Supplementary Information for

# Structural discrimination analysis for constraint selection in protein modeling

**Guilherme F. Bottino,[1,2] Allan J. R. Ferrari,[1,2] Fabio C. Gozzo,[1] and Leandro Martínez[1,2,*]**

[1]Institute of Chemistry and [2]Center for Computational Engineering & Science, University of Campinas
Campinas, SP, Brazil

*To whom correspondence should be addressed: lmartine@unicamp.br

**Index**

## Supporting Section 1

## Comments on Individual Targets

Text ST1:

The success of modeling experiments on target 1C75 was notably limited, even after constraint selection, despite showing marginally positive results after the application of $r_{pb}$ as a selection tool. We believe this is mostly due to the fact that the preliminary constraint set is of very low true-positive rate, and the selected consensus model and the best model are far from what could be considered of fold-quality. This may be surprising, since diversity $N_{eff}$ for this target is the second-highest. However, we must note that, out of the whole 131 aminoacids of the 1C75A domain, only 70% - residues 1 through 90 - map to the Cytochrom_C (PF00034) family. Also, despite the high diversity on some columns of the MSA, only a few actual direct couplings could be determined. The estimated constraint set is extremely noisy and sparse, posing a challenge to the method, and the difficulty is aggravated by a fragment library with a very low proportion of homologues. We decided to model it and report its results despite those observations, and reinforce that the increase in the modeling quality and constraint set true positive rate, although marginal, is positive.

As for target 1E6K, the situation was different. All experiments (before and after constraint selection, and also the all-native control) portrayed high success rates and little to no difference between the results, although we did remark, again, a positive marginal improvement after the application of $r_{pb}$. Upon inspecting the fragment library for 1E6K, we realized that there was strong contamination of both close and remote homologues (notably 1JBE, 1K66 and 1S8N, which also map to the PF00072 family). We performed an unconstrained modeling round and noted that even this unconstrained experiment had very good performance. After realizing this, however, we decided to keep the target in the main article, because it shows an interesting fact: **when the problem is solved sufficiently well by the fragment library and the fragment-based folding protocol, the constraint selection using $r_{pb}$ does not spoil an already good result.** The property of not worsening an unconstrained protocol is, of course, desirable to a constraint selector.

Supplementary figures for section 1:

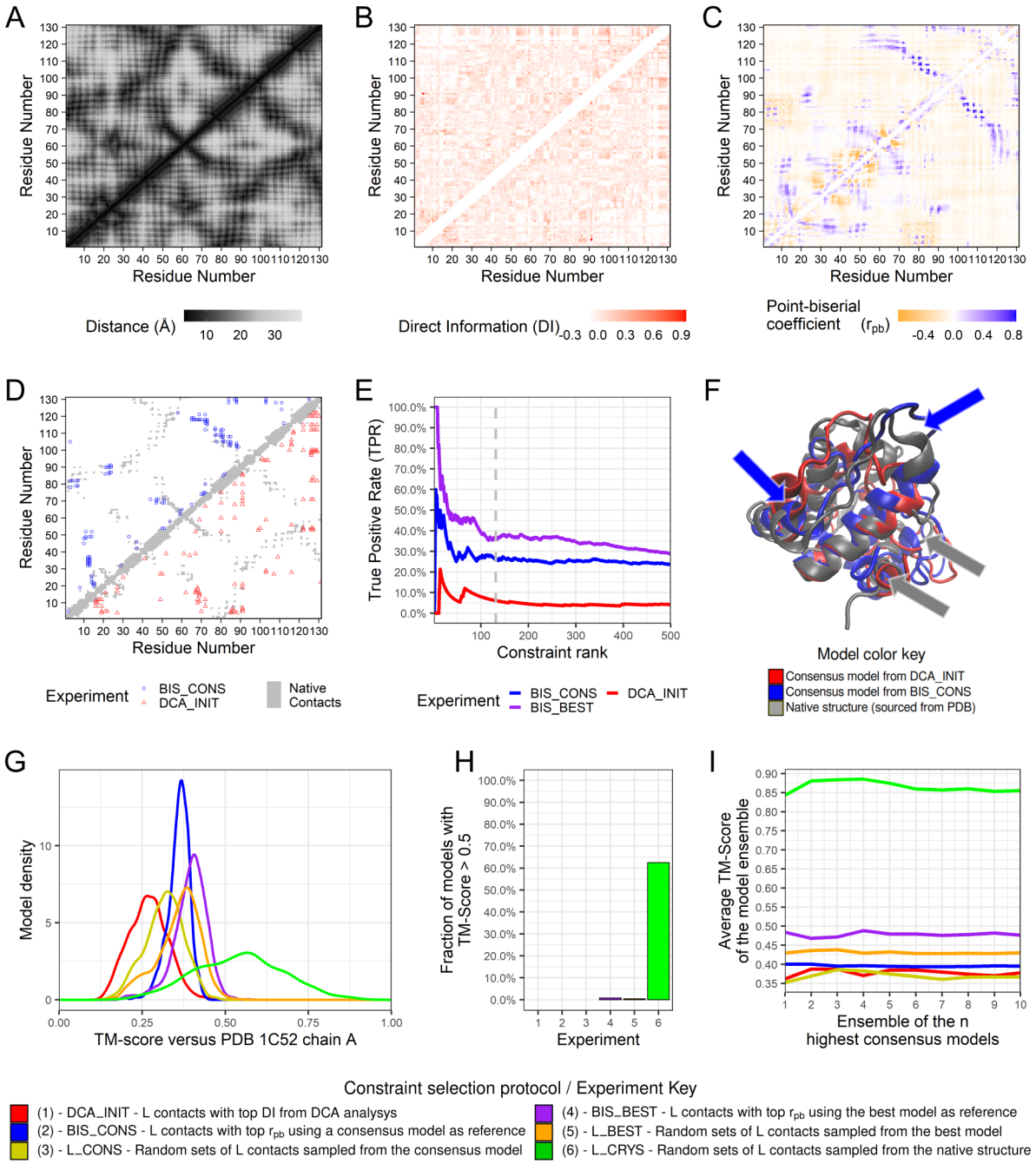


Figure SF1.1. Modeling results for target PDB_1C52_A. (A) Native distance map for target crystallographic structure. (B) Estimated DI map from DCA analysis on Target family MSA. (C) Rescored point-biserial correlation map from BIS_CONS round. (D) Selected constraints for DCA_INIT and BIS_CONS modeling rounds. (E) Cumulative True Positive Rates for sorted constraints in the DCA_INIT, BIS_CONS and BIS_BEST experiments. (F) Structural alignment of the target native structure with the consensus models elected after DCA_INIT and BIS_CONS rounds. Blue arrows indicate regions where topology improved after constraint selection; Gray arrows indicate regions wrongly sampled by reasons discussed in subsection 4.3. (G) Distribution of model qualities measured through TM-score against crystallographic structure for each modeling experiment. (H) Proportion of models with correct topology (TM-score of alignment with native structure > 0.5) for each experiment. (I) Average TM-score of the ensemble with n models recovered after the last round of each modeling experiment.

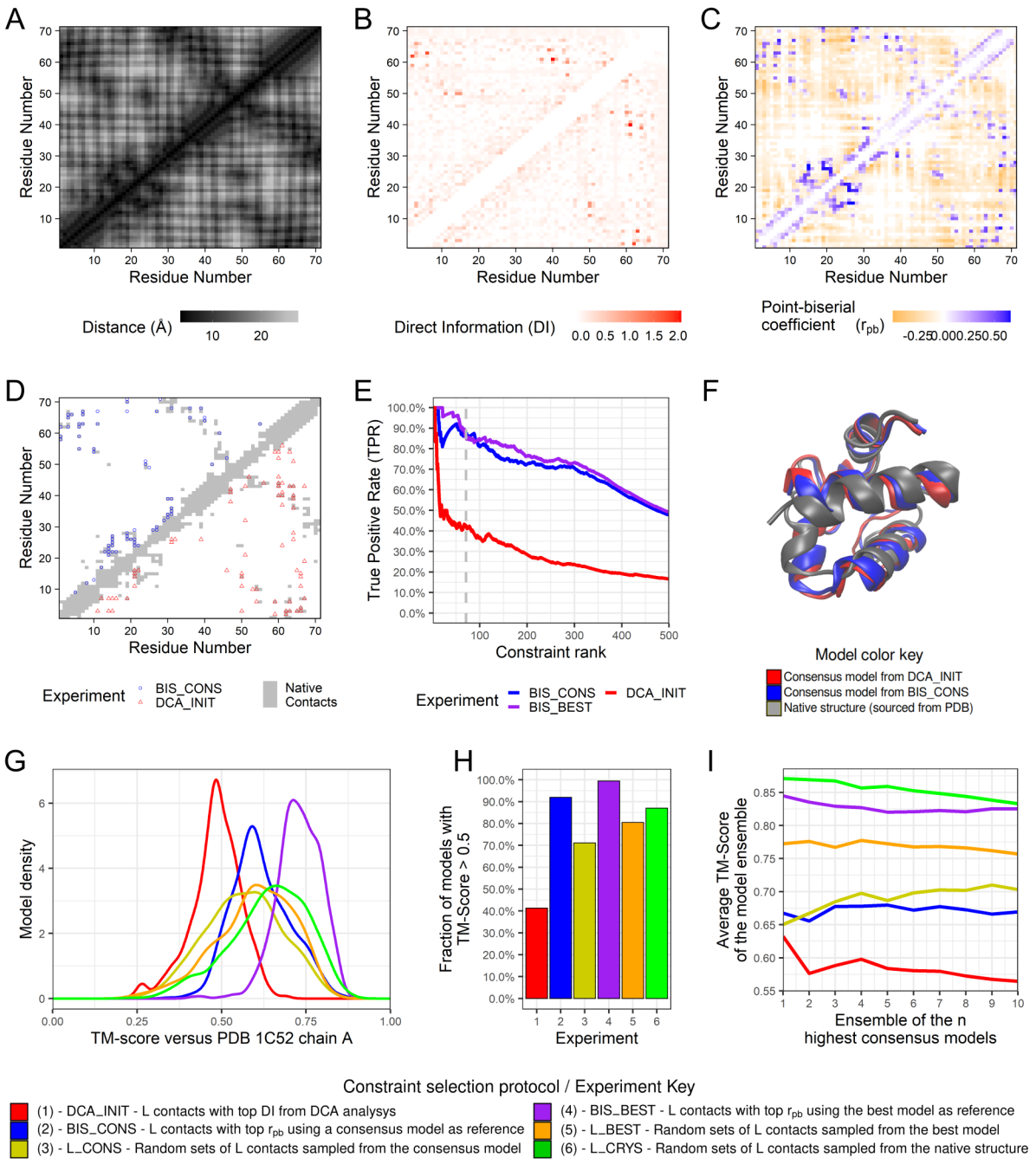


Figure SF1.2. Modeling results for target PDB_1C75_A. (A) Native distance map for target crystallographic structure. (B) Estimated DI map from DCA analysis on Target family MSA. (C) Rescored point-biserial correlation map from BIS_CONS round. (D) Selected constraints for DCA_INIT and BIS_CONS modeling rounds. (E) Cumulative True Positive Rates for sorted constraints in the DCA_INIT, BIS_CONS and BIS_BEST experiments. (F) Structural alignment of the target native structure with the consensus models elected after DCA_INIT and BIS_CONS rounds. Blue arrows indicate regions where topology improved after constraint selection; Gray arrows indicate regions wrongly sampled by reasons discussed in subsection 4.3. (G) Distribution of model qualities measured through TM-score against crystallographic structure for each modeling experiment. (H) Proportion of models with correct topology (TM-score of alignment with native structure > 0.5) for each experiment. (I) Average TM-score of the ensemble with n models recovered after the last round of each modeling experiment.

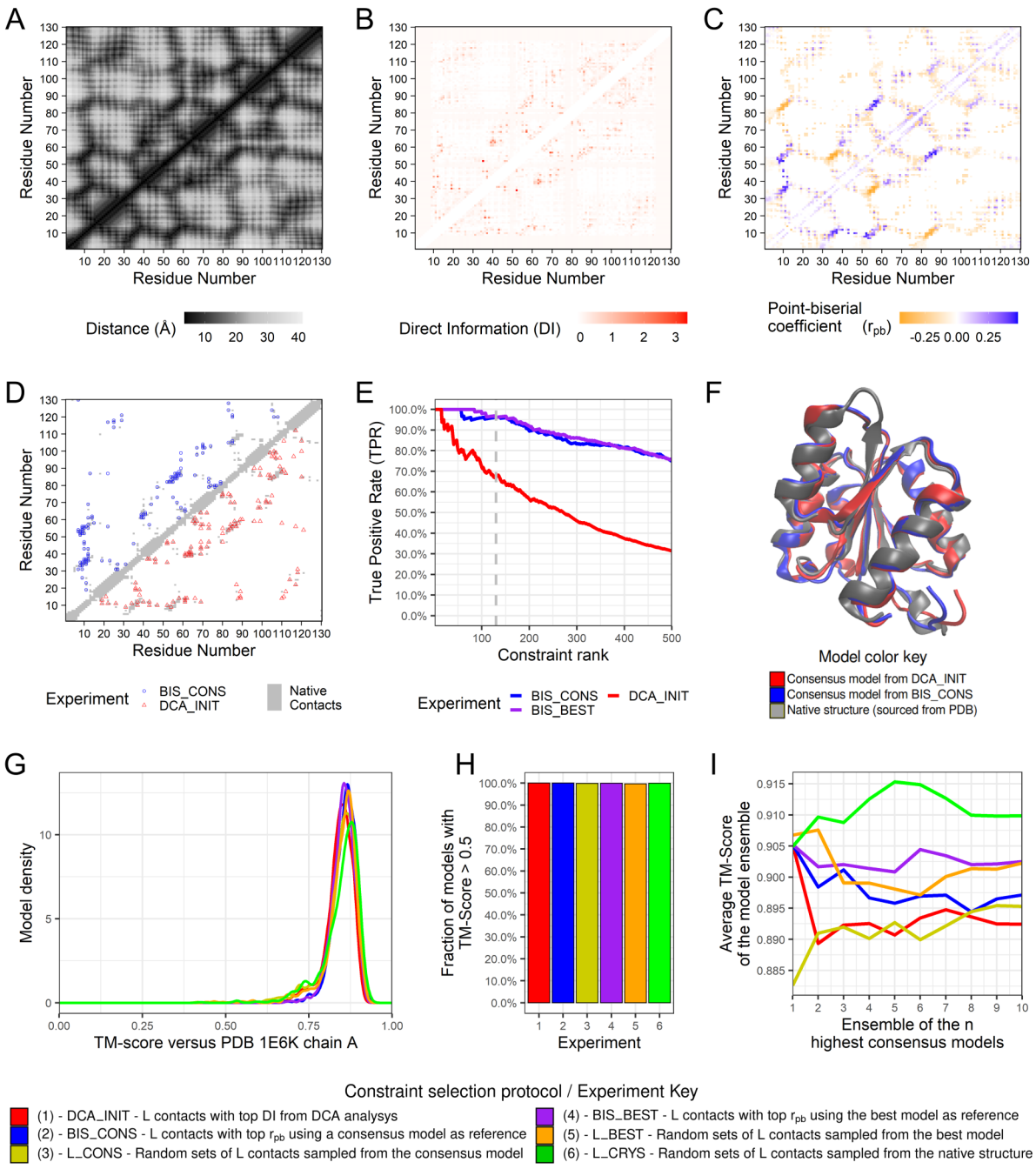


Figure SF1.3. Modeling results for target PDB_1E6K_A. (A) Native distance map for target crystallographic structure. (B) Estimated DI map from DCA analysis on Target family MSA. (C) Rescored point-biserial correlation map from BIS_CONS round. (D) Selected constraints for DCA_INIT and BIS_CONS modeling rounds. (E) Cumulative True Positive Rates for sorted constraints in the DCA_INIT, BIS_CONS and BIS_BEST experiments. (F) Structural alignment of the target native structure with the consensus models elected after DCA_INIT and BIS_CONS rounds. Blue arrows indicate regions where topology improved after constraint selection; Gray arrows indicate regions wrongly sampled by reasons discussed in subsection 4.3. (G) Distribution of model qualities measured through TM-score against crystallographic structure for each modeling experiment. (H) Proportion of models with correct topology (TM-score of alignment with native structure > 0.5) for each experiment. (I) Average TM-score of the ensemble with n models recovered after the last round of each modeling experiment.

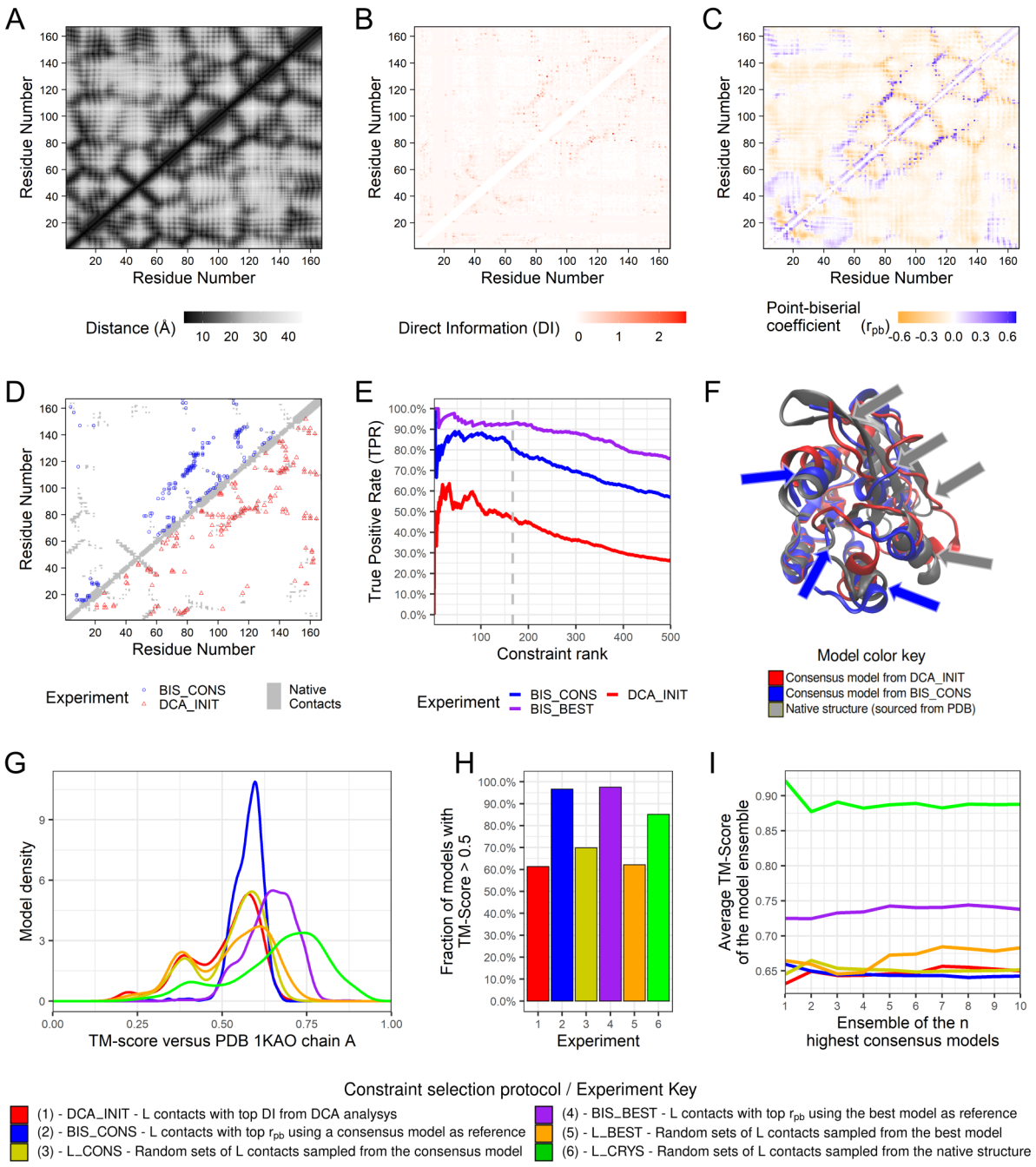


Figure SF1.4. Modeling results for target PDB_1KAO_A. (A) Native distance map for target crystallographic structure. (B) Estimated DI map from DCA analysis on Target family MSA. (C) Rescored point-biserial correlation map from BIS_CONS round. (D) Selected constraints for DCA_INIT and BIS_CONS modeling rounds. (E) Cumulative True Positive Rates for sorted constraints in the DCA_INIT, BIS_CONS and BIS_BEST experiments. (F) Structural alignment of the target native structure with the consensus models elected after DCA_INIT and BIS_CONS rounds. Blue arrows indicate regions where topology improved after constraint selection; Gray arrows indicate regions wrongly sampled by reasons discussed in subsection 4.3. (G) Distribution of model qualities measured through TM-score against crystallographic structure for each modeling experiment. (H) Proportion of models with correct topology (TM-score of alignment with native structure > 0.5) for each experiment. (I) Average TM-score of the ensemble with n models recovered after the last round of each modeling experiment.

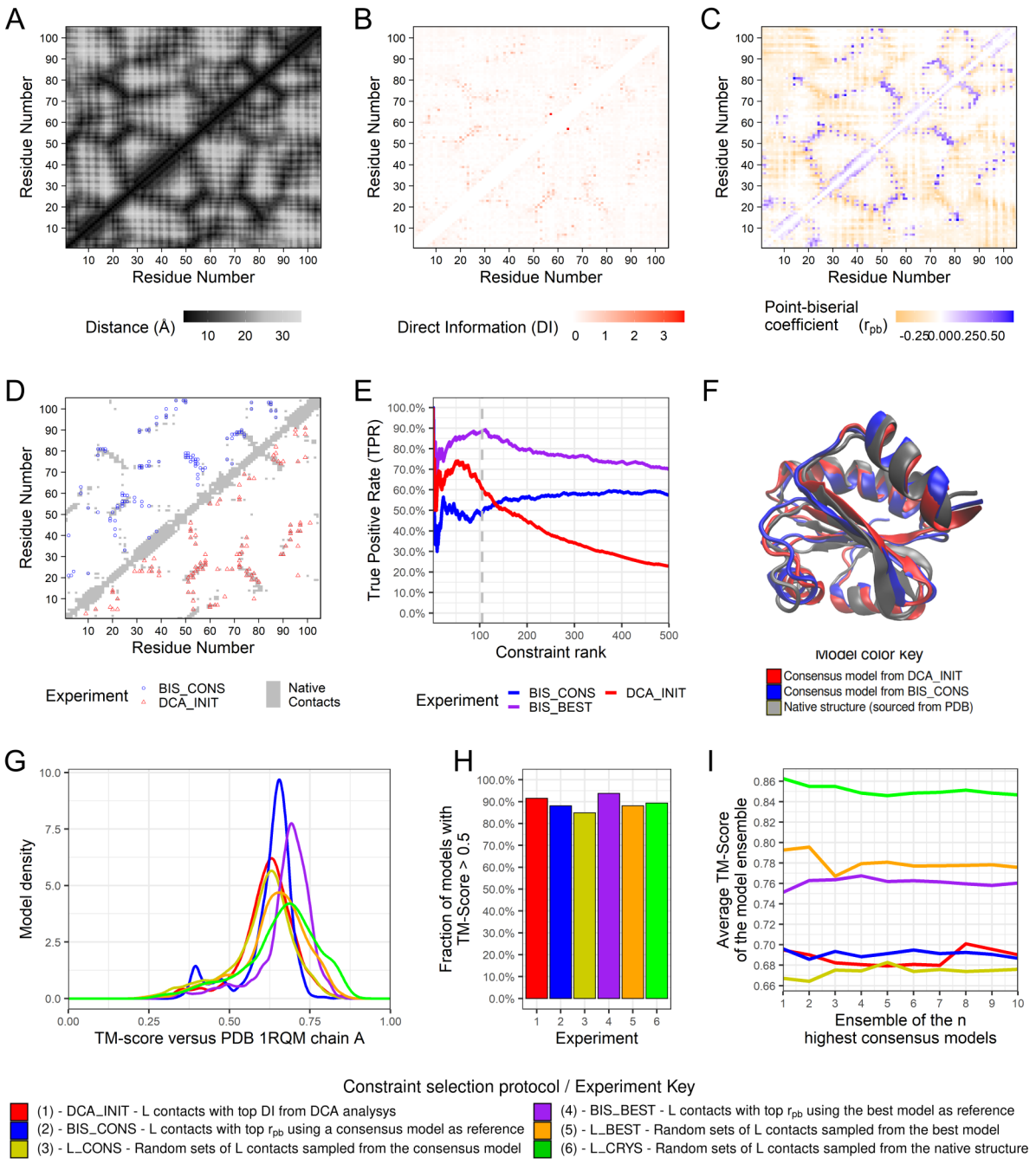


Figure SF1.5. Modeling results for target PDB_1RQM_A. (A) Native distance map for target crystallographic structure. (B) Estimated DI map from DCA analysis on Target family MSA. (C) Rescored point-biserial correlation map from BIS_CONS round. (D) Selected constraints for DCA_INIT and BIS_CONS modeling rounds. (E) Cumulative True Positive Rates for sorted constraints in the DCA_INIT, BIS_CONS and BIS_BEST experiments. (F) Structural alignment of the target native structure with the consensus models elected after DCA_INIT and BIS_CONS rounds. Blue arrows indicate regions where topology improved after constraint selection; Gray arrows indicate regions wrongly sampled by reasons discussed in subsection 4.3. (G) Distribution of model qualities measured through TM-score against crystallographic structure for each modeling experiment. (H) Proportion of models with correct topology (TM-score of alignment with native structure > 0.5) for each experiment. (I) Average TM-score of the ensemble with n models recovered after the last round of each modeling experiment.

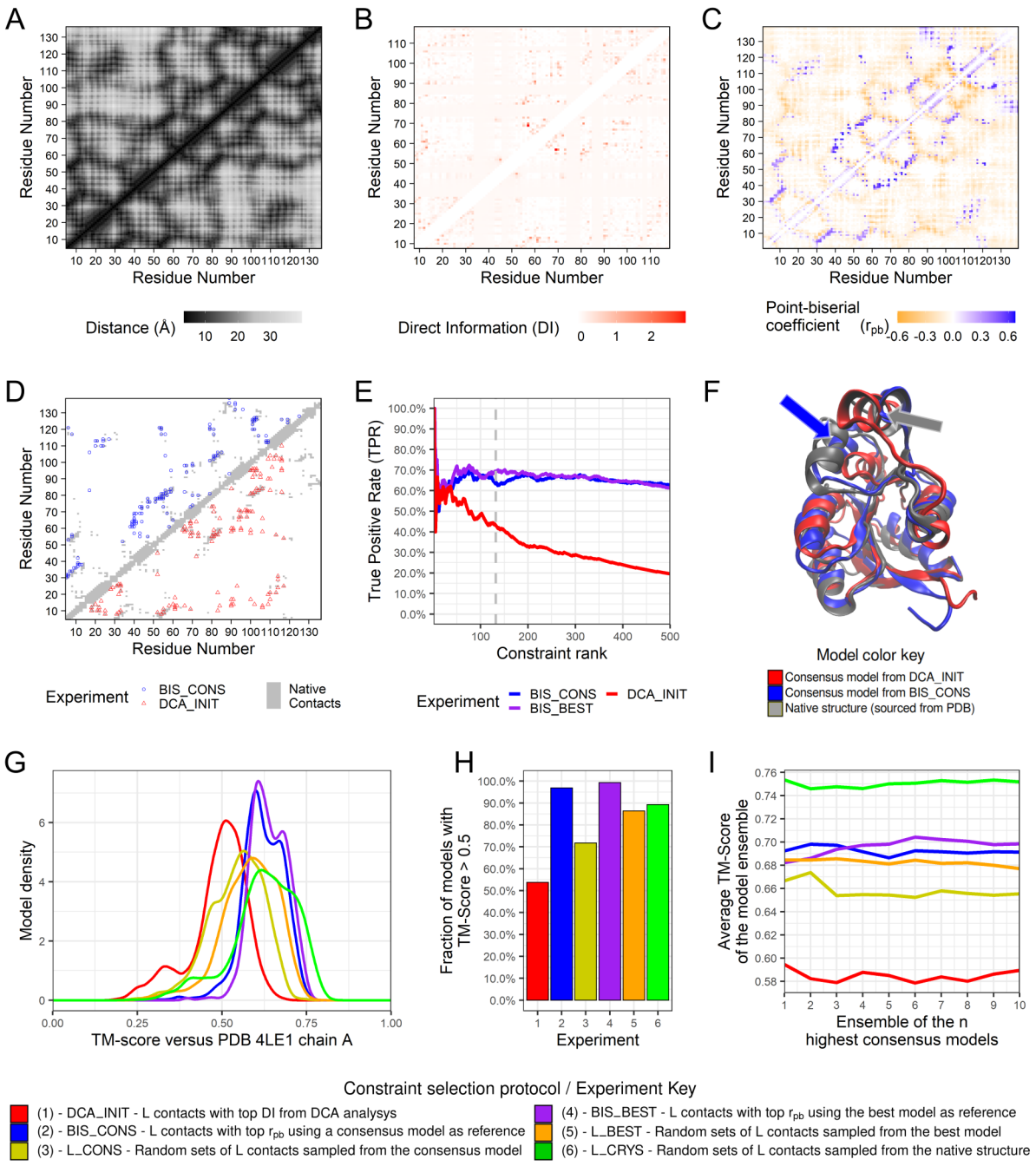


Figure SF1.6. Modeling results for target PDB_4LE1_A. (A) Native distance map for target crystallographic structure. (B) Estimated DI map from DCA analysis on Target family MSA. (C) Rescored point-biserial correlation map from BIS_CONS round. (D) Selected constraints for DCA_INIT and BIS_CONS modeling rounds. (E) Cumulative True Positive Rates for sorted constraints in the DCA_INIT, BIS_CONS and BIS_BEST experiments. (F) Structural alignment of the target native structure with the consensus models elected after DCA_INIT and BIS_CONS rounds. Blue arrows indicate regions where topology improved after constraint selection; Gray arrows indicate regions wrongly sampled by reasons discussed in subsection 4.3. (G) Distribution of model qualities measured through TM-score against crystallographic structure for each modeling experiment. (H) Proportion of models with correct topology (TM-score of alignment with native structure > 0.5) for each experiment. (I) Average TM-score of the ensemble with n models recovered after the last round of each modeling experiment.

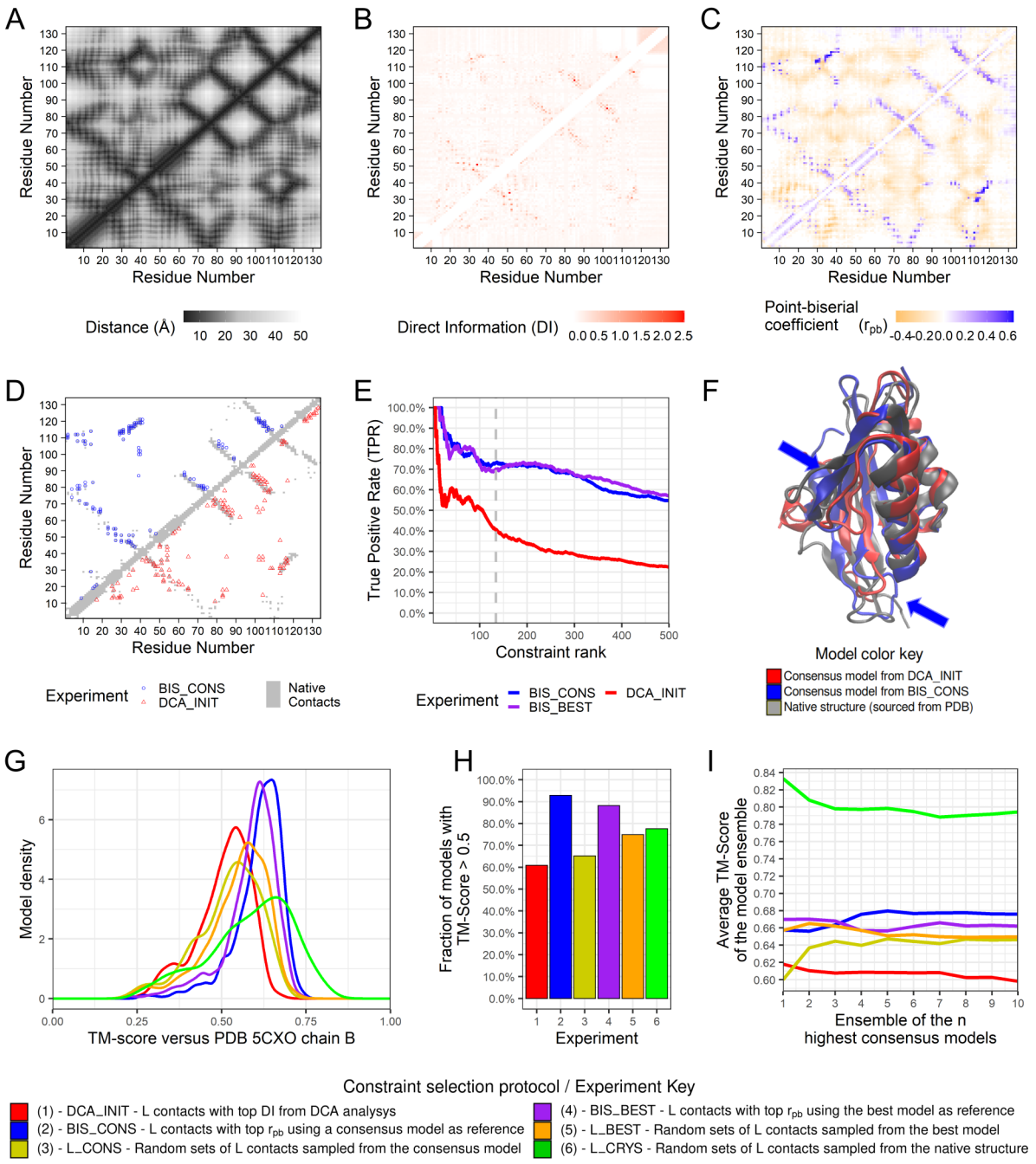


Figure SF1.7. Modeling results for target PDB_5CXO_B. (A) Native distance map for target crystallographic structure. (B) Estimated DI map from DCA analysis on Target family MSA. (C) Rescored point-biserial correlation map from BIS_CONS round. (D) Selected constraints for DCA_INIT and BIS_CONS modeling rounds. (E) Cumulative True Positive Rates for sorted constraints in the DCA_INIT, BIS_CONS and BIS_BEST experiments. (F) Structural alignment of the target native structure with the consensus models elected after DCA_INIT and BIS_CONS rounds. Blue arrows indicate regions where topology improved after constraint selection; Gray arrows indicate regions wrongly sampled by reasons discussed in subsection 4.3. (G) Distribution of model qualities measured through TM-score against crystallographic structure for each modeling experiment. (H) Proportion of models with correct topology (TM-score of alignment with native structure > 0.5) for each experiment. (I) Average TM-score of the ensemble with n models recovered after the last round of each modeling experiment.

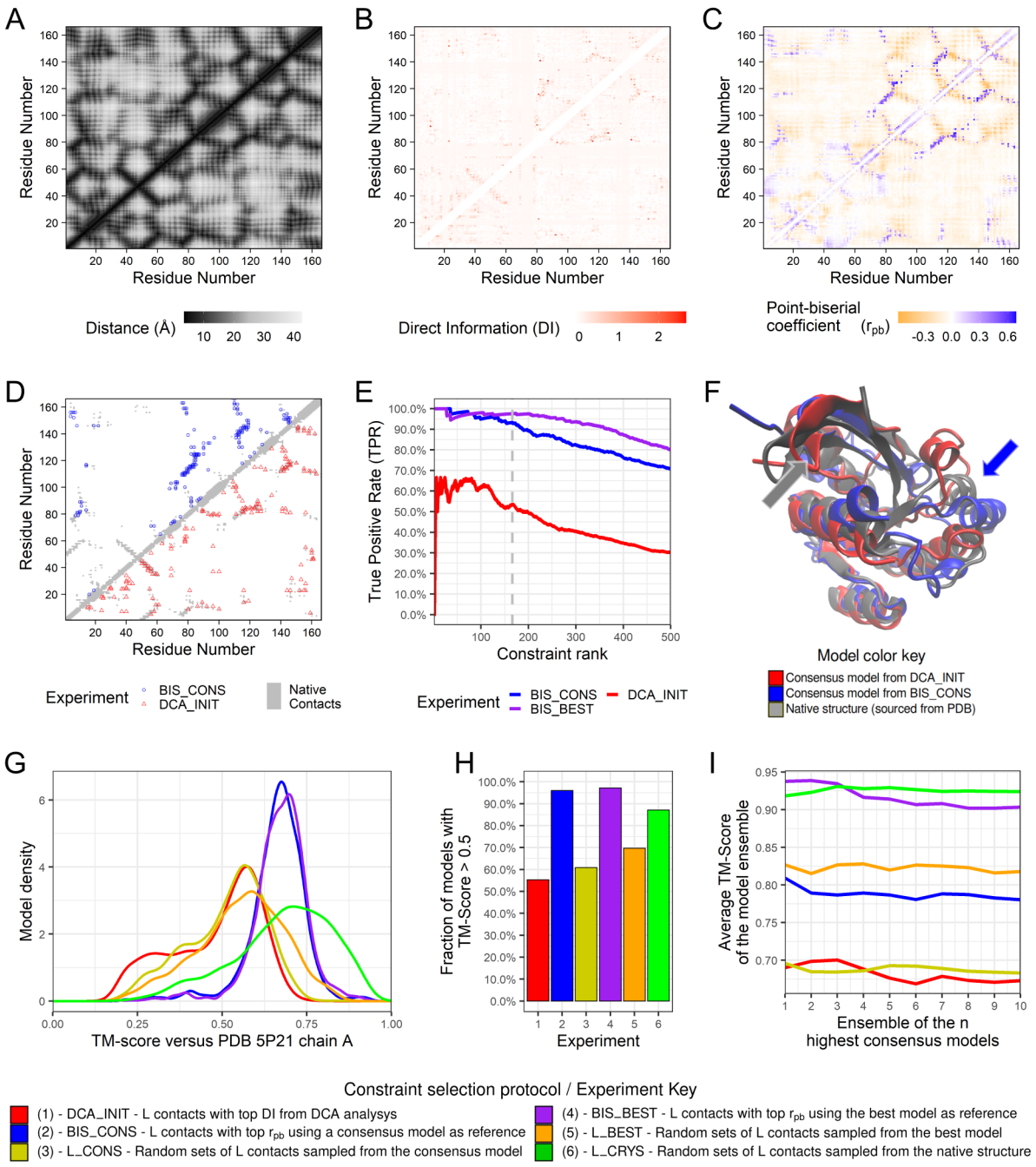


Figure SF1.8. Modeling results for target PDB_5P21_A. (A) Native distance map for target crystallographic structure. (B) Estimated DI map from DCA analysis on Target family MSA. (C) Rescored point-biserial correlation map from BIS_CONS round. (D) Selected constraints for DCA_INIT and BIS_CONS modeling rounds. (E) Cumulative True Positive Rates for sorted constraints in the DCA_INIT, BIS_CONS and BIS_BEST experiments. (F) Structural alignment of the target native structure with the consensus models elected after DCA_INIT and BIS_CONS rounds. Blue arrows indicate regions where topology improved after constraint selection; Gray arrows indicate regions wrongly sampled by reasons discussed in subsection 4.3. (G) Distribution of model qualities measured through TM-score against crystallographic structure for each modeling experiment. (H) Proportion of models with correct topology (TM-score of alignment with native structure > 0.5) for each experiment. (I) Average TM-score of the ensemble with n models recovered after the last round of each modeling experiment.

**Supporting Section 2**

**Coupling point-biserial selection with deep-learning contacts**

Text ST2:

In the main article, we defend that the principle of using discrimination analysis to enrich the constraint set is universal and could be applied modularly to different kinds of experiments. In our work, we employed DCA analysis for estimating a starting constraint set due to its simplicity, but we also acknowledge that the state-of-art in constraint-assisted Protein Structure Prediction is the use of deep-learning based approaches, as per the most recent CASP iterations. With that in mind, we wanted to illustrate with a practical example that it is possible to couple our idea of constraint selection with this kind of estimation.

The target we elected for this demonstration was the CASP13 target T0968s2 (L=116). The reasons for electing this target were: (1) That it is a mostly-beta protein, accounting for the underrepresentation of beta targets in the main article target list; (2) That its crystallographic structure is already determined and deposited in the Protein Data Bank, corresponding to Chain B of PDB 6CP9 (Gucinski *et al.*, 2019) with a high degree of coverage of its primary sequence and enough resolution; (3) That it is compatible with the length of the other targets in this study.

For the starting constraint set, instead of utilizing DCA as we did on the main paper, we recovered from the CASP Results page the predictions made by one of the groups in the CASP13 Contact Prediction rankings, RaptorX-Contact (group 498). We chose RaptorX (Wang *et al.*, 2017) because it was one of the top three groups and we were able to verify the availability of its online server for community use. The predictions file, T0968s2RR498_1-D1, shows the following figures of merit for medium and long-range contacts at an L-length list: F1-score = 56.250, Precision = 70.430; Recall = 46.820. We sorted the list of contacts by their probability, without excluding the short contacts, and extracted the first L of them for the preliminary modeling, following our protocol from that point on. Due to time constraints, only one round of our incremental constraint selection strategy was performed.

Results for PDB_6CP9_B are available in the following figure SF2.1 and table ST2.1. Figure SF2.1A shows that, along with the parallel relationships between the various beta-strands, there are other regions of contact relevant to the model topology, especially between the N-terminus and the globular core of the protein. Inspection of Figure SF2.1B shows that the initial predictions capture only partially these structural motifs, failing to sample the long-range constraints relevant to the structure. Selection of a consensus model on this set and application of constraint selection via $r_{pb}$ on the preliminary model pool gives rise to Figure SF2.1C, where we can see a strong improvement of two long-range structurally-relevant regions (circled in green). Comparison of the constraint sets before and after constraint selection Figure SF2.1D shows an increased population of constraints in these aforementioned regions, which concurs with the general observations we made in the main article. However, the true-positive rate of the constraint set after selection was lower (both in the BIS_CONS

and BIS_BEST cases) than in the starting deep-learning estimated set, as portrayed in Figure SF2.1E.

As for the output models themselves, inspection of the consensus model before and after constraint selection (Figure SF2.1F) shows that there was a major improvement in the topology of the first residues closer to the N-terminus, mostly thanks to detection of structurally-discriminating contacts between those residues and the rest of the protein (underrepresented in the initial predictions). When looking at model quality distributions in Figure SF2.1G, there is a noticeable shift in the density of high-quality models on the experiments utilizing $r_{pb}$ as selection criterion (BIS_CONS and BIS_BEST), who show large peaks with maxima around a TM-score 0.58.

This also reflects upon the proportion of correct models (Figure SF2.1H): as with other targets in the main paper, despite the quality of the preliminary constraint set, the proportion of models with a TM-score larger than 0.5 before constraint selection was around 30%, but increases to 60% and 70% on both experiments with $r_{pb}$-based constraint selection. To understand how it is possible for constraint sets with less true-positives to perform better than other ones with more true-positives (even 100% native contacts), please refer to section 4.5 of the main article. Finally, constraint selection allowed for the recovery of a better model ensemble, as illustrated by Figure Figure SF2.1I. Further results to reproduce the discussions made in sections 4.4 and 4.5 of the main article are on table SL2.1. We remark that the BIS sets again reach the highest average constraint set information, and portray the highest amount of medium+long range contacts, compared to the other sets.

With this example, we showed how our idea can be applied to another kind of preliminary constraint estimation, In fact, along with other desirable points for constraint selection highlighted in sections 2 and 3.4 of the main article, we believe that the capacity of being inserted as an “additional module” or “another piece of the puzzle” in virtually any Protein Structure Determination protocol is another requisite for any constraint selection strategy. Again, using point-biserial correlation the way we did in this work, due to its simplicity, is a viable starting point.

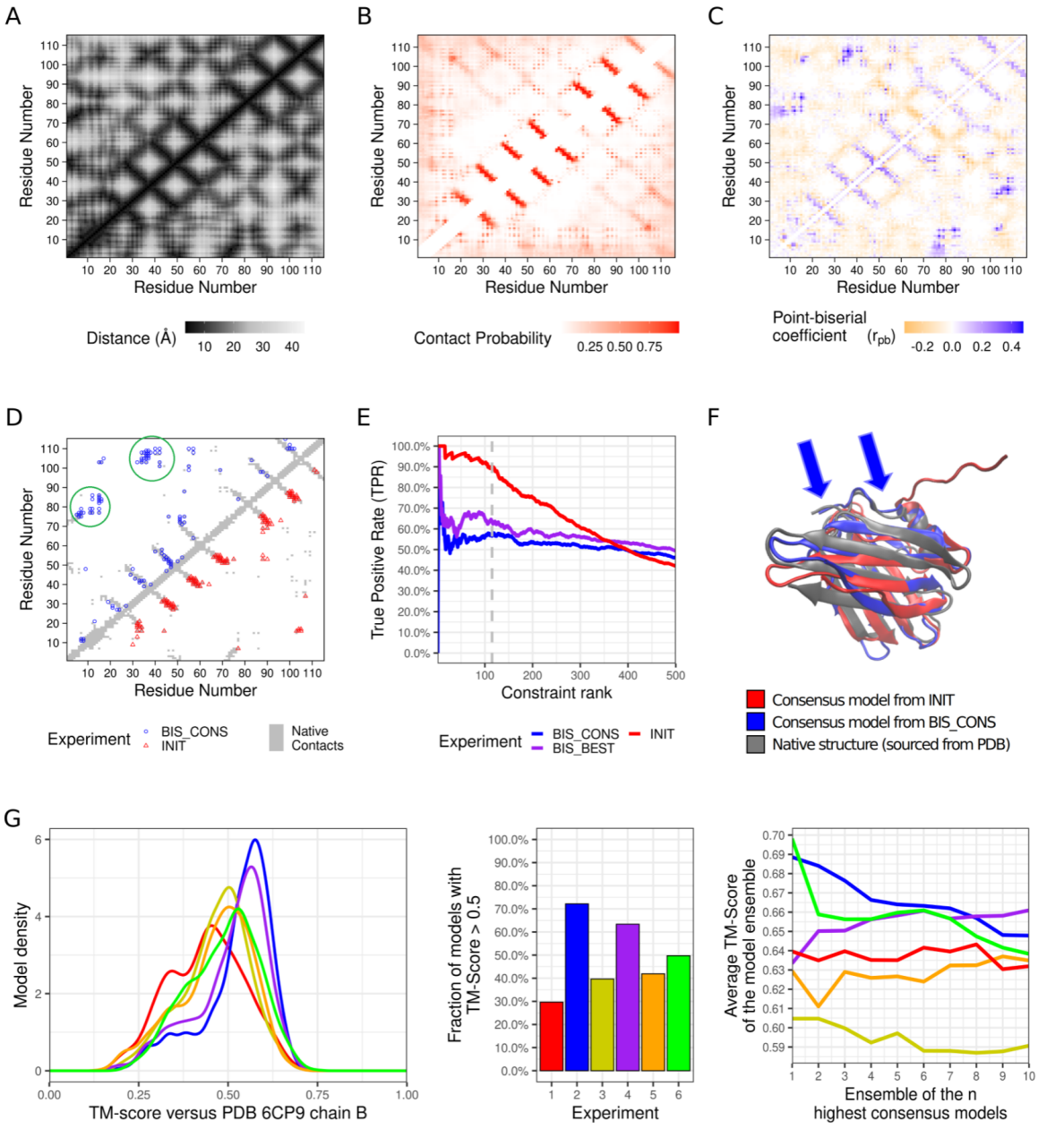


Figure SF2.1. Modeling results for target PDB_6CP9_B. (A) Native distance map for target crystallographic structure. (B) Contact probability map estimated by RaptorX-Contact, obtained from CASP13 results page, file T0968s2RR498_1-D1. (C) Rescored point-biserial correlation map after one round of BIS_CONS. (D) Selected constraints for modeling, before (INIT) and after (BIS_CONS) constraint selection. (E) Cumulative True Positive Rates for sorted constraints in the INIT, BIS_CONS and BIS_BEST experiments. (F) Structural alignment of the target native structure with the consensus models elected before (INIT) and after (BIS_CONS) constraint selection. Blue arrows indicate regions where topology improved after constraint selection. (G) Distribution of model qualities measured through TM-score against crystallographic structure for each modeling experiment. (H) Proportion of models with correct topology (TM-score of alignment with native structure > 0.5) for each experiment. (I) Average TM-score of the ensemble with n models recovered after each modeling experiment.

**Table SL2.1.** Figures of Merit for the PDB_6CP9_B example.

| Experiment | True Positive Rate | Proportion of correct models | Average TM-score of n=10 ensemble | Average Information (bits) | Proportion of constraints by range | | Contribution to TPR on each range category |
|---|---|---|---|---|---|---|---|
| INIT | 0.896 | 0.296 | 0.632 | 5.85 | short | 0.517 | 0.498 |
| | | | | | medium +long | 0.486 | 0.398 |
| BIS_CONS | 0.569 | 0.722 | 0.648 | 6.68 | short | 0.362 | 0.293 |
| | | | | | medium +long | 0.638 | 0.276 |
| L_CONS | 0.794 | 0.396 | 0.591 | 4.17 | short | 0.707 | 0.628 |
| | | | | | medium +long | 0.293 | 0.166 |
| BIS_BEST | 0.629 | 0.633 | 0.661 | 6.76 | short | 0.336 | 0.301 |
| | | | | | medium +long | 0.664 | 0.328 |
| L_BEST | 0.776 | 0.418 | 0.635 | 4.24 | short | 0.673 | 0.609 |
| | | | | | medium +long | 0.327 | 0.167 |
| L_CRYS | 1.000 | 0.497 | 0.638 | 4.52 | short | 0.687 | 0.687 |
| | | | | | medium +long | 0.313 | 0.313 |

**References**


Gucinski,G.C. *et al.* (2019) Convergent Evolution of the Barnase/EndoU/Colicin/RelE (BECR) Fold in Antibacterial tRNase Toxins. *Structure*, **27**, 1660–1674.e5.

Wang,S. *et al.* (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.*, **13**, e1005324.