

Supplementary Information:
Better Network Modelling for Link Prediction
in Protein-Protein Interaction Networks

Ho Yin Yuen, Jesper Jansson

	CN	CRA	L3	CH2_L3	Sim	L3E'(f_1)	L3E'(f_2)
BioGRID Yeast							
50% PPIs removed	4e-27	1e-29	1e-34	3e-34	4e-26	1e-31	5e-29
40% PPIs removed	1e-28	2e-29	1e-34	4e-31	4e-30	6e-33	2e-27
30% PPIs removed	5e-27	7e-29	6e-29	10e-33	10e-28	8e-30	3e-26
20% PPIs removed	8e-26	3e-24	1e-25	8e-24	2e-26	2e-25	10e-26
10% PPIs removed	1e-20	2e-22	1e-21	4e-16	3e-23	1e-23	1e-22
STRING Yeast							
50% PPIs removed	9e-38	5e-42	5e-37	1e-33	3e-36	3e-39	2e-36
40% PPIs removed	6e-38	1e-37	2e-37	6e-32	2e-36	2e-36	2e-31
30% PPIs removed	1e-35	6e-36	2e-33	9e-29	4e-35	4e-35	6e-31
20% PPIs removed	8e-33	7e-34	7e-31	3e-29	9e-31	5e-31	1e-30
10% PPIs removed	10e-31	6e-32	1e-29	1e-29	6e-30	2e-30	3e-30
MINT Yeast							
50% PPIs removed	1e-28	4e-25	1e-28	2e-28	2e-24	8e-28	3e-26
40% PPIs removed	1e-27	2e-26	1e-26	7e-19	5e-25	6e-30	4e-21
30% PPIs removed	7e-24	3e-24	8e-25	1e-20	1e-20	1e-22	3e-21
20% PPIs removed	1e-20	6e-20	3e-22	9e-18	3e-21	6e-21	2e-17
10% PPIs removed	3e-18	6e-20	2e-16	1e-16	2e-16	5e-16	2e-17

Table S1: The link predictors' p-values against negative control (the random link predictor) in terms of PR AUC for varying sample sizes of the yeast datasets (ten trials).

	CN	CRA	L3	CH2_L3	Sim	L3E'(f_1)	L3E'(f_2)
BioGRID Human							
50% PPIs removed	2e-19	1e-26	6e-24	6e-30	2e-32	4e-24	3e-25
40% PPIs removed	8e-28	4e-31	7e-19	2e-28	4e-31	1e-27	2e-24
30% PPIs removed	2e-27	9e-32	3e-29	7e-26	8e-16	1e-29	7e-32
20% PPIs removed	3e-28	5e-27	4e-21	5e-23	1e-24	2e-28	9e-26
10% PPIs removed	3e-20	1e-17	8e-21	2e-21	3e-27	2e-25	7e-21
STRING Human							
50% PPIs removed	3e-34	5e-39	1e-35	4e-29	5e-35	1e-35	2e-36
40% PPIs removed	5e-37	2e-40	9e-39	3e-37	1e-36	2e-38	4e-38
30% PPIs removed	8e-35	9e-37	1e-37	2e-34	7e-37	3e-39	2e-42
20% PPIs removed	3e-32	9e-36	4e-34	2e-34	4e-35	1e-38	2e-37
10% PPIs removed	7e-28	9e-30	9e-30	8e-30	2e-30	1e-31	4e-28
MINT Human							
50% PPIs removed	4e-16	1e-13	3e-24	1e-24	8e-18	9e-26	7e-22
40% PPIs removed	5e-15	1e-16	6e-22	3e-19	2e-21	2e-23	4e-24
30% PPIs removed	2e-16	1e-19	7e-23	1e-20	2e-20	2e-22	1e-20
20% PPIs removed	3e-17	7e-24	2e-23	2e-19	4e-18	4e-22	2e-20
10% PPIs removed	1e-15	5e-15	3e-15	3e-14	4e-12	2e-17	2e-16
HuRI Human							
50% PPIs removed	1e-23	1e-25	9e-30	8e-27	2e-25	3e-28	3e-25
40% PPIs removed	2e-27	2e-27	5e-32	4e-27	5e-28	5e-28	9e-25
30% PPIs removed	1e-23	1e-23	7e-27	1e-23	4e-24	6e-24	4e-22
20% PPIs removed	1e-21	2e-21	2e-22	5e-21	1e-22	2e-23	1e-25
10% PPIs removed	3e-20	2e-18	1e-24	2e-22	9e-20	2e-20	5e-18

Table S2: The link predictors' p-values against negative control (the random link predictor) in terms of PR AUC for varying sample sizes of the human datasets (ten trials).

AUC of PR for rand predictor					
dataset \ % of PPIs removed	50%	40%	30%	20%	10%
BioGRID Yeast	7.582e-06	4.726e-06	2.267e-06	1.262e-06	3.027e-07
STRING Yeast	2.494e-05	1.529e-05	8.408e-06	3.908e-06	9.452e-07
MINT Yeast	2.247e-06	1.009e-06	5.536e-07	4.933e-07	0.e+00
BioGRID Human	9.917e-07	8.432e-07	6.524e-07	4.690e-07	2.484e-07
STRING Human	2.805e-06	2.51e-06	2.052e-06	1.497e-06	8.719e-07
MINT Human	3.848e-07	3.310e-07	3.485e-07	2.235e-07	1.273e-07
HuRI	1.226e-06	9.521e-07	8.145e-07	3.698e-07	4.853e-07

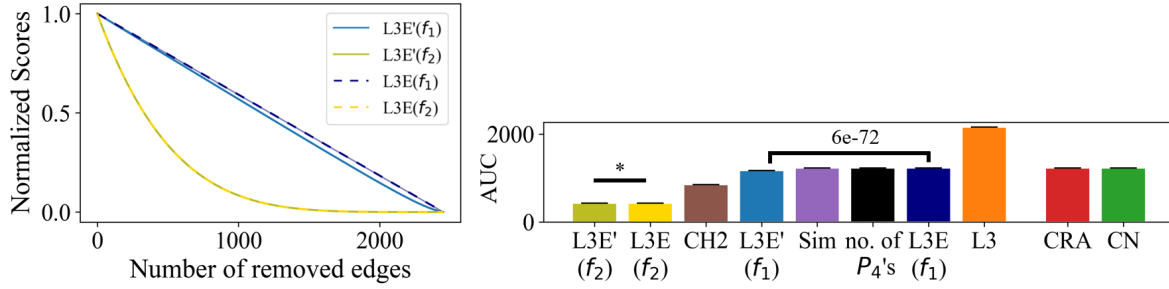
Table S3: The rand link predictor's PR AUC-values after removing varying numbers of PPIs from the datasets (ten trials for each). The table shows that it is more difficult to pick a real PPIs among the candidate PPIs as the percentage of removed PPIs decreases.

	CN-based	L3-based	CN & L3-based	CRA & L3E'(f_1)
BioGRID Yeast				
50% PPIs removed	69%	79 $\pm$ 10 %	30 $\pm$ 6 %	35%
40% PPIs removed	73%	64 $\pm$ 15 %	30 $\pm$ 9 %	41%
30% PPIs removed	74%	64 $\pm$ 14 %	31 $\pm$ 11 %	45%
20% PPIs removed	74%	59 $\pm$ 16 %	31 $\pm$ 13 %	46%
10% PPIs removed	76%	52 $\pm$ 18 %	31 $\pm$ 16 %	45%
STRING Yeast				
50% PPIs removed	89%	92 $\pm$ 2 %	72 $\pm$ 6 %	74%
40% PPIs removed	91%	81 $\pm$ 15 %	66 $\pm$ 15 %	78%
30% PPIs removed	93%	79 $\pm$ 15 %	68 $\pm$ 16 %	81%
20% PPIs removed	94%	77 $\pm$ 17 %	68 $\pm$ 18 %	83%
10% PPIs removed	94%	75 $\pm$ 16 %	68 $\pm$ 18 %	84%
MINT Yeast				
50% PPIs removed	43%	72 $\pm$ 8 %	32 $\pm$ 2 %	34%
40% PPIs removed	74%	65 $\pm$ 12 %	30 $\pm$ 6 %	35%
30% PPIs removed	61%	64 $\pm$ 14 %	39 $\pm$ 8 %	52%
20% PPIs removed	30%	66 $\pm$ 13 %	45 $\pm$ 10 %	57%
10% PPIs removed	75%	67 $\pm$ 13 %	49 $\pm$ 11 %	63%

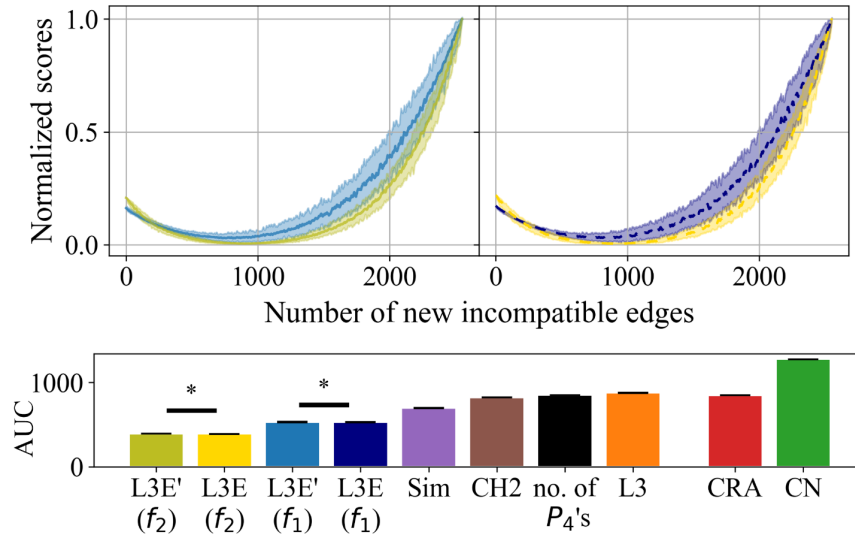
Table S4: Overlap ratios of predicted PPIs between different types of link predictors for the yeast datasets across different sample sizes. 'CN-based' and 'CRA & L3E'(f_1)' denotes the overlap ratio of the predicted PPIs between CN and CRA, and between CRA and L3E'(f_1) respectively. For 'L3-based', since there are multiple L3-based predictors (L3, CH2, Sim, L3E'(f_1), and L3E'(f_2)), we calculated the overlap ratio for each pair of predictors. We then took the mean of these ratios as the final value, and also computed the standard deviation. The same applies to 'CN & L3-based' where a CN predictor is compared to a L3-based predictor. Blue color denotes a relatively higher overlap ratio and red a relatively smaller overlap. Ratios are rounded to nearest integers.

	CN-based	L3-based	CN & L3-based	CRA & L3E'(f_1)
BioGRID Human				
50% PPIs removed	64%	69 $\pm$ 4 %	38 $\pm$ 4 %	37%
40% PPIs removed	67%	70 $\pm$ 4 %	43 $\pm$ 3 %	44%
30% PPIs removed	67%	62 $\pm$ 11 %	44 $\pm$ 6 %	48%
20% PPIs removed	64%	61 $\pm$ 6 %	44 $\pm$ 4 %	50%
10% PPIs removed	64%	57 $\pm$ 7 %	43 $\pm$ 7 %	52%
STRING Human				
50% PPIs removed	54%	58 $\pm$ 4 %	44 $\pm$ 3 %	44%
40% PPIs removed	54%	60 $\pm$ 5 %	46 $\pm$ 5 %	55%
30% PPIs removed	54%	56 $\pm$ 2 %	45 $\pm$ 3 %	51%
20% PPIs removed	56%	60 $\pm$ 8 %	46 $\pm$ 4 %	49%
10% PPIs removed	59%	55 $\pm$ 2 %	48 $\pm$ 4 %	52%
MINT Human				
50% PPIs removed	37%	71 $\pm$ 10 %	4 $\pm$ 2 %	5%
40% PPIs removed	49%	71 $\pm$ 9 %	5 $\pm$ 2 %	6%
30% PPIs removed	39%	69 $\pm$ 10 %	8 $\pm$ 3 %	11%
20% PPIs removed	31%	65 $\pm$ 11 %	10 $\pm$ 5 %	17%
10% PPIs removed	32%	56 $\pm$ 15 %	11 $\pm$ 6 %	18%
HuRI				
50% PPIs removed	64%	79 $\pm$ 7 %	24 $\pm$ 3 %	23%
40% PPIs removed	70%	79 $\pm$ 7 %	26 $\pm$ 3 %	26%
30% PPIs removed	69%	78 $\pm$ 7 %	29 $\pm$ 3 %	29%
20% PPIs removed	67%	76 $\pm$ 7 %	31 $\pm$ 4 %	33%
10% PPIs removed	61%	71 $\pm$ 9 %	36 $\pm$ 6 %	42%

Table S5: Overlap ratios of predicted PPIs between different types of link predictors for the human datasets across different sample sizes. 'CN-based' and 'CRA & L3E'(f_1)' denotes the overlap ratio of the predicted PPIs between CN and CRA, and between CRA and L3E'(f_1) respectively. For 'L3-based', since there are multiple L3-based predictors (L3, CH2, Sim, L3E'(f_1), and L3E'(f_2)), we calculated the overlap ratio for each pair of predictors. We then took the mean of these ratios as the final value, and also computed the standard deviation. The same applies to 'CN & L3-based' where a CN predictor is compared to a L3-based predictor. Blue color denotes a relatively higher overlap ratio and red a relatively smaller overlap. Ratios are rounded to nearest integers.



(a) Changes in prediction scores due to the removal of compatible edges in ideal L3 graphs



(b) Changes in prediction scores due to the addition of incompatible edges in ideal L3 graphs

Figure S1: Changes in scores for L3E link predictors when an ideal L3 graph is modified by: (a) removing compatible edges; and (b) adding incompatible edges. The shaded regions denote the variance (the minimum and maximum values) among replication experiments, and the solid lines denote the medians. The AUC bar charts correspond to the respective plots. The statistical bars denote the statistical significance (p-value) between the two corresponding samples using student's t-test. An asterisk denotes $p > 0.05$ and is therefore not statistically significant. In (b), a Savitzky–Golay filter using a polynomial of degree 3 and a window size of 21 was applied to make the curves smoother.

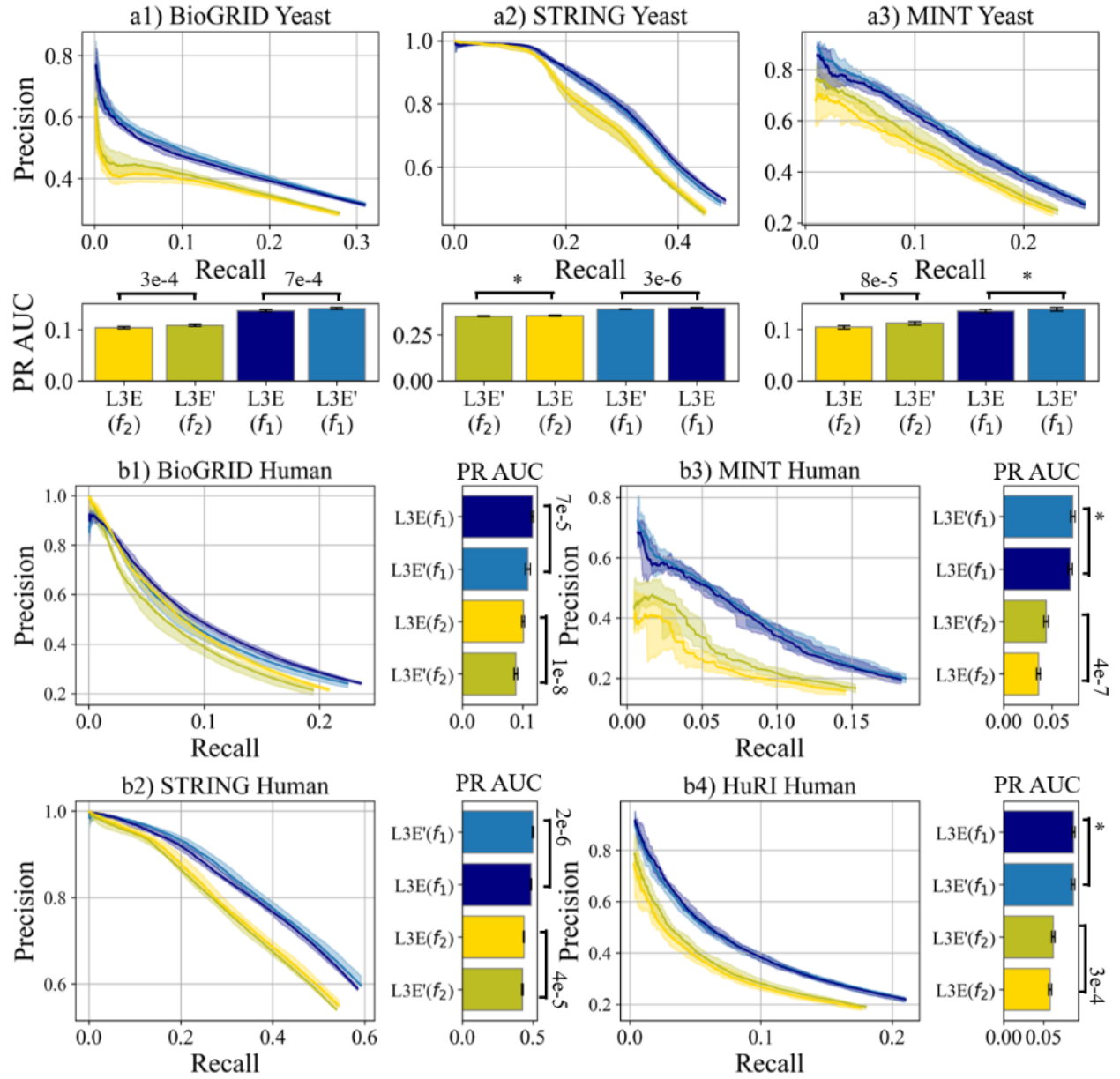


Figure S2: Precision-Recall (PR) curves of the L3E link predictors computed with **50% of the PPIs removed** in every dataset. The solid lines show the median values and the shaded regions indicate the variance (the minimum and maximum values). The accompanying bar charts show the predictors' AUC-values (the larger, the better). The statistical bars denote the statistical significance between two samples in terms of the p-value using student's t-test. An asterisk denotes $p > 0.05$ and is therefore not statistically significant.

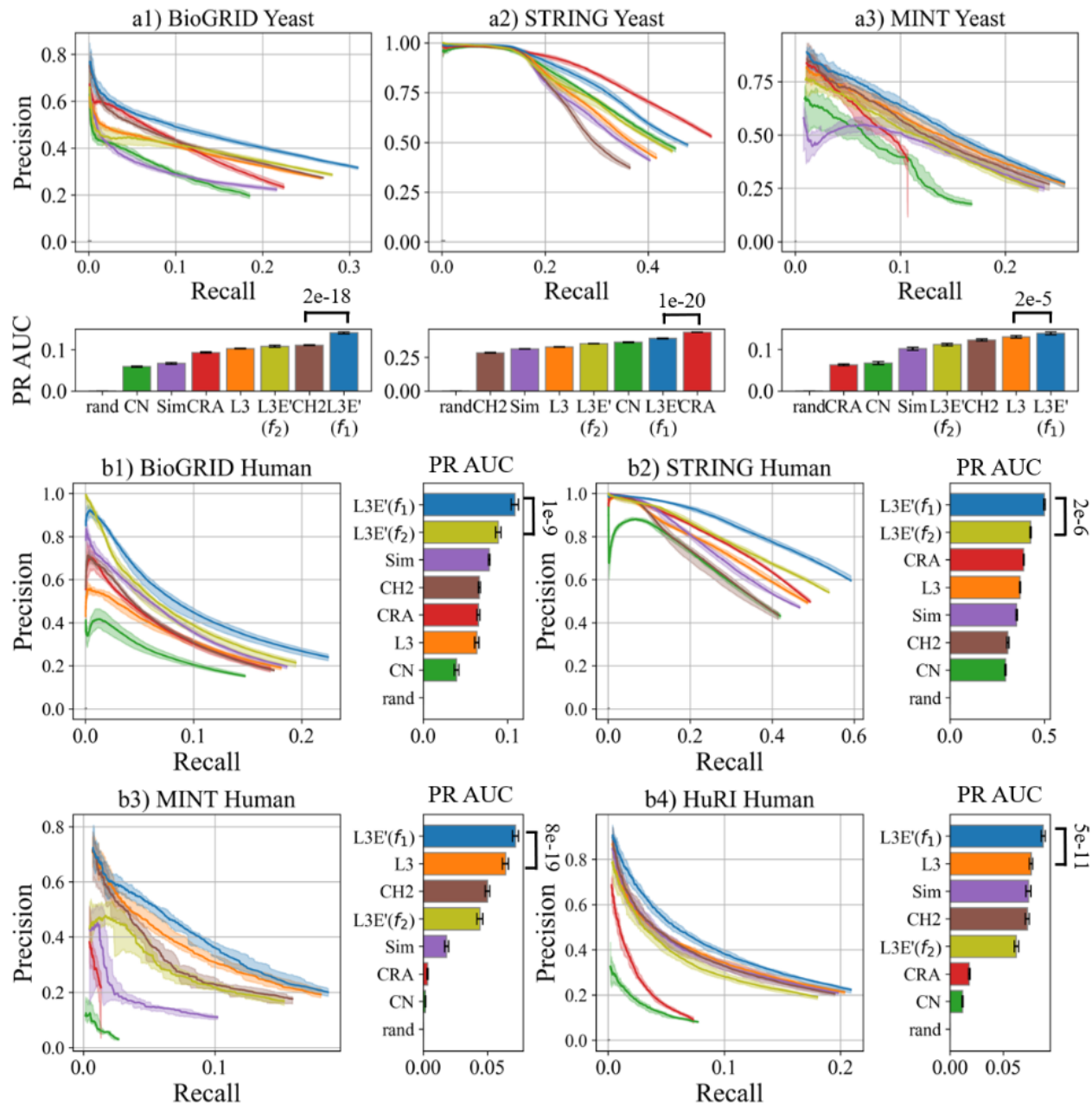


Figure S3: Precision-Recall (PR) curves of the link predictors computed with 50% of the PPIs removed in every dataset. (See the caption of Fig. S2 for explanations.)

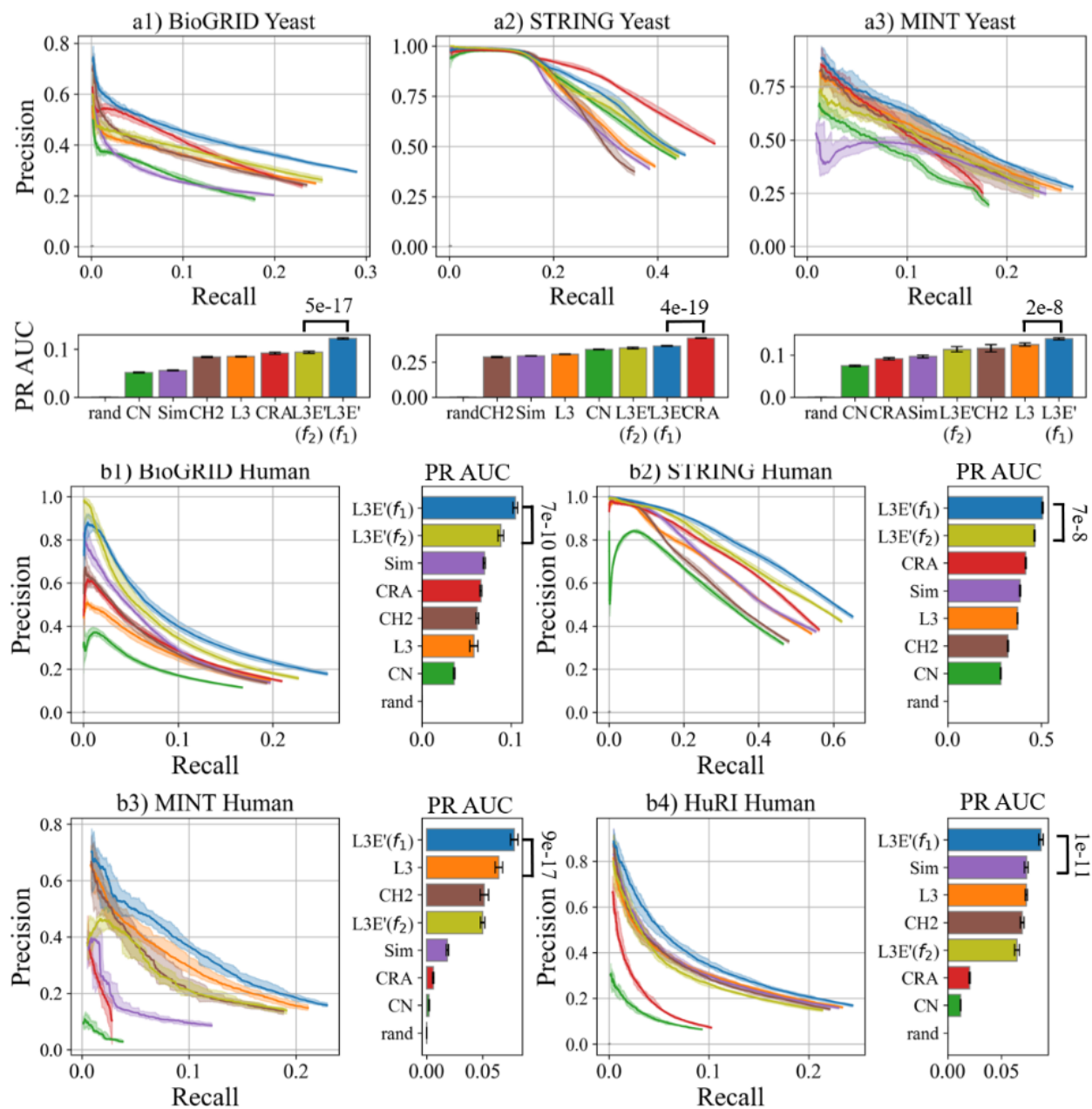


Figure S4: Precision-Recall (PR) curves of the link predictors computed with 40% of the PPIs removed in every dataset. (See the caption of Fig. S2 for explanations.)

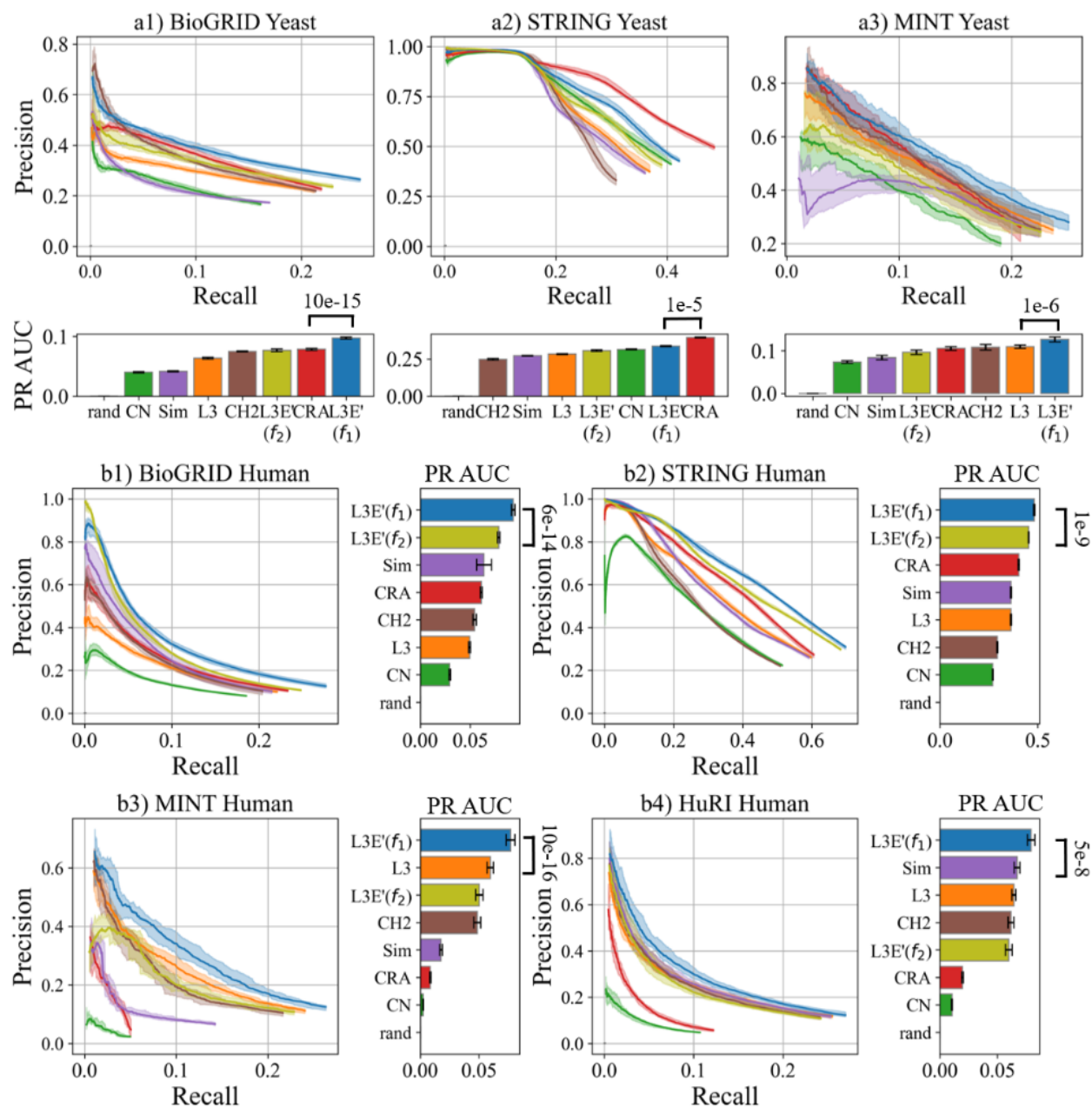


Figure S5: Precision-Recall (PR) curves of the link predictors computed with 30% of the PPIs removed in every dataset. (See the caption of Fig. S2 for explanations.)

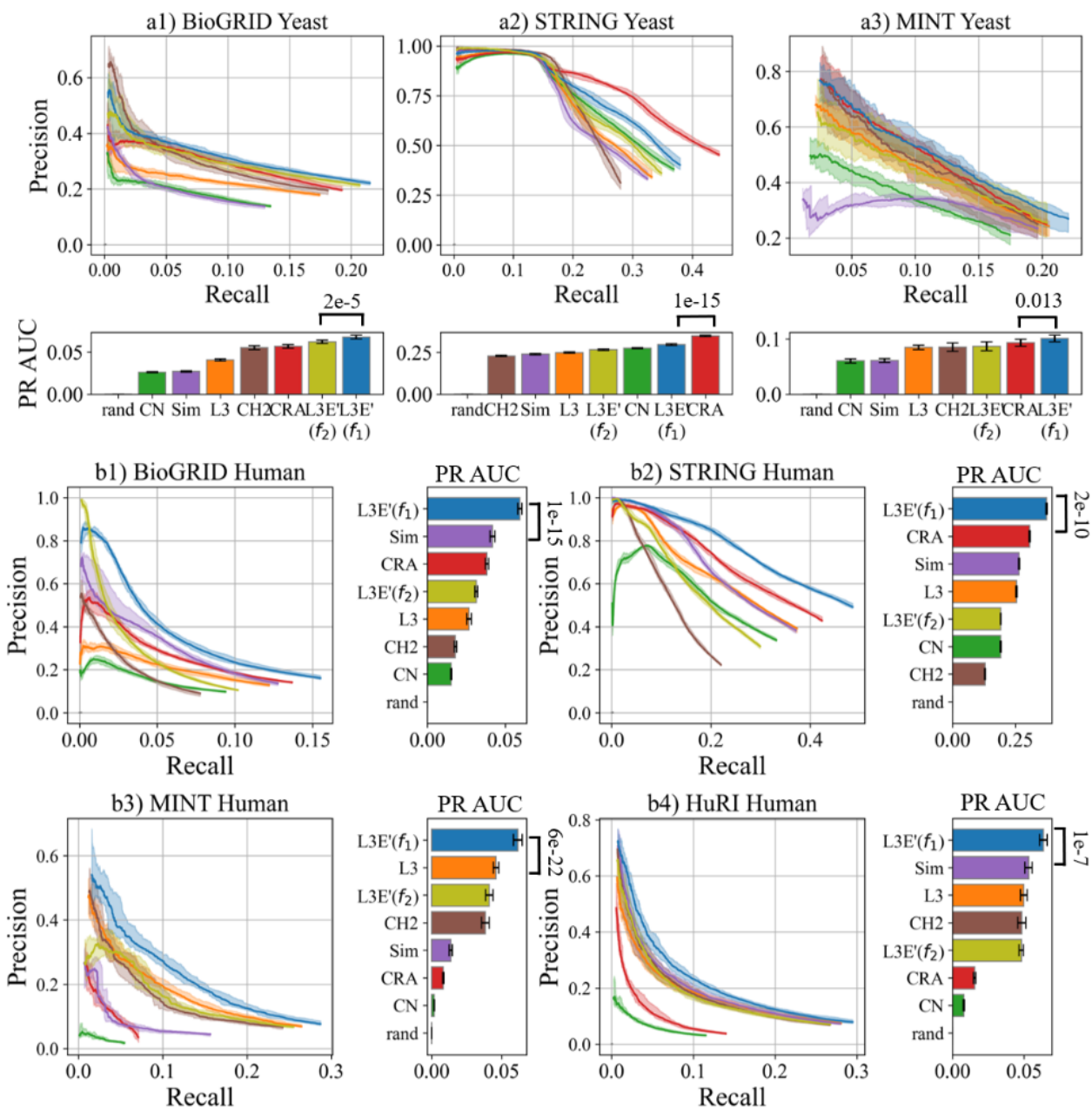


Figure S6: Precision-Recall (PR) curves of the link predictors computed with 20% of the PPIs removed in every dataset. (See the caption of Fig. S2 for explanations.)

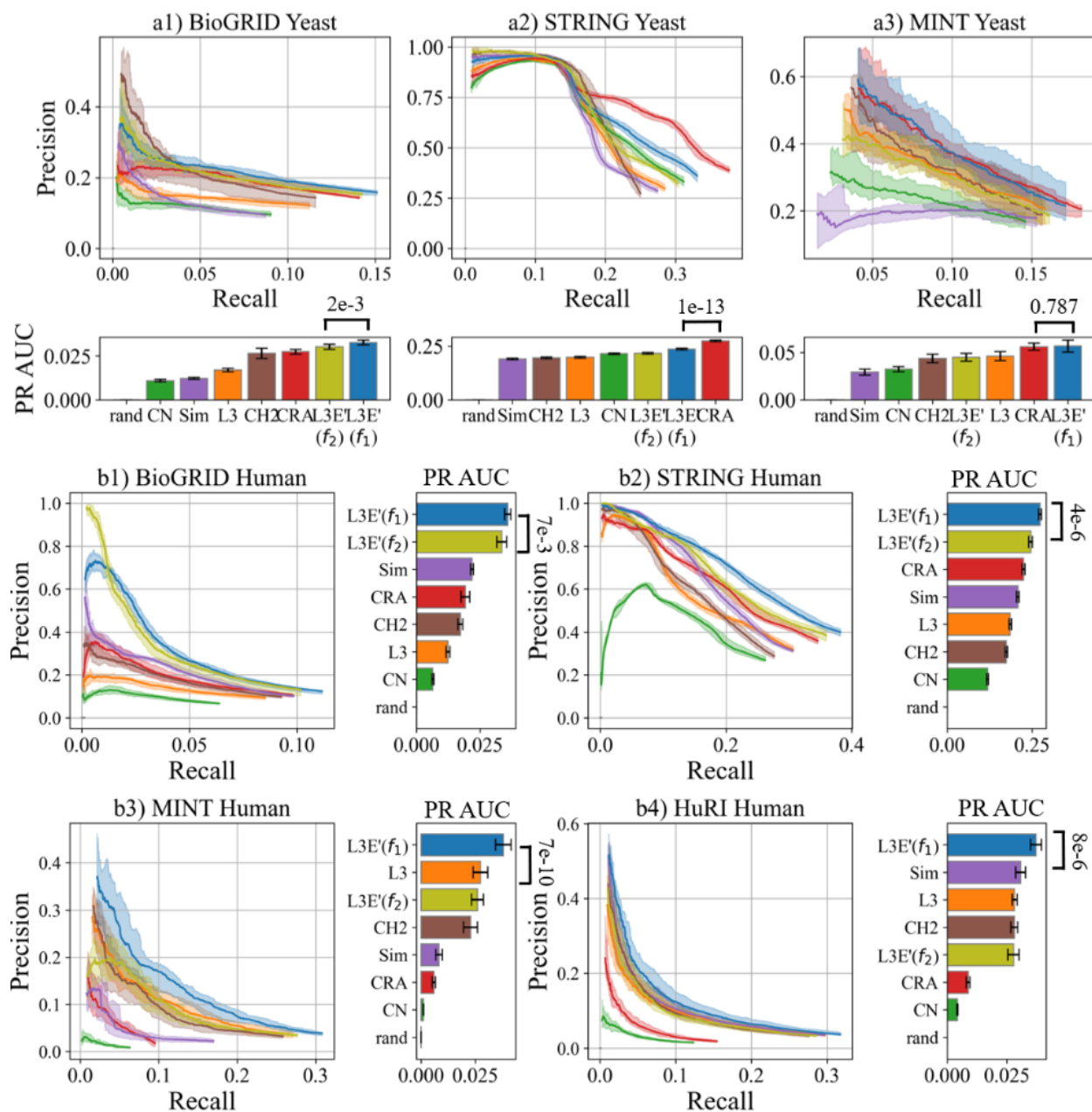


Figure S7: Precision-Recall (PR) curves of the link predictors computed with 10% of the PPIs removed in every dataset. (See the caption of Fig. S2 for explanations.)

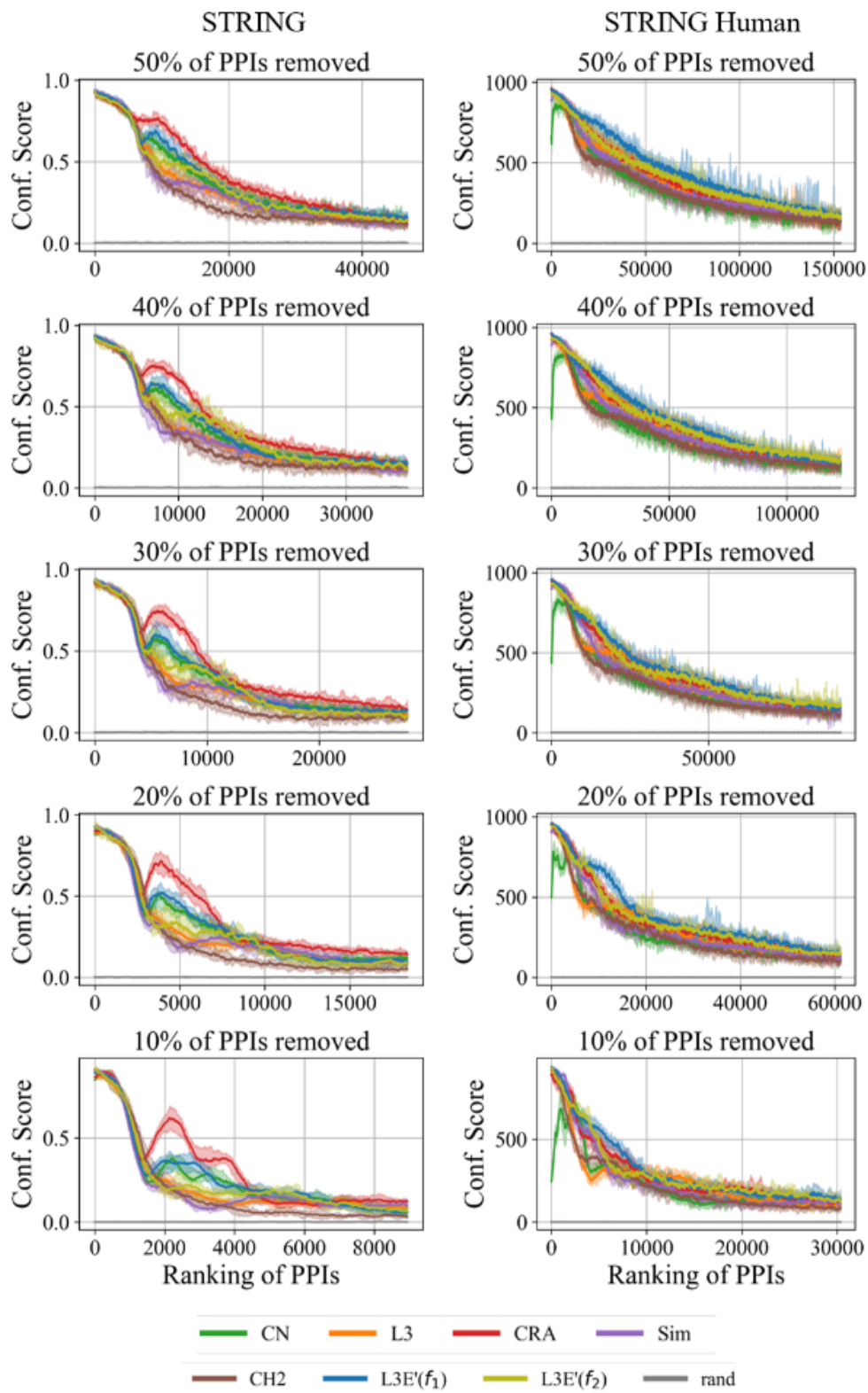


Figure S8: The moving means of the STRING confidence scores across different sample sizes using a window size of 100, with 10 steps forward in each iteration. The shaded regions illustrate the variance (the minimum and maximum values) in STRING confidence scores.

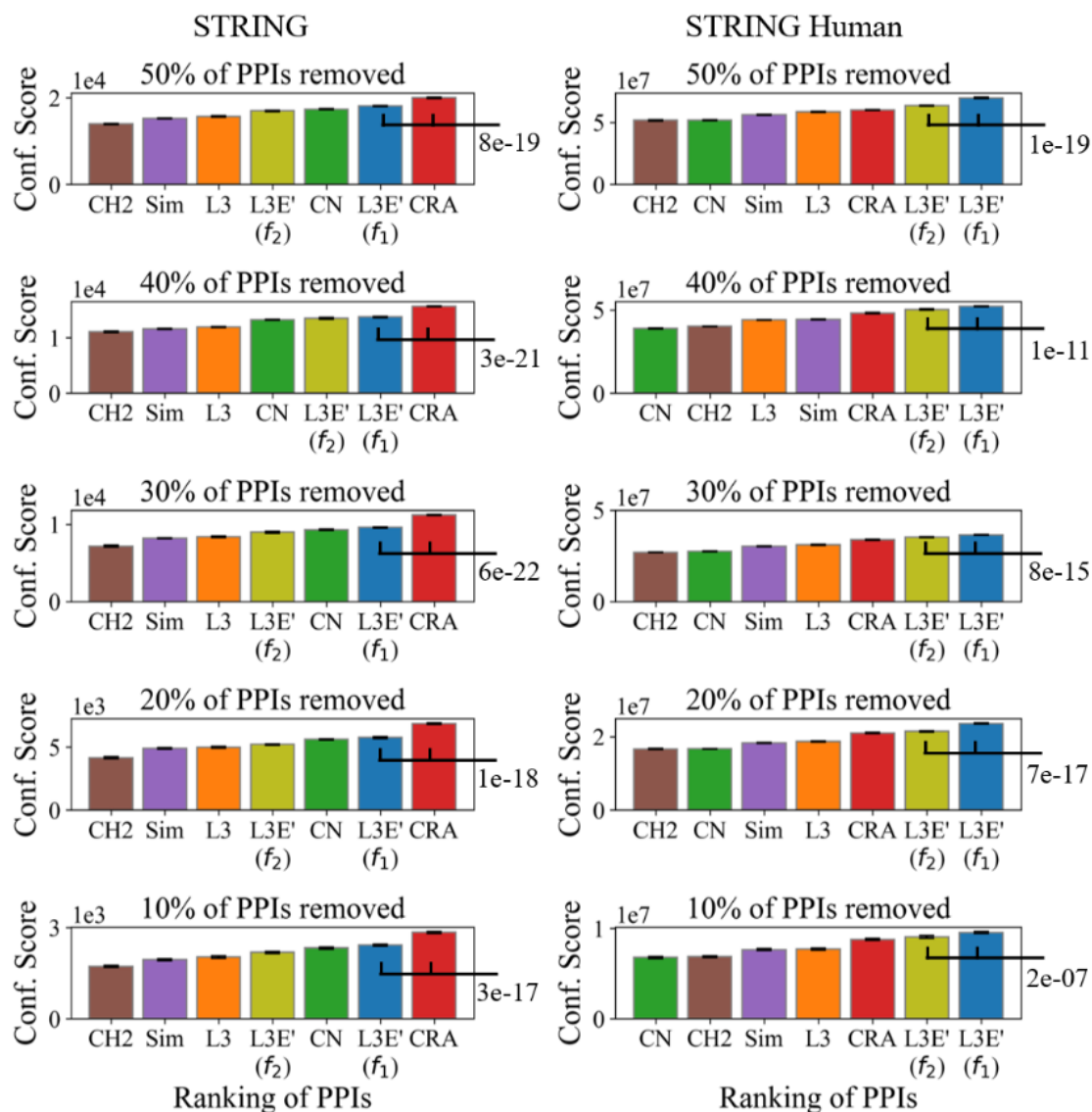


Figure S9: AUCs corresponding to the curves in Fig. S8. The predictors are sorted from left to right (worst to best) according to their AUC-values. Statistical bars denote the statistical significance between two samples in terms of p-value using student's t-test.

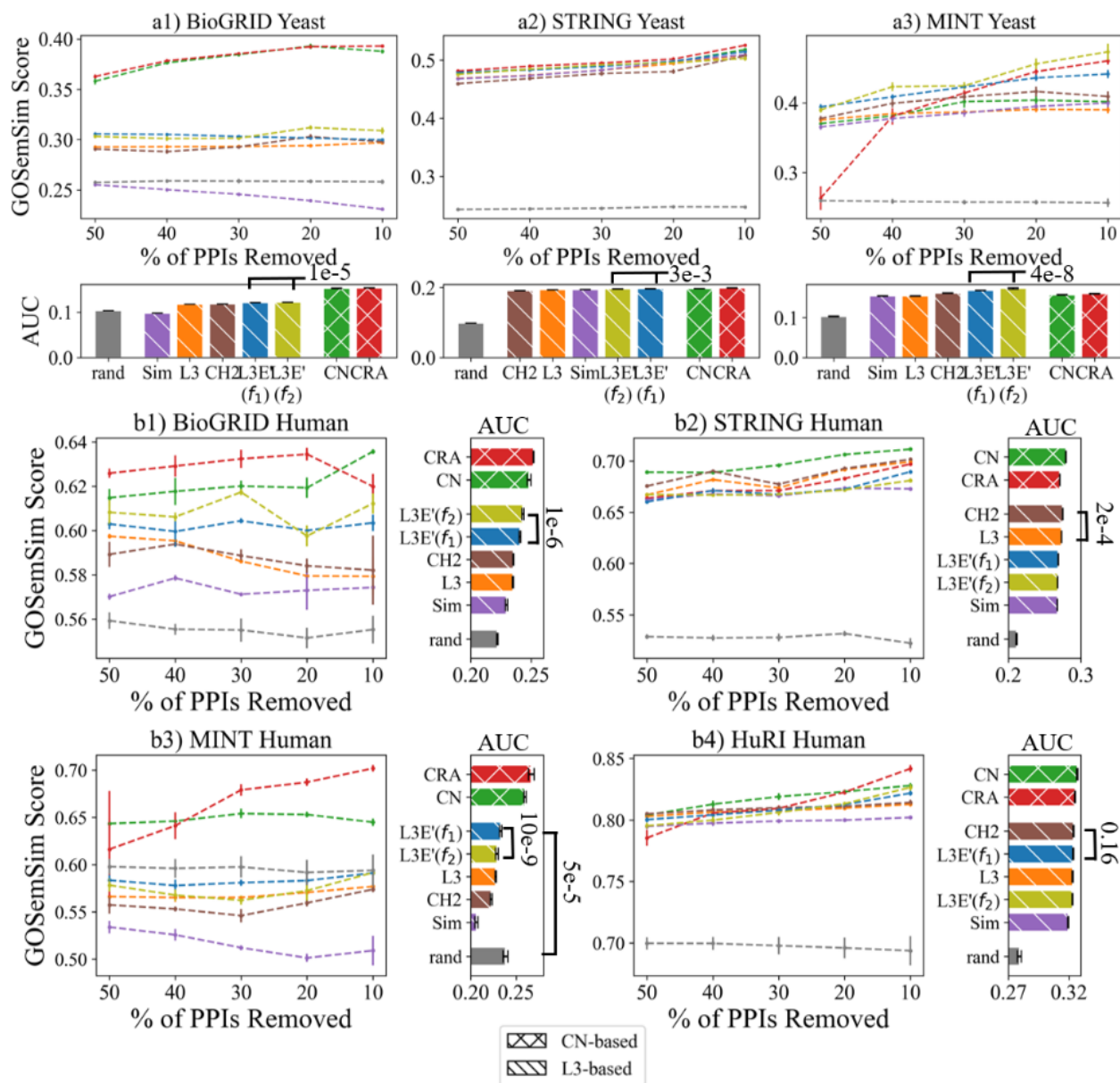


Figure S10: Illustrating how the mean GOSemSim score of the link predictors changes as the percentage of removed PPIs decreases. The dotted curves are interpolations of the data points (50%, 40%, 30%, 20%, 10%) and the vertical bars on the data points indicate the variance (the minimum and maximum values). The gray curve (rand) is the negative control. The bar charts show the AUCs of the GOSemSim scores and the statistical bars give the statistical significance between two samples in terms of the p-value using student's t-test.