Protein Preferential Solvation in Water:Glycerol Mixtures

Supporting Information

Nicolas Chéron,[†] Margaux Naepels,[†] Eva Pluhařová,[†,‡] and Damien Laage*[,†]

†*PASTEUR, Département de chimie, École Normale Supérieure, PSL University, Sorbonne Université, CNRS, 75005 Paris, France*

‡*Present address: J. Heyrovský Institute of Physical Chemistry, Czech Academy of Sciences, Dolejskova 2155/3, 182 23 Prague, Czech Republic*

E-mail: damien.laage@ens.fr

Phone: +33 (0)144322418

SI-1. Derivation of eq. 6

$$\Gamma_{GP} = \langle n_G^{\text{shell}} \rangle - \frac{\langle n_G^{\text{bulk}} \rangle}{\langle n_W^{\text{bulk}} \rangle} * \langle n_W^{\text{shell}} \rangle \tag{S1}$$

$$= \frac{\left(\langle n_G^{\text{shell}} \rangle * \langle n_W^{\text{bulk}} \rangle - \langle n_G^{\text{bulk}} \rangle * \langle n_W^{\text{shell}} \rangle \right)}{\langle n_W^{\text{bulk}} \rangle} \tag{S2}$$

$$= \frac{\left(\langle n_G^{\text{shell}} \rangle * \langle n_W^{\text{bulk}} \rangle - \langle n_G^{\text{bulk}} \rangle * \langle n_W^{\text{shell}} \rangle \right) + \left(\langle n_G^{\text{shell}} \rangle * \langle n_G^{\text{bulk}} \rangle - \langle n_G^{\text{shell}} \rangle * \langle n_G^{\text{bulk}} \rangle \right)}{\langle n_W^{\text{bulk}} \rangle} \tag{S3}$$

$$= \frac{\langle n_G^{\text{shell}} \rangle * \left(\langle n_W^{\text{bulk}} \rangle + \langle n_G^{\text{bulk}} \rangle \right) - \langle n_G^{\text{bulk}} \rangle * \left(\langle n_W^{\text{shell}} \rangle + \langle n_G^{\text{shell}} \rangle \right)}{\langle n_W^{\text{bulk}} \rangle} \tag{S4}$$

$$= \frac{\left\langle n_W^{\text{bulk}} + n_G^{\text{bulk}} \right\rangle * \left\langle n_W^{\text{shell}} + n_G^{\text{shell}} \right\rangle}{\langle n_W^{\text{bulk}} \rangle} * \left(\frac{\langle n_G^{\text{shell}} \rangle}{\langle n_W^{\text{shell}} + n_G^{\text{shell}} \rangle} - \frac{\langle n_G^{\text{bulk}} \rangle}{\langle n_W^{\text{bulk}} + n_G^{\text{bulk}} \rangle} \right) \tag{S5}$$

$$= \frac{\left\langle n_G^{\text{shell}} + n_W^{\text{shell}} \right\rangle * \left(x_G^{\text{shell}} - x_G^{\text{bulk}} \right)}{1 - x_G^{\text{bulk}}} \tag{S6}$$

SI-2. Heavy-atom based definitions

Heavy-atom based definitions are employed to determine the average numbers of water and glycerol molecules at increasing distances from the protein surface. We note OW the water oxygen atom, and O1/C1/O2/C2/O3/C3 the glycerol heavy atoms. n_X^{shell} are defined as

$$n_W^{\text{shell}}(R) = n_{OW}(R) \tag{S7}$$

$$n_G^{\text{shell}}(R) = \frac{n_{O1}(R) + n_{C1}(R) + n_{O2}(R) + n_{C2}(R) + n_{O3}(R) + n_{C3}(R)}{6} \tag{S8}$$

where $n_Y(R)$ is the number of atom Y found within the distance R from the protein.

Γ_{GP} and $\Delta x_G^{\text{shell}}$ are computed from the MD trajectory as

$$\Gamma_{GP}(R) = \left\langle n_G^{\text{shell}}(R) \right\rangle - \frac{\left\langle n_G^{\text{bulk}}(R) \right\rangle}{\left\langle n_W^{\text{bulk}}(R) \right\rangle} * \left\langle n_W^{\text{shell}}(R) \right\rangle \tag{S9}$$

$$= \left\langle n_G^{\text{shell}}(R) \right\rangle - \frac{\left\langle N_G^{\text{total}} - n_G^{\text{shell}}(R) \right\rangle}{\left\langle N_W^{\text{total}} - n_W^{\text{shell}}(R) \right\rangle} * \left\langle n_W^{\text{shell}}(R) \right\rangle \tag{S10}$$

and

$$\Delta x_G^{\text{shell}}(R) = x_G^{\text{shell}}(R) - x_G^{\text{bulk}}(R) = \frac{\left\langle n_G^{\text{shell}}(R) \right\rangle}{\left\langle n_W^{\text{shell}}(R) + n_G^{\text{shell}}(R) \right\rangle} - \frac{\left\langle n_G^{\text{bulk}}(R) \right\rangle}{\left\langle n_W^{\text{bulk}}(R) + n_G^{\text{bulk}}(R) \right\rangle} \tag{S11}$$

$$= \frac{\left\langle n_G^{\text{shell}}(R) \right\rangle}{\left\langle n_W^{\text{shell}}(R) + n_G^{\text{shell}}(R) \right\rangle} - \frac{\left\langle N_G^{\text{total}} - n_G^{\text{shell}}(R) \right\rangle}{\left\langle N_W^{\text{total}} - n_W^{\text{shell}}(R) + N_G^{\text{total}} - n_G^{\text{shell}}(R) \right\rangle} \tag{S12}$$

where $n_X^{\text{bulk}}(R) = N_X^{\text{total}} - n_X^{\text{shell}}(R)$, with N_X^{total} the total number of molecules X in the simulation box (see Table 1). The n_X^{shell} terms are distance-dependent since they depend on the choice of r^{shell}.

SI-3. Comparisons of different approaches to compute Γ_{GP}

We present in Figure S1 plots of the influence of the bulk glycerol molar fraction on the preferential interaction coefficient Γ_{GP} when it is calculated with Equation 4 and with Equation 5.[1] It appears that the values are very close (both averages and error bars) over the considered concentration range.

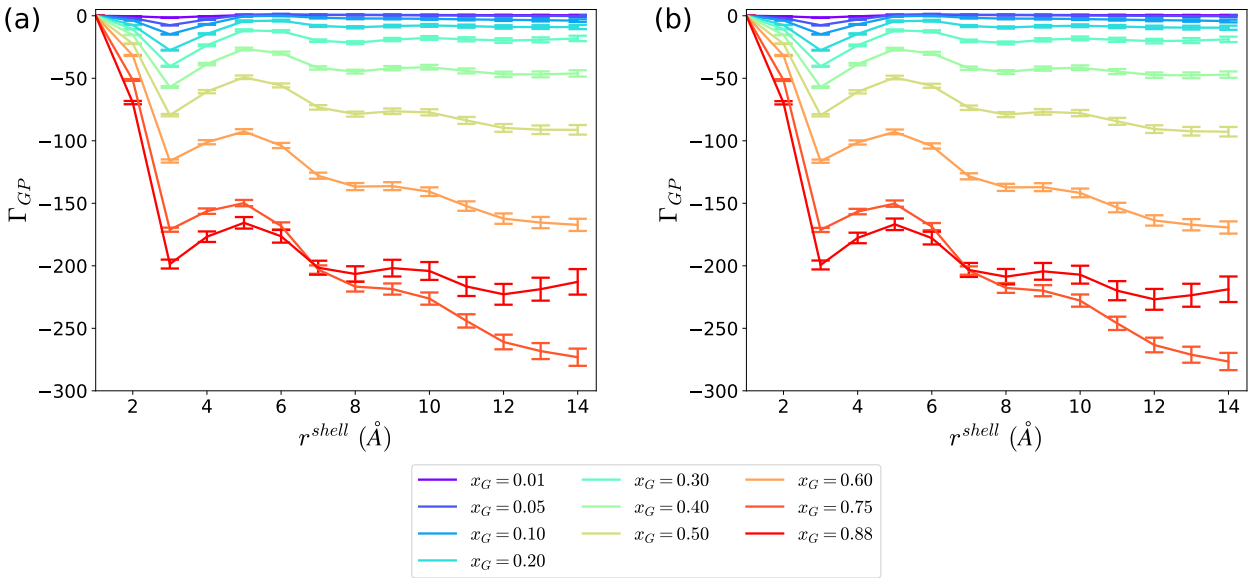

Figure S1: Comparisons of two approaches to compute the preferential interaction coefficient Γ_{GP}: (a) with Eq. 4 (this report) and (b) with Eq. 5.[1]

SI-4. Simulation convergence

Figure S2 shows the autocorrelation function of $\Gamma_{GP}(14\text{Å})$ at $x_G = 0.20$ and 300 K. The relaxation time is approximately 600 ps.

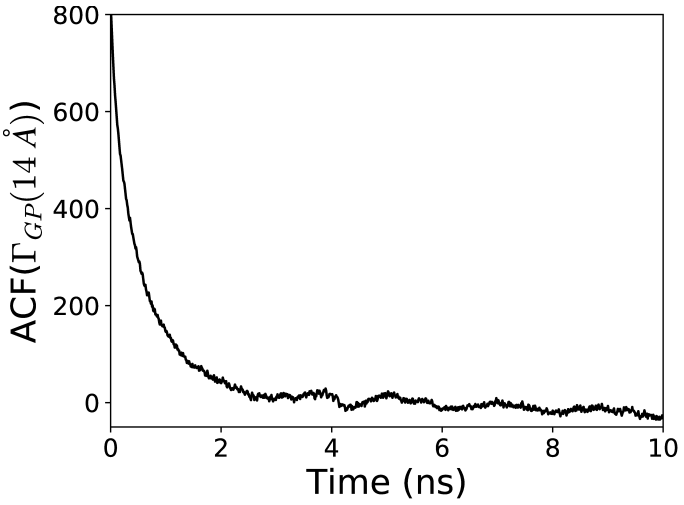

Figure S 2: Autocorrelation function of $\Gamma_{GP}(14\text{Å})$ with $x_G = 0.20$ at 300 K (averaged over 10 simulations of 100 ns).

We now compare in Figure S3 results obtained with different protocols: two independent sets of 10 replicas of 100 ns (Set1 and Set2) as well as two simulations of 1000 ns (Set3 and Set4). For 10 multiple replicas (Set1 and Set2), we define the cumulative Γ_{GP} at time t as the gathering of all data up to time $t/10$ of the simulations. For single continuous trajectories (Set3 and Set4), the cumulative Γ_{GP} at time t is simply Γ_{GP} calculated with data up to time t of the simulations. Thus, the same amount of data is used at each time of the four sets.

It appears clearly that using 10 independent replicas is less biased towards the initial configurations than using single long trajectories since values of cumulative Γ_{GP} are much closer between the two equivalent sets, eventhough at long time scale all protocols converge towards the same value. $\Gamma_{GP}(14\text{Å})$ was also calculated in independent blocks of 20 ns along the trajectories (either gathering 10 blocks of 2 ns for multiple replicas or a single block of 20 ns for long trajectories); the standard errors between all blocks were small for Set1 and Set2 (0.52 and 0.50) whereas they were higher for Set3 and Set4 (0.65 and 0.65). This

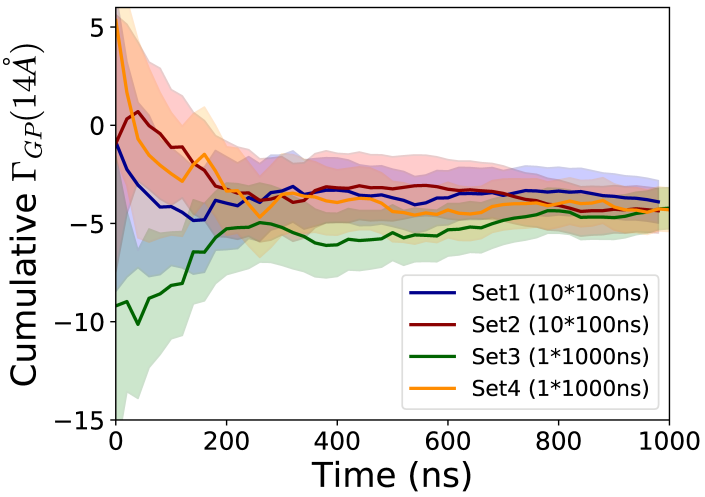

Figure S 3: Comparisons of different approaches to compute the preferential interaction coefficient Γ_{GP} with $x_G = 0.10$ at 14 Å. Set1 and Set2 represent two independent sets of 10 simulations of 100 ns (after 50 ns of equilibration for each). Set3 and Set4 represent two independent simulations of 1000 ns (after 50 ns of equilibration). Error bars are in shaded areas.

quantifies the higher variations between each time blocks obtained when using single long trajectories. Moreover, with the current computational ressources, running 10 simulations in parallel is straightforward. Thus, in the present paper we decided to average 10 independent replicas since it allows to reach convergence faster than with single long trajectories by allowing better sampling.

We observed that for low molar fractions ($x_G \leqslant 0.60$), 50 ns of equilibration are needed (see Figure 3 of main text). However, at higher molar fractions ($x_G = 0.75$ and $x_G = 0.88$) 200 ns of equilibration are needed. The reason why long simulation times are needed to obtain converged data at high molar fractions is probably linked to the viscosity as well as the way the systems were prepared: during the initial placement of glycerol molecules, they couldn't be placed too close to the protein due to a steric clash criteria and the first shells were consequently enriched in water. During the simulations, the systems equilibrate and Γ_{GP} increases. Thus, not only a single simulation is not enough to obtain proper sampling, but equilibration should be long enough.

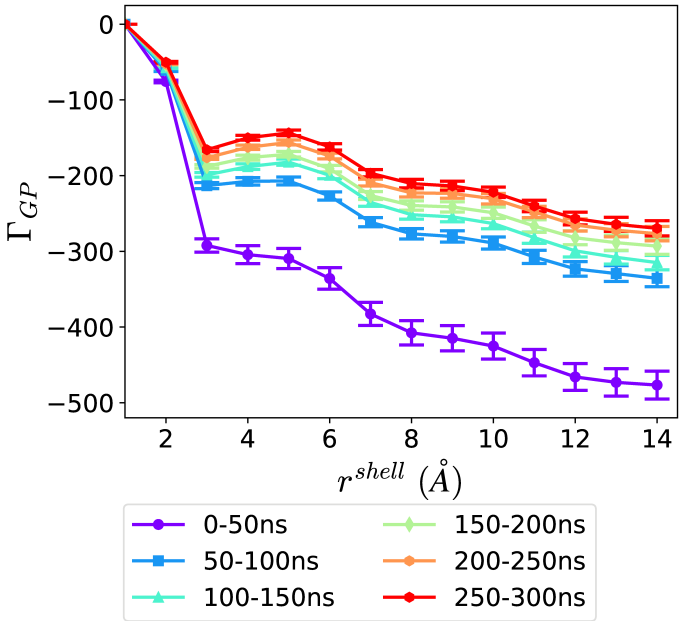

Figure S 4: Comparisons of Γ_{GP} calculated on successive time blocks for $x_G = 0.75$.

SI-5. Influence of individual residues

We present in Figure S5 the influence of each residue on Δx_G^{shell} and Γ_{GP}. At low molar fraction ($x_G = 0.10$), we find that some parts of the protein are preferentially hydrated whereas other parts are preferentially glycerated (Γ_{GP} from -3.7 to 1.0). At higher molar fraction ($x_G = 0.40$ and $x_G = 0.60$) all residues exhibit glycerol depletion, with different magnitudes though (Γ_{GP} ranging from -22.4 to -3.8 for $x_G = 0.40$, and from -58.2 to -8.8 for $x_G = 0.60$). Overall, it seems that the observed depletion of water molecules around the protein at these molar fractions is residue dependent. This result shows that the global effect observed can indeed partially be explained by chemical properties of the protein surface such as hydrophobicity or polarity. However, as stated in the main text, local residue properties are not enough and the shape of the protein plays a role. We present in Figure S6 the scatter plots related to data presented in Table 2 of main text.

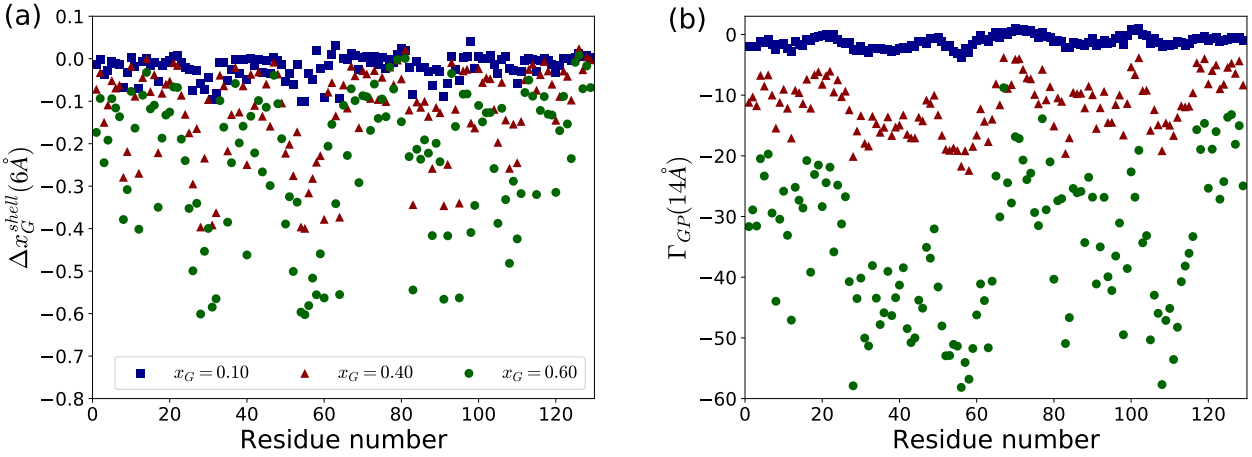

Figure S 5: Residue dependence on the difference of molar fractions and the preferential interaction coefficient for lysozyme for three values of the molar fraction x_G.

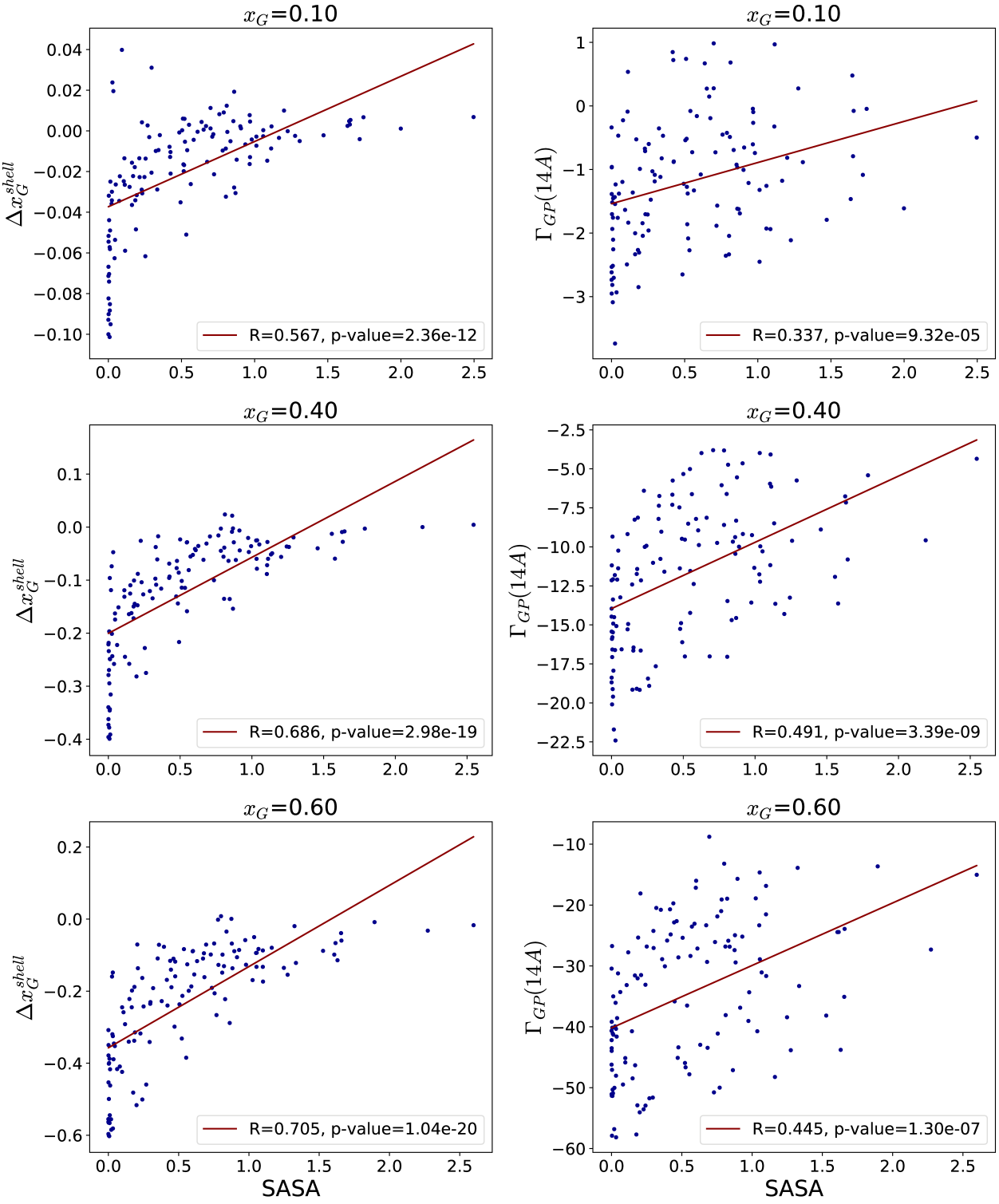

Figure S 6: Scatter plots of the correlation between Δx_G^{shell} and Γ_{GP}, and SASA (probe size of 1.4Å) at selected molar fractions for lysozyme.

SI-6. Shell volume

The volume of the solvation shell is determined as

$$V_{shell} = \left\langle n_G^{\text{shell}} \right\rangle * V_m^G + \left\langle n_W^{\text{shell}} \right\rangle * V_m^W \tag{S13}$$

$$= \left\langle N^{\text{shell}} \right\rangle * \left(x_G^{\text{shell}} * V_m^G + x_W^{\text{shell}} * V_m^W \right) \tag{S14}$$

where $V_m^{G,W}$ are the glycerol and water molar volumes, $n_{G,W}^{\text{shell}}$ are the numbers of glycerol and water molecules in the shell, and N^{shell} is the total number of solvent molecules in the shell. We consider molar volumes to be constant for all water:glycerol compositions, since for glycerol:water mixtures with no protein, they were measured to vary by less than 1.5% for water and less than 1% for glycerol.[2]

SI-7. Influence of temperature

Figures S7 and S8 show the impact of temperature on the total number of solvent molecules in the first shell and on the shell volume. The change in the number of molecules with temperature is very moderate (-4% between 300 and 350 K at both molar fractions); because the preferential solvation coefficient is proportional to the number of solvent molecules in the shell (see Eq. (6)), this result is important to discrimate the temperature effects on the shell population and on the shell composition. The temperature effect on the shell volume is much stronger: for example, the volume at $x_G = 0.05$ and 350 K is similar to the volume at $x_G = 0.75$ and 300 K.

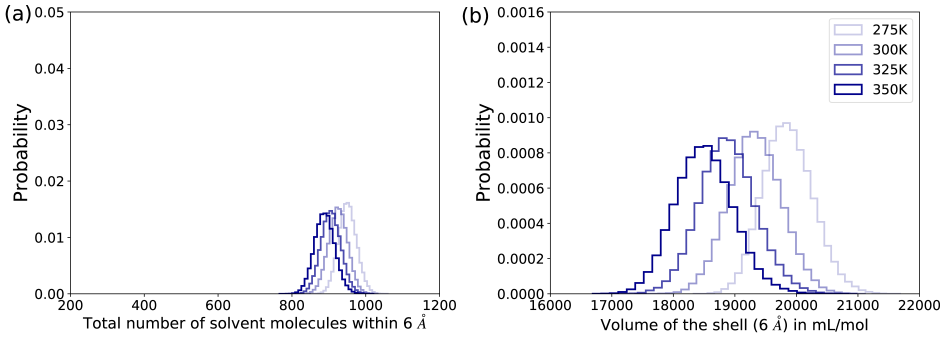

Figure S 7: (a) Number of molecules in the first shell, and (b) first shell volume for lysozyme at $x_G = 0.05$ and at different temperatures. The scale of the plot is the same as in Figure 5 of the main text.

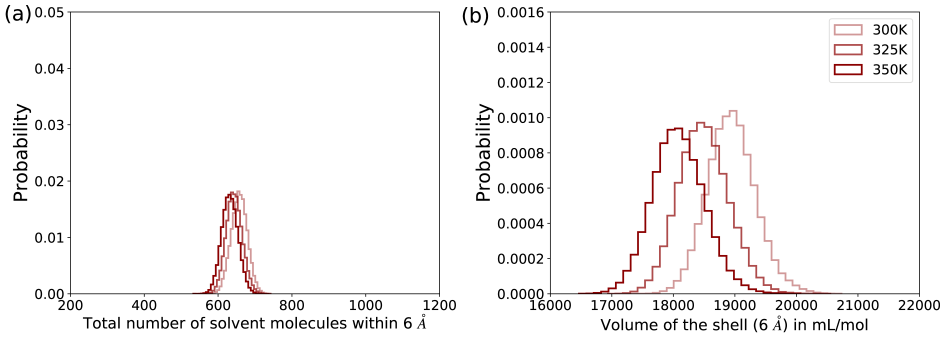

Figure S 8: (a) Number of molecules in the first shell, and (b) first shell volume for lysozyme at $x_G = 0.20$ and at different temperatures. The scale of the plot is the same as in Figure 5 of the main text.

We present below the derivation of equation 8, starting from equation 6.

$$\Gamma_{GP} = \frac{\left\langle n_G^{\text{shell}} + n_W^{\text{shell}} \right\rangle * \left(x_G^{\text{shell}} - x_G^{\text{bulk}} \right)}{1 - x_G^{\text{bulk}}} \tag{S15}$$

$$= \left\langle n_G^{\text{shell}} + n_W^{\text{shell}} \right\rangle * \frac{\left(x_W^{\text{bulk}} - x_W^{\text{shell}} \right)}{x_W^{\text{bulk}}} \tag{S16}$$

$$= \left\langle n_G^{\text{shell}} + n_W^{\text{shell}} \right\rangle * \left(1 - \frac{x_W^{\text{shell}}}{x_W^{\text{bulk}}} \right) \tag{S17}$$

$$= \left\langle n_G^{\text{shell}} + n_W^{\text{shell}} \right\rangle * \left(1 - \frac{n_W^{\text{shell}}}{n_W^{\text{bulk}}} \frac{n_W^{\text{bulk}} + n_G^{\text{bulk}}}{n_W^{\text{shell}} + n_G^{\text{shell}}} \right) \tag{S18}$$

$$= \left\langle n_G^{\text{shell}} + n_W^{\text{shell}} \right\rangle * \left(1 - \frac{K_W}{K_X} \right) \tag{S19}$$

$$= \left\langle n_G^{\text{shell}} + n_W^{\text{shell}} \right\rangle * (1 - K) \tag{S20}$$

K is the ratio between two shell-bulk partition coefficients: K_W for the exchange $W_{bulk} \rightleftarrows W_{shell}$, and K_X for the exchange $X_{bulk} \rightleftarrows X_{shell}$ where X is any solvent particle. We stress that we do not make the assumption of a 1:1 exchange. We present in Figure S9 the plots used to determine the enthalpic and entropic contributions to preferential solvation.

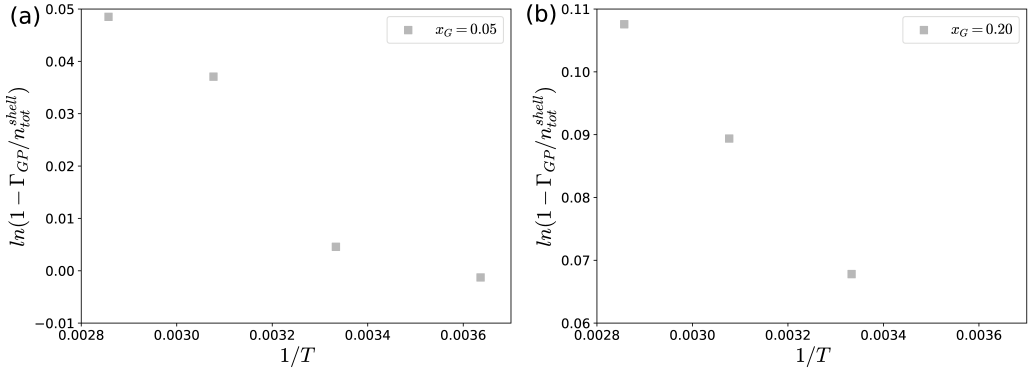

Figure S9: Plot of $1 - \frac{\Gamma_{GP}}{n_{tot}^{shell}}$ at two molar fractions.

References

(1) Baynes, B. M.; Trout, B. L. Proteins in Mixed Solvents: A Molecular-Level Perspective. *J. Phys. Chem. B* **2003**, *107*, 14058–14067.

(2) Lide, D. R. *CRC Handbook of Chemistry and Physics, 85th Edition*; CRC Press, Boca Raton, FL, 2004.