SUPPLEMENTARY INFORMATION

Carbonara: A Rapid Method for SAXS-Based Refinement of Protein Structures

Josh McKeown,[†] Arron Bale,[†] Cameron Brown,[‡] Hayden Fisher,[‡,‖] Robert P. Rambo,[¶] Jonathan W. Essex,[‡] Matteo T. Degiacomi,[*,§,⊥] and Christopher Prior[*,†]

†*Department of Mathematical Sciences, Durham University, South Road, Durham, DH1 3LE, United Kingdom*

‡*School of Chemistry and Chemical Engineering, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom*

¶*Diamond Light Source, Harwell Science and Innovation Campus, Didcot, United Kingdom*

§*EaStCHEM School of Chemistry, University of Edinburgh, Edinburgh, UK*

‖*ESRF - The European Synchrotron, 71 Avenue des Martyrs, 38043, Grenoble, France*

⊥*School of Informatics, University of Edinburgh, Edinburgh, UK*

E-mail: matteo.degiacomi@ed.ac.uk; christopher.prior@durham.ac.uk

1 Supplementary Methods

1.1 The tertiary backbone model used by Carbonara

In the following we summarise the essential components of the tertiary backbone model introduced in [1] to interpret small angle scattering BioSAXS data.

Constraints on local backbone geometry: curvature and torsion

The Cα backbone is modelled as a discrete set of three dimensional coordinates $\{\mathbf{x}_i\}$ expressed in a Cartesian coordinate system. These backbone curves are generated using Monte-Carlo sampling from a pair of distributions of two discrete geometric quantities which enforce realistic secondary geometry of this curve. The first quantity is the curvature κ, which describes the tightness of coiling of the curve (it is larger for an α-helix than a β strand) and is represented by the inverse of the sphere circumscribed on a section of four Cα coordinates, as shown in fig S1(a) (one can see a less tightly coiled curve section fig S1(b) has a larger inscribed sphere). The second quantity is the torsion τ which represents the local chiral nature of the backbone. It is positive for right handed coiling and negative for left handed coiling. On can see in fig S1(c) it has a very similar definition to the Ramachandran torsion angles, except here it represents the angle θ_n of the two two plane normals constructed from four Cα coordinates $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$. An example from a real protein structure is shown in Fig S1(d). To formally define these quantities for the set $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$ we denote the midpoints between each pair as $\mathbf{c}_i = (\mathbf{x}_i + \mathbf{x}_{i+1})/2$ and calculate (κ, τ) for this quadruplet using the following formulae.

$$\kappa(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = \frac{2|\sin(\theta_{123})|}{||\mathbf{c}_1 - \mathbf{c}_2||}, \tag{S1}$$

$$\tau(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = \sigma \frac{2}{l} \sin(\theta_n/2), \tag{S2}$$

$$l = (||\mathbf{x}_2 - \mathbf{x}_1|| + ||\mathbf{x}_3 - \mathbf{x}_2|| + ||\mathbf{x}_4 - \mathbf{x}_3||)/3.$$

where θ_{123} is the angle between the vectors $\mathbf{c}_1 - \mathbf{c}_3$ and $\mathbf{c}_2 - \mathbf{c}_3$, θ_n is the angle made by the normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ of the planes formed by the sets $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ and $(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$, and σ is 1 for right handed rotation -1 for left handed rotation.

In [1] an algorithm was defined, the discrete backbone algorithm, which inverts these formulae (eqs.S1 and S2) to generate a new Cα coordinate $\{\mathbf{x}_i\}$ from a pair of (κ_i, τ_i) values and the prior three coordinates $(\mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \mathbf{x}_{i-3})$. The C$\alpha$-C$\alpha$ distances are constrained to be $3.80\mathring{A}$. We now describe how the values of (κ, τ) are constrained to be representative of local protein geometry.

Imposing Ramachandran-like constraints on the local backbone geometry

Values of (κ, τ) pairs for all subsections of over 10000 protein structures were calculated (these proteins were chosen to have less than 70% sequence similarity). These values were then separated if they had been classed as belonging to an α-helix or a β strand. The remaining values were classed as linker sections. Probability distributions of these data sets are shown in Figure S2. For the linker distribution there are three distinct isolated peaks which correspond to the peaks in the respective α helical and β-strand distributions. These peaks were shown in the supplement of [1] to correspond to the three dominant Ramachandran angles hence we refer to them as "Ramachandran-like". One can find similar low resolution flexible Cα backbone models in a number of published studies *e.g.* [2, 3, 4, 5, 6, 7]. To the best of our knowledge ours is the only model that explicitly couples to a solution scattering model for the purpose of interpreting BioSAXS data.

Refining the Cα chain

The Cα trace is split into segments by assigning subsets of atoms to three possible secondary structure types: helices, strands, or linkers (see Section 1.4).

In any protein sequence, there are n distinct secondary structure groupings in this assignment, m of them linker sections. The geometry of an individual linker section is then altered as follows: say a given secondary grouping covers Cα 's l to m (1-3 for the first one for example), with coordinates $\{\mathbf{x}_i\}_{i=l}^{m}$, then new values $\mathbf{x_i}$ are determined as described in the previous section by choosing values $\{(\kappa_i, \tau_i)\}_{i=l}^{m}$ for each coordinate from the linker distribution, and generating new coordinates $\{\mathbf{x}'_i\}_{i=l}^{m}$ which ensure the chain locally has these curvature and torsion values. As illustrated in Figure 1 of the main text this alters the geometry of that sub-section of the curve and also then applies a translation of the coordinates $\mathbf{x}_j$ $j > m$ (as the vector separating $\mathbf{x}_{m+1}$ and $\mathbf{x}_m$ remains fixed). This process means the new configurations sampled are representative of plausible geometries and allow the side-chains to be reinserted later.

Specifying the flexibility of the molecule

This model requires that the secondary structure of the protein under study is defined. Carbonara can do this in several ways:

- Parsing PDB-defined assignment from the input structure.
- Applying DSSP to the initial structure [8].
- Applying STRIDE to the initial structure [9].
- Accepting a users-defined assignment

In this work, both examples used DSSP. This yields a set of linker sections imparting flexibility to the model. While initially every linker is defined as flexible, the user can manually lock a subset of them. For example, in the case of the IgG2 example, the linkers connecting constant and variable domains were not marked for variation, while those connecting the

inter-F(ab) domains (hinge region) were.

This selection can potentially significantly restrict the search space. To ease this choice, in the Carbonara Python front end, we provide a routine which tests which sections of the structure can be varied without altering any sheets structures (by testing if any sheet bonds are broken). This provides (if requested) an initial suggestion for which linker sections to be rendered flexible and which to keep fixed.

1.2 Validation of the Carbonara scattering model.

The following validations of the Carbonara scattering model were obtained in [1].

1. **The model was found to perform comparably to all atomistic models for SAXS structure verification**. In particular this method was applied to same set of globular proteins used in [10] to verify the FOXS method. This set includes monomer and multimer structures. In each case the model was found to fit the scattering data with a comparable accuracy. In turn the FOXS method was shown to perform comparably to both the CRYSOL [11] and the AQUASAXS [12] methods. Critically, this showed our coarse-grained model could identify the correct protein structure if the correct configuration were sampled.
2. **The effect of using an averaged scattering model for *ab initio* fitting gave a 1-2 Å variation in resolution**. Starting with a crystal structure for known scattering data (*i.e.* data we know belongs to that structure). The ensuing fit obtained with the averaged scattering model within 1-2 RMSD of the original structure. This averaged model is the one used in Carbonara.
3. **The method can achieve approximately 7.5 Å resolution fits to proteins from only a sequence and BioSAXS data**. This was demonstrated with a single protein, Lysozyme. This resolution is at the limit of what one can expect from data of $q \in [0, 0.6]\mathring{A}^{-1}$ [13] as indicated by the Shannon sampling theorem.

1.3 A description of the topological fold metric χ_t

In order to produce realistically entangled structures, a writhe-based constraint was introduced to the model. The absolute crossing number (ACN) is a positive definite measure of global self-entanglement of a curve. Intuitively, if we project a curve onto a plane, there will be points of self intersection. We can count these so-called crossings. The number of crossings will necessarily be dependent on the angle of projection chosen, so the ACN can be thought of as an average of the sum of crossings over all directions. The ACN of a smooth curve $\boldsymbol{\gamma}(t)$ is given by the Gauss linking integral [14]

$$ACN = \frac{1}{4\pi}\int_{\gamma}\int_{\gamma} \boldsymbol{T}(s)\times\boldsymbol{T}(t)\cdot\frac{|\boldsymbol{\gamma}(s)-\boldsymbol{\gamma}(t)|}{\|\boldsymbol{\gamma}(s)-\boldsymbol{\gamma}(t)\|^3}\,\mathrm{d}s\,\mathrm{d}t, \tag{S3}$$

where $T(t)$ is the tangent vector to the curve $\boldsymbol{\gamma}(t)$. Since we are considering the discrete protein backbone curve, we use the discrete analogue of the ACN given by

$$ACN(\mathcal{C}) = 2\sum_{i=2}^{n-1}\sum_{j<i}\frac{|\Omega_{ij}|}{4\pi}, \tag{S4}$$

where Ω_{ij} represents the contribution to Eq. S3 from the crossing of edges connecting $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i+1}$ and $\boldsymbol{x}_j$ to $\boldsymbol{x}_{j+1}$. There are numerous equivalent methods for computing Ω, and we will follow Method 1a given in [15]. For this, we denote by $\boldsymbol{r}_{i,j}$ the edge between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We then define the unit normal vectors:

$$\boldsymbol{n}_1 = \frac{\boldsymbol{r}_{i,j}\times\boldsymbol{r}_{i,j+1}}{\|\boldsymbol{r}_{i,j}\times\boldsymbol{r}_{i,j+1}\|}, \tag{S5}$$

$$\boldsymbol{n}_2 = \frac{\boldsymbol{r}_{i,j+1}\times\boldsymbol{r}_{i+1,j+1}}{\|\boldsymbol{r}_{i,j+1}\times\boldsymbol{r}_{i+1,j+1}\|}, \tag{S6}$$

$$\boldsymbol{n}_3 = \frac{\boldsymbol{r}_{i+1,j+1}\times\boldsymbol{r}_{i+1,j}}{\|\boldsymbol{r}_{i+1,j+1}\times\boldsymbol{r}_{i+1,j}\|}, \tag{S7}$$

$$\boldsymbol{n}_4 = \frac{\boldsymbol{r}_{i+1,j}\times\boldsymbol{r}_{i,j}}{\|\boldsymbol{r}_{i+1,j}\times\boldsymbol{r}_{i,j}\|} \tag{S8}$$

We consider the sum of the angles between these vectors

$$\Omega^* = \sin^{-1}(\boldsymbol{n}_1 \cdot \boldsymbol{n}_2) + \sin^{-1}(\boldsymbol{n}_2 \cdot \boldsymbol{n}_3) + \sin^{-1}(\boldsymbol{n}_3 \cdot \boldsymbol{n}_4) + \sin^{-1}(\boldsymbol{n}_4 \cdot \boldsymbol{n}_1). \tag{S9}$$

Then the evaluation of the Gauss integral from the crossing of $\boldsymbol{r}_{i,i+1}$ and $\boldsymbol{r}_{j,j+1}$ is given by

$$\frac{\Omega_{ij}}{4\pi} = \frac{\Omega^*}{4\pi} \operatorname{sgn}\left((\boldsymbol{r}_{j,j+1} \times \boldsymbol{r}_{i,i+1})\boldsymbol{r}_{i,j}\right) \tag{S10}$$

To compute the ACN of a protein, we define the backbone curve which is the discrete 3-dimensional curve connecting the central Cα atom of each amino acid residue. To reduce the impact of the helical nature of secondary structures on the ACN calculation, we must smooth the backbone in an appropriate manner. To do this, we apply the SKMT algorithm [16] to reduce it to a minimal representation that preserves any essential non-local entanglement of the backbone curve. Adapting the KMT algorithm [17, 18] to act solely within secondary structure elements (SSEs), we effectively replace each SSE with a straight edge, with some careful attention placed on linkers. As a result, the length of the SKMT smoothed backbone curve is proportional to the number of distinct SSEs.

In Fig.S7 we see the distribution of ACN for a representative sample of SKMT smoothed backbones from the PDB. A lower bounding curve was fit to this distribution so that 99.86% of the data falls above it. This lower bound is built into the model as a penalty during the fitting process. This ensures that any predicted structure is realistically entangled on a global scale according to the ACN.

1.4 Secondary structure assignment

SMARCAL1

The secondary structure assignment for Human SMARCAL1, obtained from DSSP, was:

```
---SSS---SSSS-----SSSS-----HHHHHHHH-----SSSHHH---S---HHHHHHHHHHH------SSSSSH
HHH---HHHH-------SSSSSSS------HHHHH---HHHHHHHHHHHHHHHS----------HHHHHHHHH---
HHHSSSSS---HHHHHHHHHHHH---HHHSHHHSSS----------------HHHH---HHHHHSSS---SS---
HH------HHHHHHHHHHH---SSSS---------HHHHHHHHHHH-------HHHHHHHHSSSSS-----------H
HHHHHHHHHHH-----HHHH---SSSSS-----SSHHH----HHHHHHHHHHHHH------HHHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHH-----SSS----HHHHHHHHHHHHH-------------HHHHHHHHHH---HH
H-----SHHH-------HHH------SSS---HHHHHHHH--------SSSSSSS------HHHHHHHHHHHHH
HHHHH-------HHHH---
```

with, `S` representing a strand helix, `H` an α-helix and `-` a linker section. The linker sections highlighted in red were those permitted to be changed during the fitting procedure.

IgG2

Using the same notation as for SMARCAL1, the four chains of the IgG2 molecule were assigned using DSSP as:

Chain 1

```
---SSSSSS-------SSSSSSSS-----HHHSSSSSSSS----SSS--SSSSHHH-SSSSSHHHHHH---SSSHH
H-SSSSSS----HHHS-SSSSSSSSS---------SSSSSSSSS-----SSSSSSSS--------------SSSS
------------HHH-----SSSSSSSSS------SSSSSS--HHHHH----SSSS--HHHSSSSSSSSSS-----
---SSSS------
```

Chain 2:

```
--SSSSS----------SSSSSSSSS-------SSSSSS----SSS-------SSS----HHH--------SSSSS
----HHHSSSSSSSS----SSSSSSSSSSS---SSSSS-SSSSSSHHHHH--SSSSSSSS-------SSSSSSHH
HSSS---SSSSSSSSSHHH-SSSSSSSS---HHHHHH----SSSS--HHH-SSSS-SSS---
```

Chain 3:

```
---SSSSSS-------SSSSSSSS-----HHHSSSSSSSS----SSSSSSSSSHHH-SSSSSHHHHHH---SSSHH
H-SSSSSS----HHHS-SSSSSSSSS---------SSSSSSSSS-----SSSSSSSS------HHHS-----SSSS
------------HHH-----SSSSSSSSS------SSSSSS--HHHHH----SSSSSSHHHSSSSSSSSSS-----
-------------
```

Chain 4:

```
--SSSSS----------SSSSSSSSS-------SSSSSS----SSS-------SSS----HHH----HHH-SSSSS
----HHHSSSSSSS-----SSSSSSSSSSS----SSSS-SSSSSSHHHHH--SSSSSSSS-------SSSS--HH
HSSS---SSSSSSSSSHHH-SSSSSSSS---HHHHHH----SSSS--HHH-SSSS-SSSS--
```

1.5 χ^2 calculation for WAXSIS

WAXSIS provides non-parametric SAXS profiles through explicit-solvent md simulation, avoiding the need for adjustable hydration layer modeling. To compare WAXSiS-computed intensities with experimental data, we first interpolated the theoretical curve onto the experimental q-values using a univariate spline interpolation. The optimal linear scaling factor c was determined by minimising the least-squares between the experimental (I_e) and theoretical (I_m) profiles

$$c = \frac{\sum I_e(q) \cdot I_m(q)}{\sum I_m(q)^2}. \tag{S11}$$

The agreement between experimental and theoretical curves was quantified using the standard χ^2 metric

$$\chi^2 = \frac{1}{N} \sum \frac{(I_e(q) - c \cdot I_m(q))^2}{\sigma(q)^2}, \tag{S12}$$

where N is the number of experimental points and $\sigma(q)^2$ is the experimental error at each q value. This approach maintains consistency with established SAXS profile comparison.

2 Supplementary Tables

Table S1: SAXS analysis and comparison between the AlphaFold structure and Carbonara model of SMARCAL1. All SAXS calculations were performed for q=[0.02, 0.22]. CRYSOL was run using 30 spherical harmonics and constant buffer subtraction enabled. FoXS and WAXSiS were run using default parameters.

		AlphaFold	**Carbonara + MODELLER**
χ^2	**CRYSOL**	13.52	1.01
	FoXS	16.86	0.97
	WAXSiS	36.7	1.77

Table S2: SAXS analysis and comparison between the crystal structure and Carbonara model of IgG2. All SAXS calculations were performed for q=[0.0, 0.2]. CRYSOL was run using 30 spherical harmonics and constant buffer subtraction enabled. CRYSOL R_g values were taken from the slope of net intensity. FoXS and WAXSiS were run using default parameters. The torsion angle was determined by calculating the dihedral formed by the centres of mass of the variable 1 (residues 1-121 + 242-349), constant 1 (residues 125-220 + 353-455), constant 2 (residues 580-676 + 808-910), and variable 2 (residues 456-576 + 697-804) domains. The hinge angle was calculated with the constant domain spanning both F(ab) arms, using the centers of mass of variable 1 (residues 1-121 + 242-349), the constant domain (residues 125-231 + 353-455 + 580-686 + 808-910), and variable 2 (residues 456-576 + 697-804).

		Crystal Structure	**Carbonara + MODELLER**
χ^2	**CRYSOL**	4.29	1.95
	FoXS	3.63	1.85
	WAXSiS	7.85	2.25
CorMap p-value		4.95e-15	2.86e-3
Rg (nm)	**CRYSOL**	3.78	3.95
	FoXS	3.63	3.88
	WAXSiS Rg	3.66	3.89
	WAXSiS Rg (solute)	3.67	3.90
Dmax (nm)	**CRYSOL**	12.35	13.08
	WAXSiS	12.53	13.41
Hinge angle °		116.4	125.3
Torsion angle °		61.4	75.8

3 Supplementary Figures

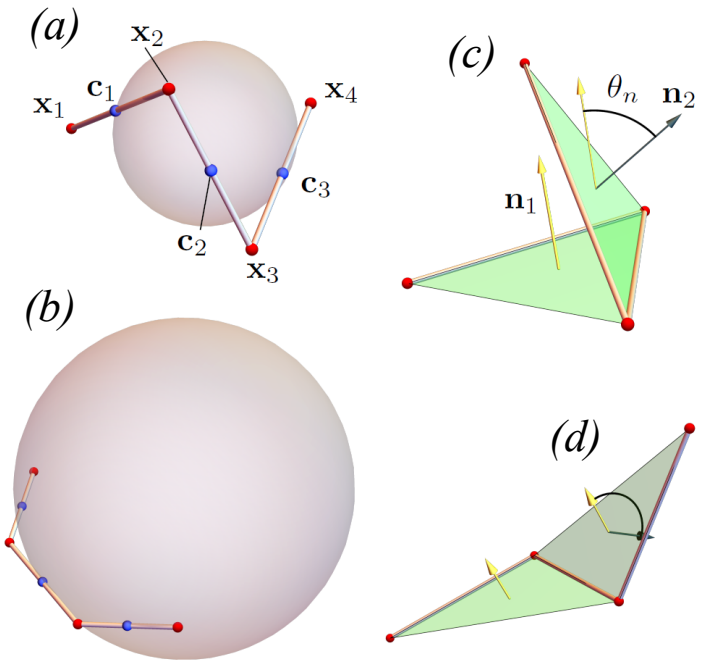

Figure S1: Illustrations of the geometrical interpretations of the curvature κ and torsion τ used to constrain the Cα backbone model. (a) four Cα coordinates $\mathbf{x}_i$ and their midpoints $\mathbf{c}_i$. There is a circumscribed sphere touching these midpoints whose inverse radius defines the curvature. (b) demonstration that a less tightly coiled section of curve than in (a) sees a larger sphere and hence smaller curvature. (c) a section of a set of 4 coordinates and the two planes they define. The normal vectors to these planes $\mathbf{n}_1$ and $\mathbf{n}_2$ are shown. The angle θ_n they make is used to define the torsion and is a measure of the non planarity of the set. (d) demonstration that θ_n can be large.

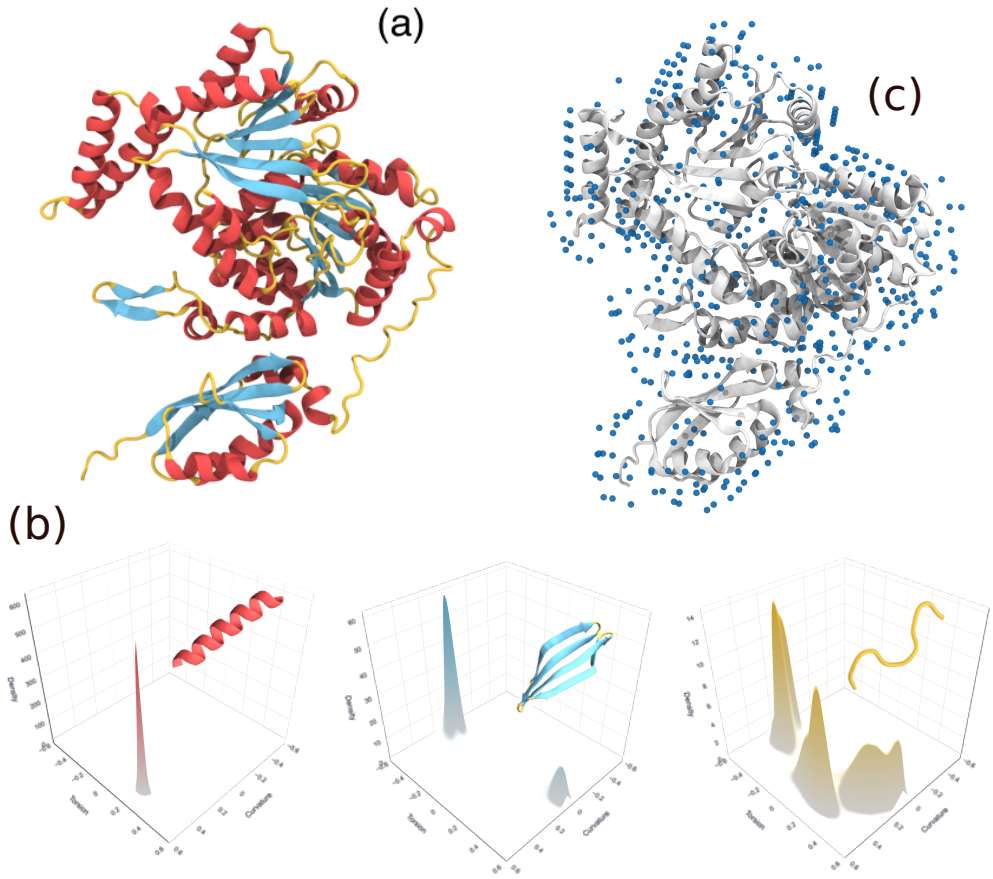

Figure S2: Illustrations of aspects of the Carbonara model. (a) a representation of a protein structure with its secondary structure clearly identified. (b) the empirically obtained probability densities for the curvature and torsion of each secondary structure type. (c) the explicit Carbonara hydration layer model for the structure shown in panel (a)

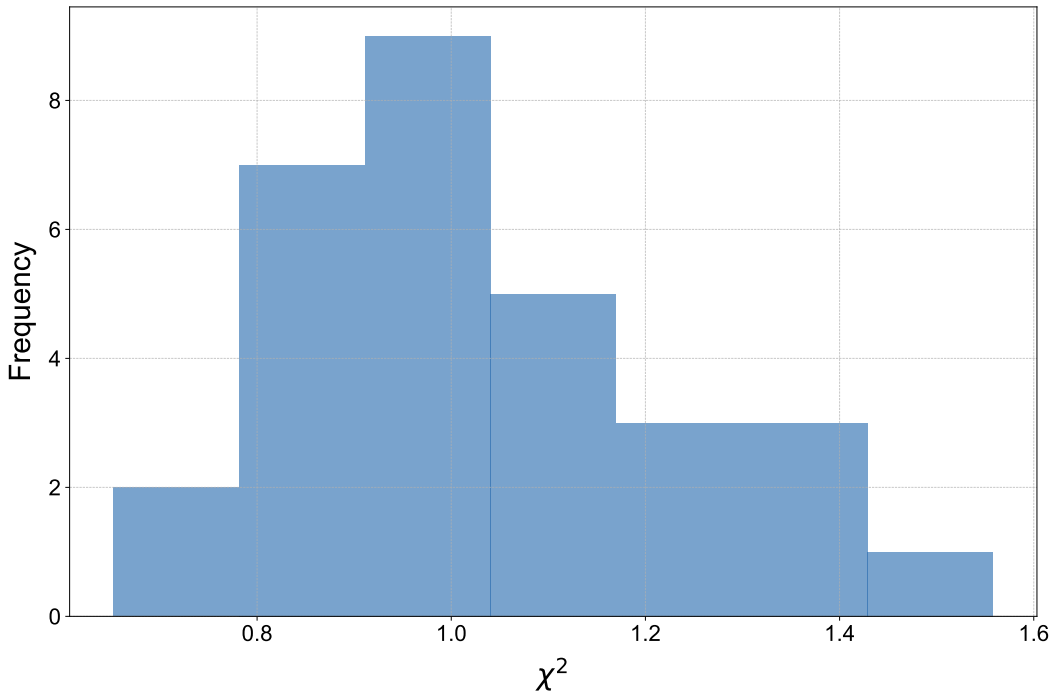

Figure S3: χ^2 distribution of Carbonara-refined SMARCAL1^{CD} structures determined by CRYSOL.

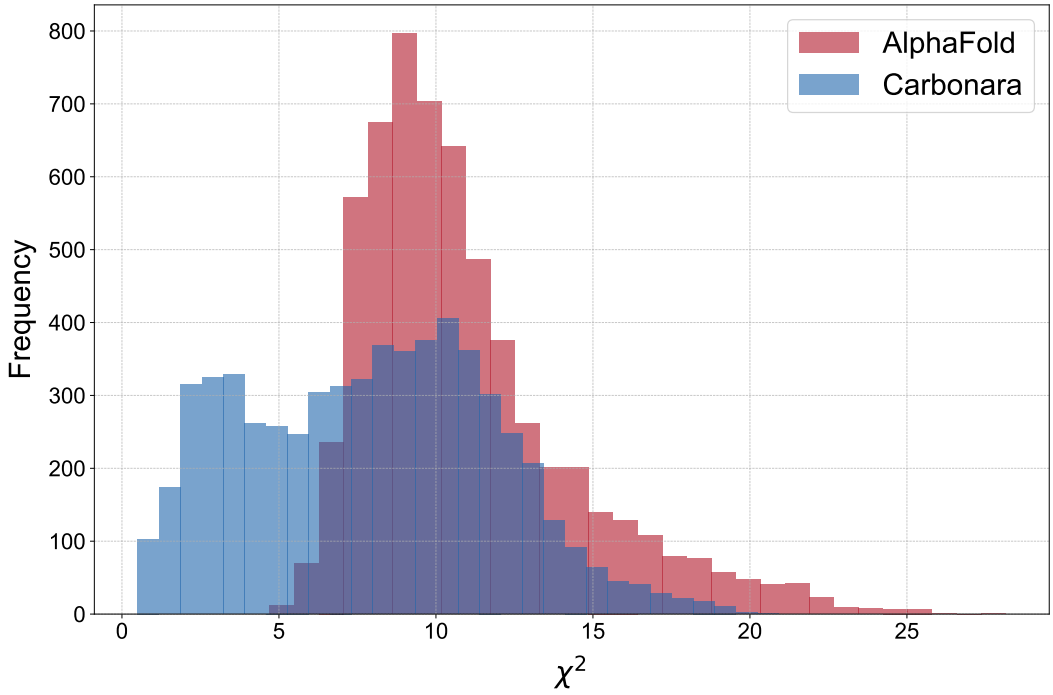

Figure S4: χ^2 distribution of SMARCAL1^{CD} AlphaFold (red) and Carbonara-refined seeded (blue) simulations calculated by CRYSOL.

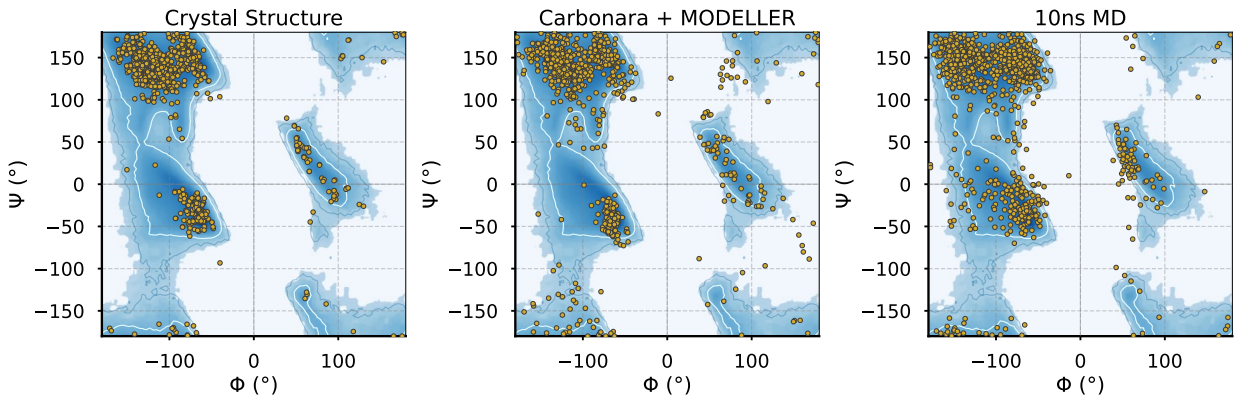

Figure S5: Ramachandran plots of the IgG2 C239S crystal structure, all-atom Carbonara output prior to MD, and all-atom Carbonara output after 10 ns of MD. Blue background densities are Ramachandran angles generated from a database of 8000 high resolution peptide crystal structures (plots generated using https://github.com/Joseph-Ellaway/Ramachandran_Plotter).

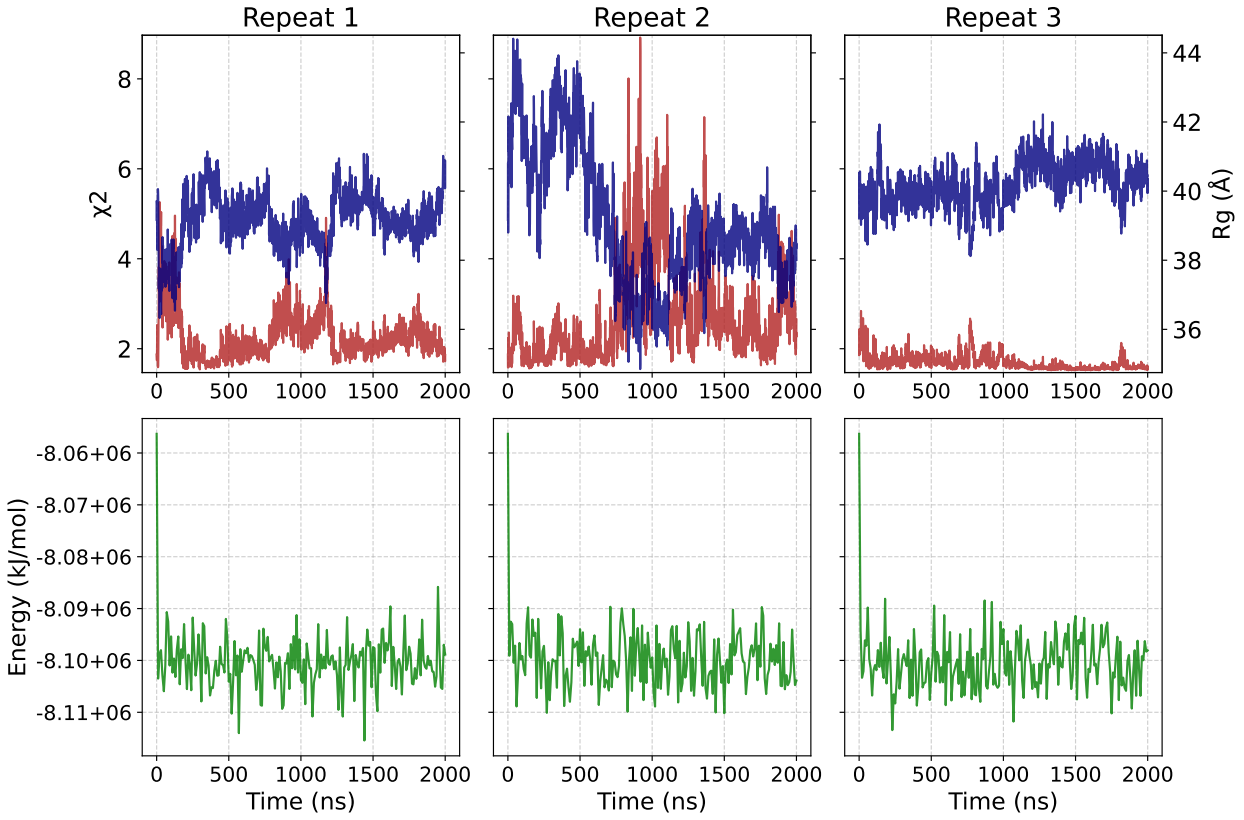

Figure S6: A plot showing R_g vs χ^2 across the three Carbonara-seeded IgG2 MD simulations as a function of time. The CRYSOL-calculated χ^2 values are shown in red, with R_g in blue. The corresponding total energy values for each production repeat are shown underneath in green, with every 10th frame plotted.

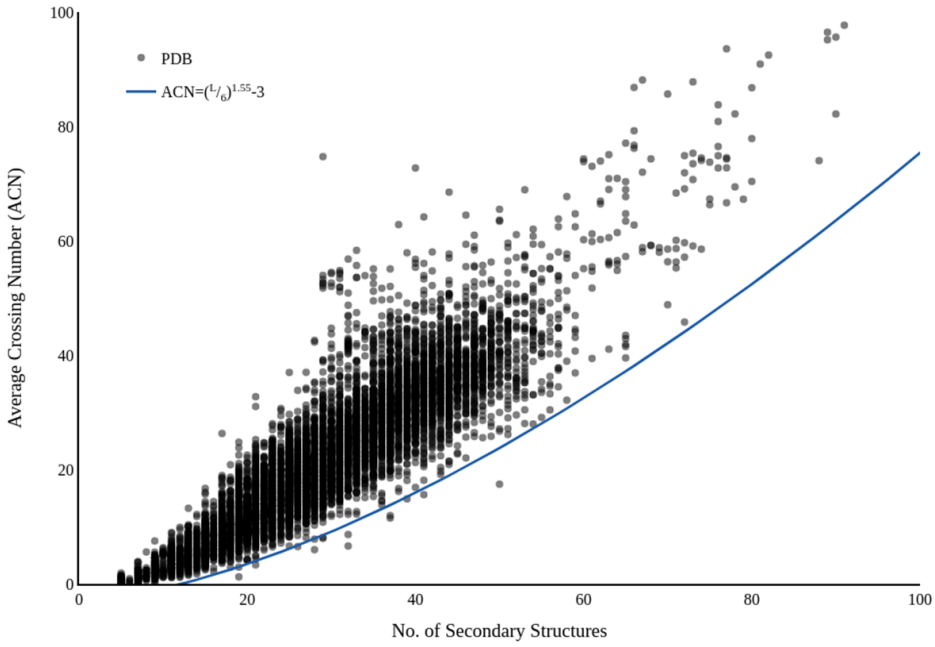

Figure S7: Distribution of absolute crossing number (ACN) of the smoothed backbones of a representative sample of proteins from the PDB. Figure modified from [19]

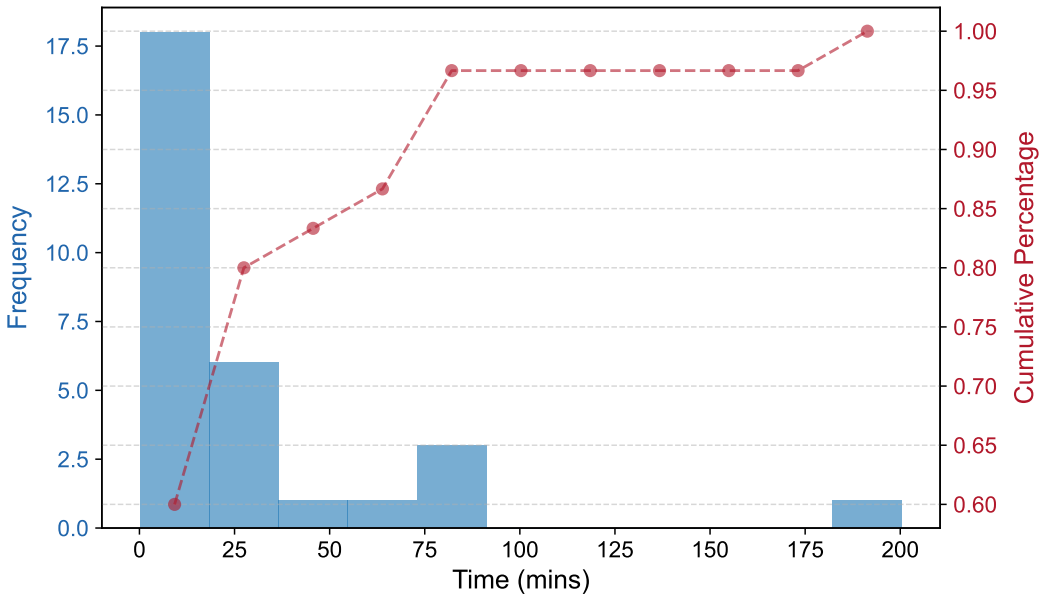

Figure S8: Time taken for Carbonara to generate each of the 30 SMARCAL1 models. The blue histogram shows the frequency of predictions across time intervals (minutes), while the red trail represents the cumulative percentage. 60% of predictions are completed within 18 minutes and approximately 80% within 36 minutes.

References

1 C. Prior, O. R. Davies, D. Bruce and E. Pohl, *Journal of chemical theory and computation*, 2020, **16**, 1985–2001.

2 A. C. Hausrath and A. Goriely, *Protein Science*, 2006, **15**, 753–760.

3 A. J. Hanson and S. Thakur, *Journal of Molecular Graphics and Modelling*, 2012, **38**, 256–278.

4 Z. Guo, E. Kraka and D. Cremer, *Journal of molecular modeling*, 2013, **19**, 2901–2911.

5 X.-B. Peng, J.-J. Liu, J. Dai, A. J. Niemi and J.-F. He, *Chinese Physics B*, 2020, **29**, 108705.

6 P. Grassein, P. Delarue, A. Nicolaı, F. Neiers, H. A. Scheraga, G. G. Maisuradze and P. Senet, *The Journal of Physical Chemistry B*, 2020, **124**, 4391–4398.

7 R. Chowdhury, N. Bouatta, S. Biswas, C. Floristean, A. Kharkar, K. Roy, C. Rochereau, G. Ahdritz, J. Zhang, G. M. Church et al., *Nature Biotechnology*, 2022, **40**, 1617–1623.

8 W. G. Touw, C. Baakman, J. Black, T. A. Te Beek, E. Krieger, R. P. Joosten and G. Vriend, *Nucleic acids research*, 2015, **43**, D364–D368.

9 M. Heinig and D. Frishman, *Nucleic acids research*, 2004, **32**, W500–W502.

10 D. Schneidman-Duhovny, M. Hammel and A. Sali, *Nucleic acids research*, 2010, **38**, W540–W544.

11 D. Svergun, C. Barberato and M. H. Koch, *Journal of applied crystallography*, 1995, **28**, 768–773.

12 F. Poitevin, H. Orland, S. Doniach, P. Koehl and M. Delarue, *Nucleic acids research*, 2011, gkr430.

13 R. P. Rambo and J. A. Tainer, *Nature*, 2013, **496**, 477–481.

14 G. Calugareanu, *Rev. Math. pures appl*, 1959, **4**.

15 K. Klenin and J. Langowski, *Biopolymers: Original Research on Biomolecules*, 2000, **54**, 307–317.

16 A. Bale, R. Rambo and C. Prior, *PLOS Computational Biology*, 2023, **19**, 1–27.

17 K. Koniaris and M. Muthukumar, *Physical review letters*, 1991, **66**, 2211.

18 W. R. Taylor, *Nature*, 2000, **406**, 916–919.

19 A. Bale, R. Rambo and C. Prior, *PLoS computational biology*, 2023, **19**, e1011248.