TransDTI: Transformer-based language models for estimating DTIs and building a drug-recommendation workflow

[Supplementary file 1]

Yogesh Kalakoti, Shashank Yadav and Durai Sundar[*]

DAILAB, Department of Biochemical Engineering & Biotechnology, Indian Institute of Technology (IIT) Delhi, New Delhi - 110 016, India

[*]Corresponding author

Email addresses:

YK: yogesh.kalakoti@dbeb.iitd.ac.in

SY: shashank.yadav1@alumni.iitd.ac.in

DS: sundar@dbeb.iitd.ac.in

1. Supplementary tables

Table S1: Paired t-test to compare docking scores of all the models under consideration for tgfb.

	ESM Family			ProtBert family						Alphafold	Validations		
	Esm1	**Esm1b**	**Esm1v**	**protbert**	**protebertbfd**	**prott5xl**	**prott5xlbfd**	**protxlnet**	**protalbert**		**DeepDTI**	**DeepConvDTI**	**DeepDTA**
Esm1	0.00	2.81	-2.43	0.63	-1.73	2.63	2.60	**-1.27**	**-4.06**	**2.01**	**11.47**	**7.42**	**4.75**
Esm1b	-2.81	0.00	**-3.78**	-0.74	**-3.65**	-0.48	-0.25	-2.80	**-5.52**	**0.20**	**8.42**	**4.84**	2.68
Esm1v	2.43	**3.78**	0.00	2.26	**1.12**	**3.64**	**3.67**	1.03	-0.91	**3.39**	**8.66**	**6.55**	**5.15**
protbert	-0.63	0.74	-2.26	0.00	-1.59	0.52	0.62	-1.39	**-3.30**	0.78	**5.57**	**3.60**	2.41
Protebertbfd*	1.73	**3.65**	-1.12	1.59	0.00	**3.51**	**3.50**	0.04	-2.37	2.95	**10.28**	**7.28**	**5.25**
prott5xl	-2.63	0.48	**-3.64**	-0.52	**-3.51**	0.00	0.22	-2.62	**-5.44**	0.52	**9.51**	**5.56**	**3.15**
prott5xlbfd	-2.60	0.25	**-3.67**	-0.62	**-3.50**	-0.22	0.00	-2.67	**-5.41**	0.37	**8.76**	**5.11**	2.88
protxlnet	1.27	2.80	-1.03	1.39	-0.04	2.62	2.67	0.00	-2.09	2.48	**8.21**	**5.89**	**4.38**
Protalbert*	**4.06**	**5.52**	0.91	**3.30**	2.37	**5.44**	**5.41**	2.09	0.00	**4.73**	**10.86**	**8.44**	**6.73**
alphafold	-2.01	**-0.20**	**-3.39**	-0.78	-2.95	-0.52	-0.37	-2.48	-4.73	0.00	**5.86**	**3.39**	1.93
DeepDTI	**-11.47**	**-8.42**	**-8.66**	**-5.57**	**-10.28**	**-9.51**	**-8.76**	**-8.21**	**-10.86**	-5.86	0.00	-2.83	**-3.95**
DeepConvDTI	**-7.42**	**-4.84**	**-6.55**	**-3.60**	**-7.28**	**-5.56**	**-5.11**	**-5.89**	**-8.44**	-3.39	2.83	0.00	-1.38
DeepDTA	**-4.75**	-2.68	**-5.15**	-2.41	**-5.25**	**-3.15**	-2.88	**-4.38**	**-6.73**	-1.93	**3.95**	1.38	0.00

*Significant scores (p-value<0.005) have been marked in **bold***

** Relatively better models based on t-test stastic*

Table S2: Paired t-test to compare docking scores of all the models under consideration for MAP2K.

	ESM Family			ProtBert family						Alphafold	Validations		
	Esm1	**Esm1b**	**Esm1v**	**protbert**	**protebertbfd**	**prott5xl**	**prott5xlbfd**	**protxlnet**	**protalbert**		**DeepDTI**	**DeepConvDTI**	**DeepDTA**
Esm1	0.00	0.00	**-6.27**	**-4.39**	**-7.51**	**-8.74**	**-21.80**	**-6.28**	**-12.51**	**-9.69**	0.81	-0.87	**-3.32**
Esm1b	0.00	0.00	**-6.27**	**-4.39**	**-7.51**	**-8.74**	**-21.80**	**-6.28**	**-12.51**	**-9.69**	0.81	-0.87	**-3.32**
Esm1v	**6.27**	**6.27**	0.00	-0.82	-1.44	-0.42	**-10.93**	0.24	**-4.43**	-2.65	**6.50**	**4.00**	1.21
protbert	**4.39**	**4.39**	0.82	0.00	-0.11	0.65	**-5.14**	0.98	-1.72	-0.77	**4.69**	3.44	1.60
protebertbfd	**7.51**	**7.51**	1.44	0.11	0.00	1.31	**-8.51**	1.71	-2.60	-1.03	**7.68**	**5.13**	2.33
prott5xl	**8.74**	**8.74**	0.42	-0.65	-1.31	0.00	**-14.12**	0.74	**-5.14**	-2.81	**8.59**	**4.94**	1.66
prott5xlbfd*	**21.80**	**21.80**	**10.93**	**5.14**	**8.51**	**14.12**	0.00	**11.78**	**7.10**	**8.29**	**20.22**	**13.78**	**9.33**
protxlnet	**6.28**	**6.28**	-0.24	-0.98	-1.71	-0.74	**-11.78**	0.00	**-4.89**	**-3.00**	**6.50**	**3.89**	1.05
protalbert	**12.51**	**12.51**	**4.43**	1.72	2.60	5.14	**-7.10**	**4.89**	0.00	1.68	**12.14**	**8.12**	**4.64**
Alphafold*	**9.69**	**9.69**	2.65	0.77	1.03	2.81	**-8.29**	**3.00**	-1.68	0.00	**9.64**	**6.43**	**3.30**
DeepDTI	-0.81	-0.81	**-6.50**	**-4.69**	**-7.68**	**-8.59**	**-20.22**	**-6.50**	**-12.14**	**-9.64**	**0.00**	-1.41	-3.71
DeepConvDTI	0.87	0.87	**-4.00**	-3.44	**-5.13**	**-4.94**	**-13.78**	**-3.89**	**-8.12**	**-6.43**	**1.41**	0.00	-2.17
DeepDTA	**3.32**	**3.32**	-1.21	-1.60	-2.33	-1.66	**-9.33**	-1.05	**-4.64**	**-3.30**	**3.71**	2.17	0.00

*Significant scores (p-value<0.005) have been marked in **bold***

** Relatively better models based on t-test stastic*

2. Supplementary figures

Figure S1: Schematic representation of the seed model architecture on which all the proposed methods are based on.

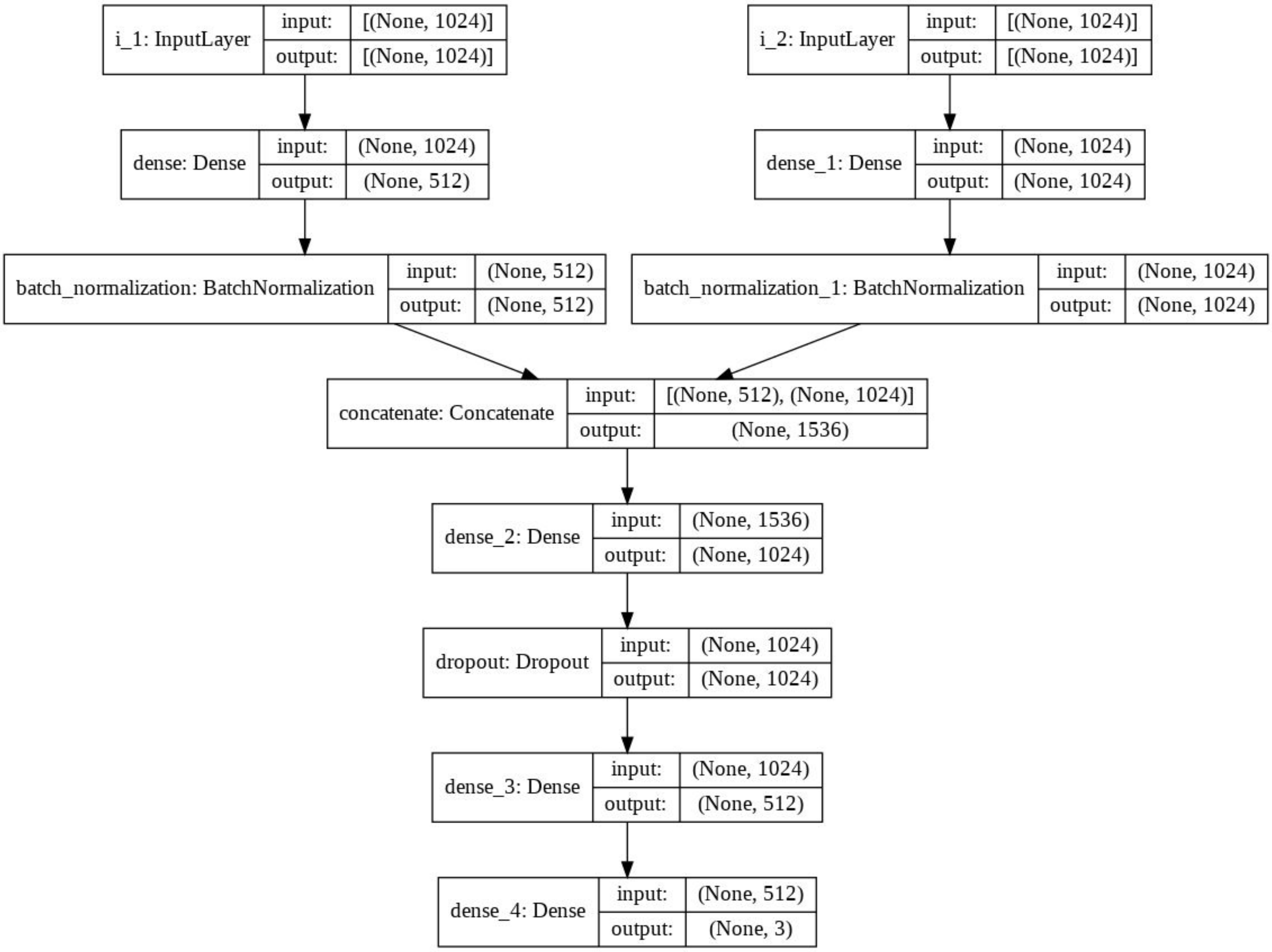


Figure S2: Interaction dynamics for map2k and tgfb from 100ns simulation

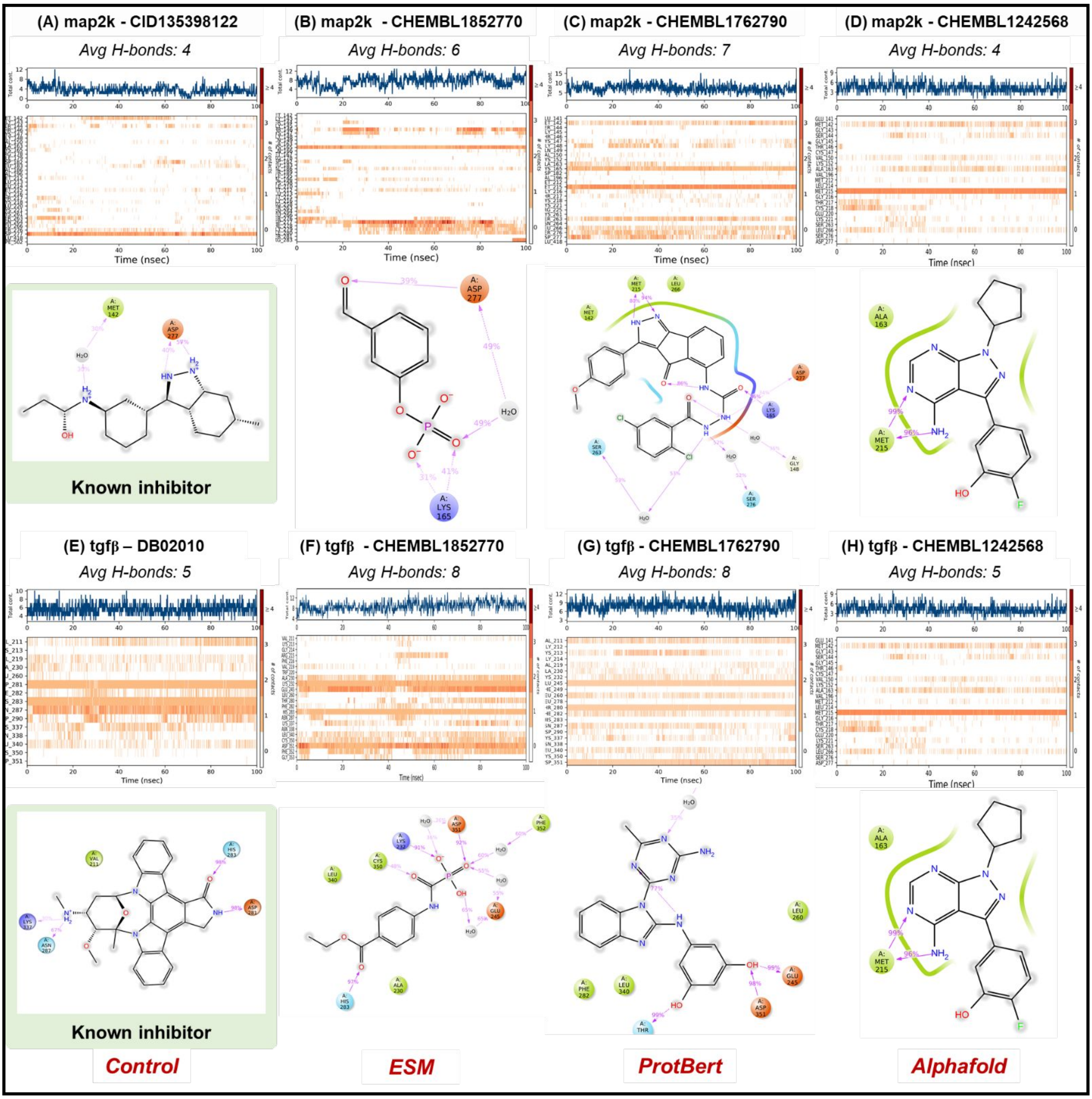


Figure S3: RMSD for the simulated complexes

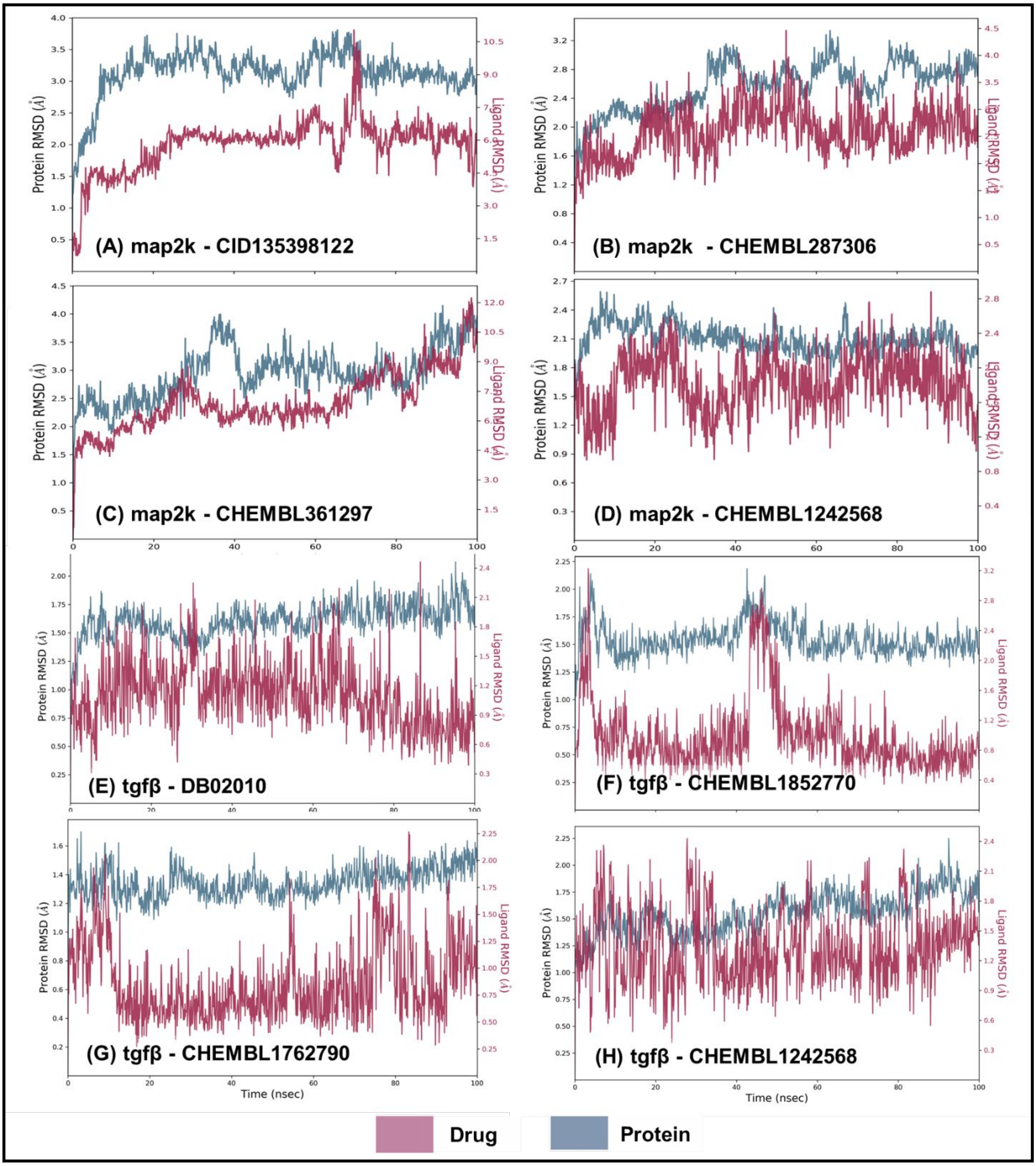
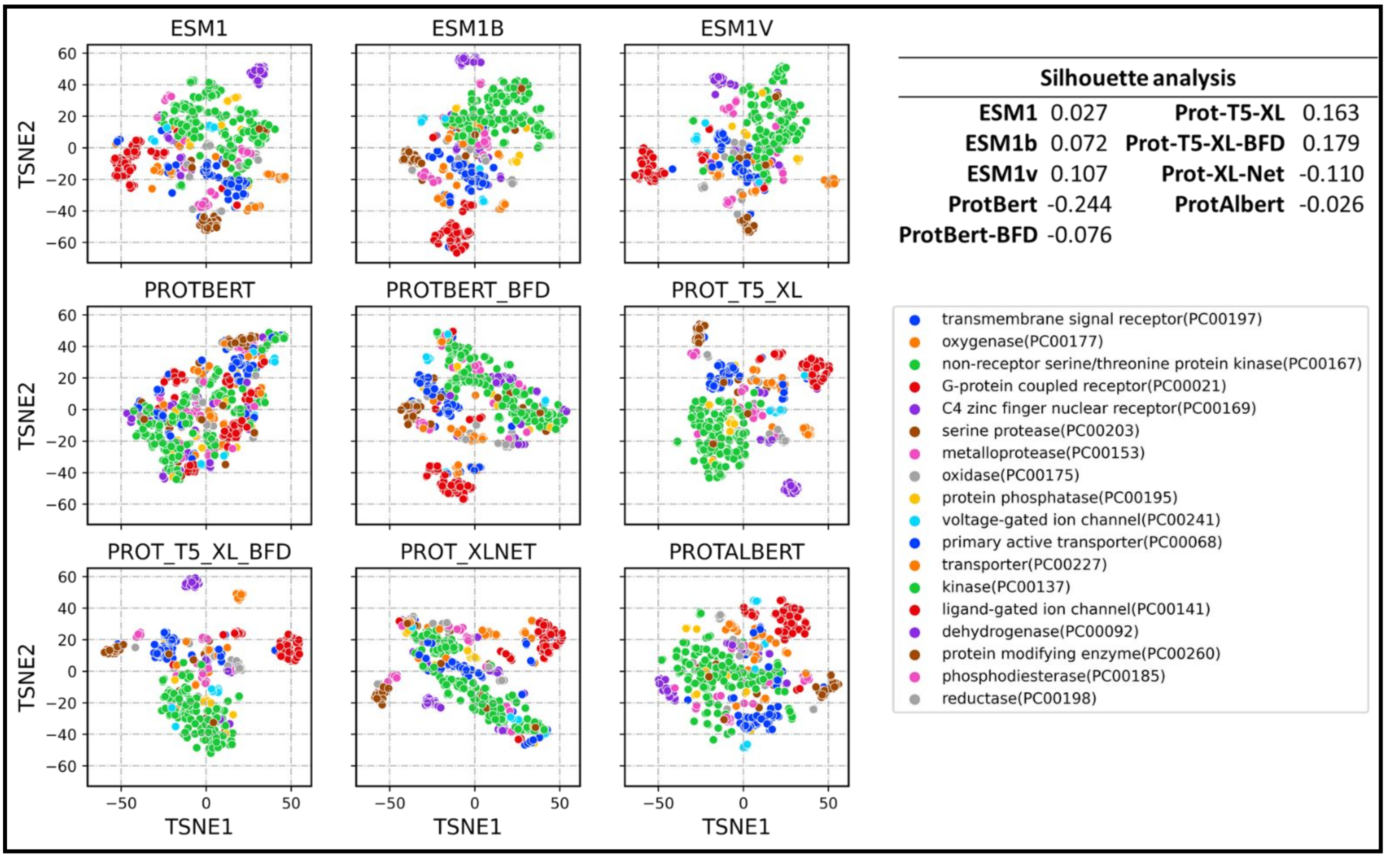


Figure S4: TSNE mappings for proteins



Silhouette analysis

ESM1	0.027	**Prot-T5-XL**	0.163
ESM1b	0.072	**Prot-T5-XL-BFD**	0.179
ESM1v	0.107	**Prot-XL-Net**	-0.110
ProtBert	-0.244	**ProtAlbert**	-0.026
ProtBert-BFD	-0.076		

Figure S5: ROC and PR curves for all the models under consideration in the classification task. auPR and auROC scores are also mentioned for each model

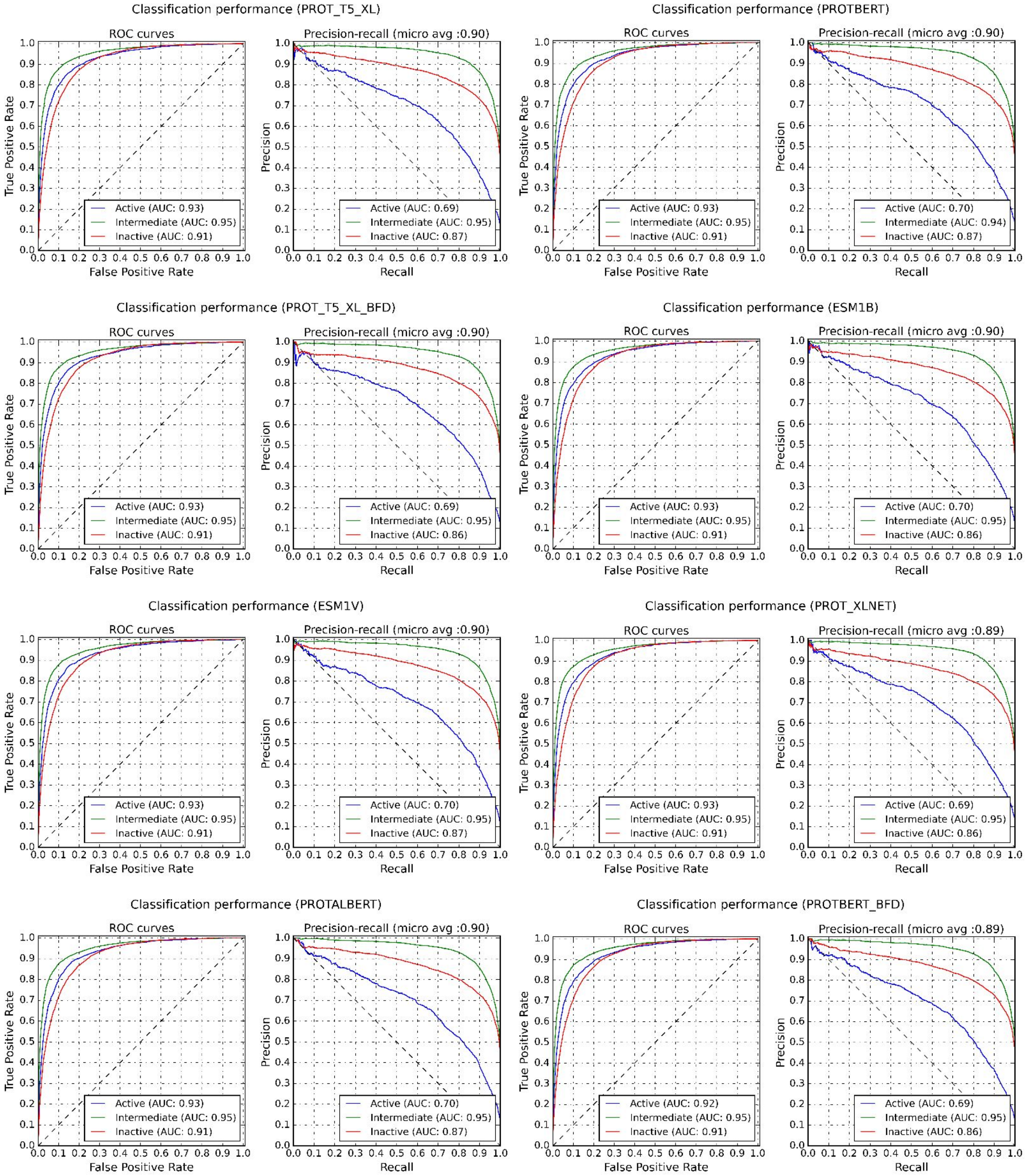


Figure S6: Coefficient of determination for all the proposed models

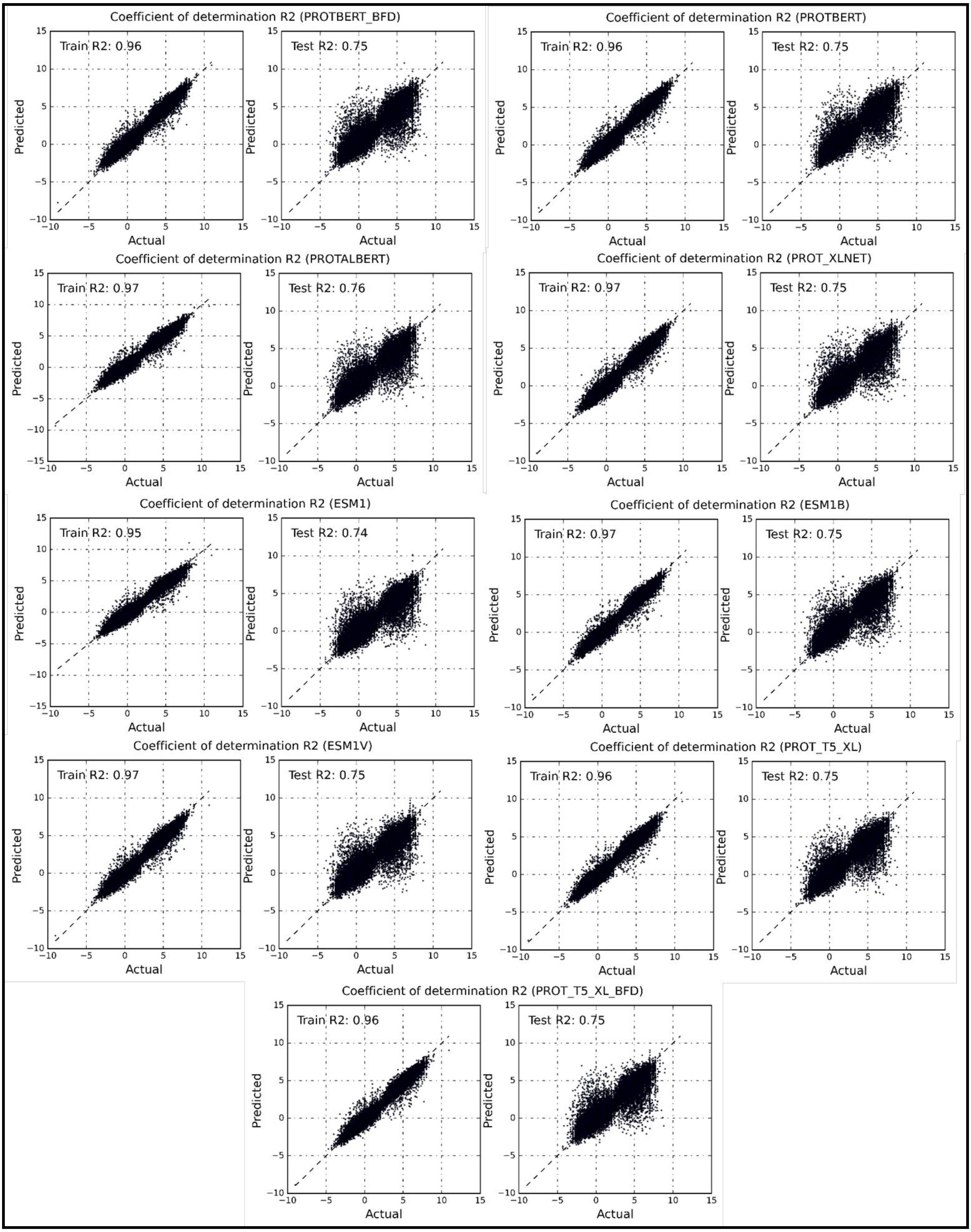


Figure S7: Training statistics for all the models under consideration shows excellent statistics and minimal overfitting

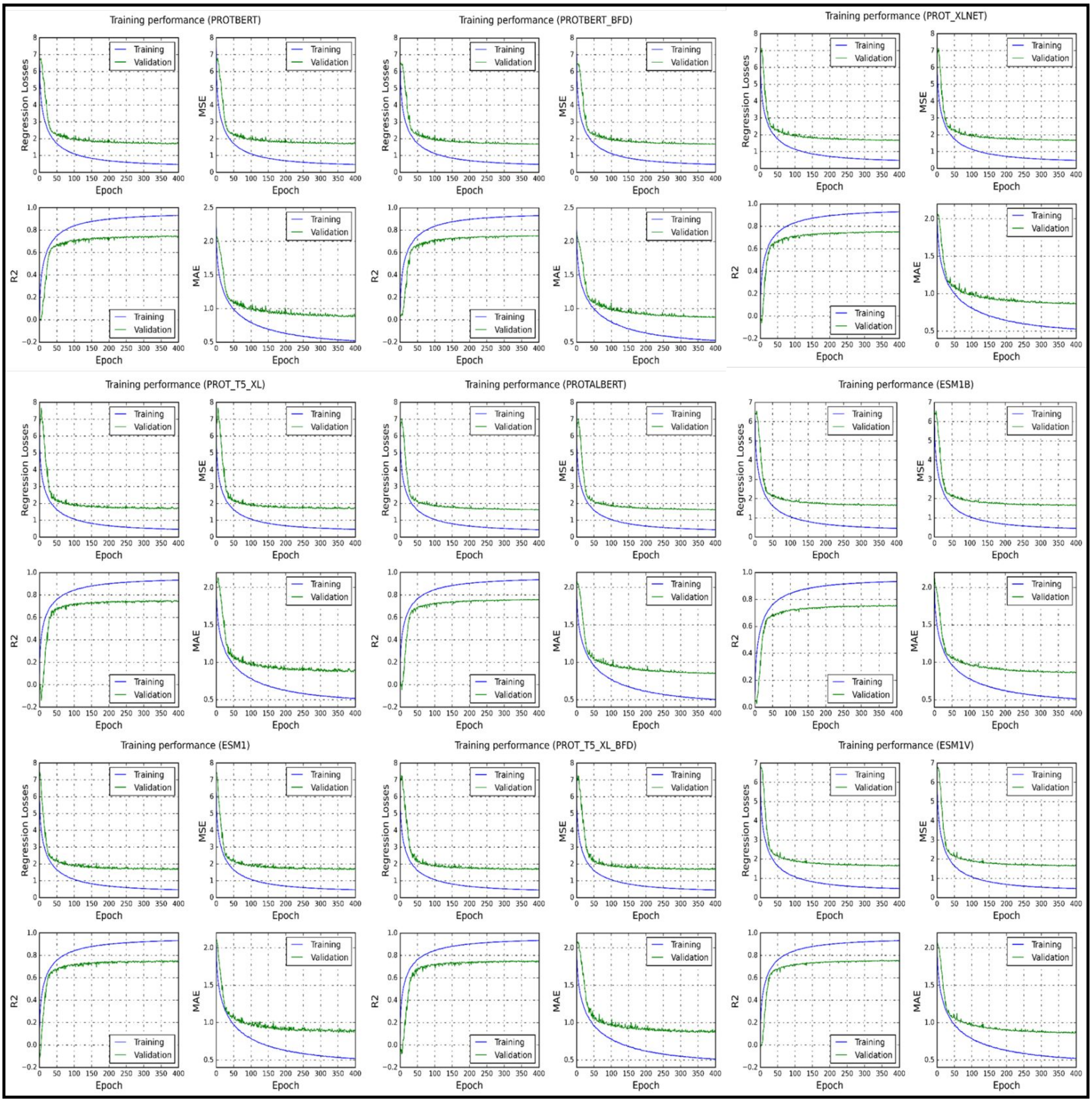
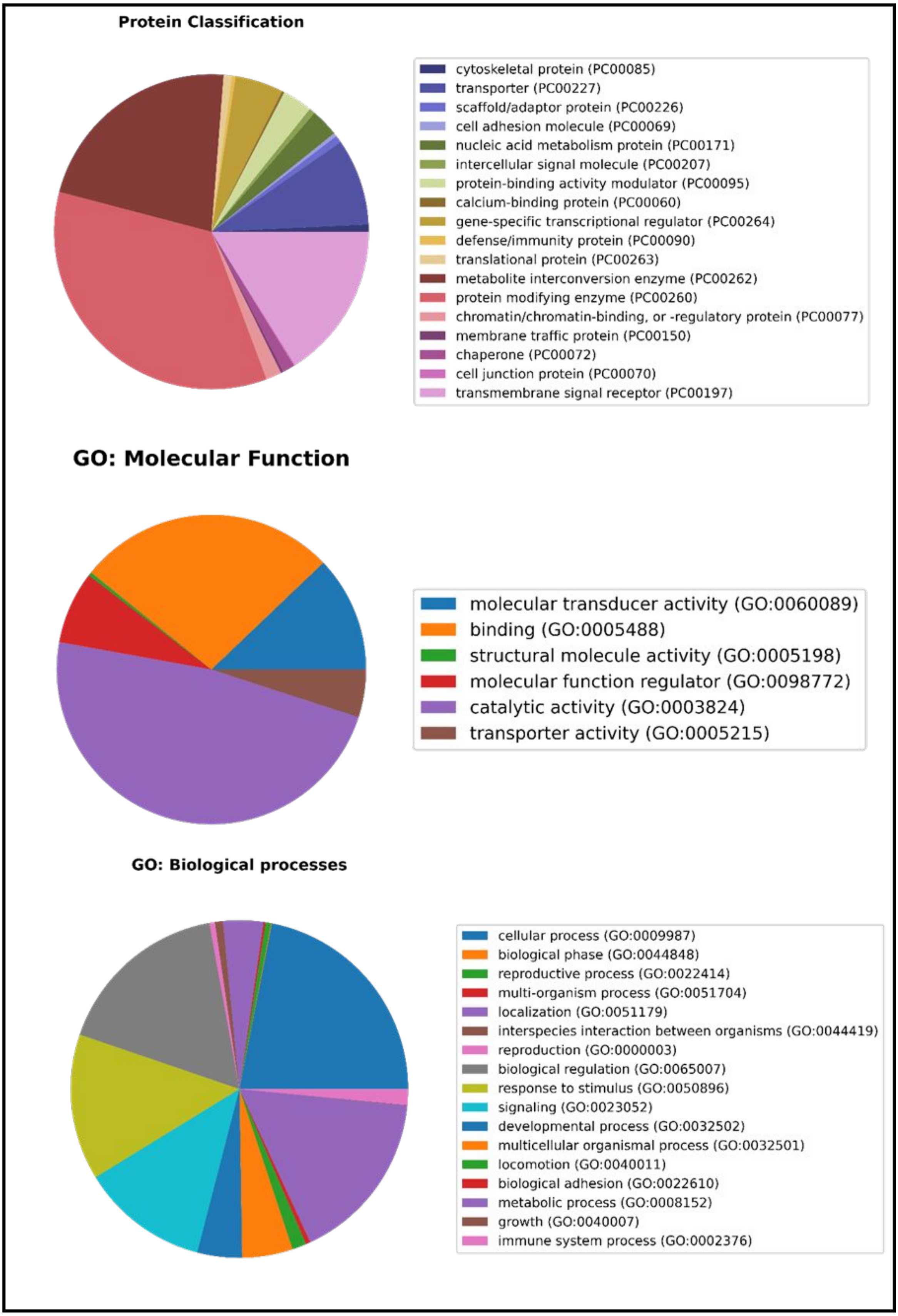


Figure S8: PANTHER enrichment results for a selected group of proteins in the data for classification, molecular function and biological processes


Protein Classification
cytoskeletal protein (PC00085)
transporter (PC00227)
scaffold/adaptor protein (PC00226)
cell adhesion molecule (PC00069)
nucleic acid metabolism protein (PC00171)
intercellular signal molecule (PC00207)
protein-binding activity modulator (PC00095)
calcium-binding protein (PC00060)
gene-specific transcriptional regulator (PC00264)
defense/immunity protein (PC00090)
translational protein (PC00263)
metabolite interconversion enzyme (PC00262)
protein modifying enzyme (PC00260)
chromatin/chromatin-binding, or -regulatory protein (PC00077)
membrane traffic protein (PC00150)
chaperone (PC00072)
cell junction protein (PC00070)
transmembrane signal receptor (PC00197)
GO: Molecular Function
molecular transducer activity (GO:0060089)
binding (GO:0005488)
structural molecule activity (GO:0005198)
molecular function regulator (GO:0098772)
catalytic activity (GO:0003824)
transporter activity (GO:0005215)
GO: Biological processes
cellular process (GO:0009987)
biological phase (GO:0044848)
reproductive process (GO:0022414)
multi-organism process (GO:0051704)
localization (GO:0051179)
interspecies interaction between organisms (GO:0044419)
reproduction (GO:0000003)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
locomotion (GO:0040011)
biological adhesion (GO:0022610)
metabolic process (GO:0008152)
growth (GO:0040007)
immune system process (GO:0002376)