Supplementary information:

Semi-supervised prediction of protein fitness for data-driven protein engineering

Alicia Olivares-Gil, José A. Barbero-Aparicio, Juan J. Rodríguez,
José F. Díez-Pastor, César García-Osorio, Mehdi D. Davari

Contents

1 Supplementary figures **4**

2 Supplementary tables **14**

3 Supplementary algorithms **15**

List of Figures

S1 **General taxonomy of semi-supervised learning classifiers** described in [18]. 4

S2 **One iteration of the evaluation process when only few labelled instances available are used for training.** The whole process, including the 5-fold cross validation for parameters tuning, is repeated 10 times for each of the possible values of N. Grid search via 5-fold cross validation is performed only for wrapper methods. The rest use default parameters for the model. 4

S3 **One iteration of the evaluation process when all the labelled instances available are used for training.** 10-fold cross-validation is used. Best parameters are calculated for each dataset as the mode of the best parameters found using only few labelled instances. 5

S4 **Absolute Spearman's ρ and absolute Weighted Spearman's ρ_w of the methods in strategies 0 and 1, using all available labelled sequences on each dataset for training and validation.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red. 6

S5 **Absolute Spearman's ρ and absolute Weighted Spearman's ρ_w of the methods in strategy 2, using all available labelled sequences on each dataset for training and validation.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red. 7

S6 **Absolute Spearman's ρ and absolute Weighted Spearman's ρ_w of the methods in strategies 3 and 4, using all available labelled sequences on each dataset for training and validation.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red. 8

S7 **Absolute Spearman's ρ of the methods in strategies 0 and 1, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red. 9

S8 **Absolute Weighted Spearman's ρ_w of the methods in strategies 0 and 1, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red. 10

S9 **Absolute Spearman's ρ and absolute Weighted Spearman's ρ_w of the methods in strategy 2, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red. 11

S10 **Absolute Spearman's ρ of the methods in strategies 3 and 4, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red. 12

S11 **Absolute Weighted Spearman's ρ_w of the methods in strategies 3 and 4, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red. 13

List of Tables

S1 **Datasets included in experimentation and their characteristics.** 14
S2 **Datasets included in experimentation and the protein function for which the fitness is given.** 15
S3 **Time and peak memory performance of the training process of the best methods found for each strategy. Calculated using 100, 1000 and 10000 labeled training sequences.** 15

List of Algorithms

S1 Original Tri-Training learning function. 16
S2 Tri-Training Regressor learning function.
Differences from Algorithm S1 are highlighted in red. 17

1 Supplementary figures

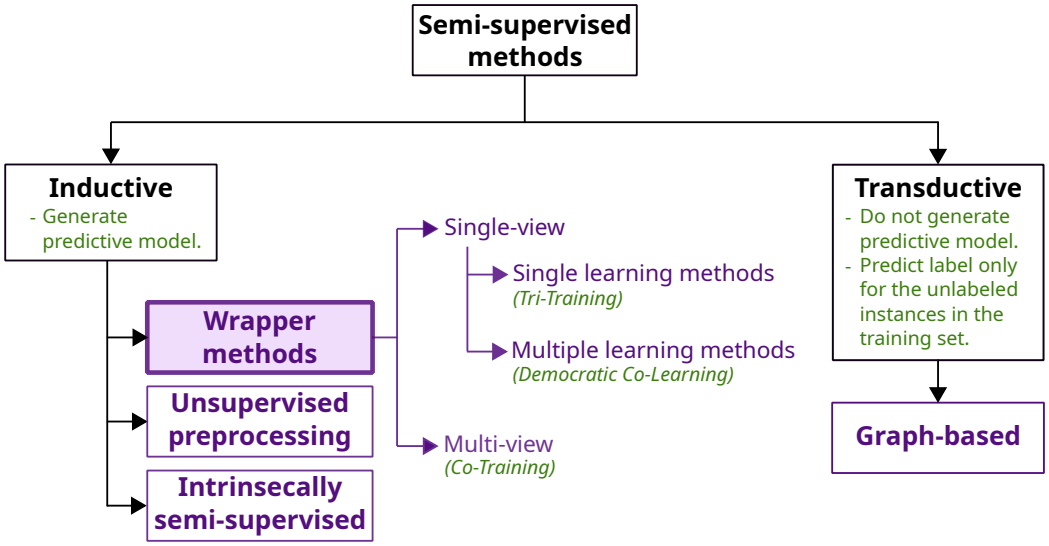

Figure S1: **General taxonomy of semi-supervised learning classifiers** described in [18].

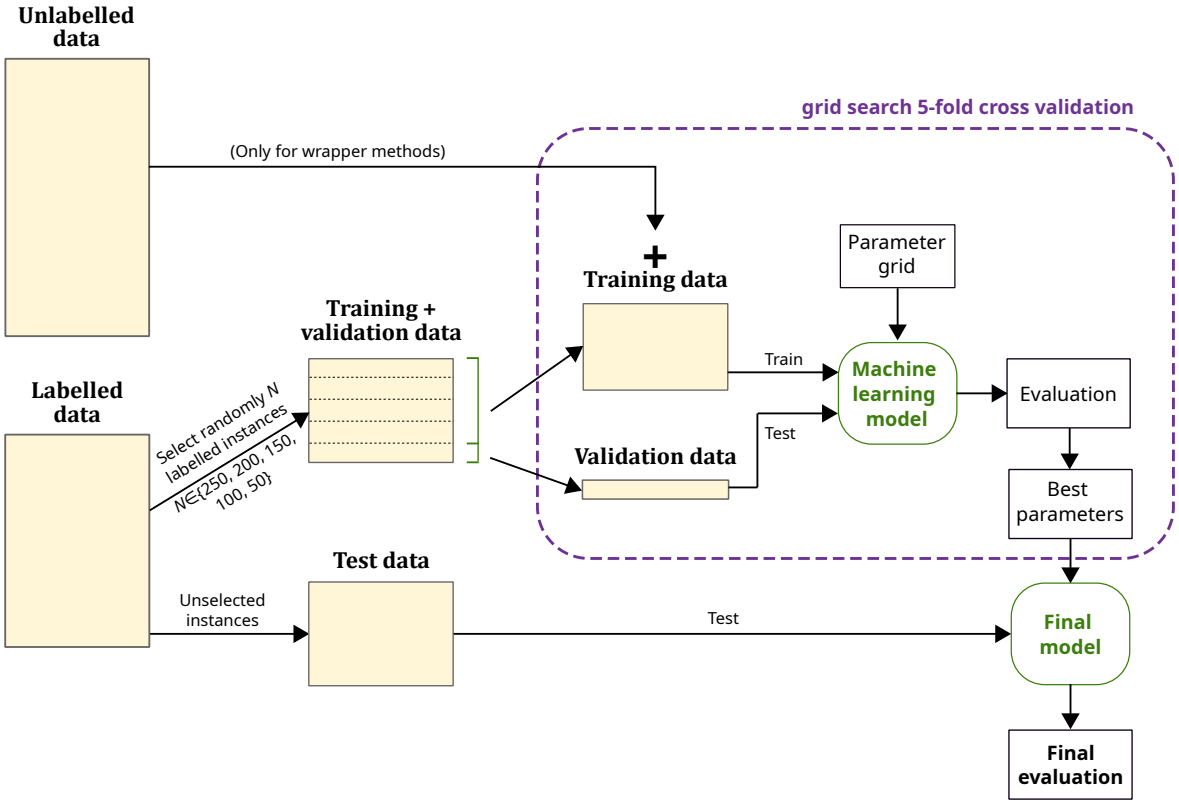

Figure S2: **One iteration of the evaluation process when only few labelled instances available are used for training.** The whole process, including the 5-fold cross validation for parameters tuning, is repeated 10 times for each of the possible values of N. Grid search via 5-fold cross validation is performed only for wrapper methods. The rest use default parameters for the model.

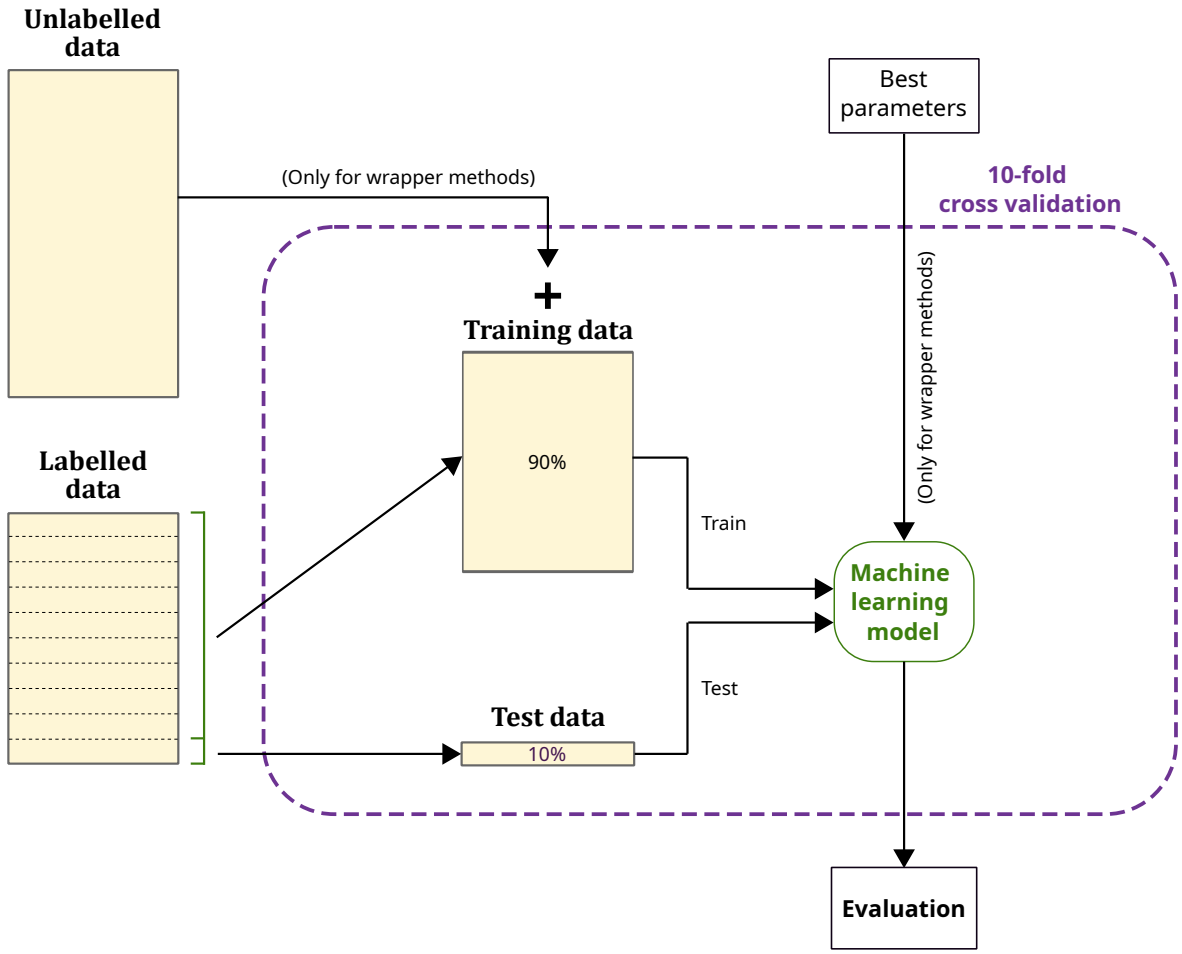

Figure S3: **One iteration of the evaluation process when all the labelled instances available are used for training.** 10-fold cross-validation is used. Best parameters are calculated for each dataset as the mode of the best parameters found using only few labelled instances.

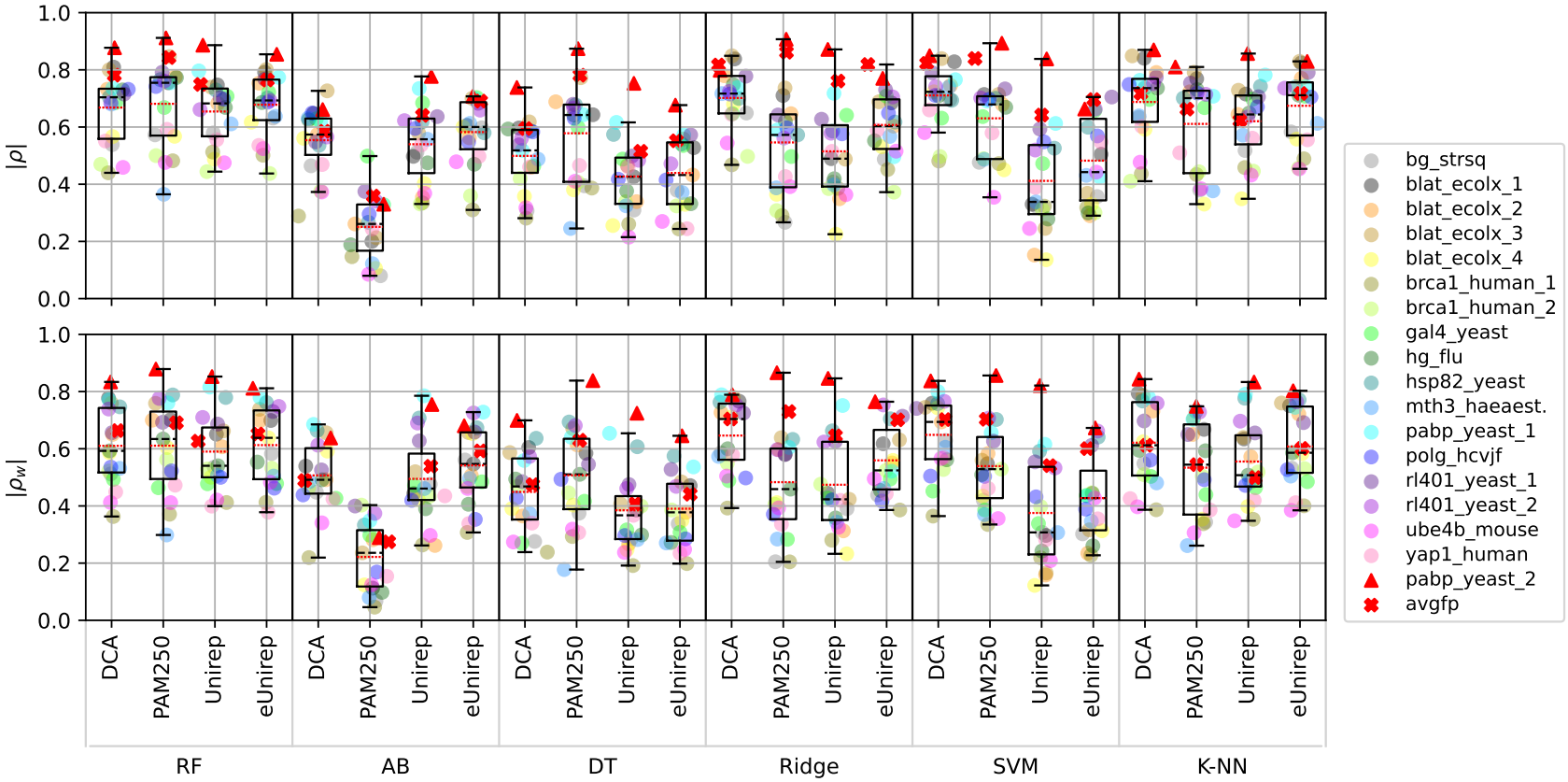

Figure S4: **Absolute Spearman's ρ and absolute Weighted Spearman's ρ_w of the methods in strategies 0 and 1, using all available labelled sequences on each dataset for training and validation.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red.

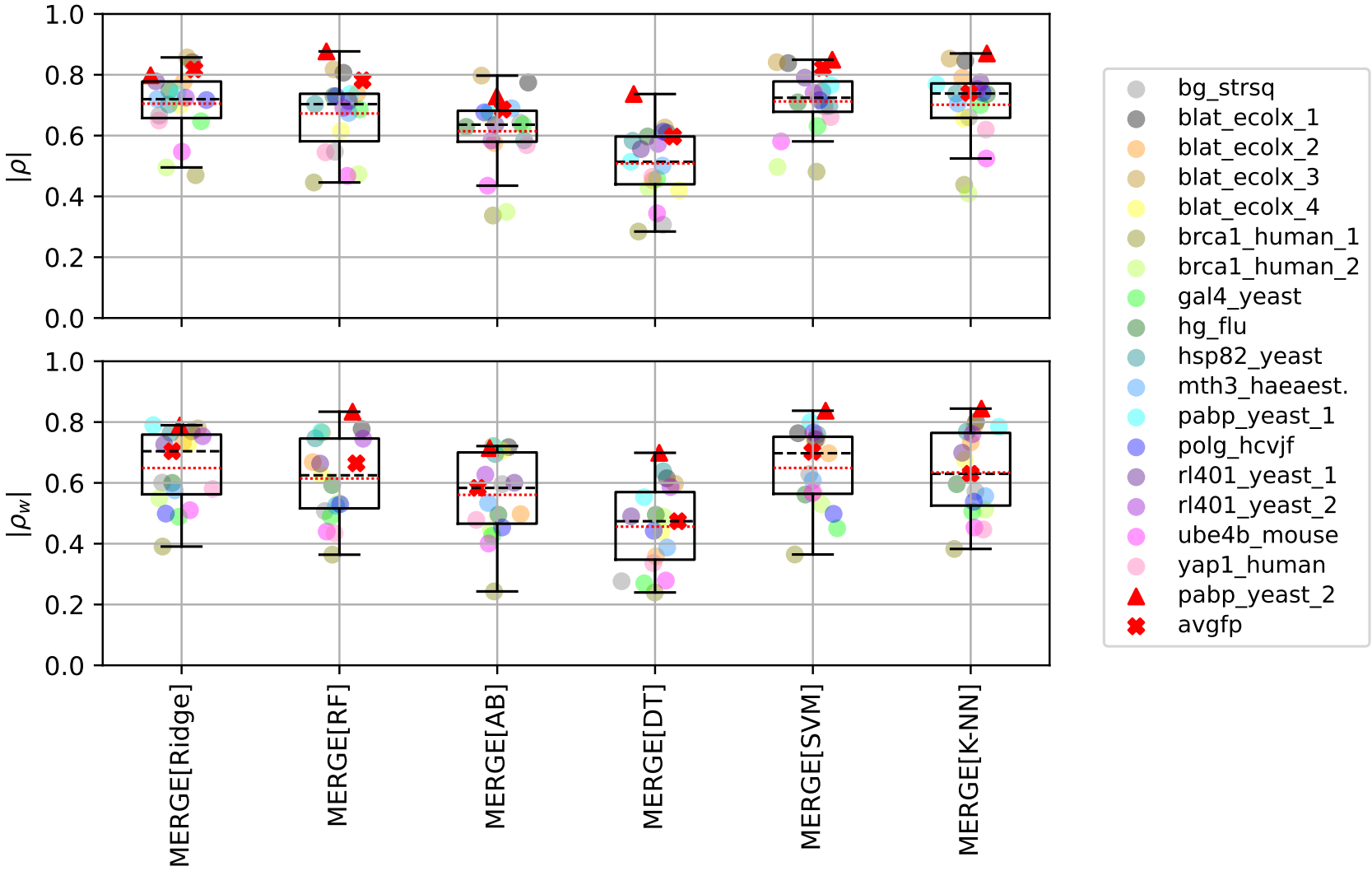

Figure S5: **Absolute Spearman's ρ and absolute Weighted Spearman's ρ_w of the methods in strategy 2, using all available labelled sequences on each dataset for training and validation.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red.

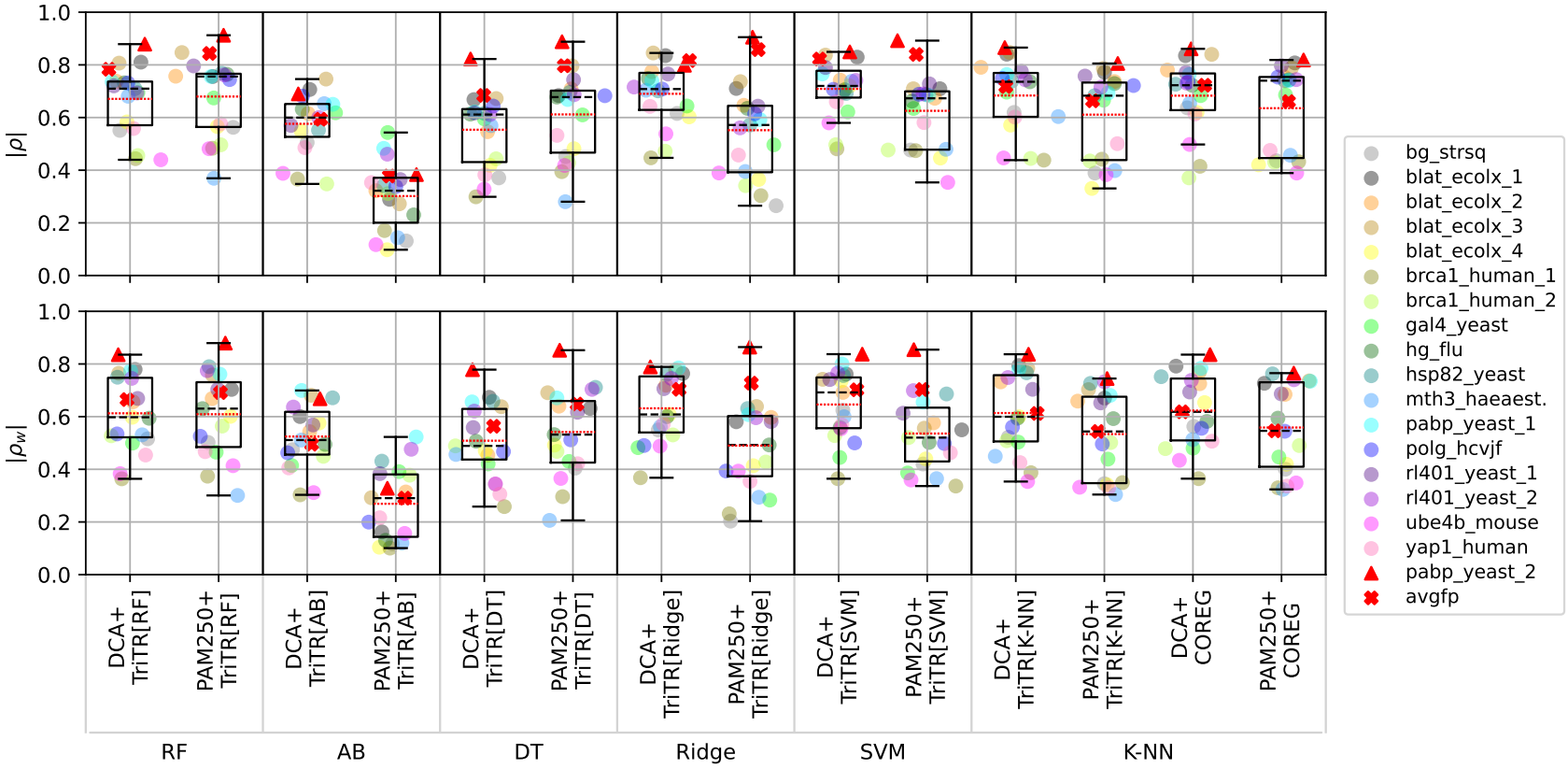

Figure S6: **Absolute Spearman's ρ and absolute Weighted Spearman's ρ_w of the methods in strategies 3 and 4, using all available labelled sequences on each dataset for training and validation.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red.

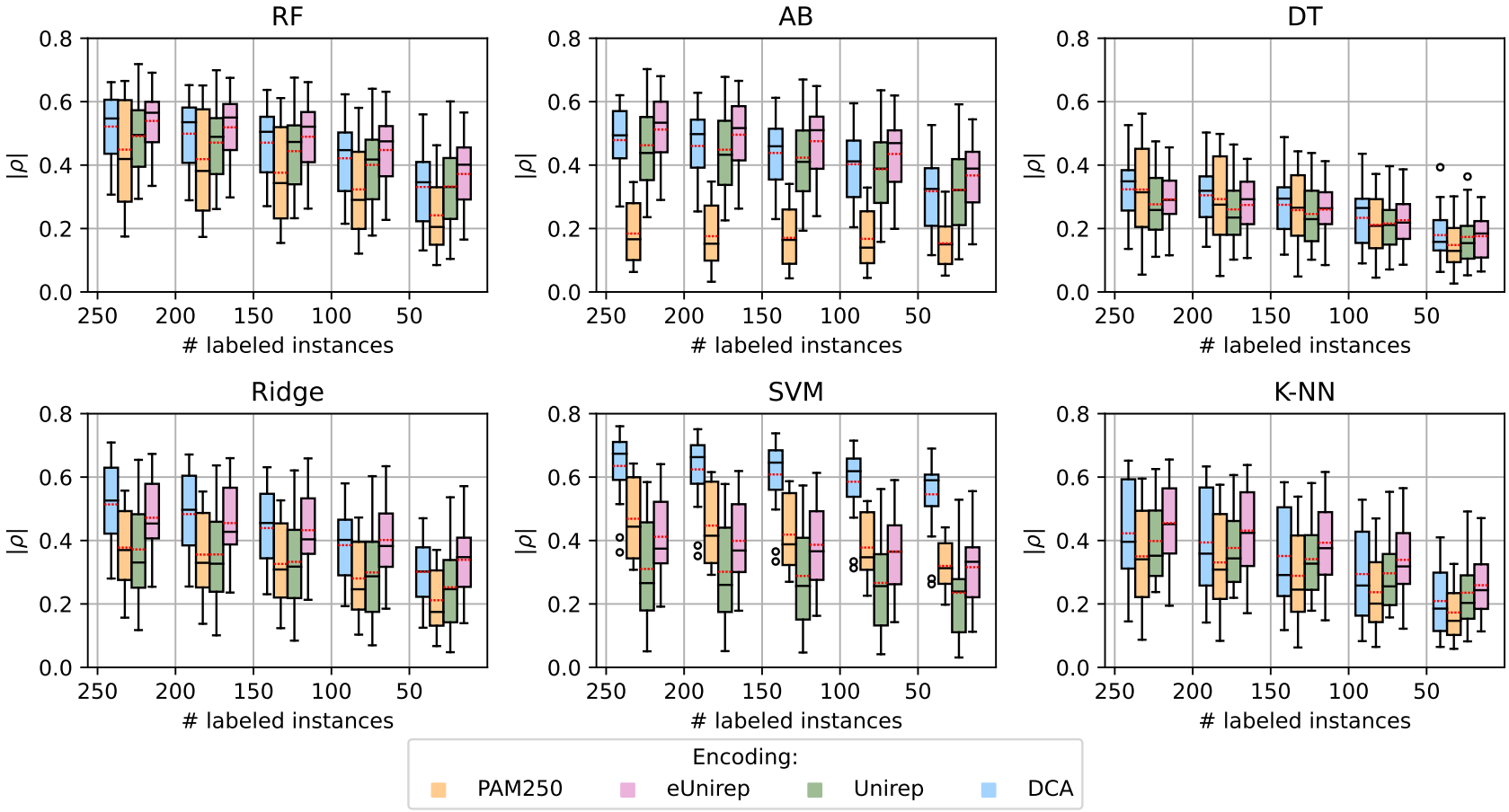

Figure S7: **Absolute Spearman's ρ of the methods in strategies 0 and 1, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red.

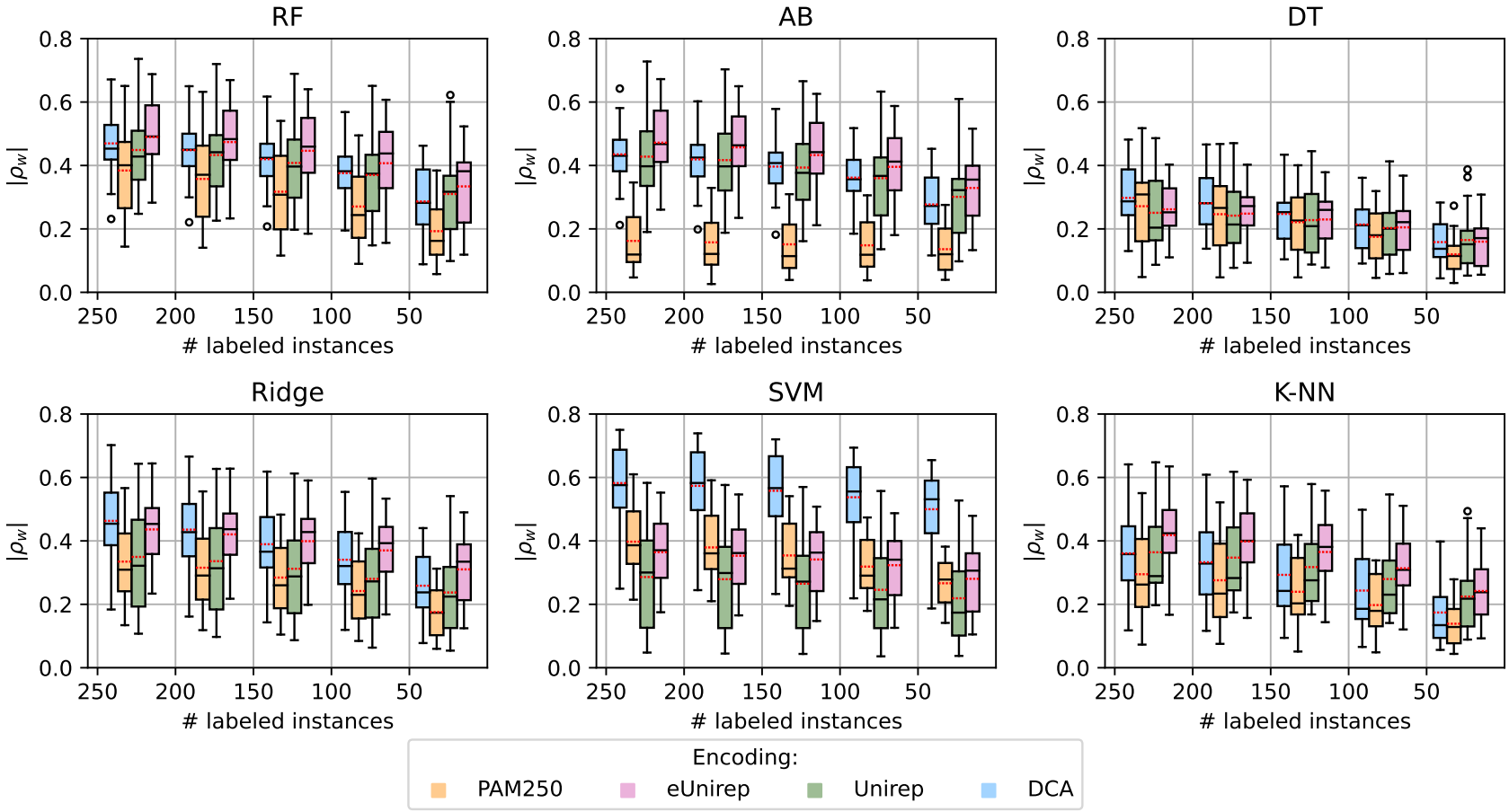

Figure S8: **Absolute Weighted Spearman's ρ_w of the methods in strategies 0 and 1, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red.

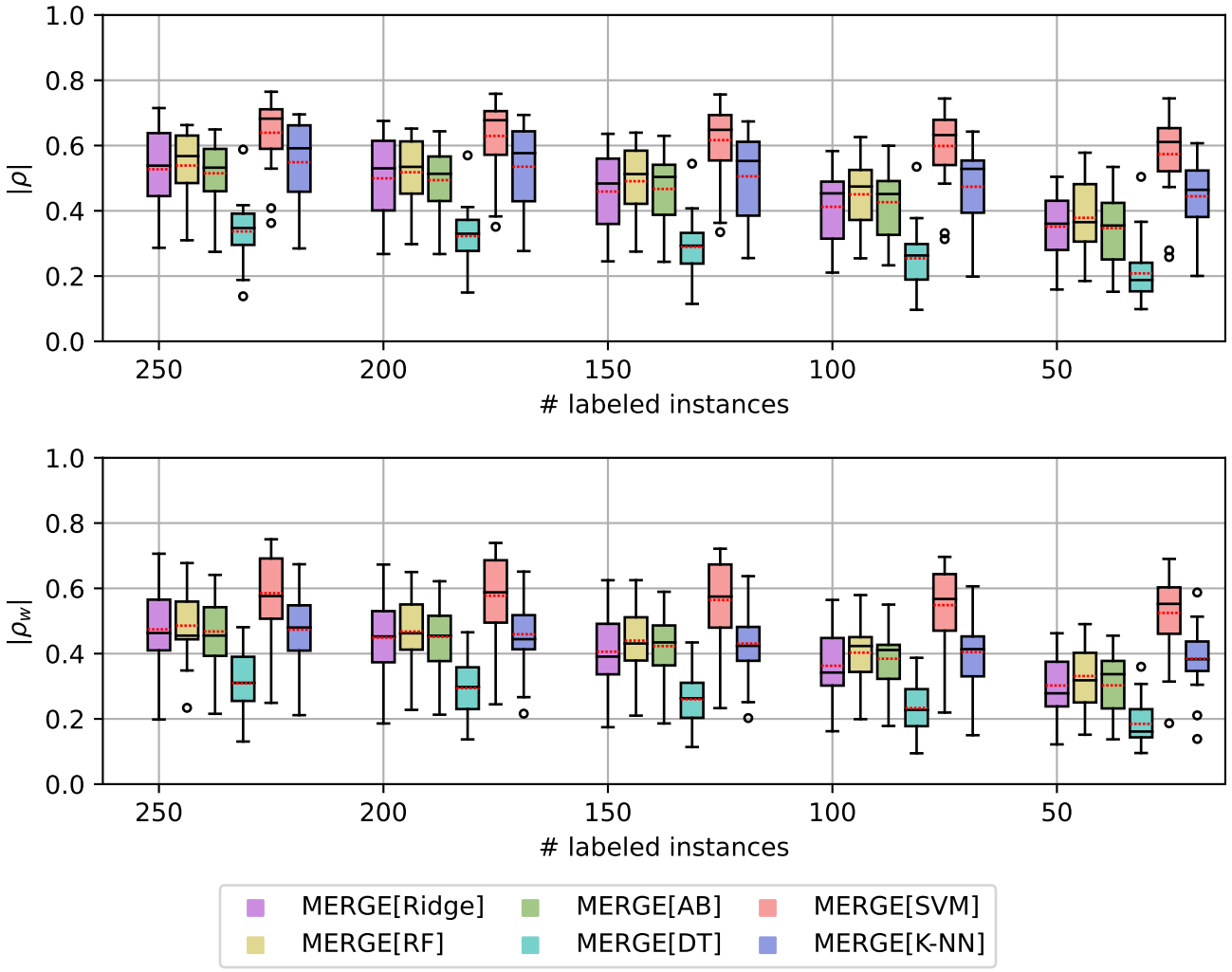

Figure S9: **Absolute Spearman's ρ and absolute Weighted Spearman's ρ_w of the methods in strategy 2, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red.

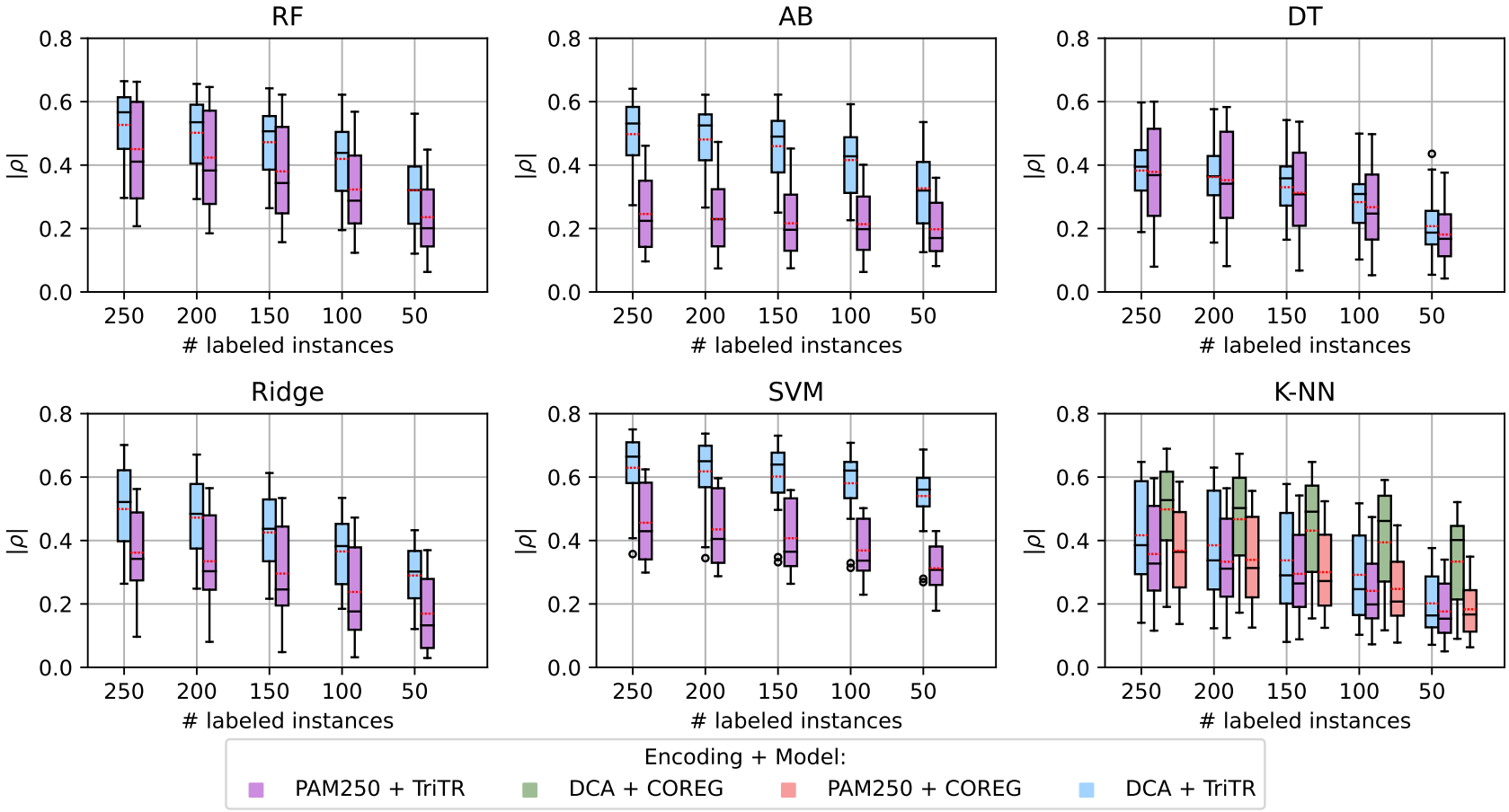

Figure S10: **Absolute Spearman's ρ of the methods in strategies 3 and 4, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red.

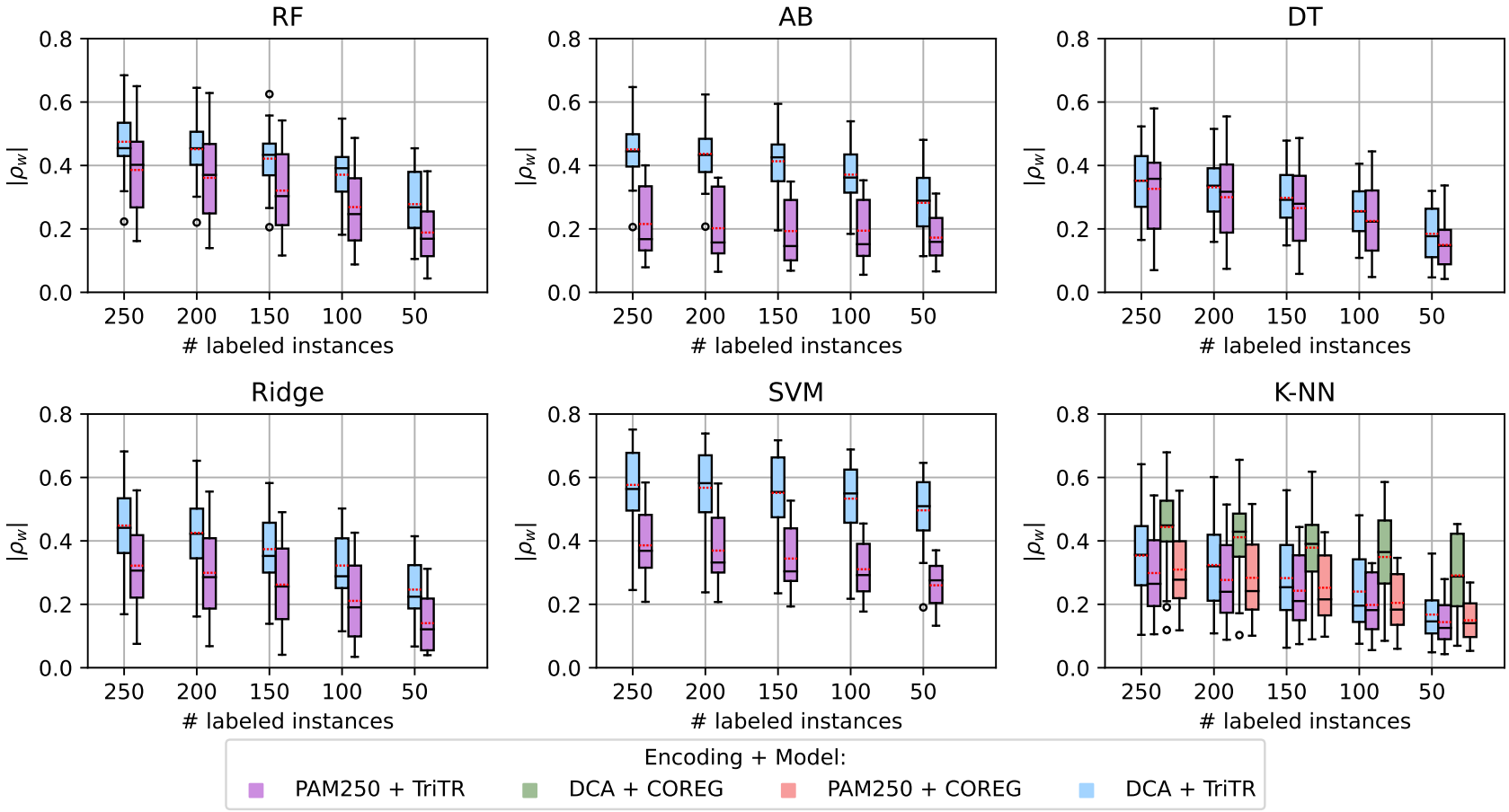

Figure S11: **Absolute Weighted Spearman's ρ_w of the methods in strategies 3 and 4, as a function of the number of labelled sequences used for training.** Coloured dots represent the different datasets. Dataset abbreviations in the legend are defined in Table S1. Lines indicating median values in black and average values in red.

2 Supplementary tables

Table S1: **Datasets included in experimentation and their characteristics**.

Dataset	Abbreviation	Length	L	U	%L	Subst.
YAP1_HUMAN_Fields2021 [1]	yap1_human	34	313	248 330	0.1	single
UBE4B_MOUSE_Klevit2013 [15]	ube4b_mouse	102	518	97 159	0.5	single
GAL4_YEAST_Shendure2015 [6]	gal4_yeast	64	803	171 882	0.5	single
BLAT_ECOLX_Tenaillon2013 [5]	blat_ecolx_4	286	975	30 356	3.1	single
PABP_YEAST_Fields2013 [7]	pabp_yeast_1	75	1 142	948 246	0.1	single
RL401_YEAST_Bolon2013 [13]	rl401_yeast_1	75	1 154	96 590	1.2	single
BRCA1_HUMAN_Fields2015_y2h [16]	brca1_human_2	303	1 278	1 909	40.1	single
RL401_YEAST_Bolon2014 [12]	rl401_yeast_2	75	1 282	96 590	1.3	single
MTH3_HAESTABILIZED_Tawfik2015 [10]	mth3_haeaest.	329	1 611	75 260	2.1	single
POLG_HCVJF_Sun2014 [9]	polg_hcvif	86	1 613	17 184	8.6	single
BG_STRSQ_Abate2015 [11]	bg_strsq	478	2 598	124 487	2.0	single
BRCA1_HUMAN_Fields2015_e3 [16]	brca1_human_1	303	2 846	1 909	59.9	single
HSP82_YEAST_Bolon2016 [8]	hsp82_yeast	230	4 065	59 490	6.4	single
BLAT_ECOLX_Ostermeier2014 [4]	blat_ecolx_1	286	4 799	30 356	13.7	single
BLAT_ECOLX_Ranganathan2015 [17]	blat_ecolx_3	286	4 921	30 356	13.9	single
BLAT_ECOLX_Palzkill2012 [2]	blat_ecolx_2	286	4 922	30 356	14.0	single
HG_FLU_Bloom2016 [3]	hg_flu	564	10 337	71 029	12.7	single
PABP_YEAST_Fields2013 [7]	pabp_yeast_2	75	33 771	948 246	3.4	double
avGFP_Kondrashov2016 [14]	avgfp	235	32 610	697	97.9	multiple

Sorted by number of labelled sequences. *Length* stands for the number of amino acids in the protein sequence, L stands for the number of labelled sequences, U stands for the number of homologous sequences found (unlabelled) and $\%L$ represents the percentage of labelled sequences out of the total number of sequences.

Table S2: **Datasets included in experimentation and the protein function for which the fitness is given**.

Dataset	Protein function for which the fitness is given
YAP1_HUMAN_Fields2021 [1]	Ability of a WW domain to bind to a peptide ligand.
UBE4B_MOUSE_Klevit2013 [15]	Interaction strength between murine ubiquitination factor E4B (Ube4b) U-box domain and auto-ubiquitination activity in the presence of the E2 UBcH5c.
GAL4_YEAST_Shendure2015 [6]	Interaction strength between the DNA-binding domain (DBD) and the yeast transcription factor Gal4.
BLAT_ECOLX_Tenaillon2013 [5]	Resistance to ampicillin of TEM-1 β-lactamase enzyme.
PABP_YEAST_Fields2013 [7]	Enrichment scores of *Saccharomyces cerivisiae* poly-A binding protein (PABP).
RL401_YEAST_Bolon2013 [13]	Effects of mutations to ubiquitin on yeast growth.
BRCA1_HUMAN_Fields2015_y2h [16]	Interaction strength between breast cancer 1 (BRCA1) protein and the BRCA1-associated RING domain protein 1.
RL401_YEAST_Bolon2014 [12]	Effects of mutations to ubiquitin on activating enzyme (E1) activity.
MTH3_HAESTABILIZED_Tawfik2015 [10]	Methylation activity of HaeIII methyltransferase.
POLG_HCVJF_Sun2014 [9]	Effects of mutations to domain IA of NS5A inhibitor on Hepatitis C virus (HCV) replication rates.
BG_STRSQ_Abate2015 [11]	Hydrolysis activity of a β-glucosidase enzyme.
BRCA1_HUMAN_Fields2015_e3 [16]	Interaction strength between breast cancer 1 (BRCA1) protein and the BRCA1-associated RING domain protein 1.
HSP82_YEAST_Bolon2016 [8]	The spread of mutant lineages of the heat shock protein 90 (Hsp90).
BLAT_ECOLX_Ostermeier2014 [4]	Resistance to ampicillin of TEM-1 β-lactamase enzyme.
BLAT_ECOLX_Ranganathan2015 [17]	Resistance to ampicillin of TEM-1 β-lactamase enzyme.
BLAT_ECOLX_Palzkill2012 [2]	Resistance to ampicillin of TEM-1 β-lactamase enzyme.
HG_FLU_Bloom2016 [3]	Effects of mutations to H1 influenza hemagglutinin (HA) on viral replication rates
PABP_YEAST_Fields2013 [7]	Enrichment scores of *Saccharomyces cerivisiae* poly-A binding protein (PABP).
avGFP_Kondrashov2016 [14]	Brightness values of *Aequorea Victoria* green fluorescent protein (avGFP).

Table S3: **Time and peak memory performance of the training process of the best methods found for each strategy. Calculated using 100, 1000 and 10000 labeled training sequences.**

Method	Time (s)			Peak memory (MB)		
	N=100	N=1000	N=10000	N=100	N=1000	N=10000
Unirep + RF	4.14	113.55	2300.79	0.87	7.37	72.85
DCA + SVM	0.0024	0.093	12.05	0.23	1.67	16.09
DCA + MERGE[SVM]	0.71	0.81	29.69	0.29	1.67	16.16
PAM250 + TriTR[RF]	10.52	385.06	20368.90	89.56	151.75	1282.42
DCA + TriTR[SVM]	0.030	3.31	235.62	1.94	11.51	102.03

3 Supplementary algorithms

Algorithm S1 Original Tri-Training learning function.

function TRI-TRAINING(L, U, *base_classifier*)
  for $i \in \{1...3\}$ **do**
    $S_i \leftarrow BootstrapSample(L)$
    $h_i \leftarrow base\_classifier(S_i)$
    $e'_i \leftarrow 0.5$
    $l'_i \leftarrow 0$
  repeat
    for $i \in \{1...3\}$ **do**
      $L_i \leftarrow \emptyset$
      $update_i \leftarrow$ FALSE
      $e_i \leftarrow \dfrac{\left|\{\mathbf{x} \in L : h_j(\mathbf{x}) = h_k(\mathbf{x}) \neq y\}\right|}{\left|\{\mathbf{x} \in L : h_j(\mathbf{x}) = h_k(\mathbf{x})\}\right|}$ ▷ Number of labelled examples on which both h_j and h_k make incorrect classification, by the number of labelled examples on which the classification made by h_j is the same as that made by h_k.
      if $e_i < e'_i$ **then**
        for each $\mathbf{x} \in U$ **do**
          if $h_j(\mathbf{x}) = h_k(\mathbf{x}) \quad (j, k \neq i)$ **then**
            $L_i \leftarrow L_i \cup \{(\mathbf{x}, h_j(\mathbf{x}))\}$
        if $l'_i = 0$ **then**
          $l'_i \leftarrow \left\lfloor \frac{e_i}{e'_i - e_i} + 1 \right\rfloor$
        if $l'_i < |L_i|$ **then**
          if $e_i|L_i| < e'_i l'_i$ **then**
            $update_i \leftarrow$ TRUE
          else if $l'_i > \frac{e_i}{e'_i - e_i}$ **then**
            $L_i \leftarrow Subsample\left(L_i, \left\lceil \frac{e'_i l'_i}{e_i} - 1 \right\rceil\right)$
            $update_i \leftarrow$ TRUE
    for $i \in \{1...3\}$ **do**
      if $update_i =$ TRUE **then**
        $h_i \leftarrow base\_classifier(L \cup L_i)$
        $e'_i \leftarrow e_i$
        $l'_i \leftarrow |L_i|$
  until none of h_i ($i \in \{1...3\}$) changes
  return $h(\mathbf{x}) \leftarrow \arg\max_{y \in label} \sum_{i:h_i(\mathbf{x})=y}$

Algorithm S2 Tri-Training Regressor learning function.
Differences from Algorithm S1 are highlighted in red.

function TRI-TRAINING(L, U, *base_regressor*, *y_tol_per*)
  for $i \in \{1...3\}$ **do**
    $S_i \leftarrow BootstrapSample(L)$
    $h_i \leftarrow base\_regressor(S_i)$
    $e'_i \leftarrow 0.5$
    $l'_i \leftarrow 0$
    $y\_tol \leftarrow Tolerance(L, y\_tol\_per)$ ▷ Calculates the tolerance as a percentage (*y_tol_per*) of the difference between the largest output value and the smallest output value.
  repeat
    for $i \in \{1...3\}$ **do**
      $L_i \leftarrow \emptyset$
      $update_i \leftarrow$ FALSE
      $e_i \leftarrow \dfrac{\left|\{\mathbf{x} \in L : |h_j(\mathbf{x}) - h_k(\mathbf{x})| < y\_tol \wedge |mean(h_j(\mathbf{x}), h_k(\mathbf{x})) - y| \geq y\_tol\}\right|}{\left|\{\mathbf{x} \in L : |h_j(\mathbf{x}) - h_k(\mathbf{x})| < y\_tol\}\right|}$ ▷ Number of labelled examples for which both h_j and h_k predict similar values but different to the actual y, divided by the number of labelled examples for which both h_j and h_k predict similar values.
      if $e_i < e'_i$ **then**
        for each $\mathbf{x} \in U$ **do**
          if $|h_j(\mathbf{x}) - h_k(\mathbf{x})| < y\_tol \quad (j, k \neq i)$ **then**
            $L_i \leftarrow L_i \cup \{(\mathbf{x}, mean(h_j(\mathbf{x}), h_k(\mathbf{x}))\}$
        if $l'_i = 0$ **then**
          $l'_i \leftarrow \left\lfloor \frac{e_i}{e'_i - e_i} + 1 \right\rfloor$
        if $l'_i < |L_i|$ **then**
          if $e_i |L_i| < e'_i l'_i$ **then**
            $update_i \leftarrow$ TRUE
          else if $l'_i > \frac{e_i}{e'_i - e_i}$ **then**
            $L_i \leftarrow Subsample\left(L_i, \left\lceil \frac{e'_i l'_i}{e_i} - 1 \right\rceil\right)$
            $update_i \leftarrow$ TRUE
    for $i \in \{1...3\}$ **do**
      if $update_i =$ TRUE **then**
        $h_i \leftarrow base\_regressor(L \cup L_i)$
        $e'_i \leftarrow e_i$
        $l'_i \leftarrow |L_i|$
  until none of h_i $(i \in \{1...3\})$ changes
  return $h(\mathbf{x}) \leftarrow mean(h_1(\mathbf{x}), h_2(\mathbf{x}), h_3(\mathbf{x}))$

References

[1] Carlos L Araya, Douglas M Fowler, Wentao Chen, Ike Muniez, Jeffery W Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences*, 109(42):16858–16863, 2012.

[2] Zhifeng Deng, Wanzhi Huang, Erol Bakkalbasi, Nicholas G Brown, Carolyn J Adamski, Kacie Rice, Donna Muzny, Richard A Gibbs, and Timothy Palzkill. Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *Journal of molecular biology*, 424(3-4):150–167, 2012.

[3] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses*, 8(6):155, 2016.

[4] Elad Firnberg, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular biology and evolution*, 31(6):1581–1592, 2014.

[5] Hervé Jacquier, André Birgy, Hervé Le Nagard, Yves Mechulam, Emmanuelle Schmitt, Jérémy Glodt, Beatrice Bercot, Emmanuelle Petit, Julie Poulain, Guilène Barnaud, et al. Capturing the mutational landscape of the beta-lactamase tem-1. *Proceedings of the National Academy of Sciences*, 110(32):13067–13072, 2013.

[6] Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields, and Jay Shendure. Massively parallel single-amino-acid mutagenesis. *Nature methods*, 12(3):203–206, 2015.

[7] Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, and Stanley Fields. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein. *Rna*, 19(11):1537–1551, 2013.

[8] Parul Mishra, Julia M Flynn, Tyler N Starr, and Daniel NA Bolon. Systematic mutant analyses elucidate general and client-specific aspects of hsp90 function. *Cell reports*, 15(3):588–598, 2016.

[9] Hangfei Qi, C Anders Olson, Nicholas C Wu, Ruian Ke, Claude Loverdo, Virginia Chu, Shawna Truong, Roland Remenyi, Zugen Chen, Yushen Du, Sheng-Yao Su, Laith Q Al-Mawsawi, Ting-Ting Wu, Shu-Hua Chen, Chung-Yen Lin, Weidong Zhong, James O Lloyd-Smith, and Ren Sun. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. *PLoS pathogens*, 10(4):e1004064, 2014.

[10] Liat Rockah-Shmuel, Ágnes Tóth-Petróczy, and Dan S Tawfik. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLoS computational biology*, 11(8):e1004421, 2015.

[11] Philip A Romero, Tuan M Tran, and Adam R Abate. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, 112(23):7159–7164, 2015.

[12] Benjamin P Roscoe and Daniel NA Bolon. Systematic exploration of ubiquitin sequence, e1 activation efficiency, and experimental fitness in yeast. *Journal of molecular biology*, 426(15):2854–2870, 2014.

[13] Benjamin P Roscoe, Kelly M Thayer, Konstantin B Zeldovich, David Fushman, and Daniel NA Bolon. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of molecular biology*, 425(8):1363–1377, 2013.

[14] Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

[15] Lea M Starita, Jonathan N Pruneda, Russell S Lo, Douglas M Fowler, Helen J Kim, Joseph B Hiatt, Jay Shendure, Peter S Brzovic, Stanley Fields, and Rachel E Klevit. Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14):E1263–E1272, 2013.

[16] Lea M Starita, David L Young, Muhtadi Islam, Jacob O Kitzman, Justin Gullingsrud, Ronald J Hause, Douglas M Fowler, Jeffrey D Parvin, Jay Shendure, and Stanley Fields. Massively parallel functional analysis of brca1 ring domain variants. *Genetics*, 200(2):413–422, 2015.

[17] Michael A Stiffler, Doeke R Hekstra, and Rama Ranganathan. Evolvability as a function of purifying selection in tem-1 β-lactamase. *Cell*, 160(5):882–892, 2015.

[18] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.