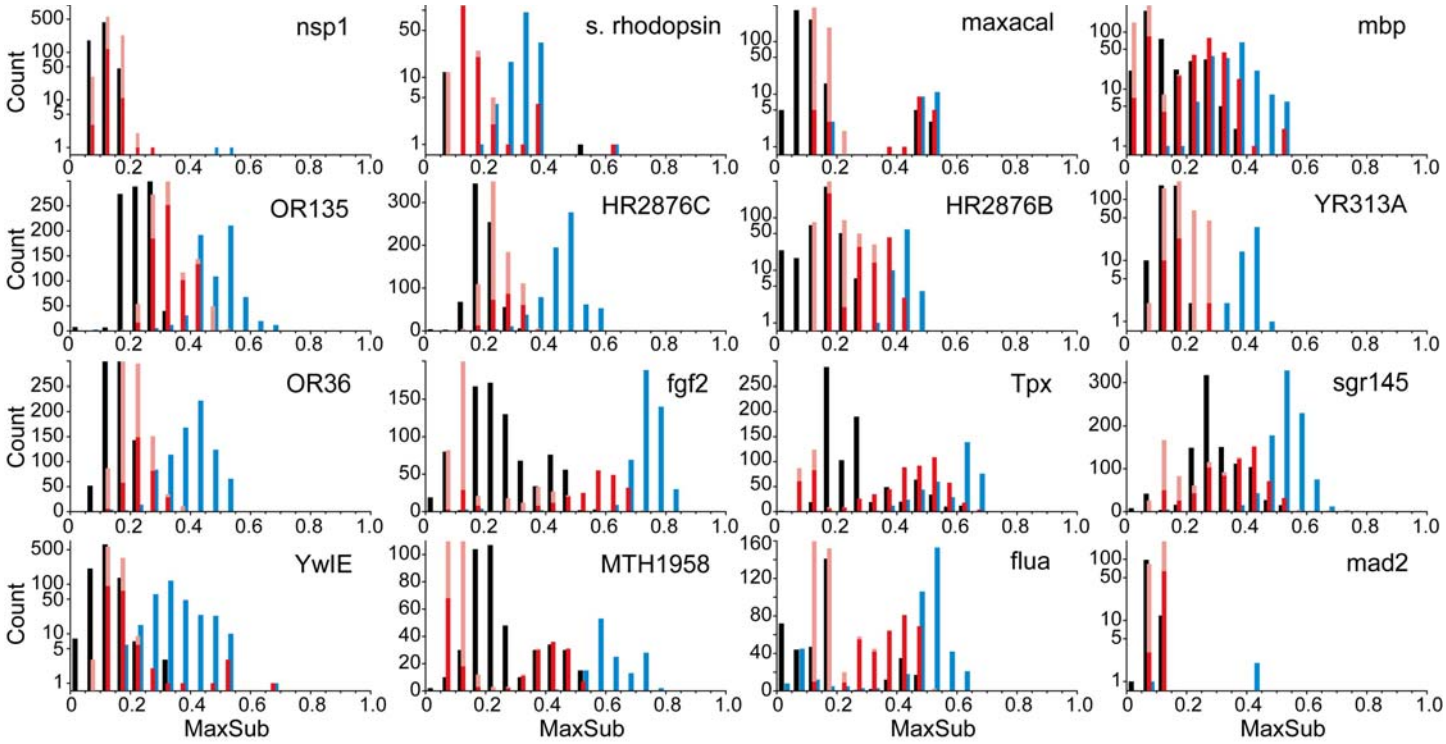
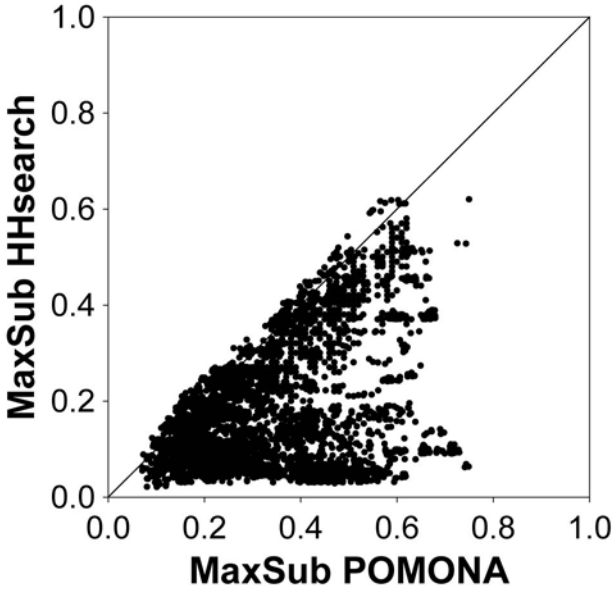


Supplementary Figure 1

Quality of protein structure alignments obtained by various methods.

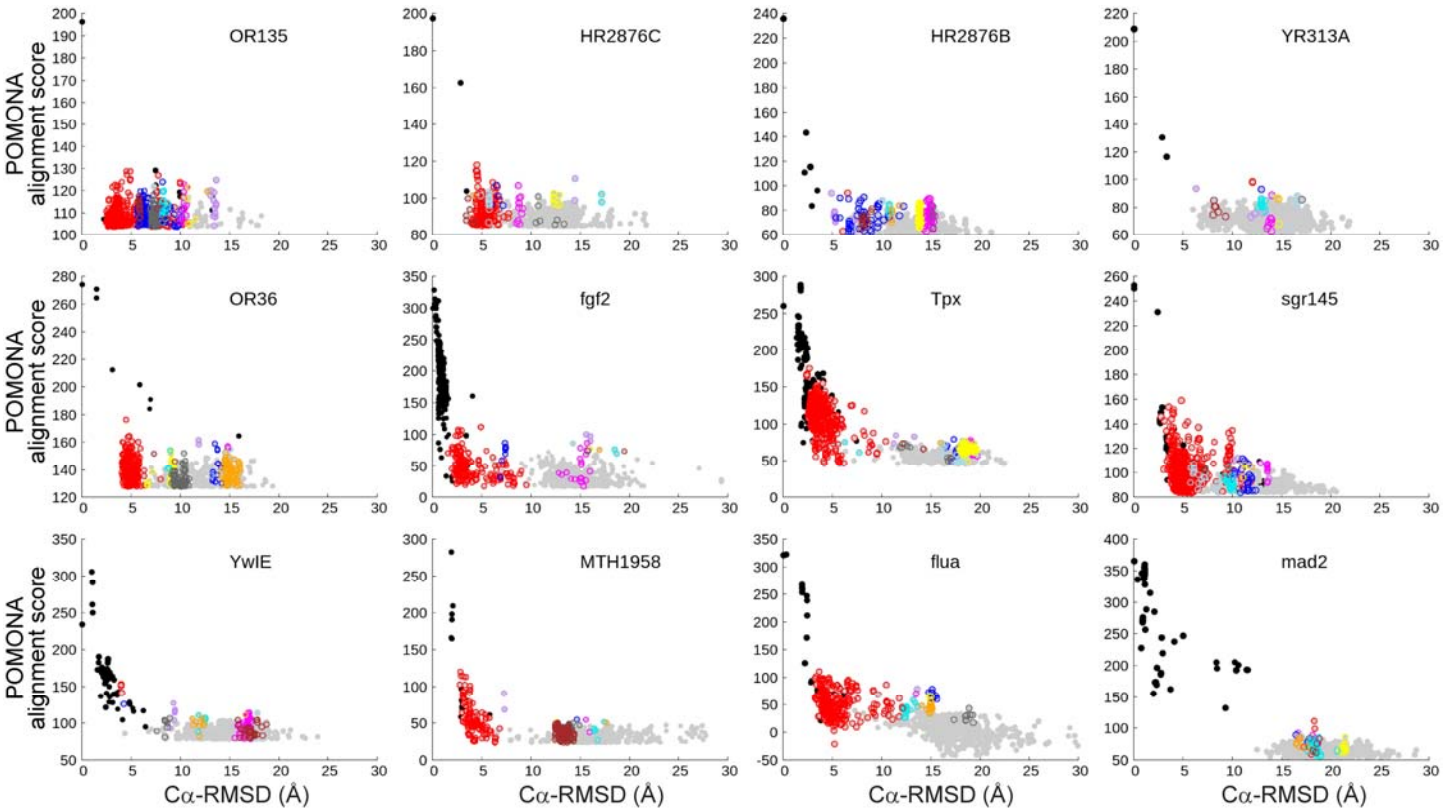
Structural quality is represented by the MaxSub score, with alignments identified by POMONA in red and by the sequence alignment method HHsearch in black, while the DALI-based structure alignment method (blue) shows the best possible alignments available in the PDB. Results are shown for each of the test proteins in **Table 1**, using a PDB from which all proteins with ≥20% sequence identity were removed. DALI and HHsearch results correspond to default thresholds of $Z > 2$ and Prob < 10%, respectively, used by these programs to identify homologs. Positive POMONA alignments are taken from the first 10 clusters (solid red bars) within the top 1,000 alignments (solid + transparent red) selected on the basis of highest H' score (equation (10)).



Supplementary Figure 2

Comparison of alignment quality obtained by sequence searching (HHsearch) and chemical shift–based alignment (POMONA).

MaxSub score of alignments obtained by HHsearch versus those obtained by POMONA (using a <20% sequence identity cutoff for both). The HHSearch alignments correspond to a setting of Prob ≥ 10%.



Supplementary Figure 3

Results of POMONA alignment for the 12 test proteins not shown in **Figure 1**.

For each protein, the database proteins with the top 1,000 POMONA alignment scores are shown as a function of C^{α} r.m.s. deviation relative to the experimental structures, with the C^{α} r.m.s. deviation restricted to the aligned parts in the target protein. Gray and black dots correspond to database proteins with a sequence identities of <20% and ≥20%, respectively. For the database hits with 20% sequence identity to the query protein, the ten clusters containing the highest alignment scores are colored according to the cluster number (red, purple, blue, magenta, light blue, yellow, cyan, orange, dark gray and brown for clusters 1–10, respectively). No NOEs were used.

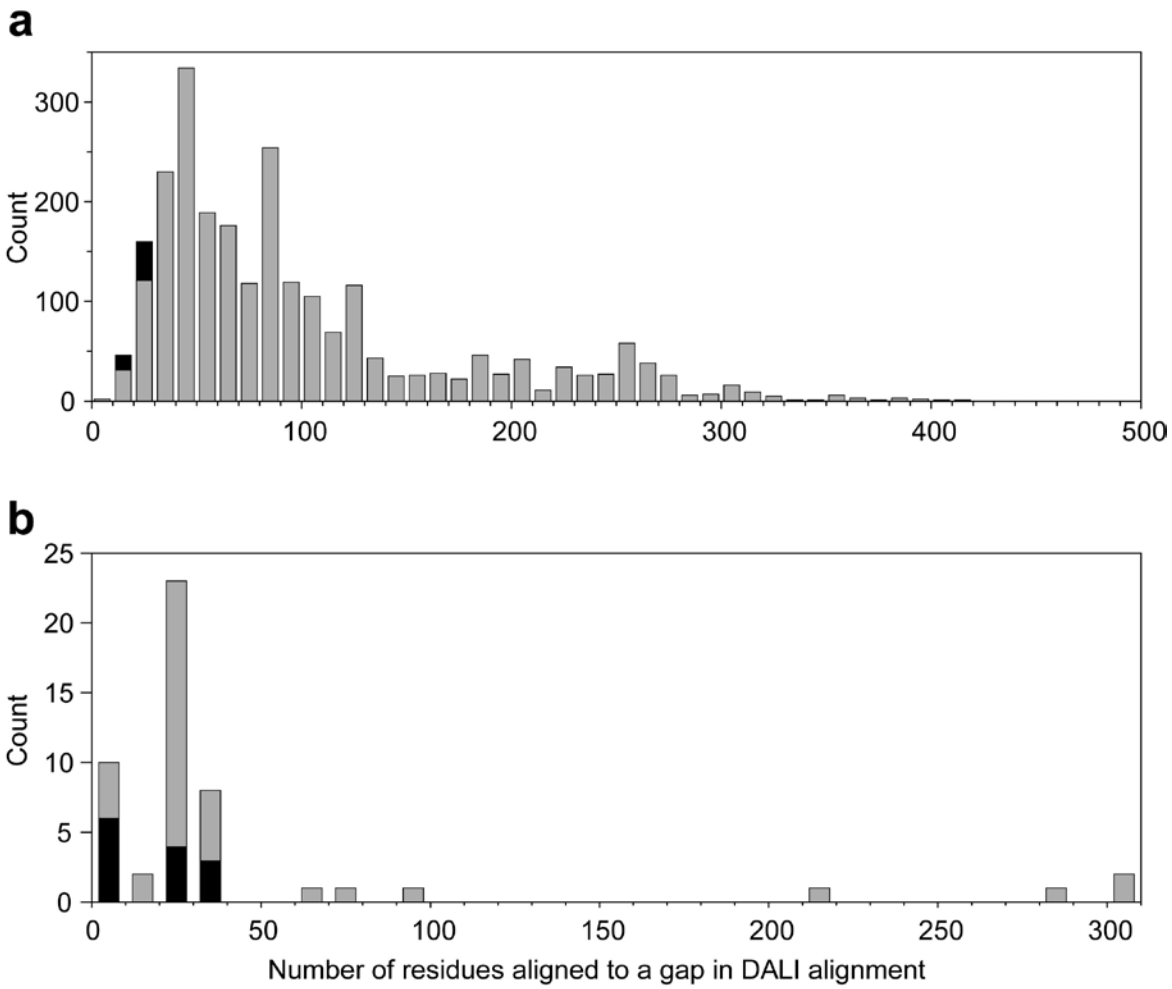
Supplementary Table 1. Performance of POMONA/CS-RosettaCM modeling for 16 test proteins when using a 30% sequence identity cut-off or when including sparse NOE data.

Name	Test 2 (CS) [a]			Test 3 (NOE) [a]			Test 4 (NOE) [a]		
	POMONA MaxSub[b]	$Rmsd_{mean}$ [c]	$Rmsd_{exp}$ [c]	POMONA MaxSub[b]	$Rmsd_{mean}$ [c]	$Rmsd_{exp}$ [c]	POMONA MaxSub[b]	$Rmsd_{mean}$ [c]	$Rmsd_{exp}$ [c]
OR135	0.56	0.79±0.16	1.43±0.11	0.52	0.98±0.27	1.46±0.32	0.52	1.01±0.34	1.51±0.24
HR2876C	0.45	1.61±0.29	2.14±0.32	0.33	1.52±0.34	1.91±0.31	0.33	1.19±0.33	1.70±0.42
HR2876B	0.62	1.40±0.22	1.86±0.25	0.41	1.91±0.32	2.63±0.54	0.41	1.47±0.27	1.95±0.27
YR313A	0.55	1.90±0.36	2.80±0.70	0.34	1.99±0.33	3.52±0.51	0.34	2.11±0.35	3.61±0.50
OR36	0.36	1.10±0.63	3.01±0.24	0.40	2.30±0.65	3.54±0.64	0.40	2.56±0.47	3.49±0.36
fgf2	0.91	0.70±0.16	1.07±0.16	0.62	1.02±0.22	1.96±0.23	0.57	1.02±0.15	1.44±0.31
tpx	0.74	1.23±0.37	1.99±0.25	0.68	0.99±0.16	1.83±0.24	0.69	0.83±0.11	1.83±0.24
sgr145	0.66	1.68±0.20	2.29±0.34	0.63	1.65±0.33	2.18±0.31	0.55	1.51±0.46	2.28±0.46
nsp1	0.30	2.18±0.63	3.30±0.73	0.42	2.01±0.31	2.49±0.49	0.41	1.44±0.37	1.90±0.29
YwlE	0.76	0.98±0.22	1.86±0.15	0.68	1.49±0.38	2.14±0.39	0.68	1.62±0.43	2.17±0.51
MTH1958	0.51	1.30±0.18	2.35±0.17	0.51	2.15±0.59	2.18±0.76	0.51	1.37±0.32	2.33±0.29
fluA	0.53	2.00±0.42	3.26±0.43	0.44	1.88±0.47	3.13±0.51	0.44	1.74±0.34	3.06±0.31
mad2	0.73	1.32±0.20	1.84±0.30	0.13	7.6±2.4	15.8±2.9	0.13	11.1±4.6	14.3±2.8
s. rhodopsin	0.85	0.91±0.20	1.61±0.19	0.41	1.84±0.34	2.52±0.38	0.62	2.13±0.38	2.75±0.51
maxacal	0.68	2.38±0.75	3.99±0.82	0.41	3.20±0.57	4.30±0.89	0.50	2.76±0.75	3.71±0.82
mbp	0.74	2.73±0.50	4.24±0.73	0.49	2.82±0.99	4.23±0.92	0.48	2.27±0.65	4.57±0.75

[a] Test 2: POMONA/CS-RosettaCM uses a < 30% sequence identity cut-off to the query protein for generating templates; Test 3 and 4: Using a < 20% sequence identity cut-off, but including N/10 long-range NOEs randomly selected from the $^{1}H^{N}$-$^{1}H^{N}$ distances <5Å in the query protein of size N residues.

[b] The highest MaxSub value observed for the alignments selected by POMONA.

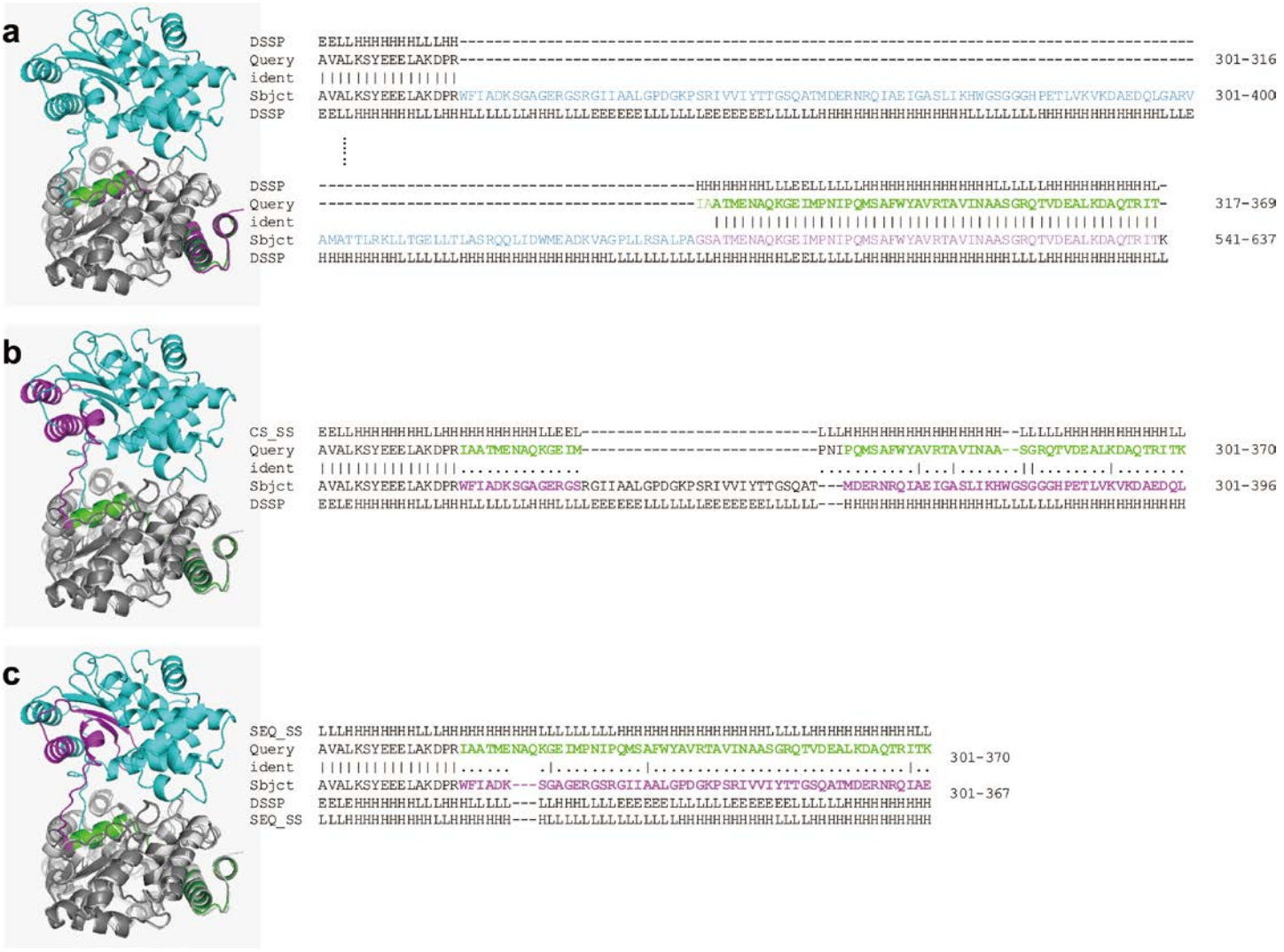
[c] C^{α}-RMSD calculated for all non-flexible residues (as defined by RCI-S^2 ≥ 0.6 (ref. [1])). $RMSD_{mean}$ is the C^{α}-RMSD between the 10 lowest-energy models and their mean coordinates. $RMSD_{exp}$ is the C^{α}-RMSD between the 10 lowest-energy models and the reference structure.



Supplementary Figure A

Distribution of the number of residues aligned to a gap (gap size) observed for DALI-identified alignments not captured by POMONA for the 16 test proteins.

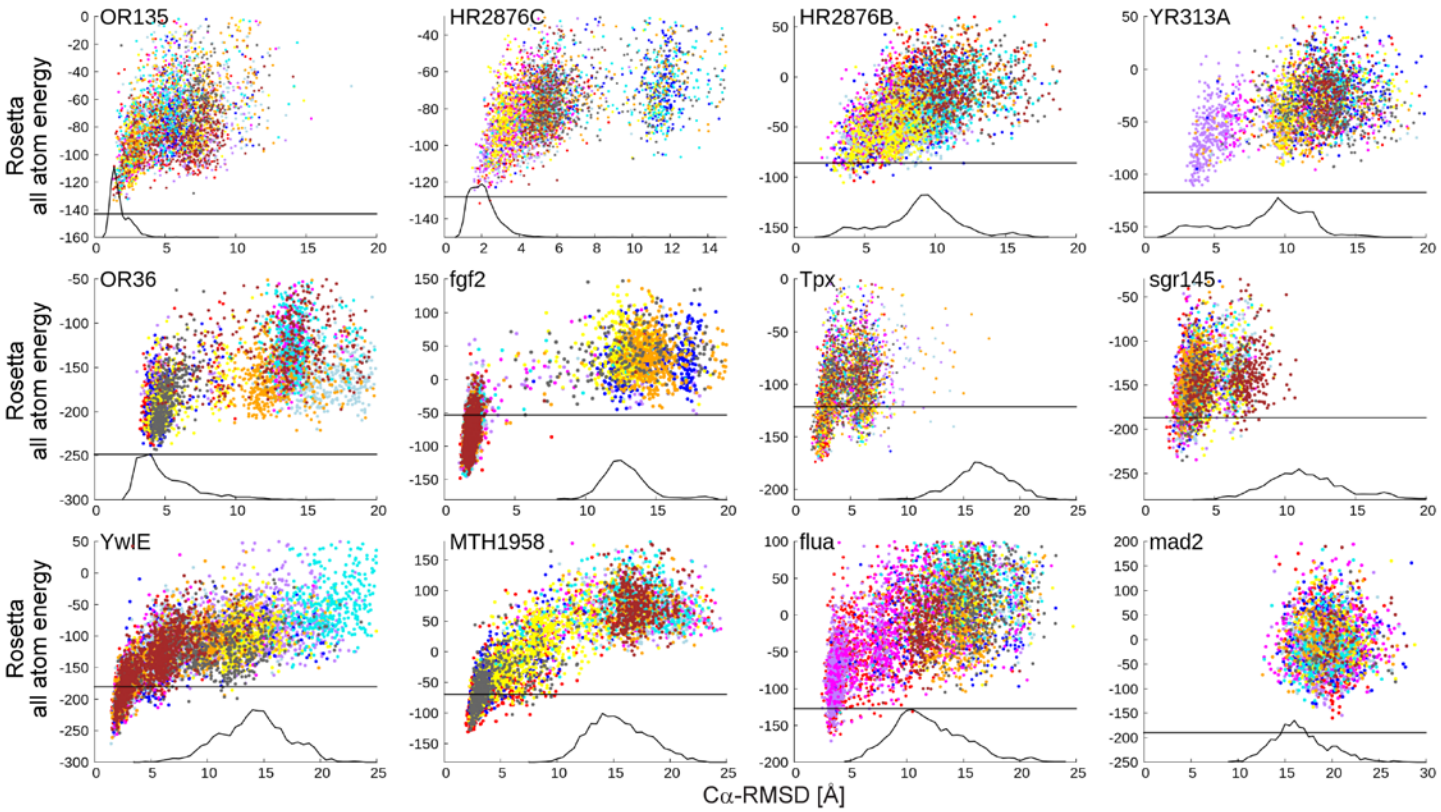
(**a**) for DALI alignments with <20% sequence identity to the query protein, and (**b**) for DALI alignments with ≥ 20% sequence identity to the query protein. Alignments to a database protein with a structure solved by NMR are in black (for (**a**), alignments with ≤30 residues aligned to a gap are included in the black bars).



Supplementary Figure B

Alignments between maltose binding protein (PDB entry 1DMBA) and protein RG13 (PDB entry 4DXBA) obtained by different methods.

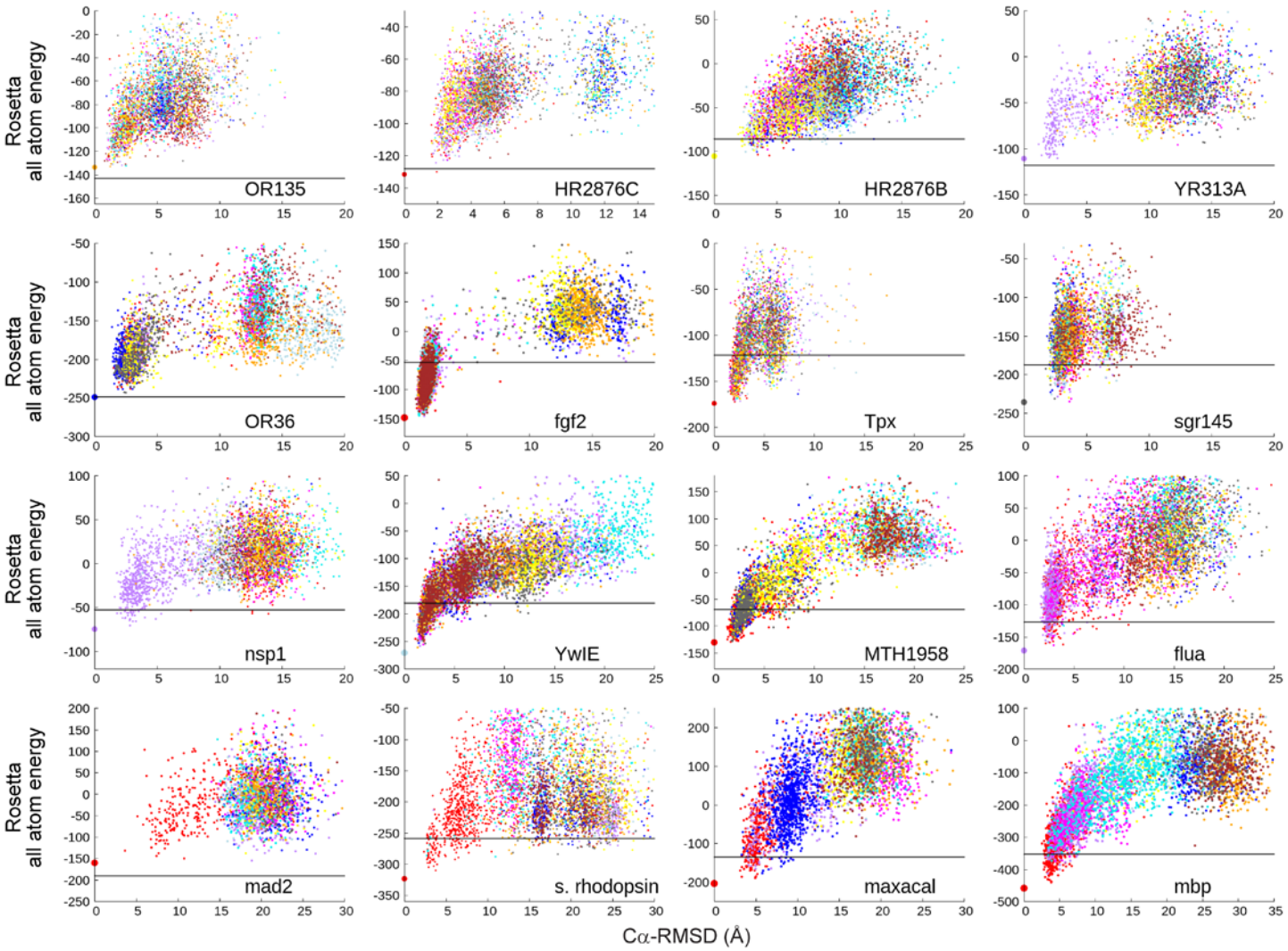
(**a**) DALI, (**b**) POMONA, and (**c**) HHsearch. (**a**) For the DALI alignment, the use of the known structure for both proteins allows a perfect alignment with a Cα-RMSD value of 1.14Å (MaxSub = 0.92) between the equivalent residues in the two proteins, i.e., residues 1-316 (of 1DMBA, grey) to 1-316 (of 4DXBA, grey) and 317-370 (1DMBA, green) to 584-637 (4DXBA, magenta), regardless the 267-residues inserted domain in 4DXBA (colored in cyan). The text version alignment on the right, represents, from top to bottom lines, the DSSP identified secondary structure for the query protein (line starting with a “DSSP” tag), the query sequence (“Query”), the sequence identity between the query and the database proteins (“ident”), the homologous protein sequence (“Sbjct”) and the DSSP designation of this homologous protein (“DSSP”), respectively. (**b**) For the POMONA alignment, the first 316 residues of 1DMBA and 4DXBA are aligned, the remainder of 1DMBA (residues 317-370, green), containing three helices, are (incorrectly) aligned to the first three helices of the inserted domain of 4DXBA (317-396, magenta). For the text version of the alignment, the TALOS-N identified secondary structure classification using the chemical shifts (line starting with a “CS_SS” tag) is listed rather than the structure-based DSSP secondary structure classification. (**c**) Use of only the sequence, HHsearch is able to align two proteins for residues 1-316 preceding the inserted domain of 4DXBA as they have identical sequences. However, the remainder of 1DMBA (residues 317-370, green), containing three helices, are (incorrectly) aligned with only a three-residue gap to the beginning of the inserted domain of 4DXBA (317-367, magenta), which includes two incorrect □-strands. The mis-aligned section results from the erroneous outcome of the sequence-based secondary structure prediction (line starting with a “SEQ_SS” tag) for the aligned part in 4DXBA, which is used by HHsearch to guide its sequence alignment.



Supplementary Figure C

Rosetta all-atom energy versus Cα-RMSD to the experimental structures.

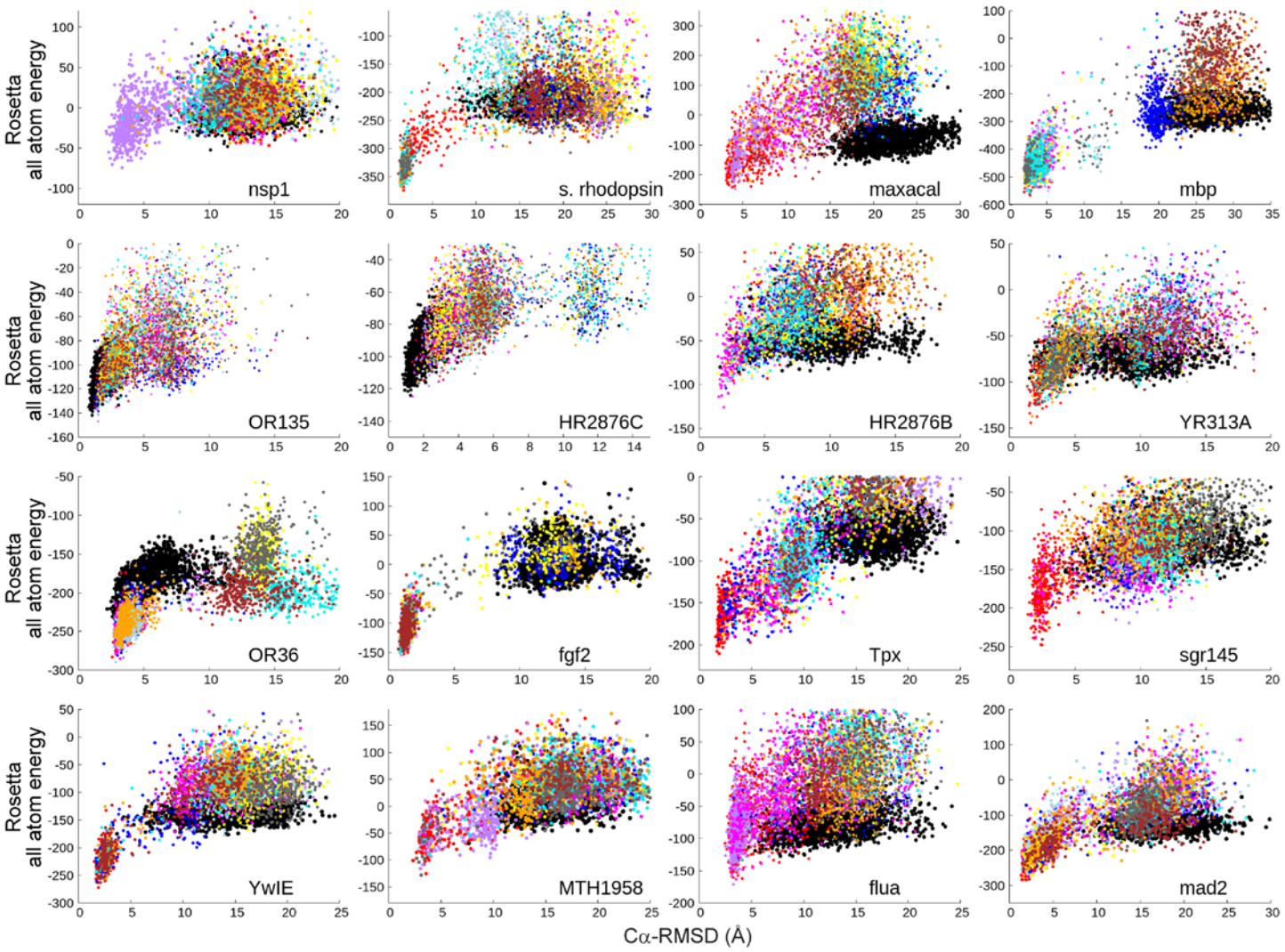
Results are shown for the 12 test proteins of **Table 1**, not shown in **Figure 1c**. For all test proteins, CS-RosettaCM models are generated by using only the backbone and 13Cβ chemical shifts as input and the POMONA-identified alignments of database proteins with a sequence identity < 20%. Each CS-RosettaCM model is colored according to the cluster number of the starting template from which it is modeled (see **Supplementary Fig. 3**). The horizontal line and the graph at the bottom represent the lowest Rosetta all-atom energy and the normalized number of structures found with a given Cα-RMSD, respectively, obtained with the csRosetta protocol.



Supplementary Figure D

Evaluation of acceptance criteria, based on convergence and energy limits, for the models generated by POMONA/CS-RosettaCM.

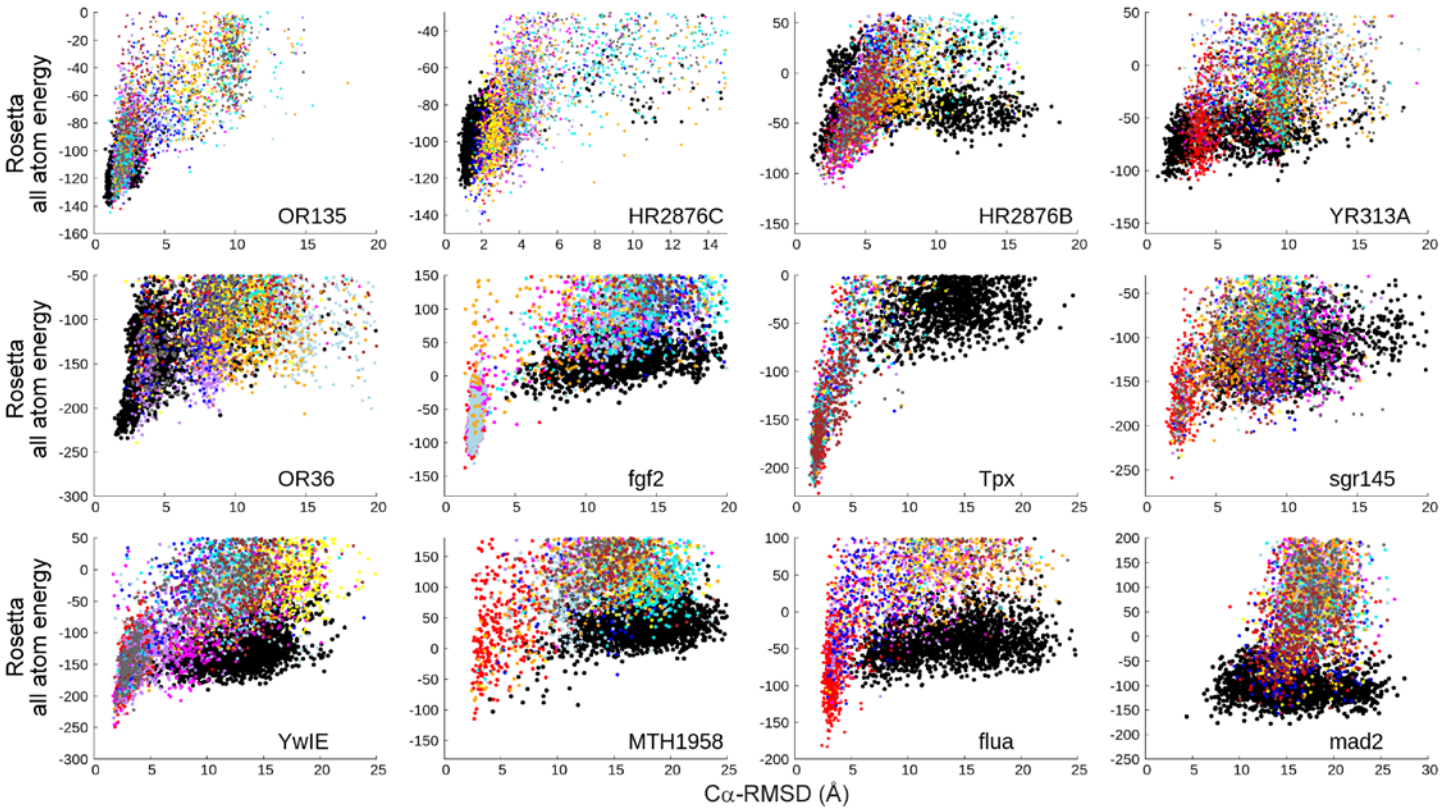
Each panel plots the all-atom energy, incl. chemical shift score, of the up to 10,000 POMONA/CS-RosettaCM models against the Cα-RMSD, calculated relative to the lowest energy model, with the color coding representing the starting template (see **Fig. 1b-c** and **Supplementary Fig. 3**). The horizontal line represents the lowest Rosetta all atom energy observed for the csRosetta structures generated by using the same inputs of chemical shifts and Rosetta energy scores. Acceptance criteria are not met for OR135, HR2876C, HR2876B, YR313A and OR36, as no improvement in energy is obtained over CS-Rosetta results, and not for mad2 both for not meeting the all-atom-energy criterion and the absence of convergence.



Supplementary Figure E

Plots of Rosetta all atom energy versus Cα-RMSD from the reference structure, when including database proteins with ≤30% sequence identity.

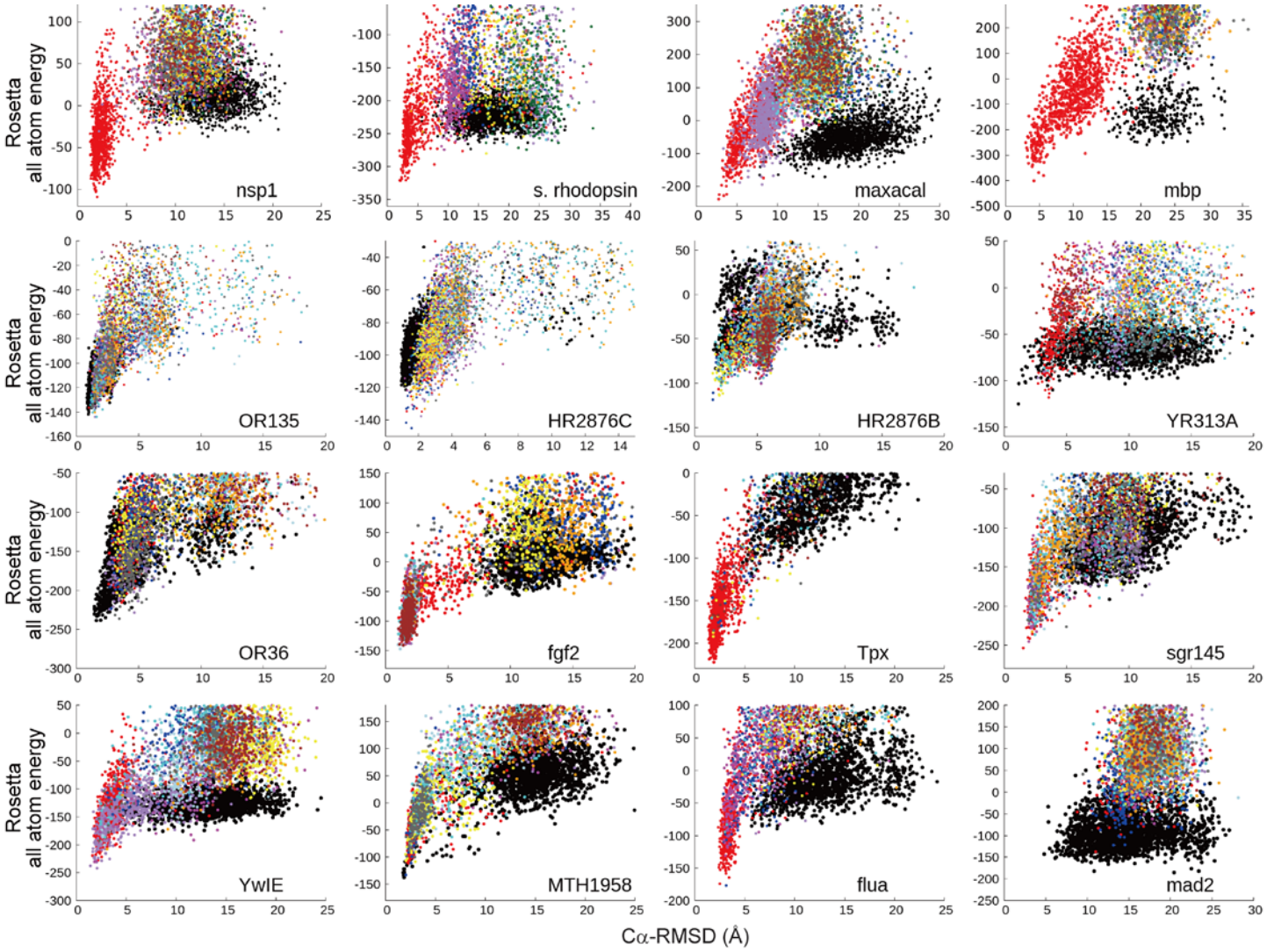
Each POMONA/CS-RosettaCM model is colored according to the cluster number from which it is derived, with red, purple, blue, magenta, light blue, yellow, cyan, orange, grey and brown for the best scoring POMONA clusters 1-10, respectively. For comparison, energies obtained with the standard CS-Rosetta protocol are shown in black.



Supplementary Figure F

Rosetta all-atom energy versus Cα-RMSD from the reference structure, when including sparse HN-HN NOEs.

Results are shown for the 12 test proteins, not presented in **Figure 1d**, main text. For all test proteins, POMONA/CS-RosettaCM models are generated by using the backbone and 13Cb chemical shifts plus additional sparse NOEs as input, using a < 20% sequence identity cut-off. Each CS-RosettaCM model is colored according to the cluster number of its starting template. For comparison, the energies of models obtained with standard CS-Rosetta are shown as black dots.



Supplementary Figure G

Rosetta all-atom energy versus Cα-RMSD from the reference structure, when including sparse HN-HN NOEs (set #2).

For all test proteins, POMONA/CS-RosettaCM models are generated by using the backbone and 13Cβ chemical shifts plus a second set of additional sparse NOEs as input, using a < 20% sequence identity cut-off. Each CS-RosettaCM model is colored according to the cluster number of its starting template. For comparison, the energies of models obtained with standard CS-Rosetta are shown as black dots.

SUPPLEMENTARY RESULTS

Chemical shift guided homology modeling of larger proteins

Yang Shen* and Ad Bax*

Performance evaluation of POMONA structure alignment

When comparing our results for the set of 16 test proteins to the theoretical limit defined by DALI-identified alignments, POMONA identifies nearly all of the 2,660 homologues with a sequence identity of ≥20% to the target proteins, missing only ~50 DALI-identified structural homologues. The POMONA-identified homologues show near optimal alignments in terms of MaxSub score compared to DALI (**Fig. 2b**), performing much better than sequence based alignment (**Supplementary Fig. 2**), such as the HHsearch method[19], one of the best modern sequence-based alignment methods. Structural homologues identified by DALI but missed by POMONA all have long alignment gaps (**Supplementary Fig. A**), resulting in depressed POMONA alignment scores that then fall below the detection threshold. Aligning two proteins with long alignment gaps is invariably challenging with a Smith-Waterman based algorithm, as too small a gap penalty would open false gaps whereas too large a gap penalty prevents the opening of gaps. In our work the implemented gap penalty function is tuned to allow the alignment to cross relatively small gaps; the largest lengths for single alignment gaps observed in the POMONA identified alignments are in the 20-30 range.

Not surprisingly, POMONA performs well for finding alignment for homologues with significant sequence identity (> 20-30%), where many other sequence alignment methods also perform well even when solely relying on sequence information. The more important question therefore is how well the program functions in finding structural homologues when there is very little or no significant sequence identity. When parameterized to detect even very weakly homologous structures, HHsearch identifies similarity in a total 10,059 protein chains with a sequence identity of <20% to the target protein for our set of 16 test proteins. Of these, only 8% are consensus with the DALI identified structural homologues that have a sequence identity of <20%, and 85% of the 5,211 DALI-identified structural homologues cannot be identified by HHsearch on the basis of sequence alone. Importantly, POMONA identifies among its positive alignments a large portion (~46%, 2,414/5,211) of DALI-identified homologues target proteins when restricting the search to proteins with <20% sequence identity (**Fig. 2a**). Structural homologues missed by POMONA nearly all exhibit long alignment gaps (≥ 30) in the DALI identified alignments (**Supplementary Fig. Aa**). We find that POMONA also missed a number of NMR-determined structures, even in the absence of large gaps. Inspection of these

Nature Methods: doi:10.1038/nmeth.3437

structures indicates that even though the fold of these proteins is close enough to register a DALI alignment, the local backbone deviates too far from ideality to allow their recognition on the basis of chemical shifts.

Evaluation of clustering and selection of POMONA alignments

When applying the selection criterion (see Online Methods) to the POMONA alignments with a sequence identity <20%, for most of our 16 test proteins the highest MaxSub score observed for this (up to) 20-member subset is comparable to that obtained for the top 1,000 positive alignments (**Table 1**). Evaluating the suitability of a given protein alignment as input for RosettaCM comparative modeling is not a straightforward problem, in particular when the protein contains gaps and/or inserts. Alignment accuracy, i.e. the RMSD between the coordinates of equivalent C^{α} atoms of corresponding residues in the database and query proteins is an important but not the only metric, and coverage can play an equally important role. For example, an aligned database protein chain with a 3-Å RMSD to the target protein and 50% alignment coverage is not necessarily better for structure modeling than one with a 5-Å RMSD but having 90% alignment coverage. An extreme example is seen when comparing the alignments between maltose binding protein, or MBP (PDB and chain id: 1dmbA), and three chains of the engineered protein RG13 (4dxbA, 4dxbB, 4dxcA) (**Fig. 1b**), which has a high POMONA alignment score but poor alignment accuracy, with a C^{α}-RMSD value of > 15 Å. RG13 is derived from MBP by substituting its residues 317 and 318 by a 267-residue domain. The MBP domain of RG13 has a sequence identity >99% and a C^{α} RMSD of only 1.14 Å relative to MBP (**Supplementary Fig. Ba**). However, due to the large penalty associated with the 267-residue alignment gap, POMONA matches the first 316 residues of these two proteins, or 85% of its total length. The C-terminal 15% fraction of the chain, consisting of three α-helices, are matched by POMONA to the first three helices of the domain inserted in RG13 (**Supplementary Fig. Bb**), resulting in > 15 Å C^{α} RMSD. However, despite this large RMSD, the final refined structures are fairly close to the MBP X-ray reference structure, which can be credited to the power of the CS-RosettaCM procedure when provided with the correct secondary structure input.

Performance evaluation of CS-RosettaCM protein structure generation

POMONA-identified alignments offer an accuracy and coverage that approaches the DALI limit. After the clustering and selection procedure, these then are used as input for CS-RosettaCM to generate complete structural models.

For comparison, the standard CS-Rosetta structure generation protocol[10] is also performed for all 16 test proteins. However, CS-Rosetta is only able to generate converged (and correct) structures for the smallest of the

proteins tested, and for the two proteins with less than 100 residues it actually outperforms the POMONA/CS-RosettaCM method. The latter reaches convergence for 15 out of the 16 proteins tested (after having removed all proteins with ≥ 20% sequence identity from the database), with all of these being close to the target structure ($RMSD_{100}$ < ~2.5Å; **Table 1**, **Supplementary Fig. C**).

The 16 proteins used to evaluate our method pose different types of challenges. First we focus on four proteins selected from structural genomics projects, incl. HR2876B, YR313A, OR36 and nsp1, that have very few sequence homologues (**Fig. 1b**, **Supplementary Fig. 3**) and no good structural homologues in the database (**Table 1**). Indeed, the best structural homologues identified by DALI for database proteins with <20% sequence identity all have MaxSub scores ≤0.5 (**Table 1**). Nevertheless, POMONA is able to identify such alignments, and reaches comparable MaxSub scores for the database proteins it selects (**Table 1**). The resulting CS-RosettaCM models for these proteins all converged quite well, with the ten lowest energy models for each of these clustering within 3Å relative to the model with the lowest Rosetta energy (**Table 1**, **Supplementary Fig. D**). However, within this group of relatively small proteins, only for nsp1 is a considerably lower total Rosetta energy obtained compared to simply using CS-Rosetta (**Fig. 1c**, **Supplementary Fig. D**). Therefore, nsp1 is the only protein in this group for which the CS-RosettaCM model is accepted. For nsp1, the only two structural homologues (3zbdA and 3zbdB; sequence identity ~15%) selected by POMONA have MaxSub scores of 0.30 (**Table 1**) and a C^{α}-RMSD of 4.5-5.0 Å (**Fig. 1b**) for their aligned regions. These input templates suffice for enabling CS-RosettaCM to generate all-atom models with a C^{α}-RMSD of ~3.3 Å to the experimental structure for its ordered regions. For the other three proteins, the CS-RosettaCM models have a Rosetta energy that is comparable to those of the *de novo* CS-Rosetta structures. Even though the folds of these CS-RosettaCM models happened to be correct, they could not be accepted as they did not meet the criterion that a substantially lower energy must be reached.

For the other four structural genomics proteins, OR135, HR2876C, sgr145 and MTH1958, DALI identified a substantial number of good structural homologues, with MaxSub scores in the 0.57-0.76 range when considering only database proteins with <20% sequence identity (**Table 1, Supplementary Fig.1**). POMONA also identifies many of these alignments, albeit with lower MaxSub scores (**Table 1**). The low energy RosettaCM structures for these four test proteins all converged to within ~2 Å relative to the structure with the lowest Rosetta energy (**Table 1**, **Supplementary Fig. D**). However, only for proteins sgr145 and MTH1958 does CS-RosettaCM reach energies substantially lower than standard CS-Rosetta (**Supplementary Figs. C** and **D**), allowing these models to be accepted. For the small OR135 and HR2876C proteins (< 90 residues), both CS-Rosetta and CS-RosettaCM generate converged and accurate structures, with comparable Rosetta energies.

The remaining eight proteins in our test set are larger (125 to 370 residues), and standard CS-Rosetta fails to converge. For seven of these, many structural homologues with <20% sequence identity and MaxSub scores in the 0.51-0.83 range are identified by DALI (**Table 1**). POMONA also identifies many of these homologous structures (**Fig. 1b**, **Supplementary Fig. 3**), but only for those without large gaps in their DALI-identified alignment. For example, for sensory rhodopsin-II, most of the 149 DALI-identified structures with < 20% identity show gaps of > 100 residues, and POMONA is only able to identify six structural homologues with the shortest gaps (dots with red circles in **Fig. 1b**). However, this provides an adequate set of starting templates for successful CS-RosettaCM modeling (**Table 1**). Note that substantial structural rearrangements can occur during the CS-RosettaCM modeling stage. For example, starting from the fifth-ranking cluster with a C^{α} RMSD of >7.5Å relative to the reference structure, refined models with a backbone RMSD <4Å are obtained by CS-RosettaCM (light-blue colors in **Fig. 1b,c,** right most panels).

For Mad2, only three suitable structural homologues with <20% sequence identity are found by DALI, all with large gaps in their alignments, and these cannot be identified by POMONA. Therefore, without even a single remote structural homologue among the POMONA-obtained templates, subsequent CS-RosettaCM modeling fails to converge (**Supplementary Fig. D**). Moreover, none of the models generated even reach an energy as low as the unsuccessful CS-Rosetta approach (**Supplementary Figs. C** and **D**), which as expected also fails to converge for this relatively large protein of 196 residues, providing an additional indication that none of the CS-RosettaCM models are of acceptable quality.

The above results demonstrate that the POMONA/CS-RosettaCM protocol performs well, provided that a reasonable structural template can be positively identified by POMONA. In practice, a template with a MaxSub score ≥ 0.3 is needed for successful modeling of all-atom models by CS-RosettaCM. When applying the protocol to a protein of unknown structure, the MaxSub score is not available, and the strict acceptance criteria defined for the CS-RosettaCM approach then are important to ensure correctness of the generated models. Importantly, CS-RosettaCM actually remodels its input template due to the hybrid fragment assembly procedure that is used for both the aligned and unaligned parts of the templates. However, even while this remodeling generally improves the agreement between the template component of the final CS-RosettaCM models and the experimentally determined reference structures, it is insufficient to find or correct the fold of the protein when no adequate structural template is available as input.

The importance of the quality of the input structural templates to CS-RosettaCM is further evaluated by extending the POMONA search to proteins with a sequence identity of up to 30%. Except for nsp1, which has no homologues in the 20-30% range in the database, POMONA identifies virtually all of these more

homologous structures it its highest scoring clusters, ensuring that at least some of these will be used as structural templates by CS-RosettaCM. With this improved template quality, reflected in higher MaxSub scores (**Supplementary Table 1**), CS-RosettaCM then converges for all 16 proteins, yielding improved structural accuracy relative to the experimental structure of the query protein (**Supplementary Fig. E**). Comparison of the structural accuracy obtained with the POMONA/CS-RosettaCM protocol when using only the <20% sequence identity templates with those available when using a 30% cutoff confirms that for about half the proteins a structure closer to the experimentally determined structure is obtained, whereas for the other half the final result remains about equal.

Evaluation of modeling performance when including sparse NOEs

Once backbone assignments have been completed it usually is straightforward to measure and assign a limited number of unambiguous backbone $^{1}H^{N}$-$^{1}H^{N}$ NOEs. The utility of such sparse NOEs is evaluated by randomly selecting for each N-residue protein several sets of N/10 long range H^{N}-H^{N} NOE distance restraints that are ≤ 5 Å in the experimental structure.

Using again the <20% sequence identity cutoff, the sparse NOEs result in improved POMONA alignments (**Supplementary Table 1**) for all test proteins, except for Mad2 which has no suitable template available in the database. For example, for nsp1, its two closest structural homologues did not yield the highest alignment scores when no NOE data were used (purple symbols in **Fig. 1b**, left most panel). With sparse NOEs included, both of these closest structural homologues now fall in the top cluster with the highest score. Subsequent CS-RosettaCM modeling yields somewhat closer agreement to the experimental reference structure (**Supplementary Table 1**), and the same applies for the other proteins.

Often, some of the residues in the query protein for which an NOE is available will be aligned by POMONA to a gap in the database protein sequence, in which case this sparse NOE will only help in restraining the sampling for such unaligned parts during the CS-RosettaCM modeling procedure, yielding substantial improvement in both convergence and accuracy of the final models (**Fig. 1d**; **Supplementary Figs. F** and **G**, **Supplementary Table 1**). In cases where the target sequence corresponds to an extreme variant of a known fold, the protocol still permits substantial reorganization of the long template fragments to accommodate these structural differences (such as β-sheet register shifts), guided by the sparse NOEs, thereby offering further improvement of the final models.

SUPPLEMENTARY REFERENCE

[1] M. Berjanskii, D. S. Wishart, *Nat. Protoc.* **1**, 683-688 (2006).