RESEARCH ARTICLE

Macromolecular Theory and Simulations
www.mts-journal.de

# Dimerization of Polyglutamine within the PRIME20 Model using Stochastic Approximation Monte Carlo

*Christian Lauer and Wolfgang Paul**

**This study presents a numerical investigation of the dimerization of polyglutamine homo-peptides of varying length. It employs the PRIME20 intermediate resolution protein model and studies it with a flat-histogram type Monte Carlo simulation that gives access to the thermodynamic equilibrium of this model over the complete control parameter range (for the simulations this is temperature). For densities comparable to typical in vitro experimental conditions, this study finds that the aggregation and folding of the polyglutamine chains occur concurrently. However, as a function of chain length the sequence of establishment of intra- and intermolecular hydrogen bonding contacts changes. Chains longer than about $N = 24$ polyglutamine repeat units fold first and then aggregate. This agrees well with the experimental finding that, beyond $N = 24$ the single polyglutamine chain is the critical nucleus for the aggregation of amyloid fibrils. A finite size scaling of the ordering temperatures reveals that for this chain length (and longer chains) folding occurs at physiological (respectively larger) temperatures, whereas shorter chains are disordered at physiological conditions.**

## 1. Introduction

Aggregation of proteins into amyloid fibrils is a phenomenon accompanying neuro-degenerative diseases.[1] The fibrils may not be the cause of the diseases, as was originally thought, but only serve as a waste dump for excess protein, removing the toxic smaller scale aggregates. However, understanding of the aggregation of proteins into clusters clearly is fundamental to get a better understanding of the origin of these neuro-degenerative diseases. One of these is Huntington's Corea, connected with the aggregation of the protein Huntingtin. Within that protein, it is elongated polyglutamine (PolyQ) stretches that lead to the aggregation.[2–6] It is therefore worthwhile to investigate the aggregation of the homo-protein polyglutamine in detail.

Recent computational studies using molecular dynamics (MD) were able to confirm $\beta$-sheet structures as stable structural motive for PolyQ aggregates that match experimental findings.[7–9] However, the configuration range accessible with MD is highly dependent on the starting configurations. For a better understanding of the aggregation transition it has proven useful to turn to broader sampling techniques, like Monte Carlo, and coarse-grained protein models.[10] In our studies we use flat-histogram Monte Carlo simulations and an intermediate resolution coarse-grained protein model, the PRIME20 model. This model has been used for a series of studies of protein folding and aggregation before[11–14] and we have recently shown that it is well able to predict and differentiate between folding structures of polyalanine, polyserine, and polyglutamine.[15] In this study, we will focus on the dimerization of polyglutamine chains of varying length. From the experimental side it is known that polyglutamine stretches in Huntingtin promote aggregation once they extend to about $N = 30$ repeat units.[5,16–18] On the other hand, from time-lag studies of polyglutamine amyloid fibril formation (which is a nucleation and growth process) it appears as if beyond a chain length of about $N = 23$ repeat units, the single chain is already the critical nucleus for further aggregation. This could be interpreted in a way that starting at this chain length polyglutamine chains are folded into $\beta$ hairpins at physiological conditions, while they are disordered for shorter chains. So it is an interesting question to ask, whether two polyglutamine chains aggregate into an amorphous globule first and then order at lower temperature (temperature will be the control parameter in the simulation) or fold first and then aggregate or do both at the same time. We will also be identifying the hydrogen bond pattern stabilizing the aggregate and the hairpin.

In section 2 we will introduce the model and simulation method. Our results will be presented in discussed in section 3 and we wrap up with some conclusions in section 4.

## 2. Model and Methods

We will first give some summary of the pertinent properties of the PRIME20 model and then explain the flat-histogram Monte Carlo

C. Lauer, W. Paul
Martin-Luther-Universität Halle-Wittenberg
Institut für Physik
06099 Halle (Saale), Germany
E-mail: wolfgang.paul@physik.uni-halle.de

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/mats.202200075



**DOI: 10.1002/mats.202200075**

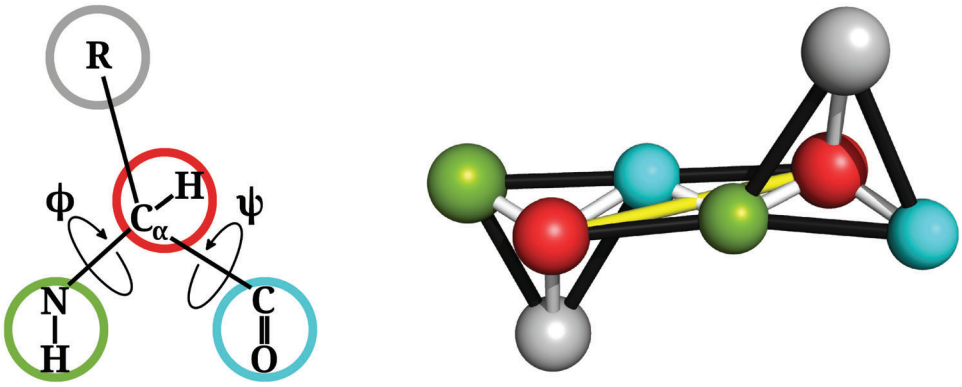



**Figure 1.** Geometry of the PRIME20 model. The backbone is represented by three beads: the NH group (green bead), the $C_\alpha$ carbon (red bead) and the CO group (cyan bead). The side chain is represented by the fourth bead (gray bead). Its position and size is specific for the individual type of amino acid. On the left the assignment of atoms to beads and the dihedral angles are shown. On the right the geometry of a PRIME20 dimer is shown. White sticks represent covalent bonds. Black and yellow sticks represent pseudo-bonds that stabilize the structure. The size of the beads is not so scale.

method called Stochastic Approximation Monte Carlo (SAMC) that we employed for this study.

### 2.1. The PRIME20 Model

We use the PRIME20 model, which is an intermediate resolution coarse-grained protein model. As such, it aims to model protein structure with sufficient detail to capture important aspects of structure formation, while at the same time maintaining reasonable computation costs. This grants access to a systems equilibrium information over a wide parameter range. The coarse-grained approach has previously proven useful in simulation studies by Chen et al.[10]

PRIME20 was first developed by Cheon et al. as an extension of the PRIME model in 2010.[11] The PRIME model was first introduced by Voegler Smith et al. in 2001 to study secondary structure formation of polyalanine.[19–21] PRIME20 expands the model to include all 20 proteinogenic amino acids. All of these 20 amino acids share the same backbone geometry. In the PRIME20 model the backbone is represented by three beads, as shown in **Figure 1**: the NH, the $C_\alpha$ and the CO bead. The fourth bead represents the side chain. Its position and size are specific for the type of amino acid.

A comprehensive collection of the parameters of the PRIME20 model can be found in the Supporting Information of [15]. In this section we focus on the parameters of glutamine.

Covalent bonds are represented by infinite well potentials around an ideal bond length. Fluctuations of 2.375% from the ideal value are allowed, which defines the width of the well potential:

$$V_{\text{bond}}(d) = \begin{cases} 0 & \text{if } d \in [d_{\text{ideal}} - \Delta, d_{\text{ideal}} + \Delta] \\ \infty & \text{otherwise} \end{cases} \quad (1)$$

where $d$ is the distance between the two bonded beads, $d_{\text{ideal}}$ is the ideal bond length and $\Delta = 0.02375\, d_{\text{ideal}}$. Bonds are shown as white sticks on the right in Figure 1. Next to these ordinary bonds, pseudo-bonds are in place between beads separated by two covalent bonds. In Figure 1, pseudo-bonds are shown as black sticks. Pseudo-bonds in the model behave just like ordinary bonds. They keep the bond angles of the peptide within their physical range. Additional pseudo-bonds are used between consecutive $C_\alpha$ beads to keep the polymer chains in *trans* configuration. The relevant bond and pseudo-bond lengths for PolyQ are listed in **Table 1**.Interactions between non-bonded beads are separated into two types. Hydrophobic interactions between backbone beads as well as between backbone and side-chain beads are treated as hard-sphere (HS) repulsions. Hydrophobic interactions involving only side-chain beads are modeled as semi-infinite square-well (SW) potentials:

**Table 1.** Bond and pseudo-bond lengths between beads of PolyQ in PRIME20. Here, the index i represents beads of the (i)th residue and the index i+1 represents beads of the (i+1)th residue. Sizes in Å.

| | | | | | | |
|---|---|---|---|---|---|---|
| Bonds | $NH_i$-$C_{\alpha,i}$ | $C_{\alpha,i}$-$CO_i$ | $CO_i$-$NH_{i+1}$ | $R_i$-$C_{\alpha,i}$ | | |
| | 1.46 | 1.51 | 1.33 | 1.60 | | |
| Pseudo-bonds | $NH_i$-$CO_i$ | $C_{\alpha,i}$-$NH_{i+1}$ | $CO_i$-$C_{\alpha,i+1}$ | $NH_i$-$R_i$ | $C_{\alpha,i}$-$C_{\alpha,i+1}$ | $CO_i$-$R_i$ |
| | 2.45 | 2.41 | 2.45 | 2.50 | 3.80 | 2.56 |

$$V_{\text{HS}}(d_{ij}) = \begin{cases} 0 & \text{if } d_{ij} > d_{ij}^{\text{HS}} \\ \infty & \text{otherwise} \end{cases} \quad (2)$$

$$V_{\text{SW}}(d_{ij}) = \begin{cases} 0 & \text{if } d_{ij} > d_{ij}^{\text{SW}} \\ \epsilon_{ij} & \text{if } d_{ij}^{\text{HS}} < d_{ij} < d_{ij}^{\text{SW}} \\ \infty & \text{otherwise} \end{cases} \quad (3)$$

where $d_{ij}$ is the distance between beads i and j, $d_{ij}^{\text{HS}}$ is the hard-sphere diameter, $d_{ij}^{\text{SW}}$ is the square-well interaction distance and $\epsilon_{ij}$ is the square-well depth. For interactions between side-chain beads, the three functional parameters ($d_{ij}^{\text{HS}}$, $d_{ij}^{\text{SW}}$ and $\epsilon_{ij}$) have specific values for each pair of interacting side-chain beads i and j, resulting in 171 values each (glycine has no side chain). For hard-sphere repulsion interactions we use the Lorenz-Berthelot combining rule to calculate $d_{ij}^{\text{HS}}$ from the beads $d^{\text{HS}}$. As side-chain diameters are only defined for side-chain-side-chain interactions,

**Table 2.** Bead diameters and square-well parameters of PolyQ in PRIME20. Sizes in Å.

| | NH | $C_\alpha$ | CO | R |
|---|---|---|---|---|
| $d^{HS}$ | 3.3 | 3.7 | 4.0 | 3.6 |
| $d^{SW}$ | 4.5 | – | 4.5 | 6.6 |
| $\epsilon$ | -1.000 | – | -1.000 | -0.080 |

**Table 3.** Squeeze factors and resulting reduced bead diameters for backbone bead interactions and interactions involving a polyglutamine side chain. Sizes in Å.

| Interactions | $C_{\alpha,i}-CO_{i+1}$ | $C_{\alpha,i}-NH_{i-1}$ | $CO_i-NH_{i+2}$ | $NH_i-NH_{i+1}$ | $CO_i-Co_{i+1}$ |
|---|---|---|---|---|---|
| original $d$ | 3.85 | 3.50 | 3.65 | 3.30 | 4.00 |
| squeeze factor | 1.1436 | 0.88 | 0.87829 | 0.8 | 0.7713 |
| squeezed $d$ | 4.40286 | 3.08 | 3.2057585 | 2-64 | 3-0852 |
| Interactions | $C_{\alpha,i-1}-R_i$ | $CO_{i-1}-R_i$ | $NH_{i+1}-R_i$ | $C_{\alpha,i+1}-R_i$ | $CO_{i-2}-R_i$ |
| original $d$ | 3.65 | 3.8 | 3.45 | 3.65 | 3.8 |
| squeeze factor | 1.407 | 1.089 | 1.158 | 1.387 | 1.316 |
| squeezed $d$ | 5.134 | 4.139 | 3.996 | 5.062 | 5.000 |

we use their self-interaction diameter for side-chain-backbone interactions. The self-interaction value of $d_{i,j}^{HS}$ and $d_{i,j}^{SW}$ are shown in **Table 2**.Hydrogen bonds form between NH and CO beads. This interaction also uses a semi-infinite square well potential with $d^{SW} = 4.5$Å. However, to achieve a realistic representation of hydrogen bonds, additional restrictions apply to their formation in PRIME20. For a detailed description, see [15] and [22].

The hydrogen bond strength $\epsilon_{HB}$ defines the energy scale of the model and is set to $-1$. All side chain interaction energies (Table 2) are given relative to $\epsilon_{HB}$. By assigning a value to $\epsilon_{HB}$ we can map the reduced energy E and Temperature T of the model to physical quantities $E' = \epsilon_{HB}E$ and $T' = \epsilon_{HB}T/k_B$. Solvent interactions are modelled implicitly in PRIME20. Thus, $\epsilon_{HB}$ has to factor in the energy gain of backbone hydrogen bonds, as well as further interactions like interactions between peptide and solvent and possible side-chain hydrogen bonds.

The model geometry described up to this point is not sufficient to model real proteins, because the large bead diameters prevent any legal configuration of the chain. Specifically along the backbone, $CO_i$ and $CO_{i+1}$ will always overlap for any position of the dihedral angle $\Phi_i$. To solve this issue, so called *squeeze factors* were introduced. They reduce the effective diameters of beads in close proximity along the chain. There are squeeze factors for ten different bead interactions. For glutamine, they are listed in **Table 3**. Squeeze factors for interactions involving side chains are specific for each amino acid. For a complete list of the parameters of the PRIME20 model, see [15].

### 2.2. Stochastic Approximation Monte Carlo

The simulation method we employ is Stochastic Approximation Monte Carlo (SAMC),[23,24] which is an advanced flat-histogram Monte Carlo method.[25] As such, it aims for a flat visitation histogram of energy states. In achieving this, it avoids getting stuck in local energy minima as can be the problem with conventional Monte Carlo (MC) algorithms. To produce a flat visitation histogram, SAMC approximates the configurational density of states (DOS) $g(U)$ as a function of potential energy $U$, and uses the approximated $g(U)$ in the MC acceptance probability calculation. The DOS describes the number of states in the interval $[U, U + \Delta U]$. An SAMC move from configuration $x$ with the energy $U(x)$ to $x'$ with the energy $U(x')$ is accepted with the probability $\min(1, \tilde{g}(U(x))/\tilde{g}(U(x')))$, where $\tilde{g}$ is the current guess for the density of states. After accepting or rejecting the move, $g(U)$ is updated according to

$$\tilde{g}(U(x_{new})) = \tilde{g}(U(x_{new})) + \gamma_t \tag{4}$$

where $x_{new} = x'$ if the move was accepted and $x_{new} = x$ if the move was rejected. The modification factor $\gamma_t$ decreases over time, which is measured in the number of MC steps passed. For $t \to \infty$, $\gamma_t$ goes to 0. For the approximation of the DOS to converge, the time series of $\gamma_t$ has to obey further conditions. This has been investigated thoroughly in [23, 24] and [26].

As the DOS varies over several orders of magnitude, the logarithm $\ln g(U)$ is used instead. This is the configurational entropy $S(U) = \ln g(U)$, if we set the Boltzmann constant to $k_B = 1$. The entropy $S(U)$ is averaged over several independent simulation runs. The flatness of the energy state visitation histogram acts as a measure for the quality of $\tilde{g}(U)$ as an approximation of $g(U)$. After a sufficiently accurate $S(U)$ is obtained, further MC simulations with the SAMC acceptance probability but with fixed $g(U)$ were performed to accumulate statistics for further structural observables as a function $U$, like the hydrogen bond contact probabilities and the tensor of gyration. The SAMC acceptance probability ensures an evenly distributed visitation of energy states of the system.

Four different MC move types are used in the simulations. The local displacement move moves a single randomly chosen bead in a random direction. The maximum displacement is small (0.02Å). The pivot rotation move randomly chooses a residue and rotates either its $\Phi$ or $\Psi$ angle by a random amount and direction. The maximum rotation angle we typically chose was $\pi/3$ to optimize the moves acceptance rate together with the chains mobility in configuration space. The last two moves are a rotation and a translation of an entire chain. As the systems contain two chains their relative position and orientation have to be modifiable by the simulation. After all moves, the new positions must conform to bond length and excluded volume constraints.

We simulated systems of two PolyQ chains of lengths (number of residues) $N = (14, 16, 18, 20, 22, 24, 26, 28, 36)$. The simulation box has length $L = 112.5$Å for $N = (14, 16, 18, 20, 22, 24, 26)$ and $L = 150.0$Å for $N = (28, 36)$. This translates to a milli-molar concentration, which is close to concentrations in in vitro experiments on PolyQ aggregation.

To validate our newly written simulation program we reproduced the densities of states of single chain polyglutamine systems with chain lengths $N = 10$ and $N = 16$ as determined by Böker et al. in the PRIME20 model using SAMC.[15] Our reproduced $\ln g(U)$ results for these two systems match the data by Böker et al. perfectly.

In the remaining sections, systems will be identified by the following code: 'CXQY', where 'X' indicates the number of chains and 'Y' indicates the number of glutamine (Q) residues per chain,

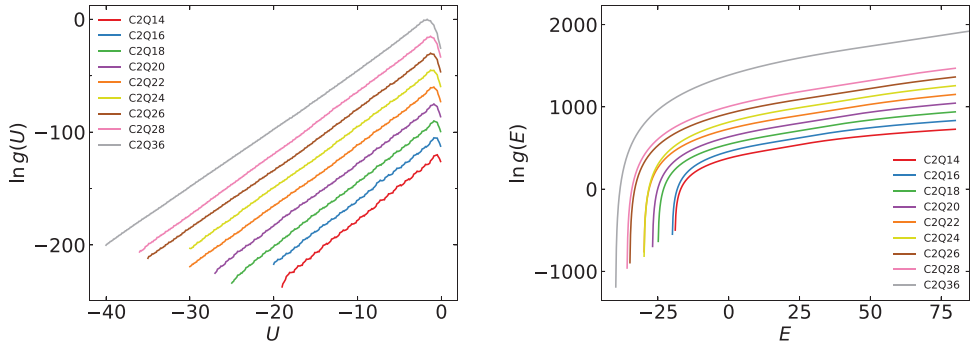



**Figure 2.** Density of states in the configurational microcanonical ensemble $g(U)$ (left figure) and in the full microcanonical ensemble $g(E)$ (right figure)

e.g., C2Q24 specifies a system of *two* chains, each with 24 glutamine residues.

## 3. Results

### 3.1. Thermodynamics

The configurational entropies ($S = \ln g(U)$) obtained from the SAMC procedure are shown on the left in **Figure 2**. As one expects, the graphs have an ascending slope, since the low energy regime puts more restrictions on configuration space. Furthermore, the entropies show an oscillating behavior that stems from the two different energy scales of the PRIME20 model.[22] These energy scales are the energy gain of a formed hydrogen bond ($\epsilon_{HB} = -1$) on the one hand, and the energy gain of a side-chain contact ($\epsilon_{SC} = -0.08$ for PolyQ) on the other hand. It is most prominent for the systems of shorter chains, because shorter chains have less side-chains available to match the energy gain of a formed hydrogen bond.

For a microcanonical thermodynamic analysis of the systems we need to consider the entropy and the structural observables as functions of the total energy $E$ as opposed to functions of only the potential energy $U$. To transform to a density of states $g(E)$ in this microcanonical ensemble we perform a convolution of $g(U)$ with the density of states of the kinetic energy as described by Shakirov et al.[27] The transformed densities of states of the *nine* different chain lengths are shown on the right in Figure 2. The lower bounds are determined by the lowest converged energies in the configurational ensemble. The upper bounds are infinity. When going to large energies the curves converge towards the density of states of the ideal gas. In order to interpret the smoothed shape of $\ln g(E)$ we look at the temperature $T(E)$ and the heat capacity $C_V(E)$ which are given by derivatives of the entropy $S(E)$:

$$T(E) = \left(\frac{\partial S}{\partial E}\right)^{-1} \tag{5}$$

$$C_V(E) = \left(\frac{\partial T}{\partial E}\right)^{-1} = -\frac{1}{T}\left(\frac{\partial^2 S}{\partial E^2}\right)^{-1} \tag{6}$$

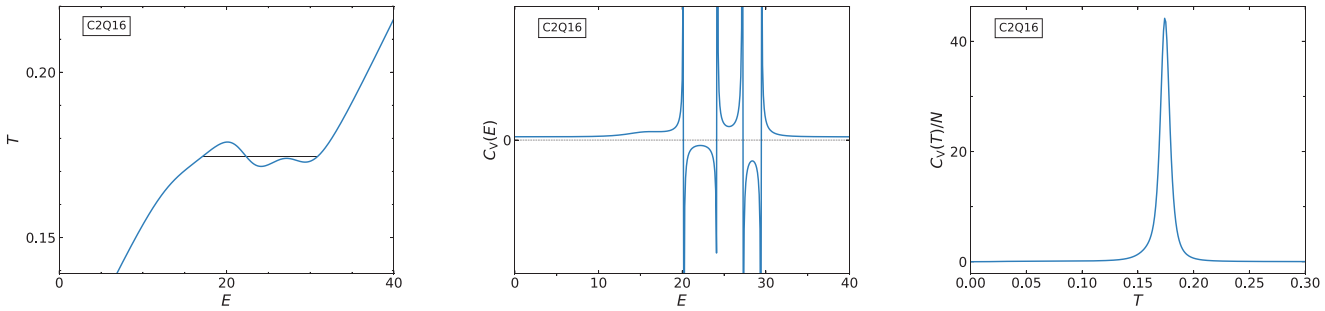
Maxima in the heat capacity indicate pseudo-phasetransitions.[28–30] The prefix "pseudo" arises because we deal with a finite system and in statistical mechanics phase transitions are defined in the thermodynamic limit, only. However, in the remainder of the text we drop the "pseudo"-prefix for better readability.

A phase transition (PT) is of first order if the slope of $T(E)$ at the corresponding inflection point $E^*$ is negative. As $C_V(E)$ is the inverse of the first derivative of $T(E)$ it holds that $C_V(E^*) < 0$. In this case the temperature curve $T(E)$ is non-monotonic and both phases coexist at the transition temperature. If the slope of $T(E^*)$ at the inflection point is positive and thus $C_V(E^*) > 0$, the two phases do not coexist. This is defined as a second-order PT. As illustration, the temperature and microcanonical heat capacity of the system of two chains of length $N = 16$ is show in **Figure 3**. The two maxima with negative value in $C_V(E)$ correspond to inflection points in $T(E)$ that belong to the same large scale oscillation of the temperature curve. They are part of the same transition in the system. The temperature of this transition can be determined by performing a Maxwell construction. Even though the two negative maxima in $C_V(E)$ are united in one temperature when using the Maxwell construction, the behavior of the microcanonical heat capacity tells us about two first-order PTs that are happening at the same temperature. These two first-order PTs are the folding of the single chain into an ordered structure and the aggregation transition.

The other chain lengths show similar oscillations in $T(E)$ to the one seen in Figure 3a. For these we also use the Maxwell construction in the T(E) graph to calculate the first-order PT temperatures. Almost all systems exhibit two negative maxima in $C_V(E)$ at the same temperature, like C2Q16 in Figure 3. The only exceptions are C2Q20 and C2Q36. In C2Q20 a shoulder for the missing' maximum is present, but the peak gets superimposed with the neighboring peak. In C2Q36 no distinction between two separate transition signatures in $C_V(E)$ can be made. A complete list of the PT temperatures identified in the microcanonical ensemble is shown in **Table 4**. Systems of all investigated chain length show signatures of first-order PTs at energies above zero. To support the transition temperatures found in the microcanonical ensemble we also look at the canonical ensemble. From the microcanonical DOS $g(E)$ we can obtain the partition function $Z(T)$ by calculating the Laplace transform to the temperature as variable:

$$Z(T) = \sum_E g(E)\exp(-E/T) \tag{7}$$

here the Boltzmann constant is again set to $k_B = 1$. The canonical partition function plays a central role in the canonical analysis, as




(a) Temperature (b) Microcanonical heat capacity (c) Canonical heat capacity

**Figure 3.** Microcanonical and canonical analysis of the entropy $S$ for system C2Q16.

**Table 4.** Phase transition temperatures for all investigated systems as obtained from microcanonical and canonical analysis of the entropy $S(E)$. $T^{\star}_{mic}$ is determined by finding maxima in the microcanonical heat capacity. $T^{\star}_{can}$ is derived from the maxima in the canonical heat capacity..

| | Transition Temperatures $T^{\star}$ | | |
|---|---|---|---|
| | | Microcanonical Ensemble $T^{\star}_{mic}$ | |
| N | Canonical Ensemble $T^{\star}_{can}$ | 1st order | 2nd order |
| 14 | $0.0600 \pm 0.0009$ | $0.16846 \pm 0.00088$ | $0.1670 \pm 0.0036$ |
| | $0.1680 \pm 0.0009$ | | |
| 16 | $0.1740 \pm 0.0025$ | $0.1746 \pm 0.0026$ | $0.1729 \pm 0.0019$ |
| 18 | $0.1750 \pm 0.0009$ | $0.17555 \pm 0.00088$ | $0.1715 \pm 0.0037$ |
| 20 | $0.1800 \pm 0.0007$ | $0.17998 \pm 0.00065$ | $0.1764 \pm 0.0040$ |
| 22 | $0.1840 \pm 0.0011$ | $0.1835 \pm 0.0014$ | $0.1868 \pm 0.0047$ |
| 24 | $0.1840 \pm 0.0008$ | $0.18369 \pm 0.00076$ | $0.1852 \pm 0.0027$ |
| 26 | $0.1870 \pm 0.0010$ | $0.1872 \pm 0.0010$ | |
| 28 | $0.1820 \pm 0.0016$ | $0.1825 \pm 0.0016$ | |
| 36 | $0.1920 \pm 0.0012$ | $0.1920 \pm 0.0011$ | |

it is used to derive further thermodynamic observables, like the canonical heat capacity $C_V(T)$. The canonical heat capacity $C_V(T)$ is defined as

$$C_V(T) = \frac{\partial \langle E \rangle}{\partial T} = \frac{1}{\partial T} \frac{1}{Z(T)} \sum_E E g(E) \exp(-E/T) \tag{8}$$

As an example, $C_V(T)$ of two chains of length $N = 16$ (C2Q16) is shown in Figure 3c. $C_V(T)$ has a distinct peak at a temperature close to $T = 0.75$. A peak in $C_V(T)$ indicates a PT at the corresponding temperature. The canonical PT temperatures for all system sizes are listed in Table 4. By comparing the results from the microcanonical and the canonical analysis for the different chain lengths, as shown in the left graph of **Figure 4**, we see that the transition temperatures from the microcanonical and the canonical ensemble are in good agreement. The deviations between the ensembles stem from the finite size of the simulation system and should disappear when going to infinite systems. The PT temperatures increase with the chain length. The data points for $N = 28$ do not follow the trend because the density of states was not converging for low enough energy for this chain length, leading to a too small transition temperature. In the thermodynamic limit, aggregation transitions are of first order and thus finite size corrections scale as $N^{-1/3}$. Folding transitions are typically of first order as well, which suggests using the same scaling behavior. This is employed on the right side of Figure 4. For $N = 14$ there is a second transition peak at a low temperature of $T = 0.060$ (see Table 4), which is not shown in Figure 4 because it belongs to a different transition, as explained in more detail in the sections below.

The high temperature transition peaks in the canonical specific heat have to be attributed to folding and aggregation of the chains. Above the transition temperature the systems form random configurations with occasional side chain interactions and a very low chance of hydrogen bonds forming (which will be discussed in more detail in section 3.2.2). Below the transition temperature, the system forms regular, folded, and aggregated structures with dominating backbone hydrogen bond interactions. To support this assignment, we turn to structural observables like the tensor of gyration and hydrogen bond contact matrices.

### 3.2. Configurations

#### *3.2.1. Tensor of Gyration*

The tensor of gyration **S** is calculated for the individual chains in the system and then averaged over both chains. In **Figure 5**, the eigenvalues of the tensors of gyration for the different systems are displayed versus temperature. The eigenvalues measure the extension of the chains in the principal axis system. In all systems, two phases are visible with $\langle \lambda_2 \rangle$ and $\langle \lambda_3 \rangle$ dropping significantly when cooling past the transition temperature $T^{\star}_{can}$ identified in the heat capacity analysis above. The inflection points of the eigenvalue curves confirm the temperatures already identified from the microcanonical and canonical heat capacities. In $\langle \lambda_1 \rangle$ the drop at $T^{\star}_{can}$ is less pronounced, yet still present. In small systems of $N = 14, 16, 18, 20$, $\langle \lambda_1 \rangle$ increases at the transition when cooling, indicating an elongation along the main axis of the gyration ellipsoid. When cooling further $\langle \lambda_1 \rangle$ decreases again below its value above the transition. This behavior is mirrored in $\langle \lambda_3 \rangle$. Here, the eigenvalue of $N = 14, 16, 18, 20$ systems drops at $T^{\star}_{can}$ when cooling, followed by a slight increase until it stabilizes at its chain length specific low temperature value. No significant changes in $\langle \lambda_1 \rangle$ and $\langle \lambda_2 \rangle$ of the $N = 14$ system can be observed at the low temperature transition $T^{\star}_{can} = 0.060$ that was

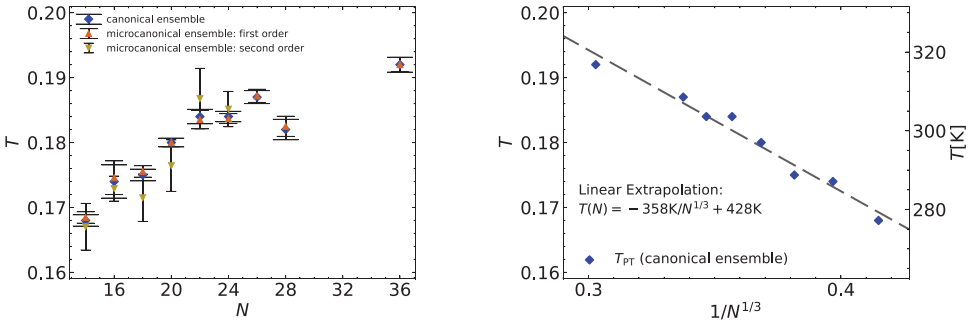



**Figure 4.** Transition temperatures as a function of chain length from canonical and microcanonical analysis. On the left side results for the different methods are compared. On the right side, the canonical transition temperatures are plotted versus $N^{-1/3}$. The dashed line is a linear extrapolated to infinite chain length. The temperature mapping that results in the temperature scale in Kelvin is discussed in Section 3.2.4 below.

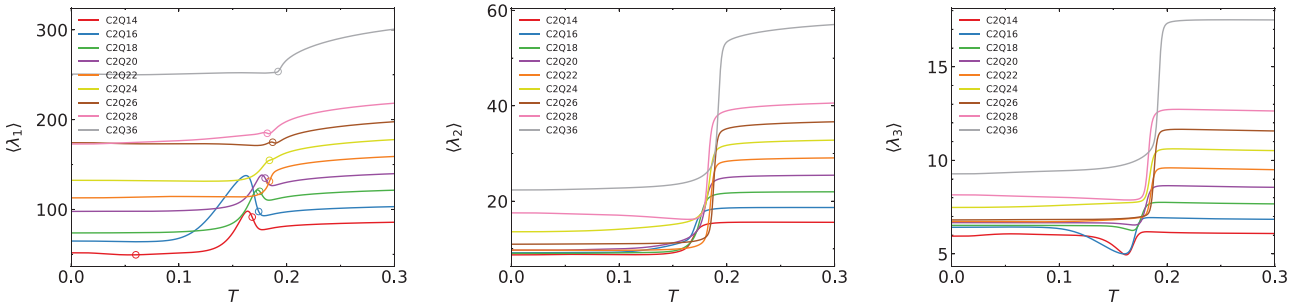



**Figure 5.** Eigenvalues of the gyration tensor, i.e., lengths of the axis of the gyration ellipsoid showing the folding transition of the chains. Phase transition temperatures from the canonical analysis are marked by circles.

identified in the canonical heat capacity analysis. However, in $\langle\lambda_3\rangle$, $T^{\star}_{\text{can}} = 0.060$ lies in the constant region below the well of the high-temperature transition. The short chain behavior might indicate an aggregation of stretched chains followed by a collapse of the aggregated conformation. In systems of longer chains no elongation at $T^{\star}_{\text{can}}$ is observed. The strong reduction in $\langle\lambda_2\rangle$ and $\langle\lambda_3\rangle$ combined with the weak reduction in $\langle\lambda_1\rangle$ fits to the formation of $\beta$-hairpin structures. Such chain length dependent aggregation behavior for PolyQ has been suggested by Chen et al. in simulation studies. They found extended structures to be the preferred form of their short monomers ($N \leqslant 20$), while longer chains ($N = 30$) show a preference for the $\beta$-hairpin structure.[10]

In small chain systems with $N = 14, 16, 18, 20, 22$, the second and third eigenvalues group around the same value in the low-temperature region with averages $\langle\lambda_2\rangle \approx 9$ and $\langle\lambda_3\rangle \approx 6.5$. In $\langle\lambda_3\rangle$, $N = 26$ also reaches this value.

From the eigenvalues of the gyration tensor we derive the relative shape anisotropy:

$$\kappa^2 = 1 - 3\frac{\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1}{(\lambda_1 + \lambda_2 + \lambda_3)^2}. \tag{9}$$

It is a measure for symmetry and dimensionality of the confirmations. $\kappa^2$ is limited between 0 and 1. It becomes 0 for highly symmetric spherical conformations and 1 for an ideal rod. As seen in **Figure 6**, when cooling, $\kappa^2$ increases at $T^{\star}_{\text{can}}$ in all systems. The inflection points are in agreement with $T^{\star}_{\text{can}}$ from the heat capacity analysis. In Figure 6 these $T^{\star}_{\text{can}}$ are marked by circles. The behavior of the relative shape anisotropy indicates a folding transition at $T^{\star}_{\text{can}}$ from random coils to $\beta$-hairpin structures. In systems of $N = 14, 16, 18, 20$, $\kappa^2$ decreases again below $T^{\star}_{\text{can}}$ down to a chain length specific value, indicating a further collapse of the conformations at low temperatures.

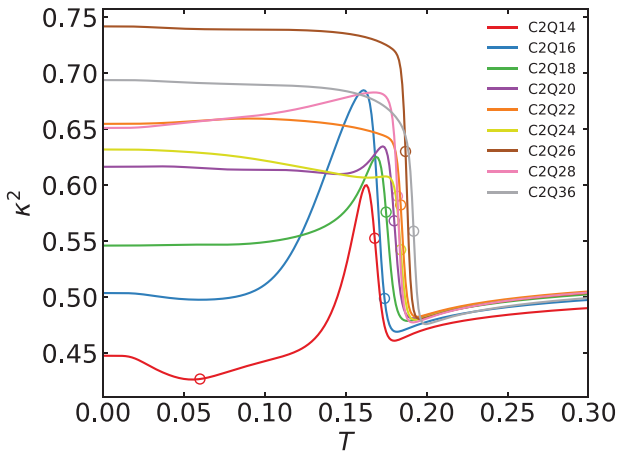



**Figure 6.** The relative shape anisotropy is a measure for symmetry and dimensionality. It is zero for spherical conformations and one for an ideal rod. Transition temperatures from the canonical heat capacity analysis ($T^{\star}_{\text{can}}$) are marked by circles.

In $N = 14$ the low temperature transition at $T^{\star}_{\text{can}} = 0.060$ marks the global minimum in $\kappa^2$ where the systems conformations are

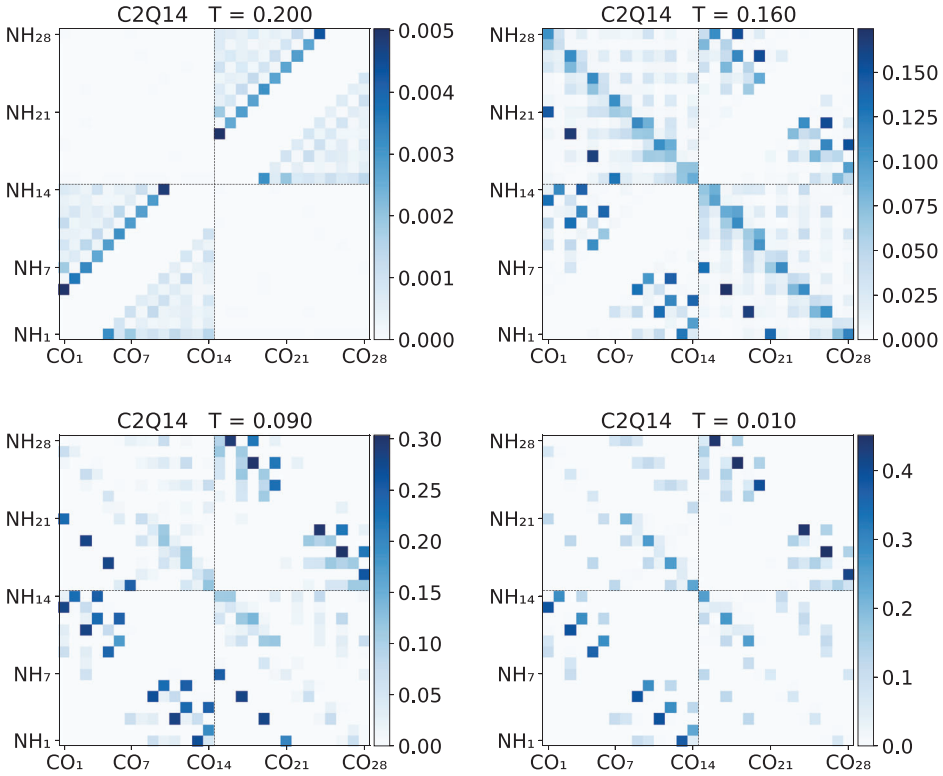



**Figure 7.** Hydrogen bond contact matrices for chain length $N = 14$ at four different temperatures. Each cell $(i, j)$ corresponds to a possible contact between the $CO_i$ and $NH_j$ bead. Darker colors mean higher probability for the respective contact to form. The colors are scaled to the highest value in the matrix. Equal colors between the matrices therefor do not mean equal contact probabilities.

most symmetric and spherical. Below this temperature $\kappa^2$ increases and the system tends to more stretched conformations.

### 3.2.2. Hydrogen Bond Contact Matrices

We use hydrogen bond contact matrices (HB matrices) to identify characteristic conformations in the phases. These matrices are contact probability maps of the hydrogen bond partners. Each cell $(i, j)$ in the $N \times N$ map corresponds to a possible contact between the $CO_i$ and $NH_j$ bead. The darker the color of the cell, the higher the probability of a hydrogen bond forming between the two corresponding beads. To include both chains in one matrix and show intra- as well as inter-chain contacts, all NH and CO beads in the system are numbered consecutively. The axes of the HB matrices have a length of $2N$ with $[1, N]$ representing beads from chain 1 and $[N + 1, 2N]$ representing beads from chain 2. It follows, that the top left and bottom right quadrants of the matrices show contact probabilities of beads of different chains (referred to in the following as *inter-quadrants*). The bottom left and top right quadrants are showing contact probabilities of beads of the same chain (referred to in the following as *intra-quadrants*).

***C2Q14***: **Figure 7** shows the HB matrices of the C2Q14 system at four different temperatures: first, at $T = 0.200$ far above the high temperature $C_V(T)$ peak ($T^\star_{\text{can}} = 0.168$); second, at $T = 0.160$ just below $T^\star_{\text{can}}$, third, at $T = 0.090$ between the two $C_V(T)$ peaks ($T^\star_{\text{can}} = 0.168$ and $T^\star_{\text{can,low}} = 0.060$) and finally at $T = 0.010$ below both $C_V(T)$ peaks. In all *four* matrices we observe that they are not symmetric, meaning cell $(i, j)$ is not necessarily equal to cell $(j, i)$. That is because CO and NH beads have different mobility properties and confinement restrictions due to their surrounding beads, leading to different abilities of forming hydrogen bonds. Furthermore, the PRIME20 model forbids hydrogen bonds to form between beads with $|i - j| < 4$. This creates a white diagonal of forbidden cells in the intra-quadrants.

The top left matrix at $T = 0.200$ represents the random coil state. The overall probability of even a single hydrogen bond forming is very low. The most commonly closed hydrogen bonds are on a diagonal in the intra-quadrants with $|i - j| = 4$. These beads are in close proximity along the chain. With only minor collapse of the chain a hydrogen bond can form between them. Thus, entropy favors these contacts, compared to other possible hydrogen bonds that put more restrictions on the conformation.

At $T = 0.160$ (Figure 7 top right), the system is in the collapsed state. The overall probability of hydrogen bonds forming is much higher when compared to the $T = 0.200$ state. Diagonals with descending slopes dominate the HB matrix. They all indicate anti-parallel alignment of polymer strands, which manifests itself as follows: if contact $(i, j)$ is closed, so is $(j, i)$. The next closed pairs in the anti-parallel structure are $(i + 2, j - 2)$ and $(i - 2, j + 2)$. We find two aggregated structures in the system. First, anti-parallel alignment of both whole PolyQ chains, which

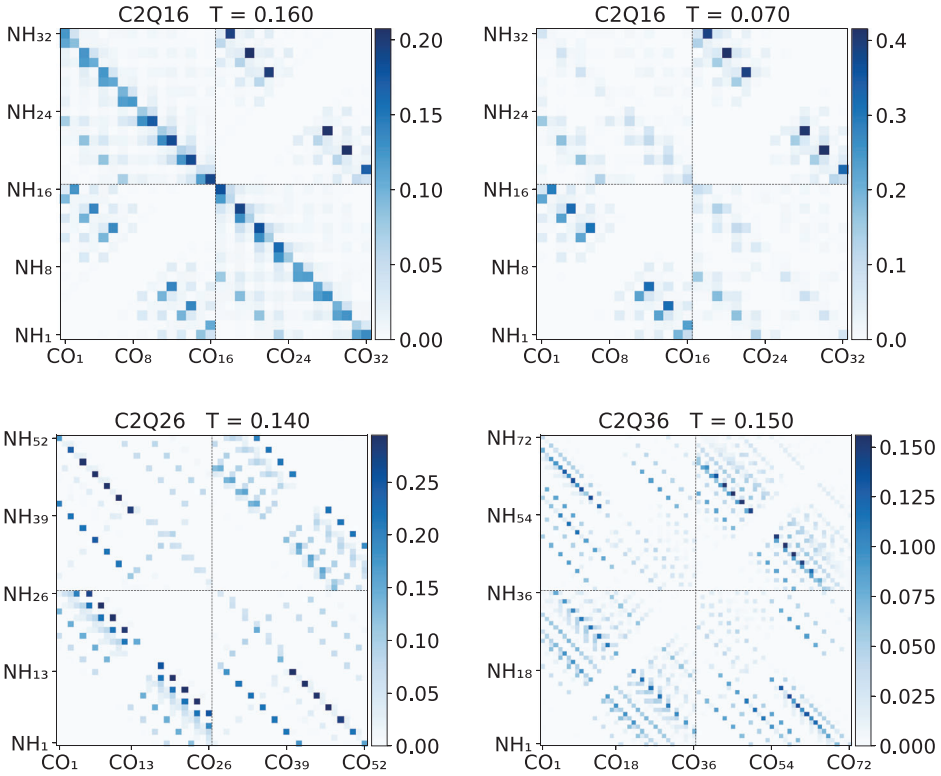



**Figure 8.** Hydrogen bond contact matrices of the folded states of C2Q16, C2Q26 and C2Q36.

belongs to the main diagonal signature in the inter-quadrants. A corresponding example configuration is shown in **Figure 8** c. The structure consists of up to seven pairs of closed hydrogen bonds, each separated by a non-bonded residue. They create a *beta*-sheet in which both $\beta$-strands consist of a whole PolyQ chain. The second aggregated structure consists of $\beta$-hairpins that aggregate into proto-fibrils of two $\beta$-hairpins. The two $\beta$-turns can be on the same or on opposite sides of the structure. The $\beta$-strands involved in inter-chain contacts align in an anti-parallel fashion. Examples are shown in Figure 8 a,b. The $\beta$-hairpins are represented by the anti-parallel diagonals in the intra-quadrants. The two strands of the $\beta$-hairpin are connected by a turn consisting of a varying number of monomers. Its lower boundary is defined by the model, as the closest possible hydrogen bond is $CO_i$-$NH_{i+4}$. The inter-chain contacts of the $\beta$-sheet are found in the diagonals in the inter-quadrants. Since both $\beta$-hairpins and aggregated $\beta$-sheets are formed, the hydrogen bonds of the monomers in each chain are assigned to either hairpin or aggregate in an alternating fashion. If monomers $i$ and $i+2$ bond to the hairpin structure, monomer $i+1$ can bond to the aggregated $\beta$-sheet, as is the case for configurations in Figure 8 a,b).

At $T = 0.09$ the probability of formed hydrogen bonds increases even further. The signatures of the full main diagonal in the inter-quadrants disappear. What remains in the inter-quadrants are signatures of half diagonals, that fit aggregated $\beta$-sheets of $\beta$-hairpins. At this temperature these are the dominant structures.

When cooling even further, the system abandons less optimal, off-center $\beta$-hairpins of only two bonded monomers, namely contacts of $Q_4$–$Q_{14}$ and $Q_6$–$Q_{12}$, and their equivalent in the other chain $Q_{18}$–$Q_{28}$ and $Q_{20}$–$Q_{26}$. What remains are $\beta$-hairpins with six hydrogen bonds that aggregate with six inter chain hydrogen bonds.

In conclusion, below the first transition temperature $T^{\star}_{\mathrm{can}} = 0.168$, aggregated $\beta$-hairpins and $\beta$-sheets of whole anti-parallel chains can be found. When cooling further, the whole-chain $\beta$-sheets disappear and the aggregated $\beta$-hairpins become the dominant structure. There is another change in structure, when going to temperatures below the PT at $T^{\star}_{\mathrm{can}} = 0.06$. At these low temperatures off-center $\beta$-hairpins disappear, to further increase the number of hydrogen bonds formed.

*C2Q16*: The C2Q16 system shows similar behavior to the C2Q14 system, as seen in the HB matrices in the top row of **Figure 9**. Below the PT temperature $T^{\star}_{\mathrm{can}} = 0.174$ two structures are present. The full main-diagonal signatures in the inter-quadrants indicate $\beta$-sheets consisting of whole chains. Half-diagonals in the inter-quadrants and full-diagonals in the intra-quadrants indicate $\beta$-hairpins that aggregate to $\beta$-sheets. When cooling further, the full main-diagonal disappears (see Figure 9 top right) and the aggregated $\beta$-hairpins become the dominant structure. An example structure for the aggregated $\beta$-hairpin is shown in Figure 8 d).

*C2Q26 and C2Q36*: The systems of longer chain length show only one PT from a random coil to a regular folded state. The HB matrices of the C2Q26 and C2Q36 system are shown in the bottom row of Figure 9 as examples. In both matrices we see $\beta$-hairpin signatures in the intra-quadrants with corresponding

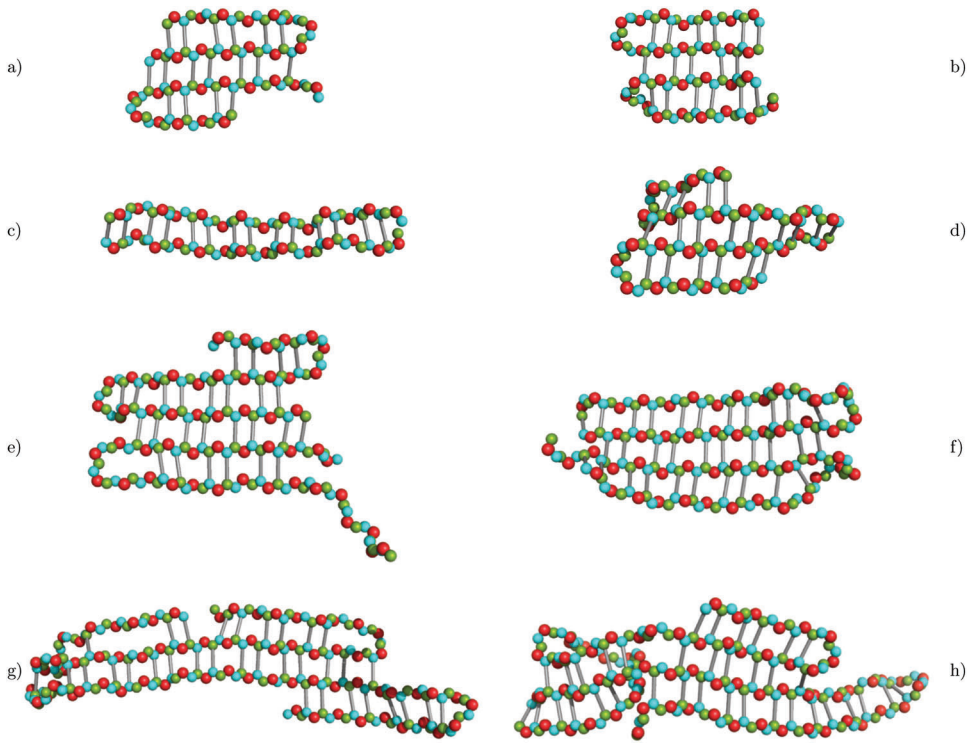



**Figure 9.** Configurations of folded states for various systems. Side-chain beads are hidden to improve visibility. The system sizes of the shown structures are: C2Q14 for a), b) and c); C2Q16 for d); C2Q26 for e) and f); C2Q36 for g) and h).

$\beta$-sheet aggregation patterns in the inter-quadrants. The full main diagonal signatures found in C2Q14 and C2Q16 are absent, suggesting that no aggregation of completely extended chains occurs for the longer chains. This again fits well with the simulation results by Chen et al.[10]

The $\beta$-sheet structures for longer chains are much more diverse when compared to the shorter chains discussed above. Most notably, a single PolyQ chain can exhibit more than one $\beta$-turn. This can lead to an S-shape with a three-stranded $\beta$-sheet, as in the top chain in Figure 8 e). Three-stranded $\beta$-sheets for longer chain length were already found by Marchut et al. for C1Q32 in PRIME20.[31] The way in which the second chain bonds to a chain with double $\beta$-turn varies and influences the double chain structure as seen for C2Q36 in Figure 8 g). Double hairpin configurations, like in Figure 8 f) are also present for longer chains.

Next to the increased number of possible $\beta$-turns, longer chains can exhibit more complicated structural motives. An example are loops in a $\beta$-strand, in which one chain changes the side of the other chain it bonds to (Figure 8 h)). Still, this configuration fits the $\beta$-sheet signature in the HB-matrices.

### 3.2.3. Intra- versus Inter-Molecular Energy

As shown in the previous sections, all systems show a PT from random coil structures to regular folded and aggregated chains. However, in all the above analyses, the two transitions of folding and aggregation appear to happen simultaneously. In order to better separate the two, we look at the number of hydrogen bonds between beads of different chains ($N_{\mathrm{HB,\,inter}}$) and of hydrogen bonds between beads of the same chain ($N_{\mathrm{HB,\,intra}}$). On the left side of **Figure 10** the difference $\Delta N_{\mathrm{HB}} = N_{\mathrm{HB,\,intra}} - N_{\mathrm{HB,\,inter}}$ is shown. Thus, values below zero indicate a dominance of inter-chain hydrogen bonds, whereas values above zero indicate a dominance of intra-chain hydrogen bonds. The PT temperatures from the canonical heat capacity analysis $T^{\star}_{\mathrm{can}}$ are indicated by circles. At high temperatures far above $T^{\star}_{\mathrm{can}}$, all chain lengths show $\Delta N_{\mathrm{HB}} > 0$ because the chains are separated and the only hydrogen bonds that occasionally form are within the same chain. When cooling toward the PT temperature, the graphs show different behavior and can be grouped into two groups. Graphs in the first group of shorter chains ($N = 14, 16, 18, 20, 24$) have a positive slope at $T^{\star}_{\mathrm{can}}$. Below $T^{\star}_{\mathrm{can}}$, $\Delta N_{\mathrm{HB}}$ reaches a minimum, below, which $\Delta N_{\mathrm{HB}}$ increases and intra HB contacts drive the transition to lower energies. The second group consists of longer chains ($N = 22, 26, 28, 36$), which have a negative slope at $T^{\star}_{\mathrm{can}}$. It is worth noting, that the chain lengths in the groups are not in consecutive order. At the border between the longer and shorter chains, $N = 22$ belongs to the group of larger chains and $N = 24$ belongs to the group of shorter chains. We conclude that there is a transition region between the two groups in which a systems group affiliation can not be assigned without ambiguity.

Graphs in group two remain at $\Delta N_{\mathrm{HB}} > 0$ over the entire energy range. The difference between the groups becomes more evident when looking at the temperature derivative of $\Delta N_{\mathrm{HB}}$, which is shown on the right side of Figure 10. For group one, $d(\Delta N_{\mathrm{HB}})/dT$ has a positive maximum close to $T^{\star}_{\mathrm{HB}}$. The driving structural change at the transition temperature is the formation of inter-chain HBs, which is the aggregation of the two chains. Below $T^{\star}_{\mathrm{can}}$, $d(\Delta N_{\mathrm{HB}})/dT$ becomes negative and intra-chain HB

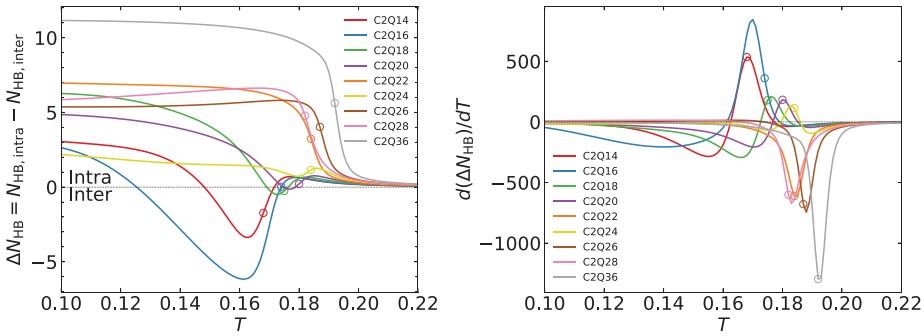



**Figure 10.** Left side: Difference of intra- and inter-molecular hydrogen bond energy as a function of temperature. Right side: Temperature derivative of the left side.

formation, meaning the folding of the chains, drives structural change. Consequently, in group two, $d(\Delta N_{HB})/dT$ has a negative minimum at $T^{\star}_{HB}$. When going to low temperatures, in all systems the intra-chain hydrogen bonds become dominant. As can be seen from the hydrogen bond contact matrices at temperatures below the transition temperature (Figure 9), the chains form hairpin structures with the $\beta$-turn at or close to the middle of the chain (centered $\beta$-hairpins). This maximizes $N_{HB,\,intra}$. The centered $\beta$-hairpins aggregate with hydrogen bonds forming between segments of anti-parallel alignment. In these configurations $N_{HB,\,intra} > N_{HB,\,inter}$.

### 3.2.4. Temperature Mapping

In the previous sections, we identified PT temperatures of PolyQ chains of different chain lengths. We identified folding and aggregation transitions for all systems in an elevated temperature range ($0.168 \leqslant T \leqslant 0.192$). For a meaningful interpretation of the results, it is necessary to convert the reduced model temperatures $T$ to actual physical temperatures $T'$. This is done via $T' = \epsilon_{HB} T/k_B$, as has already been shown in Section 2.1. The challenge is to determine $\epsilon_{HB}$, which is the effective HB strength of the system. As PRIME20 uses a mean field approach for its bead interaction potentials, $\epsilon_{HB}$ also includes peptide-solvent interactions or HBs in side chains that are not included in the model. Thus, $\epsilon_{HB}$ is specific to the peptides in the system.

Recently, the conversion to physical temperatures for PolyQ systems has been performed by Böker et al.[15] We will use the temperature conversion that was derived for model variant B, as this is the model we use in our work. With the resulting conversion formula

$$T'[\mathrm{K}] = 1650\mathrm{K} \cdot T, \tag{10}$$

where $T'$ is the physical temperature in Kelvin and $T$ is the reduced temperature in the PRIME20 model, we determine room temperature in the PRIME20 model as 300K = 0.182. We look again at the PT peak temperatures in the canonical heat capacity, now with Kelvin temperature scale, in the plot on the right in Figure 4. We observe, that the folding and aggregation transitions of all simulated system sizes occur at around room temperature. This is to be expected for PolyQ systems, due to their involvement in amyloid disorders. Looking at the temperatures in Figure 4 in more detail, we divide the systems in two groups. Using the linear extrapolation, short chain systems with $N \leqslant 21$ show the transition peak below room temperature of 300K, while long chain systems with $N \geqslant 22$ show the transition peak above room temperature. This behavior is in excellent agreement with experimental results that show that there is a critical chain length (experimentally around $N = 24$[4]) above, which a single chain can act as nucleus for amyloid aggregation.

## 4. Conclusion

We employed the SAMC method, an advanced flat-histogram Monte Carlo simulation method, to study the dimerization behavior of PolyQ chains of different chain lengths: $N$ = (14,16,18,20,22,24,26,28,36). We used the intermediate resolution protein model PRIME20, which previously proved to realistically predict folded states of PolyQ. The SAMC method aims to numerically approximate the density of states of the system. The density of states lets us perform a thorough thermodynamic analysis of the PolyQ systems. An analysis of the heat capacity in the microcanonical and the canonical ensembles revealed pseudo-phase transitions at chain length dependent transition temperatures. The conversion from reduced model temperatures to physical temperatures located the PTs to be around room temperature. More precisely, the PTs in long chain systems are above room temperature, whereas the PTs in short chain systems are situated below room temperature. PolyQ chains being in a collapsed state at room temperature has been suggested by fluorescence resonance energy transfer studies.[18]

In order to characterize structures formed in the low temperature phase, we analyzed observables derived from the tensor of gyration. They revealed elongated structures in the low temperature phase suggesting that the individual chains form $\beta$-hairpins. To confirm the single chain configurations and additionally investigate aggregated structures we used HB matrices. They showed a clear dominance of $\beta$-hairpins (intra-chain contacts) and aggregated $\beta$-strands (inter-chain contacts) in the low temperature phase. Finding aggregated $\beta$-hairpins is in good agreement with experimental results.

In the peak analysis of the canonical heat capacity, we cannot distinguish between the aggregation and the folding transition. Both are united in a single peak. In the microcanonical

heat capacity analysis, the two transitions are visible separately. However, they appear to happen simultaneously. In order to separate the folding and the aggregation transition we looked at the difference in the number of intra- and inter-chain HB contacts as a function of temperature. From this analysis we draw the conclusion that for short chains with $N \leqslant 24$ the aggregation of two chains is the driving structural change of the transition when cooling. On the other hand, for longer chains with $N > 24$ the folding of a single chain precedes the aggregation, meaning the folded single chain acts as the nucleus for the aggregation transition. We can thus confirm the experimental and computational findings that the critical nucleus size for single chain aggregation nucleation of PolyQ becomes *one* for chain lengths over 24.[4,5,10,18]

## Acknowledgements

The authors acknowledge funding by German Science Foundation (DFG) under Project no. 189853844 - TRR 102.

Open access funding enabled and organized by Projekt DEAL.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

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

## Keywords

amyloid formation, polyglutamine, PRIME20, Stochastic Approximation Monte Carlo (SAMC)

Received: November 23, 2022
Revised: March 23, 2023
Published online: June 6, 2023

[1] P. H. Nguyen, A. Ramamoorthy, B. R. Sahoo, J. Zheng, P. Faller, J. E. Straub, L. Dominguez, J.-E. Shea, N. V. Dokholyan, A. D. Simone, B. Ma, R. Nussinov, S. Najafi, S. T. Ngo, A. Loquet, M. Chiricotto, P. Ganguly, J. McCarty, M. S. Li, C. K. Hall, Y. Wang, Y. Miller, S. Melchionna, B. Habenstein, S. Timr, J. Chen, B. Hnath, B. Strodel, R. Kayed, S. Lesné, et al., *Chem. Rev.* **2021**, *121*, 2545.

[2] M. E. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, C. Lin, L. Srinidhi, G. Barnes, S. A. Taylor, M. James, N. Groot, H. MacFarlane, B. Jenkins, M. A. Anderson, N. S. Wexler, J. F. Gusella, G. P. Bates, S. Baxendale, H. Hummerich, S. Kirby, M. North, S. Youngman, R. Mott, G. Zehetner, Z. Sedlacek, A. Poustka, A.-M. Frischauf, H. Lehrach, A. J. Buckler, D. Church, L. Doucette-Stamm, et al., *Cell* **1993**, *72*, 971.

[3] C. Zuccato, M. Valenza, E. Cattaneo, *Physiol. Rev.* **2010**, *90*, 905.

[4] K. Kar, M. Jayaraman, B. Sahoo, R. Kodali, R. Wetzel, *Nat. Struct. Mol. Biol.* **2011**, *18*, 328.

[5] K. Kar, C. L. Hoop, K. W. Drombosky, M. A. Baker, R. Kodali, I. Arduini, P. C. van der Wel, W. S. Horne, R. Wetzel, *J. Mol. Biol.* **2013**, *425*, 1183.

[6] S. Chen, F. A. Ferrone, R. Wetzel, *PNAS* **2002**, *99*, 11884.

[7] M. S. Miettinen, V. Knecht, L. Monticelli, Z. Ignatova, *J. Phys. Chem. B* **2012**, *116*, 10259.

[8] Z. L. Zhou, J. H. Zhao, H. L. Liu, J. W. Wu, K. T. Liu, C. K. Chuang, W. B. Tsai, Y. Ho, *J. Biomol. Struct. Dyn.* **2011**, *28*, 743.

[9] M. Nakano, K. Ebina, S. Tanaka, *J.Mol. Model.* **2013**, *19*, 1627.

[10] M. Chen, M. Tsai, W. Zheng, P. G. Wolynes, *J. Am. Chem. Soc.* **2016**, *138*, 15197.

[11] M. Cheon, I. Chang, C. K. Hall, *Proteins: Structure, Function and Bioinformatics* **2010**, *78*, 2950.

[12] M. Cheon, I. Chang, C. K. Hall, *Biophys. J.* **2011**, *101*, 2493.

[13] V. A. Wagoner, M. Cheon, I. Chang, C. K. Hall, *J. Mol. Biol.* **2012**, *416*, 598.

[14] V. A. Wagoner, M. Cheon, I. Chang, C. K. Hall, *Proteins: Structure, Function and Bioinformatics* **2014**, *82*, 1469.

[15] A. Böker, W. Paul, *The J. Phys. Chem. B* **2022**, *126*, 7286.

[16] R. Wetzel, *J. Mol. Biol.* **2012**, *421*, 466.

[17] A. Vitalis, X. Wang, R. V. Pappu, *J. Mol. Biol.* **2008**, *384*, 279.

[18] R. H. Walters, R. M. Murphy, *J. Mol. Biol.* **2009**, *393*, 978.

[19] A. V. Smith, C. K. Hall, *Proteins: Structure, Function and Genetics* **2001**, *44*, 344.

[20] A. V. Smith, C. K. Hall, *J. Mol. Biol.* **2001**, *312*, 187.

[21] A. V. Smith, C. K. Hall, *Proteins: Structure, Function and Genetics* **2001**, *44*, 376.

[22] A. Böker, Ph.D. thesis, Martin-Luther-Universität Halle-Wittenberg, Germany **2019**.

[23] F. Liang, *J. Stat. Phys.* **2006**, *122*, 511.

[24] F. Liang, C. L. Liu, R. J. Carroll, *J. Am. Stat. Assoc.* **2007**, *102*, 305.

[25] W. Janke, W. Paul, *Soft Matter* **2016**, *12*, 642.

[26] T. Shakirov, *Comput. Phys. Commun.* **2018**, *228*, 38.

[27] T. Shakirov, S. Zablotskiy, A. Böker, V. Ivanov, W. Paul, *European Physical Journal: Special Topics* **2017**, *226*, 705.

[28] K. Qi, M. Bachmann, *Phys. Rev. Lett.* **2018**, *120*, 1.

[29] S. Schnabel, D. T. Seaton, D. P. Landau, M. Bachmann, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2011**, *84*, 1.

[30] D. Gross, *Microcanonical Thermodynamics: Phase Transitions in "small" Systems*, World Scientific lecture notes in physics, World Scientific, **2001**.

[31] A. J. Marchut, C. K. Hall, *Proteins: Structure, Function and Bioinformatics2* **2007**, *66*, 96.