The Dissertation Committee for Alec Fraser Certifies that this is the approved version of the following dissertation:

Integration of Structural Methods to Characterize the Dynamics of Macromolecular Complexes

Committee:

Petr Leiman, PhD
Supervisor

Thomas Smith, PhD
Chair

B. Montgomery Pettitt, PhD
Co-Supervisor

Gabrielle Rudenko, PhD

Marc Morais, PhD

Steven Ludtke, PhD

Integration of Structural Methods to Characterize the Dynamics of Macromolecular Complexes

by

Alec Fraser, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas Medical Branch

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas Medical Branch

November 2021

Dedication

To my mentor (Petr) for his guidance. To my lab mates for their help and friendship. To my mother (Melissa), father (Andrew) and brother (Jameson) for their encouragement and assistance. To my fiancée (Ashley) for her ever-lasting support and inspiration.

Acknowledgements

The author would like to thank his mentor Petr Leiman for his insight, support, and the resources he has provided throughout this doctoral work. He would like to thank his co-mentor B. Montgomery Pettitt for his help and knowledge.

He would also like to thank: Nikolai Prokhorov, Fang Jiao, Simon Scheuring, Maria Sokolova, Arina Drobysheva, Julia Gordeeva, Sergei Borukhov, the AlphaFold team, Konstantin Severinov, Ekaterina Knyazhanskaya and John Miller for their contributions to this work and their fruitful collaboration. He would like to thank Misha Sherman, Gillian Lynch and Ka-Yiu Wong for their help and teaching. He would like to thank his committee members: Gabrielle Rudenko, Thomas Smith, Marc Morais and Steve Ludtke for their insights and guidance. All together these individuals contributed immensely to the success of the author's research.

A special thank you to the author's family: Andrew, Melissa and Jameson for their support and compassion. Thank you to Carla, Brian and Ellyn for their kindness and generosity. Thank you to Ashley, Anders and Logan for their eternal enthusiasm, comfort, and joy.

Integration of Structural Methods to Characterize the Dynamics of Macromolecular Complexes

Publication No._____

Alec Fraser, PhD

The University of Texas Medical Branch, 2021

Supervisor: Petr Leiman

Our understanding of the mechanism of protein function has been continuously adapted to be consistent with new experimental and theoretical findings. Eventually, the idea of a relationship between the structure and function of proteins emerged. As the methods available to probe structure became higher-throughput, more precise and easier to use, the conceptual underpinnings of this relationship have grown broader and more wholistic. Our comprehension of protein structure has shifted from static globular structures towards dynamic ensembles of folded and unfolded states. Despite this paradigm shift, existing methods for the characterization of large macromolecules remain focused on the characterization of stable, low-energy states. In this work, we aim to characterize the structure, and function, of various macromolecular complexes beyond the conventional low-energy, singular state approach. To do this, we extend existing methods and create new approaches catered for the study of large-scale macromolecular dynamics. Here, we characterize the disorder-to-order transition of a 5-subunit bacteriophage AR9 RNA polymerase holoenzyme throughout the process of uracil-specific promoter recognition with cryo-EM and x-ray crystallography. Using this structural information, we assess the energetics of DNA binding to the promoter binding pocket via molecular dynamics. Furthermore, energetic, and structural insights enable the design of a mutant RNAP

holoenzyme which can recognize altered thymine-containing DNA templates. Following this, we describe the mechanism of pyocin contraction using a combination of purpose-made computational modelling, electron microscopy and biophysical assays. Together, we characterize the relevant energetics and forces generated throughout the contraction process. Moreover, we validate predictions on how the energetics of contraction can be manipulated through mutagenesis. Finally, using cryo-EM, we solve the structures of the bacteriophage A511 sheath-tube complex in the extended, contracted, and intermediate states. Using this novel structural information, we assess the relationship between the motions of sheath subunits during contraction. We find that the process of sheath contraction can be described by a combination of local, subunit-level linear and non-local global properties. With the low-energy, tertiary structure protein folding problem largely solved, we propose that similar dynamical methods to those described below can be used to further our wholistic understanding of the structure function relationship.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Illustrations

List of Abbreviations

UTMB	University of Texas Medical Branch
GSBS	Graduate School of Biomedical Science
TDC	Thesis and Dissertation Coordinator
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
Cryo-EM	Cryo-Electron Microscopy
RNAP	RNA Polymerase
XFEL	X-ray Free Electron Laser
NMR	Nuclear Magnetic Resonance
SNR	Signal to Noise Ratio
AI	Artificial Intelligence
nvRNAP	Non-virion RNAP
vRNAP	Virion RNAP
dsDNA	Double-stranded DNA
DMAD	Domain Motion in Atomic Detail
DSC	Differential Scanning Calorimetry
ITC	Isothermal Titration Calorimetry
AFM	Atomic Force Microscopy
WT	Wild Type
CD	Circular Dichroism
ROC	Receiver Operating Curve

CCD	Charge Coupled Device
NOESY	Nuclear Overhauser Exchange Spectroscopy
IDP	Intrinsically Disordered Protein
SAXS	Small Angle X-ray Scattering
QM	Quantum Mechanics
MM	Molecular Mechanics
ATP	Adenosine Triphosphate
PfCRT	Plasmodium Falciparum Chloroquine Resistance Transporter
MD	Molecular Dynamics
RMSD	Root Mean Squared Deviation
SMD	Steered Molecular Dynamics
CV	Collective Variable
ABF	Adaptive Biasing Force
ML	Machine Learning
MDFF	Molecular Dynamics Flexible Fitting
EFG	Elongation Factor G
PDB	Protein Data Bank
RMSF	Root Mean Squared Fluctuation
CNN	Convolutional Neural Network
HDX-MS	Hydrogen/Deuterium Exchange Mass Spectrometry
Cryo-ET	Cryo-Electron Tomography
MSA	Multiple Sequence Alignment
GPU	Graphical Processing Unit

AF2	AlphaFold2
FL	Full Length
TSS	Transcription Start Site
NTP	Nucleoside Triphosphate
NTD	N Terminal Domain
CTD	C Terminal Domain
TL	Trigger Loop
PPC	Promoter Pocket Conformation
PCR	Polymerase Chain Reaction
MR	Molecular Replacement
DPBB	Double Psi Beta Barrell
NCS	Non-Crystallographic Symmetry
CTF	Contrast Transfer Function
SDS-PAGE	Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
UTP	Uridine Triphosphate
GTP	Guanosine Triphosphate
CTP	Cytidine Triphosphate
TACC	Texas Advanced Computing Center
DDM	Double Decoupling Method
EMDB	Electron Microscopy Data Bank
SCSB	Sealy Center for Structural Biology
CIS	Contractile Injection System
T6SS	Type VI Secretion System

COM	Center of Mass
PVC	Photorhabdus Virulence Cassette
EM	Electron Microscopy
SASA	Solvent Accessible Surface Area
UV	Ultra-Violet
HPC	High Performance Computing
AUC	Area Under Curve

Chapter 1 Introduction

THE ESTABLISHMENT OF MOLECULAR BIOLOGY AS A SCIENTIFIC DISCIPLINE

Molecular biology aims to obtain a molecular basis for the activity of biologically relevant entities both within and between cells("Essential Cell Biology - Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter - Google Books" n.d.). As stated by William Astbury, the physicist who pioneered the use of X-rays to study biological molecules and coined the term molecular biology: "[molecular biology] implies not so much a technique as an approach, an approach from the viewpoint of the so-called basic sciences with the leading idea of searching below the large-scale manifestations of classical biology for the corresponding molecular plan. It is concerned particularly with the forms of biological molecules, and with the evolution, exploitation, and ramification of those forms in the ascent to higher and higher levels of organization. Molecular biology is predominantly three-dimensional and structural which does not mean, however, that it is merely a refinement of morphology. It must at the same time inquire into genesis and function"(WADDINGTON 1961). From this statement, two important ideas are introduced as the cornerstones of molecular biology. Firstly, there is a relationship between the structure and function of biological molecules. Secondly, Astbury acknowledges that his "basic science" approach is a reductionist one. By reducing problems to the molecular scale, one can obtain high-fidelity models of biological function. From there, these molecular models need to be integrated and refined to study higher levels of complexity and organization.

THE SUCCESS OF STRUCTURAL BIOLOGY

The methodical, reductionist approach to solving protein structure (termed structural biology, a subfield of molecular biology) has been widely successful, with nearly 180,000 experimentally solved structures in the protein data bank at the time of writing(Berman et al. 2000). Since Astbury's time, tremendous advances have been made in both the ease and accuracy of biological structure determination. Ultimately, structural biologists are now able to determine sub-atomic resolution structures. Notably, in these structures, the positions of hydrogen atoms can be determined, and the bonding character of electron/catalytic transfer sites can be obtained(Blakeley, Hasnain, and Antonyuk 2015). In one case, the structure of a NiFe hydrogenase was solved to 0.89 Å by X-ray crystallography. The detail present in the resulting electron density map showed the products of the heterolytic splitting of dihydrogen, as well as the hydrogen bonding networks, and proton transfer pathways present in the structure(Ogata, Nishikawa, and Lubitz 2015).

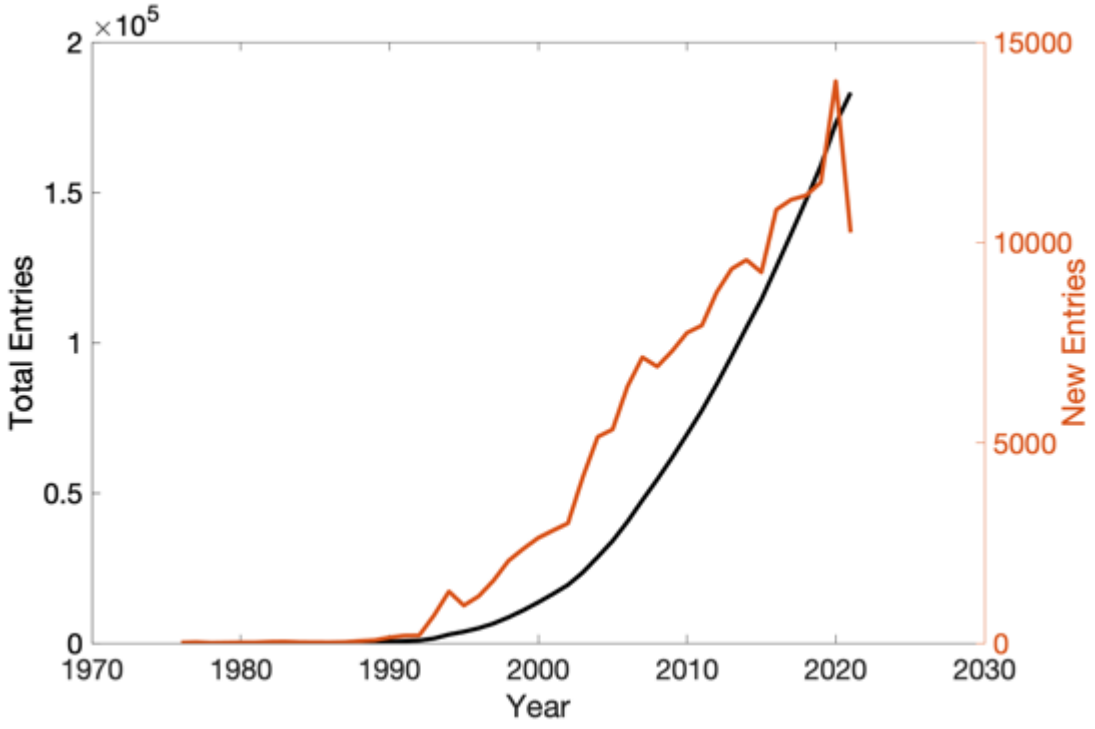


Figure. 1.1. Growth of the Protein Data Bank.

Exponential growth in PDB entries is representative of the success of structural biology. Data taken from https://www.rcsb.org/stats/growth/growth-released-structures

Recent improvements in structure determination can also be seen in the so called "resolution revolution" of Cryo-EM, whereby the field has gone from coarse-grain "blob-like" structures to near-atomic resolution models(Kühlbrandt 2014). At its inception, electron microscopes could only image large structures immersed in stain, such as viruses and bacteria, and the resulting images were extremely crude(Brenner and Horne 1959). In the 1980s, conventional vacuum systems and cold stages were developed, enabling the imaging of samples within amorphous ice, termed Cryo-EM. In the 1990s, brighter, more coherent field emission guns replaced tungsten sources. Finally, in the 2010s, electron detector technology was improved, whereby charge coupled devices (CCDs) (which must convert electrons into photons prior to detection) were replaced with direct electron detectors, greatly improving the signal to noise ratio (SNR) of Cryo-EM images(Egelman 2016). Similarly to X-ray crystallography, Cryo-EM is now capable of solving structures to atomic resolution. Recently, the atomic structure of mouse apoferritin was determined using a 1.22 Å resolution Cryo-EM map. Furthermore, using difference maps, researchers could accurately model hydrogens and thus could analyze the hydrogen bonding networks present in the structure(Nakane et al. 2020). Structural biologists will soon be able to routinely obtain near-atomic, atomic, or even subatomic resolution structures of biologically entities using various methods. As we approach the limits on the achievable resolution of molecular structure, we should reflect on the applicability and context of the reductionist approach.

THE IMPORTANCE OF DYNAMICS

Most structural studies done to date have focused on stable, low energy conformations of a molecule of interest. Despite this, it has been shown that understanding the dynamics of molecules is critical to properly describe their function("Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and ... - Alan Fersht, University Alan Fersht - Google Books" n.d.). As stated by Baldwin and Kay: "It is clear that the amino acid sequence of a protein encodes not only structural information, but also the range of dynamics accessible to the molecule, and that the interplay between structure and dynamics determines both the function of a protein and its ability to evolve and adapt to diverse sets of conditions"(Baldwin and Kay 2009). While a low-energy structure of a biomolecule can provide useful insight, it is generally insufficient to fully describe its function. Accordingly, there is a need for the integration of various structural techniques, with their respective contexts in mind, such that a more wholistic understanding of the structure-function relationship can be developed.

Conformational adaptability and the induced fit model

The idea of proteins being 'configurationally adaptable' was first proposed by Karush in the 1950s while studying the binding of anionic dye to bovine serum albumin(Karush 2002). In this study, Karush remarked on the ability of albumin to bind various ions and molecules in a wide variety of configurations. Furthermore, he was able to show that several anions could be bound in the same binding site. He suggested that in solution, albumin is in a state of thermodynamic equilibrium, populated by a variety of configurations of roughly equal energies. Upon binding, he proposed that albumin would adopt a conformation complementary to the ligand, with side chains changing conformation to stabilize the complex. Later, Koshland further developed the idea of conformational adaptability by expanding upon the well-established "lock and key" model for enzyme specificity(Fischer 1894). Koshland remarked that the "lock and key" model

didn't explain why enzymes would catalyze the reaction of smaller, roughly analogous substrates with efficiencies reduced by several orders of magnitude. Specifically, it seemed contradictory to the lock and key model that ribose-5-phosphate was hydrolyzed by 5'-nucleotidase 100x slower than adenylic acid at enzyme saturation conditions, despite the structure of ribose-5-phosphate being analogous to adenylic acid but without a purine(Koshland and Jr. 1958). Consequently, Koshland proposed that upon reaching the precise configuration necessary for enzymatic action, the substrate would modify the configuration of the active site, bringing catalytic residues into the proper configuration for reaction. In contrast, a smaller, analogous compound would not bring the same catalytic groups into formation. These ideas culminated in the infamous "induced fit" model.

Side chain dynamics are necessary for hemoglobin function

In the 1960s, the bound state of a hemoglobin analog was studied by X-ray crystallography to a resolution of 5.5 Å. While the position of the ligand in the electron density was expected, they did not expect the extent to which the binding site was buried. They showed that side chains blocked the entrance to the oxygen binding site. Upon discussion with Kendrew and Watson, the authors proposed that the side chain of histidine 7 must swing open to allow the entry of oxygen into the binding site(Perutz and Mathews 1966). Furthermore, spontaneous side chain motions were supported by reports of a null activation energy for hemoglobin-oxygen binding(Gibson and Smith 1965). Following these studies, it became commonplace to acknowledge the "conformational alteration" of proteins during allosteric regulation(Lumry, Eyeing, and 58 n.d.)·(Wolf and Briggs 1958)·(Sterman et al. n.d.).

Adaptation of nuclear magnetic resonance for the investigation of protein structure

In the 1960-70s, Nuclear Magnetic Resonance (NMR), a technique previously used in physics and chemistry, was adapted for the investigation of biological structure. Initially, biological NMR was limited to the study of paramagnetic proteins. In 1976, 2D NMR was conceived by Ernst, allowing for correlation maps between spins to be obtained(Aue, Bartholdi, and Ernst 2008) (In 1991 Ernst was awarded the Nobel prize in chemistry for his development of 2D NMR). A decade later, the structure of the first globular protein, proteinase inhibitor IIA, was solved by NMR (NOESY)(Williamson, Havel, and Wüthrich 1995). Importantly, this study represents the first atomistic description of protein dynamics in solution.

Discovery of intrinsically disordered proteins

In the 1990s, bioinformatic approaches were developed to predict structural elements of proteins based on their sequence. To the surprise of experimentalists, bioinformaticians began predicting that large segments of sequences in the SwissProt database encoded for "intrinsically disordered regions"(Wright and Dyson 1999). These proteins were predicted to lack any static tertiary structure and were termed "intrinsically disordered proteins" or IDPs(Ward et al. 2004)·(Sickmeier et al. 2007)·(Müller-Späth et al. 2010). Since the turn of the century, IDPs have been identified as crucial components in many eukaryotic processes including signal transfer, regulation and replication(Jensen et al. 2014). Furthermore, IDPs have been shown to play key roles in the progression of neurological disease(Uversky, Oldfield, and Dunker 2008)·(Babu et al. 2011)·(Uversky et al. 2014). Despite their importance, the lack of a well-defined tertiary structure has made the characterization of these proteins difficult. Nevertheless, NMR spectroscopy techniques have been adapted(Dyson and Wright 2005) to study these illusive proteins in near physiological conditions. For example, a combination of NMR and small angle X-ray scattering (SAXS) has been used to characterize Tau and alpha-synuclein(Schwalbe et al.

2014). In this study, researchers obtained amino acid specific ensemble descriptions of both IDPs. Furthermore, they found that both proteins favor the polyproline-II Ramachandran regions, especially in aggregation prone areas, which provides a structural basis for pathogenic Beta strand formation.

ADAPTATION OF STRUCTURAL BIOLOGY METHODS

Throughout the history of structural biology, our understanding of proteins has been continuously adapted to better describe their dynamic nature. However, at present, a general methodology for the experimental investigation of protein dynamics at high resolution is lacking. In particular, the characterization of large macromolecular motions has been challenging with current methods. Nevertheless, the pursuit of a dynamic description of proteins does not require the abandonment of methods which have historically thrived on homogenous, static structures (such as X-ray crystallography and Cryo-EM), but in understanding their context and limitations, and making the necessary adaptations. Certain methods have always described proteins as dynamic (NMR and Molecular Dynamics). Others have recently developed new techniques to better characterize dynamics (Cryo-EM and X-ray crystallography). At present, four main methodologies exist to study the dynamics of molecules, namely: NMR spectroscopy, time resolved X-ray crystallography, Cryo-EM, and Molecular Dynamics.

NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

NMR spectroscopy was instrumental in shifting our understanding of proteins from static globular structures towards dynamic ‘ensembles’. In contrast to X-ray crystallography, which requires proteins to adopt a stable, repeatable structure for crystal growth, NMR characterizes proteins as ensembles of folded or unfolded structures in

solution. It is the only experimental method which can characterize IDPs at high resolution(Jensen, Ruigrok, and Blackledge 2013),(Konrat 2014).

The strength of NMR lies in its ability to probe the dynamics of proteins on both long and short timescales, providing site-specific information about protein conformational changes. For instance, nuclear spin relaxation rate experiments monitor internal motions on short timescales (ps-ms), whereas proton exchange experiments track domain motions on longer (ms to days) timescales(Ishima and Torchia 2000). For example, in one study, researchers used spin relaxation experiments to investigate carbohydrate binding by a Galectin-3 domain on a picosecond timescale. Surprisingly, they found that protein conformational fluctuations were more prevalent (as opposed to quenched) in the ligand bound state, relative to the unbound state(Akke 2012),(Diehl et al. 2010). Furthermore, their analysis suggested that the entropic contribution was critical in driving carbohydrate binding and was similar in magnitude to the enthalpic component. In the case of slower motions, NMR has been successful in characterizing protein dynamics coupled to ligand binding for the T4 lysozyme. In this study, researchers investigated how a mutant lysozyme(Eriksson, Baase, and Matthews 1993) was able to bind small hydrophobic aromatic ligands in a solvent inaccessible binding cavity. Using relaxation dispersion methods, they observed significant us-ms conformational dynamics in the cavity, which enabled ligand binding. Furthermore, they showed that this binding conformer was in fact higher in energy than the ground state, due to increased disorder in the region(Mulder et al. 2001).

NMR can also be used to characterize folding intermediates. Specifically, NMR has been used in the identification of a low population state of the Fyn SH3 tyrosine kinase between the unfolded and folded states(Korzhnev et al. 2004). In this study, researchers used relaxation dispersion NMR experiments at various temperatures to characterize the energetics and kinetics of Fyn SH3 folding via an intermediate.

These studies highlight that NMR is a powerful technique for the investigation of protein dynamics on both short and long timescales. NMR is unique in its ability to evaluate the entropic contributions to various structural transitions. As a result, NMR can provide insights into both the dynamics and thermodynamics underlying biological function. Despite this, the sizes of the galectin-3 carbohydrate recognition domain, T4 lysozyme and Fyn SH3 domain are 138, 164 and 60 residues respectively, making them small enough for NMR experiment. Generally, NMR spectroscopy can only study small proteins (<35 kDa) which puts a limit on its general applicability.

TIME RESOLVED X-RAY CRYSTALLOGRAPHY

Since the structure of myoglobin(Kendrew et al. 1958) was solved in 1958 X-ray crystallography has been central to the investigation of biological structure. In most cases, X-ray crystallography aims to collect diffraction patterns of a biological molecule in a singular, low-energy state. Recently, however, crystallography has been extended to probe the dynamic properties of biological molecules. Time resolved X-ray crystallography collects a time series of diffraction patterns to investigate the conformational changes of molecules after a stimulus. Generally, in these experiments, crystals are fed through a 'pump' laser, which stimulates photo-active molecules.

Initially, these experiments were carried out at specialized synchrotrons with a temporal resolution on the order of hundreds of picoseconds(Doerr 2015). In 2003, researchers used mid infrared spectroscopy and a flash-photolyzed Myoglobin (Mb) mutant to observe the frame-by-frame evolution of MbCO binding(Schotte et al. 2003). The binding process was observed with 1.8 Å spatial resolution and 150 ps time resolution, allowing for the identification of a short-lived CO bound intermediate. Furthermore, they observed large structural changes during binding, including the correlated side chain motions responsible for the removal of CO from the primary binding site.

X-ray free-electron lasers

X-ray Free-Electron Laser (XFEL) systems can produce ultra-brilliant, coherent MHz pulses at a femtosecond time resolution(Pandey et al. 2019), enabling the study of ultra-fast structural transitions. Using XFELs, single shot diffraction patterns can be serially collected from a continuous flow of micro-crystals in random orientations(Schlichting 2015). The use of femtosecond pulses allows for data collection to occur before significant radiation damage accumulates in the crystal, suggesting that the resulting structures are "damage-free".

Time resolved XFELs

Time resolved XFEL has been used to study irreversible enzymatic reactions, whereby the "pump-probe" technique is combined with photo-sensitive caged substrates to observe enzymatic reaction intermediates. In one example, the fungal NO reductase enzyme was captured in an intermediate NO-bound conformation to a resolution of 2.1 Å(Tosha et al. 2017). Upon comparison with cryo-XFEL crystallography and QM/MM calculations, the structure was determined to have little to no radiation damage. Furthermore, the high resolution and little radiation damage present in the structure allowed for the identification of a specific bond geometry of NO in the enzymatically bound state, enabling a model for the enzymatic activity to be developed.

Nanosecond dynamics of bovine cytochrome c oxidase using time resolved XFEL

Recently, researchers investigated the dynamics of bovine cytochrome c oxidase, a protein which pumps protons across a membrane potential. It was previously known that

the oxidase must tightly regulate the timing of channel opening and closing to stop the back flow of protons through the channel. Using time resolved XFEL and infrared probing, they investigated the mechanism of channel opening (which occurs upon the release of CO from the oxidase protein), with nanosecond time resolution(Shimada et al. 2017). Their work uncovered a unique relay system, whereby 1) the redox active copper senses proton collection in the channel and then 2) binds oxygen followed by 3) binding heme a3 which then closes the channel, blocking the leakage of collected protons.

These examples illustrate that time resolved X-ray crystallography can investigate structural transitions at an ultra-fast temporal resolution with atomic spatial resolution. Furthermore, in the case of XFEL setups, the resulting structures have little-to-no radiation damage and thus like the native state. The astonishing level of detail enables the creation of precise models for protein function, such as the identification of bonding geometry for the NO bound reductase(Tosha et al. 2017). Unfortunately, the primary hurdle to the broad application of time resolved XFEL lies in the cost/scarcity of XFEL sources. Secondarily, not all reactions can be stimulated by the 'pump-probe' technique.

CRYO-ELECTRON MICROSCOPY

Over the past decade, Cryo-EM has emerged as a technique capable of generating near atomic resolution structures. Generally, its practitioners aim to generate maps of quality and resolution comparable to X-ray crystallography. Similarly to X-ray crystallography, the initial step in the standard workflow involves generating a homogenous sample in the lab. In contrast to crystallography however, which has sample homogeneity enforced via crystal packing, Cryo-EM enforces homogeneity via clustering algorithms post data collection. As a result, up to 90% of the particle images collected are discarded(Kschonsak et al. 2020),(Fitzpatrick et al. 2017),(Laverty et al. 2019). While some particles do need to be discarded due to the presence of 'junk' or radiation and air-water-

interface induced damage, numerous particle images are discarded because they do not conform with the most populous state. Moreover, many of these discarded particles are likely representative of the ensemble of biologically relevant conformations present in solution.

Various methods have been developed to characterize the dynamical information present in Cryo-EM datasets. The simplest method lies in applying conventional clustering algorithms to 'tease out' various conformational intermediates. In the best-case scenario, where datasets contain millions of particles, this technique can result in near atomic resolution structural intermediates. For instance, researchers imaged the proteasome during the breakdown of a polyubiquitylated protein(Dong et al. 2018). Using an initial dataset of nearly 3 million particles, they solved the structure of the human 26S proteasome in 7 functional states, ranging from ubiquitin recognition to substrate translocation at ~3 Å resolution. Furthermore, they proposed a structural basis for substrate regulation via the hydrolysis of ATP and rigid-body hinge-like motions of the ATPase.

Cryo-EM docking techniques

In some instances, the limited size of datasets or flexibility inherent in the sample can limit the resolution of structural intermediates. In these cases, models derived from other methods can be "flexibly fit" or "docked" into Cryo-EM maps using a rigid body fit of individual protein domains. For example, the structure of a DNA-dependent protein kinase catalytic subunit was investigated through a combined Cryo-EM/X-ray crystallography docking approach(Williams et al. 2008). In this work, a homology model was docked into a 7 Å resolution Cryo-EM map. The corresponding fit was of sufficient quality for a model to be developed for the interaction of DNA with the protein kinase catalytic subunit whereby the dsDNA enters the central channel of the kinase and interacts with a resolved alpha helical protrusion.

CryoSPARC non-uniform refinement

Recently, more sophisticated methods have been developed to characterize dynamics in Cryo-EM data. One such method, non-uniform refinement implemented in CryoSPARC, uses a novel spatially dependent regularizer depending on the SNR of the region. The result is that noise is removed from disordered regions of the map while signal is kept in well-resolved regions. This is in stark contrast to conventional methods which apply the same regularizer to the entire map. In practice, NU-refinement has been useful in studying macromolecules with significant dynamic properties, such as membrane proteins(Punjani, Zhang, and Fleet 2020). In one case, NU-refinement was implemented for the reconstruction of the plasmodium falciparum chloroquine resistance transporter (PfCRT) immersed in a lipid nano disc. PfCRT is a small, asymmetric membrane protein which is important for antimalarial treatments. When compared with traditional refinement strategies, NU-refinement improved the resolution from 6.9 Å to 3.6 Å, allowing for the unambiguous building of atomic structure(Kim et al. 2019). This remarkable improvement in resolution can be attributed to the improved alignment in NU-refinement, relative to uniformly regularized refinement(Punjani, Zhang, and Fleet 2020).

Spotiton: an experimental technique for observing transient intermediates in cryo-EM

In addition to software-driven methods, experimental grid-preparation techniques can be used to characterize protein conformational change. One interesting application, termed Spotiton, deposits multiple samples onto a grid approximately 100 ms prior to freezing(Dance 2020). The goal of this technique is to image short-lived transient intermediates which do not exist in large numbers at equilibrium. In one example, Spotiton

was used to study DNA opening by the E. Coli RNAP holoenzyme. After 150 ms of incubation prior to freezing, data processing resulted in 2D classes representing transient intermediates of the promoter bound holoenzyme(Dandey et al. 2020). In future experiments, researchers plan on vary the freezing and incubation conditions to develop an atomistic model for nucleation and the propagation of the transcription bubble.

Cryo-EM is unique in its ability to characterize many different functional intermediates from a singular dataset, such as in the case of the human 26S proteasome(Dong et al. 2018). Moreover, improvements in both hardware and software have led to higher resolution models of these structural intermediates. Despite this, the time resolution available to study such systems is generally limited by plunge freezing (~100ms) and is significantly slower than NMR or time resolved XFEL. Furthermore, it remains a challenge to study large macromolecular complexes due to the prohibitive computational cost.

MOLECULAR DYNAMICS

Molecular dynamics is a powerful tool to study the dynamic properties of molecules whose structures have been solved experimentally. Over the past 30 years, it was often the case that an "experimentalist" would use X-ray crystallography (or more recently Cryo-EM) to solve the atomic structure of a molecule of interest, then a "theorist" would use molecular dynamics to gain a deeper understanding of how it functions. The simplest and earliest form of MD is so called "brute-force" simulations. In these simulations, an experimentally derived structure is simulated in aqua (using either implicit or explicit solvent) to understand the dynamics of the molecule at equilibrium. Over sufficiently large timescales, important structural transitions can be observed, and the underlying energetics can be inferred from their Boltzmann weighted populations. Brute-force simulations on the microsecond timescale have been shown to yield microscopic properties (such as RMSD,

secondary structure and hydrogen bonding) comparable to those resulting from NMR(Spoel* and Lindahl 2003).

Enhanced sampling molecular dynamics: umbrella sampling

Unfortunately, many molecules are simply too large to simulate for biologically relevant timescales, even using modern high performance computing facilities. This disparity between the achievable and desired timescales spawned "enhanced sampling" methods. These methods add external biases to a system to focus sampling on important regions and accelerate desired structural transitions. The first enhanced sampling method was umbrella sampling as implemented in Monte Carlo simulations(Torrie and Valleau 1977). In this work, authors aimed to assess the free energy difference between a state of interest and a reference state using Monte Carlo methods (a molecular modelling technique which uses equilibrium statistical mechanics instead of simulating the dynamics of a system directly). Traditionally, Monte Carlo methods obtain estimates of free energy differences using Boltzmann-weighted sampling distributions, however, this can be extremely computationally ineffective as ergodicity is limited by the latent, uneven energetics of the system. To circumvent this issue, the authors added arbitrary sampling distributions to the system, which enhanced sampling in-key regions and dramatically improved free energy estimates. Furthermore, they were able to show that this technique was effective over a wide range of pressures and temperatures for a simple Lennard-Jones system.

Enhanced sampling molecular dynamics: steered molecular dynamics

Inspired by the experiments of atomic force microscopy, steered molecular dynamics (SMD) was developed to probe the mechanical properties of proteins in silico.

In SMD, special atoms or groups of atoms are pulled with constant forces or at constant velocities(Israelewitz, Gao, and Schulten 2001). Given sufficient sampling by SMD, the Jarzynski's equality (also known as the non-equilibrium work equation) can be used to relate the non-equilibrium forces to the work via a potential of mean force(S. Park et al. 2003)·(S. Park and Schulten 2004). In one example, constant force SMD simulations were used to characterize and identify 3 unfolding intermediates of the extracellular matrix protein Fibronectin(Gao et al. 2002). SMD experiments showed that the first and second metastable intermediates corresponded to twisted and aligned states prior to the unravelling of Fibronectin Beta strands. The third intermediate showed the A and B strands as unraveled but with the F and G strands intact. Their analysis suggested a preferred pathway for FN-III unfolding under physiological conditions. Importantly, similar unfolding pathways have been observed in single molecule atomic force microscopy experiments(Oberhauser et al. 2002).

Enhanced sampling molecular dynamics: adaptive biasing forces

Recently, collective variable (CV) methods have become prevalent in enhanced sampling MD. CVs are functions of atomic coordinates (such as radius of gyration or RMSD) which are important in describing the slow motions of the transition of interest(Yang et al. 2019). These methods work by applying "biasing potentials" along CVs to improve sampling in important regions of phase space. One useful CV-based method is that of adaptive biasing forces. In this method, adaptive biasing forces are externally applied to the system to nullify the latent energy barriers between states. Once converged, the biasing forces will near perfectly cancel the underlying free energy surface and thus the molecule will uniformly traverse the transition pathway. In one example, ABF simulations are used to study the transition of nucleic acid strands through the heptameric protein nanopore alpha-hemolysin(Martin, Jha, and Coveney 2014). In this study, they

recapitulated the experimental result that poly(A) has a larger energetic barrier to translocation than poly(dC). Furthermore, they showed that the ABF method resulted in energetic values closer to experiment than constant velocity Jarzynski equality SMD methods.

The strength of molecular dynamics lies in its ability to provide a "rationale" for the observed structural transitions of molecules. It can assess the energetic contributions underlying molecular dynamics and, as in the case of ABF and other methods, apply external forces to probe the system out of equilibrium. The main weakness of molecular dynamics is its poor scalability to larger systems. While enhanced sampling methods have somewhat alleviated this issue, the investigation of large protein complexes by molecular dynamics remains a challenge.

MACHINE LEARNING

Machine learning (ML) techniques can be separated into three general categories: supervised, unsupervised and reinforcement learning. Supervised learning techniques make predictions about data which possesses categorical "labels". For instance, a convolutional neural network can be trained to classify images ranging from "French fries" to a "washing machine"(Krizhevsky, Sutskever, and Hinton n.d.). Unsupervised learning techniques, in contrast, learn patterns about data which lacks a priori labels. As an example, a k-means clustering algorithm can be used in Cryo-EM to group particles according to their similarity("Three-Dimensional Electron Microscopy of Macromolecular Assemblies ... - Joachim Frank - Google Books" n.d.). Reinforcement learning techniques make decisions with the goal of maximizing pre-determined "rewards". In many cases, these algorithms first attempt to maximize the reward via trial and error, and eventually learn more sophisticated techniques for reward maximization. In one example, reinforcement

learning was used to train an algorithm to play Atari at a super-human level using the raw pixels as inputs and an estimated future rewards function(Mnih et al. 2013).

Deep machine learning

Deep learning techniques are a subset of ML techniques which can be applied to all three types of learning problems. They differ from other ML techniques in their ability to use less pre-processed data (e.g., images vs. tables), self-identify important features (e.g., the identification of ears for the classification of images) and leverage large amounts of data for increased performance (traditional ML techniques scale poorly with larger amounts of data). Deep learning techniques utilize large neural network architectures which consist of sets or "layers" of interconnected artificial neurons with trainable parameters, namely weights and biases. Importantly, these networks generally have multiple "hidden" layers between the initial input layer and the final output layer. The passage of information through the network involves the activation of neurons and the propagation of these activations between layers, mediated by inter-neuronal connections. The forward propagation of information (towards the output layer) consists of the features from previous layers being integrated and combined such that deeper layers can identify more complex features. Similarly, "backpropagation" and "gradient descent" describe the reverse process, whereby desired improvement in the accuracy of feature determination propagates backwards through the network, updating the weights and biases accordingly. Overall, the process of training consists of an iterative forward- and backward-propagation, until the optimal accuracy of the algorithm is achieved.

NOVEL COMBINATIONS OF STRUCTURAL BIOLOGY TECHNIQUES

To circumvent the conventional limitations of NMR, X-ray crystallography, Cryo-EM and MD, researchers have been integrating and combining structural techniques in innovative ways. In many cases, the information resulting from a singular structural technique can be "low resolution" or incomplete. By supplementing lacking information from one method, with available information from another, dynamic characterizations of molecules can be developed which would be otherwise lacking.

Molecular dynamics flexible fitting

Molecular dynamics flexible fitting (MDFF) techniques were developed to fit atomic structures into maps which are of insufficient quality for de novo model building. MDFF improves upon conventional "docking" methods by combining the coarse-grain, global information contained in the low resolution maps with the fine-grain, local information contained in molecular dynamics force fields(McGreevy et al. 2016)·(Wriggers, Milligan, and McCammon 1999). In one example, researchers investigated the conformational dynamics of Elongation factor G (EFG) binding the ribosome prior to the translocation of tRNA. Using Cryo-EM, they found that the conformation of EFG was substantially different from the X-ray crystallography structure(Tama, Miyashita, and Brooks 2004). Using MDFF, they fit the X-ray model into the Cryo-EM density and found that there were correlated rigid body motions between domains II, IV and V.

Combination of x-ray crystallography and NMR

Room temperature X-ray crystallography can be combined with NMR relaxation experiments to investigate the conformational fluctuations of proteins on the nanosecond to picosecond timescale(Fenwick et al. 2014). While studying the enzyme dihydrofolate

reductase, researchers compared backbone and side chain order parameters derived from NMR to those from room temperature X-ray crystallography and found them to be consistent, suggesting that the picosecond conformational fluctuations observed in solution are also present in the crystalline state. Accordingly, they devised an innovative approach whereby NMR and room temperature x-ray crystallography data are iteratively refined together to identify conformational substates of the enzyme. Ultimately, this approach resulted in a higher quality characterization of the spatial and temporal fluctuations of the enzyme than either individual method.

INTRODUCTION OF DEEP LEARNING TECHNIQUES TO STRUCTURAL BIOLOGY

Recently, deep machine learning techniques have been implemented as important tools for biological structure determination and dynamics analysis. While traditional machine learning techniques, such as k-means clustering in Cryo-EM classification, have been used for decades, the inclusion of deep learning techniques marks a fundamentally different approach. Traditional machine learning implementations in structural biology are generally trained on local data and result in an intermediate accuracy. In contrast, deep learning techniques are trained on large databases (such as the PDB and Genbank) and can provide higher levels of accuracy. Ultimately, deep learning techniques have been introduced to "fill in the gaps" in our experimental and theoretical knowledge of structural biology.

DefMAP: learned dynamics from cryo-EM maps

Deep learning and molecular dynamics methods have been integrated into the Cryo-EM methodology to better model the dynamic properties of macromolecules. In one example, named DefMAP(Matsumoto et al. 2021), brute-force molecular dynamics is

performed on already-solved protein structures to determine the per residue root mean squared fluctuation (RMSF). Using the resulting RMSF information, the relationship between the RMSF and the original Cryo-EM map is learned via the training of 3D convolutional neural networks (CNNs). The trained CNN can then be applied to new Cryo-EM maps to infer the dynamic properties of macromolecules without molecular dynamics simulation. To test their method, researchers compared the results of DefMAP to hydrogen deuterium exchange mass spectrometry (HDX-MS) inferred dynamics of the P-Rex1–Gβγ signaling scaffold(Cash et al. 2019). They found that the correlation between the MD-inferred and DefMAP-inferred dynamic fluctuations were r = 0.8 and 0.75, respectively, suggesting that the theoretical results are equivalent to HDX-MS experiment.

Topaz Denoise: improving the SNR of cryo-EM images

Deep learning has also been implemented to improve the SNR of Cryo-EM/CryoET micrographs. Cryo-EM and CryoET images are notoriously noisy due to the requirement of low electron dose exposures. Consequently, a deep learning method termed Topaz-denoise, has been developed to improve the SNR of Cryo-EM/CryoET micrographs. Topaz-denoise consists of a deep learning architecture which was trained on a dataset of thousands of micrographs (of various imaging conditions)(Bepler et al. 2020), such that the model has learned the Cryo-EM image formation process. Accordingly, this model can be applied to new micrographs and improve the SNR without the additional tuning of parameters. Furthermore, the application of this method results in faster data acquisition, a lower required electron dose (which reduces radiation induced damage) and improved micrograph interpretability. To test their method, the authors implemented Topaz-denoise to improve the visibility of clustered protocadherin micrographs. The improved interpretability of the micrographs resulted in 2.15x more particles being picked, with

substantially more top and oblique views. Ultimately, they were able to reconstruct the closed conformation to 12 Å, a significant improvement over the previous 35 Å model.

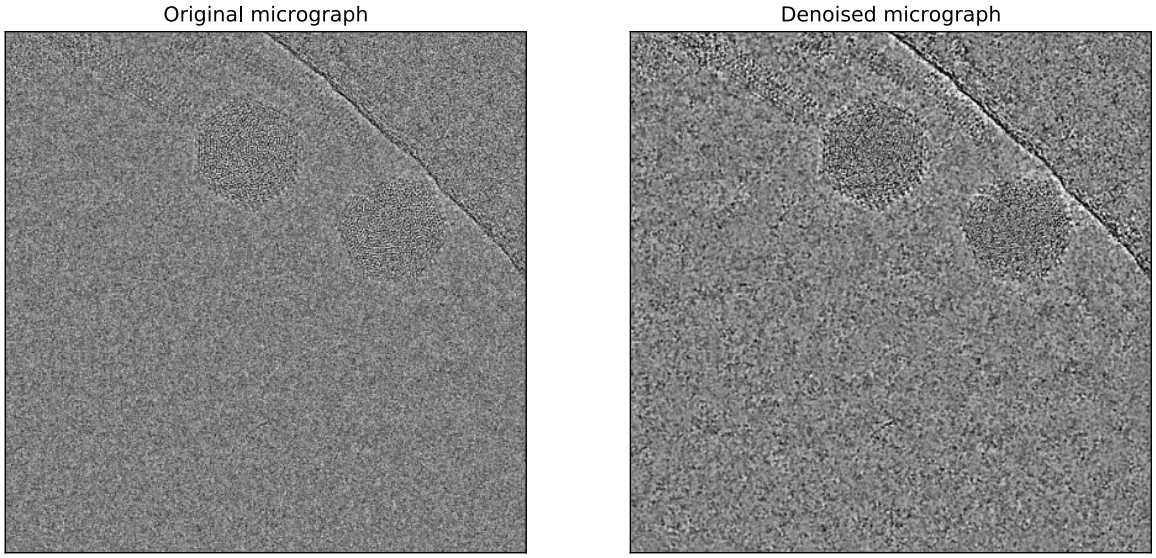


Illustration. 1.1. Example use of Topaz Denoise.

(left) Original cryo-EM micrograph of bacteriophage A511. (right) Denoised(Bepler et al. 2020) micrograph. Images were produced using cryoSPARC(Punjani et al. 2017).

Recently, novel deep learning techniques have been applied to the problem of protein folding. Historically, computational modelling of protein folds has been approached from so called “physics-based” or “evolution-based” approaches. Physics-based approaches rely upon the theoretical knowledge of protein-physics but are plagued by inaccurate or over-simplified models and computational intractability. Alternatively, evolutionary approaches have become more popular due to the ease in implementing deep learning methods and the vast amount of available bioinformatic data. Overall, deep learning methods, which extract data from large datasets, seem to work better than traditional methods which attempt to replicate the folding process using physics.

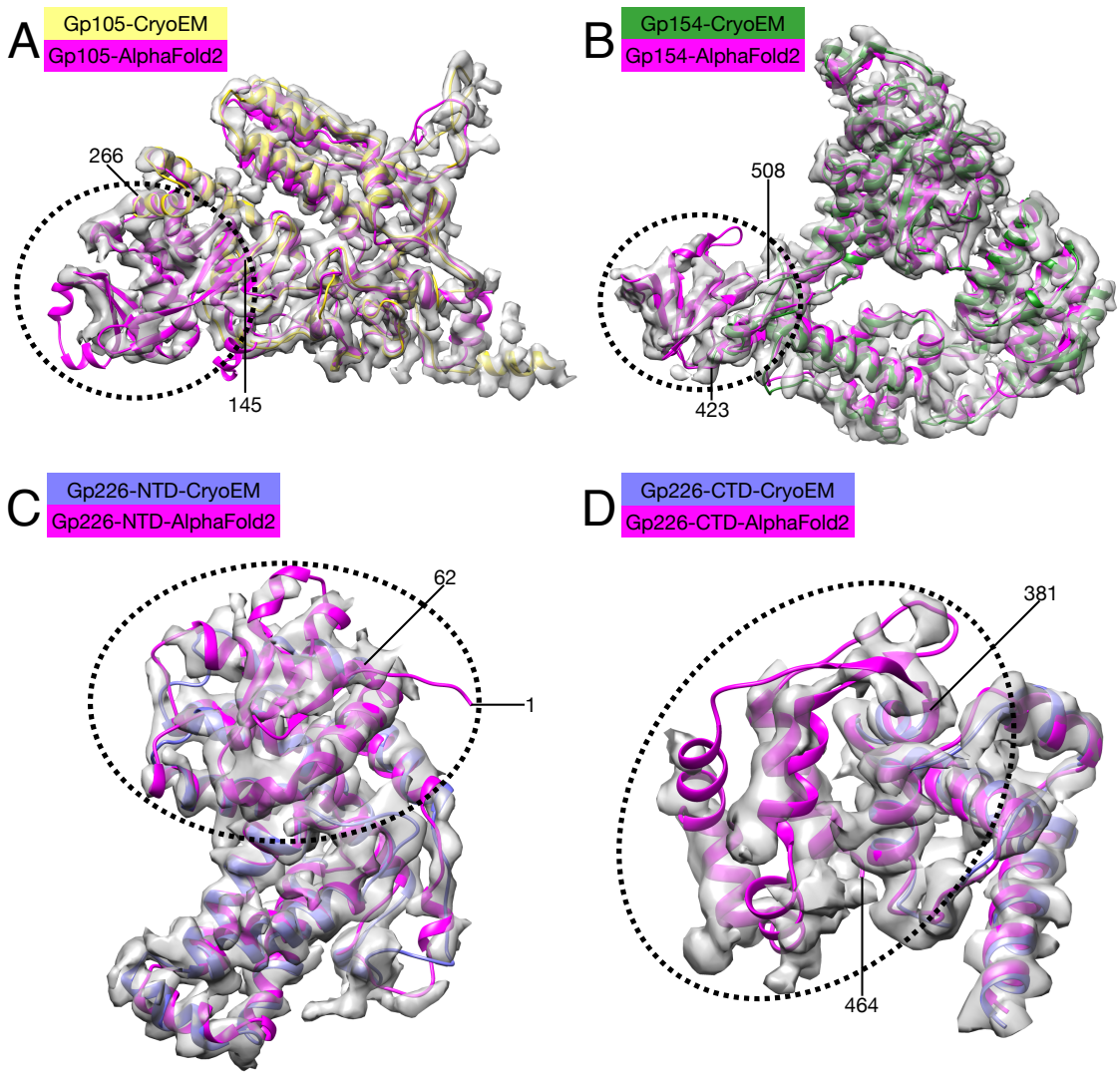


Figure 1.2. **AlphaFold2 models of AR9 nvRNAP proteins fit the cryo-EM density nearly perfectly**. The cryo-EM-derived structures of gp105, gp154, and two gp226 domains are colored according to the color code given in the upper left corner of each panel. All AlphaFold2 models are colored magenta. The electron density is contoured at 4.25 standard deviations above the mean and colored semi-transparent grey. Regions where no cryo-EM-derived structure existed prior to the availability of the AlphaFold2 models

are indicated with a dashed line and their boundary residues are labeled. Reproduced with permission from(Kryshtafovych et al. 2021).

AlphaFold2: highly accurate protein structure prediction

The first successful prediction of 3D protein structure solely from sequence was achieved by AlphaFold2 (AF2)(Jumper et al. 2021). Predictions generated by AF2 were significantly better than other methods due to the combination of physical, evolutionary, and geometric constraints in a novel neural network structure. In particular, the neural network was comprised of two portions, the Evoformer building block and the structure module, in such a way that information could be passed back and forth. The Evoformer block uses unprocessed multiple sequence alignment (MSA) information to learn the spatial and evolutionary relationships between residues via their representation as a graph, where edges are proximal residues. The structure module generates translations and rotations for each residue, whereby protein chains can break, and bond geometry isn't strictly enforced but encouraged via a loss term. Additionally, within the structure module, an equivariant SE(3) transformer network is used to refine atomic coordinates. Outputs from the network are then iteratively "cycled" by feeding them as inputs through the network. Importantly, the accuracy of predictions generated from the structure module are computed from activations at the end of the network. Remarkably, the accuracy of the AF2 network was further improved via the inclusion of unlabeled (lacking a PDB structure) data. 350k sequences were predicted using the neural network and those which the network deemed to be highly accurate were used to supplement the available PDB training data. They hypothesize that deep MSAs are necessary to develop a coarse grain structure prediction but are not necessary for fine-grained refinement of structure. Consequently, targets with a mean alignment depth of less than 30 sequences can be difficult to predict.

Furthermore, AF2 struggles to predict structures which are dictated by quaternary interactions, such as bridging linkers(Kryshtafovych et al. 2021). Overall, the AF2 method predicted CASP14 targets with a median CA RMSD of 0.96 Å and highly accurate side chains.

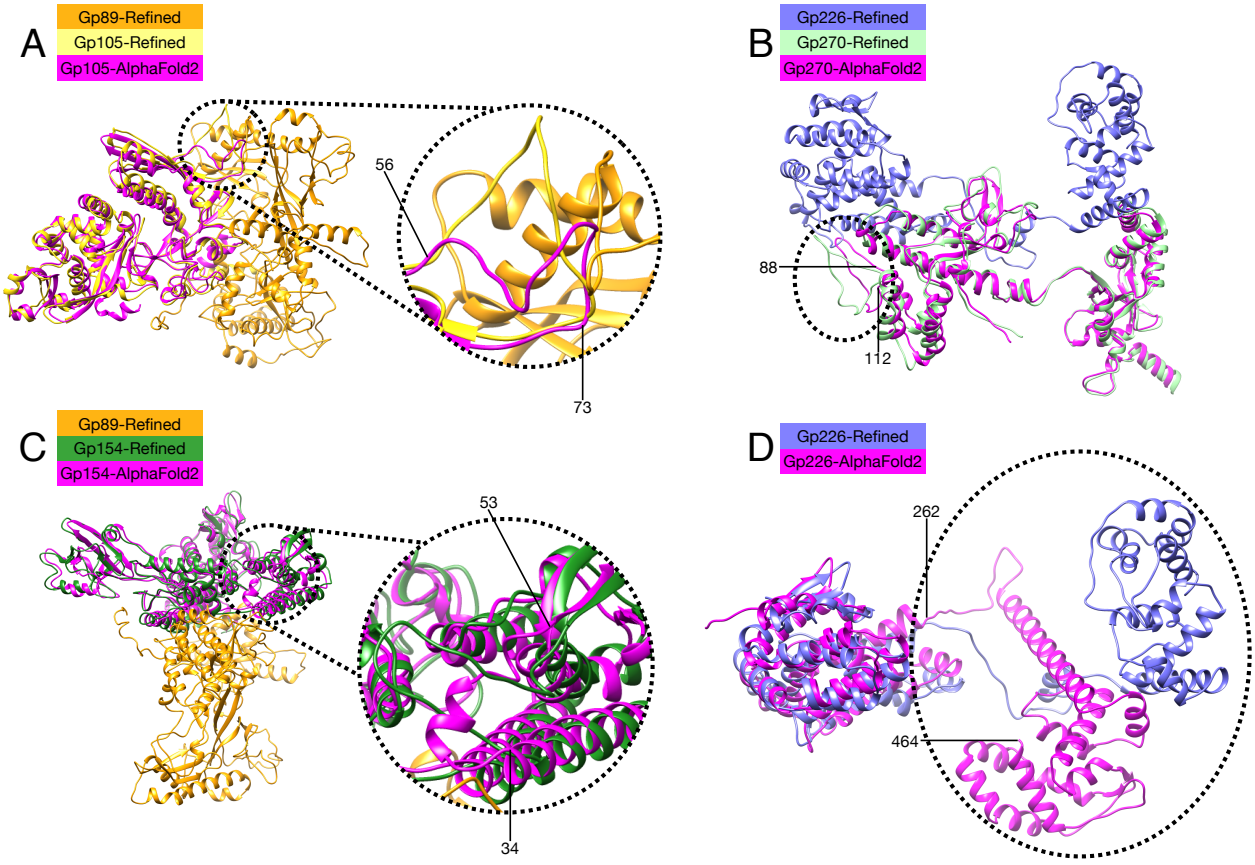


Figure 1.3. **Inaccuracies in AlphaFold2 models**. Cryo-EM-derived structures and AlphaFold2 models of several AR9 nvRNAP subunits are superimposed and regions where the conformation of the AlphaFold2 model deviates significantly from the cryo-EM-derived structure are indicated with a dashed line and their boundary residues are labeled. Note that the folds of both the N- and C-terminal domains of gp226 were predicted

correctly, but the structure of the interdomain linker and the relative orientation of the two domains were incorrect. Reproduced with permission from(Kryshtafovych et al. 2021).

RoseTTAFold: accurate and efficient protein structure prediction

Following the presentation of AF2 at CASP14(Pereira et al. 2021), another high-fidelity protein structure predictor, termed RoseTTAFold, was published(Baek et al. 2021). RoseTTAFold uses a 3-track network, combining 1D sequence, 2D distance maps and 3D coordinate information into a singular network. In their architecture, information flows between the 1D, 2D and 3D layers in a singular pass of the network, as opposed to the AF2 network, which iteratively "cycles" through the network. Since graphical processing unit (GPU) memory is limited by hardware constraints and the required GPU memory scales with sequence size, RoseTTAFold processes discontinuous segments of sequences and then combines the 1D and 2D features from each segment to create a final model. In particular, the final model is generated by two parallel approaches: 1) 1D and 2D information is fed into pyRosetta(Chaudhury, Lyskov, and Gray 2010) 2) 1D and 2D information is fed into an SE(3) equivariant layer, where the structure is generated from the neural network after end-to-end training. Importantly, the segmentation and recombination of sequences seems to improve the accuracy of final models. To investigate the predictive strength of this method, RoseTTAFold was tested on newly deposited PDB structures and outperformed all other servers. Furthermore, RoseTTAFold models were successful in solving four previously elusive crystallographic datasets using molecular replacement and one Cryo-EM dataset. In preliminary experiments, RoseTTAFold was even able to accurately model some protein complexes made up of two or three protein chains, learning accurate residue co-evolution from deep MSAs. Despite this, the accuracy of models generated by RoseTTAFold are generally worse than AF2, likely due to

limitations in available hardware, differences in architecture and the single pass vs. multiple pass approach. In the future, the authors envision that this technology will be used in the design of small molecules and proteins for the binding of medically important targets.

CONCLUDING REMARKS

Throughout the history of molecular biology, and later, structural biology, models of protein structure and function have been iteratively adapted to accommodate new information(Karush 2002),(Koshland and Jr. 1958),(Perutz and Mathews 1966),(Wright and Dyson 1999). In its earlier stages, these adaptations manifested themselves in our conceptual knowledge of protein structure, shifting our understanding from a static globular structure to a dynamic set of microstates. Despite our conceptual understanding of proteins being shifted to a more realistic and wholistic viewpoint, most methods accessible to scientists at the time did not have the resolving power necessary to probe proteins with this level of detail. Consequently, most structural studies have focused on singular, low energy states of a protein of interest.

This focus on well ordered, highly populated states has been extremely successful, with an average of 7,068 X-ray crystallography structures released in the PDB each year since 2001. Cryo-EM has seen rapid progress in its use, with 13 PDB structures released in 2001, but 2390 in 2020. NMR, in contrast, which does characterize proteins as dynamic ensembles, has only released an average of 555 structures per year over the last two decades. Interestingly, the use of multiple methods for structural determination has become more common in the PDB, with 0 structures released in 2001 but 20 in 2020.

AI-based solution to low-energy tertiary protein structure

The immense number of structures in the PDB, sequences in depositories such as Genbank and advances in neural network architectures has resulted in a solution to the protein structure prediction problem(Jumper et al. 2021),(Baek et al. 2021). To be specific, there are some instances where structure cannot be predicted reliably, such as when deep MSA information is lacking or when the protein fold is dictated by quaternary interactions. Nevertheless, in the upcoming years, the tertiary structure of most proteins will be determined with ease. Consequently, it is evident that the role of structural biology techniques will be shifted. It seems likely that structural techniques will adapt to this disruptive technology by further integrating with AI, making the structure determination process even easier. Furthermore, it will become important to extract information complementary to that derived from the AI prediction of the low energy state.

In this new paradigm, each of the four main structural biology techniques will have their own strengths and weaknesses. Techniques which can characterize macromolecular complexes will be able to add value beyond the predictions of tertiary structure(Kryshtafovych et al. 2021). Accordingly, Cryo-EM and X-ray crystallography will still be necessary for the determination of accurate quaternary structure. Furthermore, many biologically important complexes exist in several states as they function (e.g., binding ligands). As such, imaging complexes in varied conditions (e.g., with and without ligand) will provide biologically relevant information about protein conformational change, currently inaccessible to deep learning methods. Moreover, techniques provide dynamical information about proteins will be even more useful. Accordingly, NMR and Molecular Dynamics will prove to be complementary to AI methods. With the protein folding problem largely solved, it is finally time to shift the goal of structural biology towards a more wholistic description of proteins and the structure-function relationship.

Motivation and summary of original work

It seems inevitable that the future of structural biology is a multidisciplinary field, whereby experimental, theoretical, and computational methods are combined to provide wholistic descriptions of biomolecules. In this way, techniques can be combined to maximize their contributive strengths and minimize their respective weaknesses. In our original work, we combine structural biology techniques and AI methods to investigate the dynamics of macromolecular complexes. A summary follows.

In the second chapter we investigate the conformational changes of a phage encoded RNA polymerase during the process of promoter recognition. First, we solve the structure of three functional states of the enzyme (namely, a four-subunit core, a five-subunit holoenzyme, and a five-subunit holoenzyme in complex with DNA). From this structural data, we were able to develop a sequential model for promoter recognition. We then tested the predictions of this model using various in vitro assays, which verified our structure-derived hypotheses. Furthermore, we employed molecular dynamics to investigate the energetics of promoter binding and speculate on the ability of the RNAP to proceed from promoter recognition to elongation.

In the third chapter, we develop a novel method for investigating the conformational changes of large macromolecular structures(Fraser et al. 2021). In this case, we study a bacteriophage-like bacteriocin, R-type pyocin, which functions to breach bacterial membranes using a contractile sheath-rigid tube mechanism. Structures of the end states (both pre- and post-contraction) have been previously solved by Cryo-EM, but the set of intermediate structures between them or the contraction mechanism, remained poorly characterized. To investigate the contraction process, we developed a method, DMAD, which generates contraction intermediates from energetic and structural considerations. From this model we were able to make predictions on the energetics (such as the total energy of the extended state, the activation energy and the forces generated throughout) as well as the structure (when the tube comes into contract with the cell membrane and the "shape" of contraction intermediates). Furthermore, we made predictions on how the

energetics of the system would change because of mutations in key regions. To test our model, and the validity of our theoretical structural intermediates, we employed numerous biophysical assays to experimentally test the energetics, and to a lesser extent, the structural characteristics of the contraction process.

In the fourth chapter, we pursue an experimental route to the characterization of bacteriophage contraction intermediates using Cryo-EM. First, we obtained a Cryo-EM dataset with phage A511 incubated with cell membranes, such that contraction intermediates can be observed, albeit in significantly lower numbers than the pre- and post-contraction states. We then use refinement protocols to solve the structures of the pre-contraction, post-contraction, and intermediate sheaths. Finally, we compare the structures of the intermediate and contracted states to determine the relationship between subunit motions in the late stages of the contraction process and compare these to those predicted by the DMAD method.

Chapter 2 Template Strand Deoxyuridine Promoter Recognition by a Viral RNA Polymerase

The following chapter was reproduced with permission from:
Fraser, A., Sokolova, M. L., Drobysheva, A. V., Gordeeva, J. V., Borukhov, S., Artamonova, T. O., ... & Leiman, P. G. (2021). Template strand deoxyuridine promoter recognition by a viral RNA polymerase. *bioRxiv*.

ABSTRACT

Bacillus subtilis bacteriophage AR9 employs two strategies for efficient host takeover control and host defense evasion – it encodes two unique DNA-dependent RNA polymerases (RNAPs) that function at different stages of virus morphogenesis in the cell, and its double-stranded (ds) DNA genome contains uracils instead of thymines throughout. Unlike any known RNAP, the AR9 non-virion RNAP (nvRNAP), which transcribes late phage genes, can efficiently differentiate single instances of uracil to thymine substitutions in its promoters located in the template strand of dsDNA. Here, using structural data and a variety of *in vitro* transcription assays, we elucidate the basis for this unique promoter recognition mechanism. We show that the AR9 nvRNAP promoter specificity subunit gp226 is a homolog of bacterial σ factors and that the AR9 nvRNAP holoenzyme creates a canonical transcription bubble. Gp226, the nvRNAP core, and the template-strand promoter DNA motif interact to form two nucleotide base-accepting pockets whose shapes are incompatible with a thymine's C5 methyl group. The creation of this promoter-binding interface is preceded by a positional disorder-to-order transition of the gp226 N-terminal domain which itself interacts with the non-template DNA strand in the transcription bubble. Our work demonstrates the extent to which viruses can evolve new functional mechanisms to control acquired multisubunit cellular enzymes and make these enzymes serve their needs.

INTRODUCTION

Bacillus subtilis "jumbo" bacteriophage AR9 encodes two distinct multisubunit DNA-dependent RNA polymerases (RNAPs), allowing for the transcription of viral genes to proceed independently of the host RNAP(Lavysh et al. 2016)·(M. Sokolova et al. 2017)·(Lavysh et al. 2017)·(M. L. Sokolova, Misovetc, and Severinov 2020). The virion-packaged RNAP (vRNAP) is delivered into the host cell together with phage DNA at the onset of infection. The vRNAP then transcribes early phage genes, including those of the second, non-virion RNAP (nvRNAP). The nvRNAP transcribes late genes, including those coding for the vRNAP, which is packaged into progeny phage particles together with phage DNA.

The catalytically active AR9 nvRNAP core enzyme consists of four proteins that, when pairwise concatenated, show about 20% sequence identity and cover the entire lengths of the universally conserved β and β′ subunits of bacterial RNAPs (**Fig. 2.E1a**)(M. Sokolova et al. 2017). Promoter-specific transcription is performed by a five-subunit holoenzyme that in addition to the nvRNAP core contains the product of AR9 gene *226* (gp226)(M. Sokolova et al. 2017). Gp226 shows no sequence similarity to bacterial RNAP σ subunits, which mediate the recognition of promoters in bacteria, or any known transcription factor. Close orthologs of gp226 are found in the genomes of other jumbo phages that have been demonstrated or are presumed to contain uracil in their genomic DNA(Korn et al. 2021)·(Skurnik et al. 2012).

Unlike bacterial RNAPs(Bae et al. 2015), the AR9 nvRNAP holoenzyme recognizes promoters in the template strand of dsDNA and is capable of promoter-specific transcription initiation on single-stranded (ss) DNA(M. Sokolova et al. 2017). The AR9 nvRNAP template-strand promoter consensus 3′-$^{-11}$UUGU^{-8}-N_6-AU^{+1}-5′ (where N is any nucleotide and the transcription start site (TSS) coordinate is +1) contains a four-base long motif centered about 10 bases upstream of the TSS and two bases at the TSS (**Fig. 2.E1b**).

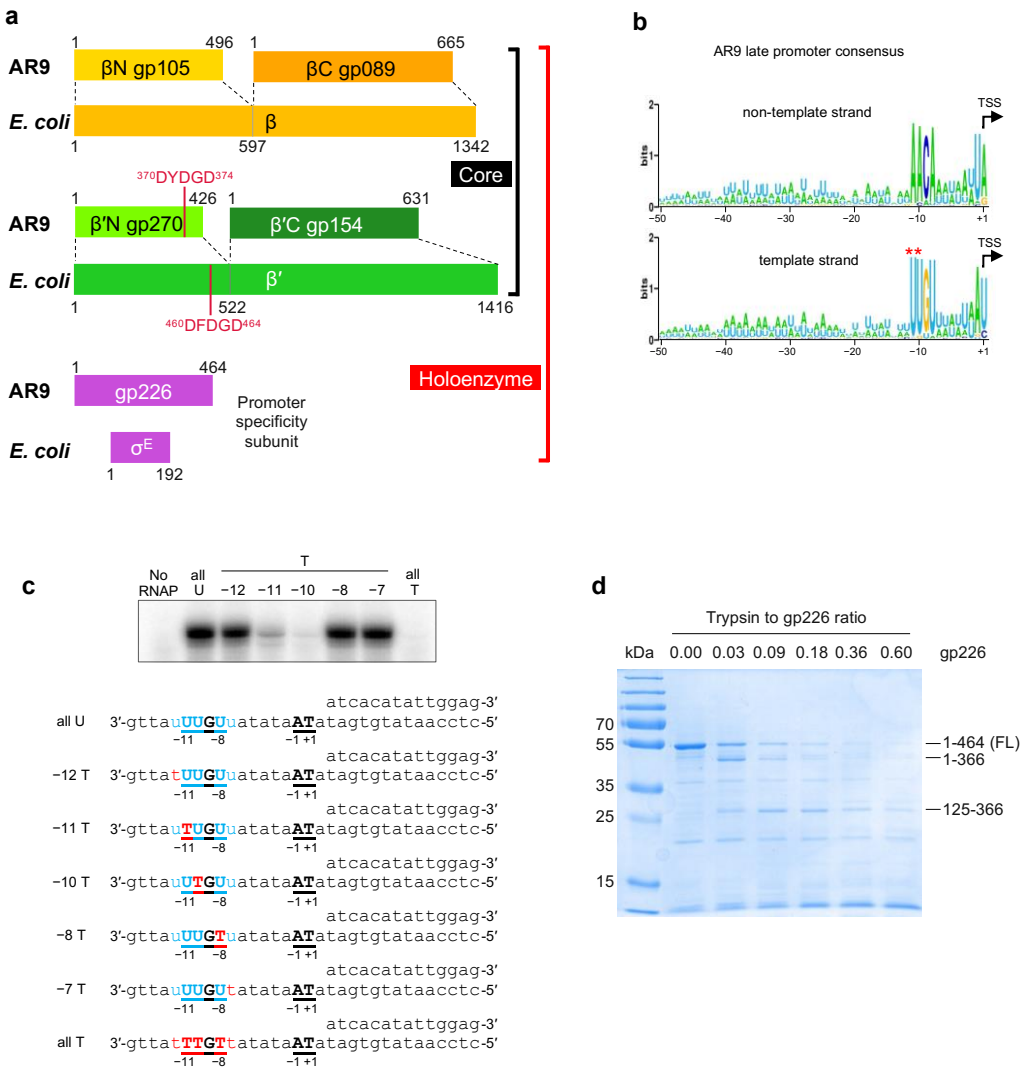


Figure. 2.E1. Organization and promoter consensus of the AR9 nvRNAP.

a, Organization of the catalytically active core and promoter initiation competent holoenzyme of the AR9 nvRNAP. A pair of genes encode a protein complex that is homologous to the bacterial subunits β or β′. The promoter specificity subunit displays no detectable sequence similarity to bacterial sigma factors. **b**, The consensus of AR9 late promoters recognized by the nvRNAP. Both DNA strands are shown. **c**, The dependance of the AR9 nvRNAP *in vitro* transcription activity on the position and number of T bases in the promoter, which is located in the template strand of DNA (bold-underlined). **d**, The resistance of recombinantly expressed gp226 to proteolysis by trypsin. The identities and

sizes of labeled major products, given as residue ranges, have been established using mass spectrometry. FL stands for the full-length protein. Two technical replicates of two biological replicates of the *in vitro* transcription and trypsin proteolysis experiments resulted in similar outcomes and one of them is shown.

Promoters with thymines in the -11^{th} and -10^{th} positions are inactive suggesting that the C5 position of the uracil's pyrimidine ring, which carries a methyl group in the thymine, plays a critical role in promoter recognition (**Fig. 2.E1c**). Despite possessing a short promoter consensus element, the AR9 nvRNAP holoenzyme protects an extensive region of DNA flanking the TSS (position −35 to +20 in the template strand and positions −29 to +17 in the non-template strand) from DNase I attack, implying additional contacts with DNA(M. Sokolova et al. 2017).

To understand the uracil-specific, template strand-dependent promoter recognition mechanism of the AR9 nvRNAP, we determined the structure of this enzyme by X-ray crystallography and cryo-electron microscopy (cryo-EM) in three states – the core, holoenzyme, and holoenzyme in complex with a 3′-overhang dsDNA oligonucleotide (historically called a "forked" or "fork" template) that mimicked the downstream half of the transcription bubble (**Fig. 2.1a**). Furthermore, we complemented this structural information by discriminative *in vitro* transcription assays. The 18 base-long ss part of the forked oligonucleotide contained the late AR9 promoter P077 while its 14 bp-long ds segment spanned positions from +3 to +16 relative to the TSS (**Fig. 2.1b**). Fortuitously, in the most populous class of the cryo-EM reconstruction and in both available crystal forms, the enzyme bound not one but two copies of this oligonucleotide – the downstream copy, as designed, and the upstream one – resulting in a superstructure that resembled the complete transcription bubble found in open complexes formed by other RNAPs. Moreover, *in crystallo* the nvRNAP molecules and the oligonucleotides formed a train in

which the upstream and downstream oligonucleotides belonging to two neighboring unit cells pi-stacked and formed a continuous double helix (**Fig. 2.E2**).

Structural similarities and differences between the AR9 nvRNAP and bacterial RNAPs

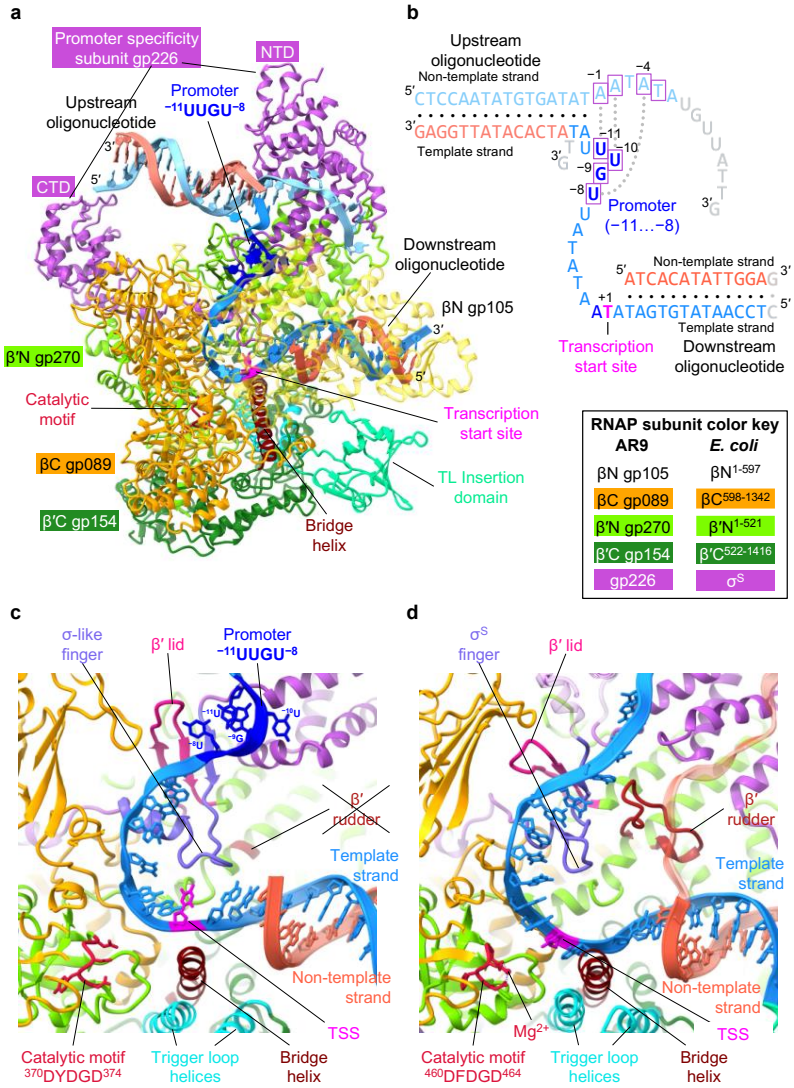


Figure. 2.1. Structure of the AR9 nvRNAP promoter complex.

a, Ribbon diagram of the crystal structure of the AR9 nvRNAP in complex with a 3′-overhang dsDNA (fork) oligonucleotide. Structural elements that are either unique to the AR9 nvRNAP or common to all RNAPs are labeled and color coded. The βN gp105 subunit is semitransparent for clarity.

b, Schematic of the two oligonucleotides that bound to one AR9 nvRNAP molecule resulting in a transcription bubble-like structure. Bases disordered in the crystal structure are rendered semitransparent. Bases in purple boxes interact with the protein.

c and **d**, Structure of the catalytic centers of the AR9 nvRNAP and *E. coli* RNAP-σ^S (PDB code 5IPM). Here and elsewhere, TSS stands for the transcription start site. The 2.4 region of σ^S, which is not present in gp226, is rendered semi-transparent. The *E. coli* RNAP-σ^S structure contains a short RNA product that is not shown for clarity. A part of the DNA non-template strand in the *E. coli* RNAP-σ^S structure is semitransparent for clarity.

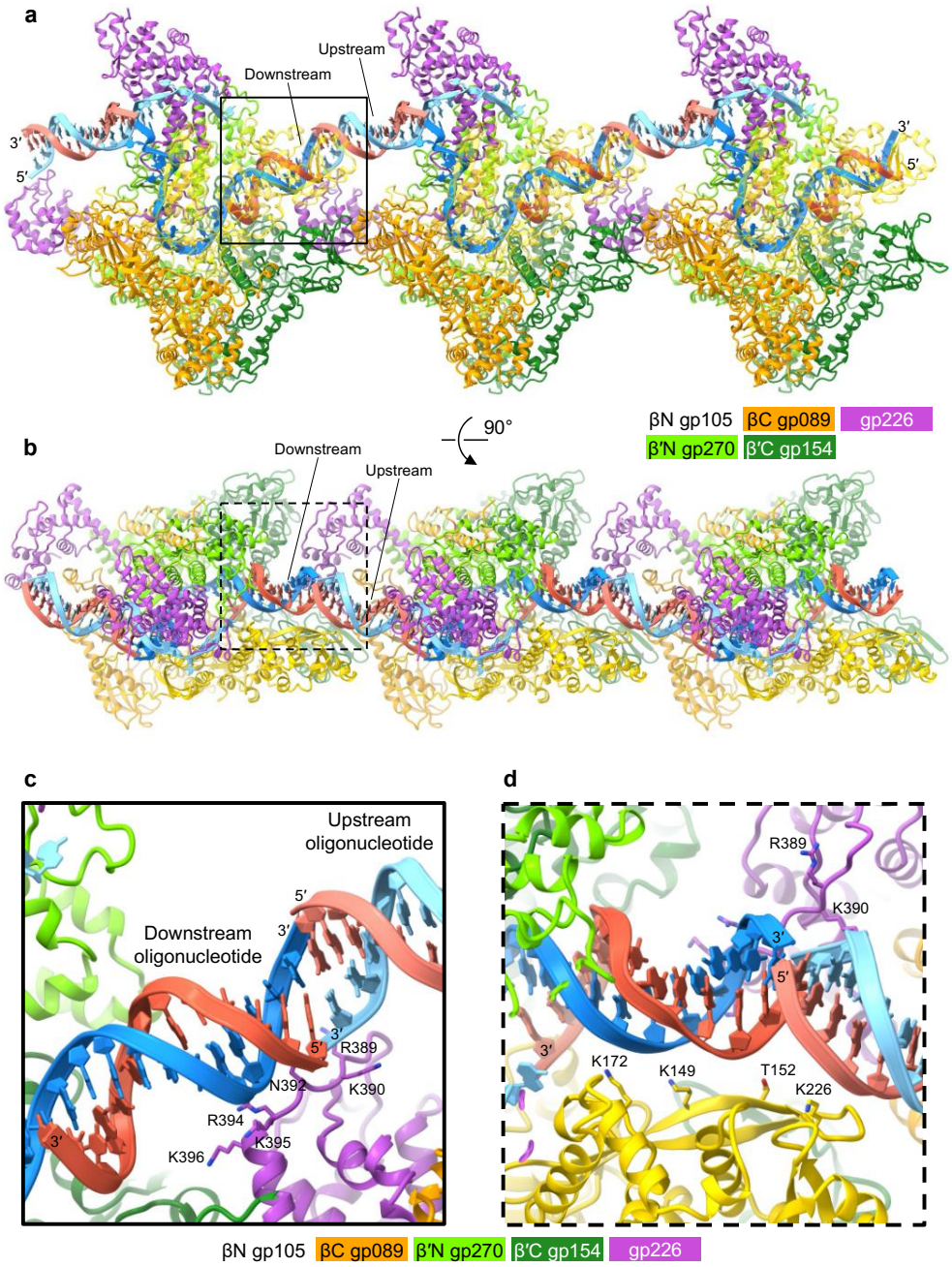


Figure. 2.E2. AR9 nvRNAPs and up- and downstream oligonucleotides form a train *in crystallo*.

a and **b**, Two orthogonal views of three nvRNAPs demonstrating the peculiar crystal packing.

c and **d**, Two areas of interest indicated with a solid and dashed line square in panels **a** and **b** that show details of pi-pi stacking interactions between the ends of the up- and

downstream oligonucleotides. Residues most proximal to the DNA are shown in a stick representation and labeled. In both panels, the color code is as in **Fig. 2.1a**.

The overall structure of the AR9 nvRNAP is a trimmed down version of a bacterial crab claw-shaped RNAP (**Fig. 2.1a**, **Fig. 2.E3**). No domain compensates for the absence of α and ω subunits that are present in all bacterial, eukaryotic, and archaeal enzymes. As a result, the AR9 nvRNAP claw is smaller and has a boxier appearance than that of its cellular counterparts. In bacterial enzymes, the α subunit dimer serves as a platform for the assembly of the β and β′ subunits(A 1981)·(Lane and Darst 2010)·(Minakhin et al. 2001). In the AR9 nvRNAP structure, the split site of the β′ subunit is spatially close to the putative β′-α^{II} interface (**Fig. 2.E3**), which suggests that this location likely represents a critical point for the formation of tertiary and quaternary structure.

Inside the catalytic cleft, the AR9 nvRNAP core contains all of the structural elements required for catalysis, stabilization of the open promoter complex, and promoter clearance found in multisubunit RNAPs(Lane and Darst 2010)·(Lee and Borukhov 2016), except for the β′ rudder (**Fig. 2.1c, 2.1d, Fig. 2.E3**). The β′ rudder is a twisted β-hairpin that is present in all known RNAPs. It extends from one of the β′ clamp α-helices and interacts with the RNA-DNA hybrid near the active site(Lane and Darst 2010). In bacterial RNAPs, deletion of the β′ rudder impairs promoter opening and destabilizes the elongation complex but does not affect the efficiency of transcription termination or the length of the RNA-DNA hybrid(Kuznedelov et al. 2002). The elongation complex of the AR9 nvRNAP must be stabilized by a different mechanism. The conformation of the ^{370}DYDGD^{374} catalytic motif of the AR9 nvRNAP, which is located near the C terminus of the β′N subunit gp270, is similar to that found in other RNAPs. The side chains of the three conserved aspartates are poised to bind a Mg^{2+} ion that is universally conserved in all nucleotidyltransferases, albeit the resolution of X-ray and cryo-EM data is insufficient for

resolving it (**Fig. 2.1c, 2.1d, Fig. 2.E3a**). The electron density of DNA at the TSS is also poor and the structural basis of high conservation of the TSS and the nucleotide preceding it (**Fig. 2.E1b**) cannot be derived.

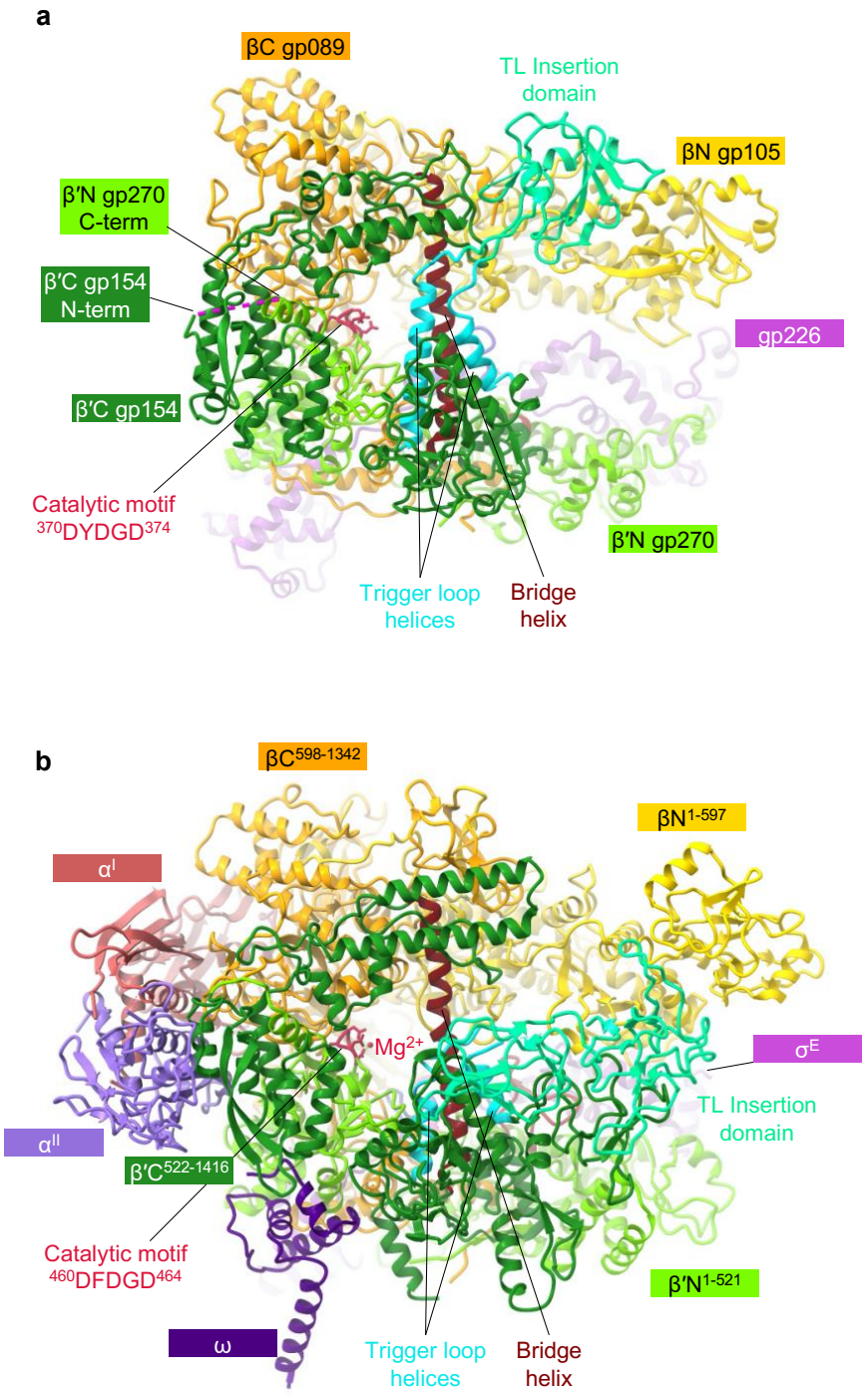


Figure. 2.E3. Comparison of the AR9 nvRNAP and *E. coli* RNAP-σ^{E} holoenzymes.

a and **b**, Ribbon diagrams of the AR9 nvRNAP and RNAP-σ^E (PDB code 6JBQ), respectively, are viewed from the NTP entrance channel. Nucleic acids are not shown for clarity. Gp226 and σ^E extend into the plane of the paper and are almost completely obscured by the depth-cueing effect. In both molecules, key elements are colored similarly and labeled. In panel **a**, the color code is as in **Fig. 2.1a**.

Similar to the *E. coli* RNAP(Campbell et al. 2002), the AR9 nvRNAP contains an insertion domain (residues 400-508 of β′C gp154) in the trigger loop (**Fig. 2.1a, Fig. 2.E3**). The trigger loop undergoes major conformational changes during the catalytic nucleotide addition cycle and template translocation(Lane and Darst 2010)·(Lee and Borukhov 2016), and the presence of an insertion domain in the *E. coli* system has not been fully reconciled with these transformations(Bao and Landick 2021). The fold of the AR9 nvRNAP insertion domain is different from that of the *E. coli* RNAP and, in fact, to any protein in the Protein Data Bank (PDB). Its sequence is also unique and found only in nvRNAPs of other jumbo phages(Korn et al. 2021)·(Skurnik et al. 2012). Its position in the structure of the nvRNAP core is also different from that of the *E. coli* RNAP. In the AR9 nvRNAP core structure, the insertion domain is located roughly in-between the β and β′ pincers where it partially obstructs the downstream DNA channel (**Fig. 2.E4a**). Furthermore, it carries a strong negative charge on its DNA-facing surface suggesting that it interacts with the downstream dsDNA in a non-sequence specific manner (**Fig. 2.E4b**). Out of 10 independent copies of the AR9 nvRNAP core molecules belonging to two different crystal forms, the insertion domain is ordered in only a singular instance. In the cryo-EM structure of DNA template-free holoenzyme, this domain is fully disordered (**Fig. 2.E5a, 2.E5b**). Considering the intrinsic propensity of this domain to large motions, it may participate in translocation by sliding on the DNA and exerting a force on the trigger loop.

The structure of promoter-specificity subunit gp226

The AR9 nvRNAP promoter-specificity subunit gp226 consists of two globular domains – a larger N-terminal domain (NTD, residues 1-264) and a smaller C-terminal domain (CTD, residues 295-464) – connected by a linker (**Fig. 2.1a, Fig. 2.E6a**).

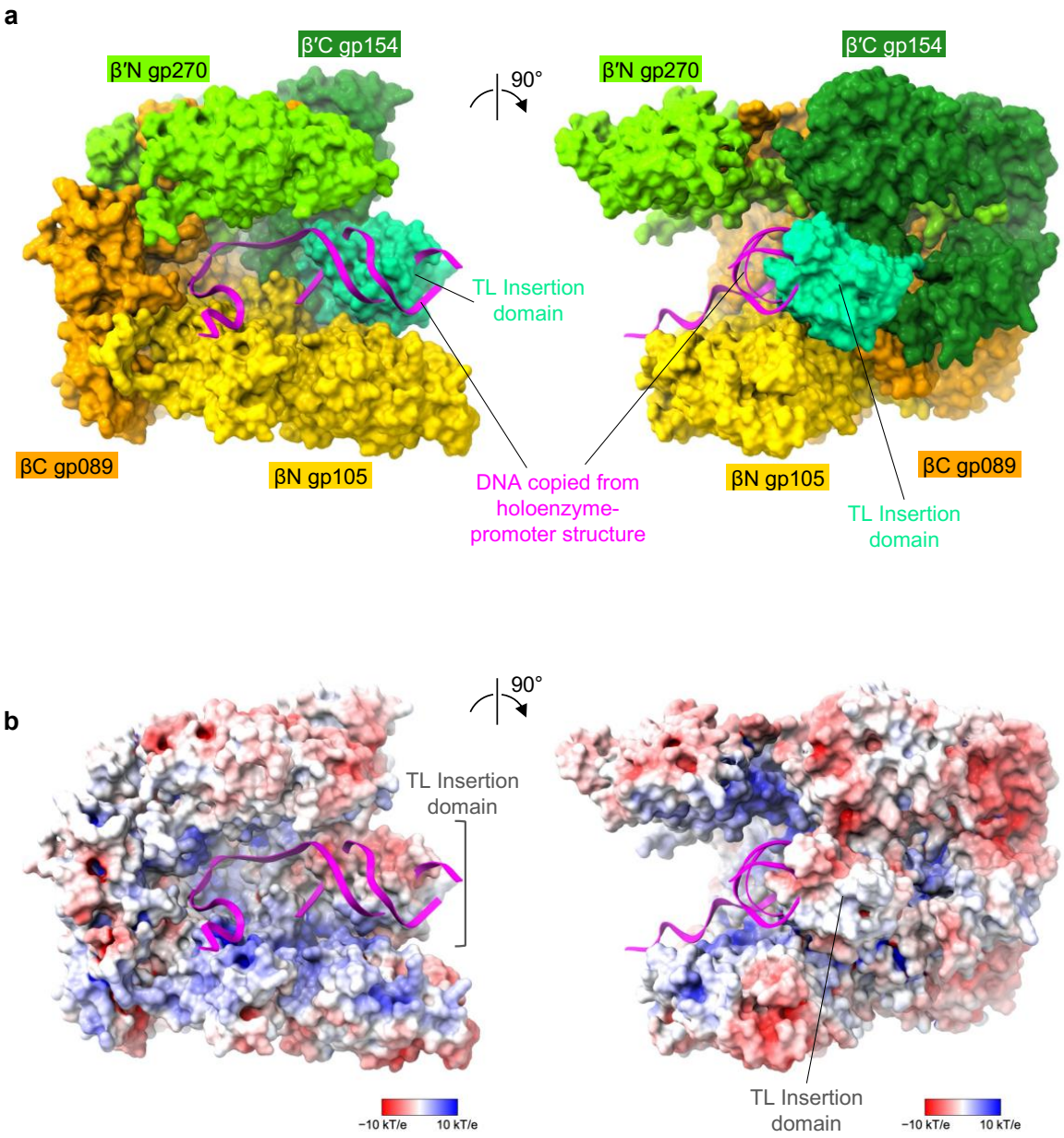


Figure. 2.E4. Structure of the AR9 nvRNAP core.

a, Molecular surface of the AR9 nvRNAP core crystal structure, with subunits colored as in **Fig. 2.1a**, with the downstream DNA oligonucleotide copied from the holoenzyme-promoter structure. The insertion domain partially obstructs the DNA binding cleft.
b, Electrostatic potential is mapped onto the molecular surface of the AR9 nvRNAP core. The orientations are identical to panels **a**.

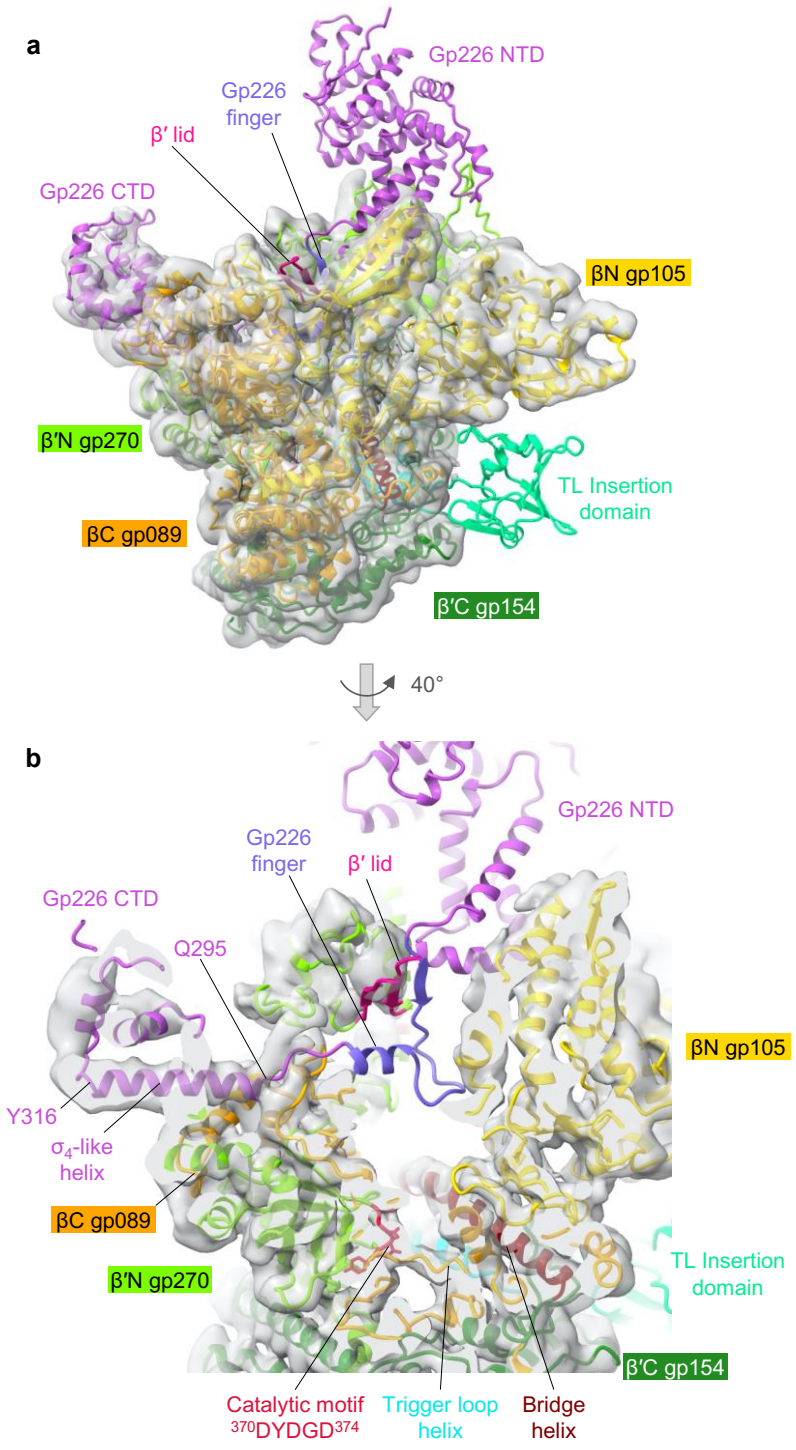


Figure. 2.E5. Cryo-EM structure of the AR9 nvRNAP holoenzyme.

a, Cryo-EM map of the AR9 nvRNAP holoenzyme contoured at 4.0 std dev above the mean (semitransparent gray) with the fitted atomic model of the promoter complex (sans DNA) colored as in **Fig. 2.1a**.

b, A zoomed-in view of the catalytic cleft demonstrating the degree of gp226 disorder.

Gp226 interacts with the nvRNAP core in a manner resembling that of bacterial σ factors(Bae et al. 2015)·(Lee and Borukhov 2016)·(B. Liu, Zuo, and Steitz 2016)·(Fang et al. 2019), and all elements that come in contact with the body of the enzyme have structural counterparts in bacterial σ factors. Residues 184-264 of gp226 fold into a σ_2-like domain (**Fig. 2.E6a**), which interacts in a sequence specific manner with the non-template strand of the −10 promoter element in bacterial σ factors and comprises their most conserved part(Feklistov and Darst 2011)·(Feklístov et al. 2014)·(Paget 2015) (**Fig. 2.E6b**). Gp226 residues 265-294 form a σ finger-like structure that invades the catalytic cleft and forms an augmented β-sheet with the β′ lid (**Fig. 2.1c**, **Fig. 2.E6a**). Gp226 residues 295-316 comprise an α-helix that matches the N-terminal α-helix of the σ_4 domain (**Fig. 2.E6a, 2.E6b**).

The most similar σ factor with a known structure, the *E. coli* σ^E, displays a Cα-Cα root mean square deviation (RMSD) of 4.2 Å and a sequence identity of 9.6% when superimposed onto residues 184-316 of gp226 which comprise its σ_2-, finger- and σ_4-like elements. Thus, the σ-like part of gp226 spans a contiguous region of 133 residues and is nearly equal in size to the entire σ^E structure (**Fig. 2.E6a, 2.E6b**). Considering the structural similarity of gp226 to the conserved part of bacterial σ factors, their similar functions, and the similar manner by which they interact with the template DNA (the structure of the transcription bubble), it follows that gp226 is a homolog of bacterial σ factors, despite their low sequence identity. The peripheral parts of the gp226 NTD and CTD have been replaced

with new folds but all elements that interact with the body of the enzyme and with DNA (albeit with some modifications for the latter) have been retained.

A Protein Data Bank-wide search(Holm and Laakso 2016) for folds resembling that of the gp226 NTD and CTD resulted in a single definitive match. The CTD of gp68, a subunit of the phage phiKZ non-virion RNAP which is required for both promoter recognition and transcription elongation(Garrido et al. 2021),(Yakunina et al. 2015), can be superimposed onto the gp226 CTD with an RMSD of 3.2 Å for 142 equivalent Cα atoms (out of 173) and a sequence identity of 11%. As the rest of the gp68 structure (residues 1-303) is disordered, we used AlphaFold(Jumper et al. 2021) Colab to model it. The local distance difference test(Mariani et al. 2013) of this model for residues 1-277 was 85.3, indicating a very high level of confidence. This model can be superimposed onto the gp226 NTD with an RMSD of 3.0 Å for 198 equivalent Cα atoms (out of 277) and a sequence identity of 8.6%. Thus, even though gp226 is as divergent from gp68 sequence-wise as it is from bacterial σ factors, their similarities in structure and location within the RNAP holoenzyme complex suggest that all these proteins have a common ancestor.

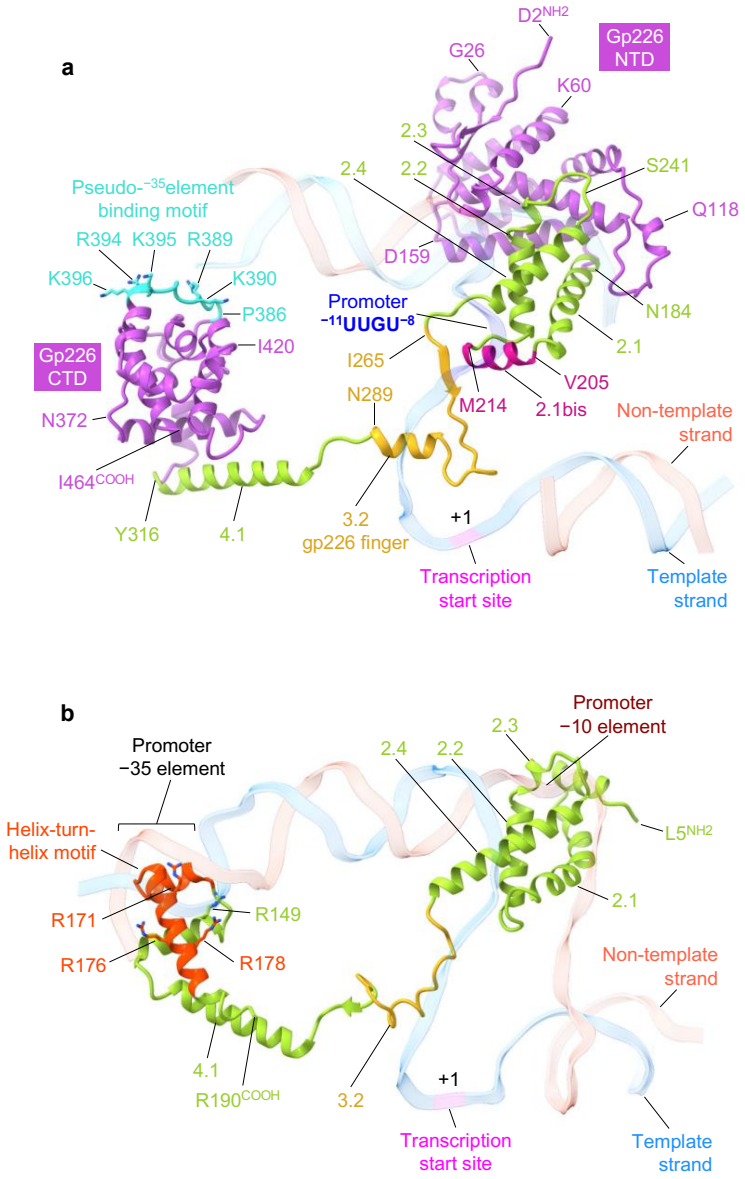


Figure. 2.E6. Structure of the promoter specificity subunit gp226.

a, Ribbon diagram of gp226 with regions structurally similar to bacterial σ factors colored in yellow green (helices 2.1 through 2.4 and 4.1) and gold (finger). The unique N- and C-terminal domains colored medium orchid. Residue numbers and identities are given at key locations. The DNA strands are colored as in **Fig. 2.1a** and **2.1b** and are semitransparent. The pseudo^{-35}-element binding motif is turquoise, and its positively charged and solvent exposed residues are shown in a stick representation.

b, Ribbon diagram of the *E. coli* σ^E factor (PDB code 6JBQ) with its helix-turn-helix motif colored in orange red. Positively charged residues that interact with the DNA are shown in a stick representation and labeled. The DNA backbone is semitransparent.

Structural adaptations of gp226 required for template strand promoter recognition

The most conserved parts of bacterial σ factors, the σ_2- and σ finger elements, are also present in gp226 (**Fig. 2.E6**). With small adaptations, these elements are responsible for the unique mode of promoter recognition through template DNA strand. A tight turn connecting helices 2.1 and 2.2 in bacterial σ factors is replaced by a short α-helix in gp226 (residues 205-214). This 2.1bis helix creates a bridge linking the two pincers of the AR9 nvRNAP claw and together with the βN gp105 subunit comprises a binding site for the promoter in the template strand of DNA (**Fig. 2.E6a**). The gp226 finger forms an augmented β sheet with the β′ lid (residues 161-177 of β′N gp270) such that the β′ lid reaches the −8 position uracil base of the 3′-$^{-11}$UUGU^{-8}-5′ promoter motif and tucks it in against the 2.1bix helix (**Fig. 2.1c, 2.1d**). The β′ lid of AR9 nvRNAP is three residues longer than its bacterial counterpart, which further enhances and facilitates this interaction (**Fig. 2.1c**).

Promoter DNA structure and the design of a T-specific enzyme

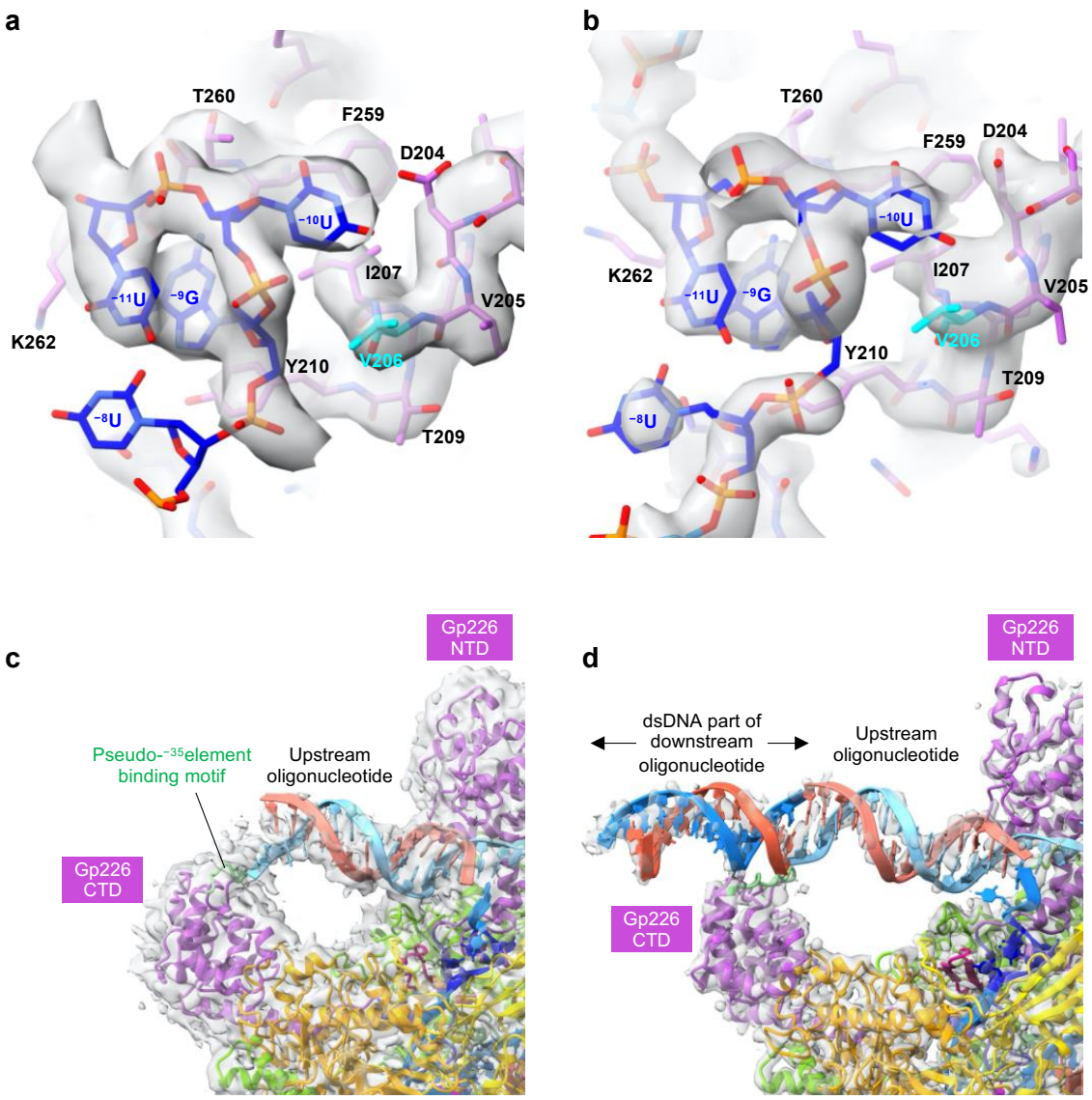


Figure. 2.E7. Electron density of the AR9 nvRNAP-DNA interacting regions.

a and **b**, Cryo-EM and composite omit X-ray electron density maps of the promoter binding pocket with refined atomic models, respectively. The cryo-EM and X-ray maps are contoured at 5.0 and 1.5 std dev above the mean, respectively. The carbon atoms are colored as in **Fig. 2.1a**. V206, which is critical for U specificity, is colored cyan.

c and **d**, Cryo-EM and composite omit X-ray electron density maps of the gp226 CTD with refined atomic models, respectively. The cryo-EM and X-ray maps are contoured at 2.0 and 1.0 std dev above the mean, respectively. The ds segment of the downstream oligonucleotide belonging to a neighboring unit cell is shown in the X-ray map (see **Fig.**

2.E2). Proteins are colored as in **Fig. 2.1a**. The pseudo-$^{-35}$element binding motif is colored light green.

The structure of the 3′-$^{-11}$UUGU^{-8}-5′ template strand promoter motif is well resolved in both cryo-EM and X-ray electron density maps, although $^{-8}$U is partially disordered in the cryo-EM map (**Fig. 2.E7a, 2.E7b).** The most critical and obligatory $^{-10}$U base, the replacement of which by a T leads to the abolishment of promoter-specific transcription(M. Sokolova et al. 2017) (**Fig. 2.E1c**), is buried in a deep pocket at the interface of the gp226 2.1bis helix and βN gp105 (**Fig. 2.2a**). In this pocket, the $^{-10}$U base is wedged between the side chains of gp226 I207 and βN gp105 R363, forming a stacking interaction with the latter. Its Watson-Crick interface forms hydrogen bonds with the side chain of βN gp105 K375 and with the main chain N of gp226 V206 and I207. Most importantly, the C5 atom of the $^{-10}$U pyrimidine ring is only 3.9 Å away from the Cβ of V206, suggesting that a C5 position methyl group would clash with the V206 side chain (**Fig. 2.2a**). Accordingly, a holoenzyme containing gp226 with a V206G substitution recognized $^{-10}$U- and $^{-10}$T-containing promoters with equal efficiencies (**Fig. 2.2b**). Notably, all close homologs of gp226 proteins in jumbo phages with deoxyuridine-containing genomic DNA(Korn et al. 2021),(Skurnik et al. 2012) display high sequence conservation of the 2.1bis helix with the critical valine being absolutely conserved. This suggests that these phages employ a similar mechanism for uracil-dependent promoter recognition.

The requirement of U vs. T in the −11th position of the promoter is nearly as strong as in the −10th position. Additionally, a G is required in the −9th position(M. Sokolova et al. 2017). However, the enzyme displays almost no U vs. T preference in the −8th position (**Fig. 2.E1c)**. In the promoter complex, the bases of $^{-11}$U and $^{-9}$G form a stacking interaction such that the C5 and C6 atoms of $^{-11}$U butt against the sugar-phosphate backbone of the

$^{-11}$UUG^{-9} segment, leaving no space for a C5 position methyl group (**Fig. 2.2a**). There is a stacking interaction between $^{-9}$G and the phenol ring of gp226 Y210 (which belongs to the 2.1bis helix), and there are three hydrogen bonds between the Watson-Crick interface of $^{-9}$G and the main chain of gp226 residues F261 and Y263, which provides a rationale for the G requirement in this position. $^{-8}$U forms one hydrogen bond with $^{-11}$U and one hydrogen bond with the tip of the β′ lid, which is longer than in its bacterial counterparts, as described above. The C5 position of $^{-8}$U points into solution and can accommodate the additional methyl group of T (**Fig. 2.2a**).

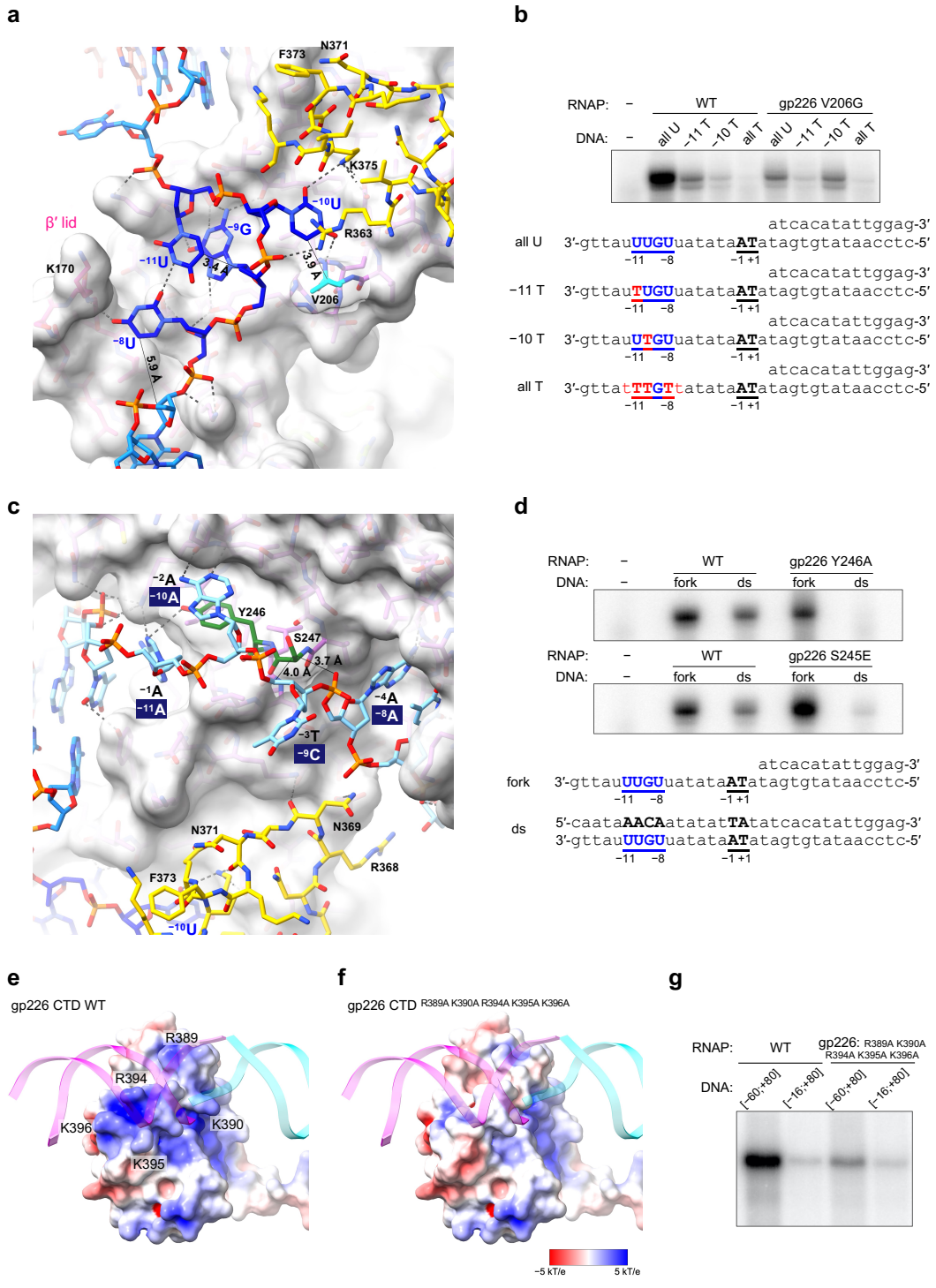


Figure. 2.2. Interaction of the AR9 nvRNAP with DNA in the promoter complex.

a, Atomic model of the AR9 nvRNAP promoter recognition element. Gp226 is shown as a semitransparent molecular surface. Only a small fragment of βN gp105 that participates in the formation of the $^{-10}$U binding pocket is shown for clarity. Interchain and DNA-intrachain hydrogen bonds are shown as dashed lines. Thin, straight lines connect the C5 atom of the uracil pyrimidine ring to the closest protein or DNA atoms which lie in-plane with the ring. The carbon atoms are colored as in **Fig. 2.1a**, except for V206, which is shown in cyan.

b, The *in vitro* transcription activity of the AR9 nvRNAP holoenzyme containing the wild type (WT) gp226 or V206G gp226 mutant have been tested using various U- and T-containing templates.

c, Interaction of the upstream oligonucleotide with the gp226 NTD. The same gp226 and βN gp105 fragments as in **Fig. 2.2a** are shown but are tilted to improve clarity. The nucleotides are numbered according to the original nomenclature of the template strand (see **Fig. 2.1b**). The putative base identities and their numbers relative to the TSS as would be found in a dsDNA transcription bubble are given in a midnight blue colored boxes.

d, The *in vitro* transcription activity of the AR9 nvRNAP holoenzyme containing gp226 mutants with an altered structure of the non-template strand binding groove.

e and **f**, The surface electrostatic potentials of the pseudo^{-35}-element binding motif in the WT gp226 and A^5 gp226 mutant.

g, The ssDNA, and dsDNA *in vitro* transcription activities of the AR9 nvRNAP holoenzyme containing the WT gp226 and A^5 gp226 mutant.

For each *in vitro* transcription experiment, two technical replicates of two biological replicates resulted in similar outcomes and one of them is shown.

Interaction with the non-template strand is essential for the transcription of dsDNA

The gp226 NTD displays several deep pockets that capture the ss part of the upstream oligonucleotide, which mimics the non-template strand of the transcription bubble (**Fig. 2.2c**). In the transcription bubble (**Fig. 2.1b**), this part of the non-template strand must have a sequence complementary to the template strand promoter motif (5′-$^{-11}$AACA^{-8}-3′, the numbering is relative to the TSS). Our oligonucleotide contained a similar motif in its ss part (5′-$^{-1}$AATA^{-4}-3′, the numbering is as in the downstream nucleotide, **Fig. 2.1b**). Together with its neighboring bases this sequence matched the appearance of the electron density. The base of the mismatched third position nucleotide ($^{-3}$T↔$^{-9}$C) does not interact with the gp226 NTD but instead protrudes into solution (**Fig. 2.2c**). This motif, despite being only partially complementary to the promoter, was likely a key determinant in the fortuitous binding of the upstream oligonucleotide.

The gp226 NTD interacts with the backbone and bases of the non-template DNA strand of the transcription bubble via pi-pi stacking, ion pairs, and hydrogen bonds. The length of this interface exceeds 30 Å. The extent of these interactions suggests that they play an important role in promoter recognition of a dsDNA template. Indeed, a Y246A substitution, which eliminated pi-pi stacking between the side chain of Y246 and the $^{-2}$A base of the $^{-1}$AATA^{-4} motif, abolished transcription on dsDNA but did not affect transcription on a fork template (**Fig. 2.2c, 2.2d**). Furthermore, a S245E substitution introduced a large, negatively charged side chain on the surface of the gp226 NTD that interfered with the trace and conformation of the sugar-phosphate backbone between $^{-2}$A and $^{-4}$A. As a consequence, the transcriptional activity of the holoenzyme containing the S245E gp226 mutant on a dsDNA template was weak, whereas its fork template activity was at or above that of WT (**Fig. 2.2d**).

In the cryo-EM structure of the DNA template-free AR9 nvRNAP holoenzyme, the NTD and σ-like finger of gp226 are disordered (**Fig. 2.E5a, 2.E5b**). Similarly to the TL insertion domain, the disorder is likely due to positional heterogeneity since both the gp226

NTD and TL insertion domain possess well-defined hydrophobic cores and are folded in other states of the nvRNAP complex. Furthermore, the NTD is resistant to proteolysis by trypsin in gp226 recombinantly expressed on its own (**Fig. 2.1d**). The order-disorder transitions of the gp226 NTD and σ-like finger play a role in the promoter recognition mechanism described below.

Gp226 CTD interacts with DNA in a non-sequence specific manner

Although weak, the cryo-EM and X-ray electron density of the upstream oligonucleotide stretches from the σ_2-like part of the gp226 NTD to the gp226 CTD. This interaction is about 35 DNA base pairs upstream from the TSS (**Fig. 2.E6a**) drawing a parallel to the −35 consensus element of bipartite bacterial promoters. The recognition of the −35 element by bacterial σ factors is mediated by a helix-turn-helix motif(Brennan and Matthews 1989), which interacts with the major groove of dsDNA(Campbell et al. 2002) in a sequence specific manner (**Fig. 2.E6b**). AR9 nvRNAP promoters, however, display no sequence conservation in this region (**Fig. 2.E1b**) and, accordingly, the gp226 CTD interacts with the minor groove of dsDNA which displays few sequence-specific features in the B form(Rohs et al. 2010). Furthermore, this interaction is mediated by a scrunched β-strand that carries several positively charged residues (R389, K390, R394, K395, K396), but not by a helix-turn-helix motif, which is absent from the gp226 structure. We termed the DNA interacting element of the gp226 CTD (amino acids 386-395) a pseudo-$^{-35}$element-binding motif.

To examine the role of the pseudo-$^{-35}$element-binding motif in promoter recognition, we removed most of the positive charge displayed on its surface by replacing R389, K390, R394, K395, K396 of gp226 with alanines (we called this mutant A^5) (**Fig. 2.2e, 2.2f**). As the AR9 nvRNAP holoenzyme containing A^5 gp226 had a lower activity

overall, we compared its activity to that of the wild type (WT) holoenzyme on two dsDNA templates that either contained or lacked the upstream part required for interaction with the pseudo-$^{-35}$element-binding motif (the [−60,+80] and [−16,+80] templates in **Fig. 2.2g**, respectively). The WT enzyme exhibited a greater decrease in activity on the [−16,+80] template compared to that of the mutant, which shows that the pseudo-$^{-35}$element-binding motif is essential for optimal promoter recognition.

The AR9 gp226 pseudo-$^{-35}$element-binding motif maps onto a disordered part of phage phiKZ gp68(Garrido et al. 2021) (residues 413-429). This region of gp68's surface carries a positive charge, akin to that of AR9 gp226 (**Fig. 2.2e**), even without the inclusion of disordered residues in electrostatic potential calculations. Again, analogous to the AR9 nvRNAP, phiKZ nvRNAP promoters show no sequence conservation 35 bases upstream of the TSS(Ceyssens et al. 2014). Considering i) the homology of the phiKZ and AR9 nvRNAPs to cellular RNAPs(Garrido et al. 2021), ii) the homology of AR9 gp226 to phiKZ gp68 described above, and iii) the presence of positive charges at equivalent locations on the surface of the AR9 gp226 and phiKZ gp68 CTDs, the phiKZ nvRNAP is thus likely to form an AR9 nvRNAP-like transcription bubble in which the CTD of gp68 participates in the binding of upstream dsDNA. Furthermore, as all these properties are seemingly conserved for such distantly related viruses as AR9 and phiKZ that infect unrelated hosts (Gram positive *B. subtilis* and Gram negative *Pseudomonas aeruginosa*, respectively), and have different genomic DNA base composition(Lavysh et al. 2016), (Mesyanzhinov et al. 2002) and genome replication strategies(Chaikeeratisak et al. 2017), the supposition of AR9 nvRNAP-like transcription bubbles can be extended to nvRNAPs of all jumbo phages.

Free energy of template strand promoter binding

To reconcile the tight integration of AR9 nvRNAP promoter DNA into the promoter complex with the transient nature of this complex (**Fig. 2.1a, 2.1c, 2.2a**) we examined the binding free energy of the three best-ordered promoter bases 3′-$^{-11}$UUG^{-9}-5′ to the AR9 nvRNAP holoenzyme by executing a double decoupling method molecular dynamics protocol(Woo and Roux 2005)·(Gumbart Benoˆıt, Roux, and Chipot 2018)·(Gumbart, Roux, and Chipot 2012)·(Phillips et al. 2005). The procedure assumes that the conformation of the enzyme does not appreciably change upon promoter binding. As such, the simulations describe a state in which the gp226 NTD has associated with the AR9 nvRNAP core.

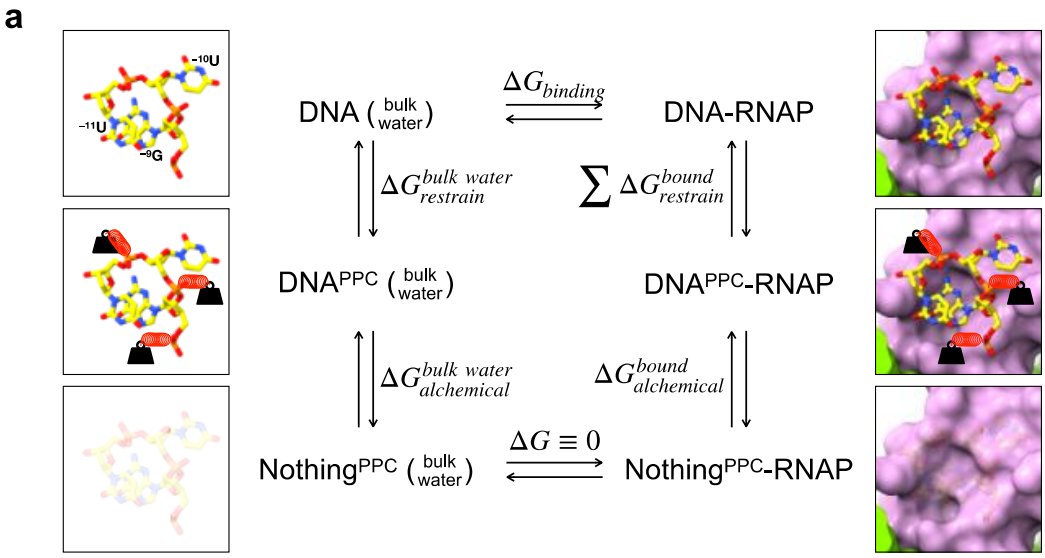
The standard binding free energy was calculated by combining the results of four separate simulations that corresponded to the vertical reactions in the thermodynamic cycle shown in **Fig. 2.E8a**. After the equilibration of the system (**Fig. 2.E8b, 2.E8c**), two types of simulations were performed: i) "alchemical transformations" in which the occupancy of the oligonucleotide located either in the promoter pocket ($\Delta G^{bound}_{alchemical}$) or in bulk water ($\Delta G^{bulk\ water}_{alchemical}$) was reduced to zero while the oligonucleotide was harmonically constrained to maintain the promoter pocket bound conformation (**Fig. 2.E8d, 2.E8e**), and ii) calculations of the entropic cost of such harmonic constraints for a oligonucleotide located in the promoter pocket ($\Delta G^{bound}_{restrain}$) (**Fig. 2.E8f-8l**) and in bulk water ($\Delta G^{bulk\ water}_{restrain}$) (**Fig. 2.E8m**). To ensure reproducibility and to minimize bias, all simulations were run bidirectionally.

The favorable energetics of promoter binding via alchemical transformation (−12.7±2.3 kcal/mol, **Fig. 2.E8d, 2.E8e**) are partially offset by the unfavorable entropic contributions of constraints on DNA conformation and position (5.8±1.5 kcal/mol, **Fig. 2.E8f-8m**). The resulting free energy gain upon complex formation is −6.9±2.8 kcal/mol, which shows that the interaction of this promoter element with the enzyme is fairly weak. Thus, despite its unusual structure in which the $^{-10}$U is buried in a deep pocket and $^{-9}$G and

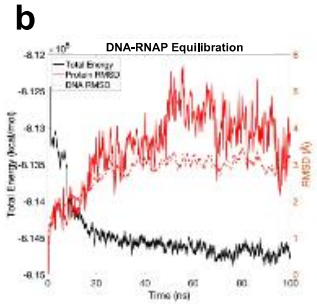
$^{-11}$U form a stacking interaction, the promoter complex is transient, and the enzyme can easily proceed towards elongation.

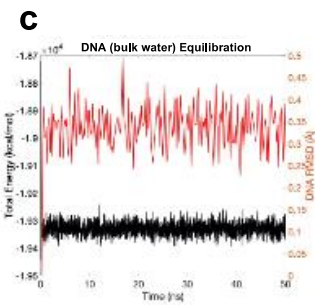
The mechanism of template strand promoter recognition in dsDNA

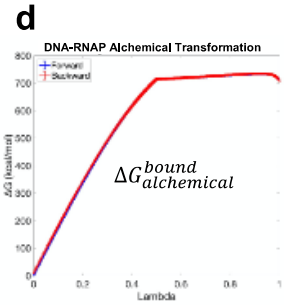
Combining these findings, we propose the following model for promoter recognition by the AR9 nvRNAP (**Fig. 2.3**). In the free state of the AR9 nvRNAP holoenzyme molecule, the NTD of gp226 is folded but does not interact with the body of the enzyme (it is positionally disordered or mobile) and the promoter-binding pocket is absent (**Fig. 2.E5**). The NTD of gp226 is attached to the CTD and the core via a linker that will eventually form a σ finger-like structure in the promoter complex. The enzyme displays two positively charged patches that have DNA binding propensity on their surface – the pseudo-$^{-35}$element-binding motif and a patch with a much stronger positive charge and better shape complementarity for the binding of a non-template strand motif that is complementary to the promoter (**Fig. 2.E9** and **State 1 in Fig. 2.3**).

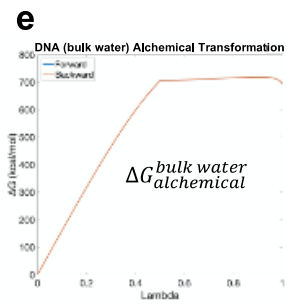


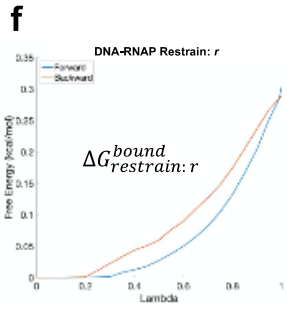
$$\Delta G_{binding} = \Delta G_{restrain}^{bulk\ water} + \Delta G_{alchemical}^{bulk\ water} - \Delta G_{alchemical}^{bound} - \sum \Delta G_{restrain}^{bound}$$

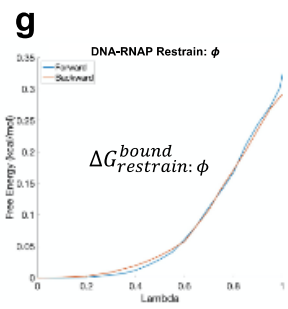


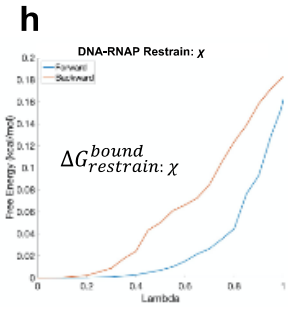


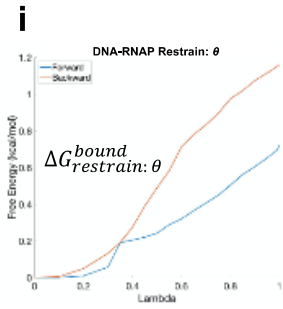


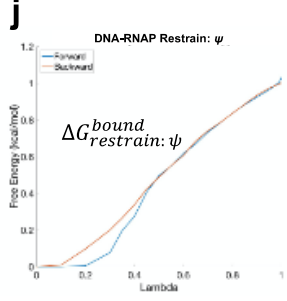


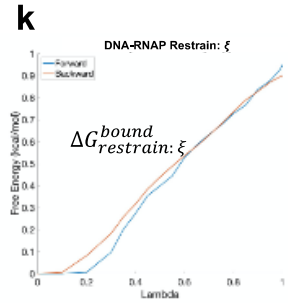


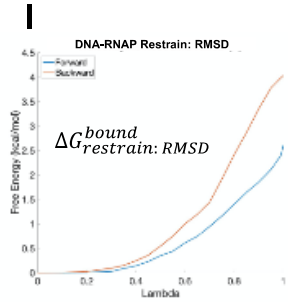












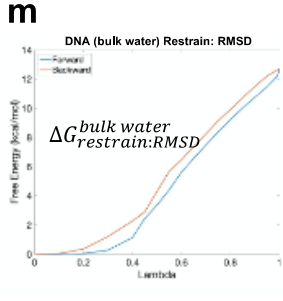


Figure. 2.E8. Derivation of the promoter binding free energy using molecular dynamics.

a, Thermodynamic cycle of promoter binding. The PPC superscript (e.g. DNA^{PPC}) stands for Promoter Pocket Conformation in regard to the structure of the 3′-$^{-11}$UUG^{-9}-5′ DNA trinucleotide.

b, Equilibration and relaxation of the cryo-EM derived atomic model of the AR9 nvRNAP holoenzyme with the 3′-$^{-11}$UUG^{-9}-5′ DNA trinucleotide bound to the promoter pocket.

c, Equilibration and relaxation of the 3′-$^{-11}$UUG^{-9}-5′ trinucleotide in bulk water.

d and **e**, Energetics of forward and backward alchemical transformations of the 3′-$^{-11}$UUG^{-9}-5′ DNA trinucleotide in the promoter pocket of the AR9 nvRNAP holoenzyme and in the PPC in bulk water, respectively.

f, g, h, i, j, k, l, Entropic cost of applying seven harmonic constraints to the 3′-$^{-11}$UUG^{-9}-5′ DNA trinucleotide to maintain it in the promoter pocket-bound state.

m, Entropic cost of the harmonic RMSD constraint on the 3′-$^{-11}$UUG^{-9}-5′ DNA trinucleotide to maintain the PPC.

The process of promoter recognition contains the following steps. 1) The gp226 NTD partially melts the template dsDNA and captures the non-template strand in a groove on its surface by recognizing a motif that may be only partially complementary to the promoter (**State 2 in Fig. 2.3**). Both events are facilitated by the high AU content of promoter-containing regions as they are likely to display transiently flipped out bases. Only three bases of this motif interact with the protein (the third position base $^{-9}$C does not), which makes the motif very common and may allow for the enzyme to scan the template. 2) The pseudo-$^{-35}$element-binding motif of the gp226 CTD interacts with dsDNA, reducing the conformational space available to the gp226 NTD and promoting its binding to the body of the enzyme (**State 3 in Fig. 2.3**). 3) The NTD of gp226 comes in contact with the

body of the enzyme, fully separating the DNA strands, forming a σ finger-like element and a transcription bubble, and placing the template strand at the [gp226]:[βN gp105] interface. This interface captures a flipped out $^{-10}$U base and buries it into the $^{-10}$U recognition pocket. Simultaneously, the DNA strand is squeezed slightly such that the bases that flank the flipped-out $^{-10}$U base form a stack and the identities of the stacked $^{-9}$G and $^{-11}$U bases are verified via geometry-sensitive interactions (hydrogen bonds and ion pairs). Additional interactions are formed at the catalytic center where the TSS is recognized (**State 4 in Fig. 2.3**). As the free energy of promoter recognition is nevertheless reasonably low and the conformation of the sugar-phosphate backbone for the four bases of the promoter motif is close to that of dsDNA, the enzyme can efficiently proceed with elongation.

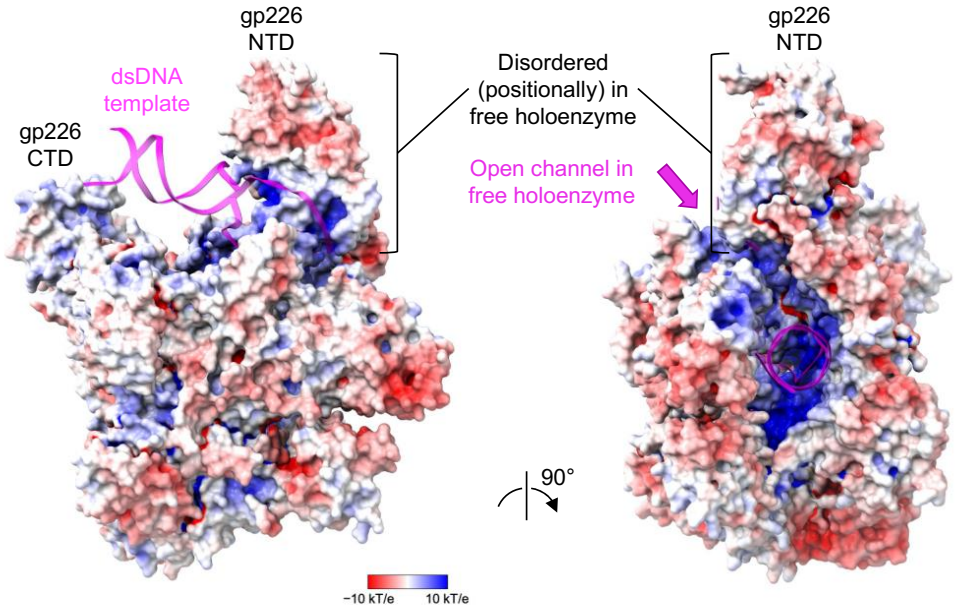


Figure. 2.E9. Distribution of electrostatic potential on the surface of the AR9 nvRNAP promoter complex.

The DNA (colored magenta) was excluded from the calculations. The orientation of the molecule is as in **Fig. 2.1a**.

CONCLUSION

Here we have explained the functional mechanism of a phage-encoded RNAP that contains a unique promoter-specificity subunit, recognizes the promoter in the template strand of DNA, requires uracil bases in the promoter, and does not use a common helix-turn-helix motif for the binding of dsDNA. Even though the AR9 nvRNAP and its promoter specificity subunit have a common ancestor with their bacterial counterparts, the extent to which the AR9 nvRNAP promoter recognition mechanism is different from any known RNAP shows that our knowledge of the structure and function of these nanomachines is far from complete.

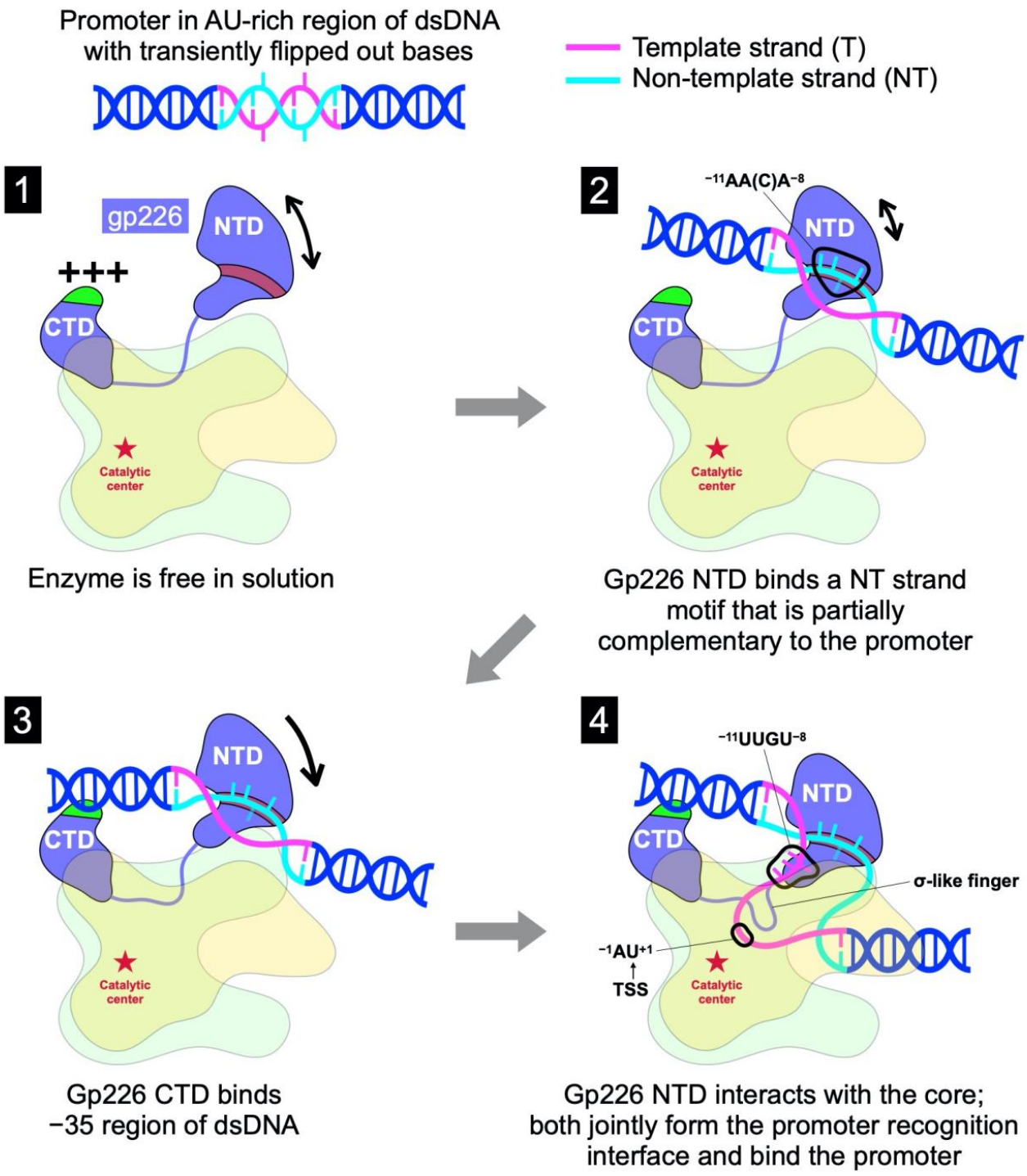


Figure. 2.3. The mechanism of template strand promoter recognition in dsDNA.

For clarity, both proteins comprising the N- and C-terminal parts of the β and β′ subunits are shown in the same color (light yellow for β and light green for β′). The putatively

degenerate, third position base $^{-9}$C of the promoter-complementary motif $^{-11}$AA(C)A^{-8} is shown as a semitransparent rod. See the main text for full explanation.

METHODS AND MATERIALS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Cloning of the AR9 nvRNAP and its mutants

Four gene-Blocks (gBlocks) encoding AR9 nvRNAP core enzyme genes optimized for expression in *E. coli* were synthetized by Integrated DNA Technologies (IDT). These gBlocks were assembled into an expression vector on the pETDuet-1 plasmid backbone with the help of the NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs). First, pETDuet-1 was digested by the NcoI and BamHI endonucleases, and gBlocks coding for N-terminally hexahistidine-tagged gp270 and gp154 were ligated. Then, this plasmid was digested by the BglII and XhoI endonucleases and ligated with two gBlocks coding for gp105 and gp089. The resulting plasmid encoded the AR9 nvRNAP core enzyme. The plasmid for expression of the AR9 nvRNAP holoenzyme was created by inserting an *E. coli*-optimized gp226 gBlock (also synthetized by ITD) into the AR9 nvRNAP core plasmid described above, which was linearized at the XhoI site. This plasmid was used as a template to create mutant versions of the AR9 nvRNAP holoenzyme by site-directed mutagenesis (the list of corresponding primers is in the Table 2.1). The plasmid encoding the tagless AR9 nvRNAP core enzyme was derived from the His-tagged AR9 nvRNAP core plasmid described above. First, a fragment that contained all the four genes but excluded the Hig-tag was PCR amplified (the primers are listed in the Table 2.1). Then,

the pETDuet-1 vector was linearized by the NcoI and XhoI endonucleases. A new plasmid was then created by ligating the PCR fragment and the linearized pETDuet-1 vector using the NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs). In all plasmids, a T7 RNAP promoter, a *lac* operator, and a ribosome binding site were located at appropriate positions upstream of each gene.

Purification of recombinant AR9 nvRNAP

Plasmids encoding AR9 nvRNAP core, tagless AR9 nvRNAP core, and holoenzyme or its mutants were transformed into BL21 Star (DE3) chemically competent *E. coli* cells. The cultures (3 L) were grown at 37°C to OD_{600} of 0.7 in LB medium supplemented with ampicillin at a concentration of 100 μg/mL, and recombinant protein overexpression was induced with 1 mM IPTG for 4 hours.

Cells containing over-expressed AR9 nvRNAP holoenzyme or its mutants were harvested by centrifugation and disrupted by sonication in buffer A (40 mM Tris-HCl pH 8, 300 mM NaCl, 3 mM β-mercaptoethanol) followed by centrifugation at 15,000 g for 30 min. Cleared lysate was loaded onto a 5 mL HisTrap sepharose HP column (GE Healthcare) equilibrated with buffer A. The column was washed with buffer A supplemented with 20 mM Imidazole. The protein was eluted with a linear 0-0.5 M Imidazole gradient in buffer A. Fractions containing AR9 nvRNAP holoenzyme or its mutants were combined and diluted with buffer B (40 mM Tris-HCl pH 8, 0.5 mM EDTA, 1 mM DTT, 5% glycerol) to the 50 mM NaCl final concentration and loaded on equilibrated 5 mL HiTrap Heparin HP sepharose column (GE Healthcare). The protein was eluted with a linear 0-1 M NaCl gradient in buffer B. Fractions containing AR9 nvRNAP holoenzyme or its mutants were pooled and concentrated (Amicon Ultra-4 Centrifugal Filter Unit with Ultracel-50 membrane, EMD Millipore) to a final concentration of 3

mg/mL, then glycerol was added up to 50% to the sample for storage at −20°C (the samples were used for transcription assays).

Samples used for crystallization and cryo-EM were produced by following a slightly different procedure. Cells containing over-expressed recombinant AR9 nvRNAP core or holoenzyme were harvested by centrifugation and disrupted by sonication in buffer C (50 mM NaH_2PO_4 pH 8, 300 mM NaCl, 3 mM β-mercaptoethanol, 0.1 mM PMSF) followed by centrifugation at 15,000 g for 30 min. Cleared lysate was loaded on 5 mL Ni-NTA column (Qiagen) equilibrated with buffer C, washed with 5 column volumes of buffer C and with 5 column volumes of buffer C containing 20 mM Imidazole. Then, elution with buffer C containing 200 mM Imidazole was carried out. Fractions containing AR9 nvRNAP core or holoenzyme were pooled and diluted ten times by buffer D (20 mM Tris pH 8, 0.5 mM EDTA, 1 mM DTT) or by buffer E (20 mM Bis-tris propane pH 6.8, 0.5 mM EDTA, 1 mM DTT) correspondingly and applied to a MonoQ 10/100 column (GE Healthcare). Bound proteins were eluted with a linear 0.25–0.45 M NaCl gradient in buffer D or E correspondingly.

Cells containing over-expressed recombinant tagless AR9 nvRNAP core were harvested by centrifugation and disrupted by sonication in buffer B followed by centrifugation at 15,000g for 30 min. An 8% polyethyleneimine (PEI) solution (pH 8.0) was added with stirring to the cleared lysate to the final concentration of 0.8%. The resulting suspension was incubated on ice for 1 hour and centrifuged at 10,000g for 15 min. The supernatant was removed, and the pellet was resuspended in buffer B containing 0.3 M NaCl. After 10 min incubation, the PEI pellet was formed by centrifugation as previously. Then, supernatant was removed, and the pellet was resuspended in buffer B containing 1 M NaCl followed by centrifugation at 10,000g for 15 min. Eluted proteins were precipitated the supernatant by addition of ammonium sulfate to 67% saturation and dissolved in buffer D and loaded on equilibrated 5 mL HiTrap Heparin HP sepharose column (GE Healthcare). The protein was eluted with a linear 0-1 M NaCl gradient in

buffer D. Fractions containing tagless AR9 nvRNAP core were pooled and subjected to anion exchange chromatography as described above for AR9 nvRNAP core.

The AR9 nvRNAP core sample was polished and buffer-exchanged using size exclusion chromatography on a Superdex 200 10/300 (GE Healthcare) column equilibrated with buffer D containing 100 mM NaCl. The tagless AR9 nvRNAP core and AR9 nvRNAP holoenzyme were not subjected to size exclusion chromatography – salt concentration in the sample was lowered during the concentration procedure.

The fractions containing AR9 nvRNAP core, tagless AR9 nvRNAP core or holoenzyme were concentrated to a final concentration of 20 mg/mL and used for crystallization or cryo-EM.

Preparation of the promoter complex for structure determination

To prepare the DNA template for crystallization and cryo-EM, two corresponding oligonucleotides (Table 2.1) that were synthetized by IDT with dual PAGE and HPLC purification at a final concentrations of 100 μM each were annealed together by mixing in a buffer containing 20 mM Bis-tris propane pH 6.8, 100 mM NaCl, 4 mM $MgCl_2$, 0.5 mM EDTA and incubating at 65 °C for 1 minute and cooling down to 4 °C by a decrement of 1°C per minute. A 1.5-fold molar excess of the DNA template was added to the holoenzyme and incubated for 30 min at room temperature (the final concentrations: 10 mg/mL of the protein (34 μM) and 50 μM of the DNA). The obtained complex was used for crystallization and cryo-EM directly.

Crystallization of AR9 nvRNAP

The initial crystallization screening was carried out by the sitting drop method in 96 well ARI Intelliwell-2 LR plates using Jena Bioscience crystallization screens at 19°C.

PHOENIX pipetting robot (Art Robbins Instruments, USA) was employed for preparing crystallization plates and setting up drops, each containing 200 nL of the protein and the same volume of well solution. Optimization of crystallization conditions was performed in 24 well VDX plates and thin siliconized cover slides (both from Hampton Research) by hanging drop vapor diffusion. The best crystals were obtained as follows: i) an 1.5 µl aliquot of AR9 nvRNAP core (4.5 mg/mL) was mixed with an equal volume of a solution containing 100 mM Tricine pH 8.8, 270 mM KNO_3, 15 % PEG 6000, 5 mM $MgCl_2$, and incubated as a hanging drop over the same solution; ii) an 1.5 µl aliquot of tagless AR9 nvRNAP core (7.5 mg/mL) was mixed with an equal same volume of a solution containing 150 mM Malic acid pH 7, 150 mM NaCl, 14 % PEG 3350 and incubated as a hanging drop over the same solution; iii) an 1.5 µl aliquot of the AR9 nvRNAP promoter complex (10 mg/mL) was mixed with an equal same volume of a solution containing 150 mM MIB pH 5, 150 mM LiCl, 13 % PEG 1500 and incubated as a hanging drop over the same solution. Some crystal reached their final size the next day and some grew for two weeks at 19 C° temperature.

Preparation of heavy-atom derivative crystals

The following compounds were tested for heavy atom derivatization of AR9 nvRNAP core crystals (by co-crystallization and soaking): $SrCl_2$, $GdCl_3$, Na_2WO_4, $HgCl_2$, $Pb(NO_3)_2$, thimerosal (2-$(C_2H_5HgS)C_6H_4CO_2Na$), 10 compounds containing Eu and Yb atoms (JBS Lanthanide Phasing Kit), three compounds containing W (JBS Tungstate Cluster Kit) and one cluster compound containing Ta (Ta_6Br_{12} JBS Tantalum Cluster Derivatization Kit). The crystals were soaked in a range of concentrations of heavy atom compounds (between 0.1 mM and 100 mM) that were added to the crystallization solution. The soaking time was varied from 2 hours to 2 days. Among all examined conditions, only solutions containing 10 mM thimerosal or 1 mM tantalum bromide resulted in heavy atom

derivatization (judging by the presence of anomalous signal in X-ray diffraction data) upon overnight soaking. To produce a Se-methionine (SeMet) derivative of the AR9 nvRNAP core enzyme, the corresponding plasmid was transformed into B834(DE3) chemically competent *E. coli* cells. The cells were first grown in LB medium until the optical density OD_{600} reached a value of 0.35. The cells were then pelleted by centrifugation at 4,000 g for 10 min at 4°C and transferred to the SelenoMet Medium (Molecular Dimensions) that was supplemented with ampicillin at a concentration of 100 μg/mL. The protein expression then proceeded according to the manufacturer's instructions. All the subsequent steps were the same as for the native protein.

X-ray data collection

Cryoprotectant solutions were prepared by replacing 25% of water in the crystallization solution (hanging drop well solution) with ethylene glycol, which was found to be the best cryoprotectant by trial and error. The crystals were either soaked for 1-5 minutes in the cryoprotectant solution or briefly dipped into it and then flash frozen in liquid nitrogen. Such frozen crystals were then transferred to a shipping Dewar and shipped to the APS (LS-CAT) or ALS (BCSB) synchrotrons for remote data collection. X-ray diffraction data and fluorescent spectra were collected in a nitrogen stream at 100 K. Heavy atom and SeMet derivative data were collected at the absorption peak wavelength (the white line, if present) of the X-ray fluorescence spectrum.

X-ray structure determination

The structure determination process spanned nearly three years. Initially, we aimed to solve the structure of the AR9 nvRNAP core enzyme by X-ray crystallography or cryo-EM and use it to solve the structure of the promoter complex. However, the atomic model

of the promoter complex obtained by cryo-EM was built first. Then, it was used to solve the X-ray structure of the core and to interpret the cryo-EM map of the holoenzyme. The path to the atomic model described below is the trunk of a tree that had many branches representing things that did not work. Many different tools were used in the determination of this structure albeit most of them ended up being dead end branches. The following procedure involves fewest datasets and fewest steps that lead to an interpretable map. None of the RNAP structures present in the PDB at the start of this project were sufficiently similar to solve the structure of the AR9 nvRNAP core by molecular replacement (MR), so crystallographic phases had to be obtained by a *de novo* phasing procedure (heavy atom isomorphous replacement or anomalous scattering). Severe anisotropy and inconsistent diffraction of AR9 nvRNAP core native crystals made this task extremely complicated and we had to screen hundreds of heavy atom-soaked crystals for diffraction. The SeMet derivative diffracted to 5.5 Å resolution, which was insufficient to solve the Se substructure using anomalous scattering. Moreover, this derivative was not isomorphous to any of the native datasets.

An interpretable map was obtained by a convoluted procedure. A map in which the characteristic features of a DNA-dependent RNAP – two adjoining double-ψ β-barrel (DPBB) domains and several large α-helices, including the bridge helix (although split in the middle) – could be discerned, but no side chain densities were present, was obtained by a multiple isomorphous replacement plus anomalous scattering method which was applied to the native, Ta_6Br_{12}, and thimerosal derivative datasets of the His-tagged AR9 nvRNAP core enzyme (see **Table 2.2**). This map was calculated by the SHARP software package that was run with mostly default settings(Vonrhein et al. n.d.). The most similar part(Zimmermann et al. 2018) of the archaeal RNAP structure(Wojtas et al. 2012) (PDB code 4ayb) was fitted into this density using Coot(Emsley, Cowtan, and IUCr 2004) and

all parts that did not fit the density and all side chains were removed. The resulting model contained 832 alanine residues.

This model was then used to solve the structure of a unique thimerosal derivative dataset that belonged to a different unit cell (**Table 2.2**) with the help of a MR procedure(Drenth and IUCr 1974). This unique dataset resulted from the crystallization of a tagless version of the AR9 nvRNAP core (all native N- and C-termini). It had a very large orthorhombic unit cell with eight molecules of the AR9 nvRNAP core in its asymmetric unit or about 17,800 amino acids. Remarkably, Phaser(McCoy et al. 2007) was able to locate all eight copies of the AR9 nvRNAP core in this 3.8 Å resolution dataset with help of the 832-residue polyalanine fragment (obtained as described before) as a search model. This polyalanine search model corresponded to about 1.8% of the total protein material in the asymmetric unit and 1% of the total asymmetric unit content if solvent atoms are considered. The density was then dramatically improved by 25 cycles of eightfold non-crystallographic symmetry (NCS) averaging using Parrot(Cowtan and IUCr 2010). The resulting density was mostly continuous and showed many bulky side chains especially in the vicinity of the DPBB domains. Buccaneer(Cowtan and IUCr 2006) was then used for automatic model building into this map. The Buccaneer model was cleaned up manually and separate chain fragments were assembled into a new intermediate AR9 nvRNAP core model that contained 995 residues of which 937 had side chains. The new intermediate model was then used as a search model in a new round of MR by Phaser that was followed by NCS averaging using Parrot.

The new density was of sufficient quality to recognize the identity of many side chains and for manual model building using Coot. Structures of the *Mycobacterium tuberculosis* and *E. coli* RNAPs (PDB codes 5ZX3(Li et al. 2019) and 6C9Y(Narayanan et al. 2018), respectively) were used to aid in chain tracing. Additionally, this thimerosal derivative dataset contained Hg atoms, identified with the help of anomalous Fourier

synthesis, that were expected to bind to cysteine side chains. Thus, the Hg atoms were used as markers to maintain the chain register.

Eventually, a large fraction of the nvRNAP core atomic model was complete. Some peripheral parts, however, and the β′C gp154 Insertion domain were too disordered for model building. While this work was in progress, an interpretable (3.8 Å resolution) cryo-EM map of the AR9 nvRNAP promoter complex was obtained. This map was of better quality than the 3.8 Å resolution, large unit cell, X-ray dataset of the tagless core enzyme, so further rounds of model building were continued using the cryo-EM map. Once the holoenzyme model was complete, the core or the entire holoenzyme (but not the DNA) were used to solve the crystal structure of the native AR9 nvRNAP core (3.3 Å resolution, the standard unit cell, **Table 2.2**) and promoter complex (3.4 Å resolution, **Table 2.2**) by MR using Phaser.

Some peripheral regions of the cryo-EM promoter complex map, namely, the β′C gp154 Insertion domain, residues 130-289 of βN gp105, and the peripheral parts of gp226, were too poor for reliable *de novo* model building. Fortunately, the structure of AR9 nvRNAP promoter complex was one of large multisubunit targets of the CASP14 protein structure prediction competition. The Google DeepMind AlphaFold2 software predicted the structure of difficult-to-build domains with excellent accuracy(Jumper et al. 2021). This allowed us to complete the AR9 nvRNAP promoter complex model in the cryo-EM map first and then use this model to solve the 3.4 Å resolution crystal structure of the promoter complex. Refinement of crystallographic and cryo-EM models was performed using Phenix(Adams et al. 2011) and Coot(Emsley, Cowtan, and IUCr 2004). Additional details describing model building and the analysis of AlphaFold2 models are given elsewhere(Kryshtafovych et al. 2021).

Cryo-EM sample preparation and data acquisition of the AR9 nvRNAP promoter complex

QUANTIFOIL 1.2/1.3 copper grids were plasma cleaned for 30s using the model 950 advanced plasma system by Gatan. 3 μL of 10 mg/mL of the AR9 nvRNAP holoenzyme (34 μM) and 50 μM of the DNA nucleotide in 20 mM Bis-tris propane pH 6.8, 100 mM NaCl, 2 mM $MgCl_2$, 0.5 mM EDTA was pipetted onto to the grid and blotted using a Vitrobot (Thermo Fischer Scientific) at 100% humidity for 5 s. Following blotting, the sample was plunged into liquid ethane cooled by liquid nitrogen.

5,351 (5,760 × 4,092 pixels) micrograph movies were collected using the EPU software on a Titan Krios 300kV electron microscope with a BioQuantum K3 imaging filter with a 20-eV slit. Each movie contained 56 frames collected over 1.5 s, with a frame dose of 0.78 e/$Å^2$ and pixel size of 1.09 Å. Movies were collected over a defocus range of −1 to −4 μm.

Cryo-EM image processing of the AR9 nvRNAP promoter complex

Image processing was performed using Eman2(Tang et al. 2007), Relion3.0(Zivanov et al. 2018) and CryoSPARC3.0(Punjani et al. 2017) (**Table 2.3**). All movies were motion corrected using MotionCor2(Zheng et al. 2017). Estimation of the contrast transfer function (CTF) parameters was performed by CTFFIND4.1(Rohou and Grigorieff 2015) over the resolution range of 5.0-30.0 Å. E2boxer(Tang et al. 2007) was used for particle picking, resulting in 420,791 particles with a box size of 300 pixels. Particle box coordinates were used by Relion3.0 to extract the boxes. 2D classification by Relion3.0 resulted in 243,976 particles belonging to high quality classes. These particles were imported into CryoSPARC3.0, where *ab initio* reconstruction was carried out using three models. Following the *ab initio* reconstruction, 3D classification was performed with five classes, a box size of 150 pixels to improve speed, a batch size of 2,000 particles per class and an assignment convergence criterion of 2%. Non-Uniform (NU)

refinement(Punjani, Zhang, and Fleet 2020) was executed using both the 106,867 particles and the map from the most populous class of 3D classification. 3D local refinement was then carried out using an alignment resolution of 0.25° and NU refinement. Subsequently, local CTF-refinement was performed using a search range of 3.5-20.0 Å. Following this, around round of NU refinement and 3D local refinement with an alignment resolution of 0.25° and NU-refinement was performed. The resulting map was sharpened with a B-factor of −138 $Å^2$.

Cryo-EM sample preparation and data acquisition of the AR9 nvRNAP holoenzyme

TED PELLA 200 mesh PELCO NetMesh copper grids were plasma cleaned for 30s using the model 950 advanced plasma system by Gatan. 3 µl of 20mg/ml His-tag 5s nvRNAP in 20 mM Bis-tris propane pH 6.8, 100 mM NaCl, 4 mM $MgCl_2$, 0.5 mM EDTA buffer was pipetted onto to the grid and blotted using a Vitrobot (Thermo Fischer Scientific) at 100% humidity for 5 s. Following blotting, the sample was plunged into liquid ethane cooled by liquid nitrogen.

2,691 (3,838 × 3,710 pixels) micrograph movies were collected using the EPU software on a Titan Krios 300kV electron microscope with a K2 Summit camera and a 20-eV slit. Each movie contained 40 frames collected over 8 s, with a frame dose of 1.08 e/$Å^2$ and pixel size of 1.08 Å. Movies were collected with a target defocus of −1.8 µm. Images were collected with a 30-degree tilt.

Cryo-EM image processing of the AR9 nvRNAP holoenzyme

Image processing was performed using Relion2.0(Zivanov et al. 2018). All movies were motion corrected using MotionCor2(Zheng et al. 2017). Estimation of the contrast

transfer function (CTF) parameters was performed by Gctf(Zhang 2016). Particle picking resulted in 227,577 particles with a box size of 200 pixels. 2D and 3D classification by Relion2.0 resulted in 104,471 particles belonging to high quality classes. The resulting particles were refined to a resolution of 4.4 Å. The resulting map was sharpened with a B-factor of −87 $Å^2$.

Gp226 cloning, purification and limited digestion with trypsin

The AR9 gene *226* was PCR amplified from AR9 genomic DNA and cloned into the pQE-2 vector (QIAGEN) between the SacI and SalI restriction sites. The resulting plasmid was transformed into BL21 (DE3) chemically competent *E. coli* cells. The culture (7 L) was grown at 37°C to an OD_{600} of 0.5 in LB medium supplemented with ampicillin at a concentration of 100 μg/mL, and recombinant protein overexpression was induced with 1 mM IPTG for 4 hours at 22°C. Cells containing over-expressed recombinant protein were harvested by centrifugation and disrupted by sonication in buffer C followed by centrifugation at 15,000 g for 30 min. Cleared lysate was loaded on a 5 mL Ni-NTA column (Qiagen) equilibrated with buffer C, washed with 5 column volumes of buffer C and with 5 column volumes of buffer C containing 20 mM Imidazole. Then, elution with buffer C containing 200 mM Imidazole was carried out. Fractions containing gp226 were pooled, concentrated and subjected to gel-filtration on a Superdex 200 10/300 (GE Healthcare) column equilibrated with buffer C. The fractions containing gp226 monomer were concentrated to a final concentration of 1 mg/mL and used for the limited proteolysis experiment.

Trypsin digestion of gp226, which had a concentration of 80 ng/μL, was carried out in 20 ul of the digestion buffer (50 mM NaH_2PO_4 pH 8.0, 300 mM NaCl) that contained a range of trypsin concentrations (Sigma-Aldrich). The trypsin to gp226 molar ratios were from 0.03 to 0.6. The reactions were allowed to proceed for 1 hour at 25°C and stopped by

the addition of Laemmli loading buffer and immediate boiling. The reaction products were analyzed by denaturing SDS polyacrylamide gel electrophoresis (SDS-PAGE) with subsequent mass-spectrometry as described previously(Lavysh et al. 2016).

DNA templates for transcription assay

Long DNA templates containing late AR9 promoters were prepared by polymerase chain reaction (PCR). PCRs were done with Encyclo DNA polymerase (Evrogen, Moscow) and the AR9 genomic DNA as a template, with a standard concentration of dNTPs (Thermo Fisher Scientific) to obtain DNA fragments with thymine or in the presence of dUTP (Thermo Fisher Scientific) in place of dTTP to obtain DNA fragments with uracil. Oligonucleotide primers used for PCR are listed in (Table 2.1).

Short double-stranded and partially single-stranded DNA templates containing the P077 promoter with uracils and thymines at certain positions were prepared by annealing of oligonucleotides ordered from Evrogen (Moscow) and listed in (Table 2.1). To prepare specific DNA templates, two corresponding oligonucleotides were annealed together by mixing in buffer containing 40 mM Tris-HCl pH 8, 10 mM $MgCl_2$ and 0.5 mM DTT, incubating at 75 °C for 1 minute and cooling down to 4 °C by a decrement of 1°C per minute.

***In vitro* transcription**

Multiple-round run-off transcription reactions were performed in 5 µL of transcription buffer (40 mM Tris-HCl pH 8, 10 mM $MgCl_2$, 0.5 mM DTT, 100 µg/mL bovine serum albumin (Thermo Fisher Scientific), and 1 U/µL RiboLock RNase Inhibitor (Thermo Fisher Scientific)) and contained 100 nM AR9 nvRNAP holoenzyme and 100 nM DNA template. The reactions were incubated for 10 min at 37°C, followed by the addition

of 100 μM each of ATP, CTP, and GTP, 10 μM UTP and 3 μCi [α-^{32}P]UTP (3000 Ci/mmol) (**Fig. 2.E1c, Figure 2.2b**) or 100 μM each of ATP, UTP, GTP, 10 μM CTP and 3 μCi [α-^{32}P]CTP (3000 Ci/mmol) (**Figure 2.2d, 2.2f**). Reactions proceeded for 30 min at 37°C and were terminated by the addition of an equal volume of denaturing loading buffer (95% formamide, 18 mM EDTA, 0.25% SDS, 0.025% xylene cyanol, 0.025% bromophenol blue). The reaction products were resolved by electrophoresis on 6-23 % (w/v) polyacrylamide gel containing 8 M urea. The results were visualized with a Typhoon FLA 9500 scanner (GE Healthcare).

Molecular dynamics general methods

Simulations were carried out on both the LS5 and Stampede2 systems at the Texas Advanced Computing Center (TACC) using NAMD 2.10(Phillips et al. 2005). The CHARMM36 force field was used(Huang and MacKerell 2013). Production runs were performed in the isothermal isobaric (NPT) ensemble using Langevin dynamics and a Langevin piston(Feller et al. 1998). Alchemical transformations were analyzed by the ParseFEP package(Pohorille, Jarzynski, and Chipot 2010). Entropic restraints were calculated via thermodynamic integration. Collective variables were implemented via the colvars module in NAMD(Fiorin, Klein, and Hénin 2013). The energy of non-bonded VdW interactions for distances exceeding 10 Å was smoothly decreased to equal zero at 12 Å. A 2 fs timestep was used in all simulations. Long range electrostatics was calculated with the help of the Particle Mesh Ewald algorithm(Darden, York, and Pedersen 1998). During alchemical transformation, a soft core VdW radius of 4 Å^2 was used to improve convergence and accuracy(Beutler et al. 1994)·(Zacharias, Straatsma, and McCammon 1998). Both alchemical and restraint calculation simulations were carried out bidirectionally. The Bennett Acceptance Ratio maximum likelihood estimate(Bennett 1976) was used to determine free energy change for alchemical transformations. The

double decoupling method (DDM) was implemented as described(Gumbart Benoˆıt, Roux, and Chipot 2018),(Gumbart, Roux, and Chipot 2012).

Molecular dynamics system setup

The protein structure description files for both structures – the AR9 nvRNAP holoenzyme in complex with the 3′-$^{-11}$UUG^{-9}-5′ oligonucleotide bound to the promoter binding pocket and for the 3′-$^{-11}$UUG^{-9}-5′ oligonucleotide in the promoter bound conformation – were generated using the psfgen plugin of VMD(Humphrey, Dalke, and Schulten 1996). Solvation was performed using TIP3 water(Harrach and Drossel 2014) in a box with 15 Å padding in each direction and ionized in 0.1 M NaCl. The holoenzyme-DNA and DNA systems were enclosed in periodic boxes with cell dimensions of (176 Å , 147 Å, 133 Å) and (42 Å , 42 Å, 41 Å), respectively, and contained 91,177 and 2,166 water molecules, respectively. Both systems were first minimized for 1,000,000 steps while restraints and constraints on the protein (in the holoenzyme system), DNA, and water atoms were gradually removed. Both systems were heated from 0 K to 300 K in 5 K increments for a total of 19.2 ns with constraints on backbone atoms. The holoenzyme-DNA and DNA systems were then equilibrated with minimal constraints in the isothermal-isobaric (NPT) ensemble for 100 ns and 50 ns, respectively.

Definition of collective variables

For the implementation of the DDM method via alchemical transformations, the system must be (harmonically) constrained such that the finite sampling can be focused on relevant regions of phase space. The entropic cost of applying these restraints is evaluated in separate independent simulations. The phase space and the entropic cost are connected

to each other through a set of collective variables that are applied to atoms during simulations.

Only one collective variable is needed to restrain the conformation of the oligonucleotide in bulk water (the bulk water DNA system): the root mean squared deviation (RMSD) of all non-H DNA atoms relative to the equilibrated state. In the holoenzyme-DNA complex, seven collective variables are required – six to define the orientation and position of the rigid DNA molecule relative to the holoenzyme complex and one to define the conformation of the DNA. Similarly to the bulk water DNA case, the RMSD of all non-H DNA atoms relative to the equilibrated state is used as the collective variable to restrain the conformation of DNA. We characterize the orientation of the rigid DNA molecule via the relative position of the backbone atoms of $^{-11}$UUG^{-9} to those of gp226 V206, gp105 N382 and gp226 K262. Accordingly, the orientation of DNA relative to the holoenzyme is characterized by six collective variables: r – the distance between $^{-11}$U and gp226 V206); ϕ – the angle between gp226 V206, $^{-11}$U, and $^{-9}$G); χ – the angle between $^{-11}$U, gp226 V206, and gp105 N382); θ – the dihedral angle between gp226 V206, $^{-11}$U, $^{-9}$G, and $^{-10}$U); ψ – the dihedral angle between gp105 N382, gp226 V206, $^{-11}$U, and $^{-9}$G); ξ – the dihedral angle between $^{-11}$U, gp226 V206, gp105 N382, and gp226 K262).

The harmonic force constraint constants applied to the distance-type (RMSD and r) and angular collective variables were 10 kcal/mol/Å^2 and 0.1 kcal/mol/deg^2, respectively. The equilibrium positions for all harmonic restraints were derived from the equilibrated holoenzyme-DNA structure. For restraint estimation simulations, harmonic forces were varied smoothly using a target force exponent value of 4.0. The lambda schedule focused near the value 1.0 to improve simulation convergence and ensure thermodynamic micro-reversibility(Roux et al. 1996)·(Gilson et al. 1997): [1.00, 0.999, 0.99, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70, 0.65, 0.60, 0.55, 0.50, 0.45, 0.40, 0.35, 0.30, 0.20, 0.10, 0.00]. The reverse sequence was used for the backward simulation.

Thermodynamic cycle

The standard binding free energy was calculated by combining the results of four separate simulations which represent the four vertical reactions of the thermodynamic cycle (**Fig. 2.E8a**). These simulations evaluate the following parameters (**Table 2.4**): 1) the entropic cost of restraining the promoter DNA to the "bound" state in the promoter binding pocket by adding/removing conformational restraints on the promoter DNA ($\Delta G^{bound}_{restrain}$); 2) the free energy of coupling/decoupling the promoter DNA from the binding pocket via alchemical transformations with restraints on the conformation of the promoter DNA ($\Delta G^{bound}_{alchemical}$); 3) the entropic cost of restraining the promoter DNA to the bound conformation in bulk water by adding/removing conformational restraints on the promoter DNA ($\Delta G^{bulk\ water}_{restrain}$); 4) the free energy of coupling/decoupling the promoter DNA from bulk water via alchemical transformations with restraints on the conformation of the promoter DNA ($\Delta G^{bulk\ water}_{alchemical}$). The change in free energy resulting from these transitions is then calculated via thermodynamic integration(Ratner, Ratner, and A. 1997) and free energy perturbation(Beveridge and Dicapua n.d.) methods. The results were validated by checking for micro-reversibility and the absence of hysteresis(Pohorille, Jarzynski, and Chipot 2010),(Bennett 1976). Following the completion of the thermodynamic cycle, the standard binding free energy of promoter DNA to the holoenzyme binding pocket was found to be −6.9 ± 2.8 kcal/mol.

Molecular dynamics error analysis

An upper bound on the error of $\Delta G^{bound}_{restrain}$, $\Delta G^{bulk\ water}_{restrain}$ and $\Delta G^{bulk\ water}_{alchemical}$ was determined by the hysteresis between backward and forward simulations(Gumbart, Roux,

and Chipot 2012). The error in $\Delta G^{bound}_{alchemical}$ was determined by performing 3 replicates of the simulation and evaluating the standard deviation (**Table 2.4**).

Data availability

All macromolecular structure data described in this paper have been deposited to the Protein Data Bank and Electron Microscopy Data Bank under the following accession numbers: PDB code 7S00 (AR9 nvRNAP core X-ray structure); PDB code 7S01 (AR9 nvRNAP promoter complex X-ray structure); EMDB code EMD-24765 (AR9 nvRNAP holoenzyme cryo-EM density); EMDB code EMD-24763 (AR9 nvRNAP promoter complex cryo-EM density). Publicly available protein atomic models with the following PDB codes were used in the study: 4AYB(Wojtas et al. 2012), 5ZX3(Li et al. 2019), 6C9Y(Narayanan et al. 2018), 5IPM(B. Liu, Zuo, and Steitz 2016), 6JBQ(Fang et al. 2019), and 7OGP(Garrido et al. 2021).

ACKNOWLEDGMENTS

This work was supported by Skoltech NGP Program (Skoltech-MIT joint project) and by the Russian Science Foundation (Grant 19-74-00011 to M. L. Sokolova). The work was also supported by the UTMB Department of Biochemistry and Molecular Biology and by the UTMB Sealy Center for Structural Biology and Molecular Biophysics. The MD work was performed using the computing facilities of the Texas Advanced Computing Center (TACC, http://www.tacc.utexas.edu) at The University of Texas for which we are very grateful. We thank the Stanford-SLAC Cryo-EM Facilities, supported by Stanford University, SLAC and the National Institutes of Health S10 Instrumentation Programs that were used to collect the AR9 nvRNAP holoenzyme cryo-EM data. We acknowledge the use of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science

User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. We thank the staff of the LS-CAT Sector 21 beamlines that is supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (Grant 085P1000817). We acknowledge the use of the Berkeley Center for Structural Biology (supported in part by the Howard Hughes Medical Institute) at the Advanced Light Source (a Department of Energy Office of Science User Facility under Contract No. DE-AC02-05CH11231) and we thank the staff of the beamline 5.0.2. Finally, we thank Dr. Mark A. White for his help and assistance with the initial crystallization and X-ray data collection of the AR9 nvRNAP core, Dr. Michael B. Sherman for his help with the cryo-EM data collection of all datasets used in this paper, and Dr. Tatyana O. Artamonova for mass-spectrometry analysis of gp226 digestion products. The research reported in this paper extensively used the facilities and resources of the UTMB SCSB Macromolecular Structure X-ray Laboratory and UTMB SCSB Cryo-EM Laboratory.

Author contributions

K.V.S. and **M.L.S.** conceived the study. **M.L.S.** cloned, purified and crystallized AR9 nvRNAP core, tagless AR9 nvRNAP core, and AR9 nvRNAP holoenzyme in complex with promoter DNA, derivatized crystals, prepared samples for cryo-EM, purified gp226 and performed limited trypsinolisis. **A.F**. obtained and analyzed all cryo-EM reconstructions, built parts of atomic models, and performed all MD work. **A.V.D.** purified AR9 nvRNAP holoenzyme and its mutants and performed *in vitro* transcription assays under the supervision of **M.L.S**. **J.G.** under the supervision of **M.L.S.** crystallized the AR9 nvRNAP holoenzyme in complex with promoter DNA. **P.G.L.** collected X-ray data, solved all crystal structures, and built and refined all atomic models. The **AF team** created models of all five AR9 nvRNAP holoenzyme proteins that were used by **P.G.L.** and **A.F.** in the

interpretation of cryo-EM and X-ray crystallography electron density maps. **A.F., M.L.S.**, **S.B.**, and **P.G.L.** analyzed the structures. **P.G.L.** and **A.F.** wrote the manuscript, which was read, edited, and approved by all authors.

Competing interests

The authors declare no competing interests.

Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to Maria L. Sokolova, Petr G. Leiman, or Konstantin V. Severinov.

Chapter 3 Quantitative Description of a Contractile Macromolecular Machine

The following chapter was published under CC license with no changes as:
Fraser A, Prokhorov NS, Jiao F, Pettitt BM, Scheuring S, Leiman PG. Quantitative description of a contractile macromolecular machine. *Sci Adv*. 2021;7(24):9601-9612. doi:10.1126/SCIADV.ABF9601

ABSTRACT

Contractile Injection Systems (CISs) (Type VI Secretion System (T6SS), phage tails, and tailocins) employ a contractile sheath-rigid tube machinery to breach cell walls and lipid membranes. The structures of the pre- and post-contraction states of several CISs are known, but the mechanism of contraction remains poorly understood. Combining structural information of the end states of the 12 MDa R-type pyocin sheath-tube complex with thermodynamic and force spectroscopy analyses and a novel modeling procedure, we describe the mechanism of pyocin contraction. We show that this nanomachine has an activation energy of 160 kcal/mol, it releases 2,160 kcal/mol of heat, and develops a force greater than 500 pN. Our combined approach provides the first quantitative and experimental description of the membrane penetration process by a CIS.

INTRODUCTION

Contractile Injection Systems (CISs), which include the bacterial Type VI Secretion System (T6SS), bacteriophage tails, R-type pyocins and other tailocins function to penetrate bacterial and eukaryotic membranes(Taylor, Raaij, and Leiman 2018)·(Patz et al. 2019)·(Marek Basler 2015)·(Cascales and Cambillau 2012). The universally conserved part of CISs consists of an external contractile sheath, an internal rigid tube, and a baseplate (**Fig. 3.1A**). The tube and the sheath are made up of sixfold symmetric layers of subunits

stacked upon each other with a twist, forming a helical structure. The initial extended conformation of the complex represents a high energy metastable state(Ge et al. 2015),(Caspar 1980). The sheath contracts to about half of its original length upon activation through a specific stimulus originating at the baseplate, e.g. attachment to the target cell surface, or spontaneously when subjected to stress or upon long storage(Leiman et al. 2004),(Guerrero-Ferreira et al. 2019). The baseplate-distal end of the tube and the sheath are fixed to each other with a capping protein (**Fig. 3.1A**). Consequently, contraction of the sheath results in the motion of the tube towards and through the target cell membrane(Leiman et al. 2004) (**Fig. 3.1A**). This process is aided by a spike-shaped protein located at the baseplate-proximal end of the tube(Browning et al. 2012),(Shneider et al. 2013) (**Fig. 3.1A**). The membrane-attacking tip of the spike protein is stabilized by an iron or zinc atom(Browning et al. 2012),(Shneider et al. 2013).

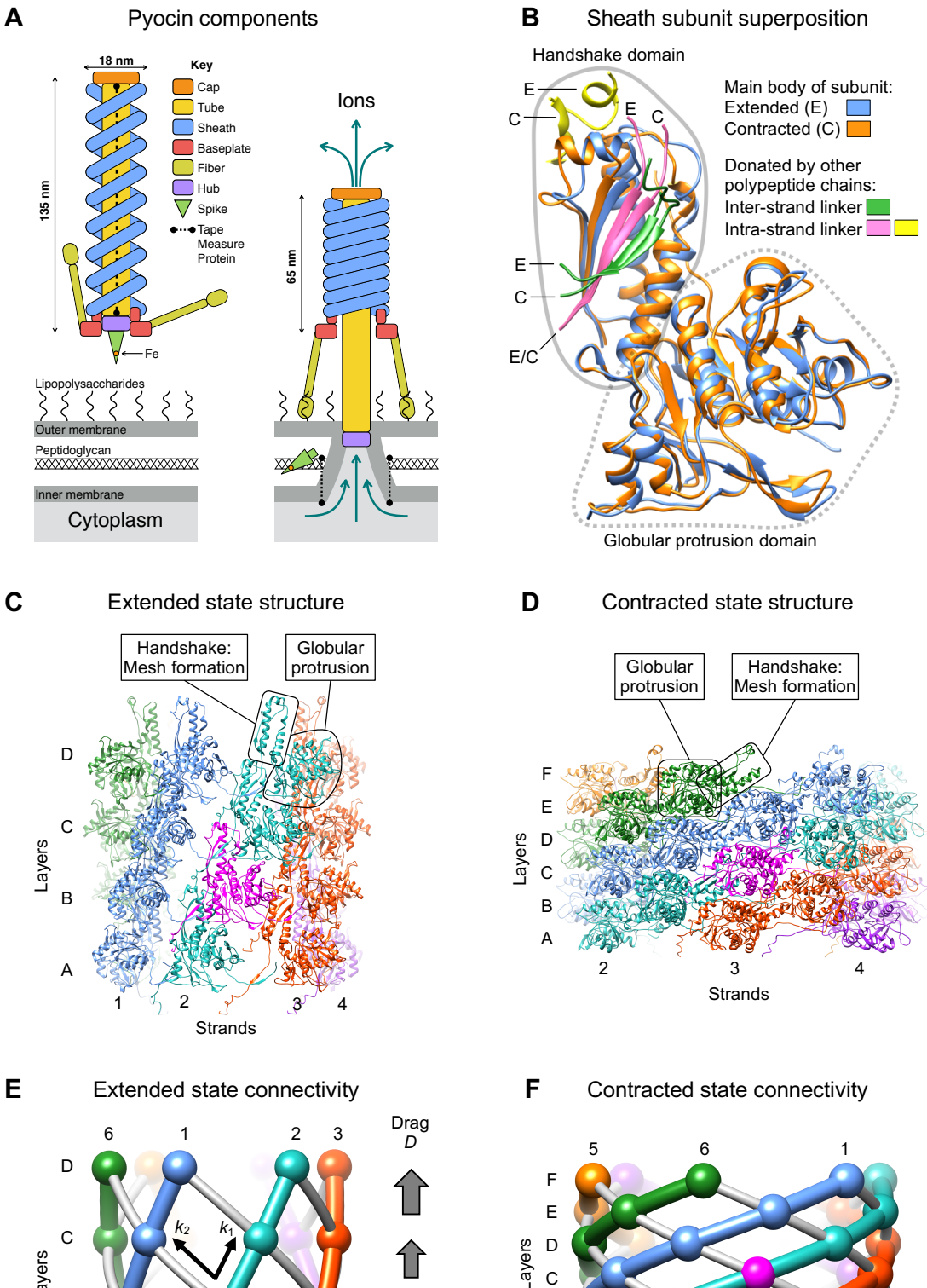

A
Pyocin components
18 nm
135 nm
Key
Cap
Tube
Sheath
Baseplate
Fiber
Hub
Spike
Tape Measure Protein
Fe
Lipopolysaccharides
Outer membrane
Peptidoglycan
Inner membrane
Cytoplasm
Ions
65 nm
B
Sheath subunit superposition
Handshake domain
E
C
E
C
Main body of subunit:
Extended (E)
Contracted (C)
Donated by other polypeptide chains:
Inter-strand linker
Intra-strand linker
E
C
E/C
Globular protrusion domain
C
Extended state structure
Handshake: Mesh formation
Globular protrusion
D
C
B
A
Layers
1
2
3
4
Strands
D
Contracted state structure
Globular protrusion
Handshake: Mesh formation
F
E
D
C
B
A
Layers
2
3
4
Strands
E
Extended state connectivity
6
1
2
3
D
C
B
A
Layers
k_2
k_1
1
2
3
4
Strands
Drag
D
F
Contracted state connectivity
5
6
1
F
E
D
C
B
A
Layers
2
3
4
Strands

Figure. 3.1. Structure of the end states and parametrization of the contraction reaction. (**A)** Schematic showing the main components and size of the pyocin particle free in solution (extended sheath) and attached to the cell surface (contracted sheath). (**B**) Superposition of sheath subunits in the extended and contracted states. The β-sheet of the handshake domain is completed by the inter- and intra-strand linkers that belong to adjacent polypeptide chains. The main body of the subunit in the extended and contracted conformations is colored in dodger blue and orange, respectively. The fragments originating from the neighboring polypeptide chains are shown in distinct colors and labeled with the letters E and C, which correspond to the extended and contracted conformations of the sheath, respectively. (**C**) Structure of a four-layer fragment of the pyocin sheath complex in the extended state(Ge et al. 2015). Each strand has a distinct color. One subunit is colored magenta to serve as a reference point. The strands are numbered, and the layers are labeled with letters. The boxed handshake domain lacks β-sheet components from adjacent subunits for clarity. (**D**) Structure of a six-layer fragment of the pyocin sheath in the contracted state(Ge et al. 2015). The color code and labeling nomenclature are as in panel (**C**). The boxed handshake domain lacks β-sheet components from adjacent subunits for clarity. **(E)** and (**F)** Diagrams demonstrating the topology and connectivity of polypeptide chains comprising the sheath in the extended and contracted states. The spheres represent subunit COMs. The sheath strands (intra-strand connections) are shown with colored tubes. The grey tubes indicate inter-strand connections. Also shown are the intra- and inter-strand coupling constants k_1 and k_2, and the drag parameter D.

The atomic structure of the sheath-tube complex of the R-type pyocin, T6SS, the Photorhabdus Virulence Cassette and its closely related Serratia Antifeeding Prophage in the extended and contracted states have been determined by cryo-electron microscopy (cryo-EM)(Ge et al. 2015),(Kudryashev et al. 2015),(Wang et al. 2017),(Jiang et al.

2019)·(Desfosses et al. 2019). Of these, the pyocin has the simplest architecture. Its sheath subunit consists of two domains, one of which is a component of an interconnected fishnet-like mesh that envelopes the tube (termed the 'handshake domain'(Kudryashev et al. 2015)) whereas the other forms a globular protrusion positioned at each node of this mesh (**Fig. 3.1B, 3.1C, 3.1D, 3.1E, 3.1F**). Other CISs build on top of this architecture by adding one or two domains to the protrusion domain and retaining all the other features of the pyocin sheath(Leiman and Shneider 2012). In all these systems, the structure of the individual subunit and the topology of the mesh connecting the handshake domains are preserved in both the contracted and extended states of the sheath(Ge et al. 2015)·(Kudryashev et al. 2015)·(Wang et al. 2017)·(Jiang et al. 2019)·(Desfosses et al. 2019). The structure of the tube is very similar in all CISs(Ge et al. 2015)·(Kudryashev et al. 2015)·(Wang et al. 2017)·(Jiang et al. 2019)·(Desfosses et al. 2019).

Despite the knowledge of the atomic structures of several CISs in the two end states, the mechanism by which chemical energy stored in the extended state is converted into the motion of the tube remains poorly understood. Previous theoretical work, which was performed before atomic structures of the sheath-tube complex became available, provides excellent insight into the contraction process but lacks quantitative details(Caspar 1980)·(Moody 1973)·(Falk and James 2006)·(Aksyuk et al. 2009). The available experimental data is sparse and somewhat contradictory. The enthalpy of sheath contraction measured for T4 ghosts (phage particles lacking DNA in which the sheath comprises less than 10% of the total material) varied by a factor of two depending on whether contraction was triggered by heat or urea(F, J, and H 1981). The activation energy of urea-induced contraction was found to be negative. The upper boundary for the timescale of contraction comes from studies of green fluorescent protein-labeled T6SS sheaths that are long enough (~10 times longer than phage tails) to be visualized in a fluorescence microscope(M. Basler et al. 2012). The actual timescale is nevertheless

unknown as contraction occurred faster than the 5 ms framerate of the microscope camera(M. Basler et al. 2012).

A few contraction intermediates of T4 and other phages have been captured in the electron microscope over the years, showing that the *in vitro* and *in vivo* triggered contractions start at the baseplate and propagate through the length of the sheath as a wave(Guerrero-Ferreira et al. 2019),(Moody 1973),(Eiserling 1967),(Donelli, Guglielmi, and Paoletti 1972). The sheath forms a narrow Christmas tree-like structure in phages targeting Gram-negative bacteria(Moody 1973) but contains a sharper transition from a wider contracted part to a narrower extended region in phages targeting Gram-positive hosts(Guerrero-Ferreira et al. 2019),(Eiserling 1967),(Donelli, Guglielmi, and Paoletti 1972). In the latter case, these intermediates are long-lived and could represent a functional state associated with the enzymatic digestion of the cell wall by enzymes located at the tip of the tail tube, an event that must precede the completion of sheath contraction during infection of a Gram-positive bacterium(Guerrero-Ferreira et al. 2019).

Recently, a computational approach that modeled the T4 sheath using Kirchhoff's rod theory with parameters derived from molecular dynamics (MD) simulations of a short fragment of the sheath has been presented(Maghsoodi et al. 2019). The elastic body calculations were parametrized to match the enthalpy of T4 sheath contraction(F, J, and H 1981) and the contraction timescale of the T6SS(M. Basler et al. 2012). Contraction was predicted to proceed via a rapid rotation of sheath subunits prior to their translation(Maghsoodi et al. 2017),(Maghsoodi et al. 2019). Such a sequence of events is incompatible with maintaining the integrity of the sheath subunit, which was implied but not validated in the approach. Furthermore, the model predicts that the free energy profile of the contraction process has an exponential form and a zero activation energy. Consequently, the forces developed in the second half of the contraction process when the

spike-tube complex comes into contact with the host membrane are vanishingly small and thus are insufficient for membrane puncture.

Here, we present a new modeling procedure that describes the free energy profile for the contraction process of the R-type pyocin sheath-tube complex in atomic detail. We developed a set of solution biophysics and single molecule experiments that characterize the activation energy E_a, the enthalpy of contraction, and the effective spring constant for the contracted sheath. Our modeling procedure correctly predicts, and solution biophysics measurements confirm, properties of pyocin mutants with an altered transition state structure.

RESULTS

Sheath contraction requires both theoretical and experimental characterization

Sheath contraction is a physicochemical reaction in which a change in the chemistry of the reactants (interactions between sheath subunits) is converted into mechanical work (motion of the tube). We characterize this process by a combination of atomic structure-based modeling and experimental measurements. Modeling aims to generate realistic sets of atomic structures which describe the contraction process. From this, the model predicts the following experimentally measurable parameters: the total free energy change, the activation energy, and the forces generated throughout contraction. Furthermore, the model describes the structure of the highest energy state (the transition state), which enables the manipulation of the activation energy by targeted mutagenesis. Therefore, modeling is integral in guiding our experiments, and consequently, the model must be described first, with experiment to follow.

Parametrization of the contraction reaction

In finding the most probable contraction pathway – a set of intermediate structures that describe the transition between the extended and contracted states – we must consider both the structural constraints and the energetics of the system. Considering that the mesh-like connectivity of the handshake domains (**Fig. 3.1C, 3.1D, 3.1E, 3.1F**) and the overall structure of the sheath subunit (**Fig. 3.1B**) are preserved in both the extended and contracted states, as well as the fact that the mesh linkers are integral parts of the handshake domain (**Fig. 3.1B**), we assumed that the fold of the subunit and, consequently, the connectivity of subunits are maintained throughout the contraction event. Furthermore, the architecture of the sheath's mesh and the near-perfect roundness of the sixfold-symmetric tube, whose structure does not change during contraction, does not allow the quaternary structure of the sheath-tube complex to significantly deviate from sixfold symmetry. Hence, we further assumed that contraction occurs in a sixfold symmetric manner.

Given these constraints, the instantaneous position of any sheath subunit in a contraction intermediate can be described by a set of six parameters (r, θ, z, ω, ϕ, κ) where (r, θ, z) are the cylindrical coordinates of the subunit's center of mass (COM) and (ω, ϕ, κ) is the set of polar angles describing the rotation of the subunit relative to the extended state. Notably, for a given subunit, the rotation axis, which is defined by the angles (ω, ϕ), is fixed throughout contraction in the COM reference frame (**Fig. 3.2A**). Accordingly, the rotation of the subunit can be effectively described by a single angle κ that spans a range of values from 0 to κ_{cnt}. Furthermore, given the constraint of preserving the connectivity and integrity of the handshake domains, the (r, θ, z, κ) parameters must

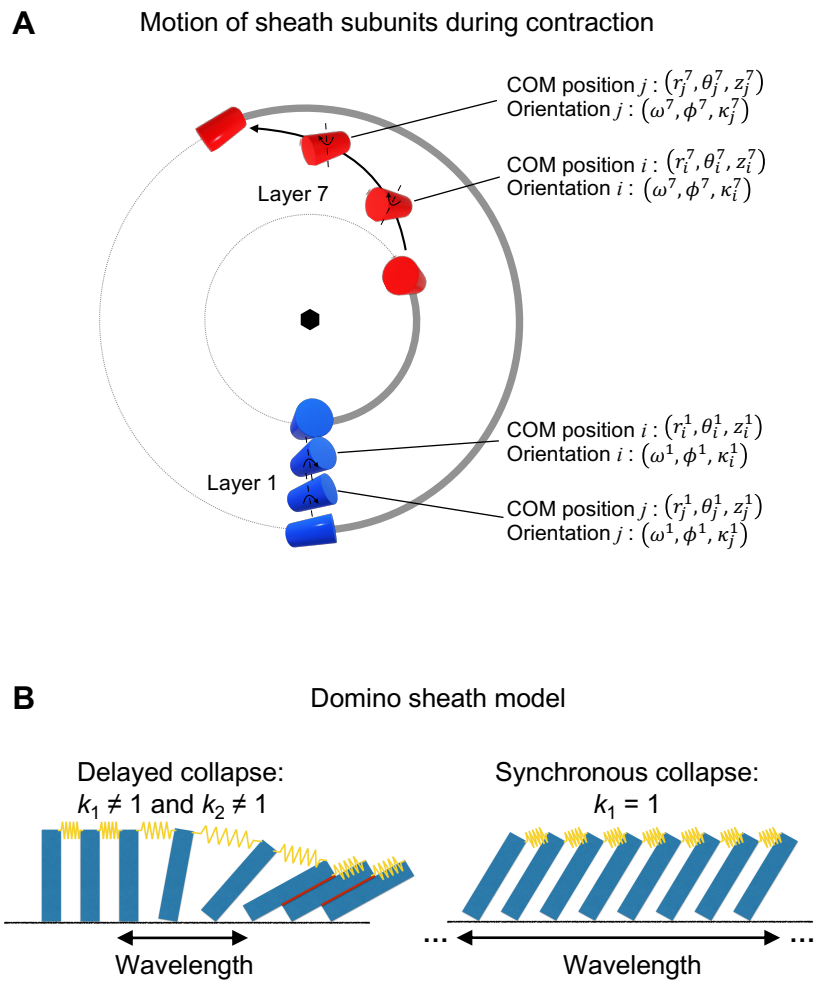


Figure. 3.2. The geometry and macroscopic approximation of sheath contraction. (A) A top view projection of the motion of the baseplate-proximal (Layer 1, blue) and a middle (Layer 7, red) sheath subunit belonging to the same strand (thick gray line) during contraction. The subunits are shown schematically as cylinders. The radial motion is exaggerated for clarity. (B) A row of dominoes connected by springs is a one-dimensional representation of the sheath. The red lines indicate favorable interactions between dominoes (or sheath subunits) that are realized in the lowest energy state (contracted state). The value of the spring constant defines whether the collapse of these dominos (or sheath contraction) is synchronous or delayed. In a delayed collapse, the contraction wavelength spans a finite number of subunits. In a synchronous collapse, the contraction wavelength is infinite.

be linearly related. Significant deviations from this linearity causes the handshake domains to disintegrate. Accordingly, we express this linear relationship by the means of a 'contracted fraction' parameter λ, with $\lambda = 0$ and $\lambda = 1$ corresponding to the extended and contracted states, respectively, such that

$$\forall\ (r, \theta, z, \kappa)\ \exists\ \lambda \in [0, 1]\ |\ (r, \theta, z, \kappa) = (1 - \lambda)\ (r_{ext}, \theta_{ext}, z_{ext}, 0) + \lambda\ (r_{cnt}, \theta_{cnt}, z_{cnt}, \kappa_{cnt}).$$

Finally, the contracted fraction of the entire sheath structure (our reaction coordinate) can be represented as an average of the contracted fractions of all sheath subunits.

An important consequence of the rigid body approximation for the motion of sheath subunits is that changes in the energetics of the sheath-tube complex are dominated by changes in interfacial interactions between subunits. The free energy of these interactions can be evaluated by summing the products of atomic solvent accessibilities and the free energy of solvation for every atom comprising the interface(Eisenberg and McLachlan 1986). Additionally, hydrogen bonds, salt bridges, and disulfide bonds (not applicable here), contribute to the interfacial energetics. Such an algorithm is implemented in PISA, which was designed for the identification of biologically meaningful interfaces in protein crystals(Krissinel and Henrick 2007). Here, we use PISA to calculate the total solvation energy of all contraction intermediates of the sheath-tube complex.

The search for a contraction pathway

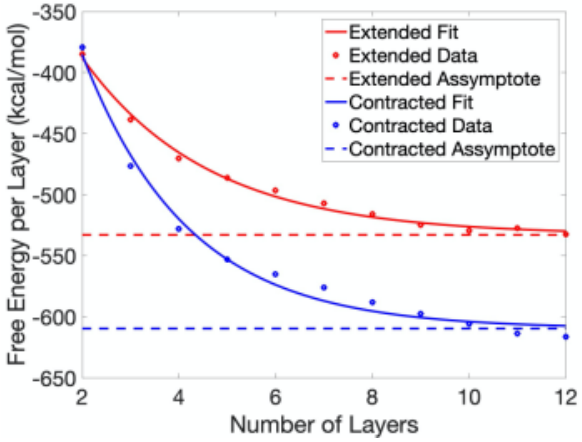
The most probable contraction pathway should exhibit the lowest activation energy and be devoid of substantial local minima since semi-contracted intermediates of pyocins have not been observed(Ge et al. 2020). We used a two-step procedure to identify such a pathway. First, we analyzed the contraction energetics of the smallest fragment of the

sheath-tube complex in which the relative influence of the solvent-exposed, terminal segments can be neglected (a 12-layer, 144-subunit segment of the sheath-tube complex) (**Fig. 3.3A**). Then, we extrapolated the best contraction pathway to the full-length 28-layer structure considering the structural constraints of the system.

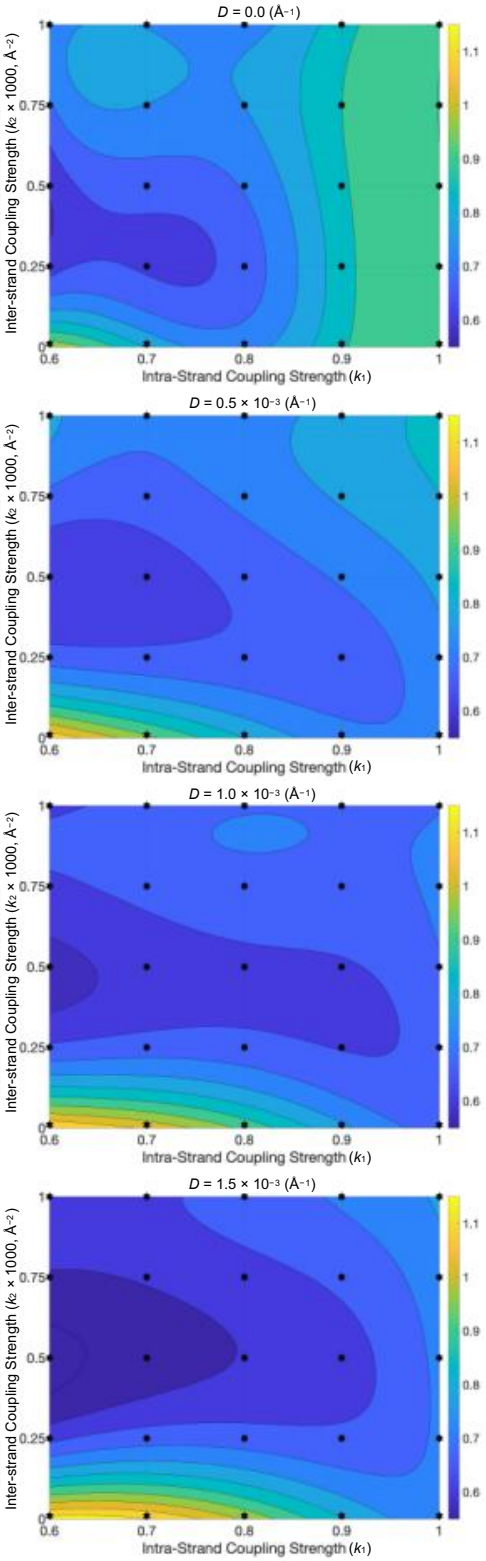
Different contraction pathways were generated by varying three parameters that determine the physical properties of the sheath structure: two spring-like constants k_1 and k_2 that describe the transfer of momentum between sheath subunits mediated by intra- and inter-strand linkers (respectively) and a drag-like parameter D that retarded the motion of baseplate-distant subunits and accounted for frictional and viscosity forces in the system (**Fig. 3.1E, 3.1F**). The range of values for each of the three parameters (k_1, k_2, D) was chosen such that the integrity of the handshake domains was maintained throughout contraction. Implementation of the above procedure as an algorithm constitutes the Domain Motion in Atomic Detail (DMAD) modeling method (see **Materials and Methods**).

One hundred contraction pathways, sampled as 5x5 matrices of (k_1, k_2) pairs for four different values of D, were generated (**Fig. 3.3B**). The pathways displayed different free energy profiles and activation energies (**Fig. 3.3C**). Notably, sheath-sheath subunit interactions dominated the energetics of the system while the initial sheath-tube subunit interfaces had negligibly small association energies (**Fig. 3.3C**). Optimal pathways occurred in a distinct region of the (k_1, k_2) plane for all values of the D parameter (**Fig. 3.3B**). This region is defined by low-to-intermediate values for both the intra- and inter-strand coupling. These contraction pathways can be described as ‘delayed’ or ‘wave-

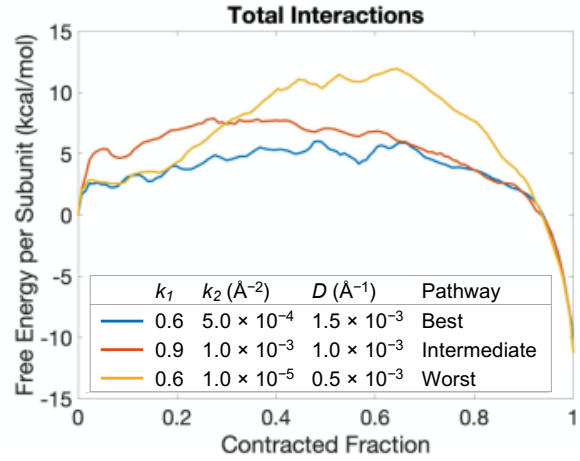
A Energy contribution per added layer



B Activation energies for different (k_1, k_2, D) parameter sets



C Free energy profiles of three contraction pathways



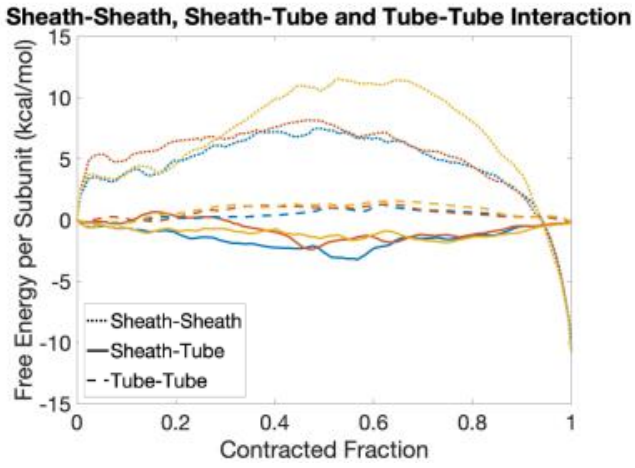


Figure. 3.3. DMAD analysis of the 12-layer sheath-tube fragment. (A) Free energy contribution per layer of sheath subunits as a function of the number of layers in the extended and contracted states. (B) Activation energies (as a fraction of ΔG) for 100 contraction pathways are plotted on the (k_1, k_2) planes for four different values of the drag parameter D. A third-degree polynomial surface is fitted to the data points. (C) Free energy profiles of contraction pathways with the lowest, intermediate, and highest activation energies. The top panel shows the total sum of all interactions. The bottom panel presents the interactions of each of the three component pairs: sheath-sheath, sheath-tube, and tube-tube.

like', akin to the collapse of dominoes connected by weak springs and surrounded by a viscous medium (**Fig. 3.2B**). These pathways are characterized by a 'contraction wavelength' defined as the smallest number of layers between near-fully contracted and near-fully extended subunits in an intermediate structure. The 'synchronous' contraction pathway is realized when the intra-strand coupling constant k_1=1, which corresponds to an infinitely stiff spring in the domino model (**Fig. 3.2B**). Its contraction wavelength is infinitely long, and its activation energy is higher than that of most delayed contraction pathways.

The best contraction pathway of the 12-layer fragment had an activation energy of 6 kcal/mol per sheath subunit (**Fig. 3.4A**). All structures in the pathway were characterized by a reasonable-to-good geometry as verified by Molprobity(Chen et al. 2009) (**Table 3.1)**. The structure of the handshake domain was maintained throughout contraction (**Fig. 3.4B**). Notably, the contraction wavelength was longer than the length of the structure (**Fig. 3.4C**). During contraction, the distances between the center of masses (COMs) of

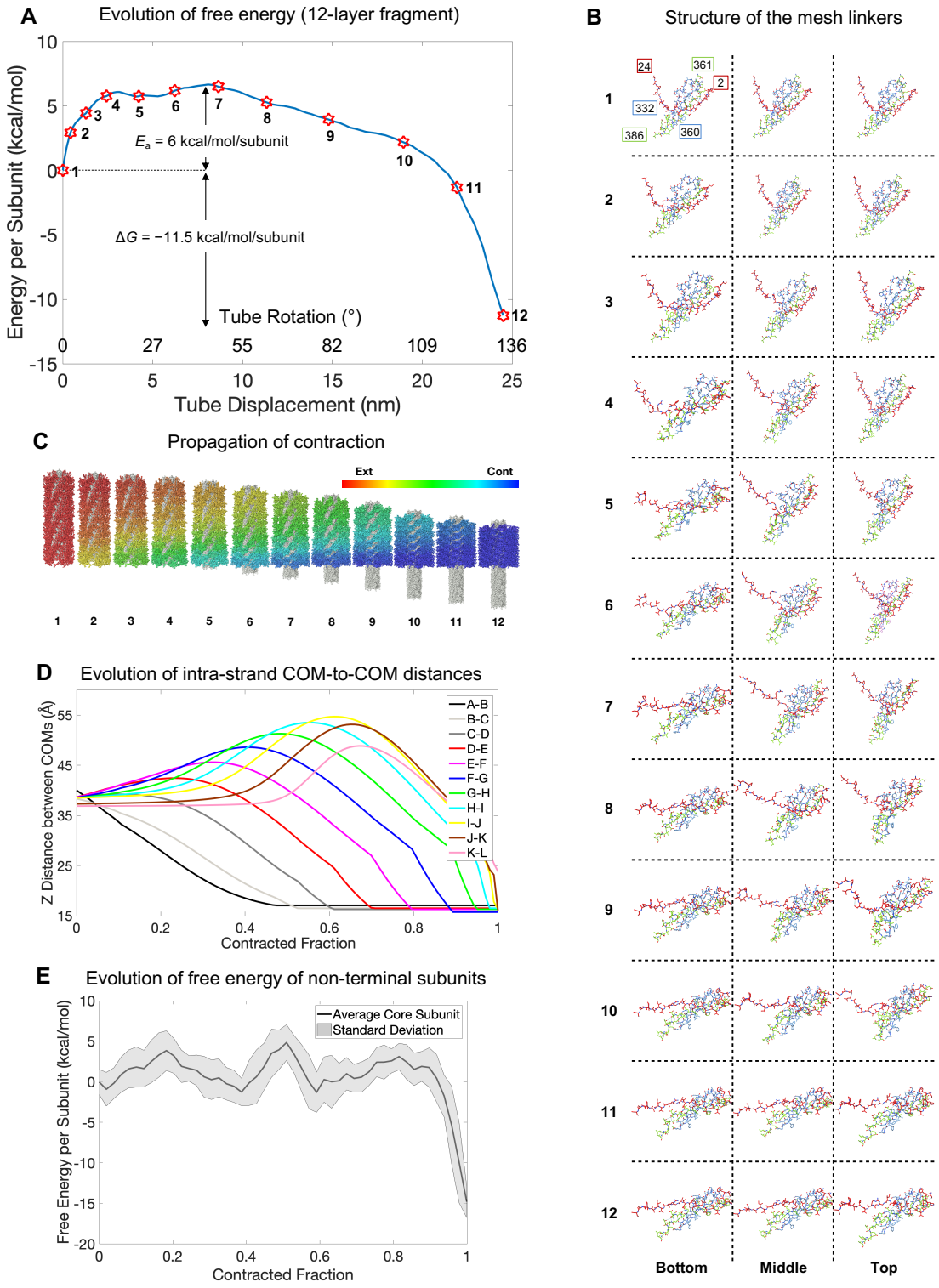

A
Evolution of free energy (12-layer fragment)
Energy per Subunit (kcal/mol)
Tube Displacement (nm)
Tube Rotation (°)
E_a = 6 kcal/mol/subunit
ΔG = −11.5 kcal/mol/subunit
B
Structure of the mesh linkers
Bottom
Middle
Top
C
Propagation of contraction
Ext
Cont
D
Evolution of intra-strand COM-to-COM distances
Z Distance between COMs (Å)
Contracted Fraction
A-B
B-C
C-D
D-E
E-F
F-G
G-H
H-I
I-J
J-K
K-L
E
Evolution of free energy of non-terminal subunits
Free Energy per Subunit (kcal/mol)
Contracted Fraction
Average Core Subunit
Standard Deviation

Figure. 3.4. DMAD-derived contraction pathway of the 12-layer fragment. (A) Free energy profile of the best contraction pathway of the 12-layer fragment. (B) Evolution of the structure of the intra- and inter-strand linkers throughout the contraction process for the best pathway. The three columns illustrate the bottom (baseplate-proximal), middle and top (baseplate-distal) layers of the structure. Each row shows one of the 12 conformations labeled with red stars in panel (A). Each polypeptide chain is in a distinct color. Residues 2-24 extend from the left-adjacent strand and one layer above subunit (inter-strand connectivity, colored red). Residues 361-386 extend from the subunit in the same strand and one layer above (intra-strand connectivity, colored green). (C) Propagation of contraction throughout the sheath. Sheath subunits are colored according to their contracted fraction with a color key given in the upper right corner of the panel. The free energy of the 12 intermediates shown are indicated with red stars in panel (A). (D) Evolution of the vertical component of the distance between COMs of sheath subunits belonging to the same strand in the 12-layer fragment throughout the best contraction pathway. Layers are labeled with consecutive letters A through L starting from the baseplate (layer A). (E) Average free energy profiles of 8 non-terminal subunits (belonging to layers 3 through 10 in the 12-layer fragment) are plotted as a function of their contracted fraction.

the sheath subunits and, consequently, the length of the sheath strands increased before collapsing into the compact contracted state (**Fig. 3.4D**). The sheath strands thus became hyper-extended, and the action of the sheath-tube system resembled that of a ballista.

Contraction of the full-length structure

Simulations of the 12-layer sheath-tube fragment revealed that given a set of (k_1, k_2, D) parameters, the free energy of non-terminal sheath subunits (layers 3 through 10 in

the 12-layer structure) is independent of the subunit's position within the strand and is uniquely determined by its contracted fraction (**Fig. 3.4E**). Similarly, given a set of (k_1, k_2, D) parameters, the contribution of the four terminal layers (the two bottom layers and the two top layers), in which not all inter-subunit interfaces are engaged in the contracted state, are the same in a structure of any length. These observations make it possible to extrapolate a contraction pathway of the 12-layer fragment to that of the full-length 28-layer structure and to obtain a free energy profile of contraction for the full-length sheath (see **Materials and Methods**). At the same time, pathways in which the vertical distance between neighboring subunit's COMs exceeds 55 Å, the maximal value found for the 12-layer fragment (**Fig. 3.4D**), cannot be realized as this is incompatible with maintaining the structural integrity of the sheath. For this reason, all but the lowest drag pathways are prohibited for the full-length pyocin structure. Furthermore, both the drag D and inter-strand constant k_2 had to be rescaled to take into account the longer paths traveled by baseplate-distal sheath subunits in the full-length structure (see **Materials and Methods**). After these considerations, optimal parameters from the 12-layer fragment simulations were used, namely, $k_1 = 0.7$ [dimensionless], $k_2 = 2.5\times10^{-4}$ Å^{-2}, and $D = 5\times10^{-4}$ Å^{-1}. This contraction pathway had an activation energy of 201±12 kcal/mol, which corresponded to ~9% of the total energy released (2,210 kcal/mol) (**Fig. 3.5A**).

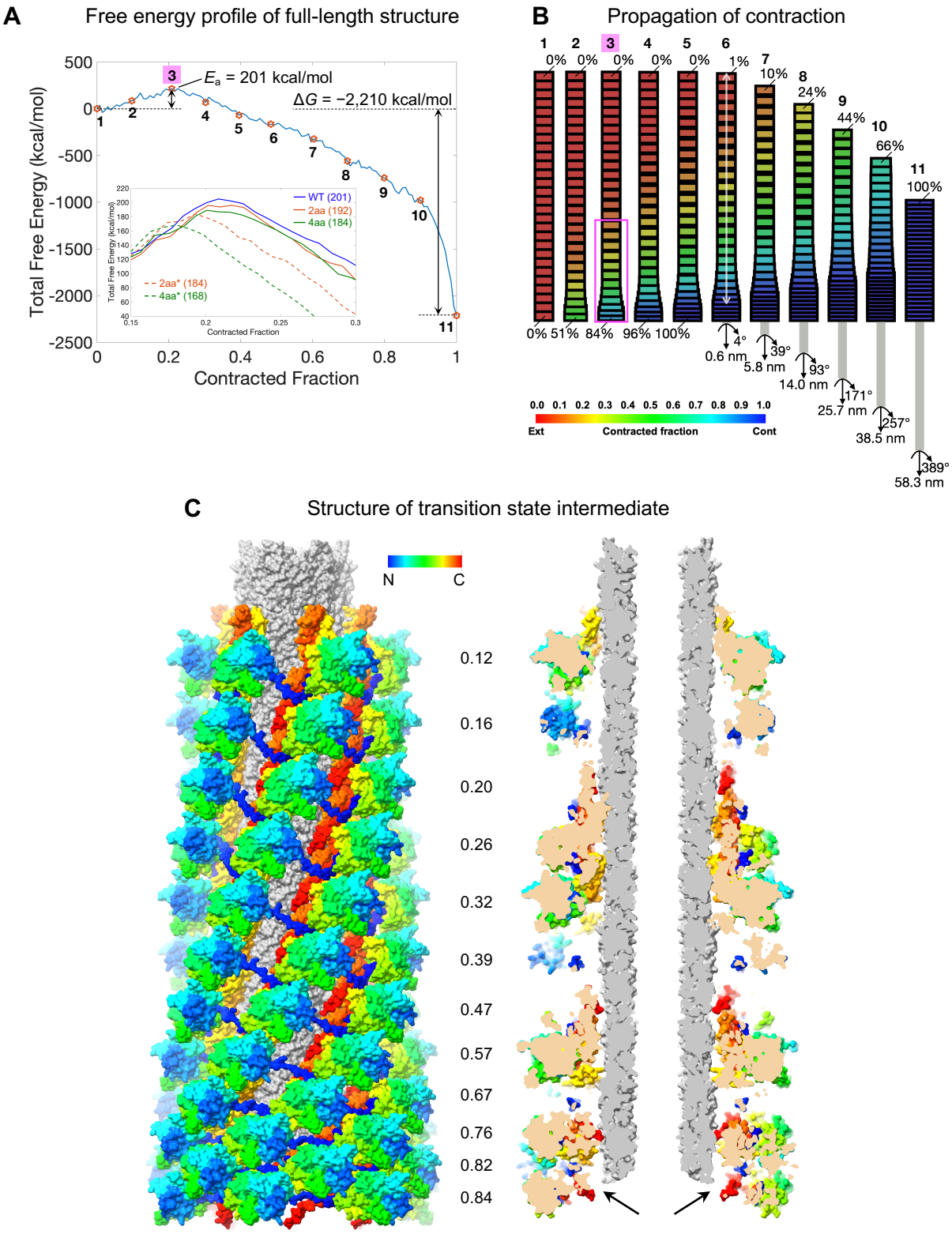

A
Free energy profile of full-length structure
Total Free Energy (kcal/mol)
Contracted Fraction
E_a = 201 kcal/mol
ΔG = −2,210 kcal/mol
WT (201)
2aa (192)
4aa (184)
2aa* (184)
4aa* (168)
B
Propagation of contraction
Ext
Contracted fraction
Cont
C
Structure of transition state intermediate
N
C

Figure. 3.5. Free energy profile of contraction and the structure of the transition state. (**A)** DMAD-derived evolution of the free energy of the sheath-tube complex during contraction. Eleven red stars mark the free energy of conformations shown schematically in panel (**B)**. Intermediate 3 (magenta background) is the transition state. The inset shows a fragment of free energy profiles for sheath mutants with inter-strand linkers carrying two and four additional residues (labeled 2aa and 4aa, respectively). The solid lines correspond to the WT simulation and modifications of the inter-strand constant k_2 described by the following (k_1, k_2, D) parameters: WT (0.7, 2.5×10^{-4} Å^{-2}, 5×10^{-4} Å^{-1}), 2aa (0.7, 1.4×10^{-4} Å^{-2}, 5×10^{-4} Å^{-1}), 4aa (0.7, 8×10^{-5} Å^{-2}, 5×10^{-4} Å^{-1}). The dashed lines to modification of both k_1 and k_2: 2aa* (0.665, 2.375×10^{-4} Å^{-2}, 5×10^{-4} Å^{-1}), 4aa* (0.63, 2.25×10^{-4} Å^{-2}, 5×10^{-4} Å^{-1}). All profiles shown in the inset have been smoothed for clarity. (**B)** Propagation of contraction throughout the pyocin structure. The sheath layers are colored according to their contracted fraction using the color code given in the lower part of the panel. The width and height of the sheath layers is also adjusted to match the contracted fraction. The transition state intermediate (3) is labeled with a magenta background. The magenta rectangle highlights the twelve baseplate-proximal layers of the transition state shown in panel (**C)**. The semitransparent white line (Intermediate 6) represents the contraction wavelength. (**C**) Structure of the twelve baseplate-proximal layers of the full-length sheath-tube complex in the transition state. The color of each sheath subunit varies along the polypeptide chain as a continuous spectrum with the N-terminus in blue and the C-terminus in red. The tube is colored gray. The cutaway view panel on the right shows baseplate-proximal sheath subunits which have dissociated from the tube (black arrows), but the tube has yet to move. The contracted fraction of each sheath layer is given between the panels.

Structure of the transition state and the wavelength of contraction

Sheath contraction is thought to be triggered by a large conformational change of the baseplate, which propagates through the sheath via direct and near-rigid body interactions between the baseplate and baseplate-proximal sheath subunits(Jiang et al. 2019),(Ge et al. 2020),(Taylor et al. 2016). The transformation of the baseplate likely provides the activation energy necessary to reach the transition state. Accordingly, a putative transition state should exhibit baseplate and baseplate-proximal sheath subunits in a near-contracted state. Additionally, the structure of the transition state should allow for a return to its original, fully extended state.

In the transition state predicted by the DMAD methodology, the baseplate-proximal ('bottom') sheath subunits of the sheath are ~84% contracted whereas the baseplate-distal ('top') subunits are fully extended (Intermediate 3 in the pathway in **Fig. 3.5A, 3.5B**). As a consequence, the bottom sheath subunits separate from the tube whereas the top part of the sheath remains in the extended state and interacts with the tube (**Fig. 3.5C**). This intermediate is comparable to the structure of the phage T4 tail with its baseplate in the post-attachment state and the tail not yet contracted, which has been imaged attached to the cell surface by cryo-electron tomography(Hu et al. 2015). Further along the contraction pathway is an intermediate in which the fifth subunit from the bottom and the top subunit are ~98% and ~1% contracted, respectively (Intermediate 6 in **Fig. 3.5A, 3.5B**). This distance corresponds to the smallest number of layers between near-fully contracted and near-fully extended subunits in an intermediate structure. Thus, the contraction wavelength spans 24 layers and is approximately equal to the size of the entire complex (**Fig. 3.5B**).

Probing contraction with solution biophysics

To gain further insight into the sheath contraction process, we probed the energetics of the system via a series of solution biophysics experiments (**Fig. 3.6**). For this, a protocol

for purification of pyocin particles of high purity in the extended and contracted states has been developed (**Fig. 3.7**). We found that pyocins contract in a narrow interval of temperatures near ~70 °C. Contraction could also be triggered by an acidic buffer (pH < 3.0). This made it possible to measure the enthalpy of heat- and pH- triggered contraction with the help of Differential Scanning Calorimetry (DSC) and Isothermal Titration Calorimetry (ITC).

DSC curves of extended and contracted pyocins contained a positive peak in the 60-80 °C interval. No morphological changes in the contracted specimen were associated with this transition, therefore this event likely corresponded to the denaturation of a baseplate or a neck component. The contribution of this transition could be accounted for by subtracting the contracted particle DSC curve from that of the extended particle. The resulting enthalpy of sheath contraction was −13.6±1.2 kcal/mol/subunit (i.e. per sheath subunit) (**Fig. 3.6A**).

The enthalpy of pH-induced contraction was measured by titrating a pyocin sample having a total ionic strength of ~2 mM into a cell that contained 50 mM Glycine-HCl pH 2.5. The contribution of solvation was accounted for by comparing the enthalpies of the extended and contracted pyocins belonging to the same biological replicate and dialyzed into the same low ionic strength buffer. In this case, the enthalpy of contraction was −10.1±1.6 kcal/mol/subunit (**Fig. 3.6B**).

Thus, the experimentally measured enthalpy associated with sheath contraction is nearly equal to the free energy difference calculated by the DMAD procedure (−13.2 kcal/mol/subunit, **Fig. 3.5A**), showing that the entropic contribution to contraction is small, further validating the DMAD approach.

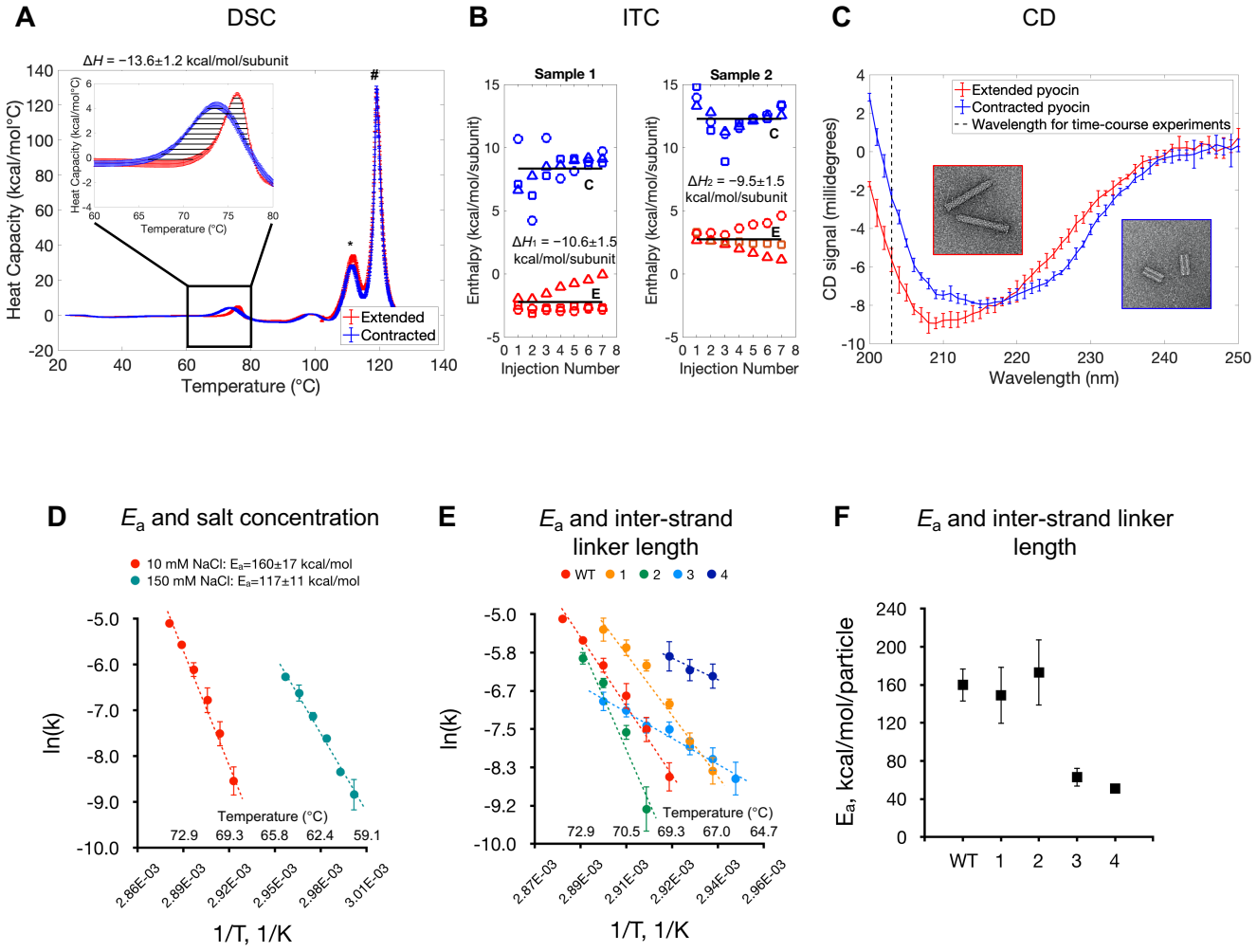


Figure. 3.6. Characterization of contraction reaction using solution biophysics. (A) DSC profiles of extended and contracted pyocins. The inset highlights the temperature range where temperature-induced sheath contraction occurs. The enthalpy of contraction is the integrated area of the striated region. The star (*) and hash (#) symbols indicate the peaks associated with heat absorbed during the denaturation of the tube and sheath, respectively. The average and standard deviations of three technical replicates are plotted. Error bars represent the standard deviation. The experiment was repeated for two biological replicates. (B) Enthalpy of pH-induced pyocin contraction is measured using an inverted ITC setup. The extended (E) and contracted (C) pyocins were titrated into a cell containing a pH 2.5 buffer. Three replicates of both, extended and contracted samples for two

biological replicates were measured (each titration is labeled with Δ's, o's or ◇'s). The averages are shown with solid lines. (C) CD spectra of extended and contracted pyocins. The dashed vertical line indicates the wavelength used in time course measurements. The insets show EM images of the samples used in these experiments. The curves are averages of three technical and three biological replicates. Error bars correspond to the standard deviation. (D) Arrhenius plot of temperature dependent contraction rates of the WT pyocin for two buffer conditions. Data points are averages of three technical and three biological replicates. Error bars represent the standard deviation. (E) Arrhenius plot of temperature dependent contraction rates for WT and four sheath mutants. Data points are averages of three technical and two biological (three for WT) replicates. Error bars represent the standard deviation. (F) Activation energy plot as a function of the inter-strand linker length. Error bars represent 95% confidence interval for the Arrhenius fit.

The folds of the sheath subunit in the extended and contracted states are very similar (**Fig**. **3.1B**), but their spatial arrangement (the helical symmetry) (**Fig**. **3.1A, 3.1C, 3.1D, 3.1E, 3.1F**), the contacts with each other (**Fig**. **3.1C, 3.1D, 3.1E, 3.1F**) and the conformation of the inter-strand linker are different (**Fig**. **3.1C, 3.1D, 3.1E, 3.1F**). The synergetic combination of these effects results in small but detectable differences in the circular dichroism (CD) spectra of contracted and extended pyocins (**Fig. 3.6C**). This property made it possible to monitor heat-induced contraction in a solution ensemble in real time. This process was found to be a first order reaction with a temperature-dependent rate (**Fig. 3.8A**). Assuming the Arrhenius model and a temperature-independent activation energy, the logarithm of the turnover rate is inversely proportional to the reaction's temperature with a coefficient of $-E_a/R$, where E_a is the activation energy and R is the universal gas constant. Such a measured activation energy was found

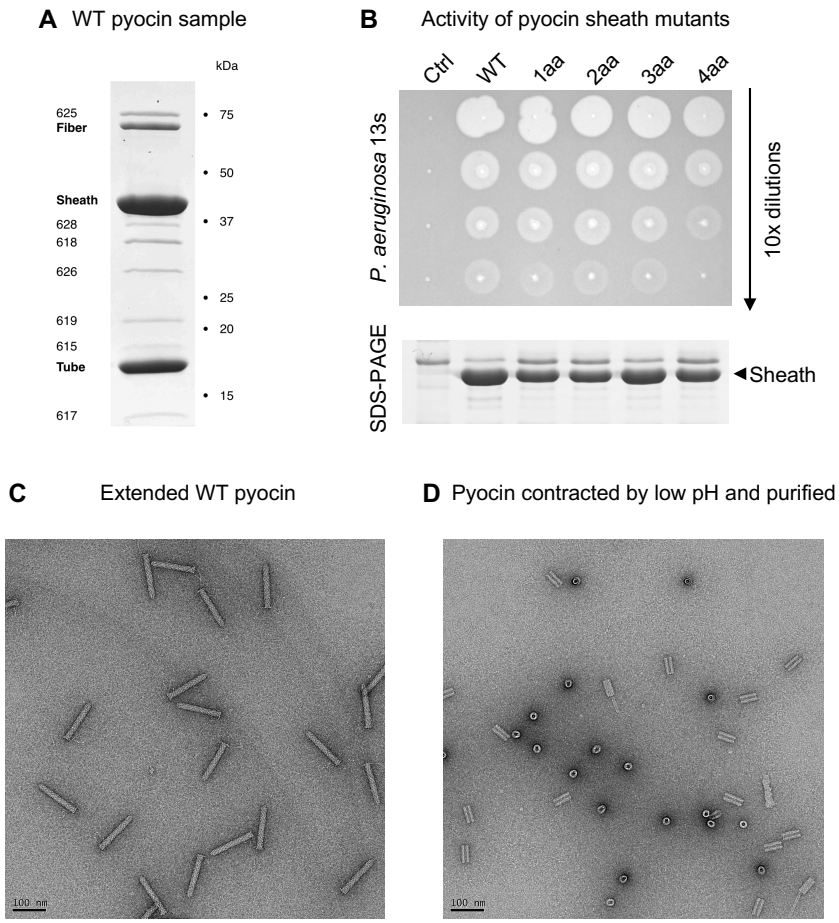


Figure. 3.7. Biophysical and functional characterization of the WT pyocin and its sheath mutants. (A) Coomassie-stained SDS-PAGE of a typical WT pyocin sample that was used in solution biophysics experiments. All bands correspond to known pyocin proteins and no detectable impurities are present. The proteins are identified with their trivial names or locus number in *P. aeruginosa* PAO1 genome (e.g. 618 stands for PA0618). PA0616 (the central spike protein, MW = 19.4 kDa, three copies per particle) and PA0627 (MW = 7.5 kDa) are not visible in this SDS-PAGE. (B) The killing activity of the WT pyocin and the four mutants with an insertion of one (1aa) to four (4aa) amino acids in the inter-strand linker is evaluated by a double agar overlay spot assay. In a dilution series up to a certain concentration, active pyocins "burn" a visible spot on a lawn of a sensitive *P. aeruginosa* 13s strain by lysing a large fraction of cells in that spot. A fragment

of a Coomassie-stained SDS-PAGE centered on the pyocin sheath band shows the relative amount of pyocins in each sample. 'Ctrl' stands for cells that contained an empty vector and did not produce pyocins (negative control sample). (C) and (D) EM images (negative staining) of extended and contracted WT pyocin particles (respectively) used in CD, DSC, and ITC measurements.

to be 160±17 kcal/mol/particle in a low salt buffer and decreased to 117±11 kcal/mol/particle as the salt concentration increased (**Fig. 3.6D**). This is in qualitative agreement with the DMAD-derived value of 201±12 kcal/mol/particle, which was calculated for the system in pure water (**Fig. 3.5A**).

Inter-strand linker length affects the activation energy

To further validate the results of biophysical experiments and DMAD modeling, we examined the activation energies of pyocin mutants with altered inter-strand linkers. We reasoned that lengthening these linkers, most of which are fully stretched in the transition state (**Fig. 3.5C**), should allow the structure to sample a larger parameter space of available contraction pathways, resulting in a lower activation energy(Drake, Harris, and Pettitt 2016).

A full, quantitative characterization of the contraction of sheath mutants with linker insertions by DMAD requires the knowledge of the structure of these linkers in both end states, which is unavailable. However, considering the geometry of the system (the linkers run at various angles to the long axis of the particle), two- and four-residue insertions into the inter-strand linker will allow for the maximum subunit COM separation to increase by ~4 Å and ~8 Å, respectively. In the DMAD procedure, this is achieved by decreasing the

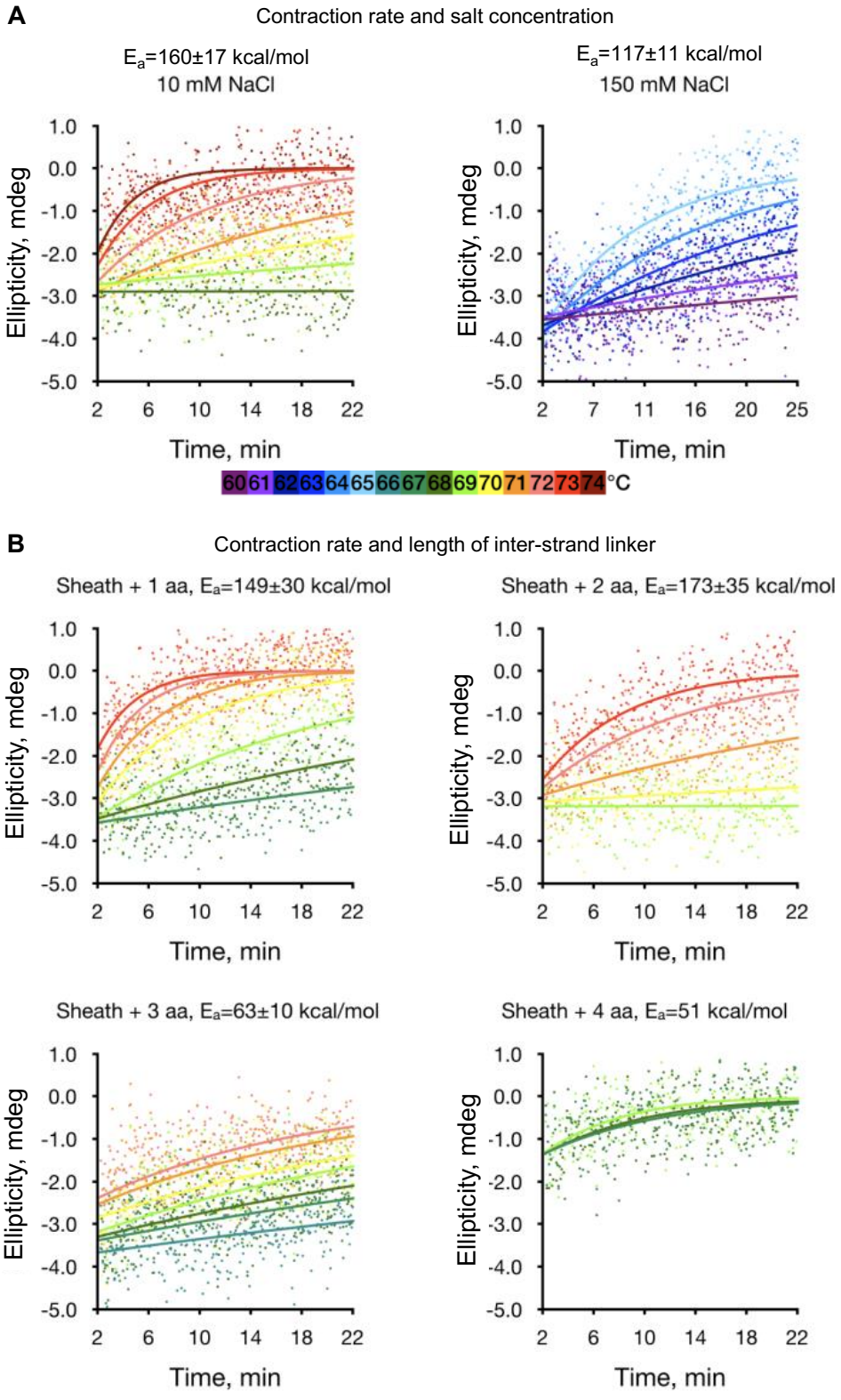

A
Contraction rate and salt concentration
Ea=160±17 kcal/mol
10 mM NaCl
Ea=117±11 kcal/mol
150 mM NaCl
Ellipticity, mdeg
Time, min
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 °C
B
Contraction rate and length of inter-strand linker
Sheath + 1 aa, Ea=149±30 kcal/mol
Sheath + 2 aa, Ea=173±35 kcal/mol
Sheath + 3 aa, Ea=63±10 kcal/mol
Sheath + 4 aa, Ea=51 kcal/mol

Figure. 3.8. CD spectroscopy of heat-induced contraction. (A) Heat-triggered contraction of WT pyocin measured by CD at a wavelength of 203 nm at various temperatures in buffers with two different salt concentrations. The temperature is color-coded according to the color bar given below the panels. The data point series are fitted with exponential functions. Each color corresponds to an average of three technical replicates. (B) Heat-triggered contraction of pyocin mutants carrying one, two, three, and four additional amino acids in the inter-strand linker. The color code is the same as in panel (A).

k_2 constant to 1.4×10^{-4} Å^{-2} and 8×10^{-5} Å^{-2}, respectively. The resulting activation energies are progressively lower, albeit within the measure of uncertainty: 192±14 kcal/mol and 186±14 kcal/mol, respectively (**Fig. 3.5A**, inset, solid lines).

Experimentally, mutants with insertions of one, three, and four amino acids in the inter-strand linkers were progressively less stable, contracted at lower temperatures (**Fig. 3.6E**), and had lower activation energies than the wild type (WT) structure (**Fig. 3.6F** and **Fig. 3.8B**). The mutant with a two-residue insertion was more stable and had a higher activation energy most likely due to a pleotropic effect. The four-residue insertion mutant was insufficiently stable to withstand the rigorous purification procedure required for solution biophysics experiments and its contraction kinetics could only be recorded for three temperature points.

All sheath mutants were functionally active (**Fig. 3.7B**). The killing capacities of mutants with one to three residue linker insertions were similar to that of the WT (**Fig. 3.7B**) while the four-residue insertion mutant was less active. The value of the activation energy, which determines the stability of the particle, affects the shape of the DMAD-derived free energy profile in the beginning of the contraction trajectory whereas the force needed to penetrate the membrane, which likely determines the killing capacity, is dictated

by the shape of the free energy profile at the end of the contraction process (**Fig. 3.5A** and **Discussion**). Hence, the comparable activities of the sheath linker insertion mutants agree with the DMAD-predicted behavior of the system.

One explanation for the non-linear behavior of these mutants is that the inter-strand linker insertions affected both the inter- and intra-strand linkers. The DMAD methodology allows us to examine this supposition. By reducing the optimal values of the intra- and inter-strand constants by 5% and 10% (k_1 from 0.7 to 0.665 and 0.63, and k_2 from 2.5×10^{-4} $Å^{-2}$ to 2.375×10^{-4} $Å^{-2}$ and 2.25×10^{-4} $Å^{-2}$) for the two- and four-residue insertion, respectively, while keeping D unchanged results in activation energies of 173±25 kcal/mol and 161±16 kcal/mol (**Fig. 3.5A**, inset, dashed lines). Interestingly, not only do the activation energies decrease but the transition state intermediate shifts to an earlier, more extended state in the contraction pathway. Such a state is likely to be attained at a temperature lower than that of the WT, as observed experimentally.

Single-molecule imaging and force measurements of sheath extension

To understand the energetics of the system at the very end of the contraction process and to probe contraction in a single particle regime, we measured the elastic properties of fully contracted sheaths by atomic force microscopy (AFM). When imaged by high-speed AFM(Ando et al. 2001), the structural features and dimensions of pyocins adsorbed to the mica surface matched those found in high resolution cryo-EM studies(Ge et al. 2015),(Ge et al. 2020) showing that the interaction with mica does not disturb the structure (**Fig. 3.9A**). Functionalizing the mica substrate with poly-lysine made it possible to orient tubeless contracted sheaths vertically (**Fig. 3.9B**). Such particles displayed two distinct

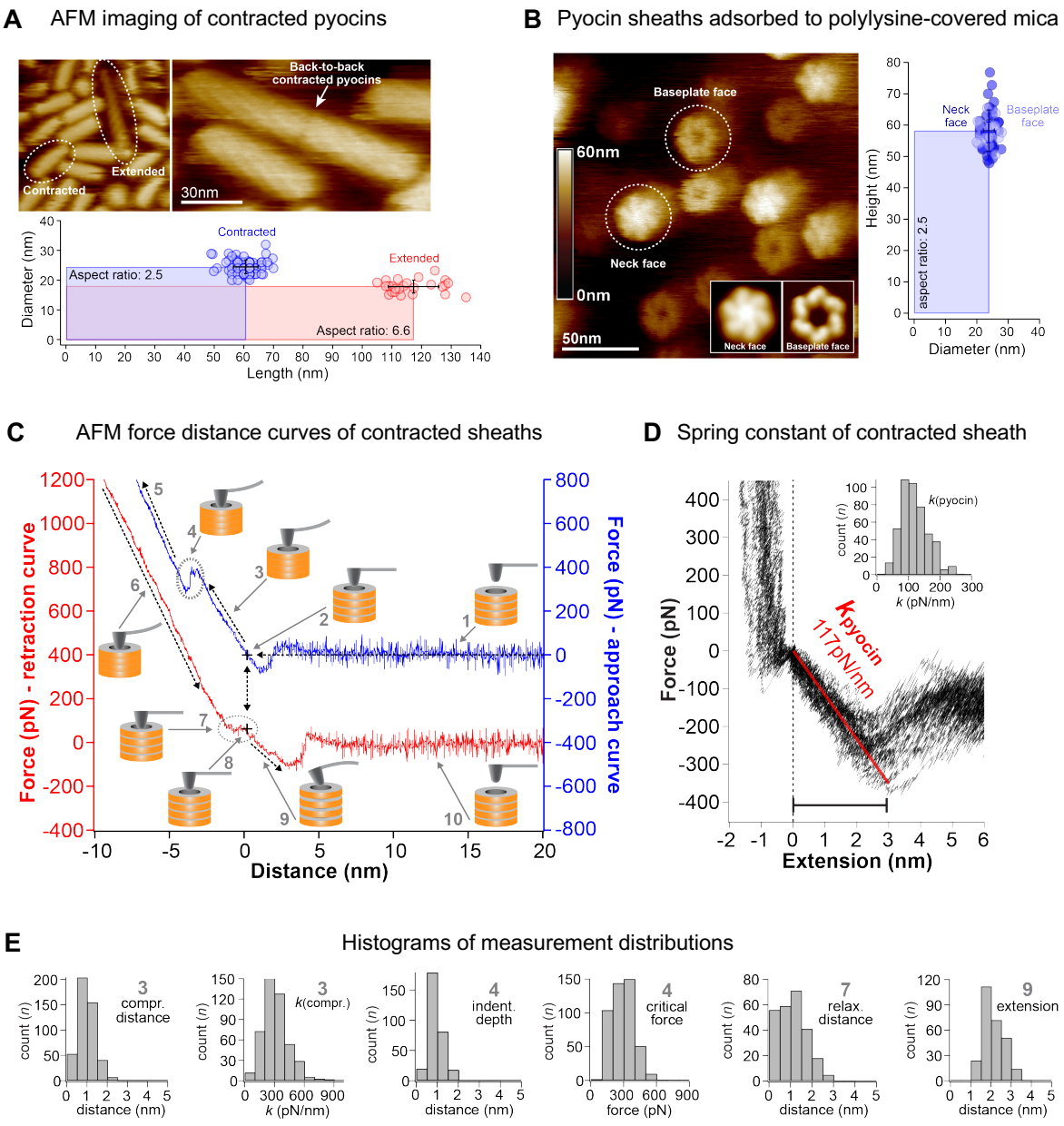


Figure. 3.9. Probing physical properties of the pyocin sheath with AFM. (A) AFM images of a pyocin sample adsorbed to the mica surface. The AFM-derived dimensions of the extended and contracted species (lower panel) match those recorded in EM. (B) AFM images of contracted pyocin sheaths standing vertically on a polylysine-treated mica surface (left subpanel). The baseplate face can be clearly distinguished from the neck face. The inset shows slightly enlarged images of the neck and baseplate faces obtained by

averaging multiple images and applying sixfold symmetry. The panel on the right shows the dimensions of the contracted sheaths in both orientations measured by AFM. (C) AFM force-distance curves of contracted sheaths. The cantilever approach and retraction curves are shown in blue and red, respectively. For clarity, they are plotted with an offset along the Y axis. The following five stages were registered during approach: 1) the tip vibrates freely before contacting the particle; 2) attractive interaction between the tip and the sheath; 3) compression of the sheath; 4) insertion of the tip into the sheath channel; 5) bending of cantilever. Retraction contained the following five stages: 6) straitening of cantilever (inverse of 5); 7) relaxation of the cantilever at zero force; 8) zero-distance adhesive force; 9) extension of the pyocin sheath; 10) the tip separates from the sheath and vibrates freely. All data was calculated by eliminating cantilever deflection. (D) Representative pyocin extension curves, with a linear fit of the extension spring constant k_{pyocin} ~ 117 pN/nm (red line) and a histogram of the spring constant distribution (inset). (E) Distributions of various parameters measured by AFM in stages 3, 4, 7, and 9 of the force distance curve (see panel C) are shown in the form of histograms. All measurements were performed for two biological replicates.

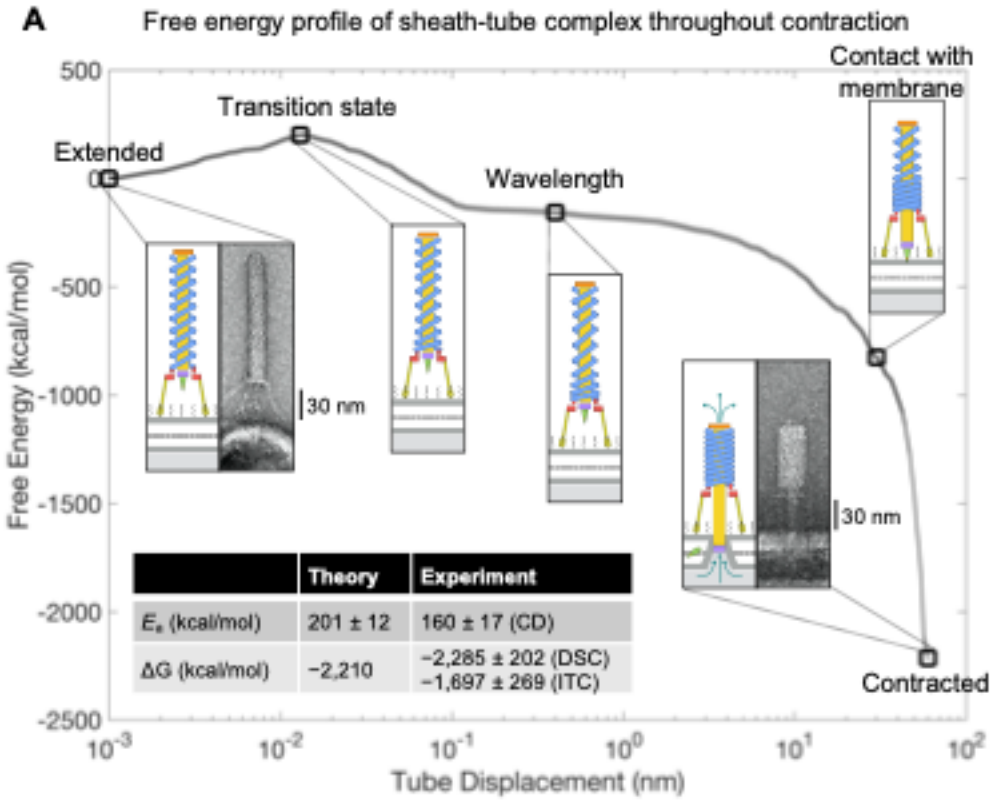
ends: a left-handed windmill hexameric structure with a ~10nm wide opening and a closed cap structure, which corresponded to the baseplate and neck faces of the contracted pyocin, face, the AFM was set to acquire force-distance cycles (**Fig. 3.9C**) by inserting the tip into the sheath opening and measuring the spring constant by stretching the particle during tip retraction (**Fig. 3.9D**).

A typical force measurement cycle consisted of 10 stages (**Fig. 3.9C**). In stage 1, the tip approached the contracted sheath, and no interaction force was detected until attractive (likely Van der Waals) forces occurred at ~1 nm tip-sheath separation. A cantilever-sheath contact at zero-force followed (stage 2, black cross on blue trace). Upon

further approach, the cantilever reported a repulsive force regime (stage 3), which corresponded to the compression of the sheath ($k_{compression}$ ~ 310 pN/nm) (**Fig. 3.9E**). The compression distance, which constituted the difference between the displacements of the piezo stage (~3.5 nm) and the cantilever deflection (~2.5 nm), was only ~1 nm, or ~1.6% of the ~60 nm height, indicating that the contracted sheath was essentially incompressible. Upon reaching a load of several hundred pNs on the cantilever ($F_{critical}$ ~ 290 pN), a relaxation (stage 4) was observed in ~70% of the approach curves (d_{indent} ~ 0.8 nm). We interpreted this event as the tip sliding into the sheath channel. Stage 5 described another linear repulsive force regime (with an up to ~2 nN applied force) with a steeper slope that matched that of the cantilever pushing against the bare mica support as recorded in control measurements before and after the stretching experiments. In stage 6, the cantilever was retracted and straightened. Upon complete cantilever relaxation, a zero-force regime was observed (stage 7) that spanned ~1 nm ($d_{relaxation}$ ~ 1.1 nm) (**Fig. 3.9E**). We interpreted this regime as the reverse process of stage 4 since the relaxation distance is very similar to the ingrain distance and terminates precisely at the point of contact (stage 2) in the approach curve (two-headed dashed arrow). Like stage 4, this regime was not found in all curves and/or it varied in length. When the tip-sample separation distance was further increased, beyond the point of the initial tip-sample contact in the approach (stage 8), an adhesive force regime was detected (stage 9). In this regime, the cantilever is stretching the sheath. The sheath could be extended by an average of ~2 nm ($d_{extension}$ ~ 2.1 nm) upon which the adhesive force on the cantilever reached ~250 pN (**Fig. 3.9E**). Upon further extension, the non-specific bond between the cantilever and the sheath broke and the cantilever snapped back to its relaxed zero-force state (stage 10). Force curve cycles could be repeated on the same particle several times until it fell over or disintegrated (**Fig. 3.9D**). Stage 9 describes the force/extension response of the pyocin sheath in its last nanometers of contraction. The spring constant of the contracted pyocin sheath is 117±20 pN/nm (**Fig. 3.9D**).

DMAD-derived force profile of the contraction reaction contains two phases

The excellent agreement between the predicted energetics of sheath contraction and experiment allows us to speculate about the forces developed by the particle throughout contraction from the DMAD theory (**Fig. 3.10**). This force can be estimated by taking the negative derivative of the free energy with respect to tube displacement (**Fig. 3.10A, 3.10B**).


A
Free energy profile of sheath-tube complex throughout contraction
Contact with membrane
Transition state
Extended
Wavelength
Free Energy (kcal/mol)
500
0
-500
-1000
-1500
-2000
-2500
30 nm
30 nm
Theory
Experiment
E_a (kcal/mol)
201 ± 12
160 ± 17 (CD)
ΔG (kcal/mol)
−2,210
−2,285 ± 202 (DSC)
−1,697 ± 269 (ITC)
Contracted
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Tube Displacement (nm)

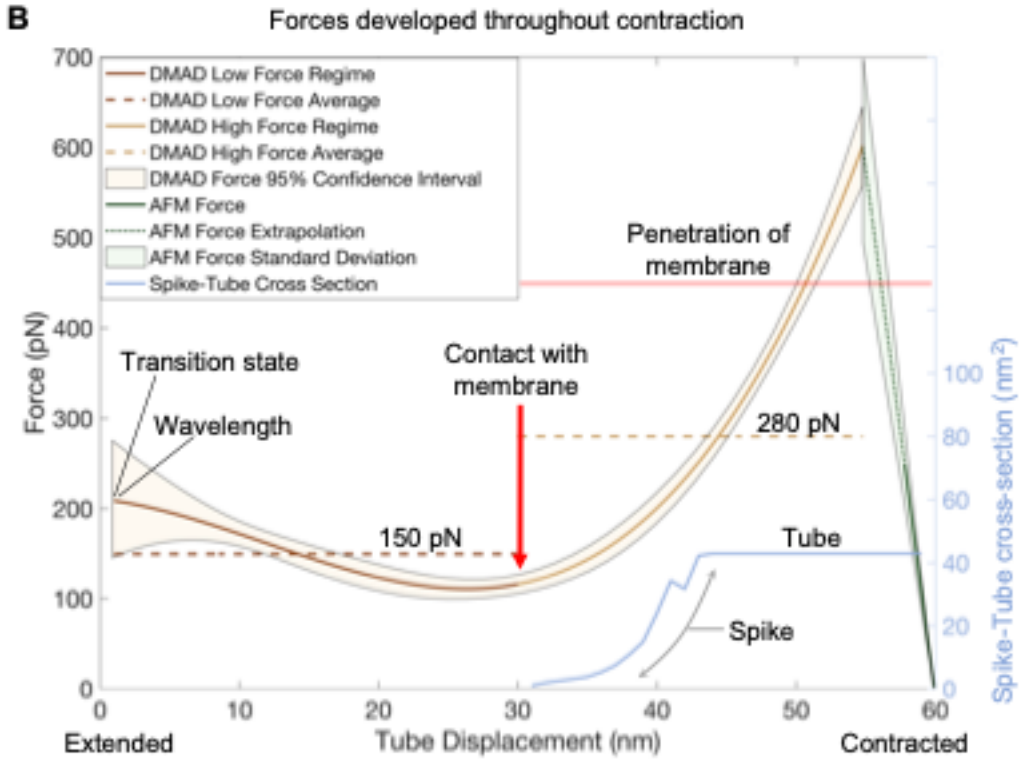

B
Forces developed throughout contraction
DMAD Low Force Regime
DMAD Low Force Average
DMAD High Force Regime
DMAD High Force Average
DMAD Force 95% Confidence Interval
AFM Force
AFM Force Extrapolation
AFM Force Standard Deviation
Spike-Tube Cross Section
Penetration of membrane
Transition state
Wavelength
Contact with membrane
280 pN
150 pN
Tube
Spike
Force (pN)
700
600
500
400
300
200
100
0
Spike-Tube cross-section (nm²)
100
80
60
40
20
0
0
10
20
30
40
50
60
Extended
Tube Displacement (nm)
Contracted

Figure. 3.10. Energies and forces developed by the pyocin sheath-tube complex throughout contraction. (A) Evolution of the DMAD-derived free energy of the pyocin sheath-tube complex throughout contraction with key structural intermediates shown in schematic form (see Fig. 3.1A for color key). The thin lines tracing the free energy curve is the standard deviation of error associated with the extrapolation from the 12-layer fragment (Fig. 3.4E) to the full-length pyocin structure. The conformations of the end states are visualized by negative stain EM of pyocin particles incubated with the *P. aeruginosa* 13s outer membrane fragments. The inset table compares the energetics of the DMAD methodology with biophysical experiment. The energetic calculations presented in this figure are performed only for the sheath-tube complex. The membrane bilayer, the baseplate, and other components are shown for geometric and illustrative purposes only. (B) Evolution of forces developed by the full-length pyocin throughout contraction. The force curve is a negative derivative of a fourth order polynomial fitted to the free energy profile (Fig. 3.5A, 3.10A). The error represents a 95% confidence interval for the DMAD derived forces and is a result of the standard propagation of error from the derivative of a polynomial fit(Cordero and Roth 2005). The extension force of contracted pyocin sheaths is derived from the spring constant measured by HS-AFM (Fig. 9D). The light blue curve shows the evolution of the cross-sectional area of the spike-tube complex as it crosses the membrane plane (fixed at the initial contact point) during sheath contraction. The cross-sectional area was extracted from atomic coordinates by calculating the polygonal area of horizontal slices of the spike-tube complex(Ge et al. 2020).

The DMAD-derived force profile can be divided into three parts: an initial regime (the first ~1 nm of tube motion), a low force regime (~1-30 nm of tube motion) and a final high force regime (~30-60 nm of tube motion) (**Fig. 3.10B**). The scale of tube motion in the first regime is too small for the force to be evaluated by DMAD. The second regime is characterized by a near-constant force of ~150 pN, which is a consequence of the wave-

like propagation of contraction. In this regime, the tube moves towards the membrane but does not yet interact with it, and the force is accordingly low. In the final regime, as the surface area of sheath subunits engaged in interfacial interactions increases dramatically, the force grows and eventually exceeds 500 pN. At the very end of the contraction process, tube displacements sampled by DMAD are too coarse to accurately evaluate the derivative and therefore the forces generated. AFM measurements show that the sheath behaves like a stiff spring, and the force decreases to zero as the sheath reaches the contracted state. The DMAD-derived and extrapolated AFM force curves intersect at ~600 pN, which likely represents the maximum force developed by the pyocin sheath. Notably, the DMAD-derived force of sheath contraction increases in line with the increase in the cross-section of the spike-tube complex as it is driven through the cell membrane (**Fig. 3.10B**). The average pressure across the cross-section of the tube is ~56 atm.

As DMAD is a deterministic method that finds the most probable path in the conformational space by employing a principle of least action-type of approach, the main source of errors in its application to pyocin contraction is in the extrapolation from the 12-layer pyocin fragment to the full-length structure. The relationship between a subunit's contracted fraction and its contribution to the total energy varies slightly depending on the layer. We express this variability via a standard deviation (**Fig. 3.4E**). This error is then propagated according to the law of propagation of uncertainties(Cordero and Roth 2005) to the energetics of the full-length structure where it is very small (**Fig 3.10A**) and to the force profile where it becomes more noticeable (**Fig. 3.10B**).

DISCUSSION

General applicability of the DMAD approach

Here, we introduced a new coarse grained modeling methodology (DMAD) that makes it possible to calculate the free energy profile of a physicochemical reaction for a multimillion atom complex. The DMAD method is innovative in that it can provide physically realistic transition intermediates for systems which are too large for a conventional MD study(Hospital et al. 2015). DMAD can also be used to bootstrap MD simulations by providing a set of physically feasible reaction coordinates, the finding of which often constitutes a problem in and of itself(Best and Hummer 2005).

DMAD can be applied to any system where conformational transitions are dominated by rigid body motions of individual protein subunits such as the contraction of the sheath or the transformation of the baseplate in CISs(Taylor et al. 2016), maturation of virus capsids(Wikoff et al. 2006), and the rotation of the stator units in the bacterial flagellum(Santiveri et al. 2020). In these cases, the first step is to determine the range of allowed motions that preserve the quaternary structure of the system. This allowed space of transitions can then be used to define a range of physical parameters such as the 'springs' employed in the pyocin sheath contraction model (**Fig. 3.1E, 3.2B, 3.3B**). The allowed space of transitions can then be systematically sampled as finely or coarsely as desired, resulting in multiple sets of transition intermediates. Next, the energetics of the transition intermediates can be evaluated by a solvent accessible surface area-based approach (such as PISA) or by other methods if the system is sufficiently small(Hou et al. 2010)·(Hénin and Chipot 2004). Finally, the local region of optimal parameters can be further sampled to find the most energetically favorable transition pathway so that the scale of expected values is determined, and an appropriate experimental methodology is selected.

One of the main sources of uncertainty in the DMAD approach is the validity of the solvent accessible surface area (SASA)-based calculations of free energy. While solvation energies are linearly related to the SASA on the macro scale, micro scale calculations require a thorough analysis of multibody correlations between the solute and

the molecular surface of the protein(Harris and Pettitt 2014),(Harris, Drake, and Pettitt 2014). SASA-based methods work particularly well when the interfaces being analyzed are large(Rajamani, Truskett, and Garde 2005) and rigid(David S. Cerutti, Lynn F. Ten Eyck, and J. Andrew McCammon† 2004),(Gohlke and Case 2004), as is the case for the pyocin sheath-tube complex. To evaluate the implicit error in approximating solvation energies using SASA methods, we applied PISA to a well-studied system, the association of the proline-rich peptide p41 with the SH3 domain of a tyrosine kinase (PDB code: 1BBZ(Pisabarro, Serrano, and Wilmanns 1998)). The results showed an association free energy of −8.2 kcal/mol, which is comparable to MD-derived values in the range of −7.7 to −7.8 kcal/mol(Chipot 2014) and in agreement with solution biophysics experimental values in the range of −7.7 to −8.0 kcal/mol(Pisabarro, Serrano, and Wilmanns 1998),(Palencia et al. 2004),(and and Serrano* 1996). Thus, PISA can provide accurate association energy estimates not only for large interfaces as in the pyocin system but also for smaller complexes.

Additional observations supporting DMAD-derived energy and force profiles

The activation energy of the sheath-tube system is ~160 kcal/mol, which manifests in high stability at room temperature. *In vivo*, contraction is triggered by the binding of multiple tail fibers to cell-surface receptors. The immobilization of the tail fibers reduces their conformational entropy, which likely results in a positive ΔG contribution and promotes a conformational change of the baseplate. The latter brings the sheath to the transition state, which then quickly contracts. Accordingly, particles with multiple fibers bound to the cell surface and extended sheaths are rarely imaged(Hu et al. 2015), likely due to their transient nature and very short lifetime.

This line of arguments is further supported by the DMAD-derived force profile of the contraction reaction and by electron microscopy imaging of pyocins (**Fig. 3.10A**) and phages T4 and P130-50. Attachment of tail fibers to the cell surface positions the baseplate ~30 nm away from the lipid membrane (**Fig. 3.10A**). Therefore, the tube does not come in contact with the cell membrane until roughly halfway through the contraction process and thus requires little force to traverse the first ~30 nm. Accordingly, in the DMAD-derived force profile of the contraction reaction the initial force is low (**Fig. 3.10B**). However, upon the tube touching the membrane, the force increases as the cross-section of the spike-tube complex, which is driven into and across the membrane, increases (**Fig. 3.10B**).

This study establishes a relationship between the functional requirements, physical parameters, and the structure of a CIS. Due to the wavelike nature of the propagation of contraction, CISs that are longer than the contraction wavelength have similar activation energies and generate similar forces for membrane puncture by the spike-tube complex. This provides a rationale for the length of the pyocin particle: it is built long enough to develop a force necessary for membrane penetration but no larger, as beyond this point additional sheath subunits do not lead to greater forces. This result is directly applicable to the function of T6SS and explains how T6SS organelles, which vary greatly in length (both within a single organism and between organisms) and can be as long as 5 μm (~37 times longer than the pyocin)(Vettiger et al. 2017)·(Stietz et al. 2020), are activated by a baseplate that resembles that of pyocin or T4 phage(Taylor et al. 2016)·(Y.-J. Park et al. 2018)·(Nazarov et al. 2018)·(Cherrak et al. 2018).

MATERIALS AND METHODS

The Domain Motion in Atomic Detail procedure

The contraction process was initiated by a small perturbation of baseplate-proximal sheath subunits (which constitutes ~1/1000th of the total motion of the subunit as it travels from the extended to the contracted state). This perturbation then propagated iteratively to all other subunits according to the algorithm described in the section "**The equations of motion of sheath subunits (propagation of contraction) in pseudocode**" below. The algorithm was implemented for a total of 100 (k_1, k_2, D) parameter sets (**Fig. 3.3B**).

In the laboratory reference frame, each subunit moved along the conical cylindric geodesic connecting the COMs of the subunit in the extended and contracted conformations (**Fig. 3.2A**). This motion was modulated by a velocity-like term that corresponded to the fraction of the geodesic an individual subunit traversed during one iteration. Several analytic expressions of the velocity curve that had vanishing velocities at the two terminal points have been tested. As none of these forms significantly lowered the activation energy or made the overall profile smoother, a parabolic form was used. The maximum velocity comprised 5% of the geodesic path (v_{max} = 0.05).

To account for small differences in the fold of the sheath subunit in the extended and contracted states (**Fig. 3.1B**), a morphing trajectory containing 99 intermediates of the sheath subunit spanning its extended (conformation 0) and contracted (conformation 100) states was generated with the help of UCSF Chimera(Pettersen et al. 2004). Accordingly, when the contracted fraction of a subunit was N % during a simulation, the coordinates of the N^{th} morphing intermediate were used as the starting conformation of the sheath fold.

As any perturbation (no matter how small) to the position of a subunit in the sheath structure resulted in a broken connectivity of the sheath's mesh, it had to be rebuilt. This task was executed with the help of a semiautomated procedure using Coot(Emsley, Cowtan, and IUCr 2004) and other programs from the CCP4 package(Winn et al. 2011). The geometry was improved, and possible clashes were removed with the help of cartesian dynamics as implemented in the Phenix software package(Adams et al. 2011). The Phenix

protocol is a fast, crude version of molecular dynamics that lacks attractive forces. Dynamics runs had to be kept short and not run at high temperatures. In practice, 400 timesteps of dynamics at 150K was best for regularizing geometry and resulted in an RMSD of < 1 Å for the complex. The resulting structures were of good quality (**Table 3.1**)(Chen et al. 2009).

Each contraction trajectory contained ~100 intermediate structures. The tube and sheath are linked to each other via a capping protein. Thus, the motion of the tube was derived from the trajectory of the uppermost sheath subunits. The free energy of the solvent-exposed and buried surfaces was calculated with PISA(Krissinel and Henrick 2007).

The equations of motion of sheath subunits (propagation of contraction) in pseudocode

Let k_1, k_2, D be the intra-strand coupling constant, the inter-strand coupling constant and the drag parameter, respectively.

Let r_{ext}^i, r_{cnt}^i be the radius of the center of mass (COM) of the i^{th} subunit ($i \in \{1,2,\dots,12\}$) in the extended and contracted states, respectively.

Let θ^i be the angle between the COMs of the i^{th} subunit in the extended and contracted states in the x-y plane (such that the bottom sheath layer COM is centered on the origin and subsequent layers COMs lie on the + z-axis).

Let θ_0^i be the angle which describes the location of the COM of the i^{th} subunit in the extended state in the x-y plane.

Let z_{ext}^i, z_{cont}^i be the height of the COM (projection onto the z-axis) of the i^{th} subunit in the extended state and contracted states, respectively.

Let b_j^i be the scalar distance between the COMs of the i^{th} subunit and the subunit connected to it by the inter-stand linker at the j^{th} iteration of the algorithm, such that b_0^i is the distance between COMs in the extended state.

Let s_j^i be the scalar difference between b_j^i and b_0^i, such that $s_j^i = \begin{cases} b_j^i - b_0^i, & b_j^i > b_0^i \\ 0, & b_j^i \leq b_0^i. \end{cases}$

Let v_j^i be the scalar 'subunit velocity' of the i^{th} subunit at the j^{th} iteration of the algorithm where the maximum velocity is v_{max}.

Let d_j^i be the scalar displacement of the COM of the i^{th} subunit between iterations j and $j-1$, such that $d_1^i = 0$.

Let $\lambda_j^i \in [0,1]$ be the 'contracted fraction' of the i^{th} subunit at the j^{th} iteration, where $\lambda = 0$ represents the extended state and $\lambda = 1$ represents the contracted state.

Let $(\omega^i, \varphi^i, \kappa^i)$ be the set of polar angles which describe the rotation of i^{th} subunit about its COM from the extended to the contracted state. (ω^i, φ^i) describe the axis of rotation in the reference frame of the COM and thus are fixed throughout the trajectory. Therefore, the orientation of the i^{th} subunit at the j^{th} iteration of the algorithm can be described by $\left(\omega^i, \varphi^i, \kappa_j^i\right)$, where $\kappa_{j=0}^i = 0$ in the extended state.

Let $r_j^i, z_j^i, \vartheta_j^i$ be the radius, height and azimuth of the COM of the i^{th} subunit at the j^{th} iteration of the algorithm, respectively.

Let P be the initial perturbation factor.

As the algorithm commences, the extended structure is perturbed slightly such that:

$$\lambda_1^i = P(k_1)^{i-1}$$

$$\left(\omega^i, \varphi^i, \kappa_1^i\right) = \left(\omega^i, \varphi^i, \lambda_1^i \kappa^i\right)$$

$$r_1^i = r_{ext}^i + \lambda_1^i (r_{cnt}^i - r_{ext}^i)$$

$$z_1^i = z_{ext}^i + \lambda_1^i (z_{cnt}^i - z_{ext}^i)$$

$$\vartheta_1^i = \theta_0^i + \lambda_1^i \theta^i$$

This defines the initial perturbed state. Now contraction progresses as follows until the fully contracted state is reached:

Set $j = 2$

While $r_j^{12} < r_{cnt}^{12}$:

For $i \in \{1,2,\dots,12\}$:

$$v_j^i = v_{max}\left(-\left(\lambda_{j-1}^i\right)^2 + \lambda_{j-1}^i\right)$$

end_For

$$\lambda_j^1 = \min\left(1, \lambda_{j-1}^1 + \frac{v_j^1 + k_1 v_j^2}{1+k_1} - k_2\left(s_{j-1}^2\right)^2 - Dd_{j-1}^1\right)$$

For $i \in \{2,\dots,11\}$:

$$\lambda_j^i = \min\left(1, \lambda_{j-1}^i + \frac{v_j^i + k_1 v_j^{i-1} + k_1 v_j^{i+1}}{1+2k_1} + k_2\left(\left(s_{j-1}^i\right)^2 - \left(s_{j-1}^{i+1}\right)^2\right) - Dd_{j-1}^i\right)$$

end_For

$$\lambda_j^{12} = \min\left(1, \lambda_{j-1}^{12} + \frac{v_j^{12} + k_1 v_j^{11}}{1+k_1} + k_2\left(s_{j-1}^{12}\right)^2 - Dd_{j-1}^{12}\right)$$

For $i \in \{1,2,\dots,12\}$:

$$\left(\omega^i, \varphi^i, \kappa_j^i\right) = \left(\omega^i, \varphi^i, \lambda_j^i \kappa^i\right)$$

$$r_j^i = r_{ext}^i + \lambda_j^i (r_{cnt}^i - r_{ext}^i)$$

$$z_j^i = z_{ext}^i + \lambda_j^i (z_{cnt}^i - z_{ext}^i)$$

$$\vartheta_j^i = \theta_0^i + \lambda_j^i \theta^i$$

$$b_j^i = \sqrt{\left(r_j^i \cos(\vartheta_j^i) - r_j^{i-1} \cos(\vartheta_j^{i-1} + \pi/3)\right)^2 + \left(r_j^i \sin(\vartheta_j^i) - r_j^{i-1} \sin(\vartheta_j^{i-1} + \pi/3)\right)^2 + \left(z_j^i - z_j^{i-1}\right)^2}$$

$$d_j^i = \sqrt{\left(r_j^i \cos(\vartheta_j^i) - r_{j-1}^i \cos(\vartheta_{j-1}^i)\right)^2 + \left(r_j^i \sin(\vartheta_j^i) - r_{j-1}^i \sin(\vartheta_{j-1}^i)\right)^2 + \left(z_j^i - z_{j-1}^i\right)^2}$$

end_For

$j = j + 1$

end_While

Extrapolation of the 12-layer fragment contraction pathway to the full length sheath

The method of extrapolating results of the 12-layer simulations to generate full-length models of sheath contraction consists of two steps:

1. Use the optimal (k_1, k_2, D) parameters found in the global search for the 12-layer fragment to calculate the position of every subunit in the full-length structure for each contraction intermediate.
2. Analyze the distance between sheath subunit COMs in these intermediates and reject pathways with values greater than 55 Å (separations of greater than 55 Å resulted in the disintegration of the handshake domain thus breaking the native sheath mesh, **Fig. 3.4D**).

Additionally, for step 1, the following properties of the system had to be taken into account.

The terms $k_2\left(\left(s_{j-1}^i\right)^2 - \left(s_{j-1}^{i+1}\right)^2\right)$ and Dd_{j-1}^i in the contraction fraction parameter formula

$$\lambda_j^i = \min\left(1, \lambda_{j-1}^i + \frac{v_j^i + k_1 v_j^{i-1} + k_1 v_j^{i+1}}{1 + 2k_1} + k_2\left(\left(s_{j-1}^i\right)^2 - \left(s_{j-1}^{i+1}\right)^2\right) - Dd_{j-1}^i\right)$$

vary with the length of the sheath due to their dependance on the path of the subunit along its geodesic trajectory. For this reason, the extrapolation from the 12-layer fragment to the full-length sheath required rescaling of the k_2 and D parameters as follows (k_1 was not rescaled as v does not depend on the path):

- k_2 was rescaled by a factor of X^2,
- D was rescaled by a factor of X,

where $X = 1/2.44$ is the ratio of the average distance traveled by all subunits on their conical geodesic trajectories in the 12-layer simulation relative to the distance traveled in the full-length simulation.

Purification of pyocins for biophysical experiments

Pyocin particles were produced in *Escherichia coli* using a pETcoco-1-based plasmid that contained the entire cluster of R2 pyocin genes (*Pseudomonas aeruginosa* PAO1 genes *pa0610-pa632*) including the regulatory genes *prtN* (*pa0610*) and *prtR* (*pa611*) and the lysis cassette (*pa0629-pa0632*). The plasmid was created by Dean Scholl and colleagues (plasmid pSW192, AvidBiotics Corp.) and its design and construction are described elsewhere(Ritchie et al. 2011).

In a medium free from arabinose, the cells carry one or two copies of the pSW192 plasmid, and the cluster is completely inhibited. The addition of arabinose increases the copy number of the pSW192 plasmid, and this activates the entire operon, including lysis genes. Thus, production of pyocin particles can be triggered by the addition of arabinose. Newly assembled pyocin particles are released from the cells by lysis. These particles were morphologically indistinguishable from pyocins produced by PAO1 treated by mitomycin C. The killing activity was quantified by spot killing assay on a sensitive *P. aeruginosa* 13s strain. EM showed that the 'recombinant' *E. coli*-produced pyocins were all extended, unlike those purified from *Pseudomonas* that contain a fraction of contracted particles,

possibly because *E. coli*-produced pyocins are never exposed to *Pseudomonas* cell fragments during the purification procedure.

Eight liters of *E. coli* BL21ΔAra(Ritchie et al. 2011) freshly transformed with pSW192 were grown in Lennox LB medium (Fisher Scientific, Cat # BP1427-500) supplemented with chloramphenicol at 11 μg/ml (Fisher Scientific, Cat # BP904-100) in eight 4 L Erlenmeyer flasks at 37°C and 240 rpm to an optical density of 1.0 at 600 nm. Five ml of 20% arabinose (Fisher Scientific, Cat # AAA1192118) and 5 ml of 80% glycerol (Fisher Scientific, Cat # G33-500) were added to each flask, the temperature was decreased to 30°C, and the cells were incubated at this temperature overnight. Debris and residual bacteria were removed from the lysate by centrifugation at 15,000g for 30 minutes in a F9-6x-1000 rotor (Fisher Scientific, Cat # 09-606-1075). The supernatant was then filtered through a 1V paper filter (GE Healthcare, Cat #1201270) and 4 mg of DNAse I and 4 mg of RNAse A (MilliporeSigma, Cat # 69182 and Cat # 556746, respectively) were added to it. The clarified lysate was supplemented with 240 g of NaCl (dry powder) and 800 g of PEG 8,000 (dry flakes) (Fisher Scientific, Cat # 18-606-422 and Cat # BP233-1, respectively). The chemicals were dissolved by stirring and then the mixture was incubated at 4°C overnight to ensure complete precipitation of pyocins. Pyocins were pelleted at 15,000g for 15 minutes in a F9-6x1000 rotor. The pellets were resuspended in 80 ml of SM buffer (8 mM $MgCl_2$, 100 mM NaCl, 50 mM Tris-HCl pH 7.5) with DNAse I and RNAse A at 1 μg/ml each. The cell membrane and other associated impurities were removed by the addition of 80 ml of chloroform (Fisher Scientific, Cat # C298-500) and centrifugation in 50 ml Falcon tubes at 15,000g for 15 minutes. The pyocin-containing aqueous phase was collected and pyocins were pelleted at 100,000g for 2 hours in a T29-8x50 rotor (Fisher Scientific, Cat # 75-003-009). The pellets were dissolved in 4 ml of SM buffer on an orbital shaker at 100 rpm overnight at 4°C. Undissolved material was removed by centrifugation at 15,000 g for 5 min in a microcentrifuge at room temperature. The soluble fraction was

divided into halves and each part was then loaded onto a premade step gradient of 10%, 20%, 30%, 40% (2 ml each) and 60% sucrose (3 ml bottom cushion) in SM buffer. The tubes were then centrifuged at 100,000g for 1 hour in a SW40ti rotor (Beckman Coulter). Bands containing pyocin particles were located in the upper part of the gradient and were visible by eye. These bands were extracted from both tubes, combined, and dialyzed against two changes of 0.1x SM buffer. The concentration of the dialyzed sample was determined based on the adsorption at 280 nm and brought to 5 mg/ml.

To obtain a sample of fully contracted pyocins, 15 µl of 3M Glycine-HCl pH 2.5 was added to 3 ml of purified extended pyocin sample at 5 mg/ml and glycine was dialyzed out using two changes of the same 0.1x SM buffer.

Design of sheath mutants

Additional residues were introduced into the inter-strand linker of the PA0622 sheath protein between Gly23 and Ser24 with the help of a modified allelic exchange procedure(Blomfield et al. 1991). For each mutant, two overlapping ~1 kb-long fragments carrying the mutation were amplified by appropriate pairs of primers (**Table 3.2**) and pSW192 as the template. The flanking primers AE21M1F and AE21M2R were common for all four linker mutants. The exchange donor vectors were assembled by NEBuilder reaction (New England Biolabs, Ipswich, MA) on the backbone of the pWM91 plasmid(Metcalf et al. 1996) in which the ampicillin resistance gene was replaced with the kanamycin resistance gene. The recipient plasmid pSW192 was maintained in RecA+ *E. coli* 4s strain(Prokhorov et al. 2017). The donor vectors were transformed into the MFDpyr *E. coli* strain(Ferrières et al. 2010) and conjugation between the donor and the acceptor strains was performed on LB agar plates overnight at 37°C. Selection for recombination products was done on LB agar plates supplemented with kanamycin at 50 µg/ml.

Counterselection for excision products, pSW192 and the mutation carrying plasmids, was done on agar plates with 1% trypton, 0.5% yeast extract and 5% sucrose (MilliporeSigma, Burlington, MA) overnight at room temperature. Colony screening was done by PCR. The presence of mutations was confirmed by Sanger sequencing.

Killing assay for pyocin sheath mutants

The WT and mutant pyocins were expressed in the *E. coli* strain BL21Δ*ara*Δ*fhuA*Δ*ompF* using the pSW192 plasmid and its sheath mutant derivatives as described above. The cells were allowed to lyse, and the particles were purified as follows. Cell debris were removed by centrifugation at 15,000g for 15 minutes using the F9-6x-1000 rotor (Fisher Scientific, Cat # 09-606-1075). The pyocins were pelleted by centrifugation at 100,000g for 6 hours using the SW-28 rotor (Beckman Coulter, Cat # 342204). The pellets were dissolved in SM buffer overnight and both centrifugation steps (medium and high speed) were repeated. The samples were kept at 4℃ throughout the purification. The samples were normalized for their protein concentration according to their UV absorbance at 280 nm. The killing activity of pyocins contained in the samples was evaluated by a double agar overlay spot assay on a lawn of *P. aeruginosa* 13s cells (R2 pyocin-sensitive strain)(Scholl and Martin 2008). Double agar overlay plates with bacterial lawns were prepared using a standard procedure(Kropinski et al. 2009). Cells transformed with an empty pETcoco-1 vector (Sigma-Aldrich, Cat # 71129) served as a negative control. Since these cells predictably did not lyse after induction, they were disrupted by ultrasound sonication and subjected to the same purification steps as the lysates containing pyocins. The sample content was verified with SDS-PAGE.

Preparation of *P. aeruginosa* 13s outer membrane fraction

The outer membrane fraction of *P. aeruginosa 13s* was purified by differential solubilization(Hobb et al. 2009) with the help of N-lauroylsarcosine sodium salt (Sigma-Aldrich). The cell culture was grown in 50 ml of LB media up to late log phage (OD_{600nm} = 1.0) at 37°C and vigorous shaking. The cells were pelleted by centrifugation at 5,000 g for 5 min. The cells were resuspended in 10 ml of 10 mM Tris-HCl pH 8.0 supplemented with DNAse I and RNAse A at 1 µg/ml each and disrupted by sonication. The total membrane fraction was pelleted by centrifugation at 100,000 g for 30 min, resuspended in the same buffer, and pelleted again. The pellet was resuspended in 1% N-lauroylsarcosine in 10 mM Tris-HCl pH 8.0 and incubated at 37°C for 30 min with moderate shaking to selectively dissolve the inner membrane. The sample was centrifugated at 100,000 g for 30 min, resuspended in 10 mM Tris-HCl pH 8.0 and pelleted again, then lyophilized and weighed, taking into account the presence of the Tris buffer in the sample.

Electron microscopy

Twenty µl of 10 mg/ml *P. aeruginosa* 13S outer membrane fragments (in 0.1x SM buffer) were incubated with 20 µl of 1 mg/ml purified extended pyocin particles (in 0.1x SM buffer) for 20 minutes. After incubation, 5 µl samples were aliquoted onto plasma cleaned Electron Microscopy Sciences CF200-CU grids, excess sample was blotted, then were stained twice with 10 µl of 7 % uranyl acetate in 50% EtOH solution. The grids were mounted into a JEOL 2100 microscope and imaged at a magnification of 40,000.

Differential scanning calorimetry and isothermal titration calorimetry

The enthalpy of heat induced pyocin contraction was measured with MicroCal PEAQ-DSC (Malvern Instruments) and NanoDSC (TA Instruments) microcalorimeters. The sample concentration was 4.5 - 5 mg/ml in 0.1x SM buffer (0.8 mM $MgCl_2$, 10 mM NaCl,

5 mM Tris-HCl pH 7.5). The scanned temperature range was from 20 to 130°C. The heating rates for the MicroCal PEAQ-DSC and NanoDSC instruments were 200°C/hr and 120°C/hr, respectively. Calculations of the molar concentration of sheath subunits took into account that the sheath constituted 55.16% of the total mass of the pyocin particle and contains 168 subunits. For instance, in a pyocin sample with a concentration of 5 mg/ml, the molar concentration of sheath subunits is 77 μM.

A MicroCal PEAQ-ITC microcalorimeter (Malvern Instruments) was used to determine the enthalpy of pyocin contraction induced by exposure to low pH buffer. The ITC experiment was conducted in an unconventional mode where the pyocin sample was titrated into the cell containing low-pH buffer. The pyocin samples (both, the extended and contracted particle) were dialyzed into 0.05x SM buffer (0.08 mM $MgCl_2$, 1 mM NaCl, 0.5 mM Tris-HCl pH 7.5). The pyocin concentration in the injection syringe was 5 mg/ml. The cuvette of the calorimeter contained 50 mM Glycine-HCl pH 2.5. The experiments consisted of 7 injections, with 60 seconds in-between. The volume of each injection was 5 μl (35 μl of a sample per run in total). The temperature was held constant at 20 °C.

Circular dichroism and contraction kinetic measurements

CD spectra and time course measurements were recorded using a JASCO J-815 CD spectrometer equipped with a Peltier cell. The concentration used for both extended and contracted samples was 0.1 mg/ml in 10 mM NaCl 10 mM phosphate buffer at pH 7.0. Prior to all measurements, the condition and conformation of the specimens (extended or contracted) was confirmed by negative stain EM and killing assay on sensitive *P. aeruginosa 13s*. Exposure of an extended pyocin sample to heat for a prolonged time changed its spectrum to a contracted-like one. EM showed that in such samples all sheaths were contracted, and some material was aggregated (some baseplates and tubes were stuck

together). Exposure of a contracted pyocin specimen to heat did not change its spectrum beyond the noise level, and the sample displayed similar aggregation.

A wavelength of 203 nm was chosen for time course measurements to maximize the difference between extended and contracted specimen spectra (**Fig. 3.6C**). The WT pyocin and its mutants displayed measurable contraction kinetics in a 60-74°C range, depending on the buffer composition.

All measurements were performed with at least three technical replicates of at least two biological replicates (three replicates were used for the WT). Each time course measurement (every mutant and every *de novo* purification) required a contracted specimen as a control. These contracted samples were measured for at least two different temperatures and in multiple technical replicates. The contracted specimen curves were averaged and subtracted from the averaged extended specimen curves to obtain the contraction time course for each of the reactions. In all datasets, the first two minutes were trimmed because they corresponded to sample heating, mixing, and equilibration. Data analysis was performed in MatLab CFTool (MathWorks).

High-speed atomic force microscopy

Pyocins were first characterized from side views when physiosorbed on mica (**Fig. 3.9A**). The head-on pyocin sheaths (**Fig. 3.9B**) were prepared by modifying the mica surface with 0.01% poly-lysine for 3 min, then depositing contracted sheaths (from 10 times diluted sample) for 10 min. All images in this study were acquired using a HS-AFM(Ando et al. 2001) (SS-NEX, RIBM, Japan) operated in amplitude modulation mode using optimized scan and feedback parameters at room temperature. Short (8μm) cantilevers (NanoWorld, Switzerland) with nominal spring constant k_c= 0.15 N/m, resonance frequency of ~0.6 MHz, and a quality factor Q_c~1.5 in buffer (50 mM Tris-HCl,

pH 7.5, 100 mM NaCl, 8 mM $MgSO_4$) were used. The energy delivered by a tip-sample interaction can be estimated by $\Delta E = (1-\alpha) * k_c(A_o^2 - A_s^2)/(2Q_c)$; with α = 0.5 being the ratio of the amplitude reduction caused by the cantilever resonance frequency shift over the total amplitude reduction. To avoid fall over or displacement of the head-on physisorbed pyocin sheaths during AFM imaging, the tip-sample interaction was minimized by using a free amplitude A_o = 1 nm and a set-point amplitude $A_s \geq 0.9$, respectively. Under such conditions the energy delivered is ~1.2 k_BT, while most of the input energy will be dissipated into the fluid between taps. Both ends of the contracted pyocin sheath – the open baseplate and closed neck end – could be easily identified during imaging. Force spectroscopy measurements were performed following the targeting of the baseplate end of single pyocin sheaths only in HS-AFM imaging mode. Upon centering on the opening in a single pyocin sheath, the setup was switched into force measurement mode and approach-retract cycles were acquired at 150 nm/s velocity.

Molecular graphics

Fig. 2.1B, 2.1C, 2.1D, 2.1E, 2.1F, 2.4C were created using UCSF Chimera(Pettersen et al. 2004). Fig. 4.4B and 4.5C were created with the help of UCSF ChimeraX(Goddard et al. 2018).

ACKNOWLEDGMENTS

General: We thank Dr. Sergey Budko (Vanderbilt University) for his insight and suggestions. We thank TACC (Texas Advanced Computing Center) at the University of Texas at Austin for their high performance computing (HPC) resources that were helpful in achieving the results presented in this paper. We thank Dean Scholl (Pylum Biosciences,

Inc.) for sharing plasmids and protocols, and Luis Marcelo F. Holthauzen (UTMB) for his help with CD and ITC experiments.

Funding: The work was supported by the UTMB Department of Biochemistry and Molecular Biology and by the Sealy Center for Structural Biology and Molecular Biophysics. AF, NSP, and PGL are supported by the NIGMS grant R01 GM139034. SS is supported by the NIH NCCIH – DP1AT010874 grant. BMP thanks the Robert A. Welch Foundation (H-0013) for partial support.

Author contributions

The modeling and AFM parts of the study were conceived by **PGL** with **AF** and **SS**, respectively. The solution biophysics experiments were conceived by **PGL** and **NSP. AF** and **PGL** developed the DMAD methodology with advice from **BMP**, and **AF** implemented it into a computer code. **NSP** developed and implemented the pyocin purification protocol, all solution biophysics experiments for the WT pyocin and all the mutants, which **NSP** also created, and performed the pyocin sheath mutant killing assay. **FJ** performed AFM measurements, and **FJ** and **SS** processed the AFM data. **PGL, AF**, **NSP**, **FJ**, and **SS** wrote the first draft of the paper, which was then read, extensively edited, and approved by all authors.

Competing interests

The authors declare no competing interests.

Data and materials availability

The extended and contracted pyocin sheath-tube complex structures are available at the Protein Data Bank under the accession numbers 3J9Q (a four-layer extended sheath-tube fragment) and 3J9R (a six-layer contracted sheath fragment)(Ge et al. 2015),(Berman et al. 2000).

Chapter 4 Structural Insights into Late-Stage Bacteriophage Contraction

Portions of the following chapter were reproduced with permission from:
Fraser, A., Prokhorov, N. S., Miller, J. M., Knyazhanskaya, E. S., & Leiman, P. G. (2021). Identification of Low Population States in Cryo-EM Using Deep Learning. *bioRxiv*

ABSTRACT

Bacteriophages employ a contractile tail to puncture host cell membranes and subsequently passage their replicative materials. The membrane puncturing process is powered by a sheath which transitions from a high energy pre-contraction to a low energy post-contraction state. While the structures of these end states have been characterized for bacteriophages and several related systems such as the bacterial type VI secretion system and tailocins, little experimental information is available about the structure of intermediates. Here, we characterize the sheath structure of a stalled contraction intermediate of bacteriophage A511 and show that the mechanism of contraction is governed by both local linear and global non-linear properties of the sheath. Furthermore, we present the first atomic model of the bacteriophage sheath during infection.

INTRODUCTION

Most known bacteriophages carry a complex multicomponent tail organelle that functions to inject phage DNA and proteins into the cytoplasm of their bacterial host cells. The tail and special proteins emanating from it must first recognize the surface of a cell, attach to it, and then create a channel spanning it(Taylor, Raaij, and Leiman 2018),(Patz et al. 2019),(Marek Basler 2015),(Cascales and Cambillau 2012). The most complicated component involved in this process, the baseplate, is responsible for the recognition and

subsequent attachment of the bacteriophage to the host cell(Bönemann, Pietrosiuk, and Mogk 2010). The second component, the sheath-tube complex(Ge et al. 2015), functions to drive a spike-shaped protein through the membrane, creating a channel through which capsid-packaged DNA and proteins can traverse. Throughout this process, both the baseplate and sheath-tube complexes demonstrate massive conformational changes spanning hundreds of nanometers on a microsecond timescale(Ge et al. 2015)·(Taylor et al. 2016).

Both components of the sheath-tube complex – the rigid tube and the contractile sheath – have matching sixfold helical symmetries in the initial, metastable, high-energy pre-attachment or 'extended' state(Ge et al. 2015). Following the recognition and attachment of the host cell surface, the 'bottom' of the sheath-tube complex and the baseplate are positioned roughly 30nm from the lipid membrane(Hu et al. 2015)·(J. Liu et al. 2011)·(Fraser et al. 2021). The baseplate changes its conformation and triggers rearrangement of the baseplate proximal sheath subunits resulting in sheath contraction. The latter is converted into a screw-like motion of the tube because the tube and sheath are fixed at the baseplate-distal end. The membrane-attacking tip of the tube is equipped with a spike protein that is thought to aid in membrane piercing(Browning et al. 2012)·(Shneider et al. 2013). Contraction will proceed until the sheath reaches the final low-energy 'contracted' state. At some point, the spike protein will separate from the tube, creating a channel connecting the interior of the bacteriophage capsid to the host cytoplasm. Finally, replicative materials will travel through the tube and across the host membrane, enabling host takeover and the creation of progeny bacteriophages.

The atomic structures of the sheath-tube complex in both the extended and contracted states have been solved for various bacteriophage-like systems, such as the R-type pyocin(Ge et al. 2015), T6SS(Kudryashev et al. 2015)·(Wang et al. 2017) and PVC(Jiang et al. 2019). Both the sheath and tube megastructures are made up of individual subunits assembled in helical strands. In all cases, the structure of the tube is extremely

similar(Ge et al. 2015; Kudryashev et al. 2015; Wang et al. 2017; Jiang et al. 2019). The individual sheath subunit is comprised of at least two domains, i) a "handshake" domain(Ge et al. 2015; Kudryashev et al. 2015; Wang et al. 2017; Jiang et al. 2019) which mediates a two-dimensional mesh of 'intra-strand' and 'inter-strand' linkers and ii) a protrusion domain, which is much larger. Such simplistic architecture is characteristic of R-type pyocins and related bacteriophages. In most systems cases, sheath proteins carry additional domains that are 'inserted' into the protrusion domain and extend roughly radially from the sheath (e.g. phage T4 contains two additional domains)(Leiman and Shneider 2012).

While both the extended and contracted states of the sheath-tube complex have been characterized for several systems(Ge et al. 2015; Kudryashev et al. 2015; Wang et al. 2017; Jiang et al. 2019), little experimental information exists about structural intermediates of this process. Early studies using negative stain electron microscopy captured low-resolution images of bacteriophages T4(Moody 1973), G(Donelli, Guglielmi, and Paoletti 1972) and PBS-1(Eiserling 1967) in putative intermediate states. These low resolution images were then used to develop geometric models for how the sheath could deform during contraction given the twist vs. rise of the helical strands(Moody 1973),(Caspar 1980). Putative intermediate structures of the bacteriophage A511 have also been imaged using cryo-electron tomography (CryoET)(Guerrero-Ferreira et al. 2019). Across all phages imaged, a "domino-like" or "wave-like" contraction scheme, whereby intermediate structures are "more contracted" at the bottom and "less contracted" at the top, were observed. Recently, a combined computational and experimental approach (called the Domain Motion in Atomic Detail, DMAD) provided an energetic rationale for the wave-like contraction(Fraser et al. 2021). Furthermore, this study made predictions for the shape of the "contraction wave" in the case of R-type pyocin. Ultimately, however, three-dimensional experimental structures of tail contraction intermediates have remained elusive.

Here, we present the structure of the bacteriophage A511 sheath in the helically symmetric extended and contracted states and the structure of a stalled contraction intermediate. We compare the observed structural transitions to those predicted by the DMAD methodology.

RESULTS

Overall structure

All conformations of the sheath are axially sixfold symmetric and are made up 6 helical strands comprised of individual sheath subunits (gp93) (Fig 4.1, Fig 4.2). In the extended and contracted states, the helical strands exhibit helical twists and rises of (21°, 39.7 Å) and (31°,19.4 Å), respectively (Fig 4.1c 4.1d). In all states, the sheath is comprised of sixfold symmetric 'discs' which are stacked on top of each other. Upon contraction, the width of the sheath increases, with the distance from the disc's center of mass (COM) to a sheath subunit's COM (i.e., the radius of the disc) increasing from 65 Å to 96 Å (Fig 4.1a 4.1b).

Structure of the sheath subunit

The individual sheath subunit is comprised of four domains: i) the handshake domain (residues 2-31, 458-562) ii) the main globular domain (residues 32-94 & 289-457), iii) the preprotrusion domain (residues 95-150 & 251-288) and iv) the protrusion domain (residues 151-250). Furthermore, the handshake domain contains N/C terminal arms (residues 2-31 & 544-562), respectively (Fig 4.1e)

In all states, the handshake domain is comprised of interactions between four separate sheath subunits (Fig 4.1f). In this arrangement, the "central" subunit (residues 533-540) forms a beta-sheet interaction with the C-terminal arm (residues 548-555) of the

sheath subunit originating within the same strand and one disc up. This arm also engages in beta-sheet interaction with the N-terminal arm (residues 15-22) of the sheath subunit originating from the clockwise adjacent subunit in the same disc. Furthermore, there is an interaction between the “central” subunit (residues 526-530) and the C-terminal arm (residues 559-562) of the sheath subunit originating within the same strand and two discs up. Altogether these components form hydrogen bonding networks and hydrophobic

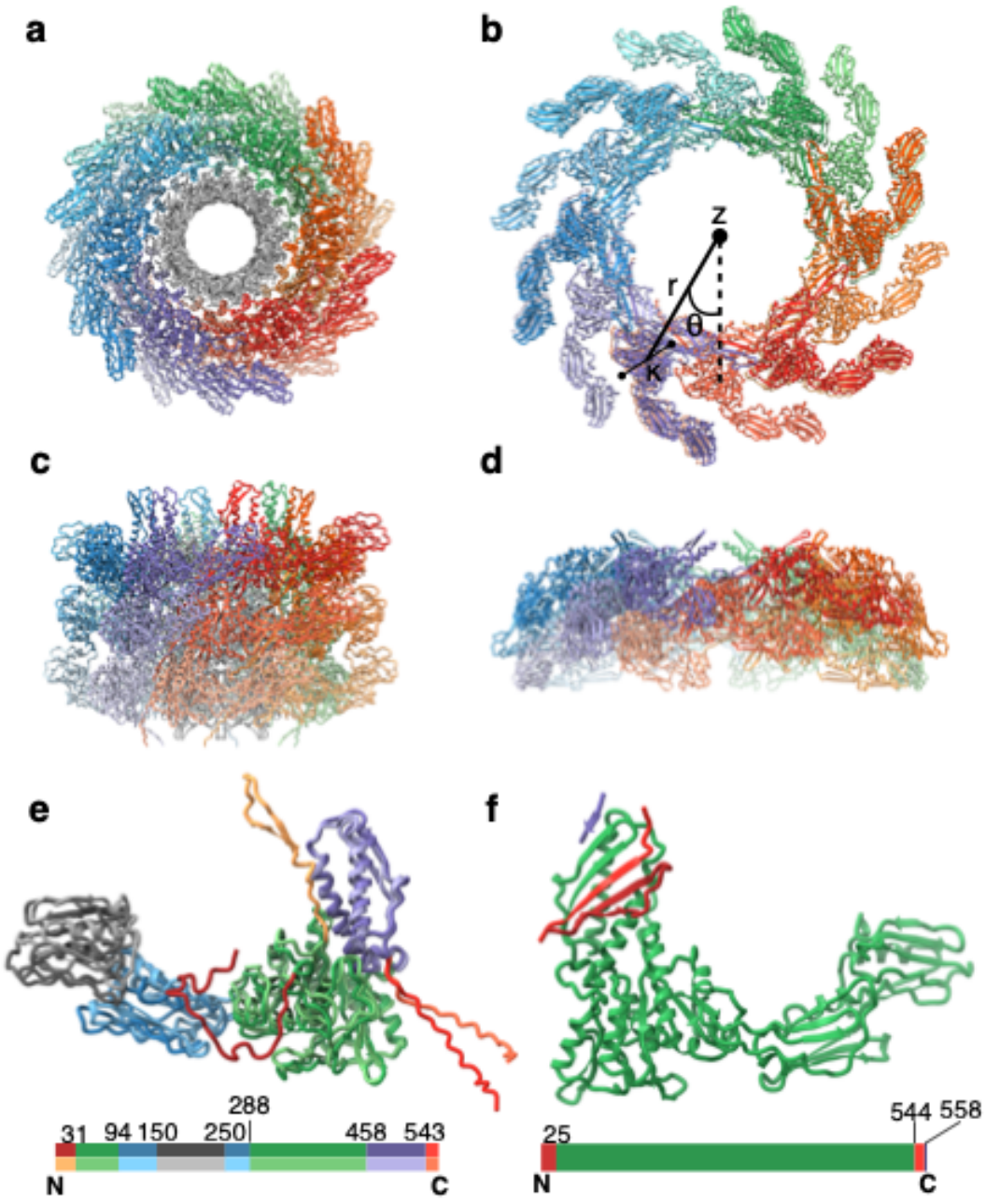


Figure 4.1: Structures of the A511 sheath

(a) Top view of the atomic model for 3 discs of the extended A511 sheath tube complex. Sheath subunits are colored according to their strand, the tube is colored in gray.

(b) Top view of the atomic model for 3 discs of the contracted A511 sheath. Sheath subunits are colored according to their strand. Lines are drawn to show (r, θ, z, κ) parameters, where the z-axis is out of the page.
(c) Side view of the atomic model for 3 discs of the extended A511 sheath tube complex. Sheath subunits are colored according to their strand, the tube is colored in gray.
(d) Side view of the atomic model for 3 discs of the contracted A511 sheath. Sheath subunits are colored according to their strand.
(e) Superposition of the extended and contracted A511 sheath subunit, with the extended state in darker color. The N-terminal arm is shown in brown, the main globular domain in green, the preprotrusion domain in blue, the protrusion domain in grey, the handshake (minus N/C-terminal arms) in purple and the C-terminal arm in red. The top and bottom N/C-terminal bar represent the colors of the extended and contracted subunits according to their residue numbers, respectively.
(f) Effective contracted subunit. The "central" subunit is shown in green. The subunit clockwise from within the same disc is shown in brown. The subunits from the same strand but one or two discs above are shown in red or purple, respectively. The N/C terminal bar represents the colors of the effective subunit according to their residue numbers.

interactions between 4 distinct subunits, and as a whole form the handshake domain(Kudryashev et al. 2015) (Fig 4.1f).

Like other systems, the conformation of the individual sheath subunit is similar between the extended and contracted states(Wang et al. 2017),(Jiang et al. 2019),(Desfosses et al. 2019) (Fig 4.1e). In particular, the handshake domain (minus the N/C-terminal arms) and the main globular domain have

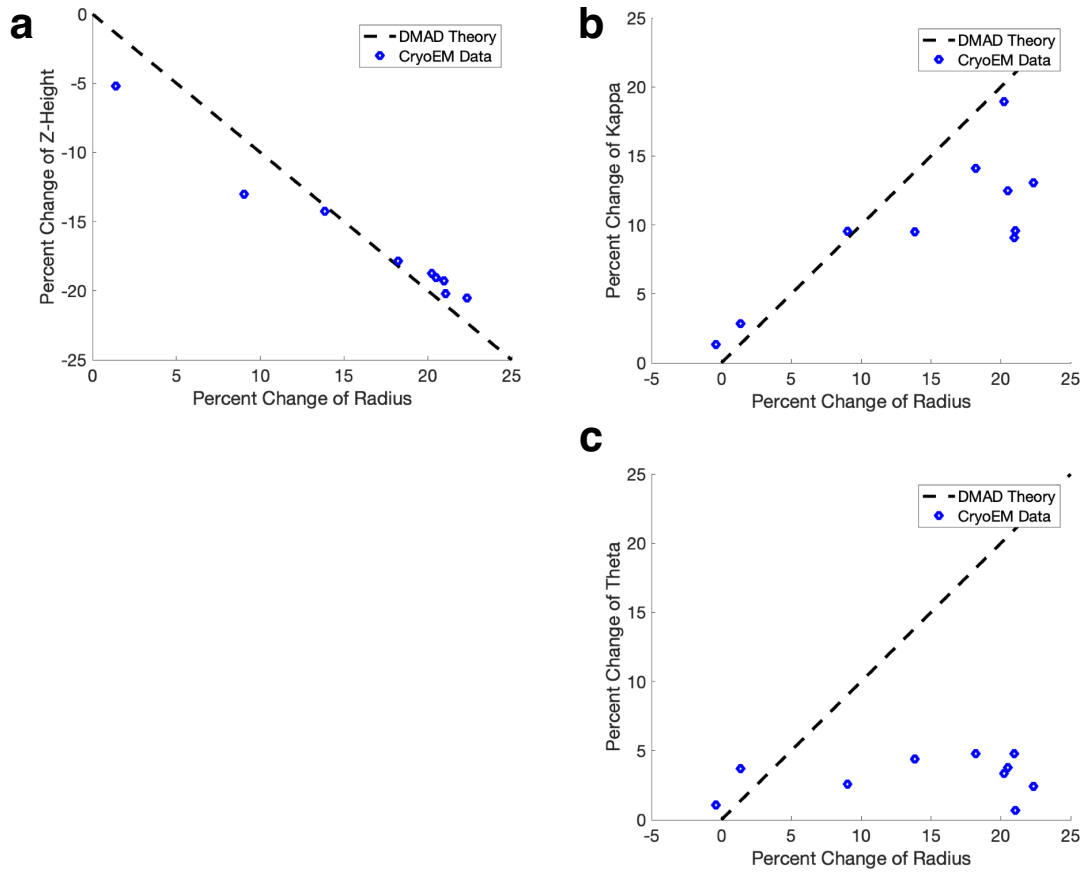


Figure 4.2: Subunit motions during late-stage contraction

Scatter plot showing the relative change in the cylindrical height (a), the polar angle kappa (b) or the cylindrical angle theta (c) vs. the radius of the ten baseplate proximal sheath discs between the intermediate and contracted states. The dashed lines indicate the linear relationship assumed in the DMAD theory.

CA RMSDs of 1.31 Å and 1.17 Å between the extended and contracted states, respectively. Despite this, there is a small change in the orientation of the handshake domain relative to the main globular domain between the extended and contracted states. The C-terminal arm (residues 544-562) is also shifted slightly between the extended and contracted states. Importantly, the N-terminal arm (residues 2-31) changes its orientation and conformation drastically to maintain the integrity of the handshake domain (Fig 4.1e). Furthermore, in

the contracted state, the C-terminal arm forms an internal beta sheet (residues 3-6 & 18-21), which is not present in the extended state. Differences in the conformation of the preprotrusion and protrusion domains could not be assessed due to insufficient map quality in the region.

Structural transitions during the late-stage contraction

Comparison of the atomic structures of the sheath between the intermediate and contracted states provides novel insights into the process of sheath contraction. The motion of a rigid sheath subunit during contraction can be wholly described by a set of four parameters(Fraser et al. 2021): (r, θ, z, κ), where (r, θ, z) describes the cylindrical coordinates of the COM of the sheath subunit and κ is the angle of rotation about the defined rotational axis between the extended and contracted states (Fig 4.1b). How these parameters change relative to each other is central to understanding the contraction process. To this end, we measured the changes in all four parameters between the intermediate and contracted states for the ten baseplate proximal discs of the sheath (Fig 4.2). We found that three of the four parameters are linearly related (r, z, κ) (Fig 4.2a 4.2b) with θ being seemingly independent of the other three (Fig 4.2c). The constant value of the percentage change in θ (termed pθ) for all ten subunits (Fig 4.2c), indicates that pθ is a seemingly global parameter. This result can be attributed to the structure of the handshake domain (Fig 4.1f). While the handshake domain can re-orient (relative to the rest of the subunit) to accommodate changes in (r, z, κ) (this re-orientation is present between the extended and contracted states), the domain cannot tolerate substantial local differences in pθ without the shear stress disintegrating the native beta sheets of the handshake domain. It follows that both linearly related subunit-level parameters: (r, z, κ) and a non-linear global parameter: θ are necessary to describe the contraction process.

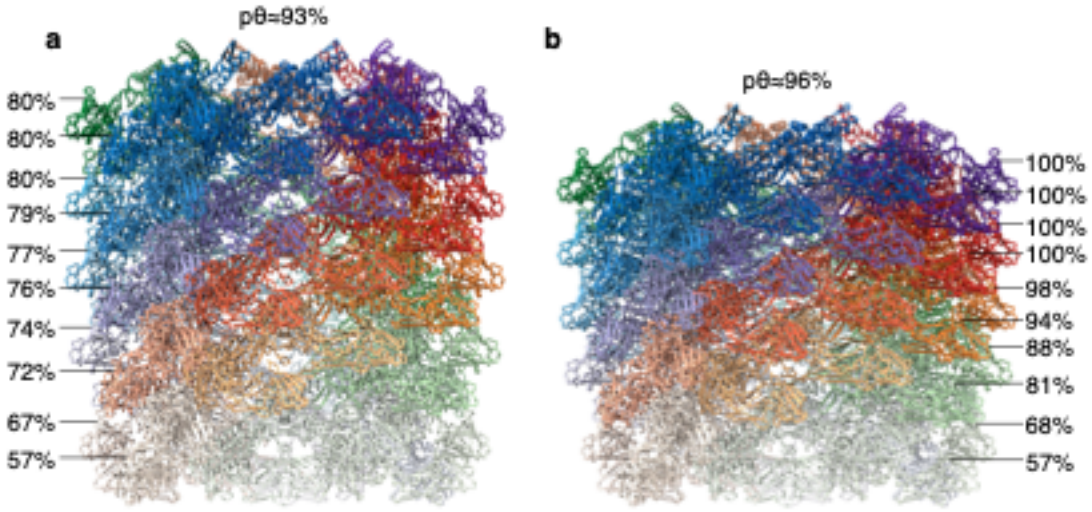


Figure 4.3: Structures of the baseplate proximal sheath.

Side view of the atomic models for the intermediate state (a) or contracted state (b) comprising the 10 most baseplate proximal discs. Sheath subunits are colored according to their strand. Lines denotes the "percentage contracted" according to the (r, z, κ) parameters.

Structure of the intermediate and contracted sheath

We can describe the structure of the 10 baseplate-proximal discs of the intermediate and contracted sheaths using both (r, z, κ) and θ parameters (Fig 4.3). Relative to a theoretical "fully contracted" sheath, with respect to the (r, z, κ) parameters, the intermediate and contracted models are 74% and 89% contracted overall, respectively, with the 1st discs (most baseplate-proximal) being both 57% contracted and the 10th discs being 80% and 100% contracted, respectively. Globally, with respect to the θ parameter and relative to a theoretical "fully contracted" sheath, the intermediate and contracted structures are ~93% and ~96% contracted, respectively.

Tolerance of the sheath to the contraction wave

The relationship between the orientations and positions of subunits in the intermediate and contracted states provides insight into the tolerance of the sheath handshake domain to deviations in the (r, z, κ) parameters. Interestingly, the deviations in the contracted state are more pronounced than in the intermediate state (Fig 4.3). We interpret this result via the energetics of the system. In the contracted state, all discs above the 7th disc are fully contracted (Fig 4.3b). These interactions between subunits are highly energetically favorable(Fraser et al. 2021) and as such apply substantial force on the baseplate-proximal subunits to contract. They remain partially extended however, due to the interaction between the baseplate proximal subunits and the sheath initiator protein. This pressure accordingly applies stress on the handshake domains, resulting in substantial deviations between adjacent subunit (r, z, κ) parameters (Fig 4.3b). As such, the deviations in (r, z, κ) parameters in the contracted state likely represents the maximal deformation present throughout the entire contraction process. In the contracted baseplate proximal structure, the bottom disc is 57% contracted, while the 7th disc is fully contracted (Fig 4.3b). Accordingly, on average, each adjacent disc can tolerate at maximum a ~6% deviation in (r, z, κ) parameters.

DISCUSSION

Extrapolation of the contraction wave to the full length intermediate

In the contracted baseplate-proximal structure, the bottom seven discs can tolerate an average deviation in (r, z, κ) parameters of ~6%. Extrapolation of this result to the entire length of the sheath structure would suggest that a theoretical maximally constrained intermediate could have fully extended discs separated from fully contracted discs by just 16 intermediate discs. Despite this theoretical lower bound on the number of discs separating fully extended from fully contracted subunits or the “contraction wave”(Fraser et al. 2021), the Cryo-EM structure of the bacteriophage contraction intermediate (Fig 4.3a)

suggests that the real value would be substantially larger. Furthermore, in our case, where baseplate proximal discs do not reach the fully contracted state (Fig 4.3b), a putative state where fully extended discs are connected to fully contracted discs via any number of intermediate discs seems unlikely.

MATERIALS AND METHODS

Bacteriophage purification

Phage A511 was propagated as described in (PMID: 30606715). *Listeria ivanovii* strain WSLC 3009 (SV 5) was grown overnight in 1/2x Brain Heart Infusion medium at 220 rpm at 30°C. 20 ml of an overnight bacterial culture was added to 1 L of pre-warmed (30°C) ½ BHI together with purified phage stocks to a final concentration of 10^5 pfu/ml. This culture was incubated to an OD_{600} of ~0.1 when additional phages were added to a final concentration of $2x10^7$ pfu/ml. Incubation continued until culture clearing for approx. 2 h when the solution was placed at 4°C. Bacterial cellular debris was removed by centrifugation at 6000×g for 15 min at 4°C. Phages were purified by addition of 10% PEG 8.000 and 1M NaCl, overnight incubation in ice water and centrifugation at 10000×g for 15 min at 4°C. Pellet was resuspended in SM buffer (50 mM Tris-HCL pH7.5, 100 mM NaCl, 8 mM $MgSO_4$) and phages were purified by CsCl gradient centrifugation (1.55 g/L) at 76000×g for 18 hours at 4°C (PMID: 18567664).

Cell membrane purification

Cell wall fragments were purified and isolated from *Listeria ivanovii* as described in (PMID: 30606715). *Listeria ivanovii* strain WSLC 3009 (SV 5) is inoculated into a 2mL ½ Brain Heart Infusion media and incubated overnight at 30°C. 1mL of the overnight culture was added to 1 L of pre-warmed (30°C). This culture was grown to an OD_{600} of

~1.0 and centrifuged at 7,000xg for 10 minutes. The pellet was suspended in 10mL of 1X SM buffer (100mM NaCl, 8mM $MgSO_4$, 50mM Tris-HCl pH 7.5). The sample was store overnight at -20C. The following day the sample was thawed and heated at 100°C for 20 minutes. The cell walls were disrupted by passing the sample through a pressure cell homogenizer three times. The sample was then centrifuged at 20,000xg for 30 minutes. The pellet was suspended in 500mL water and centrifuged at 20,000xg for 30 minutes, and repeated once more. 100ug of RNAse A and DNAse 1 were added to the sample and shaken for 3.5 hours at room temperature. At this point, 100ug of Proteinase K was added and shaken for 2 hours at room temperature. The sample was boiled in 4% SDS for 30 minutes at 100°C. The sample was centrifuged at 20,000xg for 30 minutes, and the supernatant was decanted. Pellet was washed in 500mL of water for 20 minutes at 20,000xg and repeated for a told of six washes. The pellet was suspended in 10mL SM Buffer and placed at 4°C for storage.

Cryo-EM sample preparation

Cell membranes were sonicated for 15 minutes. The bacteriophages were then incubated with the cell membranes for 3 minutes prior to grid freezing. TED PELLA 200 mesh PELCO NetMesh copper grids were plasma cleaned by the Gatan advanced plasma cleaning system. 3 ul of sample was pipetted onto the plasma cleaned grids using a Thermo Fischer Scientific Vitrobot for 20s at 100% humidity. The grids were plunged into liquid ethane. A Titan Krios 300kV electron microscope with a BioQuantum K3 imaging filter and 20eV electron slit was used for imaging. Using EPU software, ~32k micrographs (4096x4096, 59 frames pixel size 1.1 Å, and exposure time of 1.5s and a total electron dose of 40 e/ $Å^2$) were collected over a defocus range of -1.5 to -3.5 microns.

Cryo-EM image processing

Image processing, motion correction and CTF estimation were performed with CryoSARC(Punjani et al. 2017). Particle boxing was performed with CryoSPARC via a combination of manual, template-based and helical picking methods. For the extended sheath, 117,900 particles resulted from 2D classification and helical picking from a manually picked 2D classification template. These particles were refined with helical symmetry imposed, symmetry expanded and locally refined to a resolution of 3.4 Å. The resulting map was sharpened with a B factor of -154. For the contracted sheath, 64,810 particles resulted from 2D classification and helical picking from a manually picked 2D classification template. These particles were refined with helical symmetry imposed, symmetry expanded and locally refined to a resolution of 3.1 Å. The resulting map was sharpened with a B factor of -114. For the contracted baseplate proximal sheath, particles were picked with a manually picked 2D classification template, down sampled by a factor of 2 and resulted in 10,259 particles after 2D classification. These particles were subject to ab inito reconstruction with 3 classes, which resulted in one populous class with 5185 particles. These particles were refined with C6 symmetry to a resolution of 5.9A. For the intermediate baseplate proximal sheath, particles were picked manually. 1428 particles were extracted with 1133 remaining after 2D classification. Those particles were down sampled by a factor of 2 and passed through ab initio reconstruction with 3 classes, one populous class emerged with 736 particles, which was refined with C6 symmetry to a resolution of 8.5 Å. These particles were subjected to particles subtraction with a mask over the baseplate. 736 particles were locally refined with a mask over the sheath to a resolution of 8.2 Å.

Model building and refinement

The structure of the contracted sheath subunit was built first. The handshake, main globular and parts of the preprotrusion and protrusion domains were built manually using

the 3.1 Å Cryo-EM map. Peripheral regions of the preprotrusion and protrusion domains were inadequate for de novo model building and were predicted using RoseTTAFold(Baek et al. 2021). The predicted model was then docked, and real space refined in the Cryo-EM map. The model was then refined using Phenix real space refinement(Adams et al. 2011). In the case of the extended sheath subunit, the flexibility of the sheath made de novo model building difficult. The contracted sheath subunit structure was rigidly docked into the Cryo-EM map and the structure was rebuilt in regions where there were differences (mostly the N-terminal arm). The structure of the tube protein was predicted using RoseTTAFold(Baek et al. 2021) and subsequently rebuilt and refined in the extended sheath Cryo-EM map using Coot (27). For the baseplate-proximal sheath structures, the contracted sheath subunit was broken into an effective subunit (Fig 4.1f), by combining the N/C-terminal arms of adjacent subunits to form a complete handshake domain(Ge et al. 2015)·(Kudryashev et al. 2015)·(Wang et al. 2017)·(Jiang et al. 2019). The effective subunits were rigidly docked into the Cryo-EM maps. Following docking, some edge-case arms were manually moved to match the Cryo-EM structure. After this, the N/C-terminal arms were rebuilt manually in conjunction with Coot(Emsley, Cowtan, and IUCr 2004) regularization. The resulting structures were subject to Phenix(Adams et al. 2011) dynamics to improve geometry and reduce clashes.

Molecular graphics

Figures 4.1 and 4.3 were created using UCSF ChimeraX(Goddard et al. 2018). Graphs in Figure 4.2 were created using MATLAB (MathWorks).

ACKNOWLEDGEMENTS

This work was supported by the UTMB Sealy Center for Structural Biology and Molecular Biophysics and the Department of Biochemistry and Molecular Biology. We thank Dr. Michael B. Sherman for his help in Cryo-EM data collection. The work present in this paper used the resources of the UTMB SCSB Cryo-EM laboratory.

Author contributions

A.F and **P.G.L** conceived the study. **A.F.** contributed to Cryo-EM data collection, image processing, model building and structural analyses. **N.P** contributed to Cryo-EM sample preparation and Cryo-EM data collection. **J.M** contributed to cell membrane purification. **E.K** contributed to bacteriophage purification. **P.G.L** contributed to model building and structural analyses. **A.F.** and **P.G.L** wrote the manuscript, which was read, edited, and approved by all authors.

Competing interests

The authors declare no competing interests.

Correspondence and requests

Correspondences should be addressed to Petr G. Leiman.

Chapter 5 Summary and Future Directions

SUMMARY

In Chapter 1, molecular/structural biology is introduced as a scientific discipline and the historical importance of protein dynamics is outlined from conformational adaptability to intrinsically disordered proteins. The four main dynamical structural biology techniques: nuclear magnetic resonance spectroscopy, time-resolved x-ray crystallography, cryo-electron microscopy and molecular dynamics are introduced and the ability of these methods to characterize protein dynamics is assessed. The integration of machine learning tools in structural biology is introduced and finally, the future role of structural biology methods is described in the context of highly accurate protein folding algorithms such as AlphaFold2 and RoseTTAFold.

Chapter 2 introduces the "jumbo" *Bacillus subtilis* infecting bacteriophage AR9. Importantly, key properties of the phage, such as its uracil-containing dsDNA genome and its non-virion RNA polymerase are discussed. Furthermore, properties of the AR9 non virion polymerase, such as its dependance on uracil containing promoter DNA and template-strand promotor recognition are described. The AR9 nvRNAP structure is compared to *E. coli* RNAP. The components necessary for template strand promoter recognition are characterized. A mutant nvRNAP, which has its promoter specificity subunit altered such that it can recognize both T and U containing DNA templates in the -10 position, is presented. The energetics of DNA binding to the promoter pocket is characterized using double decoupling method molecular dynamics. Finally, a four step model for the disorder-to-order transition of the bacteriophage AR9 nvRNAP during promoter recognition is proposed.

Chapter 3 begins by introducing the conversed machinery of contractile injection systems. Previous experimental and theoretical attempts at characterizing macromolecular

injection systems are discussed. A computational approach to describe the mechanism of R-type pyocin contraction of a 12-layer fragment is outlined. The method is then extended to the full length structure. Solution biophysics experiments (isothermal titration calorimetry, differential scanning calorimetry and circular dichroism) are used to probe the energetics of pyocin contraction. Atomic force microscopy is used to assess the forces generated in later stages of pyocin contraction. The prediction that N-terminal sheath subunit linker mutants would have a reduced activation energy is validated by solution biophysics experiments. Finally, computational and experimental findings are combined to develop a wholistic model for the energetics and forces involved in pyocin contraction.

In Chapter 4, bacteriophage A511, a contractile phage, which has been captured in a transient contraction intermediate state, is introduced. The cryo-EM-derived structure of the extended and contracted states of the sheath-tube complex is characterized. Furthermore, the contracted sheath subunit structure is docked into the stalled contraction intermediate electron density map. This structure is then compared to the contracted structure, such that the relationship between subunit motions during the late stages of contraction is uncovered. These results show that bacteriophage contraction can be described by the rigid motions of sheath subunit with both local subunit-level and global properties.

FUTURE DIRECTIONS

The future goal is to extend the methods described in this work such that they can be applied to other systems. In particular, the methodologies of Chapter 4 can be further developed for the streamlined classification and reconstruction of low population states in Cryo-EM. The major bottleneck to the characterization of the stalled contraction intermediate of Chapter 4 was the manual picking of intermediate particles. With this dataset, conventional 2D and 3D classification algorithms were unable to separate

intermediate particles from contracted particles with sufficient fidelity for map generation. This was likely a result of various factors, namely, i) the abundance of contracted vs. intermediate particles in the dataset (the ratio was ~10:1), ii) the similarity of the baseplate proximal regions of the intermediate and contracted states and iii) the limited number of particles in the dataset. It may not be surprising that the classification algorithms were unable to separate a homogenous subset of 500+ intermediate particles from a set consisting of ~10k contracted and ~1k intermediate particles. Accordingly, a future goal is to develop a procedure which can classify low population states more accurately than conventional clustering algorithms.

To this end, a convolutional neural network was devised to identify and classify intermediate particles from a mixed particles set (such as those derived from automated picking procedures). The neural network was trained on a manually picked dataset of 400 intermediate and 400 contracted particles. Using data augmentation techniques, the number of training images was artificially expanded to 130,000 images. After 100 epochs of training, the algorithm was tested on an independent validation set and was found to be 87% accurate at classifying intermediate vs. contracted particles. The ability of the neural network to make predictions on this validation set is shown in the receiver operating characteristic (ROC) curve (Fig 5.1).

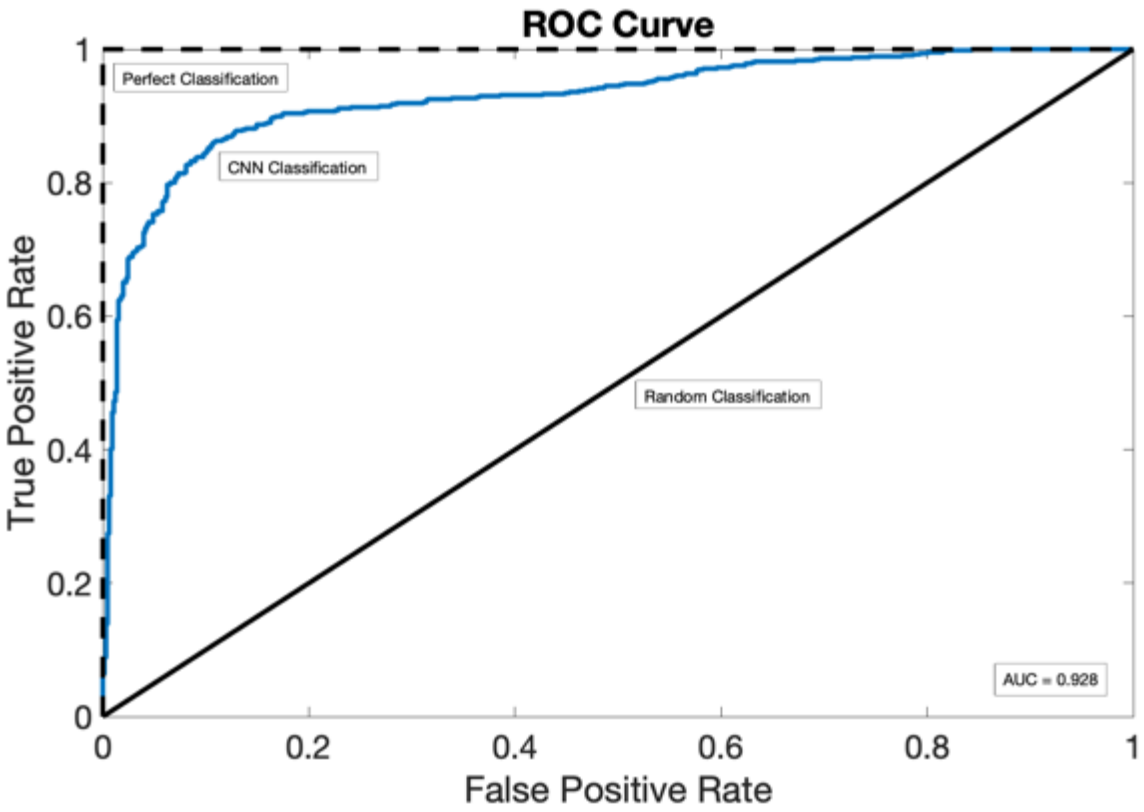


Figure 5.1: ROC curve for the identification of intermediate particles on a validation dataset.

Plot showing true positive rate (probability that an intermediate will be classified as an intermediate) and false positive rate (probability that a contracted particle will be classified as an intermediate) as metrics for model accuracy. Dashed black line shows a theoretical perfect classification result. Solid black line shows a theoretical random classification result. Blue line shows the classification result of the CNN. The area under the curve (AUC) is 0.928.

As a future goal, once the CNN reaches peak performance, it can be used to classify particle sets resulting from automated particle picking into intermediate and contracted sets. These sets can then be separately refined and the resulting electron density maps can be analyzed. In the ideal scenario, the intermediate map would be similar in quality to that

resulting from the manual picking of intermediate particles. In this case, larger datasets (consisting of 100k+ micrographs) can be processed in a near-fully automated fashion, with the goal of improving the intermediate sheath map quality from 8.2 Å to 4.5-4.0 Å resolution. In theory, this technique can be used for the classification of other systems where conventional methods are inadequate.

Secondly, in the future, PISA analysis (like that undergone in Chapter 3) will be performed on the extended, contracted, and intermediate bacteriophage A511 sheath-tube complexes. From this, a partial model for the energetics of sheath contraction at 3 locations along the free energy pathway will be created.

Given that the results of Chapter 4 show that the motion of sheath subunits during bacteriophage contraction can be characterized by both linear and non-linear properties, the modelling technique of Chapter 3 can be extended to reflect this experimental result. In principle, the DMAD methodology can be applied to the bacteriophage A511 system such that the sets of transition structures are generated using both non-linear and linear parameters. Using the same structural constraints and energetic analysis, putative transition pathways can be generated for the bacteriophage A511 system. In this case, predicted transition intermediates can be directly compared to the experimentally derived structures to further improve model accuracy.

To enable the comparison of the adapted DMAD method derived protein structures with experimental A511 reconstructions, methods need to be employed to improve the resolvability of baseplate-distal sheath subunits. In this instance, the most straightforward approach includes collecting a larger cryo-EM dataset, followed by CNN-mediated particle curation/classification then focused reconstruction. Furthermore, classification methods should be used to remove significantly flexible sheath structures which dramatically reduce map quality.

Alternatively, an analogous approach can be used to characterize the structure of the capsid proximal sheath subunits of bacteriophage A511. With this approach the capsid will serve as a reference for the alignment of particles. Together with the baseplate-proximal sheath reconstruction, this combined method should be sufficient for the characterization of approximately two thirds of the sheath subunits in the stalled contraction intermediate state of bacteriophage A511. The structure of the remaining third can be inferred via the interpolation of flanking parts of the sheath.

Appendix

Table 2.1. Oligonucleotides used in the AR9 study.

PCR primers used for site-directed mutagenesis of gp226 (Figure 2b, 2d, 2f)		
Mutation	*Oligonucleotide name*	*5′-3′ sequence*
gp226 V206G	gp226-V206X-rev	aacatctgagtactttgtctcgaagacgcga
	gp226-V206G-dir	gagacaaagtactcagatgttggaatctggacgtaccttaaaaac
A^5 mutant (gp226 R389A, K390A, R394A, K395A, K396A)	gp226-R389X-rev	cacatttggtgcgatagcgcaggataaaatctc
	gp226-A-for	gctatcgcaccaaatgtggccgcgatgaacaacgctgcagcattagttgataaaatcatc
gp226 Y246A	gp226_Y246A_F	atctccgcgcttgacgttgtgatcaaacagaaattg
	gp226_Y246A_R	gtcaagcgcggagattaccgagctattgtgtttc
gp226 S245E	gp226_S245E_F	gtaatcgaataccttgacgttgtgatcaaacag
	gp226_S245E_R	aaggtattcgattaccgagctattgtgtttcaac
PCR primers used to create the plasmid for expression of the tagless version of AR9 nvRNAP core		
	Oligonucleotide name	*5′-3′ sequence*
	nvRNAP-tag-free-for-66.5	tttaactttaagaaggagatataccatggggaaaaaattatcgttaatcgatttcaac
	nvRNAP-tag-free-rev-72.1	cagcagcggtttctttaccagactcgagttatttttcatc
DNA oligonucleotides in the AR9 nvRNAP promoter complex used for structure determination		
	Oligonucleotide name	*5′-3′ sequence*
	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand all U	ctccaatatgtgatataatatauguuuattg
DNA templates used to examine the dependence of *in vitro* transcription activity of the AR9 nvRNAP on the position and number of T bases in the promoter (Extended data figure 1c)		
Name of the template in the figure	*Oligonucleotide name*	*5′-3′ sequence*
all U	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand all U	ctccaatatgtgatataatatauguuuattg
(-12)T	[+3;+16] non-template strand	atcacatattggag

	[-16;+16] template strand T(-12)	ctccaatatgtgatataatatauuguutattg
(-11)T	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand T(-11)	ctccaatatgtgatataatatauugutuattg
(-10)T	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand T(-10)	ctccaatatgtgatataatatauugtuuattg
(-8)T	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand T(-8)	ctccaatatgtgatataatatautguuuattg
(-7)T	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand T(-7)	ctccaatatgtgatataatatatuguuuattg
all T	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand all T	ctccaatatgtgatataatatattgtttattg
DNA templates used to examine the *in vitro* transcription activity of the AR9 nvRNAP gp226 V206G mutant (Figure 2b)		
Name of the template in the figure	*Oligonucleotide name*	*5′-3′ sequence*
all U	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand all U	ctccaatatgtgatataatatauuguuuattg
(-11)T	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand T(-11)	ctccaatatgtgatataatatauugutuattg
(-10)T	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand T(-10)	ctccaatatgtgatataatatauugtuuattg
all T	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand all T	ctccaatatgtgatataatatattgtttattg

DNA templates used to examine the *in vitro* transcription activity of the AR9 nvRNAP gp226 Y246A and gp226 S245E mutants (Figure 2d)		
Name of the template in the figure	*Oligonucleotide name*	*5′-3′ sequence*
ds DNA	[-16;+16] non-template strand	caataaacaatatattatatcacatattggag
	[-16;+16] template strand all U	ctccaatatgtgatataatatauuguuuattg
fork DNA	[+3;+16] non-template strand	atcacatattggag
	[-16;+16] template strand all U	ctccaatatgtgatataatatauuguuuattg
PCR primers that were used for PCR amplification of genomic DNA fragments to examine the *in vitro* transcription activity of the A^5 mutant (Figure 2f)		
Name of the template in the figure	*Oligonucleotide name*	*5′-3′ sequence*
[-60;+80] DNA	P077-for-UP-60-63.4	taatcctcctacttatctagtctataattaattgttg
	P077-rev-ROff-80-61.9	attgcttcattaaacataaatgaagactc
[-16;+80] DNA	P077-for-UP-16-60.4	caataaacaatatattatatcacatattggagg
	P077-rev-ROff-80-61.9	attgcttcattaaacataaatgaagactc

Table 2.2. X-ray data collection and refinement statistic.

	AR9 nvRNAP core native (PDB code: 7S00)	AR9 nvRNAP core thimerosal (Hg) derivative	AR9 nvRNAP core Ta_6Br_{12} derivative	AR9 nvRNAP core thimerosal (Hg) derivative large unit cell	AR9 nvRNAP promoter complex native (PDB code: 7S01)
Data collection					
Beamline	APS 21-ID-G	APS 21ID-F	ALS 5.0.2	APS 21-ID-D	APS 21-ID-D
Detector	Rayonix MX-300	Rayonix MX-300	Dectris Pilatus3 6M 25Hz	Dectris Eiger 9M	Dectris Eiger 9M
Wavelength (Å)	0.97857	0.97872	1.25515	1.0050	0.91840
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	C2
Cell dimensions					
a, *b*, *c* (Å)	112.86, 166.27, 307.22	113.43, 169.51, 308.30	112.77, 171.51, 309.76	171.24, 231.78, 592.45	176.93, 110.46, 222.38
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 98.52, 90.00
Resolution (Å)	50.0-3.30 (3.50-3.30)*	50.0-3.60 (3.82-3.60)	50.0-4.53 (4.81-4.53)	50.0-3.79 (4.02-3.79)	50.0-3.38 (3.58-3.38)
R_{merge} (%)	12.3 (124.2)	18.2 (201.2)	20.6 (172.2)	20.9 (138.0)	15.1 (101.1)
I / σI	10.07 (1.17)	8.77 (1.13)	6.75 (1.07)	6.22 (1.02)	7.91 (1.13)
Completeness (%)	98. 6 (98.7)	99.7 (98.5)	99.6 (98.3)	99.4 (99.7)	95.3 (85.7)
Redundancy	4.55 (4.20)	6.95 (6.87)	6.25 (6.44)	6.30 (6.18)	3.33 (3.40)
$CC_{1/2}$	99.8 (42.1)	99.8 (54.8)	99.7 (47.1)	99.5 (52.4)	99.2 (65.0)
Refinement					
Resolution (Å)	49.4 – 3.30				50.0 – 3.40
No. reflections	86,075				56,728
R_{work} / R_{free}	0.2180 / 0.2580				0.2379 / 0.2921
No. atoms					
Protein	33,892				21,749
Ligand/ion	2 (Zn^{2+})				1,500 (DNA) / 96 (ions)
Water	0				0
B-factors (Å^2)					
Protein	154.65				126.15
Ligand/ion	154.85				239.32 (DNA) / 157.88 (ions)
Water	NA				NA
R.m.s. deviations					
Bond lengths (Å)	0.003				0.002
Bond angles (°)	0.57				0.462
Validation					
MolProbity score	1.46				1.34
Clashscore	8.48				6.24
Poor rotamers (%)	0.03				0.00
Ramachandran plot					
Favored (%)	98.30				98.07
Allowed (%)	1.70				1.93
Disallowed (%)	0.00				0.00

*Values in parentheses are for the highest resolution shell.

Table 2.3. Cryo-EM data collection and refinement

	AR9 nvRNAP promoter complex (EMDB-24763)	AR9 nvRNAP holoenzyme (EMDB-24765)
Data collection and processing		
Magnification	80,000	130,000
Voltage (kV)	300	300
Electron exposure (e^{-}/Å^2)	43.7	43.2
Defocus range (μm)	−1 to −4	−1 to −3
Pixel size (Å)	1.09	1.08
Symmetry imposed	C1	C1
Initial particle images (no.)	420,791	227,577
Final particle images (no.)	106,876	104,471
Map resolution (Å)	3.8	4.4
FSC threshold	0.143	0.143
Map resolution range (Å)	2.5-5.5	3.1-6.1

Table 2.4. Summary of MD simulation configurations and results.

Free energy term	Simulation Name	Window Number	Window Steps ($\times 10^3$)	Equilibration Steps per Window ($\times 10^3$)	Total Time (ns)	Energy Value (kcal/mol)
	DNA-RNAP Alchemical	400	40	8	64	705.7 ± 2.3
	DNA (bulk water) Alchemical	400	150	30	240	693.0 ± 0.4
It is a sum of seven energy terms with a combined (propagated) error	DNA-RNAP Constraint: r	20	600	120	48	0.3 ± 0.0
	DNA-RNAP Constraint: ϕ	20	60	12	4.8	0.3 ± 0.0
	DNA-RNAP Constraint: θ	20	300	60	24	0.9 ± 0.4
	DNA-RNAP Constraint: χ	20	60	12	4.8	0.2 ± 0.0
	DNA-RNAP Constraint: ψ	20	60	12	4.8	1.0 ± 0.0
	DNA-RNAP Constraint: ξ	20	60	12	4.8	0.9 ± 0.0
	DNA-RNAP Constraint: RMSD	20	600	120	48	3.3 ± 1.4
	DNA (bulk water) Constraint: RMSD	19	10,000	2,000	760	12.7 ± 0.1

Table 3.1. Properties of atomic models refined against the cryo-EM data and the transition state models obtained by the DMAD modeling procedure as analyzed by Molprobity.

Conformation of pyocin	Extended	Contracted	Transition state		
			Best	Worst	Intermediate
Source of atomic model	Cryo-EM map at 2.8 Å resolution	Cryo-EM map at 2.9 Å resolution	DMAD	DMAD	DMAD
Molprobity score	1.56	1.57	2.42	2.31	2.21
Ramachandran favored (%)	96.8	95.2	97.0	97.7	97.4
Ramachandran outliers (%)	0.0	0.0	0.2	0.1	0.1
Bad angles (%)	0.0	0.0	0.0	0.0	0.0
Favored rotamers (%)	95.3	96.0	86.1	85.8	86.3
Poor rotamers (%)	0.8	0.3	7.4	6.9	6.5

Table 3.2. Primers used to obtain mutants carrying insertions in the intra-strand linker.

Mutation	Primer	Sequence
all linker mutants	AE21M1F	GGTACCGGGCCCCCCCTCGAGGGCGGAGACTTCAACCCCAGTGA
all linker mutants	AE21M2R	CGCTCTAGAACTAGTGGATCCCCTCGACACGGAAATTGGGGTTCT
G insertion	AE21M1R	GATGACACTTGAACCGGCCGGCAGCGCGATG
	AE21M2F	GGTTCAAGTGTCATCGGCCTCTGCGATGTGTT
GA insertion	AE22M1R	GACACTTGATGCACCGGCCGGCAGCGCGATG
	AE22M2F	GGTGCATCAAGTGTCATCGGCCTCTGCGATGTGTT
GAG insertion	AE23M1R	ACTTGAACCTGCACCGGCCGGCAGCGCGATG
	AE23M2F	GGTGCAGGTTCAAGTGTCATCGGCCTCTGCGATGTGTT
GAGA insertion	AE24M1R	TGATGCACCTGCACCGGCCGGCAGCGCGATG
	AE24M2F	GGTGCAGGTGCATCAAGTGTCATCGGCCTCTGCGATGTGTT

References

A, Ishihama. 1981. “Subunit of Assembly of Escherichia Coli RNA Polymerase.” *Advances in Biophysics* 14 (January): 1–35. https://europepmc.org/article/med/7015808.

Adams, Paul D., Pavel V. Afonine, Gábor Bunkóczi, Vincent B. Chen, Nathaniel Echols, Jeffrey J. Headd, Li Wei Hung, et al. 2011. “The Phenix Software for Automated Determination of Macromolecular Structures.” *Methods* 55 (1): 94–106. https://doi.org/10.1016/J.YMETH.2011.07.005.

Akke, Mikael. 2012. “Conformational Dynamics and Thermodynamics of Protein–Ligand Binding Studied by NMR Relaxation.” *Biochemical Society Transactions* 40 (2): 419–23. https://doi.org/10.1042/BST20110750.

Aksyuk, Anastasia A, Petr G Leiman, Lidia P Kurochkina, Mikhail M Shneider, Victor A Kostyuchenko, Vadim V Mesyanzhinov, and Michael G Rossmann. 2009. “The Tail Sheath Structure of Bacteriophage T4: A Molecular Machine for Infecting Bacteria.” *The EMBO Journal* 28 (7): 821–29. https://doi.org/10.1038/EMBOJ.2009.36.

and, M. T. Pisabarro, and L. Serrano*. 1996. “Rational Design of Specific High-Affinity Peptide Ligands for the Abl-SH3 Domain†.” *Biochemistry* 35 (33): 10634–40. https://doi.org/10.1021/BI960203T.

Ando, Toshio, Noriyuki Kodera, Eisuke Takai, Daisuke Maruyama, Kiwamu Saito, and Akitoshi Toda. 2001. “A High-Speed Atomic Force Microscope for Studying Biological Macromolecules.” *Proceedings of the National Academy of Sciences* 98 (22): 12468–72. https://doi.org/10.1073/PNAS.211400898.

Aue, W. P., E. Bartholdi, and R. R. Ernst. 2008. “Two-dimensional Spectroscopy. Application to Nuclear Magnetic Resonance.” *The Journal of Chemical Physics* 64 (5): 2229. https://doi.org/10.1063/1.432450.

Babu, M. Madan, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer. 2011. “Intrinsically Disordered Proteins: Regulation and Disease.” *Current Opinion in Structural Biology* 21 (3): 432–40. https://doi.org/10.1016/J.SBI.2011.03.011.

Bae, Brian, Andrey Feklistov, Agnieszka Lass-Napiorkowska, Robert Landick, and Seth A. Darst. 2015. “Structure of a Bacterial RNA Polymerase Holoenzyme Open Promoter Complex.” *ELife* 4 (September 2015). https://doi.org/10.7554/ELIFE.08504.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, et al. 2021. “Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network.” *Science* 373 (6557): 871–76. https://doi.org/10.1126/SCIENCE.ABJ8754.

Baldwin, Andrew J, and Lewis E Kay. 2009. “NMR Spectroscopy Brings Invisible Protein States into Focus.” *Nature Chemical Biology 2009 5:11* 5 (11): 808–14. https://doi.org/10.1038/nchembio.238.

Bao, Yu, and Robert Landick. 2021. “Obligate Movements of an Active Site-Linked Surface Domain Control RNA Polymerase Elongation and Pausing via a Phe-Pocket Anchor.” *BioRxiv*, January, 2021.01.27.428476. https://doi.org/10.1101/2021.01.27.428476.

Basler, M., M. Pilhofer, G. P. Henderson, G. J. Jensen, and J. J. Mekalanos. 2012. “Type VI Secretion Requires a Dynamic Contractile Phage Tail-like Structure.” *Nature 2012 483:7388* 483 (7388): 182–86. https://doi.org/10.1038/nature10846.

Basler, Marek. 2015. “Type VI Secretion System: Secretion by a Contractile Nanomachine.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 370 (1679). https://doi.org/10.1098/RSTB.2015.0021.

Bennett, Charles H. 1976. “Efficient Estimation of Free Energy Differences from Monte Carlo Data.” *Journal of Computational Physics* 22 (2): 245–68. https://doi.org/10.1016/0021-9991(76)90078-4.

Bepler, Tristan, Kotaro Kelley, Alex J. Noble, and Bonnie Berger. 2020. “Topaz-Denoise: General Deep Denoising Models for CryoEM and CryoET.” *Nature Communications 2020 11:1* 11 (1): 1–12. https://doi.org/10.1038/s41467-020-18952-1.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. “The Protein Data Bank.” *Nucleic Acids Research* 28 (1): 235–42. https://doi.org/10.1093/NAR/28.1.235.

Best, Robert B., and Gerhard Hummer. 2005. “Reaction Coordinates and Rates from Transition Paths.” *Proceedings of the National Academy of Sciences* 102 (19): 6732–37. https://doi.org/10.1073/PNAS.0408098102.

Beutler, Thomas C., Alan E. Mark, René C. van Schaik, Paul R. Gerber, and Wilfred F. van Gunsteren. 1994. “Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations.” *Chemical Physics Letters* 222 (6): 529–39. https://doi.org/10.1016/0009-2614(94)00397-1.

Beveridge, D L, and F M Dicapua. n.d. “FREE ENERGY VIA MOLECULAR

SIMULATION: Applications to Chemical and Biomolecular Systems." Accessed October 18, 2021. www.annualreviews.org.

Blakeley, M.P., S.S. Hasnain, and S.V. Antonyuk. 2015. "Sub-Atomic Resolution X-Ray Crystallography and Neutron Crystallography: Promise, Challenges and Potential." *Urn:Issn:2052-2525* 2 (4): 464–74. https://doi.org/10.1107/S2052252515011239.

Blomfield, I. C., V. Vaughn, R. F. Rest, and B. I. Eisenstein. 1991. "Allelic Exchange in Escherichia Coli Using the Bacillus Subtilis SacB Gene and a Temperature-Sensitive PSC101 Replicon." *Molecular Microbiology* 5 (6): 1447–57. https://doi.org/10.1111/J.1365-2958.1991.TB00791.X.

Bönemann, Gabriele, Aleksandra Pietrosiuk, and Axel Mogk. 2010. "Tubules and Donuts: A Type VI Secretion Story." *Molecular Microbiology* 76 (4): 815–21. https://doi.org/10.1111/J.1365-2958.2010.07171.X.

Brennan, Richard G, and Brian W Matthews. 1989. "THE JOURNAL OF BIOLOGICAL CHEMISTRY The Helix-Turn-Helix DNA Binding Motif." https://doi.org/10.1016/S0021-9258(18)94115-3.

Brenner, S., and R. W. Horne. 1959. "A Negative Staining Method for High Resolution Electron Microscopy of Viruses." *Biochimica et Biophysica Acta* 34 (C): 103–10. https://doi.org/10.1016/0006-3002(59)90237-9.

Browning, Christopher, Mikhail M. Shneider, Valorie D. Bowman, David Schwarzer, and Petr G. Leiman. 2012. "Phage Pierces the Host Cell Membrane with the Iron-Loaded Spike." *Structure* 20 (2): 326–39. https://doi.org/10.1016/J.STR.2011.12.009.

Campbell, Elizabeth A., Oriana Muzzin, Mark Chlenov, Jing L. Sun, C. Anders Olson, Oren Weinman, Michelle L. Trester-Zedlitz, and Seth A. Darst. 2002. "Structure of the Bacterial RNA Polymerase Promoter Specificity σ Subunit." *Molecular Cell* 9 (3): 527–39. https://doi.org/10.1016/S1097-2765(02)00470-7.

Cascales, Eric, and Christian Cambillau. 2012. "Structural Biology of Type VI Secretion Systems." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1592): 1102–11. https://doi.org/10.1098/RSTB.2011.0209.

Cash, Jennifer N., Sarah Urata, Sheng Li, Sandeep K. Ravala, Larisa V. Avramova, Michael D. Shost, J. Silvio Gutkind, John J. G. Tesmer, and Michael A. Cianfrocco. 2019. "Cryo–Electron Microscopy Structure and Analysis of the P-Rex1–Gβγ Signaling Scaffold." *Science Advances* 5 (10): 8855–71. https://doi.org/10.1126/SCIADV.AAX8855.

Caspar, D. L. 1980. "Movement and Self-Control in Protein Assemblies. Quasi-Equivalence Revisited." *Biophysical Journal* 32 (1): 103–38.

https://doi.org/10.1016/S0006-3495(80)84929-0.

Ceyssens, P.-J., L. Minakhin, A. Van den Bossche, M. Yakunina, E. Klimuk, B. Blasdel, J. De Smet, et al. 2014. "Development of Giant Bacteriophage KZ Is Independent of the Host Transcription Apparatus." *Journal of Virology* 88 (18): 10501–10. https://doi.org/10.1128/JVI.01347-14.

Chaikeeratisak, Vorrapon, Katrina Nguyen, MacKennon E. Egan, Marcella L. Erb, Anastasia Vavilina, and Joe Pogliano. 2017. "The Phage Nucleus and Tubulin Spindle Are Conserved among Large Pseudomonas Phages." *Cell Reports* 20 (7): 1563–71. https://doi.org/10.1016/J.CELREP.2017.07.064.

Chaudhury, Sidhartha, Sergey Lyskov, and Jeffrey J. Gray. 2010. "PyRosetta: A Script-Based Interface for Implementing Molecular Modeling Algorithms Using Rosetta." *Bioinformatics* 26 (5): 689–91. https://doi.org/10.1093/BIOINFORMATICS/BTQ007.

Chen, V.B., W.B. Arendall, J.J. Headd, D.A. Keedy, R.M. Immormino, G.J. Kapral, L.W. Murray, J.S. Richardson, D.C. Richardson, and IUCr. 2009. "MolProbity: All-Atom Structure Validation for Macromolecular Crystallography." *Urn:Issn:0907-4449* 66 (1): 12–21. https://doi.org/10.1107/S0907444909042073.

Cherrak, Yassine, Chiara Rapisarda, Riccardo Pellarin, Guillaume Bouvier, Benjamin Bardiaux, Fabrice Allain, Christian Malosse, et al. 2018. "Biogenesis and Structure of a Type VI Secretion Baseplate." *Nature Microbiology 2018 3:12* 3 (12): 1404–16. https://doi.org/10.1038/s41564-018-0260-1.

Chipot, Christophe. 2014. "Frontiers in Free-Energy Calculations of Biological Systems." *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4 (1): 71–89. https://doi.org/10.1002/WCMS.1157.

Cordero, Raul R, and Pedro Roth. 2005. "On Two Methods to Evaluate the Uncertainty of Derivatives Calculated from Polynomials Fitted to Experimental Data." *Metrologia* 42 (1): 39. https://doi.org/10.1088/0026-1394/42/1/005.

Cowtan, K., and IUCr. 2006. "The Buccaneer Software for Automated Model Building. 1. Tracing Protein Chains." *Urn:Issn:0907-4449* 62 (9): 1002–11. https://doi.org/10.1107/S0907444906022116.

———. 2010. "Recent Developments in Classical Density Modification." *Urn:Issn:0907-4449* 66 (4): 470–78. https://doi.org/10.1107/S090744490903947X.

Dance, Amber. 2020. "Molecular Motion on Ice." *Nature Methods 2020 17:9* 17 (9): 879–83. https://doi.org/10.1038/s41592-020-0940-7.

Dandey, Venkata P., William C. Budell, Hui Wei, Daija Bobe, Kashyap Maruthi, Mykhailo Kopylov, Edward T. Eng, et al. 2020. "Time-Resolved Cryo-EM Using

Spotiton." *Nature Methods 2020 17:9* 17 (9): 897–900. https://doi.org/10.1038/s41592-020-0925-6.

Darden, Tom, Darrin York, and Lee Pedersen. 1998. "Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems." *The Journal of Chemical Physics* 98 (12): 10089. https://doi.org/10.1063/1.464397.

David S. Cerutti, *,§, ‡,§,‖ and Lynn F. Ten Eyck, and §,‖ J. Andrew McCammon†. 2004. "Rapid Estimation of Solvation Energy for Simulations of Protein−Protein Association." *Journal of Chemical Theory and Computation* 1 (1): 143–52. https://doi.org/10.1021/CT049946F.

Desfosses, Ambroise, Hariprasad Venugopal, Tapan Joshi, Jan Felix, Matthew Jessop, Hyengseop Jeong, Jaekyung Hyun, et al. 2019. "Atomic Structures of an Entire Contractile Injection System in Both the Extended and Contracted States." *Nature Microbiology 2019 4:11* 4 (11): 1885–94. https://doi.org/10.1038/s41564-019-0530-6.

Diehl, Carl, Olof Engström, Tamara Delaine, Maria Håkansson, Samuel Genheden, Kristofer Modig, Hakon Leffler, Ulf Ryde, Ulf J. Nilsson, and Mikael Akke. 2010. "Protein Flexibility and Conformational Entropy in Ligand Design Targeting the Carbohydrate Recognition Domain of Galectin-3." *Journal of the American Chemical Society* 132 (41): 14577–89. https://doi.org/10.1021/JA105852Y.

Doerr, Allison. 2015. "Protein Structure through Time." *Nature Methods 2016 13:1* 13 (1): 34–34. https://doi.org/10.1038/nmeth.3704.

Donelli, G., F. Guglielmi, and L. Paoletti. 1972. "Structure and Physico-Chemical Properties of Bacteriophage G: I. Arrangement of Protein Subunits and Contraction Process of Tail Sheath." *Journal of Molecular Biology* 71 (2): 113–25. https://doi.org/10.1016/0022-2836(72)90341-5.

Dong, Yuanchen, Shuwen Zhang, Zhaolong Wu, Xuemei Li, Wei Li Wang, Yanan Zhu, Svetla Stoilova-McPhie, Ying Lu, Daniel Finley, and Youdong Mao. 2018. "Cryo-EM Structures and Dynamics of Substrate-Engaged Human 26S Proteasome." *Nature 2018 565:7737* 565 (7737): 49–55. https://doi.org/10.1038/s41586-018-0736-4.

Drake, Justin A., Robert C. Harris, and B. Montgomery Pettitt. 2016. "Solvation Thermodynamics of Oligoglycine with Respect to Chain Length and Flexibility." *Biophysical Journal* 111 (4): 756–67. https://doi.org/10.1016/J.BPJ.2016.07.013.

Drenth, J., and IUCr. 1974. "The Molecular Replacement Method. A Collection of Papers on the Use of Non-Crystallographic Symmetry Edited by M. G. Rossmann." *Urn:Issn:0567-7394* 30 (2): 304–304. https://doi.org/10.1107/S056773947400074X.

Dyson, H. Jane, and Peter E. Wright. 2005. “Intrinsically Unstructured Proteins and Their Functions.” *Nature Reviews Molecular Cell Biology 2005 6:3* 6 (3): 197–208. https://doi.org/10.1038/nrm1589.

Egelman, Edward H. 2016. “The Current Revolution in Cryo-EM.” https://doi.org/10.1016/j.bpj.2016.02.001.

Eisenberg, David, and Andrew D. McLachlan. 1986. “Solvation Energy in Protein Folding and Binding.” *Nature 1986 319:6050* 319 (6050): 199–203. https://doi.org/10.1038/319199a0.

Eiserling, F. A. 1967. “The Structure of Bacillus Subtilis Bacteriophage PBS 1.” *Journal of Ultrastructure Research* 17 (3–4): 342–47. https://doi.org/10.1016/S0022-5320(67)80053-4.

Emsley, P., K. Cowtan, and IUCr. 2004. “Coot: Model-Building Tools for Molecular Graphics.” *Urn:Issn:0907-4449* 60 (12): 2126–32. https://doi.org/10.1107/S0907444904019158.

Eriksson, A. E., W. A. Baase, and B. W. Matthews. 1993. “Similar Hydrophobic Replacements of Leu99 and Phe153 within the Core of T4 Lysozyme Have Different Structural and Thermodynamic Consequences.” *Journal of Molecular Biology* 229 (3): 747–69. https://doi.org/10.1006/JMBI.1993.1077.

“Essential Cell Biology - Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter - Google Books.” n.d. Accessed October 18, 2021. https://books.google.com/books?hl=en&lr=&id=Cg4WAgAAQBAJ&oi=fnd&pg=PP1&dq=1.%09Alberts,+B.,+Johnson,+A.,+Lewis,+J.,+Morgan,+D.,+Raff,+M.,+Roberts,+K.,+%26+Walter,+P.+(2015).+Molecular+biology+of+the+cell+Sixth+edition.+Garland+Science+Taylor+and+Francis+Group,+New+York+NY.&ots=yf5MaK07J1&sig=SHmtD-u34KU4G-gQBfxXlwNwsRk#v=onepage&q&f=false.

F, Arisaka, Engel J, and Klump H. 1981. “Contraction and Dissociation of the Bacteriophage T4 Tail Sheath Induced by Heat and Urea.” *Progress in Clinical and Biological Research* 64 (January): 365–79. https://europepmc.org/article/med/7330053.

Falk, Wayne, and Richard D. James. 2006. “Elasticity Theory for Self-Assembled Protein Lattices with Application to the Martensitic Phase Transition in Bacteriophage T4 Tail Sheath.” *Physical Review E* 73 (1): 011917. https://doi.org/10.1103/PhysRevE.73.011917.

Fang, Chengli, Lingting Li, Liqiang Shen, Jing Shi, Sheng Wang, Yu Feng, and Yu Zhang. 2019. “Structures and Mechanism of Transcription Initiation by Bacterial ECF Factors.” *Nucleic Acids Research* 47 (13): 7094–7104.

https://doi.org/10.1093/NAR/GKZ470.

Feklistov, Andrey, and Seth A. Darst. 2011. “Structural Basis for Promoter −10 Element Recognition by the Bacterial RNA Polymerase σ Subunit.” *Cell* 147 (6): 1257–69. https://doi.org/10.1016/J.CELL.2011.10.041.

Feklístov, Andrey, Brian D. Sharon, Seth A. Darst, and Carol A. Gross. 2014. “Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective.” *Http://Dx.Doi.Org/10.1146/Annurev-Micro-092412-155737* 68 (September): 357–76. https://doi.org/10.1146/ANNUREV-MICRO-092412-155737.

Feller, Scott E., Yuhong Zhang, Richard W. Pastor, and Bernard R. Brooks. 1998. “Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method.” *The Journal of Chemical Physics* 103 (11): 4613. https://doi.org/10.1063/1.470648.

Fenwick, R. Bryn, Henry van den Bedem, James S. Fraser, and Peter E. Wright. 2014. “Integrated Description of Protein Dynamics from Room-Temperature X-Ray Crystallography and NMR.” *Proceedings of the National Academy of Sciences* 111 (4): E445–54. https://doi.org/10.1073/PNAS.1323440111.

Ferrières, Lionel, Gaëlle Hémery, Toan Nham, Anne Marie Guérout, Didier Mazel, Christophe Beloin, and Jean Marc Ghigo. 2010. “Silent Mischief: Bacteriophage Mu Insertions Contaminate Products of Escherichia Coli Random Mutagenesis Performed Using Suicidal Transposon Delivery Plasmids Mobilized by Broad-Host-Range RP4 Conjugative Machinery.” *Journal of Bacteriology* 192 (24): 6418–27. https://doi.org/10.1128/JB.00621-10.

Fiorin, Giacomo, Michael L. Klein, and Jérôme Hénin. 2013. “Using Collective Variables to Drive Molecular Dynamics Simulations.” *Https://Doi.Org/10.1080/00268976.2013.813594* 111 (22–23): 3345–62. https://doi.org/10.1080/00268976.2013.813594.

Fischer, Emil. 1894. “Einfluss Der Configuration Auf Die Wirkung Der Enzyme.” *Berichte Der Deutschen Chemischen Gesellschaft* 27 (3): 2985–93. https://doi.org/10.1002/CBER.18940270364.

Fitzpatrick, Anthony W. P., Benjamin Falcon, Shaoda He, Alexey G. Murzin, Garib Murshudov, Holly J. Garringer, R. Anthony Crowther, Bernardino Ghetti, Michel Goedert, and Sjors H. W. Scheres. 2017. “Cryo-EM Structures of Tau Filaments from Alzheimer’s Disease.” *Nature 2017 547:7662* 547 (7662): 185–90. https://doi.org/10.1038/nature23002.

Fraser, Alec, Nikolai S. Prokhorov, Fang Jiao, B. Montgomery Pettitt, Simon Scheuring, and Petr G. Leiman. 2021. “Quantitative Description of a Contractile Macromolecular Machine.” *Science Advances* 7 (24): 9601–12. https://doi.org/10.1126/SCIADV.ABF9601.

Gao, Mu, David Craig, Viola Vogel, and Klaus Schulten. 2002. “Identifying Unfolding Intermediates of FN-III10 by Steered Molecular Dynamics.” *Journal of Molecular Biology* 323 (5): 939–50. https://doi.org/10.1016/S0022-2836(02)01001-X.

Garrido, Natàlia de Martín, Mariia Orekhova, Yuen Ting Emilie Lai Wan Loong, Anna Litvinova, Kailash Ramlaul, Tatyana Artamonova, Alexei S. Melnikov, Pavel Serdobintsev, Christopher H. S. Aylett, and Maria Yakunina. 2021. “Structure of the Bacteriophage PhiKZ Non-Virion RNA Polymerase.” *BioRxiv*, April, 2021.04.06.438582. https://doi.org/10.1101/2021.04.06.438582.

Ge, Peng, Dean Scholl, Petr G Leiman, Xuekui Yu, Jeff F Miller, and Z Hong Zhou. 2015. “Atomic Structures of a Bactericidal Contractile Nanotube in Its Pre- and Postcontraction States.” *Nature Structural & Molecular Biology 2015 22:5* 22 (5): 377–82. https://doi.org/10.1038/nsmb.2995.

Ge, Peng, Dean Scholl, Nikolai S. Prokhorov, Jaycob Avaylon, Mikhail M. Shneider, Christopher Browning, Sergey A. Buth, et al. 2020. “Action of a Minimal Contractile Bactericidal Nanomachine.” *Nature 2020 580:7805* 580 (7805): 658–62. https://doi.org/10.1038/s41586-020-2186-z.

Gibson, Q. H., and M. H. Smith. 1965. “Rates of Reaction of Ascaris Haemoglobins with Ligands.” *Proceedings of the Royal Society of London. Series B. Biological Sciences* 163 (991): 206–14. https://doi.org/10.1098/RSPB.1965.0067.

Gilson, Michael K., James A. Given, Bruce L. Bush, and J. Andrew McCammon. 1997. “The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review.” *Biophysical Journal* 72 (3): 1047–69. https://doi.org/10.1016/S0006-3495(97)78756-3.

Goddard, Thomas D., Conrad C. Huang, Elaine C. Meng, Eric F. Pettersen, Gregory S. Couch, John H. Morris, and Thomas E. Ferrin. 2018. “UCSF ChimeraX: Meeting Modern Challenges in Visualization and Analysis.” *Protein Science* 27 (1): 14–25. https://doi.org/10.1002/PRO.3235.

Gohlke, Holger, and David A. Case. 2004. “Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein–Protein Complex Ras–Raf.” *Journal of Computational Chemistry* 25 (2): 238–50. https://doi.org/10.1002/JCC.10379.

Guerrero-Ferreira, Ricardo C, Mario Hupfeld, Sergey Nazarov, Nicholas MI Taylor, Mikhail M Shneider, Jagan M Obbineni, Martin J Loessner, Takashi Ishikawa, Jochen Klumpp, and Petr G Leiman. 2019. “Structure and Transformation of Bacteriophage A511 Baseplate and Tail upon Infection of Listeria Cells.” *The EMBO Journal* 38 (3): e99455. https://doi.org/10.15252/EMBJ.201899455.

Gumbart Benoˆıt, James, Benoˆıt Roux, and Christophe Chipot. 2018. “Protein:Ligand

Standard Binding Free Energies: A Tutorial for Alchemical and Geometrical Transformations." www.ks.uiuc.edu/Training/Tutorials/.

Gumbart, James C., Benoît Roux, and Christophe Chipot. 2012. "Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy?" *Journal of Chemical Theory and Computation* 9 (1): 794–802. https://doi.org/10.1021/CT3008099.

Harrach, Michael F., and Barbara Drossel. 2014. "Structure and Dynamics of TIP3P, TIP4P, and TIP5P Water near Smooth and Atomistic Walls of Different Hydroaffinity." *The Journal of Chemical Physics* 140 (17): 174501. https://doi.org/10.1063/1.4872239.

Harris, Robert C., Justin A. Drake, and B. Montgomery Pettitt. 2014. "Multibody Correlations in the Hydrophobic Solvation of Glycine Peptides." *The Journal of Chemical Physics* 141 (22): 22D525. https://doi.org/10.1063/1.4901886.

Harris, Robert C., and B. Montgomery Pettitt. 2014. "Effects of Geometry and Chemistry on Hydrophobic Solvation." *Proceedings of the National Academy of Sciences* 111 (41): 14681–86. https://doi.org/10.1073/PNAS.1406080111.

Hénin, Jérôme, and Christophe Chipot. 2004. "Overcoming Free Energy Barriers Using Unconstrained Molecular Dynamics Simulations." *The Journal of Chemical Physics* 121 (7): 2904. https://doi.org/10.1063/1.1773132.

Hobb, Rhonda I., Joshua A. Fields, Christopher M. Burns, and Stuart A. Thompson. 2009. "Evaluation of Procedures for Outer Membrane Isolation from Campylobacter Jejuni." *Microbiology (Reading, England)* 155 (Pt 3): 979. https://doi.org/10.1099/MIC.0.024539-0.

Holm, Liisa, and Laura M. Laakso. 2016. "Dali Server Update." *Nucleic Acids Research* 44 (W1): W351–55. https://doi.org/10.1093/NAR/GKW357.

Hospital, Adam, Josep Ramon Goñi, Modesto Orozco, and Josep L Gelpí. 2015. "Molecular Dynamics Simulations: Advances and Applications." *Advances and Applications in Bioinformatics and Chemistry : AABC* 8 (1): 37. https://doi.org/10.2147/AABC.S70333.

Hou, Tingjun, Junmei Wang, Youyong Li, and Wei Wang. 2010. "Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations." *Journal of Chemical Information and Modeling* 51 (1): 69–82. https://doi.org/10.1021/CI100275A.

Hu, Bo, William Margolin, Ian J. Molineux, and Jun Liu. 2015. "Structural Remodeling of Bacteriophage T4 and Host Membranes during Infection Initiation." *Proceedings*

of the National Academy of Sciences 112 (35): E4919–28. https://doi.org/10.1073/PNAS.1501064112.

Huang, Jing, and Alexander D. MacKerell. 2013. "CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data." *Journal of Computational Chemistry* 34 (25): 2135–45. https://doi.org/10.1002/JCC.23354.

Humphrey, William, Andrew Dalke, and Klaus Schulten. 1996. "VMD: Visual Molecular Dynamics." *Journal of Molecular Graphics* 14 (1): 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

Ishima, Rieko, and Dennis A. Torchia. 2000. "Protein Dynamics from NMR." *Nature Structural Biology 2000 7:9* 7 (9): 740–43. https://doi.org/10.1038/78963.

Isralewitz, B., M. Gao, and K. Schulten. 2001. "Steered Molecular Dynamics and Mechanical Functions of Proteins." *Current Opinion in Structural Biology* 11 (2): 224–30. https://doi.org/10.1016/S0959-440X(00)00194-9.

Jensen, Malene Ringkjøbing, Rob W.H. Ruigrok, and Martin Blackledge. 2013. "Describing Intrinsically Disordered Proteins at Atomic Resolution by NMR." *Current Opinion in Structural Biology* 23 (3): 426–35. https://doi.org/10.1016/J.SBI.2013.02.007.

Jensen, Malene Ringkjøbing, Markus Zweckstetter, Jie-rong Huang, and Martin Blackledge. 2014. "Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy." *Chemical Reviews* 114 (13): 6632–60. https://doi.org/10.1021/CR400688U.

Jiang, Feng, Ningning Li, Xia Wang, Jiaxuan Cheng, Yaoguang Huang, Yun Yang, Jianguo Yang, et al. 2019. "Cryo-EM Structure and Assembly of an Extracellular Contractile Injection System." *Cell* 177 (2): 370-383.e15. https://doi.org/10.1016/J.CELL.2019.02.020.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature 2021 596:7873* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Karush, Fred. 2002. "Heterogeneity of the Binding Sites of Bovine Serum Albumin1." *Journal of the American Chemical Society* 72 (6): 2705–13. https://doi.org/10.1021/JA01162A099.

Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. 1958. "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis." *Nature 1958 181:4610* 181 (4610): 662–66. https://doi.org/10.1038/181662a0.

Kim, Jonathan, Yong Zi Tan, Kathryn J. Wicht, Satchal K. Erramilli, Satish K. Dhingra, John Okombo, Jeremie Vendome, et al. 2019. “Structure and Drug Resistance of the Plasmodium Falciparum Transporter PfCRT.” *Nature 2019 576:7786* 576 (7786): 315–20. https://doi.org/10.1038/s41586-019-1795-x.

Konrat, Robert. 2014. “NMR Contributions to Structural Dynamics Studies of Intrinsically Disordered Proteins.” *Journal of Magnetic Resonance* 241 (1): 74–85. https://doi.org/10.1016/J.JMR.2013.11.011.

Korn, Abby M., Andrew E. Hillhouse, Lichang Sun, and Jason J. Gill. 2021. “Comparative Genomics of Three Novel Jumbo Bacteriophages Infecting Staphylococcus Aureus.” *Journal of Virology* 95 (19). https://doi.org/10.1128/JVI.02391-20.

Korzhnev, Dmitry M., Xavier Salvatella, Michele Vendruscolo, Ariel A. Di Nardo, Alan R. Davidson, Christopher M. Dobson, and Lewis E. Kay. 2004. “Low-Populated Folding Intermediates of Fyn SH3 Characterized by Relaxation Dispersion NMR.” *Nature 2004 430:6999* 430 (6999): 586–90. https://doi.org/10.1038/nature02655.

Koshland, D. E., and Jr. 1958. “Application of a Theory of Enzyme Specificity to Protein Synthesis.” *Proceedings of the National Academy of Sciences of the United States of America* 44 (2): 98. https://doi.org/10.1073/PNAS.44.2.98.

Krissinel, Evgeny, and Kim Henrick. 2007. “Inference of Macromolecular Assemblies from Crystalline State.” *Journal of Molecular Biology* 372 (3): 774–97. https://doi.org/10.1016/J.JMB.2007.05.022.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. n.d. “ImageNet Classification with Deep Convolutional Neural Networks.” Accessed October 18, 2021. http://code.google.com/p/cuda-convnet/.

Kropinski, Andrew M., Amanda Mazzocco, Thomas E. Waddell, Erika Lingohr, and Roger P. Johnson. 2009. “Enumeration of Bacteriophages by Double Agar Overlay Plaque Assay.” *Methods in Molecular Biology (Clifton, N.J.)* 501: 69–76. https://doi.org/10.1007/978-1-60327-164-6_7.

Kryshtafovych, Andriy, John Moult, Reinhard Albrecht, Geoffrey A. Chang, Kinlin Chao, Alec Fraser, Julia Greenfield, et al. 2021. “Computational Models in the Service of X-Ray and Cryo-Electron Microscopy Structure Determination.” *Proteins: Structure, Function and Bioinformatics*. https://doi.org/10.1002/PROT.26223.

Kschonsak, Marc, Han Chow Chua, Cameron L. Noland, Claudia Weidling, Thomas Clairfeuille, Oskar Ørts Bahlke, Aishat Oluwanifemi Ameen, et al. 2020. “Structure of the Human Sodium Leak Channel NALCN.” *Nature 2020 587:7833* 587 (7833):

313–18. https://doi.org/10.1038/s41586-020-2570-8.

Kudryashev, Mikhail, Ray Yu Ruei Wang, Maximilian Brackmann, Sebastian Scherer, Timm Maier, David Baker, Frank Dimaio, Henning Stahlberg, Edward H. Egelman, and Marek Basler. 2015. "Structure of the Type VI Secretion System Contractile Sheath." *Cell* 160 (5): 952–62. https://doi.org/10.1016/J.CELL.2015.01.037.

Kühlbrandt, Werner. 2014. "The Resolution Revolution." *Science* 343 (6178): 1443–44. https://doi.org/10.1126/SCIENCE.1251652.

Kuznedelov, Konstantin, Nataliya Korzheva, Arkady Mustaev, and Konstantin Severinov. 2002. "Structure-Based Analysis of RNA Polymerase Function: The Largest Subunit's Rudder Contributes Critically to Elongation Complex Stability and Is Not Involved in the Maintenance of RNA–DNA Hybrid Length." *The EMBO Journal* 21 (6): 1369–78. https://doi.org/10.1093/EMBOJ/21.6.1369.

Lane, William J., and Seth A. Darst. 2010. "Molecular Evolution of Multisubunit RNA Polymerases: Structural Analysis." *Journal of Molecular Biology* 395 (4): 686–704. https://doi.org/10.1016/J.JMB.2009.10.063.

Laverty, Duncan, Rooma Desai, Tomasz Uchański, Simonas Masiulis, Wojciech J. Stec, Tomas Malinauskas, Jasenko Zivanov, et al. 2019. "Cryo-EM Structure of the Human A1β3γ2 GABAA Receptor in a Lipid Bilayer." *Nature 2019 565:7740* 565 (7740): 516–20. https://doi.org/10.1038/s41586-018-0833-4.

Lavysh, Daria, Maria Sokolova, Leonid Minakhin, Maria Yakunina, Tatjana Artamonova, Sergei Kozyavkin, Kira S. Makarova, Eugene V. Koonin, and Konstantin Severinov. 2016. "The Genome of AR9, a Giant Transducing Bacillus Phage Encoding Two Multisubunit RNA Polymerases." *Virology* 495 (August): 185–96. https://doi.org/10.1016/J.VIROL.2016.04.030.

Lavysh, Daria, Maria Sokolova, Marina Slashcheva, Konrad U. Förstner, and Konstantin Severinov. 2017. "Transcription Profiling of Bacillus Subtilis Cells Infected with AR9, a Giant Phage Encoding Two Multisubunit RNA Polymerases." *MBio* 8 (1). https://doi.org/10.1128/MBIO.02041-16.

Lee, Jookyung, and Sergei Borukhov. 2016. "Bacterial RNA Polymerase-DNA Interaction—The Driving Force of Gene Expression and the Target for Drug Action." *Frontiers in Molecular Biosciences* 0 (NOV): 73. https://doi.org/10.3389/FMOLB.2016.00073.

Leiman, Petr G., Paul R. Chipman, Victor A. Kostyuchenko, Vadim V. Mesyanzhinov, and Michael G. Rossmann. 2004. "Three-Dimensional Rearrangement of Proteins in the Tail of Bacteriophage T4 on Infection of Its Host." *Cell* 118 (4): 419–29. https://doi.org/10.1016/J.CELL.2004.07.022.

Leiman, Petr G., and Mikhail M. Shneider. 2012. “Contractile Tail Machines of Bacteriophages.” *Advances in Experimental Medicine and Biology* 726: 93–114. https://doi.org/10.1007/978-1-4614-0980-9_5.

Li, Lingting, Chengli Fang, Ningning Zhuang, Tiantian Wang, and Yu Zhang. 2019. “Structural Basis for Transcription Initiation by Bacterial ECF σ Factors.” *Nature Communications 2019 10:1* 10 (1): 1–14. https://doi.org/10.1038/s41467-019-09096-y.

Liu, Bin, Yuhong Zuo, and Thomas A. Steitz. 2016. “Structures of E. Coli ΣS-Transcription Initiation Complexes Provide New Insights into Polymerase Mechanism.” *Proceedings of the National Academy of Sciences* 113 (15): 4051–56. https://doi.org/10.1073/PNAS.1520555113.

Liu, Jun, Cheng Yen Chen, Daisuke Shiomi, Hironori Niki, and William Margolin. 2011. “Visualization of Bacteriophage P1 Infection by Cryo-Electron Tomography of Tiny Escherichia Coli.” *Virology* 417 (2): 304–11. https://doi.org/10.1016/J.VIROL.2011.06.005.

Lumry, Rufus, Henry Eyeing, and Vol 58. n.d. “CONFORMATION CHANGES OF PROTEINS1.” Accessed October 18, 2021. https://pubs.acs.org/sharingguidelines.

Maghsoodi, Ameneh, Anupam Chatterjee, Ioan Andricioaei, and Noel C. Perkins. 2017. “Dynamic Model Exposes the Energetics and Dynamics of the Injection Machinery for Bacteriophage T4.” *Biophysical Journal* 113 (1): 195–205. https://doi.org/10.1016/J.BPJ.2017.05.029.

———. 2019. “How the Phage T4 Injection Machinery Works Including Energetics, Forces, and Dynamic Pathway.” *Proceedings of the National Academy of Sciences* 116 (50): 25097–105. https://doi.org/10.1073/PNAS.1909298116.

Mariani, Valerio, Marco Biasini, Alessandro Barbato, and Torsten Schwede. 2013. “LDDT: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests.” *Bioinformatics* 29 (21): 2722–28. https://doi.org/10.1093/BIOINFORMATICS/BTT473.

Martin, Hugh S. C., Shantenu Jha, and Peter V. Coveney. 2014. “Comparative Analysis of Nucleotide Translocation through Protein Nanopores Using Steered Molecular Dynamics and an Adaptive Biasing Force.” *Journal of Computational Chemistry* 35 (9): 692–702. https://doi.org/10.1002/JCC.23525.

Matsumoto, Shigeyuki, Shoichi Ishida, Mitsugu Araki, Takayuki Kato, Kei Terayama, and Yasushi Okuno. 2021. “Extraction of Protein Dynamics Information from Cryo-EM Maps Using Deep Learning.” *Nature Machine Intelligence 2021 3:2* 3 (2): 153–60. https://doi.org/10.1038/s42256-020-00290-y.

McCoy, A.J., R.W. Grosse-Kunstleve, P.D. Adams, M.D. Winn, L.C. Storoni, R.J. Read,

and IUCr. 2007. “Phaser Crystallographic Software.” *Urn:Issn:0021-8898* 40 (4): 658–74. https://doi.org/10.1107/S0021889807021206.

McGreevy, Ryan, Ivan Teo, Abhishek Singharoy, and Klaus Schulten. 2016. “Advances in the Molecular Dynamics Flexible Fitting Method for Cryo-EM Modeling.” *Methods* 100 (May): 50–60. https://doi.org/10.1016/J.YMETH.2016.01.009.

Mesyanzhinov, Vadim V., Johan Robben, Barbara Grymonprez, Victor A. Kostyuchenko, Maria V. Bourkaltseva, Nina N. Sykilinda, Victor N. Krylov, and Guido Volckaert. 2002. “The Genome of Bacteriophage ΦKZ of Pseudomonas Aeruginosa.” *Journal of Molecular Biology* 317 (1): 1–19. https://doi.org/10.1006/JMBI.2001.5396.

Metcalf, William W., Weihong Jiang, Larry L. Daniels, Soo Ki Kim, Andreas Haldimann, and Barry L. Wanner. 1996. “Conditionally Replicative and Conjugative Plasmids CarryinglacZα for Cloning, Mutagenesis, and Allele Replacement in Bacteria.” *Plasmid* 35 (1): 1–13. https://doi.org/10.1006/PLAS.1996.0001.

Minakhin, Leonid, Sechal Bhagat, Adrian Brunning, Elizabeth A. Campbell, Seth A. Darst, Richard H. Ebright, and Konstantin Severinov. 2001. “Bacterial RNA Polymerase Subunit ω and Eukaryotic RNA Polymerase Subunit RPB6 Are Sequence, Structural, and Functional Homologs and Promote RNA Polymerase Assembly.” *Proceedings of the National Academy of Sciences* 98 (3): 892–97. https://doi.org/10.1073/PNAS.98.3.892.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. “Playing Atari with Deep Reinforcement Learning,” December. https://arxiv.org/abs/1312.5602v1.

Moody, M. F. 1973. “Sheath of Bacteriophage T4: III. Contraction Mechanism Deduced from Partially Contracted Sheaths.” *Journal of Molecular Biology* 80 (4): 613–35. https://doi.org/10.1016/0022-2836(73)90200-3.

Mulder, Frans A.A., Anthony Mittermaier, Bin Hon, Frederick W. Dahlquist, and Lewis E. Kay. 2001. “Studying Excited States of Proteins by NMR Spectroscopy.” *Nature Structural Biology 2001 8:11* 8 (11): 932–35. https://doi.org/10.1038/nsb1101-932.

Müller-Späth, Sonja, Andrea Soranno, Verena Hirschfeld, Hagen Hofmann, Stefan Rüegger, Luc Reymond, Daniel Nettels, and Benjamin Schuler. 2010. “Charge Interactions Can Dominate the Dimensions of Intrinsically Disordered Proteins.” *Proceedings of the National Academy of Sciences* 107 (33): 14609–14. https://doi.org/10.1073/PNAS.1001743107.

Nakane, Takanori, Abhay Kotecha, Andrija Sente, Greg McMullan, Simonas Masiulis, Patricia M. G. E. Brown, Ioana T. Grigoras, et al. 2020. “Single-Particle Cryo-EM at

Atomic Resolution." *Nature 2020 587:7832* 587 (7832): 152–56. https://doi.org/10.1038/s41586-020-2829-0.

Narayanan, Anoop, Frank S. Vago, Kunpeng Li, M. Zuhaib Qayyum, Dinesh Yernool, Wen Jiang, and Katsuhiko S. Murakami. 2018. "Cryo-EM Structure of Escherichia Coli Σ70 RNA Polymerase and Promoter DNA Complex Revealed a Role of σ Non-Conserved Region during the Open Complex Formation." *Journal of Biological Chemistry* 293 (19): 7367–75. https://doi.org/10.1074/JBC.RA118.002161.

Nazarov, Sergey, Johannes P Schneider, Maximilian Brackmann, Kenneth N Goldie, Henning Stahlberg, and Marek Basler. 2018. "Cryo-EM Reconstruction of Type VI Secretion System Baseplate and Sheath Distal End." *The EMBO Journal* 37 (4): e97103. https://doi.org/10.15252/EMBJ.201797103.

Oberhauser, Andres F., Carmelu Badilla-Fernandez, Mariano Carrion-Vazquez, and Julio M. Fernandez. 2002. "The Mechanical Hierarchies of Fibronectin Observed with Single-Molecule AFM." *Journal of Molecular Biology* 319 (2): 433–47. https://doi.org/10.1016/S0022-2836(02)00306-6.

Ogata, Hideaki, Koji Nishikawa, and Wolfgang Lubitz. 2015. "Hydrogens Detected by Subatomic Resolution Protein Crystallography in a [NiFe] Hydrogenase." *Nature 2015 520:7548* 520 (7548): 571–74. https://doi.org/10.1038/nature14110.

Paget, Mark S. 2015. "Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution." *Biomolecules 2015, Vol. 5, Pages 1245-1265* 5 (3): 1245–65. https://doi.org/10.3390/BIOM5031245.

Palencia, Andrés, Eva S. Cobos, Pedro L. Mateo, Jose C. Martínez, and Irene Luque. 2004. "Thermodynamic Dissection of the Binding Energetics of Proline-Rich Peptides to the Abl-SH3 Domain: Implications for Rational Ligand Design." *Journal of Molecular Biology* 336 (2): 527–37. https://doi.org/10.1016/J.JMB.2003.12.030.

Pandey, Suraj, Richard Bean, Tokushi Sato, Ishwor Poudyal, Johan Bielecki, Jorvani Cruz Villarreal, Oleksandr Yefanov, et al. 2019. "Time-Resolved Serial Femtosecond Crystallography at the European XFEL." *Nature Methods 2019 17:1* 17 (1): 73–78. https://doi.org/10.1038/s41592-019-0628-z.

Park, Sanghyun, Fatemeh Khalili-Araghi, Emad Tajkhorshid, and Klaus Schulten. 2003. "Free Energy Calculation from Steered Molecular Dynamics Simulations Using Jarzynski's Equality." *The Journal of Chemical Physics* 119 (6): 3559. https://doi.org/10.1063/1.1590311.

Park, Sanghyun, and Klaus Schulten. 2004. "Calculating Potentials of Mean Force from Steered Molecular Dynamics Simulations." *The Journal of Chemical Physics* 120 (13): 5946. https://doi.org/10.1063/1.1651473.

Park, Young-Jun, Kaitlyn D. Lacourse, Christian Cambillau, Frank DiMaio, Joseph D. Mougous, and David Veesler. 2018. “Structure of the Type VI Secretion System TssK–TssF–TssG Baseplate Subcomplex Revealed by Cryo-Electron Microscopy.” *Nature Communications 2018 9:1* 9 (1): 1–11. https://doi.org/10.1038/s41467-018-07796-5.

Patz, Sascha, Yvonne Becker, Katja R. Richert-Pöggeler, Beatrice Berger, Silke Ruppel, Daniel H. Huson, and Matthias Becker. 2019. “Phage Tail-like Particles Are Versatile Bacterial Nanomachines – A Mini-Review.” *Journal of Advanced Research* 19 (September): 75–84. https://doi.org/10.1016/J.JARE.2019.04.003.

Pereira, Joana, Adam J. Simpkin, Marcus D. Hartmann, Daniel J. Rigden, Ronan M. Keegan, and Andrei N. Lupas. 2021. “High-Accuracy Protein Structure Prediction in CASP14.” *Proteins: Structure, Function, and Bioinformatics*. https://doi.org/10.1002/PROT.26171.

Perutz, M. F., and F. S. Mathews. 1966. “An X-Ray Study of Azide Methaemoglobin.” *Journal of Molecular Biology* 21 (1): 199–202. https://doi.org/10.1016/0022-2836(66)90088-X.

Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. 2004. “UCSF Chimera—A Visualization System for Exploratory Research and Analysis.” *Journal of Computational Chemistry* 25 (13): 1605–12. https://doi.org/10.1002/JCC.20084.

Phillips, James C., Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. 2005. “Scalable Molecular Dynamics with NAMD.” *Journal of Computational Chemistry* 26 (16): 1781–1802. https://doi.org/10.1002/JCC.20289.

Pisabarro, M. Teresa, Luis Serrano, and Matthias Wilmanns. 1998. “Crystal Structure of the Abl-SH3 Domain Complexed with a Designed High-Affinity Peptide Ligand: Implications for SH3-Ligand Interactions.” *Journal of Molecular Biology* 281 (3): 513–21. https://doi.org/10.1006/JMBI.1998.1932.

Pohorille, Andrew, Christopher Jarzynski, and Christophe Chipot. 2010. “Good Practices in Free-Energy Calculations.” *Journal of Physical Chemistry B* 114 (32): 10235–53. https://doi.org/10.1021/JP102971X.

Prokhorov, Nikolai S., Cristian Riccio, Evelina L. Zdorovenko, Mikhail M. Shneider, Christopher Browning, Yuriy A. Knirel, Petr G. Leiman, and Andrey V. Letarov. 2017. “Function of Bacteriophage G7C Esterase Tailspike in Host Cell Adsorption.” *Molecular Microbiology* 105 (3): 385–98. https://doi.org/10.1111/MMI.13710.

Punjani, Ali, John L Rubinstein, David J Fleet, and Marcus A Brubaker. 2017.

"CryoSPARC: Algorithms for Rapid Unsupervised Cryo-EM Structure Determination." *Nature Methods 2017 14:3* 14 (3): 290–96. https://doi.org/10.1038/nmeth.4169.

Punjani, Ali, Haowei Zhang, and David J. Fleet. 2020. "Non-Uniform Refinement: Adaptive Regularization Improves Single-Particle Cryo-EM Reconstruction." *Nature Methods 2020 17:12* 17 (12): 1214–21. https://doi.org/10.1038/s41592-020-00990-8.

Rajamani, Sowmianarayanan, Thomas M. Truskett, and Shekhar Garde. 2005. "Hydrophobic Hydration from Small to Large Lengthscales: Understanding and Manipulating the Crossover." *Proceedings of the National Academy of Sciences* 102 (27): 9475–80. https://doi.org/10.1073/PNAS.0504089102.

Ratner, Mark A., Ratner, and Mark A. 1997. "Understanding Molecular Simulation: From Algorithms to Applications, by Daan Frenkel and Berend Smit." *PhT* 50 (7): 66. https://ui.adsabs.harvard.edu/abs/1997PhT....50g..66R/abstract.

Ritchie, Jennifer M., Jennifer L. Greenwich, Brigid M. Davis, Roderick T. Bronson, Dana Gebhart, Steven R. Williams, David Martin, Dean Scholl, and Matthew K. Waldor. 2011. "An Escherichia Coli O157-Specific Engineered Pyocin Prevents and Ameliorates Infection by E. Coli O157:H7 in an Animal Model of Diarrheal Disease." *Antimicrobial Agents and Chemotherapy* 55 (12): 5469–74. https://doi.org/10.1128/AAC.05031-11.

Rohou, Alexis, and Nikolaus Grigorieff. 2015. "CTFFIND4: Fast and Accurate Defocus Estimation from Electron Micrographs." *Journal of Structural Biology* 192 (2): 216–21. https://doi.org/10.1016/J.JSB.2015.08.008.

Rohs, Remo, Xiangshu Jin, Sean M. West, Rohit Joshi, Barry Honig, and Richard S. Mann. 2010. "Origins of Specificity in Protein-DNA Recognition." *Http://Dx.Doi.Org/10.1146/Annurev-Biochem-060408-091030* 79 (June): 233–69. https://doi.org/10.1146/ANNUREV-BIOCHEM-060408-091030.

Roux, Benoît, Mafalda Nina, Régis Pomès, and Jeremy C. Smith. 1996. "Thermodynamic Stability of Water Molecules in the Bacteriorhodopsin Proton Channel: A Molecular Dynamics Free Energy Perturbation Study." *Biophysical Journal* 71 (2): 670–81. https://doi.org/10.1016/S0006-3495(96)79267-6.

Santiveri, Mònica, Aritz Roa-Eguiara, Caroline Kühne, Navish Wadhwa, Haidai Hu, Howard C. Berg, Marc Erhardt, and Nicholas M.I. Taylor. 2020. "Structure and Function of Stator Units of the Bacterial Flagellar Motor." *Cell* 183 (1): 244-257.e16. https://doi.org/10.1016/J.CELL.2020.08.016.

Schlichting, I. 2015. "Serial Femtosecond Crystallography: The First Five Years." *Urn:Issn:2052-2525* 2 (2): 246–55. https://doi.org/10.1107/S205225251402702X.

Scholl, Dean, and David W. Martin. 2008. “Antibacterial Efficacy of R-Type Pyocins towards Pseudomonas Aeruginosa in a Murine Peritonitis Model.” *Antimicrobial Agents and Chemotherapy* 52 (5): 1647–52. https://doi.org/10.1128/AAC.01479-07.

Schotte, Friedrich, Manho Lim, Timothy A. Jackson, Aleksandr V. Smirnov, Jayashree Soman, John S. Olson, Jr. George N. Phillips, Michael Wulff, and Philip A. Anfinrud. 2003. “Watching a Protein as It Functions with 150-Ps Time-Resolved X-Ray Crystallography.” *Science* 300 (5627): 1944–47. https://doi.org/10.1126/SCIENCE.1078797.

Schwalbe, Martin, Valéry Ozenne, Stefan Bibow, Mariusz Jaremko, Lukasz Jaremko, Michal Gajda, Malene Ringkjøbing Jensen, et al. 2014. “Predictive Atomic Resolution Descriptions of Intrinsically Disordered HTau40 and α-Synuclein in Solution from NMR and Small Angle Scattering.” *Structure* 22 (2): 238–49. https://doi.org/10.1016/J.STR.2013.10.020.

Shimada, Atsuhiro, Minoru Kubo, Seiki Baba, Keitaro Yamashita, Kunio Hirata, Go Ueno, Takashi Nomura, et al. 2017. “A Nanosecond Time-Resolved XFEL Analysis of Structural Changes Associated with CO Release from Cytochrome c Oxidase.” *Science Advances* 3 (7). https://doi.org/10.1126/SCIADV.1603042.

Shneider, Mikhail M., Sergey A. Buth, Brian T. Ho, Marek Basler, John J. Mekalanos, and Petr G. Leiman. 2013. “PAAR-Repeat Proteins Sharpen and Diversify the Type VI Secretion System Spike.” *Nature 2013 500:7462* 500 (7462): 350–53. https://doi.org/10.1038/nature12453.

Sickmeier, Megan, Justin A. Hamilton, Tanguy LeGall, Vladimir Vacic, Marc S. Cortese, Agnes Tantos, Beata Szabo, et al. 2007. “DisProt: The Database of Disordered Proteins.” *Nucleic Acids Research* 35 (suppl_1): D786–93. https://doi.org/10.1093/NAR/GKL893.

Skurnik, M., H. J. Hyytiainen, L. J. Happonen, S. Kiljunen, N. Datta, L. Mattinen, K. Williamson, et al. 2012. “Characterization of the Genome, Proteome, and Structure of Yersiniophage R1-37.” *Journal of Virology* 86 (23): 12625–42. https://doi.org/10.1128/JVI.01783-12.

Sokolova, Maria, Sergei Borukhov, Daria Lavysh, Tatjana Artamonova, Mikhail Khodorkovskii, and Konstantin Severinov. 2017. “A Non-Canonical Multisubunit RNA Polymerase Encoded by the AR9 Phage Recognizes the Template Strand of Its Uracil-Containing Promoters.” *Nucleic Acids Research* 45 (10): 5958–67. https://doi.org/10.1093/NAR/GKX264.

Sokolova, Maria L., Inna Misovetc, and Konstantin V. Severinov. 2020. “Multisubunit RNA Polymerases of Jumbo Bacteriophages.” *Viruses 2020, Vol. 12, Page 1064* 12 (10): 1064. https://doi.org/10.3390/V12101064.

Spoel*, David van der, and Erik Lindahl. 2003. “Brute-Force Molecular Dynamics Simulations of Villin Headpiece: Comparison with NMR Parameters.” *Journal of Physical Chemistry B* 107 (40): 11178–87. https://doi.org/10.1021/JP034108N.

Sterman, M D, J F Foster, Melvin D Sterman, and Joseph F Foster. n.d. “Vol. 78 Conformation Changes in Bovine Plasma Albumin Associated with Hydrogen Ion and Urea Binding. I. Intrinsic Viscosity and Optical Rotation1,2.” Accessed October 18, 2021. https://pubs.acs.org/sharingguidelines.

Stietz, Maria Silvina, Xiaoye Liang, Megan Wong, Steven Hersch, and Tao G. Dong. 2020. “Double Tubular Contractile Structure of the Type VI Secretion System Displays Striking Flexibility and Elasticity.” *Journal of Bacteriology* 202 (1). https://doi.org/10.1128/JB.00425-19.

“Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and ... - Alan Fersht, University Alan Fersht - Google Books.” n.d. Accessed October 18, 2021. https://books.google.com/books?hl=en&lr=&id=QdpZz_ahA5UC&oi=fnd&pg=PR20&dq=10.%09Fersht,+A.+(1999).+Structure+and+mechanism+in+protein+science:+a+guide+to+enzyme+catalysis+and+protein+folding.+Macmillan.&ots=SbqBVKhY1r&sig=dOGN2oZgsvHXm7rlGLYULQ4VJS4#v=onepage&q=10.%09Fersht%2C A. (1999). Structure and mechanism in protein science%3A a guide to enzyme catalysis and protein folding. Macmillan.&f=false.

Tama, Florence, Osamu Miyashita, and Charles L. Brooks. 2004. “Normal Mode Based Flexible Fitting of High-Resolution Structure into Low-Resolution Experimental Data from Cryo-EM.” *Journal of Structural Biology* 147 (3): 315–26. https://doi.org/10.1016/J.JSB.2004.03.002.

Tang, Guang, Liwei Peng, Philip R. Baldwin, Deepinder S. Mann, Wen Jiang, Ian Rees, and Steven J. Ludtke. 2007. “EMAN2: An Extensible Image Processing Suite for Electron Microscopy.” *Journal of Structural Biology* 157 (1): 38–46. https://doi.org/10.1016/J.JSB.2006.05.009.

Taylor, Nicholas M. I., Nikolai S. Prokhorov, Ricardo C. Guerrero-Ferreira, Mikhail M. Shneider, Christopher Browning, Kenneth N. Goldie, Henning Stahlberg, and Petr G. Leiman. 2016. “Structure of the T4 Baseplate and Its Function in Triggering Sheath Contraction.” *Nature 2016 533:7603* 533 (7603): 346–52. https://doi.org/10.1038/nature17971.

Taylor, Nicholas M. I., Mark J. van Raaij, and Petr G. Leiman. 2018. “Contractile Injection Systems of Bacteriophages and Related Systems.” *Molecular Microbiology* 108 (1): 6–15. https://doi.org/10.1111/MMI.13921.

“Three-Dimensional Electron Microscopy of Macromolecular Assemblies ... - Joachim

Frank - Google Books.” n.d. Accessed October 18, 2021. https://books.google.com/books?hl=en&lr=&id=kjGKy2LeWnUC&oi=fnd&pg=PR9&dq=66.%09Frank,+J.+(2006).+Three-dimensional+electron+microscopy+of+macromolecular+assemblies:+visualization+of+biological+molecules+in+their+native+state.+Oxford+university+press.&ots=kNocXeK1oa&sig=BcqgjiWwLJ3v7WPvWvuKIHrmEAk#v=onepage&q&f=false.

Torrie, G. M., and J. P. Valleau. 1977. “Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling.” *Journal of Computational Physics* 23 (2): 187–99. https://doi.org/10.1016/0021-9991(77)90121-8.

Tosha, Takehiko, Takashi Nomura, Takuma Nishida, Naoya Saeki, Kouta Okubayashi, Raika Yamagiwa, Michihiro Sugahara, et al. 2017. “Capturing an Initial Intermediate during the P450nor Enzymatic Reaction Using Time-Resolved XFEL Crystallography and Caged-Substrate.” *Nature Communications 2017 8:1* 8 (1): 1–9. https://doi.org/10.1038/s41467-017-01702-1.

Uversky, Vladimir N., Vrushank Davé, Lilia M. Iakoucheva, Prerna Malaney, Steven J. Metallo, Ravi Ramesh Pathak, and Andreas C. Joerger. 2014. “Pathological Unfoldomics of Uncontrolled Chaos: Intrinsically Disordered Proteins and Human Diseases.” *Chemical Reviews* 114 (13): 6844–79. https://doi.org/10.1021/CR400713R.

Uversky, Vladimir N., Christopher J. Oldfield, and A. Keith Dunker. 2008. “Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept.” *Http://Dx.Doi.Org/10.1146/Annurev.Biophys.37.032807.125924* 37 (May): 215–46. https://doi.org/10.1146/ANNUREV.BIOPHYS.37.032807.125924.

Vettiger, Andrea, Julius Winter, Lin Lin, and Marek Basler. 2017. “The Type VI Secretion System Sheath Assembles at the End Distal from the Membrane Anchor.” *Nature Communications 2017 8:1* 8 (1): 1–9. https://doi.org/10.1038/ncomms16088.

Vonrhein, Clemens, Eric Blanc, Pietro Roversi, and Gérard Bricogne. n.d. “Automated Structure Solution With AutoSHARP.” *From: Methods in Molecular Biology* 364. Accessed October 18, 2021. http://www.globalphasing.com/sharp/.

WADDINGTON, C. H. 1961. “Molecular Biology or Ultrastructural Biology?” *Nature 1961 190:4781* 190 (4781): 1124–25. https://doi.org/10.1038/1901124b0.

Wang, Jing, Maximilian Brackmann, Daniel Castaño-Díez, Mikhail Kudryashev, Kenneth N. Goldie, Timm Maier, Henning Stahlberg, and Marek Basler. 2017. “Cryo-EM Structure of the Extended Type VI Secretion System Sheath–Tube Complex.” *Nature Microbiology 2017 2:11* 2 (11): 1507–12. https://doi.org/10.1038/s41564-017-0020-7.

Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones. 2004. “Prediction

and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life." *Journal of Molecular Biology* 337 (3): 635–45. https://doi.org/10.1016/J.JMB.2004.02.002.

Wikoff, William R., James F. Conway, Jinghua Tang, Kelly K. Lee, Lu Gan, Naiqian Cheng, Robert L. Duda, Roger W. Hendrix, Alasdair C. Steven, and John E. Johnson. 2006. "Time-Resolved Molecular Dynamics of Bacteriophage HK97 Capsid Maturation Interpreted by Electron Cryo-Microscopy and X-Ray Crystallography." *Journal of Structural Biology* 153 (3): 300–306. https://doi.org/10.1016/J.JSB.2005.11.009.

Williams, Dewight R., Kyung Jong Lee, Jian Shi, David J. Chen, and Phoebe L. Stewart. 2008. "Cryo-EM Structure of the DNA-Dependent Protein Kinase Catalytic Subunit at Subnanometer Resolution Reveals α Helices and Insight into DNA Binding." *Structure* 16 (3): 468–77. https://doi.org/10.1016/J.STR.2007.12.014.

Williamson, Michael P., Timothy F. Havel, and Kurt Wüthrich. 1995. " Solution Conformation of Proteinase Inhibitor IIA from Bull Seminal Plasma by 1 H Nuclear Magnetic Resonance and Distance Geometry ," July, 319–39. https://doi.org/10.1142/9789812795830_0025.

Winn, M.D., C.C. Ballard, K.D. Cowtan, E.J. Dodson, P. Emsley, P.R. Evans, R.M. Keegan, et al. 2011. "Overview of the CCP4 Suite and Current Developments." *Urn:Issn:0907-4449* 67 (4): 235–42. https://doi.org/10.1107/S0907444910045749.

Wojtas, Magdalena N., Maria Mogni, Oscar Millet, Stephen D. Bell, and Nicola G. A. Abrescia. 2012. "Structural and Functional Analyses of the Interaction of Archaeal RNA Polymerase with DNA." *Nucleic Acids Research* 40 (19): 9941–52. https://doi.org/10.1093/NAR/GKS692.

Wolf, W. J., and D. R. Briggs. 1958. "Studies on the Cold-Insoluble Fraction of the Water-Extractable Soybean Proteins. II. Factors Influencing Conformation Changes in the 11 S Component." *Archives of Biochemistry and Biophysics* 76 (2): 377–93. https://doi.org/10.1016/0003-9861(58)90163-2.

Woo, Hyung-June, and Benoît Roux. 2005. "Calculation of Absolute Protein–Ligand Binding Free Energy from Computer Simulations." *Proceedings of the National Academy of Sciences* 102 (19): 6825–30. https://doi.org/10.1073/PNAS.0409005102.

Wriggers, Willy, Ronald A. Milligan, and J. Andrew McCammon. 1999. "Situs: A Package for Docking Crystal Structures into Low-Resolution Maps from Electron Microscopy." *Journal of Structural Biology* 125 (2–3): 185–95. https://doi.org/10.1006/JSBI.1998.4080.

Wright, Peter E., and H. Jane Dyson. 1999. "Intrinsically Unstructured Proteins: Re-

Assessing the Protein Structure-Function Paradigm." *Journal of Molecular Biology* 293 (2): 321–31. https://doi.org/10.1006/JMBI.1999.3110.

Yakunina, Maria, Tatyana Artamonova, Sergei Borukhov, Kira S. Makarova, Konstantin Severinov, and Leonid Minakhin. 2015. "A Non-Canonical Multisubunit RNA Polymerase Encoded by a Giant Bacteriophage." *Nucleic Acids Research* 43 (21): 10411–20. https://doi.org/10.1093/NAR/GKV1095.

Yang, Yi Isaac, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Qin Gao. 2019. "Enhanced Sampling in Molecular Dynamics." *The Journal of Chemical Physics* 151 (7): 070902. https://doi.org/10.1063/1.5109531.

Zacharias, M., T. P. Straatsma, and J. A. McCammon. 1998. "Separation-shifted Scaling, a New Scaling Method for Lennard-Jones Interactions in Thermodynamic Integration." *The Journal of Chemical Physics* 100 (12): 9025. https://doi.org/10.1063/1.466707.

Zhang, Kai. 2016. "Gctf: Real-Time CTF Determination and Correction." *Journal of Structural Biology* 193 (1): 1–12. https://doi.org/10.1016/J.JSB.2015.11.003.

Zheng, Shawn Q, Eugene Palovcak, Jean-Paul Armache, Kliment A Verba, Yifan Cheng, and David A Agard. 2017. "MotionCor2: Anisotropic Correction of Beam-Induced Motion for Improved Cryo-Electron Microscopy." *Nature Methods 2017 14:4* 14 (4): 331–32. https://doi.org/10.1038/nmeth.4193.

Zimmermann, Lukas, Andrew Stephens, Seung Zin Nam, David Rau, Jonas Kübler, Marko Lozajic, Felix Gabler, Johannes Söding, Andrei N. Lupas, and Vikram Alva. 2018. "A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at Its Core." *Journal of Molecular Biology* 430 (15): 2237–43. https://doi.org/10.1016/J.JMB.2017.12.007.

Zivanov, Jasenko, Takanori Nakane, Björn O. Forsberg, Dari Kimanius, Wim J.H. Hagen, Erik Lindahl, and Sjors H.W. Scheres. 2018. "New Tools for Automated High-Resolution Cryo-EM Structure Determination in RELION-3." *ELife* 7 (November). https://doi.org/10.7554/ELIFE.42166.

Vita

Alec Fraser was born in Ottawa, Ontario, Canada on October 2nd, 1994 to Andrew Fraser (father) and Melissa Hooper (mother). He attended Hopewell Avenue public school up until high school where he graduated from Glebe Collegiate Institute. He earned his Bachelor of Science in Mathematics and Physics from McGill University in Montreal, Canada. He has published in the IEEE journal of selected topics in quantum electronics, Viruses, Science Advances and Proteins.

Alec Fraser, Nikolai Prokhorov, Ekaterina Knyazhanskaya, John M. Miller, Petr G. Leiman (2021). Identification of Low Population States in Cryo-EM Using Deep Learning. *bioRxiv.*

Fraser, A., Sokolova, M. L., Drobysheva, A. V., Gordeeva, J. V., Borukhov, S., Artamonova, T. O., ... & Leiman, P. G. (2021). Template strand deoxyuridine promoter recognition by a viral RNA polymerase. *bioRxiv*.

Kryshtafovych, A., Moult, J., Albrecht, R., Chang, G. A., Chao, K., Fraser, A., ... & AlphaFold2 team. (2021). Computational models in the service of X-ray and cryo-electron microscopy structure determination. *Proteins*.

Fraser, A., Prokhorov, N. S., Jiao, F., Pettitt, B. M., Scheuring, S., & Leiman, P. G. (2021). Quantitative description of a contractile macromolecular machine. *Science Advances*, *7*(24), eabf9601.

Kanamaru, S., Uchida, K., Nemoto, M., Fraser, A., Arisaka, F., & Leiman, P. G. (2020). Structure and Function of the T4 Spackle Protein Gp61. 3. *Viruses*, *12*(10), 1070.

Vampa, G., McDonald, C., Fraser, A., & Brabec, T. (2015). High-harmonic generation in solids: Bridging the gap between attosecond science and condensed matter physics. *IEEE Journal of Selected Topics in Quantum Electronics*, *21*(5), 1-10.

Permanent address 3-220 Sunnyside Avenue, Ottawa, ON, Canada

This dissertation was typed by Alec Fraser.