Comparative study on the utility of protein spectra and protein descriptors in the analysis of sequence activity relationships

Adam McKenna[1], Dr. Sandhya Dubey[2]

[1]*School of Electronic, Electrical Engineering and Computer Science, Queen's University of Belfast*
[2]*Department of Computer Applications, Manipal Academy of Higher Education (MAHE)*

[1]amckenna41@qub.ac.uk, [2]sandhya.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

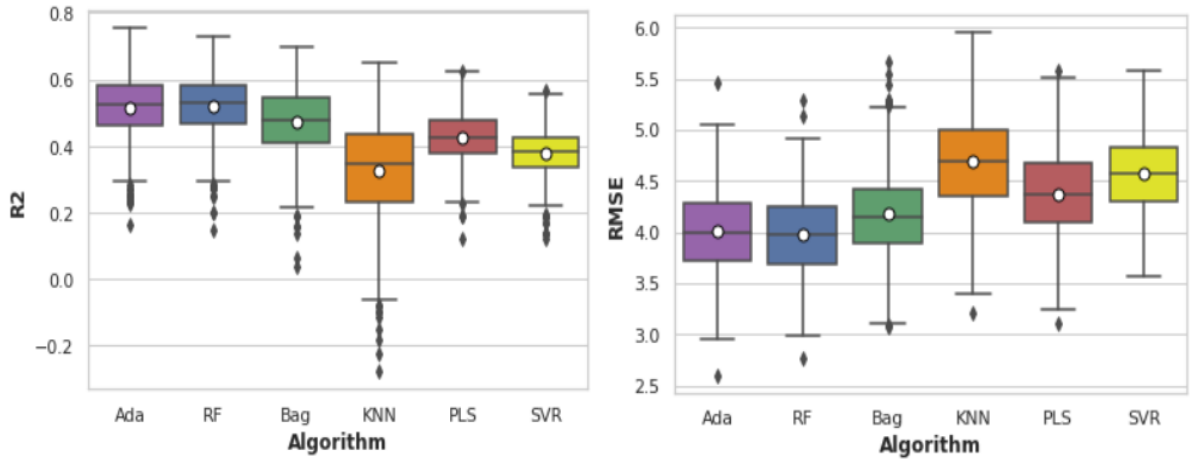

Fig. S1a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy *A(i)*, with outliers.

TABLE S2

BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY B(I)

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

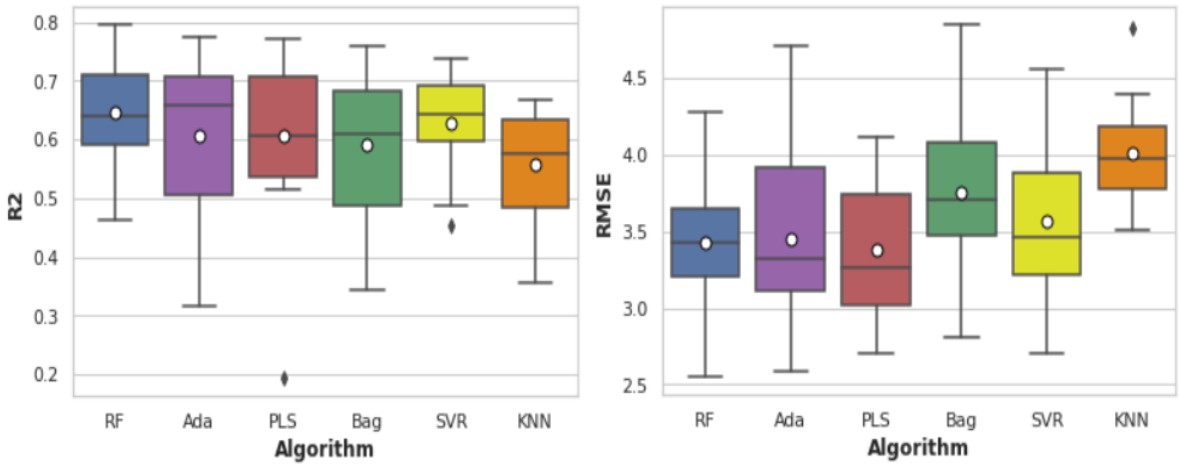

Fig. S2a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy *B(i)*, with outliers.

TABLE S3

BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY B(II)

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

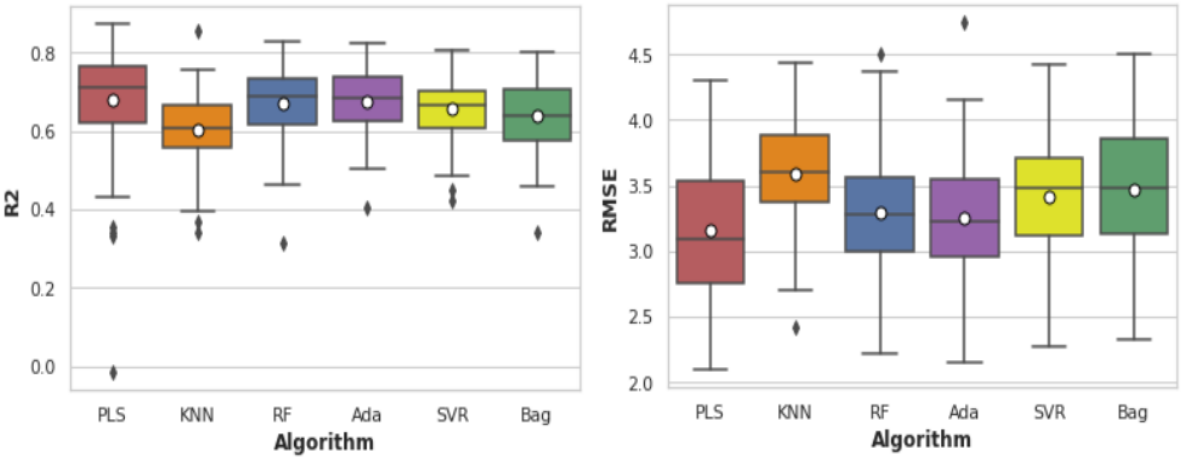

Fig. S3a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy *B(ii)*, with outliers.

TABLE S4

BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY B(III)

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

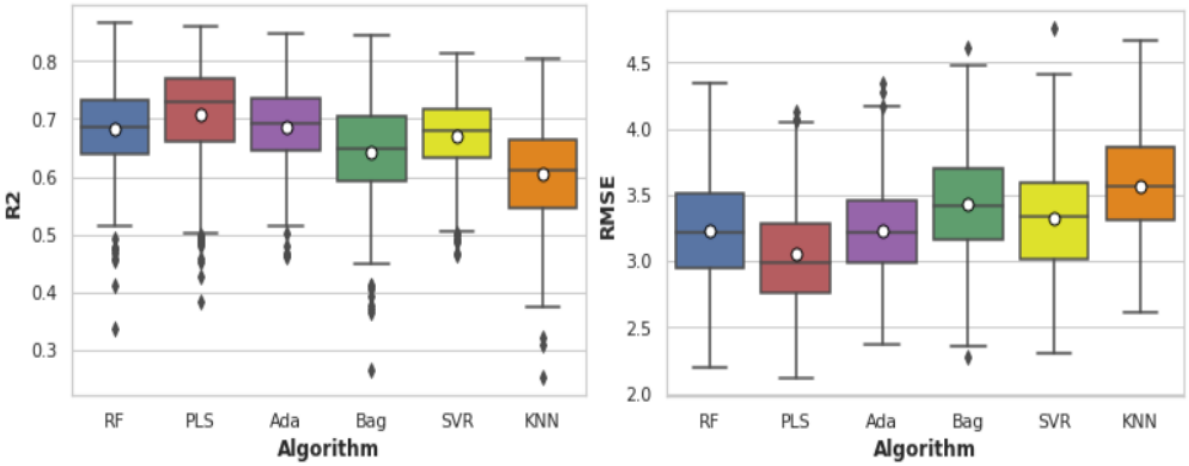

Fig. S4a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy *B(iii)*, with outliers.

TABLE S5

BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY C(I)

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

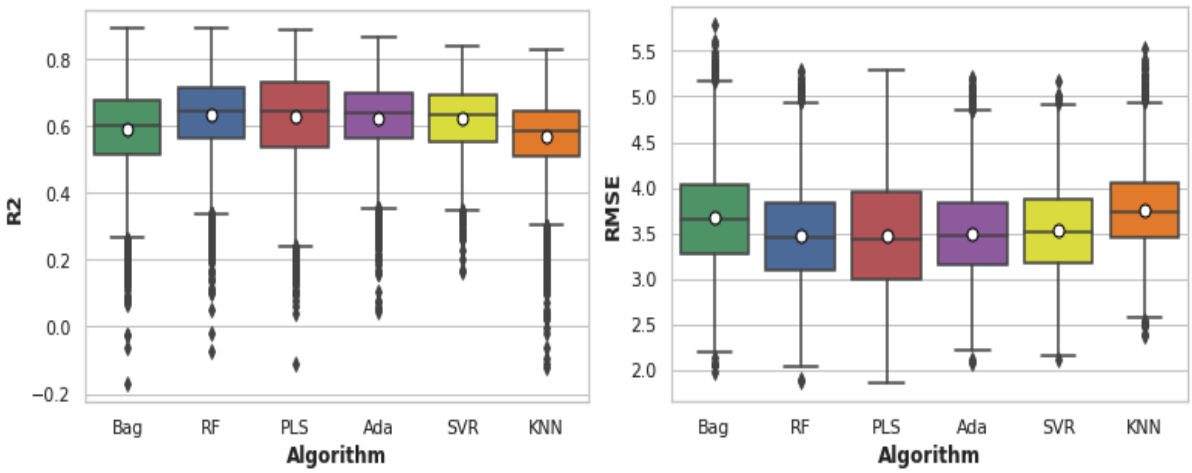

Fig. S5a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy *C(i)*, with outliers.

TABLE S6

BEST PERFORMING PREDICTIVE MODELS USING ENCODING STRATEGY C(II)

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

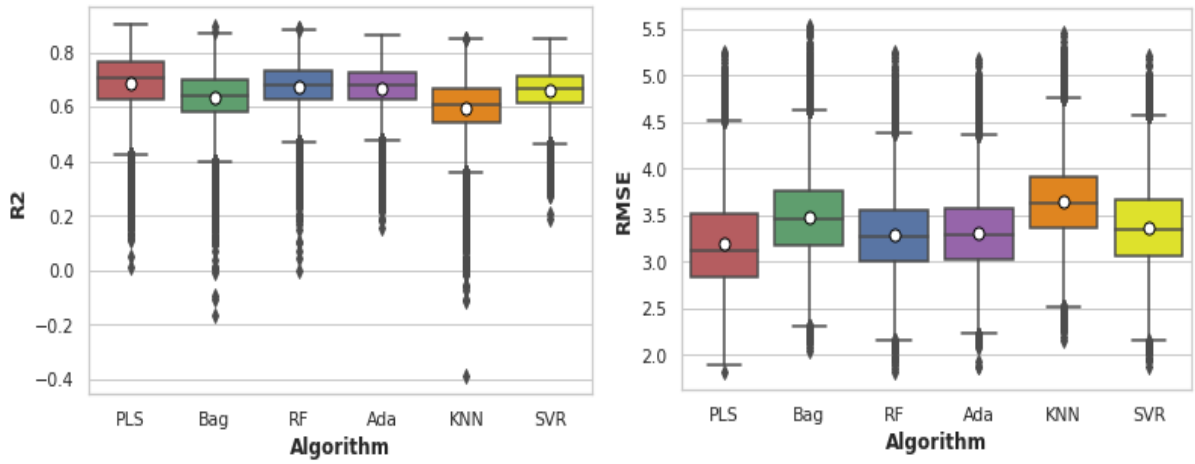

Fig. S6a, b. Boxplot of R2 and RMSE values for each algorithm, using encoding strategy *C(ii)*, with outliers.