Supporting Information:

Consistent Force Field Captures Homologue Resolved HP1 Phase Separation

Andrew P. Latham and Bin Zhang*

Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139

E-mail: binz@mit.edu

Phone: 617-258-0848

Constrained Maximum Entropy Optimization Algorithm

The constrained maximum entropy optimization algorithm is a generalization of our prior study. For the sake of completeness, below we first outline the key ideas of the original maximum entropy algorithm before introducing modifications used in the current paper.

In Ref. S1, we combined maximum entropy optimization with least square fitting to derive transferable force fields for intrinsically disordered proteins (IDPs). We start with an initial model defined as

$$U_{\mathrm{MOFF}}(\boldsymbol{r}) = U_{\mathrm{backbone}} + U_{\mathrm{memory}} + U_{\mathrm{electrostatics}} + U_{\mathrm{contact}}. \tag{S1}$$

Explicit expressions of the various terms are provided in the *Section: Mathematical Expressions of the Energy Function*. The goal of the algorithm is to introduce corrections to the the pairwise tertiary contact potential, U_{contact}, such that the model can better reproduce experimental data. For any given protein, a linear biasing term can be derived from maximum entropy optimization to correct $U_{\mathrm{MOFF}}(\boldsymbol{r})$ such that the new model, $U_{\mathrm{ME}}(\boldsymbol{r})$, reproduces its radius of gyration,[S2–S4] i.e.,

$$U_{\mathrm{ME}}(\boldsymbol{r}) = U_{\mathrm{MOFF}}(\boldsymbol{r}) + \alpha R_g(\boldsymbol{r}). \tag{S2}$$

α is a unique Lagrange multiplier and R_g is the radius of gyration of a protein configuration, $\boldsymbol{r}$. The biasing strength α can be fine tuned manually for each protein to ensure that the average R_g for simulated protein structures is within 0.5 Å of the experimental value. However, this resulting energy function U_{ME} is not transferable and cannot be directly applied to new proteins without experimental input. We instead desire a transferable energy,

$$U^{\mathrm{new}}_{\mathrm{MOFF}}(\mathbf{r}) = U_{\mathrm{MOFF}}(\mathbf{r}) + \sum_{I,J} \Delta\epsilon_{IJ} C_{IJ}(\mathbf{r}), \tag{S3}$$

which is based on the contacts ($C_{IJ}(\mathbf{r}) = \sum_{i\in I, j\in J} \mathcal{C}(r_{ij})$) between amino acid types I and J,

and the energy formed or gained ($\Delta\epsilon_{IJ}$) by creating such contacts. Here, we note the energy change in the MOFF algorithm (here denoted $\Delta\epsilon_{IJ}$), was originally termed ϵ_{IJ} in Ref. S1. The contact function between a pair of amino acids i and j separated at a distance r_{ij} is defined as

$$\mathcal{C}(r_{ij}) = \frac{1}{2}(1 + \tanh[\eta(r_{\mathrm{o}} - r_{ij})]) \tag{S4}$$

with $r_{\mathrm{o}} = 8$ Å and $\eta = 0.7$ Å^{-1}.

Previously, we then solved for $\Delta\epsilon_{IJ}$ by equating the contact energy to the maximum entropy biasing energy, which takes the form

$$\sum_{I,J} \Delta\epsilon_{IJ}[\mathcal{C}_{IJ}(\mathbf{r}_m^n) - \mathcal{C}_{IJ}^{n,\mathrm{exp}}] \equiv \alpha_n[R_g(\mathbf{r}_m^n) - R_g^{n,\mathrm{exp}}], \quad \text{for } m = 1, \cdots, M \text{ and } n = 1, \cdots, N. \tag{S5}$$

In Eq. S5, $\mathcal{C}_{IJ}^{n,\mathrm{exp}}$ was estimated by averaging the contacts for all structures where the R_g is within 0.05 nm of the original structure. $R_g^{n,\mathrm{exp}}$ is the experimental radius of gyration, and M structures for each of N proteins in our training set were used to ensure sufficient sampling in both contact and sequence space. The two additional terms $\mathcal{C}_{IJ}^{n,\mathrm{exp}}$ and $R_g^{n,\mathrm{exp}}$ were introduced to ensure that the left and right hand side of Eq. S5 reach zero at the same structures.

Eq. S5 can be simplified by casting it in matrix form as

$$\boldsymbol{\Delta\epsilon C} \equiv \boldsymbol{\alpha R_g}. \tag{S6}$$

$\boldsymbol{\Delta\epsilon}$, $\boldsymbol{C}$, $\boldsymbol{\alpha}$, and $\boldsymbol{R_g}$ are matrices for the corrections to contact energy, difference in contact number between the sampled ensemble and structures in which the R_g is within 0.05 nm of the experimental value, the Lagrange multiplier from maximum entropy optimization, and the difference in radius of gyration between the simulated and experimental value, respectively. $\boldsymbol{C}$, $\boldsymbol{\alpha}$, and $\boldsymbol{R_g}$ depend on protein structure and can be found from simulations, which allows for determination of $\boldsymbol{\Delta\epsilon}$. Previously, we utilized least squares fitting, though other fitting

procedures are possible.

Our new optimization scheme adopts the spirit of our previous strategy described above, with additional requirements. According to the energy landscape theory,[S5–S8] a gap in energy between molten globule configurations and the folded state is necessary in order for the protein to fold reliably. Optimization techniques that maximize this gap have proven quite successful at deriving transferable force fields for globular proteins.[S9–S12] Inspired by this methodology, we aimed to ensure that the contact energy of the PDB structure was lower than that of any structure sampled in simulations, up to a tolerance. This requirement takes the form

$$\boldsymbol{\epsilon'}\boldsymbol{C}_{\mathrm{PDB}} \leq \boldsymbol{\epsilon'}\boldsymbol{C}_{\mathrm{sim}} + \gamma\sigma_{\mathbf{sim}}, \tag{S7}$$

where $\boldsymbol{\epsilon'}\boldsymbol{C}_{\mathrm{PDB}}$ are the total contact energies of the PDB structure (see Table S6), and $\boldsymbol{\epsilon'}\boldsymbol{C}_{\mathrm{sim}}$ are contact energies from simulated structures. The last term is a flexible tolerance parameter based on the standard deviation of the energy distribution sampled for a particular protein in the preceding simulations ($\sigma_{\mathbf{sim}}$). Values for γ are shown in Figure S3. Here it is important to notice that the new contact energy matrix ($\boldsymbol{\epsilon'}$) differs from the change in contact energy ($\boldsymbol{\Delta\epsilon}$), and are related by $\boldsymbol{\epsilon'} = \boldsymbol{\Delta\epsilon} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the contact energy from the previous iteration.

We simultaneously solve Eq. S6 for all proteins in our training set and Eq. S7 for the ordered portion of our training set with the interior-point algorithm (Figure 1). As seen previously,[S1] the relationship between energy and contact formation is not perfectly linear, requiring this entire algorithm to be done iteratively. More details on force field optimization are provided in *Section: Simulation Details on Force Field Optimization*.

Mathematical Expressions of Energy Function

The potential energy of a protein in MOFF is defined as

$$U_{\mathrm{MOFF}}(\boldsymbol{r}) = U_{\mathrm{backbone}} + U_{\mathrm{memory}} + U_{\mathrm{electrostatics}} + U_{\mathrm{contact}}. \tag{S8}$$

As described below, U_{backbone}, U_{memory}, and $U_{\text{electrostatics}}$ describe basic features of protein structure, while the U_{contact} is the tertiary interaction optimized by our maximum entropy procedure.

U_{backbone} and U_{memory} are secondary structure potentials dependent on the input protein conformations. They are defined using the PDB structure for ordered proteins and from I-TASSER structure predictions for disordered proteins.[S13,S14] The backbone energy is defined as

$$U_{\text{backbone}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihed}}. \tag{S9}$$

The bonding potential $U_{\text{bond}} = \sum_i V_b(r_{i,i+1})$ with

$$V_b(r_{i,i+1}) = \frac{k_b}{2}(r_{i,i+1} - r_0)^2. \tag{S10}$$

We used $k_b = 1000$ kJ mol^{-1}nm^{-2} and $r_0 = 0.38$ nm. The angular potential $U_{\text{angle}} = \sum_i V_a(\theta_i)$ with

$$V_a(\theta_i) = \frac{k_a}{2}(\theta_i - \theta_0^i)^2. \tag{S11}$$

θ_i is the angle formed between the three consecutive beads i, $i+1$, and $i+2$. $k_a = 120$ kJ mol^{-1} deg^{-2}. The dihedral potential $U_{\text{dihed}} = \sum_i V_d(\phi_i)$ with

$$V_d = k_d[(1 - \cos(\phi_i - \phi_0^i)) + 0.5(1 - \cos(3(\phi_i - \phi_0^i)))]. \tag{S12}$$

ϕ_i is the dihedral angle formed between the four consecutive beads i, $i+1$, $i+2$, and $i+3$. $k_{\text{dihed}} = 3$ kJ mol^{-1}. Equilibrium values for each angle and dihedral (θ_0^i, ϕ_0^i) were taken from the PDB structures for ordered proteins, and I-TASSER structure predictions for disordered proteins.[S14]

U_{memory} takes the from $U_{\text{memory}}(\mathbf{r}) = \sum_{i,j} V_{\text{memory}}(r_{ij})$ with

$$V_{\text{memory}}(r_{ij}) = \epsilon_{\text{memory}}\left[5\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij}^0}{r_{ij}}\right)^{10}\right], \tag{S13}$$

where r_{ij} is distance between atoms i and j, and r_{ij}^0 is the equilibrium distance taken from initial structures. During training, $\epsilon_{\text{memory}} = 6$ kJ mol^{-1} in ordered proteins, and $\epsilon_{\text{memory}} = 3$ kJ mol^{-1} in disordered proteins. Equilibrium distances (r_{ij}^0) are taken from PDB structures or I-TASSER predictions.[S14] This potential was restricted to regions identified as alpha helices by the STRIDE algorithm's assessment of the input structure.[S15]

$U_{\text{electrostatics}}$ accounts for electrostatic interactions between charged residues. Based on the Debye-Hückle theory, it can be approximated as

$$U_{\text{electrostatics}} = \sum_{i,j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}\epsilon(r_{ij})} \exp(-r_{ij}/\lambda_D), \tag{S14}$$

where ϵ_0 is the permittivity of free space, λ_D is the Debye screening length, r_{ij} is the distance between particles i and j, and q_i and q_j are charges of particles i and j. To maximize accuracy at a minimal computational cost, we used a distance dependent dielectric constant, $\epsilon(r_{ij})$, to capture the change in the solvation environment upon protein folding.[S16,S17] This new dielectric takes the form

$$\epsilon(r_{ij}) = A + \frac{B}{1 + ke^{-\lambda B r_{ij}}}, \tag{S15}$$

where $A = -8.5525$, $k = 7.7839$, $\lambda = 0.03627\text{nm}^{-1}$, and $B = \epsilon_{\text{w}} - A$, where $\epsilon_{\text{w}} = 78.4$ is the dielectric constant of water. These parameters were shown to work well for describing interactions at protein surfaces,[S17,S18] and a plot for $\epsilon(r_{ij})$ is provided in Figure S1.

U_{contact} describes tertiary interactions, and is the sum of the contact energies between all pairs of amino acids i and j separated at a distance r_{ij}, given by $U_{\text{contact}} = \sum_{ij} V_{\text{nb}}(r_{ij}, \epsilon_{IJ})$. The pairwise potential is a combination of excluded volume and contact terms given by

$$V_{\text{nb}}(r_{ij}, \epsilon_{IJ}) = \frac{|\epsilon_{IJ}|\sigma_{IJ}^{12}}{r_{ij}^{12}} + \epsilon_{IJ}\mathcal{C}(r_{ij}). \tag{S16}$$

I and J correspond to the amino acid type of bead i and j. $\sigma_{IJ} = \frac{\sigma_I + \sigma_J}{2}$ is defined using the individual size of each amino acid type (Table S1).

Simulation Details on Force Field Optimization

Paramterization of MOFF force field was based on simulations of 23 sequences, 7 from ordered proteins and 16 from disordered proteins. Experimental values of radius of gyration (R_g) and ionic strength are in Table S2. To carry out simulations with MOFF in GROMACS, we implemented U_{memory}, $U_{\text{electrostatics}}$, and U_{contact} through tabulated potentials. Examples on how to generate these tables are provided in the Scripts folder of our GitHub. U_{backbone} makes use of native GROMACS functionality. SMOG was used to extract equilibrium angles, dihedrals, and distances from initial structures.[S13]

At each iteration, we carried out the following steps to update the force field.

(1): We carried out two independent replica exchange simulations for each one of the 23 proteins using U_{MOFF} and U_{ME}. Each simulation consisted of 6 replicas (300, 320, 340, 360, 380, and 400 K) and lasted for 4×10^7 steps. We used a time step of 10 fs and attempted exchanges every 100 steps, with odd pairs on odd attempts and even pairs on even attempts. The first 1×10^7 steps were excluded for equilibration, and data were recorded every 2×10^4 steps, resulting in 1500 structures from each simulation.

(i) Initial configurations for these simulations were taken from PDB structures for ordered proteins and I-TASSER structure predictions for disordered proteins. Proteins were placed in a cubic simulation box with side lengths of 50 nm. We then performed the steepest descent energy minimization for 10000 steps with a cutoff in the change in energy of 10 kJ mol^{-1} before using these structures as starting configurations.

(ii) To determine U_{ME} (Eq. S2) for each protein, we used the same force field parameters as U_{MOFF}, but placed a linear bias on the R_g using the PLUMED plugin.[S19] The values for α in Eq. S2 were determined through manual tuning. α would be positive if $R_{g,i}^{\text{sim}} > R_{g,i}^{\text{exp}}$, and negative if $R_{g,i}^{\text{sim}} < R_{g,i}^{\text{exp}}$. With this information, we began the first iteration by scanning possible values of α with 5 kJ mol^{-1} nm^{-1}

increments to find an appropriate range for each protein. We then scanned that range with 0.5 kJ mol^{-1} nm^{-1} increments. At each subsequent iteration, we began with the α from the previous iteration, and tried shifting it in 0.5 kJ mol^{-1} nm^{-1} increments. In all cases, this process was repeated until $|R_{g,i}^{\mathrm{exp}} - R_{g,i}^{\mathrm{ME}}| < 0.05$nm, where $R_{g,i}^{\mathrm{exp}}$ is the experimental radius of gyration and $R_{g,i}^{\mathrm{ME}}$ is the average R_g from the 1500 structures sampled in the maximum entropy ensemble. This optimization was done for each protein i to determine a unique α_i, and needs to be independently repeated at each iteration of force field optimization. More advanced algorithms to determine α are possible, and would be particularly useful if biases were placed on multiple variables.[S20]

Ready-to-run simulation files for the first and last iteration are available in the optimization folder of our GitHub.

(2): From the two simulations performed for each protein, we collected 3000 structures to build the list of equations in Eq. S6.

Before solving these equations, we first reduced noise and limited the change in our force field from one iteration to the next. We removed noise from our data using single value decomposition (SVD) to reconstruct the contract matrix, $\mathbf{C}$, in Eq. S6. To do this, we decompose $\mathbf{C}$ as

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{S17}$$

where $\mathbf{\Sigma}$ is a matrix whose diagonal entries, σ_i, are the singular values of $\mathbf{C}$. We then keep only the largest values of σ_i that account for 95% of the variance, and reconstruct the matrix $\mathbf{C}$ using Eq. S17, as seen in this script.

(3): For ordered proteins in our training set, we added additional equations specified in Eq. S7 for all 3000 structures of each protein. In Eq. S7, γ represents how tightly the constraint is enforced. The value for γ in Eq. S7 begins at 0, but is increased if no

solution can be found at the current value of γ. The values used in our optimization are given in Fig. S3B.

(4): We then solved the linear Eq. S6, with the constraint Eq. S7. The change in energy from one iteration to the next was limited with the requirement $|\Delta\epsilon_{IJ}| < 2\sigma_{\mathrm{MJ}}$, where σ_{MJ} is the standard deviation of all energies in the initial MJ matrix used in the first iteration of the optimization. This constraint prevents an iteration of the algorithm from giving unrealistically strong interactions to individual contact pairs. MATLAB's constrained linear least square solver solves this equation using the interior point algorithm, as seen in this script.

(5): The amino acid contact energy was updated using values obtained from solving Eq. S6 and S7, $\epsilon'_{IJ} = \epsilon_{IJ} + \Delta\epsilon_{IJ}$.

(6): We further normalize the contact energy (ϵ'_{IJ}) to ensure that amino acids maintain physiological size as the contact strength changes. Specifically, if $\epsilon'_{IJ} < 0$, the corresponding effective contact energy, $\epsilon_{IJ}^{\mathrm{new}}$ was chosen to normalize $V_{\mathrm{nb}}(r_{ij}, \epsilon_{IJ}^{\mathrm{new}})$ such that $V_{\mathrm{nb}}(\sigma_{IJ}, \epsilon_{IJ}^{\mathrm{new}}) = 0$ and $V_{\mathrm{nb}}(r_{ij}, \epsilon_{IJ}^{\mathrm{new}})$ reaches a minimum value of ϵ'_{IJ}. In this way, ϵ_{IJ} and σ_{IJ} can be interpreted as analogues of ϵ and σ in a Lennard-Jones potential. In this formalism,

$$\epsilon_{IJ}^{\mathrm{new}} = \frac{\epsilon'_{IJ} \cdot \epsilon'_{IJ}}{|V_{\mathrm{norm}}|}, \tag{S18}$$

where V_{norm} is the minimum of $V_{\mathrm{nb}}(r_{ij}, \epsilon'_{IJ})$. However, if $\epsilon'_{IJ} > 0$, then $V_{\mathrm{nb}}(r_{ij}) > 0$, and a different form of normalization is required. In this case, we chose

$$\epsilon_{IJ}^{\mathrm{new}} = \frac{\epsilon'_{IJ} \cdot \epsilon'_{IJ}}{V_{\mathrm{norm}}}, \tag{S19}$$

where $V_{\mathrm{norm}} = V_{\mathrm{nb}}(\sigma_{IJ}, \epsilon'_{IJ})$. This normalization ensures $V_{\mathrm{nb}}(\sigma_{IJ}, \epsilon_{IJ}^{\mathrm{new}}) = \epsilon'_{IJ}$. $V_{\mathrm{nb}}(r_{ij}, \epsilon_{IJ}^{\mathrm{new}})$ was then used in simulations for the next iteration. This potential is exemplified in Figure S17.

(7): Finally, the new values of $\epsilon_{IJ}^{\text{new}}$ were used as ϵ_{IJ} in Eq. S16 to update U_{MOFF} for the next iteration of simulations.

Inital contact energies for MOFF

To initialize force field optimization with a list of contact energies, we scaled the Miyazawa-Jerrigan (MJ) potential[S21] by a factor of 0.4, which was shown to best predict R_g for proteins in the training set (Figure S2). These rescaled values were then normalized according to Eq. S18 and S19 to obtain initial values for ϵ_{IJ} in Eq. S16.

Varying the secondary structure potential during optimization

During optimization, we began with weak secondary structure potentials introduced in Eq. S8 and Eq. S9, and then increased their strength to refine structure prediction. This allows our optimization to first distinguish expanded from collapsed structures, and then improve for structure prediction. At the start of the optimization, U_{angle}, U_{dihed}, and U_{memory} were only applied within individual α-helices and β-sheets of ordered proteins, which were determined by STRIDE secondary structure determination of the PDB structure.[S15] The initial energy constants for U_{angle}, U_{dihed}, and U_{memory} are $k_a = 40$ kJ mol^{-1} deg^{-2} (Eq. S11), $k_{\text{dihed}} = 1$ kJ mol^{-1} (Eq. S12), and $\epsilon_{\text{memory}} = 2$ kJ mol^{-1} (Eq. S13) respectively. From iteration 13 onward, we strengthened the potentials and applied them to the entire amino acid sequence for both ordered and disordered proteins. The resulting equations are identical to those described in *Mathematical Expressions of Energy Function*. The final U_{backbone} is the same in both ordered and disordered proteins, with $k_a = 120$ kJ mol^{-1} deg^{-2} (Eq. S11) and $k_{\text{dihed}} = 3$ kJ mol^{-1} (Eq. S12). U_{memory} is slightly weaker in disordered proteins than in ordered proteins with $\epsilon_{\text{memory}} = 6$ kJ mol^{-1} in ordered proteins and $\epsilon_{\text{memory}} = 3$ kJ mol^{-1} in disordered proteins. This allows for increased helix flexibility in disordered proteins.

Determination of T_θ

We determined the theta temperature (T_θ) for all proteins in the training set and validation set. To make this determination, we preformed replica exchange simulations at windows from 80K to 720K for disordered proteins and 200K to 720K for ordered proteins with an increment of 40K. Settings beside temperature were identical to the optimization simulations. Using the simulated protein configurations, we determined the scaling exponent ν as a function of temperature. ν was obtained by fitting the average distance between pairs of amino acids at a given sequence separation to the expression $R(|i-j|) = b|i-j|^\nu$, with $b = 0.55$ nm.[S22,S23] We then approximated T_θ as the temperature at which $\nu = 0.5$ via linear interpolation.

Folding potential for HP1 tertiary structure stabilization

As mentioned in the main text, in its current form, MOFF has not yet achieved consistent accuracy for *de novo* structure prediction. When studying large proteins with both ordered and disordered regions, it is beneficial to include biases that stabilize the tertiary structure. Therefore, in our simulations of HP1, we introduced $U_{\text{fold}}(\mathbf{r}) = \sum_{i,j} V_{\text{fold}}(r_{ij})$, where

$$V_{\text{fold}}(r_{ij}) = \epsilon_{\text{fold}} \left[5 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{10} \right], \tag{S20}$$

where r_{ij} is distance between atoms i and j, r_{ij}^0 is the equilibrium distance taken from the initial structure, and ϵ_{fold} can be determined by fitting to the RMSF of all atom simulations. For the above definition, it is clear that $U_{\text{fold}}(\mathbf{r})$ is a structure-based potential[S13] designed to suppress large fluctuations away from the PDB conformation. The final potential for HP1 simulations used a combination of MOFF and the folding potential $U_{\text{MOFF}}(\mathbf{r}) + U_{\text{fold}}(\mathbf{r})$. With the ordered regions restricted to the PDB conformations, MOFF should provide accurate description of interactions between domains within the same protein and interactions between proteins. We provide a separate script to add such a folding potential to $U_{\text{MOFF}}(\mathbf{r})$.

We applied the folding potential, $U_{\text{fold}}(\mathbf{r})$, to stabilize both the tertiary contacts found

in both chromoshadow and chromo domains, and the contacts found at the interface of the two chromoshadow domains. Contacts were determined by a shadow contact map of the heavy atom structure, with a cutoff distance of 6 Å.[S24] The strength ϵ_{fold} was determined by fitting to the root mean squared fluctuation (RMSF) of the all atom simulations of the protein dimer, as discussed further below. In all-atom simulations, the two chromoshadow domains (residues 116-175) were solvated with water molecules. Monovalent ions were added to neutralize charges of the protein with a concentration of 150mM salt. Simulations were set up with CHARMM-GUI[S25] and ran in the CHARMM36m force field[S26] with the GROMACS simulation package.[S27] Periodic boundaries were enforced with 1 nm between the protein and the nearest side of the simulation box (7.1nm $\times$ 7.1nm $\times$ 7.1nm). After energy minimization, simulations of 10 ns in NVT and 20 ns in NPT were performed for equilibration. We then carried out the NVT production simulations for 100 ns to compute the RMSF of each amino acid. For comparison, we performed five coarse-grained simulations with ϵ_{memory} at 2, 4, 6, 8, and 10 kJ mol^{-1} to compute the corresponding RMSF. The dimer was not stable at 2 and 4 kJ mol^{-1}, and we chose $\epsilon_{\text{memory}} = 6$ kJ mol^{-1} as the final value as it minimized the

$$\chi^2 = \sum_N \frac{(\text{RMSF}_i^{\text{AA}} - \text{RMSF}_i^{\text{MOFF}})^2}{N}. \tag{S21}$$

N is the number of residues in the chromoshadow domain. $\text{RMSF}_i^{\text{AA}}$ is the average RMSF of residue i from all-atom simulation and $\text{RMSF}_i^{\text{MOFF}}$ is the average RMSF of residue i from MOFF simulation with the given strength of pair potential (see Figure S18).

HP1 Slab Simulation Details

Slab simulations were used to determine the critical temperature for HP1 phase separation following the same procedure as Mittal and coworkers.[S28,S29] We began by placing 100 HP1 dimers in a large simulation box (75nm $\times$ 75nm $\times$ 75nm). Previous studies found that this system size was sufficient to prevent finite size effects.[S28] After steepest decent energy

minimization, we performed an NPT simulation for 0.1 μs at 150 K and 1 bar, using a Parrinello-rahman isotropic bariostat and time coupling constant of 1 ps. This NPT simulation results in a dense phase of HP1 dimers in a smaller simulation box (approximately 25nm $\times$ 25nm $\times$ 25nm). We then expanded the z-coordinate of the simulation box by approximately a factor of 20, to 500 nm. This created a dense phase of protein, with dilute phases on either side. Next, a 0.1 μs NVT simulation was conducted to linearly raise the temperature from 150 K to the desired temperature, with a time coupling constant of 100 ps. Finally, 2 μs of production simulations were conducted in the NVT ensemble. The first 1 μs was discarded for equilibration, and the remainder was used for analysis. For HP1α, simulations were conducted at 150 K, 200 K, 250 K, 267 K, 284 K, 300 K, 317 K, 334 K, 350 K, and 400 K. For HP1β, simulations were conducted at 150 K, 200 K, 217 K, 234 K, 250 K, 267 K, 284 K, 300 K, 350 K, and 400 K. For HP1γ, simulations were conducted at 150 K, 200 K, 234 K, 250 K, 267 K, 284 K, 300 K, 317 K, 350 K, and 400 K. These temperatures were chosen to achieve good sampling around the phase separation temperature for each protein. In MJ and MOFF-IDP, fitting was done based on simulations at 150 K, 200 K, 250 K, 300 K, and 400 K.

We then used the following procedure to monitor the collective behavior of HP1 dimers and and determine the critical temperature, T_C. We first computed a center of mass contact matrix between pairs of HP1 dimers with a distance cutoff of 8 nm. Then, using a depth first search algorithm,[S30] we identified the size of the largest cluster of HP1 dimers. The distance to the center of mass of the largest cluster along the z-axis, δz, was used to identify high ($\delta Z < 5$nm) and low density ($\delta Z > 50$nm) region. The corresponding density is denoted as ρ_H and ρ_L, respectively. The critical temperature, T_C, was determined by fitting the density at different temperatures (T) to the following expression

$$\rho_H - \rho_L = A(T_C - T)^\beta, \tag{S22}$$

with $\beta = 0.325$ is the critical exponent. A and T_C are determined by fitting. The minimum temperature is selected so that ρ_L is non-zero. Fitting was performed over a range of temperatures. As Eq. S22 is only valid below the critical temperature, the goodness of the fit deteriorates as density values at temperature higher than T_C were included (Figure S13). We determined the final values for HP1α, HP1β, and HP1γ using densities below and equal to 300 K, 250 K, and 267 K respectively.

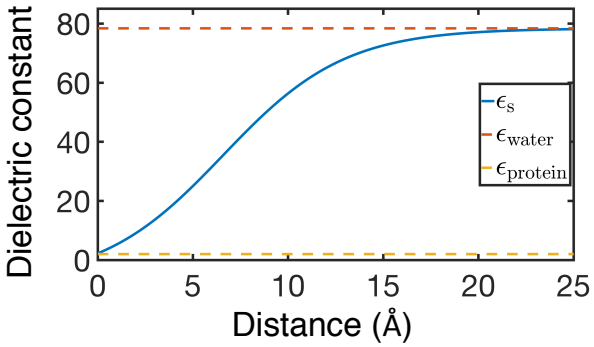

Figure S1: The dielectric constant used in this study is a function of distance (Eq. S15) and continuously switches from the value in a protein environment ($\epsilon_{protein}$, yellow) to the dielectric constant of water (ϵ_{water}, red).

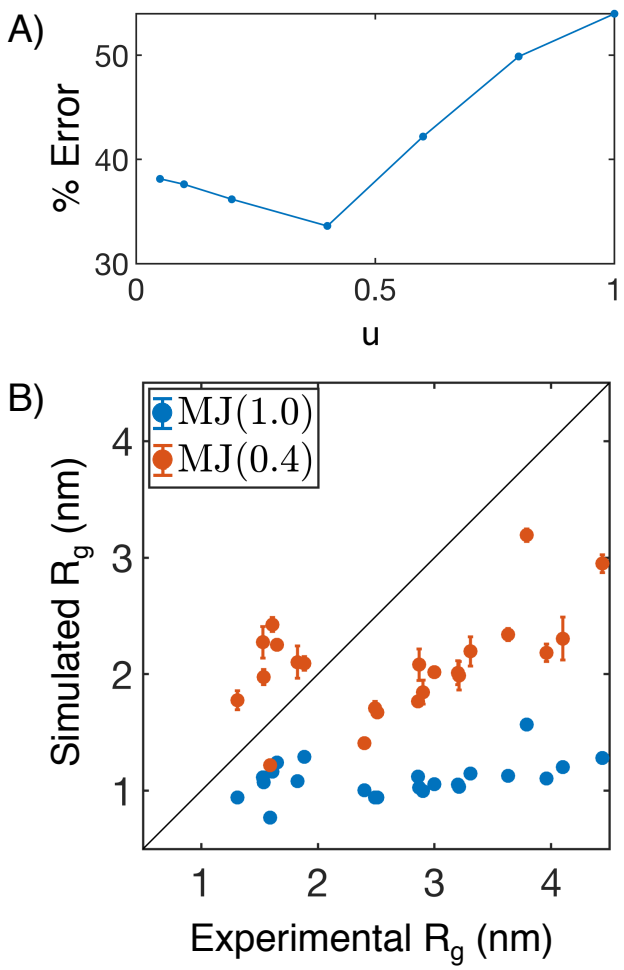

Figure S2: The MJ potential was scaled by a factor of 0.4 to initialize the contact energy (Eq. S16) for force field optimization. (A) The percent error defined in Eq. 4 of the main text as a function of the scaling factor. The error quantifies the difference between simulated and experimental radius of gyration (R_g) for proteins in the training set. It is evident that $u = 0.4$ gives the smallest error. (B) Comparison between simulated and experimental R_g for proteins included in the training set. Results obtained from simulations with the original MJ potential (MJ(1.0)) and the one scaled by a factor of 0.4 (MJ(0.4)) are both shown. Error bars represent the standard deviations after block averaging.

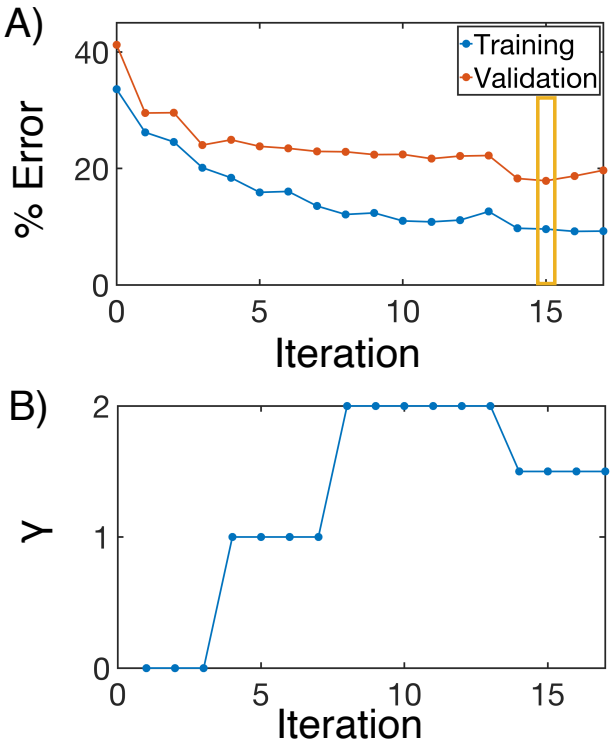

Figure S3: Data from optimization of the MOFF force field. (A) Percent error (Eq. 4 of the main text) as a function of optimization iteration for both the training set and validation set. Optimization is terminated when the following two iterations do not improve the fit to the validation set (yellow box). (B) The strength of the γ parameter was gradually increased along the iteration to relax the energy gap constraint defined in Eq. S7.

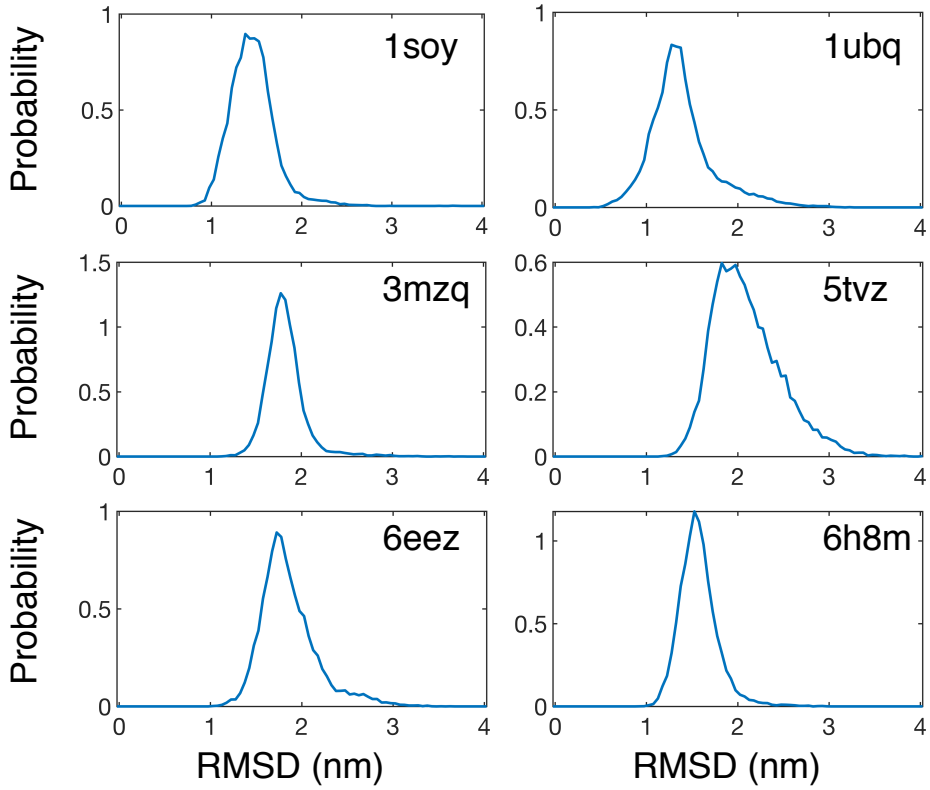

Figure S4: Probability distribution of the root mean squared displacement (RMSD) relative to the PDB structure for ordered proteins in the training set. They were calculated using replica-exchange simulations performed at iteration 15, i.e., with the converged MOFF force field. As explained in *Section: Simulation Details on Force Field Optimization*, these simulations were initialized from the PDB structure, but were given time for equilibration before taking data. We only used the configurations collected for 300 K to compute the probability distributions.

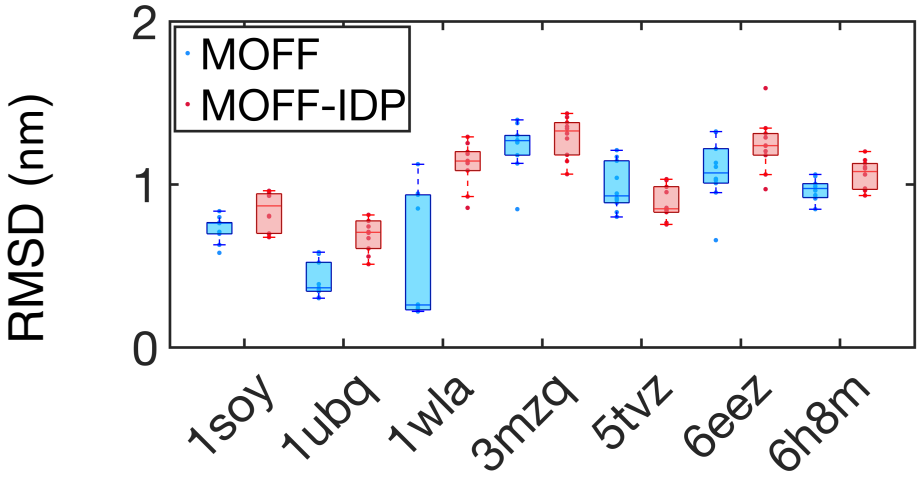

Figure S5: Simulated annealing carried out with MOFF outperforms those with MOFF-IDP in structure prediction. MOFF-IDP is a force field designed for IDPs using the maximum entropy optimization algorithm.[S1] To focus on the impact of non-bonded interactions on structural prediction, we kept the first three terms in Eq. S8 unchanged between MOFF and MOFF-IDP. The two force fields differ in the amino acid contact energy, i.e., ϵ_{IJ} in Eq. S16. Initial structures of the annealing simulations were generated from simulations of 4×10^7 steps at 1000 K. Then, we performed 8×10^7 steps of annealing simulations where the temperature was linearly lowered from 1000 to 100 K. Protein RMSD was collected over the second half of the simulation, and the lowest value was recorded. These annealing simulations were performed 10 times for each protein and force field combination.

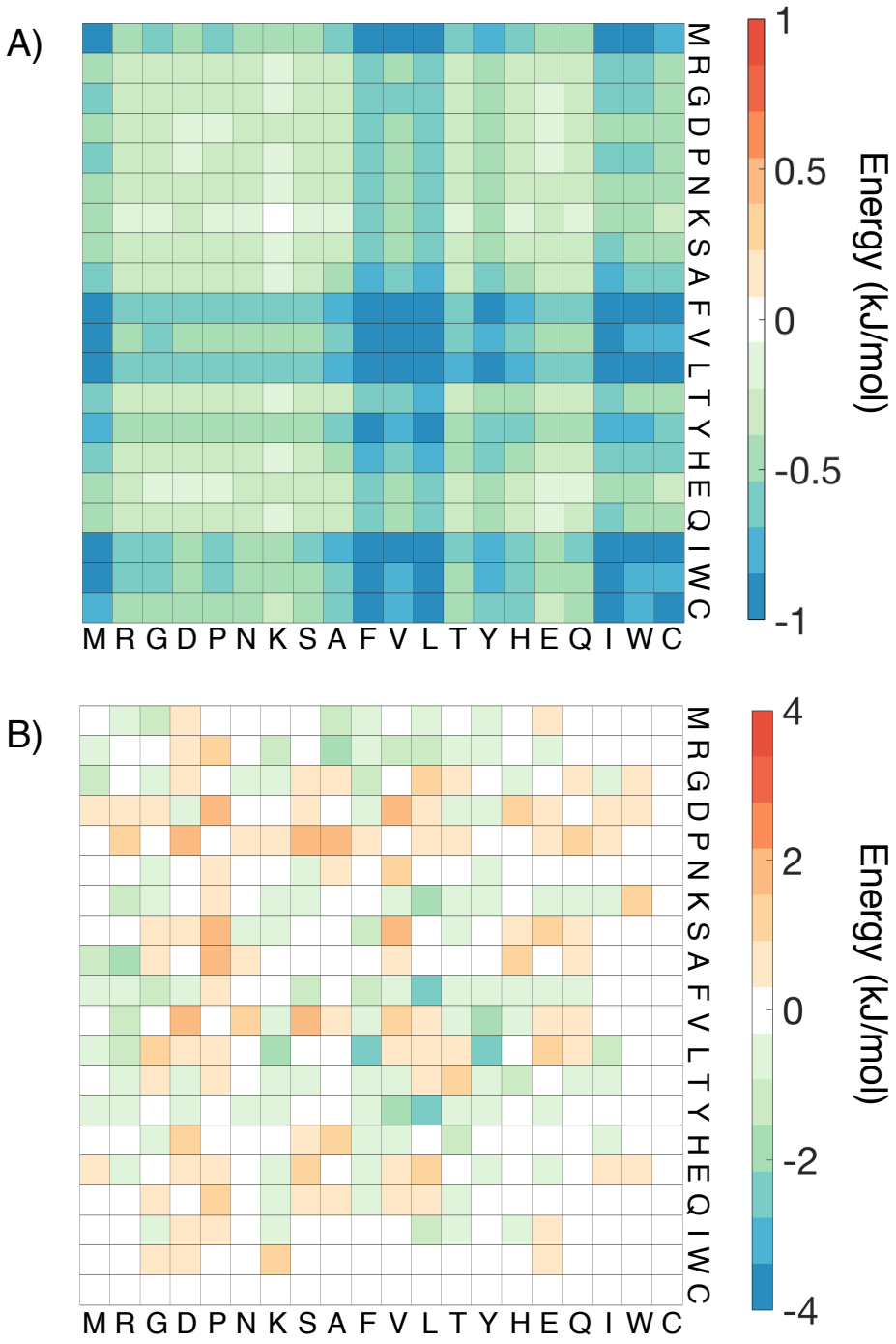

Figure S6: Contact energy matrix for MJ(0.4) (A) and MOFF-IDP (B), to compare with MOFF in Figure 4A of the main text.

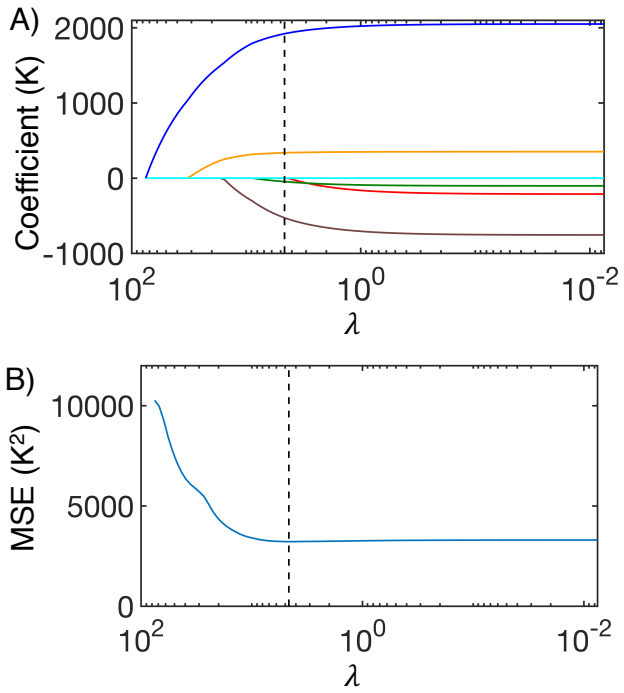

Figure S7: Fitting T_θ with the frequency of amino acids by cluster with the least absolute shrinkage and selection operator (LASSO). (A) The dependence of linear fitting coefficient for each amino acid cluster (c_i in Eq. 5 of the main text) as a function of the regularization parameter (λ). The coloring scheme is the same as that in Figure 4A of the main text (1-red, 2-brown, 3-green, 4-orange, 5-cyan, 6-blue). The dashed line represents the optimal regression value calculated by cross validation. (B) Mean-squared error (MSE) of cross validation fit as a function of λ. Optimal value of λ was chosen to minimize the MSE (black, dashed line).

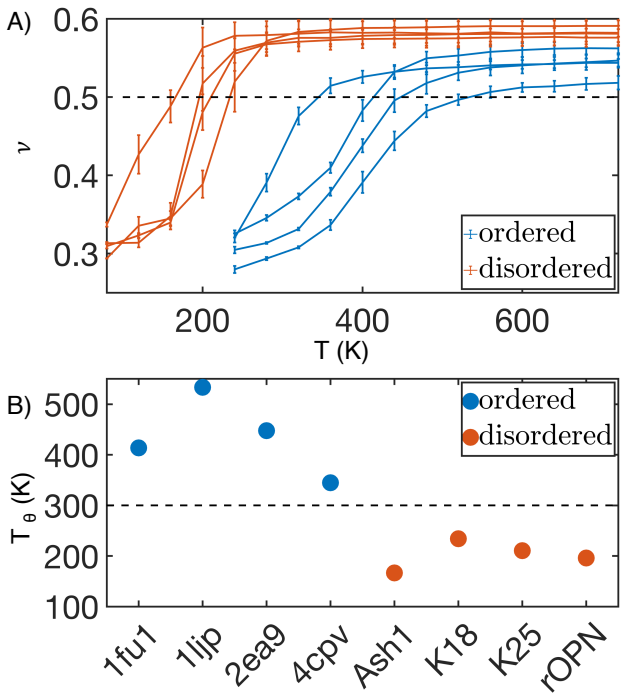

Figure S8: Theta temperature (T_θ) calculations for proteins in the test set. (A) Polymer scaling exponent (ν) as a function of T for ordered (blue) and disordered (orange) protein sequences. (B) T_θ for proteins in the test set. The 300K mark is highlighted as a guide for the eye (dashed, black).

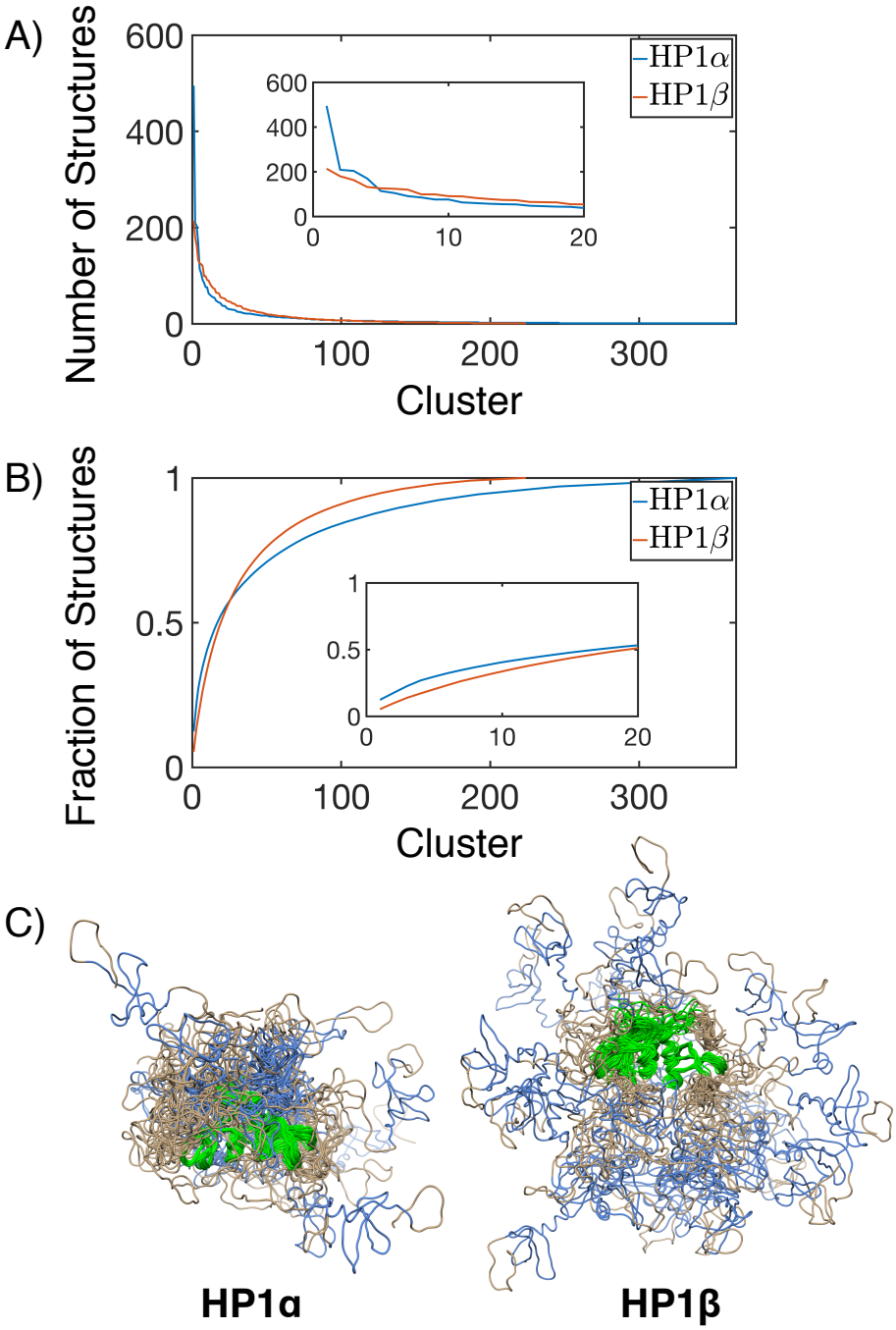

Figure S9: Clustering results for HP1 dimers. (A) Number of structures per cluster for HP1α (blue) and HP1β (red). (B) Fraction of structures assigned to a cluster for HP1α (blue) and HP1β (red). (C) Representative structures from the 20 most populated clusters. The top 20 clusters represent over 50% of the overall ensemble for both proteins.

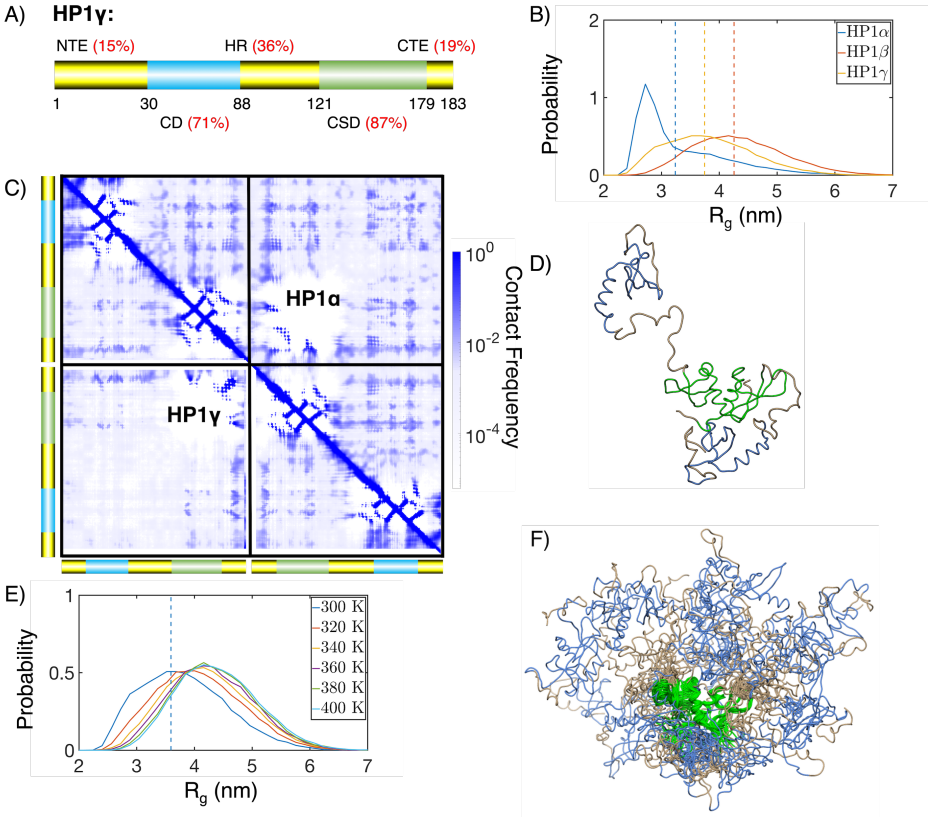

Figure S10: Simulation results for HP1γ. (A) Cartoon diagrams for HP1γ, with the disordered regions shown in yellow and ordered regions in blue and green. The red numbers indicate sequence identity to HP1α for various protein regions.[S31] (B) Probability distributions of the radius of gyration (R_g) for HP1α (blue), HP1β (orange), and HP1γ (yellow). Dashed lines show mean values of each distribution. (C) Contact maps of HP1α (top right) and HP1γ (bottom left), with cross-dimer interactions shown in the diagonal quadrants. (D) Representative structure of HP1γ selected by clustering. (E) Probability distributions of the radius of gyration (R_g) for HP1γ. (F) Representative structures from the 20 most populated clusters.

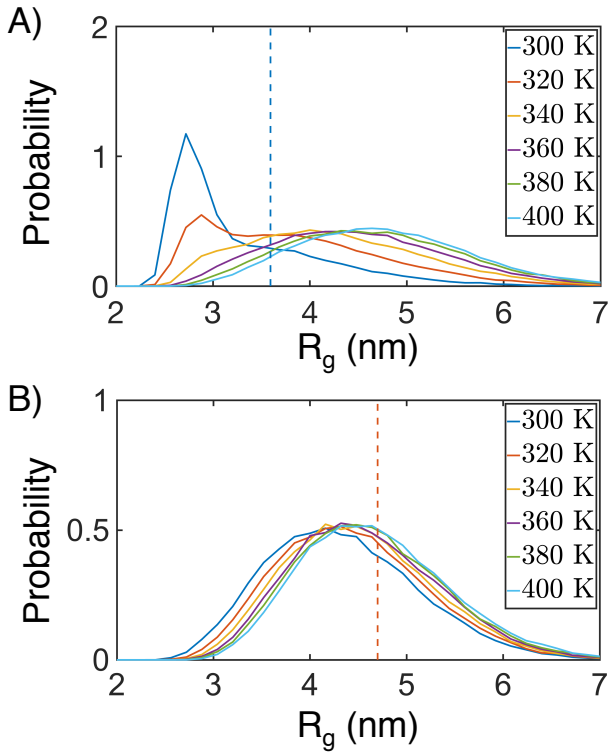

Figure S11: Probability distributions of the radius of gyration (R_g) for HP1α (A) and HP1β (B) at different temperatures. Experimental values at room temperature are shown as dashed lines.

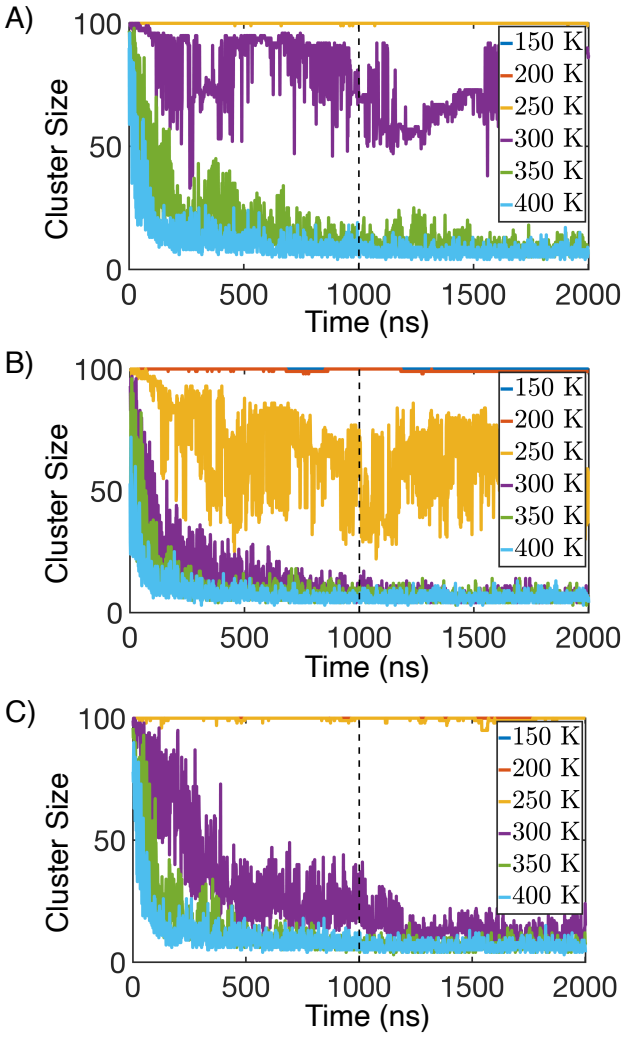

Figure S12: Size of the largest cluster formed by HP1 dimers as a function of time for HP1α (A), HP1β (B), and HP1γ (C) simulations performed at different temperatures.

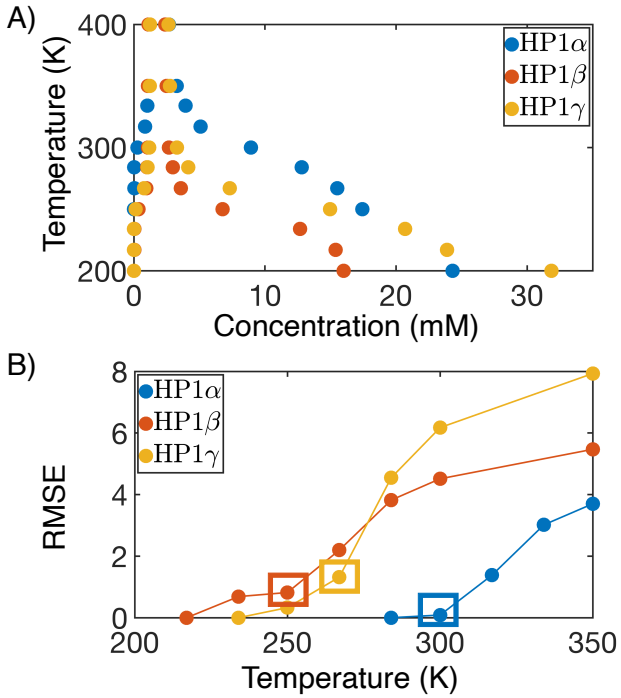

Figure S13: Determining the T_C of HP1 dimers through slab simulations. (A) Concentration as a function of temperature for both the high density (ρ_H) and low density (ρ_L) phases. This data was used to fit Eq. S22. (B) Mean squared error from the fit of Eq. S22 using data below or equal to the temperature T. The best fit was determined as the temperature value before the error increases sharply as indicated by square boxes.

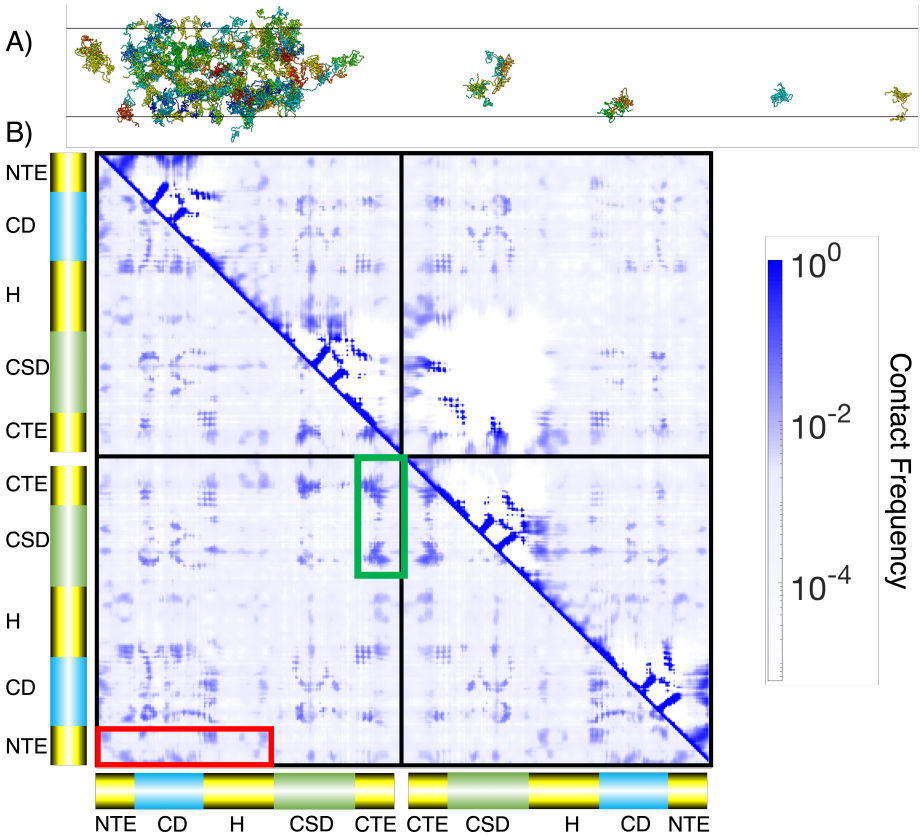

Figure S14: Comparison of inter-dimer and intra-dimer contacts in HP1α. (A) Example configuration of HP1α cluster from slab simulations performed at 300 K, below T_C. (B) Contact map for pairs of amino acids from different dimers (bottom left) and from the same dimer (top right) determined from slab simulations at 300 K. Axis is colored by the sequence diagram in Figure 6A. The disordered NTE interacting with the opposite NTE, CD, or hinge is highlighted in red, and the CTE interacting with the opposite CTE or CSD is highlighted in green.

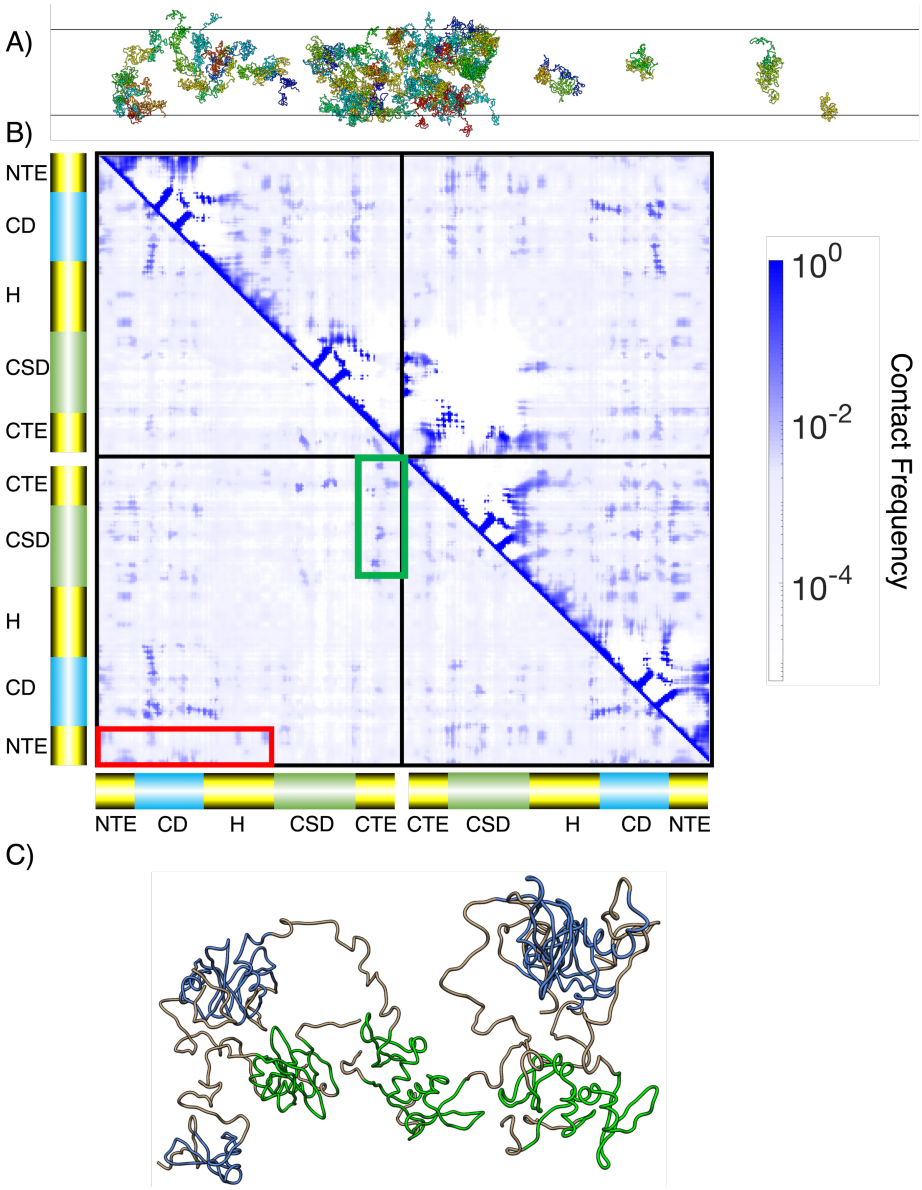

Figure S15: Comparison of inter-dimer and intra-dimer contacts in HP1β. (A) Example configuration of HP1β cluster from slab simulations performed at 250 K, below T_C. (B) Contact map for pairs of amino acids from different dimers (bottom left) and from the same dimer (top right) determined from slab simulations at 250 K. Axis is colored by the sequence diagram in Figure 6A. The disordered NTE interacting with the opposite NTE, CD, or hinge is highlighted in red, and the CTE interacting with the opposite CTE or CSD is highlighted in green. (C) Sample structure of a small cluster of HP1β at 300 K, above the phase separation temperature.

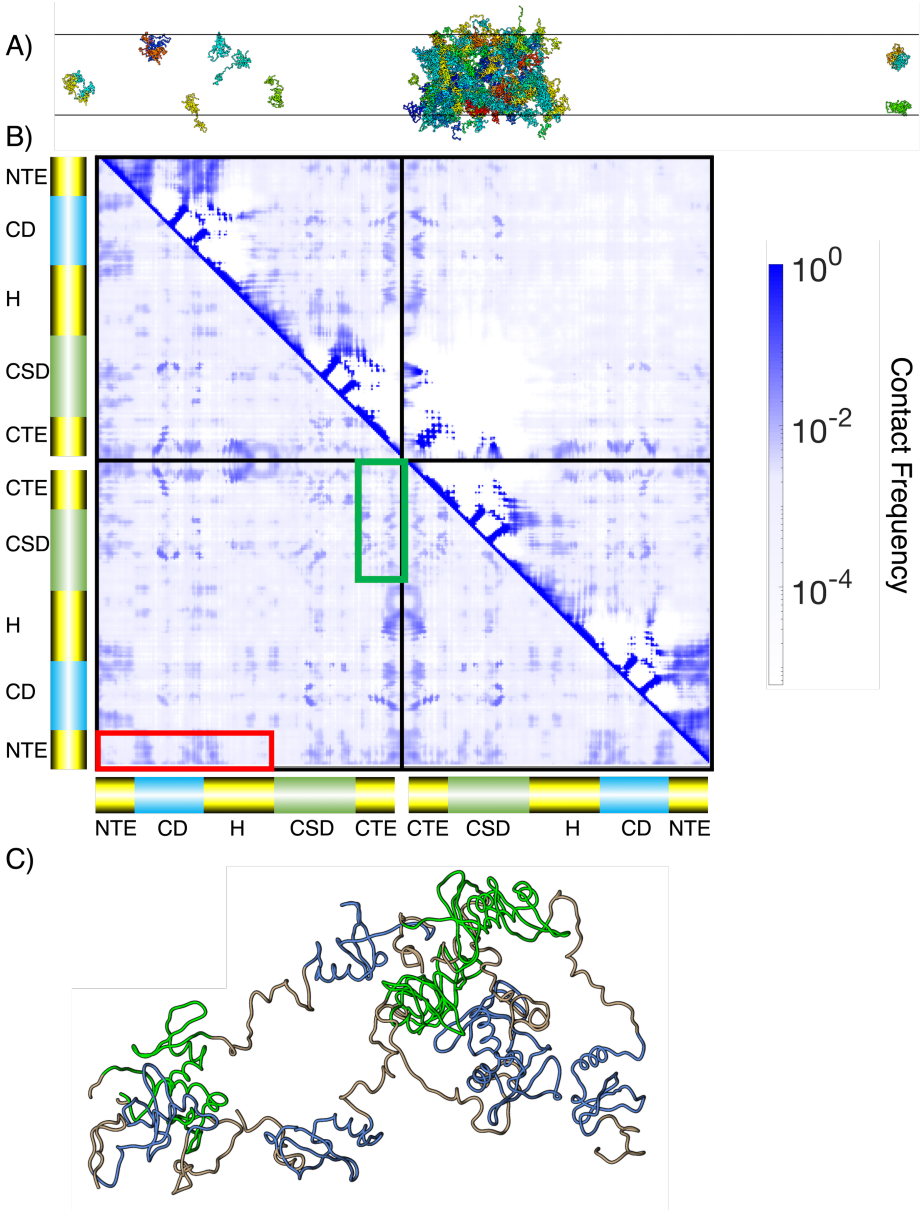

Figure S16: Comparison of inter-dimer and intra-dimer contacts in HP1γ. (A) Example configuration of HP1γ cluster from slab simulations performed at 267 K, below T_C. (B) Contact map for pairs of amino acids from different dimers (bottom left) and from the same dimer (top right) determined from slab simulations at 267 K. Axis is colored by the sequence diagram in Figure S10. The disordered NTE interacting with the opposite NTE, CD, or hinge is highlighted in red, and the CTE interacting with the opposite CTE or CSD is highlighted in green. (C) Sample structure of a small cluster of HP1γ at 300 K, above the phase separation temperature.

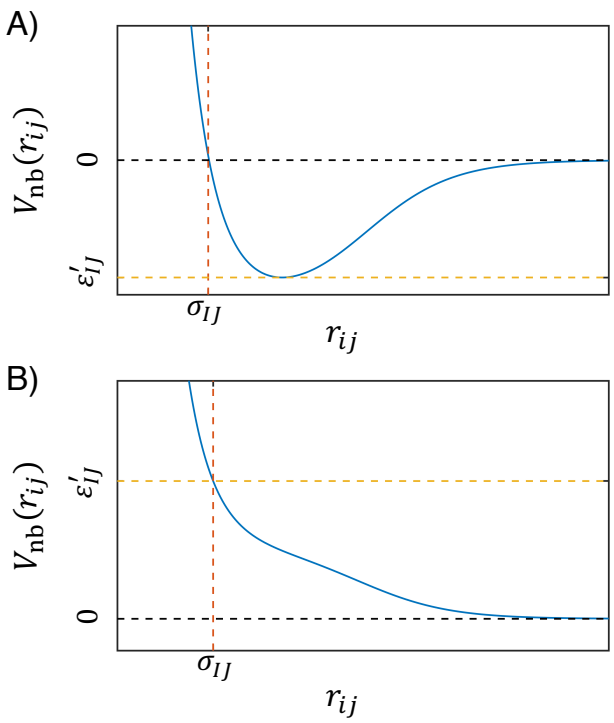

Figure S17: Illustration of $V_{\mathrm{nb}}(r_{ij}, \epsilon_{IJ}^{\mathrm{new}})$ after normalization for $\epsilon'_{IJ} < 0$ (A) and $\epsilon'_{IJ} > 0$ (B).

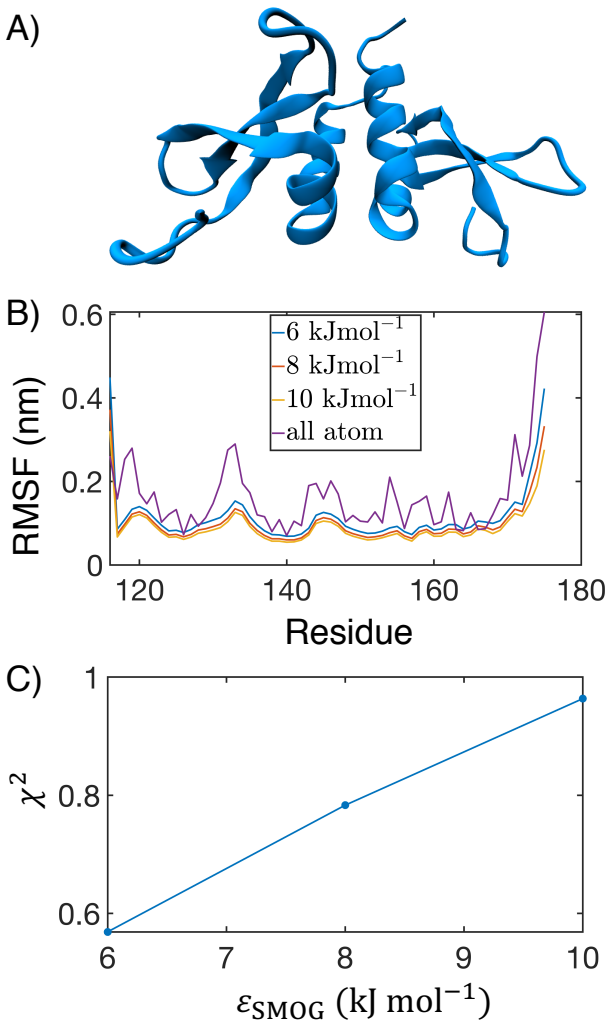

Figure S18: Determining the strength of the memory potential from all-atom simulations. (A) Structure of the CSD dimer used in all-atom and coarse-grained simulations. (B) The RMSF as a function of CSD residue is plotted for a variety of contact strengths, and compared to the RMSF from all atom simulations. (C) The χ^2, defined in Eq. S21, is plotted for each of the contact strengths, revealing 6 kJ mol^{-1} is the optimal value.

Table S1: Amino acid masses, charges, and sizes (σ) used in simulation.

Amino Acid	Mass (amu)	Charge	σ (nm)
ALA	71.08	0	0.504
ARG	156.20	1	0.656
ASN	114.10	0	0.568
ASP	115.10	-1	0.558
CYS	103.10	0	0.548
GLN	128.10	0	0.602
GLU	129.10	-1	0.592
GLY	57.05	0	0.450
HIS	137.10	0.25	0.608
ILE	113.20	0	0.618
LEU	113.20	0	0.618
LYS	128.20	1	0.636
MET	131.20	0	0.618
PHE	147.20	0	0.636
PRO	97.12	0	0.556
SER	87.08	0	0.518
THR	101.10	0	0.562
TRP	186.20	0	0.678
TYR	163.20	0	0.646
VAL	99.07	0	0.586

Table S2: Description of proteins in the training set.

Protein	sequence length	ionic strength (mM)	R_g (nm)	Secondary Structure (SS)
1soy[S32]	106	157	1.530	α/β
1ubq[S33]	76	184	1.311	α/β
1wla[S33]	153	162	1.650	α
3mzq[S34]	124	162	1.610	α/β
5tvz[S35]	103	155	1.824	β
6eez[S36]	186	166	1.884	α/β
6h8m[S37]	107	119	1.535	α/β
ACTR[S38]	71	199	2.51	
An16[S39]	185	0	4.44	
α-synuclein[S40]	140	185	3.31	
ERM TADn[S41]	122	239	3.96	
hNHE1cdt[S38]	131	199	3.63	
IBB[S42]	97	162	3.20	
N49[S42]	36	162	1.59	
N98[S42]	151	162	2.86	
NLS[S42]	44	161	2.40	
NSP[S42]	176	162	4.10	
NUL[S42]	112	162	3.00	
NUS[S42]	80	162	2.49	
P53[S43]	93	208	2.87	
ProTα[S44,S45]	111	155	3.79	
SH4-UD[S46]	85	217	2.90	
Sic[S47]	92	162	3.21	

Table S3: Description of proteins in the validation set.

Protein	sequence length	ionic strength (mM)	R_g (nm)	Secondary Structure (SS)
1fu1[S48]	117	150	1.359	α/β
1ljp[S49]	98	150	1.253	α/β
2ea9[S50]	103	150	1.267	α/β
4cpv[S51]	108	150	1.258	α
Ash1[S52]	83	150	2.85	
K18[S53]	130	163	3.8	
K25[S53]	185	163	4.4	
rOPN[S54]	163	433	3.98	

Table S4: The radius of gyration of HP1 homologs. We find MOFF-IDP underestimates the differences between HP1α and HP1β and predicts both are overly collapsed, while MJ drastically underestimates the R_g of both HP1α and HP1β. Errors represent standard deviations from five independent replica exchange simulations.

Protein	SAXS R_g (nm)	MOFF R_g (nm)	MOFF-IDP R_g (nm)	MJ R_g (nm)
HP1α	3.59^{S55}	3.24 ± 0.08	3.07 ± 0.06	2.61 ± 0.01
HP1β	4.7^{S56}	4.26 ± 0.08	3.42 ± 0.08	2.60 ± 0.02
HP1γ	Not Available	3.75 ± 0.09	3.50 ± 0.05	2.57 ± 0.06

Table S5: T_C of HP1 homologs in various models. We note that the MJ potential fails to distinguish the phase separation of HP1 homologs. Meanwhile, MOFF-IDP gets the expected trend, but underestimates the stability of the condensed phase.

Protein	MOFF T_C (K)	MOFF-IDP T_C (K)	MJ T_C (K)
HP1α	306.7	267.1	402.1
HP1β	252.9	205.8	401.5
HP1γ	268.0	200.6	411.9

Table S6: In practice, we highlight the energies $\boldsymbol{\epsilon}'\boldsymbol{C}_{\text{PDB}}$ using folded structures obtained from steepest decent energy minimization of the PDB structures. This minimization is necessary to resolve steric clashes resulted from imperfect coarse-graining. As indicated by RMSD values from the PDB structures, the changes caused by minimization is small.

Protein	RMSD (nm)
1soy	0.148
1ubq	0.132
1wla	0.105
3mzq	0.143
5tvz	0.196
6eez	0.159
6h8m	0.127

Table S7: Table to help with implementation of the MOFF force field. These values of σ (nm) can be substituted directly into Eq. S16 for force field implementation.

AA	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	ASN	GLN	ASP	GLU	HIS	ARG	LYS	PRO
CYS	5.48e-01	5.83e-01	5.92e-01	5.83e-01	5.83e-01	5.67e-01	6.13e-01	5.97e-01	5.26e-01	4.99e-01	5.55e-01	5.33e-01	5.58e-01	5.75e-01	5.53e-01	5.70e-01	5.78e-01	6.02e-01	5.92e-01	5.52e-01
MET	–	6.18e-01	6.27e-01	6.18e-01	6.18e-01	6.02e-01	6.48e-01	6.32e-01	5.61e-01	5.34e-01	5.90e-01	5.68e-01	5.93e-01	6.10e-01	5.88e-01	6.05e-01	6.13e-01	6.37e-01	6.27e-01	5.87e-01
PHE	–	–	6.36e-01	6.27e-01	6.27e-01	6.11e-01	6.57e-01	6.41e-01	5.70e-01	5.43e-01	5.99e-01	5.77e-01	6.02e-01	6.19e-01	5.97e-01	6.14e-01	6.22e-01	6.46e-01	6.36e-01	5.96e-01
ILE	–	–	–	6.18e-01	6.18e-01	6.02e-01	6.48e-01	6.32e-01	5.61e-01	5.34e-01	5.90e-01	5.68e-01	5.93e-01	6.10e-01	5.88e-01	6.05e-01	6.13e-01	6.37e-01	6.27e-01	5.87e-01
LEU	–	–	–	–	6.18e-01	6.02e-01	6.48e-01	6.32e-01	5.61e-01	5.34e-01	5.90e-01	5.68e-01	5.93e-01	6.10e-01	5.88e-01	6.05e-01	6.13e-01	6.37e-01	6.27e-01	5.87e-01
VAL	–	–	–	–	–	5.86e-01	6.32e-01	6.16e-01	5.45e-01	5.18e-01	5.74e-01	5.52e-01	5.77e-01	5.94e-01	5.72e-01	5.89e-01	5.97e-01	6.21e-01	6.11e-01	5.71e-01
TRP	–	–	–	–	–	–	6.78e-01	6.62e-01	5.91e-01	5.64e-01	6.20e-01	5.98e-01	6.23e-01	6.40e-01	6.18e-01	6.35e-01	6.43e-01	6.67e-01	6.57e-01	6.17e-01
TYR	–	–	–	–	–	–	–	6.46e-01	5.75e-01	5.48e-01	6.04e-01	5.82e-01	6.07e-01	6.24e-01	6.02e-01	6.19e-01	6.27e-01	6.51e-01	6.41e-01	6.01e-01
ALA	–	–	–	–	–	–	–	–	5.04e-01	4.77e-01	5.33e-01	5.11e-01	5.36e-01	5.53e-01	5.31e-01	5.48e-01	5.56e-01	5.80e-01	5.70e-01	5.30e-01
GLY	–	–	–	–	–	–	–	–	–	4.50e-01	5.06e-01	4.84e-01	5.09e-01	5.26e-01	5.04e-01	5.21e-01	5.29e-01	5.53e-01	5.43e-01	5.03e-01
THR	–	–	–	–	–	–	–	–	–	–	5.62e-01	5.40e-01	5.65e-01	5.82e-01	5.60e-01	5.77e-01	5.85e-01	6.09e-01	5.99e-01	5.59e-01
SER	–	–	–	–	–	–	–	–	–	–	–	5.18e-01	5.43e-01	5.60e-01	5.38e-01	5.55e-01	5.63e-01	5.87e-01	5.77e-01	5.37e-01
ASN	–	–	–	–	–	–	–	–	–	–	–	–	5.68e-01	5.85e-01	5.63e-01	5.80e-01	5.88e-01	6.12e-01	6.02e-01	5.62e-01
GLN	–	–	–	–	–	–	–	–	–	–	–	–	–	6.02e-01	5.80e-01	5.97e-01	6.05e-01	6.29e-01	6.19e-01	5.79e-01
ASP	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5.58e-01	5.75e-01	5.83e-01	6.07e-01	5.97e-01	5.57e-01
GLU	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5.92e-01	6.00e-01	6.24e-01	6.14e-01	5.74e-01
HIS	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6.08e-01	6.32e-01	6.22e-01	5.82e-01
ARG	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6.56e-01	6.46e-01	6.06e-01
LYS	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6.36e-01	5.96e-01
PRO	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5.56e-01

Table S8: Table to help with implementation of the MOFF force field. These values of ϵ_{IJ} (kJ mol^{-1}) can be substituted directly into Eq. S16 for force field implementation.

AA	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	ASN	GLN	ASP	GLU	HIS	ARG	LYS	PRO
CYS	5.4061e+00	3.0981e+00	5.1767e-01	-6.2764e-01	3.8854e+00	2.6144e+00	5.1203e-01	6.1245e+00	1.3758e+00	3.2211e-01	5.0663e+00	9.6474e-01	4.3256e+00	-1.5098e+00	1.4601e+00	3.0788e+00	2.0406e+00	2.9546e-01	2.8364e+00	1.4556e+00
MET	–	-3.0497e-01	1.5299e+00	3.7089e+00	5.9558e+00	5.0346e+00	4.7447e+00	-1.6746e-02	1.2094e+00	-2.2211e+00	2.9555e+00	-3.5017e-01	-1.2175e+00	6.7339e+00	-1.5071e+00	4.6287e+00	-5.9512e-01	-5.5113e-01	-3.6569e-01	-1.9659e+00
PHE	–	–	2.6841e+00	2.2659e+00	5.7021e+00	2.3035e+00	7.3826e+00	6.5643e+00	1.8051e+00	-2.6861e-01	3.5402e+00	1.1469e+00	-1.9925e-01	-4.0477e-01	-5.1448e-01	-5.4375e-01	2.8275e+00	-4.3225e-01	-4.8430e-02	2.2408e+00
ILE	–	–	–	3.1438e+00	5.0296e+00	3.3988e+00	-7.7766e-01	6.5300e+00	2.3397e+00	6.0256e-01	3.0813e+00	2.2460e+00	-7.6336e-01	3.4430e+00	-9.4342e-02	2.8402e+00	1.9928e+00	4.7385e+00	1.6272e+00	1.1825e+00
LEU	–	–	–	–	-9.9983e-02	6.4860e-01	6.0576e+00	-6.5870e-01	-1.1096e-01	3.3784e-01	-1.6326e-01	-1.7247e-01	-2.9341e-01	1.3129e+00	-3.9802e-01	-2.0976e-01	-2.5145e-01	1.8681e+00	1.1726e+00	-1.8712e-01
VAL	–	–	–	–	–	2.3273e+00	3.5852e+00	4.8914e-01	1.3196e+00	3.8585e-01	-1.3375e-01	2.4068e-01	-3.8288e-01	1.4348e+00	-1.5824e-01	-2.7762e-01	2.9104e+00	3.2426e+00	6.5174e-01	-1.4601e+00
TRP	–	–	–	–	–	–	5.8890e+00	7.1431e+00	3.8875e+00	2.3074e+00	2.4862e+00	4.7613e+00	3.3797e+00	-9.2620e-01	-5.2296e-01	2.3466e+00	-3.4883e-01	2.8615e+00	8.4419e+00	1.8480e-01
TYR	–	–	–	–	–	–	–	5.9014e+00	3.4674e+00	-6.5254e-01	2.2058e+00	1.5755e+00	-9.1785e-01	4.2340e+00	-9.1084e-01	1.9625e+00	-2.8072e-01	6.5026e+00	-3.7370e-02	8.5891e-01
ALA	–	–	–	–	–	–	–	–	9.0442e-01	-4.3892e-01	-2.1007e-01	-1.3621e-01	1.4128e+00	2.5452e-02	-4.3625e-01	-2.4315e-01	-2.2681e-02	2.1512e-02	-2.0147e-01	-1.7790e-01
GLY	–	–	–	–	–	–	–	–	–	1.0587e+00	8.5347e-01	-8.3220e-01	-1.6795e-01	1.1124e-01	-1.0729e+00	4.0530e-02	-9.0568e-03	2.4502e+00	5.4930e-01	-1.2797e-01
THR	–	–	–	–	–	–	–	–	–	–	7.6363e-01	1.0242e+00	-1.5589e-01	1.0138e+00	2.5847e+00	1.6356e+00	8.5508e-01	-4.1093e-01	-8.4080e-01	-4.7690e-01
SER	–	–	–	–	–	–	–	–	–	–	–	3.3690e-01	-3.1076e-01	-6.2847e-01	-2.5737e-01	-1.9516e-01	4.3766e-01	-1.6973e+00	2.8594e-01	-5.2365e-01
ASN	–	–	–	–	–	–	–	–	–	–	–	–	-4.2854e-01	4.0018e-01	1.1053e+00	-1.6814e+00	-5.0747e-01	-2.6400e+00	1.0861e+00	-2.2035e+00
GLN	–	–	–	–	–	–	–	–	–	–	–	–	–	-7.1964e-01	-3.8101e-01	-1.8400e+00	-1.9729e-01	1.5827e+00	2.3936e+00	-1.4309e+00
ASP	–	–	–	–	–	–	–	–	–	–	–	–	–	–	9.6101e-01	1.7386e+00	3.9563e-01	-5.6567e-01	-9.9272e-01	-8.0222e-01
GLU	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-1.7625e+00	-2.4092e+00	-9.1943e-01	-9.5608e-01	-7.6499e-01
HIS	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-6.7999e-02	2.0054e+00	4.2815e+00	2.3341e+00
ARG	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5.7455e+00	-1.7508e+00	-1.0810e+00
LYS	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.6775e-02	-4.0148e-01
PRO	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-1.1971e+00

Training Sequences

1soy

MNDSEFHRLADQLWLTIEERLDDWDGDSDIDCEINGGVLTITFENGSKIIINRQEPLHQV
WLATKQGGYHFDLKGDEWICDRSGETFWDLLEQAATQQAGETVSFR

1ubq

MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN
IQKESTLHLVLRLRGG

1wla

GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED
LKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKHP
GDFGADAQGAMTKALELFRNDIAAKYKELGFQG

3mzq

KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQ
KNVACKNGQTNCYQSYSTMSITDCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHF
DASV

5tvz

RVKPSASLKLHHDLKLCLGDHSSVPVALKGQGPFTLTYDIIETFSSKRKTFEIKEIKTNE
YVIKTPVFTTGGDYILSLVSIKDSTGCVVGLSQPDAKIQVRRD

6eez

SNAARDNVTKSKISQYKDQIFDLTYPYSGNENSSVIAVGFLDYSCGHCKAIKNDIKQLIN
DGKIKYIFRDAPILGNASLKAAKSALAVYFLDKEKYFDFHHAALSHKGEFSDESILDIVK
NIGIDEDDFNDSIKDNADKIEQMINNSRLLVRDLGVGGTPFLIIGDSLFVGATDLNVLRK
KVDELS

6h8m

GFPIRLVDGENKKEGRVEVFVNGQWGTICDDGWTDKHAAVICRQLGYKGPARARTMAYFG
EGKGPIHMDNVKCTGNEKALADCVKQDIGRHNCRHSEDAGVICDYLE

ACTR

GTQNRPLLRNSLDDLVGPPSNLEGQSDERALLDQLHTLLSNTDATGLEEIDRALGIPELV
NQGQALEPKQD

An16

MHHHHHHPGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPS
SQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGA
PAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTP
SSQYV

α-synuclein

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTK
EQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDP
DNEAYEMPSEEGYQDYEPEA

ERM TADn

MDGFYDQQVPFMVPGKSRSEECRGRPVIDRKRKFLDTDLAHDSEELFQDLSQLQEAWLAE
AQVPDDEQFVPDFQSDNLVLHAPPPTKIKRELHSPSSELSSCSHEQALGANYGEKCLYNY
CA

hNHE1cdt

MVPAHKLDSPTMSRARIGSDPLAYEPKEDLPVITIDPASPQSPESVDLVNEELKGKVLGL
SRDPAKVAEEDEDDDGGIMMRSKETSSPGTDDVFTPAPSDSPSSQRIQRCLSDPGPHPEP
GEGEPFFPKGQ

IBB

GCTNENANTPAARLHRFKNKGKDSTEMRRRRIEVNVELRKAKKDDQMLKRRNVSSFPDDA
TSPLQENRNNQGTVNWSVDDIVKGINSSNVENQLQAT

N49

GCQTSRGLFGNNNTNNINNSSSGMNNASAGLFGSKP

N98

GCFNKSFGTPFGGGTGGFGTTSTFGQNTGFGTTSGGAFGTSAFGSSNNTGGLFGNSQTKP
GGLFGTSSFSQPATSTSTGFGFGTSTGTANTLFGTASTGTSLFSSQNNAFAQNKPTGFGN
FGTSTSSGGLFGTTNTTSNPFGSTSGSLFGP

NLS

ACETNKRKREQISTDNEAKMQIQEEKSPKKKRKKRSSKANKPPE

NSP

GCNFNTPQQNKTPFSFGTANNNSNTTNQNSSTGAGAFGTGQSTFGFNNSAPNNTNNANSS
ITPAFGSNNTGNTAFGNSNPTSNVFGSNNSTTNTFGSNSAGTSLFGSSSAQQTKSNGTAG
GNTFGSSSLFNNSTNSNTTKPAFGGLNFGGGNNTTPSSTGNANTSNNLFGATANAN

NUL

GCGFKGFDTSSSSSNSAASSSFKFGVSSSSSGPSQTLTSTGNFKFGDQGGFKIGVSSDSG
SINPMSEGFKFSKPIGDFKFGVSSESKPEEVKKDSKNDNFKFGLSSGLSNPV

NUS

GCPSASPAFGANQTPTFGQSQGASQPNPPGFGSISSSTALFPTGSQPAPPTFGTVSSSSQ
PPVFGQQPSQSAFGSGTTPN

P53

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPL

ProTα

MSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAENEENGEQEADNEVDEEEEEG
GEEEEEEEEEGDGEEEDGDEDEEAESATGKRAAEDDEDDDVDTKKQKTDEDD

SH4-UD

MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAE
PKLFGGFNSSDTVTSPQRAGPLAGG

Sic

GSMTPSTPPRSRGTRYLAQPSGNTSSSALMQGQKTPQKPSQNLVPVTPSTTKSFKNAPLLAP
PNSNMGMTSPFNGLTSPQRSPFPKSSVKRT

Validation Sequences

1fu1

MERKISRIHLVSEPSITHFLQVSWEKTLESGFVITLTDGHSAWTGTVSESEISQEADDMA
MEKGKYVGELRKALLSGAGPADVYTFNFSKESCYFFFEKNLKDVSFRLGSFNLEKVE

1ljp

TACTATQQTAAYKTLVSILSESSFSQCSKDSGYSMLTATALPTNAQYKLMCASTACNTMI
KKIVALNPPDCDLTVPTSGLVLDVYTYANGFSSKCASL

2ea9

MSNTTWGLQRDITPRLGARLVQEGNQLHYLADRASITGKFSDAECPKLDVVFPHFISQIE
SMLTTGELNPRHAQCVTLYHNGFTCEADTLGSCGYVYIAVYPT

4cpv

AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAIIDQDKSGFIEE
DELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTALVKA

Ash1

GASASSSPSPSTPTKSGKMRSRSSSPVRPKAYTPSPRSPNYHRFALDSPPQSPRRSSNSS
ITKKGSRRSSGSSPTRHTTRVCV

K18

MQTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIINKKLDLSNVQSKCGSKDNIKHVPG
GGSVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITH
VPGGGNKKIE

K25

MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKAEEAGIGDTPSLEDEA
AGHVTQARMVSKSKDGTGSDDKKAKGADGKTKIATPRGAAPPGQKGQANATRIPAKTPPA
PKTPPSSGEPPKSGDRSGYSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSS
AKSRL

rOPN

MPVKQADSGSSEQKQLYNKYPDAVATWLNPDPSQKQNLLAPQNAVSSSDDDDFKQETLPS
KSNESHDHMDDMDDEDDDDHVDSQDSIDSNDSDDVDDTDDSHQSDESHHSDESDELVTDF
PTDLPATEVFTPVVPTVDTYDGRGDSVVYGLRSKSKKHHHHHH

HP1 Sequences

HP1α

MGKKTKRTADSSSSEDEEEYVVEKVLDRRVVKGQVEYLLKWKGFSEEHNTWEPEKNLDCP
ELISEFMKKYKKMKEGENNKPREKSESNKRKSNFSNSADDIKSKKKREQSNDIARGFERG
LEPEKIIGATDSCGDLMFLMKWKDTDEADLVLAKEANVKCPQIVIAFYEERLTWHAYPED
AENKEKETAKS

HP1β

MGKKQNKKKVEEVLEEEEEEYVVEKVLDRRVVKGKVEYLLKWKGFSDEDNTWEPEENLDC
PDLIAEFLQSQKTAHETDKSEGGKRKADSDSEDKGEESKPKKKKEESEKPRGFARGLEPE
RIIGATDSSGELMFLMKWKNSDEADLVPAKEANVKCPQVVISFYEERLTWHSYPSEDDDK
KDDKN

HP1γ

MASNKTTLQKMGKKQNGKSKKVEEAEPEEFVVEKVLDRRVVNGKVEYFLKWKGFTDADNT
WEPEENLDCPELIEAFLNSQKAGKEKDGTKRKSLSDSESDDSKSKKKRDAADKPRGFARG
LDPERIIGATDSSGELMFLMKWKDSDEADLVLAKEANMKCPQIVIAFYEERLTWHSCPED
EAQ

References

(S1) Latham, A. P.; Zhang, B. Maximum Entropy Optimized Force Field for Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **2020**, *16*, 773–781.

(S2) Roux, B.; Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* **2013**, *138*.

(S3) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.

(S4) Amirkulova, D. B.; White, A. D. Recent advances in maximum entropy biasing techniques for molecular dynamics. *Mol. Simul.* **2019**, *45*, 1285–1294.

(S5) Bryngelson, J. D. D.; Wolynes, P. G. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524–7528.

(S6) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **1995**, *21*, 167–195.

(S7) Onuchic, J. N.; Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.

(S8) Dill, K. A.; Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Mol. Biol.* **1997**, *4*, 10–19.

(S9) Eastwood, M. P.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. Statistical mechanical refinement of protein structure prediction schemes: Cumulant expansion approach. *J. Chem. Phys.* **2002**, *117*, 4602–4615.

(S10) Mirny, L. A.; Shakhnovich, E. I. How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* **1996**, *264*, 1164–1179.

(S11) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization. *J. Comput. Chem.* **1997**, *18*, 874–887.

(S12) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **2012**, *116*, 8494–8503.

(S13) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: What determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298*, 937–953.

(S14) Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: A unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010**, *5*, 725–738.

(S15) Heinig, M.; Frishman, D. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **2004**, *32*, 500–502.

(S16) Mazur, J.; Jernigan, R. L. Distance-dependent dielectric constants and their application to double-helical DNA. *Biopolymers* **1991**, *31*, 1615–1629.

(S17) Mehler, E. L.; Solmajer, T. Electrostatic effects in proteins: Comparison of dielectric and charge models. *Protein Eng. Des. Sel.* **1991**, *4*, 903–910.

(S18) Mehler, E. L.; Eichele, G. Electrostatic Effects in Water-Accessible Regions of Proteins. *Biochemistry* **1984**, *23*, 3887–3891.

(S19) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A

portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **2009**, *180*, 1961–1972.

(S20) Pitera, J. W.; Chodera, J. D. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445–3451.

(S21) Miyazawa, S.; Jernigan, R. L. Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.* **1996**, *256*, 623–644.

(S22) Zheng, W.; Zerze, G. H.; Borgia, A.; Mittal, J.; Schuler, B.; Best, R. B. Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* **2018**, *148*.

(S23) Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16155–16160.

(S24) Noel, J. K.; Whitford, P. C.; Onuchic, J. N. The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function. *J. Phys. Chem. B* **2012**, *116*, 8692–8702.

(S25) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *39*, 1859–1865.

(S26) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2016**, *14*, 71–73.

(S27) Berendsen, H. J.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(S28) Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **2018**, *14*, 1–23.

(S29) Dignon, G. L.; Zheng, W.; Best, R. B.; Kim, Y. C.; Mittal, J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 9929–9934.

(S30) Tribello, G. A.; Giberti, F.; Sosso, G. C.; Salvalaglio, M.; Parrinello, M. Analyzing and Driving Cluster Formation in Atomistic Simulations. *J. Chem. Theory and Comput.* **2017**, *13*, 1317–1327.

(S31) Canzio, D.; Larson, A.; Narlikar, G. J. Mechanisms of functional promiscuity by HP1 proteins. *Trends Cell Biol.* **2014**, *24*, 377–386.

(S32) Prischi, F.; Konarev, P. V.; Iannuzzi, C.; Pastore, C.; Adinolfi, S.; Martin, S. R.; Svergun, D. I.; Pastore, A. Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly. *Nat. Commun.* **2010**, *1*, 1–10.

(S33) Valentini, E.; Kikhney, A. G.; Previtali, G.; Jeffries, C. M.; Svergun, D. I. SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.* **2015**, *43*, D357–D363.

(S34) Franke, D.; Jeffries, C. M.; Svergun, D. I. Machine Learning Methods for X-Ray Scattering Data Analysis from Biomacromolecular Solutions. *Biophys. J.* **2018**, *114*, 2485–2492.

(S35) Upla, P.; Kim, S. J.; Sampathkumar, P.; Dutta, K.; Cahill, S. M.; Chemmama, I. E.; Williams, R.; Bonanno, J. B.; Rice, W. J.; Stokes, D. L.; Cowburn, D.; Almo, S. C.; Sali, A.; Rout, M. P.; Fernandez-Martinez, J. Molecular Architecture of the Major Membrane Ring Component of the Nuclear Pore Complex. *Structure* **2017**, *3*, 434–445.

(S36) Walden, P. M.; Whitten, A. E.; Premkumar, L.; Halili, M. A.; Heras, B.; King, G. J.; Martin, J. L. The atypical thiol–disulfide exchange protein α-DsbA2 from Wolbachia pipientis is a homotrimeric disulfide isomerase. *Acta Crystallogr. D* **2019**, *75*, 283–295.

(S37) Canciani, A.; Catucci, G.; Forneris, F. Structural characterization of the third scavenger receptor cysteine-rich domain of murine neurotrypsin. *Protein Sci.* **2019**, *28*, 746–755.

(S38) Kjaergaard, M.; Nørholm, A. B.; Hendus-Altenburger, R.; Pedersen, S. F.; Poulsen, F. M.; Kragelund, B. B. Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α-helices or loss of polyproline II? *Protein Sci.* **2010**, *19*, 1555–1564.

(S39) Balu, R.; Dutta, N. K.; Choudhury, N. R.; Elvin, C. M.; Lyons, R. E.; Knott, R.; Hill, A. J. An16-resilin: An advanced multi-stimuli-responsive resilin-mimetic protein polymer. *Acta Biomater.* **2014**, *10*, 4768–4777.

(S40) Araki, K.; Yagi, N.; Nakatani, R.; Sekiguchi, H.; So, M.; Yagi, H.; Ohta, N.; Nagai, Y.; Goto, Y.; Mochizuki, H. A small-angle X-ray scattering study of alpha-synuclein from human red blood cells. *Sci. Rep.* **2016**, *6*, 1–8.

(S41) Lens, Z.; Dewitte, F.; Monté, D.; Baert, J. L.; Bompard, C.; Sénéchal, M.; Van Lint, C.; de Launoit, Y.; Villeret, V.; Verger, A. Solution structure of the N-terminal transactivation domain of ERM modified by SUMO-1. *Biochem. Biophys. Res. Commun.* **2010**, *399*, 104–110.

(S42) Fuertes, G.; Banterle, N.; Ruff, K. M.; Chowdhury, A.; Mercadante, D.; Koehler, C.; Kachala, M.; Girona, G. E.; Milles, S.; Mishra, A.; Onck, P. R.; Gräter, F.; Esteban-Martín, S.; Pappu, R. V.; Svergun, D. I.; Lemke, E. A. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E6342–E6351.

(S43) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5762–5767.

(S44) Uversky, V. N.; Gillespie, J. R.; Millett, I. S.; Khodyakova, A. V.; Vasilenko, R. N.; Vasiliev, A. M.; Rodionov, I. L.; Kozlovskaya, G. D.; Dolgikh, D. A.; Fink, A. L.; Doniach, S.; Permyakov, E. A.; Abramov, V. M. Zn 2+ -mediated structure formation and compaction of the "natively unfolded" human prothymosin α. *Biochem. Biophys. Res. Commun.* **2000**, *267*, 663–668.

(S45) Baul, U.; Chakraborty, D.; Mugnai, M. L.; Straub, J. E.; Thirumalai, D. Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins. *J. Phys. Chem. B* **2019**, *123*, 3462–3474.

(S46) Arbesú, M.; Maffei, M.; Cordeiro, T. N.; Teixeira, J. M.; Pérez, Y.; Bernadó, P.; Roche, S.; Pons, M. The Unique Domain Forms a Fuzzy Intramolecular Complex in Src Family Kinases. *Structure* **2017**, *25*, 630–640.e4.

(S47) Mittag, T.; Marsh, J.; Grishaev, A.; Orlicky, S.; Lin, H.; Sicheri, F.; Tyers, M.; Forman-Kay, J. D. Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase. *Structure* **2010**, *18*, 494–506.

(S48) Junop, M. S.; Modesti, M.; Guarne, A.; Ghirlando, R.; Gellert, M.; Yang, W. Crystal structure of the Xrcc4 DNA repair protein and implications for end joining. *EMBO J.* **2000**, *19*, 5962–5970.

(S49) Rodrigues, M. L.; Archer, M.; Martel, P.; Jacquet, A.; Cravador, A.; Carrondo, M. A. Structure of β-cinnamomin, a protein toxic to some plant species. *Acta Crystallogr. D* **2002**, *58*, 1314–1321.

(S50) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(S51) Kumar, V. D.; Edwards, B. F.; Lee, L. Refined Crystal Structure of Calcium-Liganded Carp Parvalbumin 4.25 at 1.5-Å Resolution. *Biochemistry* **1990**, *29*, 1404–1412.

(S52) Martin, E. W.; Holehouse, A. S.; Grace, C. R.; Hughes, A.; Pappu, R. V.; Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **2016**, *138*, 15323–15335.

(S53) Svergun, D. I.; Mylonas, E.; Mandelkow, E.; Hascher, A.; Bernadó, P.; Blackledge, M. Domain Conformation of Tau Protein Studied by Solution Small-Angle X-ray Scattering. *Biochemistry* **2008**, *47*, 10345–10353.

(S54) Lenton, S.; Grimaldo, M.; Roosen-Runge, F.; Schreiber, F.; Nylander, T.; Clegg, R.; Holt, C.; Härtlein, M.; García Sakai, V.; Seydel, T.; Marujo Teixeira, S. C. Effect of Phosphorylation on a Human-like Osteopontin Peptide. *Biophys. J.* **2017**, *112*, 1586–1596.

(S55) Larson, A. G.; Elnatan, D.; Keenen, M. M.; Trnka, M. J.; Johnston, J. B.; Burlingame, A. L.; Agard, D. A.; Redding, S.; Narlikar, G. J. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* **2017**, *547*, 236–240.

(S56) Munari, F.; Rezaei-Ghaleh, N.; Xiang, S.; Fischle, W.; Zweckstetter, M. Structural Plasticity in Human Heterochromatin Protein 1β. *PLoS ONE* **2013**, *8*.