles of interest can only be achieved by a combination of automated tools and manual curation. Bottom: density map of the dome complex reconstructed by subtomogram averaging. Figure from Jensen et al. [75] (our preprint on bioRxiv).

in this chapter was the spotting of a large membrane complex located at the outer cell surface in some of the *M. pneumoniae* tomograms (see fig. 2.1). The complex resembled a large dome or cage, and it showed pseudo 3-fold symmetry, with a symmetry-breaking transmembrane component. It was observed that this complex was relatively abundant, with a count per cell of around 40 particles. Moreover, in some instances, ribosomes were found in close proximity to these complexes, together with the Sec-translocation machinery (one of the most important extracellular transport systems in bacteria [77]). In cells treated with chloramphenicol, an antibiotic that binds to the ribosome and blocks protein synthesis, the number of ribosomes interacting with the unknown complexes was noticed to increase. Besides these facts, all gathered from the visual exploration of the tomograms, nothing was known about the nature of the complexes. The biggest question concerned the function of the extracellular dome. The rest of this chapter is dedicated to how I helped identify some of the proteins in the complexes, finding out their likely function, and analyzing their evolutionary history. This work was done in collaboration with Rasmus K. Jensen, a post-doctoral researcher in the group of Julia Mahamid, with contributions from Chistian J. Somody, PhD student in the group of Peer Bork.

2.2 The major dome proteins (MDPs): MDP436, MDP444, MDP489

As a first step to try and identify the proteins of the extracellular dome, Rasmus Jensen performed a membrane-shaving experiment. This involved using a protein cleavage agent (either trypsin or proteinase K) to break off the extracellular portion of membrane proteins, followed by peptide quantification in the supernatant by mass-spectrometry. The abundance of the peptides obtained from the treated culture were compared with those obtained from a control culture where the cleavage agent was not added. Secreted proteins shouldn't change their abundance compared to the control, but membrane proteins are expected to increase after introducing the cleavage agent. This initial experiment generated a list of candidate proteins, which were then manually curated and filtered. This left only 113 compatible proteins.

Shortly after the membrane shaving experiment, in the summer of 2021, the open-source version of AlphaFold2, which had just won the CASP challenge (see section 1.3), was released. AlphaFold often achieves an accuracy comparable to experimental methods such as X-ray crystallography, and indeed it proved an invaluable tool for this project. Using the AlphaFold2 software, it was possible to obtain structure predictions for all 113 candidate protein. Subsequently, Rasmus Jensen used PowerFit [78], a software that determines the optimal placement of a protein structure inside a cryo-ET density by performing an exhaustive cross-correlation search of the three translational and three rotational degrees of freedom of the model in the density. From this rigid-body fitting analysis, three proteins stood out: MPN436, MPN444, and MPN489. At that time, they were completely uncharacter-

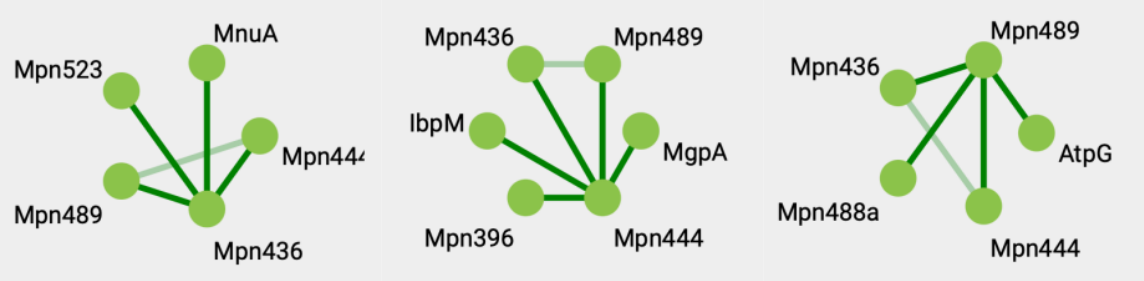


Figure 2.2: Cross-links involving MDP436, MDP444, and MDP489 (left to right). Screenshot from MycoWiki [83].

ized lipoproteins, only showing sequence similarity to other proteins of unknown function in *M. genitalium*. Based on their sequence features, they were annotated as membrane lipoproteins. According to InterPro [79, 80], they were known to harbor a PFAM domain of unknown function, DUF3713 [81] and a few disordered regions. This domain is only present in the *Mycoplasmoidaceae* family, and the proteins in this group range from 92 to 1225 amino acids long. There is one fully conserved residue, S, which could be functionally important. As *M. pneumoniae* is clinically relevant, its infections were also investigated from the medical point of view. **zhang_2016**, after analyzing several strains, found that these three lipoproteins are hypervariable, suggesting a possible role in the evasion of the host's immune system. Although these proteins are present in the STRING database [82], no physical interactions are reported for them.

Other information about these proteins can be extracted from the literature. In particular, an *in-cell* cross-linking experiment [21] found several interactions involving the major dome proteins, as reported in fig. 2.2. Proteomics studies determined the per-cell copy numbers of these proteins: 57 for MPN436, 48 for MPN444, and 48 for MPN489 [83]. A transposon insertion screening [84] identified them as essential, meaning also that knock-out experiments were out of the question as they would kill the cells. Furthermore, these three proteins are part of a paralogous group of 9 proteins in *M. pneumoniae*: MPN436, MPN442, MPN444, MPN485, MPN439, MPN438, MPN440, MPN437, MPN489. Table 2.1 summarizes other properties of these proteins that can be extracted from their sequence, and fig. 2.3 shows their sequence identities.

Given these data, the goals of the project were two: identifying the remaining members of the complex, and functionally characterizing the proteins.

2.3 Detection of remote homology with PrsA using hhblits

The first approach I tried was to use BLAST+ [85] and HMMER [86] against several reference databases to look for homologous proteins. The reference sequence databases included NCBI's non-redundant database (all non-redundant GenBank CDS translations+PDB+Swis-

Table 2.1: Main features of the nine genes in the MDP paralogous family.

Locus	Coordinates	Product	Molecular Weight	Isoelectric Point	Gene length	Protein length	Essential
MPN436	528611 → 524877 (-)	Uncharacterized lipoprotein MPN_436	136 840 Da	9.66	3735 bp	1244 aa	yes
MPN437	530638 → 528920 (-)	Uncharacterized protein MPN_437	62 920 Da	9.66	1719 bp	572 aa	no
MPN438	531893 → 530856 (-)	Uncharacterized protein MPN_438	37 950 Da	4.88	1038 bp	345 aa	no
MPN439	532662 → 531949 (-)	Uncharacterized lipoprotein MPN_439	26 070 Da	8.88	714 bp	237 aa	no
MPN440	534998 → 532818 (-)	Uncharacterized protein MPN_440	79 860 Da	10.06	2181 bp	726 aa	no
MPN442	536541 → 536089 (-)	Uncharacterized lipoprotein MPN_442	16 500 Da	10.05	453 bp	150 aa	no
MPN444	541739 → 537762 (-)	Conserved hypothetical lipoprotein MPN_444	145 750 Da	8.41	3978 bp	1325 aa	yes
MPN485	589980 → 589030 (-)	Uncharacterized protein MPN_485	34 760 Da	9.76	951 bp	316 aa	no
MPN489	596300 → 592398 (-)	Uncharacterized lipoprotein MPN_489	143 000 Da	9.68	3903 bp	1300 aa	yes

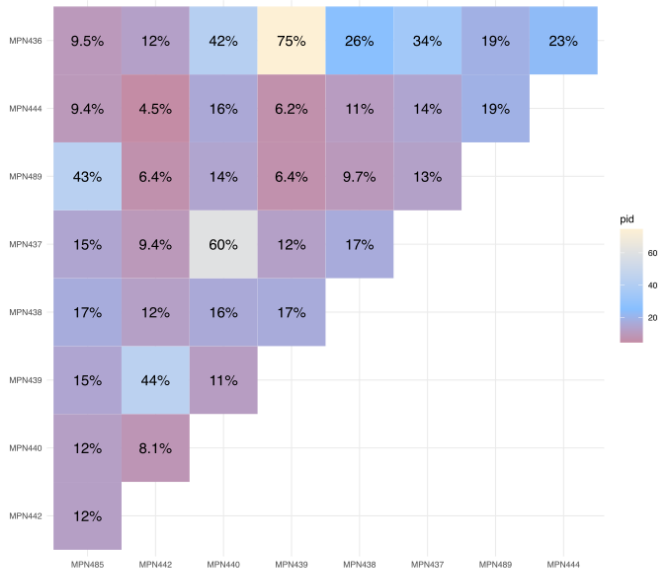


Figure 2.3: Percent identities among the sequences of the nine paralogous proteins.

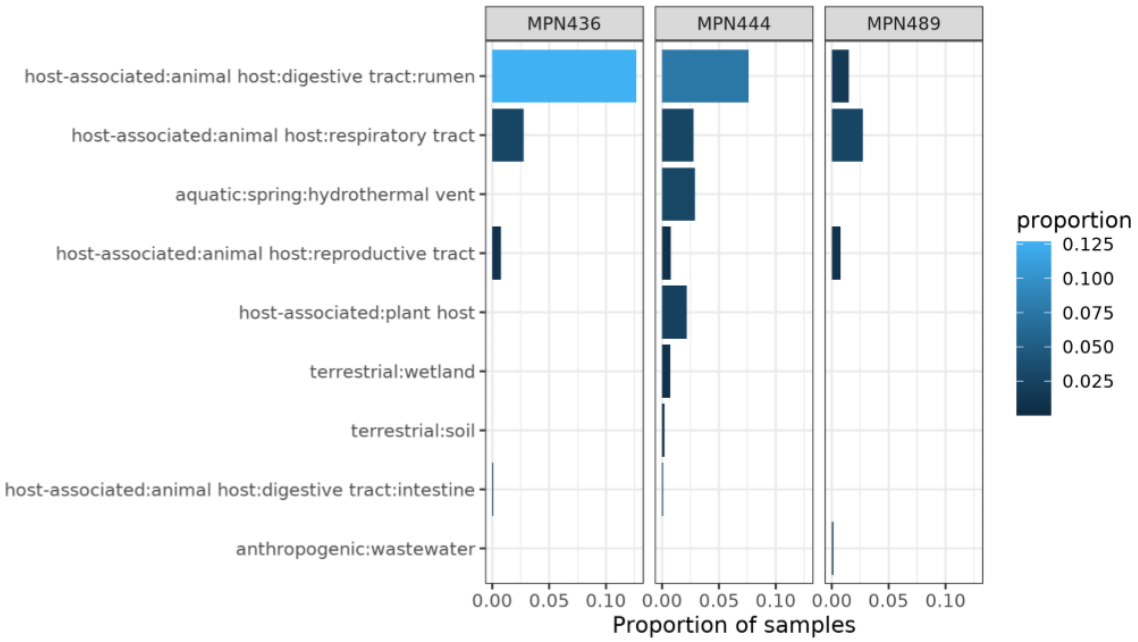


Figure 2.4: Proportion of MAGs in which the MDPs are found, relative to the total number of MAGs from each environment.

sProt+PIR+PRF, excluding environmental samples from WGS projects), the UniProtKB/Swiss-sProt database [87], and the Bork group's own non-redundant species-clustered database [88]. However, no hits besides the already known uncharacterized proteins in the *Mycoplasmoidaceae* family were reported. I also searched for these proteins in the SPIRE database [47], a large pool of annotated metagenomic sequences. Although the hits were also uncharacterized, we could perhaps find some information by looking at the environment from which the samples were collected. This analysis showed that the three proteins are mostly found in the digestive, respiratory, and reproductive tracts of animals, as well as in some hydrotermal vents (fig. 2.4). These are typical environments where Mycoplasmas are found, so, if anything, it was just a confirmation that these genes are mycoplasma-specific. Reasoning that a profile-based search tool would be better suited to look for remote sequence similarity, I turned to another sequence search tool: the HH-suite [89].

Profile-based sequence similarity tools like PSI-BLAST [90] and HMMER [86] increase the sensitivity of sequence searches by using a probabilistic representation of a group of related sequences [91]. This representation can be either a positional score matrix (for PSI-BLAST) or a hidden Markov model (for HMMER). These methods capture the conserved features of sequence families, while being relatively tolerant of insertions and deletions. While PSI-BLAST and HMMER are designed around the comparison between profiles and sequences, the newer HH-suite goes one step further by comparing profiles against profiles. Since profiles encode more information than single sequences, the HH-suite tools often

Table 2.2: Hits annotated as "Foldase protein PrsA". The results were obtained using HH-suite version 3.3.0 against the UniRef30_2021_03 [92], a database of clustered and annotated sequence profiles. Two iterations were performed, and the top 250 hits with a probability score of at least 20% were retained.

Query	Prob.	E-val	Score	A.L.[a]	Id [b]	Sim[c]	UniRef100 ID	Taxon
MPN436	86.57	1.50	59.28	103	26%	0.362	A0A022N5X4	*Enterococcus mundtii CRL35*
MPN436	82.15	3.60	55.82	71	18%	0.245	A0A061C0Y0	*Lactobacillus delbrueckii*
MPN436	64.51	20.00	50.14	74	15%	0.192	A0A023CS66	*Parageobacillus genomosp. 1*
MPN444	91.24	0.43	60.45	106	13%	0.227	A0A0D6UDC7	*Lactobacillaceae*
MPN444	90.08	0.56	62.65	113	12%	0.216	A0A061C0Y0	*Lactobacillus delbrueckii*
MPN444	86.51	1.50	59.61	106	14%	0.288	A0A022N5X4	*Enterococcus mundtii CRL35*
MPN444	79.50	4.80	55.50	119	13%	0.190	A0A011RL10	*Alkalibacterium sp. AK22*
MPN444	73.80	9.30	53.24	78	15%	0.213	A0A023CS66	*Parageobacillus genomosp. 1*
MPN489	87.78	1.00	60.39	109	14%	0.182	A0A022N5X4	*Enterococcus mundtii CRL35*
MPN489	78.93	5.50	54.38	102	17%	0.284	A0A061C0Y0	*Lactobacillus delbrueckii*
MPN489	78.45	5.40	54.75	118	18%	0.299	A0A023CS66	*Parageobacillus genomosp. 1*

[a] Alignment length
[b] Percent identity
[c] Similarity

outperform the others, especially for retrieving homologous sequences with low similarity (in the range of 20%-30% sequence identity).

Using hhblits, an HH-suite tool optimized for speed, I was indeed able to find some significant hits that had not been previously reported. Table 2.2 shows the hits annotated as "Foldase protein PrsA" for the three MDPs and some methodological details. Although the E-values are relatively high, the hhblits-computed probability of homology is close to 90% in many cases, and PrsA is the highest-scoring target excluding uncharacterized proteins. This provided a strong indication that PrsA deserved further investigation.

PrsA is a bacterial lipoprotein that plays a significant role in protein folding and secretion processes [93]. It is predominantly found in Gram-positive bacteria, where it is anchored to the inner membrane and assists in the proper folding of secreted proteins, thereby ensuring their stability and functionality. PrsA contains two distinct domains: the surA domain and the PPIC domain, each contributing to its chaperone activity. The surA domain is homologous to the SurA protein of *Escherichia coli*. This domain is responsible for recognizing and binding to the substrates of PrsA. The PPIC (peptidyl-prolyl cis-trans isomerase) domain, on the other hand, is responsible for catalyzing the isomerization of proline residues, supporting protein folding. This was highly suggestive of a putative function for the uncharacterized *M. pneumoniae* proteins, since they are also membrane lipoproteins and could reasonably be involved in secretion and/or protein folding, also in light of their interaction with ribosomes and the Sec-translocation machinery.

2.4 Confirmation of the homology of key domains by FoldSeek and DALI

In order to confirm the homology, we reasoned that we should look at the structure more in detail. While the AlphaFold-predicted structures for all *M. pneumoniae* proteins, and the database of UniProt structures had recently been released, we were lacking a tool to efficiently perform structure-based similarity searches. Although it has been possible to align two structures and calculate the TM-score for a long time, the process was too slow to be used with the AlphaFold UniProt database. Fortunately, around that time, FoldSeek became available [94]. FoldSeek is designed to compare large sets of protein structures quickly and accurately. It transforms complex 3D protein structures into simpler sequences using a special 3D interaction (3Di) alphabet, which captures how different parts of the protein interact in space. This allows FoldSeek to use fast sequence alignment methods, similar to those used for DNA, to find similarities between proteins. This approach makes it thousands of times faster than older methods like DALI [95, 96] or TM-align [97], while still being sensitive enough to catch important structural similarities. The tool can be run either in local- or global- alignment mode, and it is particularly effective with globular proteins. DALI (Distance-matrix ALIgnment) is another computational tool used for comparing protein structures [96]. It operates by constructing distance matrices that capture the pairwise distances between all residues in the protein structures being compared. By aligning these matrices, DALI can detect structurally similar regions and suggest evolutionary relationships. In the end, these two tools were instrumental in confirming the homology between the MDPs and PrsA.

I started by developing a workflow using hhblits and FoldSeek designed to increase the reach of our similarity search, hoping to find distantly related proteins that did have a functional annotation. hhblits gave us some unexpected hits for the three main proteins of interest. These hits can be seen as the neighbors of the MDPs in sequence space. With the idea of expanding the set of hits one layer further, I developed a workflow that runs FoldSeek on the hhblits hits, finding their neighbors in structure space. The final set of FoldSeek hits represents the neighbors of the neighbors of the original queries, and although it potentially contains many false positives, this approach could be useful to extend the homology search as far as possible from the original query, while still maintaining a high degree of either sequence or structure similarity. Developing this workflow presented some interesting technical challenges, so I will briefly describe it here (but see also section 3.9 in chapter 3 for a more sophisticated and successful workflow). Indeed, the "hits" from hhblits are not simple sequences, but hidden Markov model (HMM) profiles generated by a cluster of related sequences. Each profile is generated by a multiple sequence alignment (MSA) of the sequences in the cluster, and one sequence is chosen as the representative of the cluster. Thus, if the hhblits hit was, say, UniRef100_A0A023CS66, and the alignment with the query started at position 30 and ended at position 100, I would perform the following steps:

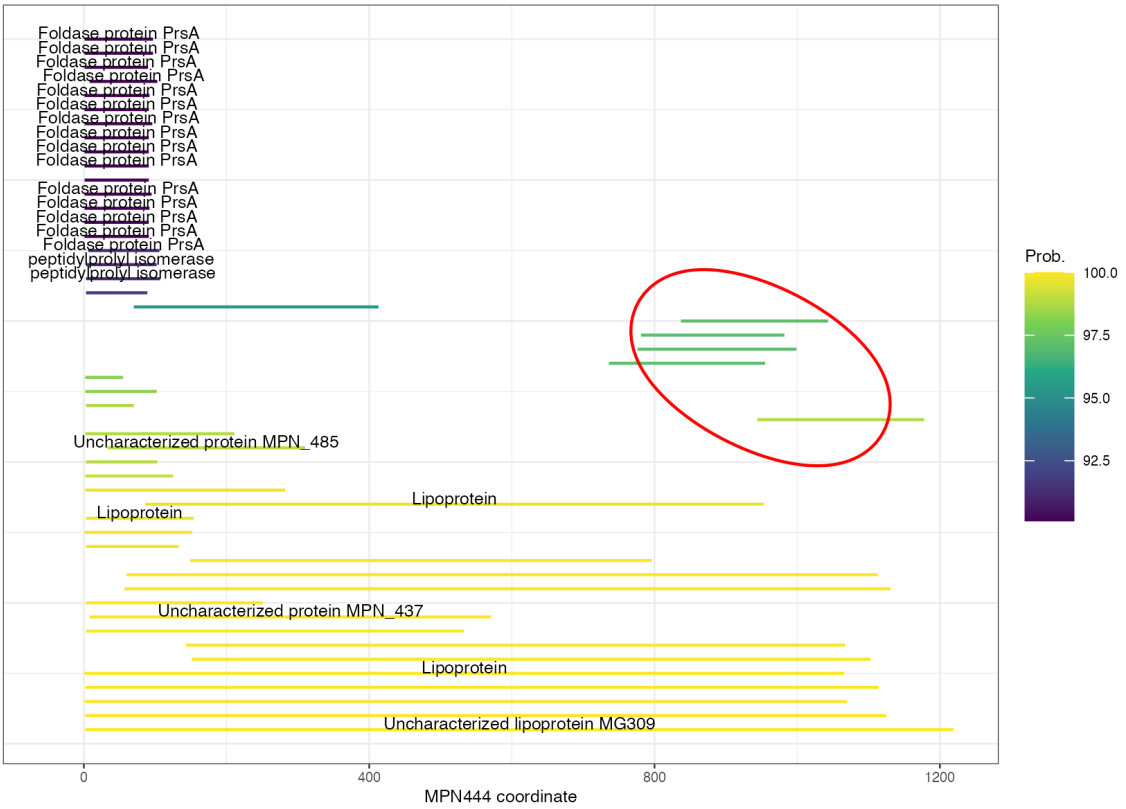


Figure 2.5: Top 50 hhblits hits for MPN444. The plot shows the local alignment between MPN444 and the target proteins: the x-axis represents the coordinates along MPN444, and each segment represents the extent of the alignment. Proteins with a given name are labelled with text. The segments are colored by homology probability, as calculated by hhblits. The encircled targets fall in the last third of the protein, denoting the potential presence of a structured domain.

- Retrieve the full MSA for cluster UniRef100_A0A023CS66;
- Map the coordinates of the alignment, 30–100, to the corresponding coordinates of the representative sequence (say, 20–70, due to gaps in the alignment);
- Fetch the PDB structure for the representative from the UniProt AlphaFold database;
- Extract the hit residues, 20–70, from the PDB;
- Submit the filtered structure to FoldSeek.

For this workflow, we used HH-suite v3.3.0 against UniRef30_2021_03 [92], and FoldSeek v3-915ef7d against the Alphafold_Uniprot50 database (UniProt predicted structures clustered at 50% identity) [56, 60], in both cases with the default significance thresholds for inclusion. The Nextflow [98] file for the workflow is available on GitHub (fmarotta/netcutter).

In the end, the workflow didn't succeed in procuring more hits than those we already knew. However, looking at where the hits for MPN444 fall (fig. 2.5) suggested that the homology with PrsA was limited to the first third of the protein. On the other hand, a few

Table 2.3: Manually defined domains in the three main MDPs. Six domains were identified in each protein. The table shows the residue ranges of the domains.

	MPN436	**MPN444**	**MPN489**
Domain 1	54–159	53–160	56–166
Domain 2	159–222, 574–627	161–228, 511–570	156–222, 543–605, 641–650
Domain 3	224–233, 273–285, 306–360, 471–478, 515–535	229–239, 277–369, 466–510	222–233, 281–293, 328–384, 455–462, 515–531
Domain 4	818–937	759–884	856–915, 957–992
Domain 5	955–962, 1037–1063, 1162–1179	903–913, 1035–1059, 1211–1245	1012–1020, 1119–1149, 1235–1245, 1272–1278
Domain 6	1065–1143	1105–1167	1163–1232

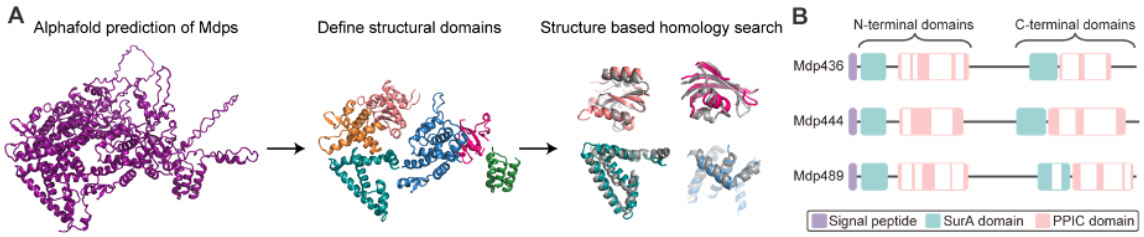
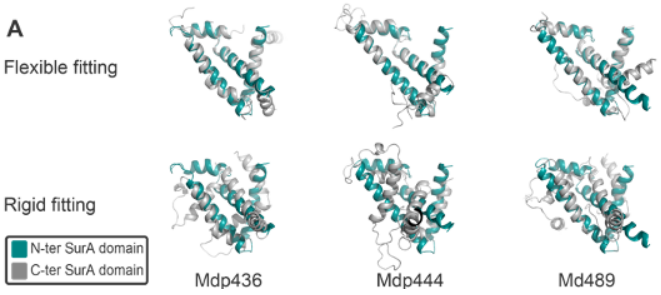


Figure 2.6: **A** The six expert-curated domains in MPN444. **B** Domain architecture of the three MDPs. Figure made in collaboration with Rasmus Jensen.

high-scoring hits fell in the last third of the protein (encircled in red in fig. 2.5). This could potentially indicate the presence of another structured domain towards the C-terminus of the protein.

Based on these results, Rasmus Jensen made a more refined analysis using FoldSeek and DALI. Leveraging his expertise in structural biology, he manually identified six domains in each of the MDPs (table 2.3 and fig. 2.6). Importantly, these domains were not necessarily contiguous in the sequence of the protein, but skipped the disordered regions. Then, each of these cut-out domains was submitted individually to FoldSeek and DALI. The results showed that domains 1 and 4 were significantly similar to a surA domain, while domains 3 and 6 were significantly similar to a PPIC domain. Both these domains are found in PrsA. Interestingly, the domain architecture (fig. 2.6 **B**) also suggested an internal duplication as a possible origin for the MDPs.



B

	p-value	RMSD / Cα	Twists	FATCAT score
Mdp436 Flexible	0.0046	2.24 Å / 91	3	139
Mdp444 Flexible	0.022	2.55 Å / 88	3	121
Mdp489 Flexible	0.042	2.81 Å / 85	3	122
Mdp436 Rigid	0.0020	5.48 Å / 82	0	133
Mdp444 Rigid	0.0078	6.14 Å / 82	0	121
Mdp489 Rigid	0.0086	5.86 Å / 91	0	121

Figure 2.7: **A** FATCAT alignments between the N- and C-terminal surA domains of each MDP. **B** Corresponding FATCAT metrics including: p-value, root mean squared deviation (RMSD), number of twists for the flexible alignment, and overall FATCAT score. Figure made in collaboration with Rasmus Jensen.

2.5 Internal duplication of the MDPs revealed by FATCAT

The MDPs are large proteins, consisting of around 1300 amino acids each. I had long hypothesized that they could have arisen from a duplication, also given that the two halves of the protein structures look similar. The domain architecture consisting of two surA-PPIC blocks seemed to go in the same direction. To perform a more quantitative analysis, I used FATCAT (Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists), an algorithm designed to align protein structures by identifying and optimizing a series of aligned fragment pairs (AFPs) between two proteins [99]. It employs a dynamic programming approach to iteratively refine the alignment, accounting for flexibility through the introduction of twists, which are small, localized conformational changes that accommodate structural variations. FATCAT enhances the detection of structural similarities by allowing slight deviations from rigid-body alignment, which is just what we needed since the MDPs in *M. pneumoniae* are interspersed with disordered fragments that do not contribute to the overall structure of the domains.

We submitted the N-terminal and C-terminal surA domains of each protein for comparison to the FATCAT web server. We tried both flexible and rigid alignment. The results, shown in fig. 2.7, showed that the N- and C-terminal regions are highly similar, lending support to the hypothesis that a duplication of PrsA led to the MDPs.

2.6 Phylogenetic tree of the MDP family of proteins and their distant homologs

Although we are now reasonably confident about the similarity between the MDPs and PrsA, we cannot be certain that the original function is conserved. Some hints towards the conservation of function emerged when we discovered an interesting crystallographic and enzymatic study of PrsA in *B. subtilis* by Jakob et al. [93]. This article investigates the

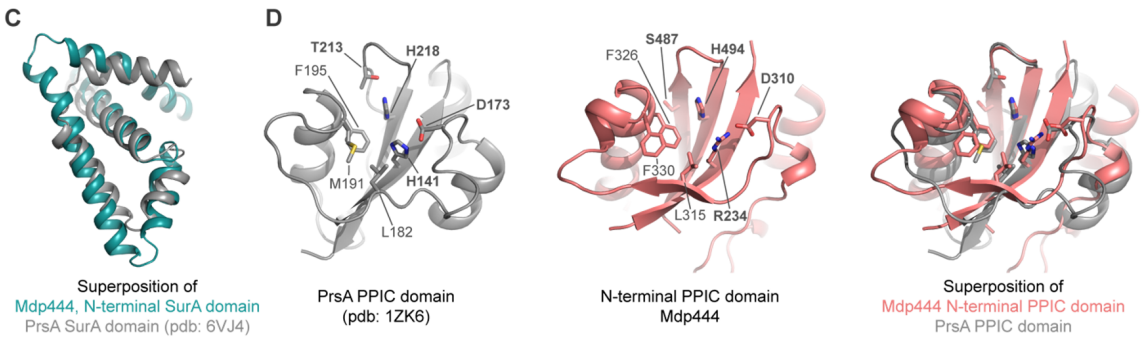


Figure 2.8: **C**: Superposition of the SurA domains from *B. subtilis* (gray) and *M. pneumoniae* (teal). **D** The PPI domains from *B. subtilis* (gray) and *M. pneumoniae* (salmon), and their superposition, highlighting the key active residues. Figure made in collaboration with Rasmus Jensen.

structure of the peptidyl-prolyl cis-trans isomerase (PPI) domain of PrsA, which belongs to the Parvulin family. The domain was shown to harbor a few active residues, including His200 and Asp155, supported by other conserved residues such as Met173, Phe177, Thr195, Tyr197, and His123. Rasmus Jensen extracted the corresponding domain from MPN444 and superimposed it to the PrsA domain from *B. subtilis*, showing that most of the key residues, including the histidine and aspartate, are present fig. 2.8. This suggested that the original isomerase function might be retained in MPN444, but the result would be even more compelling if these residues were conserved in multiple proteins. Furthermore, the presence of nine paralogs in *M. pneumoniae* signals an interesting evolutionary history for this family of proteins. Thus, as a next step, I extended the analysis from *M. pneumoniae* to the whole *Mycoplasmoidaceae* family, with the goal of investigating how widespread and conserved the catalytic residues are. I built a phylogenetic tree and performed a reconciliation analysis to infer the most likely duplication and loss events that led to the copy-number distribution of the paralogs across mycoplasmas. The multiple sequence alignment (MSA) also revealed that some key catalytic residues are conserved in at least one of the paralogs in every mycoplasmatota species except for the Ureaplasma genus.

Phylogenetic tree reconstruction consists in the inference of evolutionary relationships among organisms or genes, represented as branching diagrams known as phylogenetic trees [100]. Each branching point indicates divergence from a common ancestor, while the leaves represent extant entities. Typically, phylogenetic trees are inferred from multiple sequence alignments, using specific mathematical models of the evolution of biological sequences. The fundamental principle is that the degree of similarity between biological entities is indicative of the recency of their shared ancestry. There are many methods for inferring phylogenetic trees, including parsimony, maximum likelihood, and Bayesian methods. Here, I have only used statistical models based on maximum likelihood, *i.e.* methods that try to find the most likely evolutionary history given the current observed

data.

It is possible to build trees either at the species level or at the gene level [101]. Due to events like duplications, losses, and horizontal gene transfers, the evolutionary history of a gene does not necessarily mirror that of the species [102]. For the same reason, it is important to use single-copy marker genes when reconstructing species trees [26]. A reconciliation analysis can sort out the discrepancies and infer the evolutionary events that gave rise to the extant occurrences of genes across species. This process involves "embedding" the gene tree within the species tree, recognizing events like duplications, losses, and lateral transfers, again using probabilistic models or optimization methods [103]. Gene-species tree reconciliation is an ideal tool to investigate the evolutionary history of the MDP family.

The first step was selecting the species. I decided to restrict the analysis to the *Mycoplasmoidaceae* family, since the MDP genes are only found there. This family contains 24 genomes in NCBI RefSeq [104]. In light of the remote homology with PrsA, I also included four outgroup species which are among the closest relatives of *M. pneumoniae*, are very well studied, and contain the PrsA gene: *E. coli*, *B. subtilis*, *L. cremoris*, and *S. pneumoniae*. An outgroup is a more distantly related group of organisms that serves as a reference group when determining the evolutionary relationships of the ingroup, the set of organisms under study [105]. Outgroups are important for the rooting of phylogenetic trees, hence for assessing the direction of the flow of time in an otherwise time-reversible model [106]. The list of genome accessions used for the species tree is included in table A.1. For each of these genomes, NCBI already provides the annotation through the prokaryotic genome annotation pipeline (PGAP) [107], so I just downloaded the annotated protein sequences as well. I used the NCBI `datasets` command-line utility to download the genomes.

The next step was building the species tree. I could have used a pre-built one, such as the one generated by the Gene Taxonomy Database (GTDB) project [26], but that tree uses the FH strain of *M. pneumoniae*, whereas we work with the M129 strain. So, for consistency, I decided to build the tree from scratch, using however the same methods as the GTDB. GTDB aims to standardize bacterial taxonomy by relying on genomic sequences to classify the species. At a high level, the process used by GTDB to build the phylogenetic tree of more than 60.000 bacterial genomes is as follows. First, a set of 120 marker genes is chosen such that they are highly conserved and present in single-copy in the highest possible number of genomes. The sequences of the corresponding proteins are extracted, concatenated, and aligned using HMMER [86]. Finally, the tree is built using FastTree [108], which is based on an approximate maximum likelihood method. Conveniently, GTDB provides a software toolkit, GTDB-Tk, to reproduce the construction of their database [109]. I used the command `gtdbtk identify` to extract the reference 120 single-copy marker genes from the downloaded genomes, then `gtdbtk align` to produce a multiple sequence alignment for each gene and concatenate them into a single alignment. I then used FastTree v2.1.11 to construct the phylogenetic tree [108], with parameters *–lg –gamma*. Finally, the tree was rooted with the `ape` package in R, using *E. coli* as the outgroup, as it is known to be distantly

related to bacilli and mycoplasmas.

Subsequently, I needed to build the gene tree and perform the reconciliation. The GeneRax software can perform both steps at the same time [102], but it needs an MSA as a starting point. I used the jackhmmer program from HMMER version 3.3.2 to extract the homologous genes to MPN436, MPN444, and MPN489 in the genomes of the *Mycoplasmoidaceae* family. Jackhmmer was run separately for each protein, then the targets were combined in a single list. To these, PrsA proteins from the outgroups were added. MPN436, MPN444, and MPN489 are large proteins (for prokaryotic standards), consisting of around 1300 amino acids. On the other hand, PrsA is relatively short, around 400 amino acids. Furthermore, PrsA aligns well only to the N-terminal region of the MPN proteins. The difference in length makes it challenging for alignment tools like MAFFT [110] and MUSCLE [111]. Thus, the N-terminal and C-terminal regions were aligned separately. For the N-terminal region, I extracted residues 1-941 of MPN444, aligned all other proteins (including PrsA) to this subset of residues, and discarded the segments that aligned beyond residue 941 of MPN444. For the C-terminal region, I used the segments that were excluded from the N-terminal alignment. The two alignments were then concatenated row-wise. All alignments were made with MAFFT version 7.520 using the `linsi` option, which is supposed to give the most accurate results [110]. These manual steps were motivated by prior knowledge that the structures of the SurA and PPI domains align at specific residues (fig. 2.8).

Trimming the alignment by removing columns with too many gaps can improve the quality of the phylogenetic tree reconstruction [112]. I used trimAL version 1.4.1 with the `--automated1` option [113]. trimAL can use a variety of algorithms to decide which columns to omit. The `gappyout` algorithm uses the distribution of gaps in the alignment to choose a cutoff and remove the columns with a higher proportion of gaps. The `strict` method is the same as `gappyout`, but it also removes columns that are redundant because they are similar to one another. The `automated1` method uses a heuristic to choose between strict and gappyout, and is optimized for reconstruction of phylogenetic trees with the maximum likelihood method. The trimmed alignment was given as input to GeneRax v2.0.4, using LG+G as substitution models and parameters `--per-family-rates -r UndatedDL` [102]. After building the gene tree using a maximum likelihood method, GeneRax uses a joint model of sequence substitution and duplication/loss to evaluate the alternative reconciliation topologies, using subtree pruning and regrafting (SPR) to explore the tree space.

The final outcome is a reconciliation between the gene tree and the species tree, representing the most likely sequence of duplication, speciation, and loss events that can explain the current copy-number distribution of the genes across the species (fig. 2.9). Although the gene names are not shown in the plot, my analysis revealed, for example, that among the nine paralogs in *M. pneumoniae*, three are the "long" isoforms MPN436, MPN444, and MPN489; one, MPN485, likely arose from a duplication of MPN489, and five (MPN437, MPN438, MPN439, MPN440, and MPN442) likely arose from repeated duplications of MPN436. This would be consistent with the genome loci occupied by these proteins. More

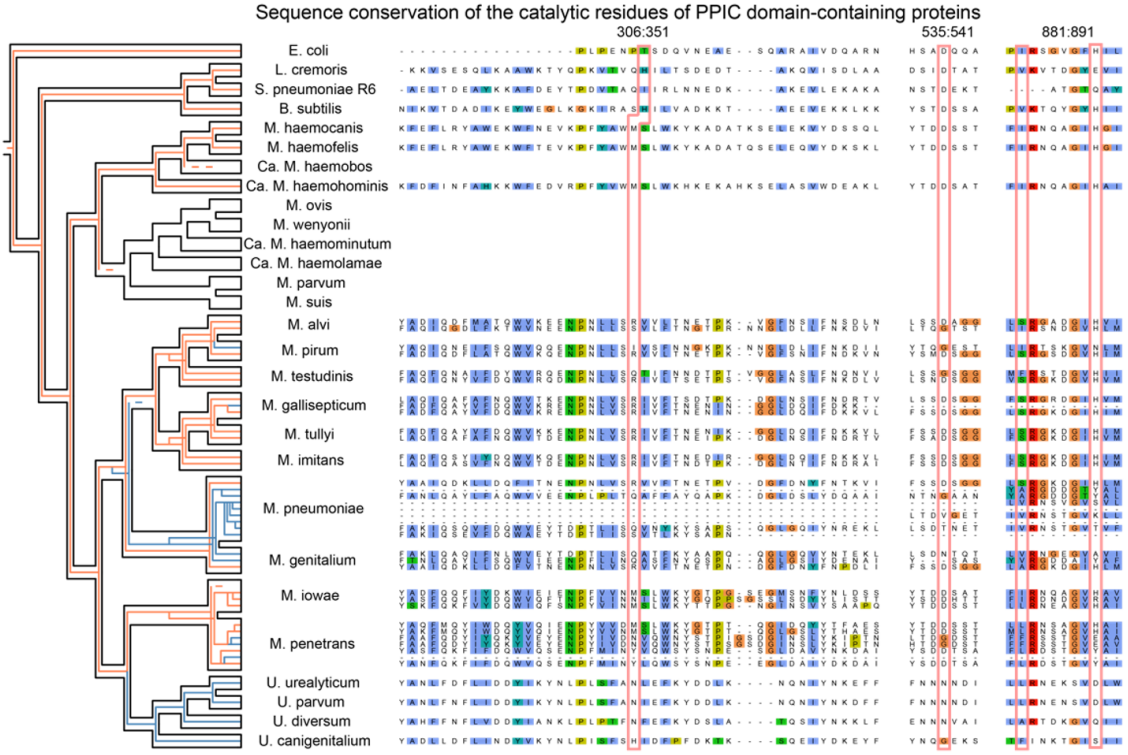


Figure 2.9: Left: the evolutionary history of PrsA and its relatives in the *Mycoplasmoidaceae* family. The black outline indicates the species tree, and the colored lines indicate the proteins. The lines are colored to indicate whether the catalytic residues are conserved (orange) or not (blue). Right: corresponding multiple sequence alignment of three selected regions: 306-351, 535-541, and 881-891, highlighting the conservation of the active residues.

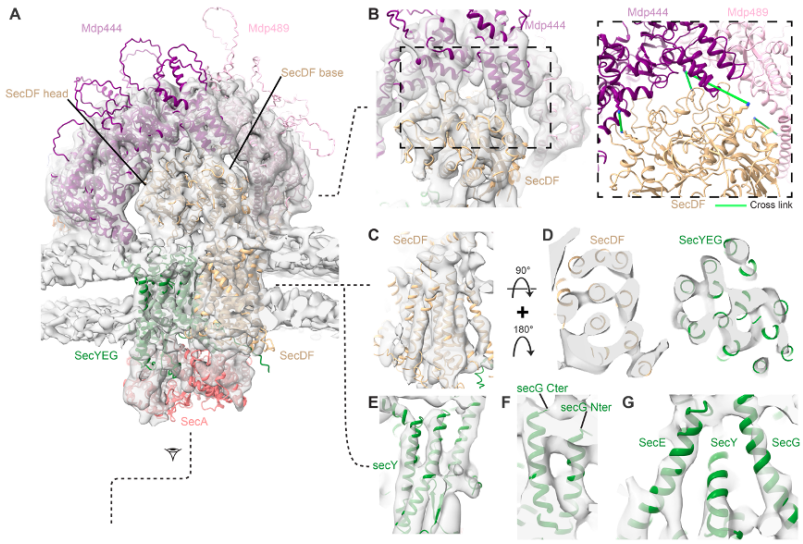


Figure 2.10: Density of the dome complex obtained by subtomogram averaging and protein structures obtained by AlphaFold2. The figure shows that the structures fit well and can explain the density. In some regions, the resolution of the density is high enough that the individual alpha-helices can be recognized.

interestingly, examining the multiple sequence alignment at the positions corresponding to the putative active residues in MPN444, we notice that these residues are conserved from PrsA in all species possessing homologs of the MDPs, with the exception of the Ureaplasmas. This provided strong evidence of selective pressure to maintain the function of the PPI domain. When a species contains multiple paralogous copies of the proteins, at least one of the copies has the active residues, raising questions about whether multiple different proteins are needed to assemble the dome complex, rather than just one as a homo trimer.

2.7 Detection of an atypical thioredoxin domain in MPN523

Another challenge in the project was the identification of the other components of the dome complex. Indeed, MPN436, MPN444, and MPN489 form the extracellular part of the dome, but the complex also has a transmembrane part (fig. 2.1). The recently acquired cross-linking dataset [21] was again instrumental for identifying the proteins that occupy the transmembrane part. MPN444 and MPN489 were both cross-linked to SecDF (MPN396), which is part of a large prokaryotic protein export complex. In most prokaryotes, the complex includes SecA, D, E, F, G, H, Y, and YajC. In *B. subtilis* and *M. pneumoniae*, however, SecD and SecF are fused in a single peptide [114]. Further analysis (from Rasmus Jensen and Liang Xue) of the cryo-ET densities of the dome complex revealed transmembrane structures that can be attributed to SecDF, SecY, SecE, and SecG (fig. 2.10). Moreover, an analysis of a cryo-ET dataset acquired after treating the cells with chloramphenicol showed

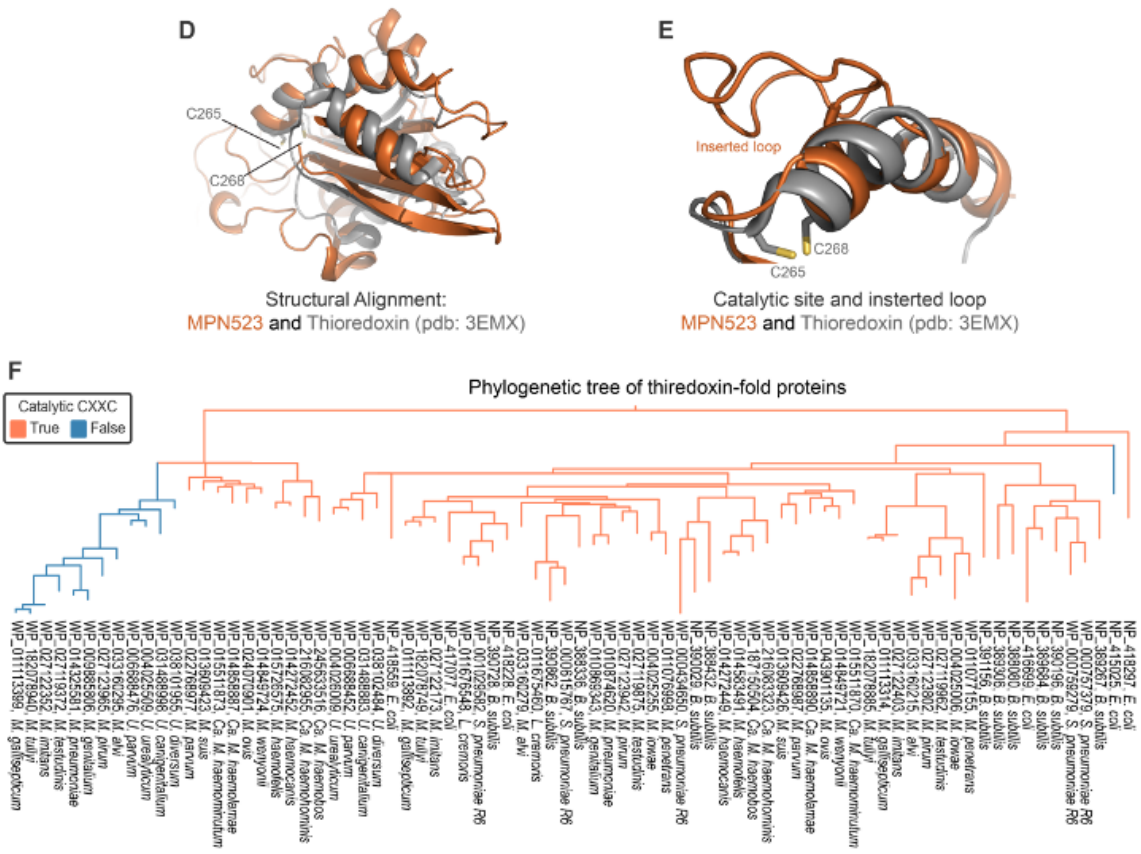


Figure 2.11: **D** Structure alignment between MPN523 (orange) and a thioredoxin from Aeropyrum pernix, the top FoldSeek hit. **E** Detail of the classic CXXC motif in the thioredoxin and the large disordered insertion in MPN523. **F** Phylogenetic tree of the MPN523-thioredoxin family, showing the presence/absence of the CXXC residues.

that SecA also interacts with the dome complex through SecDF from the intracellular side. Chloramphenicol is known to inhibit protein synthesis by blocking the ribosome, and tomograms acquired in such condition showed an enrichment in ribosomes interacting with the dome complex. This observation aligns with the known function of SecA, which binds to the nascent peptide as it exits the ribosome. SecA guides the peptide, along with the ribosome, towards the rest of the export complex, promoting its secretion. SecA is visible also in the untreated cells, but its resolution becomes higher in the chloramphenicol-treated data set. The interaction between SecA and the dome complex via SecDF suggests a direct role in the translocation process, although the dome complex is clearly different from the classic prokaryotic Sec complex.

An additional finding in the cross-linking data set was that one of the partners of MPN436 is MPN523, another uncharacterized protein. Again, a cursory HMMER search revealed no significant hits besides other uncharacterized proteins. However, after removing a large disordered insertion and submitting a clean structure to FoldSeek, we found that it bears

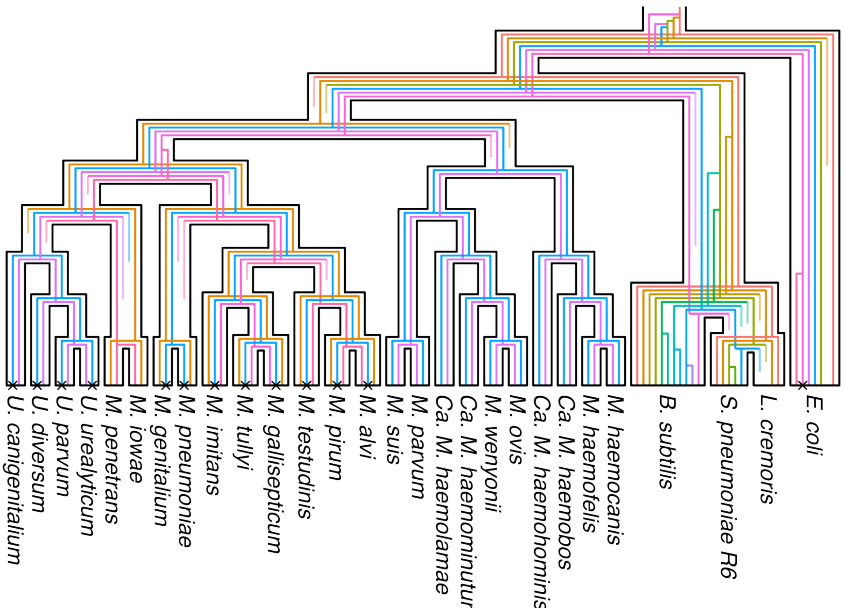


Figure 2.12: Gene-species tree reconciliation for the thioredoxin-MPN523 family.

significant similarity with a thioredoxin-like protein containing the characteristic CXXC motif typically found in Dsb chaperones (see fig. 2.11 **D** and **E**, which show the similarity between MPN523 and thioredoxin). However, MPN523 does not have the CXXC motif, suggesting that it may have lost the thioredoxin function. At first, I tried to obtain an HMM profile from the FoldSeek hits with probability score greater than 50%. However, the resulting multiple sequence alignment had too many gaps and was of poor quality. Thus, I resorted to another structure similiarity search tool: DALI [96]. I submitted the structure of MPN523 (without the disordered region) to the DALI webserver, and run a search with default parameters against the protein data bank (PDB). The result was a set of 88 pairwise alignments between MPN523 and thioredoxins from PDB. These pairwise alignments were merged into a single multiple sequence alignment using the residues of MPN523 as reference. From the alignment, I used hmmbuild and hmmsearch from HMMER v3.3.2 to build a profile and search for hits among the same genomes used in the phylogenetic analysis for MPN436, MPN444, and MPN489. This sequence database included 24 genomes from the *Mycoplasmoidaceae* family plus the four outgroups: *E. coli*, *B. subtilis*, *L. cremoris*, and *S. pneumoniae*. hmmsearch found considerably more hits, and, importantly, it captured both conventional thioredoxins (such as Dsb) and the relatives of MPN523. It should be pointed out that *M. pneumoniae* has one annotated thioredoxin, TrxA (MPN263). This new HMM profile captures both TrxA and MPN523, recognizing a similarity that was previously undetected. Figure 2.11 **F** shows that the proteins without the CXXC motif all belong to one clade, except for one protein in *E. coli*, which is unrelated.

Figure 2.12 shows the reconciliated phylogenetic tree of this thioredoxin family. One of the benefits of a reconciliated tree is that it immediately shows whether two genes are orthologs or paralogs [115]: if the node that split them most recently is a speciation event, they are orthologs, whereas if the node is a duplication event, they are paralogs. The reconciliation was performed as described in section 2.6. From my analysis, we can conclude that MPN523 is orthologous to DsbD in *E. coli*. Moreover, the split between MPN523 and TrxA happened before the divergence between *E. coli* and the bacilli/mycoplasmas clade.

2.8 Co-occurrence of the MDPs across species

One of the methods to investigate the function of unknown proteins is to study their genomic neighborhood conservation and co-occurrence [82, 116]. The biggest challenge in this case was the severe lack of significantly similar proteins in reference databases. We tried several state-of-the-art tools and services, including STRING [82], EggNOG [117], fast.genomics [118], and MGNIFY [46], but none of them was helpful as they couldn't recognize any similar proteins to the MDPs. Eventually, I came across NetCutter (Müller and Mancuso [119]) and immediately liked its approach. Although the analysis was inconclusive, like many others that I performed, I will still describe it here for posterity.

NetCutter is designed around containers and entities, aiming to find groups of entities that occur in the same container more (or less) often than expected by random chance. For example, for a gene co-occurrence analysis, the orthologous classes would be the entities, and the genomes would be the containers. In order to perform a statistical analysis of co-occurrence, we need to know the probability of occurrence of each orthologous class individually in each genome according to a null distribution. NetCutter introduced a novel approach based on edge-swapping to obtain the null distribution. The problem is modelled as a bipartite graph where an entity is connected to a container if it is present in that container. The authors investigate different methods to randomize the graph, and show that a strategy where the containers of two items are swapped, provided that the containers didn't already contain the other item, is a good approximation for the null distribution obtained by generating a complete permutation set of the bipartite graph. Thus, the process for computing the null distribution involves repeatedly swapping the edges in this way, counting how many times an entity occurs in each list, and dividing that number by the number of randomizations. This gives an estimate of the probability of occurrence of each entity in the lists. Importantly, this randomization scheme preserves the number of connections of each node, so it takes into account that some entities are potentially connected to many more containers than other entities. Once the individual occurrence probabilities are obtained, the co-occurrence probabilities under the null model are computed using the Poisson-Binomial distribution [120], a variation of the Binomial distribution which accounts for the fact that the containers have different numbers of items.

The p-value for the observed co-occurrence is calculated using such distribution. Since the reference implementation of NetCutter was only available as an old Java package and I wasn't able to make it run, I reimplemented the algorithm in an open-source R package, which is freely available on GitHub.

I then set out to apply this method to find genes that co-occur with the dome complex genes. I used the same set of genomes that I used in the phylogenetic analysis: the *Mycoplasmoidaceae* family and four "outgroups" (see section 2.6). First, I identified the orthologous classes by running an all-vs-all phmmer search (from HMMER v3.3.2). Proteins A and B were connected in a graph using the following normalized score:

$$\frac{S(A,B)S(B,A)}{S(A,A)S(B,B)} \tag{2.1}$$

where $S(X,Y)$ is the score assigned by phmmer when X is the query and Y is the target. This normalization scheme produces a number between 0 and 1. The score of an HMMER alignment depends on many factors, including the length of the proteins and the amino acid composition. Dividing by the score of the alignment of the query with itself ($S(A,A)$) provides a normalized score that measures how close the actual hit (B) is to the query. Using both the scores of A vs B and B vs A is necessary because the scores will, in general, be different, but we need a symmetric similarity measure. The edges of the graph are filtered using a threshold of 1×10^{-4} on the normalized scores. Then, the Louvain algorithm [121] as implemented in the R igraph package [122] is used to find clusters. Each cluster was identified as an orthologous class.

I built the matrix of occurrence of orthologous classes in the genomes and applied NetCutter, but, in the end, no meaningful enrichments were observed.

2.9 Discussion

Much of bioinformatics heavily relies on sequence similarity to propagate annotation, a practice rooted in the principle that orthologous sequences—those originating from a speciation event—likely share similar functions. Lack of sequence annotation and fast evolutionary rates make this process challenging. Recent advances, particularly the development of AlphaFold (Jumper et al. [56], see also section 1.3), have revolutionized bioinformatics by enabling high-accuracy protein structure prediction from the amino acid sequence alone. This means that the same principle of annotation transfer through similarity can be translated from sequence to structure. Since structure is more conserved (and more directly related to the protein's function) than sequence [49], structure-based homology methods have a great potential to expand the functional annotation of proteins.

This project started before tools like FoldSeek were available, and finding sequence similarity proved extremely challenging. Despite my efforts to decompose the proteins'

sequences into what we thought were their domains (similar to what PFAM does), the number of hits didn't increase. I attribute these challenges mainly to two factors: the rapid evolutionary changes that characterized the birth of the Mycoplasma clade, and the fact that structure is more fundamental than sequence.

Mycoplasmas are well-known for their rapid evolutionary rate [123, 31]. They are thought to have diverged around 65 million years ago from the lineage of bacilli, clostridia, streptococci, and lactobacilli [124]. This divergence was accompanied by dramatic changes including extreme genome reduction and loss of the cell wall, likely as adaptations to the parasitic, host-associated lifestyle. Moreover, Mycoplasmas have a different genetic code from that of most other bacteria: one of the stop codons, UGA, actually codes for tryptophan. The loss of a stop codon, together with the rapid genome reduction and lifestyle change, could have meant that many adjacent genes became fused into single proteins. Consequently, the genome sequence has undergone substantial changes, making sequence similarity searches more difficult.

Despite the success of sequence-based domains databases like PFAM [81], domains are really structural entities, and trying to infer them from sequence alone does not work in every case. This project provided a great example for this principle, since the structure was instrumental in confirming the homology and the internal duplication. In order to find the appropriate domains, we had to remove several "insertions"—bits and pieces that didn't alter the structure of the domain, but were interspersed within the sequence of the genes, breaking the sequence continuity of the domains.

Without the help of cryo-ET, this complex would have been much more difficult to identify and characterize structurally and functionally. Indeed, targeted approaches, while excellent for investigating deeper something that we already know, are not suited to discover something that we don't already expect. On the other hand, the possibility of looking at whole cells at such high resolution opens up the possibility of simply exploring what is out there.

Our approach integrated advanced cryo-ET imaging with computational bioinformatics, harnessing the predictive power of AlphaFold to enable structure-similarity searches across protein databases. By combining these structural insights with functional characterization and evolutionary analysis, we aimed to not only identify the proteins within the complex but also elucidate their roles and origins.

3 Aggregation, cleaning, and visualization of *M. pneumoniae* data and development of a web interface

3.1 Introduction

Due to the importance of *M. pneumoniae* as a model organism, as well as its medical relevance, a wide array of data about this organism has been collected and published. Various high-throughput "omics" experiments [125], including DNA sequencing, RNA sequencing, mass spectrometry quantification of proteins, post-translational modifications profiling, metabolic modelling, and regulatory network analysis. However, the heterogeneity and fragmentation of these data poses a significant challenge for researchers aiming to conduct integrated systems-level analyses. For someone starting a new project, even knowing what is already available is not immediate and requires deep literature reviews. This motivated me to create a unified, computationally accessible, and user-friendly framework that not only aggregates these data but also enhances their utility through interactive visualization and dynamic analysis tools. Such a framework would empower researchers to derive novel biological insights, stimulate hypothesis generation, and accelerate discoveries.

A primary obstacle in utilizing *M. pneumoniae* data is identifying and accessing relevant datasets. Data are often scattered across numerous publications, buried in supplementary files, or presented in formats that are not amenable to computational analysis. For instance, datasets embedded in PDF documents or graphical figures require labor-intensive manual extraction or specialized software for data recovery. Even when data are available in tabular form, they are typically static, lacking interactivity or searchability.

Existing databases and resources partially address this problem. Large-scale databases such as UniProt [45] and InterPro [79] provide comprehensive annotation about existing proteins and domains. BV-BRC [126] (formerly PATRIC) integrates several data modalities, but many important *M. pneumoniae* data points from the literature are missing. Other resources focused specifically on *M. pneumoniae*, also strive to aggregate different sorts of data types. The literature mentions two websites: MyMPN [127] and Mycowiki [83]. MyMPN, developed at CRG (Centre for Genomic Regulation) in Barcelona, is not accessible anymore. Mycowiki, developed at the Universty of Goẗtingen, hosts data about protein annotation, protein-protein interaction from cross-linking, gene expression across conditions, protein

abundance across conditions, metabolic reactions, and homologous genes in different organisms. However, the data is not comprehensive, and exporting the data from the website often doesn't work as expected. For instance, the metabolic network is only available as an SVG image, which is good for visual exploration, but prohibitive for someone who wants to reanalyze the network using computational tools, which expect the network as a list of edges in a CSV file, for example.

Thus, I decided to recreate a web-based resource that could address some of these limitations. In particular, my guiding principles were as follows:

FAIR principles Data should be Findable, Accessible, Interoperable, and Reusable [128].

Data aggregation Collecting data from different resources and aggregating them in one place makes exploration easier.

Interactivity Navigating across genes and across data modalities should be as easy as possible.

Scalability and extendability The resource could be extended to other model organisms for which similar data and analyses are available.

To achieve FAIRness, I strive to include data from all existing literature, allowing users to easily discover what is available without the need for extensive literature searches and reviews. Furthermore, I ensure that all data is downloadable in standard formats, such as TSV (tab-separated values) or plain text files, ensuring compatibility with all bioinformatic tools and allowing third parties to reproduce or extend the existing analyses. As regards the aggregation of data, multiple sources often provide similar types of information. For instance, protein annotation data is available from UniProt, as well as from resources like eggNOG [129] and KEGG [130]. Additionally, all these sources, along with InterPro, provide data on protein sequence domains from PFAM [81]. By aggregating data from these diverse resources into a single platform, one can glance at the annotation and compare them in a gene-centric way, without having to jump to multiple websites. Interactivity is a key feature of this resource, enabling users to seamlessly navigate across genes and data modalities. Often, the same analysis can be applied to different genes. By providing an interactive tool that allows users to select and visualize specific genes, information is conveyed much more efficiently than with a static table or figure. This interactivity allows researchers to quickly obtain insights and make informed decisions based on the data. The framework is also designed with scalability and extendability in mind. While initially focused on a specific set of data, the resource is built to accommodate expansion to other model organisms for which similar data and analyses are available. This flexibility ensures that as new data becomes available or as research interests evolve, the resource can grow and adapt to meet the needs of the scientific community.

By adhering to these guiding principles, my web-based resource aims to provide a comprehensive, user-friendly platform that facilitates data exploration and analysis of *M. pneumoniae* data. In the next sections, I will describe the different data modalities and

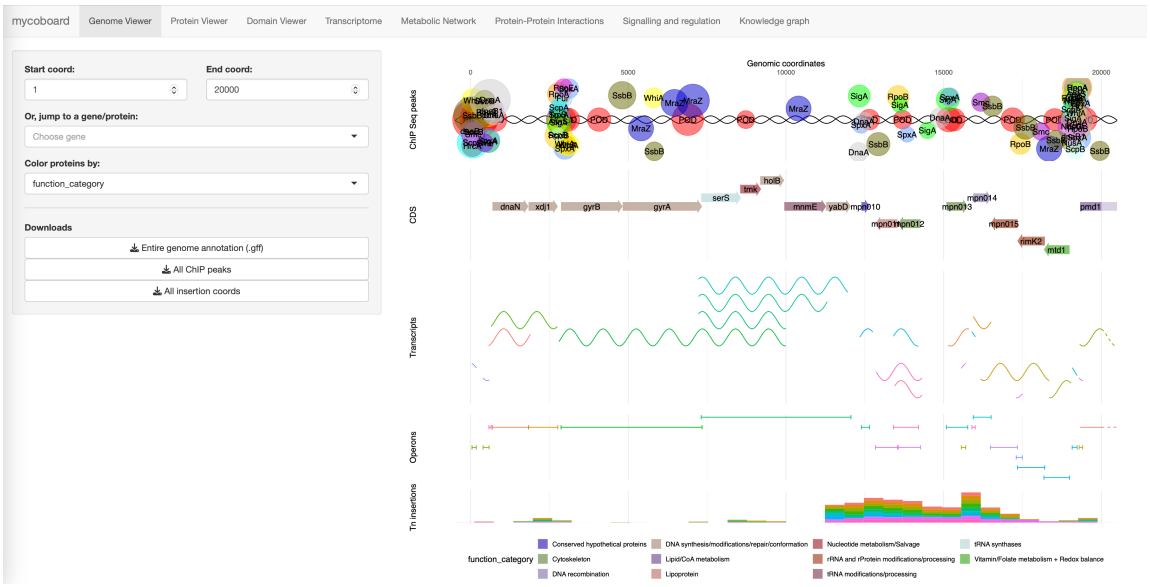


Figure 3.1: Mycoboard genome viewer.

features that the website provides. A project to annotate the uncharacterized proteins exploiting structure-based homology at the domain level will also be presented. Since the initial idea was to develop a dashboard for *Mycoplasma pneumoniae*, I called it *Mycoboard*.

3.2 Genome annotation

The landing page for Mycoboard is a genome viewer. By default, it shows five tracks: Chromatin immunoprecipitation followed by sequencing (ChIP-seq) peaks and regions of open chromatin, annotated CDS (protein-coding sequences), annotated transcripts and non-coding RNAs, annotated operons and sub-operons, and transposon insertions from different experiments. This arrangement is meant to give an overview of data that is related to the genome and can be associated with specific coordinates on the bacterium's chromosome. First and foremost, the annotation of known genes, operons, and transcripts is highlighted. Second, data from experiments that measured DNA-binding proteins, chromatin accessibility, and transposon insertion sites can be conveniently displayed. Users can navigate to specific coordinates or jump to a gene of interest through a panel on the left (fig. 3.1).

The ChIP-seq experiment was performed in the landmark work by Yus et al. [19], with the goal of reconstructing the gene regulatory network of *M. pneumoniae*. ChIP-seq uses specific antibodies to isolate DNA-associated proteins and sequence the bound DNA [131]. As such, it provides insights into the binding sites of transcription factors, DNA polymerases, RNA polymerases, and other DNA binding proteins that structurally support the chromosome. The experiment identified 23 DNA-binding proteins, their binding sites, and the sequence

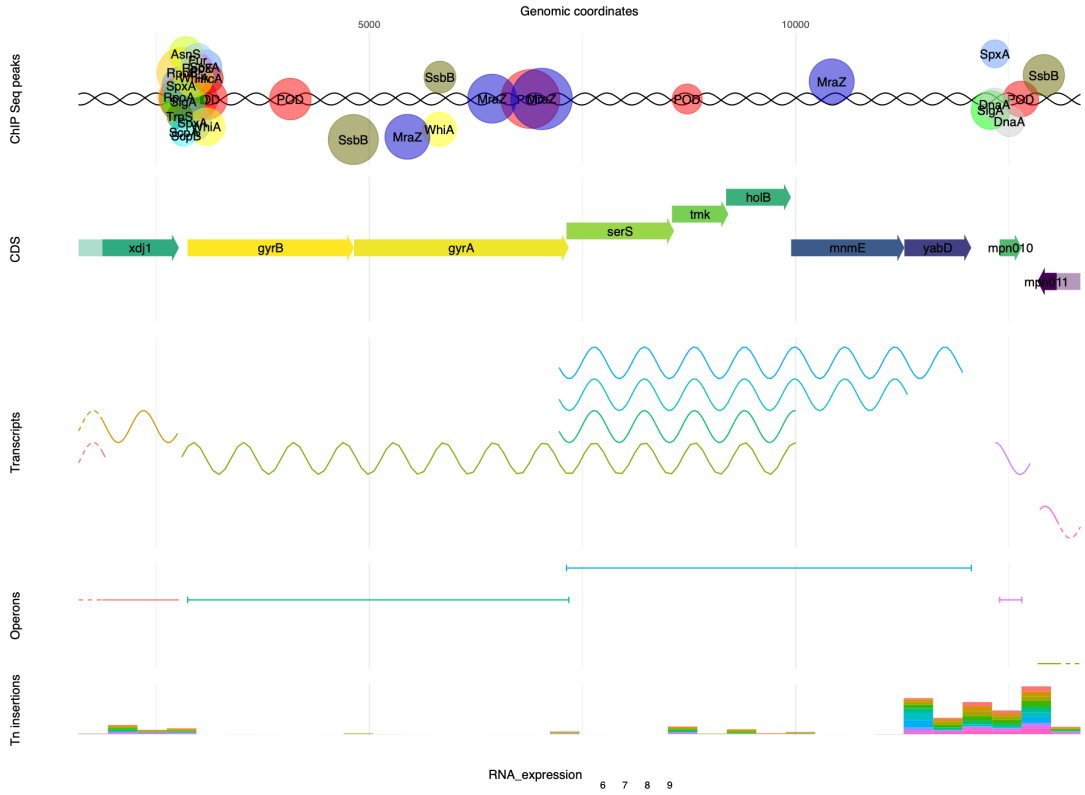


Figure 3.2: Detailed view of the staircase-like decay in gene expression for Operon 2: from gyrB to yabD, the color changes from yellow (high expression) to green (mid-low expression).

motifs that they recognize. Furthermore, Yus et al. [19] performed a DNase footprinting assay [132], which identified 428 sites as "protected regions", which usually correspond to one or more of the ChIP-seq peaks. The DNase footprinting assay consists in treating cell extracts with a DNase I enzyme, followed by sequencing of the partially digested nucleic acids; the regions that can be sequenced are those that remain intact because DNA-binding proteins protected them from the DNase. In my visualization, the DNA binding proteins and protected regions (denoted POD) are shown somewhat artistically as bubbles along the DNA double strand. The size of the bubble is proportional to the height of the peak in the ChIP-seq data: a higher peak means that more reads were mapped to that position, signifying a greater binding or presence of the protein at that locus. Often, multiple proteins are found to bind the DNA at the same position. For example, around 15 proteins are found near the replication origin, including dnaA, well-known for its role in bacterial chromosome replication. To prevent overcrowding, the y-coordinate of the bubbles is randomly jittered.

The next track shows the coding sequences, depicted as arrows pointing in the direction of the DNA strand they are in: CDS on the plus strand point to the right, while on the minus

strand they point to the left. If there is any overlap between the sequences, the downstream gene is shifted along the y-axis. By default, the CDS are coloured according to their function category, but the user can choose to change the colors according to their molecular weight, essentiality, RNA expression, protein copy number, protein half-life, or subcellular location. The RNA expression coloring is useful to visualize the well-known staircase-like decaying of expression in polycistronic operons (fig. 3.2) [16].

Next, we have two tracks showing transcripts and operons. Three main studies have analyzed gene expression in *M. pneumoniae* [16, 133, 19]. The detection of transcripts from either microarrays or RNA sequencing allowed researchers to reconstruct operon boundaries. Furthermore, Junier et al. [133] measured gene expression in several conditions, calculated the correlation matrix between pairs of genes, and applied a hierarchical clustering constrained to respect the linear organization of the genome. The resulting clusters capture multiple levels of co-transcriptional organization, from sub-operon to large-scale genomic domains. These studies also show that multiple transcript isoforms, characterized by different termination or starting sites, is relatively common, particularly due to the phenomenon of transcriptional read-through [134]. The transcripts track also shows non-coding RNAs, which contribute to the regulation of gene expression [135, 19]. It is believed that non-coding RNAs can bind to other transcripts by RNA base-pair complementarity and interfere with their translation or functioning.

Last, we have the transposon insertions track. The data come from transposon sequencing and essentiality studies [136, 137, 84]. Briefly, *M. pneumoniae* cells were transfected with a library of transposons, DNA sequences that can integrate in the bacterium's chromosome at random positions [138]. These sequences also carry some antibiotic-resistance genes so the cells that contain them can be selected. The idea of these studies is that if a transposon integrates within an essential gene, thereby disrupting its function, the cells will not survive. Thus, after several days of culture, when the DNA of the surviving cells is sequenced, the transposable elements will be identified only at the loci that are not essential for survival. The plot shows the number of inserted transposons at each genomic position. As the genomic coverage of the transfection is high (as high as 1.5bp resolution [136]), regions where no insertions are detected are likely to be essential, while regions where many insertions are found are likely to have little to no impact on the fitness. The data from different experiments was normalized for the total number of insertion sites identified, so that each experiment contributes approximately the same signal.

By having this information on the same page, one can get an initial idea of what is going on in any genomic region of interest.

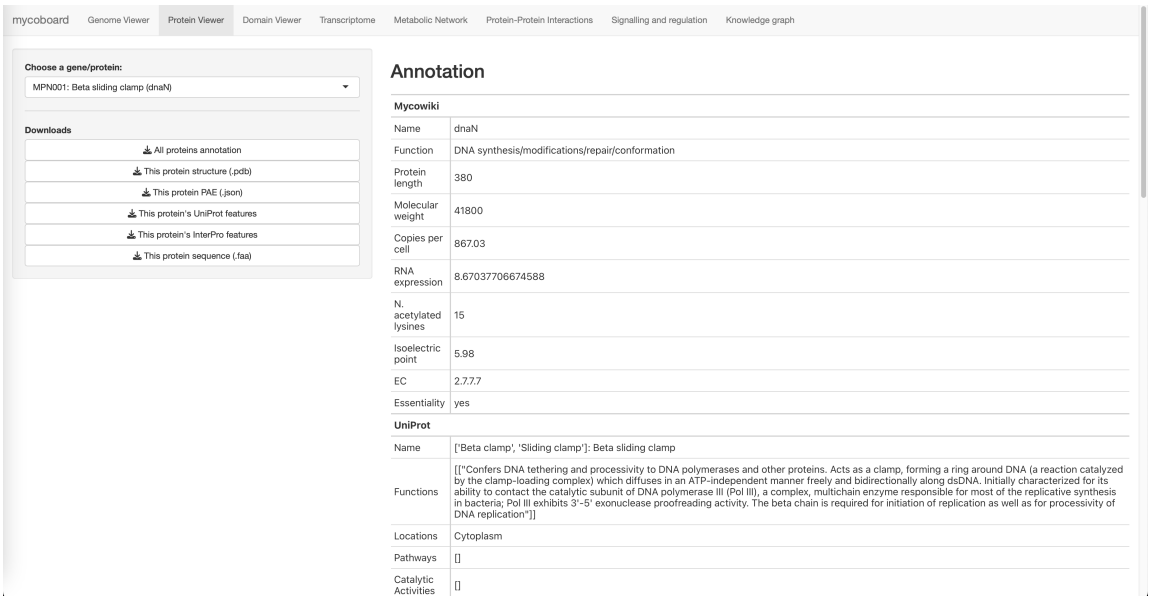


Figure 3.3: Partial view of the Annotation section for dnaN (MPN001). The table is longer than shown here, but the screenshot has been truncated for space economy.

3.3 Protein viewer

The second page of Mycoboard focuses on individual proteins, showing their physical properties and functional annotations from four different sources. This page also integrates the Mol* viewer [139] to browse the AlphaFold2-predicted structure of the protein. Furthermore, there is a visualization of the regions, domains, post-translational modifications, and other features that can be mapped to specific coordinates on the protein chain.

As regards the annotation, our sources are Mycowiki [83], UniProt [45], KEGG [130], and EggNOG [117]. The Mycowiki website provides physical properties like molecular weight and isoelectric point, curated functional annotation and enzyme class (EC number), and essentiality. To this, we add some additional information from the literature, including RNA quantification [19], protein copies per cell, and protein half life (from Burgos et al. [140]). From UniProt, we include the protein name, function description, subcellular locations, pathways, catalytic activities, and subunits. From KEGG, we have module, pathway, motifs (PFAM domains), and orthologs. EggNOG provides COG and NOG orthologous groups [141, 117], Gene ontology terms [142], enzyme class (EC number), KEGG transporter class, and CAZy category for carbohydrate metabolism [143]. The fact that some of the information is redundant, in the sense that it is provided by multiple sources, is intentional. At the end of the day, all these tools are based on the usual principle: building a database of sequences with known annotation, and using similarity search to find hits that are similar to the query protein. Nevertheless, each tool uses different methods and heuristics to obtain their annotations, hence comparing the differences can balance their

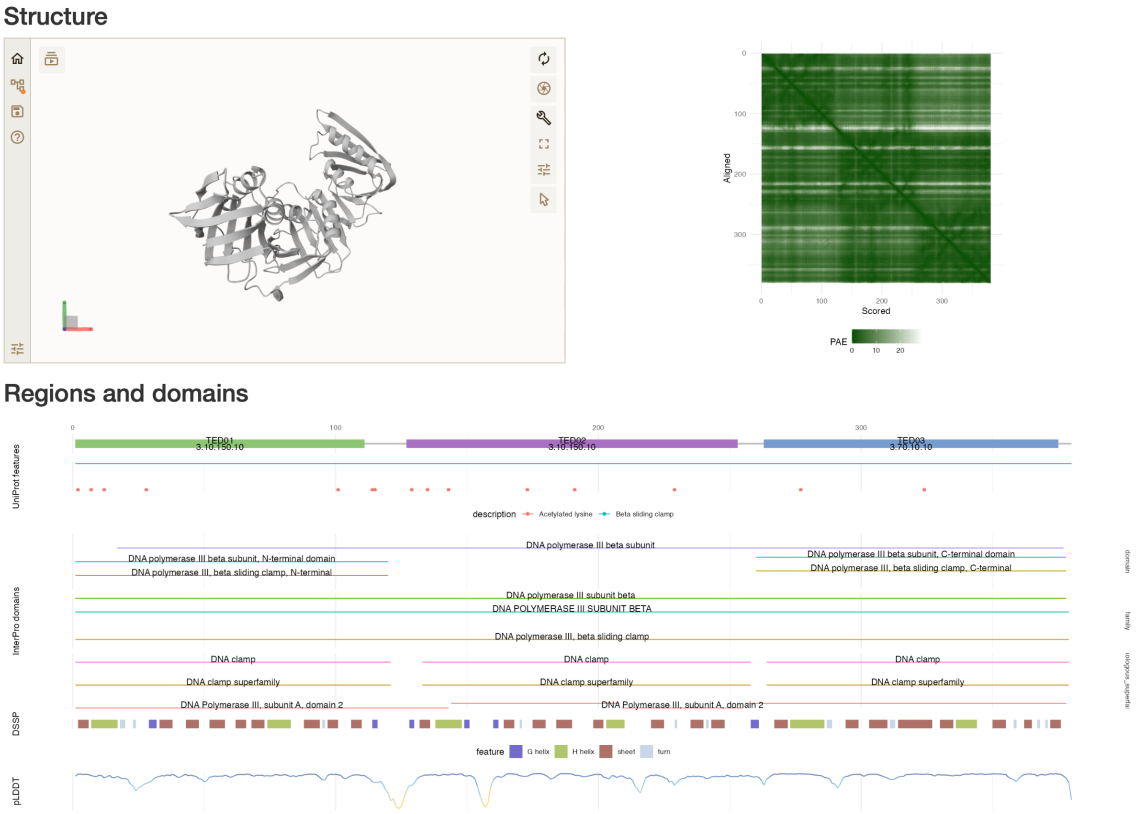


Figure 3.4: The AlphaFold2-predicted structure, predicted aligned error (PAE), and annotated sequence features of dnaN (MPN001).

strengths and weaknesses. Moreover, this data would normally be scattered in several websites or journal articles. Collecting them so that they can be viewed at a glance better reflects the interests of my colleagues who work on *M. pneumoniae* and me. We are often interested in a single protein, and the ability to view it from multiple angles is very useful. Conversely, databases like UniProt are more concerned with covering all the known proteins in a uniform way.

The next section of the protein page is an actual protein viewer, showing the structure and all features that can be mapped to the protein's sequence (fig. 3.4). The structure, which comes from the AlphaFold database [60, 56], is shown in a custom Mol* plugin (fig. 3.4) [139]. At the time of development, the only "off-the-shelf" option to use Mol* in Shiny [144] was the *shiny.molstar* package from Appsilon. However, it didn't have all the features I needed. Specifically, it didn't allow changing the coloring of custom regions of the protein's representation. Thus, I had to develop my own plugin (in TypeScript) and integrate it with Shiny (code is available on GitHub https://github.com/fmarotta/molstar-shiny). Next to the protein structure, the PAE is also shown. This plot is part of the output of AlphaFold2, and it shows the expected positional error at residue X, measured in Ångströms, if the

predicted and actual structures were aligned on residue *Y*. The PAE plot can sometimes be helpful in identifying the domains of the protein [145].

The main visualization is the regions and domains viewer. This panel shows features along the protein sequence. As for the annotation, these features also come from multiple sources. The Encyclopedia of Domains (TED) [146] is a recent project which, leveraging the large amount of protein structures in the AlphaFold database, used machine learning to segment the individual domains. Although it is not the first attempt at domain segmentation [147], it is so far the best and most comprehensive effort. The domains identified by TED have reasonably high quality, although they don't always agree with what a human expert like Rasmus would say. This is a crucial innovation compared to classic, sequence-based domain-family databases such as PFAM [81]. Sequence-based domains can only be continuous along the protein's sequence, and all sequence similarity search tools penalize gaps in the alignment, making it nigh impossible to detect domains that are discontinuous. I experienced this issue first-hand in the dome-complex project (chapter 2). Structural domains do not have this limitation: the peptide chain of a domain can be interrupted by a large disordered loop, or even by a different domain, before coming back and continuing to fold into the former domain. Thus, many TED domains are not continuous along the amino-acid sequence. This is one of the reasons why TED captures 365 million domains, around 100 million of which were undetected by sequence homology methods. Another reason is that structure is, according to some estimates, 3–10 times more conserved than sequence [49], allowing structure-based similarity methods to capture more distantly related proteins into the same family. TED domains and, if available, their CATH number, are displayed as the first track (fig. 3.4).

Next, we combine features from UniProt with features extracted from the literature. The main reference is Chen et al. [135], who measured lysine acetylation. Thus, in this track, we show post-translational modifications, signal peptides, disordered regions, and high-confidence domain annotations.

Next, we show a track from InterPro [79], a resource that already integrates data from more than 10 member databases. So far, these databases rely solely on protein sequences, and they utilize predictive models—such as profile hidden Markov models (HMM), position-specific scoring matrices, and regular expressions—to search similar sequences and assign potential functions. This is the same principle used to annotate protein functions, but in this case, the output also consists of precise coordinates along the protein sequence, denoting, for example, the start and end of the domain.

The next track is from DSSP [148], and it shows the position of helices, sheets, and turns. Last, we show the predicted local distance difference test (pLDDT), another metric provided by AlphaFold2 that measures confidence in the local structure, estimating how well the prediction would agree with an experimental structure.

3.4 Gene expression analysis and visualization

Gene expression is a fundamental process that sits midway between the genome, which carries the information or functional potential, and the final active molecules in the cell, proteins, which determine the phenotypes of the cell. It is a highly regulated process, and *M. pneumoniae* is known for utilizing several non-conventional regulatory mechanisms, such as DNA supercoiling, non-coding RNAs, codon adaptation, and GC content [19]. By studying gene expression patterns, we can gain a deeper understanding of the adaptive responses of *M. pneumoniae* to environmental changes. Mycoboard faithfully reports the results from Yus et al. [19], a landmark study which identifies regulatory proteins, reconstructs the regulatory network, and provides insight into alternative regulatory mechanisms. Moreoever, I downloaded the raw RNAseq data and reanalyzed them, enriching the original results with new analyses.

The group of Luis Serrano performed over 150 RNAseq experiments, collecting data from more than 50 environmental conditions or perturbations. Some of the perturbations consist in the over-expression or knock-out of putative regulatory genes (introduced in the cells through plasmids). Other perturbations consist in modifying the temperature or growth medium. Yet others entail administering a drug. By analyzing the changes that *M. pneumoniae* experiences after being subject to these conditions, we can learn a lot about the regulatory patterns in the cell. One of the main conclusions of the study is that alternative mechanisms, not mediated by transcription factors, are widespread.

I processed these data as follows. First, the IDs of the samples deposited in the European Nucleotide Archive (ENA) [149] were collected from the article by Yus et al. [19] and integrated with the IDs communicated by Marc Weber, a researcher in the group of Luis Serrano. Marc Weber's input was instrumental in annotating the conditions under which each experiment was performed. I used the nf-core fetchngs pipeline [98, 150] to download the FastQ files from ENA [149]. This tool requires only the sample IDs as input, and automatically downloads the sequencing data and metadata. Then, I used the nf-core rnaseq workflow version 3.5 [98, 150], a state-of-the-art pipeline to count gene/isoform abundances and perform extensive quality controls. The workflow processes raw data from FastQ inputs, performs basic quality checks, trims the adapters from the reads, aligns the reads to the reference genome, and generates relative gene counts, performing additional quality-control on the results. The pipeline is built using Nextflow, a workflow tool to run tasks across multiple compute infrastructures in a very portable manner. It also comes with docker containers making installation trivial and results highly reproducible. The input to this workflow consisted of a Fasta file for the reference *M. pneumoniae* genome, downloaded from NCBI under accession GCF_000027345.1, a general feature format (GFF) file with the genome annotation, and the raw reads in FastQ format. I crafted the GFF to include all protein-coding sequences and all known non-coding RNAs. The only non-default parameters I specified are `--skip_bbsplit` and `--skip_rseqc`. BBsplit is a tool used

Figure 3.5: PCA plot of the gene expression matrix, before and after applying SVA correction with one surrogate variable. The PCA is obtained from the variance-stabilized gene counts.

in metagenomics to remove non-reference genome reads. RSeQC is used to identify strading information in RNAseq. Both tools were deemed not useful for our purposes and omitted from the workflow.

The output of the rnaseq workflow was a file with the count of reads aligned to each gene, whether protein-coding or non-coding, for each experiment, including all biological replicates. The gene expression matrix is a basic starting point for further analyses, and is also available to download as-is in Mycoboard. Although the raw reads are available in the ENA, and in principle anyone could reproduce the analysis, starting the analysis from scratch is tedious and sometimes prohibitive. Thus, having access to the precomputed counts matrix is a step forward in the direction of reproducibility of results and flexibility in allowing other people to create their own downstream analysis. Moreover, reanalyzing the data allowed me to include non-coding RNAs in the analysis of differentially expressed genes, while the original study focused on coding sequences.

For the differential expression analysis, I used DESeq2, a state-of-the-art method that provides a comprehensive framework for analysing gene expression data (and more) [151]. Some conditions, due to the large-scale changes that they induce, were removed from the analysis to avoid artifacts that are not due to gene regulation, but to global effects of the perturbations. These conditions are: treatment with novobiocin, an antibiotic that inhibits the DNA gyrase gyrB and causes global decrease in the expression of all genes; glucose starvation, which also causes widespread drop in gene expression; and experiments

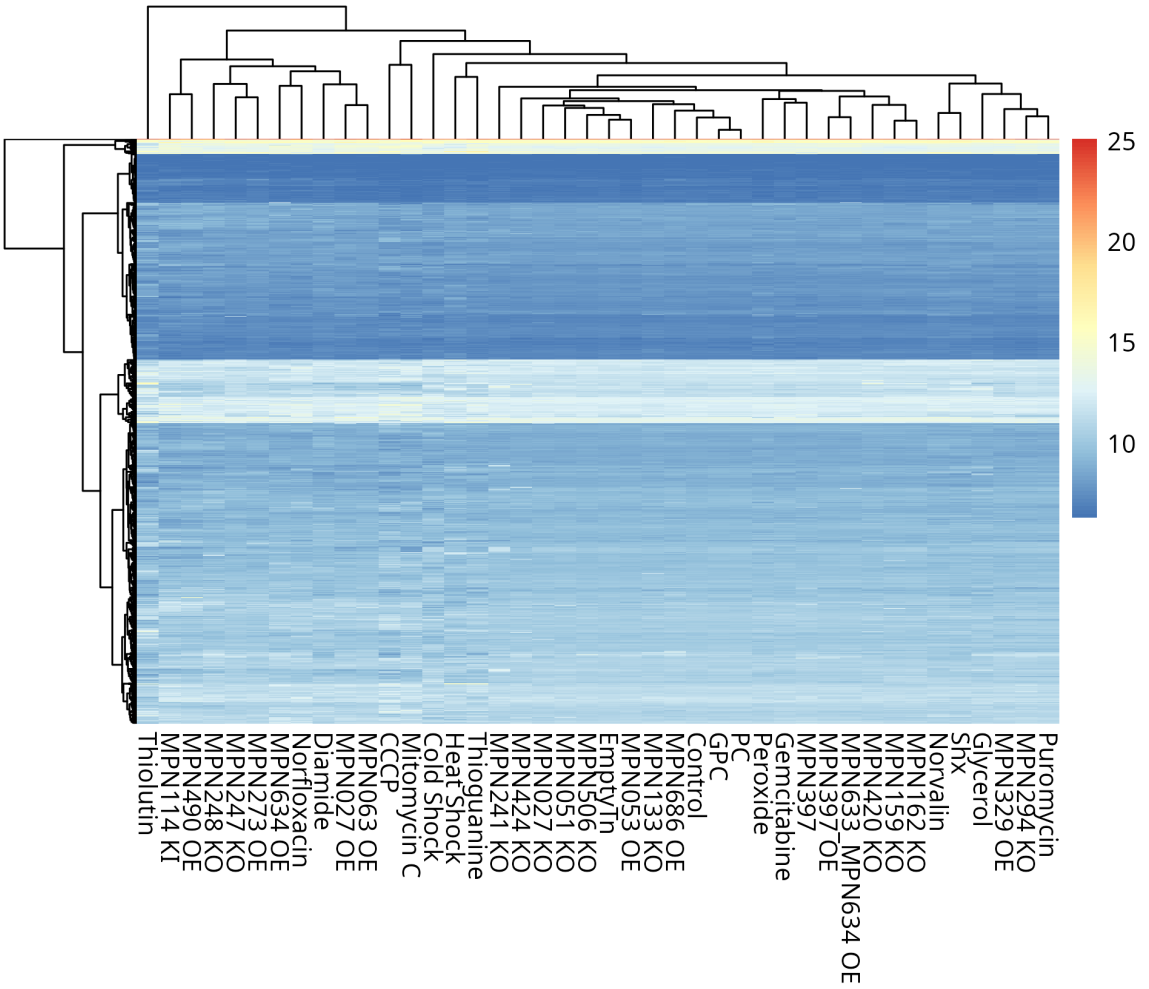


Figure 3.6: Heatmap of gene expression across conditions. Each row is a gene, and the cells of the heatmap are colored according to the TPM.

involving the knock-out or over-expression of MPN545 (ribonuclease 3), which degrades RNA. Moreover, we restricted the analysis to genes that are at least 150 bp long and have a count of more than 5 reads in at least 2 samples. Before running the differential expression analysis, it is important to remove batch effects and unwanted variation due to technical artifacts [152]. I used the SVA method, as implemented in the SVA bioconductor package, to address this issue. The package's automated analysis didn't find any significant batch effect or technical artifacts, but we still applied the SVA method assuming one surrogate variable. Based on the PCA plots, the effect of the transformation was very minor (fig. 3.5).

I then used the DESeq2 bioconductor package to find differentially expressed genes. Specifically, I computed the log2 fold-change (and associated standard error and P-value) of each gene in each condition, with respect to the "Control" condition. Rather than computing the average fold-change "manually", I used DESeq2 because of its rigorous statistical

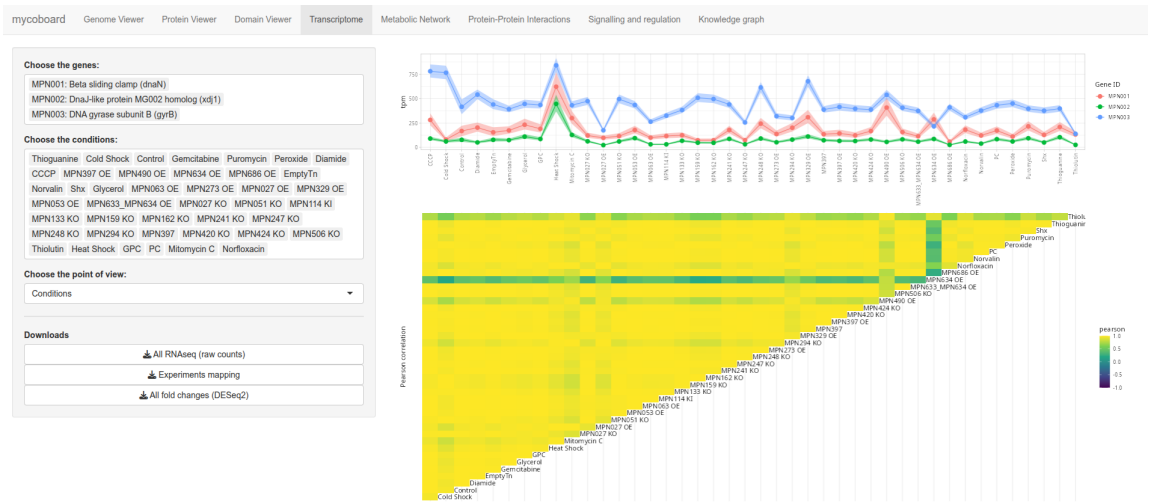


Figure 3.7: The transcriptome profile for three genes: MPN001, MPN002, and MPN003. The top part shows the TPM count across conditions. The points and solid line are the point-estimates, while the shaded ribbon shows the standard deviation. The bottom plot shows the correlations among conditions. Each condition is identified by a vector of n genes, and we compute the correlation matrix from these vectors. High values indicate that the expression of all genes changes in the same way for both conditions; low values indicate that some genes have different trends.

framework. Indeed, it employs a model based on the negative binomial distribution, which correctly handles biological replicates and accounts for biological variability across the whole data set [151]. Thus, its estimates are more accurate and less prone to biases like library size differences. The estimated fold-changes and their standard errors, together with the base mean in the control condition, were used to reconstruct the gene counts in each condition. These new values differ from the original counts in that the experiments (biological replicates) pertaining to the same condition are aggregated, based on the DESeq2 estimates of fold-change and standard error. The new gene counts were converted to TPM, which normalizes the count by the length of the gene and the total size of the library, allowing for natural cross-condition comparisons [153]. Figure 3.6 shows a heatmap of the gene expression matrix.

For Mycoboard, I provide a visualization of the expression profile of a group of genes. Users can select any number of genes, either coding or non-coding, and look at how their expression changes across conditions (fig. 3.7, top panel). Furthermore, it could be interesting to compare the similarity of the conditions with respect to the expression of the selected genes. The bottom panel of fig. 3.7 shows the Pearson correlation matrix between samples.

A complementary view of the same data can also be displayed (fig. 3.8). Here, the data is the same, but the genes, rather than the conditions, are on the x axis. The correlation matrix is also computed for the genes, using the vector of their expression across conditions.

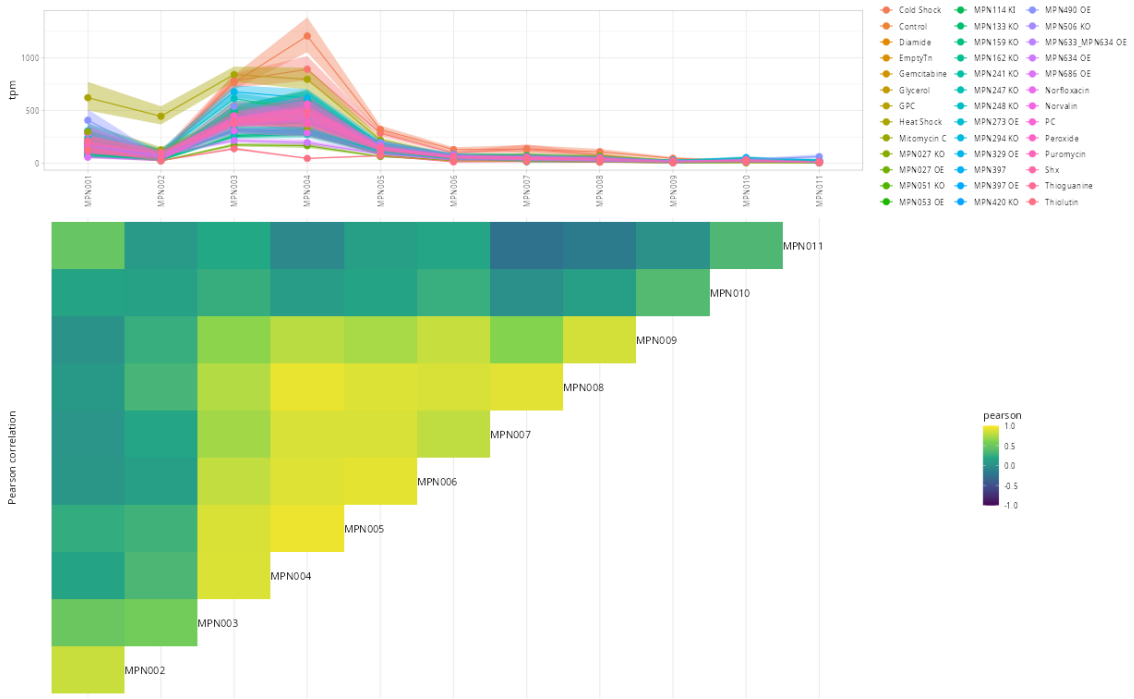


Figure 3.8: The transcriptome profile for the genes: MPN001–MPN011. See also fig. 3.7. Here, the genes MPN003-MPN009, which are in the same operon, are more highly correlated among themselves than with the other genes. The top plot also shows the staircase-like decay of expression within operons.

This view quantitatively highlights groups of genes that are potentially co-regulated, for example because they are in the same operon, or because they need to be in the same protein complex. Users can also select a subset of conditions of interest through a menu on the right of the page.

3.5 Signalling and regulation

The next page in Mycoboard shows "signalling and regulation". Here, the source of data is again Yus et al. [19], but instead of reanalyzing the data, we simply reuse what the authors of the study have reported. In particular, two key experiments were performed. First, they identified 23 high-confidence transcription regulators. Strains that overexpressed these proteins, as well as transposon or dominant-negative point mutant strains, were used to assess which genes significantly change their expression as a response to the over-expression or knocking-out of these regulators. Detecting these changes allowed the authors to identify the target genes that are modulated by each of the regulators. Second, they exposed the cells to 37 environmental perturbations, observing significant changes in gene expression for 31 of them. By analyzing which genes exhibit an alteration in their expression, it is possible to infer which genes are affected by each perturbation.

The final gene regulatory network, including 23 regulators (among which 9 are transcription factors), is shown in fig. 3.9.

Figure 3.10 shows the genes whose expression is significantly altered upon treatment with the antibiotic Spectinomycin.

Together, these two graphs, the gene regulatory network and the network of responses to perturbations, could provide insights into the function of uncharacterized proteins, as I will show later in a case study section 3.10.

3.6 Metabolic network

Metabolic modeling offers a quantitative framework to decipher the complex interplay of reactions that drive growth and homeostasis in the cell. The first study which I integrated is Yus et al. [15], who manually curated 189 reactions catalyzed by 129 enzymes. More recent studies, namely **wooke_2013** and Gaspari et al. [6], obtained improved metabolic models that include more reactions, their free energies (important to assess the direction of the reaction), and their fluxes. The data from Gaspari et al. [6] are also available in Mycoboard. For example, fig. 3.11 shows how the guanine metabolism is displayed. This visualization was obtained with the Escher software [6, 154].

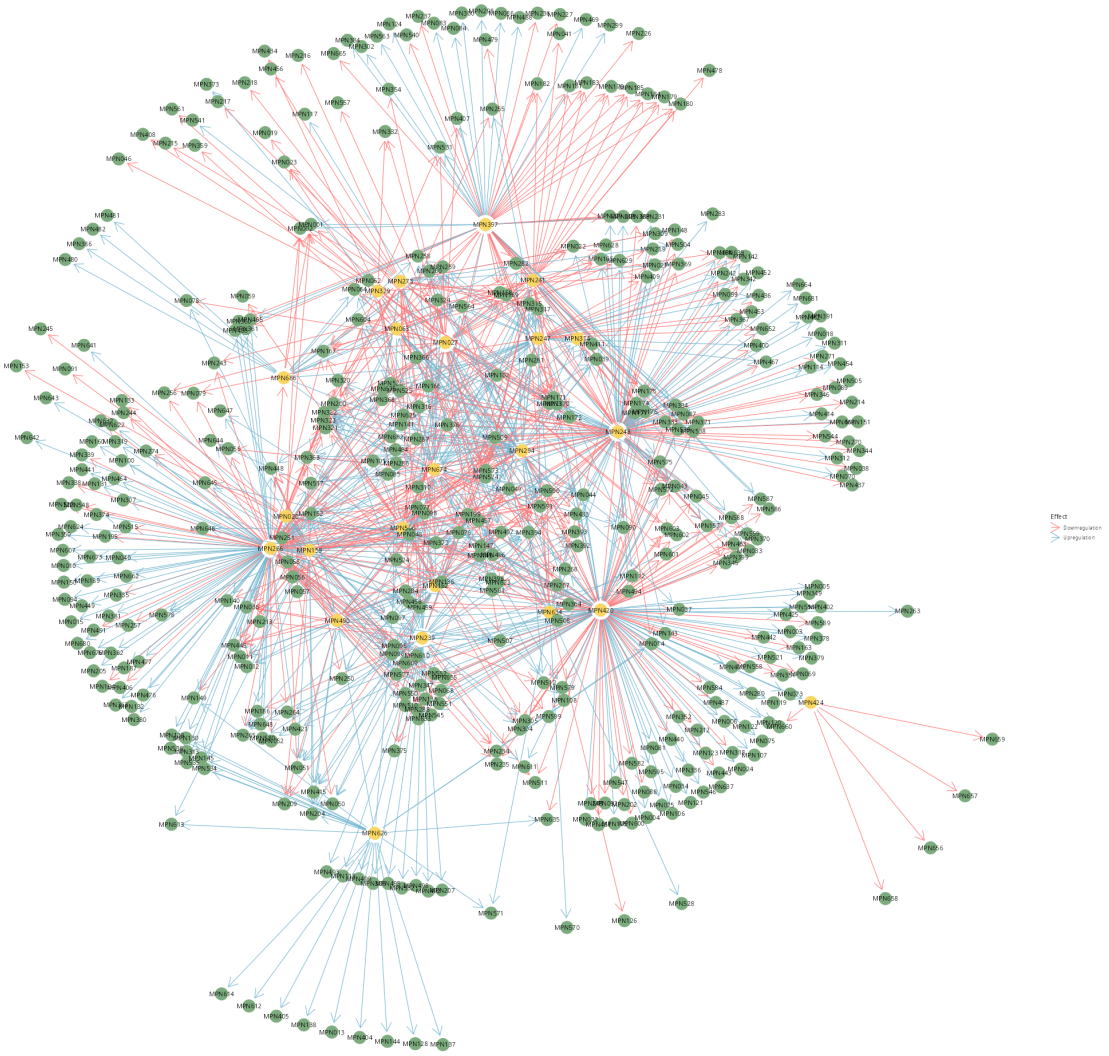


Figure 3.9: Gene regulatory network as reconstructed in Yus et al. [19]. Regulators are shown in yellow, while targets are green. Blue arrows denote upregulation, while red arrows denote downregulaton.

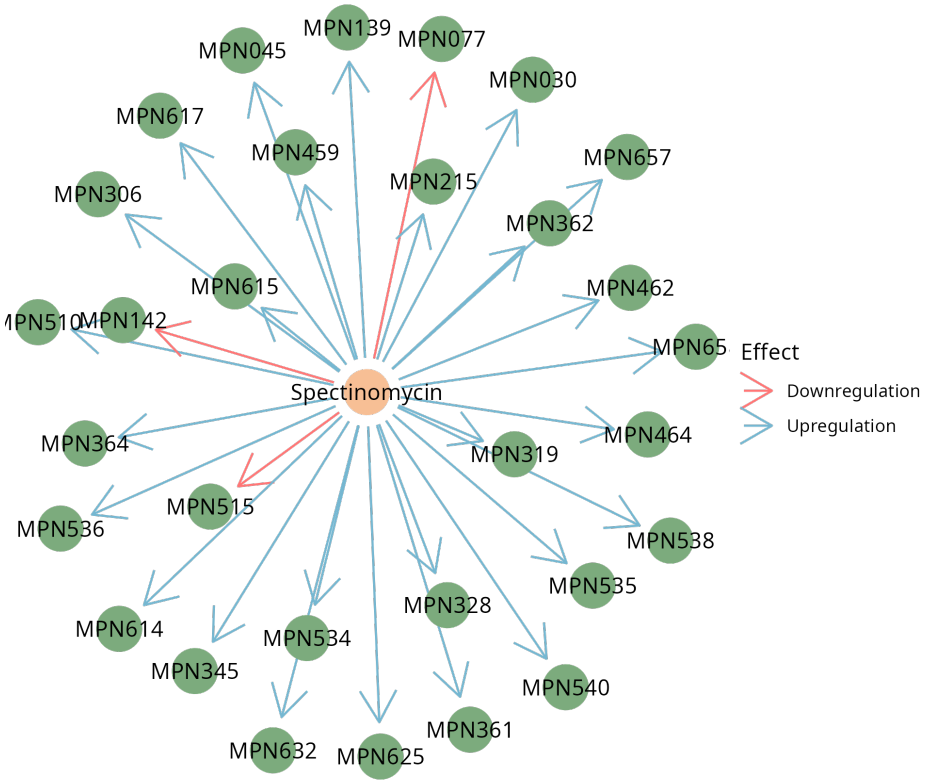


Figure 3.10: The genes whose expression is modulated by the antibiotic Spectinomycin.

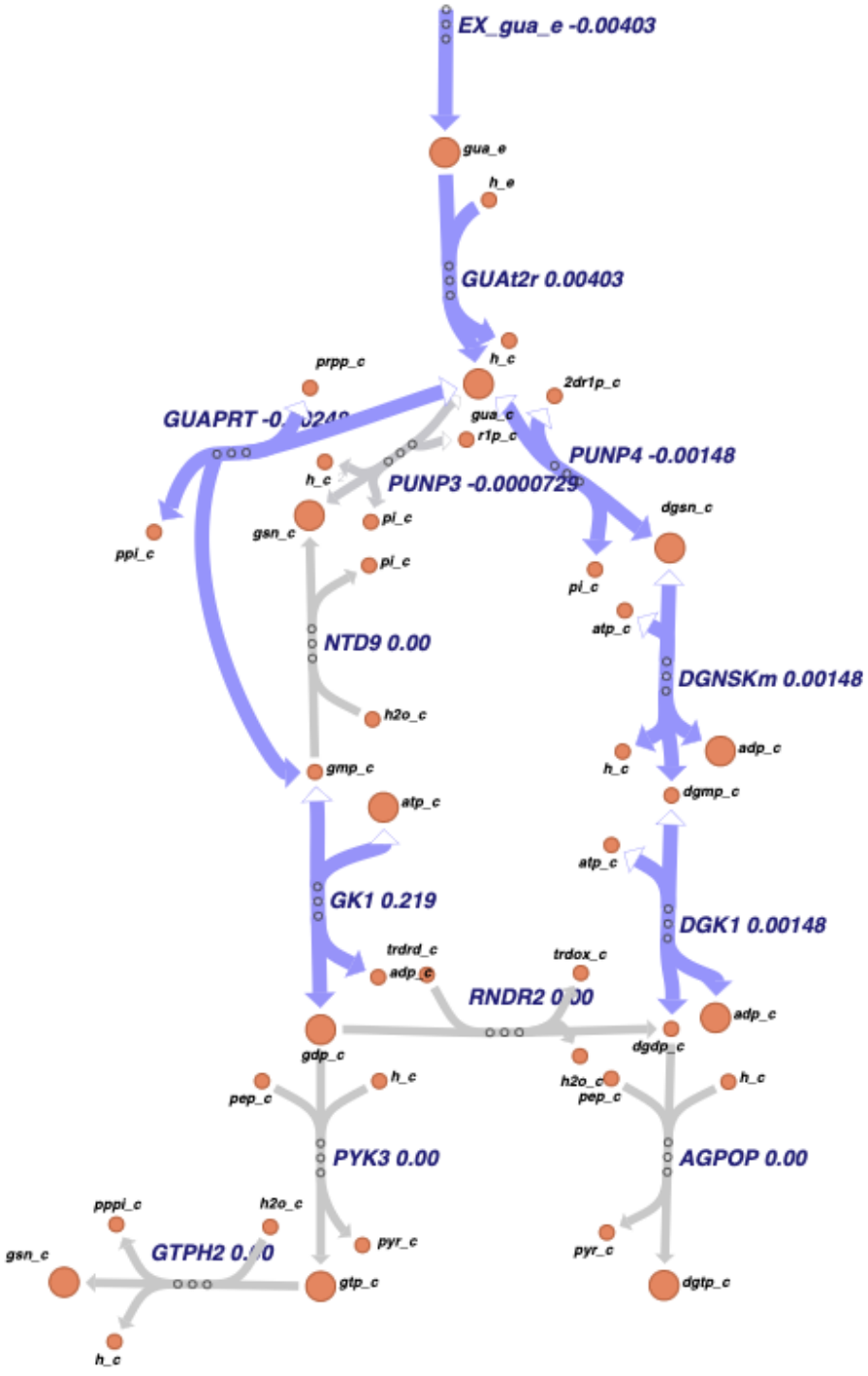


Figure 3.11: Detailed view of the guanine metabolic pathway from Gaspari et al. [6]. The color of the arrows is proportional to the flux of the reaction; gray reactions are not active in the baseline condition.

3.7 Protein-protein interaction networks

At the time of writing, two large-scale experiments investigated protein-protein interactions in *M. pneumoniae*. One is Kühner et al. [17], who used TAP. TAP is a technique used to isolate protein complexes from cells. A protein of interest is tagged at the C-terminus with a protein sequence that includes three parts: a calmodulin binding peptide, a protease cleavage site, and two Protein A domains, which bind to IgG [155]. The cell extract is first combined with agarose beads covered in IgG, then the first part of the TAP tag is cleaved, and a second filtering is performed, this time with agarose beads covered in calmodulin. The beads are separated by centrifugation, which is gentle enough to preserve the interactions between the protein of interest and its partners. For the 2009 study, the assay was performed for the whole *M. pneumoniae* proteome, collecting thousands of potential interactions. Each candidate interaction was given a socio-affinity score representing the strength of the interaction. Finally, the complexes were called by using the clique percolation algorithm on interaction network. Clique percolation starts by identifying all the k-cliques, and it proceeds by merging them if they share $k-1$ nodes. Importantly, this method allows nodes to be part of multiple communities. Indeed, one of the findings of that study was the high number of "moonlighting" proteins, *i.e.* proteins that are part of multiple complexes. Despite the two sequential purification steps, TAP still remains a relatively noisy method, where unrelated proteins have a good chance of being purified. Moreover, membrane proteins are typically underrepresented due to the challenges associated with their solubility.

The second study is O'Reilly et al. [21], who used in-cell cross-linking followed by mass-spectrometry. First, interacting proteins are chemically linked with cross-linkers that covalently bind to specific amino acid residues. The cross-linked protein complexes are then digested into peptides and analyzed using mass spectrometry to detect cross-linked peptides, which provide information about the spatial proximity of the interacting residues. By analyzing the mass spectrometry data, it is possible to infer the interaction sites and map the protein interaction network within the complex, and also to reconstruct the topology of the complex.

In Mycoboard, the protein-protein interaction viewer has three components. First, I show a summary plot of abundance vs mass for the known or predicted complexes (fig. 3.12). For some homomultimeric, and most heteromultimeric complexes, the stoichiometry of subunits is not known. Hence, the masses and abundances are only meant as an indication. For heteromultimeric complexes, the abundance is estimated as the median of the components, while the mass is estimated as the sum of the components. For homomultimeric complexes, the abundance is the abundance of the protein divided by the number of subunits, while the mass is the mass of the protein multiplied by the number of subunits. This plot should help structural biologists to prioritize uncharacterized protein complexes for further investigation. Indeed, in the tomograms, it is more likely to identify structures that are large and very abundant.

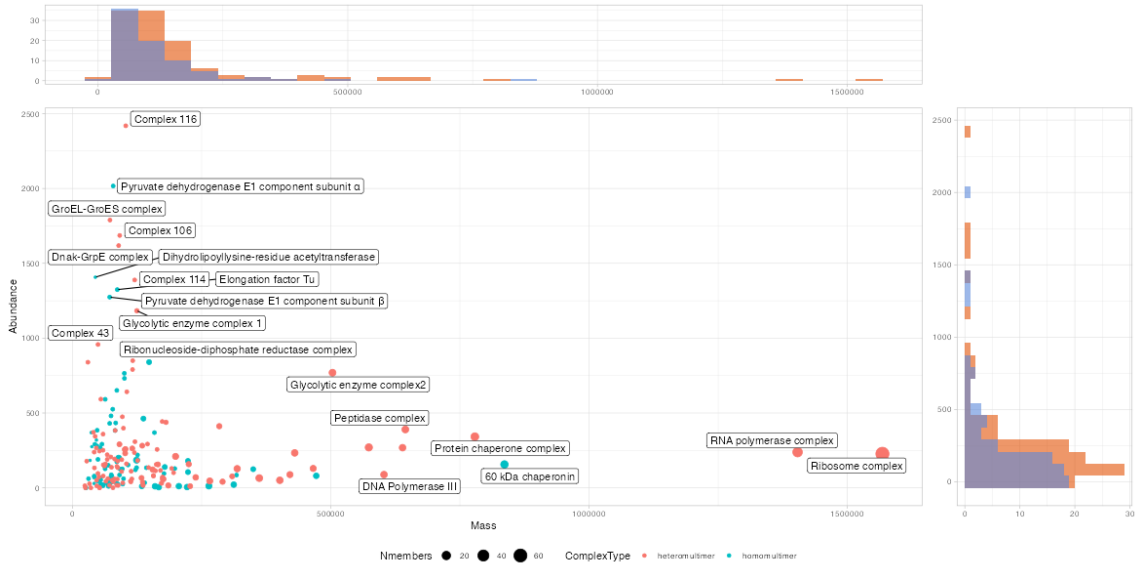


Figure 3.12: Summary of the homo- and hetero-multimeric complexes from the TAP experiment. The complexes are arranged by abundance vs mass to facilitate the prioritization of complexes that are either very abundant or very large. The histograms to the top and to the right show the distribution of mass and abundance, respectively, for homo- (blue) and hetero- (orange) multimeric complexes. Abundance is expressed in copies/cell, while mass is in dalton.

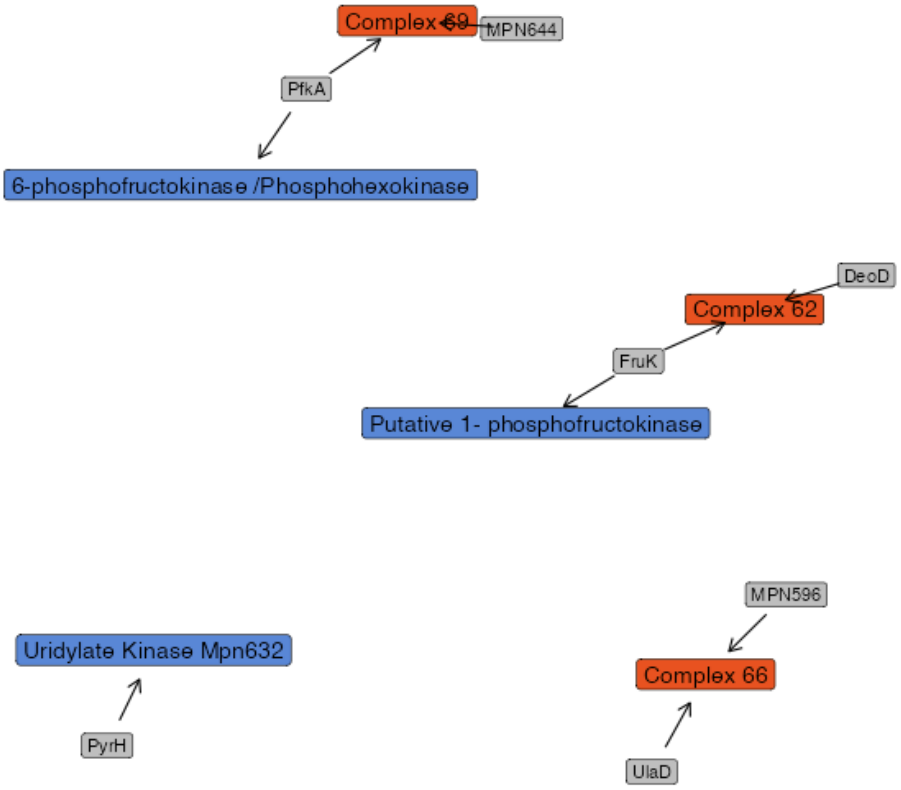


Figure 3.13: Detailed view of four TAP complexes. Heteromultimers are colored in red, homomultimers in blue, and their member proteins in gray. Users can change the color of proteins.

The second visualization shows all the complexes that were detected in the TAP experiment as a graph. The full graph is too big to show here, but fig. 3.13 presents a small corner consisting of four complexes. In this visualization, I don't depict the interactions between proteins directly. Rather, I build a bipartite graph with nodes for both proteins and complexes, and edges can only exist between a protein and a complex.

Last, I show a graph of the protein-protein interactions captured by the cross-linking mass-spectrometry experiment [21]. The data come from a combination of two cross-linking agents, BS3 or DSSO, filtered at 5% false discovery rate (FDR). Users can visualize the full network and color proteins by several features, including molecular weight, essentiality, protein copies, RNA expression, and function category (fig. 3.14).

3.8 Integrating all data modalities in a knowledge graph

So far, I have discussed each experiment in isolation. This is also the prevalent approach in the literature, where one article typically focuses on one data modality: protein-protein interactions, RNAseq, or metabolism. However, the cell is made of molecules, and they interact based on the laws of physics and chemistry irrespective of how we humans label them. Proteins can bind DNA or RNA, be modulated by ions, interact with metabolites, and so on. Thus, I advocate for a view where we strive to make a census of all the components of the cell and how they are related to each other, rather than analyzing each class on its own. This view does not preclude a multi-level approach. On the contrary, a hierarchical approach is likely to be the only way to tackle such complex systems. What I mean by this is that we recognize that the components can be broken down into smaller parts in a reductionist approach, but treating the high-level components as units in their own right is sometimes necessary. For example, proteins can be broken down as polymers of amino acids, amino acids can be seen as a group of atoms, atoms are made of subatomic particles. When we analyze a biological system such as *M. pneumoniae*, we don't have to necessarily think about electrons and quarks, but we can study the behavior of proteins treating them as coherent units. Abstractions like this are crucial in modelling and science in general. The very difficult part is to decide which levels of abstraction are important in our view of the system, and find out how they are related. To tackle this problem, I developed an integrated knowledge graph of all available data for *M. pneumoniae*.

A knowledge graph (KG) is a structured representation of data and its relationships, essentially a semantic database [156]. It organizes information by representing entities (nodes) and their relationships (edges), allowing for deeper understanding and better search capabilities. In this regard, it needs to be built on top of an ontology, which serves as the schema for the KG. The importance of ontologies in modern biology can hardly be overstated [157], as exemplified by projects such as the Gene Ontology [142]. They provide a structured approach to organizing knowledge by standardizing terminology

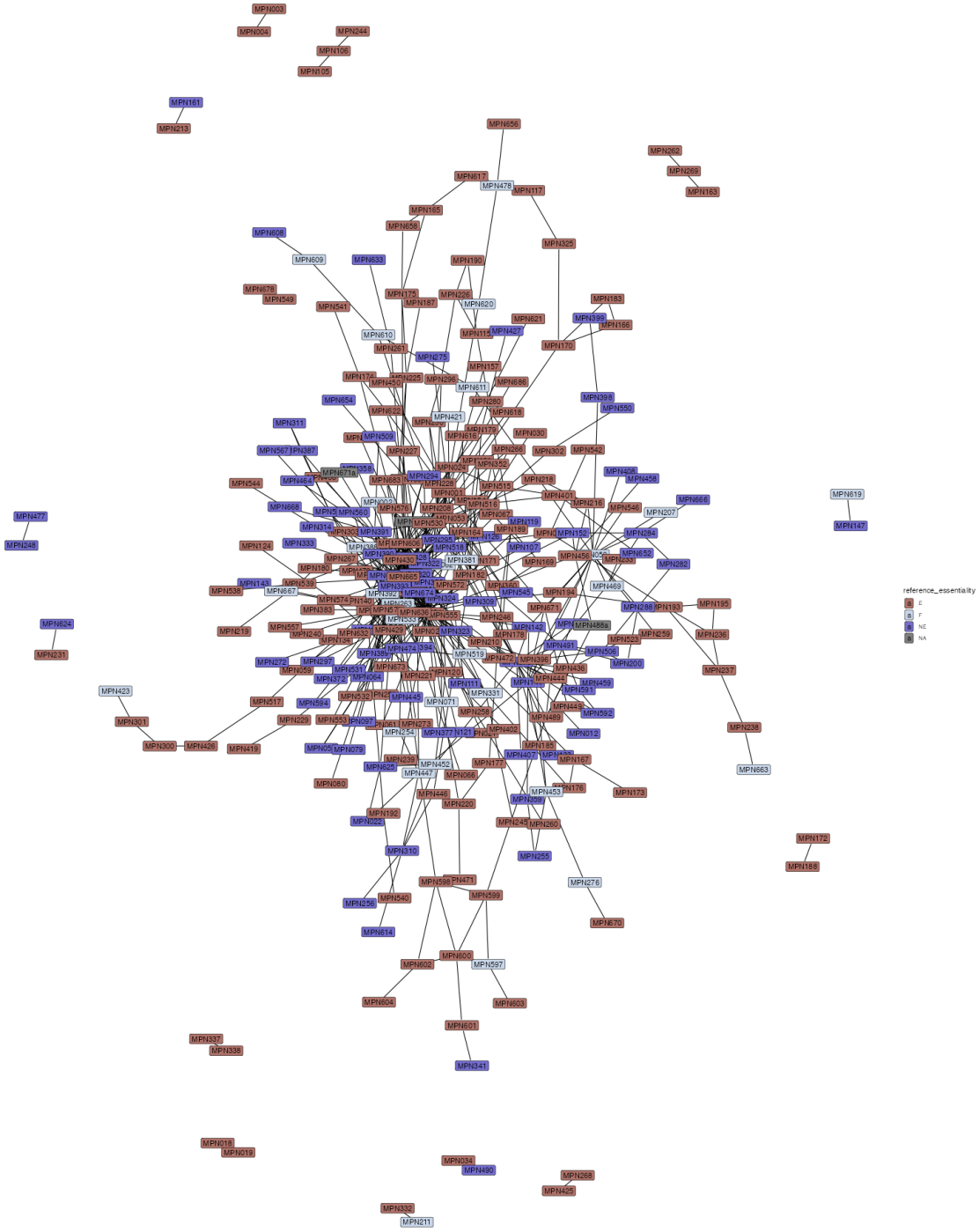


Figure 3.14: Cross-links in *M. pneumoniae*. The proteins are colored according to essentiality. E: essential; NE: not essential; F: reduced fitness phenotype; NA: missing data.

and encoding relationships such as "is_a" and "part_of", which enable hierarchical and transitive connections between concepts. This standardization addresses the challenge of navigating heterogeneous biological data, where inconsistent descriptors (*e.g.*, "bud development" versus "limb morphogenesis") hinder efficient data retrieval and reasoning. The Gene Ontology, established in 1998, exemplifies this by classifying gene products across cellular component, molecular function, and biological process domains, providing a consistent nomenclature that is friendly to both humans and computers [142]. The Gene Ontology classification is also hierarchical, which enables analyses such as finding enriched terms among the differentially expressed genes between two conditions [158]. Developing an ontology is hard because biological knowledge needs to be embedded in it. A quick literature search retrieves several examples of bespoke biological ontologies and knowledge graphs developed for the life sciences [159, 160, 161, 162, 163]. Recently, ontologies and knowledge graphs are growing more and more popular, also thanks to projects such as the Open Biological and Biomedical (OBO) foundry [164], which promotes the development of interoperable ontologies.

The relationship between an ontology and a knowledge graph is akin to the relationship between a database schema and its content. In this analogy, the ontology provides the structured framework or schema, while the knowledge graph serves as the container that holds the actual data, organized according to this framework. This structured organization of data offers several advantages. Firstly, it simplifies querying processes. By filtering based on specific entities and/or their relationships, we can efficiently extract relevant information, much like using semantic web technologies such as SPARQL [165]. This capability is particularly valuable in navigating the complex and interconnected data typical of biological systems. Moreover, graphs lend themselves very well to the visualization of data. Traditionally, knowledge graphs have been widely used by Google to aid searches and answer user queries that asked questions in natural language [166]. The slogan used by Google was "things, not strings", since the knowledge graph approach helps computers reason about concepts rather than their human language representation. Similarly, it is conceivable to develop a search engine for biological data that answers questions leveraging *M. pneumoniae* data.

Secondly, the structured nature of knowledge graphs and the integration of multiple modalities facilitates advanced analyses. Recent advancements in graph neural networks enable sophisticated tasks such as node classification and link prediction [167]. These methods can be leveraged to predict interactions within the biological system, such as protein-protein interactions, thereby enhancing our understanding of cellular processes.

Lastly, knowledge graphs can also support simulation approaches. A whole-cell model of JCVI-syn3a, an artificial minimal cell built from *Mycoplasma mycoides*, has already been built [8]. Such model uses cryo-electron tomograms for the cell geometry and ribosome distributions, and relies on kinetic models of around 2,000 reactions to capture the unfolding over time of DNA replication, transcription of all 493 genes, translation and degradation of

Table 3.1: List of entities in our knowledge graph and their associated instance count.

Entity	N
chromosome	1
curated_function	46
gene	1107
gene_ontology	362
kegg_pathway	60
metabolic_complex	5
metabolic_reaction	181
metabolite	208
operon	664
perturbation	20
protein	735
suboperon	856
TAP_heteromultimer	116
TAP_homomultimer	78
TED_domain	1266
transcript	945

all 452 mRNA, tRNA charging, and cell growth. A prerequisite for developing whole-cell models is the knowledge of all the components of the system and all possible reactions. At the same time, whole-cell models depend on quantitative rules and differential equations that, while indispensable for modelling changes over time, are not optimized for query and visualization. By abstracting the rules governing such a minimal cell system into a knowledge graph, it becomes possible to visualize and explore the data in a qualitative way that is sometimes easier to reason about. Thus, whole-cell modelling and knowledge graphs complement each other.

M. pneumoniae, with its reduced genome, is already close to being a minimal cell, with only roughly 700 protein coding genes. It serves as a good starting point for making a census of all the entities in a cell and all the ways they can interact among each other. In this work, I established the first steps for the analyses described above. Specifically, I designed an ontology based on the available data for *M. pneumoniae* and built an associated knowledge graph. The whole graph is part of Mycoboard, where it can be downloaded and conveniently visualized. In total, it consists of 6650 entities and 19265 relationships. Tables 3.1 and 3.2 show the design of the *M. pneumoniae* ontology. For the entities, I settled on 16 classes that in my opinion represent the most useful levels of abstraction, based on the available data and the expected analyses that we could perform. Importantly, the graph mixes concrete physical entities such as chromosome, proteins, and metabolites, with more abstract entities such as genes, along with completely made-up concepts like KEGG pathways and functional categories [130]. Although we will never find a physical KEGG pathway in the cell, aggregating the proteins that are annotated as belonging to the same pathway is extremely useful. In the graph, those proteins will be all linked to the KEGG

Table 3.2: List of relationships in our knowledge graph and their associated instance count.

From	To	Relationship	N
TAP_heteromultimer	chromosome	binds_DNA	1
TED_domain	protein	is_in_protein	1266
gene	chromosome	is_at_locus	1107
gene	gene	downregulates_expression_of	498
gene	gene	upregulates_expression_of	562
gene	gene	ncRNA_overlaps	352
gene	operon	is_in_operon	1092
gene	protein	produces	737
gene	suboperon	is_in_operon	1092
gene	transcript	is_in_transcript	1374
metabolic_complex	metabolic_reaction	catalyzes	6
metabolite	metabolic_reaction	is_reactant_in	718
operon	chromosome	is_at_locus	664
perturbation	gene	downregulates_expression_of	1289
perturbation	gene	upregulates_expression_of	1357
protein	TAP_heteromultimer	is_in_complex	482
protein	TAP_homomultimer	is_in_complex	78
protein	chromosome	binds_DNA	23
protein	curated_function	involved_in_function	733
protein	gene_ontology	belongs_to_gene_ontology	1108
protein	kegg_pathway	belongs_to_kegg_pathway	623
protein	metabolic_complex	is_in_complex	22
protein	metabolic_reaction	catalyzes	125
protein	protein	cross-links	1154
protein	transcript	is_translated_from	1001
suboperon	chromosome	is_at_locus	856
transcript	chromosome	is_at_locus	945

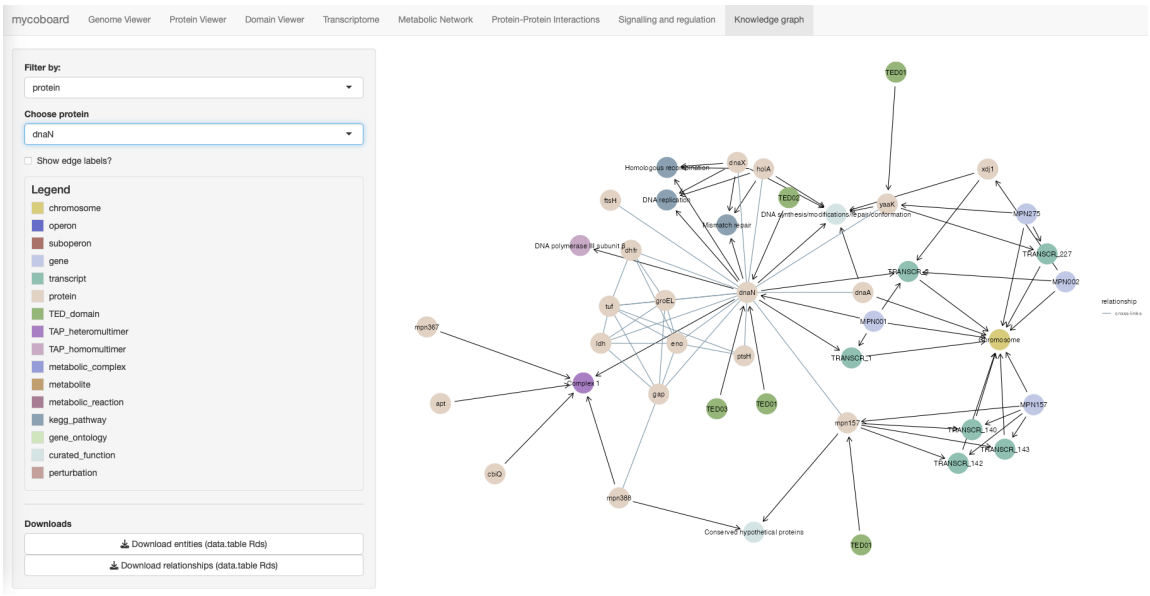


Figure 3.15: Knowledge graph view centered on the protein dnaN (MPN001).

pathway node through a "belongs_to_pathway" relationship. Therefore, those proteins will all be connected by a maximum distance of 2 in the network space, forming a cluster of interrelated entities. It then becomes possible to reason about the group of proteins as a cohesive unit. Mixing different levels of abstraction is one of the strengths of the knowledge graph.

With the relationships, I tried to capture some basic molecular biology and to include as much of the available data as possible. Thus, we have relationships between genes and transcripts ("is_in_transcript") and between transcripts and proteins ("is_translated_from") to model the central dogma of molecular biology. But we also have relationships that capture the fact that some entities are located in a precise subregion of a bigger entity. For example, protein domains are located in proteins; genes are part of operons, which are located on the chromosome. Other relationships capture dynamic processes (such as metabolic reactions) or physical links (protein-protein interactions). Navigating the graph by traversing existing relationships is a powerful way to discover new and potentially unexpected connections.

In Mycoboard, I provide interactive visualizations of subsets of the knowledge graph, where users can decide to focus on particular entities (figs. 3.15 and 3.16). For computationally-oriented researchers, download of the raw data structures is also offered. The entities are represented as an R data.table object with four columns: id, type, label, and properties. The "properties" column is a nested list that contains a variable number and type of elements depending on the type of entity. Similarly, the relationships table has four columns: from, to, relationship, and properties, where from and to contain the entity IDS, and "properties" is again a list with variable number of elements depending on the relationship type. For

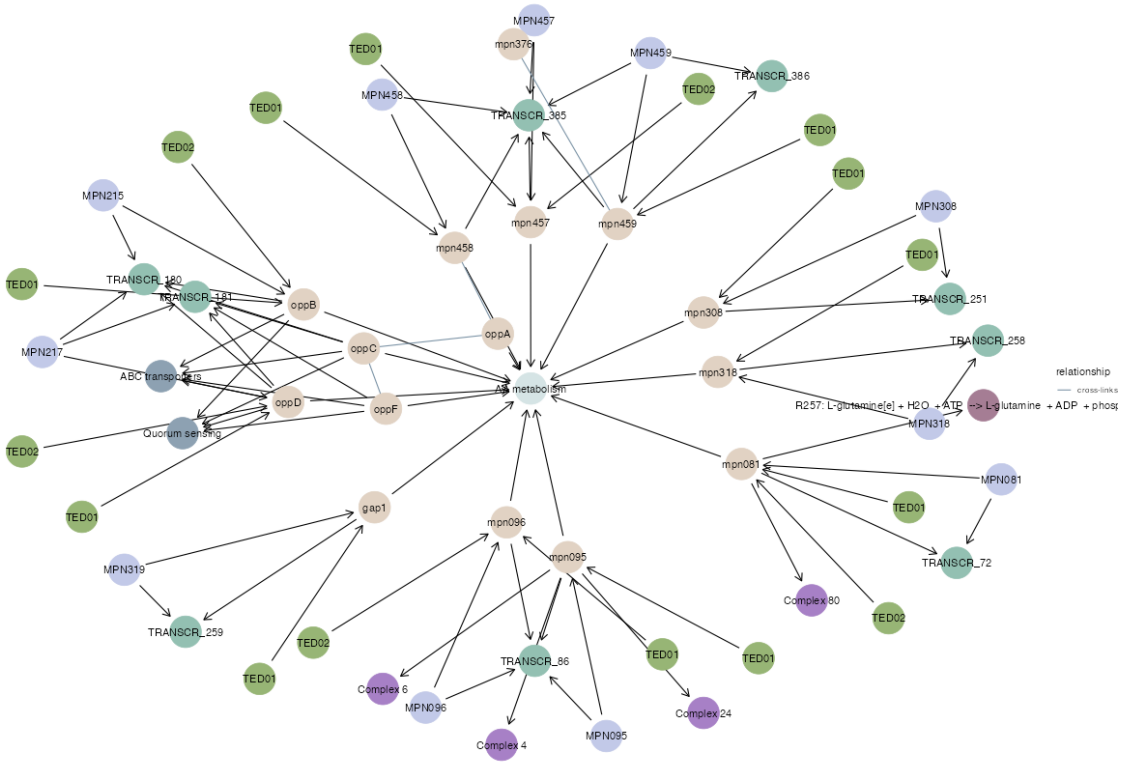


Figure 3.16: Knowledge graph view centered on an abstract entity, amino acid metabolism. The view shows all the proteins that are directly related. Navigating the graph allows users to explore indirectly related entities and discover novel connections.

example, the relationship "is_at_locus" between gene and chromosome has the start and end coordinates as properties, but these properties are not present in other relationships.

3.9 A new approach to the functional annotation of the *M. pneumoniae* proteome

Despite the vast amount of sequencing data generated by modern high-throughput technologies, the functional characterization of proteins remains a significant bottleneck in the field of molecular biology. As of now, a substantial proportion of proteins remain uncharacterized, meaning their functions are not well understood. For example, in the protein universe study [168], which generously defined the functional annotation of a protein based on how much it is covered by InterPro [79], 27% of proteins are completely mysterious, and even for proteins that are covered, many InterPro domains are DUFs (domain of unknown function). Even in the artificial cell JCVI-syn3, almost one third of the genes have unknown function [10]. This gap arises because while sequencing technologies have advanced rapidly, allowing for the identification of protein-coding genes at an unprecedented scale, the experimental methods required to elucidate protein function, such as biochemical assays, structural biology, and genetic studies, are labor-intensive, time-consuming, and often require specialized expertise. Additionally, many proteins may have context-dependent functions or participate in complex interactions that are challenging to replicate and study *in vitro*. Consequently, the disparity between the number of sequenced genes and the number of functionally characterized proteins continues to widen, highlighting the need for innovative approaches and technologies to bridge this gap and fully leverage the potential of genomic data.

This field, like many others, has recently been shaken by AlphaFold (see section 1.3). Indeed, it is generally thought that structure is more conserved than function [49]. Thus, transferring function through structural similarity rather than sequence similarity should extend our reach much further, allowing us to annotate proteins that bear very low sequence similarity with those that have already been experimentally characterized. And this is true, but, as it still took a disproportionate amount of time and effort for us to discover anything interesting about the Dome complex proteins in chapter 2, it is clear that this method has some limitations. In this section, I propose a new method for the functional annotation of proteomes, a method that is enabled by recent advances that all stemmed from AlphaFold and incorporates the lessons learned from the Dome complex project.

Functional annotation has been one of the most prominent areas of bioinformatics, and many clever techniques have been developed to tackle this problem (see Price and Arkin [118] for a recent review). Many tools have been developed by pioneers like Peer Bork [82, 117, 53]. Direct sequence or structure similarity with a protein of known function is of course the best-case scenario and can be addressed with the EggNOG mapper [169],

but it doesn't always happen. One can then look at more or less indirect evidence, such as predicting the subcellular location of the protein [170], looking at the environments or biomes where the protein is found [47, 46], or finding a conserved gene neighborhood around the protein of interest and hoping that the neighbor genes belong to a known pathway [82].

The method I propose is enabled by a recently published large-scale protein domain annotation: The Encyclopedia of Domains (TED), which was ultimately enabled by AlphaFold. I described TED in section 3.3, since it is also part of the protein viewer in Mycoboard. For the Dome complex project, after identifying the remote homology with PrsA, we confirmed this hit by cutting the domains and submitting them individually to FoldSeek [94]. In that case, our collaborator Rasmus Jensen, a structural biology expert, had to manually identify the domains and cut them out of the PDB file. Here, I developed a workflow that automates the process of retrieving the individual domains and submitting them to both FoldSeek and hhblits for homology detection. First, I downloaded the UniRef30 database version 2023_02 [92], to be used as the reference database for hhblits, and the Alphafold/UniProt50 database (AlphaFold UniProt Protein Structure Database clustered with MMseqs2 at 50% sequence identity and 90% bidirectional coverage) for FoldSeek [171, 60]. Then, I downloaded the PDB files of the protein structures in *M. pneumoniae* from the AlphaFold database and the coordinates of the TED domains. I used a custom script to trim the PDB files into the domains, and also save the amino acid sequences of the domains to separate FASTA files. Then, I run hhblits on each domain sequence against the UniRef30 database, and FoldSeek on each domain structure against the AlphaFold/UniProt50 database. The PDB files of the FoldSeek hits for each domain are also downloaded and given to FoldMason [172], a program that makes multiple structure alignments. The trick used by FoldMason is the same of FoldSeek: the protein structure is represented with a special alphabet where each letter captures a specific pattern of residue interactions. I wrote this workflow in Snakemake [173], a Python-based workflow manager inspired by Make.

The final outputs are a table of hits from hhblits, a table of hits from FoldSeek, and a multiple structure alignment from FoldMason. These visualizations are also integrated in Mycoboard, where users can select a particular protein and a domain within it (figs. 3.17 and 3.18). Since the tables contain many hits, manually going through them is tedious. Thus, I integrated an AI summary (fig. 3.19) that uses ChatGPT [174] to generate a succinct representation. Large language models (LLMs) like ChatGPT excel at summarizing text and aggregating data [175]. In this case, I give the model the description and probability score of all the hits by hhblits and FoldSeek for each domain, and prompt it to give a consensus description for the domain based on the description of the hits. The LLM is asked to give not only a consensus, but also an alternative description based on other hits that might be relevant, a confidence score representing how strongly they feel about their response, a longer text detailing their reasoning process, and pointers to potentially interesting hits that might be relevant despite having a lower score. The results are requested

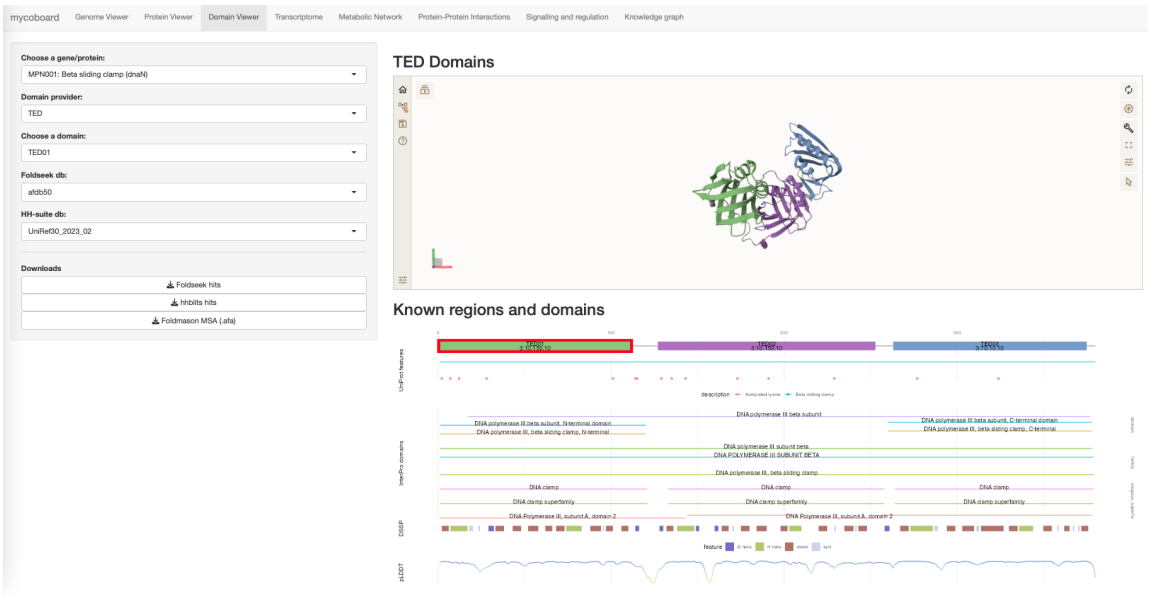


Figure 3.17: The domain viewer in Mycoboard begins with a visualization of the TED domains using the Mol* plugin. Then, it shows the known regions and domains from UniProt, InterPro, and DSSP, as in section 3.3.

automatically on demand, using the OpenAI API, when the Mycoboard page is accessed for the first time. Some prompt engineering is still needed to tune the responses, but overall the summary seems to make sense for the hits we tried. A more traditional approach to propagate annotation is to simply use the top-scoring well-characterized hit, as in Ruperti et al. [61]. However, using an LLM allows us to consider not just one, but multiple hits, and realize when they occur multiple times, which is also a strong sign of potential homology. Furthermore, another strength of the LLM approach is that it should be able to recognize multiple variations of the same name (like "DNA polymerase III subunit beta" and "DNA polymerase III beta subunit"). The LLM can also use its biological knowledge to generalize the terms and suggest interesting hits, although hallucination is always possible [176].

FoldMason also offers an interactive HTML document with visualizations of the structures and the alignment [172]. Based on its alignment, I also developed a view that shows the known UniProt and InterPro domains corresponding to the position of the query domain (fig. 3.20). The advantages of using a structure-based alignment should be clear. Firstly, structure is more conserved than sequence [49], and it is also most often what determines the function of a domain, making it potentially easier to find the correct alignment. Moreover, a structure-based alignment can ignore disordered loops and linker regions that might be misaligned by looking just at the sequence. One of the exciting applications of structural alignments is improving phylogenetic trees, which need to use high-quality alignments as a starting point.

Foldseek hits

Target	Description	Organism	Prob	Evalue	Score	Q. start
AF-Q50313-F1-model_v4	Beta sliding clamp	Mycoplasmoides pneumoniae ...	100	3.922e-18	835	1
AF-A0A3B0PQC3-F1-model_v4..	DNA polymerase III subunit beta	Mycoplasmoides gallisepticum	100	4.309e-10	455	1
AF-A0A0L0MJL4-F1-model_v4	DNA polymerase III beta subunit	Candidatus Phytoplasma phoe...	100	1.684e-9	443	1
AF-A0A292IHG5-F1-model_v4	Uncharacterized protein	Mycoplasma amphoriforme A39	100	2.808e-9	431	1
AF-X1N745-F1-model_v4	DNA_pol3_beta domain-containing protein	marine sediment metagenome	100	3.947e-9	431	1
AF-A0A800DU72-F1-model_v4	Beta sliding clamp	Campylobacterales bacterium	100	4.681e-9	406	1
AF-A0A067YNT0-F1-model_v4	DNA polymerase III beta subunit	Mycoplasmopsis synoviae	100	5.244e-9	406	1
AF-A0A6P1LEM2-F1-model_v4	DNA polymerase III subunit beta	Malacoplasma iowae 695	100	8.26e-9	386	1
AF-A0A1F5P897-F1-model_v4	Beta sliding clamp	Candidatus Doudnabacteria ba...	100	8.742e-9	405	1
AF-A0A7C5NF16-F1-model_v4	Beta sliding clamp	Epsilonproteobacteria bacterium	100	8.742e-9	391	1

1–10 of 999 rows Show 10 Previous 1 2 3 4 5 ... 100 Next

hhblits hits

Target	Description	Organism	Prob	Evalue	Score	Q. start
UniRef100_A0A257UMK6	DNA polymerase III beta sliding clamp N-te...	Acidobacteria bacterium 37-6...	100	1.4e-33	187.8	1
UniRef100_A0A0G0NMQ9	Polymerase III subunit beta protein	Berkelbacteria bacterium GW2...	100	3e-33	202.5	1
UniRef100_A0A0F9E2F3	DNA polymerase III beta sliding clamp N-te...	marine sediment metagenome	100	2e-32	198.8	1
UniRef100_A0A0F9BGN4	DNA polymerase III subunit beta (Fragment)	marine sediment metagenome	99.9	1.6e-31	207.1	1
UniRef100_A0A259F0V4	DNA polymerase III subunit beta (Fragment)	unclassified Polynucleobacter	99.9	2.9e-31	173.1	1
UniRef100_A0A090S7W4	Beta sliding clamp	Vibrio maritimus	99.9	5.6e-31	190.3	1
UniRef100_A0A0F9BVS7	DNA polymerase III beta sliding clamp cent...	marine sediment metagenome	99.9	6.5e-31	208	1
UniRef100_A0A085DTX9	Beta sliding clamp	Halomonas sp. SUBG004	99.9	7.2e-31	172.3	1
UniRef100_A0A1Q6RTD5	DNA polymerase III subunit beta (Fragment)	Firmicutes bacterium CAG:176...	99.9	1.1e-30	186.6	1
UniRef100_A0A0A0IIW3	DNA polymerase III subunit beta (Fragment)	Clostridium novyi A str. 4570	99.9	1.6e-30	200.9	1

1–10 of 999 rows Show 10 Previous 1 2 3 4 5 ... 100 Next

Figure 3.18: Results from HHblits and FoldSeek for dnaN (MPN001) as seen in Mycoboard. By default, the first 10 rows are shown, but users can navigate the tables interactively.

AI summary

	consensus	alternative	confidence	summary	interesting
TED01	DNA polymerase III beta sliding clamp	Proliferating cell nuclear antigen (PCNA) function	8.0	Consensus points strongly towards a DNA polymerase III beta sliding clamp based on several high-weight matches, including terms like 'Beta sliding clamp,' 'DNA polymerase III subunit beta', and 'DNA polymerase sliding clamp'. These all suggest the component of the DNA polymerase III holoenzyme known to act as a sliding clamp during DNA replication. While PCNA also appeared as a secondary match, it shared similarity with the sliding clamp functions in DNA replication, hence indicating potential functional overlaps that support DNA clamp activities.	Proliferating cell nuclear antigen (PCNA) related entries are potentially interesting due to shared biological roles and relevance in replication. 'MerR family transcriptional regulator' indicates a possible, albeit lower-weighted, regulatory role that could be noteworthy.
TED02	Beta sliding clamp (DNA Polymerase III subunit beta)	Proliferating cell nuclear antigen (PCNA)	9.0	The predominant weight is on ""Beta sliding clamp"" and ""DNA polymerase III subunit beta"" descriptions, indicating a function related to the sliding clamp of DNA polymerase III, crucial in DNA replication. This role is maintained by its presence in higher weights across multiple domains and fragments in the list, thus justifying the consensus. Various descriptions also mention homologous proteins like PCNA, reinforcing this functional identity and potential alternative or related activity.	Though a generally consistent pattern points to a beta sliding clamp role, the presence of 'Checkpoint protein HUS1' might imply additional roles associated with cell cycle checkpoints.
TED03	Beta Sliding Clamp (DNAP III Beta Subunit)	PCNA (Proliferating Cell Nuclear Antigen)	7.5	The query protein is most likely functioning as a beta sliding clamp due to the overwhelming prominence of descriptions related to the 'Beta Sliding Clamp,' 'DNA Polymerase III Subunit Beta,' and similar descriptions, which cumulatively have the highest weights. These terms all suggest a central role in DNA replication and repair processes, as the beta clamp facilitates polymerase processivity on DNA. The identification of homologous functions such as 'PCNA' supports the beta sliding clamp role since PCNA and bacterial beta clamps are functionally analogous, further strengthening the consensus. Despite the presence of 'Uncharacterized' and 'Hypothetical protein' descriptions, these items possess lower overall weights compared to coherent descriptions of functional relevance.	The presence of checkpoint-related proteins (e.g., 'Checkpoint protein,' 'DNA repair protein rad9') indicates a potential regulatory or protective function, such as pausing cell cycle progression in response to DNA damage or replication stress.

Figure 3.19: The summary generated by ChatGPT from the HHblits and FoldSeek hits for dnaN (MPN001).

Figure 3.20: Screenshot of the integrated alignment view. On top of the MSA, the query protein's annotations and its secondary structure from DSSP are shown.

3.10 Annotating an uncharacterized family of oxidoreductases

The workflow I developed and made available through an interactive website has enabled the discovery of a potential new family of oxidoreductases in *M. pneumoniae*, which I will mention here as a case study. The *M. pneumoniae* genome hosts a 14-member paralogous family of uncharacterized membrane lipoproteins, according to Mycowiki [83]: Mpn011, Mpn012, Mpn054, Mpn148, Mpn271, Mpn369, Mpn411, Mpn466, Mpn467, Mpn505, Mpn639, Mpn649, Mpn650, Mpn654. Of these, 8 proteins (Mpn011, Mpn012, Mpn148, Mpn271, Mpn369, Mpn466, Mpn505, and Mpn649) have "SgcJ/EcaC family oxidoreductase" as a FoldSeek hit, with probability score of at least 70%, and Mpn639 has it with a probability score of 21.4%. The SgcJ/EcaC family oxidoreductases are enzymes that play a crucial role in the biochemical pathways involving redox reactions, where they facilitate the transfer of electrons between molecules. These enzymes are part of the larger oxidoreductase family, which is responsible for catalyzing oxidation-reduction reactions often essential for various metabolic processes. In general, they have an NFT2-like fold with a characteristic beta-sheet surrounded by some alpha-helices (fig. 3.21). More specifically, the SgcJ/EcaC family is involved in the biosynthesis of specialized metabolites, such as antibiotics and other secondary metabolites, by modifying specific substrates through oxidation or reduction. This modification often alters the chemical structure and biological activity of the compounds,

Figure 3.21: The structure of MPN271.

contributing to their activity. The activity of these oxidoreductases is vital for the production of bioactive compounds with potential pharmaceutical applications, as well as for maintaining cellular redox balance and metabolic homeostasis. For example, two proteins from UniProt that are annotated in the same family are ksi from *Comamonas testosteroni* (P00947) and ecaC from *Streptomyces coelicolor* (Q8KVU1). Both seem to be steroid-delta isomerases, which are active in lipid metabolism, although the evidence for ecaC is less strong.

As a first analysis, I can simply use the relationships table of my knowledge graph (section 3.8) and filter for the IDs of these proteins, and all the information is at my fingertips. For example, Mpn012 cross-links Mpn376, another uncharacterized membrane protein. Mpn011, Mpn148, and Mpn466 are found in the same TAP complex, "Complex 4", which also includes proteins gmk, alaS, and holA (related to DNA metabolism), nrnA (related to RNA metabolism), and def (a protein synthesis factor). Almost all these proteins in this paralogous family (Mpn011, Mpn012, Mpn271, Mpn411, Mpn466, Mpn467, Mpn505, Mpn650, Mpn654) are downregulated upon cold shock perturbation. Six of them are downregulated by PrkC, a putative serine/threonine kinase.

While perhaps this is not enough to claim that we know for sure the role of these proteins in the cell, we could obtain a lot of information extremely quickly. It is interesting that our domain-based FoldSeek approach identified 8 paralogous proteins belonging to the same family. Although *M. pneumoniae* is a genome-reduced bacterium, it appears to contain many duplicated genes. This was also the case for the major dome proteins (MDPs) in chapter 2.

Another interesting case study is Mpn153. In Mycowiki [83], it is annotated as uvrD, involved in DNA synthesis/modifications/repair/conformation. In *E. coli*, uvrD is involved in the post-incision events of nucleotide excision repair and methyl-directed mismatch

repair, and probably also in repair of alkylated DNA, according to UniProt. However, our workflow identifies czcA from *Mycoplasmoides gallisepticum* as a hit for each of the 5 domains of Mpn153. The czcA protein is associated with heavy metal transport.

3.11 Discussion

Minimal cells offer a unique perspective into the fundamental mechanisms of life. By eliminating non-essential components and retaining only those necessary for cellular reproduction, the complexity of the system, though still significant, becomes more comprehensible. Initiatives such as the whole-cell modeling of JCVI-syn3A [8] demonstrate the feasibility of cataloging and characterizing the behavior of all components within a minimal cell, albeit in a simplified manner. *M. pneumoniae*, with its reduced genome, closely approximates a minimal cell while also possessing clinical significance, making it a valuable model organism. Despite extensive study, the functions of at least one-third of its genes remain unidentified. In this chapter, I described the development of a portal designed to access and visualize the available data, and a new workflow for the annotation of proteins based on TED domains, hhblits, and FoldSeek.

The workflow that submits individual domains to FoldSeek and hhblits proved useful. The main reasons why this approach works are two. First, by focusing on the domains, we can avoid insertions, disordered regions, and discontinuous sequences that would cause gaps in the sequence alignment. Having too many gaps in the sequence alignment makes it hard for sequence-based search tools to recognize hits, but querying each domain individually doesn't have this issue. Second, a protein can be seen as a group of domains with precise spatial arrangement; as such, clearly, there are many more ways to combine domains than there are domains. Thus, while a protein (*i.e.* a particular combination of domains) might be specific to a clade or even a single organism, the domains it is made of have a higher chance of existing elsewhere, and are thus easier to detect. For the dome complex proteins of chapter 2, their combination of domains, repeated twice, is only found in a handful of species, but the individual domains, as we now know, are shared with surA and prsA proteins. Of course, there are also limitations. One of the biggest ones is that, even knowing the function of most domains in the protein, we often cannot pinpoint the exact role of the protein in the cell. This is especially true for generic domains such as "DNA helicase". Another limitation is that, while the overall domain structure might be extremely conserved, even just one amino acid change in the active site is enough to prevent the usual functioning of the domain.

Ideally, this approach should be complemented by other strategies, although I believe it is not yet fully automatable. One of the strategies I envision is to incorporate JESS, a geometric hashing algorithm that can identify catalytic residues from a known template inside a protein structure [177, 178], together with a database of known active sites such as

M-CSA [179]. Then, I would be able to look for known active sites across all the protein structures in *M. pneumoniae*, potentially lending some evidence to the annotations derived by structure or sequence similarity.

A brief note on an unsuccessful approach. I obtained the proteomes from a limited selection of genomes within the Mycoplasmoidaceae family, along with four outgroup species (the same dataset used in section 2.6), and segmented all proteins in these species into their TED domains. Additionally, I constructed datasets using HHblits and FoldSeek that were restricted to these domains. Contrary to my initial hypothesis, the search outcomes were frequently inferior to those obtained using the complete UniRef30 or AlphaFoldDB databases. Given that HHblits and FoldSeek inherently perform local alignments, it is possible that segmenting both the query and target databases into domains was excessively aggressive.

The design of an effective user interface for data exploration is a critical component in bioinformatics research, as it significantly influences the accessibility and interpretability of complex datasets. Traditional methods of data presentation, such as Excel spreadsheets or static PDF documents, can be cumbersome for researchers to navigate, often requiring substantial effort to extract meaningful insights. These formats lack the intuitive engagement necessary for efficient data analysis. In contrast, an interactive interface that allows users to dynamically explore data by, for example, clicking to switch between different protein structures or alignments, can transform the experience into a more engaging and game-like process. This approach not only enhances user engagement but also facilitates a more intuitive understanding of the data, potentially attracting a broader audience to bioinformatics by lowering the barrier to entry and making data exploration more accessible and enjoyable.

4 A kinetic model for translation elongation from *in-situ* static cryo-ET data

4.1 State of the art and project overview

Biological systems often rely on molecular machines to perform useful tasks [180]. One of the fundamental processes in molecular biology is protein synthesis, which underlies all known forms of life and is critical for the proper functioning, growth, and development of cells. This process, known as translation, is carried out by the ribosome, a highly complex molecular motor [181]. Although protein synthesis is a highly regulated multistep process, arguably one of the most important and time-consuming steps is the so-called elongation cycle, where the ribosome decodes one by one the codons of the underlying mRNA molecule and adds the corresponding amino acid to the nascent peptide chain. Investigating the ribosome elongation cycle not only expands our basic understanding of molecular biology but also offers potential insights into antibiotic mode of action and development, disease mechanisms, and biotechnological applications. The kinetics of the chemical reactions or physical motions in a motor's cycle can be experimentally investigated *in vitro* or in model organisms, but it is difficult to generalize them to the *in vivo* system or to a different organism.

For decades, the individual steps of the translation elongation cycle have been studied with a variety of biochemical methods *in vitro*, mostly for the model organism *E. coli* [182, 183, 184, 185]. These experiments were complemented and enriched by X-ray crystallography and cryogenic electron microscope (cryo-EM) studies, which elucidated the structure of the ribosome's intermediate states and the binding sites of its partner molecules. More recently, time-resolved cryo-EM studies, such as Fischer et al. [186] and Dashti et al. [187], mapped the energy landscape of the process at equilibrium. These studies combine the time dimension of the biochemical experiments with the structural dimension and enable the reconstruction of the continuous trajectories of the ribosome along the free-energy landscape. Classic biochemical and cryo-EM studies share a key limitation: they have to be performed outside the living cell, sometimes in a special buffer or in the presence of drugs which stall the ribosome in a particular state or even slow down the whole translation process.

In the last few years, the "resolution revolution" experienced in cryo-EM thanks to improved sample preparation protocols and more powerful microscopes, has enabled

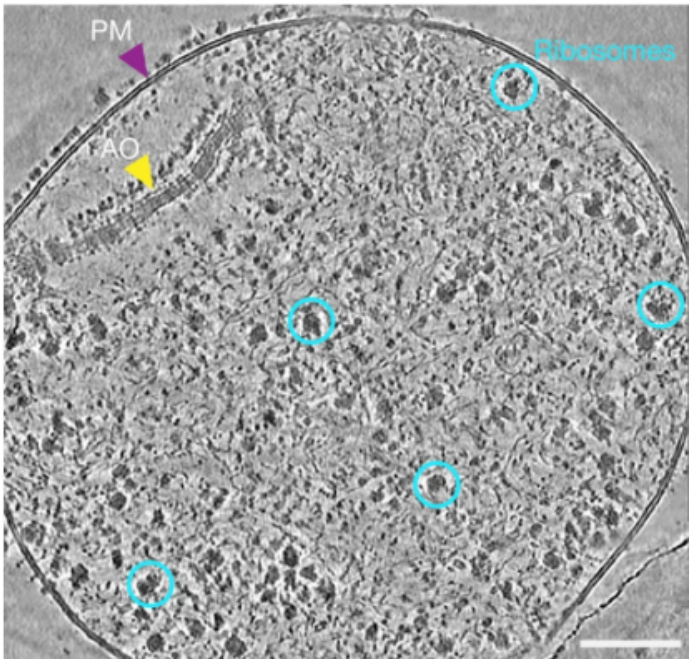


Figure 4.1: Example of a tomographic slice of a *M. pneumoniae* cell. AO: attachment organelle; PM: plasma membrane. Example ribosomes are circled. Advanced reconstruction techniques enable the classification of each ribosome in a distinct conformational class representing an intermediate in the translation-elongation cycle. Each class is identified by which tRNAs and elongation factors are bound to the ribosome, as well as the relative rotation of the small and large ribosomal subunits. Figure from Xue et al. [22]

the investigation of translation *in situ*, under near-native conditions (see section 1.2). In particular, in the model organism *Mycoplasma pneumoniae* (*M. pneumoniae*), whose small size makes thinning the samples superfluous, more than 10 distinct conformational classes have been resolved in near-atomic detail [22, 188]. The aggregation of tomograms from hundreds of cells also provided an accurate picture of the occupancy distribution of each of these states, *i.e.* the proportion of ribosomes in each class at the steady state (fig. 4.1).

In parallel to the biological experiments, numerous theoretical models of the elongation cycle have been proposed to shed light on key aspects of the process that would be otherwise inaccessible. Success stories include the calculation of competition between aa-tRNA species [189], the prediction of rates in vivo from the measured rates in vitro [190, 191], the calculation of thermodynamic properties of the cycle [192], and potential explanations for the 2-1-2 vs 2-3-2 pathways of E-site tRNA dissociation [193]. The starting point for this project was the observation that, despite the high resolution achievable with cryo-ET, so far, this method could only obtain static snapshots of translation *in situ*. Thus, we set out to develop a theoretical framework under which the dynamics of this biological process could be investigated.

Many biological processes, including translation elongation, can be understood in terms of a reaction network [194]. This is a formal framework for modeling and analyzing biochemical processes by representing them as a set of chemical reactions or conformational changes occurring among a collection of species. In these networks, nodes typically correspond to molecular species, such as proteins, nucleotides, or other biochemical entities, while the directed edges correspond to the reactions, denoting the transformation from reactants to products. There are two main objects of interest:

- The concentration of each species (as a function of time)
- The rates of the reactions

A chemical reaction network with n species can be conveniently described by a system of n differential equations, each describing the change in concentration of one species over time. A classic example is an enzymatic reaction where substrate, S, and enzyme, E, first bind to form an intermediate complex SE, then the substrate is converted to product, P, and the enzyme is recycled. In this system we have the following reactions:

1. $S + E \underset{k_2}{\overset{k_1}{\rightleftharpoons}} SE$,
2. $SE \xrightarrow{k_3} P + E$,

where k_1, k_2, and k_3 are the rate constants of the reactions. Such system is said to follow Michaelis-Menten kinetics. Letting $[X]$ denote the concentration of X at time t, the differential equations corresponding to this system are:

$$\begin{cases} \frac{\mathrm{d}[S]}{\mathrm{d}t} & = -k_1[S][E] + k_2[SE] \\ \frac{\mathrm{d}[E]}{\mathrm{d}t} & = -k_1[S][E] + (k_2 + k_3)[SE] \\ \frac{\mathrm{d}[SE]}{\mathrm{d}t} & = k_1[S][E] - (k_2 + k_3)[SE] \\ \frac{\mathrm{d}[P]}{\mathrm{d}t} & = k_3[SE]\,. \end{cases} \tag{4.1}$$

Even in apparently simple systems, the interplay between species concentrations and reaction rates gives rise to complex dynamic behavior, such as steady states, oscillations, and bifurcations [195]. In the context of translation elongation, the species are the distinct conformations of the ribosome, or its complexes with elongation factors, tRNA, and amino acids. The process can be seen as a sequence of reaction steps that include the binding of tRNA to the ribosome, peptide bond formation, and the translocation of the ribosome along the mRNA strand. Each step in this sequence constitutes a transformation that can be quantitatively modeled to elucidate the dynamics of the process (see section 4.2 for a description of reaction networks using more advanced technical tools).

In an ideal scenario, we would be able to experimentally measure as many parameters as possible to gain a comprehensive understanding of the elongation process. However, we are faced with several challenges. The first challenge is selecting an appropriate system for conducting experiments. By "system," I refer to either a defined *in vitro* solution or a model organism, such as *E. coli*. Each system, particularly living organisms, presents unique complexities and specific characteristics that complicate experimental procedures. Most methods to study the kinetics of reactions in real time, such as stopped flow [196], only work in solution. To date, the only comprehensive experimental investigations of the rates in translation elongation have been conducted *in vitro* [182, 183, 184, 185]. Replicating these experimental procedures in another system *in vivo* would be a formidable and resource-intensive task, and alas, there is often little to no incentive to replicate existing results.

Consequently, it is easier for researchers to rely on indirect measurements associated with the underlying biological processes. For example, in *E. coli*, it has been relatively straightforward to determine the average time required for a ribosome to complete a single elongation cycle by measuring the incorporation of radioactively labeled amino acids into a protein [197]. In the case of *M. pneumoniae*, an indirect measurement came to us not through biochemical methods but via cryo-ET. Thanks to the efforts from Julia Mahamid's group, we now have data on the steady-state concentrations of ribosomes in various elongation states (see fig. 4.1) [22]. The information provided by the steady-state distribution in *M. pneumoniae* alone is not enough to completely characterize the process. This is primarily because the process operates far from thermodynamic equilibrium, requiring a constant supply of energy that is ultimately provided by the hydrolysis of GTP and the formation of the peptide bonds. Consequently, understanding our *in situ* data requires putting the translation elongation cycle in the appropriate context and recognizing that the process is influenced by many external factors, including the energy state of the cell and the concentrations of elongation factors and tRNA molecules. Non-equilibrium thermodynamics is still an active field of research [198]. My model will primarily focus on a kinetic perspective, trying to elucidate the rates of the transitions.

The experimental steady-state distribution in *M. pneumoniae* cannot be directly used to calculate the rates, as there are infinitely many sets of rates that can produce the same steady-state distribution. Therefore, estimating the rates from cryo-ET data falls into the category of inverse problems [199]. (Conversely, the corresponding forward problem, which involves estimating the steady-state distribution from known rates, is totally feasible, see section 4.2.) Thus, we need to invoke additional assumptions, additional data, or both, in order to solve the problem.

In a fully uncharacterized system, where experimental data is entirely absent, reaction rates may take any value. Each conducted experiment introduces a new constraint that reduces the set of feasible reaction rates. Although these constraints are often insufficient to pinpoint a unique solution, they are instrumental in refining our understanding of the system. Such constraints allow us to narrow down the parameter space and ensure that our estimates are compatible with the current experimental data. Future experimental efforts have the potential to introduce additional constraints and could, in principle, continue to do so until reaching a singular solution. Meanwhile, generating even a preliminary estimate of the rates remains beneficial, as it can inform the design of subsequent experiments.

A prominent and ubiquitous method of addressing such inverse problems involves employing optimization techniques [200]. These techniques aim to determine an *optimal* solution from the potentially infinite set of possibilities. The primary challenge, therefore, lies in determining an appropriate optimality criterion. The central hypothesis of this project is that reaction rates in an uncharacterized system can be estimated using rates from a well-characterized "reference" system. Specifically, we define the optimal rates as those that minimize the kinetic distance metric [190] relative to the reference system. In order to honor

the constraints imposed by the experimental data, our method is based on a constrained minimization of the kinetic distance.

In this chapter, I develop a comprehensive model of the translation elongation cycle, which accounts for all the states that have been observed so far and recapitulates many of the features that ribosomes are known to possess. This process can be described as a cycle, in the sense that after going through a series of reactions, the system returns to an initial condition over and over again (once for each amino acid). Two most important properties of this system are the total elongation time, defined as the average time to return to the same state, and the steady-state distribution, defined as the proportion of time spent in each state in the limit of time going to infinity. I describe a method to estimate the kinetics of the elongation cycle in a system where direct experimental data is not available, leveraging a reference system where the kinetics have been elucidated and using only indirect data in the uncharacterized system. Such indirect data describes an aggregate property of the target system, such as the total elongation time or the steady-state distribution. I apply this strategy to estimate the rates of the transitions in the *M. pneumoniae* elongation cycle from the steady-state distribution of intermediates, using data from *E. coli*. I envision such model as a starting point for further modelling and analyses of not just the elongation cycle, but also other biological processes studied at high detail by cryo-ET.

The biggest inspiration for this project came from the PhD thesis of Sophia Rudorf, who later became my collaborator and mentor (see refs. [190, 191]). In that work, she developed a simpler model and approach to estimate the rates for *E. coli in vivo* using the rates measured *in vitro* as a reference [190, 191]. The idea is that, although we don't have complete information for the *in vivo* system, we can assume that it is reasonably similar to the *in vitro* system. After all, the *in vitro* system consists of ribosomes and other molecules purified from the living cells. Hence, we can transfer information from the known *in vitro* system to the unknown *in vivo* one. However, from measurements in the *in vivo* system, we know that the total elongation time is different from the one measured *in vitro*. To keep this difference into account, we need to slightly modify the rates. One option would be to linearly rescale them, but this is not satisfying because we have no guarantee that all rates change in the same way when going from *vitro* to *vivo*. Thus, Rudorf introduced the kinetic distance, a metric that measures the distance between two systems and can be interpreted as an approximate distance between the energy barriers of the reactions in the process [190]. By minimizing the kinetic distance of the *in vivo* rates to the *in vitro* rates, while enforcing the constraint that the elongation time *in vivo* be equal to the measured value, we achieve a non-linear update of the rates that satisfies the experimental measurements in the most parsimonious way possible. This approach is reminiscent of the Bayesian paradigm in statistics (see the book by Hoff [201]), where new data is combined with old knowledge (the prior distribution) to produce an updated state of knowledge (the posterior distribution).

Compared to the original Rudorf model, mine adds 24 new states and 14 new rates or rate constants. It is also a re-implementation completely from scratch. The addition of

the new states makes it possible to map some recently identified intermediate states onto the model, as well as to include reactions that were reported only after the original study. Specifically, my expanded model captures six translocation states rather than one, allowing me to investigate the interplay between ribosomes and EF-G. These new translocation states have been recently reported in both biochemical [185] and structural studies [22]. It also captures both 2-1-2 and 2-3-2 pathway simultaneously, which better reflects current knowledge and data [193, 22]. On the other hand, since the Rudorf model introduced important innovations and is still considered state of the art [202], I decided to keep all of its features, rather than making a simpler model tailored to our own cryo-ET data. In doing so, I kept my model implementation very modular and generalizable, so that in the future it can be extended into a framework for addressing not just translation elongation, but other biological processes as well.

All the data and code to reproduce these results are available in our EMBL internal Gitlab instance (`https://git.embl.org/grp-bork/riborates`), which also contains Systems Biology Markup Language (SBML) files that describe the models. After the submission of the thesis and associated manuscript, I plan to make everything open source. In the meantime, I am happy to provide access to the repository upon request.

4.2 Introduction to Markov processes

This section provides an overview of continuous-time Markov chains (CTMCs), focusing on their application to chemical reaction networks and in particular to the translation-elongation cycle of ribosomes. Although this material is covered in numerous textbooks [203, 204] and has already been applied to several biological processes [192, 205, 206], I will give a brief introduction to make the thesis self-contained and make sure that the methods I used are justified. I will start by describing classic ordinary differential equations (ODEs) systems and show how they are equivalent to CTMCs when viewed from a probabilistic point of view [207].

The mathematical modeling of reaction networks centers on the concentrations of the involved chemical species and how these concentrations evolve over time. The time evolution can be described through systems of ODEs. These ordinary differential equations are deterministic and capture the overall macroscopic behavior of the system by focusing on the average concentrations of the species. While ODE-based models are useful in many contexts, they also have some limitations, particularly when dealing with systems where the number of molecules of certain species is small. In such scenarios, random fluctuations at the molecular level can become significant, leading to deviations from the average behavior predicted by ODEs. Furthermore, many biochemical systems exhibit intrinsic noise, which plays a crucial role in their function and regulation. Deterministic models, by their very nature, are unable to account for this inherent stochasticity. Finally, at a more

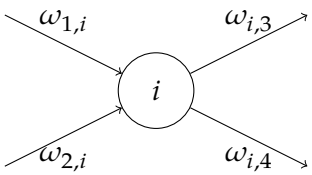


Figure 4.2: An example state i involved in four reactions, two of which produce i ($R_{1,i}$ and $R_{2,i}$) and two of which consume i ($R_{i,3}$ and $R_{i,4}$). Each arrow represents a reaction, and the arrow's label denotes the rate of the reaction.

conceptual level, it is often convenient to shift perspective and adopt the point of view of a single ribosome moving through the reaction network. At the single-ribosome level, concentrations are not relevant anymore: we view the ribosome as a finite-state machine that jumps from state to state in the reaction network. These considerations motivate the need for a probabilistic framework for reaction networks, where the focus shifts from deterministic concentrations to the probability of the system being in a specific state. In this context, a *state* is a given intermediate conformation of the ribosome (or a bundle of ribosome and other molecules such as elongation factors and tRNAs).

Following Gillespie [207] and Toral [208], we can briefly derive the probabilistic equations as follows. In the ODE setting, if a reaction R that produces i occurs at rate ω, the concentration of i in the time interval $[t, t + \mathrm{d}t)$ will increase (deterministically) by $\omega \mathrm{d}t$. In the probabilistic framework, the rate is interpreted as the probability that an individual particle will change state. Consider a system with N states labelled $1, 2, \ldots, N$ and M reactions where a reaction from state i to state j is denoted $R_{i,j}$ and occurs at rate $\omega_{i,j}$. If a reaction does not occur, we set its rate to 0. Let P_i be the probability that a particle be in state i. The probability that the particle jump from state i to state j in the time interval $[t, t + \mathrm{d}t)$ is then $\omega_{i,j}\mathrm{d}t + o(\mathrm{d}t^2)$, where $o(\mathrm{d}t)$ denotes terms that go to zero with $\mathrm{d}t$ faster than $\mathrm{d}t$ (these terms capture the cases where multiple jumps occur in the interval $\mathrm{d}t$). If we focus on a given state i produced in reactions $R_{1,i}, R_{2,i}, \ldots$ and consumed in reactions $R_{i,3}, R_{i,4}, \ldots$ (fig. 4.2), we have

$$P_i(t + \mathrm{d}t) = P_i(t)\left(1 - \sum_{j=1}^{N} \omega_{i,j}\mathrm{d}t\right) + \left(\sum_{j=1}^{N} P_j(t)\omega_{j,i}\mathrm{d}t\right) + o(\mathrm{d}t) \tag{4.2}$$

One possibility is that no reaction occurs in the interval $[t, t + \mathrm{d}t)$, which happens with probability $1 - \sum_{j=1}^{N} \omega_{i,j}\mathrm{d}t$. Then, the probability that the system be in state i equals the probability that it was already in state i. This is captured by the first term of eq. (4.2). Another possibility that influences P_i is that the system was in state j and exactly one reaction from j to i occurred. This is captured by the second term of eq. (4.2). The $o(\mathrm{d}t)$ term captures all the cases where multiple reactions occur, for example starting from i, going to j, and then back to i. These situations happen with probability smaller than $\mathrm{d}t$ as $\mathrm{d}t$ goes to

zero. Thus, if we compute the probability change $P_i(t + \mathrm{d}t) - P_i(t)$, divide by $\mathrm{d}t$, and take the limit for $\mathrm{d}t \to 0$, we obtain the differential equation

$$\frac{\mathrm{d}P}{\mathrm{d}t}_i(t) = -P_i(t) \sum_{j=1}^{N} \omega_{i,j} + \sum_{j=1}^{N} P_j \omega_{j,i} \tag{4.3}$$

The system of such differential equations for each state i, together with the constraint that $\sum_i P_i(t) = 1$, is also known as the *chemical master equation*. Given appropriate initial conditions, this equation can be solved to obtain the probabilistic evolution of the system over time.

Interestingly, this formulation of the master equation satisfies the definition of a continuous-time Markov chain (CTMC), a stochastic process where the future state of the system depends only on its current state and not on the sequence of events that preceded it. This property is called the Markov property or "memoryless" property. The processes we are concerned with are "continuous-time" because the transitions can happen at any time point, not just at discrete time intervals. They are also "discrete-valued", since there are only finitely many states. An intuitive way to see CTMCs is as follows. Suppose that a process in state i can jump to states j, k, and l with rates $\omega_{i,j}$, $\omega_{i,k}$, and $\omega_{i,k}$, respectively; when the process reaches state i, an exponential random variable with parameter $\lambda = \omega_{i,j} + \omega_{i,k} + \omega_{i,k}$ defines the time the process will spend in state i. After the time elapses, the process instantly jumps to state j with probability $\frac{\omega_{i,j}}{\lambda}$, or to state k with probability $\frac{\omega_{i,k}}{\lambda}$, or to state l with probability $\frac{\omega_{i,l}}{\lambda}$.

Switching from ODEs to CTMCs has at least two key advantages when describing biological processes. First, the stochastic nature of CTMCs allows us to understand and model cell-to-cell variability (a chemical ODE can be understood as the average approximation of a CTMC). Second, it becomes possible to reason in terms of individual particles and calculate, among other properties, the trajectory of individual particles and their first-hitting times, *i.e.* the time it takes for a particle to reach, for the first time, state j starting from state i. The average first hitting times will play an important role in my model of the elongation cycle.

A CTMC can be conveniently characterized by its infinitesimal generator matrix, usually denoted $\mathbf{Q}$, a square $N \times N$ matrix with non-diagonal entries $q_{i,j} = \omega_{i,j}$, and with diagonal entries $q_{i,i} = -\sum_{j=1}^{N} \omega_{i,j}$. In the following, I will describe four concepts which are applied in the model of the elongation cycle.

4.2.1 Steady-state distribution

For a Markov process, the steady-state distribution, also called stationary distribution and usually denoted by the vector $\boldsymbol{\pi}$, represents a probability distribution over the states of the system such that if the process reaches this distribution, it will keep it for ever. Mathematically, for a CTMC with a generator matrix $\mathbf{Q}$, the steady-state distribution $\boldsymbol{\pi}$

satisfies the condition $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$, and the elements of $\boldsymbol{\pi}$ must sum to 1, as they represent probabilities. Under certain technical conditions (ergodicity), the steady-state distribution describes the long-term probabilities of finding the system in each possible state after a sufficiently long period, regardless of the initial state of the system. In the context of the ribosome translation elongation example, the steady-state distribution would provide the probabilities of finding the ribosome in each of its conformational states under the assumption of no perturbation. Not every system has a stationary distribution, but my model of translation-elongation satisfies the properties of ergodicity, so it is guaranteed to reach a steady-state. Importantly, the stationary distribution has two interpretations. On one hand, π_i is the average proportion of time spent in state i by a single particle; on the other hand, if we have many identically distributed particles and we observe them in a snapshot at the same time, π_i is also the expected proportion of particles that we will see in state i. The stationary distribution is calculated by solving the linear system $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$ with the constraint that $\sum_i \pi_i = 1$.

4.2.2 Absorption

An absorbing state is a state in the Markov process from which there is zero probability of transitioning to any other state. Once the process enters an absorbing state, it remains there indefinitely. Absorption can be a useful concept for modeling certain types of biological phenomena, such as irreversible reactions or the formation of stable end products that cannot be converted back to their reactants within the model's scope. For the ribosome example, one could potentially define an absorbing state representing a translation elongation intermediate that is blocked by an antibiotic, such as chloramphenicol.

4.2.3 Average first-hitting time

Another important concept is the average first-hitting time. The average first-hitting time to a specific state j from a starting state i is defined as the expected amount of time it takes for the Markov process to reach state j for the first time, given that it started in state i. The average first-hitting time allows us to quantify the kinetics of reaching specific states within the system, providing insights into the typical timescales of important events in the process. In the context of the ribosome translation elongation cycle, I used the average first-hitting time to calculate the time it takes for a ribosome to complete one full elongation cycle, no matter which trajectory is taken. We denote $\tau^{x \to y}$ the average first-hitting time of state y starting from x. It is also possible to calculate the average first-hitting time for a group of states $Y = \{y_1, \dots, y_n\}$, rather than a single one; the idea is that we stop whenever we reach *any* of the states in the group. The vector of average first-hitting times to states Y starting from any state i can be computed by solving the following system of linear equations, as

proven in Norris [203]:

$$\begin{cases} \tau^{i \to Y} = 0 & \text{if } i \in Y \\ -\sum_{j=1}^{N} q_{ij} \tau^{j \to Y} = 1 & \text{if } i \notin Y\,. \end{cases} \tag{4.4}$$

A useful generalization for a chain in steady-state is $\tau^{X \to Y}$, the average first-hitting time of a group of states Y starting from *any* of the states in $X = \{x_1, \dots, x_n\}$. This can be computed as $\frac{\sum_{i=1}^{n} \pi_{x_i} \tau_i^{x \to Y}}{\sum_{i=1}^{n} \pi_{x_i}}$. For the model of fig. 4.3, the average time to complete one cycle is defined as the average first-hitting time to $\{Aco, Anr\}$ starting from $0E$ plus the average first-hitting time to $0E$ starting from $\{Aco, Anr\}$.

4.2.4 Coarse-graining

Finally, coarse-graining is a technique used to reduce the state space of a Markov process model by grouping together multiple similar states into a smaller number of effective states. This has proven crucial for modelling translation elongation: different experiments, due to different measurement resolution or focus of interest, identify different sets of intermediate states. Rather than developing a restricted model that contains the intersection of all the states from the currently available experiments, I decided to create an expanded model that contains the union of all the states identified so far. This allows me to work with a single common model onto which the states identified in the various experiments can be mapped. However, due to limitations in resolving all the intermediates, certain states become effectively indistinguishable from the point of view of a single experiment. For each experiment, I can coarse-grain the underlying expanded model to include only the states that have been observed in that experiment, lumping the states that are indistinguishable under that particular experiment. It is important to be aware that coarse-graining inevitably introduces approximations. I mitigate this by relegating the coarse-graining to the very last step, when I want to compare the theoretical predictions with the experimental data. All previous calculations are performed using the underlying expanded model.

For the computation of the coarse-grained Markov chain, I follow Baez and Courser [209] and Buchholz [210]. Briefly, given a partition Ω mapping $1, 2, \dots, N$ to $\Omega_1, \dots, \Omega_M$, with $M < N$, we first build the $N \times M$ collector matrix $\mathbf{V}$ whose entry i, j is 1 if i is mapped to j and 0 otherwise. Then, we build the $M \times N$ distributor matrix $\mathbf{W}$, whose entry i, j is α_j if j is mapped to i and 0 otherwise. The purpose of the weight w_{ij} is to represent an assumption about the contribution of state j to aggregate state i. By default, we set α_j to be the steady-state probability of state j, leading to the so-called ideal aggregate. The coarse-grained matrix $\mathbf{Q}'$ is given by

$$\mathbf{Q}' = \mathbf{WQV} \tag{4.5}$$

4.3 A comprehensive model of the elongation cycle

Our current understanding of translation in prokaryotes is reviewed by Rodnina [211], and I have tried to capture it in my model. Ribosomes have three sites to which tRNAs can bind, denoted A, P, and E. *Decoding* is the step when a loaded tRNA (in ternary complex with EF-Tu and GTP) binds to the A site and its anticodon is recognized by forming a Watson-Crick base pairing with the codon on the mRNA. *Translocation* is the step when the tRNAs shift from the A and P to the P and E sites (powered by GTP hydrolysis in cooperation with EF-G), leaving the A site free to accept a new tRNA in the next cycle. Both these steps consists of several reactions. In between decoding and translocation, the peptide bond that joins the new amino-acid to the existing chain is formed.

Thus, during the translation elongation cycle, the ribosome is known to progress through a series of intermediate states that differ either by the structural conformation of the ribosome or by the nature of the tRNAs and elongation factors bound to it. Although the interconversions among these states are thought to be continuous, certain key moments have been identified as discrete intermediate states, either due to their extended lifetimes, or because they represent a unique combination of factors bound to the ribosome at the same time, or because a key biochemical reaction takes place at those states. To date, a comprehensive *in vitro* biochemical study encompassing the entire elongation cycle has not been performed. However, some studies have focused on isolated steps, particularly decoding and translocation [182, 183, 184, 185], experimentally investigating the kinetics of the individual transition rates. Conversely, structural studies capable of examining thousands of ribosomes at high resolution have provided insights into the steady-state distribution of all intermediate states simultaneously [22, 212], but can yield little to no information about the kinetics of the process *in vivo*. Integrating the biochemical studies with the structural data is therefore a major challenge. I addressed this by developing a comprehensive model of the elongation cycle, incorporating the states identified at the single-rate level in the biochemical literature and observed in cryo-EM or cryo-ET analyses, with special regards for the recent study in *M. pneumoniae* [22] as it identified states *in vivo*. My model builds upon and extends the framework introduced by Rudorf et al. [190] and Rudorf and Lipowsky [191], henceforth referred to as the Rudorf model. Although the process of translation also includes initiation and termination steps[211], the model focuses only on the elongation cycle. All the dynamics of the reaction network will be modelled in the following sections using the Markov chain framework.

In this section, I provide a detailed description of the model, depicted in fig. 4.3, focusing on the biological point of view. Since the model is codon-specific, consider a ribosome about to decode codon i (*i.e.* a ribosome with codon i under its A site). It begins with state 0, where the P site of the ribosome is occupied by a tRNA carrying the elongating peptide chain. In this state, the ribosome is ready to accept in its A site the next aa-tRNA, a tRNA molecule loaded with the corresponding amino acid. The aa-tRNAs do not bind to the

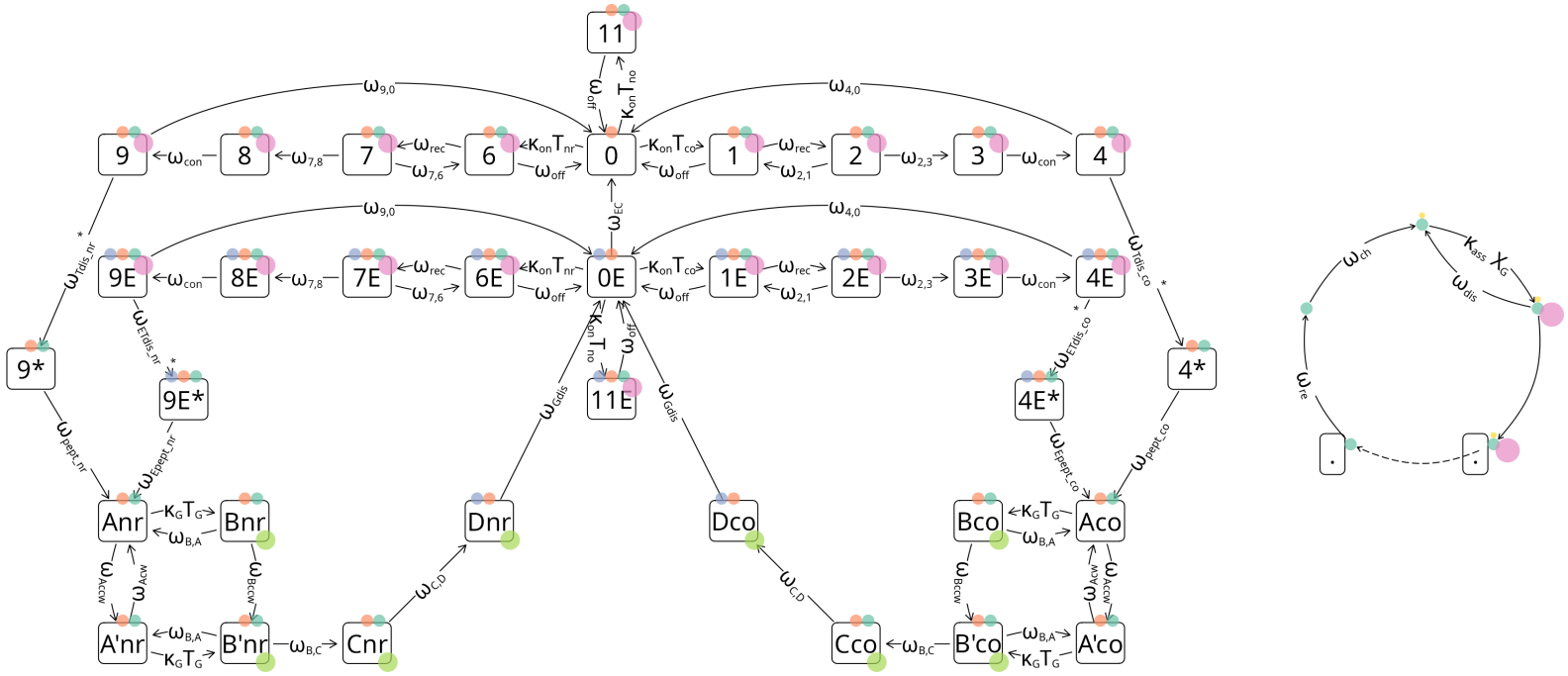


Figure 4.3: **A** Overview of the elongation cycle model used throughout this work. **B** The tRNA cycle.

ribosome alone, but as part of a ternary complex: aa-tRNA, elongation factor (EF-Tu in bacteria or EF1A in eukaryotes), and GTP. All ternary complexes compete for binding to the ribosome, but the stability of the interaction depends on Watson-Crick base pairing between the codon on the mRNA and the anticodon on the tRNA. Due to the stochastic nature of molecular motions, all ternary complexes can repeatedly bind to and unbind from the ribosome [202, 213]. I model this binding as a pseudo-first order reaction that depends only on the concentration of the aa-tRNA species. Based on experimental evidence [214, 183], the rate constant is independent of the specific codons and ternary complexes involved.

The codon-anticodon base pairing can be classified as cognate (no mismatches), near-cognate (one mismatch), or non-cognate (multiple mismatches) [190, 215]. If a non-cognate ternary complex binds, the system transitions to state 11; if a near-cognate ternary complex binds, the system moves to state 6; and if a cognate ternary complex binds, the system transitions to state 1. In the cognate branch, the ribosome and the ternary complex undergo a series of biochemical and conformational transformations: codon recognition (state 1 to state 2), GTPase activation and GTP hydrolysis (state 2 to state 3), and phosphate release along with conformational rearrangements of EF-Tu (state 3 to state 4) [216, 217]. Biochemical studies indicate that the codon recognition step is reversible; however, the subsequent transitions are not. It is important to note that irreversibility here is intended in a statistical sense, implying only that the reaction was too slow to be detected under the experimental conditions. All the rates and rate constants in this branch are taken from Rudorf et al. [190].

Despite the fact that a near- or non-cognate tRNA binds with lower energy, the energy difference alone is not enough to account for the high fidelity of protein synthesis. As such, a kinetic proofreading step is necessary [218, 219]. In line with Rudorf et al. [190], I model fidelity mechanisms both at the codon recognition level and as kinetic proofreading. Near-cognate ternary complexes are permitted to proceed past the initial binding, in a separate branch of the state space symmetrical to the cognate branch, with identical rate constants except for the two states where recognition and proofreading occur. The initial discrimination takes place in state 7 (the near-cognate equivalent of state 2): theoretically, the rate of advancing to state 8 should be lower than the corresponding rate to state 3, while the rate of reverting to state 6 should be higher than the corresponding rate in the cognate branch. This is indeed the case in *E. coli in vitro* [190]. The essence of proofreading mechanisms is to allow the wrong substrate to dissociate faster than the correct one, through energy dissipation [219]. This is modeled by letting ribosomes in state 9 (or its cognate equivalent, state 4) regress back to state 0 by releasing the ternary complex. For kinetic proofreading to function effectively, the rates of exiting states 4 and 9 must differ, and this is indeed the case *in vitro* [190].

After the aa-tRNA has been fully accommodated in the A site, the peptide bond is formed and the ribosome enters the translocation stage, whose steps have been elucidated in detail in Belardinelli et al. [185]. From this point onward in the model, the cognate and near-cognate branches are completely symmetric and share the same rates, due to lack of data

suggesting otherwise. Hence, I will describe only the cognate states, which are denoted by the "co" suffix to distinguish them from the corresponding near-cognate states, denoted by the "nr" suffix. Initially, the ribosome oscillates between a non-rotated (state Aco) and a rotated ($A'co$) state [220, 221]. In the latter, the tRNA heads in the large subunit are tilted towards the E and P sites, whereas their tails still occupy the P and A sites in the small subunit, a conformation known as "hybrid" [222]. EF-G can bind to both the rotated and the non-rotated conformations, but it promotes rotation and prevents the ribosome from reverting to the non-rotated state [185]. The binding of EF-G coincides with the transition from states $Aco/A'co$ to states $Bco/B'co$, respectively. Following Belardinelli et al. [185], I model the association of EF-G as another pseudo first order reaction that depends on the concentration of EF-G alone, whereas the dissociation happens at a fixed rate. From state $B'co$, the hydrolysis of GTP induces the transition to state Cco, and the release of orthophosphate brings the system to state Dco. At this point, translocation is essentially completed, and the tRNAs occupy sites ap/P and pe/E. Dissociation of EF-G brings the system to state $0E$, where tRNAs occupy sites P and E.

State $0E$ is a branching point. It is similar to state 0, except that the E site is still occupied by a tRNA. The precise time at which the E-site tRNA dissociates from the ribosome has been the subject of a long debate [223]. The two main possibilities are known as the 2-1-2 pathway, when the E-site tRNA leaves before the association of the next ternary complex bringing the system to state 0, and the 2-3-2 pathway, when the tRNA leaves after the association of the next ternary complex. The former scenario results in ribosomes where only the P site is occupied, whereas the latter scenario results in ribosomes with all three sites (A, P, and E) occupied by tRNAs. In the 2-3-2 pathway, the E-site tRNA leaves allosterically together with EF-Tu after states $4E$ or $9E$. The Rudorf model modelled these instances as mutually exclusive possibilities, but recent work suggests that both pathways can occur [193, 22]. The model includes both pathways simultaneously, and, for lack of data suggesting otherwise, we assume the same rates hold for the decoding step in both pathways, irrespective of whether the E-site is occupied or not.

In total, the model has 32 states and a total of 51 possible rates. As some of the rates are equal between the cognate and near-cognate branches, and between the 2-1-2 and 2-3-2 pathways, in practice only 26 rates are needed. This model allows me to capture all the states that were observed in our cryo-ET experiment and estimate the rates for an organism, *M. pneumoniae*, where they were not directly measured. Note that, while my model is depicted as a cycle, the state of the system changes from one elongation to the next, as the ribosome occupies a different position along the mRNA molecule. However, this looped representation is useful to calculate the properties of the Markov process underlying the model (see also Rudorf et al. [190]). The fact that cognate, near-cognate, non-cognate, 2-1-2, and 2-3-2 pathways are present in the same unified model also enables us to make predictions about the fidelity of the process and the relative occurrence of the 2-1-2 vs 2-3-2 pathways of E-site tRNA dissociation. Moreover, as in the Rudorf model, we distinguish

between free ternary complex concentrations and total tRNA concentrations, allowing us to model competition between different cognacy levels as well as make predictions about the codon-specific elongation times (competition between tRNAs was identified as a major factor affecting codon elongation times). The reactions that lead to the formation of the ternary complex are modelled separately (see section 4.4.1); for now, we note that what matters for mass action is not the concentration of tRNA, which is more directly measurable, but the concentration of ternary complexes. As each ternary complex species can have a different concentration, the rate of transitions $0 \to 1$, $0 \to 6$, $0 \to 11$, $0E \to 1E$, $0E \to 6E$ and $0E \to 11E$ will depend on the species of ternary complex that binds the ribosome. Therefore, the rate of one elongation cycle depends on the ternary complex concentration. Since the codon under the A site has a different set of cognate, near-cognate, and non-cognate tRNAs, codons will also have different average elongation rates and fidelities.

All the 26 rates have been measured *in vitro*, but the rates of decoding and translocation come from different experiments. Those reported for the translocation steps in Belardinelli et al. [185] are about 20 times slower than the overall translocation rate reported in other studies, from which the rates of the decoding steps were derived. This discrepancy is likely due to the fact that decoding has been investigated in the "high-fidelity" buffer [182, 184], optimized for speed and accuracy with a high concentration of Mg^{2+} ions, whereas the translocation rates were measured in a different buffer with a much lower concentration of Mg^{2+}. To avoid inconsistencies, I linearly scaled the set of translocation rates *in vitro* so that the average dwell time in the states (Aco, $A'co$, Bco, $B'co$, Cco, Dco) matched the translocation time calculated using the decoding rates in previous studies. All rates have been measured at 37 °C; thus, the estimated rates are also assumed to be valid for this temperature. The full set of rates *in vitro* is reported at the end of table A.2.

4.4 Minimization of the kinetic distance

Here, I introduce the method that allows me to estimate the reaction rates in an uncharacterized system where minimal data is available by leveraging known rates from a "reference" system. In the uncharacterized system, the available data is insufficient to uniquely identify the rates, but it can induce constraints on them, thereby specifying a space of "admissible" rates. A prime example of a reference system is the *in-vitro E. coli* ribosome, since most of the transition rates in the translation elongation have been experimentally measured, although in independent experiments. On the other hand, the *in-vivo E. coli* system is not completely characterized, as the experimental measurement of the individual rates has proven challenging. Nevertheless, the average total elongation time for one cycle has been measured, and ranges roughly between 0.045 s and 0.067 s, depending on the growth medium [224].

In this project, I aim to estimate the translation elongation reaction rates for an *in vivo* system. I use data from the target system to constrain the space of admissible rates, and,

among those, identify the set of rates that minimize the kinetic distance relative to a known reference system. Given the current state of technological advancements, this approach likely provides the most accurate estimate available until more precise methods for studying *in vivo* reaction kinetics are developed. In the original Rudorf study, the target system was *in vivo E. coli*, while the reference system was the *in vitro* solution with *E. coli* ribosomes. In my study, the target system is *in vivo M. pneumoniae*, while the reference system is the *in vivo E. coli* (I show in section 4.4.3 that using a different reference system, including *in vitro E. coli*, doesn't affect the results). Here is some intellectual acrobatics to justify this approach. In biology, the principle of minimizing the distance between two systems is a compelling choice for an extremum principle. Indeed, all living systems share a common ancestor, and biological processes are often well conserved, particularly when evolutionary divergence is recent [225]. This implies that, in the absence of specific information about a system, it is reasonable to assume its similarity to other known living systems. Consider homology modeling, which was the predominant method for predicting protein structures before the advent of AlphaFold2 [226]. This technique relies on one or more "reference" or "template" protein structures that have been solved and share sequence similarity with the protein of interest. The structure of the target protein is then estimated based on the template structure, using sequence alignment to identify equivalent amino acids. In other words, the structure information is transferred from the reference to the target protein. For the kinetic distance minimization approach, we transfer the reaction rates from the reference system to the target organism. The assumption of similarity comes from the fact that ribosomes are among the most ancient and conserved molecular machines. Indeed, ribosomal rRNAs are even used as molecular clocks and universal primers due to their extremely slow evolutionary rate [227]. Thus, minimizing the kinetic distance is akin to assuming parsimony or minimal evolution: the translation process has changed minimally, just enough to account for the differences between the two systems.

The kinetic distance was defined in Rudorf et al. [190] and is here extended to work not just with rates but also with rate constants in cases where the reactions depend on the concentration of a molecule. If ω_{ij} are the rates in the uncharacterized system and ω^*_{ij} are the corresponding rates in the reference system, the kinetic distance is

$$\mathcal{D} = \sqrt{\sum \left(\ln \frac{\omega^*_{ij}}{\omega_{ij}} \right)^2} \tag{4.6}$$

In light of the Arrhenius equation, the logarithmic differences $\ln \frac{\omega^*_{ij}}{\omega_{ij}}$ can be interpreted as an approximation of the difference between the free energy barriers of the reactions in the two systems [190]. Since this quantity will occur repeatedly, I'll introduce a new symbol for it and call it *single-barrier shift*: $\Delta_{ij} := \ln \frac{\omega^*_{ij}}{\omega_{ij}}$. In the original study, the pseudo first-order rates that depended on the concentrations of free ternary complexes were not part of the kinetic

distance, since the rate constant κ_{on} was assumed to be the same in both systems. Here, I relax this assumption and extend the kinetic distance to include reactions that depend on the concentration of reactants. In this case, the relevant terms of the kinetic distance in eq. (4.6) take the form $\left(\ln \frac{\kappa^*}{\kappa}\right)^2$.

In my model (fig. 4.3) there are 10 pseudo first-order reactions, 6 with rate constant κ_{on} and 4 with rate constant κ_G. For simplicity, I will denote with the bold $\boldsymbol{\omega}$ the vector of all rates ω_{ij} and rate constants κ_{on} and κ_G. The corresponding vector of rates and rate constants in the reference system will be denoted $\boldsymbol{\omega}^*$.

The transition rates of the Markov chain in fig. 4.3 can be conveniently collected in a so-called infinitesimal generator matrix (see section 4.2), denoted $\mathbf{Q}$. Most of the rates are zero due to the sparse structure of the reaction network; the remaining rates are of the form either ω_{ij} or $\kappa_{ij}X$, where X is the concentration of a reactant (either free ternary complexes or EF-G). Thus, $\mathbf{Q}$ is not completely equivalent to the vector $\boldsymbol{\omega}$, but it can be calculated from $\boldsymbol{\omega}$, the concentration of free ternary complexes, and the concentration of EF-G. Since the concentration of ternary complexes that are cognate, near-cognate, and non-cognate is codon-specific, the infinitesimal generator matrix is also codon specific. To account for this, I rely on abstract codon-specific models, as if all the mRNA molecules had only codons of one kind, and calculate all the properties for that codon. Then, when a property needs to be compared with actual data that is not codon-specific, the value is averaged over all codons, weighted by the usage frequency of the codon. Since each codon-specific Markov chain has a distinct infinitesimal generator matrix, I denote $\mathbf{Q}_i$ the infinitesimal generator of the chain for a ribosome with codon i under its A site. All the properties of the system can be expressed as functions of $\mathbf{Q}_i$ using standard Markov chain theory: total elongation time, first hitting times, stationary distribution, and so on.

For this study, I consider two organisms where the experimental data is adequate: *E. coli*, already analyzed by Rudorf et al. [190] and Rudorf and Lipowsky [191], where the total elongation time is available; and *M. pneumoniae*, where a coarse-grained steady-state distribution of intermediates is available from cryo-ET analysis [22, 188]. Estimating the rates in these two systems can be done by constrained minimization:

$$\begin{aligned} &\operatorname{argmin}_{\boldsymbol{\omega}} && \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\omega}^*) \\ &\text{subject to} && h(\mathbf{Q}, \boldsymbol{\theta}) \end{aligned} \tag{4.7}$$

Typically, experiments have access to only aggregated data averaged over all codons, so the constraint, or the function h here, should be a function of all the $\mathbf{Q}_i$'s and of other model parameters such as the codon usages. Here, the parameters are collectively referred to as the vector $\boldsymbol{\theta}$ (see section 4.4.2).

Notation used throughout the next sections

p_i codon usage frequency

$\tilde{P}_{i,s}$ stationary distribution for state s of a chain with codon i

$\tilde{t}_i$ elongation time of a chain with codon i

$\tau_i^{x \to Y}$ average first hitting time of states Y starting from x for a chain with codon i

$P_{i,s}$ probability of observing a ribosome in state s with codon i

$\mathbf{A}$ cognacy matrix (codons × tRNAs)

T_j free ternary complex (EF-Tu · GTP · aa-tRNA) of species j

T_i^{co} total free ternary complex cognate to codon i

$\mathcal{R}$ concentration of ribosome

$\mathcal{E}$ concentration of EF-Tu

$\mathcal{G}$ concentration of EF-G

$\mathbf{Q}_i$ infinitesimal generator of the Markov chain for codon i

$\boldsymbol{\omega}^*$ vector of known rates and rate constants in the reference system

$\boldsymbol{\omega}$ vector of rates and rate constants in the uncharacterized system (to be estimated)

Δ_{ij} single-barrier shift between the target and reference systems for reaction $i \to j$

4.4.1 The tRNA cycle

When the ribosome is in states 0 or 0E, ternary complexes (EF-Tu, GTP, and aa-tRNA) can bind to it. The state of the ribosome can be further decomposed based on the codon under the A site. Let R_i denote a ribosome with codon i under its A site. As in Rudorf et al. [190], we rely on a "cognacy matrix" $\mathbf{A}$ whose entries $a_{ij} \in \{\text{cognate}, \text{near-cognate}, \text{non-cognate}\}$ specify the relationship between codons and tRNAs. Most of the entries in the cognacy matrix are defined by the genetic code and by the rule of at most one mismatch for near-cognates, but there can be differences in affinities when multiple tRNAs carry the same amino-acid. In *E. coli*, there are 46 species of tRNA, and therefore the same number of ternary complexes. Any ternary complex j can bind to R_i, triggering the transition to either state 1, 6, or 11 (and their 2-3-2 pathway counterparts 1E, 6E, or 11E), depending on the value a_{ij} in the cognacy matrix. Let T_j be the concentration of ternary complex of species j, $X_i^{co} = \sum_j T_j \mathbb{1}_{\text{cognate}}(a_{ij})$, $X_i^{nr} = \sum_j T_j \mathbb{1}_{\text{near-cognate}}(a_{ij})$, and $X_i^{no} = \sum_j T_j \mathbb{1}_{\text{non-cognate}}(a_{ij})$, where $\mathbb{1}_k(a_{ij})$ is the indicator function for a_{ij} being equal to k. In other words, the X_i's are the total concentration of ternary complexes that are cognate, near-cognate, and non-cognate to codon i, respectively. Then, a ribosome R_i in state 0 will go to state 1 with probability $\frac{X_i^{co}}{X_i^{co}+X_i^{nr}+X_i^{no}}$, to state 6 with probability $\frac{X_i^{nr}}{X_i^{co}+X_i^{nr}+X_i^{no}}$, and to state 11 with probability $\frac{X_i^{no}}{X_i^{co}+X_i^{nr}+X_i^{no}}$.

This branching induced by the tRNA species that binds to the ribosome also explains how the model is codon-specific. It is well-known that different codons have different elongation speeds [228, 197]. In my model, the different speeds are explained by the different concentrations of cognate, near-cognate, and non-cognate ternary complexes. The rates $0 \to 1, 6, 11$ (and $0E \to 1E, 6E, 11E$) follow pseudo-first-order kinetics, *i.e.* they are proportional to the total concentration of cognate, near-cognate, and non-cognate ternary complexes, respectively. This implies that, for a given codon i, the ribosome R_i will enter the cognate branch faster when the concentration of cognate ternary complexes is higher. Conversely, if the concentration of near- or non-cognate ternary complexes is high (which in practice is almost certainly the case), the ribosome will "waste time" in the near- and non-cognate branches, potentially even incorporating the wrong amino acid in the peptide chain.

Thus, each tRNA molecule goes through different states: initially it is free, then it is loaded with an amino acid, then this aa-tRNA binds EF-Tu to form a ternary complex, then this ternary complex binds the ribosome, where the tRNA loses EF-Tu and its amino acid, and finally it is released back in the cytoplasm as a free tRNA, ready to start again. We refer to this process, which is distinct but entangled with the translation elongation cycle, as the tRNA cycle (fig. 4.3 **B**). Modelling this cycle is important because it allows us to estimate the concentration of free ternary complex of each species. The concentration of free ternary complex, in turn, determines the codon-specific transition rates in the ribosome's elongation cycle. In the rest of this section, I derive the equations to calculate the free ternary complex concentrations starting from the total tRNA concentration (which is available experimentally) and the rates of the reactions in the tRNA cycle. For context, in *M. pneumoniae*, the concentration of ribosomes is 7 μM [22], the concentration of EF-Tu is estimated to be 100 μM [137], and the concentration of individual tRNA species ranges from 0.11 μM to 14 μM [20].

Free tRNAs, as released from the ribosome's E site, are aminoacylated by the tRNA synthetase with a constant rate of ω_{re} (fig. 4.3 **B**). Following aminoacylation, they bind EF-Tu with a pseudo-first order rate constant κ_{ass}, forming ternary complexes. The ternary complexes can either get entangled with the ribosome cycle by binding to states 0 or $0E$, or EF-Tu may dissociates from these complexes with rate ω_{dis}. If the ternary complex binds to the ribosome and progresses along the elongation cycle, the EF-Tu molecule detaches from the ribosome immediately after decoding, during transitions $4 \to 4*$, $4E \to 4E*$, $9 \to 9*$, or $9E \to 9E*$ (fig. 4.3 **A**). However, the corresponding tRNA will not leave until at least two elongation cycles later, during transitions $4E* \to Aco$, $9E* \to Anr$, or $0E \to 0$, after it will have shifted to the E site. Such tRNA will not be available to form new ternary complexes as long as it is bound to the ribosome. Thus, each ribosome effectively sequesters up to three tRNAs, so that the concentration of ternary complexes is always lower than the concentration of tRNA. For both *E. coli* and *M. pneumoniae*, the rates ω_{re}, κ_{ass}, and ω_{dis} are estimated as in Rudorf and Lipowsky [191].

In practice, we need to account for the amount of tRNA of each species that is sequestered by ribosomes. Due to different codon usages, some tRNA species are sequestered more than others. Also, due to different elongation rates of different codons, translating a slow codon will sequester the tRNAs for a longer time. It is a bit messy because the tRNAs stay on the ribosome for 2-3 elongation cycles, but I will now show that the concentration of free ternary complex for each species can be derived by knowing the codon usages, the probabilities of cognate, near-cognate, and non-cognate binding, and the average elongation times for each codon. To simplify the problem, I decompose the total tRNA concentration for species j into a sum of terms, as in Rudorf and Lipowsky [191]:

$$T_j^{total} = T_j + T_j^{acc} + T_j^A + T_j^P + T_j^E + T_j^{re} + T_j^{ch}, \tag{4.8}$$

where T^{total} is the total concentration of tRNA molecules of species j and is easily accessible experimentally, T_j is the concentration of ternary complex available to bind the ribosome, T_j^{acc} is the concentration of tRNA in the process of being accommodated, T_j^A, T_j^E, and T_j^E are the concentrations of tRNA sequestered in the A, P, and E sites, respectively, T_j^{re} is the concentration of free tRNA before aminoacylation (fig. 4.3), and T^{ch} is the concentration of charged tRNA. T_j^{acc} can be further decomposed into $T_j^{acc,co} + T_j^{acc,nr} + T_j^{acc,no}$ depending on the cognacy branch where the ternary complex is accomodating. Similarly, $T_j^A = T_j^{A,co} + T_j^{A,nr}$. Assuming that both the rates of the ribosome cycle and the concentrations of ternary complexes T_j are known, it is possible to calculate the steady-state distribution and the average elongation time for the Markov chain in fig. 4.3 for each codon using standard Markov chain theory. Let $\tilde{P}_{i,s}$ be the steady-state probability for a ribosome with codon i to be in state s and $\tilde{t}_i$ be its average elongation time. We can calculate the fraction of ribosomes with codon i in each state if we know the codon usage frequencies p_i. Indeed, the probability of observing an elongating ribosome with codon i under its A site is given by

$$P_i = \frac{p_i \tilde{t}_i}{\sum_i p_i \tilde{t}_i}, \tag{4.9}$$

i.e. the relative elongation time for codon i weighted by the codon usage frequency of i. Then, the probability of observing a ribosome with codon i in state s is given by

$$P_{i,s} = P_i \tilde{P}_{i,s}. \tag{4.10}$$

These probabilities can be multiplied by the ribosome concentration $\mathcal{R}$ to obtain an estimate of the absolute concentration of ribosomes in each state for each codon. To simplify the notation, let us denote by $I_{co}(j)$, $I_{nr}(j)$, and $I_{no}(j)$ the sets of codons that are cognate, near-cognate, and non-cognate for tRNA j, respectively. Moreover, let $S_{acc,co} = \{1, 2, 3, 4, 1E, 2E, 3E, 4E\}$, $S_{acc,nr} = \{6, 7, 8, 9, 6E, 7E, 8E, 9E\}$, and $S_{acc,no} = \{11, 11E\}$ be the sets of states where accommodation occurs in the cognate, near-cognate, and non-cognate branch, let $S_{A,co} =$

$\{Aco, A'co, Bco, B'co, Cco, Dco\}$ and $S_{A,nr} = \{Anr, A'nr, Bnr, B'nr, Cnr, Dnr\}$ be the set of states where the A site is occupied by a tRNA (cognate and near-cognate branches), and let $S_E = \{0E, 1E, 2E, 3E, 4E, 5E, 6E, 7E, 8E, 9E\}$ be the states where the E-site is occupied.

We can now calculate the concentration of tRNAs in each state as follows.

$$T_j^{acc,co} = \mathcal{R} \sum_{i \in I_{co}(j)} \sum_{s \in S_{acc,co}} P_{i,s} \frac{T_j}{X_i^{co}} \tag{4.11}$$

$$T_j^{acc,nr} = \mathcal{R} \sum_{i \in I_{nr}(j)} \sum_{s \in S_{acc,nr}} P_{i,s} \frac{T_j}{X_i^{nr}} \tag{4.12}$$

$$T_j^{acc,no} = \mathcal{R} \sum_{i \in I_{no}(j)} \sum_{s \in S_{acc,no}} P_{i,s} \frac{T_j}{X_i^{no}} \tag{4.13}$$

$$T_j^{acc} = T_j^{acc,co} + T_j^{acc,nr} + T_j^{acc,no} \tag{4.14}$$

$$T_j^{A,co} = \mathcal{R} \sum_{i \in I_{co}(j)} \sum_{s \in S_{A,co}} P_{i,s} \frac{T_j}{X_i^{co}} \tag{4.15}$$

$$T_j^{A,nr} = \mathcal{R} \sum_{i \in I_{nr}(j)} \sum_{s \in S_{A,nr}} P_{i,s} \frac{T_j}{X_i^{nr}} \tag{4.16}$$

$$T_j^{A} = T_j^{A,co} + T_j^{A,nr} \tag{4.17}$$

$$T_j^{P} = \mathcal{R} \frac{T_j^A}{\sum_j T_j^A} \tag{4.18}$$

$$T_j^{E} = T_j^P \sum_i \sum_{s \in S_E} P_{i,s} \tag{4.19}$$

Note that the amounts of T_j^P depend on which codon was under the A site in the *previous* elongation cycle, and is therefore computed as a function of T_j^A.[1] Similarly, T_j^E is a function of T_j^P because the tRNAs that are now in the E site were in the P site during the previous cycle.

Calculating T_j^{re} and T_j^{ch} requires taking into account the reactions in the tRNA cycle (fig. 4.3 **B**), as well as the concentration of free ($\mathcal{E}^{fr}$) and total ($\mathcal{E}$) EF-Tu. In particular,

$$\begin{aligned}
\frac{d}{dt} T_j^{re}(t) &= T_j^E(t)\omega_{withE} - T_j^{re}(t)\omega_{re} \\
\frac{d}{dt} T_j^{ch}(t) &= T_j^{re}(t)\omega_{re} + T_j(t)\omega_{dis} - \kappa_{ass}\mathcal{E}^{fr}(t)T_j^{ch}(t)
\end{aligned}$$

where, ω_{withE} is the rate at which ribosomes lose the tRNA in their E site, calculated as the inverse of the average first hitting time of states $\{0, Aco, Anr\}$ starting from state $0E$ (denoted

[1] Every ribosome has a tRNA in its P site. The question is, how many are of species j? Because we are at steady state, the number of tRNAs of species j that is accommodated at each cycle does not change. Thus, the number of tRNAs of species j in the P site must be proportional to the number of tRNAs of species j that are accommodated in the A site.

$\tau^{0E\to\{0,Aco,Anr\}}$), weighted by codon usage:

$$\omega_{withE} = \left(\sum_i p_i \tau_i^{0E\to\{0,Aco,Anr\}} \right)^{-1}$$

At steady state, we can derive expressions for T_j^{re} and T_j^{ch}:

$$T_j^{re} = \frac{\omega_{withE}}{\omega_{re}} T_j^E \tag{4.20}$$

$$T_j^{ch} = \frac{\omega_{dis} T_j + \omega_{re} T_j^{re}}{\kappa_{ass} \mathcal{E}^{fr}} \tag{4.21}$$

$\mathcal{E}^{fr}$ can be obtained as follows. The total EF-Tu concentration ($\mathcal{E}$), which is known from experiments, can be decomposed as a sum of three terms: the free molecules ($\mathcal{E}^{fr}$), those that are part of free ternary complexes (whose concentration equals $\sum_j T_j$), and those that are part of accommodating ternary complexes ($\sum_j T_j^{acc}$). Thus, we can write

$$\mathcal{E}(t) = \mathcal{E}^{fr}(t) + \sum_j \left(T_j^{total}(t) - T_j^A(t) - T_j^P(t) - T_j^E(t) - T_j^{re}(t) - T_j^{ch}(t) \right) \tag{4.22}$$

and substitute the terms from eqs. (4.11) to (4.21) in order to calculate $\mathcal{E}^{fr}$ at steady state. We obtain a quadratic equation of form $a{\mathcal{E}^{fr}}^2 + b\mathcal{E}^{fr} + c = 0$ with coefficients

$$\begin{aligned} a &= \kappa_{ass} \\ b &= \mathcal{E} - \sum_j \left(T_j^{total} - T_j^A - T_j^P - T_j^E - T_j^{re} \right) \\ c &= -\omega_{dis} \sum_j T_j - \omega_{re} \sum_j T_j^{re} \end{aligned}$$

As eq. (4.8) must always be satisfied, we can in principle derive the concentration of free ternary complexes T_j from the terms T_j^{acc}, T_j^A, T_j^P, T_j^E, T_j^{re}, and T_j^{ch} calculated with the equations above. However, the calculation of these terms presupposes the knowledge of the transition rates of the Markov chain of fig. 4.3, hence it already presupposes the knowledge of the concentration of free ternary complexes. In other words, T_j^{acc}, T_j^A, T_j^P, T_j^E, T_j^{re}, and T_j^{ch} are all functions of T_j. This gives rise to an implicit equation that can nevertheless be solved numerically for the T_j's. Thus, eq. (4.8) should be viewed rather as an update rule that can be used to calculate new values for the free ternary complex concentrations T_j starting from some arbitrary values. In turn, with the new values of free ternary complexes, the rates of the ribosome Markov chain will need to be updated. For this reason, I use an iterative strategy that alternates the kinetic distance minimization with the calculation of new values for the free ternary complex concentrations until convergence, starting from arbitrary initial conditions (see also section 4.4.3). The model produces estimates for both

Table 4.1: Parameters that affect the model of the translation elongation cycle. See also table A.2 for the full list and associated values in *M. pneumoniae* and *E. coli*.

Symbol	Parameter
$\mathcal{R}$	Ribosome concentration
$\mathcal{E}$	EF-Tu concentration
$\mathcal{G}$	EF-G concentration
p_i	Usage frequency of each codon
T_j^{total}	Total tRNA concentrations for each species
$\mathbf{A}$	Cognacy matrix between codons and tRNA species
κ_{on}	Rate constant for ternary complex binding
κ_G	Rate constant for EF-G binding
ω_{re}	Rate of tRNA recharging
ω_{dis}	Rate of EF-Tu dissociation from ternary complexes
κ_{ass}	Rate constant of EF-Tu association to charged tRNAs

the codon-specific rates and the free ternary complex concentrations.

4.4.2 Model parameters

On top of the main reaction networks (fig. 4.3), the model depends on the intracellular concentrations of some key molecules that affect translation elongation (table 4.1). It also depends on the rates of the reactions in the tRNA cycle. These parameters can vary significantly from one organism to another and, even within the same organism, factors such as growth medium, drug treatments, temperature, and stress can influence these values. Thus, my model describes a whole family of different elongation cycles, assuming that its parameters are the only influences, and it can adapt to several organisms or growth conditions. The full set of parameters for the two main systems described in this thesis, *E. coli* and *M. pneumoniae*, are shown in table A.2.

Experiments are often performed under different conditions, identified by a specific set of parameters. I refer to such sets of parameters simply as *conditions*. In *E. coli*, experiments that measured both the average elongation rates and the other parameters were performed in four conditions [229, 224]. Each condition is characterized by a different growth medium, but, for simplicity, I label these conditions by the growth rate of *E. coli* rather than by medium composition. Specifically, the doubling rate in these four conditions is 0.7, 1.07, 1.6, and 2.5 doublings/hour, respectively. The model is applied independently for each condition. In *M. pneumoniae*, the cryo-ET experiments that measured the *in situ* steady-state ribosome intermediates were performed in three conditions: Control (cells grown in plain rich medium), Chloramphenicol (cells treated with chloramphenicol antibiotic), and Spectinomycin (cells treated with spectinomycin antibiotic). Chloramphenicol blocks the formation of the peptide bond [230, 188], spectinomycin binds to the small subunit

and inhibits the translocation step [230]. Each of these conditions is characterized by potentially different ribosome concentration, EF-Tu concentration, total tRNA concentration, and so on. Unfortunately we don't always have direct experimental measurements for all the parameters in every condition, so, I integrated the data from additional studies were the culture medium might be different [231, 19, 137]. In order to understand which parameters have an impact on the final results and to mitigate potential errors due to parameter variability, I conducted extensive sensitivity analyses (see section 4.8).

For *E. coli*, most of the parameter values were sourced from Rudorf and Lipowsky [191] and the original references cited therein, except κ_G and $\mathcal{G}$, which are from Rodnina et al. [222]. For *M. pneumoniae*, the ribosome concentration was calculated directly from the number of ribosomes per cell observed in the cryo-ET dataset provided by Xue et al. [22]. The protein copy number of elongation factors EF-Tu and EF-G were reported in a proteomics study by Maier et al. [231], and the concentrations were derived by dividing the moles of protein by the cell volume, estimated at 0.05 fL. The codon usage frequency was extracted from the protein coding genes weighted by their mRNA abundance, as measured in RNA sequencing dataYus et al. [19]. The total tRNA concentrations were recently measured by Hydro-tRNA-seq in Weber et al. [20]. The cognate relationships between codons and tRNAs were taken to respect the Mycoplasma genetic code, whereas the near-cognate matches were obtained by allowing at most one mismatch from a cognate codon, as described in Kramer and Farabaugh [232]. *M. pneumoniae* has different tRNAs molecules and even a different genetic code than *E. coli*, but the model remains nonetheless applicable to both species, provided the appropriate parameters are used. The remaining parameters κ_{on}, κ_G, ω_{re}, ω_{dis}, and κ_{ass} were assumed to be identical to the *E. coli* values; in some variations of the model, κ_{on} and κ_G were not treated as fixed parameters, but as rates to be estimated with the minimization method.

4.4.3 Model calibration

As discussed above, the model requires a set of known rates in a reference system and an experimental constraint in a target system. The rates in the target system are estimated by constrained minimization of the kinetic distance from the reference system. Ultimately, the reference rates are extracted from the extensive data from biochemical experiments for the *E. coli* elongation cycle *in vitro* [191, 185, 222].

The constraint changes from system to system and from condition to condition. Table 4.2 summarizes the six systems that I considered in this work. For *E. coli*, I considered four *in vivo* conditions, corresponding to four growth rates (0.7, 1.07, 1.6, and 2.5 doublings/hour), where the average total elongation time was measured [229, 224]. Thus, the constraint for these systems is that the total elongation time be equal to 15, 18, 22, and 22 aa/s, respectively. These are the same conditions that were used in the original work by Rudorf et al. [190].

For *M. pneumoniae*, I considered three *in vivo* systems, where coarse-grained stationary

Table 4.2: Overview of the systems for which I apply my model, and the respective constraints and reference rates used in each case.

System	Constraint	Reference rates
E. coli Growth rate 0.7 dbl/h	Total elongation time equal to 15 aa/s	*in vitro*
E. coli Growth rate 1.07 dbl/h	Total elongation time equal to 18 aa/s	*in vitro*
E. coli Growth rate 1.6 dbl/h	Total elongation time equal to 22 aa/s	*in vitro*
E. coli Growth rate 2.5 dbl/h	Total elongation time equal to 22 aa/s	*in vitro*
M. pneumoniae Control	Coarse-grained steady-state distribution as reported in Xue et al. [22]	*E. coli* 0.7 dbl/h
M. pneumoniae Chloramphenicol	Coarse-grained steady-state distribution as reported in Xue et al. [188]	*M. pneumoniae* Control
M. pneumoniae Spectinomycin	Coarse-grained steady-state distribution as reported by Dobbs 2025, unpublished data	*M. pneumoniae* Control

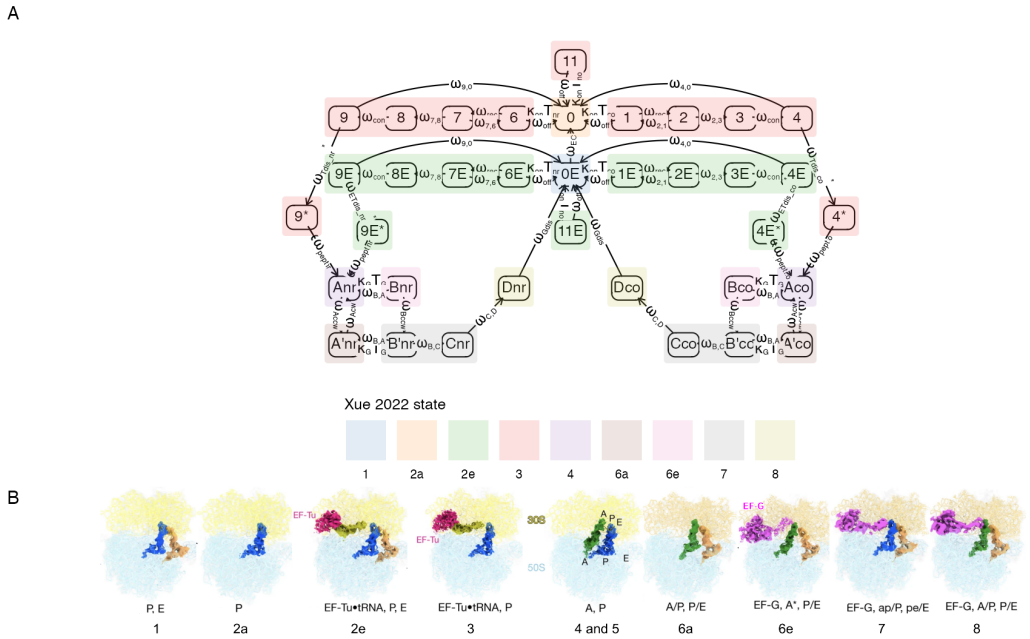


Figure 4.4: **A** Coarse-graining to the states in Xue et al. [22]. States shaded by the same color correspond to the same state. **B** Structures of the states identified in Xue et al. [22].

Table 4.3: States mapping in Xue et al. [22] and Rudorf et al. [190]. "NA" means not available, because that state was not identified in the experiment.

State	Counterpart in Rudorf et al. [190]	Counterpart in Xue et al. [22]	Counterpart in Xue et al. [188]	Description
0	0	2*a*	NA	P
1	1	3	EF-Tu· tRNA,P	P,A+EF-Tu+GTP
2	2	3	EF-Tu· tRNA,P	P,A+EF-Tu+GTP (recognized)
3	3	3	EF-Tu· tRNA,P	P,A+EF-Tu+GDP+Pi
4	4	3	EF-Tu· tRNA,P	P,A+EF-Tu+GDP
4*	5	3	A,P and a,P	P,A
0*E*	0	1	NA	E,P
1*E*	1	2*e*	EF-Tu· tRNA,P,E	E,P,A+EF-Tu+GTP
2*E*	2	2*e*	EF-Tu· tRNA,P,E	E,P,A+EF-Tu+GTP (recognized)
3*E*	3	2*e*	EF-Tu· tRNA,P,E	E,P,A+EF-Tu+GDP+Pi
4*E*	4	2*e*	EF-Tu· tRNA,P,E	E,P,A+EF-Tu+GDP
4*E**	5	2*e*	A,P,E and a,P,E	E,P,A
Aco	5	4 and 5	NA	P,A (with peptide bond)
A'co	5	6*a*	NA	P,A (rotated)
Bco	5	6*e*	NA	P,A+EF-G+GTP
B'co	5	7	NA	P,A+EF-G+GTP (rotated)
Cco	5	7	NA	P,A+EF-G+GDP+Pi
Dco	5	8	NA	E,P+EF-G+GDP
6	6	3	EF-Tu· tRNA,P	P,A+EF-Tu+GTP
7	7	3	EF-Tu· tRNA,P	P,A+EF-Tu+GTP (recognized)
8	8	3	EF-Tu· tRNA,P	P,A+EF-Tu+GDP+Pi
9	9	3	EF-Tu· tRNA,P	P,A+EF-Tu+GDP
9*	10	3	A,P and a,P	P,A
6*E*	6	2*e*	EF-Tu· tRNA,P,E	E,P,A+EF-Tu+GTP
7*E*	7	2*e*	EF-Tu· tRNA,P,E	E,P,A+EF-Tu+GTP (recognized)
8*E*	8	2*e*	EF-Tu· tRNA,P,E	E,P,A+EF-Tu+GDP+Pi
9*E*	9	2*e*	EF-Tu· tRNA,P,E	E,P,A+EF-Tu+GDP
9*E**	10	2*e*	A,P,E and a,P,E	E,P,A
Anr	10	4 and 5	NA	P,A (with peptide bond)
A'nr	10	6*a*	NA	P,A (rotated)
Bnr	10	6*e*	NA	P,A+EF-G+GTP
B'nr	10	7	NA	P,A+EF-G+GTP (rotated)
Cnr	10	7	NA	P,A+EF-G+GDP+Pi
Dnr	10	8	NA	E,P+EF-G+GDP
11	11	3	EF-Tu· tRNA,P	P,A+EF-Tu+GTP
11*E*	11	2*e*	EF-Tu· tRNA,P,E	E,P,A+EF-Tu+GTP

distributions of elongation cycle intermediates were reported: no perturbation (Control), chloramphenicol-treated, and spectinomycin-treated [22, 188, and Dobbs 2025, unpublished data]. The 10 structures identified with cryo-ET in unperturbed *M. pneumoniae* were mapped to the states in my model as shown in fig. 4.4 and table 4.3. The constraint for the kinetic distance minimization was that this coarse-grained steady-state distribution be respected. In the chloramphenicol-treated *M. pneumoniae* data set, the intermediate states that were detected are somewhat different from those in the unperturbed study, and the mapping to the states in my model is shown in table 4.3. For this system, as well as for the spectinomycin-treated system, I used the estimated rates in unperturbed *M. pneumoniae* as reference.

Rather than simplifying the model to only include the cryo-ET structures, I chose to keep the underlying model as close as possible to the states identified by biochemical experiments. This makes it easier to map new structures onto my model as they become available. However, some of the states in the model give rise to structures that are either too short-lived or structurally indistinguishable at the current cryo-ET resolutions. Therefore, some of the 10 structures were mapped to more than one state. For example, the cognate and near-cognate branches of the cycle are highly symmetrical and the conformations taken by the ribosomes are thought to differ only at the mRNA codon and tRNA species level, which cannot yet be resolved by cryo-ET. When minimizing the kinetic distance, I constrain the sum of the steady-state distribution of all the states that map to the same structure to be equal to the observed proportion of the structure. This approach is reminiscent of Hidden Markov Models: there is a "real", latent process that occurs in the cells, but the cryo-ET experiments cannot identify the latent states directly due to limitations in the resolution of the experiment; rather, they can identify some states that are indirectly related to the latent ones. In my model, I try to capture directly the latent process, making sure that it represents the experimental results when it is coarse-grained.

In general, the constraints are calculated using the theoretical tools described in section 4.2: first hitting time, stationary distribution, and coarse-graining. For the *E. coli* models, the constraint was the total elongation time of one cycle, averaged over all codons. This is obtained as follows. I start by building the codon-specific transition matrix for codon i, $\mathbf{Q}_i$. I then calculate the average first-hitting time to reach either state *Dco* or state *Dnr* starting from 0*E*, and sum this to the average first-hitting time to reach state 0*E* starting from either *Dco* or *Dnr*. This sum covers one full cycle starting and ending in state 0*E*, and this is how I define the elongation time. Importantly, the elongation time is an average over all possible paths or trajectories, weighted by the probability of following that trajectory. After calculating the elongation time for each codon independently, I then calculate the average over all codon, weighted by the frequency of the codons in the RNA pool. As such, the calculated elongation time should match what we would observe from an experiment that measures the incorporation of amino-acids over time, for example with pulse-chase radioactive labelling as in Sørensen and Pedersen [197].

In practice, most programming languages for scientific computing have routines to perform constrained minimization. I implemented the model in the Julia language [233], and I used the package NLopt for the constrained minimization [234, 235]. This package expects a function that calculates the objective function to minimize (in this case, I wrote a function that computes the kinetic distance) and a function that calculates the constraint (in this case, I wrote a function that calculates the average elongation time as described above). Both functions take the rates as input, and the constraint can be specified either as an equality or an inequality. The package then takes care of calculating the optimal rates, *i.e.* those that minimize the kinetic distance while making sure that the average elongation time is equal to the experimentally measured one. Our model is further complicated by the fact that optimizing the rates leads to a change in the free ternary complex (EF-Tu · GTP · aa-tRNA) concentrations, which means that the rates need to be updated again, and so on. Calculating the free ternary complex concentrations requires solving a non-linear system of equations, for which I use the NLsolve package [236]. The minimization-recalculation loop is carried out until convergence.

For fitting the *M. pneumoniae* models, the general idea is the same. The constraint is that the steady-state distribution matches the observations from cryo-ET. Again, I first calculate the codon-specific steady-state distribution from the transition matrix $\mathbf{Q}_i$, then average over all codons, and finally coarse-grain to reduce the states to the observable ones. At the same time, I also want to enforce high-fidelity in the translation process. Indeed, we do know that translation is a highly reliable process, making one error every 1000–10 000 amino acids [237]. From cryo-ET, we cannot distinguish between ribosomes bound to cognate and near-cognate tRNAs. Thus, I cannot estimate the fidelity directly. However, I can make the model reflect the known error rates by splitting the observed probability mass for the translocation part of the cycle unevenly between the cognate and near-cognate branches, allocating $\frac{1}{1000}$ of the steady-state mass to the near-cognate states, and the rest to the cognate states, while still maintaining the coarse-grained probabilities of each state. This additional constrain, although motivated by biological knowledge, is somewhat arbitrary, so I investigated the effect of different values of this parameter. The results are shown in fig. 4.5. The only rates that change dramatically as the fraction allocated to the near-cognate branch increases are ω_{Tdis_nr} and ω_{Tdis_co}, which are the rates of the dissociation of EF-Tu in the near-cognate and cognate branch, respectively (panel **A**). These rates influence the proofreading step of decoding [218, 219]: as the near-cognate rate gets higher and the cognate rate gets lower, ribosomes have fewer time to discriminate between cognate and near-cognate tRNAs, leading to more mistakes. Higher error rates also increases the average elongation rate (panel **B**), but at some point the cost for the cell would be too high: allocating a fraction of 0.02 to the near-cognate branch would lead to an error rate of around 5%, with potentially disastrous consequences for protein synthesis.

Once the rates have been estimated, it is also possible to calculate the fidelity of the translation elongation process in a "forward" way. To do so, I rely again on the codon-

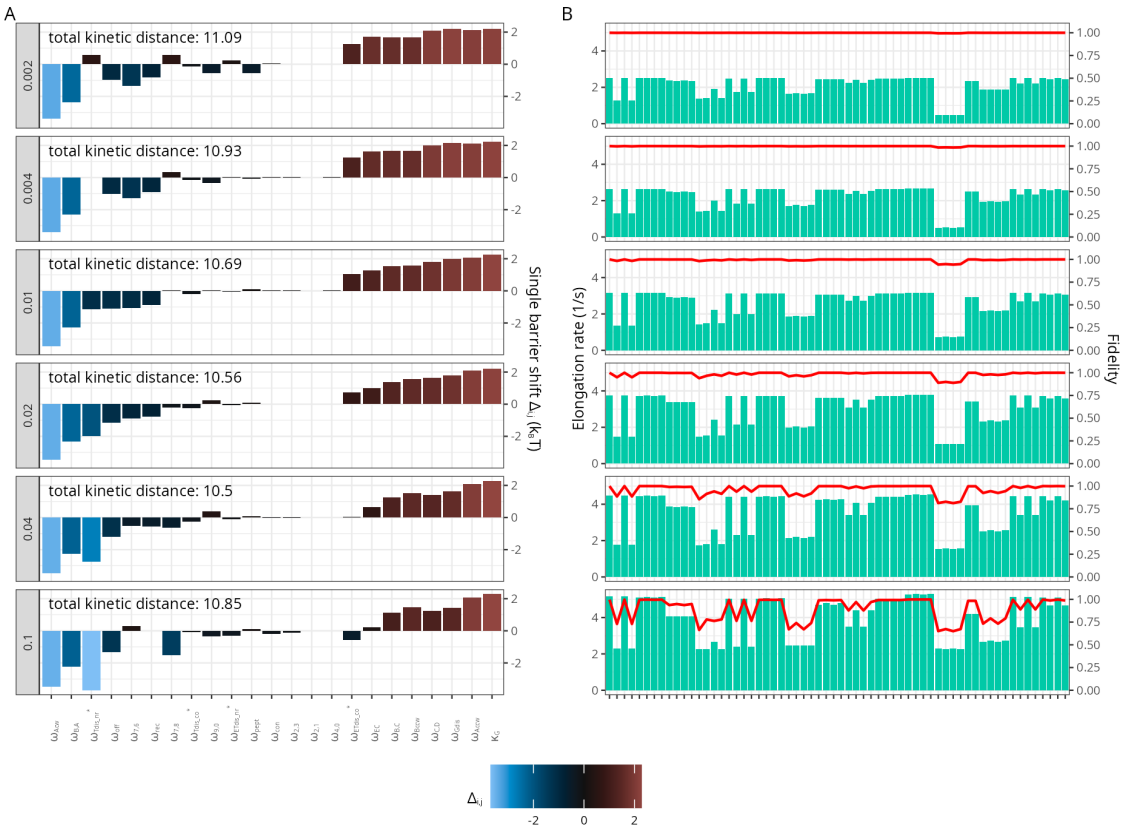


Figure 4.5: Each row corresponds to a different value for the fraction of occupancy allocated to the near-cognate branch (roughly proportional to the expected error rate), as indicated in the gray bars on the left. **A** Estimated single-barrier-shifts, representing the differences between the rates in *M. pneumoniae* and *E. coli*. Each bar corresponds to one rate. **B** Codon-specific elongation rates (green bars) and fidelities (red line). Each bar corresponds to one of the 62 non-stop codons.

specific steady-state distribution of the Markov chain. I find the total probability mass in the cognate states after the point of no-return (state Aco), denoted $\tilde{P}_{i,co}$: all the ribosomes that reach these states will inevitably incorporate the correct (cognate) amino acid.

$$\tilde{P}_{i,co} = \tilde{P}_{i,Aco} + \tilde{P}_{i,A'co} + \tilde{P}_{i,Bco} + \tilde{P}_{i,B'co} + \tilde{P}_{i,Cco} + \tilde{P}_{i,Dco}.$$

Similarly, I find the total probability mass in the near-cognate states after the point of no-return (state Anr), denoted $\tilde{P}_{i,nr}$: all the ribosomes that reach these states will inevitably incorporate the near-cognate amino acid.

$$\tilde{P}_{i,nr} = \tilde{P}_{i,Anr} + \tilde{P}_{i,A'nr} + \tilde{P}_{i,Bnr} + \tilde{P}_{i,B'nr} + \tilde{P}_{i,Cnr} + \tilde{P}_{i,Dnr}.$$

One definition of fidelity is simply $\tilde{P}_{i,co}/(\tilde{P}_{i,co} + \tilde{P}_{i,nr})$. However, sometimes a near-cognate tRNA can still carry the correct amino-acid, due to the redundancy of the genetic code. Thus, a better strategy is to decompose the steady-state probability mass of the near-cognate branch by tRNA, according to their relative concentration, and consider only the contribution of the tRNA with the wrong amino-acid. Unless otherwise noted, when I mention the fidelity I refer to this latter definition.

4.4.4 Dealing with errors: uncertainty estimation and sensitivity analyses

The rates in the reference system and the model parameters are all affected by uncertainty. The uncertainty comes from different sources of noise and errors. First, any experimental measurement is subject to errors and statistical noise. These affect any parameter that was experimentally measured, such as the rates of the *in-vitro* system and the concentrations of ribosome, tRNA, and elongation factors. Furthermore, cells grown in different conditions might have different parameter values for biological reasons. Unfortunately, the experiments in the literature are rarely so comprehensive as to encompass all the parameters of interest for the model, so our data necessarily comes from multiple sources. For example, for *M. pneumoniae*, the steady-state distribution comes from cryo-ET experiments performed in the Mahamid lab at the EMBL in Heidelberg [22], but the tRNA concentrations were measured in the Serrano lab in Barcelona [20]. These labs use different media to grow the bacteria, so there could be differences. Here, I illustrate the methods I used to estimate errors and uncertainty. The relevant results for *E. coli* and *M. pneumoniae* will be described in later sections.

To assess the impact of errors and noise in the reference rates, I generated samples of fictitious values based on the reported standard deviations and fit the model from scratch using the fictitious values. The purpose of this analysis is to evaluate the robustness of the estimates as the reference rates are changed. The original rates can be expressed as $\omega^*_{ij} \pm \delta_{ij}$, where δ_{ij} is the standard deviation for rate ω^*_{ij}. To generate the fictitious values, I considered three points for each rate: $\omega^*_{ij} - \delta_{ij}$, ω^*_{ij}, and $\omega^*_{ij} + \delta_{ij}$. Generating all possible $3^{|\omega^*|}$

combinations would be computationally intractable, so I adopted a Monte Carlo approach where I take random samples from the set of all combinations. A statistical analysis of the results obtained with these samples allows me to compute an "uncertainty interval" for my estimates.

To assess the impact of different biological conditions and of errors in the measurement of the model parameters, I conducted comprehensive sensitivity analyses. For each parameter (and for some pairs or triplets), I defined a grid of plausible values and repeated the fit from scratch for each value in the grid. This allowed me to analyze the robustness of the estimates to changes in the parameters (see section 4.8).

Whenever minimization and non-linear systems are involved, there is always the risk of having multiple solutions, possibly dependent on the initial conditions. For this reason, I fit all the models 4096 times starting from random initial values for the rates and concentrations of free ternary complexes. I keep the solution with the smallest kinetic distance, provided that it was observed at least 30 times in runs that converged without errors. In all cases I encountered so far, there was only one solution that is observed multiple times, and it was always the one of smallest kinetic distance. To check for convergence, I employ the so-called waterfall methodology: the kinetic distance is plotted as a function of the (sorted) run index, where each run starts from randomized initial conditions, and the presence of plateaus in the waterfall plot is indicative of convergence [238].

For *M. pneumoniae*, but not for *E. coli*, many of the initial conditions lead to invalid runs due to reaching a state that is physically impossible (*e.g.*, a negative concentration of a chemical species in the model). This indicates that the model, which is highly non-linear, is sensitive to the initial conditions. However, for a broad range of initial conditions, the waterfall plots show that it always converges without errors to the same value. The kinetic distance between the reference system and the target system is relatively small for *E. coli*, where the jump is between the ribosomes *in vitro* and *in vivo*. For *M. pneumoniae*, however, the jump is from one organism to another, and the kinetic distance across which we need to minimize is far greater. This is the likely cause of the sensitivity to the initial conditions.

Another source of uncertainty concerns the structure of the model itself, in terms of which states and transitions should be included. The experimental data in the literature is still inconclusive about some of the details of the elongation cycle. For example, it is not clear when the E-site tRNA leaves the ribosome. The existence of a structure with the E-site occupied but without EF-Tu in the Cm-treated dataset suggests that the E-site leaves after EF-Tu in the 2-3-2 pathway, and this is the scenario that I have used for the model. However, this is just one possibility, and it doesn't exclude that the E-site tRNA could leave early, in which case the 2-3-2 pathway would merge into the 2-1-2 pathway. I have also developed variations of the main model to assess the impact of model structure on the results. In total, I have six variations:

Simple: The original Rudorf model with 12 states (used as control and for replicating the

original study).

Expanded: An extension of the simple model where states 5 and 10 are replaced by six translocation states each, and both 2-1-2 and 2-3-2 pathways are present simultaneously (the main innovations I introduced).

Free κ_G: Same as the expanded model, but κ_G is estimated by kinetic distance minimization instead of being constant.

Free κ_{on} and κ_G: Same as the expanded model, but κ_{on} and κ_G are estimated by kinetic distance minimization instead of being constant.

With APE: Same as the "Free κ_G" model, but it contains additional states between decoding and translocation to account for structures without EF-Tu but with the E-site occupied (the most general model, used throughout this chapter unless otherwise noted).

One pept: Same as the "With APE" model, but the rate of peptide bond formation is the same for all four cases (cognate, near cognate, 2-1-2, and 2-3-2 branches).

The main conclusions do not change by using different variants, so, in the rest of the thesis, I focus on the model with APE, which is the most general.

4.5 *In-vivo* rates for *E. coli* using the new general model

Compared to the original Rudorf model [191], mine adds 24 new states and 14 new rates or rate constants. My model captures some recently identified intermediates in the translocation step [185, 22], and could help elucidate the role of EF-G in translocation as well as the relative proportions of the 2-1-2 and 2-3-2 pathways. In order to leverage these new features, I need to estimate the *in vivo* rates for the new model. Since I reimplemented all calculations from scratch with different methods, it is also necessary to see how my model compares with the Rudorf model. The comparison is possible by coarse-graining my model to the Rudorf one, as described in section 4.2. In this section, I replicate that work using my own model, explore the advantages, and add novel analyses of the relationships between the parameters.

As in Rudorf and Lipowsky [191], I used four elongation times as constraint, corresponding to four different growth conditions for *E. coli*: 0.7, 1.07, 1.6, and 2.5 doublings/hour. Here I only show the results for the 0.7 doublings/hour (see also table A.3), but the results are the same for the other conditions. When the system is coarse-grained to the Rudorf model, the rates are not significantly different from those obtained in the original study (fig. 4.6). Similarly, the estimated concentrations of free ternary complexes are very close to those obtained with the original Rudorf model. This shows that the kinetic distance minimization yields robust estimates that are not affected by the presence of additional states and rates. Furthermore, the original study was successfully replicated using my independent method, a result that is not always guaranteed in science [239].

One advantage of the present model is that it allows the investigation of the steady-state

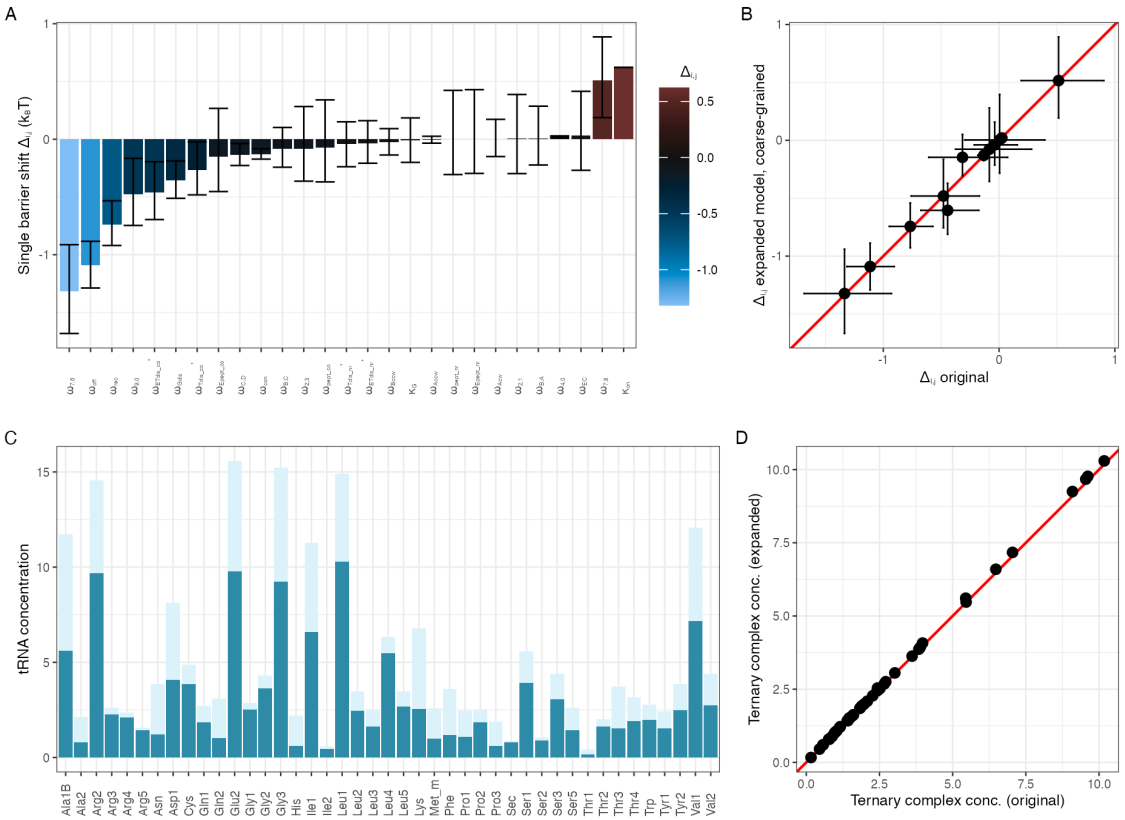


Figure 4.6: Rates and ternary complex concentrations estimated for *E. coli*. **A** Single barrier shifts compared to the *in vitro* reference elongation cycle. **B** I coarse-grained my model to the original states from Rudorf et al. [190] and compared them to the rates estimated with the original model. This figure shows the single-barrier shifts in a scatter plot, with the identity line in red. **C** Estimated ternary complex concentrations *in vivo* (dark blue) and measured total tRNA concentrations (light blue). **D** Scatter plot between the predicted ternary complex concentrations with my model and with the original Rudorf model, and identity line in red.

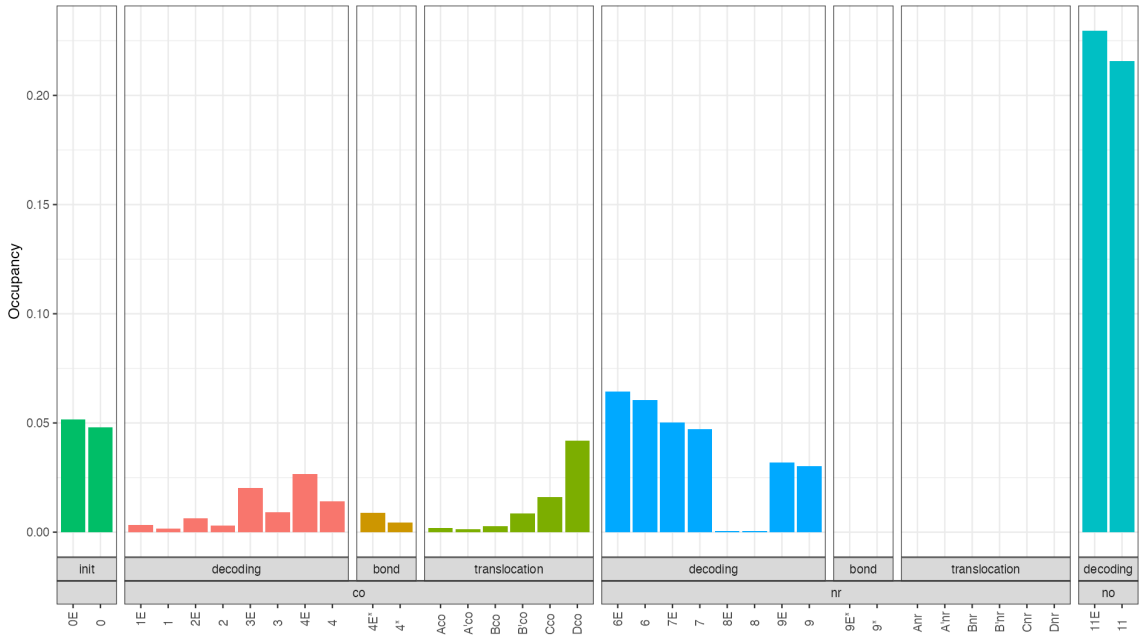


Figure 4.7: Predicted steady-state distribution (occupancies) in *E. coli*, grouped by elongation cycle phase (initiation, decoding, peptide bond formation, or translocation) and by cognacy branch (cognate [co], near-cognate [nr], non-cognate [no]).

distribution of the elongation cycle intermediates in greater detail (fig. 4.7). For example, roughly 45% of ribosomes are expected to be found in the 2-3-2 branch (decoding with the E-site occupied), 33% in the 2-1-2 branch (decoding with a free E-site), and 9% in the translocation phase. Interestingly, the most prevalent states are 11 and 11*E*, the states corresponding to the binding of a non-cognate ternary complex, which account for more than 40% of the probability mass. These results should be compared with the steady-state distribution measured by cryo-ET in *E. coli*; however, a detailed cryo-ET analysis of ribosomes is not yet available for this organism. I am aware of ongoing efforts from the group of Julia Mahamid to identify intermediate states in *E. coli*, and I am looking forward to exploring those data as soon as they become available.

To explore the relationships between the main variables in the model, I created a matrix of all the pairwise scatter plots and correlations between the variables of interest. As expected, there is a strong relationship between the codon-specific elongation times and the corresponding amount of cognate ternary complex. A high concentration of free ternary complex will boost the pseudo-first-order reactions that lead from state 0 to 1 and from 0*E* to 1*E* (cfr. fig. 4.3). Moreover, there is a strong positive correlation between fidelity and probability of following the 2-3-2 pathway, both of which are strongly negatively correlated with the elongation time (fig. 4.8).

Since the model has so many parameters, it is interesting to investigate the relationships between them. This can confirm known facts from biology or even suggest new hypotheses.

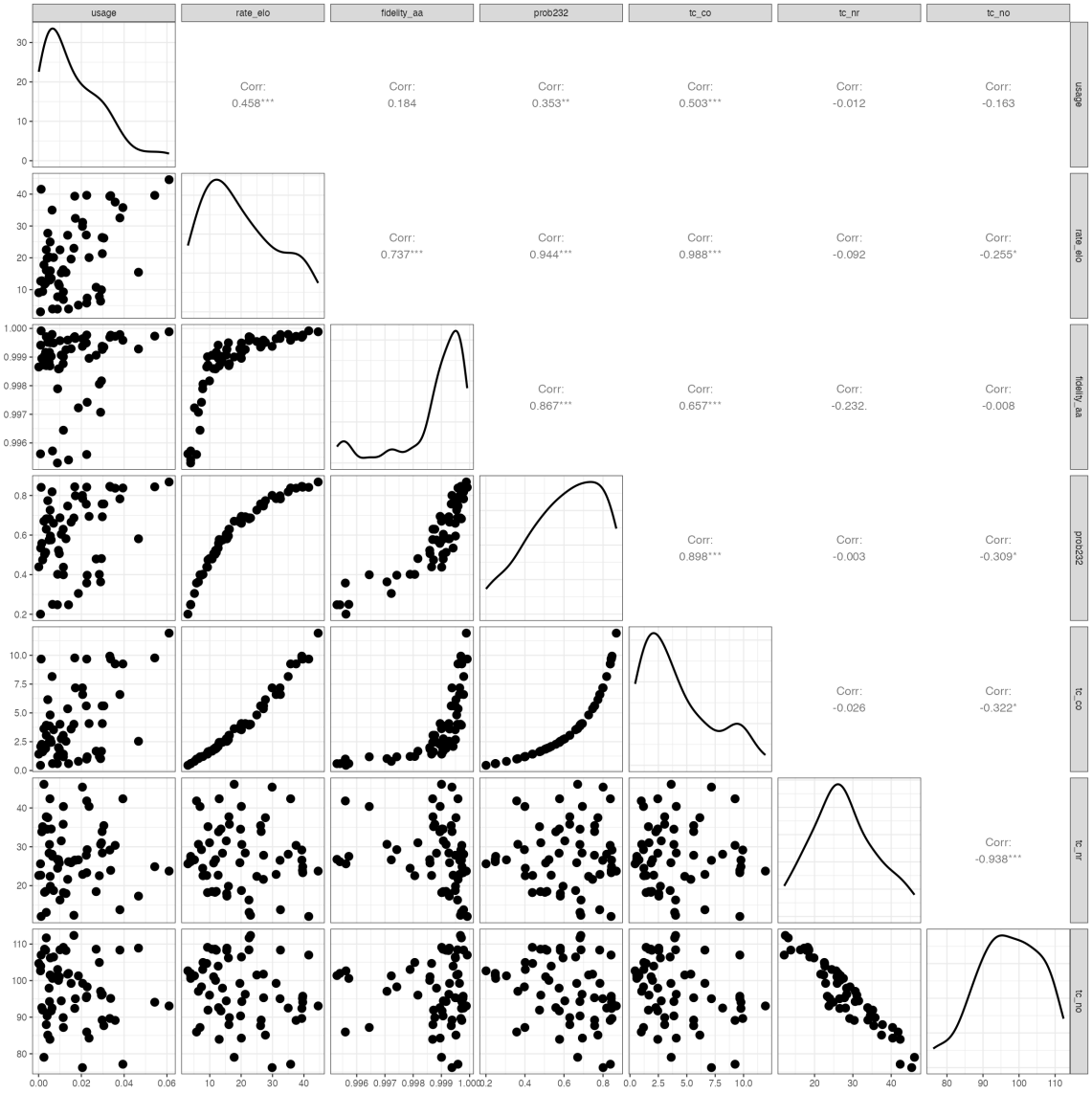


Figure 4.8: Correlations between the main variables in the model across codons. Each black dot represents one of the 61 non-stop codons in *E. coli*. In particular, the concentration of cognate ternary complex (tc_co) shows significant correlation with codon usage (indicating an evolutionary optimization), elongation rate, fidelity, and probability of following the 2-3-2 pathway.

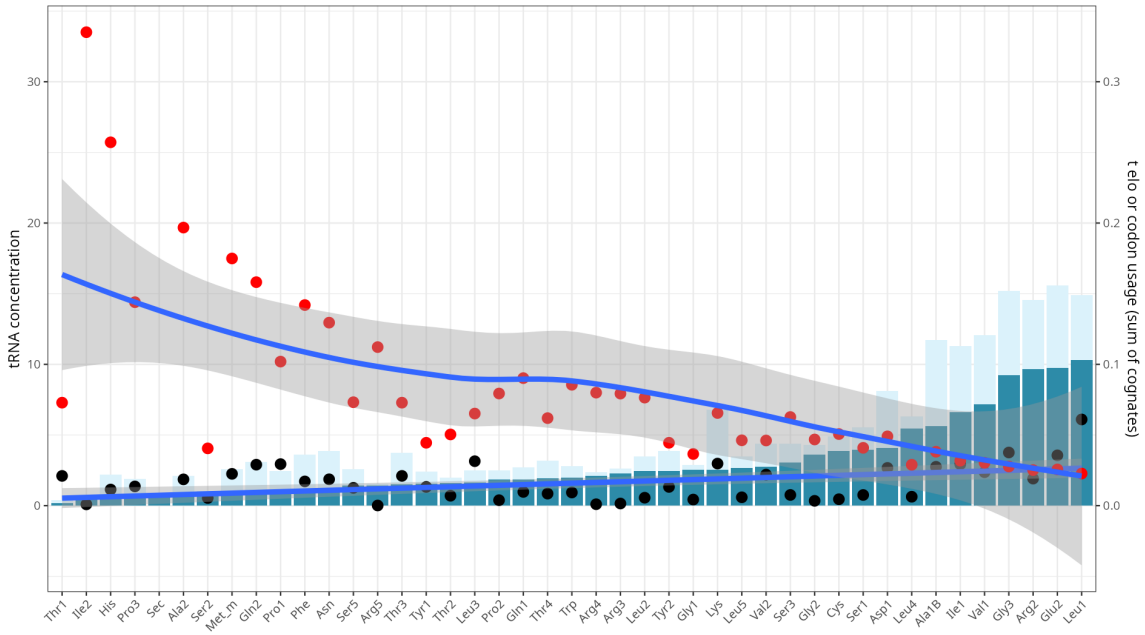


Figure 4.9: Relationship between total tRNA concentration (light blue bars), free ternary complex concentration (light blue bars), average codon usage of the cognate codons (black dots), and average elongation time of the cognate codons (red dots).

In any case, it helps to understand the inner workings of the model. For example, in fig. 4.9, I plot the relationship between total tRNA concentration, free ternary complex concentration, average codon usage, and average elongation time. Codons with higher usage frequencies also tend to have higher concentration of cognate tRNAs and, hence, faster elongation times.

One of the interesting features of the model is that it is codon-specific. Only the rates of six transitions in the whole cycle depend on the codon, but this is enough to make the overall elongation rate codon-specific. These six transitions are those coming out of states 0 and $0E$, as those rates depend on a shared rate constant (κ_{on}) and on the concentration of cognate, near-cognate, and non-cognate ternary complexes. The relationship between codons and tRNAs induced by the genetic code is quite complex, but I tried to summarize it in fig. 4.10. In general, codons whose cognate ternary complexes have a higher abundance will progress more readily in the forward direction of the elongation cycle. However, the relative usage of a codon (*i.e.* its proportion in the total mRNA pool of the cell) is also important. Indeed, all codons compete for the same pool of free ternary complexes. In the top-right corner of fig. 4.10, we can observe a correlation between the codon usage probability and the abundance of cognate ternary complexes (Pearson: 0.51, p-value=2.668×10^{-6}). This is likely an evolutionary adaptation for fast-growing organisms and has been described before [229]. Such trend is apparent across all codons, but it is also visible within the families of codons that encode for the same amino-acid. Due to the redundancy of the genetic code, some codons are synonymous for the same amino-acid. Within a family

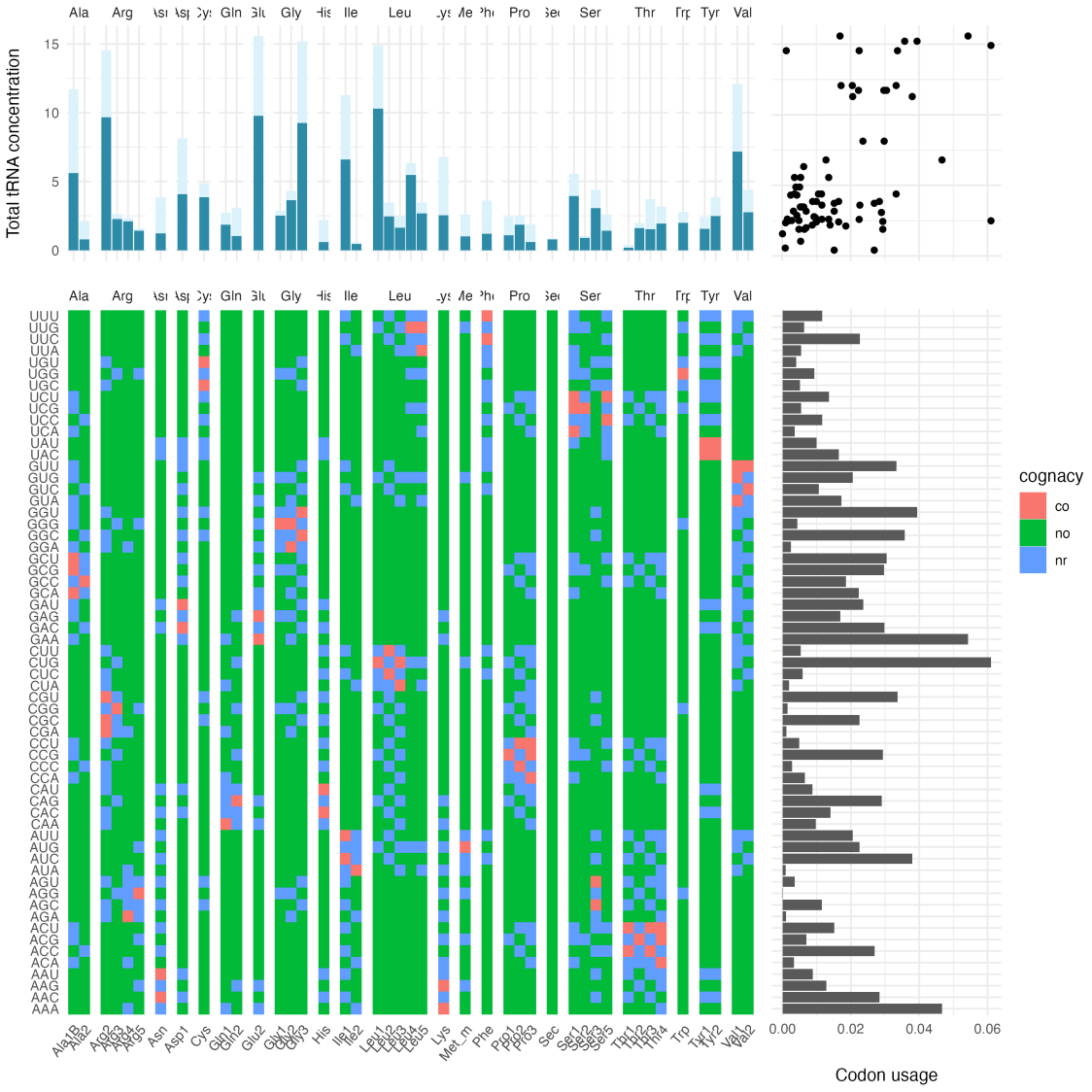


Figure 4.10: Cognacy relationships between codon and tRNAs (bottom left), codon usage (bottom right), tRNA and ternary complex concentrations (top left), and scatter plot between the usage probability and the concentration of corresponding cognate ternary complexes (top right).

Table 4.4: Summary of the parameters for the *E. coli* (0.7 dbl/h) and *M. pneumoniae* (wild type) models.

Parameter	Value in *E. coli*	Value in *M. pneumoniae*
N. tRNAs	43	35
Total tRNA concentration	209.4 μM [191]	110 μM [137]
Ribosome concentration	18.57 μM [191]	7 μM [22]
EF-Tu concentration	152.2 μM [191]	100 μM [137]
EF-G concentration	10 μM [185]	10 μM [137]

of synonymous codons, each codon has a different usage frequency (sometimes called codon usage bias). One of the forces that stimulates the codon usage bias is translational efficiency [240]. My model can then be used to investigate this phenomenon by making predictions on the elongation rates. Furthermore, the model makes it is possible to estimate the overall translation speed of a gene by summing the elongation rate of all the codons in its sequence.

4.6 Estimation of the rates in *M. pneumoniae*

Having estimated the rates in *E. coli*, I then set out to do the same in *M. pneumoniae*. Recently, the collection of a large cryo-ET dataset enabled the resolution of the structure of 10 intermediate states of the elongation cycle *in situ* [22]. Such dataset enables us to measure the relative proportion of each intermediate by taking a snapshot of a large population of ribosomes at a fixed time. Because the cells were growing undisturbed for a few days before being imaged in the electron microscope, we can assume that the relative proportions observed are representative of the steady-state distribution. We can thus use this data as the new constraint in the minimization of the kinetic distance method. This time, rather than using the rates of translation *in vitro*, which are not available for *M. pneumoniae*, I decided to use the estimated rates in *E. coli* growing at 0.7 dbl/h as a reference. I had the option of choosing either the *E. coli in vitro* system or one of the estimated *in vivo* systems, but I reasoned that the rates in *M. pneumoniae* would probably be closer to the *E. coli* rates *in vivo* than those *in vitro*, so I opted to use the estimated rates of the slowest-growing *E. coli* system (0.7 doublings/hour) as the reference system for unperturbed *M. pneumoniae*. After doing all the fitting using the 0.7 dbl/h as the reference, I investigated the effect of using different reference systems, and I found that the estimated rates in *M. pneumoniae* essentially do not change (fig. 4.11). In fact, surprisingly, the smallest kinetic distance is achieved when the fastest-growing *E. coli*, at 2.5 doublings/hour, is used as the reference.

Nevertheless, when using the *E. coli* rates to estimate those in *M. pneumoniae*, it is important to account for the differences between these two organisms. *M. pneumoniae* has a different set of tRNAs, and even a different genetic code, where one of the classic stop codons

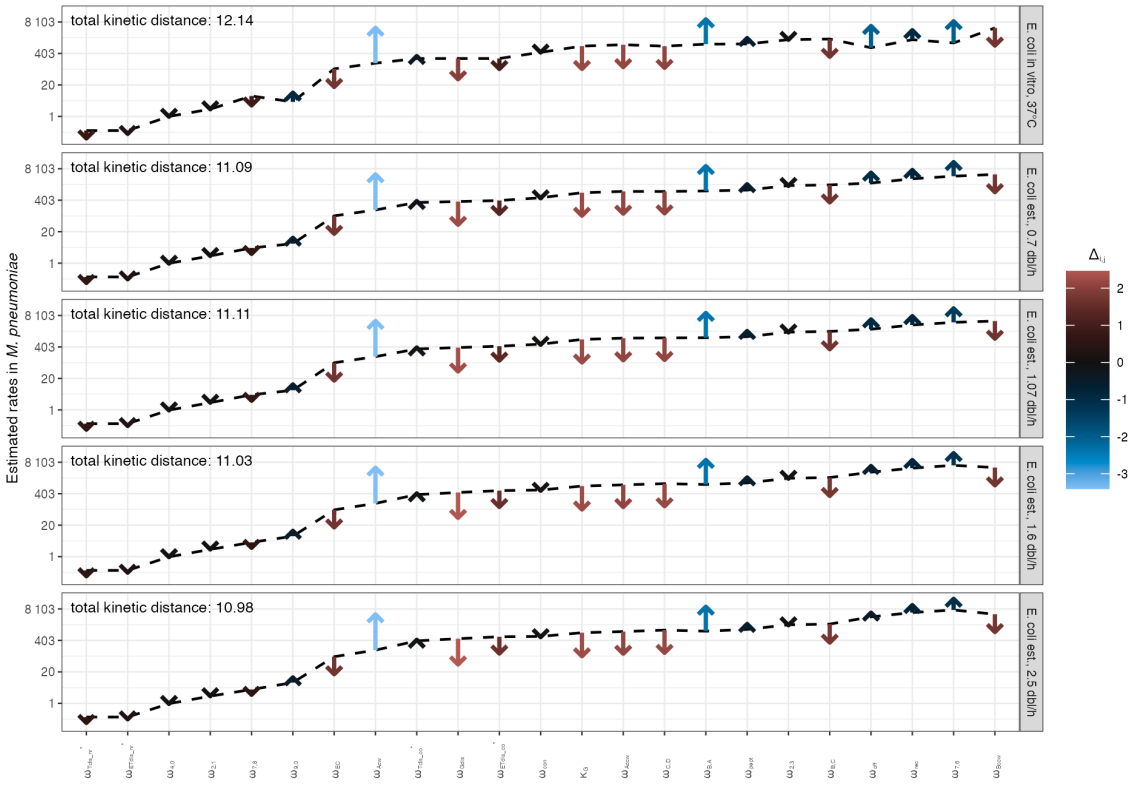


Figure 4.11: Estimated rates in *M. pneumoniae* corresponding to five different reference systems. The black dashed line crosses the reference rates, while the arrows point to the estimated *M. pneumoniae* rates. Since the y-axis is in log scale, the length of the arrows is proportional to the single barrier shifts $\Delta_{i,j}$, which is also encoded in the arrow's color.

Table 4.5: Comparison between the experimentally measured low resolution steady-state distribution in *E. coli*, the predicted steady-state distribution in *E. coli* from my model, and the measured steady-state distribution in unperturbed *M. pneumoniae*.

State description	*E. coli* K-12 30min LB [241]	*E. coli* 2.5 dbl/h [191]	*M. pneumoniae* wildtype [22]
P	14%	5.2%	8.5%
E,P,A+EFTu	38%	46%	6.3%
E,P+EFG	48%	8.2%	11%

(UGA) actually codes for tryptophan. While *E. coli* has 43 distinct tRNAs, *M. pneumoniae* has only 35. Since the cognacy relationships, to the best of my knowledge, have not been accurately determined in mycoplasma, I built the cognacy matrix relying only on the genetic code, assigning to each codon all the tRNAs that carry the corresponding amino acid. The concentration of ribosomes was estimated directly from the count of ribosomes in the tomograms [22], whereas the concentrations of EF-Tu and EF-G were obtained by mass spectrometry as reported in [137]. table 4.4 shows the main parameters used for the systems described here: the unperturbed *M. pneumoniae* and the reference *E. coli* growing at 0.7 dbl/h. Except for the concentration of EF-G, all values are lower in *M. pneumoniae*, consistently with the extremely slow growth rate of this organism. These parameters are sometimes known only as rough approximations; moreover, they heavily depend on the growth conditions of the culture, and in the literature it is difficult to find two studies that use exactly the same media. Other parameters (κ_{on}, κ_G, ω_{re}, ω_{dis}, and κ_{ass}) are completely unknown for *M. pneumoniae*, and I estimate them to be the same as *E. coli*. I try to mitigate these issues by repeating the analysis for different parameter values and assess the robustness of the predictions (section 4.8), as prescribed in Villaverde et al. [238].

It is also important to investigate differences in the steady-state distribution of elongation intermediates between these two organisms. So far, the steady-state distribution has not yet been measured in *E. coli* at the same level of completeness that was achieved in *M. pneumoniae*. There is only one report in the literature for a distribution of intermediates at very low resolution for two *E. coli* strains [241]. It could still be interesting to compare the steady state distribution measured in that study (Khusainov et al. [241]) with the one measured in *M. pneumoniae* (Xue et al. [22]) and with the one predicted in *E. coli* by my model. Some general limitations do remain. First and foremost, the Khusainov et al. [241] study identified only three intermediates, compared to 10 in *M. pneumoniae*. Second, the distributions for the two strains of *E. coli* grown under the same conditions, are already qualitatively quite different, with the P class going from 14% in the K-12 strain to 1% in the ED1a strain. This suggests that steady-state distributions might be extremely variable even across strains of the same species, so it wouldn't be surprising if there were major differences between *E. coli* and *M. pneumoniae*. Third, the growth medium used in Khusainov

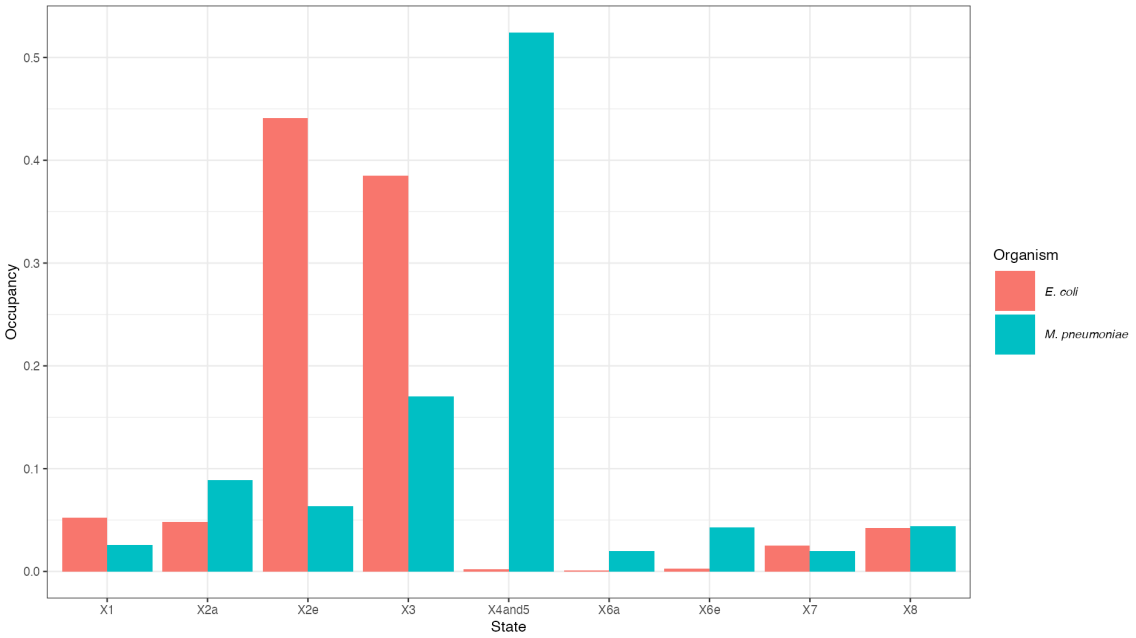


Figure 4.12: Comparison between steady-state distribution of intermediates between *E. coli* (estimated *in vivo* with my model) and *M. pneumoniae* (measured by cryo-ET [22]). The *E. coli* model was constrained to include only the states reported in *M. pneumoniae*.

et al. [241], LB, is different from the media used to measure the elongation rates in the studies used to fit my model in *E. coli*, which could lead to differences in the steady-state distribution of ribosome intermediates. The comparison between these systems, including only the three states that are common to all of them, is shown in table 4.5. For my predictions in *E. coli* and the observations in *M. pneumoniae*, the proportions are calculated only from the states that could be mapped to the corresponding state in Khusainov et al. [241], and therefore do not add up to 100%.

If we exclude Khusainov et al. [241] and only compare my model's prediction in *E. coli* with the experimental results in *M. pneumoniae*, we can conduct a more extensive and fair comparison, since all the states in the model can be mapped to the experimentally resolved structures. Figure 4.12 shows the abundances of the intermediates in *M. pneumoniae*, along with the corresponding abundances for *E. coli* when the model is coarse-grained to the same structures. In *E. coli*, the prevalent states are those in the decoding phase of the elongation cycle, *i.e* those when the ribosome is bound to the ternary complex. In *M. pneumoniae*, on the other hand, the most abundant states are those just after decoding and before the binding of EF-G.

Now, given that the steady-state profile is probably different between these two organisms, I asked what would my model predict about the rates. The lifestyle and growth rate of these bacteria are quite different, but the differences at the molecular level need not be large. Furthermore, I was interested in whether the differences, if any, are driven by parameters

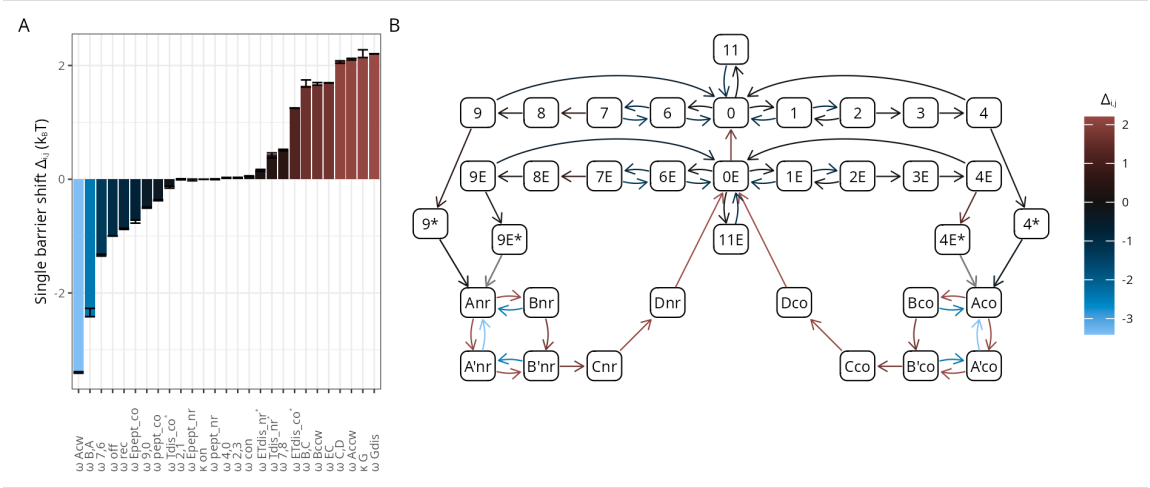


Figure 4.13: **A** Single-barrier shifts representing the logarithmic difference between the rates in *M. pneumoniae* and *E. coli*. Negative values denote reactions that are faster in *M. pneumoniae*; positive values denote reactions that are faster in *E. coli*. **B** Single-barrier shifts mapped onto the elongation cycle model.

like concentration of elongation factors, or rather intrinsic properties of the ribosomes such as the structure, or perhaps something that we haven't parametrized in our model like the ion concentrations.

After applying the kinetic distance minimization using the observed steady state distribution in *M. pneumoniae* as constraint, I obtained the estimate shown in fig. 4.13 for the rates (see also table A.3). The rates of the decoding step (top half of fig. 4.13 **B**) do not substantially differ from those in *E. coli*. However, translocation (bottom half of fig. 4.13 **B**) is significantly slowed down. Specifically, the *M. pneumoniae* ribosomes have a tendency to remain in the non-rotated pre-translocation state. For example, the rate of the counter-clockwise rotation from state *A* to state *Aco* is $900\,s^{-1}$ in *E. coli*, but only $100\,s^{-1}$ in *M. pneumoniae*. Conversely, the "reverse" reaction from *Aco* to *A* shifts from $200\,s^{-1}$ in *E. coli* to $8000\,s^{-1}$ in *M. pneumoniae*. The binding of EF-G shows a similar pattern, with the rate constant of binding being 10 times slower, and the rate of unbinding being 50 times faster in *M. pneumoniae* compared to *E. coli*. It is important to note that the model may not have enough resolution to uniquely identify both the forward and reverse rate of a reversible reaction. The forward and reverse rate could be scaled differently and still yield a similar same overall steady-state distribution. This is not just a problem within reversible reactions, but it affects every reaction. There simply are not enough constraints to uniquely identify the rates, and this is precisely why we have to use a minimization method. The value of the

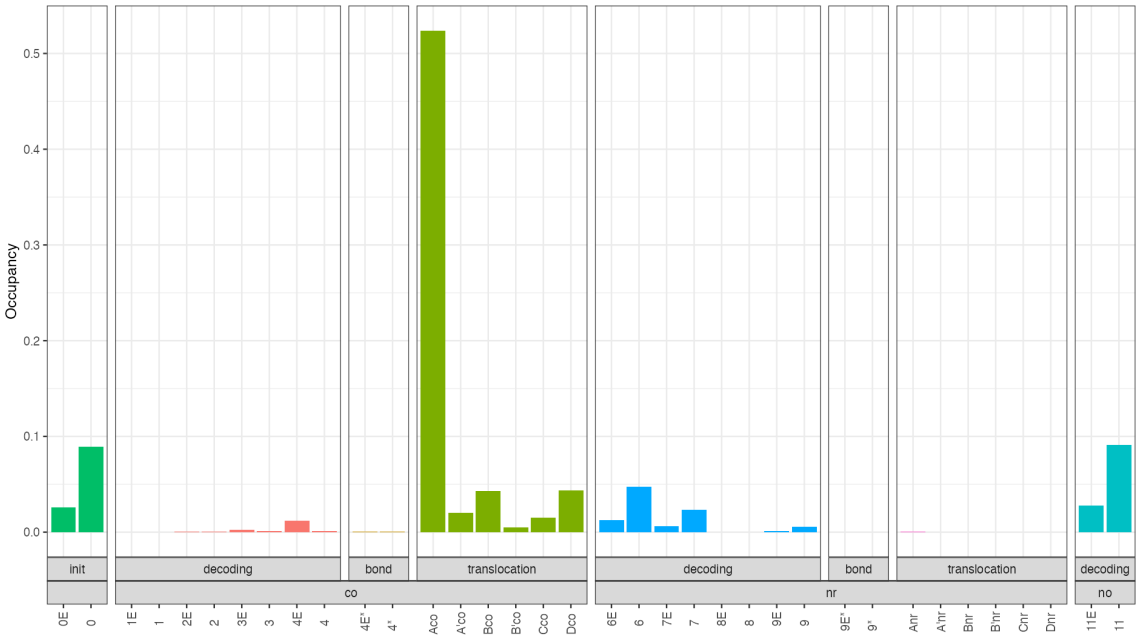


Figure 4.14: Predicted steady-state abundances (occupancies) in *M. pneumoniae*, grouped by elongation cycle phase (initiation, decoding, peptide bond formation, or translocation) and by cognacy branch (cognate [co], near-cognate [nr], non-cognate [no]).

minimization method is that it finds the most parsimonious set of rate changes that can explain the data.

The steady-state distribution of intermediate states is dominated by state *Aco* (fig. 4.14, to be compared with fig. 4.7 on page 124). The decrease of the number of ribosomes in states 11 and 11*E* compared to *E. coli* makes sense in light of the lower concentration of non-cognate tRNAs in *M. pneumoniae*. Another noticeable difference is the proportion of ribosomes in the 2-3-2 branch (states whose name ends in "E"), 20% in *M. pneumoniae*, whereas it was 50% in *E. coli*.

An important question is whether the slower translocation rates are driven by intrinsic structural properties of the ribosomes or by the concentration of EF-G. Indeed, EF-G binds to the ribosome exactly at the beginning of translocation. Here, the binding of EF-G is modelled as a pseudo-first-order reaction, *i.e.* the rate is given by the product between a rate constant κ_G and the concentration of EF-G. The model predicts a κ_G of ~800 s^{-1} μM^{-1} in *E. coli*, and of ~100 s^{-1} μM^{-1} in *M. pneumoniae* (table A.3). The concentration of EF-G is taken to be 10 μM for both organisms (table 4.4). Thus, when we multiply the respective κ_G and concentration, we get that the rates are ~8000 s^{-1} in *E. coli* an ~1000 s^{-1} in *M. pneumoniae*, concluding that the slowdown effect cannot be explained by EF-G concentration alone. In section 4.8 I test the effect of possible mistakes in the parameter estimates by running a sensitivity analysis.

As I did for *E. coli*, I also investigated the relationships between some important model

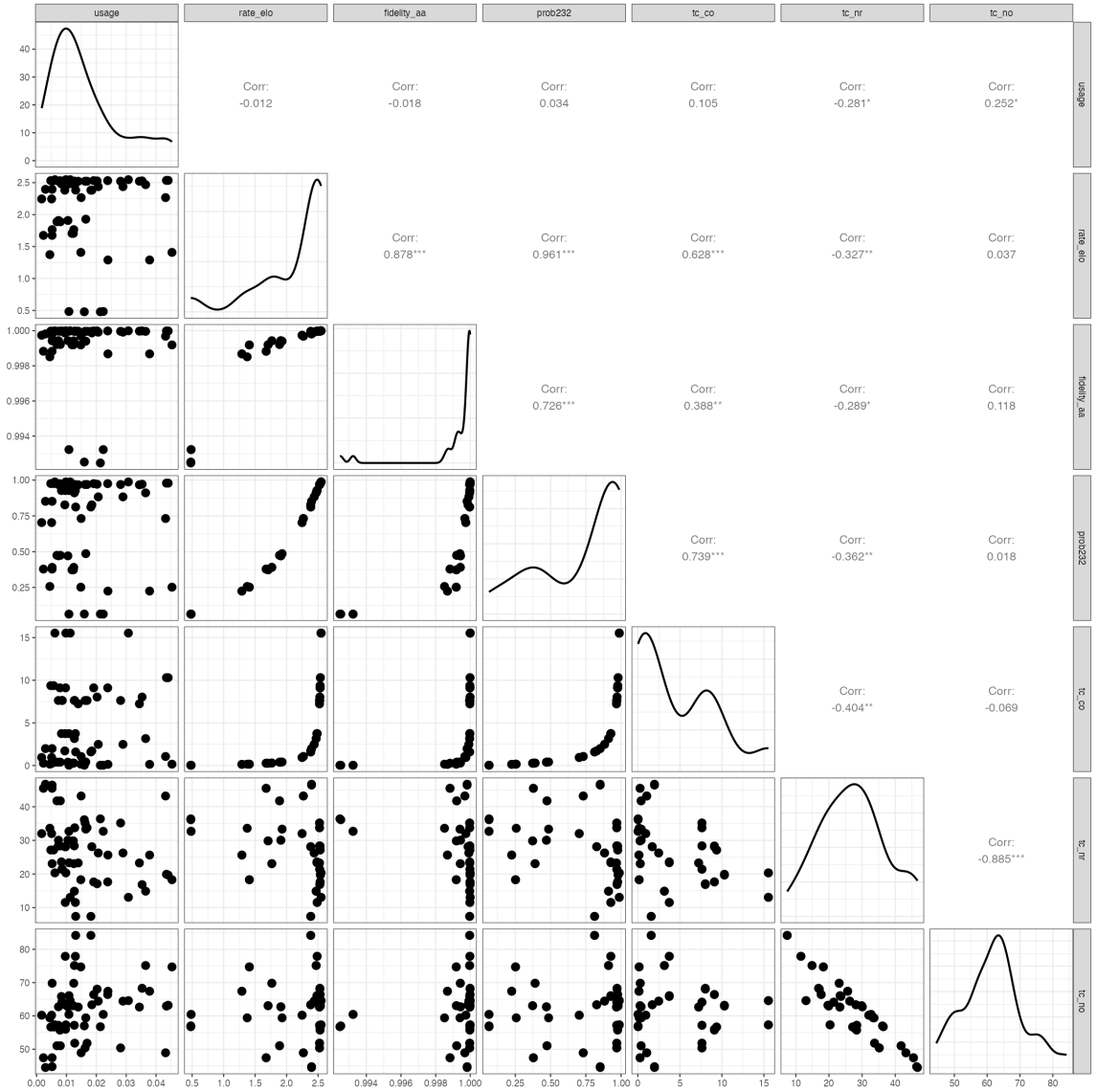


Figure 4.15: Correlations between the main variables in the model across codons. Each black dot represents one of the 62 non-stop codons in *M. pneumoniae*. In particular, the concentration of cognate ternary complex (tc_co) shows significant correlation with elongation rate, fidelity, and probability of following the 2-3-2 pathway, but not codon usage.

quantities in *M. pneumoniae*. Figure 4.15 shows the codon-specific usage bias, elongation rate, fidelity, probability of going through 2-3-2 pathway, and concentration of cognate, near-cognate, and non-cognate ternary complexes. We can observe the same positive relationship between elongation rate, fidelity, and concentration of cognate ternary complexes. However, the correlation between codon usage and cognate ternary complexes is much lower (0.105 compared to 0.503 in *E. coli*). This suggests that the ability of ribosomes to find the correct ternary complex is not under evolutionary pressure, possibly because it is not the limiting factor in the slow growth rate of this bacterium.

4.7 Experimental validations

Every model prediction should be experimentally validated to make sure that it is sensible. In this section, I discuss potential ways in which the model described in this chapter has been or could be validated. Unfortunately, most relevant experiments on the biochemistry of translation have been done *in vitro* using *E. coli* components, while *M. pneumoniae* has not received the same level of attention until recently with the cryo-ET [22, 188] and ribosome profiling experiments [20]. Moreover, experimental manipulation of *M. pneumoniae* is considerably more difficult than *E. coli* due to pathogenicity (S2 class) and lack of established protocols for genetic manipulation.

In section 4.5 I showed that my generalized model replicates the results of the original Rudorf model, meaning that the predictions about the elongation rates are essentially the same. Two *in vivo* experiments have been used in the original study to confirm the model's prediction for *E. coli* [190]. One measured the relative speed of translation versus frameshifting, the other used a radioactive pulse-chase strategy to measure the incorporation of amino acids over time. In both cases, the experiment showed good agreement with the theory.

In *M. pneumoniae*, such experiments have not been performed, so I had to rely on more indirect data. The model predicts elongation rates ranging from 0.48 aa/s to 2.55 aa/s, with an average (weighted by codon usage frequency) of 2.1 aa/s. This value is quite far from the 15 aa/s of the reference *E. coli*. Importantly, the only constraint in the model is the steady-state distribution of intermediates, so the model is oblivious to any information about elongation rate. Yet, this estimate of the average elongation rate is remarkably similar to an independent estimate made by ribosome profiling [20], where the estimated elongation rates ranged from 0.55 aa/s to 4.91 aa/s with a weighted average of 1.81 aa/s. Although the average rates estimated with ribosome profiling and our model are in good agreement, the codon-specific rates are not. The Pearson correlation between the 62 rates is -0.02 (p-value 0.84). Such discrepancy might be explained by the different growth media used in different experiments: rich buffer for cryo-ET, and minimal medium for the ribosome profiling.

As another validation of the average elongation rate, Joe Dobbs from the Mahamid

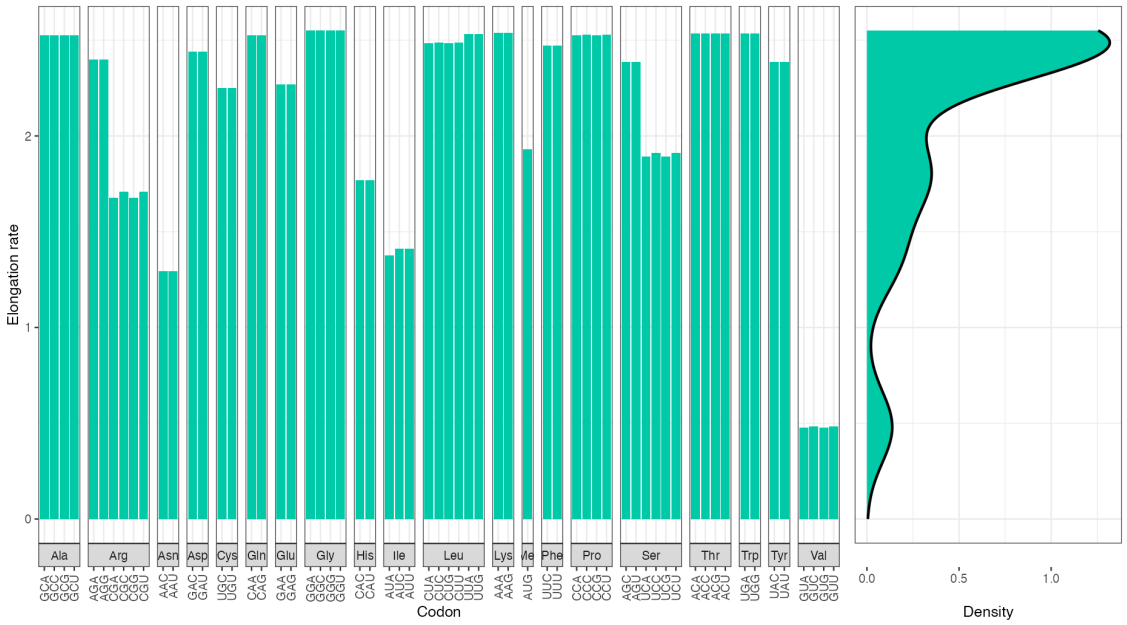


Figure 4.16: Codon-specific elongation rates predicted in *M. pneumoniae*, grouped by amino acid (left), and kernel density estimate of the distribution of elongation rates (right).

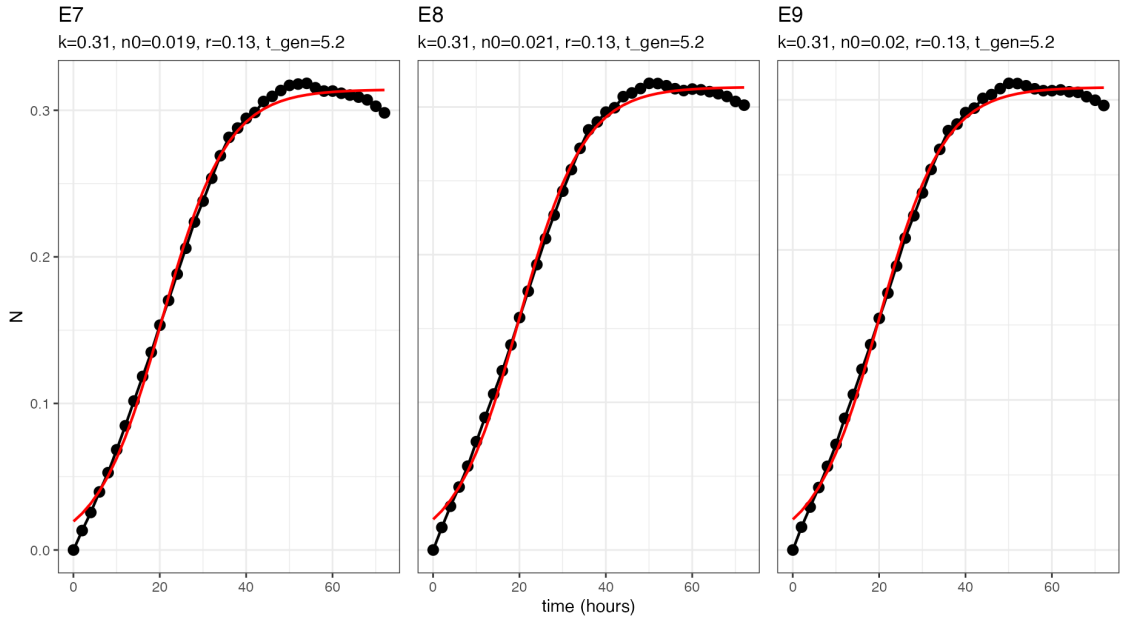


Figure 4.17: Three technical replicates of a growth curve experiment hosted in wells E7, E8, and E9 of a 96-well culture plate. The experiment used a plate reader to measure the pH of the medium every 2 hours for 72 hours in total. Medium acidification is a strong indicator of bacterial cell growth. After correcting the data by subtracting the background signal obtained from a well with pure medium, I processed the data with the growthcurver v0.3.1 package in R, which estimates the growth rate by fitting a logistic curve.

group and the Microbial Automation and Culturomics Core Facility at EMBL performed a growth curve experiment, and I analyzed the data. Assuming that the cells grow with an exponential rate, and also assuming that protein synthesis is the limiting factor for growth, it is known that the growth rate is directly related to the translation elongation rate [242]. The following relationship can be derived from mass-balance principles:

$$r_g = \frac{r_t \mathcal{R} \beta}{N_P} \tag{4.23}$$

where r_g is the growth rate, r_t the translation elongation rate, $\mathcal{R}$ the concentration of ribosomes, β the fraction of actively translating ribosomes on average, and N_P the total number of peptide bonds in the cell. For mycoplasma, the growth culture experiment yields a growth rate of 0.13 per hour. cryo-ET experiments show that the number of active ribosomes in the cell is around 250. And mass spectrometry experiments reveal that the protein content in this bacterium is about 107 g L^{-1} [231]. Considering a volume of 0.5 fL and an average amino acid weight of 110 Da, we estimate that the total number of incorporated amino acids (and therefore of peptide bonds) is ~3×10^7. Plugging everything in the equation above, we get an estimate of 4.4 aa/s. This number is higher than the 2.1 aa/s predicted by my model, but a few remarks are in order. First, 0.13 is the theoretical maximum rate at which the cells can grow, assuming unlimited resources and no competition; in typical situations, the rates would probably be lower. Second, the cells were grown in 96-well plates for the growth curve experiment, but for cryo-ET, they are grown directly on the grid that is used to hold the sample in the electron microscope. This different environment may well lead to a lower growth (and, consequently, elongation) rates.

If, in the future, more experiments are performed in the *M. pneumoniae* system, that data can be used to further refine the estimated rates or validate the predictions.

4.8 Sensitivity analysis

Uncertainty quantification are integral parts of the modelling process [238]. In its most general sense, sensitivity analysis is the study of how the outputs of a system are related to its inputs. This is a highly nontrivial task, especially in complex non-linear systems with many parameters. Uncertainty in the inputs does not directly translate in uncertainty in the predictions. The input parameters are affected by biological noise, biases and limitations of experimental protocols, different values in different growth conditions, and so on. Sensitivity analysis is instrumental in evaluating our conclusions, helping us judge which statements hold true across broad ranges of parameters and which are less robust. Besides mitigating errors, sensitivity analysis can also provide testable hypotheses about the behaviour of the system. For example, we could ask "what if" questions such as what would happen if the concentration of elongation factors was two times higher. The model can thus be used as

an *in silico* platform to provide interesting hypotheses. Given that ribosomes are among the main targets for bacterial antibiotics [230], having access to a comprehensive model of the elongation cycle could also enable clinical applications.

Due to the complexity of the model and the computational cost of its fitting, analyzing changes in all parameters at once would prove extremely challenging due to the exponential increase in the number of parameter combinations. Thus, the general strategy I adopted is to asses changes in turn for one, two, or at most three parameters, keeping the others constant. Moreover, I tried to optimize my code as much as possible, and I used parallelization to run multiple instances of the model at the same time. To manage the large number of runs (multiple organisms, multiple model versions, multiple conditions, multiple parameters), avoiding to unnecessarily regenerate the output files unless their inputs had changed, and further scaling the jobs across multiple compute nodes, I wrote a Snakemake workflow [173] which ended up having more than 1000 job steps. In this thesis I will only present the results for one model version ("with APE", see section 4.4.3), unless otherwise specified, but the same general conclusions hold for the other versions.

4.8.1 Steady-state proportions

First, I assessed the impact of variability in the steady-state proportions of ribosomes in each state. These data are used as hard constraint in the model, meaning that the predicted steady-state distribution will always exactly match the experimental data, when the model is coarse-grained. The cryo-ET dataset from Xue et al. [22] contains a total of 355 tomograms. For simplicity, we can assume that each tomogram contains one cell, and this is indeed mostly the case. However, sometimes the cell is incomplete, and sometimes there are two or more cells in the field of view of the microscope. This is, in general, not a problem since we expect a spatially uniform distribution of ribosomes both within and across cells, so I will use the terms tomogram and cell interchangeably. The count of ribosomes in each state exhibits some cell-to-cell variability. Similarly, fluctuations in the concentration of elongation factors and tRNAs from cell to cell are also to be expected. Moreover, the cells had not been synchronized, so they were all in different phases of the cell cycle when they were imaged. To try and understand the effects of such variability, I built cell-specific models and tried to correlate the estimated rates with other features that can be derived from the tomograms, such as cell volume and total number of ribosomes.

Figure 4.18 shows histograms of the number of ribosomes in each state across cells. The average cell has 250 ribosomes in total, but each cell can have a different proportions of intermediate states. In the previous section, I used the average steady-state distribution as constraint; here, I repeat the fit independently for each cell and investigate the variability in the predictions.

After discarding 41 cells where one or more intermediates could not be detected, I applied the kinetic distance minimization method to estimate the rates in *M. pneumoniae* using the

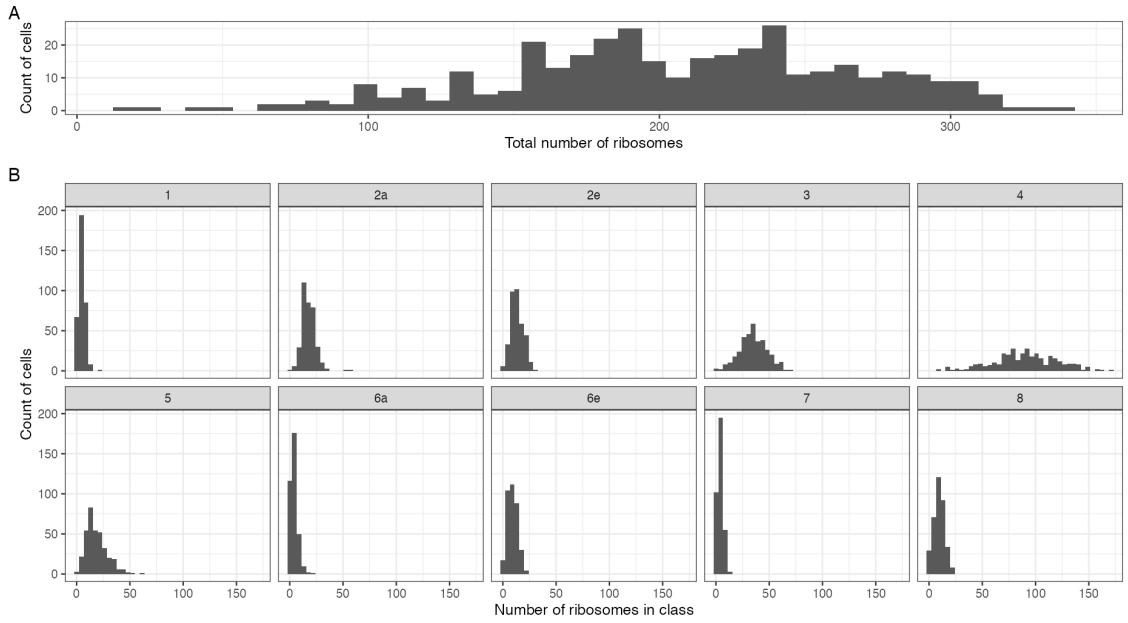


Figure 4.18: **A** Histogram of the number of ribosomes per cell across 355 tomograms. **B** Histograms of the number of ribosomes in each intermediate state.

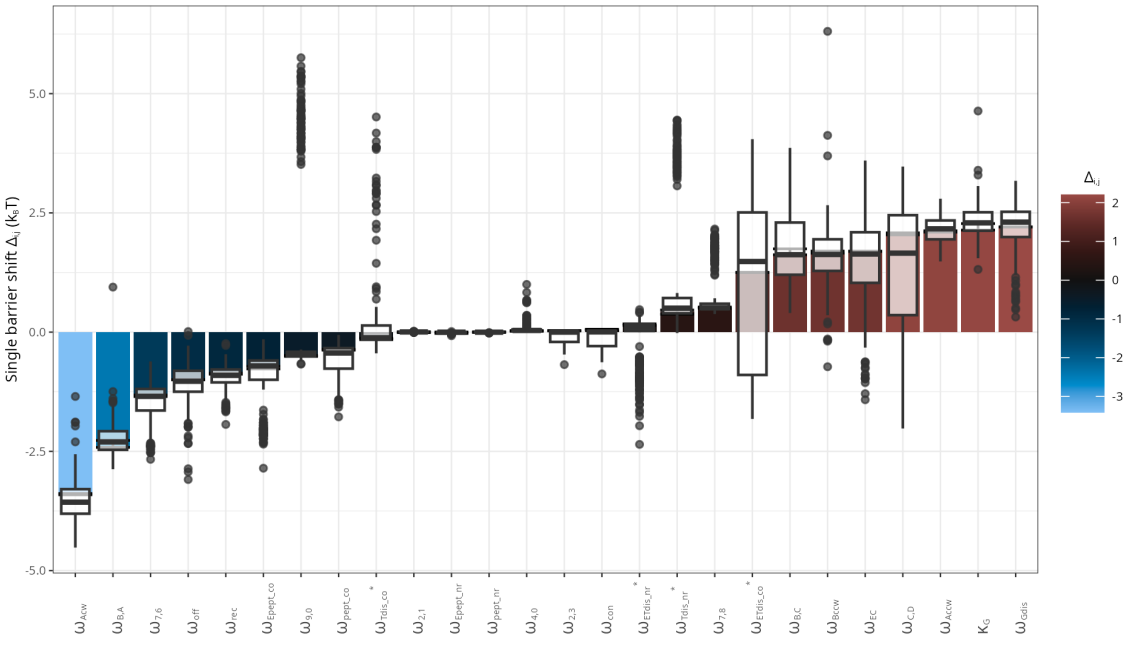


Figure 4.19: Variability in single-barrier shifts reflecting the cell-to-cell variability in the steady state distribution. The colored bars in the background show the estimate with the average steady-state distribution. The white box plots in the foreground show the distribution of predicted rates (one dot = one cell).

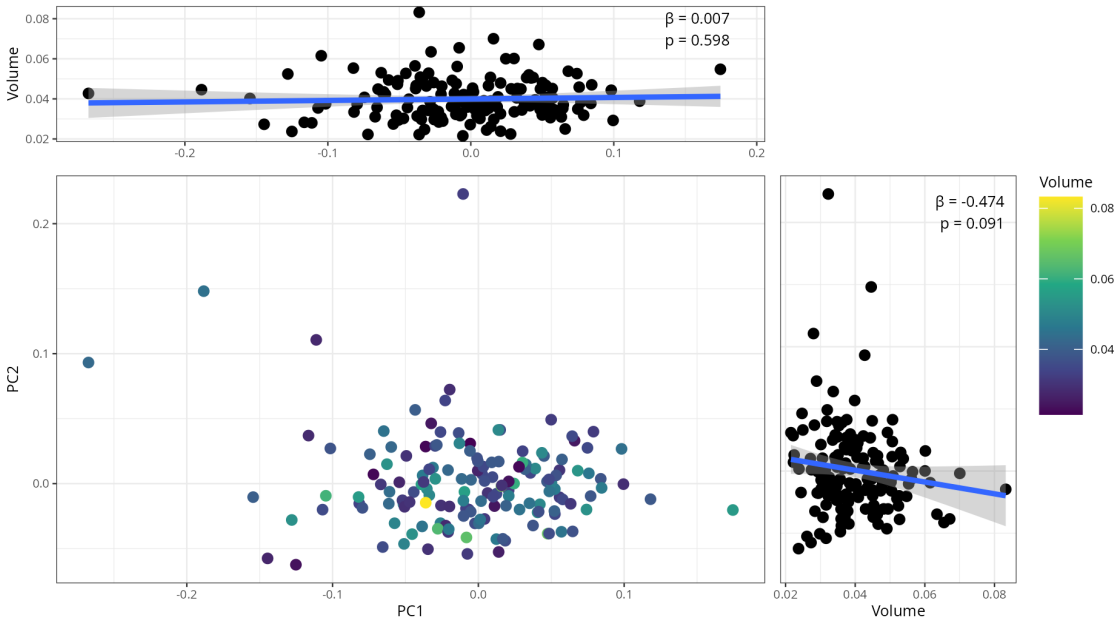


Figure 4.20: First two principal components of the steady-state distribution across cells, colored by cell volume. The insets to the top and right show scatter plots of the components versus the cell volume, along with a linear regression line, coefficient (β), and p-value.

cell-specific steady-state distribution as the constraint. The model failed to converge for 11 cells, likely due to insufficient number of initial conditions probed. For the 303 cells where the model was able to estimate the rates, the results are shown in fig. 4.19. Although some rates show high variability, the overall trend reflects the estimate achieved using the average steady-state distribution. There are, however, some clusters of tomograms where specific rates have a different value from the population average.

I then investigated the correlations between the steady state distributions and the cell volume. For this analysis, it was important to avoid using tomograms containing more than one cell or only part of a cell. Therefore, I used only a subset of the dataset containing "nice" cells. This dataset was created by Joe Dobbs (PhD student in the group of Julia Mahamid), who also calculated the cell volume, and it includes 153 cells. As the steady state distribution is a complex multidimensional object, I used PCA [243] and considered only the first two components. These components represent the axes with the most variability. Figure 4.20 shows the first two principal components (bottom left), along with the linear regression of the volume onto each of them (top and right). No clear association with volume can be seen.

Although the total number of ribosomes is moderately correlated with the cell volume (Pearson coefficient 0.45), it seems to be more directly associated with the steady state distribution. The proportion of ribosomes in state 4 (just before translocation) increases with the total count of ribosomes, while the ribosomes in states 5 (just before translocation) and

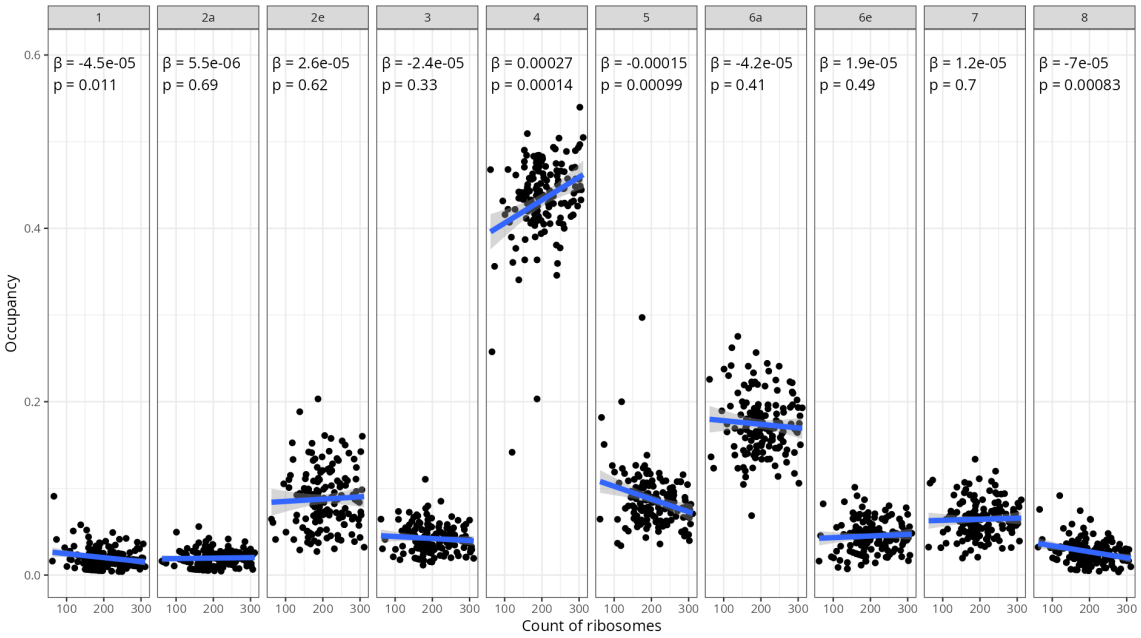


Figure 4.21: Occupancy of the 10 intermediate states versus total count of ribosomes across 153 cells for which the volume is available. Blue line: linear regression, also showing the coefficient (β) and p-value at the top.

8 (last step of translocation, just before release of EF-G) decrease as the count of ribosomes increases (fig. 4.21).[2] Furthermore, the average elongation rate is negatively correlated with the total count of ribosomes (Pearson coefficient -0.34, p-value 1.76×10^{-5}). Taken together, these observations suggests that state 4 (mapped to *Aco* and *Anr* in my model) is a bottleneck for translation elongation, as cells with a higher proportion of ribosomes in that state have slower elongation rates. The number of ribosomes is also related to the cell cycle phase: younger cells should have fewer ribosomes than older ones. Interestingly, the model predicts that older cells have slower elongation rates. Perhaps cells slow down elongation rate while getting older and preparing to divide, and they may do so by stalling ribosomes at state 4 with some unknown regulatory mechanism. Unfortunately we don't know how the concentrations of EF-G and other parameters change during the cell cycle.

Last, I analyzed how a higher occupancy of a particular state could influence the kinetics of translation elongation. For each intermediate state from Xue et al. [22], I binned the steady-state proportion across cells and grouped the single-barrier shifts for each rate arising from cells within each bin. The full matrix of plots is shown in fig. A.1 We can appreciate that the steady-state proportion of an intermediate tends to have the biggest impact on the rates that flow directly into or out of that state. For example, ω_{EC} is sensitive to the proportion of ribosomes in state 1 (fig. 4.22). This transition, which "produces"

[2] Perhaps confusingly, some state names in this study overlap with the state names in Xue et al. [22], see fig. 4.4. Here I refer to the states in Xue et al. [22]. See also fig. 4.4.

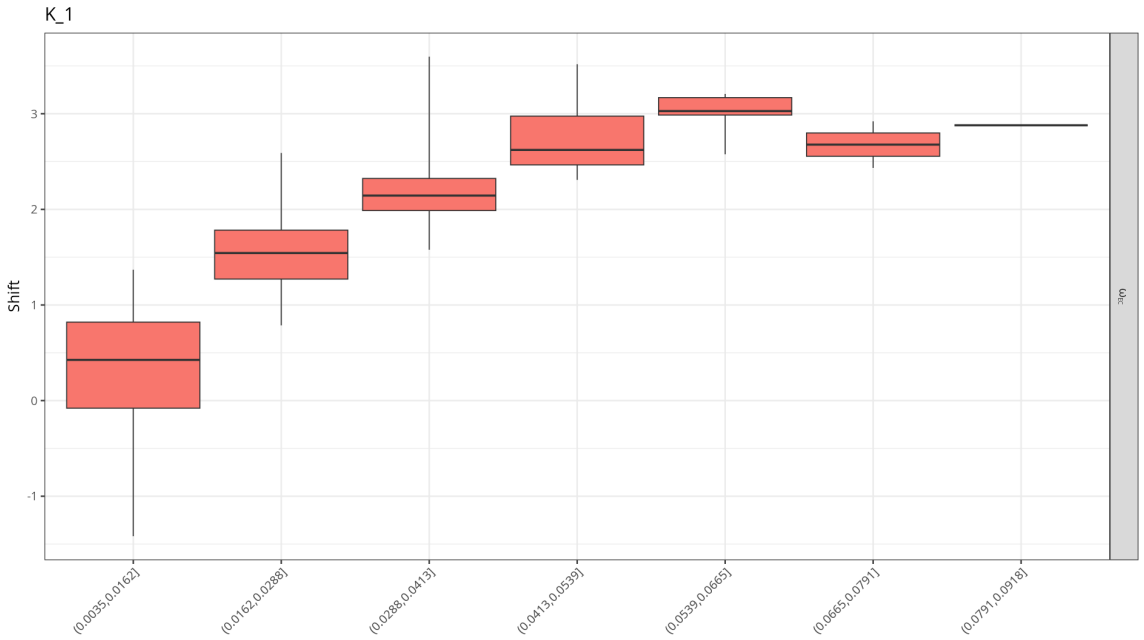


Figure 4.22: Single-barrier shift for rate ω_{EC} as a function of the occupancy of state 1.

intermediates of state 1, is faster as the occupancy increases.

4.8.2 Ribosome concentration and EF-G concentration

Next, I assessed the effect of the concentration of ribosomes and EF-G by varying these parameters in a grid of physiologically relevant values centered on the default estimates (table 4.4). The ribosome concentration, which is estimated to be 7 μM *in vivo*, has been probed between 2 μM and 8 μM. EF-G concentration, which is estimated to be 10 μM, from 4.7 μM to 64 μM.

The elongation rate steadily decreases as the concentration of ribosomes increases (as also observed in the previous section), while the concentration of EF-G has no effect on the elongation times (fig. 4.23 B). Overall, the total protein synthesis rate, defined as the product between the ribosome concentration and the elongation rate, peaks when the ribosome concentration is around 5 μM (fig. 4.23 A). The total protein synthesis rate expresses the capacity of the cell to synthesize proteins. Note that the profile is slightly skewed: the decrease in synthesis rate is sharper than the increase, signifying a non-linear relationship between ribosome concentrations and elongation time. One way to explain this is that when there are more active ribosomes, more tRNAs are "sequestered", thus unable to contribute to the pool of free ternary complexes.

Although EF-G has no effect on the estimated elongation time, it does affect some rates (fig. A.2). In particular, κ_G (the rate constant for the binding of EF-G) and ω_{BA} (the rate of EF-G unbinding) appear to change, across all concentrations of ribosomes, to compensate for the

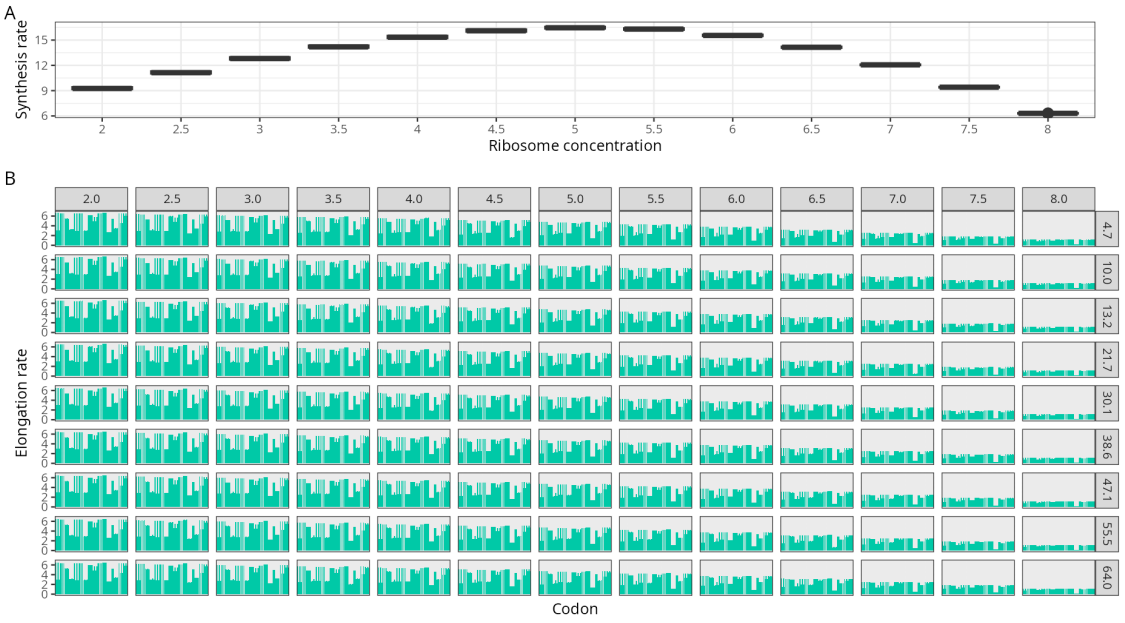


Figure 4.23: **A** Estimated synthesis rate (ribosome concentration times average elongation rate) as a function of the ribosome concentration. **B** Elongation rate as a function of the ribosome concentration (increasing along the columns) and EF-G concentration (increasing down the rows). Each bar is a codon.

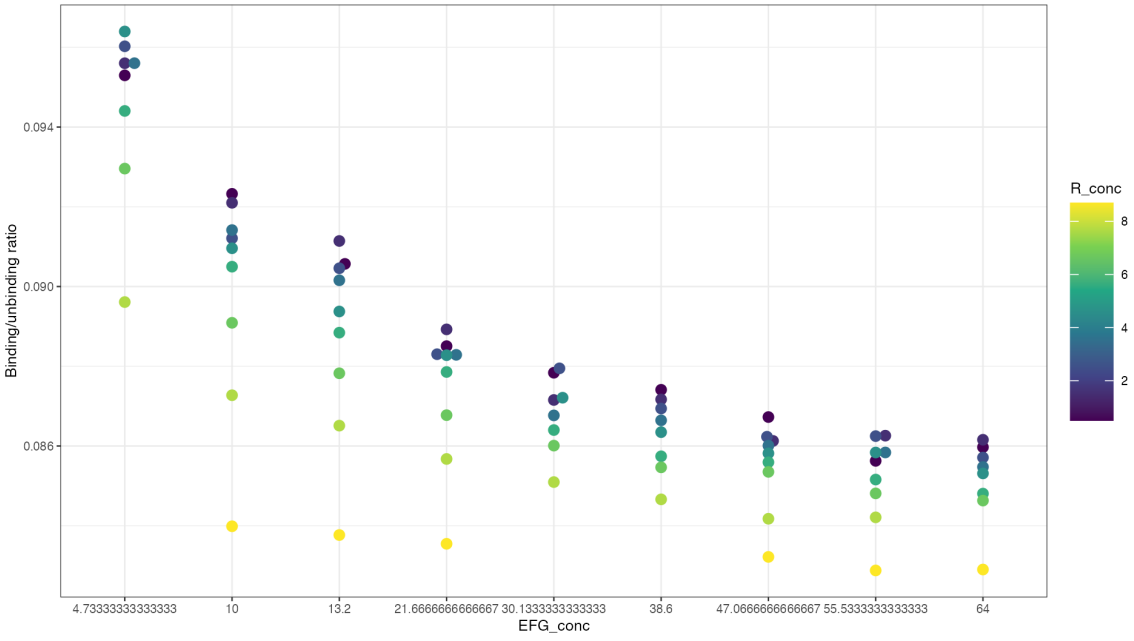


Figure 4.24: Ratio between the estimated rates of EF-G binding and unbinding as a function of the input value for the concentration of EF-G (x-axis) and ribosome (color).

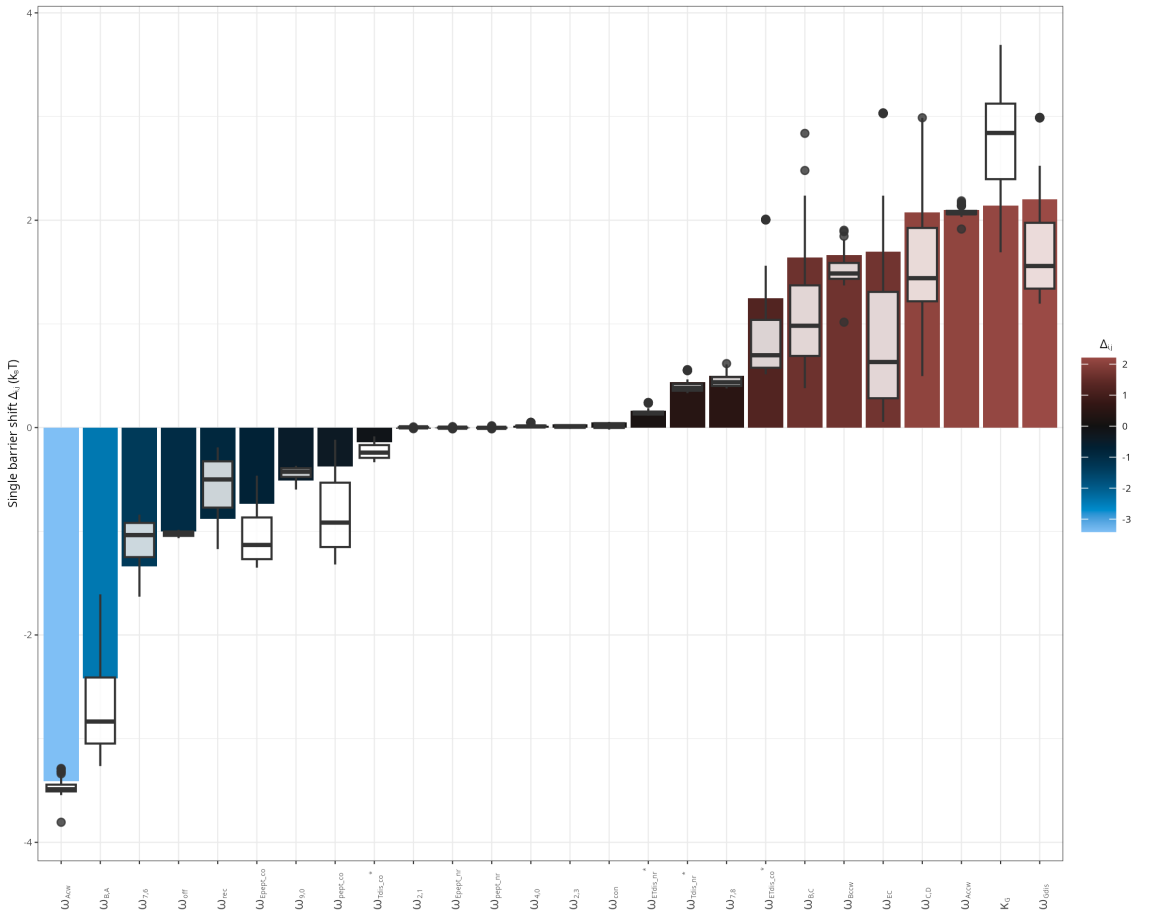


Figure 4.25: Variability in single-barrier shifts reflecting changes in the concentration of ribosomes and EF-G. The colored bars in the background show the estimate with the assumed default concentrations. The white box plots in the foreground show the distribution of predicted rates (one dot = one combination of ribosome and EF-G concentration different from the default values).

increased concentration of EF-G: the ratio between the forwarad and reverse reactions, $\frac{\kappa_G G}{\omega_{BA}}$, is relatively stable between 0.083 and 0.97 despite a 10-fold change in EF-G concentration (fig. 4.24).

It is important to remark that this analysis tells us how the model fitting procedure reacts to changes in the input concentration of the parameters. It does not tell us how does the cell react if we keep the elongation rates fixed and change the concentration of EF-G *in vivo*.

Although the concentration of EF-G is known only approximately, my sensitivity analysis shows that the results wouldn't change much by using a different value. Figure 4.25 shows boxplots of the rates estimated for all the combination of ribosome concentration and EF-G concentration that I probed, with the point-estimates in the background. Even assuming concentrations at the limits of what is physiologically reasonable, the rates of the decoding

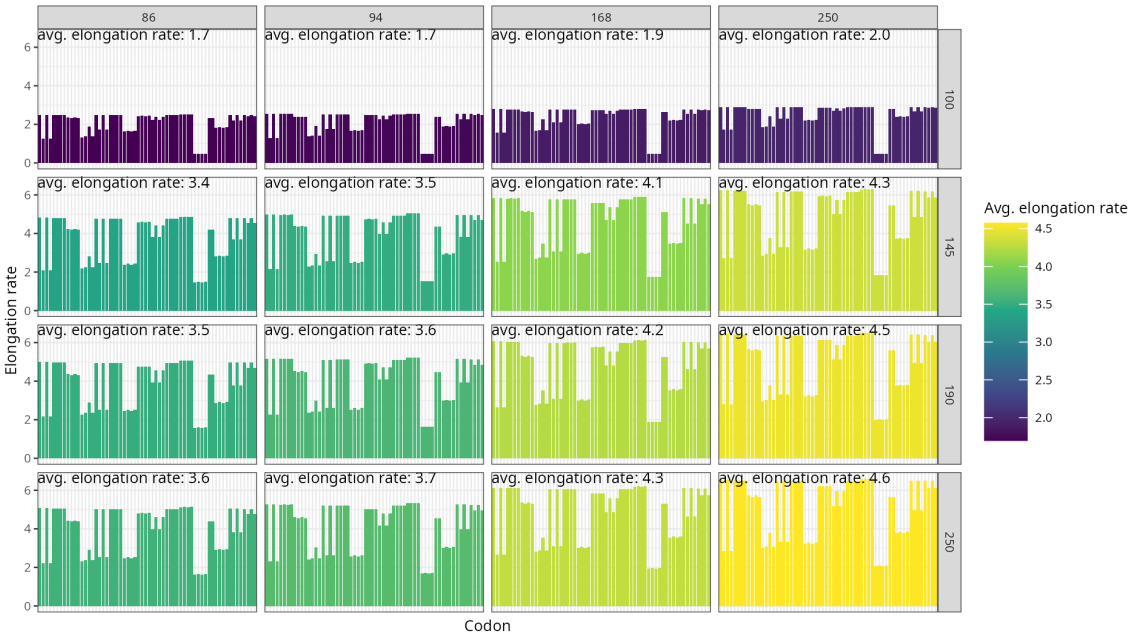


Figure 4.26: Elongation rate as a function of value of the κ_{on} rate (increasing along the columns) and EF-Tu concentration (increasing down the rows). Each bar is a codon.

step are still mostly similar to those in *E. coli*, while tanslocation is significantly slower.

4.8.3 EF-Tu concentration and κ_{on}

Other parameters that could have an impact on the estimated rates are related to the binding of ternary-complexes. These are κ_{on}, the rate constant of the binding of ternary complexes to ribosomes in states 0 or $0E$; and the concentration of EF-Tu, the elongation factor that carries the aa-tRNA to the ribosome and provides energy through the hydrolysis of GTP. A key feature of our model is that the concentration of free ternary complexes (EF-Tu · GTP · aa-tRNA) is identified as a parameter on its own right, distinct from the total concentration of tRNA. Indeed, tRNA molecules need to be loaded with an amino acid and bound to an EF-Tu molecule before being available to bind the ribosome (fig. 4.3 **B**). Still, the concentrations of EF-Tu and total tRNA influence the steady-state concentration of free ternary complexes, and are therefore important parameters to investigate.

Since we have no prior data on the value of κ_{on} in *M. pneumoniae*, I took it to be the same as *E. coli*, namely $94\,s^{-1}\,\mu M^{-1}$. The default values for EF-Tu concentration is 100 μM, and for the total tRNA concentration (summed over all 35 species) 110 μM, were reported in Weber et al. [20] and Miravet-Verde et al. [137]. Here, I fit the model for values of κ_{on} ranging from $87\,s^{-1}\,\mu M^{-1}$ to $250\,s^{-1}\,\mu M^{-1}$, EF-Tu concentration ranging from 100 μM to 220 μM (the fit errors out for values lower than 100), and tRNA concentration from 99 μM to 112 μM. The impact of κ_{on} on the elongation time is very small but becomes more noticeable at higher

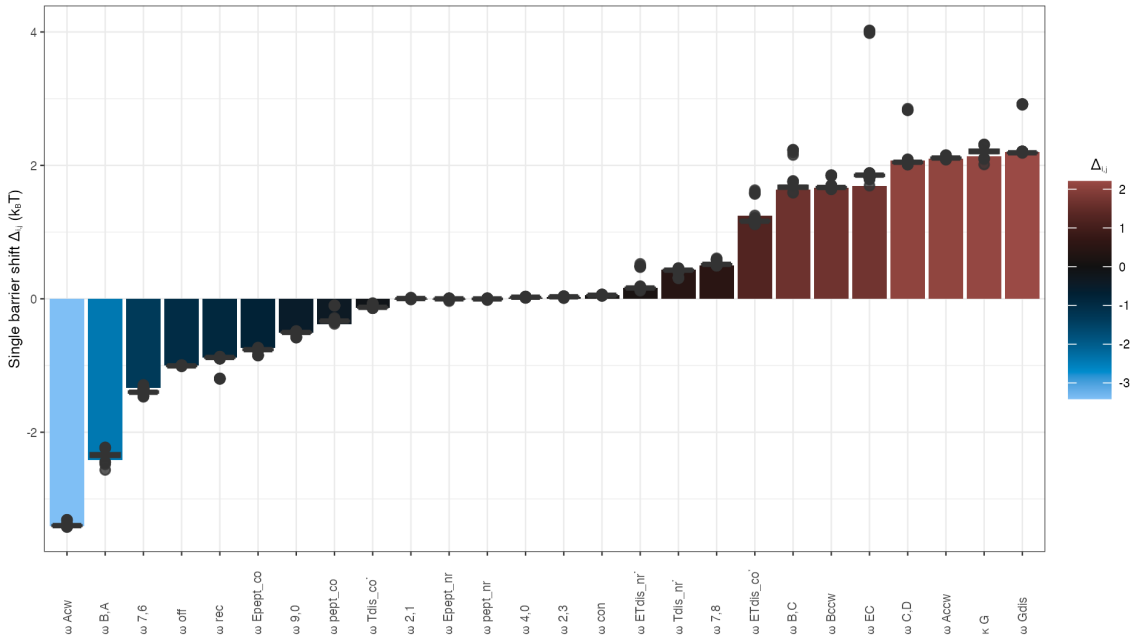


Figure 4.27: Sensitivity to cognacy matrix. The bars in the background show the default estimate, while the white boxplots in the foreground show the estimates obtained with randomized cognacy matrices, which do not deviate significantly from the default estimates.

concentrations of EF-Tu (fig. 4.26). On the other hand, increasing EF-Tu from 100 µM to 145 µM more than doubles the predicted average elongation rate. Interestingly, no strong compensatory effects on the rates are apparent for these two parameters, in contrast with EF-G as discussed above.

4.8.4 tRNA concentrations and cognacy matrix

Finally, I was interested in the sensitivity to the concentration of individual tRNA species and to the relationships between codons and tRNAs. *M. pneumoniae* has 35 tRNA species whose relative concentrations have been measured by Hydro-tRNA-seq in Weber et al. [20]. The cognacy relationships between codons and tRNAs can be inferred from the genetic code. When two or more tRNAs for the same amino acid are present, the codon-anticodon base pairing also matters. If codon and anticodon differ only at the third position, a recognition is still likely (the so-called wobbly base pairing) [244]. If they differ at any other position, the matching becomes much less likely, but can of course still happen.

To investigate the effect of uncertainty in the specific codon-anticodon binding affinities, I generated 1000 samples of randomly shuffled cognacy matrices. The shuffling procedure keeps the genetic code relationships unaltered, but it changes the codon-tRNA assignments when multiple tRNA for the same amino acid are available. As a consequence, any correlation between codon usage and abundance of cognate tRNAs would be broken. In

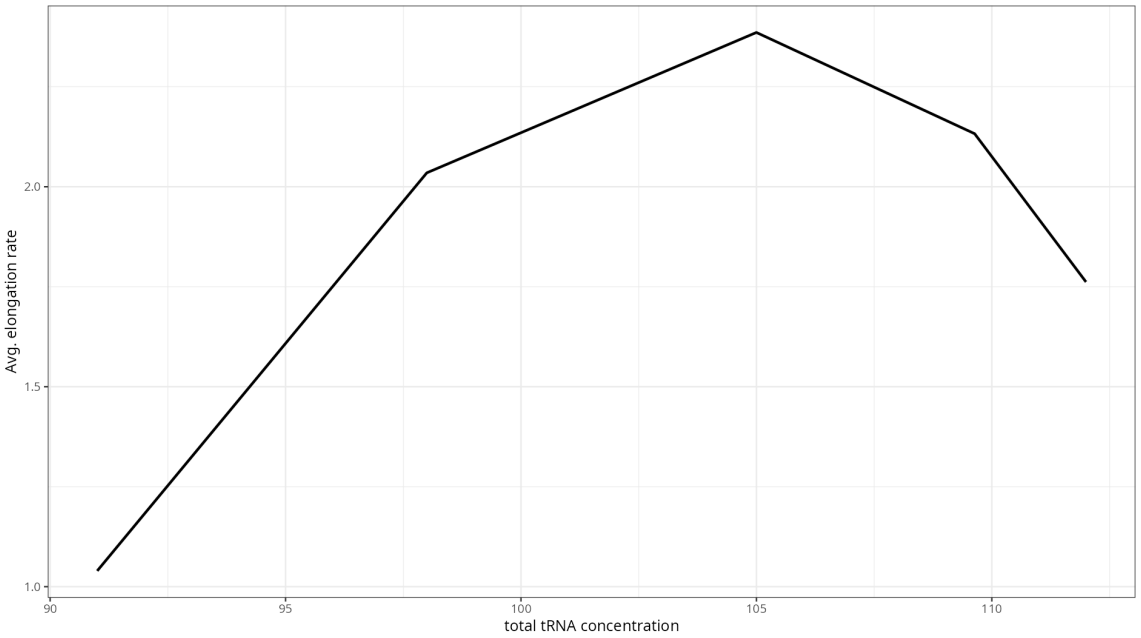


Figure 4.28: Estimated average elongation rate as a function of the total tRNA concentration (summed over all tRNA species).

M. pneumoniae, such correlation is not as strong as in *E. coli* (figs. 4.8 and 4.15). Figure 4.27 shows that the model is robust to misspecifications of the cognacy matrix, since the rates estimated with the 1000 randomized samples closely align with the rates estimated with the default matrix. Similarly, the elongation rates are barely affected: the mean of the distribution is $2.2\,\mathrm{s}^{-1}$ and the standard deviation only $0.083\,\mathrm{s}^{-1}$.

The relationship between elongation rate and total tRNA concentration is also non-linear (fig. 4.28), initially increasing and then starting to decrease after $105\,\mu\mathrm{M}$. This is due to the fact that non-cognate tRNAs increase faster than cognate tRNAs, leading to higher competition for the initial binding and requiring more proofreading. Once again, it is interesting that the default value (which I estimated using data from the literature and I fixed before doing any sensitivity analysis) is close to the peak of the curve, meaning that it is close to the optimal value for elongation speed.

4.9 The effect of antibiotics

One of the cryo-ET data sets that was collected included 64 tomograms of cells treated with Chloramphenicol, an antibiotic that is known to bind ribosomes and inhibit protein synthesis by blocking the formation of the peptide bond [188]. Moreover, the mechanism of action seems to be conditional on the sequence of the protein being translated: the inhibition occurs only when the amino acid to be incorporated is an Ala and the previous amino

acid in the peptide chain is not a Gly [245]. The steady-state occupancies observed in the treated data set appear quite different from those found in the unperturbed cells, with some intermediates vanishing and a new one becoming visible.

One of the potential applications where the model could be useful is in predicting or simulating the mechanism of action of ribosome antibiotics. As a proof of concepts, I took an antibiotic with a known mechanism of action, chloramphenicol, for which the cryo-ET data is available, and fit the model to the observed steady-state distribution. I reasoned that, if the model could estimate slower rates for the reactions that are indeed inhibited by this drug, it would be a good indication that the model is capable of suggesting hypotheses about mechanisms of action. Thus, I applied my model using the steady-state occupancies under chloramphenicol as the constraint and the estimated rates in the unperturbed *M. pneumoniae* system as the reference. The presence of extra states that were not reported in the unperturbed cryo-ET dataset is not a problem for my model, which is general enough that these new experimental structures can be mapped to its states. However, the fact that multiple states of the elongation cycle had not been detected in these tomograms posed a challenge, since using an occupancy of zero would keep the Markov chain stuck in some absorbing states. Thus, for the undetected states, I estimated their abundance with a value smaller than the approximate detection limit of intermediate states, but proportional to the abundances in the unperturbed condition. The model correctly estimates that the rate of peptide-bond formation undergoes a slowdown of several orders of magnitude (fig. 4.29).

An even better validation would be if we could artificially slow down the rate of peptide bond formation and reproduce the observed steady-state distribution. I try to do this in fig. 4.30, where the rate of the chloramphenicol-inhibited reaction is slowed down either 10.000 times compared to the value in the control condition, while the other rates are left untouched. I then calculate the steady-state distribution and coarse-grain it to the intermediate states that have been observed. The distributions become more similar, with the Kullback-Leibler divergence decreasing from 7.1 bits to 0.7 bits.

However, this approach suffers from several limitations. First and foremost, adding chloramphenicol doesn't only affect the ribosome in isolation, but it has a deep impact on the whole cell. Even neglecting potential off-target effects of the drug itself, blocking protein synthesis is a major stress that triggers the stringent response and its signalling nucleotides, pppGpp and ppGpp [19, 246]. This, in turn, causes a cascade of signaling and responses that alters virtually all biological processes. As such, the distance between the reference system (the unperturbed cells) and the target system is bound to increase dramatically; the kinetic distance minimization method, like any non-linear system, is likely to become less and less effective as the distance between reference and target system increases (see Strogatz and Fox [195] for an introduction to non-linear systems and chaos). Moreover, changes in the intracellular environment, such as different pH or concentrations of ions, could also have an effect on the ribosome kinetics. Mg^{2+} ions, in particular, are known to be a major factor influencing elongation speed [184]. All these effects make the estimation

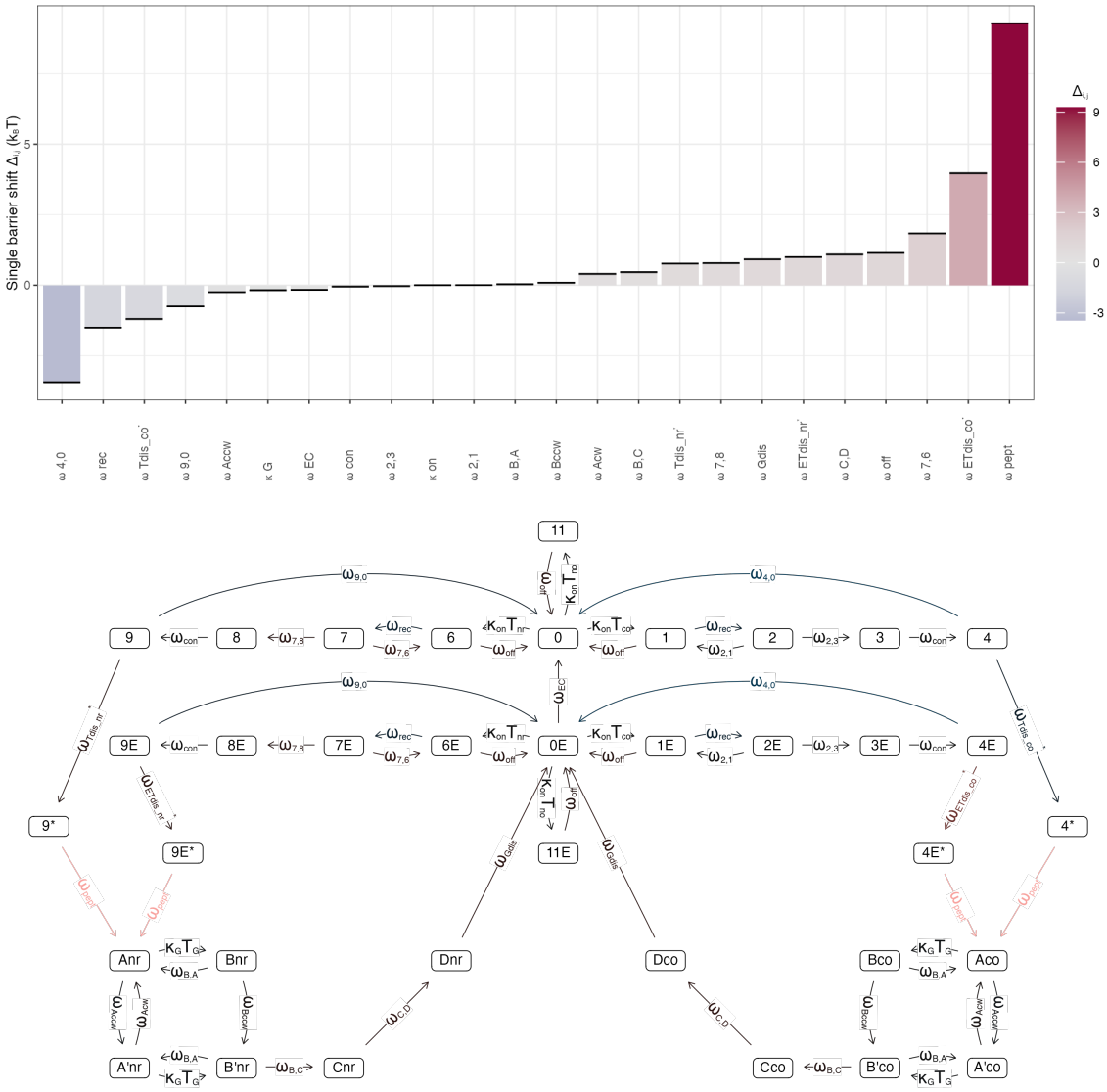


Figure 4.29: Single-barrier shifts estimated from steady-state occupancies of cells treated with chloramphenicol, which blocks the formation of the peptide bond. The estimated rates indicate a large slowdown of the corresponding rate in the model.

Figure 4.30: Comparison of the experimental steady-state occupancies with the occupancies predicted by the model. Top: prediction with the rates estimated in unperturbed *M. pneumoniae*. Bottom: prediction with the same rates, except that the rate of peptide bond formation was artificially slowed down by 10 000 times. The experimental occupancies do not change, but the predicted distribution becomes more similar to it.

of the rates more difficult, because they are not included as part of the model. Indeed, even in fig. 4.29, we see that the model predicts a significant slowing down for another rate, the dissociation of EF-Tu from ribosomes in state 4*E*, and, even more unexpectedly, it predicts the speeding up (by almost 2 times) of another rate, $\omega_{4,0}$. Furthermore, the simulation of the inhibition of the peptide bond formation doesn't perfectly reproduce the observed occupancies, suggesting that other, unmodelled effects are at play.

Another antibiotic-treated condition for which tomograms are available is Spectinomycin. This antibiotic binds to the small subunit of the ribosome, but its mechanism is less understood than for chloramphenicol. Despite the limitations mentioned above, I tried to repeat the fit using the steady-state distribution under spectinomycin. This time, I added an additional constraint that all single-barrier shifts should be non-increasing; this forces the model to fit the data using by only slowing rates down. The model predicts two rates to slow down the most: the dissociation of EF-Tu and the hydrolysis of GTP during translocation (fig. 4.31).

4.10 Discussion

In this chapter, I developed a model of the elongation cycle and generalized an approach based on constrained minimization to estimate the kinetic rates. Time-resolved biochemical studies *in vitro* are so far the only reliable way to measure the rates across the whole elongation cycle. *In vivo* approaches are considerably more challenging from the experimental

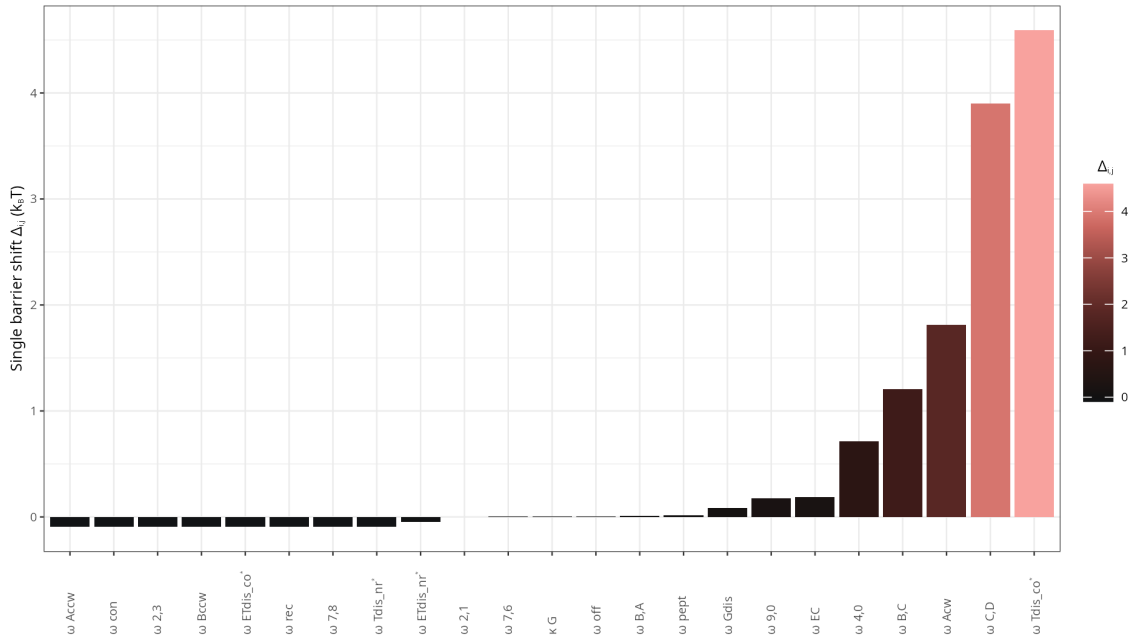


Figure 4.31: Single-barrier shifts estimated from steady-state occupancies of cells treated with spectinomycin. The estimated rates indicate a large slowdown of the rate of dissociation of EF-Tu in the cognate branch, and a slightly smaller slowdown of the rate of GTP hydrolysis in translocation.

point of view, and so far have not been able to measure rates at the resolution of the individual transitions. Rather, even with modern single-molecule fluorescent methods, it has only been possible to measure the total elongation time, *i.e.* the time taken by ribosomes to go through the full cycle and incorporate one amino acid in the nascent peptide chain [247, 248]. Here, I was interested in a finer resolution. cryo-ET data allows us to resolve the intermediate states *in situ*, therefore the reaction rates of the individual transitions become much more interesting. Furthermore, *in vitro* experiments are typically carried out in controlled conditions, often at or near thermodynamic equilibrium. However, in living cells, the process is far from equilibrium and the conditions are dynamic. cryo-ET is, so far, the only window we have into the native cellular environment. As the technology to investigate single rates *in vivo* is not yet available, it makes sense to use theoretical models to tackle the problem; after all, quoting Cohen [42], "mathematics is biology's next microscope".

This was a complex project from the technical point of view, and it required cooperation between biologists and theorists. To summarize, I updated and generalized the method introduced in Rudorf and Lipowsky [191]. The model of the translation elongation cycle now includes recently characterized intermediates and reactions. Furthermore, my implementation is very generic with respect to the constraint and the model structure, meaning that it can easily be extended to other experiments or even other biological processes. I also introduced a coarse-graining method that makes it possible to map the states between

various experiments and models. I initially replicated the original Rudorf model, estimating the rates for *E. coli in vivo* using the rates *in vitro* as reference. Then, I extended the study to unperturbed *M. pneumoniae*, using the rates in *E. coli* as reference, predicting that the translocation reactions are comparatively slow. The main validation comes from predicting an average total elongation rate that is consistent with independent experiments. I studied the effect of changes in parameters, showing that the estimates are overall robust, and discovering some non-linear relationships between elongation rates and concentration of ribosomes and tRNAs. Finally, I applied the model to extreme conditions, *M. pneumoniae* cells treated with antibiotics that blocks protein synthesis, providing insights into the mechanisms of action of these drugs.

A potentially hidden assumption is that the same intermediate states and the transitions among them exist *in vitro*, in *E. coli*, and in *M. pneumoniae*. So far, there has been a remarkable agreement between the states identified by cryo-ET and those reported from biochemical experiments *in vitro*. However, mapping experimental structures to the model's states is a critical step. The cryo-ET structure reconstruction by subtomogram averaging and classification (see section 1.2) is still a semi-manual process that requires human intervention. As such, it may include biases, although researchers are extremely careful in avoiding that. Moreover, ribosomes and their binding partners are thought to undergo smooth, continuous transitions from one state to the next, whereas in particle classification (and in my model) we collapse the continuum into a set of discrete intermediate states. This procedure is also affected by uncertainty. Furthermore, since the particle reconstruction and classification is performed *de novo* for any new dataset, it is impossible to directly compare structures from different studies. My model could also help in this regard, by providing a common ground of consensus intermediate states onto which the experimental structures can be mapped.

The rates estimated by the model are those that can explain the observed experimental data with in the most parsimonious way, *i.e.* requiring the minimal amount of change from the rates that are known. As my model is not a full molecular dynamics (MD) model, it doesn't capture the full physics of the process. Although coarse-grained MD simulations of ribosomes are becoming increasingly feasible [249], it would be a completely different project. Here, I focused on a more abstract model that can help elucidate the kinetics of the process in a rigorous way. The simplifying assumptions made for this model imply that the estimated rates are to be interpreted only as "effective rates", that is, they are approximations of the true rates, potentially depending on unmodelled parameters. Moore [202] recently reanalyzed the original Rudorf model and raised the important issue that the rate constants for the steps of elongation seem to vary with growth rate, even within the same organism (in this case, *E. coli*). This is true even for those steps that do not depend on the concentration of ternary complexes or elongation factors, suggesting that ribosomes in fast-growing cells are different from ribosomes in slow-growing cells. In light of the effective-rate interpretation, the fact that ribosomes decay into the next state with different

rates could be explained by different ion concentrations or different regulatory strategies in the various growth conditions. Although they are not explicitly modeled, their effect manifests itself in the different estimates for the rate values.

For my model, I tried to keep the design limited to the features that we could observe from the tomograms. I focused on capturing as many intermediate states as possible, provided that they had also been investigated *in vitro*. However, it is worth mentioning some features that have not been included. Protein synthesis is a highly regulated process, with multiple checkpoints before, during, and after [250]. Furthermore, it is entangled with other biological processes such as transcription [21]. My model only focuses on elongation, including ternary complexes, EF-Tu, and EF-G. It ignores the effect of frameshifting and mRNA structure on protein synthesis [251]. Finally, it is known that about 5% of ribosomes in *M. pneumoniae* are engaged in polysomes, interacting chains of 2-5 ribosomes that translate the same mRNA molecule in sync [22]. These interactions are known to streamline the translation process, resolve situation of stale, and increase the overall throughput of protein synthesis by parallelization [252].

Despite these limitations, the model is the most comprehensive and general abstract description of the translation-elongation cycle based on Markov chains. Its inclusion of both cognate and near-cognate branches enables the study of error rates, and it more closely reflects the underlying biology of the process. Its inclusion of both 2-1-2 and 2-3-2 pathways enables the investigation of the factors that could influence this choice [253, 193]. Using biochemical data and cryo-ET, it allows researchers to gain a dynamic understanding of the system starting from static snapshots. Although the data and the model are specific to translation-elongation, the same framework can be extended to other biological processes.

5 Conclusion

Although the trajectory of my PhD has not been linear, it is easy to identify the common themes that tie my work together. First, the model organism *M. pneumoniae*. Its resemblance to the minimal cell made it ideal for a variety of large-scale systems biology experiments. My supervisors Peer Bork and Luis Serrano had been working on this bacterium since the year 2000, so a great amount of data was already available. The first landmark *in situ* cryo-ET study of *M. pneumoniae* was published around the time when I joined EMBL [21]. From that study, it became clear that it would be possible to identify large macromolecular assemblies from the tomograms, *in primis* ribosomes, RNA polymerases, and GroEL-GroES complexes, but many more unknown assemblies were also visible. This realization started a quest to annotate all known complexes and protein-protein interactions in this organism. One of my very first projects was a random-forest machine learning model to predict protein-protein interactions from gene expression data. This project didn't make it in the final cut of the thesis because it became apparent that structure-based methods would perform best. Indeed, that was also the time when AlphaFold2 became available, triggering a global paradigm shift in the field of structural biology. Without AlphaFold, nobody knows how these projects would have turned out.

The identification of the dome-shaped complex in the tomograms demanded bioinformatic analyses, which is how I started to collaborate with Rasmus Jensen. Functional characterization of unknown proteins is a cornerstone of bioinformatics, a field to which the Bork group made substantial contributions through the development of tools such as STRING [82] and EGGNog [117]. Traditional approaches are based on sequence similarity between the uncharacterized protein and known, well-studied proteins. Such approaches have again proven useful in this instance, but it wasn't until we incorporated relatively novel, structure-similarity based approaches that we could really be confident in our results. In particular, focusing on core structural domains, deprived of disordered loops, was key to identifying homologous proteins.

Another activity that I started early on was an attempt to curate the available data for *M. pneumoniae*. Large amounts of data are not useful if they are impossible to find, buried in the supplementary materials of decade-old articles. My interest in the FAIR principles (Findability, Accessibility, Interoperability, and Reusability, [128]) led me to start working on a dashboard to present the data to the community working on *M. pneumoniae* at EMBL. Gradually, this project evolved into an idea to also aggregate all the data under a unified framework. Network science has proven an invaluable tool in understanding complex

systems [254]. Therefore, it was natural to aggregate all data in a unified heterogeneous network. The potential applications are endless: clustering, simulation of dynamic behavior, inference of new links, and more.

It is striking that, despite the large amount of sequencing data that is now available (over 25 trillion worth of nucleotides in the ENA [255]), the function of so many genes remains completely mysterious. In *M. pneumoniae*, which is considered a well-studied organism, at least one third of the genes have unknown function. The same goes for the artificial minimal cell, Syn3A [10]. This suggests the existence of completely new and unexpected biological processes, which are likely to be as impactful as CRISPR and RNA interference when they are discovered. The lessons learned from the dome complex project gave me inspiration to develop an automated pipeline for the annotation of the *M. pneumoniae* proteome using the annotation of individual domains. Extending this workflow to other organisms will provide valuable.

Given this starting point, in retrospect, I should have probably started to go in the direction of whole-cell modelling early on, inspired by Karr et al. [12] and Thornburg et al. [8]. Despite my interest in systems biology and mathematical modelling, I was not fully aware of those developments. Besides, the Mahamid group had a new cryo-ET dataset that called for a different analysis. Thus, rather than developing a broad, coarse-grained model of the entire cell, I developed a much more fine-grained model of a specific biological process, translation elongation. The idea stemmed from the availability of the abundances of the intermediate states of the elongation cycle. In whole-cell models, protein synthesis is typically modelled as a polymerization reaction with a uniform rate [8]. My model has tens of reactions, potentially with codon-specific rates. One of its predictions is that the rate of the translocation step (after the binding of EF-G) is slower in *M. pneumoniae* than *E. coli*, although it is still challenging to devise an *in vivo* validation for this claim. The acquisition of cryo-ET data in *E. coli* and perhaps new technological developments may allow us to refine the model in the future. In any case, I envision it as a starting point for bridging the gap between static snapshots and dynamic behavior, with a framework that can potentially be extended to other biological processes.

List of scientific contributions

From chapter 2 (manuscript on BiorXiv by Rasmus Jensen et al. [75] in preparation, I am listed as third author):

- Identification of the remote homology between an uncharacterized protein and the foldase PrsA;
- Identification of conserved catalytic residues in such protein, confirming its function;
- Phylogenetic analysis of this protein's family;

- Implementation of an open-source R package to perform co-occurrence analysis (`https://github.com/fmarotta/netcutter`);
- Implementation of an open-source R package to visualize reconciliated phylogenetic trees (`https://github.com/fmarotta/recPhylo`);

From chapter 3 (manuscript in preparation as first author):

- Creation of a web app to display and analyze *M. pneumoniae* data (the app is available internally at EMBL, but I plan to open-source it and make it publicly available);
- Development of a Snakemake workflow to split proteins into structural domains, annotate each domain individually, and aggregate the results (I plan to generalize it for different species, include some benchmarks, and also make it publicly available after my thesis submission);
- Implementation of a small R package to bridge the mol* JavaScript library into a Shiny app (`https://github.com/fmarotta/molstar-shiny`);
- Implementation of a circular slider plugin for Dash-Plotly (`https://github.com/fmarotta/dash-cisl`);

From chapter 4 (manuscript in preparation as first author):

- Design of a comprehensive kinetic model of the elongation cycle based on Markov chains;
- Implementation of a model calibration method based on constrained minimization in the Julia programming language (I plan to make the package available open-source, as well as the SBML file describing the model);
- Formal prediction that in *M. pneumoniae* the translocation step is slower than in *E. coli* (to be verified or disputed);
- Development of a Snakemake workflow to orchestrate more than 1000 jobs reflecting fitting of the model and sensitivity analysis for various organisms, conditions, and parameter values;

From chapter B (manuscript in preparation together with the challenge organizers):

- Development of a machine-learning model based on random forest and polygenic risk scores to predict the risk of high-cholesterol from genetic data and health survey;

Although not described in this thesis, I collaborated in the development of an R package to analyze the functional composition of a metagenomic sample (published by Zhao, Marotta, and Wu [256], I am listed as second author):

- Contribution to the software implementation and data visualization

Bibliography

[1] M D Eaton, M D Beck, and H E Pearson. "A virus from cases of atypical pneumonia : relation to the viruses of meningopneumonitis and psittacosis." In: *The Journal of Experimental Medicine* 73.5 (Apr. 1941), pp. 641–654. DOI: 10.1084/jem.73.5.641. URL: http://dx.doi.org/10.1084/jem.73.5.641 (visited on 06/17/2025).

[2] G Meiklejohn, M D Eaton, and W van Herick. "A clinical report on cases of primary atypical pneumonia caused by a new virus." In: *The Journal of Clinical Investigation* 24.2 (Mar. 1945), pp. 241–250. DOI: 10.1172/{JCI101600}. URL: http://dx.doi.org/10.1172/%7BJCI101600%7D (visited on 06/17/2025).

[3] R M Chanock, L Hayflick, and M F Barile. "Growth on artificial medium of an agent associated with atypical pneumonia and its identification as a PPLO." In: *Proceedings of the National Academy of Sciences of the United States of America* 48 (Jan. 1962), pp. 41–49. DOI: 10.1073/pnas.48.1.41. URL: http://dx.doi.org/10.1073/pnas.48.1.41 (visited on 06/17/2025).

[4] R Himmelreich et al. "Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae." In: *Nucleic Acids Research* 24.22 (Nov. 1996), pp. 4420–4449. DOI: 10.1093/nar/24.22.4420. URL: http://dx.doi.org/10.1093/nar/24.22.4420 (visited on 04/10/2023).

[5] T Dandekar et al. "Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames." In: *Nucleic Acids Research* 28.17 (Sept. 2000), pp. 3278–3288. DOI: 10.1093/nar/28.17.3278. URL: http://dx.doi.org/10.1093/nar/28.17.3278 (visited on 05/25/2025).

[6] Erika Gaspari et al. "Model-driven design allows growth of Mycoplasma pneumoniae on serum-free media." In: *NPJ Systems Biology and Applications* 6.1 (Oct. 2020), p. 33. DOI: 10.1038/s41540-020-00153-7. URL: http://dx.doi.org/10.1038/s41540-020-00153-7 (visited on 05/25/2025).

[7] G Biberfeld and P Biberfeld. "Ultrastructural features of Mycoplasma pneumoniae." In: *Journal of Bacteriology* 102.3 (June 1970), pp. 855–861. DOI: 10.1128/jb.102.3.855-861.1970. URL: http://dx.doi.org/10.1128/jb.102.3.855-861.1970 (visited on 06/29/2025).

[8] Zane R Thornburg et al. "Fundamental behaviors emerge from simulations of a living minimal cell." In: *Cell* 185.2 (Jan. 2022), 345–360.e28. ISSN: 00928674. DOI: 10.1016/j.cell.2021.12.025. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867421014884 (visited on 05/25/2025).

[9] John I Glass et al. "Essential genes of a minimal bacterium." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.2 (Jan. 2006), pp. 425–430. DOI: 10.1073/pnas.0510013103. URL: http://dx.doi.org/10.1073/pnas.0510013103 (visited on 05/25/2025).

[10] Clyde A Hutchison et al. "Design and synthesis of a minimal bacterial genome." In: *Science* 351.6280 (Mar. 2016), aad6253. DOI: 10.1126/science.aad6253. URL: http://dx.doi.org/10.1126/science.aad6253 (visited on 03/17/2022).

[11] Matthew J Betts and Robert B Russell. "The hard cell: from proteomics to a whole cell model." In: *FEBS Letters* 581.15 (June 2007), pp. 2870–2876. DOI: 10.1016/j.febslet.2007.05.062. URL: http://dx.doi.org/10.1016/j.febslet.2007.05.062 (visited on 05/25/2025).

[12] Jonathan R Karr et al. "A whole-cell computational model predicts phenotype from genotype." In: *Cell* 150.2 (July 2012), pp. 389–401. ISSN: 00928674. DOI: 10.1016/j.cell.2012.05.044. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092867412007763 (visited on 05/21/2015).

[13] Martina Maritan et al. "Building structural models of a whole mycoplasma cell." In: *Journal of Molecular Biology* 434.2 (Jan. 2022), p. 167351. ISSN: 00222836. DOI: 10.1016/j.jmb.2021.167351. URL: https://linkinghub.elsevier.com/retrieve/pii/%7BS002228362100588X%7D (visited on 05/25/2025).

[14] Anja Seybert, Richard Herrmann, and Achilleas S Frangakis. "Structural analysis of Mycoplasma pneumoniae by cryo-electron tomography." In: *Journal of Structural Biology* 156.2 (Nov. 2006), pp. 342–354. DOI: 10.1016/j.jsb.2006.04.010. URL: http://dx.doi.org/10.1016/j.jsb.2006.04.010 (visited on 01/17/2022).

[15] Eva Yus et al. "Impact of genome reduction on bacterial metabolism and its regulation." In: *Science* 326.5957 (Nov. 2009), pp. 1263–1268. DOI: 10.1126/science.1177263. URL: http://dx.doi.org/10.1126/science.1177263 (visited on 01/18/2022).

[16] Marc Güell et al. "Transcriptome complexity in a genome-reduced bacterium." In: *Science* 326.5957 (Nov. 2009), pp. 1268–1271. ISSN: 1095-9203. DOI: 10.1126/science.1176951. URL: http://dx.doi.org/10.1126/science.1176951 (visited on 06/03/2021).

[17] Sebastian Kühner et al. "Proteome organization in a genome-reduced bacterium." In: *Science* 326.5957 (Nov. 2009), pp. 1235–1240. DOI: 10.1126/science.1176343. URL: http://dx.doi.org/10.1126/science.1176343 (visited on 11/05/2021).

[18] Vera van Noort et al. "Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium." In: *Molecular Systems Biology* 8 (Feb. 2012), p. 571. DOI: 10.1038/msb.2012.4. URL: http://dx.doi.org/10.1038/msb.2012.4 (visited on 05/25/2025).

[19] Eva Yus et al. "Determination of the Gene Regulatory Network of a Genome-Reduced Bacterium Highlights Alternative Regulation Independent of Transcription Factors." In: *Cell Systems* 9.2 (Aug. 2019), 143–158.e13. ISSN: 24054712. DOI: 10.1016/j.cels.2019.07.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2405471219302327 (visited on 11/05/2021).

[20] Marc Weber et al. "Comprehensive quantitative modeling of translation efficiency in a genome-reduced bacterium." In: *Molecular Systems Biology* 19.10 (Oct. 2023), e11301. DOI: 10.15252/msb.202211301. URL: http://dx.doi.org/10.15252/msb.202211301 (visited on 05/17/2024).

[21] Francis J O'Reilly et al. "In-cell architecture of an actively transcribing-translating expressome." In: *Science* 369.6503 (July 2020), pp. 554–557. ISSN: 0036-8075. DOI: 10.1126/science.abb3758. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.abb3758 (visited on 01/18/2022).

[22] Liang Xue et al. "Visualizing translation dynamics at atomic detail inside a bacterial cell." In: *Nature* 610.7930 (Oct. 2022), pp. 205–211. ISSN: 0028-0836. DOI: 10.1038/s41586-022-05255-2. URL: https://www.nature.com/articles/s41586-022-05255-2 (visited on 04/10/2023).

[23] W G Weisburg et al. "A phylogenetic analysis of the mycoplasmas: basis for their classification." In: *Journal of Bacteriology* 171.12 (Dec. 1989), pp. 6455–6467. DOI: 10.1128/jb.171.12.6455-6467.1989. URL: http://dx.doi.org/10.1128/jb.171.12.6455-6467.1989 (visited on 05/25/2025).

[24] Conrad L Schoch et al. "NCBI Taxonomy: a comprehensive update on curation, resources and tools." In: *Database: the Journal of Biological Databases and Curation* 2020 (Jan. 2020), baaa062. DOI: 10.1093/database/baaa062. URL: http://dx.doi.org/10.1093/database/baaa062 (visited on 05/25/2025).

[25] Matthias Wolf et al. "Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data." In: *International Journal of Systematic and Evolutionary Microbiology* 54.Pt 3 (May 2004), pp. 871–875. DOI: 10.1099/ijs.0.02868-0. URL: http://dx.doi.org/10.1099/ijs.0.02868-0 (visited on 04/14/2025).

[26] Donovan H Parks et al. "GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy." In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D785–D794. DOI:

10.1093/nar/gkab776. URL: http://dx.doi.org/10.1093/nar/gkab776 (visited on 05/05/2024).

[27] C R Woese, J Maniloff, and L B Zablen. "Phylogenetic analysis of the mycoplasmas." In: *Proceedings of the National Academy of Sciences of the United States of America* 77.1 (Jan. 1980), pp. 494–498. DOI: 10.1073/pnas.77.1.494. URL: http://dx.doi.org/10.1073/pnas.77.1.494 (visited on 05/25/2025).

[28] Radhey S Gupta et al. "Phylogenetic framework for the phylum Tenericutes based on genome sequence data: proposal for the creation of a new order Mycoplasmoidales ord. nov., containing two new families Mycoplasmoidaceae fam. nov. and Metamycoplasmataceae fam. nov. harbouring Eperythrozoon, Ureaplasma and five novel genera." In: *Antonie Van Leeuwenhoek* 111.9 (Sept. 2018), pp. 1583–1630. DOI: 10.1007/s10482-018-1047-3. URL: http://dx.doi.org/10.1007/s10482-018-1047-3 (visited on 05/25/2025).

[29] Mitchell Balish et al. "Recommended rejection of the names Malacoplasma gen. nov., Mesomycoplasma gen. nov., Metamycoplasma gen. nov., Metamycoplasmataceae fam. nov., Mycoplasmoidaceae fam. nov., Mycoplasmoidales ord. nov., Mycoplasmoides gen. nov., Mycoplasmopsis gen. nov. [Gupta, Sawnani, Adeolu, Alnajar and Oren 2018] and all proposed species comb. nov. placed therein." In: *International Journal of Systematic and Evolutionary Microbiology* 69.11 (Nov. 2019), pp. 3650–3653. DOI: 10.1099/ijsem.0.003632. URL: http://dx.doi.org/10.1099/ijsem.0.003632 (visited on 05/25/2025).

[30] Radhey S Gupta and Aharon Oren. "Necessity and rationale for the proposed name changes in the classification of Mollicutes species. Reply to: 'Recommended rejection of the names Malacoplasma gen. nov., Mesomycoplasma gen. nov., Metamycoplasma gen. nov., Metamycoplasmataceae fam. nov., Mycoplasmoidaceae fam. nov., Mycoplasmoidales ord. nov., Mycoplasmoides gen. nov., Mycoplasmopsis gen. nov. [Gupta, Sawnani, Adeolu, Alnajar and Oren 2018] and all proposed species comb. nov. placed therein', by M. Balish et al. (Int J Syst Evol Microbiol, 2019;69:3650-3653)." In: *International Journal of Systematic and Evolutionary Microbiology* 70.2 (Feb. 2020), pp. 1431–1438. DOI: 10.1099/ijsem.0.003869. URL: http://dx.doi.org/10.1099/ijsem.0.003869 (visited on 05/25/2025).

[31] Rong Chen et al. "Evidence for the rapid and divergent evolution of mycoplasmas: structural and phylogenetic analysis of enolases." In: *Frontiers in molecular biosciences* 8 (2021), p. 811106. DOI: 10.3389/fmolb.2021.811106. URL: http://dx.doi.org/10.3389/fmolb.2021.811106 (visited on 04/09/2024).

[32] Martin Turk and Wolfgang Baumeister. "The promise and the challenges of cryo-electron tomography." In: *FEBS Letters* 594.20 (Oct. 2020), pp. 3243–3261. DOI: 10.

1002/1873-3468.13948. URL: http://dx.doi.org/10.1002/1873-3468.13948 (visited on 06/29/2025).

[33] Ye Hong et al. "Cryo-Electron Tomography: The Resolution Revolution and a Surge of In Situ Virological Discoveries." In: *Annual review of biophysics* (Jan. 2022). DOI: 10.1146/annurev-biophys-092022-100958. URL: http://dx.doi.org/10.1146/annurev-biophys-092022-100958 (visited on 06/29/2025).

[34] Fergus Tollervey et al. "Molecular architectures of centrosomes in C. elegans embryos visualized by cryo-electron tomography." In: *Developmental Cell* 60.6 (Mar. 2025), 885–900.e5. DOI: 10.1016/j.devcel.2024.12.002. URL: http://dx.doi.org/10.1016/j.devcel.2024.12.002 (visited on 06/29/2025).

[35] W Kühlbrandt. "The resolution revolution". In: *Science* 343.6178 (Mar. 2014), pp. 1443–1444. DOI: 10.1126/science.1251652. URL: http://dx.doi.org/10.1126/science.1251652 (visited on 08/08/2022).

[36] Ka Man Yip et al. "Atomic-resolution protein structure determination by cryo-EM." In: *Nature* 587.7832 (Nov. 2020), pp. 157–161. ISSN: 0028-0836. DOI: 10.1038/s41586-020-2833-4. URL: http://www.nature.com/articles/s41586-020-2833-4 (visited on 06/29/2025).

[37] J P Abrahams et al. "Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria." In: *Nature* 370.6491 (Aug. 1994), pp. 621–628. DOI: 10.1038/370621a0. URL: http://dx.doi.org/10.1038/370621a0 (visited on 03/30/2025).

[38] Friedrich Förster and Ariane Briegel, eds. *Cryo-Electron Tomography: Structural Biology in situ*. Vol. 11. Focus on structural biology. Cham: Springer International Publishing, 2024. ISBN: 978-3-031-51170-7. DOI: 10.1007/978-3-031-51171-4. URL: https://link.springer.com/10.1007/978-3-031-51171-4 (visited on 06/29/2025).

[39] Irene de Teresa-Trueba et al. "Convolutional networks for supervised mining of molecular patterns within cellular context." In: *Nature Methods* 20.2 (Feb. 2023), pp. 284–294. ISSN: 1548-7091. DOI: 10.1038/s41592-022-01746-2. URL: https://www.nature.com/articles/s41592-022-01746-2 (visited on 04/01/2025).

[40] Takanori Nakane and Sjors H W Scheres. "Multi-body Refinement of Cryo-EM Images in RELION." In: *Methods in Molecular Biology* 2215 (2021), pp. 145–160. DOI: 10.1007/978-1-0716-0966-8_7. URL: http://dx.doi.org/10.1007/978-1-0716-0966-8%5C_7 (visited on 06/29/2025).

[41] Joseph Yoniles et al. "Time-resolved cryogenic electron tomography for the study of transient cellular processes." In: *Molecular Biology of the Cell* 35.7 (July 2024), mr4. DOI: 10.1091/mbc.E24-01-0042. URL: http://dx.doi.org/10.1091/mbc.E24-01-0042 (visited on 03/30/2025).

[42] Joel E Cohen. "Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better." In: *PLoS Biology* 2.12 (Dec. 2004), e439. DOI: 10.1371/journal.pbio.0020439. URL: http://dx.doi.org/10.1371/journal.pbio.0020439 (visited on 03/26/2025).

[43] C. Levinthal. "How to fold graciously". In: (1969). URL: https://www.semanticscholar.org/paper/How-to-fold-graciously-Levinthal/1ef89dfb1e3404f4ace99399ce582b2bc982d0bf (visited on 03/30/2025).

[44] wwPDB consortium. "Protein Data Bank: the single global archive for 3D macromolecular structure data." In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D520–D528. DOI: 10.1093/nar/gky949. URL: http://dx.doi.org/10.1093/nar/gky949 (visited on 06/06/2020).

[45] UniProt Consortium. "Uniprot: the universal protein knowledgebase in 2025." In: *Nucleic Acids Research* 53.D1 (Jan. 2025), pp. D609–D617. DOI: 10.1093/nar/gkae1010. URL: http://dx.doi.org/10.1093/nar/gkae1010 (visited on 05/14/2025).

[46] Lorna Richardson et al. "MGnify: the microbiome sequence data analysis resource in 2023." In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D753–D759. DOI: 10.1093/nar/gkac1080. URL: http://dx.doi.org/10.1093/nar/gkac1080 (visited on 03/30/2025).

[47] Thomas S B Schmidt et al. "SPIRE: a Searchable, Planetary-scale mIcrobiome REsource." In: *Nucleic Acids Research* 52.D1 (Jan. 2024), pp. D777–D783. DOI: 10.1093/nar/gkad943. URL: http://dx.doi.org/10.1093/nar/gkad943 (visited on 03/30/2025).

[48] In-Geol Choi and Sung-Hou Kim. "Evolution of protein structural classes and protein sequence families." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.38 (Sept. 2006), pp. 14056–14061. DOI: 10.1073/pnas.0606239103. URL: http://dx.doi.org/10.1073/pnas.0606239103 (visited on 03/30/2025).

[49] Kristoffer Illergård, David H Ardell, and Arne Elofsson. "Structure is three to ten times more conserved than sequence–a study of structural response in protein cores." In: *Proteins* 77.3 (Nov. 2009), pp. 499–508. DOI: 10.1002/prot.22458. URL: http://dx.doi.org/10.1002/prot.22458 (visited on 04/09/2024).

[50] Szymon Kaczanowski and Piotr Zielenkiewicz. "Why similar protein sequences encode similar three-dimensional structures?" In: *Theoretical chemistry accounts* 125.3-6 (Mar. 2010), pp. 643–650. ISSN: 1432-881X. DOI: 10.1007/s00214-009-0656-3. URL: http://link.springer.com/10.1007/s00214-009-0656-3 (visited on 03/30/2025).

[51] C A Orengo et al. "CATH–a hierarchic classification of protein domain structures." In: *Structure* 5.8 (Aug. 1997), pp. 1093–1108. DOI: 10.1016/s0969-2126(97)00260-8. URL: http://dx.doi.org/10.1016/s0969-2126(97)00260-8 (visited on 03/30/2025).

[52] Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. "SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures." In: *Nucleic Acids Research* 42.Database issue (Jan. 2014), pp. D304–9. DOI: 10.1093/nar/gkt1240. URL: http://dx.doi.org/10.1093/nar/gkt1240 (visited on 03/30/2025).

[53] Ivica Letunic, Supriya Khedkar, and Peer Bork. "SMART: recent updates, new developments and status in 2020." In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D458–D460. DOI: 10.1093/nar/gkaa937. URL: http://dx.doi.org/10.1093/nar/gkaa937 (visited on 03/30/2025).

[54] Torsten Schwede et al. "SWISS-MODEL: An automated protein homology-modeling server." In: *Nucleic Acids Research* 31.13 (July 2003), pp. 3381–3385. DOI: 10.1093/nar/gkg520. URL: http://dx.doi.org/10.1093/nar/gkg520 (visited on 03/30/2025).

[55] J Moult et al. "A large-scale experiment to assess protein structure prediction methods." In: *Proteins* 23.3 (Nov. 1995), pp. ii–v. DOI: 10.1002/prot.340230303. URL: http://dx.doi.org/10.1002/prot.340230303 (visited on 03/30/2025).

[56] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold." In: *Nature* 596.7873 (Aug. 2021), pp. 583–589. ISSN: 0028-0836. DOI: 10.1038/s41586-021-03819-2. URL: http://www.nature.com/articles/s41586-021-03819-2 (visited on 10/01/2021).

[57] David Silver et al. "Mastering the game of Go with deep neural networks and tree search." In: *Nature* 529.7587 (Jan. 2016), pp. 484–489. DOI: 10.1038/nature16961. URL: http://dx.doi.org/10.1038/nature16961 (visited on 05/03/2016).

[58] Oriol Vinyals et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning." In: *Nature* 575.7782 (Nov. 2019), pp. 350–354. ISSN: 0028-0836. DOI: 10.1038/s41586-019-1724-z. URL: http://www.nature.com/articles/s41586-019-1724-z (visited on 05/06/2023).

[59] Mihaly Varadi et al. "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models." In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D439–D444. DOI: 10.1093/nar/gkab1061. URL: http://dx.doi.org/10.1093/nar/gkab1061 (visited on 05/23/2025).

[60] Mihaly Varadi et al. "AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences." In: *Nucleic Acids Research* 52.D1 (Jan. 2024), pp. D368–D375. DOI: 10.1093/nar/gkad1011. URL: http://dx.doi.org/10.1093/nar/gkad1011 (visited on 07/01/2024).

[61] Fabian Ruperti et al. "Cross-phyla protein annotation by structural prediction and alignment." In: *Genome Biology* 24.1 (May 2023), p. 113. DOI: 10.1186/s13059-023-02942-9. URL: http://dx.doi.org/10.1186/s13059-023-02942-9 (visited on 03/30/2025).

[62] Longxing Cao et al. "Design of protein-binding proteins from the target structure alone." In: *Nature* 605.7910 (May 2022), pp. 551–560. DOI: 10.1038/s41586-022-04654-9. URL: http://dx.doi.org/10.1038/s41586-022-04654-9 (visited on 03/25/2022).

[63] Michael Jendrusch, Jan O. Korbel, and S. Kashif Sadiq. "AlphaDesign: A de novo protein design framework based on AlphaFold". In: *BioRxiv* (Oct. 2021). DOI: 10.1101/2021.10.11.463937. URL: http://biorxiv.org/lookup/doi/10.1101/2021.10.11.463937 (visited on 03/30/2025).

[64] Rui Yin et al. "Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants." In: *Protein Science* 31.8 (Aug. 2022), e4379. ISSN: 0961-8368. DOI: 10.1002/pro.4379. URL: https://onlinelibrary.wiley.com/doi/10.1002/pro.4379 (visited on 03/30/2025).

[65] Josh Abramson et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3." In: *Nature* 630.8016 (June 2024), pp. 493–500. ISSN: 0028-0836. DOI: 10.1038/s41586-024-07487-w. URL: https://www.nature.com/articles/s41586-024-07487-w (visited on 05/15/2024).

[66] Nicola Bordin et al. "AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms." In: *Communications Biology* 6.1 (Feb. 2023), p. 160. DOI: 10.1038/s42003-023-04488-9. URL: http://dx.doi.org/10.1038/s42003-023-04488-9 (visited on 05/23/2025).

[67] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." In: *Science* 379.6637 (Mar. 2023), pp. 1123–1130. ISSN: 0036-8075. DOI: 10.1126/science.ade2574. URL: https://www.science.org/doi/10.1126/science.ade2574 (visited on 03/04/2024).

[68] Ashish Vaswani et al. "Attention is all you need". In: *arXiv* (2017). DOI: 10.48550/arxiv.1706.03762. URL: https://arxiv.org/abs/1706.03762 (visited on 11/15/2022).

[69] Jeremy Gunawardena. "Models in biology: 'accurate descriptions of our pathetic thinking'." In: *BMC Biology* 12 (Apr. 2014), p. 29. ISSN: 1741-7007. DOI: 10.1186/1741-7007-12-29. URL: http://www.biomedcentral.com/1741-7007/12/29 (visited on 03/31/2025).

[70] Nicholas Rhind. "Cell-size control." In: *Current Biology* 31.21 (Nov. 2021), R1414–R1420. DOI: 10.1016/j.cub.2021.09.017. URL: http://dx.doi.org/10.1016/j.cub.2021.09.017 (visited on 05/23/2025).

[71] Nitzan Rosenfeld and Uri Alon. "Response delays and the structure of transcription networks." In: *Journal of Molecular Biology* 329.4 (June 2003), pp. 645–654. DOI: 10.1016/s0022-2836(03)00506-0. URL: http://dx.doi.org/10.1016/s0022-2836(03)00506-0 (visited on 05/23/2025).

[72] Rob Phillips. "Theory in biology: figure 1 or figure 7?" In: *Trends in Cell Biology* 25.12 (Dec. 2015), pp. 723–729. DOI: 10.1016/j.tcb.2015.10.007. URL: http://dx.doi.org/10.1016/j.tcb.2015.10.007 (visited on 05/23/2025).

[73] Alfred Korzybski. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics Sixth Edition*. 6th ed. New York, New York, USA: Institute of General Semantics, 2023, p. 982. ISBN: 9781970164220. (Visited on 05/25/2025).

[74] Julia Mahamid et al. "Visualizing the molecular sociology at the HeLa cell nuclear periphery." In: *Science* 351.6276 (Feb. 2016), pp. 969–972. DOI: 10.1126/science.aad8857. URL: http://dx.doi.org/10.1126/science.aad8857 (visited on 03/22/2025).

[75] Rasmus K. Jensen et al. "In-cell discovery and characterization of a non-canonical bacterial protein translocation-folding complex". In: *bioRxiv* (Jan. 2025). URL: https://www.biorxiv.org/content/10.1101/2025.04.25.650208v1 (visited on 06/30/2025).

[76] Friedrich Förster, Bong-Gyoon Han, and Martin Beck. "Visual proteomics." In: *Methods in Enzymology* 483 (2010), pp. 215–243. DOI: 10.1016/S0076-6879(10)83011-3. URL: http://dx.doi.org/10.1016/S0076-6879(10)83011-3 (visited on 04/01/2025).

[77] David J F du Plessis, Nico Nouwen, and Arnold J M Driessen. "The Sec translocase." In: *Biochimica et Biophysica Acta* 1808.3 (Mar. 2011), pp. 851–865. DOI: 10.1016/j.bbamem.2010.08.016. URL: http://dx.doi.org/10.1016/j.bbamem.2010.08.016 (visited on 06/30/2025).

[78] Gydo van Zundert and Alexandre Bonvin. "Powerfit Software". In: *Zenodo* (2016). DOI: 10.5281/zenodo.1037228. URL: https://zenodo.org/record/1037228 (visited on 04/02/2025).

[79] Matthias Blum et al. "InterPro: the protein sequence classification resource in 2025." In: *Nucleic Acids Research* 53.D1 (Jan. 2025), pp. D444–D456. DOI: 10.1093/nar/gkae1082. URL: http://dx.doi.org/10.1093/nar/gkae1082 (visited on 04/14/2025).

[80] Philip Jones et al. "InterProScan 5: genome-scale protein function classification." In: *Bioinformatics* 30.9 (May 2014), pp. 1236–1240. DOI: 10.1093/bioinformatics/btu031. URL: http://dx.doi.org/10.1093/bioinformatics/btu031 (visited on 11/08/2021).

[81] Jaina Mistry et al. "Pfam: The protein families database in 2021." In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D412–D419. DOI: 10.1093/nar/gkaa913. URL: http://dx.doi.org/10.1093/nar/gkaa913 (visited on 08/03/2023).

[82] Damian Szklarczyk et al. "The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest." In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D638–D646. ISSN: 0305-1048. DOI: 10.1093/nar/gkac1000. URL: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 (visited on 10/22/2023).

[83] Christoph Elfmann et al. "MycoWiki: Functional annotation of the minimal model organism Mycoplasma pneumoniae." In: *Frontiers in Microbiology* 13 (July 2022), p. 935066. DOI: 10.3389/fmicb.2022.935066. URL: http://dx.doi.org/10.3389/fmicb.2022.935066 (visited on 04/14/2025).

[84] Maria Lluch-Senar et al. "Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium." In: *Molecular Systems Biology* 11.1 (Jan. 2015), p. 780. DOI: 10.15252/msb.20145558. URL: http://dx.doi.org/10.15252/msb.20145558 (visited on 04/10/2023).

[85] Christiam Camacho et al. "BLAST+: architecture and applications." In: *BMC Bioinformatics* 10 (Dec. 2009), p. 421. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-421. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-421 (visited on 01/25/2016).

[86] Sean Eddy and The team. *Biological sequence analysis using profile hidden Markov models*. WEBSITE. URL: https://hmmer.org (visited on 04/14/2025).

[87] Elisabeth Coudert et al. "Annotation of biologically relevant ligands in UniProtKB using ChEBI." In: *Bioinformatics* 39.1 (Jan. 2023). DOI: 10.1093/bioinformatics/btac793. URL: http://dx.doi.org/10.1093/bioinformatics/btac793 (visited on 06/18/2024).

[88] Daniel R Mende et al. "proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes." In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D529–D534. DOI: 10.1093/nar/gkw989. URL: http://dx.doi.org/10.1093/nar/gkw989 (visited on 04/14/2025).

[89] Martin Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation." In: *BMC Bioinformatics* 20.1 (Sept. 2019), p. 473. DOI: 10.1186/s12859-019-3019-7. URL: http://dx.doi.org/10.1186/s12859-019-3019-7 (visited on 04/14/2025).

[90] S F Altschul et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." In: *Nucleic Acids Research* 25.17 (Sept. 1997), pp. 3389–3402. DOI: 10.1093/nar/25.17.3389. URL: http://dx.doi.org/10.1093/nar/25.17.3389 (visited on 12/12/2016).

[91] Ganapathi Varma Saripella, Erik L L Sonnhammer, and Kristoffer Forslund. "Benchmarking the next generation of homology inference tools." In: *Bioinformatics* 32.17 (Sept. 2016), pp. 2636–2641. DOI: 10.1093/bioinformatics/btw305. URL: http://dx.doi.org/10.1093/bioinformatics/btw305 (visited on 03/30/2022).

[92] Milot Mirdita et al. "Uniclust databases of clustered and deeply annotated protein sequences and alignments." In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D170–D176. DOI: 10.1093/nar/gkw1081. URL: http://dx.doi.org/10.1093/nar/gkw1081 (visited on 04/15/2025).

[93] Roman P Jakob et al. "Dimeric structure of the bacterial extracellular foldase prsa." In: *The Journal of Biological Chemistry* 290.6 (Feb. 2015), pp. 3278–3292. DOI: 10.1074/jbc.M114.622910. URL: http://dx.doi.org/10.1074/jbc.M114.622910 (visited on 04/12/2020).

[94] Michel van Kempen et al. "Fast and accurate protein structure search with Foldseek." In: *Nature Biotechnology* 42.2 (Feb. 2024), pp. 243–246. ISSN: 1087-0156. DOI: 10.1038/s41587-023-01773-0. URL: https://www.nature.com/articles/s41587-023-01773-0 (visited on 06/06/2023).

[95] Liisa Holm. "Using DALI for protein structure comparison." In: *Methods in Molecular Biology* 2112 (2020), pp. 29–42. DOI: 10.1007/978-1-0716-0270-6_3. URL: http://dx.doi.org/10.1007/978-1-0716-0270-6%5C_3 (visited on 04/15/2025).

[96] Liisa Holm et al. "DALI shines a light on remote homologs: One hundred discoveries." In: *Protein Science* 32.1 (Jan. 2023), e4519. DOI: 10.1002/pro.4519. URL: http://dx.doi.org/10.1002/pro.4519 (visited on 04/16/2025).

[97] Yang Zhang and Jeffrey Skolnick. "TM-align: a protein structure alignment algorithm based on the TM-score." In: *Nucleic Acids Research* 33.7 (Apr. 2005), pp. 2302–2309. DOI: 10.1093/nar/gki524. URL: http://dx.doi.org/10.1093/nar/gki524 (visited on 03/02/2017).

[98] Paolo Di Tommaso et al. "Nextflow enables reproducible computational workflows." In: *Nature Biotechnology* 35.4 (Apr. 2017), pp. 316–319. DOI: 10.1038/nbt.3820. URL: http://dx.doi.org/10.1038/nbt.3820 (visited on 07/13/2018).

[99] Zhanwen Li et al. "FATCAT 2.0: towards a better understanding of the structural diversity of proteins." In: *Nucleic Acids Research* 48.W1 (July 2020), W60–W64. DOI: 10.1093/nar/gkaa443. URL: http://dx.doi.org/10.1093/nar/gkaa443 (visited on 04/16/2025).

[100] Ziheng Yang. *Molecular evolution: A statistical approach*. Oxford University Press, May 2014. ISBN: 9780199602605. DOI: 10.1093/acprof:oso/9780199602605.001.0001. URL: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199602605.001.0001/acprof-9780199602605 (visited on 04/16/2025).

[101] R Nichols. "Gene trees and species trees are not the same." In: *Trends in Ecology & Evolution* 16.7 (July 2001), pp. 358–364. DOI: 10.1016/s0169-5347(01)02203-0. URL: http://dx.doi.org/10.1016/s0169-5347(01)02203-0 (visited on 04/16/2025).

[102] Benoit Morel et al. "GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss." In: *Molecular Biology and Evolution* 37.9 (Sept. 2020), pp. 2763–2774. DOI: 10.1093/molbev/msaa141. URL: http://dx.doi.org/10.1093/molbev/msaa141 (visited on 04/11/2024).

[103] Bastien Boussau and Celine Scornavacca. "Reconciling Gene trees with Species Trees". In: *DataverseNL* (2024). DOI: 10.34894/vq1dja. URL: https://dataverse.nl/citation?%7BpersistentId%7D=doi:10.34894/%7BVQ1DJA%7D (visited on 04/16/2025).

[104] Nuala A O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D733–45. DOI: 10.1093/nar/gkv1189. URL: http://dx.doi.org/10.1093/nar/gkv1189 (visited on 04/11/2024).

[105] J. S. Farris. "Outgroups and Parsimony". In: *Systematic Biology* 31.3 (Sept. 1982), pp. 328–334. ISSN: 1063-5157. DOI: 10.1093/sysbio/31.3.328. URL: https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/31.3.328 (visited on 04/16/2025).

[106] John P Huelsenbeck, Jonathan P Bollback, and Amy M Levine. "Inferring the root of a phylogenetic tree." In: *Systematic Biology* 51.1 (Feb. 2002), pp. 32–43. DOI: 10.1080/106351502753475862. URL: http://dx.doi.org/10.1080/106351502753475862 (visited on 04/16/2025).

[107] Tatiana Tatusova et al. "NCBI prokaryotic genome annotation pipeline." In: *Nucleic Acids Research* 44.14 (Aug. 2016), pp. 6614–6624. DOI: 10.1093/nar/gkw569. URL: http://dx.doi.org/10.1093/nar/gkw569 (visited on 04/16/2025).

[108] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. "FastTree 2 —approximately maximum-likelihood trees for large alignments." In: *Plos One* 5.3 (Mar. 2010), e9490. DOI: 10.1371/journal.pone.0009490. URL: http://dx.doi.org/10.1371/journal.pone.0009490 (visited on 12/18/2017).

[109] Pierre-Alain Chaumeil et al. "GTDB-Tk v2: memory friendly classification with the genome taxonomy database." In: *Bioinformatics* 38.23 (Nov. 2022), pp. 5315–5316. DOI: 10.1093/bioinformatics/btac672. URL: http://dx.doi.org/10.1093/bioinformatics/btac672 (visited on 04/11/2024).

[110] Kazutaka Katoh and Daron M Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." In: *Molecular Biology and Evolution* 30.4 (Apr. 2013), pp. 772–780. DOI: 10.1093/molbev/mst010. URL: http://dx.doi.org/10.1093/molbev/mst010 (visited on 02/28/2017).

[111] Robert C. Edgar. "Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny". In: *Nature Communications* 13.1 (Nov. 2022), p. 6968. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34630-w. URL: https://www.nature.com/articles/s41467-022-34630-w (visited on 05/08/2025).

[112] Gerard Talavera and Jose Castresana. "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments." In: *Systematic Biology* 56.4 (Aug. 2007), pp. 564–577. DOI: 10.1080/10635150701472164. URL: http://dx.doi.org/10.1080/10635150701472164 (visited on 05/29/2022).

[113] Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón. "trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses." In: *Bioinformatics* 25.15 (Aug. 2009), pp. 1972–1973. DOI: 10.1093/bioinformatics/btp348. URL: http://dx.doi.org/10.1093/bioinformatics/btp348 (visited on 12/12/2016).

[114] A Bolhuis et al. "SecDF of Bacillus subtilis, a molecular Siamese twin required for the efficient secretion of proteins." In: *The Journal of Biological Chemistry* 273.33 (Aug. 1998), pp. 21217–21224. DOI: 10.1074/jbc.273.33.21217. URL: http://dx.doi.org/10.1074/jbc.273.33.21217 (visited on 05/09/2025).

[115] David M Emms and Steven Kelly. "OrthoFinder: phylogenetic orthology inference for comparative genomics." In: *Genome Biology* 20.1 (Nov. 2019), p. 238. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1832-y. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y (visited on 02/05/2020).

[116] Price et al. "Interactive tools for functional annotation of bacterial genomes". In: *Database* 2024 (Feb. 2024). URL: https://academic.oup.com/database/article/doi/10.1093/database/baae089/7750433 (visited on 05/09/2025).

[117] Jaime Huerta-Cepas et al. "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses." In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D309–D314. ISSN: 0305-1048. DOI: 10.1093/nar/gky1085. URL: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky1085/5173662 (visited on 12/03/2021).

[118] Morgan N Price and Adam P Arkin. "A fast comparative genome browser for diverse bacteria and archaea." In: *Plos One* 19.4 (Apr. 2024), e0301871. DOI: 10.1371/journal.pone.0301871. URL: http://dx.doi.org/10.1371/journal.pone.0301871 (visited on 05/09/2025).

[119] Heiko Müller and Francesco Mancuso. "Identification and analysis of co-occurrence networks with NetCutter." In: *Plos One* 3.9 (Sept. 2008), e3178. DOI: 10.1371/journal.pone.0003178. URL: http://dx.doi.org/10.1371/journal.pone.0003178 (visited on 07/04/2023).

[120] Wenpin Tang and Fengmin Tang. "The poisson binomial distribution—old & new". In: *Statistical Science* 38.1 (Feb. 2023). ISSN: 0883-4237. DOI: 10.1214/22-{STS852}. URL: https://projecteuclid.org/journals/statistical-science/volume-38/issue-1/The-Poisson-Binomial-Distribution-Old--New/10.1214/22-%7BSTS852%7D.full (visited on 07/03/2025).

[121] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. URL: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008?key=crossref.46968f6ec61eb8f907a760be1c5ace52 (visited on 12/07/2016).

[122] Gábor Csárdi et al. "igraph for R: R interface of the igraph library for graph theory and network analysis". In: *Zenodo* (2025). DOI: 10.5281/zenodo.14736815. URL: https://zenodo.org/doi/10.5281/zenodo.14736815 (visited on 05/11/2025).

[123] Nigel F Delaney et al. "Ultrafast evolution and loss of CRISPRs following a host shift in a novel wildlife pathogen, Mycoplasma gallisepticum." In: *PLoS Genetics* 8.2 (Feb. 2012), e1002511. DOI: 10.1371/journal.pgen.1002511. URL: http://dx.doi.org/10.1371/journal.pgen.1002511 (visited on 04/14/2025).

[124] Shlomo Trachtenberg. "Mollicutes." In: *Current Biology* 15.13 (July 2005), R483–4. DOI: 10.1016/j.cub.2005.06.049. URL: http://dx.doi.org/10.1016/j.cub.2005.06.049 (visited on 04/14/2025).

[125] Mario Vailati-Riboni, Valentino Palombo, and Juan J. Loor. "What are omics sciences?" In: *Periparturient diseases of dairy cows*. Ed. by Burim N. Ametaj. Cham: Springer International Publishing, 2017, pp. 1–7. ISBN: 978-3-319-43031-7. DOI: 10.1007/978-3-319-43033-1_1. URL: http://link.springer.com/10.1007/978-3-319-43033-1%5C_1 (visited on 05/12/2025).

[126] Robert D Olson et al. "Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR." In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D678–D689. DOI: 10.1093/nar/gkac1003. URL: http://dx.doi.org/10.1093/nar/gkac1003 (visited on 03/04/2025).

[127] Judith A H Wodke et al. "MyMpn: a database for the systems biology model organism Mycoplasma pneumoniae." In: *Nucleic Acids Research* 43.Database issue (Jan. 2015), pp. D618–23. DOI: 10.1093/nar/gku1105. URL: http://dx.doi.org/10.1093/nar/gku1105 (visited on 06/03/2021).

[128] Mark D Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship." In: *Scientific data* 3 (Mar. 2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: http://www.nature.com/articles/sdata201618 (visited on 07/13/2018).

[129] Ana Hernández-Plaza et al. "eggNOG 6.0: enabling comparative genomics across 12 535 organisms." In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D389–D394. DOI: 10.1093/nar/gkac1022. URL: http://dx.doi.org/10.1093/nar/gkac1022 (visited on 05/14/2025).

[130] M Kanehisa and S Goto. "KEGG: Kyoto encyclopedia of genes and genomes." In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. DOI: 10.1093/nar/28.1.27. URL: http://dx.doi.org/10.1093/nar/28.1.27 (visited on 12/21/2020).

[131] Peter J Park. "ChIP-seq: advantages and challenges of a maturing technology." In: *Nature Reviews. Genetics* 10.10 (Oct. 2009), pp. 669–680. DOI: 10.1038/nrg2641. URL: http://dx.doi.org/10.1038/nrg2641 (visited on 05/14/2025).

[132] M Brenowitz, D F Senear, and R E Kingston. "DNase I footprint analysis of protein-DNA binding." In: *Current Protocols in Molecular Biology* Chapter 12 (May 2001), Unit 12.4. DOI: 10.1002/0471142727.mb1204s07. URL: http://dx.doi.org/10.1002/0471142727.mb1204s07 (visited on 05/14/2025).

[133] Ivan Junier et al. "Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium." In: *Cell Systems* 2.6 (June 2016), pp. 391–401. DOI: 10.1016/j.cels.2016.04.015. URL: http://dx.doi.org/10.1016/j.cels.2016.04.015 (visited on 05/15/2025).

[134] Jörg Stülke. "Control of transcription termination in bacteria by RNA-binding proteins that modulate RNA structures." In: *Archives of Microbiology* 177.6 (June 2002), pp. 433–440. DOI: 10.1007/s00203-002-0407-5. URL: http://dx.doi.org/10.1007/s00203-002-0407-5 (visited on 05/15/2025).

[135] Wei-Hua Chen et al. "Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances." In: *Nucleic Acids Research* 44.3 (Feb. 2016), pp. 1192–1202. DOI: 10.1093/nar/gkw004. URL: http://dx.doi.org/10.1093/nar/gkw004 (visited on 05/15/2025).

[136] Samuel Miravet-Verde et al. "FASTQINS and ANUBIS: two bioinformatic tools to explore facts and artifacts in transposon sequencing and essentiality studies." In: *Nucleic Acids Research* 48.17 (Sept. 2020), e102. DOI: 10.1093/nar/gkaa679. URL: http://dx.doi.org/10.1093/nar/gkaa679 (visited on 05/15/2025).

[137] Samuel Miravet-Verde et al. "ProTInSeq: transposon insertion tracking by ultra-deep DNA sequencing to identify translated large and small ORFs." In: *Nature Communications* 15.1 (Mar. 2024), p. 2091. DOI: 10.1038/s41467-024-46112-2. URL: http://dx.doi.org/10.1038/s41467-024-46112-2 (visited on 06/26/2025).

[138] Martín Muñoz-López and José L García-Pérez. "DNA transposons: nature and applications in genomics." In: *Current Genomics* 11.2 (Apr. 2010), pp. 115–128. DOI: 10.2174/138920210790886871. URL: http://dx.doi.org/10.2174/138920210790886871 (visited on 03/23/2022).

[139] David Sehnal et al. "Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures." In: *Nucleic Acids Research* 49.W1 (July 2021), W431–W437. DOI: 10.1093/nar/gkab314. URL: http://dx.doi.org/10.1093/nar/gkab314 (visited on 10/05/2023).

[140] Raul Burgos et al. "Protein quality control and regulated proteolysis in the genome-reduced organism Mycoplasma pneumoniae." In: *Molecular Systems Biology* 16.12 (Dec. 2020), e9530. DOI: 10.15252/msb.20209530. URL: http://dx.doi.org/10.15252/msb.20209530 (visited on 05/15/2025).

[141] R L Tatusov et al. "The COG database: a tool for genome-scale analysis of protein functions and evolution." In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 33–36. DOI: 10.1093/nar/28.1.33. URL: http://dx.doi.org/10.1093/nar/28.1.33 (visited on 05/15/2025).

[142] The Gene Ontology Consortium. "The Gene Ontology Resource: 20 years and still GOing strong." In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D330–D338. DOI: 10.1093/nar/gky1055. URL: http://dx.doi.org/10.1093/nar/gky1055 (visited on 01/27/2021).

[143] Elodie Drula et al. "The carbohydrate-active enzyme database: functions and literature." In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D571–D577. DOI: 10.1093/nar/gkab1045. URL: http://dx.doi.org/10.1093/nar/gkab1045 (visited on 05/15/2025).

[144] Winston Chang et al. "shiny: Web Application Framework for R". In: *The R Foundation* (2024). DOI: 10.32614/cran.package.shiny. URL: https://%7BCRAN%7D.R-project.org/package=shiny (visited on 05/15/2025).

[145] Hao-Bo Guo et al. "AlphaFold2 models indicate that protein sequence determines both structure and dynamics." In: *Scientific Reports* 12.1 (June 2022), p. 10696. ISSN: 2045-2322. DOI: 10.1038/s41598-022-14382-9. URL: https://www.nature.com/articles/s41598-022-14382-9 (visited on 05/15/2025).

[146] Andy M Lau et al. "Exploring structural diversity across the protein universe with The Encyclopedia of Domains." In: *Science* 386.6721 (Nov. 2024), eadq4946. ISSN: 0036-8075. DOI: 10.1126/science.adq4946. URL: https://www.science.org/doi/10.1126/science.adq4946 (visited on 05/15/2025).

[147] Jing Zhang et al. "DPAM: A domain parser for AlphaFold models." In: *Protein Science* 32.2 (Feb. 2023), e4548. DOI: 10.1002/pro.4548. URL: http://dx.doi.org/10.1002/pro.4548 (visited on 05/15/2025).

[148] Robbie P Joosten et al. "A series of PDB related databases for everyday needs." In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D411–9. DOI: 10.1093/nar/gkq1105. URL: http://dx.doi.org/10.1093/nar/gkq1105 (visited on 09/19/2019).

[149] Josephine Burgin et al. "The european nucleotide archive in 2022." In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D121–D125. DOI: 10.1093/nar/gkac1051. URL: http://dx.doi.org/10.1093/nar/gkac1051 (visited on 05/16/2025).

[150] Philip A Ewels et al. "The nf-core framework for community-curated bioinformatics pipelines." In: *Nature Biotechnology* 38.3 (Mar. 2020), pp. 276–278. ISSN: 1087-0156. DOI: 10.1038/s41587-020-0439-x. URL: http://www.nature.com/articles/s41587-020-0439-x (visited on 05/24/2021).

[151] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." In: *Genome Biology* 15.12 (2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8 (visited on 04/25/2016).

[152] Jeffrey T Leek. "svaseq: removing batch effects and other unwanted noise from sequencing data." In: *Nucleic Acids Research* 42.21 (Dec. 2014), e161. DOI: 10.1093/nar/gku864. URL: http://dx.doi.org/10.1093/nar/gku864 (visited on 04/29/2020).

[153] Yingdong Zhao et al. "TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository." In: *Journal of Translational Medicine* 19.1 (June 2021), p. 269. DOI: 10.1186/s12967-021-02936-w. URL: http://dx.doi.org/10.1186/s12967-021-02936-w (visited on 05/16/2025).

[154] Zachary A King et al. "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways." In: *PLoS Computational Biology* 11.8 (Aug. 2015), e1004321. DOI: 10.1371/journal.pcbi.1004321. URL: http://dx.doi.org/10.1371/journal.pcbi.1004321 (visited on 05/19/2025).

[155] G Rigaut et al. "A generic protein purification method for protein complex characterization and proteome exploration." In: *Nature Biotechnology* 17.10 (Oct. 1999), pp. 1030–1032. DOI: 10.1038/13732. URL: http://dx.doi.org/10.1038/13732 (visited on 05/19/2025).

[156] Ciyuan Peng et al. "Knowledge graphs: opportunities and challenges." In: *Artificial Intelligence Review* (Apr. 2023), pp. 1–32. DOI: 10.1007/s10462-023-10465-9. URL: http://dx.doi.org/10.1007/s10462-023-10465-9 (visited on 05/20/2025).

[157] Jonathan B L Bard and Seung Y Rhee. "Ontologies in biology: design, applications and future challenges." In: *Nature Reviews. Genetics* 5.3 (Mar. 2004), pp. 213–222. DOI: 10.1038/nrg1295. URL: http://dx.doi.org/10.1038/nrg1295 (visited on 07/04/2025).

[158] Eran Eden et al. "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists." In: *BMC Bioinformatics* 10 (Feb. 2009), p. 48. DOI: 10.1186/1471-2105-10-48. URL: http://dx.doi.org/10.1186/1471-2105-10-48 (visited on 05/12/2017).

[159] Tiffany J Callahan et al. "An open source knowledge graph ecosystem for the life sciences." In: *Scientific data* 11.1 (Apr. 2024), p. 363. DOI: 10.1038/s41597-024-03171-w. URL: http://dx.doi.org/10.1038/s41597-024-03171-w (visited on 05/20/2025).

[160] Benjamin J Stear et al. "Petagraph: A large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data." In: *Scientific data* 11.1 (Dec. 2024), p. 1338. DOI: 10.1038/s41597-024-04070-w. URL: http://dx.doi.org/10.1038/s41597-024-04070-w (visited on 05/20/2025).

[161] Yuan Zhang et al. "A comprehensive large-scale biomedical knowledge graph for AI-powered data-driven biomedical research". In: *Nature Machine Intelligence* (Mar. 2025). ISSN: 2522-5839. DOI: 10.1038/s42256-025-01014-w. URL: https://www.nature.com/articles/s42256-025-01014-w (visited on 05/20/2025).

[162] Emanuele Cavalleri et al. "An ontology-based knowledge graph for representing interactions involving RNA molecules." In: *Scientific data* 11.1 (Aug. 2024), p. 906. DOI: 10.1038/s41597-024-03673-7. URL: http://dx.doi.org/10.1038/s41597-024-03673-7 (visited on 05/20/2025).

[163] Alberto Santos et al. "A knowledge graph to interpret clinical proteomics data." In: *Nature Biotechnology* 40.5 (May 2022), pp. 692–702. ISSN: 1087-0156. DOI: 10.1038/s41587-021-01145-6. URL: https://www.nature.com/articles/s41587-021-01145-6 (visited on 09/12/2022).

[164] Jackson et al. "OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies". In: *Database* 2021 (Sept. 2021). URL: https://academic.oup.com/database/article/doi/10.1093/database/baab069/6410158 (visited on 05/20/2025).

[165] Axel Polleres. *SPARQL*. Springer New York, Jan. 2014, pp. 1960–1966. ISBN: 9781461461692. URL: https://link.springer.com/referenceworkentry/10.1007/978-1-4614-6170-8%5C_124 (visited on 05/20/2025).

[166] *Introducing the Knowledge Graph: things, not strings*. WEBSITE. May 2012. URL: https://blog.google/products/search/introducing-knowledge-graph-things-not/ (visited on 05/20/2025).

[167] Benjamin Sanchez-Lengeling et al. "A gentle introduction to graph neural networks". In: *Distill* 6.8 (Aug. 2021). ISSN: 2476-0757. DOI: 10.23915/distill.00033. URL: https://distill.pub/2021/gnn-intro (visited on 05/20/2025).

[168] Janani Durairaj et al. "Uncovering new families and folds in the natural protein universe." In: *Nature* 622.7983 (Oct. 2023), pp. 646–653. DOI: 10.1038/s41586-023-06622-3. URL: http://dx.doi.org/10.1038/s41586-023-06622-3 (visited on 07/04/2025).

[169] Carlos P Cantalapiedra et al. "eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale." In: *Molecular Biology and Evolution* 38.12 (Dec. 2021), pp. 5825–5829. DOI: 10.1093/molbev/msab293. URL: http://dx.doi.org/10.1093/molbev/msab293 (visited on 12/03/2021).

[170] Felix Teufel et al. "SignalP 6.0 predicts all five types of signal peptides using protein language models." In: *Nature Biotechnology* 40.7 (July 2022), pp. 1023–1025. ISSN: 1087-0156. DOI: 10.1038/s41587-021-01156-3. URL: https://www.nature.com/articles/s41587-021-01156-3 (visited on 05/20/2025).

[171] Martin Steinegger and Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets." In: *Nature Biotechnology* 35.11 (Nov. 2017), pp. 1026–1028. DOI: 10.1038/nbt.3988. URL: http://dx.doi.org/10.1038/nbt.3988 (visited on 07/25/2023).

[172] Cameron L.M. Gilchrist, Milot Mirdita, and Martin Steinegger. "Multiple Protein Structure Alignment at Scale with FoldMason". In: *BioRxiv* (Aug. 2024). DOI: 10.1101/2024.08.01.606130. URL: http://biorxiv.org/lookup/doi/10.1101/2024.08.01.606130 (visited on 05/21/2025).

[173] Felix Mölder et al. "Sustainable data analysis with Snakemake." In: *F1000Research* 10 (Jan. 2021), p. 33. ISSN: 2046-1402. DOI: 10.12688/f1000research.29032.2. URL: https://f1000research.com/articles/10-33/v1 (visited on 06/28/2025).

[174] Open AI. *Introducing ChatGPT |OpenAI*. WEBSITE. URL: https://openai.com/index/chatgpt/ (visited on 05/21/2025).

[175] Tom B. Brown et al. "Language models are few-shot learners". In: *arXiv* (2020). DOI: 10.48550/arxiv.2005.14165. URL: https://arxiv.org/abs/2005.14165 (visited on 11/15/2022).

[176] Jerome Goddard. "Hallucinations in chatgpt: A cautionary tale for biomedical researchers." In: *The American Journal of Medicine* 136.11 (Nov. 2023), pp. 1059–1060. DOI: 10.1016/j.amjmed.2023.06.012. URL: http://dx.doi.org/10.1016/j.amjmed.2023.06.012 (visited on 05/21/2025).

[177] Jonathan A Barker and Janet M Thornton. "An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis." In: *Bioinformatics* 19.13 (Sept. 2003), pp. 1644–1649. DOI: 10.1093/bioinformatics/btg226. URL: http://dx.doi.org/10.1093/bioinformatics/btg226 (visited on 07/04/2025).

[178] Ioannis G Riziotis and Janet M Thornton. "Capturing the geometry, function, and evolution of enzymes with 3D templates." In: *Protein Science* 31.7 (July 2022), e4363. ISSN: 0961-8368. DOI: 10.1002/pro.4363. URL: https://onlinelibrary.wiley.com/doi/10.1002/pro.4363 (visited on 07/04/2025).

[179] António J M Ribeiro et al. "Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites." In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D618–D623. DOI: 10.1093/nar/gkx1012. URL: http://dx.doi.org/10.1093/nar/gkx1012 (visited on 05/21/2025).

[180] R Dean Astumian. "Design principles for Brownian molecular machines: how to swim in molasses and walk in a hurricane." In: *Physical Chemistry Chemical Physics* 9.37 (Oct. 2007), pp. 5067–5083. DOI: 10.1039/b708995c. URL: http://dx.doi.org/10.1039/b708995c (visited on 06/02/2025).

[181] Bruce Alberts et al. *Molecular Biology of the Cell, Fourth Edition*. 4th ed. New York: Garland Science, Mar. 2002, p. 1616. ISBN: 0-8153-3218-1. (Visited on 07/04/2025).

[182] Kirill B Gromadski and Marina V Rodnina. "Kinetic determinants of high-fidelity tRNA discrimination on the ribosome." In: *Molecular Cell* 13.2 (Jan. 2004), pp. 191–200. DOI: 10.1016/s1097-2765(04)00005-x. URL: http://dx.doi.org/10.1016/s1097-2765(04)00005-x (visited on 08/26/2024).

[183] Kirill B Gromadski, Tina Daviter, and Marina V Rodnina. "A uniform response to mismatches in codon-anticodon complexes ensures ribosomal fidelity." In: *Molecular Cell* 21.3 (Feb. 2006), pp. 369–377. DOI: 10.1016/j.molcel.2005.12.018. URL: http://dx.doi.org/10.1016/j.molcel.2005.12.018 (visited on 08/26/2024).

[184] Ingo Wohlgemuth, Corinna Pohl, and Marina V Rodnina. "Optimization of speed and accuracy of decoding in translation." In: *The EMBO Journal* 29.21 (Nov. 2010), pp. 3701–3709. DOI: 10.1038/emboj.2010.229. URL: http://dx.doi.org/10.1038/emboj.2010.229 (visited on 08/26/2024).

[185] Riccardo Belardinelli et al. "Choreography of molecular movements during ribosome progression along mRNA." In: *Nature Structural & Molecular Biology* 23.4 (Apr. 2016), pp. 342–348. DOI: 10.1038/nsmb.3193. URL: http://dx.doi.org/10.1038/nsmb.3193 (visited on 05/04/2023).

[186] Niels Fischer et al. "Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy." In: *Nature* 466.7304 (July 2010), pp. 329–333. ISSN: 1476-4687. DOI: 10.1038/nature09206. URL: http://dx.doi.org/10.1038/nature09206 (visited on 06/13/2021).

[187] Ali Dashti et al. "Trajectories of the ribosome as a Brownian nanomachine." In: *Proceedings of the National Academy of Sciences of the United States of America* 111.49 (Dec. 2014), pp. 17492–17497. DOI: 10.1073/pnas.1419276111. URL: http://dx.doi.org/10.1073/pnas.1419276111 (visited on 05/16/2024).

[188] Liang Xue et al. "Structural insights into context-dependent inhibitory mechanisms of chloramphenicol in cells". In: *Nature Structural & Molecular Biology* (Dec. 2024). ISSN: 1545-9993. DOI: 10.1038/s41594-024-01441-0. URL: https://www.nature.com/articles/s41594-024-01441-0 (visited on 12/12/2024).

[189] Aaron Fluitt, Elsje Pienaar, and Hendrik Viljoen. "Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis." In: *Computational biology and chemistry* 31.5-6 (Oct. 2007), pp. 335–346. DOI: 10.1016/j.compbiolchem.2007.07.003. URL: http://dx.doi.org/10.1016/j.compbiolchem.2007.07.003 (visited on 05/17/2024).

[190] Sophia Rudorf et al. "Deducing the kinetics of protein synthesis in vivo from the transition rates measured in vitro." In: *PLoS Computational Biology* 10.10 (Oct. 2014), e1003909. DOI: 10.1371/journal.pcbi.1003909. URL: http://dx.doi.org/10.1371/journal.pcbi.1003909 (visited on 04/10/2023).

[191] Sophia Rudorf and Reinhard Lipowsky. "Protein Synthesis in E. coli: Dependence of Codon-Specific Elongation on tRNA Concentration and Codon Usage." In: *Plos One* 10.8 (Aug. 2015), e0134994. DOI: 10.1371/journal.pone.0134994. URL: http://dx.doi.org/10.1371/journal.pone.0134994 (visited on 04/10/2023).

[192] Annwesha Dutta, Gunter M Schütz, and Debashish Chowdhury. "Stochastic thermodynamics and modes of operation of a ribosome: A network theoretic perspective." In: *Physical review. E* 101.3-1 (Mar. 2020), p. 032402. DOI: 10.1103/{PhysRevE}.101.032402. URL: http://dx.doi.org/10.1103/%7BPhysRevE%7D.101.032402 (visited on 02/15/2024).

[193] Xiao-Xuan Shi, Hong Chen, and Ping Xie. "Dynamics of tRNA dissociation in early and later cycles of translation elongation by the ribosome." In: *Bio Systems* 172 (Oct. 2018), pp. 43–51. DOI: 10.1016/j.biosystems.2018.08.008. URL: http://dx.doi.org/10.1016/j.biosystems.2018.08.008 (visited on 05/01/2023).

[194] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. Second edition. |Boca Raton, Fla. : CRC Press, [2019]: Chapman and Hall/CRC, July 2019. ISBN: 9780429283321. DOI: 10.1201/9780429283321. URL: https://www.taylorfrancis.com/books/9781000001327 (visited on 04/06/2025).

[195] Steven H. Strogatz and Ronald F. Fox. "nonlinear dynamics and chaos: with applications to physics, biology, chemistry and engineering". In: *Physics today* 48.3 (Mar. 1995), pp. 93–94. ISSN: 0031-9228. DOI: 10.1063/1.2807947. URL: http://physicstoday.scitation.org/doi/10.1063/1.2807947 (visited on 04/06/2025).

[196] Thomas Engel, Gary Drobny, and Philip J. Reid. *Physical Chemistry for the Life Sciences*. illustrated. Pearson Prentice Hall, 2008. ISBN: 9780805382778. (Visited on 04/06/2025).

[197] Michael A. Sørensen and Steen Pedersen. "Absolute in vivo translation rates of individual codons in Escherichia coli". In: *Journal of Molecular Biology* 222.2 (Nov. 1991), pp. 265–280. ISSN: 00222836. DOI: 10.1016/0022-2836(91)90211-N. URL: https://linkinghub.elsevier.com/retrieve/pii/%7B002228369190211N%7D (visited on 04/06/2025).

[198] P Ván. "Nonequilibrium thermodynamics: emergent and fundamental." In: *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 378.2170 (May 2020), p. 20200066. DOI: 10.1098/rsta.2020.0066. URL: http://dx.doi.org/10.1098/rsta.2020.0066 (visited on 07/04/2025).

[199] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics, Jan. 2005. ISBN: 978-0-89871-572-9. DOI: 10.1137/1.9780898717921. URL: http://epubs.siam.org/doi/book/10.1137/1.9780898717921 (visited on 09/04/2024).

[200] H. Kunze, D. La Torre, and M. Ruiz Galán. "Optimization methods in inverse problems and applications to science and engineering". In: *Optimization and Engineering* 22.4 (Dec. 2021), pp. 2151–2158. ISSN: 1389-4420. DOI: 10.1007/s11081-021-09688-y. URL: https://link.springer.com/10.1007/s11081-021-09688-y (visited on 04/13/2025).

[201] Peter D. Hoff. *A First Course in Bayesian Statistical Methods* (*Springer Texts in Statistics*). 1st ed. 2009. London: Springer, June 2009, p. 280. ISBN: 978-0-387-92299-7. (Visited on 07/02/2025).

[202] Peter B Moore. "On the response of elongating ribosomes to forces opposing translocation." In: *Biophysical Journal* 123.18 (Sept. 2024), pp. 3010–3023. DOI: 10.1016/j.bpj.2024.05.032. URL: http://dx.doi.org/10.1016/j.bpj.2024.05.032 (visited on 08/26/2024).

[203] J. R. Norris. *Markov Chains*. Cambridge University Press, Feb. 1997. ISBN: 9780511810633. DOI: 10.1017/{CBO9780511810633}. URL: https://www.cambridge.org/core/product/identifier/9780511810633/type/book (visited on 04/13/2025).

[204] Samuel Karlin and Howard M. Taylor. *A first course in stochastic processes*. Elsevier, 1975. ISBN: 9780080570419. DOI: 10.1016/C2009-1-28569-8. URL: https://linkinghub.elsevier.com/retrieve/pii/C20091285698 (visited on 04/13/2025).

[205] Johan Paulsson. "Models of stochastic gene expression". In: *Physics of life reviews* 2.2 (June 2005), pp. 157–175. ISSN: 15710645. DOI: 10.1016/j.plrev.2005.03.003. URL: http://linkinghub.elsevier.com/retrieve/pii/S1571064505000138 (visited on 04/13/2025).

[206] Matthew C Gibson et al. "The emergence of geometric order in proliferating metazoan epithelia." In: *Nature* 442.7106 (Aug. 2006), pp. 1038–1041. ISSN: 1476-4687. DOI: 10.1038/nature05014. URL: http://dx.doi.org/10.1038/nature05014 (visited on 04/13/2025).

[207] Daniel T. Gillespie. "A rigorous derivation of the chemical master equation". In: *Physica A: Statistical Mechanics and its Applications* 188.1-3 (Sept. 1992), pp. 404–425. ISSN: 03784371. DOI: 10.1016/0378-4371(92)90283-V. URL: http://linkinghub.elsevier.com/retrieve/pii/%7B037843719290283V%7D (visited on 04/07/2025).

[208] Raul Toral. *Introduction to master equations*. WEBSITE. 2014. URL: https://ifisc.uib-csic.es/users/raul/%7BCURSOS%7D/%7BSP%7D/Introduction%5C_to%5C_master%5C_equations.pdf (visited on 04/07/2025).

[209] John C. Baez and Kenny Courser. "Coarse-Graining Open Markov Processes". In: *arXiv* (2017). DOI: 10.48550/arxiv.1710.11343. URL: https://arxiv.org/abs/1710.11343 (visited on 04/11/2025).

[210] Peter Buchholz. "Exact and ordinary lumpability in finite Markov chains". In: *Journal of applied probability* 31.1 (Mar. 1994), pp. 59–75. ISSN: 0021-9002. DOI: 10.2307/3215235. URL: https://www.cambridge.org/core/product/identifier/S0021900200107338/type/journal%5C_article (visited on 04/11/2025).

[211] Marina V Rodnina. "Translation in Prokaryotes." In: *Cold Spring Harbor Perspectives in Biology* 10.9 (Sept. 2018). DOI: 10.1101/cshperspect.a032664. URL: http://dx.doi.org/10.1101/cshperspect.a032664 (visited on 07/02/2025).

[212] Elmar Behrmann et al. "Structural snapshots of actively translating human ribosomes." In: *Cell* 161.4 (May 2015), pp. 845–857. DOI: 10.1016/j.cell.2015.03.052. URL: http://dx.doi.org/10.1016/j.cell.2015.03.052 (visited on 05/20/2015).

[213] Mainak Mustafi and James C Weisshaar. "Simultaneous Binding of Multiple EF-Tu Copies to Translating Ribosomes in Live Escherichia coli." In: *mBio* 9.1 (Jan. 2018). DOI: 10.1128/{mBio}.02143-17. URL: http://dx.doi.org/10.1128/%7BmBio%7D.02143-17 (visited on 04/13/2025).

[214] M V Rodnina et al. "Initial binding of the elongation factor Tu.GTP.aminoacyl-tRNA complex preceding codon recognition on the ribosome." In: *The Journal of Biological Chemistry* 271.2 (Jan. 1996), pp. 646–652. DOI: 10.1074/jbc.271.2.646. URL: http://dx.doi.org/10.1074/jbc.271.2.646 (visited on 04/13/2025).

[215] Dylan Girodat et al. "Geometric alignment of aminoacyl-tRNA relative to catalytic centers of the ribosome underpins accurate mRNA decoding." In: *Nature Communications* 14.1 (Sept. 2023), p. 5582. DOI: 10.1038/s41467-023-40404-9. URL: http://dx.doi.org/10.1038/s41467-023-40404-9 (visited on 08/25/2024).

[216] T Pape, W Wintermeyer, and M Rodnina. "Induced fit in initial selection and proofreading of aminoacyl-tRNA on the ribosome." In: *The EMBO Journal* 18.13 (July 1999), pp. 3800–3807. DOI: 10.1093/emboj/18.13.3800. URL: http://dx.doi.org/10.1093/emboj/18.13.3800 (visited on 08/25/2024).

[217] Ingo Wohlgemuth et al. "Evolutionary optimization of speed and accuracy of decoding on the ribosome." In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 366.1580 (Oct. 2011), pp. 2979–2986. DOI: 10.1098/rstb.2011.0138. URL: http://dx.doi.org/10.1098/rstb.2011.0138 (visited on 08/25/2024).

[218] J J Hopfield. "Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity." In: *Proceedings of the National Academy of Sciences of the United States of America* 71.10 (Oct. 1974), pp. 4135–4139. DOI: 10.1073/pnas.71.10.4135. URL: http://dx.doi.org/10.1073/pnas.71.10.4135 (visited on 08/25/2024).

[219] Hinrich Boeger. "Kinetic Proofreading." In: *Annual Review of Biochemistry* 91 (June 2022), pp. 423–447. DOI: 10.1146/annurev-biochem-040320-103630. URL: http://dx.doi.org/10.1146/annurev-biochem-040320-103630 (visited on 08/25/2024).

[220] Riccardo Belardinelli et al. "Translocation as continuous movement through the ribosome." In: *RNA Biology* 13.12 (Dec. 2016), pp. 1197–1203. DOI: 10.1080/15476286.2016.1240140. URL: http://dx.doi.org/10.1080/15476286.2016.1240140 (visited on 08/25/2024).

[221] Chunlai Chen et al. "Single-molecule fluorescence measurements of ribosomal translocation dynamics." In: *Molecular Cell* 42.3 (May 2011), pp. 367–377. DOI: 10.1016/j.molcel.2011.03.024. URL: http://dx.doi.org/10.1016/j.molcel.2011.03.024 (visited on 08/25/2024).

[222] Marina V Rodnina et al. "Converting GTP hydrolysis into motion: versatile translational elongation factor G." In: *Biological Chemistry* 401.1 (Dec. 2019), pp. 131–142. DOI: 10.1515/hsz-2019-0313. URL: http://dx.doi.org/10.1515/hsz-2019-0313 (visited on 08/25/2024).

[223] D N Wilson and K H Nierhaus. "The E-site story: the importance of maintaining two tRNAs on the ribosome during protein synthesis." In: *Cellular and Molecular Life Sciences* 63.23 (Dec. 2006), pp. 2725–2737. DOI: 10.1007/s00018-006-6125-4. URL: http://dx.doi.org/10.1007/s00018-006-6125-4 (visited on 04/23/2023).

[224] S T Liang et al. “mRNA composition and control of bacterial gene expression.” In: *Journal of Bacteriology* 182.11 (June 2000), pp. 3037–3044. DOI: 10.1128/{JB}.182.11.3037-3044.2000. URL: http://dx.doi.org/10.1128/%7BJB%7D.182.11.3037-3044.2000 (visited on 04/13/2025).

[225] T Dobzhansky. “Nothing in Biology Makes Sense except in the Light of Evolution.” In: *The American biology teacher* 35.3 (Mar. 1973), pp. 125–129. ISSN: 00027685. DOI: 10.2307/4444260. URL: http://abt.ucpress.edu/cgi/doi/10.2307/4444260 (visited on 08/11/2015).

[226] M A Martí-Renom et al. “Comparative protein structure modeling of genes and genomes.” In: *Annual review of biophysics and biomolecular structure* 29 (2000), pp. 291–325. DOI: 10.1146/annurev.biophys.29.1.291. URL: http://dx.doi.org/10.1146/annurev.biophys.29.1.291 (visited on 07/02/2025).

[227] J Rajendhran and P Gunasekaran. “Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond.” In: *Microbiological Research* 166.2 (Feb. 2011), pp. 99–110. DOI: 10.1016/j.micres.2010.02.003. URL: http://dx.doi.org/10.1016/j.micres.2010.02.003 (visited on 07/02/2025).

[228] J F Curran and M Yarus. “Rates of aminoacyl-tRNA selection at 29 sense codons in vivo.” In: *Journal of Molecular Biology* 209.1 (Sept. 1989), pp. 65–77. DOI: 10.1016/0022-2836(89)90170-8. URL: http://dx.doi.org/10.1016/0022-2836(89)90170-8 (visited on 04/13/2025).

[229] H Dong, L Nilsson, and C G Kurland. “Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates.” In: *Journal of Molecular Biology* 260.5 (Aug. 1996), pp. 649–663. DOI: 10.1006/jmbi.1996.0428. URL: http://dx.doi.org/10.1006/jmbi.1996.0428 (visited on 06/23/2025).

[230] Daniel N Wilson. “Ribosome-targeting antibiotics and mechanisms of bacterial resistance.” In: *Nature Reviews Microbiology* 12.1 (Jan. 2014), pp. 35–48. DOI: 10.1038/nrmicro3155. URL: http://dx.doi.org/10.1038/nrmicro3155 (visited on 06/28/2025).

[231] Tobias Maier et al. “Quantification of mRNA and protein and integration with protein turnover in a bacterium.” In: *Molecular Systems Biology* 7 (July 2011), p. 511. DOI: 10.1038/msb.2011.38. URL: http://dx.doi.org/10.1038/msb.2011.38 (visited on 05/17/2024).

[232] Emily B Kramer and Philip J Farabaugh. “The frequency of translational misreading errors in E. coli is largely determined by tRNA competition.” In: *RNA (New York)* 13.1 (Jan. 2007), pp. 87–96. DOI: 10.1261/rna.294907. URL: http://dx.doi.org/10.1261/rna.294907 (visited on 05/16/2024).

[233] Jeff Bezanson et al. "Julia: A fresh approach to numerical computing". In: *SIAM Review* 59.1 (Jan. 2017). ISSN: 0036-1445. DOI: 10.1137/141000671. URL: http://epubs.siam.org/doi/10.1137/141000671 (visited on 06/26/2022).

[234] jump_nlopt. *jump-dev/NLopt.jl: A Julia interface to the NLopt nonlinear-optimization library*. WEBSITE. URL: https://github.com/jump-dev/%7BNLopt%7D.jl (visited on 06/23/2025).

[235] stevengj. *stevengj/nlopt: library for nonlinear optimization, wrapping many algorithms for global and local, constrained or unconstrained, optimization*. WEBSITE. URL: https://github.com/stevengj/nlopt (visited on 06/23/2025).

[236] nlsolve. *JuliaNLSolvers/NLsolve.jl: Julia solvers for systems of nonlinear equations and mixed complementarity problems*. WEBSITE. URL: https://github.com/%7BJuliaNLSolvers%7D/%7BNLsolve%7D.jl (visited on 06/23/2025).

[237] Hani S Zaher and Rachel Green. "Fidelity at the molecular level: lessons from protein synthesis." In: *Cell* 136.4 (Feb. 2009), pp. 746–762. DOI: 10.1016/j.cell.2009.01.036. URL: http://dx.doi.org/10.1016/j.cell.2009.01.036 (visited on 06/20/2025).

[238] Alejandro F Villaverde et al. "A protocol for dynamic model calibration." In: *Briefings in Bioinformatics* 23.1 (Jan. 2022). DOI: 10.1093/bib/bbab387. URL: http://dx.doi.org/10.1093/bib/bbab387 (visited on 06/20/2025).

[239] Timothy M Errington et al. "Investigating the replicability of preclinical cancer biology." In: *eLife* 10 (Dec. 2021). ISSN: 2050-084X. DOI: 10.7554/{eLife}.71601. URL: https://elifesciences.org/articles/71601 (visited on 06/23/2025).

[240] Premal Shah and Michael A Gilchrist. "Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.25 (June 2011), pp. 10231–10236. DOI: 10.1073/pnas.1016719108. URL: http://dx.doi.org/10.1073/pnas.1016719108 (visited on 06/26/2025).

[241] Iskander Khusainov et al. "Bactericidal effect of tetracycline in E. coli strain ED1a may be associated with ribosome dysfunction." In: *Nature Communications* 15.1 (June 2024), p. 4783. DOI: 10.1038/s41467-024-49084-5. URL: http://dx.doi.org/10.1038/s41467-024-49084-5 (visited on 06/26/2025).

[242] Nathan M Belliveau et al. "Fundamental limits on the rate of bacterial growth and their influence on proteomic composition." In: *Cell Systems* 12.9 (Sept. 2021), 924–944.e2. ISSN: 24054712. DOI: 10.1016/j.cels.2021.06.002. URL: https://linkinghub.elsevier.com/retrieve/pii/%7BS240547122100209X%7D (visited on 06/27/2025).

[243] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis (6th Edition)*. 6th ed. Upper Saddle River, N.J: Pearson, Apr. 2007, p. 808. ISBN: 9780131877153. (Visited on 06/28/2025).

[244] F H Crick. "Codon–anticodon pairing: the wobble hypothesis." In: *Journal of Molecular Biology* 19.2 (Aug. 1966), pp. 548–555. ISSN: 00222836. DOI: 10.1016/S0022-2836(66)80022-0. URL: http://linkinghub.elsevier.com/retrieve/pii/S0022283666800220 (visited on 06/30/2025).

[245] James Marks et al. "Context-specific inhibition of translation by ribosomal antibiotics targeting the peptidyl transferase center." In: *Proceedings of the National Academy of Sciences of the United States of America* 113.43 (Oct. 2016), pp. 12150–12155. DOI: 10.1073/pnas.1613055113. URL: http://dx.doi.org/10.1073/pnas.1613055113 (visited on 06/30/2025).

[246] Salzer et al. "Role of (p)ppGpp in antibiotic resistance, tolerance, persistence and survival in Firmicutes". In: *microLife* 4 (Jan. 2023). URL: https://academic.oup.com/microlife/article/doi/10.1093/femsml/uqad009/7076323?login=true (visited on 06/30/2025).

[247] Xiaowei Yan et al. "Dynamics of Translation of Single mRNA Molecules In Vivo." In: *Cell* 165.4 (May 2016), pp. 976–989. DOI: 10.1016/j.cell.2016.04.034. URL: http://dx.doi.org/10.1016/j.cell.2016.04.034 (visited on 07/01/2025).

[248] Maximilian F Madern et al. "Long-term imaging of individual ribosomes reveals ribosome cooperativity in mRNA translation." In: *Cell* 188.7 (Apr. 2025), 1896–1911.e24. DOI: 10.1016/j.cell.2025.01.016. URL: http://dx.doi.org/10.1016/j.cell.2025.01.016 (visited on 07/01/2025).

[249] Lars V Bock et al. "Simulation of complex biomolecular systems: the ribosome challenge." In: *Annual review of biophysics* (Jan. 2022). DOI: 10.1146/annurev-biophys-111622-091147. URL: http://dx.doi.org/10.1146/annurev-biophys-111622-091147 (visited on 07/01/2025).

[250] Rodney Tollerson and Michael Ibba. "Translational regulation of environmental adaptation in bacteria." In: *The Journal of Biological Chemistry* 295.30 (July 2020), pp. 10434–10445. DOI: 10.1074/jbc.{REV120}.012742. URL: http://dx.doi.org/10.1074/jbc.%7BREV120%7D.012742 (visited on 07/01/2025).

[251] Jan-Hendrik Trösemeier et al. "Optimizing the dynamics of protein expression." In: *Scientific Reports* 9.1 (May 2019), p. 7511. DOI: 10.1038/s41598-019-43857-5. URL: http://dx.doi.org/10.1038/s41598-019-43857-5 (visited on 07/01/2025).

[252] Joel D Richter and Jeff Coller. "Pausing on polyribosomes: make way for elongation in translational control." In: *Cell* 163.2 (Oct. 2015), pp. 292–300. DOI: 10.1016/j.cell.2015.09.041. URL: http://dx.doi.org/10.1016/j.cell.2015.09.041 (visited on 07/01/2025).

[253] Chunlai Chen et al. "Allosteric vs. spontaneous exit-site (E-site) tRNA dissociation early in protein synthesis." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.41 (Oct. 2011), pp. 16980–16985. DOI: 10.1073/pnas.1106999108. URL: http://dx.doi.org/10.1073/pnas.1106999108 (visited on 08/25/2024).

[254] Albert-László Barabási. *Network Science*. 1st ed. Cambridge, United Kingdom: Cambridge University Press, Aug. 2016, p. 475. ISBN: 1107076269. (Visited on 07/01/2025).

[255] Colman O'Cathail et al. "The european nucleotide archive in 2024." In: *Nucleic Acids Research* 53.D1 (Jan. 2025), pp. D49–D55. DOI: 10.1093/nar/gkae975. URL: http://dx.doi.org/10.1093/nar/gkae975 (visited on 07/01/2025).

[256] Zhe Zhao, Federico Marotta, and Min Wu. "Thanos: An R Package for the Gene-Centric Analysis of Functional Potential in Metagenomic Samples." In: *Microorganisms* 12.7 (June 2024). DOI: 10.3390/microorganisms12071264. URL: http://dx.doi.org/10.3390/microorganisms12071264 (visited on 07/03/2025).

[257] Pablo Meyer and Julio Saez-Rodriguez. "Advances in systems biology modeling: 10 years of crowdsourcing DREAM challenges." In: *Cell Systems* 12.6 (June 2021), pp. 636–653. ISSN: 24054712. DOI: 10.1016/j.cels.2021.05.015. URL: https://linkinghub.elsevier.com/retrieve/pii/S2405471221002015 (visited on 07/01/2025).

[258] Deepak Bhatnagar, Handrean Soran, and Paul N Durrington. "Hypercholesterolaemia and its management." In: *BMJ (Clinical Research Ed.)* 337 (Aug. 2008), a993. DOI: 10.1136/bmj.a993. URL: http://dx.doi.org/10.1136/bmj.a993 (visited on 08/19/2024).

[259] Paul Durrington. "Dyslipidaemia." In: *The Lancet* 362.9385 (Aug. 2003), pp. 717–731. DOI: 10.1016/S0140-6736(03)14234-1. URL: http://dx.doi.org/10.1016/S0140-6736(03)14234-1 (visited on 08/19/2024).

[260] Sabina O Beheshti et al. "Worldwide Prevalence of Familial Hypercholesterolemia: Meta-Analyses of 11 Million Subjects." In: *Journal of the American College of Cardiology* 75.20 (May 2020), pp. 2553–2566. ISSN: 07351097. DOI: 10.1016/j.jacc.2020.03.057. URL: https://linkinghub.elsevier.com/retrieve/pii/S0735109720347501 (visited on 08/19/2024).

[261] Samuel A Lambert et al. "The Polygenic Score Catalog: new functionality and tools to enable FAIR research." In: *medRxiv* (May 2024). DOI: 10.1101/2024.05.29.24307783. URL: http://medrxiv.org/lookup/doi/10.1101/2024.05.29.24307783 (visited on 08/19/2024).

[262] Christopher C Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. In: *GigaScience* 4.1 (Feb. 2015), s13742–015–0047–8. DOI: 10.1186/s13742-015-0047-8. URL: http://dx.doi.org/10.1186/s13742-015-0047-8 (visited on 06/16/2020).

[263] Laurens F Reeskamp et al. “Differential DNA methylation in familial hypercholesterolemia.” In: *EBioMedicine* 61 (Nov. 2020), p. 103079. DOI: 10.1016/j.ebiom.2020.103079. URL: http://dx.doi.org/10.1016/j.ebiom.2020.103079 (visited on 08/19/2024).

A Additional figures and tables

Table A.1: List of species and genomes included in the phylogenetic trees displayed throughout chapter 2.

Species name	RefSeq genome ID [104]
Bacillus subtilis subsp. subtilis str. 168	GCF_000009045.1
Candidatus Mycoplasma haemobos	GCF_001645765.1
Candidatus Mycoplasma haemohominis	GCF_902712995.1
Candidatus Mycoplasma haemolamae	GCF_000281235.1
Candidatus Mycoplasma haemominutum	GCF_000319365.1
Escherichia coli strain K-12 substr. MG1655	GCF_000005845.2
Lactococcus cremoris	GCF_000014545.1
Malacoplasma iowae	GCF_900660615.1
Malacoplasma penetrans	GCF_000011225.1
Mycoplasma haemocanis	GCF_000238995.1
Mycoplasma haemofelis	GCF_000200735.1
Mycoplasma ovis	GCF_000508245.1
Mycoplasma parvum	GCF_000477415.1
Mycoplasma suis	GCF_000179035.2
Mycoplasma testudinis	GCF_000687795.1
Mycoplasmatota	GCF_000518305.1
Mycoplasma tullyi	GCF_014068355.1
Mycoplasma wenyonii	GCF_000277795.1
Mycoplasmoides alvi	GCF_000701785.1
Mycoplasmoides gallisepticum	GCF_900476085.1
Mycoplasmoides genitalium	GCF_000027325.1
Mycoplasmoides pirum	GCF_000685905.1
Mycoplasmoides pneumoniae	GCF_900660465.1
Streptococcus pneumoniae	GCF_000007045.1
Ureaplasma canigenitalium	GCF_000712165.1
Ureaplasma diversum	GCF_000731915.1
Ureaplasma parvum	GCF_000019345.1
Ureaplasma urealyticum	GCF_000169535.1

Table A.2: Parameter values for the two main systems used in the thesis, *E. coli* growing at 0.7 doublings/hour and *M. pneumoniae* growing in rich medium. NA means not available (The tRNA species are different in these two organisms). The values come from multiple references [191, 185, 19, 22, 20, 137]

	E. coli, growth rate 0.7 dbl/h	M. pneumoniae, control
General concentrations (μM)		
Ribosome concentration	19	7
EF-Tu concentration	150	100
EF-G concentration	10	10
General rate constants (1/(μM s))		
κ_on	94	94
κ_G	440	440
ω_re	100	100
ω_dis	0.01	0.01
κ_ass	1	1
Codon usage frequency		
AAA	0.047	0.044
AAC	0.028	0.038
AAG	0.013	0.043
AAU	0.0089	0.024
ACA	0.0033	0.0097
ACC	0.027	0.024
ACG	0.007	0.0078
ACU	0.015	0.019
AGA	0.001	0.0052
AGC	0.012	0.0095
AGG	0.0001	0.003
AGU	0.0036	0.019
AUA	0.0009	0.0045
AUC	0.038	0.015
AUG	0.022	0.016
AUU	0.021	0.045
CAA	0.0097	0.034
CAC	0.014	0.012
CAG	0.029	0.014
CAU	0.0088	0.0052
CCA	0.0065	0.013

CCC	0.0028	0.0085
CCG	0.029	0.0073
CCU	0.0049	0.0086
CGA	0.0012	0.0023
CGC	0.022	0.012
CGG	0.0015	0.0052
CGU	0.034	0.012
CUA	0.0019	0.011
CUC	0.0059	0.013
CUG	0.061	0.0083
CUU	0.0053	0.0096
GAA	0.054	0.043
GAC	0.03	0.021
GAG	0.017	0.015
GAU	0.024	0.029
GCA	0.022	0.017
GCC	0.019	0.016
GCG	0.03	0.013
GCU	0.03	0.028
GGA	0.0025	0.0061
GGC	0.036	0.011
GGG	0.0043	0.0098
GGU	0.039	0.031
GUA	0.017	0.016
GUC	0.011	0.011
GUG	0.02	0.021
GUU	0.033	0.022
UAC	0.016	0.018
UAU	0.01	0.013
UCA	0.0036	0.008
UCC	0.012	0.01
UCG	0.0054	0.0068
UCU	0.014	0.0073
UGC	0.0051	0.0017
UGG	0.0093	0.0057
UGU	0.004	0.0051
UUA	0.0055	0.035
UUC	0.023	0.013
UUG	0.0063	0.02
UUU	0.012	0.037

UGA	NA	0.0048
Total tRNA concentration (μM)		
Ala1B	12	NA
Ala2	2.1	NA
Arg2	15	NA
Arg3	2.6	NA
Arg4	2.4	NA
Arg5	1.6	NA
Asn	3.9	NA
Asp1	8.1	NA
Cys	4.9	NA
Gln1	2.7	NA
Gln2	3.1	NA
Glu2	16	NA
Gly1	2.9	NA
Gly2	4.3	NA
Gly3	15	NA
His	2.2	NA
Ile1	11	NA
Ile2	0.56	NA
Leu1	15	NA
Leu2	3.5	NA
Leu3	2.5	NA
Leu4	6.3	NA
Leu5	3.5	NA
Lys	6.8	NA
Met_m	2.6	NA
Phe	3.6	NA
Pro1	2.4	NA
Pro2	2.5	NA
Pro3	1.9	NA
Sec	0.86	NA
Ser1	5.6	NA
Ser2	1	NA
Ser3	4.4	NA
Ser5	2.6	NA
Thr1	0.41	NA
Thr2	2	NA
Thr3	3.7	NA

Thr4	3.2	NA
Trp	2.8	NA
Tyr1	2.4	NA
Tyr2	3.9	NA
Val1	12	NA
Val2	4.4	NA
MPNt01	NA	8.9
MPNt02	NA	0.92
MPNt03	NA	2.2
MPNt04	NA	6.5
MPNt05	NA	1.1
MPNt06	NA	8.3
MPNt07	NA	0.66
MPNt08	NA	0.19
MPNt09	NA	0.11
MPNt11	NA	3.3
MPNt12	NA	3.9
MPNt13	NA	0.57
MPNt14	NA	2.2
MPNt15	NA	0.46
MPNt16	NA	0.36
MPNt17	NA	0.17
MPNt18	NA	15
MPNt19	NA	0.27
MPNt20	NA	8.4
MPNt21	NA	8.1
MPNt22	NA	2.1
MPNt23	NA	9.3
MPNt24	NA	0.24
MPNt25	NA	0.53
MPNt27	NA	0.92
MPNt28	NA	3.4
MPNt29	NA	3
MPNt30	NA	1.1
MPNt31	NA	0.63
MPNt32	NA	1.9
MPNt33	NA	1
MPNt34	NA	0.56
MPNt35	NA	3.5
MPNt36	NA	8.7

MPNt37	NA	1.7
Cognacy		
Ala1B	GCA,GCG,GCU	NA
Ala2	GCC	NA
Arg2	CGA,CGC,CGU	NA
Arg3	CGG	NA
Arg4	AGA	NA
Arg5	AGG	NA
Asn	AAC,AAU	NA
Asp1	GAC,GAU	NA
Cys	UGC,UGU	NA
Gln1	CAA	NA
Gln2	CAG	NA
Glu2	GAA,GAG	NA
Gly1	GGG	NA
Gly2	GGA,GGG	NA
Gly3	GGC,GGU	NA
His	CAC,CAU	NA
Ile1	AUC,AUU	NA
Ile2	AUA	NA
Leu1	CUG	NA
Leu2	CUC,CUU	NA
Leu3	CUA,CUG	NA
Leu4	UUG	NA
Leu5	UUA,UUG	NA
Lys	AAA,AAG	NA
Met_m	AUG	NA
Phe	UUC,UUU	NA
Pro1	CCG	NA
Pro2	CCC,CCU	NA
Pro3	CCA,CCG,CCU	NA
Sec		NA
Ser1	UCA,UCG,UCU	NA
Ser2	UCG	NA
Ser3	AGC,AGU	NA
Ser5	UCC,UCU	NA
Thr1	ACC,ACU	NA
Thr2	ACG	NA
Thr3	ACC,ACU	NA

Thr4	ACA,ACG,ACU	NA
Trp	UGG	NA
Tyr1	UAC,UAU	NA
Tyr2	UAC,UAU	NA
Val1	GUA,GUG,GUU	NA
Val2	GUC,GUU	NA
MPNt01	NA	GCA,GCC,GCG,GCU
MPNt02	NA	AUA,AUC,AUU
MPNt03	NA	AGC,AGU
MPNt04	NA	ACA,ACC,ACG,ACU
MPNt05	NA	UGC,UGU
MPNt06	NA	CCA,CCC,CCG,CCU
MPNt07	NA	AUG
MPNt08	NA	AUA,AUC,AUU
MPNt09	NA	UCA,UCC,UCG,UCU
MPNt11	NA	GAC,GAU
MPNt12	NA	UUC,UUU
MPNt13	NA	CGA,CGC,CGG,CGU
MPNt14	NA	GGA,GGC,GGG,GGU
MPNt15	NA	AGA,AGG
MPNt16	NA	UGA,UGG
MPNt17	NA	CGA,CGC,CGG,CGU
MPNt18	NA	GGA,GGC,GGG,GGU
MPNt19	NA	UUA,UUG
MPNt20	NA	AAA,AAG
MPNt21	NA	CAA,CAG
MPNt22	NA	UAC,UAU
MPNt23	NA	UGA,UGG
MPNt24	NA	UCA,UCC,UCG,UCU
MPNt25	NA	UCA,UCC,UCG,UCU
MPNt27	NA	CUA,CUC,CUG,CUU
MPNt28	NA	AAA,AAG
MPNt29	NA	ACA,ACC,ACG,ACU
MPNt30	NA	GUA,GUC,GUG,GUU
MPNt31	NA	ACA,ACC,ACG,ACU
MPNt32	NA	GAA,GAG
MPNt33	NA	AAC,AAU
MPNt34	NA	CAC,CAU
MPNt35	NA	CUA,CUC,CUG,CUU
MPNt36	NA	UUA,UUG

MPNt37	NA	AGA,AGG
Reference rates (1/s)		
ω off	700	2100
ω rec	1500	3100
ω 2,1	2	2
ω 2,3	1500	1600
ω con	450	510
ω 4,0	1	0.98
ω Tdis_co*	250	330
ω ETdis_co*	250	390
ω pept_co	1000	1100
ω Epept_co	1000	1200
ω 7,6	1100	4000
ω 7,8	7	4.3
ω 9,0	4	6.4
ω Tdis_nr*	0.26	0.27
ω ETdis_nr*	0.26	0.27
ω pept_nr	1000	1000
ω Epept_nr	1000	1000
κ G	810	820
ω B,A	980	970
ω Accw	930	930
ω Acw	160	160
ω Bccw	4600	4700
ω B,C	1600	1700
ω C,D	810	930
ω Gdis	250	360
ω EC	93	91

Table A.3: Estimated *in vivo* rates and free ternary complex concentrations for the two main systems used in the thesis, *E. coli* growing at 0.7 doublings/hour and *M. pneumoniae* growing in rich medium. NA means not available (The tRNA species are different in these organisms).

	E. coli, growth rate 0.7 dbl/h	M. pneumoniae, control
Estimated rates (1/s)		
ω off	2100	5600
ω rec	3100	7500
ω 2,1	2	2
ω 2,3	1600	1600
ω con	510	480
ω 4,0	0.98	0.96
ω Tdis_co*	330	390
ω ETdis_co*	390	110
ω pept_co	1100	1600
ω Epept_co	1200	2400
ω 7,6	4000	15000
ω 7,8	4.3	2.6
ω 9,0	6.4	11
ω Tdis_nr*	0.27	0.17
ω ETdis_nr*	0.27	0.23
ω pept_nr	1000	1000
ω Epept_nr	1000	1000
κ G	820	97
ω B,A	970	11000
ω Accw	930	110
ω Acw	160	4800
ω Bccw	4700	890
ω B,C	1700	340
ω C,D	930	120
ω Gdis	360	39
ω EC	91	17
Estimated free ternary complex concentration (μM)		
Ala1B	5.60	NA
Ala2	0.80	NA
Arg2	9.70	NA
Arg3	2.30	NA
Arg4	2.10	NA

Arg5	1.40	NA
Asn	1.20	NA
Asp1	4.10	NA
Cys	3.90	NA
Gln1	1.90	NA
Gln2	1.00	NA
Glu2	9.80	NA
Gly1	2.50	NA
Gly2	3.60	NA
Gly3	9.20	NA
His	0.60	NA
Ile1	6.60	NA
Ile2	0.45	NA
Leu1	10.00	NA
Leu2	2.40	NA
Leu3	1.60	NA
Leu4	5.50	NA
Leu5	2.70	NA
Lys	2.50	NA
Met_m	1.00	NA
Phe	1.20	NA
Pro1	1.10	NA
Pro2	1.80	NA
Pro3	0.60	NA
Sec	0.79	NA
Ser1	3.90	NA
Ser2	0.89	NA
Ser3	3.10	NA
Ser5	1.40	NA
Thr1	0.17	NA
Thr2	1.60	NA
Thr3	1.50	NA
Thr4	1.90	NA
Trp	2.00	NA
Tyr1	1.50	NA
Tyr2	2.50	NA
Val1	7.20	NA
Val2	2.80	NA
MPNt01	NA	7.60
MPNt02	NA	0.13

MPNt03	NA	1.70
MPNt04	NA	5.90
MPNt05	NA	0.94
MPNt06	NA	7.60
MPNt07	NA	0.41
MPNt08	NA	0.028
MPNt09	NA	0.049
MPNt11	NA	2.50
MPNt12	NA	3.20
MPNt13	NA	0.21
MPNt14	NA	2.00
MPNt15	NA	0.43
MPNt16	NA	0.35
MPNt17	NA	0.063
MPNt18	NA	14.00
MPNt19	NA	0.24
MPNt20	NA	7.30
MPNt21	NA	7.20
MPNt22	NA	1.60
MPNt23	NA	9.00
MPNt24	NA	0.11
MPNt25	NA	0.24
MPNt27	NA	0.77
MPNt28	NA	3.00
MPNt29	NA	2.70
MPNt30	NA	0.032
MPNt31	NA	0.56
MPNt32	NA	1.10
MPNt33	NA	0.14
MPNt34	NA	0.29
MPNt35	NA	3.00
MPNt36	NA	7.80
MPNt37	NA	1.60

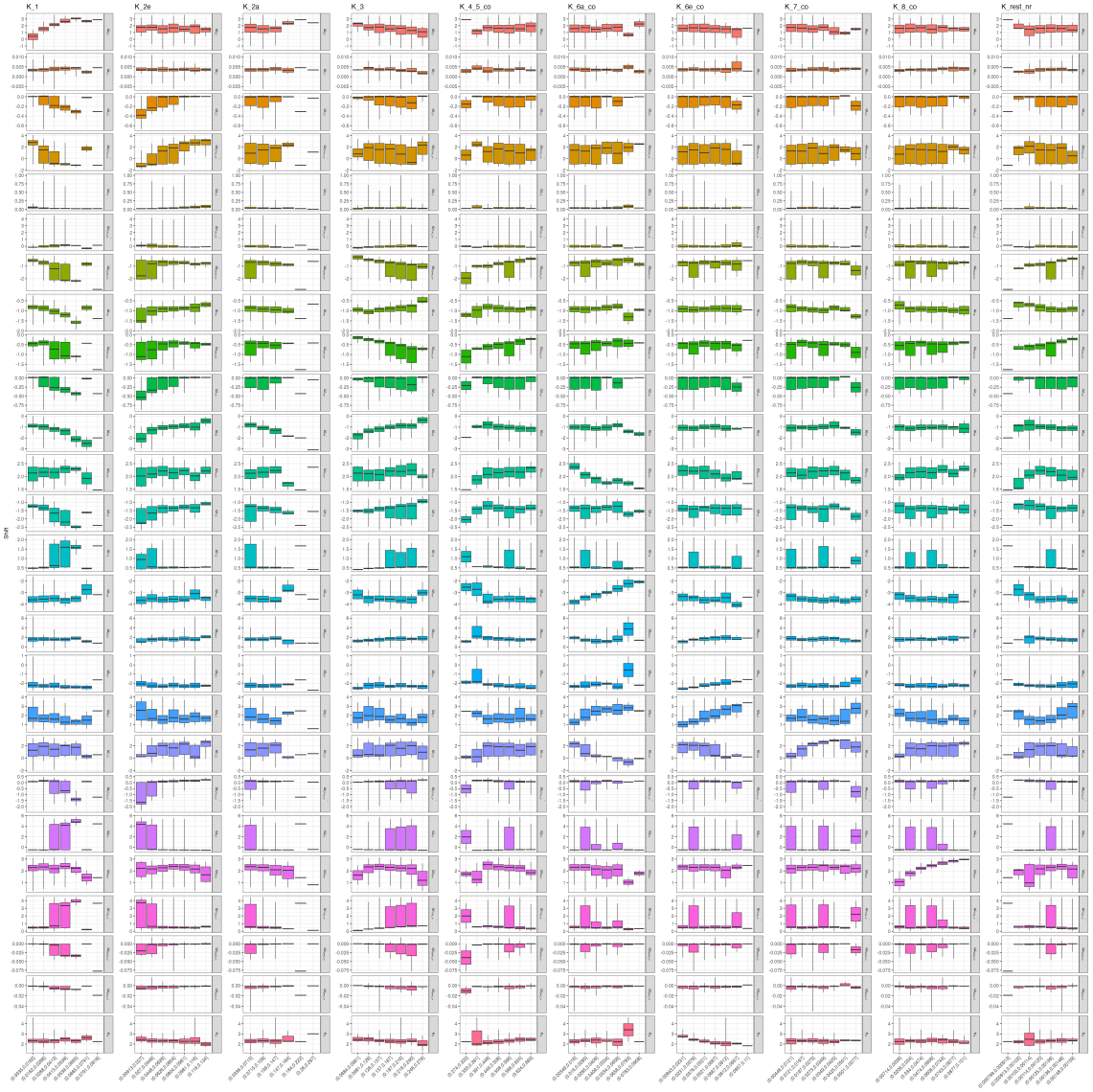


Figure A.1: For each intermediate state from Xue et al. [188], I binned the steady-state proportion across cells and grouped the single-barrier shifts for each rate arising from cells within each bin. Each column is an intermediate, with the x-axis showing the occupancy bins. Each row is a rate, with the y-axis showing the single-barrier shift.

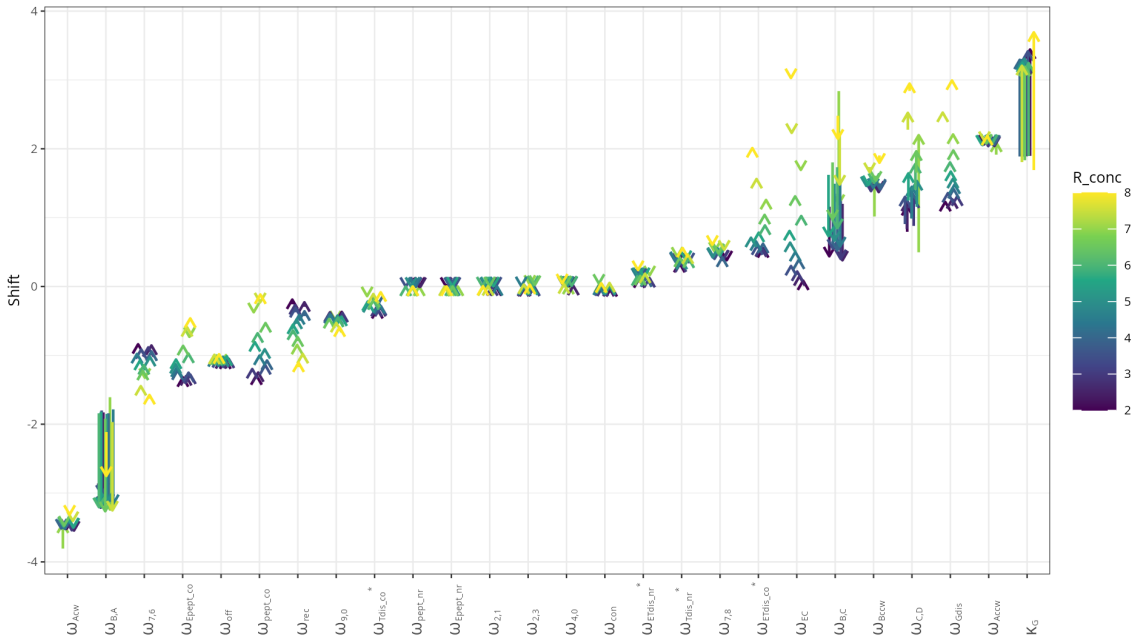


Figure A.2: Single-barrier shift change as a function of ribosome concentration (arrow color) and EF-G concentration (arrow length).

B The PEGS DREAM Challenge

This chapter is dedicated to a project that I carried out together with Alessandro Lussana, PhD student at EMBL-EBI. Because it is thematically different from the rest of the thesis, it is only reported here in the appendix. Briefly, DREAM Challenges (Dialogue for Reverse Engineering Assessment and Methods) is a non-profit initiative for advancing biomedical and systems biology research via crowd-sourced competitions [257]. The PEGS DREAM Challenge (`https://www.synapse.org/Synapse:syn52817032/wiki/624336`) leveraged data from the Personalized Environment and Genes Study (PEGS) sponsored by the National Institute of Environmental Health Sciences (NIEHS), which is part of the National Institutes of Health (NIH). Health, exposure, geospatial, and genomic data are available for the PEGS cohort. The motivations behind the PEGS DREAM challenge include understanding disease risk factors, improving disease classification, and promoting method development. I was nerd-sniped into joining this challenge, and we ended up winning second place among all competing teams.

Background

Hypercholesterolemia, characterized by elevated levels of LDL cholesterol in the blood, poses significant risks for cardiovascular diseases such as heart attacks and strokes [258]. As a complex trait influenced by genetics, diet, and lifestyle, its risk factors are challenging to predict, but large, multimodal data sets provide a unique opportunity to improve disease classification. Our approach began with a simple model based on the Health and Exposure (H&E) Survey data, which we focused on first because, unlike other data modalities, they are available for the whole PEGS cohort, providing the largest possible sample size. Recognizing that hypercholesterolemia is often linked to other diseases such as diabetes and hypertension [259], we hypothesized that the H&E profile of an individual, which provides a glimpse into their overall current health, is likely correlated with high cholesterol. This first model performed remarkably well for its simplicity, so we decided to complement it with genetic data as secondary predictors. Single Nucleotide Variants (SNVs) are not only informative for familiar hypercholesterolemia (caused by mutations in the LDLR, APOB, and PCSK9 genes) [260], but they also offer insights into the polygenic nature of hypercholesterolemia in general. By leveraging existing polygenic risk scores, we effectively borrowed publicly available data from a much larger sample size, thus enhancing the accuracy of our estimate. We employed a random forest classifier to generate an initial

Table B.1: Variables removed from the training data set and corresponding reason why.

Removed variable	Reason
he_flag8843	Survey metadata
he_phase	Survey metadata
he_version	Survey metadata
he_bmi_derived	Redundant

risk score based on the H&E data, taking advantage of its inherent capability to perform feature selection. For individuals with available genetic data, we calculated 12 Polygenic Risk Scores using weights downloaded from the PGS Catalog, and combined them with the baseline random forest score through logistic regression. To the best of our knowledge, the logistic regression approach to combine the random forest score with multiple Polygenic Risk Scores is novel and shows high potential for success in clinical applications.

Methods

Our model makes use of two components from the PEGS data: the Health and Exposure (H&E) Survey, and the Single Nucleotide Variants (SNVs) genetic data. Below, we describe the training and prediction procedures.

Training phase

We processed the H&E Survey data as follows. Variables of class "character" and "factor" were deemed too sparse and removed. Binary variables denoting whether another question was answered, identified by the "RESPPROV" label in the "sas_format" field of the metadata table, were removed as not informative. Additionally, all the variables in table B.1 were removed as not informative or redundant. The remaining variables of class "numeric" were imported as such, those of class "binary" were parsed as categorical, and those of class "ordered factor" were treated as ordered categorical. All missing, skipped, nullified, unknown, or not applicable answers were left as missing values, and variables with at least 5% of missing data were eliminated. This left us with 3062 observations of 239 variables, identified by the EPR number.

The response variable for this classification task was "he_b008_high_cholesterol", described as "ever diagnosed with high cholesterol". 33 individuals were missing this information, and were removed altogether from the training data set. Missing data in the other fields were imputed as the column median (for numeric variables) or mode (for categorical). After preprocessing, we trained a random forest soft classifier to predict the probability of high cholesterol. We also extracted the fitted probabilities for the training individuals, as they would be needed in a subsequent training step.

In order to make use of the genetic data, we downloaded 12 sets of weights from the Polygenic Score Catalog [261] that were associated with hypercholesterolemia (listed in table B.2). The weights we downloaded were already harmonized and lifted to the GRCh38 assembly of the human genome, the same as the genetic data provided by the PEGS. We further processed them to remove duplicated records and retain only three columns: SNV ID, effect allele, and weight. For the ID field, if the variants already had an rsID from dbSNP, it was used as-is; if not, we generated an ID in the form `<chr_name>:<chr_position>`, the same convention used for the PEGS SNV data. We copied the processed weights, together with the plink2 v5.12 binary [262], as static assets in the Docker container that we submitted for the challenge.

Table B.2: IDs of the polygenic score weights downloaded from the PGS Catalog.

Polygenic Score
PGS000936
PGS002334
PGS002406
PGS002455
PGS002504
PGS002553
PGS002602
PGS002651
PGS002700
PGS002764
PGS004783
PGS004784

As the PEGS genetic data underwent a thorough QC and have been obtained at 30x sequencing depth, we did not perform additional filtering or imputations on the SNV data. We directly ran plink2 to calculate 12 polygenic risk scores for the full training cohort, translating all dosages to mean zero using the ”center” modifier and using the allele frequencies computed from the data themselves. The individuals with available genetic data are only a subset of those who participated in the H&E Survey, leaving us with a PGS data set of 1515 samples and 12 variables. To this data set, we added as an additional variable the fitted disease probabilities obtained earlier from the random forest model, matching the EPR numbers. We then trained a logistic regression classifier on these 13 scores using again the ”he_b008_high_cholesterol” variable as the response. In principle, the PGS can be used directly as a measure of disease risk. However, this additional step brought several benefits: * Detecting which combination of the 12 polygenic scores works better and assigning it more weight in the final prediction; * Combining the PGS-based predictions with the H&E-based prediction while regularizing the latter; * Mitigating potential errors due to the wrong effect allele being reported, which can result in scores with the opposite sign.

Prediction phase

Having trained the random forest and the logistic classifiers, we went ahead to process the validation data set. The H&E Survey data were processed as for the training cohort, except that columns with missing data were not removed to avoid filtering out some of the variables that were already included in the model. We applied the random forest classifier and predicted the baseline probability scores for the validation cohort.

For individuals with available genetic data, we calculated the polygenic risk scores with plink2 as described above, and recreated a data set that combines the PGS with the random forest predicted probabilities. We used our logistic classifier to predict a refined probability score for these individuals. This refined score became the final disease probability for individuals with genetic data, while the baseline score was used for the other individuals.

Conclusion

Our initial strategy was to build a complex model incrementally, starting with a simple baseline and gradually adding more predictors. However, we found that our first model, which used only the H&E data, performed remarkably well, despite its simplicity. Adding the PGS, which is an already established and proved method, with publicly available data for hypercholesterolemia, was a natural choice and led to an even better performance. In the limited time we could dedicate to the challenge, we did try to include more predictors in the model, such as the Exposome Surveys data; we also tested an approach that included the SNVs falling in genes associated to hypercholesterolemia according to the Open Targets platform [**ochoa_2020**]. We believe that, given more time to explore the data and find a suitable feature selection scheme, these avenues could enhance the predictive performance, but so far we found that the initial H&E + PGS model was the best one. Another promising future direction is investigating the effect of methylation; there has been at least one study which did this [263], highlighting genome-wide methylation profiles as a potentially insightful source of signal. All in all, our model is a testament to the facts that simple models often outperform complex ones and, when limited resources are available, a small amount of high quality data can trump vast quantities of convoluted or less relevant data.

Acknowledgements

The past four years have been a challenging journey, and I am very grateful to those who have supported and guided me along the way. First and foremost, I would like to express my gratitude to my supervisors and TAC members: Peer, Julia, Maria, Rob, and Luis. Your guidance has been instrumental in shaping my interests and career. I am especially grateful to Peer for granting me the freedom to explore side-quests while ensuring the main projects stayed on track.

A special mention goes to Sophia Rudorf, who, although not one of my official advisors, was a great inspiration and made our collaboration both enjoyable and productive. I am also grateful to Rasmus, who imparted invaluable lessons on professionalism and project management.

To my lab-mates, thank you for the shared office space and the fun activities that made the long hours at EMBL much more enjoyable. Your companionship has been an essential part of this journey. I am especially indebted to my fellow PhD students: Chris, Askarbek, Marisa, Jonas, and Anastasia. Thanks also to Matej, Mahdi, Anthony, and Yan, for sharing geeky interests with me. Thank you, Renato, for introducing me to Bio-IT and its wonderful community.

I would also like to extend my heartfelt appreciation to my friends, who provided emotional support and encouragement, especially during the long running, bouldering, or home-made pizza sessions to relieve stress.

I am indebted to the developers of AlphaFold, HHblits, and all the heavy-duty software tools that I rely on daily. Without them, I simply couldn't have carried out my research.

Finally, to my beautiful wife Zhe, words cannot express my gratitude for your unwavering support, especially during the challenging times of writing this thesis. Your encouragement has been my anchor.