Supporting information for: Maximum Entropy Optimized Force Field for Intrinsically Disordered Proteins

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

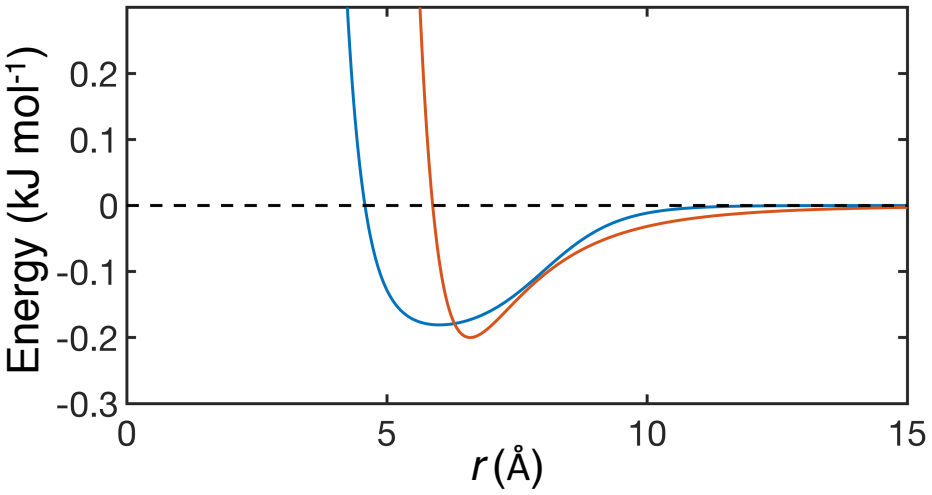

Figure S1: Comparison between the nonbonded poteintial ($V_{\mathrm{nb}}(r_{ij})$) defined in Eq. 9 of the main text (blue) and the Lennard-Jones potential of equal ϵ_{IJ} (orange).

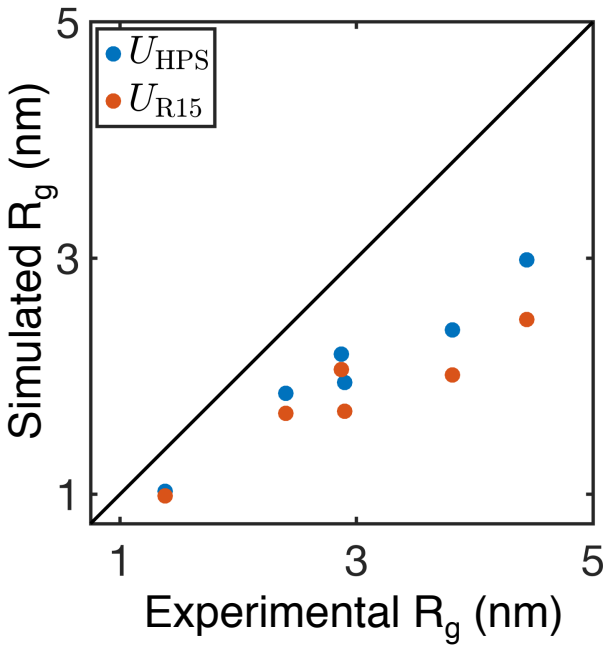

Figure S2: Application of the force field $U_{\mathrm{R15}}(\mathbf{r})$ optimized for a single protein to the set of test proteins listed in Table 2. The corresponding R_g predicted by the hydrophobic scale model ($U_{\mathrm{HPS}}(\mathbf{r})$) is shown in blue for comparison.

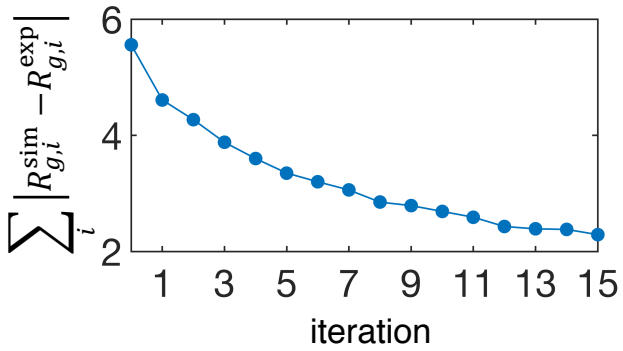

Figure S3: Simulation error ($\sum_i |R_{g,i}^{\mathrm{sim}} - R_{g,i}^{\mathrm{exp}}|$) for test proteins decreases monotonically as a function of the number of iterations carried out for MOFF optimization.

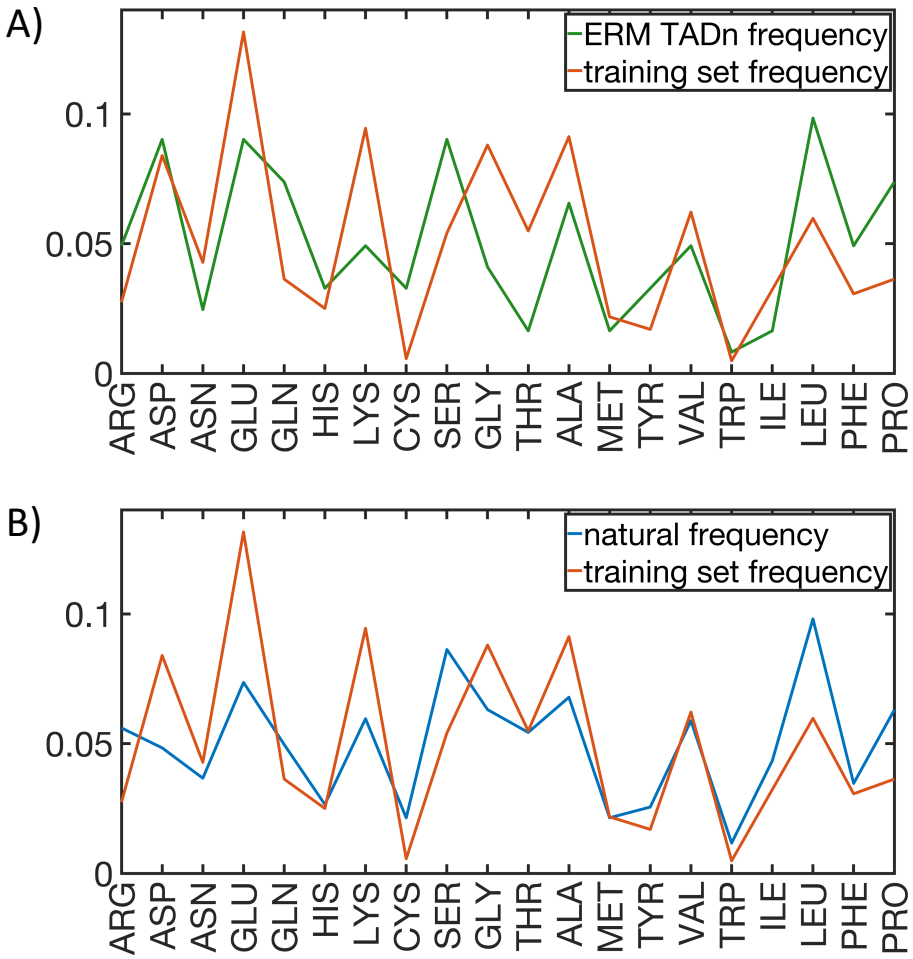

Figure S4: Analysis of amino acid frequency. A) Frequency of amino acids in our training set (orange) compared to that from the protein ERM TADn (green).[S1] B) Frequency of amino acids in our training set (orange) compared to the naturally occurring frequency of codons for those amino acids (blue).[S2] Amino acids on the x-axis are sorted from least to most hydrophobic.[S3]

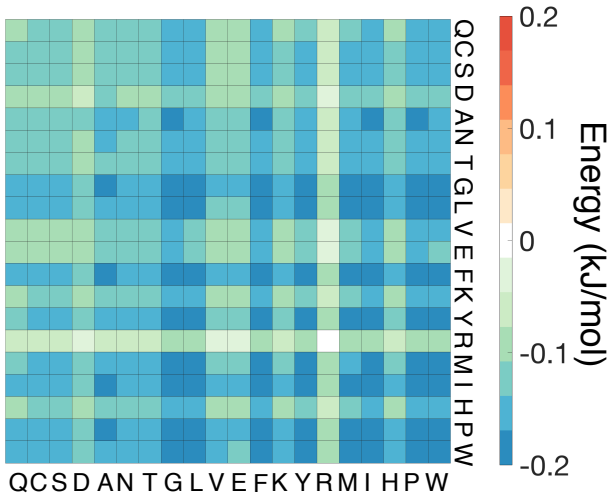

Figure S5: Contact energy between amino acids for the hydrophobic scale model.[S3] Amino acids are ordered according to the MOFF clusters defined in Figure 5 of the main text. The energy scale is reduced from that in Figure 5 due to the smaller magnitude of energies in the hydrophobic scale model relative to MOFF, and ranges from red (most repulsive) to blue (most attractive).

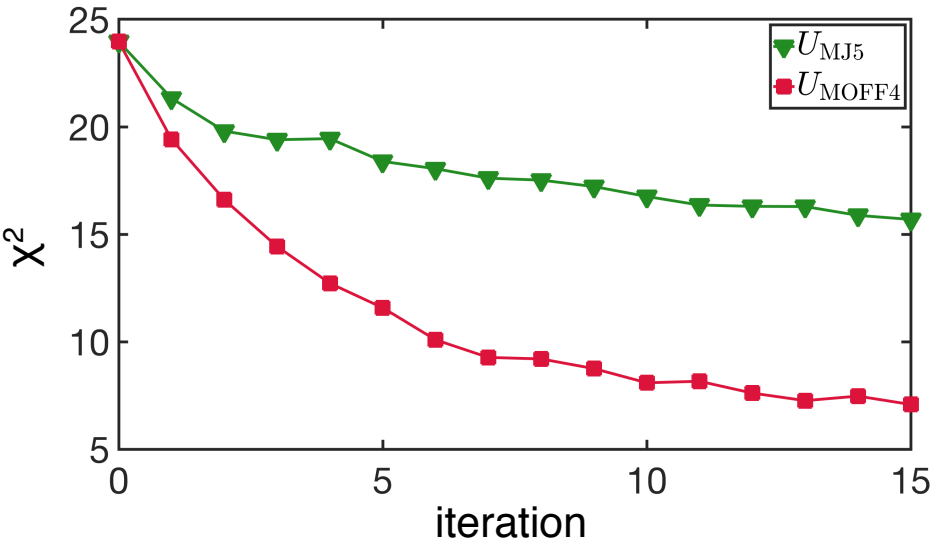

Figure S6: The normalized difference between simulated and experimental R_g values (χ^2 defined in Eq. 11 of main text) as a function of the number of iterations for the optimization of U_{MOFF4} and U_{MJ5}.

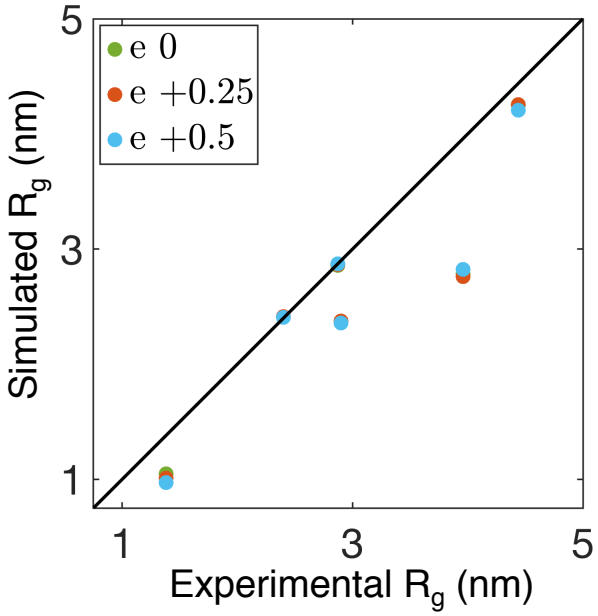

Figure S7: Varying the charge of the histidine residue incurs minimal changes in the simulated R_g values of test proteins. In the results presented in the main text, we used a charge of 0.25 (red). We carried out additional simulations in which the charge was changed to 0 (green) and 0.5 (blue).

Table S1: Amino acid masses and charges used in simulation.

Amino Acid	Mass (amu)	Charge
ALA	71.08	0
ARG	156.20	1
ASN	114.10	0
ASP	115.10	-1
CYS	103.10	0
GLN	128.10	0
GLU	129.10	-1
GLY	57.05	0
HIS*	137.10	0.25
ILE	113.20	0
LEU	113.20	0
LYS	128.20	1
MET	131.20	0
PHE	147.20	0
PRO	97.12	0
SER	87.08	0
THR	101.10	0
TRP	186.20	0
TYR	163.20	0
VAL	99.07	0

*The charge of the histidine residue can vary from 0 to 0.5 at different pH values (7.5 to 6.0). Since experiments were mostly performed at pH 7, we set the charge as +0.25. Varying this exact number appears to have minimal effect the simulated R_g values for proteins studied here (see Figure S7).

Training Sequences

CspTm

GPGMRGKVKWFDSKKGYGFITKDEGGDVFVHWSAIEMEGFKTLKEGQVVEFEIQEGKKGG
QAAHVKV

IN

GSHCFLDGIDKAQEEHEKYHSNWRAMASDFNLPPVVAKEIVASCDKCQLKGEAMHGQVDC

ProTα-N

GPSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAENEENGEQEADNEVDEECEE
GGEEEEEEEEEGDGEEEDGDEDEEAESATGKRAAEDDEDDDVDTKKQKTDEDD

ProTα-C

MAHHHHHHSAALEVLFQGPMSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNANE
ENGEQEADNEVDEECEEGGEEEEEEEEEGDGEEEDGDEDEEAESATGKRAAEDDEDDDVDT
KKQKTDEDD

R15

KLKEANKQQNFNTGIKDFDFWLSEVEALLASEDYGKDLASVNNLLKKHQLLEADISAHED
RLKDLNSQADSLMTSSAFDTSQVKDKRETINGRFQRIKSMAAARRAKLNESHRL

R17

RLEESLEYQQFVANVEEEEAWINEKMTLVASEDYGDTLAAIQGLLKKHEAFETDFTVHKD
RVNDVAANGEDLIKKNNHHVENITAKMKGLKGKVSDLEKA

hCyp

SSFHRIIPGFMSQGGDFTRHNGTGGKSIYGEKFEDENFILKHTGPGILSMANAGPNTNGS
QFFISTAKTEFLDGKHVVFGKVKEGMNIVEAMERFGSRNGKTSKKITIADSGQLE

Protein-L

MEEVTIKANLIFANGSTQTAEFKGTFEKATSEAYAYADTLKKDNGEWTVDVADKGYTLNI
KFAG

ACTR

GTQNRPLLRNSLDDLVGPPSNLEGQSDERALLDQLHTLLSNTDATGLEEIDRALGIPELV
NQGQALEPKQD

hNHE1cdt

MVPAHKLDSPTMSRARIGSDPLAYEPKEDLPVITIDPASPQSPESVDLVNEELKGKVLGL
SRDPAKVAEEDEDDDGGIMMRSKETSSPGTDDVFTPAPSDSPSSQRIQRCLSDPGPHPEP
GEGEPFFPKGQ

sNase

ATSTKKLHKEPATLIKAIDGDTVKLMYKGQPMTFRLLLVDTPETKHPKKGVEKYGPEASA
FTKKMVENAKKIEVEFDKGQRTDKYGRGLAYIYADGKMVNEALVRQGLAKVAYVYKPNNT
HEQHLRKSEAQAKKEK

α-synuclein

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTK
EQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDP
DNEAYEMPSEEGYQDYEPEA

Test Sequences

An16

MHHHHHHPGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPS
SQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGA
PAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTP
SSQYV

ERM TADn

MDGFYDQQVPFMVPGKSRSEECRGRPVIDRKRKFLDTDLAHDSEELFQDLSQLQEAWLAE
AQVPDDEQFVPDFQSDNLVLHAPPPTKIKRELHSPSSELSSCSHEQALGANYGEKCLYNY
CA

Histatin-5

DSHAKRHHGYKRKFHEKHHSHRGY

Nucleoporin 153

GCPSASPAFGANQTPTFGQSQGASQPNPPGFGSISSSTALFPTGSQPAPPTFGTVSSSSQ
PPVFGQQPSQSAFGSGTTPNA

p53

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPL

SH4-UD

MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAE

PKLFGGFNSSDTVTSPQRAGPLAGG

References

(S1) Lens, Z.; Dewitte, F.; Monté, D.; Baert, J. L.; Bompard, C.; Sénéchal, M.; Van Lint, C.; de Launoit, Y.; Villeret, V.; Verger, A. *Biochem. Biophys. Res. Commun.* **2010**, *399*, 104–110.

(S2) Athey, J.; Alexaki, A.; Osipova, E.; Rostovtsev, A.; Santana-Quintero, L. V.; Katneni, U.; Simonyan, V.; Kimchi-Sarfaty, C. *BMC Bioinform.* **2017**, *18*, 1–10.

(S3) Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. *PLOS Comput. Biol.* **2018**, *14*, 1–23.