ABSTRACT

OWEIDA, THOMAS JOSEPH. Predicting the Structure and Self-Assembly of Polyelectrolytes through Molecular Modeling and Machine Learning. (Under the direction of Dr. Yaroslava G. Yingling).

Polyelectrolytes play a key role in designing responsive materials as they exhibit a controllable response to external stimuli such changes in salt or pH. DNA is a unique polyelectrolyte that displays molecular recognition through Watson-Crick base pairs allowing DNA-based materials to have tunable properties along with programmable self-assembly at the nanoscale. However, the behavior and structure of polyelectrolytes in solution is not as well understood compared to their neutral counterparts. In particular, the flexibility and diverse molecular interactions found in ssDNA limit the resolution and characterization of ssDNA using experimental techniques. This work investigates the structure of ssDNA in solution using simulations with atomistic resolution. Simulation results are connected back to experiments through a data-driven approach that serves to bridge gaps in the differing sizes and timescales between techniques. Procedures for validating all-atom simulations of ssDNA are set forth along with recommendations and improvements for future protocols and analysis. Simulations of ssDNA are also performed to identify how its degradation in biological environments is related to its structure and dynamics.

A more general molecular modeling approach is performed to analyze the self-assembly of polyelectrolyte block copolymers. These polyelectrolyte-based block copolymers exhibit properties of surfactants while maintaining the salt responsive behavior of the polyelectrolyte. This portion of the study focuses on understanding the influence of the block copolymer's molecular architecture and solvent ionic strength on the equilibrium morphologies, aggregation numbers, and size of micelles. Ultimately, a generalized phase diagram is provided.

© Copyright 2021 by Thomas J. Oweida

All Rights Reserved

Predicting the Structure and Self-Assembly of Polyelectrolytes through Molecular Modeling and Machine Learning

by
Thomas Joseph Oweida

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Materials Science and Engineering

Raleigh, North Carolina
2021

APPROVED BY:

Dr. Yaroslava G. Yingling Committee Chair	Dr. Brian Reich
Dr. Albena Ivanisevic	Dr. Doug Irving

DEDICATION

"Everything that living things do can be understood in terms of the jigglings and wigglings of atoms." – Richard P. Feynman

To my family and friends who have supported me throughout life.

BIOGRAPHY

I was born and raised in Charlotte, North Carolina with three older sisters and a younger brother. I was fortunate enough to have a loving mother, Lynne Oweida, and father, Sami Oweida who provided a platform in which I could explore various avenues and interests of life. I took to sports where I settled down as a competitive gymnast, ultimately retiring after one season at the United States Air Force Academy. Although my parents encouraged my active lifestyle, there was always an emphasis on academic achievement and pursuing higher education.

My higher education began at the Air Force Academy in Colorado Springs, Colorado in 2012. However, after completing my first year, I transferred and enrolled at North Carolina State University (NCSU) where I eventually received my bachelor's degree in Materials Science and Engineering. During my enrollment, I performed research in various labs around campus; the bulk of my research centering on theory and simulation in the Yingling Research Group. After graduation, I was presented with the opportunity to join Dr. Yingling's group as a graduate student, which I happily accepted.

The theme and tools I used for research varied significantly once I transitioned to life as a graduate student. I became a SEAS fellow, which focused on adopting data-driven methodologies to handle materials data. In addition, I adopted new simulation techniques to study a diverse range of materials. Under the guidance of Dr. Yingling, I was able to merge multidisciplinary approaches into a unified research approach.

ACKNOWLEDGMENTS

I would like to thank my advisor, lab-mates, mentors, family, and friends for their support and guidance throughout my graduate career. Without them, my accomplishments over the past four years would not have been possible. In particular, I would like to thank Dr. Yaroslava G. Yingling who molded the landscape of my graduate research. Her guidance provided me with countless opportunities to expand my skillset as she pushed me to pursue interdisciplinary areas of research. Perhaps her most influential contribution came in the form of assistantship that allowed me to become a fellow in the SEAS program.

As a SEAS fellow, I also gained invaluable mentorship from my peers and the program coordinators along with considerable amounts of technical knowledge. I would especially like to thank program coordinator, Dr. Ashleigh Wright, who always made herself available to discuss the volatility of life in research. Similarly, I would like to thank my lab-mates in the Yingling research group (Dr. Abhishek Singh, Dr. Hoshin Kim, Dr. Albert Kwansa, Dr. James Peerless, Dr. Sanket Deshmukh, Dr. Nan K. Li, Dr. Jessica Nash, Dr. Yuxin Xie, Dr. Alexey Gulyuk, Tad Deaton, Matthew Manning, Akhlak Ul Mahmood, Sabila Pinky, Sergei Rigin, and Naimul Haque) who were all a part of countless discussions inside and outside the realm of research. I would like to give Dr. Kwansa and Dr. Kim additional thanks for their contributions in my development as a researcher. Dr. Kwansa maintained the lab's hardware and software that provided the fundamental platform required to conduct intensive computations. Dr. Kim was my mentor during my time as an undergraduate researcher and served as a significant resource early in my graduate career. I would also like to thank two of my mentees, Johnny Donald and Ibrahim Ahmad, for their dedication and hard work on their respective projects while working with me.

Lastly, I would like to thank my family and friends who provided consistent support and motivation throughout my time as a graduate student. They provided an outlet to relieve stress that was necessary to navigate graduate school, while not allowing me to lose focus on my goals and career.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

1.1 Polyelectrolytes and Self-Assembly

Polyelectrolytes are defined as polymers carrying a charged, ionizable group in its repeat unit.[1] In solution, polyelectrolytes behave differently than neutral polymers as charge repulsion plays a vital role in dictating its conformations. In general, polyelectrolytes in low-salt solutions behave as directed random walks as charge repulsion dominates its behavior. This result in an elongated chain whose length is proportional to the number of Kuhn segments in the chain.[1–3] However, polyelectrolytes are salt-responsive materials as counterions can screen charge-charge interactions. In high salt concentrations, where the Debye screening length is smaller than the electrostatic blob size of the polyelectrolyte, the polyelectrolyte no longer behaves as a directional random walk but behaves as a neutral polymer in a good solvent. At intermediate salt conditions, the polyelectrolyte will behave as a semiflexible polymer. Other factors that influence polyelectrolyte behavior include the concentration of the polyelectrolyte (dilute vs. semi-dilute), charge density, solvent quality, and solvent dielectric constant.[3,4]

The responsive nature of polyelectrolytes presents a desirable opportunity to create salt and pH responsive materials. Polyelectrolytes such as DNA are of particularly high interest due to its programmable nature in terms of molecular recognition among Watson-Crick base pairs.[5] Furthermore, DNA's biocompatibility is desirable for drug delivery applications as natural, tunable degradation can serve to release cargo and minimize toxic accumulation of foreign materials.[6] However, the behavior of ssDNA is not well understood in solution limiting its success in applications. For example, persistence-length measurements, a measure of molecular stiffness, vary drastically based on the underlying theory and method used. In part, this is due to the flexibility of the molecule which makes solution structures difficult to resolve.[7–10] Furthermore,

the chain dynamics are influenced by salt concentration, valency of ions, and sequence of bases which affects are particularly complicated.[11–13] Specifically, the bases in ssDNA are known to affect chain properties due to variable degree of base-stacking and thus self-interaction, ultimately making properties a challenge to predict.[14,15]

Polyelectrolytes can also be covalently bonded to hydrophobic molecules to create polyelectrolyte block copolymers (PBCs). In aqueous solution, amphiphilic PBCs can spontaneously aggregate to form a variety of organized structures in a process known as self-assembly. This process is largely driven by hydrophobic interactions and is well studied for non-ionic block copolymers. Linear, flexible block copolymers are described by the packing parameter which is an empirical relationship between the amphiphile structure and morphology upon self-assembly. However, when amphiphile structures becomes more complex, the packing parameter begins to fail. For example, in PBCs are complex amphiphiles that are poorly predicted via the traditional packing parameter. This is due to the high dimensional design space that is reliant on the solvent, solvent ionic strength, ratio of blocks in the PBCs, and other structural and chemical factors such as branching that are poorly captured in the existing predictive model.[16]

1.2 All-Atom Molecular Dynamics

Molecular Dynamics (MD) is a widely used *in-silico* technique based on Newtonian physics. This classical view on physics is embedded in a force field, which ultimately dictates the bonding, angles, and dihedral angles of atoms in a molecule and the non-bonded interactions of atoms between molecules. Inside of each force field there are multiple approximations that are made in how the physics are described, thus, there can be many force fields to choose form with varying strengths and weaknesses. In addition, there are various packages and platforms to run molecular

dynamics. Here, we will introduce molecular dynamics in the framework of the Assisted Model Building with Energy Refinement (AMBER) package which was constructed in the 1970s.[17]

AMBER has become a state-of-the-art molecular dynamics package for biomolecules such as proteins, nucleic acids, and carbohydrates. With today's computing power, it is possible to run simulations with millions of atoms at microsecond level timescales with fully atomistic resolution.[18] The simulation process consists of iteratively integrating Newton's equations of motion to update the positions and momentum of atoms over time. There are numerous statistical ensembles that can be specified for each simulation, however the NVT and NPT ensembles are most common. In these simulations, the number of particles and temperature is always kept constant, however, the NVT ensemble maintains constant volume while the NPT ensemble maintains constant pressure through a barostat. To start a simulation, the positions of atoms are specified, and atomic velocities are randomly assigned from a maxwell distribution that is coupled to temperature. There are numerous thermostats available to control the velocity of atoms and maintain temperature throughout the simulation. At each timestep an integration algorithm is applied. One common selection is the Velocity-Verlet algorithm

$$p_i\left(t+\frac{1}{2}\delta t\right) = p_i(t) + \frac{1}{2}\delta t f_i(t) \tag{1.2.1}$$

$$r_i(t+\delta t) = r_i(t) + \delta t p_i(t+\frac{1}{2}\delta t)/m_i \tag{1.2.2}$$

$$p_i(t+\delta t) = p_i\left(t+\frac{1}{2}\delta t\right) + \frac{1}{2}\delta t f_i(t+\delta t) \tag{1.2.3}$$

where the potential energy and thus force evaluation is performed after equation 1.2.2. This potential energy calculation comes from a potential energy function

$$U_{total} = \sum_{bond} k_b(r - r_0)^2 + \sum_{angle} k_\theta(\theta - \theta_0)^2 + \sum_{dihedral} [k_\phi(1 + \cos n\phi - \delta)] + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \quad (1.2.4)$$

where AMBER force fields contain necessary parameters for each atom present. The first three summation terms encompass all bonded interactions of the simulation while the fourth summation term contains a Lennard-Jones potential and Coulombic interaction term, which together, represent all nonbonded interactions.[19]

While each force field contains a set of parameters for given atoms, it is not uncommon for simulations to require further parameterization. In these instances, quantum mechanical approaches can be used. While the level of theory directly impacts the accuracy of calculations, another large source of error comes from the handling of the electron cloud. In molecular dynamics, the electron cloud is collapsed down to a point charge that sits at the center of the atom, which is termed to be the partial charge of that atom. The method by which this is done can affect the partial charges of atoms in a molecule, and thus the dynamics of the simulation. It has recently been shown that bulk properties calculated from molecular dynamic simulations are significantly impacted by variations in partial charge, however, simulations are less sensitive to variations in partial charge of buried atoms of molecules.[20]

Traditionally, simulations are performed in explicit solvent where each solvent molecule's atoms are represented by the potential energy described by equation 1.2.4. This provides the most accurate results but comes at a higher computational cost. An alternative method is implicit solvent simulations where no solvent atoms are present in the system. Instead, the solvent is treated as a continuum that is an approximate solution to the Poisson-Boltzmann equation. Equations 1.2.5 and

1.2.6 describe the generalized born implicit solvent approach which is dependent upon factors such as solvent dielectric constant and atomic radii of the solute used in the calculations. Along with a decrease in computational cost, conformational sampling is increased because there is no more viscosity from the solvent. Thus, in certain instances, implicit solvation is the practical choice.[21,22]

$$U^{elec} = \frac{1}{2}\sum_{i \neq j} \frac{q_i q_j}{r_{ij}} - \frac{1}{2}(1 - \frac{1}{\varepsilon_w}) \sum_{ij} \frac{q_i q_j}{f_{ij}^{GB}} \tag{1.2.5}$$

$$f^{GB} = \left[r_{ij}^2 + R_i R_j \exp\left(-\frac{r_{ij}^2}{4R_i R_j}\right) \right]^{\frac{1}{2}} \tag{1.2.6}$$

The AMBER package also contains built in analysis modules with a program called CPPTRAJ. This tool provides a standardized approach to analyze the large amounts of data stored in common AMBER file types. The breadth of applications is quite varied and invaluable in determining if each simulation has been sufficiently performed including metrics for convergence and verification that the thermostat and barostat have maintained the simulation as expected.[17,19]

1.3 Dissipative Particle Dynamics

In the early 1990's, the computational cost of molecular dynamics and limited success of rheological studies by setting up and solving differential equations prevented accurate predictions of hydrodynamic behavior. To reduce computational cost but retain a higher accuracy of predictions for hydrodynamic behavior, particle-based simulations that quantitatively satisfy the Navier Stokes equations for hydrodynamic behavior were developed. These particle-based simulations significantly reduced computational cost by decreasing the necessary calculations at each time-step while simultaneously increasing the size of the time-step by orders of magnitude. However, early particle-based models such as lattice gas automata (LGA) required particle movement on a regular lattice limiting the complexity of systems that could be studied. Namely,

the movement of particles restricted to a lattice led to violations in isotropy and Galilean invariance which are fundamental flaws in the scheme. For special cases, these flaws could be overcome by selecting a lattice with sufficient rotational symmetry and Galilean invariance could be satisfied by rescaling velocities of particles at each timestep.

To overcome problems of MD and LGA methods in predicting hydrodynamic behavior, Hoogerbrugge and Koelman developed the Dissipative Particle Dynamics (DPD) simulation method. Ultimately, DPD takes an LGA-type timestep and introduces it into a MD scheme. This leads to a computationally efficient, stochastic, particle-based model for isothermal fluids that is fully isotropic, Galilean invariant, and flexible to the addition of model features. As in MD, the DPD model has a set number of N particles with specified positions r and momenta p. As in LGA, the DPD model is updated at discrete timesteps δt which consists of a collision phase followed by a propagation. During the collision phase, the momenta are simultaneously updated in such a way that momentum is conserved for the entire system. This conservative force is supplemented by two additional terms that serve as a thermostat for the model; a dampening term that represents viscosity in the fluid and a stochastic term that leads to correct pressure effects. Brownian dynamics are sufficiently described with these 3 force terms, and in conjuncture with conserved momentum, Hoogerbrugge and Koelman successfully created a model for hydrodynamic behavior as described in equation 1.3.1

$$p_i' = p_i + \sum_j \Omega_{ij} e_{ij} \tag{1.3.2}$$

where p_i is momentum of particle i, Ω_{ij} specifies the momentum transferred between particle j and particle i, and e_{ij} is the unit vector pointing from particle j to particle i.[23]

From here, the DPD method was shown to be capable of studying dilute polymer solutions where polymers were represented by a bead and spring type model. However, in this case, each bead was not viewed as a representation of a molecule, but as soft sphere equal to a Kuhn segment of the polymer that carries momentum. In this scheme, the beads of the polymer behave just as other fluid particles during the collision and propagation phase of DPD, but exchange momentum according to an elastic spring force creating a linear chain. The elastic spring force was modeled by a Fraenkel spring

$$F_{ij} = K(|r_i - r_j| - r_{eq})e_{ij} \tag{1.3.3}$$

where F_{ij} is the force, K is the spring constant, $|r_i - r_j|$ is the distance between particles i and j, r_{eq} is the equilibrium spring length, and e_{ij} is the unit vector pointing from particle j to particle i. Ultimately, scaling laws for radius of gyration R_g and relaxation times for the DPD model were compared to results from experiments of polymers in good solvent and results from the Rouse and Zimm models. The scaling laws suggested that excluded volume and hydrodynamic interactions were evident, and the degree to which they are present can be altered by DPD interactions and spring parameters.[24]

In 1995, Español and Warren noted that while the original DPD algorithm was for an isothermal liquid, there was no expression relating the temperature of the system to the dampening term or stochastic noise that served as the system's thermostat. Thus, it was necessary to formulate a continuous stochastic differential equation that could relate the DPD framework to a Fokker-Planck equation that could be used to study the equilibrium solution. Upon doing this, Español and Warren provided a clear definition of temperature in DPD and illustrated that the time-step must be sufficiently small to obtain the correct equilibrium results. The resulting DPD algorithm was rephrased as a continuous version where total force on a given particle is

$$\dot{p}_i = \sum_{j \neq i} F_{ij}^C + \sum_{j \neq i} F_{ij}^D + \sum_{j \neq i} F_{ij}^R \tag{1.3.4}$$

where F_{ij}^C is the conservative force, F_{ij}^D is the dissipative force and F_{ij}^R is the random force.

As before, the conservative force represents the repulsive potential exerted on particle *i* by the *j*-th particle, the dissipative force represents the drag, and in this case is linearly dependent on the momentum, and the random force is independent of momentum. To satisfy Galilean invariance, the drag force and random force can depend only on the position and velocity vectors of particle *i* and to meet the isotropy requirement for hydrodynamic behavior the forces should transform under rotations as vectors. Español and Warren established the form of the drag and random forces as

$$F_{ij}^D = -\gamma \omega_D(r_{ij})(e_{ij}.v_{ij})e_{ij} \tag{1.3.5}$$

$$F_{ij}^R = \sigma \omega_R(r_{ij})e_{ij}.\zeta_{ij} \tag{1.3.6}$$

where $r_{ij}=|r_i - r_j|$ and $e_{ij}=^{r_i - r_j}/_{r_{ij}}$ are unit vectors from the *j*-th to the *i*-th particle. In the drag force, γ is interpreted as the friction coefficient. In the random force, the ζ_{ij} term is a Gaussian white noise variable that has a $\zeta_{ij} = \zeta_{ji}$ property that ensures the conservation of momentum while σ is interpreted as the amplitude of the noise. Lastly, the ω_D and ω_R terms are weight functions that provide a range of interactions for the particles drag and random force.

When plugging force equations 1.3.2 and 1.3.3 into Newton's second law, Español and Warren showed that the subsequent Langevin equations were

$$dr_i = \frac{p_i}{m_i} dt \tag{1.3.7}$$

$$dp_i = \left[\sum_{j\neq i} F_{ij}^{C}(r_{ij}) + \sum_{j\neq i} -\gamma\omega_D(r_{ij})(e_{ij}.v_{ij})e_{ij}\right]dt + \sum_{j\neq i}\sigma\omega_R(r_{ij})e_{ij}dW_{ij} \tag{1.3.8}$$

The Langevin equations describe particle position (equation 1.3.6) and particle momentum (equation 1.3.7) where m_i is the mass of particle *i* and $dW_{ij} = dW_{ij}$ and are independent increments of the Wiener process, which is the integral of the Gaussian white noise process. Through the corresponding Fokker-Planck equation where $k_B$is Boltzmann's constant and T is equilibrium temperature, it is found that

$$\omega_R(r) = \omega_D^{1/2}(r) \tag{1.3.9}$$

$$\sigma = (2k_BT\gamma)^{1/2} \tag{1.3.10}$$

Thus, the fluctuation-dissipation theorem for the DPD method ends up having the exact same structure as the fluctuation-dissipation theorem for traditional Brownian motion. To satisfy equation 1.3.8, Español and Warren modified Hoogerbrugge and Koelman's original DPD algorithm by inserting an extra factor in the dissipative term.

The original DPD algorithm and the modified algorithm were then compared to the temperature in equation 1.3.9 and equipartition which was obtained from the steady-state Fokker-Planck equation where the solution was the Gibbs canonical ensemble. At timesteps too large, neither DPD algorithm performed well which was attributed to the fact that at such time steps, a particle was traveling distances near the length of calculated interactions. However, at sufficiently small timesteps, the modified algorithm was within 5% error of temperature from equation 8 and

within statistical error for equipartition while the original algorithm was still in violation of equipartition.[25]

Groot and Warren further developed the physical definitions of the DPD model parameters and provided a context in which one can interpret simulation results by linking DPD to Flory-Huggins theory of polymers. For this study, Groot and Warren utilized Español and Warren's DPD framework where the total force is represented by equation 1.3.3, the drag force represented by equation 1.3.4, and the random force is represented by equation 1.3.5. Groot and Warren define the conservative force with the repulsive potential

$$F_{ij}^{C} = a_{ij}(1 - r_{ij})\, e_{ij} \tag{1.3.11}$$

where a_{ij} is a maximum repulsion between particle *i* and particle *j*, r_{ij} is the distance between particle *i* and particle *j*, and e_{ij} are unit vectors from the *j*-th to the *i*-th particle. Equation 1.3.10 holds true for all $r_{ij} < 1$ while $F_{ij}^{C} = 0$ for all $r_{ij} \geq 1$. To update particle positions at each time step, Groot and Warren did not use the Euler algorithm as previous studies, but instead chose to use a modified velocity-Verlet algorithm

$$r_i(t + \Delta t) = r_i(t) + \Delta t v_i(t) + \frac{1}{2}(\Delta t)^2 f_i(t) \tag{1.3.12}$$

$$\tilde{v}_i(t + \Delta t) = v_i(t) + \lambda \Delta t f_i(t) \tag{1.3.13}$$

$$f_i(t + \Delta t) = f_i\big(r(t + \Delta t), \tilde{v}_i(t + \Delta t)\big) \tag{1.3.14}$$

$$v_i(t + \Delta t) = v_i(t) + \frac{1}{2}\Delta t(f_i(t) + f_i(t + \Delta t)) \tag{1.3.15}$$

where the traditional velocity-Verlet algorithm can be achieved by setting λ=1/2. It was shown that by using the velocity-Verlet algorithm one can use a significantly larger time-step in DPD compared to the Euler algorithm while still accurately representing temperature. Ultimately, using

this algorithm, a step size $\Delta t = 0.04$, noise amplitude $\sigma = 3$, and λ=1/2 were recommended for fast DPD simulations that have a stable and physically relevant meaning.

With the selection of parameters for the thermostat set, the lone parameter left is a_{ij} which describes the magnitude of repulsion between particles. The selection of this parameter was determined by the compressibility of water at room temperature using the Weeks-Chandler-Anderson perturbation theory of liquids. Ultimately, the selection of a_{ij} for water is calculated to be $a_{ij} = 75\,{}^{k_B T}/_{\rho}$ for particle density greater than 2. It is noted that the calculations per timestep increases with the square of density, thus $\rho = 3$ is a popular choice.

From here, Groot and Warren mapped a_{ij} to Flory-Huggins solution theory. This was done by performing simulations of monomeric mixtures with 50% particle A and 50% particle B. In the instance of simulations where $\rho = 3$, $a_{AA} = a_{BB} = 25$ and $a_{AB} > 25$. The density profile showed some mixing at the interface between the mixtures, with increased mixing at the interface for the lower a_{AB} values. The Flory Huggins interaction parameter, χ, was found by calculating averages particle density by type. Specifically, the fraction of volume consisting of purely particle A was substituted for ϕ_A in equation 1.3.15.

$$\chi N_A = \frac{\ln\left[\frac{1-\phi_A}{\phi_A}\right]}{1-2\phi_A} \tag{1.3.16}$$

For $\chi > 3$ it was shown the correlation between χ and a_{AB} was linear. The linear mapping between parameters varies based on density and is described by equation 1.3.16 for $\rho = 3$ and equation 1.3.17 for $\rho = 5$.

$$\chi = (0.286 \pm 0.002)\Delta a \tag{1.3.17}$$

$$\chi = (0.689 \pm 0.002)\Delta a \tag{1.3.18}$$

Groot and Warren expanded this parameterization to polymer systems with bonds represented as springs (Eq. 1.3.2) with a spring constant of 2. Using sample polymer lengths ranging from N=2 to N=10, it was shown that $\frac{\chi N}{\Delta a}$ is linearly related to the length of polymer N by equation 1.3.18.[26]

$$\frac{\chi N k_B T}{\Delta a} = (0.306 \pm 0.003)N \qquad (1.3.19)$$

This parameterization gave way to studies of more complex systems with DPD. One highly studied system is that of diblock copolymers (A_nB_m) which were shown to self-assemble into various morphologies as the length and ratio between the 2 blocks varied. Furthermore, it was shown that the solubility of each particle influences the self-assembled structure as expected. When compared to other mesoscale simulation methods, DPD was found to better predict equilibrium structures as hydrodynamic forces can play a crucial part in microphase separation. Specifically, the hydrodynamic forces were shown to affect the kinetics and pathway along which block copolymers find equilibrium structures. The ability to observe these dynamical pathways with DPD truly separates this method from other computational techniques.[27,28]

The validity of this parameterization has since been questioned. In 2001, a Gibbs ensemble Monte Carlo simulation with a soft repulsive potential and a modified identity change mechanism was used to study polymer-solvent systems. The results were mapped back to Flory-Huggins solution theory as was done by Groot and Warren with the DPD method. When comparing the a_{ij} parameter to the Flory-Huggins interaction parameter, it was demonstrated that polymer systems did not exhibit a linear mapping. Instead, the length of the polymer affects the a_{ij} parameter. However, it should be noted that this study used a different method than DPD which involved a stiffer harmonic for polymer bonds and a different repulsive potential.[29] In 2007, it was shown that DPD maintains Zimm like behavior of polymers in most instances. However, for simulations with

low Schmidt numbers, polymer behavior can deviate towards Rouse behavior indicating the hydrodynamics are slightly underdeveloped for these simulations.[30]

In 2015, efficient parameterization of polyelectrolytes was introduced to DPD using the implicit solvent ionic strength (ISIS) method. ISIS-DPD is based on an approximation of mean-field theory to reproduce the effects of solvent ionic strength. The second virial coefficient was calculated to be

$$\mathrm{V} = V_A + {\alpha^2}/{C_s} \tag{1.3.19}$$

where $V_A \leq 1$ is the non-electrostatic contribution to the second virial coefficient, C_s is solvent ionic strength, and α is the degree of ionization. This representation is analogous to the repulsive parameters between polyelectrolyte beads in DPD

$$a_{pp} = a_{ii} + a_{elec} \tag{1.3.20}$$

where $a_{elec} \approx C_s^{-1}$ and a_{ii} is the traditional DPD parameterization for a given density. As solvent ionic strength increases, $a_{pp} \rightarrow a_{ii}$ where the polyelectrolyte behaves like a neutral polymer in good solvent. The implementation of the ISIS-DPD method was tested using a single polyelectrolyte chain in solution as well as a polyelectrolyte based diblock copolymer. The single polyelectrolyte chain showed a scaling relation for end-to-end distance and radius of gyration that was in-line with previous molecular dynamics studies. The self-assembly of the polyelectrolyte diblock copolymer also agreed with developed theories and experimental observations as increasing solvent ionic strength changed micelle morphology from spherical, to cylindrical, to worm-like. In addition, the increasing solvent ionic strength was correlated with an increase in aggregation number.[31]

Overall, DPD is a soft sphere model that can represent liquids and solids with a purely repulsive potential. DPD has had its equilibrium calculations validated by other theoretical calculations ranging from Flory Huggins theory to mean-field theory. While the prediction of equilibrium structures is not unique to DPD, the speed and conservation of hydrodynamics separate this method from the rest. While DPD is a useful technique for the study of liquid-liquid or liquid-solid interfaces, it should be mentioned that liquid-vapor interfaces are not feasible to simulate with this method and caution should be used for weak polyelectrolytes at low solvent ionic strengths.

1.4 Materials Informatics

* This section is a reproduction for a portion of the manuscript in the published work:

Thomas J. Oweida, Akhlak Mahmood, Matthew D. Manning, Sergei Rigin, and Yaroslava G. Yingling. *MRS Advances* **2020** *5* (7), 329-346.

DOI: 10.1557/adv.2020.171

1.4.1 Overview

Materials-based research has adopted the use of machine learning (ML) as an analytical tool. ML encompasses any algorithm whose performance will improve, or learn, as it is exposed to or trained on larger quantities of quality data. Implementing these ML algorithms with a specific workflow to overcome the unique challenges of materials-based research is termed materials informatics (MI). The application of ML tools to materials science data and the use of MI workflow to design new materials and techniques has shown exponential growth with over 2,000 publications during the past decade. The United States leads the global effort in MI with almost half of these publications. To date, most of the publications have largely focused on facilitating materials design, parameterizing potentials for in silico techniques, and optimizing materials

characterization techniques [32–37]. For example, researchers have started to employ ML algorithms to process undetected or complex trends in databases containing first principle calculations data [38]. Ultimately, this has led to the proposal and synthesis of promising surface coatings [39], alloys[40–43], perovskites [44], and composites [33] that meet specified target properties for a specific application. MI has not only been useful in designing and predicting properties of new materials, but has also been vital in the recent development of new potentials used for in silico approaches via rapid parameterization [45,46] at a reduced computational expense [46,47] and the development of completely data driven potentials [48,49]. These advancements have been utilized to push the length and time scales of the current capabilities of simulations while maintaining the same level of accuracy as higher resolution simulation techniques. The MI framework to develop these new potentials has already been laid out in multiple studies [50–53]. For example, Huan et al. discuss a universal approach for creating atomistic force fields via ML [53]. MI has also been used to analyze phenomenological parameters such as the work performed by Miles, Leon, Smith, and Oates that looked at the uncertainty and sensitivity of parameters in a ferroelectric continuum model for lead titanate [54,55]. Experimental characterization techniques are also a beneficiary of MI approaches. For example, a Bayesian inference approach was shown to provide many advantages for X-ray diffraction peak fitting over traditional approaches such as Reitveld refinement. Specifically, the Bayesian approach has the ability to escape from false minima, incorporate prior knowledge of the material into analysis, and provide uncertainty quantification [56]. MI has also been used to analyze position averaged convergent beam electron diffraction patterns with a convolutional neural network that achieved great speed and accuracy compared to brute force methods [57]. Overall, the rate of adoption of MI workflow to speed up characterization, simulations and materials discovery has been remarkable.

1.4.2 Challenges

MI is undeniably a valuable tool for materials scientists as the modern pace of materials innovation has become intractable by traditional approaches. The success of MI in academia and industry has only reinforced this truth as structure and property predictions across vast chemical spaces become simultaneously cheaper and more accurate. However, materials informatics is still lagging behind other fields that have adopted data science approaches due to the unique challenges inherent to materials datasets. One of the most impactful processes in each MI approach is the user-dependent choice of material descriptors. In general, these descriptors need to sufficiently identify unique atomic environments, while being invariant to transformations such as translation, rotation, and permutations of like elements [48]. However, these descriptors can quickly become computationally expensive, which is especially true for soft matter as the exploration space is inherently highly dimensional [58,59]. These materials can have properties heavily reliant upon this design space as their sequence, environment, length, chemical composition, density, etc. can drastically change morphology and non-bonded interactions [60–62]. Thus, developing a framework that can identify the optimal material descriptors for each MI application can help overcome one of the biggest barriers that has kept MI from realizing its full potential. Early works have already targeted this issue through the development of standard notation such as SMILES for molecules and BigSMILES for macromolecules [63–65]. In addition, some works have performed analysis on the influence of numerous materials descriptors ranging from crystal chemistry to electronic structure descriptors used to predict multiple properties of intermetallic compounds [42].

Databases would seem to be the easy solution to standardize the structure of reported data, however, current databases are for particular purposes or limited to specific materials class(es) limiting their viability for use in MI studies. In addition, most databases do not report material

processing details resulting in a possible disconnect between the structure-property relationship that is fundamental to material science. This is especially important for less-ordered materials commonly found in soft materials and glasses. For more information on the additional challenges for disordered materials we recommend referencing "Soft Matter Informatics: Current Progress and Challenges" by Peerless et al. [59] and "Data-driven glass/ceramic science research: Insights from the glass and ceramic and data science/informatics communities" by De Guire et al. [66].

Collecting data from previous publications also possess significant challenges. In a recent paper, the quality of data reported was highly concerning for inorganic materials synthesis recipes. Through a text mining approach, it was found that the overall extraction yield was 28% of total papers. Out of the successfully mined publications, 30% of papers did not contain a complete set of starting materials and final products, thus reconstruction of the reaction was not possible. Lastly, 42% of potential reactions were not reconstructed due to an incomplete or overcomplete set of extracted precursor/target materials, or a failure to correctly parse chemical composition [67]. Thus, the already limited materials datasets are further reduced in size due to poor data quality and lack of standards for reporting data.

For future database development, the materials informatics community needs to follow the examples of well-established databases such as the Protein Data Bank (PDB) [68–70]. The development of the PDB has created a culture, incentives and level of prestige that benefits each researcher that successfully submits a protein structure in this database for others to use. This effective data sharing in PDB database resulted in the growth and development of structural bioinformatics field. In addition to centralizing large amounts of data, PDB has implemented a data quality metric that ensures only quality data exists within the database effectively reducing the burden and time it takes to pre-process and filter the data for analysis.

For databases specific to materials science, The National Institute of Materials Science (NIMS) deserves special recognition as being one of the front runners for database development for MI applications [71]. NIMS is the co-copyright owner of databases such as the Pauling Files [72] which provides reliable data on crystal structure and phase diagrams in addition to being the owner of other respected databases such as PoLyInfo which provides data for polymeric materials design [73]. These databases are traditionally curated by hand as NIMS employees comb through literature daily assessing accurate information for entry into a database. Thus, NIMS curation of data has resulted in the development of databases that have been successfully used as the source of information in numerous MI studies.

CHAPTER 2: Methods for Analyzing Simulations of Flexible Polymers

2.1 Assessment of AMBER Force Fields for Simulations of ssDNA

* This section is a reproduction of the manuscript for the published work:
Thomas J. Oweida, Ho Shin Kim, Johnny M. Donald, Abhishek Singh, and Yaroslava G. Yingling. *Journal of Chemical Theory and Computation* **2021** *17* (2), 1208-1217.
DOI: 10.1021/acs.jctc.0c00931

2.1.1 Introduction

The knowledge of double-stranded DNA (dsDNA) structure, resolved by high-resolution methods such as X-ray Diffraction (XRD) and Nuclear Magnetic Resonance Spectroscopy (NMR), has enabled numerous investigations of dsDNA structure and properties.[74] While single-stranded DNA (ssDNA) has been used for applications in medicine, bioelectronics, and DNA nanotechnology [75–83], this research is currently bottlenecked by the lack of high-resolution structure in solution due to its inherent dynamic nature. For example, studies of ssDNA persistence length indicated a 2 to 4 times difference in measurements at a given concentration depending on experimental methodology.[7,8,84] A deeper understanding of ssDNA's structure and behavior will help clarify confounding property predictions of ssDNA and facilitate the development of ssDNA-based materials.

Typically, investigations of the structure and dynamics of biomolecules are performed using experimental methods such as XRD, NMR, Forster Resonance Energy Transfer (FRET), and Small Angle X-Ray Scatting (SAXS).[7,85] XRD and NMR have the potential ability to resolve high resolution structures of ssDNA, however, these techniques require well-defined structures. Thus, the intrinsic flexibility of ssDNA limits the use of these techniques to studies of nucleic acids bound to proteins or other macromolecules, ultimately providing little to no information of ssDNA structure in solution.[86–90] To overcome the limitations of XRD and NMR, researchers have utilized lower resolution tehniques such as FRET and SAXS that are capable of measuring structural

information of intrinsically disordered molecules. FRET technique has been widely used to measure the end-to-end distance (R_{ee}) of ssDNA by tracking the positon of flourescent molecules bound to the 3' and 5' end of the DNA. However, the presence of the flourescnet molecules on the ends of ssDNA may interfere with the natural dynamics to some degree.[7,91–93] SAXS is a more powerful technique that produces a spatially averaged intensity distribution originating from the contrast in electron density between the molecule and surrounding solvent.[94] The final intensity curve provides information about the average size, shape, and structure, which can be used to derive specific structural information such as radius of gyration (R_g), volume, and molecule flexibility through the use of analytical tools such as Kratky plots or the Porod-Debye law.[85,94–96] In recent years, SAXS has been employed to obtain information on the length, nucleobase, and ionic strength dependence of single-stranded nucleic acids.[12,13,93,97,98] However, the uncertainty underlying the ensemble average values obtained from FRET and SAXS and variations in empirical formulas have propagated into ssDNA persistence length predictions and has led to a large disparity in values that range from approximately 9 Å to 25 Å in 100 mM solution.[7,8,84] Currently, the structure and dynamics of ssDNA in solution remains unresolved which has prompted the use of *in-silico* techniques to obtain a higher resolution understanding.[9,99]

Similar to experiment, the use of *in-silico* techniques has been highly successful for studying dsDNA due to the development of reliable force fields for all atom molecular dynamics (MD).[100] The accuracy of MD for dsDNA is highlighted in a recent study by Galindo-Murillo et al. which provides a detailed evaluation on the progression and accuracy of AMBER force fields for dsDNA in explicit solvent.[101] This study illustrates that the two most recent DNA force fields, ff99 with bsc1 corrections and OL15, are state-of-the-art when comparing the simulated dsDNA structure to XRD and NMR experiments.[101] Gaillard et al. also performed a detailed structural analysis of

dsDNA in a variety of generalized born implicit solvents.[102] This study showed that implicit solvent is in reasonable agreement with dsDNA's structure from explicit solvent simulations and experimental data.[102] The thorough evaluations performed in both studies clearly demonstrate the robust nature and reliability of AMBER force fields and protocol for MD simulations of dsDNA. However, it is important to note that these force fields are specifically parameterized for dsDNA with the goal to maintain helical stability. However, studies involving ssDNA are also usually performed with the same force fields because there are no alternative force fields specific to ssDNA.

To date, there has been limited evaluation on the performance of AMBER force fields for simulating ssDNA.[103] The limited understanding of how these force fields represent ssDNA structure has resulted in a wide range of simulation protocols being employed to study ssDNA's structure,[9,104,105] interactions,[106–108] and binding mechanisms.[109–112] The intrinsic disorder and flexibility of ssDNA has made validating simulation protocol particularly difficult as researchers have limited experimental data for comparison. In addition, identifying convergence in the simulations becomes challenging due to the large degrees of freedom of ssDNA and the presence of metastable low energy states. Complex ssDNA-ion interactions also obscure ssDNA's behavior as sequence, ionic strength, and the valency of ions in solution can uniquely impact ssDNA structure.[13]

Here, we perform a comprehensive evaluation on the applicability of four widely used AMBER force fields in all-atom MD simulations of ssDNA in explicit and implicit environment. The force fields and protocol used in this study were chosen due to their recent development or recent use in computational studies of ssDNA.[104,111,113,114] We chose 30-mer poly-thymine (polyT) as our model material, due to availability of experimental data, such as FRET, SAXS and force

extension methods. The simulations results were first evaluated using comparison to traditional low-resolution measurements such as R_g and R_{ee} from FRET and SAXS. Additionally, we compared simulations using a more rigorous validation process by comparing theoretical and experimentally generated SAXS curves.[115,116] Overall, our study provides guidance and insight on the current limitations of the use of current force fields for ssDNA studies.

2.1.2 Methods

2.1.2.1 All-Atom MD Simulations

Figure 2.1 provides an overview of the 10 simulations performed in this study. Nucleic Acid Builder (NAB) software was used to generate a 30-mer of the adenine-thymine duplex in the Arnott B-DNA canonical structure.117 The complementary adenine strand was subsequently deleted using BIOVIA Discover Studio to create the initial structure of single-stranded thymine.118 All initial simulation setups were prepared using xLeAP in the AMBER18 software package.19 The AMBER18 software package contains parameters for all three solvation environments and all four of the different DNA force fields tested. The four DNA force field parameters tested were parm99 (ff99)119, parm99 with bsc0 corrections (bsc0)120, parm99bsc0 with bsc1 corrections (bsc1)121, and parm99bsc0 with OL15 corrections (OL15).113,114 To examine DNA in the implicit and explicit environment, these four force fields were tested in both the Hawkins, Cramer, Truhlar Generalized Born (GB) implicit water model with parameters described by Tsui and Case (IGB1)21,122,123 and the TIP3P explicit water model124 with the explicit NaCl ions described by the Joung and Cheatham parameters.125,126 The salt concentration for the implicit and explicit water models was set to 100 mM to match experimental measurements of 30-mer polyT. TIP3P was chosen due to the parameterization of the AMBER DNA force fields within this water model,

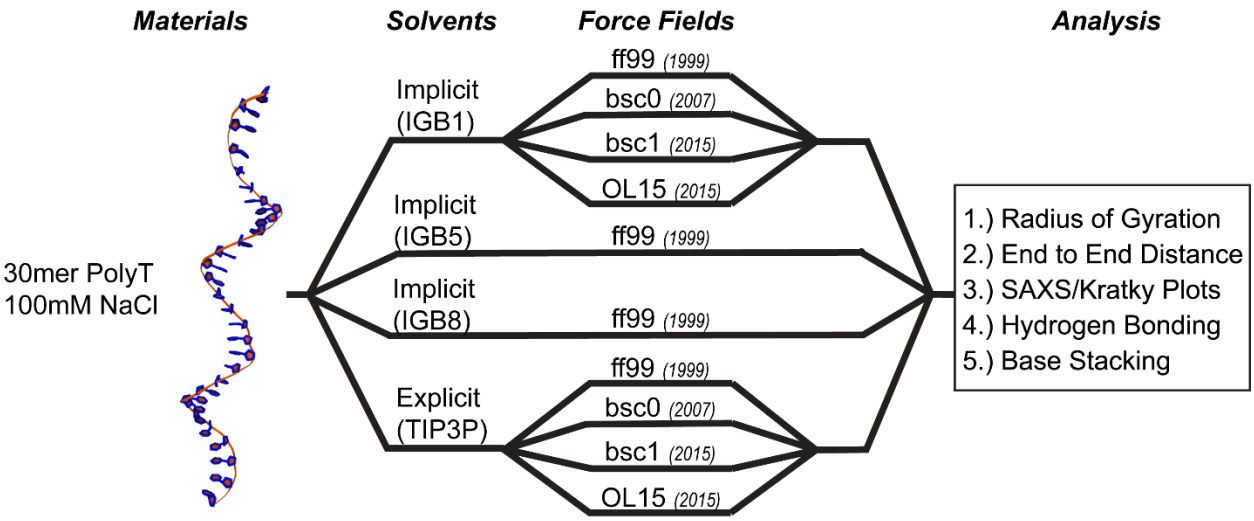


Figure 2.1. Schematic representation of a workflow for materials and methods. From left to right, the figure shows the ssDNA sequence, MD choices for solvent and force field, and concludes with the analysis done on the simulation results.

while IGB1 was chosen due to its previous use in studies and computational efficiency as compared to explicit solvent. Implicit solvent models are a popular choice as its significantly reduces computational cost by approximating water as a dielectric continuum where the electric potential is a function of distance and the Born radius of atoms. Ultimately, there is more electrostatic screening for atoms exposed to waters dielectric continuum due to the higher dielectric constant used to describe water. The absence of solvent viscosity in implicit solvent allows for rapid exploration of conformational space by ssDNA. To account for variations in IGB parameterization, the best performing DNA force field in IGB1 solvent, ff99, was also simulated in a modified GB model (IGB5) and the GB-neck2 model (IGB8) which includes a physically motivated correction to better mimic the molecular surface compared to IGB1 and IGB5. This newer implicit water models have modified atomic radii and parameters specific to nucleic acids, with IGB8 producing more stable DNA duplexes.127,128 IGB8 has also improved accuracy in protein parameters, which will be valuable for simulations containing both DNA and proteins.128

For simulations in explicit solvent, Langevin dynamics was used with a 1 ps^{-1} collision frequency, a 10 Å cutoff for non-bonded interactions, and Particle Mesh Ewald (PME) method for long-range electrostatics.[101] PolyT was first neutralized with 29 sodium counterions whose placement was calculated using a Coulombic potential on a grid. A truncated octahedral TIP3P water box with a 15 Å buffer distance between the polyT chain and the edge of the box was then generated to avoid self-contacts across periodic boundary conditions. Lastly, an excess of NaCl ions were added randomly to achieve a final salt concentration of ~100 mM. The system was energetically minimized in four stages, with each stage using the steepest descent method for 5000 steps followed by the conjugate gradient method for another 5000 steps. A harmonic restraint of 10, 10, 5, and 0 kcal mol^{-1} $Å^{-2}$ was used to restrain polyT for each minimization stage, respectively. After the first minimization stage, the harmonic restraints were used only on non-hydrogen atoms of polyT. Following minimization, the system was heated for 100 ps to 300 K using a timestep of 0.5 fs and a harmonic restraint of 10 kcal mol^{-1} $Å^{-2}$ on polyT. Next, three equilibration steps were performed for 1 ns with a harmonic restraint of 5, 1, and 0.5 kcal mol^{-1} $Å^{-2}$ on polyT respectively. The production run of all TIP3P solvent simulations was performed at 300 K for 500 ns using a 2 fs timestep in the NPT ensemble. The SHAKE algorithm was used to constrain all bonds involving hydrogen atoms while temperature was regulated using Langevin thermostat and pressure was regulated using the Berendsen barostat.

For implicit solvent simulations, we used a cutoff of 999 Å for nonbonded interactions and the default of 78.5 dielectric constant for water and 1.0 dielectric constant for solute. The initial minimization of polyT was performed using the steepest descent method for 1000 steps followed by the conjugate gradient method for 9000 steps. During minimization, a harmonic restraint of 5 kcal mol^{-1} $Å^{-2}$ was applied to all non-hydrogen atoms. Next, Langevin dynamics with

a collision frequency of 5 ps^{-1} and 1 fs timestep was used to heat the system to 300 K. The SHAKE algorithm was implemented during heating to restrain the elongation of bonds with hydrogen atoms. The system was then equilibrated in three stages for 50 ps each with a harmonic restraint of 5, 1, and 0.1 kcal mol^{-1} $Å^{-2}$ used to restrain polyT for each equilibration stage, respectively. The production run of all implicit solvent simulations was performed at 300 K for 500 ns with a Langevin thermostat, collision frequency of 5 ps^{-1} and a 1fs timestep. The ff99 force field in IGB5 and IGB8 implicit solvent simulations were each extended to 2 μs (Figure S2.2).

2.1.2.2 PolyT Analysis

The CPPTRAJ module in the AMBER18 package was used to analyze the mass weighted R_g and R_{ee} between the 3' and 5' ends center of mass of polyT for every 100 ps over the last 100 ns of implicit and explicit solvent simulations. Small angle X-ray scattering (SAXS) curves were generated using FoXS software[129,130] and fit to the experimental SAXS curve provided by the SASDBD39 entry on the Small Angle Scattering Biological Data Bank (SASBDB).[131] PDBs for this SAXS analysis were generated every 50 ps during the last 100ns of implicit and explicit solvent. Base stacking between adjacent thymine bases was analyzed with an in-house TCL script and VMD 1.9.3.[132] Intramolecular hydrogen bonds of polyT were measured for the entire simulation with the N3 atom being the donor and the O2 and O4 atoms being the acceptor atoms. The hydrogen bonds were measured through the CPPTRAJ module in AMBER18 with a donor-acceptor distance of 3.0 Å and an angle cutoff of 45 degrees. This criteria for hydrogen bonding has been successfully used for analysis in previous studies.[133]

2.1.3 Results and Discussion

Low-resolution structural properties, such as R_g and R_{ee}, can be calculated from MD simulations and from FRET and SAXS. The R_g and R_{ee} of polyT were calculated over the last 100 ns of each force field MD simulation and were plotted as bivariate contour plots to obtain a better view of the overall chain dynamics (Figure 2.2). The color intensity in the contour plots on Figure 2.2 varies with frequency of occurrence. The experimental values of R_g and R_{ee} were obtained from Plumridge et al.[13] where average values are represented as black dashed lines and the green

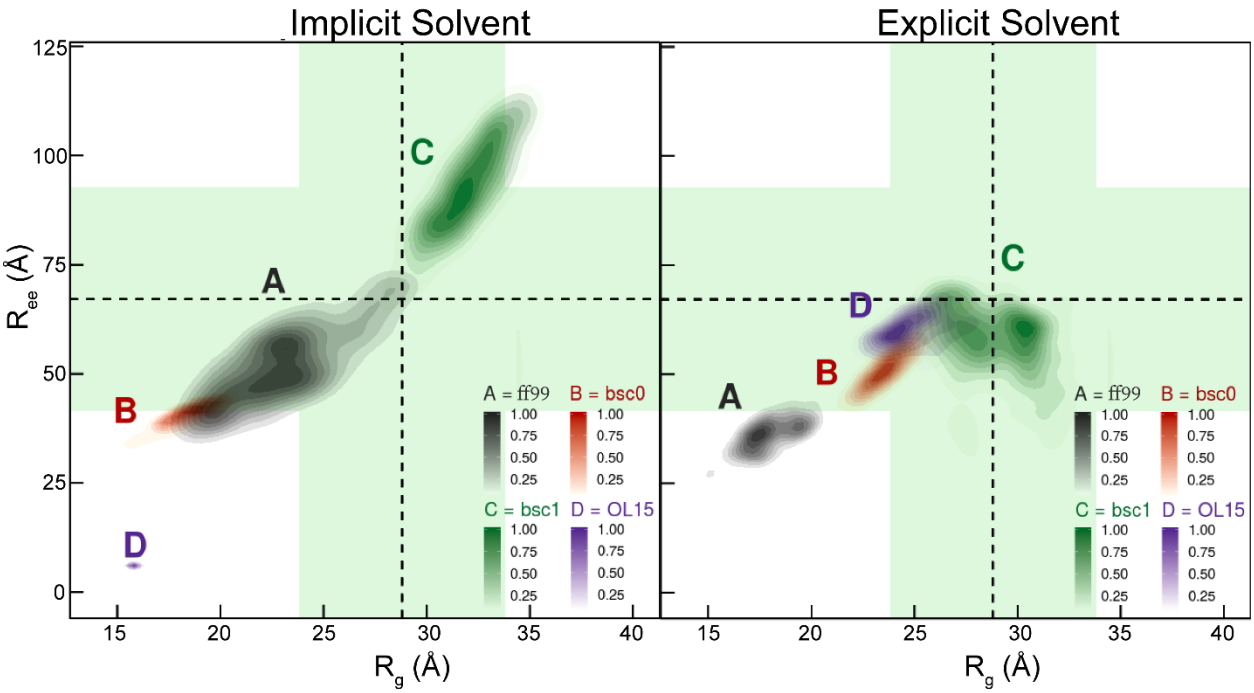


Figure 2.2. Comparison of the Rg and Ree ssDNA values obtained from MD simulations and SAXS and FRET experiments. The data is from the last 100 ns of a 500 ns simulation that started in the B-DNA canonical structure. The simulated force fields are represented by color with ff99 (black), bsc0 (red), bsc1 (green) and OL15 (purple). Experimental values and errors are represented by the dashed black line and shaded green areas. The legend indicates frequency of occurrence with dark colors (1.00) being highest relative frequency and light colors indicating lower frequency of occurrence.

bars represent experimental error. The results from Figure 2.2 indicates that there is an unclear picture on what force field and solvent model produces the most accurate representation for ssDNA structure and dynamics. In general, the implicit solvent simulations appear to sample a larger

conformational space near experimental values which is likely due to the lack of viscosity in these simulations. While this can be advantageous due to the inherent flexibility of ssDNA, only ff99 and bsc1 appear to be reasonable force field choices while bsc0 and OL15 appear outside of the experimental validation zone. In addition, neither force field appears to be in exact agreement with experiment as ff99 appears to favor conformations on the lower end of experimental measurements and bsc1 favors conformations on the higher end of experimental measurements. The simulation results differ in explicit solvent as bsc0, bsc1, and OL15 fall within expected values of experiment with ff99 being the only force field that does not produce reasonable results. While all three force fields sample similar conformations, bsc1 appears to be the most accurate in representing experimental values of R_{ee} and R_g. Figure S2.1 provides temporal profiles of R_g for each simulation and the corresponding ssDNA snapshots indicating the approximate path from the B-DNA canonical structure to the sampled relaxed ssDNA states. Ultimately, these analysis serves as a useful screening approach but is inadequate to serve as a method for validation due to its limited ability to identify the strengths and weaknesses of the ssDNA structures that fall in and out of the experimental range. Thus, a more robust approach should be adopted that compares more than the low-resolution structural features of ssDNA to one another.

SAXS presents a more holistic approach that determines the size, shape, and interface of molecules based on the scattering angle of an X-ray. To augment the initial conformational analysis in Figure 2.2, a theoretical SAXS curve was calculated from MD simulations and then

$$I_m(q) = \sum_{j=1}^{N_A}\sum_{i=1}^{N_A} f_i(q)f_j(q)\frac{\sin(qd_{ij})}{qd_{ij}} \tag{2.1.1}$$

$$x = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{I_{exp}(q_i) - cI(q_i)}{\sigma(q_i)}\right)^2} \tag{2.1.2}$$

compared to an experimental SAXS intensity curve. Equation 2.1.1, shows the formula used to perform the calculation from MD simulations where I(q) is the scattering intensity, f(q) is the form factor for each atom, $q=(4\pi/\lambda)\sin(\theta)$, and d_{ij} is the Euclidean distance between atoms i and j. Equation 2.1.2 illustrates the goodness of fit calculation for comparing scattering intensity from MD simulations to the experimentally generated one where $I_{exp}(q)$ is the experimental intensity, I(q) is the calculated intensity, c is a scaling factor to account for solute concentration differences, σ(q) is the experimental error, and M is the number of data points in the SAXS curves.[129,130]

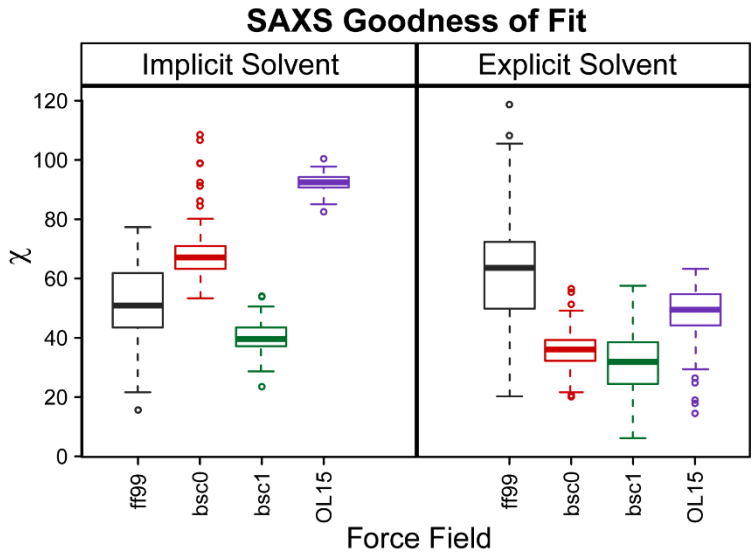


Figure 2.3. Calculated SAXS goodness of fit for 1000 simulated ssDNA structures from each force field in implicit and explicit solvent. Higher χ values indicate a worse fit of theoretical SAXS curves compared to experiment as calculated by Eq. (2).

Figure 2.3 shows the box plots for the calculated goodness of fit for each force field in the two different solvation environments and represents the deviation of the calculated SAXS curve to the experimental curve. All structures used in the SAXS calculations come from the last 100 ns of simulations at 0.1 ns increments. The distribution of χ values indicates that very few of the ssDNA simulations reproduce a structure that agrees well with experimental data. It should be noted, however, that the experimental SAXS curve represents the ensemble average of numerous low-energy ssDNA structures. Therefore, we analyze the structures with the minimum χ values in

Figure 2.3. The results indicate that the simulations using ff99 force field in implicit solvent and bsc1 force field in explicit solvent both produce the ssDNA structures with the lowest χ values in their respective solvation environments. This corresponds well with the R_g and R_{ee} analysis in Figure 2.1 as these same force fields sample conformational space that fall closest to the expected experimental values. This trend is also seen for the other force fields whose increase in χ is correlated with an increase in error found in R_g and R_{ee} values in Figure 2.2. For example, OL15 in implicit solvent has an extremely high χ value and extreme error in the simulated R_g and R_{ee} values as compared to experiment.

While this analysis provides insight on what force fields best capture the structure of ssDNA, it is important to note that magnitude of the χ values for ff99 in implicit solvent indicate the ssDNA structures from this simulation still rather poorly represents experiment, despite appearing to be the best protocol. Alternatively, the single digit χ values for bsc1 in explicit solvent indicate that these ssDNA structures reasonably agree with experiment and contain smaller amounts of error. Thus, the bsc1 force field in explicit solvation appears to be the overall best protocol to use for simulation of solvated ssDNA. However, while the comparison of SAXS curves is useful in determining the best simulation protocol, it does little to highlight the origins of error in the ssDNA structures. SAXS curves can be broken into three regions, each which contain information on the R_g, shape, or interface of the molecule. A goodness of fit metric such as χ will not tell you if goodness of fit error is originating from the R_g, shape of the molecule, or interfacial features. Looking at Figure 2.2, most simulations resulted in R_g and R_{ee} values that fell within experimental errors, so it is likely that the high χ values originate from the shape or interface regions of the SAXS measurements. Thus, it is necessary to explore how the shape of our simulated ssDNA

corresponds to experiment, as well as analyzing smaller features to determine possible root causes of these high χ values.

To assess the quality of ssDNA's overall fold and shape in our simulations, the scattering intensity of the calculated SAXS curve was multiplied by q^2 and plotted against q to create a Kratky Plot, which analyzes the decay in intensity of the curve. Generally, in a Kratky plot, the linear behavior at the larger q values is indicative of a disordered molecular structure, such as a worm

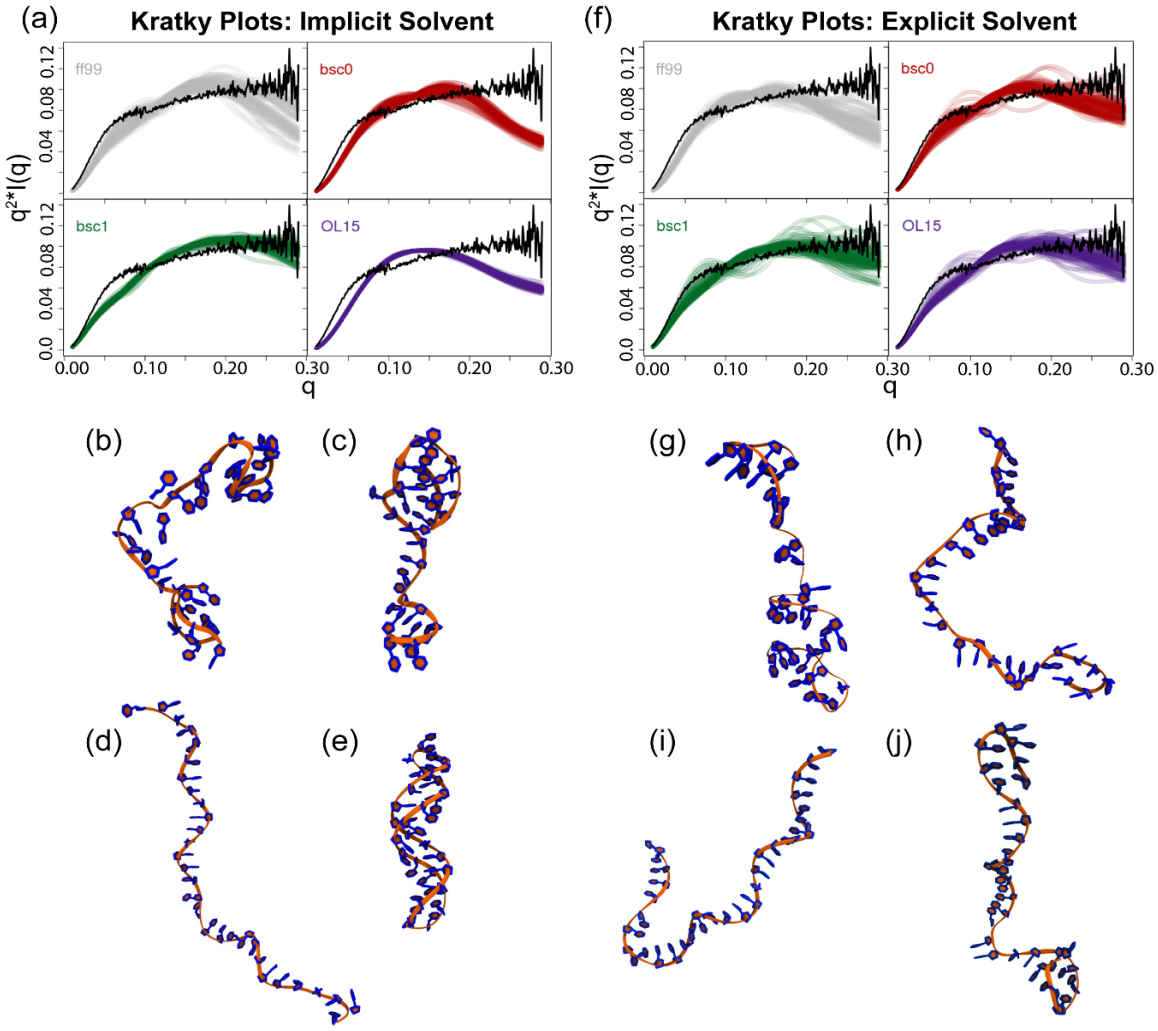


Figure 2.4. Kratky plots of ssDNA structures obtained from (a) implicit and (f) explicit solvent simulations with ff99 in grey, bsc0 in red, bsc1 in green, OL15 in purple, and experimental results in black. Representative snapshots of ssDNA structure in implicit solvent simulations with (b) ff99, (c) bsc0, (d) bsc1, and (e) OL15 and explicit solvent simulations with (g) ff99, (h) bsc0, (i) bsc1, and (j) OL15.

like chain; whereas parabolic behavior is representative of partial or complete folding in the structure, such as globular proteins.[94] Figure 2.4(a) and 2.4(f) shows the comparison between the Kratky plots for ssDNA structures from MD simulations of the implicit and explicit solvent (collected during the last 100 ns of the trajectory) and the experimental SAXS curves (black color).[131] The experimental curves indicate that there is no folding of the ssDNA chain. However, most of the simulated ssDNA structures, except for bsc1 in implicit and explicit solvent, appear to have structures that produce a parabolic Kratky plot to some degree. This indicates that despite solvation, ff99, bsc0, and OL15 are all biased towards having some order and folding in their structures. Representative snapshots of the ssDNA structures can be visualized for implicit solvent in Figures 2.4(b)-2.4(e) and for explicit solvent in Figures 2.4(g)-2.4(j). These snapshots illustrate self-folding in ssDNA structures resulting in the strong parabolic features seen in the corresponding Kratky plots, such as OL15 in implicit solvent (Fig. 2.4(e)). In addition, the snapshots of bsc1 in implicit (Fig. 2.4(d)) and explicit (Fig. 2.4(i)) solvent do not exhibit any folding which is also corroborated by the Kratky plots, however, the bsc1 fits do indicate some bias as compared to experiment, especially in implicit solvent. Ultimately, the errors observed in the Kratky plots in Figure 2.4 are most likely due to the fact the force fields were parameterized for dsDNA, which is a molecule containing high order.

To assess the extent of the folding in ssDNA structures as a function of force field and solvent model, the atomistic features of ssDNA, such as base stacking and base pairing, were analyzed. Figure 2.5(a) demonstrates a clear trend that the newer force fields are correlated with a higher degree of base-stacking in ssDNA. Although it is difficult to quantify the base stacking from force extension experiments, it is estimated that polyT has little-to-no base stacking,[15] which is a stark

contrast the base-stacking seen throughout all MD simulations. Ultimately, for implicit and explicit solvent, the oldest force field, ff99, produces the base-stacking closest to what one would expect in ssDNA while the re-parameterization of dihedral angles in the newer force fields lends itself to form a helical structure and high base stacking. Specifically, the base stacking is particularly low for ff99 in implicit solvent, which likely plays a role in its enhanced performance relative to the newer force fields.

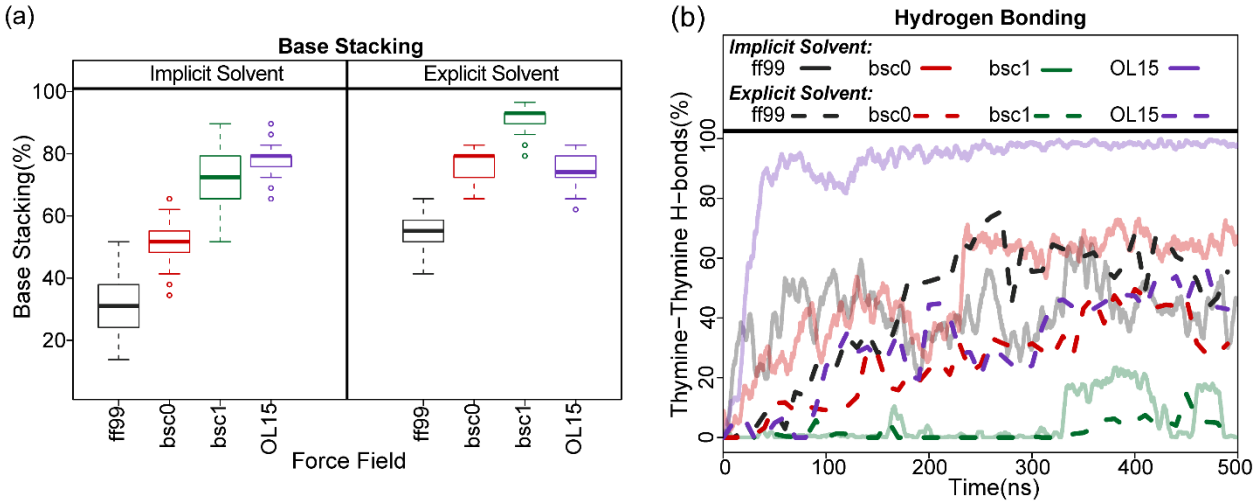


Figure 2.5. Assessment of ssDNA atomistic features, such as (a) base stacking during the last 100 ns and (b) temporal profile of thymine-thymine hydrogen bonding. Dashed lines represent explicit solvent while solid lines represent implicit solvent.

Figure 2.5(b) illustrates the base-pairing in ssDNA by tracking hydrogen bond formations between thymines. Thymine-thymine interactions are infrequent but can occur in the formation of loops.[134] This is supported from the observation that force fields with the largest percentage of thymine-thymine hydrogen bonding are those that have the largest error when compared to experimental SAXS and FRET data. Namely OL15 in implicit solvent, bsc0 in implicit solvent, and ff99 in explicit solvent simulations have the highest number of thymine hydrogen bonding. These three simulations are also the same three simulation that produce macroscopic chain

conformations that fall out of the expected of R_g and R_{ee} ranges according to SAXS and FRET experiments (Figure 2.2).

Although base stacking and hydrogen bonds are features too small to observe in SAXS, these atomistic features influence the chain dynamics and conformations at scales that can ultimately be measured through SAXS. For example, bsc1 has the least amount of hydrogen bonding out of any force field which is ultimately correlated to the lack of folding through self-interaction in the ssDNA chain. This likely accounts for the accuracy in R_g and R_{ee} values seen in Figure 2.2 and correlates well with the minimal parabolic structure seen in the Kratky plots in Figure 2.4. However, it should be noted that bsc1's propensity to maintain high levels of base stacking disagrees with force extension experiments, indicating bsc1 overestimates base stacking in ssDNA. Ultimately, it appears that formation of an excessive number of hydrogen bonds within ssDNA's structure in simulation is the most significant cause of error when compared to experiment as it directly influences the shape and size of the ssDNA molecule. However, it is possible that excessive base stacking also plays an influential role on the chain conformations explored in simulation. The high degree of base pairing and base stacking is critical to maintain stability within dsDNA. Since all DNA force fields are parameterized for dsDNA, these same force fields cannot accurately capture all atomistic features of ssDNA.

Overall, we found that all force fields, despite solvation method, have limitations in capturing atomistic features of ssDNA, however, macroscopic chain conformations appear to be relatively accurate in terms of R_g and R_{ee}. Our current recommended protocol for ssDNA is to use the ff99 force field in implicit solvent or bsc1 force field in explicit solvent. Other force field selections can be defended to some degree, such as bsc1 in implicit solvent, bsc0 in explicit solvent, or OL15 in explicit solvent, however, each of these force fields appear to have a slight disadvantage to the

recommended force fields. For example, bsc1 in implicit solvent makes the ssDNA chain too stiff and biases it towards a helical structure of DNA and high levels of base stacking (Figure 2.4(d)). The bsc0 and OL15 force fields in explicit solvent appear to fall farther away from the expected R_g and R_{ee} values as compared to bsc1, but their base stacking propensity appears to be marginally better than what is seen in bsc1.

2.1.4 Effect of Generalized Born Models

Since our results indicated that the ff99 force field in implicit solvent produced the best agreement with the experiment, we next examine the effects of implicit solvation models parameters, such as IGB1, IGB5, IGB8, on ssDNA structure. The IGB5 and IGB8 models were selected because of the updated parameters for nucleic acids. Moreover, they contain a different set of Born radii for calculations of the electric potential. IGB8 also has an additional correction aimed to better mimic the molecular surface. Our results indicate that the change of solvation model from IGB1 to IGB5 and IGB8 models improve the agreement with experimental observations (Figure 2.6). For example, the use of these models results in a slightly stiffer ssDNA chain that pushes the R_g and R_{ee} simulation values closer towards the experimental average (Figure 2.6(a)). The contour plots illustrate that the IGB5 simulation explores chain conformations up to the boundaries of the expected experimental values, indicating reduced bias towards an overly compact structure as seen with ff99 in IGB1. IGB8 also has notable improvements to IGB1 and falls within experimental expectations but appears to take on a slightly more compact structure as compared to IGB5 simulations. Figure 2.6(b) indicated that the goodness of fit for calculated SAXS profiles is also improved for IGB5 and IGB8 implicit solvent models as compared to IGB1 which is in line with results from Figure 2.6(a). The χ values (Figure 2.6(b)) and Kratky plots

(Figure 2.6(c)) demonstrate that the IGB5 and IGB8 produce better agreement with experiment as

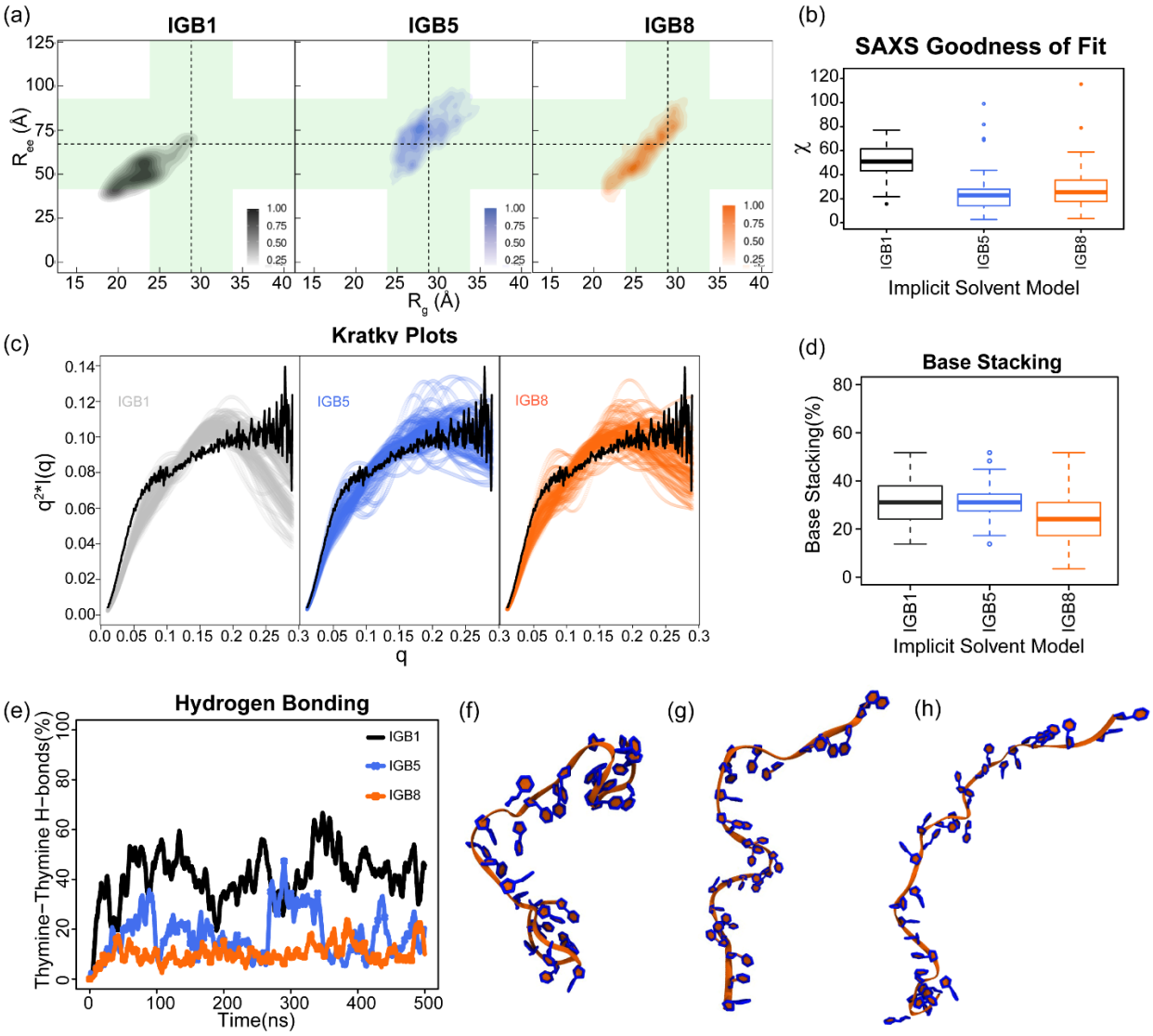


Figure 2.6. Performance of ff99 with IGB1, IGB5 and IGB8 implicit solvent models. (a) Comparison of the R_g and R_{ee} values for ssDNA obtained from simulations and SAXS and FRET experiments. Experimental values and errors are represented by the dashed black line and shaded green areas. The legends range from 0 to 1, with 1 being the darkest shade that indicates a higher relative frequency of occurrence. (b) SAXS calculated goodness of fit, (c) Kratky plots from the simulations and experiments, (d) base stacking for the last 100ns of simulations, (e) thymine-thymine hydrogen bonding, and representative snapshots of ssDNA described by the ff99 force field in (f) IGB1, (g) IGB5 and (h) IGB8 implicit solvents.

compared to IGB1 implicit solvent. However, IGB5 is in slightly better agreement with experiment than IGB8. Figure 2.6(f-h) corroborates this finding through representative snapshots of the

ssDNA where the IGB5 (Fig. 2.6(g)) ssDNA structure did not exhibit folding at the 3' and 5' ends, while the IGB8 (Fig. 2.6(h)) ssDNA structure shows slight folding at one end of ssDNA, and the IGB1 (Fig. 2.6(f)) simulation is the most compact and folded of all three.

Overall, the use of ff99 with either IGB5 or IGB8 implicit solvent are excellent choices for capturing the macroscopic structure and dynamics of ssDNA. However, as we analyze the atomistic features such as base stacking and thymine-thymine hydrogen bonding, IGB8 appears to have an overall advantage as compared to IGB5. Figure 2.6(d) shows that IGB8 has the lowest base stacking, which indicates a better agreement with force extension experiments.[15] Thymine-thymine hydrogen bonding propensity (Figure 2.6(e)) is significantly reduced in both IGB5 and IGB8 as compared to IGB1. It is likely that the reduction in hydrogen bonding propensity is the root cause for the improved accuracy in macroscopic chain motions for IGB5 and IGB8. This supports the argument that excess hydrogen bonding in ssDNA is a significant cause of error when compared to experiment.

Overall, simulations with ff99 in IGB5 and IGB8 implicit solvent outperformed simulations of ff99 in IGB1 implicit solvent. While both simulations appear to be reasonable choices for simulating ssDNA, IGB8 has slight advantages when looking at atomistic features. In addition, the improved parameters in IGB8 for simulations of proteins will likely provide a more applicable approach for studies of protein-DNA complexes.[135]

2.1.5 Conclusion

Overall, the results indicate that the ff99 force field with IGB5 or IGB8 implicit solvent have a strong agreement with experimental data in terms of the R_g, R_{ee}, Kratky plots, and atomistic features. For explicit solvent simulations, the bsc1 force field with the TIP3P water model appears to be the best; however, bsc0 and OL15 force fields can be viable options as well. While the

macroscopic chain conformations appear to be adequately captured in these recommended force fields, there are limitations on how much the ssDNA structure is able to deviate from the double helix due to particular parameterization of all force fields for dsDNA. This can be exemplified by a high propensity for base-stacking in ssDNA structure for all force fields. Excessive thymine-thymine hydrogen bonding can also bias simulation results in explicit solvent simulations using bsc0 and OL15 force fields.

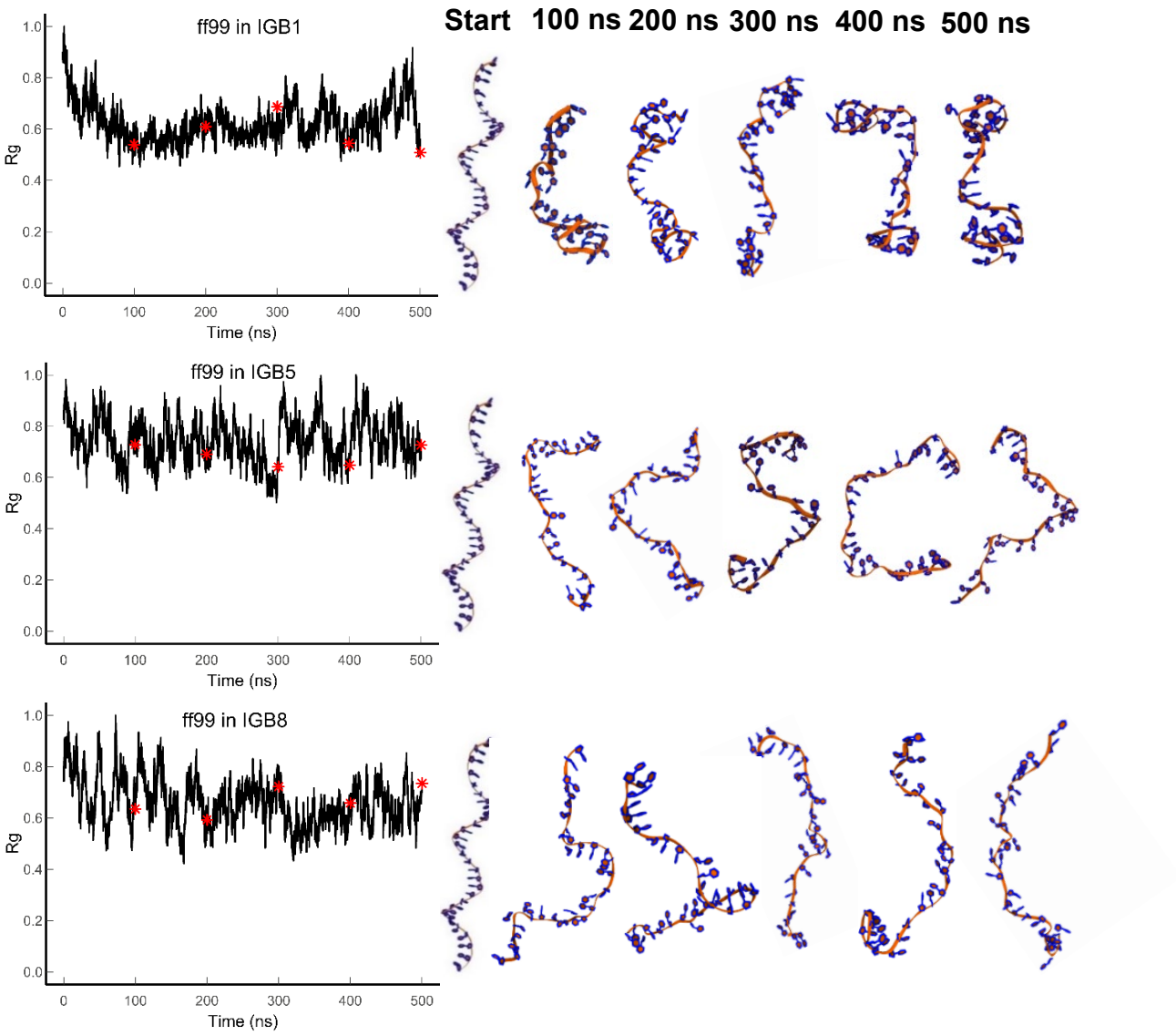
While we find the results in this study to be indicative of what force fields tend to perform best for ssDNA, there are limitations to our findings. First, it is important to note that the starting conformation of ssDNA is identical for all simulations in this study and that simulations starting from other conformations and more folded states could alter force field performance as the potential energy landscape could be different. Second, any sequence deviating from polyT could alter force field performance. As more experimental data becomes available on other sequences, it is recommended that researchers follow the protocol for validation laid out in this study, particularly using metrics for comparison beyond R_g and R_{ee}. Lastly, it is important to note that while the length of MD simulation to converge ssDNA is not known, it can be assumed that due to the increased flexibility as compared to dsDNA, ssDNA simulation times will exceed the microseconds time scale that is required to converge dsDNA.[136] Thus, the 500 ns time scale used in this study will clearly not capture the entire ensemble of structures. To verify that 500 ns results are still meaningful, we extended the simulation time for ff99 in IGB5 and IGB8 solvents to 2 μs (Figure S2.2). The results indicate that longer simulations did not alter the conclusions in this paper. However, it may be beneficial for individual studies to perform simulations beyond the 500 ns time scale to capture a more complete ensemble. Employing new strategies rooted in machine learning is a possible route to assure the convergence.[59,137] For example, machine-learned force

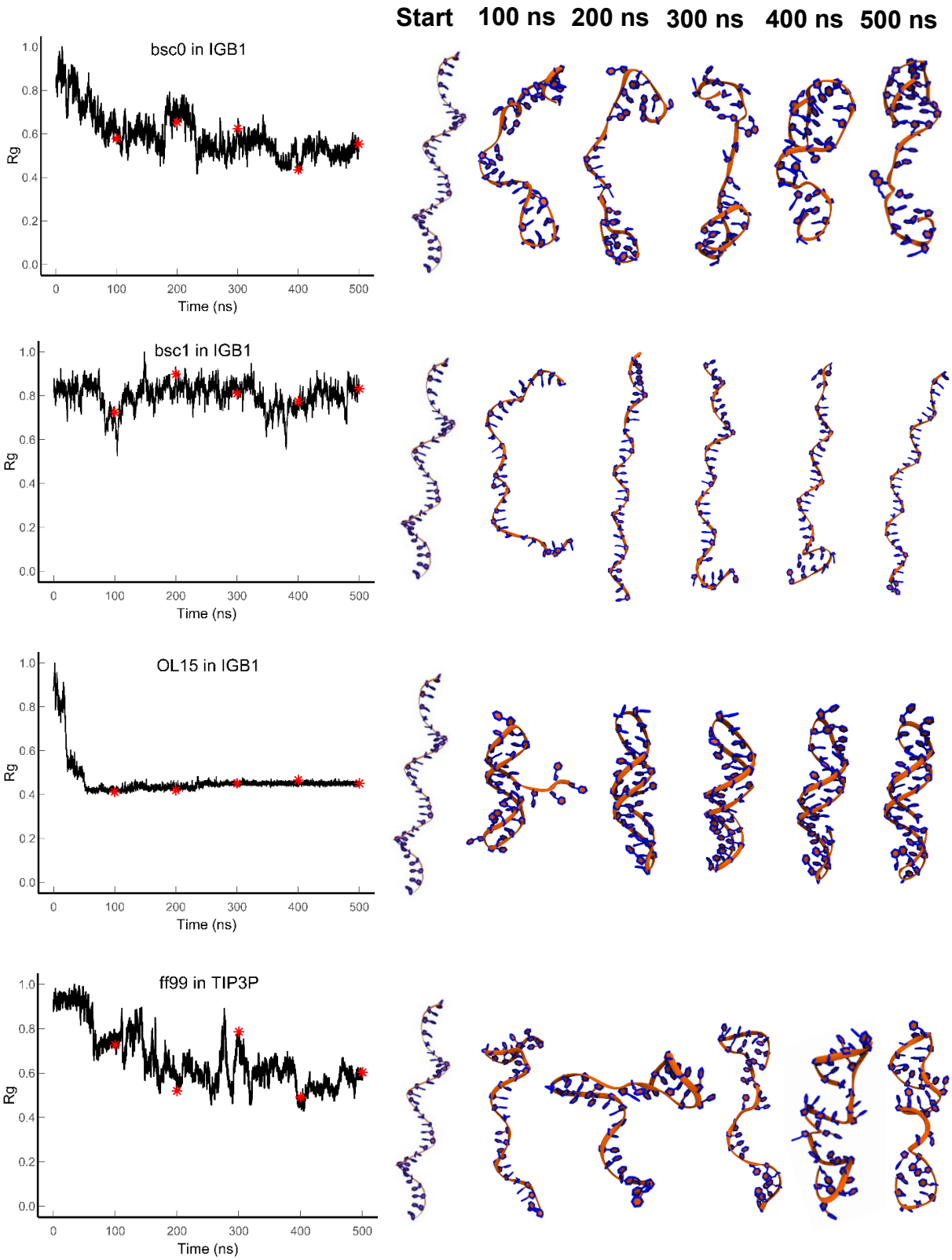
field parameters could lead to the development of a more accurate force field or a fitness function could be used to bias simulations towards experimental results at a quicker pace than traditional MD.

2.1.6 Supplemental Information

Figure S2.1 shows the dynamic nature of ssDNA in simulations such as ff99 in IGB1, IGB5, and IGB8. Simulations such as OL15 in IGB1 are more static which is unexpected due to the instability of ssDNA lending itself to fluctuate as opposed to reaching a single low energy conformation. Thus, at each point there could be a more compact or elongated structure than in the previous time step. The snapshots show an approximate path of ssDNA in each simulation. The red asterisk correspond to the snapshots taken throughout the simulation.

Figure S2.1. This figure provides a temporal profile of the normalized radius of gyration (left). This figure also provides a visual representation of how ssDNA looks throughout each simulation. Each simulation starts in the B-DNA canonical structure. The red asterisk corresponds to the structures at 100, 200, 300, 400, and 500 ns respectively.


ff99 in IGB1
Start
100 ns
200 ns
300 ns
400 ns
500 ns
Rg
Time (ns)
ff99 in IGB5
Rg
Time (ns)
ff99 in IGB8
Rg
Time (ns)


Start
100 ns
200 ns
300 ns
400 ns
500 ns
bsc0 in IGB1
bsc1 in IGB1
OL15 in IGB1
ff99 in TIP3P
Rg
Time (ns)
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500

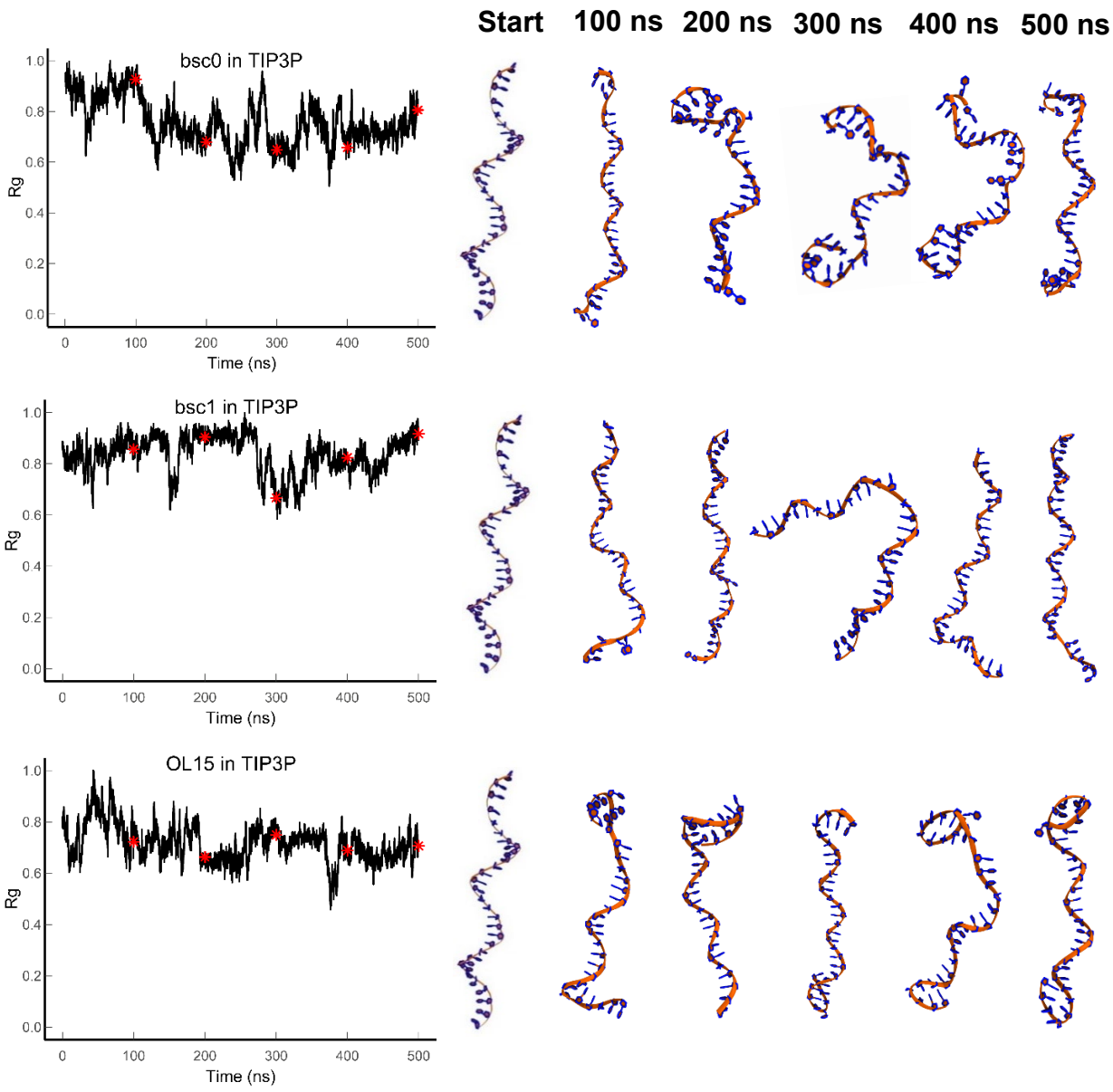

Start
100 ns
200 ns
300 ns
400 ns
500 ns
bsc0 in TIP3P
bsc1 in TIP3P
OL15 in TIP3P
Rg
Time (ns)
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0

Figure S2.2 shows an investigation of how simulation length might impact our conclusions. For this comparison, simulations of ff99 in IGB5 and IGB8 impicit solvent were extended from 500 ns to 2 μs. These simulations were chosen due to their performance being the best compared to experiment. In addition, explicit solvent simulations were not chosen due to the additional computational cost associated with caculations of water molecules surrounding a 30mer ssDNA molecule. A shorter strand would lend itself to fewer water molecules and longer simulation times for this type of comparison. To compound the computational cost and storage of the simulations, the processing time for computing analysis on explicit solvent simulations is significanly longer. In total our study performed 2 μs of simulations in explicit solvent and 6 μs of simulations in implicit solvent. Thus, we believe our choice to extend only the best performing implicit solvent simulations is reasonable.

Overall, there are minimal differences in the simulations frun for 500 ns and 2 μs. This supports our conclusions in the main text based off simulations run for 500 ns. The main difference can be seen on the comparison of R_g and R_{ee} in Figure S2.2 (a). As expected there is a larger ensemble with longer simulation time. This has advantages, however, the shorter simulations

explored identical chain conformations, just at slightly different frequencies. Figures S2.2 (b-e)

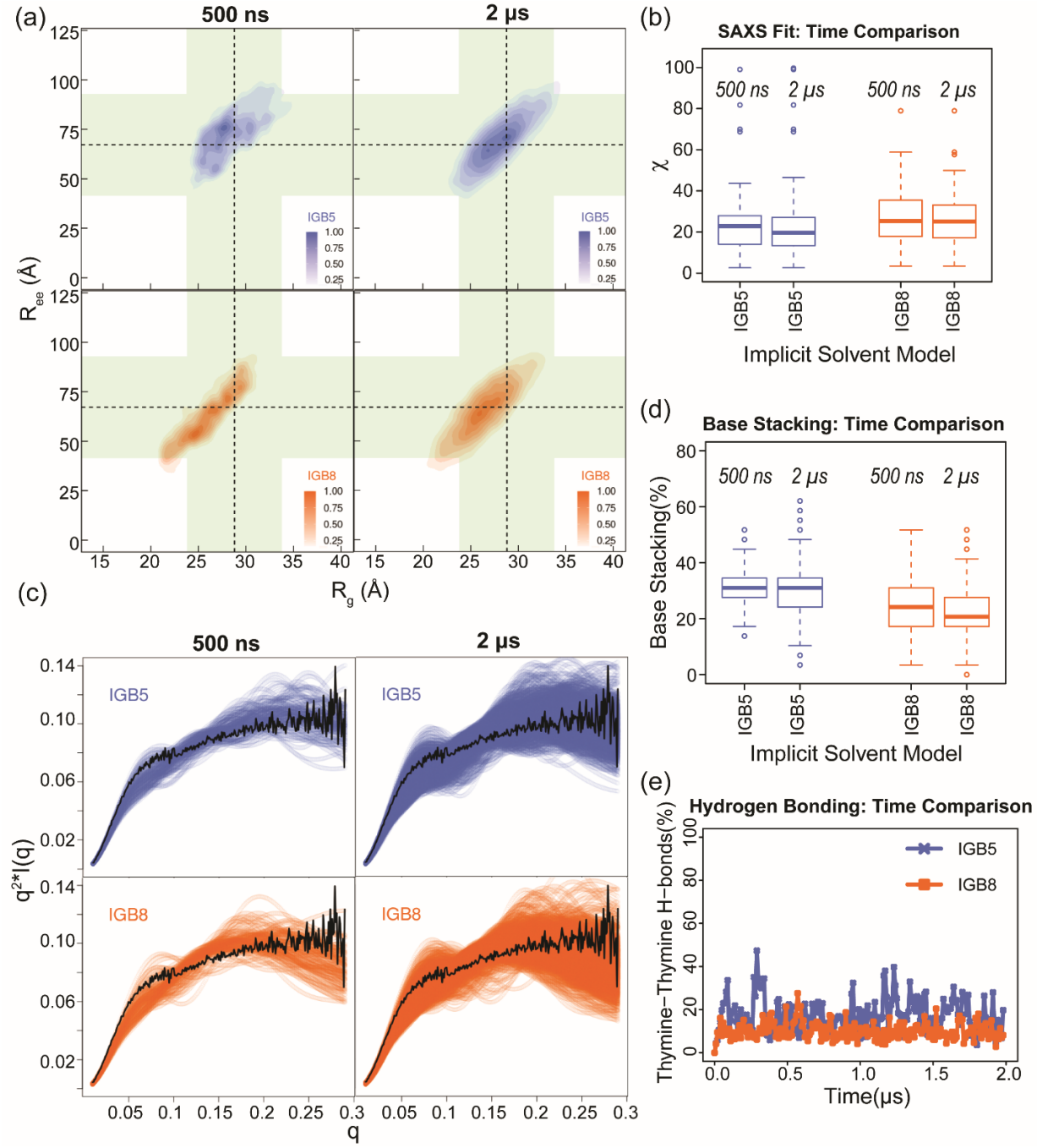


Figure S2.2. This figure provides a time comparison for ff99 in IGB5 and IGB8 (500 ns and 2 µs) of (a) Rg and Ree, (b) SAXS goodness of fit, (c) Kratky plots, (d) base stacking, and (e) hydrogen bonding.

demonstrate that the differences in the explored ensemble seen in Figure S2.2 (a) do not alter the range of explored values, thus our conclusions do not change.

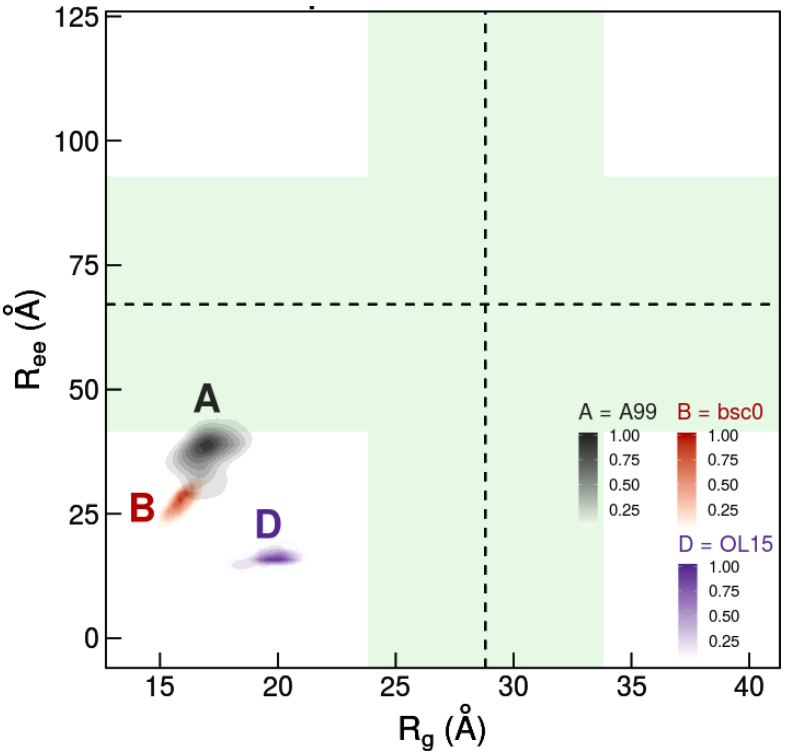


Figure S2.3. Comparison of the R_g and R_{ee} ssDNA values obtained from explicit refinement MD simulations and SAXS and FRET experiments. The data is from the last 100 ns of a 300 ns simulation that started from the final structure of a 500 ns implicit solvent simulation. The simulated force fields are represented by color with ff99 (black), bsc0 (red), and OL15 (purple). Experimental values and errors are represented by the dashed black line and shaded green areas. The legend indicates frequency of occurrence with dark colors (1.00) being highest relative frequency and light colors indicating lower frequency of occurrence.

2.1.7 Explicit Solvent Refinement

Explicit solvent refinement is a process that uses implicit solvent molecular dynamics to efficiently explore conformational space followed by an explicit solvent molecular dynamics step to more accurately calculate the lowest energy structures that were achieved via the implicit solvation calculations.[138] Ultimately, this allows for structures to converge faster with hopes of obtaining the same high accuracy of fully explicit solvent MD. This portion of the study assesses the viability of this approach for studying ssDNA in solution. Specifically, the ff99, bsc0, and

OL15 force fields were simulated in igb1 implicit solvent for 500ns as described in section 2.1.2 The final structure from these implicit solvent simulations served as the starting structure for the explicit solvent simulations, which were run for 300ns following the same protocol as described in section 2.1.2 with the respective ff99, bsc0, and OL15 force fields in TIP3P water.

Figure S2.3 shows a bivariate contour plot for ssDNA R_g and R_{ee} values for each explicit refinement simulation. From this plot the structure of ssDNA appears to significantly deviate from expected experimental values as the ssDNA is very compact. The ff99 and bsc0 force field perform better in implicit solvent before refinement, as the introduction of explicit solvent compacts the ssDNA further through self-interaction and folding. OL15 performs better using explicit refinement than using OL15 in implicit solvent, however, the ssDNA structure is still folded and overly compact. Thus, for OL15, explicit solvent MD from the B-DNA starting conformation performs the best. The Kratky plots shown in Figure S2.4 support the findings that each explicit refinement simulation leads to overly compact ssDNA structures. The presence of the parabolic peak indicates that the ssDNA is ordered to some extent through self-interaction.

These finding suggest explicit solvent refinement is a poor method for studying ssDNA in solution, however this study fails to provide a full scope assessment on the viability of the method. For instance, explicit refinement may be a more useful method for simulations of ssDNA binding to proteins or other substrates. Furthermore, the force field used to describe ssDNA could be changed between implicit and explicit solvent simulations allowing for the best performing force fields to work during each respective step. This would involve using the older ff99 force field in igb5 or igb8 implicit solvent and using the newer bsc0, bsc1, or OL15 force fields in the explicit solvent simulations.

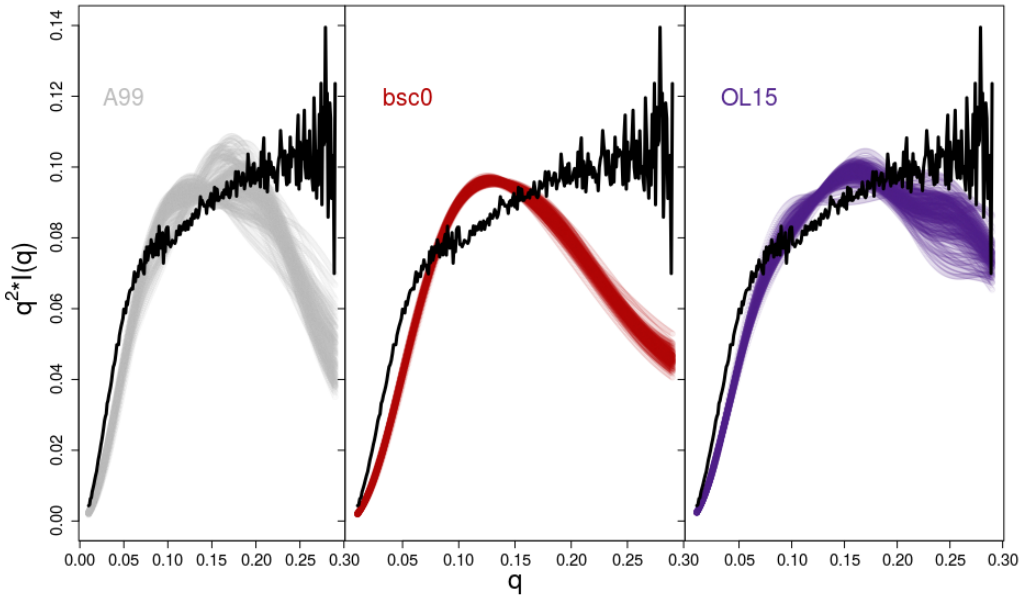


Figure S2.4. Kratky plots of ssDNA structures obtained from explicit refinement simulations with ff99 in grey, bsc0 in red, OL15 in purple, and experimental results in black.

CHAPTER 3: Controlling DNA Structure and Interactions

3.1 Enzymatic Synthesis of Nucleobase-Modified Single-Stranded DNA Offers Tunable Resistance to Nuclease Degradation

* This section is a reproduction of the manuscript for the published work:
Renpeng Gu, Thomas Oweida, Yaroslava G. Yingling, Ashutosh Chilkoti, and Stefan Zauscher. *Biomacromolecules* **2018** *19* (8)*,* 3525-3535.
DOI: 10.1021/acs.biomac.8b00816

3.1.1 Introduction

Nucleic acid-based materials are exploited in a wide range of applications, including detection of mRNA and tumor cells [139–141], gene regulation [142–145], as targeting elements [146–148] and as nano-carriers for drug delivery [149–152]. However, a critical issue for the *in vivo* use of nucleic acid-based materials is their fast degradation by nucleases. In human serum the main extracellular nucleases are 3'-5' exonuclease [153] and endonuclease [154]. To address this issue, two strategies are currently used to enhance DNA's resistance to exo- and endonuclease degradation. In the first strategy, nucleic acid strands are densely packed, such as in spherical nucleic acids (SNAs) and DNA micelles, to enhance the resistance to nuclease degradation by creating steric hindrance and a cloud of monovalent counterions around them [155–159]. In the second strategy, the nucleic acid structure is directly modified to block the binding between nuclease and nucleic acids [160–162]. To date, the chemical modification of nucleic acids mainly focused on the sugar-phosphate backbone, as nucleases degrade the nucleic acid chains via hydrolysis of the phosphate bond [163,164]. Several approaches implement this strategy, including substitution of a non-bridging oxygen in the phosphate backbone with a sulfur atom [165–167], substitution of 2'-H of the ribose sugar with a functional moiety (*i.e.,* 2'-F, 2'-OCH_3 or 2'-NH_2 *etc.*) [168–176], and use of locked nucleic acids [177–180]. Generally, these chemically modified nucleic acids are a poor enzymatic substrate for nuclease binding, leading to retarded degradation compared to unmodified nucleic acids.

An aspect that has been largely overlooked to date is that nucleobase modified nucleic acids offer an effective means to increase resistance to nuclease degradation in general, as we show here. Previously a variety of nucleobase-modified nucleotides and base-functionalized nucleic acids have been made by enzymatic incorporation of unnatural nucleotides [181–184]. Most previous studies of base-functionalized nucleic acids focused on their application in bioanalysis [184–189] and chemical biology, such as the selection of DNAzymes, development of aptamers with improved affinity and thermodynamic stability and small interfering RNAs (siRNAs) with improved potency [182,184,190–195], and only a few studies investigated the influence of base-modified nucleotides, incorporated into DNA chains, on cleavage by restriction endonucleases.[196–198]

We present here an efficient one-pot method to synthesize nucleobase-modified, single-stranded DNA (ssDNA) using terminal deoxynucleotidyl transferase (TdT) enzymatic polymerization. This method allows the incorporation of both natural and base-modified unnatural nucleotides into long ssDNA chains with good control over the molecular weight (MW) and polydispersity of the ssDNA [181,183,199,200]. We systematically investigated the effect of size and density of unnatural nucleobases on ssDNA resistance to degradation upon exposure to Exonuclease I (3'-5' exonuclease), DNase I (endonuclease), and human serum. In addition, we used molecular dynamics (MD) simulations to explore the mechanism underlying the resistance to nuclease degradation.

3.1.2 Methods

3.1.2.1 Materials

All oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA, USA). Recombinant terminal deoxynucleotidyl transferase was purchased from Promega (Madison, WI, USA). Aminoallyl-dUTP (NH_2-dUTP), fluorescein-12-dUTP (FITC-dUTP),

dTTP, agarose, and 2% SYBR Green II were purchased from Thermo Fisher Scientific (Waltham, MA, USA). 5-dibenzylcyclooctyne-dUTP (DBCO-dUTP) and alkyne-dUTP were purchased from Jena Bioscience (Jena, Germany). Aldehyde-dUTP (CHO-dUTP) was synthesized by Duke University's Small molecule synthesis facility. Cyanine 3-dUTP (Cy3-dUTP) was purchased from Enzo Life Science, Inc. (Farmingdale, NY, USA). Sulfo-Cyanine3 N-hydroxysuccinimide (NHS) ester and dimethylformamide (labelling grade) were purchased from Lumiprobe (Hallandale Beach, FL, USA). Exonuclease I (*E. coli*) and DNase I (RNase-free) were purchased from New England BioLabs Inc, (Ipswich, MA, USA). Human serum (male AB clotted whole blood) was purchased from Sigma-Aldrich (St. Louis, MO, USA). Illustra ProbeQuant G-50 microcolumns were purchased from GE Healthcare Life Sciences (Pittsburgh, PA, USA). Microcon-10kDa centrifugal filter unit was purchased from EMD Millipore (Billerica, MA, USA). 10% Mini-PROTEAN TBE Gels, 15 well, 15 µl were purchased from Bio-Rad (Hercules, CA, USA). NANOpure water (18.2 ΩM) was used for all aqueous solution reactions.

3.1.2.2 Synthesis of ssDNA (polyT) using TdT

The reaction mixture consisted of 1 µM Cy5-labeled oligonucleotide initiator 5'-Cy5-dT_{10}-3', 1 mM dTTP and 1 U/µl TdT in TdT buffer (1×, 100 mM potassium cacodylate, 1 mM $CoCl_2$, and 0.2 mM DTT, pH 7.2). The reaction mixture was incubated at 37 °C for 2 h and terminated by heating at 90 °C for 3 min, followed with purification using a Microcon-10kDa centrifugal filter unit.[181,183,200]

3.1.2.3 End-functionalization of ssDNA using TdT

The reaction mixture consisted of 1 µM synthesized 5'-Cy5-polyT-3', 20 µM NH_2-dUTP (aldehyde-dUTP, alkyne-dUTP, DBCO-dUTP, FITC-dUTP, or Cy3-dUTP) and 1 U/µl TdT in

TdT buffer. The reaction mixture was incubated at 37 °C for 2 h and terminated by heating at 90 °C for 3 min, followed with purification using a Microcon-10kDa centrifugal filter unit.

3.1.2.4 Synthesis of ssDNA with various densities of unnatural nucleotides (NH_2-dUTP, CHO-dUTP, alkyne-dUTP, DBCO-dUTP, FITC-dUTP and Cy3-dUTP)

The reaction mixture consisted of 1 μM 5'-Cy5-dT_{10}-3', 1 mM nucleotides (dTTP + unnatural nucleotides) with a range of ratios of unnatural nucleotides to dTTP (0.1, 0.2, 0.5, 1.0, and 2.0 (only for NH_2-dUTP)) and 1 U/μl TdT in TdT buffer. The reaction mixture was incubated at 37 °C for 2 h and terminated by heating at 90 °C for 3 min, followed by purification using a Microcon-10kDa centrifugal filter unit.

3.1.2.5 Sulfo-Cy3 NHS ester and 3-Sulfo-N-succinimidyl benzoate coupling on synthesized poly(T-co-NH_2)

The reaction mixture consisted of 1.25 μM poly(T-co-NH_2) (feeding ratio of NH_2-dUTP/dTTP = 0.1, 0.2, 0.5, 1.0 and 2.0) and sulfo-Cy3 NHS ester or 3-Sulfo-N-succinimidyl benzoate (300 μM, 600 μM, 1.125 mM, 1.67 mM, and 2.2 mM) in 9/10 reaction volume of 100 mM sodium bicarbonate buffer (pH=8.5) and 1/10 reaction volume of DMF. The reaction mixture was incubated at room temperature in a shaker overnight, followed by purification using illustra ProbeQuant G-50 micro columns and a Microcon-10kDa centrifugal filter unit.

3.1.2.6 Copper-catalyzed click reaction on synthesized poly(T-co-alkyne)

The reaction mixture consisted of 0.25 μM poly(T-*co*-alkyne), 0.1 M triethylammonium acetate buffer, pH 7.0, 0.1 mM azide-Cy3, 0.5 mM ascorbic acid and 0.5 mM Copper(II)-TBTA in water and DMSO (1:1 in volume). The reaction mixture was degassed by bubbling N_2 gas for 2 mins and then incubated at room temperature in a shaker overnight, followed by purification using illustra ProbeQuant G-50 micro columns and a Microcon-10kDa centrifugal filter unit.

3.1.2.7 Copper free click reaction on synthesized poly(T-co-DBCO)

The reaction mixture consisted of 0.35 μM poly(T-*co*-DBCO) and 62.5 μM azide-Cy3 in H_2O and t-BuOH (1:1 in volume). The reaction mixture was incubated at room temperature in a shaker overnight, followed with purification using illustra ProbeQuant G-50 micro columns and a Microcon-10kDa centrifugal filter unit.

3.1.2.8 Degradation of ssDNA in the presence of exo- and endonucleases

The degradation reaction mixture consisted of 64 ng/μl (or 100 ng/μl) synthesized ssDNA, 0.02 U/μl Exonuclease I (or 0.02 U/μl DNase I) in Exonuclease I reaction buffer (1×, 67 mM Glycine-KOH, 6.7 mM $MgCl_2$, 10 mM β-ME, pH 9.5 @ 25 °C) or DNase I reaction buffer (1×, 10 mM Tris-HCl, 2.5 mM $MgCl_2$, 0.5 mM $CaCl_2$, pH 7.6 @ 25 °C). The mixture was incubated at 37 °C. At each time point, 4 μl was taken and added into 2 μl 100 mM EDTA followed by heating at 90 °C for 3 min. All samples were analysed by polyacrylamide gel electrophoresis.

3.1.2.9 Stability of ssDNA in human serum

The degradation reaction mixture consisted of 30 ng/μl synthesized polyT (poly(T-*co*-NH_2) 0.5 or poly(T-*co*-Cy3) 0.5) in 85% human serum. The mixture was incubated at 37 °C for up to 3 days. At each time point, 1 μl was taken and added to 4 μl 1×TBE and 5 μl 2×Urea sample loading buffer (1×, 89 mM Tris-HCl, 89 mM boric acid, 2 mM EDTA, 7 M urea, 12% Ficoll, pH 8.0). All samples were analysed by agarose gel electrophoresis.

3.1.2.10 Characterization

Gel electrophoresis was conducted on a mini-PROTEAN® tetra vertical electrophoresis cell by loading a 2 μl sample (~0.2 μM) and 2×RNA loading buffer (loading dyes were removed) into a 10% Mini-PROTEAN® TBE gel purchased from Bio-Rad Labs, and then applying 110 V for 60 min. The gels were imaged with a Typhoon 9410 scanner (GE Healthcare Life Science,

Piscataway, NJ) at 633 nm (Cy5) or 532 nm (Cy3) laser excitation. The fluorescence intensity of CHO, FITC, Cy5 and Cy3 was measured on a Nanodrop^{TM} 3300 fluorescence spectrometer (Thermo Scientific). The concentration of ssDNA was measured on a Nanodrop^{TM} 1000 UV-vis spectrophotometer (Thermo Scientific).

3.1.2.11 All-atom molecular dynamics (MD) simulations

Nucleic Acid Builder (NAB) software was used to generate a 40-mer of the B-form adenine-thymine double helix [201]. The complimentary adenine strand was subsequently deleted using BIOVIA Discover Studio, to create a 40-mer single stranded thymine DNA molecule (polyT). Using the same software, poly(T-*co*-NH_2) and poly(T-*co*-Cy3) were created by replacing the methyl group of the thymine with -NH_2 and -Cy3, respectively [202]. Poly(T-*co*-NH_2) and poly(T-*co*-Cy3) were formed with unnatural nucleotide densities of 5%, 7.5%, 20%, 35%, and 45% to stay consistent with experiment. The unnatural nucleotide placement was chosen randomly but kept consistent between poly(T-*co*-NH_2) and poly(T-*co*-Cy3). The partial charges for poly(T-*co*-NH_2) and poly(T-*co*-Cy3) were calculated using GAMESS-US and the RESP charge method (Red Server) [203].

PolyT, poly(T-*co*-NH_2) and poly(T-*co*-Cy3) were simulated with the AMBER99 force field and the Generalized Born (GB) implicit solvent model at a 50 mM salt concentration. While there has not been a systemic evaluation of DNA in implicit solvent, improvements in the GB model have been shown to accurately predict the energetics and structures of nucleic acids [102,204]. The generalized amber force field (GAFF) was used for the unnatural nucleotides [205]. All starting structures for implicit simulations were subjected to minimization for 10,000 steps, followed by an unconstrained heating to 300 K, and an MD equilibration using a Berendsen thermostat. The production MD runs were performed at 300 K for 300 ns using a 1 fs time step. Protocol details

for simulations were reported previously [10,133,206]. Our simulation setup was validated via the agreement between the simulations and Forster Resonance Energy Transfer (FRET) measurements of the polyT end-to-end distance as well as the agreement between the simulations and Small Angle X-ray Scattering (SAXS) measurements of radius of gyration (R_g) [12,13,93,207].

3.1.2.12 Analysis of all-atom MD simulations

The mass-weighted root-mean-square fluctuations (RMSF) was calculated per residue for each ssDNA strand's backbone with the CPPTRAJ 15 module in the AMBER 16 package [208]. RMSF represents the positional standard deviation with reference to the average coordinates over the last 100 ns of the simulation trajectory. The per-residue, backbone RMSF values were summed for all modified nucleotides and their adjacent nucleotides in each ssDNA strand. A ΔRMSF value is reported based on the equation

$$\Delta RMSF = \sum_{R_N} RMSF_{PolyT_M(R_N)} + RMSF_{PolyT_M(R_N-1)} + RMSF_{PolyT_M(R_N+1)} - \sum_{R_N} RMSF_{PolyT(R_N)} + RMSF_{PolyT(R_N-1)} + RMSF_{PolyT(R_N+1)}, \quad (3.1.1)$$

where R_N represents the set of modified nucleobases, $polyT_M$ represents the modified ssDNA strands (*i.e.*, poly(T-*co*-NH_2) and poly(T-*co*-Cy3)), and polyT represents unmodified ssDNA strands. Each summation counted each residue a maximum of one time, thus $R_N \neq R_{N-1} \neq R_{N+1}$ across each set of modified nucleotides. Solvent accessible surface area (SASA) was calculated via an in-house TCL script and VMD 1.9.2 [209]. The SASA was calculated for the ssDNA backbones with a commonly used probe radius of 1.4 Å to elucidate the steric hindrance associated with ssDNA's conformation and incorporation of unnatural nucleotides.

3.1.3 Results and Discussion

3.1.3.1 3' end-functionalized ssDNA in the presence of Exonuclease I

To investigate the effects of unnatural nucleobase size on ssDNA stability upon exposure to exonuclease, we separately modified the 3'-end of polyT with six different types of unnatural nucleotides, *i.e.*, NH_2-dUTP, CHO-dUTP, alkyne-dUTP, DBCO-dUTP, FITC-dUTP and Cy3-dUTP [181,210–214] (Figure 1A) using TdT enzymatic polymerization. Similar to previous studies, our polyTs were successfully tail-functionalized with these base-modified nucleotides at the 3'-end.[183,215,216] These 3'-end functionalized ssDNAs were treated with Exonuclease I. As shown in Figure 2, unmodified polyT was degraded fastest while ssDNA with bulky nucleobases at their 3'-end (DBCO, FITC and Cy3 modified polyT) were degraded significantly slower. These results suggest that bulky nucleobases increase the resistance of 3' end-functionalized ssDNA to Exonuclease I degradation.

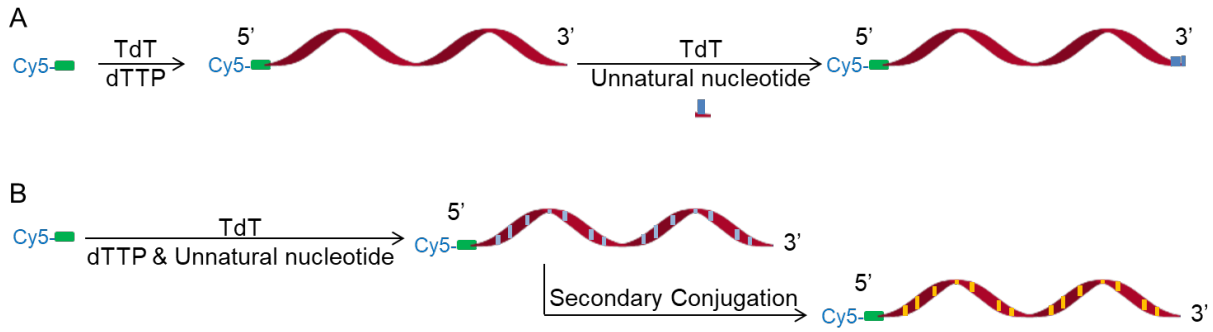


Figure 3.1.1. Illustration of (**A**) synthesis of 3' end-functionalized ssDNA and (**B**) internal-functionalized ssDNA via TdT enzymatic polymerization and secondary conjugation.

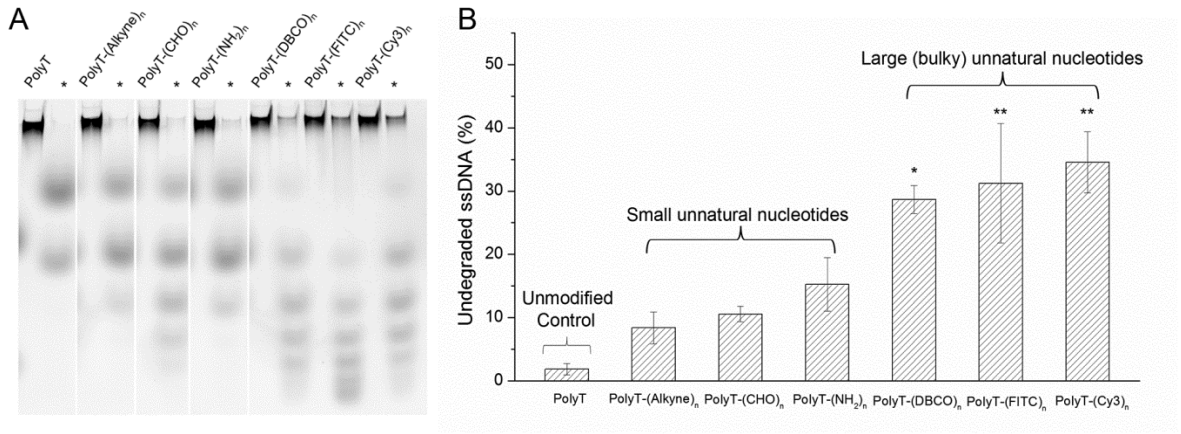


Figure 3.1.2. (**A**) Gel electrophoresis image showing Exonuclease I degradation of 3'-end functionalized ssDNA (* Incubation time: 30 mins). (**B**) Effect of unnatural nucleobase size on the stability of 3'-end functionalized ssDNA in the presence of Exonuclease I for 30 min incubation. (* $p<0.05$, ** $p<0.01$, one-way ANOVA with post-hoc Tukey HSD test compared to the polyT control)

Exonuclease I is 3'-5' exonuclease, and thus has to interact with the 3'-terminal nucleotide first before sequentially cleaving subsequent nucleotides in a ssDNA chain. During the hydrolytic degradation, Exonuclease I interacts with both the ssDNA backbone and the nucleobases; more specifically, the terminal nucleobase binds to a hydrophobic pocket formed by the side chains of Exonuclease I [217]. Therefore, certain unnatural nucleobases present at the 3'-end may not fit the hydrophobic pocket of Exonuclease I well. This disrupts enzyme binding with the ssDNA backbone and thus prevents the effective hydrolysis of the phosphodiester bond. Compared to the relatively small, alkyne-, CHO- and NH_2- functionalized nucleobases, the bulky, DBCO-, FITC- and Cy3- functionalized nucleobases fit more poorly into the enzyme's hydrophobic pocket, this interferes with nuclease binding even more significantly, which is manifest in an enhanced stability of ssDNA upon exposure to Exonuclease I. These two scenarios are illustrated by the cartoon in Figure 3. Similar phenomena were also observed in other studies where a restriction endonuclease cleaved specific DNA sequences that contained nucleobase-modified nucleotides.[196–198] In those studies, most endonucleases tolerated a small base modification in the specific sequence, but not

a bulky one, which is in good agreement with the findings in our study. Although Exonuclease I is not a sequence-specific nuclease, the size of the nucleobase modification still plays a significant role for the DNA cleavage by exonuclease.

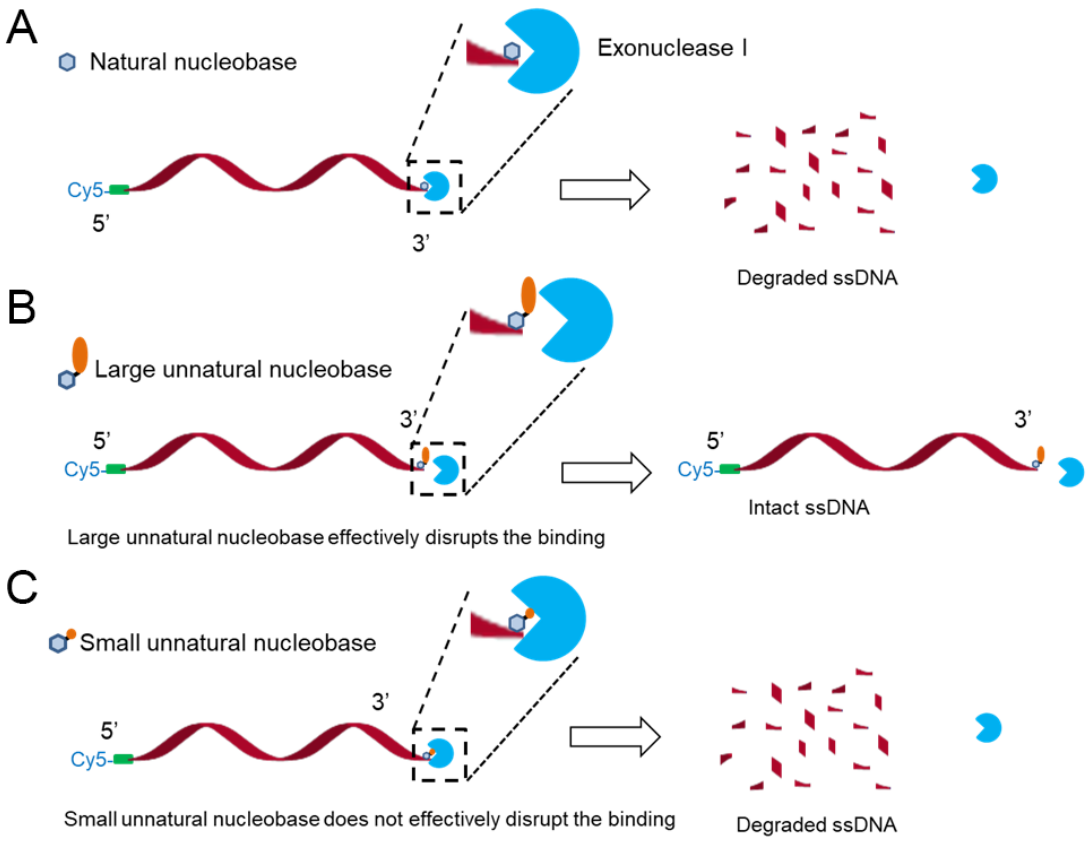


Figure 3.1.3. Schematic illustration of Exonuclease I degradation of (**A**) ssDNA, and of ssDNA with a (**B**) large and (**C**) a small unnatural nucleobase at the 3'-end.

In addition, according to our previous experience, 3'-end functionalized ssDNA amphiphiles can self-assemble into micellar structures [183,218]. We note that DBCO and FITC are both hydrophobic, which may induce ssDNA self-assembly. If that were the case, the 3'-end of the ssDNA is buried in the hydrophobic core of the resulting DNA micelles, and the observed enhanced resistance to nuclease degradation could thus largely arise from the micellar self-assembly rather than from the poor fit between an unnatural nucleobase and the exonuclease. To

rule out this possibility, we decreased the concentration of polyT-$(DBCO)_n$ and polyT-$(FITC)_n$ in the degradation system to prevent self-assembly (no assemblies were detected using dynamic light scattering, data not shown here) (Figure S1). Compared to polyT control, both functionalized ssDNAs at low concentration ─where the ssDNA would exist as free polymer chains─ still showed significant stability to degradation by Exonuclease I (Figure S2). Hence, it is likely that the unnatural nucleobases present at the 3'-end disrupt DNA-exonuclease binding and retard DNA degradation rather than the lack of access of nuclease to the 3'-end of the ssDNA.

3.1.3.2 Internally-functionalized ssDNA in the presence of Exonuclease I and DNase I

To investigate the effects of unnatural nucleobase density and size on ssDNA stability upon exposure to both exo- and endonuclease, we synthesized ssDNA with a controlled density of unnatural nucleobases along the polynucleotide chain. We refer to this type of ssDNA as "internally-functionalized ssDNA". We first studied the incorporation efficiency of six different, unnatural nucleotides. In these experiments, we kept the nucleotide (monomer)/oligonucleotide (initiator) ratio constant at 1000/1, and systematically varied the feed ratio of unnatural nucleotides/dTTP from 0.1 to 2.0 (Figure 1B). As the feed ratio of unnatural nucleotides/dTTP increases (as shown in Table 1 and Figure S3), the MW of the synthesized ssDNA decreases dramatically, except for the incorporation of NH_2-dUTP (see supplementary materials for calculation details). Some bulky, unnatural nucleotides (DBCO-dUTP and FITC-dUTP) nearly prohibited enzymatic polymerization when the unnatural nucleotide/dTTP ratio reached 0.5.

Table 3.1.1. Average degree of polymerization of ssDNA with different types of incorporated unnatural nucleotides (calculated from Figure S3).

	Feed ratio of unnatural nucleotide/dTTP				
	0.1	0.2	0.5	1.0	2.0
Unnatural nucleotides	Average number of nucleotides per chain (degree of polymerization)				
NH_2-dUTP	~730	~740	~750	~690	~640
CHO-dUTP	~350	~260	~150	~100	-
Alkyne-dUTP	~640	~480	~200	~130	-
DBCO-dUTP	~260	~110	-	-	-
FITC-dUTP	~180	~130	-	-	-
Cy3-dUTP	~490	~260	~180	~140	-
No unnatural nucleotides (PolyT)	~800				

In addition, we calculated the density of incorporated unnatural nucleotides (Table 2, see supplementary materials for calculation details). Compared to the other unnatural nucleotides, NH_2-dUTP had the highest incorporation density, especially when the ratio of unnatural nucleotide/dTTP was equal or larger than 0.5. Differences in the incorporation efficiency between natural and unnatural nucleotides are not unexpected. For other DNA polymerases it has been demonstrated that the modification of the nucleobase structure impacts the incorporation efficiency of such modified nucleotides into the DNA strand [219–222]. In many cases, the structural change of

the enzyme-substrate complex renders nucleobase-modified nucleotides a worse substrate for DNA polymerases. However, other nucleobase-modified nucleotides were tolerated or even more efficiently processed, because the interactions between modified nucleobase and amino acids near the active site stabilize the enzyme-substrate complex [221]. A similar argument may be invoked for TdT to explain why one nucleobase-modified nucleotide is better accepted than another. However, to gain structural insights into the processing of nucleobase-modified nucleotides by TdT requires further crystal structure analysis of the TdT-substrate complex.

Table 3.1.2. Incorporation density of different types of unnatural nucleotides in ssDNA.

	Feed ratio of unnatural nucleotide/dTTP				
	0.1	0.2	0.5	1.0	2.0
Unnatural nucleotides	Density of incorporated unnatural nucleotides (per 100 nucleotides)				
NH_2-dUTP	4.4 ± 0.2	7.3 ± 1.3	20.8 ± 3.9	36.2 ± 0.9	46.3 ± 2.0
CHO-dUTP	6.3	4.2	2.7	1.9	-
Alkyne-dUTP	2.2	2.2	-	-	-
DBCO-dUTP	2.7	-	-	-	-
FITC-dUTP	10.2	7.2	-	-	-
Cy3-dUTP	4.3	7.2	4.1	2.6	-

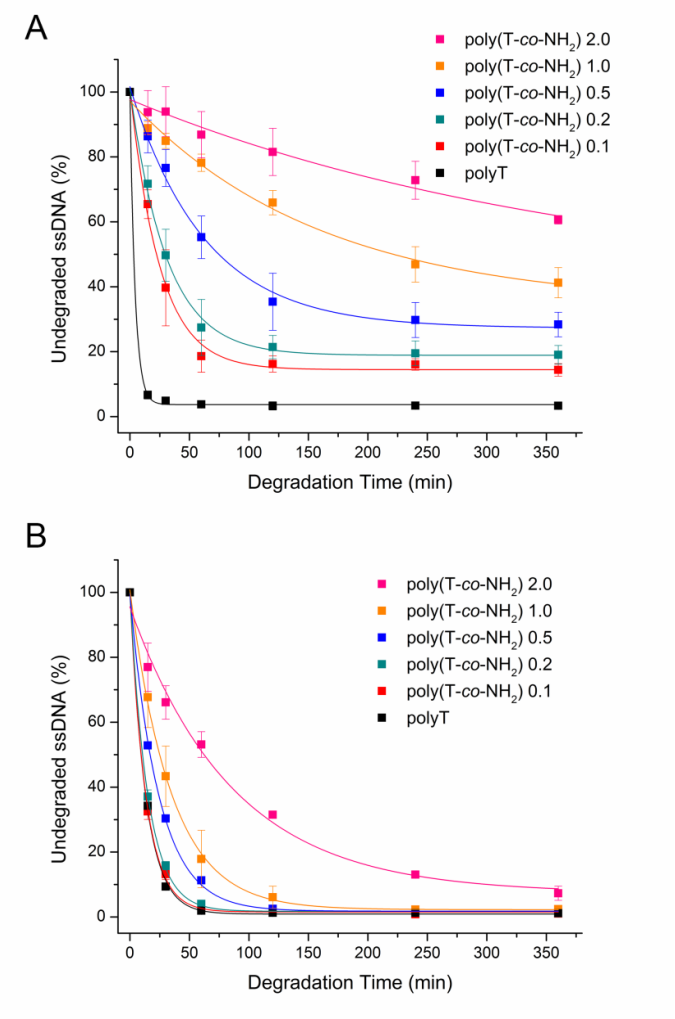


Figure 3.1.4. Degradation kinetics of polyT and different poly(T-co-NH2) by (A) Exonuclease I and (B) DNase I.

Because NH_2-dUTP exhibited the highest incorporation efficiency and because it can be readily functionalized to tune the size of unnatural nucleobase, we used it for the degradation study of internally-functionalized ssDNA. To investigate the effects of NH_2-functionalization density on ssDNA resistance to nuclease degradation, we treated the NH_2- internally-functionalized ssDNA with two types of nucleases: Exonuclease I and DNase I. As shown in Figure 4, Figures S4-7 and Tables 3 and 4, after treatment with the two nucleases, the rate and extent of degradation correlates inversely with the incorporation density of NH_2-dUTP. Although more complex models for exonuclease degradation are available [223,224], we chose a simple exponential decay function fit as

we are interested in comparing the half-life of different synthesized ssDNA. As shown in Figure 4, this simple model fits the data well and yields an estimate of the half-life which thus provides a single, useful parameter that enables the quantitative comparison across different polynucleotides.

Table 3.1.3. Half-life of poly(T-*co*-NH_2) and poly (T-*co*-Cy3) upon exposure to Exonuclease I.

	t_{half} (min)		t_{half} (min)
polyT	3.2 ± 0.2	polyT	3.2 ± 0.2
poly(T-*co*-NH_2) 0.1	23.1 ± 5.6	poly(T-*co*-Cy3) 0.1	76.4 ± 4.1
poly(T-*co*-NH_2) 0.2	38.4 ± 11.7	poly(T-*co*-Cy3) 0.2	126.6 ± 55.2
poly(T-*co*-NH_2) 0.5	104.9 ± 31.0	poly(T-*co*-Cy3) 0.5	265.3 ± 36.1
poly(T-*co*-NH_2) 1.0	273.2 ± 37.5	poly(T-*co*-Cy3) 1.0	> 360
poly(T-*co*-NH_2) 2.0	> 360	poly(T-*co*-Cy3) 2.0	> 360

Poly(T-*co*-NH_2) X refers to ssDNA synthesized with a feed ratio of NH_2-dUTP/dTTP=X. Poly(T-*co*-Cy3) X refers to ssDNA after Cy3 NHS-ester conjugation to poly(T-*co*-NH_2) X (assuming a conjugation efficiency of 100%).

Table 3.1.4. Half-life of poly(T-*co*-NH_2) and poly (T-*co*-Cy3) upon exposure to DNase I.

	t_{half} (min)		t_{half} (min)
polyT	9.3 ± 0.2	polyT	9.3 ± 0.2
poly(T-*co*-NH_2) 0.1	9.4 ± 0.7	poly(T-*co*-Cy3) 0.1	-
poly(T-*co*-NH_2) 0.2	10.6 ± 0.5	poly(T-*co*-Cy3) 0.2	17.0 ± 3.5
poly(T-*co*-NH_2) 0.5	17.0 ± 0.2	poly(T-*co*-Cy3) 0.5	25.4 ± 3.2
poly(T-*co*-NH_2) 1.0	34.2 ± 9.2	poly(T-*co*-Cy3) 1.0	30.4 ± 1.6
poly(T-*co*-NH_2) 2.0	73.2 ± 13.2	poly(T-*co*-Cy3) 2.0	63.2 ± 12.9

The data for NH_2- internally-functionalized ssDNA in Table 3 show that the half-life increases with increasing incorporation density of NH_2-dUTP, which indicates the increased resistance of ssDNA to exonuclease degradation with increasing NH_2-dUTP incorporation density. This is because that for internally-functionalized ssDNA, as the degradation proceeds, each incorporated NH_2-dUTP along the ssDNA chain will interfere with enzyme binding and will lead to a stepwise, retarded degradation (Figure 5A). Therefore, when the density of incorporated NH_2-dUTP increases, the degradation kinetics of Exonuclease I will decrease.

In contrast, as shown in Figure 4B, in the endonuclease treatment, significant resistance to degradation was achieved only at high NH_2-dUTP incorporation density. DNase I indiscriminately cleaves the phosphate bonds within a ssDNA chain (Figure 5B). Previous research showed that when DNase I binds the DNA backbone, the nucleobases typically point away from the DNase I without specific interactions [225–227]. Although the nucleobases seem not to interact with DNase I directly, our DNase I degradation results (Figure 4B) show that increasing the incorporation

density of NH_2-dUTP enhances ssDNA resistance to endonuclease degradation significantly, especially at high incorporation density. In addition, ssDNA degradation by DNase I is likely correlated with local DNA flexibility [228], and several studies showed that increased conformational constraint of ssDNA can lead to enhanced resistance to DNase I degradation [229,230]. Therefore, the mechanism underlying the enhanced resistance to DNase I degradation in our study likely arises

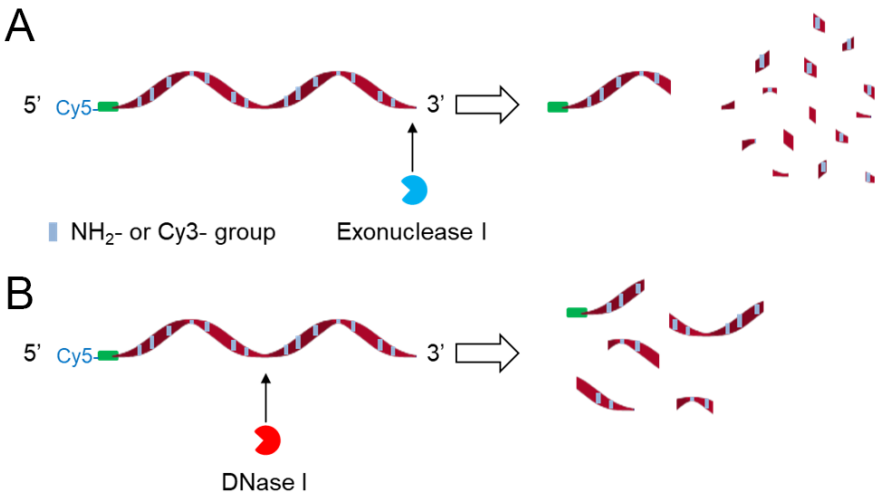


Figure 3.1.5. Schematic illustration of (**A**) Exonuclease I and (**B**) DNase I degradation of NH_2-internally functionalized polyT.

from the local, conformational constraint caused by the incorporated NH2-dUTP. To elucidate the local chain flexibility of ssDNA as a function of unnatural nucleobase (-NH2) density, MD simulations were used to determine the local, mass-weighted root-mean-square fluctuations (RMSF), associated with the unnatural nucleotides in the ssDNA backbone (Figure 6A). Figure 6B shows that the ΔRMSF of poly(T-co-NH2), compared to polyT, decreases with increasing unnatural nucleobase density, which means more local regions of poly(T-co-NH2) become inflexible. Therefore, when the incorporation density of NH2-dUTP is low (4.4 ± 0.2 % and 7.3 ± 1.3% as shown in Table 4), DNase I has a high probability of binding to a conformationally

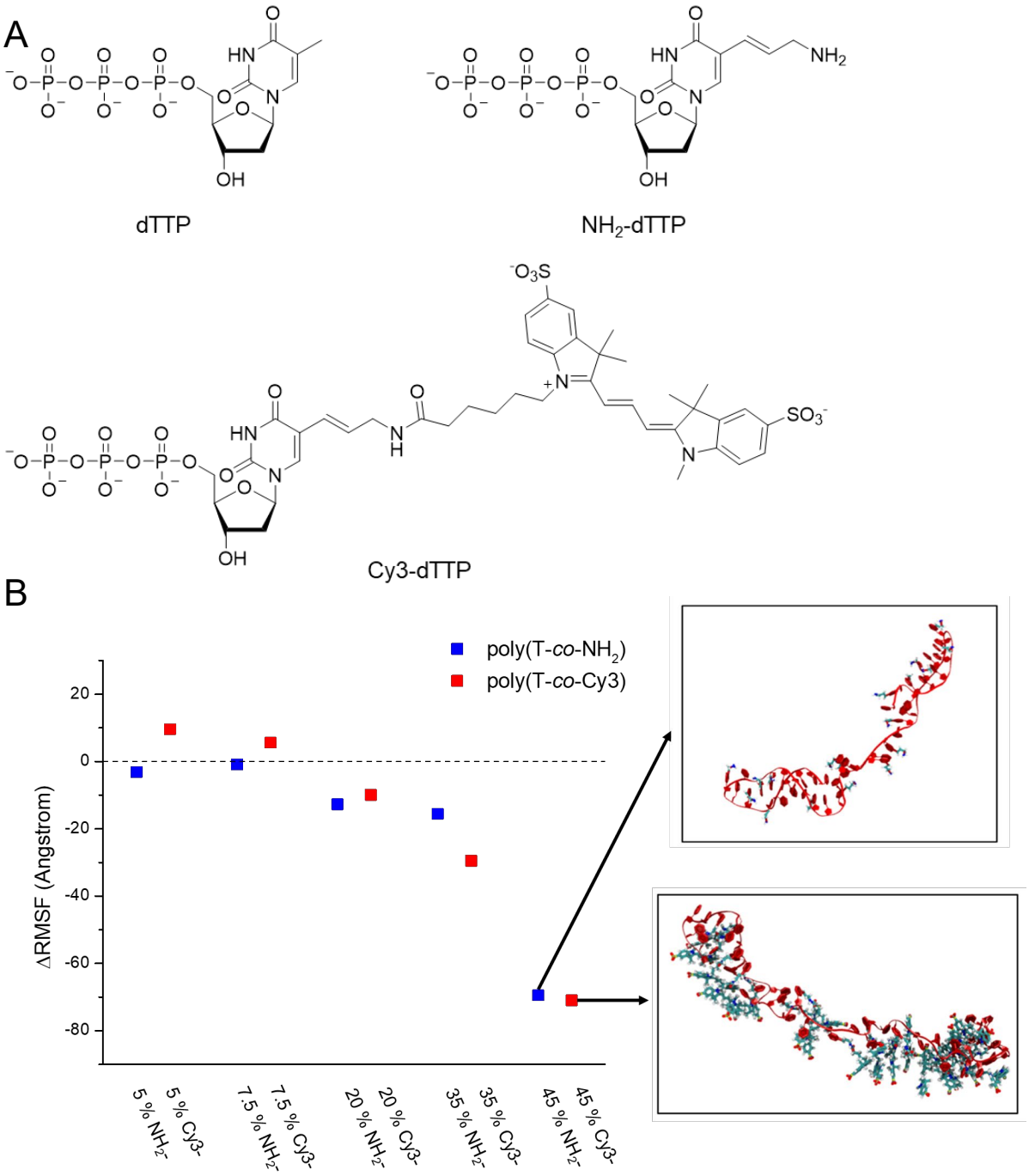


Figure 3.1.6. (A) Chemical structures of dTTP, NH2-dUTP and Cy3-dUTP used in the simulations. (B) Simulation results showing the change in local flexibility (ΔRMSF) of poly(T-co-NH2) and poly(T-co-Cy3) compared to that of polyT, as a function of incorporation density and type of unnatural nucleobases. The simulation snapshots at (45 % incorporation density) depict the modified ssDNA structures, where the sugars, the phosphate groups, and the natural bases are colored red, and where for the modified nucleobase structures, the carbon atoms are colored cyan, hydrogen atoms are colored white, nitrogen atoms are colored blue, oxygen atoms are colored red, and sulfur atoms are colored yellow. (B) Chemical structure of dTTP, NH2-dUTP and Cy3-dUTP used in simulations.

unconstrained ssDNA chain segment, and thus readily cleaves the ssDNA strand into smaller ssDNA fragments. As the NH_2-dUTP incorporation density increases, the increasing local conformational constraints in ssDNA make effective DNase I binding less likely and thus prevent ssDNA cleavage.

Furthermore, to investigate the size effect of unnatural nucleobase on resistance of internally-functionalized ssDNA to nuclease degradation, we conjugated Cy3 groups onto the poly(T-*co*-NH_2) via NHS ester coupling chemistry to increase the unnatural nucleobase size (Figure 6A). After this modification, the ssDNA resistance to both exo- and endonuclease degradation were further enhanced (Tables 3 and 4, Figures S8-10). Specifically, in exonuclease degradation, as shown in Figure 7A, the change in the degradation rate for poly(T-*co*-NH_2) is larger than that for poly(T-*co*-Cy3), which suggests that exonuclease faces additional difficulties in binding to an unnatural nucleotide with large nucleobase modification, as schematically shown in Figures 3B and C. In contrast, for endonuclease degradation (Figure 7B), the change in the degradation rate for poly(T-*co*-NH_2) is only slightly greater than that for poly(T-*co*-Cy3). This suggests that the size of the unnatural nucleobase may not play a significant role in determining the resistance of ssDNA to endonuclease degradation. This is also supported by our simulation results (Figure 6B) which show that the substantial size difference between the Cy3 moiety and NH_2 moiety on the nucleobase does not affect the local chain flexibility (ΔRMSF) significantly.

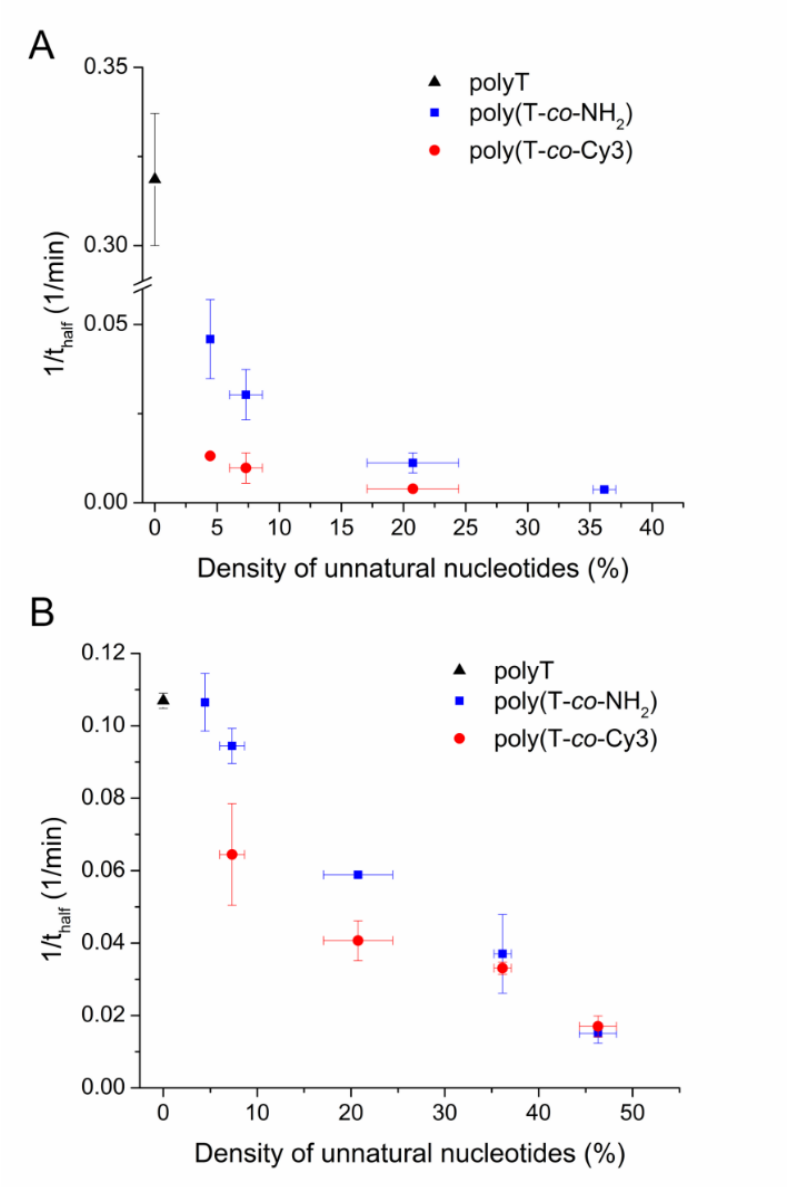


Figure 3.1.7. (**A**) Exonuclease I and (**B**) DNase I degradation rates of polyT, poly(T-*co*-NH_2) and poly(T-*co*-Cy3) plotted as a function of incorporation density of NH_2- and Cy3- moieties.

However, we also noticed that unlike Exonuclease I which degrades ssDNA into single nucleotides, DNase I degrades ssDNA into small fragments (Figures S6-7). During the degradation process, DNase I not only cleaves intact ssDNA, but also further cleaves degraded ssDNA fragments. Therefore, besides the fraction of residual, intact ssDNA, the fraction and size of the degraded ssDNA fragments are also an important parameter to assess the resistance of ssDNA to

DNase I degradation. However, our synthesized polyT and poly(T-*co*-NH_2) are only labelled with a Cy5 molecule at the 5' end, and thus the degraded fragments, without the Cy5 label, do not show up in the gel chromatogram. To visualize these degraded fragments by gel electrophoresis, we replaced a fraction of dTTP with other natural nucleotides (dATP, dCTP and dGTP) in the enzymatic polymerization reaction, so that the resulting ssDNA and also the degraded fragments now have secondary structure which can trap SYBR Green staining dye. We first tested different combinations of the natural nucleotides for efficient ssDNA synthesis (Tables S1-2). We found that all ssDNA reaction products could be stained (Figures S11-12), and that the combination of dTTP, dATP and dCTP (2' & 4' in Table S1) and NH_2-dUTP, dATP and dCTP (2 & 4 in Table S2) resulted in the narrowest MW distributions. These latter ssDNAs were chosen for Cy3 NHS-ester conjugation (Figure S13) and subsequent DNase I degradation analysis.

Although, after incorporating NH_2- or Cy3- unnatural nucleobases, the amount of intact ssDNA remaining after degradation with DNase I is about the same for all three types of co-polynucleotides, the degradation patterns are very different (Figures S14-17). While degraded poly(Cy3-*co*-A-*co*-C) contains many large ssDNA fragments, degraded poly(NH_2-*co*-A-*co*-C) contains fewer large ssDNA fragments, and degraded poly(T-*co*-A-*co*-C) does not contain any large ssDNA fragments. This suggests that as the size of the unnatural nucleobases increases, the number of DNase I induced cleavages of the ssDNA backbone decreases. Therefore, by only considering the remaining amount of intact ssDNA, we underestimate the poly(T-*co*-Cy3) resistance to endonuclease degradation. Compared to poly(T-*co*-NH_2), poly(T-*co*-Cy3) still demonstrated enhanced resistance to endonuclease degradation.

Since our ΔRMSF values from simulation of poly(T-*co*-NH_2) and poly(T-*co*-Cy3) do not show significant difference, the enhanced poly(T-*co*-Cy3) resistance to endonuclease degradation

does most likely not arise from the change of local chain flexibility. Thus we hypothesized that the bulky Cy3 moiety provides stronger steric hindrance than the NH_2 moiety, which makes endonuclease-ssDNA backbone binding more difficult and enhances the resistance to endonuclease degradation. To evaluate the steric hindrance provided by the incorporated unnatural nucleobases (NH_2- and Cy3-), we measured the solvent accessible surface area (SASA) of the ssDNA backbone using MD simulations. A lower SASA thus suggests that the ssDNA backbone is less accessible to a solvent molecule. Since an endonuclease molecule is much larger than a solvent molecule, a lower SASA value also suggests that the ssDNA backbone is less accessible to endonuclease. As shown in Figure 8, with increasing incorporation density of Cy3 unnatural nucleobases, the SASA of the ssDNA backbone decreases significantly. This is in contrast to ssDNA that contains NH_2-dUTP. Therefore, the incorporation of Cy3 moieties makes the ssDNA backbone less accessible to endonuclease binding. Thus, both density and size of the unnatural nucleobases are the key factors determining the internally-functionalized ssDNA resistance to endonuclease degradation.

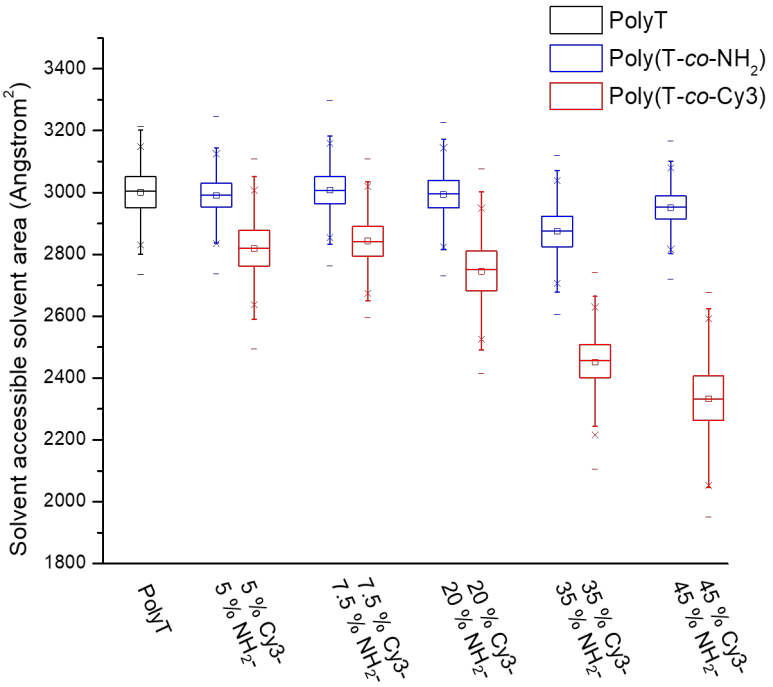


Figure 3.1.8. Simulation results showing the SASA as a function of incorporation density of poly(T-*co*-NH_2) and poly(T-*co*-Cy3).

3.1.3.3 Stability of internally-functionalized ssDNA in human serum

At last, we evaluated the stability of the synthesized ssDNA in human serum mimicking the *in vivo* environment that contains both exonuclease and endonuclease. We chose one set of internally-functionalized ssDNA (poly(T-*co*-NH_2) 0.5 and poly(T-*co*-Cy3) 0.5) and incubated them in 85% human serum for up to 3 days. As shown in Figure 9, Figure S18 and Table S3, the half-life of polyT is ~4 h (control), while the half-life of poly(T-*co*-NH_2) 0.5 increases to ~9 h. After Cy3 conjugation, the half-life of poly(T-*co*-Cy3) further increases to ~15 h. Thus, the NH_2- and Cy3- internally-functionalized ssDNA are significantly more stable than unfunctionalized polyT towards degradation by human serum. This result is consistent with our stability investigations of the synthesized ssDNA in the presence of Exonuclease I or DNase I.

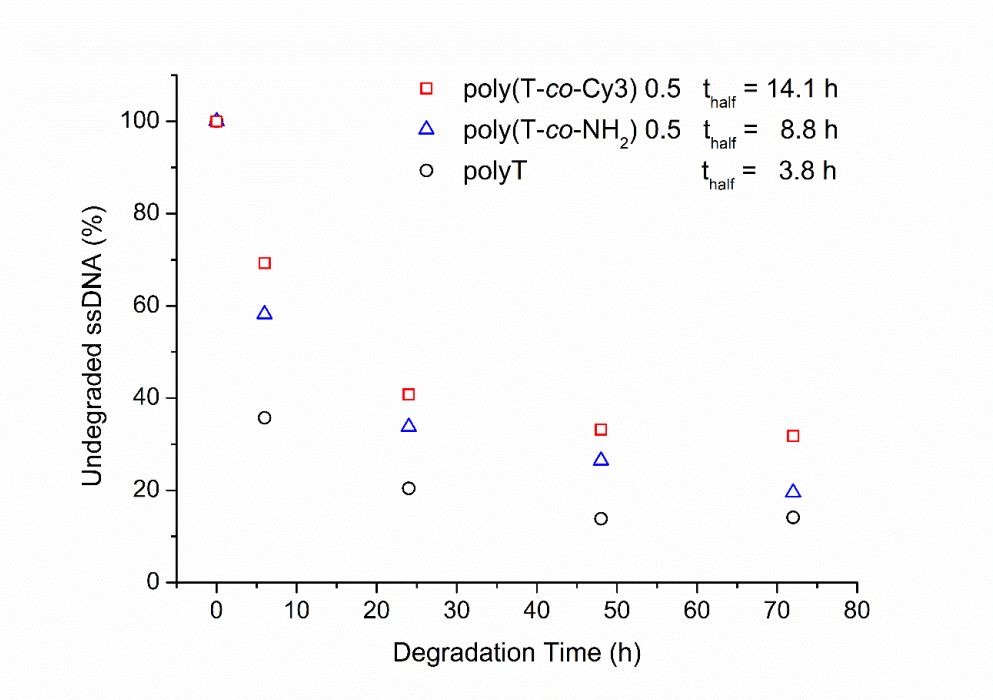


Figure 3.1.9. Degradation kinetics of polyT, poly(T-*co*-NH_2) 0.5 and poly(T-*co*-Cy3) 0.5 in human serum.

3.1.4 Conclusion

In summary, we demonstrated that different nucleotides with modified nucleobase structure can be easily and effectively incorporated into growing ssDNA using TdT enzymatic polymerization. Although the nuclease degradation of ssDNA is due to the hydrolysis of the phosphate backbone, we demonstrated that the presence of unnatural nucleobases can also impede the degradation, and does not entail modification of the sugar-phosphate backbone structure. We discovered that both the size and density of unnatural nucleobases are key factors that determine the half-life of ssDNA in nuclease degradation. MD simulations show that the incorporated unnatural nucleobases induce a decrease in local ssDNA chain flexibility and accessibility of nucleases to the ssDNA backbone with increasing size and incorporation density of unnatural nucleobases. Finally, unnatural nucleobase functionalized ssDNA also showed improved resistance to nuclease degradation in human serum, which contains both exonuclease and

endonuclease. Therefore, the enzymatic incorporation of unnatural nucleotides with modified nucleobase structures into ssDNA can be an effective and powerful strategy to tune and prolong the lifetime of DNA-based materials for *in vivo* applications.

CHAPTER 4: Self-Assembly of Amphiphiles

4.1 The Effect of Hydrophobic Tail Stiffness and Length on Polyelectrolyte Diblock Copolymer Self Assembly

* This section is a manuscript in preparation by:
Thomas J. Oweida, Tad Deaton, Lukas Harries, Ibrahim Ahmad, and Yaroslava G. Yingling.

4.1.1 Introduction

Polyelectrolyte diblock copolymers (PDCs) consist of one hydrophobic polymer block that is covalently bound to one hydrophilic polymer block carrying charged ionizable groups.[62] This ultimately combines the pH and salt responsive nature of polyelectrolytes with the self-assembly, or spontaneous aggregation, capabilities of surfactants.[231] When these molecules are exposed to an aqueous environment, the amphiphilic nature of the PDCs will allow them to self-assemble and form a variety of morphologies including, but not limited to, vesicles, cylindrical micelles, and spherical micelles. These assemblies are promising candidates as carriers for drug and gene delivery, however, the size and morphology of the assemblies play a critical role in determining their transportation dynamics and delivery capabilities.[62,232] The presence of the polyelectrolyte block in PDCs creates a complex interplay between the long-range electrostatics and the short-range molecular attractions that are governed by the molecular architecture. Thus, the self-assembly behavior of PDCs is much more complicated than their neutral counterparts.[62,232–234]

Despite the complexity, mean-field theories and scaling theories have been developed and applied to predict the properties of PDCs with some success.[233–235] One of the most notable theories for PDCs was developed by Borisov and Zhulina. This mean-field theory describes the properties of micelles in terms of the length of the core-forming block (N_B), the length of the coronal polyelectrolyte block (N_A), and several other parameters that describe the coulombic and excluded volume interactions of monomers as a function of salt concentration. Although this model is

functional, the segmentation of scaling laws into multiple regimes indicates that the theory needs to be further developed.

In silico techniques are a powerful tool used to study the driving forces and thermodynamics that control PDC self-assembly. Coarse-grained molecular dynamics (CG-MD) relies on classical mechanics to capture the attractive and repulsive forces of PDCs, where atoms are grouped to form a bead-spring chain. This simulation technique has recently been used to study the effects of monovalent and multivalent counterions on PDC self-assembly.[232] However, CG-MD simulations are too computationally expensive for broad-scope studies with multiple variables. Dissipative Particle Dynamics (DPD) is a coarse-grained simulation technique that also relies on classical mechanics but uses a soft repulsive potential which significantly reduces computational cost. The implicit solvent ionic strength method allows the DPD technique to be used for PDCs. Recently, Li et al. used the implicit solvent ionic strength DPD method to study the equilibrium structures of PDCs as a function of polyelectrolyte length and salt concentration. The observations made by Li et al. matched the scaling laws laid out by Borisov and Zhulina's theory.[218]

This study uses the implicit solvent ionic strength DPD method to perform broad scope generalized analysis of PDC self-assembly. Specifically, this study expands upon the previous work by Li et al. and captures the influence of polyelectrolyte length, solvent ionic strength, hydrophobic segment length, and hydrophobic segment rigidity on PDC self-assembly. Generalized phase diagrams are created while micelle characteristics are quantified by aggregation number and radius of gyration to capture size.

4.1.2 Methods

4.1.2.1 Dissipative Particle Dynamics (DPD)

All simulation were performed via LAMMPS using Dissipative Particle Dynamics (DPD) simulations. DPD is a coarse-grained technique where DPD beads are a representation of a molecular fragment.[236] The DPD beads move according to Newton's equations of motion and the force between DPD beads are described by a soft sphere model and interact through a purely repulsive potential defined by equation 4.1.1. F_{ij}^{C} is the conservative force, F_{ij}^{D} is the dissipative force and F_{ij}^{R} is the random force. The conservative force represents the repulsive potential exerted on particle *i* by the *j*-th particle, the dissipative force represents the drag, and in this case is linearly dependent on the momentum, and the random force is sampled from a zero-mean gaussian independent of momentum. The conservative force, dissipative force, and random force are described by equations 4.1.2, 4.1.3, and 4.1.4 respectively.[25,26]

$$f_{ij} = \sum_{j \neq i} F_{ij}^{C} + \sum_{j \neq i} F_{ij}^{D} + \sum_{j \neq i} F_{ij}^{R} \tag{4.1.1}$$

In equations 4.1.2-4.1.4, a_{ij} is the maximum repulsion between beads i and j and vanishes beyond a cutoff radius; $r_{ij} = r_i - r_j$, $v_{ij} = v_i - v_j$, and $\bar{r}_{ij}$ is the unit vector directed from j to i. The strengths of the dissipative and random force are characterized by the coefficients γ and σ, where $\gamma = \frac{\sigma^2}{2k_B T}$. The iteration time step is represented by Δt and the zero-mean symmetric random variable is represented by $\theta_{ij}(t)$.[25,26]

$$F_{ij}^{C} = \begin{cases} a_{ij}(1 - \frac{r_{ij}}{r_c})\bar{r}_{ij}, & r_{ij} < r_c \\ 0, & r_{ij} \geq r_c \end{cases} \tag{4.1.2}$$

$$F_{ij}^{D} = -\gamma \omega_D(r_{ij})(\bar{r}_{ij} . v_{ij})\bar{r}_{ij} \tag{4.1.3}$$

$$F_{ij}^{R} = \sigma\omega_R(r_{ij})\theta_{ij}\Delta t^{1/2}\bar{r}_{ij} \quad (4.1.4)$$

The a_{ij} parameters for each DPD bead are set following the parameterization established by Groot and Warren for a simulation box with a density of beads set to 3.[26] Overall, each simulation box dimension was set to be 36x36x36 resulting in a total of 139968 DPD beads per simulation. Table 4.1 provides the details on all a_{ij} parameter for the DPD beads. It should be noted that the a_{pp} parameter represents the interaction between polyelectrolyte beads. This interaction parameter is selected for implementation using the implicit solvent ionic strength DPD method which implicitly captures the ionization of the polyelectrolyte and the salt concentration or solvent ionic strength. The parameter is analogous to the second virial coefficient which includes non-electrostatic contributions and electrostatic contributions. When the salt concentration is much greater than the available counterions, the electrostatic repulsion within the polyelectrolyte segment of the PDCs is screened. Thus, the repulsive parameter represents bare, non-electrostatic contributions (a_{pp}=25). As the salt concentration is reduced, the repulsive parameter in the polyelectrolyte segment increases to account for electrostatic contributions and reduced charge screening.[31]

Table 4.1.1 Lists the repulsive parameters between the DPD beads present in the simulation. Ultimately, there are 3 bead types that make up water, the hydrophobic segment of the PDC, and the polyelectrolyte segment of the PDC.

DPD Repulsive Parameters	
Water-Water (aww)	25
PE-PE (app)	25, 30, 40, 50, 60, 90
Hydrophobic-Hydrophobic (ahh)	25
PE-Water (apw)	26
Hydrophobic-Water (ahw)	100
Hydrophobic-PE (ahp)	90

Adjacent beads inside PDC chains are covalently bonded by a harmonic potential that has a characteristic equilibrium length (r_{eq}) and energy potential (C) that represents how difficult it is for the bond to elongate or contract away from the equilibrium length. Equation 4.1.5 describes the harmonic spring.

$$F_{ij} = C(|r_i - r_j| - r_{eq}) \tag{4.1.5}$$

A harmonic potential was also applied to control the angle between 3 adjacent beads in the hydrophobic segment of the PDC chains. The harmonic potential is described by equation 4.1.6.

$$E = K(\theta - \theta_0)^2 \tag{4.1.6}$$

The prefactor, K, represents how difficult it is for the angle between 3 hydrophobic beads to deviate from the equilibrium angle θ_0. The units for K are effectively energy per radian^2.[236]

The interaction parameters, harmonic spring, and harmonic angle potential were applied to create a total of 240 DPD simulations which consisted of 300 PDC chains in water. Each simulation was performed for 4.0 x 10^6 steps with a timestep of 0.5 to attain a thermodynamic equilibrium. The first 36 simulations serve as the baseline for a comparison on how hydrophobic

segment length and stiffness influences the self-assembly of PDCs. Li et al. created a morphological phase diagram of PDCs that had a hydrophobic segment 4 beads long and a polyelectrolyte segment that was either 4, 10, 30, 50, 70, or 90 segments long. For each of the 6 lengths of PDC, the interaction parameter (a_{pp}) was ranged from 25-90 as described in table 4.1.1 creating a total of 36 simulations. These simulations had a harmonic spring constant of 2.0 for the hydrophobic segment and 25.0 for the polyelectrolyte segment and an equilibrium bond length of 0. The harmonic angle prefactor is equivalent to being 0 so the hydrophobic tail is fully flexible.[62]

Using the same 6 PDC lengths as the diagram created by Li et al., the stiffness of the hydrophobic segment was modified. The next 36 simulations arose from a modification to the harmonic spring of the hydrophobic segment. The spring constant was set to 100.0 with an equilibrium bond distance of 0.5. The polyelectrolyte harmonic potential was unchanged. The harmonic potential for the angle between 3 hydrophobic beads was set to an equilibrium angle of 180 degrees and a prefactor of 0.5 which enforced linearity of the hydrophobic segment. Another set of 36 simulations were generated for these PDCs with identical harmonic spring potentials of C=100.0 and r_{eq}=0.5. For this set of simulations, the prefactor for the harmonic angle was changed to 1.0 increasing the energy required for the hydrophobic segment to deviate from its linear 180-degree angle. It should be noted that the DPD harmonic spring coefficient is not known to affect qualitative trends in the self-assembly of PDCs.[62]

To investigate the effects of hydrophobic segment length on the self-assembly of PDCs, the next set of 36 simulations had a hydrophobic segment length of 8 instead of 4. The shortest polyelectrolyte length coincided this variation and was set to 8 instead of 4 to maintain a symmetric PDC starting point. As with the work previously performed by Li et al., the hydrophobic segment was fully flexible with an effective harmonic angle prefactor of 0. The harmonic spring constant

and equilibrium bond distance of the hydrophobic tail were set to 100.0 and 0.5 to stay consistent with the DPD simulations performed in this work. Another set of 36 simulations were performed on the PDCs with a hydrophobic segment length of 8. However, this set of 36 simulations varied only in the harmonic angle. The reinforce linearity in the hydrophobic segment, the harmonic angle prefactor was set to 1.0 and the equilibrium angle was set to 180 degrees.

Following the same process as before, the hydrophobic segment was lengthened to 24 beads per PDC. The shortest polyelectrolyte was also set to 24 making a total of 5 different PDCs with a hydrophobic segment of 24 beads. The harmonic spring was maintained to have parameters of C=100.0 and r_{eq}=0.5 as with the other simulations in this work. The hydrophobic segment is also set to be fully flexible with an effective prefactor of 0. The variation of the polyelectrolyte interaction parameter from 25-90 as seen in table 4.1.1 creates a total of 30 simulations for this set of PDCs. The last 30 simulations are identical to the PDCs with a hydrophobic segment of 24 just described, however, the harmonic angle prefactor is set to 1.0 and the equilibrium angle is set to 0. Ultimately, the 250 DPD simulations allow for the study of how hydrophobic segment length and stiffness affect the self-assembly of PDCs with varying polyelectrolyte lengths and solvent ionic strength conditions.

4.1.2.1 Analysis

The last 1.0 x 10^6 timesteps were analyzed for each simulation, which corresponds to the dynamics of the system at thermodynamic equilibrium. An in-house Perl script was used to calculate the aggregation number (P), radius of gyration of the micelle ($R_{g,m}$), radius of gyration of the micelle core ($R_{g,c}$), and relative shape anisotropy of the micelle (κ). The aggregation number describes the number of PDC chains that form a micelle. A gyration tensor was used to calculate the radius of gyration of the core and micelle along with the shape anisotropy of the micelle. The

gyration tensor is described in equation S4.5.1. The principal moments of the micelle are determined from the gyration tensor and is described in equation S4.1.2. From the principal moments, the radius of gyration can be calculated for the micelle or core based on equation S4.1.3 and the shape anisotropy can be calculated using equation S4.1.4. Ultimately, the radius of gyration represents the size of the micelle by accounting for the distance of every bead from the center of mass of the micelle. Lastly, the shape anisotropy produces a normalized value that can be serve as a classifier for micelle morphology. If the shape anisotropy is 0, the DPD beads in a micelle form a perfect sphere and if the shape anisotropy is equal to 1, the DPD beads in micelle form a perfect line. The cylindrical morphology was classified for shape anisotropy values greater than 0.15 to remain consistent with previous work.[62] A vesicle or lamellar classification is indicated by the aggregation of 95% or more of PDC chains into a single micelle and a shape anisotropy less than 0.15. The median values from each simulation were used for the final classification. Any micelles with aggregation numbers less than 10 were dropped from the dataset.

4.1.3 Results and Discussion

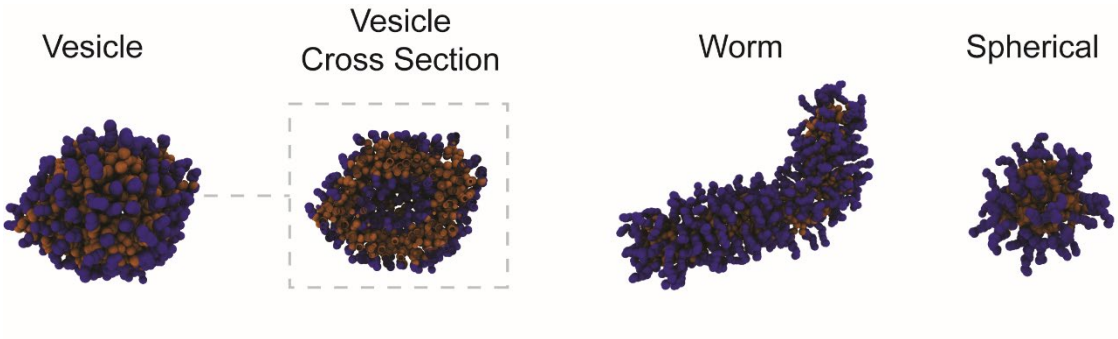


Figure 4.1.1. Provides a visual representation of the general morphology classifications observed in the DPD simulations.

The DPD simulations yielded numerous thermodynamically stable morphologies based on the solvent ionic strength of solution, length of the hydrophilic polyelectrolyte segment, length of the hydrophobic segment, and rigidity of the hydrophobic segment. Figure 4.1.1 illustrates the general

morphological classifications observed which range from vesicle to worm-like/cylindrical to spherical micelles. Inside of each classification there are nuances to the morphological structure. For instance, the vesicles can become more disc-like or planar, the worms can vary in aspect ratio, and the spherical micelles can vary in size of the corona creating a crew-cut style micelle for corona sizes smaller than the core and star-like micelles for corona sizes much larger than the core. Despite these variations, the generalized morphologies are depicted in the phase-diagram because the DPD method serves as a generalized simulation technique for PDCs. In this study, the DPD parameters are not mapped back to experiments of a particular PDC, although it has been done in the past for some simulation parameters tested.

The generalized phase diagrams provided in Figure 4.1.2 show a traditional transition from vesicles to cylinders to spherical micelles as the salt concentration and polyelectrolyte length increases with each respective hydrophobic segment. Although not depicted, the prominence of vesicles that form more oblong or disc-like morphologies increases as the proximity to the worm-like phase boundary increases. This indicates that PDC asymmetry and ionic strength affect the formation of vesicles. The formation of disc-like micelles near the phase boundary of worm-like micelles has been previously observed and has been shown to be dependent on polymer concentration and molecular architecture.[237]

Figure 4.1.2 (a) shows the phase diagrams for the PDCs with a hydrophobic segment length of 4 beads. The induced stiffness of the hydrophobic segment increases from left to right which corresponds with a reduction in the formation of worm-like micelles. The induced stiffness in the hydrophobic tail appears to particularly influence the formation of vesicles over cylinders. This effectively means that worm-like morphologies can only be maintained with an increase in PDC asymmetry or decrease in salt concentration as the hydrophobic segment becomes more rigid. To

a lesser extent, the cylindrical-spherical micelle boundary is altered with increasing rigidity. Namely, only one transition from cylinder to spherical micelle was observed. This simulation had a polyelectrolyte length of 90 beads and a salt concentration that effectively removed any columbic contributions from the polyelectrolyte.

Figure 4.1.2 (b) shows the generalized phase diagrams for PDCs with a hydrophobic segment of 8 beads. Comparing the phase diagram of the flexible hydrophobic segment in Figure 4.1.2 (b) to the phase diagram of the flexible hydrophobic segment in Figure 4.1.2 (a), it is clear the length of the hydrophobic segment influences both the vesicle-cylinder phase boundary and cylinder-sphere phase boundary. As the length of the hydrophobic segment increases, vesicles and spheres become more prominent at high salt concentrations and the cylindrical region narrows at both boundaries. However, there are no morphological phase changes observed for data points at low salt concentration. Low salt concentration morphological phase transitions are observed only when rigidity is introduced to the hydrophobic segment with 8 beads. As observed previously, a worm-like to spherical transition occurs with a slight shift in the phase boundary. In this instance, coulombic contributions are present in simulations undergoing phase transitions, which indicates that phase transitions due to rigidity in the hydrophobic segment are not specific to high-salt conditions. It should be noted that for the shorter hydrophobic segment, rigidity also influenced the vesicle-cylindrical micelle phase boundary, which is not observed at longer hydrophobic segment lengths. This indicates that the length of the hydrophobic segment may be the dominant factor influencing the vesicle-cylinder phase transition while rigidity plays a secondary role with diminishing influence as the hydrophobic segment length increases.

Figure 4.1.2 (c) shows the phase diagram for PDCs with a hydrophobic segment length of 24 beads. As previously observed, the increased length of the hydrophobic segment shifts the vesicle-

cylinder phase boundary. The transition is again more localized at high salt concentrations indicating that the unscreened charge on polyelectrolytes makes vesicle formation unlikely, especially for longer polyelectrolytes where charge density effectively increases. There is no change in the phase diagram when rigidity is introduced into the hydrophobic segment with 24 beads. This supports the previous theory that hydrophobic segment length is the dominant factor influencing that phase boundary of the diagram, while rigidity plays a lesser role in comparison and particularly for longer hydrophobic segments. With the given datapoints, there is no cylindrical morphology present when the hydrophobic bead length is 24, however, it is possible that a cylindrical regime is present at increments not tested. In this instance, it could be expected that rigidity may influence the cylindrical-spherical boundary as previously identified at shorter hydrophobic segment lengths. However, it is possible the formation of cylinders is entirely removed given the breadth of the region at hydrophobic segment lengths of 8 beads. If this is the case, rigidity appears to be non-influential in determining morphology once a specific hydrophobic segment length is reached.

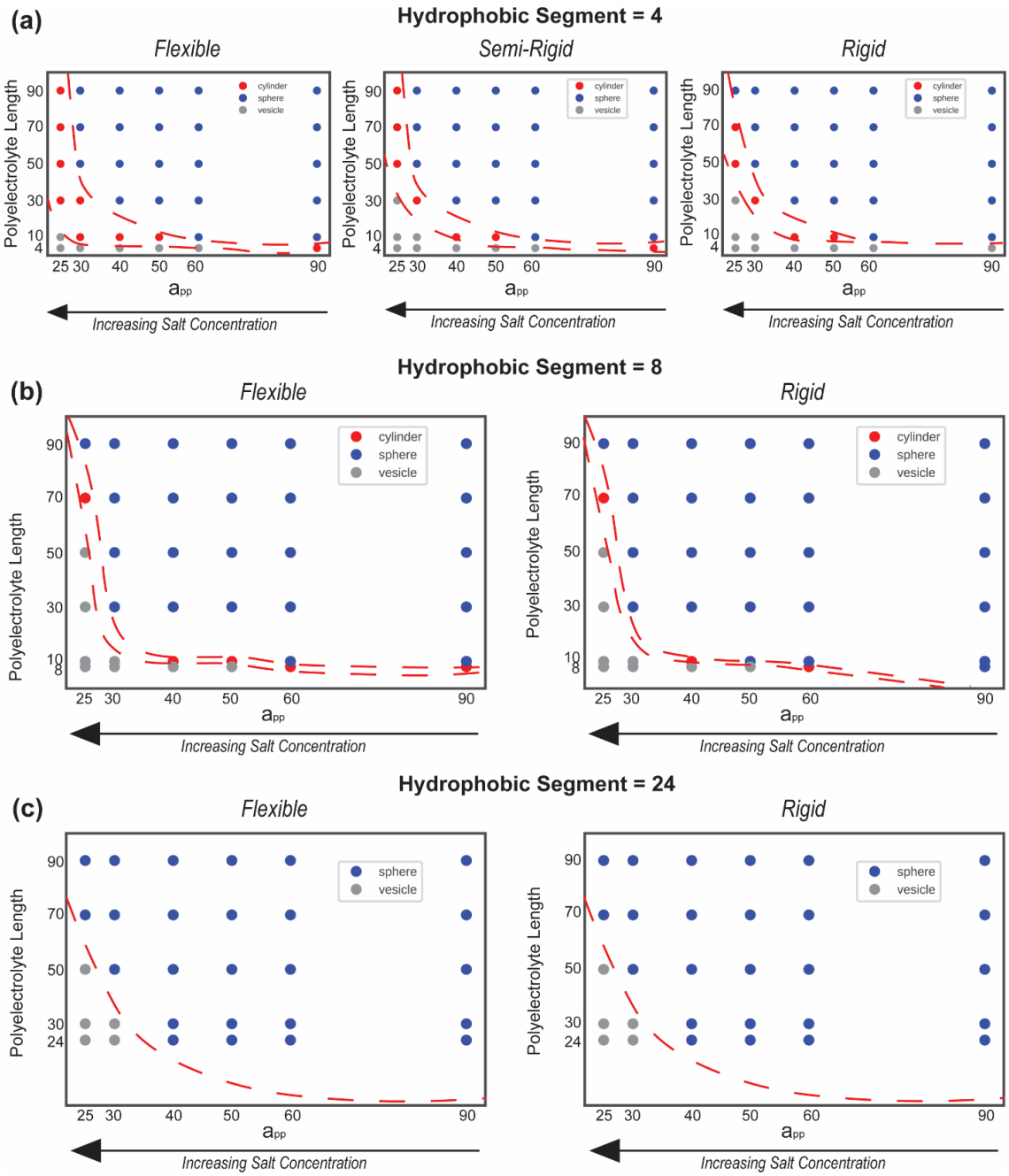


Figure 4.1.2. (a) generalized phase diagrams with a hydrophobic segment length of 4 beads and increasing rigidity in the hydrophobic segment from left to right (b) generalized phase diagrams with a hydrophobic segment length of 8 beads and increasing rigidity in the hydrophobic segment from left to right and (c) generalized phase diagrams with a hydrophobic segment length of 24 beads and increasing rigidity in the hydrophobic segment from left to right. The red dashed line indicates phase boundaries.

Overall, an increase in hydrophobic length and rigidity makes the formation of cylindrical micelles increasingly difficult as the cylindrical micelle region narrows at both the vesicle and

spherical phase boundaries. Increasing the length of the hydrophobic segment appears to be the dominant factor in determining micelle morphology while rigidity plays a secondary role. At short hydrophobic segment lengths, rigidity influenced both phase boundaries narrowing the cylindrical regime. However, as hydrophobic segment length increased, rigidity induced a cylindrical to spherical transition only. Ultimately, the influence of rigidity appears to diminish, and no discernable difference is detected in the given data points with a hydrophobic segment length of 24 beads.

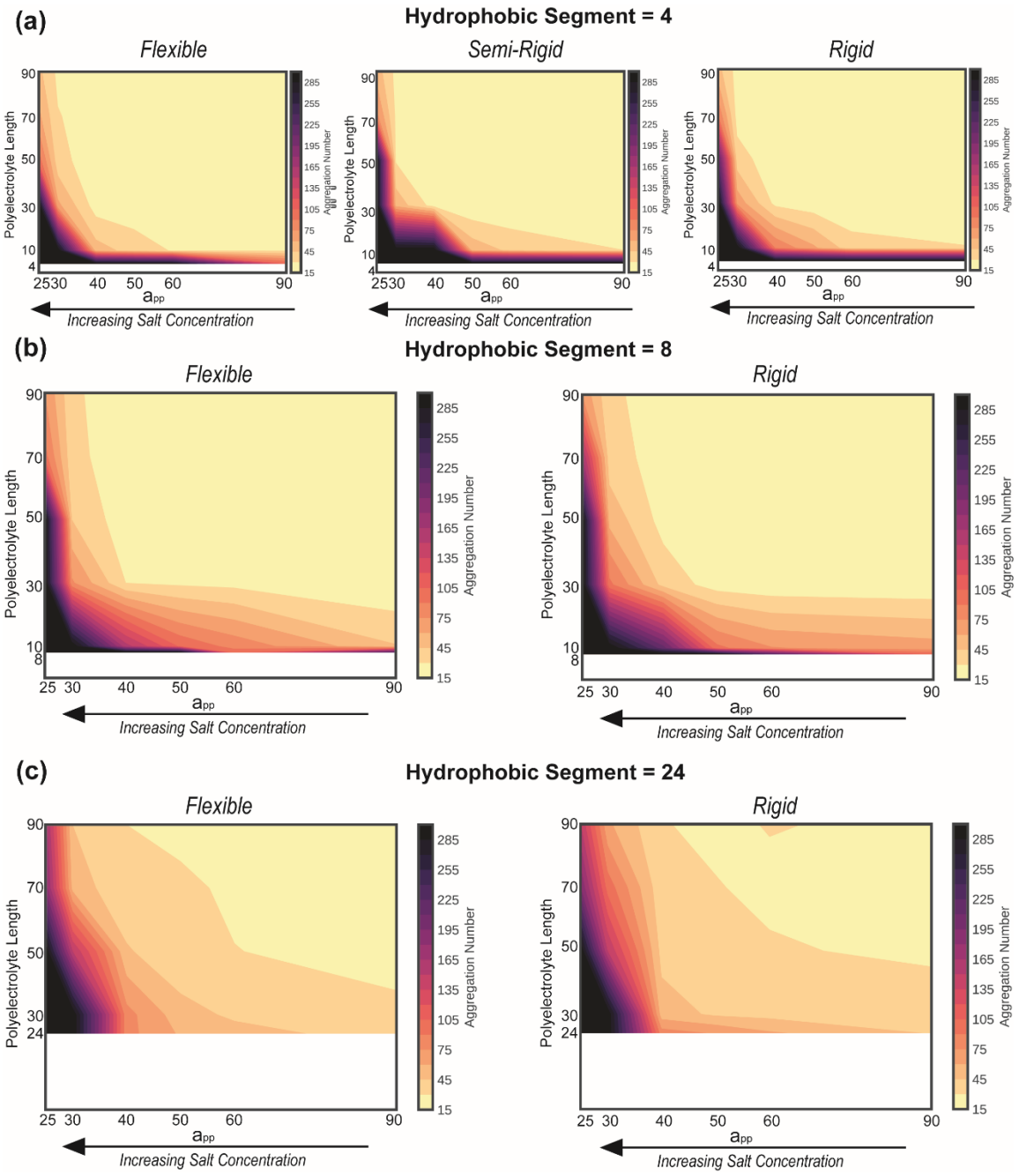


Figure 4.1.3. Represents the median aggregation number of micelles as contour plots for (a) PDCs with a hydrophobic segment length of 4 beads and increasing rigidity in the hydrophobic segment from left to right (b) PDCs with a hydrophobic segment length of 8 beads and increasing rigidity in the hydrophobic segment from left to right and (c) PDCs with a hydrophobic segment length of 24 beads and increasing rigidity in the hydrophobic segment from left to right.

Figure 4.1.3 shows the median aggregation number of micelles in each respective DPD simulation. As before, the figure breaks down simulations by hydrophobic segment length and hydrophobic tail rigidity while salt concentration and polyelectrolyte length are plotted against one another as contour plots. In general, the aggregation number correlates well with morphological classifications identified in Figure 4.1.2. For larger aggregation numbers, a vesicle can be expected while the morphology transition to cylinders and then spheres as the aggregation number decreases. The aggregation number of micelles is mainly driven by solvent ionic strength and polyelectrolyte length. Large aggregation numbers, and thus vesicles, correspond to conditions that limit the coulombic effects of the polyelectrolyte. Namely, through reducing polyelectrolyte length or increasing solvent ionic strength which both decrease charge-charge repulsion by either reducing charge density or increasing charge screening affects.

Figure 4.1.3 (a) highlights the effects of hydrophobic segment rigidity for PDCs with a short hydrophobic segment. As the rigidity of the hydrophobic segment increases, the aggregation number tapers off at a slower rate. Ultimately, this means that a greater shift in polyelectrolyte length or salt concentration is required to reduce aggregation number. The increased aggregation numbers correspond well with Figure 4.1.2 that shows increased prominence of vesicles as hydrophobic segment rigidity increases.

As hydrophobic segment length increases, the aggregation number also shifts to larger values. As before, the larger aggregation numbers correspond to increased prominence of vesicles in the morphological phase diagram. In addition, there appears to be a reduced effect of the tail rigidity

on the aggregation numbers at longer hydrophobic segment lengths. This supports the theory that

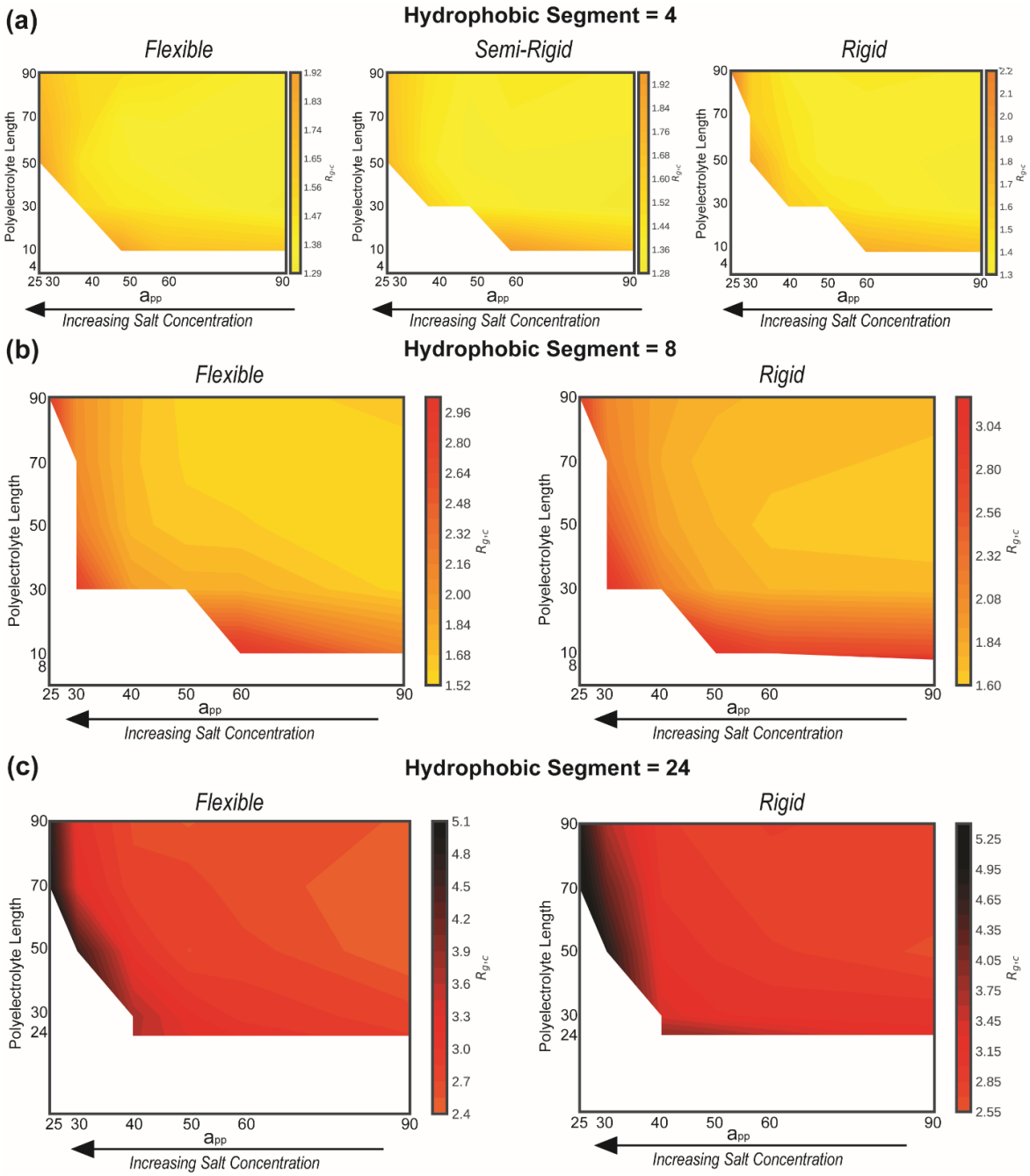


Figure 4.1.4. Represents the median radius of gyration of micelle cores (Rg,c) as contour plots for (a) PDCs with a hydrophobic segment length of 4 beads and increasing rigidity in the hydrophobic segment from left to right (b) PDCs with a hydrophobic segment length of 8 beads and increasing rigidity in the hydrophobic segment from left to right and (c) PDCs with a hydrophobic segment length of 24 beads and increasing rigidity in the hydrophobic segment from left to right.

hydrophobic segment rigidity becomes less influential as hydrophobic segment length increases; however, it does appear that rigidity does still increase aggregation numbers to some extent.

Figure 4.1.4 highlights the median radius of gyration of micelle cores ($R_{g,c}$) for all simulations performed in this study. For the most part, Figure 4.1.4 follows the same trend as described with aggregation number where the size of the micelle core decreases with the the number of chains making up a micelle. Logically it makes sense that as fewer chains pack together, the smaller the micelle core will become. There is, however, a notable deviation from this trend in the low salt, long polyelectrolyte regime.

Figure 4.1.3 shows that the aggregation number plateaus or continues to decrease as polyelectrolyte length increases and salt concentration decreases. For the hydrophobic segment length of 4 beads in Figure 4.1.4 (a), the same trend is observed. Figure 4.1.4 (b) and Figure 4.1.4 (c) deviate from this trend in the low salt, long polyelectrolyte simulations. In this case, as aggregation number decreases, the size of the micelle core increases. This indicates that for long hydrophobic tails, the packing becomes less efficient, particularly when charge density is high in low salt conditions. As rigidity of the hydrophobic segment increases, the packing inefficiency becomes exacerbated, and the micelle core becomes even larger. Any effects from hydrophobic segment rigidity appear to be negligible for the short hydrophobic tail depicted in Figure 4.1.4 (a).

While the $R_{g,c}$ provides insight on the packing of the hydrophobic segment of the micelle, the radius of gyration of the whole micelle ($R_{g,m}$) is physically relevant. Figure 4.1.5 shows the $R_{g,m}$ for all simulations performed in this study. In general, the $R_{g,m}$ increases as polyelectrolyte length increase and salt concentration decreases. Thus, despite aggregation numbers decreasing, the size of micelles are increasing. The dominating factors for determining $R_{g,m}$ is polyelectrolyte and hydrophobic segment length in PDCs. There is a significant increase in micelle size along the

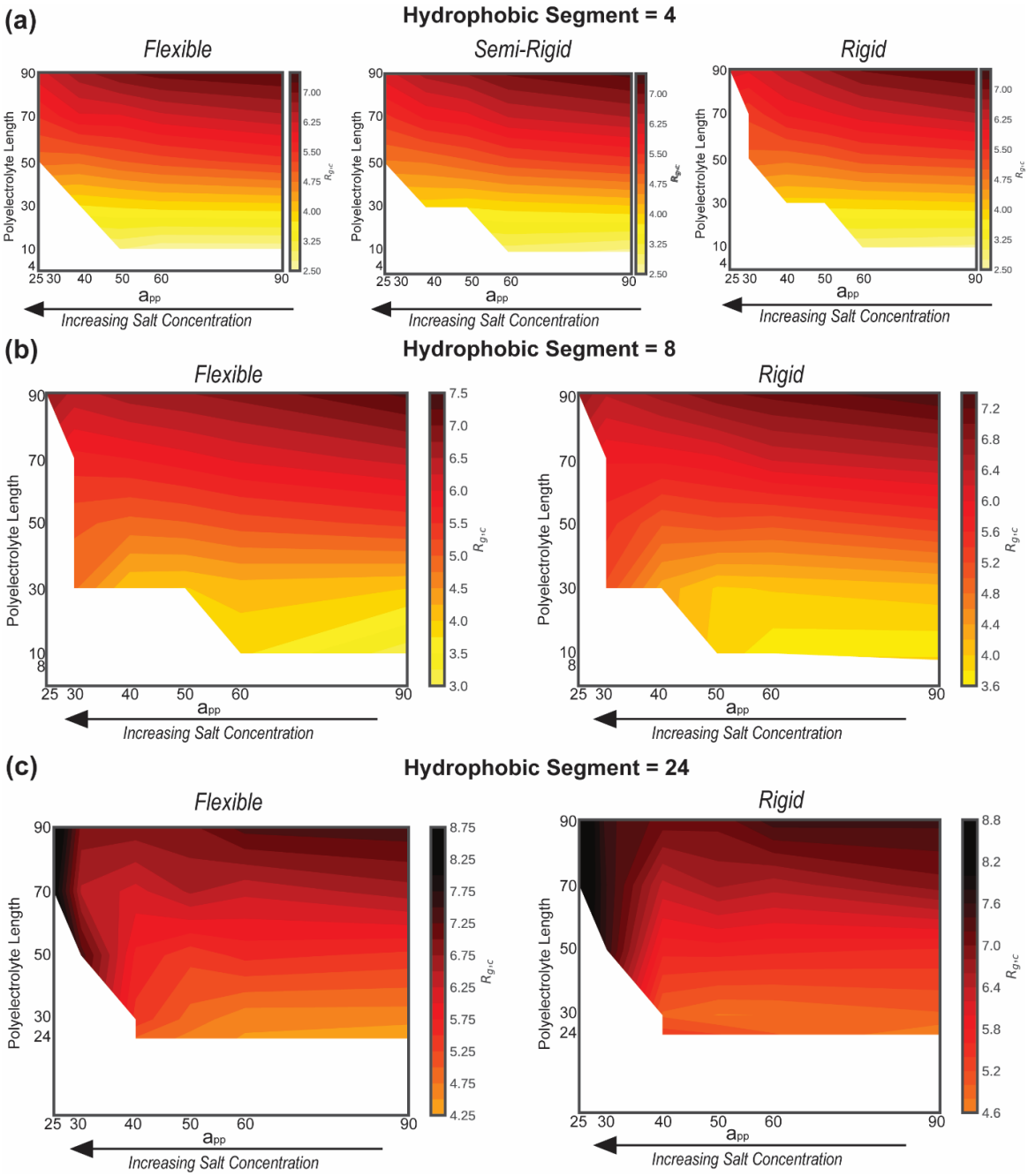


Figure 4.1.5. Represents the median radius of gyration of micelles ($R_{g,m}$) as contour plots for (a) PDCs with a hydrophobic segment length of 4 beads and increasing rigidity in the hydrophobic segment from left to right (b) PDCs with a hydrophobic segment length of 8 beads and increasing rigidity in the hydrophobic segment from left to right and (c) PDCs with a hydrophobic segment length of 24 beads and increasing rigidity in the hydrophobic segment from left to right.

polyelectrolyte length axis and a significant increase in micelle size as hydrophobic length increases.

The contour plots in Figure 4.1.4 and Figure 4.1.5 indicate that salt concentration and hydrophobic segment rigidity are secondary factors in determining the overall size of micelles. Referring to Figure 4.1.4, the size of the micelle core increases as rigidity in induced in long hydrophobic segments. Thus, in those regions, stiffness would contribute to overall micelle size as well. Figure 4.1.5 shows that micelle size increases as salt concentration decreases. This likely arises from polyelectrolyte beads acting as more of a directional random walk and less like a random coil effectively increasing the average size of the micelle corona.

4.1.4 Conclusion

This study investigates how self-assembly of polyelectrolyte diblock copolymers (PDCs) are affected by molecular architecture and solvent ionic strength. Overall, a total of 250 simulations were performed to create generalized phase diagrams of PDCs as a function of polyelectrolyte length, hydrophobic segment length, hydrophobic segment rigidity, and solvent ionic strength. As previously reported, it was observed that vesicles form for PDCs with short polyelectrolyte lengths and high salt concentrations. As polyelectrolyte length increases or salt concentration decreases, the vesicles transition into cylindrical or worm-like micelles before transitioning again into spherical micelles. It is observed that increasing the length of the hydrophobic segment significantly hinders the formation of cylindrical micelles and the formation of vesicles and spheres becomes more prominent at each respective phase boundary. Hydrophobic segment rigidity was observed to have a significant impact on morphology at short polyelectrolyte lengths, but the effects appeared to become secondary to hydrophobic segment length. Overall, increasing rigidity hindered cylindrical micelle formation.

The median aggregation number of micelles was observed to correspond with micelle morphology, where high aggregation numbers indicated vesicle formation and reducing

aggregation numbers led to cylindrical and ultimately spherical micelles. The aggregation number has a direct impact on micelle core size. However, at longer hydrophobic segments, the packing of chains in the core appeared to become less efficient as micelle core size began to increase despite a reduction in aggregation number. The introduction of rigidity in long hydrophobic segments exacerbated the inefficient chain packing, further increasing the micelle core size. The overall size of micelles was observed to be mostly dependent on the length of the polyelectrolyte and hydrophobic segment. However, salt concentration and rigidity play a secondary role.

4.1.5 Supplemental Information

Equations S4.1.1 show the gyration tensor which is calculated from the XYZ coordinates of particles in each DPD simulations. The gyration tensor describes the second moments of positions for particles using a coordinate system. From the gyration tensor, the diagonalization can be formed into a symmetric 3x3 matrix with principal moments that serve as eigenvectors of the mass as shown in Equation S4.1.2.

$$S = \frac{1}{N}\begin{bmatrix} \sum_i (x_i - x_{cm})^2 & \sum_i (x_i - x_{cm}) - (y_i - y_{cm}) & \sum_i (x_i - x_{cm}) - (z_i - z_{cm}) \\ \sum_i (x_i - x_{cm}) - (y_i - y_{cm}) & \sum_i (y_i - y_{cm})^2 & \sum_i (y_i - y_{cm}) - (z_i - z_{cm}) \\ \sum_i (x_i - x_{cm}) - (z_i - z_{cm}) & \sum_i (y_i - y_{cm}) - (z_i - z_{cm}) & \sum_i (z_i - z_{cm})^2 \end{bmatrix} \quad \text{(S4.1.1)}$$

$$S = \begin{bmatrix} \lambda_x^2 & 0 & 0 \\ 0 & \lambda_y^2 & 0 \\ 0 & 0 & \lambda_z^2 \end{bmatrix} \quad \text{(S4.1.2)}$$

During simulations and analysis, the gyration tensor was used to quantify various metrics of the size and shape of micelles. Equations S4.1.3 and S4.1.4 establish two metrics that can be calculated through the principal moments. Figure S4.1.3 calculates radius of gyration (R_g) which measures the size of a particle. Specifically, the larger the value, the larger the micelle as particles

are farther from the center of mass. Figure S4.1.4 describes relative shape anisotropy ($\boldsymbol{\kappa}^2$) which has possible values falling between zero and one.

$$R_g^2 = \lambda_x^2 + \lambda_y^2 + \lambda_z^2 \tag{S4.1.3}$$

$$k^2 = \frac{3}{2}\frac{\lambda_x^4 + \lambda_y^4 + \lambda_z^4}{(\lambda_x^2 + \lambda_y^2 + \lambda_z^2)^2} - \frac{1}{2} \tag{S4.1.4}$$

In our DPD simulations, a relative shape anisotropy equal to 0 indicates a perfectly spherical particle, while a relative shape anisotropy equal to 1 indicates particles in a straight line. In this study, the relative shape anisotropy was used to label micelle shape as worm-like or spherical.

CHAPTER 5: Materials Informatics Approaches

5.1 Connecting Experiments and Simulation: A Genetic Algorithm for Flexible Polyelectrolyte Analysis

* This section is a manuscript in preparation by:
Thomas J. Oweida and Yaroslava G. Yingling.

5.1.1 Introduction

Single-stranded DNA (ssDNA) plays a significant role in biological processes and has drawn significant interest for nanotechnology applications due to its programmable self-assembly and tunable properties.[5] Specifically, ssDNA's structure in solution can control properties such as its degradation rate, molecular binding, and hybridization mechanisms.[104,238] However, research on ssDNA is currently bottlenecked by the lack of understanding surrounding its structure and dynamics due to its inherent flexibility. For example, the persistence length of ssDNA, a metric for polymer chain stiffness, has been empirically calculated to be as low as ~9 Å and as high as ~25 Å in 100 mM solution based on experimental measurements.[7,8,84] Revealing the fundamental structure of ssDNA in solution will provide insight on the physical constraints of these molecules and provide a framework on how to control and manipulate the self-assembly and structure of ssDNA and ssDNA-based materials.

To advance the understanding of ssDNA structure in solution, it is critical to extract information from both experimental and computational methodologies and unite them to provide a cohesive prediction. Small-angle X-ray scattering (SAXS) and Forster resonance energy transfer (FRET) are two experimental techniques frequently used to study ssDNA in solution. SAXS measures the contrast in electron density between a molecule of interest and surrounding solvent, ultimately providing a scattering intensity that represent the time and spatial average of millions of molecules. This signal provides information on the average size, shape, and flexibility of ssDNA

but does not provide information on singular structures.[85,94,96] FRET measures the end-to-end distance (R_{ee}) of ssDNA by tracking the distance between fluorescent molecules bound to the 3' and 5' ends of ssDNA. However, these fluorescent molecules may interact with the ssDNA molecule affecting the natural dynamics to some degree.[7,91–93] On the computational side, molecular dynamics (MD) simulations can be used to predict the atomistic structure of complex molecules, providing a higher resolution structure as compared to experimental measurements.[104,239,240] However, the quality of such predictions depend on the quality of the used force field. In a recent study, the accuracy and applicability of force fields for DNA were comprehensively assessed for use on ssDNA. A validation process was laid out for comparing the MD simulations of ssDNA to experimental measurements including SAXS and FRET, with a few force fields recommended for use on ssDNA.[239]

Ultimately, the comparisons of simulation and experiment for ssDNA still suffer from the difference in size and timescale. For example, SAXS experiments measure a time-averaged ensemble of ssDNA structures, while simulations might attempt to map a single representative structure to fit the experimentally captured ensemble.[94] This traditional methodology can be misleading on the physical structures that are present in solution, in particular for flexible polymers such as ssDNA.[116] Thus, there is a need to further analyze MD simulations of ssDNA to better compare computational results to the time-averaged ensembles captured in experimental data.

A genetic algorithm (GA) is an evolutionary algorithm that can find the optimal subset of provided simulation structures that most closely represents the experiment ensemble. Ultimately, this allows for a group of structures with atomistic resolution to be connected to the lower resolution experiment. Specifically, this deconvolution of the experimental SAXS curve is done by calculating a SAXS scattering intensity for individual structures in simulation followed by

searching for which group of calculated SAXS intensities can be combined to minimize the error when compared to experiment.[13,116,130] The implementation of a GA is required for this process as a brute force method requires a test of 12,497,500 possible combinations with just 5,000 simulated structures and an ensemble containing a set of 2 structures. Considering the size of the ensemble is related to the molecules flexibility, it can be expected that the ensemble for ssDNA is larger than 2, increasing the number of possible combinations that would have to be tested without inducing a selection bias with a GA.[116]

Overall, this study relies on the full representation of Newtonian physics in each simulation, without a dependence on high temperature MD simulations or a reduction in non-bonded interactions to explore conformational space. This preserves ssDNA chain dynamics and produces reliable and physically relevant structures in each simulation. The simulations of ssDNA are then analyzed with a GA to (1) further validate force field selection for ssDNA in various types of solvent, (2) provide an accurate ensemble of atomistic ssDNA structures in solution, and (3) highlight a promising parameterization pathway for the development of more accurate force fields for ssDNA. The implementation of machine learning to connect computation and experiment has allowed for the best possible predictions of ssDNA structure given the available data.

5.1.2 Methods

5.1.2.1 Simulations

This study utilizes a total of 7 simulations performed on a 30-mer of poly(thymine) (polyT). Nucleic Acid Builder (NAB) software was used to generate a B-DNA adenine-thymine complex.[117] The complementary adenine strand was subsequently deleted using Discover Studio, resulting in the initial polyT conformation used in this study.[118] All initial simulation setups were prepared using xLeAP in the AMBER18 software package.[19] The AMBER18 software package

contains parameters for all solvation environments and different DNA force fields used. Specifically, the parm99 (ff99)[119], parm99 with bsc0 corrections (bsc0)[120], parm99bsc0 with bsc1 corrections (bsc1)[121], and parm99bsc0 with OL15 corrections (OL15) force fields were used.[113,114] In addition, these force fields were used in either implicit or explicit solvent. For implicit solvent, either the Hawkins, Cramer, Truhlar Generalized Born (GB) implicit water model (IGB1)[21,122,123], modified GB model (IGB5), or the GB-neck2 model (IGB8) were used.[127,128] All explicit solvent simulations were described by the TIP3P explicit water model[124] with the NaCl ions described by the Joung and Cheatham parameters.[125,126] The salt concentration for the implicit and explicit water models was set to 100 mM to match experimental measurements of 30-mer polyT.

Overall, the force field combinations were chosen based on a recent assessment study that compared simulations of ssDNA to experiment. Table 5.1.1 shows all combinations of solvent and DNA force fields recommended for use based on that study.[239] In implicit IGB1 solvent, the ff99 and bsc1 force field were both shown to perform the best, thus both force field combinations were chosen for further examination. In the IGB5 and IGB8 implicit solvents, ff99 was shown to perform extraordinarily well for polyT as compared to experiment and are also chosen for further examination. Lastly, in explicit TIP3P solvent, bsc0, bsc1, and OL15 were deemed the top 3 performing force fields for polyT. Although the performance of polyT in explicit solvent is not objectively good, it is the current benchmark for simulations.

Table 5.1.1. Shows the combination of solvent and DNA force fields to be examined.

SIMULATIONS	
SOLVENT	DNA FORCE FIELD
IGB1	ff99 bsc1
IGB5	ff99
IGB8	ff99
TIP3P	bsc0 bsc1 OL15

For implicit solvent simulations, a 999 Å cutoff was used for nonbonded interactions and the default dielectric constants of 78.5 and 1.0 were used for the solvent and solute, respectively. The energy minimization of polyT was performed using the steepest descent method for 1000 steps followed by a 9000-step minimization with the conjugate gradient method. During minimization, a harmonic restraint of 5 kcal mol^{-1} $Å^{-2}$ was applied to all non-hydrogen atoms. Next, Langevin dynamics was used to heat the system to 300 K with a collision frequency of 5 ps^{-1} and 1 fs timestep with bonded hydrogens restrained with the SHAKE algorithm. An equilibration was carried out in three stages for 50 ps each. Each stage used a harmonic restraint of 5, 1, and 0.1 kcal mol^{-1} $Å^{-2}$ to restrain polyT, respectively. The production run of all implicit solvent simulations was performed in an NPT ensemble at 300 K for 500 ns with a Langevin thermostat. The thermostat used a collision frequency of 5 ps^{-1} and a 1fs timestep.

The simulations of ssDNA in explicit solvent used Langevin dynamics with a 1 ps^{-1} collision frequency, a 10 Å cutoff for non-bonded interactions, and Particle Mesh Ewald (PME) method for long-range electrostatics.[101] The charge of polyT was first neutralized with 29 sodium counterions whose placement was calculated using a Coulombic potential on a grid. A truncated octahedral TIP3P water box with a 15 Å buffer distance between the polyT chain and periodic boundary was generated to avoid self-interaction of polyT across periodic boundary conditions. Lastly, ~100mM of NaCl ions were added randomly to achieve a final salt concentration. Energy minimization was carried out in four stages, with each stage using the steepest descent method for 5000 steps followed by the conjugate gradient method for another 5000 steps. A harmonic restraint of 10, 10, 5, and 0 kcal mol^{-1} $Å^{-2}$ was used on all polyT atoms in each minimization stage, respectively. After the first minimization stage, the harmonic restraints were applied only to non-hydrogen atoms of polyT. Following minimization, the system was heated to 300 K over 100 ps with a 0.5 fs timestep

and a harmonic restraint of 10 kcal mol^{-1} $Å^{-2}$ on polyT. The water, ions, and polyT were equilibrated for a total of 3 ns. The 3 ns were split into a 1 ns step where a harmonic restraint of 5, 1, and 0.5 kcal mol^{-1} $Å^{-2}$ were used to restrain polyT. The production run of all explicit solvent simulations was performed at 300 K for 500 ns. A 2 fs timestep and SHAKE algorithm were used in the NPT ensemble while temperature was regulated using a Langevin thermostat and pressure was regulated using a Berendsen barostat.[239]

5.1.2.2 Small Angle X-ray Scattering (SAXS) Calculations

FoXS software[129,130] was used to calculate small angle X-ray scattering (SAXS) intensities for polyT in each simulation using equation 5.1.1 where I(q) is the scattering intensity, $q=(4\pi/\lambda)\sin(\theta)$, f(q) are the atomic form factors that take into account excluded volume and solvent, and d is the distance between atoms.

$$I_m(q) = \sum_{j=1}^{N_A}\sum_{i=1}^{N_A} f_i(q)f_j(q)\frac{\sin(qd_{ij})}{qd_{ij}} \tag{5.1.1}$$

Specifically, intensities were calculated from PDB files which were generated from simulations every 100 ps over the entire 500 ns trajectory resulting in 5000 structures for each simulation. The calculated intensities are compared to experimental intensity, SASDBD39, which is an entry on the Small Angle Scattering Biological Data Bank (SASBDB).[131] This experimental intensity is for polyT in 100 mM NaCl, which corresponds directly to the simulations. The goodness of fit between the calculated and experimental profiles is quantified through a χ^2 metric as shown in equation 5.1.2.

$$x^2 = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{I_{exp}(q_i) - cI(q_i, c_1, c_2)}{\sigma(q_i)}\right)^2 \tag{5.1.2}$$

In equation 5.1.2, M is the number of datapoints in the experimental SAXS profile, $I_{exp}(q)$ is the experimental intensity, σ(q) is the experimental error, and I(q) is the calculated intensity with 3

fitting parameters performed by FoXS software. The c parameter is a scaling factor, the c_1 parameter scales the excluded volume of atoms, and the c_2 parameter scales for the density of a water in the first hydration layer. In each SAXS calculation, the c_1 and c_2 parameters were set to a constant value of 1.05 and 4, which indicates FoXS software is adding electron density. Although these values are the maximum allowed by FoXS software, it provided the best fit between computation and experiment. The high charge density of ssDNA could explain why the fitting parameters adjust for a larger electron density in the SAXS calculations.

5.1.2.3 Genetic Algorithm (GA)

A genetic algorithm (GA) was used to select the appropriate ensemble of ssDNA structures to better compare polyT simulations to SAXS profiles. Specifically, a GA is an evolutionary algorithm that was used to find a subset of the provided simulation structures that most closely represents the experiment ensemble captured via SAXS. Figure 5.1.1 illustrates the GA scheme used in this study. The initial population is the total set of 5000 ssDNA structures provided from each simulation where each individual ssDNA structure is termed a gene. These genes are grouped into subsets called chromosomes as seen encompassed by a red rectangle in Figure 5.1.1. Each chromosome contains 7 genes, which was selected based on its performance and discussed in Figure S5.1.1. The population in each generation contains 400 chromosomes for a total of 2800 ssDNA structures.

The population in each generation is selected via a tournament selection process. In this method, 2 chromosomes are randomly selected to compete against one another. The winner is selected to move to the next generation, while the loser is discarded. The tournament selection process is performed until there are enough winners to create the next generation's population of

400 chromosomes. Winners and loser are decided by using a fitness function as described by Equation 5.1.3, where the lowest χ^2 value indicates a better fit to experiment.

$$x^2 = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{I_{exp}(q_i) - c\sum_n \omega_n I_n\,(q_i, c_1, c_2)}{\sigma(q_i)}\right)^2 \qquad (5.1.3)$$

Equation 5.1.3 is a modified form of Equation 5.1.2, as it considers numerous structures contributing to a single SAXS intensity. The calculated SAXS intensity for each ssDNA structure is weighted and summed to form a single calculated SAXS scattering intensity which can be compared to experiment. As mentioned, the c_1 and c_2 parameters are kept constant at 1.05 and 4. The weight of contribution for each ssDNA structure, ω_n, is fit for each of the 7 structures and must be between 0 and 1. An additional constraint is that the sum of weights for the 7 genes in each chromosome must equal 1. The scaling parameter, c, is constant for each gene in a chromosome. This means that the initial scaling factor for each gene provided by the FoXS software must be removed and rescaled to be made constant. All parameter fitting in the GA process was performed using a least squares optimization with the SciPy package in python.

Once enough chromosomes are selected for the next generation, 2 processes take place. The first process is elitism where the 5 best fit chromosomes are immediately moved into the population of the next generation without undergoing any breeding. Elitism was implemented to maximize the exploitation of chromosomes with a good fit to experiment. The remaining chromosomes undergo a breeding process which consists of crossover and mutation. Every chromosome undergoes a single-point crossover as depicted in Figure 5.1.1. In the 7 gene chromosome, this involves swapping the first 3 genes of one chromosome with the first 3 genes of another chromosome. Ultimately, this results in 2 new chromosomes termed the offspring, while the original 2 chromosomes are termed parents. The parent chromosomes are discarded, and the offspring are prepared to form the next generations population. However, before the next

generation's population is set, there is a 30% chance a gene is mutated in an offspring chromosome. This involves swapping out a gene in a chromosome with a randomly selected gene from the initial pool of 5000 ssDNA structures. The selection of mutation rate is not shown to be influential in the results of this study as shown in Figure S5.1.2. Furthermore, mutation is guaranteed to take place if a chromosome is already in existence for the next generation. This was decided to maximize exploration in the GA.

Overall, this GA undergoes 250 generations, which is shown to be a sufficient selection to converge results (Figure S5.1.3). A GA is required for this type of ensemble optimization because with the selected 7 genes per chromosome and a total of 5000 genes, there are over 1.5×10^{22} possible combinations to test using a brute force method. Furthermore, it should be noted that although 7 genes provide the best results for our GA, each chromosome contains genes that are assigned negligible weights. The 7 genes per chromosome likely provided the best results due to enhanced exploration in the GA, while the number of genes with significant weights is indicative of the flexibility and number of states present for polyT in solution.

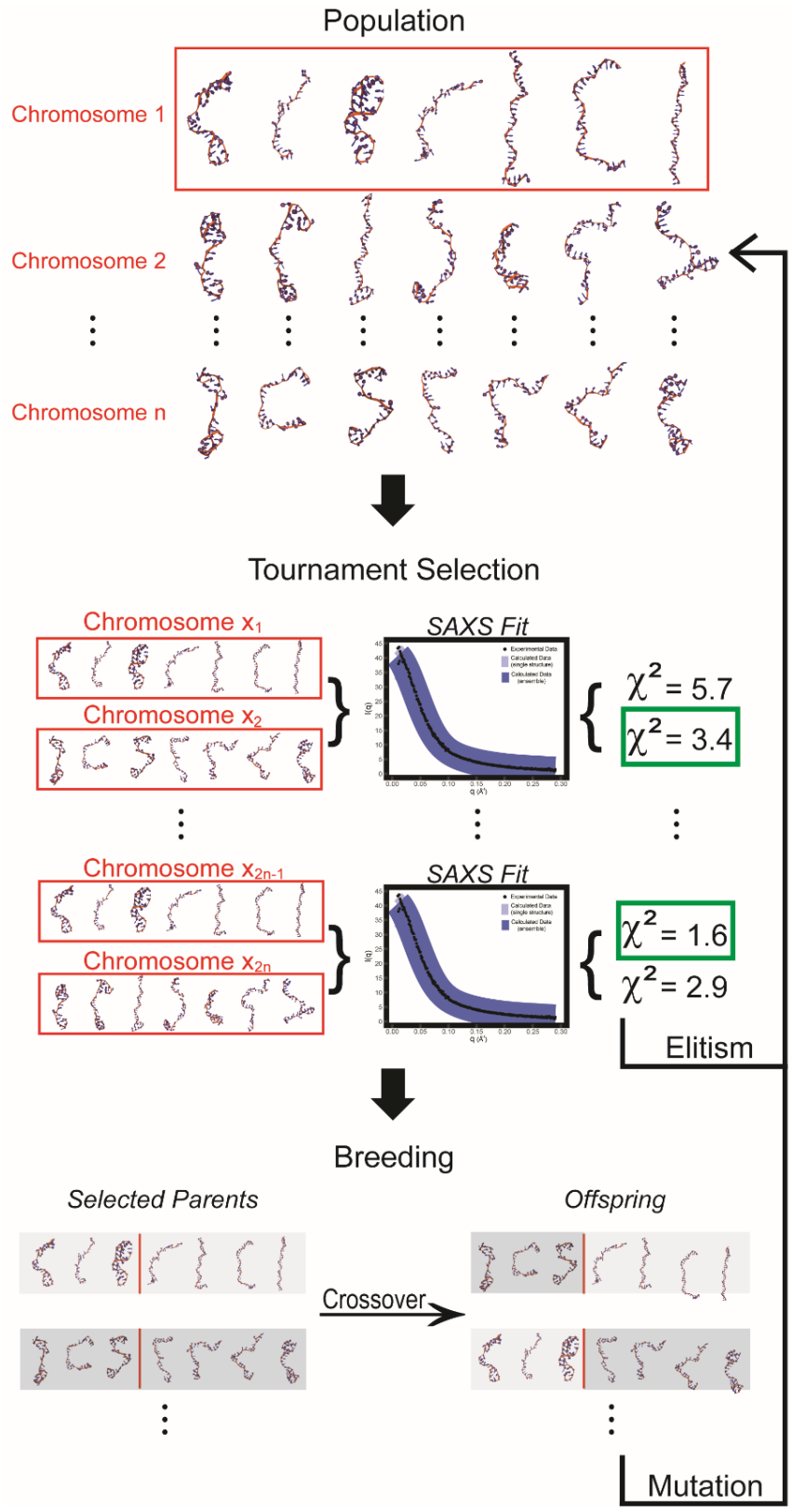


Figure 5.1.1. Illustrates the genetic algorithm scheme used for ensemble optimization.

5.1.2.4 Analysis

The CPPTRAJ module in the AMBER18 package was used to analyze the mass weighted radius of gyration (R_g) and end-to-end distance (R_{ee}) between the 3' and 5' ends center of mass of polyT. This analysis was performed every 100 ps over the entire 500 ns of the simulation.[239] In addition, this analysis was done for every ensemble of ssDNA structures in the last generation of the GA. All 5000 structures within the 500 ns simulations were viable choices for the GA. The 6 dihedral angles in the backbone of ssDNA (α, β, γ, δ, ε, ζ) along with the χ dihedral angle of the polyT nucleobase were calculated using CPPTRAJ. This analysis was applied to each monomer of polyT containing all 7 dihedral angles of interest. This analysis was done over the last 250 ns of implicit and explicit solvent simulations at 100 ps intervals. The dihedral angles were also calculated for the polyT structures found in the last generation of the GA for comparison to simulation.

5.1.3 Results and Discussion

5.1.3.1 Implicit Solvent

The structure of polyT is compared between simulation and the GA's best-fit ensembles through analysis of polyT's R_g and R_{ee}. Figure 5.1.2 shows the four combinations of force fields tested in implicit solvent where (a) is the ff99 force field in IGB1 solvent, (b) is the ff99 force field in IGB5 solvent, (c) is the ff99 force field in IGB8 solvent, and (d) is the bsc1 force field in IGB1 solvent. Experimental R_g and R_{ee} values are plotted for reference, where the black dashed line represents the experimental mean and shaded green bars indicate experimental error from SAXS and FRET. On the left side of Figure 5.1.2, the contour plot represents the MD simulation where

transparency indicates relative frequency of occurrence in simulation. The best ensemble from

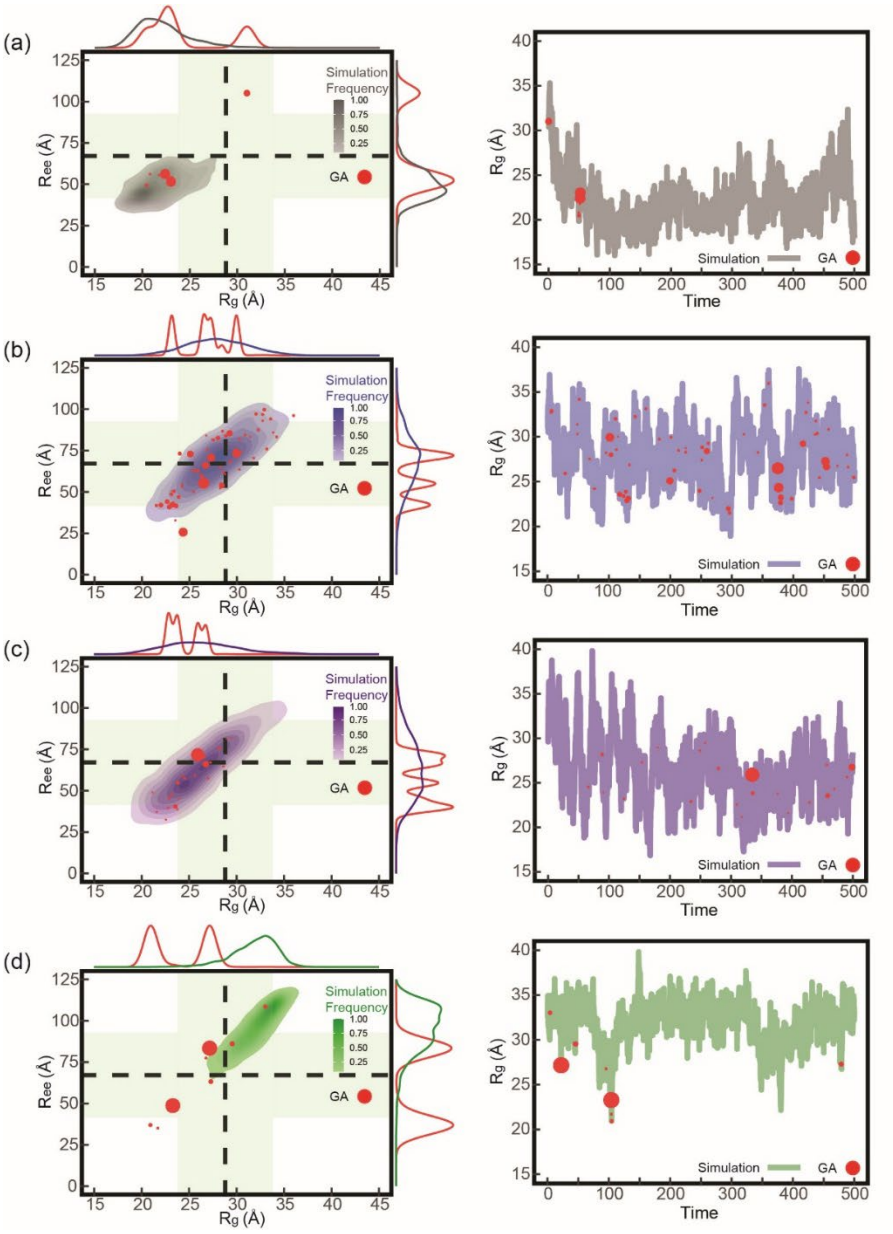


Figure 5.1.2. Shows a comparison between the R_g and R_{ee} calculated from MD simulation (bivariate contour plot) where transparency is correlated with relative frequency of sampling and the GA (red circles) where size is correlated with weight. Density plots are in the margin to compare conformational distributions between simulation and the GA. Experimental values and error are depicted by the black dashed line and light green rectangles. The temporal profile of R_g (right) show which part of the MD simulations the GA was choosing structures from. The simulations represented are (a) ff99 with igb1 (b) ff99 with igb5 (c) ff99 with igb8 and (d) bsc1 with igb1.

each simulation was then determined by the GA where each structure is assigned a weighted contribution representing its prominence in solution. The larger the red circle, the more weight and contribution that structure has to the final scattering intensity of the ensemble. It should be noted that not all red circles overlay the conformations from MD simulations represented by the bivariate contour plots. This indicates that the GA is choosing a structure from a conformation that is rarely explored in simulation and the relative frequency is negligible compared to the bulk of the simulation. This implies a mismatch between the conformations of polyT regularly explored in the MD simulation and the conformation of polyT that best match experiment as determined by the GA. In the margins, density plots are provided for a comparison of how the conformations in simulation compare to the conformations selected by the GA. On the right side of the figure, the temporal profile of polyT's R_g is provided by a solid line. The temporal profile highlights at which point the simulation produced structures that compare best to experiment. The polyT structures selected by the GA are again represented by red circles where weight and size are correlated.

Figure 5.1.2 (a) represents the ff99 force field for DNA in IGB1 solvent. Looking at the left side of the figure, the red circles do not coincide well with the dark, frequently explored areas of the contour plot indicating that structures chosen by the GA are not well represented by the dynamics carried out in simulation. In particular, there is one GA chosen structure that lies well outside of the shaded region of the contour plot. Looking at the right side of Figure 5.1.2 (a) most of the selected structures come from a localized time frame and from early in the simulation. The temporal profile shows structures are chosen by the GA when the R_g is still decreasing towards a lower equilibrium value. Thus, the ff99 force field appears to poorly represent the structure and dynamics of polyT in solution. Specifically, the simulation appears to explore conformations that

are too collapsed compared to experiment. This is corroborated by the bivariate contour plot being below the experimentally derived R_g and R_{ee} values.

The MD simulation for the ff99 force field in IGB5 implicit solvent (Fig 5.1.2 (b)) appears to frequently sample structures that represent experimental (black dashed lines) values well and traverse the expected deviation from the mean (green error bars). Furthermore, the ensembles of polyT chosen by the GA appear to be well dispersed among the simulation with most of the prominent structures falling near heavily sampled conformations in the MD simulation. Most lesser-weighted conformations of polyT coincide with more infrequent sampling in the simulation. This indicates that the simulation produces chain conformations and dynamics that compare well to the best-fit ensemble selected by the GA. This is further corroborated by the temporal profile (right) showing the GA selects an ensemble of structures across the bulk of the simulation. However, it should be noted that there are a few highly prominent structures that do not correspond well with frequently visited conformation throughout the ff99-IGB5 simulation. Thus, there may be a slight mismatch between the simulation's representation of polyT and the desired representation of polyT that best match experimental measurements.

The ff99 force field in IGB8 implicit solvent (Fig 5.1.2 (c)) appears to have a strong match between the MD simulation and GA selected ensembles. The heavily weighted structures chosen by the GA all lie near frequently explored conformations of the MD simulation. Moreover, none of the ssDNA conformations selected by the GA come from sparsely explored conformations in simulation. The temporal profile shows most structures chosen by the GA occur and are well dispersed in the latter half of the simulation, which is when the simulation would be most converged. Although the simulation and conformations of polyT selected by the GA are inside of the expected experimental range, it appears to be on the lower end of the expected experimental

values. Thus, the ff99-IGB8 simulation may have a slight tendency to produce overly compact structures of polyT.

The bsc1 force field in IGB1 implicit solvent also appears to poorly represent ssDNA (Fig 5.1.2 (d)). In this instance, the MD simulation explores ssDNA conformations that are too elongated as the bivariate contour plot is on the upper end of the experimental R_g and R_{ee} values. Furthermore, the structures chosen from the GA do not line up well with the frequently visited conformations in the bsc1-IGB1 MD simulation (left) and are chosen mostly from rarely sampled conformations with the smallest R_g values (right). Thus, the bsc1-IGB1 MD simulation does not represent the best-fit ensemble selected from the GA.

Overall, Figure 5.1.2 indicate that the ff99 force field in IGB5 and IGB8 implicit solvent are superior to the ff99 force field and bsc1 force field in IGB1 implicit solvent. The best-fit ensembles selected by the GA from the ff99 force field in IGB5 and IGB8 implicit solvent appear to reasonably agree with the frequency of conformations explored in simulation and agree with experimental FRET and SAXS values. This conclusion is supported by the χ^2 values reported in Table 5.1.2 which shows the best-fit ensembles selected from the ff99 simulations in IGB5 and IGB8 solvent match the experimental ensemble well. The higher χ^2 values for ff99 and bsc1 in IGB1 solvent indicate both simulations poorly represent the experimental ensemble of polyT structures.

Table 5.1.2. Shows the goodness of fit for the ensemble of polyT structures chosen by the GA for each implicit solvent simulation

Goodness of fit (χ^2)			
ff99 IGB1	ff99 IGB5	ff99 IGB8	bsc1 IGB1
5.84 $\pm$ 0.1805	1.26 $\pm$ 0.3491	1.75 $\pm$ 0.0428	8.95 $\pm$ 1.4208

Additional insight on the structure of polyT in solution and the performance of each MD simulation is achieved by analyzing the dihedral angles of the ssDNA in each of the best-fit

ensembles. Specifically, the α, β, γ, δ, ε, ζ, and χ dihedral angles are examined. Their placement in the ssDNA chain is illustrated in Figure 5.1.3. Understanding how the distribution of these dihedral angles change and how that affects the χ^2 goodness of fit metric for each ensemble presents a more fundamental understanding of how the ssDNA structure changes compared to its double stranded counterpart. Ultimately, the analysis of dihedral angles in the ssDNA chain provides guidance on the source of flexibility for ssDNA and provides a potential guideline for future parameterization of simulations to match these distributions.

The explored space for each dihedral angle is plotted against one another as a bivariate contour plot in Figure 5.1.4. It should be noted that Figure 5.1.4 corresponds only to the ssDNA structures selected by the GA, which is represented by the red circles in Figure 5.1.2. Thus, the frequency and distribution of the dihedral angles throughout the entire MD simulation might vary from the ensemble of structures represented in Figure 5.1.4. The χ^2 goodness of fit metric is calculated (Eq. 5.1.3) for the polyT structures selected by the GA for each simulation. The values are reported at the top of Figure 5.1.4. The mean goodness of fit for all chromosomes in the final generation of the GA indicate that the MD simulation for ff99 in IGB5 implicit solvent (χ^2=1.26) provides the best-fit ensemble of structures. This is followed by the ff99 force field in IGB8 implicit solvent

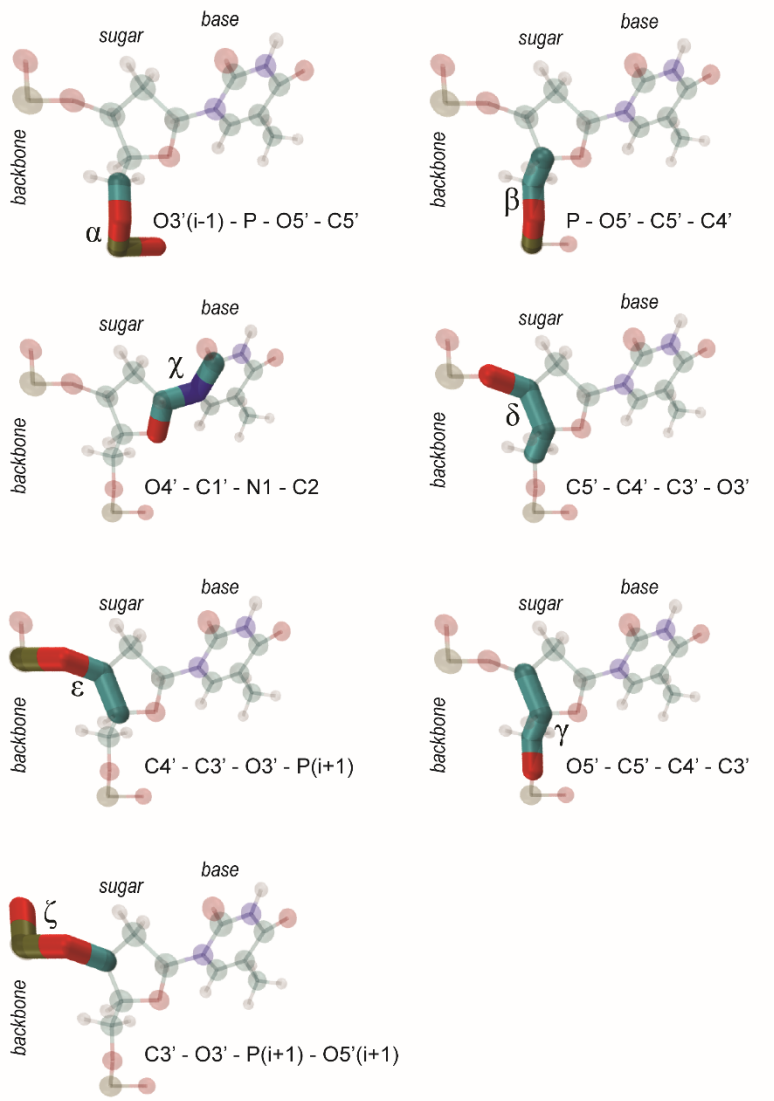
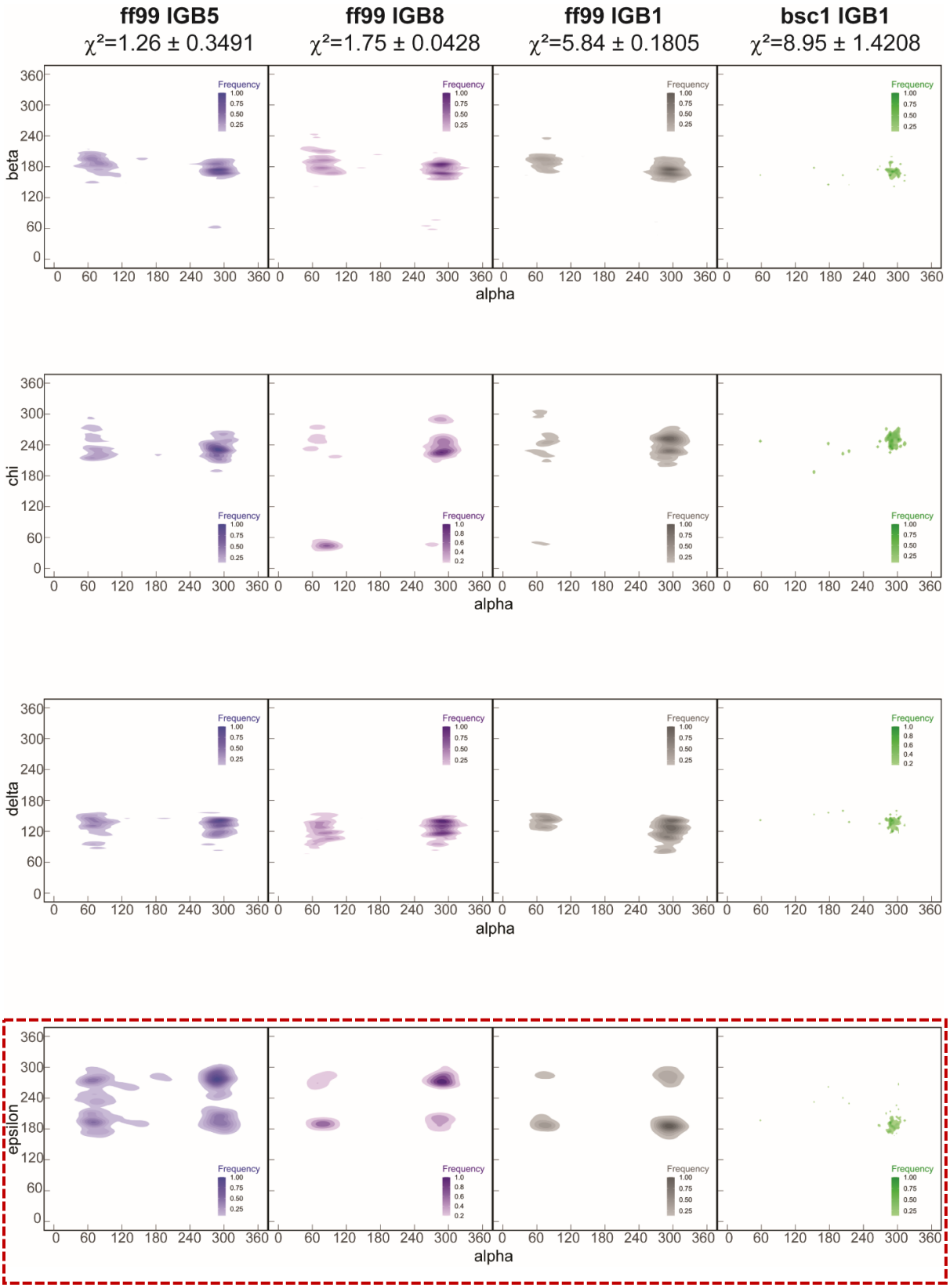
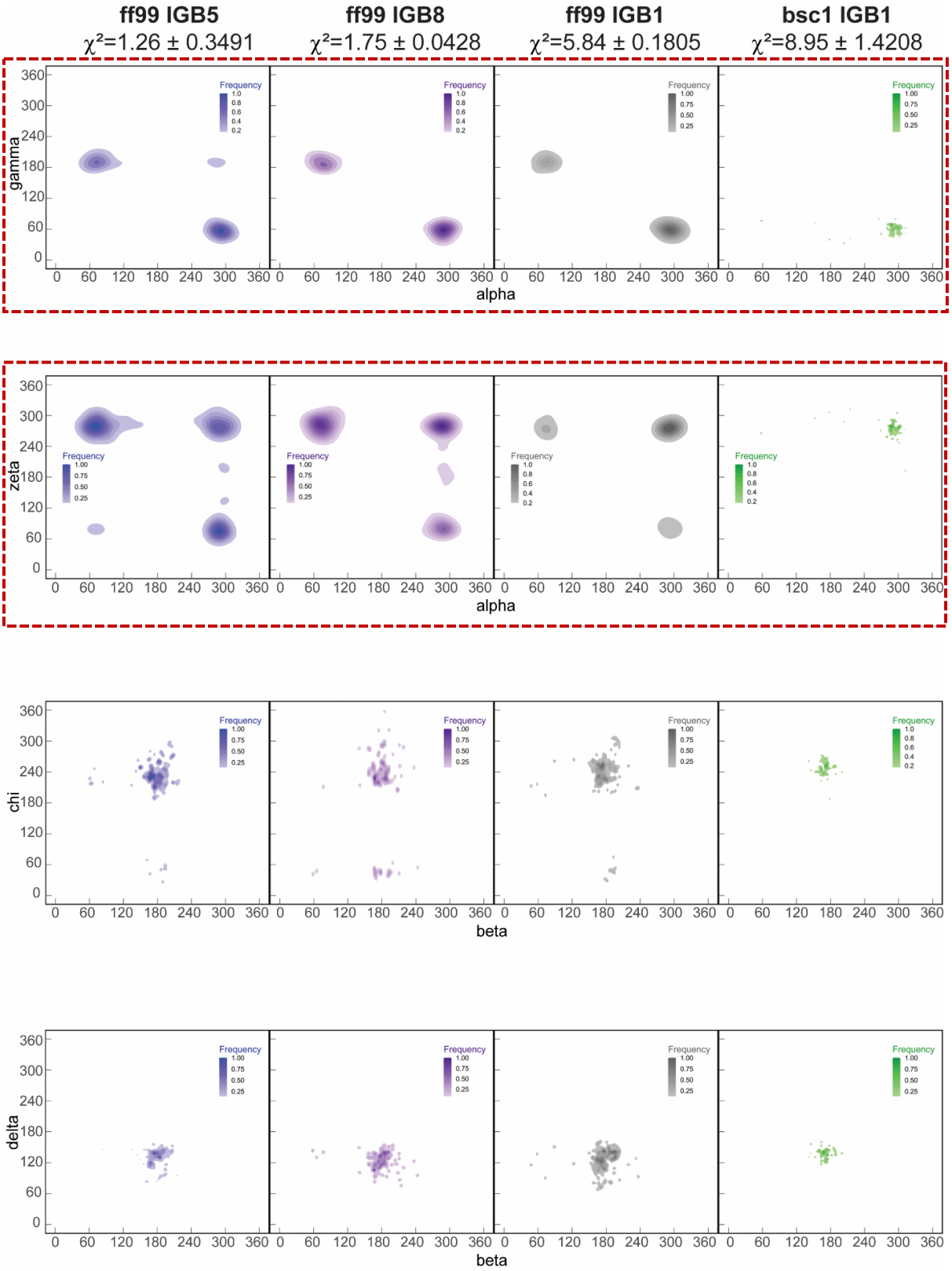


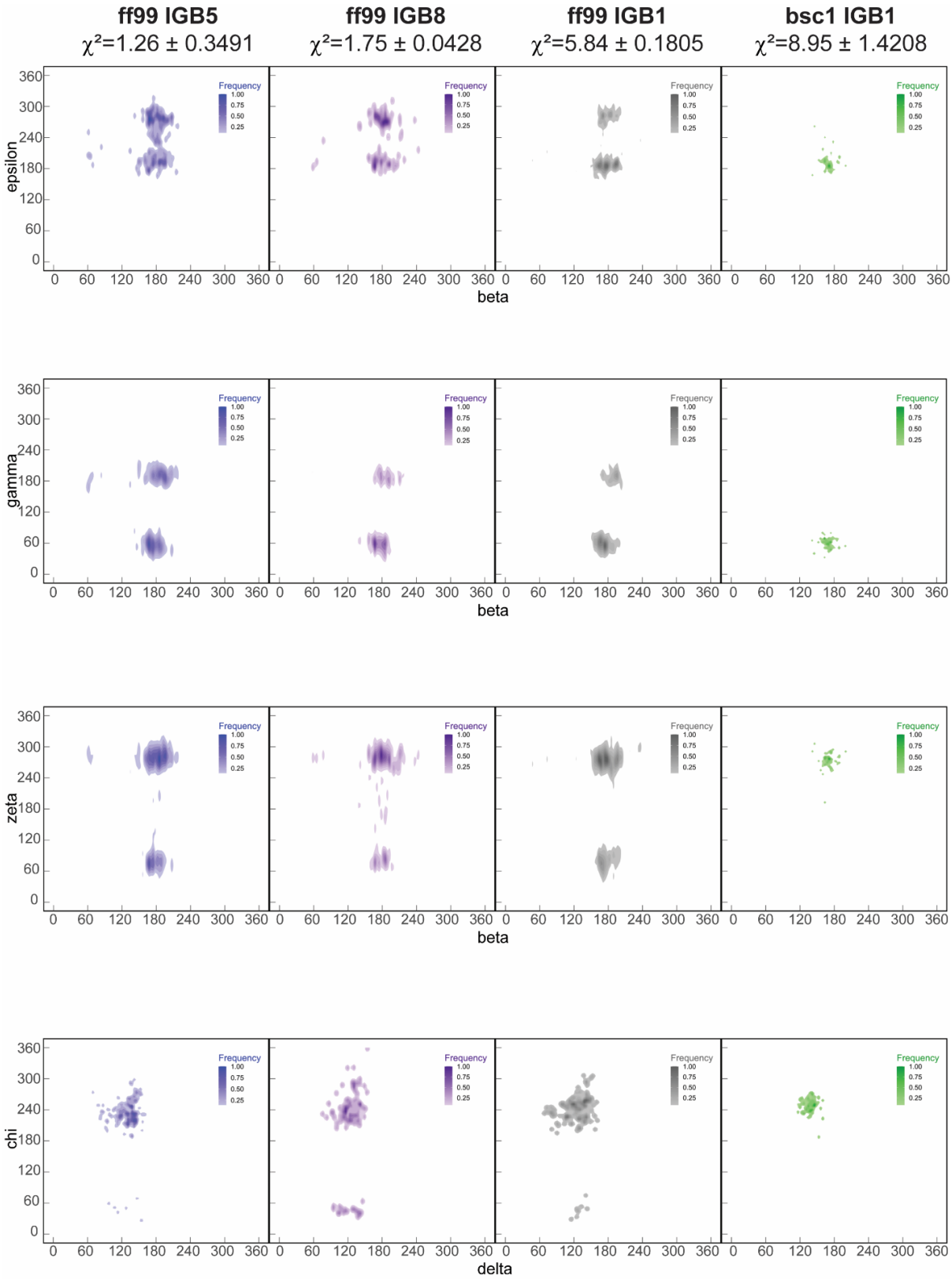
Figure 5.1.3. Illustrates the location and atoms that make up each of the dihedral angles analyzed in the ssDNA structures. All non-contributing atoms for a dihedral angle of interest are transparent. The backbone, sugar, and base of DNA are indicated for reference.

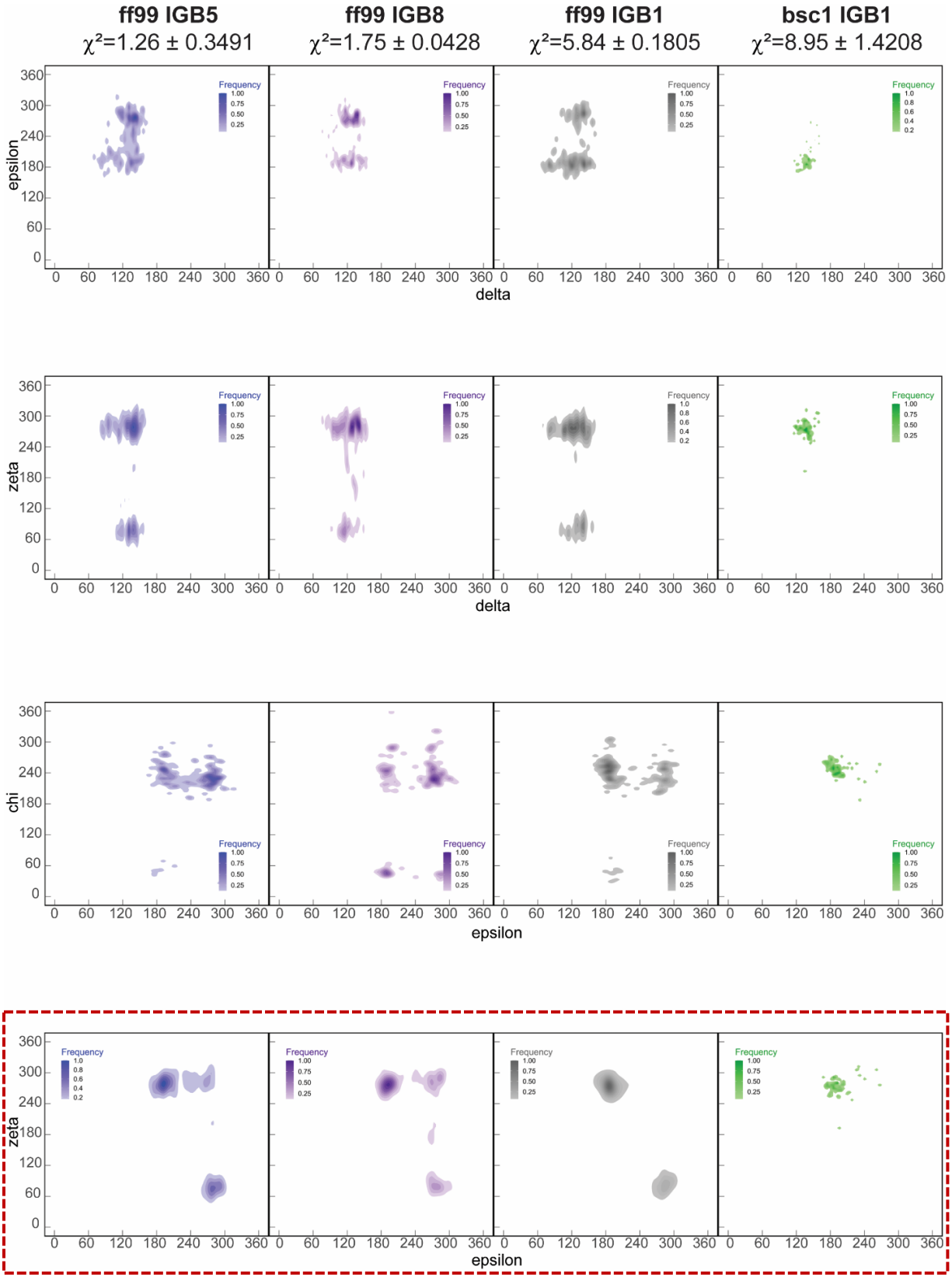
(χ^2=1.75), ff99 force field in IGB1 implicit solvent (χ^2=5.84) and lastly the bsc1 force field in IGB1 implicit solvent (χ^2=8.95).

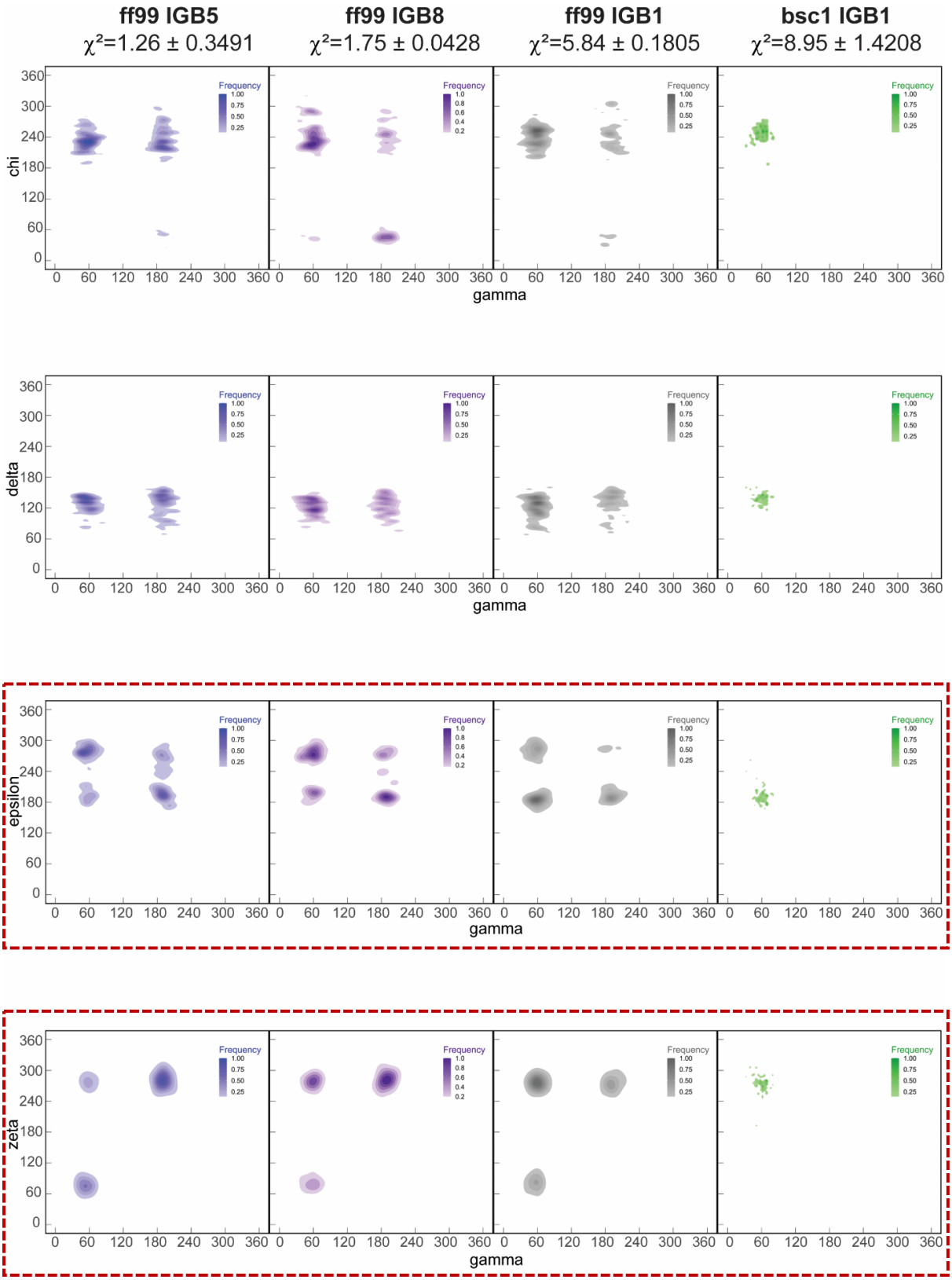
Figure 5.1.4. Shows bivariate contour plots for the alpha, beta, chi, delta, epsilon, gamma, and zeta dihedral angles. The dihedral angles are calculated from the ensemble of structures chosen by the GA. The GA was applied to select structures from MD simulations with the ff99 force field in IGB5, IGB8 and IGB1 implicit solvent and from an MD simulation with the bsc1 force field in IGB1 implicit solvent. The goodness of fit (χ^2) for the ensemble of structures chosen by the GA is provided for each simulation. The lower the value, the better the fit to experiment. The red dashed box indicates bivariate plots that likely have significant influence on the goodness of fit.

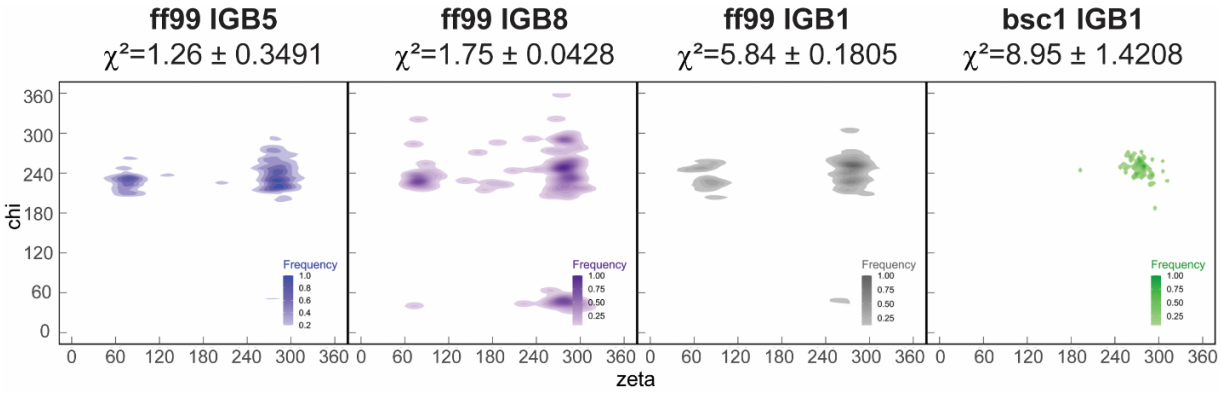

ff99 IGB5
χ²=1.26 ± 0.3491
ff99 IGB8
χ²=1.75 ± 0.0428
ff99 IGB1
χ²=5.84 ± 0.1805
bsc1 IGB1
χ²=8.95 ± 1.4208
beta
chi
delta
epsilon
alpha
Frequency


ff99 IGB5
χ²=1.26 ± 0.3491
ff99 IGB8
χ²=1.75 ± 0.0428
ff99 IGB1
χ²=5.84 ± 0.1805
bsc1 IGB1
χ²=8.95 ± 1.4208
gamma
alpha
zeta
alpha
chi
beta
delta
beta
Frequency


ff99 IGB5
χ²=1.26 ± 0.3491
ff99 IGB8
χ²=1.75 ± 0.0428
ff99 IGB1
χ²=5.84 ± 0.1805
bsc1 IGB1
χ²=8.95 ± 1.4208
Frequency
1.00
0.75
0.50
0.25
epsilon
gamma
zeta
chi
beta
delta
0
60
120
180
240
300
360


ff99 IGB5
χ²=1.26 ± 0.3491
ff99 IGB8
χ²=1.75 ± 0.0428
ff99 IGB1
χ²=5.84 ± 0.1805
bsc1 IGB1
χ²=8.95 ± 1.4208
epsilon
delta
zeta
delta
chi
epsilon
zeta
epsilon
Frequency


ff99 IGB5
χ²=1.26 ± 0.3491
ff99 IGB8
χ²=1.75 ± 0.0428
ff99 IGB1
χ²=5.84 ± 0.1805
bsc1 IGB1
χ²=8.95 ± 1.4208
chi
delta
epsilon
zeta
gamma
Frequency


ff99 IGB5
χ²=1.26 ± 0.3491
ff99 IGB8
χ²=1.75 ± 0.0428
ff99 IGB1
χ²=5.84 ± 0.1805
bsc1 IGB1
χ²=8.95 ± 1.4208
chi
zeta
0
60
120
180
240
300
360
Frequency

Across Figure 5.1.4, the bsc1 force field in IGB1 implicit solvent is the most restricted in terms of exploring a range of dihedral angle values. This is likely due to the re-parameterization of this force field as compared to ff99. In particular, the re-parameterization was to maintain the appropriate base-stacking and double helix conformation that is characteristic of double stranded DNA. Thus, the dihedral angles likely have a difficult time transitioning from the energy minimum determined for B-DNA. The ensemble of structures chosen from the bsc1 force field in IGB1 implicit solvent has a significantly worse goodness of fit than the ensemble of structures chosen from the ff99 force field in all solvents, indicating that ssDNA deviates from the dihedral angles that are characteristic of double-stranded DNA. This is not surprising given that ssDNA is known to be more flexible than its double-stranded counterpart.

The chromosomes selected by the GA for the ff99 force field in IGB5, IGB8, and IGB1 implicit solvent contain largely similar distributions of dihedral angles across structures of polyT. In particular, the β, δ, and χ dihedral angles do not appear to have a clear trend that explain changes in the χ^2 goodness of fit metric across simulations. However, the α, γ, ε, and ζ dihedral angles appear to correlate with one another in a way that can explain why chromosomes selected from simulations of ff99 with IGB5 solvent match experiment better than chromosomes from ff99 with IGB8 solvent and much better than chromosomes from ff99 and bsc1 with IGB1 solvent. Coupled dihedral angle with discernable trends are highlighted by a red, dashed box in Figure 5.1.4 and include all combinations of the α, γ, ε, and ζ dihedral angles (α-ε, α-γ, α-ζ, ε-ζ, γ-ε, and γ-ζ). In general, the dihedral angles are increasingly destabilized for polyT as the goodness of fit to experiment increases. The α-γ and ε-ζ bivariate contour plots can be used as a representative example for discussion.

In the instance of the α-γ bivariate contour plot, it can be clearly seen that there is limited exploration in the dihedral angles for bsc1 in IGB1 implicit solvent. This is associated with an average χ^2 value of 8.95, which indicates the group of structures inside each chromosome are a very poor fit compared to experiment. As the average χ^2 value of chromosomes drops to 5.84 with the ff99 force field in IGB1 solvent, the dihedral angles transition into a new state creating two distinct states for the α-γ dihedral angle pair. However, this new state appears in limited frequency and the bulk of the dihedral angles remain in the state seen in the simulation with the bsc1 force field. Thus, the χ^2 value indicates the chromosomes from the ff99 force field in IGB1 solvent still do not fit experimental structures well, even though there is an improvement associated with the presence of the second dihedral angle state.

It appears that the goodness of fit for chromosomes is sensitive to the frequency with which this new α-γ state occurs. This is supported by an increased goodness of fit for chromosomes in the ff99 simulation with IGB8 implicit solvent which has an observable increase in the presence of the second α-γ state. However, considering the interrelation between all dihedral angles, it is difficult to definitively say to what extent the χ^2 value is influenced by changes to this dihedral angle specifically. While the increased presence of the second α-γ state is discernable, it appears to be minimal compared to the drop in χ^2 value from 5.84 to 1.75. This indicates other structural changes to the ensemble of polyT aid in creating a better match to experiment.

The correlation between a decreased χ^2 value and destabilization of the α-γ dihedral angle is further established for the simulation of polyT with the ff99 force field and IGB5 implicit solvent. In this simulation, the GA selects chromosomes that introduce a transition to a third α-γ state, albeit infrequently. The frequency of the two previously established α-γ states does not appear to deviate from what was observed in the ff99 simulation with IGB8 implicit solvent. Thus, the introduction

of this third α-γ state appears to play a role in the χ^2 value decreasing to 1.26 for the ff99 force field in IGB5 implicit solvent.

As previously discussed, structural changes associated with other dihedral angles are happening concurrently with the transitions identified for the α-γ dihedral angles. For example, the ε-ζ dihedral angles follow a similar trend of destabilization as the structures of polyT in selected chromosomes better match those observed experimentally. In a comparable fashion to the α-γ dihedral angle, the presence of a second state in the ε-ζ dihedral angles corresponds with a large decrease in the average χ^2 value for chromosomes selected from the ff99-IGB1 (5.84) simulation compared to the bsc1-IGB1 simulation (8.95). Thus, the presence of this second state appears to better match the experimental structure of polyT in solution. Furthermore, a third ε-ζ state emerges for chromosomes selected from the ff99-IGB8 simulation. It is important to note this third state emerges concurrently with a notable shift in frequency for the two a α-γ dihedral angle states highlighted earlier. However, because the frequency shift for each α-γ state appeared to be disproportionate to the increased match to experiment, the emergence of a third state in the ε-ζ dihedral angles appears to be the major driver in why the chromosomes better match experimental observations. This does not discredit the importance of accurately capturing the frequency of dihedral angle states, it merely indicates the frequency of visited states plays a secondary role to the emergence of the new states.

Lastly, Figure 5.1.4 shows the χ dihedral angle has increased flexibility in the MD simulation with ff99 in IGB8 implicit solvent. There is a non-negligible presence of a second state for this dihedral angle, however, it is difficult to determine if it has any significant impact on the χ^2 goodness of fit metric. A lack of significance could stem from the χ dihedral angle being related to the base and sugar of polyT, not the backbone which would more directly impact the chain's

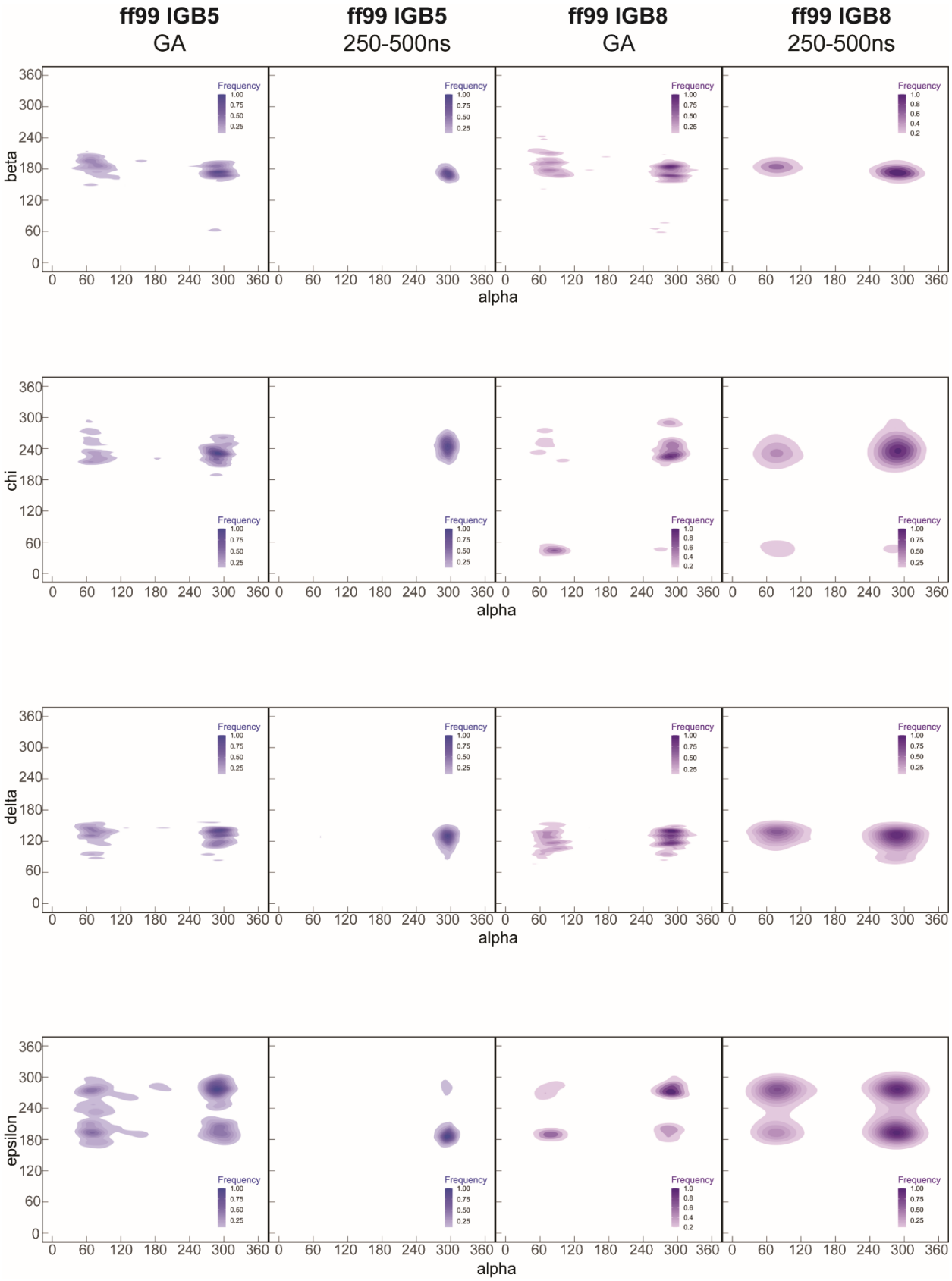
conformations. However, a previous assessment showed that the ff99 force field in IGB8 implicit solvent had the lowest base stacking compared to any other tested MD simulation.[239] The increased flexibility in this dihedral angle likely impacts the base stacking observed in the polyT chain given its location in the ssDNA chain (Fig 5.1.3). While experimental observations have indicated base stacking is low in polyT, it is difficult to quantify. Thus, it can be concluded that the destabilization of the χ dihedral angle is supported by experimental observations, but the degree to which it should be destabilized is still unclear.

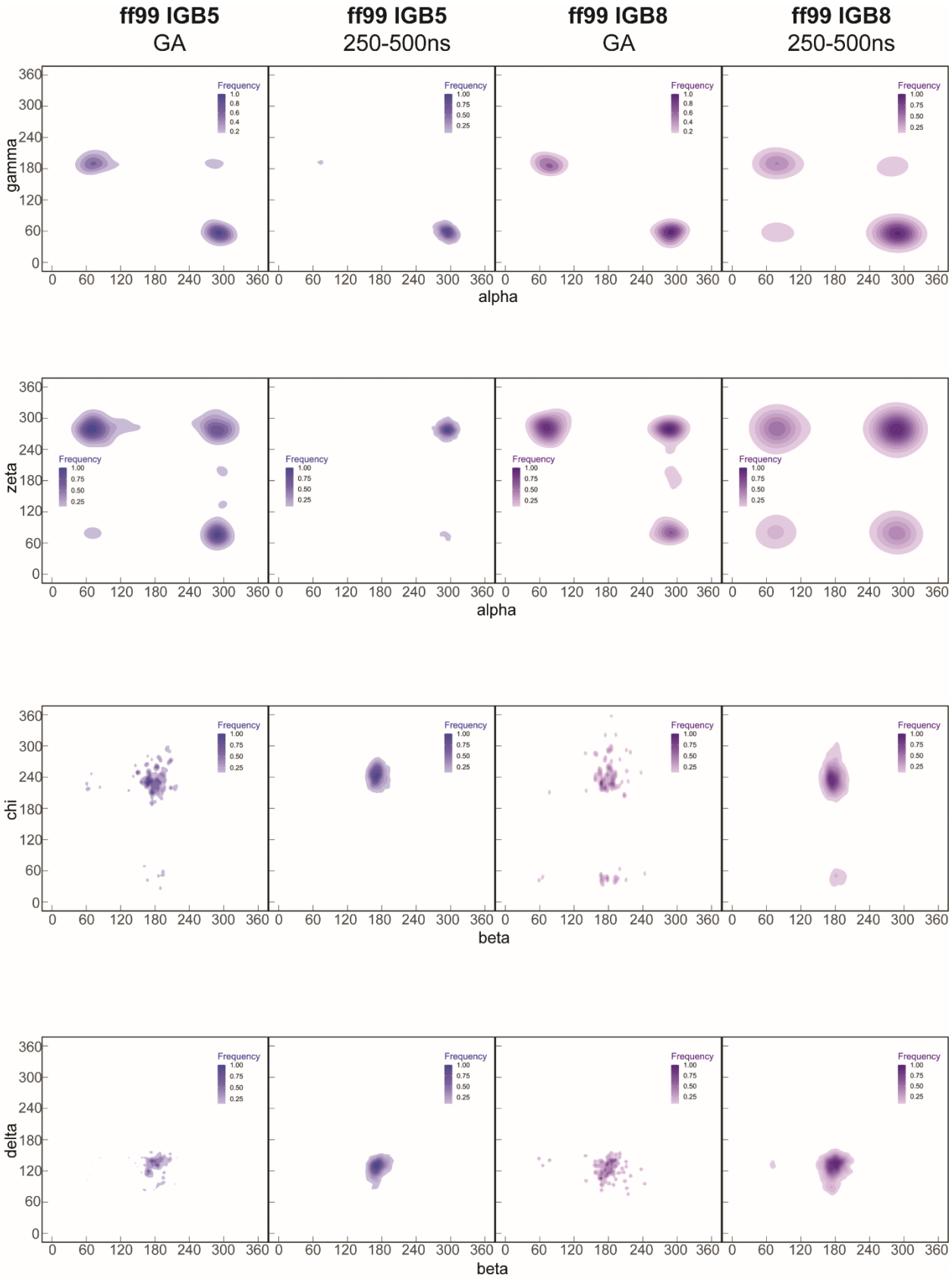
Overall, it is difficult to determine the significance of each individual shift among the dihedral angles shown in Figure 5.1.4. However, destabilization of the α, γ, ε, and ζ dihedral angles in the backbone of DNA appears to be essential in replicating the experimentally observed structure of polyT in solution. Figure 5.1.4 highlights the emergence of important states for each combination of the α, γ, ε, and ζ dihedral angles, although the exact frequency with which each state should be present is still unclear. In large part, this comes from difficulty in decoupling all dihedral angles from one another. The 2D bivariate plots do not capture the multi-dimensional nature of this problem. Furthermore, the frequency and location with which each state of dihedral angles is present along the 30mer chain is unclear. The existence of one set of dihedral angles could prohibit the existence of another set of dihedral angles in an adjacent position along the polyT chain.

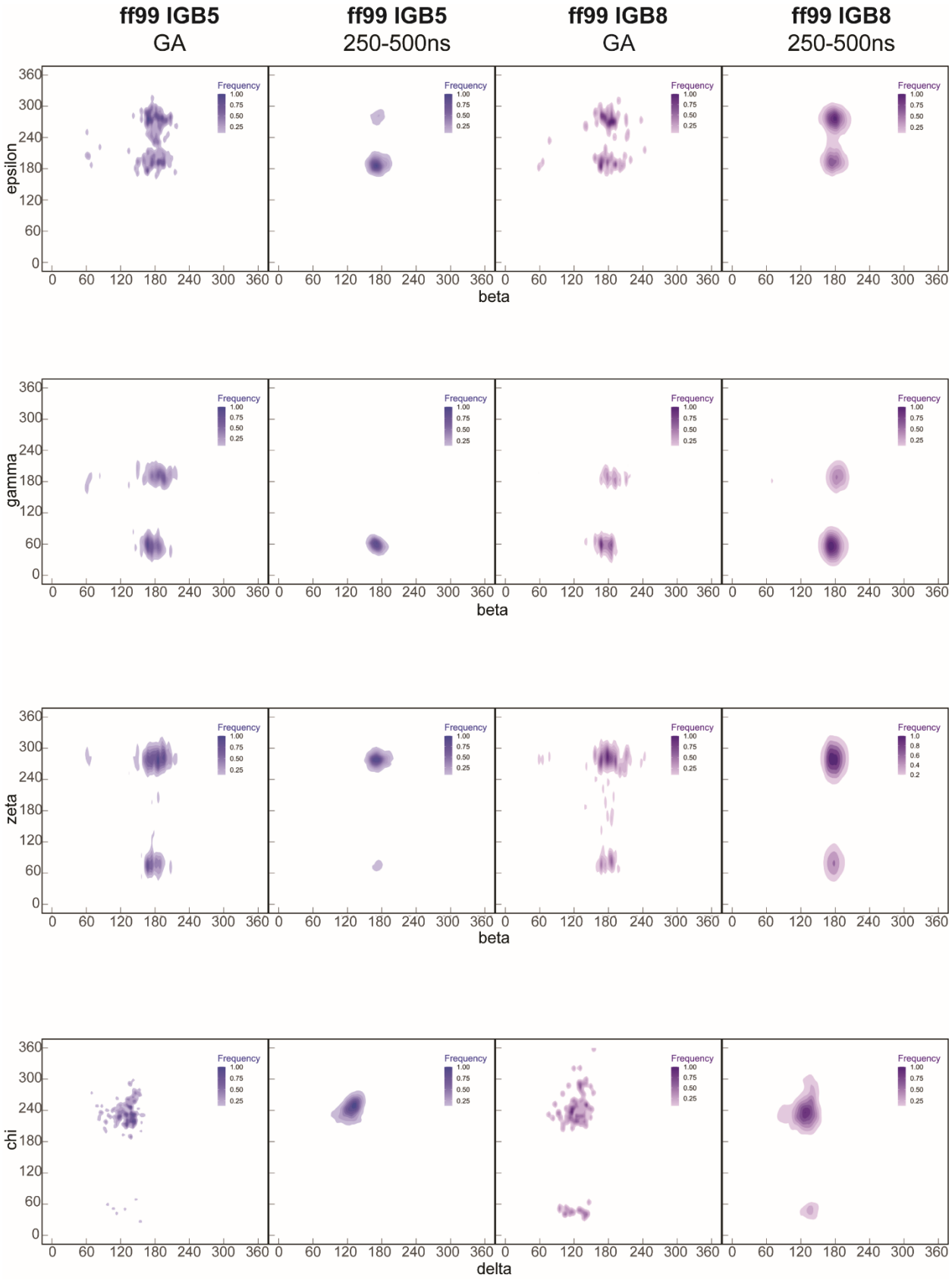
Figure 5.1.4 established the desired distribution of dihedral angles to accurately capture the structure of polyT in solution. However, as established in Figure 5.1.2, the weight and frequency of polyT conformations chosen by the GA does not necessarily line up with the prominence of those conformations in simulation. Figure 5.1.5 compares the dihedral angles of structures chosen from the GA and the dihedral angles predominantly present in the MD simulation. The dihedral

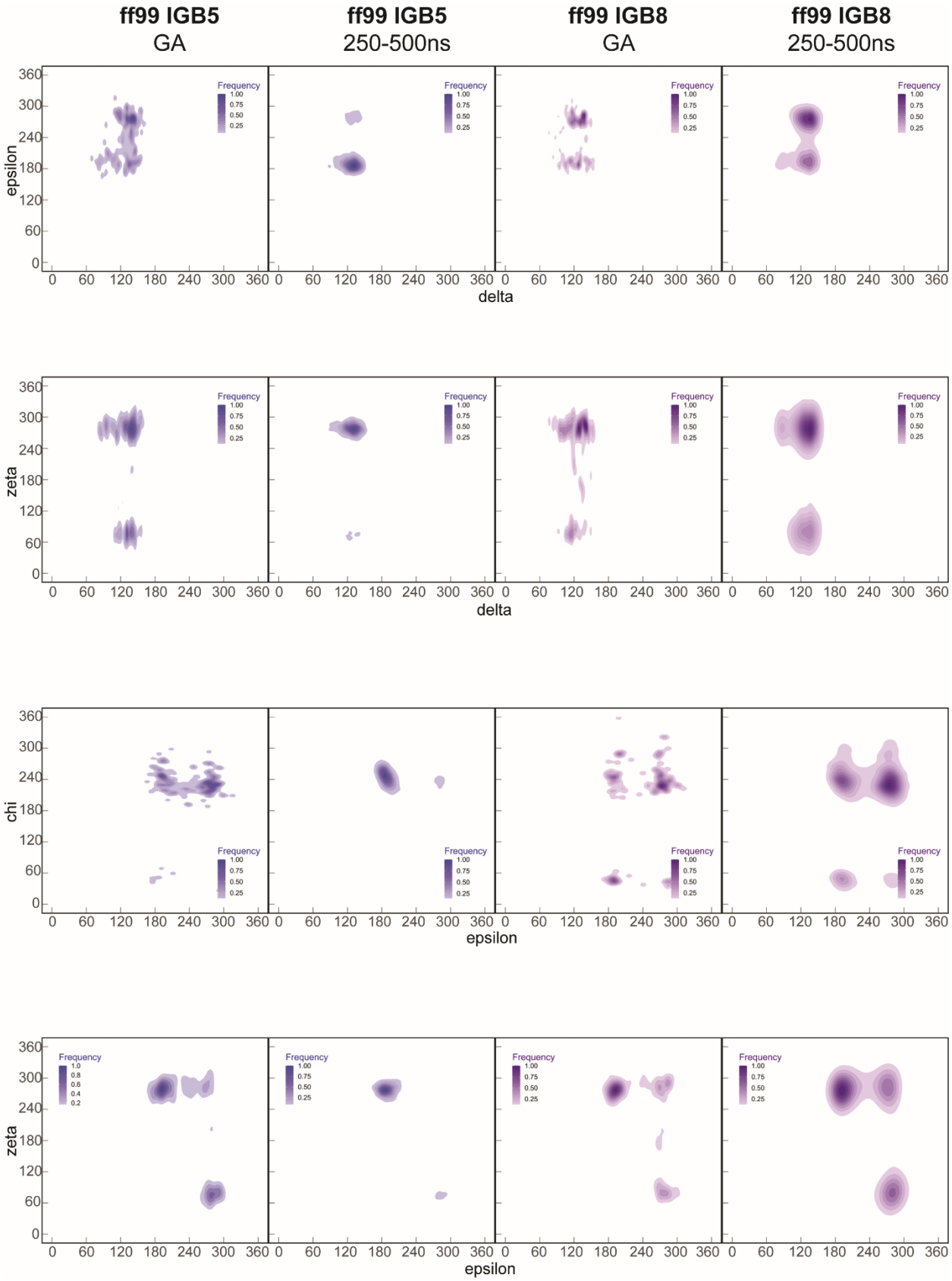
angles from each MD simulation should mirror the distributions for structures selected by the GA if the simulation accurately depicts the conformation of polyT in solution.

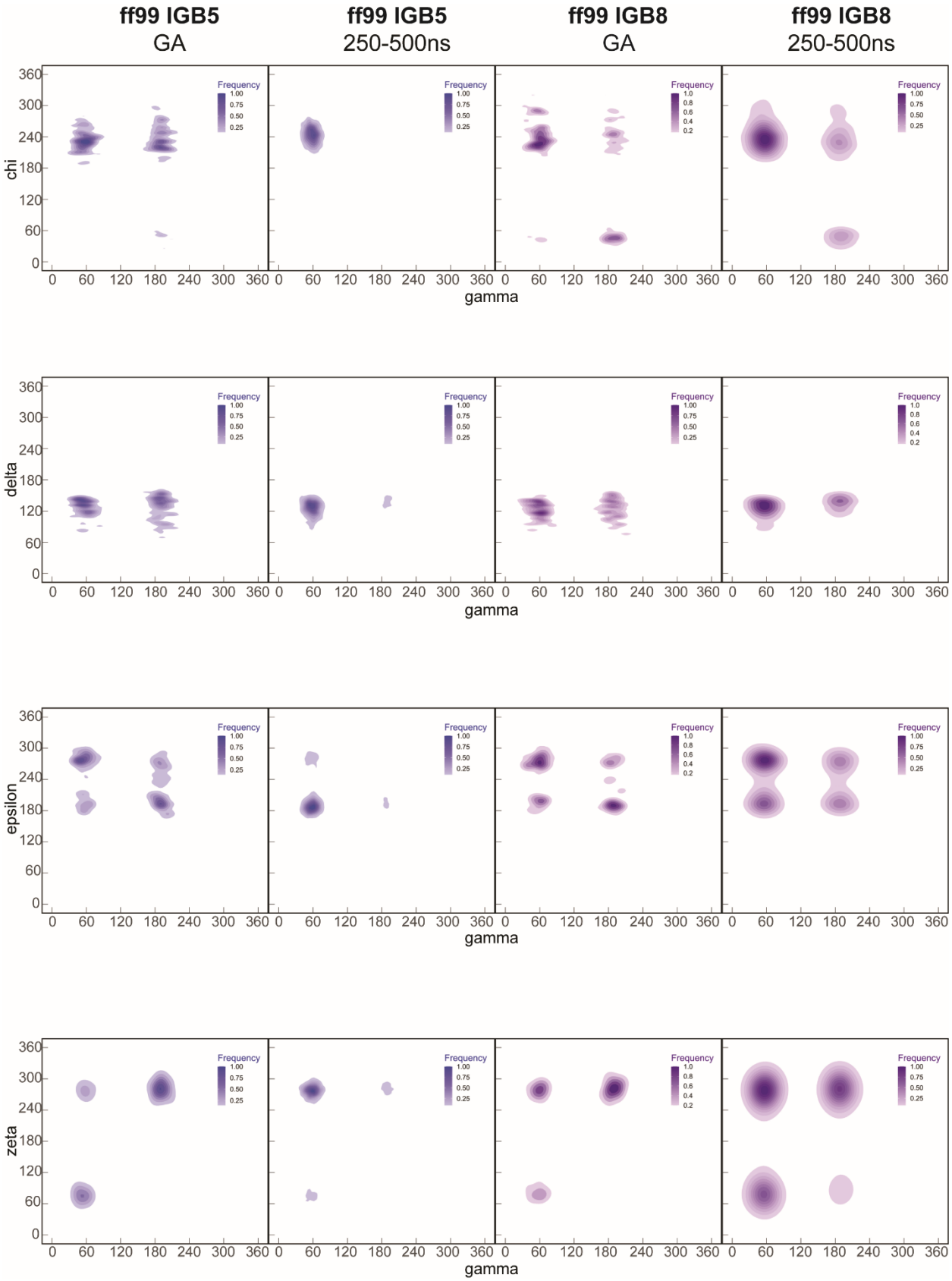
Figure 5.1.5. Shows bivariate contour plots for the alpha, beta, chi, delta, epsilon, gamma, and zeta dihedral angles. The dihedral angles are calculated from the ensemble of structures chosen by the GA and compared to the dihedral angles calculated from MD simulation. The simulations with the ff99 force field in IGB5 and IGB8 implicit solvent are shown due to their strong performance compared to experiment. The selection of presenting 250-500 ns is discussed in Figure S5.1.4.


ff99 IGB5
GA
ff99 IGB5
250-500ns
ff99 IGB8
GA
ff99 IGB8
250-500ns
beta
chi
delta
epsilon
alpha
Frequency


ff99 IGB5
GA
ff99 IGB5
250-500ns
ff99 IGB8
GA
ff99 IGB8
250-500ns
gamma
alpha
zeta
alpha
chi
beta
delta
beta
Frequency
0
60
120
180
240
300
360


ff99 IGB5
GA
ff99 IGB5
250-500ns
ff99 IGB8
GA
ff99 IGB8
250-500ns
epsilon
gamma
zeta
chi
beta
delta
Frequency
0
60
120
180
240
300
360


ff99 IGB5
GA
ff99 IGB5
250-500ns
ff99 IGB8
GA
ff99 IGB8
250-500ns
epsilon
delta
zeta
delta
chi
epsilon
zeta
epsilon
Frequency
0 60 120 180 240 300 360


ff99 IGB5
GA
ff99 IGB5
250-500ns
ff99 IGB8
GA
ff99 IGB8
250-500ns
chi
delta
epsilon
zeta
gamma
Frequency

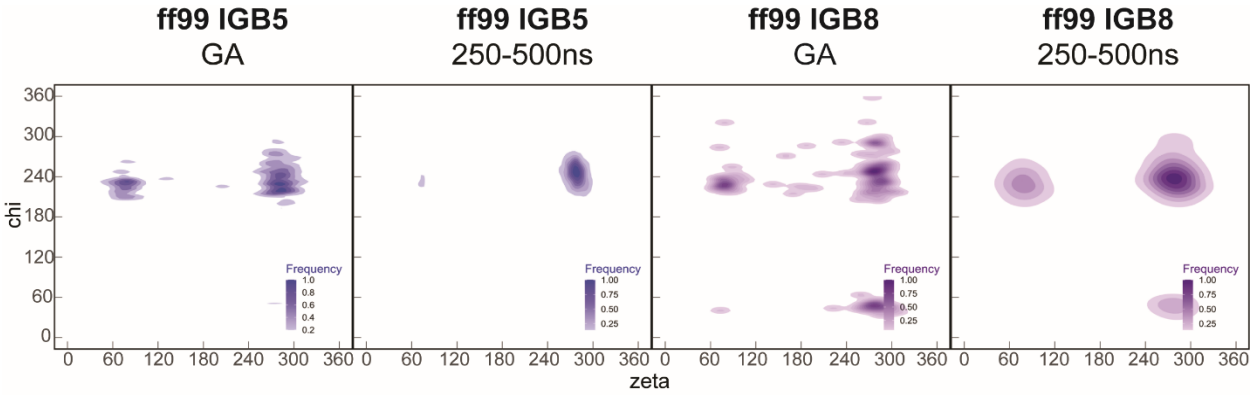

ff99 IGB5
GA
ff99 IGB5
250-500ns
ff99 IGB8
GA
ff99 IGB8
250-500ns
chi
zeta
Frequency
0
60
120
180
240
300
360

Figure 5.1.5 shows that the dihedral angles from simulation do not always replicate the desired distribution of dihedral angles found in structures selected by the GA. The simulation of ff99-IGB5 deviates farther from the desired distribution than the ff99-IGB8 simulation. This indicates that while the best representation of polyT in solution was chosen from the ff99-IGB5 simulation, the ff99-IGB8 simulation protocol may be the most practical for MD studies of ssDNA. This is further corroborated by referring to Figure 5.1.2 which shows heavily weighted polyT conformations selected by the GA fall near heavily sampled conformations of the simulation. On the contrary, the GA selected and heavily weighted some polyT conformations that correspond to infrequently sampled conformations in the ff99-IGB5 simulation.

Figure 5.1.5 highlights some problematic dihedral angles for the MD simulations with the ff99 force field in IGB5 and IGB8 implicit solvent. The simulation of ff99 in IGB5 implicit solvent appears too restricted in its exploration of dihedral angles. Specifically, all four dihedral angles identified as important for replicating the structure of polyT in solution (α, γ, ε, and ζ) appear to be restricted in the simulation compared to the desired target distribution. This is evident by the lack of dihedral angle states on the contour plots from the ff99-IGB5 simulation. This is problematic considering the GA algorithm indicates those dihedral angle states should be more frequent for polyT in solution. At a minimum, the relative frequency should be significant enough to appear on the simulation contour plots. Furthermore, the χ dihedral angle appears slightly restricted in the ff99-IGB5 MD simulation compared to the target distribution indicated by the GA. This suggests the χ dihedral angle should be further destabilized as well, potentially reducing base stacking.

The simulation of ff99 in IGB8 implicit solvent appears to mostly sample the same dihedral angle states as the ensemble of polyT structures selected by the GA. Any deviation in sample states

suffers the opposite problem than what was observed for the ff99-IGB5 simulation. For instance, the α-γ, α-ζ, and γ-ζ dihedral angle pairs all have additional states sampled in simulation compared to structures selected as the best-fit ensemble by the GA. This indicates the ff99 force field overly relaxed the α, γ, and ζ dihedral angles, but has represented the ε dihedral angle reasonably well.

Although dihedral angles are specifically parameterized in the ff99 force field, the difference between the ff99-IGB5 and ff99-IGB8 simulations arise from changes in the non-bonded interactions that are altered by the solvent representation and atomic radii. It is likely that the modified solvent representation allowed polyT to overcome dihedral angle energy barriers and transition to desirable states. Furthermore, this analysis highlights which dihedral angles are influential in replicating experimental structures of polyT in solution and how each simulation compares to the dihedral angles of polyT structures that match experiment best. Ultimately, this provides guidance on what force field combinations should be tested in the future. Namely, newer force fields should be tested with IGB8 implicit solvent. The ff99-IGB8 simulation overly relaxed the α, γ, and ζ dihedral angles. These dihedral angles have been targeted and re-parameterized in the bsc0, bsc1, and OL15 AMBER force fields. These newer force fields could outperform the ff99 force field in IGB8 implicit solvent producing a baseline for accurate MD simulations with ssDNA. The force fields should specifically hinder the transition of the dihedral angles seen in ff99, which would be in-line with desired results based on this study. Furthermore, the IGB8 solvent is the superior implicit solvent for simulations of other molecules such as proteins.

If the newer DNA force fields perform poorly in IGB8 implicit solvent, the ff99 force field can serve as the baseline for the parameterization of the first AMBER force field specific to ssDNA. Specifically, re-parameterization could target the energy required for the α, γ, and ζ dihedral angles to transition into certain states. A more complicated route could target the

parameterization of a new implicit solvent or new set of atomic radii. Given that the ff99-IGB5 simulation appears to restrict the α, γ, ε, and ζ dihedral angles and the ff99-IGB8 simulation appears to overly relax the α, γ, and ζ dihedral angles, it is reasonable to assume a more accurate parameterization of the solvent and atomic radii lie between the two existing force fields.

5.1.3.2 Explicit Solvent

The structure of polyT is compared between simulation and the GA's best-fit ensembles through analysis of polyT's R_g and R_{ee}. Figure 5.1.6 shows the three DNA force fields tested in explicit TIP3P solvent where (a) is the bsc0 force field, (b) is the bsc1 force field, and (c) is the OL15 force field. Experimental R_g and R_{ee} values are plotted for reference, where the black dashed line represents the experimental mean and shaded green bars indicate experimental error from SAXS and FRET. On the left side of Figure 5.1.6, the contour plot represents the MD simulation where increased transparency indicates reduced relative frequency of occurrence in simulation. The best-fit ensemble of polyT structures selected from the GA are shown as red circles, where the size is positively correlated with the weight assigned to that gene in the GA. The larger the red circle, the more weight and contribution that structure has to the final scattering intensity of the ensemble. In the margins, density plots are provided for a comparison of how the conformations in simulation compare to the conformations selected by the GA. On the right side of the structure, the temporal profile of polyT's R_g is provided by a solid line. The polyT structures selected by the GA are again represented by red circles where weight and size are correlated.

Figure 5.1.6 (a) represents the bsc0 force field for DNA in TIP3P solvent. Looking at the left side of the figure, the presence of large red circles over the dark, frequently explored areas of the contour plot indicates that those structures chosen by the GA are well represented by the dynamics

carried out in simulation. However, the presence of heavily weighted red circles that fall outside

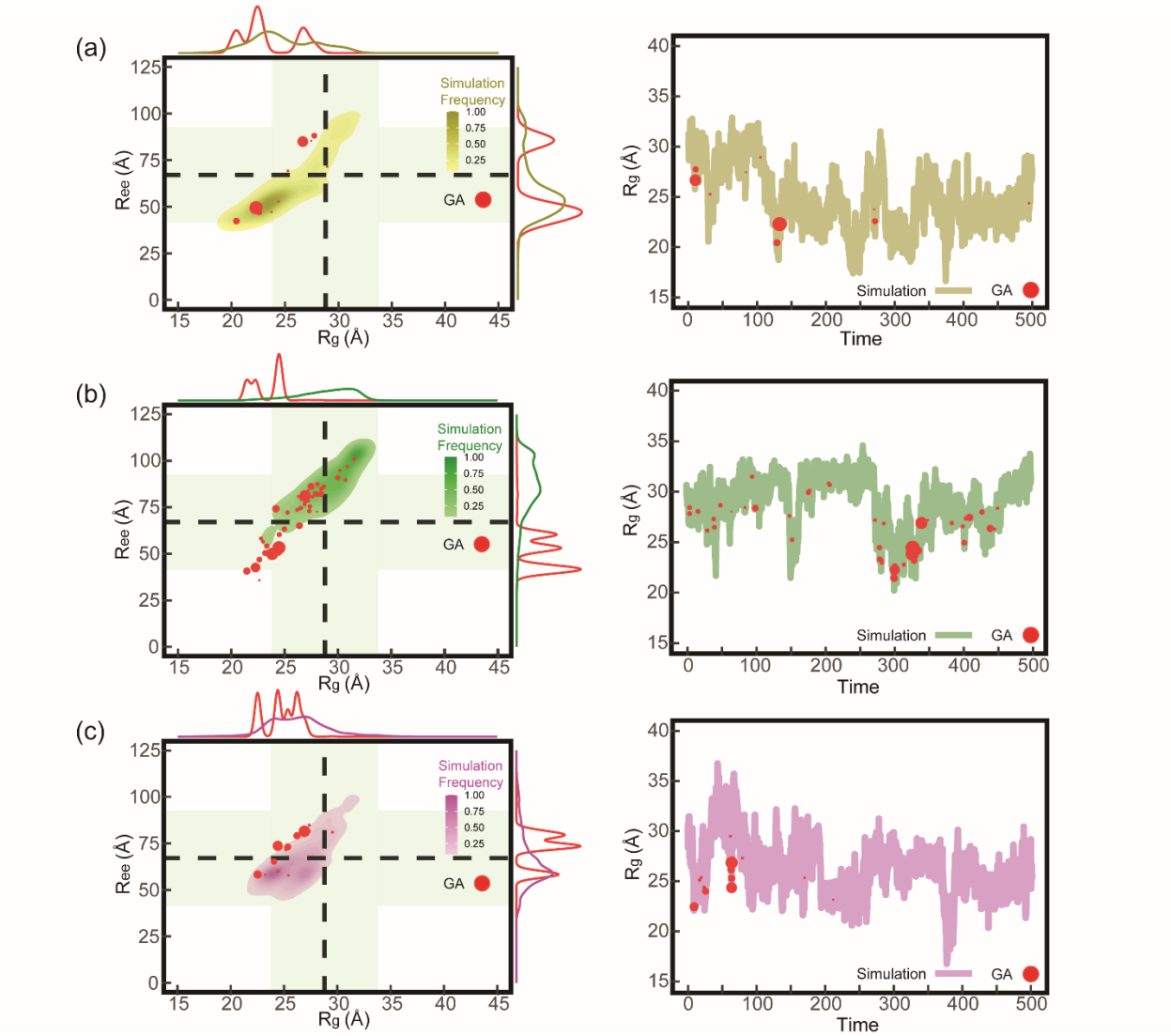


Figure 5.1.6. Shows a comparison between the Rg and Ree calculated from MD simulation (bivariate contour plot) where transparency is correlated with relative frequency of sampling and the GA (red circles) where size is correlated with weight. Density plots are in the margin to compare conformational distributions between simulation and the GA. Experimental values and error are depicted by the black dashed line and light green rectangles. The temporal profile of Rg (right) show which part of the MD simulations the GA was choosing structures from. The simulations represented are (a) bsc0 with tip3p (b) bsc1 with tip3p and (c) OL15 with tip3p.

the shaded regions of the contour plot indicates that the simulation has neglected to explore a conformation of polyT that the GA has deemed significant compared to other conformations of polyT in the simulation. The right side of the figure provides some guidance as to why these

conformations are poorly explored. In general, the GA selects structures from the early part of the simulation, which contains conformations of polyT that are not converged. The R_g of polyT is higher in these earlier parts of the simulation due to a more elongated structure, closer to its starting point in the B-DNA conformation. Thus, it appears that the bsc0 produces overly compact structures of polyT compared to the array of polyT structures in the experimental ensemble. This is indicative that the bsc0 force field in TIP3P solvent should not be used in future studies.

Figure 5.1.6 (b) shows the conformational comparison between the bsc1 force field for DNA in TIP3P solvent and polyT selected by the GA. Ultimately, the contour plot shows that the simulation tends to accurately match the experimental R_g but tends to overestimate the R_{ee}. The GA selects structures that vary from the simulation in this regard, as the selected polyT conformations are well dispersed across the expected experimental R_g and R_{ee} values. The deviation of GA structures from the frequently explored structures in simulation indicates the simulation does not capture the structure and dynamics of polyT in solution, however, the temporal profile on the right side of Figure 5.1.6 (b) indicates this could be a convergence issue for the simulation. Towards the latter half of the simulation, the polyT structure undergoes significant fluctuations in its R_g. In this region, there are heavily weighted structures of polyT selected by the GA which coincide with structures that fall outside of the shaded region of the contour plot. As the simulation is extended, the simulation may better represent experiment and deviate from the overly elongated structures of polyT. This is especially relevant for analysis of explicit solvent simulations as the explicit representation of solvent reduces conformational exploration compared to implicit solvent simulations. Thus, the 500ns of simulation may be insufficient for polyT to sample a representative ensemble of what would be found experimentally. Further study should

be performed on the bsc1 force field in TIP3P solvent. It is possible that the fit between simulation and the GA structures improves and that the GA structures improve in their fit to experiment.

Figure 5.1.6 (c) shows the OL15 force field for DNA in TIP3P solvent. The left side of the figure indicates that the simulation and structures selected by the GA reasonably agree with the experimentally expected R_g and R_{ee} values. The conformations of polyT selected by the GA mostly overlay infrequently sampled conformations in the simulation, however, the points still fall within the shaded area of the contour plot. As with the other two tested force fields in explicit solvent, this indicates that there is some mismatch between the structure and chain dynamics captured in simulation and the desired structure of polyT that the GA selects as the best matches to experiment. The biggest concern for the OL15 force field arises when examining the temporal profile on the right side of Figure 5.1.6 (c). All structures selected from the GA occur early in the simulation, indicating that as the OL15 force field moves closer to equilibrium, the representation of polyT is getting worse. This is in direct contrast with the conclusions made about the bsc1 force field shown in Figure 5.1.6 (b). Thus, the bsc1 force field appears to be the most promising for accurately representing ssDNA given a sufficient simulation time.

Overall, Figure 5.1.6 indicates that the bsc0 and OL15 force field will not improve their fit to experiment as the simulation is continued and performance is likely worse than the ensemble from the GA indicates. Considering the respective χ^2 fits of 5.84 and 2.76 reflect a poor fit to experiment, they are not suggested for MD studies of ssDNA. The bsc1 force field has an average χ^2 fit of 3.74 for the GA selected ensemble of structures. The analysis of Figure 5.1.6 (b) establishes that the simulation represents the quality dictated by this metric as the structures of the GA come from the more converged part of the simulation. Despite this being a poor fit to experiment, the potential to further improve as the simulation is extended makes it the suggested force field for use. However,

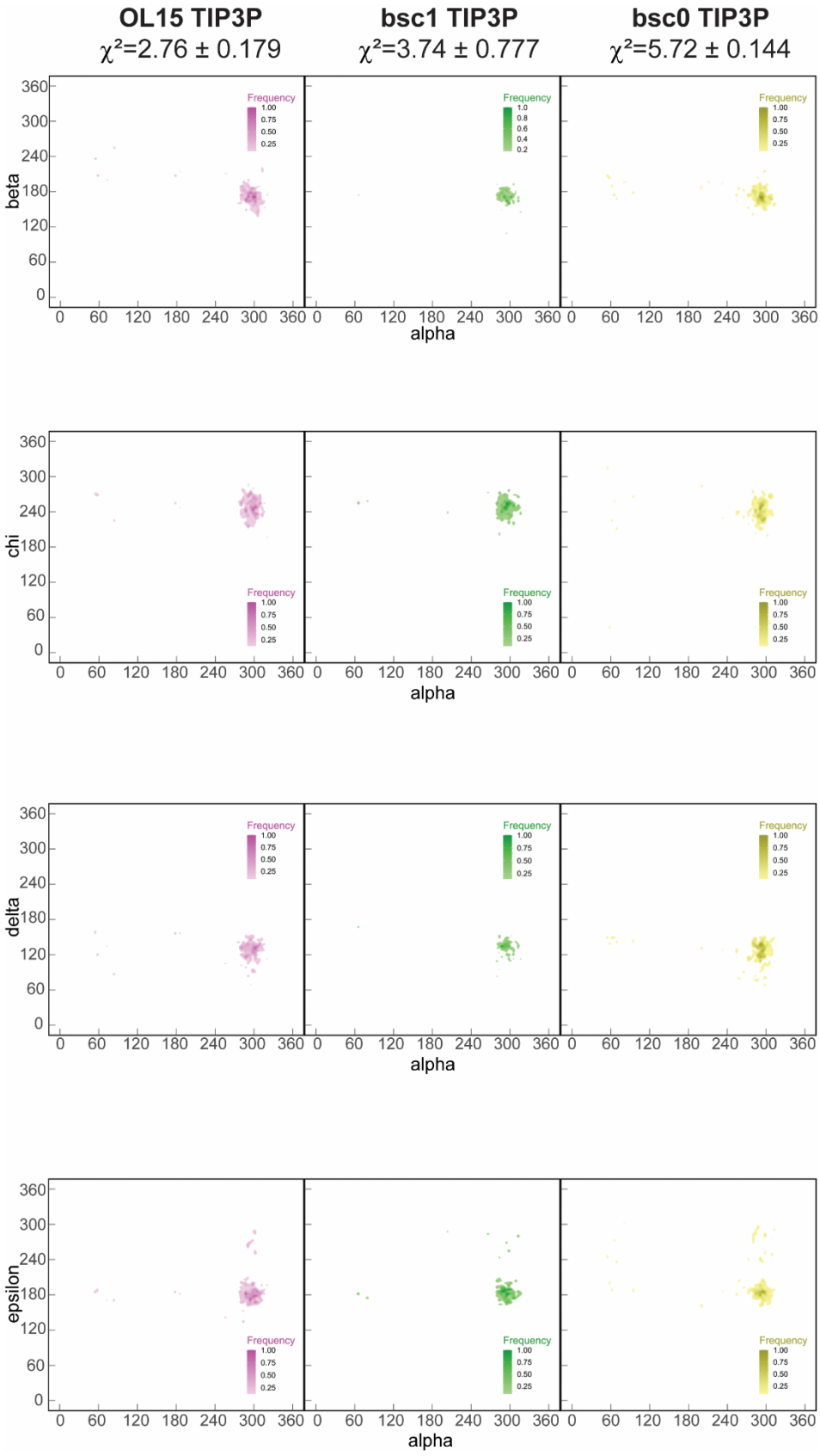
further study is required to determine the necessary simulation time required for this simulation to sufficiently converge and to verify if the results improve as the simulation is continued.

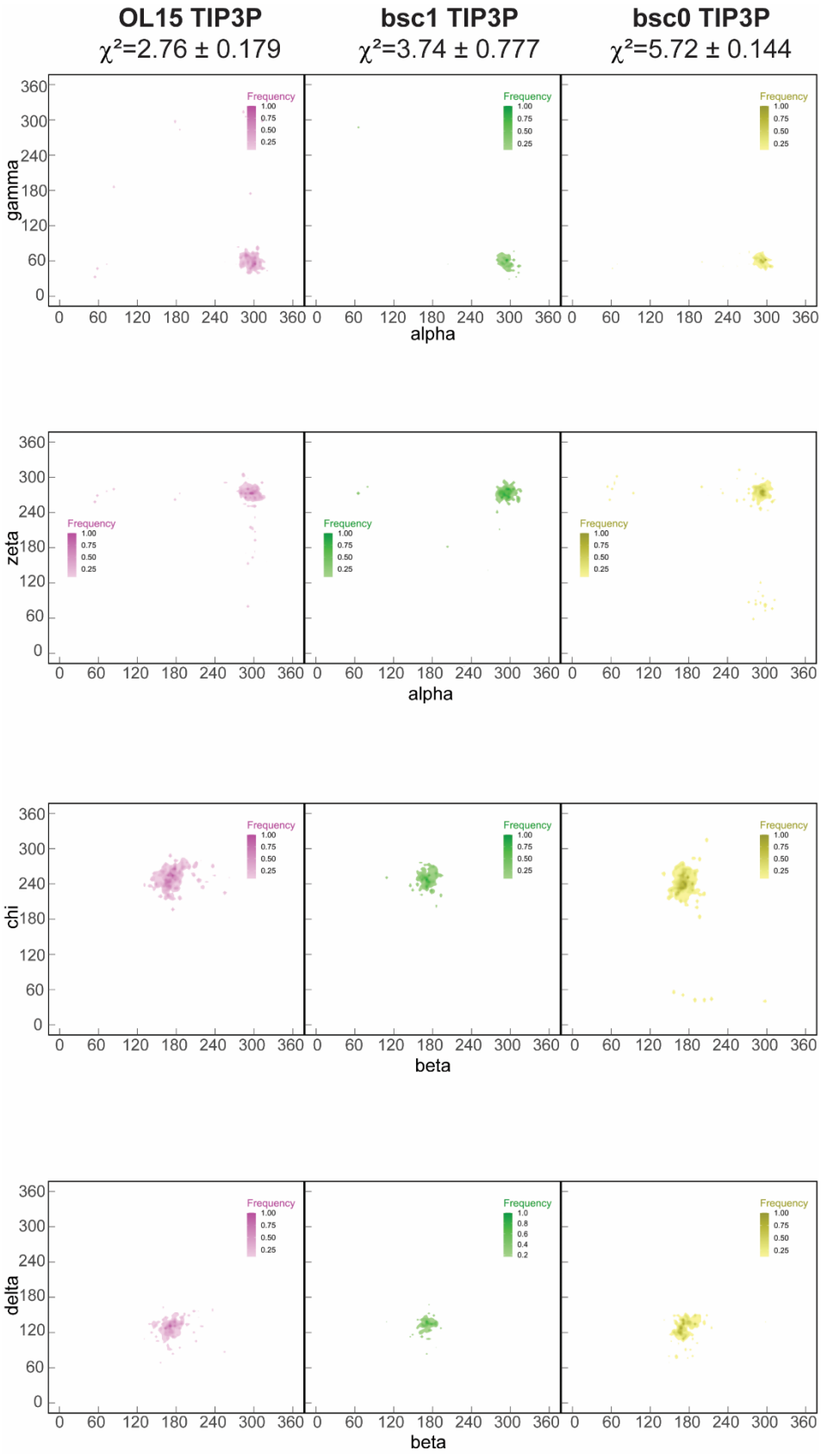
Table 5.1.3. Shows the goodness of fit for the ensemble of polyT structures chosen by the GA for each explicit solvent simulation

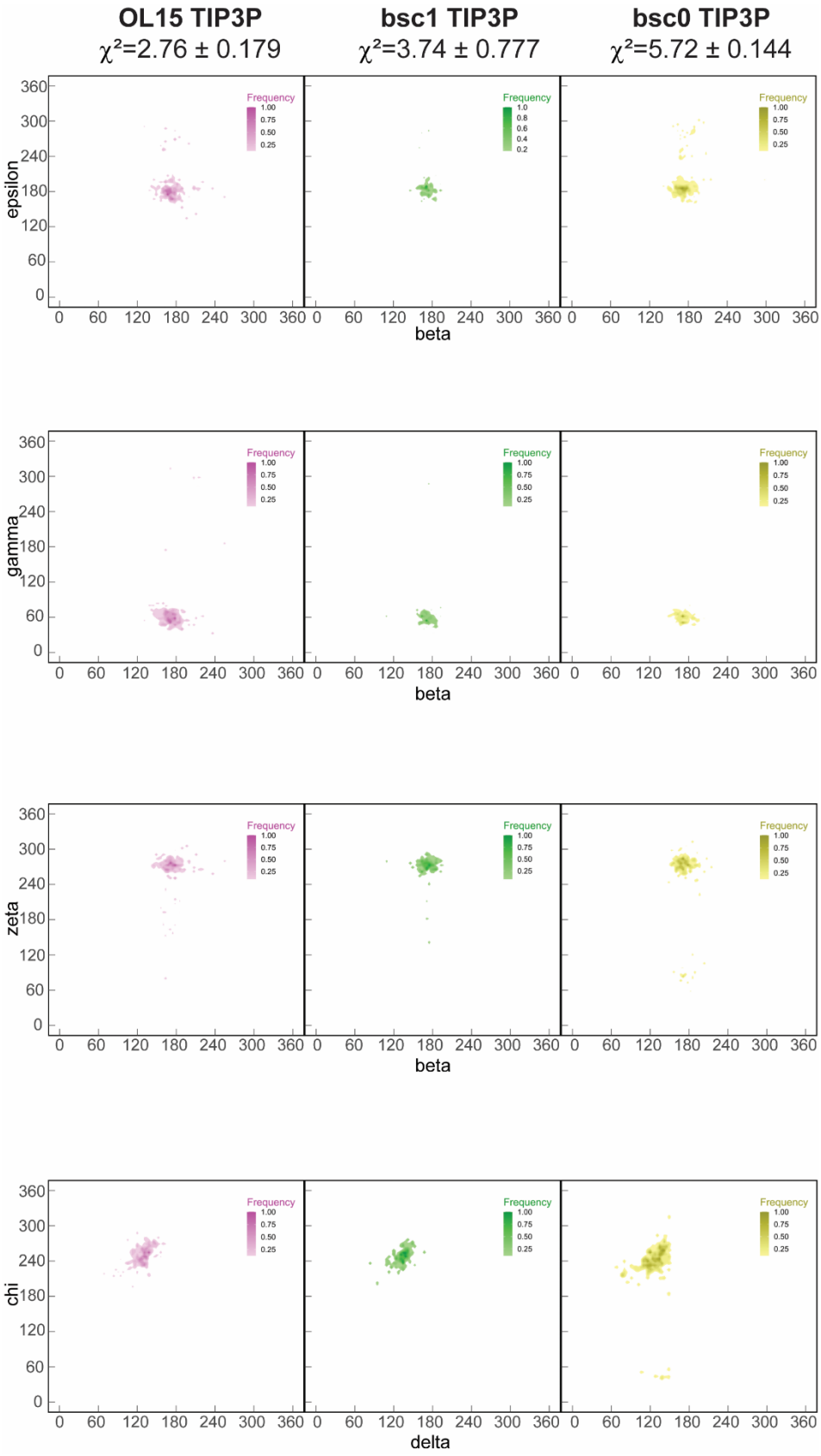
Goodness of fit (χ^2)		
bsc0 TIP3P	bsc1 TIP3P	OL15 TIP3P
5.84 ± 0.144	3.74 ± 0.777	2.76 ± 0.179

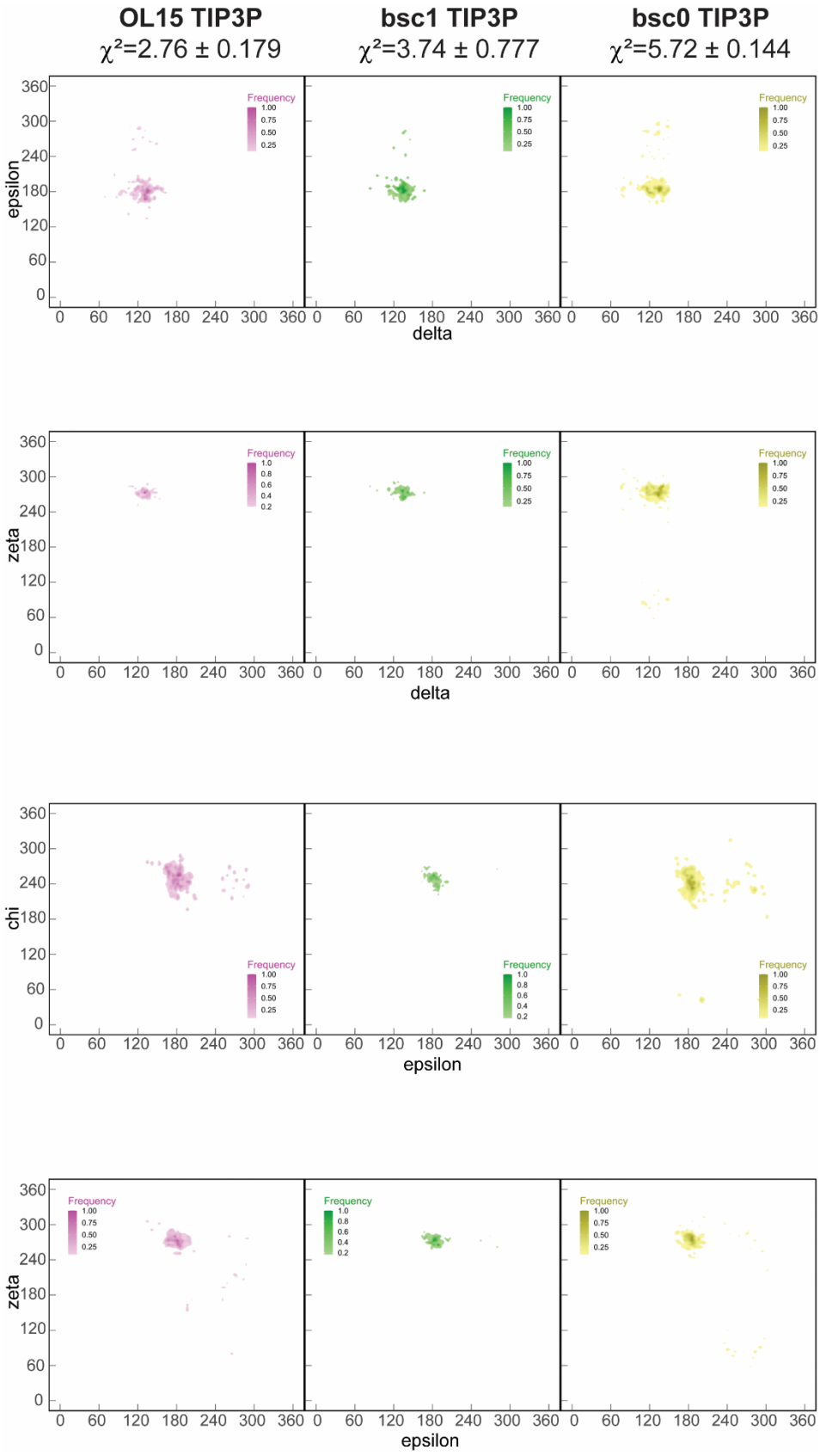
As with the implicit solvent simulations, additional insight on the structure of polyT in solution and the performance of each MD simulation is achieved by analyzing the dihedral angles of the ssDNA in each of the best-fit ensembles. Specifically, the α, β, γ, δ, ε, ζ, and χ dihedral angles are examined. Their placement in the ssDNA chain is illustrated in Figure 5.1.3. The explored space for each dihedral angle is plotted against one another as a bivariate contour plot in Figure 5.1.7. It should be noted that Figure 5.1.7 corresponds only to the ssDNA structures selected by the GA, which is represented by the red circles in Figure 5.1.6. As discussed, the structures chosen by the GA in explicit solvent line up poorly with the frequently explored conformations in simulation, thus the frequency and distribution of the dihedral angles throughout the entire MD simulation likely varies from the ensemble of structures represented in Figure 5.1.7.

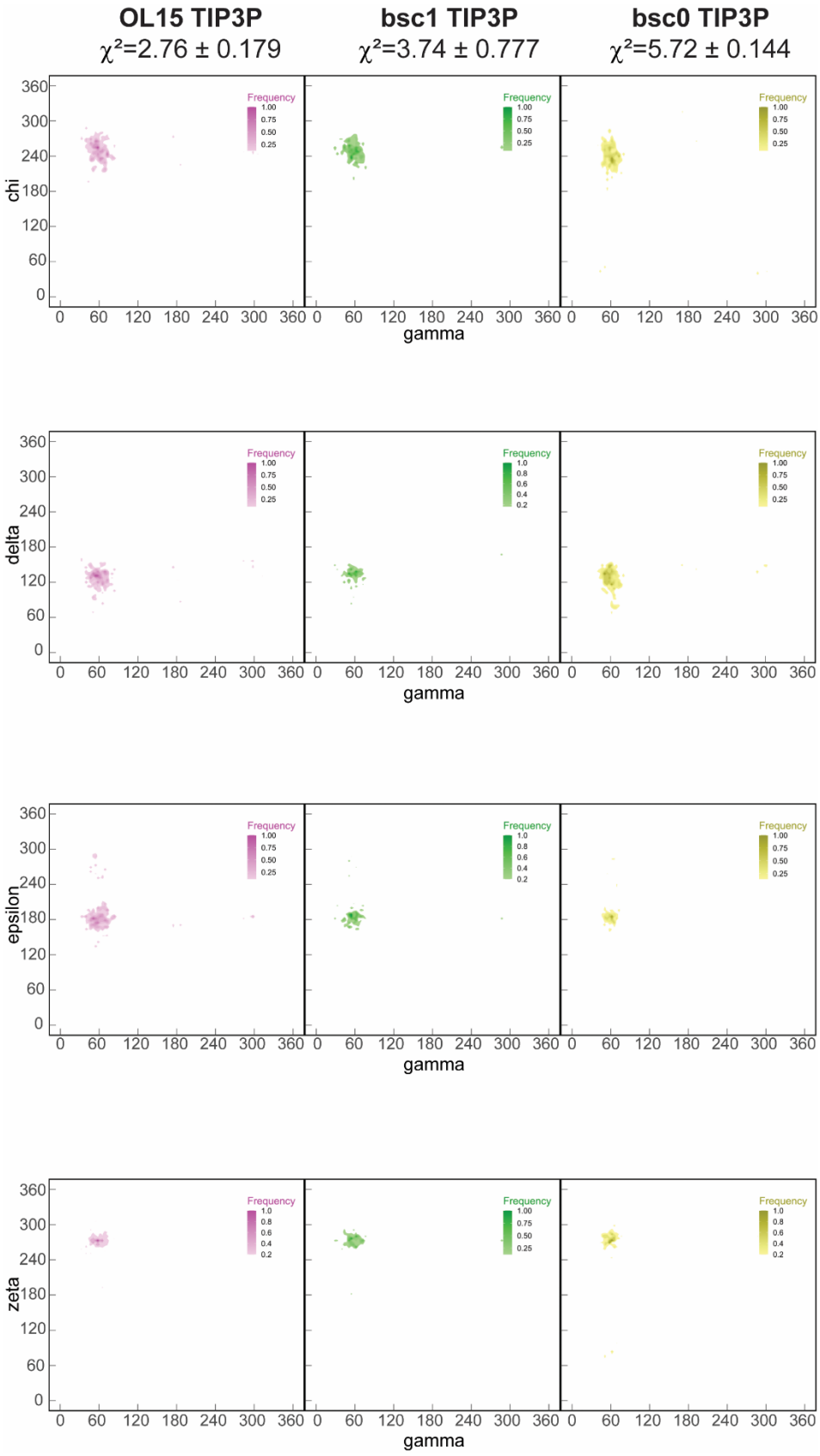
Figure 5.1.7. Shows bivariate contour plots for the alpha, beta, chi, delta, epsilon, gamma, and zeta dihedral angles. The dihedral angles are calculated from the ensemble of structures chosen by the GA. The GA was applied to select structures from MD simulations with the bsc0, bsc1, and OL15 force field in TIP3P solvent. The goodness of fit (χ^2) for the ensemble of structures chosen by the GA is provided for each simulation. The lower the value, the better the fit to experiment.


OL15 TIP3P
χ²=2.76 ± 0.179
bsc1 TIP3P
χ²=3.74 ± 0.777
bsc0 TIP3P
χ²=5.72 ± 0.144
beta
chi
delta
epsilon
alpha
Frequency


OL15 TIP3P
χ²=2.76 ± 0.179
bsc1 TIP3P
χ²=3.74 ± 0.777
bsc0 TIP3P
χ²=5.72 ± 0.144
Frequency
gamma
zeta
chi
delta
alpha
beta
0
60
120
180
240
300
360


OL15 TIP3P
χ²=2.76 ± 0.179
bsc1 TIP3P
χ²=3.74 ± 0.777
bsc0 TIP3P
χ²=5.72 ± 0.144
epsilon
gamma
zeta
chi
beta
delta
Frequency


OL15 TIP3P
χ²=2.76 ± 0.179
bsc1 TIP3P
χ²=3.74 ± 0.777
bsc0 TIP3P
χ²=5.72 ± 0.144
epsilon
delta
zeta
delta
chi
epsilon
zeta
epsilon
Frequency


OL15 TIP3P
χ²=2.76 ± 0.179
bsc1 TIP3P
χ²=3.74 ± 0.777
bsc0 TIP3P
χ²=5.72 ± 0.144
chi
delta
epsilon
zeta
gamma
Frequency

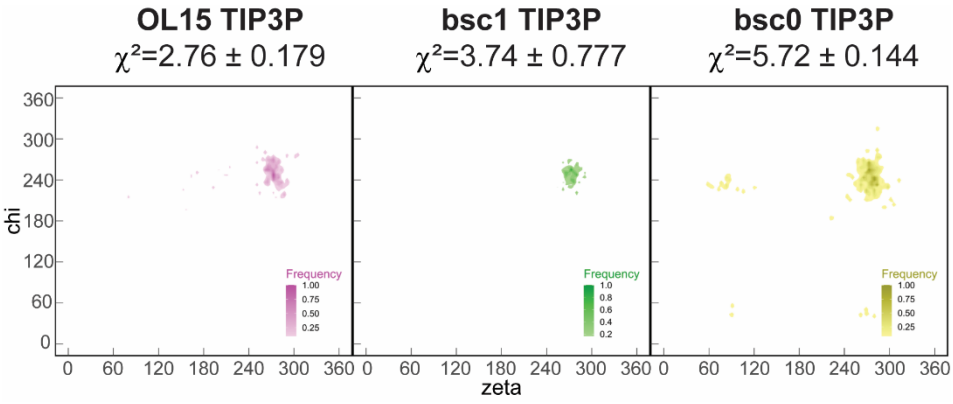

OL15 TIP3P
χ²=2.76 ± 0.179
bsc1 TIP3P
χ²=3.74 ± 0.777
bsc0 TIP3P
χ²=5.72 ± 0.144
chi
zeta
0
60
120
180
240
300
360
Frequency
1.00
0.75
0.50
0.25
1.0
0.8
0.6
0.4
0.2

Figure 5.1.7 does not show a clear and discernable difference between specific dihedral angle pairs across any of the simulations in TIP3P solvent. Compared to the implicit solvent simulations, the dihedral angles are poorly explored. This is likely the reason that the best-fit ensembles in the last generation of the GA poorly represent experiment for the bsc0, bsc1, and OL15 simulations.

The similarity between dihedral angle distributions, but variation in goodness of fit does provide unique insight. Specifically, this provides support for the previously mentioned multi-dimensionality of this problem. Specifically, the dihedral angles of the backbone are all correlated to one another to form the overall structure of ssDNA, thus the 2D representation in Figure 5.1.7 is insufficient to capture the entire picture. This indicates that further study is required. One possible option would be to break the polyT chains down into smaller building blocks such as dimers or trimers. A clustering algorithm could then be applied to identify the sets of dihedral angles present in each building block unit and do so in the high dimensional framework this problem requires. Once the unique set of building blocks has been identified, the frequency and location of each block can be studied and compared to one another.

Another potential issue with the current analysis of polyT in TIP3P is the potential lack of convergence. As discussed in Figure 5.1.6, the bsc0 and OL15 simulations are unlikely to improve as the simulation continues. However, the bsc1 force field appears to be improving towards the end of the simulation. Additional simulation time could lead to the further exploration of dihedral angles. Eventually, this simulation may reveal the 2D dependence of dihedral angles captured in implicit solvent (Figure 5.1.4). This is further supported by Figure S5.1.4 which shows that as the implicit solvent simulations continued, the prominence of the varied dihedral angel states became more prominent. The increase in relative frequency of new states would be delayed in explicit solvent compared to implicit solvent, thus the explicit solvent simulations should be run longer

than implicit solvent simulations. Further study of bsc1 in TIP3P solvent is required to test this theory. If the dihedral angle transitions do not occur, it is possible that the TIP3P solvent is over stabilizing the dihedral angles. In this instance, other explicit solvents should be explored for use with AMBER force fields.

5.1.3.3 Structural Analysis

The best ensemble of polyT structures is obtained from applying the GA to the ff99-IGB5 simulation. Figure 5.1.8 shows a single best fit chromosome in the final generation of the GA.

The individual components in Figure 5.1.8 (b) sum to fit the experimental intensity profile very well as indicated by the χ^2 fit (1.15). This indicates that the chosen ensemble of polyT matches the size and shape of polyT found in solution experimentally. At first glance, the empirically derived R_g and R_{ee} values obtained for the ensemble appear to fall below the experimentally obtained values of R_g and R_{ee} which are 28.8Å and 67.1Å, respectively. However, it is important to note that the experimental R_g is extracted from the Guinier region of the SAXS intensity profile which includes portions of the hydration shell and the R_{ee} values are obtained from fluorescent molecules which are attached outside of the 3' and 5' ends of the ssDNA. This not only adds length to the chain but could also influence dynamics. Thus, it is reasonable to believe that the true R_g and R_{ee} fall closer to the empirically derived values of 27.04Å and 60.88Å, respectively. However, these values are still dependent on the quality of the SAXS intensity profile used in the fitness function of the GA.

Although the best fit ensemble comes from the ff99-IGB5 simulation, it has been established the ff99-IGB8 simulation appears to better capture the structure and dynamics of polyT. Therefore, the expected values and representative structures from the best-fit ensemble in the ff99-IGB8

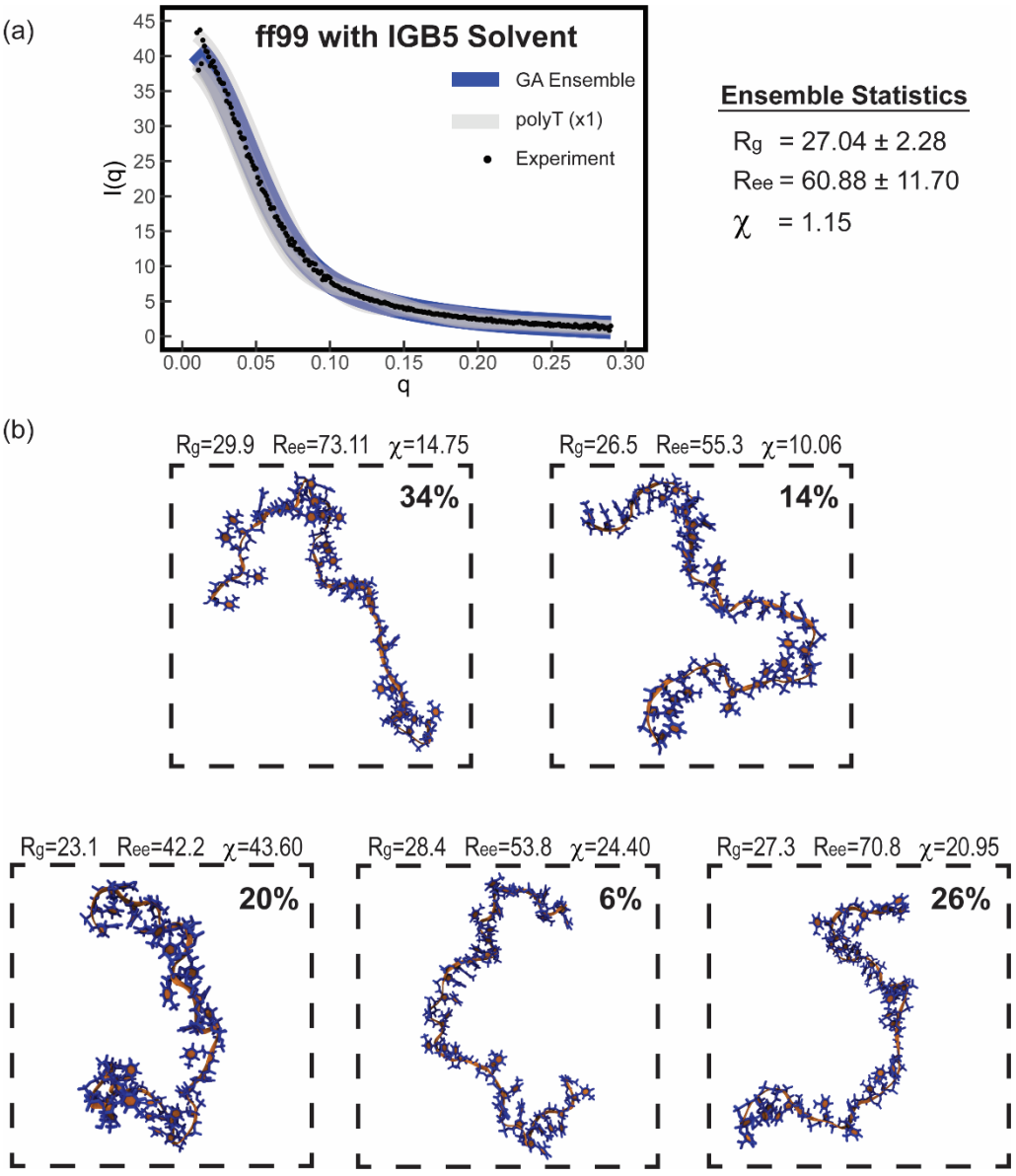


Figure 5.1.8. (a) Shows the SAXS intensity for the best-fit chromosome selected from the ff99-IGB5 simulation. The SAXS plot shows the ensemble intensity, individual polyT structure intensities, and the experimental intensity. Ensemble statistics are provided to the right. (b) Shows individual polyT structures and their corresponding statistics. Only structures with a significant weight (>1%) are shown.

simulation are provided in Figure 5.1.9. Figure 5.1.9 (a) shows how the scattering intensity of the GA selected ensemble compares to the experimental ensemble and provides empirically derived ensemble statistics. Figure 5.1.9 (b) shows the individual polyT structures that contribute to the ensemble.

The individual components in Figure 5.1.9 (b) sum to fit the experimental intensity profile with minimal errors as indicated by the χ^2 fit (1.73). Looking at the low q range of the SAXS plot in Figure 5.1.9 (a), there is some deviation from experimental values. This region directly corresponds to the size and specifically R_g of structures in the ensemble. Comparing the empirically derived values of R_g and R_{ee} in this ensemble to the empirically derived values obtained from the ff99-IGB5 ensemble (Fig 5.1.8), it is evident the polyT has become more compact. This supports the previous hypothesis stemming from Figure 5.1.5 that states the ff99-IGB8 simulation has overly relaxed the dihedral angles of polyT, influencing its structure in a negative manner. Ultimately, it appears that one affect is the sampling of structures that are too collapsed compared to experimental observations.

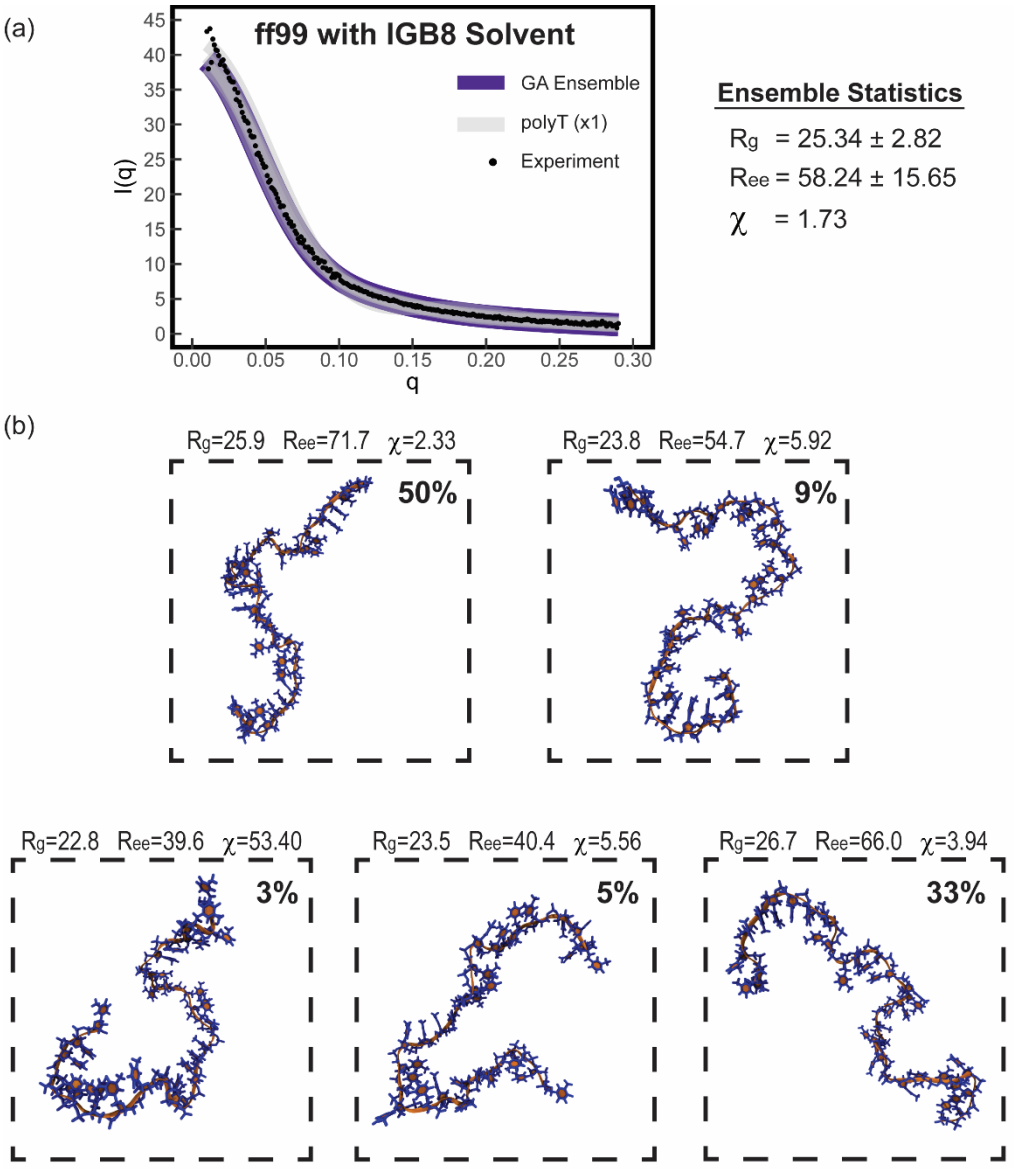


Figure 5.1.9. (a) Shows the SAXS intensity for the best-fit chromosome selected from the ff99-IGB8 simulation. The SAXS plot shows the ensemble intensity, individual polyT structure intensities, and the experimental intensity. Ensemble statistics are provided to the right. (b) Shows individual polyT structures and their corresponding statistics. Only structures with a significant weight (>1%) are shown.

Figure 5.1.10 shows the comparison of SAXS intensities for experiment, the bsc1-TIP3P ensemble, and individual contributing structures of polyT. As with the ff99-IGB8 ensemble (Fig 5.1.9), the ensemble fit selected by the GA clearly deviates from the experimental intensity profile in the low q range (Fig 5.1.10 (a)). The ensemble statistics again indicate that the empirically

derived R_g value is lower than the best-fit ensemble for the ff99-IGB5 simulation (Fig 5.1.8), which serves as the best representation of experiment. The pathway for which this deviation occurs from experiment in the bsc1-TIP3P ensemble is unlike the pathway for the ff99-IGB8 simulation, which was hypothesized to be from overly relaxed dihedral angles. In fact, the exact opposite trend was observed for dihedral angles in the bsc1-TIP3P simulation, and the dihedral angles were not well explored (Fig 5.1.7).

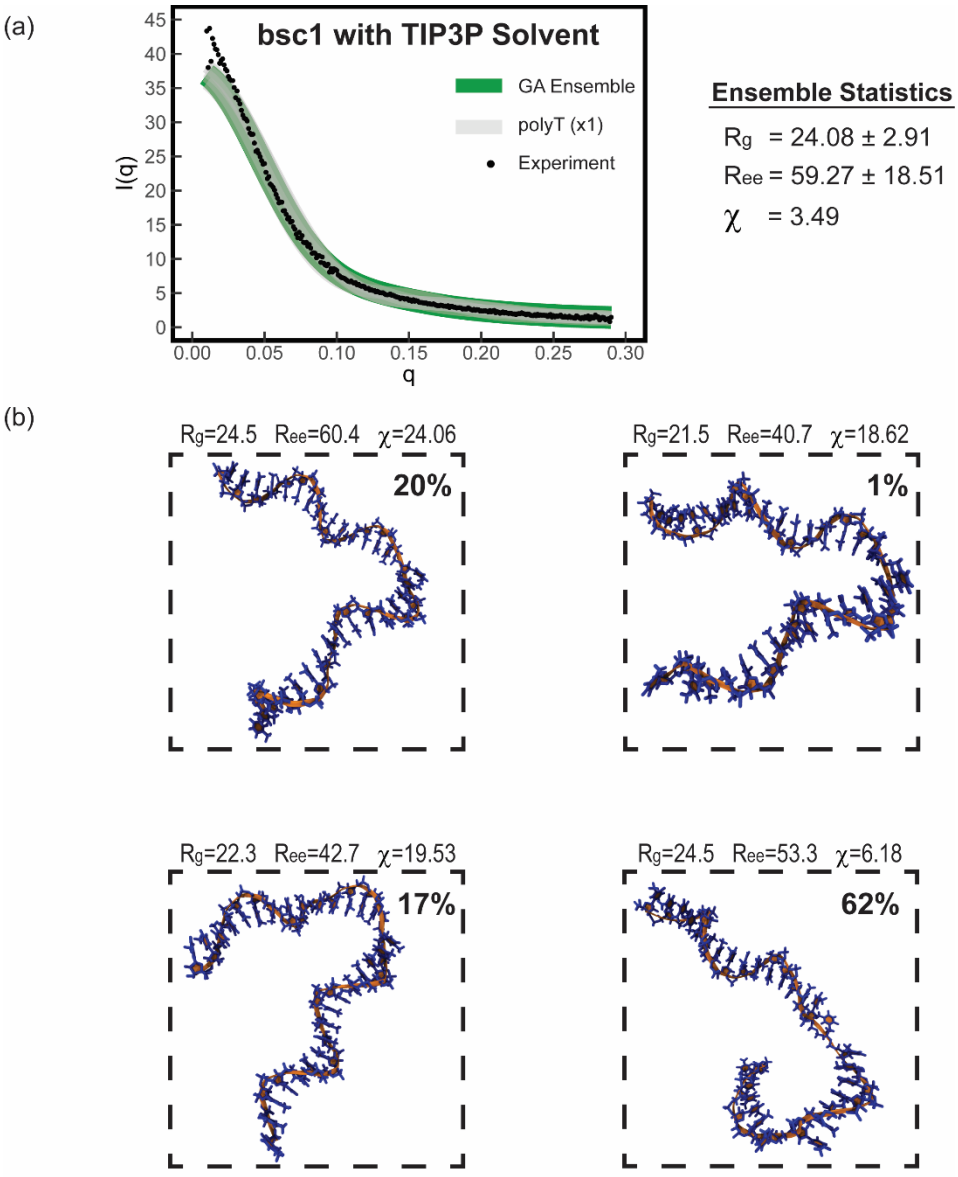


Figure 5.1.10. (a) Shows the SAXS intensity for the best-fit chromosome selected from the bsc1-TIP3P simulation. The SAXS plot shows the ensemble intensity, individual polyT structure intensities, and the experimental intensity. Ensemble statistics are provided to the right. (b) Shows individual polyT structures and their corresponding statistics. Only structures with a significant weight (>1%) are shown.

Figure 5.1.10 (b) helps to visualize the pathway of deviation for the bsc1-TIP3P ensemble, although remains highly qualitative. Specifically, the polyT structures from the bsc1-TIP3P simulation have one significant point in the chain where the structure deviates. This is characterized by a high degree turn, while the polyT chain on both sides of it remains straight and

rigid with a helical twist. This behavior of polyT is not seen in the ff99-IGB5 or ff99-IGB8 simulation (Fig 5.1.8 and Fig 5.1.9 (b)). In those simulations, the helical structure of polyT breaks down and the chain behaves more like a directional random walk with many turns and kinks. The structure of polyT shown in Figure 5.1.10 (b) illustrates why the bsc1-TIP3P dihedral angle contour plots (Fig 5.1.7) indicate poor exploration of dihedral angles, while the statistical ensemble appears overly compact. The bulk of the polyT monomers retain the dihedral angles characteristic of the B-DNA form they started in, while a select, small group of monomers deviate significantly. Thus, in the scheme of all dihedral angles, this deviation is highly infrequent and would not be present based on how the dihedral angles are depicted in analysis (Fig 5.1.7). This behavior is not a good representation of what is expected of polyT in solution. The poor χ^2 fit of 3.49 supports that this ensemble has significant errors in depicting an accurate ensemble of polyT.

5.1.4 Conclusion

Ultimately, the GA applied to the ff99 force field in IGB5 implicit solvent provided the most accurate ensemble of structures compared to experiment. However, the ff99 force field in IGB8 implicit solvent and bsc1 force field in TIP3P solvent are the current recommendations for use in MD studies of ssDNA. This is due to these simulations ability to frequently explore conformations similar to the structures of polyT selected by the GA, indicating the ssDNA chain dynamics are better represented in these simulations. It should be noted however, that the structure of polyT is poorly represented in explicit TIP3P solvent.

The α, γ, ε, and ζ dihedral angles were identified as influential parameters in replicating the structure of polyT in solution. These dihedral angles were identified from the analysis of implicit solvent simulations. However, analysis of the explicit solvent simulations identified the current 2D analysis performed in this study is insufficient to capture the interdependence of all dihedral

angles. This is due to the dihedral angles of ssDNA being interrelated to one another in a dimensionality greater than the 2D analysis performed.

Moving forward, newer force fields of DNA should be tested with the IGB8 implicit solvent simulation as they could yield a better match to experiment than the ff99 force field. Multi-dimensional analysis of the dihedral angles could be performed on ssDNA by breaking the chain into building blocks and using a clustering algorithm. This could ultimately be used to identify the unique sets of dihedral angles present in the building blocks of polyT and the frequency with which they appear.

This studies guidance on the current usage of DNA force fields on ssDNA is not without limitations. Namely, the results reported in this study can change as the nucleotide sequence varies from polyT. Furthermore, all simulations started with polyT being in the canonical B-DNA conformation. Varying the starting point of the simulation can alter simulation performance and results. Lastly, the best-fit ensembles selected by the GA implemented in this study are highly dependent on the quality of the experimental SAXS intensity profile used for comparison.

5.1.5 Supporting Information

5.1.5.1 Genetic Algorithm (GA) Hyperparameter Tuning

The GA used in this study was tuned with respect to the number of genes per chromosome, the mutation rate, and number of generations per run. These parameters were tuned with ssDNA structures obtained from a simulation using the ff99 force field and IGB5 implicit solvent. Figure S5.1.1 shows the minimum χ^2 value for each GA run, which varies only in how many genes are grouped into a single chromosome. The minimum ensemble that achieves the lowest χ^2 value is selected as the appropriate number of genes per chromosome. Ultimately, this results in 7 genes per chromosome being selected as the appropriate parameter. Generally, previous works test 1-5

structures per ensemble, however, because ssDNA is a complexly flexible molecule, the number of genes per chromosome was extended beyond that.

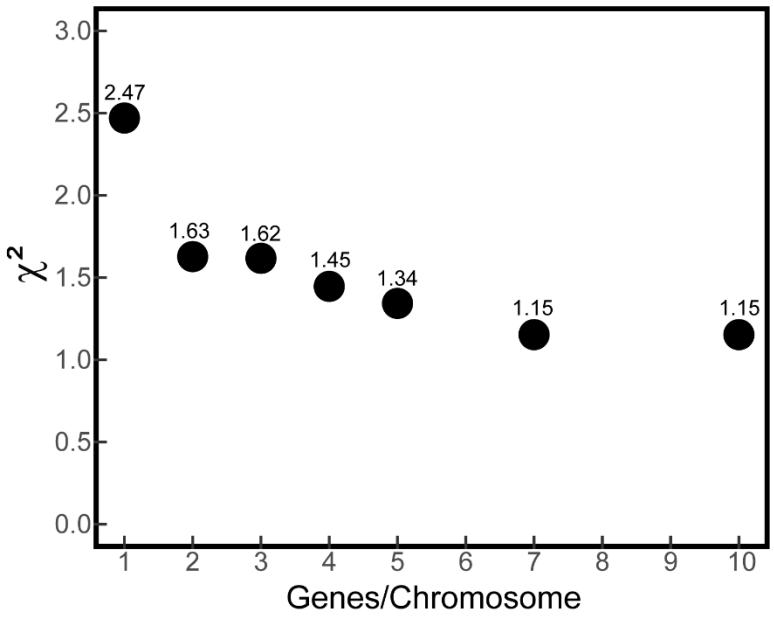


Figure S5.1.1. Shows the minimum χ^2 value achieved for each GA run. This data is for simulation structures modeled by the ff99 force field in IGB5 implicit solvent.

Figure S5.1.2 shows the how mutation rate affects the convergence rate of the GA as well as the overall result with the minimum χ^2 value. The mutation rate was varied from probability as low as 0.1 and as high as 0.4. Overall, it is seen that the mutation rate does not have a significant effect on the convergence or overall χ^2 values obtained in our GA with 7 genes per chromosome. Overall, a mutation rate of 0.3 was chosen because it was the initial default value that was used. The indifference in selecting the mutation rate is likely tied to the implementation of the GA. In our GA, we maximize exploration by mutating any duplicate chromosomes in a single generation. Thus, the mutation rate is likely higher than all the tested mutation rates selected. Although convergence is achieved with all the tested mutation rates, a second GA optimization was

performed on the ff99 simulation with IGB5 implicit solvent. Figure S5.1.3 shows the second GA

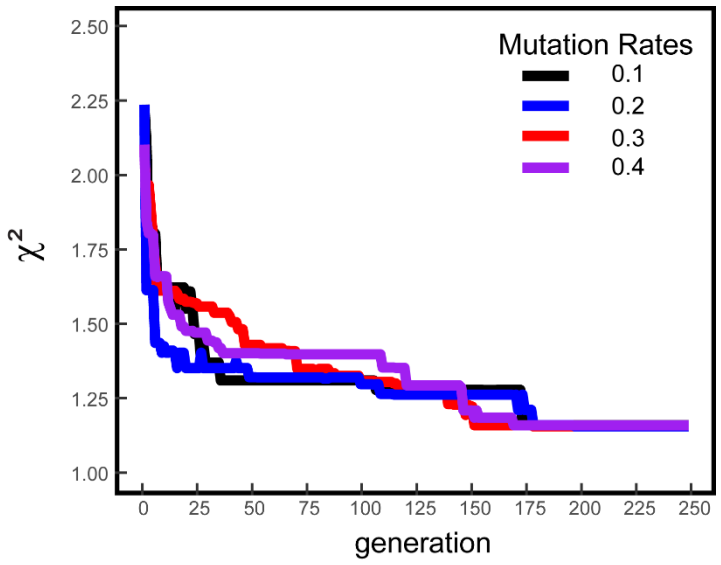


Figure S5.1.2. Shows the minimum χ^2 value in each generation for 4 independent GA optimizations. Each optimization has 7 genes per chromosome. Each run varies in mutation rate and uses simulation structures modeled by the ff99 force field in IGB5 implicit solvent.

optimization results and clearly demonstrates the optimization process is complete within 250 generation of our GA.

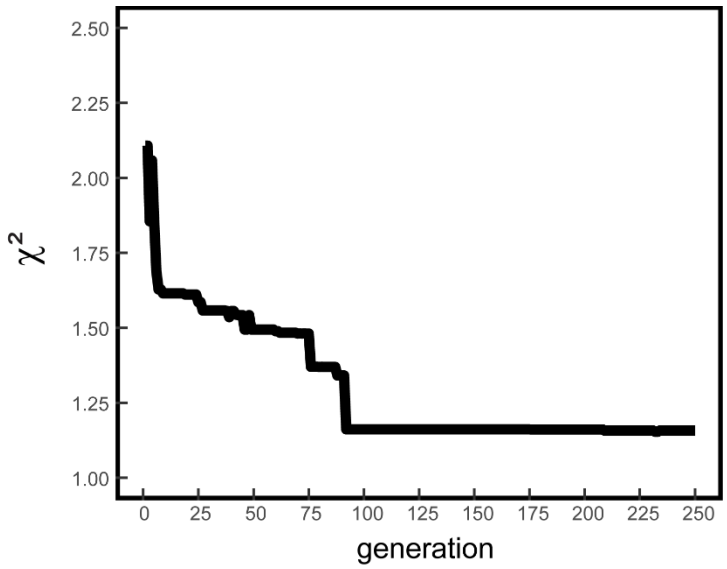
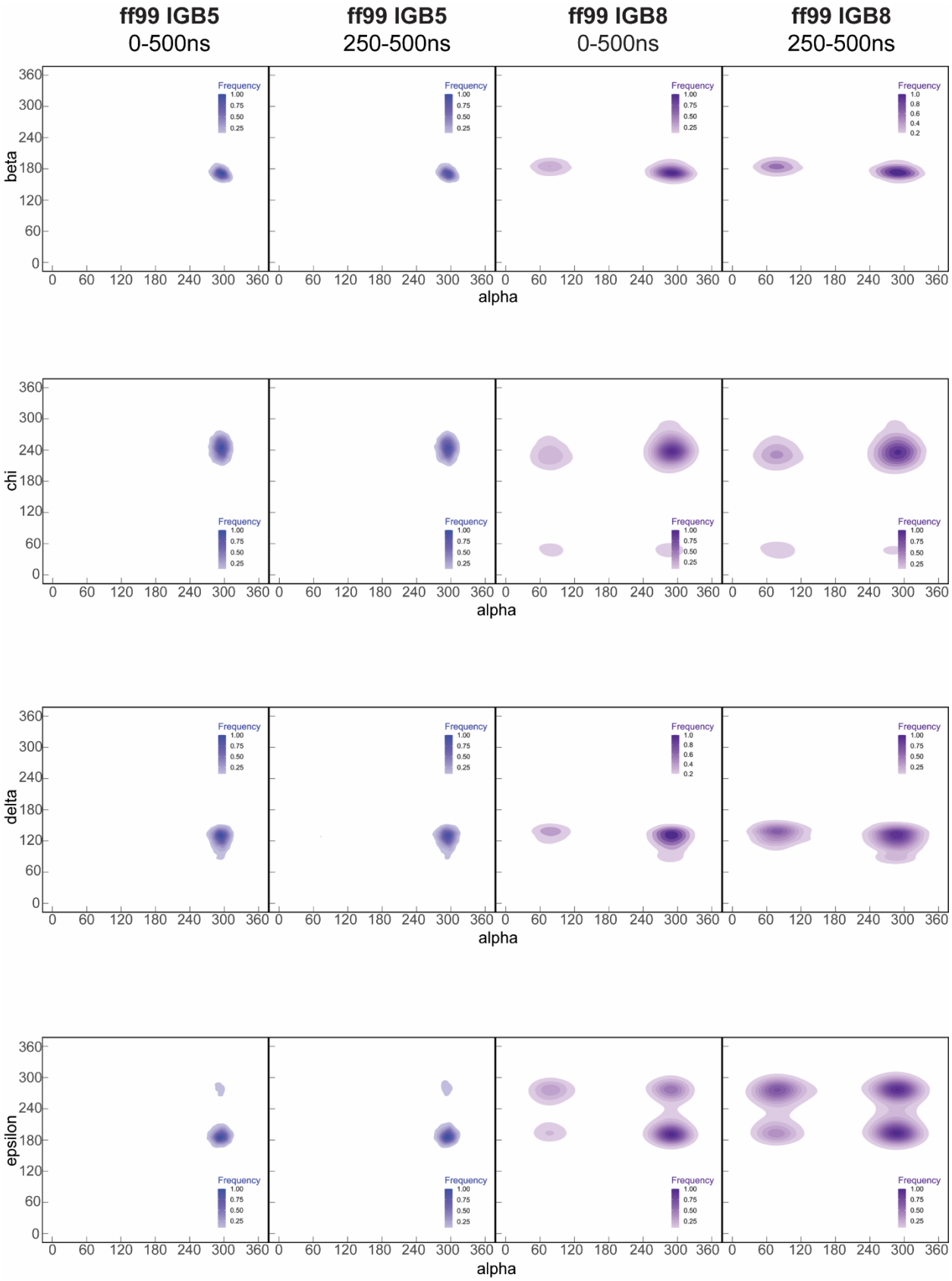
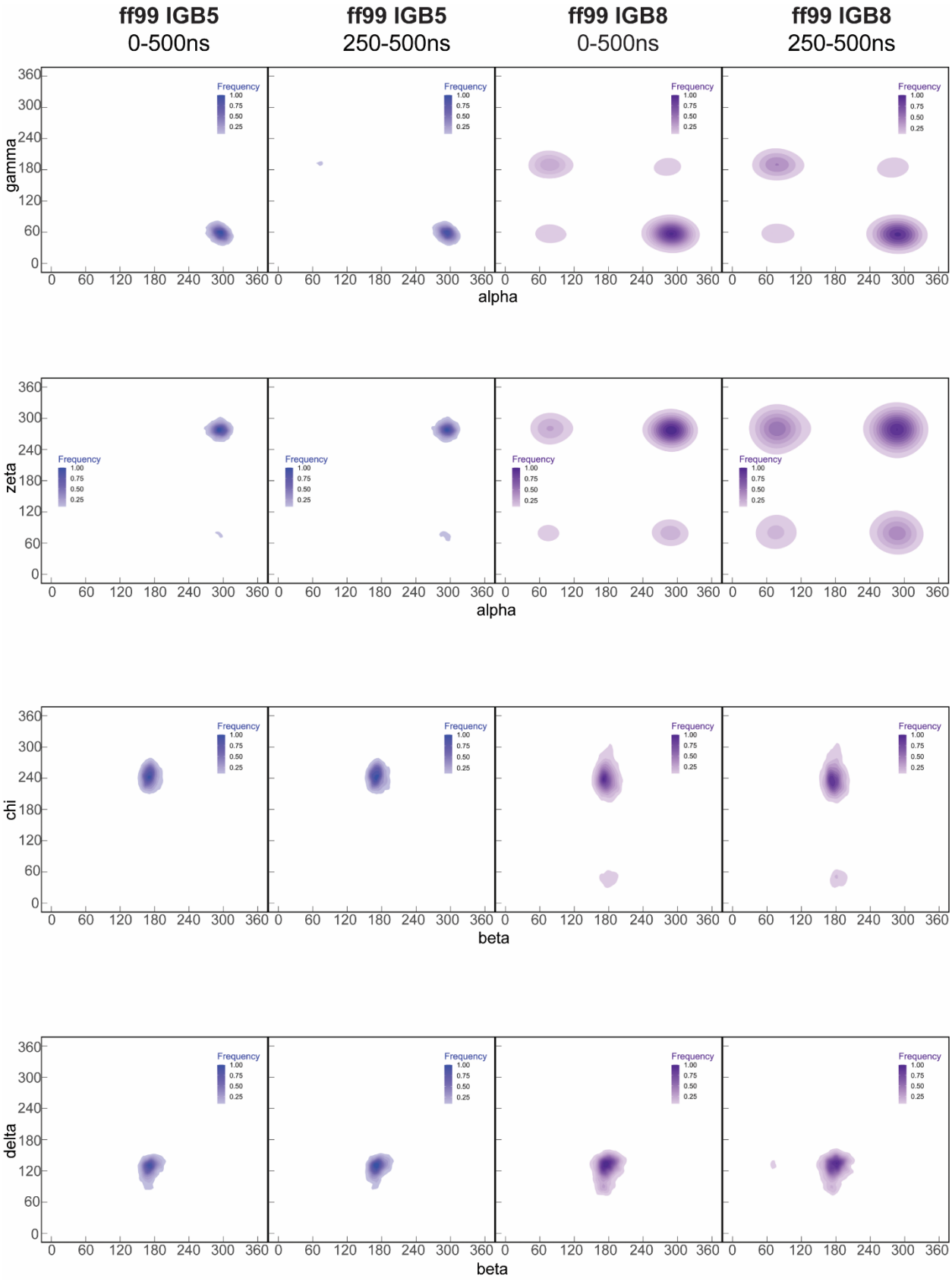


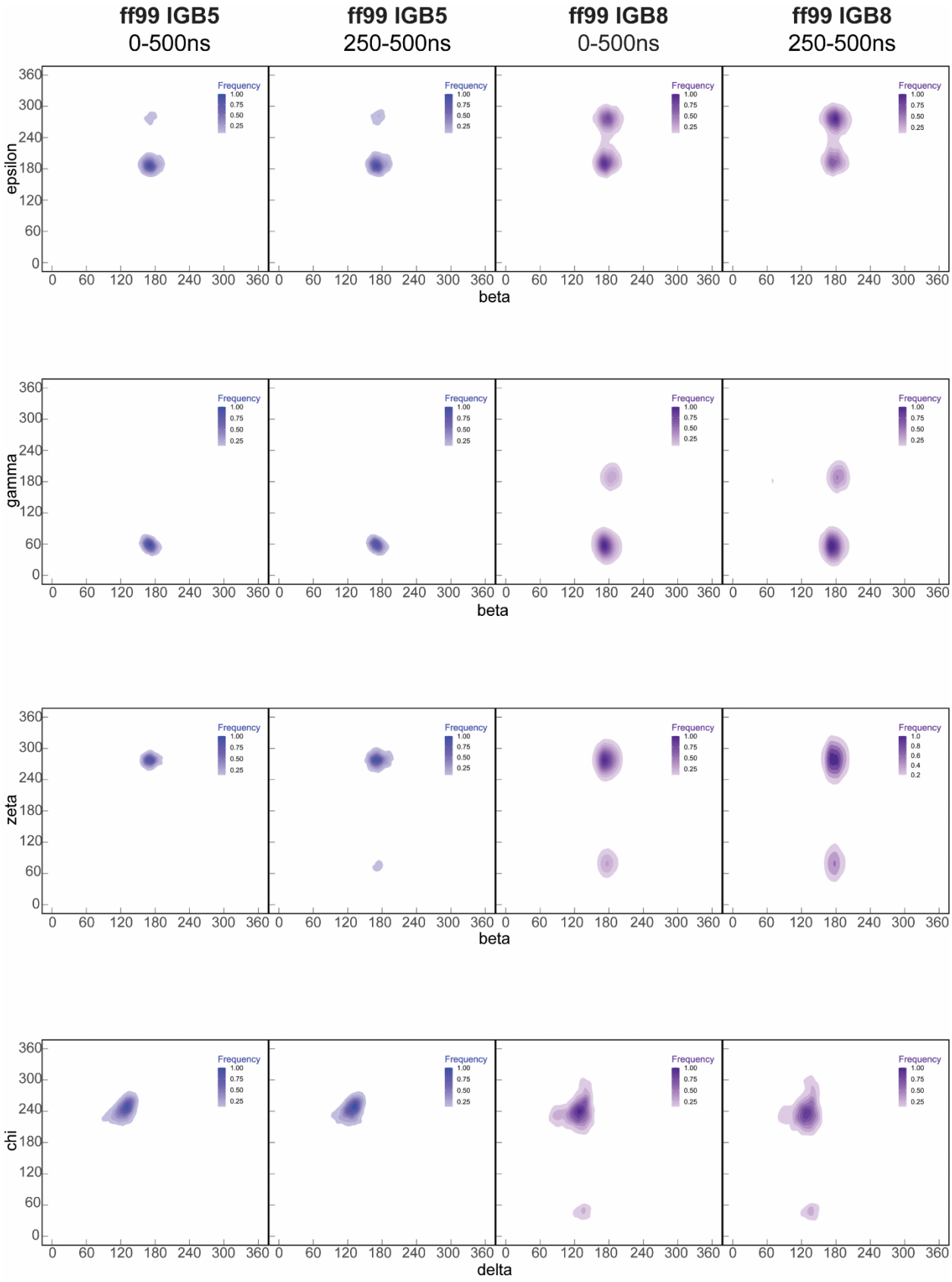
Figure S5.1.3. Shows the minimum χ^2 value in each generation.

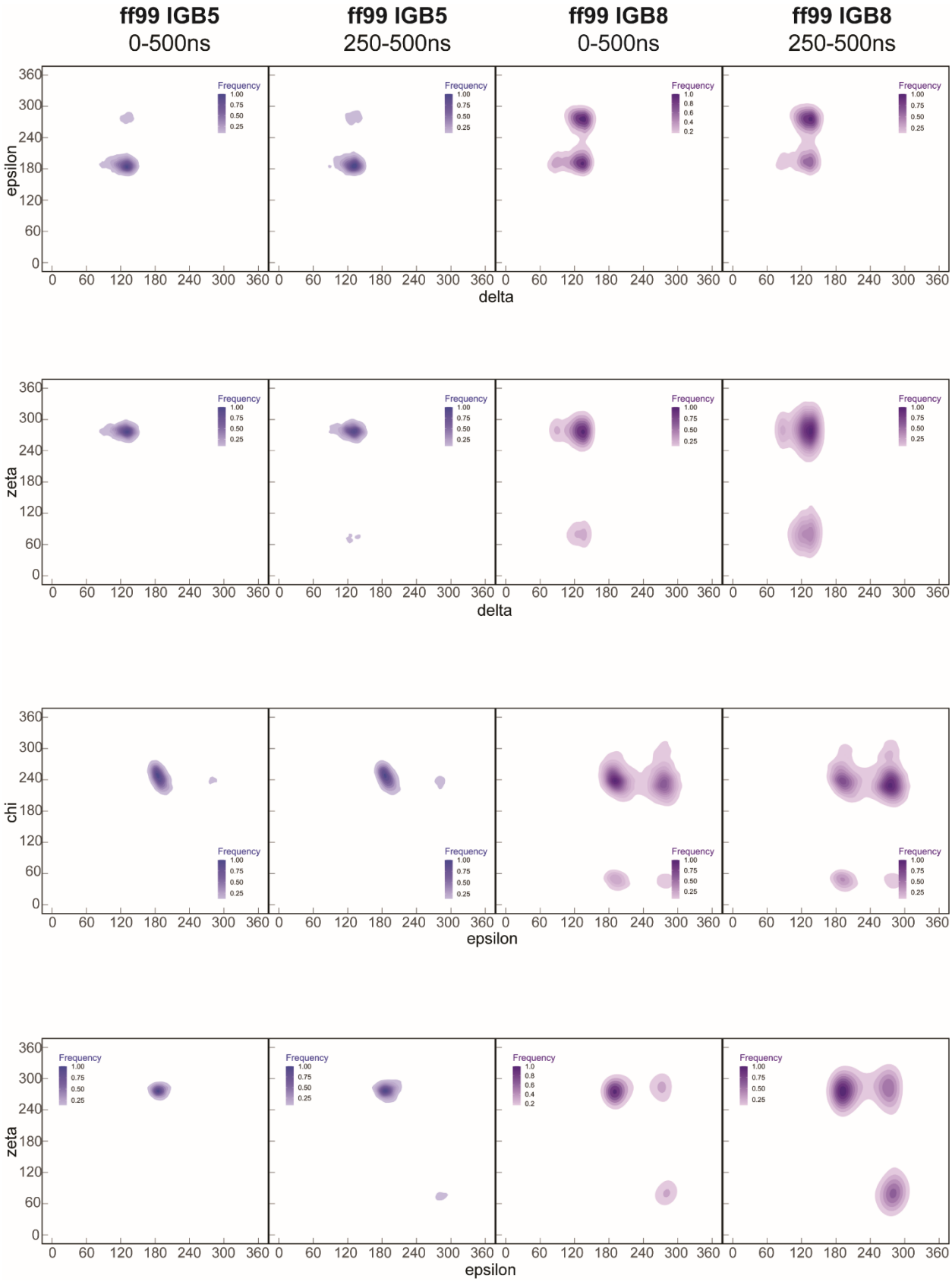
Figure S5.1.4 and S5.1.6 show contour plots for dihedral angles in the two best implicit and explicit solvent simulations. For each simulation, the plots represent ssDNA structures from every 0.1ns across 0-500ns or 250-500ns. Overall, there are minor differences in the bivariate contour plots based on the time in which simulation data is taken. Deviation between plots of the same simulation likely comes from ssDNA sampling structures farther away from the canonical B-DNA structure, which the simulations started in prior to equilibration. Considering the latter half of the simulation represents the conformational sampling of more converged structures, those bivariate plots are used for further analysis between the GA and simulation.

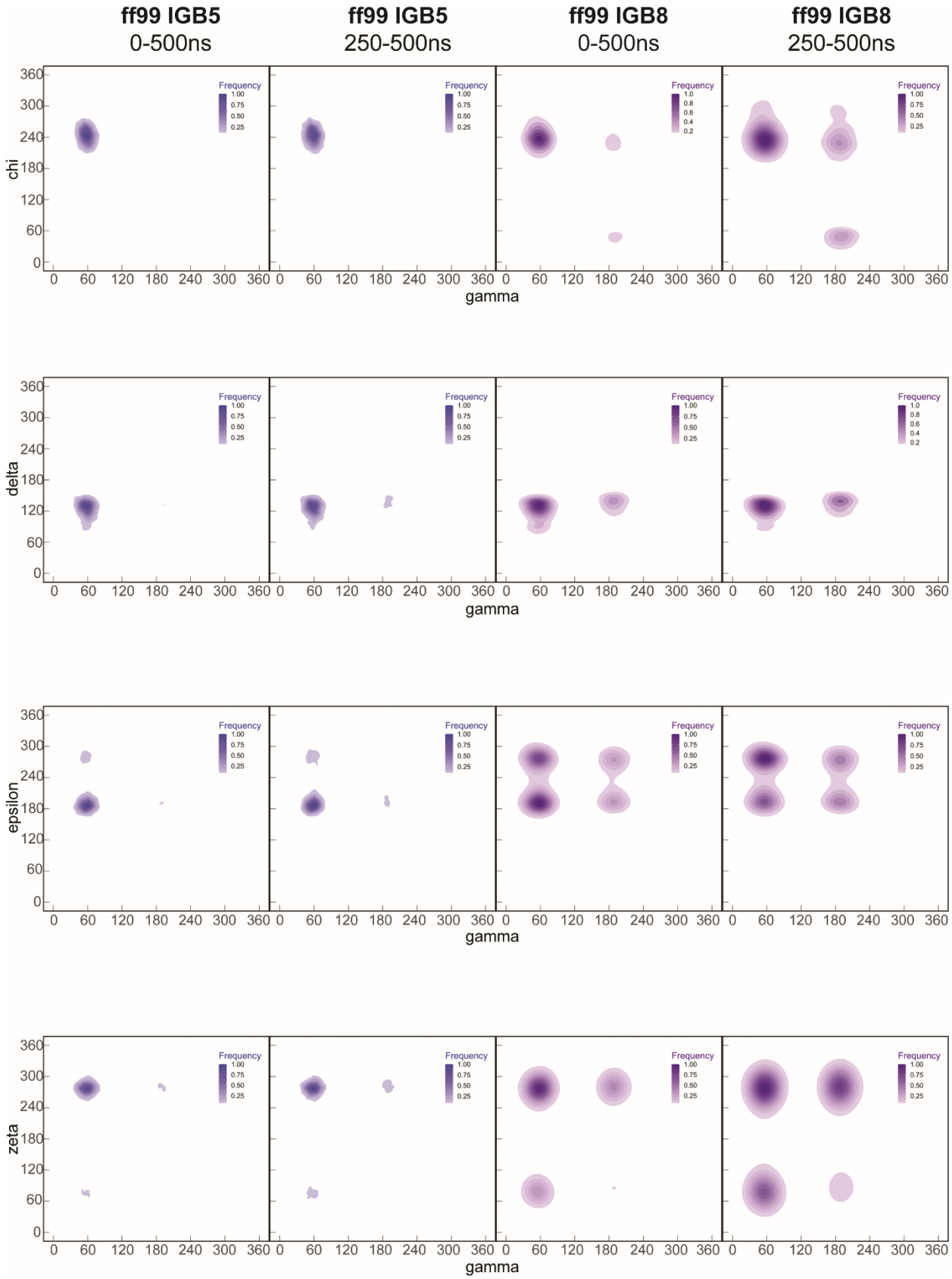
Figure S5.1.4. Shows bivariate contour plots for the alpha, beta, chi, delta, epsilon, gamma, and zeta dihedral angles. The dihedral angles are calculated from the ensemble of structures chosen by the GA and compared to the dihedral angles calculated from MD simulation. The simulations with the bsc1 and OL15 force field in TIP3P solvent are shown due to their performance compared to experiment.

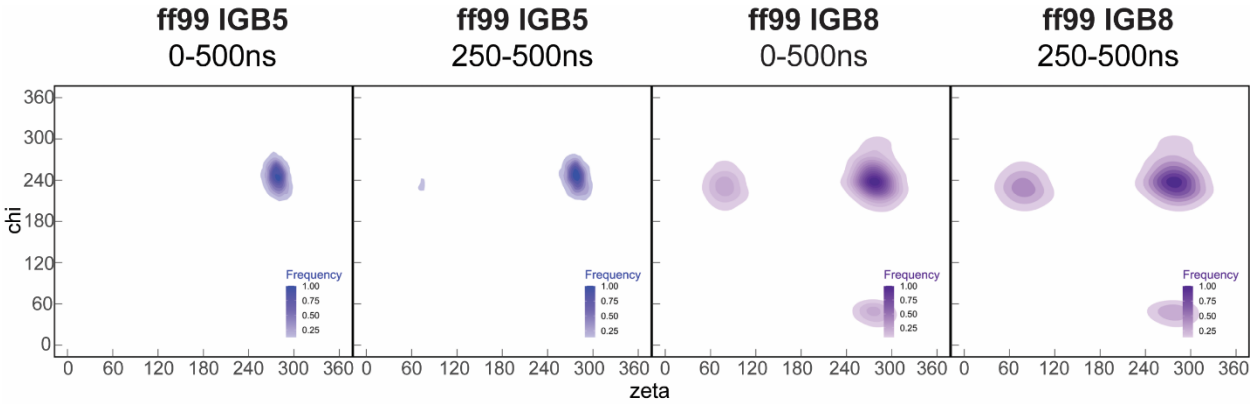

ff99 IGB5
0-500ns
ff99 IGB5
250-500ns
ff99 IGB8
0-500ns
ff99 IGB8
250-500ns
beta
chi
delta
epsilon
alpha
Frequency
0
60
120
180
240
300
360


ff99 IGB5
0-500ns
ff99 IGB5
250-500ns
ff99 IGB8
0-500ns
ff99 IGB8
250-500ns
gamma
alpha
zeta
alpha
chi
beta
delta
beta
Frequency
1.00
0.75
0.50
0.25
0
60
120
180
240
300
360


ff99 IGB5
0-500ns
ff99 IGB5
250-500ns
ff99 IGB8
0-500ns
ff99 IGB8
250-500ns
epsilon
gamma
zeta
chi
beta
beta
beta
delta
Frequency

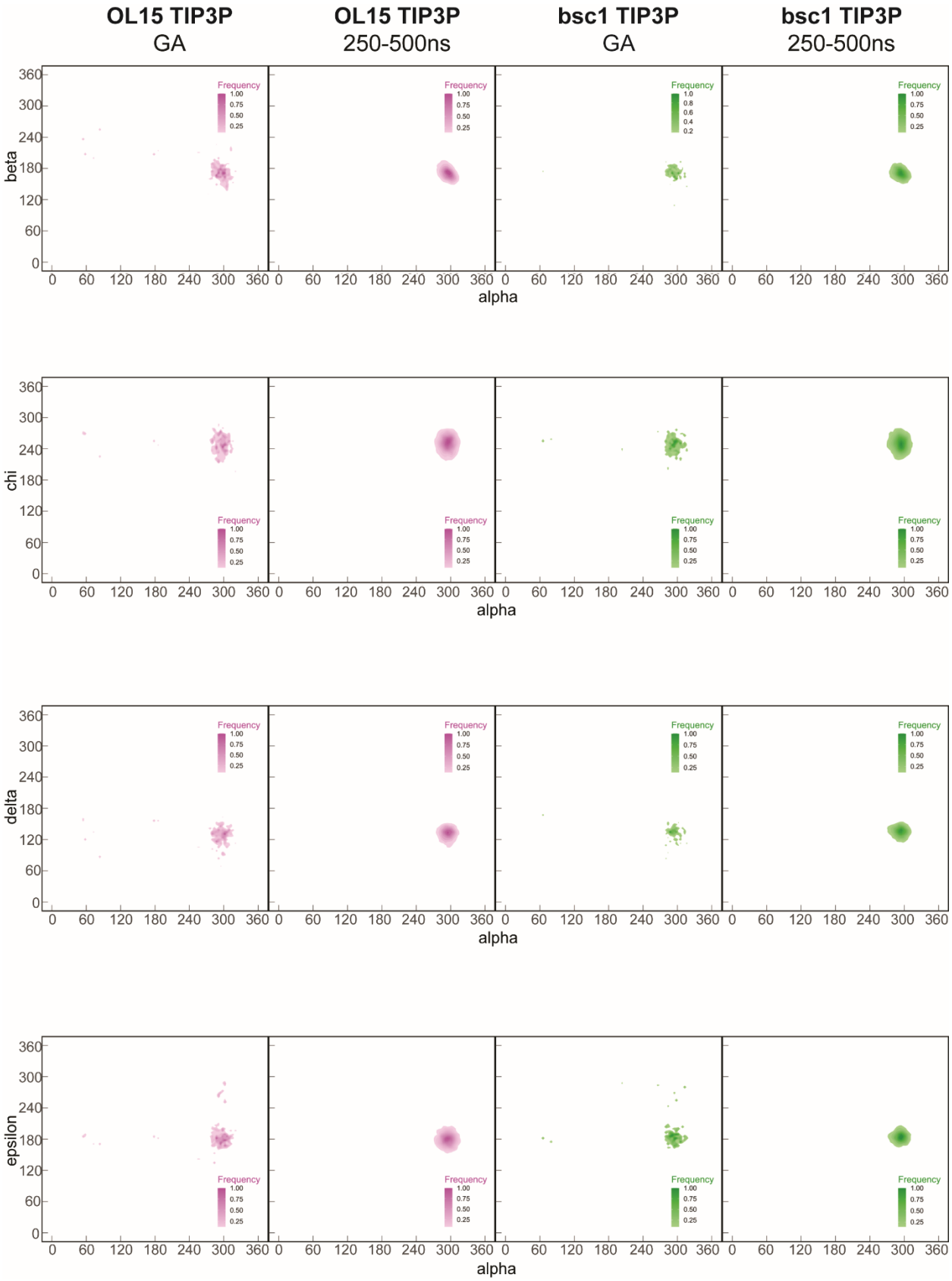

ff99 IGB5
0-500ns
ff99 IGB5
250-500ns
ff99 IGB8
0-500ns
ff99 IGB8
250-500ns
epsilon
delta
zeta
delta
chi
epsilon
zeta
epsilon
Frequency

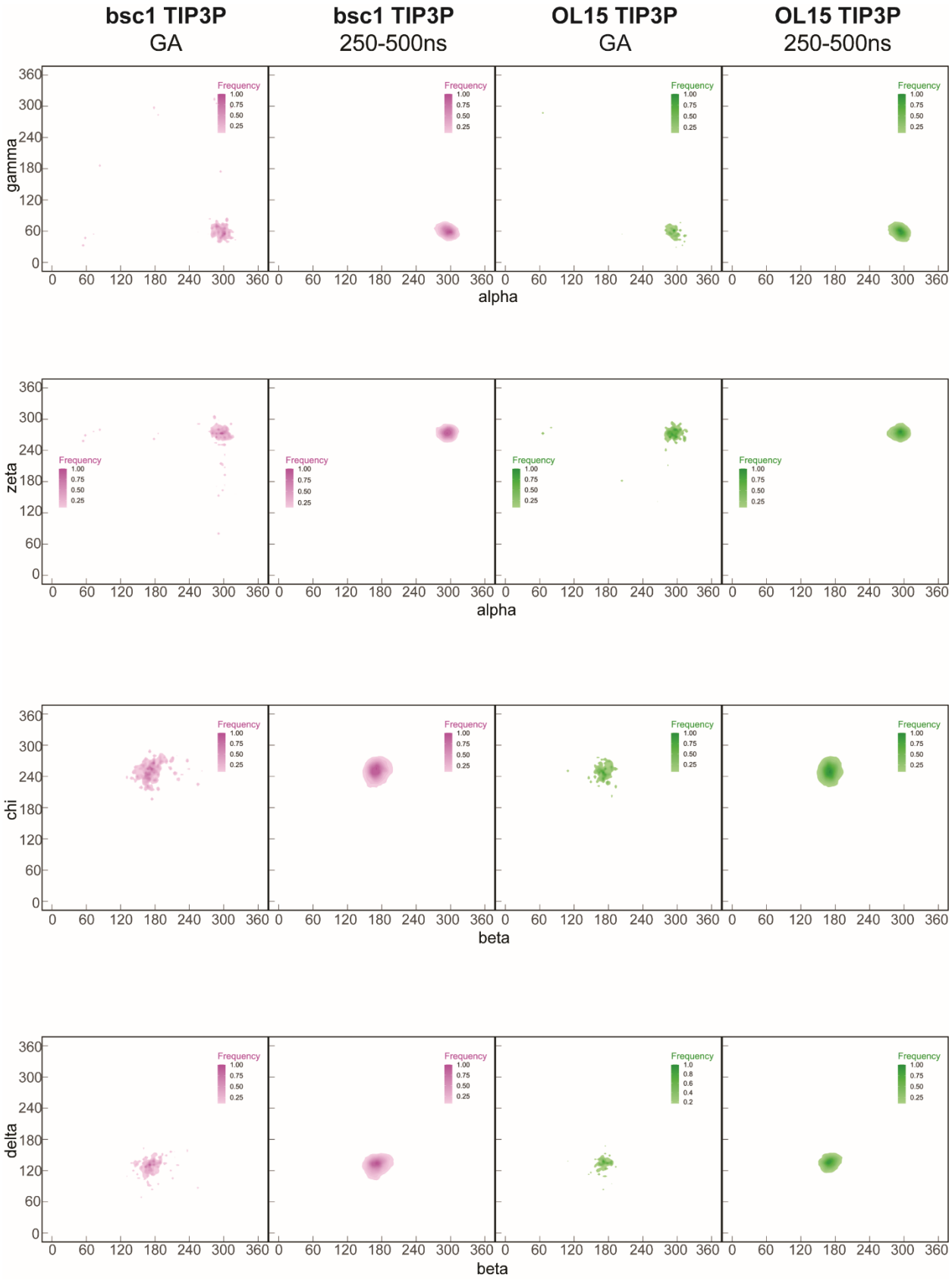

ff99 IGB5
0-500ns
ff99 IGB5
250-500ns
ff99 IGB8
0-500ns
ff99 IGB8
250-500ns
chi
delta
epsilon
zeta
gamma
Frequency

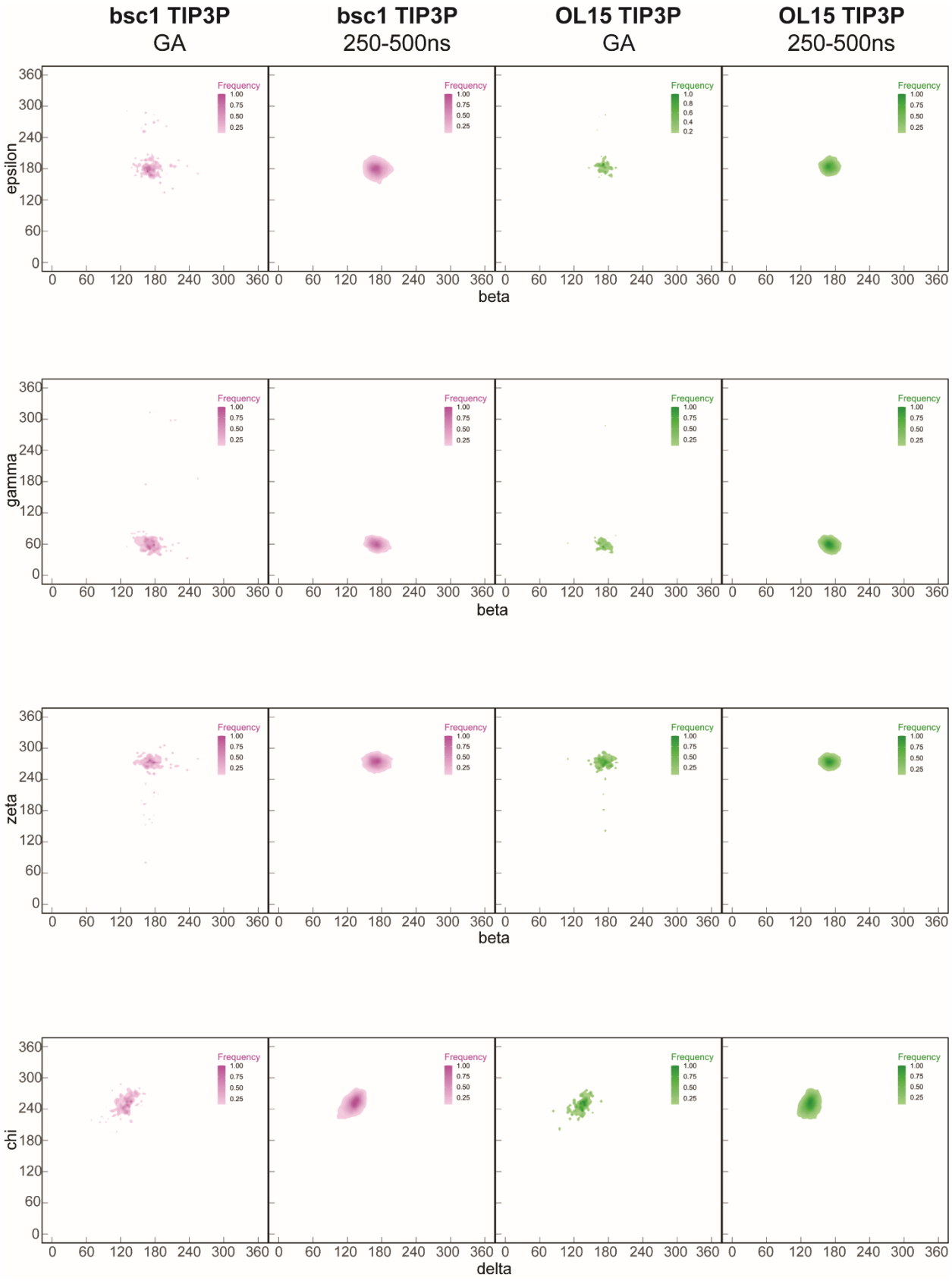

ff99 IGB5
0-500ns
ff99 IGB5
250-500ns
ff99 IGB8
0-500ns
ff99 IGB8
250-500ns
chi
zeta
Frequency
1.00
0.75
0.50
0.25
0
60
120
180
240
300
360

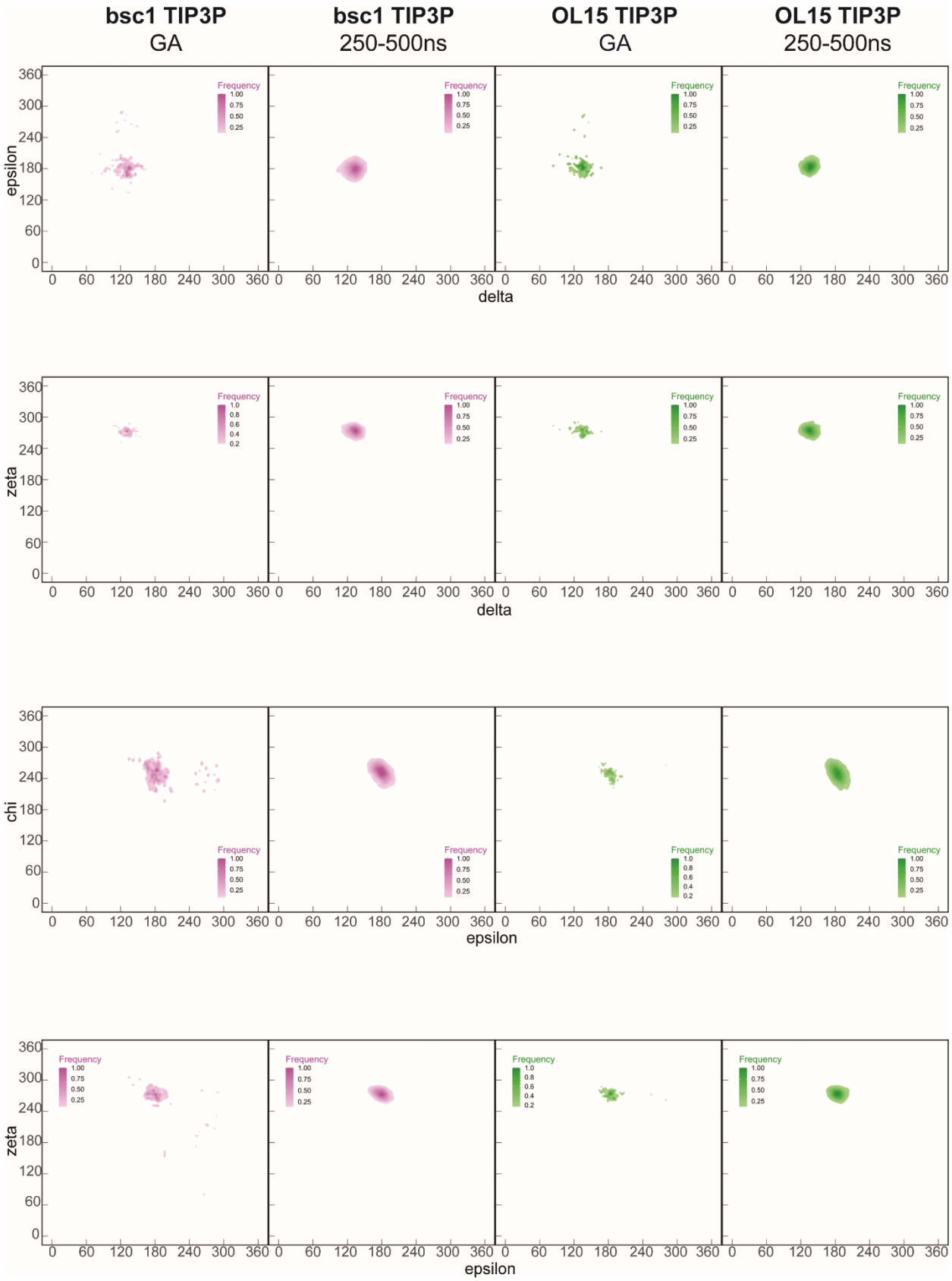
The bivariate plots in Figure S.5.1.4 vary in transparency based on the frequency of sampling. Thus, if a point is sampled infrequently enough compared to the bulk of the simulation, it will not be visible. The alpha-zeta contour plot in Figure S5.1.4 illustrates this as it appears there is a reduced sampling of this dihedral angle over the entire simulation compared to the last 250ns. The entire sampling range is present over the simulation, the only difference is the relative frequency of datapoints when comparing data over the time ranges. Figure S.5.1.6 has less variation than Figure S.5.1.4. The lack of dihedral angle sampling variation in explicit solvent could be due to the parameterization of solvent further stabilizing the B-DNA conformation or could be due to the viscosity of explicit solvent slowing and conformational changes.

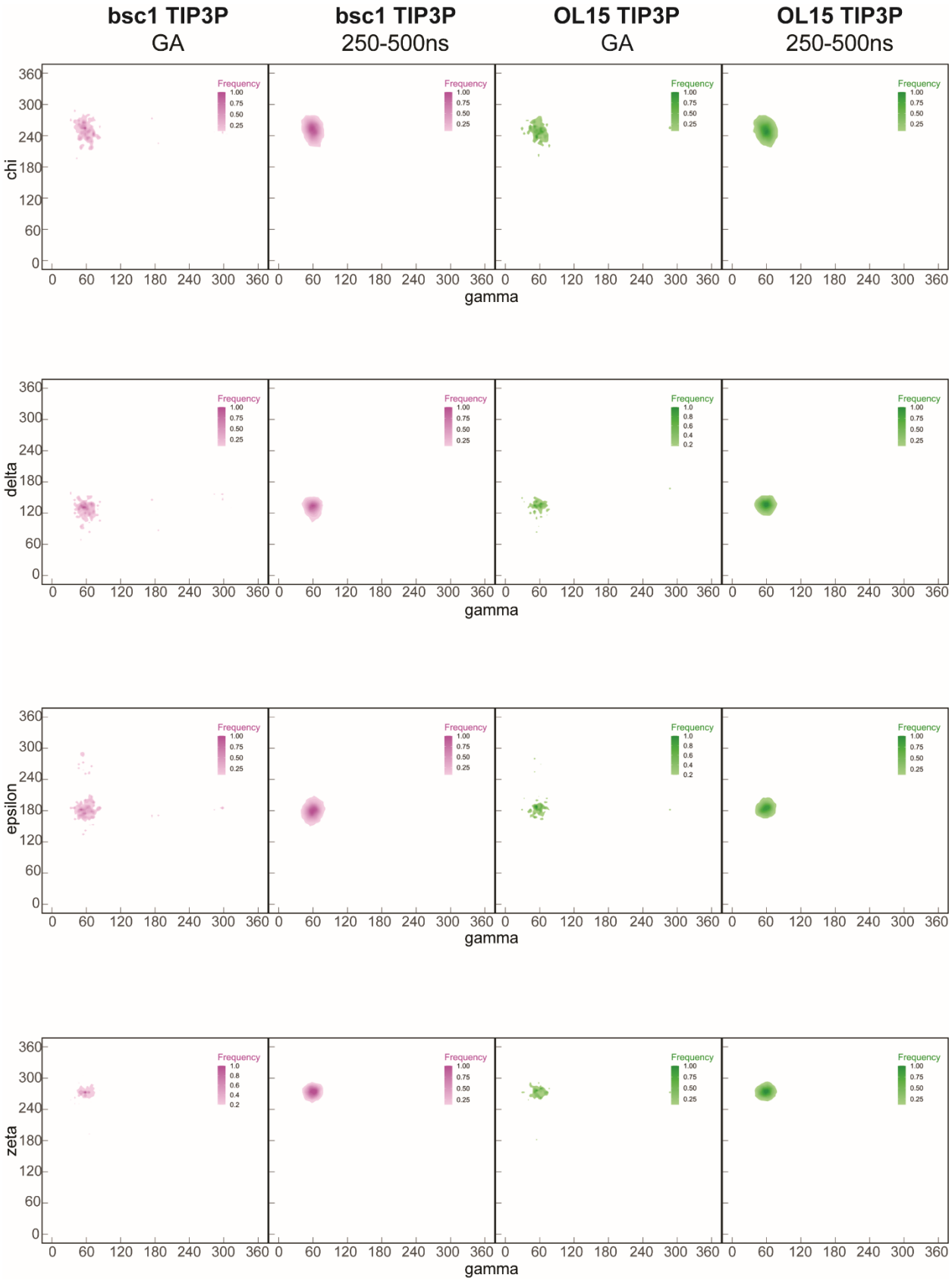
Figure S5.1.5. Shows bivariate contour plots for the alpha, beta, chi, delta, epsilon, gamma, and zeta dihedral angles. The dihedral angles are calculated from the ensemble of structures chosen by the GA and compared to the dihedral angles calculated from MD simulation. The simulations with the bsc1 and OL15 force field in TIP3P solvent are shown due to their performance compared to experiment. The selection of presenting 250-500 ns is discussed in Figure S5.1.6.


OL15 TIP3P
GA
OL15 TIP3P
250-500ns
bsc1 TIP3P
GA
bsc1 TIP3P
250-500ns
beta
chi
delta
epsilon
alpha
Frequency
0
60
120
180
240
300
360


bsc1 TIP3P
GA
bsc1 TIP3P
250-500ns
OL15 TIP3P
GA
OL15 TIP3P
250-500ns
gamma
alpha
zeta
alpha
chi
beta
delta
beta
Frequency
1.00
0.75
0.50
0.25
0
60
120
180
240
300
360


bsc1 TIP3P
GA
bsc1 TIP3P
250-500ns
OL15 TIP3P
GA
OL15 TIP3P
250-500ns
epsilon
gamma
zeta
chi
beta
beta
beta
delta
Frequency
0
60
120
180
240
300
360


bsc1 TIP3P
GA
bsc1 TIP3P
250-500ns
OL15 TIP3P
GA
OL15 TIP3P
250-500ns
epsilon
delta
zeta
delta
chi
epsilon
zeta
epsilon
Frequency


bsc1 TIP3P
GA
bsc1 TIP3P
250-500ns
OL15 TIP3P
GA
OL15 TIP3P
250-500ns
chi
delta
epsilon
zeta
gamma
Frequency

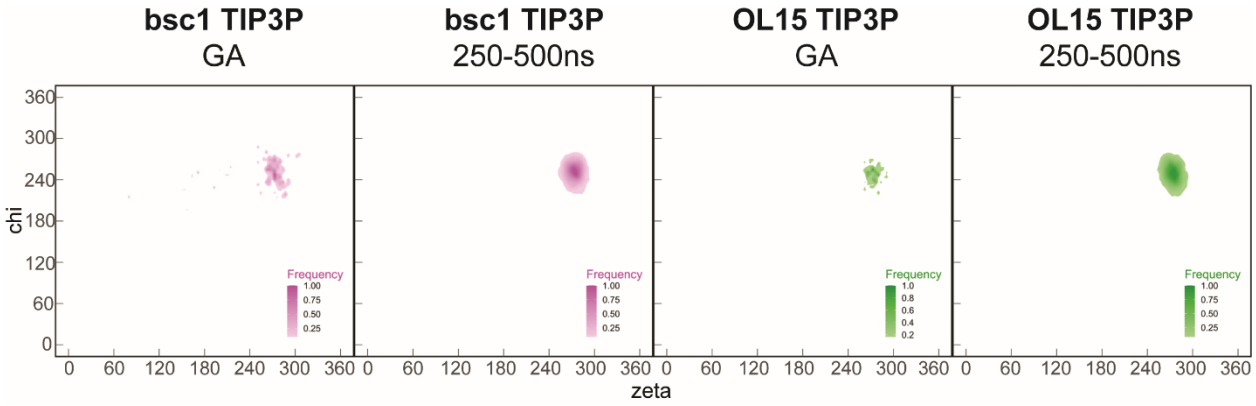
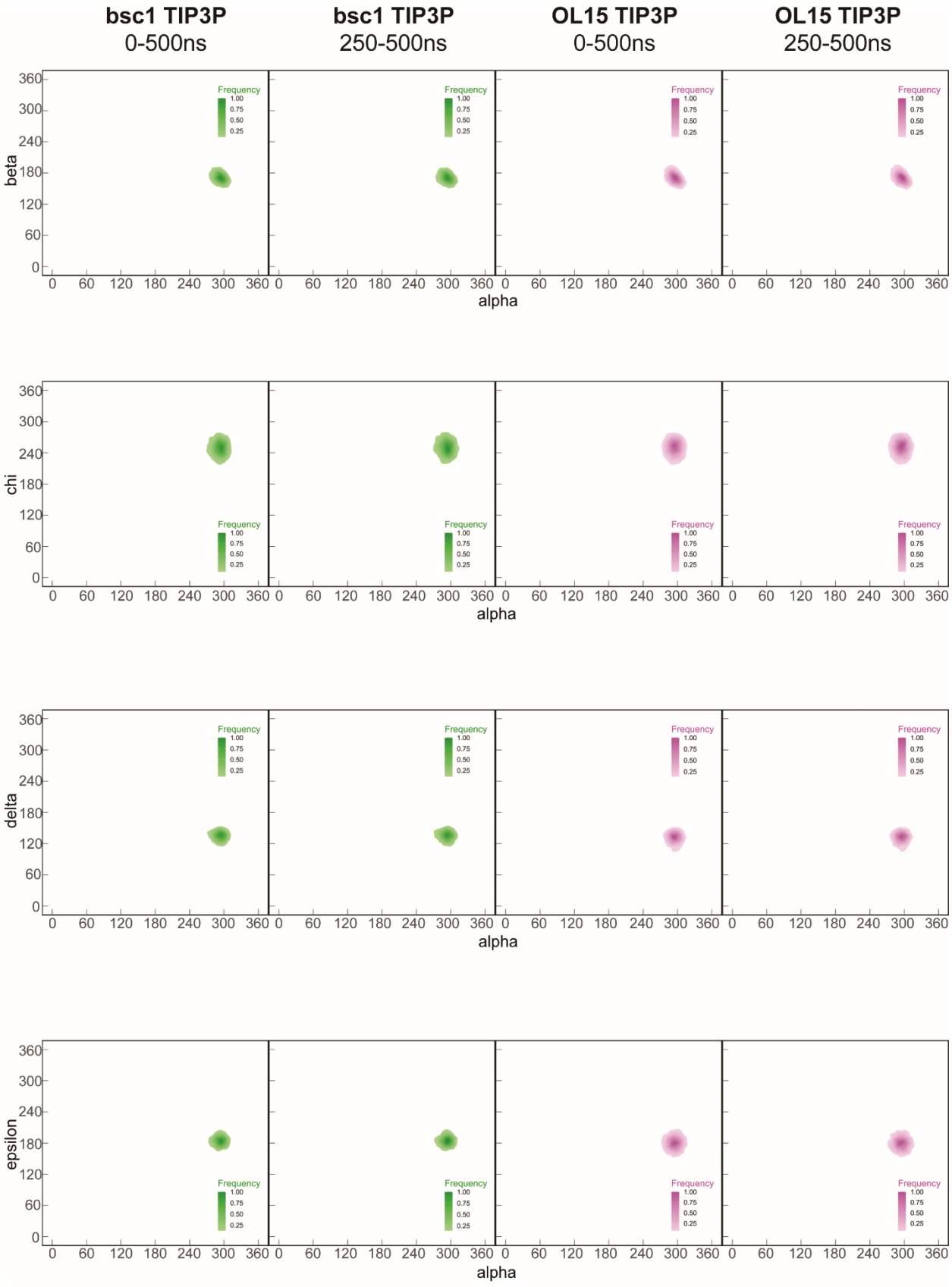
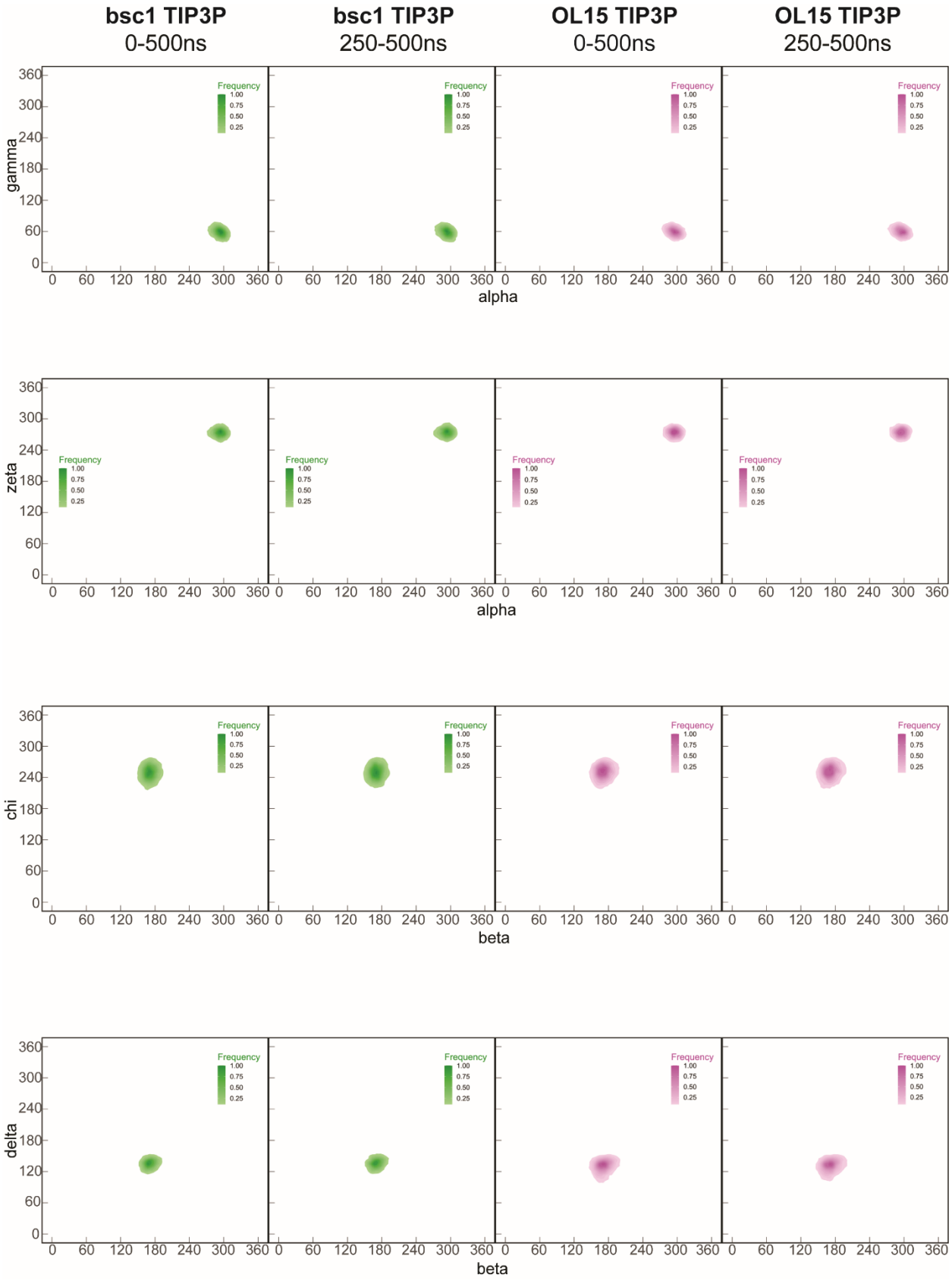
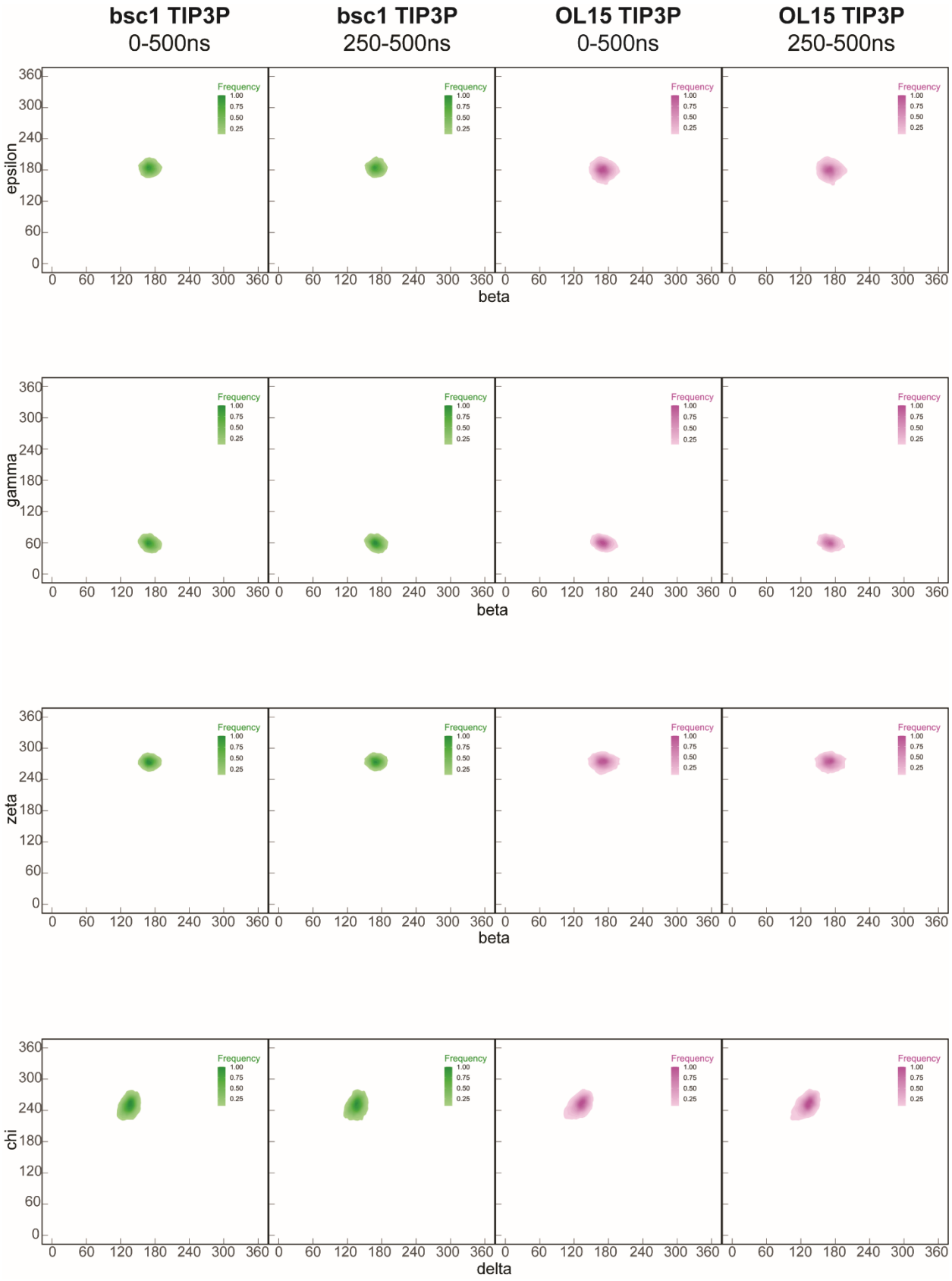

bsc1 TIP3P
GA
bsc1 TIP3P
250-500ns
OL15 TIP3P
GA
OL15 TIP3P
250-500ns
chi
zeta
Frequency
0
60
120
180
240
300
360

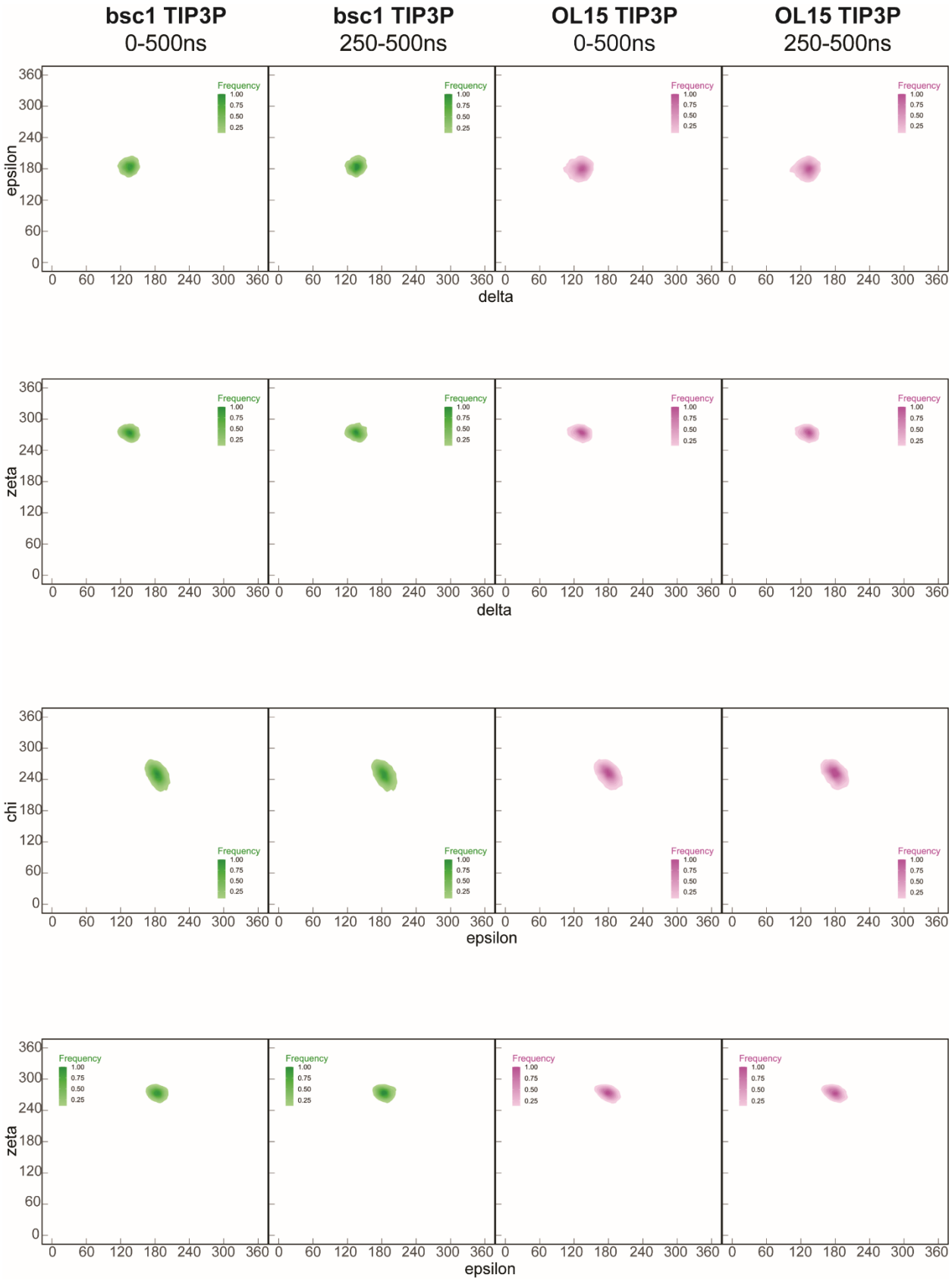
Figure S5.1.5 shows the limited range of dihedral angle exploration in explicit solvent. Considering explicit solvent simulations are known to require more time to converge and reach equilibrium, it is possible that more time is needed to explore the conformations of ssDNA more appropriately. This is supported by the shift in the contour plots seen in Figure S5.1.4. The latter half of implicit solvent simulations showed more dihedral angle exploration as the canonical B-form was abandoned. Thus, the relative frequency of the newly explored states, closer to convergence were registered in the contour plots. Figure S5.1.6 shows no notable differences between the 0-500ns distribution and 250-500ns distribution in explicit solvent.

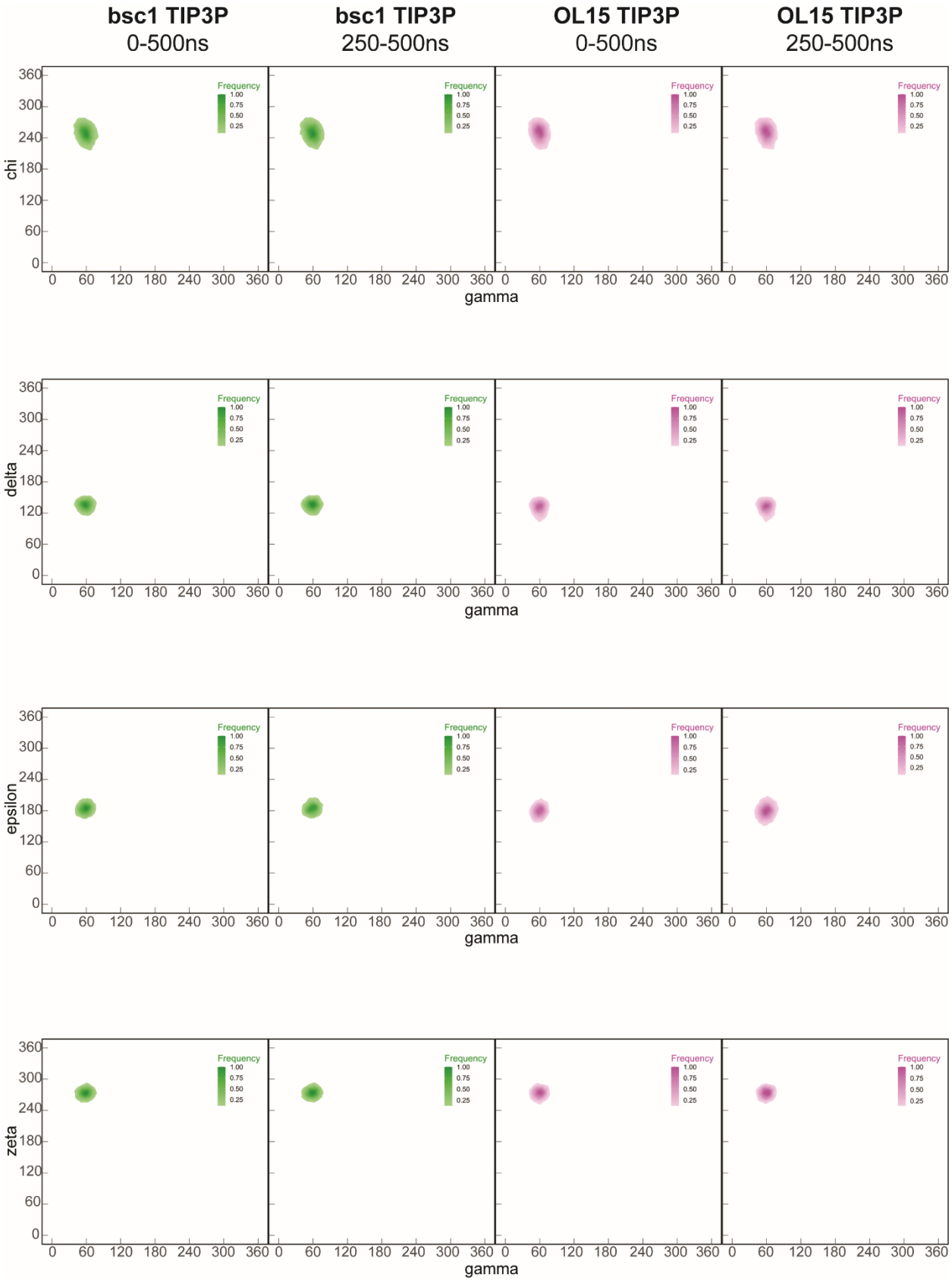
Figure S5.1.6. Shows the explored dihedral angle space of polyT in MD simulation using bsc1 and OL15 force field in TIP3P solvent. The dihedral angles relative frequency of occurrence is simulation is represented by contour plots. The more transparent the region, the lesser the frequency.

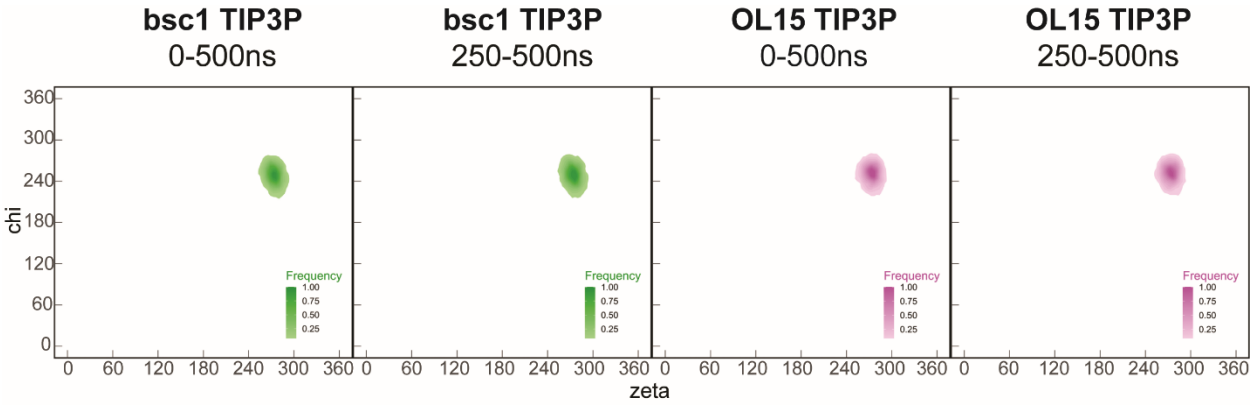

bsc1 TIP3P
0-500ns
bsc1 TIP3P
250-500ns
OL15 TIP3P
0-500ns
OL15 TIP3P
250-500ns
beta
chi
delta
epsilon
alpha
Frequency
1.00
0.75
0.50
0.25
0
60
120
180
240
300
360


bsc1 TIP3P
0-500ns
bsc1 TIP3P
250-500ns
OL15 TIP3P
0-500ns
OL15 TIP3P
250-500ns
gamma
zeta
chi
delta
alpha
beta
Frequency
1.00
0.75
0.50
0.25
0
60
120
180
240
300
360


bsc1 TIP3P
0-500ns
bsc1 TIP3P
250-500ns
OL15 TIP3P
0-500ns
OL15 TIP3P
250-500ns
epsilon
gamma
zeta
chi
beta
delta
Frequency
1.00
0.75
0.50
0.25
0
60
120
180
240
300
360


bsc1 TIP3P
0-500ns
bsc1 TIP3P
250-500ns
OL15 TIP3P
0-500ns
OL15 TIP3P
250-500ns
epsilon
delta
zeta
delta
chi
epsilon
zeta
epsilon
Frequency
1.00
0.75
0.50
0.25
0
60
120
180
240
300
360


bsc1 TIP3P
0-500ns
bsc1 TIP3P
250-500ns
OL15 TIP3P
0-500ns
OL15 TIP3P
250-500ns
chi
delta
epsilon
zeta
gamma
Frequency
1.00
0.75
0.50
0.25
0
60
120
180
240
300
360


bsc1 TIP3P
0-500ns
bsc1 TIP3P
250-500ns
OL15 TIP3P
0-500ns
OL15 TIP3P
250-500ns
chi
zeta
Frequency
1.00
0.75
0.50
0.25
0
60
120
180
240
300
360

5.1.6 Acknowledgements

Farshad Saberi-Movahed and Leila Khalili

CHAPTER 6: Outlook

6.1 Materials Informatics

* This section is a reproduction from parts of the manuscript for the published works:
James S. Peerless, Nina J. B. Milliken, Thomas J. Oweida, Matthew D. Manning, Yaroslava G. Yingling. *Advanced Theory and Simulations* **2018** *2* (1)*,* 1800129.
DOI: 10.1002/adts.201800129

Matthew D. Manning, Albert L. Kwansa, Thomas Oweida, James S. Peerless, Abhishek Singh, Yaroslava G. Yingling. *Biointerphases* **2018** *13* (6)*,* 06D502.
DOI: 10.1116/1.5044381

6.1.1 Machine Learning for Polymer Systems

ML techniques applied specifically to polymer systems have had a relatively long history. Much of this work utilizes ML to inform the construction of quantitative structure–property relations (QSPRs), which can be thought of as precise mathematical forms of a PSPP relation.[241] QSPRs may or may not incorporate ML algorithms, but their overall goal of connecting structural descriptors to macroscopic properties can often be aided by ML techniques. As early as 1994, Sumpter and Noid showed a neural network approach improved predictions of glass transition temperatures from polymer structural descriptors over more traditional QSPR methods.[242] Glass transition temperature has proved to be a fertile test case for ML application, with many studies building on the work of Sumpter and Noid by more advanced applications of neural networks to larger and more diverse data sets.[243–247] Yu et al. applied a multiple linear regression (MLR) technique to the glass transition temperature of a set of 107 polymers, yet this ML technique requires sufficient domain knowledge in the selection and construction of descriptors.[248] This work was subsequently built on by Yu in subsequent research where 1664 structural descriptors were automatically produced and selected, followed by the application support vector machine (SVM) algorithm to construct a non-linear QSPR relationship.[249] More recently, Pei et al. applied SVR (the regression form of SVM) with a particle swarm optimization (PSO) method to further improve

a QSPR for glass transition temperature based on quantum descriptors of 19 polymers.[250,251] Chen et al. used a genetic algorithm (GA) and an MLR technique to build a QSPR to predict glass transition temperature of a diverse set of polymers. The GA-MLR technique found that a model consisting of seven variables provided the best fit for the training and test data sets without overfitting. The nature of these structural descriptors indicated that the glass transition temperature could be governed by the electronegative groups on polymers.[252]

In addition to glass transition temperatures, researchers have focused on predicting properties such as refractive indices and dielectric constants.[253–260] Jabeen et al. used four descriptors selected via GA in an MLR-derived QSPR model and highlight polarizability and sp2 hybridized carbon atoms as the most influential descriptors.[254] This model was used to test four small virtual libraries of novel polymers, which identified several polymers that had refractive indices higher than the highest refractive indices in the training data set. Venkatraman et al. also used a QSPR model built on linear partial least squares regression and ensemble tree-based random forests to rapidly test polymers for high refractive indices.[253] The researchers used geometrical quantum chemistry-based descriptors consisting of the highest occupied and lowest unoccupied molecular orbital energies, charges, polarizabilities, super delocalizabilities, and radial distribution function indices. The authors mention that while their model is sufficiently predictive, there are limitations on the extrapolation that can be performed resulting in high variance between the predictive model and experimental observations. This highlights the importance and difficulty associated with choosing the appropriate descriptors for data-driven techniques.

In the field of polymer dielectrics, Mannodi-Kanakkithodi et al. recently used chemical blocks of CH2, CO, NH, C6H4, C4H2S, CS, O to form 284 symmetrically unique 4-block polymers.[256] These studies utilized data mining of ab initio data, kernel ridge regression (KRR), and GA to

design simple polymer dielectrics with higher energy densities than the current state-of-the-art biaxially oriented polypropylene (BOPP).[255,256,259,261] Based on the data from the predictive model, a polyurea polymer, polyimide polymer, and polythiourea polymer were synthesized, characterized, and proven to have higher energy densities than BOPP, providing potential alternatives for polymer dielectrics beyond BOPP. The successful predictions of organic polymers led the authors to expand their exploration of chemical space and include organometallic polymers, which indicated that the ionic contribution of polymers with metals such as Sn could provide significant improvements in polymer dielectric properties.[255,259,261] However, combining organic and organometallic polymers into the same data set led to a decrease in prediction accuracy, due to the wider range of chemical space.[255] The prediction accuracy could most likely be improved by enhancing the number of organometallic polymers in the training data set because the ML algorithm may have viewed the organometallic polymers as having extremal properties.[255,262] Overall, data-driven techniques have guided researchers to promising polymer chemistries for dielectric constants and bandgaps, but properties such as dielectric loss, dielectric breakdown, resistance to degradation, film formability, and mechanical behavior are relevant to the application of dielectric polymers. Thus, it can be concluded that ML techniques are necessary but not alone sufficient for the rapid exploration of highly divergent materials.[257]

Overall, polymer property predictions and designs have benefited from the use of data-driven techniques. However, polymers present inherent difficulties that can limit the accuracy of its predictions. For instance, polymer properties can vary with molecular weight, which is typically described by a distribution.[263–265] Wu et al. attempt to address this issue by introducing infinite chain descriptors. This idea is based on the convergence of some polymeric properties as molecular weight increases.[266] While this eliminates the effects of calculations using end-capped monomers,

data-driven approaches to polymer prediction and design are still unable to deal with chain branching, cross linking, and multiple monomers.[263,266] This increasingly poses a challenge as the synthesis of polymers becomes more advanced and the production of materials such as aperiodic copolymers become more prominent, which creates a new class of polymers that does not fall into the standard block, graft, alternating, or random copolymer categories.[267]

As discussed in a review by Ramprasad et al., the reduction of each input material to a unique, descriptive representation takes domain expertise and is the most influential step on each data-driven strategy, while the mapping between the descriptors and target property requires little to no domain expertise. Furthermore, the size scale of the descriptors must be taken into account, as the properties to be predicted may be a function of organization at multiple size scales.[268] As an example, a polyethylene sample could be represented in terms of its amount of crystallinity, repeat unit -[CH2-CH2]n-, its chemical block -CH2-, or its fundamental atomic or molecular fragments such as fourfold carbon and onefold hydrogen.[261] Huan, Mannodi-Kanakkithodi, and Ramprasad address this in a recent study by applying a hierarchical fingerprinting approach to small-molecule KRR of various bulk properties that have had some success in translation to longer polymer chains.[255,267,269] However, due to the organization of polymers at multiple length scales, application of this fingerprinting technique invokes an exponentially larger parameter space when capturing larger-scale descriptions such as copolymerization, branching, and crystalline phases. Yet, the application of a hierarchical fingerprinting approach is still a promising one for polymers going forward.

Moving forward in ML for polymer systems, researchers should continue to explore a broader chemical space for linear polymers, as Mannodi-Kanakkithodi et al. did with organic and organometallic polymers for dielectric applications.[255] While investigating this wider range of

chemical space decreased accuracy, the improvement of polymeric descriptors and larger data sets should provide a higher degree of accuracy for these more universal models. However, expanding data sets should be done strategically by employing techniques such as active learning. Zhao et al. demonstrated a sixfold reduction in the number of DPD runs required by using an active learning strategy based on a Gaussian process regression to help select which simulations were most necessary to perform.[270] In addition, researchers should creatively apply algorithms that will allow the prediction and design of polymers that meet multiple criteria or perform constrained optimization. This can further facilitate the novel application of polymers by employing a more application-specific screening process. Researchers will also have to overcome the challenge of developing polymer descriptors that capture hierachical polymer structures (and the distributions thereof), such as degree of crystallinity and molecular weight. The hierarchical fingerprinting technique developed by Huan et al. is a promising methodology for this, yet it will likely take a considerable effort to further extend the method to complex polymer formulations such branches and copolymers.[255,269] Data-driven techniques have also been used to create force fields for computational studies. To date, the parameterization of force fields has been confined to elemental solids; however, approaches have been developed that can be universally applied to multicomponent systems. Specifically, AGNI force fields have been shown to rival the accuracy of computations performed by DFT with a significant increase in the speed of the computations given a sufficient training set of atomic arrangements.[271–273] Currently, a large barrier for AGNI force fields in soft materials is the increased number of atomic arrangements possible in these materials. Finally, and most importantly, polymer chemists, materials scientists, and computer scientists must continue to communicate their work such that a shared language for ML techniques for soft matter applications may be developed. Often researchers are daunted by the volume and

selection of ML techniques that are avail- able and may have concerns about their connection to underlying chemistry or physics. Yet, as further interdisciplinary work is performed, it is hoped that ML techniques will further become part of the general materials science lexicon and will seem less like black-box approaches, similar to the adoption of simulation techniques seen in the previous decade. Drastic improvements in ML techniques for polymer systems, however, are contingent on the availability of well-organized and curated polymer data on which to train more precise algorithms. The current state of available data and the challenges intrinsic in its improvement are discussed in length below.

6.1.2 Polymer Databases

6.1.2.1 Currently Available Databases

There are a number of currently available polymer databases that attempt to organize polymer material properties with varying degrees of generality and coverage. Broadly, polymer databases can be divided into two categories: i) databases that attempt to organize a variety of chemical, mechanical, electronic, and thermal properties; and ii) databases that specialize in a few select properties such as Flory–Huggins χ parameters, NMR spectra, or MALDI methods.[274–277]

Databases such as National Institute for Materials Science (NIMS) PoLyInfo, CROW Polymer Properties Database, the MatWeb Material Property Database, Citrine Informatics' Citrination platform, and the MakeItFrom Materials Properties Database organize a multitude of polymer properties.[73,278–281] For a commonly found polymer material, poly(methyl methacrylate) for example, these databases contain between 21 (Citrination) and 75 (PoLyInfo) properties such as tensile strength, refractive index, specific heat capacity, and thermal diffusivity.

Even more databases are dedicated to one or two specific material properties. For example, the Polymer Property Predictor and Database developed by the Center for Hierarchical Materials Design

records only Flory–Huggins χ parameters and glass transition temperatures for various polymers; the Polymer Science Learning Center Spectral Database developed at the University of Wisconsin–Stevens Point and the ACD/Labs NMR Database principally catalog polymer IR and NMR spectra. These specialized databases are often supplemented with additional functionality such as property predictor engines informed by hosted data.[274–276]

6.1.2.2 Limitations of Current Databases

The unique characteristics of polymeric materials make the development of cohesive databases challenging compared to other materials such as biomaterials, metals, or ceramics. Therefore, any advancement made in the development of polymer databases is predicated on a consensus resolution of the following challenges.

Polymeric materials are intrinsically complex and algorithmic methods of naming materials composed of even somewhat complex macromolecules become arduous when one considers the enormous diversity of polymer morphology. Beyond chemical composition, polymer categorization must consider copolymerization, polymer blending, linear versus branched polymers, the extensive library of polymer endcaps, the fact that multiple polymers can be synthesized from the same monomers, and differences in stereochemistry. When considering the class of generic polymers, IUPAC polymer naming conventions are sufficient for uniquely describing most of these polymers, albeit with sometimes inconveniently long names.[282] Unfortunately, nomenclature standardization is made almost entirely intractable by the enormity of trade names that are used to describe common commercial polymers. For example, poly(methyl methacrylate) can be described by more than 928 trade names.[283] In this way, polymer property information may be effectively hidden behind an indiscernible proprietary name, complicating use for informatics processing. Moreover, while information on proprietary materials may be available under that specific trade name, information on

additives or processing techniques that affect the reported material properties may be intentionally omitted.

Polymer database design is further complicated by the inherently distributional nature of polymeric materials. Polymeric material properties are intrinsically distributional because polymer samples are very rarely monodisperse and are composed of macromolecular chains which exhibit some distribution in length. The dispersity of a polymer sample can have a significant impact on the measured properties of the sample. This has been known for decades with investigations on the effect of polydispersity on the linear viscoelastic properties of entangled linear polybutadiene rubber in 1985.[284] This dependence on dispersity, which in and of itself is a function of a distribution of weight-average molar mass and number-average molar mass, necessitates that measured properties should be accompanied by information on the dispersity of the polymer sample used in the measurement of the property or described by a distribution of values measured from samples of varying dispersity. As an additional example, as illustrated in the case study described below, glass transition temperature can vary significantly with number-average molecular weight—a relationship describable by the empirical Flory–Fox equation.[285] Any database that reports material properties without a requisite description of the dispersity fails to capture the full nature of the material property. Some databases, such as PoLyInfo, provide this data but its textual format compromises its machine-readability and makes it difficult to incorporate into informatics processing.

Beyond the dispersity of a sample, by omitting information on a sample's processing history, one fails to capture an essential aspect of polymeric materials—that properties can vary significantly with processing history and measurement methods. Many databases will provide values for fundamental physical properties such as density, heat capacity, and tensile modulus. However, it has been shown that even these basic properties will vary with processing history.[286,287] In much the same way, the

method of measurement will have an effect on the measured value of a property but very rarely will information on the measurement method be provided with the reported property value.[286] A notable exception to this trend is the commercial plastics database CAMPUSplastics which provides a notation on the International Standards Organization (ISO) standard used to acquire the reported property value.[288,289] ISO has developed more than 22,000 standards for measurement and production, many of which are available for free. The compliance to and citation of relevant ISO standards is an elegant method of condensing of- ten complex processing information into a compact and easily communicable format. However, the standards are not without their disadvantages—specifically their human-readable format. Organized in an extensive essay style format, the lack of process codification is difficult, if not impossible, to be interpreted in an algorithmic fashion. It is a boon to the field that such standards have been developed, but additional work must be done to ex- tract the most useful aspects of processing and measurement in a machine-readable format. Without this attendant metadata on how a property was measured, cohesiveness between databases is compromised and informatics techniques that rely on consistent and cohesive data are undermined.

6.1.2.3 The Future of Polymer Databases

Applying materials informatics processing techniques to polymer materials will be difficult until proper information organization can be accomplished; such organization is predicated on the development of universal standards for naming complex polymer architectures, adequately describing the distributional nature of their properties, and compactly and comprehensively reporting process history and testing methods. It has already been suggested that IUPAC naming conventions, while not comprehensive, need to be encouraged in all published database entries while these conventions are expanded and refined.[263] This relatively small addition to commercial plastics

databases will make property data visible for informatics engines. To improve processing and sample metadata, sample dispersity and average molecular weights could be reported for all property measurements to provide the proper context for the value. Additionally, citation of test and fabrication standards, such as those developed by the ISO, should be included in a standardized template for property measurement, so that more specific or variational metadata descriptors can be cited by researchers involved in novel characterization.

Beyond supplementing current databases with consistent naming schemes, processing, and sample metadata descriptors, researchers should be discouraged from generating more databases that curate one or two specific polymer properties to prevent the landscape from becoming too sprawling. Rather, when the aggregate recording of a new property becomes pressing, current databases such as PoLyInfo and MatWeb should have established avenues available for the modification and expansion of their already extensive data sets. Moreover, alongside the expansion and sophistication of these established databases, researchers should be encouraged to submit their data to these databases or, for enhanced long-term effectiveness, make the data available in a public, machine-readable format for its assimilation into these databases via automated interfaces. Human collation and entry into these databases, while useful, can never exceed the efficiency of computer-aided curation. Thus, priority needs to be given to formulating standardized, machine-readable formats for polymer informatics to enhance database development, expansion, and utility.

6.1.3 Ligand Design for Nanoparticles with Biological Interfaces

Nanoparticles (NPs) have shown great potential as versatile drug and gene-delivery platforms. Experimental and computational studies have demonstrated that hydrophobicity and charge are critical factors in controlling the biocompatibility and efficiency of NPs. However, more work is needed to understand the role that NP shape and surface patterning have on their biological

properties. Complex ligands, such as highly branched or multivalent, and mixed monolayers have shown improved efficiency, but the reasons for this are still unclear. Furthermore, the interactions between even simple NPs and the broader biological milieu are poorly understood, including responses to multivalent ions, pH, temperature, and small molecules.[290–293]

The use of in silico tools provides a way for quick, inexpensive screening of potential ligand designs. Simulations can predict the performance of both the final structure, precursors, and assembly conditions (such as solvent choice), thereby accelerating synthesis efforts. However, the vast design space of functionalized nanoparticles and the continuous increase in available computational power call for tools that go beyond simple statistical models to uncover complex and nonintuitive design principles. While machine-learning (ML) tools, such as artificial neural networks, have long been a focus in computer science, their use in the design of biomimetic materials has been less widespread. The availability of open-source software packages has made these tools accessible to the broader materials science community, but fundamental challenges centered on data generation, organization, and analysis, the shape of the hypothesis space, and interactions with experimental work require a tailored approach to the use of ML tools in in silico materials design. Improvements in search efficiency will be necessary to generate sufficient high-quality data for training ML models. Further improvements can be made by using a multiscale, multiresolution model. We believe that tight integration of ML tools into the simulation workflow will become an essential part of future high- throughput in silico materials design.

6.2 Convergence Informatics

* This section is a manuscript in preparation by:
Thomas J. Oweida, Alexey Gulyuk, Akhlak Mahmood, and Yaroslava G. Yingling.

6.2.1 Convergence Informatics

Nowadays it is impossible to imagine a single sphere of life that does not involve handling enormous information flows. While numerous databases have emerged to store data categorically, each database is not standardized as developers have implemented unique data gathering, data handling, and data interpretation processes.[59] Certain data-driven approaches were implemented recently, and this helped to overcome some bottlenecks of established workflows.[294,295] In materials informatics, this has led to rapid materials discovery.[44,296] However, despite the facilitated development of new materials, acceleration in the utility by industrial consumers is absent. In large part, this is due to the variability and incongruence of available data, which has limited the types of information that can be considered in individual materials optimization processes.

Here, we introduce the concept of convergence informatics as a comprehensive solution for robust engineering projects and provide an outlook on state-of-the-art data storage and processing tools that can serve as a platform to streamline the novel materials research and development. Convergence Informatics is defined as a set of holistic, data-driven approaches that can seamlessly analyze a wide range of data types that connect material's properties, performance in various environments, and its application-specific characteristics. The convergence informatics workflow first involves a traditional materials informatics approach that takes existing materials process-structure-property from available databases, and subsequently analyzes the data to accelerate the design of the new material. However, an additional component that is centered around downstream

data is integrated into the materials informatics approach to develop materials with immediate utility for specific applications.

This convergence informatics concept will require the further development of conventional tools in materials science, specifically as it relates to materials data. Currently, apart from being accumulated in various non-linked databases, experimental and computational data is stored in different formats such as numerical, text, graphical, or other types. Additionally, the data can be completely different in its nature as it pertains to the types of measurements performed. As an example, the data can be static (like an SEM scan), or time, unit-dependent (like dynamic measurements of XRD spectra). This type of data is inherent to materials science and it becomes challenging to link using data science tools, particularly when the data is stored in various formats. The variations in format and data types can be defined as heterogeneous data as it is not readily comparable to one another. The prominence of heterogeneous data in materials science is discussed below.

6.2.2 Heterogeneous Data

Materials science is a broad field that generates numerous types of data for the study of a single material. The type of data generated is dependent upon the characterization method used for study and can vary in terms of materials being measured (thin films, surface, or bulk properties), the environment conditions (ambient, gaseous, vacuum, aqueous, etc.), the resolution of measurement (electronic structure to mesoscale), and the time scale (static or time dependent). As a particular research example of how this heterogeneous data impacts studies in material science, we discuss the rational design of drug delivery vehicles. This topic attempts to connect molecular characteristics to macromolecular structure and performance. The first step is to characterize the base material through techniques such as FTIR, XPS, MS (helps to assess fundamental properties

like chemical composition), and Chromatography (provides the degree of polymerization profile). At the macromolecular scale the drug delivery, vehicle's shape, and size can play a critical role in circulation time and cell internalization.[60] Common characterization techniques to provide such information include SAXS, DLS, and TEM. Even though such methods are complementary to each other, each has a distinct underlying data model that is complex, high dimensional and difficult to combine with others. For instance, SAXS and DLS would help to assess the time averaged ensemble in solution, while TEM provides similar information through a static image of a small sample on a grid. As a result, datasets studying the same materials properties are not compatible with each other even when the target property of interest is the same.

In addition to the heterogeneity of measured properties in experiments, comparisons to theoretical work provides a further challenge. Namely, computational studies cannot account for all the variables present in the experiment. For example, in the context of drug delivery vehicles, polydispersity can alter the carrier's shape, size, and stability, making an objective assessment of material performance without this data difficult. However, polydispersity is often neglected in computational approaches even at the mesoscale level.[218] This inherently means that experiment and computations cannot easily be compared to one another, as the studied materials or environments are not exactly the same. Even if the materials could be made identical, the time or size scales are likely different due to the limitations of modern computers.[116,239] Although this limitation exists, time-averaged ensembles from computational studies at the nanosecond scale are compared against experiments whose measurements are on the order of seconds.[13,239] In addition, the molecular ensemble from computation would contain a single or at most a handful of drug carriers which are being compared to experimental measurements based on millions of molecules. Thus, detailed information about the variable nature of materials is often lost in computations.

It is evident that comparisons using different techniques can be challenging. At the same time, the materials comparison for the same characterization technique becomes a non-trivial task. For example, computational techniques have different levels of accuracy based on the basis sets, functional forms or force fields used.[20,239] While theoretical results need to be validated against experimental findings, the experimental data can also suffer from instrumental limitations, measurement noise, and failure to follow proper steps doing the measurements. This gets even more complicated when methods of experiments are not clearly published for the datasets or simply unreliable. Validation of results is not practical when the data available are collected in different experimental conditions. In many cases, exact experimental data acquiring conditions differ between each method of study and experimental validation of theoretical results becomes impossible due to difficulty or impossibility of direct measurements.

At this point, the handling of materials data for its performance in various environments and its application-specific characteristics have been ignored. In the scope of drug delivery vehicles, this encompasses drug carriers' data from industry and user specific data from clinical trials. With current approaches in materials science, it has been shown that *in-vitro*, bench-top studies do not translate well to *in vivo* results.[297] In large part this is due to the neglected effects of alternative stimuli in biological environments. As an example, we will look at the physical nature of controlling a drug's release, which is commonly achieved via molecular recognition or the material's response to changes in temperature or pH. While this information can be accounted for based on the chemical makeup of the material, it has been well documented that upon exposure to a biological environment, proteins adhere to the drug vehicle's surface, forming a complex and dynamic protein corona. This ultimately influences the drug carrier's circulation time, biodistribution, and targeting capabilities.[298] Furthermore, the dynamic flow experienced by a drug

carrier alters the composition of the protein corona, thus the drug carrier will behave differently depending whether it is circulating in arteries, veins, or capillaries.[298,299] In addition to flow rates, the consideration of material performance at biologically relevant temperature and pH is necessary. The organism's temperature can fluctuate with infection while pH naturally varies by location in the body or even by cell type given cancerous cells can be more acidic than native healthy cells.[300] Lastly, issues such as ease of synthesis, cost, and stability or shelf-life of a drug carrier can indicate if manufacturing is a worthy endeavor. Moving beyond material performance, the inclusion of user specific information such as a patient's clinical, genetic, genomic, and environmental information should be considered. In this instance, materials development and patient treatment can be optimized based on characteristics unique to individual populations.[301,302] Ultimately, with the correct handling of heterogeneous data spanning material's properties, its performance, and user data, a drug carrier will no longer be designed to optimize *in-vitro* properties, but will be optimized with respect to its *in-vivo* results instead.

6.2.3 Heterogeneous Data Collection and Management

Collecting this breadth of data requires involvement of academic and industrial experimentation installments ranging from studies of fundamental materials properties to clinical trials and application-based use. As mentioned earlier, the non-standardized development of databases provides a challenge in linking this data, but at a deeper level, data sharing is the most formidable challenge. Specifically, in industrial experimentation, proprietary or personal information is intentionally withheld making direct comparisons to alternative experiments difficult.[59,303] While complete data sharing continues to be pushed for at the academic level, industrial experiments will almost always contain missing data because of the exclusion of

proprietary data. This poses a challenge with identifying what data is missing or incomplete before data imputation strategies can be implemented.

As it can be understood from the discussion, the integration of data for a portfolio of materials, implemented in a wide range of devices, and utilized by various users becomes a non-trivial task. The datasets of different formats and from different sources (users, databases, data collection hubs etc.) must be implemented into a single set of heterogeneous data and processed accordingly. This may seem like a problem with no clear solution at this point. However, the concept of managing and analyzing the heterogeneous data, characterized by high variability of data types and formats, is not new for the data science community.[304] Thus, there are numerous tools and algorithms that have already been developed, resulting in most of the challenges being associated with proper data integration and curation. Data curation models for Big Data have been developed with the goals of maximizing the data quality and usability from multiple sources, however, such data management schemes in materials-related fields have not yet been implemented.[305] Furthermore, it is important to note that existing machine learning and data analytics methods for heterogeneous data still tend to fall short because of the diversity of database types, thus improvements to machine learning for heterogeneous data is an active area of research. Current research generally involves developing approaches that use Distributed Data Mining (DDM) and Collective Data Mining (CDM).[304] Although materials data has unique challenges embedded inside each dataset, it is imperative that the implementation of data-driven techniques are informed from the successes and failures in adjacent disciplines of study.

REFERENCES

1. Dobrynin, A. V., Colby, R. H. & Rubinstein, M. Scaling Theory of Polyelectrolyte Solutions. *Macromolecules* (1995) doi:10.1021/ma00110a021.

2. Colby, R. H. Structure and linear viscoelasticity of flexible polymer solutions: Comparison of polyelectrolyte and neutral polymer solutions. *Rheologica Acta* (2010) doi:10.1007/s00397-009-0413-5.

3. Dobrynin, A. V. & Rubinstein, M. Theory of polyelectrolytes in solutions and at surfaces. *Progress in Polymer Science (Oxford)* (2005) doi:10.1016/j.progpolymsci.2005.07.006.

4. Dobrynin, A. V. Theory and simulations of charged polymers: From solution properties to polymeric nanomaterials. *Current Opinion in Colloid and Interface Science* (2008) doi:10.1016/j.cocis.2008.03.006.

5. Pinheiro, A. V., Han, D., Shih, W. M. & Yan, H. Challenges and opportunities for structural DNA nanotechnology. *Nature Nanotechnology* (2011) doi:10.1038/nnano.2011.187.

6. Becker, A. L., Johnston, A. P. R. & Caruso, F. Layer-by-layer-assembled capsules and films for therapeutic delivery. *Small* (2010) doi:10.1002/smll.201000379.

7. Murphy, M. C., Rasnik, I., Cheng, W., Lohman, T. M. & Ha, T. Probing Single-Stranded DNA Conformational Flexibility Using Fluorescence Spectroscopy. *Biophys. J.* **86**, 2530–2537 (2004).

8. Kuznetsov, S. V., Shen, Y., Benight, A. S. & Ansari, A. A semiflexible polymer model applied to loop formation in DNA hairpins. *Biophys. J.* (2001) doi:10.1016/S0006-3495(01)75927-9.

9. Doose, S., Barsch, H. & Sauer, M. Polymer properties of polythymine as revealed by translational diffusion. *Biophys. J.* (2007) doi:10.1529/biophysj.107.107342.

10. Kim, H. S., Huang, S. M. & Yingling, Y. G. Sequence dependent interaction of single stranded DNA with graphitic flakes: atomistic molecular dynamics simulations. *MRS Adv.* **1**, 1883–1889 (2016).

11. Bao, L., Zhang, X., Jin, L. & Tan, Z. J. Flexibility of nucleic acids: From DNA to RNA. in *Chinese Physics B* vol. 25 (2015).

12. Sim, A. Y. L., Lipfert, J., Herschlag, D. & Doniach, S. Salt dependence of the radius of gyration and flexibility of single-stranded DNA in solution probed by small-angle x-ray scattering. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **86**, (2012).

13. Plumridge, A., Meisburger, S. P., Andresen, K. & Pollack, L. The impact of base stacking on the conformations and electrostatics of single-stranded DNA. *Nucleic Acids Res.* **45**, 3932–3943 (2017).

14. McIntosh, D. B., Duggan, G., Gouil, Q. & Saleh, O. A. Sequence-dependent elasticity and electrostatics of single-stranded DNA: Signatures of base-stacking. *Biophys. J.* (2014) doi:10.1016/j.bpj.2013.12.018.

15. Ke, C., Humeniuk, M., S-Gracz, H. & Marszalek, P. E. Direct measurements of base stacking interactions in DNA by single-molecule atomic-force spectroscopy. *Phys. Rev. Lett.* (2007) doi:10.1103/PhysRevLett.99.018302.

16. Foster, J. C. *et al.* 100th Anniversary of Macromolecular Science Viewpoint: The Role of Hydrophobicity in Polymer Phenomena. *ACS Macro Letters* vol. 9 1700–1707 (2020).

17. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).

18. Go, A. W. *et al.* Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs . 1 . Generalized Born. (2012).

19. Case, D. A. Amber 18. *Univ. California, San Fr.* (2018).

20. Peerless, J. S., Kwansa, A. L., Hawkins, B. S., Smith, R. C. & Yingling, Y. G. Uncertainty Quantification and Sensitivity Analysis of Partial Charges on Macroscopic Solvent Properties in Molecular Dynamics Simulations with a Machine Learning Model. *J. Chem. Inf. Model.* acs.jcim.0c01204 (2021) doi:10.1021/acs.jcim.0c01204.

21. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* (1995) doi:10.1016/0009-2614(95)01082-K.

22. Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A. & Onufriev, A. Generalized born model with a simple, robust molecular volume correction. *J. Chem. Theory Comput.* (2007) doi:10.1021/ct600085e.

23. Hoogerbrugge, P. J. & Koelman, J. M. V. A. Simulating microscopic hydrodynamic phenomena with dissipative particle dynamics. *EPL* (1992) doi:10.1209/0295-5075/19/3/001.

24. Schlijper, A. G., Hoogerbrugge, P. J. & Manke, C. W. Computer simulation of dilute polymer solutions with the dissipative particle dynamics method. *J. Rheol. (N. Y. N. Y).* (1995) doi:10.1122/1.550713.

25. Espanol, P. & Warren, P. Statistical mechanics of dissipative particle dynamics. *EPL* (1995) doi:10.1209/0295-5075/30/4/001.

26. Groot, R. D. & Warren, P. B. Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. *J. Chem. Phys.* **107**, 4423–4435 (1997).

27. Groot, R. D. & Madden, T. J. Dynamic simulation of diblock copolymer microphase separation. *J. Chem. Phys.* (1998) doi:10.1063/1.476300.

28. Groot, R. D., Madden, T. J. & Tildesley, D. J. On the role of hydrodynamic interactions in block copolymer microphase separation. *J. Chem. Phys.* (1999) doi:10.1063/1.478939.

29. Wijmans, C. M., Smit, B. & Groot, R. D. Phase behavior of monomeric mixtures and polymer solutions with soft interaction potentials. *J. Chem. Phys.* (2001) doi:10.1063/1.1362298.

30. Jiang, W., Huang, J., Wang, Y. & Laradji, M. Hydrodynamic interaction in polymer solutions simulated with dissipative particle dynamics. *J. Chem. Phys.* (2007) doi:10.1063/1.2428307.

31. Li, N. K., Fuss, W. H. & Yingling, Y. G. An Implicit Solvent Ionic Strength (ISIS) method to model polyelectrolyte systems with dissipative particle dynamics. *Macromol. Theory Simulations* (2015) doi:10.1002/mats.201400043.

32. Chan, H. *et al.* Machine learning coarse grained models for water. *Nat. Commun.* (2019) doi:10.1038/s41467-018-08222-6.

33. Chen, C.-T. & Gu, G. X. Composite Materials: Effect of Constituent Materials on Composite Performance: Exploring Design Strategies via Machine Learning (Adv. Theory Simul. 6/2019). *Adv. Theory Simulations* **2**, (2019).

34. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, (2016).

35. Hill, J. *et al.* Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bull.* **41**, 399–409 (2016).

36. Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nat. Mater.* 191–201 (2013) doi:10.1038/nmat3568.

37. Liu, Y. *et al.* Materials discovery and design using machine learning. *J. Mater.* **3**, 159–177 (2017).

38. Takahashi, K. & Tanaka, Y. Material synthesis and design from first principle calculations and machine learning. *Comput. Mater. Sci.* **112**, 364–367 (2016).

39. Zhao, L. R., Chen, K., Yang, Q., Rodgers, J. R. & Chiou, S. H. Materials informatics for the design of novel coatings. *Surf. Coatings Technol.* **200**, 1595–1599 (2005).

40. Zeng, S., Li, G., Zhao, Y., Wang, R. & Ni, J. Machine Learning-Aided Design of Materials with Target Elastic Properties. *J. Phys. Chem. C* **123**, 5042–5047 (2019).

41. Liu, R. *et al.* A predictive machine learning approach for microstructure optimization and materials design. *Sci. Rep.* **10**, (2015).

42. Srinivasan, S. *et al.* Mapping Chemical Selection Pathways for Designing Multicomponent Alloys: An informatics framework for materials design. *Sci. Rep.* (2015)

doi:10.1038/srep17960.

43. Kulik, H. J. Making machine learning a useful tool in the accelerated discovery of transition metal complexes. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* (2019) doi:10.1002/wcms.1439.

44. Kim, C., Pilania, G. & Ramprasad, R. Machine Learning Assisted Predictions of Intrinsic Dielectric Breakdown Strength of ABX3 Perovskites. *J. Phys. Chem. C* **120**, 14575–14580 (2016).

45. Nakata, H. & Bai, S. Development of a new parameter optimization scheme for a reactive force field based on a machine learning approach. *J. Comput. Chem.* **40**, 2000–2012 (2019).

46. Wang, P., Shao, Y., Wang, H. & Yang, W. Accurate interatomic force field for molecular dynamics simulation by hybridizing classical and machine learning potentials. *Extrem. Mech. Lett.* **24**, 1–5 (2018).

47. Chen, C. *et al.* Accurate force field for molybdenum by machine learning large materials data. *Phys. Rev. Mater.* **1**, (2017).

48. Botu, V. & Ramprasad, R. Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B - Condens. Matter Mater. Phys.* **92**, (2015).

49. Wood, M. A., Cusentino, M. A., Wirth, B. D. & Thompson, A. P. Data-driven material models for atomistic simulation. *Phys. Rev. B* **99**, (2019).

50. Bleiziffer, P., Schaller, K. & Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **58**, 579–590 (2018).

51. Chmiela, S., Sauceda, H. E., Müller, K. R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* (2018) doi:10.1038/s41467-018-06169-2.

52. Li, Y. *et al.* Machine Learning Force Field Parameters from Ab Initio Data. *J. Chem. Theory Comput.* **13**, 4492–4503 (2017).

53. Huan, T. D. *et al.* A universal strategy for the creation of machine learning-based atomistic force fields. *npj Comput. Mater.* (2017) doi:10.1038/s41524-017-0042-y.

54. Miles, P., Leon, L., Smith, R. C. & Oates, W. S. Analysis of a multi-axial quantum informed ferroelectric continuum model: Part 1—uncertainty quantification. *J. Intell. Mater. Syst. Struct.* **29**, 2823–2839 (2018).

55. Leon, L., Smith, R. C., Oates, W. S. & Miles, P. Analysis of a multi-axial quantum-informed ferroelectric continuum model: Part 2—sensitivity analysis. *J. Intell. Mater. Syst. Struct.* **29**, 2840–2860 (2018).

56. Paterson, A. R., Reich, B. J., Smith, R. C., Wilson, A. G. & Jones, J. L. Bayesian approaches to uncertainty quantification and structure refinement from X-ray diffraction. in *Springer Series in Materials Science* 81–102 (2018). doi:10.1007/978-3-319-99465-9_4.

57. Xu, W. & LeBeau, J. M. A Convolutional Neural Network Approach to Thickness Determination using Position Averaged Convergent Beam Electron Diffraction. *Microsc. Microanal.* **23**, (2017).

58. Venkatraman, V. & Alsberg, B. Designing High-Refractive Index Polymers Using Materials Informatics. *Polymers (Basel).* (2018) doi:10.3390/polym10010103.

59. Peerless, J. S., Milliken, N. J. B., Oweida, T. J., Manning, M. D. & Yingling, Y. G. Soft Matter Informatics: Current Progress and Challenges. *Adv. Theory Simulations* **2**, 1–12 (2019).

60. Manning, M. D. *et al.* Progress in ligand design for monolayer-protected nanoparticles for nanobio interfaces. *Biointerphases* **13**, (2018).

61. Nash, J. A., Kwansa, A. L., Peerless, J. S., Kim, H. S. & Yingling, Y. G. Advances in molecular modeling of nanoparticle-nucleic acid interfaces. *Bioconjug. Chem.* **28**, 3–10 (2017).

62. Li, N. K. *et al.* Prediction of solvent-induced morphological changes of polyelectrolyte diblock copolymer micelles. *Soft Matter* **11**, 8236–8245 (2015).

63. Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).

64. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101 (1989).

65. Lin, T.-S. *et al.* BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **5**, 1523–1531 (2019).

66. De Guire, E. *et al.* Data-driven glass/ceramic science research: Insights from the glass and ceramic and data science/informatics communities. *J. Am. Ceram. Soc.* **102**, 6385–6406 (2019).

67. Kononova, O. *et al.* Text-mined dataset of inorganic materials synthesis recipes. *Sci. data* (2019) doi:10.1038/s41597-019-0224-1.

68. Berman, H. M. *et al.* The Protein Data Bank (www.rcsb.org). *Nucleic Acids Res.* (2000) doi:10.1093/nar/28.1.235.

69. Bernstein, F. C. *et al.* The Protein Data Bank. *Eur. J. Biochem.* **80**, 319–324 (1977).

70. Burley, S. K. *et al.* RCSB Protein Data Bank: Biological macromolecular structures

enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2019).

71. Source: National Institute for Materials Science. https://www.nims.go.jp/eng/.

72. Villars, P. *et al.* The Pauling File, Binaries Edition. in *Journal of Alloys and Compounds* (2004). doi:10.1016/j.jallcom.2003.08.058.

73. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. PoLyInfo: Polymer database for polymeric materials design. in *Proceedings - 2011 International Conference on Emerging Intelligent Data and Web Technologies, EIDWT 2011* (2011). doi:10.1109/EIDWT.2011.13.

74. Nakamura, S., Yang, H., Hirata, C., Kersaudy, F. & Fujimoto, K. Development of 19F-NMR chemical shift detection of DNA B-Z equilibrium using 19F-NMR. *Org. Biomol. Chem.* (2017) doi:10.1039/c7ob00706j.

75. Chen, Z., Liu, C., Cao, F., Ren, J. & Qu, X. DNA metallization: Principles, methods, structures, and applications. *Chemical Society Reviews* (2018) doi:10.1039/c8cs00011e.

76. Rafiee-Pour, H. A., Behpour, M. & Keshavarz, M. A novel label-free electrochemical miRNA biosensor using methylene blue as redox indicator: Application to breast cancer biomarker miRNA-21. *Biosens. Bioelectron.* (2016) doi:10.1016/j.bios.2015.09.025.

77. Diculescu, V. C., Chiorcea-Paquim, A. M. & Oliveira-Brett, A. M. Applications of a DNA-electrochemical biosensor. *TrAC - Trends in Analytical Chemistry* (2016) doi:10.1016/j.trac.2016.01.019.

78. Liu, B., Salgado, S., Maheshwari, V. & Liu, J. DNA adsorbed on graphene and graphene oxide: Fundamental interactions, desorption and applications. *Current Opinion in Colloid and Interface Science* (2016) doi:10.1016/j.cocis.2016.09.001.

79. Wei, B., Dai, M. & Yin, P. Complex shapes self-assembled from single-stranded DNA tiles. *Nature* (2012) doi:10.1038/nature11075.

80. Sun, H. & Zu, Y. Aptamers and their applications in nanomedicine. *Small* (2015) doi:10.1002/smll.201403073.

81. Li, D. *et al.* When biomolecules meet graphene: From molecular level interactions to material design and applications. *Nanoscale* (2016) doi:10.1039/c6nr07249f.

82. Xi, Z. *et al.* Selection of HBsAg-specific DNA aptamers based on carboxylated magnetic nanoparticles and their application in the rapid and simple detection of hepatitis b virus infection. *ACS Appl. Mater. Interfaces* (2015) doi:10.1021/acsami.5b01180.

83. Liu, M. *et al.* Aptamer selection and applications for breast cancer diagnostics and therapy. *Journal of Nanobiotechnology* (2017) doi:10.1186/s12951-017-0311-4.

84. Bosco, A., Camunas-Soler, J. & Ritort, F. Elastic properties and secondary structure formation of single-stranded DNA at monovalent and divalent salt conditions. *Nucleic Acids Res.* (2014) doi:10.1093/nar/gkt1089.

85. Hammel, M. Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS). 789–799 (2012) doi:10.1007/s00249-012-0820-x.

86. Choi, S. J. & Ban, C. Crystal structure of a DNA aptamer bound to PvLDH elucidates novel single-stranded DNA structural elements for folding and recognition. *Sci. Rep.* (2016) doi:10.1038/srep34998.

87. Chen, X. *et al.* Hairpins are formed by the single DNA strands of the fragile X triplet repeats: Structure and biological implications. *Proc. Natl. Acad. Sci. U. S. A.* (1995) doi:10.1073/pnas.92.11.5199.

88. Braddock, D. T., Baber, J. L., Levens, D. & Clore, G. M. Molecular basis of sequence-specific single-stranded DNA recognition by KH domains: Solution structure of a complex between hnRNP K KH3 and single-stranded DNA. *EMBO J.* (2002) doi:10.1093/emboj/cdf352.

89. Scholl, Z. N. *et al.* Origin of Overstretching Transitions in Single-Stranded Nucleic Acids. **188302**, 1–5 (2013).

90. Santhana Mariappan, S. V. *et al.* Solution Structures of the Individual Single Strands of the Fragile X DNA Triplets (GCC)n{middle dot}(GGC)n. *Nucleic Acids Res.* (1996) doi:10.1093/nar/24.4.784.

91. Kang, J., Jung, J. & Kim, S. K. Flexibility of single-stranded DNA measured by single-molecule FRET. *Biophysical Chemistry* (2014) doi:10.1016/j.bpc.2014.08.004.

92. Laurence, T. A., Kong, X. & Ja, M. Probing structural heterogeneities and fluctuations of nucleic acids and denatured proteins. (2005).

93. Meisburger, S. P. *et al.* Polyelectrolyte properties of single stranded DNA measured using SAXS and single-molecule FRET: Beyond the wormlike chain model. *Biopolymers* **99**, 1032–1045 (2013).

94. Rambo, R. P. & Tainer, J. A. Characterizing Flexible and Intrinsically Unstructured Biological Macromolecules by SAS Using the Porod-Debye Law. **95**, (2011).

95. Boldon, L., Laliberte, F. & Liu, L. relevant integrated application. **1**, 1–21 (2015).

96. Mylonas, E. & Petoukhov, M. V. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. 5656–5664 (2007) doi:10.1021/ja069124n.

97. Meisburger, S. P., Pabit, S. A. & Pollack, L. Determining the Locations of Ions and Water around DNA from X-Ray Scattering Measurements. *Biophysj* **108**, 2886–2895 (2015).

98. Chen, H. *et al.* Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc. Natl. Acad. Sci.* **109**, 799–804 (2012).

99. Rechendorff, K. *et al.* Persistence length and scaling properties of single-stranded DNA adsorbed on modified graphite Persistence length and scaling properties of single-stranded DNA adsorbed. **095103**, (2013).

100. Li, N. K., Kim, H. S., Nash, J. A., Lim, M. & Yingling, Y. G. Progress in molecular modelling of DNA materials. *Mol. Simul.* **40**, 777–783 (2014).

101. Galindo-murillo, R. *et al.* Assessing the Current State of Amber Force Field Modifications for DNA. (2016) doi:10.1021/acs.jctc.6b00186.

102. Gaillard, T. & Case, D. A. Evaluation of DNA Force Fields in Implicit Solvation. 3181–3198 (2011) doi:10.1021/ct200384r.

103. Guy, A. T., Piggot, T. J. & Khalid, S. Single-Stranded DNA within Nanopores : Conformational Dynamics and Implications for Sequencing ; a Molecular Dynamics Simulation Study. **103**, 1028–1036 (2012).

104. Gu, R., Oweida, T., Yingling, Y. G., Chilkoti, A. & Zauscher, S. Enzymatic Synthesis of Nucleobase-Modified Single-Stranded DNA Offers Tunable Resistance to Nuclease Degradation. *Biomacromolecules* (2018) doi:10.1021/acs.biomac.8b00816.

105. Johnson, R. R., Johnson, a T. C. & Klein, M. L. Probing the Structure of DNA− Carbon Nanotube Hybrids with Molecular Dynamics. *Nano Lett.* **8**, 69–75 (2008).

106. Vilhena, J. G. *et al.* Stick–Slip Motion of ssDNA over Graphene. *J. Phys. Chem. B* acs.jpcb.7b06952 (2017) doi:10.1021/acs.jpcb.7b06952.

107. Andrews, C. T., Campbell, B. A. & Elcock, A. H. Direct Comparison of Amino Acid and Salt Interactions with Double-Stranded and Single-Stranded DNA from Explicit-Solvent Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **13**, 1794–1811 (2017).

108. Prior, C., Danilane, L. & Oganesyan, V. S. All-atom molecular dynamics simulations of spin labelled double and single-strand DNA for EPR studies. *Phys. Chem. Chem. Phys.* (2018) doi:10.1039/c7cp08625c.

109. Nash, J. A., Tucker, T. L., Therriault, W. & Yingling, Y. G. Binding of single stranded nucleic acids to cationic ligand functionalized gold nanoparticles. *Biointerphases* **11**, (2016).

110. Oprzeska-Zingrebe, E. A. & Smiatek, J. Preferential Binding of Urea to Single-Stranded DNA Structures: A Molecular Dynamics Study. *Biophys. J.* (2018) doi:10.1016/j.bpj.2018.02.013.

111. Kim, H. S., Farmer, B. L. & Yingling, Y. G. Effect of Graphene Oxidation Rate on Adsorption of Poly-Thymine Single Stranded DNA. *Adv. Mater. Interfaces* (2017)

doi:10.1002/admi.201601168.

112. Fan, J., Zhang, H., Mu, Y. & Zheng, Q. Studies the Recognition Mechanism of TcaR and ssDNA Using Molecular Dynamic Simulations. *J. Mol. Graph. Model.* (2017) doi:10.1016/j.jmgm.2017.12.001.

113. Zgarbová, M. *et al.* Toward improved description of DNA backbone: Revisiting epsilon and zeta torsion force field parameters. *J. Chem. Theory Comput.* (2013) doi:10.1021/ct400154j.

114. Zgarbová, M. *et al.* Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.* (2015) doi:10.1021/acs.jctc.5b00716.

115. Chen, P. & Hub, J. S. Interpretation of Solution X-Ray Scattering by Explicit-Solvent Molecular Dynamics. *Biophysj* **108**, 2573–2584 (2015).

116. Martin Pelikan, Greg L. Hura, M. H. Structure and flexibility within proteins as identified through small angle X-ray scattering. **28**, 174–189 (2013).

117. Leontis, N. B., SantaLucia, J. & Staff, A. C. S. *Molecular modeling of nucleic acids*. (ACS Publications, 1998).

118. BIOvIA, D. S. Discovery studio modeling environment. (2015).

119. Cheatham, T. E., Cieplak, P. & Kollman, P. A. A modified version of the cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* (1999) doi:10.1080/07391102.1999.10508297.

120. Pérez, A. *et al.* Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophys. J.* (2007) doi:10.1529/biophysj.106.097782.

121. Ivani, I. *et al.* Parmbsc1: A refined force field for DNA simulations. *Nat. Methods* (2015) doi:10.1038/nmeth.3658.

122. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* (1996) doi:10.1021/jp961710n.

123. Tsui, V. & Case, D. A. Theory and applications of the Generalized Born solvation model in macromolecular simulations. *Biopolymers* (2000) doi:10.1002/1097-0282(2000)56:4<275::AID-BIP10024>3.0.CO;2-E.

124. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* (1983) doi:10.1063/1.445869.

125. Joung, I. S. & Cheatham, T. E. Determination of alkali and halide monovalent ion

parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* (2008) doi:10.1021/jp8001614.

126. Joung, S. & Cheatham, T. E. Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J. Phys. Chem. B* (2009) doi:10.1021/jp902584c.

127. Onufriev, A., Bashford, D. & Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins Struct. Funct. Genet.* (2004) doi:10.1002/prot.20033.

128. Nguyen, H., Pérez, A., Bermeo, S. & Simmerling, C. Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and Their Complexes with Proteins. *J. Chem. Theory Comput.* (2015) doi:10.1021/acs.jctc.5b00271.

129. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* (2016) doi:10.1093/nar/gkw389.

130. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* (2013) doi:10.1016/j.bpj.2013.07.020.

131. Plumridge, A., Meisburger, S. P. & Pollack, L. Visualizing single-stranded nucleic acids in solution. **45**, 1–13 (2017).

132. Humphrey, W., Dalke, A. & Schulten, K. VMD - Visual Molecular Dynamics. 33–38 (1996).

133. Kim, H. S., Oweida, T. J. & Yingling, Y. G. Interfacial stability of graphene-based surfaces in water and organic solvents. *J. Mater. Sci.* **53**, 5766–5776 (2018).

134. Keniry, M. A., Owen, E. A. & Shafer, R. H. The contribution of thymine-thymine interactions to the stability of folded dimeric quadruplexes. *Nucleic Acids Res.* (1997) doi:10.1093/nar/25.21.4389.

135. Yoo, J., Winogradoff, D. & Aksimentiev, A. Molecular dynamics simulations of DNA–DNA and DNA–protein interactions. *Curr. Opin. Struct. Biol.* **64**, 88–96 (2020).

136. Galindo-Murillo, R., Roe, D. R. & Cheatham, T. E. Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochim. Biophys. Acta - Gen. Subj.* (2015) doi:10.1016/j.bbagen.2014.09.007.

137. Oweida, T. J., Mahmood, A., Manning, M. D., Rigin, S. & Yingling, Y. G. Merging Materials and Data Science: Opportunities, Challenges, and Education in Materials Informatics. *MRS Adv.* (2020) doi:10.1557/adv.2020.171.

138. Kim, H. S. Understanding the Role of Surface and Solvent on Biomolecular Structure and

Dynamics. (North Carolina State Univeristy, 2017).

139. Zheng, D., Seferos, D. S., Giljohann, D. A., Patel, P. C. & Mirkin, C. A. Aptamer nano-flares for molecular detection in living cells. *Nano Lett.* **9**, 3258–3261 (2009).

140. Seferos, D. S., Giljohann, D. A., Hill, H. D., Prigodich, A. E. & Mirkin, C. A. Nano-flares: probes for transfection and mRNA detection in living cells. *J. Am. Chem. Soc.* **129**, 15477–9 (2007).

141. Halo, T. L. *et al.* NanoFlares for the detection, isolation, and culture of live tumor cells from human blood. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 17104–17109 (2014).

142. Rosi, N. L. *et al.* Oligonucleotide-modified gold nanoparticles for intracellular gene regulation. *Science* **312**, 1027–1030 (2006).

143. Calabrese, C. M. *et al.* Biocompatible Infinite-Coordination-Polymer Nanoparticle-Nucleic-Acid Conjugates for Antisense Gene Regulation. *Angew. Chemie* **127**, 486–490 (2015).

144. Zhang, C. *et al.* Biodegradable DNA-Brush Block Copolymer Spherical Nucleic Acids Enable Transfection Agent-Free Intracellular Gene Regulation. *Small* **11**, 5360–5368 (2015).

145. Young, K. L. *et al.* Hollow Spherical Nucleic Acids for Intracellular Gene Regulation Based upon Biocompatible Silica Shells. *Nano Lett.* **12**, 3867–3871 (2012).

146. Bagalkot, V., Farokhzad, O. C., Langer, R. & Jon, S. An aptamer-doxorubicin physical conjugate as a novel targeted drug-delivery platform. *Angew. Chemie - Int. Ed.* **45**, 8149–8152 (2006).

147. Kortylewski, M. *et al.* In vivo delivery of siRNA to immune cells by conjugation to a TLR9 agonist enhances antitumor immune responses. *Nat Biotechnol* **27**, 925–932 (2009).

148. Xu, W. *et al.* Aptamer-conjugated and doxorubicin-loaded unimolecular micelles for targeted therapy of prostate cancer. *Biomaterials* **34**, 5244–5253 (2013).

149. Chang, M., Yang, C.-S. & Huang, D.-M. Aptamer-Conjugated DNA Icosahedral Nanoparticles As a Carrier of Doxorubicin for Cancer Therapy. *ACS Nano* **5**, 6156–6163 (2011).

150. Zhu, G. *et al.* Self-Assembled Aptamer-Based Drug Carriers for Bispecific Cytotoxicity to Cancer Cells. *Chem. - An Asian J.* **7**, 1630–1636 (2012).

151. Zhu, G. *et al.* Noncanonical Self-Assembly of Multifunctional DNA Nanoflowers for Biomedical Applications. *J. Am. Chem. Soc.* **135**, 16438–16445 (2013).

152. Zhu, G. *et al.* Self-assembled, aptamer-tethered DNA nanotrains for targeted transport of molecular drugs in cancer theranostics. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 7998–8003

(2013).

153. Dagle, J. M., Weeks, D. L. & Walder, J. A. Pathways of degradation and mechanism of action of antisense oligonucleotides in Xenopus laevis embryos. *Antisense Res. Dev.* **1**, 11–20 (1991).

154. Eder, P. S., DeVine, R. J., Dagle, J. M. & Walder, J. A. Substrate specificity and kinetics of degradation of antisense oligonucleotides by a 3' exonuclease in plasma. *Antisense Res. Dev.* **1**, 141–151 (1991).

155. Seferos, D. S., Prigodich, A. E., Giljohann, D. A., Patel, P. C. & Mirkin, C. A. Polyvalent DNA nanoparticle conjugates stabilize nucleic acids. *Nano Lett.* **9**, 308–311 (2009).

156. Cutler, J. I. *et al.* Polyvalent nucleic acid nanostructures. *J. Am. Chem. Soc.* **133**, 9254–9257 (2011).

157. Giljohann, D. A., Seferos, D. S., Prigodich, A. E., Patel, P. C. & Mirkin, C. A. Gene regulation with polyvalent siRNA-nanoparticle conjugates. *J. Am. Chem. Soc.* **131**, 2072–2073 (2009).

158. Rush, A. M., Thompson, M. P., Tatro, E. T. & Gianneschi, N. C. Nuclease-resistant DNA via high-density packing in polymeric micellar nanoparticle coronas. *ACS Nano* **7**, 1379–1387 (2013).

159. Tan, X. *et al.* Light-Triggered, Self-Immolative Nucleic Acid-Drug Nanostructures. *J. Am. Chem. Soc.* **137**, 6112–6115 (2015).

160. Ni, S. *et al.* Chemical Modifications of Nucleic Acid Aptamers for Therapeutic Purposes. *Int. J. Mol. Sci.* **18**, 1683 (2017).

161. Watts, J. K., Deleavey, G. F. & Damha, M. J. Chemically modified siRNA: tools and applications. *Drug Discov. Today* **13**, 842–855 (2008).

162. Selvam, C., Mutisya, D., Prakash, S., Ranganna, K. & Thilagavathi, R. Therapeutic potential of chemically modified siRNA: Recent trends. *Chem. Biol. Drug Des.* **90**, 665–678 (2017).

163. Brautigam, C. A. & Steitz, T. A. Structural principles for the inhibition of the 3'-5' exonuclease activity of Escherichia coli DNA polymerase I by phosphorothioates. *J. Mol. Biol.* **277**, 363–377 (1998).

164. Campbell, V. W. & Jackson, D. A. The effect of divalent cations on the mode of action of DNase I. The initial reaction products produced from covalently closed circular DNA. *J. Biol. Chem.* **255**, 3726–3735 (1980).

165. Nakamaye, K. L. & Eckstein, F. Inhibition of restriction endonuclease Nel I cleavage by phosphorothioate groups and its application to oligonucleotide-directed mutagenesis. *Nucleic Acids Res.* **14**, 9679–9698 (1986).

166. Spitzer, S. & Eckstein, F. Inhibition of deoxyribonucleases by phosphorothioate groups in oligodeoxyribonucleotides. *Nucleic Acids Res.* **16**, 11691–11704 (1988).

167. Yang, X. *et al.* Gene silencing activity of siRNA molecules containing phosphorodithioate substitutions. *ACS Chem. Biol.* **7**, 1214–1220 (2012).

168. Prakash, T. P., Kawasaki, a M., Fraser, a S., Vasquez, G. & Manoharan, M. Synthesis of 2 '-O-[2-[(N,N-Dimethylamino)Oxy]Ethyl] Modified Nucleosides and Oligonucleotides. *J. Org. Chem.* **67**, 357–369 (2002).

169. Keefe, A. D., Pai, S. & Ellington, A. Aptamers as therapeutics. *Nat. Rev. Drug Discov.* **9**, 537–550 (2010).

170. Shigdar, S. *et al.* Aptamers as theranostic agents: modifications, serum stability and functionalisation. *Sensors (Basel).* **13**, 13624–13637 (2013).

171. Xiang, D. *et al.* Nucleic acid aptamer-guided cancer therapeutics and diagnostics: the next generation of cancer medicine. *Theranostics* **5**, 23–42 (2015).

172. Sun, H. & Zu, Y. A Highlight of Recent Advances in Aptamer Technology and Its Application. *Molecules* **20**, 11959–11980 (2015).

173. Lakhin, A. V, Tarantul, V. Z. & Gening, L. V. Aptamers: problems, solutions and prospects. *Acta Naturae* **5**, 34–43 (2013).

174. Wu, S. Y. *et al.* 2′-OMe-phosphorodithioate-modified siRNAs show increased loading into the RISC complex and enhanced anti-tumour activity. *Nat. Commun.* **5**, 3459 (2014).

175. Fauster, K. *et al.* 2′-Azido RNA, a Versatile Tool for Chemical Biology: Synthesis, X-ray Structure, siRNA Applications, Click Labeling. *ACS Chem. Biol.* **7**, 581–589 (2012).

176. Martínez-Montero, S. *et al.* Locked 2′-Deoxy-2′,4′-Difluororibo Modified Nucleic Acids: Thermal Stability, Structural Studies, and siRNA Activity. *ACS Chem. Biol.* **10**, 2016–2023 (2015).

177. Kurreck, J., Wyszko, E., Gillen, C. & Erdmann, V. A. Design of antisense oligonucleotides stabilized by locked nucleic acids. *Nucleic Acids Res.* **30**, 1911–1918 (2002).

178. Schmidt, K. S. *et al.* Application of locked nucleic acids to improve aptamer in vivo stability and targeting function. *Nucleic Acids Res.* **32**, 5757–5765 (2004).

179. Elm n, J. *et al.* Locked nucleic acid (LNA) mediated improvements in siRNA stability and functionality. *Nucleic Acids Res.* **33**, 439–447 (2005).

180. Mook, O. R., Baas, F., de Wissel, M. B. & Fluiter, K. Evaluation of locked nucleic acid-modified small interfering RNA in vitro and in vivo. *Mol. Cancer Ther.* **6**, 833–843 (2007).

181. Tjong, V., Yu, H., Hucknall, A., Rangarajan, S. & Chilkoti, A. Amplified on-chip fluorescence detection of DNA hybridization by surface-initiated enzymatic polymerization. *Anal. Chem.* **83**, 5153–5159 (2011).

182. Hollenstein, M. & Marcel. Nucleoside Triphosphates — Building Blocks for the Modification of Nucleic Acids. *Molecules* **17**, 13569–13591 (2012).

183. Tang, L. *et al.* Enzymatic polymerization of high molecular weight DNA amphiphiles that self-assemble into star-like micelles. *Adv. Mater.* **26**, 3050–3054 (2014).

184. Hocek, M. Synthesis of Base-Modified 2′-Deoxyribonucleoside Triphosphates and Their Use in Enzymatic Synthesis of Modified DNA for Applications in Bioanalysis and Chemical Biology. *J. Org. Chem.* **79**, 9914–9921 (2014).

185. Balintová, J. *et al.* Anthraquinone as a Redox Label for DNA: Synthesis, Enzymatic Incorporation, and Electrochemistry of Anthraquinone-Modified Nucleosides, Nucleotides, and DNA. *Chem. - A Eur. J.* **17**, 14063–14073 (2011).

186. Balintová, J. *et al.* Benzofurazane as a New Redox Label for Electrochemical Detection of DNA: Towards Multipotential Redox Coding of DNA Bases. *Chem. - A Eur. J.* **19**, 12720–12731 (2013).

187. Riedl, J., Pohl, R., Rulíšek, L. & Hocek, M. Synthesis and Photophysical Properties of Biaryl-Substituted Nucleos(t)ides. Polymerase Synthesis of DNA Probes Bearing Solvatochromic and pH-Sensitive Dual Fluorescent and ^{19} F NMR Labels. *J. Org. Chem.* **77**, 1026–1044 (2012).

188. Riedl, J. *et al.* Labelling of nucleosides and oligonucleotides by solvatochromic 4-aminophthalimide fluorophore for studying DNA–protein interactions. *Chem. Sci.* **3**, 2797 (2012).

189. Riedl, J. *et al.* GFP-like Fluorophores as DNA Labels for Studying DNA–Protein Interactions. *J. Org. Chem.* **77**, 8287–8293 (2012).

190. Lee, K. Y. *et al.* Bioimaging of Nucleolin Aptamer-Containing 5-(*N* -benzylcarboxyamide)-2′-deoxyuridine More Capable of Specific Binding to Targets in Cancer Cells. *J. Biomed. Biotechnol.* **2010**, 1–9 (2010).

191. Kimoto, M., Yamashige, R., Matsunaga, K., Yokoyama, S. & Hirao, I. Generation of high-affinity DNA aptamers using an expanded genetic alphabet. *Nat. Biotechnol.* **31**, 453–457 (2013).

192. Stovall, G. M. *et al.* In vitro selection using modified or unnatural nucleotides. *Curr. Protoc. Nucleic Acid Chem.* **56**, 9.6.1-9.6.33 (2014).

193. Sefah, K. *et al.* In vitro selection with artificial expanded genetic information systems. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 1449–1454 (2014).

194. Yang, Z. *et al.* Conversion Strategy Using an Expanded Genetic Alphabet to Assay Nucleic Acids. *Anal. Chem.* **85**, 4705–4712 (2013).

195. Chumakov, A. M., Yuhina, E. S., Frolova, E. I., Kravchenko, J. E. & Chumakov, S. P. Expanding the application potential of DNA aptamers by their functionalization. *Russ. J. Bioorganic Chem.* **42**, 1–13 (2016).

196. Macíčková-Cahová, H. & Hocek, M. Cleavage of adenine-modified functionalized DNA by type II restriction endonucleases. *Nucleic Acids Res.* **37**, 7612–7622 (2009).

197. Macíčková-Cahová, H., Pohl, R. & Hocek, M. Cleavage of Functionalized DNA Containing 5-Modified Pyrimidines by Type II Restriction Endonucleases. *ChemBioChem* **12**, 431–438 (2011).

198. Mačková, M., Boháčová, S., Perlíková, P., Poštová Slavětínská, L. & Hocek, M. Polymerase Synthesis and Restriction Enzyme Cleavage of DNA Containing 7-Substituted 7-Deazaguanine Nucleobases. *ChemBioChem* **16**, 2225–2236 (2015).

199. Bollum, F. J. 5. Terminal Deoxynucleotidyl Transferase. *Enzym.* **10**, 145–171 (1974).

200. Tang, L., Navarro, L. A., Chilkoti, A. & Zauscher, S. High-Molecular-Weight Polynucleotides by Transferase-Catalyzed Living Chain-Growth Polycondensation. *Angew. Chemie Int. Ed.* **56**, 6778–6782 (2017).

201. *Molecular Modeling of Nucleic Acids*. vol. 682 (American Chemical Society, 1997).

202. Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Release 2017, San Diego: Dassault Systèmes, 2016.

203. Vanquelef, E. *et al.* R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res.* **39**, W511–W517 (2011).

204. Li, N. K., Kim, H. S., Nash, J. A., Lim, M. & Yingling, Y. G. Progress in molecular modelling of DNA materials. *Mol. Simul.* **40**, 777–783 (2014).

205. 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.* **25**, 1157–1174 (2004).

206. Railsback, J. G. *et al.* Weakly Charged Cationic Nanoparticles Induce DNA Bending and Strand Separation. *Adv. Mater.* **24**, 4261–4265 (2012).

207. Chen, H. *et al.* Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 799–804 (2012).

208. Case, D. A. *et al.* Amber 2017 Reference Manual. *AMBER 17* Unversity of California, San Francisco (2017).

209. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

210. Langer, P. R., Waldrop, A. A. & Ward, D. C. Enzymatic synthesis of biotin-labeled polynucleotides: novel nucleic acid affinity probes. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 6633–7 (1981).

211. Gierlich, J. *et al.* Synthesis of Highly Modified DNA by a Combination of PCR with Alkyne-Bearing Triphosphates and Click Chemistry. *Chem. - A Eur. J.* **13**, 9486–9494 (2007).

212. Yu, H. *et al.* Cyanine dye dUTP analogs for enzymatic labeling of DNA probes. *Nucleic Acids Res.* **22**, 3226–3232 (1994).

213. Marciel, A. B., Mai, D. J. & Schroeder, C. M. Template-Directed Synthesis of Structurally Defined Branched Polymers. *Macromolecules* **48**, 1296–1303 (2015).

214. Selleri, L. *et al.* Detection and characterization of "Chimeric" yeast artificial chromosome clones by fluorescent in Situ suppression hybridization. *Genomics* **14**, 536–541 (1992).

215. Horáková, P. *et al.* Tail-labelling of DNA probes using modified deoxynucleotide triphosphates and terminal deoxynucleotidyl tranferase. Application in electrochemical DNA hybridization and protein-DNA binding assays. *Org. Biomol. Chem.* **9**, 1366 (2011).

216. Röthlisberger, P. *et al.* Facile immobilization of DNA using an enzymatic his-tag mimic. *Chem. Commun.* **53**, 13031–13034 (2017).

217. Korada, S. K. C. *et al.* Crystal structures of Escherichia coli exonuclease I in complex with single-stranded DNA provide insights into the mechanism of processive digestion. *Nucleic Acids Res.* **41**, 5887–5897 (2013).

218. Li, N. K. *et al.* Prediction of solvent-induced morphological changes of polyelectrolyte diblock copolymer micelles. *Soft Matter* **11**, 8236–8245 (2015).

219. Kielkowski, P., Fanfrlík, J. & Hocek, M. 7-Aryl-7-deazaadenine 2′-Deoxyribonucleoside Triphosphates (dNTPs): Better Substrates for DNA Polymerases than dATP in Competitive Incorporations. *Angew. Chemie Int. Ed.* **53**, 7552–7555 (2014).

220. Cahová, H., Panattoni, A., Kielkowski, P., Fanfrlík, J. & Hocek, M. 5-Substituted Pyrimidine and 7-Substituted 7-Deazapurine dNTPs as Substrates for DNA Polymerases in Competitive Primer Extension in the Presence of Natural dNTPs. *ACS Chem. Biol.* **11**, 3165–3171 (2016).

221. Hottin, A. & Marx, A. Structural Insights into the Processing of Nucleobase-Modified Nucleotides by DNA Polymerases. *Acc. Chem. Res.* **49**, 418–427 (2016).

222. Kropp, H. M., Betz, K., Wirth, J., Diederichs, K. & Marx, A. Crystal structures of ternary complexes of archaeal B-family DNA polymerases. *PLoS One* **12**, e0188005 (2017).

223. Brodys, R. S., Doherty, K. G. & Zimmerman, P. D. Processivity and Kinetics of the Reaction of Exonuclease I from Escherichia coli with Polydeoxyribonucleotides. *J. Biol. Chem.* **261**, 7136–7143 (1986).

224. Lu, D., Myers, A. R., George, N. P. & Keck, J. L. Mechanism of Exonuclease I stimulation by the single-stranded DNA-binding protein. *Nucleic Acids Res.* **39**, 6536–6545 (2011).

225. Suck, D. & Oefner, C. Structure of DNase I at 2.0 Å resolution suggests a mechanism for binding to and cutting DNA. *Nature* **321**, 620–625 (1986).

226. Lahm, A. & Suck, D. DNase I-induced DNA conformation. *J. Mol. Biol.* **222**, 645–667 (1991).

227. Pan, C. Q., Uumer, J. S., Herzka, A. & Lazarus, R. A. Mutational analysis of human DNase I at the DNA binding interface: Implications for DNA recognition, catalysis, and metal ion dependence. *Protein Sci.* **7**, 628–636 (1998).

228. Hogan, M. E., Robersont, M. W. & Austint, R. H. DNA flexibility variation may dominate DNase I cleavage. *Biophysics (Oxf).* **86**, 9273–9277 (1989).

229. Tang, Z. *et al.* Constraint of DNA on Functionalized Graphene Improves its Biostability and Specificity. *Small* **6**, 1205–1209 (2010).

230. Zhu, G. *et al.* Nuclease-resistant synthetic drug-DNA adducts: programmable drug-DNA conjugation for targeted anticancer drug delivery. *NPG Asia Mater.* **7**, e169 (2015).

231. Hales, K. & Pochan, D. J. Using polyelectrolyte block copolymers to tune nanostructure assembly. *Current Opinion in Colloid and Interface Science* (2006) doi:10.1016/j.cocis.2006.12.004.

232. Liu, L. Y., Xia, G., Feng, Z. J., Hao, Q. H. & Tan, H. G. Self-assembly of polyelectrolyte diblock copolymers at monovalent and multivalent counterions. *Soft Matter* **15**, 3689–3699 (2019).

233. Borisov, O. V. & Zhulina, E. B. Morphology of micelles formed by diblock copolymer with a polyelectrolyte block. *Macromolecules* **36**, 10029–10036 (2003).

234. Zhulina, E. B. & Borisov, O. V. Theory of Block Polymer Micelles: Recent Advances and Current Challenges. (2012) doi:10.1021/ma300195n.

235. Borisov, O. V, Zhulina, E. B., Leermakers, F. A. M. & Müller, A. H. E. Self-Assembled Structures of Amphiphilic Ionic Block Copolymers: Theory, Self-Consistent Field Modeling and Experiment. in *Self Organized Nanostructures of Amphiphilic Block Copolymers I* (eds. Müller, A. H. E. & Borisov, O.) 57–129 (Springer Berlin Heidelberg, 2011). doi:10.1007/12_2011_114.

236. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput.*

Phys. **117**, 1–19 (1995).

237. Xu, J. *et al.* Dissipative particle dynamics simulations reveal the pH-driven micellar transition pathway of monorhamnolipids. *J. Colloid Interface Sci.* **506**, 493–503 (2017).

238. Jin, R. & Maibaum, L. Mechanisms of DNA hybridization: Transition path analysis of a simulation-informed Markov model. *J. Chem. Phys.* (2019) doi:10.1063/1.5054593.

239. Oweida, T. J., Kim, H. S., Donald, J. M., Singh, A. & Yingling, Y. G. Assessment of AMBER Force Fields for Simulations of ssDNA. *J. Chem. Theory Comput.* (2021) doi:10.1021/acs.jctc.0c00931.

240. Tu, Q. *et al.* Interfacial Mechanical Properties of Graphene on Self-Assembled Monolayers: Experiments and Simulations. *ACS Appl. Mater. Interfaces* **9**, 10203–10213 (2017).

241. Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **112**, 2889–2919 (2012).

242. Sumpter, B. G. & Noid, D. W. Neural networks and graph theory as computational tools for predicting polymer properties. *Macromol. Theory Simulations* **3**, 363–378 (1994).

243. Joyce, S. J., Osguthorpe, D. J., Padgett, J. A. & Price, G. J. Neural network prediction of glass-transition temperatures from monomer structure. *J. Chem. Soc. Faraday Trans.* **91**, 2491 (1995).

244. Mattioni, B. E. & Jurs, P. C. Prediction of Glass Transition Temperatures from Monomer and Repeat Unit Structure Using Computational Neural Networks. *J. Chem. Inf. Comput. Sci.* **42**, 232–240 (2002).

245. Chen, X., Sztandera, L. & Cartwright, H. M. A neural network approach to prediction of glass transition temperature of polymers. *Int. J. Intell. Syst.* **23**, 22–32 (2008).

246. Xu, J. *et al.* Prediction of glass transition temperatures for polystyrenes from cyclic dimer structures using artificial neural networks. *Fibers Polym.* **13**, 352–357 (2012).

247. Ning, L. Artificial neural network prediction of glass transition temperature of fluorine-containing polybenzoxazoles. *J. Mater. Sci.* **44**, 3156–3164 (2009).

248. Yu, X., Wang, X., Li, X., Gao, J. & Wang, H. Prediction of Glass Transition Temperatures for Polystyrenes by a Four-Descriptors QSPR Model. *Macromol. Theory Simulations* **15**, 94–99 (2006).

249. Yu, X. Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers. *Fibers Polym.* **11**, 757–766 (2010).

250. Kennedy, J. & Eberhart, R. Particle swarm optimization. *Proc. IEEE Int. Conf. Neural*

Networks **4**, 1942–1948 (1995).

251. Pei, J.-F., Cai, C.-Z., Zhu, Y.-M. & Yan, B. Modeling and Predicting the Glass Transition Temperature of Polymethacrylates Based on Quantum Chemical Descriptors by Using Hybrid PSO-SVR. *Macromol. Theory Simulations* **22**, 52–60 (2013).

252. Chen, M., Jabeen, F., Rasulev, B., Ossowski, M. & Boudjouk, P. A computational structure-property relationship study of glass transition temperatures for a diverse set of polymers. *J. Polym. Sci. Part B Polym. Phys.* **56**, 877–885 (2018).

253. Venkatraman, V. & Alsberg, B. Designing High-Refractive Index Polymers Using Materials Informatics. *Polymers (Basel).* **10**, 103 (2018).

254. Jabeen, F., Chen, M., Rasulev, B., Ossowski, M. & Boudjouk, P. Refractive indices of diverse data set of polymers: A computational QSPR based study. *Comput. Mater. Sci.* **137**, 215–224 (2017).

255. Mannodi-Kanakkithodi, A., Huan, T. D. & Ramprasad, R. Mining Materials Design Rules from Data: The Example of Polymer Dielectrics. *Chem. Mater.* **29**, 9001–9010 (2017).

256. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **6**, 20952 (2016).

257. Mannodi-Kanakkithodi, A. *et al.* Rational Co-Design of Polymer Dielectrics for Energy Storage. *Adv. Mater.* **28**, 6277–6291 (2016).

258. Huan, T. D. *et al.* Advanced polymeric dielectrics for high energy density applications. *Prog. Mater. Sci.* **83**, 236–269 (2016).

259. Treich, G. M. *et al.* A rational co-design approach to the creation of new dielectric polymers with high energy density. *IEEE Trans. Dielectr. Electr. Insul.* **24**, 732–743 (2017).

260. Huan, T. D. *et al.* A polymer dataset for accelerated property prediction and design. *Sci. Data* **3**, 160012 (2016).

261. Mannodi-Kanakkithodi, A. *et al.* Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **21**, 785–796 (2018).

262. Patra, T. K., Meenakshisundaram, V., Hung, J.-H. & Simmons, D. S. Neural-Network-Biased Genetic Algorithms for Materials Design: Evolutionary Algorithms That Learn. *ACS Comb. Sci.* **19**, 96–107 (2017).

263. Audus, D. J. & de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).

264. Li, T. *et al.* Effect of conjugated polymer poly (9,9-dioctylfluorene) (PFO) molecular

weight change on the single chains, aggregation and β phase. *Polym. (United Kingdom)* **103**, 299–306 (2016).

265. Cravero, F., Martínez, M. J., Vazquez, G. E., Díaz, M. F. & Ponzoni, I. Feature Learning applied to the Estimation of Tensile Strength at Break in Polymeric Material Design. *J. Integr. Bioinform.* **13**, 286 (2016).

266. Wu, K. *et al.* Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials. *J. Polym. Sci. Part B Polym. Phys.* **54**, 2082–2091 (2016).

267. Lutz, J.-F. Aperiodic Copolymers. *ACS Macro Lett.* **3**, 1020–1023 (2014).

268. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).

269. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **92**, 014106 (2015).

270. Zhao, L., Li, Z., Caswell, B., Ouyang, J. & Karniadakis, G. E. Active learning of constitutive relation from mesoscopic dynamics for macroscopic modeling of non-Newtonian flows. *J. Comput. Phys.* **363**, 116–127 (2018).

271. Huan, T. D. *et al.* A universal strategy for the creation of machine learning-based atomistic force fields. *npj Comput. Mater.* **3**, 37 (2017).

272. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).

273. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: Construction, validation, and outlook. *J. Phys. Chem. C* (2017) doi:10.1021/acs.jpcc.6b10908.

274. Polymer Property Predictor and Database.

275. Badger, R. ACD/Labs NMR Databases. https://www.acdlabs.com/%0Aproducts/dbs/nmr_db/.

276. Polymer Science Learning Center Spectral Database.

277. NIST Synthetic Polymer MALDI Recipes Database.

278. Chemical Retrieval on the Web (CROW).

279. MATWEB Material Property Data.

280. Citrination.

281. MakeItFrom.com: Material Properties Database. https://www.makeitfrom.com/.

282. Hiorns, R. C. *et al.* A brief guide to polymer nomenclature from IUPAC. *Colloid Polym. Sci.* **291**, 457–458 (2013).

283. Common Chemistry - Substance Details - 9011-14-7. http://www.commonchemistry.org/ChemicalDetail.aspx?ref=9011-14-7.

284. Struglinski, M. J. & Graessley, W. W. Effects of polydispersity on the linear viscoelastic properties of entangled polymers. 1. Experimental observations for binary mixtures of linear polybutadiene. *Macromolecules* **18**, 2630–2643 (1985).

285. Fox, T. G. & Flory, P. J. Second-order transition temperatures and related properties of polystyrene. I. Influence of molecular weight. *J. Appl. Phys.* **21**, 581–591 (1950).

286. Backfolk, K. *et al.* Determination of the glass transition temperature of latex films: Comparison of various methods. *Polym. Test.* **26**, 1031–1040 (2007).

287. Jordens, K. The influence of molecular weight and thermal history on the thermal, rheological, and mechanical properties of metallocene-catalyzed linear polyethylenes. *Polymer (Guildf).* **41**, 7175–7192 (2000).

288. mbH, C. CAMPUSplastics. *CAMPUS® - a material information system for the plastics industry* (1988).

289. ISO - International Organization for Standardization. https://www.iso.org/home.html.

290. García-Álvarez, R., Hadjidemetriou, M., Sánchez-Iglesias, A., Liz-Marzán, L. M. & Kostarelos, K. In vivo formation of protein corona on gold nanoparticles. The effect of their size and shape. *Nanoscale* **10**, 1256–1264 (2018).

291. Schöttler, S. *et al.* Protein adsorption is required for stealth effect of poly(ethylene glycol)- and poly(phosphoester)-coated nanocarriers. *Nat. Nanotechnol.* **11**, 372–377 (2016).

292. Gref, R. *et al.* 'Stealth' corona-core nanoparticles surface modified by polyethylene glycol (PEG): Influences of the corona (PEG chain length and surface density) and of the core composition on phagocytic uptake and plasma protein adsorption. *Colloids Surfaces B Biointerfaces* **18**, 301–313 (2000).

293. Ding, H. M. & Ma, Y. Q. Design strategy of surface decoration for efficient delivery of nanoparticles by computer simulation. *Sci. Rep.* **6**, (2016).

294. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.* (2015) doi:10.1002/qua.24836.

295. Kim, C., Pilania, G. & Ramprasad, R. From Organized High-Throughput Data to Phenomenological Theory using Machine Learning: The Example of Dielectric Breakdown. *Chem. Mater.* (2016) doi:10.1021/acs.chemmater.5b04109.

296. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* (2016) doi:10.1038/srep20952.

297. Cedervall, T. *et al.* Detailed Identification of Plasma Proteins Adsorbed on Copolymer Nanoparticles. *Angew. Chemie Int. Ed.* **46**, 5754–5756 (2007).

298. Corbo, C. *et al.* The impact of nanoparticle protein corona on cytotoxicity, immunotoxicity and target drug delivery. *Nanomedicine* **11**, 81–100 (2016).

299. Caracciolo, G., Farokhzad, O. C. & Mahmoudi, M. Biological Identity of Nanoparticles In Vivo: Clinical Implications of the Protein Corona. *Trends Biotechnol.* **35**, 257–264 (2017).

300. Gillies, R. J., Pilot, C., Marunaka, Y. & Fais, S. Targeting acidity in cancer and diabetes. *Biochimica et Biophysica Acta - Reviews on Cancer* vol. 1871 273–280 (2019).

301. Fröhlich, H. *et al.* From hype to reality: Data science enabling personalized medicine. *BMC Med.* **16**, 1–15 (2018).

302. Cirillo, D. & Valencia, A. Big data analytics for personalized medicine. *Current Opinion in Biotechnology* vol. 58 161–167 (2019).

303. Mello, M. M., Lieou, V. & Goodman, S. N. Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing. *N. Engl. J. Med.* **378**, 2202–2211 (2018).

304. Wang, L. Heterogeneous Data and Big Data Analytics. *Autom. Control Inf. Sci.* **3**, 8–15 (2017).

305. Cavanillas, J. M., Curry, · Edward & Wahlster, W. *New Horizons for a Data-Driven Economy A Roadmap for Usage and Exploitation of Big Data in Europe*.