Build of a model for the detection of signal peptides in proteins - supplementary materials – Piccolo Gianluca

FEATURE SET	C	γ	MCC	SENSITIVITY	PRECISION	ACCURACY
AA	4	0.5	0.756	0.901	0.836	0.877
AA + HYDROPATHY	2	1.0	0.792	0.921	0.855	0.895
AA + CHARGE	1	0.5	0.752	0.917	0.820	0.873
AA + HELIX PROP	1	0.5	0.763	0.909	0.836	0.881
AA + TM PROP	8	0.5	**0.825**	**0.925**	**0.883**	**0.913**
AA + HYDRO + CHARGE	8	0.5	0.784	0.920	0.847	0.891
AA + HYDRO + HELIX	4	Scale	0.803	0.917	0.868	0.901
AA + TM PROP (DUPLICATE)	8	0.5	**0.825**	**0.925**	**0.883**	**0.913**
AA + HYDRO + HELIX (DUPLICATE)	4	Scale	0.803	0.917	0.868	0.901
AA + CHARGE + HELIX	4	0.5	0.755	0.905	0.832	0.876
AA + CHARGE + TM	2	0.5	0.821	0.921	0.883	0.911
AA + ALL FEATURES	4	0.5	0.818	0.917	0.883	0.910

***Table 1**. Best performing model uses AA composition + Transmembrane Propensity features (MCC = 0.825). Adding more features doesn't always improve performance. Most models perform best with γ = 0,5. C parameters varies between 1-8, with higher values generally preferred. There are two duplicate feature combinations in the results. All models achieve > 90% sensitivity (recall).*

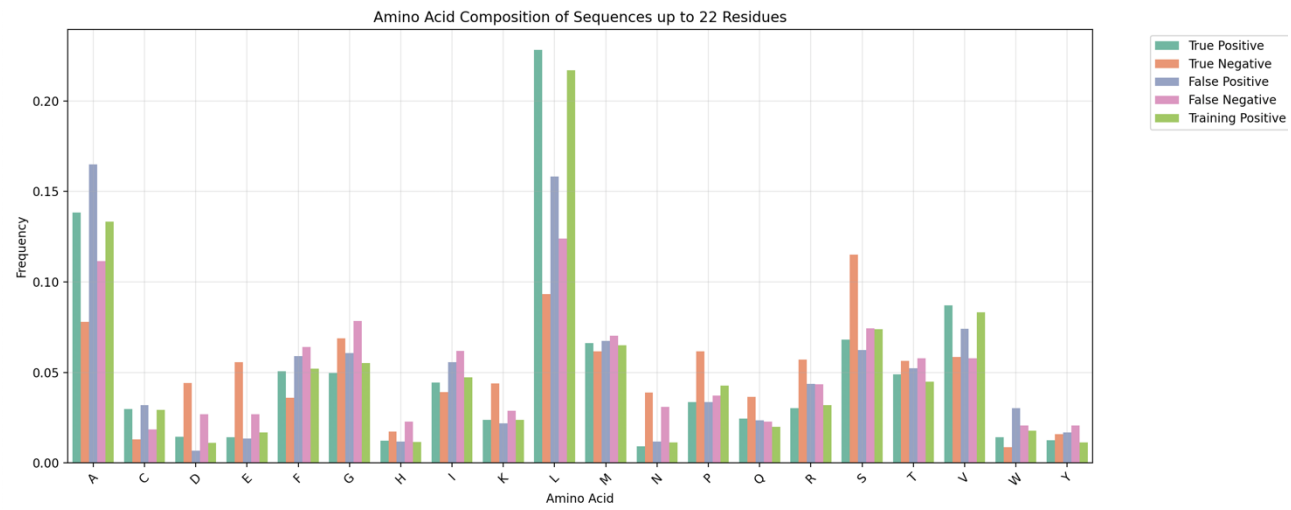
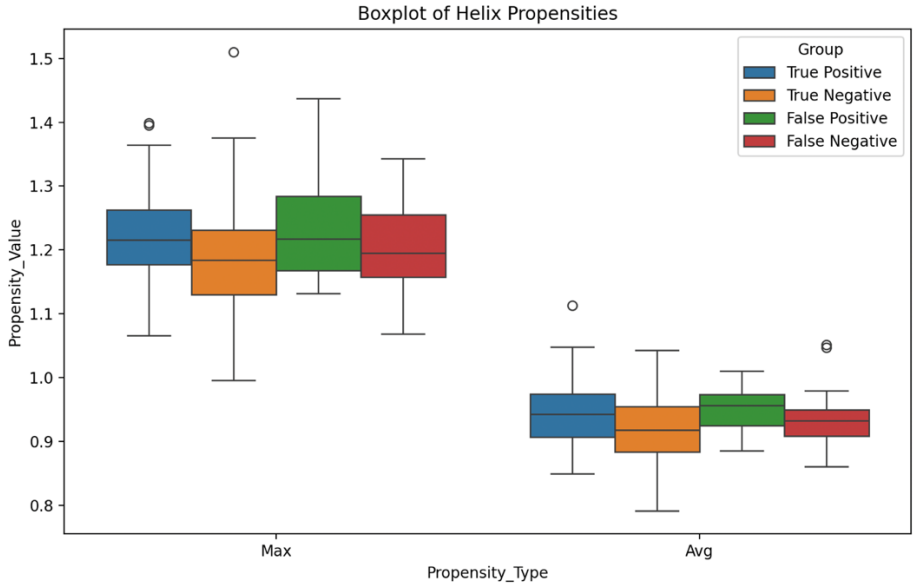


Figure 1. *Aminoacidic Composition comparison between different datasets. Highest peaks appear for Leucine across all categories, which is a characteristic of signal peptides due to their hydrophobic core. True Positives and Training Positives show similar patterns, suggesting good model learning. Alanine shows notable differences between positive and negative cases. Some amino acids show clear discriminative power between positive and negative cases.*

***Figure 2**. Metazoa consistently represents the largest proportion across all categories, but with varying percentages. The model seems to have more false predictions in Metazoa, possibly due to its larger representation in the dataset. Fungi shows relatively consistent proportions between false positives and negatives. The presence of "Others" category in some distributions but not all suggests potential bias in prediction for less represented kingdoms.*



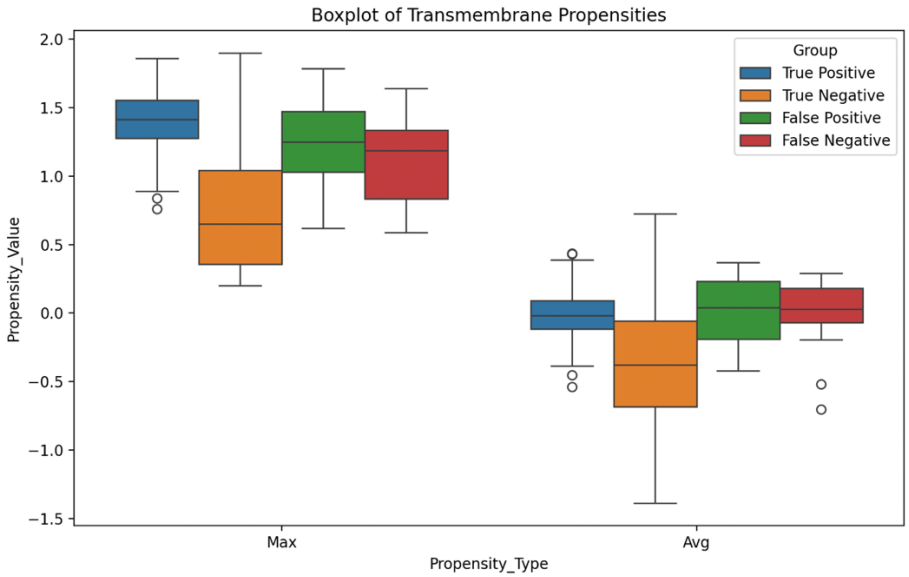


***Figure 3**. Boxplots of Transmembrane Propensity and Helix Propensity among different results. Transmembrane propensities are more discriminative than helix propensities. The model appears to rely more heavily on transmembrane properties for classification. False predictions often show intermediate properties between true positives and negatives. Helix propensities alone might not be sufficient for accurate prediction. These suggest that transmembrane propensity is a strong predictor of signal peptides; the model correctly captures the importance of transmembrane regions. Cases with intermediate properties are more likely to be misclassified. Combining both properties might help reduce false predictions.*