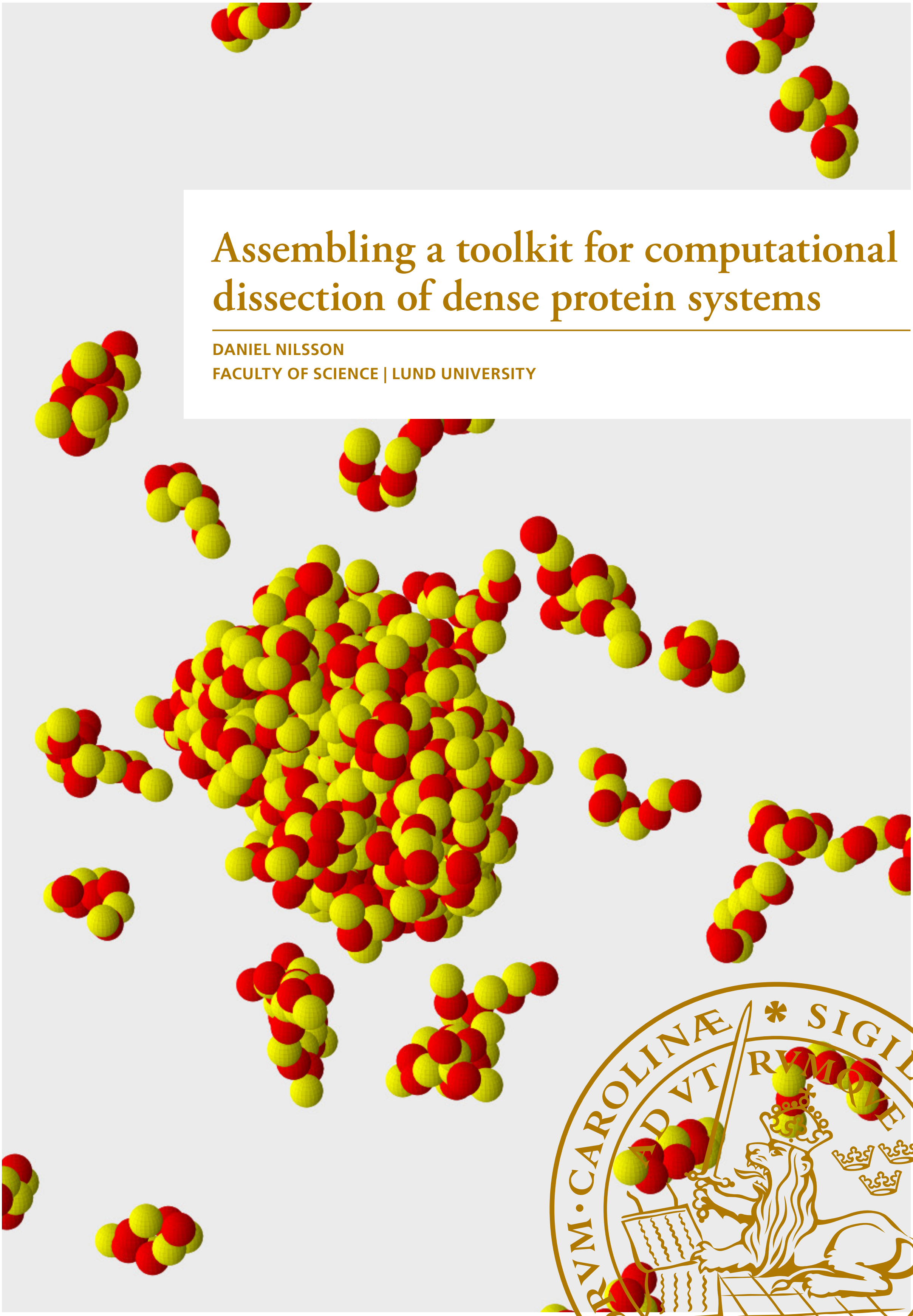
Assembling a toolkit for computational dissection of dense protein systems

Nilsson, Daniel

2021



Assembling a toolkit for computational dissection of dense protein systems

DANIEL NILSSON

FACULTY OF SCIENCE | LUND UNIVERSITY

LUND UNIVERSITY

Lund University
Faculty of Science
Department of Astronomy and Theoretical Physics
Computational Biology and Biological Physics
ISBN 978-91-8039-018-7

9 789180 390187

Printed by Media-Tryck, Lund 2021 NORDIC SWAN ECOLABEL 3041 0903

Assembling a toolkit for computational dissection of dense protein systems

Assembling a toolkit for computational dissection of dense protein systems

by Daniel Nilsson

Thesis for the degree of Doctor of Philosophy
Thesis advisor: Anders Irbäck
Faculty opponent: Peter Virnau

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in Lundmarksalen at the Department of Astronomy and Theoretical Physics on Friday, the 22nd of October 2021 at 10:00.

Organization
LUND UNIVERSITY

Department of Astronomy and Theoretical Physics
Box 43
SE–221 00 LUND
Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2021-10-22

Sponsoring organization

Author(s)
Daniel Nilsson

Title and subtitle
Assembling a toolkit for computational dissection of dense protein systems

Abstract

The cellular interior is a dense environment. Understanding how such an environment impacts the properties of proteins and other macromolecules, as well as how weak, non-specific interactions drive processes such as protein droplet formation through liquid-liquid phase separation, is a major challenge in biological physics. The complexity of this environment often makes experimental studies extremely challenging, leaving an important niche to be filled by simulation studies.

Simulations do, however, have their own set of challenges, and to use them to their full potential, a suitable set of computational tools must be developed. Such a toolset must include accurate yet computationally affordable force fields, computationally efficient simulation algorithms, and analysis tools that allow for the extraction of meaningful information from the simulation results.

In this thesis, a number of tools for all three areas are developed and/or evaluated. We present an atom level, implicit solvent force field, as well as a coarse-grained continuous HP model which we use for droplet formation studies. We investigate sampling issues in field theory simulations with the complex Langevin equation. We use finite-size scaling analysis to analyse simulations of liquid-liquid phase separation, and Markov state modeling to analyse crowding simulations.

DOKUMENTDATABLAD enl SIS 61 41 21

Key words
macromolecular crowding, liquid-liquid phase separation, Monte Carlo simulation, time-lagged independent component analysis, finite-size scaling, polymer field theory, protein force field

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-8039-018-7 (print)
978-91-8039-019-4 (pdf)

Recipient's notes

Number of pages
128

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature Daniel Nilsson

Date 2021-09-12

Assembling a toolkit for computational dissection of dense protein systems

by Daniel Nilsson

LUND
UNIVERSITY

A doctoral thesis at a university in Sweden takes either the form of a single, cohesive research study (monograph) or a summary of research papers (compilation thesis), which the doctoral student has written alone or together with one or several other author(s).

In the latter case the thesis consists of two parts. An introductory text puts the research work into context and summarizes the main points of the papers. Then, the research publications themselves are reproduced, together with a description of the individual contributions of the authors. The research papers may either have been already published or are manuscripts at various stages (in press, submitted, or in draft).

Cover illustration front: A snapshot from one of the simulations done for paper II. Red spheres are H beads, and yellow ones P beads.

Funding information: The thesis work was in part funded by the Swedish strategic research program eSSENCE and the Swedish Research Council (Grant No. 621-2018-04976).

Faculty of Science, Department of Astronomy and Theoretical Physics

ISBN: 978-91-8039-018-7 (print)
ISBN: 978-91-8039-019-4 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2021

Contents

List of Papers iii
Acknowledgements iv
Populärvetenskaplig sammanfattning v

Assembling a toolkit for computational dissection of dense protein systems **1**
1 Introduction 1
2 Dense protein systems 2
3 Force fields 3
4 Simulation methods 6
5 Data analysis 10
6 Summary 13

Scientific publications **21**
Paper i: Markov modeling of peptide folding in the presence of protein crowders 25
Paper ii: Finite-size scaling analysis of protein droplet formation 41
Paper iii: Finite-size shifts in simulated protein droplet phase diagrams 51
Paper iv: Limitations of field-theory simulation for exploring phase separation: the role of repulsion in a lattice protein model 61
Paper v: An effective potential for atomic-level simulation of structured and unstructured proteins 89

List of Papers

This thesis is based on the following papers, referred to by their Roman numerals:

I **Markov modeling of peptide folding in the presence of protein crowders**

D. Nilsson, S. Mohanty, A. Irbäck
The Journal of Chemical Physics **148**, 055101 (2018)

II **Finite-size scaling analysis of protein droplet formation**

D. Nilsson, A. Irbäck
Physical Review E **101**, 022413 (2020)

III **Finite-size shifts in simulated protein droplet phase diagrams**

D. Nilsson, A. Irbäck
The Journal of Chemical Physics **154**, 235101 (2021)

IV **Limitations of field-theory simulation for exploring phase separation: the role of repulsion in a lattice protein model**
D. Nilsson, B. Bozorg, S. Mohanty, B. Söderberg, A. Irbäck
Submitted to *The Journal of Chemical Physics*

V **An effective potential for atomic-level simulation of structured and unstructured proteins**
D. Nilsson, S. Mohanty, A. Irbäck
Manuscript

Papers II and III are reproduced under a Creative Commons Attribution 4.0 International license. Paper I is reproduced with permission.

Publications not included in this thesis:

When a foreign gene meets its native counterpart: computational biophysics analysis of two *PgiC* loci in the grass *Festuca ovina*
Y. Li, S. Mohanty, D. Nilsson, B. Hansson, K. Mao and A. Irbäck
Scientific Reports **10**, 18752 (2020)

Peptide folding in cellular environments: a Monte Carlo and Markov modeling approach
D. Nilsson, S. Mohanty and A. Irbäck
In *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes*, ed. A. Liwo, pp. 453–466 (Springer, Second Edition, 2019)

Acknowledgements

I have never been a fan of long-winded thank yous, so I will try to keep this section brief. Nevertheless, there are some people I cannot in good conscience leave unthanked.

First and foremost, I must of course thank my supervisor, Anders Irbäck. Thank you for always providing a good mix of guidance and freedom, and for occasionally telling me that my results are not actually that bad.

I wish to thank the rest of the people I have collaborated closely with during my PhD studies, Sandipan Mohanty, Yuan Li, Behruz Bozorg and Bo Söderberg. Each have brought their own unique mix of knowledge, curiosity, enthusiasm and friendliness. I have enjoyed working with all of you.

I also want to extend my thanks to everyone else I have gotten to meet through my research, with whom I have had stimulating discussions, or just a good time. A special thanks to Carsten Peterson for his comments on the thesis text.

Finally, I want to thank my family for their support, and for providing me with an escape from my research whenever I needed one.

Populärvetenskaplig sammanfattning

Vi består alla av celler, och att förstå vad som händer inuti cellerna är viktigt bland annat för att förstå vad som orsakar olika sjukdomar. Men cellernas inre är en stökig miljö. En stor mängd så kallade makromolekyler - huvudsakligen proteiner, RNA och DNA - upptar en stor del av utrymmet, kanske så mycket som en tredjedel. Var och en av dessa makromolekyler har en specifik uppgift i cellens maskineri, men de kan också växelverka med varandra, mer eller mindre slumpmässigt. Att förstå hur den här sortens oavsiktlig växelvekan påverkar molekylernas huvudsakliga funktion är en svår men viktig uppgift.

Trots detta skenbara kaos är cellerna internt organiserade. Denna organisation består delvis av membranomgärdade så kallade organeller, men det finns även grupper av molekyler som "spontant" bildar strukturer utan hjälp av membran. Att förstå de krafter som styr den här typen av strukturbildning är också viktigt.

Att studera hur enskilda molekyler i en cell beter sig vore som att försöka följa en enskild person i en storstad från yttre rymden. Därför utförs experiment på makromolekyler oftast i en vattenlösning. Det blir då lättare att urskilja molekylen, men har man otur kan det man ser skilja sig helt från hur det verkligen ser ut innuti en cell.

Ett viktigt komplement till direkta experiment på den här sortens svårstuderade system är datorsimuleringar. I simuleringar är det enkelt att zooma in på precis de aspekter man vill titta på. Men för att kunna utföra simuleringar krävs att man löser ett antal viktiga problem.

För det första måste man utveckla en modell, och se till att simuleringar med denna faktiskt ger resultat som faktiskt stämmer med verkligheten. Det är i sig långt ifrån enkelt, och kräver att man kan jämföra åtminstone några av resultaten med experiment.

För det andra måste simuleringarna vara tillräckligt snabba. Även om datorer utvecklats i en rasande fart kan en enskild simulering av detta slag utan problem ta flera veckor eller ännu längre. Snabbare simuleringsprogram gör också att man kan studera större och mer detaljerade modeller.

Slutligen måste man kunna dra slutsatser från den data som kommer ut. Det är visserligen rätt enkelt om man redan från början vet vilka sorters processer man vill studera. Men det är långtifrån säkert att man vet det när man studerar system med många makromolekyler. Ett annat problem är att resultaten ibland kan påverkas av hur stort det simulerade systemet är – och det är än så länge omöjligt att simulera system som är lika stora som verkliga celler.

Alla ovan beskrivna problem måste lösas om datorsimuleringar fullt ut skall kunna användas för att studera cell-liknande miljöer. Den här avhandlingen utgör mitt bidrag till att lösa dem.

Assembling a toolkit for computational dissection of dense protein systems

1 Introduction

The question of what separates the living from the non-living has long fascinated mankind. Over the last hundred-or-so years, advances in imaging has allowed for a vastly improved understanding of the structure, function, and organisation of biological matter. We now know that living organisms are made up of cells, and that the interior of cells consists of large numbers of chain molecules, such as DNA, RNA, and proteins.

Traditionally, the sequence of a protein has been seen as determining it's three-dimensional shape, and thereby its function. Many proteins fold to a unique, so called native, conformation. Thus, by determining the native state of a protein, either through experiment or through computation, it should be possible to determine the functionality of the protein. However, as our understanding of the cellular interior has improved, new challenges have emerged.

First, protein function can not be fully understood without understanding the environment within which it operates. For imaging of macromolecules to be at all feasible, they are usually studied in a dilute water solution. This environment is significantly different from the cellular interior, where macromolecules often fill as much as a quarter of the available volume [1]. Understanding if and how such a crowded environment affects protein function is therefore crucial [2–4].

Second, non-specific interactions between proteins (and/or nucleic acids) can provide large-scale structure to the cellular interior, as evidenced by recent studies of biomolecular condensates. In contrast to membrane-enclosed organelles, these condensates appear to form as liquid droplets [5, 6]. They are often rich in so-called intrinsically disordered proteins, which do not fold to a unique conformation. Understanding the conformational ensembles of these proteins, and how their sequence properties affect droplet formation is a key question.

Because the dense systems where the above described phenomena take place are difficult to study using traditional experimental techniques, computer simulations have an important role to play in the field. Yet such simulations are not without challenges of their own, and in order to use them to their full potential, a suitable set of tools must be developed. Broadly interpreted, such a toolkit will have to include

- Force fields/models which are simple enough that simulations are feasible, yet accurate enough to allow conclusions about reality to be drawn.
- Computationally efficient sampling algorithms.
- Data analysis tools which allow meaningful information to be extracted from simulations.

This thesis contains work on all three of these categories of tools. Below, I will first give a brief biological background, including the two challenges mentioned above. Thereafter, I will discuss the three categories in relation to my research.

2 Dense protein systems

Proteins are one of the most common types of macromolecules found in living organisms. Proteins are chain molecules built up of amino acids, the sequence of amino acids determining the structure of the protein. Many proteins fold to a single well-defined three-dimensional structure, encoded by its sequence. In recent years, however, it has become increasingly clear that many proteins exhibit significant conformational flexibility, so-called intrinsically disordered proteins. A key challenge within molecular biophysics is to understand how the amino acid sequence encodes the properties of (folded or intrinsically disordered) proteins.

2.1 Macromolecular crowding

The relationship between protein sequence on one hand, and structure and function on the other, has traditionally been studied using protein molecules in dilute solutions. In such an environment, it is easy to separate the experimental signal from the background response. On the other hand, a dilute environment need not be a particularly good approximation of the cellular interior, where as much as a quarter of the volume is occupied by different macromolecules [1].

The presence of large numbers of surrounding macromolecules can affect the conformational properties of proteins in several ways. The most straightforward way is through steric interactions, which should universally cause proteins to adopt more compact conformations. Thus, folded or bound conformations should be more favored in such an environment, than in dilute solution.

However, recent studies have indicated that environments rich in macromolecules can either stabilize or destabilize folded states, depending on both the protein studied and on the types of (macro)molecules in the environment [7, 8].

The difficulties involved in experimentally probing protein behaviour in complex, cell-like environments open the possibility of using computer simulation to answer important questions about how macromolecular crowding affects protein behaviour [9–11].

2.2 Biomolecular condensates

It has been known for more than a century that the cellular interior contains membrane-bound compartments known as organelles. In addition to these, cells also contain membrane-less assemblies of biological macromolecules, sometimes referred to as biomolecular condensates [5, 6]. Intrinsically disordered proteins are thought to often play an important role in the formation of these assemblies [12–14].

In recent years, it has been shown that macromolecular condensates exhibit properties typical of liquid droplets, such as concentration-dependent formation/dissolution, coalescence, and wetting [5]. These observations strongly suggest that macromolecular condensates form through a liquid-liquid phase separation (LLPS) process.

Biomolecular condensates presumably play important roles in regulating cell function, and may also play a role in the development of certain diseases. Understanding the forces driving the droplet formation process is therefore crucial, and also here simulation has an important role to play [15–18].

3 Force fields

In order to gain useful insights from simulation, the choice of an appropriate force field is crucial. In choosing a force field, two main and competing goals have to be fulfilled.

1. The force field needs to be detailed enough that the phenomena of interest can be observed.

2. At the same time, the force field must be simple enough to permit calculations to finish within a reasonable time.

In the literature, the range of complexity of force fields is vast, from highly coarse-grained models such as the lattice-based HP model of Lau and Dill [19], to extremely detailed descriptions which include quantum-mechanical effects [20]. The choice of model for a particular study depends on the problem studied and the available resources.

Within my work I have been using two classes of models. In Paper I, we used an all-atom implicit solvent model. The same model was further developed in Paper V. In Papers II-IV, we instead used coarse-grained models, in which the basic entities are amino acids rather than atoms.

Below I will describe the approach taken in each case.

3.1 All-atom implicit-solvent force fields

In the force fields used in Papers I and V, all atoms in the protein molecules are explicitly represented. This detail of representation is often required in order to be able to differentiate the physical properties of different amino acids. Compared to more coarse-grained models [21–23], we expect it to be better able to describe the conformational distribution of specific proteins. In particular, a fully atomistic description facilitates the description of secondary structure.

In contrast to the protein chain, the surrounding water is described only implicitly. Compared to explicit-solvent models [24–26], we need to represent far fewer atoms in the simulations, which means our computational demands are lower. This holds especially when simulating proteins with extended conformations. With our choice of model, we are able to simulate multiple folding/unfolding events in a single simulation run, at least for small proteins ($\sim$50 residues long). Two force fields with a similar level of detail as ours are ABSINTH [27] and AWSEM [28].

Since the force field is intended to be able to reproduce experimental results, it must be carefully calibrated. Our approach has been to calibrate the force field against experimental data for short peptides. For the version used in Paper I, the data came from a set of folded peptides, while in Paper V, we revised the force field using data also from a set of unstructured peptides.

Below, an outline of the revised force field is given, see Paper V for a full description. A description of the force field used in Paper I can be found in ref. [29]. While many parts of the two versions differ, the basic form of the two force fields is similar, with four terms $E = E_{\mathrm{ev}} + E_{\mathrm{loc}} + E_{\mathrm{hb}} + E_{\mathrm{sc}}$.

The first of the four terms is an excluded volume potential, which ensures that collisions between atoms do not happen.

The second term is needed to produce an accurate description of local distributions of torsion angles. While several features of these distributions can be understood in terms of excluded volume effects, others are hard to rationalize based on physical principles. The local term we use is therefore an effective term, fitted based on observed torsion angle distributions of folded proteins in the Protein Data Bank [30].

The third term represents hydrogen bonding. Hydrogen bonding is one of the major stabilizing forces in protein folding, and is largely responsible for the formation of the common secondary structures seen in folded proteins. They are formed through the interaction between a hydrogen atom bound to an electronegative atom (called the donor) and a lone pair of electrons on a different electronegative atom (the acceptor).

The strength of a single hydrogen bond in our model depends on the bond distance, as well as on orientation. Rather than directly summing the individual hydrogen bond contributions, however, we also explicitly ensure that each hydrogen atom participates in at most one bond, and each acceptor in at most two (one for each lone electron pair). This requires matching donors and acceptors such that each donor/acceptor is paired with the available acceptor/donors that give the lowest possible energy contribution.

The fourth and final term represents interactions between the protein side chains, due to electrostatics and (more importantly) hydrophobicity. Hydrophobic interactions play an important role in determining the global, tertiary structure of proteins, since hydrophobic parts of the protein chain "dislikes" contact with water and non-hydrophobic parts of the protein.

We calculate the degree of "buriedness" of a hydrophobic atom i as a sum of contact measures C_{ij} with other hydrophobic atoms j. To avoid overly concentrated conformations, however, C_{ij} is reduced if atom i also makes contact with an atom near or in the same amino acid as atom j.

The electrostatic energy is for simplicity calculated based on the same type of contact measure as the hydrophobicity. In this case, however, the pairwise terms are combined by a simple sum.

3.2 Coarse-grained force fields

For the Papers about liquid-liquid phase separation (II-IV), we need to run simulations with larger numbers of chains (at least ∼100). In order to ensure that the simulations remain feasible we therefore use simpler HP-type lattice or off-lattice models. Because of their simplicity, detailed comparisons with experiments on specific proteins are typically not fruitful. Nevertheless, they may be useful for identifying important factors governing large-scale properties.

In the HP model, each amino acid is represented by a single bead, and there are only two types of amino acid in the model (hydrophobic and polar, hence the name). The original model was defined on a lattice [19], with each bead occupying a lattice site, and beads adjacent along the chain occupying adjacent lattice sites. The interaction energy was taken as $E_{\mathrm{HP}} = -\epsilon N_{\mathrm{HH}}$, with N_{HH} being the number of hydrophobic beads which are adjacent on the lattice but not along the chain.

In Papers II and III we use an off-lattice HP model. The bonds have a fixed length, b, while the bond angles can vary freely. The bead-bead interaction has a square-well dependence on the distance r_{ij},

$$E_{ij} = \begin{cases} \infty, & \text{if } r_{ij} < d_{\mathrm{ev}} \\ \epsilon_{ij}, & \text{if } d_{\mathrm{ev}} < r_{ij} < \Lambda \\ 0, & \text{if } r_{ij} > \Lambda \end{cases}, \qquad (1)$$

where $d_{\mathrm{ev}} = 0.75b$, and $\Lambda = 2.0b$. The bead-bead interaction strength is set to $\epsilon_{ij} = \epsilon < 0$ for HH pairs, and $\epsilon_{ij} = 0$ otherwise.

In Paper IV, we use a variant of the lattice HP model, with finite same-site repulsion. For further details, see section 4.2 on field theory simulation.

4 Simulation methods

Along with the choice of force field, selecting an appropriate sampling technique is key to acquiring useful simulation results. A major consideration when selecting a method is of course the computational efficiency, but there are also other considerations. For instance, the choice may differ if one is interested in how a system evolves in time, or if one only wants to sample from an equilibrium distribution. Depending on the problem, several approaches are in use, including molecular dynamics, Monte Carlo and Langevin dynamics.

We have mostly used Monte Carlo methods, such as the Metropolis algorithm [31]. Among the advantages of these methods are that the move set can be customized to produce computationally efficient simulations. They also by design sample the desired probability distribution exactly (without e.g. time discretization errors). On the other hand, the updates are not meant to mimic the precise time-evolution of the system, so trajectories may be less well represented. Nevertheless, for large-scale processes associated with the crossing of free-energy barriers the trajectories should still be informative, at least as long as only local updates are used.

4.1 Monte Carlo sampling

The aim of any Monte Carlo algorithm is to sample conformations $\mathbf{r}$, from a probability distribution $P(\mathbf{r}) \propto e^{-\beta E(\mathbf{r})}$. For complicated, high-dimensional distributions, generating independent samples is often not feasible. For these cases, the Metropolis algorithm [31] offers an attractive alternative.

The Metropolis algorithm falls into the wider category of Markov Chain Monte Carlo (MCMC) algorithms. Here, each new sample $\mathbf{r}'$ is generated based only on the immediately preceding conformation $\mathbf{r}$. The particular algorithm can be specified through the (conditional) transition probability $W(\mathbf{r} \to \mathbf{r}')$.

When constructing an MCMC algorithm, it is common to enforce the condition

$$W(\mathbf{r} \to \mathbf{r}')P(\mathbf{r}) = W(\mathbf{r}' \to \mathbf{r})P(\mathbf{r}'), \tag{2}$$

known as detailed balance. Detailed balance ensures that the balance of probabilities between any pair of "nearby" conformations (i.e. any pair with non-zero transition probability) remains the same once the target distribution is reached. Assuming that the Markov chain is ergodic, detailed balance is sufficient but not necessary to ensure that the correct distribution is sampled.

In the Metropolis algorithm, detailed balance is enforced by splitting the transition probability as $W(\mathbf{r} \to \mathbf{r}') = F(\mathbf{r} \to \mathbf{r}')A(\mathbf{r} \to \mathbf{r}')$, with

$$A(\mathbf{r} \to \mathbf{r}') = \min\left(1, \frac{F(\mathbf{r}' \to \mathbf{r})}{F(\mathbf{r} \to \mathbf{r}')}\frac{P(\mathbf{r}')}{P(\mathbf{r})}\right). \tag{3}$$

The so-called proposal probabilities, F, can then be chosen arbitrarily, so long as they are ergodic, i.e. any conformation can be reached starting from any other conformation. The acceptance probabilities, A, ensure that detailed balance is obeyed.

Move set

In order to specify the proposal probabilities in the Metropolis algorithm, one has to select a suitable set of elementary moves. In order to achieve an efficient sampling, we use a mix of large-scale and local updates. The large-scale updates include rigid-body translations and rotations, as well as so-called pivot rotations. In a pivot rotation a one part of the molecule is rotated around an arbitrary axis relative to the rest of the molecule.

The local updates in the coarse-grained HP models are movements of one or a few consecutive beads. In the atomic models, local updates include rotations of individual side chains. In addition we use semi-local moves to update a few consecutive torsion angles along a chain. With only torsional degrees of freedom, a purely local update would require the iterative solution of a trigonometric equation, so we instead use a computationally cheaper first order approximation called biased Gaussian steps [32].

In the droplet simulations we also use a cluster update to move multiple chains at the same time [33]. In this update, inspired by an algorithm originally proposed for the Ising spin model [34], detailed balance is achieved by means of a stochastic cluster construction process rather than a Metropolis-style acceptance step.

Generalized ensembles

The basic Metropolis algorithm can of course be used with any target probability. Sometimes, sampling efficiency can be improved by simulating a different probability distribution from the one you eventually want to study. Some commonly used algorithms exploiting this freedom are:

- Simulated tempering [35–37], where the target probability is $P(\mathbf{r}, \beta) \propto e^{-\beta E(\mathbf{r})+g(\beta)}$. An important feature of this algorithm is that the inverse temperature $\beta = 1/k_{\mathrm{B}} T$ is treated as a dynamic variable. The free parameters $g(\beta)$ are often chosen so that the marginal distribution $P(\beta)$ is flat.
- Parallel tempering [38–40]. In this algorithm, several copies of the same system are simulated in parallell at different temperatures. Apart from conformational updates, one also attempts to swap pairs of temperatures between systems.
- The Wang-Landau algorithm [41, 42], where the target ensemble is $P(\mathbf{r}) \propto 1/g(E(\mathbf{r}))$ and the density of states, $g(E)$, is determined iteratively. The desired canonical ensemble is recovered through reweighting. There are also related methods, where the target probability is chosen to further optimize sampling efficiency [43, 44].

4.2 Field theory simulation

While the Monte Carlo-based simulation algorithms described above do work well for most of our simulations, the system sizes we simulate are still limited by computational considerations. A recently proposed alternative to conventional particle based simulation for droplet formation, is to rewrite the system as a field theory and simulate that [45, 18]. An appealing feature of this method is that the particle number appears only as a parameter, potentially making simulations with large numbers of particles less costly.

We use a lattice HP model with finite same-site repulsion, and nearest-neighbour interactions between pairs of H beads. For simplicity, we consider a system consisting of N identical HP chains. In order to transform the model to a field theory, the energy E, is rewritten in terms of the occupancies of all beads, $n(\mathbf{r})$, and of H-beads, $n_{\mathrm{H}}(\mathbf{r})$,

$$E = \frac{\Lambda}{2} \sum_{\mathbf{r}} n(\mathbf{r})^2 - \frac{1}{2} \sum_{\mathbf{r}} \sum_{k=\mathrm{x,y,z}} n_{\mathrm{H}}(\mathbf{r}) n_{\mathrm{H}}(\mathbf{r} + \hat{\mathbf{e}}_k). \tag{4}$$

This model can be transformed to a field theory through the Hubbard-Stratonovich method if each term is quadratic. To this end, the second term is written as a square of the vector field $\tilde{\mathbf{n}}(\mathbf{r}) = \sum_k (\alpha n_{\mathrm{H}}(\mathbf{r}) - \alpha^* n_{\mathrm{H}}(\mathbf{r} + \hat{\mathbf{e}}_k))\hat{\mathbf{e}}_k$, where $\alpha = (1 + i)/\sqrt{2}$.

In the Hubbard-Stratonovich method, the quadratic dependencies are eliminated by introducing auxiliary fields $w, \boldsymbol{\phi}$. After some calculations, we are left with a field theory partition function, $Z_{\mathrm{FT}} = e^{-H}$, with an effective Hamiltonian

$$H = \frac{1}{2\Lambda\beta} \sum_{\mathbf{r}} w(\mathbf{r})^2 + \frac{1}{2\beta} \sum_{\mathbf{r}} \boldsymbol{\phi}(\mathbf{r})^2 - N \log Q. \tag{5}$$

Here, Q is the partition function of a *single* chain, where each monomer, m, interacts with the fields through the energy $E_Q = i[w(\mathbf{r}_m) + \sigma_m \sum_k (\alpha \phi_k(\mathbf{r}_m - \hat{\mathbf{e}}_k) - \alpha^* \phi_k(\mathbf{r}))]$, and can be calculated analytically. σ_m indicates whether the monomer is hydrophobic ($\sigma_m = 1$) or not ($\sigma_m = 0$). As can be seen, the field theory Hamiltonian is complex-valued.

Complex Langevin simulation

Since the field-theory Hamiltonian is complex-valued, normal simulation methods, such as the Monte Carlo simulation techniques described in section 4.1, cannot be used. A possible way around the issue is to use complex Langevin sampling [46–48].

In this sampling algorithm, the system evolves according to the Langevin equation

$$\dot{w}(\mathbf{r}, t) = -\frac{\partial H}{\partial w(\mathbf{r}, t)} + \sqrt{2}\xi_w(\mathbf{r}, t) \tag{6}$$

$$\dot{\boldsymbol{\phi}}(\mathbf{r}, t) = -\frac{\partial H}{\partial \boldsymbol{\phi}(\mathbf{r}, t)} + \sqrt{2}\boldsymbol{\xi}_{\boldsymbol{\phi}}(\mathbf{r}, t) \tag{7}$$

where ξ_w and $\boldsymbol{\xi}_{\boldsymbol{\phi}}$ are sources of zero-mean unit-variance white noise. Langevin dynamics is a well-established method for sampling the Boltzmann distribution $\propto e^{-\beta H}$ when the Hamiltonian H is real-valued. For a complex-valued Hamiltonian, means of analytic functions can often be obtained from the same type of time-evolution, by allowing the simulated variables to drift off into the complex plane. However, it is known that this sometimes gives rise to instabilities [49].

5 Data analysis

Of course, performing an accurate and efficient simulation is not necessarily sufficient to gain insights into the underlying system. To do that, one also has to be able to extract relevant data from the simulation, which is not necessarily trivial.

Below, I will address two different challenges related to data extraction. First, the use of methods such as time-lagged independent component analysis (TICA) and Markov state modeling to systematically extract important coordinates from simulation data. Second, the use of finite-size scaling analysis for studying the transition in phase-separating systems.

5.1 Identifying slow-changing coordinates

The macromolecular systems we wish to study contain a vast number of degrees of freedom. Even a single short protein can contain hundreds, if not thousands of atoms. To figure out which coordinates correspond to physically and biologically relevant processes is not trivial. In folding or binding studies, we often start from an educated guess based on experimental data, but for multiprotein systems with many non-specific interactions, guessing the interactions is rarely feasible. In those situations, a systematic way of identifying the most relevant degrees of freedom is highly useful.

An important insight is that the most relevant coordinates in biomolecular systems are often those which change slowly [50]. Such coordinates can be constructed by combining the trajectories from a set of observables, $\{o_i(t)\}$, from the simulations, to some function, $u(\{o_i\})$, such that the normalized autocorrelation $C_u(\tau)/C_u(0)$ is maximal. Here,

$$C_u(\tau) = \langle u(t)u(t+\tau)\rangle - \langle u(t)\rangle\langle u(t+\tau)\rangle, \tag{8}$$

and the lag-time τ is a free parameter.

TICA

Perhaps the simplest way to construct the function u is to take it as a linear function of the input variables, $u = \sum a_i o_i$. The a_i are here parameters to be optimized. This approach, known as time-lagged independent component analysis [51, 52], has a lot in common with the well-known principal component analysis (PCA). The difference being that TICA determines high-autocorrelation rather that high-variance coordinates.

To find the slowest component, we need simply solve a linear optimization problem, using a Lagrange multiplier to handle the regularization. The problem is thus to find

$$\mathrm{argmax}_{\mathbf{a}} \mathbf{a}^T \mathsf{C}(\tau)\mathbf{a} - \lambda \mathbf{a}^T \mathsf{C}(0)\mathbf{a} \tag{9}$$

where $\mathbf{a}$ is a vector of the coefficients a_i, and $\mathsf{C}(\tau)$ is a matrix with elements $C_{ij}(\tau) = \langle o_i(t) o_j(t+\tau)\rangle - \langle o_i(t)\rangle\langle o_j(t+\tau)\rangle$. Solving for $\mathbf{a}$ gives the generalized eigenvalue equation

$$\mathsf{C}(\tau)\mathbf{a} = \lambda \mathsf{C}(0)\mathbf{a}. \tag{10}$$

Just as in PCA, the subsequent components can be found as solutions of the same equation, with each additional component corresponding to a lower value for the eigenvalue λ.

Markov state models

Of course, linear combinations of the input coordinates may not always be an ideal choice when constructing slow-changing functions. A popular method for constructing nonlinear functions has been to build so-called Markov state models [53, 50], wherein one first tesselates the conformation space into cells $\mathcal{C}_i$ through a clustering procedure. Then, one constructs a set of indicator coordinates $\theta_i(\mathbf{r})$ which depend on the conformation $\mathbf{r}$. The coordinates are such that $\theta_i = 1$, if the conformation is in cell i, and $\theta_i = 0$ otherwise. Slow coordinates can then be constructed as linear combinations of the indicator variables, in the same way as for TICA.

This construction also has an appealing interpretation, wherein each cell defines a state. The autocorrelation matrix is then closely related to the transition matrix with elements corresponding to the probability of moving from state i to state j in the chosen lag-time τ. If the number of states is small, this interpretation may yield a useful intuitive view of the dynamics.

In practice, the number of states required to achieve a good description is typically large, thereby limiting the interpretability of the Markov state model. Even then though, it can be used e.g. to achieve a more precise determination of the time scales involved in the slow processes.

5.2 Finite-size effects

When studying biomolecular condensates, the underlying concentration-temperature phase diagram is of key interest. From simulations, points on the phase boundary can of course be determined, e.g. as the droplet formation temperature in constant-density simulations. However, simulations necessarily deal with finite systems, where the phase diagram turns out to be distorted. Developing methods to minimize and control these finite-size effects is important.

The main source of finite-size effects is the interplay between the bulk free energies of the two phases, and the interfacial free energy, which scale differently with system size [54–56]. A simple but useful phenomenological ansatz for understanding finite-size effects, based on the different scaling of the bulk and interface terms with system size, is given by [55, 57]

$$F(\rho_\ell, \rho_h, V_\ell, V_h) = f_\ell(\rho_\ell)V_\ell + f_h(\rho_h)V_h + \gamma A_{\ell h}. \tag{11}$$

Here ρ_i, V_i and f_i ($i = h, \ell$) are the single-phase densities, volumes and free-energy densities respectively. $A_{\ell h}$ is the interface area and γ is the surface tension. To find the single-phase densities and volumes for a system with given volume V and density ρ, this expression should be minimized subject to the constraints $V = V_\ell + V_h$ and $\rho V = \rho_\ell V_\ell + \rho_h V_h$. The large-volume limit corresponds to neglecting the surface term, and leading order corrections can be found by assuming that this term, as well as the density shifts, are small (for more detail on this calculation, see Paper III). The first-order correction to the single-phase densities turns out to be proportional to $dA_{\ell h}/dV_h$. For a spherical droplet this correction is therefore inversely proportional to the droplet radius, while it vanishes for a slab-like droplet (see Paper III for further details).

Equation 11 can also be used to determine the finite-size scaling of the transition densities, see e.g. [55]. In this case, however, the relevant physics can also be inferred from the following back-of-an-envelope calculation [54].

To determine whether a droplet forms at density $\rho = \rho^\infty + \delta\rho$ (where ρ^∞ is the infinite-size droplet formation density), we must determine whether the free-energy cost of using the excess density $\delta\rho$ to form a droplet is lower than the cost of absorbing it into the dilute phase. The cost for absorbing the excess scales (to lowest order) as $\delta\rho^2 V$, whereas the cost of making a droplet is proportional to the interface area, $A_{\ell h} \sim V_h^{2/3} \sim (\delta\rho V)^{2/3}$. Droplet formation should set in when the two costs are comparable, i.e. when $\delta\rho \sim V^{-1/4}$.

When doing the full calculation [54, 55], it turns out that even when a droplet does form, some of the excess density will remain in the background. Despite this, the basic scaling law remains the same.

It is interesting to note that in both the above cases, the finite-size shifts are inversely proportional to the droplet radius. The different scalings (as $V^{-1/3}$ for the single-phase densities, and $V^{-1/4}$ for the transition densities) are due to the fact that the relative size of the droplet, V_h/V, at the transition density decreases with system size.

In the large-V limit, the transition and single-phase densities both approach the coexistence densities. The above scaling relations thus offer two ways to determine coexistence densities in large systems. Similar considerations apply when studying other observables.

6 Summary

In this thesis, I have developed and evaluated biophysical models, simulation algorithms and data analysis tools for exploring new and challenging areas in biomolecular physics, such as crowding effects in dense protein systems and liquid-liquid phase separation. These efforts can be summarized as follows.

- Development of an all-atom force field designed to be able to handle intrinsically disordered as well as globular proteins.
- Identification of important non-specific intermolecular interactions in simulations of crowded protein systems through data analysis methods (TICA and Markov state modeling).
- Exploration of the potential and limitations of field-theory simulations of biomolecular liquid-liquid phase separation.
- Investigation of finite-size effects on phase diagrams from simulations of protein droplet formation and their dependence on system geometry.

References

[1] Vendeville, A., D. Larivière, and E. Fourmentin, 2011. An inventory of the bacterial macromolecular components and their spatial organization. *FEMS Microbiol. Rev.* 35:395–414.

[2] Zhou, H.-X., G. Rivas, and A. P. Minton, 2008. Macromolecular Crowding and Confinement: Biochemical, Biophysical, and Potential Physiological Consequences. *Annu. Rev. Biophys.* 37:375–397.

[3] Rivas, G., and A. P. Minton, 2016. Macromolecular Crowding In Vitro, In Vivo, and In Between. *Trends Biochem. Sci.* 41:970–981.

[4] Ostrowska, N., M. Feig, and J. Trylska, 2019. Modeling crowded environment in molecular simulations. *Front. Mol. Biosci.* 6:86.

[5] Brangwynne, C. P., C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, and A. A. Hyman, 2009. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* 324:1729–1732.

[6] Banani, S. F., H. O. Lee, A. A. Hyman, and M. K. Rosen, 2017. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* 18:285–298.

[7] Guzman, I., H. Gelman, J. Tai, and M. Gruebele, 2014. The extracellular protein VlsE is destabilized inside cells. *J. Mol. Biol.* 426:11–20.

[8] Danielsson, J., X. Mu, L. Lang, H. Wang, A. Binolfi, F.-X. Theillet, B. Bekei, D. T. Logan, P. Selenko, H. Wennerström, and M. Oliveberg, 2015. Thermodynamics of protein destabilization in live cells. *Proc. Natl. Acad. Sci. USA* 112:12402–12407.

[9] McGuffee, S. R., and A. H. Elcock, 2010. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.* 6:e1000694.

[10] Yu, I., T. Mori, T. Ando, R. Harada, J. Jung, Y. Sugita, and M. Feig, 2016. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *Elife* 5:e19274.

[11] Bille, A., K. S. Jensen, S. Mohanty, M. Akke, and A. Irbäck, 2019. Stability and Local Unfolding of SOD1 in the Presence of Protein Crowders. *J. Phys. Chem. B* 123:1920–1930.

[12] Nott, T. J., E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, and A. J. Baldwin, 2015. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Makromol. Chem., Theory Simul.* 57:936–947.

[13] Molliex, A., J. Temirov, J. Lee, M. Coughlin, A. P. Kanagaraj, H. J. Kim, T. Mittag, and J. P. Taylor, 2015. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* 163:123–133.

[14] Burke, K. A., A. M. Janke, C. L. Rhine, and N. L. Fawzi, 2015. Residue-by-residue View of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II. *Makromol. Chem., Theory Simul.* 60:231–241.

[15] Dignon, G. L., W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal, 2018. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* 14:e1005941.

[16] Das, S., A. N. Amin, Y.-H. Lin, and H. S. Chan, 2018. Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. *Phys. Chem. Chem. Phys.* 20:28558–28574.

[17] Robichaud, N. A. S., I. Saika-Voivod, and S. Wallin, 2019. Phase behavior of blocky charge lattice polymers: Crystals, liquids, sheets, filaments, and clusters. *Phys. Rev. E* 100:052404.

[18] McCarty, J., K. T. Delaney, S. P. O. Danielsen, G. H. Fredrickson, and J.-E. Shea, 2019. Complete phase diagram for liquid–liquid phase separation of intrinsically disordered proteins. *J. Phys. Chem. Lett.* 10:1644–1652.

[19] Lau, K. F., and K. A. Dill, 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22:3986–3997.

[20] Senn, H. M., and W. Thiel, 2009. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Edit.* 48:1198–1229.

[21] Cragnell, C., E. Rieloff, and M. Skepö, 2018. Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions. *J. Mol. Biol.* 430:2478–2492.

[22] Latham, A. P., and B. Zhang, 2019. Maximum entropy optimized force field for intrinsically disordered proteins. *J. Chem. Theory Comput.* 16:773–781.

[23] Regy, R. M., J. Thompson, Y. C. Kim, and J. Mittal, 2021. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* 30:1371–1379.

[24] Schmid, N., A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, 2011. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* 40:843–856.

[25] Huang, J., and A. D. MacKerell Jr, 2013. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.* 34:2135–2145.

[26] Maier, J. A., C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* 11:3696–3713.

[27] Choi, J.-M., and R. V. Pappu, 2019. Improvements to the ABSINTH Force Field for Proteins Based on Experimentally Derived Amino Acid Specific Backbone Conformational Statistics. *J. Chem. Theory Comput.* 15:1367–1382.

[28] Davtyan, A., N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, 2012. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B* 116:8494–8503.

[29] Irbäck, A., S. Mitternacht, and S. Mohanty, 2009. An effective all-atom potential for proteins. *PMC biophysics* 2:1–24.

[30] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242. `www.rcsb.org`.

[31] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* 21:1087–1092.

[32] Favrin, G., A. Irbäck, and F. Sjunnesson, 2001. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.* 114:8154–8158.

[33] Irbäck, A., S. Æ. Jónsson, N. Linnemann, B. Linse, and S. Wallin, 2013. Aggregate geometry in amyloid fibril nucleation. *Phys. Rev. Lett.* 110:058101.

[34] Swendsen, R. H., and J.-S. Wang, 1987. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* 58:86–88.

[35] Lyubartsev, A. P., A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, 1992. New approach to Monte Carlo calculation of the free energy: method of expanded ensembles. *J. Chem. Phys.* 96:1776–1783.

[36] Marinari, E., and G. Parisi, 1992. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* 19:451–458.

[37] Irbäck, A., and F. Potthast, 1995. Studies of an off‐lattice model for protein folding: Sequence dependence and improved sampling at finite temperature. *J. Chem. Phys.* 103:10298–10305.

[38] Tesi, M. C., E. J. J. van Rensburg, E. Orlandini, and S. G. Whittington, 1996. Monte Carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys.* 82:155–181.

[39] Hukushima, K., and K. Nemoto, 1996. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. (Jap)* 65:1604–1608.

[40] Hansmann, U. H., 1997. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* 281:140–150.

[41] Wang, F., and D. P. Landau, 2001. Efficient, multiple-range random walk algorithm to calculate density of states. *Phys. Rev. Lett.* 86:2050–2053.

[42] Jónsson, S. Æ., S. Mohanty, and A. Irbäck, 2011. Accelerating atomic-level protein simulations by flat-histogram techniques. *J. Chem. Phys.* 135:125102.

[43] Trebst, S., D. A. Huse, and M. Troyer, 2004. Optimizing the ensemble for equilibration in broad-histogram Monte Carlo simulations. *Phys. Rev. E* 70:046701.

[44] Lindahl, V., J. Lidmar, and B. Hess, 2018. Riemann metric approach to optimal sampling of multidimensional free-energy landscapes. *Phys. Rev. E* 98:023312.

[45] Fredrickson, G. H., V. Ganesan, and F. Drolet, 2002. Field-theoretic computer simulation methods for polymers and complex fluids. *Macromolecules* 35:16–39.

[46] Parisi, G., 1983. On complex probabilities. *Phys. Lett. B* 131:393–395.

[47] Klauder, J. R., 1983. A Langevin approach to fermion and quantum spin correlation functions. *J. Phys. A: Math. Gen.* 16:L317–L319.

[48] Söderberg, B., 1988. On the complex Langevin equation. *Nucl. Phys. B* 295:396–408.

[49] Ambjørn, J., M. Flensburg, and C. Peterson, 1986. The complex Langevin equation and Monte Carlo simulations of actions with static charges. *Nucl. Phys. B* 275:375–397.

[50] Pérez-Hernández, G., F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, 2013. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* 139:015102.

[51] Molgedey, L., and H. G. Schuster, 1994. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* 72:3634–3637.

[52] Naritomi, Y., and S. Fuchigami, 2013. Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *J. Chem. Phys.* 139:215102.

[53] Schwantes, C. R., and V. S. Pande, 2013. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.* 9:2000–2009.

[54] Biskup, M., L. Chayes, and R. Kotecký, 2002. On the formation/dissolution of equilibrium droplets. *Europhys. Lett.* 60:21–27.

[55] MacDowell, L. G., P. Virnau, M. Müller, and K. Binder, 2004. The evaporation/condensation transition of liquid droplets. *J. Chem. Phys.* 120:5293–5308.

[56] Zierenberg, J., and W. Janke, 2015. Exploring different regimes in finite-size scaling of the droplet condensation-evaporation transition. *Phys. Rev. E* 92:012134.

[57] Schrader, M., P. Virnau, and K. Binder, 2009. Simulation of vapor-liquid coexistence in finite volumes: a method to compute the surface free energy of droplets. *Phys. Rev. E* 79:061104.

Scientific publications

Paper I

Markov modeling of peptide folding in the presence of protein crowders
D. Nilsson, S. Mohanty, A. Irbäck
The Journal of Chemical Physics **148**, 055101 (2018)

In this Paper, we applied Markov state modeling techniques to crowding simulations of the GB1m3 peptide in the prescence of either BPTI of GB1 crowders. These systems had been studied in the group before I joined as a PhD student, and the paper could be seen as a follow-up of this work. The aim of the paper was to see whether Markov state modeling could be useful in analysing crowding simulations. We found that dimensional reduction through TICA was very helpful for characterizing the systems, while the further analysis provided only limited insights.

I performed all the simulations as well as the TICA and Markov state modeling calculations. The simulations were performed using code written and maintained by Sandipan Mohanty. The manuscript was mostly written by Anders Irbäck, with input from the other authors.

Paper II

Finite-size scaling analysis of protein droplet formation
D. Nilsson, A. Irbäck
Physical Review E **101**, 022413 (2020)

This paper was our first foray into simulation of protein droplet formation. The aim was to probe the usefulness of finite-size scaling analysis in determining the values of biophysical observables for large systems, from simulations which are necessarily performed at finite size. To do this, we simulated systems of two different model proteins in a continuous hydrophobic-polar model, and could show that one of them phase-separated while the other did not.

I formulated the model in discussion with Anders Irbäck. I wrote the simulation code, ran the simulations, and analysed the data. The manuscript was mostly written by Anders Irbäck, with input from me.

Paper III

Finite-size shifts in simulated protein droplet phase diagrams
D. Nilsson, A. Irbäck
The Journal of Chemical Physics **154**, 235101 (2021)

When determining the droplet formation temperature in Paper II, we identified the maximum of the specific heat, at constant density. Another commonly used method of determining the same quantity is to measure the coexistence densities of the two phases in simulations at constant temperature. After discussions with Anders Irbäck about how using this method would affect the finite-size shifts, I came up with an expression for the finite-size shifts (eq. 4 in the paper). The form of this expression also naturally lead us to the question of how the simulation geometry influences the size of the shifts. We found that when determining coexistence densities through direct measurement of equilibrium densities, an elongated simulation geometry results in drastically smaller shifts.

I derived the expression for finite-size shifts of the coexistence densities. I ran the simulations, which were run using the code developed for Paper II, and analysed the data. I wrote the first draft of the manuscript, which was then revised by Anders Irbäck and me.

Paper IV

Limitations of field-theory simulation for exploring phase separation: the role of repulsion in a lattice protein model
D. Nilsson, B. Bozorg, S. Mohanty, B. Söderberg, A. Irbäck
Submitted to *The Journal of Chemical Physics*

This project was started because we wanted to try out field theory simulations, which were proposed as a fruitful way of studying LLPS in [18]. To this end, we initially tried performing field-theory simulations using a continuous HP model similar to the model used in that paper. However, we had difficulties getting the simulations to output sensible results, and therefore formulated a lattice model that had fewer free parameters and could be exactly compared to particle-based simulations. We found that the results from field-theory and particle-based simulations coincided only when the repulsion strength was so low that the model droplets collapsed to an artificially dense state. We set up a toy model to investigate the sampling issues, and identified a loss of ergodicity as a possible cause.

Initial code development, simulations and analysis were done by Behruz Bozorg, with contributions from Sandipan Mohanty to the code development. By the time I started to become actively involved, the cause of the discrepancies between field theory and particle based simulations was still unclear. I helped adapting the field theory formulation to the lattice model, together with Bo Söderberg and Anders Irbäck. I formulated the toy model. Bo Söderberg did additional calculations with the toy model. Field theory simulations for the lattice model were performed by Anders Irbäck, and particle-based simulations were performed by Anders Irbäck and Sandipan Mohanty. The paper was mostly written by Anders Irbäck and Bo Söderberg, with input from the other authors.

Paper v

An effective potential for atomic-level simulation of structured and unstructured proteins
D. Nilsson, S. Mohanty, A. Irbäck
Manuscript

In this project, we updated the protein folding model developed in the group. The main aim was to improve the performance of the model when sampling intrinsically disordered peptides. We did this by comparing model results to experimental data for a set of small peptides, both ordered and disordered.

Design of the model was done mostly by me, with important contributions from Anders Irbäck. I did most of the coding and parameterization of the force field. Sandipan Mohanty reimplemented the model in the protein simulation software package PROFASI, which among other things enabled multi-threaded simulation of the final model. I wrote the first draft of the manuscript, which was then revised by all authors.

Paper 1

Markov modeling of peptide folding in the presence of protein crowders

Daniel Nilsson,[1,a)] Sandipan Mohanty,[2,b)] and Anders Irbäck[1,c)]
[1]*Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden*
[2]*Institute for Advanced Simulation, Jülich Supercomputing Centre, Forschungszentrum Jülich, D-52425 Jülich, Germany*

(Received 23 November 2017; accepted 10 January 2018; published online 1 February 2018)

We use Markov state models (MSMs) to analyze the dynamics of a β-hairpin-forming peptide in Monte Carlo (MC) simulations with interacting protein crowders, for two different types of crowder proteins [bovine pancreatic trypsin inhibitor (BPTI) and GB1]. In these systems, at the temperature used, the peptide can be folded or unfolded and bound or unbound to crowder molecules. Four or five major free-energy minima can be identified. To estimate the dominant MC relaxation times of the peptide, we build MSMs using a range of different time resolutions or lag times. We show that stable relaxation-time estimates can be obtained from the MSM eigenfunctions through fits to autocorrelation data. The eigenfunctions remain sufficiently accurate to permit stable relaxation-time estimation down to small lag times, at which point simple estimates based on the corresponding eigenvalues have large systematic uncertainties. The presence of the crowders has a stabilizing effect on the peptide, especially with BPTI crowders, which can be attributed to a reduced unfolding rate k_{u}, while the folding rate k_{f} is left largely unchanged. *Published by AIP Publishing.* https://doi.org/10.1063/1.5017031

I. INTRODUCTION

In the crowded interior of living cells, proteins are surrounded by high concentrations of macromolecules, which leads to a reduction of the volume available to a given protein. Under such conditions, steric interactions would universally favor more compact structures. A growing body of evidence indicates, however, that the effects of macromolecular crowding on properties such as protein stability cannot be explained in terms of steric repulsion alone.[1–3] To understand the role of other interactions, in recent years, there have been increasing efforts to perform computer simulations with realistic crowder molecules,[4–11] rather than hard-sphere crowders. When analyzing these large systems, a major challenge lies in identifying the main states and dynamical modes, which may not be easily anticipated. One possible approach to this problem is provided by Markov modeling techniques,[12–16] which in recent years have found widespread use in studies of biomolecular processes such as folding and binding.[17,18] Most of these studies dealt with data from molecular dynamics simulations, but the methods are general and can be used on Monte Carlo (MC) data as well.

In this article, we use Markov modeling, along with time-lagged independent component analysis (TICA),[19–22] to analyze data from MC simulations of a test peptide in the presence of interacting protein crowders, for two different types of crowder proteins. We show that the major free-energy minima and slow dynamical modes of these high-dimensional systems can be identified in a systematic manner using TICA and Markov state models (MSMs). We further show that the dominant MC relaxation times of the peptide can be robustly estimated from the constructed MSMs, although simple estimates based on the MSM eigenvalues are subject to well-known systematic uncertainties. Our procedure for relaxation-time estimation uses the MSM eigenfunctions and autocorrelation fits, rather than the eigenvalues.

As a test molecule, we use the β-hairpin-forming GB1m3 peptide.[23] The peptide is simulated in homogeneous crowding environments, where either the bovine pancreatic trypsin inhibitor (BPTI) or the B1 domain of streptococcal protein G (GB1) serves as a crowding agent. Both these proteins are thermally highly stable[24,25] and therefore modeled using a fixed-backbone approximation, whereas the GB1m3 peptide is free to fold and unfold in the simulations. The simulations are conducted using MC dynamics at a constant temperature. Recently, we studied the same systems using MC replica-exchange methods and found that both BPT1 and GB1 have a stabilizing effect on GB1m3.[26]

II. METHODS

A. Simulated systems

The simulated systems consist of one GB1m3 molecule and eight crowder molecules, enclosed in a periodic box with side length 95 Å. The eight crowder molecules are copies of a single protein, either BPTI or GB1. This setup yields crowder densities of ~100 mg/mL, whereas the macromolecule densities in cells can be ~300–400 mg/mL.[27] The volume fraction occupied by the crowders is around 7%. The simulation temperature is set to 332 K, which is near the melting temperature of the free GB1m3 peptide.[23] For reference, simulations of the free peptide are also performed, using the same temperature.

[a)]Electronic mail: daniel.nilsson@thep.lu.se
[b)]Electronic mail: s.mohanty@fz-juelich.de
[c)]Electronic mail: anders@thep.lu.se

The GB1m3 peptide is an optimized variant of the second β-hairpin (residues 41–56) in protein GB1, with enhanced stability.[23] It differs from the original sequence at 7 of 16 positions. To the best of our knowledge, no experimental structure is available for GB1m3, but its native fold is expected to be similar to the parent β-hairpin in GB1.

B. Biophysical model

Our simulations are based on an all-atom protein representation with torsional degrees of freedom and an implicit solvent force field.[28] A detailed description of the interaction potential can be found elsewhere.[28] In brief, the potential consists of four main terms, $E = E_{\mathrm{loc}} + E_{\mathrm{ev}} + E_{\mathrm{hb}} + E_{\mathrm{sc}}$. One term ($E_{\mathrm{loc}}$) represents local interactions between atoms separated by only a few covalent bonds. The other, non-local terms represent excluded-volume effects (E_{ev}), hydrogen bonding (E_{hb}), and residue-specific interactions between pairs of side-chains, based on hydrophobicity and charge (E_{sc}). In multi-chain simulations, intermolecular interaction terms have the same form and strength as the corresponding intramolecular ones. The potential is an effective energy function, parameterized through folding thermodynamics studies for a structurally diverse set of peptides and small proteins.[28] Previous applications of the model include folding/unfolding studies of several proteins with >90 residues.[29–34] Recently, it was used by us to simulate the peptides trp-cage and GB1m3 in the presence of protein crowders.[8,26]

Our simulations use the same fully atomistic representation of both the GB1m3 peptide and the crowder proteins. However, because of their high thermal stability,[24,25] the crowder proteins are modeled with a fixed backbone and thus with side-chain rotations as their only internal degrees of freedom. The assumed backbone conformations of BPTI and GB1 are model approximations of the Protein Data Bank (PDB) structures 4PTI and 2GB1, derived by MC with minimization. The structures were selected for both low energy and high similarity to the experimental structures. The root-mean-square deviations (RMSDs) from the experimental structures (calculated over backbone and C^{β} atoms) are $\lesssim 1$ Å.

C. MC simulations

The systems are simulated using MC dynamics. The simulations are done in the canonical rather than some generalized ensemble. Also, only "small-step" elementary moves are used so that the system cannot artificially jump between free-energy minima, without having to climb intervening barriers. With these restrictions, the simulations should capture some basics of the long-time dynamics.[35] Despite the restrictions, the methods are sufficiently fast to permit the study of the folding and binding thermodynamics of the peptide, through simulations containing multiple folding/unfolding and binding/unbinding events.

Our move set consists of four different updates: (i) the semi-local Biased Gaussian Steps (BGSs) method[36] for backbone degrees of freedom in the peptide, (ii) simple single-angle Metropolis updates in side chains, (iii) small rigid-body translations of whole chains, and (iv) small rigid-body rotations of whole chains. The "time" unit of the simulations is MC sweeps, where one MC sweep consists of one attempted update per degree of freedom. Specifically, each MC sweep consists of 74 attempted moves in the crowder-free system, whereas the corresponding numbers are 1208 and 1328 with BPTI and GB1 crowders, respectively. Note that the average number of attempted conformational updates of the peptide per MC sweep is the same in all three cases. In the simulations with crowders, the relative fractions of BGS moves, side-chain updates, rigid-body translations, and rigid-body rotations are approximately 0.02, 0.94, 0.02, and 0.02, respectively.

All simulations are run with the program PROFASI,[37] using both vector and thread parallelization. To gather statistics, a set of independent runs is generated for each system. The number of runs is 16 with BPTI crowders, 62 with GB1 crowders, and 30 for the isolated peptide. Each run comprises 40×10^6 MC sweeps if crowders are present and 10×10^6 MC sweeps without crowders. Compared to the longest relaxation times in the respective systems (see below), the individual runs are a factor >20 longer.

Several different properties are recorded during the simulations. As a measure of the nativeness of the peptide, the number of native H bonds present, n_{hb}, is computed, assuming that the native H bonds are the same as in the full GB1 protein (PDB code 2GB1). The interaction of the peptide with surrounding crowder molecules is studied by monitoring intermolecular H bonds and C^{α}–C^{α} contacts. A residue pair is said to be in contact if their C^{α} atoms are within 8 Å.

As input for our TICA and MSM analyses (see below), two sets of parameters are stored at regular intervals during the course of the simulations. The first set consists of all (non-constant) intramolecular C^{α}–C^{α} distances within the peptide, called r_{ij}. The second set consists of intermolecular distances between the peptide and the crowders, called d_{ij}. Specifically, d_{ij} denotes the shortest C^{α}–C^{α} distance between peptide residue i and residue j in any of the crowder molecules.

D. TICA and MSM analysis

TICA can be used as a dimensionality reduction method. It is somewhat similar to the principal component analysis but identifies high-autocorrelation (or slow) rather than high-variance coordinates. Given time trajectories of a set of parameters, $\{o_n\}$ (in our case, the distances r_{ij} and d_{ij}, see above), one constructs the time-lagged covariance matrix $c_{nm}(\tau_{\mathrm{cm}}) = \langle o_n(t) o_m(t+\tau_{\mathrm{cm}})\rangle_t - \langle o_n(t)\rangle_t \langle o_m(t+\tau_{\mathrm{cm}})\rangle_t$, where τ_{cm} is the lag time and $\langle\cdot\rangle_t$ denotes an average over time t. By solving the (generalized) eigenvalue problem $\mathbf{C}(\tau_{\mathrm{cm}})\hat{\mathbf{v}}_i = \hat{\lambda}_i \mathbf{C}(0)\hat{\mathbf{v}}_i$, slow linear combinations of the original parameters can be identified. A more advanced method for identifying slow modes is to construct MSMs.

To build an MSM, the state space needs to be discretized. In our calculations, following Ref. 22, the discretization is achieved by clustering the data with the k-means algorithm[38] in a low-dimensional subspace spanned by slow TICA coordinates. By computing the probabilities of transition among these clusters in a time τ_{tm} (which, like τ_{cm}, is an adjustable parameter), a transition matrix is obtained. Assuming Markovian dynamics, the eigenvectors of this matrix have relaxation times given by

$$\tilde{t}_i = -\tau_{\text{tm}} / \ln \tilde{\lambda}_i(\tau_{\text{tm}}), \tag{1}$$

where $1 = \tilde{\lambda}_0 > \tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots > 0$ are the eigenvalues. The eigenvalue $\tilde{\lambda}_0$ corresponds to a stationary distribution ($\tilde{t}_0 = \infty$), whereas all other eigenvalues correspond to relaxation modes with finite time scales $\tilde{t}_i$. The time scales obtained using Eq. (1) are expected to reproduce the dominant relaxation times of the full system if the discretization is sufficiently fine[39,40] or if the lag time is sufficiently large.[40,41]

There are several software packages available for TICA and MSM analysis.[42–45] Our calculations are done using the pyEMMA software.[42]

E. Time scales from autocorrelations of MSM eigenfunctions

Another way of estimating relaxation times from an MSM is to compute autocorrelations of the eigenfunctions. The (normalized) autocorrelation function of a general property f is given by $C_f(\tau) = [\langle f(t)f(t+\tau)\rangle_t - \langle f(t)\rangle_t \langle f(t+\tau)\rangle_t]/\sigma_f^2$, where σ_f^2 is the variance of f. Let ψ_i^{MSM} be the ith eigenfunction of a given MSM, and let ψ_i be the true ith eigenfunction of the system's time transfer operator.[16] The autocorrelation function of ψ_i^{MSM}, $C_i(\tau)$, may be expanded as

$$C_i(\tau) = \sum_j c_j \mathrm{e}^{-\tau/t_j}, \tag{2}$$

where t_j is the exact jth relaxation time. The coefficients c_j are given by $c_j = |\langle \psi_j, \psi_i^{\text{MSM}} \rangle|^2$, where the overlap $\langle \psi_j, \psi_i^{\text{MSM}} \rangle$ can be expressed as an average with respect to the stationary distribution, $\mu(x)$: $\langle \psi_j, \psi_i^{\text{MSM}} \rangle = \int dx \mu(x) \psi_j(x) \psi_i^{\text{MSM}}(x)$. Note that ψ_j and ψ_i^{MSM} have mean zero and unit norm. In Sec. III, overlaps between pairs of general functions are computed in the same way, after shifting and normalizing the functions to zero mean and unit norm.

Now, if ψ_i^{MSM} is a good approximation of ψ_i, then $c_j \ll c_i$ for $j \neq i$. If this holds, $C_i(\tau)$ decays approximately as $\mathrm{e}^{-\tau/t_i}$ for not too large τ (compared to t_i) so that t_i can be estimated through a simple exponential fit.

The calculations discussed below use data for $C_i(\tau)$ in the range of τ where $0.2 < C_i(\tau) < 0.8$. Over this range, $C_i(\tau)$ is to a good approximation single exponential for all MSM eigenfunctions studied. The upper bound on τ is set primarily by statistical uncertainties, rather than by deviations from single-exponential behavior.

III. RESULTS

Our analysis of the GB1m3 peptide in the three simulated systems (with BPTI crowders, with GB1 crowders, without crowders) can be divided into two parts. First, equilibrium free-energy surfaces are constructed, using TICA coordinates. Second, the dynamics are investigated, using MSM techniques.

A. Free-energy landscapes

It is instructive to begin with the isolated GB1m3 peptide, whose folding thermodynamics have been studied before using the same model.[28] This study found that the isolated peptide folds in a cooperative manner, and that the number of native H bonds present, n_{hb}, is a useful folding coordinate that has a bimodal distribution at the melting temperature. Figure 1(a) shows the free energy of the isolated GB1m3, calculated as a function of the two slowest TICA coordinates, TIC0 and TIC1. The free-energy surface exhibits two major minima, labeled I and II, which are well separated in the TIC0 direction. From Fig. 1(b), it can be seen that this coordinate is strongly (anti-) correlated with n_{hb}. This correlation implies that the peptide is native-like in free-energy minimum I and unfolded in minimum II.

We now turn to the system where GB1m3 is surrounded by BPTI crowders. Here, the TICA coordinates are linear combinations of both intra- and intermolecular distances (r_{ij} and d_{ij}; see Sec. II C). Calculated as a function of the two slowest TICA coordinates, the free energy exhibits four major minima, labeled I–IV [Fig. 2(a)]. To characterize the minima, an interpretation of the TIC0 and TIC1 coordinates is needed. As in the previous case, TIC0 is strongly correlated with n_{hb} [Fig. 2(b)] and thus linked to the degree of nativeness. Inspection of the eigenvector corresponding to TIC1 suggests that this coordinate depends strongly on certain peptide-crowder distances d_{ij} involving the BPTI residue Pro8, which is part of a sticky patch on the BPTI surface.[26] Motivated by this observation, Fig. 2(c) shows the TIC0,TIC1-dependence of a function defined to be unity whenever there is at least one residue-pair contact between the peptide and a Pro8 BPTI residue and zero otherwise (smoothing is used). This function is indeed strongly correlated with TIC1. Therefore, the main free-energy minima can be classified based on whether or not the peptide is native-like, and whether or not the peptide forms any Pro8 BPTI contact. The peptide is native-like and bound in minimum I, which actually can be split into two distinct subminima, corresponding to two preferred orientations of the folded and bound peptide. In the remaining three main minima, the peptide is either unfolded and bound (minimum II), native-like and unbound (minimum III), or unfolded and unbound (minimum IV).

With GB1 crowders, the free energy of GB1m3 exhibits five well-separated and easily visible minima [Fig. 3(a)] when

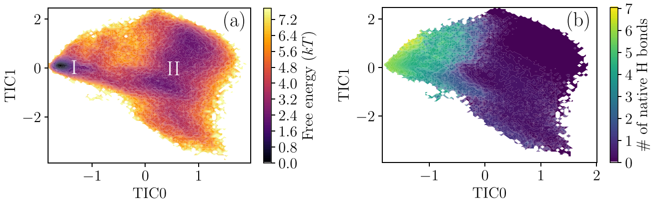


FIG. 1. (a) Free energy of the isolated GB1m3 peptide, calculated as a function of the two slowest TICA coordinates, TIC0 and TIC1. Major minima are labeled by Roman numerals. (b) The dependence of the number of native H bonds, n_{hb}, on these coordinates. Here, each stored conformation is represented by a point in the TIC0, TIC1-plane, in a color determined by the value of n_{hb}. Smoothing is applied to improve readability. The TICA lag time is set to $\tau_{\text{cm}} = 10^3$ MC sweeps.

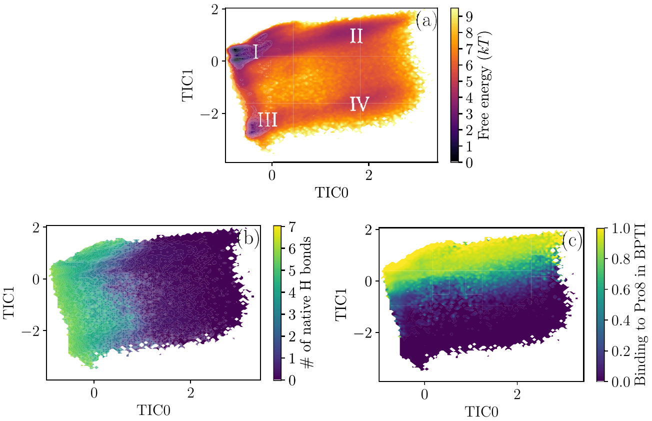


FIG. 2. Characterization of the GB1m3 peptide in the presence of BPTI crowders, using the two slowest TICA coordinates, TIC0 and TIC1. (a) Free energy. Major minima are labeled by Roman numerals. (b) The number of native H bonds present in the peptide, n_{hb}. (c) A function which is unity whenever there is at least one residue-pair C^{α}–C^{α} contact between the peptide and a Pro8 BPTI residue and zero otherwise (drawn using smoothing). The contact cutoff distance is 8 Å. The TICA lag time is set to $\tau_{cm} = 10^3$ MC sweeps.

calculated as a function of the slowest and third-slowest TICA coordinates. The TIC0, TIC2-plane is used here because two of the minima (III and IV) cannot be distinguished in the TIC0, TIC1-plane (see the supplementary material, Fig. S1). The TIC0 coordinate is again correlated with the degree of nativeness of the peptide [Fig. 3(b)]. Proper interpretation of the TIC2 coordinate requires knowledge of the preferred peptide-crowder binding modes. It turns out that there are two preferred binding modes, called B1 and B2. In both cases, binding occurs through β-sheet extension; the edge strand β3 (residues 42–46) of GB1 binds to either the first or the second strand of the folded GB1m3 β-hairpin. The binding modes can be described in terms of the H bonds involved (see the supplementary material, Fig. S2). Figures 3(c) and 3(d) show how the presence of H bonds associated with the respective modes vary with TIC0 and TIC2. Apparently, low and high TIC2 signal B1 and B2 binding, respectively. A similar analysis of TIC1 shows that this coordinate separates bound and unbound states but discriminates poorly between the B1 and B2 modes [see the supplementary material, Figs. S1(c) and (d)]. The isolated island at low TIC0 and intermediate TIC2 stems from simultaneous binding of the peptide via both modes, to two crowder molecules. Based on the above observations, the free-energy minima in Fig. 3(a) can be described as follows. In minima I and II, the peptide is unfolded and native-like, respectively, and neither B1 nor B2 binding occurs. In the remaining three minima, the peptide is native-like and bound. The mode of binding is either B1 (minimum III), B2 (minimum IV) or both (minimum V).

It is worth noting that the interpretation of the TIC0 coordinates of the BPTI and GB1 systems is not necessarily the same although TIC0 is a strongly correlated with folding in both cases. In the GB1 system, TIC0 is correlated not only with folding but also with double binding [Figs. 3(c) and 3(d)]. By contrast, in the BPTI system, the correlation between TIC0 and the binding coordinate is weak [Fig. 2(c)].

To sum up, the results of this section show that TICA provides useful coordinates for describing the free energy of the peptide in the different systems. Using a few slow TICA coordinates, the main free-energy minima can be identified.

B. Dynamics

TICA provides a first approximation of the slow modes. For a more detailed investigation of the dynamics of the peptide in our simulations with crowders, MSMs are constructed

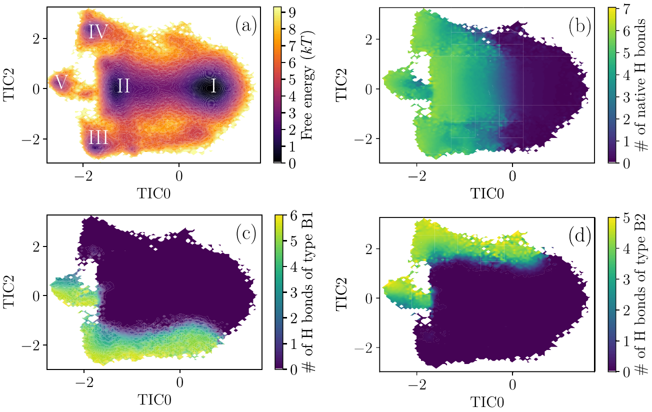


FIG. 3. Characterization of the GB1m3 peptide in the presence of GB1 crowders, using the slowest and third-slowest TICA coordinates, TIC0 and TIC2. (a) Free energy. Major minima are labeled by Roman numerals. (b) The number of native H bonds present in the peptide, n_{hb} [(c) and (d)] the numbers of present H bonds associated with the peptide-crowder binding modes B1 and B2 (see the supplementary material, Fig. S2), respectively. The TICA lag time is set to $\tau_{cm} = 20 \times 10^3$ MC sweeps.

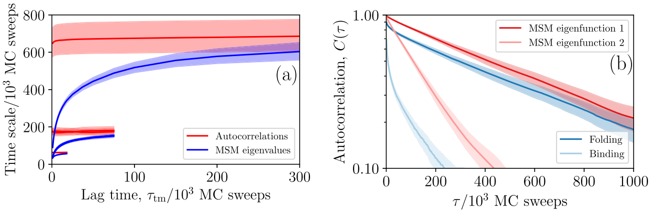


FIG. 4. Long-time dynamics of GB1m3 in the presence of BPTI crowders. Shaded areas indicate statistical 1σ errors. (a) Estimates of the four longest relaxation times, as obtained using MSM eigenvalues [Eq. (1); blue curves] and autocorrelation analysis (Sec. II E; red curves). The data are plotted against the lag time τ_{tm} of the MSM transition matrix. The second and third longest time scales are very similar. In building the MSMs, data were clustered in the space spanned by the four slowest TICA modes (using $\tau_{cm} = 10^3$ MC sweeps), into 800 clusters. (b) Autocorrelation functions, $C(\tau)$, for the two slowest MSM eigenfunctions ($\tau_{tm} = 25 \times 10^3$ MC sweeps), the folding variable n_{hb} [Fig. 2(b)], and the binding variable studied in Fig. 2(c).

as described in Sec. II D, for a range of lag times τ_{tm}. Relaxation times are estimated by two methods: (i) from MSM eigenvalues [Eq. (1)] and (ii) by fits to autocorrelation data for MSM eigenfunctions (Sec. II E). Illustrations of how the main MSM eigenfunctions are related to the TICA modes discussed above can be found in the supplementary material (Figs. S3–S6).

Figure 4(a) shows estimates of the four longest relaxation times in the system with BPTI crowders, as obtained by the above-mentioned methods. As expected, the eigenvalue-based estimates have systematic errors for small lag times τ_{tm}. To keep this error low, τ_{tm} has to be comparable to the time scale in question. The estimates based on autocorrelation analysis depend, by contrast, only very weakly on τ_{tm}. This behavior suggests that the true relaxation times can be estimated from the MSM eigenfunctions even if τ_{tm} is relatively small. Consistent with this, a further test shows that the shape of the slowest MSM eigenfunction depends only weakly on τ_{tm}. Here, pairwise overlaps (see Sec. II E) were computed between variants of this function obtained for different τ_{tm}. The overlap was $\geq$0.96 for all possible pairs of τ_{tm}.

Figure 4(b) compares the raw autocorrelation functions for the two slowest MSM eigenfunctions to those for the folding and binding coordinates studied in Figs. 2(b) and 2(c), respectively. One observation that can be made is that the autocorrelations of the folding and binding coordinates, not unexpectedly, show clear deviations from single-exponential behavior at small τ. The MSM eigenfunctions are, as intended by construction, much closer to single exponential, which facilitates the extraction of relaxation times.

Another observation from Fig. 4(b) is that, except at small τ, the autocorrelations of the first MSM eigenfunction and the folding coordinate decay at very similar rates. A close relationship between these two functions is indeed suggested from a comparison of Figs. 2(b) and S4(a) (see the supplementary material). This conclusion is further strengthened by their overlap (about 0.88). The autocorrelation function for the second MSM eigenfunction somewhat resembles that for the binding coordinate [Fig. 4(b)], but the overlap is not very large (about 0.36); the binding coordinate overlaps significantly with other MSM eigenfunctions as well. Thus, while the second eigenfunction probably is related to binding, that relationship is not fully captured by the binding coordinate.

Figure 5 shows data from our simulations with GB1 crowders. The statistical uncertainties are larger for this system. The main reason for this is that transitions to and from free-energy minimum V [Fig. 3(a)], where the peptide simultaneously binds two crowder molecules, occur only rarely in the simulations. Nevertheless, after increasing the number of runs from 16 for the BPTI system to 62, our total data set contains about 30 independent visits to this minimum, and some clear trends can be seen. The estimated relaxation times follow the same pattern as with BPTI crowders; the estimates based on MSM eigenvalues converge only slowly with increasing τ_{tm}, whereas those based on autocorrelation analysis are essentially constant down to small τ_{tm} [Fig. 5(a)]. However, in the GB1

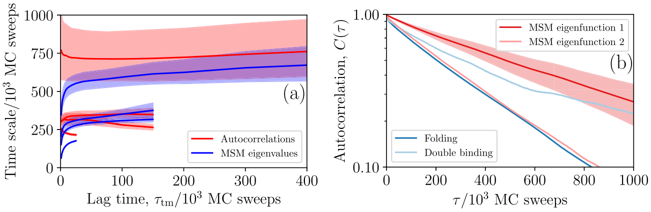


FIG. 5. Long-time dynamics of GB1m3 in the presence of GB1 crowders. Shaded areas indicate statistical 1σ errors. (a) Estimates of the four longest relaxation times, as obtained using MSM eigenvalues [Eq. (1); blue curves] and autocorrelation analysis (Sec. II E; red curves). The data are plotted against the lag time τ_{tm} of the MSM transition matrix. In building the MSMs, data were clustered in the space spanned by the three slowest TICA modes (using $\tau_{cm} = 20 \times 10^3$ MC sweeps), into 1574 clusters. (b) Autocorrelation functions, $C(\tau)$, for the two slowest MSM eigenfunctions ($\tau_{tm} = 25 \times 10^3$ MC sweeps), the folding variable n_{hb} [Fig. 3(b)], and the binding variable χ_b (see text). For clarity, statistical errors are shown only for one of the four functions. The statistical uncertainties are somewhat larger for the binding variable χ_b than they are for the other three functions.

system, the first MSM eigenfunction is more closely linked to binding than to folding. To show this, a binary function sensitive to simultaneous binding of the peptide to two crowder molecules is calculated. Specifically, this function is defined as $\chi_b = \chi_1\chi_2$, where χ_i is unity if at least three of the H bonds associated with binding mode i (see the supplementary material, Fig. S2) are present, and $\chi_i = 0$ otherwise. Figure 5(b) shows autocorrelation data for the two slowest MSM eigenfunctions, the folding coordinate (n_{hb}), and the function χ_b. The n_{hb} and χ_b functions are natural candidates for the slowest modes since they are both highly correlated with TIC0. It turns out that the autocorrelation function of χ_b decays slower than that of n_{hb}, and at a rate comparable to that for the first MSM eigenfunction [Fig. 5(b)]. Consistent with this, the first MSM eigenfunction has a larger overlap with the binding function χ_b (about 0.76) than it has with the folding coordinate (about 0.44).

Finally, we compute and compare the folding and unfolding rates of the peptide, k_f and k_u, in the three simulated environments. To this end, we determine the native-state probability, P_n (with the peptide being defined as native if $n_{hb} \geq 3$), and the apparent folding/unfolding rate, $k = k_f + k_u$. The rate k is obtained by a fit to autocorrelation data for the folding coordinate n_{hb} (Fig. 6). Knowing k and P_n and assuming a simple folded/unfolded two-state behavior, k_f and k_u can be computed ($k_f = kP_n$, $k_u = k - k_f$). Our data for P_n, k, k_f, and k_u are summarized in Table I. The BPTI crowders cause a considerable stabilization of the peptide (increased P_n) and a marked decrease in k. The decrease in k can be attributed to a lower k_u; no significant change in k_f is observed. With GB1 crowders, a similar pattern is observed although the stabilization of the peptide is much weaker in this case. Again, a markedly reduced k_u is observed, whereas the change in k_f is smaller. Therefore, in both the BPTI and GB1 simulations, the peptide seems to interact more efficiently with the crowders when folded than when unfolded. At the same time, the peptide-crowder interaction is different in character in the BPTI and GB1 cases (see above). Note therefore that the folding of the peptide to its native state entails the formation of both β-sheet structure and a hydrophobic side-chain cluster, which may enhance the interaction with GB1 and BPTI, respectively.

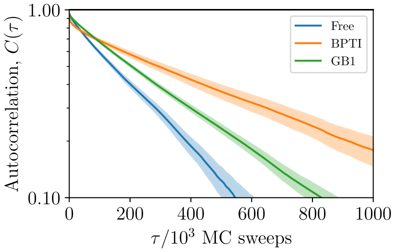


FIG. 6. Autocorrelation function, $C(\tau)$, for the folding variable n_{hb} (the number of native H bonds present in the peptide), as obtained without crowders, with BPTI crowders, and with GB1 crowders. Table I shows apparent folding rates k obtained by exponential fits to the data. Shaded areas indicate statistical 1σ errors.

TABLE I. Folding and unfolding rates of the GB1m3 peptide, k_f and k_u, in our three simulated systems. The rates are computed from the apparent rate constant $k = k_f + k_u$ and the native-state probability, P_n. The peptide is taken as native if at least three native H bonds are present, and k is obtained by fits to the data in Fig. 6. Rates are in units of $(10^6$ MC sweeps$)^{-1}$.

System	P_n	k	k_f	k_u
No crowders	0.30 ± 0.01	3.8 ± 0.3	1.1 ± 0.1	2.7 ± 0.2
BPTI crowders	0.72 ± 0.02	1.5 ± 0.2	1.1 ± 0.1	0.4 ± 0.1
GB1 crowders	0.33 ± 0.01	2.8 ± 0.1	0.9 ± 0.1	1.9 ± 0.1

IV. DISCUSSION AND SUMMARY

In this article, we have analyzed the interplay between peptide folding and peptide-crowder interactions in MC simulations of the GB1m3 peptide with protein crowders, using TICA and MSM techniques. A common major advantage of these methods is that they can be used to search for key coordinates of complex systems in an unsupervised manner. We used the simpler TICA method to explore the free-energy landscape of the peptide. Using a few slow TICA coordinates, it was possible to identify the major free-energy minima of the peptide in the presence of the crowders.

In order to quantitatively analyze the dynamics of the peptide in the simulations, we built MSMs. MSMs offer a convenient method for estimating relaxation times, from the eigenvalues via Eq. (1). However, this method is subject to well-known systematic uncertainties. In particular, it assumes effectively Markovian dynamics, which, at a given level of coarse graining, need not hold for small lag times τ_{tm}. Unfortunately, in our systems, τ_{tm} had to be comparable to the relaxation time in question to keep the systematic error low. Instead, we therefore estimated relaxation times by a procedure based on fits to autocorrelation data for the MSM eigenfunctions. The estimates obtained this way show essentially no τ_{tm}-dependence. This robustness suggests that the calculated MSM eigenfunctions maintain significant overlaps with the respective true eigenfunctions down to the smallest τ_{tm} values used.

It is, of course, also possible to estimate relaxation times from autocorrelation data for other functions than the MSM eigenfunctions. However, the autocorrelation of a general function is a multi-exponential whose parameters may be statistically challenging to determine. The autocorrelation of an MSM eigenfunction should, by contrast, be close to single-exponential over a range of τ, if this eigenfunction approximates the true eigenfunction sufficiently well (at low and high τ, deviations will occur since the approximation is not perfect). The autocorrelations of our MSM eigenfunctions showed this behavior, and relaxation times could therefore be estimated by single-exponential fits in an intermediate range of τ (where $0.2 < C(\tau) < 0.8$). If general functions rather than the MSM eigenfunctions had been used, our possibilities to estimate relaxation times would have been much more limited.

Our simulations further suggest that the GB1m3 peptide interacts more efficiently with both BPTI and GB1 when folded than when unfolded. The addition of either of the crowders led to a reduced unfolding rate k_u, while the change in the folding rate k_f was smaller, especially with BPTI crowders.

SUPPLEMENTARY MATERIAL

See supplementary material for illustrations of (i) the free energy of GB1m3 with GB1 crowders as a function of the TIC0 and TIC1 coordinates (Fig. S1), (ii) the preferred GB1m3-GB1 binding modes (Fig. S2), and (iii) the character of the leading MSM eigenfunctions in the different systems (Figs. S3–S6).

ACKNOWLEDGMENTS

This work was in part supported by the Swedish Research Council (Grant No. 621-2014-4522) and the Swedish strategic research program eSSENCE. The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at LUNARC, Lund University, Sweden, and Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany.

[1] I. Guzman, H. Gelman, J. Tai, and M. Gruebele, J. Mol. Biol. **426**, 11 (2014).
[2] W. B. Monteith, R. D. Cohen, A. E. Smith, E. Guzman-Cisneros, and G. J. Pielak, Proc. Natl. Acad. Sci. U. S. A. **112**, 1739 (2015).
[3] J. Danielsson, X. Mu, L. Lang, H. Wang, A. Binolfi, F.-X. Theillet, B. Bekei, D. T. Logan, P. Selenko, H. Wennerström, and M. Oliveberg, Proc. Natl. Acad. Sci. U. S. A. **112**, 12402 (2015).
[4] S. R. McGuffee and A. H. Elcock, PLoS Comput. Biol. **6**, e1000694 (2010).
[5] M. Feig and Y. Sugita, J. Phys. Chem. B **116**, 599 (2012).
[6] A. V. Predeus, S. Gul, S. M. Gopal, and M. Feig, J. Phys. Chem. B **116**, 8610 (2012).
[7] B. Macdonald, S. McCarley, S. Noeen, and A. E. van Giessen, J. Phys. Chem. B **119**, 2956 (2015).
[8] A. Bille, B. Linse, S. Mohanty, and A. Irbäck, J. Chem. Phys. **143**, 175102 (2015).
[9] I. Yu, T. Mori, T. Ando, R. Harada, J. Jung, Y. Sugita, and M. Feig, eLife **5**, 18457 (2016).
[10] M. Feig, I. Yu, P.-h. Wang, G. Nawrocki, and Y. Sugita, J. Phys. Chem. B **121**, 8009 (2017).
[11] S. Qin and H.-X. Zhou, Curr. Opin. Struct. Biol. **43**, 28 (2017).
[12] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).
[13] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).
[14] N.-V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[15] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, J. Chem. Phys. **131**, 124101 (2009).
[16] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**, 174105 (2011).
[17] J. D. Chodera and F. Noé, Curr. Opin. Struct. Biol. **25**, 135 (2014).
[18] F. Noé and C. Clementi, Curr. Opin. Struct. Biol. **43**, 141 (2017).
[19] L. Molgedey and H. G. Schuster, Phys. Rev. Lett. **72**, 3634 (1994).
[20] Y. Naritomi and S. Fuchigami, J. Chem. Phys. **139**, 215102 (2013).
[21] C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).
[22] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).
[23] R. M. Fesinmeyer, F. M. Hudson, and N. H. Andersen, J. Am. Chem. Soc. **126**, 7238 (2004).
[24] E. Moses and H.-J. Hinz, J. Mol. Biol. **170**, 765 (1983).
[25] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, Science **253**, 657 (1991).
[26] A. Bille, S. Mohanty, and A. Irbäck, J. Chem. Phys. **144**, 175105 (2016).
[27] A. Vendeville, D. Larivière, and E. Fourmentin, FEMS Microbiol. Rev. **35**, 395 (2011).
[28] A. Irbäck, S. Mitternacht, and S. Mohanty, BMC Biophys. **2**, 2 (2009).
[29] S. Mitternacht, S. Luccioli, A. Torcini, A. Imparato, and A. Irbäck, Biophys. J. **96**, 429 (2009).
[30] S. Æ. Jónsson, S. Mohanty, and A. Irbäck, Proteins **80**, 2169 (2012).
[31] S. Mohanty, J. H. Meinke, and O. Zimmermann, Proteins **81**, 1446 (2013).
[32] A. Bille, S. Æ. Jónsson, M. Akke, and A. Irbäck, J. Phys. Chem. B **117**, 9194 (2013).
[33] S. Æ. Jónsson, S. Mitternacht, and A. Irbäck, Biophys. J. **104**, 2725 (2013).
[34] J. Petrlova, A. Bhattacherjee, W. Boomsma, S. Wallin, J. O. Lagerstedt, and A. Irbäck, Protein Sci. **23**, 1559 (2014).
[35] G. Tiana, L. Sutto, and R. A. Broglia, Phys. A **380**, 241 (2007).
[36] G. Favrin, A. Irbäck, and F. Sjunnesson, J. Chem. Phys. **114**, 8154 (2001).
[37] A. Irbäck and S. Mohanty, J. Comput. Chem. **27**, 1548 (2006).
[38] S. Lloyd, IEEE Trans. Inf. Theory **28**, 129 (1982).
[39] S. Kube and M. Weber, J. Chem. Phys. **126**, 024103 (2007).
[40] N. Djurdjevac, M. Sarich, and C. Schütte, Multiscale Model. Simul. **10**, 61 (2012).
[41] J.-H. Prinz, J. D. Chodera, and F. Noé, Phys. Rev. X **4**, 011020 (2014).
[42] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, J. Chem. Theory Comput. **11**, 5525 (2015).
[43] M. Seeber, A. Felline, F. Raimondi, S. Muff, R. Friedman, F. Rao, A. Caflisch, and F. Fanelli, J. Comput. Chem. **32**, 1183 (2010).
[44] X. Biarnés, F. Pietrucci, F. Marinelli, and A. Laio, Comput. Phys. Commun. **183**, 203 (2012).
[45] M. P. Harrigan, M. M. Sultan, C. X. Hernández, B. E. Husic, P. Eastman, C. R. Schwantes, K. A. Beauchamp, R. T. McGibbon, and V. S. Pande, Biophys. J. **112**, 10 (2017).

Supplementary material for:
Markov modeling of peptide folding in the presence of protein crowders

Daniel Nilsson,[1, a] Sandipan Mohanty,[2, b] and Anders Irbäck[1, c]

[1] *Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden*

[2] *Institute for Advanced Simulation, Jülich Supercomputing Centre, Forschungszentrum Jülich, D-52425 Jülich, Germany*

[a] Electronic mail: daniel.nilsson@thep.lu.se

[b] Electronic mail: s.mohanty@fz-juelich.de

[c] Electronic mail: anders@thep.lu.se

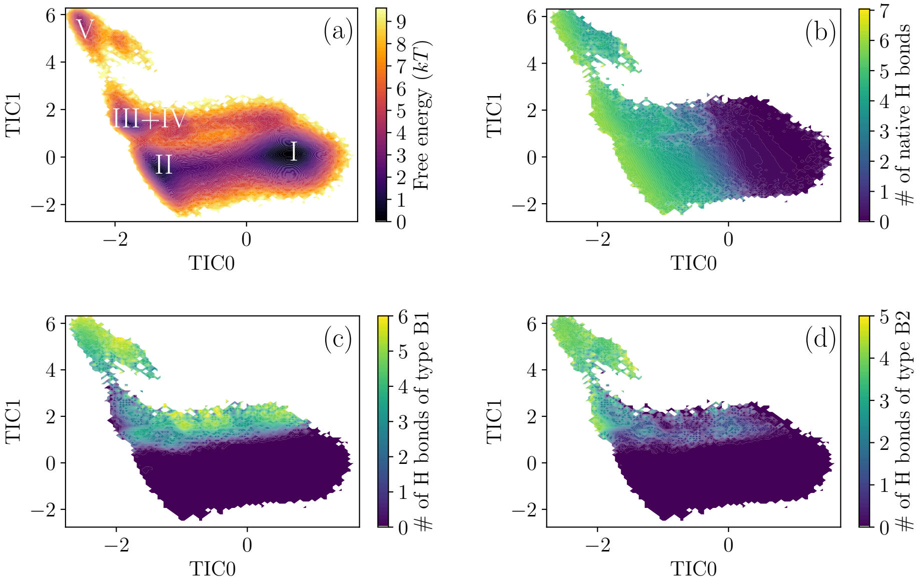


FIG. S1. Characterization of the GB1m3 peptide in the presence of GB1 crowders, using the two slowest TICA coordinates, TIC0 and TIC1. (a) Free energy. The minima identified in Figure 3 are indicated. Minima III and IV cannot be distinguished, while being well separated in the TIC0,TIC2 coordinates used in the main text (Figure3). (b) The number of native H bonds present in the peptide, n_{hb} (c) The number of present H bonds associated with the peptide-crowder binding mode B1 (Figure S2). (d) The number of present H bonds associated with the peptide-crowder binding mode B2 (Figure S2).

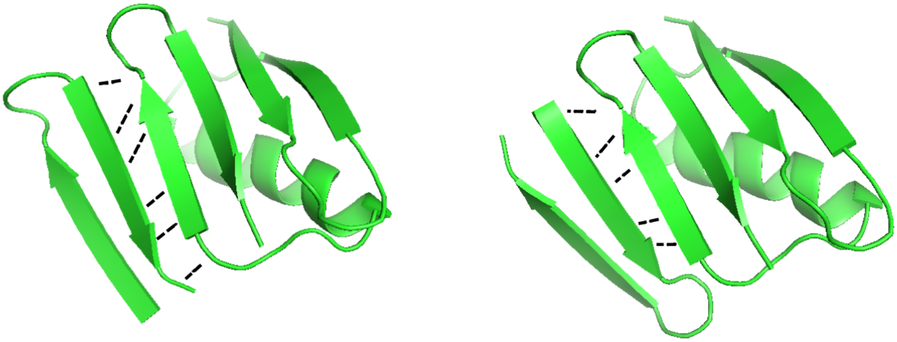


FIG. S2. Preferred binding modes between the peptide GB1m3 and the crowder protein GB1, B1 (left) and B2 (right). B1 involves residues 10–16 on the peptide, whereas B2 involves residues 1–5. In both cases, the peptide binds to residues 41–47 on the crowder protein. Dashed black lines indicate intermolecular H bonds. Structures drawn with PyMOL.[1]

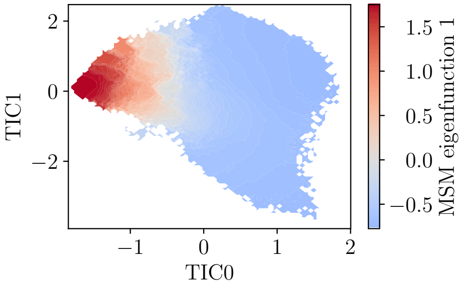


FIG. S3. The dependence of the slowest MSM eigenfunction on the two slowest TICA coordinates, TIC0 and TIC1, for the isolated GB1m3 peptide. The TICA and MSM lag times are set to $\tau_{\rm cm} = \tau_{\rm tm} = 10^3$ MC sweeps. In building the MSM, data were clustered in the space spanned by the three slowest TICA modes, into 547 clusters.

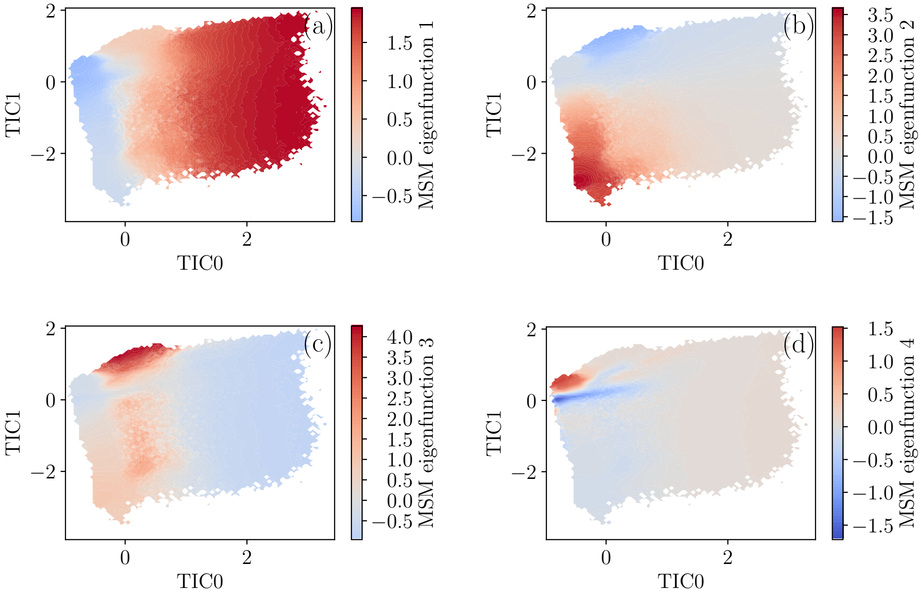


FIG. S4. The four slowest MSM eigenfunctions for GB1m3 in the presence of BPTI crowders, viewed as functions of the two slowest TICA coordinates, TIC0 and TIC1. The TICA and MSM lag times are set to $\tau_{\mathrm{cm}} = 10^3$ MC sweeps and $\tau_{\mathrm{tm}} = 25 \times 10^3$ MC sweeps, respectively. In building the MSM, data were clustered in the space spanned by the four slowest TICA modes, into 800 clusters.

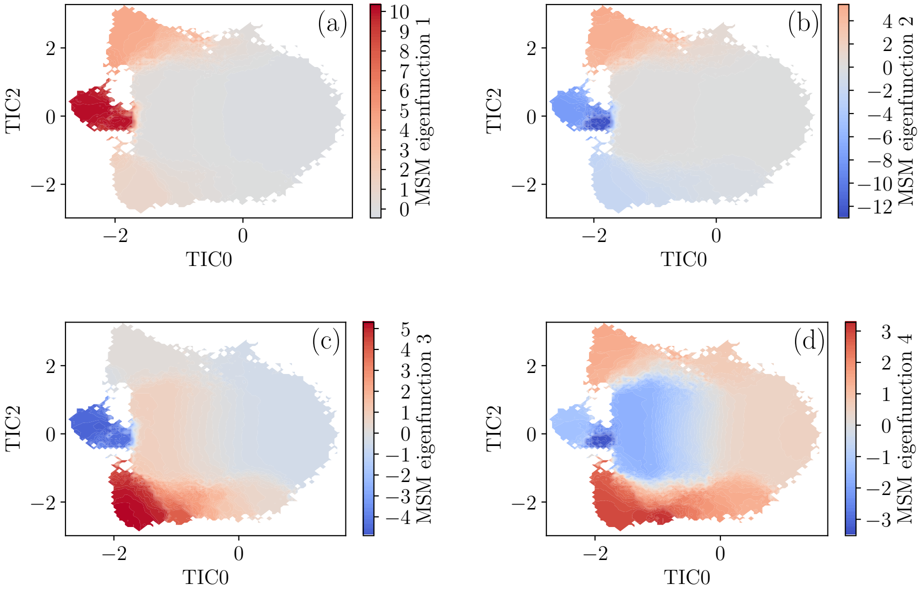


FIG. S5. The four slowest MSM eigenfunctions for GB1m3 in the presence of GB1 crowders, viewed as functions of the slowest and third-slowest TICA coordinates, TIC0 and TIC2. The TICA and MSM lag times are set to $\tau_{\mathrm{cm}} = 20 \times 10^3$ MC sweeps and $\tau_{\mathrm{tm}} = 25 \times 10^3$ MC sweeps, respectively. In building the MSM, data were clustered in the space spanned by the three slowest TICA modes, into 1574 clusters.

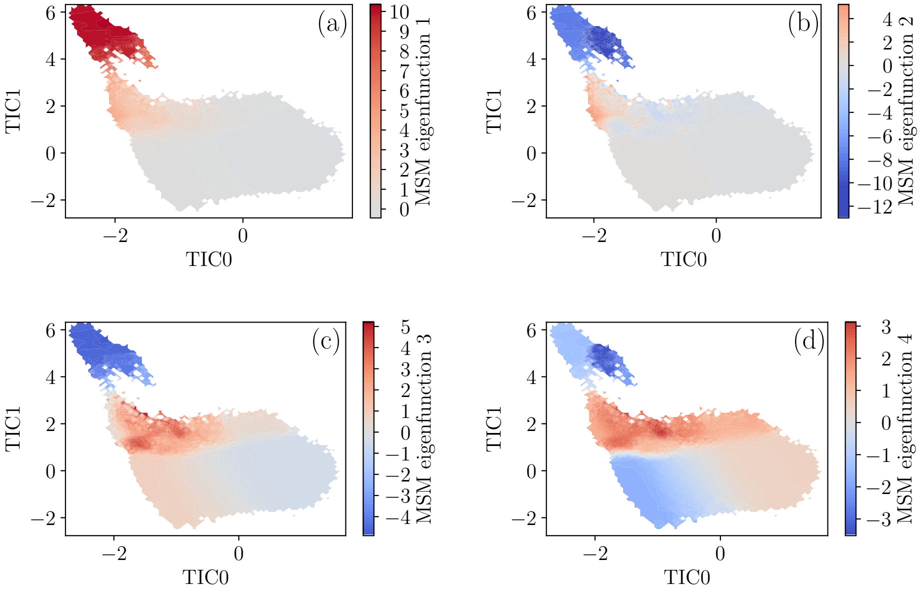


FIG. S6. The four slowest MSM eigenfunctions for GB1m3 in the presence of GB1 crowders, viewed as functions of the two slowest and TICA coordinates, TIC0 and TIC1. The TICA and MSM lag times are set to $\tau_{\rm cm} = 20 \times 10^3$ MC sweeps and $\tau_{\rm tm} = 25 \times 10^3$ MC sweeps, respectively. In building the MSM, data were clustered in the space spanned by the three slowest TICA modes, into 1574 clusters.

REFERENCES

[1] W. L. DeLano, "The PyMOL molecular graphics system," (2002), San Carlos, CA: DeLano Scientific.

Paper II

Finite-size scaling analysis of protein droplet formation

Daniel Nilsson and Anders Irbäck
Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University,
Sölvegatan 14A, SE-223 62 Lund, Sweden

(Received 15 October 2019; accepted 17 January 2020; published 18 February 2020)

The formation of biomolecular condensates inside cells often involve intrinsically disordered proteins (IDPs), and several of these IDPs are also capable of forming dropletlike dense assemblies on their own, through liquid-liquid phase separation. When modeling thermodynamic phase changes, it is well known that finite-size scaling analysis can be a valuable tool. However, to our knowledge, this approach has not been applied before to the computationally challenging problem of modeling sequence-dependent biomolecular phase separation. Here we implement finite-size scaling methods to investigate the phase behavior of two 10-bead sequences in a continuous hydrophobic-polar protein model. Combined with reversible explicit-chain Monte Carlo simulations of these sequences, finite-size scaling analysis turns out to be both feasible and rewarding, despite relying on theoretical results for asymptotically large systems. While both sequences form dense clusters at low temperature, this analysis shows that only one of them undergoes liquid-liquid phase separation. Furthermore, the transition temperature at which droplet formation sets in is observed to converge slowly with system size, so that even for our largest systems the transition is shifted by about 8%. Using finite-size scaling analysis, this shift can be estimated and corrected for.

DOI: 10.1103/PhysRevE.101.022413

I. INTRODUCTION

Advances over the past decade have shown that, in addition to classical membrane-bound organelles, various membraneless liquidlike droplets of proteins and nucleic acids can be found within living cells [1,2]. The droplets form through a liquid-liquid phase separation (LLPS) process, also called coacervation, in which intrinsically disordered proteins (IDPs) often play a key role. Furthermore, it has been demonstrated *in vitro* that several of these IDPs are able to phase separate on their own [3–5], depending on the solution conditions. Phase-separating IDPs can be rich in charged residues [3] but can also be dominated by polar and aromatic residues [5].

To rationalize the phase behavior of IDPs and its dependence on solution conditions, a variety of theoretical and computational methods have been employed. A widely used method is Flory-Huggins mean-field theory [6,7] and its extension to polyelectrolytes by Voorn and Overbeek [8]. However, this approach is sensitive only to the overall composition of amino acids but not to their ordering along the chains. One way to overcome this shortcoming without resorting to explicit-chain simulation is offered by the random-phase approximation [9], which has been applied to model the phase-separating ability of IDPs with different charge patterns [10].

By turning to molecular simulation with explicit chains, key approximations made in the above approaches can be removed. In addition, structural properties become readily accessible. Therefore, despite being computationally costly, recent years have seen a growing number of explicit-chain simulation studies of biomolecular LLPS [11–18]. In particular, there have been simulations based on coarse-grained lattice or continuous representations to elucidate sequence determinants of phase-separating IDPs [11–14].

Another approach, recently applied for the first time to biomolecular LLPS [19,20], is to rewrite the original polymer problem as a statistical field-theory problem that can be investigated by field-theory simulation. This approach has the potential to open for studies of system sizes that are inaccessible with explicit-chain simulation.

Yet, regardless of whether explicit-chain or field-theory methods are used, the simulated systems are finite and as such there is a need to understand how measured properties depend on system size. Fortunately, tools for this purpose are available in the form of finite-size scaling theory for droplet formation by phase separation [21–24]. These tools have previously been applied to analyze droplet formation in simpler systems such as the lattice gas and the Lennard-Jones fluid [24–26], but we are not aware of any prior study of biomolecular LLPS using these ideas.

In this paper, we implement finite-size scaling methods to assess the phase behavior of two short model proteins, which provide an instructive testbed for the analysis methods. While several previous computational studies of IDP phase separation focused on the role of charge-charge interactions, we here consider a hydrophobic-polar (H-P) protein model. One of the sequences we study, called A, is alternating (HPHPHPHPHP), whereas the other, called B, has a block structure

(HHHHHPPPPP). Using Monte Carlo (MC) methods, we perform canonical simulations of these sequences for a range of system sizes, with up to 640 chains. Both sequences form dense multichain assemblies surrounded by a dilute background at low temperatures, while only small clusters are present at high temperatures. However, the sequences differ in phase behavior. We show that their phase behavior can be assessed in a systematic fashion by finite-size scaling analysis of the simulation data. This analysis demonstrates that one of the sequences, A, undergoes LLPS, whereas the other, B, does not.

II. METHODS

A. Biophysical model

We study systems consisting of N copies of a polypeptide enclosed in a periodic cubic box with volume V. The polypeptide is represented as a string of n hydrophobic (H) or polar (P) beads. The length of the bond between two consecutive beads, b, is kept fixed, while the polar and azimuthal bond angles are both free to vary. In the absence of interactions, the bonds have a spherically uniform distribution.

The beads interact through a pairwise additive potential, $E = \sum_{i<j} E_{ij}$, where the sum runs over all intra- and intermolecular pairs of beads in the system. All beads have a diameter of $r_{\text{ev}} = 0.75b$. When two beads i and j are at a distance $r_{ij} < r_{\text{ev}}$ from each other, the pair potential E_{ij} becomes infinite. Additionally, each HH pair interacts through a soft attractive potential with interaction range $r_{\text{hp}} = 2b$. If $r_{\text{ev}} < r_{ij} < r_{\text{hp}}$, then the interaction energy is set to $-\epsilon$ (with $\epsilon > 0$). Thus, the pair potential can be summarized as

$$E_{ij} = \begin{cases} \infty, & \text{if } r_{ij} < r_{\text{ev}} \\ u_{ij}, & \text{if } r_{\text{ev}} < r_{ij} < r_{\text{hp}}, \\ 0, & \text{if } r_{ij} > r_{\text{hp}} \end{cases} \tag{1}$$

where $u_{ij} = -\epsilon$ when beads i and j are both hydrophobic and $u_{ij} = 0$ otherwise.

Throughout this article, lengths and energies are given in units of b and ϵ, respectively.

B. MC simulations

We investigate the thermodynamics of droplet formation in this model by using MC methods to generate samples from the canonical (NVT) ensemble. Of particular interest is the behavior at the onset of droplet formation. Therefore, given N and V, the temperature T is chosen close to the maximum of the heat capacity, by an iterative procedure. Simulations at one or several additional temperatures are performed when needed to ensure an accurate description of the heat capacity throughout the transition region. The temperature dependence of the heat capacity is computed by reweighting techniques [27], using data from all simulated temperatures.

The efficiency of MC simulations depends strongly on the choice of move set. We use a set of six elementary moves. Two of the moves update the internal structure of individual chains. The first of these is a single-bead move, which turns a randomly selected nonend bead about the axis through its two nearest neighbors. The second one is a pivot-type rotation, where the rotation axis goes through a randomly selected nonend bead in a random direction. Beads on one side of the selected one are turned as a rigid body about this axis.

The other four moves are rigid-body translations and rotations of either a single chain or a cluster of chains. In the cluster moves, the clusters are constructed probabilistically using a Swendsen-Wang-type algorithm [28,29]. The construction is recursive and begins by picking a random first cluster member, i. Then each chain j that has an interaction energy $E_{ij} < 0$ with chain i is added to the cluster with probability $p_{ij} = 1 - \mathrm{e}^{\beta E_{ij}}$, where $\beta = 1/k_B T$ is inverse temperature (k_B is Boltzmann's constant). This step is iterated until the list of potential further cluster members is empty. Finally, the resulting cluster is subject to a trial rigid-body move. The form of the statistical weight p_{ij} is such that no Metropolis accept-reject criterion is needed; any sterically allowed move is accepted.

For each choice of N, V, T, and HP sequence, a set of one to eight trajectories is generated, each comprising 10^7 MC sweeps, where one MC sweep corresponds to nN elementary updates. Multiple runs are used for the largest systems to ensure statistical significance. Statistical uncertainties are computed using a jackknife procedure [30].

C. Finite-size scaling theory

Droplet formation by phase separation in finite systems is a topic that has been extensively studied over the years [21–24]. This body of research provides a general framework for finite-size scaling analysis, which has been tested on systems such as the lattice gas and the Lennard-Jones fluid [24–26]. This section outlines some key results that will be used in Sec. III.

We consider a d-dimensional system of N particles in a volume V at temperatures T below an assumed critical temperature T_c. A schematic phase diagram can be found in Fig. 1. Under grand-canonical conditions, for a given $T < T_c$ and large system size, the system can be in one of two bulk phases with respectively low (ρ_L) and high (ρ_H) density, depending on the chemical potential. At some value of the chemical potential, a first-order transition occurs between these phases. Under canonical conditions, for $T < T_c$ and densities ρ such that $\rho_L(T) < \rho < \rho_H(T)$, the system is in a mixed two-phase regime, bounded by the binodal curve, $T_b(\rho)$ (Fig. 1).

Consider now a finite but large system under canonical conditions, for a given $T < T_c$ and ρ just above $\rho_L(T)$ (Fig. 1). At some $\rho_L^{(N)}(T) > \rho_L(T)$, the system transitions from a supersaturated dilute state to a mixed two-phase state. It has been shown that this mixed state consists of a single large droplet of the high-density phase in a low-density background, and that the linear dimension R of the droplet scales as $R \sim N^{1/(d+1)}$ with N [21–23]. This result can be rigorously proven for the two-dimensional lattice gas [31]. In brief, the size of the critical droplet can be viewed as the result of two competing mechanisms for handling a particle excess, δN. One is that the particle excess is absorbed as a density fluctuation in the low-density phase, the free-energy cost of which scales as $(\delta N)^2/N$. The other mechanism is that a finite fraction of the particle excess forms a dense droplet, the free-energy cost of which scales as the surface area of the droplet, that is, $(\delta N)^{(d-1)/d}$. Assuming that droplet formation sets in when these two costs become comparable, one finds that the

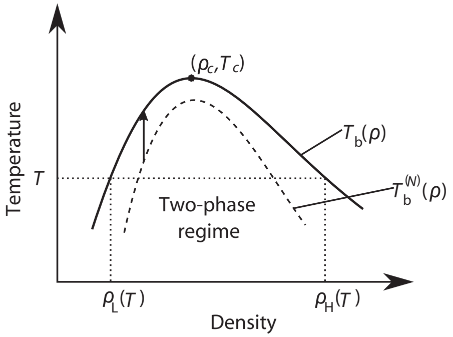


FIG. 1. Schematic temperature-density phase diagram of a system that undergoes phase separation below an upper critical temperature, T_c, into two phases with respectively low (ρ_L) and high (ρ_H) densities. In other systems, phase separation may occur above a lower critical temperature. Below the so-called binodal curve, $T_b(\rho)$, the low- and high-density phases coexist. At the left branch of the curve, the system transitions between the low-density phase and a mixed two-phase regime. In finite systems, the transition temperature, $T_b^{(N)}(\rho)$, is shifted (dashed line). Finite-size scaling theory predicts $T_b^{(N)}(\rho)$ to converge toward $T_b(\rho)$ following the scaling relation in Eq. (2) (arrow).

linear size of the critical droplet scales as $R \sim (\delta N)^{1/d} \sim N^{1/(d+1)}$ [21–23].

Using this result, it follows that the finite-size shift of the transition density scales as $\rho_L^{(N)}(T) - \rho_L(T) \propto N^{-1/(d+1)}$. Correspondingly, with ρ rather than T fixed, the transition temperature has a finite-size shift, given by

$$T_b^{(N)}(\rho) - T_b(\rho) \propto N^{-1/(d+1)}. \qquad (2)$$

Note that this relation implies that the convergence of $T_b^{(N)}$ toward its value for infinite system size, T_b, is slow. For comparison, the finite-size shift of a regular temperature-driven first-order phase transition scales as N^{-1} [32].

In finite systems, the transition is not only shifted but also smeared. At fixed ρ, the smearing, or width, of the transition, w_T, may be estimated as the temperature interval over which $|\beta\Delta F| \lesssim 1$ [23,26], where ΔF is the free-energy difference between the states with and without a droplet. Since ΔF vanishes at $T_b^{(N)}$, a Taylor expansion yields $\beta\Delta F = -[\Delta E/k_BT^2]_{T=T_b^{(N)}}[T - T_b^{(N)}]$ to leading order. Here ΔE is the energy gap, which, assuming that particle interactions are negligible in the low-density phase, should scale as the droplet volume, that is,

$$\Delta E \sim N^{d/(d+1)}. \qquad (3)$$

It then follows that the smearing of the transition scales as

$$w_T \propto N^{-d/(d+1)}. \qquad (4)$$

When analyzing the droplet transition, a useful property is the specific heat, C_V/N, which exhibits a peak at the transition and can be computed from the energy fluctuations, using $C_V = (\langle E^2\rangle - \langle E\rangle^2)/k_BT^2$. The transition temperature, $T_b^{(N)}$, and smearing, w_T, may be defined as, the position and width of the specific heat peak, respectively. With increasing N, the width of the peak, w_T, decreases [Eq. (4)], whereas the height of the peak, $C_{V,\max}/N$, increases. With a two-state approximation, one has $C_{V,\max} \approx (\Delta E)^2/4k_BT^2$, where ΔE, as before, is the energy gap. Using this relation along with Eq. (3), one finds that

$$C_{V,\max}/N \sim N^{(d-1)/(d+1)}. \qquad (5)$$

A slightly different behavior, namely $C_{V,\max}/N \sim N^{d/(d+1)}$, has been suggested [26], based on the assumption that $C_{V,\max}/N$ scales inversely proportional to w_T. However, unlike at a regular temperature-driven first-order phase transition, in the case of droplet formation, the area under the specific heat peak vanishes in the large-N limit, since $\Delta E/N$ does so. Hence $C_{V,\max}/N$ should grow slower than $w_T^{-1} \sim N^{d/(d+1)}$ with N, as it does in Eq. (5).

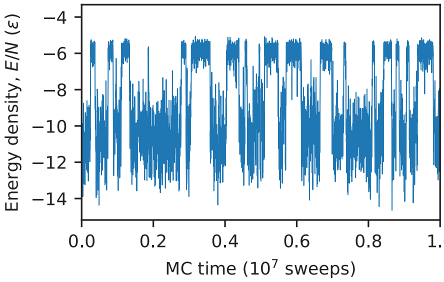


FIG. 2. MC evolution of the energy density E/N in a run with $N = 640$, $T \approx T_b^{(N)}$, and $\rho_b = 0.025b^{-3}$, for sequence A. Low and high energies correspond to states with and without a droplet, respectively. During the course of the run, droplet formation and dissolution occur several times.

III. RESULTS

Using the model and MC methods described in Sec. II, we conduct thermodynamic simulations of droplet formation by the two sequences A (HPHPHPHPHP) and B (HHHHH-PPPPP) for a range of system sizes, with up to $N = 640$ chains. The volume V is adjusted so as to have a given bead density $\rho_b = nN/V$. Most of the calculations are for a bead density of $\rho_b = 0.025b^{-3}$, where b is the link length of the chains. For comparison, some data for $\rho_b = 0.0125b^{-3}$ and $\rho_b = 0.0375b^{-3}$ are also included. The simulations focus on temperatures near the onset of droplet formation and were sufficiently fast for droplets to form and dissolve several times during the course of a run, even for the largest systems, as illustrated by Fig. 2.

A. Overall characterization

At high temperatures, the simulated systems are in a disordered state, with only small clusters present ($\lesssim$ 10 chains). As the temperature is reduced, markedly larger clusters, or droplets, appear. Their formation sets in abruptly, in a narrow temperature interval, where states both with and without droplets are observed. Figure 3 shows representative snapshots of configurations with droplets for both sequences, from

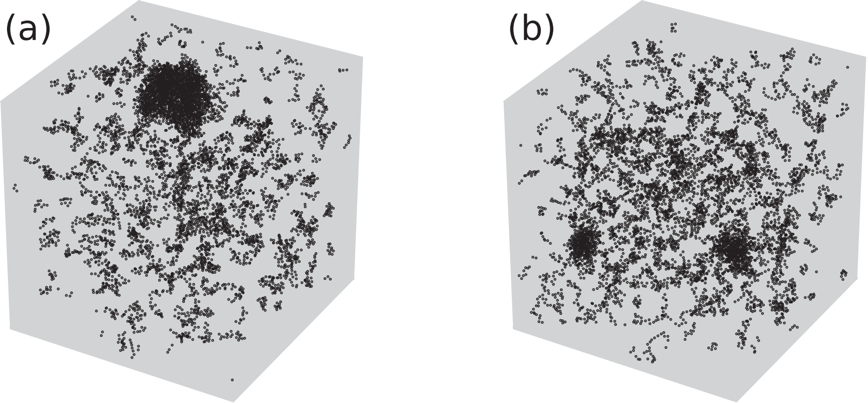


FIG. 3. Snapshots showing representative droplet-containing configurations from simulations near the temperature at which droplet formation sets in for $N = 640$ and $\rho_b = 0.025b^{-3}$. Each bead is shown as a dot. (a) Sequence A, for which a single large droplet is observed. (b) Sequence B, which typically forms a few smaller droplets.

simulations with 640 chains. For sequence A, a single large droplet can be seen, in a dilute background with only small clusters. For sequence B, more than one droplet is often present, and the droplets are smaller than those formed by sequence A. A single large droplet is what one expects to observe if droplet formation occurs through phase separation [21–23].

If phase separation occurs, then the onset of droplet formation is, furthermore, expected to be associated with a divergence in the specific heat (Sec. II C). Consistent with this, for sequence A, specific-heat data from simulations with 10–640 chains show a peak that steadily gets higher and narrower with increasing system size [Fig. 4(a)]. The corresponding data for sequence B follow the same trend for small systems [Fig. 4(b)]. However, for this sequence, at some system size (around 80 chains), the specific heat stops growing higher and becomes multimodal. This behavior reflects the fact that sequence B forms more than one droplet in the larger systems (Fig. 3) and implies that this sequence does not undergo LLPS.

B. Finite-size scaling analysis

The above results indicate that, unlike sequence B, sequence A may undergo LLPS. To determine whether this is the case, we next compare simulation data for several quantities with predictions from finite-size scaling theory (Sec. II C), focusing on sequence A.

At the onset of droplet formation, due to the coexistence of states with and without a droplet, the probability distribution of energy should be bimodal, as it is at a regular temperature-driven first-order phase transition. In the latter case, the energy gap between the two phases scales linearly with system size, corresponding to a nonzero specific latent heat. However, at the droplet transition, the energy gap ΔE should scale as the critical droplet volume or $\Delta E \sim N^{3/4}$ [Eq. (3)].

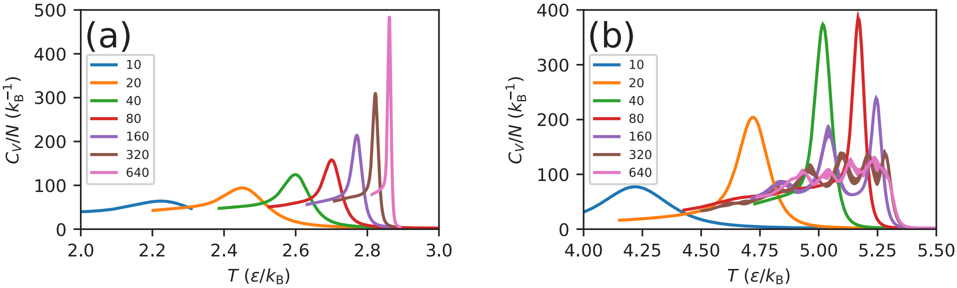


FIG. 4. Temperature dependence of the specific heat, C_V/N, from simulations with 10–640 chains for fixed $\rho_b = 0.025b^{-3}$. The curves are computed by reweighting methods [27] using data from canonical MC simulations at several temperatures. Shaded bands indicate statistical uncertainties but are in many cases too narrow to be visible. (a) For sequence A, the specific heat exhibits a single peak that steadily gets higher and narrower with increasing system size. (b) For sequence B, the same trend is observed but only for small systems. In the larger systems, sequence B forms more than one droplet, which leads to a multimodal specific heat.

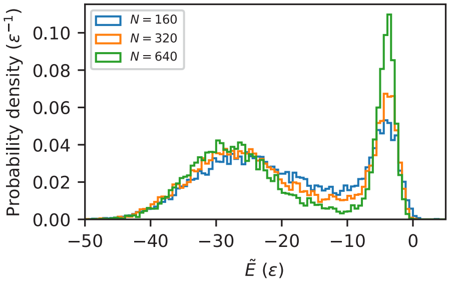


FIG. 5. Probability distribution of the shifted and rescaled energy $\tilde{E} = (E + a)/N^{3/4}$ (with $a = 5N\epsilon$) from simulations with 160, 320, and 640 chains for sequence A at $T \approx T_b^{(N)}$ and $\rho_b = 0.025b^{-3}$. Consistent with the predicted scaling relation $\Delta E \sim N^{3/4}$ [Eq. (3)], the gap between the two peaks in $\tilde{E}$ stays essentially constant, whereas the statistical suppression of intermediate energies gets stronger with increasing system size.

Figure 5 shows the probability distribution of the shifted and rescaled energy $\tilde{E} = (E + a)/N^{3/4}$, where a is a parameter independent of E, for $T \approx T_b^{(N)}$ and $\rho_b = 0.025b^{-3}$, for our three largest systems. With larger system size, the probability distribution of $\tilde{E}$ becomes increasingly bimodal in character, due to a stronger suppression of intermediate energies. By contrast, the gap between the two peaks in $\tilde{E}$ stays essentially unchanged, in perfect agreement with the predicted scaling of ΔE [Eq. (3)].

Assuming this scaling of ΔE with N [Eq. (3)], the maximum specific heat, $C_{V,\max}/N$, should scale as $N^{1/2}$ [Eq. (5)]. Figure 6(a) shows $C_{V,\max}/N$ data against N in a log-log plot, for three bead densities ρ_b. Not surprisingly, the data for small systems do not follow the predicted scaling relation for large N. However, the data for the four largest systems ($N = 80$–640) match well with the predicted form for all three bead densities.

Figure 6 also illustrates the finite-size smearing and shift of the transition, for the same three bead densities. The smearing w_T is expected to scale inversely proportional to the energy gap ΔE or $w_T \sim N^{-3/4}$ [Eq. (4)]. From the log-log plot in Fig. 6(b), it can be seen that the data for w_T indeed are consistent with the predicted scaling for large N.

The finite-size shift of the transition temperature, $T_b^{(N)} - T_b$, is predicted to scale as $N^{-1/4}$ [Eq. (2)]. Therefore, Fig. 6(c) shows the data for $T_b^{(N)}$ plotted against $N^{-1/4}$. As can be seen from this figure, fits of the form $T_b^{(N)} = T_b + cN^{-1/4}$, with T_b and c as parameters, indeed provide a good description of the large-N data ($80 \leqslant N \leqslant 640$). It is worth noting that the scaling of the shift as $N^{-1/4}$ or inversely proportional to the linear size of the critical droplet, implies a slow convergence

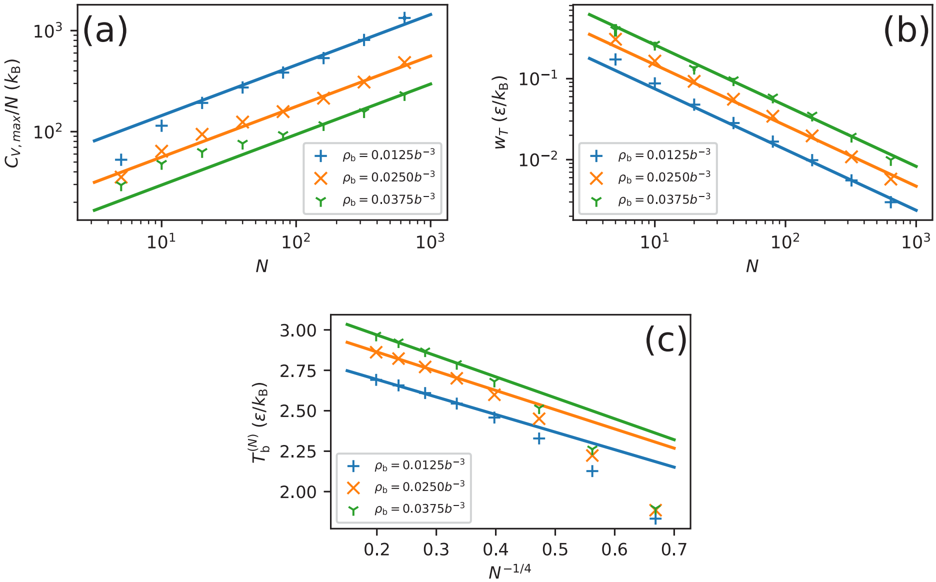


FIG. 6. Finite-size scaling analysis at three bead densities ρ_b ($0.0125b^{-3}$, $0.0250b^{-3}$, $0.0375b^{-3}$) for sequence A using data from simulations with 5–640 chains. Lines represent fits of predicted scaling expressions from Sec. II C to data for the four largest system sizes. (a) Log-log plot of the maximum specific heat, $C_{V,\max}/N$, against N. The lines are fits of the form $C_{V,\max}/N \sim N^{1/2}$ [Eq. (5)]. (b) Log-log plot of the finite-size smearing of the transition, w_T, against N, where w_T is computed as the length of the temperature interval over which $C_V > 0.8C_{V,\max}$. The lines are fits of the form $w_T \sim N^{-3/4}$ [Eq. (4)]. (c) The transition temperature $T_b^{(N)}$ plotted as a function of $N^{-1/4}$. The lines are fits of the form $T_b^{(N)} = T_b + cN^{-1/4}$ [Eq. (2)], with c and the transition temperature for infinite system size, T_b, as fit parameters. The fitted values of T_b are $T_bk_B/\epsilon = 2.92$, 3.10, and 3.23 for $\rho_b b^3 = 0.0125$, 0.025, and 0.0375, respectively.

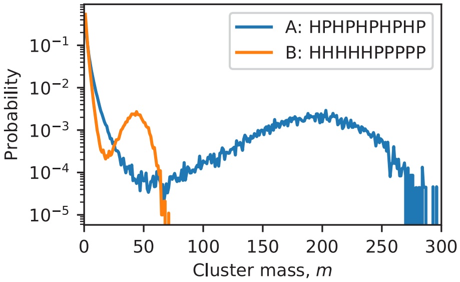


FIG. 7. Mass fraction of clusters with m chains, $P(m)$, as obtained using $N = 640$, $\rho_b = 0.025b^{-3}$ and a temperature near the onset of droplet formation. Alternatively expressed, $P(m)$ is the probability that a randomly selected chain belongs to a cluster with m chains. Sequence A forms droplets containing roughly 200 of the 640 chains, whereas intermediate-mass clusters are statistically suppressed.

of $T_b^{(N)}$ toward T_b with increasing N. In fact, for our largest systems with 640 chains, $T_b^{(N)}$ is still about 8% smaller than the fitted value of T_b.

To summarize the above analysis, for all properties studied, we find that the simulation data for sequence A are consistent with the theoretical predictions, which provides strong evidence that this sequence indeed undergoes LLPS.

C. Droplet size and structure

The specific heat data discussed in Secs. III A and III B show that the sequences A and B, despite sharing the same length and composition, have different phase behaviors. To understand this difference, we next examine some basic structural properties of the droplets formed by these sequences. Throughout this section, we focus on data obtained using $N = 640$, $\rho_b = 0.025b^{-3}$ and a temperature near the onset of droplet formation.

One important characteristic is the mass of the droplets or the number of chains that they contain. It was already noted that sequence A forms more massive droplets than sequence B (Fig. 3). To quantify this assertion, Fig. 7 shows cluster mass distributions for both sequences. From this figure, it can be seen that, in these systems, a typical sequence A droplet accommodates about 200 chains, whereas the corresponding number for sequence B is less than 50. Also worth noting is the statistical suppression of intermediate-mass clusters, which is particularly pronounced for sequence A. If phase separation occurs, then one expects to observe a single dominant droplet [21–23], as is the case for sequence A.

Another basic characteristic is the density of the droplets. Figure 8 shows average bead-density profiles around the center of mass of large clusters. Here a given cluster is defined as large if the number of chains exceeds a threshold (75 for sequence A and 20 for sequence B), and the density is calculated as a function of the distance from its center of mass, $r_{\text{c.m.}}$, counting all beads, whether or not they belong to a chain in the cluster. The total density is split into H and P densities. The calculated density profiles for sequence A are essentially flat at both small and large $r_{\text{c.m.}}$ [Fig. 8(a)], suggesting that these densities are representative for the interior of droplets and the dilute background, respectively. Using this property, we find that the density inside droplets is more than a factor 40 higher than that of the dilute background (where the total bead density is $0.019b^{-3}$). Note also that the droplets are homogeneous in composition; the H to P ratio is virtually independent of $r_{\text{c.m.}}$.

The droplets formed by sequence B exhibit, by contrast, a micellar structure, with a high-density core composed almost exclusively of H beads and a corona of mainly P beads [Fig. 8(b)]. The formation of a hydrophobic core is possible due to the block structure of this sequence. However, as the sequence is short and contains a stretch of P beads, this core can only accommodate a small number of chains, which explains the low mass of droplets formed by this sequence (Fig. 7). The mechanisms of micelle formation by block copolymers have been extensively studied by both theory and simulation [33–35].

While we have seen above that sequence A phase separates, it is still not immediately clear whether the dense phase is liquidlike. Therefore, we end with a brief assessment of the

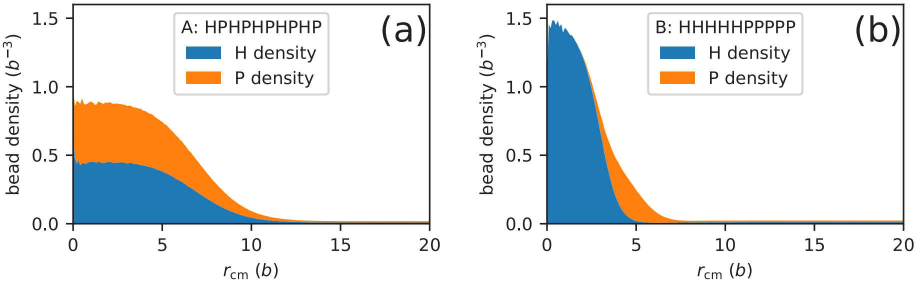


FIG. 8. Bead-density profiles calculated as a function of the distance $r_{\text{c.m.}}$ from the center of mass of large clusters for (a) sequence A and (b) sequence B. The data were obtained using $N = 640$, $\rho_b = 0.025b^{-3}$ and a temperature near the onset of droplet formation. A cluster is defined as large if the number of chains is above a cutoff (75 and 20 for sequences A and B, respectively). The total density is split into H and P densities. For comparison, a perfect close-packing of the beads would give a total density of $3.35b^{-3}$.

mobility of the chains in droplets formed by this sequence. The analysis uses configurations stored at a time interval of 10^3 MC cycles, which is much shorter than the average droplet lifetime of about 2×10^5 MC cycles. As before, a droplet is a cluster with more than 75 chains. We first consider the exchange of chains between droplets and their surroundings. To this end, whenever two consecutive snapshots both contain droplets, the chain contents of the droplets are compared. Over this timescale (10^3 MC cycles), it turns out that, on average, 44% of the chains present in the original droplet are lost, indicating a fast exchange with the surroundings compared to the lifetime of a droplet.

To get a measure of whether also the internal structure of a droplet is dynamic, we monitor changes in chain-chain contacts within droplets. To this end, given a droplet-containing snapshot, we identify all pairs of chains in the droplet that are in contact (interaction energy < 0) and where each chain also interacts with at least 15 other chains. The latter condition serves to focus the analysis on chain pairs buried in the interior of the droplet. Whenever a droplet is present also in the next snapshot (10^3 MC cycles later), we check the fate of the contacts identified in the first snapshot. On average, we find that 54% of the pairs remain in contact, whereas only about 11% are broken due to at least one of the chains leaving the droplet. This leaves 34% of the pairs separating due to internal rearrangements of the droplet, showing that the internal structure is far from rigid. Thus, the droplets are dynamic with respect to both exchange with the surroundings and their internal organization.

IV. DISCUSSION AND CONCLUSIONS

It is well known that finite-size scaling theory provides a powerful tool for analyzing phase transitions in spin models as well as vapor-to-droplet transitions in simple liquids. In this manuscript, we have applied these ideas to investigate the sequence-dependent phase behavior of a simple explicit-chain model for protein droplet formation.

Of the two specific sequences studied, the block sequence B turned out not to undergo LLPS. It is worth noting that from data for small systems, one may be led to the opposite conclusion. In particular, the observed peak in the specific heat is for small systems higher for sequence B than it is for the alternating sequence A, which does phase separate. However, above some system size (about 80 chains), the maximum specific heat does not increase further for sequence B, in contrast to what is observed for sequence A and to what one expects if phase separation takes place.

For sequence B, we observed micelle formation rather than the formation of a droplet of a dense bulk phase. Micelle formation was found to set in at a kT of about 5. Note that the system need not remain micellar in character well below this temperature. In particular, it is conceivable that the global free-energy minimum of this system contains bilayer structures at low temperatures. However, a proper investigation of the low-temperature phase structure is computationally challenging and beyond the scope of the present article.

To determine whether sequence A phase separates, simulation data for several properties and a range of system sizes were compared with predictions from finite-size scaling theory. In this way, the phase behavior can, in principle, be investigated in a systematic fashion, but it must be remembered that the theoretical results are leading-order predictions for large systems and therefore not necessarily valid for the system sizes amenable to simulation. It turned out, however, that a scaling behavior consistent with the predicted asymptotic one could be observed for all properties studied. Hence, taken together, the results of this analysis leave little doubt that sequence A does indeed phase separate.

It is worth noting that sequences with alternating hydrophobic and polar residues tend to have a high β-sheet propensity [36,37]. The biophysical model used in our present calculations cannot describe β-sheet formation, due to the lack of hydrogen bonding. However, it has been shown that droplet formation through LLPS sometimes is followed by maturation into a solidlike state containing amyloid fibrils [38]. In this case, LLPS represents a first step toward β-sheet formation.

Among the specific scaling relations studied, the finite-size shift of the transition temperature deserves special attention. This shift scales inversely proportional to the linear size, rather than the volume, of the critical droplet, so that $T_b^{(N)} - T_b \sim N^{-1/4}$ [Eq. (2)]. This slow convergence of the transition temperature $T_b^{(N)}$ toward its value for infinite system size, T_b, makes finite-size scaling analysis an important ingredient when determining the phase diagram from simulation data. This conclusion is highlighted by the magnitude of the relative shift of the transition temperature for sequence A, which was found to still be ∼8% for the largest systems with 640 chains.

Simulation methods, based on explicit-chain or field-theory representations, offer some distinct advantages over mean-field methods in the study of sequence-dependent biomolecular phase separation. However, to exploit the full potential of the simulations, the system-size dependence of the generated data needs to be understood and accounted for. The results presented here demonstrate that a systematic analysis of the system-size dependence can be both feasible and rewarding.

Note added in proof. We recently became aware of Ref. [39]. This article studied finite-size effects on pair distribution functions in a model for biomolecular LLPS. It did not use the theoretical framework employed in the present article.

ACKNOWLEDGMENTS

We thank Behruz Bezorg, Sandipan Mohanty, and Bo Söderberg for stimulating discussions. This work was in part supported by the Swedish Research Council (Grants No. 621-2014-4522 and No. 2018-04976) and the Swedish strategic research program eSSENCE. The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at LUNARC, Lund University, Sweden.

[1] C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, and A. A. Hyman, Science **324**, 1729 (2009).

[2] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, Nat. Rev. Mol. Cell Biol. **18**, 285 (2017).

[3] T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, and A. J. Baldwin, Mol. Cell **57**, 936 (2015).

[4] A. Molliex, J. Temirov, J. Lee, M. Coughlin, A. P. Kanagaraj, H. J. Kim, T. Mittag, and J. P. Taylor, Cell **163**, 123 (2015).

[5] K. A. Burke, A. M. Janke, C. L. Rhine, and N. L. Fawzi, Mol. Cell **60**, 231 (2015).

[6] M. L. Huggins, J. Chem. Phys. **9**, 440 (1941).

[7] P. J. Flory, J. Chem. Phys. **10**, 51 (1942).

[8] J. T. G. Overbeek and M. J. Voorn, J. Cell Comp. Physiol. **49**, 7 (1957).

[9] J. Wittmer, A. Johner, and J. F. Joanny, Europhys. Lett. **24**, 263 (1993).

[10] Y.-H. Lin, J. D. Forman-Kay, and H. S. Chan, Phys. Rev. Lett. **117**, 178101 (2016).

[11] G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal, PLoS Comput. Biol. **14**, e1005941 (2018).

[12] G. L. Dignon, W. Zheng, R. B. Best, Y. C. Kim, and J. Mittal, Proc. Natl. Acad. Sci. USA **115**, 9929 (2018).

[13] S. Das, A. Eisen, Y.-H. Lin, and H. S. Chan, J. Phys. Chem. B **122**, 5418 (2018).

[14] S. Das, A. N. Amin, Y.-H. Lin, and H. S. Chan, Phys. Chem. Chem. Phys. **20**, 28558 (2018).

[15] N. A. S. Robichaud, I. Saika-Voivod, and S. Wallin, Phys. Rev. E **100**, 052404 (2019).

[16] T. S. Harmon, A. S. Holehouse, M. K. Rosen, and R. V. Pappu, eLife **6**, e30294 (2017).

[17] T. S. Harmon, A. S. Holehouse, and R. V. Pappu, New J. Phys. **20**, 045002 (2018).

[18] S. Qin and H.-X. Zhou, J. Phys. Chem. B **120**, 8164 (2016).

[19] J. McCarty, K. T. Delaney, S. P. O. Danielsen, G. H. Fredrickson, and J.-E. Shea, J. Phys. Chem. Lett. **10**, 1644 (2019).

[20] Y. Lin, J. McCarty, J. N. Rauch, K. T. Delaney, K. S. Kosik, G. H. Fredrickson, J.-E. Shea, and S. Han, eLife **8**, e42571 (2019).

[21] K. Binder and M. H. Kalos, J. Stat. Phys. **22**, 363 (1980).

[22] M. Biskup, L. Chayes, and R. Kotecký, Europhys. Lett. **60**, 21 (2002).

[23] K. Binder, Physica A **319**, 99 (2003).

[24] T. Neuhaus and J. S. Hager, J. Stat. Phys. **113**, 47 (2003).

[25] M. Schrader, P. Virnau, and K. Binder, Phys. Rev. E **79**, 061104 (2009).

[26] J. Zierenberg and W. Janke, Phys. Rev. E **92**, 012134 (2015).

[27] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).

[28] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **58**, 86 (1987).

[29] A. Irbäck, S. Æ. Jónsson, N. Linnemann, B. Linse, and S. Wallin, Phys. Rev. Lett. **110**, 058101 (2013).

[30] R. G. Miller, Biometrika **61**, 1 (1974).

[31] M. Biskup, L. Chayes, and R. Kotecký, Commun. Math. Phys. **242**, 137 (2003).

[32] C. Borgs and R. Kotecký, Phys. Rev. Lett. **68**, 1734 (1992).

[33] L. Leibler, H. Orland, and J. C. Wheeler, J. Chem. Phys. **79**, 3550 (1983).

[34] R. Nagarajan and K. Ganesh, J. Chem. Phys. **90**, 5843 (1989).

[35] A. Milchev, A. Bhattacharya, and K. Binder, Macromolecules **34**, 1881 (2001).

[36] M. W. West, W. Wang, J. Patterson, J. D. Mancias, J. R. Beasley, and M. H. Hecht, Proc. Natl. Acad. Sci. USA **96**, 11211 (1999).

[37] N. B. Hung, D.-M. Le, and T. X. Hoang, J. Chem. Phys. **147**, 105102 (2017).

[38] A. Patel, H. O. Lee, L. Jawerth, S. Maharana, M. Jahnel, M. Y. Hein, S. Stoynov, J. Mahamid, S. Saha, T. M. Franzmann, A. Pozniakovski, I. Poser, N. Maghelli, L. A. Royer, M. Weigert, E. W. Myers, S. Grill, D. Drechsel, A. A. Hyman, and S. Alberti, Cell **162**, 1066 (2015).

[39] J. M. Choi, F. Dar, and R. V. Pappu, PLoS Comput. Biol. **15**, e1007028 (2019).

Paper III

Finite-size shifts in simulated protein droplet phase diagrams

Cite as: J. Chem. Phys. **154**, 235101 (2021); doi: 10.1063/5.0052813
Submitted: 1 April 2021 • Accepted: 27 May 2021 •
Published Online: 15 June 2021

View Online Export Citation CrossMark

Daniel Nilsson[a] and Anders Irbäck[b]

AFFILIATIONS

Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

[a] Electronic mail: daniel.nilsson@thep.lu.se
[b] Author to whom correspondence should be addressed: anders@thep.lu.se

ABSTRACT

Computer simulation can provide valuable insight into the forces driving biomolecular liquid–liquid phase separation. However, the simulated systems have a limited size, which makes it important to minimize and control finite-size effects. Here, using a phenomenological free-energy ansatz, we investigate how the single-phase densities observed in a canonical system under coexistence conditions depend on the system size and the total density. We compare the theoretical expectations with results from Monte Carlo simulations based on a simple hydrophobic/polar protein model. We consider both cubic systems with spherical droplets and elongated systems with slab-like droplets. The results presented suggest that the slab simulation method greatly facilitates the estimation of the coexistence densities in the large-system limit.

I. INTRODUCTION

Liquid–liquid phase separation (LLPS) has recently been identified as an important driver of compartmentalization in living cells.[1,2] Through LLPS, membraneless liquid droplets with high concentrations of proteins and nucleic acids are formed. Intrinsically disordered proteins (IDPs) often play an important role in this process, and several IDPs have been shown to phase separate on their own *in vitro*.[3–5]

Recent years have seen a growing number of theoretical and computational investigations of biomolecular LLPS with some emphasis on IDPs. These studies have provided insights into the forces driving biomolecular LLPS and especially into the sequence-dependence of IDP LLPS. A mainly analytical method that has been adopted for this purpose is the random-phase approximation[6] by which sequence-determinants of polyampholyte LLPS were elucidated.[7] A computationally demanding but more general approach is to use numerical simulation of coarse-grained models based on either explicit-chain[8–21] or field-theoretic[22–24] representations of the biomolecular systems. The most widely used of these alternatives is explicit-chain simulation, which provides a versatile method for exploring the sequence-dependence of IDP LLPS as well as the basic structural properties of condensates. However, both explicit-chain and field-theory simulations tend to become time-consuming for large systems, which makes it important to be able to minimize and control the system-size dependence of the results.

The temperature-density (T-ρ) phase diagram of a phase-separating sequence can be investigated in a systematic manner by performing simulations close to the phase boundary that defines the coexistence region, for different system sizes, followed by a finite-size scaling (FSS) extrapolation[25–27] to the large-system limit. This approach provides information on both the character and location of the transition that occurs at the phase boundary and was recently tested by us on a simple continuous hydrophobic/polar explicit-chain model.[18] However, if the main focus is to locate the phase boundary, then a common choice is to rely on (canonical) simulations in the coexistence region, where the dense and dilute phases are both present. The single-phase densities of such a system provide estimates of the transition densities at the temperature used. The simulated systems may be cubic with spherical droplets. An often used alternative for phase-separation studies is to adopt elongated geometries, thereby causing droplets to take on a slab-like shape.[10,28–32]

In this article, we investigate how the asymptotic coexistence densities are shifted in finite-size canonical systems with cubic and elongated geometries. Using a phenomenological free-energy ansatz for the mixed two-phase regime,[33,34] we derive an expression describing how the single-phase densities depend on the system size and the total density. Comparing with Monte Carlo simulations based on the hydrophobic/polar model in Ref. 18, we find that this expression can be used to rationalize data for both cubic and elongated systems. In line with the analytical result, the simulations show that using an elongated geometry greatly reduces the finite-size effects on the coexistence densities.

II. METHODS

A. Biophysical model

Our simulations are performed using a minimal off-lattice protein model[18] in which each protein chain is represented as a string of m hydrophobic (H) or polar (P) beads. The simulated systems consist of N chains sharing the same HP sequence, which are enclosed in a periodic box with volume V. The box is either cubic ($V = L^3$) or elongated in one of the dimensions ($V = L_1 L_2^2$, with $L_1 > L_2$).

The length of the bond connecting two consecutive beads in a chain, b, is kept fixed, while the polar and azimuthal bond angles are both free to vary. In the absence of interactions, the bonds have a spherically uniform distribution. All beads are assigned a common hard-sphere diameter, d_{ev}.

The interaction potential is a sum over all intra- and intermolecular pairs of beads in the system, $E = \sum_{i<j} E_{ij}$, where E_{ij} has a square-well shape with depth ε for HH pairs, whereas HP and PP pairs interact through a pure hard-sphere potential. For a bead pair at distance r_{ij} from each other, the pair potential is given by

$$E_{ij} = \begin{cases} \infty & \text{if } r_{ij} < d_{\text{ev}} \\ \varepsilon_{ij} & \text{if } d_{\text{ev}} < r_{ij} < \Lambda \\ 0 & \text{if } r_{ij} > \Lambda, \end{cases} \tag{1}$$

where $d_{\text{ev}} = 0.75b$, $\Lambda = 2b$, and $\varepsilon_{ij} = -\varepsilon$ (HH pairs) or 0 (HP and PP pairs).

Throughout this article, lengths and energies are given in units of b and ε, respectively.

B. Phase-separation phenomenology

In this section, we consider a general system of particles or chains at some fixed temperature T at which we assume that two bulk phases with densities ρ_ℓ^c and ρ_h^c coexist. Coexistence occurs for total densities ρ in the interval $\rho_\ell^c < \rho < \rho_h^c$ in the limit of infinite system size.

In large but finite systems, where interface effects cannot be neglected, the mixed two-phase behavior sets in at slightly shifted densities, ρ_ℓ^t and ρ_h^t. The finite-size effects on the character and location of the transition to the mixed two-phase regime have been extensively investigated.[25–27] In particular, it was shown that the finite-size shift $\rho_i^t - \rho_i^c$ $(i = \ell, h)$ scales as $V^{-1/4}$. This FSS framework, which can be used to study transition temperatures as well, has been applied to analyze simulations of, for example, the lattice gas,[35] the Lennard–Jones fluid,[36] polymer cluster formation,[37] and recently also the protein model studied in this paper.[18]

Here, we focus entirely on the problem of estimating the infinite-system-size transition densities ρ_ℓ^c and ρ_h^c. This problem can be approached by using simulations near the transition densities for different system sizes in combination with FSS analysis. A common alternative is to use simulations at total densities ρ well into the regime where both phases are present and measure the single-phase densities $\rho_\ell(\rho, V)$ and $\rho_h(\rho, V)$. For large V, the densities $\rho_\ell(\rho, V)$ and $\rho_h(\rho, V)$ provide accurate estimates of the respective asymptotic densities ρ_ℓ^c and ρ_h^c, independent of the precise choice of ρ.

The finite-size shifts $\delta\rho_i = \rho_i(\rho, V) - \rho_i^c$ $(i = \ell, h)$ can be estimated by adopting a phenomenological free-energy ansatz for the mixed two-phase system, given by[33,34]

$$F(\rho_\ell, \rho_h, V_\ell, V_h) = f_\ell(\rho_\ell)V_\ell + f_h(\rho_h)V_h + \gamma A_{\ell h}, \tag{2}$$

where $f_i(\rho_i) = \mu_i(\rho_i)\rho_i - p_i(\rho_i)$ denotes the free-energy density of bulk phase i $(i = \ell, h)$, $A_{\ell h}$ is the interface area, and γ is the surface tension. For simplicity, we assume a constant surface tension, which should be a good approximation unless the droplet of the minority phase is small.[34] By minimizing the free energy $F(\rho_\ell, \rho_h, V_\ell, V_h)$ in Eq. (2) subject to the constraints $N = \rho_\ell V_\ell + \rho_h V_h$ and $V = V_\ell + V_h$, one can determine how the N particles and the volume V are partitioned between the two phases and thus obtain the finite-size densities $\rho_\ell(\rho, V)$ and $\rho_h(\rho, V)$.

In the large-V limit, the surface term in Eq. (2) can be neglected. In its absence, the constrained minimization of F leads to the conditions

$$f'_\ell(\rho_\ell) = f'_h(\rho_h) = \frac{f_h(\rho_h) - f_\ell(\rho_\ell)}{\rho_h - \rho_\ell}, \tag{3}$$

which simply says that the two phases share the same chemical potential $\mu = f'_i(\rho_i)$ and the same pressure $p = \mu\rho_i - f_i(\rho_i)$ $(i = \ell, h)$. The conditions in Eq. (3) are thus fulfilled when the densities take their asymptotic values ρ_ℓ^c and ρ_h^c, as they should be.

For large but finite volumes, the mass balance condition remains unchanged, $\mu_\ell(\rho_\ell) = \mu_h(\rho_h)$, whereas the volume balance condition picks up an additional surface term to become the Young–Laplace equation, $p_h(\rho_h) = p_\ell(\rho_\ell) + \gamma dA_{\ell h}/dV_h$. This extra term leads to a shift of the volumes and densities of the two phases. To leading order, the density shifts are given by

$$\delta\rho_i = \rho_i(\rho, V) - \rho_i^c = \frac{\gamma\kappa_i^c\rho_i^{c\,2}}{\rho_h^c - \rho_\ell^c}\frac{dA_{\ell h}}{dV_h} \qquad (i = \ell, h), \tag{4}$$

where $\kappa_i^c = 1/(f''_i(\rho_i^c)\rho_i^{c\,2})$ is the isothermal compressibility of phase i at coexistence. The derivative $dA_{\ell h}/dV_h$ in Eq. (4) depends on V_h, which to leading order, by the lever rule, can be written as $V_h = V(\rho - \rho_\ell^c)/(\rho_h^c - \rho_\ell^c)$.

Under conditions such that the dense phase forms a spherical droplet ($dA_{\ell h}/dV_h \propto V_h^{-1/3}$), it follows that the density shifts in Eq. (4) are positive and scale as

$$\delta\rho_i \propto \frac{1}{V^{1/3}(\rho - \rho_\ell^c)^{1/3}} \qquad (i = \ell, h). \tag{5}$$

This result suggests that the single-phase densities $\rho_i(\rho, V)$ both decrease with increasing total density, which, at first glance, may seem paradoxical. This behavior is possible because the volume filled by the dense phase increases and can be seen as a consequence of the shape of the chemical potential $\mu(\rho)$ in finite systems.[34]

At fixed ρ, the finite-size shift of the single-phase densities scales as $V^{-1/3}$ [Eq. (5)]. This may be compared with the finite-size shift of the transition densities, which, as indicated above, scales as $V^{-1/4}$.[25–27] In both cases, the approach to the asymptotic densities ρ_ℓ^c and ρ_h^c is rather slow.

Both these inferred system-size dependencies are for the case that the minority phase forms a spherical droplet or bubble. However, the density shifts predicted by Eq. (4) depend on the "droplet" geometry through the derivative $dA_{\ell h}/dV_h$. In particular, if the minority phase forms a slab, extending over the periodic boundary in two of the three dimensions, then the interface area becomes independent of the droplet volume so that the leading-order density shifts vanish [Eq. (4)]. The formation of slab-like droplets can be promoted by using boxes that are elongated in one of the dimensions, which is a common choice in phase-separation studies.[10,28,30–32]

C. Monte Carlo simulations

We investigate the thermodynamics of the HP protein model described above through equilibrium Monte Carlo simulations in the canonical (NVT) ensemble, using a set of six elementary moves. Two of these moves alter the internal structure of individual chains, either through a single-bead move or through a pivot-type rotation of part of the chain relative to the rest. The other moves are rigid-body translations and rotations of either a single chain or a cluster of chains.

The clusters are constructed stochastically, following a Swendsen–Wang-type procedure.[38,39] Here, a random chain is selected as the first cluster member. New chains are then added iteratively with probability $p_{ij} = 1 - e^{\beta E_{ij}}$, where E_{ij} (≤ 0) is the interaction energy between chain i in the cluster and chain j not (yet) added to the cluster ($\beta = 1/k_B T$). This step is repeated until all potential additions to the cluster have been tested. Finally, the resulting cluster is subject to a trial rigid-body move. The probabilistic construction of the cluster is such that the move can be accepted whenever sterically allowed, without invoking any Metropolis accept/reject step. For a cluster rotation, the periodic boundary conditions may cause the E_{ij} values to change upon the proposed move. If so, the move has to be rejected. Note also that the particular form we use for the chain-addition probability p_{ij} assumes that all $E_{ij} \leq 0$, as is the case in our model.

The simulated systems consist of 320 chains in boxes with varying volume and either cubic or elongated shape. In the elongated systems, the volume $V = L_1 L_2^2$ is altered by changing the longitudinal size, L_1, while keeping the transverse size, L_2, fixed. All slab simulation results quoted below are for $L_2 = 12b$. Additional control simulations were performed using several different L_2 values ranging from $L_2 = 8b$ to $L_2 = 16b$. The results obtained did not show any statistically significant L_2-dependence.

Each simulation run consists of 10^7 sweeps, where one sweep corresponds to mN attempted elementary moves. All simulations are started from random initial states. The thermalization period required for a dense-phase droplet to form represents a negligible fraction of the total simulation time.

D. Density profiles

To determine the single-phase densities ρ_ℓ and ρ_h, we compute the bead density distribution around the center of dense-phase droplets. In a given snapshot, the droplet is identified by clustering the chains,[40,41] in our case, based on the criterion that two chains with non-zero interaction energy must be in the same cluster. For most snapshots, this procedure gave a single dominant cluster, which is taken to define the droplet. However, some snapshots from the cubic simulations (<1%) contained no cluster with more than 50 chains. These snapshots were omitted in constructing average density profiles. A droplet was identified in all snapshots from the slab simulations.

Given a snapshot with an identified droplet, we compute the distribution of beads around the center of mass of the droplet, counting all beads, whether or not they belong to a chain in the droplet. Thus, the sole purpose of the droplet identification is to determine the droplet center. The final density distributions are obtained by averaging over the snapshots. In the cubic systems, with spherical droplets, the density distribution is calculated as a function of the radial distance to the droplet center, $r_{\text{c.m.}}$. In the elongated systems, we use the projected distance onto the elongated direction, $z_{\text{c.m.}}$.

From the thus calculated profiles, the single-phase densities ρ_ℓ and ρ_h can be obtained using the data at large and small distances, respectively.

III. RESULTS

Using the model and methods described in Sec. II, we study systems composed of multiple copies of the same ten-bead hydrophobic/polar chain, HPHPHPHPHP. In previous work,[18] we showed by FSS analysis that this alternating sequence, unlike the block sequence HHHHHPPPPP, undergoes phase separation. The block sequence forms miscelles of limited size rather than droplets of a dense bulk phase. For the alternating sequence studied here, in the (ρ, T) plane, mixed two-phase behavior is observed under the coexistence, or binodal, curve, $T_b(\rho)$.

In this work, we focus on the problem of determining the asymptotic coexistence densities ρ_ℓ^c and ρ_h^c, at a given temperature, from finite-size simulation data taken in the coexistence region. The densities ρ_ℓ^c and ρ_h^c coincide with the total density values between which coexistence is observed and thus provide two points on the phase boundary $T_b(\rho)$. We consider both cubic systems with spherical droplets and elongated systems with slab-like droplets. The simulation results are compared with the finite-size shifts of the coexistence densities predicted by Eq. (4). Where relevant we also compare the simulation results with simulation data from our previous study.[18]

Throughout our calculations, the number of chains is the same, namely, $N = 320$. The total density ρ is varied by changing the volume. All densities quoted are bead, rather than chain, densities.

Figure 1 shows representative snapshots from simulations in the cubic and elongated geometries. In both cases, there is a dense droplet present in a dilute background. The dense minority phase

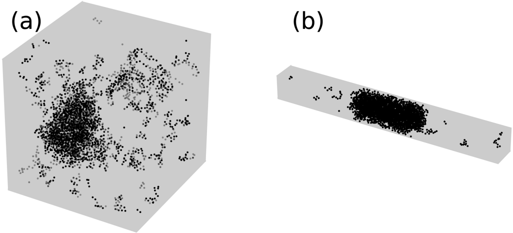


FIG. 1. Representative snapshots from simulations using cubic and elongated boxes at $T = 2.86\varepsilon/k_B$ and $\rho = 0.02\,375b^{-3}$. Both a dense and a dilute phase are present in both cases. The dense phase forms (a) an approximately spherical droplet in the cubic box and (b) a slab in the elongated box, which, for clarity, has been truncated in the longitudinal direction.

forms an approximately spherical droplet in the cubic box and a slab in the box with one elongated side.

A. Extracting finite-size coexistence densities

In our simulations, we determine the dilute- and dense-phase densities, ρ_ℓ and ρ_h, from the bead density distribution around the center of mass of the droplets, as described in Sec. II D. Figure 2 shows representative examples of such density profiles from simulations of both cubic and elongated systems for three total densities ρ and $T = 2.86\varepsilon/k_B$. The calculated density profiles level off at both small and large distances, which implies that ρ_ℓ and ρ_h can be estimated from the data at large and small distances, respectively. The finite-size single-phase densities ρ_ℓ and ρ_h at given ρ, V, and T are predicted by Eq. (4) to depend on whether the droplets are spherical or slab-like. Specifically, this equation predicts the finite-size shifts $\delta\rho_i = \rho_i - \rho_i^c$ $(i = \ell, h)$ to be positive if the droplets are spherical while vanishing for a slab-like droplet whose surface area does not change with the droplet volume. Consistent with this, the data in Fig. 2 suggest that both ρ_ℓ and ρ_h are higher in the cubic systems than in the elongated ones. With a dilute minority phase, one would instead expect ρ_ℓ and ρ_h to be lower in cubic systems with spherical bubbles than in elongated systems with slab-like bubbles [Eq. (4)].

It is also instructive to look at how the single-phase densities in a given geometry depend on the total density, ρ (Fig. 2). In the cubic case, with a spherical droplet, the positive shifts $\delta\rho_i$ predicted by Eq. (4) scale as $\delta\rho_i \sim V_h^{-1/3}$ $(i = \ell, h)$ with the droplet volume, V_h, which, in turn, increases with ρ, by the lever rule. Hence, the finite-size single-phase densities should decrease with increasing ρ, as is indeed observed in Fig. 2(a). A more detailed discussion of this ρ-dependence will be given toward the end of this section. In the elongated geometry, the simulated single-phase densities show no detectable dependence on ρ [Fig. 2(b)]. This behavior is also consistent with Eq. (4) since the predicted leading-order shifts vanish for a slab-like droplet.

B. Estimating asymptotic coexistence densities

Having seen that the simulations in both cubic and elongated geometries yield single-phase densities ρ_ℓ and ρ_h that qualitatively match well with the finite-size shifts predicted by Eq. (4), we now turn to the problem of estimating the asymptotic densities ρ_ℓ^c and ρ_h^c. Knowledge of these densities at a given temperature gives two points on the coexistence curve $T_b(\rho)$.

We first discuss the slab simulations in elongated geometries. As indicated above, our data do not reveal any significant finite-size shifts of the single-phase densities in the elongated systems. Therefore, we simply take the results obtained for some suitable choice of ρ and V as estimates of the asymptotic densities, without invoking any extrapolation. The results obtained, for seven choices of T, are displayed in Fig. 3, along with three data points from our previous study.[18] The latter were obtained by determining the finite-size transition temperature in cubic systems at a given ρ for several different volumes, followed by an FSS extrapolation.[18] These data are in approximate but not perfect agreement with those from the slab simulations, with the former falling a few percent above the latter. This discrepancy can be clearly seen in Fig. 4(a), which provides a zoomed-in view of the low-density branch of the coexistence curve.

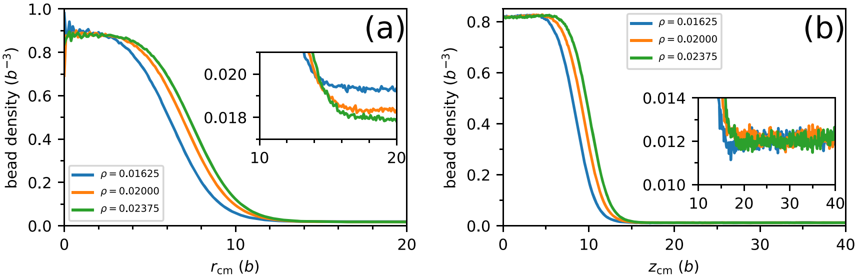


FIG. 2. Average bead density distributions around the center of mass of large clusters in (a) cubic and (b) elongated geometries, for three total densities ρ ($0.01\,625b^{-3}, 0.02\,000b^{-3}, 0.02\,375b^{-3}$), at $T = 2.86\varepsilon/k_B$. The single-phase densities ρ_h and ρ_ℓ can be estimated from the plateaus at small and large distances, respectively. Insets are zoomed-in views of the tails of the distributions. In the cubic geometry, the single-phase density shifts decrease with increasing ρ, i.e., increasing droplet volume, as expected from Eq. (4). In the elongated geometry, where the first-order shifts vanish, no statistically significant ρ dependence can be detected.

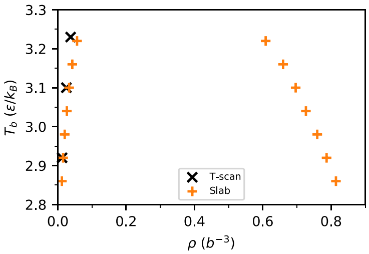


FIG. 3. Data for the coexistence curve, $T_b(\rho)$. Orange symbols indicate single-phase densities from slab simulations at seven temperatures at a total density of ρ = 0.015b^{-3}. To be able to unambiguously measure the single-phase densities at higher T, closer to the critical point, simulations of larger systems would be required. Black symbols represent data based on cubic simulations and FSS analysis.[18] A zoomed-in view of the low-density branch of the coexistence curve can be found in Fig. 4. The statistical errors are smaller than the symbol size.

The small but systematic shift between these two datasets prompted us to reanalyze the data for the finite-size transition temperatures, $T_b^{(N)}$, from Ref. 18. In that study, the asymptotic transition temperatures, T_b, were estimated for three values of ρ by seemingly good fits of the leading-order form[25–27,36] $T_b^{(N)} - T_b \propto N^{-1/4}$. However, the rather slow decay of the finite-size transition temperature shift makes the determination of T_b a delicate task. Therefore, to test the robustness of the extrapolation, given data for $T_b^{(N)}$ and $T_b^{(N/2)}$, we compute the quantity $\tilde{T}_b{}^{(N)} = (2^{1/4}T_b^{(N)} - T_b^{(N/2)})/(2^{1/4} - 1)$, which becomes equal to T_b if the finite-size transition temperature shift scales as $N^{-1/4}$. The data for $\tilde{T}_b{}^{(N)}$ shows a clear N-dependence [Fig. 4(b)], indicating that there are non-negligible higher-order corrections to the leading-order form $T_b^{(N)} - T_b \propto N^{-1/4}$. The shape of the data suggests that fits of this form may overestimate T_b, unless restricted to sufficiently large N. On the other hand, for large N, the $\tilde{T}_b{}^{(N)}$ values fall close to T_b estimates based on slab simulation data, which are indicated by the horizontal lines in Fig. 4(b). This observation, in particular, provides further support for the conclusion that the finite-size effects on the coexistence densities are small in the slab simulations.

Finally, we discuss the possibility of determining the asymptotic coexistence densities, ρ_ℓ^c and ρ_h^c, from cubic simulations with spherical droplets by means of Eq. (4). The predicted finite-size coexistence density shifts are, in this case, inversely proportional to the linear size of the droplet or equivalently $\delta\rho_i = \rho_i - \rho_i^c \sim N_h^{-1/3}$ ($i = \ell, h$), where N_h is the number of chains in the droplet. One way to increase the droplet size is by increasing the total number of chains, N, at fixed total density ρ [Eq. (5)]. A computationally appealing alternative is to keep N fixed and decrease V. Obviously, V cannot be reduced indefinitely or the spherical droplet is lost. Still, one may ask whether a sufficiently large range of droplet sizes can be covered to permit extrapolation of ρ_ℓ^c and ρ_h^c.

To assess the feasibility of this approach, we consider data for the dilute-phase density, ρ_ℓ, from cubic simulations with N = 320, $T = 2.86\varepsilon/k_B$, and varying V. We focus on ρ_ℓ, which is easier to determine than ρ_h. As expected, when decreasing V, we find that the droplet size, N_h, increases while ρ_ℓ decreases. At the smallest V studied, the droplet contains just above 220 chains. We stop at this V to ensure that the simulated droplet never extends over the periodic boundary, although it might be possible to further slightly reduce V without losing the spherical droplet. Figure 5 shows the data for ρ_ℓ, which can be quite well described by a fit of the form $\rho_\ell - \rho_\ell^c \propto N_h^{-1/3}$ [Eq. (4)]. However, the fit yields $\rho_\ell^c = 0.009\, b^{-3}$, whereas slab simulation data suggest that $\rho_\ell^c = 0.012\, b^{-3}$ [Fig. 2(b)]. This rather poor agreement is not surprising, given the limited range of droplet sizes covered by the data (Fig. 5). Unfortunately, for an accurate extrapolation of ρ_ℓ^c, or ρ_h^c, one would have to use much larger droplets, which requires prohibitively time-consuming simulations with much larger N.

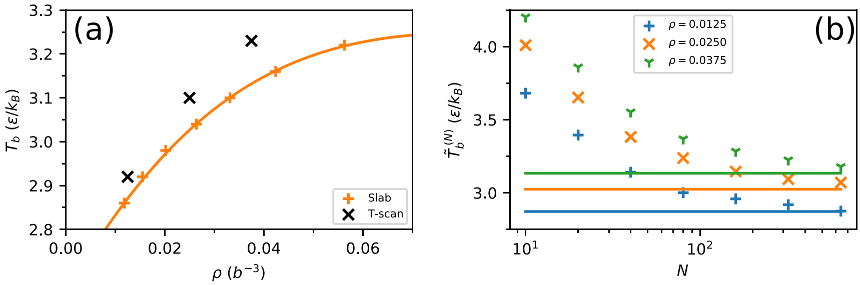


FIG. 4. (a) Zoomed-in view of the low-density branch of the coexistence curve (Fig. 3). Symbols are as in Fig. 3. The line is a fit of the form $|\rho - \rho_c| = A(T_c - T)^{0.32642}$ with which we interpolate the slab simulation data (orange symbols) to enable comparison with results from cubic simulations[18] (black symbols). (b) The quantity $\tilde{T}_b{}^{(N)} = (2^{1/4}T_b^{(N)} - T_b^{(N/2)})/(2^{1/4} - 1)$, which depends on the two finite-size transition temperatures $T_b^{(N)}$ and $T_b^{(N/2)}$, plotted against N, using data from the cubic simulations in Ref. 18 for three values of ρ. Horizontal lines indicate transition temperatures based on slab simulation data, extracted through the fit shown in (a). The N-dependence of $\tilde{T}_b{}^{(N)}$ indicates that higher-order corrections to the leading-order form $T_b^{(N)} - T_b \propto N^{-1/4}$ are non-negligible and explains why fits of this form may overestimate T_b. Note also that the data from the large cubic systems are close to the slab simulation data. The statistical errors are smaller than the symbol size.

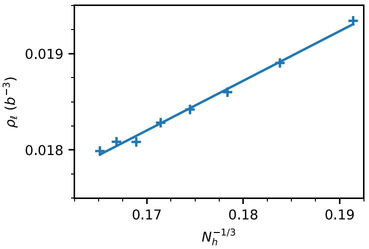


FIG. 5. Dilute-phase density, ρ_ℓ, in cubic systems with a spherical droplet for $T = 2.86\varepsilon/k_B$, $N = 320$, and varying total density ρ ($0.01\,625b^{-3} < \rho < 0.025b^{-3}$). The data are plotted against $N_h^{-1/3}$, where N_h is the number of chains in the droplet. This number, N_h, is computed by following the clustering procedure described in Sec. II D for individual snapshots of the system and then averaging over snapshots. The line represents a fit of the form $\rho_\ell = \rho_\ell^c + cN_h^{-1/3}$ [Eq. (4)]. The fit is reasonable, but the range of $N_h^{-1/3}$ covered is too small to permit an accurate extrapolation of the asymptotic density ρ_ℓ^c. The fitted value is $\rho_\ell^c = 0.009\,b^{-3}$, whereas the slab simulations suggest that $\rho_\ell^c = 0.012\,b^{-1/3}$ [Fig. 2(b)]. The statistical errors are smaller than the symbol size.

IV. DISCUSSION AND SUMMARY

In computational studies of phase-separating systems, a widely used method for estimating the phase diagram is to conduct simulations under coexistence conditions and measure the single-phase densities. In this paper, we have analyzed simulated single-phase densities based on a phenomenological ansatz for the free energy of a mixed two-phase system, $F(\rho_\ell, \rho_h, V_\ell, V_h)$.[33,34] Minimizing the free energy, subject to the constraints $N = \rho_\ell V_\ell + \rho_h V_h$ and $V = V_\ell + V_h$, yields a simple and general leading-order expression for the finite-size shifts of the coexistence densities [Eq. (4)], which depends on the droplet geometry through the derivative $dA_{\ell h}/dV_h$. In the case of spherical droplets, the shifts are rather slowly decaying functions of the droplet volume ($\propto V_h^{-1/3}$). If, on the other hand, the droplets are slab-like, then changes in the droplet volume leave the interfacial surface area unchanged so that $dA_{\ell h}/dV_h = 0$, which makes the predicted finite-size shifts vanish.

The results of our simulations, based on a simple HP protein model, are fully consistent with the predictions by Eq. (4). In cubic systems with spherical droplets, we thus observe significant finite-size shifts of the coexistence densities. In fact, the magnitude and slow decay of these shifts make accurate extrapolation of the asymptotic coexistence densities a challenge. By contrast, in the slab simulations, we do not find any detectable finite-size shifts of the coexistence densities. The slab simulation data for the coexistence curve were also compared with previous results based on temperature scans in cubic systems with varying size.[18] A reanalysis of the previous data shows that the transition temperatures in large cubic systems match well with the slab simulation results, which further strengthens the conclusion that the finite-size shifts are small in the slab simulations.

It should be noted that our study has only looked at single-phase densities in finite systems and their relation to the coexistence densities in the large-system limit. Other properties may be less well suited to slab simulation, especially near the transition to the coexistence region. Thus, slab simulation may be more useful for locating this transition than for investigating its character.

Nevertheless, when it comes to exploring the shape of the phase diagram, the smallness of the finite-size coexistence density shifts makes slab simulation a very attractive method, which may open for studies of LLPS in systems that would otherwise be too demanding to simulate.

ACKNOWLEDGMENTS

We thank Behruz Bozorg, Sandipan Mohanty, and Bo Söderberg for stimulating discussions. This work was, in part, supported by the Swedish Research Council (Grant No. 621-2018-04976) and the Swedish strategic research program eSSENCE. The simulations were, in part, performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at LUNARC, Lund University, Sweden.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, and A. A. Hyman, "Germline P granules are liquid droplets that localize by controlled dissolution/condensation," Science **324**, 1729–1732 (2009).

[2] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, "Biomolecular condensates: Organizers of cellular biochemistry," Nat. Rev. Mol. Cell Biol. **18**, 285–298 (2017).

[3] T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, and A. J. Baldwin, "Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles," Mol. Cell **57**, 936–947 (2015).

[4] A. Molliex, J. Temirov, J. Lee, M. Coughlin, A. P. Kanagaraj, H. J. Kim, T. Mittag, and J. P. Taylor, "Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization," Cell **163**, 123–133 (2015).

[5] K. A. Burke, A. M. Janke, C. L. Rhine, and N. L. Fawzi, "Residue-by-residue view of *in vitro* FUS granules that bind the C-terminal domain of RNA polymerase II," Mol. Cell **60**, 231–241 (2015).

[6] J. Wittmer, A. Johner, and J. F. Joanny, "Random and alternating polyampholytes," Europhys. Lett. **24**, 263–268 (1993).

[7] Y.-H. Lin, J. D. Forman-Kay, and H. S. Chan, "Sequence-specific polyampholyte phase separation in membraneless organelles," Phys. Rev. Lett. **117**, 178101 (2016).

[8] T. S. Harmon, A. S. Holehouse, M. K. Rosen, and R. V. Pappu, "Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins," eLife **6**, e30294 (2017).

[9] T. S. Harmon, A. S. Holehouse, and R. V. Pappu, "Differential solvation of intrinsically disordered linkers drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins," New J. Phys. **20**, 045002 (2018).

[10] G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal, "Sequence determinants of protein phase behavior from a coarse-grained model," PLoS Comput. Biol. **14**, e1005941 (2018).

[11] G. L. Dignon, W. Zheng, R. B. Best, Y. C. Kim, and J. Mittal, "Relation between single-molecule properties and phase behavior of intrinsically disordered proteins," Proc. Natl. Acad. Sci. U. S. A. **115**, 9929–9934 (2018).

[12] S. Das, A. Eisen, Y.-H. Lin, and H. S. Chan, "A lattice model of charge-pattern-dependent polyampholyte phase separation," J. Phys. Chem. B **122**, 5418–5431 (2018).

[13] S. Das, A. N. Amin, Y.-H. Lin, and H. S. Chan, "Coarse-grained residue-based models of disordered protein condensates: Utility and limitations of simple charge pattern parameters," Phys. Chem. Chem. Phys. **20**, 28558–28574 (2018).

[14] V. Nguemaha and H.-X. Zhou, "Liquid-liquid phase separation of patchy particles illuminates diverse effects of regulatory components on protein droplet formation," Sci. Rep. **8**, 6728 (2018).

[15] N. A. S. Robichaud, I. Saika-Voivod, and S. Wallin, "Phase behavior of blocky charge lattice polymers: Crystals, liquids, sheets, filaments, and clusters," Phys. Rev. E **100**, 052404 (2019).

[16] W. Zheng, G. L. Dignon, N. Jovic, X. Xu, R. M. Regy, N. L. Fawzi, Y. C. Kim, R. B. Best, and J. Mittal, "Molecular details of protein condensates probed by microsecond long atomistic simulations," J. Phys. Chem. B **124**, 11671–11679 (2020).

[17] B. Xu, G. He, B. G. Weiner, P. Ronceray, Y. Meir, M. C. Jonikas, and N. S. Wingreen, "Rigidity enhances a magic-number effect in polymer phase separation," Nat. Commun. **11**, 1561 (2020).

[18] D. Nilsson and A. Irbäck, "Finite-size scaling analysis of protein droplet formation," Phys. Rev. E **101**, 022413 (2020).

[19] G. Krainer, T. J. Welsh, J. A. Joseph, J. R. Espinosa, S. Wittmann, E. de Csilléry, A. Sridhar, Z. Toprakcioglu, G. Gudiškytė, M. A. Czekalska, W. E. Arter, J. Guillén-Boixet, T. M. Franzmann, S. Qamar, P. S. George-Hyslop, A. A. Hyman, R. Collepardo-Guevara, S. Alberti, and T. P. J. Knowles, "Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions," Nat. Commun. **12**, 1085 (2021).

[20] Y. Xing, A. Nandakumar, A. Kakinen, Y. Sun, T. P. Davis, P. C. Ke, and F. Ding, "Amyloid aggregation under the lens of liquid–liquid phase separation," J. Phys. Chem. Lett. **12**, 368–378 (2021).

[21] M. K. Hazra and Y. Levy, "Biophysics of phase separation of disordered proteins is governed by balance between short- and long-range interactions," J. Phys. Chem. B **125**, 2202–2211 (2021).

[22] J. McCarty, K. T. Delaney, S. P. O. Danielsen, G. H. Fredrickson, and J.-E. Shea, "Complete phase diagram for liquid–liquid phase separation of intrinsically disordered proteins," J. Phys. Chem. Lett. **10**, 1644–1652 (2019).

[23] Y. Lin, J. McCarty, J. N. Rauch, K. T. Delaney, K. S. Kosik, G. H. Fredrickson, J.-E. Shea, and S. Han, "Narrow equilibrium window for complex coacervation of tau and RNA under cellular conditions," eLife **8**, e42571 (2019).

[24] J. Wessén, T. Pal, S. Das, Y.-H. Lin, and H. S. Chan, "A simple explicit-solvent model of polyampholyte phase behaviors and its ramifications for dielectric effects in biomolecular condensates," J. Phys. Chem. B **125**, 4337–4358 (2021).

[25] K. Binder and M. H. Kalos, "'Critical clusters' in a supersaturated vapor: Theory and Monte Carlo simulation," J. Stat. Phys. **22**, 363–396 (1980).

[26] M. Biskup, L. Chayes, and R. Kotecký, "On the formation/dissolution of equilibrium droplets," Europhys. Lett. **60**, 21–27 (2002).

[27] K. Binder, "Theory of the evaporation/condensation transition of equilibrium droplets in finite volumes," Physica A **319**, 99–114 (2003).

[28] K. Binder, "Monte Carlo calculation of the surface tension for two- and three-dimensional lattice-gas models," Phys. Rev. A **25**, 1699–1709 (1982).

[29] P. Virnau, M. Müller, L. G. MacDowell, and K. Binder, "Phase behavior of n-alkanes in supercritical solution: A Monte Carlo study," J. Chem. Phys. **121**, 2169–2179 (2004).

[30] F. J. Blas, L. G. MacDowell, E. de Miguel, and G. Jackson, "Vapor-liquid interfacial properties of fully flexible Lennard-Jones chains," J. Chem. Phys. **129**, 144703 (2008).

[31] P. V. Ramírez-González, S. E. Quiñones-Cisneros, and U. K. Deiters, "Chemical potentials and phase equilibria of Lennard-Jones chain fluids," Mol. Phys. **113**, 28–35 (2014).

[32] K. S. Silmore, M. P. Howard, and A. Z. Panagiotopoulos, "Vapour–liquid phase equilibrium and surface tension of fully flexible Lennard–Jones chains," Mol. Phys. **115**, 320–327 (2017).

[33] L. G. MacDowell, P. Virnau, M. Müller, and K. Binder, "The evaporation/condensation transition of liquid droplets," J. Chem. Phys. **120**, 5293–5308 (2004).

[34] M. Schrader, P. Virnau, and K. Binder, "Simulation of vapor-liquid coexistence in finite volumes: A method to compute the surface free energy of droplets," Phys. Rev. E **79**, 061104 (2009).

[35] T. Neuhaus and J. S. Hager, "2D crystal shapes, droplet condensation, and exponential slowing down in simulations of first-order phase transitions," J. Stat. Phys. **113**, 47–83 (2003).

[36] J. Zierenberg and W. Janke, "Exploring different regimes in finite-size scaling of the droplet condensation-evaporation transition," Phys. Rev. E **92**, 012134 (2015).

[37] J. Zierenberg, P. Schierz, and W. Janke, "Canonical free-energy barrier of particle and polymer cluster formation," Nat. Commun. **8**, 14546 (2017).

[38] R. H. Swendsen and J.-S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," Phys. Rev. Lett. **58**, 86–88 (1987).

[39] A. Irbäck, S. Æ. Jónsson, N. Linnemann, B. Linse, and S. Wallin, "Aggregate geometry in amyloid fibril nucleation," Phys. Rev. Lett. **110**, 058101 (2013).

[40] F. H. Stillinger, "Rigorous basis of the Frenkel-band theory of association equilibrium," J. Chem. Phys. **38**, 1486–1494 (1963).

[41] P. R. ten Wolde and D. Frenkel, "Computer simulation study of gas–liquid nucleation in a Lennard-Jones system," J. Chem. Phys. **109**, 9901–9918 (1998).

[42] A. Pelissetto and E. Vicari, "Critical phenomena and renormalization-group theory," Phys. Rep. **368**, 549–727 (2002).

Paper IV

Limitations of field-theory simulation for exploring phase separation: the role of repulsion in a lattice protein model

Daniel Nilsson,[1, a)] Behruz Bozorg,[1, b)] Sandipan Mohanty,[2, c)] Bo Söderberg,[1, d)] and Anders Irbäck[1, e)]

[1)] *Computational Biology & Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden.*

[2)] *Institute for Advanced Simulation, Jülich Supercomputing Centre, Forschungszentrum Jülich, D-52425 Jülich, Germany*

(Dated: 7 September 2021)

Field-theory simulation by the complex Langevin method offers an alternative to conventional sampling techniques for exploring the forces driving biomolecular liquid-liquid phase separation. Such simulations have recently been used to study several polyampholyte systems. Here, we formulate a field theory corresponding to the hydrophobic/polar HP lattice protein model, with finite same-site repulsion and nearest-neighbor attraction between HH bead pairs. By direct comparison with particle-based Monte Carlo simulations, we show that complex Langevin sampling of the field theory reproduces the thermodynamic properties of the HP model only if the same-site repulsion is not too strong. Unfortunately, the repulsion has to be taken weaker than what is needed to prevent condensed droplets from assuming an artificially compact shape. Analysis of a minimal and analytically solvable toy model hints that the sampling problems caused by repulsive interaction may stem from a loss of ergodicity.

[a)] Electronic mail: daniel.nilsson@thep.lu.se.

[b)] Electronic mail: behruz.bozorg@thep.lu.se.

[c)] Electronic mail: s.mohanty@fz-juelich.de.

[d)] Electronic mail: bo.soderberg@thep.lu.se.

[e)] Electronic mail: anders.irback@thep.lu.se.

I. INTRODUCTION

Advances over the past 15 years have identified liquid-liquid phase separation (LLPS) as a driver of compartmentalization in living cells[1,2]. Through LLPS, membraneless droplets are formed, with high concentrations of proteins and nucleic acids. In this process, it has been found that intrinsically disordered proteins (IDPs) often play a key role, and several such IDPs have been shown to phase separate on their own[3–5].

To gain insight into the forces driving IDP LLPS, a broad set of theoretical and computational methods have been employed. The Flory-Huggins[6,7] and Voorn-Overbeek[8] mean-field methods provide useful analytical estimates, which, however, are insensitive to the ordering of the amino acids along the protein chains. By using the random phase approximation[9,10], the sequence dependence of polyampholytes can be explored without resorting to extensive simulations, at the price of assuming Gaussian chains. To be able to avoid approximations made in the above methods, there have also been many studies of biomolecular LLPS based on explicit-chain simulation[11–18]. In particular, using various coarse-grained models, the sequence determinants of polyampholyte LLPS were elucidated[11–15]. However, particle-based simulation (PBS), with explicit chains, becomes computationally expensive for large systems, even with coarse-grained models.

Another approach to dense polymer systems is to use field-theory simulation (FTS)[19], which has recently been applied for the first time to biomolecular LLPS[20]. Here, by means of a Hubbard-Stratonovich transformation, the original polymer system is reformulated as a statistical field theory, which can be investigated by simulation[19,20]. This approach has the advantage of removing direct interchain interactions, which makes it, at least formally, easy to increase the number of chains in the simulations. A disadvantage is that the effective energy of the field theory is complex-valued, which renders standard sampling techniques inadequate. A potential solution to this problem is offered by the complex Langevin method[21–23]. Indeed, using this method, several investigations of biomolecular LLPS in both one- and two-component systems have been reported[20,24–28]. In all these systems, phase separation was driven by Coulomb interactions, which are well suited for field-theoretic treatment.

In this article, we test the FTS approach on a hydrophobic/polar (HP) protein model, where phase separation is driven by short-range hydrophobic attraction rather than electrostatics. Since the FTS method requires the introduction of an auxiliary spatial grid, we deliberately consider a lattice-based protein model. In this way, it becomes possible to compare FTS and PBS results in a direct fashion, without having to extrapolate FTS results to the limit of vanishing lattice spacing. Specifically, we consider a variant of the well-known HP lattice model for protein folding[29], with a finite same-site repulsion strength, Λ. As will be shown below, this model can be mapped onto a field theory with a simple structure. Note that with a lattice-based protein model, particle densities are, by construction, smeared. With this implicit smearing present, there is no need for the explicit Gaussian smearing typically used in continuous models.

The strength of the same-site repulsion, Λ, is a critically important parameter. On physical grounds, Λ has to be sufficiently large to prevent condensed clusters from collapsing to an artificially compact shape. If, on the other hand, Λ is taken too large, it turns out that the complex Langevin method breaks down. To elucidate these two conflicting requirements, we simulate and analyze in some detail a lattice gas, consisting of H particles rather than HP chains. Comparing FTS and PBS data, we find that the two Λ regions where the respective requirements are met, unfortunately, do not overlap. To determine whether the sampling problems that we observe at large Λ is a peculiarity of our particular model or a more general problem associated with repulsive interactions, we construct a minimal toy model, which can be solved analytically. In the presence of strong repulsive interactions, we find that the complex Langevin method fails in this toy model as well. Finally, we present some examples of FTS results for systems of HP chains, which, again, are compared with PBS data for the same systems.

II. METHODS

A. Biophysical model

We consider a system of N linear chains with M beads each, on a simple cubic lattice with volume V and periodic boundaries in all three directions. The beads can be either hydrophobic (H) or polar (P). For simplicity, we assume that all N chains share the same sequence, which we write as $\sigma = (\sigma_1, \ldots, \sigma_M)$, where $\sigma_m = 1$ for an H bead and $\sigma_m = 0$ for a P bead. Throughout the paper, we use dimensionless values for energy and length, with the lattice spacing set to unity.

The interaction potential is pairwise additive, $U = \sum_{i<j} u_{ij}$, where the sum runs over all pairs of beads, both intra- and interchain pairs. The pair potential u_{ij} has a repulsive part, which assigns an energy penalty $\Lambda > 0$ to any pair of beads residing on the same lattice site. In addition, there is an attractive nearest-neighbor interaction, which is felt only by HH pairs. In total, thus, the pair potential is given by

$$u_{ij} = \begin{cases} \Lambda\,, & \text{if beads } i \text{ and } j \text{ are on the same site;} \\ -\sigma_i\sigma_j\,, & \text{if beads } i \text{ and } j \text{ are nearest neighbors;} \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{1}$$

The full potential may thus be written as

$$U = U_{\mathrm{r}} + U_{\mathrm{a}}\,, \tag{2}$$

where $U_{\mathrm{r}} = \Lambda \times \{\text{number of same-site pairs}\}$ and $U_{\mathrm{a}} = -\{\text{number of nearest-neighbor HH pairs}\}$.

The thermodynamic behavior of the system at inverse temperature β is determined by the partition function

$$Z = \sum_C e^{-\beta U}\,, \tag{3}$$

where the sum runs over all possible configurations C of the N-chain system.

To obtain a field theory representation of this particle-based system, we first express the potential U in terms of bead counts rather than bead positions. To this end, we consider the ansatz

$$U_{\mathrm{bc}} = \frac{\Lambda}{2}\sum_{\mathbf{r}} n(\mathbf{r})^2 + \frac{1}{2}\sum_{\mathbf{r},\hat{\mu}} \tilde{n}_{\mathrm{H}}(\mathbf{r},\hat{\mu})^2\,, \quad \text{with} \quad \tilde{n}_{\mathrm{H}} = \alpha n_{\mathrm{H}}(\mathbf{r}+\hat{\mu}) - \alpha^* n_{\mathrm{H}}(\mathbf{r})\,, \tag{4}$$

where $\mathbf{r}$ denotes a lattice site, $\hat{\mu}$ is one of three lattice unit vectors, α is a complex parameter, $n(\mathbf{r})$ is the total number of beads at site $\mathbf{r}$, and $n_{\mathrm{H}}(\mathbf{r})$ is the number of H beads at site $\mathbf{r}$. The first sum is over lattice sites, while the second is over links. Note that the second sum does not have the common form $\sum_{\mathbf{r}}[\sum_{\mathbf{r}'}\Gamma(\mathbf{r},\mathbf{r}')n_{\mathrm{H}}(\mathbf{r}')]^2$, where Γ is a smearing matrix and the outer sum is over sites rather than links. The link-based form in Eq. 4 makes it possible to avoid interactions beyond nearest-neighbor distance.

The bead counts n and n_{H} can be written as

$$n(\mathbf{r}) = \sum_i \delta(\mathbf{r},\mathbf{r}_i) \qquad \text{and} \qquad n_{\mathrm{H}}(\mathbf{r}) = \sum_i \sigma_i\delta(\mathbf{r},\mathbf{r}_i)\,, \tag{5}$$

where $\mathbf{r}_i$ and σ_i denote, respectively, the location and type of bead i, and δ is a Kronecker delta. Using Eq. 5, U_{bc} (Eq. 4) can be rewritten as a sum over bead pairs. One finds that the n-dependent part of U_{bc} is equal to U_{r} plus a constant self-energy term, given by $NM\Lambda/2$. The $\tilde{n}_{\mathrm{H}}$-dependent part of U_{bc} generally contains both same-site and nearest-neighbor interactions. However, this mixing can be avoided by choosing the parameter $\alpha = e^{i\pi/4}$. With this α, this part of U_{bc} becomes equal to U_{a}, which means that

$$U_{\mathrm{bc}} = U + \frac{NM\Lambda}{2}. \tag{6}$$

All numerical results presented below were obtained using $\alpha = e^{i\pi/4}$.

B. Field theory

It follows from the above that, adding the constant self-energy term to U, the partition function (Eq. 3) can be expressed as

$$Z = \sum_{C} e^{-\beta U_{\mathrm{bc}}}, \tag{7}$$

where U_{bc} (Eq. 4) depends quadratically on both the site variables $n(\mathbf{r})$ and the link variables $\tilde{n}_{\mathrm{H}}(\mathbf{r},\hat{\mu})$. These quadratic dependencies can be linearized by introducing auxiliary fields, $w(\mathbf{r})$ and $\varphi(\mathbf{r},\hat{\mu})$, by means of the Hubbard-Stratonovich method. Note that since $\tilde{n}_{\mathrm{H}}(\mathbf{r},\hat{\mu})$ is associated with links, the corresponding field $\varphi(\mathbf{r},\hat{\mu})$ can be seen as a discrete version of a vector field, living on the links of the lattice.

Specifically, the fields are introduced through the relations

$$\begin{aligned} \exp\left(-\frac{\beta\Lambda}{2} n(\mathbf{r})^2\right) &\propto \int dw(\mathbf{r}) \exp\left(-\frac{1}{2\beta\Lambda} w(\mathbf{r})^2 - iw(\mathbf{r})n(\mathbf{r})\right), \\ \exp\left(-\frac{\beta}{2}\tilde{n}_{\mathrm{H}}(\mathbf{r},\hat{\mu})^2\right) &\propto \int d\varphi(\mathbf{r},\hat{\mu}) \exp\left(-\frac{1}{2\beta}\varphi(\mathbf{r},\hat{\mu})^2 - i\varphi(\mathbf{r},\hat{\mu})\tilde{n}_{\mathrm{H}}(\mathbf{r},\hat{\mu})\right). \end{aligned} \tag{8}$$

This yields the partition function

$$Z \propto Z_{\mathrm{FT}} = \int \prod_{\mathbf{r}} dw(\mathbf{r}) \prod_{\mathbf{r},\hat{\mu}} d\varphi(\mathbf{r},\hat{\mu}) e^{-H[w,\varphi]}, \tag{9}$$

where the effective energy $H[w,\varphi]$ is given by

$$H[w,\varphi] = \frac{1}{2\nu}\sum_{\mathbf{r}} w(\mathbf{r})^2 + \frac{1}{2\eta}\sum_{\mathbf{r},\hat{\mu}} \varphi(\mathbf{r},\hat{\mu})^2 - N\ln Q[w,\varphi], \tag{10}$$

with $\nu = \beta\Lambda$ and $\eta = \beta$. Here, $Q[w,\varphi]$ is a conditional single-chain partition function, given by

$$Q[w,\varphi] = \sum_{C_1} \exp\left[-i\sum_{m=1}^{M}\left(w(\mathbf{r}_m) + \sigma_m \sum_{\hat{\mu}}[\alpha\varphi(\mathbf{r}_m-\hat{\mu},\hat{\mu}) - \alpha^*\varphi(\mathbf{r}_m,\hat{\mu})]\right)\right], \tag{11}$$

where the outer sum is over single-chain configurations $C_1 = (\mathbf{r}_1,\ldots,\mathbf{r}_M)$, corresponding to a random walk on the cubic lattice with $M-1$ unit steps. The evaluation of Q, given w and φ, can be conveniently organized by rewriting Eq. 11 in the form

$$Q[w,\varphi] = \sum_{C_1} \prod_{m=1}^{M} \chi_{\sigma_m}(\mathbf{r}_m) = \sum_{\mathbf{r}_1,\ldots,\mathbf{r}_M} \chi_{\sigma_M}(\mathbf{r}_M)\ldots T(\mathbf{r}_3,\mathbf{r}_2)\chi_{\sigma_2}(\mathbf{r}_2)T(\mathbf{r}_2,\mathbf{r}_1)\chi_{\sigma_1}(\mathbf{r}_1)\,, \tag{12}$$

where $T(\mathbf{r},\mathbf{r}') = 1$ if $\mathbf{r}$ and $\mathbf{r}'$ are nearest neighbors, $T(\mathbf{r},\mathbf{r}') = 0$ otherwise, and

$$\chi_\sigma(\mathbf{r}) = \begin{cases} e^{-iw(\mathbf{r})}\,, & \text{if } \sigma = 0\,; \\ e^{-iw(\mathbf{r})-i\sum_{\hat{\mu}}[\alpha\varphi(\mathbf{r}-\hat{\mu},\hat{\mu})-\alpha^*\varphi(\mathbf{r},\hat{\mu})]}\,, & \text{if } \sigma = 1\,. \end{cases} \tag{13}$$

C. Extracting polymer properties from the fields

In the field representation, the original bead count variables are not readily available, but hidden in the conditional partition function Q. However, it is possible to derive useful identities between bead count and field correlations[19]. A whole series of such identities can be derived by noting that the fields $w(\mathbf{r})$ and $\varphi(\mathbf{r},\hat{\mu})$ (Eq. 8) can be expressed as

$$\begin{aligned} w(\mathbf{r}) &= u(\mathbf{r}) - i\nu n(\mathbf{r})\,, \\ \varphi(\mathbf{r},\hat{\mu}) &= u_{\mathrm{H}}(\mathbf{r},\hat{\mu}) - i\eta\tilde{n}_{\mathrm{H}}(\mathbf{r},\hat{\mu})\,, \end{aligned} \tag{14}$$

where $u(\mathbf{r})$ and $u_{\mathrm{H}}(\mathbf{r},\hat{\mu})$ are auxiliary zero-mean Gaussian fields with $\langle u(\mathbf{r})u(\mathbf{r}')\rangle = \nu\,\delta(\mathbf{r},\mathbf{r}')$ and $\langle u_{\mathrm{H}}(\mathbf{r},\hat{\mu})u_{\mathrm{H}}(\mathbf{r}',\hat{\mu}')\rangle = \eta\,\delta(\mathbf{r},\mathbf{r}')\delta(\hat{\mu},\hat{\mu}')$. At the one- and two-point levels, one finds the identities

$$\begin{aligned} \langle w(\mathbf{r})\rangle &= -i\nu\langle n(\mathbf{r})\rangle \qquad (= -i\nu NM/V)\,, \\ \langle \varphi(\mathbf{r},\hat{\mu})\rangle &= -i\eta\langle\tilde{n}_{\mathrm{H}}(\mathbf{r},\hat{\mu})\rangle \qquad (= 2\eta\,\mathrm{Im}(\alpha)NM_{\mathrm{H}}/V)\,, \\ \langle w(\mathbf{r})w(\mathbf{r}')\rangle &= \nu\delta(\mathbf{r},\mathbf{r}') - \nu^2\langle n(\mathbf{r})n(\mathbf{r}')\rangle\,, \\ \langle \varphi(\mathbf{r},\hat{\mu})\varphi(\mathbf{r}',\hat{\mu}')\rangle &= \eta\delta(\mathbf{r},\mathbf{r}')\delta(\hat{\mu},\hat{\mu}') - \eta^2\langle\tilde{n}_{\mathrm{H}}(\mathbf{r},\hat{\mu})\tilde{n}_{\mathrm{H}}(\mathbf{r}',\hat{\mu}')\rangle\,, \end{aligned} \tag{15}$$

where M_{H} denotes the number of H beads per chain. From the last two of these identities, it follows that

$$\begin{aligned} \langle U_{\mathrm{r}}\rangle + \frac{NM\Lambda}{2} &= \frac{V}{2\beta} - \frac{1}{2\beta^2\Lambda}\sum_{\mathbf{r}}\langle w(\mathbf{r})^2\rangle\,, \\ \langle U_{\mathrm{a}}\rangle &= \frac{3V}{2\beta} - \frac{1}{2\beta^2}\sum_{\mathbf{r},\hat{\mu}}\langle\varphi(\mathbf{r},\hat{\mu})^2\rangle\,. \end{aligned} \tag{16}$$

The average particle-based total energy $U = U_{\mathrm{r}} + U_{\mathrm{a}}$ (Eq. 3) therefore can be obtained as the field-theory average of the estimator

$$U_{\mathrm{FT}} = \frac{2V}{\beta} - \frac{NM\Lambda}{2} - \frac{1}{2\beta^2}\left[\frac{1}{\Lambda}\sum_{\mathbf{r}} w(\mathbf{r})^2 + \sum_{\mathbf{r},\hat{\mu}}\varphi(\mathbf{r},\hat{\mu})^2\right]\,. \tag{17}$$

When studying phase separation, a common choice is to use elongated simulation boxes, with volume $V = L_z L^2$, in which droplets tend to be slab-like rather than spherical. Droplets can then be detected by determining the density profile $\rho(z) = L^{-2}\sum_{x,y} n(x,y,z)$. The simultaneous presence of two bulk phases with different densities leads to a large spatial variance of $\rho(z)$, defined as

$$\sigma_\rho^2 = \frac{1}{L_z - 1}\sum_{z=1}^{L_z}(\rho(z) - \overline{\rho})^2 = \frac{1}{L_z - 1}\left(\sum_{z=1}^{L_z}\rho(z)^2 - L_z\overline{\rho}^2\right), \tag{18}$$

where $\overline{\rho} = NM/V$ denotes total density. Using Eq. 15, it can be easily verified that the ensemble average of this quantity, $\langle\sigma_\rho^2\rangle$, can be determined by using the field-theoretic estimator

$$\sigma_{\rho,\mathrm{FT}}^2 = \frac{1}{L_z - 1}\left(\frac{L_z}{L^2\beta\Lambda} - L_z\overline{\rho}^2 - \frac{1}{(\beta\Lambda)^2}\sum_z \rho_w(z)^2\right), \tag{19}$$

where $\rho_w(z) = L^{-2}\sum_{x,y} w(x,y,z)$.

D. Complex Langevin sampling

The statistical field theory defined by Eq. 9 has a complex-valued effective energy $H[w,\varphi]$, and therefore a complex weight function e^{-H}, which renders sampling techniques such as Markov chain Monte Carlo inadequate. In principle, this problem can be overcome by sampling the distribution $e^{-\mathrm{Re}H}$ and using reweighting methods. However, this approach typically requires estimating rapidly fluctuating observables, which makes it inefficient. A potentially useful alternative is to use Langevin dynamics[21–23], defined by

$$\begin{aligned}\dot{w}(\mathbf{r}) &= -\frac{\partial H}{\partial w(\mathbf{r})} + \sqrt{2}\,\Xi_w(\mathbf{r},t),\\ \dot{\varphi}(\mathbf{r},\hat{\mu}) &= -\frac{\partial H}{\partial \varphi(\mathbf{r},\hat{\mu})} + \sqrt{2}\,\Xi_\varphi(\mathbf{r},\hat{\mu},t),\end{aligned} \tag{20}$$

where t is Langevin time, a dot indicates time derivative, Ξ_w is standard Gaussian noise with zero mean and correlations given by $\langle\Xi_w(\mathbf{r},t)\Xi_w(\mathbf{r}',t')\rangle = \delta(\mathbf{r},\mathbf{r}')\delta(t-t')$, and similarly for Ξ_φ. In a simulation, these continuous-time equations have to be discretized. A simple discrete form is

$$\begin{aligned}w(\mathbf{r})_{k+1} &= w(\mathbf{r})_k - dt\,\frac{\partial H}{\partial w(\mathbf{r})}\Big|_k + \sqrt{2dt}\,\xi_w(\mathbf{r},t_k)\\ &= (1-\nu\,dt)\,w(\mathbf{r})_k + dt\,\frac{N}{Q}\frac{\partial Q}{\partial w(\mathbf{r})}\Big|_k + \sqrt{2dt}\,\xi_w(\mathbf{r},t_k),\\ \varphi(\mathbf{r},\hat{\mu})_{k+1} &= \varphi(\mathbf{r},\hat{\mu})_k - dt\,\frac{\partial H}{\partial \varphi(\mathbf{r},\hat{\mu})}\Big|_k + \sqrt{2dt}\,\xi_\varphi(\mathbf{r},\hat{\mu},t_k)\\ &= (1-\eta\,dt)\,\varphi(\mathbf{r},\hat{\mu})_k + dt\,\frac{N}{Q}\frac{\partial Q}{\partial \varphi(\mathbf{r},\hat{\mu})}\Big|_k + \sqrt{2dt}\,\xi_\varphi(\mathbf{r},\hat{\mu},t_k),\end{aligned} \tag{21}$$

where dt is the time step and k a time index, while $\xi_w(\mathbf{r}, t_k)$ and $\xi_\varphi(\mathbf{r}, \hat{\mu}, t_k)$ are two sets of independent Gaussian random variables with zero mean and unit variance. In Eq. 21, it is possible and potentially advantageous to use different time steps for the w and φ fields, dt_w and dt_φ, depending on ν and η[30,31]. However, throughout this paper, we use the same dt for all degrees of freedom.

Due to the complex nature of H, the fields will not be restricted to real values when evolving according to Eq. 21, but will wander off into the complex plane. Thus, we will have a probability distribution over complex-valued fields, or, equivalently, a joint probability distribution over their real and imaginary parts. Under fairly general conditions, the Langevin dynamics allows for this distribution over complex fields to converge to one that mimics the formal, complex-valued Boltzmann distribution over real fields, e^{-H}, in the sense that expectation values of *analytic* functions of the fields will converge to the correct values. However, it is well-known that the success of the method is system-dependent[32,33].

E. Simulation details

We test the FTS method on systems consisting of H particles or multiple copies of one of two different 10-bead HP chains. For comparison, we apply PBS techniques to the same systems, to generate reference data.

The FTS results are time averages over Langevin trajectories, generated using Eq. 21 with a fixed step size in the range $5 \times 10^{-6} \leq dt \leq 10^{-4}$. The simulations are started from randomly perturbed uniform field configurations. Each run covers a total Langevin time of 4×10^3 (H particles) or 5×10^3 (HP chains), the first 20% of which is discarded for thermalization.

The PBS results are obtained using Monte Carlo methods. For H particles, a single type of move is employed, namely displacement of individual particles to nearest-neighbor sites on the lattice. A majority of the results are from fixed-temperature simulations with the Metropolis algorithm. However, near the condensation/evaporation transition, this sampling method becomes inefficient, because transitions between states with and without a droplet are rare. To overcome this problem, some of our simulations use the Wang-Landau algorithm[34,35], which, in particular, facilitates the determination of the condensation/evaporation temperature.

The PBS results for HP chains are based on a set of three elementary moves. The first move alters the internal structure of a random chain, by rotating one of its $M-1$ bond vectors. The second move is a rigid-body translation or rotation of an individual chain. The third and final move is a rigid-body translation of a cluster of chains. The construction of the cluster to be moved is stochastic, following a Swendsen-Wang type procedure[36,37]. All three moves are subject to a Metropolis accept/reject test.

III. RESULTS

Above we gave a field-theoretic representation (Sec. II B) of the HP lattice protein model with finite same-site repulsion (Sec. II A). In this section, we evaluate to what extent simulation of this field theory by the complex Langevin method (Sec. II D) reproduces the thermodynamic properties of the HP model, using reference data obtained by conventional particle-based Monte Carlo simulation. First, we investigate in some detail the case of a lattice gas, where the system consists of (one-bead) H particles. To shed some light upon the findings for the lattice gas, we then introduce a minimal, analytically solvable toy model, whose behavior under Langevin dynamics can be analyzed and understood. Finally, we present some results from simulations with 10-bead HP chains.

A. H particles

Throughout this subsection, we consider systems consisting of 64 H particles at fixed density $\rho = N/V = 0.125$. The lattice used is either cubic (8^3) or elongated in one direction (32×4^2). In the latter case, condensed droplets assume a slab-like rather than spherical shape. We study how the ability of the FTS method to reproduce results obtained using conventional PBS techniques depends on the two parameters of the model, the repulsion strength Λ and the inverse temperature β. Another important issue is how large Λ has to be taken in order to prevent condensed droplets from becoming artificially compact.

Figures 1A,C compare FTS and PBS data for the repulsive and attractive energies U_r and U_a, respectively, for different Λ at fixed $\beta = 0.3$, on a cubic lattice. At this β, the system is in an uncondensed gas state for all Λ values considered. The FTS results are in agreement with the PBS data for $\Lambda \lesssim 3$, but deviations develop as Λ is increased. That the FTS method suffers from sampling errors at large Λ is underscored by the fact that the, by definition, positive quantity U_r turns negative.

The corresponding data at $\beta = 0.4$ follow a similar pattern (Figs. 1B,D), although the accuracy

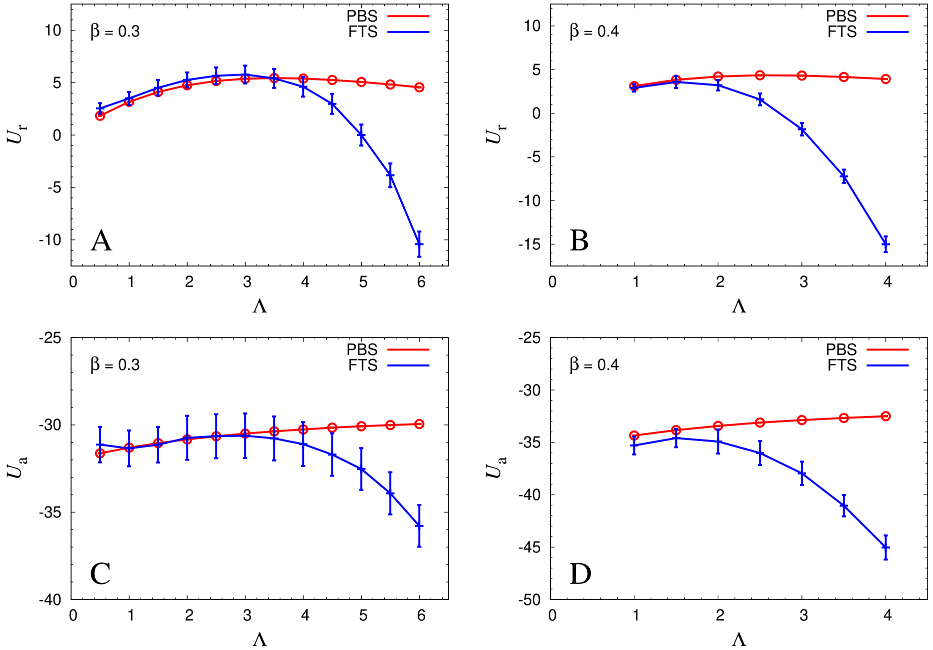


FIG. 1. Λ-dependence of the repulsion and attraction energies U_{r} and U_{a} at $\beta = 0.3$ and $\beta = 0.4$, for a system of 64 H particles on an 8^3 lattice, as obtained using PBS (red) and FTS (blue). Lines are drawn to guide the eye. (A) U_{r} at $\beta = 0.3$ (B) U_{r} at $\beta = 0.4$. (C) U_{a} at $\beta = 0.3$, (D) U_{a} at $\beta = 0.4$.

of the FTS method starts deteriorating at a lower Λ in this case. At $\beta = 0.4$, we omitted data obtained for $\Lambda = 0.5$. The reason for this is that a condensation transition takes place as Λ is reduced from 1 to 0.5, which leads to U_{r} and U_{a} values far outside the plotted ranges.

For fixed $\Lambda = 0.5$, the above results imply that a temperature-induced condensation transition occurs as β is increased from 0.3 to 0.4. This transition is illustrated in Fig. 2, which shows the β-dependence of the total energy $U = U_{\mathrm{r}} + U_{\mathrm{a}}$. By PBS, we estimate that the condensation transition occurs at $\beta_{\mathrm{t}} \approx 0.375$, with β_{t} defined by having the maximum heat capacity. The curve representing PBS data in Fig. 2 is computed by using the Wang-Landau algorithm[34,35], along with reweighting techniques. This algorithm turns out to be much more efficient than standard constant-temperature Monte Carlo, which becomes very slow in the vicinity of the condensation transition. This slowdown can be linked to a strongly bimodal energy distribution, $P(U)$ (Fig. 3A), where the two peaks correspond to states with and without a droplet, respectively. At the transition

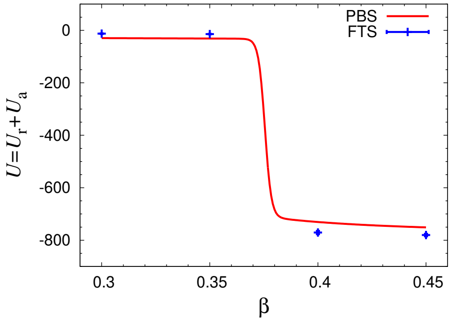


FIG. 2. β-dependence of the total energy $U = U_r + U_a$ for fixed $\Lambda = 0.5$ in a system of 64 H particles on an 8^3 lattice. The red curve represents PBS data, obtained with the Wang-Landau algorithm[34,35]. Blue symbols indicate FTS results.

temperature, the valley between the two peaks is statistically suppressed by about eight orders of magnitude, despite the modest size of the system.

We now turn to the FTS method, which, for this Λ, captures the transition quite well. Figure 2 shows FTS results at four different β, two on each side of the transition. All four data points fall close to the curve obtained by PBS. It is worth noting that all the FTS results are from runs started from random initial field configurations. Therefore, for $\beta > \beta_t$, the Langevin dynamics has to bring the system from a random state to field configurations corresponding to a droplet-containing state. Figure 3B shows the time evolution of the field-theoretic energy estimator U_{FT} (Eq. 17) in a run at $\beta = 0.4 \approx 1.07\beta_t$. The system initially spends a period of time in a high-U_{FT} state, followed by a sudden jump to a low-U_{FT} state. This behavior matches well with PBS data for $P(U)$ at $\beta = 0.4$ (Fig. 3A). Here, although the low-energy peak dominates, the high-energy peak is still present. Consistent with this, in the FTS run, there is a waiting time before the system escapes from the initial high-U_{FT} state (Fig. 3B).

Next, we investigate the droplet condensation transition in some more detail for three values of Λ (0.5, 2.0, 5.0), using an elongated simulation box (32×4^2). To this end, we first consider the longitudinal distribution of particles, $\rho(z) = L^{-2}\sum_{x,y} n(x,y,z)$. In particular, we compute the spatial variance of this distribution, σ_ρ^2 (Eq. 18), and its field-theoretic estimator, $\sigma_{\rho,\mathrm{FT}}^2$ (Eq. 19). The formation of a dense droplet in a dilute backgrund leads to an increased spatial variance σ_ρ^2. Figure 4A shows PBS and FTS data for σ_ρ^2 and $\sigma_{\rho,\mathrm{FT}}^2$, respectively, in the vicinity of the inverse

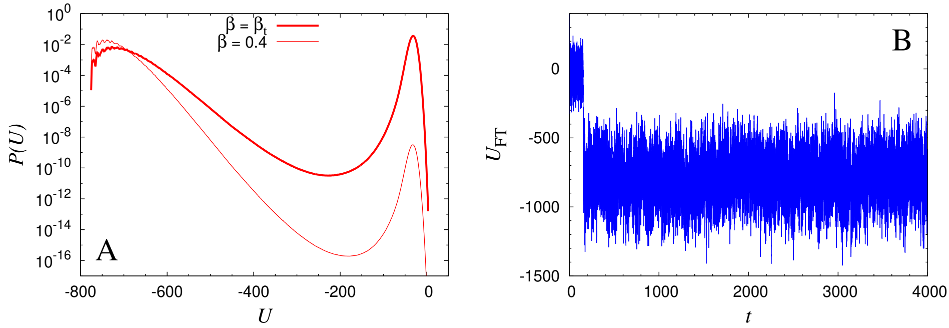


FIG. 3. Droplet condensation in a system of 64 H particles on an 8^3 lattice for $\Lambda = 0.5$. (A) Energy distribution, $P(U)$, in logscale at $\beta = 0.375 \approx \beta_\mathrm{t}$ (thick line) and $\beta = 0.4$ (thin line), based on PBS data obtained with the Wang-Landau algorithm[34,35]. (B) Time evolution of the field-theoretic estimator U_FT of U (Eq. 17) in an FTS run at $\beta = 0.4$.

transition temperature, β_t, for all three choices of Λ. As in Fig. 2, the FTS data agree quite well with the PBS data for $\Lambda = 0.5$. By contrast, but not surprisingly given the data in Fig. 1, the FTS method fails to properly describe the condensation transition for $\Lambda = 2.0$ and $\Lambda = 5.0$.

The change in σ_ρ^2 near β_t is abrupt and large for $\Lambda = 0.5$, while becoming less drastic as Λ is increased (Fig. 4A). The abruptness of the transition for small Λ is linked to the collapse of condensed droplets, which leads to artificially low energies for droplet-containing configurations. Figure 4B shows the total number of lattice sites hosting at least one of the 64 particles in the system. For $\Lambda = 0.5$, it can be seen that this number, n_s, drops from ≈ 60 to ≈ 10 upon droplet condensation. By contrast, for $\Lambda = 5.0$, n_s stays above 63 throughout the β range studied, $0.93 \leq \beta/\beta_\mathrm{t} \leq 1.07$.

In summary, in the lattice gas studied here, in order for the FTS sampling errors to stay small, the repulsion strength Λ must not be too large. At the same time, in order to prevent the formation of artificially compact droplets, Λ must not be too small. Unfortunately, at least with the standard Langevin scheme used here, the Λ regions where these two requirements are met do not overlap, as is illustrated by the results for $\Lambda = 2.0$ in Fig. 4. This Λ is too large to avoid large sampling errors (Fig. 4A), but still too small to prevent condensed droplets from collapsing (Fig. 4B).

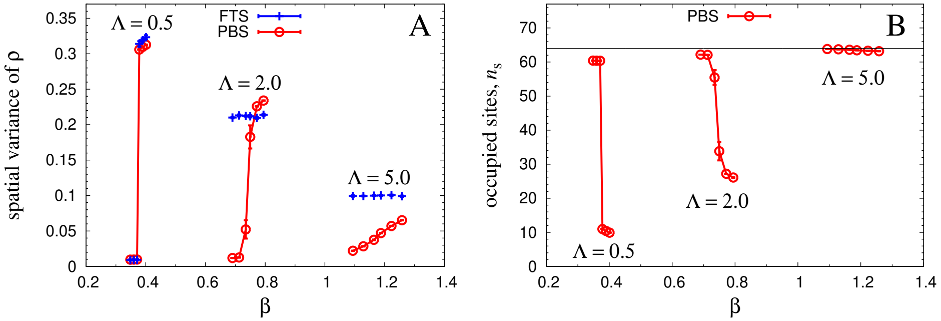


FIG. 4. Droplet condensation in a system of 64 H particles on a 32×4^2 lattice for three values of Λ (0.5, 2.0, 5.0), studied using PBS (red) and FTS (blue). For each Λ, data were acquired for six values of β/β_t (0.93, 0.96, 0.99, 1.01, 1.04, 1.07), where β_t is the inverse transition temperature, defined as the heat capacity maximum. Lines are drawn to guide the eye. (A) Spatial variance of the density $\rho(z)$, calculated using Eq. 18 (PBS) or Eq. 19 (FTS). (B) The number of lattice sites hosting at least one particle, n_s, with its maximal value (64) indicated by the horizontal line.

B. Toy model

In this subsection, we turn to a minimal toy model, to elucidate how increased repulsion strength can cause sampling problems in the FTS approach.

Thus, we consider a single particle (or a gas of N identical ones) on a lattice with only two sites, labelled 1 and 2, respectively, with two possible types of same-site pair interactions, of an either repulsive or attractive nature. The repulsive interaction gives a penalty of $\nu \geq 0$ for each same-site pair, while the attractive one instead gives a reward $\eta \geq 0$, as expressed by the respective interaction energies

$$\begin{aligned} \beta U_\mathrm{r} &= \frac{\nu}{2}\left(n_1^2 + n_2^2\right), \\ \beta U_\mathrm{a} &= -\frac{\eta}{2}\left(n_1^2 + n_2^2\right). \end{aligned} \tag{22}$$

In a similar way as for the main model, these systems can be transformed into field theories, with the respective effective energies

$$\begin{aligned} H_\mathrm{r}(w_1, w_2) &= \frac{1}{2\nu}\left(w_1^2 + w_2^2\right) - N \log Q_\mathrm{r}, \\ H_\mathrm{a}(\varphi_1, \varphi_2) &= \frac{1}{2\eta}\left(\varphi_1^2 + \varphi_2^2\right) - N \log Q_\mathrm{a}. \end{aligned} \tag{23}$$

Here, w_1, w_2 and φ_1, φ_2 are site fields for the repulsive and attractive cases, respectively, while Q_r and Q_a are the conditional, single particle partition functions, given by $Q_\mathrm{r} = e^{-iw_1} + e^{-iw_2}$, $Q_\mathrm{a} = e^{-\varphi_1} + e^{-\varphi_2}$.

Conveniently, H in each case separates in terms of the sum and difference of the fields on the two sites, given by $W = w_1 + w_2$, $w = w_1 - w_2$ for the repulsive case, and $\Phi = \varphi_1 + \varphi_2$, $\varphi = \varphi_1 - \varphi_2$ for the attractive one. The quadratic terms then become $\left(W^2 + w^2\right)/4\nu$ and $\left(\Phi^2 + \varphi^2\right)/4\eta$, respectively, while the conditional partition functions factorize as, respectively, $Q_\mathrm{r} = e^{-iW/2}\left(e^{-iw/2} + e^{iw/2}\right)$ and $Q_\mathrm{a} = e^{-\Phi/2}\left(e^{-\varphi/2} + e^{\varphi/2}\right)$.

As a result, the summed fields, W or Φ, have quadratic effective energies,

$$\begin{aligned} h_\mathrm{r}(W) &= \frac{W^2}{4\nu} + iN\frac{W}{2}, \\ h_\mathrm{a}(\Phi) &= \frac{\Phi^2}{4\eta} + N\Phi, \end{aligned} \tag{24}$$

and become simple Gaussian variables with rather trivial Langevin dynamics. Neglecting these, we can focus on the non-trivial difference fields, w or φ, with the effective energies

$$\begin{aligned} H_\mathrm{r}(w) &= \frac{w^2}{4\nu} - N\log\cos\left(\frac{w}{2}\right), \\ H_\mathrm{a}(\varphi) &= \frac{\varphi^2}{4\eta} - N\log\cosh\left(\frac{\varphi}{2}\right). \end{aligned} \tag{25}$$

Applying conventional (complex) Langevin dynamics to the original fields leads to the following dynamics for the difference fields:

$$\begin{aligned} \dot{w} &= -\frac{w}{\nu} - N\tan\left(\frac{w}{2}\right) + 2\,\Xi_\mathrm{r}\,, \\ \dot{\varphi} &= -\frac{\varphi}{\eta} + N\tanh\left(\frac{\varphi}{2}\right) + 2\,\Xi_\mathrm{a}\,, \end{aligned} \tag{26}$$

where Ξ_r is a standard Gaussian noise with zero mean and $\langle \Xi_\mathrm{r}(t)\Xi_\mathrm{r}(t')\rangle = \delta(t-t')$, and similarly for Ξ_a. As before, the continuous-time evolution in Eq. 26 has to be approximated by discrete time equations.

As it turns out, the dynamics differs significantly between the two cases, and we will therefore consider them separately.

1. Repulsive case

For the repulsive case, the target distribution on the real w line, determined by H_{r} (Eq. 25), reads

$$P(w) \propto e^{-w^2/2\nu} \cos^N(w/2)\,, \tag{27}$$

which is real on the real line, but with a varying sign, at least for odd N, due to the cosine factor. Henceforth, we will assume $N = 1$. Figure 5 illustrates the drift in the complex w plane for two values of ν.

The zeros of the cosine at odd multiples of π define poles of H, at which the drift term in the Langevin equation for w diverges (Eq. 26). This leads to wild behavior, unless regulated, e.g., with a dynamical time step. Between the poles, the drift is smooth, and leaves the real line an invariant manifold that attracts the motion. On the real w line, the poles at odd multiples of π are repulsive under the drift, and alternate with attracting fixed points. The noise term, however, spreads out the trajectories.

Thus, it is clear that the real line acts as an attractor for the Langevin dynamics. In computer simulation with a finite time step, trajectories will be trapped on the real line, in the intervals between consecutive poles, and only occasionally pass to a neighboring interval. Within each interval, the resulting distribution will be proportional to $|P(w)|$, but with different random normalization constants in the different intervals, in a manner that depends on the particular simulation details.

Due to the Gaussian factor in P, for small enough repulsion strength, $\nu \ll 1$, the distribution is dominated by the central peak around the fixed point at $w = 0$, and the error in the Gaussian tail can be neglected. Hence, we would expect essentially correct long-term averages from computer simulations of the Langevin dynamics for small enough ν, while they would deteriorate for larger ν.

2. Attractive case

For the attractive case, on the other hand, the target distribution on the real φ line, determined by H_{a} (Eq. 25), reads

$$P(\varphi) \propto e^{-\varphi^2/2\eta} \cosh^N(\varphi/2)\,, \tag{28}$$

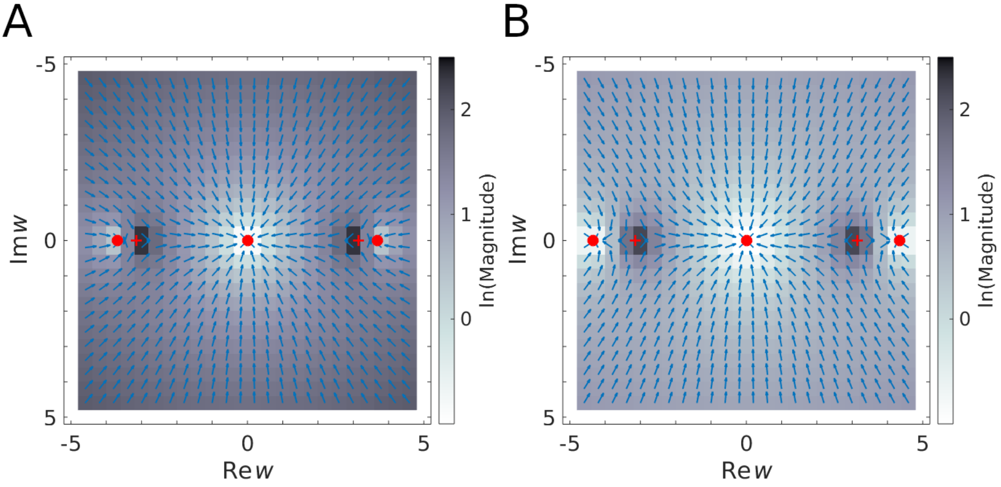


FIG. 5. Drift in the complex w plane of the repulsive toy model with $N = 1$, for (A) $\nu = 1$ and (B) $\nu = 3$. The arrows are normalized and indicate only the direction of the drift. The magnitude of the drift is indicated by the background color. Red symbols indicate attractive fixed points (filled circles) and poles (plus signs). The latter are repelling/attracting in the real/imaginary direction, respectively.

which is real and positive on the entire real line. For simplicity, we again focus on the case $N = 1$. Figure 6 illustrates the drift in the complex φ plane for two values of η.

In Eq. 28, the cosine of Eq. 27 is replaced by a cosh, which means that $P(\varphi)$ instead has zeroes on the imaginary axis. These zeroes again correspond to poles of H, but are less disturbing, being away from the real φ line. As in the w case, the real line is invariant, but the drift term now is smooth there (Eq. 26). However, the dynamics close to the real line depends on the size of the attraction strength η.

For $\eta < 2$, the real line is everywhere attracting, and the drift has a single attractive fixed point there, $\varphi = 0$, with a basin of attraction containing the whole real line. This indicates that a numerical simulation of the Langevin dynamics (Eq. 26) will result in long-term averages consistent with $P(\varphi)$.

At $\eta = 2$, the system undergoes a pitchfork bifurcation, where the central fixed point at $\varphi = 0$ turns unstable, while a previously repelling pair of fixed points on the imaginary line have closed in on the origin, and instead becomes a pair of attracting fixed points on the real line, on either

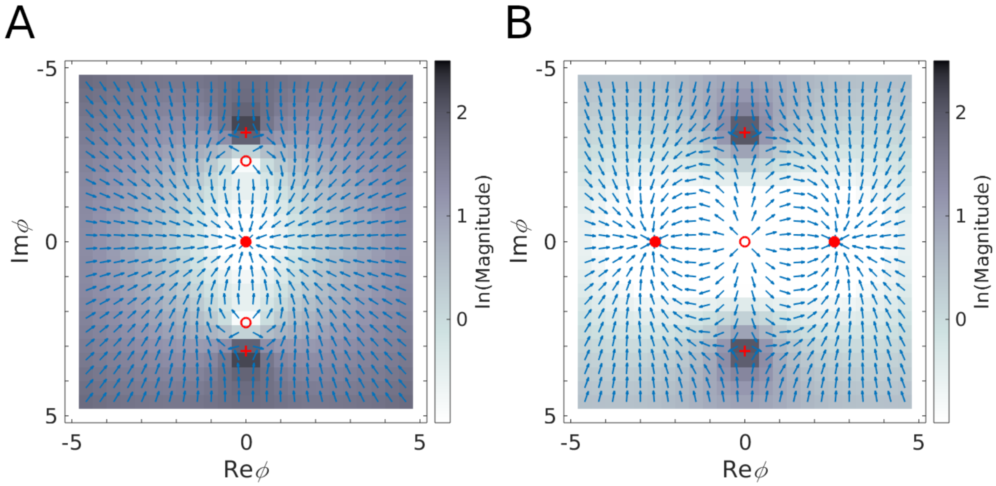


FIG. 6. Drift in the complex φ plane of the attractive toy model with $N = 1$, for (A) $\eta = 1$ and (B) $\eta = 3$. The arrows are normalized and indicate only the direction of the drift. The magnitude of the drift is indicated by the background color. Red symbols indicate attractive fixed points (filled circles), repulsive fixed points (open circles) and poles (plus signs). The latter are repelling/attracting in the real/imaginary direction, respectively.

side of the origin.

For $\eta > 2$, the real line is locally attracting only outside a pair of points lying inside the new attracting fixed points; however, a strip around the real line, $|\mathrm{Im}\,\varphi| < \pi$, is attracting. Within this strip there are no sampling barriers in the real direction, indicating that Langevin sampling may not suffer from the same problems as in the repulsive case.

3. *Numerical results and implications*

We have performed a set of simulations to probe the performance of the complex Langevin method for both the repulsive and the attractive toy model, using $N = 1$. Figure 7A shows the second moment of w for the repulsive model, as compared to the correct value $\langle w^2\rangle = 2\nu - \nu^2$. Likewise, Fig. 7B shows the second moment of φ for the attractive model, as compared to the correct value $\langle \varphi^2\rangle = 2\eta + \eta^2$. The simulations indeed confirm that the method significantly deteriorates in the repulsive case for $\nu \gtrsim 0.5$, while no noticeable deviation from the correct values is

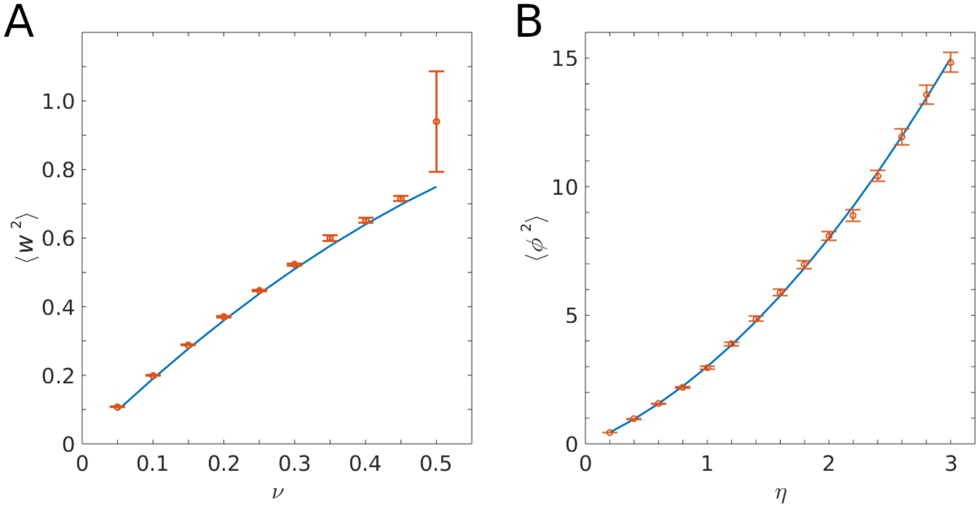


FIG. 7. Simulation data (red symbols) versus theoretical results (blue lines) for the toy model with $N = 1$. (A) Repulsive case: the second moment of w as a function of ν. Simulation data are well behaved for small ν, but deteriorates at $\nu \approx 0.5$. This erratic behavior will only become worse for higher ν values, and is conjectured to be due to loss of ergodicity. (B) Attractive case: the second moment of φ as a function of η. Simulation data follow the theoretical curve over the whole range.

seen in the attractive case.

This toy model illustrates how the Langevin dynamics yields correct results for an attractive pair interaction, but deteriorates for a strong enough repulsive one, due to a loss of ergodicity, in this case caused by poles on the real line. This behavior is qualitatively similar to what we observe in the larger model, where the complex Langevin dynamics fails to yield correct results when the repulsive part of the interaction is too large as compared to the attractive part.

The similarity in behavior suggests that also for the larger model, the problems might be due to loss of ergodicity. Note that the zeros of Q form pole manifolds of H with complex codimension one, corresponding to real codimension two. Normally, this should not jeopardize ergodicity. However, in case there exists an attractor in field space of real codimension one or more – like the real line in the repulsive toy model – ergodicity could be destroyed.

We speculate that the failure of the complex Langevin algorithm for strong repulsion might be due to a bifurcation to a situation with a codimension-one attractor, inside which the pole manifold may have a real codimension one. This would be enough to block trajectories and destroy ergodicity for the exact continuum version of the complex Langevin dynamics. In computer simulations with finite time step, trajectories may actually jump over the pole blockage, but in a way leading to erroneous probabilities.

C. HP chains

We now return to the HP lattice model, with repulsion strength Λ. In the lattice gas (Sec. III A), we saw that $\Lambda = 2$ was too small to prevent condensed droplets from collapsing, whereas this problem was significantly alleviated when using $\Lambda = 5$. In this subsection, we present simulation results for two 10-bead HP sequences, obtained with $\Lambda = 5$. We wish to explore how the FTS method performs when applied to chain systems at this Λ.

The two HP sequences considered are the alternating sequence $(\mathrm{HP})_5$, called A, and the block sequence $\mathrm{H}_5\mathrm{P}_5$, called B, which share the same composition. These two sequences have previously been studied using a coarse-grained continuous model[38,39], where cluster formation was found to set in at a higher temperature for sequence B than for sequence A. However, the clusters formed by sequence B were micelle-like, and therefore did not represent a bulk phase. By contrast, sequence A did phase separate[38,39].

Here, we consider systems consisting of 64 copies of either the A or the B sequence, on an elongated 36×12^2 grid. Figure 8 shows the β-dependence of the longitudinal bead density distribution, $\rho(z)$, using PBS data obtained with the Wang-Landau algorithm[34,35]. Here, before averaging over snapshots, the distribution $\rho(z)$ in a given snapshot is shifted, in such a way that if a single droplet is present, then its center of mass ends up close to the center of the box (in the z direction). From Fig. 8 it can be seen that cluster formation indeed sets in at a lower β for sequence B than for sequence A, as in previous work[38,39]. We estimate that $\beta_\mathrm{t} \approx 0.77 \pm 0.01$ for sequence B and $\beta_\mathrm{t} \approx 1.31 \pm 0.03$ for sequence A.

We test the FTS method using $\beta = 0.5$ and $\beta = 1.0$. Table I compares FTS data for the energy estimator U_FT (Eq. 17) with PBS data for the energy U. The β_t estimates above imply that a large cluster is present in only one of the four systems studied, namely for sequence B at $\beta = 1.0$. Therefore, the energy U is much lower in this system than in the other three. In all four cases, we find that the FTS method severely underestimates U, which is not unexpected given that $\Lambda = 5$ (cf. Fig. 1). Now, one may argue that the energy is a model-dependent quantity and therefore less

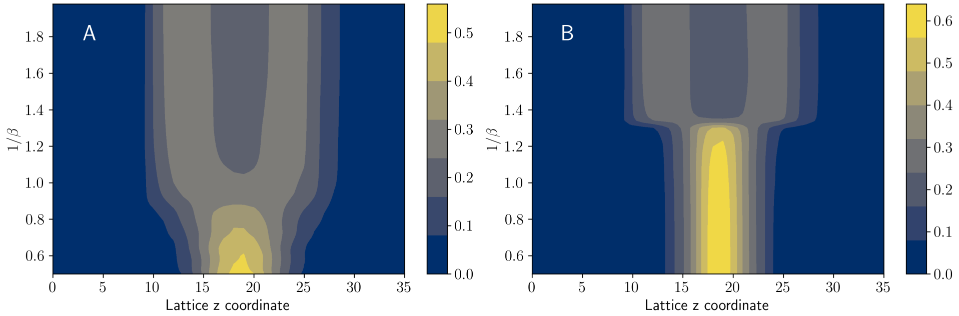


FIG. 8. Heat maps showing the temperature ($1/\beta$)-dependence of the bead density profile $\rho(z)$, for (A) sequence A and (B) sequence B. At a given β, $\rho(z)$ is either clearly unimodal, indicating the presence of a single dominant droplet, or weakly bimodal. In the latter case, the system tends to exhibit two main clusters. The data are from PBS simulations with the Wang-Landau algorithm[34,35]. The simulated systems consist of 64 chains on a 36×12^2 lattice, for $\Lambda = 5$.

interesting than basic structural properties, such as the presence or absence of large clusters.

In Fig. 9, we therefore also compare bead density profiles, $\rho(z)$, obtained using FTS and PBS, respectively, at the same two β values. As expected, the PBS profiles show that a large droplet is present for sequence B at $\beta = 1.0$, but not in any of the other three systems studied. In sharp contrast, the FTS data erroneously indicate that a large cluster is present in all four systems. Thus, in the systems studied, there is clear tendency for FTS sampling errors to cause a bias toward cluster formation. We also note that the maximal (averaged aligned) densities $\rho(z)$ from the FTS runs tend to be high, with values exceeding unity for sequence B. A value of unity corresponds to one bead per site.

TABLE I. Estimates of the energy U obtained with FTS and PBS for the 10-bead HP sequences A and B at $\beta = 0.5$ and $\beta = 1.0$, for $\Lambda = 5$ and 64 chains on a 36×12^2 lattice.

	Sequence A		Sequence B	
	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$
FTS	-4558 ± 16	-2992 ± 15	-7858 ± 14	-6253 ± 12
PBS	-102 ± 15	-12 ± 9	-549 ± 15	-22 ± 15

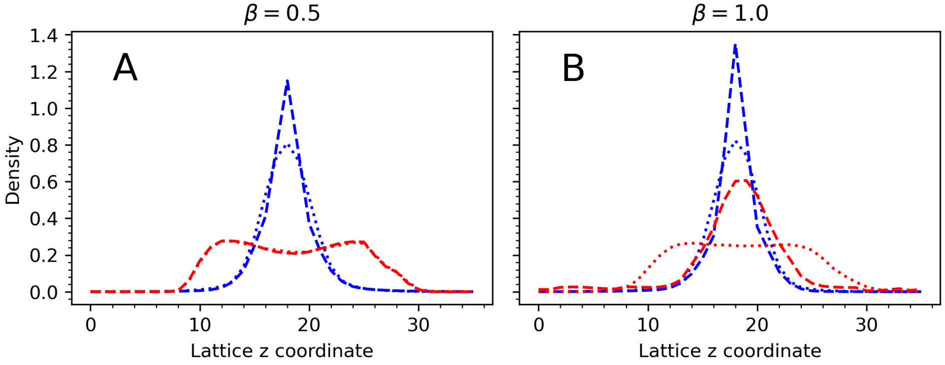


FIG. 9. Bead density profiles, $\rho(z)$, calculated using FTS (blue) and PBS (red) for the sequences A (dots) and B (dashes), for (A) $\beta = 0.5$ and (B) $\beta = 1.0$. The profiles are either clearly unimodal, indicating the presence of a single dominant droplet, or weakly bimodal. The latter systems tend to exhibit two main clusters. The FTS results for $\rho(z)$ are obtained using a field-theoretic estimator derived from Eq. 15. The simulated systems consist of 64 chains on a 36×12^2 lattice, for $\Lambda = 5$.

IV. DISCUSSION

FTS offers a new tool for investigating the mechanisms of biomolecular LLPS, with potential advantages over traditional PBS. The FTS approach has previously been used to investigate various systems where phase separation is driven by electrostatics[20,24–28]. In this paper, we have studied systems where phase separation is driven by short-range hydrophobic attraction. In preliminary work, we considered a continuous protein model similar to those in previous FTS studies[20,24–28], but with Coulomb interaction replaced by an effective attraction between hydrophobic beads. For simplicity, we decided, however, to focus on a lattice-based HP protein model, with finite same-site repulsion and nearest-neighbor attraction between HH pairs. We showed that this model can be mapped onto a field theory with a simple structure, by using an unconventional link-based form for the HH attraction. One advantage of choosing this lattice-based protein model is that FTS results can then be directly compared with PBS data, without having to extrapolate the FTS results to the continuum limit.

For a protein model to be amenable to standard FTS techniques, its excluded-volume repulsion has to be soft. However, if this repulsion is made too soft, one risks affecting the phase behavior[14,20,27]. A previous FTS study found that the excluded-volume strength affected the phase separation propensity in one-component polyampholyte systems[20], in line with theoretical results[40]. Using both FTS and PBS, another study found that the excluded-volume strength affected demixing in two-component polyampholyte systems[27].

In this paper, we have investigated the ability of the FTS method to accurately capture the thermodynamic behavior of the HP lattice protein model, which depends on the strength of the same-site repulsion, Λ. To this end, we examined in some detail the special case of a lattice gas, consisting of (one-bead) H particles. In particular, we asked whether, at a given Λ, the FTS method can provide accurate results at sufficiently large β to permit the study of droplet condensation. For this, Λ must not be too large. Unfortunately, we find that FTS accurately describes the droplet condensation transition only for Λ values that are still too small to prevent the droplets from becoming artificially compact (Fig. 4).

To get an idea of the origin and generality of the FTS sampling problems observed at large Λ in the lattice gas, we introduced a minimal two-site toy model with either attractive or repulsive interaction, which can be solved analytically. We find that this model can be simulated using the complex Langevin method if the interaction is attractive, whereas sampling problems arise if the interaction is repulsive and strong. We thus observe the same trends as in our lattice gas simulations. This similarity hints that the FTS sampling problems might be of the same nature for the lattice gas as in the toy model, where they can be linked to a loss of ergodicity.

Finally, we also presented results from some simulations of HP chains. Here, we wanted to explore the size and nature of the FTS sampling errors in chain systems with a significant same-site repulsion. We studied 64-chain systems for two 10-bead HP sequences, and and observed a clear tendency for FTS sampling errors to cause a bias toward droplet formation. In particular, using a temperature at which both systems should be free from large clusters, FTS data instead indicated the presence of a high-density droplet in both cases.

In our systems, we thus find that complex Langevin sampling fails when the repulsion is made sufficiently strong to prevent condensed droplets from assuming an artificially compact shape. It should be remembered, however, that we have in this paper limited ourselves to a simple standard implementation of complex Langevin dynamics. Furthermore, the gap between the two Λ regions with acceptably strong repulsion and acceptable FTS sampling errors, respectively, is not huge. To us, it would seem premature to rule out the possibility that this gap can be bridged by fine-tuning the FTS approach.

ACKNOWLEDGMENTS

This work was in part supported by the Swedish Research Council (Grant Nos. 621-2014-4522 and 621-2018-04976) and the Swedish strategic research program eSSENCE. Part of the simulations was performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at LUNARC, Lund University, Sweden.

AUTHOR DECLARATIONS

Conflicts of interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, and A. A. Hyman, "Germline P granules are liquid droplets that localize by controlled dissolution/condensation," Science **324**, 1729–1732 (2009).

[2] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, "Biomolecular condensates: organizers of cellular biochemistry," Nat. Rev. Mol. Cell Biol. **18**, 285–298 (2017).

[3] T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, and A. J. Baldwin, "Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles," Mol. Cell **57**, 936–947 (2015).

[4] A. Molliex, J. Temirov, J. Lee, M. Coughlin, A. P. Kanagaraj, H. J. Kim, T. Mittag, and J. P. Taylor, "Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization," Cell **163**, 123–133 (2015).

[5] K. A. Burke, A. M. Janke, C. L. Rhine, and N. L. Fawzi, "Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II," Mol. Cell **60**, 231–241 (2015).

[6] M. L. Huggins, "Solutions of long chain compounds," J. Chem. Phys. **9**, 440 (1941).

[7] P. J. Flory, "Thermodynamics of high polymer solutions," J. Chem. Phys. **10**, 51–61 (1942).

[8] J. T. G. Overbeek and M. J. Voorn, "Phase separation in polyelectrolyte solutions. Theory of complex coacervation," J. Cell Comp. Physiol. **49**, 7–26 (1957).

[9] J. Wittmer, A. Johner, and J. F. Joanny, "Random and alternating polyampholytes," Europhys. Lett. **24**, 263–268 (1993).

[10] Y.-H. Lin, J. D. Forman-Kay, and H. S. Chan, "Sequence-specific polyampholyte phase separation in membraneless organelles," Phys. Rev. Lett. **117**, 178101 (2016).

[11] G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal, "Sequence determinants of protein phase behavior from a coarse-grained model," PLoS Comput. Biol. **14**, e1005941 (2018).

[12] G. L. Dignon, W. Zheng, R. B. Best, Y. C. Kim, and J. Mittal, "Relation between single-molecule properties and phase behavior of intrinsically disordered proteins," Proc. Natl. Acad. Sci. USA **115**, 9929–9934 (2018).

[13] S. Das, A. Eisen, Y.-H. Lin, and H. S. Chan, "A lattice model of charge-pattern-dependent polyampholyte phase separation," J. Phys. Chem. B **122**, 5418–5431 (2018).

[14] S. Das, A. N. Amin, Y.-H. Lin, and H. S. Chan, "Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters," Phys. Chem. Chem. Phys. **20**, 28558–28574 (2018).

[15] N. A. S. Robichaud, I. Saika-Voivod, and S. Wallin, "Phase behavior of blocky charge lattice polymers: crystals, liquids, sheets, filaments, and clusters," Phys. Rev. E **100**, 052404 (2019).

[16] T. S. Harmon, A. S. Holehouse, M. K. Rosen, and R. V. Pappu, "Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins," eLife **6**, e30294 (2017).

[17] T. S. Harmon, A. S. Holehouse, and R. V. Pappu, "Differential solvation of intrinsically disordered linkers drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins," New J. Phys. **20**, 045002 (2018).

[18] S. Qin and H.-X. Zhou, "Fast method for computing chemical potentials and liquid–liquid phase equilibria of macromolecular solutions," J. Phys. Chem. B **120**, 8164–8174 (2016).

[19] G. H. Fredrickson, V. Ganesan, and F. Drolet, "Field-theoretic computer simulation methods for polymers and complex fluids," Macromolecules **35**, 16–39 (2002).

[20] J. McCarty, K. T. Delaney, S. P. O. Danielsen, G. H. Fredrickson, and J.-E. Shea, "Complete phase diagram for liquid–liquid phase separation of intrinsically disordered proteins," J. Phys. Chem. Lett. **10**, 1644–1652 (2019).

[21] G. Parisi, "On complex probabilities," Phys. Lett. B **131**, 393–395 (1983).

[22] J. R. Klauder, "A langevin approach to fermion and quantum spin correlation functions," J. Phys. A: Math. Gen. **16**, L317–L319 (1983).

[23] B. Söderberg, "On the complex Langevin equation," Nucl. Phys. B **295**, 396–408 (1988).

[24] Y. Lin, J. McCarty, J. N. Rauch, K. T. Delaney, K. S. Kosik, G. H. Fredrickson, J.-E. Shea, and S. Han, "Narrow equilibrium window for complex coacervation of tau and RNA under cellular conditions," eLife **8**, e42571 (2019).

[25] S. P. O. Danielsen, J. McCarty, J.-E. Shea, K. T. Delaney, and G. H. Fredrickson, "Molecular design of self-coacervation phenomena in block polyampholytes," Proc. Natl. Acad. Sci. USA **116**, 8224–8232 (2019).

[26] S. Najafi, Y. Lin, A. P. Longhini, X. Zhang, K. T. Delaney, K. S. Kosik, G. H. Fredrickson, J. Shea, and S. Han, "Liquid–liquid phase separation of Tau by self and complex coacervation," Protein Sci. **30**, 1393–1407 (2021).

[27] T. Pal, J. Wessén, S. Das, and H. S. Chan, "Subcompartmentalization of polyampholyte species in organelle-like condensates is promoted by charge-pattern mismatch and strong excluded-volume interaction," Phys. Rev. E **103**, 042406 (2021).

[28] J. Wessén, T. Pal, S. Das, Y.-H. Lin, and H. S. Chan, "A simple explicit-solvent model of polyampholyte phase behaviors and its ramifications for dielectric effects in biomolecular condensates," J. Phys. Chem. B **125**, 4337–4358 (2021).

[29] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," Macromolecules **22**, 3986–3997 (1989).

[30] G. Batrouni, G. Katz, A. Kronfeld, G. Lepage, B. Svetitsky, and K. Wilson, "Langevin simulations of lattice field theories," Phys. Rev. D **32**, 2736–2747 (1985).

[31] A. Irbäck, "Hybrid Monte Carlo simulation of polymer chains," J. Chem. Phys. **101**, 1661–1667 (1994).

[32] J. Ambjørn, M. Flensburg, and C. Peterson, "The complex Langevin equation and Monte Carlo simulations of actions with static charges," Nucl. Phys. B **275**, 375–397 (1986).

[33]E. Seiler, “Status of Complex Langevin,” EPJ Web of Conferences **175**, 01019 (2018).

[34]F. Wang and D. P. Landau, “Efficient, multiple-range random walk algorithm to calculate density of states,” Phys. Rev. Lett. **86**, 2050–2053 (2001).

[35]S. Æ. Jónsson, S. Mohanty, and A. Irbäck, “Accelerating atomic-level protein simulations by flat-histogram techniques,” J. Chem. Phys. **135**, 125102 (2011).

[36]R. H. Swendsen and J.-S. Wang, “Nonuniversal critical dynamics in Monte Carlo simulations,” Phys. Rev. Lett. **58**, 86–88 (1987).

[37]A. Irbäck, S. Æ. Jónsson, N. Linnemann, B. Linse, and S. Wallin, “Aggregate geometry in amyloid fibril nucleation,” Phys. Rev. Lett. **110**, 058101 (2013).

[38]D. Nilsson and A. Irbäck, “Finite-size scaling analysis of protein droplet formation,” Phys. Rev. E **101**, 022413 (2020).

[39]D. Nilsson and A. Irbäck, “Finite-size shifts in simulated protein droplet phase diagrams,” J. Chem. Phys. **154**, 235101 (2021).

[40]S. L. Perry and C. E. Sing, “PRISM-based theory of complex coacervation: excluded volume versus chain correlation,” Macromolecules **48**, 5040–5053 (2015).

Paper v

LU TP 21-37
September 12, 2021

An effective potential for atomic-level simulation of structured and unstructured proteins

Daniel Nilsson[1], Sandipan Mohanty[2] and Anders Irbäck[1]*

[1]Computational Biology & Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden.
[2]Institute for Advanced Simulation, Jülich Supercomputing Centre, Forschungszentrum Jülich, D-52425 Jülich, Germany.

ABSTRACT

Intrinsicially disordered proteins play important roles in processes such as protein aggregation and biomolecular liquid-liquid phase separation, which are computationally challenging to investigate and therefore studied using biophysical models at varying levels of detail, depending on the question at hand. Here, we develop an effective interaction potential for all-atom protein simulations with implicit solvent, by revising an earlier model by us. The previous model was designed and parametrized through studies of a structurally diverse set of folded peptides. In developing the revised model, we use an expanded set of peptides, which also contains experimentally characterized disordered peptides. We demonstrate that the revised model provides an adequate description of the structure and thermal stability a broad set of folded and disordered peptides with 9–32 residues.

*E-mail: anders@thep.lu.se. Phone: +46 46 2223493.

1 INTRODUCTION

For a molecular understanding of cellular processes, there is a need for methods to model the dynamics and interactions of proteins. A long-standing and notoriously difficult problem is to simulate on the computer how proteins fold to their native states. While this problem remains a challenge, the methods and hardware have today reached a stage such that fully atomistic simulations of the folding of small proteins, in explicit solvent, are becoming possible [1]. The past two decades have also seen a growing interest in the properties of intrinsically disordered proteins (IDPs), which lack a well-defined 3D fold, in part due to their prominent roles in protein aggregation and biomolecular liquid-liquid phase separation. Simulations of IDPs are sometimes carried out using explicit solvent and fully atomistic force fields such as AMBER, CHARMM or GROMOS [2–4]. However, long IDPs require large simulation boxes, which can make the calculations challenging. Therefore, it is not uncommon for IDP simulations to rely on coarse-grained protein models without explicit solvent [5–10], sometimes tailored specifically toward IDPs.

In this paper, we develop a biophysical model intended for simulations of both folded proteins and IDPs. For computational speed, the effects of the surrounding solvent are modeled implicitly, through an effective interaction potential. By contrast, the model retains an all-atom protein representation, which in particular facilitates the modeling of the forces driving secondary-structure formation. The model builds on previous work by us [11–13]. Like previous versions, the model presented here uses an effective potential, developed by testing results from thermodynamic simulations against experimental data for a selected set of small polypeptides. In our earlier work, the polypeptides used for this purpose all had a native 3D fold. Since then, a large amount of experimental data on IDPs has become available [14]. Here, we revise the interaction potential in Ref. [13] using data for both structured and unstructured polypeptides.

The revision covers all parts of the interaction potential. The most important change is in the local potential. For the description of local properties of protein chains, the use a grid correction map (CMAP) in the Ramachandran ϕ, ψ space has proven very useful [15, 16]. In previous versions of our model, the local potential was based on a simple ansatz. In the revised model presented here, we adopt a CMAP-like procedure for deriving the local potential, based on data from the Protein Data Bank (PDB).

Our previous model, as implemented in the program package PROFASI [17], has proven useful for studies of some large-size problems, such as the folding of top7 [18], the local unfolding of SOD1 in the presence protein crowders [19], and the mechanical stability of PgiC dimers with $>$1000 residues [20]. It has also been used to study IDPs such as Aβ [21, 22] and α-synuclein [23]. The main aim of the present paper is to improve the description of IDPs, by including disordered peptides in the set of polypeptides used for calibration.

2 BIOPHYSICAL MODEL

We use the same all-atom protein representation as in previous versions of the model [11–13]. We thus constrain bond lengths, bond angles and ω backbone torsion angles to fixed values, which leaves us with the Ramachandran backbone angles ϕ, ψ and sidechain torsion angles χ as the degrees of freedom. Bond lengths and bond angles are as previously described [11–13].

As in the previous models, the effective interaction potential, E, can be split into five major terms, $E = E_{\mathrm{loc}} + E_{\mathrm{ev}} + E_{\mathrm{hb}} + E_{\mathrm{hp}} + E_{\mathrm{ch}}$. One term ($E_{\mathrm{loc}}$) represents local interactions between atoms separated by only a few covalent bonds. The other, non-local terms represent excluded-volume effects (E_{ev}), H bonding (E_{hb}), and residue-specific interactions between pairs of sidechains based on hydrophobicity (E_{hp}) and charge (E_{ch}). In multi-chain simulations, intermolecular interaction terms have the same form and strength as the corresponding intramolecular ones.

In revising the potential, all the five major terms have undergone change. As indicated above, the most important change is in the local potential, E_{loc}, which we now determine by a CMAP-like procedure, based on PDB data. We also modify the form of both the H bonding (E_{hb}) and sidechain-sidechain (E_{hp}, E_{ch}) potentials, whereas the changes of the excluded-volume potential (E_{ev}) are modest.

In the following, we describe the five major terms of the potential. All energies are given in a unit called eu, which is the thermal energy kT at 315 K, corresponding to 0.6260 kcal/mol.

Excluded volume

The repulsive excluded-volume potential is taken to have the form

$$E_{\mathrm{ev}} = \kappa_{\mathrm{ev}} \sum_{i<j} \left(\frac{\sigma_i + \sigma_j}{r_{ij}} \right)^{12} \tag{1}$$

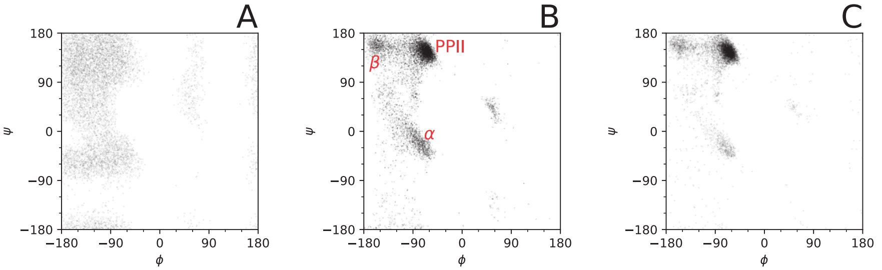


FIGURE 1: Ramachandran maps for (A) the alanine residue in the peptide GAG using our model without any local potential, (B) alanine residues in the PDB database classified as coil by STRIDE [24], and (C) the alanine residue in the peptide GAG using our model including the local potential. The ϕ,ψ regions corresponding to α-helix, β-sheet and polyproline II (PPII) structure are indicated in (B).

where $\kappa_{\text{ev}} = 0.04$ eu and the radius parameter σ_i is set to $\sigma_i = 1.77$, 1.75, 1.53, 1.42 and 1.00 Å for S, C, N, O and H atoms, respectively. The sum runs over all atom pairs with a non-fixed separation r_{ij}, except those H-O pairs that are capable of forming H bonds. The latter pairs can be excluded because the H bonding potential contains a short-range hard-core repulsion (see below). We also note that previous versions of our model [11–13] used a reduced excluded-volume repulsion for atom pairs separated by three covalent bonds. This reduction has now been removed. Each term in Eq. 1 is evaluated using a cutoff of 4.3 Å.

Local potential

While the excluded-volume term alone is able to qualitatively describe several key features of observed Ramachandran ϕ,ψ distributions, in some regions this term fails to match PDB data, either qualitatively or quantitatively (Figs. 1A,B). Some important examples are as follows.

- Angle pairs (ϕ,ψ) to the left of the α-helix region are not statistically suppressed by steric interactions (Fig. 1A), while in the PDB very few residues are found in this part (Fig. 1B).
- Steric interactions do not separate the β and polyproline-II (PPII) regions (Fig. 1A).
- With only steric interactions, the β region is less asymmetric around the $\phi = -\psi$ line than seen in the PDB (Figs. 1A,B). This asymmetry is important because it is linked to the

TABLE 1: Reweighting factors used when generating target probabilities, $P_{\text{tg}}(\phi,\psi,\text{X})$, for the local potential. Any point in the specified ϕ,ψ regime is multiplied by this factor. The functions used are defined by $\ln u(\phi,\psi) = 1.21 + 0.49\{1+\exp[-0.4(0.87(\phi+67.5^\circ) - 0.50(\psi-150^\circ)+12^\circ)]\}^{-1}$ and $\ln v(\phi,\psi) = 1.21\{1+\exp[-0.06(\phi-\psi+45^\circ)]\}^{-1}$

ϕ-range	ψ-range	Reweighting factor	Amino acids
$\phi < 0^\circ$	$50^\circ < \psi$	$u(\phi,\psi)$	all
$-150^\circ < \phi < 0^\circ$	$-100^\circ < \psi < 50^\circ$	$v(\phi,\psi)$	all but proline
		$e^{2.91}v(\phi,\psi)$	proline
$0^\circ < \phi < 150^\circ$	$-50^\circ < \psi < 100^\circ$	$e^{2.18}$	glycine

observed twist of β-sheets.

These observations suggest that local non-steric interactions play an important role. Due to their importance, rather than using some *ad hoc* ansatz for such interactions, we here implement a CMAP-like approach for determining amino acid-specific local interaction energies, based on PDB data. For this purpose, we collected a set of 7133 protein X-ray structures using the PISCES web server [25].

Each amino acid i, except the two terminal ones, is taken to contribute a term $e_{\text{loc}}(\phi_i,\psi_i;\text{X}_i)$ to the total local potential E_{loc}, where ϕ_i and ψ_i are the Ramachandran angles and X_i stands for amino acid type. In the Ramachandran plane, the function $e_{\text{loc}}(\phi,\psi;\text{X})$ is piecewise constant on $2^\circ \times 2^\circ$ squares. Our procedure for determining $e_{\text{loc}}(\phi,\psi;\text{X})$ is as follows.

First we generate a target distribution $P_{\text{tg}}(\phi,\psi;\text{X})$ for each amino acid type X, based on PDB data. This problem is solved in two steps. The aim of the first step is to obtain an accurate description of the shape of the major free-energy minima, like α, β and PPII, for different amino acids X. To this end, we use the Ramachandran angles of all amino acids in our database that are classified as coil by STRIDE [24]. Additionally, we include data for amino acids classified as 3^{10}- or α-helix, with a weight of 1/50 and 1/300, respectively, relative to amino acids classified as coil. The addition of 3^{10}- and α-helix data improves the description of these otherwise weakly populated regions. Our second and final step in determining $P_{\text{tg}}(\phi,\psi;\text{X})$ is to adjust the relative weights of the major free-energy minima, by multiplying the original distributions with a few reweighting factors. These factors are selected by trial and error, and can be found in Table 1.

Given the target distributions $P_{\text{tg}}(\phi,\psi;X)$, we determine $e_{\text{loc}}(\phi,\psi;\text{X})$ using Monte Carlo simulations of GXG tripeptides, at 275 K. We want the Ramachandran distribution obtained from such a simulation, $P_{\text{sim}}(\phi,\psi;X;E_{\text{loc}})$, to closely resemble the target distribution, $P_{\text{tg}}(\phi,\psi;X)$. To this end, we determine $e_{\text{loc}}(\phi,\psi,\text{X})$ by iteratively minimizing the cost function

$$\sum_{\phi,\psi}[P_{\text{sim}}(\phi,\psi;\text{X};e)-P_{\text{tg}}(\phi,\psi;\text{X})]^2+\lambda\sum_{\phi,\psi}\nabla^2 e_{\text{loc}}(\phi,\psi,\text{X}), \tag{2}$$

where the second term serves to smoothen the final, and otherwise spiky, $e_{\text{loc}}(\phi,\psi,\text{X})$. In Eq. 2, ∇^2 is a discrete Laplace operator, acting on the $2^\circ\times 2^\circ$ grid.

The procedure above works well for most amino acids, but for the four amino acids aspartic acid, asparagine, glycine and proline, some special considerations apply.

Asparagine and aspartic acid require special consideration because their sidechains can easily H bond with the backbone NH group on the amino acid two steps downstream, forming a so-called Asx turn. This possibility biases the observed Ramachandran distribution. To remedy this, we exclude all Asn/Asp residues i with any of their sidechain oxygen atoms within 4 Å of the backbone nitrogen of amino acid $i+2$ when generating the target distribution.

Glycine is unique in that its Ramachandran distribution contains heavily populated free-energy minima with $\phi > 0^\circ$. These minima are different in form for glycine residues classified as coil and as turn. For this reason, in the first step of the construction of $P_{\text{tg}}(\phi,\psi;\text{G})$, we weighted in Ramachandran pairs classified as turn, and located in the region $0^\circ < \phi < 150^\circ$, $-50^\circ < \psi < 100^\circ$ with a weight of 1/4. Finally, the distribution was explicitly symmetrized, $P(\phi,\psi) = P(-\phi,-\psi)$.

For proline, the ϕ angle is fixed because of the assumed geometry of our model. We thus generate a local term for only the ψ angle, following a procedure similar to that for other residues. The target distribution is generated for the full ϕ,ψ plane as above, and then maginalized to retrieve a target distribution for ψ alone. Note that the reweighting factors (Table 1) are modifed for both proline and glycine.

For the sidechain angles χ_i, we use the same local potential as in Ref. [13]. The sidechain contribution to the total local potential E_{loc} is

$$E_{\text{loc}}^{\text{sc}} = \sum_i \kappa_i \cos n_i\chi_i, \tag{3}$$

TABLE 2: Groups participating in H bonding in the model. An asterisk indicates sp^3 hybridized N atoms.

Donor (NH) groups	Acceptor (CO) groups
Backbone	Backbone
Arg sidechain	Asp sidechain
Trp sidechain	Glu sidechain
His sidechain	Asn sidechain
Asn sidechain	Gln sidechain
Gln sidechain	C-terminus
Lys sidechain*	
N-terminus*	

where the parameters κ_i and n_i depend on the type of sidechain angle and have the same values as in Ref. [13].

H bonding

The H bond potential, E_{hb}, consists of explicit H bonding terms between NH donor groups and CO acceptor groups. Both backbone and side-chain NH and CO groups can participate in H bonding (see Table 2), but mainchain-mainchain H bonding between adjacent peptide units along the chain are disallowed.

The interaction energy of a pair I of one NH and one CO group is taken to have the form

$$e_{\text{hb},I} = \varepsilon_{\text{hb}} \left[5 \left(\frac{\sigma_{\text{hb}}}{r_I} \right)^{12} - 6 \left(\frac{\sigma_{\text{hb}}}{r_I} \right)^{10} \right] \psi_I \tag{4}$$

where r_I is the HO distance, $\sigma_{\text{hb}} = 2.0\,\text{Å}$, $\varepsilon_{\text{hb}} = 3.4\,\text{eu}$, and ψ^I represents a directional dependence. For the radial dependence in Eq. 4, a cutoff of 4.5 Å is used. For most donor-acceptor pairs, the function ψ_I depends on both the NHO angle, α_I, and the HOC angle, β_I, and is given by

$$\psi_I = \begin{cases} (\cos\alpha_I \cos\beta_I)^{1/2} & \text{if } \alpha_I, \beta_I > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

However, for a mainchain-mainchain H bond connecting the NH group of amino acid i and the CO group of amino acid $i-3$, the dependence on α_I is omitted, such that $\psi_I = (-\cos\beta_I)^{1/2}$ when

$\beta_I > 90^\circ$ and $\psi_I = 0$ otherwise. In this case, the strength parameter is reduced to $\varepsilon_{\text{hb}} = 2.8\,\text{eu}$. This exception is made to promote the formation of certain β turns that otherwise were only rarely observed, probably due to our use of fixed bond lengths and bond angles.

The total H bond energy is usually defined as a simple sum of single-bond energies over all possible combinations of donor and acceptor groups. However, each NH group should be able to participate in at most one proper H bond, and each CO group in at most two such bonds. To strictly enforce this condition, rather than relying on indirect steric effects, we define the total H bond energy as a constrained sum of single-bond energies,

$$E_{\text{hb}} = \sum_I{}' e_{\text{hb},I} \tag{6}$$

In this constrained summation, indicated by a prime, all positive energies $e_{\text{hb},I}$ and a selected set of negative energies $e_{\text{hb},I}$ are included. The selection of which negative energies to include proceeds as follows.

1. All NH,CO pairs I with a negative $e_{\text{hb},I}$ are listed as potential bonds.
2. For each donor and acceptor group, among all potential bonds I involving this group, the one with lowest energy is determined, and marked as the preferred bond of the group.
3. All potential bonds I which are preferred bonds by both their donor and acceptor groups are accepted as bonds and removed from the list of potential bonds. Their energies $e_{\text{hb},I}$ are added to the total energy E_{hb}.
4. Any remaining non-accepted potential bond I whose donor group already participates in one accepted bond or whose acceptor group participates in two accepted bonds is removed from the list of potential bonds. Potential bonds where the acceptor group participates in one bond are reduced in strengths by a factor $\gamma_{\text{db}} = 0.5$.
5. Repeat from step 2 until the list of potential bonds is empty.

Hydrophobicity

The E_{hp} potential provides an effective attraction between hydrophobic sidechains. The atoms in each sidechain are classified in three categories: strongly hydrophobic, weakly hydrophobic, and not hydrophobic. Table 3 shows all atoms classified as strongly or weakly hydrophobic.

TABLE 3: Atoms defined as strongly or weakly hydrophobic in the various amino acids.

Amino acid	Strongly hydrophobic atoms	Weakly hydrophobic atoms
Ala	C_β	
Val	$C_{\gamma 1}$,$C_{\gamma 2}$	C_β
Leu	C_γ,$C_{\delta 1}$,$C_{\delta 2}$	C_β
Ile	$C_{\gamma 1}$,$C_{\gamma 2}$,C_δ	C_β
Met	S_δ,C_ε	C_β,C_γ
Pro	C_β,C_γ,C_δ	
Phe	$C_{\delta 1}$,$C_{\delta 2}$,$C_{\varepsilon 1}$,$C_{\varepsilon 2}$,C_ζ	C_β,C_γ
Tyr	$C_{\delta 1}$,$C_{\delta 2}$,$C_{\varepsilon 1}$,$C_{\varepsilon 2}$	C_β,C_γ,C_ζ
Trp	$C_{\delta 1}$,$C_{\varepsilon 3}$,$C_{\zeta 2}$,$C_{\zeta 3}$,$C_{\eta 2}$	C_β,C_γ,$C_{\delta 2}$,$C_{\varepsilon 2}$

For each pair i, j of strongly or weakly hydrophobic atoms not in the same residue, a contact measure C_{ij} is calculated, such that $C_{ij} = 0$ (no contact) if the separation is greater than $a = 5$ Å, $C_{ij} = 1$ (full contact) if the separation is less than $b = 3.5$ Å, and $C_{ij} = (a^2 - r_{ij}^2)/(a^2 - b^2)$ in between.

A simple choice of energy function based on these contacts would be one proportional to the sum of all contacts between pairs of hydrophobic atoms, where at least one atom is strongly hydrophobic. However, such an energy function might encourage conformations, wherein many hydrophobic atoms are located in close proximity, thus creating a large number of contacts. To discourage such conformations, we adjust this simple formulation in two ways.

First, for any pair i, j where both atoms are also in contact with at least one other atom, we use a modified contact measure, C'_{ij}, defined as

$$C'_{ij} = C_{ij} \prod_K [1 - \max_{k \in K}(C_{ik}C_{jk})/4], \tag{7}$$

where the product is over amino acids K. This modified contact measure is meant to mimic the effect of one hydrophobic atom “shielding” another.

Second, we remove the contribution from having multiple contacts with the same amino acid by calculating the final energy as

$$E_{\text{hp}} = -\varepsilon_{\text{hp}} \sum_{J>I+1} \sum_{i \in I} \max_{j \in J}(C'_{ij}), \tag{8}$$

TABLE 4: Sidechain atoms defined as charged, and their assigned charges (in units of the elementary charge e).

Amino acid	Atoms	Charge
Asp	$O_{\delta 1}$,$O_{\delta 2}$	$-1/2$
Glu	$O_{\varepsilon 1}$,$O_{\varepsilon 2}$	$-1/2$
Lys	N_{ζ}	$+1$
Arg	$N_{\varepsilon 1}$,$N_{\eta 1}$,$N_{\eta 2}$	$+1/3$

where the outer sum is over all pairs I,J of amino acids that are not nearest neighbors along the chain, i runs over strongly hydrophobic atoms in amino acid I, and j runs over all (strongly or weakly) hydrophobic atoms in amino acid J. The strength parameter is set to $\varepsilon_{\mathrm{hp}} = 0.45\,\mathrm{eu}$.

Note that hydrophobic atoms with a given class play the same role irrespective of what amino acid type they belong to. Therefore, after specifying the hydrophobicity class of all individual atoms, there is only one free strength parameter in E_{hp}, namely the overall strength $\varepsilon_{\mathrm{hp}}$. This represents a major reduction of the number of parameters compared to Ref. [13].

Electrostatics

The last term of the potential, E_{ch}, represents electrostatic energy between charged amino acids. Assuming approximately neutral pH, there are two positively charged amino acids, arginine and lysine, and two negatively charged ones, aspartic and glutamic acid. Furthermore, it is assumed that the Coulomb interactions are screened by salt. For simplicity, we then model electrostatic interactions using the same measure of contact, C_{ij}, as in the hydrophobic potential. Our contact-based definition of E_{ch} reads

$$E_{\mathrm{ch}} = \varepsilon_{\mathrm{ch}} \sum_{i \neq j} q_i q_j C_{ij}\,, \tag{9}$$

where q_i and q_j are atomic charges. The partial charges carried by atoms in the charged amino acids can be found in Table 4. The two chain ends are treated in the same way as the lysine and aspartic/glutamic acid sidechains, respectively. The strength parameter is set to $\varepsilon_{\mathrm{ch}} = 0.8\,\mathrm{eu}$.

Simulations details

Our simulations are performed using Monte Carlo-based conformational sampling, along with simulated [26–28] or parallell [29–31] tempering. A typical peptide simulation ran for 5.5×10^9

elementary Monte Carlo updates.

Our move set consists of the following three elementary updates: (i) pivot, where a single backbone angle is rotated, (ii) BGS [32], where a series of consecutive backbone angles are updated in such a way that the ends of this chain segment are kept approximately fixed, and (iii) rotation of a single sidechain angle.

Model development and most of the peptide simulations were performed using a C simulation program. The final model was independently reimplemented in the PROFASI C++ package [17]. It was checked that the two implementations produced the same results.

Analysis

Our simulations of small folded proteins give conformational ensembles that contain both folded and unfolded structures. If an experimentally known structure is available, we quantify the nativeness of the folded subensemble by computing the average root-mean-square deviation (RMSD) from the experimental structure over the folded subensemble.

For some peptides forming a simple α-helix, we use the helix content, q_{h}, as a measure of nativeness. A residue is defined as helical if its Ramachandran angle pair is in region the $-90^\circ < \phi < 30^\circ$, $-77^\circ < \phi < 17^\circ$. A given conformation is deemed helical if $>60\%$ of the residues are helical (excluding the two end residues), and q_{h} is the fraction of simulated conformations that are helical.

For some peptides forming β-structure, we use a nativeness measure, q_{hb}, based on H bonding. Specifically, we define q_{hb} as the fraction of simulated conformations in which at most two the native H bonds are missing. An H bond is said to be formed if its energy is $<-\varepsilon_{\mathrm{hb}}/3$.

In evaluating simulated ensembles for disordered peptides, we compute ${}^3J_{\mathrm{HNH}\alpha}$ scalar couplings using the Karplus equation [33], with coefficients derived by Ref. [34], which we compare to NMR data.

3 RESULTS

This section provides a summary of simulation results obtained with the biophysical model presented in Sec. 2, to illustrate the potential and limitations of this model. We consider a set of about 30 polypeptides with 9–42 residues, some of which have a native fold, while others are disordered. The amino acid sequences of these polypeptides can be found in Appendix A.

TABLE 5: Summary of simulation results for seven helical peptides, and comparison to experimental results.

Peptide	Observable	Experiment	Simulation
trp-cage	Structure	1L2Y [35]	RMSD 1.8 Å
	Melting temperature	315 K [35]	297 K
E6apn1	Structure	1RIJ [36]	RMSD 2.0 Å
	Melting temperature	305 K [36]	309 K
C	Helix content	Partially helical (273 K) [37]	$h = 0.43$ (275 K)
EK	Helix content	Partially helical (273 K) [38]	$h = 0.96$ (275 K)
F_s	Melting temperature	303 K (CD) [39] 308 K (CD) [40] 334 K (IR) [41]	329 K
GCN4tp	Structure	2OVN [42]	RMSD 1.4 Å
	Helix content	0.6 (278 K, NMR, CD) [42]	$h = 0.65$ (275 K)
	Melting temperature		287 K
HPLC-6	Structure	1WFA [43]	RMSD 1.2 Å
	Helix content	0.10 (CD; 343 K) [44]	0.13 (336 K)
	Melting temperature		327 K

Folded peptides

We begin with a set of 17 peptides which all have a folded 3D structure, although their thermal stability varies. This set was used when parametrizing the previous version of our interaction potential [13]. It consists of seven helical peptides and 10 peptides forming β-hairpins or three-stranded β-sheets. A summary of the results obtained for these peptides with the current version of the interaction potential can be found in Tables 5 and 6, which show data for helical and β-sheet peptides, respectively.

As can be seen from these tables, after fine-tuning the new interaction potential, our model remains able to fold all these 17 peptides, and the thermal stability for the most part shows good agreement with experimental results. For a few peptides, the deviations from the experimental results are somewhat large, with the simulated thermal stability being either higher (EK peptide) or lower (chignolin and trpzip1) than experimental data. However, it should be noted that precise comparisons of simulated and experimental thermal stabilities are difficult for small polypep-

TABLE 6: Summary of simulation results for 10 β-sheet peptides, and comparison to experimental results.

Peptide	Observable	Experiment	Simulation
chignolin	Structure	1UAO [45]	RMSD 0.8 Å
(hairpin)	Melting temperature	311-315 K (CD/NMR) [45]	279 K
MBH12	Structure	1J4M [46]	RMSD 0.8 Å
(hairpin)	Melting temperature		305 K
GB1p	Native population	0.42 (CD/NMR; 278 K) [47]	$q_{hb} = 0.94$ (275 K)
(hairpin)		0.30 (CD/NMR; 298 K) [48]	$q_{hb} = 0.63$ (298 K)
	Melting temperature	297 K (Trp flourescence) [49]	303 K
GB1m2	Native population	0.74 (CD/NMR; 298 K) [48]	$n_s = 0.85$ (298 K)
(hairpin)	Melting temperature	320K (CD/NMR) [48]	316 K
GB1m3	Native population	0.86 (CD/NMR; 298 K) [48]	$q_{hb} = 0.95$ (298 K)
(hairpin)	Melting temperature	333 K (CD/NMR) [48]	325 K
trpzip1	Structure	1LE0 [50]	RMSD 0.9 Å
(hairpin)	Melting temperature	323 K (CD) [50]	287 K
trpzip2	Structure	1LE1 [50]	RMSD 0.7 Å
(hairpin)	Melting temperature	345 K (CD) [50]	325 K
		290–335 K (various) [51]	
betanova	Structure		RMSD 2.8 Å
(three-stranded sheet)	Native population	0.08 (NMR; 283 K) [52]	$q_{hb} = 0.09$ (286 K)
LLM	Structure		RMSD 1.8 Å
(three-stranded sheet)	Native population	0.36 (NMR; 283 K) [53]	$q_{hb} = 0.44$ (286 K)
beta3s	Structure		RMSD 1.8 Å
(three-stranded sheet)	Native population	0.13–0.31 (NMR; 283 K) [54]	$q_{hb} = 0.19$ (273 K)

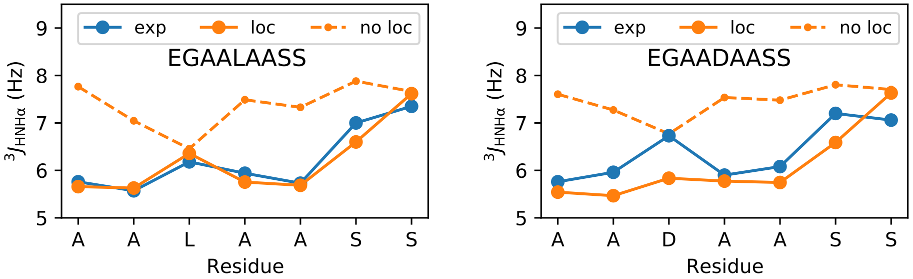


FIGURE 2: $^3J_{\mathrm{HNH}\alpha}$ scalar couplings from simulations (at 294 K) and experiments (at 298 K) [55] for two disordered EGAAXAASS peptides. Experimental data are read off from figure 3 in Ref. [55]. In addition to simulation results obtained using the full potential (“loc”), we also show data from simulations without the local potential ("no loc”). (Left) EGAALAASS and (Right) EGAADAASS.

tides like these, as the melting temperature sometimes shows a significant dependence on the monitored observable. For instance, a study of trpzip2 measured melting temperatures varying between 290 K and 335 K depending on observable [51]. This uncertainty limits the possibilities to further fine-tune the model based on thermal stability data.

Disordered EGAAXAASS peptides

To assess the ability of our model to describe IDPs, we first consider a set of 14 disordered nine-residue peptides with sequences of the form EGAAXAASS, for different choices of the residue “X”. This set of peptides has been studied by NMR [55]. Figure 2 compares simulation and experimental data for $^3J_{\mathrm{HNH}\alpha}$ scalar couplings of two such peptides, with respectively L and D at the variable position. Similar plots for the remaining 12 peptides can be found in Appendix B. All the 14 peptides are indeed disordered in the simulations, and the simulated $^3J_{\mathrm{HNH}\alpha}$ scalar couplings show an adequate, albeit not perfect, agreement with experimental data. In Fig. 2, we also include data from otherwise identical simulations with the local potential switched off, which underscore the importance of this potential. The agreement with experimental data deteriorates considerably when leaving this term out. Figure 3 shows a scatter plot of simulated versus experimental $^3J_{\mathrm{HNH}\alpha}$ values, for all the $^3J_{\mathrm{HNH}\alpha}$ scalar couplings that were determined experimentally [55]. From this figure, it can be seen that while there is a slight tendency for simulation

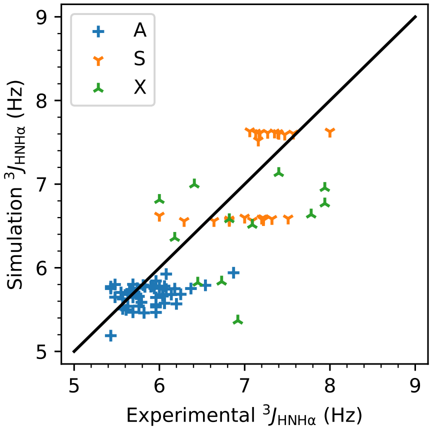


FIGURE 3: Simulation results (at 294 K) versus experimental data (at 298 K) for all the $^3J_{\mathrm{HNH\alpha}}$ scalar couplings measured experimentally [55] in 14 EGAAXAASS peptides. The experimental data are read off from figure 3 in Ref. [55].

data to fall below experimental data, a clear correlation exists between the two data sets.

The experimental data on these EGAAXAASS peptides [55] have previously been used to evaluate a few force fields for molecular dynamics simulations with explicit solvent [56]. These authors computed the RMSD of simulated $^3J_{\mathrm{HNH\alpha}}$ scalar couplings from their experimental values for five of the EGAAXAASS peptides, with I, V, D, G and W at the variable position. The best results were obtained with the AMBER03-ILDN force field [57, 58], which gave an average RMSD of 0.45 Hz over the five peptides. Table 7 shows RMSD values from simulations with our model for all the 14 EGAAXAASS peptides studied in Ref. [55]. The average RMSD over the above-mentioned set of five peptides is 0.54 Hz, which is slightly worse than for the best force

TABLE 7: RMSDs of simulated $^3J_{\mathrm{HNH\alpha}}$ scalar couplings from their experimental values [55] for 14 EGAAXAASS peptides.

Residue X	I	V	L	N	Q	T	D	E	K	G	P	W	Y	H	Avg.
RMSD (Hz)	0.57	0.58	0.21	0.25	0.39	0.23	0.52	0.80	0.34	0.47	0.73	0.54	0.51	0.47	0.50

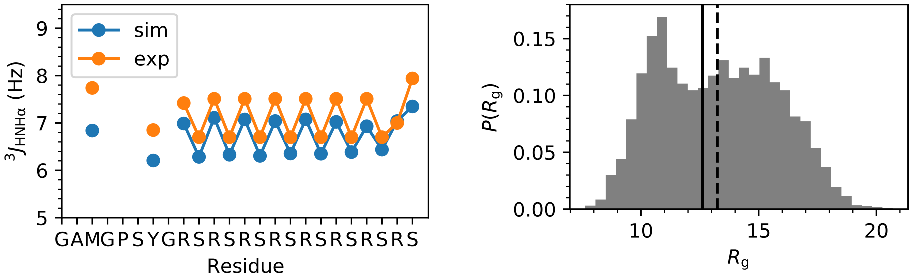


FIGURE 4: Comparison of simulation and experimental results for the RS peptide. (Left) $^3J_{\mathrm{HNH}\alpha}$ scalar couplings from our simulations (at 298 K) and the experiments (at 298 K) of Ref. [59]. (Right) Probability distribution of the radius of gyration, R_g, from our simuations (at 298 K). The solid line indicates the experimentally determined mean of R_g (at 298 K) [60], while the dashed line represents the mean from our simulation.

field in Ref. [56]. The average RMSD over all the 14 peptides is 0.50 Hz. Overall, we conclude that our model, despite its lower level of detail, is able to achieve results that are close to those obtained with the best of the explicit-solvent force fields tested in Ref. [56].

The disordered RS peptide

Another disordered peptide that has served as benchmark in evaluating the performance of other force fields is the 24-residue RS peptide [59, 60]. Figure 4 (left panel) shows our simulated $^3J_{\mathrm{HNH}\alpha}$ couplings for this peptide, along with experimentally measured values [59]. The average unsigned error of our simulation results is 0.43 Hz, which is about 30% higher than what was obtained with the CHARMM22 [61] force field in Ref. [60], but lower than the errors found with all the other force fields tested in Ref. [60]. It worth noting that the overall shape of the experimental $^3J_{\mathrm{HNH}\alpha}$ profile is accurately reproduced by our model. A large part of the deviations from the experimental values corresponds to a simple constant shift by about 0.4 Hz.

For the RS peptide, the chain extension has also been investigated experimentally, by small-angle X-ray scattering (SAXS) [60]. Through the SAXS data, the mean radius of gyration, R_g, was measured [60], and simulation results obtained with different force fields were compared to this measured value [60]. The best result was again obtained with the CHARMM22 [61] force field, which gave an R_g value consistent with the measured value. Figure 4 (right panel) shows the

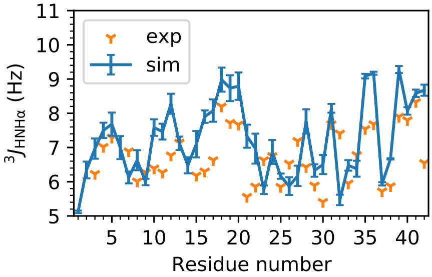


FIGURE 5: $^3J_{\mathrm{HNH\alpha}}$ scalar couplings for Aβ42 from our simulations (at 273 K) and the experiments (at 273 K) of Ref. [62].

probability distribution of R_g in our model. As can be seen, the distribution is weakly bimodal, and the mean value falls slightly above the experimentally determined value. Interestingly, a nearly bimodal distribution of R_g was also seen for the best performing force field in Ref. [60].

Summarizing the results for the EGAAXAASS and RS peptides, we find that, despite the absence of explicit solvent, our model gives results comparable with the best performing force fields in previous tests using these peptides [56, 60]. It is also worth noting that the only force fields that beat us in each case (AMBER03 for EGAAXAASS and CHARMM22 for RS) performed relatively poorly on the other benchmark.

Aβ42

All systems discussed so far were part of the iterative process by which the new potential was derived. Finally, we briefly discuss some preliminary results obtained with this potential for an additional peptide, the disordered Aβ42 peptide with 42 residues, which has been studied with our previous model [21].

Figure 5 shows $^3J_{\mathrm{HNH\alpha}}$ scalar couplings from our Aβ42 simulations, along with experimental data from Ref. [62]. The agreement is not perfect, with a Pearson correlation coefficient of 0.61 and an RMSD of 0.90 Hz, computed over all the 38 experimentally determined couplings. For instance, the data suggest that the model overestimates the probability of β-hairpin formation in the C-terminal part. Nevertheless, some trends seen in the experimental data are captured by the simulations. Note also the largest discrepancy is for the C-terminal residue 42, for which we have not included any local potential in the current simulations.

4 DISCUSSION AND SUMMARY

We have in this paper formulated and parameterized an effective interaction potential for all-atom protein simulations with implicit solvent, starting from an earlier potential developed by us [13]. While the development of this predecessor was based on comparisons to experimental data on folded peptides only [13], we have here used experimental data on both folded and disordered peptides, the main aim being to improve the description of IDPs.

Our revision of the potential entails many changes, several of which are physically motivated, and some of which are made just to simplify the potential, when possible. The most important change is the adoption of a CMAP-like approach for determining an amino acid-specific local potential, based on PDB data. Through the use of PDB data, it is possible to derive a local potential that is able to accurately reproduce the amino acid-dependent shape of the populated regions in the Ramachandran ϕ,ψ space, which is hard to achieve using some *ad hoc* ansatz. However, the PDB-based analysis alone cannot be expected to provide accurate relative weights of the populated Ramachandran regions, like the α/β ratio. Therefore, in a final second step, the relative weigths of these regions were manually adjusted (Table 1), based on simulation data.

The simulation results presented here demonstrate that the new potential retains the ability of the old one [13] to adequately describe a structurally diverse set of 17 folded peptides. In addition, simulations with the new potential yield results in approximate agreement with experimental data on the disordered EGAAXAASS and RS peptides. With regard to these disordered peptides, the new potential represents a significant improvement over the old one. How useful the model is for larger polypeptides remains to be seen. The study of larger polypeptides is beyond the scope of the present paper.

The interaction potential presented in this paper has been implemented in the open source program package PROFASI [17].

ACKNOWLEDGEMENTS

This work was in part supported by the Swedish Research Council (Grant no. 621-2018-04976) and the Swedish strategic research program eSSENCE. The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at LUNARC, Lund University, Sweden, and the Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany.

APPENDIX A: AMINO ACID SEQUENCES

TABLE A1: Amino acid sequences of the polypeptides studied. Capping groups are indicated when used. Ac and Suc stand for acetyl and succinylic acid, respectively.

Polypeptide	PDB ID	Sequence
trp-cage	1L2Y	NLYIQ WLKDG GPSSG RPPPS
E6apn1	1RIJ	Ac–ALQEL LGQWL KDGGP SSGRP PPS–NH_2
C		Ac–KETAA AKFER AHA–NH_2
EK		Ac–YAEAA KAAEA AKAF–NH_2
F_s		Suc–AAAAA AAARA AAARA AAARA A–NH_2
GCN4tp	2OVN	NYHLE NEVAR LKKLV GE
HPLC-6	1WFA	DTASD AAAAA ALTAA NAKAA AELTA ANAAA AAAAT AR–NH_2
chignolin	1UAO	GYDPE TGTWG
MBH12	1J4M	RGKWT YNGIT YEGR
GB1p		GEWTY DDATK TFTVT E
GB1m2		GEWTY NPATG KFTVT E
GB1m3		KKWTY NPATG KFTVQ E
trpzip1	1LE0	SWTWE GNKWT WK–NH_2
trpzip2	1LE1	SWTWE NGKWT WK–NH_2
betanova		RGWSV QNGKY TNNGK TTEGR
LLM		RGWSL QNGKY TLNGK TMEGR
beta3s		TWIQN GSTKW YQNGS TKIYT
RS		GAMGP SYGRS RSRSR SRSRS RSRS
Aβ		DAEFR HDSGY EVHHQ KLVFF AEDVG SNKGA IIGLM VGGVV IA

APPENDIX B: $^3J_{\mathrm{HNH}\alpha}$ SCALAR COUPLINGS FOR EGAAXAASS PEPTIDES

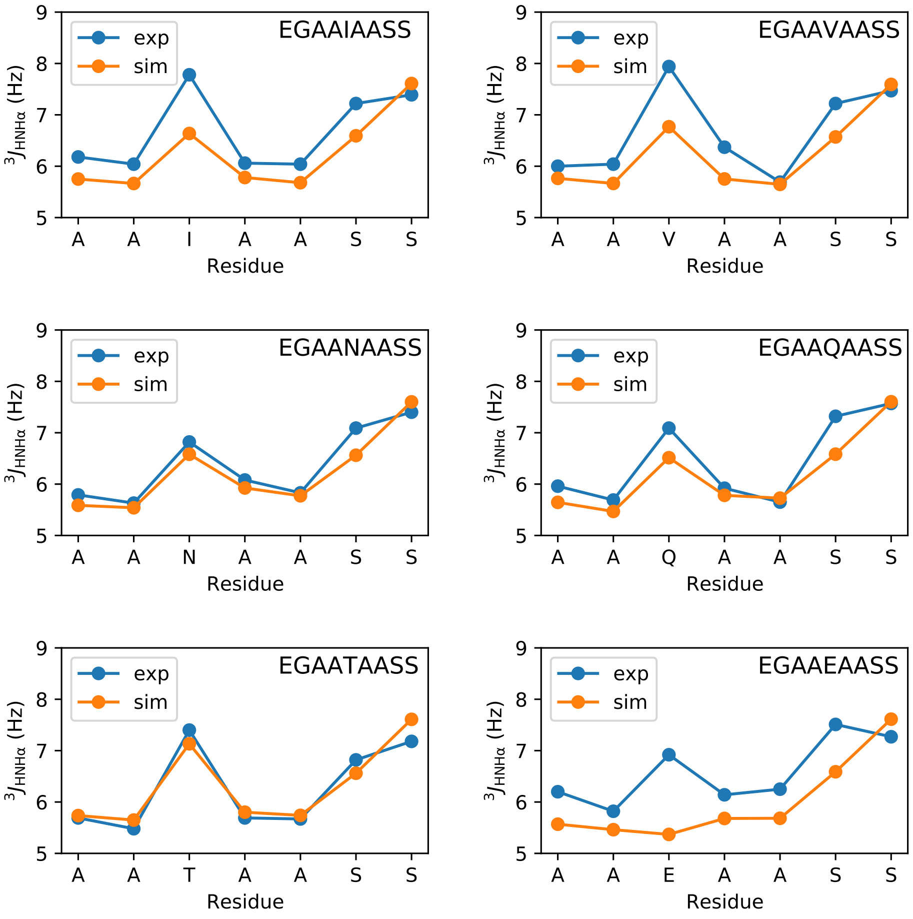


FIGURE A1: $^3J_{\mathrm{HNH}\alpha}$ scalar couplings from simulations (at 294 K) and experiments (at 298 K) [55] for EGAAXAASS peptides with I, V, N, Q, T and E at the variable position. Experimental data are read off from figure 3 in Ref. [55].

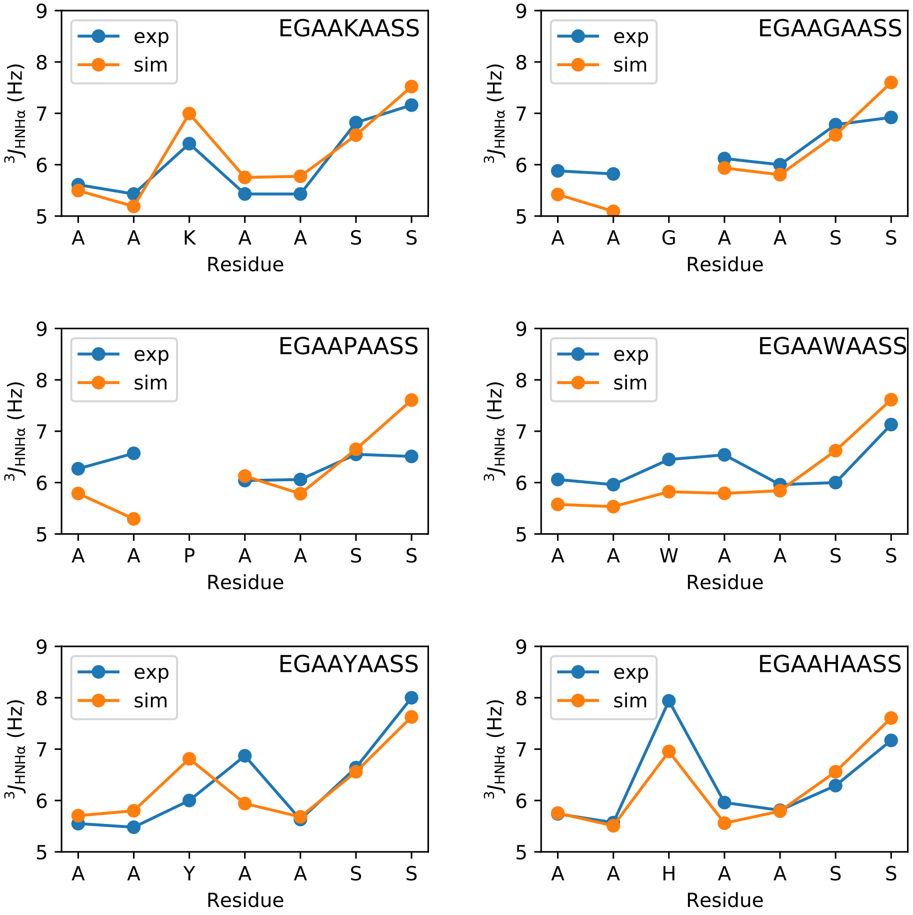


FIGURE A2: $^3J_{\mathrm{HNH}\alpha}$ scalar couplings from simulations (at 294 K) and experiments (at 298 K) [55] for EGAAXAASS peptides with K, G, P, W, Y and H at the variable position. Experimental data are read off from figure 3 in Ref. [55].

REFERENCES

1. Lindorff-Larsen, K., S. Piana, R. Dror, and D. Shaw, 2011. How fast-folding proteins fold. *Science* 334:517 – 520.
2. Maier, J. A., C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, 2015. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* 11:3696–3713.
3. Huang, J., and A. D. MacKerell Jr, 2013. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* 34:2135–2145.
4. Schmid, N., A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, 2011. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* 40:843–856.
5. Cragnell, C., E. Rieloff, and M. Skepö, 2018. Utilizing coarse-grained modeling and Monte Carlo simulations to evaluate the conformational ensemble of intrinsically disordered proteins and regions. *J. Mol. Biol.* 430:2478–2492.
6. Wu, H., P. G. Wolynes, and G. A. Papoian, 2018. AWSEM-IDP: a coarse-grained force field for intrinsically disordered proteins. *J. Phys. Chem. B* 122:11115–11125.
7. Latham, A. P., and B. Zhang, 2019. Maximum entropy optimized force field for intrinsically disordered proteins. *J. Chem. Theory Comput.* 16:773–781.
8. Choi, J.-M., and R. V. Pappu, 2019. Improvements to the ABSINTH force field for proteins based on experimentally derived amino acid specific backbone conformational statistics. *J. Chem. Theory Comput.* 15:1367–1382.
9. Heilmann, N., M. Wolf, M. Kozlowska, E. Sedghamiz, J. Setzler, M. Brieg, and W. Wenzel, 2020. Sampling of the conformational landscape of small proteins with Monte Carlo methods. *Sci. Rep.* 10:1–13.
10. Regy, R. M., J. Thompson, Y. C. Kim, and J. Mittal, 2021. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* 30:1371–1379.
11. Irbäck, A., B. Samuelsson, F. Sjunnesson, and S. Wallin, 2003. Thermodynamics of α- and β-structure formation in proteins. *Biophys. J.* 85:1466–1473.
12. Irbäck, A., and S. Mohanty, 2005. Folding thermodynamics of peptides. *Biophys. J.* 88:1560–1569.
13. Irbäck, A., S. Mitternacht, and S. Mohanty, 2009. An effective all-atom potential for proteins. *BMC Biophys.* 2:2.

14. Huang, J., and A. D. MacKerell, 2018. Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 48:40–48.

15. MacKerell, A. D., M. Feig, and C. L. Brooks, 2004. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.* 126:698–699.

16. Mackerell, A. D., M. Feig, and C. L. Brooks, 2004. Extending the treatment of backbone energetics in protein force fields: Limitations of gas–phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* 25:1400–1415.

17. Irbäck, A., and S. Mohanty, 2006. PROFASI: a Monte Carlo simulation package for protein folding and aggregation. *J. Comput. Chem.* 27:1548–1555.

18. Mohanty, S., J. H. Meinke, and O. Zimmermann, 2013. Folding of Top7 in unbiased all-atom Monte Carlo simulations. *Proteins* 81:1446–1456.

19. Bille, A., K. S. Jensen, S. Mohanty, M. Akke, and A. Irbäck, 2019. Stability and local unfolding of SOD1 in the presence of protein crowders. *J. Phys. Chem. B* 123:1920–1930.

20. Li, Y., S. Mohanty, D. Nilsson, B. Hansson, K. Mao, and A. Irbäck, 2020. When a foreign gene meets its native counterpart: computational biophysics analysis of two *PgiC* loci in the grass *Festuca ovina*. *Sci. Rep.* 10:18752.

21. Mitternacht, S., I. Staneva, T. Härd, and A. Irbäck, 2010. Comparing the folding free-energy landscapes of Aβ42 variants with different aggregation properties. *Proteins* 78:2600–2608.

22. Mitternacht, S., I. Staneva, T. Härd, and A. Irbäck, 2011. Monte Carlo study of the formation and conformational properties of dimers of Aβ42 variants. *J. Mol. Biol.* 410:357–367.

23. Jónsson, S. Æ., S. Mohanty, and A. Irbäck, 2012. Distinct phases of free α-synuclein — a Monte Carlo study. *Proteins* 80:2169–2177.

24. Frishman, D., and P. Argos, 1995. Knowledge-based protein secondary structure assignment. *Proteins* 23:566–579.

25. Wang, G., and R. L. Dunbrack, 2003. PISCES: a protein sequence culling server. *Bioinformatics* 19:1589–1591.

26. Lyubartsev, A. P., A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, 1992. New approach to Monte Carlo calculation of the free energy: method of expanded ensembles. *J. Chem. Phys.* 96:1776–1783.

27. Marinari, E., and G. Parisi, 1992. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* 19:451–458.

28. Irbäck, A., and F. Potthast, 1995. Studies of an off-lattice model for protein folding: sequence dependence and improved sampling at finite temperature. *J. Chem. Phys.* 103:10298–10305.

29. Tesi, M. C., E. J. J. van Rensburg, E. Orlandini, and S. G. Whittington, 1996. Monte Carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys.* 82:155–181.

30. Hukushima, K., and K. Nemoto, 1996. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. (Jap)* 65:1604–1608.

31. Hansmann, U. H. E., 1997. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* 281:140–150.

32. Favrin, G., A. Irbäck, and F. Sjunnesson, 2001. Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. *J. Chem. Phys.* 114:8154–8158.

33. Karplus, M., 1959. Contact electron-spin coupling of nuclear magnetic moments. *J. Chem. Phys.* 30:11–15.

34. Vuister, G. W., and A. Bax, 1993. Quantitative *J* correlation: a new approach for measuring homonuclear three-bond $J(\mathrm{H}^{\mathrm{N}}\mathrm{H}^{\alpha})$ coupling constants in 15N-enriched proteins. *J. Am. Chem. Soc.* 115:7772–7777.

35. Neidigh, J. W., R. M. Fesinmeyer, and N. H. Andersen, 2002. Desiging a 20-residue protein. *Nat. Struct. Biol.* 9:425–430.

36. Liu, Y., Z. Liu, E. Androphy, J. Chen, and J. D. Baleja, 2004. Design and characterization of helical peptides that inhibit the E6 protein of papillomavirus. *Biochemistry* 43:7421–7431.

37. Bierzynski, A., P. S. Kim, and R. L. Baldwin, 1982. A salt bridge stabilizes the helix formed by isolated C-peptide of RNase A. *Proc. Natl. Acad. Sci. USA* 79:2470–2474.

38. Scholtz, J. M., D. Barrick, E. J. York, J. M. Stewart, and R. L. Baldwin, 1995. Urea unfolding of peptide helices as a model for interpreting protein unfolding. *Proc. Natl. Acad. Sci. USA* 92:185–189.

39. Thompson, P. A., W. A. Eaton, and J. Hofrichter, 1997. Laser temperature jump study of the helix$\rightleftharpoons$coil kinetics of an alanine peptide interpreted with a 'kinetic zipper' model. *Biochemistry* 36:9200–9210.

40. Lockhart, D. J., and P. S. Kim, 1993. Electrostatic screening of charge and dipole interactions with the helix backbone. *Science* 260:198–202.

41. Williams, S., T. P. Causgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff, and R. B. Dyer, 1996. Fast events in protein folding: helix melting and formation in a small peptide. *Biochemistry* 35:691–697.

42. Steinmetz, M. O., I. Jelesarov, W. M. Matousek, S. Honnappa, W. Jahnke, J. H. Missimer, S. Frank, A. T. Alexandrescu, and R. A. Kammerer, 2007. Molecular basis of coiled-coil formation. *Proc. Natl. Acad. Sci. USA* 104:7062–7067.

43. Sicheri, F., and D. S. C. Yang, 1995. Ice-binding structure and mechanism of an antifreeze protein from winter flounder. *Nature* 375:427–431.

44. Chakrabartty, A., V. S. Ananthanarayanan, and C. L. Hew, 1989. Structure-function relationships in a winter flounder antifreeze polypeptide. *J. Biol. Chem.* 264:11307–11312.

45. Honda, S., K. Yamasaki, Y. Sawada, and H. Morii, 2004. 10 residue folded peptide designed by segment statistics. *Structure* 12:1507–1518.

46. Pastor, M. T., M. López de la Paz, E. Lacroix, L. Serrano, and E. Pérez-Payá, 2002. Combinatorial approaches: a new tool to search for highly structured β-hairpin peptides. *Proc. Natl. Acad. Sci. USA* 99:614–619.

47. Blanco, F. J., G. Rivas, and L. Serrano, 1994. A short linear peptide that folds into a native stable β-hairpin in aqueous solution. *Nat. Struct. Mol. Biol.* 1:584–590.

48. Fesinmeyer, R. M., F. M. Hudson, and N. H. Andersen, 2004. Enhanced hairpin stability through loop design: The case of the protein G B1 domain hairpin. *J. Am. Chem. Soc.* 126:7238–7243.

49. Muñoz, V., P. A. Thompson, J. Hofrichter, and W. A. Eaton, 1997. Folding dynamics and mechanism of β-hairpin formation. *Nature* 390:196–199.

50. Cochran, A. G., N. J. Skelton, and M. A. Starovasnik, 2001. Tryptophan zippers: stable monomeric β-hairpins. *Proc. Natl. Acad. Sci. USA* 98:5578–5583.

51. Yang, W. Y., J. W. Pitera, W. C. Swope, and M. Gruebele, 2004. Heterogeneous folding of the trpzip hairpin: full atom simulation and experiment. *J. Mol. Biol.* 336:241–251.

52. Kortemme, T., M. Ramírez-Alvarado, and L. Serrano, 1998. Design of a 20-amino acid, three-stranded β-sheet protein. *Science* 281:253–256.

53. López de la Paz, M., E. Lacroix, M. Ramírez-Alvarado, and L. Serrano, 2001. Computer-aided design of β-sheet peptides. *J. Mol. Biol.* 312:229–246.

54. de Alba, E., J. Santorio, M. Rico, and M. A. Jimenez, 1999. De novo design of a monomeric three-stranded antiparallel β-sheet. *Protein Sci.* 8:854–865.

55. Dames, S. A., R. Aregger, N. Vajpai, P. Bernado, M. Blackledge, and S. Grzesiek, 2006. Residual dipolar couplings in short peptides reveal systematic conformational preferences of individual amino acids. *J. Am. Chem. Soc.* 128:13508–13514.

56. Palazzesi, F., M. K. Prakash, M. Bonomi, and A. Barducci, 2015. Accuracy of current all-atom force-fields in modeling protein disordered states. *J. Chem. Theory Comput.* 11:2–7.

57. Duan, Y., C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* 24:1999–2012.

58. Lindorff-Larsen, K., S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78:1950–1958.

59. Xiang, S., V. Gapsys, H.-Y. Kim, S. Bessonov, H.-H. Hsiao, S. Möhlmann, V. Klaukien, R. Ficner, S. Becker, H. Urlaub, R. Lührmann, B. de Groot, and M. Zweckstetter, 2013. Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. *Structure* 21:2162–2174.

60. Rauscher, S., V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller, 2015. Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J. Chem. Theory Comput.* 11:5513–5524.

61. MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102:3586–3616.

62. Roche, J., Y. Shen, J. H. Lee, J. Ying, and A. Bax, 2016. Monomeric Aβ1–40 and Aβ1–42 peptides in solution adopt very similar Ramachandran map distributions that closely resemble random coil. *Biochemistry* 55:762–775.