

**Table S8:** Results on the 90 proteins that were non-redundant to the training set of DeepPotential. For the mean TM-scores, the  $p$ -values were calculated using paired, two-sided Student’s t-tests, while the  $p$ -values for the median TM-scores were calculated using two-sided, non-parametric Wilcoxon signed rank tests.

Method	Mean TM-score ( $p$ -value)	Median TM-score ( $p$ -value)	Correct Folds*	$TM_{\text{DeepFold}} >$ $TM_{\text{Method}}^\ddagger$
I-TASSER	0.384 (8.5E-31)	0.346 (4.7E-16)	28.9%	93.3%
C-I-TASSER	0.580 (3.0E-20)	0.605 (7.7E-15)	66.7%	93.3%
DMPfold	0.643 (6.4E-16)	0.700 (2.6E-15)	76.7%	90.0%
trRosetta	0.688 (7.5E-09)	0.719 (2.6E-09)	87.8%	83.3%
DeepFold	<b>0.730 (*)</b>	<b>0.778 (*)</b>	<b>90.0%</b>	-

\* This column represents the percent of proteins with TM-scores  $\geq 0.5$ .

‡ This column indicates the percent of test proteins for which DeepFold generated a model with a higher TM-score than the control method.