Molecular Graph Convolutions: Moving Beyond Fingerprints
Appendix

Steven Kearnes
Stanford University
kearnes@stanford.edu

Kevin McCloskey
Google Inc.
mccloskey@google.com

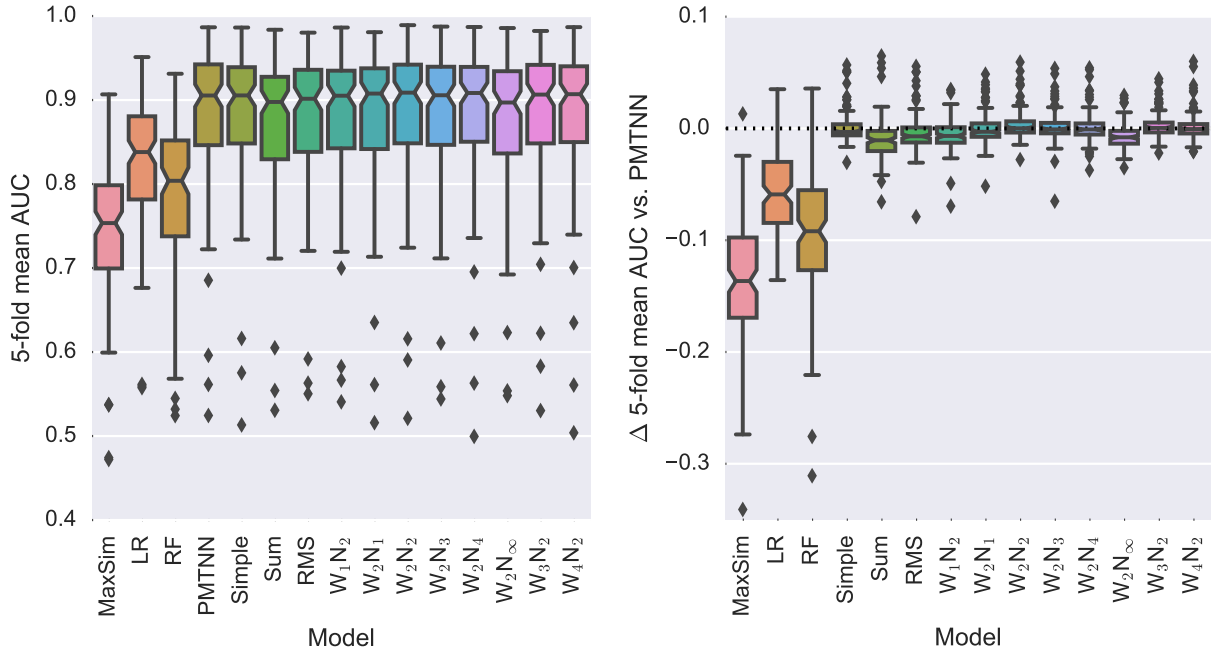
Marc Berndl
Google Inc.
marcberndl@google.com

Vijay Pande
Stanford University
pande@stanford.edu

Patrick Riley
Google Inc.
pfr@google.com

A Appendix: Model comparison

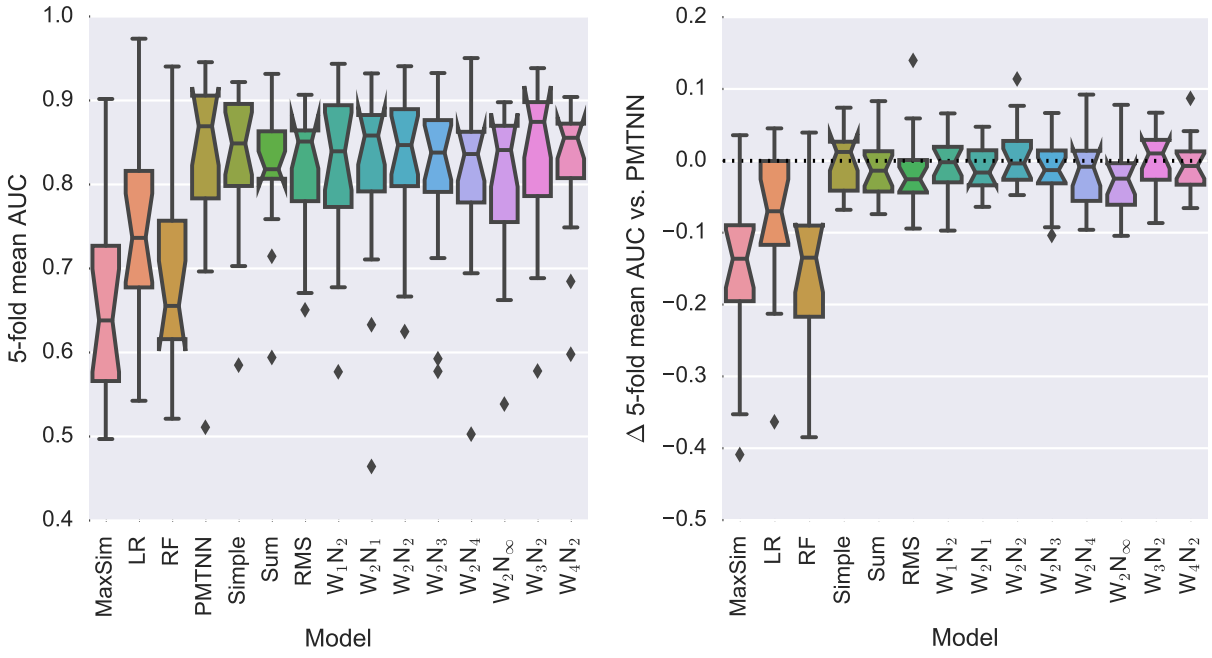
The following figures are box plot representations of the data summarized in Table 4, organized by dataset group. We provide (a) box plots for absolute 5-fold mean AUC scores for each model and (b) difference box plots showing differences in 5-fold mean AUC scores against the pyramidal (2000, 100) multitask neural network (PMTNN) baseline model. The difference box plots are visual analogs of the sign test confidence intervals reported in Table 4. Note, however, that the confidence intervals on box plot medians (calculated as $\pm 1.57 \times \mathrm{IQR}/\sqrt{N}$ (McGill et al., 1978)) do not necessarily correspond to the sign test confidence intervals.



(A) Full box plot.

(B) Difference box plot vs. PMTNN.

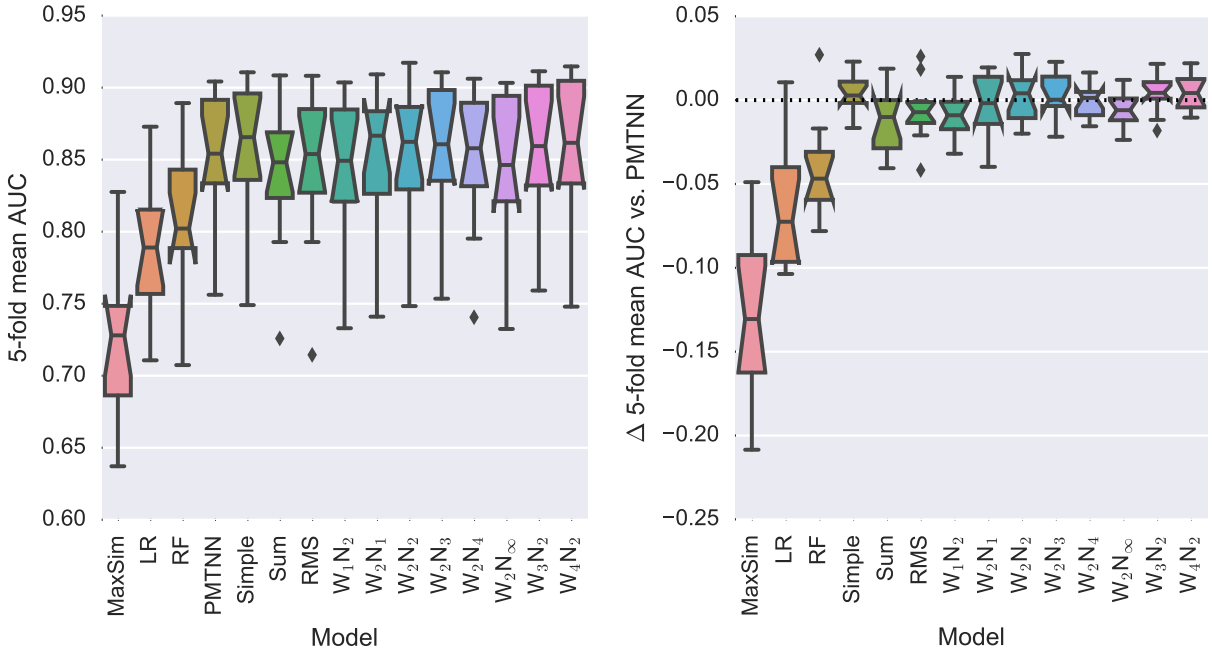
FIGURE S1: Model performance on PCBA datasets.



(A) Full box plot.

(B) Difference box plot vs. PMTNN.

FIGURE S2: Model performance on MUV datasets.



(A) Full box plot.

(B) Difference box plot vs. PMTNN.

FIGURE S3: Model performance on Tox21 datasets.

B Appendix: ROC enrichment

The following tables report ROC enrichment (Jain and Nicholls, 2008) scores for baseline and graph convolution models. Each metric was optimized separately using the held-out validation set for each model, such that ROC AUC or ROC enrichment scores at different false positive rates (FPRs) are not necessarily derived from predictions using the same set of model training checkpoints.

TABLE S1: Median 5-fold mean ROC enrichment values for reported models at 1% FPR ($E_{1\%}$). For each model, we report the median $\Delta E_{1\%}$ and the 95% Wilson score interval for a sign test estimating the probability that a given model will outperform the PMTNN baseline (see Section 3.7). Bold values indicate sign test confidence intervals that do not include 0.5.

Model	PCBA ($n = 128$)			MUV ($n = 17$)			Tox21 ($n = 12$)		
	Median $E_{1\%}$	Median $\Delta E_{1\%}$	Sign Test 95% CI	Median $E_{1\%}$	Median $\Delta E_{1\%}$	Sign Test 95% CI	Median $E_{1\%}$	Median $\Delta E_{1\%}$	Sign Test 95% CI
MaxSim	24.1	**−16.2**	**(0.04, 0.13)**	13.3	−3.3	(0.22, 0.64)	12.8	**−13.0**	**(0.00, 0.24)**
LR	20.2	**−18.8**	**(0.01, 0.08)**	16.7	0.0	(0.28, 0.72)	17.8	**−5.1**	**(0.05, 0.45)**
RF	34.5	**−6.9**	**(0.12, 0.25)**	23.3	−3.3	(0.23, 0.67)	26.4	−0.2	(0.25, 0.75)
PMTNN	43.7			30.0			28.1		
W_2N_2-simple	42.3	**−1.6**	**(0.15, 0.29)**	30.0	−3.3	(0.14, 0.56)	24.7	−1.1	(0.19, 0.68)
W_2N_2-sum	34.5	**−6.5**	**(0.05, 0.15)**	16.7	**−13.3**	**(0.03, 0.36)**	17.2	**−9.8**	**(0.01, 0.35)**
W_2N_2-RMS	39.2	**−3.5**	**(0.04, 0.14)**	13.3	**−6.7**	**(0.01, 0.30)**	21.2	**−4.3**	**(0.05, 0.45)**
W_1N_2	38.3	**−3.6**	**(0.05, 0.15)**	20.0	**−3.3**	**(0.08, 0.48)**	22.6	−4.7	(0.09, 0.53)
W_2N_1	40.9	**−2.2**	**(0.17, 0.31)**	16.7	−6.7	(0.14, 0.56)	25.6	−2.7	(0.09, 0.53)
W_2N_2	42.2	**−0.8**	**(0.30, 0.46)**	26.7	**−3.3**	**(0.07, 0.45)**	26.2	1.6	(0.47, 0.91)
W_2N_3	42.0	**−0.9**	**(0.18, 0.33)**	26.7	**−3.3**	**(0.10, 0.49)**	25.5	2.4	(0.39, 0.86)
W_2N_4	42.0	**−0.7**	**(0.23, 0.39)**	23.3	**−6.7**	**(0.08, 0.48)**	23.5	−0.4	(0.25, 0.75)
W_2N_∞	38.8	**−2.7**	**(0.06, 0.17)**	20.0	−3.3	(0.14, 0.56)	23.4	−1.1	(0.09, 0.53)
W_3N_2	42.1	**−1.0**	**(0.19, 0.34)**	26.7	0.0	(0.25, 0.70)	24.8	0.5	(0.32, 0.81)
W_4N_2	40.6	**−1.2**	**(0.22, 0.38)**	23.3	**−3.3**	**(0.08, 0.48)**	24.8	−0.9	(0.09, 0.53)

TABLE S2: Median 5-fold mean ROC enrichment values for reported models at 5% FPR ($E_{5\%}$). For each model, we report the median $\Delta E_{5\%}$ and the 95% Wilson score interval for a sign test estimating the probability that a given model will outperform the PMTNN baseline (see Section 3.7). Bold values indicate sign test confidence intervals that do not include 0.5.

Model	PCBA ($n = 128$)			MUV ($n = 17$)			Tox21 ($n = 12$)		
	Median $E_{5\%}$	Median $\Delta E_{5\%}$	Sign Test 95% CI	Median $E_{5\%}$	Median $\Delta E_{5\%}$	Sign Test 95% CI	Median $E_{5\%}$	Median $\Delta E_{5\%}$	Sign Test 95% CI
MaxSim	8.5	**−4.4**	**(0.01, 0.08)**	6.0	**−3.3**	**(0.03, 0.34)**	6.7	**−3.9**	**(0.00, 0.24)**
LR	8.8	**−3.6**	**(0.02, 0.09)**	6.0	−2.0	(0.14, 0.56)	8.3	**−1.9**	**(0.01, 0.35)**
RF	10.2	**−2.5**	**(0.06, 0.17)**	6.0	−2.0	(0.14, 0.56)	9.6	**−1.0**	**(0.05, 0.45)**
PMTNN	13.5			10.7			10.3		
W_2N_2-simple	13.4	**−0.3**	**(0.19, 0.34)**	10.0	−1.3	(0.22, 0.64)	10.1	−0.2	(0.19, 0.68)
W_2N_2-sum	12.3	**−0.9**	**(0.12, 0.25)**	7.3	**−2.0**	**(0.04, 0.38)**	8.8	**−1.9**	**(0.01, 0.35)**
W_2N_2-RMS	12.9	**−0.7**	**(0.12, 0.25)**	8.0	**−2.0**	**(0.06, 0.41)**	9.4	**−1.4**	**(0.01, 0.35)**
W_1N_2	13.0	**−0.5**	**(0.13, 0.27)**	9.3	**−2.0**	**(0.10, 0.49)**	9.9	−0.8	(0.09, 0.53)
W_2N_1	13.3	**−0.4**	**(0.20, 0.35)**	8.7	**−0.7**	**(0.01, 0.33)**	10.4	−0.4	(0.14, 0.61)
W_2N_2	13.6	**−0.1**	**(0.30, 0.47)**	10.0	**−1.3**	**(0.10, 0.49)**	10.4	0.0	(0.28, 0.79)
W_2N_3	13.3	**−0.2**	**(0.24, 0.40)**	8.7	−1.3	(0.12, 0.55)	10.5	−0.2	(0.19, 0.68)
W_2N_4	13.3	**−0.2**	**(0.25, 0.41)**	8.7	−1.3	(0.13, 0.53)	10.2	−0.2	(0.14, 0.61)
W_2N_∞	12.8	**−0.5**	**(0.06, 0.16)**	8.7	**−1.3**	**(0.03, 0.34)**	10.4	−0.2	(0.15, 0.65)
W_3N_2	13.6	**−0.1**	**(0.26, 0.43)**	9.3	0.0	(0.16, 0.61)	10.4	−0.2	(0.14, 0.61)
W_4N_2	13.3	**−0.1**	**(0.29, 0.46)**	8.0	−1.3	(0.14, 0.56)	10.5	0.0	(0.25, 0.75)

TABLE S3: Median 5-fold mean ROC enrichment values for reported models at 10% FPR ($E_{10\%}$). For each model, we report the median $\Delta E_{10\%}$ and the 95% Wilson score interval for a sign test estimating the probability that a given model will outperform the PMTNN baseline (see Section 3.7). Bold values indicate sign test confidence intervals that do not include 0.5.

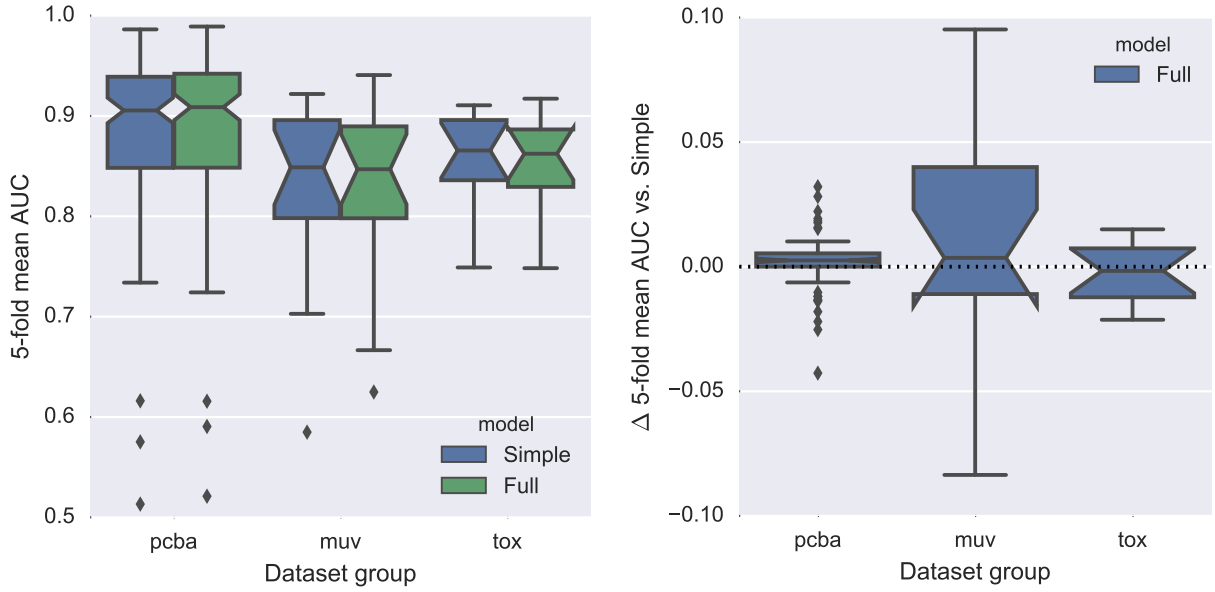
Model	PCBA ($n = 128$)			MUV ($n = 17$)			Tox21 ($n = 12$)		
	Median $E_{10\%}$	Median $\Delta E_{10\%}$	Sign Test 95% CI	Median $E_{10\%}$	Median $\Delta E_{10\%}$	Sign Test 95% CI	Median $E_{10\%}$	Median $\Delta E_{10\%}$	Sign Test 95% CI
MaxSim	5.1	**−2.2**	**(0.00, 0.06)**	3.3	**−2.0**	**(0.04, 0.38)**	4.3	**−2.1**	**(0.00, 0.24)**
LR	5.9	**−1.4**	**(0.01, 0.08)**	4.7	−0.7	(0.26, 0.69)	5.2	**−1.1**	**(0.00, 0.24)**
RF	6.0	**−1.3**	**(0.04, 0.14)**	3.7	−1.0	(0.13, 0.53)	5.8	**−0.7**	**(0.05, 0.45)**
PMTNN	7.8			6.3			6.4		
W_2N_2-simple	7.7	**−0.1**	**(0.26, 0.42)**	5.7	−0.7	(0.15, 0.58)	6.3	0.0	(0.25, 0.75)
W_2N_2-sum	7.2	**−0.4**	**(0.12, 0.25)**	5.3	−0.7	(0.13, 0.53)	5.9	**−0.6**	**(0.05, 0.45)**
W_2N_2-RMS	7.5	**−0.2**	**(0.13, 0.26)**	5.3	**−1.0**	**(0.07, 0.45)**	5.9	**−0.4**	**(0.05, 0.45)**
W_1N_2	7.5	**−0.2**	**(0.12, 0.25)**	5.0	**−1.0**	**(0.10, 0.49)**	6.2	**−0.2**	**(0.05, 0.45)**
W_2N_1	7.6	**−0.1**	**(0.21, 0.37)**	6.0	−0.7	(0.11, 0.52)	6.3	−0.1	(0.09, 0.53)
W_2N_2	7.7	**0.0**	**(0.28, 0.44)**	5.7	−0.3	(0.18, 0.61)	6.2	0.0	(0.25, 0.75)
W_2N_3	7.7	**0.0**	**(0.28, 0.45)**	5.7	**−0.7**	**(0.10, 0.49)**	6.3	0.1	(0.35, 0.85)
W_2N_4	7.7	**−0.1**	**(0.25, 0.41)**	5.7	−0.7	(0.13, 0.53)	6.4	0.0	(0.25, 0.75)
W_2N_∞	7.4	**−0.3**	**(0.09, 0.20)**	5.0	−1.0	(0.13, 0.53)	6.3	−0.1	(0.09, 0.53)
W_3N_2	7.8	0.0	(0.34, 0.51)	6.0	−0.3	(0.17, 0.59)	6.2	0.0	(0.25, 0.75)
W_4N_2	7.7	**0.0**	**(0.29, 0.46)**	5.7	−0.7	(0.13, 0.53)	6.3	0.1	(0.32, 0.81)

TABLE S4: Median 5-fold mean ROC enrichment values for reported models at 20% FPR ($E_{20\%}$). For each model, we report the median $\Delta E_{20\%}$ and the 95% Wilson score interval for a sign test estimating the probability that a given model will outperform the PMTNN baseline (see Section 3.7). Bold values indicate sign test confidence intervals that do not include 0.5.

Model	PCBA ($n = 128$)			MUV ($n = 17$)			Tox21 ($n = 12$)		
	Median $E_{20\%}$	Median $\Delta E_{20\%}$	Sign Test 95% CI	Median $E_{20\%}$	Median $\Delta E_{20\%}$	Sign Test 95% CI	Median $E_{20\%}$	Median $\Delta E_{20\%}$	Sign Test 95% CI
MaxSim	3.0	**−1.1**	**(0.00, 0.03)**	2.2	**−1.0**	**(0.03, 0.34)**	2.8	**−1.1**	**(0.00, 0.24)**
LR	3.6	**−0.5**	**(0.03, 0.11)**	3.0	−0.5	(0.18, 0.61)	3.2	**−0.5**	**(0.01, 0.35)**
RF	3.4	**−0.7**	**(0.03, 0.11)**	2.5	**−0.7**	**(0.03, 0.36)**	3.4	**−0.4**	**(0.01, 0.35)**
PMTNN	4.2			3.8			3.7		
W_2N_2-simple	4.3	**0.0**	**(0.30, 0.46)**	3.3	**−0.3**	**(0.10, 0.49)**	3.8	0.0	(0.32, 0.81)
W_2N_2-sum	4.2	**−0.1**	**(0.17, 0.31)**	3.3	**−0.3**	**(0.07, 0.43)**	3.7	−0.1	(0.09, 0.53)
W_2N_2-RMS	4.2	**−0.1**	**(0.19, 0.34)**	3.5	−0.2	(0.11, 0.52)	3.8	−0.1	(0.09, 0.53)
W_1N_2	4.2	**−0.1**	**(0.19, 0.34)**	3.7	−0.3	(0.14, 0.56)	3.7	0.0	(0.14, 0.61)
W_2N_1	4.3	**0.0**	**(0.32, 0.49)**	3.5	−0.2	(0.23, 0.67)	3.9	0.0	(0.25, 0.75)
W_2N_2	4.3	0.0	(0.38, 0.55)	3.5	−0.3	(0.17, 0.59)	3.9	0.1	(0.35, 0.85)
W_2N_3	4.3	0.0	(0.35, 0.52)	3.3	−0.3	(0.26, 0.69)	3.8	0.0	(0.32, 0.81)
W_2N_4	4.3	**0.0**	**(0.28, 0.45)**	3.3	**−0.3**	**(0.10, 0.47)**	3.8	0.0	(0.25, 0.75)
W_2N_∞	4.2	**−0.1**	**(0.12, 0.25)**	3.3	**−0.3**	**(0.07, 0.43)**	3.8	0.0	(0.19, 0.68)
W_3N_2	4.3	0.0	(0.37, 0.54)	3.5	−0.2	(0.23, 0.67)	3.8	0.1	(0.32, 0.81)
W_4N_2	4.3	0.0	(0.34, 0.51)	3.7	−0.2	(0.16, 0.61)	3.8	0.1	(0.47, 0.91)

C Appendix: Input featurization

For each of the experiments described in Section 4.2, we provide figures showing (a) box plots for absolute 5-fold mean AUC scores for each model and (b) difference box plots showing differences in 5-fold mean AUC scores against a baseline model (without any y-axis cropping).



(A) Full box plot.

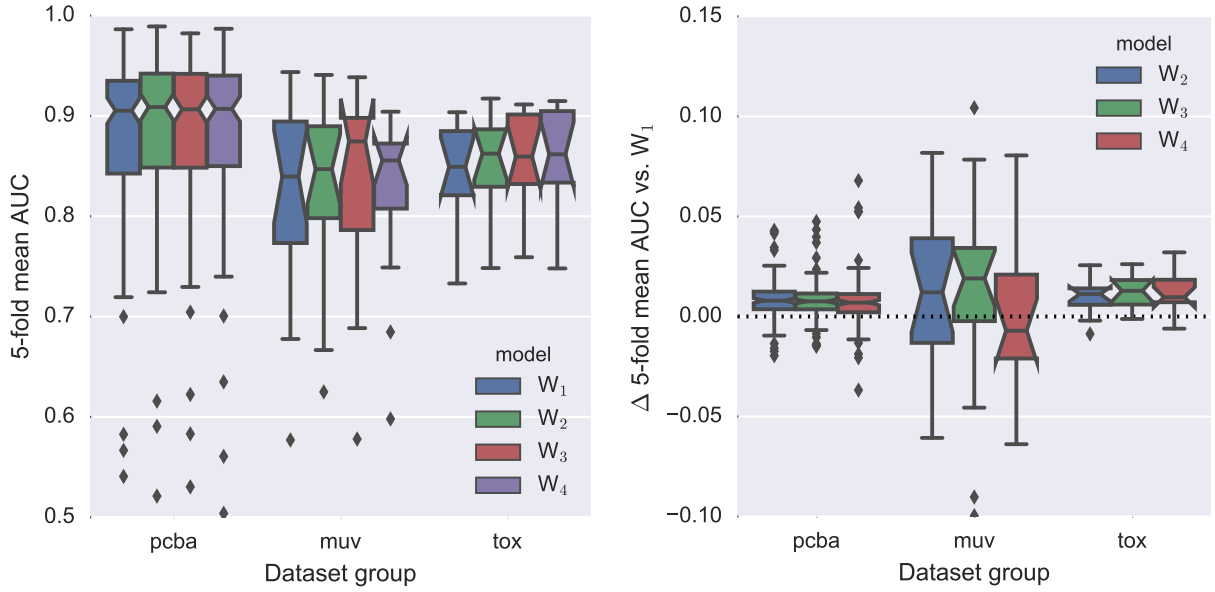
(B) Difference box plot vs. "simple" featurization.

FIGURE S4: Comparison of models with "simple" and "full" input featurizations.

D Appendix: Hyperparameter sensitivity

For each of the experiments described in Section 4.3, we provide figures showing (a) box plots for absolute 5-fold mean AUC scores for each model and (b) difference box plots showing differences in 5-fold mean AUC scores against a baseline model (without any y-axis cropping).

D.1 Number of Weave modules

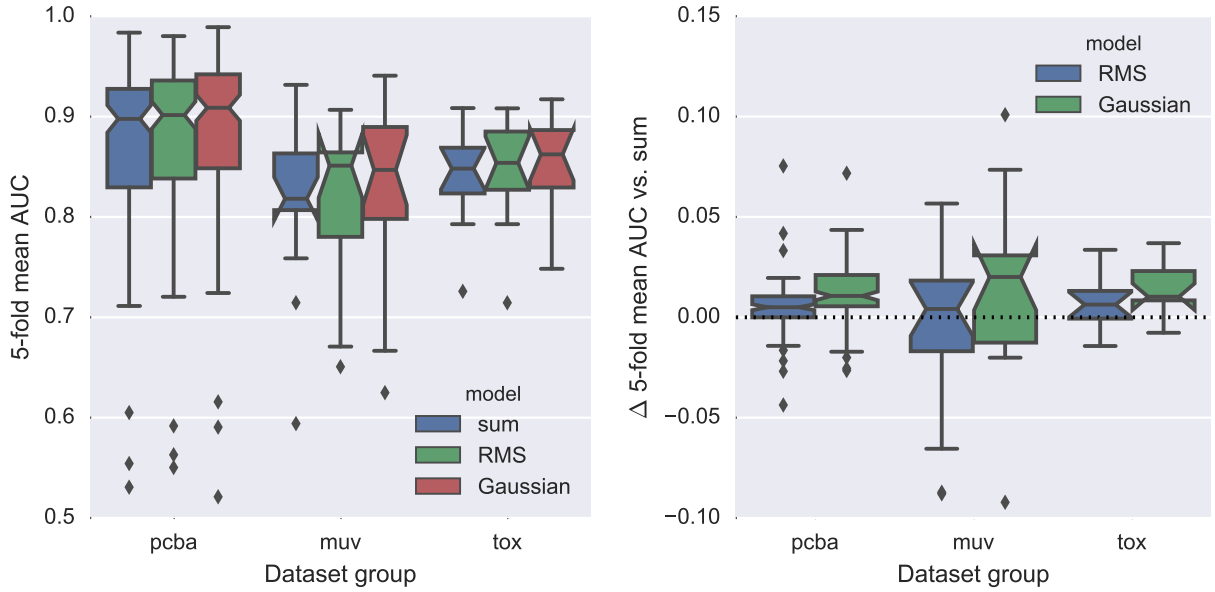


(A) Full box plot.

(B) Difference box plot vs. W_1 model.

FIGURE S5: Comparison of models with different numbers of Weave modules.

D.2 Alternative feature reductions

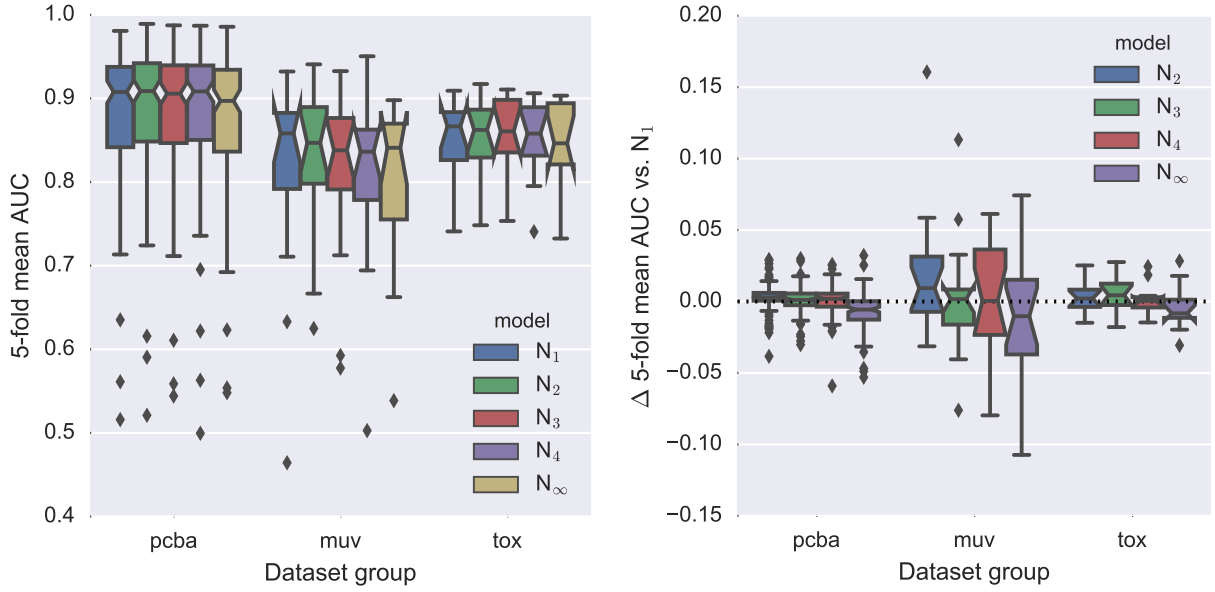


(A) Full box plot.

(B) Difference box plot vs. sum reduction.

FIGURE S6: Comparison of models with different feature reduction methods.

D.3 Distance-dependent pair features



(A) Full box plot.

(B) Difference box plot vs. N_1 model.

FIGURE S7: Comparison of models with different maximum atom pair distances.

E Appendix: Atom pair feature evolution

Figure 8 showed the evolution of atom features at different stages of a graph convolution model (after subsequent Weave modules). The following figures show the evolution of atom pair features from the same models, using both the "full" and "simple" input featurization. As in Figure 8, the initial pair features describe ibuprofen. Most of the initial featurization describes the graph distance between the atoms in the pair (see Table 3). There are many blank rows since pairs separated by more than the maximum atom pair distance are masked. Note that only unique pairs are represented (i.e. (a, b) but not (b, a)). As the pair features move through the graph convolution network, it can be seen that similar initial featurizations diverge as a consequence of Weave module operations.

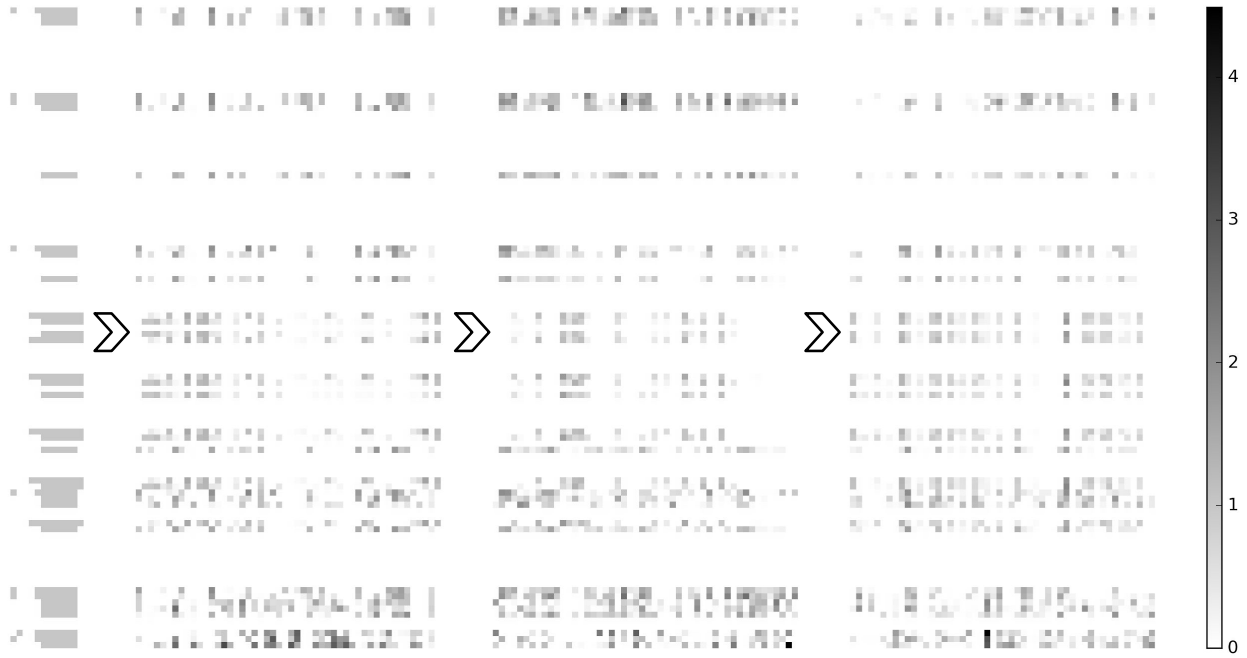


FIGURE S8: Graph convolution atom pair feature evolution using the "full" featurization in a W_3N_2 architecture. Unique atom pairs are on the y-axis (one atom pair per row). Initial pair features are shown on the left, with whitespace separating subsequent Weave module outputs.

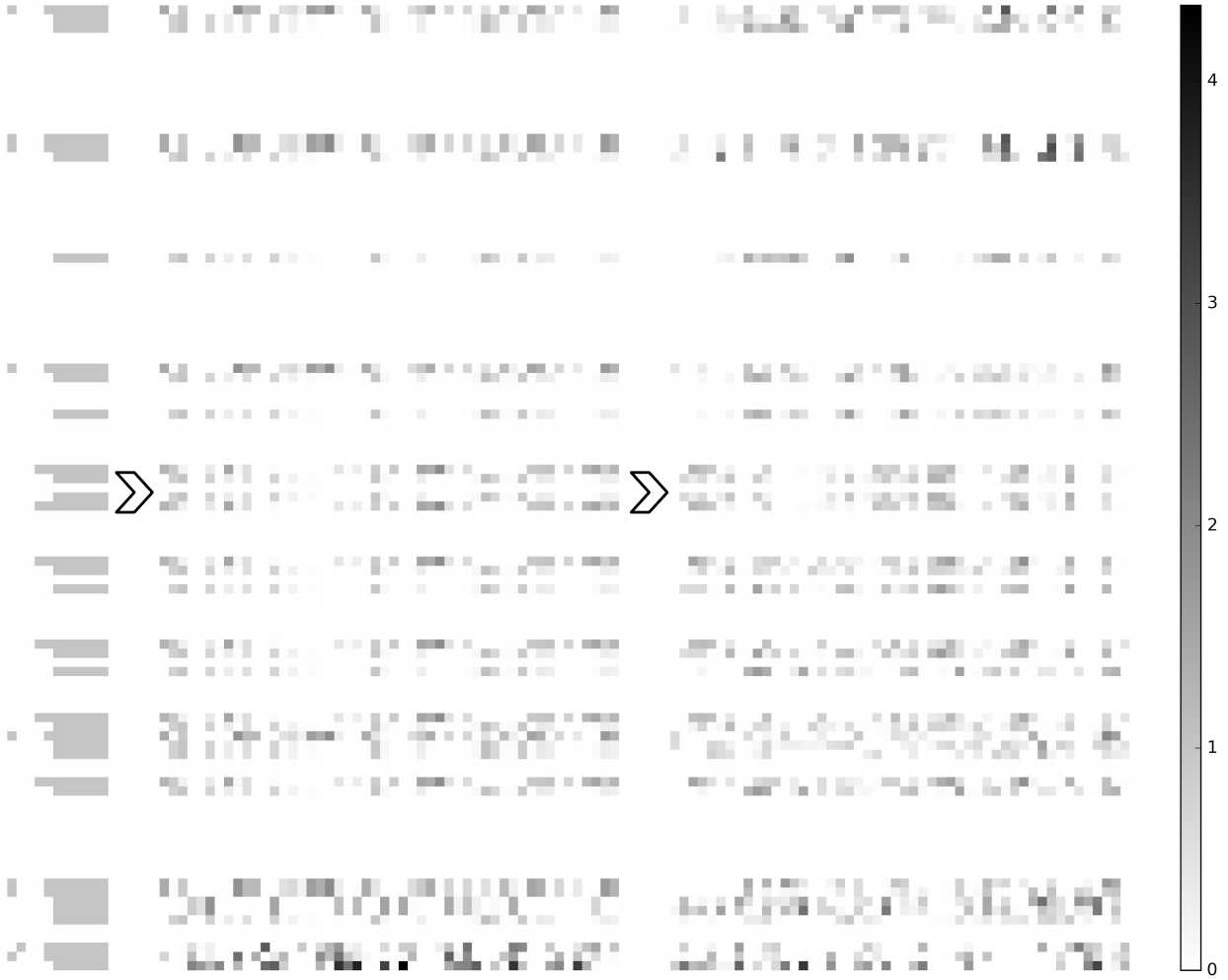


FIGURE S9: Graph convolution atom pair feature evolution using the "simple" featurization in a W_2N_2 architecture. Unique atom pairs are on the y-axis (one atom pair per row). Initial pair features are shown on the left, with whitespace separating subsequent Weave module outputs.

F Appendix: Gaussian histogram membership functions

TABLE S5: Gaussian membership functions.

Mean	Variance
−1.645	0.080
−1.080	0.029
−0.739	0.018
−0.468	0.014
−0.228	0.013
0.000	0.013
0.228	0.013
0.468	0.014
0.739	0.018
1.080	0.029
1.645	0.080

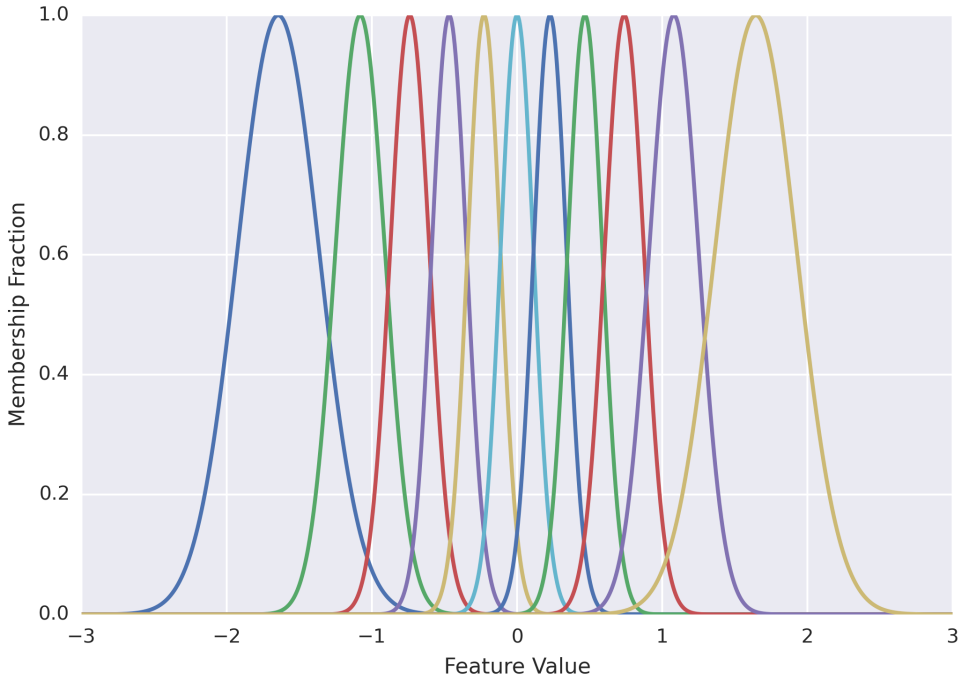


FIGURE S10: Visualization of the Gaussian membership functions.

References

Ajay N Jain and Anthony Nicholls. Recommendations for evaluation of computational methods. *Journal of computer-aided molecular design*, 22(3-4):133–139, 2008.

Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.