UNIVERSITAT DE BARCELONA

Free energy and information-content measurements in thermodynamic and molecular ensembles

Álvaro Martínez Monge

ÁLVARO MARTÍNEZ MONGE

FREE ENERGY AND INFORMATION-CONTENT MEASUREMENTS IN THERMODYNAMIC AND MOLECULAR ENSEMBLES

FREE ENERGY AND INFORMATION-CONTENT MEASUREMENTS IN THERMODYNAMIC AND MOLECULAR ENSEMBLES

ÁLVARO MARTÍNEZ MONGE

DEPARTAMENT DE FÍSICA DE LA MATÈRIA CONDENSADA
FACULTAD DE FÍSICA
UNIVERSIDAD DE BARCELONA

UNIVERSITAT DE BARCELONA

Memoria presentada para optar al título de
DOCTOR EN FÍSICA.

Tesis realizada bajo la supervisión del Dr. Fèlix Ritort Farran y la Dra. María Mañosas Castejón.

 Tesis realizada gracias a la concesión de una ayuda predoctoral del Programa Estatal de Promoción del Talento y su Empleabilidad en I+D+i (BES-2014-068730).

El éxito consiste en ir de fracaso en fracaso sin perder el entusiasmo.
— Winston Churchill

AGRADECIMIENTOS

A lo largo de todos estos años siempre me he preguntado qué pensaré cuando eche la vista atrás y mire con perspectiva esta etapa que ahora concluye. Ahora estoy seguro que recordaré todos estos años con un cariño muy especial. No únicamente por lo que me ha aportado la tesis en sí, si no por todas esas personas que he me he encontrado a lo largo de esta aventura. Sí, porque ha sido una aventura con sus momentos buenos —muchos— y sus momentos malos —otros tantos—, pero me gusta pensar que esto, en el fondo, ha sido cumplir un sueño.

Gracias de corazón a mi director, Fèlix, por confiar en mí hace seis años y darme la oportunidad de sumergirme en la ciencia de verdad. Gracias por valorar siempre mi trabajo y por enseñarme a ser meticuloso y preciso. Por nunca perder ese ímpetu tan tuyo, ni en las buenas ni en las malas. Por enseñarme a hacerme las preguntas adecuadas. Espero que los dos guardemos con mucho cariño en la memoria esta etapa trabajando juntos.

Gracias de corazón también a mi otra directora, María. Apareciste cuando solo tenía nubarrones negros sobre la cabeza y fuiste capaz de darme un empujón cuando más lo necesitaba. Gracias por tu paciencia y por tus buenos consejos.

También recuerdo de manera muy especial a mi "padre" científico, Marco. Fuiste una persona muy importante al principio de mi carrera. Tú me enseñaste que, esencialmente, la vida en la ciencia es una montaña rusa. Estás un día arriba y tres abajo. Aún así, gracias a ti sé que hay que aprovechar y disfrutar al máximo de las pequeñas cosas que descubrimos. Gracias porque, de verdad, sin ti esto no hubiera sido posible.

Si echo la vista atrás siempre sonrío al recordar a una persona que pasó de ser mi profesora a una muy buena amiga y un gran ejemplo en el día a día en el laboratorio. Gracias Anna por todo lo que me has enseñado, por tu energía vital, tus buenos consejos y por saber que siempre estás ahí. Tú también tienes gran parte de "culpa" de esta tesis.

De toda esta historia lo que más valoro es haberme encontrado con personas aparentemente muy diferentes a mí, pero especiales y únicos a partes iguales. Muchas veces únicamente gracias a ellos he tenido fuerzas para continuar. Por eso, Marc, Xavi e Isabel sois una parte imprescindible de esta historia. Gracias Marc por esas innumerables horas de charlas de cualquier cosa, esos piques futboleros, esas cuentas atrás tan eternas en la pizarra para Star Wars, por las cervezas en tu terraza con Anna, por esas sesiones de despotrique en las que no queda títere con cabeza y tantas otras cosas. Gracias Xavi por dejarme ver la persona tan especial que eres, por tus consejos, por hacerme los días más amenos con esas cosas tan… especiales… ya sabes, por Twitter. Me demostraste ser una de las personas más valientes que he conocido. No por salir a parar el tráfico en un puente en la A2 en hora punta, si no por seguir siendo tú mismo día a día sin importar que vengan mal dadas. Gracias a Isabel, mi querida gata madrileña, por tantos ratos a tu lado en el laboratorio, por animarnos el uno al otro siempre a seguir partiéndonos la cara cada día y… por dejarme achuchar a Enzo y Lola, claro. Gracias a los tres por convertiros en parte de mis mejores amigos. Nunca os podré agradecer lo suficiente los buenos ratos que hemos pasado.

No puedo olvidarme de muchas otras personas que también han pasado por el Small Biosystems Lab con las que he compartido muchos momentos únicos. Marta, mujer cañera donde las haya. Detrás de su fachada se oculta una de las personas más bonitas que conozco. Gracias por tu transparencia, por siempre devolverme una sonrisa en mis días refunfuñones —que son muchos—. Gracias también por todos los momentos juntos. También gracias a Aurélien que, a parte de ser un físico brillante es una persona increíble. Merci pour tout mon ami. Merci pour tes précieux conseils (personnelles et scientifiques), merci pour les soirs de Tarantino, pizza, bière et chansons des Beatles. Tu es une des plus belles personnes que je connaisse. ¡Mucho fuego para siempre! Una de personas más entrañables de todas las que han pasado por el laboratorio es Laura. Mi querida Laura, gracias por todos los momentos contigo, por las risas a tu lado y por siempre demostrarme cariño. También parte de esta tesis es gracias a Ainara. Qué pena que no hayamos podido compartir más momentos juntos en el laboratorio.

No hay que olvidarse del relevo generacional, siempre tan ignorado. A Paolo, compañero de despacho en esta etapa final. Eres un fenómeno.

Espero que en el futuro compartamos experiencias juntos tanto online como offline. A Jaime, mi relevo merengue en el laboratorio, quien, aún llevando poco tiempo, ya nos ha ganado a todos y ha hecho que nos desternillemos de risa en más de una ocasión.

Soy muy afortunado porque he podido disfrutar de la experiencia de la docencia universitaria. Gracias Matteo, Jesús, Agustí, Federico, Arantxa, Miquel. Todos me habéis aportado perspectiva y me habéis ayudado con vuestros buenos consejos y palabras. Gracias a Iván, que, a parte de ayudarme con unas difíciles clases de FOFT, tú me enseñaste a tener mano izquierda en muchas ocasiones.

También me gustaría mencionar y recordar a todas esas personas que, sin compartir tu día a día, han sido capaces de arrancarme una sonrisa en mis días más grises. Gracias Lucas, Luisen, Isa, Eloy, Joan, Helena, Bea, Miriam y Paola.

No quiero olvidar a toda esa gente que me hizo amar la ciencia como tal. Sin publicaciones, ni proyectos, ni nada por el estilo. Recuerdo con especial cariño a Miquel. Aún con su mirada fija y expresión imperturbable era capaz de mostrarte la física más básica con imaginativos experimentos adornados con un toque de humor. Esto también va por ti, Miquel.

Lejos del ámbito académico agradezco el apoyo sincero a mi familia. A mis tíos Jordi, Conchita y Alicia, por su constante interés y sus muy necesarios ánimos. Que sepáis que esto lo celebramos con un *torrenillo*. No importa cuando leáis esto. También a Jesús y Rosamari que, gracias a vuestro inmenso cariño desde la distancia, también tenéis gran parte de "culpa" de que lo haya conseguido. No me olvido nunca de las personas que ya no están. Todo lo que soy, para bien y para mal, también se lo debo a ellos.

Mamá, gracias por estar siempre ahí, gracias por enseñarme la belleza de las palabras. Por mostrarme lo bello de la literatura y cómo, en parte, la ciencia es también poesía. Gracias por siempre maravillarte cuando te explico cosas de física. Eso es lo que hace mantener la ilusión viva cada día. Gracias Papá por enseñarme a valorar el trabajo bien hecho y esas dosis de perfeccionismo tan necesarias en este mundo. Os debo mucho más de lo que pensáis.

Gracias a María, mi hermanita. Gracias por ser siempre mi crítica más exigente y siempre regalarme tus sabios consejos. Por estar en los

buenos y en los malos momentos con un café cuando más se necesita. Esta tesis también va por ti.

Finalmente, a ti, mi Andrea. Siempre a ti. Gracias por aparecer en mi vida y hacerme sentir tan especial. Gracias por levantarme del suelo todos estos años cuando yo no tenía fuerzas para hacerlo. Gracias por tu inmensa paciencia conmigo, por tu apoyo, por entenderme —muchas veces mejor que yo mismo—, por consolarme y quererme. Tú has impedido que olvide que esta tesis es por y para nosotros. Por todo esto y todo lo que está por venir a tu lado, muchas gracias Andrea.

Barcelona, 4 de marzo de 2019

CONTENTS

ACRONYMS

AFM Atomic Force Microscopy

AFS Acoustic Force Spectroscopy

CEBA Continous Effective Barrier Analysis

CFT Crooks Fluctuation Theorem

dsDNA double-stranded DNA

ECDF Empirical Cumulative Distribution Function

ECFT Extended Crooks Fluctuation Theorem

EWD Ensemble Work Distribution

ExtEns Extensional Ensemble

FDCs Force-Distance Curves

FEL Free Energy Landscape

FJC Freely Jointed Chain

ForceEns Force Ensemble

JE Jarzynski Equality

LOT Laser Optical Tweezers

MT Magnetic Tweezers

SME Single Molecule Experiments

ssDNA single-stranded DNA

ssRNA single-stranded RNA

WLC Worm-Like Chain

RESUMEN DE LA TESIS EN CASTELLANO

El siglo XX marcó un punto de inflexión en el desarrollo de la física. Concretamente, la física evolucionó para llegar a ser una ciencia capaz de promover el desarrollo tecnológico. Como paradigmas de este hecho podemos mencionar el desarrollo de la teoría de la relatividad por parte de Albert Einstein a principios de siglo y la eclosión de la teoría cuántica. Pese a que muchas de las aplicaciones tecnológicas de la mecánica cuántica aún permanecen en estado latente, otras muchas han supuesto una revolución sin precedentes en muchos campos. Un ejemplo de ello ha sido la invención del láser. El láser —acrónimo de *light amplification by stimulated emission of radiation*— es un dispositivo basado en un efecto puramente mecano-cuántico —la emisión estimulada— y es ampliamente utilizado hoy en día. Una de las aplicaciones más importantes en el mundo científico y tecnológico del láser fue la invención de los instrumentos de pinzas ópticas por parte de Arthur Ashkin en 1970. Precisamente gracias a este hecho le fue otorgado el premio Nobel de Física en el año 2018.

Gracias a los estudios y desarrollos de Ashkin, junto con la sofisticación de las técnicas en el campo de la biofísica molecular, se produjo una revolución en el campo de la biofísica. Por ejemplo, mediante el uso de los instrumentos de molécula individual se ha conseguido, con una resolución espacial y temporal sin precedentes, medir y observar reacciones moleculares otrora impensable. Este hecho tiene muchas aplicaciones, tanto a nivel biológico, como físico. Por un lado, desde una perspectiva biológica, los dispositivos de molécula individual han permitido estudiar el plegamiento de las proteínas y los ácidos nucleicos, la unión de ligandos o iones a sistemas moleculares e, incluso, la observación y seguimiento de motores moleculares sobre sustratos in vivo. Por otro lado, desde el punto de vista físico, el poder estudiar y experimentar con sistemas de molécula individual ha permitido desarrollar el campo de la física de los llamados sistemas pequeños. Las dimensiones de estos sistemas abarcan desde unos pocos nanómetros —una millonésima parte del metro— hasta varios cientos de nanómetros. Además, los sistemas pequeños están lejos del llamado límite

termodinámico y están dominados por las fluctuaciones térmicas del entorno. Por lo tanto, debido a estas peculiaridades, el estudio de sistemas pequeños mediante los instrumentos de molécula individual está permitiendo impulsar y extender los horizontes de la física de no equilibrio.

Esta tesis doctoral se ha llevado a cabo empleando dos de los instrumentos de molécula individual más conocidos: las pinzas ópticas y las pinzas magnéticas. Ambas son técnicas que permiten la aplicación controlada de fuerzas mecánicas a los extremos de una molécula individual. Una molécula en cada realización experimental. El poder aplicar fuerzas a sistemas moleculares permite llevar a cabo una profunda caracterización de las propiedades físicas de dichos sistemas. Por ejemplo, mediante experimentos de desnaturalización mecánica, es posible inducir transiciones entre el estado nativo y el estado desplegado de horquillas de ácidos nucleicos o proteínas. Además, las pinzas ópticas y magnéticas permiten medir distancias con una precisión nanométrica, con lo que se pueden realizar medidas de extensiones moleculares con un grado de precisión incomparable. Además, no únicamente permiten medir distancias moleculares, si no que permiten controlarlas. Es decir, los instrumentos utilizados permiten fijar, de manera controlada, la distancia entre los dos extremos de una molécula. Dicha distancia está estrechamente relacionada con la extensión de la molécula.

El estudio y comparación de ambos regímenes —fuerza y extensión controlada— es precisamente el objeto de estudio de la primera parte de la tesis. Mediante el uso de una horquilla de ADN bien caracterizada a nivel de molécula individual, hemos demostrado cómo ambas situaciones —controlar fuerza o controlar distancia— no son equivalentes. En particular, la energética de ambas situaciones no es la misma. Fundamentalmente, ambas magnitudes son diferentes. Mientras que la distancia es una magnitud extensiva, la fuerza es una magnitud física intensiva. Este hecho conlleva que la descripción termodinámica —el colectivo estadístico— de los sistemas moleculares en uno u otro contexto sea diferente. Pese a ello, hemos mostrado cómo es posible conectar ambas descripciones termodinámicas en un contexto general. Adicionalmente, hemos observado cómo dicha no equivalencia también tiene un fuerte impacto a nivel cinético, abriendo un amplio abanico de interesantes preguntas para investigar en el campo de la biofísica celular.

La segunda parte de la tesis está basada en el desarrollo de un método sistemático para medir contenidos de información en sistemas moleculares. La conexión de la física estadística con la teoría de la información se inició hace casi 75 años, gracias a los trabajos de Claude E. Shannon. Posteriormente, científicos como Jaynes desarrollaron una conexión directa de la física estadística de equilibrio y de no equilibrio con la teoría de la información. De nuevo, gracias a los instrumentos de molécula individual y a los controles de retroalimentación en dichos instrumentos —en inglés, feedback—, recientemente se consiguió realizar experimentalmente un demonio de Maxwell, así como también se demostró la posibilidad de asociar una información termodinámica a un sistema fuera de equilibrio. Estos trabajos permitieron demostrar cómo el hecho de poseer información tiene implicaciones termodinámicas. Nosotros, con el desarrollo de esta tesis, hemos puesto la primera piedra para relacionar la termodinámica con la información. En particular, hemos demostrado la posibilidad de realizar medidas de contenidos de información —con una precisión de unos pocos bits— en sistemas moleculares únicamente mediante el estudio de cantidades termodinámicas medibles. Las aplicaciones de este método son muy amplias, ya que abre la posibilidad medir contenidos de información en procesos, por ejemplo, de evolución molecular dirigida.

Finalmente, la tercera parte de la tesis se basa en la medición de energías de unión específicas de iones magnesio con un sustrato de ARN de gran importancia biológica. Hasta ahora, la posibilidad de medir directamente la energía específica de unión de un ion metálico a una molécula de ARN con experimentos de molécula individual no existía. Este hecho era debido a la imposibilidad de discernir las contribuciones específicas y no específicas a la energía de unión. Gracias a recientes estudios experimentales con pinzas ópticas, se demostró la equivalencia entre dos condiciones iónicas, a priori, diferentes. Basándonos en estos estudios, hemos sido capaces de llevar a cabo experimentos de molécula individual sobre una molécula que contiene el nexo de tres hélices de ARN con los que hemos sido capaces de medir, con una precisión de unos pocos $k_{\mathrm{B}}T$, la contribución específica de unión por parte del magnesio.

Los resultados de esta tesis tienen implicaciones a nivel físico y biológico. Por un lado, los resultados correspondientes a la primera y segunda parte son de mucha relevancia a nivel físico. El estudio de la

no equivalencia entre colectivos estadísticos es un tema candente en la actualidad precisamente gracias a la posibilidad de manipular sistemas nano y microscópicos de modos —hoy en día— irrealizables en sistemas macroscópicos. Del mismo modo, poder realizar mediciones de contenido de información y demostrar la conexión entre la termodinámica y la información es una de las líneas de investigación más actuales en el ámbito de la física estadística. Por otro lado, los resultados de la última parte de la tesis son de gran impacto a nivel biológico y biofísico. Las medidas de energías de unión específicas, tanto de iones como de ligandos, tradicionalmente se han llevado a cabo utilizando las llamadas técnicas de volumen, siendo este tipo de medidas particularmente complicadas. Por lo tanto, nuestro estudio nos ha permitido obtener un resultado sin precedentes en el campo de los experimentos de molécula individual.

Part I

PRELIMINARIES

1

GENERAL INTRODUCTION

> Depending on how you look at it, statistical mechanics is either the least fundamental or most fundamental of all fields of physics. That is because it is not really science at all. It is pure mathematics.
>
> Peter Eastman

The interest that classical thermodynamics aroused from the 17th to the 19th centuries is, probably, only surpassed by the birth of quantum physics in the beginning of the 20th century. Indeed, thermodynamics fuelled the biggest boost of humanity, being the fiercest exponent of it the development of sophisticated heat engines by the mid-1800s. The revolution triggered by thermodynamics and the impact on mankind is not yet comparable nowadays with any recent scientific discovery.

As a major feature, thermodynamics settled the possibility of studying *macroscopic* physical systems regardless of their inner structure or their behavior, considering only the relations with their surroundings. It was not until the probabilistic interpretation of the Second Law developed by Boltzmann [1] (1844 – 1906) that the microscopic nature of physical systems was added as a new ingredient to the theory. For the first time, the entropy S is related to statistical considerations of the accessible *microstates* of the system (rather than being introduced axiomatically as Carathéodory or phenomenologically as Clausius), resulting in the well-known expression [2]:

$$S = k_{\mathrm{B}} \log W \,, \tag{1.1}$$

where k_B is the Boltzmann constant[1] and W the so-called thermodynamic probability of a macrostate (i.e. the number of available microstates of the system).

Boltzmann also introduced the concept of thermodynamic equilibrium from a probabilistic point of view. Statistical mechanics was formally born. As a matter of fact, he laid the foundations of ensemble theory, predicted the so-called equilibrium statistical fluctuations (being the Brownian motion the most noticeable effect of equilibrium thermal fluctuations) and he investigated, for the first time, the non-equilibrium regime with his H-theorem.

Figure 1.1: **Statistical mechanics fathers.** Portrait of Ludwig Boltzmann (1844 – 1906), who introduced probabilistic concepts into thermodynamics (left picture). Josiah Willard Gibbs (1839 – 1903), founder of modern ensemble theory (right picture).

Modern ensemble theory was developed by Gibbs (1839 – 1903). He was able to connect the properties of statistical ensembles to the laws of thermodynamics and, moreover, he formalized statistical mechanics as a general theory that can be applied to all kind of physical systems.

During the course of the 20th century non-equilibrium phenomena and information theory attracted the attention of statistical mechanics. Spanning from the study of stochastic processes as Brownian motion via the Fokker-Planck equation [3] and the study of dissipative systems (or structures) [4] in non-equilibrium contexts, through the foundation of information theory [5] and its latter connection to thermodynamics [6, 7], statistical thermodynamics has successfully expanded its "old"

1 Although Eq. (1.1) was formally written by Planck (1858 – 1947), Boltzmann showed that the entropy S is proportional to the $6N$–dimensional phase space volume Ω occupied by the corresponding macrostate of an N–particle system.

perspective towards new horizons, such as biological, evolutionary systems or even non-Hamiltonian systems (complex systems).

1.1 FLUCTUATIONS AND THE PHYSICS OF SMALL SYSTEMS

Statistical mechanics defines a general framework in which the connection with thermodynamics (i.e. the macroscopic behavior) is well settled in the so-called *thermodynamic limit*. It corresponds to the idealized limit in which a system composed by an arbitrary large number N of *units* (e.g. atoms, particles, etc.) occupies an infinite volume V but the ratio N/V is kept constant. Macroscopic measurable thermodynamic quantities are the result of the average over all the possible states of the N molecules forming the system. On the other hand, fluctuations are inherent to all physical systems and their description arises naturally within statistical mechanics framework. Indeed, it can be shown that the relative fluctuations[2] of thermodynamic quantities (such as energy or entropy) are of the order $1/\sqrt{N}$, vanishing (or being extremely hard to measure) in the thermodynamic limit.

One of the most challenging aspects of statistical mechanics is the study of physical systems with a small number N of particles, far away from the thermodynamic limit. A paradigmatic example of small systems are single molecule systems, where $N \sim 1$ and, hence, $1/\sqrt{N} \sim 1$. Moreover, in small systems, typical energetic exchanges between the system and the environment are of the order of Brownian fluctuations (i.e. $\sim k_B T$). Small systems thermodynamics provides the precise framework that allows us to understand the physics of small systems. Applications of small systems thermodynamics range from molecular systems (such as nucleic acids, proteins or molecular machines) up to self-propelled organisms (cells or active colloids).

The study of small systems and the small N regime, yet seemed unrealisable in the dawn of statistical mechanics, has become a trending topic due to the recent developments of micromanipulation techniques (together with the progress of stochastic thermodynamics), such as Single Molecule Experiments (SME).

The importance of fluctuations in ensemble theory is deeply discussed in chapter 3 of the present thesis.

2 Defined as the root mean square on the scale of the mean.

1.2 MOLECULAR BIOPHYSICS

Biophysics is a bridge science. It is the field of science that applies the physical theories and methodologies to study biological systems. Moreover, the enterprise of biophysicists is to find out which are the physical laws that govern life. In this context, the scope of biophysics spans from molecules, to cells, tissues and even populations and evolutionary systems.

Figure 1.2: **Discovery of DNA structure.** James Watson and Francis Crick showing the three-dimensional structure of DNA in 1953 (left picture). Rosalind Franklin (central picture), author of the X-Ray crystallography image crucial for the inference of the double helix structure of DNA. The so-called photo 51 is the nickname of the image of crystallized DNA obtained in 1952 (right picture).

Since the mid-1800s, some physicists were interested on applying physical theories to biological systems [8]. The interest on biological systems, and life in particular, started to grow up during the second half of the 20th century, probably thanks to the inspirational book wrote in 1944 by Erwin Schrödinger: *What is Life?* [9]. Few years later, the double helix structure of DNA was discovered by James Watson and Francis Crick in 1953 [10] by using the X-Ray diffraction images obtained by Rosalind Franklin. Schrödinger's book, together with the structural resolution of the most important molecule in biology, burgeoned the interest of late-1900s scientists onto molecular biophysics.

The revolution that entailed molecular biophysics has not yet come to its end. The boost of biochemical procedures and the development of single molecule techniques have led to a significant breakthrough in the field of non-equilibrium statistical physics.

1.3 SINGLE MOLECULE EXPERIMENTS

One of the main results of the participation of physicists in biological issues is the development of extremely precise experimental techniques. Besides crystallographic techniques and imaging techniques [11], where the interaction with the biological systems is pursued to be as minimal as possible, single molecule assays emerge as one of the most powerful techniques to study the behavior of biopolymers that are crucial for life, such as nucleic acids or proteins. As a matter of fact, in SME the studied systems are externally (and individually) perturbed and their response is measured with an unprecedented accuracy, allowing experimentalists to characterize the kinetics and energetics of individual molecules.

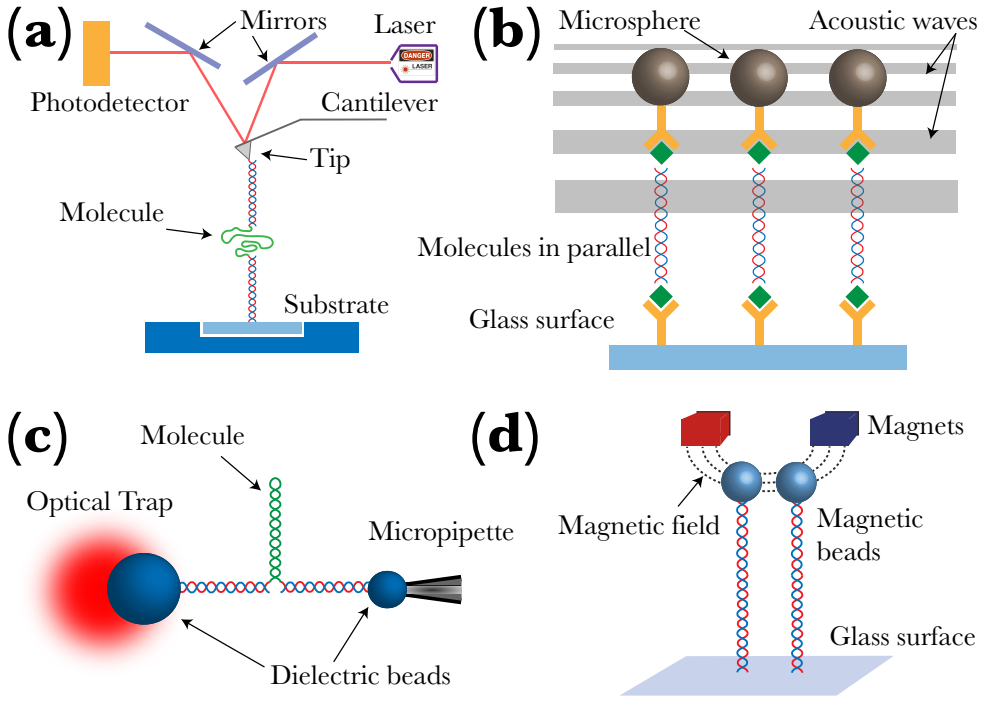

Figure 1.3: **Schematics of single molecule devices.** **(a)** - Atomic Force Microscope. **(b)** - Acoustic Force Spectroscopy. **(c)** - Laser Optical Tweezers. **(d)** - Magnetic Tweezers.

Among all single molecule techniques, we will focus in force spectroscopy techniques (they allow to exert mechanical forces on individual molecules). As major protagonists of force techniques we highlight: Atomic Force Microscopy (AFM), Acoustic Force Spectroscopy (AFS), Laser Optical Tweezers (LOT) and Magnetic Tweezers (MT), all of them allow experimentalists to manipulate single molecules. Nevertheless,

there are some differences that make some techniques more suitable than others depending on the studied system or the goal of the experiment. An schematic cartoon of each experimental set-up is shown in Fig. 1.3. Now we will briefly summarize the main features and drawbacks of each instrument [12, 13].

- Although **AFM** is widespread used in imaging, it can be used as a force spectroscopy tool. In AFM the molecules are adsorbed on a planar surface (substrate) that can move relative to a metallic cantilever. The tip of the cantilever is coated with molecules that can recognize and bind (either specifically or non-specifically) a site of the molecules adsorbed on the surface. Therefore, by moving the surface vertically, forces can be exerted on the molecule. The force is measured by recording the deflection of the cantilever using a laser beam focused to the edge of the cantilever (see Fig. 1.3(a)). The force can be obtained in real time as a function of the molecular extension. Cantilevers need to be soft enough to detect typical molecular forces but, generally, AFM force resolution covers from 10 to 10^4 pN. Considering a stiffness as small as $k \sim 10$ pN/nm, the typical spatial resolution can be estimated from Equipartition law, giving: $\sqrt{\langle \Delta x^2 \rangle} = \sqrt{k_B T / k} \sim 1$ Å.

  While AFM instruments are the perfect tool to study strong covalent interactions (~ 1 nN = 10^{-9} N $= 10^3$ pN), they present several major drawbacks in single molecule studies. For instance, the exploration of the low force regime (< 10 pN) turns out to be difficult. As a matter of fact, there can happen many undesired interactions when the cantilever is approached to the surface (this fact can be overcame by using single molecule markers). On the other hand, cantilevers are very fragile, turning AFM one of the most time-expensive techniques.

- **AFS** is based on the principle that a microsphere experiences a force within a planar acoustic standing wave. The z-component of the acoustic force equals to:

$$f = -V \frac{\partial}{\partial z} \left(\frac{\kappa_m - \kappa_p}{4} p^2 - \frac{\rho_p - \rho_m}{2\rho_p + \rho_m} \rho_m v^2 \right), \tag{1.2}$$

being V the volume of the microsphere; κ_m and κ_p the compressibility of the medium and the microsphere, respectively; and ρ_m and ρ_p the density of the medium and the trapped microsphere, respectively; p the acoustic pressure and v the acoustic velocity. For standard polystyrene beads, the second term of Eq. (1.2) is small as compared to the pressure term [14]. As a consequence, the force is dominated by the pressure gradient, driving the microspheres towards acoustic pressure nodes. The typical applied forces range from 1 - 10^2 pN. On the other hand, the spatial fluctuations are dominated by the image sampling bandwidth and they are around 10 nm [15].

AFS permits high-throughput single molecule measurements (i.e. measure several molecules in parallel). It means that it is possible to measure hundreds of molecules at a time. Additionally, AFS is a relative simple technique allowing a straightforward implementation in lab-on-a-chip devices. It is important to mention, however, that the spatial resolution is poorer as compared to other single molecule techniques and specific limitations due to image processing methods (i.e. bandwidth limitation, etc.). The temporal resolution strongly depends on the wave frequencies, on the viscosity of the surrounding medium and the number of tracked beads. In Fig. 1.3(b) we show a schematic cartoon of the experimental construction of AFS, emphasizing its paralleling capabilities.

- The physical principle behind **LOT** is the optical gradient force. It is created by the deflection of a light beam focused on an object with an index of refraction higher than the one of the surrounding medium (see section 2.2 for further details). On the one hand, a biochemically modified micron-sized bead is captured in the focus of an optical trap. On the other hand, another bead containing in its surface a molecular construct is held in the tip of a micropipette by air suction[3]. Both beads are specifically biochemically labelled so that the molecular construct form a tether between the two beads. The construction contains the single molecule that will be studied and another pair of molecules that act as spacers between

3 Although it is possible to implement a second optical trap, our description stands for the single trap construction.

the beads, as can be seen in Fig. 1.3(c). The usual force range in LOT is 10^{-1} - 10^{2} pN. Such range depends, essentially, on two factors: the bead size and the laser power. The spatial resolution is dominated by thermal fluctuations but since the typical stiffness of LOT is about $k_{\mathrm{LOT}} \sim 10^{-2}$ pN/nm $10^{-3} \cdot k_{\mathrm{AFM}}$ it fairly reaches nanometric precision.

The main advantage of LOT relies on their remarkable force precision (typically $\sim 10^{-1}$ pN) and the possibility, depending on the setup, of a direct measurement of the force (without post-processing techniques). This fact, combined with the sub-milisecond temporal resolution, turns LOT into one of the most versatile single molecule techniques. The disadvantage of LOT is the complicated setup (as well as the optical aligning) and the difficulty of carrying out high-throughput measurements.

- **MT** are based on the physical phenomenon that a superparamagnetic bead becomes magnetized when subjected to an external magnetic field **B**. And the force that the magnetized bead feels is proportional to the magnetic field gradient (see section 2.3 for a detailed description). Molecular extensions are determined by image analysis and the force is determined by equipartition law: $f = k_{\mathrm{B}}T\frac{x}{\langle \Delta x^2 \rangle}$, being x the molecular extension and $\langle \Delta x^2 \rangle$ the average extension fluctuations (which depend on the molecule). The stiffness of magnetic traps are on the order of 10^{-4} pN/nm, allowing sub-piconewton exploration and controlled force measurements. Typical operating forces cover: $10^{-2} - 10^{2}$ pN.

  An important feature of MT is, besides its low force capabilities, the possibility of exert torques on molecules. This fact allows to study elastic and torsional properties of DNA and molecular motors. Like AFS, MT allows for high-throughput measurements. Yet, again, the main drawback of MT technique relies on the fact that measurements are limited by the acquisition rate of the tracked elements.

As a final summary, in Table 1.1 we summarize the main characteristics of each experimental technique. Parameters were obtained from Refs. [12, 14, 16].

We have briefly mentioned four of the single-molecule techniques capable of exerting forces on molecules (other techniques are be related

	AFM	AFS	LOT	MT
Force range [pN]	10^1 - 10^4	10^1 - 10^3	10^{-1} - 10^2	10^{-2} - 10^1
Spatial resolution [nm]	0.1	10	1	1
Stiffness [pN/nm]	10^1 - 10^5	10^{-2}	10^{-2} - 10^0	10^{-4}
Temporal resolution [s]	10^{-3}	10^{-2}	10^{-4}	10^{-1} - 10^{-3}
Paralleling?	✗	✓	✗	✓

Table 1.1: **Single molecule techniques.** Comparison of single molecule force spectroscopy techniques. ✓= yes, ✗= no

to fluorescence and holographic methods, just to mention a few). In combination, the four force spectroscopy methods that we highlighted, form the perfect lineup to face a broad range of molecular problems in a wide force regime (see Fig. 1.4), as well as the exploration of the conditions where thermal fluctuations really matter and new non-equilibrium physics is still to be discovered.

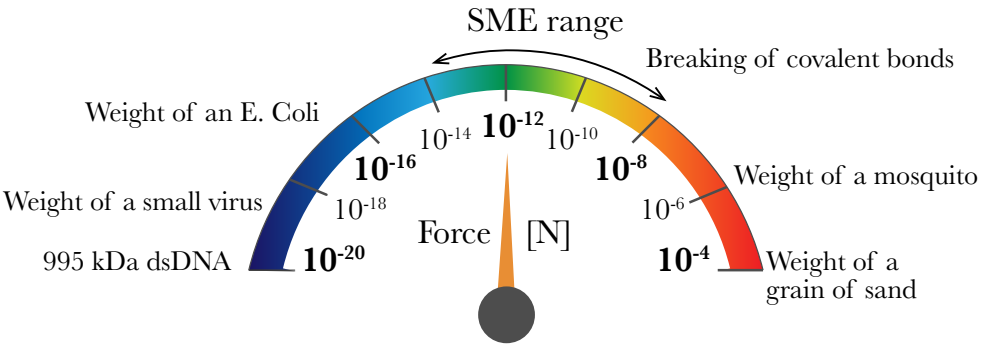

Figure 1.4: **Force range.** Illustration of SME force range compared to different examples of forces at different orders of magnitude.

2

EXPERIMENTAL SETUP

This thesis addresses several questions in the field of non-equilibrium statistical mechanics. As we already discussed in chapter 1, SME are the perfect playground for studying non-equilibrium phenomena.

The present chapter aims to introduce, from a biological and chemical perspective, the biological molecules that are employed as physical systems (i.e. nucleic acids) and to perform a detailed description of the single molecule devices (i.e. LOT and MT), covering from their physical principles, the experimental set-up and their typical calibration procedures.

2.1 NUCLEIC ACIDS

The term *nucleic acid* is the overall name for deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Among the most important existing macromolecules, nucleic acids are essential for life.

They are called nucleic acids for two reasons: first, they were discovered in the nucleus of white blood cells and they present some acidic properties. The discovery of nucleic acids was done by Friedrich Miescher (1844 – 1895) in 1869, who called them nuclein [17]. He realized that the molecules he discovered, were phosphate rich and quite similar to proteins [18]. Miescher's discovery was done in a period where biology shifted its attention focus from organisms to cells. Indeed, nuclein was found a few years after Charles R. Darwin (1809 – 1882) published his famous book *On the Origin of the Species by Means of Natural Selection* [19]. Miescher's findings paved the way to relate inheritance to DNA. Richard Altmann (1852 – 1900) renamed nucleins to nucleic acids after he discovered their acidic properties in 1889. Nevertheless, the chemical and biological differences between DNA and RNA were not well established until the first part of the 20th century.

Nowadays we know that nucleic acids are essential to store, copy and transmit genetic information. They are polymeric molecules formed by *nucleotides*, which are macromolecular compounds comprising a

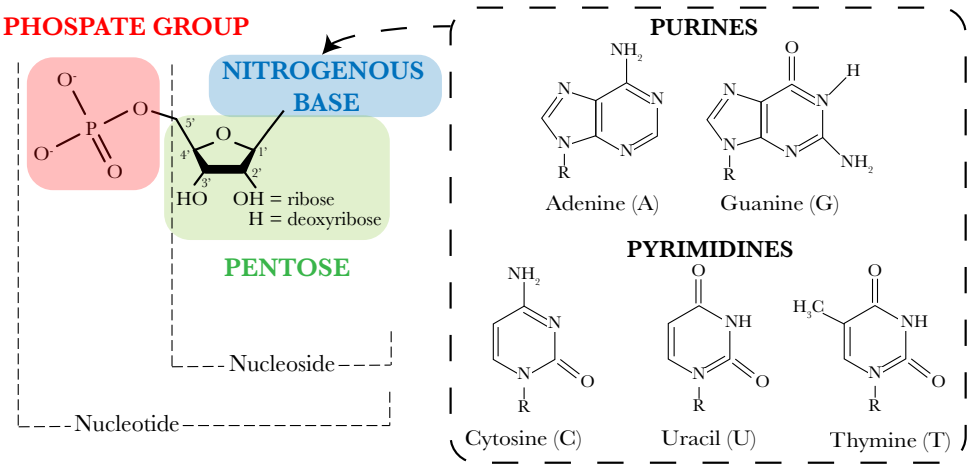

Figure 2.1: **Components of nucleic acids.** Chemical structure of a nucleotide, formed by a phosphate group plus a 5-carbon sugar (pentose) and a nitrogenous base.

nitrogenous base (also called nucleobase), pentose (i.e. a five-carbon sugar), and a phosphate group. The sugar can be either a deoxyribose in DNA or a ribose in RNA. There exist five nitrogenous bases: adenine (A), cytosine (C), guanine (G), thymine (T) and uracil (U). While A,C,G are common in DNA and RNA, T is only found in DNA and U is only found in RNA. A and G nucleobases are classified as purines while C, U, T bases are called pyrimidines (Fig. 2.1).

Both in DNA and RNA, individual nucleotides link each other by a phosphodiester bond from the 3' sugar carbon of one nucleotide to the 5' sugar carbon of the following nucleotide. The prime notation indicates the directionality of the molecule. Typically, the carbon linked to the nucleobase is labelled as 1' while the rest of the carbons are labelled as 2', 3', etc., increasing from clockwise direction starting from the 1' (see scheme in Fig. 2.1).

As can be seen in Fig. 2.2 the phosphodiester bond takes place in the phosphate group of each nucleotide. This bond is formed in the process of DNA synthesis by the DNA polymerase enzyme [20].

Regarding the structure of nucleic acids, it is traditionally ranked into four levels [21]:

i. The **primary structure**, corresponds to the nucleotide (for the case of nucleic acids) or aminoacid sequence (for the case of proteins).

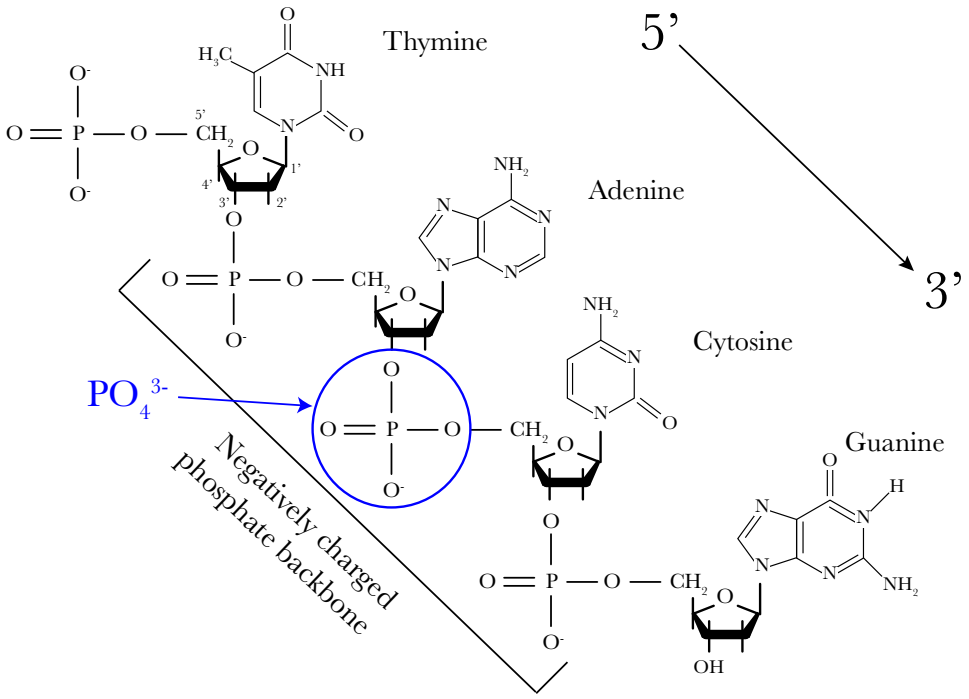

Figure 2.2: **Example of a single strand of DNA.** Sequences are written from 5' to 3'. This sequence would read as TACG. DNA is negatively charged due to the phosphate backbone. Adapted from [20].

ii. The **secondary structure**, refers to the structural motifs.

iii. The **tertiary structure**, indicates the three-dimensional structure of the molecule.

iv. The **quaternary structure**, corresponds to the assembly of different tertiary structures.

2.1.1 *Deoxyribonucleic acid (DNA)*

Eukaryotes[1] store the major part of the DNA inside the cell nucleus and some part in the mitochondria. DNA, together with packaging proteins (histones), form structures called *chromosomes*. They are responsible of carrying the major part of the genetic information of organisms.

The three-dimensional structure (i.e. the tertiary structure) of DNA was finally resolved in 1953 by James Watson and Francis Crick. They

1 Organisms whose cells contain a nucleus, mitochondria and an endomembrane system, dividing the cell into functional compartments (such as the Golgi apparatus and chloroplasts in the case of some plants and algae).

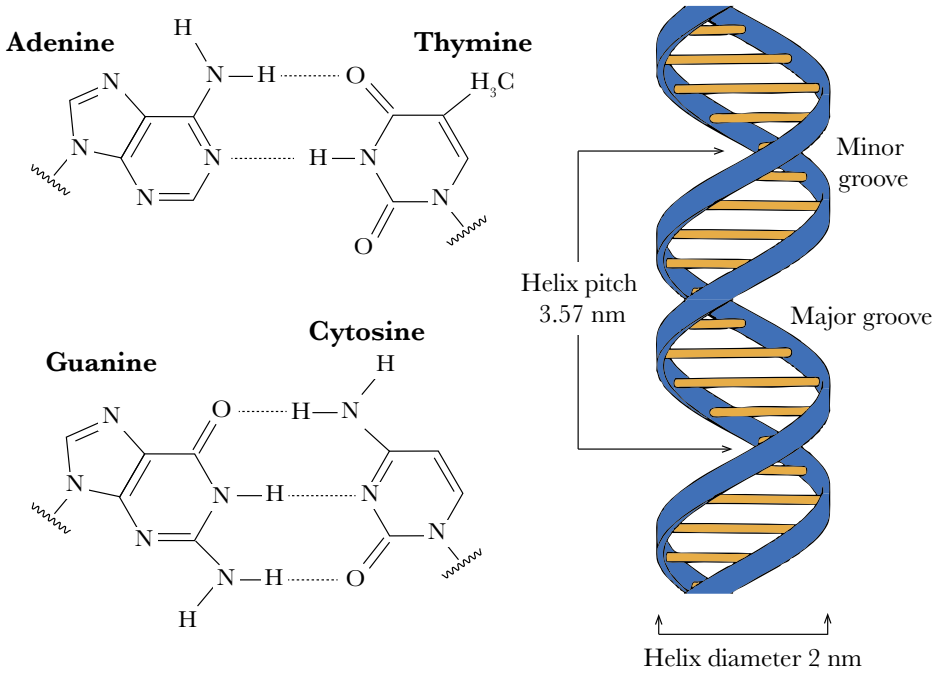

Figure 2.3: **Structure of DNA.** Canonical Watson-Crick basepairs where hydrogen bonds are shown as dashed lines (left figure). Sketch of the helical structure of a DNA molecule and structural parameters (right figure).

showed, thanks to the crucial work of Rosalind Franklin and Maurice Wilkins, how the DNA forms a right-handed helix keeping the negatively charged phosphate backbone in the outer part of the helix (major groove). It is formed by two individual antiparallel strands (i.e. one strand is oriented in the 5' $\rightarrow$ 3' direction whereas the other runs in the 3' $\rightarrow$ 5' direction). Bases are located in the inner part of the helix. Therefore, DNA structure physically protects the nucleobases from the effects of the outer environment. Pairing between bases follow the complementary rules (see Fig. 2.3):

- **A** interacts with **T** via **two hydrogen bonds**: A=T
- **C** interacts with **G** via **three hydrogen bonds**: G≡G

Bases that follow the previous pairing rules are coined as *canonical or Watson-Crick basepairs*. Nevertheless, there can happen other possible non-canonical pairings, such as Hoogsteen or Wobble bonds (for details see Ref. [20]). Despite the weakness of hydrogen bonds (for DNA they

are around $2 - 3$ kcal/mol ($3.4 - 5.1\ k_{\mathrm{B}}T$) [20]), DNA is a quite stable molecule. This is due to the fact that two consecutive bases are able to stack on one another near the center of the helix. Therefore, the cooperative effect of all stacking contributions yield a remarkably stable molecule.

A schematic cartoon of the three-dimensional structure of a right-handed DNA helix is shown in Fig. 2.3. The size of the helix diameter is 2 nm, the axial rise (distance between consecutive bases) is 0.34 nm and the helix pitch equals to 3.57 nm (i.e. 10 basepairs per turn). This is the standard structure of DNA and it is called the B-form. Nevertheless, there are other helical structures, such as the A-form (right-handed helix with an axial rise of 0.21 nm and 11 basepairs per helix turn) and the Z-form (left-handed double helix with a distance of 0.38 nm between consecutive bases and 12 basepairs per turn), that are formed in certain conditions.

2.1.2 *Ribonucleic acid (RNA)*

Structurally the ribonucleic acid (RNA) shares several features with DNA, however it also presents substantial differences. For instance, whereas DNA is usually found in the double stranded conformation forming long molecules, RNA molecules are much shorter and typically single stranded. Despite of that, two single-stranded RNA molecules can 'hybridize'. That is, spontaneously join together to form a double-helical three-dimensional structure [22]. This fact was discovered by Alexander Rich and David R. Davies in 1956, only three years after the resolution of DNA structure by Watson and Crick. On the other hand, ribonucleotides are formed by a pentose sugar containing ribose, a phosphate backbone and a nucleobase (A, G, C or U).

RNA is chemically more reactive as compared to DNA. Indeed, RNA is essentially involved in short-term functions rather than storing long-term information as DNA. It is mainly involved in protein synthesis and regulatory processes. It is present in eukaryotes and prokaryote and in some viruses. The importance of RNA in protein synthesis was discovered by Severo Ochoa (1905 – 1993). There are three main types of RNA and all of them are involved in protein synthesis:

- Messenger RNA (**mRNA**) transfers the information coming from the DNA to a ribosome[2] and acts as a template for the protein that will be synthesized.
- Ribosomal RNA (**rRNA**) is the major constituent of ribosomes.
- Transfer RNA (**tRNA**) is a small RNA molecule that inserts the aminoacids in the correct location of the protein that is being synthesized.

Moreover, many small RNAs have been discovered during the past years. From RNA molecules that parcticipate in gene regulation, such as microRNAs (miRNA) or small interfering RNAs (siRNA), to non-coding RNA molcules, like piwi-interacting RNAs (piRNAs) or RNA thermosensors.

2.2 OPTICAL TWEEZERS

Optics and mechanics are the two oldest scientific disciplines. The study of light was laid in the ancient Greece. Indeed the book *Optics* by Euclid (330 – 275 B.C.) collects all the existing knowledge, to that time, about the geometry of vision (including reflection and diffusion). Euclid's work remained silent until Ibn al-Haytham (Alhazen) (965 – 1040), who was also known as the "father of modern optics", performed a large set of experiments and observations on light reflection and refraction effects using lenses and mirrors. Al-Haytham had a great influence on the later development of modern optics.

The invention of optical microscopes by Zacharias Janssen (1585 – 1632) in 1590 turned out to be a tipping-point in science. The use of microscopes opened up a new scientific dimension, in particular, for biology. It is worthwhile to mention the works of Robert Hooke (1635 – 1703), who, using a microscope, was able to directly observe cells for the first time. Later, Anton van Leeuwenhoek (1632 – 1723) popularised the use of microscopes to view biological structures.

Although being well-known for developing the field of classical mechanics and, together with Leibniz, for developing the infinitesimal calculus, Isaac Newton (1642 – 1727) also made outstanding contributions in optics. As a matter of fact, he collected the existing technology

2 Macromolecular complexes of RNA and proteins

on lenses, prisms, mirrors, telescopes and microscopes. Newton showed that white light is formed by the mixture of different colors with different refractivity. Contemporary to Newton, Christiaan Huygens (1629 – 1695) proposed that light is actually a wave, becoming his theory of light a mainstream topic in physics. Huygens' wave theory experienced a boost thanks to the experimental proofs made by Thomas Young (1773 – 1829) and Augustin Fresnel (1788 - 1827). Optics was later unified to electromagnetism by James Clerk Maxwell (1831 – 1879). Maxwell's equations laid the foundations of modern electrodynamics.

The ultimate boost in optics took place in 1900, when Max Planck finally described the black body radiation showing that the energy exchanges between light and matter only occurs for discrete packages of energy equal to: $h\nu$ (called *quanta*), being h the Planck constant. Albert Einstein (1879 – 1955) proposed the existence of light quanta, the so-called *photons*. The energy of a single photon is given by the Planck–Einstein relation, which reads as:

$$E = h\nu\,, \tag{2.1}$$

where h is the Planck constant and ν is the frequency of the photon. Regardless of having zero mass, photons are full-fledged particles. Indeed, the linear momentum p of a photon can be obtained by combining the Planck–Einstein equation (Eq. (2.1)) with the following relativistic relation:

$$E^2 = m^2c^4 + p^2c^2\,. \tag{2.2}$$

Being m the mass and c the speed of light. Hence, setting $m = 0$, the energy of a photon can be related to its energy as:

$$E = pc\,. \tag{2.3}$$

Therefore, by combining Eq. (2.1) and (2.3), the linear momentum of a photon p yields:

$$p = \frac{h}{\lambda} = \frac{h\nu}{c}\,. \tag{2.4}$$

Einstein's predictions were experimentally verified by Arthur Compton (1892 - 1962) in 1923. Since then, thanks to the development of quantum mechanics, we know that undulatory and corpuscular properties of light will everlastingly walk hand by hand.

The full comprehension of the interaction between light and matter yielded the development of quantum optics. This brand new field of science aroused remarkable interest thanks to the inventions of the maser (acronym for Microwave Amplification by Stimulated Emission of Radiation) in 1954 and the laser (acronym for Light Amplification by Stimulated Emission of Radiation) in 1960. The physical principle underlying both devices is the phenomenon of stimulated emission, stated by Einstein in 1916. Laser rapidly emerged as a perfect tool for addressing new and challenging physical problems and for designing sophisticated instrumental devices.

2.2.1 *Principles of optical trapping*

As we introduced, if a ray of light is composed by N photons, carries a linear momentum proportional to the energy of each photon as:

$$\mathbf{p} = N\frac{h\nu}{c}\hat{\mathbf{u}}_p\,, \tag{2.5}$$

where $\hat{\mathbf{u}}_p$ is a unit vector that indicates the direction of the linear momentum (i.e. parallel to the direction of propagation of the ray). The conservation of momentum is the physical principle of optical trapping.

Let us consider an incoming ray of light from a laser which has a Gaussian intensity profile[3] and it interacts with a transparent bead. When the laser beam reaches the object, the light rays are deflected according to the laws of reflection and refraction. There are two reasons that explain the change of the linear momentum in Eq. (2.5): first, according to the Snell law, rays are refracted, entailing a change of $\hat{u}_p$; second, since there is also reflection on the bead interfaces, the number of incident photons on a surface is not equal to the number of exiting photons. Both ingredients may yield a substantial change in the linear momentum of light.

3 Roughly speaking, the intensity of the light is higher at the center of the beam than at the edges.

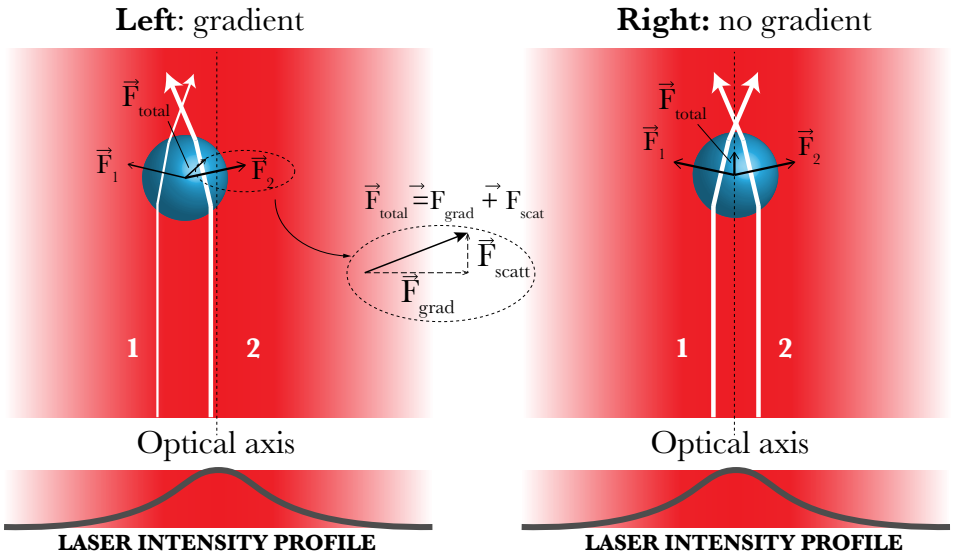

Figure 2.4: **Effect of gradient and scattering forces.** Ray optics sketch of the effect of scattering (F_{scatt}) and gradient forces (F_{grad}) when a Gaussian laser beam interacts with a transparent bead. In the situation depicted in the left figure, gradient force pulls the bead towards the center of the laser beam, whereas in the situation corresponding to the left figure, there is no gradient force.

In particular, the resultant force ($\vec{F}_{\text{total}}$) from all such rays can be separated into two components: the scattering force and the gradient force. The scattering force points in the same direction of the incident light and tends to push the bead away the light source. On the other hand, the scattering force is due to the fact that the light near the center of the Gaussian beam is more intense than the light of the edges. Therefore, this generates an extra component for the change of the linear momentum of the bead that points towards the center of the beam. Indeed, gradient force is a restoring force that pulls the bead to the most intense region of the beam. In Fig. 2.4 there are shown two schematic depictions of the forces that undergoes a bead that is displaced from the beam center (left figure) and one located in the beam center (right figure). Whilst in the first case the gradient force is intense and tends to move the bead towards the center, in the second situation the gradient force does not play a significant role. Therefore the bead only feels the effect of the scattering force.

Forces due to radiation are small and their effects are only relevant in microscopic scales. In order to illustrate this, let us consider a macroscopic mirror which is illuminated by a regular 60W light bulb.

Assuming all light rays are parallel and interact perpendicularly to the surface of the mirror, the net force due on the mirror to total reflection can be obtained from Eq. (2.3) as:

$$F = \frac{dp}{dt} = \frac{2}{c} W\,, \tag{2.6}$$

where W is the power of the light and the 2 prefactor takes into account for the fact that the net force is due to the incident and reflected photons by the mirror. According to 2.6 the force that the mirror feels is of order 10^{-7} N = 100 nN. Hence, radiation-due forces are only sensitive for microscopic objects (with a mass lower than, approximately, 10^{-2} μg).

The relevance of radiation pressure for tiny objects was noticed, for the first time, by Arthur Ashkin in 1970 [23]. He experimentally tested the possibility of accelerating micron-sized particles (of a radius of 2.68 μm) by radiation pressure forces using a 514.5 nm wavelength laser with milliwatt power. Additionally, a few years later, in other experiments he demonstrated the possibility of trapping and levitating micrometric objects (for a thorough review see [24]).

Later, in 1986, Ashkin and collaborators used a lens in order to focus a 514.5 nm wavelength laser, showing how an *optical trap* emerges near to the focusing point [25]. As a matter of fact, they were able to trap particles with sizes that were four order of magnitudes different (from 25 nm up to 10 μm). They realized that traditional geometrical optics does not properly describe the optical trapping phenomenon when the size (diameter) of the object is similar (or smaller) to the wavelength of the used light[4]. Ashkin's groundbreaking invention was called *optical tweezers* and triggered a revolution in biophysics [28] and nanotechnology [29]. Arthur Ashkin was awarded the 2018 Nobel Prize in Physics for his invention of optical tweezers.

Since the typical working forces of optical tweezers are in the order of piconewtons (i.e. 10^{-12} N) and they are designed with the capability of measuring nanometric displacements, optical tweezers have became the perfect tool to explore weak molecular forces.

4 For a detailed discussion of the existing regimes: the Mie regime (i.e. the size of the objects is larger than the light wavelength) and the Rayleigh regime (the size of the objects is smaller than the light wavelength) check Refs. [26, 27]

2.2.2 *The miniTweezers setup*

All the experimental results regarding LOT data of this thesis were obtained using the miniTweezers optical tweezers setup. It was designed by Steve B. Smith and Carlos Bustamante in 2003 [30]. MiniTweezers allow for a direct force measurements, resulting in a clear advantage as compared to most LOT setups. Moreover, miniTweezers is an instrument capable of reaching piconewtons, sub-nanometer and millisecond resolution.

MiniTweezers consist of two counter-propagating 845 nm wavelength laser beams focused into a spot located inside a microfluidics chamber. The beams are generated by two 200 mW laser diodes and the focusing is done by means of two high numerical aperture objective lenses (Olympus UPlanSApo 60x/1.20). The location of the optical trap, and hence the deflection of the laser beams, is set by means of two piezoelectric actuators (fiber wigglers) located on the tip of the laser fibers. A simplified scheme of the miniTweezers instrument is shown in Fig. 2.5. In what follows, we briefly describe the optical path followed by one laser (the other laser follows the same path but in opposite direction) and the details of the microfluidics chamber.

Optical path prior to entering the microfluidics chamber

The light emitted by the 845 nm wavelength laser diode is directed through a single mode optical fiber. The laser position is controlled by a fiber wiggler capable of reaching high response frequencies (> 2 kHz) and a position range of, approximately, 11 μm. As soon as the laser exits the optical fiber, a $\sim$ 4 % of the light is reflected by means of a pellicle mirror and sent to a position-photo sensitive detector (PSD) that measures the position of the optical trap. The remaining $\sim$ 96 % of the light is collimated with a lens before it reaches a polarizing beam splitter (PBS). Thanks to the PBS, the laser beam is completely reflected and linearly polarized. Afterwards, the beam passes through a quarter-wave plate, becoming now circularly polarized. Then, the beam is focused by means of a microscope objective lens to a focal point located inside a microfluidics chamber (where the optical trap is generated).

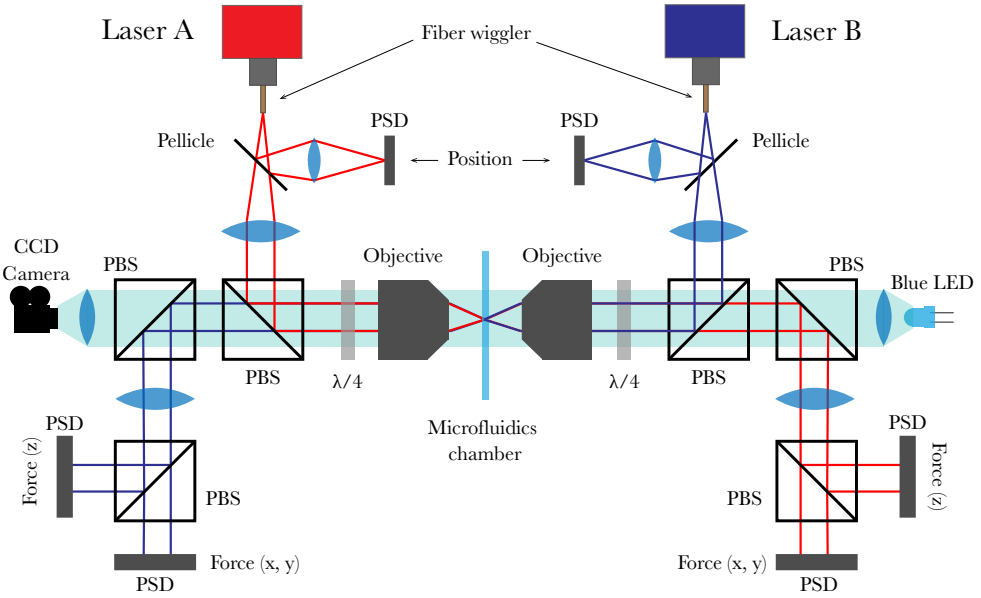

Figure 2.5: **MiniTweezers schematics** The optical path of the counter propagating laser beams are shown as blue and red lines. Images of the microfluidics chamber are obtained using a CCD camera, for which a blue LED is used as the source of illumination in a Köhler configuration.

Optical path after exiting the microfluidics chamber

After coming out from the microfluidics chamber, the laser light is collected by a second microscope objective and it is immediately restored to linear polarization conditions thanks to the action of another quarter-wave plate. The linearly polarized beam passes through a PBS and it is entirely transmitted to a second PBS that changes the direction of propagation of the beam. After traversing a relay lens it is redirected to a final PBS that splits the beam into two perpendicular beams. The laser beam that has not changed its direction of propagation is finally directed to a PSD that directly measures the (x, y) components of the force exerted by the optical trap. The $z-$force is obtained by measuring, by means of a PSD, the diameter of the laser beam.

Microfluidics chamber

The microfluidics chamber is the physical site at which single-molecule measurements are actually done. It is inserted between the two micro-

scope objectives (see Fig. 2.5). A sketch of the internal setup of the microfluidics chamber can be seen in Fig. 2.6.

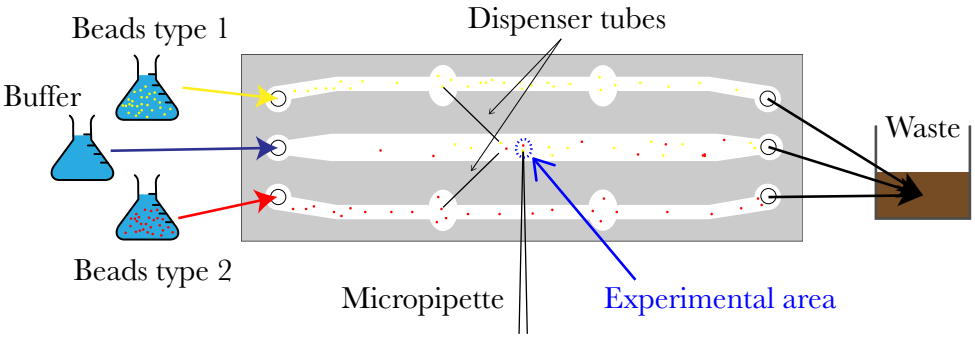

Figure 2.6: **Schematics of the microfluidics chamber.** It is composed by three microfluidic channels. The two lateral channels are conected to the central one by two glass dispenser tubes of 0.040(6) mm of internal diameter. Experiments are performed in the dashed blue region, close to the tip of the micropipette.

Microfluidics chambers are handmade by joining two glass surfaces (coverslips) and a plastic paraffin film (parafilm) layer forming a sandwich-like structure. The parafilm (grey zone of Fig. 2.6) is cut using a laser in order to create the three-channel shape inner structure of the microfluidics chamber. The lateral channels are connected to the central channel using two glass dispenser tubes (Garner Glass CO.) with an outer diameter of 0.10 ± 0.01 mm and an inner diameter of 0.040 ± 0.006 mm. A glass pipette with a tip of size $\sim \mu$ is made using a pipette puller. The glass micropipette is manually inserted before glueing both glass coverslips.

In the central channel it is flowed buffer[5], while in the lateral channels it is flown a mixture of buffer and beads. In one of the lateral channels it is flowed a mixture of buffer and beads type 1 (e.g. streptavidin-coated beads). Part of the beads eventually end in the central channel, where one of them is captured by the optical trap and later lead to the tip of the micropipette, where it is held by air suction. On the other hand, in the other lateral channel, it is flowed a mixture of beads type 2 (e.g. antidigoxigenin-coated beads) with the molecule of interest attached to its surface and buffer. Some of these beads reach the central channel through the dispenser tube (Fig. 2.6) and one is captured by the optical trap and, subsequently, approached to the experimental area located

5 Buffer depends on the experiment.

close to the tip of the pipette by moving the trap. Experiments are carried out in the experimental area of Fig. 2.6.

It is important to recall that the optical trap is generated in the central region of the microfluidics chamber, far enough from the surfaces in order to avoid undesired hydrodynamic effects on the trapped objects.

Force measurement using miniTweezers

The main advantage of miniTweezers instrument is the possibility of directly measuring the mechanical force by means of the linear momentum of light. Indeed, the linear momentum of the light that forms the optical trap is measured by means of PSD detectors. PSD values are related to the real mechanical force f (plus a force offset) in the x (y) direction through:

$$f_{x(y)} = C_{x(y)}\mathrm{PSD}_{x(y)}\,, \tag{2.7}$$

where $C_{x(y)}$ is a calibration factor and $\mathrm{PSD}_{x(y)}$ is the sum of the PSD values of both lasers (in analog to digital units so that the product $C_{x(y)}\mathrm{PSD}_{x(y)}$ equals a force). The calibration factors $C_{x(y)}$ are independent of the performed experiment, therefore their precise measurement is done by applying a well-known force on a trapped bead and measuring the PSD response. In what follows we briefly describe several force calibration methods that are typically used in the miniTweezers instrument.

- **Stokes' law** relates the drag force f_d that undergoes a small spherical object (of radius R) moving through a fluid with a shear viscosity μ at constant speed v:

  $$f_\mathrm{d} = \gamma v = 6\pi\mu R v\,, \tag{2.8}$$

  being $\gamma = 6\pi\mu R$ the viscosity coefficient. Stokes' law is valid in laminar flows, for homogeneous objects and in absence of hydrodynamic interactions. Hence, using distilled water (whose shear viscosity has been accurately measured) and particles with a known radius, by moving the microfluidics chamber at velocity v, it is possible to use Eq. (2.8) to obtain the mechanical force as

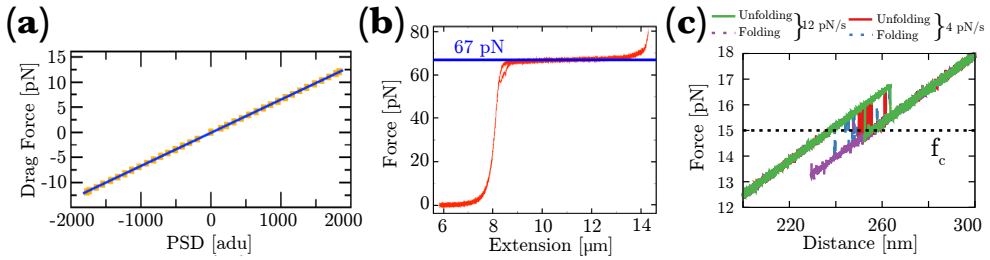

Figure 2.7: **Force calibration methods. (a)** - Stokes' law calibration. Drag force (obtained using Eq. (2.8)) as a function of PSD values. Orange squares are experimental measurements and the blue solid line a linear fit yielding the calibration factor. **(b)** - Overstretching transition of a λ-DNA molecule. It occurs around 67 pN. Previous figures have been obtained from Ref. [26]. **(c)** - Unfolding-folding cycles on CD_4 DNA at different pulling speeds. Coexistence force (f_c) is around 15 pN.

a function of v. Then, by considering the linear relation between the force and the PSD values (i.e. Eq. (2.7)), the calibration factors $C_{x(y)}$ can be finally obtained, as can be seen in Fig. 2.7(a).

- The **overstretching transition of a B-DNA molecule** is commonly accepted to occur at 65 - 67 pN at room temperature and 500 mM NaCl [31] (see Fig. 2.7(b)). It is attributed to a conformational change of the DNA molecule, where it changes from the B-DNA conformation to the S-DNA conformation (see Sec. 2.1.1). As a result at 67 pN, with a little increase in the force, the double-stranded DNA (dsDNA) extension is increased by around 1.7 times its contour length. Therefore, due to the reproducibility of this phenomenon, the force value at which the overstretching transition occurs is usually taken as a referential value.

- The coexistence force of the **20 basepairs CD_4-DNA molecule** falls between 14 - 16 pN at room temperature and 1M NaCl [32] (see Fig. 2.7(c)). The coexistence force corresponds to the force value at which the molecule can be found in the folded state or in the folded state with equal probability. This procedure has been recently developed in our lab and due its reproducibility has become a benchmark when calibrating miniTweezers.

These three methods are currently the most used when calibrating miniTweezers due to their straightforward implementation and the possibility of performing cross-check between them.

Stiffness of miniTweezers

A key parameter in optical tweezers instruments is the trap rigidity or trap stiffness. For the usual range of forces or small displacements, the optical trap behaves as an harmonic potential ($U_{\mathrm{ot}} = \frac{1}{2}k_{\mathrm{b}}x^2$). Therefore, in the overdamped limit, the Langevin equation that governs the motion of a Brownian particle (i.e. the bead) in a harmonic potential is:

$$\gamma\frac{dx}{dt} = -k_{\mathrm{b}}x + \eta(t)\,, \tag{2.9}$$

where x is the relative distance of the position of the bead to the center of the optical trap (where the force is zero), k_{b} is the stiffness of the optical trap and $\eta(t)$ is the stochastic force acting on the bead due to Brownian fluctuations. By performing a Fourier transform on the force correlation function obtained by solving Eq. (2.9), the so-called power spectrum of the force $S_f(\nu)$ is obtained, yielding:

$$S_f(\nu) = \frac{k_{\mathrm{B}}T}{2\pi^2\gamma}\frac{k_{\mathrm{b}}^2}{\nu^2+\nu_c^2}\,, \tag{2.10}$$

where ν_c is the so-called corner frequency ($\nu_c = k_{\mathrm{b}}/(2\pi\gamma)$). The trap stiffness can be obtained as follows: first, record the Brownian-induced displacements of a trapped bead conducting high bandwidth measurements (e.g. 50 kHz). Next, compute the power spectrum by using a fast Fourier transform algorithm (FFT) on the force correlation function. Finally, using a Levenberg-Marquadt algorithm fit the value of k_{b} using Eq. (2.10). Figure 2.8 shows the power spectrum of a bead captured in the optical trap at zero force and a fit to Eq. (2.10). For the case shown, we get: $k_{\mathrm{b}} = 0.079 \pm 0.010$ pN/nm and $\gamma = (3.1 \pm 0.1) \times 10^{-5}$ pN·s/nm.

2.3 MAGNETIC TWEEZERS

Force spectroscopy techniques are, nowadays, one of the most powerful approaches to study the behavior of single molecules, macro-

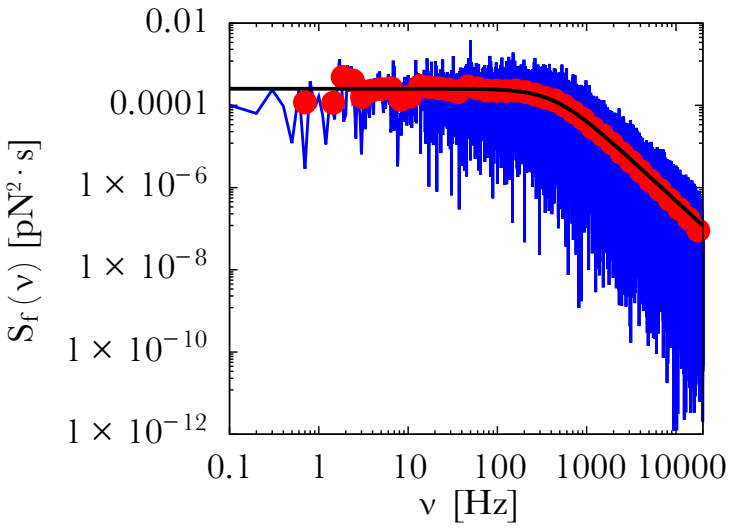

Figure 2.8: **Power spectrum of a bead in the optical trap at zero force.** Raw data shown as blue lines, the exponential average are the red circles and the fit to Eq. (2.10) is the black solid line.

molecular assemblies and cells, with applicable results in the field of non-equilibrium statistical physics or biophysics. Soon thereafter the invention of optical tweezers, MT were developed to manipulate paramagnetic objects using magnetic field gradients.

The first MT devices were developed in the 1990s and their usefulness was originally similar to the then-emerging LOT instruments. As LOT they allow applying piconewton forces and measuring displacements in the scale of the nanometer. First MT assays provided insights on the elastic response of DNA [33, 34]. In MT setup, a superparamagnetic bead is held inside a magnetic field gradient generated by a pair of permanent magnets. Then, by controlling the position of the magnets' stage, forces can be directly modulated. Therefore, MT naturally operate in the force controlled scheme, without a force feedback.

2.3.1 *Physical principles of magnetic tweezers*

The key ingredients of MT are the use of *superparamagnetic beads* and *magnetic field gradients*. In what follows, we briefly describe each element.

Superparamagnetism

Superparamagnetism is a form of magnetism that displays some features of ferromagnetism and paramagnetism appearing in sufficiently small ferromagnetic particles. Usually, the lowest energy state of ferromagnetic samples (in absence of an external field) corresponds to the demagnetized state. This is not the case of small magnetic systems, where the energetic cost to form domain walls is larger than volume energies[6], therefore they will present no domain walls, behaving like a small permanent magnet or a single big magnetic moment (sum of all the individual magnetic moments carried by the atoms of the nanoparticle). In order to illustrate this fact, let us consider two of the possible magnetization configurations for a small spherical ferromagnetic particle shown in Fig. 2.9.

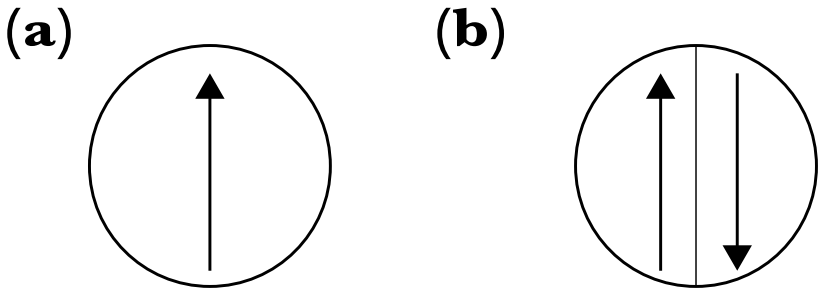

Figure 2.9: **Two possible configurations of a spherical ferromagnetic particle.** Arrows indicate the direction of the magnetic moment. Vertical line of right panel is a Bloch wall, creating two magnetic domains inside the particle. Example reproduced from Ref. [35].

The energy of the (a)-configuration, formed by a single magnetic domain equals to:

$$E_{(a)} = -\mu_0 \int_V d^3r\, \mathbf{M} \cdot \mathbf{H}_a - \frac{\mu_0}{2} \int_V d^3r\, \mathbf{M} \cdot \mathbf{H}_d\,, \tag{2.11}$$

where μ_0 is the magnetic permeability of free space, $\mathbf{M}$ the magnetization, $\mathbf{H}_a$ the applied magnetic field and $\mathbf{H}_d$ the so-called demagnetizing field (i.e. the field generated by the own magnetization of the magnet). We note that both integrals are done over the volume V of the magnet. The first term of Eq. (2.11) is the Zeeman energy (i.e. the energy of a

6 Surface energies $\sim$ (size)^2 and volume energies $\sim$ (size)^3.

magnetized system inside an external field), whereas the second term is the energy due to the own magnetic field of the system (magnetostatic energy or dipolar energy). Considering a uniformly magnetized ferromagnetic sphere of radius R, the demagnetizing fields is equal to $H_d = -M/3$ [35]. Moreover, when there is no external magnetic field, $H_a = 0$, Eq. (2.11) becomes:

$$E_{(a)} = \frac{2}{9}\mu_0 \pi M^2 R^3 \,. \tag{2.12}$$

On the other hand, the energy of the (b)-configuration will have two contributions. First, it will contain a magnetostatic-like term yet taking into account that every domain occupies one half of the total volume of the magnet. Finally, it will have an additional contribution due to the formation of the domain wall (see Fig. 2.9(b)). Unlike magnetostatic energies, domain wall formation energies are proportional to the magnets' surface. Hence, putting all the pieces together, the energy yields:

$$E_{(b)} = \frac{1}{9}\mu_0 \pi M^2 R^3 + \pi R^2 \epsilon \,, \tag{2.13}$$

where the term ϵ is the energy of formation of a vertical domain wall per unit area. For sufficiently small particles, the (a)-configuration is more favourable than the (b)-configuration. The critical radius R_c can be found by investigating the regime where $E_{(a)} < E_{(b)}$, obtaining:

$$R_c = \frac{9\epsilon}{\mu_0 M^2} \,. \tag{2.14}$$

Then, for particle radius $R < R_c$, the (a)-configuration becomes more stable than the (b)-configuration. Considering $\epsilon \sim 10^{-2}$ Jm^{-2} and $\mu_0 M \sim 1$ T, the critical radius is around 0.1 μm. We emphasize that the critical size can be tuned by using different magnetic materials.

Another key feature of superparamagnetism is the fact that the average magnetization of superparamagnetic beads is zero. Therefore, by terms of an external magnetic field they can become magnetized, as it happens in a paramagnetic system[7]. This fact can be understood

7 However, their magnetic susceptibility is much higher as compared to typical paramagnetic systems.

by considering the characteristic time τ at which magnetic moments spontaneously flip due to thermal fluctuations. It is given by the Néel relaxation time, τ_N:

$$\tau_N = \tau_0 \exp\left(\frac{KV}{k_B T}\right), \tag{2.15}$$

where τ_0 is the attempt time that is material-dependent, V is the volume and K the so-called anisotropy constant, being the product KV the energy barrier associated with the change in the magnetization. For bulk systems, the exponential dependence on the volume makes the flipping probability negligible as compared to small particles. By setting $\tau_0 \sim 10^{-9}$ s and considering small particles where the product $KV \ll k_B T$, the relaxation time can be as small as few nanoseconds. Therefore, for typical measurement times t (around milliseconds) we have $t \gg \tau_N$. Hence, the magnetization will flip several times during the measurement (i.e. the average magnetization will be zero), but not in the presence of an external field.

Magnetic fields

Now we will consider the effects of a magnetic field $\mathbf{B}$ when a superparamagnetic (or paramagnetic) bead is placed inside the field. The bead will become magnetized with a net magnetic moment equal to $\mathbf{m}(\mathbf{B})$ and its energy due to the presence of the magnetic field will be [36]:

$$U = -\frac{1}{2}\mathbf{m}(\mathbf{B}) \cdot \mathbf{B}\,. \tag{2.16}$$

Therefore, the bead will experience a force that is given by the gradient of the energy with a minus sign:

$$\mathbf{f} = -\boldsymbol{\nabla} U = \frac{1}{2}\boldsymbol{\nabla}\left(\mathbf{m}(\mathbf{B}) \cdot \mathbf{B}\right). \tag{2.17}$$

Besides, the bead not only feels a force that points in the same direction than the gradient of the magnetic field, but also a torque $\boldsymbol{\Pi}$ given by:

$$\boldsymbol{\Pi} = \mathbf{m}(\mathbf{B}) \times \mathbf{B}\,. \tag{2.18}$$

The magnetization of the beads as a function of the applied field is well described by the Langevin function:

$$M(B) = M_{\text{sat.}} \left(\coth\left(\frac{B}{B_0}\right) - \frac{1}{\frac{B}{B_0}} \right), \tag{2.19}$$

where $M_{\text{sat.}}$ is the saturation magnetization. For low fields, the magnetization of the beads is linear with the applied field, yielding a force: $\mathbf{f} \propto \boldsymbol{\nabla}|\boldsymbol{B}|^2$. However, for high enough fields the magnetization quickly tends to the saturation value, $M_{\text{sat.}}$.

On the other hand, the magnetic field of a permanent magnet (like the ones used in the MT setup) can be computed using the Biot-Savart law using the method of equivalent currents[37]:

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \mathbf{I}_{\text{equi}} \frac{\hat{\mathbf{dl}} \times \hat{\mathbf{r}}}{r^2}, \tag{2.20}$$

where $\mathbf{I}_{\text{equi.}} = \mathbf{M} \times \hat{\mathbf{n}}$, being $\hat{\mathbf{n}}$ the surface normal unitary vector, $\hat{\mathbf{l}}$ a unit vector pointing in the direction of the equivalent current (for details see Ref. [38]) and $\hat{\mathbf{r}}$ the unitary vector that points from the equivalent current to the point at which the magnetic field is calculated. We note that Eq. (2.20) cannot be analytically solved in general since it depends on the magnet geometry.

2.3.2 *The picoTwist setup*

The experiments involving MT throughout this thesis were done using a picoTwist instrument developed by Gosse and Croquette in 2002 [39]. PicoTwist is a very low drift and robust MT apparatus. Additionally, picoTwist is easily portable and plug-and-play. Unlike optical tweezers, where temperature control is harder and needs for a specific setup [40], a single Peltier system allows for direct temperature control in picoTwist. Also, it allows the measurement of up to 100 molecules in parallel. In addition, since picoTwist uses a pair of magnets, it is possible to exert torques on individual molecules, becoming the perfect tool for the study of molecular motors that modify the DNA topology (such as topoisomerases) or the coiling of biological molecules.

A schematic illustration of picoTwist is shown in Fig. 2.10. The sample (magnetic bead + molecule) is flowed, together with reagents,

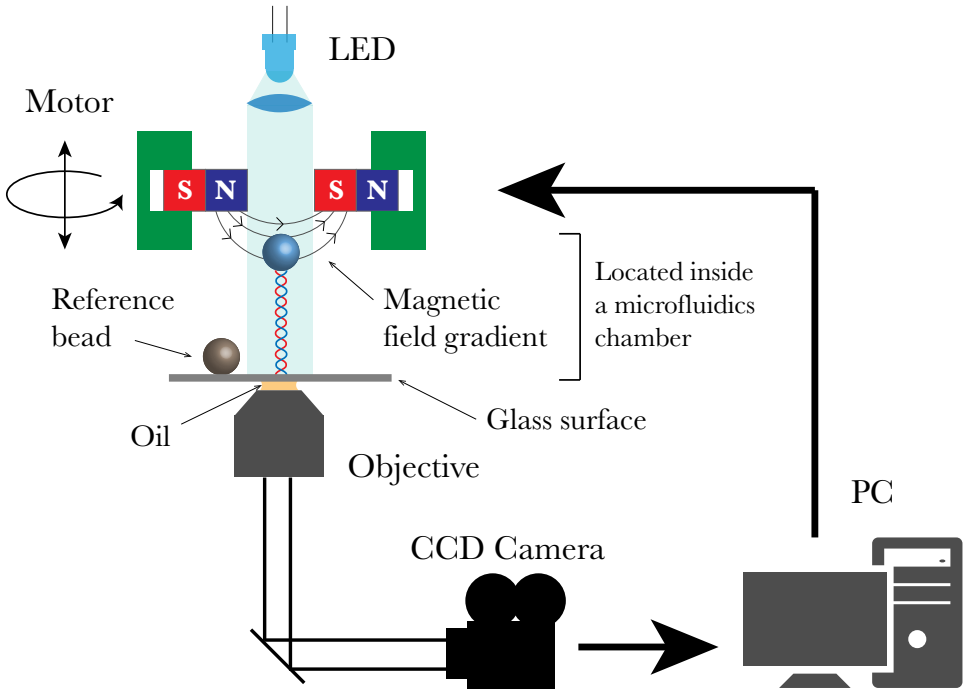

Figure 2.10: **Schematic diagram of picoTwist.** The superparmagnetic bead(s) is(are) located within a strong magnetic field gradient. A reference bead is fixed at the surface via non-specific interactions. In order to measure the position of the pulled beads, a differential measurement with the fixed bead is done to reduce drift effects. Magnets can be either translated or rotated by means of piezoelectric actuators that are externally controlled. The computer processes the images obtained by the CCD camera in order to get the positions of the beads in real time via an analysis of the diffraction rings. A LED illuminates the sample through the gap between the magnets.

inside a monochannel microfluidics chamber. In what follows we briefly describe the individual parts of the setup.

Magnetic trap and force measurement

The magnetic trap is generated by a pair of small rare earth permanent magnets (NdFeB). While the generated magnetic field is horizontal, the magnetic field gradient is vertical and so the force, according to Eq. (2.17). With picoTwist we are able to exert forces down to 50 fN up to $\sim$ 20 pN [39] with standard $\sim$ 1 μm beads (or $\sim$ 100 pN using 2.7 μm beads). The maximum achievable force can be increased by using bigger beads.

Unlike the employed LOT instrument in this thesis (miniTweezers, described in Sec. 2.2.2), picoTwist does not allow of a direct measurement of the mechanical force. This is due to the intrinsic variability in the magnetization of commercial beads [41]. Therefore, the force measurement is done by relating the thermal motion of the bead at several positions of the magnets and constructing an empirical law for the force as a function of the magnets' position, $f(z_{\text{mag.}})$. In particular, a magnetic bead that is anchored to a surface through a molecule of length l (see Fig. 2.11), behaves like a pendulum feeling a vertical force due to the presence of the magnetic field.

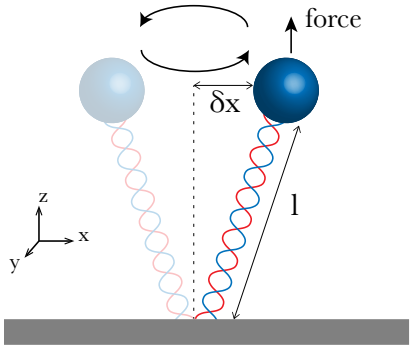

Figure 2.11: **Force measurement in magnetic tweezers.** A magnetic bead tethered to a molecule is undergoes transverse Brownian fluctuations and it feels a vertical force due to the presence of a magnetic field. The magnetic field gradient goes in the z-direction (not shown).

Therefore, the mean transverse position fluctuations, $\langle \delta x^2 \rangle$, are used to characterize the mechanical force as:

$$f = \frac{k_{\text{B}} T}{\langle \delta x^2 \rangle} l \,, \tag{2.21}$$

where there has been used the law of Equipartition and the fact that f/l is the lateral stiffness. Then the force vs. magnets position profile can be characterized[8]. It has been shown that the $f(z_{mag.})$ curve is well fitted by an exponential function as follows:

$$f(z_{\text{mag.}}) = f_{\text{max}} \exp\left(-a z_{\text{mag.}} + b z_{\text{mag.}}^2\right) \,, \tag{2.22}$$

8 A more accurate force calibration is done in the frequency domain by means of the power-spectrum. For a detailed discussion see Ref. [41].

where it has been found that, for typical $\sim$ 1 μm beads, $f_{\text{max.}}$ is around 21 pN, $a = 3.53\ \text{mm}^{-1}$ and $b = 0.66\ \text{mm}^{-2}$ [42]. The typical position range[9] of the magnets span from zero (touching the outer layer of the microfluidics chamber) up to few millimetres. The dominant term in the aforementioned distance range is $az_{\text{mag.}}$ (Eq. (2.22)) being a of order 1 mm^{-1}. Therefore, the position of the magnets, $z_{\text{mag.}}$, is insensitive to changes of the position of the bead due to conformational changes of molecules (e.g. unfolding-folding transitions), which are typically in the nm - μm range. Hence, MT are high-precise natural force clamps with typical stiffnesses of $\sim 10^{-4}$ pN/nm [39].

Bead tracking and extension determination

The positions of the magnetic beads are recorded by a videotracking system. The volume sample is vertically illuminated by using a LED as a parallel light source. Then, since the size of the beads is comparable with the wavelength of the incident light, significant diffraction effects appear. Beads are observed by an oil immersion objective located under the tethering surface and diffraction pattern are recorded by means of a 60 Hz 576p CCD camera (see Fig. 2.10) and subsequently analysed by a computer program. Examples of diffraction rings are shown in Fig. 2.12.

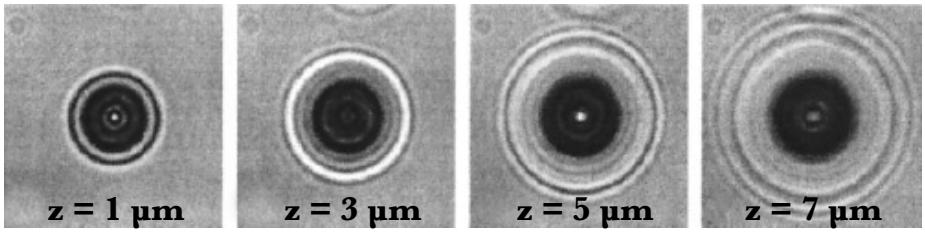

Figure 2.12: **Diffraction rings of a 4.5 μm magnetic bead.** Diffraction rings are significantly different depending on the position of the microscope focus plane, z. Figure adapted from [39].

The center of the diffraction rings correspond to the tethering point of the magnetic bead to the surface. On the other hand, the vertical position of the pulled beads is determined by comparing to a set of reference beads that are kept immobilized on the surface. The calibration procedure is done prior to the actual experimental essay. It consists

9 For typical experimental purposes.

on obtaining a calibration image by translating the objective (i.e. the focal plane) by means of piezoelectric elements at known positions. This protocol is done, typically, from 0 to 10 μm height and with a characteristic resolution of 2 - 10 nm [43].

The position and, subsequently, molecular extension determination is done in real time by using a customized software. For each frame the x, y, z positions are determined for one or several beads simultaneously. By using a static reference to measure positions, drift effects are significantly reduced. Nevertheless, as a main drawback of picoTwist, it is fair to mention that spatial and time resolution is poorer than other techniques that do not rely on image tracking.

2.4 EXPERIMENTAL CONFIGURATION

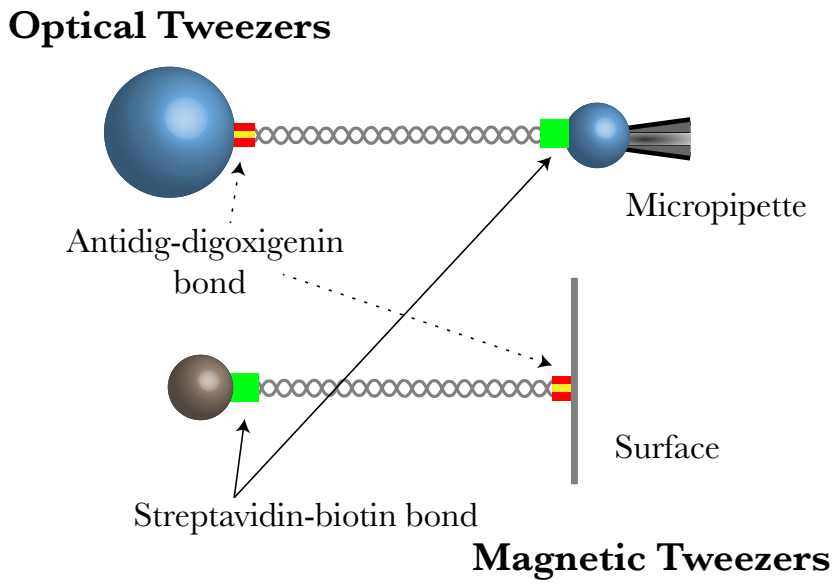

Figure 2.13: **Experimental configuration and schematics of bead coating.** Experimental setup for LOT (top) and MT (bottom) experiments.

Regardless they were performed using LOT or MT, the common denominator of the experiments done in the course of this thesis, is the use of functionalized micrometer-sized beads. In particular, beads are coated with a chemical substance that can specifically bind to its complementary molecule. This is the case, for instance, of streptavidin, a protein that binds via non-covalent bonds to biotin, a vitamine. By la-

belling the interested molecule (DNA or RNA) with a biotin at one of its ends, it can bind to the streptavidin-coated beads through a strong non-covalent bond. Streptavidin-coated beads are used both in LOT and MT experiments. Nevertheless, in this latter case beads have, furthermore, superparamagnetic characteristics.

On the other hand, the remaining free end of the molecule is biochemically modified in order to recognize another molecular complex and avoid undesired bead attachments. For the case of LOT experiments, molecules are merged with a tail of digoxigenins so now they bind to an anti-digoxigenin functionalized bead via a high specific antigen-antibody bond. For the case of MT experiments, there is no second bead. In place, the bottom surface of the microfluidics chamber in which the experiments are performed is also functionalized with digoxigenin antibody, as the second bead in LOT (details can be found in Ref. [42]).

3

BASICS OF STATISTICAL MECHANICS OF POLYMER ENSEMBLES

3.1 GIBBS AND HELMHOLTZ STATISTICAL ENSEMBLES

In statistical mechanics, the thermodynamic behavior of physical systems is obtained by modelling, through probability theory and statistics, the dynamics of fluctuating microscopic states. A macroscopic state is specified by the values of a set of measurable physical parameters, like temperature or pressure. Such quantities are externally imposed to the system and they do not fluctuate in time. Each set of controlled parameters defines a statistical ensemble. For instance, when a gas formed by N particles is in contact with a heat bath at a given temperature T and it is kept at a fixed volume V, its equilibrium macroscopic state is described by these three quantities. This situation corresponds to the so-called *canonical ensemble*, NVT ensemble or Helmholtz ensemble.

The connection with thermodynamics is done by defining thermodynamic potentials that are, in general, functions of the k controlled parameters: $\Psi\left(X_1, \cdots, X_k\right)$. Their successive derivatives, $\partial_{X_i}\Psi$, are physically measurable quantities.

Also, in any ensemble, a *conjugate statistical ensemble* is generated by Legendre transforming the thermodynamic potential Ψ via a pair of conjugate variables[1], X_i and $\partial_{X_i}\Psi$ (with $i \in \{1, \cdots, k\}$) [44]. In particular, choosing the volume V and pressure $P = -\left(\frac{\partial F}{\partial V}\right)_T$ as conjugate pairs, the NPT ensemble (or Gibbs ensemble) is generated. The thermodynamic potential associated with the NPT ensemble is the $G\left(N, P, T\right)$ potential or Gibbs free energy:

$$G\left(N, P, T\right) = F\left(N, V, T\right) + PV\,, \tag{3.1}$$

1 Conjugate pairs are conjugate are conjugate variables with respect to energy.

where $F(N,V,T)$ is the thermodynamic potential associated to the canonical ensemble or Helmholtz free energy defined as:

$$F(N,V,T) = -k_{\mathrm{B}}T \log Z(N,V,T) \ , \tag{3.2}$$

where $Z(N,V,T)$ is the canonical partition function of the system. It is important to remark, though, than in the NPT ensemble (i.e. Eq. (3.1)) the volume V stands for the ensemble average of the volume of the system, because in the NPT ensemble the pressure, rather than the volume, is fixed.

Partition functions of conjugate ensembles are related by Laplace transformations. Indeed, the partition function of the NPT ensemble, $\Xi(N,P,T)$, is the Laplace transform of the canonical partition function $Z(N,V,T)$:

$$\Xi(N,P,T) = \int_0^\infty dV \, Z(N,V,T) \, e^{-PV/k_{\mathrm{B}}T} \ . \tag{3.3}$$

In bulk systems, the equation of state does not depend on the ensemble (in the thermodynamic limit). Hence, conjugate statistical ensembles are *equivalent*. For instance, the equation of state of a gas in a piston at fixed volume yield the same result than if the applied pressure is controlled. Since fluctuations of the uncontrolled variables vanish in the thermodynamic limit, the outcome of physical measures done on macroscopic systems are not able to distinguish a situation where, for instance, the pressure or the volume is fixed.

3.2 CHAIN ENSEMBLES

Due to the development of micromanipulation techniques, the issue of ensemble inequivalence has become a hot topic in polymer systems, such as synthetic polymers (e.g. as polyethylene or synthetic rubber) or biopolymers (e.g. as nucleic acids or proteins). These techniques allow researchers to observe the behavior corresponding to different statistical ensembles. For instance, one can work in the Helmholtz ensemble or extensional ensemble (hereafter referred to as Extensional Ensemble (ExtEns)) by tethering a single polymer between two points, keeping its extension fixed. On the other hand, if a constant force is applied to

the free ends of the polymer, the Gibbs or force ensemble (hereafter referred to as Force Ensemble (ForceEns)) is generated. Extension and force in polymers are the analogous parameters to volume and pressure in liquids or P-V systems. Indeed, the main difference between the ExtEns and the ForceEns is the nature of the control parameter: while in the ExtEns the control parameter is extensive (it scales with the length of the polymer), in the ForceEns the control parameter is an intensive quantity.

According to Flory [45], in the thermodynamic limit, where the number of monomers N constituting the polymer tends to infinity keeping the ratio N/L constant (L is the contour length of the polymer), the fixed extension scheme is also indistinguishable from the situation in which the force is fixed. Nevertheless, single polymers, as an example of small systems, are far away from the thermodynamic limit. Then, their thermodynamic behavior depends on the boundary conditions that are set on the polymer (i.e. the extensive/intensive nature of the control parameter).

In what follows, we perform a brief description of the physical principles that underlie the ExtEns and the ForceEns and, as a proof of ensemble equivalence, we will demonstrate the equivalence of both thermodynamic potentials (i.e. $F(N,V,T,)$ and $G(N,P,T)$) in the thermodynamic limit.

3.2.1 *Fixed-extension ensemble*

Let us consider a polymer, such as a protein or a nucleic acid, with a fixed extension $\mathbf{X}_{\text{tot.}} = (X_{\text{tot}},0,0)$ in contact with a heat bath at a temperature T. This situation corresponds to the ExtEns. In Fig.3.1 it is shown an schematic comparison between the classical canonical NVT ensemble and the ExtEns ensemble for polymers.

The polymer is assumed to be composed of N monomers and the dynamics of the system can be described by a Hamiltonian or energy function $\mathcal{H}(\mathbf{x},\mathbf{p})$, where $\mathbf{x} = \{\mathbf{x}_1,\cdots,\mathbf{x}_N\}$ are the positions of the N monomers and $\mathbf{p} = \{\mathbf{p}_1,\cdots,\mathbf{p}_N\}$ the linear momenta of the monomers. Hence,

$$\mathcal{H}(\mathbf{x},\mathbf{p}) = \sum_{k=1}^{N} \frac{\mathbf{p}_k\cdot\mathbf{p}_k}{2m} + \mathcal{U}(\mathbf{x}_1,\cdots,\mathbf{x}_N)\,, \tag{3.4}$$

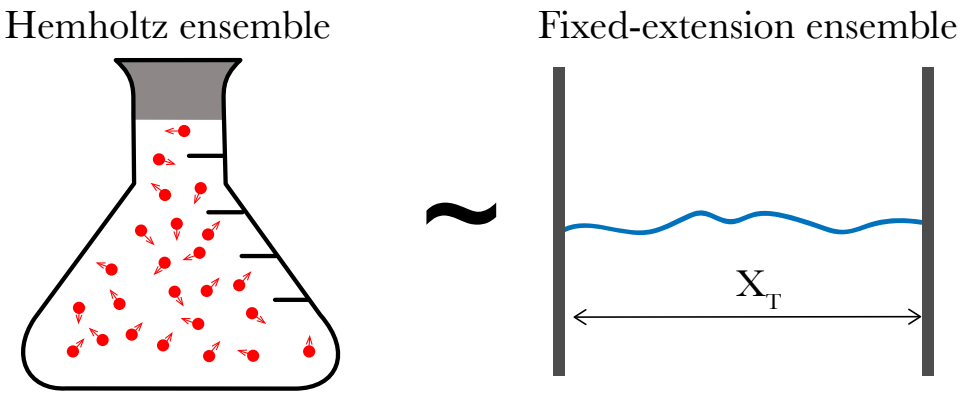

Figure 3.1: **Fixed-volume and fixed-extension analogy.** Analogy between a gas at fixed volume V and a polymer with fixed extension X_{tot}.

where m is the mass of the monomers and $\mathcal{U}(\mathbf{x}_1, \cdots, \mathbf{x}_N)$ is the potential that mediates the interaction of the monomers.

In the ExtEns the positions of the first and the last monomers are fixed, so we can set the position of the first and the N-th monomer as: $\mathbf{x}_1 = \mathbf{0}$ and $\mathbf{x}_N = \mathbf{X}_{\text{tot.}}$, respectively (so that $\mathbf{x}_N - \mathbf{x}_0 = \mathbf{X}_{\text{tot.}}$). Additionally, as both ends are fixed, $\mathbf{p}_1 = \mathbf{p}_N = \mathbf{0}$. Hence, the set of position and momenta become: $\mathbf{x} = (\mathbf{x}_1 = \mathbf{0}, \mathbf{x}_2, \cdots, \mathbf{x}_N = \mathbf{X}_{\text{tot.}})$, $\mathbf{p} = (\mathbf{p}_1 = \mathbf{0}, \mathbf{p}_2, \cdots, \mathbf{p}_N = \mathbf{0})$

Within this scheme, the canonical partition Z_X function of the system yields:

$$Z_X = \int \int_{\mathbb{R}^{6(N-2)}} e^{-\mathcal{H}(\mathbf{x},\mathbf{p})/k_B T} d\mathbf{x} d\mathbf{p} \,. \tag{3.5}$$

And the free energy at constant extension (Eq. (3.2)) equals to[2]:

$$F(\mathbf{x}) = -k_B T \log Z_X \,. \tag{3.6}$$

From $F(\mathbf{x})$ we can obtain the ensemble average of the mechanical force (i.e. equilibrium force) acting on the system as:

$$\langle \mathbf{f}(\mathbf{x}) \rangle = \frac{\partial F(\mathbf{x})}{\partial \mathbf{x}} = -k_B T \frac{\partial \log Z_X}{\partial \mathbf{x}} \,. \tag{3.7}$$

2 Since temperature is constant, its explicit dependence on the thermodynamic quantities is omitted.

3.2.2 *Fixed-force ensemble*

Now let us assume that the polymer, rather than being fixed by its both edges, has a dangling end where a controlled force f is applied This situation corresponds to the ForceEns and it is the analogous of a gas in contact with a bariostat (see Fig. 3.2).

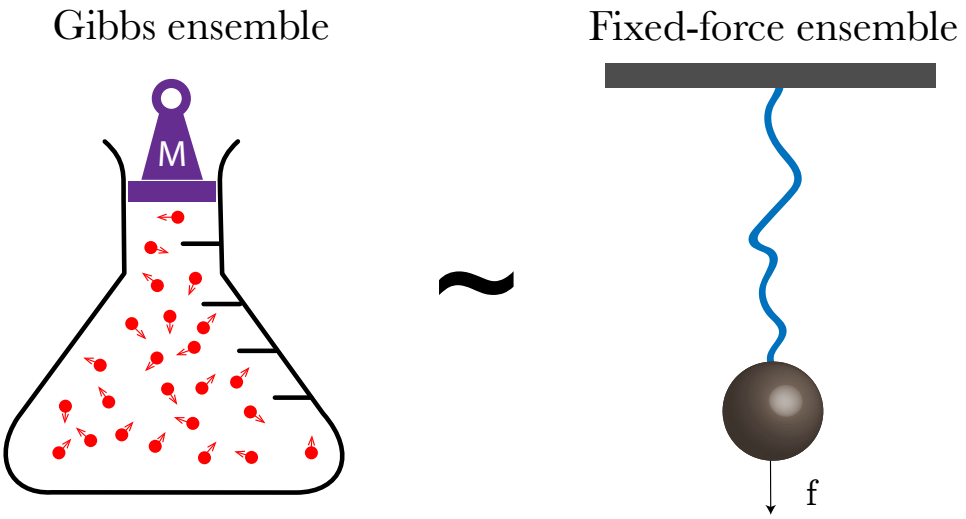

Figure 3.2: **Fixed-pressure and fixed-force analogy.** Analogy between a gas at fixed pressure P and a polymer with a constant force f applied at its free end.

Now, when an external force is applied, the Hamiltonian of the system becomes:

$$\widetilde{\mathcal{H}}(\mathbf{x},\mathbf{p}) = \mathcal{H}(\mathbf{x},\mathbf{p}) - \mathbf{f}\cdot\mathbf{x}\,. \tag{3.8}$$

Therefore, the partition function in the ForceEns can be calculated as:

$$Z_f = \int \int_{\mathbb{R}^{6(N-1)}} e^{-\widetilde{\mathcal{H}}(\mathbf{x},\mathbf{p})/k_B T} d\mathbf{x} d\mathbf{p}\,, \tag{3.9}$$

where we have considered that the set of position and momenta are: $\mathbf{x} = (\mathbf{x}_1 = \mathbf{0}, \mathbf{x}_2, \cdots, \mathbf{x}_N)$, $\mathbf{p} = (\mathbf{p}_1 = \mathbf{0}, \mathbf{p}_2, \cdots, \mathbf{p}_N)$; where now only the first monomer is fixed, resulting in an increase on the number of the degrees of freedom of the whole system, as compared to the ExtEns. In the example shown in Fig. 3.2 the force is applied to the polymer via a bead that is coupled, for instance, to a harmonic

oscillator. Additionally, the kinetic energy of the bead in Fig. 3.2 is neglected because its mass is assumed to be much larger than the whole polymer.

Now, in the ForceEns the Gibbs free energy is directly obtained from the partition function Z_f as:

$$G(\mathbf{f}) = -k_{\mathrm{B}}T \log Z_f \,. \tag{3.10}$$

And the equation of state of the system is obtained by differentiating with respect to the mechanical force f,

$$\langle \mathbf{x}\,(\mathbf{f})\rangle = -\frac{\partial G(\mathbf{f})}{\partial \mathbf{f}} = k_{\mathrm{B}}T \frac{\partial \log Z_f}{\partial \mathbf{f}} \tag{3.11}$$

3.2.3 *Relation between the free energies in the two ensembles*

As in classical thermodynamics, the Helmholtz free energy $F(\mathbf{x})$ obtained in the ExtEns is related to the Gibbs free energy $G(\mathbf{f})$ obtained in the ForceEns. Considering Eqs. (3.6) and (3.10) we can express the corresponding partition functions as follows:

$$Z_X(\mathbf{x}) = e^{-\frac{F(\mathbf{x})}{k_{\mathrm{B}}T}} \,, \tag{3.12}$$

$$Z_f(\mathbf{f}) = e^{-\frac{G(\mathbf{f})}{k_{\mathrm{B}}T}} \,. \tag{3.13}$$

Inserting in Eq. (3.13) the explicit expression for the Hamiltonian of the system (Eq. (3.8)), we may write (omitting prefactor constant terms):

$$e^{-\frac{G(\mathbf{f})}{k_{\mathrm{B}}T}} = Z_f(\mathbf{f}) \sim \int_{\mathbb{R}^3} Z_X(\mathbf{x}) e^{\frac{\mathbf{f}\cdot\mathbf{x}}{k_{\mathrm{B}}T}} d\mathbf{x} \,. \tag{3.14}$$

Then, from Eq. (3.12) we have:

$$\int_{\mathbb{R}^3} Z_X(\mathbf{x}) e^{\frac{\mathbf{f}\cdot\mathbf{x}}{k_{\mathrm{B}}T}} d\mathbf{x} = \int_{\mathbb{R}^3} e^{-\frac{F(\mathbf{x})}{k_{\mathrm{B}}T}} e^{\frac{\mathbf{f}\cdot\mathbf{x}}{k_{\mathrm{B}}T}} d\mathbf{x} \,. \tag{3.15}$$

Now, considering the extensive character of the thermodynamic potentials we write: $F(\mathbf{x}) = Nf_N(\mathbf{l})$, $G(\mathbf{f}) = Ng_N(\mathbf{f})$, $\mathbf{f}\cdot\mathbf{x} = N\mathbf{f}\cdot\mathbf{l}$ and $d\mathbf{x} = Nd\mathbf{l}$. Note that $\mathbf{l} = \mathbf{x}/N$. As a consequence, Eq. (3.14) becomes:

$$\exp\left(-N\frac{g_N(\mathbf{f})}{k_{\mathrm{B}}T}\right) \sim \int_{\mathbb{R}^3} Nd\mathbf{l}\, \exp\left(-N\frac{f_N(\mathbf{l}) - \mathbf{f}\cdot\mathbf{l}}{k_{\mathrm{B}}T}\right) . \quad (3.16)$$

The integral can be evaluated in the large N limit using the saddle-point method, obtaining:

$$\exp\left(-N\frac{g_N(\mathbf{f})}{k_{\mathrm{B}}T}\right) \sim N\exp\left(-N\min_{\mathbf{l}}\left[\frac{f_N(\mathbf{l}) - \mathbf{f}\cdot\mathbf{l}}{k_{\mathrm{B}}T}\right]\right) , \quad (3.17)$$

which, extracting logarithms and dividing by $-N/k_{\mathrm{B}}T$ yields,

$$g_N(\mathbf{f}) = \min_{\mathbf{l}}\left[f_N(\mathbf{l}) - \mathbf{f}\cdot\mathbf{l}\right] + \mathcal{O}\left(\frac{\log N}{N}\right) . \quad (3.18)$$

Since the $\log(N)/N$ corrections to Eq. (3.18) vanish in the large N limit we can confirm that, $g_N(\mathbf{f})$ is the Legendre transformation of $f_N(\mathbf{l})$ using $\mathbf{f}$ and $\mathbf{l}$ as conjugate pairs. This transformation preserves the convexity of thermodynamic potentials with respect to its own variables (i.e. Helmholtz free energy is a convex function of the extensive variable and so it is the Gibbs free energy with respect to the intensive conjugate variable).

Equation (3.18) is exact in the large N limit. Therefor, the main message of Eq. (3.18) is that Gibbs and Helmholtz ensembles, for polymer systems, are equivalent when $N \to \infty$. This fact has been recently demonstrated for the case of real polymers, showing that the mechanical response (i.e. force-extension curves) are ensemble-independent for single infinite polymers [46].

3.3 STABILITY CRITERIA IN POLYMER ENSEMBLES

As it is well known in classical thermodynamics, the equilibrium conditions of a physical system subjected to one or some external constraints (e.g. by fixing its temperature, volume, etc.) are given by the extrema

(maximum or minimum) of the thermodynamic potentials [44]. In particular, in equilibrium, a system at constant volume (pressure), will be found in its Helmholtz (Gibbs) free energy minimum. Moreover, minima conditions also impose constraints in the curvature of the thermodynamic potentials with respect to their control parameters. As a matter of fact, Helmholtz (Gibbs) free energy is a convex function[3] with respect to the volume (pressure).

Convexity conditions are denoted by the negative behavior of second derivatives of thermodynamic potentials at their extrema. Also, classical thermodynamics relate these second derivatives with thermodynamic measurable quantity. For instance, in the Helmholtz ensemble, where the equation of the systems is given by:

$$P = -\left(\frac{\partial F}{\partial V}\right)_T , \tag{3.19}$$

taking the derivative with respect to the pressure yields:

$$\frac{\partial P}{\partial V} = -\left(\frac{\partial^2 F}{\partial V^2}\right)_T . \tag{3.20}$$

And, according to the minimum criterion, the quantity $\left(\frac{\partial^2 F}{\partial V^2}\right)_T$ must be positive. The isothermal compressibility is defined as $\kappa_T = -(1/V)(\partial V/\partial P)_T$, so that we find:

$$\left(\frac{\partial^2 F}{\partial V^2}\right)_T = \frac{1}{V\kappa_T} , \tag{3.21}$$

resulting in the well-known condition which states that isothermal compressibility must be positive to guarantee thermodynamic stability. Similarly, the isothermal compressibility is related to the Gibbs free energy in fixed-pressure ensemble through:

$$\left(\frac{\partial^2 G}{\partial P^2}\right)_T = V\kappa_T . \tag{3.22}$$

3 Also called concave upward, meaning that the line segment between any two points of the function lies above or on the function.

Therefore, Gibbs ensemble also preserves the positive behavior of the isothermal compressibility. Interestingly, as in standard P-V systems in statistical mechanics, κ_T is also related to fluctuations of the free (non-controlled) parameters in the polymer ensembles. In the following sections we will develop the relations between the isothermal compressibility in the usual polymer ensembles (ExtEns and ForceEns) and the fluctuations of the free parameters.

3.3.1 *Force fluctuations in the extensional ensemble*

In the ExtEns, as described in Sec. 3.2.1, the force acting on the polymer ends is a fluctuating quantity while the total extension X_{tot} is a fixed parameter. For the sake of simplicity, throughout this section we will treat the polymers as one-dimensional systems.

Without loss of generality, let us consider a polymer that can be found into two conformations[4]: folded (F) or unfolded (U). Furthermore, regard that the polymer is subjected to isometric conditions (i.e. ExtEns), fixing its total extension. The equilibrium probabilities of states F and U are given by the Boltzmann-Gibbs factor:

$$p_{\text{F(U)}} = \frac{\exp\left(-\frac{F_{\text{F(U)}}}{k_{\text{B}}T}\right)}{Z_X}\,, \tag{3.23}$$

where $F_{\text{F(U)}}$ is the partial free energy of F (U) state and $Z_X = e^{-F_{\text{F}}/k_{\text{B}}T} + e^{-F_{\text{U}}/k_{\text{B}}T}$ is the partition function of the system calculated in the ExtEns scheme.

The ensemble (or equilibrium) force of the system at a given extension X_{tot} is given by:

$$\langle f \rangle = \frac{\langle f_{\text{F}} \rangle e^{-F_{\text{F}}/k_{\text{B}}T} + \langle f_{\text{U}} \rangle e^{-F_{\text{U}}/k_{\text{B}}T}}{Z_X}\,, \tag{3.24}$$

and the second moment of the force $\langle f^2 \rangle$ equals to:

$$\langle f^2 \rangle = \frac{\langle f_{\text{F}}^2 \rangle e^{-F_{\text{F}}/k_{\text{B}}T} + \langle f_{\text{U}}^2 \rangle e^{-F_{\text{U}}/k_{\text{B}}T}}{Z_X}\,. \tag{3.25}$$

4 This is the actual situation for small size polymers and it is analogous of a a two-state system in statistical mechanics.

We note that $\langle f_{\mathrm{F}} \rangle$ and $\langle f_{\mathrm{U}} \rangle$ are the average over the partially equilibrated F and U states and are defined as: $\langle f_{\mathrm{F(U)}} \rangle = \partial_{X_{\mathrm{tot}}} F_{\mathrm{F(U)}}$.

Multiplying at both sides of Eq. (3.24) by Z_X and differentiating at both sides with respect to X_{tot} yields:

$$\begin{aligned} \frac{\partial \langle f \rangle}{\partial X_{\mathrm{tot}}} Z_X + \langle f \rangle \frac{\partial Z_X}{\partial X_{\mathrm{tot}}} &= \frac{\partial \langle f_{\mathrm{F}} \rangle}{\partial X_{\mathrm{tot}}} e^{-F_{\mathrm{F}}/k_{\mathrm{B}}T} + \frac{\partial \langle f_{\mathrm{U}} \rangle}{\partial X_{\mathrm{tot}}} e^{-F_{\mathrm{U}}/k_{\mathrm{B}}T} \\ -\frac{1}{k_{\mathrm{B}}T} \left(\langle f_{\mathrm{F}} \rangle \frac{\partial f_{\mathrm{F}}}{\partial X_{\mathrm{tot}}} e^{-F_{\mathrm{F}}/k_{\mathrm{B}}T} + \langle f_{\mathrm{U}} \rangle \frac{\partial f_{\mathrm{U}}}{\partial X_{\mathrm{tot}}} e^{-F_{\mathrm{U}}/k_{\mathrm{B}}T} \right) &= \\ = \langle k_{\mathrm{F}} \rangle e^{-F_{\mathrm{F}}/k_{\mathrm{B}}T} + \langle k_{\mathrm{U}} \rangle e^{-F_{\mathrm{U}}/k_{\mathrm{B}}T} - \frac{\langle f \rangle^2}{k_{\mathrm{B}}T} Z_X \,. & \end{aligned} \tag{3.26}$$

In previous equation (Eq. (3.26)) we have identified the terms $\frac{\partial \langle f_{\mathrm{F}} \rangle}{\partial X_{\mathrm{tot}}}$ and $\frac{\partial \langle f_{\mathrm{U}} \rangle}{\partial X_{\mathrm{tot}}}$ as the stiffness of the F and U states (i.e. the slope of the $f_{\mathrm{F(U)}}(X_{\mathrm{tot}})$ curve). The equilibrium stiffness $\langle k \rangle$ is defined as:

$$\langle k \rangle = \frac{1}{Z_X} \left(\langle k_{\mathrm{F}} \rangle e^{-F_{\mathrm{F}}/k_{\mathrm{B}}T} + \langle k_{\mathrm{U}} \rangle e^{-F_{\mathrm{U}}/k_{\mathrm{B}}T} \right) . \tag{3.27}$$

Dividing both sides of Eq. (3.26) by Z_X and inserting the definition of $\langle k \rangle$ (Eq. (3.27)) we obtain:

$$\frac{\partial \langle f \rangle}{\partial X_{\mathrm{tot}}} - \frac{\langle f \rangle^2}{k_{\mathrm{B}}T} = \langle k \rangle - \frac{\langle f^2 \rangle}{k_{\mathrm{B}}T} \,, \tag{3.28}$$

where we have also used the following equality:

$$\frac{1}{Z_X} \frac{\partial Z_X}{\partial X_{\mathrm{tot}}} = \frac{\partial \log Z_X}{\partial X_{\mathrm{tot}}} = \frac{-1}{k_{\mathrm{B}}T} \frac{\partial F}{\partial X_{\mathrm{tot}}} = -\frac{\langle f \rangle}{k_{\mathrm{B}}T} \,, \tag{3.29}$$

being F the total free energy of the system. The average force fluctuations $(\Delta f)^2$ are equal to: $(\Delta f)^2 = \langle f^2 \rangle - \langle f \rangle^2$. Hence, Eq. (3.28) can be rewritten as:

$$\frac{(\Delta f)^2}{k_{\mathrm{B}}T} = \langle k \rangle - \frac{\partial \langle f \rangle}{\partial X_{\mathrm{tot}}} := \langle k \rangle - k^x_{\mathrm{eff}} \,, \tag{3.30}$$

where the term k^x_{eff} is the effective stiffness[5] of the system (i.e. the slope of the force-extension $\langle f(X_{\mathrm{tot}}) \rangle$ curve). A beautiful consequence

5 It is called *effective* due to the fact that, in typical experimental setups, the polymer is attached in series with more molecules, resembling as a collection of springs in series.

of Eq. (3.30) is that negative compressibilities (i.e. negative k^x_{eff}) are allowed in the ExtEns. Unlike typical P-V systems, where negative compressibilities are forbidden by stability conditions of thermodynamic potentials, in polymer systems we can actually measure negative response parameters yet fulfilling the constraint that fluctuations of the uncontrolled parameters are positive.

We point that last equation (Eq. (3.30)) can be generalized for more than two states. For an N-state system, the average stiffness would read as: $\langle k \rangle = (1/Z_X) \sum_{i=1}^{N} \frac{\partial f_i}{\partial X_{\text{tot}}} e^{-F_i/k_{\text{B}}T}$, where now the limit of summation is the number of states N that the polymer may explore and Z_X is the partition function of the system .

3.3.2 *Extension fluctuations in the force ensemble*

Consider now that the mechanical force at the ends of the polymer is controlled, as discussed in Sec. 3.2.2. In the ForceEns the extension of the polymer, rather than the force, fluctuates. As we have previously shown, the partition function in this configuration is given by Eq. (3.14).

$$Z_f = \int dx\, e^{-\frac{F(x)}{k_{\text{B}}T}} e^{\frac{fx}{k_{\text{B}}T}} = e^{-\frac{G(f)}{k_{\text{B}}T}}\,, \tag{3.31}$$

where $F(x)$ is the Helmholtz free energy (i.e. in the ExtEns) and $G(f)$ is the Gibbs free energy (i.e. in the ForceEns). Now, it is straightforward to show that the n-th derivative of Z_{F} with respect to the control parameter f is related to the n-th moment of the extension as:

$$\langle x^n \rangle = \frac{(k_{\text{B}}T)^n}{Z_f} \frac{\partial^n Z_f}{\partial f^n} = (k_{\text{B}}T)^n \frac{\partial^n \log Z_f}{\partial f^n} \;\; \forall n \geq 1\,. \tag{3.32}$$

Hence, the average extension fluctuations, $(\Delta x)^2 = \langle x^2 \rangle - \langle x \rangle^2$, can be related to the slope of the equilibrium force-extension curve as follows:

$$\frac{(\Delta x)^2}{k_{\text{B}}T} \frac{1}{\langle x \rangle} = \frac{-1}{\langle x \rangle} \frac{\partial^2 G(f)}{\partial f^2} = \frac{1}{\langle x \rangle} \frac{\partial \langle x \rangle}{\partial f} := \frac{1}{\langle x \rangle k^f_{\text{eff}}}\,, \tag{3.33}$$

where we have identified the effective stiffness in the ForceEns, k^f_{eff}, with:

$$\frac{\partial\langle x\rangle}{\partial f} = \frac{1}{\partial f/\partial\langle x\rangle} = \frac{1}{k^f_{\text{eff}}}\,. \tag{3.34}$$

Again, the effective stiffness is the slope of the extension-force curve ($\langle x(f)\rangle$ curve).

We note that Eq. (3.33) is a fluctuation-dissipation-like relation. Moreover, Eq. (3.33) indicates us that in the ForceEns, the effective stiffness is always positive. This effect opposites with the fact than in the ExtEns, isothermal compressibility can be negative (see Eq. (3.30)), whereas in the ForceEns is always positive as required for stability.

Part II

ENSEMBLE INEQUIVALENCE

4

FLUCTUATION THEOREM IN THE FORCE ENSEMBLE

4.1 FLUCTUATION THEOREMS

Until the dawn of stochastic thermodynamics at the end of the 20th Century, the measurement of equilibrium thermodynamic quantities in physical systems was only restricted to experimental assays performed in static equilibrium conditions (quasistatic conditions). Typically, experiments are done by modifying an externally controlled physical quantity λ, the so-called *control parameter*. The assumption of quasistaticity implies that the thermodynamic system must relax much faster than the rate at which the external parameter is varied. Therefore, a quasistatic process must satisfy the following relation:

$$\left(\frac{d\lambda}{dt}\right) \ll \left(\frac{\Delta\lambda}{\tau}\right), \tag{4.1}$$

where τ is the relaxation time of the system, t is time and $\Delta\lambda$ is the variation step of the control parameter. This condition guarantees that the system passes through an infinite set of equilibrium states, ensuring that the measured quantities are purely the equilibrium ones.

In general, the required energy to drive the system from an arbitrary state labelled with a value of the control parameter λ_0 up to a λ_1 state is given by:

$$W_\lambda = \int_{\lambda_0}^{\lambda_1} d\lambda \left(\frac{\partial \mathcal{H}(\lambda, t)}{\partial \lambda}\right), \tag{4.2}$$

where W_λ is the mechanical work and $\mathcal{H}$ is the Hamiltonian or energy function. The previous equation holds for arbitrary condictions. Nevertheless, if the process is carried out quasistatically and, additionally, if it is possible to perform the time-inversion process at each infinitesimal variation of the control parameter (i.e. there is no time arrow) the process is said to be *reversible*. In such conditions, the energy required to modify the state of the system from λ_0 to λ_1 is equal to the

free energy ΔG difference between the states labeled with λ_0 and λ_1 [47]:

$$W_{\text{rev.}} = \Delta G = G(\lambda_1) - G(\lambda_0)\,. \tag{4.3}$$

The free energy change ΔG measures the net amount of energy exchanged between the system and its surroundings along an arbitrary equilibrium pathway and, as we mentioned, its measurement is constrained by equilibrium conditions. However, the development of fluctuation theorems in the 90s changed the situation: equilibrium free energy differences can be measured from irreversible processes.

Fluctuation theorems are mathematical identities that allow the recovery of equilibrium thermodynamic quantities in nonequilibrium experiments in driven microsystems. Although their experimental applicability range is very wide (ranging from mechanical oscillators, colloids, biological systems up to electric circuits) [48] and their validity is general in stochastic thermodynamics [49], it is in biomolecular systems where their potential is more exploited, highlighting the free energy recovery of molecular structures in nonequilibrium experiments as their major feature. For instance, they have been successfuly applied to determine the free energies of formation of DNA and RNA hairpins [50] or proteins [51]. Moreover, their applicability have been extended to non full equilibrium states and non native states, such as kinetic intermediate states [52], plus the measurement of binding energy of small ligands that bind to nucleic acids [53] or even the measurement of mechanical torque in molecular rotatory motors [54]. Also, due to the connection of stochastic thermodynamics with information theory [55], fluctuation theorems are currently extended to explore information-to-energy conversion in systems with feedback control [56] and the measurement of information content in molecular systems (see chapter 7).

4.1.1 *Nonequilibrium work relations. The Crooks fluctuation relation and the Jarzynski equality*

Among all the existing fluctuation theorems, the Crooks Fluctuation Theorem (CFT) [57] and its corollary, the Jarzynski Equality (JE) [58], are

the most used to relate *irreversible* work measurements with equilibrium free energy differences.

Consider a system initially in thermal equilibrium at a given state at time t_0 where the control parameter fixed at a value $\lambda(t_0) = \lambda_0$. Then, an arbitrary experimental protocol $\lambda(t)$ is applied on the system during a time interval Δt. Afterwards, at the end of the protocol, at time $t_1 = t_0 + \Delta t$, the system will be found in the state corresponding to $\lambda(t_1) = \lambda_1$. This process corresponds to the so-called *forward* (F) process and the exerted mechanical work is given by Eq. (4.2).

Now let us suppose that the *reversed* (R) protocol is subsequently implemented. The system is equilibrated now at λ_1 and, following the time-mirrored image of the experimental path ($\tilde{\lambda}(t) = \lambda(\Delta t - t)$), the state λ_0 is recovered after a time interval $\Delta t = t_1 - t_0$. In Fig. 4.1(a) we show a schematic depiction of an arbitrary nonequilibrium experiment where a control parameter is changed from λ_0 to λ_1 in a time interval Δt, while in Fig. 4.1(b) it is shown its time-reversed process.

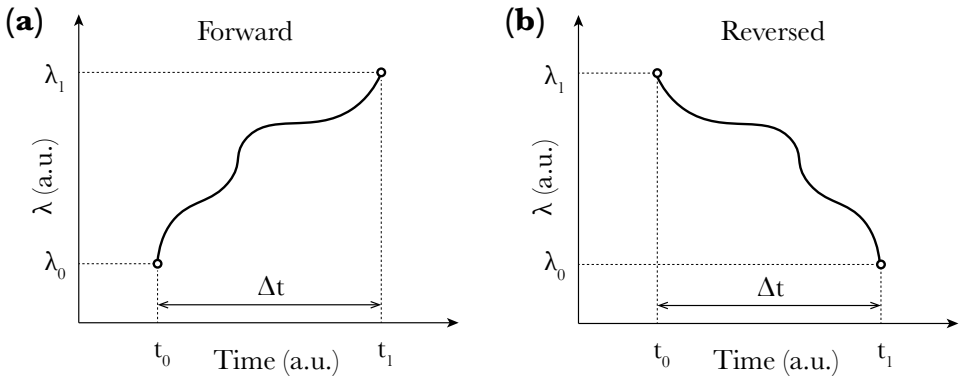

Figure 4.1: **Experimental protocol.** Schematic depiction of an arbitrary forward protocol **(a)** and its time-reversed path **(b)**.

Since the energies involved in the typical processes of small systems are on the same scale than Brownian fluctuations (i.e. $\sim k_B T$), the configurations that a small system may explore in the F and R processes can be significantly different (and in subsequent experimental realizations). As a matter of fact, the term $\partial_\lambda \mathcal{H}$ of Eq. (4.2) is a fluctuating quantity, so the mechanical work is path-dependent in small systems.

The relation between the mechanical work in the F and R processes is done via the CFT, which reads as:

$$\frac{P_F(W)}{P_R(-W)} = \exp\left(\frac{W - \Delta G}{k_B T}\right), \tag{4.4}$$

where $P_F(W)$ is the probability density function of the work done in the F process (forward protocol), $P_R(-W)$ is the probability density function of the mechanical work done along the R process (reversed protocol) and ΔG is the free energy difference given by Eq. (4.3).

Additionally, the JE can be easily obtained from the CFT by multiplying both sides of Eq. (4.4) by $P_R(-W)$ and integrating with respect to W. Therefore:

$$\left\langle \exp\left(-\frac{W}{k_B T}\right)\right\rangle_F = \exp\left(-\frac{\Delta G}{k_B T}\right), \tag{4.5}$$

where $\langle \cdots \rangle_F$ denote the average over the forward process: $\langle(\cdots)\rangle_F = \int dW\,(\cdots)P_F(W)$. The JE is a corollary of the CFT and, since it does not carry any information about the reverse process, it is called a *unidirectional free energy estimator*.

It has been already shown that the CFT holds for systems initially in equilibrium and independently of how far from equilibrium the system is driven [59]. In general, Eq. (4.4) holds when the full work is measured, while it does not hold when partial work measurements are done [60] or when the transferred, rather than the accumulated, work is measured in controlled extension protocols using LOT experiments [61, 62]. Moreover, the CFT holds under general assumptions of microscopic reversibility and detailed balance.

4.2 WORK DEFINITION IN THE FORCE ENSEMBLE

Despite the general validity of the CFT and some experiments using driven oscillators [63], it has not been previously tested in the case of a force-controlled SME. As we commented in the previous section, the CFT has been widely tested and used for free energy recovery in several experimental scenarios. However, all these studies possess a common factor: the use of extensive control parameters. For instance, in single

molecules pulled by LOT and AFM the optical trap-bead distance and the cantilever-surface distance scale proportionally to the polymer length (see Sec. 1.3 for a brief introduction on these techniques). On the other hand, MT and AFS are high-throughput techniques that manipulate multiple molecules in parallel where force is the natural control parameter (magnetic field in MT and acoustic pressure in AFS). Note that in the latter case, the control parameter is an intensive variable, and so it does not scale with the system size.

As we already discussed in Chapter 3, the selection of a certain control parameter defines the proper statistical ensemble [64]. For the case of macroscopic systems, the thermodynamic description is independent of the ensemble. However, in small systems, where fluctuations dominate the microscopic behaviour, this fact is no longer true. Two conjugate statistical ensembles are not equivalent, in general. We will consider the case of the mechanical work as the paradigm of ensemble inequivalence.

To illustrate how the control parameter choice constraints the physical description of the system, let us consider a single polymer with controlled extension. Hence, $\lambda = x$. If the polymer is stretched by increasing the extension from x_0 to x_1, the mechanical work given by Eq. (4.2) reduces to the well-known classical work expression [65]:

$$W_x = \int_{x_0}^{x_1} f(x')\, dx' \,, \tag{4.6}$$

where $f = \partial_x \mathcal{H}$ is the mechanical force acting on the ends of the polymer. If the mechanical force is controlled ($\lambda = f$), the performed mechanical work in a protocol where the force is changed from f_0 to f_1 is given by:

$$W_f = -\int_{f_0}^{f_1} x(f')\, df' \,, \tag{4.7}$$

with $x = -\partial_f \mathcal{H}$. The first situation corresponds to the ExtEns (Eq. (4.6)), while the latter (Eq. (4.7)) corresponds to the ForceEns (see chapter 3 for a discussion on both ensembles). Both work definitions are related by boundary terms of a Legendre transformation using extension and force as conjugate pairs:

$$W_x = W_f + \Delta\,(xf) = W_f + (x_1 f_1 - x_0 f_0)\,. \tag{4.8}$$

Although the correctness of the theoretical work definition in the ForceEns, Eq. (4.7), is widely accepted by the scientific community by now, there has been controversy in this regard [66–69]. For this reason, we decided to test the validity of the work definition in two conjugated ensembles using the CFT. Moreover, testing the validity of the CFT in the ForceEns is crucial to extend the applicability of free-energy recovery methods to high-throughput single-molecule techniques.

4.3 EXPERIMENTAL TEST OF THE CROOKS FLUCTUATION THEOREM IN THE FORCE ENSEMBLE

In this section we show the experimental validation of the CFT in the ForceEns by computing the mechanical work according to Eq. (4.7). This test has been carried out using MT and LOT with force-feedback control [70] .

4.3.1 *Results with Magnetic Tweezers*

We first tested the validity of the CFT in MT experiments. To do so, we performed bidirectional pulling experiments on the CD_4 20-bp DNA hairpin (whose sequence is shown in Fig. 4.2(a)). Also, the DNA hairpin was flanked by two 29-bp dsDNA handles [71]. The whole molecular construct (handles + DNA) was tethered between a glass surface and a superparamagnetic 1-μm bead that is captured in a magnetic trap generated by a pair of permanent magnets. The mechanical force is directly controlled by modulating the magnetic field gradient that increases as the magnets approach the glass surface.

Bidirectional pulling experiments consist of consecutive cycles of stretching/releasing cycles of the molecule. The stretching (or unfolding) protocol is identified with the F process (see Sec. 4.1.1), whereas the releasing (or folding) protocol is identified with the R process.

In the unfolding process, the system is equilibrated at a given value of the force f_0. Thus, by approaching the magnets to the glass surface at constant velocity the DNA hairpin is stretched until it unfolds and a final force f_1 is reached. Due to the Brownian nature of the system, the unfolding force (i.e. the force at which the DNA hairpin can not withstand the tension in the folded conformation) is stochastic. The unfolding results in a sudden increase of the molecular extension cor-

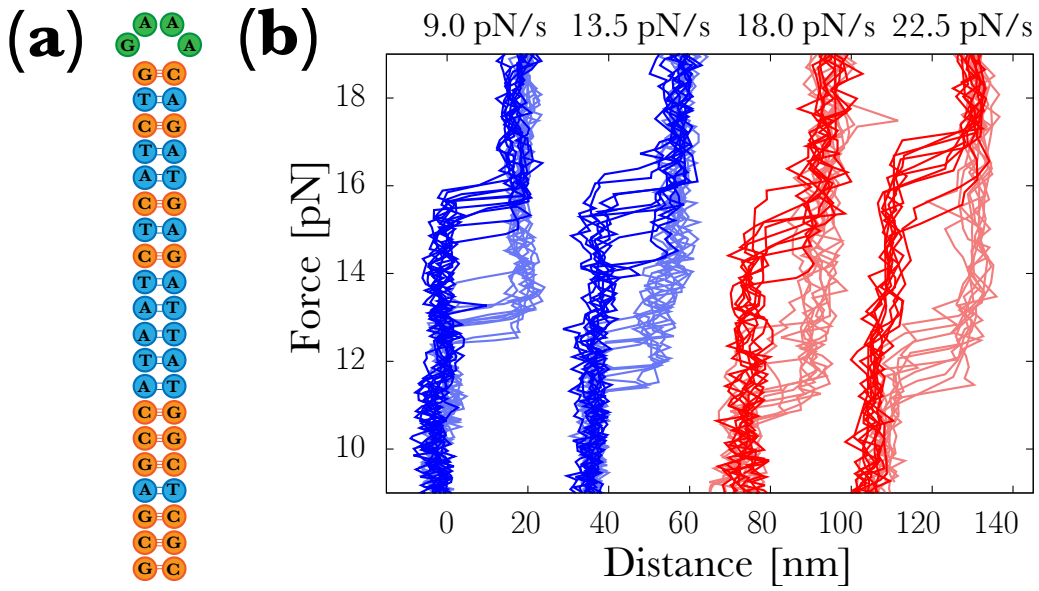

Figure 4.2: **CD_4 hairpin and bidirectional pulling experiments in Magnetic Tweezers. (a) -** Sequence and structure of CD_4 DNA hairpin. **(b) -** Unfolding (folding) trajectories are plot as dark (light) curves. Examples of force-distance cycles at different pulling rates. Curves were shifted for the sake of clarity.

responding to the release of the single-stranded DNA (ssDNA) that is associated with the unfolding process.

On the other hand, in the folding process, the unfolded ssDNA is in equilibrium at f_1 and then, moving away the magnets from the glass surface following the time-reversed protocol the force f_0 is finally reached. The refolding of the molecule is observed as a sudden absorption of a certain molecular extension corresponding to the generation of the original helical structure.

In Fig. 4.2 we show typical Force-Distance Curves (FDCs) of bidirectional pulling experiments performed in MT at different pulling rates [1]. Hysteresis effects increase as the pulling rate r increases: the dispersion in unfolding/folding forces grows as r increases.

According to Eqs. (4.6) and (4.7), W_x and W_f are given by the shaded areas in Fig. 4.3.

$P_F(W)$ and $P_R(-W)$ are shown in Fig. 4.4(a) when the mechanical work is obtained using Eq. (4.7). Interestingly, although the presence of high dissipation effects, the work value at the crossing point between both distributions does not change with the pulling rate, as expected

1 Pulling rate r is defined as the temporal derivative of the force, $r = \dot{f}$

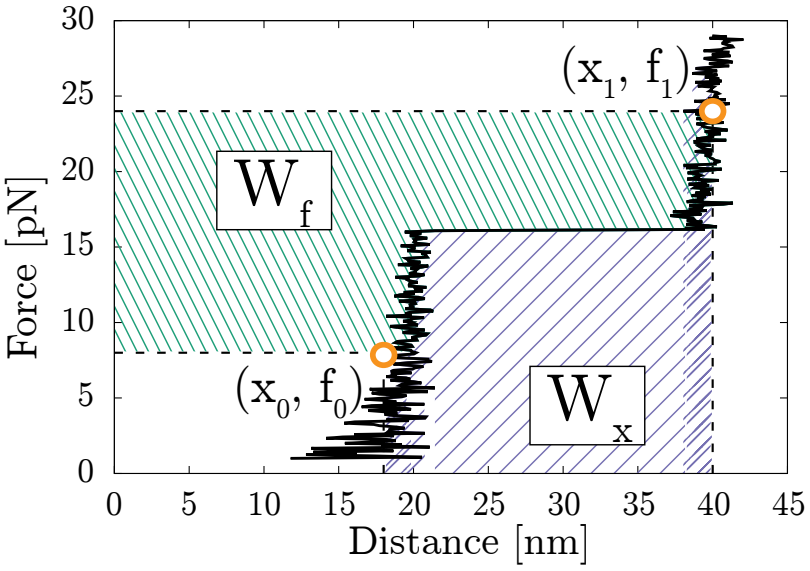

Figure 4.3: **Measurement of the mechanical work in the ForceEns and in the ExtEns.** Work value corresponding to the ForceEns and ExtEns correspond to the shaded area at the left and below the curve, respectively.

[57]. According to the CFT, Eq. (4.4), such work value is equal to ΔG_f (Eq. (4.3))[2]. The measurement of the crossing point of work distributions obtained at 9.0 pN/s and 22.5 pN/s gives $\Delta G_f = -36 \pm 6$ and $-35 \pm 6\ k_B T$, respectively.

When the work is computed according to Eq. (4.6) notwithstanding the fact that force, rather than molecular extension, is the control parameter, distributions of W_x also present intersecting points that are independent of the pulling rate (Fig. 4.4(b)). In this case, however, the CFT is not fulfilled. The CFT can be validated by extracting logarithms in both sides of Eq. (4.4), yielding:

$$\log\left(\frac{P_F(W)}{P_R(-W)}\right) = \frac{W}{k_B T} - \frac{\Delta G}{k_B T}, \tag{4.9}$$

and performing a linear fit to the left-hand side of Eq. (4.9) as a function of $W/\,k_B T$. When the CFT holds, data falls in a straight line of slope 1 and y-intercept equal to $-\Delta G$, both in $k_B T$ units.

In Fig. 4.5 it is shown how the CFT is fulfilled for W_f (i.e. Eq. (4.7)). However, the CFT is not satisfied for W_x (i.e. Eq. (4.6)): the slopes of

2 The subscript f in the Gibbs free energy will be used to distinguish between the ForceEns and ExtEns values.

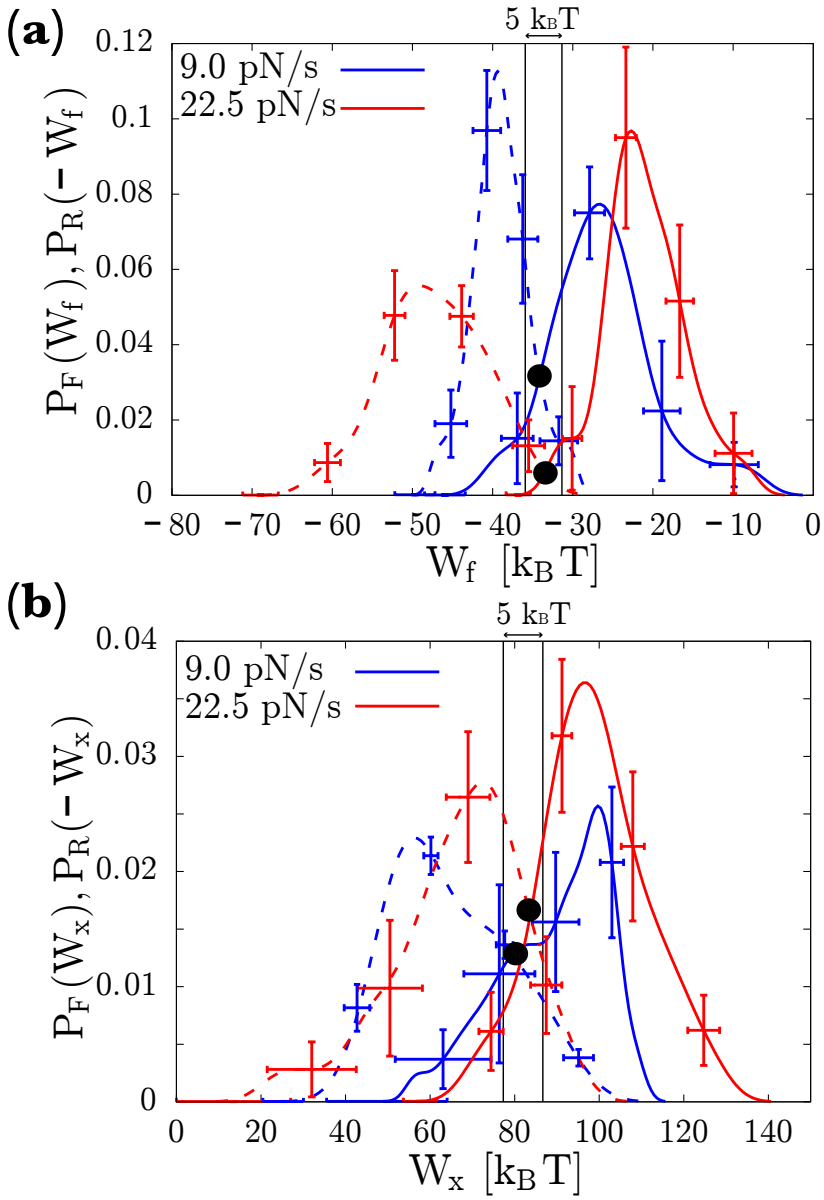

Figure 4.4: **Work distributions in MT experiments. a -** Distributions obtained computing W in the ForceEns scheme (i.e. Eq. (4.7). **b -** Distributions obtained computing W in the ExtEns scheme (i.e. Eq. (4.6). Error bars have been obtained using the Bootstrap method.

the linear fits of Eq. (4.9) are 0.075 ± 0.010 and 0.33 ± 0.04 for the 9.0 pN/s and 22.5 pN/s pulling rates, respectively.

The breakdown of the CFT symmetry indicates that W_x does not measure the correct thermodynamic work in the ForceEns. In fact, the missing contribution in W_x is the boundary term: $\Delta(xf) = x_1 f_1 - x_0 f_0 = W_x - W_f$. This term is not constant but fluctuates over different

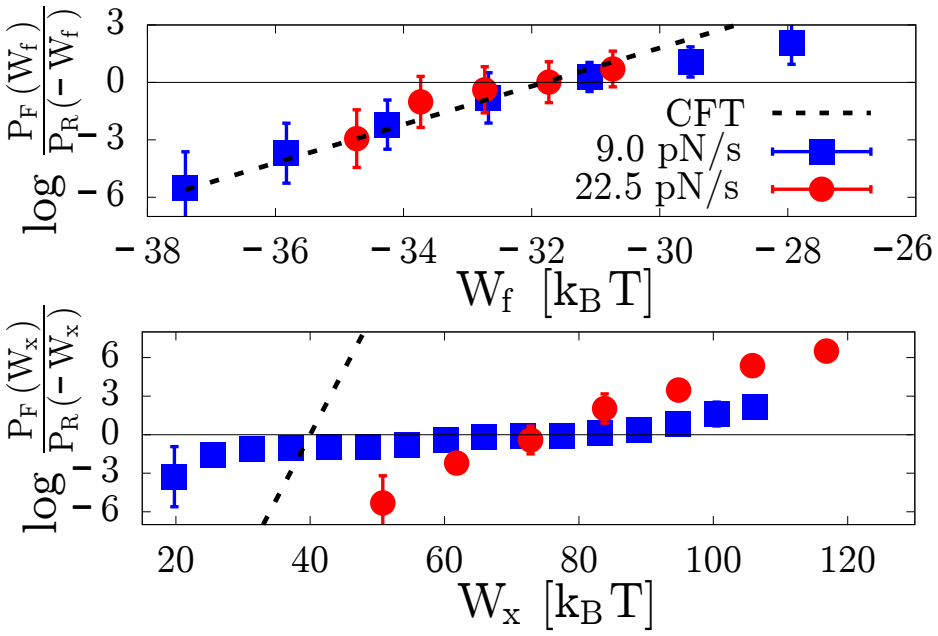

Figure 4.5: **CFT test plot in MT experiments.** Results corresponding to the ForceEns (top panel) and the ExtEns (bottom panel). Dashed straight black lines have slopes equal to 1 in $k_B T$ units.

pulling cycles as the initial and final extensions x_0, x_1 are fluctuating variables (whereas f_0, f_1 are fixed). In other words, the boundary term $\Delta(xf)$ is a stochastic variable that contributes to the tails of the work distributions that are crucial for testing the validity of the CFT in the work crossing region. In Sec. 4.4 we discuss the contribution of the work boundary terms deeper.

$\Delta G_f\ [k_B T]$	$\Delta G_x\ [k_B T]$	$\Delta G_0\ [k_B T]$
-32 ± 5	80 ± 5	49 ± 5

Table 4.1: **Fluctuation theorem and free energy recovery.** The free energy in the ForceEns is obtained by terms of the CFT (Eq. (4.4) with W_f), whereas ΔG_x is obtained by subtracting the boundary term $\langle \Delta(xf) \rangle$. Folding free energy ΔG_0 is finally obtained when all the elastic contributions are subtracted (see Appendix C for details). Previous results are the average over 7 molecules and the error bar corresponds to the propagation of the standard error of the mean and the error of the free energy estimator.

Moreover, as the mechanical work, the free energy difference in the ForceEns, ΔG_f, is also related to the free energy difference in the ExtEns, ΔG_x, via: $\Delta G_x = \Delta G_f + \langle \Delta (xf) \rangle$, where angular brackets denote the average over all experimental realizations. In Table 4.1 we report the results for the free energies in MT experiments. The obtained value for the folding free energy at zero force, ΔG_0, is in very good agreement with the predicted value using the Nearest-Neighbour model for DNA [72, 73], giving: $\Delta G_0 = 51\, k_B T$.

4.3.2 *Results with Laser Optical Tweezers*

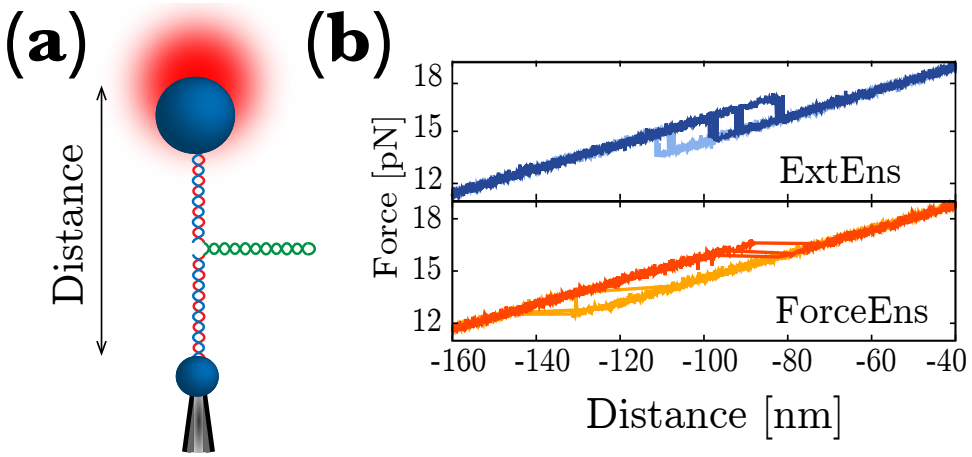

Figure 4.6: **Schematics of LOT experiments and FDCs. (a)** - Schematic depiction of the experimental setup in LOT experiments. Distance corresponds to the relative distance between the center of the optical trap and the tip of the micropipette. **(b)** - Data corresponding to the ExtEns (top graph) and the ForceEns (bottom graph).

In LOT the position of the optical trap is the natural control parameter (see Fig. 4.6(a)), whereas the molecular extension and the force are fluctuating quantities[3]. However, using force feedback control the position of the trap is actively rectified while the force is kept constant. This process is done by implementing the following feedback loop: first, at a sampling rate of 1 kHz, the force acting on the bead captured in the optical trap is measured as a time-average. Then, depending whether the measured value is higher (lower) than the desired value, the position

3 Although the molecular extension is not directly controlled, the position of the optical trap is an extensive quantity (it is directly related to the molecular extension).

of the optical trap is decreased (increased), so that an approximately constant force is maintained[70, 74].

We performed experiments in LOT in the standard passive mode (ExtEns) and in the active feedback mode (ForceEns). The implemented experimental protocol is equivalent to the one we used in MT experiments (see section 4.3.1).

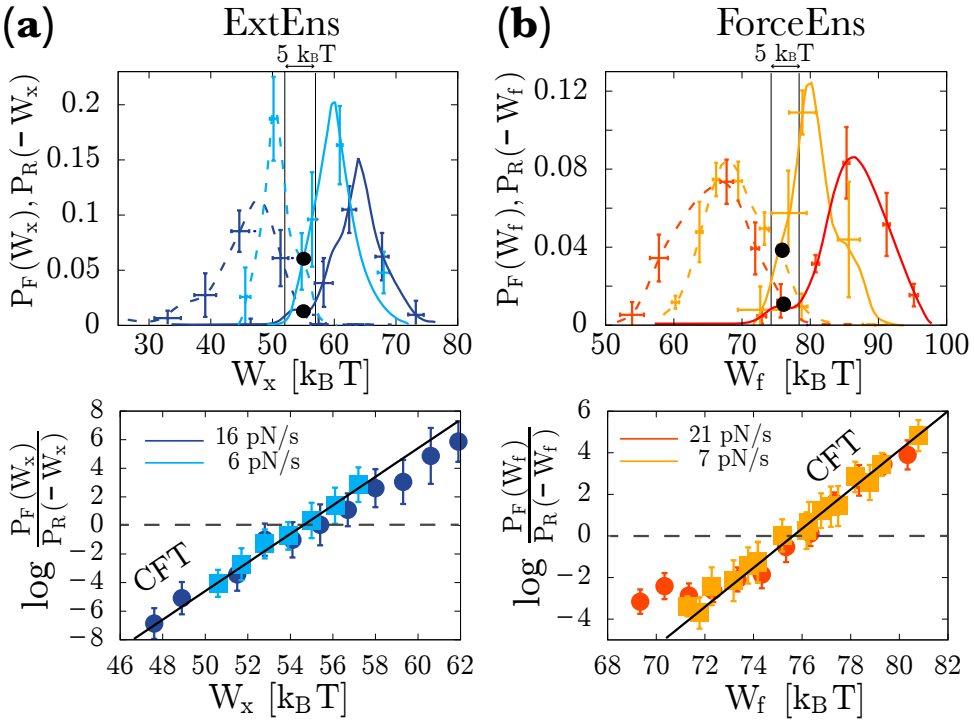

Figure 4.7: **Crooks Fluctuation Theorem for LOT experiments (a)** Work probability distributions and CFT (top and bottom panel, respectively) in the ExtEns (Eq. (4.7)) for ExtEns experiments. **(b)** Work probability distributions and CFT (top and bottom panel, respectively) in the ExtEns (Eq. (4.6)) for ForceEns experiments. In both upper panels solid (dashed) lines correspond to F (R) distributions, while vertical lines correspond to the free energy uncertainty. In both lower panels, solid line corresponds to a straight line with slope equal to 1 and y-intercept equal to -ΔG, in $k_B T$ units.

Typical FDCs fot LOT in the ExtEns (ForceEns) mode are shown in top (bottom) graph of Fig. 4.6(b). Unfolding (folding) events in the ExtEns (top graph of Fig. 4.6(b)) may be seen as sudden force drops (rises) due to the release (absorption) of the molecular extension corresponding to the unwinding of the double helix structure of the DNA hairpin. In the ForceEns (bottom graph of Fig. 4.6(b)) the feature of the transitions

corresponding to the switching between the folded and unfolded conformations occurs at constant force, as we described for the case of MT (see the previous section), leading to a sudden increase/decrease in the molecular extension

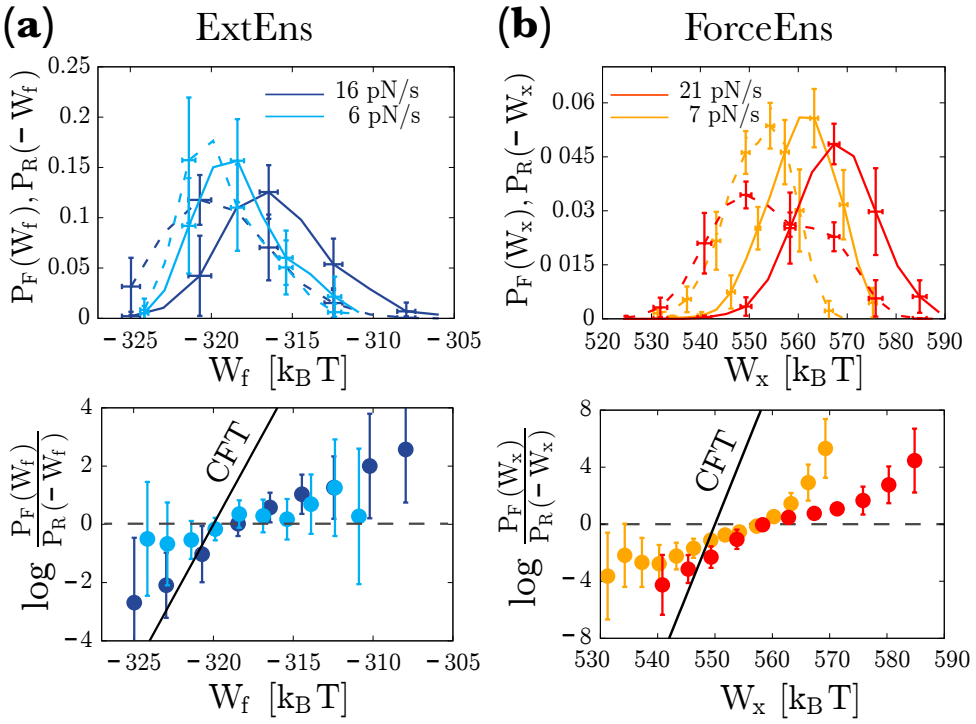

Figure 4.8: **Breakdown of CFT symmetry. (a)** Work probability distributions and CFT (top and bottom panel, respectively) in the ForceEns (Eq. (4.6)) for ExtEns experiments. **(b)** Work probability distributions and CFT (top and bottom panel, respectively) in the ForceEns (Eq. (4.7)) for ForceEns experiments. In both upper panels solid (dashed) lines correspond to F (R) distributions. In both lower panels, solid line corresponds to a straight line with slope equal to 1 in $k_B T$ units.

In all cases, the CFT prediction is fulfilled using the appropriate work definition. In upper panels of Fig. 4.7 we show the work distributions, after subtracting stretching contributions, obtained according to the proper statistical ensemble. The CFT test is shown in the lower panels of Fig. 4.7. We stress that the obtained free energy value is compatible with the theoretical prediction if the work is properly calculated: $\Delta G_0^{\text{ExtEns}} = 52 \pm 5\, k_B T$, $\Delta G_0^{\text{ForceEns}} = 51 \pm 5\, k_B T$.

On the other hand, in Fig. 4.8 we show how the CFT fails when the wrong work definition is used. In particular, in top panel of Fig. 4.8(a) we have obtained the work probability distributions obtained

using the ForceEns work definition (Eq. (4.7)) for the case of experiments performed in the ExtEns. Clearly, the CFT is not satisfied due to the use of a wrong work definition for the ExtEns (bottom panel of Fig. 4.8(a)). The slopes for the 6 and 16 pN/s pulling rates are, respectively, 0.11 ± 0.03 and 0.31 ± 0.02 (both in $k_B T$ units). Top panel of Fig. 4.8(b) shows the distributions obtained using the ExtEns work definition in the case of active mode (ForceEns) experiments. Again, the CFT is clearly not satisfied (bottom panel of Fig. 4.8(b)). The slopes for the 7 and 21 pN/s pulling rates are, respectively, 0.19 ± 0.02 and 0.17 ± 0.01 (in $k_B T$ units).

4.4 CHARACTERIZATION OF THE BOUNDARY TERMS OF THE THERMODYNAMIC WORK

As we already widely discussed, the breakdown of CFT symmetry when the work is not properly computed (i.e. regardless of the statistical ensemble) is due to the effect of the work boundary term: $\Delta(xf) = x_1 f_1 - x_0 f_0$. The measurement of this contribution, yet being direct in MT and LOT experiments, might not be feasible in all scenarios. We show, both for LOT and MT, how it is possible to infer statistical properties of the boundary terms by quantifying the breakdown of the CFT symmetry when an ensemble-wrong work definition is used.

We state the problem as follows. In general, in each type of experiment, only a single definition of the mechanical work satisfies the CFT. While the experiments performed in the ExtEns, require the use of W_x (Eq. (4.6)), the experimental assays done in the ForceEns, require the use of W_f (Eq. (4.7)) instead. As we discussed, both work definitions are related by a Legendre transform. Indeed, Eq. (4.8), can interpreted in terms of sum (and difference) of random variables. In particular, let us assume that work distributions are described by Gaussians (an exact result in the linear response limit). Hence, we can write:

$$p(W_x) = \mathcal{N}\left(\langle W_x \rangle, \sigma^2_{W_x}\right) , \tag{4.10}$$

$$p(W_f) = \mathcal{N}\left(\langle W_f \rangle, \sigma^2_{W_f}\right) , \tag{4.11}$$

where $\langle(\cdots)\rangle$ denotes the average value and $\sigma^2_{(\cdots)}$ the variance of each distribution. Moreover, the moments of the $p(W_x)$ distribution (Eq. (4.10)) are related to the moments of the $p(W_f)$ distribution (Eq. (4.11)) [75]:

$$\langle W_x \rangle = \langle W_f \rangle + \langle \Delta\,(xf) \rangle\,, \tag{4.12}$$

$$\sigma^2_{W_x} = \sigma^2_{W_f} + \sigma^2_{\Delta(xf)} + 2\rho\sigma_{W_f}\sigma_{\Delta(xf)}\,, \tag{4.13}$$

where $\langle \Delta\,(xf) \rangle$ and $\sigma^2_{\Delta(xf)}$ are the mean and variance of the boundary terms and $\rho := \mathrm{Cov}(W_f, \Delta_{xf})/\sigma_{W_f}\sigma_{\Delta(xf)}$ is the correlation coefficient ($\rho \in [-1, 1]$). Note that, for correlated random variables, the variance of their sum might be smaller than the sum of the individual variances due to a negative correlation, $\rho < 0$ (Eq. (4.13)). On the other hand, the average of the sum it is not affected by the non-independence of the random variables. Once set the framework, we divide the following part of the discussion depending whether the experiments have been performed using MT or LOT. We emphasize that our goal is the inference of $\sigma^2_{\Delta(xf)}$ and ρ by studying the breakdown of the CFT due to the use of ensemble-work definitions.

4.4.1 *Boundary terms in Magnetic Tweezers experiments*

We have seen in section 4.3.1 how, for MT experiments (i.e. ForceEns) the CFT is satisfied for W_f (Eq. (4.7)) and it is not for W_x (Eq. (4.6)). Nevertheless, for this latter case, we can consider that W_x does satisfy an effective-CFT that can be written as:

$$\frac{P_F(W_x)}{P_R(-W_x)} = \exp\left(\frac{W_x - \Delta G_x}{k_\mathrm{B} T_\mathrm{eff}}\right) \quad , \quad x := T/T_\mathrm{eff}\,, \tag{4.14}$$

where F and R denote the standard forward and reversed distributions and T_eff is an effective temperature. The x parameter is related to the fluctuation-dissipation ratio in glassy systems [76]. Note that Eq. (4.14) can be rearranged as:

$$\exp\left(-\frac{W_x}{k_\mathrm{B} T_\mathrm{eff}}\right) P_F(W_x) = P_R(-W_x) \exp\left(-\frac{\Delta G_x}{k_\mathrm{B} T_\mathrm{eff}}\right)\,. \tag{4.15}$$

Thus, by integrating over W both sides of Eq. (4.15) we obtain:

$$\left\langle \exp\left(-\frac{W_x}{k_{\mathrm{B}}T_{\mathrm{eff}}}\right)\right\rangle_F = \exp\left(-\frac{\Delta G_x}{k_{\mathrm{B}}T_{\mathrm{eff}}}\right), \tag{4.16}$$

where again $\langle(\cdots)\rangle_F$ denote the average over the F distribution. Since we are considering $p(W_x)$ (and $p(W_f)$) as Gaussian distributions, left-hand side of previous equation (Eq. (4.16)) can be analytically computed using the moment-generating function[4]. After some straightforward algebraic steps we obtain:

$$\begin{aligned}\frac{\sigma^2_{W_x}}{2k_{\mathrm{B}}T_{\mathrm{eff}}} = \langle W_x\rangle - \Delta G_x &= x\frac{\sigma^2_{W_x}}{2\,k_{\mathrm{B}}T} \\ &= x\frac{\sigma^2_{W_f} + \sigma^2_{\Delta(xf)} + 2\rho\sigma_{W_f}\sigma_{\Delta(xf)}}{2\,k_{\mathrm{B}}T}\,.\end{aligned} \tag{4.17}$$

Recalling that in ForceEns W_f fulfils the CFT, for a Gaussian $p(W_f)$ the following expression holds:

$$\frac{\sigma^2_{W_f}}{2\,k_{\mathrm{B}}T} = \langle W_f\rangle - \Delta G_f\,. \tag{4.18}$$

Note that in previous equation the effective temperature does not appear since W_f is the suitable work for the ForceEns. We must also have in mind the fact the relation between ΔG_x and ΔG_f:

$$\Delta G_x = \Delta G_f + \langle\Delta\,(xf)\rangle\,. \tag{4.19}$$

Finally, by subtracting Eq. (4.18) from Eq. (4.17) and using the relations (4.12) and (4.19) we obtain:

$$\sigma^2_{W_f}(x-1) + x\sigma^2_{\Delta(xf)} + 2x\rho\sigma_{W_f}\sigma_{\Delta(xf)} = 0\,. \tag{4.20}$$

Equation (4.20) allows us to link ρ and $\sigma_{\Delta(xf)}$ with x. Furthermore, it can be used as a constraint in the inference procedure (see below).

4 $\mathbb{E}[e^{\pm tX}] = e^{\pm t\langle X\rangle + \frac{1}{2}\sigma^2 t^2}$, being t a parameter and $\langle X\rangle$ and σ^2 the mean and the variance of X, respectively.

The inference of the aforementioned parameters is done using a Maximum Likelihood estimation. Such procedure consists on finding the set of parameters (in our case ρ and $\sigma_{\Delta(xf)}$) that maximize the so-called likelihood of a given statistical model. The likelihood function, $\mathcal{L}\left(\rho, \sigma_{\Delta(xf)} \mid \{W_x\}\right)$, is defined as:

$$\mathcal{L}\left(\rho, \sigma_{\Delta(xf)} \mid \{W_x\}\right) = p\left(\{W_x\} \mid \rho, \sigma_{\Delta(xf)}\right), \tag{4.21}$$

where $\{W_x\}$ are the set of N_{exp} measured W_x and $p\left(\{W_x\} \mid \rho, \sigma_{\Delta(xf)}\right)$ is the joint density function of W_x (which we have assumed to be Gaussian, Eq. (4.10)). Then, for our N_{exp} independent measurements, Eq. (4.21) becomes:

$$\begin{aligned} p\left(W_x^{(1)}, \cdots, W_x^{(W_x^{(1)})} \mid \rho, \sigma_{\Delta(xf)}\right) &= \prod_{k=1}^{N_{\text{exp}}} p\left(W_x^{(k)} \mid \rho, \sigma_{\Delta(xf)}\right) \\ &= \left(2\pi\sigma_{W_x}^2\right)^{-N_{\text{exp}}/2} \exp\left(-\frac{\sum_{k=1}^{N_{\text{exp}}}\left(W_x^{(k)} - \langle W_x\rangle\right)^2}{2\sigma_{W_x}^2}\right). \end{aligned} \tag{4.22}$$

For convenience, we extract the logarithm of Eq. (4.22). The log-likelihood, $\log\mathcal{L}\left(\rho, \sigma_{\Delta(xf)} \mid \{W_x\}\right)$, becomes:

$$\log\mathcal{L}\left(\rho, \sigma_{\Delta(xf)} \mid \{W_x\}\right) = -\frac{N_{\text{exp}}}{2}\log\left(2\pi\sigma_{W_x}^2\right) - \frac{1}{2\sigma_{W_x}^2}\sum_{k=1}^{N_{\text{exp}}}\left(W_x^{(k)} - \langle W_x\rangle\right)^2. \tag{4.23}$$

Since the values which maximize the likelihood also maximize its logarithm, we have numerically maximized Eq. (4.23) imposing the constraint found in Eq (4.20) in order to obtain a simultaneous estimation of ρ and $\sigma_{\Delta(xf)}$. It is important to mention that we have used the experimental value for $\langle W_x\rangle$, so we have not estimated it. On the other hand, we stress that the value of $\sigma_{W_x}^2$ is given by Eq. (4.13).

In Fig. 4.9 we show the F (solid lines) and R (dashed lines) work boundary terms for the experimental data obtained at two different pulling rates (9.0 and 22.5 pN/s). We note that F and R distributions

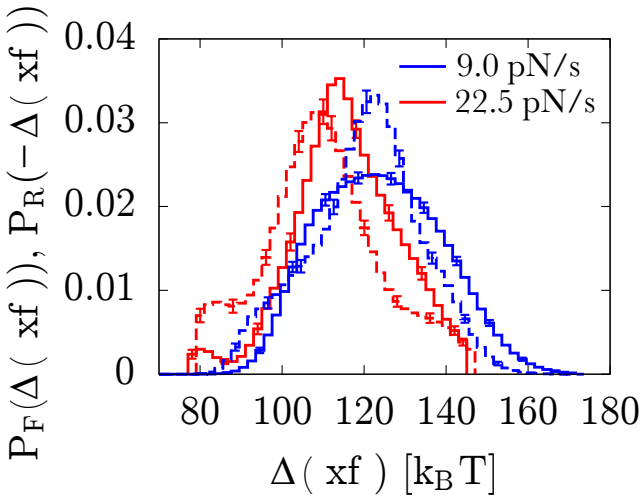

Figure 4.9: **Distribution of the boundary terms in MT experiments.** Probability distributions of the work boundary terms in the F process (solid lines) and R process (dashed lines) for ForceEns experiments.

are almost overlapped, indicating us that there is nearly no dispersion between the F and R boundary terms in MT: $|\Delta(xf)_F| \approx |\Delta(xf)_R|$.

We measured x from the slope of the effective-CFT (see section 4.3.1). We obtained: $x = 0.075 \pm 0.010$ and $x = 0.33 \pm 0.04$ for 9.0 and 22.5 pN/s, respectively. In Table 4.2 we report the values for the correlation ρ and the inferred values of $\sigma^2_{\Delta(xf)}$ (denoted as $\sigma^2_{\Delta^{\text{inf}}_{(xf)}}$) obtained by maximizing Eq. (4.23) for MT experiments when the work is computed according to Eq. (4.6). Furthermore, we also compare the inferred and the experimental ($\sigma^2_{\Delta^{\text{exp}}_{(xf)}}$) values.

r [pN/s]	$\sigma^2_{\Delta^{\text{exp}}_{(xf)}}$ [$(k_B T)^2$]	$\sigma^2_{\Delta^{\text{inf}}_{(xf)}}$ [$(k_B T)^2$]	ρ [ad.]
9.0	254 ± 20	267 ± 50	0.7 ± 0.2
22.5	205 ± 60	188 ± 60	-0.8 ± 0.1

Table 4.2: **Fluctuations of work boundary terms in MT.** The term $\sigma^2_{\Delta^{\text{exp}}_{xf}}$ have been obtained calculating the average between the mean values of F and R distributions and the values $\sigma^2_{\Delta^{\text{inf}}_{xf}}$, ρ have been obtained from the maximization of Eq. (4.23) using the constraint (4.20).

Interestingly, inferred values are in good agreement with the experimental measurements. Surprisingly, correlation coefficient ρ changes its sign upon increasing the pulling speed. On the other hand, fluctuations of the boundary term become smaller as the pulling speed increases. This decreasing trend might indicate us that in the infinite pulling rate regime, $r \to \infty$, relative fluctuations of the boundary term might become negligible as compared to W_x and W_f fluctuations. Hence, in this regime, the CFT might be satisfied both for W_x or W_f.

4.4.2 *Boundary terms in Laser Optical Tweezers experiments*

In what follows we perform the same analysis for LOT experiments. Nevertheless, we must take into account the fact that, using LOT, we carried out experiments both in the ExtEns and in the ForceEns. While in the first case the study is done by investigating the breakdown of the CFT for W_f, in the second case it is done by considering that the CFT does not hold for W_x. For the sake of clarity, we split the discussion in two parts, depending whether the experiments were done in the ForceEns or in the ExtEns.

ExtEns experiments. Breakdown of the CFT symmetry for W_f

According to the framework we discussed in section 4.4.1, for the ExtEns experiments (i.e. the CFT is fulfilled for W_x) we consider that W_f satisfy an effective-CFT given by:

$$\frac{P_F(W_f)}{P_R(-W_f)} = \exp\left(\frac{W_f - \Delta G_f}{k_B T_{\text{eff}}}\right) \quad , \quad x := T/T_{\text{eff}}\,, \tag{4.24}$$

and, consequently:

$$\left\langle \exp\left(-\frac{W_f}{k_B T_{\text{eff}}}\right)\right\rangle_F = \exp\left(-\frac{\Delta G_f}{k_B T_{\text{eff}}}\right)\,. \tag{4.25}$$

Moreover, the Gaussian approximation for $p(W_f)$ (Eq. (4.11)) implies:

$$\begin{aligned}\frac{\sigma^2_{W_f}}{2k_{\mathrm{B}}T_{\mathrm{eff}}} = \langle W_f\rangle - \Delta G_f = x\frac{\sigma^2_{W_f}}{2\,k_{\mathrm{B}}T} \\ = x\frac{\sigma^2_{W_x} + \sigma^2_{\Delta(xf)} - 2\rho\sigma_{W_x}\sigma_{\Delta(xf)}}{2\,k_{\mathrm{B}}T}\,.\end{aligned} \tag{4.26}$$

We point that in previous equation (Eq. (4.26)) the sign of the rightmost term of the final equality has been changed from positive to negative[5]. On the other hand, we can write the analogous expression of Eq. (4.18) for the ExtEns as:

$$\frac{\sigma^2_{W_x}}{2\,k_{\mathrm{B}}T} = \langle W_x\rangle - \Delta G_x\,. \tag{4.27}$$

In order to obtain the equivalent expression of Eq. (4.20), we subtract Eq. (4.27) from Eq. (4.26) and we use the relation between the free energies (Eq. (4.19)) and the result of Eq. (4.12):

$$\sigma^2_{W_x}(x-1) + x\sigma^2_{\Delta(xf)} - 2x\rho\sigma_{W_x}\sigma_{\Delta(xf)} = 0\,. \tag{4.28}$$

As we expected, Eq. (4.28) allows us to impose an additional constraint for the numerical estimation of ρ and $\sigma_{\Delta(xf)}$. The inference of these parameters has been done following the same procedure we explained in section 4.4.1. We note that the log-likelihood function now reads as:

$$\begin{aligned}\log\mathcal{L}\left(\rho,\sigma_{\Delta(xf)} \mid \{W_f\}\right) = & -\frac{N_{\mathrm{exp}}}{2}\log\left(2\pi\sigma^2_{W_f}\right) \\ & -\frac{1}{2\sigma^2_{W_f}}\sum_{k=1}^{N_{\mathrm{exp}}}\left(W_f^{(k)} - \langle W_f\rangle\right)^2\,,\end{aligned} \tag{4.29}$$

where $\sigma^2_{W_f}$ is given by:

$$\sigma^2_{W_f} = \sigma^2_{W_x} + \sigma^2_{\Delta(xf)} - 2\rho\sigma_{W_x}\sigma_{\Delta(xf)}\,. \tag{4.30}$$

5 This is due to the fact that $W_f = \Delta(xf) - W_x$.

In Fig. 4.10(a) we show the F (solid lines) and R (dashed lines) work boundary terms for the experimental data obtained at two different pulling rates (6.0 and 16.5 pN/s). The results we obtained for ExtEns experiments in LOT are reported in Table 4.3. Finally, we indicate that the slopes of the effective-CFT (i.e. x) we used are: 0.11 ± 0.03 and 0.31 ± 0.02, for the 6 and 16 pN/s pulling rates, respectively.

ForceEns experiments. Breakdown of the CFT symmetry for W_x

Since the goal of this part is the study of the breakdown of the CFT due to the use of W_x in ForceEns experiments, the situation is completely analogous to that already explained for MT. Hence, the estimation of the parameters has been done as we explained in section 4.4.1. We note that the slopes of the effective-CFT (i.e. x) we used are: 0.19 ± 0.02 and 0.17 ± 0.01, for the 7 and 21 pN/s pulling rates, respectively.

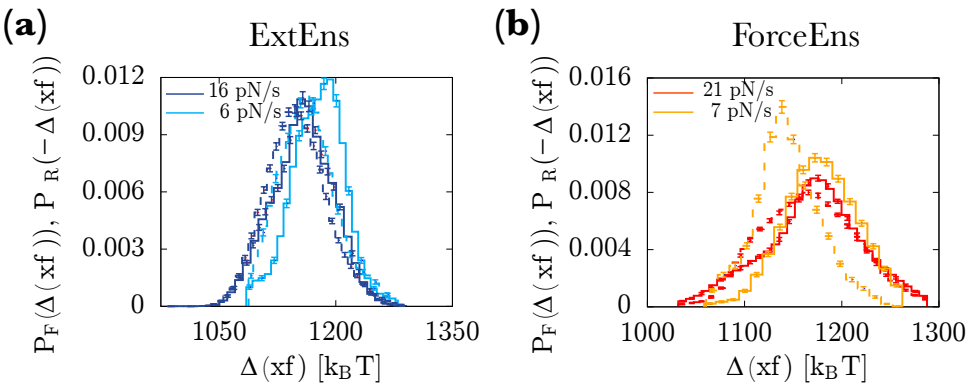

Figure 4.10: **Work boundary term for LOT experiments.** Distributions of the work boundary terms in the F process (solid lines) and R process (dashed lines) for ForceEns experiments **(a)** and for ExtEns experiments **(b)**.

Figure 4.10(b) shows the F (solid lines) and R (dashed lines) work boundary terms (obtained at the aforementioned pulling rates) for ForceEns experiments in LOT. On the other hand, Table 4.3 contains the results of the estimation of $\sigma^2_{\Delta^{\text{inf}}_{xf}}$, ρ based on the numerical maximization of the log-likelihood given by Eq. (4.23).

Summary of results

We note that the fluctuations of the boundary term in LOT are typically ten times bigger than in MT. This might be a result of the fact that the

	r [pN/s]	$\sigma^2_{\Delta^{\mathrm{exp}}_{xf}}$ [$(k_\mathrm{B}T)^2$]	$\sigma^2_{\Delta^{\mathrm{inf}}_{xf}}$ [$(k_\mathrm{B}T)^2$]	ρ [ad.]
ExtEns	6	1265 ± 180	2800 ± 600	0.96 ± 0.04
	16	2460 ± 300	2340 ± 180	0.61 ± 0.02
ForceEns	7	1050 ± 140	2000 ± 250	-0.98 ± 0.02
	21	1560 ± 170	2970 ± 140	-0.99 ± 0.01

Table 4.3: **Fluctuations of the work boundary terms in LOT.** The term $\sigma^2_{\Delta^{\mathrm{exp}}_{xf}}$ have been obtained calculating the average between the mean values of F and R distributions and the values $\sigma^2_{\Delta^{\mathrm{inf}}_{xf}}$, ρ have been estimated by numerically maximizing of Eq. (4.29) using the constraint (4.28). The slopes of the effective-CFT for ExtEns (ForceEns) experiments obtained using W_f (W_x) are: $x = 0.11 \pm 0.03$ ($x = 0.19 \pm 0.02$) and $x = 0.31 \pm 0.02$ ($x = 0.17 \pm 0.01$) for 6 and 16 pN/s (7 and 21 pN/s), respectively.

distance x in the LOT setup, rather than being directly the molecular extension as in MT, is the sum of several fluctuating quantities: the molecular extension (x_m) plus the extension of the dsDNA handles (x_h) plus the displacement of the bead in the optical trap (x_b). While in the ExtEns experiments the total distance $x = x_\mathrm{m} + x_\mathrm{h} + x_\mathrm{b}$ is fixed, its individual components fluctuate. This, added to their different elastic response might induce higher fluctuations in the energetics of the systems as compared to the MT setup. On the other hand, notoriously, the correlation coefficient ρ we infer have two different behaviors depending on the ensemble. Whereas for ExtEns ρ decreases as the pulling rate increases (like we obtained in MT, see Table 4.2), in ForceEns ρ is approximately constant for all pulling rates.

The model we derived is able to reproduce the typical order of magnitude of the work boundary fluctuations. Nevertheless, we must bear in mind that the Gaussian assumption for the work distributions might not be realistic in all situations. Indeed, the distributions corresponding to the natural ensembles of each experimental system (Fig. 4.9 for ForceEns in MT and Fig. 4.10(a) for ExtEns in LOT) are symmetric (as the Gaussian assumption requires), while the corresponding one to the ForceEns in LOT (Fig. 4.10(b)) is not.

4.5 CONCLUSIONS

In this chapter we have addressed the issue of ensemble inequivalence from a thermodynamic perspective. We performed nonequilibrium bidirectional pulling experiments in a small DNA hairpin and we have explored two conjugate ensembles: ExtEns and ForceEns. In particular, we carried out experiments in the ForceEns, both with MT and LOT with force feedback, and in the ExtEns with LOT. This has allowed us to perform the definitive experimental verification of Eq. (4.7) by using the CFT, indicating that in the ForceEns Eq. (4.7) measures the correct thermodynamic work and that it is not equivalent to using Eq. (4.6).

Moreover, by comparing the ForceEns and the ExtEns we have shown the importance of the often neglected boundary terms of the measured mechanical work. They play a pivotal role when testing the CFT and, since they strongly depend on the experimental conditions, their study may allow experimentalists to gather useful information about fluctuations of the different parts of experimental setup. Moreover, we have exemplified this fact by developing a solvable model that allows us to infer the statistical properties of the work boundary terms.

Our study paves the way to the extension of free energy recovery methods using fluctuation theorems in situations in which only intensive variables (such as force) are controlled.

5

KINETICS AND DISSIPATION IN THE FORCE ENSEMBLE

In the preceding chapter we have explored the issue of ensemble inequivalence in small systems by exploring several thermodynamic quantities in two conjugate statistical ensembles (ExtEns vs. ForceEns).

As a matter of fact, we found the effect of Brownian fluctuations goes beyond thermodynamic effects. In the present chapter we will analyze how the folding/unfolding kinetics of DNA hairpins are affected by whether intensive variables (e.g. force or pressure) rather than extensive ones (e.g. extension or volume) are controlled, causing also subsequently notorious differences in dissipation.

The chapter is organized as follows: in the first section we will briefly expose the concept of the Free Energy Landscape (FEL) and how the molecular folding/unfolding problem is mapped onto it. Then, in the second section the theory beneath the folding/unfolding kinetics is introduced, leading to the quantification of dissipation and irreversibility effects in the ForceEns and the ExtEns. Then, the implications to liquid systems are finally discussed.

5.1 THE FREE ENERGY LANDSCAPE: A BRIEF REMINDER

The search for specific (or targeted) molecular conformations has drawn attention since the birth of molecular biophysics. As pointed out by Levinthal [77], even for a small protein, the number of possible states that the molecule can explore is gigantic. Additionally, the possibility of exploring a certain molecular state does not only depend on its thermodynamic stability but also on kinetic considerations. The boost of SME has fuelled the study of molecular folding problem [78, 79]. As a matter of fact, the feasibility of applying mechanical force to biomolecules has allowed experimentalists to explore kinetic states that often remain hidden in bulk assays [80].

In statistical physics the concept of FEL is widely used to obtain a relation between the free energy of a system as a function of all its

available configurations. It is widely used as a tool to predict reaction pathways in chemical reactions, to infer thermodynamically stable states in molecular systems (such as proteins or nucleic acids) or even to study glassy systems [81] and mathematical optimization problems in computer science [82].

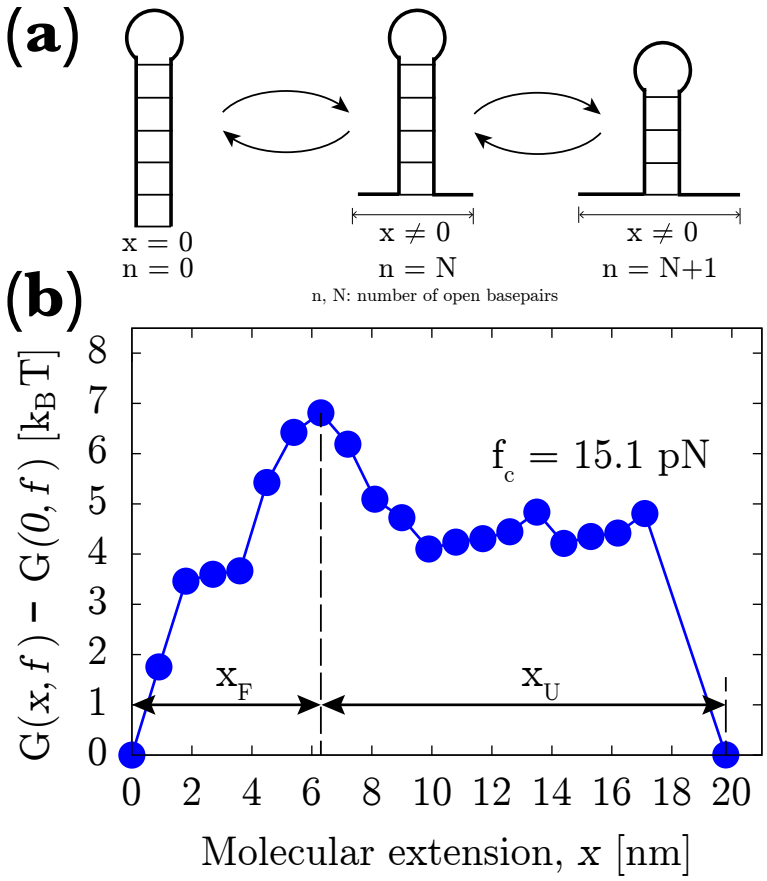

Figure 5.1: **Configurations and free energy landscape of CD4 DNA hairpin. (a)** - Sketch of the sequential configurations that appear when unzipping the hairpin. Leftmost configuration corresponds to the situation in which there are no free bases, whereas the following ones correspond to situations in which there are $n = N$ and $n = N + 1$ released bases. x denotes the molecular extension corresponding to n. **(b)** - FEL evaluated using Eq. (5.1) at the coexistence force, f_c. Using the NN parameters [73]: $f_c = 15.1$ pN at T = 298 K and 1M NaCl. The vertical dashed line corresponds to the position of the transition state and $x_{F(U)}$ are the distances from the folded (unfolded) state to the transition state.

The FEL is, mathematically, a continuous non-bijective function that sets a correspondence between each configuration of a physical system

and its free energy. Configurations are labelled according to a *reaction coordinate*. When applying a mechanical force at the extremities of a molecule (e.g. such as in LOT or MT experiments), the good reaction coordinate is the molecular extension. Nevertheless, for DNA hairpins in which the unzipping process is sequential, the number n of released (or unpaired) bases is also a good reaction coordinate [32]. To calculate the FEL when an external force f is applied to the hairpin we used the following expression:

$$G(n,f) = G_0^n + G_{\text{stret.}}(n,f) + G_{\text{diam.}}(f)\,, \tag{5.1}$$

where the term G_0^n accounts for the free energy of formation at zero force of the configuration in which n sequential basepairs are unpaired (see Fig. 5.1(a)). It is fully sequence-dependent and it is computed according to the Nearest-Neighbor model for DNA [72, 83] as:

$$G_0^n = \sum_{k=n+1}^{N} g_{k,k+1} + (1-\delta_{n,N})g_{\text{loop}}\,, \tag{5.2}$$

where the terms in the summation, $g_{k,k+1}$ are the basepair free energies, g_{loop} is the free energy of formation of the loop and $\delta_{n,N}$ is the Kronecker delta. The used values for $g_{k,k+1}$ are obtained from the Mfold web server [73].

The elastic response of the released single-stranded nucleic acid is modelled according to the Worm-Like Chain (WLC) model [84]. Thus, the stretching free energy at a fixed force f is given by:

$$G_{\text{stret.}}(n,f) = -\int_0^f x_n(f')\,df'\,, \tag{5.3}$$

where $x_n(f)$ is the extension of the ssDNA at the force f when n bases are released. $x_n(f')$ is calculated as the inverse function of Eq. (B.11) using $P = 1.35$ nm and $d_b = 0.59$ nm/base [85]. The contour length of the ssDNA equals to $L_c^n = (n + n_{\text{loop}}\delta_{n,N})d_b$, being n_{loop} the number of bases that form the loop ($n_{\text{loop}} = 4$ for the hairpin shown in Fig. 4.2(a)).

Finally, the energy cost to orientate the double helix diameter ($d = b = 2$ nm) along the direction of the force is evaluated using the Freely Jointed Chain (FJC) model [78], giving:

$$G_{\text{diam.}}(f) = -k_{\text{B}}T \log\left(\frac{k_{\text{B}}T}{fd}\sinh\left(\frac{fd}{k_{\text{B}}T}\right)\right) . \tag{5.4}$$

In Fig. 5.1(b) it is shown the calculated FEL for the CD_4 DNA hairpin at the coexistence force f_c. That is, the force at which the unfolded and folded states have the same energy: $G(n = 0, f_c) = G(n = N, f_c)$. From the estimated FEL, that presents only two minima separated by a single barrier, we can infer that the CD_4 DNA hairpin behaves as a two-state system with two stable conformations: the folded state (corresponding to $n = 0$) and the unfolded state ($n = N$).

5.2 TWO-STATE KINETIC RATES IN A NUTSHELL

As we already commented, small DNA hairpins (such as CD_4) behave like two-state systems. Therefore, under the action of an external force f, the DNA can switch between two states: the folded (F) and the unfolded (U) state (for a schematic depiction see Fig. 5.2).

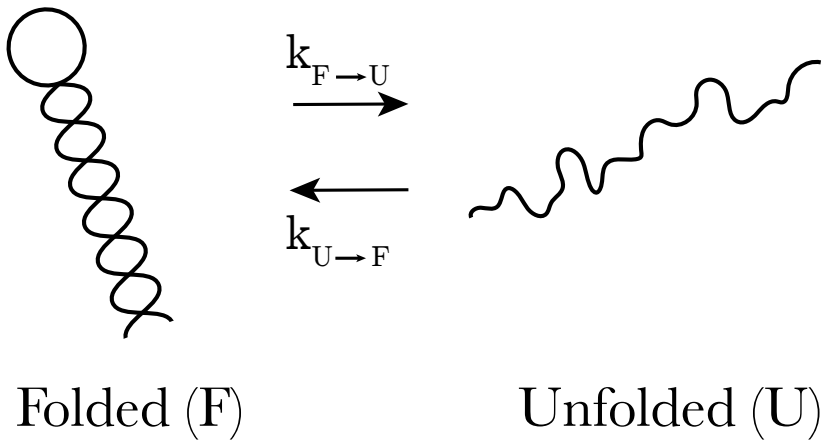

Figure 5.2: **Schematics of two-state systems.** Switching between the folded and the unfolded conformation of a DNA hairpin according to the unfolding kinetic rate, $k_{\text{F}\to\text{U}}$, and the folding kinetic rate, $k_{\text{U}\to\text{F}}$.

Since in the FEL it appears a single kinetic barrier (i.e. the state with highest free energy along the reaction coordinate, dashed line in Fig. 5.1), the hairpin is considered to behave like a two-state system

[86]. Unfolding and folding rates are usually described according to Kramers Bell-Evans theory [87–90] with kinetic rates $k_{F\to U}$ and $k_{U\to F}$ for transitions F $\to$ U and U $\to$ F, respectively. Rates can be written as:

$$k_{\mathrm{F}\to\mathrm{U}}(f) = k_m \exp\left(\frac{f x_{\mathrm{F}}}{k_{\mathrm{B}} T}\right), \tag{5.5}$$

$$k_{\mathrm{U}\to\mathrm{F}}(f) = k_m \exp\left(\frac{\Delta G_{\mathrm{FU}} - f x_{\mathrm{U}}}{k_{\mathrm{B}} T}\right), \tag{5.6}$$

where k_m is the unfolding kinetic rate at zero force, $\Delta G_{\mathrm{FU}} = f_c x_m$ is the free energy difference between states F and U and $x_m = x_{\mathrm{F}} + x_{\mathrm{U}}$ is the molecular extension at the force f_c.

5.3 AVERAGE DISSIPATED WORK

The Second Law of thermodynamics sets the free energy, ΔG, as a lower bound for the average mechanical work, $\langle W \rangle$, done over a set of different experimental realizations of an arbitrary experimental protocol as: $\langle W \rangle \geq \Delta G$. The excess work $\langle W \rangle - \Delta G$ is often referred to as the *dissipated work* and it strongly depends on the experimental conditions, $\langle W_{\mathrm{dis}} \rangle = \langle W \rangle - \Delta G$.

It has been recently shown that the dissipated work provides a measure of distinguishability between forward and backward trajectories in the phase space [91], providing a direct link to information thermodynamics. From a different perspective, we found that irreversibility effects and dissipation are another sign of ensemble inequivalence.

In our bidirectional pulling experiments we can estimate the average dissipated work per cycle as:

$$\langle W_{\mathrm{dis}} \rangle \simeq \frac{\langle W \rangle_F - \langle W \rangle_R}{2}, \tag{5.7}$$

where $\langle W \rangle_{F(R)}$ is the mean value of F(R) work distribution. The main advantage of using Eq. (5.7) is that the neither the knowledge of the free energy nor the stretching contributions are required.

5.3.1 *Experimental results for dissipation and kinetic rescaling*

In Fig. 5.3 it is shown $\langle W_{\text{dis.}} \rangle$ as a function of the pulling rate r for all experiments: ForceEns in MT; and ForceEns and ExtEns in LOT. Note that, under equivalent pulling rate conditions, dissipation is systematically lower in the ExtEns as compared to the ForceEns.

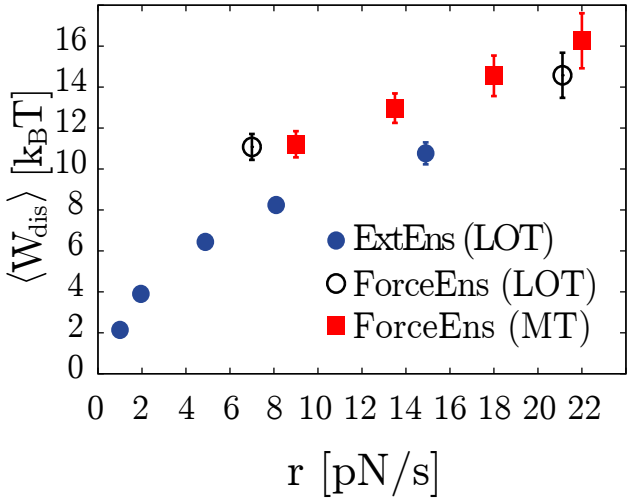

Figure 5.3: **Dissipation in the ForceEns and in the ExtEns.** Comparison between the average dissipated work in the ForceEns with MT (red full squares), ExtEns (full dark blue circles) and the ForceEns in LOT (empty light circles) (obtained from Ref. [74]).

The difference found in the average dissipation between the ForceEns and the ExtEns relies on the molecular kinetics. In the ForceEns, the unfolding-folding transitions occur at constant force, keeping the kinetics unchanged. In contrast, in the ExtEns, every unfolding-folding event is followed by a force jump, speeding up the kinetics as compared to the ForceEns. In Fig. 5.4 there are shown schematic depictions of an arbitrary unfolding event (solid lines) in the ForceEns and in the ExtEns when kinetics are described according to Eqs. (5.5) and (5.6). Kinetic rates (and hence the overall relaxation time) are always higher in the ExtEns, which leads to lower dissipation.

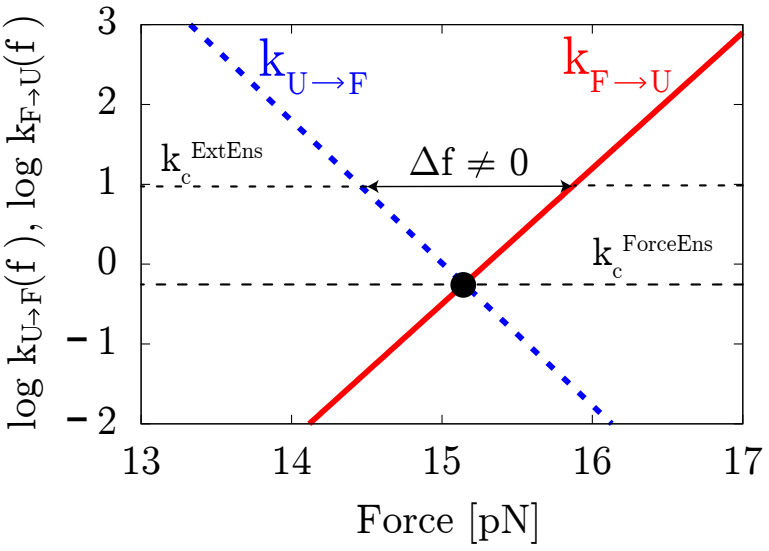

Figure 5.4: **Illustration of ensemble dependence of coexistence kinetic rates.** Hopping kinetics at coexistence in the ForceEns (fixed point) and the ExtEns (two arrow line), Δf is the force jump when the molecule unfolds.

In Ref. [92] an expression for the average dissipated work, $\langle W_{\text{dis.}} \rangle$, has been derived for a two-state DNA hairpin if an external mechanical force is varied at a constant pulling rate f. It reads as:

$$\frac{\langle W_{\text{dis.}} \rangle}{k_{\text{B}}T} = \int_{-\infty}^{\infty} dx \int_{-\infty}^{x} \frac{dy}{\cosh^2(y)} \exp\left(\frac{-1}{\tilde{r}} \int_{y}^{x} dz\, e^{\mu z} \cosh z \right), \tag{5.8}$$

being μ the molecular fragility [93] defined as:

$$\mu = \frac{x_{\text{F}} - x_{\text{U}}}{x_{\text{F}} + x_{\text{U}}} = \frac{x_{\text{F}} - x_{\text{U}}}{x_m}, \tag{5.9}$$

and the dimensionless rate $\tilde{r}$:

$$\tilde{r} = \frac{x_m}{4\, k_{\text{B}} T k_c} r, \tag{5.10}$$

where k_c is the so-called critical coexistence rate of the F and U states (i.e. $k_{F\to U}(f_c) = k_{U\to F}(f_c) := k_c$). Interestingly, although Eq. (5.8) was derived in the ForceEns scheme, through a kinetic rescaling, it can be

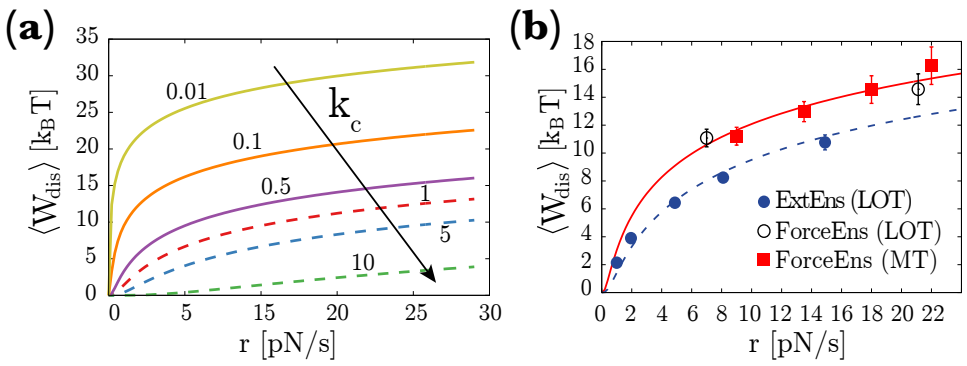

Figure 5.5: **Kinetic rescaling for average dissipated work. (a)** $\langle W_{\text{dis.}} \rangle$ as a function of the pulling rate r for different values of k_c (indicated in the graph). **(b)** Theoretical prediction for ExtEns data (dashed line) and ForceEns data (solid line). $\langle W_{\text{dis.}} \rangle$ has been obtained by numerically integrating Eq. (5.8) using $x_m = 19.8 \pm 0.9$. $\mu =$ -0.3 $\pm$ 0.1 and $k_B T = 4.11$ pN nm at 298 K.

used to characterize dissipation in the ExtEns. In Fig. 5.5(a) it is shown how the average dissipated work (obtained by numerically integrating Eq. (5.8)) spans from few $k_B T$s up to tens of $k_B T$s by varying the value of k_c. As a matter of fact, in order to reproduce the ExtEns behavior using Eq. (5.8), it has been shown that kinetic rates at the coexistence transition must be appropriately rescaled as [92]:

$$k^{\text{ForceEns}} = \Omega k^{\text{ExtEns}} , \tag{5.11}$$

where the Ω factor equals to:

$$\Omega = \exp\left(-\frac{1-\mu^2}{8} \frac{x_m |\Delta f|}{k_B T} \right) < 1 . \tag{5.12}$$

Since the folded-unfolded transition in the ForceEns occurs at constant force ($\Delta f = 0$), the force jump must be measured in the ExtEns. For the CD_4 DNA hairpin (Fig. 4.2(a)) we obtain $x_m = 19.8 \pm 0.9$ nm, $\Delta f = 1.1 \pm 0.1$ and $\mu =$ -0.3 $\pm$ 0.1, leading to $\Omega = 0.55 \pm 0.02$ at $T = 298$ K. Moreover, the kinetic rate at coexistence was measured from hopping experiments in the ExtEns [71]: $k^{\text{ExtEns}} = 1.3 \pm 0.2\ \text{s}^{-1}$, giving $k^{\text{ForceEns}} = 0.72 \pm 0.11\ \text{s}^{-1}$. Using these values we find good agreement between theory and experiments, as it can be seen in Fig. 5.5(b).

5.4 DISCUSSION

We found that the ensemble inequivalence phenomenon is also present at the level of molecular kinetics. We showed that the average dissipated work, which is essentially governed by the molecular kinetics, strongly depends on the nature of the control parameter.

In general, fluctuations of intensive variables in the ExtEns leads to effective higher kinetic rates in thermally activated processes. The characteristic Arrhenius dependence of kinetic rates, $k \sim \exp\left(-B/\,k_{\mathrm{B}}T\right)$, and the fluctuating nature of the kinetic barrier, B, together with Jensen's inequality[1] give:

$$k^{\mathrm{ExtEns}} \sim \langle \exp\left(-B/\,k_{\mathrm{B}}T\right)\rangle > \exp\left(-\langle B\rangle/\,k_{\mathrm{B}}T\right) \sim k^{\mathrm{ForceEns}}\,. \quad (5.13)$$

In turn, in linear response, the average dissipated work is expected to scale like: $\langle W_{\mathrm{dis.}}\rangle \sim P/k$, with P a characteristic driving power ($\sim x_m r$ in our pulling experiments), giving $\langle W_{\mathrm{dis.}}^{\mathrm{ExtEns}}\rangle < \langle W_{\mathrm{dis.}}^{\mathrm{ForceEns}}\rangle$. Note that Eq. (5.12) can be written as: $\Omega = \exp\left(-a|\langle \Delta x \Delta f\rangle|/\,k_{\mathrm{B}}T\right)$, with $\Delta x = x_m$ and $a = (1-\mu^2)/8 \sim \mathcal{O}(1)$.

We believe that the conclusions of our single-molecule study might be generalized to other physical contexts. As a matter of fact, in the pressure-volume context of liquids we argue that the Ω factor would read as:

$$\Omega = \exp\left(-b\frac{|\langle \Delta V \Delta P\rangle|}{k_{\mathrm{B}}T}\right) = \exp\left(-b\frac{(\Delta P)^2 V \kappa_T}{k_{\mathrm{B}}T}\right)\,, \quad (5.14)$$

where $b \sim \mathcal{O}(1)$, V being the volume, ΔP the root-mean square deviation of pressure fluctuations and κ_T the isothermal compressibility.

An important consequence of Eq. (5.14) is that, given the fundamental thermodynamic uncertainty relation between pressure and volume fluctuations [94]: $|\langle \Delta P \Delta V\rangle| > k_{\mathrm{B}}T$, it is expected that the ForceEns and the ExtEns will recover the property of equivalence among them in the high T limit.

Ensemble inequivalence might be important in *in vivo* molecular reactions. As a matter of fact, Eq. (5.14) allows us to exemplify it. Let us consider a cell of typical size 10 μm with fixed volume $V =$

1 $e^{\mathbb{E}[X]} \leq \mathbb{E}\left[e^{X}\right]$.

$10^3 \mu\text{m}^3$, osmotic pressure fluctuations $\Delta P \approx 100$ Pa (osmotic pressure differences can be as large as 300 Pa [95]) and isothermal compressibility κ_T of water as small as 4×10^{-10} Pa^{-1}. Inserting these values in Eq. (5.14) with $k_\text{B}T = 4.11$ pN nm $= 4 \times 10^{-21}$ N m at $T = 298$ K we obtain: $\Omega \approx \exp(-b)$, which is of order 1 if b is of order 1, as assumed. This result suggests that the kinetics of molecular reactions inside cellular compartments [96, 97] might be strongly sensitive to the ensemble.

Since the right-hand side of Eq. (5.14) is strongly sensitive to the three terms appearing in the exponent: ΔP, V and κ_T, the figures employed in the previous expression for Ω should be taken only as a guide. Indeed, for the case of molecular reactions in much smaller compartments, V can be a thousand times smaller. Nevertheless, the magnitude of the pressure fluctuations, ΔP, can be comparatively larger. Also, κ_T must not be necessarily as small as for pure water[2], the bulk modulus of the cellular solvent could be larger at finite frequencies under nonequilibrium conditions. In this regard, SME of molecular folding in crowded environments offer an interesting research track to follow.

2 $\kappa_T = 4 \times 10^{-10}$ Pa^{-1}

Part III

INFORMATION-CONTENT OF MOLECULAR ENSEMBLES

6

SINGLE-MOLECULE CHARACTERIZATION OF HETEROGENEOUS NEUTRAL MOLECULAR ENSEMBLES

6.1 MOTIVATION

Through complex non-equilibrium dynamics, living organisms can grow, undergo metabolism, reproduce or even evolve. Nonetheless, those features are hard to explain in a classical thermodynamic scenario and concepts like information must be added as an ingredient. Indeed, living systems behave like a Maxwell demon, as they can measure and exploit information from their surroundings. From chemotaxis (i.e. the ability of bacteria to move towards regions with nutrien-richt concentrations) [98, 99], to communication [100] or even to the adaptation to changing environments [101], living organisms are able of harvesting information from their surroundings and, subsequently, of effectively transducing the obtained information into useful energy.

In the early years of information theory, information-to-energy conversions were assumed to be only *gedankenexperiments*. Indeed, quoting Schrödinger [102]: "*We never experiment with just one electron or atom or (small) molecule. In thought experiments we sometimes assume that we do; this invariably entails ridiculous consequences (...) In the first place it is fair to say that we cannot experiment with single particles, any more than we can raise Ichtkyosauria in the zoo*". Fortunately this situation has not become true. Nowadays, experimental measurements of information-contents are a hot topic in the field of statistical mechanics. From the experimental demonstration of information-to-work conversion [56], to the verification of the Landauer principle [103], the link between information theory and thermodynamics is finally well-built [55]. In this regard, the measurement of information content of populations has become essential due to its applications in directed evolution experiments.

Evolution is the natural process that, through genotipic variations, generates phenotipic variations. Natural evolution has two main in-

gredients: variability and selection. Whereas selection is related to the change of heritable traits of a population over time, variability brings out evolution. Moreover, evolutionary processes occur across all spatial and temporal scales, from species and organisms down to the molecular level. In the context of information theory, evolving populations are continuously adapting to changing environment, so they must extract and store information from their surroundings in order to improve their adaptability. Therefore, a precise knowledge (and quantification) of the information content of evolving systems is essential to understand the information-to-energy trade-offs that underlie evolutionary dynamics.

In this part of the thesis we prove how the information content of a DNA molecular ensemble can be directly measured by extending the traditional combination of precise single-molecule measurements and fluctuation relations for mechanical work. This paves the way for the resolution of the long-standing research question of whether information content is or is not a physically measurable quantity. Our method is built on what we call **ensemble force spectroscopy**, a powerful systematic experimental procedure that allows to overcome the evident difficulty of single-molecule methods: sampling enough molecules in order to have a precise measurement of ensemble properties (sample versus population characterization). In this context, there is a fundamental interest in characterizing heterogeneous ensembles. First, and foremost, heterogeneity has an enormous (and often overlooked) impact on biophysical systems: from spatial conformations of molecules (such as proteins, nucleic acids or even viruses) [104–107], up to oncologic implications at the cellular level [108–110]. Moreover, since there is no systematic procedure to study heterogeneity at the single-molecule level, our work aims to fill this gap by establishing a framework in which structural and kinetic properties of molecular ensembles can be obtained by means of thermodynamic measurements in SME. The precise knowledge of thermodynamic properties of the molecular ensembles will, ultimately, allow us to relate the information content with physical properties of the molecular ensembles.

This chapter is focused on the thermodynamic ensemble characterization, aiming to provide a complete and systematic description of an heterogeneous DNA ensemble. Building on this result, the connection between energy and information in molecular ensembles is discussed in the chapter 7.

6.2 MOLECULAR ENSEMBLES

In order to pose the long-standing question of whether information content is an actual physically measurable quantity in molecular ensembles, we need to aim for a suitable physical system. For the case of molecules studied by means of single-molecule assays, the possibility of conducting extremely high accurate thermodynamic measurements (e.g. measuring free energy differences) suggest that they are the perfect playground to explore the connection between energy and information. Nevertheless, in the best tradition of SME, only one molecule is studied at a time. Hence, the information gained at each experimental realization is the same (within experimental uncertainty), yielding a zero formation content of an homogeneous sample. In order to generate the suitable system to carry out information-content measurements, variability must be added as an ingredient.

First of all, we argue that that folding free energy is the key quantity that we must look for. Folding free energy is one of the most straightforward measurable quantity in SME, becoming a useful phenotype (i.e. an observable trait) in many situations. Through free energy measurements, further thermodynamic quantities can be obtained. This is the case of energy dissipation, which is of remarkable importance for living systems (see section 5.3) or the quantification of specific ligand binding energy of, for instance, ions or small ligands (see later discussion of Sec. 8.4). We will show how the measurement of free energy differences in molecular ensembles allows us, ultimately, to conduct systematic information content measurements in heterogeneous molecular ensembles.

Variability is introduced in the molecular sample by performing random uniform[1] point mutations in some specific bases of a given DNA molecule. In particular, we have used as a "template" molecule, the CD_4 DNA hairpin (the same we used in chapter 4). Since force kinetics and energetics of CD_4 DNA are have been widely measured by means of, bulk and single-molecule measurements, CD_4 is a benchmark molecule for a large variety of experimental assays [32]. For this reason we have used CD_4 as the base molecule for generating three different heterogeneous samples: consisting on four, eight and twelve randomized

1 In the sense that the insertion of any nucleotide is equally probable. Therefore, $p(A) = p(C) = p(G) = p(T) = 1/4$.

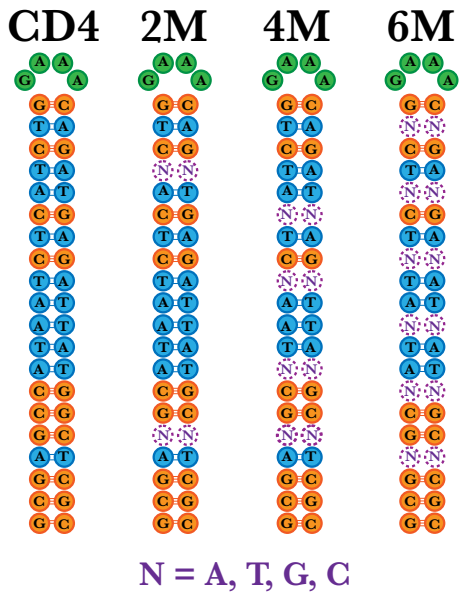

Figure 6.1: **Schematics of molecular ensembles.** Each hairpin corresponds to a different molecular ensemble. Orange bases correspond to GC canonical basepairs whereas blue correspond to AT canonical basepairs. Green bases are loop bases and purple N bases can be either A, G, C or T.

Ensemble	Population, Ω
CD4	$4^0 = 1$
2M	$4^4 = 256$
4M	$4^8 = 65536$
6M	$4^{12} = 16777216$

Table 6.1: **Summary of molecular ensembles and populations.** List of the number of different molecules that are compatible with a given number of randomized nucleotides.

nucleotides (for the specific positions, see Fig. 6.1). Each ensemble is coined as 2M, 4M and 6M, respectively. In Table 6.1 we report the number of different molecules that are compatible with a given number of mutations. For an arbitrary number N of mutated bases, the total amount of existing molecules equals to: $\Omega = 4^N = 2^{2N}$.

Clearly, for traditional single-molecule techniques (no high-troughput or paralleling) the possibility of measuring all (or a significant fraction) of the molecules existing in a mutational ensemble becomes unattainable as N grows (see Table 6.1). The necessity of overcoming this difficulty is essential for measuring ensemble quantities (e.g. information content) of real molecular systems.

We note that our heterogeneous DNA ensembles are analogous to DNA ensembles undergoing evolutionary dynamics without selecting

forces. That is, obtained only by means of genetic drift (i.e. neutral ensembles).

6.3 ENSEMBLE FORCE SPECTROSCOPY

The measurement of information contents of molecular ensembles is done by characterizing the folding free energy spectrum of the molecular ensembles. Here we propose a systematic methodology that allows us to quantify the folding free energy distributions of heterogeneous ensembles by the combination of classical nonequilibrium pulling experiments and the CFT. With our methodology we are able to extract the folding free energy spectra of neutral molecular ensembles. In the best tradition of statistical physics, a pool of several tens of molecules out of a large population it is sufficient to extract the information-content.

Pulling experiments were performed analogously as explained in section 4.3.1. Additionally, DNA hairpins are also attached to the same two 29-bp dsDNA handles we used in the experiments described in chapter 4 and the molecular construct is inserted between two micron-sized polystyrene beads, as shown in Fig. 6.2(a). While one bead is held in the tip of a micropipette by air suction, the other one is captured in the center of the optical trap. The distance λ is the relative distance between the center of the optical trap and the tip of the micropipette. All the assays are performed in the ExtEns, where λ is the control parameter.

The protocol applied for characterizing the ensembles is as follows. Every sampled DNA hairpin is subjected to bidirectional nonequilibrium pulling experiments. Molecules are stretched by increasing the distance λ, whereas they are relaxed by decreasing λ. The stretching (releasing) path is identified with the forward (reversed), F (R), process. At the beginning of the F process, the molecule is in thermal equilibrium at a given value of $\lambda := \lambda_0$ in the folded conformation and it is mechanically unfolded by performing a controlled increase of λ. When the molecule does no longer withstand the mechanical tension, it unfolds to its single-stranded conformation. Unfolding events can be seen as sudden force drops in the F process. On the other hand, in the R process, the molecule is initially found in equilibrium at $\lambda := \lambda_1$ in the unfolded conformation and the time reversed protocol is applied until again λ_0 is reached. The footprint of the refolding of the molecules is

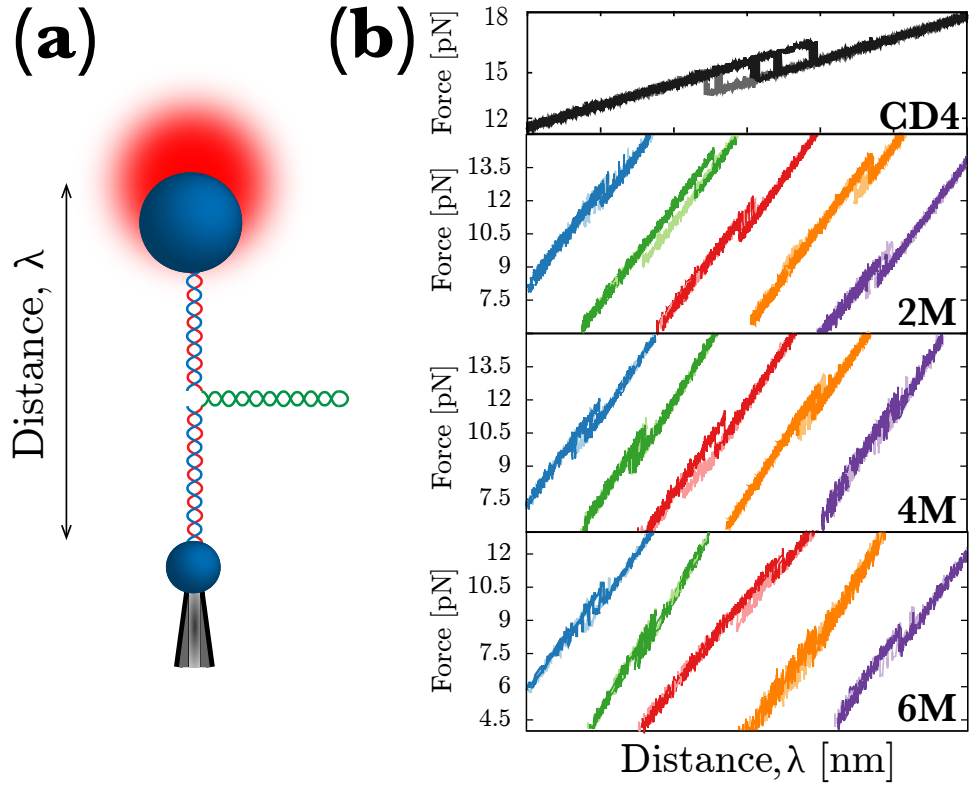

Figure 6.2: LOT **experimental setup and experimental evidence of heterogeneity.** **(a)** - Schematics of the experimental setup of LOT experiments and representation of the control parameter λ. **(b)** - Examples of FDCs of CD4 (top panel) and different molecules of the three ensembles we studied (from top to bottom: 2M, 4M, 6M; see Fig. 6.1(a)). Curves were shifted for the sake of clarity. Dark colors corresponding to unfolding process and light colors correspond to the refolding paths.

an abrupt force increase, corresponding to the formation of the original double helical structure.

In Fig. 6.2(b) we show several FDCs for different molecules and different molecular ensembles (indicated in the right bottom side of each panel). All molecules were pulled at the same conditions (Tris-HCl 1M NaCl buffer, 200 nm/s of pulling speed), so the mechanical response and, in particular, the unfolding/folding forces is only dependent on the sequence. Nevertheless, we point that our assays are sequence-blindfolded, so we are not able to determine the sequence we are measuring by means of standard pulling experiments.

A key quantity when studying molecular ensembles in evolutionary processes is the molecular stability. Due to the features of our experi-

mental procedure, the CFT (Eq. (4.4)) is the perfect tool to recover the individual molecular folding free energy, ΔG_0, using irreversible work measurements. Also, we must take into account two aspects of the mechanical work W. First, W has to be calculated according to the ExtEns scheme (Eq. (4.6)). Additionally, when comparing the mechanical work (or related quantities) of different molecules (e.g. work distributions or dissipation), stretching contributions must be subtracted as described in appendix C.

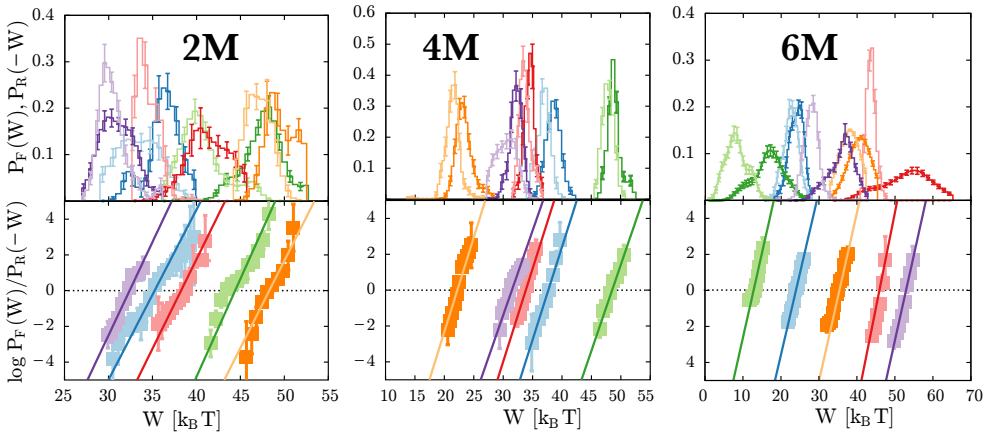

Figure 6.3: **Work distributions and CFT verification. Top panels:** forward (dark colors) and reversed (light colors) work distributions of a selection of molecules belonging to each ensemble (written as in top graphs). **Bottom panels:** CFT test for the corresponding upper distributions. Solid lines are linear fits to the experimental data. All slopes are approximately 1 in $k_B T$ units. Error bars have been calculated using the Bootstrap method.

In upper panels of Fig. 6.3 we show the measured work distributions for different molecules belonging to the same molecular ensemble (the ensemble is indicated in the graph), whereas lower panels show the usual CFT test (i.e. plotting $\log P_F(W)/P_R(-W)$ and fitting a straight line to the experimental data). Interestingly, even though the work distributions can be significantly different among molecules of the same ensemble, all of them fulfil the CFT. We remark that, in order to compare the work distributions of different molecules, for each value of W, the intrinsic elastic contributions of the experimental setup have been subtracted as explained in appendix C. Hence, according to Eq. (4.4), the crossing point of the work distributions of upper panels of Fig. 6.3 (or the x-intercept of each data set shown in lower panels of Fig.

6.3) equals to the folding free energy of the corresponding molecule ΔG_0.

6.3.1 *Folding free energy spectra*

We extracted the folding free energy of every molecule that we measured (40, 93 and 54 molecules for the 2M, 4M and 6M ensembles, respectively) by means of the CFT (see Fig. 6.3). In order to obtain ΔG_0, the energetic contributions of the elements forming the experimental setup have been subtracted to the individual ΔGs. We mention that we have used the effective stiffness approximation for the energetic contribution of the handles plus the optical trap (see Sec. C.0.1) and the ssDNA elastic parameters reported in Ref. [85] (i.e. persistence length equal to 1.35 nm and contour length equal to 0.59 nm/base). Despite there are sophisticated single-molecule studies regarding sequence-dependence elasticity of ssDNA [111], we assumed an homogeneous elastic response when stretching the ssDNA since we cannot directly identify the sequence we pull.

In Fig. 6.4 we show the obtained folding free energy histograms for the measured molecules (solid lines) and we compare them with the theoretical prediction obtained using Mfold (dashed lines). Mfold data have been obtained by numerically folding all the existing 256 molecules for the 2M case, whereas for the 4M and 6M ensembles we have numerically folded, respectively, 5000 and 50000 different molecules.

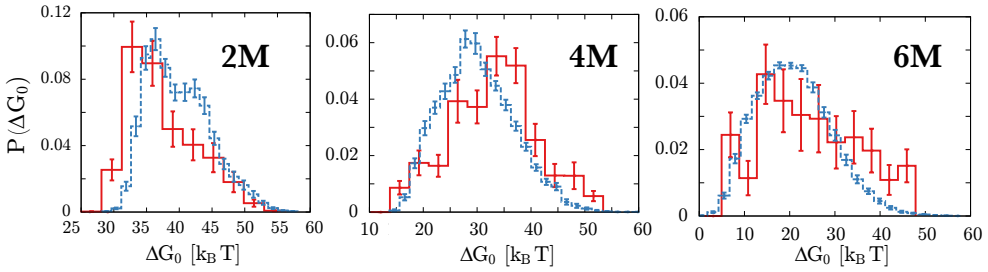

Figure 6.4: **Experimental and theoretical folding free energy spectra.** Experimental folding free energy histograms (red data) and theoretical folding free energy histograms (blue data) obtained using Mfold. Error bars have been obtained by Bootstrap resampling.

Notoriously, our measured folding free energies seem to be in good agreement with the nearest-neighbor model prediction using Mfold [73]. Nevertheless, eyeball comparison between histograms may lead to false conclusions[2]. Hence, we performed a statistical test in order to unveil the compatibility (or not) between folding free energy distributions (the experimental and the theoretical prediction). In particular, we carried out the non-parametric two-sample Kolmogorov–Smirnov test (hereafter referred to as K-S test) for checking whether two underlying one-dimensional probability distributions differ [112]. Briefly, the K-S statistic is built by evaluating the maximum absolute difference between the Empirical Cumulative Distribution Function (ECDF) of the two studied datasets, $D^\star$. Then, the rejection of the null hypothesis at level α is done if the following inequality holds:

$$D^\star > c(\alpha)\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}\,. \tag{6.1}$$

Being $c(\alpha) = \sqrt{\frac{-\log\alpha}{2}}$ and N_1, N_2 the number of points of each dataset (details and mathematical considerations are shown in section E.1). α accounts for the probability of incorrectly rejecting the null hypothesis.

Our (null) hypothesis is:

Null hypothesis (H_0). *Assuming that, for each ensemble, Mfold folding free distribution contains all existing molecules in the ensemble, the free energies that we experimentally measure are drawn from the Mfold distribution.*

In order to test this hypothesis, the ECDF of both datasets (the experimental measured values of ΔG_0 and the obtained using Mfold) must be obtained first. The ECDF of ΔG_0, $\hat{F}_N(\Delta G_0)$, is defined as:

$$\begin{aligned}\hat{F}_N(\Delta G_0) &= \frac{\text{Number of molecules with free energy} \leq \Delta G_0}{N} \\ &= \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{\Delta G_0^{(j)} \leq \Delta G_0}\,.\end{aligned} \tag{6.2}$$

2 Histograms depend on the number of bins and their size. Often these parameters are set by a rule of thumb, leading to biased and unreal results.

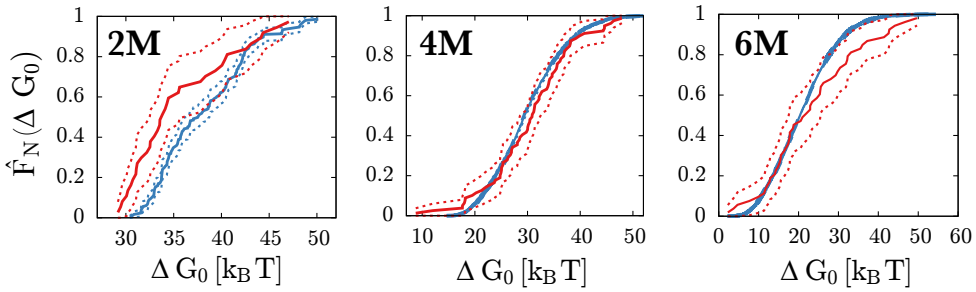

Figure 6.5: **Empirical cumulative distribution function.** ECDF obtained for experimental data (red, sharp line) and for theoretical prediction (blue, smoother line). Dashed lines correspond to the 95% lower and upper confidence bounds for the ECDF.

Where, for a given ensemble, N is the number of measured molecules and $\mathbb{1}_{\Delta G_0^{(j)} \leq \Delta G_0}$ is the indicator function of the event $\Delta G_0^{(j)} \leq \Delta G_0$. In Fig. 6.5 we show the calculated ECDF for both, the experimental values of ΔG_0 and the theoretical ones (red and blue lines, respectively). Dashed lines correspond to the 95% confidence bounds, obtained by applying the Greenwod's Formula [113]. In Table 6.2 we report the K-S statistic and the corresponding p-value, as well as the conclusion for the hypothesis.

	N_1	N_2	$D^\star$	$c(\alpha)\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$	p-value	Compatible?
2M	40	256	0.2867	0.2080	0.0019	✗
4M	94	5000	0.0891	0.1274	0.4461	✓
6M	53	50000	0.1527	0.1682	0.1415	✓

Table 6.2: **Kolmogorov-Smirnov test for folding free energy distributions.** Results of the K-S test and conclusion about the hypothesis (N_1 corresponds for the experimental data, whereas N_2 corresponds to the Mfold prediction). Hypothesis test has been performed with a significance level α equal to 5%.

Interestingly, we note that only the 2M ensemble is not compatible with the folding free energy distribution predicted by Mfold. Both the K-S test and the corresponding p-value[3], indicate that the measured folding free energies are not drawn from the Mfold distribution (which

3 The p-value indicates the probability of, assuming that both distributions are drawn from the same distribution, what is the probability of the two ECDF are as far apart

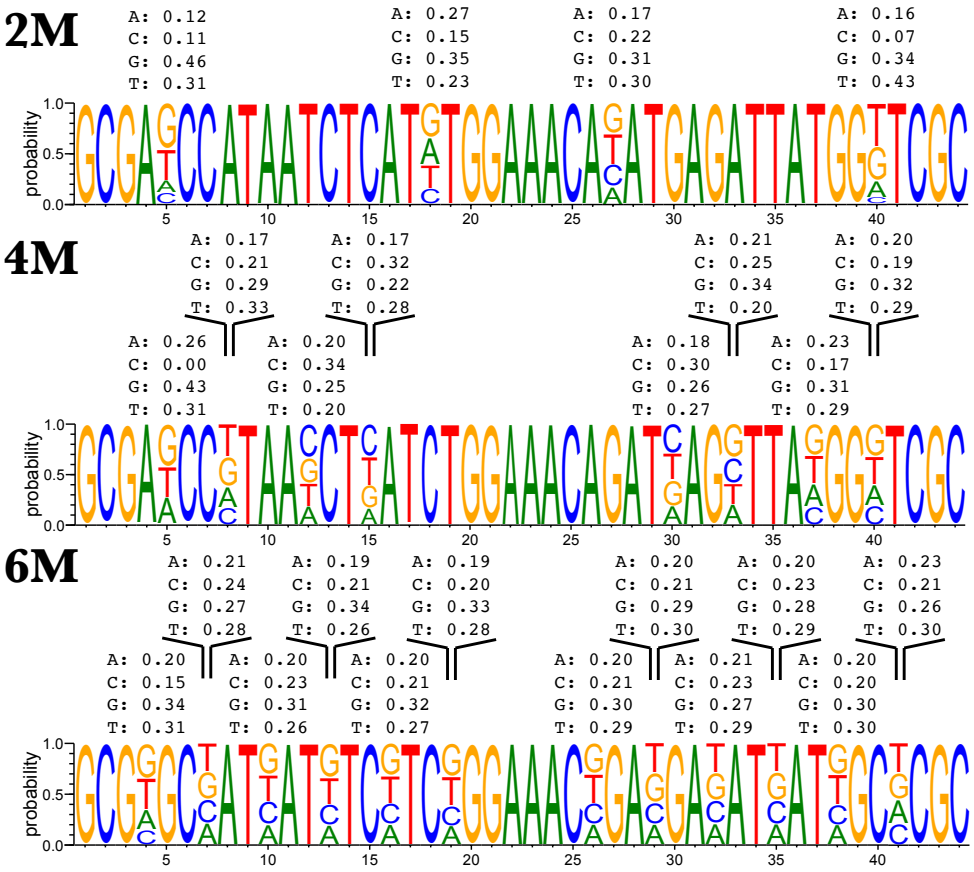

Figure 6.6: **Sequence logo of all molecular ensembles.** Graphical representation of the sequence conservation of the molecular ensembles (from top to bottom: 2M. 4M and 6M). The height of each base indicates its frequency of appearance in the sequence. For randomized positions, its frequencies of appearance are shown in the graph. Data obtained by D. Incarnato.

contains all the existing sequences compatible with 2 point mutations). For the rest of the molecular ensembles, the K-S test and the corresponding p-values indicate that the measured folding free energies are drawn from their respective Mfold distributions.

Spurred by the discrepancies that we found in the 2M sample, we sequenced all the molecular ensembles. Sequencing results are shown in Fig. 6.6, where we show the sequence logo of each molecular ensemble. We highlight several important facts. First and foremost, the nucleotide frequencies are closer to the desired ones (i.e. 0.25) in ensembles containing molecules with more mutational sites (see Table 6.3). In contrast,

as observed. Hence, low p-values indicate that two distributions are not likely to be compatible.

	νA [ad.]	νC [ad.]	νG [ad.]	νT [ad.]
2M	0.18 ± 0.03	0.14 ± 0.03	0.37 ± 0.03	0.31 ± 0.04
4M	0.20 ± 0.01	0.22 ± 0.04	0.30 ± 0.02	0.27 ± 0.02
6M	0.200 ± 0.003	0.21 ± 0.01	0.30 ± 0.01	0.285 ± 0.004

Table 6.3: **Summary of sequencing results.** Average of nucleotide frequencies (ν N, with N = A, C, G, T) for all the randomized positions of each ensemble. Uncertainty in each case corresponds with the standard error of the mean. Ideally, all frequencies should be equal to 0.25.

for ensembles containing less mutations, the nucleotides frequencies do not satisfy the ordered weights. This fact becomes crucial in the 2M sample, where the measured nucleotide frequencies are far away from the target frequencies. This fact might explain the discrepancies we found between the measured values of ΔG_0 and the ones predicted by Mfold. This fact will be further discussed below. Interestingly, the G, T nucleotide frequencies are significantly higher as compared to the A, C nucleotides. Finally, we also note that in the 4M sample, there is a missing nucleotide in the 5th position. Hence, this results might be a fingerprint of the difficulty when synthesizing small DNA molecules and the importance of having a precise knowledge of the sequence in order to correctly relate sequence and physical properties of molecules in SME.

We modified the *theoretical* prediction by including the actual nucleotide frequencies provided by the sequencing (Fig. 6.6). Afterwards, we computed again the folding free energy distributions for the same number of molecules than before (i.e. 256, 5000 and 50000 for the 2M, 4M and 6M ensemble, respectively). Finally, we carried out again the K-S test in order to discern whether the experimental data is compatible with the corrected Mfold predictions.

In Fig. 6.7 we show the ECDF calculated for the experimentally measured ΔG_0 (red curves, same as in Fig. 6.5) and the theoretical ECDF (green curves) obtained for an equivalent amount of molecules as in Fig. 6.5 but setting the nucleotide frequencies obtained by sequencing each ensemble (Fig. 6.6). On the other hand, in Table 6.4 we summarize the results regarding the K-S test as well as the p-values for the hypothesis test.

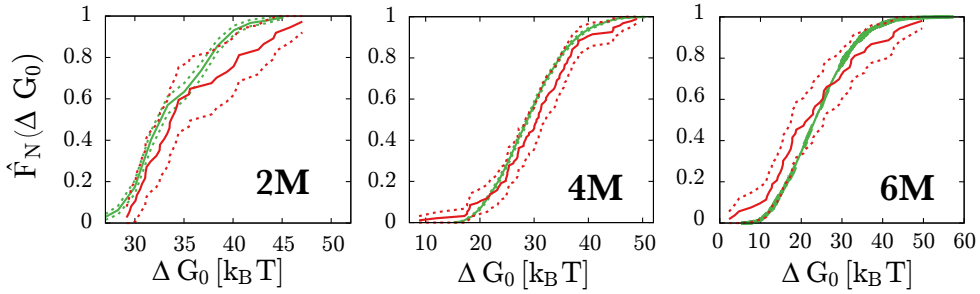

Figure 6.7: **New ECDF with corrected weights.** ECDF obtained for experimental data (red, sharp line) and for theoretical prediction (green, smoother line) with the corrected nucleotide frequencies. Dashed lines correspond to the 95% lower and upper confidence bounds for the ECDF.

	N_1	N_2	$D^\star$	$c(\alpha)\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$	p-value	Compatible?
2M	40	256	0.2003	0.2080	0.2189	✓
4M	94	5000	0.1193	0.1274	0.1434	✓
6M	53	50000	0.0917	0.1682	0.1081	✓

Table 6.4: **New Kolmogorov-Smirnov test for folding free energy distributions after sequencing.** Results of the K-S test and conclusion about the hypothesis (N_1 corresponds for the experimental data, whereas N_2 corresponds to the Mfold prediction with corrected frequencies). Hypothesis test has been performed with a significance level α equal to 5%.

Now, we find that all the experimental ECDF are fully compatible with the corrected theoretical predictions. Besides 2M ensemble, where the theoretical prediction was not fully compatible with the actual molecular ensemble, in 4M and 6M ensembles the differences between theory and experiments were not significant. We conclude that, the assumption that all samples come from a neutral mutational ensemble (where any nucleotide can be found in the randomized positions with equal probability) is fair enough for further considerations, yet always recalling that the comparison between theory and experiments in the 2M ensemble must be delicate.

6.3.2 *Comment on the sample size*

Overwhelmed by the astonishing number of different molecules existing in an NM molecular ensemble (where $\Omega_N = 4^{2N}$), one may ask what is the required sample size to correctly estimate some statistical quantities of such ensemble. Among all of the statistical properties of a distribution, the mean and the variance are two of the most important. Higher order moments, such as the statistical skewness or kurtosis, are usually hard to estimate since they are very sensitive to data outliers. Indeed, histogram method is usually the best method to unravel symmetry properties of the data. On the other hand, the estimation of the mean requires a knowledge of the expected sample standard deviation. This parameter, in general, is not known in our purposes. Therefore, the present discussion is restricted on the sample size required for the estimation of the population variance (or standard deviation). Indeed, variance is one of the key quantities in statistics. From statistical inference, hypothesis testing up to Monte Carlo techniques, having a good estimation of the variance is essential. Nevertheless, its accurate estimation is often a difficult task.

Let us suppose that we want to estimate the population variance of the folding free energy, denoted as $\sigma^2_{\Delta G_0}$. A typical procedure consists on the measurement of a subset N_{exp} ($< \Omega_N$) of folding free energies of molecules belonging to the population, $\Delta G_0^{(1)}, \cdots, \Delta G_0^{(N_{\text{exp}})}$, and its latter estimation of the population variance as:

$$s^2_{\Delta G_0} = \frac{1}{N_{\text{exp}} - 1} \sum_{k=1}^{N_{\text{exp}}} (\Delta G_0^{(k)} - \Delta G_0^{\star})^2 , \tag{6.3}$$

where $\Delta G_0^{\star} = \frac{1}{N_{\text{exp}}} \sum_{k=1}^{N_{\text{exp}}} \Delta G_0^{(k)}$ is the sample mean of the measured folding free energies. By repeating the number of experiments a large number of times ($N_{\text{exp}} \gg 1$), the sample variance yields: $\mathbb{E}[s^2_{\Delta G_0}] = \sigma^2_{\Delta G_0}$. Our goal is to foresee how many molecules we need to pull in order to have a precise estimation of the population folding free energy variance. In order to solve this question, we will focus on the relative difference between $s_{\Delta G_0}$ and $\sigma_{\Delta G_0}$ [114]. Let us assume that the measured $\Delta G_0^{(k)}$ ($1 \leq k \leq N_{\text{exp}}$) are i.i.d. Gaussian random

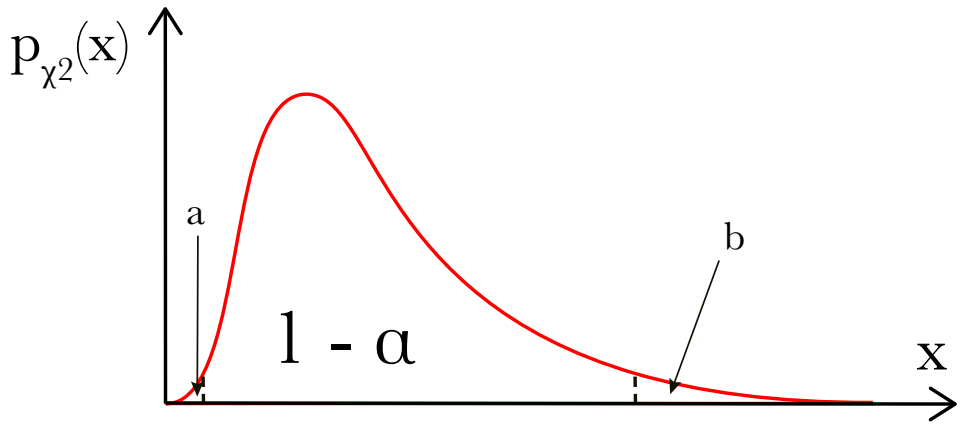

Figure 6.8: **Illustration of χ^2 distribution.** The a and b parameters shown in the graph are the areas that are pointed by the arrows and they correspond to the probabilities given by Eq. (6.9) and Eq. (6.10).

variables drawn from a $\mathcal{N}(\mu_{\Delta G_0}, \sigma^2_{\Delta G_0})$ distribution. We define the relative deviation between $s_{\Delta G_0}$ are $\sigma_{\Delta G_0}$ as:

$$u := \frac{|s_{\Delta G_0} - \sigma_{\Delta G_0}|}{\sigma_{\Delta G_0}} . \tag{6.4}$$

Note that, for $s_{\Delta G_0} < \sigma_{\Delta G_0}$:

$$u = -\frac{s_{\Delta G_0} - \sigma_{\Delta G_0}}{\sigma_{\Delta G_0}} , \tag{6.5}$$

whereas for $s_{\Delta G_0} > \sigma_{\Delta G_0}$:

$$u = \frac{s_{\Delta G_0} - \sigma_{\Delta G_0}}{\sigma_{\Delta G_0}} , \tag{6.6}$$

so $0 < u < 1$. Then, the probability that the relative deviation between $s_{\Delta G_0}$ and $\sigma_{\Delta G_0}$ lies within a fraction $0 < u < 1$ can be written as:

$$P\{s_{\Delta G_0} < (1-u)\sigma_{\Delta G_0}\} := a , \tag{6.7}$$

and

$$P\{s_{\Delta G_0} > (1+u)\sigma_{\Delta G_0}\} := b . \tag{6.8}$$

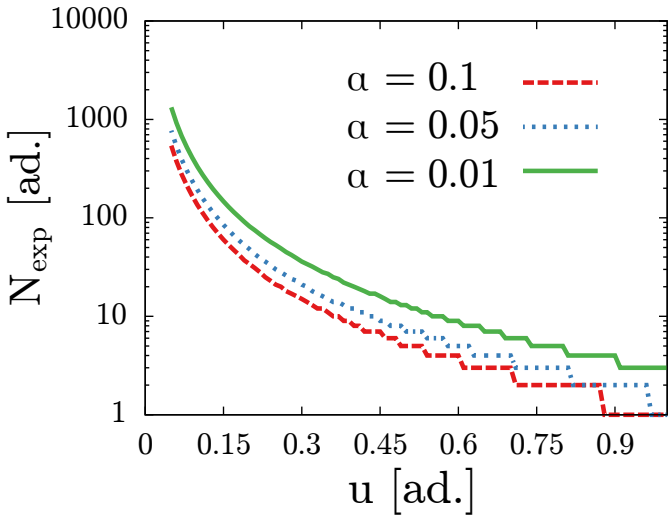

Figure 6.9: **Sample size as a function of the relative deviation between the sample and the population standard deviation, u.** N_{exp} obtained by numerically solving Eq. (6.11) for different values of the significance level, α.

Equivalently, previous probabilities can be written as:

$$P\left\{\frac{(N_{exp}-1)s^2_{\Delta G_0}}{\sigma^2_{\Delta G_0}} < (N_{exp}-1)(1-u)^2\right\} = a\,, \tag{6.9}$$

$$P\left\{\frac{(N_{exp}-1)s^2_{\Delta G_0}}{\sigma^2_{\Delta G_0}} > (N_{exp}-1)(1+u)^2\right\} = b\,. \tag{6.10}$$

Equations (6.9), (6.10) define the confidence level c of the variance estimation and it is related to the significance level α as $1-\alpha = c$. In Fig. 6.8 we show an illustration of a and b from Eq. (6.9) and Eq. (6.10). The area pointed by both arrows correspond to a and b.

Note that the quantity $\frac{(N_{exp}-1)s^2_{\Delta G_0}}{\sigma^2_{\Delta G_0}}$ follows a chi-squared distribution with $N_{exp}-1$ degrees of freedom (for a formal proof see section E.2). Then, the addition of Eqs. (6.9), (6.10) yields:

$$1-\alpha = F_{\chi^2_{(N_{exp}-1)}}((N_{exp}-1)(1+u)^2) - F_{\chi^2_{(N_{exp}-1)}}((N_{exp}-1)(1-u)^2)\,, \tag{6.11}$$

where α is the significance level[4] and $F_{\chi^2_{(N_{\text{exp}}-1)}}$ is the cumulative distribution function of the chi-squared distribution. Then, Eq. (6.11) can be numerically solved in order to estimate the number of molecules (i.e. sample size), N_{exp}, that are needed to estimate $s_{\Delta G_0}$ with a relative deviation from $\sigma_{\Delta G_0}$ equal to u.

Figure 6.9 highlights the difficulty involving a precise determination of a population variance (i.e. the squared value of the standard deviation). For instance, for a 5% significance level, in order to decrease u from 0.3 down to 0.15, implies an increase of N_{exp} by a factor of 4. This effect becomes stronger when lowering α and for decreasing u.

6.4 AVERAGE ENSEMBLE DISSIPATION AND KINETIC PROPERTIES

As we discussed in Sec. 5.3, molecular kinetics governs the behavior of the average dissipated work. Moreover, kinetics are strongly sequence-dependent. Indeed, recent studies related the unfolding kinetic rate at zero force (k_m of Eqs. (5.5) and (5.6)) with the sequence, showing a clear increase of the spontaneous unfolding kinetic rate (i.e. at zero force) as the AT-content of the sequence also increases [32]. Despite of that, there are no studies relating the average dissipation with molecular sequences.

In contrast to in section 5.3, the experiments presented throughout this chapter were done at a constant and unique pulling speed. Hence, now we are not able to extract the molecular kinetic properties by studying the average dissipation in different pulling regimes, as done before. Nevertheless, in what follows, we prove that characteristic kinetic properties can be also obtained by means of the ensemble average dissipation. In particular, we have developed an analytical model that allows us to quantify ensemble kinetic properties from the knowledge of the folding free energy distribution for each ensemble.

The average dissipated work, defined as: $\langle W_{\text{dis}} \rangle = \langle W \rangle - \Delta G$, has been obtained for every molecule using Eq. (5.7). We note that in the previous definition of the dissipated work, if stretching contributions are subtracted to $\langle W \rangle$, instead of ΔG, ΔG_0 must be used. In Fig. 6.10 we show, for each molecular ensemble, the probability distribution of

4 Intuitively, the significance level accounts for the probability that the pattern of the data is due to chance.

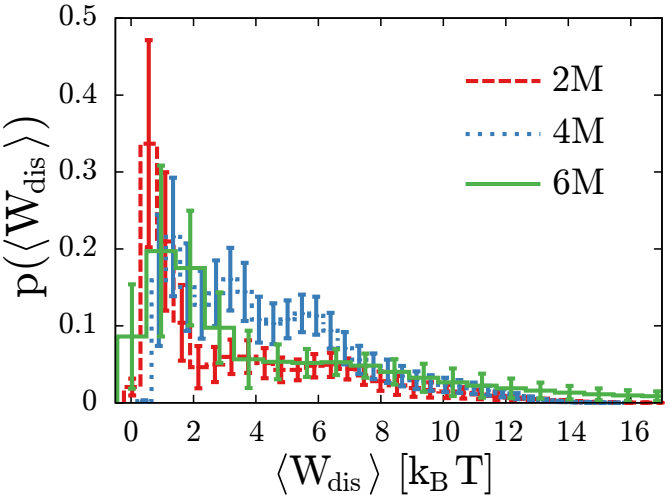

Figure 6.10: **Probability distribution of the average dissipated work.** Probability density function of the average dissipated work for each ensemble (indicated in the graph). Error bars were obtained by Bootstrap resampling and the number of molecules are the reported in the preceding section.

the average dissipated work obtained by measuring, for each molecule, $\langle W_{\text{dis}} \rangle$. Interestingly, average dissipation spans from few $k_{\text{B}}T$s up to 20 $k_{\text{B}}T$s. Since experimental conditions were identical for all the molecules, the differences found in dissipation are clearly due to the sequence variability from molecule to molecule.

Having in mind that we cannot know the molecule we are pulling, we aim to develop a model for $\langle W_{\text{dis}} \rangle$ that allows us, through the knowledge of the particular folding free energy distribution of each ensemble ($p(\Delta G)$, see Fig. 6.4), to characterize the shape of the $p(\langle W_{\text{dis}} \rangle)$. Since probability is conserved, $p(\Delta G)$ can be transformed to $p(\langle W_{\text{dis}} \rangle)$ as:

$$|p(\langle W_{\text{dis}} \rangle) d\langle W_{\text{dis}} \rangle| = |p(\Delta G) d\Delta G| \,. \tag{6.12}$$

Then, in order to characterize $p(\langle W_{\text{dis}} \rangle)$ we need to relate $\langle W_{\text{dis}} \rangle$ with ΔG. In the linear response limit, the average total dissipated work of a two-state system is given by [92]:

$$\langle W_{\text{dis}} \rangle = \frac{\pi x_m}{4 k_{\text{B}} T k_c} \frac{1-\mu^2}{\cos \frac{\pi \mu}{2}} r + \mathcal{O}(r^2) \,. \tag{6.13}$$

Where μ is the molecular fragility defined in Eq. (5.9), r is the loading rate, k_c the unfolding-folding kinetic rate at the coexistence force (i.e.

the force at which folded and unfolded state are equally populated), x_m is the molecular extension at the coexistence force (same definitions as in Section 5.3.1). Equation (6.13) is obtained by integrating Eq. (5.8) using the saddle-point method. We point out that Eq. (6.13) accounts for the total average dissipation, the sum of the average dissipated work in the F process and the R process[5].

Equation (6.13) depends on the free energy via the critical kinetic constant k_c. Since Eq. (5.6) at f_c can be written as:

$$k_{U\to F}(f_c) := k_c = k_m \exp \frac{\Delta G - f_c x_U}{k_B T}, \tag{6.14}$$

where at first order approximation ΔG equals: $f_c x_m$. Then, inserting the previous relation and Eq. (6.14) in Eq. (6.13), we obtain:

$$\langle W_{\text{dis}} \rangle = \frac{\pi}{4\, k_B T} \frac{1-\mu^2}{\cos \frac{\pi\mu}{2}} \frac{x_m r}{k_m} \exp \left[\frac{-\Delta G}{k_B T} \left(\frac{1+\mu}{2} \right) \right]. \tag{6.15}$$

To obtain the previous result we have also used the fact that the fragility can be written as: $\mu = 1 - 2\frac{x_U}{x_m}$.

Equation (6.15) allows us to analytically differentiate $\langle W_{\text{dis}} \rangle$ with respect to ΔG. By considering $p(\Delta G) = \mathcal{N}(\Delta G^\star, \sigma^2_{\Delta G})$, Eq. (6.15) also allows us to explicitly write the dependence of ΔG on $\langle W_{\text{dis}} \rangle$. Finally, after straightforward algebraic steps, $p(\langle W_{\text{dis}} \rangle)$ yields:

$$p(\langle W_{\text{dis}} \rangle) = \frac{1}{\sqrt{2\pi}} \frac{\Omega}{\langle W_{\text{dis}} \rangle} \exp \left(\frac{-\Omega^2}{2} \left(\log \langle W_{\text{dis}} \rangle - m_{\langle W_{\text{dis}} \rangle} \right)^2 \right). \tag{6.16}$$

Where the parameter Ω is defined as:

$$\Omega^2 = \frac{(k_B T)^2}{\sigma^2_{\Delta G}} \frac{(1+\mu)^2}{4}, \tag{6.17}$$

and $m_{\langle W_{\text{dis}} \rangle}$ equals to:

$$m_{\langle W_{\text{dis}} \rangle} = \log \left(\frac{\pi}{4\, k_B T} \frac{1-\mu^2}{\cos \frac{\pi\mu}{2}} \frac{x_m r}{k_m} \right) - \frac{1+\mu}{2} \frac{\Delta G^\star}{k_B T}. \tag{6.18}$$

5 In the R process $\mu \to -\mu$

We note that Eq. (6.16) allow us to estimate important kinetic properties of each molecular ensemble. This is the case of, for instance, the molecular fragility μ and the unfolding kinetic rate at zero force, k_m. On the one hand μ gives us information about the position of the kinetic barrier ($\mu \in [-1, 1]$) and, on the other hand, k_m provides us insights about the height of the kinetic barrier.

The estimation of μ using Eq. (6.16) is complex, in general. The fragility appears in the Ω parameter (Eq. (6.17)), which essentially governs the kurtosis of the distribution (via an exponential dependence) [115]. Kurtosis is the quantity that measures the "tailedness" of the distribution. Therefore, since tails are extremely hard to characterize, obtaining a good and reliable estimation for μ turns out to be an arduous task. Hence, we decided to fix this parameter to $\mu = 0$ due to several reasons. First and foremost, the value $\mu = 0$ corresponds to the measured (and predicted) value for CD_4, our template molecule [32]. Secondly, since we are sequence-blindfolded and $\mu \in [-1, 1]$, we approximate the probability distribution of the molecular fragility, $p(\mu)$, as a continuous uniform probability distribution as:

$$p(\mu) = \begin{cases} 0 & \text{if } \mu < -1\,, \\ \frac{1}{2} & \text{if } -1 \leq \mu \leq 1\,, \\ 0 & \text{if } \mu > 1\,. \end{cases} \tag{6.19}$$

In this approximation, the average value of μ equals 0. It is worth mentioning that assuming a molecule-independent fragility implies that the position of the kinetic barrier is the same for all the molecules. Furthermore, setting $\mu = 0$ means that the kinetic barrier is equidistant between the folded and the unfolded state. A molecular fragility $\mu = -1$ implies that the barrier is located in the folded state, whereas $\mu = 1$ implies that the barrier is located in the unfolded state.

The knowledge of k_m allows us to, ultimately, estimate the height of the kinetic barrier. k_m estimation has been done by maximizing the logarithm of the Maximum Likelihood function of Eq. (6.16) (see

section 4.4.1). For a series of $W_{\mathrm{dis}}^{(k)}$ ($k = 1, \cdots, N$), the logarithm of the Maximum Likelihood is defined as:

$$\log \mathcal{L}\left(k_m \mid \{\langle W_{\mathrm{dis}}\rangle\}\right) = \log \prod_{k=1}^{N} p(\langle W_{\mathrm{dis}}^{(k)}\rangle \mid k_m) = \sum_{k=1}^{N} \log p(\langle W_{\mathrm{dis}}^{(k)}\rangle \mid k_m)\,, \tag{6.20}$$

where the multiplication and the sum runs for all the measured molecules, N. From Eq. (6.16) we insert $p(\langle W_{\mathrm{dis}}\rangle)$ into Eq. (6.20), yielding:

$$\log \mathcal{L}\left(k_m \mid \{\langle W_{\mathrm{dis}}\rangle\}\right) = -\frac{N}{2}\log 2\pi + N\log\Omega - N\sum_{k=1}^{N}\log\langle W_{\mathrm{dis}}^{(k)}\rangle - \frac{\Omega^2}{2}\sum_{k=1}^{N}\left(\log\langle W_{\mathrm{dis}}^{(k)}\rangle - m_{\langle W_{\mathrm{dis}}\rangle}\right)^2. \tag{6.21}$$

Equation (6.21) has been numerically maximized for k_m. In Table 6.5 we report the obtained values for each molecular ensemble. The uncertainty associated to each value of k_m is the standard error of the Maximum Likelihood. That is, the square root of the numerical Hessian. We emphasize that the maxima we find are reproducible: multiple runs from different starting points yield the same maxima.

Interestingly, the kinetic constants we obtain (Table 6.5) are several orders of magnitude above the measured value of CD_4 [32]. This is due to the fact that it is more likely to measure molecules with non-canonical basepairs (i.e. non complementary) rather than fully canonical molecules. Hence, DNA hairpins with less complementary basepairs tend to have smaller kinetic barriers. This fact can be illustrated considering that the unfolding kinetic constant at zero force, k_m has an exponential dependence on the kinetic barrier as: $k_m = k_0 \exp\left(-B/\,k_{\mathrm{B}}T\right)$, being k_0 the so-called attempt frequency at zero force. Then, for $k_0 \sim 10^4\ \mathrm{s}^{-1}$, a typical value for hairpins [32], the kinetic barrier B can decrease from almost 20 $k_{\mathrm{B}}T$ for CD_4, to few $k_{\mathrm{B}}T$ for molecular ensembles with higher number of mutations. Of course, we are still able to measure molecules spontaneously forming hairpins (i.e. with higher kinetic barriers as compared to Brownian fluctuations), hence the figures employed in the previous discussion should be taken only

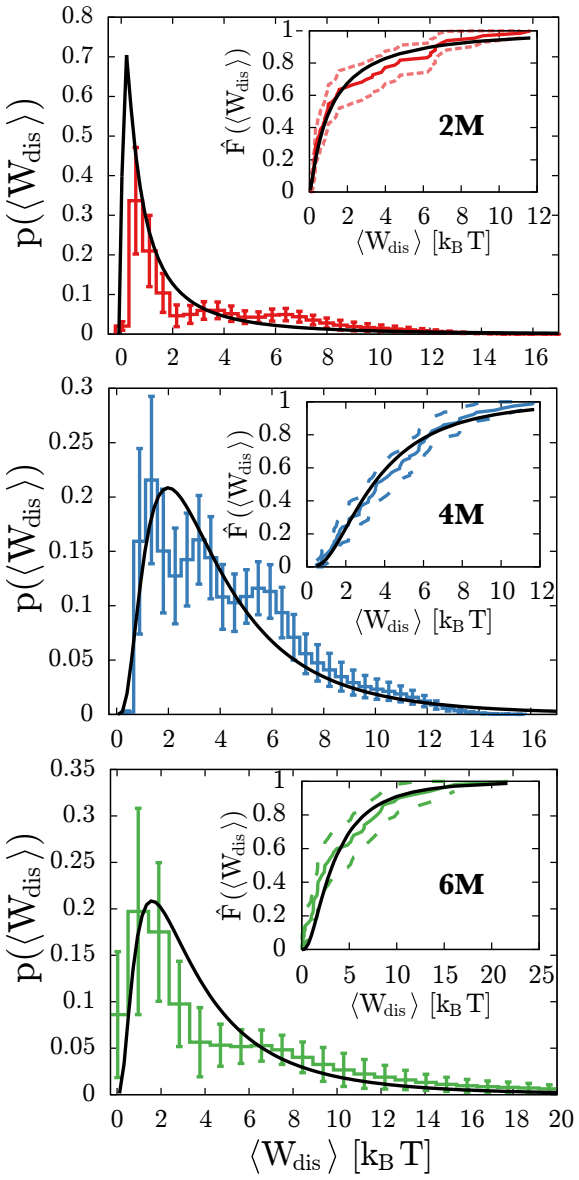

Figure 6.11: **Theoretical prediction of average dissipated work.** Comparison between the experimental probability distribution function of the average dissipated work (color lines) and the fit to Eq. (6.16) (black solid lines). Insets are a comparison between the ECDF of the experimental data and the cumulative distribution function calculated according to Eq. (6.16). Error bars are obtained by resampling, whereas dashed lines in the ECDF correspond to the 95% lower and upper confidence bounds.

as a guide. Indeed, we think that the kinetic constants we estimate using Eq. (6.21) are an upper bound for the k_m that we may find for each individual molecule. This is due to the logarithmic dependence of

	CD_4	2M	4M	6M
k_m [s^{-1}]	$(5 \pm 5) \times 10^{-12}$	$(5 \pm 2) \times 10^{-7}$	$(3 \pm 1) \times 10^{-6}$	$(7 \pm 3) \times 10^{-5}$

Table 6.5: **Estimated unfolding kinetic rates at zero force.** Results of the Maximum Likelihood estimation of k_m via the ensemble average dissipated work (Fig. 6.11). The value corresponding to CD_4 was measured in Ref. [32].

$m_{\langle W_{\text{dis}} \rangle}$ (Eq. (6.18)), where higher k_m dominate when maximizing the Maximum Likelihood function (Eq. (6.21)).

Figure 6.11 shows the comparison between the experimental probability distributions of the average dissipated work and the theoretical ones obtained using Eq. (6.16) and the estimated values of k_m (Eq. (6.21)). As insets, we show the comparison between the ECDF for the experimental data (color lines) and the theoretical cumulative distribution function calculated using Eq. (6.16) with the estimated k_m (Table 6.5). Remarkably, our model can reproduce the shape of the experimental distribution, as well as the location of the mode of $\langle W_{\text{dis}} \rangle$.

6.5 CONCLUSIONS

This chapter aimed for setting the grounds for the development of a systematic procedure for measuring the information-content of a molecular ensemble. Such method will be described in the following chapter, however a previous and exhaustive characterization of the molecular ensembles was required.

We have presented a novel experimental system for usual SME: heterogeneous molecular pools. Typically, in single-molecule assays, the studied molecules are always known. This is not our case, where in each experimental realization the molecule is unknown (and, very likely, different). This fact, rather than being a hindrance, allows us to harvest a remarkable amount of information of the molecular ensembles. We have demonstrated that with some tens of molecules, we are able to determine the folding free energy spectra of the ensembles. Furthermore, we detected discrepancies between the theoretical and the actual construction of the molecular ensembles. Such discrepancies, until now, were only noticeable when sequencing the samples and might have been masked when only considering individual molecular

properties, rather than ensemble molecular properties (e.g. ΔG_0 vs. $p(\Delta G_0)$). Moreover, by studying ensemble thermodynamic quantities (e.g. dissipation), we have been able to measure characteristic kinetic properties of the molecular ensembles via a solvable model.

Finally, it is worth mentioning that, among all the results we presented in the present chapter, the verification of the CFT for arbitrary randomized molecules must not be underestimated. We confirmed that the CFT is fulfilled for a large set of (very) different molecules, being this, probably, the SME study involving the largest set of molecules testing the CFT.

7

INFORMATION-CONTENT MEASUREMENT OF MOLECULAR ENSEMBLES

7.1 INTRODUCTION

The connection between statistical mechanics and information theory sprang forth almost 75 years ago. Claude E. Shannon set the mathematical foundations of information theory in a landmark paper published in 1948 in the context of communication theory [5]. Shannon defined the so-called information by a very familiar formula to the one defining the entropy in statistical physics: the Gibbs formula. While the Gibbs formula reads as:

$$S = -k_{\mathrm{B}} \sum_i p_i \log p_i \,, \tag{7.1}$$

the Shannon entropy H equals to:

$$H = -\sum_i p_i \log_b p_i \,. \tag{7.2}$$

Although previous equations have clear similarities, they are contextually different. For instance, while in the Gibbs formula (Eq. (7.1)), p_i denotes the probability of the microstate i, in Shannon's expression (Eq. (7.2)) p_i accounts for the probability of receiving the i message. Furthermore, one may note that the base of the logarithm in Eq. (7.2) is b, whereas in the Gibbs entropy (Eq. (7.1)) the logarithm is a natural logarithm. Hence, the logarithm prefactor between both expressions is also different. Shannon's entropy, in its most basic terms (setting $b = 2$) accounts for the number of binary digits required to encode a message. Nevertheless, it is possible to make a direct connection between Gibbs and Shannon formulae. In the view of Jaynes, Eqs. (7.1) and (7.2) are two sides of the same token [6, 7]. He argued that statistical mechanics can be regarded as an application of a more general theory containing logical inference and information theory.

Shannon entropy has recently acquired a thermodynamic meaning in small systems under feedback control [116–118]. Recent experimental realizations of information-to-energy converting devices have demonstrated the close connection between information and energy [48], showing how the possession (or lack) of information might have thermodynamic consequences. Paradigmatic examples of the connection between information and energy are the Szilard's engine [119] and the Landauer's principle [120, 121].

Szilard aimed to resolve the famous Maxwell's paradox: the Maxwell demon. The Maxwell's demon refers to a *Gedankenexperiment* in which an intelligent being (the demon) is able to monitor the individual molecules of a gas contained in two neighboring chambers. The demon is able to gather information about the state of the particles in order to sort them according to their velocity. Ultimately, only by means of the demon, the system ends up in a situation in which fast molecules (higher temperature) are in one chamber whereas slow molecules (low temperature) are in the other side of the vessel. Hence, the entropy of the system is decreased with no external action. Szilard showed that a one-particle device is able to perform useful work only by receiving information, rather than energy (like the Maxwell demon). Nowadays, there are many experimental realizations of the Szilard's engine both in its classical and quantum version [56, 122], strengthening the fact that information is a physically measurable quantity. On the other hand, according to the Laudauer's principle, every process involving the erasure of information dissipates some heat to the environment. The amount of entropy generated upon erasing one bit of information is set by the Landauer's limit: $k_{\mathrm{B}} \log 2$. Hence, the dissipated energy energy is $E \geq k_{\mathrm{B}} T \log 2 \approx 0.69\, k_{\mathrm{B}} T$.

Summing up, information-to-energy conversion is a well-established topic. However, what about the reverse? Is it possible to convert energy into information? The possibility of obtaining information-contents in physical systems through thermodynamic energetic measurements opens a wide range of exciting possibilities. For instance, in molecular systems, measuring the information-content of protein families or studying the information-content production of molecular ensembles in directed evolutionary processes, just to mention a few. In this chapter we show how the information-content of heterogeneous popu-

lations can be robustly defined by only means of free energy differences measurements.

7.2 INFORMATION-CONTENT OF MOLECULAR ENSEMBLES

Let us consider an heterogeneous population (or ensemble ε) of individuals. Each individual can be characterized (or identified) by its phenotype α in a population of M phenotypes ($1 \leq \alpha \leq M$). The fraction of individuals with a given phenotype, p_α define the phenotypic frequencies in ε. Clearly: $\sum_\alpha p_\alpha = 1$.

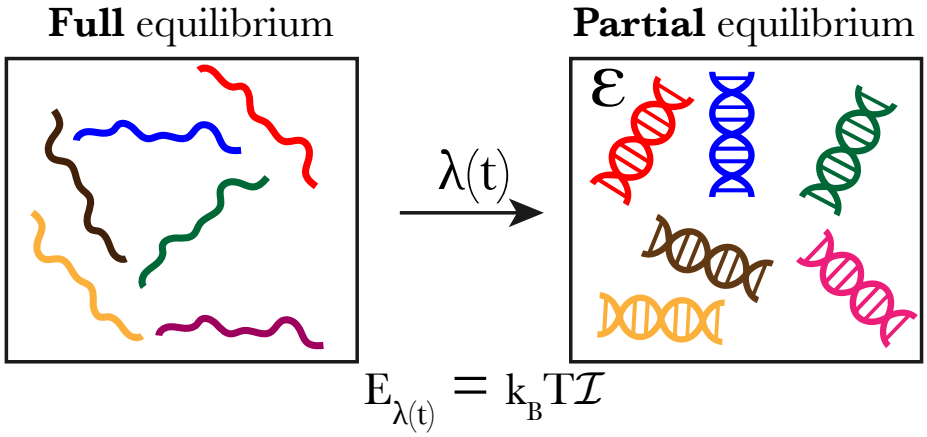

Figure 7.1: **Information-content of heterogeneous ensembles.** The information-content of ε is the minimum amount of energy, $k_B T \mathcal{I}$, required to build the heterogeneous partially equilibrated population (right) starting from a population of individuals in full thermodynamic equilibrium (left) via a $\lambda(t)$ protocol.

We define the information-content of ε at temperature T as the minimum free energy cost required to build the population of partially equilibrated individuals (defined by the set $\{p_\alpha\}$) starting from a population of individuals in full thermodynamic equilibrium (see Fig. 7.1).

An ensemble ε is in equilibrium *if and only if*:

$$p_\alpha = \frac{e^{-G_\alpha / k_B T}}{\mathcal{Z}} \quad , \quad \mathcal{Z} = \sum_\alpha e^{-G_\alpha / k_B T} = e^{-\mathcal{G} / k_B T} \, , \tag{7.3}$$

being G_α the partial free energy of individuals with phenotype α, $\mathcal{Z}$ the partition function of the system and $\mathcal{G}$ the ensemble free energy.

Let us now consider an isothermal thermodynamic transformation $0 \to 1$ applied to all individuals of the ensemble where one or more

control parameters λ are varied between an initial (λ_0) and final (λ_1) values following an arbitrary $\lambda_{\rightarrow}(t)$ protocol in a time Δt. We define the Ensemble Work Distribution (EWD) as:

$$\mathcal{P}_{\rightarrow}(W) = \sum_{\alpha} p_{\alpha} P_{\rightarrow}^{(\alpha)}(W)\,, \tag{7.4}$$

where $P_{\rightarrow}^{(\alpha)}(W)$ is the work distribution corresponding to individual α calculated in the $0 \rightarrow 1$ transformation. The EWD fulfils a fluctuation theorem (see section E.3 of appendix E for a mathematical demonstration):

$$\frac{\mathcal{P}_{\rightarrow}(W)}{\mathcal{P}_{\leftarrow}(-W)} = \exp\left(\frac{W - \Delta\mathcal{G} + k_{\mathrm{B}}T\mathcal{I}}{k_{\mathrm{B}}T}\right)\,, \tag{7.5}$$

where $\mathcal{P}_{\leftarrow}(-W)$ stands for the EWD in the reverse process ($0 \leftarrow 1$), defined as that process applied on the same phenotypic ensemble ($\{p_{\alpha}\}$) where the control parameter is varied following the time-reversed path of the forward one ($\lambda_{\leftarrow}(t) = \lambda_{\rightarrow}(\Delta t - t)$). On the other hand, $\Delta\mathcal{G}$ is the ensemble free energy difference, defined as: $\Delta\mathcal{G} = \mathcal{G}(\lambda_1) - \mathcal{G}(\lambda_0) = -k_{\mathrm{B}}T\log\left(\mathcal{Z}_{\lambda_1}/\mathcal{Z}_{\lambda_0}\right)$.

We note that the reversed EWD can be expressed as:

$$\mathcal{P}_{\leftarrow}(-W) = \sum_{\alpha} \hat{p}_{\alpha} P_{\leftarrow}^{(\alpha)}(-W)\,, \tag{7.6}$$

being $\hat{p}_{\alpha}$ positive normalized frequencies ($\sum_{\alpha}\hat{p}_{\alpha} = 1$) defined as:

$$\hat{p}_{\alpha} = p_{\alpha}\frac{e^{-\Delta G_{\alpha}/\,k_{\mathrm{B}}T}}{\sum_{\alpha} p_{\alpha} e^{-\Delta G_{\alpha}/\,k_{\mathrm{B}}T}}\,. \tag{7.7}$$

Here $\Delta G_{\alpha} = G_{\alpha}(\lambda_1) - G_{\alpha}(\lambda_0)$. Equation (7.5) defines a fluctuation theorem for a phenotypic ensemble and the information-content $\mathcal{I}$ of the ensemble ε equals to:

$$k_{\mathrm{B}}T\mathcal{I} = \Delta\mathcal{G} + k_{\mathrm{B}}T\log\left(\sum_{\alpha} p_{\alpha}\exp\left(-\frac{\Delta G_{\alpha}}{k_{\mathrm{B}}T}\right)\right)\,. \tag{7.8}$$

For an equilibrium phenotypic ensemble the probabilities are given by: $p_\alpha = e^{-G_\alpha(\lambda_0)/k_\mathrm{B}T}/Z_{\lambda_0}$, yielding $\mathcal{I} = 0$, as expected. However, we note that Eq. (7.8) is not uniquely defined as it depends on the final state. This issue can be solved by considering a specific final state of the $0 \to 1$ transformation where all the phenotypes have the same free energy, $G_\alpha(\lambda_1) := G(\lambda_1)$. In this latter case, Eq. (7.8) reduces to:

$$\begin{aligned}\mathcal{I} &= \log\left(\sum_\alpha p_\alpha \exp\left(-\frac{G_\alpha(\lambda_0)}{k_\mathrm{B}T}\right)\frac{1}{M}\sum_\alpha \exp\left(\frac{G_\alpha(\lambda_0)}{k_\mathrm{B}T}\right)\right)\\ &= \log\left(\sum_\alpha p_\alpha \exp\left(-\frac{G_\alpha(\lambda_0)}{k_\mathrm{B}T}\right)\right) + \log\left(\frac{1}{M}\sum_\alpha \exp\left(\frac{G_\alpha(\lambda_0)}{k_\mathrm{B}T}\right)\right).\end{aligned} \tag{7.9}$$

Equation (7.9) has two fundamental properties. First, it is uniquely defined for a given ensemble ε. Second, its average over all possible phenotypic ensembles $\{p_\alpha\}$ is always positive, $\langle\mathcal{I}\rangle \geq 0$. Moreover, Eq. (7.9) provides a simple way to unambiguously measure the information-content of populations by only using thermodynamic considerations, circumventing the use of the Shannon information or other information measures based on distributions of arbitrary quantities across the population. Our derivation of the information-content uses the extended fluctuation theorem [52, 123] to extract free energy differences between partially equilibrated Gibbs states. This requires the knowledge of the partial free energies $G_\alpha(\lambda_0)$ of the different phenotypes in the population, a task that can be accomplished using single-molecule methods on different phenotype individuals.

7.2.1 *Upper bound for information-content in molecular ensembles*

While Eq. (7.9) is fully general, we found an upper bound for the information-content of a molecular ensemble. The derivation of the upper bound of the information-content $\mathcal{I}$ only requires two ingredients: the application of the Jensen's inequality and the assumption that $G_\alpha(\lambda_0)$ are i.i.d. Gaussian random variables with an average equal to the arithmetic mean of $G_\alpha(\lambda_0)$ ($\langle G_\alpha(\lambda_0)\rangle := G^\star$). The Gaussian approximation may be justified by arguing that the Gaussian entropy has the maximum entropy relative to all probability distributions in $\mathbb{R}$ (and, moreover, finite moments).

We rewrite the first term of Eq. (7.9) in a more recognizable way:

$$\log\left(\sum_{\alpha} p_{\alpha} \exp\left(-\frac{G_{\alpha}(\lambda_0)}{k_{\mathrm{B}}T}\right)\right) := \log\left(\langle \exp\left(-\frac{G_{\alpha}(\lambda_0)}{k_{\mathrm{B}}T}\right)\rangle\right). \tag{7.10}$$

Then, for Gaussian distributions, the following relation holds:

$$\log\left(\langle \exp\left(-\frac{G_{\alpha}(\lambda_0)}{k_{\mathrm{B}}T}\right)\rangle\right) = -\frac{G^{\star}}{k_{\mathrm{B}}T} + \frac{\sigma_G^2}{2(k_{\mathrm{B}}T)^2}, \tag{7.11}$$

where σ_G^2 is the variance of the folding free energy distribution. On the other hand, the application of Jensen's inequality to the second term of Eq. (7.9) yields:

$$\log\left(\langle \exp\left(\frac{G_{\alpha}(\lambda_0)}{k_{\mathrm{B}}T}\right)\rangle\right) \geq \log\left(\exp\left(\frac{\langle G_{\alpha}(\lambda_0)\rangle}{k_{\mathrm{B}}T}\right)\right) = \frac{G^{\star}}{k_{\mathrm{B}}T}. \tag{7.12}$$

We stress that the latter expected value is done over the uniform distribution, so the Gaussian approximation does not affect the (7.12) result. Then, by summing Eq. (7.11) and (7.12) and inserting the result in Eq. (7.9) we obtain the upper bound for $\mathcal{I}$:

$$k_{\mathrm{B}}T\mathcal{I} \geq \frac{\sigma_G^2}{2\,k_{\mathrm{B}}T}. \tag{7.13}$$

Interestingly, the bound set by Eq. (7.13) only depends on the variance of the free energies. This result looks reasonable, since the information-content of a molecular ensemble should not depend on the molecule (i.e. the mean free energy) but on the widespread of the free energy spectrum (for a discussion of the sample size in variance estimation see Sec. 6.3.2).

7.3 RESULTS

7.3.1 *Information-content measurement*

To show the applicability of Eq. (7.9) we extract the information-content of the molecular ensembles presented in chapter 6, whose folding free

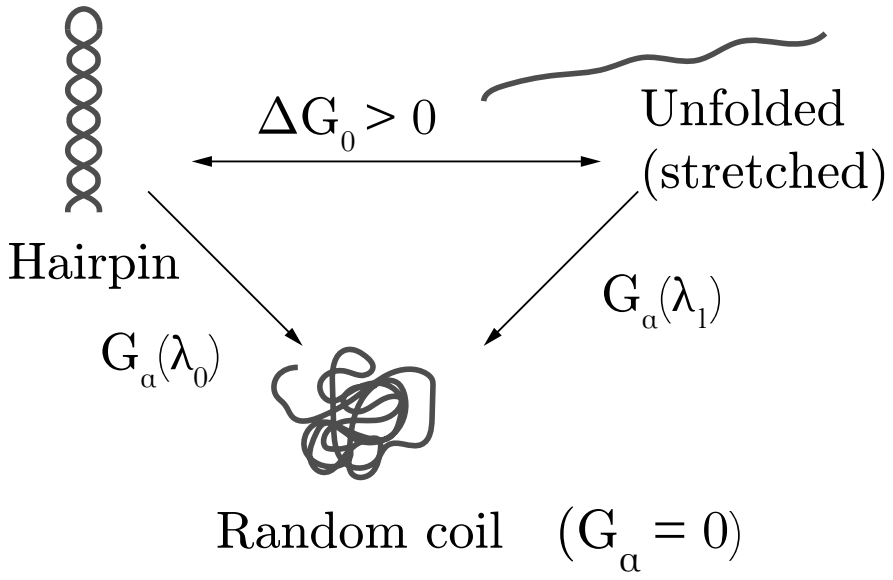

Figure 7.2: **Reference states for free energy measurement.** $G_\alpha(\lambda_0)$ and $G_\alpha(\lambda_1)$ are the energies of the hairpin state and the unfolded (stretched) state, respectively. Both quantities are measured with respect to the random coil state and $\Delta G_0 = G_\alpha(\lambda_1) - G_\alpha(\lambda_0) > 0$.

energies spectra were obtained in Sec. 6.3.1. We recall that the number of molecules we have used in the present study are 40, 93 and 54 for the 2M, 4M and 6M molecular ensembles, respectively. Moreover, for every molecule we have used a similar number of pulls (~ 100) and we subtracted the stretching contributions to every W as explained in appendix C. Also, it is important to bear in mind that all the folding free energies we extract (as in chapters 4 and 8) are measured with respect to the random coil state (see Fig. 7.2). Now, in the present framework, $G_\alpha(\lambda_0)$ is also measured with respect to the random coil. Therefore, $G_\alpha(\lambda_0) = -\Delta G_0$, where ΔG_0 (> 0) is the free energy of formation of the hairpin structure (Fig. 7.2).

In Fig. 7.3 we show the EWD distributions computed according to Eq. (7.4) (F distribution) and (7.6) (R distribution) using, in all cases, $p_\alpha = 1/M$. On the other hand, $\hat{p}_\alpha$ are given by Eq. (7.7). In Fig. 7.3(a) we show the EWD we obtained for the three molecular ensembles. Now, according to Eq. (7.5), the crossing point between F and R distribution corresponds to the work value equal to $\Delta\mathcal{G} - k_B T\mathcal{I}$, rather than the bare ΔG value as it is the case when using the standard CFT for individual molecules.

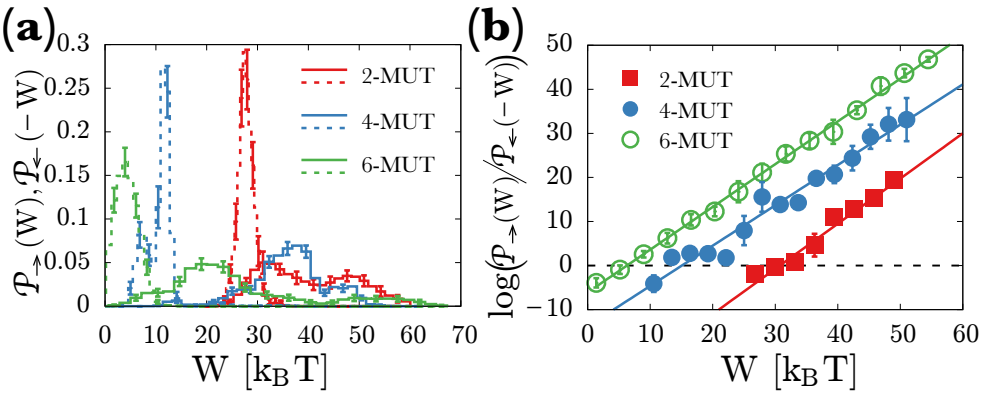

Figure 7.3: **Experimental verification of the information-content fluctuation theorem. (a)** - Forward (solid lines) and reversed (dashed lines) work distributions obtained using Eqs. (7.4), (7.6) and imposing the relation given by Eq. (7.7) with $p_\alpha = 1/M$. **(b)** - Logarithm of the ratio of the work distributions shown in the (a) panel and plot of straight lines with slope equal to 1 (in $k_B T$ units). X-intercepts correspond to the work values, according Eq. (7.5), equal to: $\Delta\mathcal{G} - k_B T\mathcal{I}$. In all cases, error bars have been calculated using the Bootstrap method.

On the other hand, the validity of the information-content fluctuation theorem can be tested in the usual way. By extracting logarithms at both sides of Eq. (7.5) we have:

$$\log\left(\frac{\mathcal{P}_{\rightarrow}(W)}{\mathcal{P}_{\leftarrow}(-W)}\right) = \frac{W}{k_B T} - \frac{\Delta\mathcal{G} - k_B T\mathcal{I}}{k_B T}\,. \tag{7.14}$$

Then, if the fluctuation theorem is satisfied, the logarithm of the ratio of the EWD will follow a straight line with slope equal to 1 and x-intercept equal to $\Delta\mathcal{G} - k_B T\mathcal{I}$ (both in $k_B T$ units). In Fig. 7.3(b) we show the experimental validation of the information-content for the molecular ensembles we studied.

Having in mind the validity of the information-content fluctuation theorem, we aim to measure the precise value of $\mathcal{I}$. The measurement can be done using three distinct approaches. First, we can use the closed formula for $\mathcal{I}$ (Eq. (7.9)) inserting the measured $p(\Delta G_0)$ (see previous chapter). Second, the information-content can be also measured in the framework of partial measurements and thermodynamic inference. Third, we show that $\mathcal{I}$ can be estimated numerically. In what follows, we apply all three methods in order to quantify $\mathcal{I}$ for all the molecular ensembles.

Theoretical prediction for the information-content $\mathcal{I}$

We note that the information-content $\mathcal{I}$ (Eq. (7.9)) can be written as:

$$\begin{aligned}\mathcal{I} = &\log\left(\int d\left(\Delta G_0\right)\ p(\Delta G_0)\exp\left(\frac{\Delta G_0}{k_{\mathrm{B}}T}\right)\right)\\ &+\log\left(\int d\left(\Delta G_0\right)\ p(\Delta G_0)\exp\left(\frac{-\Delta G_0}{k_{\mathrm{B}}T}\right)\right).\end{aligned} \tag{7.15}$$

where $p(\Delta G_0)$ is the folding free energy distribution. This method requires the measurement of the folding free energy spectrum of each molecular ensemble, which can characterized with few tens of molecules (see Sec. 6.3.1). The simplicity of this method lies in the fact that it only requires the calculation of two expected values. In Table 7.1 we report the values for the information-content obtained using Eq. (7.15).

Thermodynamic inference of the information-content $\mathcal{I}$

The information-content fluctuation theorem is valid only when the weights p_α and $\hat{p}_\alpha$ fulfil the relation given by Eq. (7.7). Hence, one may ask what is the effect of ignoring such constraint between the aforementioned frequencies. Let us consider the simplest case, equal and constant (yet normalized) weights equal to $p_\alpha = \hat{p}_\alpha = 1/M$. In this scenario, the EWD are equal to the white average of the individual work distributions:

$$\mathcal{P}^{\mathrm{white}}_{\rightarrow}(W) = \frac{1}{M}\sum_{\alpha} P^{(\alpha)}_{\rightarrow}(W)\,, \tag{7.16}$$

$$\mathcal{P}^{\mathrm{white}}_{\leftarrow}(-W) = \frac{1}{M}\sum_{\alpha} P^{(\alpha)}_{\leftarrow}(-W)\,. \tag{7.17}$$

In Fig. 7.4(a) we show the EWD obtained using Eqs. (7.16), (7.17). While R distributions (dashed lines) are the same than in Fig. 7.3, F distributions are significantly different from the ones in Fig. 7.3. In the present case, all molecules are equally weighted due to the factor $1/M$, whereas in Fig. 7.3 the molecules with higher free energies dominate the shape of the distribution. On the other hand, the information-content fluctuation theorem does not hold for the previous EWD since they are

wrong. This fact can be seen in Fig.7.4(b), where the solid line is the same line than in Fig. 7.3(b) and the slopes of the experimental data are clearly lower than 1. The slopes we measure for the experimental data are 0.11 ± 0.03, 0.074 ± 0.010 and 0.05 ± 0.01 for the 2M, 4M and 6M ensembles, respectively.

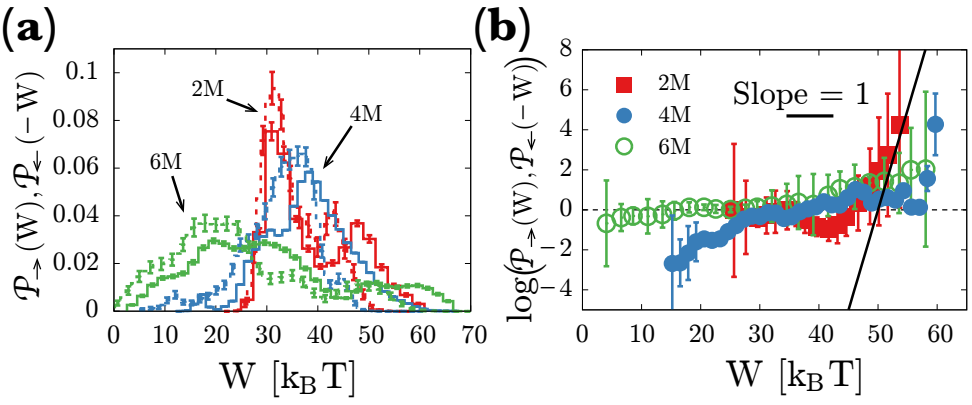

Figure 7.4: **Breakdown of the information-content fluctuation theorem. (a)** - Forward (solid lines) and reversed (dashed lines) work distributions obtained using Eqs. (7.16), (7.17). **(b)** - Logarithm of the ratio of the work distributions shown in the (a) panel and plot of the same straight line of Fig. 7.3 (slope equal to 1 in $k_B T$ units). In all cases, error bars have been calculated using the Bootstrap method.

Despite Eqs. (7.16) (7.17) correspond to the wrong EWD they allow us to infer the information-content of each ensemble. The breakdown of the fluctuation theorem can be analytically quantified from Eqs. 7.16, 7.17. To do so, we consider that the individual work distributions (both F and R) are Gaussian distributions with identical variances equal to σ_W^2. In this conditions, the CFT is satisfied for every molecule. Moreover let us consider that the ensemble folding free energy distribution is well-reproduced by a Gaussian distribution $\mathcal{N}(G^\star, \sigma_G^2)$. In this approximation, the EWD (Eqs. (7.16), (7.17)), fulfil a quasi-fluctuation theorem (explicit calculations can be found in section E.4) given by:

$$\frac{\mathcal{P}_{\rightarrow}^{\text{white}}(W)}{\mathcal{P}_{\leftarrow}^{\text{white}}(-W)} = \exp\left(x\frac{W - G^\star}{k_B T}\right) \quad , \quad x = \frac{1}{1 + \frac{\sigma_G^2}{\sigma_W^2}} . \tag{7.18}$$

Where x is a similar parameter to the one found in Eq. (4.14). Indeed x governs the breakdown of the fluctuation theorem. We note that

the traditional CFT symmetry will be restored in the limit $\sigma_G^2 \to 0$ (i.e. an homogeneous or equilibrium phenotypic ensemble) or in the regime where $\sigma_W^2 >> \sigma_G^2$. In this latter case, work fluctuations mask heterogeneous effects, yielding a similar situation to the one explained in Sec. 4.4, where work fluctuations in the limit of infinite pulling speed are so high that ExtEns and ForceEns become equivalent. Interestingly, our approximation for x (Eq. (7.18)) is in good agreement with the experimental data. The slopes predicted by x are equal to 0.06 ± 0.02, 0.063 ± 0.010 and 0.044 ± 0.010 for the 2M, 4M and 6M ensembles, respectively. We remind that the experimental slopes are equal to 0.11 ± 0.03, 0.074 ± 0.010 and 0.05 ± 0.01, respectively.

In the context of thermodynamic inference we can argue that the fluctuation theorem is not satisfied because we are not measuring the correct mechanical work, so $\mathcal{P}_{\rightarrow}(W)$ and $\mathcal{P}_{\leftarrow}(-W)$ are not the full work[1] distributions [60, 124]. In the Gaussian approximation, the full work distributions can be inferred imposing that $\mathcal{P}_{\rightarrow}(W)$ and $\mathcal{P}_{\leftarrow}(-W)$ satisfy the fluctuation theorem with the same free energy $G^\star$. Within our scheme, the full work distributions can be recovered by adding to each W the following quantity (explicit calculations can be found in section E.5):

$$\Delta = \frac{\sigma_G^2}{2\, k_B T}\,. \tag{7.19}$$

This quantity is reminiscent of the upper bound of the information-content (Eq. (7.13)) so we may write: $\Delta = k_B T \mathcal{I}$. In Eq. (7.19) we replace σ_G^2 by its value appearing in x in the Gaussian approximation (Eq. (7.18)) so we rewrite previous equation as:

$$k_B T \mathcal{I} = \frac{1-x}{2x}\,\frac{\sigma_W^2}{k_B T}\,. \tag{7.20}$$

Hence, the information-content can be inferred, rather than directly measured, as follows: first, quantify the breakdown of the fluctuation theorem (i.e. measuring x) when the EWD are calculated according to Eqs. (7.16), (7.17). Then, insert the variance of the work distributions, σ_W^2 and x in Eq. (7.20). The calculation of Eq. (7.20) yields the

1 So an energetic contribution to W is systematically lost (or not measured).

information-content of the molecular ensemble. The values of the variances of the work distributions are: 3.4 ± 0.4, 3.8 ± 0.4 and 4.0 ± 1.4 $(k_{\mathrm{B}}T)^2$ for the 2M, 4M and 6M ensemblee, respectively. In Table 7.1 we report the values of the information-content obtained using Eq. (7.20).

Numerical estimation of the information-content $\mathcal{I}$

Within the Gaussian scheme, the information-content $\mathcal{I}$ relies on the assumption that the function $H(W)$, defined as:

$$H(W) = \log\left(\frac{\mathcal{P}_{\rightarrow}(W)}{\mathcal{P}_{\leftarrow}(-W)}\right), \tag{7.21}$$

is linear on the mechanical work W. This supposition may not be true in all circumstances. For instance, the data we shown in Fig. 7.4(b) corresponding to the 2M ensemble, is not a linear function of W at all. Thus, the quantification of the information-content via the slope of the fluctuation theorem is hard. In order to circumvent this situation, let us propose that the uniform (i.e. $p_\alpha = \hat{p}_\alpha = 1/M$) EWD fulfil an information-content fluctuation theorem given by:

$$\frac{\mathcal{P}^{\mathrm{white}}_{\rightarrow}(W)}{\mathcal{P}^{\mathrm{white}}_{\leftarrow}(-W)} = \exp\left(\frac{W - G^\star + k_{\mathrm{B}}T\mathcal{I}}{k_{\mathrm{B}}T}\right). \tag{7.22}$$

Where again $\mathcal{I}$ is the information-content of the molecular ensemble and $G^\star$ is the mean free energy of the molecular ensemble. We note that, in this scenario, $\Delta\mathcal{G}$ (Eq. 7.5) is equal to $G^\star$. Equation (7.22) can be rearranged as:

$$\exp\left(-\frac{W - G^\star}{k_{\mathrm{B}}T}\right)\mathcal{P}^{\mathrm{white}}_{\rightarrow}(W) = \exp\left(\mathcal{I}\right)\mathcal{P}^{\mathrm{white}}_{\leftarrow}(-W). \tag{7.23}$$

Then, integrating over W in Eq. (7.23) we have a Jarzynski-like relation:

$$\left\langle \exp\left(-\frac{W - G^\star}{k_{\mathrm{B}}T}\right)\right\rangle_{\rightarrow} = \exp\left(\mathcal{I}\right), \tag{7.24}$$

where $\langle\cdots\rangle_{\rightarrow}$ denotes the average over the white F EWD (Eq. (7.16)). Hence, by numerically solving for $\mathcal{I}$ Eq. (7.24), we obtain an estimation

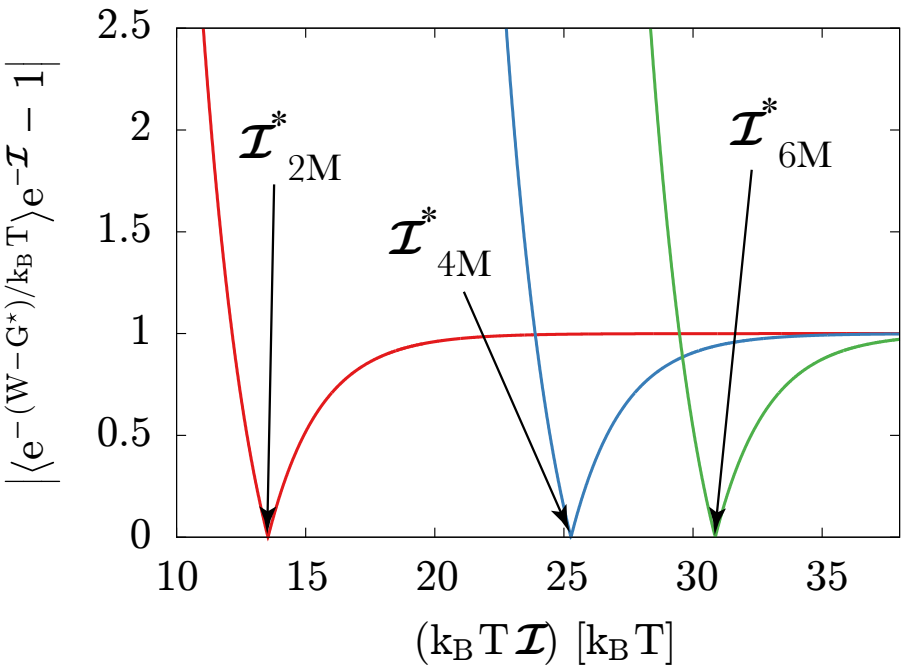

Figure 7.5: **Numerical estimation of information-content.** Plot of Eq. (7.24) as a function of $k_B T\mathcal{I}$. Minima ($\mathcal{I}^\star$) correspond to the information-content of each molecular ensemble.

of the information-content of each molecular ensemble. In Table 7.1 we show the obtained values of the information-content when solving Eq. (7.24) and in Fig. 7.5 we show the numerical minimization of Eq. (7.24).

7.3.2 *Summary of results*

In this section we summarize and report the results for the information-content $\mathcal{I}$ we obtained for the three molecular ensembles (2M, 4M and 6M). Values can be found in Table 7.1.

Interestingly, all the results are compatible among them. This supports our theoretical findings regarding the information-content of an arbitrary molecular ensemble (Eq. (7.9)). The Gaussian approximation of $\mathcal{I}$ is less accurate for the 6M ensemble, highlighting the fact that a large number of molecules are required in order to have a good estimation of the profile of the $H(W)$ function (Eq. (7.22)) and, in consequence, of its slope. Moreover, it is worth mentioning that the upper bound of the information-content provides a good estimation of the

	$k_B T\mathcal{I}_{exp}$ [$k_B T$]	$k_B T\mathcal{I}_{Gaussian}$ [$k_B T$]	$k_B T\mathcal{I}_{JE}$ [$k_B T$]	Upper bound [$k_B T$]
2M	12 ± 2	14 ± 2	13 ± 1	12 ± 1
4M	26 ± 6	24 ± 2	25 ± 1	23 ± 1
6M	31 ± 5	38 ± 4	31 ± 1	33 ± 1

Table 7.1: **Information-content measurement.** Summary of the results of the information-content measurement using the three methods described in Sec. 7.3.1. $k_B T\mathcal{I}_{exp}$ corresponds to Eq. (7.15), $k_B T\mathcal{I}_{Gaussian}$ has been obtained using Eq. (7.20), $k_B T\mathcal{I}_{JE}$ using Eq. (7.24) and the upper bound has been obtained using Eq. (7.13) with the variance predicted by Mfold when numerically folding 256, 5000 and 50000 molecules for the 2M, 4M and 6M ensemble, respectively. CD4 data corresponds to the non-mutated ensemble (i.e. the native CD4, see chapter 4). Errors have been obtained by propagation of the experimental uncertainties.

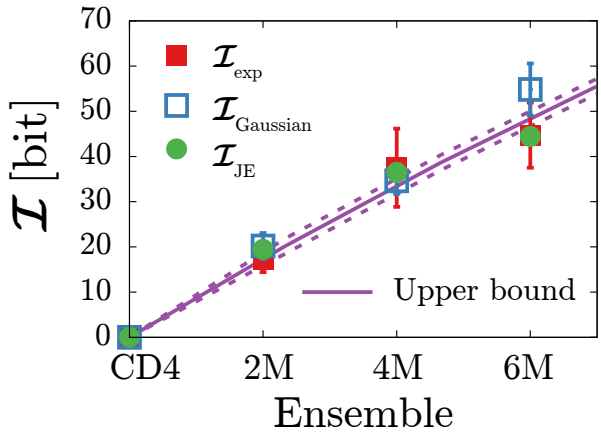

Figure 7.6: **Experimental energy-to-information conversion.** Conversion of the information-content of each molecular ensemble from energy to bits (1 bit $= k_B T \log 2$). Solid line corresponds to the upper bound of the information-content and the dashed lines are its uncertainty obtained by propagation.

information-content, as well. Hence, the Gaussian approximation is reasonable for estimating information-contents in molecular ensemble.

In Fig. 7.6 we plot the energy-to-information conversion, where the information-content of each ensemble is shown in bits. As expected, information-content is a monotonous increasing function with the number of mutations. Regarding the energy-to-information conver-

sion, we have used the equivalence set by the Landauer limit (1 bit $= k_B T \log 2 \approx 0.69\ k_B T$).

7.4 CONCLUSIONS

In this chapter we have set the theoretical basis for establishing the connection between energy and information. We have shown that in SME this procedure can be efficiently implemented in order to conduct systematic energy-to-information measurements. Moreover, the information-content $\mathcal{I}$ can be unambiguously defined by using only thermodynamic considerations. Likewise, the information-content is well defined regardless experimental conditions as it only depends on thermodynamic equilibrium (measurable) quantities.

This chapter must be taken, together with chapter 6, as a single package. While in the preceding chapter 6 we characterized the variability of heterogeneous molecular ensembles and discussed the biological and medical implications, in the present chapter we have demonstrated how not only molecular ensemble have an information-content that is intrinsic to their nature, but also how to measure it. To the best of our knowledge, this kind of information-content measurements have never been carried out. The work presented in this chapter aims to spur new research lines in which information-content measurements can be combined with thermodynamic and kinetic measurements in order to grant access to a full and unprecedented level of characterization of biophysical systems.

Part IV

SPECIFIC BINDING

8

EXPERIMENTAL MEASUREMENT OF THE SPECIFIC BINDING ENERGY OF MAGNESIUM CATIONS TO AN RNA THREE-WAY JUNCTION

8.1 INTRODUCTION

One of the most important ingredients regulating intracellular biomolecular reactions are electrostatic forces. The structure of biopolymers is essentially determined by their ionic charge and the concentration of dissociated ions in the surrounding environment (solvent). Both factors also affect the binding strength to ligands. As we explained in Sec. 2.1, nucleic acids have a net charge due to the presence of phosphate groups in the outer backbone. Indeed, they are one of the most densely charge polymers of all. DNA and RNA have a linear charge density of $2e^-$ every, approximately, 3 Å (see Fig. 2.2), resulting in a repulsive force per basepair in water of order $\sim$ 1 pN. Such repulsive force is counterbalanced by base stacking interactions that stabilize the double helical structure. Base interactions can be either specific or non specific. Examples of specific interactions are the interaction with metal ions, where the hydration sphere of the ion coordinates with charged grups of the different nucleobases. Moreover, since the negative charge of the phosphate backbone is screened by the net positive charge of the ion, the interaction with metal ions also has a non-specific aspect.

Non-specific electrostatic interactions can be described phenomenologically using generalized activity theories of electrolytes, mean field approaches such as the Debye-Hückel theory, Gouy-Chapman, Poisson-Boltzmann [125, 126], the tightly bound ion (TBI) model [127] or the DLVO Theory [128, 129]. On the other hand, much less is known for specific charge interactions between metal ions binding to DNA and RNA structures [130, 131].

Divalent ions, such as magnesium, play a leading role in RNA folding [132]. The strong repulsive forces between basepairs difficult RNA to fold into a compact structure, but thanks to the surrounding positive ions, the folding is promoted by the reduced repulsion between the

charges of the phosphate groups. Indeed, millimolar concentrations of Mg^{2+} are able to stabilize RNA tertiary structures (see Sec. 2.1.1) [133]. However, the effect of Mg^{2+} in RNA folding goes beyond charge screening as they can specifically bind to RNA. Magnesium ions act as a major driving force for tertiary structure formation since the two free positive charges of magnesium are able to specifically recognise the negatively charged hydroxyl group of the ribose and, hence, two distant nucleotides can be brought together in order to form a stable tertiary structure [134]. Regarding non-specific electrostatic screening effects, they can be described using the 100/1 phenomenological rule which states that the non-specific binding affinity of a given concentration of divalent cations is equal to that of 100-fold times large concentration of monovalent cations[1] [135–137]. This rule has been recently verified in SME using RNA [138] and DNA hairpins [139].

In this chapter we aim to disentangle the non-specific and specific electrostatic contributions of magnesium to the stabilization energy of a 3-way helix RNA junction (hereafter referred to as 3WJ) using SME. The chapter is organized as follows: in the following section the biological context and relevance of the 3WJ is explained, as well as the effects that divalent cations have on the 3WJ structure. Next, the 3WJ is studied with classical Dynamic Force Spectroscopy (DFS), characterizing the folding-unfolding pathways of the 3WJ. In the final section, the free energy of formation of the 3WJ molecule is extracted by combining irreversible work measurements and the extended fluctuation theorem.

It is worth mentioning that although it is not the first time that 3WJ is a target of single molecule studies, previous experimental assays done with LOT instruments were unable to separate the two electrostatic contributions and only the full folding free energy was measured [50].

8.2 WHY THE 3WJ RNA?

As we already mentioned in the general introduction of the thesis (see Chapter 2), RNA is an essential molecule participating in, mainly, protein synthesis and regulatory processes. In particular, proteins are assembled thanks to complex molecular machineries, such as the ribosomes. Ribosomes are formed by a mix of proteins and RNA (see Sec.

1 For instance, the stabilizing contribution to RNA of 10mM $MgCl_2$ is equivalent to that of 1M NaCl.

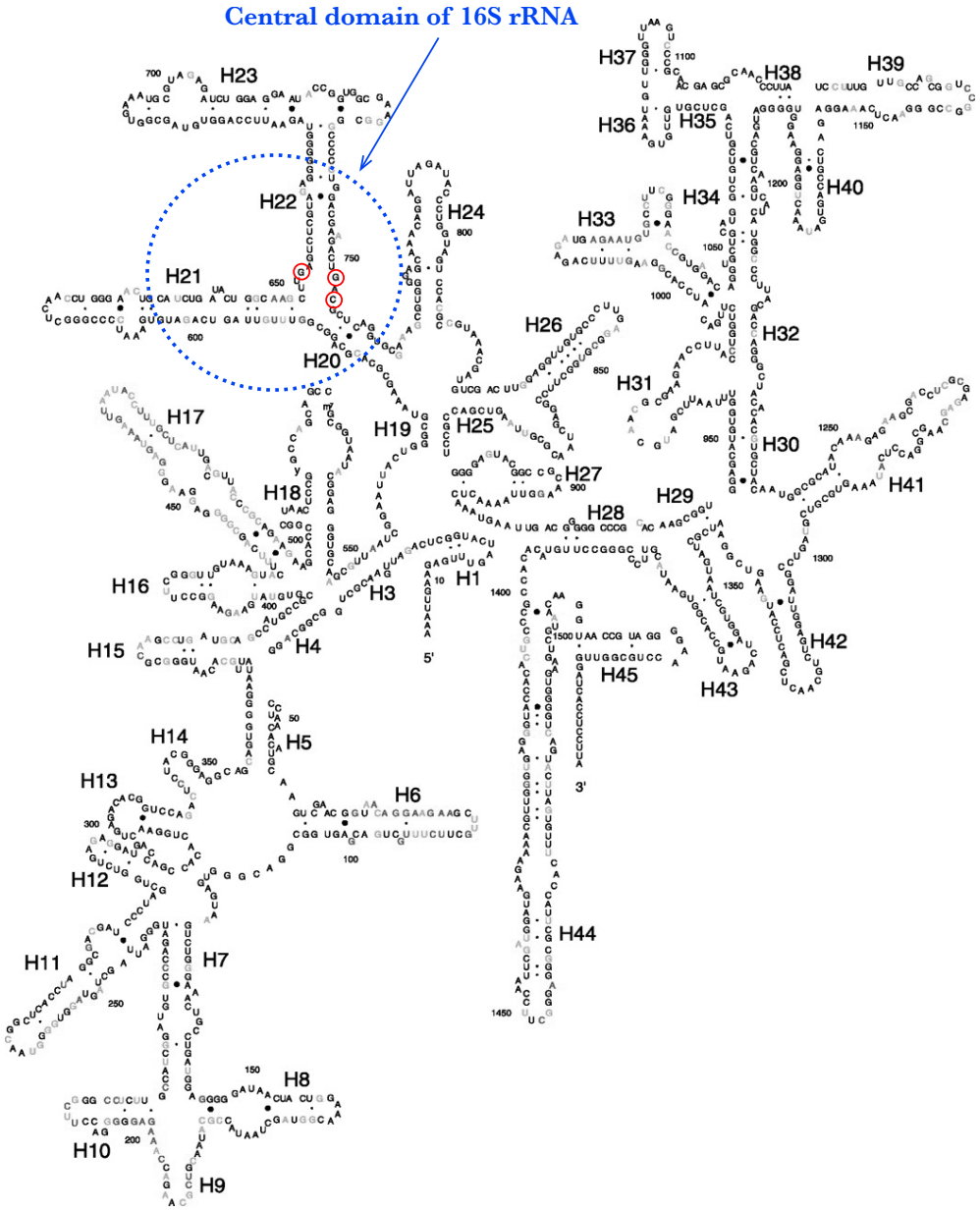

Figure 8.1: **Secondary structure of 16S rRNA.** Helices are indicated with a capital H and a number. Red circles mark the positions C754-G654-C752 (magnesium binding site). Figure adapted from Ref. [146].

2.1.2). The ribosomal RNA (rRNA) consist of two major components: the small ribosomal units (which read the mRNA) and the big ribosomal units (where the machinery essential for protein synthesis is).

Prokaryotes have 70S[2] ribosomes, each one consisting of a 50S large and a 30S small subunits. While the large subunit is formed by two types of rRNAs, the 30S subunit is formed by a 16S rRNA of, approximately, 1540 nucleotides long. In Fig. 8.1 we show the secondary structure of 16S rRNA, each helix is labelled according to the standard nomenclature. The central domain of 16S rRNA (H20, H21 and H22 helices in Fig. 8.1) is a highly-conserved site shared above 95% across all known eubacterial sequences [140]. This fact makes 16S rRNA particularly important for tracing studies and reconstructing phylogenetic trees [141].

Even though the central domain of 16S rRNA is usually found in extended conformations, the crystal structure reveals its three-helix junction structure [142]. Moreover, the central domain of 16S rRNA is able to bind to the 89 aminoacids small protein S15, a key protein for the assembly of the whole ribosome. S15 protein interacts with a G-U/G-C motif in the 3WJ and this interaction is mediated by magnesium. Indeed, the presence of magnesium ions enhances the characteristic binding rates of S15 to the 3WJ RNA [143]. Upon adding magnesium, specific binding of three Mg^{2+} ions to the G754, G654 and G752 nucleotides [144] in the 3WJ (indicated with red circles in Fig. 8.1) induce a conformational change in the 3WJ by changing the relative angular positions of helices H20, H21 and H22 [145].

The 3WJ molecule, together with its biological importance and all the existing studies about its structure, is a useful molecular system that allow us to dig into the effects of magnesium ions in kinetic and structural properties.

8.3 FORCE SPECTROSCOPY OF THE 3WJ MOLECULE

Force spectroscopy experiments were performed on an RNA molecule (i.e. the 3WJ RNA) containing the highly-conserved site of a 16S rRNA complex [147] using LOT. In previous single-molecule assays it was discovered that the 3WJ RNA molecule has a force-induced misfolded state, unnoticed by bulk techniques [50, 148]. The study of misfolded

2 S stands for Svedberg unit. It accounts for a particle's size based on its sedimentation rate.

molecular structures has turned out to be a hot topic in biophysics for their potential in the development of diseases. A remarkable case is, for instance, the case of prions (i.e. misfolded proteins), which are responsible of several neurodegenerative (fatal) diseases [149]. Misfolding happens because there is a large number of competing structures in the folding pathway that can kinetically trap the molecule. In our particular system, the application of force cycles, favours the formation of a stable secondary structure rather different from the native one. Moreover, it was demonstrated that the misfolding probability of the RNA 3WJ strongly depends on the experimental conditions in a nontrivial fashion [148]. In the left panel of Fig. 8.2 we show the native structure of the RNA 3WJ molecule whereas in the right panel we represent the misfolded structure proposed in Ref. [148]. Misfolded structure is composed by two short RNA hairpins (H_1^M and H_2^M) connected in series by three unpaired bases (see Fig. 8.2, right).

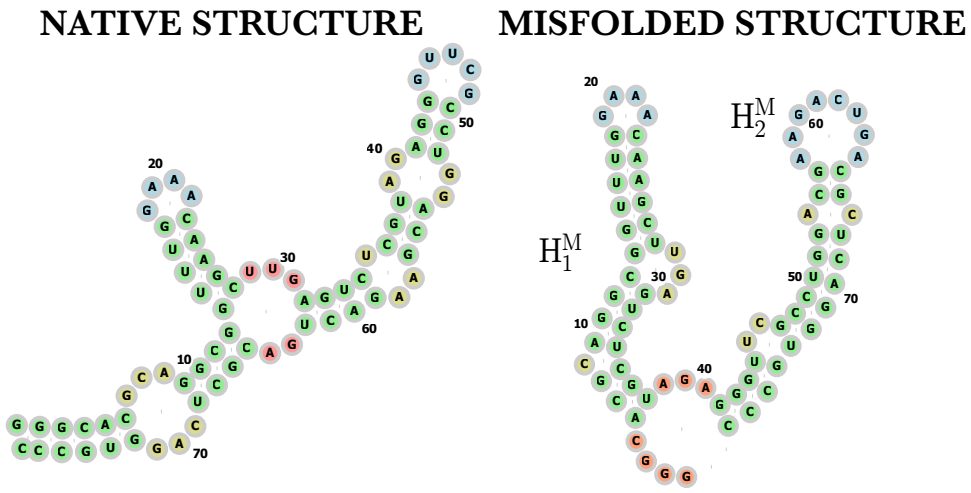

Figure 8.2: **Structures of 3WJ.** Native structure of the RNA 3WJ (left figure) and proposed misfolded structure (right figure). Colors indicate the type of motif: green color indicate that bases are forming Watson-Crick basepairs, blue indicates the formation of outer loops, brown bases are inner loop bases and red color indicates that bases are forming a single-stranded chain.

Given an RNA sequence, the folding free energy and the corresponding secondary structure can be predicted using the Vienna package or Mfold software [73, 150]. We have, for the native 3WJ structre, a folding free energy of $\Delta G_0^N = -39.5\,\text{kcal/mol} = -67\,k_B T$ at standard conditions ($T = 298$ K, [NaCl] = 1M). Regarding the misfolded structure,

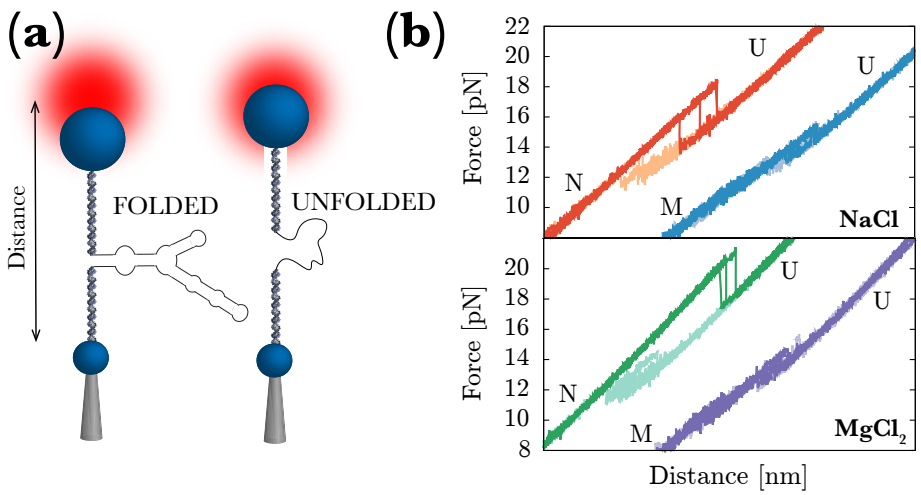

Figure 8.3: **Dynamic force spectroscopy on RNA 3WJ. (a)** - LOT experimental setup. The distance between the micropipette and the center of the optical trap is the control parameter of the experiment. **(b)** - Typical FDC when unfolding the native structure (leftmost curves of both panels) and the misfolded structure (rightmost curves of both panels). Top panel corresponds to experiments in monovalent conditions whereas bottom panel corresponds to experiments performed in divalent salt conditions.

the free energy of formation is equal to $\Delta G_0^{\mathrm{M}} = -29.6\,\mathrm{kcal/mol} = -50\,k_{\mathrm{B}}T$, resulting in a difference with respect to the folding free energy of the native structure of less than $20\,k_{\mathrm{B}}T$.

Upon performing nonequilibrium pulling experiments (same protocol as described in Sec. 6.3 see Fig. 8.3(a)) we observe two types of unfolding/folding patterns that we interpret as the stretching the native structure or the misfolded structure of the 3WJ.

Most of the times ($\simeq$ 90%), the unfolding curves display a single force jump event (leftmost curves of Fig. 8.3(b)) resulting of the cooperative unfolding in all the 77 bases forming the 3WJ. Therefore, we associate this kind of FDCs with the unfolding of the native 3WJ. The refolding pathway of the native structure display two events: one minor folding event resulting in the formation of a partial structure and another minor event in which the original native structure is recovered.

Less frequently ($\simeq$ 10% of the times), we observe FDCs that we associate to the unfolding of the misfolded structure of the 3WJ (rightmost curves of Fig. 8.3(b)). Previous studies proposed that the behavior we observe when unfolding the misfolded structure is due, first of all, to

the non-cooperative unfolding (like a zipper) of H_1^M hairpin (around $10 - 12$ pN) and, finally, to the cooperative unfolding of the H_2^M (force rip around ~ 15 pN) [148].

In the following sections we study the force-dependent kinetics of both, native and misfolded structure, in order to obtain the maximum amount of information of each structure in different salt conditions. Then, combining the results from the DFS experiments with the theoretical prediction for the FEL are able to unravel some structural properties of the 3WJ molecule.

Our experiments were performed in two distinct, salt conditions. On the one hand, experiments were done in an aqueous buffer containing 1 M NaCl (hereafter referred to as monovalent conditions). On the other hand, experiments were repeated using an aqueous buffer that contains 50 mM NaCl and 10 mM $MgCl_2$ (hereafter referred as divalent conditions). Note that applying the 100/1 rule between monovalent and divalent conditions, both salt conditions are equivalent from the ionic strength point of view.

Folding-unfolding kinetics can be investigated by means of equilibrium hopping experiments or non-equilibrium pulling experiments. In hopping experiments, the control parameter (i.e. the distance between the trap and the micropipette, labelled as "Distance" in Fig. 8.3) is kept fixed and the system is able to explore equilibrium states (i.e. sampled according to the Boltzmann-Gibbs distribution). By monitoring the force changes as a function of the time (hopping traces), kinetics can be obtained. The main drawback is that kinetic barriers between different conformational sates might be large (specially for long molecules) and therefore the kinetics are too slow to be explored by means of hopping experiments. From the pulling experiments, force-dependent kinetics are obtained via first-rupture forces measurements. A molecule that is subjected to a pulling experiment undergoes stochastic transitions over different conformations at different forces. By recording the first rupture forces and extracting the survival probabilities of the corresponding states can be obtained (see appendix D).

8.3.1 *Native structure*

Throughout this section we will only focus in trajectories involving the **native structure**. Moreover, despite the fact that the footprints of the

unfolding of the native 3WJ are sudden force jumps (f_U in Fig. 8.4(a)), folding is less well-defined. We recorded the first-refolding forces (f_F in Fig. 8.4(a)) despite there is a gentle previous non-cooperative folding, as can be noticed in light curve of top panel of Fig. 8.3(b). For this reason, the folding kinetics we measured do not correspond to U → N transitions. They correspond, instead, to transitions between a kinetic intermediate state (labelled as I_N) and the native state: $I_N \rightarrow N$.

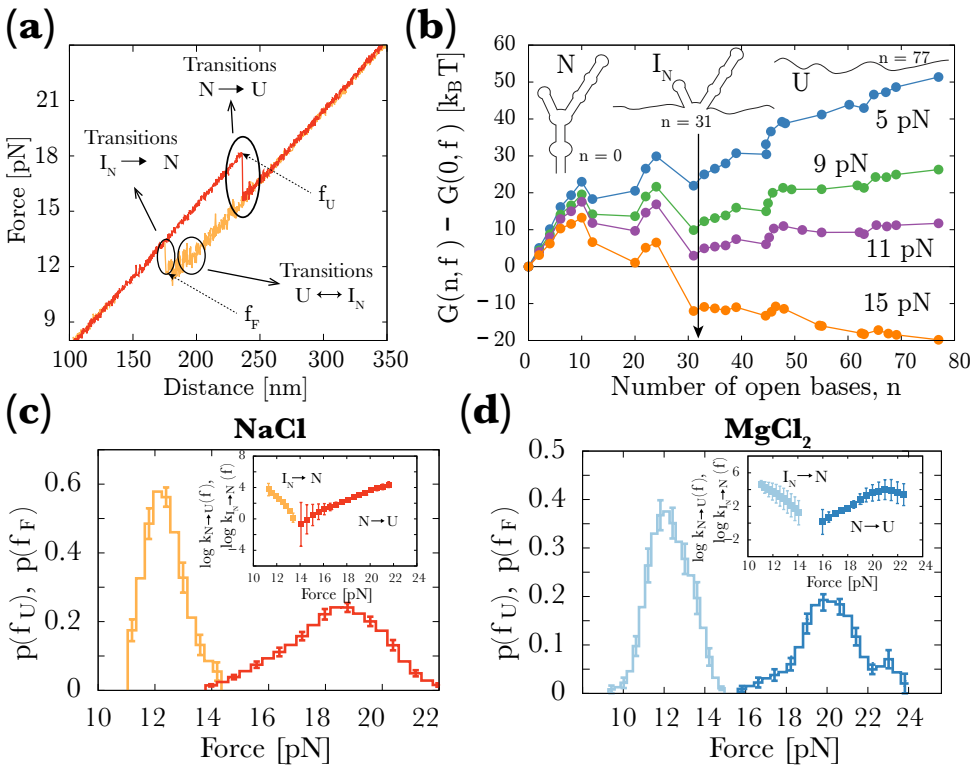

Figure 8.4: **Unfolding-folding pathways, FEL and force kinetics of the native structure of the 3WJ RNA. (a)** - Example of an unfolding (dark curve) trajectory and a folding (light curve) trajectory. The structures that the molecule explores are shown as cartoons in the graph. **(b)** - FEL of native 3WJ at different forces and sketches of intermediate kinetic states. **(c)** - Distributions of first-rupture and first-folding forces. Bell-Evans kinetics are shown as inset. $k_{N\rightarrow U}$ are shown as dark symbols whereas $k_{I_N\rightarrow N}$ are shown as light symbols. **(d)** - Distributions of first rupture (and refolding) forces and Bell-Evans kinetics (inset) in divalent salt conditions. In both graphs dark curves correspond to unfolding forces and light curves correspond to folding forces. Regarding kinetics, the same color code holds than in **(c)** panel. In both cases, errors were computed using the Bootstrap method. Pulling speed is 200 nm/s for both cases.

	$x_{\text{N-TS}}$ [nm]	$x_{\text{I}_\text{N}\text{-N}}$ [nm]
NaCl	5.5 ± 0.7	6.7 ± 0.4
$MgCl_2$	5.3 ± 0.4	5.4 ± 1.0

Table 8.1: **Fit of the kinetic rates for the native 3WJ to the Bell-Evans model.** Parameters are obtained by linear fitting Eqs. (5.5) and (5.6) to the data shown in Fig. 8.4(c,d). Results obtained by averaging the individual obtained parameters of four different pulling speeds (50, 100, 200 and 500 nm/s) for six different molecules.

The FEL is a useful tool when studying molecular folding dynamics. It allows us to relate molecular structural properties with the experimentally measured kinetic rates. We have calculated the theoretical profile of the FEL (see section 5.1) for the native 3WJ molecule using the elastic parameters of the RNA molecule and the handles reported in Ref. [148]. In Fig. 8.4(b) we show the FEL calculated at different forces as a function of the unpaired bases. Interestingly, for forces around 11 - 12 pN, the FEL has a local minimum corresponding to an intermediate state composed by, approximately, 46 bases, which might be associated with the I_N state observed in the FDCs.

Figure 8.4(c) shows the rupture forces distributions and Bell-Evans kinetics (already introduced in Sec. 5.2) for monovalent conditions, whereas the analogous plots in divalent conditions are shown in Fig. 8.4(d). Kinetics have been obtained as explained in appendix D. By comparing rupture force distributions, we note that the presence of divalent ions increase the mean unfolding force by ~ 2 pN. This effect points towards the stabilizing role of magnesium ions in RNA. In contrast, folding distributions do not seem affected by the presence of magnesium ions in the buffer. This fact might indicate us the magnesium binds to the 3WJ after forming the native structure.

Transition state distances (i.e. the distance from the native, N, and unfolded, U, state to the kinetic barrier) are obtained by performing linear fits of Eqs. (5.5) and (5.6) to the data shown as insets in Fig. 8.4(c,d) and the results are reported in Table 8.1. It is important to have in mind that Bell-Evans approach is only valid for two-state systems and force ranges close to the coexistence region between the two considered states. Since, the folding-unfolding process of the 3WJ is not a two-state process (see previous discussion). The molecule does not directly

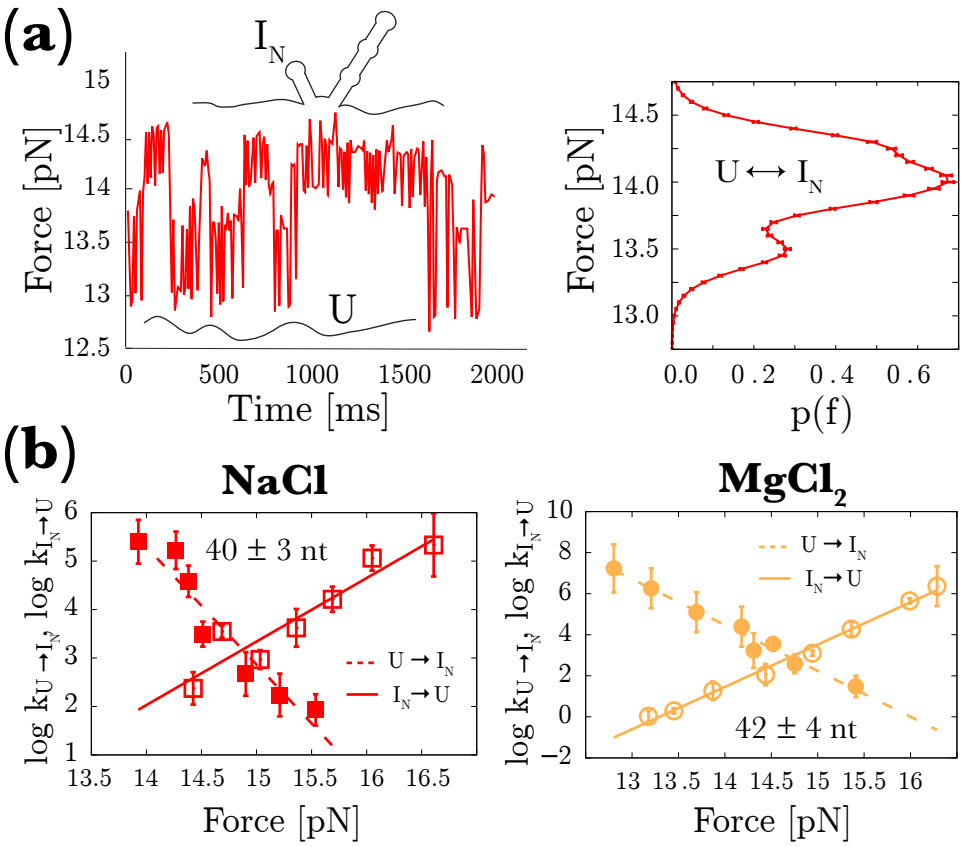

Figure 8.5: **Folding pathway of native 3WJ. (a)** - Example of a hopping trace between the unfolded and the intermediate structure (left). Force distribution function obtained for the hopping trace (right). **(b)** - Hopping kinetics and number of released nucleotides in the unfolded-intermediate transitions for the case of monovalent (left) and divalent (right) salt conditions. Empty symbols correspond to $I_N \leftrightarrow U$ kinetics, whereas full symbols are $U \leftrightarrow I_N$ kinetics.

switch from the unfolded to the folded state. It explores a kinetic intermediate state (I_N) before recovering the native structure. Thus, note that: $x_{N\text{-}TS} + x_{I_N\text{-}N} \neq x_{N\text{-}U}$. Nevertheless, from the force jump in the $N \rightarrow U$ transition (see Fig. 8.4(a)), we find that 73 ± 4 nucleotides are released, which agrees with the total number of nucleotides forming the molecule (77 nucleotides), therefore it is consistent with the unfolding of the native 3WJ. In what follows we perform a deeper analysis of the folding pathway of the 3WJ in order to obtain the maximum amount of information about the kinetic intermediate state I_N.

We studied the I_N state by performing hopping experiments in the region where $U \leftrightarrow I_N$ transitions take place (see Fig. 8.4(a)). An example of a hopping trace is shown in left panel of Fig. 8.5(a). The force

	$x_{I_N\text{-TS}}$ [nm]	$x_{\text{TS-U}}$ [nm]
NaCl	6.3 ± 1.3	11.0 ± 1.0
$MgCl_2$	9.8 ± 0.5	9.1 ± 0.5

Table 8.2: **Bell-Evans analysis for the intermediate-unfolded transition.** Parameters are obtained by linear fitting Eqs. (5.5) and (5.6) to the data shown in Fig. 8.5(b). Results obtained by averaging the individual obtained parameters of four different pulling speeds (50, 100, 200 and 500 nm/s) for six different molecules.

distribution has a two-state structure (right panel), indicating that the molecule only jumps between two well-defined states: the U and I_N states. Kinetic rates in hopping experiments are obtained by measuring the inverse average lifetime of each state. By measuring several forces, the force profile of $k_{U\to I_N}$ and $k_{I_N\to U}$ is obtained, as we show in Fig. 8.5(b). Kinetic rates are modelled according to the Bell-Evans scheme, so that the distances from the I state to the U state can be directly obtained.

In Table 8.2 we report the obtained results for the transition state distances between the I_N and U states. These results allow us to estimate the number of released (or absorbed) nucleotides in the $U \leftrightarrow I_N$ transitions, obtaining 40 ± 3 and 42 ± 4 for the case of monovalent and divalent conditions, respectively. Then, recalling the number of released nucleotides in the $N \to U$ transition is equal to 73 ± 4 nucleotides, we infer that in the $I_N \to N$ transition, 31 ± 3 nucleotides are absorbed. This result is is in very good agreement with the value predicted by the FEL (Fig. 8.4(b)). The scheme in Fig. 8.6 summarizes the unfolding-folding pathway of the native 3WJ molecule.

Finally, the folding free energy of the intermediate structure has been obtained using the Continous Effective Barrier Analysis (CEBA). While Bell-Evans theory considers a single kinetic barrier whose position does not depend on the mechanical force, the CEBA method is able to reproduce a realistic behavior of the force-dependent kinetic barrier [93]. Briefly, within CEBA framework, the kinetic rates are written as:

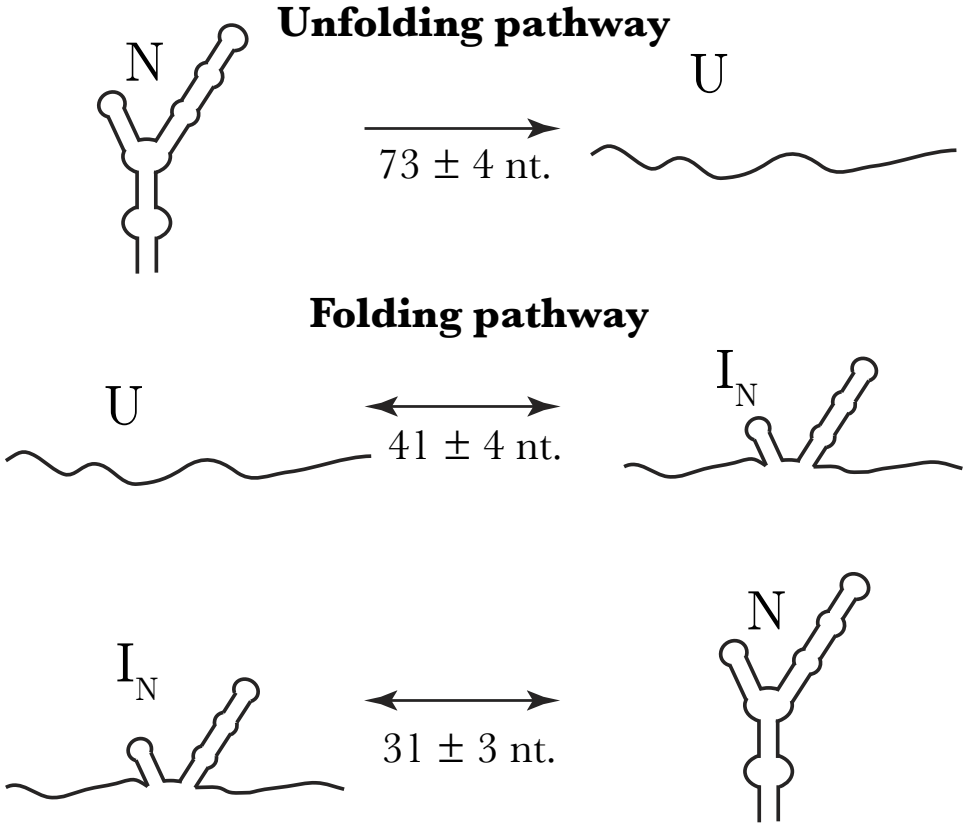

Figure 8.6: **Unfolding-folding pathway of the 3WJ native structure.** The folded native structure switches to the unfolded state through no intermediate state (U → N transitions are not observed due to the height of the kinetic barrier), whereas the folding process occurs through a kinetic intermediate state (I_N). Number of released nucleotides are indicated in each scheme.

$$k_{I_N \to U}(f) = k_0 \exp\left(-\frac{B_{\text{eff}}(f)}{k_B T}\right), \tag{8.1}$$

$$k_{U \to I_N}(f) = k_0 \exp\left(-\frac{B_{\text{eff}}(f) - \Delta G_{UI_N}(f)}{k_B T}\right). \tag{8.2}$$

Here, k_0 is the kinetic rate at zero force, $B_{\text{eff}}(f)$ is the force-dependent kinetic barrier and $\Delta G_{UI_N}(f)$ is the free energy difference between the I_N and U state at a force f. We notice that the term $\Delta G_{UI_N}(f)$ contains several energetic contributions due to the unfolded hairpin in the U

state and the diameter contribution in the I_N state (both calculated in the ForceEns). Then:

$$\Delta G_{UI_N}(f) = \Delta G^0_{UI_N} - \int_0^f x_U(f')\,df' + \int_0^f x_d(f')\,df', \qquad (8.3)$$

where $\Delta G^0_{UI_N}$ is the free energy difference between I_N and U states at zero force, $x_U(f)$ is the extension of the unfolded hairpin at a force f (obtained as the inverse function of Eq. (B.11)) and the term $x_d(f)$ is the extension of two serially connected hairpin diameters forming the I_N state (Eq. (B.5)).

We note that Eq. (8.1) can be written as:

$$\frac{B_{eff}(f)}{k_B T} = \log k_0 - \log k_{I_N \to U}(f)\,, \qquad (8.4)$$

whereas from Eq. (8.1) we have:

$$\frac{B_{eff}(f)}{k_B T} = \log k_0 - \log k_{U \to I_N}(f) + \frac{\Delta G_{UI_N}(f)}{k_B T}\,. \qquad (8.5)$$

Hence, from $k_{I_N \to U}(f)$ (Fig. 8.5(b)) we can obtain: $-\log k_{I_N \to U}(f) = B_{eff}(f)/k_B T - \log k_0$. Finally, by imposing analytical continuity of Eqs. (8.4) and (8.5) we can estimate $\Delta G^0_{UI_N}$. That is, imposing that the two kinetic barriers (Eqs. (8.4) and (8.5)) follow the same curve.

In Fig. 8.7 we show the results for experiments performed in monovalent conditions (left panel) and in divalent salt conditions (right panel). Remarkably, we find that magnesium stabilizes the secondary structure by increasing the folding free energy of the intermediate structure by almost $10\,k_B T$.

We stress that Bell-Evans analysis was not suitable for obtaining thermodynamic parameters of the native structure since folded-unfolded transitions do not behave as a two-state process. Nevertheless, the analysis we performed allowed us to gain some insights on the effects of magnesium in the native structure, pointing towards the stabilization of the native structure due to specific binding.

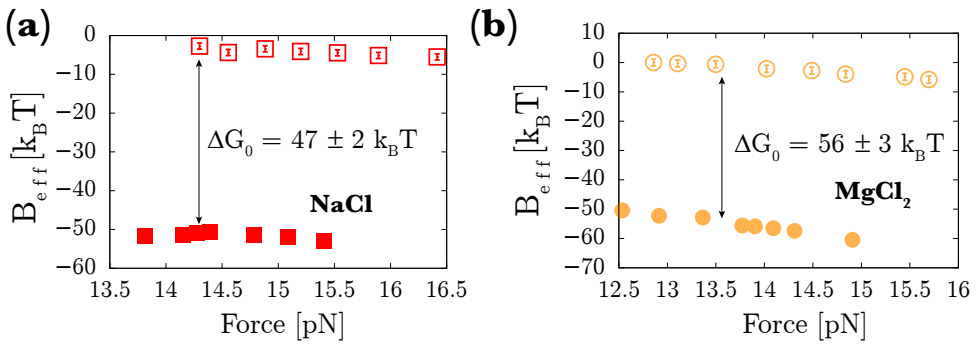

Figure 8.7: **Force-dependent kinetic barrier and measurement of ΔG^0_{UI}.** Analytical continuation of Eqs. (8.4) and (8.5). Empty symbols correspond to the experimental values of: $-\log k_{I_N\to U}(f) = B_{eff}(f)/k_BT - \log k_0$ while full symbols are obtained by: $-\log k_{U\to I_N}(f) - \int_0^f x_U(f')\,df' + \int_0^f x_d(f')\,df' = B_{eff}(f)/k_BT - \log k_0 - \Delta G^0_{UI_N}$. Values of $\Delta G^0_{UI_N}$ are shown as insets for monovalent conditions (left panel) and divalent conditions (right panel).

8.3.2 *Misfolded structure*

Whereas in the preceding section we have only focused in the folding-unfolding pathways of trajectories departing from the native structure, in the present section we will repeat the same analysis but only considering the trajectories involving the **misfolded structure**.

Misfolding in the 3WJ RNA is a force-induced effect, consequence of the competition between the formation of two smaller hairpins that cannot coexist in the same conformation. Previous studies proposed the misfolded state has the structure shown in the right panel of Fig. 8.2, where H_1^M and H_2^M are the two non-native hairpins [148]. Along the unfolding pathway of the misfolded structure we observe two distinct unfolding patterns depending whether H_1^M or H_2^M unfolds (see Fig. 8.8(a)). The different unfolding patterns can be understood in terms of the FEL of both hairpins (see Fig. 8.8(b)). At the same force, while the FEL of H_1^M is pretty flat, the one corresponding to H_2^M has a more abrupt profile. This fact allows us to foresee a smoother unfolding of H_1^M as compared to the unfolding of H_2^M.

For the sake of clarity, we divide the following discussion depending whether we study the unfolding of H_1^M or H_2^M.

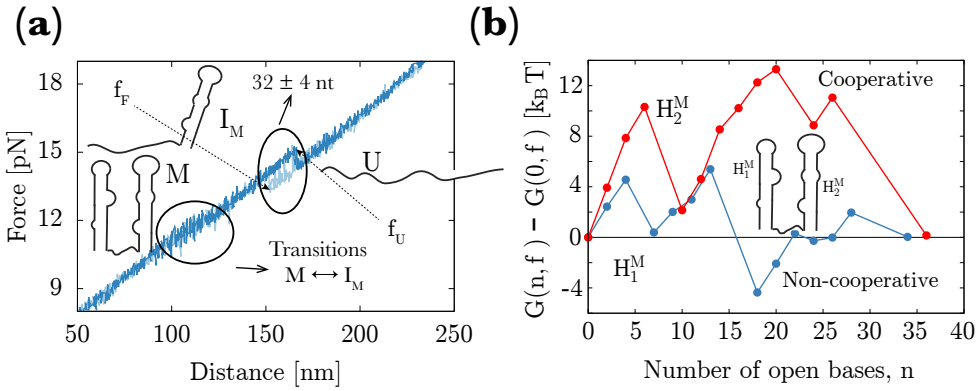

Figure 8.8: **Folding pathway of misfolded structure of 3WJ. (a)** - Example of an unfolding-folding curve (dark and light curve, respectively) and schematic depiction of the structures participating in the unfolding-folding pathway. **(b)** - FEL of the hairpins forming the misfolded structure. FEL calculated at coexistence forces (11.6 pN for H_1^M and 11.4 pN for H_2^M). Note the higher barrier for the H_2^M hairpin.

Cooperative unfolding of H_2^M hairpin

In Fig. 8.9 we show the rupture forces distributions obtained for the cooperative unfolding of the misfolded structure. Bell-Evans kinetic rates are shown as insets and the results are reported in Table 8.3.

Interestingly, transition state distances and free energies do not change with the presence of magnesium. Hence, the misfolded structure is not affected by the presence of magnesium ions. This fact is in contrast with the results that we found for the native structure (Fig. 8.4(c,d)), where we noticed that the presence of magnesium increases the average unfolding forces. Also, coexistence forces are equal with and without magnesium, and are compatible with the experimental observations (see circles in Fig. 8.9(a) and (b)). On the other hand, we find that the observed extension jump in the cooperative unfolding is consistent with a release of 32 ± 4 bases. Since the size of the H_2^M hairpin is 34 nucleotides, our findings are consistent with the prediction that the H_2^M hairpin unfolds all at once (see Fig. 8.8(a)). Thus, the situation in which H_1^M hairpin is unfolded but the hairpin H_2^M is still formed acts as an intermediate state (I_M) between the M and U states.

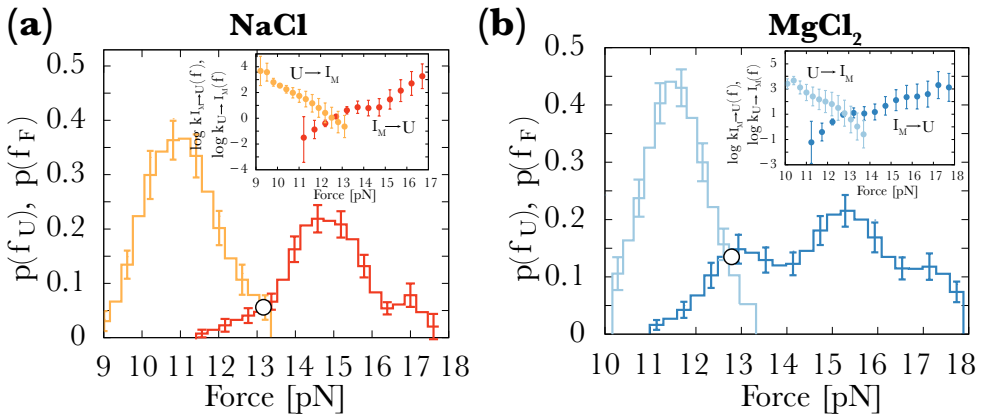

Figure 8.9: **First rupture force distributions and kinetics for the cooperative unfolding of H_2^M hairpin. (a)** - Distributions of first rupture forces and Bell-Evans kinetics (inset) in monovalent conditions. **(b)** - Distributions of first rupture forces and Bell-Evans kinetics (inset) in divalent conditions. In both graphs dark curves correspond to unfolding forces and light curves correspond to folding forces. Errors were computed using the Bootstrap method. Graphs correspond to a pulling speed equal to 200 nm/s. The circles highlight the crossing point of the distributions (coexistence force).

	x_{I_M-U} [nm]	x_{U-I_M} [nm]	ΔG_{I_MU} [k_BT]	f_c [pN]
NaCl	3.3 ± 0.5	4.7 ± 0.2	24.6 ± 1.6	12.6 ± 0.4
$MgCl_2$	3.4 ± 0.2	4.7 ± 0.1	26 ± 1	13.1 ± 0.6

Table 8.3: **Bell-Evans analysis for the misfolded structure in the folding-unfolding transition.** Parameters are obtained by linear fitting Eqs. (5.5) and (5.6) to the data shown in Fig. 8.9. Results obtained by averaging the individual obtained parameters of four different pulling speeds (50, 100, 200 and 500 nm/s) for six different molecules.

Non-cooperative unfolding of H_1^M hairpin

Next, the non-cooperative unfolding of the rest of the molecule can be studied by performing equilibrium hopping experiments in the region where the transition takes place. Non-cooperative unfolding (i.e. $I_M \leftrightarrow$ M transitions of Fig. 8.8(a)) are consistent with the unfolding of H_1^M for two reasons. First, when computing the number of released nucleotides in the $I_M \leftrightarrow$ U transition (see previous discussion), is compatible with

the whole unfolding of H_2^M. Moreover, the FEL of H_1^M in the observed force range (i.e. 11.5 - 12 pN) is pretty flat, with an intermediate located at $n \simeq 17 - 18$ bases, as can be seen in Fig. 8.8(b). Hence, thermal fluctuations make feasible the equilibrium non-cooperative transition between the folded and unfolded state of H_1^M.

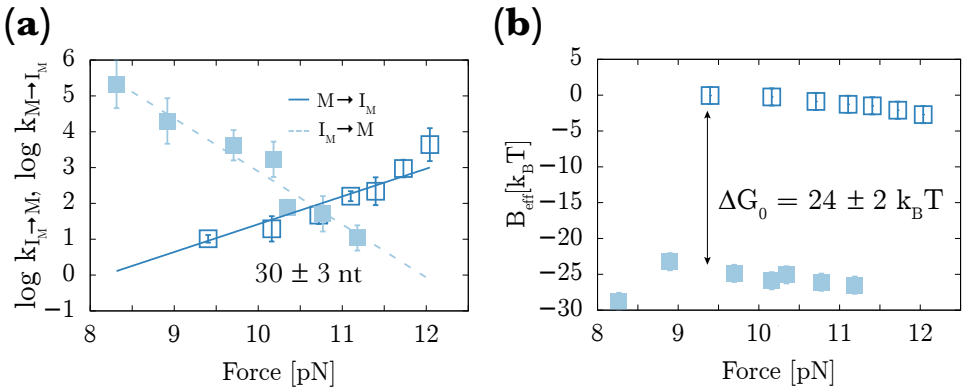

Figure 8.10: **Hopping kinetics and effective barrier for folded-intermediate transitions of misfolded structure in monovalent conditions (folding-unfolding of H_1^M hairpin).** **(a)** - Unfolding kinetics (empty symbols) from the misfolded state to the intermediate state (where the hairpin H_1^M is unfolded) and folding kinetics (full symbols). Number of released/absorbed nucleotides in the M $\leftrightarrow$ I_M transition is shown as inset. **(b)** - Continuous effective barrier analysis for the M $\leftrightarrow$ I_M transition. Colour criteria is the same as in Fig. 8.7. Free energy of formation of I_M state is shown in the graph.

In Fig. 8.10(a) we show the hopping kinetics measured in the M $\leftrightarrow$ I_M in monovalent conditions. By analysing the transition state distances we obtain that the non-cooperative transition is compatible with a release/absorption of 30 ± 3 bases, reinforcing the hypothesis that the opening of H_1^M hairpin acts as an intermediate state between the fully unfolded 3WJ molecule and the misfolded state. On the other hand, we applied once again the CEBA, as in the previous section, in order to extract the free energy difference (or free energy of formation) between the misfolded state and the I state, giving $\Delta G_{MI_M} = 24 \pm 2\ k_B T$. This latter value is compatible with the free energy of formation of hairpin H_1^M predicted by Mfold, which is equal to $\Delta G_{H_1^M} = 24.3\ k_B T$.

Interestingly, in a buffer containing magnesium we are not able to extract the kinetics of the I_M $\leftrightarrow$ M transition. Only in divalent conditions, after some time measuring hopping traces we observe a

sudden force increase. After this event, hopping does not longer occur. We believe that this phenomenon is due to the rescue of the native 3WJ structure by the magnesium ions. We stress that the rescue takes place in time scales shorter than those we need in order to accurately measure hopping transitions. In the following section we delve into this guess by performing further experimental assays.

It is worth mentioning that the kinetic analysis done by Bell-Evans approach (Table 8.3) plus the CEBA method (result in graph of Fig. 8.10(b)) reports a free energy of formation of the misfolded structure equal to: $\Delta G^0_{\mathrm{MU}} = 49 \pm 3\, k_{\mathrm{B}}T$, whereas the Mfold prediction is equal to $50\, k_{\mathrm{B}}T$. Hence, our experimental findings are in good agreement with the theoretical prediction.

8.3.3 *Mg^{2+} rescue experiments*

Upong adding magnesium, we could not observe hopping kinetics between the intermediate and the misfolded state. Indeed, when performing hopping experiments to characterize $\mathrm{I_M} \leftrightarrow \mathrm{M}$ transition (i.e. for the misfolded structure), after a certain time (the so-called rescue time), a sudden force jump is observed and $\mathrm{I_M} \leftrightarrow \mathrm{M}$ transitions are not anymore observed. The sudden force range is associated with the formation of the native structure (see below) and the process is called rescue.

In divalent conditions, for usual force ranges ($\sim$ 10 pN), the rescue time is as small as 2 seconds, short enough to avoid obtaining accurate kinetic measurements by means of hopping experiments. However, rescue experiments might provide complementary insights on the folding procedure of misfolded structure, so we designed a new experimental procedure. We point out that in monovalent conditions this phenomenon also takes place. Nevertheless, in monovalent conditions the rescue time can be around 5 minutes, long enough to obtain accurate hopping measurements.

Rescue experiments are composed by the following steps. First, when refolding the misfolded structure, the position (i.e. the relative distance between the optical trap and the tip of the micropipette, see Fig. 8.3(a)) is kept fixed at a value in which the molecule is in the intermediate state prior to form the misfolded structure. Then, the time-evolution of the force is monitored, as in a typical hopping experiment, until a

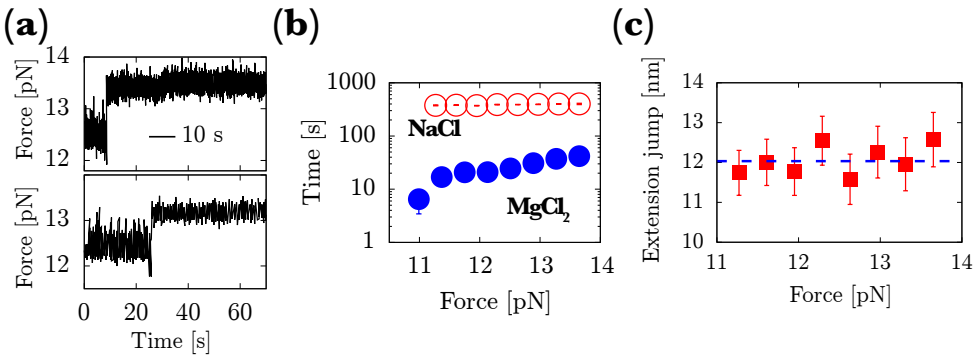

Figure 8.11: **Rescue experiments in $MgCl_2$. (a)** - Examples of two typical force-time trace . **(b)** - Evolution of the rescue time as a function of the force. Empty symbols correspond to monovalent conditions whereas full symbols correspond to divalent conditions. **(c)** - Absorbed extension in the rescue events as a function of the force. Error bars are the standard error after averaging the results of six different molecules in divalent conditions.

force jump is observed. This protocol is repeated for different values of the trap-micropipette distances (and hence, forces). In Fig. 8.11(a) we show typical force-time traces upon performing rescue experiments. The rescue time is stochastic, but with a slight tendency to increase as the force increases, as can be seen in Fig. 8.11(b) (empty symbols correspond to monovalent conditions and full symbols correspond to divalent conditions). The force jump can be related to an absorption of a certain molecular extension. Hence, according to the scheme shown in Fig. 8.3(a), the distance (hereafter denoted as x) can be decomposed as:

$$x(f) = x_{\mathrm{b}}(f) + x_{\mathrm{handles}}(f) + x_{\mathrm{mol}}(f)\,, \tag{8.6}$$

where $x_{\mathrm{b}}(f)$ is the distance from the center of the bead to the center of the optical trap at a force f, $x_{\mathrm{handles}}(f)$ is the extension of the handles and $x_{\mathrm{mol}}(f)$ is the molecular extension. Hence, by considering that the rescue happens at constant x, from Eq. (8.6) we can obtain absorbed molecular extension Δx_{mol} as:

$$\Delta x_{\mathrm{mol}} = -(\Delta x_{\mathrm{b}} + \Delta x_{\mathrm{handles}})\,, \tag{8.7}$$

where Δx_{b} and $\Delta x_{\mathrm{handles}}$ denote the difference of $x_{\mathrm{b}}(f)$ and $x_{\mathrm{handles}}(f)$ after the force jump. In Fig. 8.11(c) we show the extension jump calcu-

lated according to the previous scheme as a function of the force prior to the jump. We note that the absorbed extension is equal to 12 ± 1 nm and it is nearly independent of the mechanical force (in the range [11 - 14 pN]), yielding a value of 25 ± 3 nucleotides absorbed. This value is compatible with a configuration in which the H_2^M hairpin is completely formed whereas the H_1^M hairpin is only formed by the six bases prior to the GAAA loop. After the force jump, the remaining 26 nucleotides are absorbed and the native structure is rescued. This situation is depicted in Fig. 8.12.

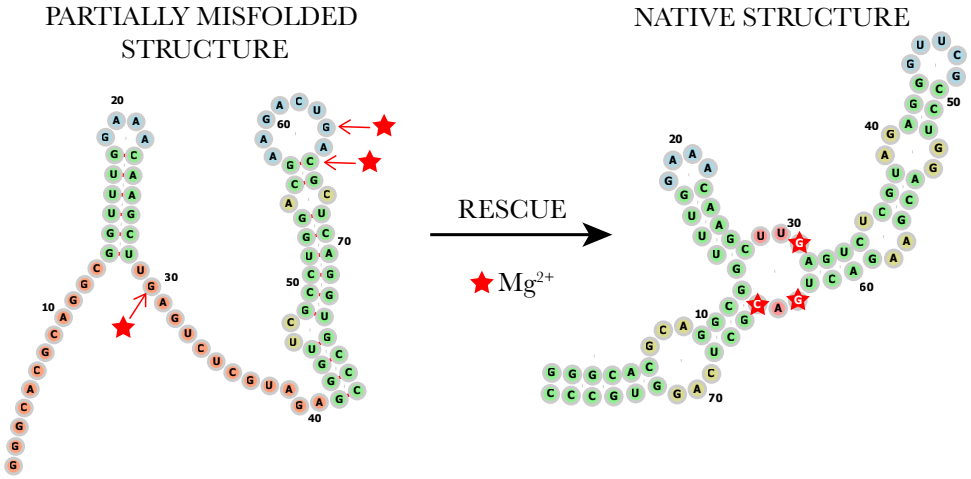

Figure 8.12: **Proposed partially misfolded structure and rescue of native structure in $MgCl_2$.** The partially misfolded structure has 26 free bases. Binding sites of magnesium cations are marked with stars. After magnesium binding, the native structure is rescued.

Magnesium ions can specifically recognise the negatively charged hydroxyl group of the bases so they can bring together distant RNA nucleotides in order to form stable tertiary structures. We believe that this is, in fact, the effect that triggers the rescue of the native structure in the presence of magnesium. Since the misfolded structure does not contain the specific binding motif of magnesium (see Fig. 8.1), the native structure cannot be rescued. However, in the folding pathway, the partially misfolded structure shown in left Fig. 8.12 can be eventually formed, so magnesium is able to bind to some specific locations (red stars) [145]. Then, via electrostatic interactions, the native structure can be again restored.

8.4 DETERMINATION OF THE SPECIFIC BINDING ENERGY OF Mg^{2+}

8.4.1 *Free energy determination of kinetic states*

An important quantity when studying ligand interactions is the binding strength. This topic is particularly interesting when studying RNA folding. Indeed, due to the strong negativity of RNA molecules, the interaction with positive-charged ligands is crucial for correctly folding RNA molecules (i.e. in their native structure). In particular, in physiologic conditions, the concentration of free Mg^{2+} is of the order of 1 mM [151] and, as we demonstrated in the preceding section, magnesium cations are essential for the correct folding of 3WJ RNA. Indeed, magnesium ions stabilize the native structure and they are able to rescue the native structure from misfolded structures. The next question is immediate: how much is the specific binding energy of Mg^{2+} to the 3WJ?

The relation between Mg^{2+} (and other ions) and RNA folding is still a hot topic in the field [152, 153] and, moreover, a precise quantification of the binding energy of Mg^{2+} with RNA is still under the spotlight [154]. We have found that the specific binding energy of Mg^{2+} to the 3WJ is measurable by means of fluctuation relations. We have used the CFT throughout the thesis in order to either determine folding free energy differences in different statistical ensembles or to perform information-content measurements. In both frameworks, the CFT is the suitable framework since the molecules we used have well-established equilibrium states. This is not the case of the 3WJ RNA molecule, where misfolded state is not an equilibrium state. Hence, the traditional CFT (Eq. (4.4)) will not provide us the correct free energy. In order to take into account non native states, the Extended Crooks Fluctuation Theorem (ECFT) must be used instead [52]. In particular, the ECFT reads as[3]:

$$\frac{\phi_F^{A\to B}}{\phi_R^{B\to A}} \frac{P_F^{A\to B}(W)}{P_R^{B\to A}(-W)} = \exp\left(\frac{W - \Delta G_{AB}}{k_B T}\right), \tag{8.8}$$

where A and B denote two kinetic states (i.e. partially equilibrated), $\Delta G_{AB} = G_B(\lambda_1) - G_A(\lambda_0)$ is the free energy difference between A

3 As always, λ denotes the control parameter of the experiment.

state at λ_0 and B state at λ_1, $P_F^{A\to B}(W)$ and $P_R^{B\to A}(-W)$ are the partial forward and reversed work distributions (i.e. the work distributions for the processes that start at A and end at B), respectively. Finally, $\phi_F^{A\to B}$ and $\phi_R^{B\to A}$ are the fraction of trajectories starting in A state (or B) at λ_0 (or λ_1) and ending in B state (or A) at λ_1 (or λ_0). The use of the ECFT, rather than the CFT, is crucial for the correct determination of the free energy differences of non-native states. Indeed, the free energy appearing in Eq. (8.8) is related to the free energy difference appearing in Eq. (4.4) (i.e. ΔG from Eq. (4.3)) via:

$$\Delta G_{AB} = \Delta G - k_B T \log \frac{\phi_F^{A\to B}}{\phi_R^{B\to A}} . \tag{8.9}$$

Which, in some cases, is a significant correction to ΔG. Either we start at the native or misfolded state, all trajectories end in the unfolded state, so: $\phi_F^{N,M\to U} = 1$. Then, the overlooking of misfolding probability (i.e. the fraction of trajectories starting in the unfolded conformation ending in the misfolded state: $\phi_R^{U\to M}$) would lead to a remarkable free energy underestimation. In fact, the misfolding probability we observe is around 5 - 10%, for low and high pulling speeds, respectively, leading to a free energy underestimation of 3 $k_B T$. We note that the misfolding probability we observe is compatible with the model developed in Ref. [148].

First, the folding free energy difference between the native and the unfolded state, ΔG_{NU}, has been determined by performing nonequilibrium pulling experiments (as described in Sec. 6.3) and using the ECFT (Eq. (8.8)). As usual, the unfolding process is identified with the forward protocol (F), whereas the folding process is identified with the reversed protocol (R). Since the control parameter is the relative distance trap-micropipette (x in Eq. (8.6)), the work W is calculated according to the ExtEns scheme (Eq. (4.6)). We stress that, even though $\phi_F^{N\to U} = 1$, $\phi_R^{U\to N} \neq 1$. Not all of the refolding trajectories end up in the native state. Nevertheless, the classification of the trajectories is straightforward from the pattern of the FDCs (see Fig. 8.3). In Fig. 8.13(a) we show the partial work distributions obtained after classifying the trajectories involving the native structure for the case of experiments done in monovalent conditions, whereas Fig. 8.13(b) show the same information obtained in divalent conditions.

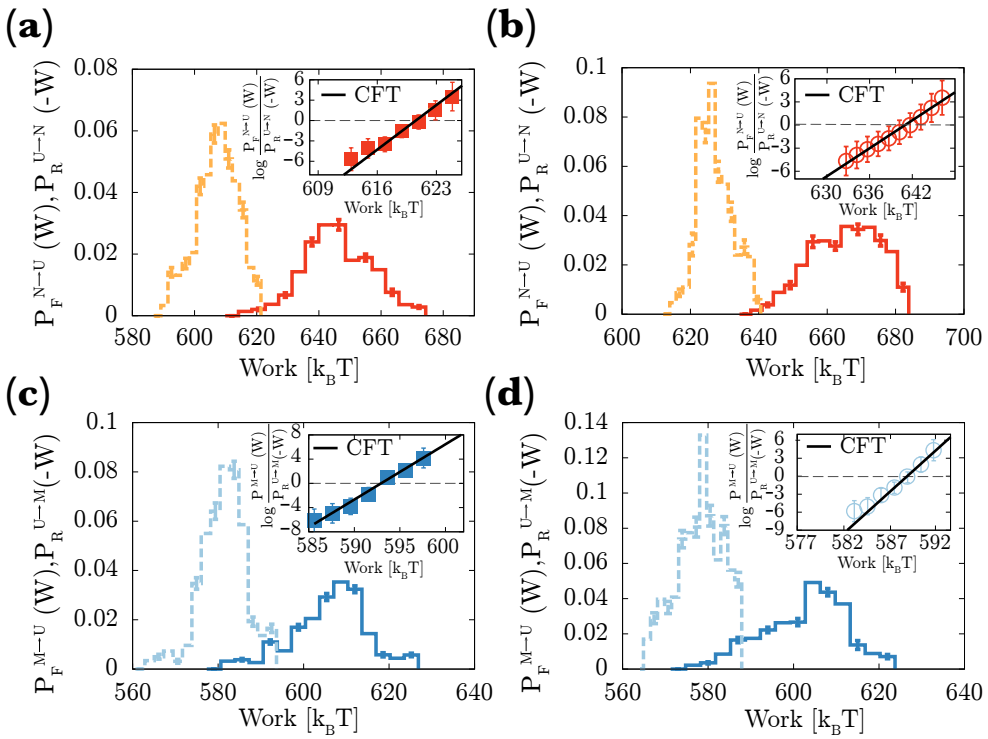

Figure 8.13: **ECFT applied to the native (top panels) and misfolded (bottom panels) structure of 3WJ RNA.** Results of the partial work distributions obtained in monovalent salt conditions (left panels) and divalent salt conditions (right panels) for a pulling speed equal to 50 nm/s. The integration range is the same for all conditions. Forward distributions are plotted as solid lines whereas reversed distributions are plotted as dashed lines. Top panels (**(a), (b)**) correspond to partial work distributions for trajectories starting in the native state ending in the unfolded state, whereas bottom panels (**(c), (d)**) correspond to partial work distributions for trajectories starting in the misfolded state ending in the unfolded state. The ECFT verification is shown as inset in both graphs, were solid line represents a straight line with slope 1 and y-intercept equal to $-\Delta G_{XU}$, both in $k_B T$ units, being $X = N, M$. In all cases, N stands for native, M for misfolded and U for unfolded. Error bars are obtained using Bootstrap method.

We emphasize the validity of the ECFT in both cases as insets of Fig. 8.13, where we have plotted the logarithm of the ratio of the F and R work distributions plus the logarithm of the term $\frac{\phi_F^{A\to B}}{\phi_R^{B\to A}}$ as a function of the work in $k_B T$ units. As predicted by Eq. (8.8), the slope is 1 for both salt conditions (solid lines in both insets of Fig. 8.13).

The same procedure has been repeated in order to obtain the free energy difference between the misfolded and the unfolded state, ΔG_{MU}. Again, we must take into account the fact that $\phi_F^{M\to U} = 1, \phi_R^{U\to M} \neq 1$. In Fig. 8.13(c) we show the partial work distributions obtained after classifying the trajectories involving the misfolded structure for the case of monovalent salt conditions and divalent ionic conditions (Fig. 8.13(d)). Again, the ECFT verification is shown as inset in both panels.

The energetic contributions inherent to the experimental setup (i.e. displacement of the bead in the optical trap, stretching of the handles and the released ssRNA) are subtracted to the free energy determined by means of the ECFT, as we describe in Appendix C. We point up that the obtained free energies are measured with respect to the energy of the random coil state at zero force. In Table 8.4 we report the measured values for ΔG^0_{XU}, being $X = N$ (native) or $X = M$ (misfolded).

		ΔG_{XU} [k_BT]	$\Delta W^{rev.}_{stret.}$ [k_BT]	$\Delta W^{rev.}_{handles\ +\ bead}$ [k_BT]	ΔG^0_{XU} [k_BT]
X = N	NaCl	615 ± 4	21 ± 2	524 ± 4	70 ± 4
	$MgCl_2$	637 ± 4	21 ± 2	531 ± 3	87 ± 4
X = M	NaCl	595 ± 4	21 ± 2	520 ± 4	54 ± 4
	$MgCl_2$	588 ± 4	21 ± 2	515 ± 4	52 ± 4

Table 8.4: **Experimental measurement of ΔG^0_{NU} and ΔG^0_{MU}**. The values for ΔG_{XU} are obtained using the ECFT. Error bars contain statistical and systematic errors. Experiments were performed at four different pulling speeds (50, 100, 200 and 500 nm/s) for six different molecules.

We find that the folding free energy of the misfolded structure does not change in monovalent or divalent salt conditions. This result is due to two key factors. First and foremost, our experiments are performed in equivalent monovalent/divalent salt conditions. As a matter of fact, our findings serve as another experimental validation of the empirical 100/1 rule regarding the non-specific contribution of monovalent and divalent salt to the free energy of formation of RNA molecules. Finally, since the misfolded structure does not contain the minimal binding site of magnesium, there is no contribution of specific binding involving magnesium ions.

On the other hand, we find that the folding free energy of the native structure is considerably higher upon adding magnesium. Since experiments are performed in equivalent salt conditions, the difference found in the free energy of formation is, undoubtedly, due to the specific

binding of three divalent cations to the native structure. Moreover, the difference of both folding free energies: $\Delta\Delta G_{\mathrm{NU}} = \Delta G_{\mathrm{NU}}^{\mathrm{MgCl_2}} - \Delta G_{\mathrm{NU}}^{\mathrm{NaCl}}$ quantifies the specific binding strength of three Mg^{2+} to the 3WJ structure. We obtain: $\Delta\Delta G_{\mathrm{NU}} = 17 \pm 5\, k_{\mathrm{B}}T$ and, hence, a binding strength per ion of $6 \pm 2\ k_{\mathrm{B}}T$.

It is important to emphasize that the free energy obtained using the ECFT is consistent with the folding free energy predicted by Mfold, which is equal to $\Delta G_{\mathrm{NU}}^{0} = 67\, k_{\mathrm{B}}T$ for the native structure and $\Delta G_{\mathrm{MU}}^{0} = 50\, k_{\mathrm{B}}T$ for the misfolded structure.

8.5 CONCLUSIONS

Summarizing, in this chapter we have explored the thermodynamic and kinetic behavior of a three-helix RNA junction molecule and we provided a direct measurement of the specific contribution of magnesium ions binding to the tertiary RNA structure. In particular, we studied the highly-conserved site of the 16S rRNA. Such RNA fragment has a three-way junction structure acting as the binding site of S15 protein. The interaction between the 3WJ RNA and S15 is promoted by the presence of magnesium cations in the solute. Moreover, magnesium stabilizes the native structure by inducing a conformational change in the molecule. Also, previous force-spectroscopy assays showed that 3WJ RNA molecule is able to explore a force-induced state.

Even though the secondary structure of the native and misfolded structure was already characterized by means of DFS, there was a gap in the study of folding and unfolding pathways. We have performed a detailed DFS characterization in order to determine the kinetics of the unfolding-folding pathways of the native and misfolded structure in equivalent monovalent and divalent salt conditions (according to the 100/1 rule). In particular, we unravelled the kinetic structures acting as an intermediate states when mechanically unfolding and refolding the molecule. Such studies allowed us to gain some insights on the role of magnesium ions on the 3WJ RNA molecule. In fact, we discovered that magnesium is not only able to stabilize the native structure, but also to rescue the native structure from the misfolded state in measurable time scales (on the order of a few seconds).

Moreover, we have used the ECFT to measure the free energy of formation of the native and the misfolded structure of the 3WJ RNA. The

results we found investigating the thermodynamics of both conformations have remarkable implications. First and foremost, we confirmed, at the single-molecule precision, that magnesium is not able to bind to the misfolded RNA since the binding domain does not exist in the misfolded structure. Finally, thanks to having tested the 100/1 rule, we have been able to perform a direct quantification of the specific energy of magnesium ions binding to the 3WJ molecule, which we found to be $6 \pm 2\ k_B T$ per magnesium cation, unprecedented result in single-molecule assays.

Part V

FINAL CONCLUSIONS

FINAL CONCLUSIONS

The development of quantum theory and relativity at the beginning of the XXth Century shook the world of physics. The advent of quantum mechanics and relativity caused a paradigm shift in physics. As a consequence of this revolution, the scope of modern physics became significantly different. Physics grown into a science capable of promoting technological developments. Among many others, for instance, the invention of the laser turned out to be a milestone for modern and contemporary physics. One of the most relevant and influential applications of the laser is the development of optical tweezers by Arthur Ashkin in 1970. For this invention Arthur Ashkin was awarded the 2018 Nobel Prize in Physics. Thanks to Ashkin's invention, nowadays, physics can cope with new problems, which were unimaginable few decades ago. For instance, the field of biophysics experienced a breakthrough due to the development of single-molecule instruments capable of exerting forces on individual molecules.

Single-molecule experiments have emerged as a powerful tool that allow researchers to investigate the physical behavior of individual molecules with unprecedented resolution. The feasibility of exerting forces at the piconewton scale (10^{-12} N) and measuring nanometric displacements in the sub-millisecond scale, offers a widespread range of exciting possibilities. This fact is attractive both from a biological and a physical perspective. On the one hand, from the biological perspective, the possibility of manipulating individual molecules to induce their mechanical denaturation may allow researchers to get insights about the origin —and hopefully the cure— of many diseases. On the other hand, from a physical perspective, the study of individual molecules is also attractive for physicists and chemists. Indeed, energetic exchanges of molecular systems are of the order $\sim$ nm pN $\sim$ kT, that is, of the order of Brownian fluctuations. As a consequence of this fact, most of the quantities that we are able to measure in single-molecule experiments have an inherent stochastic nature. Therefore, single-molecule systems are very attractive to theoretical physicists to test and discovery new physical laws of non-equilibrium processes.

The major part of this thesis is devoted to address fundamental topics of statistical physics using single-molecule experiments. In particular, in Part II, we aimed to study one of the eldest questions in statistical mechanics: the issue of ensemble inequivalence. Statistical physics sets a bridge between the microscopic and the macroscopic behavior —thermodynamics— and its modern conception is based on the Gibbs ensemble theory. Essentially, the approach of statistical mechanics consists in studying the average behavior of the individual elements of a statistical ensemble when some external constraints are imposed to the system. One of the most remarkable —and polemical— result of the ensemble theory is the phenomenon of ensemble equivalence. In general, for a given statistical ensemble it is possible to build a conjugate statistical ensemble by performing a Legendre transform using two conjugate variables with respect to energy (for instance, the pressure p and the volume V). Mathematically, two conjugate ensembles are equivalent in the thermodynamic limit. Nevertheless, this fact is not always true. In macroscopic magnetic systems, it has been experimentally observed that controlling an extensive quantity (like the volume) is not equivalent to controlling an intensive quantity (like the pressure). In our case, by performing single-molecule experiments on a well-known molecule—CD_4 DNA hairpin—, we have been able to explore two conjugate ensembles: the fixed-extension and the fixed-force ensemble. Both ensembles are conjugate with respect to energy since the product force times extension equals has energy dimensions. We carried out experiments in the fixed-force ensemble using both optical tweezers and magnetic tweezers, and in the fixed-extension using optical tweezers. We have found that these two conjugate ensembles are not equivalent at the level of thermodynamics and kinetics. Moreover, we showed that the often-neglected boundary terms in the definition of the thermodynamic work are essential for the validity of the fluctuation theorem. The main consequences of our studies are: first, the possibility of extending free energy recovery methods to statistical ensembles in which only intensive variables can be controlled and, second, the resolution of the controversial question of whether the work definition of the fixed-force is indeed a correct thermodynamic work definition or not. On the other hand, our findings in ensemble inequivalence at the level of molecular kinetics arise interesting questions from a biophysical perspective. For instance, what is the suitable statistical

ensemble of crowded environments like cells? How do molecular reactions behave in different statistical ensembles? All of these questions are an interesting research track to follow.

The second part of this thesis is also merely theoretical. Recent single-molecule assays confirmed the connection between information theory and statistical physics. This historical pursuit sprang forth thanks to the works of Claude E. Shannon, who laid the foundations of information theory, and the works of Edwin T. Jaynes, who spurred the connection between statistical physics and information theory. The development of fluctuation theorems and stochastic thermodynamics have provided a general framework in which the thermodynamics of information naturally appears. Moreover, single-molecule experiments have turned out to be the perfect playground to explore the thermodynamic implications of having —or lacking— information. It is worthwhile to mention the experimental realization of the Szilard engine and the experimental verification of Landauer's limit. With the current existing results, the information-to-energy connection is well established. We have been able to experimentally demonstrate, for the first time, the reversed implication. We have been able to quantify the information-content of neutral molecular ensembles by means of thermodynamic measurements. That is, we experimentally demonstrated the energy-to-information conversion. Our works are built on what we call ensemble force spectroscopy, a systematic procedure capable of obtaining a robust characterization of molecular ensembles by measuring, in the best tradition of statistical physics, just a few tens of molecules. We think that our work paves the way to study the information-content production of systems undergoing evolutionary dynamics. Despite our experimental system was a neutral molecular ensemble without selection forces, the framework is fully general and suitable for studying real evolutionary systems (such as molecular systems under directed molecular evolution). We think that we have before us a large number of exciting questions ready to answer. For instance, how does information evolve in time? Is there a competition between information and energy in evolution?

In the final part of the thesis (Part III) we aimed to measure the specific binding energy of a metallic ion to the tertiary structure of a three-way RNA junction belonging to the central domain of the 16S ribosomal RNA (rRNA). This study has remarkable physical and biological consid-

erations. The central structure of the 16S rRNA is able to bind the small S15 protein, an essential element for assembling the whole ribosome. From the physics perspective, to the best of our knowledge, this is the first time we have been able to discern the free energy contribution due to the specific binding of magnesium ions to an RNA substrate by means of single-molecule assays. On the other hand, such molecule is able to form, besides its native conformation, a force-induced misfolded state. Despite this fact was already pointed out in previous single-molecule studies, there was a lack of knowledge regarding the molecular kinetics and the folding pathway of the three-helix junction. Aiming to fill this gap, we performed a thorough study of the three-helix RNA junction using dynamic force spectroscopy. As a result, we have characterized the full folding pathway of the molecule, including both the native and the misfolded structure. Furthermore, we have experimentally confirmed the fact that the presence of magnesium promotes the stabilization of the native structure and we have measured this contribution. We have found that magnesium is able to rescue the native structure from the misfolded structure via electrostatic interactions due to magnesium binding. This fact is biologically relevant, since we have been able to characterize the conditions in which a misfolded molecule is able to recover its native conformation. We hope that our findings will spur further single-molecule assays in this direction.

Part VI

APPENDIXES

A

MOLECULAR SYNTHESIS OF DNA/RNA CONSTRUCTS

The present appendix briefly summarizes the steps to synthesize the DNA and RNA hairpins we have used throughout the thesis. This includes short DNA hairpins, short randomized DNA hairpins and RNA constructs.

A.1 RANDOMIZED AND NON-RANDOMIZED DNA HAIRPINS WITH SHORT HANDLES

Following, the steps to synthesize short DNA hairpins with short 29-basepairs long dsDNA handles are summarized [71]. We point that the following protocol is valid either for fully-complementary DNA hairpins and hairpins presenting unpaired bases. The handles are the same for every hairpin and their sequence is:

5' – AGTTAGTGGTGGAAACACAGTGCCAGCGC – 3'

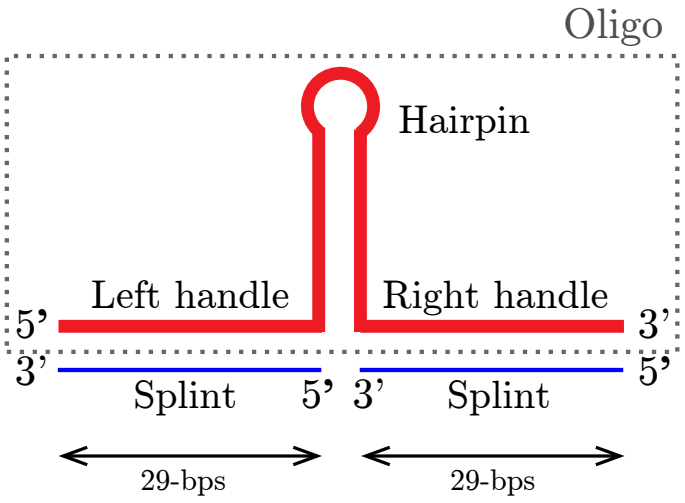

Figure A.1: **Sketch of the molecular construct composed by a DNA hairpin linked to two handles.** Dashed box indicates the oligonucleotide we use to attach the splints.

The oligonucleotide that is able to pair up with each handle is the so-called "splint". Its sequence is the following one:

5' – GCGCTGGCACTGTGTTTCCACCACTAACT – 3'

Note that the sequence of the splint is fully complementary (from 3' to 5') to the sequence of the handles.

DNA hairpins are synthesized by hybridizing the oligonucleotides shown in Fig. A.1. The first oligonucleotide (grey dashed box) is composed by the two handles plus the hairpin (which contains the suitable sequence for each assay) and the two splint molecules. All of these oligonucleotides are supplied by a specialized company (as Sigma-Aldrich or Invitrogen) and, usually, they are freeze dried.

We must have in mind that in order to allow the binding of the molecular construct (Fig. A.1) with the elements involved in each single-molecule setup, at the 5' end of the left handle contains a biotin, whereas the 3' end of the right handle is modified with a digoxigenin tail. The biotin labelling is indicated when buying the oligonucleotides, but the digoxigenin tailing is done as follows:

Digoxigenin of the 3' end of the oligonucleotide

Using the *Oligonucleotide Tailing Kit* (Roche), a tail of an average of Digoxigenin-dUTP (Dig-dUTP) nucleotides is added to the 3' end of the oligonucleotide. The steps of this process are:

i. Dissolve the supplied oligonucleotide with double distilled water (ddH_2O) until a 100 μM concentration is reached. The required water is specified in the oligonucleotide tube. After that, spin down the oligonucleotide tube.

ii. Mix the following components in a sterile Eppendorf tube:

iii. Incubate for 15 minutes at 37°C.

iv. Purify the mixture using the *Qiaquick Nucleotide Purification Kit* (QUIAGEN). Follow the instructions of the kit.

v. Keep the final sample at -20°C.

8 μl	ddH_2O
1 μl	oligonucleotide 100 μM (100 pmol)
4 μl	$CoCl_2$
4 μl	Reaction Buffer x5
1 μl	dATP
1 μl	Dig-dUTP
1 μl	Terminal transferase (enzyme)
20μl	

Annealing of the whole molecular construct

In what follows, we describe the protocol that allows for the hybridization of the two splint molecules and the oligonucleotide (which has been previously Dig-tailed). At the end of the process, the molecular construct will be available for performing SME.

i. Spin down all the oligonucleotide tubes (oligo and splint).

ii. Mix the following components in a sterile Eppendorf tube:

10 μl	Dig-tailed oligonucleotide (5 pmol)
4 μl	splint (10 pmol)
1 μl	Tris 1 M pH7.5
1 μl	NaCl 5 M
x μl	ddH_2O
1 μl	Dig-dUTP
30μl	

iii. Incubate for 2 hours at 42°C.

iv. Decrease the temperature at a constant rate equal to 1°C/min until room temperature (25°C) is reached.

v. If desired,dialysis of the mixture for 30 min in 50 ml of 10 mM NaCl, 10 mM Tris pH 7.5, 1 mM EDTA. Recover the maximum amount of the mixture.

vi. Keep the DNA at 4°C.

A.2 SYNTHESIS OF THE RNA THREE-HELIX JUNCTION

Now we briefly summarize the steps required to synthesize the RNA three-helix junction (hereafter to referred as 3WJ). A thorough and precise description about all the intermediate steps of the synthesis can be found in Ref. [53].

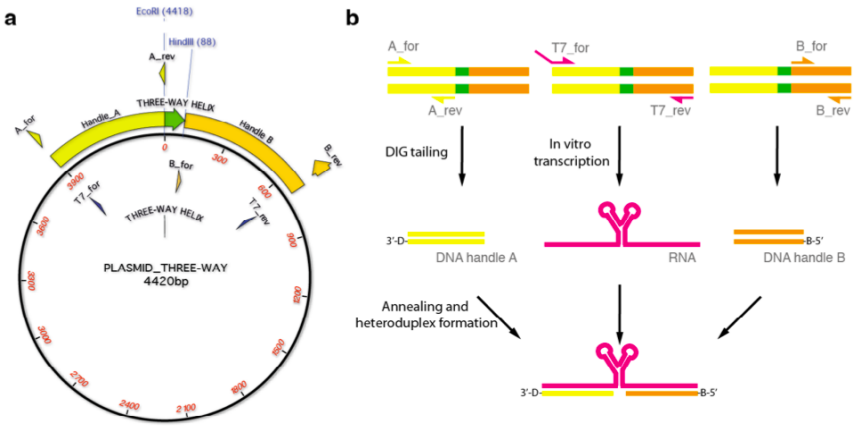

Figure A.2: **Scheme of RNA synthesis. (a)** - Plasmid containing the 3WJ (green) and the required flanking regions (yellow and orange). **(b)** - PCR amplification (top) and tethering of the handles (central) for forming the final construct (bottom). Figure obtained from Ref. [53].

We used pBR322 plasmid with 16S-3WJ sequence inserted between EcoRI and HindIII restriction sites (see Fig. A.2). The plasmid was purchased at Eurofins. After extracting the region of interest from the plasmid, we performed a PCR amplification in order to obtain a template for the in vitro transcription that contains, besides the 3WJ sequence, 500 extra bases at each end, which will be used to form the handles needed to connect the molecules to the beads used in the LOT experiments.

After performing the in vitro transcription, biotin or digoxigenin labelled complementary handles to the template are hybridized to the single-stranded RNA (ssRNA) molecule at each end.

B

ELASTIC MODELS OF LINEAR POLIMERS

B.1 FREELY JOINTED CHAIN (FJC) MODEL

The Freely Jointed Chain (FJC) model is a simplified model for the structural properties of a linear polymer. In this model, the polymer is assumed to be formed by N rigid linear monomers of length b (the so-called Kuhn length). Moreover, the N monomers are joined together at their ends by freely rotating hinges. As the joints can freely rotate, we do not consider excluded volume interactions.

Now, let us consider that a FJC polymer that is kept fixed by one of its ends while the remaining end is subjected to an external force f. The external force creates an effective potential energy equal to $-fx$, being x be the end-to-end distance of the polymer(see Fig. B.1(a)).

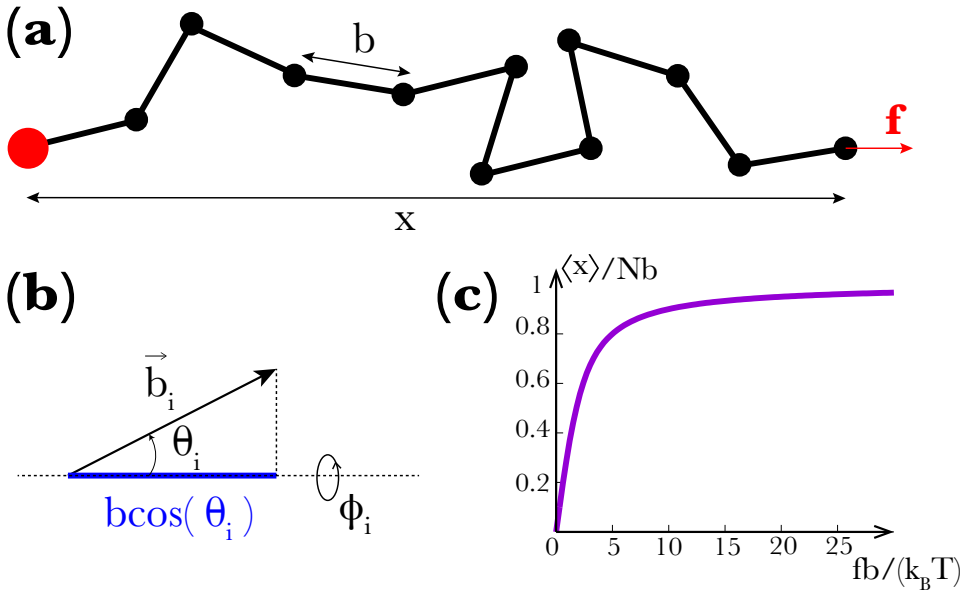

Figure B.1: **Freely Jointed Chain model. (a)-** Illustration of a linear polymer made by N monomers of length b under the action of an external force f. Big red leftmost ball indicates that this edge is fixed. **(b)-** Contribution of each monomer to the end-to-end distance x. **(c)-** Rescaled average end-to-end distance as a function of the dimensionless force $fb/\,k_BT$.

At zero force, due to thermal fluctuations, the polymer explores a large number of different configurations (microstates). However, as force increases, the number of available microstates decreases. Indeed, when the polymer becomes fully extended, there is only one available microstate. Therefore, the elastic response of polymers is due to entropic effects: competition between the external action that forces the chain to adopt entropically less favourable conformations and the force that tends to collapse the chain.

Due to the linearity of the polymer, the contribution to the end-to-end distance from i-th monomer equals to $b\cos\theta_i$. Therefore, the end-to-end distance x equals to:

$$x(\theta_1,\cdots,\theta_N,\phi_1,\cdots,\phi_N) = b\sum_{i=1}^{N}\cos\theta_i\,. \tag{B.1}$$

The ϕ angles take into account the freely rotating ends. According to Fig. B.1(b), $\theta \in [0,\pi]$ and $\phi \in [0,2\pi]$. Therefore, the partition function in the Force Ensemble equals to:

$$Z_f = \sum_{\{s\}} e^{\frac{fb}{k_B T}\sum_{i=1}^{N}\cos\theta_i}\,, \tag{B.2}$$

where $\{s\}$ denotes the set of available microstates. Indeed, the sum over microstates can be replaced by the following integral:

$$Z_f = \prod_{i=1}^{N}\int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta \, \exp\left(\frac{fb}{k_B T}\cos\theta\right)\,. \tag{B.3}$$

Hence, the average end-to-end distance $\langle x\rangle$ at a force f can be obtained via:

$$\langle x\rangle = k_B T\frac{\partial \log Z_f}{\partial f} = Nb\left(\frac{1}{\tanh\left(\frac{fb}{k_B T}\right)} - \frac{k_B T}{fb}\right)\,. \tag{B.4}$$

Usually, the FJC model is written in the following way:

$$x(f) = L_c\left(\frac{1}{\tanh\left(\frac{fb}{k_B T}\right)} - \frac{k_B T}{fb}\right)\,, \tag{B.5}$$

being $L_c = Nb$ the contour length of the polymer. A schematic illustration of the behavior of Eq. (B.5) is shown in Fig.B.1(c), where it is plot $\langle x \rangle / Nb$ as a function of $fb/\, k_B T$.

B.1.1 *Low and high force regimes*

For the **low force regime** (or high temperature, as $fb/\, k_B T \ll 1$) it is possible to perform a Taylor expansion of the hyperbolic function around zero:

$$x(f) = L_c \left(\left(\frac{k_B T}{fb} + \frac{fb}{3\, k_B T} + \mathcal{O}(f^3) \right) - \frac{k_B T}{fb} \right) \approx L_c \frac{fb}{3\, k_B T} \,. \quad \text{(B.6)}$$

Hence, for small forces, the polymer behaves like a Hookean spring (i.e. $f = k_{FJC} x$) with an elastic constant equal to: $k_{FJC} = \frac{3\, k_B T}{L_c b}$. Indeed, for high temperatures, entropic effects become more notorious, rendering more difficult to pull the polymer.

On the other hand, for **high forces** (or low temperatures, as $fb/\, k_B T \gg 1$) the hyperbolic tends to unity very fast. Therefore, Eq. B.5 becomes:

$$x(f) = L_c \left(1 - \frac{k_B T}{fb} \right) \,. \quad \text{(B.7)}$$

Previous equation indicates that an infinite force is required to fully stretch the polymer or, equivalently, entropic effects cannot be ignored in any regime, yielding, in general, $x(f) < L_c$.

B.1.2 *Extensible Freely Jointed Chain (EFJC) model*

In the FJC model, bending effects have not been considered. Moreover, monomers are assumed to be inextensible. This is not the case, though, of real polymers, where monomers are more compliant to extend at high enough forces. Then, in order to take this effect into account, a modification of Eq. (B.5) was proposed *ad hoc* by Smith, Cui and Bustamante in 1996 [155] to characterize the overstretching transition of B-DNA (see Sec. 2.1.1 and 2.2.2 of the present thesis for a brief

description of the effect). The Extensible Freely Jointed Chain (EFJC) reads as:

$$x(f) = L_c \left(\frac{1}{\tanh\left(\frac{fb}{k_B T}\right)} - \frac{k_B T}{fb} \right) \left(1 + \frac{f}{S}\right), \tag{B.8}$$

being now S the stretching modulus of the polymer in units of force. Despite that Eq. (B.8) has been widely used to characterize the elastic response of several polymers, it has not been until recently that it has been derived from statistical mechanics principles [156].

B.2 WORM-LIKE CHAIN (WLC) MODEL

A more detailed and realistic description of polymers is done through the Worm-Like Chain (WLC) model. Polymers now are assumed to be isotropic homogeneous semiflexible rods. Therefore, bending effects are now considered and a sort of cooperative effects between monomers appear (nearby segments are roughly aligned, see Fig. B.2(a)).

In the WLC model, the energetic cost of bending the polymer can be written as:

$$\mathcal{H}_{\text{bend.}} = \frac{P\, k_B T}{2} \int_0^{L_c} ds \left(\frac{\partial^2 \vec{r}(x)}{\partial s^2} \right)^2 = \frac{P\, k_B T}{2} \int_0^{L_c} ds \left(\frac{\partial \hat{t}(x)}{\partial s} \right)^2, \tag{B.9}$$

where P is the persistence length of the polymer, $\vec{r}(s)$ is the position vector along the chain, $\hat{t}(s) = \frac{\partial \vec{r}(s)}{\partial s}$ is the unit tangent vector to the chain at the point s (see Fig. B.2(b)) and L_c is the contour length of the polymer. The end-to-end distance can be obtained via: $x = \int_0^{L_c} \hat{t}(s) ds$. The persistence length P measures the tangent-tangent correlation function of the polymer at zero force as: $\langle \hat{(t)}(s) \cdot \hat{t}(s + \Delta s) \rangle = e^{-|\Delta s|/P}$. Also, it quantifies the bending effects due to thermal fluctuations and it is the key parameter of the WLC model.

As we explained for the FJC model, when the polymer is stretched, the number of accessible states of the polymer is reduced, causing an

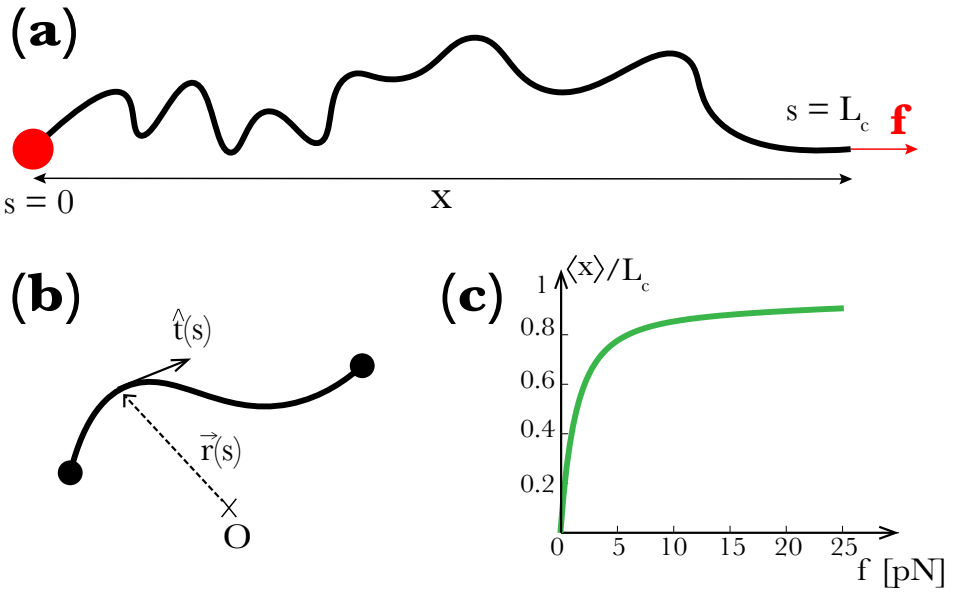

Figure B.2: **Worm-Like Chain model. (a)-** Illustration of a continuous linear polymer under the action of an external force f. Big red leftmost ball indicates that this edge is fixed. **(b)-** Position ($\vec{r}(s)$-) and unit tangent vector ($\hat{u}(s)$) along the contour length s of a segment of the whole polymer. **(c)-** Relative extension as a function of the mechanical force f.

entropic force against the external force f. Then, the Hamiltonian of the polymer can be written as:

$$\mathcal{H} = \mathcal{H}_{\text{bend.}} + \mathcal{H}_{\text{stret.}} = \frac{P\,k_B T}{2} \int_0^{L_c} ds \left(\frac{\partial^2 \vec{r}(x)}{\partial s^2} \right)^2 - xf\,, \quad \text{(B.10)}$$

being now x the extension of the polymer. The WLC model has no analytical solution but using the previous Hamiltonian as an energy functional, the partition function of the system can be minimized and the following interpolating formula for the force-extension curve can be found [84]:

$$f = \frac{k_B T}{4P} \left(\left(1 - \frac{\langle x \rangle}{L_c} \right)^{-2} - 1 + 4\frac{\langle x \rangle}{L_c} \right) . \quad \text{(B.11)}$$

The behavior of the relative extension (i.e. $\langle x \rangle / L_c$) as a function of the applied force f is shown in Fig. B.2(c). It is important to mention several improvements of Eq. (B.11) have been developed. For

instance, by adding a polynomial correction [157] or non-polynomial more sophisticated corrections [158].

B.2.1 *Extensible Worm-Like Chain (EWLC) model*

As we discussed in the previous section, polymers elongate due to external forces. This enthalpic[1] effect can be taken into account by adding an extra term to the Hamiltonian of the polymer as follows:

$$\mathcal{H} = \mathcal{H}_{\text{bend.}} + \mathcal{H}_{\text{enthal.}} + \mathcal{H}_{\text{stret.}}\,, \tag{B.12}$$

where now $\mathcal{H}_{\text{enthal.}} = \frac{Sx^2}{2L_c}$. Then, for the low force regime ($f < 10$ pN) the Extensible Worm-Like Chain (EWLC) model yields [157]:

$$f = \frac{k_{\text{B}}T}{4P}\left(\left(1 - \frac{\langle x\rangle}{L_c} + \frac{f}{S}\right)^{-2} - 1 + 4\left(\frac{\langle x\rangle}{L_c} - \frac{f}{S}\right)\right)\,. \tag{B.13}$$

Again, there are more accurate interpolation formulas for the EWLC model [158] or approximations for higher-force regimes [159]. Nevertheless, throughout this thesis we have used the standard WLC model given by Eq. (B.11) or, when required, the EWLC given by Eq. (B.13).

1 It is said to be an enthalpic effect because the elongation of the monomers due to the external force changes the average energy of the system at controlled force.

C

FOLDING FREE ENERGY RECOVERY

The free energy difference ΔG determined in nonequilibrium pulling experiments by terms of the Crooks Fluctuation Theorem (CFT) contains several inherent elastic contributions due to the different elements of the experimental set-up. Therefore,

$$\Delta G_x = \begin{cases} \Delta G_0 + \Delta W_{\text{st}}^{\text{rev}} + W_{\text{handles}}^{\text{rev}}. & \text{for MT experiments} \\ \Delta G_0 + \Delta W_{\text{st}}^{\text{rev}} + W_{\text{handles}}^{\text{rev}} + \Delta W_{\text{ot}}^{\text{rev}}. & \text{for LOT experiments} \end{cases} \tag{C.1}$$

Being ΔG_x the free energy difference between the unfolded (U) state and the folded (F) state calculated in the extensional ensemble (ExtEns). ΔG_0 is the folding free energy at zero force. $\Delta W_{\text{stret.}}^{\text{rev.}} = W_{\text{stret.}}^{U} - W_{\text{stret.}}^{F}$ is the difference between the reversible work required to stretch the unfolded single-stranded DNA molecule from o up to a maximum force f_{max} (molecular extension at f_{max}: x_U) and the reversible work needed to align the folded DNA hairpin along the force axis from o to f_{min} (molecular extension at f_{min}: x_F):

$$\Delta W_{\text{stret.}}^{\text{rev.}} = \int_0^{x_U(f_{\text{max}})} f_U(x')\,dx' - \int_0^{x_F(f_{\text{min}})} f_F(x')\,dx'\,. \tag{C.2}$$

Where $f_U(x)$ $[f_F(x)]$ and the inverse function $x_U(f)$ $[x_F(f)]$ are the equation of state of the unfolded (folded) DNA. The first integral is calculated using the WLC model given by Eq. (B.11) and setting $\langle x \rangle = x_n$ as the extension of the n bases of released single-stranded nucleic acid, $L_c^n = nd_b$ the contour length and d_b equal to the average interphosphate distance. On the other hand, the second integral is computed according to the FJC model (Eq. (B.5)) considering the hairpin as a single dipole with fixed diameter $d = 2$ nm and equal Kuhn length [78].

The term $\Delta W^{\text{rev.}}_{\text{handles}}$ is the reversible work needed to stretch the handles from f_{min} to f_{max}:

$$\Delta W^{\text{rev.}}_{\text{handles}} = \int_{x_{\text{handles}}(f_{min})}^{x_{\text{handles}}(f_{max})} f_{\text{handles}}(x')\, dx' \quad . \tag{C.3}$$

The handles are modeled according to the EWLC model (Eq. (B.13)) using the Bouchiat interpolation formula [157].

Finally, $\Delta W^{\text{rev.}}_{\text{bead}}$ is the reversible work needed to pull the optically trapped bead from f_{min} to f_{max}:

$$\Delta W^{\text{rev}}_{\text{ot}} = \int_{x_{\text{bead}}(f_{min})}^{x_{\text{bead}}(f_{max})} f(x')\, dx' = \int_{f_{min}}^{f_{max}} \frac{1}{k_{\text{ot}}(f')} df' \quad , \tag{C.4}$$

where $k_{\text{ot}}(f)$ is the force-dependent stiffness of the optical trap determined for the miniTweezers instrument [71]. Since in MT the bead is always in the equilibrium position of the trap, there is no energy contribution due to the displacement of the bead in the trap.

Note that: $\Delta G_x = \Delta G_f + \langle \Delta\,(xf) \rangle$, where ΔG_f is the free energy difference measured in the force ensemble (ForceEns) and $\langle \Delta\,(xf) \rangle$ is the average over all experimental realizations of the force and extension boundary terms.

C.0.1 *Effective stiffness approximation*

Typically, in LOT experiments, the stiffness of the optical trap (i.e. k_{ot}) is not known. It depends on the size of the bead, the laser power, the surrounding medium, etc. Therefore, for short ranges of integration, it is possible to perform an approximation that bypasses the precise knowledge of the elastic response of the handles and the bead in the optical trap [32]. The sum of the contributions $\Delta W^{\text{rev.}}_{\text{handles}}$ and $\Delta W^{\text{rev.}}_{\text{bead}}$ (i.e. Eqs. (C.3), (C.4)) can be written as:

$$\Delta W^{\text{rev.}}_{\text{handles}} + \Delta W^{\text{rev.}}_{\text{ot}} = \int_{f_{min}}^{f_{max}} f' \left(\frac{1}{k_{\text{handles}}} + \frac{1}{k_{\text{ot}}} \right) df' \tag{C.5}$$

$$\approx \int_{f_{min}}^{f_{max}} f' \left(\frac{1}{k^F_{\text{eff}}} \right) df' \tag{C.6}$$

$$= \frac{f^2_{\text{max}} - f^2_{\text{min}}}{2k^F_{\text{eff}}} \, , \tag{C.7}$$

where k_{eff}^{F} is the effective stiffness of the folded branch. Here we have performed two approximations: first, we have considered that the k_{ot} is independent of the force and, second, that the folded molecule is much stiffer than the combination of the handles and the bead, so that it does not contribute to the total stiffness of the system.

D

FORCE KINETICS

The goal of this appendix is to summarize and present the concepts involved in first-rupture forces analysis. A thorough theoretical description of kinetic rates (and beyond Bell-Evans theory) can be found in Ref. [32].

D.1 RUPTURE FORCES, SURVIVAL PROBABILITIES AND KINETIC RATES

Let us consider a nonequilibrium pulling experiment in which a short hairpin (can be either DNA or RNA) experiences a structural transition (the molecule unfolds or folds). In standard extension controlled experiments (i.e. ExtEns), the fingerprint of such transition is a force jump. The force at which the hairpin unfolds for the first time in an experimental realization will be referred to as unfolding force (f_U) while the first force at which it refolds will be referred to as folding force (f_F). Due to the Brownian nature of the system, the unfolding-folding forces are stochastic. Therefore, they are different in each experimental realization and force histograms $\rho\left(f_U\right), \rho\left(f_F\right)$ can be easily obtained.

The survival probability (or survival function), $S_{F(U)}(f)$, quantifies the probability that a molecule remains in the $F(U)$ state at a force f. They can be obtained from the force histograms $\rho\left(f_U\right), \rho\left(f_F\right)$ according to:

$$S_F(f) = 1 - P(\{f < f_U\}) = 1 - \int_0^f \rho\left(f_U\right) df_U\,, \tag{D.1}$$

$$S_U(f) = P(\{f < f_F\}) = 1 - \int_f^\infty \rho\left(f_F\right) df_F\,. \tag{D.2}$$

On the other hand, the time (or force) evolution of both survival probabilities, $S_{F(U)}(f)$, can be obtained by considering that they are

described by a first-order Markov process. For the case of the unfolding process:

$$\frac{dS_F(f)}{df} = -\frac{k_{\mathrm{F}\to U}(f)}{r} S_F(f)\,, \tag{D.3}$$

whereas for the folding process:

$$\frac{dS_U(f)}{df} = -\frac{k_{\mathrm{U}\to F}(f)}{r} S_U(f)\,, \tag{D.4}$$

where $r = \dot{f}$ is the loading rate and $k_{\mathrm{F}\to U}(f)$, $k_{\mathrm{U}\to F}(f)$ are the force-dependent folding-unfolding kinetic rates of the hairpin. Hence, they are obtained as:

$$k_{\mathrm{F}\to U}(f) = -r\frac{1}{S_F(f)}\frac{dS_F(f)}{df}\,, \tag{D.5}$$

$$k_{\mathrm{U}\to F}(f) = -r\frac{1}{S_U(f)}\frac{dS_U(f)}{df}\,. \tag{D.6}$$

Since the survival probability of the folded state decreases with force, the kinetic rate $k_{\mathrm{F}\to U}(f)$ increases with the force, while $k_{\mathrm{U}\to F}(f)$ decreases.

E

MATHEMATICAL METHODS AND DEMONSTRATIONS

E.1 KOLMOGOROV–SMIRNOV STATISTIC

The goal of this section is to introduce the statistic used in the Kolmogorov-Smirnov (K-S) test. K-S test is a statistical test that aims to unveil whether an empirical cumulative distribution function $\hat{F}_n$ is drawn from a known cumulative distribution function F.

The statistical hypotheses are defined as:

$$H_0 : \hat{F}_n = F \quad H_1 : \hat{F}_n \neq F \,.$$

On one hand, $\hat{F}_n(x)$ is the empirical distribution function of $X_1, \cdots, X_n$ of n i.i.d. random variables defined as:

$$\hat{F}_n(x) := \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}\{X_i \leq x\} \,, \tag{E.1}$$

where $\mathbb{1}$ is the indicator function. On the other hand, $F(x)$ is the *real* cumulative distribution function of $\{X\}$. We note that $\hat{F}_n(x)$ is a consistent:

$$\mathbb{E}[\hat{F}_n(x)] = \frac{1}{n}\mathbb{E}[n\hat{F}_n(x)] = F(x) \,, \tag{E.2}$$

and an unbiased estimator:

$$\text{Var}\left(\hat{F}_n(x)\right) = \frac{1}{n^2}\text{Var}\left(n\hat{F}_n(x)\right) = \frac{F(x)\left(1 - F(x)\right)}{n} \,, \tag{E.3}$$

of $F(x)$. We point out that previous expressions have been obtained by recalling that $\sum_{k=1}^{n} \mathbb{1}\{X_i \leq x\}$ is the sum of n independent Bernoulli random variables, so $n\hat{F}_n(x)$ is a binomial random variable. According to the Central Limit Theorem, the following equality holds:

$$\sqrt{n}\left(\hat{F}_n(x) - F(x)\right) \to^d \mathcal{N}\left(0, F(x)\left(1 - F(x)\right)\right) \,. \tag{E.4}$$

Where d means that they converge in distribution. Equally, the same convergence in distribution holds for: $\sqrt{n}\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F(x)\right|$.

Theorem without proof. *For $n\to\infty$ we have:*

$$P\left(\sqrt{n}\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F(x)\right|\leq c\right)\to 1-2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2c^2k^2}\,. \quad \text{(E.5)}$$

We note that the c parameter is the same than $c(\alpha)$ in Eq. (6.1) (main text of the thesis).

Let us define the following statistic (Kolmogorov-Smirnov statistic):

$$D_n := \sup_{x\in\mathbb{R}}\left(\hat{F}_n(x)-F(x)\right)\,. \quad \text{(E.6)}$$

It is important to have in mind two important properties of D_n (enunciated without proof):

i. As $n\to\infty$, $D_n\to 0$ almost surely. This property can be proven using the Glivenko-Cantelli theorem [160].

ii. The distribution of D_n is the same for all continuous underlying distribution functions F (The Distribution-Free property).

We note that we can find the threshold c by recalling the definition of the significance level α:

$$\alpha = P(D_n > c \mid H_0)\,. \quad \text{(E.7)}$$

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.073	1.224	1.358	1.517	1.628	1.858

Table E.1: **Values of significance level α and threshold $c(\alpha)$.** Values for $c(\alpha)$ have been obtained by numerically solving Eq. (E.8). A good approximation for $c(\alpha)$ is $c(\alpha)=\sqrt{\frac{-\log\alpha}{2}}$ [161].

Then, from Eq. (E.5) and considering $n \to \infty$:

$$P\left(\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x) - F(x)\right| > \frac{c}{\sqrt{n}}\right) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2c^2k^2} = \alpha\,. \quad \text{(E.8)}$$

The threshold c clearly depends on α, so strictly it must read as $c(\alpha)$ (as it does in Eq. (6.1)). In Table E.1 we report some of the most typical values for α and the corresponding $c(\alpha)$ obtained from Eq. (E.8).

We conclude this section by mentioning that Kolmogorov-Smirnov statistic can be used in order to unveil the compatibility between two distributions (as we did in section 6.3.1). Note that since both samples may have different sizes, $\sqrt{n}$ must read as $\sqrt{\frac{nm}{n+m}}$, being n and m the sizes of both samples.

E.2 PROBABILITY DISTRIBUTION OF SAMPLE VARIANCE

In this section we will proof the following statement:

$$\frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{(N-1)}\,. \quad \text{(E.9)}$$

Where $S^2 = (N-1)^{-1}\sum_i(X_i - \bar{X})^2$ is the sample variance of N random and independent observations drawn from a $\mathcal{N}(\mu, \sigma^2)$ distribution and $\chi^2_{(N-1)}$ is the chi-squared distribution with $N-1$ degrees of freedom.

First of all we need to bear in mind that $\bar{X}$ (i.e. the sample mean) and S^2 are independent[1]. Then, let us consider the following function:

$$W = \sum_{i=1}^{N}\left(\frac{X_i - \mu}{\sigma}\right)^2, \quad \text{(E.10)}$$

now let us add and subtract $\bar{X}$ in the numerator of the right hand side of last equation as:

$$W = \sum_{i=1}^{N}\left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma}\right)^2. \quad \text{(E.11)}$$

1 According to Cochran's theorem

By expanding the square we obtain:

$$W = \sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^{N} \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 + 2 \left(\frac{\bar{X} - \mu}{\sigma} \right) \sum_{i=1}^{N} (X_i - \bar{X}) \,. \tag{E.12}$$

Note that the rightmost term is equal to zero since: $\sum_i (X_i - \bar{X}) = N\bar{X} - N\bar{X} = 0$. Thus, Eq. (E.12) becomes:

$$W = \sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + N \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \,. \tag{E.13}$$

Note that the second term of Eq. (E.12) does not depend on i, so the sum equals to N. Now, inserting the definition of S^2 (see above) in Eq. (E.13) we have:

$$W = \frac{(N-1)S^2}{\sigma^2} + \frac{N(\bar{X} - \mu)^2}{\sigma^2} \,. \tag{E.14}$$

Now we have to recall that since $W = \sum_{i=1}^{N} \left(\frac{X_i - \mu}{\sigma} \right)^2$ and $X_1, \cdots X_N$ are drawn from a $\mathcal{N}(\mu, \sigma^2)$, the quantity inside the parenthesis is a standardized variable[2]. Then, defining $Z_i = (X_i - \bar{X})/\sigma$, we have:

$$W = \sum_{i=1}^{N} Z_i^2 \,. \tag{E.15}$$

The moment-generating function of W, equals to:

$$m_W := \langle e^{tW} \rangle = \langle e^{tZ_1^2} \rangle \cdots \langle e^{tZ_N^2} \rangle = m_{Z_1^2} \cdots m_{Z_N^2} \,. \tag{E.16}$$

The function m_W can be explicitly calculated:

$$m_W = \int_{-\infty}^{\infty} dz \, f(z) e^{tz} = (1 - 2t)^{-1/2} \quad \text{for } t < \frac{1}{2} \quad , \tag{E.17}$$

2 Also called z-score. It is a rescaled variable with zero mean and standard deviation equal to 1.

where we have used the fact that $f(z) = (2\pi)^{-1}\exp(-z^2/2)$ for standardized variables. We note that the result of Eq. (E.17) is precisely the moment-generating function of the chi-squared distribution [162]. As a consequence, and after performing the product of the N moment-generating functions in Eq. (E.16), we can see that W (Eq. (E.10)) is chi-squared distributed with N degrees of freedom (i.e. N Z_i's).

From Eq. (E.14) we can write:

$$m_W = \langle e^{tW} \rangle \tag{E.18}$$

$$= \left\langle \exp\left(t\frac{(N-1)S^2}{\sigma^2}\right) \exp\left(t\frac{N(\bar{X}-\mu)^2}{\sigma^2}\right) \right\rangle \tag{E.19}$$

$$= m_{\frac{(N-1)S^2}{\sigma^2}}\, m_{\frac{N(\bar{X}-\mu)^2}{\sigma^2}}\,. \tag{E.20}$$

We have used the fact that $\bar{X}$ and S^2 are independent (so their functions). On the other hand, using the same argument than before (Eq. (E.16)), the quantity $\left(\frac{\bar{X}-\mu}{\sigma}\right)^2$ is also chi-squared distributed with 1 degree of freedom. Therefore, Eq. (E.20) yields:

$$(1-2t)^{-N/2} = m_{\frac{(N-1)S^2}{\sigma^2}}\,(1-2t)^{-1/2}\,. \tag{E.21}$$

Therefore, solving Eq. (E.21) for $m_{\frac{(N-1)S^2}{\sigma^2}}$ we have:

$$m_{\frac{(N-1)S^2}{\sigma^2}} = (1-2t)^{-(N-1)/2} \quad \text{for } t < 1/2\,. \tag{E.22}$$

Which is again a chi-squared distribution but with $N-1$ degrees of freedom. The uniqueness property of the moment-generating function indicates us that:

$$\frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{(N-1)}\,, \tag{E.23}$$

as we wanted to proof.

E.3 MATHEMATICAL PROOF OF THE INFORMATION-CONTENT FLUCTUATION THEOREM

Let us consider an heterogeneous system in which there there are M observable traits (i.e. phenotypes or physically measurable quantities). Moreover, let us suppose that we perform an isothermal thermodynamic transformation $0 \to 1$ applied to all individuals of the ensemble through the variation of a control parameter λ from λ_0 to λ_1 in a $\lambda_{\to}(t)$ protocol in a time Δt.

The forward (F) and reversed (R) ensemble work distributions (EWD) are defined as:

$$\mathcal{P}_{\to}(W) = \sum_{\alpha} p_{\alpha} P_F^{(\alpha)}(W)\,, \tag{E.24}$$

$$\mathcal{P}_{\leftarrow}(-W) = \sum_{\alpha} \hat{p}_{\alpha} P_R^{(\alpha)}(-W)\,, \tag{E.25}$$

where $P_F^{(\alpha)}(W)$, $P_R^{(\alpha)}(-W)$ are the F and R work distribution corresponding to an individual α calculated in the $0 \to 1$ transformation. On the other hand, p_α, $\hat{p}_\alpha$ are the probability of finding an individual with a given phenotype. Note that individual work distributions fulfil the Crooks Fluctuation Theorem (CFT):

$$\frac{P_F^{(\alpha)}(W)}{P_R^{(\alpha)}(-W)} = \exp\left(\frac{W - \Delta G_\alpha^{01}}{k_{\mathrm{B}} T}\right)\,, \tag{E.26}$$

where $\Delta G_\alpha^{01} = G_\alpha^1 - G_\alpha^0 := G_\alpha(\lambda_1) - G_\alpha(\lambda_0)$ is the free energy difference between λ_1 and λ_0 states.

In Eq. (E.24) we substitute the value of $P_F^{(\alpha)}(W)$ given by Eq. (E.26). Then, Eq. (E.24) becomes:

$$\mathcal{P}_{\to}(W) = \sum_{\alpha} p_{\alpha} P_R^{\alpha}(-W) \exp\left(\frac{W - \Delta G_\alpha^{01}}{k_{\mathrm{B}} T}\right) \tag{E.27}$$

$$= e^{\frac{W}{k_{\mathrm{B}} T}} \sum_{\alpha} p_{\alpha} P_R^{\alpha}(-W) e^{-\frac{G_\alpha^1 - G_\alpha^0}{k_{\mathrm{B}} T}}\,. \tag{E.28}$$

Now we note that, in order to have a fluctuation theorem-symmetry for the EWD, p_α and $\hat{p}_\alpha$ must be related as:

$$p_\alpha \exp\left(-\frac{G^1_\alpha - G^0_\alpha}{k_B T}\right) = \hat{p}_\alpha \exp\left(-\frac{\Delta\mathcal{G} - k_B T\mathcal{I}}{k_B T}\right), \tag{E.29}$$

where $\Delta\mathcal{G} = -k_B T \log \frac{Z_1}{Z_0}$, being Z_0 and Z_1 the partition functions given by:

$$Z_0 = \sum_\alpha e^{-G^0_\alpha / k_B T} \qquad Z_1 = \sum_\alpha e^{-G^1_\alpha / k_B T}. \tag{E.30}$$

On the other hand $k_B T\mathcal{I}$ is the information-content of the ensemble defined as the minimum free energy cost required to generate the population of partially equilibrated individuals (defined by p_α) starting from a population in full thermodynamic equilibrium. Inserting the result of Eq. (E.29) in Eq. (E.28) we obtain:

$$\mathcal{P}_\rightarrow(W) = e^{\frac{W}{k_B T}} \left(\sum_\alpha \hat{p}_\alpha P^\alpha_R(-W)\right) e^{-\frac{\Delta\mathcal{G} - k_B T\mathcal{I}}{k_B T}} = e^{\frac{W}{k_B T} - \frac{\Delta\mathcal{G} - k_B T\mathcal{I}}{k_B T}} \mathcal{P}_\leftarrow(-W). \tag{E.31}$$

Finally, Eq. (E.31) can be rewritten in order to obtain the information-content fluctuation theorem:

$$\frac{\mathcal{P}_\rightarrow(W)}{\mathcal{P}_\leftarrow(-W)} = \exp\left(\frac{W - \Delta\mathcal{G} + k_B T I}{k_B T}\right). \tag{E.32}$$

We note that, for $\mathcal{I} = 0$ we obtain the result corresponding to the equilibrium phenotypic ensemble. We can prove it by summing for all α on both sides of Eq. (E.29) (and setting $\mathcal{I} = 0$):

$$\sum_\alpha p^{\text{eq}}_\alpha \exp\left(-\frac{G^1_\alpha - G^0_\alpha}{k_B T}\right) = \exp\left(-\frac{\Delta\mathcal{G}}{k_B T}\right) \sum_\alpha \hat{p}^{\text{eq}}_\alpha. \tag{E.33}$$

Recalling that the frequencies p^{eq}_α, $\hat{p}^{\text{eq}}_\alpha$ are normalized[3], Eq. (E.33) becomes:

$$\sum_\alpha p^{\text{eq}}_\alpha \exp\left(-\frac{G^1_\alpha - G^0_\alpha}{k_B T}\right) = \exp\left(-\frac{\Delta\mathcal{G}}{k_B T}\right). \tag{E.34}$$

3 $\sum_\alpha p^{\text{eq}}_\alpha = \sum_\alpha \hat{p}^{\text{eq}}_\alpha = 1$.

Now, inserting the equilibrium probabilities ($p_\alpha^{\mathrm{eq}} = e^{-G_\alpha^0/k_\mathrm{B}T} / \sum_\alpha e^{-G_\alpha^0/k_\mathrm{B}T}$):

$$\sum_\alpha \frac{\exp\left(-\frac{G_\alpha^0}{k_\mathrm{B}T}\right)}{\sum_\alpha \exp\left(-\frac{G_\alpha^0}{k_\mathrm{B}T}\right)} \exp\left(-\frac{G_\alpha^1 - G_\alpha^0}{k_\mathrm{B}T}\right) = \exp\left(-\frac{\Delta\mathcal{G}}{k_\mathrm{B}T}\right) . \quad \text{(E.35)}$$

Expanding the left-hand side of Eq. (E.35) we obtain:

$$\frac{\sum_\alpha \exp\left(-\frac{G_\alpha^1}{k_\mathrm{B}T}\right)}{\sum_\alpha \exp\left(-\frac{G_\alpha^0}{k_\mathrm{B}T}\right)} = \frac{Z_1}{Z_0} = \exp\left(-\frac{\Delta\mathcal{G}}{k_\mathrm{B}T}\right) , \quad \text{(E.36)}$$

as required for an equilibrium phenotypic ensemble.

Finally we point out that, summing for all α in Eq. (E.29) we obtain the reported equation for the information-content (Eq. (7.8)):

$$\sum_\alpha p_\alpha \exp\left(-\frac{G_\alpha^1 - G_\alpha^0}{k_\mathrm{B}T}\right) = \exp\left(-\frac{\Delta\mathcal{G} - k_\mathrm{B}T\mathcal{I}}{k_\mathrm{B}T}\right) \sum_\alpha \hat{p}_\alpha^{\mathrm{eq}} , \quad \text{(E.37)}$$

or, equivalently:

$$k_\mathrm{B}T\mathcal{I} = \Delta\mathcal{G} + k_\mathrm{B}T \log\left(\sum_\alpha p_\alpha \exp\left(-\frac{G_\alpha^1 - G_\alpha^0}{k_\mathrm{B}T}\right)\right) . \quad \text{(E.38)}$$

E.4 FLUCTUATION THEOREM FOR WHITE AVERAGED INDIVIDUAL WORK DISTRIBUTIONS

In this section we show the explicit calculations for the case that the EWD are the white average of the individual work distributions. That is, $p_\alpha = \hat{p}_\alpha = 1/M$, where M is the number of phenotypes. Then, the white average EWD are:

$$\mathcal{P}_{\rightarrow}^{\mathrm{white}}(W) = \frac{1}{M} \sum_\alpha P_F^{(\alpha)}(W) , \quad \text{(E.39)}$$

$$\mathcal{P}_{\leftarrow}^{\mathrm{white}}(-W) = \frac{1}{M} \sum_\alpha P_R^{(\alpha)}(-W) , \quad \text{(E.40)}$$

where the individual work distributions satisfy the CFT (Eq. (E.26)). Under Gaussian assumption, the folding free energy spectrum is Gaussian distributed according to:

$$Q(G) = \frac{1}{\sqrt{2\pi\sigma_G^2}} e^{-\frac{(G-G^\star)^2}{2\sigma_G^2}} . \tag{E.41}$$

Likewise, so are individual work distributions:

$$P^{(\alpha)}(W) = \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(W-\langle W\rangle)^2}{2\sigma_W^2}} . \tag{E.42}$$

Keep in mind the relation given by Jarzynski equality (Eq. (4.5)) and the moment-generating function for Gaussian variables[4]:

$$G = \langle W\rangle - \frac{\sigma_W^2}{2\,k_B T} . \tag{E.43}$$

In the continuum limit, the ratio between $\mathcal{P}_{\rightarrow}^{\text{white}}(W)$ and $\mathcal{P}_{\leftarrow}^{\text{white}}(-W)$ (Eqs. (E.39), (E.40)) can be written as:

$$\frac{\mathcal{P}_{\rightarrow}^{\text{white}}(W)}{\mathcal{P}_{\leftarrow}^{\text{white}}(-W)} = \frac{\int_{-\infty}^{\infty} dG P_G^R(-W) e^{(W-G)/\,k_B T} Q(G)}{\int_{-\infty}^{\infty} dG P_G^R(-W) Q(G)} , \tag{E.44}$$

Note that in previous equation we have used Eq. (E.26) and, since we integrate over G, α does not appear anymore. For convenience, using again the CFT (Eq. (E.26)) we rewrite Eq. (E.44) as:

$$\frac{\mathcal{P}_{\rightarrow}^{\text{white}}(W)}{\mathcal{P}_{\leftarrow}^{\text{white}}(-W)} = \frac{\int_{-\infty}^{\infty} dG P_G^F(W) Q(G)}{\int_{-\infty}^{\infty} dG P_G^F(W) e^{-(W-G)/\,k_B T} Q(G)} , \tag{E.45}$$

Now, in order to calculate both the numerator and the denominator of Eq. (E.45) we need to recall several results. First, $P_G(W)$ are given by (E.42). Second, the result from the Jarzynski equality: Eq. (E.43).

4 $\langle e^{-\frac{W}{k_B T}} \rangle = e^{-\frac{\langle W\rangle}{k_B T} + \frac{\sigma_W^2}{2(k_B T)^2}}$.

Third, considering that the work variances are molecule-independent and equal for the F and R distributions. Finally, we use the result for the Gaussian integral:

$$\int_{-\infty}^{\infty} e^{-ax^2+bx+c}dx = \sqrt{\frac{\pi}{a}}e^{\frac{b^2}{4a}+c} \quad \text{with} \quad a,b,c \in \mathbb{R}\,. \tag{E.46}$$

Now, putting all the pieces together, the integral of the numerator I_N, equals to:

$$\begin{aligned} I_N =& A \cdot \sqrt{\frac{2\pi}{\frac{1}{\sigma_G^2}+\frac{1}{\sigma_W^2}}} \cdot \exp\left(\frac{\sigma_G^2\sigma_W^2}{2(\sigma_G^2+\sigma_W^2)}\left(\frac{W}{\sigma_W^2}+\frac{G^\star}{\sigma_G^2}\right)^2\right)\cdot \\ & \exp\left(\frac{-\beta\sigma_G^2\sigma_W^2}{2(\sigma_G^2+\sigma_W^2)}\left(\frac{W}{\sigma_W^2}+\frac{G^\star}{\sigma_G^2}\right)\right)\cdot \\ & \exp\left(\frac{\beta^2\sigma_G^2\sigma_W^2}{8(\sigma_G^2+\sigma_W^2)}\right), \end{aligned} \tag{E.47}$$

where $\beta = (k_B T)^{-1}$. On the other hand, the integral of the denominator I_D, equals to:

$$\begin{aligned} I_D =& e^{-\beta W} A \cdot \sqrt{\frac{2\pi}{\frac{1}{\sigma_G^2}+\frac{1}{\sigma_W^2}}} \cdot \exp\left(\frac{\sigma_G^2\sigma_W^2}{2(\sigma_G^2+\sigma_W^2)}\left(\frac{W}{\sigma_W^2}+\frac{G^\star}{\sigma_G^2}\right)^2\right)\cdot \\ & \exp\left(\frac{\beta\sigma_G^2\sigma_W^2}{2(\sigma_G^2+\sigma_W^2)}\left(\frac{W}{\sigma_W^2}+\frac{G^\star}{\sigma_G^2}\right)\right)\cdot \\ & \exp\left(\frac{\beta^2\sigma_G^2\sigma_W^2}{8(\sigma_G^2+\sigma_W^2)}\right). \end{aligned} \tag{E.48}$$

Both in Eq. (E.47) and in Eq. (E.48), A is a real parameter that equals to:

$$A = \frac{e^{-\frac{(W-\frac{\beta}{2}\sigma_W^2)^2}{2\sigma_W^2}}}{2\pi\sigma_G\sigma_W} e^{-(G^\star)^2/2\sigma_G^2}\,. \tag{E.49}$$

Then, by dividing Eqs. (E.47), (E.48) we obtain the effective-fluctuation theorem (Eq. (7.18)):

$$\frac{\mathcal{P}^{\text{white}}_{\rightarrow}(W)}{\mathcal{P}^{\text{white}}_{\leftarrow}(-W)} = \exp\left(x\frac{W-G^{\star}}{k_{\text{B}}T}\right) \quad , \quad x = \frac{1}{1+\frac{\sigma_G^2}{\sigma_W^2}} \, . \tag{E.50}$$

E.5 INFERENCE OF FULL ENSEMBLE WORK DISTRIBUTIONS

In the thermodynamic inference context [60], the full work distributions can be recovered by looking for the quantity Δ that satisfies:

$$\frac{\mathcal{P}_{\rightarrow}(W)}{\mathcal{P}_{\leftarrow}(-W)} = \frac{\mathcal{P}'_{\rightarrow}(W-\Delta)}{\mathcal{P}'_{\leftarrow}(-W-\Delta)} = \exp\left(\frac{W-G^{\star}}{k_{\text{B}}T}\right) , \tag{E.51}$$

where the prime (') distributions are the ones that fulfil the effective-CFT (Eq. (E.50)). We point out that, for the sake of clarity, we have omitted the label *white* for all the distributions. Nevertheless, bear in mind that all the distributions of the present section correspond to the white averaged ones.

Multiplying at both sides of Eq. (E.50) by $\mathcal{P}'_{\leftarrow}(-W)$, integrating over W and using the Gaussian moment-generating function we have:

$$G^{\star} = \langle W \rangle - x\frac{\sigma_W^2+\sigma_G^2}{2\,k_{\text{B}}T} \, . \tag{E.52}$$

Then, by repeating the same procedure in Eq. (E.51) we obtain the following result:

$$G^{\star} = \langle W \rangle + \Delta - \frac{\sigma_W^2+\sigma_G^2}{2\,k_{\text{B}}T} \, . \tag{E.53}$$

In both cases the average $\langle \cdots \rangle$ runs over the $\rightarrow$ distributions. Then, after simple algebraic steps in Eqs. (E.52), (E.53) and substituting x (Eq. (E.50)) we obtain:

$$\Delta = \frac{\sigma_G^2}{2\,k_{\text{B}}T} \, . \tag{E.54}$$

BIBLIOGRAPHY

[1] Kim Sharp and Franz Matschinsky. "Translation of Ludwig Boltzmann's Paper "On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium" Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissen Classe. Abt. II, LXXVI 1877, pp 373-435 (Wien. Ber. 1877, 76: 373-435). Reprinted in Wiss. Abhandlungen, Vol. II, reprint 42, p. 164-223, Barth, Leipzig, 1909." In: *Entropy* 17.4 (2015), pp. 1971–2009.

[2] Pierre Perrot. *A to Z of Thermodynamics*. Oxford University Press on Demand, 1998.

[3] Crispin W Gardiner. "Handbook of stochastic methods for physics, chemistry and the natural sciences." In: *Applied Optics* 25 (1986), p. 3145.

[4] Ilya Prigogine. "Time, structure, and fluctuations." In: *Science* 201.4358 (1978), pp. 777–785.

[5] Claude E Shannon. "A mathematical theory of communication." In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[6] Edwin T Jaynes. "Information theory and statistical mechanics." In: *Physical Review* 106.4 (1957), p. 620.

[7] Edwin T Jaynes. "Information theory and statistical mechanics. II." In: *Physical Review* 108.2 (1957), p. 171.

[8] Marco Bischof. "Some remarks on the history of biophysics and its future." In: *Current Development of Biophysics* 22 (1996).

[9] Erwin Schrödinger. *What is life?: With mind and matter and autobiographical sketches*. Cambridge University Press, 1992.

[10] James D Watson, Francis HC Crick, et al. "Molecular structure of nucleic acids." In: *Nature* 171.4356 (1953), pp. 737–738.

[11] Armen R Kherlopian et al. "A review of imaging techniques for systems biology." In: *BMC Systems Biology* 2.1 (2008), p. 74.

[12] Felix Ritort. "Single-molecule experiments in biological physics: methods and applications." In: *Journal of Physics: Condensed Matter* 18.32 (2006), R531.

[13] Helen Miller et al. "Single-molecule techniques in biophysics: a review of the progress in methods and applications." In: *Reports on Progress in Physics* 81.2 (2017), p. 024601.

[14] Gerrit Sitters et al. "Acoustic force spectroscopy." In: *Nature Methods* 12.1 (2015), p. 47.

[15] Douwe Kamsma. "Acoustic Force Spectroscopy (AFS): From single molecules to single cells." In: *PhD Thesis* (2018).

[16] Keir C Neuman and Attila Nagy. "Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy." In: *Nature Methods* 5.6 (2008), p. 491.

[17] Ralf Dahm. "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research." In: *Human Genetics* 122.6 (2008), pp. 565–581.

[18] Ralf Dahm. "Friedrich Miescher and the discovery of DNA." In: *Developmental Biology* 278.2 (2005), pp. 274–288.

[19] Charles Darwin. "On the origins of species by means of natural selection." In: *London: Murray* 247 (1859), p. 1859.

[20] Richard R Sinden. *DNA structure and function*. Elsevier, 2012.

[21] Harvey Lodish. *Biología celular y molecular*. Ed. Médica Panamericana, 2005.

[22] Alexander Varshavsky. "Discovering the RNA double helix and hybridization." In: *Cell* 127.7 (2006), pp. 1295–1297.

[23] Arthur Ashkin. "Acceleration and trapping of particles by radiation pressure." In: *Physical Review Letters* 24.4 (1970), p. 156.

[24] Arthur Ashkin. "Optical trapping and manipulation of neutral particles using lasers." In: *Proceedings of the National Academy of Sciences* 94.10 (1997), pp. 4853–4860.

[25] Arthur Ashkin et al. "Observation of a single-beam gradient force optical trap for dielectric particles." In: *Optics Letters* 11.5 (1986), pp. 288–290.

[26] Josep Maria Huguet. "Statistical and thermodynamic properties of DNA unzipping experiments with optical tweezers." PhD thesis. Doctoral thesis, Barcelona, 2010.

[27] David S Bradshaw and David L Andrews. "Manipulating particles with light: radiation and gradient forces." In: *European Journal of Physics* 38.3 (2017), p. 034008.

[28] Karel Svoboda and Steven M Block. "Biological applications of optical forces." In: *Annual Review of Biophysics and Biomolecular Structure* 23.1 (1994), pp. 247–285.

[29] Keir C Neuman and Steven M Block. "Optical trapping." In: *Review of scientific instruments* 75.9 (2004), pp. 2787–2809.

[30] Steven B Smith, Yujia Cui, and Carlos Bustamante. "Optical-trap force transducer that operates by direct measurement of light momentum." In: *Methods in Enzymology* 361 (2003), pp. 134–162.

[31] Mark C Williams, Ioulia Rouzina, and Micah J McCauley. "Peeling back the mystery of DNA overstretching." In: *Proceedings of the National Academy of Sciences* 106.43 (2009), pp. 18047–18048.

[32] Anna Alemany. "Dynamic force spectroscopy and folding kinetics in molecular systems." PhD thesis. Doctoral thesis, Barcelona, 2014.

[33] Steven B Smith, Laura Finzi, and Carlos Bustamante. "Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads." In: *Science* 258.5085 (1992), pp. 1122–1126.

[34] Terence R Strick et al. "The elasticity of a single supercoiled DNA molecule." In: *Science* 271.5257 (1996), pp. 1835–1837.

[35] Stephen Blundell. *Magnetism in condensed matter*. 2003.

[36] John David Jackson. *Classical electrodynamics*. John Wiley & Sons, 2012.

[37] Jan Lipfert, Xiaomin Hao, and Nynke H Dekker. "Quantitative modeling and optimization of magnetic tweezers." In: *Biophysical Journal* 96.12 (2009), pp. 5040–5049.

[38] Scipione Bobbio et al. "Equivalent sources methods for the numerical evaluation of magnetic force with extension to nonlinear materials." In: *IEEE Transactions on magnetics* 36.4 (2000), pp. 663–666.

[39] Charlie Gosse and Vincent Croquette. "Magnetic tweezers: micromanipulation and force measurement at the molecular level." In: *Biophysical Journal* 82.6 (2002), pp. 3314–3329.

[40] Sara de Lorenzo et al. "A temperature-jump optical trap for single-molecule manipulation." In: *Biophysical Journal* 108.12 (2015), pp. 2854–2864.

[41] Rupa Sarkar and Valentin V Rybenkov. "A guide to magnetic tweezers and their applications." In: *Frontiers in Physics* 4 (2016), p. 48.

[42] Maria Manosas et al. "Magnetic tweezers for the study of DNA tracking motors." In: *Methods in Enzymology*. Vol. 475. Elsevier, 2010, pp. 297–320.

[43] Igor D Vilfan et al. "Magnetic tweezers for single-molecule experiments." In: *Handbook of single-molecule biophysics*. Springer, 2009, pp. 371–395.

[44] Herbert B Callen. *Thermodynamics and an Introduction to Thermostatistics*. 1998.

[45] Paul J Flory. *Statistical mechanics of chain molecules*. Wiley Online Library, New York, 1969.

[46] Fabio Manca et al. "On the equivalence of thermodynamics ensembles for flexible polymer chains." In: *Physica A: Statistical Mechanics and its Applications* 395 (2014), pp. 154–170.

[47] Mark Waldo Zemansky. *Heat and thermodynamics: an intermediate textbook*. McGraw-Hill, 1968.

[48] Sergio Ciliberto. "Experiments in stochastic thermodynamics: Short history and perspectives." In: *Physical Review X* 7.2 (2017), p. 021051.

[49] Udo Seifert. "Stochastic thermodynamics, fluctuation theorems and molecular machines." In: *Reports on Progress in Physics* 75.12 (2012), p. 126001.

[50] Delphine Collin et al. "Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies." In: *Nature* 437.7056 (2005), p. 231.

[51] Anna Alemany et al. "Mechanical folding and unfolding of protein barnase at the single-molecule level." In: *Biophysical Journal* 110.1 (2016), pp. 63–74.

[52] Anna Alemany et al. "Experimental free-energy measurements of kinetic molecular states using fluctuation theorems." In: *Nature Physics* 8.9 (2012), p. 688.

[53] Joan Camunas-Soler, Anna Alemany, and Felix Ritort. "Experimental measurement of binding energy, selectivity, and allostery using fluctuation theorems." In: *Science* 355.6323 (2017), pp. 412–415.

[54] Kumiko Hayashi et al. "Fluctuation theorem applied to F 1-ATPase." In: *Physical Review Letters* 104.21 (2010), p. 218103.

[55] Juan MR Parrondo, Jordan M Horowitz, and Takahiro Sagawa. "Thermodynamics of information." In: *Nature Physics* 11.2 (2015), p. 131.

[56] Shoichi Toyabe et al. "Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality." In: *Nature Physics* 6.12 (2010), p. 988.

[57] Gavin E Crooks. "Path-ensemble averages in systems driven far from equilibrium." In: *Physical Review E* 61.3 (2000), p. 2361.

[58] Christopher Jarzynski. "Nonequilibrium equality for free energy differences." In: *Physical Review Letters* 78.14 (1997), p. 2690.

[59] Chris Jarzynski. "Nonequilibrium work theorem for a system strongly coupled to a thermal environment." In: *Journal of Statistical Mechanics: Theory and Experiment* 2004.09 (2004), P09005.

[60] Marco Ribezzi-Crivellari and Felix Ritort. "Free-energy inference from partial work measurements in small systems." In: *Proceedings of the National Academy of Sciences* 111.33 (2014), E3386–E3394.

[61] J Michael Schurr and Bryant S Fujimoto. "Equalities for the nonequilibrium work transferred from an external potential to a molecular system. Analysis of single-molecule extension experiments." In: *The Journal of Physical Chemistry B* 107.50 (2003), pp. 14007–14019.

[62] Alessandro Mossa et al. "Measurement of work in single-molecule pulling experiments." In: *The Journal of Chemical Physics* 130.23 (2009), p. 234116.

[63] Frédéric Douarche et al. "Work fluctuation theorems for harmonic oscillators." In: *Physical Review Letters* 97.14 (2006), p. 140603.

[64] Terrell L Hill. *Thermodynamics of small systems*. Courier Corporation, 1963.

[65] Frédéric Douarche, Sergio Ciliberto, and Artyom Petrosyan. "Estimate of the free energy difference in mechanical systems from work fluctuations: experiments and models." In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.09 (2005), P09011.

[66] Onuttom Narayan and Abhishek Dhar. "Reexamination of experimental tests of the fluctuation theorem." In: *Journal of Physics A: Mathematical and General* 37.1 (2003), p. 63.

[67] Luca Peliti. "On the work–Hamiltonian connection in manipulated systems." In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.05 (2008), P05002.

[68] Jose MG Vilar and J Miguel Rubi. "Failure of the work-Hamiltonian connection for free-energy calculations." In: *Physical Review Letters* 100.2 (2008), p. 020601.

[69] Gavin E Crooks. "Comment regarding "On the Crooks fluctuation theorem and the Jarzynski equality"[J. Chem. Phys. 129, 091101 (2008)] and "Nonequilibrium fluctuation-dissipation theorem of Brownian dynamics"[J. Chem. Phys. 129, 144113 (2008)]." In: *The Journal of Chemical Physics* 130.10 (2009), 03B801.

[70] Eckhard Dieterich et al. "Control of force through feedback in small driven systems." In: *Physical Review E* 94.1 (2016), p. 012107.

[71] Nuria Forns et al. "Improving signal/noise resolution in single-molecule experiments using molecular constructs with short handles." In: *Biophysical Journal* 100.7 (2011), pp. 1765–1774.

[72] John SantaLucia. "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics." In: *Proceedings of the National Academy of Sciences* 95.4 (1998), pp. 1460–1465.

[73] Michael Zuker. "Mfold web server for nucleic acid folding and hybridization prediction." In: *Nucleic Acids Research* 31.13 (2003), pp. 3406–3415.

[74] Marc Rico-Pasto, Isabel Pastor, and Felix Ritort. "Force feedback effects on single molecule hopping and pulling experiments." In: *The Journal of Chemical Physics* 148.12 (2018), p. 123327.

[75] George Casella and Roger L Berger. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA, 2002.

[76] Andrea Crisanti, Marco Picco, and Felix Ritort. "Derivation of the spin-glass order parameter from stochastic thermodynamics." In: *Physical Review E* 97.5 (2018), p. 052103.

[77] Cyrus Levinthal. "How to fold graciously." In: *Mossbauer spectroscopy in biological systems* 67 (1969), pp. 22–24.

[78] Michael T Woodside et al. "Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins." In: *Proceedings of the National Academy of Sciences* 103.16 (2006), pp. 6190–6195.

[79] Michael T Woodside and Steven M Block. "Reconstructing folding energy landscapes by single-molecule force spectroscopy." In: *Annual Review of Biophysics* 43 (2014), pp. 19–39.

[80] Dustin B Ritchie and Michael T Woodside. "Probing the structural dynamics of proteins and nucleic acids with optical tweezers." In: *Current Opinion in Structural Biology* 34 (2015), pp. 43–51.

[81] David Wales. *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.

[82] Andrew J Ballard et al. "Energy landscapes for machine learning." In: *Physical Chemistry Chemical Physics* 19.20 (2017), pp. 12585–12603.

[83] Simona Cocco, John F Marko, and Rémi Monasson. "Theoretical models for single-molecule DNA and RNA experiments: from elasticity to unzipping." In: *Comptes Rendus Physique* 3.5 (2002), pp. 569–584.

[84] John F Marko and Eric D Siggia. "Stretching DNA." In: *Macromolecules* 28.26 (1995), pp. 8759–8770.

[85] Anna Alemany and Felix Ritort. "Determination of the elastic properties of short ssDNA molecules by mechanically folding and unfolding DNA hairpins." In: *Biopolymers* 101.12 (2014), pp. 1193–1199.

[86] Alessandro Mossa et al. "Dynamic force spectroscopy of DNA hairpins: I. Force kinetics and free energy landscapes." In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.02 (2009), P02060.

[87] George I Bell. "Models for the specific adhesion of cells to cells." In: *Science* 200.4342 (1978), pp. 618–627.

[88] Evan Evans and Ken Ritchie. "Dynamic strength of molecular adhesion bonds." In: *Biophysical Journal* 72.4 (1997), pp. 1541–1555.

[89] Rudolf Merkel et al. "Energy landscapes of receptor–ligand bonds explored with dynamic force spectroscopy." In: *Nature* 397.6714 (1999), p. 50.

[90] Evan Evans. "Probing the relation between force—lifetime—and chemistry in single molecular bonds." In: *Annual Review of Biophysics and Biomolecular Structure* 30.1 (2001), pp. 105–128.

[91] Ryoichi Kawai, Juan M R Parrondo, and Christian Van den Broeck. "Dissipation: The phase-space perspective." In: *Physical Review Letters* 98.8 (2007), p. 080602.

[92] Maria Manosas et al. "Dynamic force spectroscopy of DNA hairpins: II. Irreversibility and dissipation." In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.02 (2009), P02061.

[93] Anna Alemany and Felix Ritort. "Force-Dependent Folding and Unfolding Kinetics in DNA Hairpins Reveals Transition-State Displacements along a Single Pathway." In: *The Journal of Physical Chemistry Letters* 8.5 (2017), pp. 895–900.

[94] Lev Landau and Evgeny M Lifshitz. "Statistical physics, vol. 5." In: *Course of theoretical physics* 30 (1980).

[95] Philip Nelson. *Biological physics*. WH Freeman New York, 2004.

[96] Allen P Minton. "The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media." In: *Journal of Biological Chemistry* 276.14 (2001), pp. 10577–10580.

[97] Anthony A Hyman, Christoph A Weber, and Frank Jülicher. "Liquid-liquid phase separation in biology." In: *Annual Review of Cell and Developmental Biology* 30 (2014), pp. 39–58.

[98] Andre C Barato, David Hartich, and Udo Seifert. "Efficiency of cellular information processing." In: *New Journal of Physics* 16.10 (2014), p. 103024.

[99] Sosuke Ito and Takahiro Sagawa. "Maxwell's demon in biochemical signal transduction with feedback loop." In: *Nature Communications* 6 (2015), p. 7498.

[100] William Bialek. *Biophysics: searching for principles*. Princeton University Press, 2012.

[101] John Maynard Smith. "The concept of information in biology." In: *Philosophy of Science* 67.2 (2000), pp. 177–194.

[102] Erwin Schrödinger. "Are there quantum jumps? Part II." In: *The British Journal for the Philosophy of Science* 3.11 (1952), pp. 233–242.

[103] Antoine Bérut et al. "Experimental verification of Landauer's principle linking information and thermodynamics." In: *Nature* 483.7388 (2012), p. 187.

[104] Zheng Xie et al. "Single-molecule studies highlight conformational heterogeneity in the early folding steps of a large ribozyme." In: *Proceedings of the National Academy of Sciences* 101.2 (2004), pp. 534–539.

[105] Yuanjie Pang et al. "Optical trapping of individual human immunodeficiency viruses in culture fluid reveals heterogeneity with single-molecule resolution." In: *Nature Nanotechnology* 9.8 (2014), p. 624.

[106] Sofie L Noer et al. "Folding dynamics and conformational heterogeneity of human telomeric G-quadruplex structures in Na+ solutions by single molecule FRET microscopy." In: *Nucleic Acids Research* 44.1 (2015), pp. 464–471.

[107] Michael Hinczewski, Changbong Hyeon, and D Thirumalai. "Directly measuring single-molecule heterogeneity using force spectroscopy." In: *Proceedings of the National Academy of Sciences* 113.27 (2016), E3852–E3861.

[108] Isaiah J Fidler. "Tumor heterogeneity and the biology of cancer invasion and metastasis." In: *Cancer Research* 38.9 (1978), pp. 2651–2660.

[109] Daniel L Dexter and John T Leith. "Tumor heterogeneity and drug resistance." In: *Journal of Clinical Oncology* 4.2 (1986), pp. 244–257.

[110] Andriy Marusyk and Kornelia Polyak. "Tumor heterogeneity: causes and consequences." In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1805.1 (2010), pp. 105–117.

[111] Xavier Viader-Godoy, Maria Manosas, and Felix Ritort. "Length dependence of elastic properties of single stranded DNA." In: *In preparation* (2019).

[112] William Templeton Eadie, Daniel Drijard, and Frederick E James. "Statistical methods in experimental physics." In: *Amsterdam: North-Holland, 1971* (1971).

[113] David Roxbee Cox. *Analysis of survival data*. Routledge, 2018.

[114] Joseph A Greenwood and Marion M Sandomire. "Sample size required for estimating the standard deviation as a per cent of its true value." In: *Journal of the American Statistical Association* 45.250 (1950), pp. 257–260.

[115] Narayanaswami Balakrishnan and WS Chen. *Handbook of tables for order statistics from lognormal distributions with applications*. Springer Science & Business Media, 1999.

[116] Francisco J Cao and Manuel Feito. "Open Problems on Information and Feedback Controlled Systems." In: *Entropy* 14.4 (2012), pp. 834–847.

[117] Takahiro Sagawa and Masahito Ueda. "Nonequilibrium thermodynamics of feedback control." In: *Physical Review E* 85.2 (2012), p. 021104.

[118] Momčilo Gavrilov, Raphaël Chétrite, and John Bechhoefer. "Direct measurement of weakly nonequilibrium system entropy is consistent with Gibbs–Shannon form." In: *Proceedings of the National Academy of Sciences* 114.42 (2017), pp. 11097–11102.

[119] Charles H Bennett. "Demons, engines and the second law." In: *Scientific American* 257.5 (1987), pp. 108–117.

[120] Rolf Landauer. "Irreversibility and heat generation in the computing process." In: *IBM journal of research and development* 5.3 (1961), pp. 183–191.

[121] Charles H Bennett. "Notes on Landauer's principle, reversible computation, and Maxwell's Demon." In: *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics* 34.3 (2003), pp. 501–510.

[122] Tamir Admon, Saar Rahav, and Yael Roichman. "Experimental Realization of an Information Machine with Tunable Temporal Correlations." In: *Physical review letters* 121.18 (2018), p. 180601.

[123] Ivan Junier et al. "Recovery of free energy branches in single molecule experiments." In: *Physical Review Letters* 102.7 (2009), p. 070602.

[124] Anna Alemany, Marco Ribezzi-Crivellari, and Felix Ritort. "From free energy measurements to thermodynamic inference in nonequilibrium small systems." In: *New Journal of Physics* 17.7 (2015), p. 075009.

[125] Meyer B Jackson. *Molecular and cellular biophysics*. Cambridge University Press, 2006.

[126] Margaret Robson Wright. *An introduction to aqueous electrolyte solutions*. John Wiley & Sons, 2007.

[127] Zhi-Jie Tan and Shi-Jie Chen. "Electrostatic correlations and fluctuations for ion binding to a finite length polyelectrolyte." In: *The Journal of Chemical Physics* 122.4 (2005), p. 044903.

[128] Boris V Deraguin and Lev Landau. "Theory of the stability of strongly charged lyophobic sols and of the adhesion of strongly charged particles in solution of electrolytes." In: *Acta Physicochim: USSR* 14 (1941), pp. 633–662.

[129] Evert Johannes Willem Verwey, J Th G Overbeek, and Jan Theodoor Gerard Overbeek. *Theory of the stability of lyophobic colloids*. Courier Corporation, 1999.

[130] Anna Pyle. "Metal ions in the structure and function of RNA." In: *JBIC Journal of Biological Inorganic Chemistry* 7.7-8 (2002), pp. 679–690.

[131] Jan Lipfert et al. "Understanding nucleic acid–ion interactions." In: *Annual Review of Biochemistry* 83 (2014), pp. 813–841.

[132] Li-Zhen Sun, Dong Zhang, and Shi-Jie Chen. "Theory and modeling of RNA structure and interactions with metal ions and small molecules." In: *Annual Review of Biophysics* 46 (2017), pp. 227–246.

[133] David E Draper. "A guide to ions and RNA structure." In: *Rna* 10.3 (2004), pp. 335–343.

[134] Ignacio Tinoco Jr and Carlos Bustamante. "How RNA folds." In: *Journal of Molecular Biology* 293.2 (1999), pp. 271–281.

[135] Susan L Heilman-Miller, D Thirumalai, and Sarah A Woodson. "Role of counterion condensation in folding of the Tetrahymena ribozyme. I. Equilibrium stabilization by cations1." In: *Journal of Molecular Biology* 306.5 (2001), pp. 1157–1166.

[136] Susan L Heilman-Miller et al. "Role of counterion condensation in folding of the Tetrahymena ribozyme II. Counterion-dependence of folding kinetics1." In: *Journal of Molecular Biology* 309.1 (2001), pp. 57–68.

[137] Zhi-Jie Tan and Shi-Jie Chen. "Predicting electrostatic forces in RNA folding." In: *Methods in Enzymology*. Vol. 469. Elsevier, 2009, pp. 465–487.

[138] Cristiano V Bizarro, Anna Alemany, and Felix Ritort. "Non-specific binding of Na+ and Mg 2+ to RNA determined by force spectroscopy methods." In: *Nucleic Acids Research* 40.14 (2012), pp. 6922–6935.

[139] Josep Maria Huguet et al. "Derivation of nearest-neighbor DNA parameters in magnesium from single molecule experiments." In: *Nucleic acids research* 45.22 (2017), pp. 12921–12931.

[140] Sultan C Agalarov et al. "Structure of the S15, S6, S18-rRNA complex: assembly of the 30S ribosome central domain." In: *Science* 288.5463 (2000), pp. 107–112.

[141] Carl R Woese and George E Fox. "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." In: *Proceedings of the National Academy of Sciences* 74.11 (1977), pp. 5088–5090.

[142] Alexei Nikulin et al. "Crystal structure of the S15–rRNA complex." In: *Nature Structural and Molecular Biology* 7.4 (2000), p. 273.

[143] Robert T Batey and James R Williamson. "Effects of polyvalent cations on the folding of an rRNA three-way junction and binding of ribosomal protein S15." In: *Rna* 4.8 (1998), pp. 984–997.

[144] Liliana R. Stefan. "MeRNA: a database of metal ion binding sites in RNA structures." In: *Nucleic Acids Research* 34.90001 (2006), pp. D131–D134. ISSN: 1362-4962. DOI: 10.1093/nar/gkj058. URL: http://dx.doi.org/10.1093/nar/gkj058.

[145] Harold D Kim et al. "Mg2+-dependent conformational change of RNA studied by fluorescence correlation and FRET on immobilized single molecules." In: *Proceedings of the National Academy of Sciences* 99.7 (2002), pp. 4284–4289.

[146] Rebecca J Case et al. "Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies." In: *Applied and Environmental Microbiology* 73.1 (2007), pp. 278–288.

[147] Jan Liphardt et al. "Reversible unfolding of single RNA molecules by mechanical force." In: *Science* 292.5517 (2001), pp. 733–737.

[148] Maria Manosas, Ivan Junier, and Felix Ritort. "Force-induced misfolding in RNA." In: *Physical Review E* 78.6 (2008), p. 061925.

[149] Stanley B Prusiner. "Molecular biology of prion diseases." In: *Science* 252.5012 (1991), pp. 1515–1522.

[150] Ivo L Hofacker. "Vienna RNA secondary structure server." In: *Nucleic Acids Research* 31.13 (2003), pp. 3429–3431.

[151] Ryota Yamagami et al. "Cellular conditions of weakly chelated magnesium ions strongly promote RNA stability and catalysis." In: *Nature Communications* 9.1 (2018), p. 2149.

[152] Vinod K Misra and David E Draper. "The linkage between magnesium binding and RNA folding." In: *Journal of Molecular Biology* 317.4 (2002), pp. 507–521.

[153] Nina M Fischer et al. "Influence of Na+ and Mg2+ ions on RNA structures studied with molecular dynamics simulations." In: *Nucleic Acids Research* 46.10 (2018), pp. 4872–4882.

[154] Natalia A Denesyuk, Naoto Hori, and Devarajan Thirumalai. "Molecular Simulations of Ion Effects on the Thermodynamics of RNA Folding." In: *The Journal of Physical Chemistry B* 122.50 (2018), pp. 11860–11867.

[155] Steven B Smith, Yujia Cui, and Carlos Bustamante. "Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules." In: *Science* 271.5250 (1996), pp. 795–799.

[156] Alessandro Fiasconaro and Fernando Falo. "Exact analytical solution of the extensible freely jointed chain model." In: *arXiv preprint arXiv:1805.01499* (2018).

[157] Claude Bouchiat et al. "Estimating the persistence length of a worm-like chain molecule from force-extension measurements." In: *Biophysical Journal* 76.1 (1999), pp. 409–413.

[158] Rafayel Petrosyan. "Improved approximations for some polymer extension models." In: *Rheologica Acta* 56.1 (2017), pp. 21–26.

[159] Theo Odijk. "Stiff chains and filaments under tension." In: *Macromolecules* 28.20 (1995), pp. 7016–7018.

[160] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[161] Donald Ervin Knuth. *The art of computer programming*. Vol. 3. Pearson Education, 1997.

[162] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Vol. 55. Courier Corporation, 1965.