Supporting information for:

Temperature controlled liquid-liquid phase separation of disordered proteins

Gregory L. Dignon,†§ Wenwei Zheng,†§ Young C. Kim,¶ and Jeetain Mittal†

†Department of Chemical and Biomolecular Engineering, Lehigh University, 111 Research Dr. Bethlehem, PA 18015, United States
‡College of Integrative Sciences and Arts, Arizona State University, Mesa, Arizona 85212, United States
¶Center for Materials Physics and Technology, Naval Research Laboratory, Washington, DC 20375, United States
§Authors contributed equally to this work
1. Supporting Methods

1.1 List of Sequences used for training HPS-T model

CspTm

MRGKVKFDFS KKGYGFITKD EGGDVFVHWS AIEMEGKTL KEGQTVFEI QEGKKGQAA
HVKVVE

Integrase

CFLDGIDKAQ EEHEKYHSNW RAMASDFNLP PVVAKGIVAS CDKCQLKGEA MHQVDC

λ-repressor

GPCLTQEQL E DARRLKAIYE KKKKELGLSQ ESVADKMGMG QSGVGMFLNG INALNAYNA
LLAKILKSVV EEFPSIARE CR

ProTαC

CEEGGEEEE EEEGDEEED GDDEEAEAESA TGKRAEDDE DDDVDTKKQK TDEDC

ProTαN

CDAAVDTI SE ITKDLKEKK EVVEEAENGR DAPANGNAEN EENGEQEADN EVCEEC

1.2 Fitting all-atom data to experimental data

The FRET experimental data\(^{S1}\) is first analyzed using the new SAW-\(\nu\) model.\(^{S2}\) Since the experimental data only covers a small range of temperatures (285K-350K), we opted to use all-atom data which gave many more data points in a larger range of temperatures (275K-500K). The simulations were conducted using the Amber ff03w,\(^{S3}\) which has been shown to give overly collapsed configurations for proteins.\(^{S4}\) Toward this end, we applied a scaling and shift to the data set:

\[
R'_g(T) = R_g(T) \ast n_1 + n_2
\]  
\[(S1)\]
where \( n_1 \) and \( n_2 \) are fitting parameters used to minimize the difference from the experimental data from Wuttke et al.\textsuperscript{S1} on the same protein sequences (Fig. S1).

### 1.3 Fitted functional forms to data from Van Dijk et al.\textsuperscript{S5}

\[
E_H(T) = -22.657 + 0.15379T - 0.00025597T^2 \quad (S2a)
\]
\[
E_A(T) = -23.364 + 0.15876T - 0.00026696T^2 \quad (S2b)
\]
\[
E_O(T) = 2.1607 - 0.015064T + 0.000026T^2 \quad (S2c)
\]
\[
E_P(T) = 10.475 + 0.071482T + 0.0001201T^2 \quad (S2d)
\]
\[
E_C(T) = 8.5997 - 0.057676T + 0.00009317T^2 \quad (S2e)
\]

### 1.4 Implementation of Dill-Alonso-Hutchinson-based model

To apply the thermodynamics-based model from Dill, Alonso and Hutchinson,\textsuperscript{S6} we use the following functional form from their work:

\[
\chi(T) = \frac{1.4}{kT}(\Delta H^\circ + \Delta C_p(T - T_0) - T[\Delta S^\circ + \Delta C_p\ln(T/T_0)]) \quad (S3)
\]

where the factor of 1.4 accounts for the number of amino acids per lattice segment from the original derivation of this equation,\textsuperscript{S7} \( T_0 = 298K \), and \( \Delta H^\circ, \Delta S^\circ \) and \( \Delta C_p^\circ \) are obtained from experimental measurements of hydrophobic amino acid side chain analogues\textsuperscript{S8} and are equal to 0.0 cal, -6.7 cal/(mol·K), and 55 cal/(mol·K) respectively. We adopt the same set of solvophobic amino acids from that work A, F, I, L, M, V, W, Y as hydrophobic. We then apply this functional form to the hydrophobicity model as:

\[
\lambda(T) = \lambda_{HPS} + \frac{kT}{\epsilon}(\chi(T) - \chi(T_{\text{ref}})) \quad (S4)
\]

where \( T_{\text{ref}} \) is set to 300K where the original HPS model was parameterized, and \( \epsilon \) is 0.2
kcal/mol as in the original HPS model. Multiplying by kT results in the $\chi kT$ functional form presented in that work.\textsuperscript{S6} For amino acids not considered hydrophobic, we do not impose any temperature dependence on $\lambda$.

### 1.5 Simplified functional forms of the optimized 15-parameter model.

$$\lambda_{i,H}(T) = \lambda_{i,HPS} - 34.690 + 0.20242T - 0.00025463T^2,$$

(S5a)

$$\lambda_{i,A}(T) = \lambda_{i,HPS} - 63.201 + 0.36955T - 0.00053392T^2,$$

(S5b)

$$\lambda_{i,O}(T) = \lambda_{i,HPS} + 6.8820 - 0.040528T + 0.000052T^2,$$

(S5c)

$$\lambda_{i,P}(T) = \lambda_{i,HPS} + 32.994 - 0.19100T + 0.0002402T^2,$$

(S5d)

$$\lambda_{i,C}(T) = \lambda_{i,HPS} + 21.768 - 0.13373T + 0.00018663T^2,$$

(S5e)

### 1.6 Method for sampling sequences.

To sample sequences with amino acid compositions, we start from the CspTm sequence with a chain length of 66. We then pick 10 positions randomly along the sequence and mutate each to other residues, with probabilities adjusted according to the relative abundances of different amino acids in IDP sequences.\textsuperscript{S9} We repeat this procedure 1 million times, allowing mutations to accumulate, thus generating 1 million sequences, corresponding to the average amino acid composition of IDPs. The errors of the probabilities of the phase-diagram shape and the abundance of amino acids in one phase-diagram shape are estimated using a block average with five blocks. Results are listed in Table S3.

We further test whether differing overall amino acid abundance has an appreciable effect on the abundance of amino acids in each phase-diagram shape. We have generated additional sequences (1 million each) using uniformly weighted amino acid probability
(Table S4) and probability based on amino acid composition of structures from the protein data bank (Table S5). The probabilities of sampling different phase-diagram shapes however differ considerably from those tested using amino acid abundances from IDPs. For both cases, the closed-loop shape is with a much higher probability whereas single-UCST is almost completely absent. The abundance of amino acids with different phase-diagram shape does qualitatively agree among all the three cases using different amino acid composition for generating the sequences.
2. Supporting Figures

Figure S1: Reference data set containing temperature-dependent $R_g$ of five sequences. The experimental data$^{s1}$ analyzed with the newly proposed method$^{s2}$ is shown in black; the all-atom simulation$^{s10}$ in red; the fitted data set by shifting the simulation data to minimize its difference from experiment in green; and the fitted data set by shifting and rescaling the simulation data to minimize its difference from experiment in blue. For each model we parameterize to reference data in this work, we use the blue curve except for Fig. S6 and S7, in which we use the green curve and the black curve respectively.
Figure S2: A) Model resulting from directly using functional forms from van Dijk et al \textsuperscript{511} rescaled to units of 0.2 kcal/mol so that well depth follows the same temperature dependence. Wellness of fit to reference data for proteins B) CspTm, C) Integrase, D) λ-repressor, E) ProTαC, and F) ProTαN.
Figure S3: Model optimized to training data using independent free parameters for each of the five amino acid types, for a total of 15 free parameters as listed in Table 2.

Figure S4: Application of Dill-Alonso-Hutchinson temperature-dependent model\textsuperscript{56} into the hydrophobic scaling.
Figure S5: Model parameterized to fit the reference data set at temperatures below 400K. Fitted parameters are: $\alpha = 0.898$, $T_{\text{ref}} = 289.9$, $T_{\text{shift}} = 60.32$.

Figure S6: Model parameterized to fit reference data calculated by shifting all-atom data up to account for collapsed configurations from older force field, and to preserve the change in compaction as a function of temperature. Fitted parameters are: $\alpha = 0.3929$, $T_{\text{ref}} = 294.3$, $T_{\text{shift}} = 92.5$. 
Figure S7: Model parameterized to fit the $R_g$ data directly from experiment. Fitted parameters are: $\alpha = 2.450$, $T_{\text{ref}} = 285.9$, $T_{\text{shift}} = 29.6$.

Figure S8: Simulation results of QC sequences using Dill-Alonso-Hutchinson model\textsuperscript{S6} for sequences showing A) LCST, B) UCST, and C) no phase separation from experimental results.\textsuperscript{S12}
Figure S9: A). Phase diagram of QC37 and slab simulation configurations at different temperatures. B). At very high temperatures, QC37 forms densely-packed assemblies. C) Collapsed QC37 globules forming a less densely-packed assembly. D). Collapsed globules weakly associate but remain overall dispersed. E). Above $T_{\theta}$ QC37 is still dispersed, even though chains are mostly collapsed. F). Supercritical fluid of extended protein chains. G). Slab configuration with condensed liquid-like phase and coexisting vapor phase.

Figure S10: Average hydrophobicity of amino acids as a function of distance from the COM of the QC37 chain, calculated from single chain simulation.
Figure S11: Average (solid line) and standard deviation (dashed line) of $\lambda$ for all residues of QC21 and QC37 sequences over the range of temperatures tested. Blue dashed lines indicate single-chain $T_\theta$ values.

Figure S12: Mean squared displacement of QC37 within the condensed phase below UCST (left) and above LCST (right). Diffusion coefficients are reported in parentheses in units of $cm^2/s$. 
Figure S13: The probability of H/A/O/P/C type amino acids in a typical IDP sequence from a bioinformatics study.\cite{S9}

Figure S14: Predicting $R_g$ and $\nu$ from the temperature, mean hydrophobicity $<\lambda>$ and the chain length.
Figure S15: Contour lines of $R_g$ and $\nu$ in the predictor as a function of temperature and mean hydropathy $<\lambda>$. 
Figure S16: $R_g$ of 2000 random sequences generated using the amino acid propensity of IDPs.\textsuperscript{9} A) The comparison between the simulated $R_g$ using LAMMPS and the $R_g$ from the predictor (see Methods). B) The relative difference as a function of the net charge.
3. Supporting Tables

Table S1: Hydrophobicity and size parameters from original HPS model.

| Amino Acid | $\lambda_{HPS}$ | $\sigma$ (Å) | Amino Acid | $\lambda_{HPS}$ | $\sigma$ (Å) |
|------------|-----------------|-------------|------------|-----------------|-------------|
| A          | 0.730           | 5.04        | M          | 0.838           | 6.18        |
| C          | 0.595           | 5.48        | N          | 0.432           | 5.68        |
| D          | 0.378           | 5.58        | P          | 1.000           | 5.56        |
| E          | 0.459           | 5.92        | Q          | 0.514           | 6.02        |
| F          | 1.000           | 6.36        | R          | 0.000           | 6.56        |
| G          | 0.649           | 4.50        | S          | 0.595           | 5.18        |
| H          | 0.514           | 6.08        | T          | 0.676           | 5.62        |
| I          | 0.973           | 6.18        | V          | 0.892           | 5.86        |
| K          | 0.514           | 6.36        | W          | 0.946           | 6.78        |
| L          | 0.973           | 6.18        | Y          | 0.865           | 6.46        |

Table S2: Sequences from Garcia-Quiroz et al.\textsuperscript{512} and labels used for this work, (Quiroz-Chilkoti/QC sequences). Groups 1, 2 and 3 correspond to sequences which undergo LCST, UCST and no phase separation respectively.

| Name  | Length | Sequence     | Group | Name  | Length | Sequence     | Group |
|-------|--------|--------------|-------|-------|--------|--------------|-------|
| QC1   | 150    | [AVPGVG]\textsuperscript{25} | 1     | QC2   | 390    | [TVPGVG]\textsuperscript{65} | 1     |
| QC3   | 330    | [TVPGAG]\textsuperscript{55} | 3     | QC4   | 180    | [GVPGAV]\textsuperscript{30} | 1     |
| QC5   | 300    | [GVPGVA]\textsuperscript{50} | 1     | QC6   | 120    | [VAPGVG]\textsuperscript{20} | 3     |
| QC7   | 225    | [APGVG]\textsuperscript{45} | 3     | QC8   | 175    | [VPGVA]\textsuperscript{35} | 1     |
| QC9   | 150    | [VPGVG]\textsuperscript{30} | 1     | QC10  | 125    | [VHPGVG]\textsuperscript{25} | 1     |
| QC11  | 175    | [VAPVG]\textsuperscript{35} | 1     | QC12  | 150    | [VPGVV]\textsuperscript{30} | 1     |
| QC13  | 150    | [VPAGVG]\textsuperscript{25} | 1     | QC14  | 240    | [VPTGVG]\textsuperscript{40} | 1     |
| QC15  | 210    | [APVGVG]\textsuperscript{35} | 1     | QC16  | 125    | [VRPVG]\textsuperscript{25} | 3     |
| QC17  | 240    | [APVGLG]\textsuperscript{40} | 1     | QC18  | 225    | [VPAVG]\textsuperscript{45} | 1     |
| QC19  | 200    | [VPHVG]\textsuperscript{40} | 1     | QC20  | 120    | [VGPVAV]\textsuperscript{20} | 1     |
| QC21  | 150    | [VPNAVG]\textsuperscript{25} | 1     | QC22  | 180    | [TPVAVG]\textsuperscript{30} | 1     |
| QC23  | 189    | [VPSALYG]\textsuperscript{21} | 1     | QC24  | 320    | [GRGNSPYG]\textsuperscript{40} | 2     |
| QC25  | 448    | [RGDSPHG]\textsuperscript{64} | 2     | QC26  | 160    | [GRGDSPYU]\textsuperscript{20} | 2     |
| QC27  | 160    | [GRDGSPYU]\textsuperscript{20} | 2     | QC28  | 168    | [RGDSSPYG]\textsuperscript{24} | 2     |
| QC29  | 160    | [GRGDSPF]\textsuperscript{20} | 2     | QC30  | 160    | [GRGESYU]\textsuperscript{20} | 2     |
| QC31  | 192    | [RGDAPYU]\textsuperscript{24} | 2     | QC32  | 224    | [RGDAPYU]\textsuperscript{28} | 2     |
| QC33  | 192    | [QYPSDGGR]\textsuperscript{24} | 2     | QC34  | 140    | [RGDSYPYU]\textsuperscript{20} | 2     |
| QC35  | 320    | [VHPHSRNG]\textsuperscript{40} | 3     | QC36  | 108    | [VPSTDYG]\textsuperscript{12} | 2     |
| QC37  | 160    | [GRPSDSYU]\textsuperscript{20} | 1     | QC38  | 261    | [VPSDDYG]\textsuperscript{29} | 3     |
| QC39  | 180    | [VPSDDYG]\textsuperscript{20} | 3     |
Table S3: Possible states in sequence space with different phase behaviors using sequences generated with amino acid compositions based on their relative abundances in IDPs. Numbers in brackets show the errors of the last digit.

| Phase-diagram shape | P(shape) (%) | H     | A     | O     | P     | C     |
|---------------------|--------------|-------|-------|-------|-------|-------|
| IDP                 |              | 0.289 | 0.068 | 0.160 | 0.241 | 0.242 |
| none                | 31.42(4)     | 0.2631(1) | 0.0627(1) | 0.1533(1) | 0.2501(1) | 0.2707(1) |
| single-UCST         | 8.48(5)      | 0.20887(7) | 0.04863(7) | 0.1712(1) | 0.2892(2) | 0.2821(2) |
| closed-loop         | 59.63(9)     | 0.3154(2) | 0.0736(1) | 0.1618(2) | 0.2282(1) | 0.2208(2) |
| hourglass           | 0.468(8)     | 0.1607(2) | 0.0355(4) | 0.1774(9) | 0.3293(9) | 0.2971(6) |

Table S4: Possible states in sequence space with different phase behaviors using sequences generated with unbiased amino acid compositions. Numbers in brackets show the errors of the last digit.

| Phase-diagram shape | P(shape) (%) | H     | A     | O     | P     | C     |
|---------------------|--------------|-------|-------|-------|-------|-------|
| IDP                 |              | 0.1903 | 0.1567 | 0.1541 | 0.2395 | 0.2594 |
| none                | 7.145(5)     | 0.1903(3) | 0.1567(3) | 0.1541(6) | 0.2395(4) | 0.2594(2) |
| single-UCST         | 0.166(8)     | 0.146(1) | 0.114(2) | 0.1762(8) | 0.292(2) | 0.272(2) |
| closed-loop         | 92.69(6)     | 0.2547(1) | 0.2034(1) | 0.1497(2) | 0.1969(2) | 0.1952(2) |
| hourglass           | 0.0022(5)    | 0.124(4) | 0.072(4) | 0.17(1) | 0.34(1) | 0.29(1) |

Table S5: Possible states in sequence space with different phase behaviors using sequences generated with amino acid compositions based on their relative abundances in the protein data bank. Numbers in brackets show the errors of the last digit.

| Phase-diagram shape | P(shape) (%) | H     | A     | O     | P     | C     |
|---------------------|--------------|-------|-------|-------|-------|-------|
| IDP                 |              | 0.318 | 0.114 | 0.139 | 0.199 | 0.230 |
| none                | 10.25(7)     | 0.2501(1) | 0.0912(1) | 0.1405(2) | 0.2302(3) | 0.2879(2) |
| single-UCST         | 0.49(1)      | 0.1955(6) | 0.0677(6) | 0.1595(5) | 0.274(1) | 0.303(1) |
| closed-loop         | 89.26(8)     | 0.3265(1) | 0.1167(9) | 0.1388(8) | 0.19500(8) | 0.2229(1) |
| hourglass           | 0.0055(9)    | 0.151(5) | 0.051(3) | 0.158(6) | 0.315(9) | 0.325(9) |
References

(S1) Wuttke, R.; Hofmann, H.; Nettels, D.; Borgia, M. B.; Mittal, J.; Best, R. B.; Schuler, B. Temperature-dependent solvation modulates the dimensions of disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* 2014, 111, 5213–5218.

(S2) 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, 123329.

(S3) Best, R. B.; Mittal, J. Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse. *J. Phys. Chem. B* 2010, 114, 14916–14923.

(S4) Best, R. B.; Zheng, W.; Mittal, J. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theor. Comput.* 2014, 10, 5113–5124.

(S5) van Dijk, E.; Varilly, P.; Knowles, T. P.; Frenkel, D.; Abeln, S. Consistent treatment of hydrophobicity in protein lattice models accounts for cold denaturation. *Phys. Rev. Lett.* 2016, 116, 078101.

(S6) Dill, K. A.; Alonso, D. O.; Hutchinson, K. Thermal stabilities of globular proteins. *Biochemistry* 1989, 28, 5439–5449.

(S7) Dill, K. A. Theory for the folding and stability of globular proteins. *Biochemistry* 1997, 24, 1501–1509.

(S8) Nozaki, Y.; Tanford, C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions establishment of a hydrophobicity scale. *Journal of Biological Chemistry* 1971, 246, 2211–2217.
(S9) Coeytaux, K.; Poupon, A. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics* **2005**, *21*, 1891–1900.

(S10) Zerze, G. H.; Best, R. B.; Mittal, J. Sequence-and temperature-dependent properties of unfolded and disordered proteins from atomistic simulations. *J. Phys. Chem. B* **2015**, *119*, 14622–14630.

(S11) van Dijk, E.; Hoogeveen, A.; Abeln, S. The hydrophobic temperature dependence of amino acids directly calculated from protein structures. *PLoS Comput. Biol.* **2015**, *11*, e1004277.

(S12) Quiroz, F. G.; Chilkoti, A. Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers. *Nat. Mater.* **2015**, *14*, 1164.