Connecting Longitudinal and Transverse Relaxation Rates in Live-Cell NMR

SARAH LEEB, FAN YANG, MIKAEL OLIVEBERG* AND JENS DANIELSSON*

Department of Biochemistry and Biophysics, Arrhenius Laboratories of Natural Sciences, Stockholm University, 106 91 Stockholm, Sweden.

* **Corresponding authors**: Mikael Oliveberg and Jens Danielsson
email: mikael@dbb.su.se, jens.danielsson@dbb.su.se

Supporting information

Supporting information and controls.

SI 1. Relaxation rates as function of τ_c and their conversion to functions of M_w

The NMR relaxation rates are tightly linked to molecular motions, where the motions on different geometrical dimensions are described by its corresponding correlation time. Typically, the overall correlation time, τ_c, is the harmonic mean of the global rotational correlation time, τ_r, the local motional correlation time, τ_{loc}, and chemical exchange contributions, τ_{ex}. The correlation time is, in turn, linked to the longitudinal (R_1) and transverse (R_2) relaxation rates for amide nitrogen atoms in the protein backbone via the motional spectral density functions, $J(\omega)$, by [1]:

$$F_{R_1}(\tau_c) = \frac{\pi}{5} b_{NH}^2 [J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H + \omega_N)] + \frac{4\pi}{15} c_N^2 J(\omega_N) \quad \text{(Eq. S1)}$$

$$F_{R_2}(\tau_c) = \frac{\pi}{10} b_{NH}^2 [4J(0) + J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H) + 6J(\omega_H + \omega_N)] + \frac{2\pi}{45} c_N^2 [4J(0) + 3J(\omega_N)] + R_{ex} \quad \text{(Eq. S2)}$$

where b_{NH} and c_N are constants of dipole-dipole interactions (DD) and chemical shielding anisotropy (CSA) respectively and $J(\omega)$ is the spectral density at frequency ω (in the case above the Larmor frequencies of ^1H and ^{15}N nuclei). The spectral density is a quantity obtained by Fourier transforming the auto-correlated time function of a particle's motion defined by the overall motional correlation time τ_c. The correlation time dependence of DD and CSA relaxation contributions is described by an auto-correlated spherical harmonic of rank 2 (l=2)[1] and the corresponding spectral density function becomes

$$J(\omega) = \frac{1}{2\pi} \frac{\tau_c}{1 + \omega^2 \tau_c^2} . \quad \text{(Eq. S3)}$$

The DD interaction strength constant b_{NH} is defined by the following expression:

$$b_{NH} = -\frac{\mu_0 \gamma_H \gamma_N \hbar}{4\pi r_{NH}^3} \quad \text{(Eq. S4)}$$

where μ_0 is the magnetic permeability in vacuum, γ_H and γ_N are the gyromagnetic ratios of ^1H and ^{15}N, $\hbar$ is the Planck constant divided by 2π and r_{NH} is the average bond length between an

amide nitrogen and its amide proton, which was set to 104 pm[1]. The CSA interaction strength constant c_N is given by:

$$c_N = \omega_N \Delta\sigma \quad \text{(Eq. S5)}$$

with ω_N being the Larmor frequency of nitrogen and $\Delta\sigma$ is the chemical shielding anisotropy, which here is set to -160 ppm[1].

In addition, the transverse relaxation rate R_2 is crucially sensitive to chemical exchange processes in the *μs* to *ms* time range[1-2], and relaxation rate contributions – if present – are typically represented by an additive term R_{ex}.

Provided that rotational correlation time (τ_r) dominates relaxation processes, a closed-form expression that estimates τ_r from R_1 and R_2 exists for large molecules (*e.g.* proteins) that rotate slowly and are thus within the NMR slow motion regime ($(\omega_0 \tau_c)^2 >> 1$)[3]:

$$\tau_r \approx \frac{1}{4\pi\nu_N}\sqrt{6\frac{R_2}{R_1} - 7} \quad \text{(Eq. S6)}$$

with ν_N being the Larmor frequency of a the ^{15}N nucleus given in Hz. The rotational correlation time τ_r, was calculated from the R_1 and R_2 values of both *in-vitro* glycerol references as well as in-cell data. The relaxation rates were then plotted against τ_r and while the predicted rates fitted well with the observed ones in the case of the *in-vitro* glycerol references, the in-cell rates showed substantial deviations from the theory (Fig. 2).

To assign relaxation rates to distributions of differently sized proteins, we estimated τ_r from the molecular weight. For this, we first calculated the radius of gyration (r_G) from the number of residues N in a protein [4]:

$$r_G(N) = 2.24\, N^{0.392} \quad \text{(Eq. S7)}$$

If the number of residues was not known, the M_w was divided by 110 Da, the average molecular weight of an amino acid. To further convert r_G to the hydrodynamic radius (r_H) we use the following empirical correlation [4b, 5]:

$$r_H = 1.45\, r_G \quad \text{(Eq. S8)}$$

Based on the assumption that folded proteins behave roughly like hard spheres, we then used r_H to get τ_c through the Stokes-Einstein-Debye relationship[6]

$$\tau_c = \frac{4\pi\eta r_H^3}{3k_B T} \quad \text{(Eq. S9)}$$

where k_B is the Boltzmann constant, T the temperature and η the viscosity of the solution.

SI 2. Conversion of R_i to apparent viscosity η^{app}

To translate in-cell relaxation rates to apparent viscosity, both R_1 and R_2 was measured for each of the three types of protein (TTHA^{pwt}, HAH1^{pwt} and SOD1^{barrel}) in solutions made increasingly viscous through the addition of deuterated glycerol (Table S1). The viscosity of the solutions was estimated by calculating τ_r from the obtained relaxation rates and using the Stokes-Einstein-Debye relationship (Eq. S9), while keeping r_H fixed for each protein. The relaxation rates were then plotted as a function of viscosity and the theoretically derived relaxation rate functions for each probe protein (Eq. S1-S2) overlaid (Fig. S4). Since relaxation parameters are so sensitive to the rotational tumbling rate and thus size, each of the three types of protein shows a slightly different dependence on viscosity (Fig. S4).

Using these reference curves, each in-cell relaxation parameter could be assigned a particular apparent viscosity (Fig. S4, Table 2) and in doing so, the retardation effects of cytosolic crowding on differently sized proteins were made comparable. Since protein size is not changing substantially by a point mutation, the same reference curves were used for the charge variants TTHA^{E32K}, HAH1^{K57E} and SOD1^{R100E}.

SI 3. The binding models

Model S1. In the first interaction model we assume that the probe proteins are in fast exchange with an encounter complex with a molecular species of a particular averaged mass M^{av}. R_i^{bound} ($i \in 1||2$), is a function of the combined mass of probe protein and interaction partner ($M_j + M^{av}$), and R_i^{free} is a function solely of the probe protein's mass, given by Eq. S1 and S2 (SI1). The population weighted observed relaxation rate is then given by:

$$R_i^{OBS} = \quad p_B R_i^{bound}(M_j + M^{av}) + (1 - p_B)R_i^{free}(M_j) \qquad \text{(Eq. S10)}$$

This results in a set of six equations, one for each probe protein. Since their mass M_j is known to us, the optimisation problem has in total two fitting parameters: the population of bound protein p_B, which is optimised locally for each probe protein and M^{av}, which is

optimised globally for the entire array of probe proteins. With M^{av} = 143 kDa, in-cell R_1 and R_2 can be closely reproduced with a *rmsd* of 0.26 (Fig. 3). The populations of bound species for each probe protein are listed in Table S2.

Model S2. The second binding model is conceptually similar to *model S1*, with the sole difference that M^{av} is not optimised, but fixed to the average molecular mass of the UniProt-derived size distribution for human cytosolic proteins. The average molecular mass $M^{av}_{database}$ (~73 kDa) is used as a proxy for a generic cytosolic binding partner, leaving an optimisation problem with only a single fitting parameter p_B for each probe protein.

$$R_i^{OBS} = \; p_B R_i^{bound}(M_j + M^{av}_{database}) + (1 - p_B)R_i^{free}(M_j) \qquad \text{(Eq. S11)}$$

The optimised populations of bound species as well as the correspondingly calculated relaxation rates are summarized in Table S2 and result in a much poorer agreement with the in-cell measurements (*rmsd* = 0.54).

Model S3. To expand the binding models to a more realistic scenario, in which the probe proteins are encountering a variety of interaction partners, we make use of a distribution of sizes instead of a single average mass. Two different types of distributions can be used for approximating this range of sizes: the lognormal and the Γ-distribution (see *model S4*)[7]. Both describe the cytosolic protein molecular weight distribution fairly well, however the lognormal resulted in a somewhat better fit, and henceforth we use the lognormal distribution. The fitted distributions $\rho_{database}(M_w)$ were then used to determine an expression for R_i^{bound} for each probe protein:

$$R_i^{bound}(M_j + M_w) = \frac{\int_0^\infty \rho_{database}(M_w) F_{R_i}(M_j + M_w) dM_w}{\int_0^\infty \rho_{database}(M_w) dM_w} \qquad \text{(Eq. S12)}$$

And similar as in *model S2*, only one parameter – p_B – is optimised for each probe protein:

$$R_{\mathrm{i}}^{\mathrm{calc}} = \; p_{\mathrm{B}} R_{\mathrm{i}}^{\mathrm{bound}}(M_{\mathrm{j}} + M_{\mathrm{w}}) + (1 - p_{\mathrm{B}}) R_{\mathrm{i}}^{\mathrm{free}}(M_{\mathrm{j}}) \qquad \text{(Eq. S13)}$$

The obtained populations of bound protein and relaxation rates are summarized in Table S2. The *rmsd* of 0.60 in the case of a fitted lognormal distribution is similar to *model S2*, however still far from the fairly accurate predictions of the much simpler *model S1*.

Model S4. In the fourth and last binding model, we use the same set-up of equations as in *model S3*. However, ρ(M_{w}) is not constrained to the data-base values, but is itself subjugated to optimization, *i.e.* the range of sizes of interaction partners and their abundance are varied until in-cell R_1 and R_2 measurements are optimally approximated. In the case of the lognormal distribution

$$\rho^{\mathrm{lognormal}}(M_{\mathrm{w}}) = \frac{1}{M_{\mathrm{w}}\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{(\ln(M_{\mathrm{w}})-\mu)^2}{2\sigma^2}} \qquad \text{(Eq. S14)}$$

this meant that both σ and μ were optimized globally for the whole array of probe proteins, while in the case of the Γ-distribution

$$\rho^{\Gamma}(M_{\mathrm{w}}) = \frac{1}{\Gamma(\mathrm{k})\theta^{\mathrm{k}}}\, M_{\mathrm{w}}^{\mathrm{k}-1}\, \mathrm{e}^{-\frac{M_{\mathrm{w}}}{\theta}} \qquad \text{(Eq. S15)}$$

the shape parameter k and the scale parameter θ had to be optimized. With these expressions for ρ(M_{w}) and thus $R_{\mathrm{i}}^{\mathrm{bound}}$ (Eq. S12-S13) the population of bound species p_{B} was simultaneously and locally optimised for each probe protein. The results are summarised in Table S2. Figure 3 shows the calculated reduced relaxation rates derived from the optimised lognormal distribution. The Γ-distribution gave almost identical results and both have a *rmsd* = 0.26. In summary, *model S4* is far superior than *model S3* in predicting in-cell relaxation rates and just as good as the simpler *model S1*, however it provides more information and describes a physiologically more relevant scenario.

SI 4. Validation of interaction model with lysozyme.

In order to validate the interaction model, both R_1 and R_2 of TTHA^{pwt} were measured in solutions with 50 mg/ml and 150 mg/ml lysozyme (Fig. S5). Despite lacking physiological relevance with regard to being a highly positively charged, secretory protein, we know from previous studies that it is forming charge-mediated transient complexes with net negatively charged probe proteins [8]. Several studies point to the formation of short-lived intracellular transient encounter complexes being of electrostatic nature, rendering lysozyme with known molecular weight (~16.5 kDa) an ideal candidate to test the interaction model. First, we confirmed the weak, transient interaction by fluorescence spectroscopy, where titration of TTHA^{pwt} onto lysozyme, revealed an apparent dissociation constant in the low millimolar regime, K_D = 5.0 mM ± 0.7 mM (Fig S5). Under conditions with 50 mg/ml lysozyme, using Eq. 1, both R_1 and R_2 can be well accounted for, if 25 % of TTHA^{pwt} is bound to monomeric lysozyme. Increasing the crowder concentration three-fold to 150 mg/ml leads to a population of 46 % bound to monomeric lysozyme. These populations correspond to $K_D \approx 10$ mM, in relatively good agreement with the fluorescence data.

Finally, we used the two pairs of relaxation data to determine the mass distribution of lysozyme, according to Eq. 3. At these concentrations, lysozyme occurs mainly as a monomer, and in accordance with this we find that a narrow distribution centered at 15.1 kDa fulfils the observed relaxation rates (Fig. S5).

SI 5. Surface net charge density and size distribution

To test the assumption, that the size composition of interaction partners is independent of net charge, we split the set of proteins, given in the database by Geiger et al.[9], into three size categories: proteins smaller than 70 kDa, those larger than 70 kDa but smaller than 140 kDa and those larger than 140 kDa. For these three intervals the net charge density was calculated and plotted in a histogram (Fig. S7). A Lorentzian fitted on top revealed the mean charge densities for each category to be at -0.04, -0.08 and -0.11 e/nm^2 respectively. Although not identical, we conclude that the overall pattern with net negatively charged densities is valid for proteins along the extension of the size distribution.

SI 6. The maximum entropy approach to derive a minimum information distribution

Solving for $\rho(M_w)$ in Eq. S12-S13 results in a multitude of solutions (Fig. S6). In all these solutions the calculated R_1 and R_2 values are identical for both distribution types (lognormal and Γ, *model S4*) even though the population of bound protein could vary in some cases up to 1 % (Table S2).

Due of the multitude of equivalent solutions, additional criteria to select possible representatives for the molecular weight distribution of the cytosolic proteome were introduced. One criterion is based on choosing the distribution that deviated the least from the data base derived fit to the cytosolic proteome. Another criterion is based on choosing the distribution whose information content is minimal by maximizing its information entropy S (Fig. 4, Fig S9). The information entropy of any distribution ρ is defined as [10]:

$$S = \int_{-\inf}^{\inf} \rho \log(\rho)\, dx \qquad \text{(Eq. S16)}$$

The resulting distributions are quite similar to the database-derived ones, however once again, they show a distinct tail in the high-mass region corresponding to *e.g.* larger intracellular molecular assemblies (Fig. S9).

SI 7. Validation of the self-consistency between in-cell R_1 and R_2 data.

Due to the short inter-scan delay of 1 s, a full recovery to equilibrium magnetisation is not achieved between scans of varying relaxation delay lengths. This may underlie the systematic deviations seen in the R_1 intensity attenuation curves (Fig. S1). Consequently, we wanted to see if re-fitting the in-cell R_1 data to only two of the three data points could still give reasonable results. Omitting the first data point at 10 ms relaxation delay led to overall prolonged longitudinal relaxation times and thus lower R_1 values. However, re-running the calculations according to interaction *model S4* with a lognormal distribution resulted in *rmsd* = 0.86 and $r^2 = 0.27$ (Fig. S10). This is an even worse match between prediction and measurement than that obtained from interaction *models S2* and *S3* with the original R_1 data set (Fig. 4). This example illustrates well that not just any set of relaxation rates can be made compatible with Eq. S12 and S13 by adjusting distribution shape parameters and p_B and

confirms that our in-cell R_1 values are globally – over the set of six reporter proteins – consistent with our in-cell R_2 measurements.

SI Methods and Materials

Protein mutagenesis, expression and purification

The plasmids carrying a carbenicillin resistance marker and the target gene were transformed into *E. coli* BL21(DE3) expression strains. For ^{15}N-isotope enriched protein production, minimal medium (0.02 M KH_2PO_4, 0.04 M Na_2HPO_4, 0.1 M NaCl, 2 mM $MgSO_4$, 0.4 % (w/v) glucose, M2 trace metal mix) – supplemented with carbenicillin and 0.1 % (w/v) $^{15}NH_4Cl$ – was inoculated and the cells grown at 37 °C while shaking until they reached an OD_{600} between 0.6 – 0.8. To induce over-expression of the target protein, 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) was added. After 4 more hours of incubation at 37 °C, the cells were harvested at 5000g for 10 min at 4 °C. The cell pellet was resuspended in 50 mM Tris-HCl buffer (pH 7.4) and stored at – 80 °C. Point mutations were introduced through site-directed mutagenesis. The sequence identity was verified by DNA sequencing after amplification in *E. coli* XL Blue cells.

For protein purification the cell pellet was thawed and the cells lysed through sonication. Cell debris was removed by spinning the sonicate at 39 000g for 30 min. The supernatant was then used for further purification usually including a heat denaturation step, an ammonium sulphate precipitation step, ion exchange chromatography and gel filtration. The purification protocol of $SOD1^{barrel}$ (and charge variants thereof) is described in detail in Danielsson *et al.* [11], while detailed protocols for $TTHA^{pwt}$ and $HAH1^{pwt}$ (and charge variants thereof) can be found in Mu *et al.* [12] and Leeb *et al.* [13]. During the whole purification procedure, the protein and buffer solutions were kept at 4 °C. Protein purity was eventually confirmed by SDS-PAGE using 4 - 20 % precast gels (BioRad, California, USA).

Protein Transfer into mammalian cells for in-cell NMR

Human ovary adenocarcinoma A2780 were grown in RPMI 1640 growth medium (Life Technologies, California, USA) supplemented with 10 % Fetal Bovine Serum (Life Technologies), 1 % Antibiotics-Antimycotics mixture (Life Technologies) and 0.45 μg/ml Plasmocin (InvivoGen, California, USA). When the cells reached 70 – 90 % confluence, they were detached from the plates by trypsin (Trypsin/EDTA in Dulbecco's phosphate buffered saline (DPBS), Life Technologies) and divided onto three fresh growth plates. The cells were normally trypsinated every other day and never on two consecutive days.

A detailed protocol regarding the transfer of ^{15}N-isotope enriched protein into mammalian cells through electroporation (NEPA21 Super Electroporator, Nepa Gene Co., Ichikawa, Japan) can be found in Leeb *et al.* [13]. In short, the cells were prepared by first washing them twice with DPBS and then harvesting them with trypsin. The trypsin was deactivated by adding RPMI growth medium and the cell suspension subsequently washed twice with OptiMem (Life Technologies). For each washing step the cells were gently centrifuged at 200×g for 5 min and the supernatant was discarded. The cells were counted using Trypan Blue staining. About 60 x 10^6 cells were re-suspended in OptiMem supplemented with protein in DPBS to a final protein concentration equalling 1.5 mM. 2 ml were then evenly distributed to 20 electroporation cuvettes (3×10^6 cells per cuvette). For the electroporation, poring pulse lengths between 14 to 16 ms at 115 V were used followed by a series of five 50 ms long transfer pulses at 20 V interceded by 50 ms delays. With these settings the total energy applied to the cell solution was typically between 4 – 4.5 J. After electroporation the cells were washed once more with OptiMem and finally plated in RPMI growth medium. After 5 hours recovery, dead cells were removed by washing the plates twice with DPBS. The surviving, re-attached cells were then harvested with trypsination. Trypsin was once more deactivated with RPMI growth medium and the cells washed twice with OptiMem. The pelleted cell slurry was resuspended in about 400 μl OptiMem supplemented with 10 % D_2O and transferred to a 4 mm flat-bottomed NMR tube (BMS-004B, Shigemi Inc., Tokyo, Japan) for subsequent data acquisition.

In-cell relaxation measurements

All NMR data was acquired using a Bruker Avance III 700 MHz spectrometer with a cryogenically cooled triple-resonance probe, at 37 °C using an 'interleaved' acquisition method. This form of data acquisition averages any time-dependent processes *i.e.* cell packing events that would temporarily change the protein concentration in the detection volume and therefore distort the relaxation decay curve [13]. Both R_2 and R_1 measurements were carried out using one-dimensional ^{15}N-filtered heteronuclear single quantum coherence (HSQC)-based pulse sequences with three relaxation delays each, ranging between 0 and 68 ms in case of R_2 and 10 and 500 ms in the case of R_1. With an inter-scan delay of 1 s, an acquisition time of 125 ms and a total amount of 5120 scans, R_2 experiments lasted for ~ 5 hours. R_1 experiments with an inter-scan delay of 1 s, an acquisition time of 104 ms and a total amount of 4200 scans took approximately the same amount of time to run (~ 5 hours).

To be able to survey sample changes during the hours-long relaxation data acquisition and to quantify potentially leaked reporter protein, 13 min short one-dimensional ^{15}N-filtered ^{1}H-band-selective optimized flip-angle short-transient heteronuclear multiple quantum coherence (SOFAST-HMQC) spectra were recorded immediately before and after the relaxation experiments adding ~ 26 min to the total measurement time.

Leakage of reporter protein into the interstitial fluid surrounding the cells was quantified by carefully removing the cell slurry from the NMR tube. After spinning 5 min at 200 g, the supernatant was transferred to a fresh NMR tube and a 1D ^{1}H-SOFAST-HMQC was recorded. Overlays of in-cell and supernatant spectra are shown in Figure S2. Integrating over the same spectral regions in both samples and then calculating the ratio after correcting for dilution in the supernatant samples showed that protein leakage typically didn't surpass 10 %. In addition, there is evidence that most of the leakage is introduced during sample preparation of the supernatant [12]. The supernatant sample of TTHA^{E32K}, for instance, sticks out as having considerably more leakage than the other samples (Fig. S2), yet the in-cell longitudinal relaxation rate obtained from that sample falls well in line with our expectations. Due to technical issues with our equipment, this particular sample was prepared at 2 times the g-force normally applied, putatively inducing post-experimental leakage. This underlines that gentle handling during sample preparation is paramount for keeping the cells' structural integrity intact.

Relaxation measurements of *in-vitro* glycerol series

Samples contained: 200 µM protein, 10 mM MES pH 6.5, 10 % (v/v) D_2O and increasing amounts of deuterated glycerol-d8 (98 % D) from 0 – 50 % (v/v). Both transverse and longitudinal relaxation rates were measured with the exact same acquisition parameters that were used for the in-cell equivalent with the sole difference that spectra for a total of 6 different relaxation delays (0, 34, 51, 68, 85 and 102 ms) in the case of R_2 and 10 different relaxation delays (10, 100, 200, 400, 600, 800, 1000, 1200, 1600 and 2000 ms) in the case of R_1 were recorded. In addition, the number of scans was reduced to 64, due to the substantially higher protein concentration.

Determination of relaxation rates

The NMR raw data was processed and phase corrected with TopSpin 4.0.6 (Bruker, Massachusetts, USA). The data was further processed with in-house MATLAB (MathWorks, MA, USA) scripts, applying linear base line correction and integration over a particular spectral portion. For $HAH1^{pwt}$, $TTHA^{pwt}$ and their charge variants spectral regions between 8.6 and 9.4 ppm were used, while for $SOD1^{barrel}$ and its charge variant the region between 8.9 and 9.7 ppm was chosen. Using the most downfield-shifted spectral region of protein signal ensured that mainly signals stemming from the ^{15}N-enriched reporter proteins were used for data analysis. Due to the large amount of intracellular protein in the in-cell NMR samples, naturally abundant ^{15}N shows small amounts of signal in the more central regions of the typical amide proton signal range, which may distort the relaxation decay curves. For details, see Leeb *et. al.* [13].

The integrals of the spectral portions were normalized and fitted to a single exponential decay. The relaxation rate errors were determined by using the standard deviation of the distribution of decay rates obtained through repeatedly fitting decay curves (40 000 - 60 000 times) to data points that were randomly varied within their error region (Fig. S1). The error region of each data point, in turn, was defined by the signal-to-noise ratio of the corresponding spectrum. Naturally, the individual data points' errors increase with relaxation delay times as the signal becomes more attenuated.

Curating and analysing the cytosolic proteome database

To estimate the proteomic composition of the mammalian cytosol, we retrieved a list of 5217 human cytosolic proteins from the UniProt database [14]. The charge density of each protein was estimated by counting and assigning the acidic residues glutamate and aspartate with a negative charge and the basic residues lysine and arginine with a positive charge. Histidine residues were assumed to be neutral at physiological conditions, an assumption in line with net charge calculations performed with the PROPKA3 software [15]. The determined net charge for each protein was then normalized with its surface area, A, to obtain the charge density. The surface area was estimated from the radius of gyration, r_g determined by Eq. S7 (SI1):

$$A_{\text{protein surface}} = 4\pi r_g^2 \qquad \text{(Eq. S16)}$$

Furthermore, to get an impression of how much this size distribution changes (Fig. S6), if the relative abundance of individual proteins is taken into account the published database by Geiger et al.[9], where lysate proteins of eleven different human cancer cell lines were identified and quantified with mass spectrometry, was analysed. The database contains among other things the identity of each protein in form of a UniProt-ID, their molecular weight and their abundance. For our purposes, we used the abundance-weighted average of all eleven cell lines. Of the 11 731 proteins in the list, about 4.2 % were given UniProt-IDs that had either been removed or altered without providing a replacement ID. As a result, these sequences were omitted in our analysis.

Supporting References

1. Kowalewski, J.; Mäler, L., *Nuclear Spin Relaxation in Liquids: Theory, Experiments, and Applications, Second Edition*. CRC Press: 2017.

2. Kleckner, I. R.; Foster, M. P., An introduction to NMR-based approaches for measuring protein dynamics. *Biochim Biophys Acta* **2011,** *1814* (8), 942-68.

3. Kay, L. E.; Torchia, D. A.; Bax, A., Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **1989,** *28* (23), 8972-9.

4. (a) Hong, L.; Lei, J., Scaling law for the radius of gyration of proteins and its dependence on hydrophobicity. *Journal of Polymer Science Part B: Polymer Physics* **2009,**

47 (2), 207-214; (b) Dill, K. A.; Ghosh, K.; Schmit, J. D., Physical limits of cells and proteomes. *Proc Natl Acad Sci U S A* **2011,** *108* (44), 17876-82.

5. Tyn, M. T.; Gusek, T. W., Prediction of diffusion coefficients of proteins. *Biotechnol Bioeng* **1990,** *35* (4), 327-38.

6. Einstein, A., Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik* **1905,** *322* (8), 549-560.

7. Tiessen, A.; Perez-Rodriguez, P.; Delaye-Arredondo, L. J., Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res Notes* **2012,** *5*, 85.

8. Danielsson, J.; Mu, X.; Lang, L.; Wang, H.; Binolfi, A.; Theillet, F. X.; Bekei, B.; Logan, D. T.; Selenko, P.; Wennerstrom, H.; Oliveberg, M., Thermodynamics of protein destabilization in live cells. *Proc Natl Acad Sci U S A* **2015,** *112* (40), 12402-7.

9. Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M., Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* **2012,** *11* (3), M111 014050.

10. Shannon, C. E., A mathematical theory of communication. *Bell system technical journal* **1948,** *27* (3), 379-423.

11. Danielsson, J.; Kurnik, M.; Lang, L.; Oliveberg, M., Cutting off functional loops from homodimeric enzyme superoxide dismutase 1 (SOD1) leaves monomeric beta-barrels. *J Biol Chem* **2011,** *286* (38), 33070-83.

12. Mu, X.; Choi, S.; Lang, L.; Mowray, D.; Dokholyan, N. V.; Danielsson, J.; Oliveberg, M., Physicochemical code for quinary protein interactions in Escherichia coli. *Proc Natl Acad Sci U S A* **2017,** *114* (23), E4556-E4563.

13. Leeb, S.; Sörensen, T.; Yang, F.; Mu, X.; Oliveberg, M.; Danielsson, J., Diffusive protein interactions in human versus bacterial cells. *Current Research in Structural Biology* **2020**.

14. Consortium, T. U., UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **2018,** *47* (D1), D506-D515.

15. Olsson, M. H.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H., PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J Chem Theory Comput* **2011,** *7* (2), 525-37.

Supporting Figures

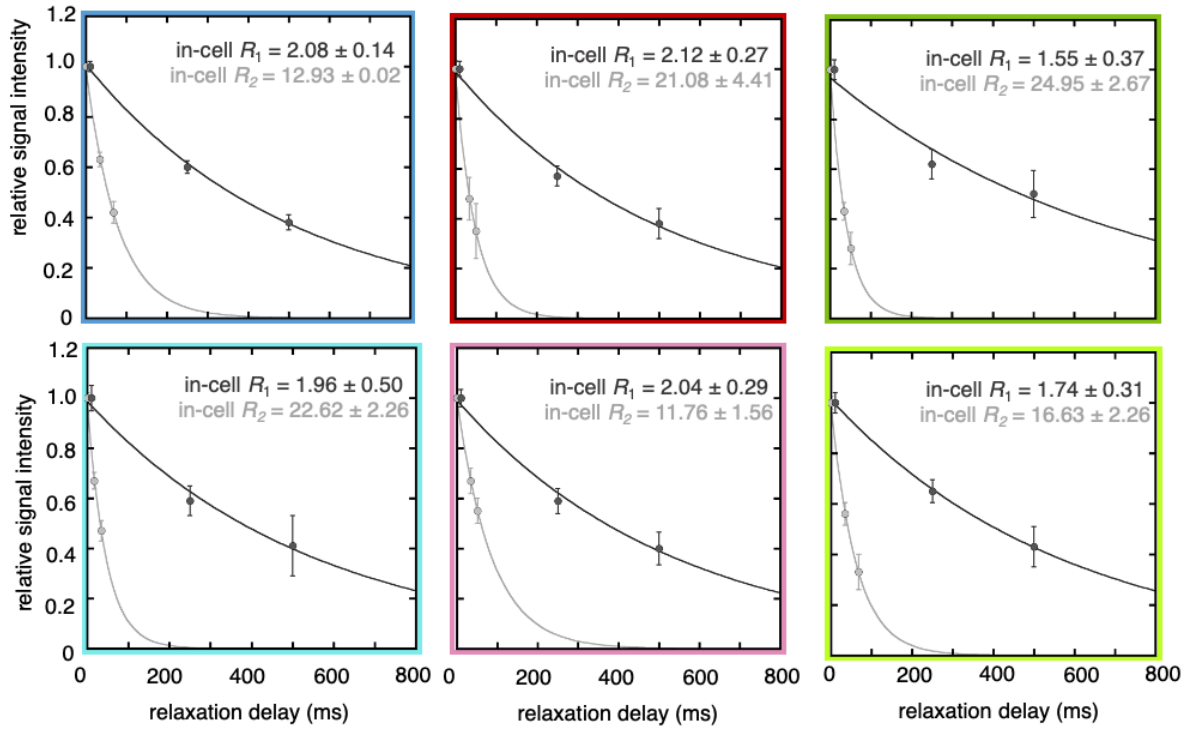

Figure S1: *In-cell transverse (R_2) and longitudinal (R_1) relaxation for the six reporter protein variants.* The signal intensity attenuation obtained from the R_1 (dark grey) and R_2 (light grey) experiments are shown for each variant. The fitted single exponential decays are shown as lines. The color code for the three types of protein are $TTHA^{pwt}$ (blue), $HAH1^{pwt}$ (red) and $SOD1^{barrel}$ (green). Their surface charge variants ($TTHA^{E32K}$, $HAH1^{K57E}$ and $SOD1^{R100E}$) are depicted in a lighter version of the corresponding color. The error bars for the individual data points represents the signal-to-noise ratio for that particular spectrum.

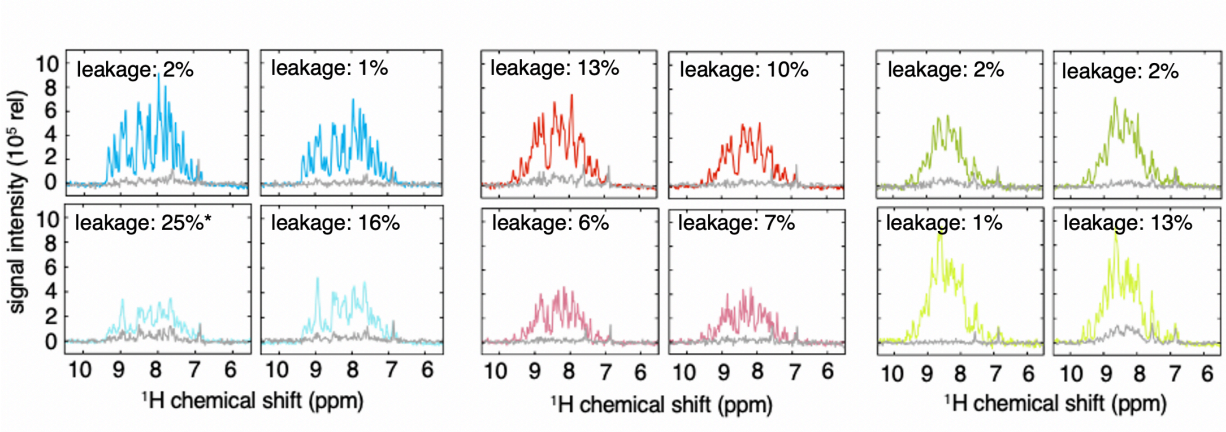

Figure S2: *Protein leakage quantification*. Overlay of amide proton spectral region from in-cell and supernatant samples. ^{15}N-filtered ^{1}H-spectra for in-cell longitudinal (left) and in-cell transverse (right) relaxation rates. The colour code for the three types of protein is TTHA^{pwt} (blue), HAH1^{pwt} (red), and SOD1^{barrel} (green) is. Their surface charge variants (TTHA^{E32K}, HAH1^{K57E} and SOD1^{R100E}) are depicted in a lighter version of the same colour. The grey spectra are the supernatant spectra at each condition, indicating leaked protein, and the intensity fraction is given in each panel. The supernatant sample of TTHA^{E32K} R_1 (marked by *), for instance, sticks out as having considerably more leakage than the other samples, yet the in-cell longitudinal relaxation rate obtained from that sample falls well in line with our expectations. This particular sample was prepared at 2 times the g-force normally applied, possibly inducing post-experimental leakage. This underlines that gentle handling during sample preparation is paramount for keeping the cells' structural integrity intact.

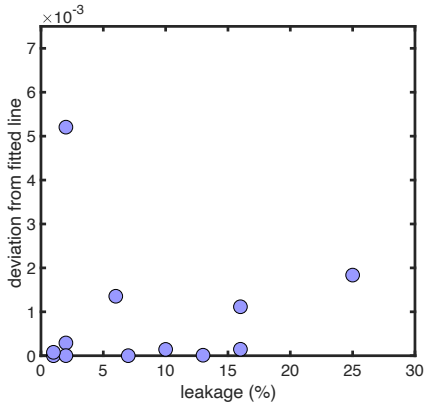

Figure S3. *Correlation between leakage and deviation from single exponentiality.* The magnitude of deviation from mono-exponentiality was quantified by the residual square sum (RSS) and the leakage was determined as described in material and methods, shown in Figure S2. We find no systematic correlation between leakage and RSS in this data set.

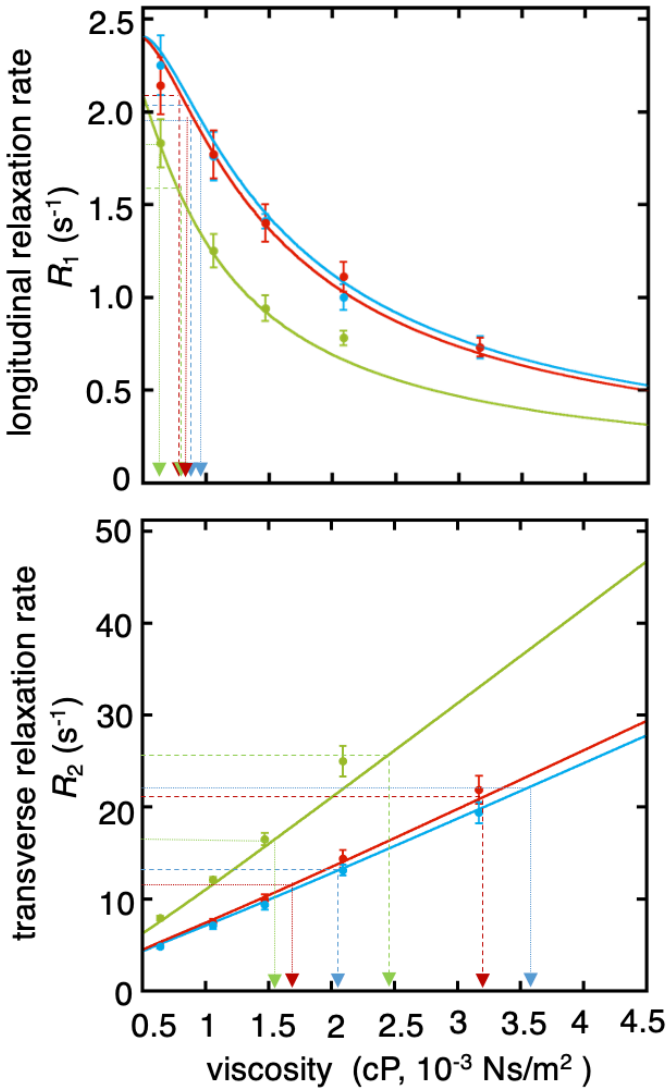

Figure S4: *Apparent viscosity reference curves for longitudinal (top) and transverse (bottom) relaxation.* The differently colored curves show how R_i changes with increasing viscosity for TTHA^{pwt} (blue), HAH1^{pwt} (red) and SOD1^{barrel} (green). The dashed and dotted lines show graphically how in-cell parameters are converted into apparent viscosity. Dashed lines trace the in-cell relaxation parameters of TTHA^{pwt}, HAH1^{pwt} and SOD1^{barrel}, while dotted lines trace their surface charge mutants TTHA^{E32K}, HAH1^{K57E} and SOD1^{R100E}.

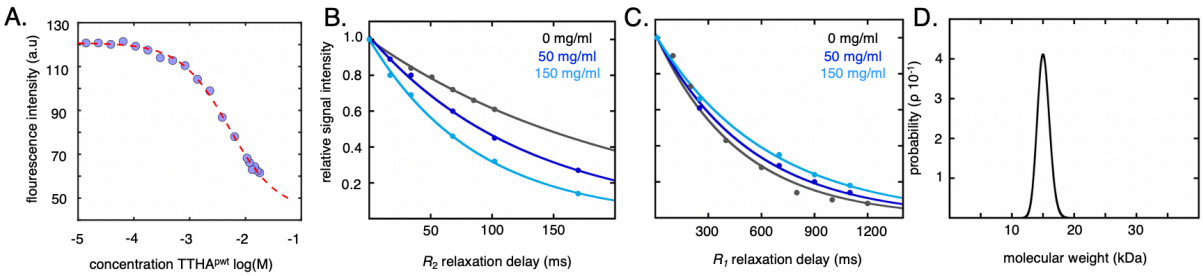

Figure S5. Benchmarking the method using weak interaction between TTHA^{pwt} and human lysozyme. The net negative TTHA^{pwt} interacts weakly with the positive lysozyme. A. Fluorescence signal intensity of lysozyme is reduced upon addition of TTHA^{pwt} in large excess. The apparent dissociation constant, corresponding to the red curve, is 5.0 mM ± 0.5 mM. B. Transverse relaxation rate R_2, determined at different concentrations of lysozyme, showing a marked increase in R_2, indicating transient increase in apparent size. C. The corresponding data for longitudinal relaxation, R_1, shows a slight decrease in relaxation rate, in full agreement with transient binding of the reporter protein TTHA^{pwt} to lysozyme. D. The determined distribution of masses using Eq 3. and including R_1 and R_2 from both lysozyme concentrations shows a narrow distribution centered at 15.1 kDa, close to the expected 16.5 kDa.

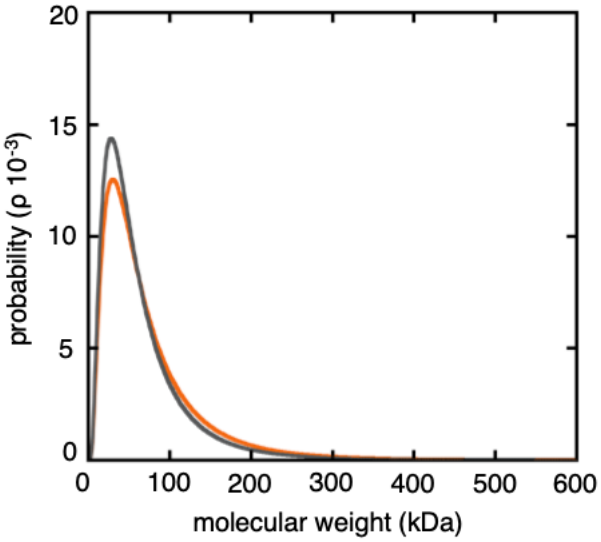

Figure S6: *Comparison of the cytosolic proteome size-distributions*. The fitted lognormal distributions to the molecular weight histogram of the UniProt-derived list of human cytosolic proteins (orange) and of abundance-weighted lysate proteome based on the data base by Geiger et al.[9] (SI *Methods and Materials*) (grey) differ only marginally. The abundance-weighted size distribution shows a larger frequency of lower-molecular weight proteins, however the peak maxima are almost identical in both distributions. We therefore conclude that using the simple UniProt-derived list of cytosolic proteins is valid as a first approximation to the protein interactome of a mammalian cell.

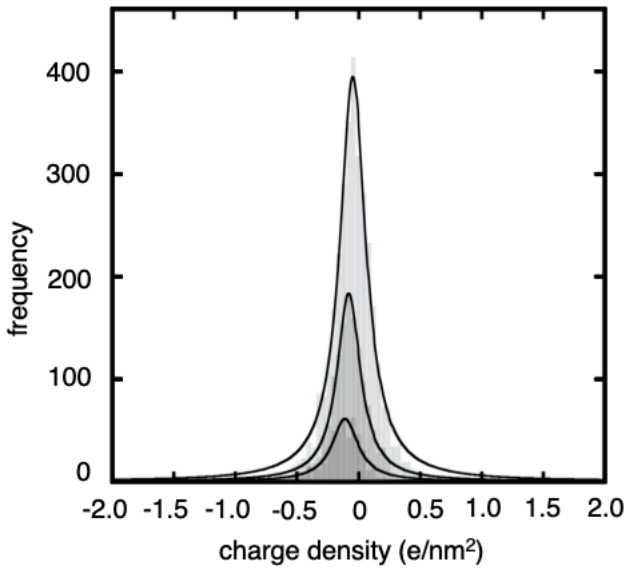

Figure S7: *Histogram of surface charge density for proteins of three different size ranges.* The first size range encompasses proteins smaller than 70 kDa (light grey), the second, proteins between 70 and 140 kDa (medium grey) and the third, proteins larger than 140 kDa (dark grey). All three size categories show symmetric and similarly distributed charge densities. A Lorentzian-shaped fit to the histograms results in central peak positions at -0.04, -0.08 and -0.11 e/nm^2 respectively.

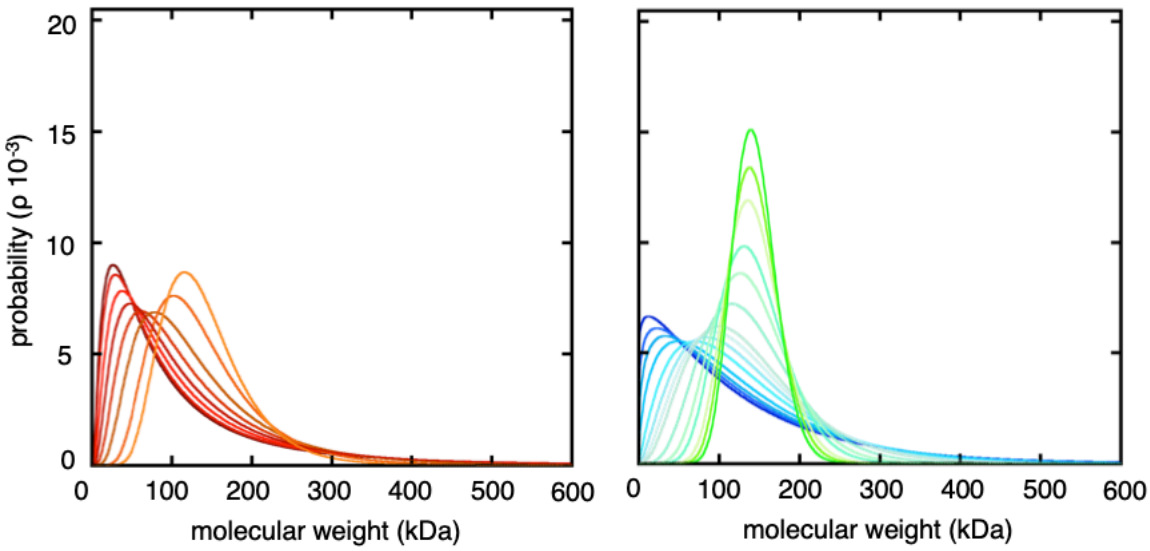

Figure S8: *Family of solutions for possible size distributions of the cytosolic interactome.* Multiple distributions are giving identical or similar results, when back-calculating in-cell R_1 and R_2 with Eq. 1 and 2. In the left panel a subset of solutions for the lognormal distribution and in the right panel a subset of solutions for the Γ-distribution are shown. All depicted distributions resulted in identical R_1 and R_2 values for both the lognormal and the Γ-distributions – albeit the population of bound protein could vary by up to 1 % (Table S2). The solutions had an *rmsd* = 0.26 when calculated reduced relaxation rates were compared with measured reduced in-cell relaxation rates (Fig. 3). Interestingly, the peak maxima are wandering slowly towards the average molecular weight of the interaction partner in *model S1* (~143 kDa), while the distributions are becoming more and more symmetric around this peak maximum. This is well in agreement with the results of *model S1*, where a single, average molecular weight was able to unify the two in-cell relaxation parameters as long as the amount of bound protein was kept free to vary for each protein.

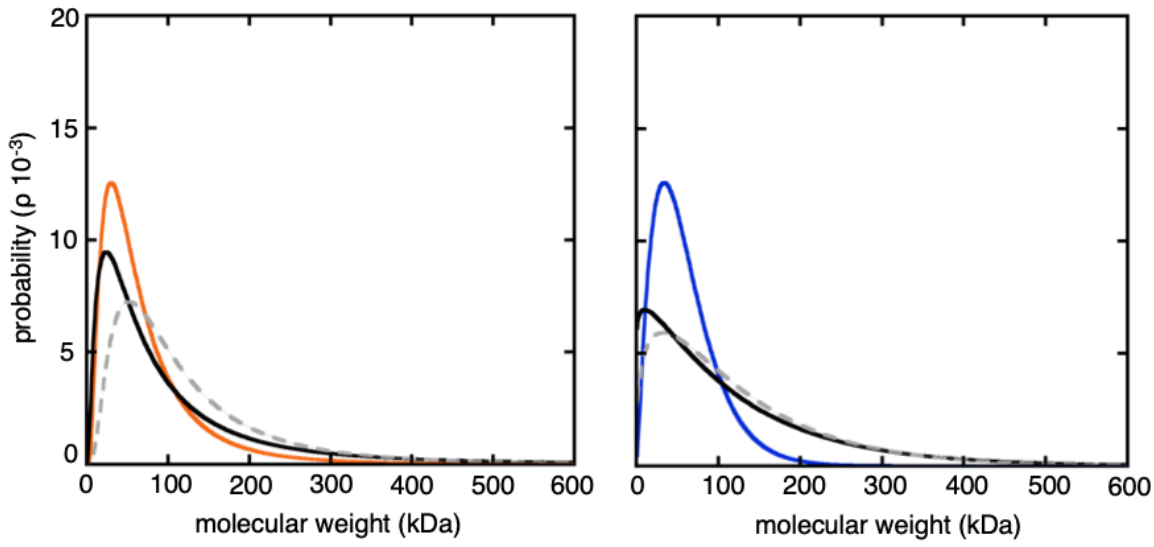

Figure S9: *Selected representatives for the cytosolic interactome.* Lognormal (left) and Γ-distributions (right) were selected based on two criteria: (i) those that showed minimal deviation from the fit to the database-derived set (black) and (ii) those whose information entropy was maximal (dashed grey) (SI6). The fitted curves to the database-derived cytosolic protein sizes are shown in orange and blue for the respective type of distribution.

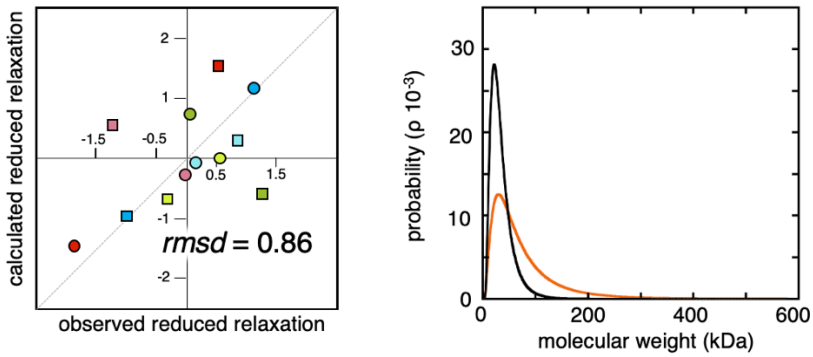

Figure S10: *Relaxation rate parameters need to be globally consistent to give reasonable results.* The left panel shows a correlation plot between observed and calculated reduced relaxation rates, for a data set where R_1 values were taken from fits that omitted the first data point (SI7). As a result, R_1 was in all 6 cases underestimated (SI7) which lead to a low agreement in the correlation plot with *rmsd* = 0.86 and r^2 = 0.27. The right panel depicts the corresponding optimised size-distribution that is closest to the one obtained from the UniProt list of human cytosolic proteins (orange). Not only is it very different from the solutions obtained with the original R_1 data set (Fig. 4, Fig. S7), it is also not plausible to assume that abundant proteins between 100 and 200 kDa (Fig. S4) would not be part of the interactome. This shows reassuringly, that despite only using three different relaxation delay times, estimations for in-cell R_1 and R_2 are consistent for all 6 proteins and lead to reasonable solutions regarding interactome size-distributions.

Supporting Tables

Table S1. Longitudinal and transverse relaxation rates measured for the three probe proteins HAH1^{pwt}, TTHA^{pwt} and SOD1^{barrel}.

% (v/v) glycerol	TTHA^{pwt}		HAH1^{pwt}		SOD1^{barrel}	
	R_1	R_2^*	R_1	R_2^*	R_1	R_2^*
0	2.18 ± 0.22	4.84 ± 0.26	2.06 ± 0.16	4.88 ± 0.31	1.83 ± 0.13	7.91 ± 0.26
20	1.76 ± 0.13	7.13 ± 0.42	1.77 ± 0.13	7.42 ± 0.42	1.25 ± 0.09	12.1 ± 0.28
30	1.41 ± 0.04	9.37 ± 0.55	1.40 ± 0.10	9.92 ± 0.59	0.94 ± 0.07	16.51 ± 0.66
40	1.00 ± 0.07	13.14 ± 0.58	1.11 ± 0.08	14.38 ± 0.95	0.78 ± 0.04	25.00 ± 1.69
50	0.73 ± 0.06	19.4 ± 1.14	0.73 ± 0.05	21.87 ± 1.53	-	-

* data taken from Leeb et al.[13]

Table S2. Collected results for the calculations of both relaxation rates R_1 and R_2 and the population of bound protein p_B for all six protein variants after optimization of different parameters specified in the various binding models S1-S4. The relaxation rates are converted back from reduced to ordinary rates for better comparability with in-cell measurements.

Model		TTHA^{pwt}	TTHA^{E32K}	HAH1^{pwt}	HAH1^{K57E}	SOD1^{barrel}	SOD1^{R100E}
S1	R_1 (s^{-1})	2.15	2.01	2.02	2.13	1.54	1.64
	R_2 (s^{-1})	13.54	23.07	20.25	12.49	24.90	16.00
	$p_B$$^{\#}$ (%)	5.4 (-0.5, +0.3)	11.8 (-1.9, +1.8)	9.8 (-2.5, +3.0)	4.6 (-1.4, +1.1)	11.1 (-2.2, +1.6)	5.2 (-1.6, +1.7)
S2	R_1 (s^{-1})	2.05	1.82	1.89	2.04	1.39	1.58
	R_2 (s^{-1})	12.48	20.16	17.05	11.80	22.96	14.58
	$p_B$$^{\#}$ (%)	10.4 (-1.0, +1.6)	21.7 (-3.6, +6.5)	16.8 (-3.2, +8.0)	9.1 (-3.2, +3.6)	21.4 (-4.6, +5.1)	9.4 (-3.4, +4.9)
S3a* $\sigma^{**} = 0.78$ $\mu^{**} = 10.92$	R_1 (s^{-1})	2.08	1.86	1.91	2.06	1.43	1.59
	R_2 (s^{-1})	12.90	21.20	18.08	12.10	23.69	15.06
	$p_B$$^{\#}$ (%)	9.9 (-0.9, +1.2)	21.0 (-3.6, +4.9)	16.5 (-2.4, +7.0)	8.5 (-2.8, +3.0)	20.3 (-4.3, +4.1)	9.1 (-2.9, +4.2)
S3b* $k^{**} = 2.31$ $\theta^{**} = 2.60E4$	R_1 (s^{-1})	2.04	1.80	1.87	2.03	1.38	1.57
	R_2 (s^{-1})	12.24	19.56	16.48	11.61	22.57	14.34
	$p_B$$^{\#}$ (%)	12.0 (-1.2, +2.0)	24.8 (-4.3, +8.3)	18.9 (-3.2, +10.1)	10.4 (-3.4, +4.5)	24.8 (-5.3, +6.3)	10.8 (-4.1, +6.2)
S4a* $\sigma = 1.03$ $\mu = 11.16$	R_1 (s^{-1})	2.15	2.01	2.02	2.13	1.54	1.64
	R_2 (s^{-1})	13.54	23.07	20.24	12.49	24.90	16.01
	$p_B$$^{\#}$ (%)	6.0 (-0.6, +0.3)	13.1 (-2.2, +1.9)	10.8 (-2.8, +3.4)	5.1 (-1.6, +1.2)	12.3 (-2.4, +1.8)	5.8 (-1.7, +1.9)
S4b* $k = 1.09$ $\theta = 1.11E5$	R_1 (s^{-1})	2.15	2.01	2.02	2.13	1.54	1.64
	R_2 (s^{-1})	13.54	23.07	20.24	12.49	24.90	16.01
	$p_B$$^{\#}$ (%)	6.1 (-0.6, +0.3)	13.2 (-2.1, +2.0)	11.0 (-2.8, +3.4)	5.2 (-1.6, +1.2)	12.5 (-2.5, +1.8)	5.9 (-1.8, +1.9)

* (a) denotes results based on lognormal distributions, (b) denotes results based on Γ-distributions

** these parameters were obtained from curve fitting to the histogram in Figure 2B (see *model S3*)

$^{\#}$ errors estimated by re-optimising p_B 10 000 times for normally distributed R_1 and R_2 using their respective errors as standard deviation. Upper and lower limits of p_B errors correspond to the quantile at 16% and 84% of the obtained skewed distribution.