# Decoding Nonspecific Interactions from Nature
# Electronic Supplementary Information

Andrew D White, Ann K Nowinksi, Wenjun Huang, Andrew J Keefe,
Fang Sun, Shaoyi Jiang

## Expected Number of Amino Acid Pairs

To estimate the number of residues which should be next to each other in sequence if the sequence is random, a background model was calculated. It is a multinomial model. The number of pairs can be calculated as:

$$E\left[N_{ij}\right] = N\hat{p}_i\hat{p}_j,\ i \neq j \quad (1)$$

$$E\left[N_{ii}\right] = \frac{1}{2}N\hat{p}_i^2 \quad (2)$$

$$\hat{p}_i = \frac{N_i}{N} \quad (3)$$

$N_{ij}$ is the number of sequence residue pairs between types $i$ and $j$ on the surface, $N_i$ is the number of residues of type $i$ on the surface, and $N$ is the total number of residues on the surfaces. The uncertainty in this estimate, relying on a truncated Taylor expansion, is:

$$\sigma^2_{N_{ij}} = N^2\hat{p}_i^2\sigma^2_{\hat{p}_j} + N\hat{p}_j^2\sigma^2_{\hat{p}_i}\ i \neq j \quad (4)$$

$$\sigma^2_{N_{ii}} = N^2\hat{p}_i^2\sigma^2_{\hat{p}_i} \quad (5)$$

where $\sigma^2_{\hat{p}_i}$ is the sample variance with each protein treated as an observation. This estimate is reasonably valid if the number of categories is high, which in this case is true with 20 amino acids.

## Sensitivity to Surface Cutoff

The effect of changing the surface cutoff is shown in Figure S1. The residue fractions are relatively stable to cutoffs, with lysine (K) and glutamic acid (E) being the most sensitive. The total change across the cutoffs considered here is 2% for the E and K fractions. The cutoff chosen in text was 0.3. Hydrophilic residues tend to increase as the cutoff is increased, which is expected since hydrophilic residues are generally more solvent exposed than hydrohpobic residues.

## Further Detail On Interior Surface Identification

Occluded atoms are atoms which have another atom occluding its orthogonal vector to the principal axis of the protein ($\vec{S}$). The orthogonal vector is occluded when another atom, $O$, has an occlusion line segment $\overline{o}$ which is less than the occlusion margin, $M$, away from $\overline{r}$ (see Figure S2). $\overline{r}$ is given by the formula for the projection of a point onto a line:

$$P_B = \vec{S}\,\frac{P_A \cdot \vec{S}}{|S|}$$

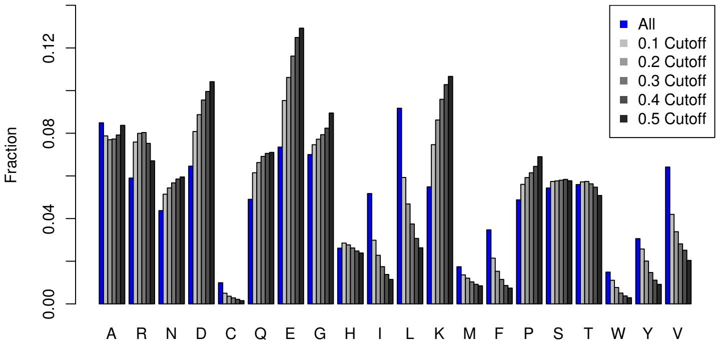


Figure S1: The *E. Coli* residue distribution as a function of surface cutoff. The total residue distribution (no cutoff) is shown in blue. As the cutoff changes, the residue fractions change slightly. There is a tend towards more hydrophilic residues as the cutoff increases, which is expected since hydrophilic residues are generally more surface exposed than hydrophobic residues. A cutoff of 0.3 was chosen for all calculations.

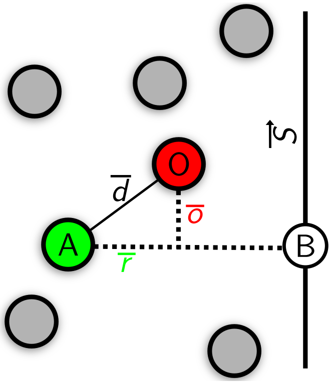


Figure S2: A graphic showing how to test if atom A is occluded by atom O. Atom A's point is projected onto $S$ to obtain line $\overline{r}$. By the Pythagorean theorem, $\overline{o}$ may be found knowing $\overline{r}$ and $\overline{d}$, the distance from Atom A to atom O. If $\overline{o}$ is shorter than the occlusion margin, $M$, atom O is occluding atom A.

| Short Name | PDB ID | Cutoff | $h$ | $M$[Å] | $\vec{S}$[Å] | $\epsilon$ | $r_{max}$ [Å] |
|---|---|---|---|---|---|---|---|
| GroEL Open | 1SX4 | 0.3 | 2 | 2 | $(0, 0, 1)$ | $(-25, 35)$ | 40 |
| GroEL Close | 1SX4 | 0.3 | 2 | 2 | $(0, 0, 1)$ | $(-70, 35)$ | 60 |
| Thermo GroEL Open | 1WE3 | 0.3 | 2 | 1.8 | $(-0.01, -0.68, 0.73)$ | $(-15, 23)$ | 45 |
| Thermo GroEL Close | 1WE3 | 0.3 | 2 | 2.2 | $(-0.01, -0.68, 0.73)$ | $(-75, 30)$ | 60 |
| Group II Close | 3KFB | 0.3 | 2 | 1.8 | $(0.71, 0, 0.71)$ | $(-35, 20)$ | 70 |
| HSP90 Close | 2CG9 | 0.3 | 2 | 1.3 | $(1, 0, 0)$ | $(-30, 45)$ | 22 |
| Yeast CCT | 3P9D | 0.3 | 3 | 1.8 | $(0.83, 0.33, 0.45)$ | $(-61, 61)$ | $\infty$ |

Table S1: The parameters used for identifying interior residues for the various chaperone proteins.

$$\bar{r} = P_B - P_A,\ \bar{d} = P_A - P_O$$

Using the Pythagorean theorem, $\bar{o}$, can be found:

$$|\bar{o}|^2 = \left|\bar{d}\right|^2 - \frac{\left(\bar{d}\cdot\bar{r}\right)^2}{|\bar{r}|^2}$$

where the last term is from the projection of atom O onto $\bar{r}$. If $|\bar{o}| < M$, the atom is occluded. Only heavy atoms are tested and they are tested against only heavy atoms. After tabulating all occluded atoms, if a residue contains greater than or equal to $h$ occluded heavy atoms, it is considered an exterior residue.

A few additional details are necessary as well. First, the occluding atom $O$ must not lie on the other side of $\vec{S}$ from the atom being considered $A$. $O$ must not be behind $A$ ($\bar{d}\cdot\bar{r} > 0$). Residues whose projection onto $\vec{S}$ exceeds an extent, $\epsilon\vec{S}$, are considered occluded. Finally, residues which are farther away than $r_{max}$ are considered occluded. A table containing the specific parameters used for each chaperone protein are shown in Table S1.

# Glycine-Serine Pair

As mentioned in the main text discussing Figure 3a, the glycine-serine pair is order specific and the SG order is 2.5 times more favored than the GS. If we examine the mode of the $\phi - \psi$ angles for that ordering, it is clear that the most commonly observed secondary structure is Type-II turns. The modes were calculated as the density maximum after using a kernel density estimator (bw= $10°$, grid= $1000 \times 1000$) as seen in Figure S4. The serine has a mode $(\phi, \psi)$ of $(-57°, 132°)$ and the glycine has a mode of $(88°, -2°)$. The ideal Type-II turn values are $(-60°, 131°)$ followed by $(84°, 1°)$. Type-II turns are secondary structure motifs and we do not consider them relevant to understanding non-specific interactions.

# Interaction Energy Definition

The interaction energies are derived from counts of residue contacts and total number of residues. The interaction energy is defined as:

$$\chi_{AB} \equiv U_{AB} - U_A - U_B$$

where $U$ indicates energy and $A$ and $B$ are residue types. Now, substituting the Boltzmann distribution:

$$\chi_{AB} = -\frac{1}{\beta}\ln\left(e^{-\beta U_{AB}}\right) + \frac{1}{\beta}\ln\left(e^{-\beta U_A}\right) + \frac{1}{\beta}\ln\left(e^{-\beta U_B}\right), \quad \beta = kT$$

where $k$ is Boltzmann's constant and $T$ is the temperature. Replacing the Boltzmann distribution with the probabilities:

$$\chi_{AB} = -\frac{1}{\beta}\ln P_{AB} + \frac{1}{\beta}\ln P_A + \frac{1}{\beta}\ln P_B = -\frac{1}{\beta}\ln\frac{P_{AB}}{P_A P_B} \qquad (6)$$

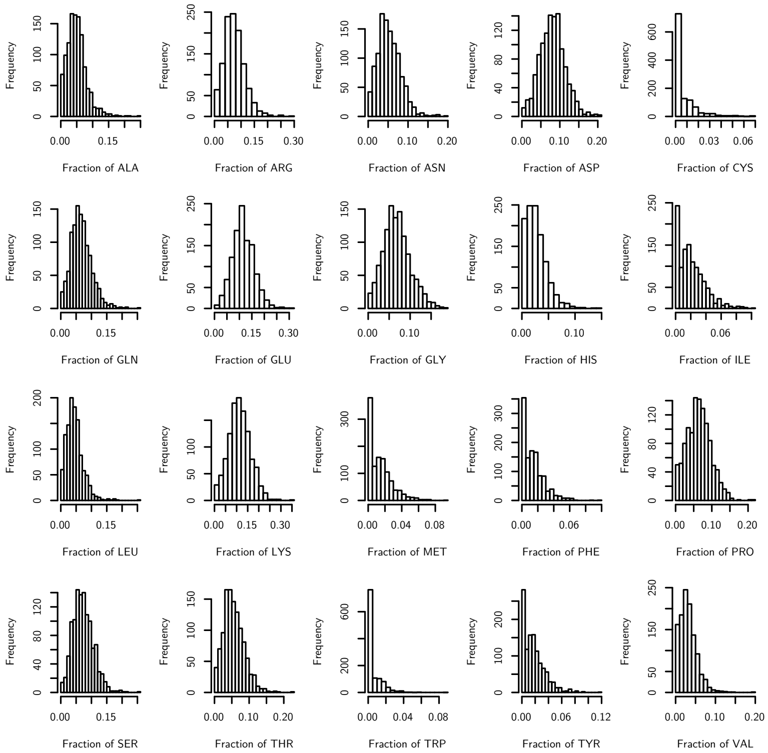


Figure S3: A histogram of the 20 amino acid surface fractions on protein surfaces. This plot shows that the assumption of normality is true for the more commonly observed residues.

The probabilities may be found using maximum likelihood estimators[1] and shown to be:

$$\hat{P}_A = \frac{N_A}{\sum_i N_i}, \quad \hat{P}_{AB} = \frac{N_{AB}}{N_{\text{Free A}} + \sum_y N_{Ay}} \tag{7}$$

where $N_A$ indicates the number of residues of type $A$, $N_{AB}$ indicates the number of $A, B$ pairs and 'Free' indicates unpaired $A$. The summation in the denominator is across all pairs where $A$ is part of the pair.

Residue contacts were calculated by finding the pair-wise distance between each side-chain heavy atom on each residue. Neighboring residues, as determined by residue indices in PDB files, were excluded from being in contact. If any of the heavy-atom pairs were below the cutoff distance, the side-chain pair is said to be in contact. The heavy atom pair cutoffs were taken to be Van der Waals energy minimum radii. The following Van der Waal radii were used: nitrogen: 3.25Å, oxygen 2.96Å, and sulfur 3.55Å. The following mixing rule was applied:

$$r_{ij} = 2^{1/6}\sqrt{\sigma_i \sigma_j} \tag{8}$$

where $\sigma$ is the Van der Waals radius.

## Error Analysis

Figure 1 uses standard error for the error bars and the amino acid surface fractions errors reported in text were standard errors. Standard error is calculated as:

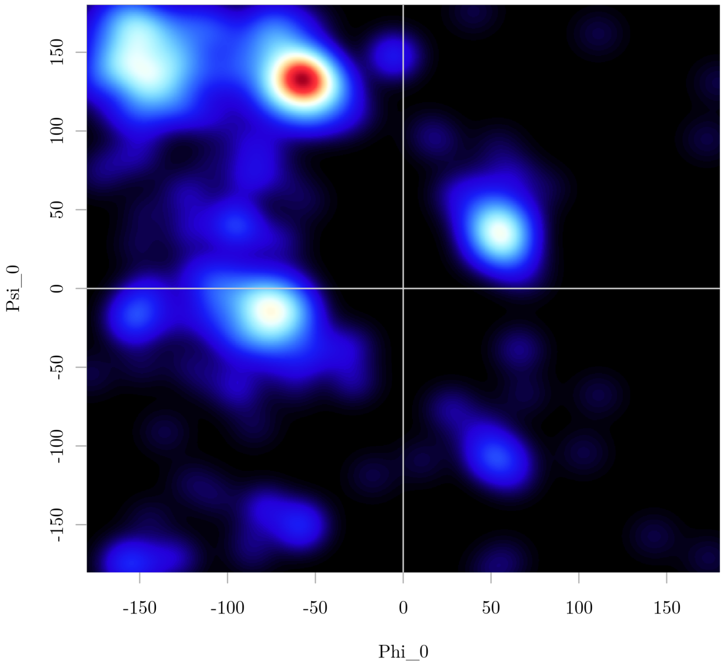
$$\sigma_i = \frac{1}{\sqrt{N}}\sqrt{\frac{\sum_j (p_{ij} - \hat{p}_j)^2}{N-1}} \tag{9}$$

where $i$ is the residue type, $j$ is the protein, $p_{ij}$ is the fraction of residue type $i$ on protein $j$, $\hat{p}_i$ is the average residue fraction of type $i$, and $N$ is the number of protein. Notice that the standard error calculations were done by considering each residue fraction on each protein as a single observation. The choice of standard error means the error considered is uncertainty in the mean. This is different from standard deviation. This is because our design principles should be based on our materials being in contact with a population of proteins, not a single protein. Thus, uncertainty in the mean is most important. The use of these statistics assumes normality, which can be seen visually in Figure S2. As one can see, the higher residue fractions are normal and the assumption is valid. For the more rarely observed residues, for example cysteine, the assumption of normality is not correct and the error bars may not be exact in Figures 1 and 2. Note that this does not affect the values themselves.

A sensitivity analysis similar to bootstrap resampling was used for the error analysis called "bootstrap error," specifically in Figure 3, except for the error bars in Figure 3a. The sensitivity analysis was done by treating each protein as an independent observation. Pseudo-replicates of the data were created by sampling from the protein dataset with replacement. Each pseudo-replicate has the same number of proteins as the original dataset; there are repeats and omissions in the pseudo-replicates. The sampling was done 500 times for each statistic. Quantiles were calculated on the pseudo-replicates to obtain errors. 95% confidence intervals are shown as error bars in Figures 3b, 3c, and 3d.

## References

[1] Guttorp, P. Stochastic Modeling of Scientific Data (Chapman and Hall/CRC, London, 1995).



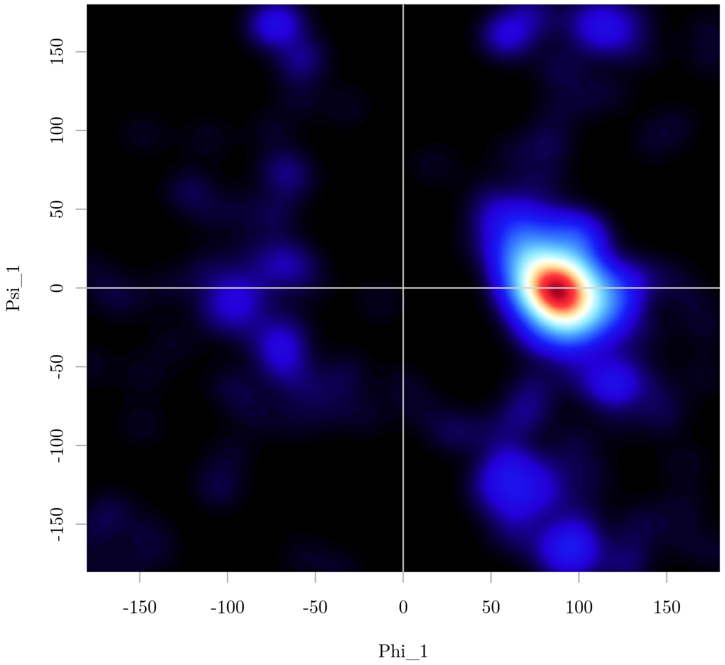
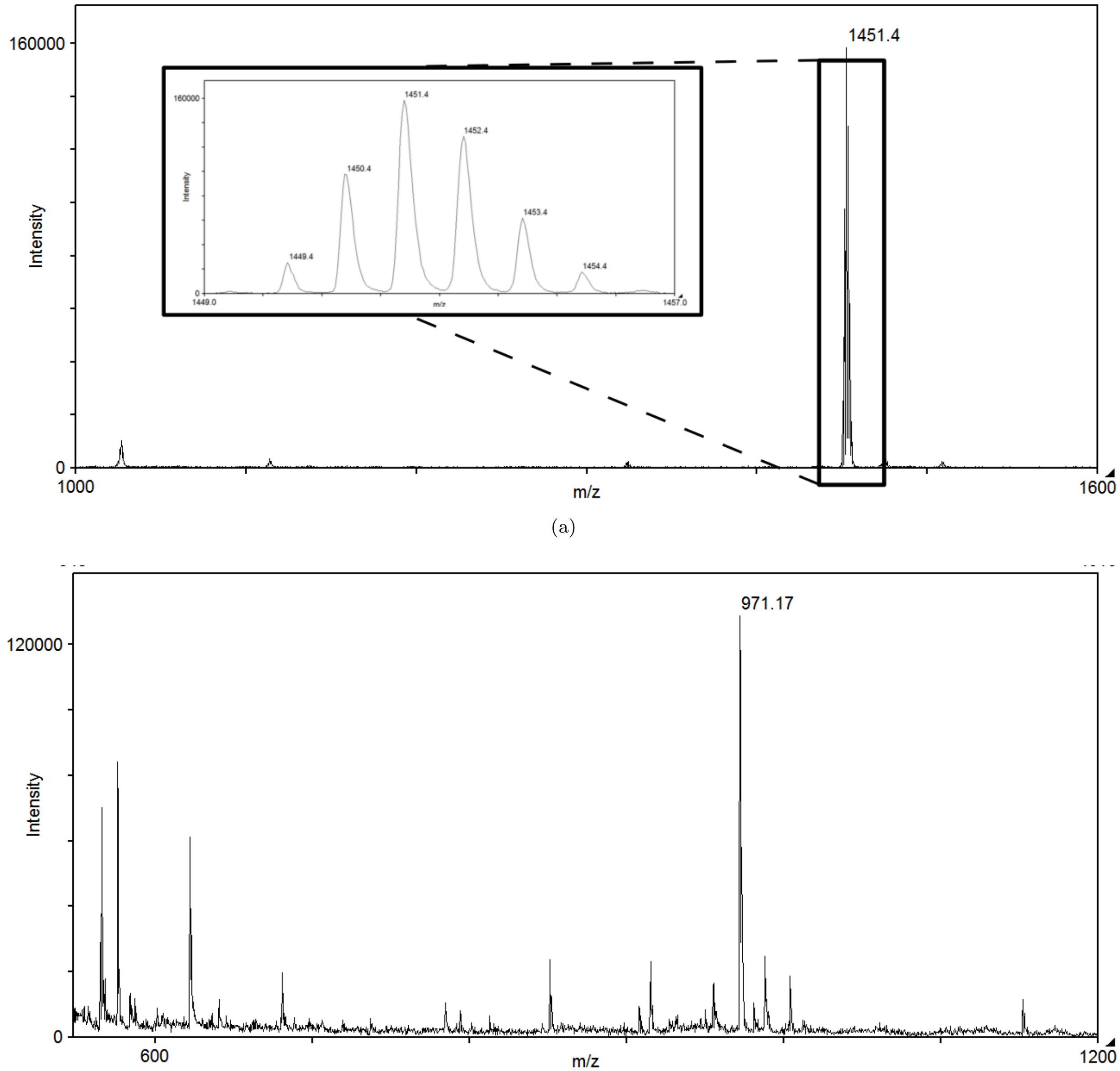


Figure S4: Dihedral $\phi$, $\psi$ plots for pairs of serine glycine, in that order. Notice that the density maximum is the value for a Type-II turn.



(a)

(b)

Figure S5: MALDI-TOF mass spectra of (a) Ac-[EK]$_7$PPPPC-Am, where the brackets indicate random ordering of K and E, and (b) Ac-GGGGGGGPPPPC-Am. The expected molecular weight of Ac-[EK]$_7$PPPPC-Am is about 1451. The expected binomial distribution of molecular weights around this value can be seen in the MALDI spectra. The expected molecular weight of Ac-GGGGGGGPPPPC-Am is 950.03. The molecular weight of Ac-GGGGGGGPPPPC-Am is higher by 21.14, which corresponds to a positive sodium ion conjugated to the peptide and the loss of a proton.

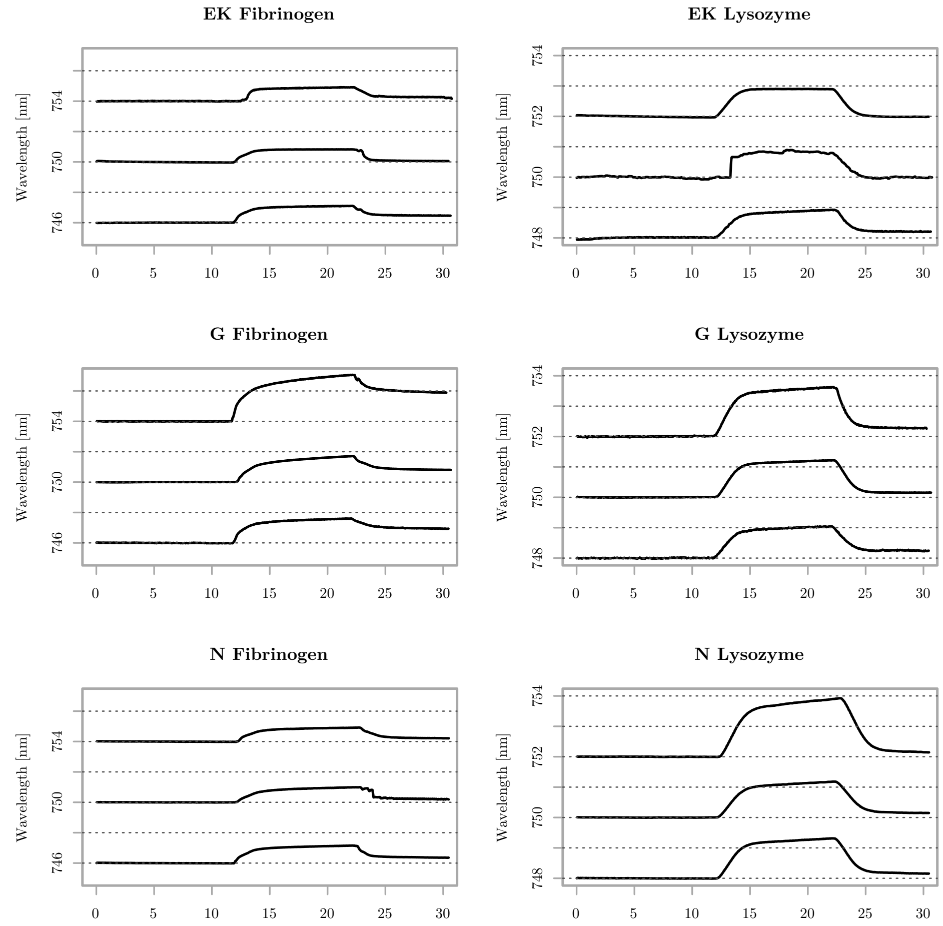


Figure S6: SPR sensograms of protein adsorption for the Ac-[EK]$_7$PPPPC-Am (EK), Ac-GGGGGGGPPPPC-Am (G), and Ac-CPPPPNNNNNNN-Am (N) sequences. ng/cm$^2$ protein adsorption can be obtained by multiplying the wavelength shift by 17. The values for EK are $4.4 \pm 3.0\,\mathrm{ng/\,cm^2}$ and $1.7 \pm 1.6\,\mathrm{ng/\,cm^2}$ for fibrinogen and lysozyme, respectively. The values for G are $17.0 \pm 7.0\,\mathrm{ng/\,cm^2}$ and $3.2 \pm 1.0\,\mathrm{ng/\,cm^2}$ for fibrinogen and lysozyme, respectively. The values for N are $5.2 \pm 1.0\,\mathrm{ng/\,cm^2}$ and $2.7 \pm 0.4\,\mathrm{ng/\,cm^2}$ for fibrinogen and lysozyme, respectively.