Biophysical Journal, Volume *121*

Supplemental information

DeepTracer-ID: De novo protein identification from cryo-EM maps

Luca Chang, Fengbin Wang, Kiernan Connolly, Hanze Meng, Zhangli Su, Virginija Cvirkaite-Krupovic, Mart Krupovic, Edward H. Egelman, and Dong Si

Supporting Material

DeepTracer ID: De Novo Protein Identification from Cryo-EM Maps

Luca Chang[1,†], Fengbin Wang[2,†,*], Kiernan Connolly[1], Hanze Meng[3], Zhangli Su[4], Virginija Cvirkaite-Krupovic[5], Mart Krupovic[5], Edward H. Egelman[2,*], Dong Si[1,*]

[1]Division of Computing and Software Systems,

University of Washington Bothell, Bothell, WA 98011, USA

[2]Department of Biochemistry and Molecular Genetics,

University of Virginia School of Medicine, Charlottesville, VA 22903, USA

[3]Department of Mathematics,

University of Washington, Seattle, WA 98105, USA

[4]Department of Genetics,

University of Alabama at Birmingham, Birmingham, AL 35233, USA

[5]Institut Pasteur, Université de Paris, CNRS UMR6047,

Archaeal Virology Unit, 75015 Paris, France

[†]These authors contributed equally

Table S1. Cryo-EM and Refinement Statistics of *A. pernix* flagellum and AFV6 filaments

Parameter	***A. pernix* flagellum**	**AFV6**
Data collection and processing		
Voltage (kV)	300	300
Electron exposure (e^{-} Å^{-2})	50	50
Pixel size (Å)	1.08	1.4
Particle images (n)	59,338	78,141
Shift (pixel)	8	10
Helical symmetry		
Point group	C1	C1
Helical rise (Å)	5.52	5.75
Helical twist (°)	108.0	38.46
Map resolution (Å)		
Map:map FSC (0.143)	3.5	3.9
Model:map FSC (0.38)	3.7	4.2
d_{99}	3.9	4.1
Refinement and Model validation		
Ramachandran Favored (%)	93.2	93.1
Ramachandran Outliers (%)	0.5	0.6
RSCC	0.82	0.85
Clashscore	9.6	12.6
Bonds RMSD, length (Å)	0.004	0.006
Bonds RMSD, angles (°)	0.732	0.781
Deposition ID		
PDB (model)	7TXI	7TXJ
EMDB (map)	EMD-26158	EMD-26159

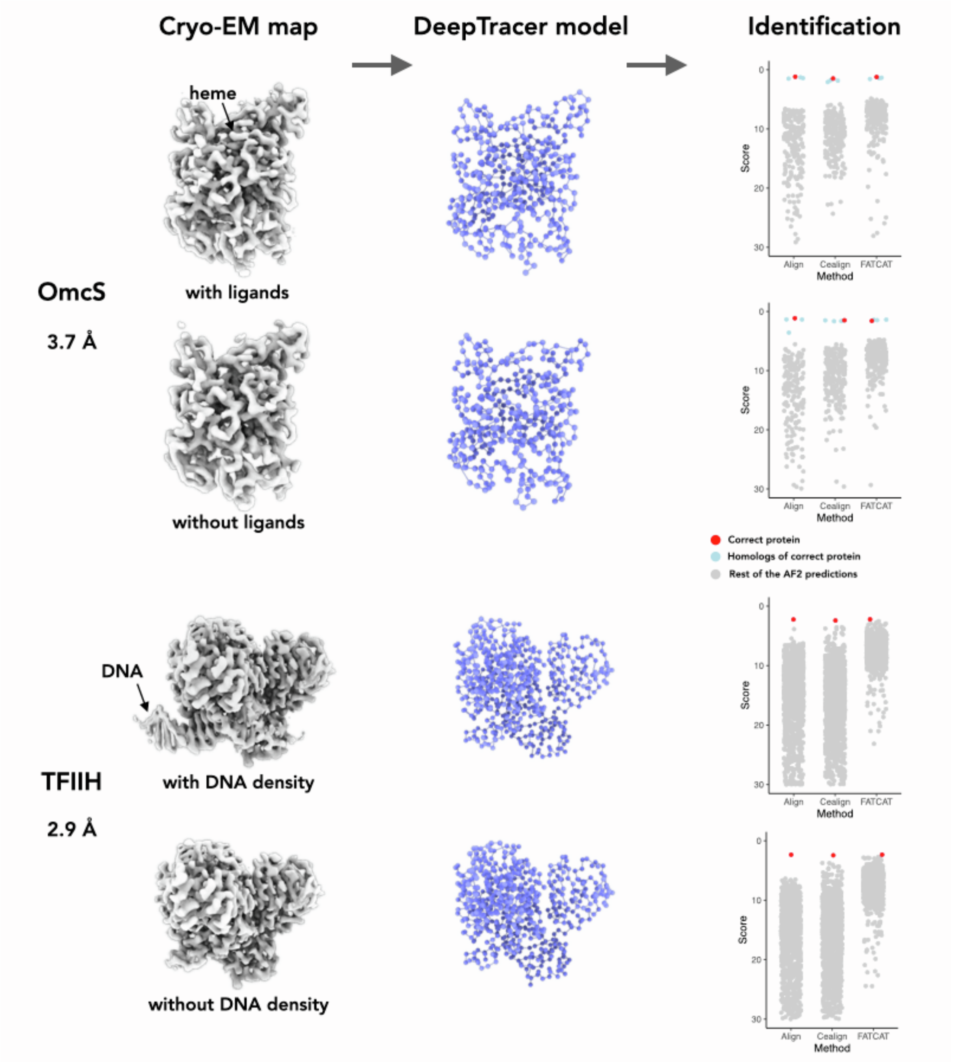

Figure S1 Protein identification is not affected by ligand or nucleic acids densities

Left, cryo-EM maps used for *de novo* protein identification with their reported resolution. OmcS has six hemes per protein subunit. TFIIH protein shown with and without bound dsDNA. Middle, the Cα backbone of the model generated by DeepTracer, from the maps on left. Some extra resides were assigned in the heme area by DeepTracer, while the dsDNA densities are recognized as non-protein area so very few residues were placed there. Right, the DeepTracer-ID scores of AF2 predictions. The correct protein is shown by a red dot, the proteins with significant structural similarity to the correct protein are shown as blue dots, and the remaining AF2 predictions are shown as grey dots. The size of AF2 library and the corresponding organism for the eight benchmark datasets are: OmcS (*G. sulfurreducens PCA,* N=226), and TFIIH (*H. sapiens*, N=1347).

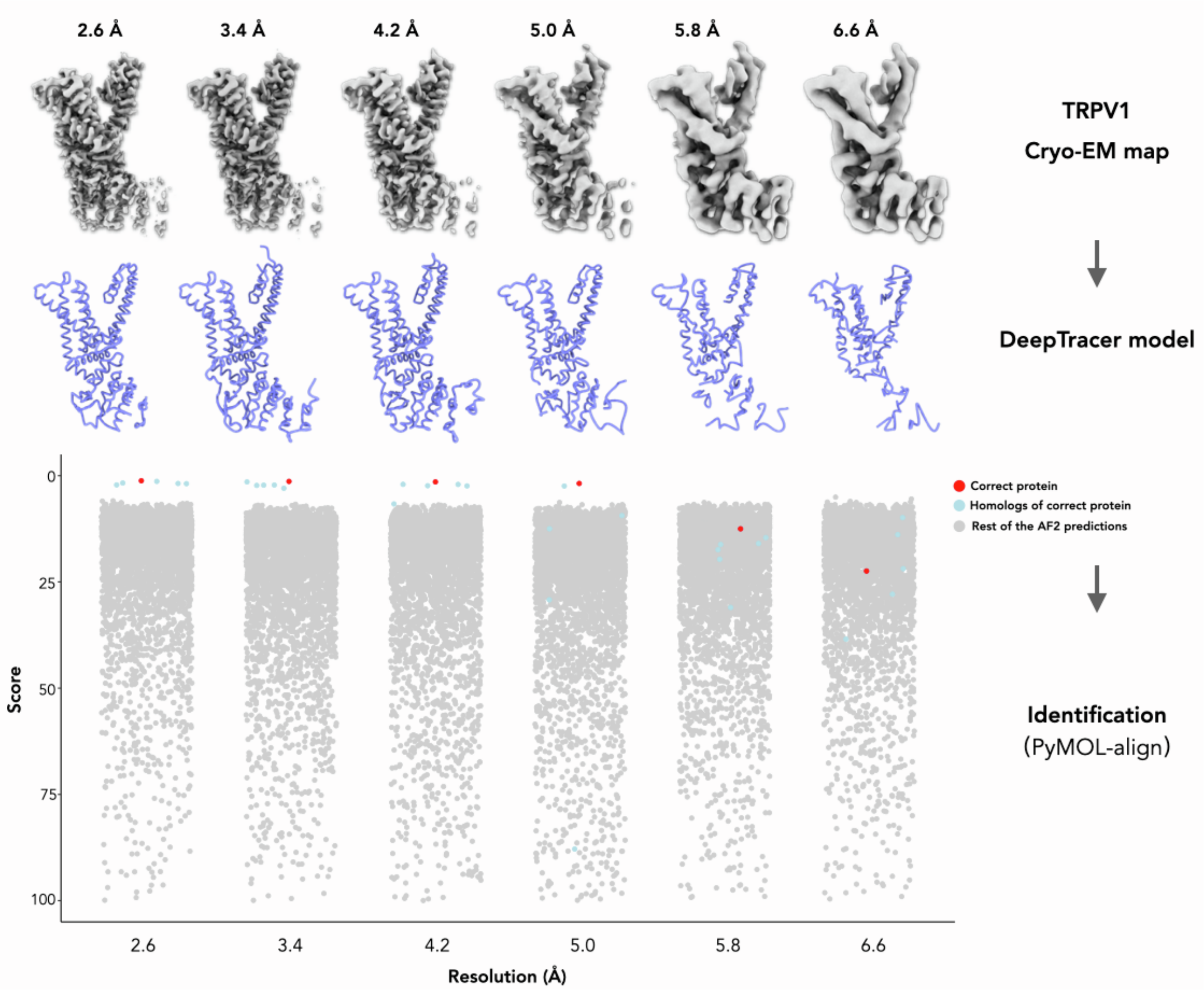

Figure S2 DeepTracer-ID working on TRPV1 maps filtered to different resolution

Top, the segmented TRPV1 maps at different resolutions. Middle, the corresponding DeepTracer model generated from the maps on top. Bottom, the DeepTracer-ID scores of AF2 predictions using PyMOL-align method. (*R. norvegicus*, N=3679)

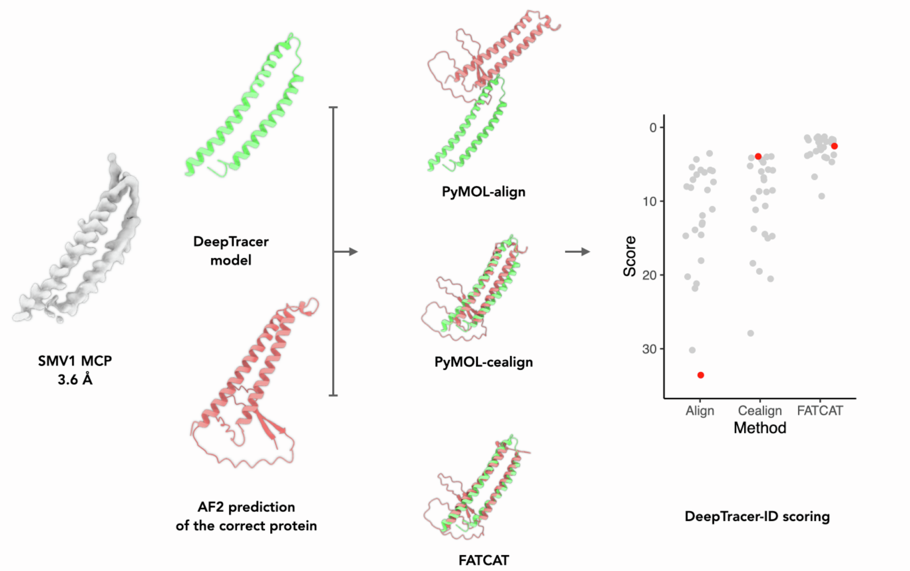

Figure S3 Identifying very small proteins relies on the initial 3D alignments

Far Left, the segmented cryo-EM map of SMV1 major capsid protein with the reported resolution. Left, the DeepTracer model (green) and the AF2 model of the correct protein (red). Right, How the AF2 model is aligned to the DeepTracer model using three different approaches. Far right, the DeepTracer-ID scores of AF2 predictions (SMV1, N=28).