Majority Vote Cascading:
A Semi-Supervised Framework For Improving Protein Function Prediction.

Supplementary Material

John Lazarsfeld Jonathan Rordríguez Mert Erden Yuelin Liu Lenore J. Cowen

---

Supplementary Table S1

Table S1: Summary of results for *S. cerevisiae* using cascading across label type, prediction method, and CV fold size. We report mean and standard deviation across 10 random splits of testing and training data. A fixed cascading setting (12 rounds, 35% high confidence threshold, **CC confidence function**) is used.

			2-Fold		**4-Fold**		**6-Fold**	
			acc.	F1	acc.	F1	acc.	F1
MIPS1	cDSD-MV		66.71 ± 0.57	41.43 ± 0.29	57.96 ± 0.32	40.43 ± 0.23	49.38 ± 0.31	37.86 ± 0.26
	cDSD-MV	*casc.*	67.46 ± 0.33	41.79 ± 0.18	62.46 ± 0.21	41.85 ± 0.16	58.92 ± 0.36	40.43 ± 0.22
	cDSD-MV-Known		67.65 ± 0.52	41.69 ± 0.25	64.21 ± 3.24	41.89 ± 0.15	62.34 ± 0.23	41.01 ± 0.06
	cDSD-MV-Known	*casc.*	68.11 ± 0.35	41.68 ± 0.18	64.05 ± 0.38	41.84 ± 0.11	60.89 ± 0.44	40.80 ± 0.15
	cDSD-WMV		66.71 ± 0.42	42.07 ± 0.24	57.83 ± 0.52	41.07 ± 0.22	49.06 ± 0.24	38.26 ± 0.22
	cDSD-WMV	*casc.*	68.05 ± 0.35	42.37 ± 0.21	62.67 ± 0.58	42.33 ± 0.28	58.74 ± 0.33	40.51 ± 0.25
	cDSD-WMV-Known		68.31 ± 0.36	42.40 ± 0.31	64.91 ± 0.30	42.74 ± 0.14	63.06 ± 0.28	41.99 ± 0.15
	cDSD-WMV-Known	*casc.*	68.66 ± 0.45	42.40 ± 0.10	64.22 ± 0.39	42.49 ± 0.15	61.05 ± 0.35	41.38 ± 0.13
MIPS2	cDSD-MV		53.50 ± 0.71	32.55 ± 0.22	44.32 ± 0.29	30.30 ± 0.09	37.29 ± 0.38	27.85 ± 0.23
	cDSD-MV	*casc.*	54.43 ± 0.45	32.80 ± 0.29	48.25 ± 0.23	31.32 ± 0.22	43.89 ± 0.27	29.18 ± 0.13
	cDSD-MV-Known		54.70 ± 0.39	32.87 ± 0.23	50.39 ± 0.18	31.70 ± 0.09	48.17 ± 0.19	30.36 ± 0.14
	cDSD-MV-Known	*casc.*	55.30 ± 0.42	33.09 ± 0.19	50.92 ± 0.37	31.70 ± 0.18	47.18 ± 0.36	29.93 ± 0.16
	cDSD-WMV		55.02 ± 0.58	33.57 ± 0.24	45.94 ± 0.47	31.38 ± 0.18	38.03 ± 0.33	28.70 ± 0.15
	cDSD-WMV	*casc.*	55.36 ± 0.64	33.81 ± 0.20	48.73 ± 0.40	31.88 ± 0.19	43.80 ± 0.29	29.39 ± 0.17
	cDSD-WMV-Known		55.85 ± 0.59	33.87 ± 0.19	51.86 ± 0.38	32.93 ± 0.17	49.42 ± 0.21	31.51 ± 0.11
	cDSD-WMV-Known	*casc.*	55.74 ± 0.48	33.98 ± 0.19	51.41 ± 0.29	32.72 ± 0.15	47.63 ± 0.56	30.93 ± 0.15
MIPS3	cDSD-MV		47.95 ± 0.49	27.06 ± 0.22	39.54 ± 0.28	26.76 ± 0.30	32.59 ± 0.28	25.31 ± 0.27
	cDSD-MV	*casc.*	49.34 ± 0.53	27.34 ± 0.24	42.90 ± 0.52	26.94 ± 0.22	38.80 ± 0.34	25.28 ± 0.15
	cDSD-MV-Known		49.15 ± 0.43	27.46 ± 0.17	44.42 ± 0.22	26.84 ± 0.18	41.48 ± 0.26	25.50 ± 0.11
	cDSD-MV-Known	*casc.*	49.51 ± 0.49	27.38 ± 0.19	45.03 ± 0.45	26.72 ± 0.20	41.00 ± 0.43	25.13 ± 0.11
	cDSD-WMV		50.03 ± 0.34	28.41 ± 0.24	41.31 ± 0.35	27.79 ± 0.27	34.07 ± 0.38	26.00 ± 0.27
	cDSD-WMV	*casc.*	50.11 ± 0.62	28.45 ± 0.19	43.77 ± 0.26	27.82 ± 0.18	39.18 ± 0.31	25.93 ± 0.15
	cDSD-WMV-Known		50.70 ± 0.46	28.33 ± 0.11	46.64 ± 0.27	28.13 ± 0.11	43.66 ± 0.33	26.83 ± 0.11
	cDSD-WMV-Known	*casc.*	50.97 ± 0.40	28.66 ± 0.27	45.72 ± 0.44	27.92 ± 0.20	41.72 ± 0.41	26.29 ± 0.15
GO-MFBP-4	cDSD-MV		47.41 ± 0.59	13.24 ± 0.18	37.98 ± 0.29	10.77 ± 0.11	31.19 ± 0.44	9.03 ± 0.07
	cDSD-MV	*casc.*	47.98 ± 0.25	13.26 ± 0.05	39.80 ± 0.31	11.08 ± 0.06	33.90 ± 0.36	9.38 ± 0.06
	cDSD-MV-Known		49.07 ± 0.57	13.68 ± 0.10	44.09 ± 0.34	12.48 ± 0.11	41.31 ± 0.18	11.76 ± 0.06
	cDSD-MV-Known	*casc.*	49.39 ± 0.60	13.67 ± 0.13	44.72 ± 0.38	12.30 ± 0.08	41.58 ± 0.25	11.35 ± 0.07
	cDSD-WMV		49.41 ± 0.41	13.67 ± 0.10	39.48 ± 0.43	11.23 ± 0.10	31.98 ± 0.36	9.35 ± 0.10
	cDSD-WMV	*casc.*	49.43 ± 0.31	13.79 ± 0.09	40.80 ± 0.32	11.44 ± 0.06	34.27 ± 0.38	9.59 ± 0.09
	cDSD-WMV-Known		50.49 ± 0.45	14.06 ± 0.16	45.99 ± 0.36	12.93 ± 0.07	43.42 ± 0.21	12.16 ± 0.08
	cDSD-WMV-Known	*casc.*	50.52 ± 0.48	14.03 ± 0.13	45.90 ± 0.45	12.78 ± 0.09	42.61 ± 0.26	11.72 ± 0.08

Supplementary Table S2

Table S2: Summary of results for *S. cerevisiae* using cascading across label type, prediction method, and CV fold size. We report mean and standard deviation across 10 random splits of testing and training data. A fixed cascading setting (12 rounds, 35% high confidence threshold, **WCC confidence function**) is used.

			2-Fold		**4-Fold**		**6-Fold**	
			acc.	F1	acc.	F1	acc.	F1
MIPS1	cDSD-MV		66.71 ± 0.57	41.43 ± 0.29	57.96 ± 0.32	40.43 ± 0.23	49.38 ± 0.31	37.86 ± 0.26
	cDSD-MV	*casc.*	67.58 ± 0.34	41.72 ± 0.19	62.47 ± 0.47	41.79 ± 0.10	58.84 ± 0.31	40.25 ± 0.21
	cDSD-MV-Known		67.65 ± 0.52	41.69 ± 0.25	64.21 ± 3.24	41.89 ± 0.15	62.34 ± 0.23	41.01 ± 0.06
	cDSD-MV-Known	*casc.*	68.40 ± 0.47	41.84 ± 0.15	63.79 ± 0.24	41.76 ± 0.12	61.05 ± 0.39	40.78 ± 0.11
	cDSD-WMV		66.71 ± 0.42	42.07 ± 0.24	57.83 ± 0.52	41.07 ± 0.22	49.06 ± 0.24	38.26 ± 0.22
	cDSD-WMV	*casc.*	68.11 ± 0.55	42.38 ± 0.21	62.45 ± 0.29	42.29 ± 0.18	58.80 ± 0.48	40.66 ± 0.19
	cDSD-WMV-Known		68.31 ± 0.36	42.40 ± 0.31	64.91 ± 0.30	42.74 ± 0.14	63.06 ± 0.28	41.99 ± 0.15
	cDSD-WMV-Known	*casc.*	68.87 ± 0.37	42.45 ± 0.18	64.21 ± 0.28	42.45 ± 0.18	61.18 ± 0.33	41.36 ± 0.14
MIPS2	cDSD-MV		53.50 ± 0.71	32.55 ± 0.22	44.32 ± 0.29	30.30 ± 0.09	37.29 ± 0.38	27.85 ± 0.23
	cDSD-MV	*casc.*	54.69 ± 0.64	32.83 ± 0.32	48.14 ± 0.78	30.86 ± 0.36	43.90 ± 0.37	28.76 ± 0.15
	cDSD-MV-Known		54.70 ± 0.39	32.87 ± 0.23	50.39 ± 0.18	31.70 ± 0.09	48.17 ± 0.19	30.36 ± 0.14
	cDSD-MV-Known	*casc.*	55.42 ± 0.44	33.04 ± 0.34	51.09 ± 0.33	31.64 ± 0.15	46.98 ± 0.57	29.69 ± 0.17
	cDSD-WMV		55.02 ± 0.58	33.57 ± 0.24	45.94 ± 0.47	31.38 ± 0.18	38.03 ± 0.33	28.70 ± 0.15
	cDSD-WMV	*casc.*	55.49 ± 0.43	33.69 ± 0.21	48.88 ± 0.24	31.74 ± 0.12	44.14 ± 0.33	29.29 ± 0.16
	cDSD-WMV-Known		55.85 ± 0.59	33.87 ± 0.19	51.86 ± 0.38	32.93 ± 0.17	49.42 ± 0.21	31.51 ± 0.11
	cDSD-WMV-Known	*casc.*	55.92 ± 0.51	33.96 ± 0.31	51.32 ± 0.50	32.58 ± 0.19	47.78 ± 0.62	30.80 ± 0.20
MIPS3	cDSD-MV		47.95 ± 0.49	27.06 ± 0.22	39.54 ± 0.28	26.76 ± 0.30	32.59 ± 0.28	25.31 ± 0.27
	cDSD-MV	*casc.*	49.43 ± 0.44	27.42 ± 0.18	43.14 ± 0.28	26.98 ± 0.17	38.67 ± 0.23	25.28 ± 0.13
	cDSD-MV-Known		49.15 ± 0.43	27.46 ± 0.17	44.42 ± 0.22	26.84 ± 0.18	41.48 ± 0.26	25.50 ± 0.11
	cDSD-MV-Known	*casc.*	49.83 ± 0.71	27.53 ± 0.29	45.16 ± 0.56	26.69 ± 0.13	41.10 ± 0.41	25.24 ± 0.16
	cDSD-WMV		50.03 ± 0.34	28.41 ± 0.24	41.31 ± 0.35	27.79 ± 0.27	34.07 ± 0.38	26.00 ± 0.27
	cDSD-WMV	*casc.*	50.21 ± 0.60	28.43 ± 0.36	43.74 ± 0.30	27.81 ± 0.16	39.22 ± 0.24	25.96 ± 0.14
	cDSD-WMV-Known		50.70 ± 0.46	28.33 ± 0.11	46.64 ± 0.27	28.13 ± 0.11	43.66 ± 0.33	26.83 ± 0.11
	cDSD-WMV-Known	*casc.*	50.90 ± 0.49	28.54 ± 0.23	45.62 ± 0.28	27.89 ± 0.16	41.59 ± 0.41	26.32 ± 0.11
GO-MFBP-4	cDSD-MV		47.41 ± 0.59	13.24 ± 0.18	37.98 ± 0.29	10.77 ± 0.11	31.19 ± 0.44	9.03 ± 0.07
	cDSD-MV	*casc.*	48.04 ± 0.55	13.37 ± 0.14	39.50 ± 0.44	11.16 ± 0.07	34.07 ± 0.30	9.60 ± 0.07
	cDSD-MV-Known		49.07 ± 0.57	13.68 ± 0.10	44.09 ± 0.34	12.48 ± 0.11	41.31 ± 0.18	11.76 ± 0.06
	cDSD-MV-Known	*casc.*	49.28 ± 0.59	13.60 ± 0.11	44.83 ± 0.38	12.30 ± 0.09	41.82 ± 0.32	11.40 ± 0.07
	cDSD-WMV		49.41 ± 0.41	13.67 ± 0.10	39.48 ± 0.43	11.23 ± 0.10	31.98 ± 0.36	9.35 ± 0.10
	cDSD-WMV	*casc.*	48.95 ± 0.40	13.71 ± 0.08	40.40 ± 0.61	11.47 ± 0.11	34.60 ± 0.45	9.83 ± 0.09
	cDSD-WMV-Known		50.49 ± 0.45	14.06 ± 0.16	45.99 ± 0.36	12.93 ± 0.07	43.42 ± 0.21	12.16 ± 0.08
	cDSD-WMV-Known	*casc.*	50.75 ± 0.44	14.06 ± 0.10	45.91 ± 0.40	12.73 ± 0.10	42.59 ± 0.38	11.72 ± 0.07

Supplementary Table S3

Table S3: Avg. Distribution of accuracy among high and low confidence predictions after 1 round of cascading across confidence functions, using high-confidence thresholds of 10%, 25%, 40%, and 50%; Results shown for *S. cerevisiae* using 2-Fold CV (50% of training data) and cDSD-MV. We report mean and standard deviation across 10 random splits of testing and training data.

		MIPS1		**MIPS2**		**MIPS3**		**GO-MFBP-4**	
		high-conf	*low-conf*	*high-conf*	*low-conf*	*high-conf*	*low-conf*	*high-conf*	*low-conf*
[**CC**]	10%	83.0 ± 1.1	62.8 ± 0.5	82.0 ± 0.4	46.7 ± 0.5	79.7 ± 0.9	40.9 ± 0.7	61.9 ± 1.6	44.5 ± 0.3
	25%	78.3 ± 0.8	57.5 ± 0.8	70.5 ± 0.6	39.4 ± 0.7	68.0 ± 1.0	32.5 ± 0.6	56.0 ± 1.0	41.9 ± 0.6
	40%	74.8 ± 0.6	54.7 ± 1.2	64.1 ± 0.6	35.0 ± 1.2	61.3 ± 0.4	25.1 ± 1.0	53.2 ± 0.6	39.3 ± 1.2
	50%	72.3 ± 0.5	54.2 ± 1.4	61.0 ± 0.5	33.1 ± 0.8	57.5 ± 0.6	20.6 ± 0.9	52.1 ± 0.5	35.7 ± 1.4
[**WCC**]	10%	84.5 ± 1.3	62.5 ± 0.4	84.1 ± 1.1	46.5 ± 0.5	79.7 ± 0.7	40.3 ± 0.7	63.4 ± 1.6	44.0 ± 0.8
	25%	79.4 ± 0.5	57.4 ± 0.7	72.9 ± 0.9	39.2 ± 0.5	69.6 ± 1.1	31.9 ± 0.9	56.7 ± 0.9	41.7 ± 0.7
	40%	74.3 ± 1.0	53.9 ± 1.1	65.1 ± 0.5	34.1 ± 0.8	62.1 ± 0.7	24.9 ± 0.2	53.3 ± 1.1	38.9 ± 1.3
	50%	71.8 ± 0.3	51.8 ± 1.0	62.2 ± 0.5	31.6 ± 1.0	58.4 ± 0.3	19.8 ± 1.0	53.3 ± 0.7	35.2 ± 1.5
[**EC**]	10%	87.1 ± 1.0	61.9 ± 0.5	85.9 ± 0.7	46.4 ± 0.5	83.7 ± 1.1	39.6 ± 0.6	76.8 ± 1.2	41.1 ± 0.5
	25%	79.9 ± 0.6	56.4 ± 0.6	77.1 ± 1.0	35.7 ± 0.6	75.7 ± 0.7	27.4 ± 0.7	69.5 ± 0.7	32.0 ± 0.6
	40%	75.7 ± 0.6	51.5 ± 1.0	69.4 ± 0.9	27.5 ± 0.9	65.6 ± 0.9	18.7 ± 1.5	61.8 ± 0.6	25.3 ± 1.1
	50%	72.7 ± 0.7	50.1 ± 1.5	64.3 ± 0.5	23.9 ± 1.8	60.1 ± 0.6	16.2 ± 1.2	57.3 ± 0.5	22.1 ± 1.0
[**RC**]	10%	66.5 ± 2.1	67.1 ± 0.4	53.8 ± 2.0	53.8 ± 0.5	47.7 ± 1.9	48.1 ± 0.5	48.3 ± 2.3	47.1 ± 1.5
	25%	66.8 ± 0.6	67.2 ± 0.9	53.8 ± 1.0	54.0 ± 0.7	47.9 ± 1.0	48.0 ± 0.6	47.8 ± 0.9	48.8 ± 0.9
	40%	66.8 ± 0.8	67.9 ± 0.5	54.2 ± 0.8	54.7 ± 1.3	48.3 ± 0.7	48.5 ± 1.0	48.0 ± 0.8	48.2 ± 1.3
	50%	67.2 ± 0.6	67.1 ± 1.2	53.8 ± 0.7	54.5 ± 0.8	48.1 ± 0.9	47.9 ± 1.3	47.9 ± 1.0	48.4 ± 1.1

Supplementary Table S4

Table S4: Summary of results for *D. melanogaster* using cascading across label type, prediction method, and CV fold size. We report mean and standard deviation across 10 random splits of testing and training data. A fixed cascading setting (**EC confidence function**, 12 rounds, 35% high confidence threshold) is used.

			2-Fold		**4-Fold**		**6-Fold**	
			acc.	F1	acc.	F1	acc.	F1
GO-MFBP-4	cDSD-MV		38.00 ± 0.52	10.33 ± 0.11	28.58 ± 0.34	9.45 ± 0.10	22.46 ± 0.49	8.20 ± 0.08
	cDSD-MV	*casc.*	39.90 ± 0.55	10.51 ± 0.13	32.78 ± 0.47	10.00 ± 0.12	28.61 ± 0.36	8.95 ± 0.07
	cDSD-MV-Known		40.36 ± 0.40	10.92 ± 0.12	37.43 ± 0.31	11.90 ± 0.06	35.97 ± 0.31	11.94 ± 0.05
	cDSD-MV-Known	*casc.*	40.52 ± 0.60	10.88 ± 0.11	37.67 ± 0.09	11.69 ± 0.09	35.98 ± 0.25	11.54 ± 0.07
	cDSD-WMV		38.56 ± 0.50	10.26 ± 0.13	29.35 ± 0.37	9.48 ± 0.09	22.93 ± 0.33	8.23 ± 0.08
	cDSD-WMV	*casc.*	40.01 ± 0.77	10.48 ± 0.13	32.94 ± 0.33	10.03 ± 0.12	28.81 ± 0.38	8.96 ± 0.07
	cDSD-WMV-Known		41.85 ± 0.48	11.22 ± 0.10	38.86 ± 0.52	12.18 ± 0.06	37.14 ± 0.37	12.10 ± 0.09
	cDSD-WMV-Known	*casc.*	41.63 ± 0.31	10.96 ± 0.08	38.39 ± 0.40	11.83 ± 0.10	36.39 ± 0.30	11.63 ± 0.08

Supplementary Table S5

Table S5: Summary of results for *S. cerevisiae* using cascading across prediction method and CV fold size, where the number of neighbors considered in each voting method (k parameter) varies. We report mean and standard deviation across 10 random splits of testing and training data. A fixed cascading setting (EC confidence function, 10 rounds, 35% high confidence threshold) is used. Only MIPS1 labels are considered.

MIPS1 labels		**2-fold**		**4-fold**		**6-fold**	
		acc.	F1	acc.	F1	acc.	F1
cDSD-MV	k=8	67.30 ± 0.30	41.89 ± 0.19	62.31 ± 0.36	41.72 ± 0.20	58.67 ± 0.39	40.21 ± 0.15
	k=9	67.37 ± 0.51	41.86 ± 0.21	62.40 ± 0.54	41.85 ± 0.16	58.83 ± 0.27	40.27 ± 0.21
	k=10	67.54 ± 0.25	41.99 ± 0.20	62.30 ± 0.36	42.08 ± 0.24	59.05 ± 1.44	40.67 ± 0.21
	k=15	67.29 ± 0.33	41.83 ± 0.19	62.48 ± 0.35	41.80 ± 0.20	58.99 ± 0.50	40.31 ± 0.18
	k=20	67.45 ± 0.47	41.98 ± 0.21	62.47 ± 0.33	41.91 ± 0.18	59.13 ± 0.50	40.36 ± 0.20
cDSD-MV-known	k=8	67.88 ± 0.38	41.90 ± 0.19	65.09 ± 0.36	42.23 ± 0.16	63.20 ± 0.21	41.39 ± 0.13
	k=9	67.99 ± 0.64	41.87 ± 0.28	64.98 ± 0.30	42.20 ± 0.14	63.09 ± 0.33	41.33 ± 0.18
	k=10	67.96 ± 0.40	42.03 ± 0.22	65.10 ± 0.25	42.42 ± 0.14	62.80 ± 0.26	41.54 ± 0.17
	k=15	68.04 ± 0.47	41.83 ± 0.23	64.92 ± 0.31	42.16 ± 0.13	63.07 ± 0.23	41.31 ± 0.16
	k=20	67.94 ± 0.42	42.03 ± 0.30	64.98 ± 0.31	42.18 ± 0.14	63.17 ± 0.29	41.35 ± 0.12
cDSD-WMV	k=8	67.93 ± 0.76	42.50 ± 0.16	62.81 ± 0.40	42.26 ± 0.24	59.22 ± 0.76	40.71 ± 0.17
	k=9	67.72 ± 0.45	42.46 ± 0.18	62.75 ± 0.45	42.33 ± 0.17	58.84 ± 0.35	40.59 ± 0.21
	k=10	67.86 ± 0.58	42.59 ± 0.26	62.76 ± 0.63	42.41 ± 0.14	59.10 ± 0.56	40.98 ± 0.11
	k=15	68.11 ± 0.50	42.32 ± 0.25	62.50 ± 0.43	42.17 ± 0.24	58.99 ± 0.50	40.54 ± 0.20
	k=20	67.85 ± 0.28	42.51 ± 0.25	62.52 ± 0.58	42.19 ± 0.24	58.68 ± 0.40	40.58 ± 0.20
cDSD-WMV-known	k=8	68.48 ± 0.64	42.40 ± 0.30	65.55 ± 0.36	42.68 ± 0.16	63.41 ± 0.39	41.83 ± 0.14
	k=9	68.43 ± 0.47	42.28 ± 0.16	65.54 ± 0.31	42.80 ± 0.18	63.47 ± 0.24	41.87 ± 0.19
	k=10	68.37 ± 0.46	42.34 ± 0.12	65.43 ± 0.37	42.77 ± 0.17	63.63 ± 0.53	41.94 ± 0.15
	k=15	68.29 ± 0.42	42.40 ± 0.20	65.42 ± 0.28	42.75 ± 0.14	63.44 ± 0.20	41.82 ± 0.16
	k=20	68.31 ± 0.57	42.30 ± 0.26	65.48 ± 0.25	42.74 ± 0.18	63.56 ± 0.32	41.87 ± 0.16