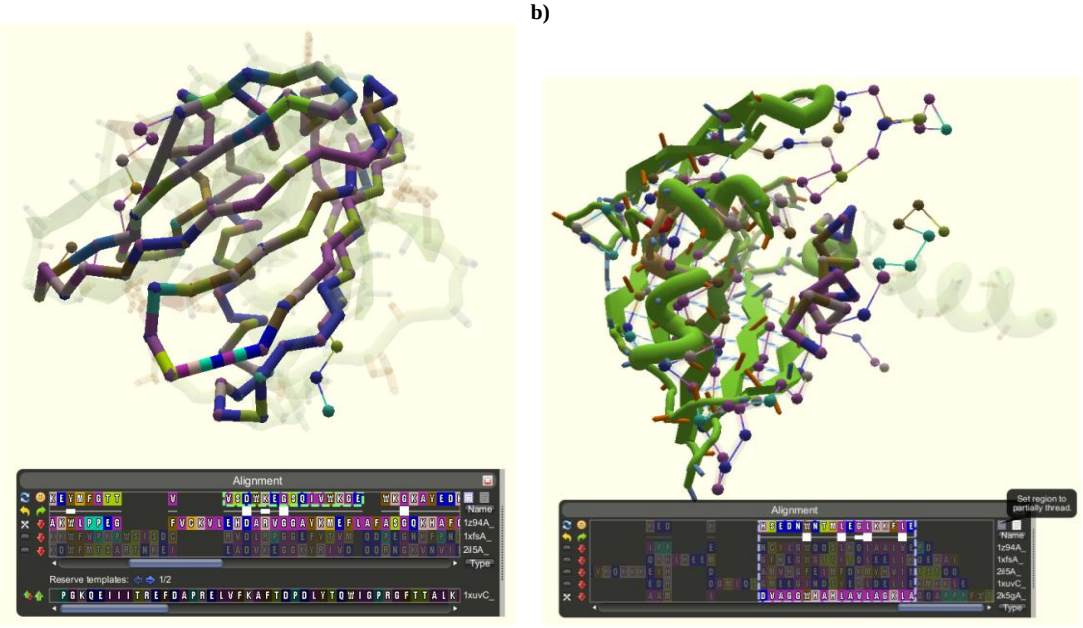
Crystal structure of a monomeric retroviral protease solved by protein folding game players

Firas Khatib, Frank DiMaio, Foldit Contenders Group, Foldit Void Crushers Group, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, Mariusz Jaskolski, David Baker

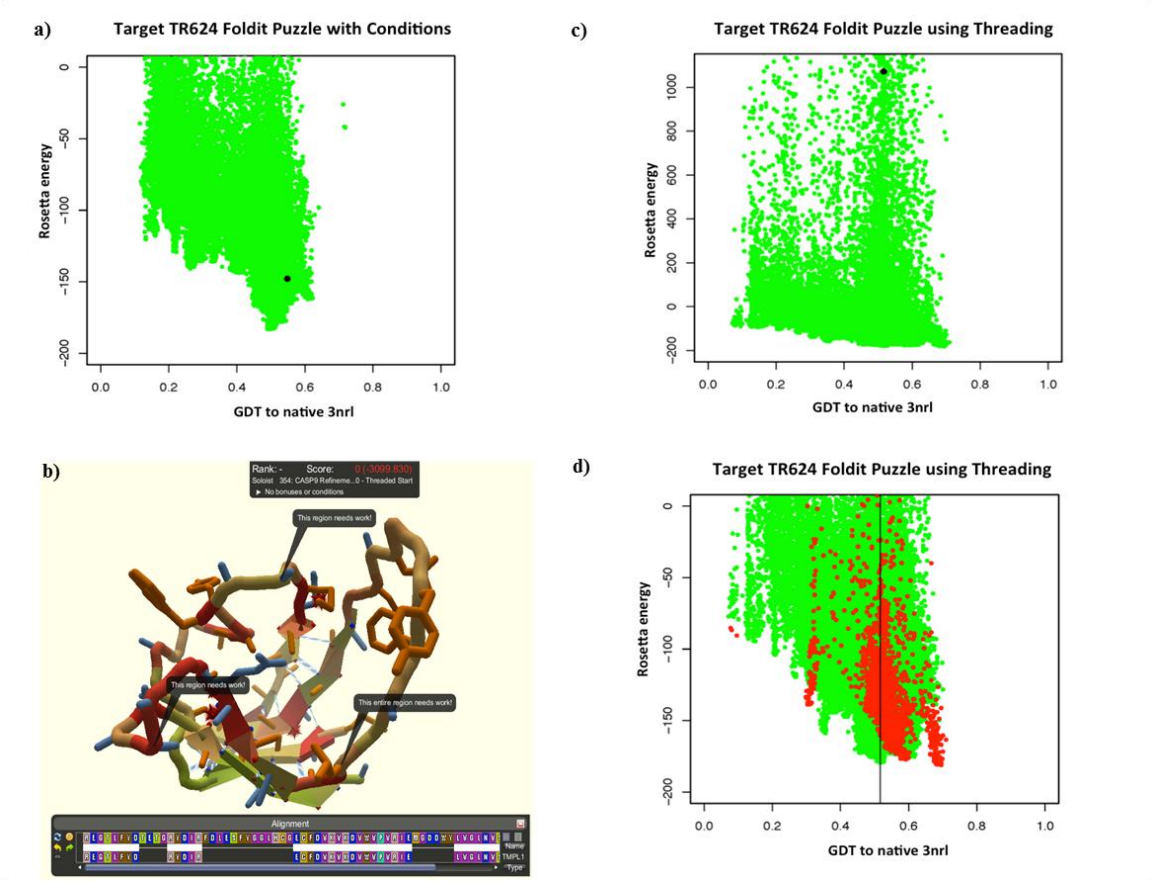
Supplementary Figures:

a) **b)**



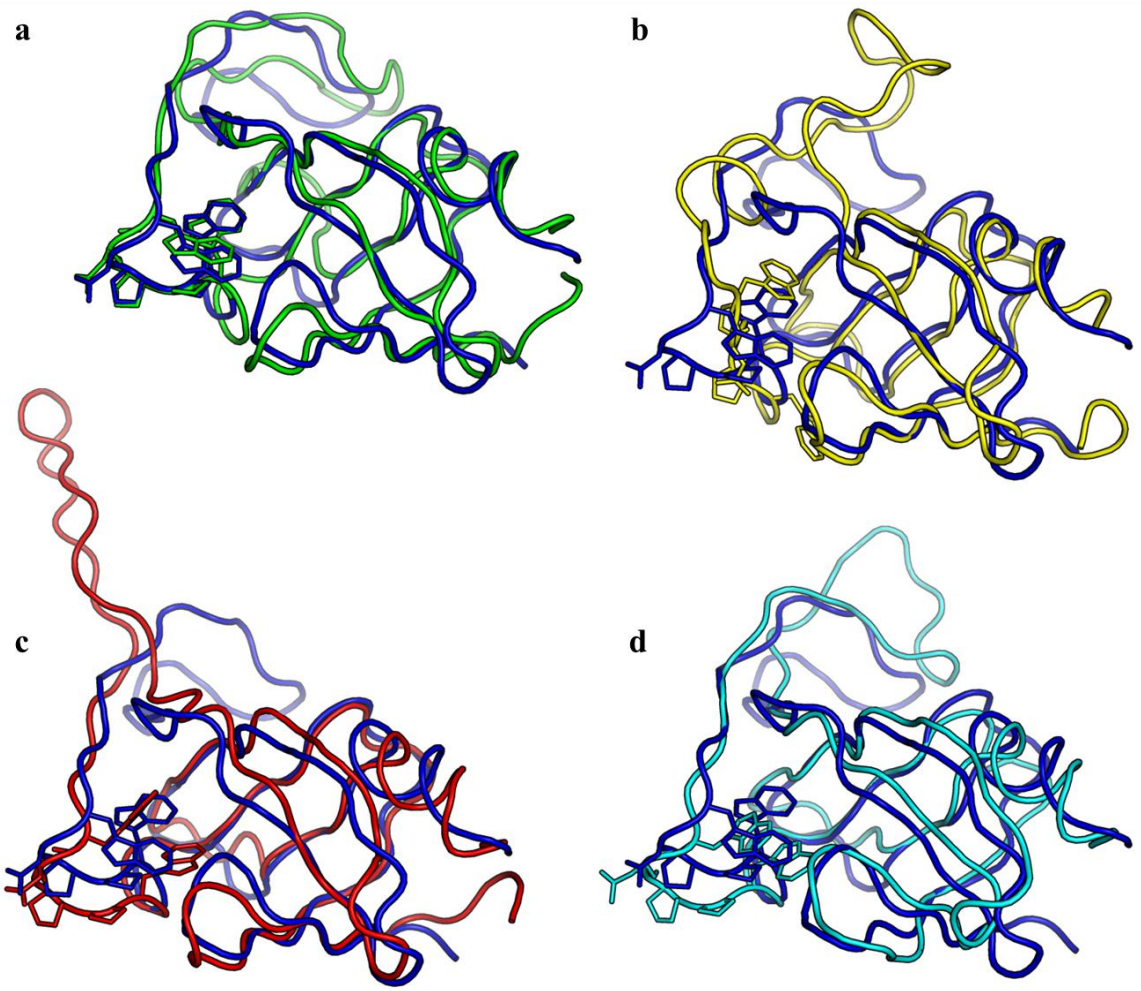
Supplementary Figure 1 Foldit screenshot of the Alignment Tool. The Alignment Tool allows Foldit players to load in different templates and manually move alignments. Players are then able to thread their sequence onto the structures of these known homologs. (**a**) When a template is selected, the aligned regions are represented as cylinders in the game while any unaligned regions are shown as spheres connected by lines; these graphical representations change in real time as players select residues and move the alignments around in the Alignment Tool. (**b**) During CASP9, Foldit players requested the ability to thread only a specific region from one template so partial threading was added to the Alignment Tool; this allows players to combine different regions from multiple templates into one hybrid model.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2119



Supplementary Figure 2 (**a**) **Foldit results for CASP9 Refinement target TR624 using RMSD conditions.** Foldit player predictions are shown in green. The starting refinement model is shown as a black dot, with the *x*-axis representing GDT (Global Distance Test, where a prediction of 1.0 would match the native conformation) and the *y*-axis representing Rosetta energy (the same energy function used in the game). Foldit players had trouble improving the score because the starting Rosetta model had already been minimized with the Rosetta energy function. The top Foldit scoring model (lowest Rosetta energy) was farther from the native structure than the starting model. (**b**) **CASP9 refinement puzzle for TR624 created using the Alignment Tool.** We reposted refinement target TR624 as a separate Foldit puzzle using a different starting model. Instead of the Rosetta model provided by the CASP organizers, we used the Alignment Tool to only align the regions of that model that were deemed correct by the organizers; the unaligned portions were rebuilt randomly, resulting in poor energies and encouraging diversification of models in these incorrect regions. (**c**) **Foldit results for CASP9 Refinement target TR624 using the Alignment Tool as a puzzle start.** Foldit player predictions are shown in green. The starting model, created using the Alignment Tool, is shown in black, with the *x*-axis representing GDT and the *y*-axis representing Rosetta energy (note the poor starting energy). (**d**) **Zoomed in plot of the top scoring Foldit solutions in (c) highlighting solutions by the Foldit Void Crushers Group in red.** Using this modified starting model, the Void Crushers Group was able to come up with a top scoring Foldit model that was closer to the native structure than any other CASP9 prediction submitted for TR624.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2119



Supplementary Figure 3 Superpositions of different best attempts to solve the structure of M-PMV PR by molecular replacement. Selected models shown were the best predictions by optimally-superimposed Phaser LLG. The later determined crystal structure is shown in blue in all four panels. (**a**) In green: Foldit prediction by mimi (same model shown in **Figure 2d**) highlighting the accuracy of the core side-chain overlap in the loop at the bottom left. (**b**) In yellow: closest Rosetta prediction from the rebuild-and-refine protocol[1] starting from the NMR ensemble. (**c**) In red: closest Rosetta prediction from the relax protocol[2] starting from the NMR ensemble. (**d**) In cyan: closest Rosetta prediction from CS-Rosetta[3] starting from the NMR ensemble and using chemical shifts as restraints.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2119

Supplementary Table:

Table 1 Data collection and refinement statistics (molecular replacement)

	M-PMV PR
Data collection	
Space group	$P2_1$
Cell dimensions	
a, *b*, *c* (Å)	26.76, 86.62, 39.31
α, β, γ (°)	90, 104.6, 90
Resolution (Å)	43.3-1.63(1.73-1.63)*
R_{merge}	0.068(0.752)
$I / \sigma I$	14.9(1.9)
Completeness (%)	99.0(96.3)
Redundancy	4.7(4.2)
Refinement	
Resolution (Å)	28.6-1.63
No. reflections	21,365
R_{work} / R_{free}	0.1694 / 0.2124
No. atoms	
Protein	1,527
Ligand/ion	-
Water	154
B-factors (Å^2)	
Protein	28.4
Ligand/ion	-
Water	34.6
R.m.s. deviations	
Bond lengths (Å)	0.018
Bond angles (°)	1.77

*One single crystal used for data collection. *Values in parentheses are for highest-resolution shell.

Supplementary Discussion:

The most important structural improvement made by the Contenders Foldit Group for M-PMV PR was tucking in the loop at the top (shown in **Figure 2d**). This loop was the major difference between all the predictions generated by Rosetta, which failed to solve the structure using MR[4], and the successful Foldit prediction that solved the crystal structure. This long loop is shown at the top of each panel in **Supplementary Figure 3**, where the best models (by optimally-superimposed Phaser LLG) using four different prediction methods are each superimposed onto the recently solved crystal structure (shown in each panel in blue). The green model in **Supplementary Figure 3a** is the Foldit prediction that was used to solve the structure; it shows the accurate side-chain placements in the loop at the bottom left. This loop was another region where players from the Contenders Foldit Group were able to outperform the automated Rosetta methods. None of these best Rosetta predictions by optimally-superimposed Phaser LLG (**Supplementary Fig. 3b-d**) were able to accurately predict the conformation of this loop.

Supplementary References:

1. B. Qian *et al.*, *Nature*. **450**, 259-264 (2007).
2. B. Kuhlman and D. Baker, *Proc. Natl. Acad. Sci.* **97**, 10383-10388 (2000).
3. Y. Shen *et al.*, *Proc. Natl. Acad. Sci.* **105**, 4685-4690 (2008).
4. F. DiMaio *et al.*, *Nature* **473**, 540-543 (2011).

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2119