DDGEmb

## The S2450 dataset

This dataset has been obtained by mapping variations included in the well-known S2648 dataset (Dehouck et al., 2009) on full-length UniProt sequences. Moreover, proteins sharing more 30% sequence identity on 40% alignment coverage with any protein in the S669 test set (Pancotti et al., 2022) were also excluded.

The final dataset contains 2450 single-point variations endowed wuth experimental ΔΔG on 115 protein sequences.

Download TSV

Download FASTA

## The S669 dataset

This dataset has been obtained by mapping variations included in the S669 test set (Pancotti et al., 2022) on full-length UniProt sequences.

The final dataset contains 669 single-point variations endowed with experimental ΔΔG on 87 unique protein sequences.

Download TSV

Download FASTA

## The ptMUL-NR dataset

This dataset has been obtained by mapping multi-site variations included in the PTmul test set (Montanucci et al., 2019) on full-length UniProt sequences. Moreover, the dataset has been homology-reduced (30% sequence identity on 40% alignment coverage) with respect to the S2450 training set.

The final dataset contains 82 multi-point variations endowed with experimental ΔΔG on

Download TSV

Download FASTA

13 protein sequences.

## References


- Dehouck, Y. et al. (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, **12**, 151.
- Pancotti, C. et al. (2022) Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Brief. Bioinform.*, **23**, bbab555.
- Montanucci, L. et al. (2019) DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics*, **20**, 335.