SUPPORTING INFORMATION

Structural and Sequence Similarity Makes Significant Impact on Machine Learning-Based Scoring Functions for Protein-Ligand Interactions

Yang Li[†,‡] *and Jianyi Yang*[*,‡]

[†]College of Life Sciences, Nankai University, Tianjin 300071, China

[‡]School of Mathematical Sciences, Nankai University, Tianjin 300071, China

*To whom correspondence should be addressed: yangjy@nankai.edu.cn

Table S1. The performance of RF-Score and X-Score re-trained at different levels of similarity. The training sets are constructed using equation (1), with TM-score as the similarity measure, where the TM-score for two complexes is defined as the **highest** pairwise chains TM-score.

Cutoff	#Train	TC	RF-Score			X-Score		
			R_p	R_s	RMSE	R_p	R_s	RMSE
1.000	1105	0.620	0.783	0.769	1.542	0.643	0.707	1.867
0.999	955	0.605	0.726	0.721	1.685	0.645	0.707	1.863
0.998	852	0.601	0.692	0.698	1.758	0.645	0.707	1.862
0.997	792	0.602	0.675	0.686	1.804	0.647	0.708	1.864
0.996	767	0.600	0.672	0.682	1.815	0.647	0.708	1.860
0.995	745	0.597	0.660	0.670	1.825	0.648	0.707	1.858
0.990	658	0.594	0.624	0.622	1.885	0.648	0.707	1.852
0.985	622	0.591	0.626	0.625	1.885	0.647	0.709	1.863
0.980	574	0.593	0.610	0.604	1.907	0.645	0.708	1.860
0.975	562	0.594	0.608	0.600	1.904	0.645	0.709	1.856
0.970	546	0.592	0.606	0.599	1.913	0.645	0.707	1.863
0.965	533	0.592	0.604	0.597	1.917	0.644	0.707	1.864
0.960	529	0.592	0.588	0.577	1.939	0.643	0.706	1.868
0.955	520	0.592	0.598	0.586	1.927	0.644	0.706	1.858
0.950	514	0.590	0.590	0.580	1.940	0.644	0.706	1.852
0.900	475	0.583	0.583	0.577	1.949	0.643	0.702	1.846
0.850	440	0.585	0.588	0.582	1.949	0.643	0.704	1.845

0.800	425	0.584	0.587	0.583	1.951	0.643	0.707	1.844
0.750	409	0.585	0.605	0.612	1.935	0.645	0.709	1.838
0.700	385	0.587	0.625	0.640	1.906	0.644	0.709	1.836
0.650	376	0.587	0.605	0.617	1.937	0.645	0.709	1.839
0.600	358	0.585	0.605	0.615	1.932	0.644	0.709	1.835
0.550	324	0.592	0.599	0.605	1.939	0.646	0.708	1.835
0.500	284	0.588	0.585	0.594	1.956	0.646	0.708	1.822
0.450	217	0.594	0.557	0.581	1.992	0.645	0.708	1.843
0.400	110	0.571	0.552	0.576	2.000	0.632	0.674	1.863

Table S2. The performance of RF-Score and X-Score trained at different similarity levels. The training sets are constructed using equation (1), with sequence identity as the similarity measure, where the sequence identity for two complexes is defined as the **highest** pairwise chains sequence identity.

Cutoff	#Trian	TC	RF-Score			X-Score		
			R_p	R_s	RMSE	R_p	R_s	RMSE
1.000	1105	0.620	0.783	0.769	1.542	0.643	0.707	1.867
0.999	767	0.602	0.650	0.655	1.839	0.644	0.710	1.875
0.998	766	0.601	0.659	0.662	1.826	0.644	0.711	1.875
0.997	762	0.600	0.654	0.658	1.833	0.644	0.711	1.874
0.996	727	0.600	0.648	0.649	1.845	0.647	0.708	1.858
0.995	707	0.596	0.633	0.640	1.867	0.646	0.708	1.860
0.990	679	0.595	0.624	0.625	1.877	0.646	0.707	1.859
0.985	665	0.596	0.614	0.618	1.896	0.646	0.709	1.860
0.980	657	0.595	0.620	0.621	1.887	0.647	0.709	1.859
0.975	627	0.589	0.605	0.604	1.904	0.647	0.707	1.854
0.970	625	0.588	0.608	0.608	1.902	0.647	0.708	1.854
0.965	592	0.592	0.606	0.603	1.906	0.646	0.706	1.860
0.960	592	0.592	0.606	0.603	1.906	0.646	0.706	1.860
0.955	570	0.590	0.598	0.592	1.922	0.645	0.706	1.870
0.950	570	0.590	0.598	0.592	1.922	0.645	0.706	1.870
0.900	549	0.593	0.594	0.586	1.928	0.644	0.706	1.857
0.850	537	0.594	0.593	0.586	1.926	0.645	0.705	1.851
0.800	531	0.594	0.603	0.598	1.916	0.644	0.705	1.854
0.750	524	0.594	0.598	0.597	1.925	0.643	0.703	1.856
0.700	520	0.594	0.589	0.585	1.935	0.643	0.702	1.851
0.650	512	0.592	0.578	0.586	1.950	0.643	0.703	1.852

0.600	505	0.594	0.561	0.563	1.973	0.642	0.702	1.856
0.550	497	0.595	0.573	0.572	1.962	0.642	0.700	1.859
0.500	484	0.590	0.561	0.557	1.972	0.641	0.698	1.857
0.450	467	0.590	0.566	0.556	1.973	0.640	0.696	1.858
0.400	414	0.590	0.576	0.586	1.962	0.645	0.705	1.856
0.350	308	0.588	0.564	0.579	2.014	0.645	0.707	1.870
0.300	149	0.583	0.471	0.456	2.126	0.648	0.701	1.837

Table S3. The performance of RF-Score trained with similar samples only. In the training sets, the samples are required to have **sequence** similarity (with the test samples) higher than the specified cutoffs.

Cutoff	#Train	TC	RF-Score		
			R_p	R_s	RMSE
0.100	1105	0.620	0.783	0.769	1.542
0.150	1104	0.620	0.784	0.769	1.541
0.200	1091	0.621	0.780	0.765	1.551
0.250	1049	0.623	0.781	0.765	1.549
0.300	924	0.627	0.778	0.755	1.554
0.350	755	0.634	0.768	0.741	1.577
0.400	654	0.640	0.783	0.764	1.561
0.450	597	0.643	0.780	0.768	1.575
0.500	583	0.646	0.782	0.768	1.570
0.550	570	0.643	0.781	0.771	1.569
0.600	563	0.644	0.784	0.777	1.567
0.650	560	0.645	0.780	0.767	1.578
0.700	553	0.643	0.781	0.770	1.574
0.750	552	0.643	0.775	0.758	1.585
0.800	551	0.642	0.782	0.766	1.570
0.850	547	0.642	0.784	0.770	1.572
0.900	540	0.643	0.781	0.771	1.584
0.950	521	0.648	0.787	0.780	1.567
0.955	521	0.648	0.787	0.780	1.567
0.960	499	0.648	0.782	0.769	1.578
0.965	499	0.648	0.782	0.769	1.578

0.970	467	0.656	0.784	0.773	1.578
0.975	465	0.656	0.781	0.770	1.577
0.980	435	0.651	0.774	0.768	1.585
0.985	427	0.651	0.777	0.770	1.580
0.990	413	0.653	0.771	0.757	1.590
0.995	379	0.656	0.756	0.737	1.617
0.996	359	0.652	0.753	0.729	1.634
0.997	324	0.656	0.753	0.733	1.623
0.998	320	0.654	0.750	0.733	1.644
0.999	319	0.653	0.753	0.729	1.632