Supporting Information: Artificial Neural Network to Predict Structure-based Protein-protein Free Energy of Binding from Rosetta-calculated Properties

Matheus V. F. Ferraz,[†,‡,¶] José C.S. Neto,[§] Roberto D. Lins,[†,‡] and Erico S. Teixeira[*,§]

[†]*Aggeu Magalhães Institute, Oswaldo Cruz Foundation, FIOCRUZ, Recife, PE, 50670-465, Brazil*

[‡]*Department of Fundamental Chemistry, Federal University of Pernambuco, UFPE, Recife, PE, 50740-670, Brazil*

[¶]*Heidelberg Institute for Theoretical Studies,HITS, Heidelberg, 69118, Heidelberg, Germany*

[§]*Recife Center for Advanced Studies and Systems, CESAR, Recife, PE, 50040-220, Brazil*

E-mail: est@cesar.school

Content

Rosetta parsed command lines

Energy minimization:

```
$ ./minimize.macosclangrelease -l [list-of-pdbs] -min_all_jumps true -run::min_type lbfgs_armijo
_nonmonotone -use_input_sc true -ex1 -ex2 -extrachi_cutoff 1 -no_his_his_pairE true -no_optH
false -ignore_unrecognized_res -ndruns 5
```

Properties calculations:

```
$ ./rosetta_scripts.macosclangrelease -l [list-of-minimized-pdb] -parser:protocol interface_analysis.xml
-ignore_unrecognized_res -no_his_his_pairE -out:file:score_only ifa.sc -no_optH false -ex1 -ex2
-use_input_sc -run::min_type lbfgs_armijo_nonmonotone -extrachi_cutoff 1 -linmem_ig 10 -
atomic_burial_cutoff 0.01 -sasa_calculator_probe_radius 1.2
```

Rosetta scripts in XML format

XML to calculate interface properites (interface_analysis.xml)

```
< ROSETTASCRIPTS >
< SCOREFXNS >
< ScoreFunctionname = ”ref2015”weights = ”ref2015”/ >
< /SCOREFXNS >
< FILTERS >
< ShapeComplementarityname = ”Sc”min_sc = ”2.0”write_int_area = ”1”jump =
”1”confidence = ”0”/ >
< Ddgname = ”ddg”scorefxn = ”ref2015”threshold = ”0”jump = ”1”repeats =
”5”repack = ”1”repack_bound = ”0”confidence = ”0”/ >
< /FILTERS >
< MOVERS >
< InterfaceAnalyzerMovername = ”ifa”scorefxn = ”ref2015”pack_separated =
”1”pack_input = ”1”tracer = ”0”interface_sc = ”1”interface = ”A_B”/ >
```

```
< /MOVERS >

< PROTOCOLS >

< Addmover = ”ifa”/ >

< Addfilter = ”Sc”/ >

< Addfilter = ”ddg”/ >

< /PROTOCOLS >

< /ROSETTASCRIPTS >
```

Polar atom definition

The SASA for a polar atom is calculated as the sum of the SASA for that specific atom and the SASA for any bound hydrogen. Polar atoms presenting SASA smaller than 0.1 Å^2 are considered buried. Hydrogen bonds between the donor and acceptors atoms with a SASA smaller than 3.0 Å^2 are considered buried. Atomic radii from the Reduce software (1) and a water probe radius of 1.2 Å^2 were employed to map buried polar atoms and hydrogen bonds. These values were reasoned by probability distributions of hydration water molecules around polar atoms from data collection of high-resolution PDB structures.(2)

Table S1. Calculated $R_{Pearson}$ in ascending order for the correlation between the features value and the experimental ΔG

Feature	$R_{Pearson}$
dslf_fa13	0.336668
hbond_bb_sc	0.252102
hbond_lr_bb	0.215789
p_aa_pp	0.210008
lk_ball_wtd	0.113917
hbond_sc	0.111662
fa_atr	0.109970
fa_elec	0.101131
complex_normalized	0.055793
total_score	0.047824
side1_score	0.046541
omega	0.031989
side2_score	0.029075
fa_rep	0.022766
sc_value	0.012860
Sc	0.004894
dG_cross	-0.033080
pro_close	-0.037478
dG_separated	-0.037859
side2_normalized	-0.048827
fa_dun	-0.053386
ddg	-0.053916

Feature	$R_{Pearson}$
per_residue_energy_int	-0.056136
fa_intra_rep	-0.056630
hbond_sr_bb	-0.056821
side1_normalized	-0.061778
dG_separated/dSASAx100	-0.064651
dG_cross/dSASAx100	-0.065108
fa_intra_sol_xover4	-0.068651
hbond_E_fraction	-0.070499
rama_prepro	-0.078543
fa_sol	-0.096505
nres_all	-0.121934
ref	-0.270164
hbonds_int	-0.346316
delta_unsatHbonds	-0.378498
dSASA_polar	-0.397664
dSASA_hphobic	-0.439358
nres_int	-0.451539
dSASA_int	-0.458725
Sc_int_area	-0.532643

Table S2. Comparison of the predicted ΔG of binding using the ANN and experimental ΔG of binding for the 19 cases of the metadynamics-validation set.

PDB ID	Experimental ΔG (kcal.mol^{-1})	ANN ΔG (kcal.mol^{-1})
1ACB	13.76	-11.254782
1AY7	13.76	-11.054798
1BVN	15.65	-11.545321
1EMV	19.32	-14.301220
1FFW	8.33	-8.465515
1KAC	11.11	-9.067882
1KTZ	9.27	-10.862952
1QA9	7.16	-8.139755
1R0R	14.94	-12.371928
1US7	8.28	-10.642823
2C0L	9.88	-12.066045
2OOB	5.99	-8.733976
2PTC	18.75	-13.219584
2UUY	11.7	-11.982295
3A4S	7.87	-8.636804
3BZD	9.95	-9.275232
3F1P	8.3	-9.549908
3LVK	9.25	-10.150698
3SGB	15.24	-11.496317

Table S3. Calculated Rosetta folding and interface properties. Short description of the features based on the Rosetta package energy function. Only features representing energetic and/or geometric terms were considered.

Feature	**Description**
dslf_fa13	Disulfide geometry potential
hbond_bb_sc	Energy of backbone-side chain hydrogen bonding
hbond_lr_bb	Energy of long-range hydrogen bonding
p_aa_pp	Probability of amino acid at ϕ/ψ
lk_ball_wtd	Orientation-dependent solvation of polar atoms
hbond_sc	Energy of side chain to side chain hydrogen bonding
fa_atr	Attractive energy between two atoms on different residues separated by a given distance
fa_elec	Coulombic potential energy for two atoms separated by a given distance
complex_normalized	Average energy of a residue in the entire complex
total_score	Relative folding free energy
side1_score	Folding energy of the first interface
omega	Omega dihedral in the backbone
side2_score	Folding energy of the second interface
fa_rep	Lennard-Jones repulsive between atoms in different residues
Sc	Shape complementarity
dG_cross	Interaction energy
pro_close	Proline ring closure energy
dG_separated	Binding free energy
side2_normalized	Average per-residue energy on the second interface
fa_dun	Probability of a chosen rotamer is native-like conformation given backbone ϕ, ψ angles
ddg	Change in the binding free energy

Feature	**Description**
per_residue_energy_int	Average energy of each residue at the interface
fa_intra_rep	Intra-residue repulsive component
hbond_sr_bb	Energy of short-range hydrogen bonding
side1_normalized	Average per-residue energy on the first interface
dG_separated/dSASAx100	Binding free energy divided by the total solvent accessible surface area multiplied by 100
dG_cross/dSASAx100	Interaction energy divided by the total solvent accessible surface area multiplied by 100
fa_intra_sol_xover4	Gaussian exclusion implicit solvation energy
hbond_E_fraction	Contribution of the hydrogen bonding potentials to the binding energy
rama_prepro	Backbone torsion preference term
fa_sol	Gaussian exclusion implicit solvation energy
nres_all	Total number of residues
ref	Reference energy for each amino acid relatively to unfolding.
hbonds_int	Number of hydrogen bonds in the interface
delta_unsatHbonds	Number of buried hydrogen bonds in the interface
dSASA_polar	Polar solvent accessible surface area
dSASA_hphobic	Hydrophobic solvent accessible surface area
nres_int	Number of residues in the interface
dSASA_int	Total solvent accessible surface area
Sc_int_area	Shape complementarity divided by interface area

Table S4. Codes of the PDB used for the test set along with its binding affinity in kcal.mol^{-1}. Binding affinities were retrieved from the PDBind data set in form of k_D and converted using thermodynamic relationships

PDB ID	k_D **(kcal.mol^{-1}))**	**PDB ID**	k_D **(kcal.mol^{-1}))**
2WH6	-10.5	5H3J	-8.95
2WP3	-8.31	5INB	-9.42
3V1C	-10.25	5MA4	-14.02
3VFN	-9.17	5NT7	-6.78
3WQB	-11.92	5TZP	-10.25
4B1Y	-8.95	5V5H	-9.27
4CJ0	-9.55	5XCO	-10.97
4CJ2	-10.85	5YWR	-10.1
4K5A	-10.89	6B6U	-6.4
4KT3	-13.06	6E3I	-11.58
4LZX	-11.31	6E3J	-12.07
4M0W	-7.05	6HER	-10.08
4NL9	-9	6JB2	-8.09
4PJ2	-14.08	6FU9	-9.99
4QLP	-13.14	6FUB	-10.27
4UYP	-14.58	6FUD	-9.73
4WND	-10.2	6J14	-11.46
4X33	-9	5IMK	-8.27
4YL8	-7.18	5IMM	-11.52
4Z99K	-11.8	5KXH	-8.6
5B78	-7.8	5KY4	-7.95
5DC4	-10.16	5KY5	-8.32
5DJT	-10.59	6DDM	-12.78
5E95	-10.64	6FG8	-8.19
5EP6	-8.37	6NE2	-12.11

Figure S1. Histogram containing all the standardized range value for all features without outliers

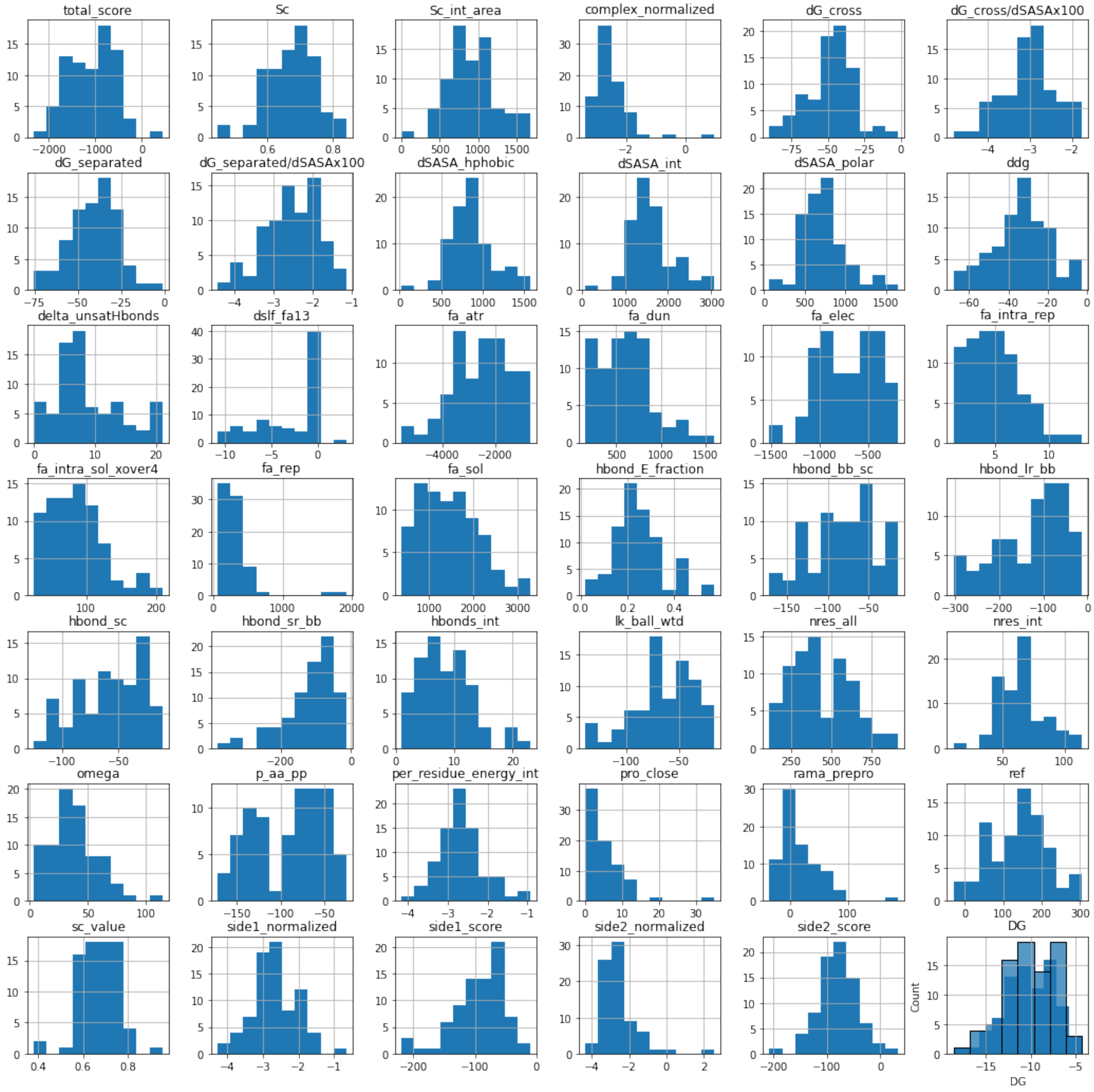


Figure S2. Correlation between the predicted and experimental ΔG of binding for the separated training sets using the ANN and PRODIGY methods

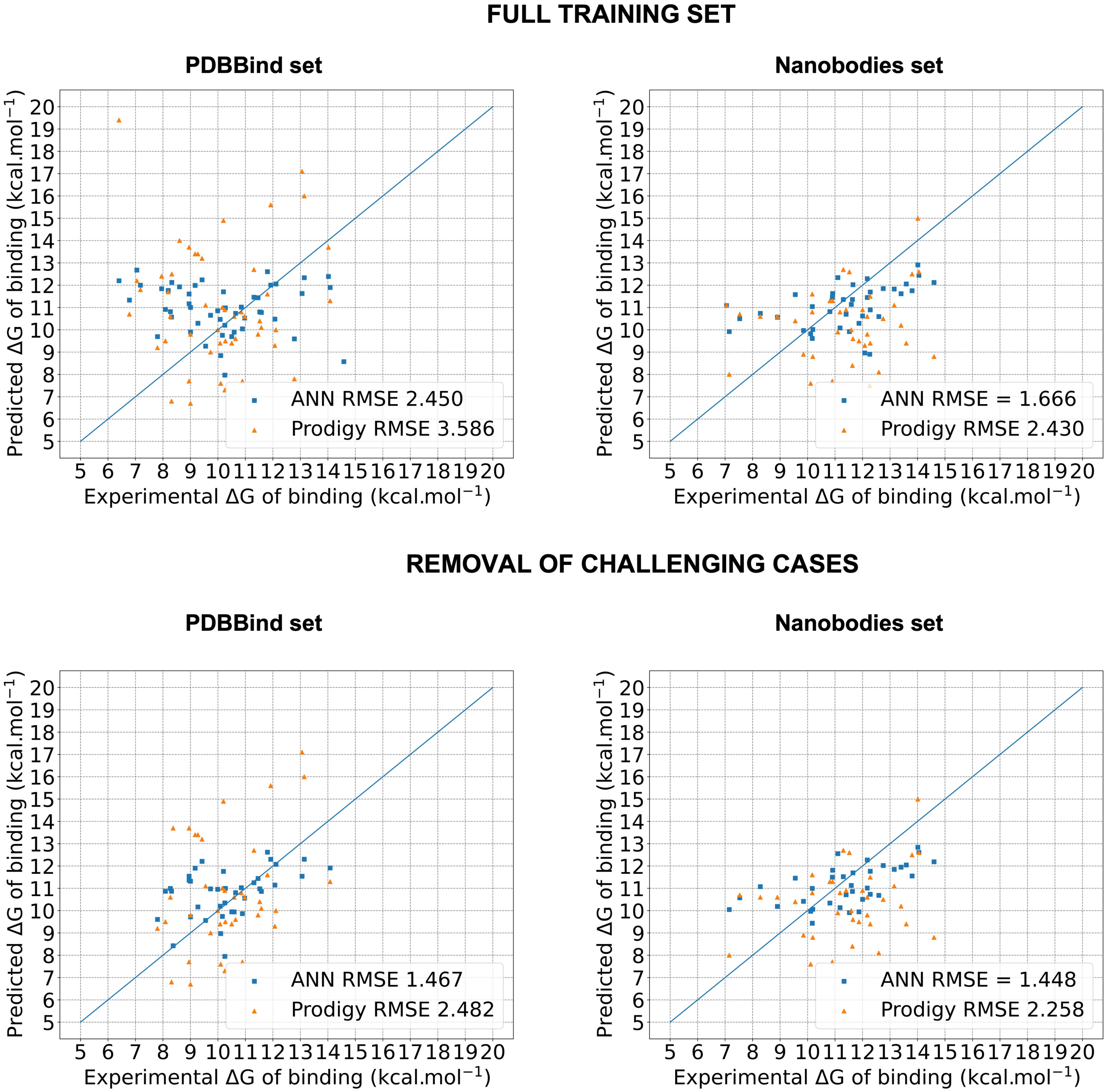


Figure S3. Feature importance score for all the features

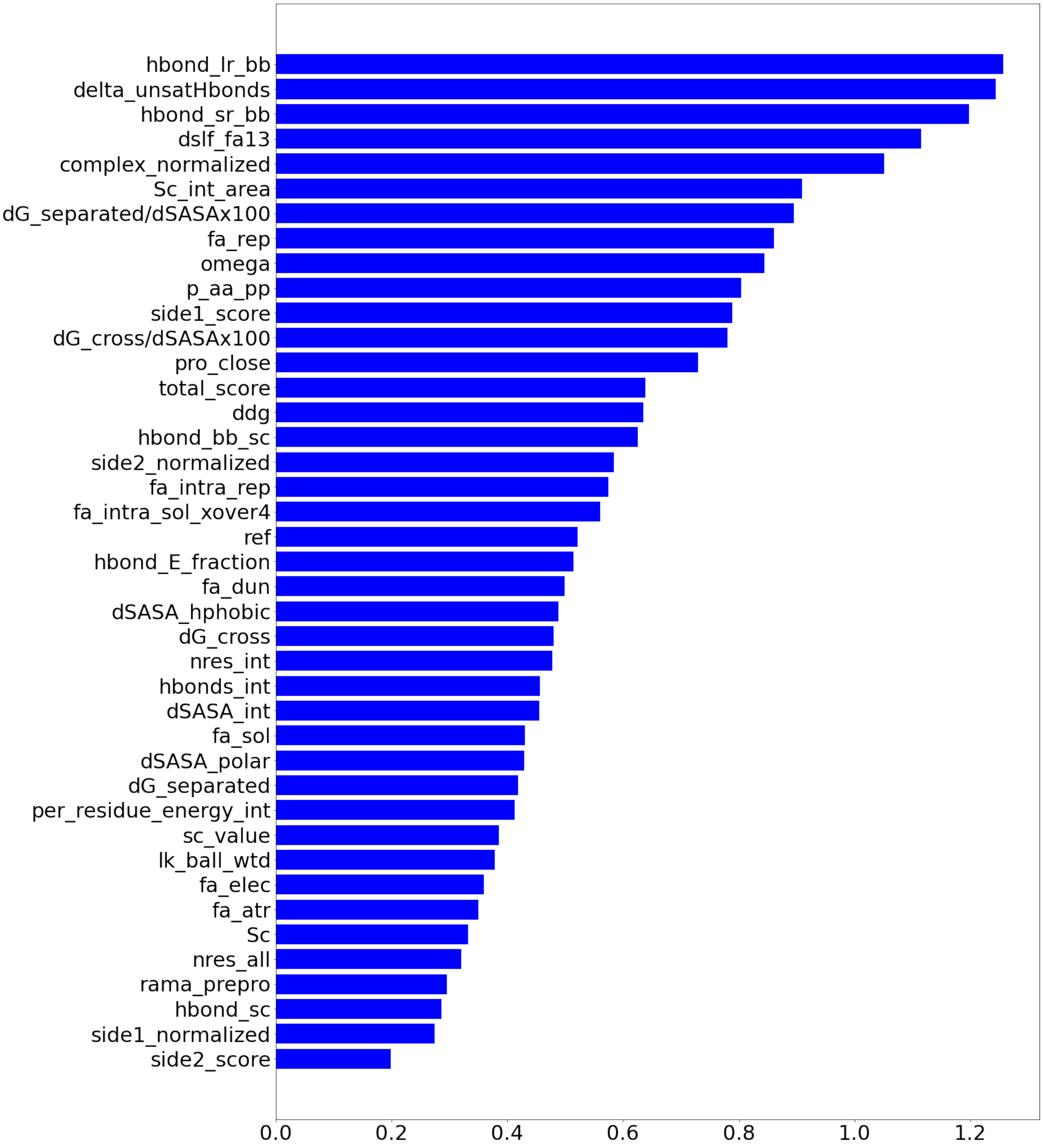


Figure S4. Histogram containing all the original range value for all features

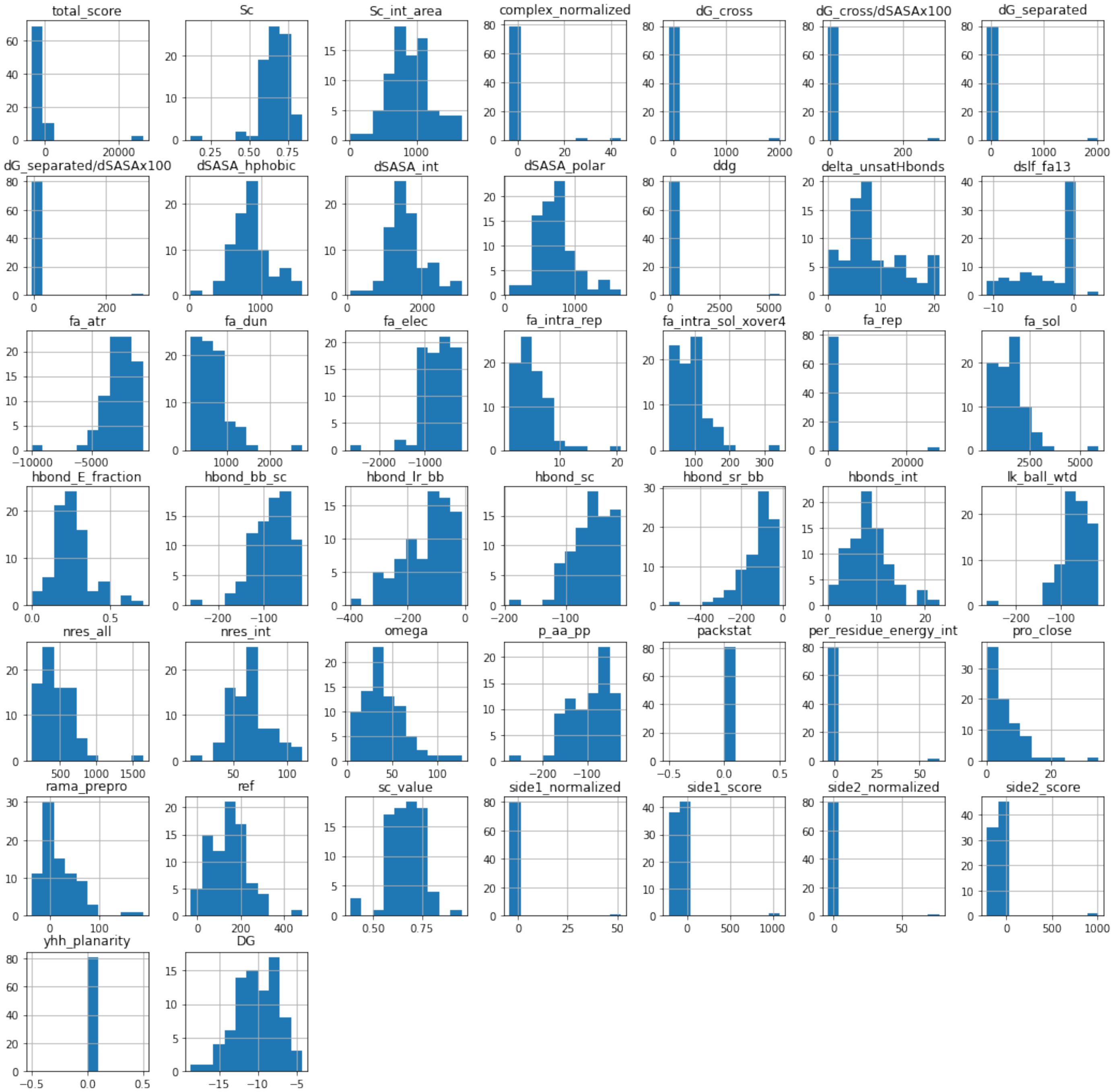


Figure S5. Histogram containing all the standardized range value for all features

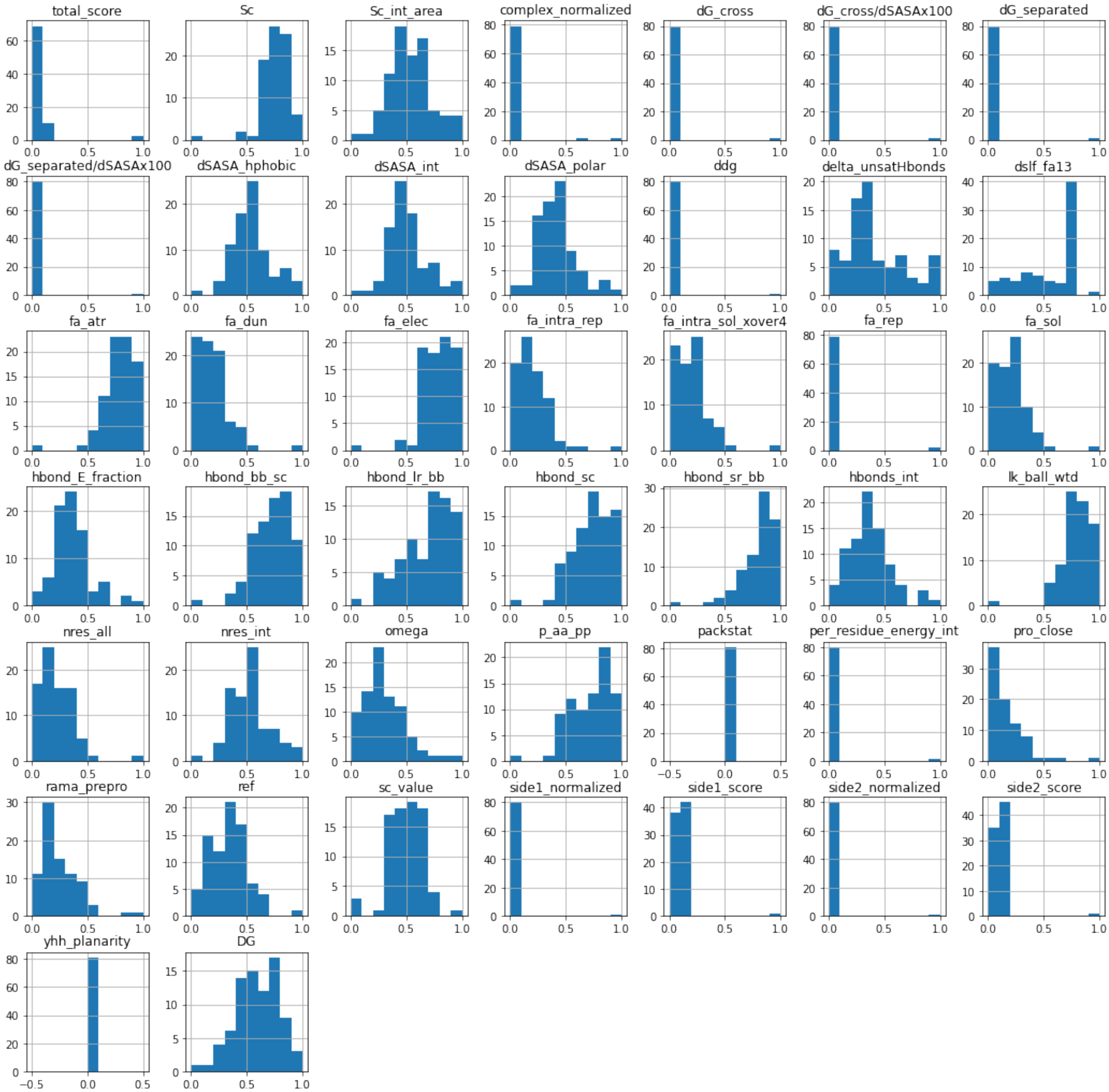
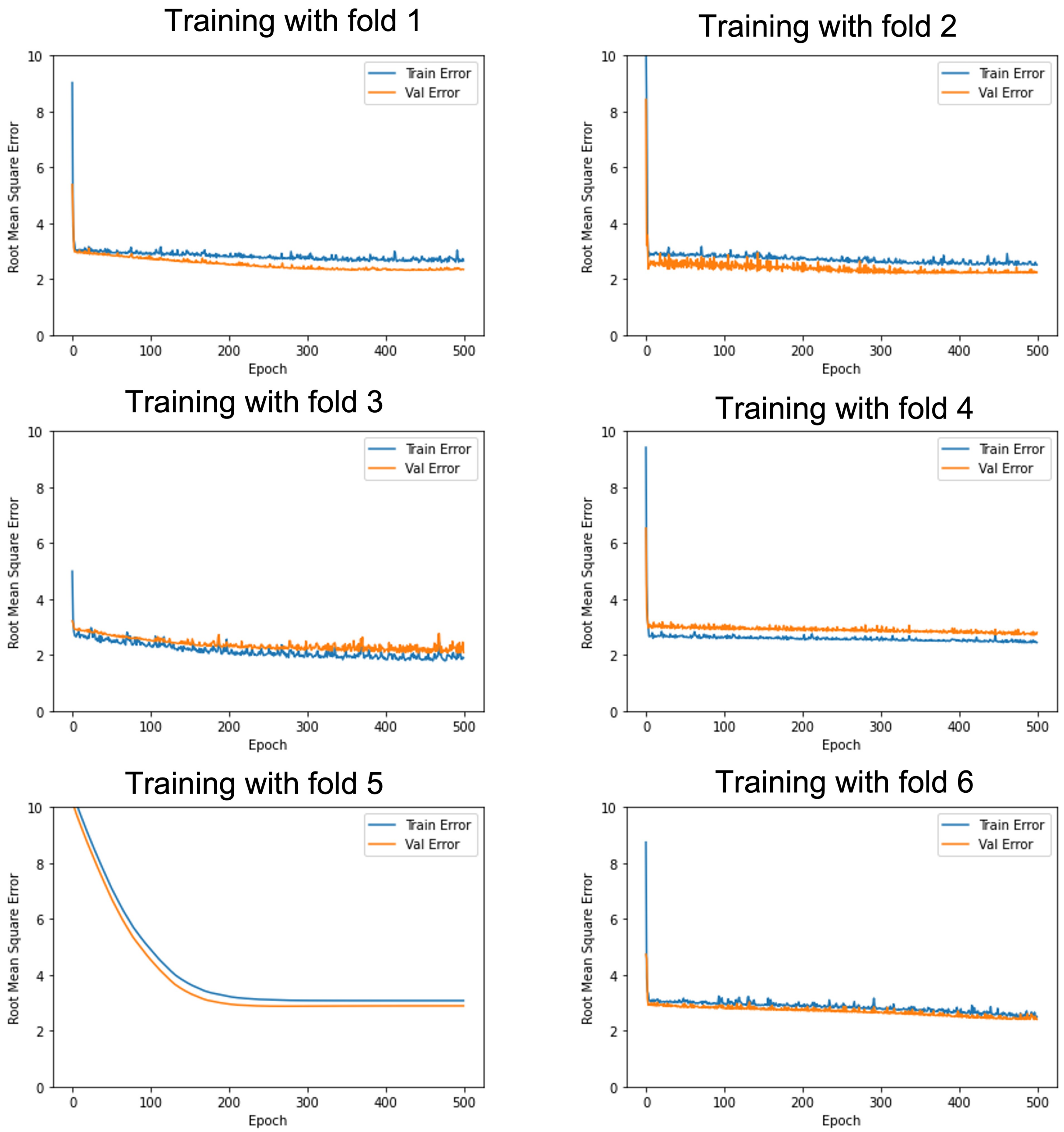
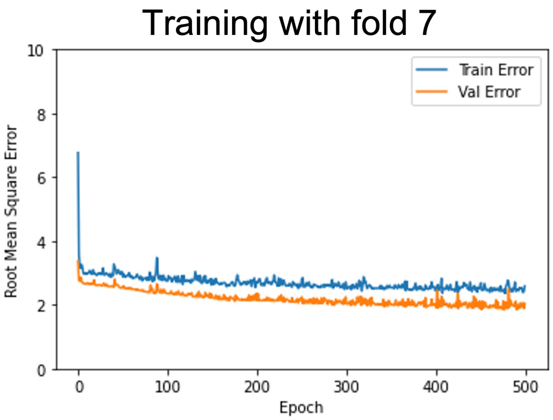
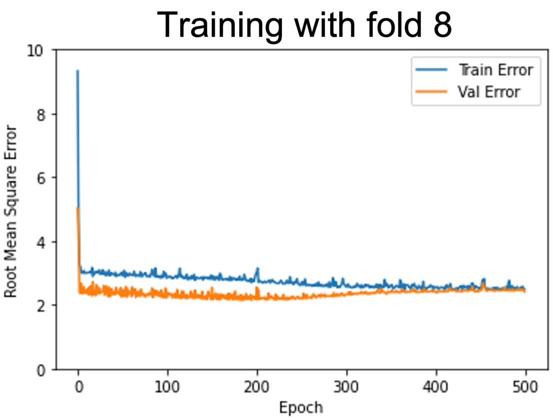


Figure S6. Evaluation of the number of epochs as a function of the root mean square error for a k-fold training where $k \in \{1, .., 10\}$

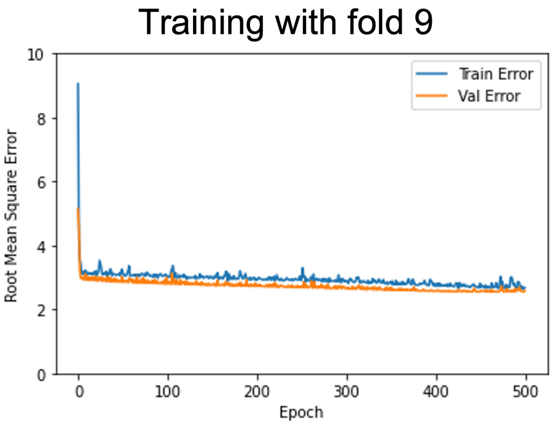




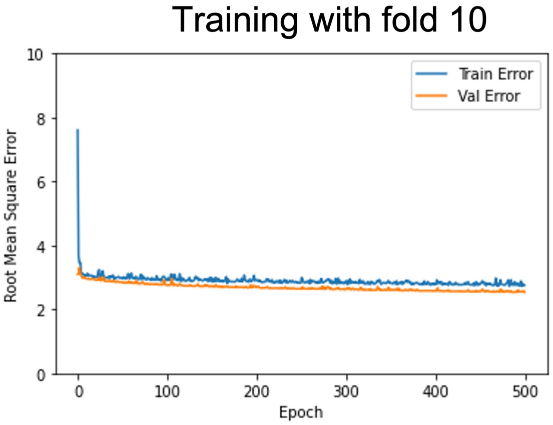
Training with fold 7



Training with fold 8



Training with fold 9



Training with fold 10

References

.

(1) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. Journal of molecular biology 1999, 285, 1735–1747.

(2) Matsuoka, D.; Nakasako, M. Probability distributions of hydration water molecules around polar protein atoms obtained by a database analysis. The Journal of Physical Chemistry B 2009, 113, 11274–11292.