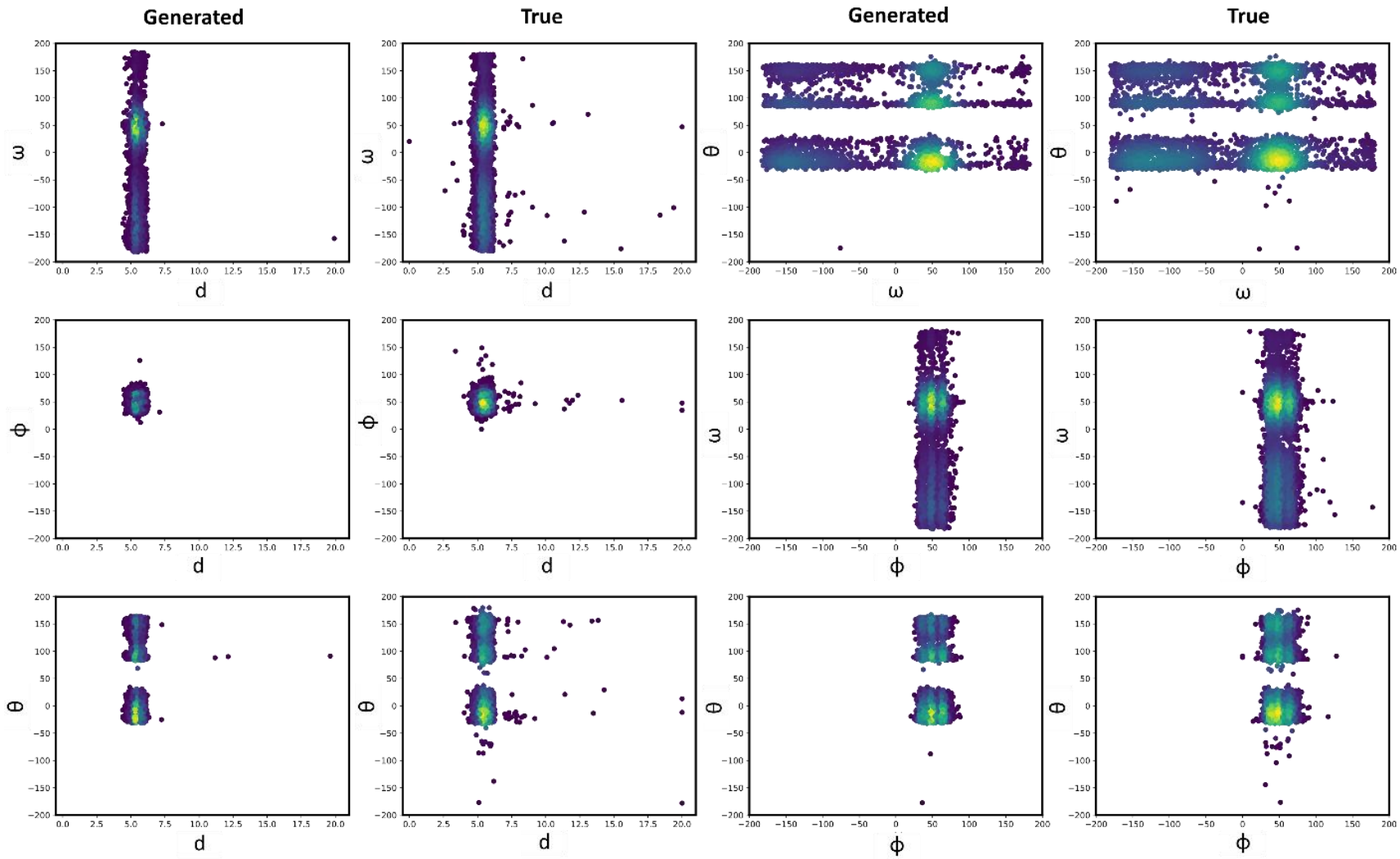
nature computational science

Article

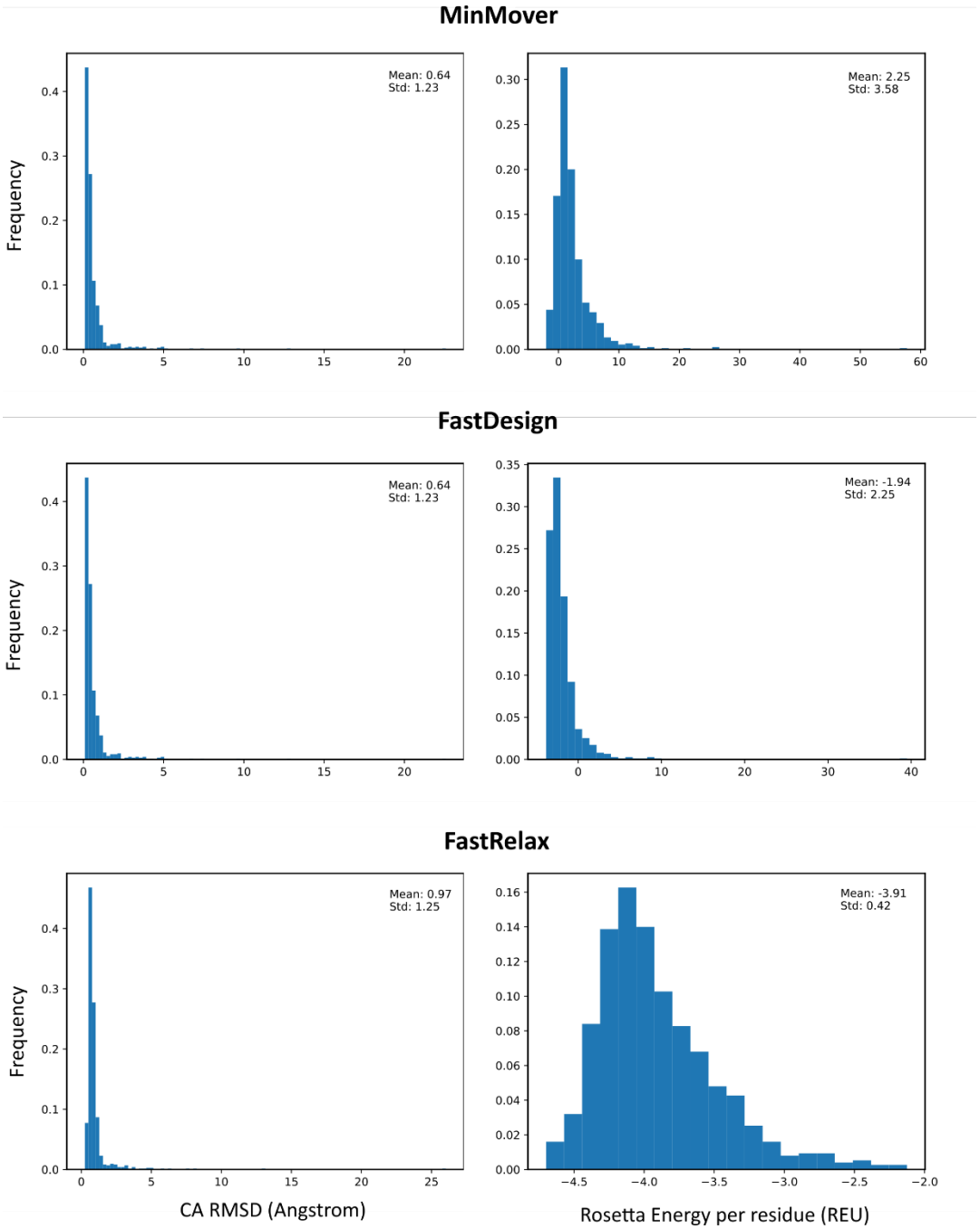
https://doi.org/10.1038/s43588-023-00440-3

Score-based generative modeling for de novo protein design

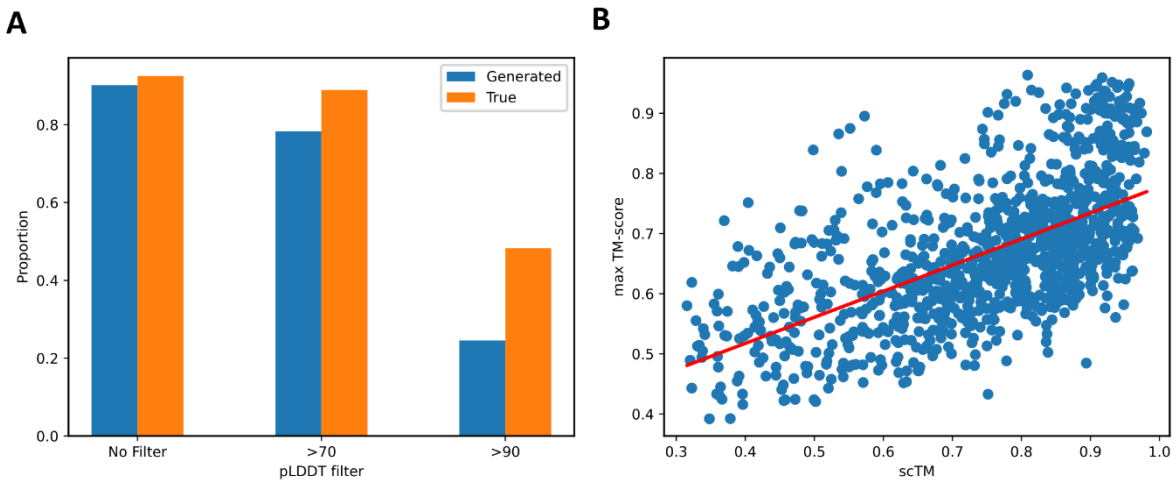
In the format provided by the authors and unedited



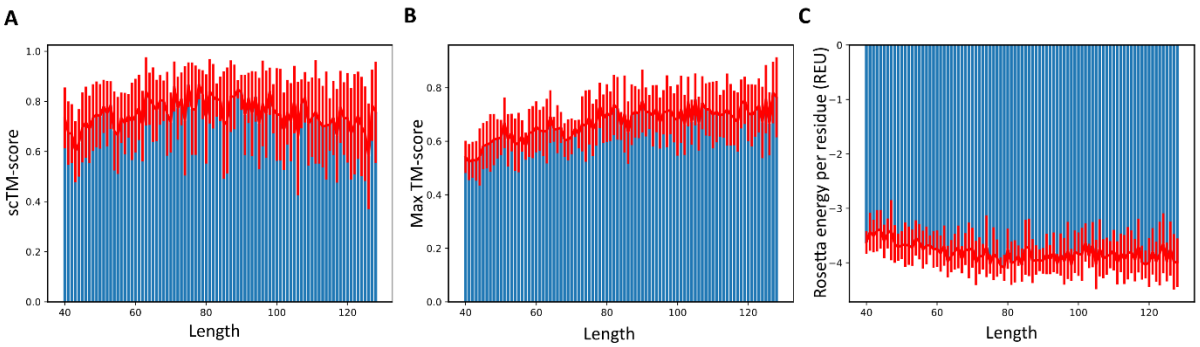
Supplementary Figure 1. 2D matrix distributions. We analyzed joint distributions of 6D coordinates in adjacent residues between true and generated samples to further assess generation quality of ProteinSGM. We see a clear concordance between generated and true samples, suggesting that ProteinSGM has learned native biophysical constraints.



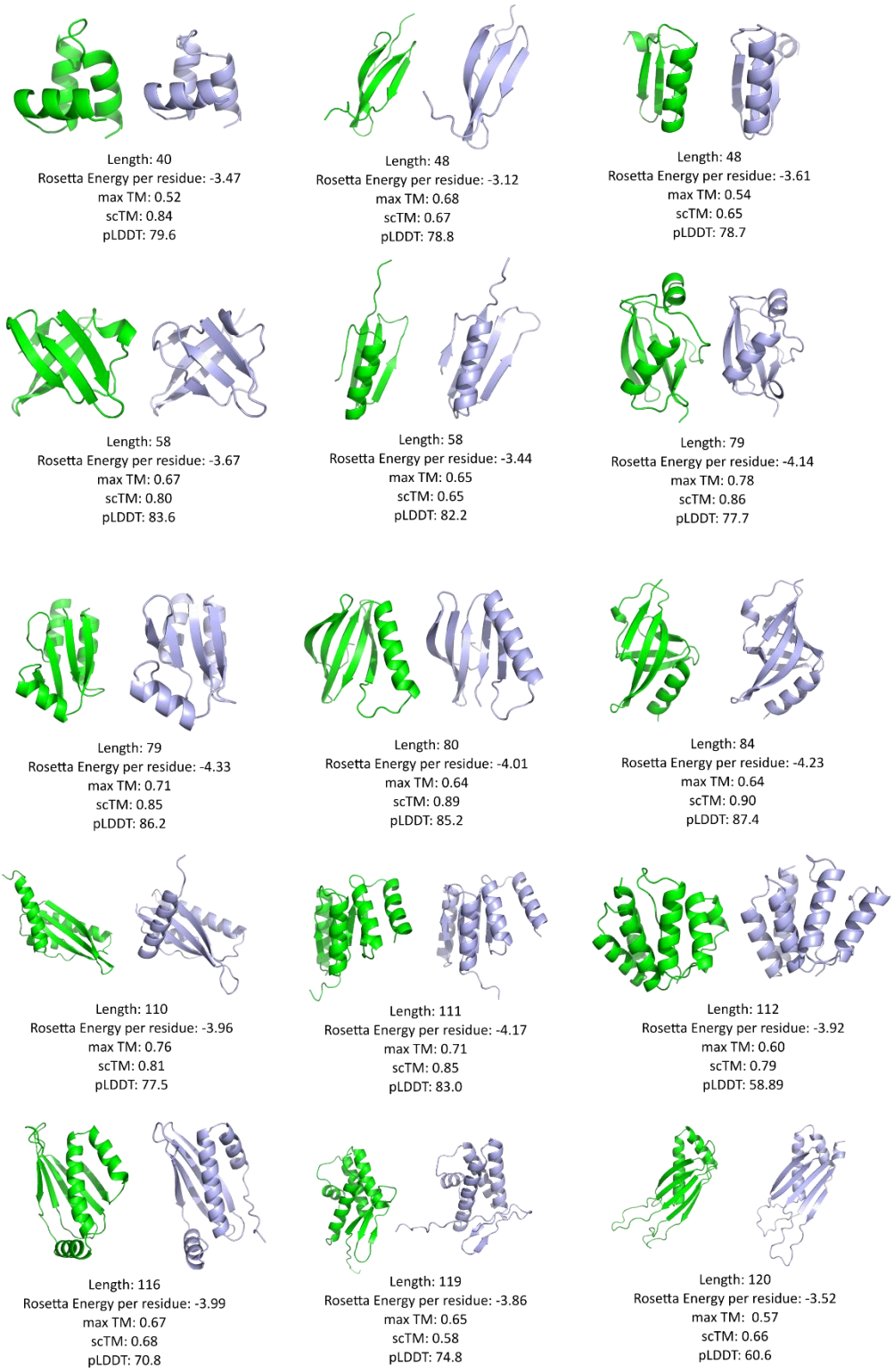
Supplementary Figure 2. Rosetta energies and Ca RMSDs with Rosetta pipeline. (left) Rosetta energy before FastRelax remains relatively high for generated structures, but full-atom relaxation with constraints allows effective energy minimization to bring energies closer to the true distribution. (right) Minimized structures from 6D coordinates were compared with their native structures to assess fidelity of the Rosetta protocol. We observe that across all steps, mean Ca RMSDs are lower than 1A, suggesting the protocol is suited for reproducible generation of structures from 6D coordinates.



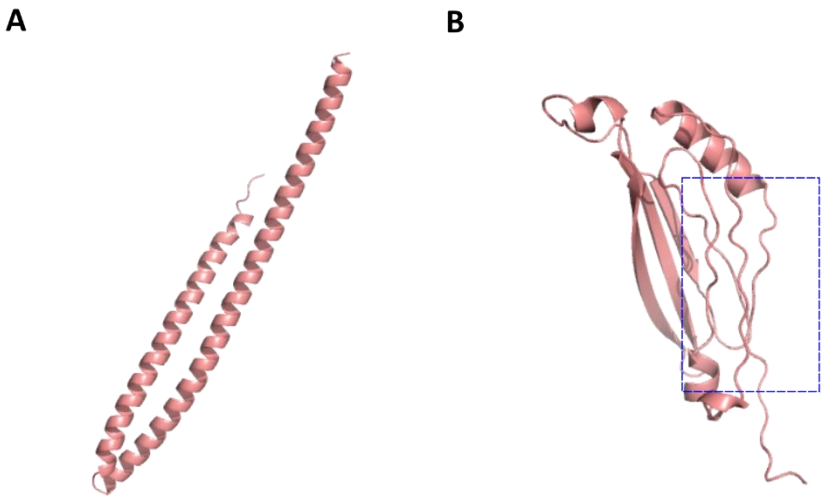
Supplementary Figure 3. Additional scTM analysis. (A) Since predicted structures with low pLDDT (low confidence) are generally unreliable, we filter OmegaFold structures at different cutoffs and analyze changes to the scTM > 0.5 proportion. With no filtering, a > 70 pLDDT filter, and a > 90 pLDDT filter, we obtain 90.2%, 78.3%, and 24.5% scTM > 0.5 in the generated samples and 92.5%, 88.9%, and 48.3% in the true samples, respectively. Though there is a notable difference between generated and true samples with more stringent pLDDT filters, we observe that at a > 70 pLDDT filter corresponding to confident predictions, more than 78% of the generated samples are designable with a protein sequence. (B) We analyze the relationship between scTM (designability) and max-TM (similarity to training data), and observe that there is a positive correlation between scTM and max TM-score. This is expected since structures with closer similarity to PDB structures are, by definition, realizable with a protein sequence, and therefore should have an scTM > 0.5.



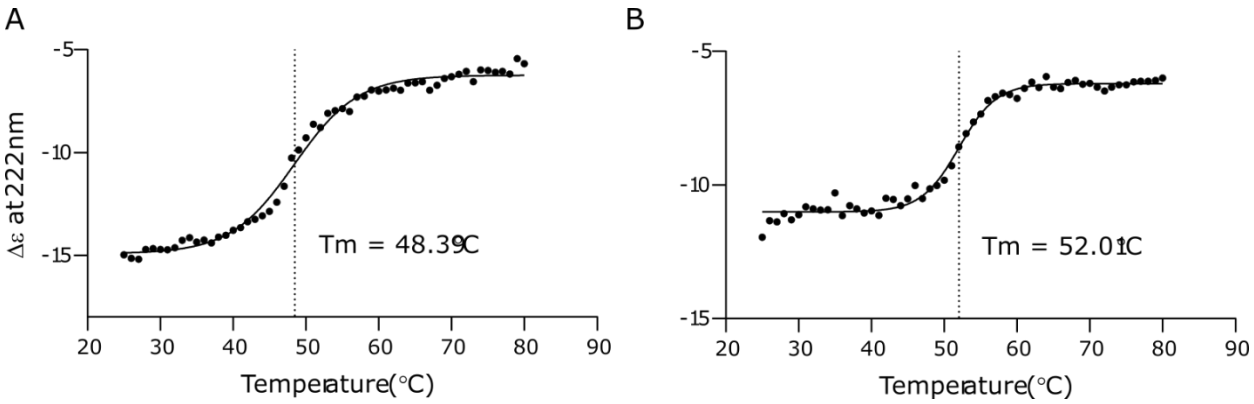
Supplementary Figure 4. Per-length analysis of (A) scTM, (B) max TM-score, and (C) Rosetta energy. We observe that there is consistency across all metrics between structures of different lengths. However, there is a slightly lower max TM-score and higher Rosetta energy for shorter structures, which may be explained by the lower number of short domains in the training set and the lower number of stabilizing interactions in shorter proteins, respectively.



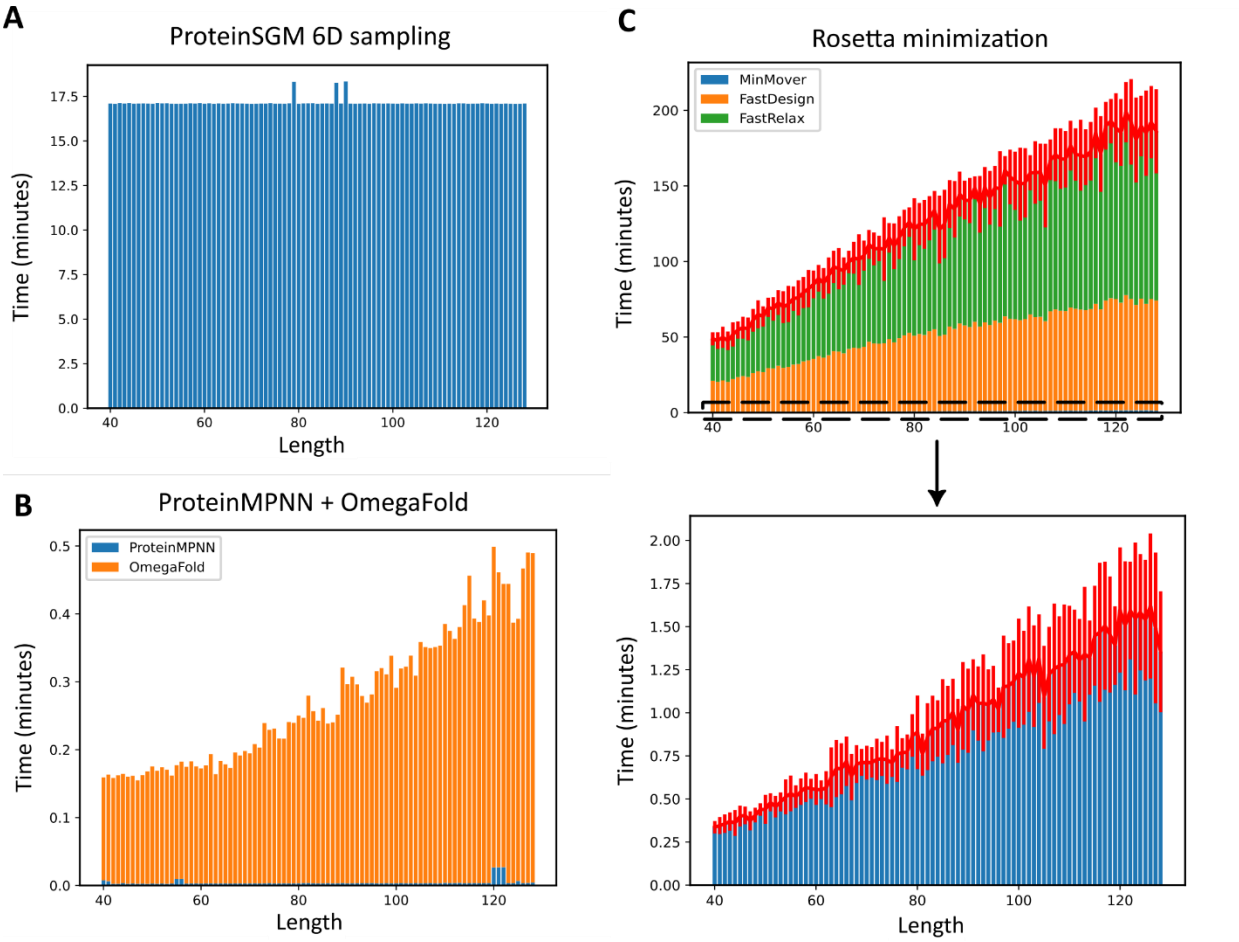
Supplementary Figure 5. Examples of generated structures. Rosetta-generated backbones are in green, and the best OmegaFold structure is shown in blue.



Supplementary Figure 6. Failure modes of ProteinSGM. (A) We observe that in a few cases, ProteinSGM generates long helices that may not exist in nature. (B) Since beta sheets are more difficult to model, some generated backbones display loop-like structures that are not visualized as beta sheets due to incorrect placement of hydrogen bonds.



Supplementary Figure 7. Melting curves of two α proteins (A and B) monitoring the variation of the CD signal at 222 nm are shown with the temperature (25-80 °C). The number in the graph indicates the melting temperature (Tm) values of each protein.



Supplementary Figure 8. Runtime analysis by length. (A) Sampling from ProteinSGM is uniform across all lengths since the dimensionality of the generated matrices is fixed. (B) With the ProteinMPNN + OmegaFold pipeline, OmegaFold dominates most of the runtime, yet is relatively fast at < 30 seconds per structure. (C) With the Rosetta pipeline, initial backbone generation with MinMover is relatively fast at < 2 minutes per structure, but FastDesign and FastRelax can take up to 3 hours per structure. ProteinSGM, ProteinMPNN, and OmegaFold runtimes were measured using a single NVIDIA V100, while Rosetta was run with 2 cores and 8GB of RAM. Standard deviations for Rosetta iterations are shown in red. ProteinSGM was sampled with batch size 16, and ProteinMPNN was run in batch mode with 12 input structures and 8 sampled sequence per structure, so runtime may vary with different batch sizes.

Supplementary Table 1. Mean absolute error of input 6D coordinates and 6D coordinates recovered after MinMover and FastRelax.

	6D coordinate	MinMover	FastRelax
Generated	d (Cβ - Cβ)	0.373Å	0.469Å
	ω	14.6°	23.2°
	θ	9.38°	15.3°
	φ	4.43°	7.29°
True	d (Cβ - Cβ)	0.147Å	0.349Å
	ω	4.97°	18.83°
	θ	4.04°	12.33°
	φ	2.05°	6.07°

Supplementary Table 2. Estimated secondary structure content (%) of unconditional samples by BeStSel [29]. Structure ID corresponds to the subfigure denoted in **Figure 4**.

		Structure ID				
Secondary Structure		A	B	C	D	E
a-helix	Helix1 (regular)	77.4	78.3	43.7	73.5	49.1
	Helix2 (distorted)	21.8	21.1	23.2	17.1	21.6
Anti-parallel beta sheet	Anti1 (left-twisted)	0	0	0	0	0
	Anti2 (relaxed)	0	0	0.8	0	0
	Anti3 (right-twisted)	0.9	0.6	28.7	2.5	0
Parallel beta sheet		0	0	0	0	0
Turn		0	0	0	1.5	0
Others*		0	0	3.5	5.3	29.3

* 3,10-helix, π-helix, β -bridge, bend, loop/irregular and invisible regions