Training and validation accuracy of a ProteinSolver network optimized to reconstruct amino acid sequences

A — Accuracy on missing residues vs. Number of training data points (1e8); training (blue), validation (orange).

B — Number of sequences vs. Sequence identity to reference; oneshot (accuracy: 27.29%), incremental (accuracy: 26.80%).

C — Number of sequences (normalized) vs. Sequence identity to reference; 0% residues present, 50% residues present, 80% residues present.
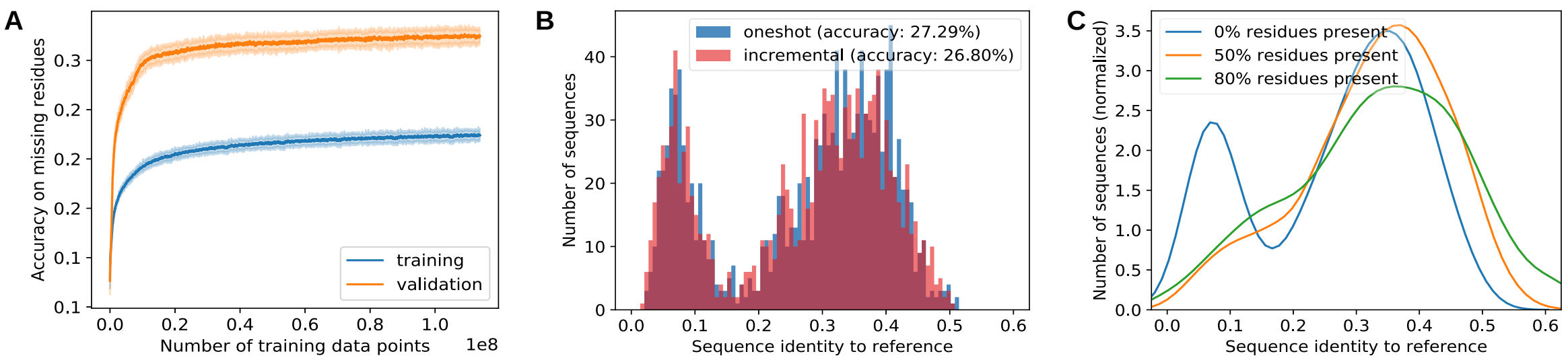
Independent validation of the trained network using stability measurements of mutated or designed proteins

Legend: ProteinSolver, Ingraham et al., Rosetta (score), Rosetta (ddg_monomer), Rosetta (cartesian_ddg)

D — Spearman's ρ; Protherm.

E — Spearman's ρ; Known (Pin1, hYAP65, villin), HHH (rd2-0134, rd3-0138), HEEH (rd2-0779, rd3-0223, rd3-0726, rd3-0872), EEHEE (rd3-0037, rd3-1498, rd3-1702, rd3-1716).

F — Spearman's ρ; HHH (Round 1–4), HEEH (Round 1–4), EHEE (Round 1–4), EEHEE (Round 1–4).
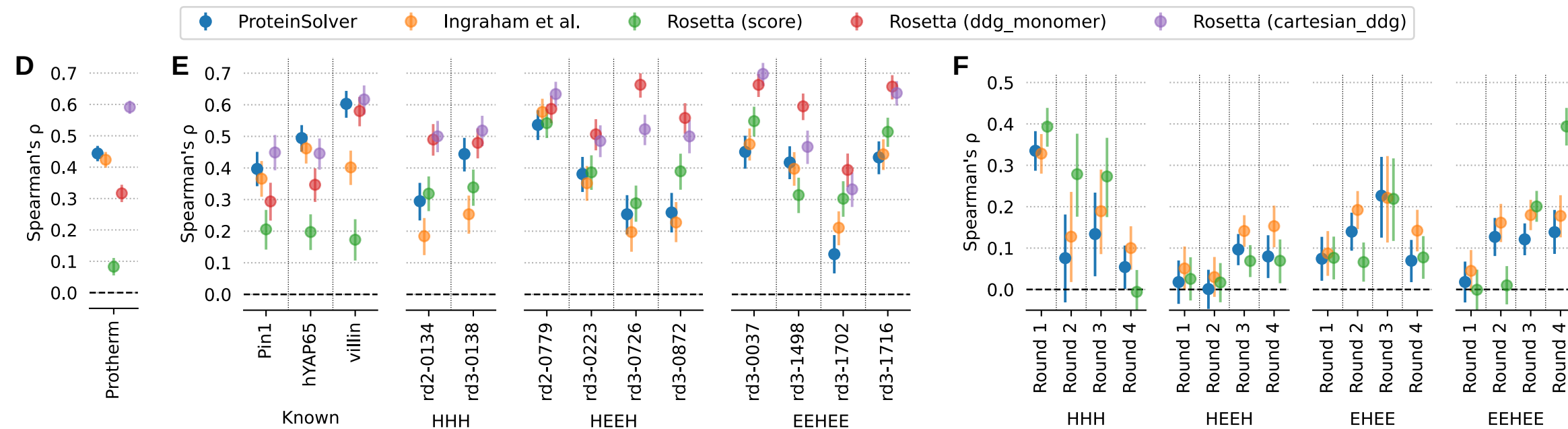
Computational and experimental validation of sequences generated to match the architecture of a *serum albumin*

G — Amino acid position vs. Amino acid position; 1 / distance (Å⁻¹).

H — Network score vs. Sequence identity; all residues (R² = 0.189), exposed residues (R² = 0.027), intermediate residues (R² = 0.047), buried residues (R² = 0.142).

I — bits; KFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVEND

J — probability; CCC HHHHHHHHHHHHHHH CC CCCCC HHHHHHHHHHHHHHHHH CCCCC HHHHHHHHHHHHHHHH CCCCCCC HHHHHHH CCCCC HHHHHHHHH CCCC

K — Modeller (molpdf), Modeller (dope_norm), Modeller (dope_hr), Modeller (ga341), Rosetta (score).

L — protein structure.

M — rmsf (Angstroms) vs. amino-acid position; WildType, Design 1, Design 2.

N — Molar ellipticity (deg cm² dmol⁻¹) vs. Wavelength (nm).