SUPPORTING INFOMATION

Improved Structure Sampling Strategies for Protein Model Refinement based on Molecular Dynamics Simulation

Lim Heo, Collin F. Arbour, Giacomo Janson, and Michael Feig*

Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA

*corresponding author:
603 Wilson Road, Room 218 BCH
East Lansing, MI 48824, USA
mfeiglab@gmail.com
+1-517-432-7439

Running Title: *Refinement via Molecular Dynamics*

SUPPLEMENTARY FIGURES

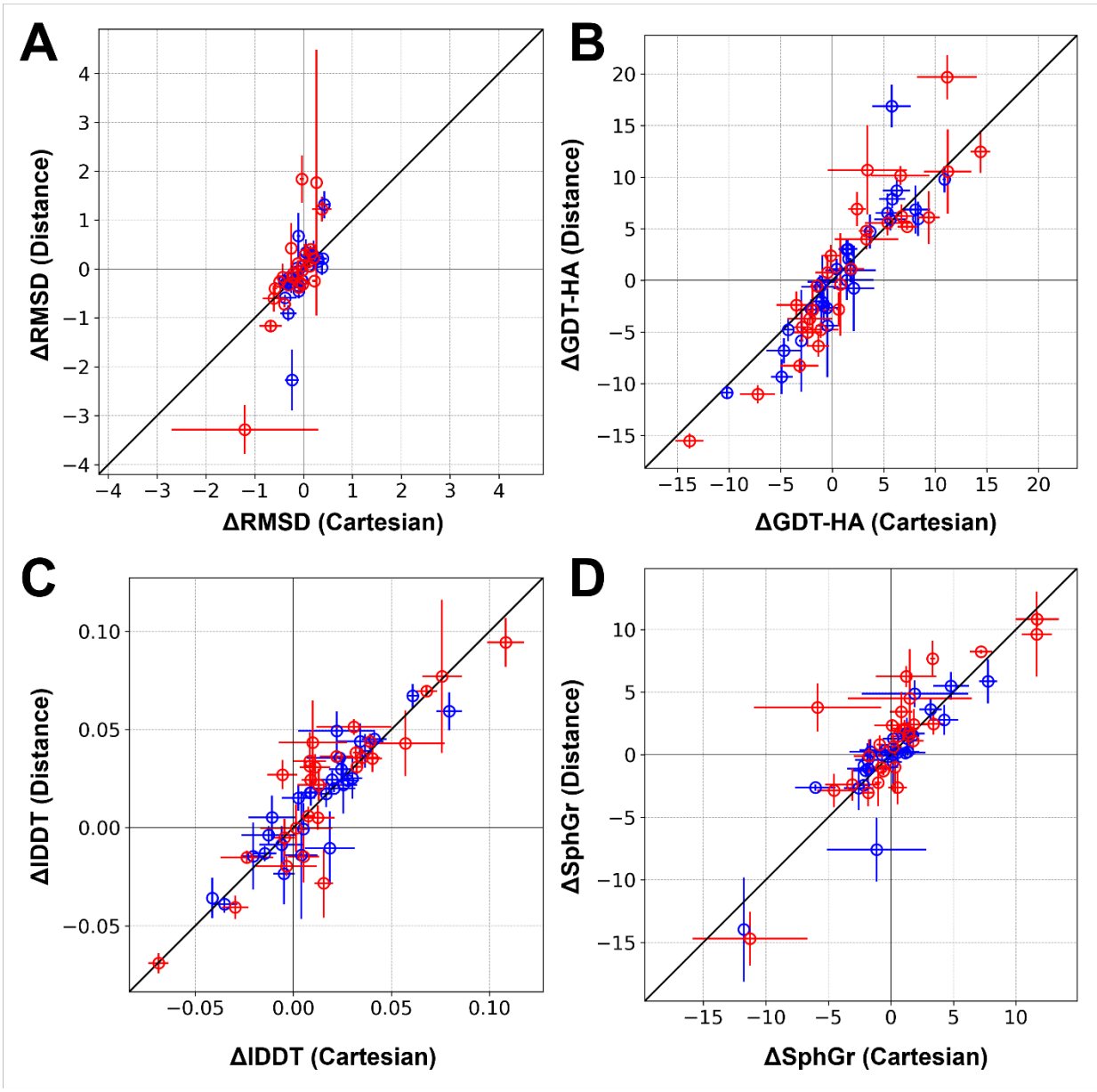

Figure S1. Refinement performance comparisons between refinements with different types of restraints. Model quality changes were compared in RMSD (A), GDT-HA (B), lDDT (C), and SphereGrinder (D). Refinement results at 298.15 K and 360 K are shown in blue and red, respectively.

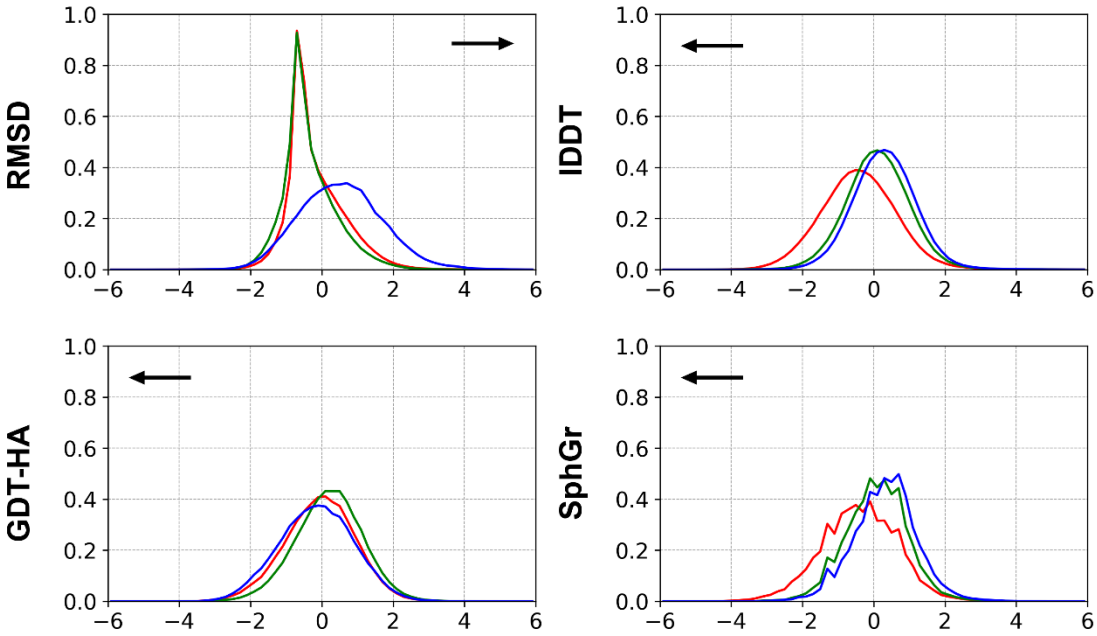

Figure S2. Distribution of sampled structures' structural similarity with respect to the initial model with different restraint types. The distributions represent target-averaged probability distributions in Z-score for each measure. Z-score is used for the analysis to get a target-averaged distribution. The distributions for sampling with Cartesian, combined, and distance restraints are shown in red, green, and blue, respectively. The direction of reduced accuracy based on each measure relative to the values of the initial structure is indicated with black arrows.

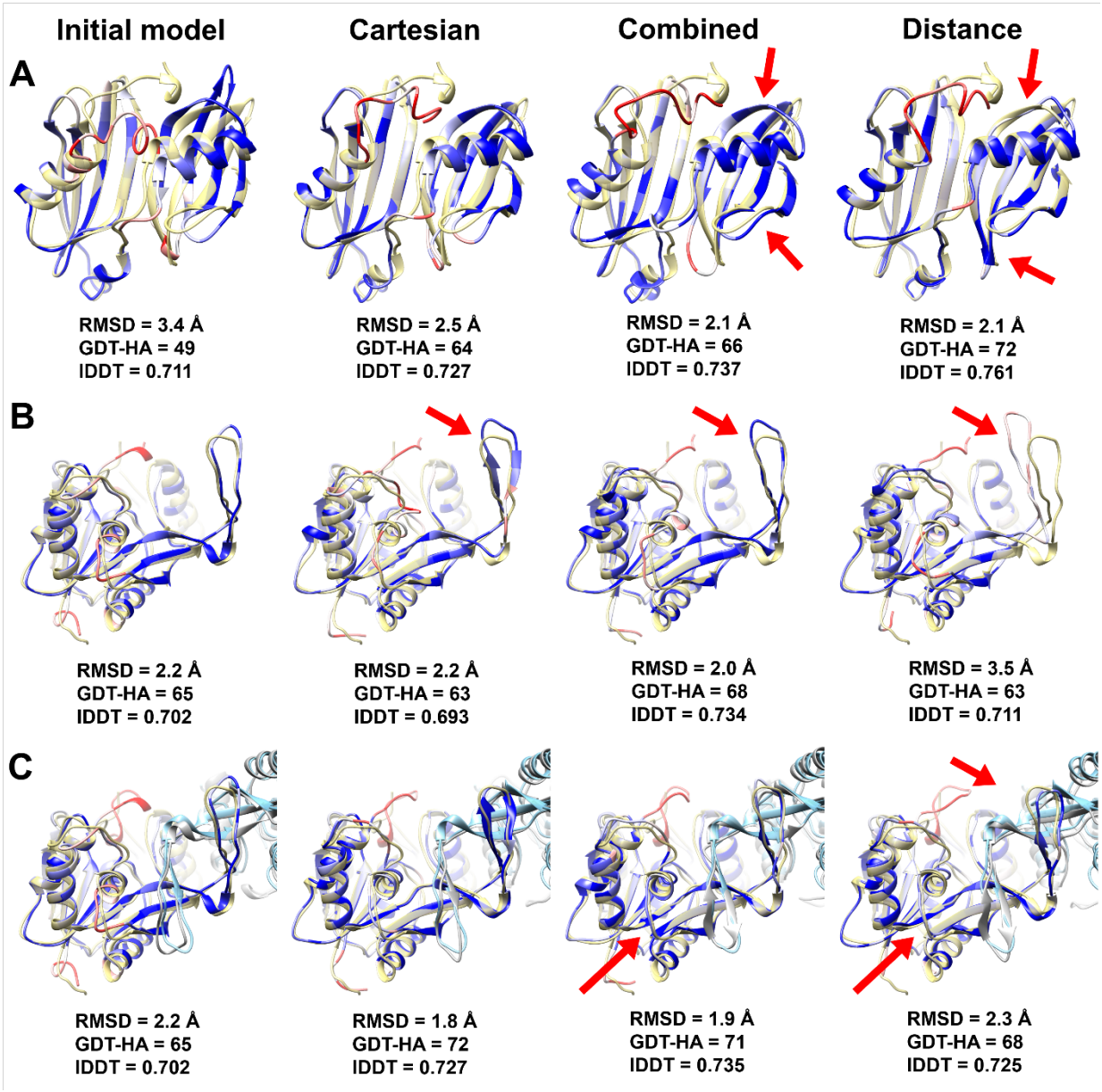

Figure S3. Examples of refinements with different types of restraints. (A) TR663, (B) TR699, and (C) TR699 as dimer. Model structures are shown in cartoon diagrams with blue-to-red colors. The colors represent residuewise lDDT score; blue and red are for high (> 0.8) to low (< 0.3) scores. The experimental structures are overlaid in transparent yellow and light blue cartoon diagram.

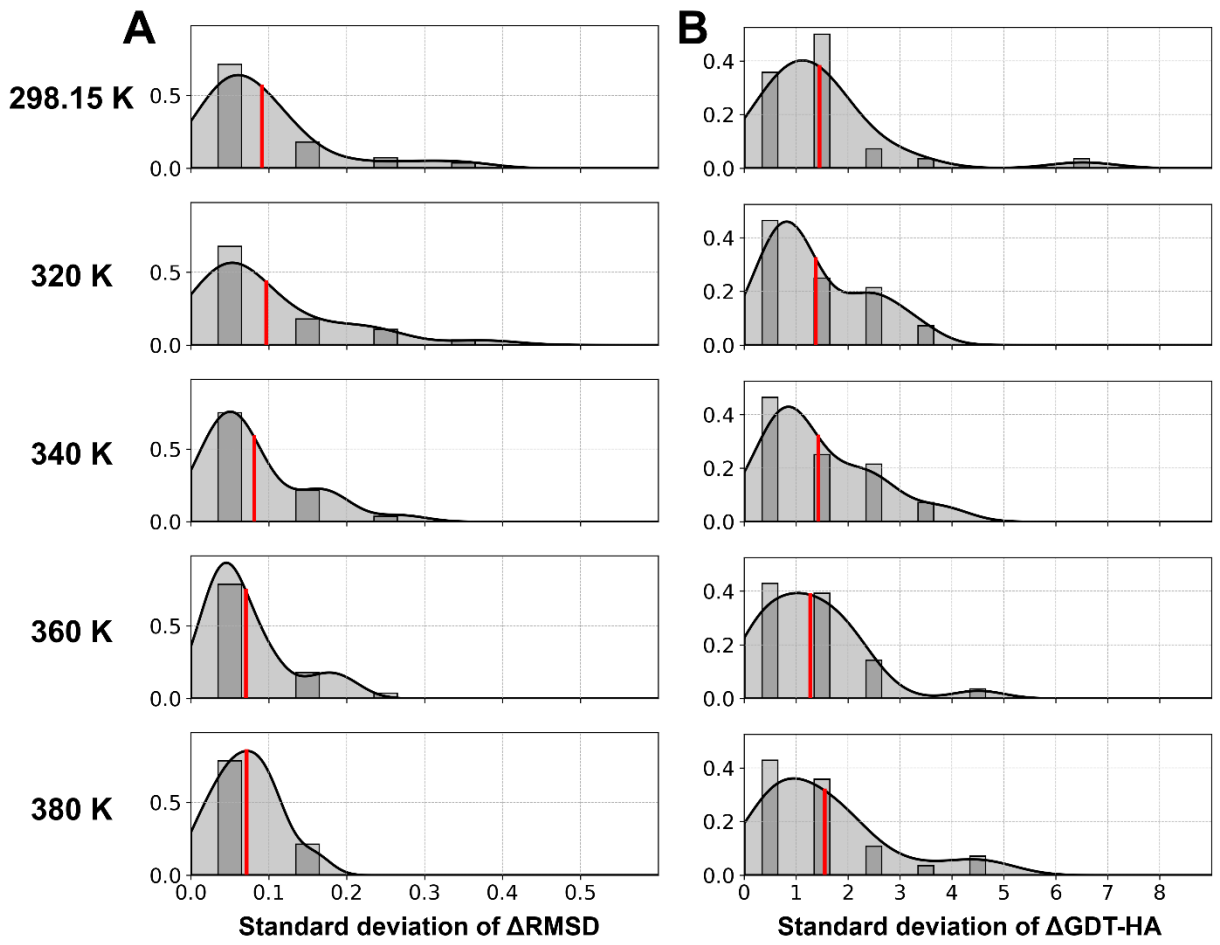

Figure S4. Refinement performance consistency at different temperatures. Standard deviation of refined model qualities from three runs of refinement are shown as histograms and density distributions with Gaussian kernels. Averages of the standard deviations are shown as red vertical lines.

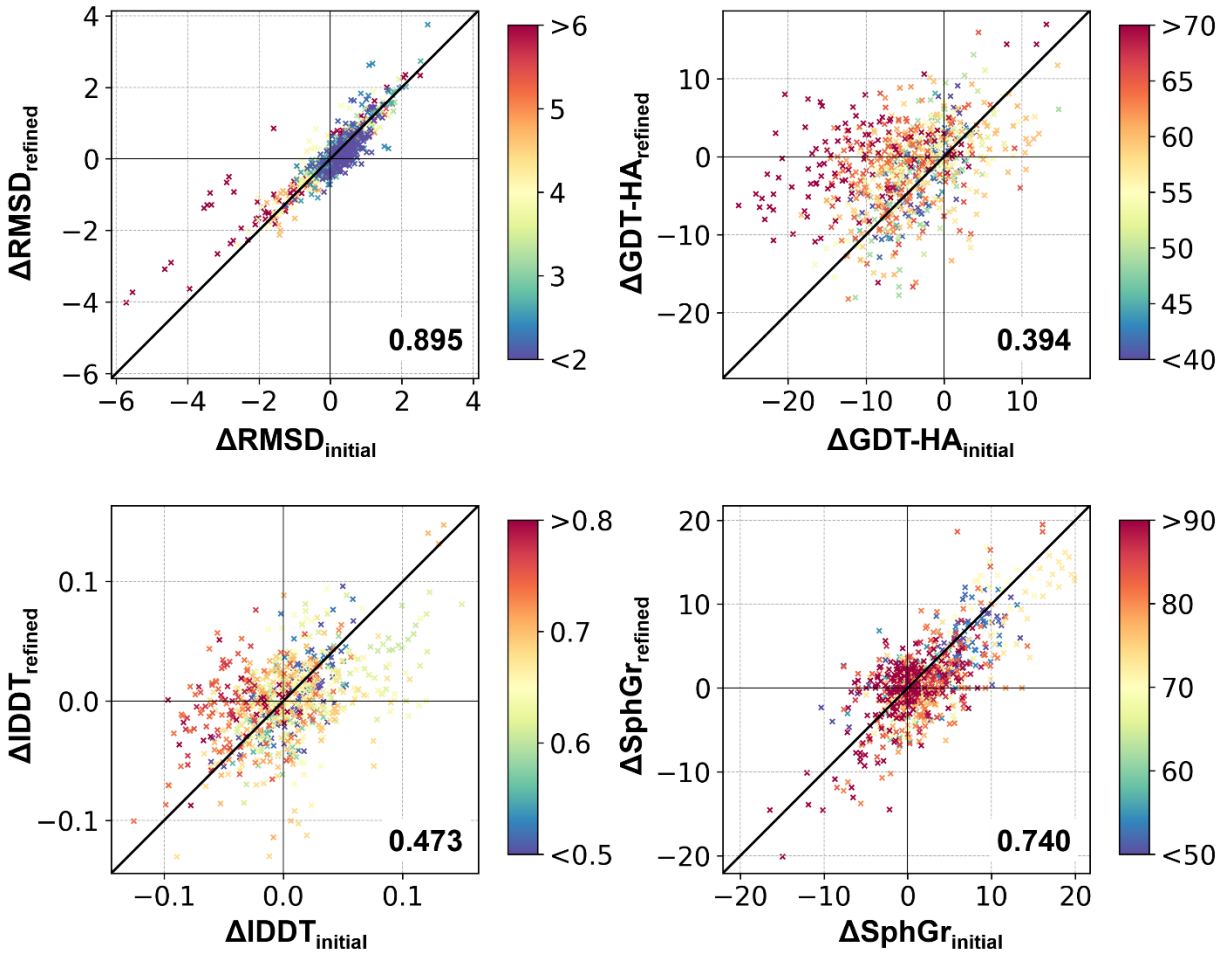

Figure S5. Refinement results dependence on alternative initial model qualities. In each panel, each model quality difference between the original initial model and alternative models, $\Delta Quality_{initial}$ (alternative-original) are on the X-axes, and difference between their refined models, $\Delta Quality_{refined}$ (alternative-original) are on the Y-axes. Points are colored based on the original initial model qualities. Pearson's correlation coefficients are shown in each panel.

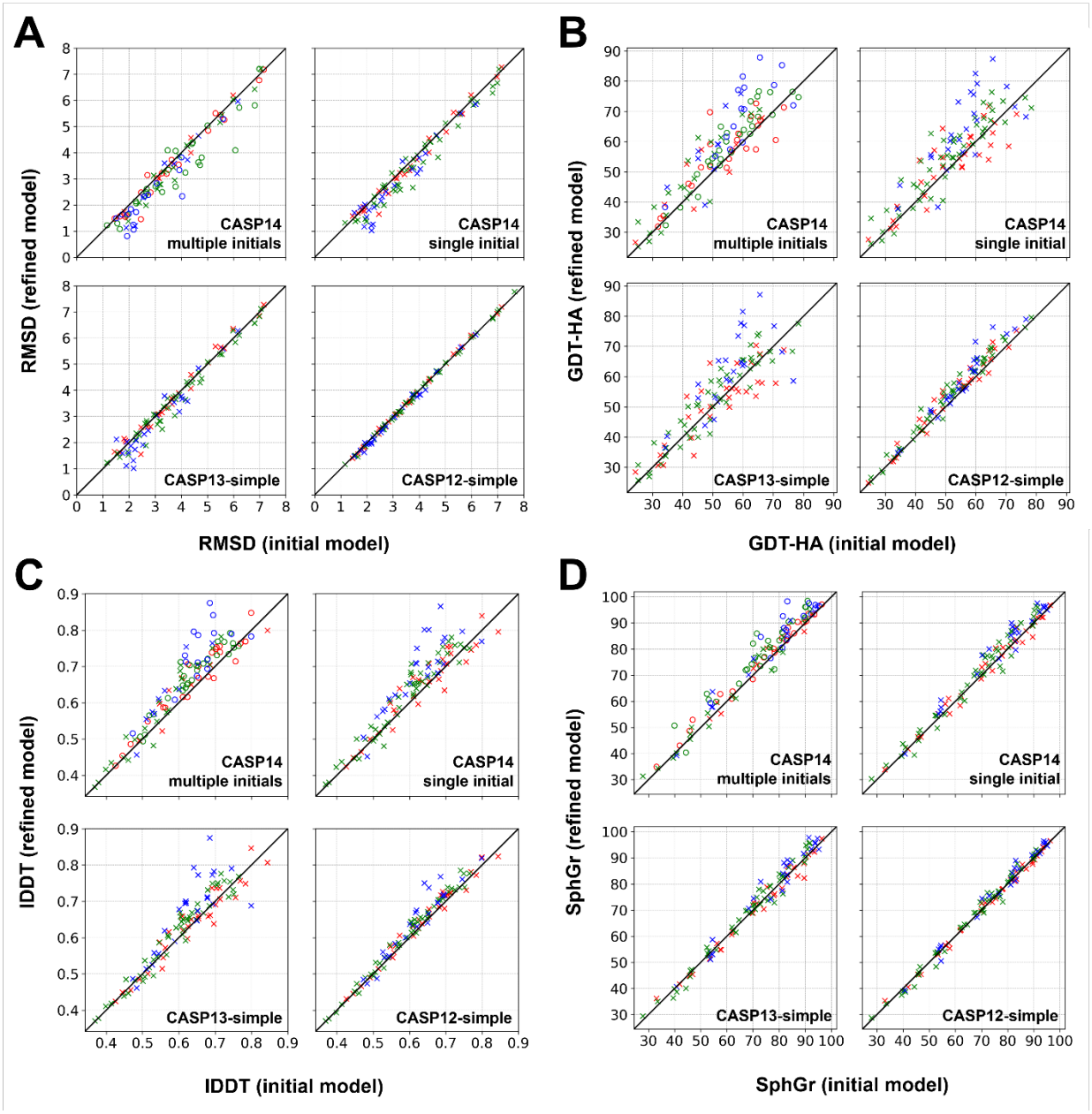

Figure S6. Refinement results for the benchmarked protocols, model 1. Model qualities for the initial model and refined models (the averaged values of three independent tests) with each protocol are compared as scatter plots: Cα-RMSD (A), GDT-HA (B), lDDT (C), and Sphere Grinder (D). For the CASP14 protocol with multiple initial models, targets refined with multiple initial models are presented as circles, while targets which were not able to generate multiple initial models were shown as Xs, and the corresponding data is the same as the CASP14 protocol with single initial model. Colors represents the target protein sizes; blue, green, and reds points correspond to small (< 100 residues), medium (≥ 100 and < 200 residues), and large (≥ 200 residues) proteins.

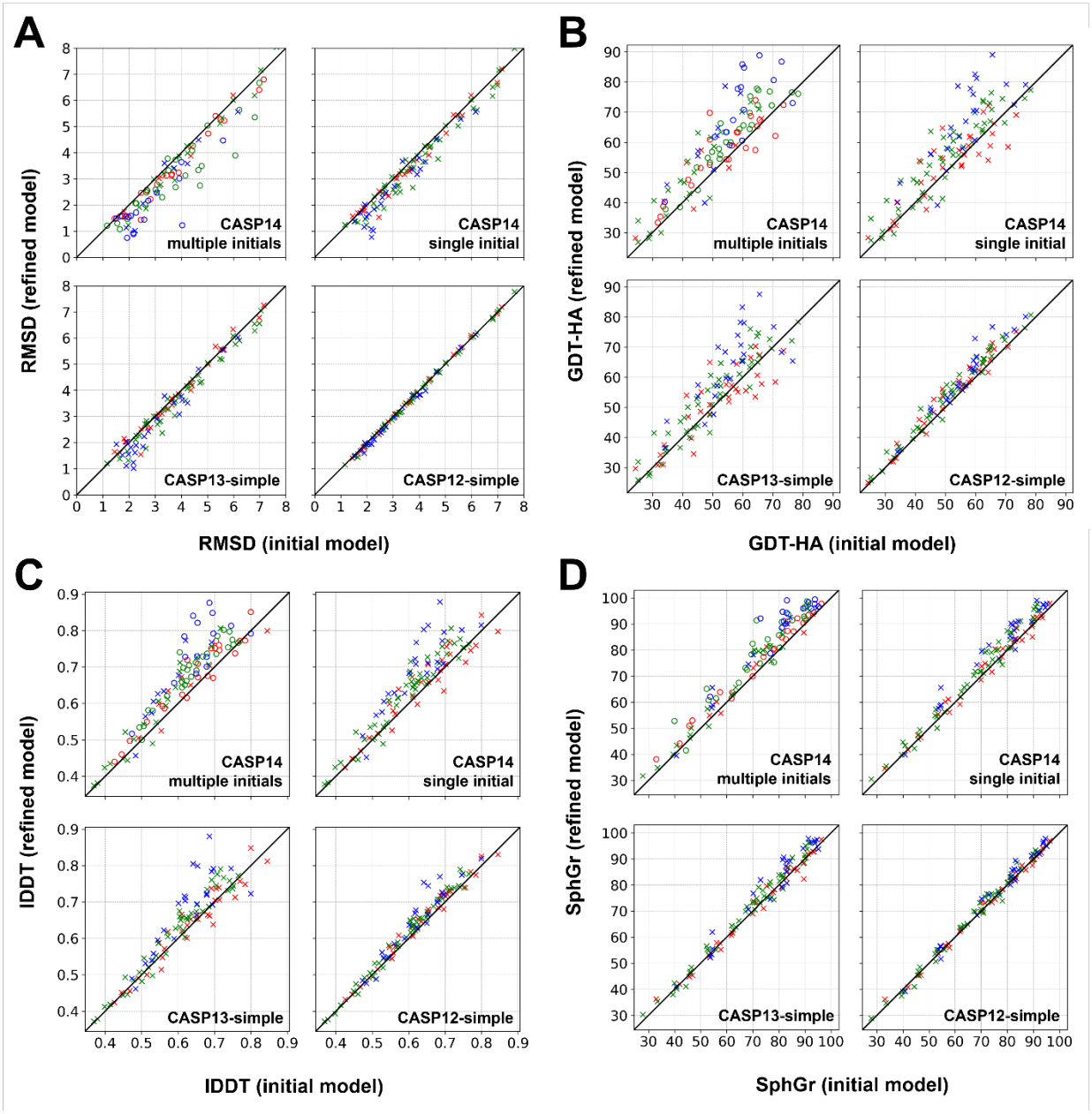

Figure S7. Refinement results for the benchmarked protocols, best out of 5 models. See **Figure S6** for the figure descriptions.

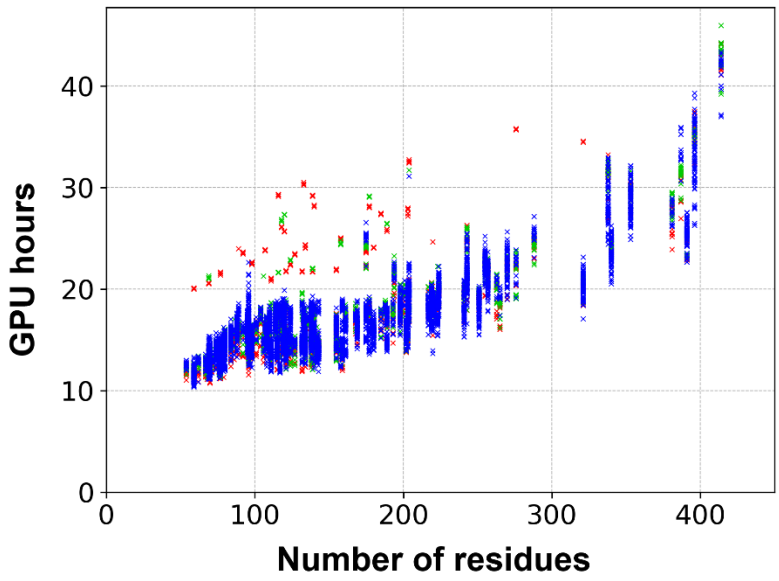

Figure S8. Computational cost for structure sampling via molecular dynamics simulations. GPU hours on an RTX-2080Ti for MD simulations from an initial model is shown as a function of protein size. Sampling protocols named CASP12, CASP13, and CASP14 are presented in red, green, and blue, respectively.

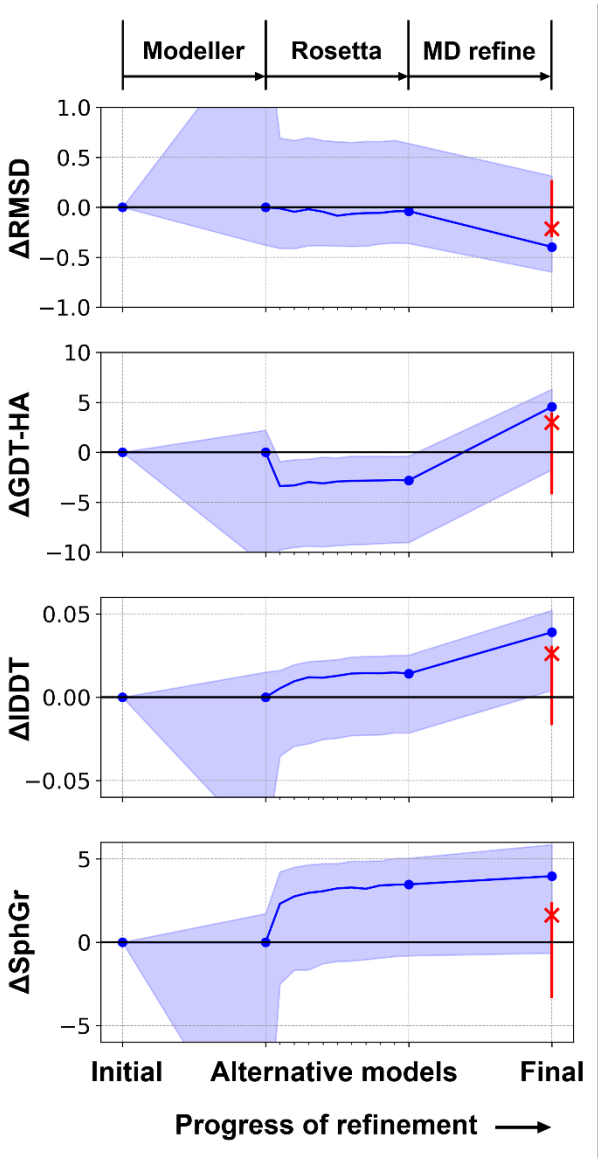

Figure S9. Progress in model qualities for refinement with additional alternative initial models. Average model 1 qualities for refinement with additional alternative initial models are shown in blue lines. Methods used for each step are shown at the top: "Modeller" for single template-based modeling by Modeller, "Rosetta" for "Iterative Hybridize" protocol of Rosetta, and "MD refine" for MD sampling and followed ensemble-averaging. For the alternative model generation step, the lines track the lowest Rosetta energy conformations. Quality ranges between the best and worst among the top 5 models are presented as blue shaded regions. Average model 1 qualities for refinement without alternative initial models are shown in red points, and quality ranges for the method are presented as error bars.

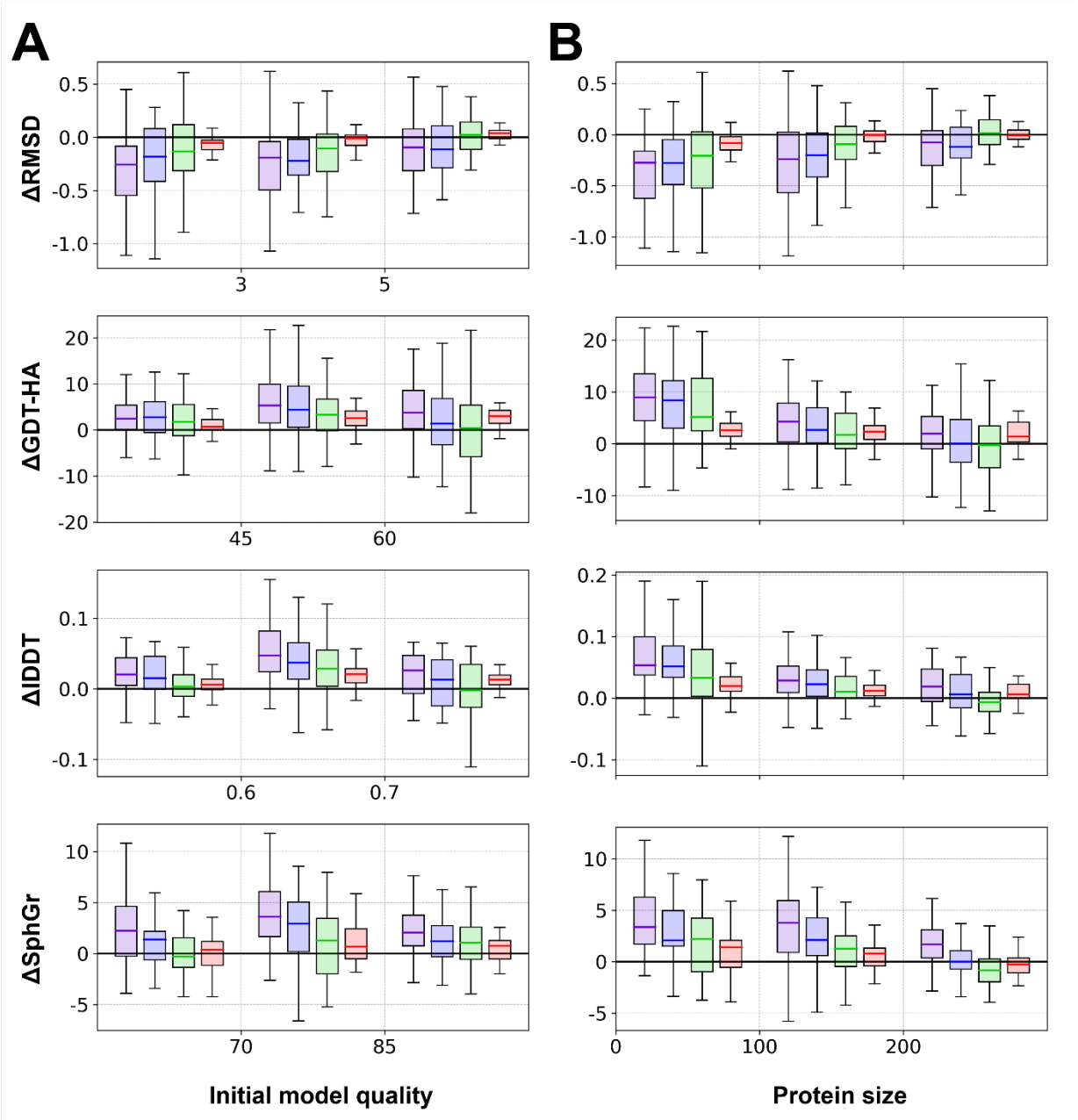

Figure S10. Model 1 refinement performance dependence on initial model quality (A) and protein size (B). Statistical analysis for each bin is shown as a boxplot for "CASP14" with multiple initial models (violet), "CASP14" with single initial model (blue), "CASP13-simple" (green), and "CASP12-simple" (red).

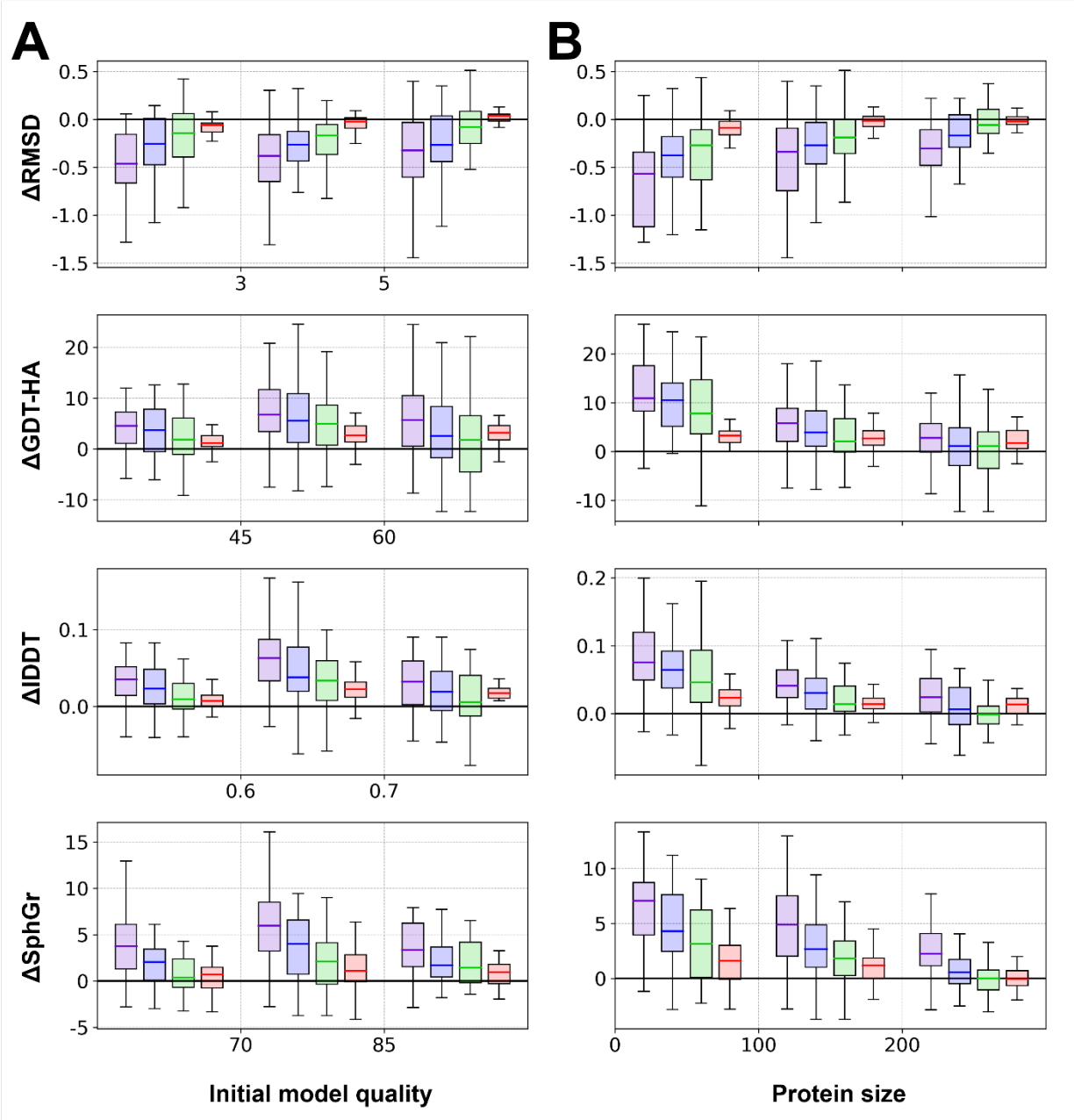

Figure S11. Refinement performance of the best out of five models dependence on initial model quality (A) and protein size (B). See **Figure S10** for the figure descriptions.

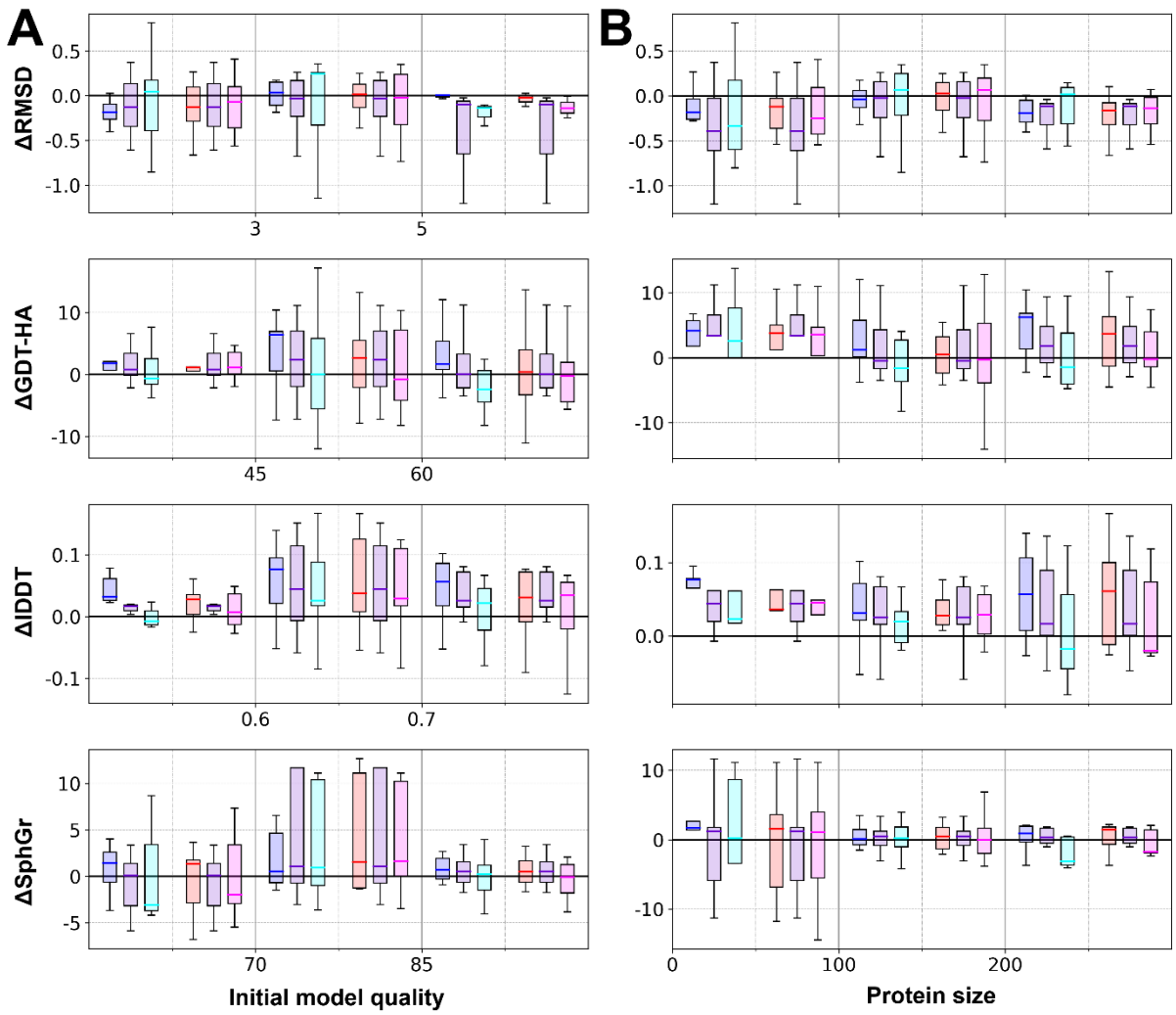

Figure S12. Refinement performance dependence on initial model quality (A) and protein size (B) for various selections of the Cartesian restraint parameters, k_0 and b_{flat} of Eq. 1. Statistical analysis for each bin is shown as a boxplot. On the left side of each bin, results with k_0=0.025 kcal/mol/Å^2 and b_{flat}=2, 4, and 6 Å are shown in blue, purple, and cyan, respectively. On the right side, results with b_{flat}=4 Å and k_0=0.05, 0.025, and 0.01 kcal/mol/Å^2 are shown in red, purple, and magenta, respectively.

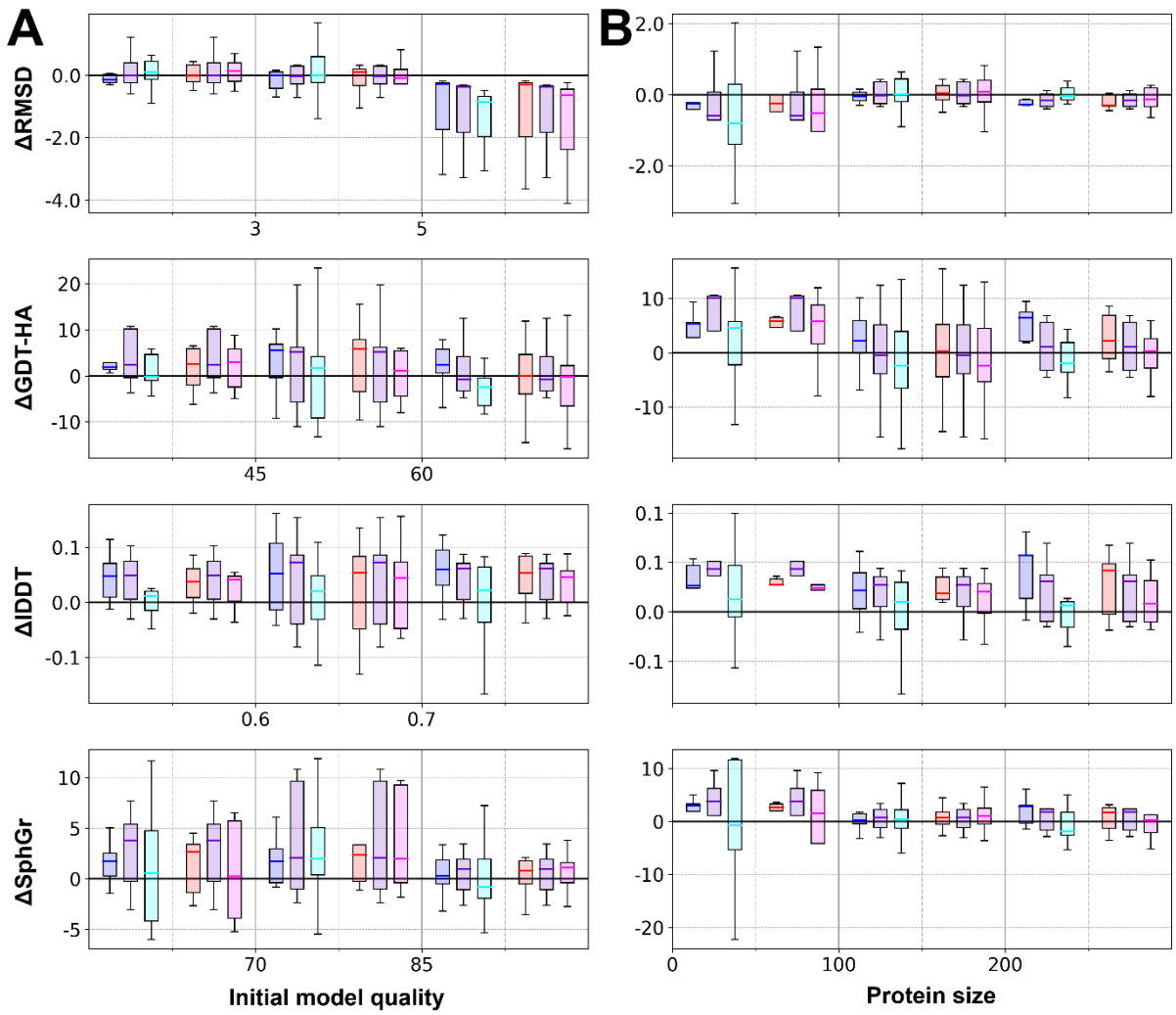

Figure S13. Refinement performance dependence on initial model quality (A) and protein size (B) for various selections of the distance restraint parameters, k_0 and b_{flat} of Eq. 2. Statistical analysis for each bin is shown as a boxplot. On the left side of each bin, results with k_0=0.05 kcal/mol/Å^2 and b_{flat}=0, 2, and 4 Å are shown in blue, purple, and cyan, respectively. On the right side, results with b_{flat}=2 Å and k_0=0.1, 0.05, and 0.025 kcal/mol/Å^2 are shown in red, purple, and magenta, respectively.

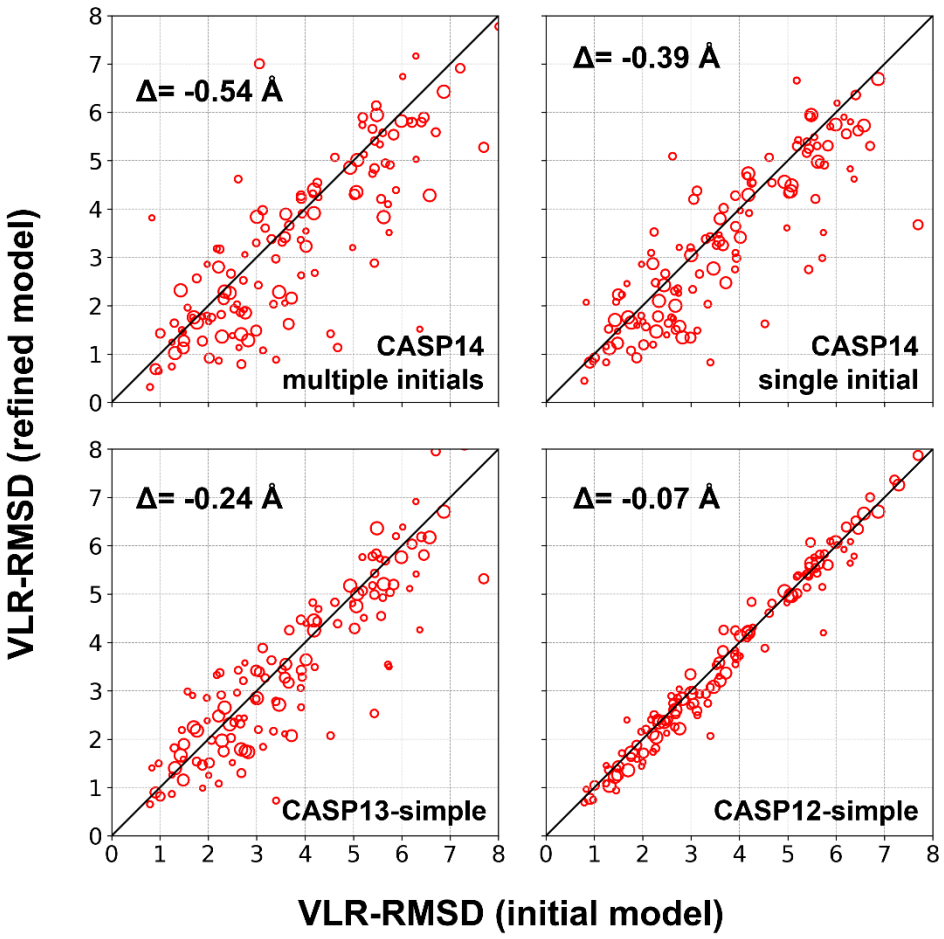

Figure S14. Refinement results on variable local regions (VLRs) for the benchmarked protocols. Cα-RMSDs for the initial and refined models are compared as scatter plots; the circle size is proportional to the VLR length. Per-VLR average changes are shown at the top left of each panel.

SUPPLEMENTARY TABLES

Table S1. Ligand binding site accuracy in backbone RMSD

Target	Ligand[1]	Initial[2]	CASP14[3] (multiple initial models[4])	CASP14[3] (single initial models)	CASP13-simple[3]	CASP12-simple[3]
R0999-D3	SKM	0.68	1.36 (+0.68)	1.36 (+0.68)	1.27 (+0.59)	0.86 (+0.18)
TR520	COI	1.42	1.44 (+0.02)	1.57 (+0.15)	1.72 (+0.30)	1.50 (+0.08)
	Ca^{2+}	1.17	0.54 (-0.63)	0.43 (-0.74)	0.97 (-0.20)	1.18 (+0.01)
TR786	Zn^{2+}	1.21	0.56 (-0.66)	1.01 (-0.21)	1.39 (+0.18)	1.40 (+0.18)
TR810	Zn^{2+}	2.46	2.14 (-0.32)	2.16 (-0.31)	2.44 (-0.02)	2.51 (+0.05)
TR823	G4P	2.83	2.40 (-0.43)	2.40 (-0.43)	2.29 (-0.54)	2.80 (-0.03)
TR879	Zn^{2+}	0.83	1.48 (+0.65)	1.65 (+0.82)	1.51 (+0.69)	1.00 (+0.17)
TR891	Heme	1.23	1.54 (+0.30)	1.54 (+0.31)	1.53 (+0.29)	1.24 (+0.01)
TR893	ADP, Mg^{2+}	0.60	1.71 (+1.11)	1.60 (+0.99)	1.39 (+0.78)	0.77 (+0.16)
TR920	CYT	1.32	1.75 (+0.43)	1.71 (+0.39)	1.92 (+0.60)	1.42 (+0.10)
TR922	Ca^{2+} (201)[5]	0.47	0.38 (-0.09)	0.40 (-0.06)	0.94 (+0.48)	0.57 (+0.10)
	Ca^{2+} (202) [5]	1.40	1.28 (-0.12)	1.47 (+0.08)	1.49 (+0.09)	1.48 (+0.08)

[1] Biologically relevant ligands were only considered for the evaluation. We defined ligand binding site residues as protein residues that are within 5 Å from the ligand molecule.

[2] Binding sites that already have a high quality in the initial model (< 1 Å) were shaded.

[3] Backbone RMSD changes are shown in the parentheses. Improvements and deteriorations more than 0.1 Å are highlighted in blue and red, respectively.

[4] Single initial model-based procedure was used for targets that were not able to generate alternative initial models.

[5] There were two Ca^{2+} binding sites. Residue numbers for each Ca^{2+} ions are shown in the parentheses.