Predict RMSD from Protein Physiochemical Properties

RMSD-Size of the residue.

F1 - Total surface area. F2 - Non polar exposed area. F3 - Fractional area of exposed non polar residue. F4 - Fractional area of exposed non polar part of residue. F5 - Molecular mass weighted exposed area. F6 - Average deviation from standard exposed area of residue. F7 - Euclidian distance. RMSD is a measure of how well a predicted protein's structure fits to an experimental structure, with a value of zero being a perfect fit. F1-F7 correspond to other features about the protein's structure. The goal here is to use the other numerical features to predict the target value RMSD.

TPOT Analysis was done to find a strong model for modelling. Extra Trees Regressor Model was used. It had the following test dataset (10% of total data) metrics:

R-squared for Extra Trees Regressor model 1 is: 0.71

MSE for Extra Trees Regressor model is 11.03

RMSE for Extra Trees Regressor model is 3.32

Below sample values can be input to get a prediction for RMSD. There are plots showing the relationship between RMSD and the three most correlated features, as well as a plot of F3 vs F2, as F2 is the strongest correlator to RMSD.

Enter the protein characteristics to get the predicted RMSD

Predicted RMSD: 8

Sample Values

RMSD	F1	F2	F3	F4	F5	F6	F7	F8	F9
0	14636.4	2928.43	0.2001	197.99	2037005.2065	269.467	4719.28	89	46.5464
4.495	8836.87	2592.51	0.2934	41.7062	1234444.5347	92.7959	4060.94	100	32.1182
1.722	7644.94	1994.08	0.2608	65.3175	1047191.6269	108.621	3606.68	21	37.5168
14.399	8310.5	3763.58	0.4529	67.4463	1172630.836	127.032	4032.74	55	35.6014
16.293	19006.6	5509.93	0.2899	208.633	2694229.2165	299.982	5893.21	245	20.9852
1.732	6997.32	2024.58	0.2893	66.5503	961785.5523	93.0358	3088.28	51	39.34
11.302	4074.48	1301.24	0.3194	37.3653	548468.1286	51.2057	1391.21	29	44.6544
1.541	10193.6	3150.95	0.3091	127.562	1363098.367	166.662	4019.18	37	33.5786
20.924	16257.9	4636.65	0.2852	154.928	2201503.9306	221.028	5337.84	91	28.9365
17.792	7425.18	2749.14	0.3702	37.6903	1012163.3314	113.233	3124.74	36	40.6634

F1:

1

F2:

1

F3:

1

F4:

1

F5:

1

F6:

1

F7:

1

F8:

1

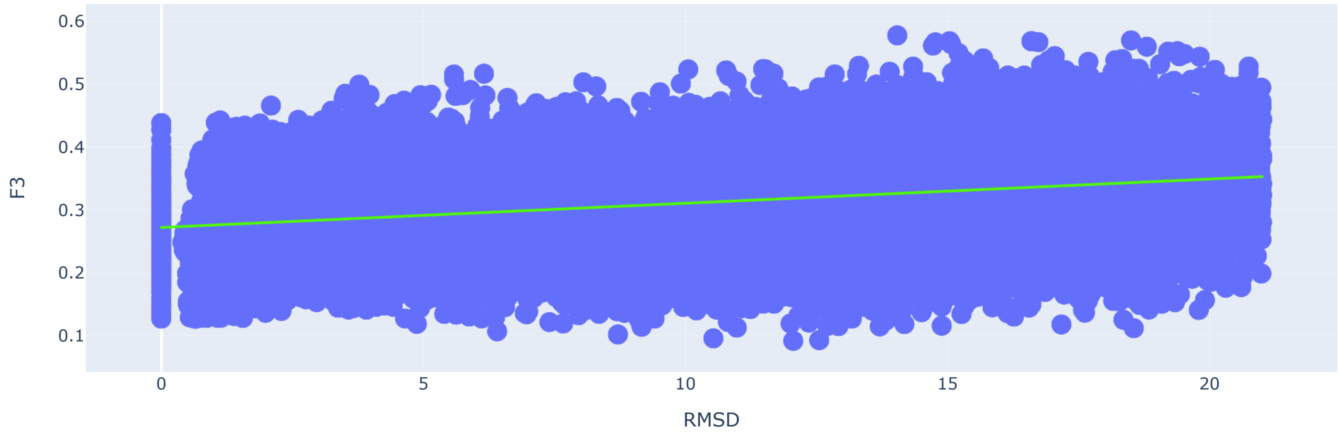
F9:

1

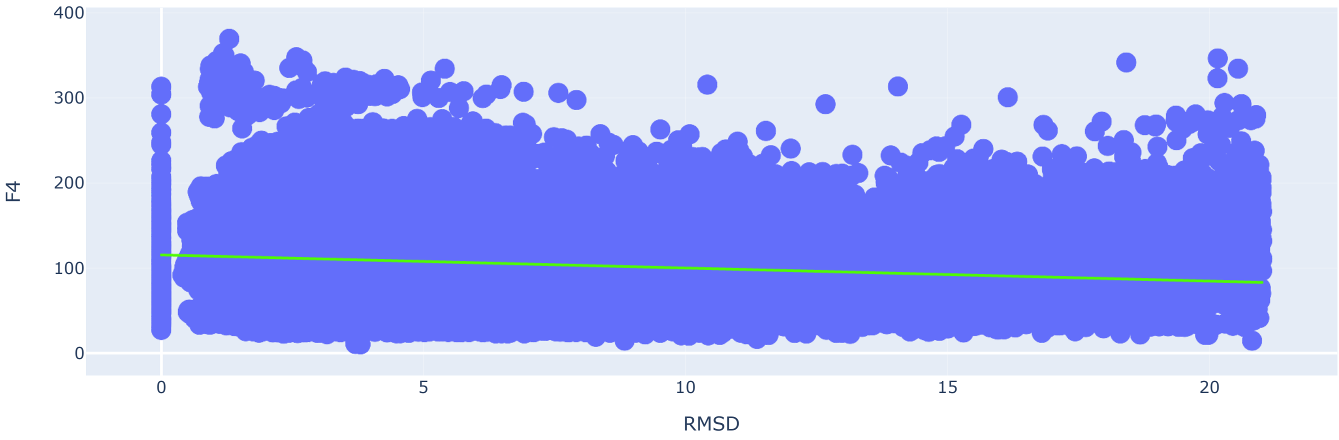
Correlation Values:

RMSD	F1	F2	F3	F4	F5	F6	F7	F8	F9
1	-0.0151	0.1569	0.3743	-0.1698	-0.014	-0.0361	-0.0033	0.0003	0.0628
-0.0151	1	0.9066	0.1263	0.9311	0.9982	0.9675	0.554	0.6513	-0.8982
0.1569	0.9066	1	0.5026	0.7931	0.9029	0.9084	0.5159	0.5842	-0.7862
0.3743	0.1263	0.5026	1	0.0312	0.1226	0.2007	0.0801	0.0953	-0.069
-0.1698	0.9311	0.7931	0.0312	1	0.9257	0.9381	0.4852	0.6769	-0.8918
-0.014	0.9982	0.9029	0.1226	0.9257	1	0.9618	0.5537	0.643	-0.8978
-0.0361	0.9675	0.9084	0.2007	0.9381	0.9618	1	0.5382	0.6626	-0.882
-0.0033	0.554	0.5159	0.0801	0.4852	0.5537	0.5382	1	0.347	-0.5211
0.0003	0.6513	0.5842	0.0953	0.6769	0.643	0.6626	0.347	1	-0.6373
0.0628	-0.8982	-0.7862	-0.069	-0.8918	-0.8978	-0.882	-0.5211	-0.6373	1

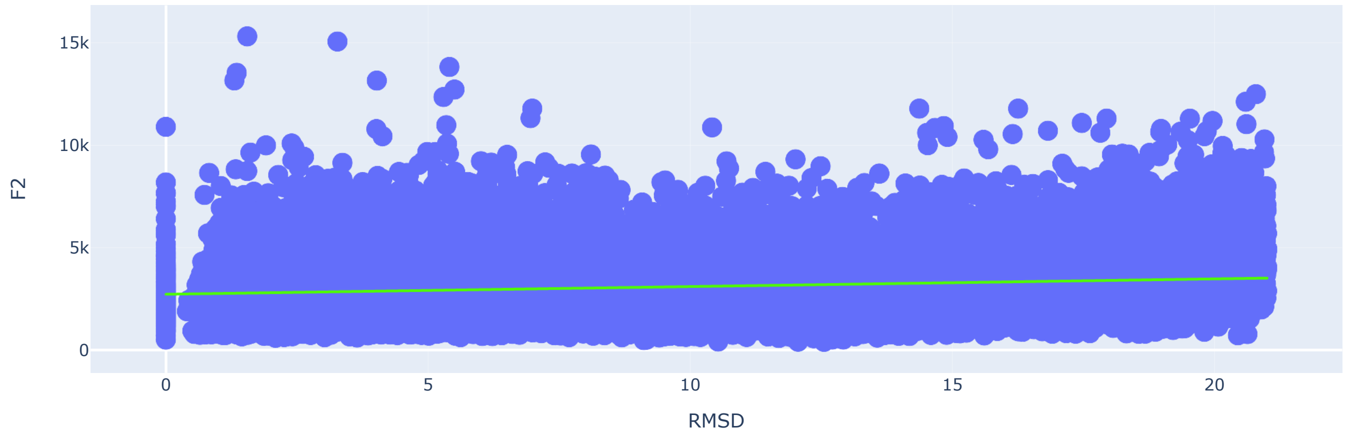
RMSD vs. F3, Correlation: +0.374



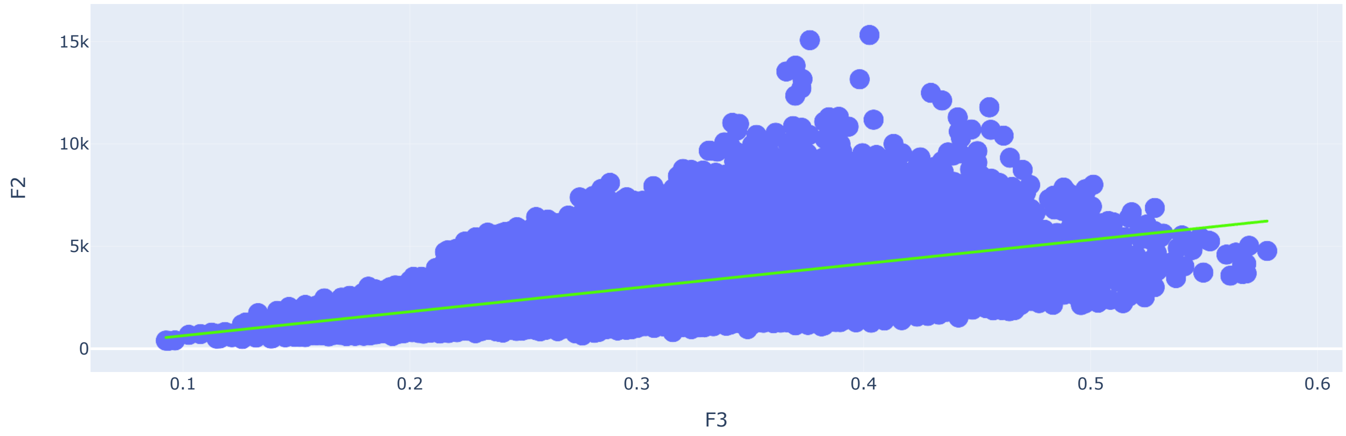
RMSD vs F4, Correlation: -0.170



RMSD vs F2, Correlation: +0.157



F3 vs. F2, Correlation: +0.503


F2
15k
10k
5k
0
0.1
0.2
0.3
0.4
0.5
0.6
F3