Supporting Information for

Experimentally derived and computationally optimized backbone conformational Statistics for blocked amino acids

Jeong-Mo Choi and Rohit V. Pappu*
Department of Biomedical Engineering and Center for Biological Systems Engineering
Washington University in St. Louis
One Brookings Drive, Campus Box 1097
St. Louis, Missouri 63130

*Email: pappu@wustl.edu

Table S1. (separate Excel file) Basin statistics for coil-library landscapes of 20 amino acids. Bolded numbers indicate the basins with weights > 15%. (Column A) amino acid type; (column B) basin ID; (column C) basin weight; (columns D-E) position of basin centers on the (ϕ, ψ) space; (column F) basin depth in terms of PMF; (column G) basin area in units of $(2.5°)^2$; (column H) manual assignment of the three major basins.

Table S2. (separate Excel file) Experimental and calculated *J* coupling constants for various combinations of candidates for the three major basins. (Column A) amino acid type; (column B) candidate number for each amino acid; (columns C-H) positions of basin centers on the (ϕ, ψ) space; (column I) experimental *J* coupling constants; (column J) calculated *J* coupling constants; (column K) absolute errors between the experimental and calculated *J* coupling constants; (column L) combinations that provide the smallest absolute errors.

AminoAcidData.zip. A zipped archive comprising of individual comma-separated text files with Optimized populations for all 20 amino acids on the 2.5° × 2.5° grid. Columns and rows correspond to ϕ and ψ bins, respectively. The total size of each data set is normalized to 5×10^5. There may be some fractional deviations due to round-off.

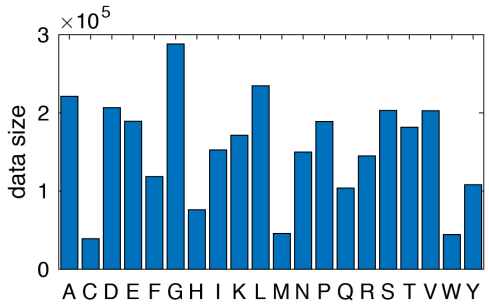

Figure S1. Sizes of the data sets used in this work, according to their residue types.

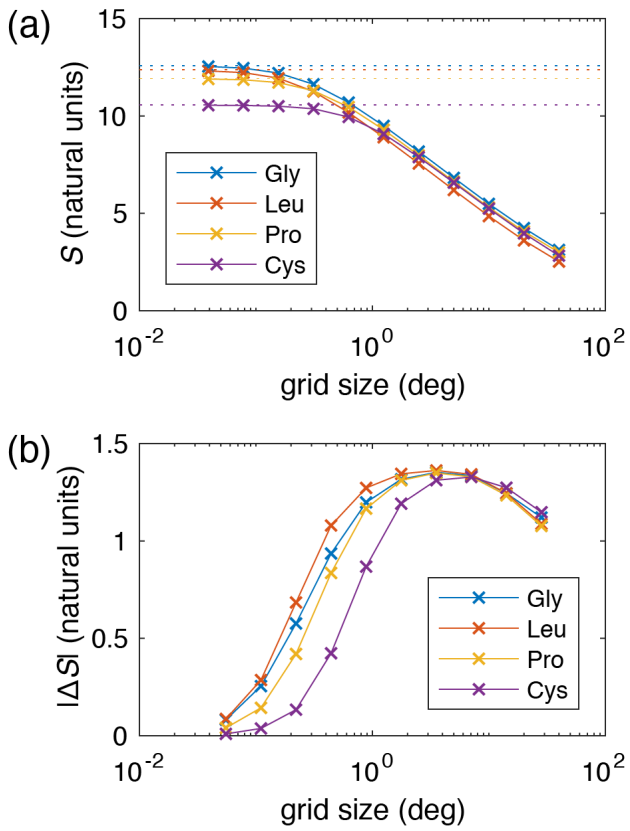

Figure S2. Analysis of the information content as a function of grid size. (a) Shannon entropy *S* as a function of grid size for four prototypical amino acids, namely, Gly, Leu, Pro, and Cys. The units of *S* are natural units of information. The dotted lines show the saturation values of *S* for each of the datasets, indicated by their own colors. (b) Absolute difference in Shannon entropy as a function of grid size.

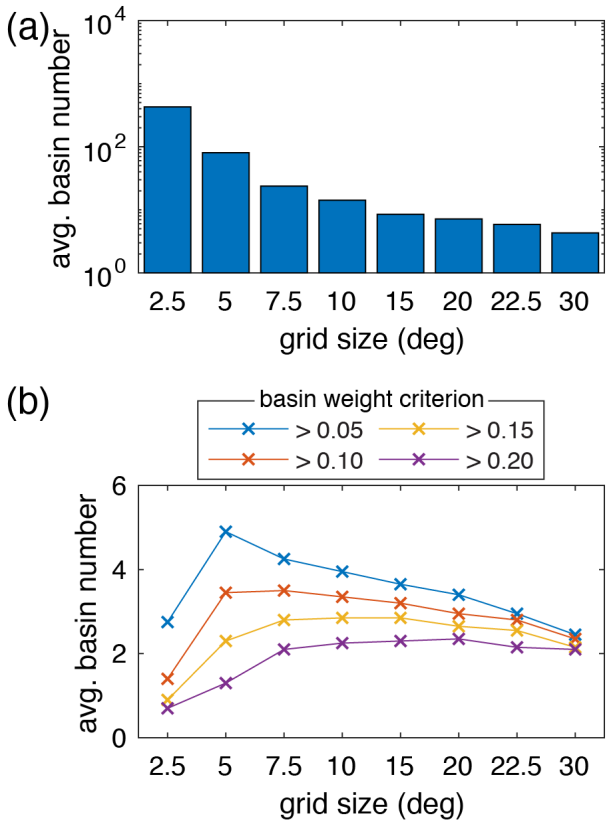

Figure S3. Average numbers of basins for different coarse-grained grid sizes. Averaging was performed over all twenty amino acids. (a) Average number of all basins. (b) Average number of basins with different basin weight criteria, as indicated by different colors.

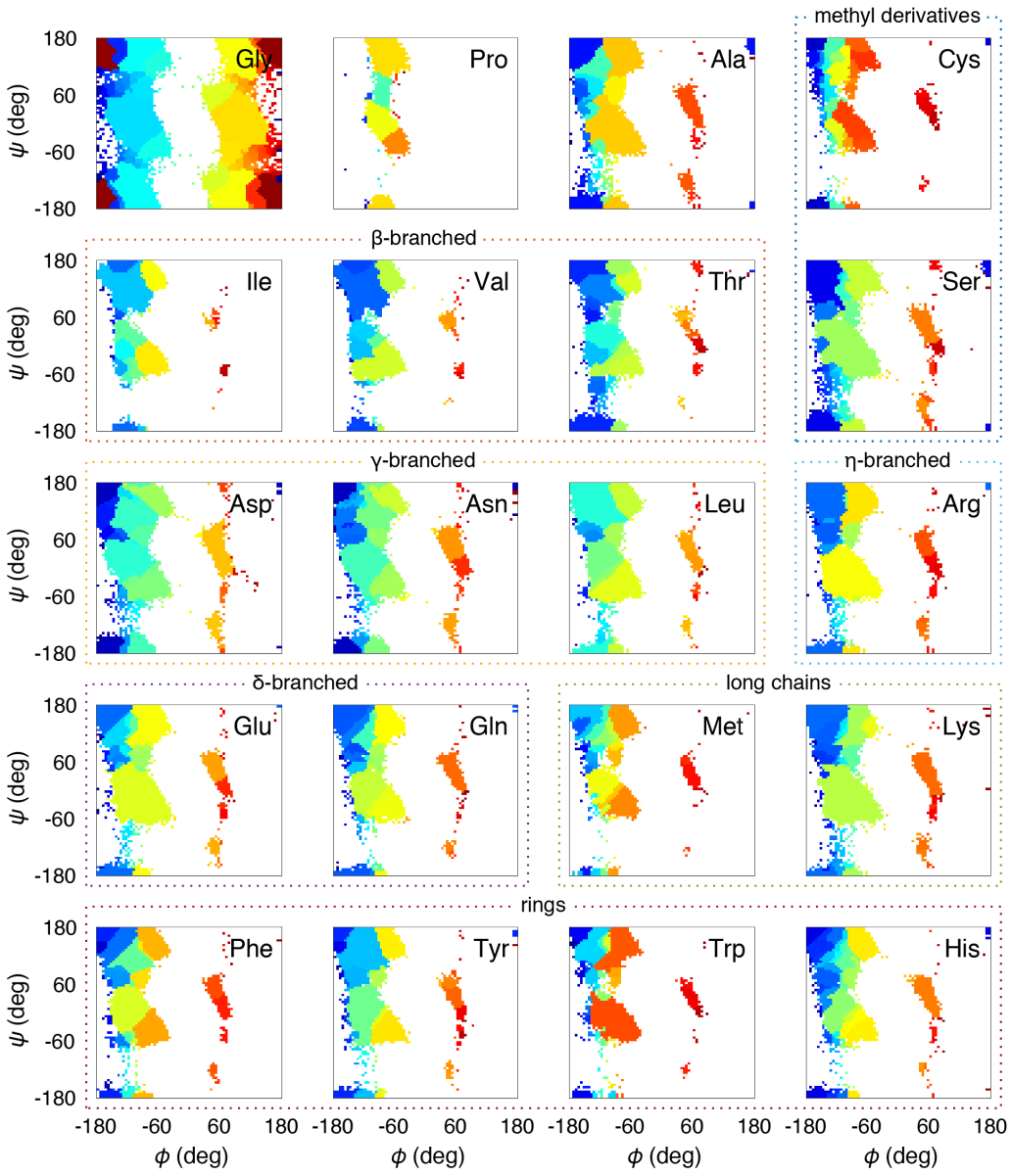

Figure S4. Basin structures for coil-library landscapes of 20 amino acids, generated by a 5° × 5° coarse-grained grid system.

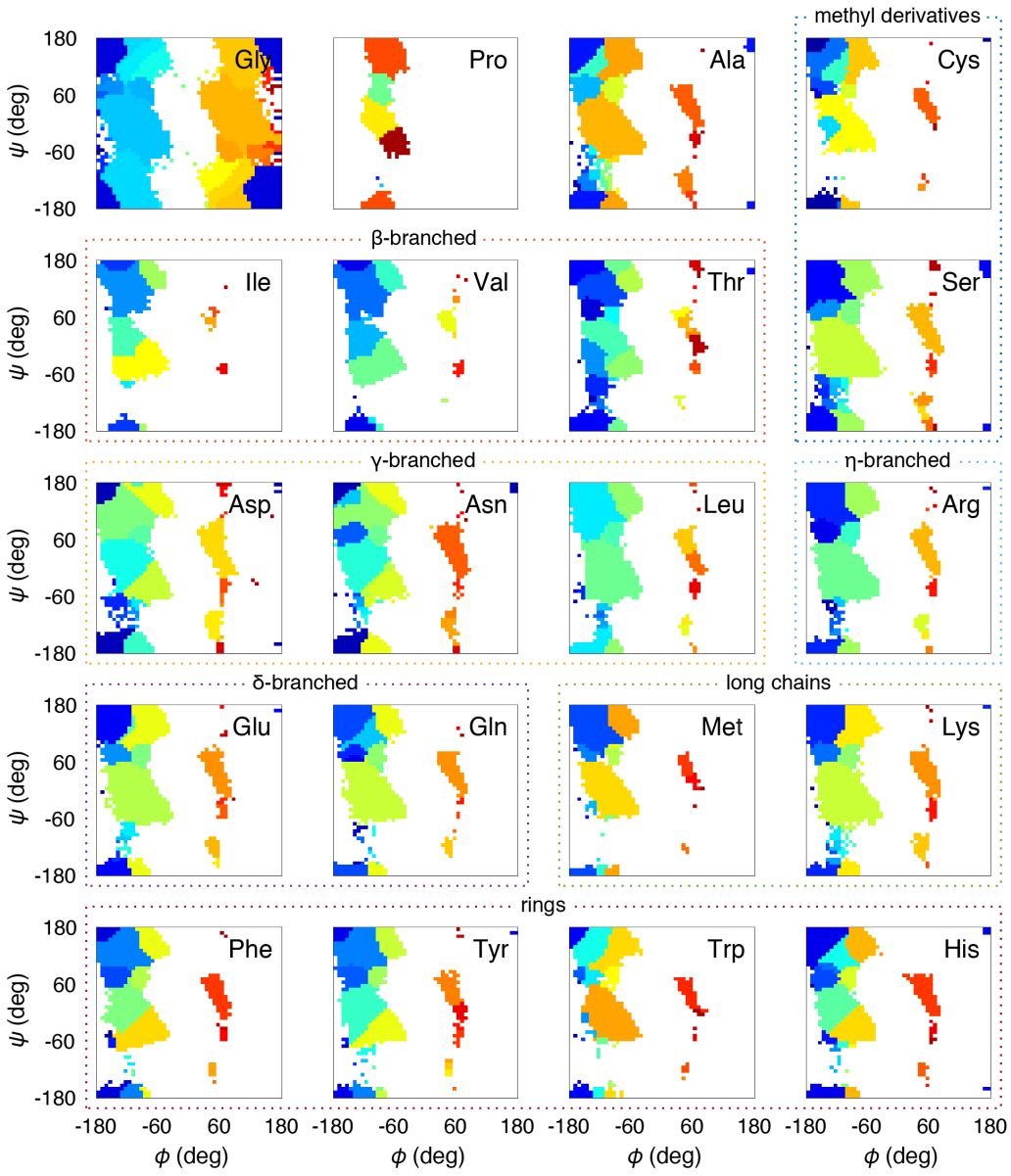

Figure S5. Basin structures for coil-library landscapes of 20 amino acids, generated by a 7.5° × 7.5° coarse-grained grid system.

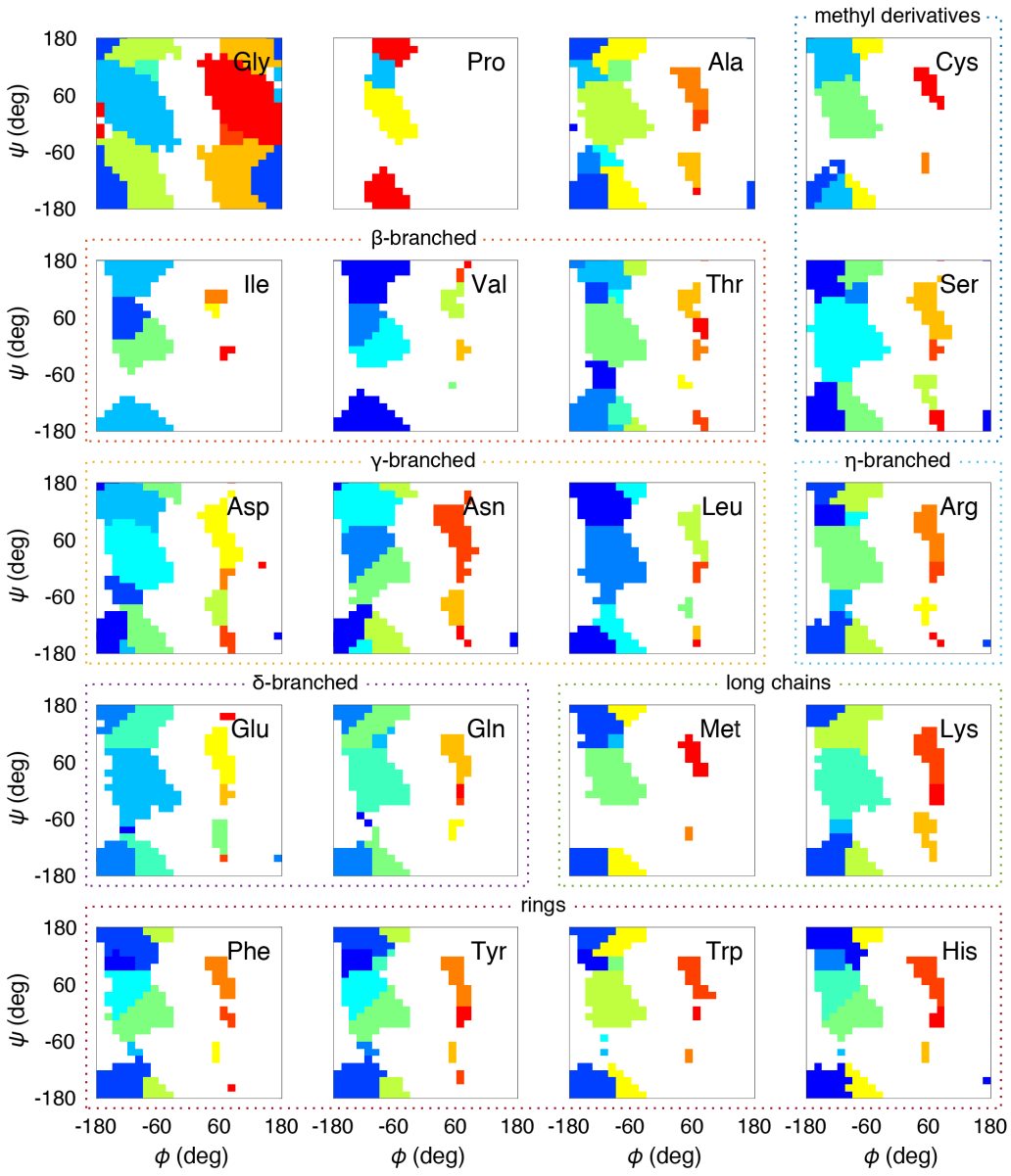

Figure S6. Basin structures for coil-library landscapes of 20 amino acids, generated by a 15° × 15° coarse-grained grid system.

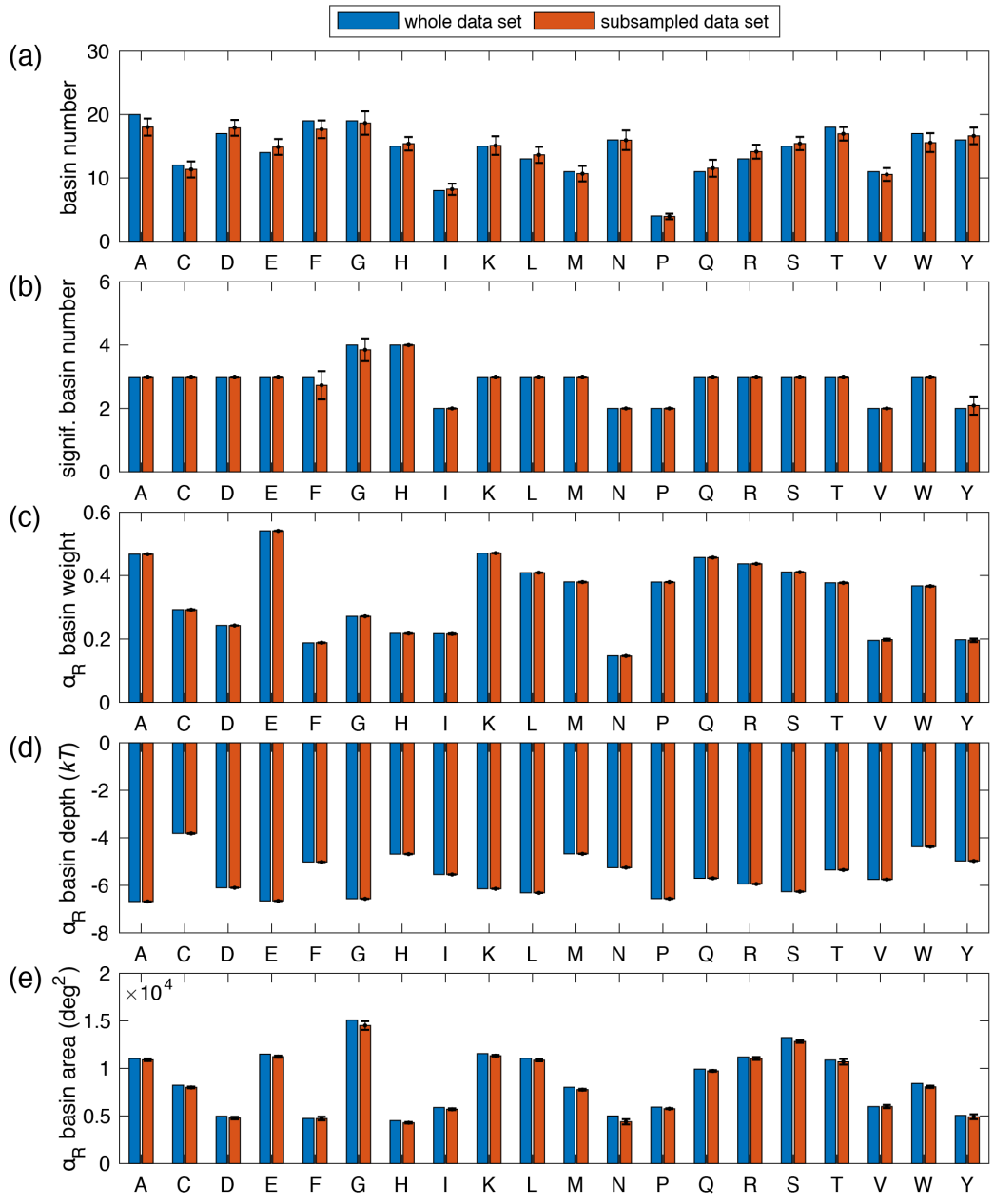

Figure S7. Basin analysis statistics of the whole data set (blue) and average statistics of 100 subsampled data sets (red) for all 20 amino acids. Error bars indicate standard deviations. (a) Total number of basins. Maximum standard deviation (σ_{max}) = 1.85 (for G); minimum standard deviation (σ_{min}) = 0.43 (for P). (b) Number of significant basins, defined as basins with weight larger than 0.15. σ_{max} = 0.44 (for F); σ_{min} = 0.0 (for most amino acids). (c) Weight of the α_R basin. σ_{max} = 5.4×10^{-3} (for Y); σ_{min} = 3.2×10^{-4} (for L). (d) Local minimum depth of the α_R basin in units of k_BT. σ_{max} = 0.013 k_BT (for C); σ_{min} = 0.0026 k_BT (for P). (e) Area of the α_R basin in units of deg^2. σ_{max} = 72 deg^2 (for G); σ_{min} = 7.5 deg^2 (for P).

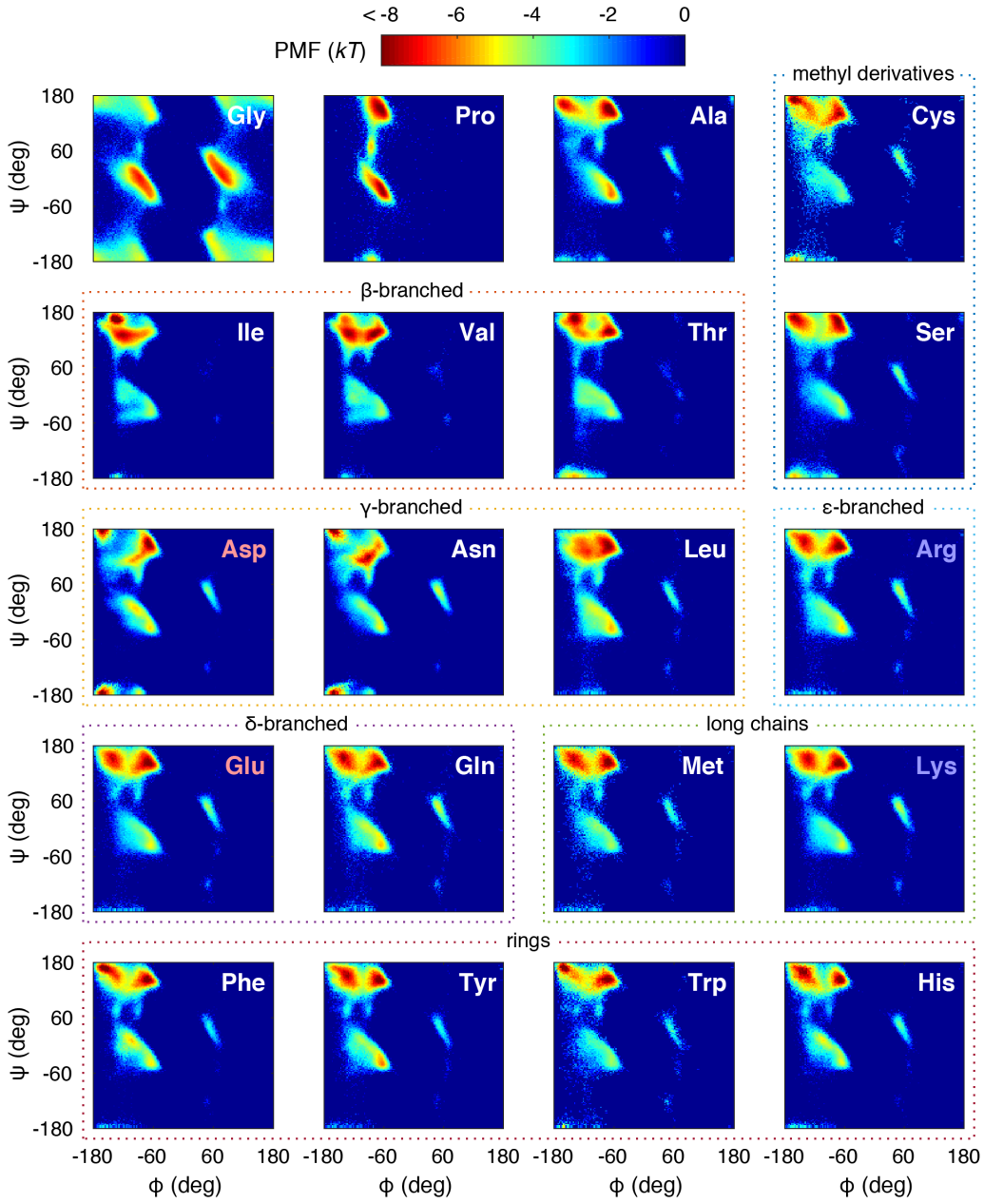

Figure S8. Optimized effective free energy landscapes with parameter $c = 0.975$ for all twenty amino acids with color scale in $\underline{kT}$ (top). The total data size for each amino acid is normalized to 5×10^5. See also **Figure 8**.

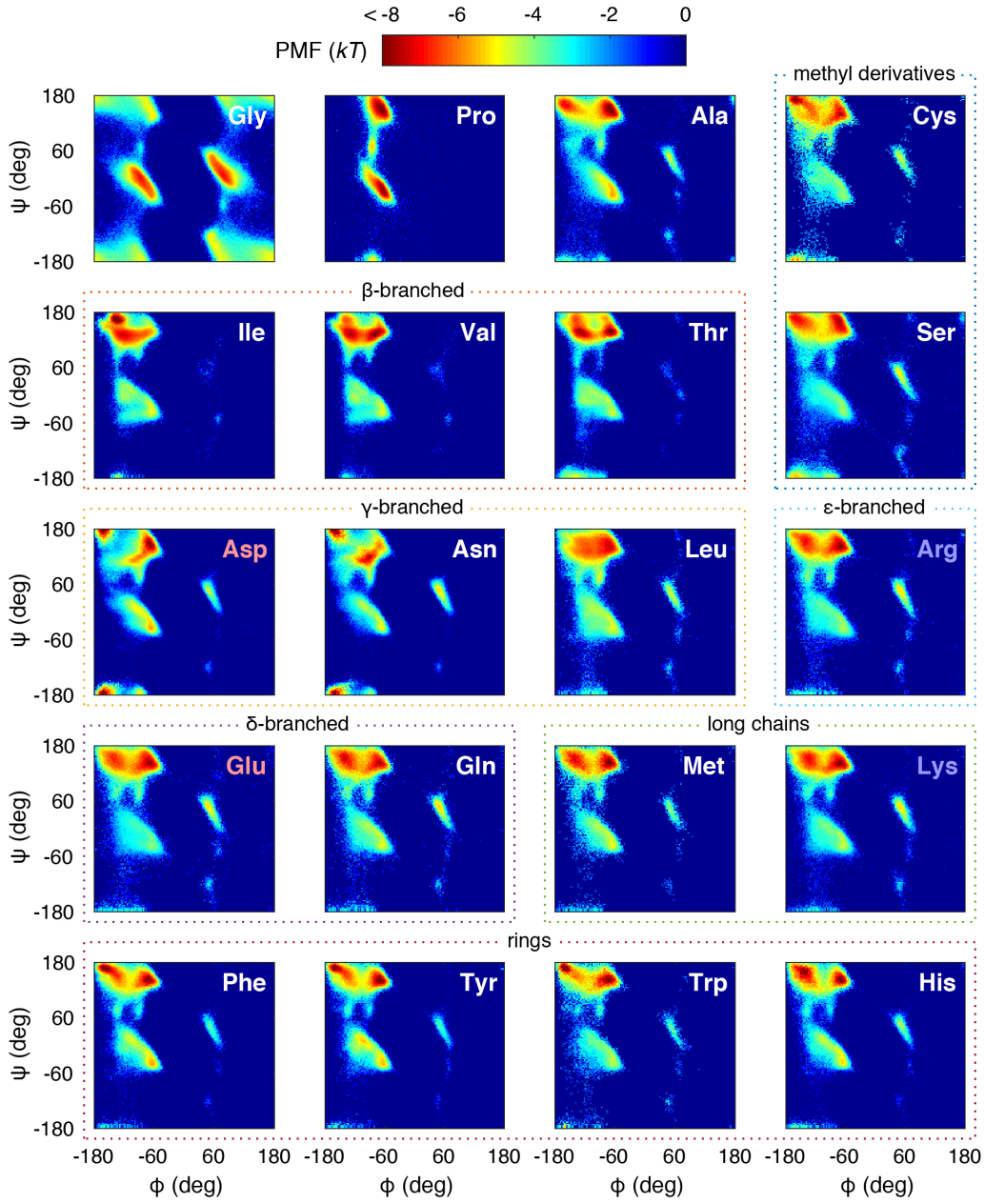

Figure S9. Optimized effective free energy landscapes with parameter c = 0.95 for all twenty amino acids with color scale in $\underline{kT}$ (top). The total data size for each amino acid is normalized to 5×10^5. See also **Figure 8**.