SUPPLEMENTAL DATA

Supplemental Figure S1. We benchmark the performance of various manifold visualization methods using 6048 protein domain sequence datasets from Pfam. This is an expansion of the benchmarks shown in Figure 3A. The dotted lines in gray are a visual guide indicating x=0.8 and x=0.9. As an aside, we note the apparent performance of MDS in Spearman rank correlation. Although MDS has the ability to capture global structure, we demonstrate that the method is not good at identifying distinct clusters, quantified by silhouette score in Supplemental Figure S2.

Supplemental Figure S2. Manifold scatter plot visualizations for the phosphatases, protein kinases, and radical SAM datasets using various dimensionality reduction methods. This is an expansion of the scatter plots shown in Figure 4A, 6A, and 8A. Color labels are maintained, corresponding to their respective figures in the main text.

Supplemental Figure S3. Embedding-based tree of human phosphatases. VIBEs are shown on each hierarchical cluster. This is the full data for the tree shown in Figure 4B.

Supplemental Figure S4. Embedding-based tree of human phosphatases. VIBEs are shown on each hierarchical cluster. This is the full data for the tree shown in Figure 5.

Supplemental Figure S5. Embedding-based tree of human protein kinases. VIBEs are shown on each hierarchical cluster. This is the full data for the tree shown in Figure 6B.

Supplemental Figure S6. Embedding-based tree of human protein kinases. VIBEs are shown on each hierarchical cluster. This is the full data for the tree shown in Figure 7.

Supplemental Figure S7. Embedding-based tree of the radical SAM enzymes. VIBEs are shown on each hierarchical cluster. This is the full data for the tree shown in Figure 8A and 8B.

representation	distance metric	silhouette score
mean_special_tokens	TS_SS	0.2873
beginning_of_sequence	TS_SS	0.2851
beginning_of_sequence	cosine	0.2751
mean_special_tokens	cosine	0.2607
mean_residue_tokens	TS_SS	0.2375
mean_residue_tokens	cosine	0.2372
beginning_of_sequence	manhattan	0.2062
beginning_of_sequence	euclidean	0.1932
end_of_sequence	cosine	0.1886
mean_special_tokens	manhattan	0.1874
mean_special_tokens	euclidean	0.1865
end_of_sequence	TS_SS	0.1856
mean_residue_tokens	manhattan	0.1804
mean_residue_tokens	jensenshannon	0.1706
mean_residue_tokens	euclidean	0.1703
end_of_sequence	manhattan	0.1588
end_of_sequence	euclidean	0.1554
end_of_sequence	jensenshannon	0.1495
mean_special_tokens	jensenshannon	0.0557
beginning_of_sequence	jensenshannon	-0.3880

Supplemental Table S1. Evaluating methods for calculating embedding distances within the human phosphatase dataset. Starting from full-sized embeddings of the phosphatase sequences, we evaluate different strategies for producing a fixed-size embedding (the representation column) and distance metrics. The resulting distance matrices from each combination were evaluated using the silhouette score, given the phosphatase structural fold labels: CC1, CC2, CC3, HAD, HP, AP, PPL, and PPM. The RTR1 and PHP fold labels were not considered for calculating silhouette score as they only contain one example each.

representation	distance metric	silhouette score
end_of_sequence	TS_SS	0.3793
mean_special_tokens	TS_SS	0.3621
end_of_sequence	cosine	0.3210
mean_special_tokens	cosine	0.2978
beginning_of_sequence	TS_SS	0.2225
beginning_of_sequence	cosine	0.2128
end_of_sequence	jensenshannon	0.2031
end_of_sequence	euclidean	0.2029
end_of_sequence	manhattan	0.2013
mean_special_tokens	euclidean	0.1873
mean_special_tokens	manhattan	0.1872
mean_residue_tokens	cosine	0.1840
mean_residue_tokens	TS_SS	0.1650
beginning_of_sequence	manhattan	0.1433
beginning_of_sequence	euclidean	0.1389
mean_residue_tokens	manhattan	0.1303
mean_residue_tokens	jensenshannon	0.1292
mean_residue_tokens	euclidean	0.1283
mean_special_tokens	jensenshannon	0.1148
beginning_of_sequence	jensenshannon	-0.2399

Supplemental Table S2. Evaluating methods for calculating embedding distances within the human protein kinase dataset. Starting from full-sized embeddings of the protein kinase sequences, we evaluate different strategies for producing a fixed-size embedding (the representation column) and distance metrics. The resulting distance matrices from each combination were evaluated using the silhouette score, given the kinase evolutionary group labels: TK, TKL, RGC, NEK, STE, AGC, CAMK, CMGC, CK1, and Atypical which contains divergent lipid and small molecule kinases. The Others group was not considered for calculating silhouette score as it reflects evolutionary intermediates which could not be classified into any of the major evolutionary groups.

representation	distance metric	silhouette score
mean_special_tokens	TS_SS	0.5419
end_of_sequence	TS_SS	0.5317
beginning_of_sequence	TS_SS	0.5168
mean_special_tokens	cosine	0.4814
end_of_sequence	cosine	0.4744
beginning_of_sequence	cosine	0.4724
mean_residue_tokens	TS_SS	0.4186
mean_residue_tokens	cosine	0.3837
end_of_sequence	euclidean	0.3376
end_of_sequence	jensenshannon	0.3374
end_of_sequence	manhattan	0.3369
mean_special_tokens	euclidean	0.3338
mean_special_tokens	manhattan	0.3334
beginning_of_sequence	euclidean	0.3248
beginning_of_sequence	manhattan	0.3246
mean_special_tokens	jensenshannon	0.2918
mean_residue_tokens	manhattan	0.2738
mean_residue_tokens	euclidean	0.2686
mean_residue_tokens	jensenshannon	0.2643
beginning_of_sequence	jensenshannon	-0.1061

Supplemental Table S3. Evaluating methods for calculating embedding distances within the radical SAM dataset. Starting from full-sized embeddings of the radical SAM sequences, we evaluate different strategies for producing a fixed-size embedding (the representation column) and distance metrics. The resulting distance matrices from each combination were evaluated using the silhouette score, given the radical SAM family labels: LipA, MTTase, MTaseA, B12-binding, HemN, Elp3, BATS, Viperin, SPASM, Activating Enzyme, QueE, TYW1, and PHP-dependent.