

Machine Learning Based Predictive Model for the Analysis of Sequence Activity Relationships Using Protein Spectra and Protein Descriptors

Adam McKenna¹, Dr. Sandhya Dubey²

¹ School of Electronics, Electrical Engineering and Computer Science, Queen's University of Belfast, University Road, BT7 1NN, Belfast, United Kingdom, Belfast

² Department of Data Science and Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), Manipal, Karnataka 576104, India

¹amckenna41@qub.ac.uk, ²sandhya.dubey@manipal.edu

TABLE S1

BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY A(i)

Predictive Model	Index Category	R2	RMSE	MSE	RPD	MAE	Explained Variance
AA _{PONJ960101} _Bag	Geometry	0.749	2.938	8.632	1.994	2.461	0.752
AA _{MEIH800101} _RF	Geometry	0.730	2.942	8.654	1.924	2.424	0.730
AA _{ARGP820101} _RF	Hydrophobic	0.730	3.286	10.796	1.923	2.784	0.730
AA _{RACS820109} _Bag	Geometry	0.722	3.035	9.213	1.898	2.366	0.726
AA _{QIAN880133} _RF	Sec Struct	0.720	3.039	9.235	1.889	2.497	0.720
AA _{AURR980113} _Ada	Sec Struct	0.716	3.241	10.502	1.877	2.562	0.716
AA _{KIMC930101} _RF	Sec Struct	0.714	3.544	12.562	1.870	2.659	0.718
AA _{PONP800101} _Ada	Hydrophobic	0.714	3.544	12.557	1.869	2.919	0.714
AA _{KARP850101} _Ada	Flexibility	0.712	3.490	12.179	1.864	2.636	0.727
AA _{GUYH850102} _Bag	Hydrophobic	0.710	3.299	10.881	1.858	2.650	0.713

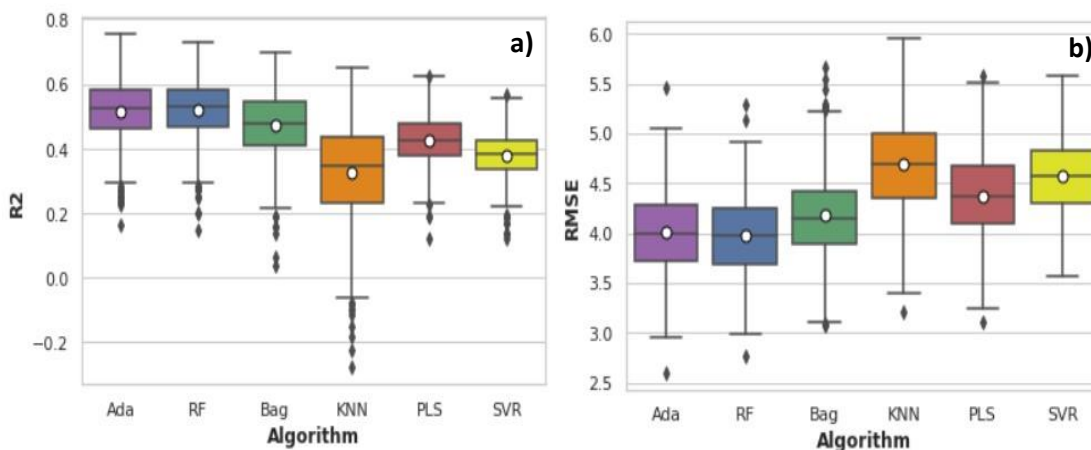


Fig. S1a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy A(i), with outliers.

TABLE S2

BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY B(i)

<i>Predictive Model</i>	Descriptor Group	R²	RMSE	MSE	RPD	MAE	Explained Variance
<i>DPComp_RF</i>	Composition	0.796	2.547	6.487	2.216	1.910	0.797
<i>TPComp_RF</i>	Composition	0.789	2.634	6.938	2.179	2.019	0.790
<i>TPComp_Ada</i>	Composition	0.774	2.587	6.695	2.103	2.020	0.779
<i>GAuto_PLS</i>	Autocorrelation	0.773	2.749	7.556	2.097	2.180	0.773
<i>TPComp_PLS</i>	Composition	0.772	2.702	7.301	2.095	2.173	0.773
<i>CTriad_PLS</i>	Conjoint Triad	0.771	2.914	8.492	2.091	2.475	0.773
<i>MAuto_Bag</i>	Autocorrelation	0.760	3.005	9.031	2.042	2.355	0.769
<i>MAuto_Ada</i>	Autocorrelation	0.758	2.932	8.595	2.034	2.337	0.758
<i>CTriad_Ada</i>	Conjoint Triad	0.754	2.599	6.755	2.0151	2.085	0.755
<i>TPComp_Bag</i>	Composition	0.796	2.547	6.488	2.215	1.901	0.797

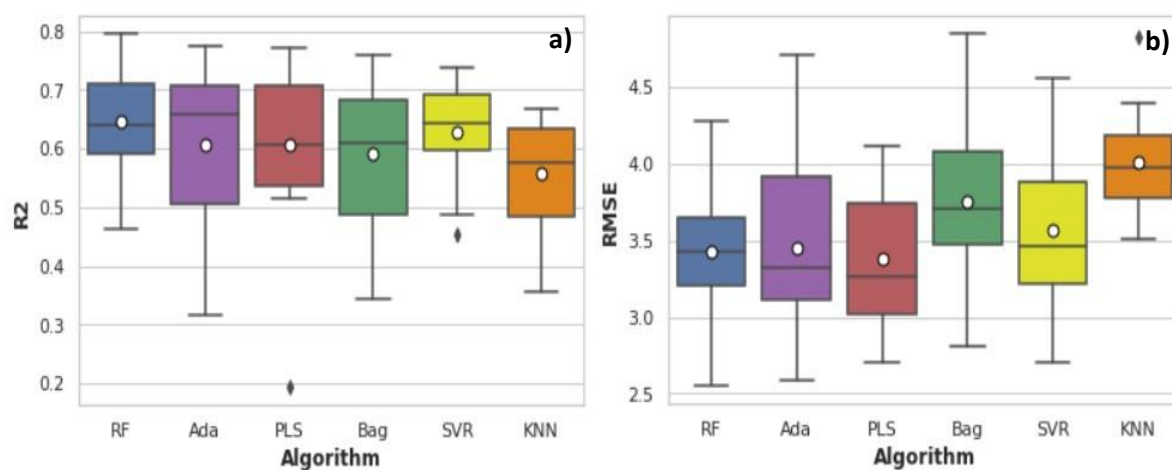
Fig. S2a, b. Boxplot of R² and RMSE values for each algorithm, using encoding strategy B(i), with outliers.

TABLE S3
BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY B(ii)

<i>Predictive Model</i>	<i>Descriptor Group</i>	R2	RMSE	MSE	RPD	MAE	Explained Variance
<i>TPComp^T_CTD_PLS</i>	Composition CTD	0.874	2.09	4.377	2.823	1.758	0.876
<i>DPComp^QOrder_PLS</i>	Composition Quasi-sequence-order	0.858	2.391	5.719	2.651	1.956	0.871
<i>AAComp^C_CTD_KNN</i>	Composition CTD	0.856	2.419	5.854	2.632	1.947	0.857
<i>TPComp^CTD_PLS</i>	Composition CTD	0.854	2.481	6.154	2.621	1.971	0.855
<i>TPComp^APAAComp_PLS</i>	Composition Pseudo-composition	0.837	2.455	6.025	2.477	1.921	0.844
<i>CTriad^PAAComp_PLS</i>	Conjoint Triad Pseudo-composition	0.832	2.428	5.894	2.442	1.952	0.833
<i>TPComp^GAuto_RF</i>	Composition Autocorrelation	0.828	2.220	4.930	2.414	1.860	0.829
<i>MAuto^CTriad_PLS</i>	Autocorrelation Conjoint Triad	0.825	2.526	6.379	2.394	1.992	0.825
<i>TPComp^C_CTD_Ada</i>	Composition CTD	0.824	2.333	5.445	2.382	1.903	0.824
<i>CTriad^QOrder_PLS</i>	Conjoint Triad Quasi-sequence-order	0.874	2.092	4.377	2.823	1.758	0.876

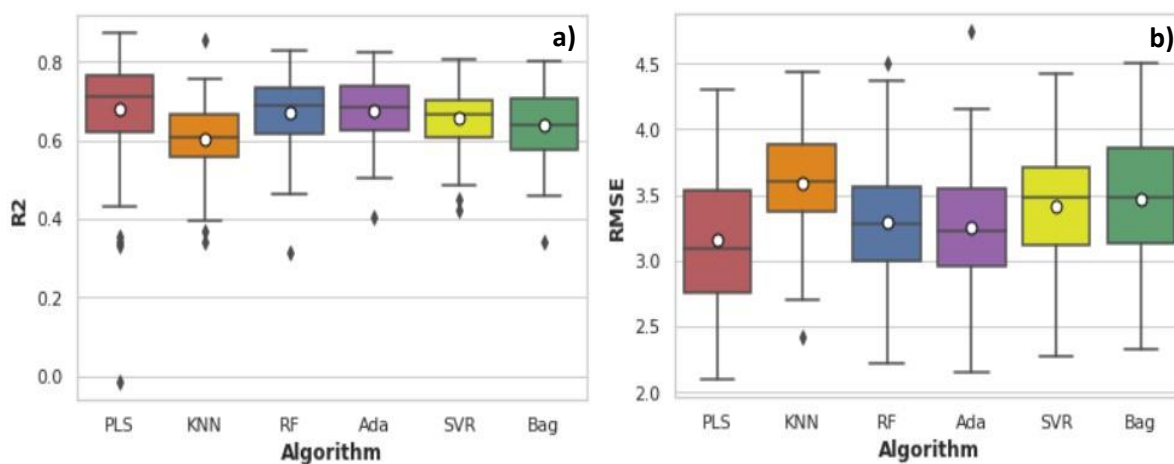


Fig. S3a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy B(ii), with outliers.

TABLE S4
BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY B(iii)

<i>Predictive Model</i>	Descriptor Group	R2	RMSE	MSE	RPD	MAE	Explained Variance
<i>GAuto^CTriad^APAAComp_RF</i>	Autocorrelation Conjoint Triad Pseudo-composition	0.867	2.192	4.801	2.740	1.841	0.869
<i>DPComp^CTD^CTriad_PLS</i>	Composition CTD Conjoint Triad	0.861	2.395	5.736	2.685	1.830	0.863
<i>TPComp^C_CTD^CTriad_PLS</i>	Composition CTD Conjoint Triad	0.856	2.229	4.969	2.632	1.840	0.856
<i>TPComp^GAuto^QOrder_PLS</i>	Composition Autocorrelation Quasi-sequence-order	0.852	2.393	5.727	2.602	1.945	0.854
<i>GAuto^T_CTD^PAAComp_PLS</i>	Autocorrelation CTD Pseudo-composition	0.852	2.521	6.356	2.596	2.105	0.857
<i>GAuto^CTD^T_CTD_PLS</i>	Autocorrelation CTD CTD	0.850	2.109	4.447	2.582	1.668	0.857
<i>AAComp^TPComp^GAuto_PLS</i>	Composition Composition Autocorrelation	0.847	2.699	7.285	2.557	2.144	0.847
<i>DPComp^MAuto^CTriad_Ada</i>	Composition Autocorrelation Conjoint Triad	0.846	2.431	5.910	2.550	1.999	0.847
<i>T_CTD^D_CTD^CTriad_RF</i>	CTD CTD Conjoint Triad	0.845	2.401	5.764	2.544	1.900	0.851
<i>TPComp^D_CTD^CTriad_Bag</i>	Composition CTD Conjoint Triad	0.845	2.357	4.804	2.740	1.841	0.869

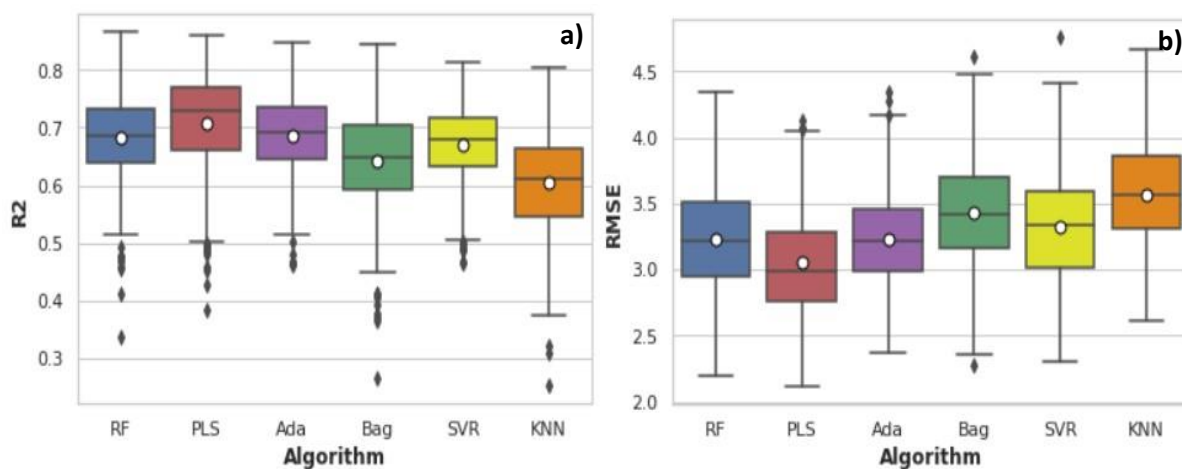


Fig. S4a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy B(iii), with outliers.

TABLE S5
BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY C(i)

<i>Predictive Model</i>	Descriptor Group	Index Category	R2	RMSE	MSE	RPD	MAE	Explained Variance
<i>AA_{KIDA850101}</i> [^] <i>DPComp_Bag</i>	Composition	Hydrophobic	0.894	1.973	3.892	3.073	3.073	1.493
<i>AA_{PALJ810114}</i> [^] <i>DPComp_RF</i>	Composition	Sec Struct	0.890	1.900	3.610	3.018	3.018	1.547
<i>AA_{YUTK870102}</i> [^] <i>DPComp_PLS</i>	Composition	Observable	0.887	2.038	4.152	2.974	2.974	1.570
<i>AA_{RICJ880108}</i> [^] <i>CTriad_PLS</i>	Conjoint Triad	Sec Struct	0.887	1.852	3.430	2.971	2.971	1.466
<i>AA_{NAKH920103}</i> [^] <i>CTriad_RF</i>	Conjoint Triad	Composition	0.885	1.882	3.540	2.943	2.943	1.440
<i>AA_{ROBB760111}</i> [^] <i>DPComp_PLS</i>	Composition	Sec Struct	0.876	2.051	4.205	2.836	2.836	1.653
<i>AA_{ISOY800108}</i> [^] <i>CTriad_Bag</i>	Conjoint Triad	Sec Struct	0.873	2.074	4.301	2.809	2.809	1.707
<i>AA_{AVBF000108}</i> [^] <i>TPComp_PLS</i>	Composition	Polar	0.873	2.198	4.831	2.804	2.804	1.792
<i>AA_{RADA880101}</i> [^] <i>GAuto_PLS</i>	Autocorrelation	Hydrophobic	0.870	2.456	6.030	2.775	2.775	2.060
<i>AA_{AURR980106}</i> [^] <i>DPComp_PLS</i>	Composition	Sec Struct	0.869	2.021	3.892	2.759	3.073	1.493

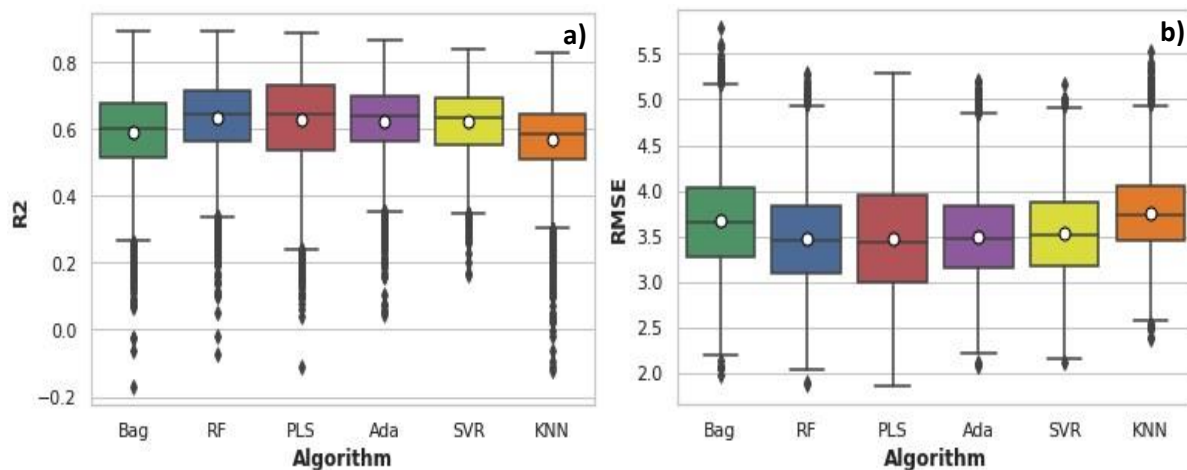


Fig. S5a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy C(i), with outliers.

TABLE S6
BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY C(ii)

<i>Predictive Model</i>	Index Category	Descriptor Group	R2	RMSE	MSE	RPD	MAE	Explained Variance
<i>AA_{RACS820114}^TPComp^D_CTD_PLS</i>	Geometry	Composition CTD	0.903	2.070	3.758	3.073	1.601	0.910
<i>AA_{JANJ790101}^AComp^TPComp_PLS</i>	Hydrophobic	Composition Composition	0.899	1.808	3.672	3.077	1.654	0.906
<i>AA_{WILM950103}^TPComp^CTriad_PLS</i>	Hydrophobic	Composition Conjoint Triad	0.898	2.089	4.022	3.012	1.701	0.898
<i>AA_{CHAM820101}^AComp^DPComp_PLS</i>	Polar	Composition Composition	0.897	2.137	4.301	3.030	1.756	0.897
<i>AA_{QIAN880110}^TPComp^QSOOrder_PLS</i>	Sec_struct	Composition Quasi-sequence-order	0.897	2.072	4.465	3.035	1.804	0.894
<i>AA_{TANS770110}^DPComp^CTriad_Bag</i>	Sec_struct	Composition Conjoint Triad	0.896	2.048	3.822	3.011	1.679	0.893
<i>AA_{QIAN880120}^TPComp^C_CTD_PLS</i>	Sec_struct	Composition CTD	0.894	2.013	3.201	2.993	1.504	0.890
<i>AA_{OLSK800101}^DPComp^T_CTD_PLS</i>	Geometry	Composition CTD	0.894	2.053	3.876	2.907	1.550	0.886
<i>AA_{GEIM800103}^DPComp^PAAComp_PLS</i>	Sec_struct	Composition Pseudo-Composition	0.893	2.022	4.014	2.898	1.467	0.886
<i>AA_{LIFS790101}^AComp^TPComp_PLS</i>	Sec_struct	Composition Composition	0.893	2.113	3.772	3.003	1.541	0.885

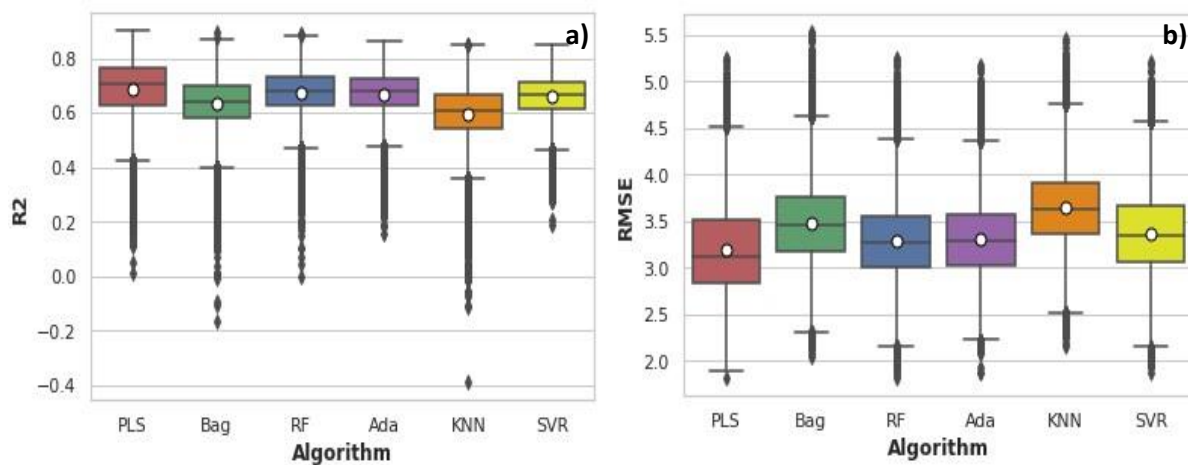


Fig. S6a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy C(ii), with outliers.