Legend for Supplementary Data files

Supplementary Data 1. SCOPe fold types poorly predicted by SPIRED. The table lists SCOPe fold types that are predicted by SPIRED with the average TM-score < 0.5. The "Fold ID" and "Fold Name" columns denote the ID and name of SCOPe fold types. Numbers in the columns show the average TM-score of all domains of individual SCOPe fold types, as predicted by SPIRED, OmegaFold and ESMFold, respectively.

Supplementary Data 2. SCOPe fold types poorly predicted by OmegaFold. The table lists SCOPe fold types that are predicted by OmegaFold with the average TM-score < 0.5. The "Fold ID" and "Fold Name" columns denote the ID and name of SCOPe fold types. Numbers in the columns show the average TM-score of all domains of individual SCOPe fold types, as predicted by SPIRED, OmegaFold and ESMFold, respectively.

Supplementary Data 3. SCOPe fold types poorly predicted by ESMFold. The table lists SCOPe fold types that are predicted by ESMFold with the average TM-score < 0.5. The "Fold ID" and "Fold Name" columns denote the ID and name of SCOPe fold type. Numbers in the columns show the average TM-score of all domains of individual SCOPe fold types, as predicted by SPIRED, OmegaFold and ESMFold, respectively.

Supplementary Data 4. ProteinGym assays chosen for performance evaluation in the zero-setting mode. The assays are selected by: 1) protein sequence identity < 50% with the original training set of the SPIRED-Fitness (Stage 2) model; 2) protein length < 900 residues. The total of 50 assays are further clustered into 10 groups (see the column of "cluster index") based on protein sequence similarity, with the inter-group protein sequence identity < 50%. According to our training protocol, each cluster of assays correspond to a specific SPIRED-Fitness zero-shot model. Hence, in order to replicate the zero-shot prediction results of SPIRED-Fitness on an assay, the zero-shot model with the same ID as the cluster ID should be used.

Supplementary Data 5. ProteinGym assays chosen for performance evaluation of multiple mutational effects in the supervised mode. We slightly modified the model architecture of SPIRED-Fitness (Supplementary Figure S10) to allow the prediction of multiple mutational effects in the supervised setting of ProteinGym. However, this version of SPIRED-Fitness has to be independently trained for each individual assay. In order to avoid the prohibitive computational costs, 61 DMS assays in the table are chosen for evaluating the prediction of multiple mutational effects by SPIRED-Fitness.

Supplementary Data 6. CAMEO proteins adopted for evaluating the protein structure prediction algorithms. The total of 680 proteins are collected from the CAMEO competition between Aug. 2022 and Aug. 2023. The "Sample" column denotes the sample names of proteins from CAMEO test set. Taking 7YFV_B as an example, 7YFV is the PDB ID, and B means the chain B of this protein. The "L" column denotes the protein length.

Supplementary Data 7. CASP15 protein domains adopted for evaluating the protein structure prediction algorithms. The total of 45 protein domains are collected from the CASP15 official website, with both sequences and ground-truth structures released. The "Sample" column denotes the target ID of proteins from CASP15 test set. The "L" column denotes the protein length.

Supplementary Data 8. Protein assays in the MaveDB/DeepSequence datasets for fitness prediction. These assays were used for the training, validation and preliminary testing of our SPIRED-Fitness model. The superscript in the dataset name represents the source of the data, where [1] denotes data from DeepSequence and [2] denotes data from MaveDB.