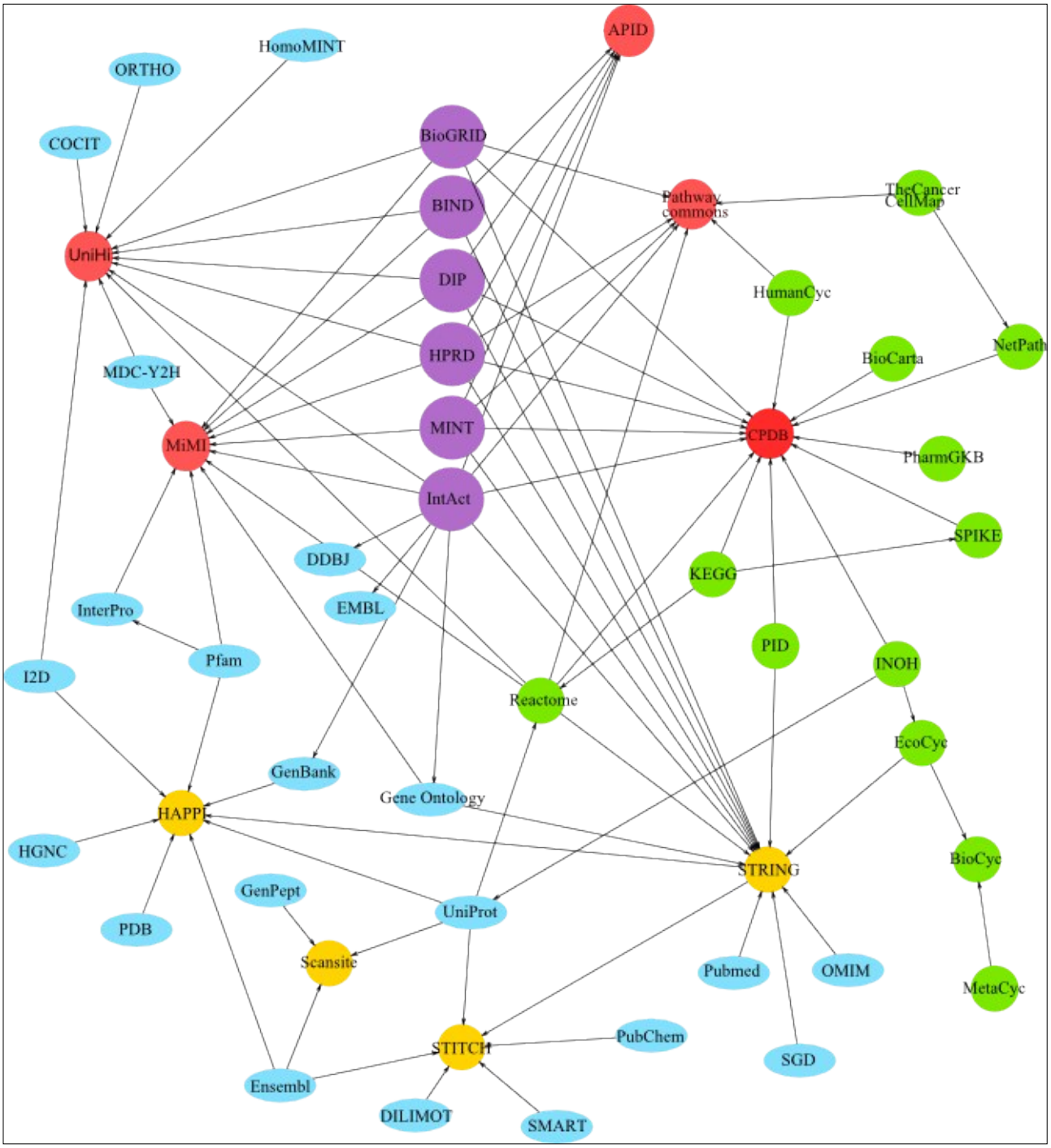
Electronic Supplementary Material (ESI) for Molecular BioSystems.



Supplementary Figure 1. Relation map of PPI databases with several other databases: Purple color nods represents primary protein interaction database, red nodes shows secondary databases, green circles are pathway databases, blue nodes are feature databases and lastly orange nods represents computationally predicted database. The figure shows the detailed interconnections between various

databases with PPI repositories. As it is visible that meta-mining PPI databases are not only deriving data from primary PPI databases (purple circles) but they are also using various pathway information (green circles), as well as genomic and proteomic features (blue ovals).

Table A: Optimizing cutoff for COR value in NIPs

Bin no.	COR bin Size	PPI Count	NIP count Samples				Average NIP count	PPI percentage	NIP average percentage
			(I)	(II)	(III)	(IV)			
268	0.33to0.335	376	100	66	97	67	82.50	82.00	17.89
269	0.335to0.34	333	77	97	65	80	79.75	80.67	19.26
270	0.34to0.345	371	71	77	67	86	75.25	83.13	16.84
271	0.345to0.35	322	73	68	68	65	68.50	82.45	17.54
272	0.35to0.355	316	82	55	64	71	68.00	82.29	17.65
273	0.355to0.36	316	67	52	47	56	55.50	85.06	14.90
274	0.36to0.365	286	57	51	55	52	53.75	84.17	15.81
275	0.365to0.37	271	42	48	49	61	50.00	84.42	15.53
276	0.37to0.375	272	45	47	55	50	49.25	84.66	15.32
277	0.375to0.38	253	42	42	47	38	42.25	85.69	14.30
278	0.38to0.385	259	48	29	28	35	35.00	88.09	11.84
279	0.385to0.39	279	36	33	37	40	36.50	88.43	11.56
280	0.39to0.395	250	46	33	29	36	36.00	87.41	12.54
281	0.395to0.40	261	30	28	41	20	29.75	89.76	10.17
282	0.4to0.405	224	27	25	23	26	25.25	89.87	10.12
283	0.405to0.410	220	31	34	30	27	30.50	87.82	12.16
284	0.41to0.4150	213	29	32	25	26	28.00	88.38	11.60
285	0.415to0.420	201	17	39	28	17	25.25	88.84	11.02
286	**0.42to0.4250**	**206**	**17**	**21**	**19**	**28**	**21.25**	**90.65**	**9.32**
287	0.425to0.430	203	22	21	14	18	18.75	91.54	8.43
288	0.43to0.4350	198	16	16	22	20	18.50	91.45	8.53
289	0.435to0.440	185	17	18	13	19	16.75	91.69	8.29
290	0.44to0.4450	204	14	10	16	10	12.50	94.22	5.75
291	0.445to0.450	166	19	12	13	7	12.75	92.86	7.08
292	0.45to0.4550	176	12	8	12	9	10.25	94.49	5.49
293	0.455to0.460	179	9	8	11	11	9.75	94.83	5.16
294	0.46to0.4650	176	9	12	13	8	10.50	94.36	5.61
295	0.465to0.470	168	6	7	11	4	7.00	96.10	3.98
296	0.47to0.4750	159	8	7	9	7	7.75	95.35	4.64

297	0.475to0.480	177	16	4	7	5	8.00	95.67	4.26
298	0.48to0.4850	173	8	7	4	1	5.00	97.19	2.78
299	0.485to0.490	141	5	1	5	6	4.25	97.07	2.90
300	0.49to0.4950	132	6	9	5	3	5.75	95.82	4.15
301	0.495to0.500	150	2	6	5	6	4.75	96.93	3.06
302	0.5to0.5050	157	6	1	3	7	4.25	97.36	2.61
303	0.505to0.510	129	0	4	3	1	2.00	98.47	1.51
304	0.51to0.5150	134	5	3	6	3	4.25	96.92	3.06
305	0.515to0.520	145	2	5	4	2	3.25	97.80	2.18
306	0.52to0.5250	124	1	2	1	3	1.75	98.60	1.38
307	0.525to0.530	128	5	4	4	5	4.50	96.60	3.39
308	0.53to0.5350	117	2	0	3	5	2.50	97.90	2.07
309	0.535to0.540	136	1	5	0	2	2.00	98.55	1.43
310	0.54to0.5450	114	4	2	1	5	3.00	97.43	2.54
311	0.545to0.550	129	1	1	0	2	1.00	99.23	0.76
312	0.55to0.5550	121	0	0	1	0	0.25	99.79	0.20
313	0.555to0.560	110	1	1	1	1	1.00	99.09	0.91
314	0.56to0.5650	117	0	1	0	1	0.50	99.57	0.42
315	0.565to0.570	145	1	0	1	0	0.50	99.65	0.34
316	0.57to0.5750	106	1	1	3	2	1.75	98.37	1.61
317	0.575to0.580	112	0	1	1	1	0.75	99.33	0.66
318	0.58to0.5850	117	0	2	1	1	1.00	99.15	0.84
319	0.585to0.590	116	1	1	3	2	1.75	98.51	1.47
320	0.59to0.5950	112	1	1	0	3	1.25	98.89	1.09
321	0.595to0.600	100	0	0	0	1	0.25	99.75	0.24
322	0.6to0.6050	122	0	0	0	0	0.00	0.00	0.00
323	0.605to0.610	103	1	0	1	0	0.50	99.51	0.48

In table 1, only a specific section (from bin number 268 to 323) is shown to explain the frequency distribution. The COR value from -1 to +1 was divided into 400 bins each having equal range (column 2). To take equal number of NIP pairs as with total PPI set, random shuffling was done to create four unbiased equal sets of NIP samples (column 4-7) having 68,265 pairs each out of 13,523,822 NIPs. The ratio percentage of PPI vs NIP (column 3 and 8 respectively) was evaluated. In row number 286 of the table it is shown that the PPI:NIP ratio percentage is 90.64: 9.32 within a COR bin range of 0.42 to 0.425, Therefore the upper limit was taken as cut off. Hence 43,676 Negative pairs which were exhibiting correlation value ≥ 0.425 were removed from the NIP set.

Table B: List of all attributes used for Feature Selection

No.	Feature	Name
Features derived probabilistic modelling of DDIs		
1.	Number of domains mapped in Protein-A	numDomainP1
2.	Number of domains mapped in Protein-B	numDomainP2
3.	Number of inferred domain-domain pair	numDDI
4.	Maximum frequency of domain pair found for protein pair	Max_freq_DDI
5.	Minimum frequency of domain pair found for protein pair	Min_freq_DDI
6.	Minimum probability score of a domain pair found for protein pair	Min_Zscore
7.	Maximum probability score of a domain pair found for protein pair	Max_Zscore
8.	Probability Score of most frequent domain pair (feature no. 4)	Zscore_of_mostfreq_DDI
Features derived from Network Analysis		
9.	Betweenness of vertex protein-A	btw_P1
10.	Betweenness of vertex protein-B	btw_P2
11.	Vertex degree of protein-A	dgr_P1
12.	Vertex degree of protein-B	dgr_P2
13.	Closeness of Vertex, protein-A	cls_P1
14.	Closeness of Vertex, protein-B	cls_P2
15.	Eccentricity of Vertex protein-A	ecc_P1
16.	Eccentricity of Vertex protein-B	ecc_P2
17.	First order of vertex neighborhood for protein-A	nb1_P1
18.	First order of vertex neighborhood for protein-B	nb1_P2
19.	Second order of vertex neighborhood for protein-A	nb2_P1
20.	Second order of vertex neighborhood for protein-B	nb2_P2
21.	Third order of vertex neighborhood for protein-A	nb3_P1
22.	Third order of vertex neighborhood for protein-B	nb3_P2
23.	Fourth order of vertex neighborhood for protein-A	nb4_P1
24.	Fourth order of vertex neighborhood for protein-B	nb4_P2
25.	Fifth order of vertex neighborhood for protein-A	nb5_P1
26.	Fifth order of vertex neighborhood for protein-A	nb5_P2
27.	Closeness Centrality for Vertex protein-A	CC_P1
28.	Closeness Centrality for Vertex protein-B	CC_P2
29.	Eigenvector centralities of vertex protein-A within network graph.	EvCV_P1
30.	Eigenvector centralities of vertex protein-B within	EvCV_P2

	network graph.	
31.	The first Eigen value of the adjacency matrix for protein-A and B of the network graph	Eigen_value
Features based on number of Amino acids and protein disordered region		
32.	Total number of amino acids in protein-A	totalAA_p1
33.	Total number of amino acids in protein-B	totalAA_p2
34.	Total percentage disorder in protein-A	disorder_p1
35.	Total percentage disorder in protein-B	disorder_p2
36.	Total no. of disordered regions greater than 30 amino acids in protein-A	greater_than_30aa_p1
37.	Total no. of disordered regions greater than 30 amino acids in protein-B	greater_than_30aa_p2
38.	Total no. of disordered regions greater than 50 amino acids in protein-A	greater_than_50aa_p1
39.	Total no. of disordered regions greater than 50 amino acids in protein-B	greater_than_50aa_p2
40.	Number of disordered segments in protein-A	disordered_segments_p1
41.	Number of disordered segments in protein-A	disordered_segments_p2
Features derived from Gene Coexpression		
42.	Mutual Rank score of Coexpression correlation for protein-A with protein-B	MR
43.	Coexpression Correlation value for protein-A with protein-B	COR

Table C: Monte Carlo feature selection RI scores

	Relative Importance (RI)		
Sample in each set →	**10,000**	**20,000**	**30,000**
numDomainP1	0.011125	0.011405	0.011533
numDomainP2	0.010958	0.010945	0.010808
numDDI	0.011705	0.011262	0.011506
Max_freq_DDI	0.094794	0.095964	0.091623
Min_freq_DDI	0.038702	0.037525	0.037113
Min_Zscore	0.102666	0.10289	0.106265
Max_Zscore	0.158712	0.159373	0.162039
Zscore_of_mostfreq_DDI	0.065379	0.065063	0.064829
btw_P1	0.04852	0.049285	0.04842
btw_P2	0.042709	0.043263	0.04468
dgr_P1	0.127083	0.128434	0.126189
dgr_P2	0.095592	0.092601	0.09182

cls_P1	**0**	**0**	**0**
cls_P2	**0**	**0**	**0**
ecc_P1	0.016583	0.015596	0.016341
ecc_P2	0.020448	0.021143	0.020051
nb1_P1	0.140609	0.136645	0.135385
nb1_P2	0.102071	0.095198	0.094426
nb2_P1	0.078105	0.077251	0.076784
nb2_P2	0.072151	0.072604	0.075024
nb3_P1	0.070988	0.073346	0.070529
nb3_P2	0.06622	0.064964	0.068029
nb4_P1	0.067792	0.068395	0.069977
nb4_P2	0.063984	0.066885	0.065548
nb5_P1	0.068985	0.065779	0.060934
nb5_P2	0.071152	0.073381	0.072992
CC_P1	**0.000773**	**0.000745**	**0.00068**
CC_P2	**0.00231**	**0.002062**	**0.001838**
EvCV_P1	0.05412	0.056981	0.057277
EvCV_P2	0.040393	0.042157	0.042534
Eigen_value	0.008167	0.008237	0.008396
totalAA_p1	0.007211	0.007165	0.007378
totalAA_p2	0.007367	0.007197	0.00726
disorder_p1	0.007082	0.007008	0.006752
disorder_p2	**0.006258**	**0.006364**	**0.006531**
greater_than_30aa_p1	0.009178	0.009056	0.008659
greater_than_30aa_p2	0.009055	0.009205	0.008998
greater_than_50aa_p1	0.008414	0.008043	0.007459
greater_than_50aa_p2	0.007746	0.008073	0.007864
disordered_segments_p1	0.010064	0.009955	0.009854
disordered_segments_p2	0.009737	0.009748	0.00979
MR	0.036752	0.036774	0.03495
COR	0.031597	0.029808	0.02947

The Relative Importance (RI) calculated for all attributes is given below in table 3.6.3. The values written in bold are the one which were lower than the calculated RI cutpoint and hence were rejected.

Table D: Boruta analysis of balanced subset of 20,000 training protein pairs

Feature name	meanZ	medianZ	minZ	maxZ	normHits	decision
Min_freq_DDI	107.4076	107.488	101.5251	111.8468	1	Confirmed
btw_P1	41.8486	41.8441	37.81631	47.4557	1	Confirmed
btw_P2	34.6266	34.5838	30.2344	39.0896	1	Confirmed
nb1_P1	29.5187	29.5488	26.118	32.1054	1	Confirmed
Min_Zscore	27.3161	27.8625	20.1584	31.7243	1	Confirmed
dgr_P1	26.7139	26.7385	23.7017	28.8826	1	Confirmed
nb1_P2	25.6636	25.6391	23.02444	27.8038	1	Confirmed
Max_freq_DDI	24.9243	24.9594	23.26345	27.0220	1	Confirmed
EvCV_P1	23.4507	23.3654	21.7353	25.2942	1	Confirmed
dgr_P2	23.2052	23.3140	21.1034	25.2808	1	Confirmed
Zscore_of_mostfreq_DDI	22.6571	22.7747	11.7359	28.1174	1	Confirmed
EvCV_P2	20.9027	20.8579	19.1110	22.5786	1	Confirmed
Max_Zscore	19.5804	19.8451	14.1055	23.2826	1	Confirmed
numDDI	19.4480	19.4527	17.0649	22.2929	1	Confirmed
COR	19.0620	18.9098	15.7164	21.9557	1	Confirmed
nb2_P2	17.0405	17.0515	15.5234	19.0421	1	Confirmed
nb4_P2	14.5390	14.4850	13.0172	16.1486	1	Confirmed
CC_P2	14.2185	14.1828	12.4218	15.8264	1	Confirmed
MR	14.1638	13.8948	12.2590	20.4912	1	Confirmed
nb2_P1	14.1068	14.1711	12.4378	16.2413	1	Confirmed
nb5_P2	13.4432	13.4810	12.0933	14.8004	1	Confirmed
nb3_P2	12.6757	12.6490	11.3621	14.2029	1	Confirmed
nb3_P1	12.4684	12.4637	10.6805	14.0684	1	Confirmed
CC_P1	12.3555	12.4279	10.6423	14.1354	1	Confirmed
nb4_P1	12.3199	12.3125	11.1192	13.7349	1	Confirmed
nb5_P1	11.0801	10.9565	8.5732	14.3850	1	Confirmed
numDomainP2	8.5666	8.6151	6.4594	10.4765	1	Confirmed
disorder_p1	8.5624	8.5389	6.6577	10.8721	1	Confirmed
numDomainP1	7.9007	7.9222	5.5389	9.8261	1	Confirmed
Eigen_value	7.4277	7.4408	5.3306	10.1259	1	Confirmed
ecc_P2	6.6311	6.6202	3.2722	10.2478	0.9752	Confirmed
totalAA_p1	6.1831	6.1817	3.4220	8.6904	0.9752	Confirmed
ecc_P1	5.4984	5.6192	2.0521	7.8477	0.9338	Confirmed
greater_than_50aa_p1	3.8705	3.8249	1.4221	5.9820	0.7933	Confirmed
cls_P2	3.7913	3.7844	2.5174	4.9920	0.8016	Confirmed
disorder_p2	3.3984	3.4408	1.1452	5.7096	0.6528	Tentative
cls_P1	3.3271	3.3234	2.0081	4.4789	0.7024	Confirmed

totalAA_p2	3.2726	3.2366	0.6650	5.6274	0.5289	Tentative
greater_than_50aa_p2	2.7770	2.7862	-0.1207	4.8621	0.4545	Tentative
greater_than_30aa_p1	2.7336	2.7035	-0.1327	4.7099	0.4132	Tentative
greater_than_30aa_p2	2.6980	2.6873	0.1640	5.0577	0.3719	Tentative
disodered_segments_p1	2.0013	2.2358	-0.0124	3.2046	0.0578	Rejected
disodered_segments_p2	1.2287	1.1790	-0.2771	2.5781	0.0082	Rejected

The scores shown above were obtained by testing a subset of 10,000 samples from each positive and negative protein pairs, which means 20,000 protein pairs in total. From all the 43 attributes tested, the features like number of amino-acids or disordered region percentage failed to qualify the test, whereas domain frequency and interaction probability scores performed very well. Similarly Boruta algorithm also tested for 40,000 samples, shown in table table E.

Table E: Boruta analysis of balanced subset of 40,000 training protein pairs

	meanZ	**medianZ**	**minZ**	**maxZ**	**normHits**	**decision**
Min_freq_DDI	70.7829	70.9649	65.2305	74.011	1	Confirmed
btw_P1	40.5735	40.4710	35.8467	44.6133	1	Confirmed
btw_P2	39.3431	38.9904	36.2540	43.1266	1	Confirmed
Max_freq_DDI	35.3506	35.3627	33.8445	36.7156	1	Confirmed
MR	33.4810	34.2939	25.3160	37.7033	1	Confirmed
Min_Zscore	31.1634	31.0913	28.7869	34.3296	1	Confirmed
COR	27.7230	27.8417	22.5308	31.4857	1	Confirmed
nb1_P2	27.6031	27.4768	25.6606	29.7106	1	Confirmed
nb1_P1	27.3765	27.5030	24.9908	29.5615	1	Confirmed
dgr_P2	26.9429	26.8306	25.2152	28.9279	1	Confirmed
Max_Zscore	26.6451	26.8764	23.3484	29.3012	1	Confirmed
dgr_P1	26.3846	26.2192	25.0059	28.7566	1	Confirmed
Zscore_of_mostfreq_DDI	24.9415	25.2390	20.1361	28.3939	1	Confirmed
EvCV_P1	23.2232	23.2909	21.5356	25.1090	1	Confirmed
EvCV_P2	22.1707	22.1386	20.5202	23.7938	1	Confirmed
nb3_P2	21.7620	21.8946	18.7462	24.9241	1	Confirmed
nb3_P1	21.3202	21.4794	19.1779	23.3595	1	Confirmed
CC_P1	21.0002	20.9110	19.7514	22.2395	1	Confirmed
CC_P2	20.8715	20.9203	18.8872	22.8138	1	Confirmed
nb2_P2	20.8104	20.9022	19.4528	22.0838	1	Confirmed
nb4_P2	20.5999	20.6136	19.1195	22.4359	1	Confirmed
nb4_P1	20.5641	20.8727	18.8923	22.2365	1	Confirmed
nb2_P1	20.5043	20.4122	19.2228	22.8296	1	Confirmed
nb5_P2	19.9141	19.9581	18.2968	21.6909	1	Confirmed
nb5_P1	19.2840	19.3426	17.0462	21.3040	1	Confirmed

numDDI	16.8599	16.9867	13.1143	20.0232	1	Confirmed
numDomainP1	12.3790	12.3342	10.0122	14.5476	1	Confirmed
totalAA_p1	11.3611	11.4002	8.8062	13.2869	1	Confirmed
numDomainP2	11.0648	11.0732	8.5690	13.2446	1	Confirmed
cls_P2	9.4823	9.4748	7.6232	11.4344	1	Confirmed
cls_P1	8.9993	9.0747	8.1565	9.9931	1	Confirmed
ecc_P1	8.2658	8.2467	7.1144	9.2619	1	Confirmed
disorder_p1	8.2124	8.5463	2.4914	11.7818	1	Confirmed
totalAA_p2	8.2034	8.2404	6.4514	11.0636	1	Confirmed
ecc_P2	7.4781	7.3311	6.6412	8.4872	1	Confirmed
disorder_p2	7.1540	7.4243	2.5201	9.9458	0.9791	Confirmed
Eigen_value	7.0418	6.9905	4.6064	9.5802	1	Confirmed
greater_than_50aa_p2	4.9315	5.0178	2.8754	6.6550	1	Confirmed
greater_than_30aa_p1	4.5612	4.5586	1.5119	6.4407	0.9791	Confirmed
disordered_segments_p1	4.3247	4.5559	1.4684	7.4819	0.8958	Confirmed
greater_than_50aa_p1	4.3118	4.3443	1.9915	6.0095	0.9791	Confirmed
greater_than_30aa_p2	4.0057	4.0736	1.7924	5.8105	0.9375	Confirmed
disordered_segments_p2	3.9973	4.07150	2.12387	6.18051	0.9375	Confirmed

Table F: Performance of machine learning (in %) for sample with 4000 protein pairs in balanced class.

Method used	**Precision**	**Recall**	**Standard Deviation for precision**	**Standard Deviation for Recall**
Nearest Neighbors	90.69	82.26	4.74	13.7
Linear SVM	92.93	85.58	3.92	16.8
RBF SVM	92.31	83.33	3.74	15.11
Decision Tree	94.83	97.04	1.44	1.19
Random Forest	95.15	97.34	1.54	1.43
AdaBoost	91.74	82.26	4.94	13.7
Naive Bayes	91.05	79.35	5.08	18.8
LDA	90.66	79.35	4.96	17.98
QDA	90.99	75.35	4.82	18.57
MLP	87.39	75.12	1.59	3.71

Table G: Performance of machine learning (in %) for sample with 8000 protein pairs in balanced class

Method used	Precision	Recall	Standard Deviation for precision	Standard Deviation for Recall
Nearest Neighbors	90.21	82.71	5.02	13.28
Linear SVM	92.64	86.44	4.04	15.96
RBF SVM	92.02	83.93	3.81	14.54
Decision Tree	94.92	97.14	0.82	1.22
Random Forest	95.34	97.67	1.06	1.1
AdaBoost	91.45	85.36	5.37	13.5
Naive Bayes	90.83	79.86	5.23	18.38
LDA	90.35	78.33	5.11	17.69
QDA	90.72	76.01	4.94	17.96
MLP	87.95	77.37	2.22	3.71

Table H: Performance of machine learning (in %) for sample with 20000 protein pairs in balanced class

Method used	Precision	Recall	Standard Deviation for precision	Standard Deviation for Recall
Nearest Neighbors	91.12	84.52	4.88	11.88
Linear SVM	93.15	87.93	4.73	14.29
RBF SVM	92.66	85.51	4.21	13.08
Decision Tree	96.37	97.62	1.09	0.62
Random Forest	96.4	98.02	0.83	0.66
AdaBoost	92.21	86.89	5.08	12.06
Naive Bayes	91.63	81.11	4.94	18.02
LDA	91.12	79.51	4.83	17.39
QDA	91.53	76.83	4.7	18.07
MLP	89.55	80.44	1.99	3.71

Table I: Performance of machine learning (in %) for sample with 50000 protein pairs in balanced class

Method used	**Precision**	**Recall**	**Standard Deviation for precision**	**Standard Deviation for Recall**
Nearest Neighbors	91.23	85.52	4.83	11.23
Linear SVM	93.38	89.05	4.59	13.35
RBF SVM	92.78	86.72	4.12	12.26
Decision Tree	96.56	98.14	0.22	0.33
Random Forest	96.61	98.47	0.24	0.44
AdaBoost	92.22	87.66	4.93	11.32
Naive Bayes	91.66	81.94	4.76	17.51
LDA	91.17	80.22	4.65	12.02
QDA	91.54	77.63	4.51	17.63
MLP	87.28	85.13	2.09	3.21

Table J: Precision values obtained under different class ratio testing (all values, %)

Sample ratio	**Decision Tree**	**SD (+/-)**	**Random Forest**	**SD (+/-)**	**Linear SVM**	**SD (+/-)**
1:1	94.92	0.77	95.34	1.01	92.64	3.74
1:5	94.25	0.77	94.61	1.01	92.75	2.92
1:10	92.63	1.63	93.61	1.71	93.88	1.74

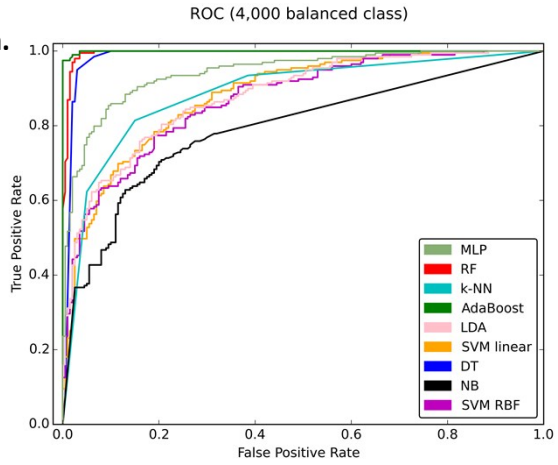
Table K: Recall values obtained under different class ratio testing (all values, %)

Sample ratio	**Decision Tree**	**SD (+/-)**	**Random Forest**	**SD (+/-)**	**Linear SVM**	**SD (+/-)**
1:1	97.14	1.19	97.67	1.07	86.44	13.80
1:5	91.47	1.19	92.41	1.07	75.91	17.80
1:10	87.77	2.37	88.79	2.20	68.23	19.90

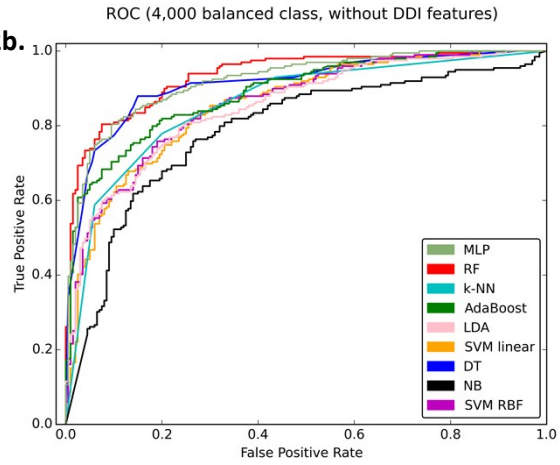
Table L. AUC values with and without DDI features for 4000 and 50000 samples

	4000 Samples		50000 Samples	
ML Methods	**All features**	**Without DDI features**	**All features**	**Without DDI features**
Random Forest (RF)	99.51%	93.42%	98.92%	91.99%
Nearest Neighbors (k-NN)	88.78%	85.91%	88.40%	86.41%
AdaBoost	99.93%	88.77%	98.92%	91.99%
LDA	88.02%	85.24%	88.24%	86.63%
linear SVM	88.11%	85.52%	90.48%	87.74%
Decision Tree (DT)	98.45%	91.26%	98.06%	90.98%
Naïve Bayes (NB)	79.21%	78.81%	80.81%	82.86%
RBF SVM	87.03%	86.00%	90.17%	88.64%
Multilayer Perceptron (MLP)	94.26%	92.15%	95.55%	92.34%

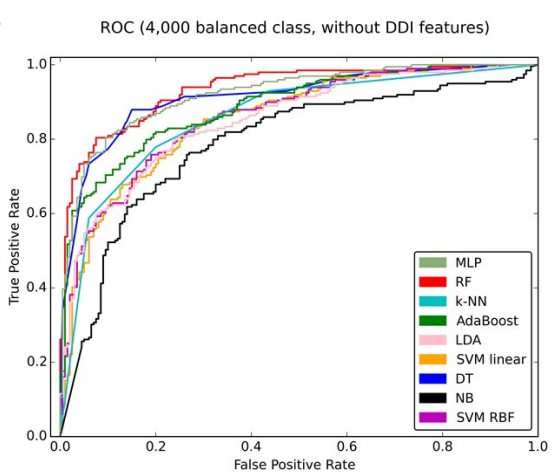
2a.



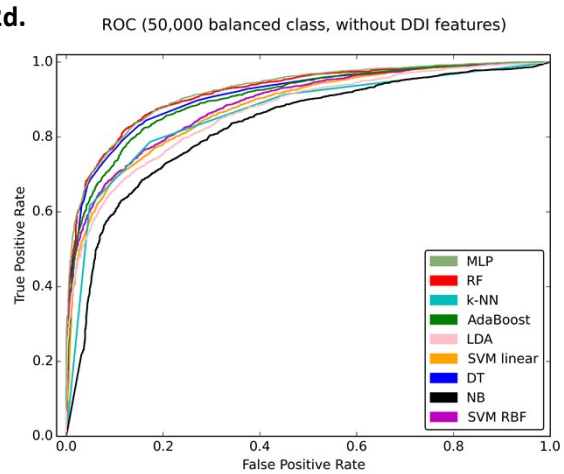
2b.



2c.



2d.



Supplementary figure 2: Comparing ROC curves for balanced class with and without DDI features.
To establish the effect of DDI features on machine learning performance for low (2a, 2b) and high (2c 2d) sample size (4000 and 50000 protein pair of balance class respectively). Figure 2a and 2c shows ROC curves for ML methods using all the features whereas figure 2b and 2d shows ROC curves for ML methods after removing DDI features. The latter case shows the decrease in AUC values. Refer to table M for values.

The parameters used in the classification methods are reported below:

1) **Decision Tree:**
   Criterion (function to measure the quality of a tree split): *Gini function*
   Maximum tree depth: *5*
   Minimum number of samples to split an internal node: *2*
   Minimum number of samples in newly created leaves: *1*
   Strategy to choose the split at each node: *best split*

2) **Random Forest:**
   Criterion: *Gini function*
   Maximum tree depth: *5*
   Minimum number of samples to split an internal node: *2*
   Minimum number of samples in newly created leaves: *1*
   Strategy to choose the split at each node: *best split*
   Number of estimators: *40*
   Bootstrap samples (boolean): True

3) **AdaBoost:**
   Base estimators: *Decision Trees (with DT parameters as in point 1)*
   Number of estimators: *40*
   Learning Rate Coefficient: *1*

4) **Linear SVM:**
   Kernel: linear
   Penalty parameter C of the error term: *1*
   Loss function: *squared hinge*
   Penalization norm: *L2*
   Shrinking heuristic (boolean): *True*
   Tolerance for stopping criteria: *1e-4*
   Maximum number of iterations: *No limit*

5) **RBF SVM:**
   Kernel: *Radial Basis Function (RBF)*

Penalty parameter C of the error term: *1*
Gamma coefficient for RBF: *1/number of features*
Shrinking heuristic (boolean): *True*
Tolerance for stopping criteria: *1e-4*
Maximum number of iterations: *No limit*

6) **k-NN:**
Number of neighbours: *5*
Distance metric: *Minkowski*
Power parameter for the Minkowski metric: *2 (Euclidean metric)*
Prediction weights: *uniform (all points in each neighborhood are weighted equally)*
Algorithm: *Optimized selection of Ball-tree or KD-tree algorithms*
Leaf size: *30*

7) **Linear Discriminant Analysis:**
Solver: *Singular value decomposition(SVD)*
Threshold used for SVD rank estimation: *0.0001*

8) **Naïve Bias:**
Method*: Gaussian Naïve Bias*

9) **MLP (Multi-layer preceptron):**
Learning_rule: *stochastic gradient descend (sgd)*
Learning_rate: *0.001*
Learning_momentum: *0.9*
Number of iterations: *25*
N_stable (number of interations after which training should return when the validation error remains (near) constant): *10*
F_stable (threshold under which the validation error change is assumed to be stable, to be used in combination with *n_stable*): *0.001*
MLP Layers:

1. *Maxout layer:*
   1.1. *Number of neurons: 50*
   1.2. *Number of piecewise linear segments in the Maxout activation: 2*
2. *Softmax output layer activation*