Received: 21 June 2017 | Revised: 1 August 2017 | Accepted: 18 August 2017

DOI: 10.1002/prot.25373

RESEARCH ARTICLE

WILEY PROTEINS

Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent

Genki Terashi[1] | Daisuke Kihara[1,2]

[1]Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907

[2]Department of Computer Science, Purdue University, West Lafayette, Indiana 47907

Correspondence
Daisuke Kihara, Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907,
Email: dkihara@purdue.edu

Funding information
Division of Biological Infrastructure, Grant number: DBI1262189; National Institute of General Medical Sciences, Grant number: R01GM123055; Division of Information and Intelligent Systems, Grant number: IIS1319551; Division of Integrative Organismal Systems, Grant number: IOS1127027; Division of Mathematical Sciences, Grant number: DMS1614777

Abstract

Protein structure prediction has matured over years, particularly those which use structure templates for building a model. It can build a model with correct overall conformation in cases where appropriate templates are available. Models with the correct topology can be practically useful for limited purposes that need residue-level accuracy, but further improvement of the models can allow the models to be used in tasks that need detailed structures, such as molecular replacement in X-ray crystallography or structure-based drug screening. Thus, model refinement is an important final step in protein structure prediction to bridge predictions to real-life applications. Model refinement is one of the categories in recent rounds of critical assessment of techniques in protein structure prediction (CASP) and has recently been drawing more attention due to its realized importance. Here we report our group's performance in the refinement category in CASP12. Our method is based on inexpensive short molecular dynamics (MD) simulations in implicit solvent. Our performance in CASP12 was among the top, which was consistent with the previous round, CASP11. Our method with short MD runs achieved comparable performance with other methods that used longer simulations. Detailed analyses found that improvements typically occurred in entire regions of a structure rather than only in flexible loop regions. The remaining challenge in the structure refinement includes large conformational refinement which involves substantial motions of secondary structure elements or domains.

KEYWORDS
CASP, computational method, molecular dynamics, protein structure prediction, protein structure refinement, structure modeling, template-based modeling

1 | INTRODUCTION

Methodology of protein structure prediction has been intensively studied over decades from various angles, such as bioinformatics, physics, chemistry, statistics, and robotics. Although there are still some areas that need further development, for example, template-free (often also called ab initio or de novo) modeling,[1,2] structure prediction has become practical in several situations, which include cases where structures can be modeled using templates (template-based modeling).[3–5] The progress of the protein structure modeling field has been objectively monitored in the critical assessment of techniques for protein structure prediction (CASP) from 1994,[6] a community-wide assessment of prediction methods that is held every 2 years. In CASP, participating prediction groups/methods are evaluated based on structure models they build for protein target sequences, which are presented by the organizers before determination of their tertiary structures. From CASP7 in 2006, assessors' evaluation reports show that performance of template-based modeling has not changed much partly reflecting maturity of the methods.[4,7–10] Now, biologists routinely use structure prediction to interpret or design experiments.[11,12]

In the case of template-based modeling, obtaining a structure model of the correct fold, that is, a structure that has a root-mean square deviation (RMSD) of 3–6 Å, can be expected if the template structures found in a structure library has a reasonably high confidence score.[5,13] Improving the structure model further through model refinement, an additional procedure to gain a couple of more angstroms in RMSD to the native structure, is critical for bringing a computational model with the correct fold to a level that it is practically useful in various applications. For example, a model at a 5 Å RMSD would be only used for indicating residue positions in the protein such as interpreting

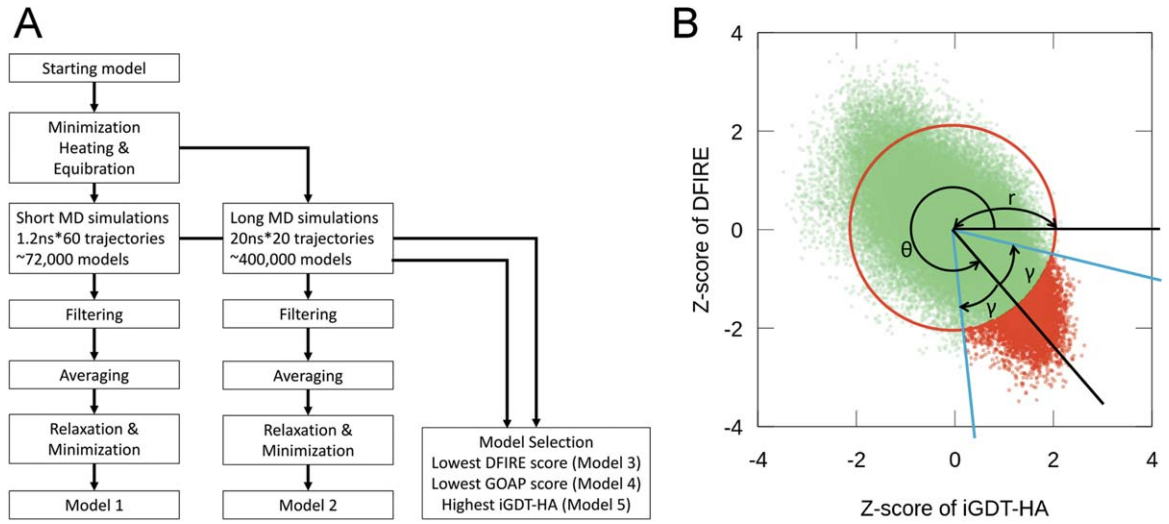

FIGURE 1 The model refinement protocol of the Kiharalab group in CASP12. **A**, the flow chart of our refinement protocol. Before performing MD simulations, the energy of the staring model was minimized, and then subjected to the next heating and equilibration step. The equilibrated structure was used as the starting model for the two types of production MD runs (short MD and long MD simulations). In the filtering step, subset of models extracted from the MD trajectories were selected by the filtering criteria. The coordinates of selected models were averaged and then relaxed. **B**, Definition of the parameters used in the filtering step. The *x* axis is the *Z* score of GDT-HA of a model from the initial structure and the *y* axis is the *Z* score of the DFIRE scoring function. The distribution shown is from extracted models from short MD trajectories of TR520. Red points represent selected models by the filtering with $r = 2.0$, $\theta = 310$, and $\gamma = 35$

and designing residue mutagenesis experiments. However, if the model was refined to within 1.5 Å RMSD, it can be used for molecular replacement in experimental structure determination, virtual drug screening, and investigating enzymatic reactions.[14] Thus, model refinement has been one of the foci of recent method developments in the CASP community.[4,15]

However, model refinement is still not easy. It is well known that running naïve molecular dynamics (MD) simulations on a structure model do not improve model consistently. Rather, it deteriorates the model for almost half of the cases.[16] In CASP, until CASP9 held in 2010, there were no methods that could refine models consistently with statistical significance.[17] This situation changed in CASP10, when the FEIG and Seok groups showed improvements on the starting models in majority of their cases.[18] Particularly, the FEIG group significantly outperformed all the other groups in their Global Distance Test-High Accuracy (GDT-HA) score[19] improvement. The FEIG's approach was based on MD: from the MD trajectories starting of a model to be refined, structures were chosen that did not deviate much from the initial structure to avoid degraded structures and further filtered by using additional scoring function.[20] In CASP11, many top performing groups, including FEIG,[15,21] used MD-based approaches with some variations, for example, using support vector machine (a machine learning method) to select models after MD,[22] remodeling using multiple templates before MD-based refinement,[23] taking consensus with homologous structures,[24] or using multiple rounds of relaxation.[25]

Here, we report our group's method and performance in the model refinement category in CASP12. Our approach is based on MD trajectories, following the FEIG group's success, with several critical differences: We used an implicit solvent model rather than explicit water molecules in simulation, and moreover, the length of the trajectories is significantly shorter than those used in the FEIG approach. In spite of the computationally inexpensive strategy, our approach was ranked high among participants in CASP11[15] as well as in CASP12. Differences of our approach in CASP11 and CAS12 are that we optimized parameters of the method carefully for CASP12 and also changed the implicit solvent model used in the MD simulations. In CASP12, our approach refined 29 out of 42 targets successfully in terms of the GDT-HA score and 24 in terms of the CASP12 assessors' score that is a linear combination of five scores (http://predictioncenter.org/casp12/zscores_final_refine.cgi?formula = assessors) . We showed that the refinement occurred not only at loop regions but also overall in structures cores, and the approach mainly expanded structures (showing that the improvement of structure evaluation scores are not merely due to compression of structures). Drawbacks of the approach are also discussed.

2 | MATERIALS AND METHODS

2.1 | Overview of the refinement protocol

Our refinement protocol used in CASP12 is shown in Figure 1. Figure 1A shows the overall flowchart. We performed short and long MD simulations after the energy minimization was applied to a structure model to refine. The short MD consists of sixty 1.2 nanosecond (ns) MD simulations with restraints of increasing strength (0.1, 0.2, 0.4 kcal mol^{-1} $Å^{-2}$) applied to Cα atom positions, which was increased every 400 pico seconds (ps). The long MD consists of twenty 20 ns MD simulations with weak restraints (0.05 kcal mol^{-1} $Å^{-2}$)[21] on Cα atoms positions. After the MD simulations, a subset of structures in

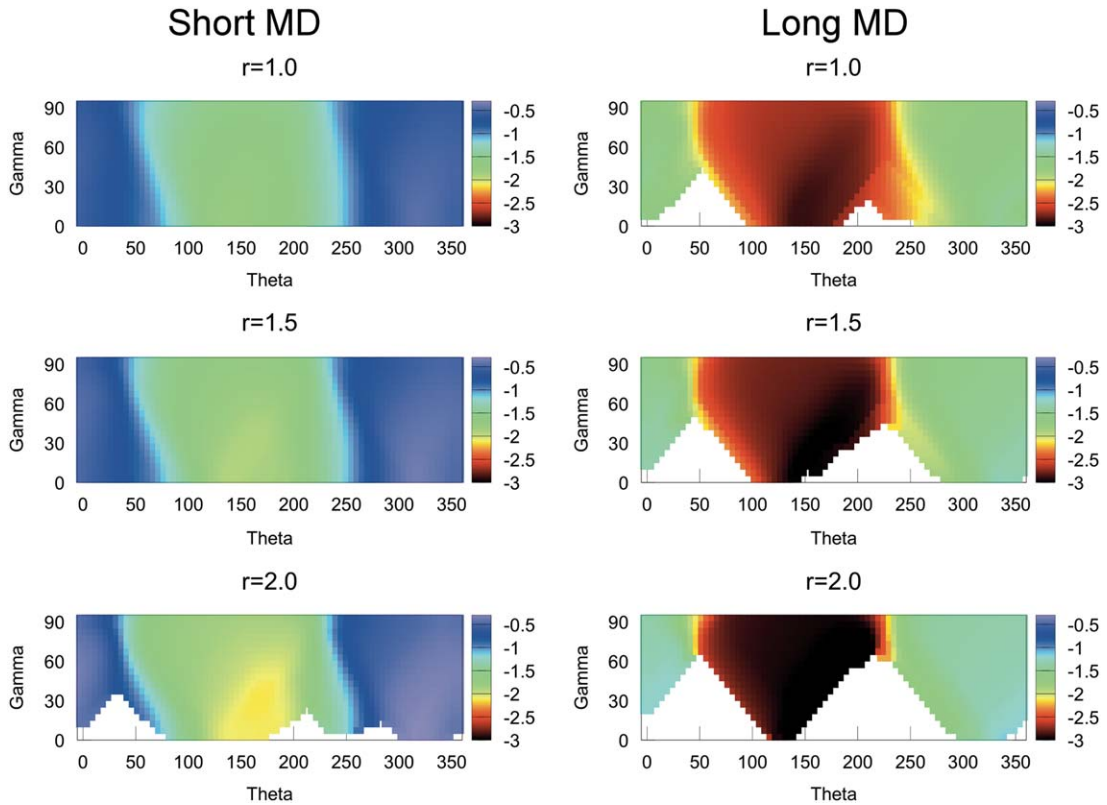

FIGURE 2 Results of model selection from MD trajectories with different parameter settings, *r*, θ, and γ. For the explanation of the parameters, see Figure 1B. The distribution of the average dGDT-HA of selected models from the short MD (left) and the long MD (right) trajectories are shown in a color scale from dark red to purple for negatively large dGDT-HA, −3 to over −0.5. dGDT-HA is the difference of the GDT-HA to the native structure of the initial model to that of the average of the selected models using a corresponding parameter set. A white region means that no models exist for the corresponding parameter combinations

the trajectories were selected by considering the deviation from the starting model and the statistical potential score (Figure 1B). The selected models were averaged on the Cartesian coordinates of atoms, which were then minimized and relaxed with a10 ps MD simulation.

2.2 | Setting of MD simulation

MD runs were performed by CHARMM MD program version 38b2 with CHARM22/CMAP force field. We used a 2 femto seconds (fs) time step for all of simulations. The non-bond interactions were listed using a 14 Å cutoff. Electrostatic interactions were calculated with a shifting function applied to the potential energy at 12 Å. van der Waals interactions were calculated with a switching function applied to the potential energy between 10 and 12 Å. The solvent effect was computed by the FACTS implicit solvent model[26] with its default parameters. In CASP11, we used SCPISM[27] for implicit solvent but changed it to FACTS following a comparison study of implicit solvents by Hua et al.[28] Before running MD simulations, hydrogen atoms were added to the starting model by the CHARMM HBUIld command. To fix the length of bonds involving hydrogen atoms, the CHARMM SHAKE command was used. Energy minimization was performed in total of 12,100 steps of the Adopted Basis Newton-Raphson (ABNR) method. In the first 100 steps of the minimization, the position of all non-hydrogen atoms was fixed, and then we applied harmonic constants that were subsequently decreased from 20.0, 10.0, 5.0, 2.0, 1.0, to 0.5 kcal mol^{-1} $Å^{-2}$ for every 2 000 steps (4 ps). The minimized protein model was subjected to the next heating and equilibration step. The temperature of the system was gradually increased from 50 K to 298 K in 200 000 steps (400 ps) with harmonic restraints of 0.05 kcal mol^{-1} $Å^{-2}$ on Cα atom positions. Then, the equilibrated structure was used as the starting model for production MD runs. All trajectories were calculated with the leapfrog verlet algorithm at 298 K. Model structures are saved every 500 steps (1ps) in each trajectory. In the short and the long MDs, a total of 72 000 (1200*60) and 400 000 (20 000*20) models were extracted from trajectories, respectively.

2.3 | Model filtering

Following the FEIG method,[20,21] extracted models from MD trajectories were cross-evaluated by a statistical knowledge-based potential (we used DFIRE[29]). We have also calculated the GDT-HA between the starting model and the extracted models after MD (denoted as *iGDT_HA*). These two parameters for the filtering are illustrated in Figure 1B. The DFIRE score and *iGDT_HA* of each selected model were normalized by computing *Z* score, using a distribution of structures from each MD trajectory as follows:

$$Z\_iGDT\_HA_m = \frac{iGDT\_HA_m - \overline{iGDT\_HA}}{\sigma_{iGDT\_HA}} \quad (1)$$

$$Z\_DFIRE_m = \frac{DFIRE_m - \overline{DFIRE}}{\sigma_{DFIRE}} \quad (2)$$

where $\overline{iGDT\_HA}$ and $\overline{DFIRE}$ are average values, and σ_{iGDT_HA} and σ_{DFIRE} are standard deviations of *iGDT_HA* and DFIRE. For DFIRE, structures with negative *Z* scores are those which are more geometrically favorable than the average (*Z* score = 0). For *iGDT-HA*, structures with a positive value are those which are more similar to the initial structure than the average.

To select models, we applied a filter (Figure 1B) that extracts a pie-shaped area between angle $\theta \pm \gamma$ degree and the radial distance *r* from the center of distribution of $Z\_iGDT_HA_m$ and Z_DFIRE_m. The criteria for selecting model *m* were

$$Z\_iGDT\_HA_m{}^2 + Z\_DFIRE_m{}^2 > r \quad (3)$$

and

$$arccos\left(\frac{Z\_iGDT\_HA_m cos\theta + Z\_DFIRE_m cos\theta}{Z\_iGDT\_HA_m{}^2 + Z\_DFIRE_m{}^2}\right) < \gamma \quad (4)$$

which corresponds to the lower-right region in red in Figure 1B. Thus, the intention is that, among models from MD trajectories, we would like to select those which have relatively low (better) statistical potential and less deviation from the initial model.

For the extracted models from the Short MD trajectories, we used $r = 2.0$, $\theta = 310$ and $\gamma = 35$ in the filtering step. For the Long MD, we used $r = 1.5$, $\theta = 325$ and $\gamma = 30$. These parameters were chosen based on a benchmark study performed on the CASP11 dataset (discussed in Results). After the filtering, 1 000∼4 500 and 1 000∼25 000 models were selected from the short MD and the long MD trajectories. Selected models were averaged, which were subject to the structure relaxation and energy minimization (Model 1 and Model 2 in Figure 1A).

This protocol is different from the FEIG method[21] used in CASP11 in the following ways: First, the MD runs in our protocol were much shorter. We used 1.2 ns * 60 trajectories (the short MD runs) and 20 ns * 20 trajectories (the long MD runs), thus in total of 472 ns MD runs, while the FEIG method used 1200 ns long runs (40 ns * 30). Second, we used an implicit solvent model while the FEIG method used explicit solvent. These two differences make our protocol more computationally inexpensive and affordable. There are also other technical differences, described in the following. Third, as mentioned above, we applied Cα restraints in the MD runs that increased over time from 0.1 to 0.4 kcal mol^{-1} $Å^{-2}$ where FEIG used a constant value of 0.05 kcal mol^{-1} $Å^{-2}$. Fourth, the interval of structure snap shots taken was 1 ps in our protocol, while FEIG used 40 ps. Fifth, for the filtering (Figure 1B), we used iGDT-HA and DFIRE, while the FEIG method used iRMSD and RW+.[30]

2.4 | Structure relaxation and energy minimization

Averaged models underwent energy minimization with 500 steps of the steepest descent algorithm and 4 500 steps of ABNR. The minimized models were relaxed with a 40 ps MD simulation. In the minimization and relaxation, we used strong harmonic restraints of 100 kcal mol^{-1} $Å^{-2}$ on C-α atom positions. The model 3–5 were selected from all extracted models with the lowest DFIRE, the lowest GOAP[31] score and the highest iGDT-HA (Figure 1A).

TABLE 1 Average performance of our protocol in comparison with top10 groups in CASP11

Group[a]	Group ID	GDT-TS	GDT-HA	RMS_CA
FEIG	288	74.63	56.45	3.58
Seok	296	72.45	53.71	3.62
Seok-refine	423	72.61	53.56	3.59
Schroderlab	396	73.27	55.07	3.67
PRINCETON_TIGRESS	106	72.75	53.98	3.65
Kiharalab	333	72.88	54.22	3.57
LEE	169	72.14	53.31	3.65
BAKER	64	71.20	52.25	3.86
nns	038	71.57	52.51	3.66
Seok-server	011	72.66	53.93	3.64
Average of all 53 Groups	n/a	69.13	50.08	3.99
Starting model	n/a	71.96	53.03	3.69
Short MD	*n/a*	*73.26*	*54.27*	*3.62*
Long MD[a]	*n/a*	*70.48*	*50.82*	*3.84*

The names of top 10 groups were obtained ranked according to the CASP11 web site, assessors' formula.
(http://predictioncenter.org/casp11/zscores_final_refine.cgi?formula = assessors).
[a]The results are for 31 out of the 36 targets for which the long MD runs finished before the CASP12 has started to release refinement targets.

2.5 | Computational costs

For a 200-residues target protein, the single 0.4 ns equilibration and 1.2 ns MD simulation took about 13.2 and 40 CPU hours, respectively. The total computational cost of the short MD simulations (60 trajectories) was about 3,200 CPU hours per target (2.7 hours on 1 200 cores of Intel Xeon-E5 CPU). A single 20 ns MD simulation took about 640 CPU hours. Consequently, the total computational costs of the long MD (20*20 ns MDs) was about 12,800 CPU hours for a refinement target (64 hours on 200 cores of Intel Xeon-E5 CPU).

3 | RESULTS AND DISCUSSIONS

First we describe benchmark results on the CASP11 refinement dataset, then show the results of our group in CASP12.

3.1 | Parameter optimization for model selection using the CASP11 dataset

As a preparation for participating in CASP12, we optimized parameters, *r*, θ, and γ that were used in the filtering protocol (Figure 1B) on the 36

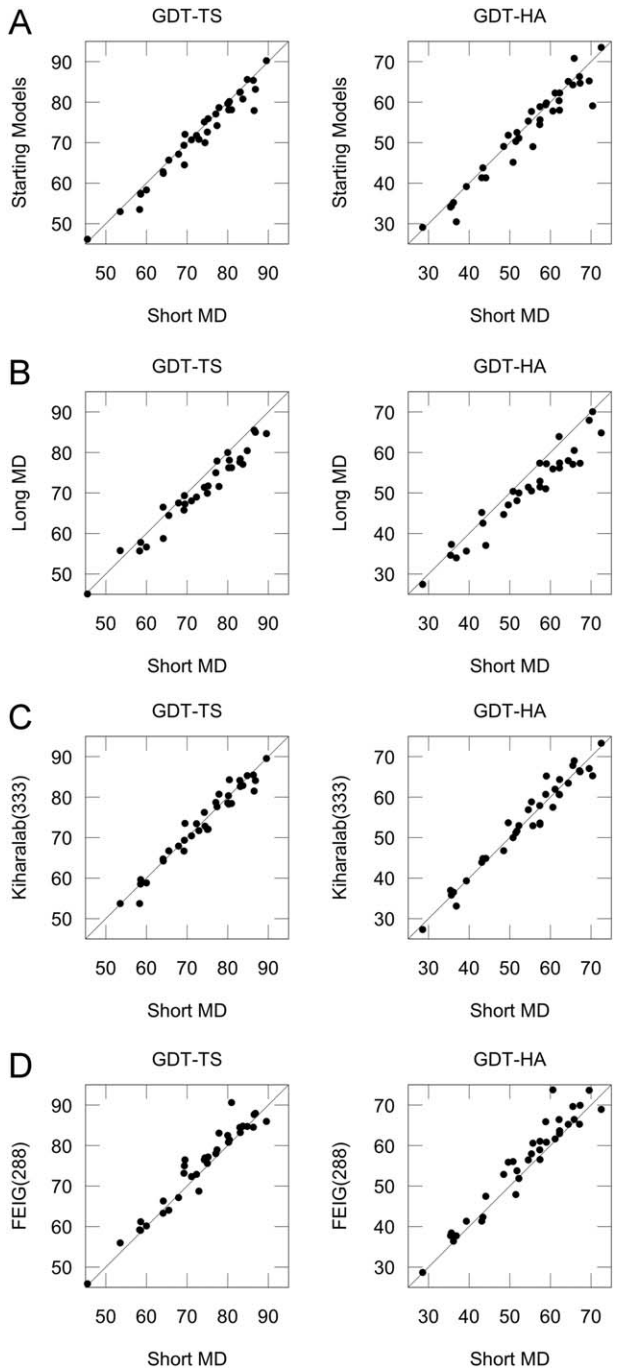

FIGURE 3 Comparison of the refined models by the short MD protocol on the CASP11 refinement targets. Models were evaluated by GDT-TS in the left column and by GDT-HA in the right column. **A**, refined models by the short MD in comparison with starting models. **B**, the short MD results compared with the long MD results. **C**, The refined models with the short MD runs compared with models submitted by Kiharalab (Group number 333) in CASP11. **D**, the short MD results compared with models by FEIG (Group number 288) in CASP11

refinement targets from CASP11. The dataset included the target TR217 to TR857 excluding TR795, whose native structure was not available at the time of the work. The starting models and the native structure of these targets were downloaded from the CASP11 website (http://predictioncenter.org/casp11/). The size of the target proteins ranged from 62 to 288 residues (average: 155). The Cα RMSD of the starting models to their native structures ranged from 1.45 to 12.45 Å (average: 3.69 Å), GDT-TS ranged from 46.17 to 90.24 (average: 71.96), and GDT-HA were from 29.10 to 73.51 (average: 53.03).

The parameter optimization was performed separately for the short and the long MD runs. For the short MD runs we generated forty 1.2 ns-long trajectories while for the long MD runs we computed up to three 20 ns-long trajectories for the starting models of the targets. Long MD runs were finished only for 31 out of 36 targets before the first refinement target was released in CASP12.

We first tested in total of 3888 (= 3*72*18) combinations of the three parameters (r, θ, γ) for selecting models (Figure 1B), which come from three values, 1.0, 1.5, and 2.0 for r, 72 values from 0 to 355 with an interval of 5 for θ, and 18 values from 5 to 90 with an interval of 5 for γ. We examined the average change of GDT-HA of the selected models with each parameter set from that of the initial model (dGDT-HA). The results are visualized with a color scale in Figure 2. Higher in the color scale, purple to blue, are better than dark red that is at the bottom of the scale. Several important trends were observed: (1) Interestingly, as shown in Figure 2, none of the parameter combination gave on average better GDT-HA (i.e., positive dGDT-HA) than the starting model. However, as we show later, averaging structures selected by a good parameter set improved GDT-HA from the starting model. (2) Also, it was evident that the short MD runs (left panels) gave better results than the long MD runs (right panels). (3) The parameter space that gave better results locates at the right bottom corner of the panel, where θ is around 300 to 350 and γ is around 0 to 30 degrees. This region corresponds to the red regions in Figure 1B. (4) For this region, using a larger radius r gave better results. (5) The parameter space that did not perform well are at the middle bottom of the panel, where θ is around 150 and γ is around 0 to 50 degrees. This is the opposite region from the good performing region shown in red in Figure 1B.

We selected 76 and 40 parameter combinations with largest dGDT-HA for the short and long MD runs. Then, for each parameter combination, we generated an averaged structure model from the selected models by the combination. The generated models were evaluated by dGDT-HA. A combination of $r = 2.0$, $\theta = 310$, and $\gamma = 35$ was found to be optimal for the short MD trajectories and a $r = 1.5$, $\theta = 325$, and $\gamma = 30$ combination was best for the long MD runs.

3.2 | Refinement results on the CASP11 targets

Table 1 summarizes the performance of our protocol using the optimized refinement protocol with short and long MDs (Figure 1) on the CASP11 refinement targets. For comparison, results of the top 10 groups in CASP11 including our group (Kiharalab) are shown. Results are on the first model (Model 1) of the groups. In the table, the average value of three scores, GDT-TS, GDT-HA, and Cα-RMSD (RMS_CA) are shown, which were major evaluation scores used by the CASP assessors. GDT-TS (Global Distance Test-Total Score) and GDT-HA (Global Distance Test-High Accuracy) are the average of the percentage of Cα

TABLE 2 Average performances of the top10 groups in the CASP12 refinement targets

Group[a]	Group ID	Assessors' formula[b]	GDT-TS	GDT-HA	RMS_CA	MolProbity[c]	SphGr[d]	QCS[e]
GOAL	220	21.79	67.08	50.13	5.14	2.06	69.05	80.23
Seok	023	18.39	67.79	50.93	5.43	1.11	68.37	81.22
BAKER	247	17.73	67.11	50.48	5.36	0.87	69.28	81.02
Seok-server	250	16.91	67.40	50.34	5.42	1.11	68.05	78.57
SVMQA	208	15.01	67.72	51.01	5.48	1.71	67.30	80.70
FEIG	204	13.66	67.58	50.89	5.51	0.94	66.67	79.81
LEEab	450	13.58	64.94	47.42	5.37	1.83	68.13	78.53
Kiharalab	102	12.38	67.33	50.46	5.52	1.45	66.80	80.44
GOAL_COMPLEX	430	11.87	67.33	50.78	5.52	1.80	66.67	80.45
LEE	011	11.75	65.76	48.42	5.32	1.80	67.66	80.21
Average of All 39 Groups	n/a	−11.57	62.85	45.85	6.14	2.08	62.87	76.14
Starting model	*n/a*	*n/a*	*66.93*	*49.76*	*5.50*	*1.77*	*66.99*	*79.74*

[a]The names of Top 10 groups were obtained from CASP12 result page (http://predictioncenter.org/casp12/zscores_final_refine.cgi?formula = assessors) ranked by the assessors' formula.
[b]Combined *Z* score of GDT-HA, RMS_CA, SphGr, QCS and MolProbity score defined as;
$0.17Z_{GDT-HA} - 0.46Z_{RMS_CA} + 0.20Z_{SphGr} + 0.15Z_{QCS} - 0.02Z_{MolProbity}$.
[c]MolProbity considers the number of steric clashes and the percentages of outliers in rotamer and backbone conformations. A low MolProbity score indicates that a model is more physically favorable.
[d]SphereGrinder evaluates local structural similarity between a model and the native structure.
[e]Quality Control Score. QCS evaluates the correctness of secondary structure element predictions in a model.

atoms within four distance cutoffs from the corresponding Cα atoms in the native structure after structure superimposition. GDT-TS uses 1.0, 2.0, 4.0, and 8.0 Å while GDT-HA uses 0.5, 1.0, 2.0, and 4.0 Å for the distance cutoffs. Both scores ranges from 0 to 100 where 100 is the best score for a model. RMS_CA is RMSD of Cα atoms of a model to the native, which was calculated with the LGA program.[32]

Seven groups, FEIG, Seok, Seok-refine, Schroderlab, PRINCETON_TIGRESS, Kiharalab, and Lee showed improvement over the starting model in terms of all three scores, GDT-TS, GDT-HA, and RMS_CA (Table 1). In CASP11, our group (Kiharalab) used twenty10 ns MD runs with the Screened Coulomb Potential implicit solvent model (SCPISM). The current protocol using the short and long MD runs are shown at the bottom of the table. The Short MD runs performed very well, ranking the third in terms of GDT-HA (54.27) and GDT-TS (73.26) and the fourth for RMS_CA (3.62). The long MD runs performed worse than the short runs. On average, it failed to improve the starting models and ranked lower than the top 10 groups when GDT-TS and GDT-HA were considered, and was ranked ninth for RMS_CA.

In Figure 3, we examined refinement results with the short MD protocol in comparison with other refinement protocols. Figure 3A shows that short MD improved from the starting models in majority of the cases, 26 (72.2%) and 21 (58.3%) out of 36 targets in terms of GDT-TS and GDT-HA, respectively. It is also observed in Figure 3A that improvement by the short MD did not depend on the initial quality of the starting models. Next, we compared with the results with the long MD runs (Figure 3B). Consistent with the data shown in Table 1 and Figure 2, the short MD performed substantially better than the long MD. In terms of GDT-TS and GDT-HA, the models of the short MD were better for 26 and 27 targets (83.9% and 87.1%) than the long MD models among the 31 targets for which we had data with the long MD runs. In the next two panels, we compared the short MD refinement with the models submitted in CASP11 by Kiharalab (Figure 3C) and FEIG (Figure 3D). Compared with CASP11 Kiharalab models, the short MD's results were comparable. For about half of targets (17/36 targets), the short MD models had better GDT-TS and GDT-HA than the Kiharalab models. On the other hands, for only 7 (19.4%) targets the short MD models had better GDT-TS and GDT-HA than the FEIG models (Figure 3D). However, it is worth noting that the computational cost for the short MD protocol is much smaller than FEIG. The former used in total of 48 ns (40*1.2 ns) MD runs with implicit solvent while the latter used 1.2 μs MD runs with the TIP3P explicit solvent model.[21]

3.3 | Results in CASP12

Now we discuss our group's performance in CASP12. In CASP12, we applied our refinement protocol to all 42 refinement targets from TR520 to TR948. The starting models were obtained from CASP website (http://predictioncenter.org/casp12/). The residue length of the targets were 54–414 residues (average: 193). The initial quality of the starting models ranged from 1.15 to 13.86 Å Cα RMSD (average: 5.5 Å), 37.03 to 92.07 in terms of GDT-TS (average: 66.93), and GDT-HA ranged from 24.33 to 78.36 (average: 49.76).

First we see our performance in CASP12 relative to other groups. Table 2 summarize the average performance of top10 groups in

TABLE 3 Head-to-head comparison of the top 10 groups in CASP12

(A) Paired *t*-test on GDT-TS											
ID	**220**	**023**	**247**	**250**	**208**	**204**	**450**	**102**	**430**	**011**	**Starting model**
220	–	0.84	0.52	0.68	0.82	0.75	**0.02**	0.65	0.65	**0.02**	0.40
023	0.16	–	0.18	**0.03**	0.38	0.31	**0.00**	0.05	**0.04**	**0.00**	**0.00**
247	0.48	0.82	–	0.67	0.82	0.74	**0.01**	0.64	0.63	**0.03**	0.39
250	0.32	0.97	0.33	–	0.87	0.65	**0.01**	0.40	0.40	**0.00**	**0.04**
208	0.18	0.63	0.18	0.13	–	0.34	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**
204	0.25	0.69	0.26	0.35	0.66	–	**0.01**	0.22	0.21	**0.00**	0.05
450	0.99	1.00	0.99	0.99	1.00	1.00	–	0.99	0.99	0.90	0.98
102	0.35	0.95	0.36	0.60	0.99	0.78	**0.01**	–	0.48	**0.00**	**0.03**
430	0.35	0.96	0.37	0.60	1.00	0.79	**0.01**	0.52	–	**0.00**	**0.01**
011	0.98	1.00	0.98	1.00	1.00	1.00	0.10	1.00	1.00	–	0.98
Starting model	0.60	1.00	0.61	0.96	1.00	0.95	**0.02**	0.97	0.99	**0.02**	–
(B) Paired *t* test on GDT-HA											
ID	**220**	**023**	**247**	**250**	**208**	**204**	**450**	**102**	**430**	**011**	**Starting model**
220	–	0.86	0.66	0.61	0.87	0.81	**0.01**	0.69	0.82	**0.01**	0.30
023	0.14	–	0.29	**0.01**	0.61	0.47	**0.00**	0.09	0.31	**0.00**	**0.00**
247	0.34	0.71	–	0.43	0.76	0.69	**0.00**	0.49	0.67	**0.01**	0.18
250	0.39	0.99	0.58	–	0.98	0.81	**0.00**	0.66	0.92	**0.00**	**0.04**
208	0.13	0.39	0.24	**0.02**	–	0.40	**0.00**	**0.02**	0.13	**0.00**	**0.00**
204	0.19	0.53	0.31	0.19	0.60	–	**0.00**	0.17	0.41	**0.00**	**0.03**
450	1.00	1.00	1.00	1.00	1.00	1.00	–	1.00	1.00	0.94	0.99
102	0.31	0.91	0.51	0.34	0.99	0.83	**0.00**	–	0.92	**0.00**	**0.02**
430	0.18	0.69	0.34	0.08	0.87	0.59	**0.00**	0.08	–	**0.00**	**0.00**
011	0.99	1.00	0.99	1.00	1.00	1.00	0.06	1.00	1.00	–	0.97
Starting model	0.71	1.00	0.82	0.97	1.00	0.97	**0.01**	0.98	1.00	**0.03**	–

The performances of the top 10 groups were compared using paired Student's t-tests. Statistically significant wins (P values < 0.05) of the group listed in the first column from the left against each group listed in the first raw are shown in bold.

CASP12. Refined model quality was evaluated by GDT-TS, GDT-HA, RMS_CA, MolProbity,[33] SphereGrinder (SphGr),[34] and Quality Control Score (QCS),[35] which were the scores used in assessors' evaluation. Kiharalab was ranked eighth according to the asessors' formula (see Table 2 caption). In terms of individual scores, we were ranked fifth in GDT-TS, sixth in GDT-HA, ninth in RMS_CA, fifth in MolProbity, eighth in SphGr, and fifth in QCS. These ranking results are slightly worse than in CASP11, where our group was ranked at fourth (the CASP11 evaluation paper,[15] Table 2).

It is noted that the differences of individual scores between the top 10 groups are small. Interestingly, the difference between Kiharalab and FEIG is very small, and indeed became smaller in CASP12 relative to CASP11. For example, the average difference of GDT-HA between Kiharalab and FEIG was 2.23 in CASP11, which became as small as 0.43 this time. To reveal the significance of the differences of the group performance, we applied the paired Student's t test on the common set of predicted targets in CASP12. Table 3 shows the results on GDT-TS and GDT-HA. According to the test, our group (Kiharalab, Group ID: 102) showed significantly better results (P values < 0.05) than LEEab (Group ID: 450), LEE (Group ID: 011), and the starting model for both GDT-TS and GDT-HA. On the other hands, only SVMQA (Group ID: 208) performed significantly better than our group, and the other seven groups were indistinguishable in this test with our group.

3.4 | Analyses of short MD (model 1) models in CASP12

From this section, we analyze our submitted models in details. In Figure 4, we examined our Model 1 models generated by the short MD protocols. First we checked how many cases the Model 1 improved over starting models (Figure 4A). When evaluated by GDT-TS and GDT-HA,

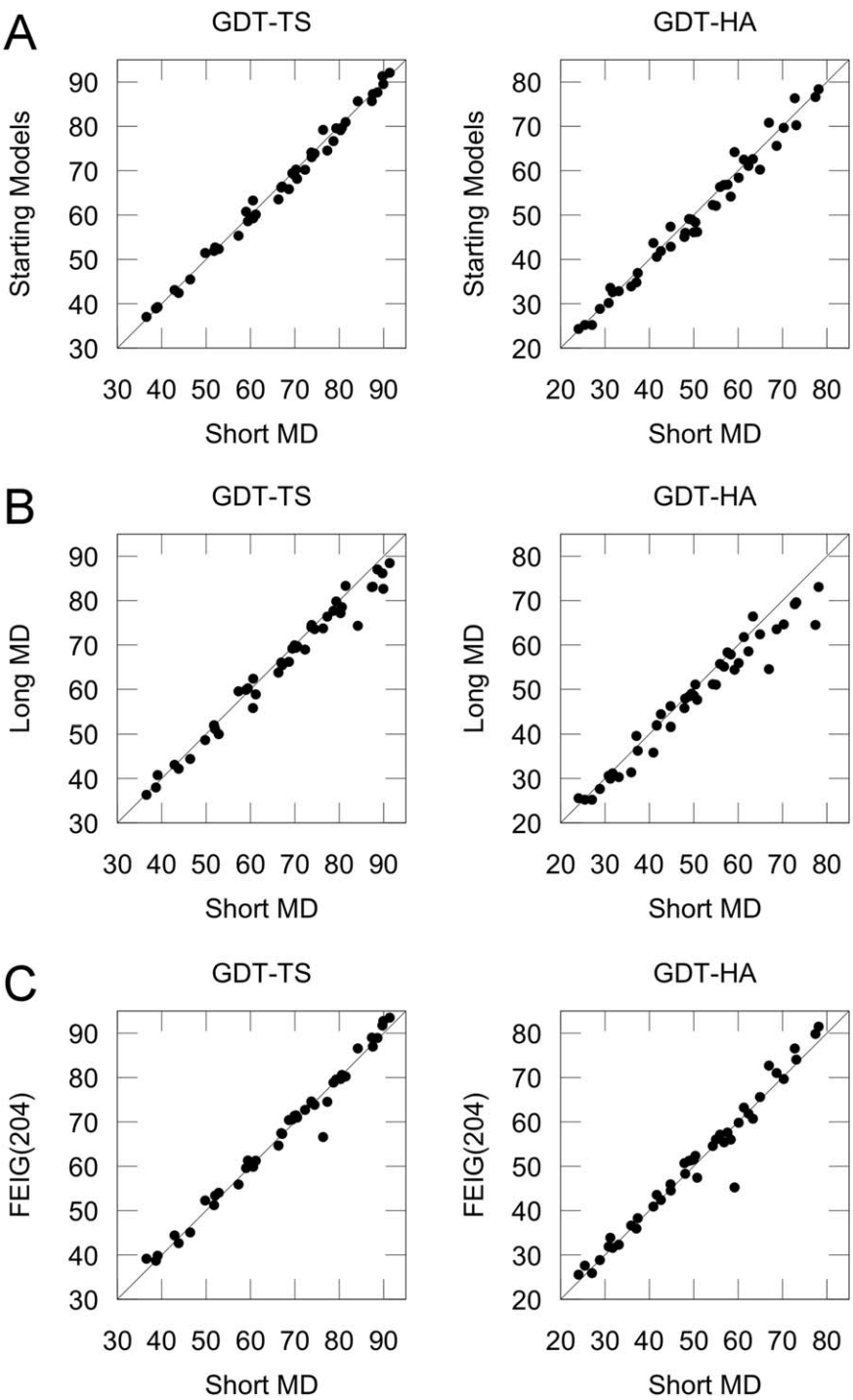

FIGURE 4 Comparison of short MD models submitted as Model 1 with other models in CASP12. GDT-TS (on the left column) and GDT-HA (right) are used for evaluation. **A**, comparison with the starting models. **B**, compared with refined models with the long MD runs that were submitted as Model 2. **C**, comparison with Model 1 models from the FEIG group

25 (59.5%) and 29 (69.0%) out of 42 targets were improved, respectively. These fractions of improved targets are similar to those observed on the CASP11 dataset (Figure 3A). Consistent with what was observed in the CASP11 dataset (Figure 3A), improvements were observed to starting models of various initial quality.

When compared with models from the long MD runs, which were submitted as Model 2 models, the short MD models were better for 30 (71.4%) and 33 (78.6%) out of 42 targets. This result is qualitatively consistent with the observation on the CASP11 dataset, but the fractions of the wins by the short MD models are lower than before (Figure 3B). To understand the better performance of the short MD over the long MD, we performed short-MD-based refinement using the constraints of 0.05 kcal mol^{-1} $Å^{-2}$, the relatively weak constraint we used in the long-MD refinement. This comparison was performed on

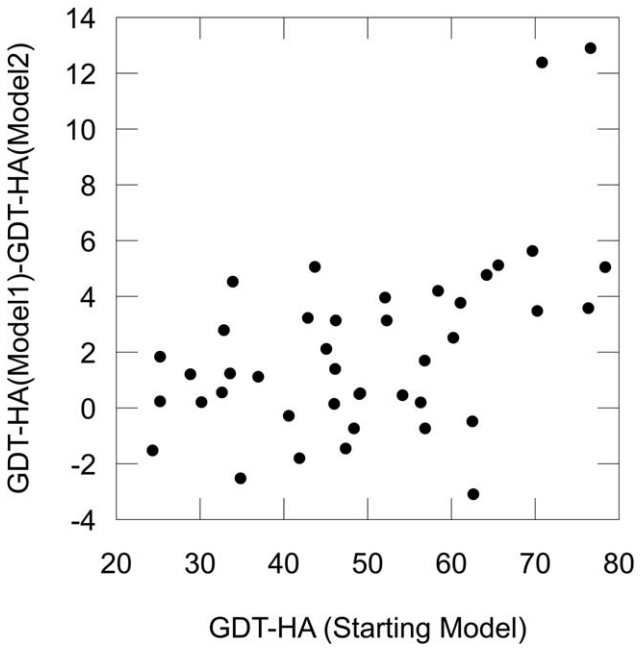

FIGURE 5 GDT-HA difference between short MD and long MD models relative to the initial quality of the targets. The x axis shows GDT-HA of the starting models. The y axis shows the difference of GDT-HA of Model 1 (short MD) and Model 2 (long MD). The positive value indicates that GDT-HA of Model 1 was higher than Model 2

36 CASP12 refinement targets that had their native structures available. TR944 had its native structure in PDB but was excluded from the analysis because its trajectory files used in CASP12 were corrupted and could not be used at the time of the post-analysis. With the increasing constraint strength of 0.1, 0.2 to 0.4 kcal mol^{-1} $Å^{-2}$, which was used in CASP12, the average GDT-TS and GDT-HA obtained for the 36 targets were 67.32 and 50.11, respectively. These values were decreased to 66.29, and 48.60 for GDT-TS and GDT-HA, respectively, when the constant 0.05 kcal mol^{-1} $Å^{-2}$ was used. Thus, we would have obtained better results with long-MD-based refinement if a stronger constraint had been used.

Long MD performed substantially worse than short MD for high-quality targets whose initial GDT-HA were >70 (Figure 5). This is probably because moving target structures far away from initial structures by the long MD was more harmful for targets that were already in high quality.

We also examined if the oligomeric state of targets affected to the refinement results. In Table 4, we show the average GDT-HA of monomer and oligomer targets in CASP12. Both short MD and long MD performed worse on the oligomer targets and the difference between the short MD- and the long MD-based refinement was smaller for the oligomer targets. Thus, it is not the case that the long MD worked particularly worse on the oligomer targets. The reason of the worse performance on the oligomer targets would probably because we applied the refinement protocol to a single target protein even for an oligomer target without considering physical interaction to other proteins. This would mean, conversely, a structure would be better refined when its interacting proteins, either other subunits in a complex if there are any, or crystal contacts in the protein crystal is considered in refinement.

TABLE 4 Average GDT-HA of monomer and oligomer targets in CASP12

	No. of Targets	Short MD	Long MD	Difference
Monomers	24	52.69	49.82	2.88
Oligomers	18	47.49	46.32	1.17

Among the 48 refinement targets, the following targets are oligomers: TR520, 594, 694, 862, 866, 870, 875, 877, 881, 887, 893, 896, 909, 912, 913, 917, 945, 947 (all IDs with TR as prefix).

Next, we compared with the Model 1 models from FEIG, which used a similar MD-based refinement protocol (Figure 4C). The FEIG models were better than Kiharalab in 27 (64.3%) and 25 (59.5%) out of 42 targets in terms of GDT-TS and GDT-HA, respectively. This difference between the two groups is much smaller than on the CASP11 dataset (Figure 3D), which is consistent with the results shown as overall average scores in Table 2 and the results in Table 3 that shows the two groups were statistically indistinguishable.

3.5 | Examples of refined models

Examples of models with successful and failed refinement are shown in Figure 6. Improved and deteriorated regions in a model are colored in blue and red, the darker more substantial. Green regions are those which did not move >0.1 Å. The first example in the top row is TR948. For this target, the overall improvement of GDT-HA (dGDT-HA) was large, 4.70, and this model was ranked sixth among all the first models submitted by the 36 groups. As shown in the color code, improvement occurred at almost all the helical regions in the structure and degradation occurred at ends of some helices. The next one, TR912, is another successful example, where dGDT-HA of 4.58 was achieved. This model was the best in GDT-HA among all the 31 groups' Model 1 models. Similar to the previous example, improvement occurred globally, at almost all β-strands in the structure. In TR868, the third example, there was a modest improvement where dGDT-HA at 0.72 was observed. This model was ranked seventh among 34 groups who submitted models for this target. In this refined model, improved and degraded regions co-exist, mixed in the structure, which is typically observed in models with a small improvement of <1.0 dGDT-HA. The last one, TR891, is a case that our refinement failed. The model had a dGDT-HA of −3.57 and ranked 18th among 36 groups. As shown, the refinement protocol moved all the β-sheets away from the native structure. We observed failure for other three targets of β-barrel structures, too (TR879, dGDT-HA of −5.0; TR891, −3.57; TR928, −2.79). Overall, these examples illustrate that improvement and deterioration occurred globally in a model including the structure core rather than merely moving flexible loop regions.

3.6 | Relaxing or compressing?

Observing that the structural change occurs globally to a model by our refinement protocol, we questioned whether the observed

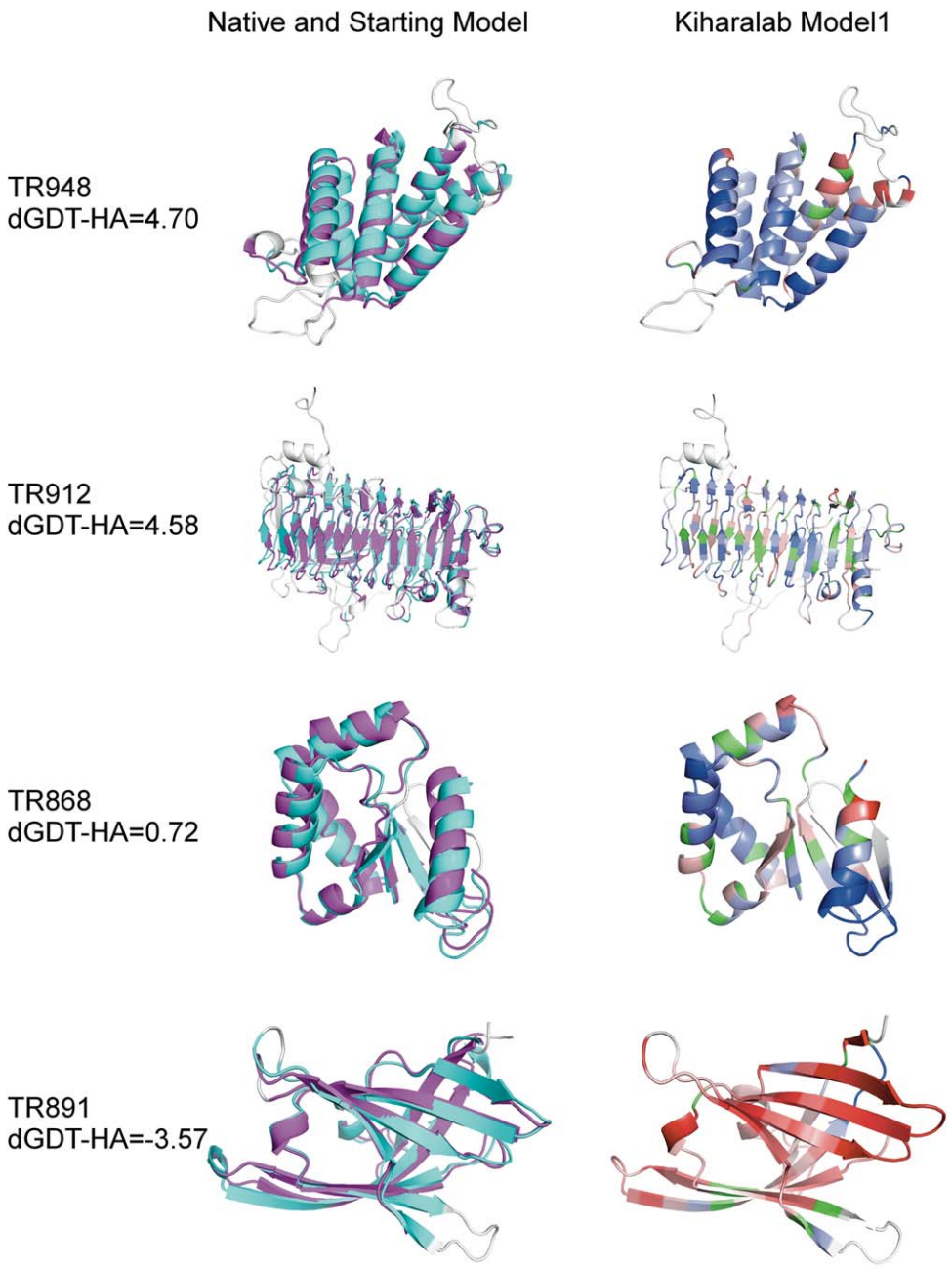

FIGURE 6 Examples of successful and failed refinement by our group. The left column shows the native structures (magenta) and the starting models (cyan). The right column shows our Model 1 models that were refined with the short MD runs. The refined models are colored according to the degree of improvement of Cα atom positions from the starting models. Improved regions in a model are colored from light blue to dark blue for small to large improvements. On the other hand, deteriorated regions are colored from light to dark red for slight to large deteriorations. Green represents regions that did not change >0.1 Å. Deviation of the Cα positions of a model from the starting structure was judged after superimposing the starting structure and the model to the native structure using the LGA program with a 4.0 Å threshold

improvement was due to simply compression of the structure since it is widely known that compression of Cα coordinates of a model decreases the radius of gyration, which contributes to apparent improvement of some quality assessment scores, such as GDT-TS, GDT-HA, and RMSD. To answer this question, we evaluated a compactness of refined models and the starting models by comparing the radius of gyration (R_g) of the structures. It is defined as:

$$R_g = \sqrt{\frac{\sum_i^N |v_i - \overline{v}|^2}{N}} \quad (5)$$

where N is a total number of Cα atoms in the model, v_i is the coordinate of ith Cα atom, $\overline{v}$ is the average coordinate of all Cα atoms of the model. In this calculation, we ignored largely incorrect regions in the

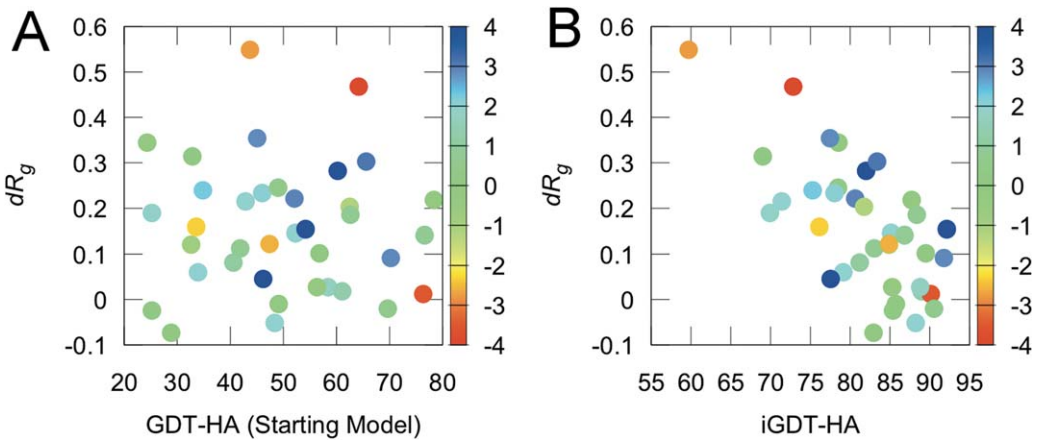

FIGURE 7 Refinement results relative to the change of radius of gyration of models. Improvement of models (dGDT-HA) is shown in color code relative to the change of the radius of gyration (dR_g) by the short MD refinement protocol. dGDT-HA of over −4.0 to >4.0 is shown in a color scale from red to blue. Each data point represents a Model 1 refined model for the 37 targets that have their native structure available for the analysis. **A**, The *x* axis shows the quality, GDT-HA of starting models. **B**, The *x* axis is the structural deviation of refined models form the starting model (iGDT-HA)

starting model where the distance between the corresponding Cα atom positions to the native structure was larger than 4.0 Å, because these regions are highly unlikely to influence refined model's GDT-TS and more so for GDT-HA. To assess the compression or relaxation of refined models, we computed the difference of R_g between the starting model and the refine model (dR_g), $dR_g = R_g$(refined model) − R_g(starting model). A positive dR_g indicates that the refined model was relaxed (expanded) while a negative value shows the model was compressed from the starting model.

In Figure 7A, the improvement of models (dGDT-HA) was presented in a color code relative to dR_g and GDT-HA of starting models. First, by examining dR_g, models for 32 out of 37 targets (five targets were excluded from this analysis because their native structures were not available for computing GDT-HA of their starting models) have positive values, indicating that the refinement actually expanded (or relaxed), not compressed the structures. This figure also shows that the degree of the relaxation did not depend on the quality of the starting models and larger improvement (points in blue) occurred for starting models with a middle range GDT-HA and dR_g, namely about 50 and 0.2, respectively. Figure 7B compares dGDT-HA with dR_g and the deviation of refined models from their starting structures (iGDT-HA). There is an obvious correlation between iGDT-HA and dR_g, which simply shows that the model drifted away from the initial structure as it expanded (larger dR_g). An interesting observation is that two

TABLE 5 Average performance of Kiharalab model 1–5 in CASP12

	GDT-TS	GDT-HA	RMS_CA	MolProbity	SphGr	QCS
Model 1[a]	**67.33**	**50.46**	5.52	**1.45**	66.80	**80.44**
Model 2[b]	65.84	48.32	5.51	1.80	66.11	79.76
Model 3[c]	65.29	47.19	5.59	2.01	65.85	78.79
Model 4[d]	64.37	46.22	5.63	1.94	65.40	76.73
Model 5[e]	66.16	48.38	5.56	2.10	65.59	79.84
Starting Model	66.93	49.76	**5.50**	1.77	**66.99**	79.74

The best result among Model 1–5 and the starting model for each score is shown in bold.
[a]Averaged and relaxed model generated from the subset of short MD trajectories.
[b]Averaged and relaxed model generated from the subset of long MD trajectories.
[c]The lowest DFIRE model.
[d]The lowest GOAP model.
[e]The highest iGDT-HA model.

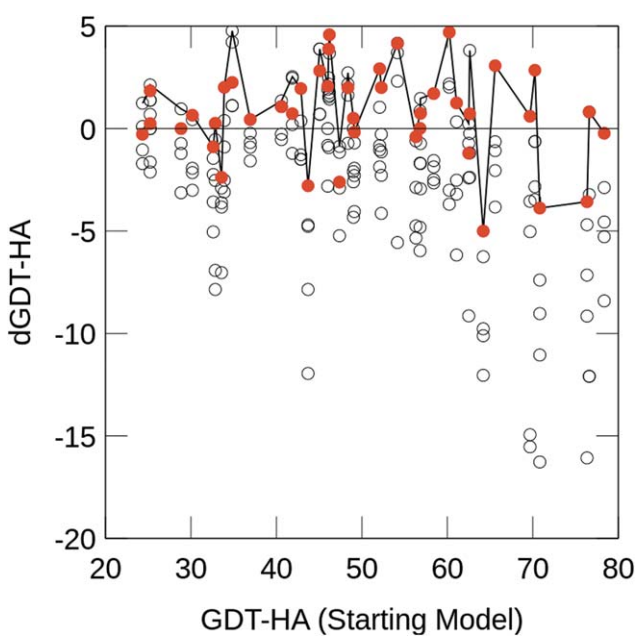

FIGURE 8 Improvement of GDT-HA (dGDT-HA) relative to the quality of starting models (GDT-HA(Starting Model)). Red points represent Model 1 models and open circles represent Model 2 to Model 5. The best dGDT-HA models among the five models for each target are connected with a line

TABLE 6 Results of using different structure sampling form short MD runs

	Cα restraints[a]	GDT-TS	GDT-HA
Section 1 (400 ps)	0.1	66.73	49.21
Section 1, 2 (800 ps)	0.1, 0.2	67.06	49.70
Section 1, 2, 3 (1.2 ns)	0.1, 0.2, 0.4	67.32	50.11
Section 2 (400 ps)	0.2	67.11	49.78
Section 3 (400 ps)	0.4	67.27	50.08
Starting model	n/a	67.06	49.80

In the short MD runs of 1.2 ns in total, as described in Methods, we applied an increasing Cα constraints from 0.1, 0.2, to 0.4 to three 400 ps-long subsections sequentially. In this small experiment, structures are sampled from different subsections of the MD runs, which underwent the structure averaging and relaxing procedure to yield a final model. GDT-TS and GDT-HA are average of models for 36 targets, which had the native structure available from PDB. TR944 has its native structure in PDB but was excluded from this data, because the trajectory files used in CASP12 were corrupted and could not be used.
[a]The unit is kcal/mol/Å^2. If more than one value is shown, each was applied subsequently to each subsection in MD trajectories.

unsuccessful refined models (TR879, dGDT-HA: −5.0; TR928, dGDT-HA: −2.8), which are shown in red and orange, are found at high dR_g (0.47 and 0.55, respectively) and low iGDT-HA (59.7 and 73.8, respectively), distinct from the other models, in Figure 7B. This result suggests that models of unsuccessful refinement may be better identified by the combination of iGDT-HA and dR_g rather than only using iGDT-HA. This idea works particularly well for distinguishing TR879 (the red data point) from the other models that have a similar iGDT-HA value.

3.7 | What went right and what went wrong

Following the tradition of the CASP predictors' reports, we discuss things that worked well and those which need improvement.

One thing which clearly worked well was the ranking of the submitted models. Table 5 summarizes the average of evaluation scores of Model 1 to 5 and Figure 8 presents dGDT-HA of individual models relative to the quality (GDT-HA) of the starting models. In the figure, Model 1 models are shown in red and the best model (i.e., the model with the largest dGDT-HA) for each target are connected with lines. From Table 5, Model 1 models were on average the best among the five submitted models for all the evaluation scores except for that of the RMS_CA, where Model 1 was ranked second, following Model 2. Figure 8 visualizes the same conclusion; 28 out of 42, of the Model 1 models were the best for the targets, and if not they were close to the best.

Second, as discussed with Figure 4A, our refinement protocol with short MD runs improved models for most of the cases. Additionally, the structure sampling from short MD runs with an increasing Cα constraints, which increased from 0.1, 0.2 to 0.4 kcal mol^{-1} Å^{-2} for every 400 ps, worked well. As shown in Table 6, sampling structures from different portions of MD runs we tried all worked worse than the method we used. Thus, overall, we can conclude that we were successful in exploiting inexpensive short MD runs with an implicit solvent model very effectively. This point becomes evident when our group's results were compared with a contrasting approach that used significantly more expensive MD runs but had similar performance (Table 2).

On the other hand, by design our protocol could not make large refinement to models due to the use of short MD simulations. To make substantial corrections to a model conformation, such as rearrangement of secondary structures or domain moves, a completely different algorithm design, probably with a different protein model, such as a coarse-grained model,[2,36,37] is obviously needed. Indeed, this is the challenge left for the whole CASP refinement community.

4 | CONCLUSION

We discussed our group's performance in the CASP12 refinement category. Our protocol makes use of inexpensive short MD simulations with implicit solvent and successfully showed consistent improvements to starting models regardless of the quality of the starting models. By examining submitted models, we found that achieved improvements are due to relaxation of structures rather than compression, which also suggested that the degree of relaxation (dR_g) could be another metric to eliminate unsuccessful refined models. However, the protocol does not make large conformational refinement by design due to the use of short MD trajectories, which is still the goal of further development.

ACKNOWLEDGMENTS

The authors thank Kevin Shim for proofreading the manuscript. This work was partly supported by National Institutes of Health (R01GM123055) and National Science Foundation (IIS1319551, DBI1262189, IOS1127027, DMS1614777).

ORCID

Daisuke Kihara http://orcid.org/0000-0003-4091-6614

REFERENCES

[1] Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV. Evaluation of free modeling targets in CASP11 and ROLL. *Proteins*. 2016;84(Suppl 1):51–66.

[2] Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci USA*. 2001;98(18):10125–10130.

[3] Modi V, Xu Q, Adhikari S, Dunbrack RL Jr.Assessment of template-based modeling of protein structure in CASP11. *Proteins*. 2016;84 (Suppl 1):200–220.

[4] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction: progress and new directions in round XI. *Proteins*. 2016;84(Suppl 1):4–14.

[5] Qu X, Swanson R, Day R, Tsai J. A guide to template based structure prediction. *Curr Protein Peptide Sci*. 2009;10(3):270–285.

[6] Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. *Proteins*. 1995;23(3):2–4.

[7] Kryshtafovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. *Proteins*. 2007;69(Suppl 8):194–207.

[8] Kryshtafovych A, Fidelis K, Moult J. CASP8 results in context of previous experiments. *Proteins*. 2009;77(Suppl 9):217–228.

[9] Kryshtafovych A, Fidelis K, Moult J. CASP9 results compared to those of previous CASP experiments. *Proteins*. 2011;79(Suppl 10): 196–207.

[10] Kryshtafovych A, Fidelis K, Moult J. CASP10 results compared to those of previous CASP experiments. *Proteins*. 2014;82(Suppl 2): 164–174.

[11] Kihara D, ed. *Protein Structure Prediction*. New York: Humana Press; 2014.

[12] Padilla-Sanchez V, Gao S, Kim HR, et al. Structure-function analysis of the DNA translocating portal of the bacteriophage T4 packaging machine. *J Mol Biol*. 2014;426(5):1019–1038.

[13] Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*. 2000;29:291.

[14] Baker D, Sali A. Protein structure prediction and structural genomics. *Science*. 2001;294(5540):93–96.

[15] Modi V, Dunbrack RL Jr. Assessment of refinement of template-based models in CASP11. *Proteins*. 2016;84(Suppl 1):260–281.

[16] an H, Mark AE. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci*. 2004;13 (1):211–220.

[17] MacCallum JL, Perez A, Schnieders MJ, Hua L, Jacobson MP, Dill KA. Assessment of protein structure refinement in CASP9. *Proteins*. 2011;79(Suppl 10):74–90.

[18] Nugent T, Cozzetto D, Jones DT. Evaluation of predictions in the CASP10 model refinement category. *Proteins*. 2014;82(Suppl 2):98–111.

[19] Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins*. 2007;69(Suppl 8):38–56.

[20] Mirjalili V, Noyes K, Feig M. Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. *Proteins*. 2014;82(Suppl 2):196–207.

[21] Feig M, Mirjalili V. Protein structure refinement via molecular-dynamics simulations: What works and what does not? *Proteins*. 2016;84(Suppl 1):282–292.

[22] Khoury GA, Smadbeck J, Kieslich CA, et al. Princeton_TIGRESS 2.0: High refinement consistency and net gains through support vector machines and molecular dynamics in double-blind predictions during the CASP11 experiment. *Proteins*. 2017;85(6):1078–1098.

[23] Joung I, Lee SY, Cheng Q, et al. Template-free modeling by LEE and LEER in CASP11. *Proteins*. 2016;84(Suppl 1):118–130.

[24] Della Corte D, Wildberg A, Schroder GF. Protein structure refinement with adaptively restrained homologous replicas. *Proteins*. 2016;84(Suppl 1):302–313.

[25] Lee GR, Heo L, Seok C. Effective protein model structure refinement by loop modeling and overall relaxation. *Proteins*. 2016;84 (Suppl 1):293–301.

[26] Haberthur U, Caflisch A. FACTS: fast analytical continuum treatment of solvation. *J Computat Chem*. 2008;29(5):701–715.

[27] Hassan SA, Guarnieri F, Mehler EL. A general treatment of solvent effects based on screened coulomb potentials. *J Phys Chem B*. 2000;104:6478–6489.

[28] Hua DP, Huang H, Roy A, Post CB. Evaluating the dynamics and electrostatic interactions of folded proteins in implicit solvents. *Protein Sci*. 2016;25(1):204–218.

[29] Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci 2002*.11(11):2714–2726.

[30] Zhang J, Zhang Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS One*. 2010;5(10):e15386.

[31] Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys J*. 2011;101(8):2043–2052.

[32] Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003;31(13):3370.

[33] Chen VB, Arendall WB III, Headd JJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallograph Sect D Biol Crystallogr*. 2010;66(Part 1):12–21.

[34] Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins*. 2014;82(Suppl 2):7–13.

[35] Cong Q, Kinch LN, Pei J, et al. An automatic method for CASP9 free modeling structure prediction assessment. *Bioinformatics*. 2011; 27(24):3371–3378.

[36] Kolinski A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol*. 2004;51(2):349–371.

[37] Sieradzan AK, Krupa P, Scheraga HA, Liwo A, Czaplewski C. Physics-based potentials for the coupling between backbone- and side-chain-local conformational states in the united residue (UNRES) force field for protein simulations. *J Chem Theory Computat*. 2015; 11(2):817–831.

How to cite this article: Terashi G, Kihara D. Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. *Proteins*. 2018;86:189–201. https://doi.org/10.1002/prot.25373