

Making Soup: Preparing and Validating Molecular Simulations of the Bacterial Cytoplasm

Leandro Oliveira Bortot,^{†,¶} Zahedeh Bashardanesh,^{‡,¶} and David van der
Spoel^{*,‡}

[†]*Laboratory of Biological Physics, School of Pharmaceutical Sciences of Ribeirão Preto,
University of São Paulo, Ribeirão Preto, Brazil*

[‡]*Science for Life Laboratory, Department of Cell and Molecular Biology. Uppsala
University, SE-751 05 Uppsala, Sweden*

[¶]*These authors contributed equally to this work*

E-mail: david.vanderspoel@icm.uu.se

Abstract

Biomolecular crowding affects the biophysical and biochemical behavior of macromolecules when compared to the dilute environment present in experiments made with isolated proteins. Computational modeling and simulation are useful tools to study how crowding affects the structural dynamics and biological properties of macromolecules. As computational power increased, modeling and simulating large scale all-atom explicit solvent models of the prokaryote cytoplasm become possible. In this work, we build an atomistic model of the cytoplasm of *Escherichia coli* composed of 1.5 million atoms and submit it to a total of 3 μ s of molecular dynamics simulations. The properties of biomolecules under crowding conditions are compared to those from simulations of the individual compounds under dilute conditions. The simulation model is found to be consistent with experimental data about the diffusion coefficient and stability of macromolecules under crowded conditions. In order to stimulate further work we provide a Python script and a set of files that enables other researchers to build their own *E. coli* cytoplasm models to address questions related to crowding.

Introduction

The intracellular environment is markedly different from the *in vitro* model systems from which most of our knowledge about the biochemical and biophysical behavior of macromolecules has been derived.¹ The *in vivo* microenvironment is heterogeneous with high concentrations of many different compounds, leading to spatial constraints that directly affects biological activities². This effect, called macromolecular crowding, influences the stability of many macromolecules because the equilibrium is shifted towards a stable native structure as a result of the entropic component given by the excluded volume effect.³ At the same time, an effective destabilization due to the enthalpic component of the Gibbs energy has been observed, due to replacing interactions with water by weaker interactions with the crowders.⁴⁻⁶ Additionally, crowding affects the structure and dynamics of water around proteins, which has a significant effect on their biological activities.^{7,8} While the structure and dynamics of biomolecules is well characterized *in vitro*, the understanding of the influences of crowded environments *in vivo* are still evolving. Techniques such as nuclear magnetic resonance^{9,10} and fluorescence spectroscopy¹¹⁻¹³ are among the most relevant to the field due to their ability to probe cells in a non destructive way.

A powerful class of methods is computational modeling and simulation. As the available computational power increases, models become increasingly complex and are better representations of reality. The most simple class of computational models apply Brownian Dynamics calculations to probe macromolecules immersed in a solution of crowders that are treated as hard spheres with varying radii to represent different types of molecules.¹⁴ These models usually use an implicit representation of the solvent.¹⁵ More recently, all-atom explicit solvent molecular dynamics simulations have been employed on crowded systems composed of several copies of one or a few macromolecules.¹⁶ However, these models do not reflect the true heterogeneity of biologically relevant crowded environments such as the cytoplasm. This was addressed by the development of a detailed model for the cytoplasm of the bacterium *Mycoplasma genitalium* containing 103 million atoms.¹⁷ This model, as well as a smaller

version with 13 million atoms, was submitted to molecular dynamics simulations.¹⁸ Due to the size of these systems, the simulation times were very short, 20 ns for the complete model and 140 ns for a smaller version, precluding an accurate characterization of dynamical effects that can take place in such a large scale. It was estimated, by extrapolating the advances in computational power in the last decades to the future, that the atomistic simulation of an entire *Escherichia coli* bacterium (about 10^{11} atoms) for 1 ns will be possible around the year 2034.¹⁹ This prediction may be too pessimistic, seeing that “only” a factor of 50 is lacking between the simulations by Feig and co-workers¹⁷ (from 2015) and the target of a whole bacterium for 1 ns. It was noted already some years ago, however, that even microsecond simulations may be too short to probe the dynamics of large complexes like viruses^{20,21} since dynamics is slower in larger systems in general. Although a meaningful simulation of a bacterium at the atomistic level does not seem to be within reach in the foreseeable future, simulations of parts of a cell may still be useful to address biological questions.

In this work, we report on an all-atom model of part of the cytoplasm of *Escherichia coli*, built to reflect the real biological system, in particular as regards a realistic composition of this “soup”. We discuss the difficulties that arise when building such a system and provide a set of Python scripts and files that can be used to build crowded systems with custom parameters and compounds. Additionally, we performed long molecular dynamics simulations of our cytoplasm model, reaching 3 μ s of total simulation time. To the best of our knowledge, this work is the first to explore the structural dynamics of an all-atom explicit solvent cytoplasm model with microsecond scale molecular dynamics simulations.

Materials and Methods

Initial structures

The protein and tRNA structures were downloaded from the Protein Data Bank (PDB).²² Proteins structures that were either from *E. coli* or its closest homologues (Table 1) were

selected for this work. In case of PDB IDs 1U22 (MetE protein) and 2EIP (Ppa protein), we used a loop-closure modeling tool based on the Random Coordinate Descent (RCD) method²³ to model missing residues. The four metabolites were parametrized using the General Amber Force Field (GAFF)²⁴ and Antechamber²⁵ using RESP charges²⁶ that were derived from electron density calculated at the HF/6-31G* theory level with Gaussian16. Properties of compounds modeled using the GAFF have been evaluated in a number of papers²⁷⁻³¹ and strengths and weaknesses are well understood. Here, GAFF is used for compatibility with the Amber99SB-ws force field³² used for the biomolecules.

General simulation setup

The cytoplasm model was built at a 30% biomolecular mass fraction and a physiological salt concentration of 0.15 mol/L KCl. Droplets containing each component of the model surrounded by water and counter-ions were inserted in an empty box in random positions and orientations. Since the volume of this box was initially bigger than what was strictly necessary to accomodate all components, the system was allowed to shrink in a short simulation. Details about how the simulation box was built and optimized are in the results section. All simulations were performed with the GROMACS 2018 package^{33,34} using the Amber99SB-ws force field,³² a modified version of the Amber99SB force field, in combination with the TIP4P/2005 water model³⁵ and improved K⁺ parameters³⁶ to avoid the formation of KCl crystals that is observed when using the default parameters of the Amber99SB force field.³⁷ All interactions were calculated explicitly inside a 1 nm radius and long range electrostatic interactions were treated using the particle mesh Ewald algorithm.³⁸ Corrections due to long range effects of dispersion interactions were made to both energy and pressure.³⁹ All chemical bonds were constrained at their equilibrium length using the parallel LINCS algorithm^{40,41} allowing an integration time step of 2 fs. Temperature was kept at 310 K by the v-rescale thermostat⁴² with a coupling constant of 0.5 ps. Pressure was kept at 1 bar using the Berendsen barostat⁴³ with the coupling constant of 10.0 ps during thermaliza-

tion steps and by the Parrinello-Rahman barostat⁴⁴ with the same coupling constant during production steps.

All systems were initially submitted to energy minimization with the steepest descent algorithm until convergence to machine (single) precision. For the cytoplasm model, a short shrinking step of 500 ps was performed in order to allow the box volume to adjust to its optimal value. For all systems, initial velocities were sampled from a Maxwell-Boltzmann distribution at 310 K and we ran a 1 ns NVT thermalization step followed by 10 ns of NPT equilibration. Production runs for the dilute systems, in which each crowder is simulated alone in a water box, had a length of 200 ns. The cytoplasm model was run for 1 μ s. All systems were simulated in three replicas. While the replicas for the diluted systems were generated by different initial velocities, each replica of the cytoplasm model is a completely different box in which the components of the model are in different positions and orientations.

Analyses

For statistics, the average and standard error of the three trajectories are reported. For values that are specific to each chain, we used the average over the individual chains in each trajectory of the replicas. Additionally, since there is more than one copy of some crowders in the cytoplasm models, average values and standard errors are calculated considering each copy of each replica. As in the dilute system, separate chains of each copy of the crowders were considered as independent samples when running analyses that focus on one single chain. The first half of all simulations was discarded as equilibration time and analyses used only the second half.

Analyses were done using GROMACS tools after removing artifacts generated by periodic boundary conditions from the trajectories. A Mean Square Displacement (MSD) analysis was performed to calculate the translational diffusion coefficient, which were extracted by a linear fit to MSD by averaging blocks with a length of 10 ns for dilute systems and 100 ns for cytoplasm system.³⁹ In principle, diffusion coefficient needs to be corrected for finite

size effects.^{45,46} Due to relatively large simulation boxes this correction is negligible. Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) calculations considered the C α atoms only for analysis in each trajectory. The chain interface area of a given oligomer in a given frame was calculated as the difference between the solvent exposed surface area (SASA) of the oligomer in that frame and the sum of the SASAs of each of its chains independently in that same frame. Other analyses, such as box volume and Radial Distribution Function (RDF), followed standard procedures.

Results

Cytoplasm model

In this section we describe the rationale behind the composition of our model, which consists of five fractions: protein, RNA, metabolites, water and ions. It should be noted that we did not add lipids and DNA to our model because we consider only elements that are free to diffuse through the cytosol. We gathered data from several sources in order to build a computational model that is representative of the cytoplasm of *Escheria coli*.^{15,47–49}

Protein fraction

Eight different proteins were selected that together account for 50% of the non-ribosomal proteins in the cytoplasm of *Escherichia coli*.⁴⁹ The two most abundant proteins, TufA and MetE, account for about 20% and 12% of the total protein abundance in *E. coli*, respectively, while the other six proteins contribute with less than 5% each (Table 1). The least abundant protein was present with 1 copy and the amount of each of the other proteins was taken to be an integer number of oligomers. This was done because the original oligomeric state for each protein was kept as reported in their crystallographic structure (Table 1).

Table 1: Protein, PDB ID, number N of proteins in the simulation. Fraction of the total abundance of non-ribosomal proteins in the cytosol of *E. coli* K12 and oligomeric state of each protein that was selected to compose our cytoplasm model.

Protein	PDB	N	Fraction [%]	State
Elongation factor TU (TufA)	1DG1 ⁵⁰	6	19.7	dimer
Cobalamin-independent methionine synthase (MetE)	1U22 ⁵¹	7	11.6	monomer
Isocitrate dehydrogenase (IcdA)	1P8F ⁵²	2	4.7	dimer
Alkyl hydroperoxide reductase subunit C (AhpC)	1YEP ⁵³	1	4.1	decamer
Cold-shock protein (CspC)	1MJC ⁵⁴	3	4.0	monomer
Pyrophosphatase (Ppa)	2EIP ⁵⁵	1	2.9	hexamer
Glyceraldehyde 3-phosphate dehydrogenase A (GapA)	1S7C ⁵⁶	1	2.1	tetramer
Enolase (Eno)	1E9I ⁵⁷	1	1.9	dimer
Total		48	51.1	

RNA fraction

Transporter RNAs (tRNAs) account for 74% of the dry weight of non-ribosomal RNAs.⁵⁸ Thus, we chose to model the RNA presence in the cytoplasm with tRNA molecules. Specifically, the tRNA(Phe) molecule as a representative of tRNAs due to the availability of a recent crystal structure (4YCO⁵⁹). The protein and RNA content of the total dry weight of *E.coli* is 55% and 2.9%, respectively.⁵⁸ That is, the total RNA weight corresponds to 5% of the total protein weight. This protein/RNA weight ratio was used to calculate the correct number of tRNA(Phe) molecules that were added to the cytoplasm model Table 2.

Metabolites fraction

The most abundant molecules of each metabolite class was used to represent the class, i.e. Glutamate (GLT) for amino acids, ATP for nucleotides, Fructose-2,6-Biphosphate (FBP) for central carbon intermediates and Glutathione (GSH) redox cofactors.⁴⁸ The total number of molecules was calculated considering data showing that the number of metabolite molecules in the cytoplasm of *E. coli* is ≈ 43 times higher than the number of proteins.⁴⁸ The copy number for each molecule was calculated from the ratios between their experimentally observed concentration in *E. coli*.⁴⁸

Water fraction

The number of water molecules was calculated according to the desired biomolecular concentration, which is a parameter of the cytoplasm model. Macromolecular concentration ranges from 300 to 400 g/L in biological systems such as *E. coli* cytoplasm,⁶⁰ the highest concentration corresponds to volume occupation as high as 40%.⁶¹ In our case, we choose a biomolecular concentration of 30%, that is, the number of molecules necessary to reach a ratio of total biomolecular mass to water mass of 30% was inserted into the cytoplasm model.

Inorganic ions fraction

Finally, Mg^{2+} was used as counter-ions for tRNA and ATP. K^+ and Cl^- were added to neutralize the charges of the simulation box and to reach the ionic strength of 0.150 mol/L by substitution of randomly selected water molecules.

Table 2: Number of copies for non-protein components of the cytoplasm model built at the biomolecular fraction of 30%.

Class	Name	Number
RNA	tRNA ^{Phe}	5
Metabolite	GLT	1436
	ATP	144
	FBP	225
	GSH	255
Solvent	Water	306221
Inorganic Ion	K^+	4602
	Mg^{2+}	400
	Cl^-	1320

Building the simulation box

All components can be put in the same simulation box by inserting each of them in random orientations in a cubic volume of side L that is initially empty. However, that is not a trivial process. We need to use a box size that is big enough to allow the random insertions to

succeed without structural overlapping and without creating artificial interactions between the elements by placing them too close from each other. On the other hand, as the box gets bigger, it gets harder to equilibrate the system because the empty space between the components will induce the barostat to reduce the box volume quickly during the simulation.

We devised an iterative process that solves both problems simultaneously. We start with a box size L that is too small to allow all components to fit in the box by random insertion. In our case, we start with $L = 30$ nm. Then, we allow 100 insertion trials for each element. If all trials fail for any of them, we start the whole process again with an empty cubic box that is larger by a step size dL of 1 nm. We repeat this process until all insertions succeed. In our case, all insertions succeeded after increasing L to 35 nm. Additionally, instead of adding only the protein, tRNA or metabolite molecules in the empty box in each trial, we actually add a droplet of water and counter ions in which the molecule of interest is embedded. Such droplets are taken from molecular dynamics simulations in which each component was previously equilibrated. The benefits of using such droplets are threefold: *i.* it acts as natural protecting layer that prevents artificial contacts between the components that could be created due to the random insertions. *ii.* it is a natural way to place water molecules and counter-ions in the simulation box around each component. *iii.* the components are already pre-equilibrated, which will help us to perform the equilibration of the whole cytoplasm model.

In order to do this, the droplets around each component are also constructed iteratively. The number of water molecules that we must place in the simulation box is known from the total biomolecular mass of the system and the desired biomolecular fraction, which is a parameter of the model. However, we don't know the thickness of the water layer around each component, l , that accounts for such amount of water molecules *a priori*. Since l depends on a series of factors such as the shape, size and abundance of each element, we define it iteratively. We start by taking a droplet of thickness $l = 3$ Å around all elements and counting the number of water molecules that we would add to the box with such droplets.

If it is smaller than the number we need, l is increased by $dl = 1 \text{ \AA}$. If it is larger, we reduce the droplet thickness by dl , reduce dl by 10 times and then increase l with the new smaller dl . We can carry this process until a user-defined precision cutoff, here 5%, for the number of waters in the system is satisfied.

After all droplets are successfully inserted in a simulation box by the iterative process described above, we proceed to add ions to neutralize the net charge of the system and to reach the desired ionic strength. Then, we perform energy minimization and a short simulation of 500ps in which all molecules of the box are free to move. In this step the box shrinks to its optimum size. In our case, the simulation box shrank from the initial box size of 35 nm to 22.9 nm, which corresponds to a volume change from 42875 nm³ to 12009 nm³. From this point, the system is ready to be submitted to the default simulation steps such as thermalization and production run (please check the materials section for details about the parameters used).

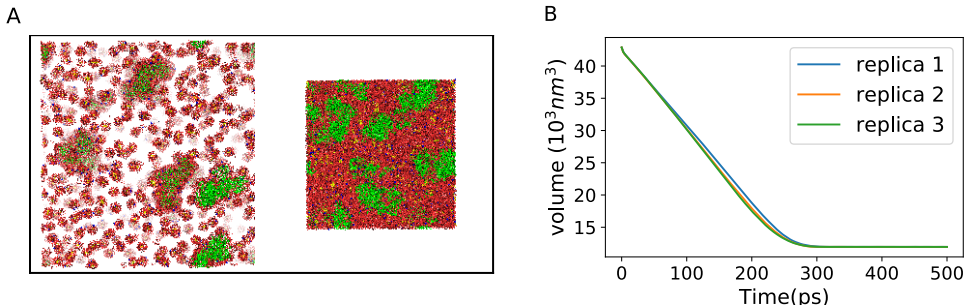


Figure 1: A) The initial simulation box with $L = 35.0$ nm and the box after the shrinking step, $L = 22.9$ nm. B) The evolution of the volume of the simulation box during the shrinking simulations. Each replica is shown in different colors.

Effects of macromolecular crowding

In order to investigate the effects of crowding in the *E. coli* cytoplasm model on the structural integrity and dynamics of its elements, we constructed three completely independent cytoplasm models that have different orientations for each of its elements. We submitted these systems, each composed of more than 1.5 million atoms, to molecular dynamics simu-

lations of 1 μ s. We also performed 200 ns molecular dynamics simulations for each isolated element (proteins and tRNA) in order to represent the non-crowded, i.e. dilute environment (please check the materials section for details about the parameters).

Translational diffusion

Under crowded conditions, molecules are confined to a smaller effective volume by the other components of the environment. The extent of this effect is evident when we compare the translational diffusion constant of each component in the cytoplasm model, D_{cyto} , and in a dilute system, D_{dil} (Fig. 2A). The ratio of these quantities is independent of the protein size and is always close to 0.13 (Fig. 2B). However, the effect of crowding on the translational diffusion of tRNA is two times larger. Its D_{cyto}/D_{dil} is 0.07, indicating that its movement is especially constrained in the crowded environment (Fig. 2B, crowder with 25 kDa), however, see below.

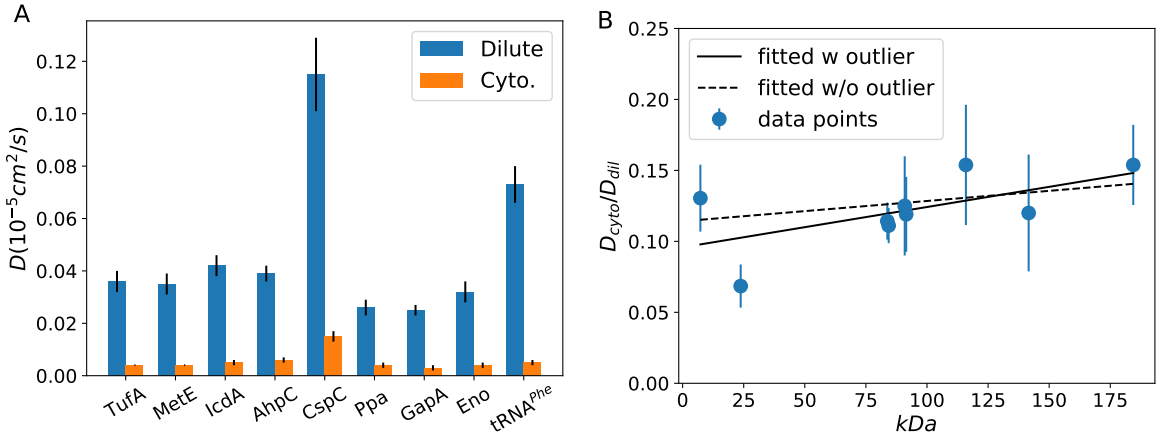


Figure 2: A) Diffusion coefficient of crowders from simulations under diluted conditions (blue bars) and cytoplasm simulations (orange bars). B) The ratio between diffusion coefficient obtained from simulations under both conditions. The crowders are sorted according to their sizes on the x-axis.

tRNA is aggregating

After inspecting the trajectories of the cytoplasm models, we found that the reason why the translational diffusion constant for tRNA was reduced to a greater extent than for the other

components of the cytoplasm model is that it was forming aggregates with Mg^{2+} , ATP and FBP (Fig. 3). Thus, we will omit the tRNA molecules in the analyses about the structural integrity of the crowders. In the next sections we will further investigate such aggregation and we will show a way to prevent it.

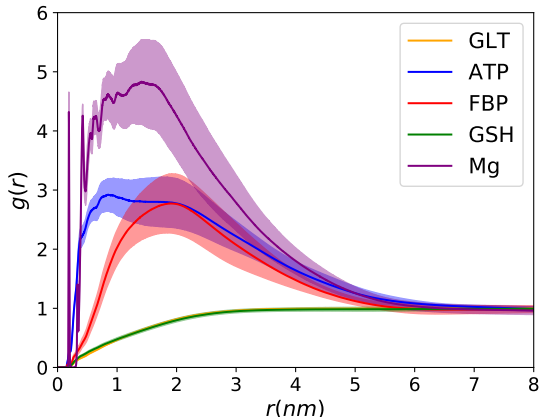


Figure 3: Radial distribution function showing the probability of finding metabolites or Magnesium cations around tRNA molecules in the cytoplasm model. Note that ATP and FBP are much more concentrated around tRNA than the other metabolites, GSH and GLT.

Structural integrity

The structure of the individual chains of the crowders that were used in our cytoplasm model were not affected to a great extent by the crowded environment. The values of their root mean square deviation considering only their C α in the cytoplasm model increased by less than 11% when compared to the dilute condition, except for TufA, MetE and IcdA, which have RMSD values 105, 25 and 33% higher in the crowded environment than in the dilute condition (Fig. 4A). Visual inspection of the final frames of the simulations of these crowders in the cytoplasm model and in dilute condition revealed that, for TufA, the conformational differences are on the β -hairpin around residues 40-62. This region is called Switch I and is known to be flexible and to display conformational changes due to interactions with the ribosome or other partners.⁵⁰ For MetE, there are differences in the conformation of the loop around residues 450-460 and in the C-terminal α -helices. This loop is known to be flexible and was not modeled in the crystallographic structure due to such flexibility.⁵¹ For IcdA,

the differences are located at the N-terminal coil formed by residues 1-8.

The effect of crowded environment on the structural integrity of oligomers can be evaluated by their chain interface area as well. This quantity is the solvent accessible surface area that is removed from access to the solvent by oligomerization. Our results show that oligomers are not significantly disturbed in the cytoplasm model when compared to the diluted condition. Similarly to what can be seen by the RMSD values, which reflect the structural integrity of individual chains, the oligomers that undergo the most change is TufA (Fig. 4B).

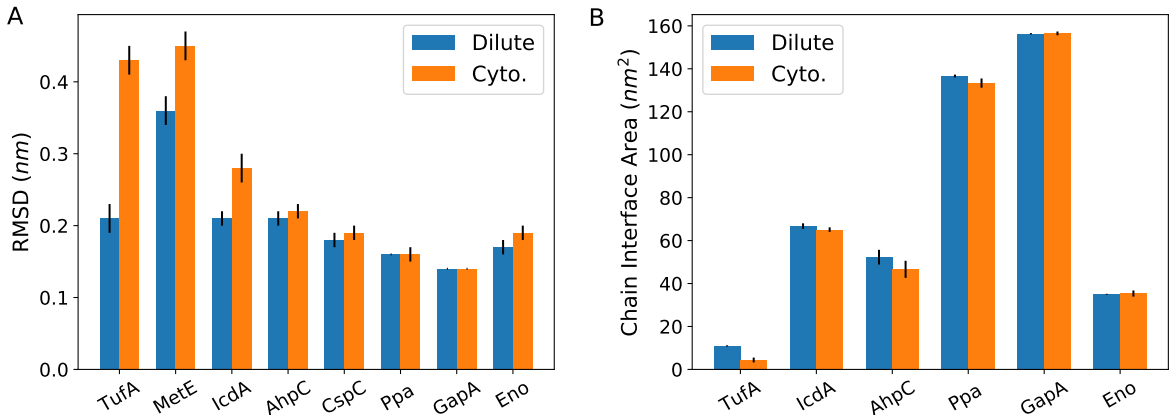


Figure 4: A) RMSD for all protein crowders from simulations under diluted conditions (blue bars) and in the cytoplasm model (orange bars). This value reflects the structural integrity of individual chains of each crowder. B) Chain interface area for oligomers under diluted conditions (blue bars) and in the cytoplasm model (orange bars). This value reflects the structural integrity of oligomers of the crowders.

Structural dynamics

Root mean square fluctuation (RMSF) calculations show that, overall, the shape of residue-wise flexibility profile of the proteins is not affected by crowding, which is in agreement with our results showing that their structural integrity does not change significantly in the cytoplasm model in most cases (Fig. 5). Accordingly, we can see significant changes in the shape of the flexibility profile for regions that undergo conformational changes, such as the Switch I β -hairpin of TufA around residues 40-62 and the flexible loop of MetE around residues 450-460 (Fig. 5).

In most cases (TufA, MetE, AhpC, CspC and Ppa), the residue-wise RMSF values are higher in the cytoplasm model than under diluted conditions, which indicates that proteins are more flexible in the cytoplasm model (Fig. 5). This is in agreement with computational and experimental data showing that nonspecific interactions with crowders can lead to destabilization of proteins.⁴⁻⁶

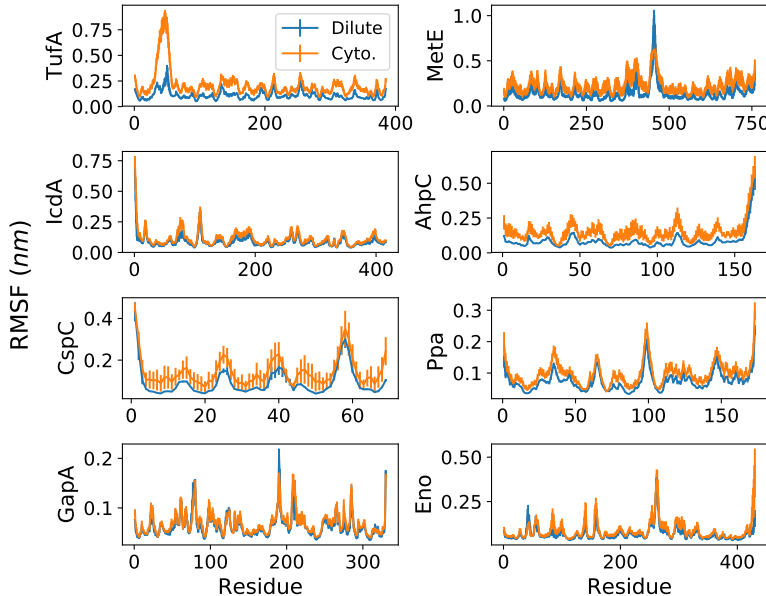


Figure 5: Root mean squared fluctuations (RMSF) from simulations under diluted conditions (blue bars) and in the cytoplasm model (orange bars). The x-axis shows the residues for each chain of proteins.

Aggregation can be avoided by protonating the metabolites

The initial detection of aggregation in the cytoplasm model was due to the outlier behavior of the translational diffusion constant of tRNA when compared to the other crowders. However, tRNA was not the cause of this phenomenon. Instead, it was triggered by the exaggerated interactions between the phosphate-containing metabolites, ATP and FBP, with Mg^{2+} (Fig. 6A). tRNA gets involved in the aggregates because Mg^{2+} is present in the box only as counter-ion for ATP and tRNA, and so tRNA is one of the few “sources” of Mg^{2+} in the whole system. We looked for ways of avoid aggregation by using simulations of small systems composed of these metabolites in high concentration, Mg^{2+} and water (please check

the methods section for details).

We found that completely protonating the phosphate groups of ATP and FBP is enough to prevent their aggregation with Mg^{2+} . While RDFs show that ATP^{3-} and FBP^{4-} aggregate around Mg^{2+} (Fig. 6A), this is not observed for ATPH_3 and FBPH_4 (Fig. 6B-C).

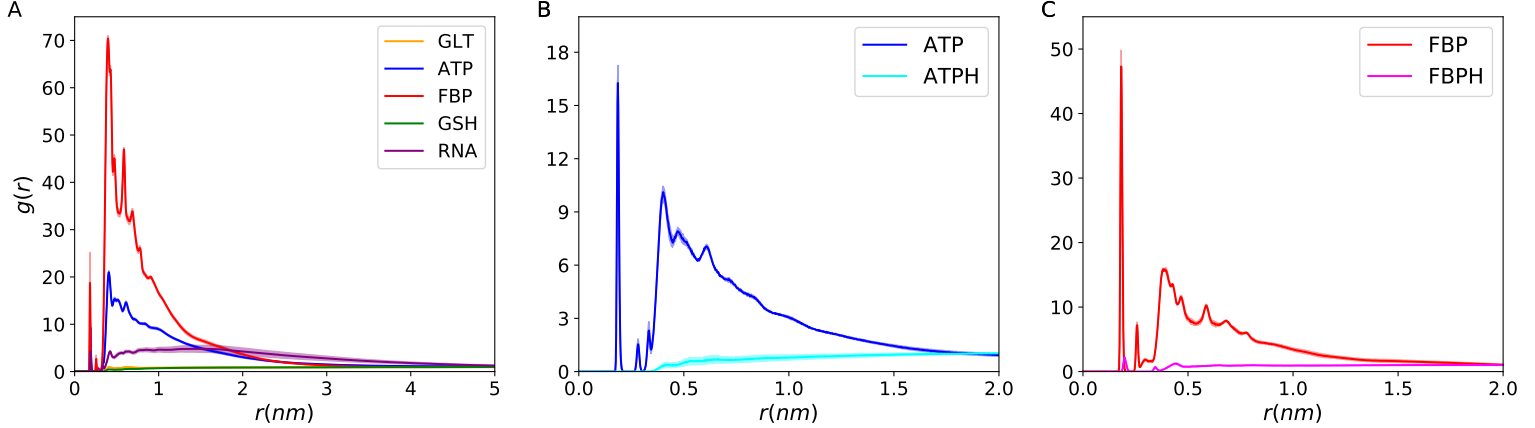


Figure 6: RDFs centered around Mg^{2+} showing the probability of finding A) metabolites in the cytoplasm model, B) ATP^{3-} and ATPH_3 or C) FBP^{4-} and FBPH_4 in small independent systems (please see the methods section for details).

Discussion

The main difference in behavior of the macromolecules that compose our model between the crowded system and the dilute condition is their translational diffusion coefficient, which drops by around 85% for the proteins and 93% for the tRNA (Fig. 2A). The slope of the line generated by fitting a linear regression model to values of $D_{\text{cyto}}/D_{\text{dil}}$ is close to zero (slope = 0.00028), showing that the drop in diffusion coefficient is independent of the size of the macromolecules. We found that the aggregation of ATP^{3-} and FBP^{4-} around Mg^{2+} , which was used as counter-ion for tRNA, leads to an exaggerated confinement of the tRNA molecules. Therefore, tRNA is considered an outlier in the diffusion calculations. In fact, the slope of the linear regression line fitted to $D_{\text{cyto}}/D_{\text{dil}}$ is reduced by 50% and even closer to zero (slope = 0.00014) after removing the diffusion coefficient of tRNA from the data. These data are in agreement with experimental results showing that the translational

diffusion coefficient of Green Fluorescent Protein (GFP) in *E. coli*, as measured by fluorescence recovery after photo-bleaching (FRAP), decreased 90% when compared to *in vitro* measurements.^{62,63}

It has been demonstrated that the force fields that are most widely used to perform molecular dynamics simulations of macromolecules yield different results for highly concentrated amino acid solutions.⁶⁴ Moreover, it has been reported that simulations of a highly concentrated environment leads to aggregation when additive force fields are used^{65–67} and several approaches have been proposed to eliminate this problem.^{32,68,69} Apart from protein-water interactions, the London-dispersion coefficients in the force fields have been investigated as well but their effect on protein-protein interactions seems to be minor.^{69–71} A comparison at the amino acid level shows that none of the Amber family of force fields combined with different water model can simultaneously produce high accuracy Gibbs energies of hydration and amino acid side chain mobility.⁷² Of the different combinations tested in that work the Amber99sb-ws³² in conjunction with the TIP4P/2005 water model³⁵ was deemed the best compromise.⁷² Nevertheless, in a related paper we find signs that the increased interaction between protein and water may be slightly too strong as it leads to partial unfolding of a protein.⁷³

Previous all-atom simulations of large scale cytoplasm models used parameters from the CHARMM family of force fields.¹⁸ In this work we use a modified version of the AMBER FF99SB force field in which the solvent-solute interactions are 10% stronger³² to study crowding effects on a cytoplasm model. This force field was chosen because it was corrected to reproduce experimental data when there are many proteins in the same simulation box. This force field or similarly adjusted one were also used for studying crowding effect on simpler systems of high concentration of proteins.^{67,73,74} All-atom explicit solvent molecular dynamics simulations have been reported for large-scale cytoplasm models of *Mycoplasma genitalium* composed of 103- and 12-million atoms that were simulated for 20 and 60 ns, respectively.¹⁸ The cytoplasm model described in this work has 1.5 million atoms and was

simulated for the total of 3 μ s, an unprecedented time scale to study crowding effects on a heterogeneous cytoplasm model in atomic detail.

Three crowders (TufA, MetE and IcdA) displayed greater deviation from their crystallographic structure in the cytoplasm model when compared to the dilute condition, indicating that there was some degree of local unfolding (Fig. 4A). Interestingly, oligomers were marginally less stable in the cytoplasm model than in the dilute condition. (Fig. 4B) and the flexibility profiles calculated considering the C α for each chain of the elements of our cytoplasm model show that, in the crowded condition, most of the proteins (TufA, MetE, AhpC, CspC and Ppa) are more flexible than in the dilute condition (Fig. 5). This is in agreement with previous simulations⁴ and experiment^{5,6} showing that crowders can destabilize the native structure of proteins due to the numerous unspecific short-lived interactions that are formed between them in a crowded environment. A study of protein properties at increasing concentration showed only few interactions between proteins of the same type.⁷³ The only specific interaction we anticipate between the components of our cytoplasm model is between TufA and tRNA (PDB ID 1TTT is an example of such complex).⁷⁵ The TufA-tRNA complex is not present in our simulation boxes even after 1 μ s of simulation starting from random orientations.

Magnesium cations were added to our cytoplasm model as counter-ions for tRNA and ATP. However, as the simulation progressed, FBP and ATP aggregated around Mg²⁺ and, by consequence, around tRNA. After identifying this problem, we used simulations of small systems containing only the metabolites, Mg²⁺ and water to test different strategies to prevent aggregation in future studies using long simulations of cytoplasm models. We found that protonating the phosphate-containing metabolites is enough to reach this goal (Fig. 6). Even though ATPH₃ and FBPH₄ are not the biologically relevant forms of these molecules in physiological pH, using their protonated forms may be an acceptable compromise between parametrization complexity and practical results. The biggest impact of this approach is on long range interactions of these metabolites with other molecules due to the neutraliza-

tion of their net charge. However, molecules in a crowded environment are restrained to smaller volumes and shorter inter molecular distances, which mitigates the effect of charge neutralization. It has been suggested that charged species should employ scaled charges in order to be compatible with the empirical water models.^{76,77} Due to the high local density of negative charges in FBP and ATP this might indeed be an alternative solution, barring use of polarizable models.⁷⁸⁻⁸¹

To the best of our knowledge, there is only one published study on large scale cytoplasm models that included metabolites with all-atom explicit solvent molecular dynamics.¹⁸ Although these authors did not report aggregation, they mention that the translational diffusion of highly charged phosphate-containing metabolites is “much slower” than expected in their system. They attributed this observation to non-specific interactions between such metabolites and proteins. Maybe aggregation also contributed to their slower diffusion but it was not evident during their 140 ns simulation.

In this work we built an all-atom explicit solvent cytoplasm model composed of proteins, tRNA, metabolites, inorganic ions and water for a total of 1.5 million atoms (Table 2). The identity and amount of each component of the model was chosen based on experimental data about the composition of the cytoplasm of *E. coli*.^{15,47-49}

We submitted this model to a total of 3 μ s of molecular dynamics simulations using a modified AMBER FF99SB force field that was tuned to reproduce experimental data of systems composed of several macromolecules.³² The model was validated by comparing to experimental data about the reduction of the translational diffusion coefficient of proteins in the cytoplasm of *E. coli* when compared to the dilute condition (Figure 2).^{62,63} Additionally, our data is in agreement with recent simulation work and experimental data that show that the unspecific interactions between macromolecules in a crowded environment can destabilize them.⁴⁻⁶ Specifically, we can see that, despite the overall stability of our cytoplasm model even in the μ s timescale, the proteins are slightly less stable in the cytoplasm model than in dilute condition (Figures 4 and 5).

As part of this work we are providing Python scripts with a set of structure and topology files that any researcher can use to build their own *E. coli* cytoplasm models. Such models can be used to investigate research questions about the effects of crowding on specific systems of interest or about crowding itself. With the files we are providing it is possible, for example, to add a probe protein to investigate the effect of crowding on it, to add new macromolecular or small crowders, evaluate the effects of temperatures and change the biomolecular concentration to increase or decrease the intensity of the crowding effect. We highlight that, since we have already validated it, researchers can take advantage from building their own cytoplasm models even if they do not have access to the infrastructure that is necessary to perform long molecular dynamics simulations. With modest computational power it is possible, for example, to build many independent snapshots of the cytoplasm of *E. coli*. All files are publicly available at <http://github.com/dspoel/soup> facilitating contributions to the scripts and data files from other workers in the field.

Acknowledgement

The Swedish research council is acknowledged for a grant of computer time (SNIC2018-2-42) through the High Performance Computing Center North in Umeå, Sweden. Funding from eSENCE - The e-Science Collaboration (Uppsala-Lund-Umeå, Sweden) is gratefully acknowledged.

References

- (1) Feig, M.; Yu, I.; Wang, P.-h.; Nawrocki, G.; Sugita, Y. Crowding in Cellular Environments at an Atomistic Level from Computer Simulations. *J. Phys. Chem. B.* **2017**, *121*, 8009–8025, PMID: 28666087.

- (2) Ostrowska, N.; Feig, M.; Trylska, J. Modeling crowded environment in molecular simulations. *Front. Mol. Biosci.* **2019**, *6*, 86.
- (3) Cheung, M. S.; Klimov, D.; Thirumalai, D. Molecular crowding enhances native state stability and refolding rates of globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4753–4758.
- (4) Feig, M.; Sugita, Y. Variable interactions between protein crowders and biomolecular solutes are important in understanding cellular crowding. *J. Phys. Chem. B.* **2011**, *116*, 599–605.
- (5) Miklos, A. C.; Sarkar, M.; Wang, Y.; Pielak, G. J. Protein crowding tunes protein stability. *J. Amer. Chem. Soc.* **2011**, *133*, 7116–7120.
- (6) Wang, Y.; Sarkar, M.; Smith, A. E.; Krois, A. S.; Pielak, G. J. Macromolecular crowding and protein stability. *J. Amer. Chem. Soc.* **2012**, *134*, 16614–16618.
- (7) Harada, R.; Sugita, Y.; Feig, M. Protein crowding affects hydration structure and dynamics. *J. Amer. Chem. Soc.* **2012**, *134*, 4842–4849.
- (8) King, J. T.; Arthur, E. J.; Brooks III, C. L.; Kubarych, K. J. Crowding induced collective hydration of biological macromolecules over extended distances. *J. Amer. Chem. Soc.* **2013**, *136*, 188–194.
- (9) Reckel, S.; Hänsel, R.; Löhr, F.; Dötsch, V. In-cell NMR spectroscopy. *Prog. Nucl. Mag. Res. Sp.* **2007**, *2*, 91–101.
- (10) Pielak, G. J.; Li, C.; Miklos, A. C.; Schlesinger, A. P.; Slade, K. M.; Wang, G.-F.; Zigoneanu, I. G. Protein nuclear magnetic resonance under physiological conditions. *Biochemistry* **2008**, *48*, 226–234.
- (11) Ignatova, Z.; Gierasch, L. M. Monitoring protein stability and aggregation in vivo by real-time fluorescent labeling. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 523–528.

- (12) Xie, X. S.; Choi, P. J.; Li, G.-W.; Lee, N. K.; Lia, G. Single-molecule approach to molecular biology in living bacterial cells. *Annu. Rev. Biophys.* **2008**, *37*, 417–444.
- (13) English, B. P.; Hauryliuk, V.; Sanamrad, A.; Tankov, S.; Dekker, N. H.; Elf, J. Single-molecule investigations of the stringent response machinery in living bacterial cells. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, E365–E373.
- (14) Ando, T.; Skolnick, J. Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 18457–18462.
- (15) McGuffee, S. R.; Elcock, A. H. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.* **2010**, *6*, e1000694.
- (16) Wang, P.-H.; Yu, I.; Feig, M.; Sugita, Y. Influence of protein crowder size on hydration structure and dynamics in macromolecular crowding. *Chem. Phys. Lett.* **2017**, *671*, 63–70.
- (17) Feig, M.; Harada, R.; Mori, T.; Yu, I.; Takahashi, K.; Sugita, Y. Complete atomistic model of a bacterial cytoplasm for integrating physics, biochemistry, and systems biology. *J. Mol. Graph. Mod.* **2015**, *58*, 1–9.
- (18) Yu, I.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *eLife* **2016**, *5*.
- (19) van Gunsteren, W. F. et al. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 4064–4092.
- (20) Larsson, D. S. D.; Liljas, L.; van der Spoel, D. Virus capsid dissolution studied by microsecond molecular dynamics simulations. *PLoS Comput. Biol.* **2012**, *8*, e1002502.
- (21) Larsson, D. S. D.; van der Spoel, D. Screening for the Location of RNA using the

- Chloride Ion Distribution in Simulations of Virus Capsids. *J. Chem. Theory Comput.* **2012**, *8*, 2474–2483.
- (22) Westbrook, J.; Feng, Z.; Jain, S.; Bhat, T. N.; Thanki, N.; Ravichandran, V.; Gilliland, G. L.; Bluhm, W.; Weissig, H.; Greer, D. S.; Bourne, P. E.; and Helen M. Berman, The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* **2002**, *30*, 245–248.
- (23) Chys, P.; Chacón, P. Random coordinate descent with spinor-matrices and geometric filters for efficient loop closure. *J. Chem. Theory Comput.* **2013**, *9*, 1821–1829.
- (24) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General AMBER Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (25) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Antechamber, an Accessory Software Package for Molecular Mechanical Calculations. *J. Comput. Chem.* **2005**, *25*, 1157–1174.
- (26) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- (27) Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Compressibility, Expansion Coefficient and Dielectric Constant. *J. Chem. Theory Comput.* **2012**, *8*, 61–74.
- (28) Sprenger, K.; Jaeger, V. W.; Pfaendtner, J. The general AMBER force field (GAFF) can accurately predict thermodynamic and transport properties of many ionic liquids. *J. Phys. Chem. B.* **2015**, *119*, 5882–5895.

- (29) Fischer, N. M.; van Maaren, P. J.; Ditz, J. C.; Yildirim, A.; van der Spoel, D. Properties of liquids in Molecular Dynamics Simulations with explicit long-range Lennard-Jones interactions. *J. Chem. Theory Comput.* **2015**, *11*, 2938–2944.
- (30) Zhang, J.; Tuguldur, B.; van der Spoel, D. Force field benchmark II: Gibbs energy of solvation of organic molecules in organic liquids. *J. Chem. Inf. Model.* **2015**, *55*, 1192–1201.
- (31) van der Spoel, D.; Ghahremanpour, M. M.; Lemkul, J. Small Molecule Thermochemistry: A Tool For Empirical Force Field Development. *J. Phys. Chem. A* **2018**, *122*, 8982–8988.
- (32) Best, R. B.; Zheng, W.; Mittal, J. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **2014**, *10*, 5113–5124.
- (33) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.
- (34) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.
- (35) Abascal, J. L. F.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.
- (36) Dang, L. X.; Kollman, P. A. Free energy of association of the K^+ : 18-crown-6 complex in water: a new molecular dynamics study. *J. Phys. Chem.* **1995**, *99*, 55–58.

- (37) Auffinger, P.; Cheatham, T. E.; Vaiana, A. C. Spontaneous formation of KCl aggregates in biomolecular simulations: a force field issue? *J. Chem. Theory Comp.* **2007**, *3*, 1851–1859.
- (38) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8592.
- (39) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publications: Oxford, 1987.
- (40) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
- (41) Hess, B.; Kutzner, C.; Van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
- (42) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
- (43) Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (44) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.
- (45) Yeh, I.-C.; Hummer, G. System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions. *J. Phys. Chem. B.* **2004**, *108*, 15873–15879.
- (46) Abraham, M. J.; Hess, B.; the GROMACS development team, Gromacs User Manual version 2016. 2018.

- (47) Dong, H.; Nilsson, L.; Kurland, C. G. Co-variation of trna abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* **1996**, *260*, 649–663.
- (48) Bennett, B. D.; Kimball, E. H.; Gao, M.; Osterhout, R.; Van Dien, S. J.; Rabinowitz, J. D. Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.* **2009**, *5*, 593–599.
- (49) Link, A. J.; Robison, K.; Church, G. M. Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **1997**, *18*, 1259–1313.
- (50) Abel, K.; Yoder, M. D.; Hilgenfeld, R.; Journak, F. An α to β conformational switch in EF-Tu. *Structure* **1996**, *4*, 1153–1159.
- (51) Ferrer, J.-L.; Ravanel, S.; Robert, M.; Dumas, R. Crystal structures of cobalamin-independent methionine synthase complexed with zinc, homocysteine, and methyltetrahydrofolate. *J. Biol. Chem.* **2004**, *279*, 44235–44238.
- (52) Mesecar, A. D.; Koshland Jr, D. E. Structural biology: A new model for protein stereospecificity. *Nature* **2000**, *403*, 614–616.
- (53) Parsonage, D.; Youngblood, D. S.; Sarma, G. N.; Wood, Z. A.; Karplus, P. A.; Poole, L. B. Analysis of the link between enzymatic activity and oligomeric state in AhpC, a bacterial peroxiredoxin. *Biochemistry* **2005**, *44*, 10583–10592.
- (54) Schindelin, H.; Jiang, W.; Inouye, M.; Heinemann, U. Crystal structure of CspA, the major cold shock protein of *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5119–5123.
- (55) Kankare, J.; Salminen, T.; Lahti, R.; Cooperman, B.; Baykov, A.; Goldman, A. Structure of *Escherichia coli* inorganic pyrophosphatase at 2.2 Å resolution. *Acta Crystallogr. D* **1996**, *52*, 551–563.

- (56) Shin, D.; Thor, J.; Yokota, H.; Kim, R.; Kim, S. Crystal structure of MES buffer bound form of glyceraldehyde 3-phosphate dehydrogenase from *Escherichia coli*. To be published.
- (57) Kühnel, K.; Luisi, B. F. Crystal structure of the *Escherichia coli* RNA degradosome component enolase. *J. Mol. Biol.* **2001**, *313*, 583–592.
- (58) Phillips, R.; Theriot, J.; Kondev, J.; Garcia, H. *Physical biology of the cell*; Garland Science, 2012.
- (59) Byrne, R. T.; Jenkins, H. T.; Peters, D. T.; Whelan, F.; Stowell, J.; Aziz, N.; Kasatsky, P.; Rodnina, M. V.; Koonin, E. V.; Konevega, A. L. Major reorientation of tRNA substrates defines specificity of dihydrouridine synthases. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 6033–6037.
- (60) Zimmerman, S. B.; Trach, S. O. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of *Escherichia coli*. *J. Mol. Biol.* **1991**, *222*, 599–620.
- (61) Ellis, R. J.; Minton, A. P. Join the crowd. *Nature* **2003**, *425*, 27–28.
- (62) Elowitz, M. B.; Surette, M. G.; Wolf, P.-E.; Stock, J. B.; Leibler, S. Protein Mobility in the Cytoplasm of *Escherichia coli*. *J. Bacteriol.* **1999**, *181*, 197–203.
- (63) Konopka, M. C.; Shkel, I. A.; Cayley, S.; Record, M. T.; Weisshaar, J. C. Crowding and confinement effects on protein diffusion in vivo. *J. Bacteriol.* **2006**, *188*, 6115–6123.
- (64) Andrews, C. T.; Elcock, A. H. Molecular dynamics simulations of highly crowded amino acid solutions: comparisons of eight different force field combinations with experiment and with each other. *J. Chem. Theory Comput.* **2013**, *9*, 4585–4602.
- (65) Petrov, D.; Zagrovic, B. Are current atomistic force fields accurate enough to study proteins in crowded environments? *PLoS Comput. Biol.* **2014**, *10*, e1003638.

- (66) Abriata, L. A.; Dal Peraro, M. Assessing the potential of atomistic molecular dynamics simulations to probe reversible protein-protein recognition and binding. *Sci. Rep.* **2015**, *5*, 10549.
- (67) Nawrocki, G.; Wang, P.-h.; Yu, I.; Sugita, Y.; Feig, M. Slow-down in diffusion in crowded protein solutions correlates with transient cluster formation. *J. Phys. Chem. B.* **2017**, *121*, 11072–11084.
- (68) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113–5123.
- (69) Bashardanesh, Z.; van der Spoel, D. Impact of Dispersion Coefficient on Simulations of Proteins and Organic Liquids. *J. Phys. Chem. B.* **2018**, *122*, 8018–8027.
- (70) Mohebifar, M.; Johnson, E. R.; Rowley, C. N. Evaluating Force-Field London Dispersion Coefficients Using the Exchange-Hole Dipole Moment Model. *J. Chem. Theory Comput.* **2017**, *13*, 6146–6157.
- (71) Walters, E.; Mohebifar, M.; Johnson, E. R.; Rowley, C. N. Evaluating the London Dispersion Coefficients of Protein Force Fields Using the Exchange-Hole Dipole Moment Model. *J. Phys. Chem. B.* **2018**, *122*, 6690–6701.
- (72) Zhang, H.; Yin, C.; Jiang, Y.; van der Spoel, D. Force Field Benchmark of Amino acids: I. Hydration and Diffusion in Different Water Models. *J. Chem. Inf. Model.* **2018**, *58*, 1037–1052.
- (73) Bashardanesh, Z.; Elf, J.; Zhang, H.; van der Spoel, D. Rotational and translational diffusion of proteins as a function of concentration. *ACS Omega* **2019**, Submitted.
- (74) Nawrocki, G.; Karaboga, A.; Sugita, Y.; Feig, M. Effect of protein–protein interactions

- and solvent viscosity on the rotational diffusion of proteins in crowded environments. *Phys. Chem. Chem. Phys.* **2019**, *21*, 876–883.
- (75) Nissen, P.; Kjeldgaard, M.; Thirup, S.; Polekhina, G.; Reshetnikova, L.; Clark, B. F.; Nyborg, J. Crystal structure of the ternary complex of Phe-tRNAPhe, EF-Tu, and a GTP analog. *Science* **1995**, *270*, 1464–1472.
- (76) Leontyev, I.; Stuchebrukhov, A. Electronic continuum model for molecular dynamics simulations. *J. Chem. Phys.* **2009**, *130*, 02B609.
- (77) Leontyev, I.; Stuchebrukhov, A. Accounting for electronic polarization in non-polarizable force fields. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2613–2626.
- (78) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr., R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.
- (79) Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, J., Alexander D. Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J. Chem. Theory Comput* **2013**, *9*, 5430–5449.
- (80) Ghahremanpour, M. M.; van Maaren, P. J.; Caleman, C.; Hutchison, G. R.; van der Spoel, D. Polarizable Drude Model with s-type Gaussian or Slater Charge Density for General Molecular Mechanics Force Fields. *J. Chem. Theory Comput.* **2018**, *14*, 5553–5566.
- (81) Walz, M. M.; Ghahremanpour, M. M.; van Maaren, P. J.; van der Spoel, D. Phase-Transferable Force Field for Alkali Halides. *J. Chem. Theory Comput.* **2018**, *14*, 5933–5948.