Supplemental Material: A Direct Approach to Estimate the Anisotropy of Protein Structures from SAXS

Biel Roig-Solvas, Dana H. Brooks and Lee Makowski

1 Effect of noise on the ellipsoid estimation

To quantify the effect of noise on the semiaxes estimated by our approach, we carried out a noise study, where additive Gaussian noise was added at increasing standard deviations to the PDB-generated curves, previously normalized to $I(0) = 0$. The proposed algorithm was applied to these noisy curves, using polynomial orders of both 4 and 6, and repeated over 100 noise realizations for each noise level. For each noise level, the mean and standard. deviation over the 100 realizations of the estimated semiaxes values were computed. These results are shown in the following figure, where for each PDB file and each candidate (oblate or prolate), the mean semiaxis estimates are shown as a function of noise standard deviation, together with a shaded region representing +/- 1 standard deviation:

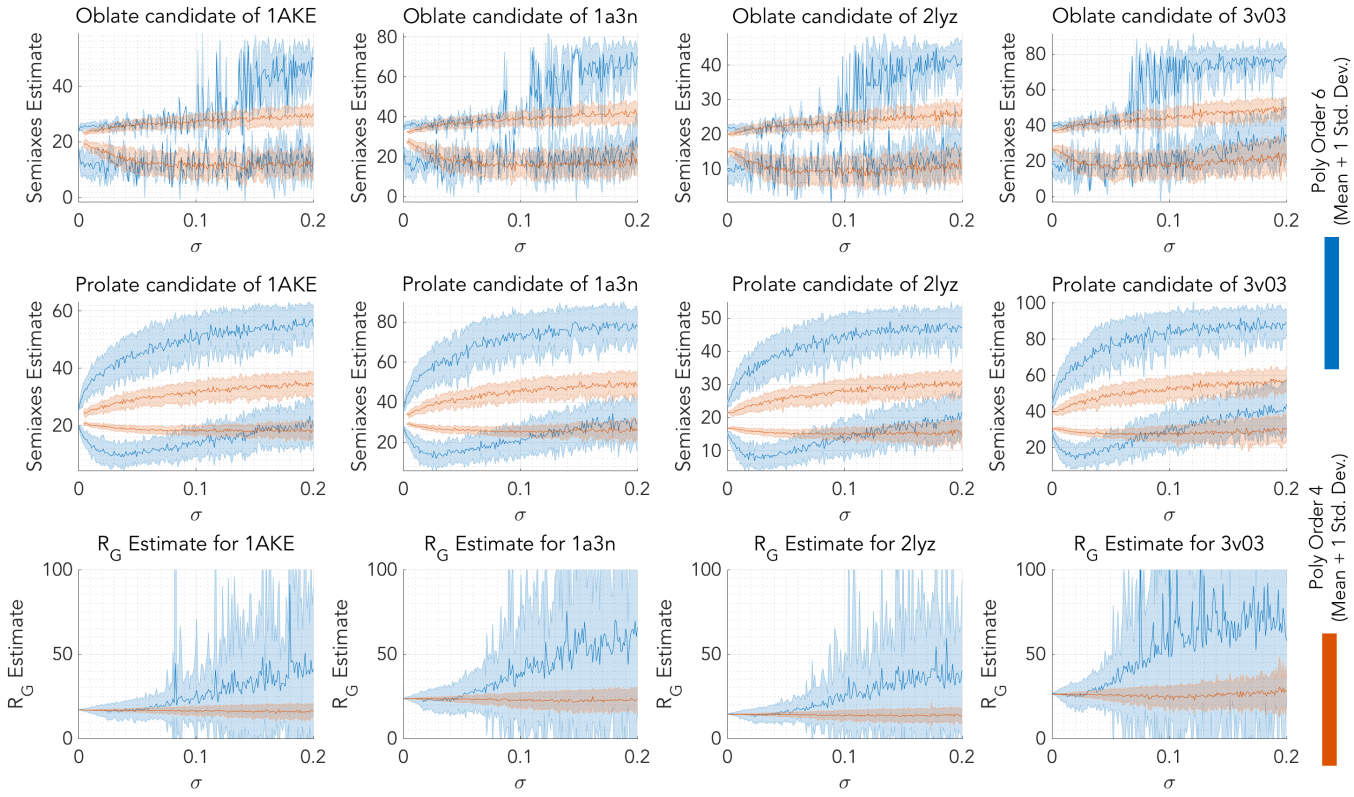

Figure S1: Effect of noise on the proposed algorithm. Each row shows the mean and standard deviation of the oblate axial estimates (first column), prolate axial estimates (second column) and R_G (last column), as a function of the noise standard deviation, for each PDB case considered in the manuscript.

For the semiaxes estimated using the 4th order fit (orange curves), one can observe a clear correlation with the noise level: as noise increases, the estimated anisotropy seems to increase (semiaxes means separate from each other). The R_G estimate, however, stays close to constant and only its standard deviation seems affected by the noise increase. For the semiaxes estimated using the 6th order fit (blue curves), we see again a an increase in apparent anisotropy, both as a function of noise and with respect to the 4th order estimates, at least at low noise levels, when the R_G estimate remains close to constant. These high order estimates also present a bias to overestimate R_G as a function of noise level σ.

We would however like to point out that the impact of noise on the estimated anisotropy is a general feature of the nature of the data and not a reflection of any particular algorithm. Added noise will inevitably increase the apparent

anisotropy of the scattering object. Given that $p(r)$ can be thought of as essentially the spectrum of the data - and D_{max} a measure of the bandwidth of the data - any addition of noise will increase the apparent bandwidth (and apparent D_{max}). For a fixed R_G (R_G is highly robust to added noise, as shown in Figure S1) the result is an increase in the apparent anisotropy.

2 Effect of $q\,R_G$ ranges on the estimation

To validate the choice of the $q\,R_G$ fitting range, we performed a series of Monte Carlo simulations analyzing the performance of the algorithm as a function of $[q\,R_G]_{min}$ and $[q\,R_G]_{max}$ across noise levels and across the 4 PDB-generated curves used in the previous section. The range $q\,R_G$ was sampled from 0 to 4 in steps of 0.05, while the noise standard deviation was sampled in 5 uniform steps from 0.005 to 0.02. For each possible combination of $[q\,R_G]_{min}$ and $[q\,R_G]_{max}$ within the $q\,R_G$ range, the proposed algorithm was used to estimate the ellipsoid parameters of the 4 aforementioned $I(q)$ curves, repeating that estimation 20 times for different noise realizations at each noise level, and across all noise levels, leading to 400 estimations per each $q\,R_G$ and over a million estimations total.

For each estimation, we recorded if the algorithm provided a valid result, where an experiment was considered to yield valid information if the recovered parameters where real numbers, and invalid if the estimated parameters were complex numbers. Additionally, for each valid result, we analyzed how much the estimated semiaxes differed from the correct semiaxes, where we took the correct semiaxes to be the ones estimated by BODIES using the noiseless curves. Specifically, we measured the norm of the difference between each experiment and the BODIES result, divided over the norm of the BODIES result, yielding a relative measure of the estimation error (as opposed to an absolute one). The results of these experiments can be seen in Figure S2:

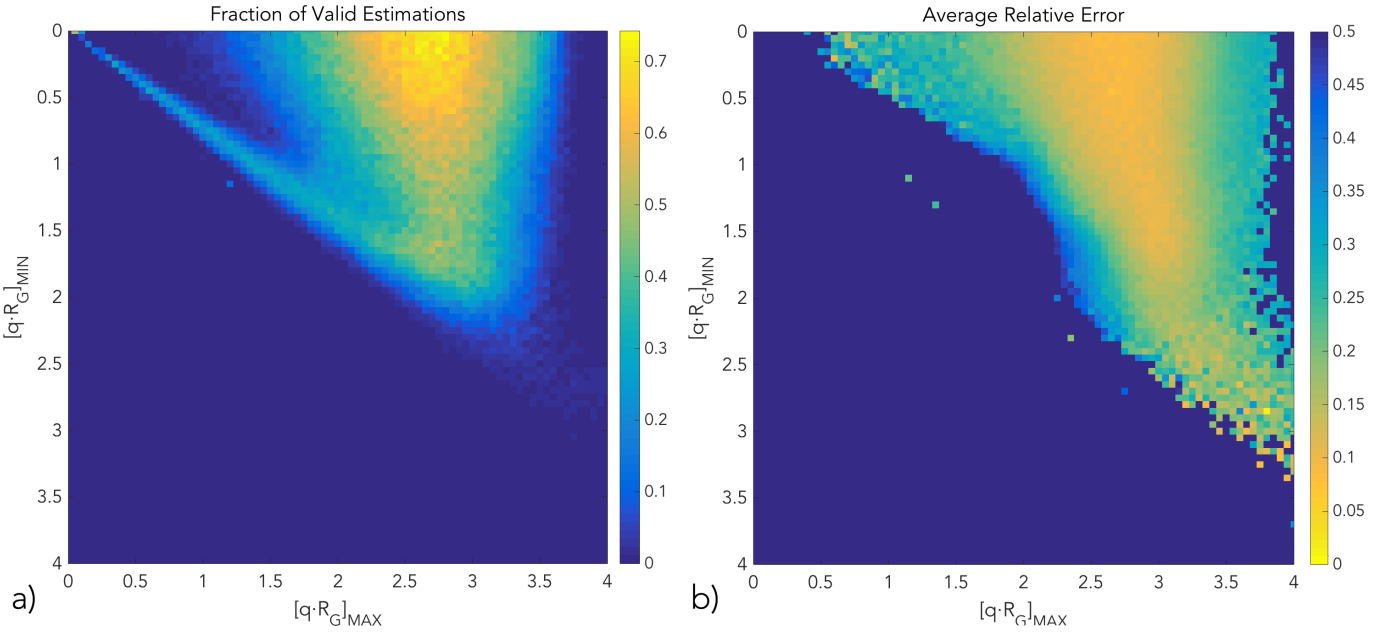

Figure S2: a) Fraction of valid estimations as a function of $[q\,R_G]_{min}$ and $[q\,R_G]_{max}$. b) Average relative error over valid estimations as a function of $[q\,R_G]_{min}$ and $[q\,R_G]_{max}$

As can be observed in Figure S2.a, ranges in the region $[0,1] \leq q\,R_G \leq [2,3.5]$ provide the highest rate of valid results for the proposed algorithm, while, in Figure S2.b, the narrower band $[0,2] \leq q\,R_G \leq [2.5,3]$ provides the best agreement with BODIES, with an average relative error around 10%. As expected from the use of a Taylor approximation, the estimation in general improves when the lower bound tends to 0, for any given upper bound. From the plot in Figure S2.a, we can observe that the upper bound $q\,R_G < 3$ seems to be the most adequate to carry out the estimation. The fact that the estimation gets worse for both larger and smaller upper bounds seems to be in direct relation to the high order terms of the Taylor series: the required higher order terms might not be strong enough at lower upper bounds, but additional terms might show up for larger upper bounds, both phenomena impacting the successful estimation of the approximation parameters.

While the choice of 0 as the lower bound of the qR_g range seems natural both from the theorerical perspective and from the results in Figure S2, we show next that this is not always the best approach for experimental SAXS curves. In Figure S3 we show the amount of valid estimations for the experimental datasets studied in the manuscript, as a function of the chosen qR_g range. Each experiment can either yield a valid estimate for both oblate and prolate cases, only for prolate or an invalid estimate for both. For each qR_g range, we indicate these 3

possible outcomes by a different color. The results of S3 show that in some cases, extending the lower bound of qR_g to 0 might result in a worse performance of the estimation. We believe this to be due the fact that the least squares estimator used is the optimal one under additive white Gaussian noise (AWGN), which might not be the correct choice of noise model for SAXS curves. Behaviors that deviate from this AWGN model, like heteroskedasticity or multiplicativity, have been indeed proposed for SAXS experiments (e.g. Minh and Makowski, 2013 and Onuk et al, 2015). The fact the the method performs as expected in the results from Figure S2 using additive Gaussian noise, but doesn't perform as well for the experimental datasets seems to hint that a more noise-informed estimation could improve the reported estimation of the anisotropy parameters of molecules from their solution-SAXS curves. Given these results, we have chosen the range $1 \leq qR_g \leq 3$ to carry out the estimation for all SAXS curves in the manuscript, as it seems to be a range that presents a good performance for both computed and experimentally measured curves.

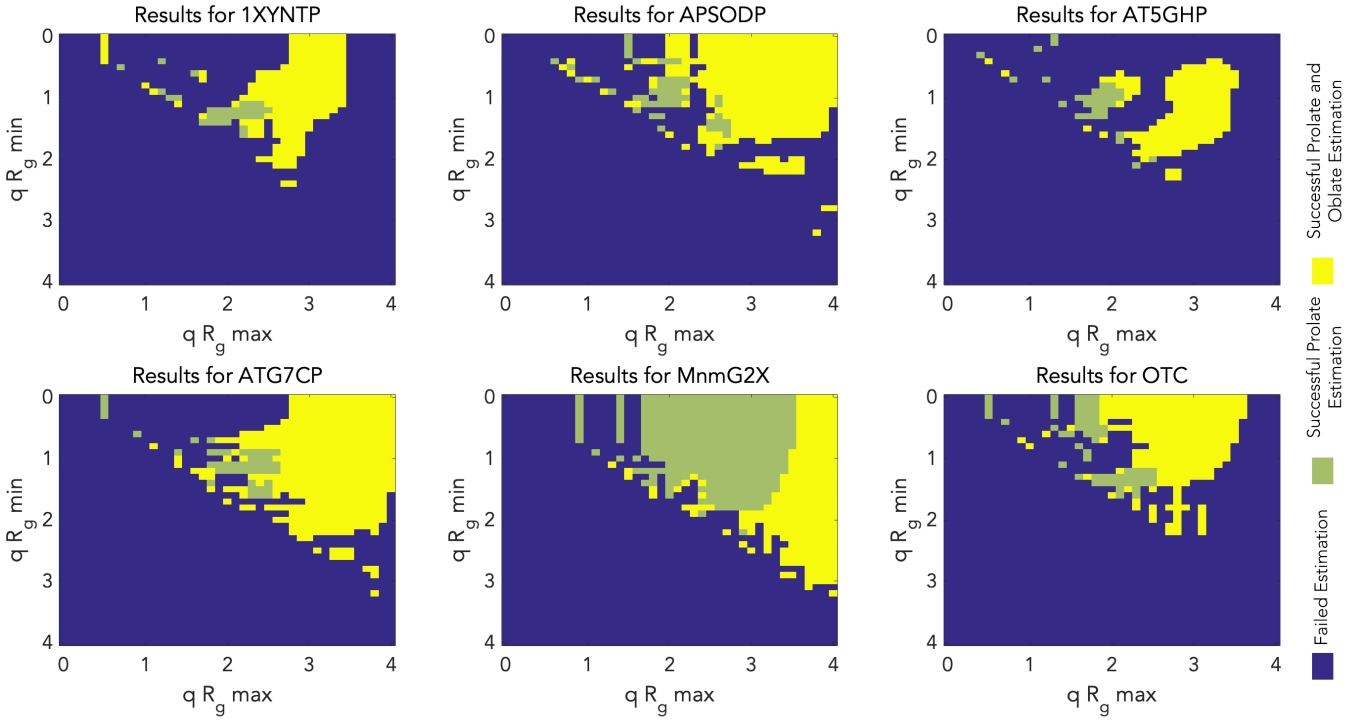

Figure S3: Valid estimations as a function of $[q\,R_G]_{min}$ and $[q\,R_G]_{max}$ for the six datasets studied in the main manuscript. Color indicates no valid estimation (blue), only prolate valid estimation (green) and both prolate and oblate valid estimations (yellow).