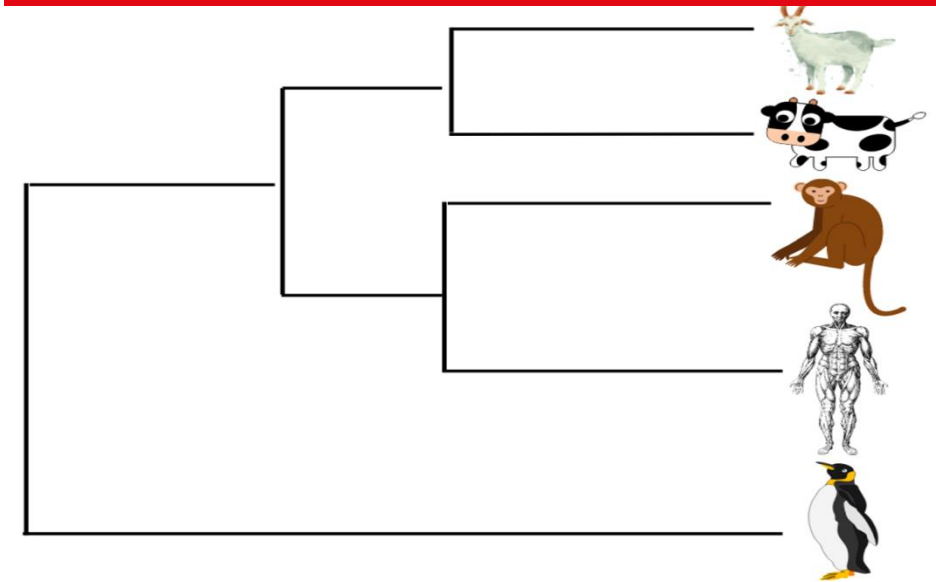
Inferring phylogenetic relationships using MSA Transformer derived embeddings



Marija Zelic

Section of Life Sciences Engineering MA3

Lab Immersion

Bitbol Lab

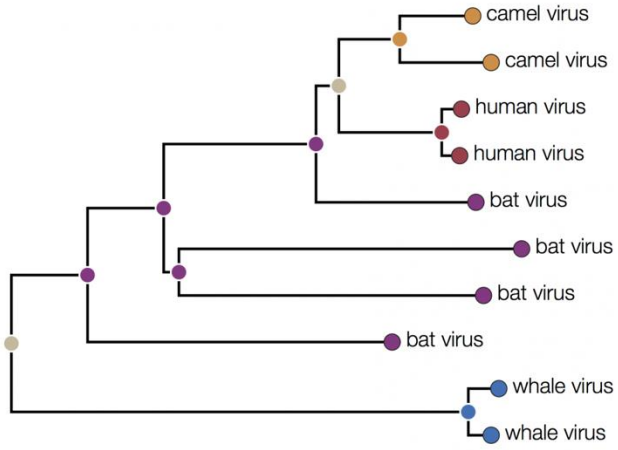
January 2025

Content

- Introduction and Problem Statement
- Generating Synthetic Data
  - bmDCA
  - ESM2
- Models
  - Regression
  - Fully-connected Neural Network
  - Fine-tuning MSA Transformer
- Next Steps
- Conclusion
- Appendix

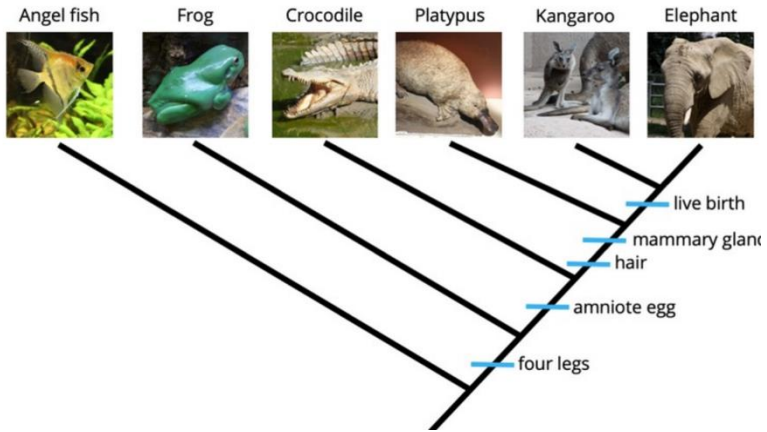
Marija Zelic

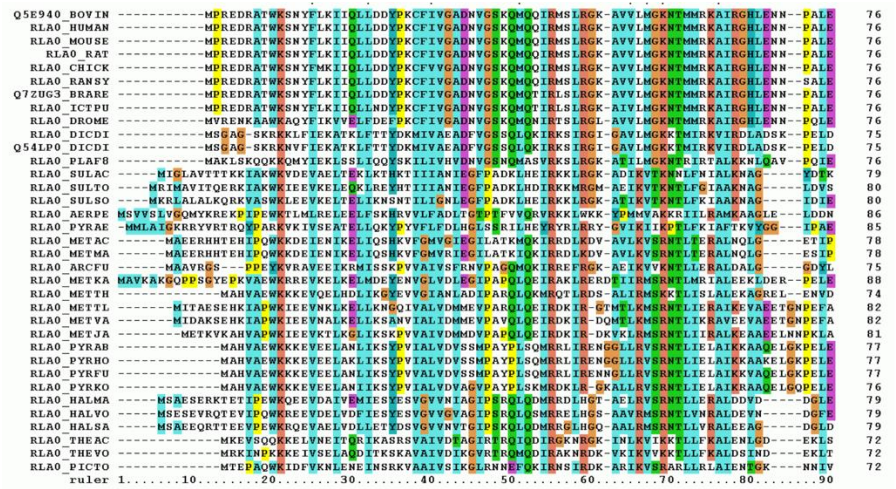
Introduction and Problem Statement



Phylogeny

- Branch of biology that focuses on studying evolution of species and genes.
- **Goal**: Infer **evolutionary tree** (phylogenetic tree).
- Now: mostly based on MSA alignments of homologous sequences.
- Methods:
  - distance-based methods
  - character-based methods
- Deep learning?





Starting Point

Article | Open access | Published: 22 October 2022

Protein language models trained on multiple sequence alignments learn phylogenetic relationships

Umberto Lupo, Damiano Sgarbossa & Anne-Florence Bitbol

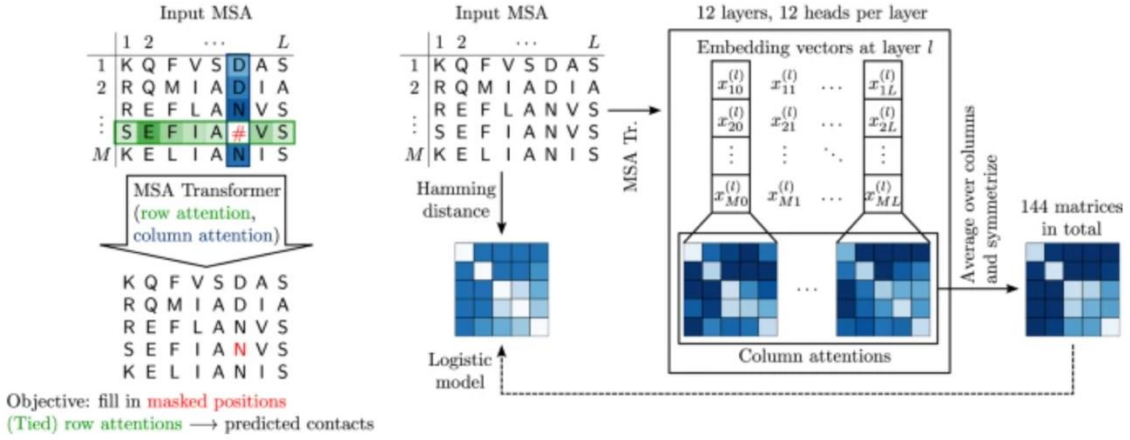
Nature Communications **13**, Article number: 6298 (2022) | Cite this article

14k Accesses | **26** Altmetric | Metrics

- "Simple and universal combinations of MSA Transformer's column attentions strongly correlate with Hamming distances between sequences in MSAs."

Starting Point

- "Simple and universal combination of MSA Transformer's column attentions…"



Visualization taken from https://doi.org/10.1038/s41467-022-34032-y

- "... strongly correlate with Hamming distances between MSAs."

AGATC
AGGCA

Hamming distance counts sites that differ between two sequences.

Why and How to go beyond this?

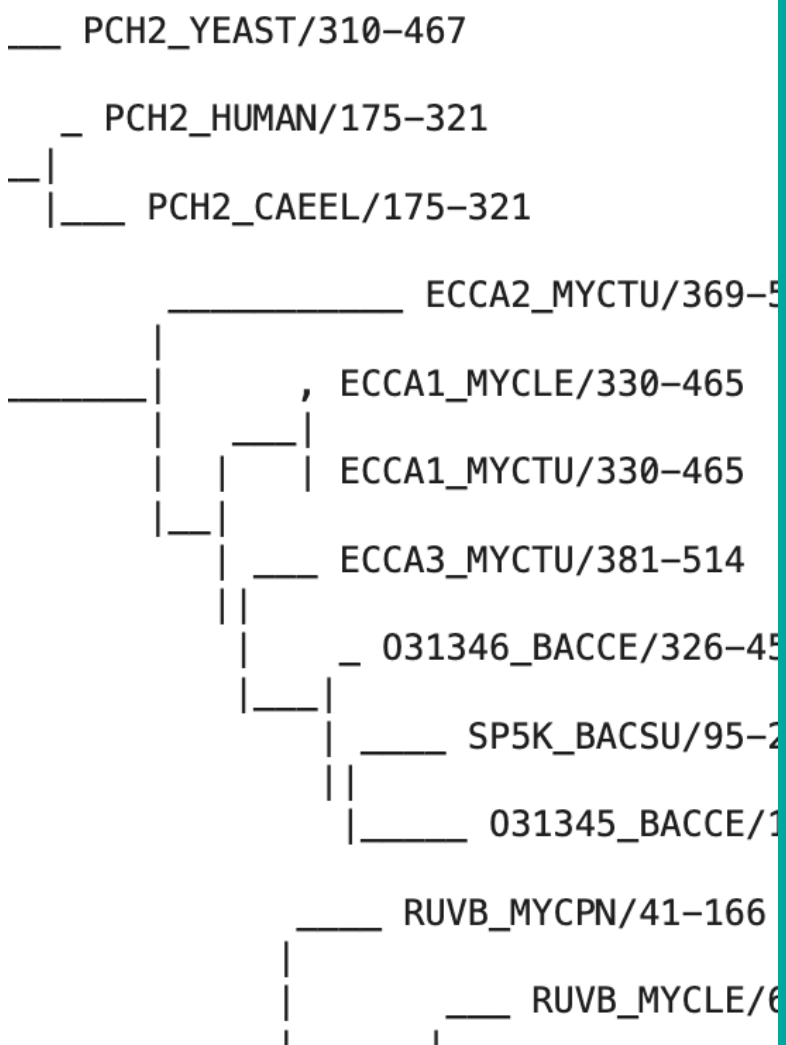
- Regression is performed on MSAs from 15 protein families (both individually and jointly).
- R^2 captured is very high.
- We wonder if we can go beyond this.

Family	R^2
PF00004	0.97
PF00005	0.99
PF00041	0.98
PF00072	0.99
PF00076	0.98
PF00096	0.94

Snapshot from https://doi.org/10.1038/s41467-022-34032-y

- **Why?**
  - Our 'ground truth' distances for regression are coming from Hamming distances – **but they are very simple proxy for phylogenetic relationship**.
  - Without the **true** phylogenetic tree, we lack an accurate 'ground truth' for distances.
- **How?**
  - Generate synthetic sequences along the existing (known) tree and use *patristic distances* derived from the tree to analyze if MSA Transformer based embeddings can capture them.

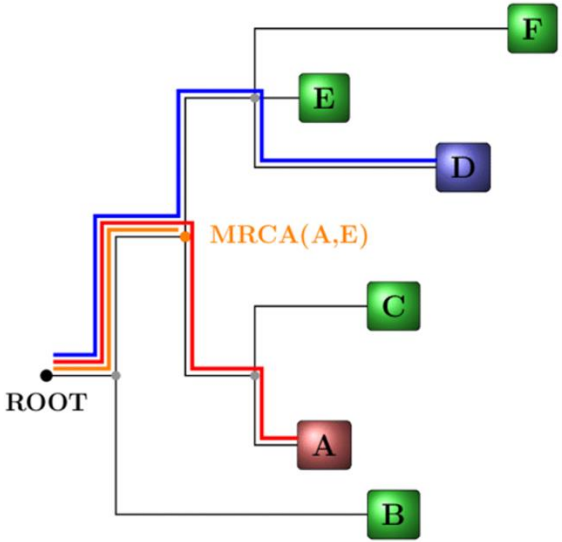
Marija Zelic



Generating Synthetic Data

Patristic Distance

- *Patristic distance* is the sum of the lengths of the branches that link two nodes in a phylogenetic tree, where those nodes are typically leaf nodes.



Tree distance

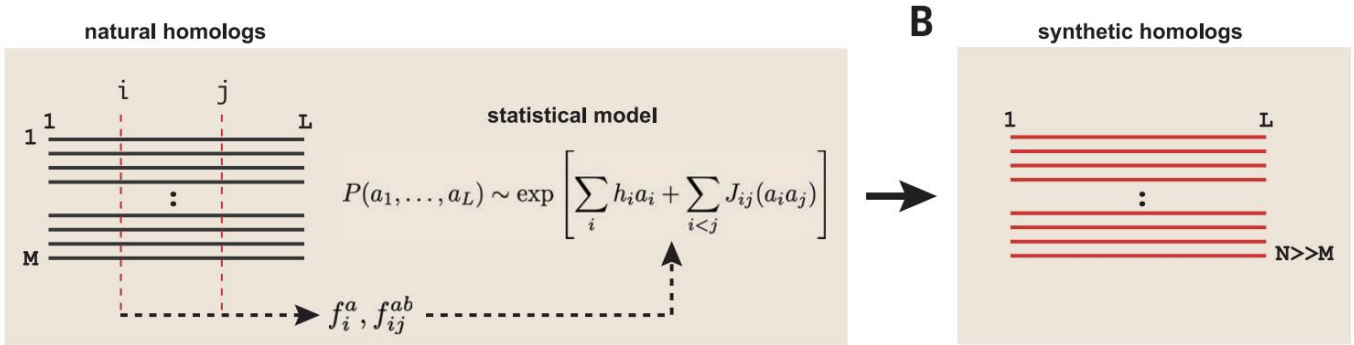
$$d_{\mathrm{AD}}^{\mathrm{Tree}} = c_{\mathrm{AA}} + c_{\mathrm{DD}} - 2c_{\mathrm{AE}}$$

- c_{AA} length of path connecting ROOT and leaf A
- c_{DD} length of path connecting ROOT and leaf D
- c_{AD} length of path connecting ROOT and MRCA(A,D), the More Recent Common Ancestor of A and D

Image taken from 19. 10.1186/s12859-018-2033-5.

Generating Synthetic Sequences - bmDCA

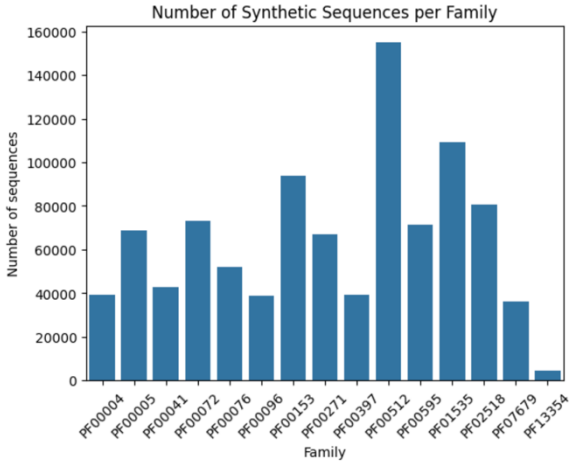
- Generosity of the Lab. Generated by using two different approaches: **bmDCA** and **ESM2 Transformer**.
- **Boltzmann Machine Direct Coupling Analysis (bmDCA):**
  1. Generates a phylogenetic tree from a natural MSA using either FastTree or IQTree.
  2. Inferrs statistical model with bmDCA method.
  3. Uses Monte Carlo (MC) sampling from the model to generate synthetic sequences.

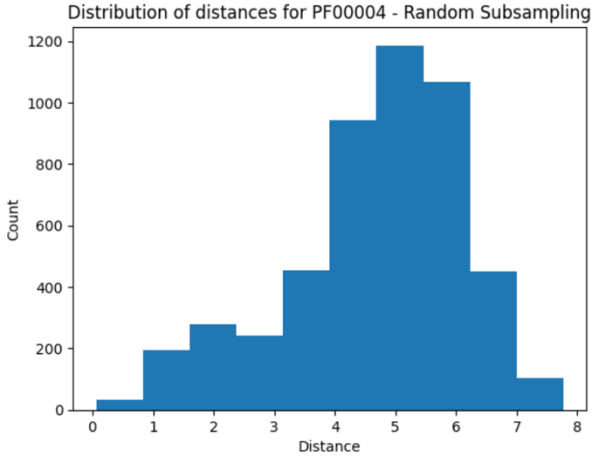


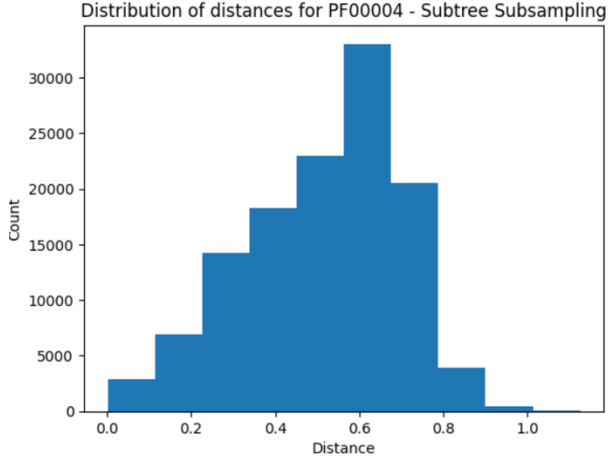
Generating Synthetic Sequences - bmDCA

Marija Zelic

- Varying number of synthetic sequences across families.
- Two different approaches for *subsampling:*
  - ***Randomly*** (100)
  - ***Subtree*** (~500)

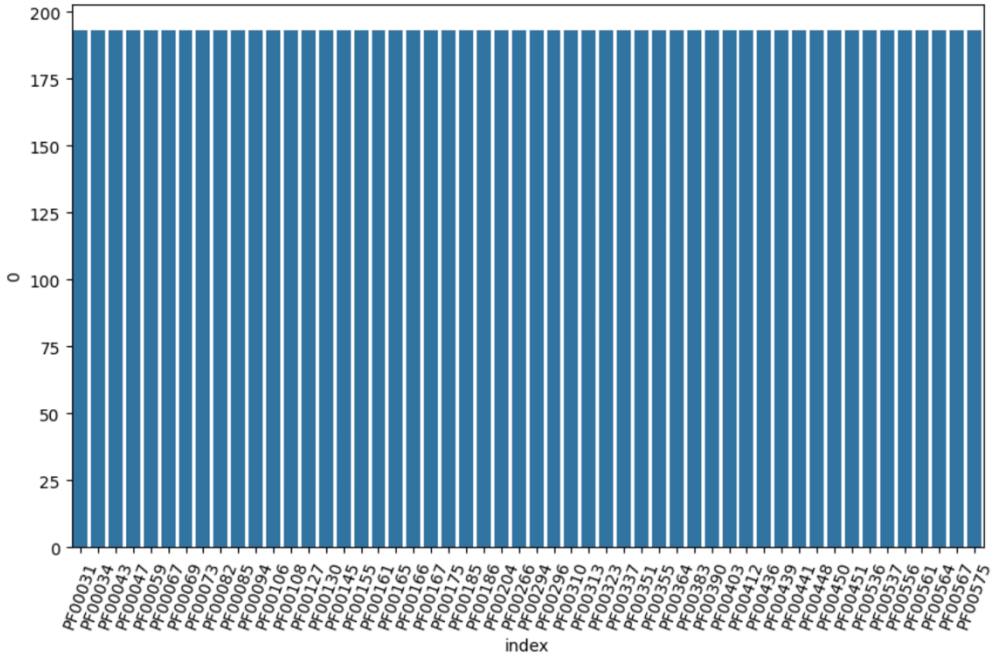






Generating Synthetic Sequences – ESM Transformer

- **Evolutionary Scale Modeling Transformer:**
  1. Generates a phylogenetic tree from a natural MSA using either FastTree or IQTree.
  2. Produces synthetic MSA through a Metropolis-Hastings algorithm for Markov Chain Monte Carlo (MCMC) sampling employing probabilities given by ESM2 model.
- Total dataset: over 10k small trees (~50 sequences) coming from over 20 different families.



Models

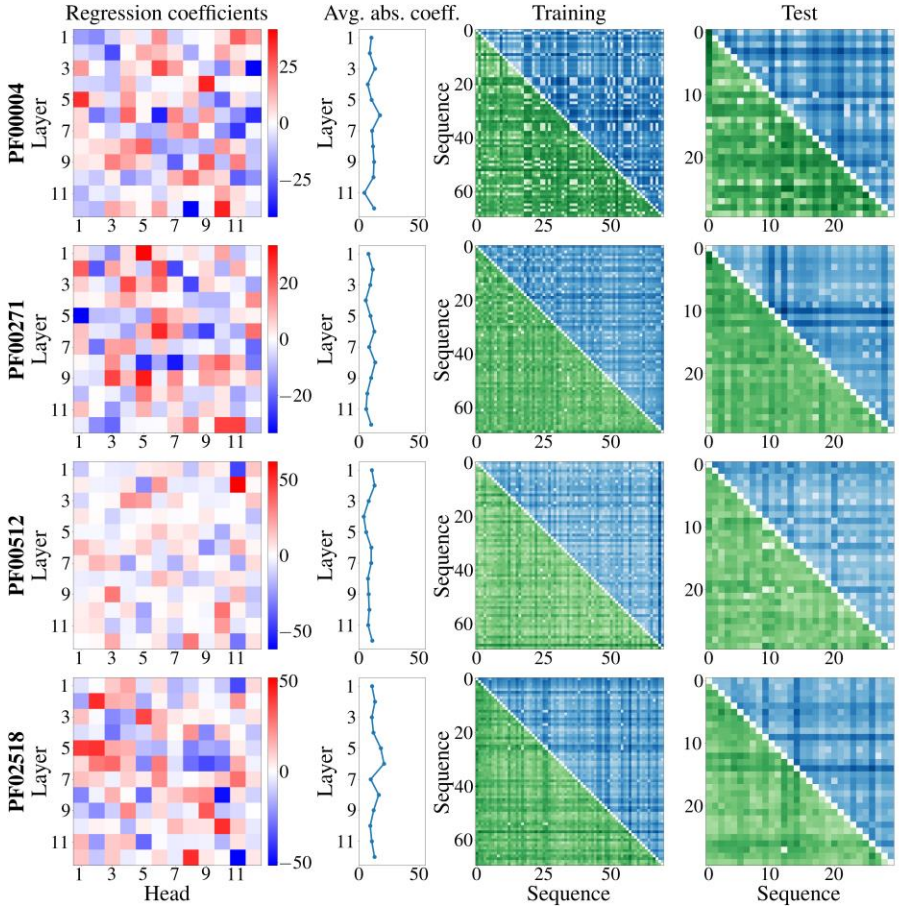
Regression

- **Idea**: Recap the results from the Starting Point paper but now using Synthetic Sequences and Patristic Distances.
- Total of 4 combinations depending on *subsampling approach* and *whether we fit one regression per family or for all families combined.*
- Per family: Train/test ratio 70:30.
- Combined families: Families PF02518, PF07679 and PF13345 are left out for testing.
- Distances are **normalized**!

Combination 1: Random Subsampling & Regression per Family	**Combination 2:** Random Subsampling & Regression for Combined Familes
Combination 3: Subtree Subsampling & Regression per Family	**Combination 4:** Subtree Subsampling & Regression for Combined Families

Marija Zelic

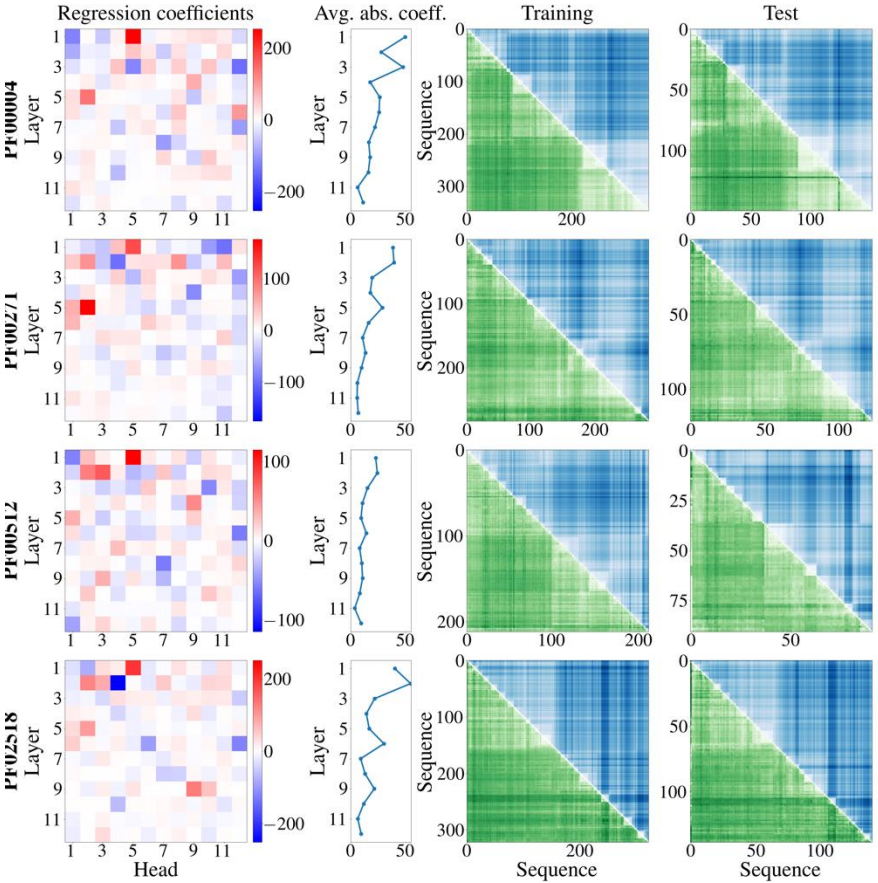
Random Subsampling & Regression per Family



	Depth	mean (training)	mean (test)	std (training)	std (test)	RMSE (training)	RMSE (test)	MAE (training)	MAE (test)	R^2 (test)	Pearson (test)	Slope (test)
PF00004	100.0	0.585225	0.542525	0.218867	0.210729	0.081116	0.151500	0.059222	0.106144	0.483133	0.744935	0.741293
PF00005	100.0	0.625269	0.605651	0.166095	0.191008	0.087620	0.140096	0.067128	0.109203	0.462042	0.719502	0.687343
PF00041	100.0	0.581015	0.568382	0.142442	0.198819	0.068240	0.116044	0.052843	0.087846	0.659332	0.816192	0.667526
PF00072	100.0	0.609785	0.599449	0.153634	0.182447	0.080631	0.101521	0.061552	0.080477	0.690371	0.839493	0.792141
PF00076	100.0	0.575087	0.551773	0.150752	0.183231	0.077602	0.120108	0.059961	0.093720	0.570320	0.772426	0.707614
PF00096	100.0	0.393803	0.483795	0.157811	0.200784	0.086778	0.138830	0.068382	0.105303	0.521907	0.741466	0.629536
PF00153	100.0	0.531441	0.590439	0.134826	0.191154	0.062278	0.120892	0.048104	0.089589	0.600030	0.793648	0.677088
PF00271	100.0	0.530376	0.507762	0.150247	0.185976	0.071963	0.128697	0.055519	0.101159	0.521128	0.732082	0.619037
PF00397	100.0	0.367436	0.421765	0.164472	0.196317	0.071745	0.133782	0.054099	0.098684	0.535616	0.747401	0.658835
PF00512	100.0	0.456902	0.434617	0.150521	0.160272	0.080767	0.122591	0.061769	0.094277	0.414935	0.687968	0.628356
PF00595	100.0	0.599864	0.579627	0.161139	0.194507	0.063323	0.110245	0.048931	0.084263	0.678748	0.827182	0.738576
PF01535	100.0	0.557785	0.513592	0.157965	0.178738	0.092003	0.117825	0.073733	0.087748	0.565450	0.772614	0.733256
PF02518	100.0	0.529058	0.496693	0.154671	0.175698	0.077469	0.120317	0.059637	0.095459	0.531058	0.753795	0.669750
PF07679	100.0	0.627187	0.554862	0.142469	0.172145	0.064371	0.112570	0.048663	0.086105	0.572385	0.791145	0.805002
PF13354	100.0	0.549692	0.516323	0.194861	0.203312	0.072949	0.110924	0.057332	0.086451	0.702339	0.847584	0.797117

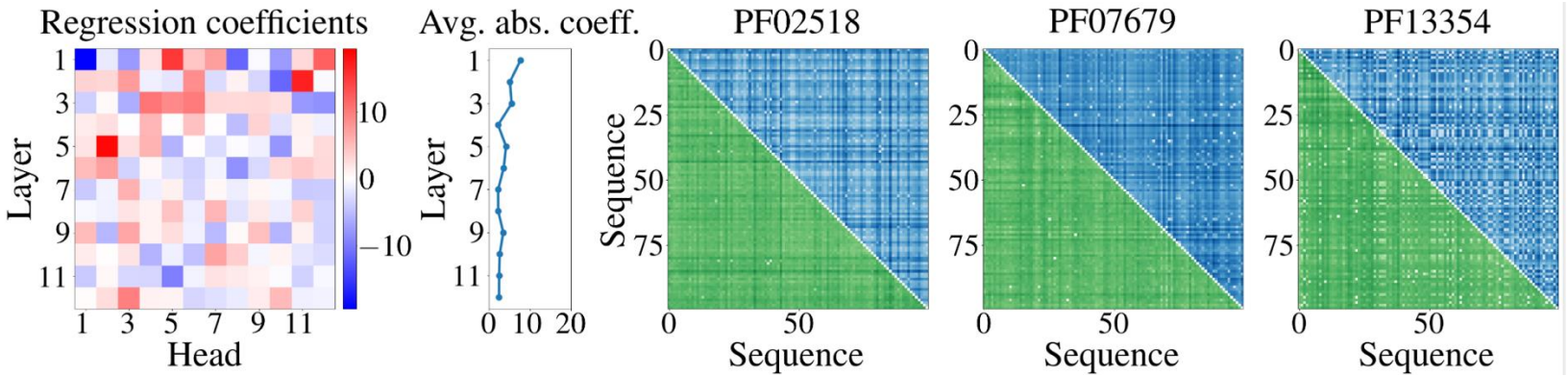
Marija Zelic

Subtree Subsampling & Regression per Family



	Depth	mean (training)	mean (test)	std (training)	std (test)	RMSE (training)	RMSE (test)	MAE (training)	MAE (test)	R^2 (test)	Pearson (test)	Slope (test)
PF00004	497.0	0.461581	0.434894	0.166836	0.169105	0.055506	0.065033	0.042938	0.049872	0.852106	0.925919	0.909082
PF00005	496.0	0.543482	0.530608	0.133622	0.144489	0.073968	0.072922	0.058513	0.058011	0.745287	0.866954	0.764818
PF00041	444.0	0.530650	0.542943	0.177000	0.194658	0.086737	0.097993	0.066563	0.074759	0.746575	0.867057	0.716773
PF00072	420.0	0.532032	0.511381	0.204912	0.197377	0.075022	0.081273	0.058317	0.063167	0.830451	0.912194	0.846003
PF00076	417.0	0.437901	0.429270	0.211744	0.213320	0.073420	0.077375	0.054478	0.057016	0.868434	0.932442	0.894402
PF00096	452.0	0.394309	0.385203	0.139154	0.129786	0.084329	0.081824	0.066025	0.063237	0.602524	0.782678	0.666935
PF00153	499.0	0.376404	0.351045	0.244250	0.241863	0.077167	0.093393	0.054989	0.060564	0.850895	0.922571	0.858903
PF00271	406.0	0.441302	0.408974	0.145749	0.142160	0.066045	0.071201	0.051340	0.055551	0.749144	0.868226	0.795934
PF00397	478.0	0.230729	0.218145	0.296291	0.285276	0.067140	0.080476	0.036820	0.042423	0.920420	0.960603	0.961499
PF00512	303.0	0.436765	0.427348	0.148949	0.159675	0.073402	0.110479	0.057649	0.078938	0.521278	0.731757	0.586441
PF00595	427.0	0.319019	0.338447	0.229491	0.232058	0.032384	0.040554	0.022235	0.027933	0.969459	0.984845	0.959082
PF01535	431.0	0.500402	0.513881	0.161561	0.181300	0.095205	0.101070	0.075656	0.080200	0.689225	0.846261	0.650155
PF02518	461.0	0.514261	0.504423	0.156052	0.156252	0.073558	0.083271	0.057804	0.061346	0.715990	0.852957	0.786619
PF07679	480.0	0.460923	0.463548	0.167969	0.167055	0.090752	0.099761	0.070855	0.076129	0.643385	0.807322	0.712878
PF13354	455.0	0.757360	0.751609	0.107783	0.135922	0.028077	0.029757	0.021662	0.022770	0.952070	0.976102	0.960502

Random Subsampling & Regression for Combined Families

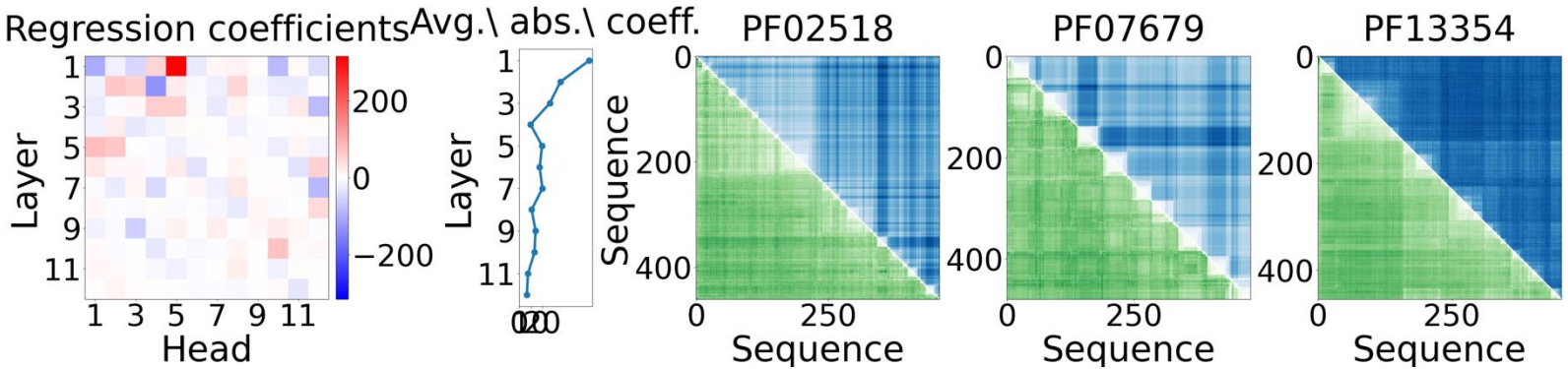


	Depth	RMSE	Std	Pearson	Slope	R^2
PF02518	100.0	0.113831	0.145782	0.632813	1.011546	0.390304
PF07679	100.0	0.113215	0.130254	0.740202	1.028508	0.244526
PF13354	100.0	0.127507	0.184766	0.732667	1.182788	0.523758

- Substantial Pearson correlation coefficient, but also lower R^2 metric compared to the results reported in [1] where the average value was around 0.6.

Marija Zelic

Subtree Subsampling & Regression for Combined Families

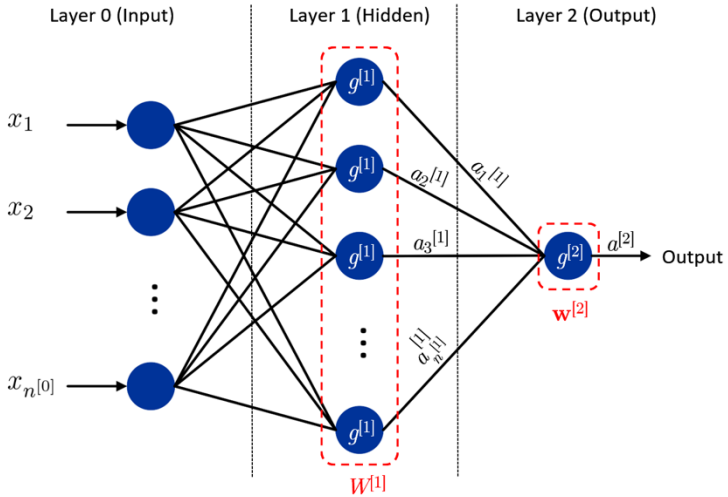


	Depth	RMSE	Std	Pearson	Slope	R^2
PF02518	461.0	0.130856	0.153137	0.785327	1.002512	0.269827
PF07679	480.0	0.125657	0.165138	0.675257	0.863871	0.421001
PF13354	455.0	0.344328	0.105648	0.832437	0.710768	-9.622452

- Training and testing approach not suitable?
- Can we do better?

Fully-connected Neural Network

Input:144 coefficients derived from MSA Transformer's column attentions characterizing one pair of sequences

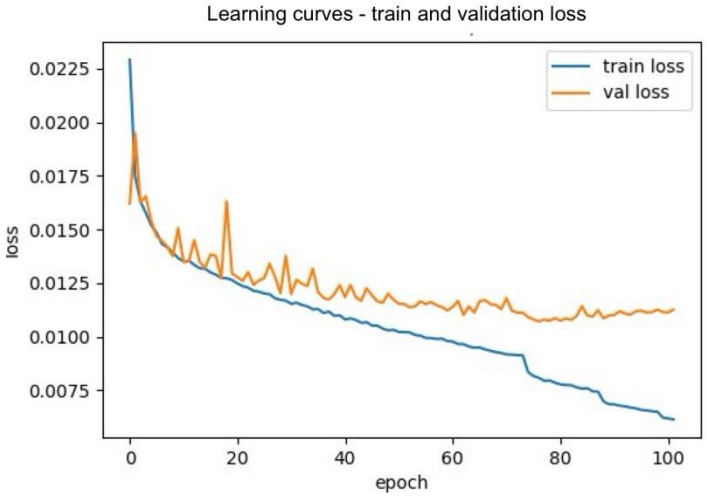


Output: Predicted Patristic Distance for that pair of sequences

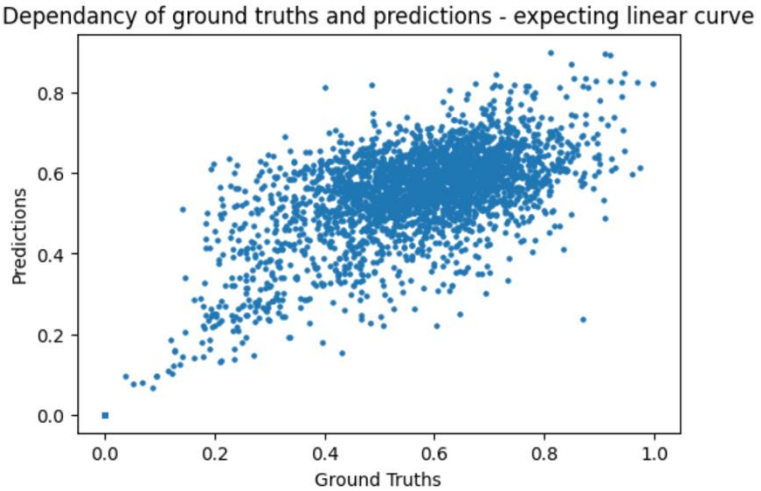
- Dataset: Random and Subtree subsampled synthetic MSA sequences.
- 80% of every family is kept for training, while 20% is used for testing.
- 5-fold cross-validation with hyperparameter tuning (network architecture, learning rate, etc.).
- MSE Loss.
- Issues with overfitting: Early stopping and Learning rate scheduler to prevent it.

Fully-connected Neural Network – Random Subsampling

Marija Zelic



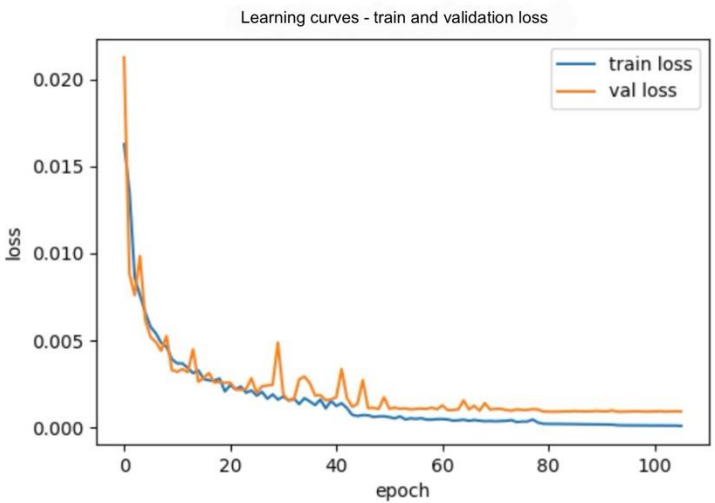
Final MSE loss on test data: 0.0146.

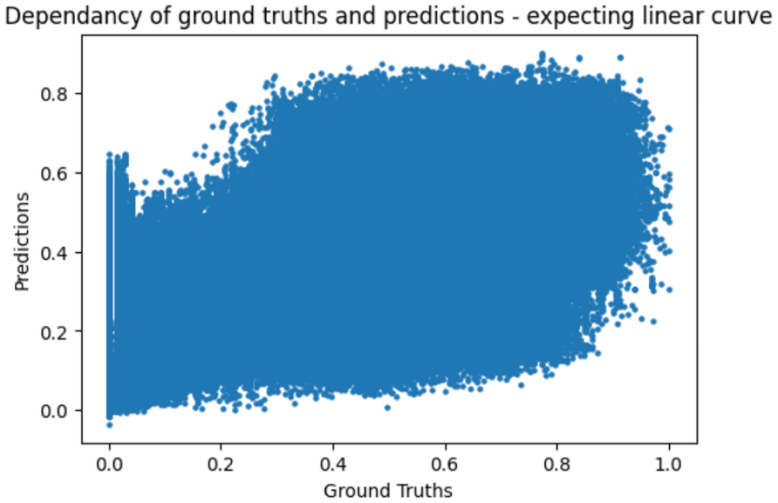


R^2 captured on test data: 0.7.

Marija Zelic

Fully-connected Neural Network – Subtree Subsampling



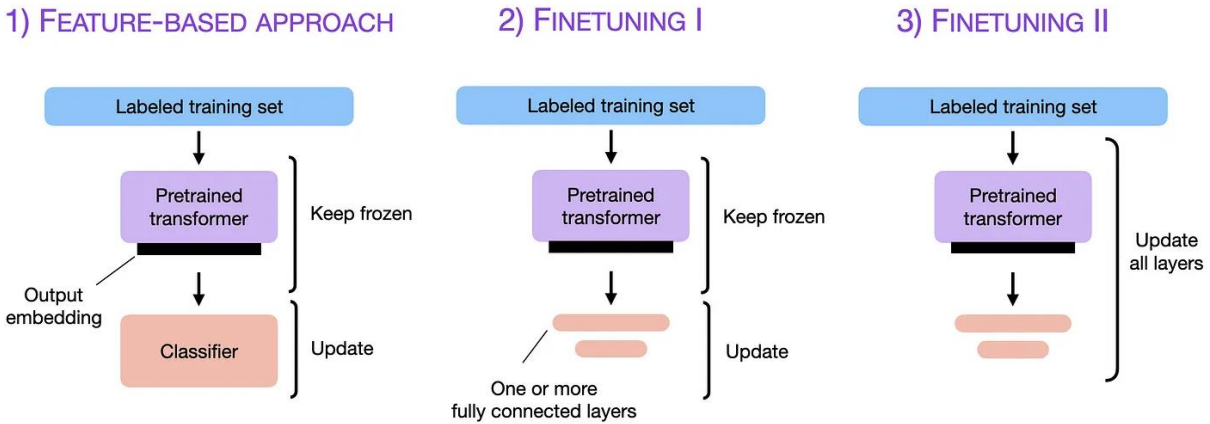


Final MSE loss on test data: 0.0318.

R^2 captured on test data: 0.36.

Fine-tuning MSA Transformer with bmDCA generated data

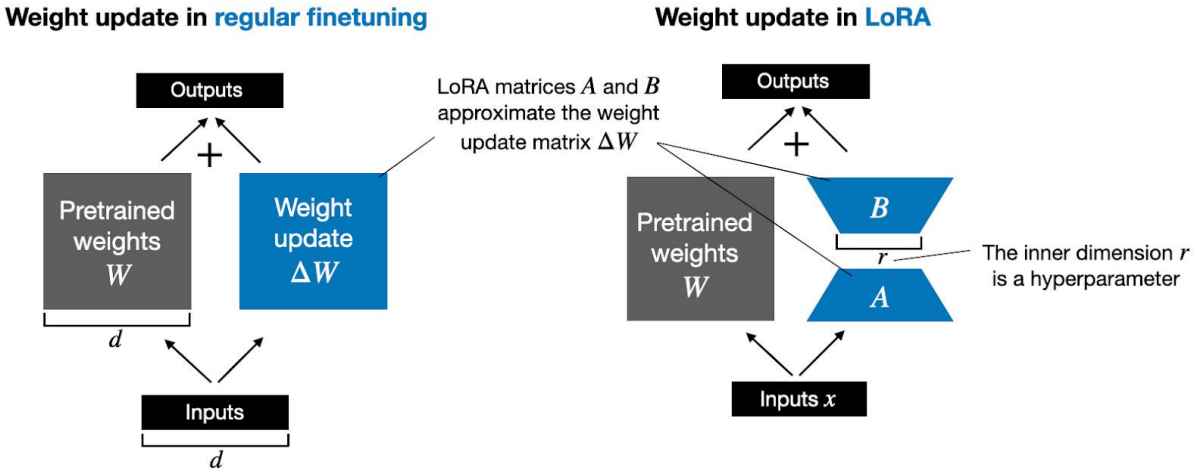
- **Fine-tuning** (in realm of (protein) language models): adding a task specific head, if necessary, and updating the weights of network through backpropagation during the training process.
- MSA Transformer: 100M parameters, trained on 4.3TB of data.



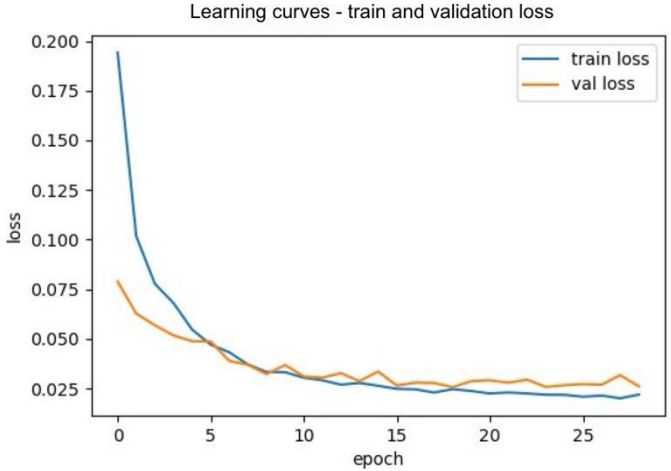
The 3 conventional feature-based and finetuning approaches.

LoRA (Low Rank Adaptation)

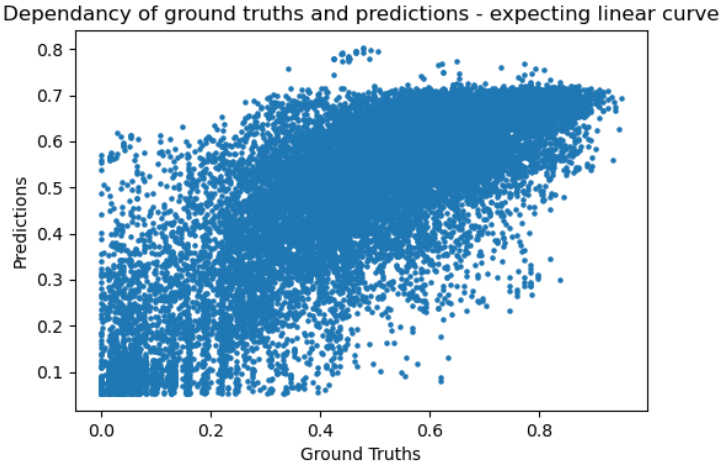
- **LoRA**: instead of fine-tuning all the weights that constitute the weight matrix of pre-trained LM, two smaller matrices that approximate this large matrix are fine-tuned.
- We only apply LoRA to weight matrices corresponding to column attentions' weights (key, query, value matrices in each layer and head).
- Opted for $r = 16$. Reduced the number of trainable parameters to ~1.4M.



Fine-tuning MSA Transformer with bmDCA generated data (Subtree approach) - Results



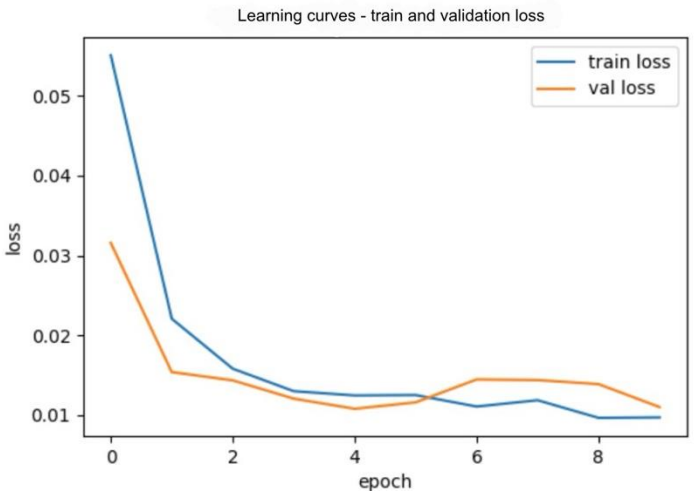
Final MSE loss on test data: 0.0222.



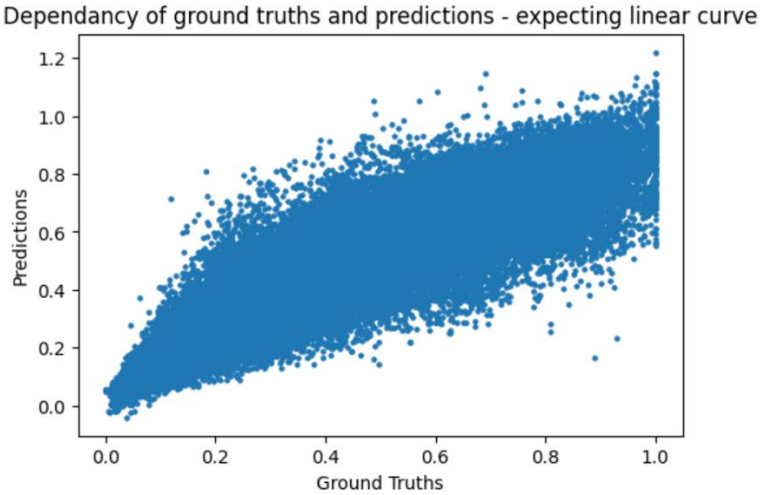
R^2 captured on test data: 0.59.

- Same train/test split approach as previously.
- New: Custom batching procedure.
- Non-exhaustive grid search on hyperparameters (architecture of custom head and learning rate).
- MSE Loss.
- Early stopping and Learning rate scheduler to prevent overfitting.

Fine-tuning MSA Transformer with ESM2 generated data



Final MSE loss on test data: 0.0099.



R^2 captured on test data: 0.79.

- Every tree is one batch.
- Non-exhaustive grid search on hyperparameters (architecture of custom head and learning rate).
- MSE Loss.
- Early stopping and Learning rate scheduler to prevent overfitting.

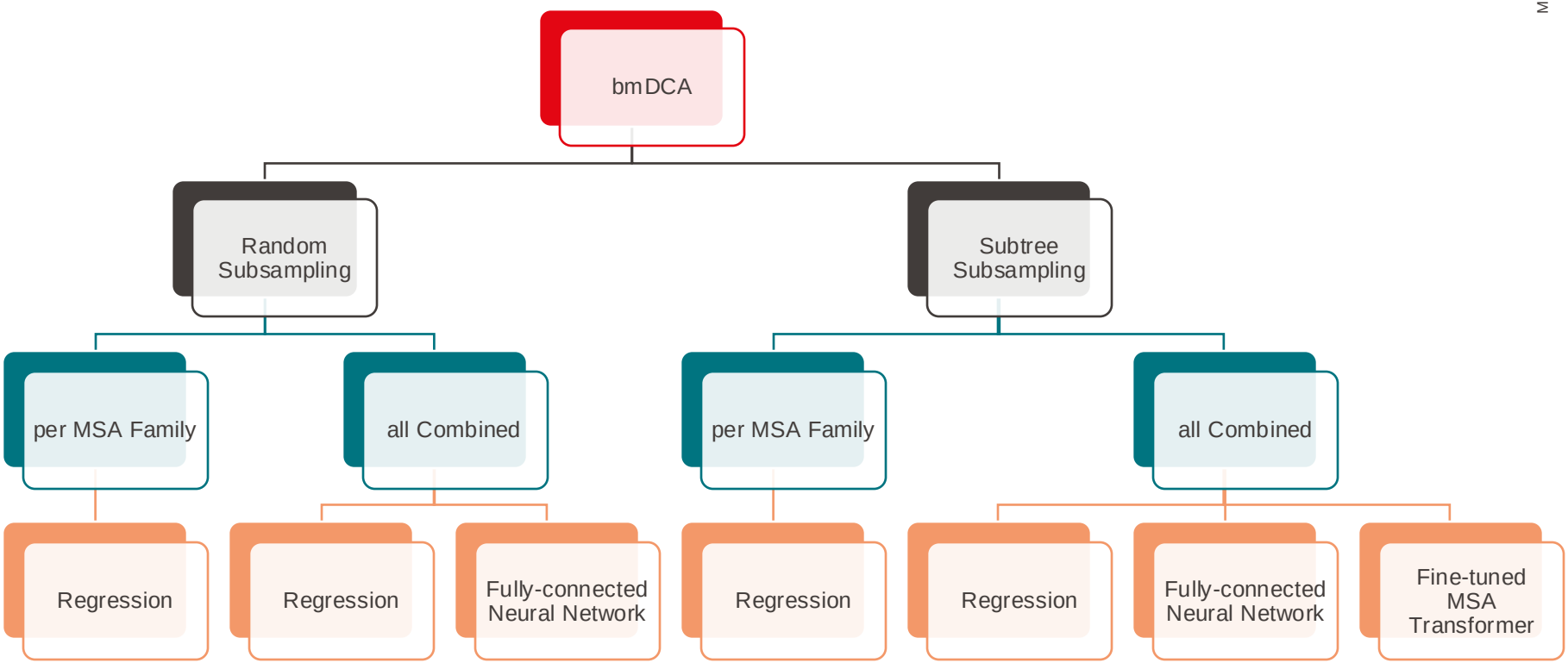
Next Steps

If we still had time…

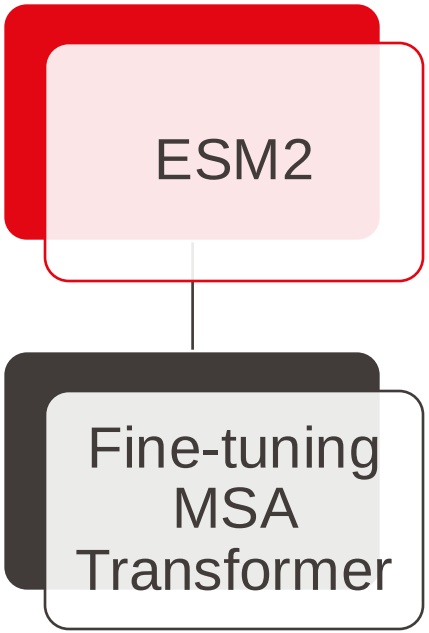
1. LoRA fine-tuning applied only on weights of column attentions.
2. More exhaustive search on hyperparameters and cross-validation.
3. Models' performance is evaluated on the synthetic data generated using the same approach (bmDCA or ESM2) as the fine-tuning data. Interesting to investigate whether trained models can generalize to synthetic data generated by different method.
4. Disentangle whether performance of model fine-tuned on ESM2 data is due to data quantity or inherent difference from data synthesized by bmDCA approach.
5. Understand performance of fine-tuned models across different families. Simpson's paradox?
6. bmDCA fine-tuned model has only less distant sequences. Comparable performance?
7. Natural sequences?

Conclusion

bmDCA


bmDCA
Random Subsampling
Subtree Subsampling
per MSA Family
all Combined
per MSA Family
all Combined
Regression
Regression
Fully-connected Neural Network
Regression
Regression
Fully-connected Neural Network
Fine-tuned MSA Transformer

ESM2



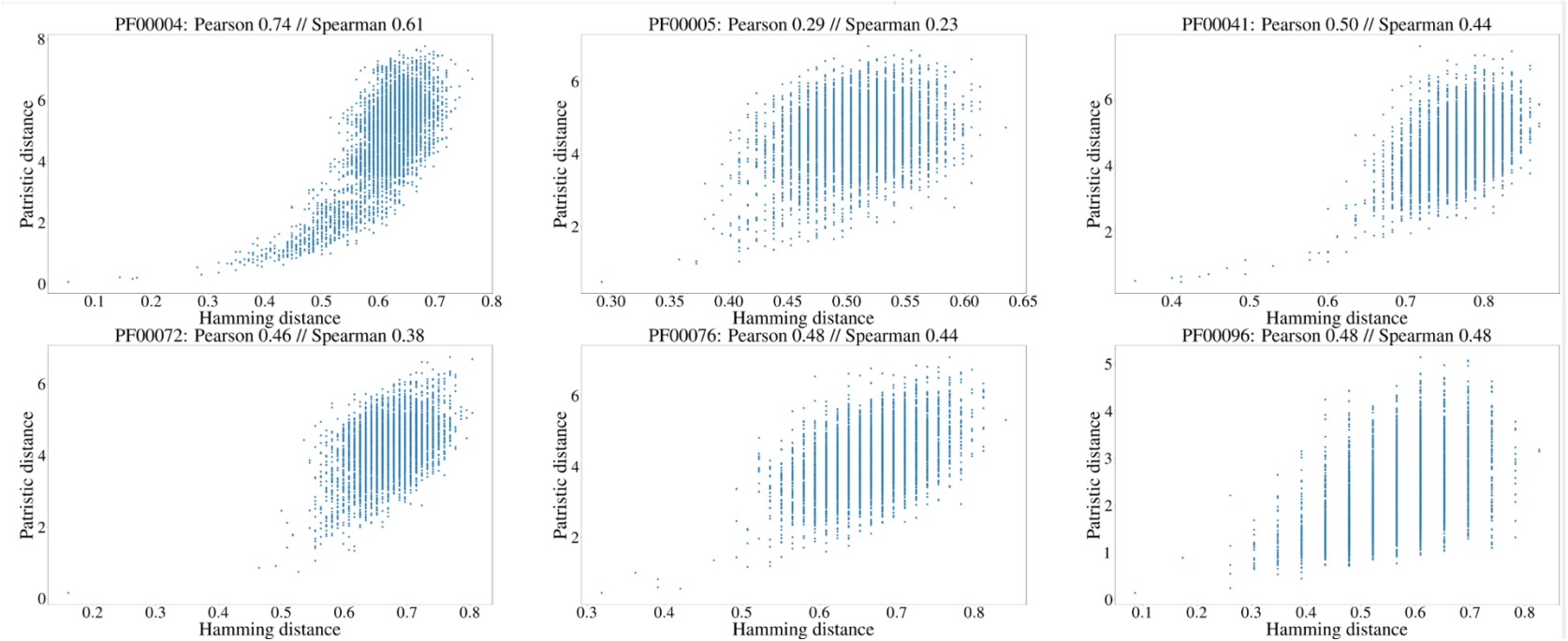
References

[1] Lupo, U., Sgarbossa, D. & Bitbol, AF. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nat Commun* **13**, 6298 (2022). https://doi.org/10.1038/s41467-022-34032-y

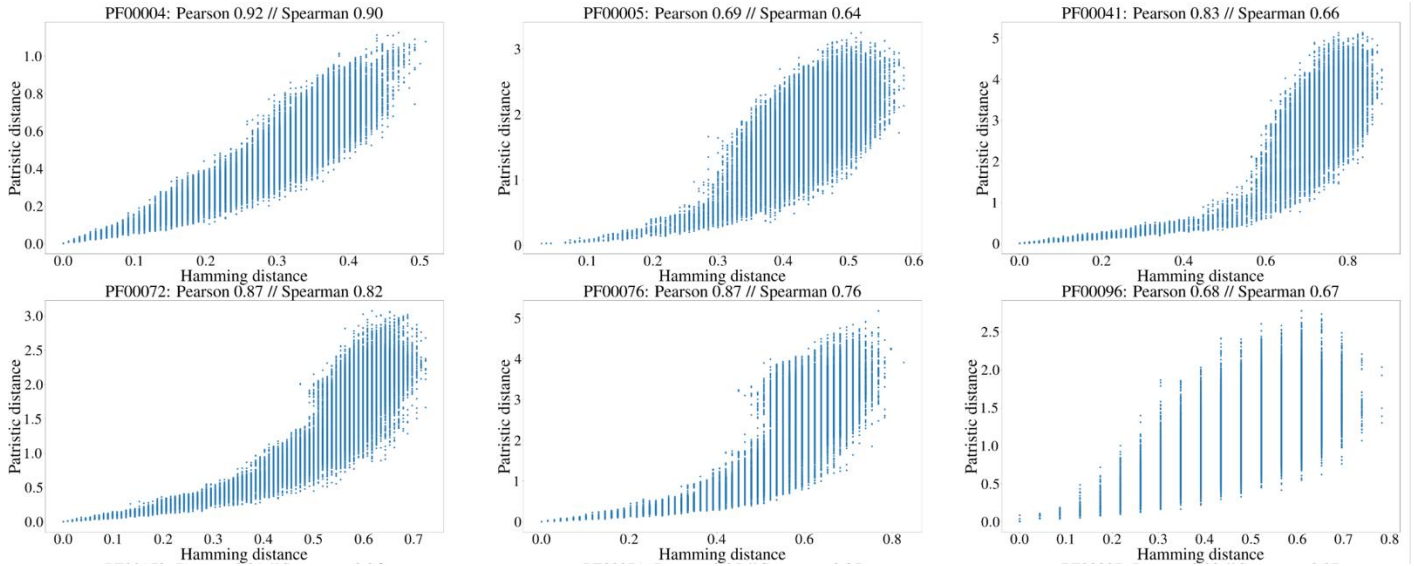
Thank you for your attention! :)
You can check out the code on:
Fine-tuning-MSA-Transformer

Appendix

Pearson and Spearman Correlation Coefficient for Random Subsampling between Patristic and Hamming distance (some families)



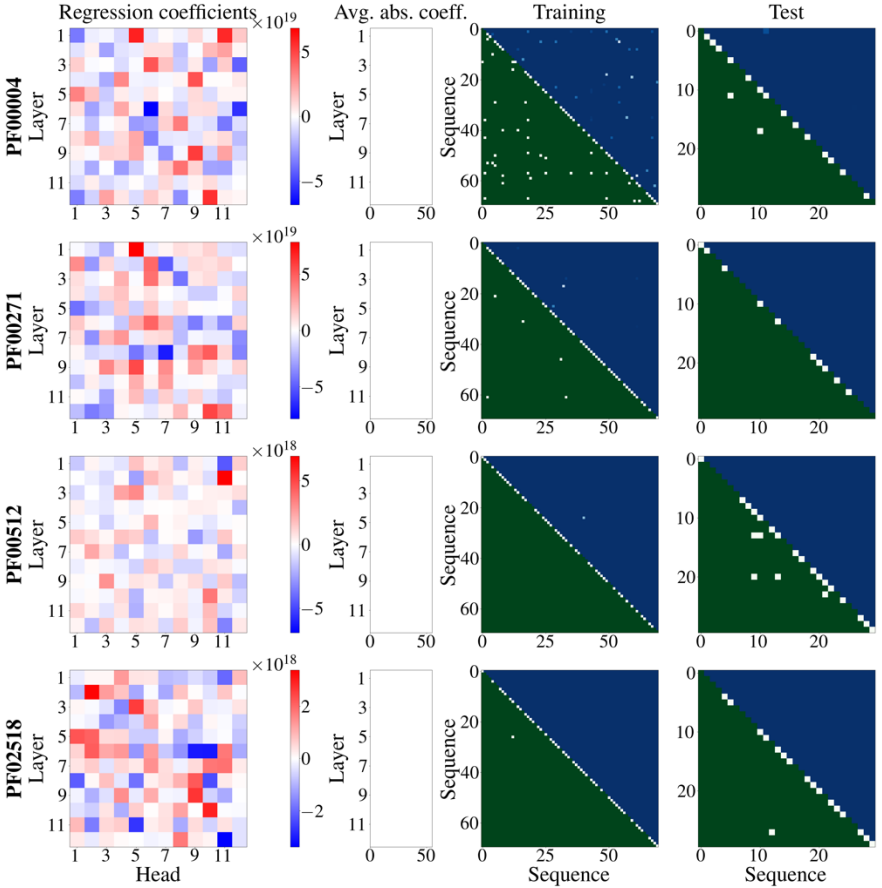
Pearson and Spearman Correlation Coefficient for Subtree Subsampling between Patristic and Hamming distance (some families)



- Noticeably higher values of Pearson and Spearman correlation coefficient between Patristic and Hamming distance in case of Subtree sampling.
- **Possible interpretation**: Hamming distance is better at capturing phlogenetic relationship between sequences that are less distant, then the once that are more apart.

Marija Zelic

Random Subsampling & Regression per Family (without normalization)

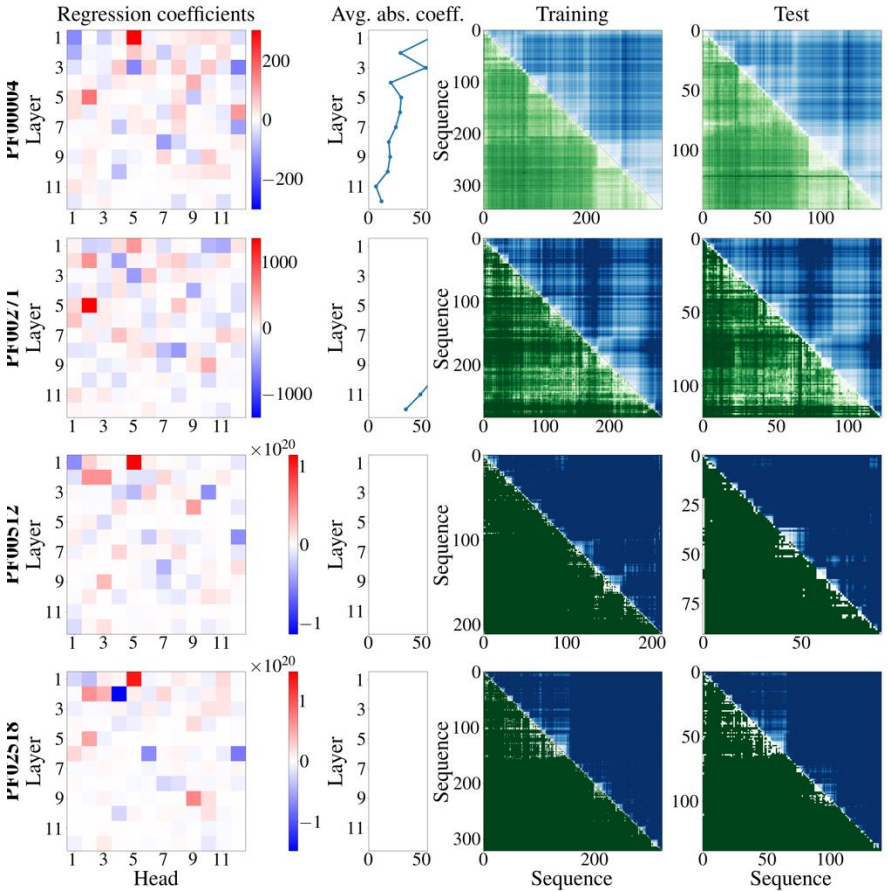


	Depth	mean (training)	mean (test)	std (training)	std (test)	RMSE (training)	RMSE (test)	MAE (training)	MAE (test)	R^2 (test)	Pearson (test)	Slope (test)
PF00004	100.0	4.545982	4.214292	1.700138	1.636923	3.931997	3.603751	3.602294	3.318041	-3.846779	0.468837	0.052206
PF00005	100.0	4.360504	4.223693	1.158315	1.332055	3.551809	3.482199	3.397167	3.311866	-5.833810	0.654455	0.097263
PF00041	100.0	4.420353	4.324235	1.083694	1.512611	3.585092	3.646303	3.457968	3.410674	-4.810997	0.606062	0.081289
PF00072	100.0	4.115994	4.046225	1.037018	1.231501	3.280956	3.279844	3.152131	3.136548	-6.093114	0.659323	0.103276
PF00076	100.0	4.076254	3.911000	1.068541	1.298749	3.253274	3.180150	3.111594	2.992720	-4.995768	0.671076	0.109700
PF00096	100.0	2.024900	2.487630	0.811450	1.032412	1.303900	1.801539	1.078603	1.580104	-2.044960	0.479017	0.091852
PF00153	100.0	4.748584	5.275751	1.204712	1.708018	3.934791	4.600306	3.788899	4.368224	-6.254185	0.601696	0.066114
PF00271	100.0	5.123037	4.904604	1.451270	1.796391	4.369218	4.295515	4.162241	4.012131	-4.717794	0.404760	0.032685
PF00397	100.0	2.588195	2.970887	1.158531	1.382846	1.960060	2.397599	1.653095	2.081855	-2.006116	0.525521	0.090403
PF00512	100.0	5.131445	4.881163	1.690496	1.800010	4.461895	4.279929	4.170142	3.984389	-4.653579	0.523652	0.061763
PF00595	100.0	3.334047	3.221569	0.895611	1.081072	2.495118	2.462376	2.381716	2.313799	-4.188000	0.638344	0.122616
PF01535	100.0	4.322277	3.979826	1.224074	1.385038	3.538004	3.281500	3.359701	3.070533	-4.613339	0.634167	0.101301
PF02518	100.0	4.652743	4.368112	1.360238	1.545159	3.895410	3.701312	3.690167	3.464886	-4.738062	0.542793	0.065928
PF07679	100.0	4.212461	3.726689	0.956884	1.156200	3.348723	2.955152	3.248062	2.819124	-5.532714	0.646804	0.107914
PF13354	100.0	4.241852	3.984350	1.503699	1.568915	3.571977	3.367274	3.292515	3.089726	-3.606365	0.482382	0.057704

R^2 metric is negative! Regression on column attention derived coeffiecients is completely unable to capture the relationship with Patristic Distance.

Marija Zelic

Subtree Subsampling & Regression per Family (without normalization)



	Depth	mean (training)	mean (test)	std (training)	std (test)	RMSE (training)	RMSE (test)	MAE (training)	MAE (test)	R^2 (test)	Pearson (test)	Slope (test)
PF00004	497.0	0.519417	0.489386	0.187741	0.190294	0.063275	0.073405	0.048963	0.056670	0.851203	0.925155	0.904290
PF00005	496.0	1.766427	1.724581	0.434299	0.469619	0.879484	0.857923	0.797309	0.776525	-2.337378	0.547581	0.212398
PF00041	444.0	2.723425	2.786513	0.908404	0.999032	1.940372	2.039726	1.803108	1.883242	-3.168551	0.550474	0.110123
PF00072	420.0	1.634417	1.570977	0.629495	0.606348	0.885254	0.819830	0.777613	0.708459	-0.828116	0.538179	0.218993
PF00076	417.0	2.266275	2.221608	1.095843	1.103997	1.665779	1.638343	1.459978	1.429853	-1.202286	0.596457	0.171366
PF00096	452.0	1.093830	1.068569	0.386019	0.360031	0.405677	0.371840	0.317338	0.290718	-0.066676	0.444359	0.334163
PF00153	499.0	0.826697	0.771000	0.536446	0.531204	0.454297	0.463895	0.397730	0.393928	0.237365	0.588601	0.515198
PF00271	406.0	0.777424	0.720473	0.256760	0.250437	0.178261	0.187843	0.140680	0.148378	0.437405	0.754717	0.834970
PF00397	478.0	0.605789	0.572751	0.777926	0.749005	0.563975	0.544247	0.404516	0.378875	0.472015	0.714650	0.472493
PF00512	303.0	1.731326	1.693998	0.590430	0.632949	0.939394	0.960868	0.799222	0.799528	-1.304572	0.296352	0.104029
PF00595	427.0	0.130929	0.138902	0.094186	0.095239	0.015296	0.019278	0.010104	0.012855	0.959027	0.979643	0.955432
PF01535	431.0	2.106557	2.163299	0.680126	0.763225	1.295191	1.385592	1.160524	1.242309	-2.295837	0.486073	0.121301
PF02518	461.0	1.545062	1.515503	0.468847	0.469448	0.722396	0.705429	0.613355	0.600856	-1.258040	0.456658	0.217792
PF07679	480.0	2.547888	2.562403	0.928498	0.923447	1.800219	1.812794	1.612506	1.644216	-2.853654	0.501724	0.111967
PF13354	455.0	0.508731	0.504869	0.072400	0.091301	0.017682	0.018154	0.013718	0.013982	0.960466	0.980262	0.964951

- Results are way worse compared to the case when we perform normalization!
- Interestingly for the families for which there is high Pearson correlation coefficent between Hamming and Patristic distances they are still plausible, even very good.