Comparative Analysis of Nucleotide Translocation through Protein Nanopores using Steered Molecular Dynamics and an Adaptive Biasing Force

Hugh S.C. Martin,[†] Shantenu Jha,[*,‡] and Peter V. Coveney[*,†]

Centre for Computational Science and Department of Chemistry, UCL, 20 Gordon Street, London, UK, and Department of Electrical Engineering Building, 94 Brett Road, Piscataway, New Jersey, USA

E-mail: shantenu.jha@rutgers.edu; p.v.coveney@ucl.ac.uk

Abstract

The translocation of nucleotide molecules across biological and synthetic nanopores has attracted attention as a next generation technique for sequencing DNA. Computer simulations have the ability to provide atomistic-level insight of important states and processes, delivering a means to develop a fundamental understanding of the translocation event, for example, by extracting meaningful quantities such as the free energy of the process. Even on modern supercomputing facilities, the simulation of many-atom systems in fine detail is limited to shorter timescales than the real events they attempt to recreate. This imposes the need for enhanced simulation techniques that expand the scope of investigation in a given timeframe. In the case of nucleotide translocation, such a technique would, for example, increase the speed of translocation. There are numerous free energy calculation and translocation methodologies available,

[*]To whom correspondence should be addressed
[†]University College London
[‡]Rutgers University

and it is by no means clear which method is best applied to a particular problem. This paper explores the use of two popular free energy calculation methodologies in a nucleotide-nanopore translocation system, using the α-hemolysin nanopore. The first uses constant velocity-steered molecular dynamics in conjunction with Jarzynski's Equality. The second applies an adaptive biasing force, which has not previously been applied to the nucleotide-nanpore system. The purpose of the present study is to provide a unique, detailed, and comprehensive comparison of the two methodologies, allowing for a detailed comparative assessment of the scientific merits, the computational cost, and the statistical quality of the data obtained from each technique. We find that the adaptive biasing force method produces results that are closer to experimental measurements than with constant velocity-steered molecular dynamics, while the net errors are smaller for the same computational cost. We determine that the advantages of the adaptive biasing force methodology are likely to be applicable to a significant number of systems studied using molecular dynamics.

1 Introduction

The translocation of nucleic acid strands through confined protein pores has substantial biological relevance, for example the transfer of antibiotic resistance genes between bacteria,[1–3] phageinfection ,[4] and the uptake of oligonucleotides into kidney tissue.[5] Moreover, the passage of nucleic acids through pores is also of biotechnological and diagnostic relevance; for these applications, a single nanopore is inserted into a lipid bilayer, and individual negatively charged nucleic acids are electrophoretically driven through the pore. The passage of strands leads to detectable fluctuations in the ionic pore current. Data from these single channel current recording (SCCR) experiments provide information on polymer length, orientation and composition for polymers such as single-stranded DNA and RNA.[6–11] The capability of SCCR to reveal information on translocating DNA strands has long been under investigation as an avenue for faster and cheaper genetic sequencing.[12] In recent years, it has been demonstrated that SCCR has sequencing capabilities,[13,14] and in February 2012 Oxford Nanopore Technologies demonstrated a fully functional genetic sequencing

device, due to be available commercially in 2012.[15]

Understanding the microscopic processes of nucleic acid translocation through nanopores is crucial in improving SCCR techniques and apparatus for sequencing DNA. Using molecular dynamics (MD) simulations of the translocation process, it is possible to retrieve kinetic and structural information that cannot be obtained solely through experiment. Experiments investigating the translocation of nucleic acid under the influence of a transmembrane potential indicate that the process typically takes hundred of microseconds to tens of milliseconds.[7] But accurately simulating biological processes and systems with atomistic resolution remains a challenge for many reasons, not least of which are the substantial computational resources required. Even with state-of-the-art high-end computers, performing simulations with atomistic resolution for such large systems over the required time-scales remains infeasible at present. Simple approaches to circumventing this issue can give rise to undesirable consequences, for example, the application of an artificially high transmembrane potential to induce faster translocation causes disruption of the lipid membrane; applying a high uniform electrostatic field to only the translocating atoms fails to translocate nucleic acid polymers through the protein nanopore.[16] Thus, if these events are to be effectively investigated using simulation, novel approaches and better algorithms are required in order to bridge the gap between time-scales over which the translocation events occur and those that are accessible using simple equilibrium simulations.

It is desirable to examine the free energy behaviour of chemical processes in order to fully understand them. By computing the free energy difference associated with a change of state, it is often possible to establish stable states, their thermodynamics properties, the kinetics of transitions between states, and indeed to infer how stable states are altered by external conditions. Such changes of state include protein mutation, protein-ligand binding, conformational changes, and molecule translocation. It is, of course, both possible and valuable to calculate experimental free energy changes, and there has recently been a considerable amount of research dedicated to measuring free energy changes in simulated systems.

There are several well established methods for extracting free energy from MD simulations.

These include history-dependent methods such as metadynamics,[17] self-healing umbrella sampling,[18] and the adaptive biasing force method (ABF),[19] which can bias a translocating molecule along a reaction coordinate. Other methods such as constant velocity-steered molecular dynamics (cv-SMD) or contact force-steered molecular dynamics[20] may be used to entice a molecule along a reaction coordinate, based on the behaviour of which free energy calculation methods such as Jarzynski's equality (JE)[21] or Crooks fluctuation theorem[22] may be used to extract the free energy.

Cv-SMD/JE and the ABF methodology are two well-established and widely used translocation/free energy calculation methods that serve as exemplary methodologies for the purposes of such a comparison. The methodologies have key similarities, yet important differences in their "dynamics". It is the aim of this paper to explore these similarities and differences. We believe that conclusions from this investigation can be extrapolated to many other translocation and free energy calculation methods.

In cv-SMD, the translocating molecule of interest is attached via a harmonic spring to a point in space that is pulled at constant velocity. Using the force experienced by the spring, the free energy of translocation may be determined using JE to equate the free energy to the work done.

In the ABF methodology, the translocating molecule of interest is encouraged along the reaction coordinate by inserted a biasing force into the equations of motion for an atom or group of atoms in the molecule. This biasing force opposes the free energy estimate for a section of the reaction coordinate and is calculated using the instantaneous forces acting on the atom(s) in question.

A major benefit of algorithms such as cv-SMD and ABF is that they permit larger and/or more complex systems to be investigated using a given computational budget (comprising the hardware and computational hours available). It is therefore pertinent to choose a system of considerable size and complexity for this study, as the behaviour of the algorithms at these limits has been hitherto unclear. The system should also have experimental or biological relevance, in order that we may draw comparisons with experimental data, and any insight we gain may have relevance to other studies and future research

The translocation system we have chosen to investigate is the passage of nucleotide molecules through the protein nanopore α-hemolysin (αHL), depicted in Figure 1. αHL is a heptameric protein-pore that has been extensively studied in experiments and computer simulations.[10,11,16,23–29] We explore the protein-pore translocation of the nucleic acid strands polyadenosine, poly(A), and polydeoxycytidine, poly(dC), which are single strands of RNA and DNA respectively. Poly(A) and poly(dC) molecules of 100-200 bases in length exhibit a 20-fold difference in translocation time through αHL in SCCR experiments.[7] We also translocate single nucleotides A_1 and dC_1 to discern their relative contributions to the free energy profiles.

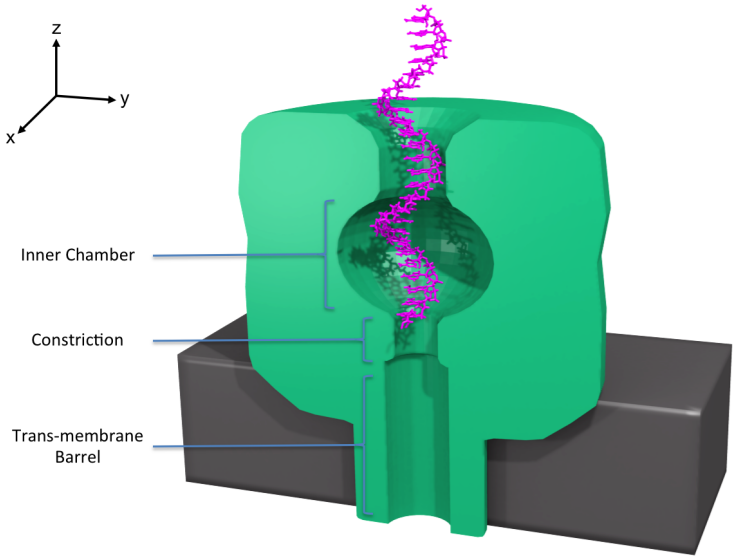

Figure 1: Figure representing a cross-section of the protein pore αHL, and the starting configuration for the simulations studied here. The heptameric αHL protein pore (green) is inserted into a lipid bilayer (black). The *cis*-entrance at the top of the protein pore is about 28 Å in diameter and the *trans*-entrance at the bottom of the pore is about 20 Å. Key features inside the pore interior include the wide inner chamber (up to 46 Å wide), a constriction about half way down the pore (14 Å wide), followed by the trans-membrane barrel (20 Å wide) that spans the lipid bilayer. The translocating molecule, in this example a polynucleotide (pink), is positioned with the 3′-end at the top of the constriction.

We recently published an investigation into the nucleotide-nanopore system using cv-SMD/JE.[30] The study applied the cv-SMD translocation technique in a system of unprecedented size, revealing new insight into the translocation process. In that study, we identified the existence and signifi-

cance of a phosphate-lysine interaction, Bond *et al.* have since verified this interaction in a separate study.[31] They had performed nucleotide translocation simulations through a simplified αHL pore using an applied transmembrane potential and determined that the phosphate-lysine interaction plays a major role.

In this paper, we use the ABF methodology to investigate the nucleotide-nanopore system and provide a comprehensive comparison of the two methodologies. By performing simulations using cv-SMD/ JE and ABF under comparable conditions, we are able to make direct comparisons of the data quality and associated errors, the modes of translocation, the free energy calculations and the computational resources that each method requires.

In Section 2 we describe the theoretical background of the cv-SMD and ABF methodologies. In Section 3 we provide details of the model and techniques used to perform our simulations. In Section 4 and Section 5 we present analyses of simulations of single and polynucleotide translocation through wild type αHL, for cv-SMD and ABF respectively. In Section 6 we compare cv-SMD/JE to ABF for the nucleotide-nanopore system. Finally, in Section 7 we present our conclusions.

2 Theoretical Background

In this section we provide an account of the theoretical background to the cv-SMD/JE and ABF translocation methodologies and their associated means of determining the free energy of translocation.

2.1 Constant Velocity-Steered Molecular Dynamics

SMD provides a means of exploring longer translocation trajectories on a smaller timescale while retaining atomistic detail. This method is inherently non-equilibrium in nature. Such non-equilibrium approaches are often used instead of reducing the detail of the model to a coarse-grained one in order to increase the scope of the investigation. MD programs such as NAMD[32] allow simulated processes to be steered by introducing non-equilibrium forces. In SMD, an atom can have a di-

rectional force applied to it, causing the atoms (and any other particles coupled to them) to move with the direction of the force. The consequence of this applied force is that a significant degree of movement can be induced in a relatively short timeframe, and the full atomistic dynamics of the process can be examined.

By applying SMD to threading nucleic acid chains through nanopores, the translocation process can be replicated in an atomistically detailed model within a timeframe that is feasible to simulate. By pulling the nucleic acid strand at constant velocity (known as constant velocity steered MD or cv- SMD[32]), a translocation of known distance can be performed. In cv-SMD, an atom or the centre of a group of atoms is harmonically restrained to a point in space that is shifted in the chosen direction. The harmonic restraint can be thought of as a spring attached to a dummy atom, the strength of the restraint being given by a force constant k. The simulation outputs the force in pico-newtons (pN) experienced by the spring in the direction of pulling (the reaction coordinate).

We want to convert the force output of the simulation to the free energy of translocation, giving free energy profiles which can tell us about energetic barriers to translocation. Free energy, however, is an equilibrium property while cv-SMD is a non-equilibrium technique. To overcome this, Jarzynski's Equality[21] can be used. JE equates equilibrium free energy to the ensemble average of the exponential work from a non-equilibrium process. The work can be calculated from the force output using a force-distance integral.

Consider a process changing a parameter, λ, of a system at an initial equilibrium state at time zero, to a final state at time t. The second law of thermodynamics states that the average work done during a process, $\langle W \rangle$, on the system cannot be smaller than the Helmholtz free-energy difference between the initial and final states of λ:

$$\Delta F = F(\lambda_t) - F(\lambda_0) \leq \langle W \rangle \tag{1}$$

Here, and subsequently, angular brackets $\langle \cdots \rangle$ imply an ensemble average in the statistical mechanical sense. Jarzynski determined that the free-energy difference between the two states can be related to the work, W, by the following equality:

$$e^{-\beta\Delta F} = \left\langle e^{-\beta W} \right\rangle \tag{2}$$

Here, β is the inverse temperature, $(1/k_BT)$, where k_B is Boltzmann's constant.

2.2 Adaptive Biasing Force

Adaptive biasing force is an advanced form of umbrella sampling (US) which estimates the free energy landscape during the sampling simulation and applies biasing forces which allow for effective sampling. The algorithm used to determine the adaptive biasing force was developed by Darve and Pohorille [19,33] and implemented within the NAMD code by Hénin and Chipot.[34]

In US, an external potential is applied to allow for the exploration of higher energy configurations and the states that they separate. Here, the free energy along a chosen reaction coordinate (ξ) is given by:

$$F(\xi) = \frac{1}{\beta} ln \mathscr{P}_\xi - \mathscr{U}_{bias} + F_0 \tag{3}$$

where $F(\xi)$ is the free energy of the state at a particular value of ξ; $\mathscr{P}_\xi$ is the probability density of finding the system at ξ; $\mathscr{U}_{bias}$ the applied external potential; β the inverse temperature, and F_0 is a constant. If $\mathscr{U}_{bias}$ is tuned to be the exact opposite of the free energy, $-F(\xi)$, then the free energy landscape will effectively be flattened out. This then permits uniform sampling along the reaction coordinate. Knowing the landscape of $-F(\xi)$ implies an *a priori* knowledge of the free energy landscape. If the $\mathscr{U}_{bias}$ deviates from $-F(\xi)$ then the reaction coordinate will be poorly sampled. Thus, when using US for unfamiliar systems, good sampling of the reaction coordinate is very difficult to realise.

Adaptive biasing force (ABF) is a method of reaction coordinate sampling which applies a continuous biasing force, which is tuned during the simulation to the cumulative estimate of the free energy landscape. Uniform sampling can thus be achieved by on-the-fly calculation of $F(\xi)$ and the implementation of this information in the form of an external bias. The implementation of

ABF, therefore, requires no knowledge of the free energy landscape prior to the simulation.

The free energy in an ABF simulation can be calculated from the forces acting on the biased atom. The derivative of the free energy of translocation with respect to the reaction coordinate can be expressed in terms of configurational averages at constant ξ:

$$\frac{dF(\xi)}{d\xi} = \left\langle \frac{\partial \mathscr{V}(x)}{\partial \xi} - \frac{1}{\beta}\frac{\partial ln|J|}{\partial \xi} \right\rangle_{\xi} = -\left\langle f_{\xi} \right\rangle_{\xi}. \tag{4}$$

The $\partial \mathscr{V}(x)/\partial \xi$ term represents the physical forces exerted on the system, as derived from the potential energy function, $\mathscr{V}(x)$. The $(1/\beta)(\partial ln|J|/\partial \xi)$ term represents a geometric correction where $|J|$ is the determinant of the Jacobian for the transformation from generalised to Cartesian coordinates; it accounts for the difference in phase space availability as the reaction coordinate varies. $\left\langle f_{\xi} \right\rangle_{\xi}$ is the average force acting along reaction coordinate ξ, derived from instantaneous force components f_{ξ}.

The biasing force of the ABF method is calculated as the negative of the average force acting along reaction coordinate ($\left\langle f_{\xi} \right\rangle_{\xi}$) derived from instantaneous force components (f_{ξ}). For its practical implementation, f_{ξ} is accumulated in small bins along the reaction coordinate ($\Delta\xi$) to provide an estimate of the change in free energy with respect to the reaction coordinate, $dF(\xi)/d\xi$. The adaptive biasing force, $\mathbf{f}^{ABF}$, is thus defined as:

$$\mathbf{f}^{ABF} = \frac{dF(\xi)}{d\xi} = -\left\langle f_{\xi} \right\rangle_{\xi}, \tag{5}$$

where $F(\xi)$ is the current estimate of the free energy as a function of the reaction coordinate. The calculated biasing force, $\mathbf{f}^{ABF}$, is introduced into the equations of motion during a simulation. With the biasing force applied, the overall forces acting on the biased atom(s) along the reaction coordinate within a bin average to zero over time. This is because the biasing force approximately opposes the energetic barriers to translocation. As the applied biasing force roughly matches the free energy landscape, evolution along the reaction coordinate is largely governed by a process of self-diffusion.

The instantaneous force, f_ξ, fluctuates to a high degree; thus the average force will initially take inaccurate and physically meaningless values. Therefore, the biasing force is not implemented within a particular bin until a threshold of force measurements has been accumulated; the biasing force is then introduced gradually via a linear ramp. The threshold value, ζ, is an input parameter in units of timesteps and its optimal value is dependent on the nature of the system. A higher value of ζ will help in the reduction or removal of excessive non-equilibrium effects; therefore, it will need to be larger for systems with slowly relaxing degrees of freedom, for example, those systems with large, flexible translocating molecules. ζ heavily influences the translocation/simulation time, though it is not the only contributing factor due to the dependence of the motion of the biased atom(s) on self-diffusion. This process is described in Figure 2.

To keep the atom(s) within the bounds of the reaction coordinate, harmonic boundary forces are applied at either end. The position of the boundaries and the value of their force constants are selected as simulation input parameters. The boundary forces are rigidly implemented and may influence the biased atom when it is near to them,[35] therefore it is beneficial in this respect to ensure that the distance between the boundaries is as long as possible. However, as we shall see later in this paper, there are benefits to segmenting the full reaction coordinate into several shorter intervals.

The ABF reaction coordinate is often defined in terms of relative atomic positions, that is, the distance between selected atoms or groups of atoms. In the latest implementation of ABF in NAMD, the reaction coordinate may be defined in various ways, including bond angles, radii, and RMSD values.[36] For the studies in this paper, the reaction coordinate is most appropriately designated in terms of inter-atomic distances in the z-axis direction. Since the reaction coordinate is relative to two atoms or groups of atoms, the Cartesian coordinates of the moving and reference atoms or groups must be converted into generalised coordinates, as outlined in Section 2.2.

By imposing a biasing force on the molecule of interest, and simultaneously allowing the free energy of translocation to be calculated, ABF presents itself as an ideal candidate for the nucleotide-nanopore translocation system.

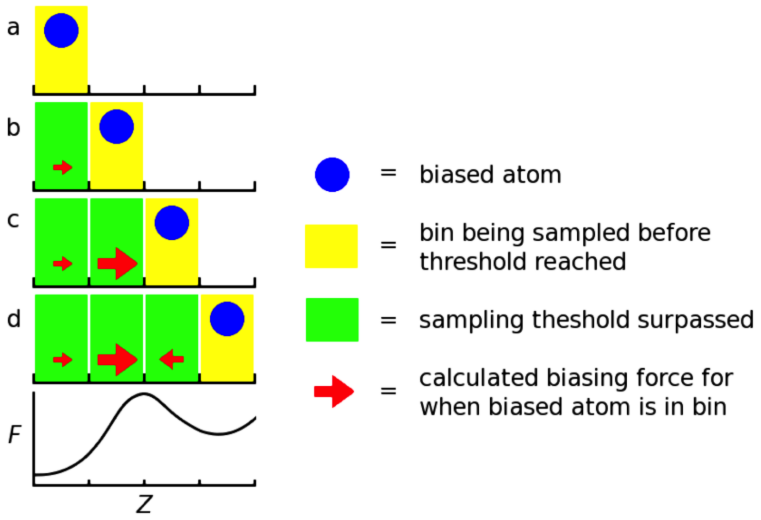

Figure 2: Illustration of the ABF methodology. The horizontal axes represent the translocation reaction coordinate, which is the distance along the pore. The reaction coordinate is discretised into bins of equal size. The ABF methodology can be summarised through the time sequence represented by illustrations (a)-(d) and the free energy profile at the bottom. (a) The biased atom (blue circle) is within the first bin of the reaction coordinate. The biasing force is being estimated (yellow rectangle). (b) The biased atom has been sampled more than the threshold, ζ, in the first bin and the biasing force (red arrow) has allowed the atom to move into the subsequent bin. (c) The biased atom has been sampled more than the threshold ζ in the second bin and has moved to the subsequent bin. The biasing force to move the atom from the second bin was larger than in the first, due to the free energy landscape. (d) The biased atom has reached the end of the reaction coordinate; it may continue to diffuse backwards across the reaction coordinate if the simulation continues. In this illustration, the atom has tended to move forward at each stage, but it may also diffuse in the reverse direction at any time, since the biasing force is intended to match rather than exceed the free energy barriers.

3 Method

Martin *et al.* exhaustively described the details of the model construction and simulation parameters.[30] The cv-SMD method section described there[30] is applicable to the cv-SMD simulations in this paper; therefore only an overview of this method, along with some additional points of note, will be described here.

For the ABF simulations reported in this paper, the majority of the parameters and model construction from the cv-SMD method also apply, with some key exceptions. In this section we describe and justify the ABF-specific parameters that we have chosen and explored.

The αHL crystallographic structure coordinates were taken from Protein Data Bank (PDB) entry 7AHL. The protein was inserted into a patch of 150 Å x 150 Å pre-equilibrated and solvated 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayer using the VMD plug-in *membrane*, aligned to the *xy*-plane plane. The centre of mass of the hydrophobic belt of αHL (residues 118-126 and 132-142) was aligned with the centre of mass of the lipid bilayer. The system was solvated in a water box of pre-equilibrated water molecules and the aqueous solution was set at 1M NaCl. Figure 1 shows αHL inserted in a lipid membrane as it appears in our models. The protonation states chosen are consistent with the typical SCCR recording pH range of around pH 8.0.[6,11] Key protonation states include: protonated, positively charged amine groups of lysine and arginine residues; unprotonated, negatively charged inter-chain phosphate groups; and unprotonated, doubly negatively charged terminal phosphate groups on the single nucleotide molecules.

The poly(A) and poly(dC) molecules were constructed using the AMBER module *nucgen*[37] to 25 bases in length. Single nucleotide PDB files of adenosine (A_1) and deoxycytidine (dC_1) monophosphates were obtained from the Protein Data Bank (PDB identifiers AMP and DCM respectively). The topology files were modified accordingly to produce accompanying PSF files. The final models consisted of 328,000 and 262,000 atoms for the 25-base polynucleotide and single nucleotide models respectively. The nucleotide molecules were orientated with the C3′-carbon atom of the leading residue was aligned with the centre of the alpha carbon atoms (C_α) of protein residue 111. The nucleotide molecules were pulled or biased from this starting position towards the *trans*-entrance of the pore. A partial translocation of the leading residue through the constriction was performed to maximise the number of translocation samples performed given a finite computational budget, Martin *et al.* justify this selection of reaction coordinate in detail.[30] An example of the starting position of the polynucleotides is shown in Figure 1.

Simulations were performed using the molecular dynamics simulation package NAMD version 2.71b.[32] The CHARMM[38] forcefield was applied using all-hydrogen parameter files for CHARMM22 proteins and CHARMM27 lipids and nucleic acids.

In order to gather a set of samples to form an ensemble, multiple simulations of the nucleic acid molecule translocating past the same section of the pore were required. The initial configurations used to perform these translocation samples were obtained by capturing snapshots of the atomic positions and velocities, separated by 0.2 ns at equilibrium, with the SMD atom position fixed and the C_α protein atoms restrained.

Unless otherwise stated, harmonic constraints of 0.5 N/m were placed on the C_α atoms of the protein amino acid residues in order to prevent translocation of the protein. In cv-SMD simulations, this allows the reaction coordinate to indicate specific protein-nucleic acid interactions. In ABF simulations of the type described in this paper, the relationship between the reaction coordinate and the pore interior is maintained regardless of shifts in the protein's location.

3.1 cv-SMD Method

Translocation was limited to 1 ns per simulation due to resource constraints and to control binning errors. An overlap of 0.2 ns between sequential simulations was performed to enable removal of startup artefacts. The SMD atom was pulled at 0.04 Å/ps and the SMD spring constant was set to 100 kcal/mol, which Martin *el al.*[30] established as suitable for these molecular models.

3.2 ABF Method

The biasing force was applied to the C3′-atom of the leading residue. The adaptive biasing force implemented using the *colvar* module.[39] Force measurements were accumulated in bins of 0.25 Å (unless otherwise stated) for 16 Å length trajectories. The reaction coordinate in the ABF methodology was calculated as a function of distance from the translocating molecule to a reference set of atoms in the protein-pore, in contrast to cv-SMD. This relative definition of the reaction coordinate allows for the protein to be left unconstrained in ABF simulations. Unless otherwise stated, the simulations reported in this section were performed with an unconstrained αHL pore.

While only the z-axis separation was controlled by the biasing force, the steric constraints of the pore interior were sufficient to keep each sample trajectory within the desired xy-boundaries. The

biased atom was kept within the outer z-axis boundaries of the reaction coordinate by a harmonic force implemented at either end.

The length of the reaction coordinate was set to 16 Å, spanning the length of the constriction. It is possible to split reaction coordinates into segments and construct free energy profiles from each of the segments. Splitting the reaction coordinate can help prevent the biased atom getting stuck. However, this is not necessary with a reaction coordinate length as short as 16 Å. Furthermore, introducing too many segments can cause the harmonic restraints at the ends of each segment to significantly impact the free energy values, and so it should be implemented with caution. The number of simulated timesteps required to sample the reaction coordinate depends on the force measurement threshold parameter and the diffusion time, which varies between simulations. The simulations were therefore performed in blocks of 100,000 to 1 million MD integration timesteps until the full reaction coordinate was sampled. The width of the bins along the reaction coordinate and the force measurements threshold parameter (ζ) are investigated in the next section.

3.3 Summary of the Models and Simulations Performed

This subsection summarises the key configuration details in the simulations represented in this paper. This is presented in the form of a table (Table 1), in order for the reader to be able to quickly refer to and understand our data, particularly when comparisons are being drawn between multiple figures.

4 Adenine and Deoxycytosine Translocation Using cv-SMD

In our previous study,[30] we used single nucleotides and polynucleotides in wild type and mutated αHL nanopores to gain insight into the translocation process. We found that a phosphate-lysine electrostatic interaction at the pore constriction played a key role in translocation, proving its significance by mutating the lysine residue in question, which significantly impacted the free energy profiles. The extent to which this interaction occurred for a particular nucleotide molecule was

Table 1: Table listing key components of the simulated systems from the profiles in this paper.

Configuration Name	Pulling Method	Nucleotide Base	Nucleotide Bases	Samples Performed	Protein Constraints	ABF Threshold	Translocation Distance (Å)	Figure Number
A_{25}-cvSMD-48Å	cv-SMD	Adenine	25	16	Constrained	N/A	48	Figure 3
dC_{25}-cvSMD-48Å	cv-SMD	Deoxycytosine	25	16	Constrained	N/A	48	
A_1-cvSMD	cv-SMD	Adenine	1	16	Constrained	N/A	16	Figure 4
dC_1-cvSMD	cv-SMD	Deoxycytosine	1	16	Constrained	N/A	16	
A_{25}-ABF-uncon	ABF	Adenine	25	4	Unconstrained	5,000	16	Figure 5(a)
dC_{25}-ABF-uncon	ABF	Deoxyctosine	25	4	Unconstrained	5,000	16	
A_{25}-ABF-con	ABF	Adenine	25	4	Constrained	5,000	16	Figure 5(b)
dC_{25}-ABF-con	ABF	Deoxyctosine	25	4	Constrained	5,000	16	
A_{25}-ABF-20k	ABF	Adenine	25	4	Unconstrained	20,000	16	Figure 6
dC_{25}-ABF-20k	ABF	Deoxycytosine	25	4	Unconstrained	20,000	16	
A_1-ABF-5k	ABF	Adenine	1	4	Unconstrained	5,000	16	Figure 7
dC_1-ABF-5k	ABF	Deoxycytosine	1	4	Unconstrained	5,000	16	
A_{25}-cvSMD-16Å-4s	cv-SMD	Adenine	25	4	Constrained	N/A	16	Figure 8(a)
dC_{25}-cvSMD-16Å-4s	cv-SMD	Deoxycytosine	25	4	Constrained	N/A	16	
A_{25}-ABF-16Å-4s	ABF	Adenine	25	4	Constrained	5,000	16	Figure 8(b)
dC_{25}-ABF-16Å-4s	ABF	Deoxycytosine	25	4	Constrained	5,000	16	

highlighted as a potential cause for the discrimination of poly(A) and poly(dC) translocation. With a demonstrated dependence of the interaction on local solvation ionic environments, it was deemed necessary to increase the sampling of the reaction coordinate to give dependable insight.

In this section, we extend our previous investigation by comparing the free energy profiles with significantly greater sampling for A_{25}, dC_{25}, A_1 and dC_1 translocation through wild type αHL using cv-SMD. By providing a set of highly sampled profiles in this way, we can use the dataset as a reference point for the validation of new data that has not been permitted the same sampling budget.

4.1 Polynucleotides in Wild Type αHL

The translocation of poly(A) and poly(dC) is shown in Figure 3. The figure shows the free energy profiles for a translocation over 48 Å for A_{25} and dC_{25} with 16 samples used for the calculation of each profile. Here, the SMD atom at the $3'$-end of the nucleic acid polymer was pulled from the

top of the constriction to the bottom of the transmembrane barrel. Given the pore dimensions, as listed in Figure 1, the steric barriers to translocation occur mainly within this region.

The free energy plots from Figure 3 show that A_{25} displays a higher free energy profile than dC_{25}, with non-overlapping error bars from 11 Å onwards. The separation between the profiles continues to grow throughout the translocation process with the the free energy estimate for A_{25} being ~30% higher than that of dC_{25} at the end of the 48 Å reaction coordinate. The higher free energy values for A_{25} compared to dC_{25} is in qualitative agreement with the longer experimental translocation times for A_{25}.[7]

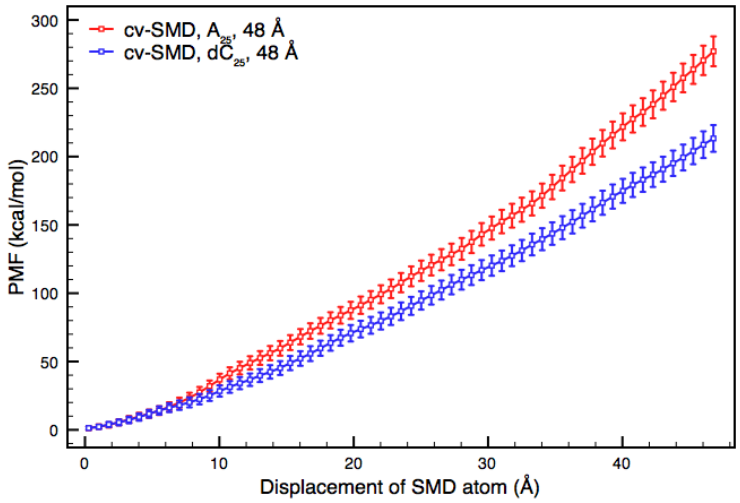

Figure 3: Free energy profiles of A_{25} and dC_{25} translocation from a set of cv-SMD simulations. The reaction coordinate spans 48 Å from the top of the constriction to the bottom of the *trans*-entrance of wild type αHL. Each profile was derived from 16 samples, calculated using a bin width of 0.75 Å. The free energy estimate for A_{25} is ~30% higher than that of dC_{25} at the end of the 48 Å reaction coordinate. The plots show discrimination of A_{25} and dC_{25} with non-overlapping error bars after 11 Å of translocation.

4.2 Single Nucleotides in Wild Type αHL

Using single nucleotide translocation simulations, we can obtain a clear picture as to what kind of molecular interactions give rise to energy barriers to translocation. This is because the contributions to translocation barriers are reduced to those attributable to the small molecule, whose size

and relative simplicity make it straight forward to inspect visually. With a polymeric molecule, numerous steric and electrostatic interactions occur along its length, making it difficult to identify major points of interest. By comparing the single nucleotide to polynucleotide translocation, we can also infer the degree to which non-equilibrium effects impact the polynucleotide free energy profiles.

Figure 4 shows the free energy profiles from A_1 and dC_1 translocation through wild type αHL. Our previous study showed that an electrostatic interaction between the nucleotide phosphate (negatively charged) and the protein lysine 147 (positively charged) skewed the values of these single nucleotide profiles in unexpected ways. The result of this is higher free energy values in the dC_1 profile due a particularly strong phosphate-lysine contribution. The consequence of the upshifted dC_1 profile is the barely distinguishable A_1 and dC_1 profiles shown in Figure 4.

The phosphate-lysine interaction and the small size of the single nucleotide molecule also gives rise to the distinct profile shape we see in Figure 4. Compared to the polynucleotide profiles, which exhibit a relatively consistent gradient throughout the reaction coordinate, both single nucleotide profiles exhibit a distinctive curve. The single nucleotide profiles show a rapid rise of gradient after 4 Å, then a significant reduction in the gradient after 10 Å, effectively levelling off. This shape corresponds to steric and electrostatic interactions reaching a maximum between 4 and 10 Å of translocation, after which the nucleotide molecule exits the constriction into the wider and uncharged transmembrane barrel, offering little resistance to ongoing translocation.

5 Adenine and Deoxycytosine Translocation Using ABF

In this section, we employ ABF as an alternative means of investigating the translocation process. Since the ABF method has not been used for this system previously, it is important to fully establish the optimum parameters, and validate the results in comparison to the heavily sampled cv-SMD data, as well as in comparison to experimental findings. We first explore the interesting option of leaving the protein completely unconstrained, which is permitted due to the nature of the

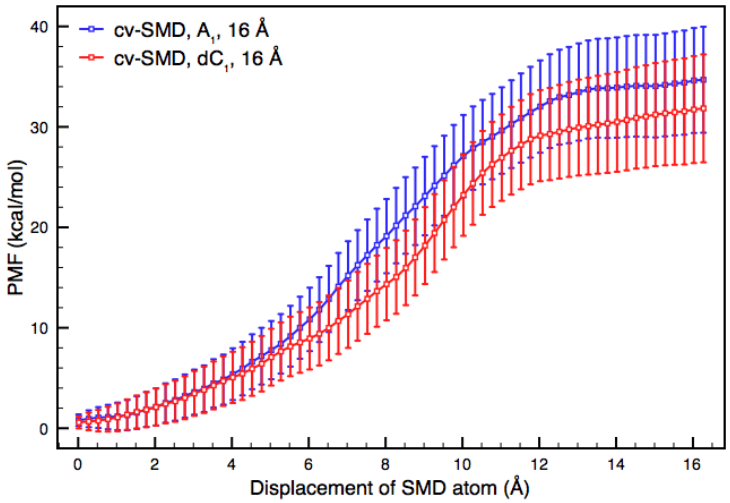

Figure 4: Free energy profiles of A_1 and dC_1 translocation from a set of cv-SMD simulations. The reaction coordinate spans 16 Å through the constriction of wild type of wild type αHL. Each profile was derived from 16 samples, calculated using a bin width of 0.25 Å. The two free energy profiles do not show discrimination outside of the error bars. Compared to the polynucleotide profiles, the single nucleotide profiles exhibit a distinctive shape, showing a rapid rise of gradient after 4 Å, then a large reduction in gradient after 10 Å, effectively levelling off. This corresponds to steric and electrostatic interactions reaching a maximum between 4 and 10 Å; thereafter the molecule exits the constriction into the wider uncharged transmembrane barrel, giving little resistance to ongoing translocation.

calculation of the reaction coordinate in ABF simulations. We then examine A_{25}, dC_{25}, A_1, and dC_1 translocation using ABF and find that it qualitatively reproduces the experimental findings of poly(A) and poly(dC) translocation and the major observations from the highly sampled cv-SMD data.

5.1 Polynucleotide Translocation with an Adaptive Biasing Force

As with the cv-SMD simulations in Section 4, our primary measure of the validity of these simulations is the reproduction of qualitative experimental findings from poly(A) and poly(dC) translocation through αHL. This subsection presents multi-sample free energy profiles for the 16 Å reaction coordinates of A_{25} and dC_{25}. As we saw in the cv-SMD investigation, 16 Å is sufficient to examine translocation at the pore-constriction, and doing so allows for well sampled data within a reason-

able computational budget and time-frame.

However, before we can examine ABF-induced polynucleotide translocation in detail, it is important to recall that the reaction coordinate of the ABF methodology is calculated as a function of distance from the translocating molecule to a reference atom or set of atoms. In cv-SMD, the reaction coordinate is a function of the Cartesian coordinates of the system. If the protein were to move during the simulation, the free energy profile would no longer be an accurate function of the pore interior; therefore pin-pointing the free energy profile to specific pore residues would not be possible. This necessitates the constraining of the αHL pore in cv-SMD simulations. For ABF simulations, if the reference atoms are reasonably close to the translocating molecule, the free energy profile will be an accurate function of the length of the pore interior, regardless of the movement of the protein. This aspect of ABF provides an opportunity to leave the protein completely unconstrained, allowing the simulation to further approach conditions found in experiment.

Leaving the αHL protein unconstrained could have important consequences. In a comprehensive study of the αHL pore in MD simulations, the transmembrane barrel was found, on average, to be elliptical in shape.[23] The study found that the pore adopted seven different orientations of this shape, alternating between them on the scale of several nanoseconds. Constraining the protein prevents this kind of activity, and the consequences this might have on the simulations performed in this paper are unknown. Therefore, avoiding the constraining of the protein is could be important for accuracy.

Figure 5 shows free energy profiles from ABF simulations of A_{25} and dC_{25} translocation through an unconstrained (Figure 5(a)) and constrained (Figure 5(b)) αHL pore, both with a ζ value of 5000. Each profile is constructed from four samples and the error bars represent the sample-to-sample fluctuations. At the end of the 16 Å reaction coordinate in both cases, there is a separation of the A_{25} and dC_{25} profiles with non-overlapping error bars, with A_{25} showing a higher cumulative free energy of translocation. The figure also shows the average number of force measurements per bin from the four samples, in the form of histograms.

While the simulations of the unconstrained and constrained systems give similar results in

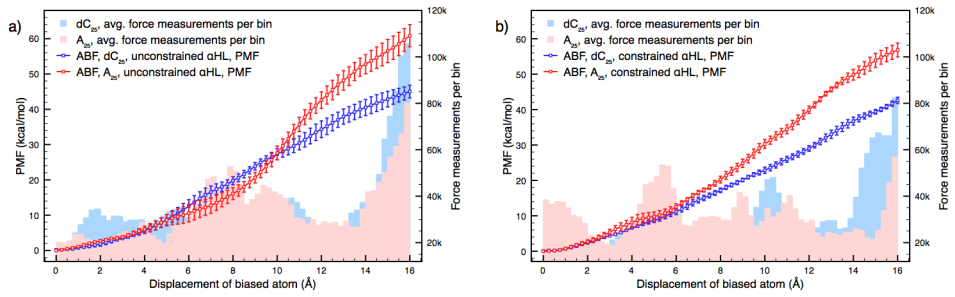

Figure 5: Free energy profiles of A_{25} and dC_{25} translocation through wild type αHL from a set of ABF simulations with the protein constrained and unconstrained. The reaction coordinate is from the centre of the alpha carbon atoms of protein residue 111 at the top of the constriction to 16 Å into the transmembrane barrel from that point. The timestep threshold parameter was set to 5,000 for these simulations. Each free energy profile was constructed from four samples, the error bars representing the sample-to-sample variation. The histograms represent the number of instantaneous force measurements per bin. (a) Free energy profiles from simulations with the αHL pore completely unconstrained. The free energy profiles are separated with non-overlapping error bars after 11 Å of translocation. At the end of the reaction coordinate, the free energy values of A_{25} translocation are approximately 45% higher than those of dC_{25}. (b) Free energy profiles from simulations with the C_α atoms of αHL constrained. The free energy profiles are separated with non-overlapping error bars after 6 Å of translocation. At the end of the reaction coordinate, the free energy values of A_{25} translocation are approximately 33% higher than those of dC_{25}. Taking an average of the error bars across the whole reaction coordinate, the errors are approximately 70% larger when unconstrained for A_{25}, and approximately 125% larger when unconstrained for dC_{25}. The A_{25} profile exhibits a curved shape in the unconstrained system, and a more uniform gradient in the constrained system.

terms of their end-points (Figure 5(a) and Figure 5(b) respectively), they also differ in some respects. Firstly, the histograms representing the instantaneous force measurements per bin (plotted against the right-hand *y*-axes) for the translocation of both polynucleotides have, on average, higher values in the unconstrained protein system, showing that the simulations need to be run for more timesteps in order to sample the full reaction coordinate. This is likely a consequence of the protein being allowed to shift position, to a degree in response to the translocating polynucleotide. Such shifts in the protein position appear to have a minimal impact on the total free energy difference across the reaction coordinate; as Figure 5 shows, the profile end points overlap to within errors when comparing the same translocating molecule (for example, comparing the A_{25} profiles in constrained and unconstrained protein conditions).

Secondly, the free energy error bars are larger with the unconstrained protein. Taking an average of the size of the error bars across the whole reaction coordinate, the errors are approximately 70% larger unconstrained compared to constrained for A_{25} translocation, and approximately 125% larger for dC_{25} translocation. When the protein is unconstrained, the scope of the phase space that can be explored becomes greater, therefore the sample-to-sample fluctuations are expected to be larger for a finite amount of sampling.

Thirdly, the separation of the A_{25} and dC_{25} profiles with non-overlapping error bars begins at different points in the constrained system (6 Å) compared to the unconstrained one (11 Å). The shapes of the profiles also play a role in this observation; the A_{25} profile exhibits a curved shape in the unconstrained system, while this is absent from the constrained system. Given the expected impact of the phosphate-lysine interaction as previously demonstrated, the absence of curves in the profiles could be interpreted as a loss of detail; however, the presence of curves in the free energy profiles from the unconstrained system may be due to the shifting position of the protein, thus being a consequence of the force from the translocating molecule, this would be an undesirable effect.

Finally, the separation of the A_{25} and dC_{25} profiles is larger for the unconstrained system than for the constrained system. At the end of the reaction coordinate, the cumulative free energy value of A_{25} translocation is approximately 45% higher than that of dC_{25} for the unconstrained system, and approximately 33% higher in the constrained system. This difference can be attributed to error, however, as the profile end-points lie within error bars of each other.

In deciding whether to constrain the protein or not, one must consider the impact of a shifting protein, both in terms of conformation and overall position. Based on the results in Figure 5, the differences in the data give rise to a trade-off between the size of the errors and the accurate re-creation of experimental conditions. For the αHL -polynucleotide system at least, the choice does not dramatically impact the end result of the free energy profiles.

Another major parameter of particular interest in ABF simulations is the value of ζ. As discussed in Section 2.2, this has a major impact on the translocation time and the influence of

non-equilibrium effects; we therefore investigate this parameter in great detail in the Supporting Information. Our investigations show that simulations where ζ=5000 are expected to contain a significant degree of non-equilibrium contributions in the free energy profiles. At ζ=20000 the non-equilibrium contributions are expected to be much lower, while the computational expense of running simulations using this parameter is significantly increased.

Figure 6 shows free energy profiles from ABF simulations of A_{25} and dC_{25} translocation with a ζ value of 20000. The profiles show good agreement with experimental observations of higher resistance to translocation for poly(A) than for poly(dC). They also exhibit agreement with the highly sampled cv-SMD data, showing consistently higher free energy values for A_{25} than dC_{25}. The separation between the free energy profiles is the largest of those represented in this paper with non-overlapping error bars for the vast majority of the reaction coordinate. This figure could also be viewed as representing conditions most similar to those found in experiments, given that the average translocation speed is slower as ζ increases, and that the αHL pore in this system is unconstrained. The figure also shows that the error bars are greater for A_{25} than for dC_{25}; this finding was also observed in the cv-SMD simulations in Section 4. As shown by the comparison of ζ=20000 and ζ=80000 in the Supporting Information, the ζ=20000 profiles here are still likely to contain some residual non-equilibrium effects. However, at ζ values higher than 20000, the computational expense of producing multi-sample profiles becomes too great to fully investigate.

Figure 6 also shows the average force measurements per bin from the 4 samples plotted as a histogram. Here, the average force measurements per bin are generally higher for A_{25} than for dC_{25}. It is interesting to note that the comparison of the two profiles may be considered not entirely on equivalent grounds due to the difference in the amount of sampling between the two and the impact that this has on removing non-equilibrium effects. To remedy this, the value of ζ could be increased for dC_{25} or decreased for A_{25}; this would make the average force samples per bin more alike across the reaction coordinate. It is also worth noting that doing so would almost certainly improve the separation between the free energy profiles, as all the data so far in this section indicates. It would be interesting to test this hypothesis, but it is unnecessary for the current

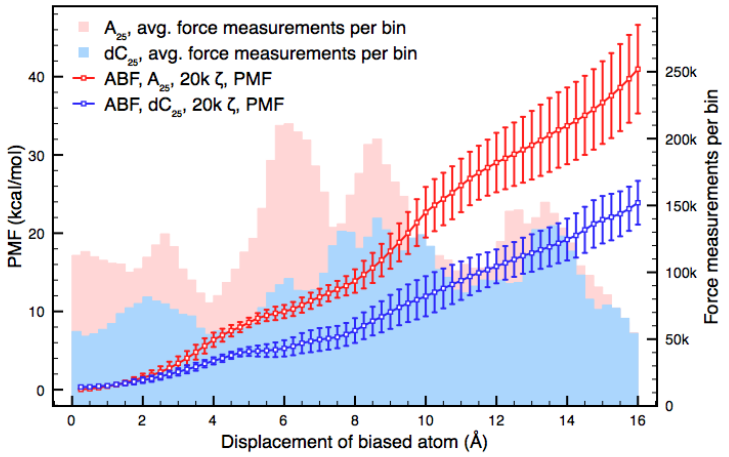

Figure 6: The free energy profiles of A_{25} and dC_{25} translocation through wild type αHL from a set of ABF simulations where the conditions have been set to obtain high quality data, requiring a large computational budget. The reaction coordinate is from the centre of the alpha carbons of protein residue 111 at the top of the constriction to 16 Å into the transmembrane barrel from that point. The timestep threshold parameter was 20,000 for these simulations. Each free energy profile was constructed from four samples, the error bars representing the sample-to-sample variation. The histograms represent the number of instantaneous force measurements per bin. The profiles show good agreement with the highly sampled cv-SMD data, exhibiting higher free energy values for A_{25} than dC_{25}. The free energy profiles are separated with non-overlapping error bars after just 3 Å of translocation. At the end of the reaction coordinate, the free energy value of A_{25} translocation is approximately 70% higher than that of dC_{25}.

study as it would provide no new insight while consuming a great deal of computational resources, due to the difficulty in predicting values of ζ that would shift the average force measurements per bin by a given amount.

5.2 Single Nucleotide Translocation with an Adaptive Biasing Force

The comparison of A_1 and dC_1 from Section 4 is revisited here, this time using the ABF methodology instead of cv-SMD. Given the impact of slow-relaxing forces in the polynucleotide chain, it is important to investigate ABF using smaller molecules such as single nucleotides, giving insight into the impact of non-equilibrium effects on the polynucleotide data.

Figure 7 shows free energy profiles from ABF simulations of A_1 and dC_1 translocation with a ζ value of 5000. The profiles are constructed from four samples per profile and the error bars

represent the sample-to-sample error. As indicated by the data in the Supporting Information, a higher value of ζ is not as important in reducing non-equilibrium contributions for smaller translocating molecules. The rapid rise in profile gradient after 5 Å and the subsequent levelling out after 11 Å corresponds well to the single nucleotide molecule leaving the confines of the αHL constriction, as was observed in the cv-SMD data. The strong phosphate-lysine interaction found in cv-SMD simulations for dC_1 is shown to be contributing similarly here as the dC_1 free energy profile shows a higher cumulative free energy than with A_1. The histograms showing the average force measurements per bin are largely similar for both nucleotides.

Unlike the cv-SMD profiles of A_1 and dC_1 translocation, separation between the two profiles is observed in Figure 7, with dC_1 showing higher free energy values. It is clear, then, that the greater propensity of dC_1 to experience a strong electrostatic interaction is observed when using ABF, just as it is when using cv-SMD. Additional samples would be needed to confirm if the dC_1 electrostatic interaction was experienced to a greater degree with ABF, as is suggested in the figure.

6 Comparison of Constant Velocity-Steered Molecular Dynamics with the Adaptive Biasing Force Method

Section 4 and Section 5 explored the cv-SMD and ABF methodologies for nucleotide translocation through αHL. Both approaches provided qualitative agreement with the experimental finding that A_{25} experiences greater barriers to translocation than dC_{25}. This section explores which of the two translocation methods is better suited to explore the nanopore-nucleotide system. First, we compare cv-SMD to ABF based on the general methodological differences; we look at the mode of translocation, consistency with experiment and constraints on the system. We then compare the results from simulations using each methodology in terms of the re-creation of experimental conditions and in data quality, the free energy profile shapes and the free energy profile separation between A_{25} and dC_{25}. We also consider the computational efficiency of each approach. The section finishes by extrapolation of our findings to other systems.

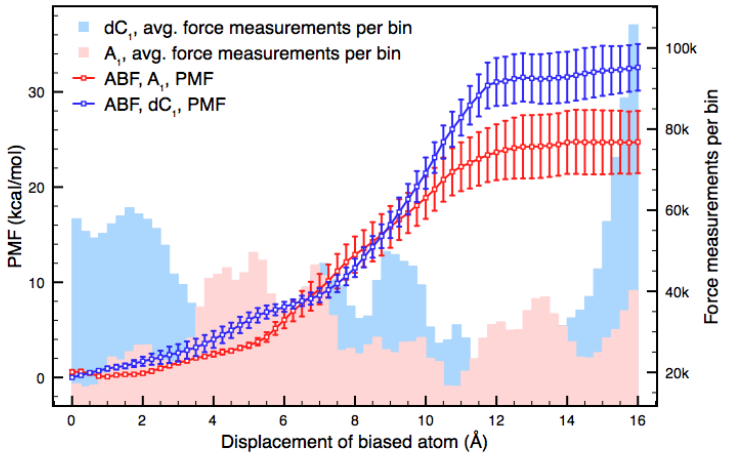

Figure 7: The free energy profiles of A_1 and dC_1 translocation through wild type αHL from a set of ABF simulations. The reaction coordinate is from the centre of the alpha carbon atoms of protein residue 111 at the top of the constriction to 16 Å into the transmembrane barrel from that point. The timestep threshold parameter was 5,000 for these simulations. Each free energy profile was constructed from four samples, the error bars represent the sample to sample variation. The histograms represent the number of instantaneous force measurements per bin. The free energy profiles are separated by non-overlapping error bars after 10.5 Å of translocation. At the end of the reaction coordinate, the free energy of A_{25} translocation is approximately 33% higher than that of dC_{25}.

6.1 Methodological Comparison

Single channel current recording experiments involve the translocation of a polymer through a protein-pore; this is a non-equilibrium process, though it is in a steady-state due to the constant transmembrane potential. This potential drives the polymer through the pore, and the driving force acts on the entire length of the polymer at all times. The free energy landscape of the solvated and ionised molecular system with respect to the translocating molecule, combined with the applied potential, determines the translocation time (a measurable quantity). So it is this free energy landscape that we wish to estimate using simulation, the difference in translocation time between poly(A) and poly(dC) being a measure that we use to validate our simulations, the translocation time being a reflection of the free energy landscape. Therefore, one key point of comparison between cv-SMD/JE and ABF is how closely the methodology matches the experimental process.

In cv-SMD/JE, the molecule is pulled in a non-equilibrium state and, while the method causes

the molecule to move at constant velocity, the applied force varies in response to the free energy landscape. The driving force is therefore different to experiment in this way. Another key difference to experiment is that the driving force is applied to the leading atom of the polymer, whereas experimentally it is applied to the whole molecule. During simulations, pulling a polymeric molecule by its leading end can result in deformation from the equilibrium conformation.[16] Deformation of the translocating molecule is expected to occur experimentally due to the dimensions of the pore,[7,16] but as a response to the steric hindrance of the constricting pore dimensions, rather than due to being dragged through the solvent. This artifactual form of deformation can be reduced by using a smaller driving force, where relaxation forces have time to act on the molecule. Furthermore, since the reaction coordinate of the ABF methodology is calculated as a function of distance relative to other reference atoms, the free energy profile will be an accurate function of the length of the pore interior, regardless of the movements of the protein. This allows the protein to be completely unconstrained, as discussed in SubSection 5.1.

The two methodologies also differ from each other in several other respects. Firstly, the ABF reaction coordinate is one-dimensional and therefore it is not restricted to axes orthogonal to the reaction coordinate; cv-SMD, on the other hand, is restricted to such orthogonal axes, and so, assuming a stiff spring constant (required in order to use JE), the SMD atom may not stray from a precisely chosen course. In this respect, ABF is closer to experiment than cv-SMD, where under experimental conditions the molecule is free to explore the full internal dimensions of the pore, and the translocation time is a measure of its transmembrane progression (a one-dimensional quantity).

Secondly, the direction of translocation along the reaction coordinate is not consistent in ABF simulations; therefore deviations from expected structural conformations of the polynucleotide can vary significantly from sample-to-sample. Such deviations result in a systematic error in the free energy profiles. It is quite straight forward to extrapolate the effect that this has in cv-SMD simulations due to the error being proportional to the consistent pulling speed. With data from several pulling speeds, one could extrapolate what the free energy difference would be at infinitesimally small translocation speed; this would be more difficult in ABF.

Thirdly, while using cv-SMD/JE requires a balance of statistical and truncation errors in equating the work done to the free energy, ABF involves no such approximations, due to its calculating the free energy directly from the system forces, and applying the biasing force directly into the biased atom's equations of motion.

The ABF methodology, while fundamentally different from cv-SMD in principle and in practice, is nevertheless closely related in certain respects. In ABF for instance, the molecule is permitted to diffuse along the reaction coordinate by a force that is adjusted in response to energetic barriers to translocation. In cv-SMD, since the leading molecule is being forcibly relocated, the actual force applied to it scales in response to the energetic barriers to the relocation, so the process is not in a steady-state in terms of the driving force. In this sense, cv-SMD could be considered more closely related to ABF than it is to constant force-SMD. Additionally, for the polynucleotide-nanopore system, the driving force is applied to a leading residue rather than the whole polymeric molecule, as in cv-SMD simulations. This makes the two methodologies more closely related to each other than they are to methods that use a transmembrane potential as a driving force such as, for example, grid-SMD.[16]

6.2 Data Comparison

When analysing simulation results in relation to experimental results, quantitative comparisons are difficult to draw without key data such as friction coefficients, and full pore length translocation data. However, qualitative comparisons may be drawn quite readily. When considering the simulation pulling methods in relation to each other, we may perform a rigorous analysis by drawing comparisons between simulation conditions, error bars, profile shapes, profile separation, free energy values and computational efficiency.

A direct comparison of ABF and cv-SMD for the translocation of A_{25} and dC_{25} is shown in Figure 8. Each profile is the average of four sample trajectories, each spanning the full 16 Å reaction coordinate across the pore-constriction. The bin width was set to 0.25 Å and the C_α atoms of the αHL pore were constrained in all instances. The ABF parameters were set to ζ=5000 with

a bin width of 0.25 Å, while the cv-SMD parameters were set to a pulling speed of 0.04 Å/ps. These methodology specific parameters equated to roughly the same average translocation speed. The profiles show that both methodologies exhibit higher free energy values for A_{25} than for dC_{25}. Additionally, the mean free energy values at the end of the reaction coordinate for ABF are within error bars for cv-SMD for the same polynucleotide.

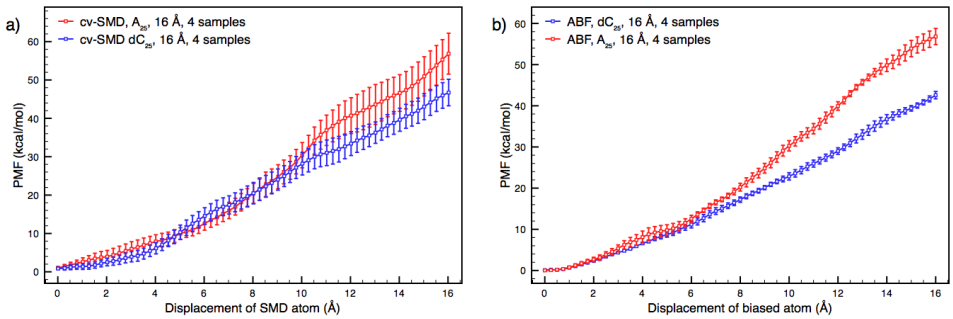

Figure 8: Free energy profiles from cv-SMD (Figure a) and ABF (Figure b) simulations of A_{25} and dC_{25} translocation under comparable conditions. Each profile is calculated from the average of 4 sample trajectories spanning the full 16 Å reaction coordinate. (a) The ABF parameters were ζ=5000 and bin width 0.25 Å. The free energy profiles are separated with non-overlapping error bars after 11 Å of translocation. At the end of the reaction coordinate, the free energy value of A_{25} translocation is approximately 45% higher than that of dC_{25}. (b) The cv-SMD parameters were a pulling speed of 0.04 Å/ps and bin width 0.25 Å. The free energy profiles are separated with non-overlapping error bars after 15.5 Å of translocation. At the end of the reaction coordinate, the mean of the free energy values of A_{25} translocation is approximately 20% higher than that of dC_{25}. The two methodologies exhibit a greater free energy profile for A_{25} than for dC_{25}. Compared to the cv-SMD profiles, the ABF data show greater separation at the end of reaction coordinate; the separation occurs throughout a larger proportion of the reaction coordinate, and the errors are substantially smaller. Taking an average of the error bars across the whole reaction coordinate, the errors in the cv-SMD profiles are approximately 185% larger than the ABF profiles for A_{25}, and approximately 270% larger for dC_{25}.

While the ABF and cv-SMD free energy profiles from Figure 8 are similar in some respects, the figure also shows that use of each methodology leads to notable differences. The ABF method manifests a greater separation between the free energy profiles of A_{25} and dC_{25} by the end of the reaction coordinate. At the end of the reaction coordinate, the free energy value of A_{25} translocation is approximately 20% higher than that of dC_{25} when cv-SMD is used. The difference is approximately 33% when ABF is used. The separation is also aided by the considerably smaller

error bars in the ABF profiles. By contrast, the error bars on the A_{25} and dC_{25} free energy profiles can be seen to overlap in the case of cv-SMD for the majority of the reaction coordinate. The separation between A_{25} and dC_{25} at the end of the reaction coordinate is 15.8 ± 4.9 kcal/mol in ABF and 10.1 ± 8.8 kcal/mol using cv-SMD, therefore the separation of the profile means is larger in addition to having smaller errors using ABF. Taking an average of all the free energy profile error bars across the reaction coordinate, the average errors in the cv-SMD profiles are approximately 185% larger than with ABF for A_{25}, and approximately 270% larger for dC_{25}.

The error bars are likely smaller when using ABF for a couple of reasons. Firstly, as demonstrated earlier in the section, the binning error is negligible due the large number of measurements taken per bin, improving the statistical quality of the calculations. Secondly, single samples of ABF, with its translocative motion determined largely by self-diffusion and not by being forced along the reaction coordinate, may be more representative of the true free energy landscape, and so the sample-to-sample fluctuations are lower. Consider that an infinitesimally slowly moving molecule is likely to fully sample all accessible phase space configurations and energy values to a degree which is fully representative of the free energy landscape, and therefore multiple samples of infinitesimally slowly moving trajectories will have zero sample to sample free energy profile fluctuation. Equally, a fast moving entity will sample less of the accessible phase space; therefore more samples will be required to construct a meaningful free energy profile. It follows, then, that a methodology which samples the phase space more effectively will represent the free energy landscape better per sample, and so the sample-to-sample variation will be less. It is likely that being unconstrained in axes orthogonal to the reaction coordinate also contributes to this effect.

6.3 Computational Efficiency

To fully compare each method, one must also look at the computational cost under comparable conditions, in addition to the quality of the output. In general there is roughly a 3.5% increase in computation time for an ABF simulation compared to a cv-SMD simulation for a fixed number of timesteps with the same number of cores on the same system (tested on the XSEDE machine

Kraken at 576 processors). This is because an ABF simulation must perform additional calculations for the generalised coordinates of the biased and reference atoms, and calculate the average instantaneous force acting on the biased atom. Calculations based on the cv-SMD harmonic spring and the position of the reference atom are comparatively simple, and therefore less computationally demanding.

For the ABF simulations that give rise to the profiles in Figure 8, a bin width of 0.25 Å and a ζ value of 5000 is used. This requires roughly 2 million timesteps per sample trajectory at a total cost of roughly 25,000 CPU hours for a four sample profile. Here, each sample trajectory is produced from two or more simulations in blocks of 100,000 to 1 million timesteps per simulation until the full reaction coordinate is sampled. With a cv-SMD pulling speed of 0.04 Å/ps, for a 16 Å translocation, 2.4 million timesteps are required per sample at a total cost of 29,000 CPU hours for a four sample ensemble average. Here, each sample is produced from four simulations, the combined simulations covering the full reaction coordinate. There is additional computational time required in cv-SMD simulations under the conditions we have used in order to produce the reaction coordinate segment overlap; the explanation for this is provided by Martin *et al.*[30]

It should be noted that a relatively consistent progression along the reaction coordinate for the ABF simulations under these conditions is aided by undesirable slow non-equilibrium relaxational effects. With smaller translocating molecules, or higher ζ values to allow more time for the conformations to relax (thus producing a more correct profile), the number of timesteps required to sample the whole trajectory would increase and be difficult to predict. As shown in the Supporting Information, where ζ=80000, sampling the reaction coordinate requires roughly 16 million timesteps for a polymeric chain and 20 million timesteps for a single nucleotide. In cv-SMD simulations, the quality of the data may also be improved by slowing down the translocation. In the case of cv-SMD, the increase in computational cost is precise and therefore straightforward to plan and manage.

For the conditions given for this comparison, ABF displays numerous advantages; it possesses fewer sources of errors, smaller errors, better separation of free energy profiles, lower computa-

tional cost, fewer constraints, and greater freedom in axes orthogonal to the reaction coordinate.

6.4 Extrapolating to Other Systems

The question remains as to whether this comparison would hold up in other systems/conditions. To answer this we must consider individual contributors to each free energy profile. In cv-SMD/JE there are two sources of error from the implementation of the methodology, the harmonic spring and the truncation of the cumulative term in the use of Jarzynski's identity. The latter will have a contribution in alternative systems, regardless of size or pulling speed. The harmonic spring leads to an increase in the statistical noise of the output as the harmonic spring constant is increased, yet it must be high enough to approximate a stiff spring. For larger translocating molecules, the spring constant must be scaled up in order to continue approximating a stiff spring, hence it becomes necessary to introduce more statistical noise. The higher statistical noise will increase the binning error in the free energy profiles. Therefore, the cv-SMD error would be expected to increase for larger translocating molecules. This scaling of binning error may also be affected by the pulling speed, where faster pulling speeds require higher spring constants in order to approximate a stiff spring, thereby increasing the error contribution.

Even if the binning error was completely negated in the cv-SMD profiles, the sample-to-sample contributions to the errors are larger than those of the ABF profiles. This may be surprising as, for the ABF simulations, the reaction coordinate is not restrained in axes orthogonal to it. This lack of restraint increases accessible regions of the phase space, which one would expect to increase the sample-to-sample fluctuations. The opposite case is observed, where each sample appears to represent the free energy landscape well, resulting in low sample-to-sample fluctuations. It is possible that the constrained reaction coordinate in the cv-SMD case imposes certain conformations in the translocating molecule, to a degree which may not be proportionally representative of the ensemble phase space, thereby resulting in more varied individual samples. It is moreover feasible that the sampling of phase space is also improved by the translocative motion in ABF simulations being determined largely by self-diffusion rather than rigidly implemented relocation, again leading to

lower sample-to-sample fluctuation. For these advantages in the ABF sampling to be allowed to flourish, the translocating molecule must be permitted sufficient time within each bin along the reaction coordinate, whereas the time spent in each bin would be reduced if the average translocation speed would increase. Therefore, at higher speeds, one might expect the sample-to-sample fluctuations to occur to a similar degree in both methodologies, whereas at slower speeds, the ABF methodology would produce better data for a given computational budget. Further investigation - at great computational - expense would be required in order to fully answer the question of how the ABF and cv-SMD methodologies compare in other systems and/or conditions; it is nonetheless clear that, for the translocation of polynucleotides through the αHL protein pore, ABF stands out as the methodology of choice.

7 Conclusions

Using constant velocity-steered molecular dynamics in conjunction with Jarzynski's Equality, free energy profiles were produced from translocation simulations. These simulations covered single nucleotides and polynucleotides translocation through wild type and mutated αHL protein-nanopores.

Free energy profiles from cv-SMD translocation of A_{25} and dC_{25} through wild type αHL were consistent with the experimentally observed trend that A_{25} experiences greater barriers to translocation than dC_{25}. Due to a high number of samples, the data is significantly more reliable than any previous simulations of the nucleotide-nanopore system using cv-SMD. The free energy of A_{25} translocation was shown to be approximately 30% higher than that of dC_{25} after 48 Å of translocation.

Free energy profiles from single nucleotide translocation were used to highlight the strongest contribution to polynucleotide free energy profiles; the data was calculated from an unprecedented amount of samples. The free energy profiles displayed a distinctive shape that we attribute to a strong phosphate-lysine interaction at the constriction.

Using the adaptive biasing force methodology, free energy profiles were produced from translocation simulations in the nucleotide-nanopore system for the first time. These simulations spanned single nucleotide and polynucleotide translocation through the wild type αHL protein-nanopore.

As with the cv-SMD data, the free energy profiles from ABF simulations of A_{25} and dC_{25} translocation through wild type αHL exhibited the experimentally observed trend that A_{25} experiences greater barriers to translocation than dC_{25}. Unlike with cv-SMD, the ABF methodology allows for the protein-pore to be left unconstrained, permitting conditions more akin to those found in nanopore current recording experiments. By comparing simulations using a constrained and unconstrained αHL pore, it was found that constraining the protein led to smaller errors but also smaller separation of the A_{25} and dC_{25} profiles. It was also demonstrated that when the simulation conditions were optimised to reduce non-equilibrium contributions, the free energy of A_{25} translocation was shown to be approximately 70% higher than that of dC_{25} after 16 Å of translocation.

To best compare cv-SMD/JE to ABF for the nanopore-nucleotide system, simulations were performed with conditions across both methodologies tuned to be as similar as possible. The resulting free energy profiles from both methodologies were within error bars in terms of their end-point values after 16 Å of translocation. Additionally, the error bars were found to be notably smaller and the separation between the free energy profiles of A_{25} and dC_{25} translocation was larger when using ABF. Given that ABF presents these advantages in the statistical quality of the data, and under our conditions is less computationally intensive for obtaining similar results to cv-SMD, the ABF method is a natural choice for future work of this type.

Additionally, the ABF method has some intrinsic advantages, including the biased atom being permitted more fully to explore the internal dimensions of the pore, as one would expect under experimental nanopore recording conditions. Furthermore, in considering the choice of methodology for other systems, there are many more forms of reaction coordinate available to exploit when using ABF. Additionally, ABF generates fewer sources of error and uncertainty, eliminating the need for cumulant expansions (needed when using JE), and the need for carefully selecting stiff springs (as in cv-SMD). These factors further strengthen ABF as the recommended method.

However, cv-SMD retains a major advantage over ABF in that it has a set number of timesteps required to traverse a reaction coordinate distance, allowing precise planning of simulation time and a computational budget.

With ABF established as the preferred method, future investigations could aim to compare ABF to alternative translocation methods, particularly metadynamics and/or grid-SMD. With Oxford Nanopore Technologies making major progress in the field of nanopore sequencing, it would also be of great interest to reconstruct their most successful αHL nanopores in simulations that harness such translocation methods. The insight gained could be used to improve the experimental system, while the race for cheaper and faster sequencing technologies goes on.

Acknowledgements

This work was supported by the UK Engineering and Physical Science Research Council (EPSRC) which provided access to HPCx (http://www.hpcx.ac.uk/). Computing resources were made possible via NSF TRAC award TG-MCB090174 and LONI resources.

8 Supporting Information Available

See Supporting Information for data and discussion of the displacement of the biased atom in ABF simulations with respect to time at different values of ζ, and investigation of single and polynucleotide translocation free energy profiles at different values of ζ. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

[1] Mullen, C.; Kilstrup, M.; Blaese, R. *Proc. Natl. Acad. Sci.* **1992**, *89*, 33–37.

[2] Davies, J. *Sci.* **1994**, *264*, 375–382.

[3] Koonin, E.; Makarova, K.; Aravind, L. *Ann. Rev. Microbiol.* **2001**, *55*, 709–742.

[4] Dreiseikelmann, B. *Microbiol. Mol. Biol. Rev.* **1994**, *58*, 293–316.

[5] Hanss, B.; Leal-Pinto, E.; Bruggeman, L.; Copeland, T.; Klotman, P. *Proc. Natl. Acad. Sci.* **1998**, *95*, 1921–1926.

[6] Kasianowicz, J.; Brandin, E.; Branton, D.; Deamer, D. *Proc. Natl. Acad. Sci.* **1996**, *93*, 13770–13773.

[7] Akeson, M.; Branton, D.; Kasianowicz, J.; Brandin, E.; Deamer, D. *Biophys. J.* **1999**, *77*, 3227–3233.

[8] Meller, A.; Nivon, L.; Brandin, E.; Golovchenko, J.; Branton, D. *Proc. Natl. Acad. Sci.* **2000**, *97*, 1079–1084.

[9] Meller, A.; Nivon, L.; Branton, D. *Phys. Rev. Lett.* **2001**, *86*, 3435–3438.

[10] Butler, T.; Gundlach, J.; Troll, M. *Biophys. J.* **2006**, *90*, 190–199.

[11] Mathe, J.; Aksimentiev, A.; Nelson, D.; Schulten, K.; Meller, A. *Proc. Natl. Acad. Sci.* **2005**, *102*, 12377–12382.

[12] Branton, D.; Deamer, D.; Marziali, A.; Bayley, H.; Benner, S.; Butler, T.; Di Ventra, M.; Garaj, S.; Hibbs, A.; Huang, X. *Nat. biotechnol.* **2008**, *26*, 1146–1153.

[13] Clarke, J.; Wu, H.; Jayasinghe, L.; Patel, A.; Reid, S.; Bayley, H. *Nat. Nanotechnol.* **2009**, *4*, 265–270.

[14] Lieberman, K.; Cherf, G.; Doody, M.; Olasagasti, F.; Kolodji, Y.; Akeson, M. *J. Am. Chem. Soc.* **2010**, *132*, 17961–17972.

[15] Hayden, E. C. Nanopore Genome Sequencer Makes Its Debut. (2012, accessed date 21/11/2012); `http://www.nature.com/news/nanopore-genome-sequencer-makes-its-debut-1.10051`.

[16] Wells, D.; Abramkina, A.; A, A. *J. Chem. Phys.* **2007**, *127*, 125101–1–125101–10.

[17] Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562.

[18] Marsili, S.; Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V. *J. Phys. Chem. B* **2006**, *110*, 14011–14013.

[19] Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9169.

[20] Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. *J. Comp. Chem.* **2005**, *26*, 1781–1802.

[21] Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

[22] Crooks, G. *J. Stat. Phys.* **1998**, *90*, 1481–1487.

[23] Aksimentiev, A.; Schulten, K. *Biophys. J.* **2005**, *88*, 3745–3761.

[24] Maglia, G.; Restrepo, M.; Mikhailova, E.; Bayley, H. *Proc. Natl. Acad. Sci.* **2008**, *105*, 19720–19725.

[25] Howorka, S.; Cheley, S.; Bayley, H. *Nature Biotechnology* **2001**, *19*, 636–639.

[26] Mitchell, N.; Howorka, S. *Angew. Chem. Int. Ed.* **2008**, *47*, 5565–5568.

[27] Ashkenasy, N.; Sánchez-Quesada, J.; Ghadiri, M.; Bayley, H. *Angew. Chem. Int. Ed.* **2005**, *44*, 1401–1404.

[28] Hornblower, B.; Coombs, A.; Whitaker, R.; Kolomeisky, A.; Picone, S.; Meller, A.; Akeson, M. *Nat. Methods* **2007**, 315–317.

[29] Astier, Y.; Braha, O.; Bayley, H. *Sci.* **2005**, *309*, 1728–1732.

[30] Martin, H.; Jha, S.; Howorka, S.; Coveney, P. *J. Chem. Theor. Comp.* **2009**, *5*, 2135–2148.

[31] Bond, P.; Guy, A.; Heron, A.; Bayley, H.; Khalid, S. *Biochem.* **2011**, *50*, 3777–3783.

[32] Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J. Comp. Phys.* **1999**, *151*, 283–312.

[33] Darve, E.; Wilson, M.; Pohorille, A. *Mol. Sim.* **2002**, *28*, 113–144.

[34] Hénin, J.; Chipot, C. *J. Chem. Phys.* **2004**, *121*, 2904.

[35] Chipot, C.; Hénin, J. *J. Chem. Phys.* **2005**, *123*, 244906.

[36] Hénin, J.; Fiorin, G.; Chipot, C.; Klein, M. *J. Chem. Theor. Comp.* **2010**, *6*, 164–170.

[37] Pearlman, D.; Case, D.; Caldwell, J.; Ross, W.; Cheatham III, T.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comp. Phys. Comm.* **1995**, *91*, 1–41.

[38] Feller, S.; MacKerell, A. *J. Phys. Chem. B* **2000**, *104*, 7510–7515.

[39] Bhandarkar, M.; Bhatele, A.; Bohm, E.; Brunner, R.; Buelens, F.; Chipot, C.; Dalke, A.; Dixit, S.; Fiorin, G.; Freddolino, P. *Urbana* **2009**, *51*, 61801.