

Table 1: RMSE on three regression tasks

	Morgan RF	Morgan XGB	Morgan DNN	MP graph NEF*	MP graph GCNN	MP graph Weave*	N-gram graph RF	N-gram graph XGB	N-gram graph DNN
delaney	1.311 ± 0.174	1.110 ± 0.129	1.231 ± 0.109	0.520 ± 0.137	0.913 ± 0.061	0.460 ± 0.157	0.773 ± 0.076	0.700 ± 0.091	0.699 ± 0.046
malaria	1.028 ± 0.029	1.008 ± 0.031	1.052 ± 0.051	1.160 ± 0.059	1.055 ± 0.042	1.070 ± 0.118	1.030 ± 0.023	1.010 ± 0.026	1.119 ± 0.027
cep	1.642 ± 0.026	1.410 ± 0.040	1.477 ± 0.042	1.430 ± 0.176	1.184 ± 0.061	1.100 ± 0.118	1.379 ± 0.013	1.290 ± 0.026	1.365 ± 0.034

- Mean and standard deviation on 5-fold cross validation.
- Top three results are **bolded**.
- Baseline results (marked in *) are from [1][2].
- For model **Morgan DNN**, we were using the results from [1] in the submitted version. Then we got a hyperparameter set that performs better on all three tasks after submission, so we replace it here. (It was 1.4, 1.13, 2.0 on three tasks.)
- MP graph is short for Message-passing graph.

[1] Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. 2224–2232.

[2] Kusner, M. J.; Paige, B.; and Hernández-Lobato, J. M. 2017. Grammar variational autoencoder. arXiv preprint arXiv:1703.01925.

Table 4: AUC[ROC] on Tox21

	Morgan RF	Morgan XGB	Morgan DNN	MP graph GCNN	MP graph Weave	N-gram graph RF	N-gram graph XGB	N-gram graph DNN
NR-AR	0.787 ± 0.065	0.777 ± 0.062	0.756 ± 0.057	0.793 ± 0.068	0.796 ± 0.057	0.802 ± 0.082	0.790 ± 0.073	0.795 ± 0.062
NR-AR-LBD	0.864 ± 0.056	0.852 ± 0.050	0.817 ± 0.057	0.858 ± 0.037	0.816 ± 0.047	0.844 ± 0.039	0.858 ± 0.027	0.853 ± 0.034
NR-AhR	0.903 ± 0.028	0.900 ± 0.022	0.854 ± 0.036	0.896 ± 0.018	0.869 ± 0.036	0.890 ± 0.020	0.898 ± 0.020	0.869 ± 0.023
NR-Aromatase	0.827 ± 0.070	0.802 ± 0.063	0.742 ± 0.099	0.824 ± 0.052	0.830 ± 0.050	0.845 ± 0.066	0.852 ± 0.053	0.830 ± 0.065
NR-ER	0.724 ± 0.019	0.721 ± 0.022	0.692 ± 0.018	0.734 ± 0.036	0.729 ± 0.025	0.727 ± 0.043	0.733 ± 0.035	0.712 ± 0.024
NR-ER-LBD	0.815 ± 0.051	0.783 ± 0.055	0.772 ± 0.019	0.805 ± 0.024	0.804 ± 0.030	0.810 ± 0.062	0.819 ± 0.036	0.787 ± 0.038
NR-PPAR-gamma	0.839 ± 0.042	0.793 ± 0.092	0.756 ± 0.037	0.821 ± 0.105	0.803 ± 0.064	0.801 ± 0.104	0.825 ± 0.104	0.783 ± 0.106
SR-ARE	0.818 ± 0.039	0.809 ± 0.039	0.781 ± 0.046	0.782 ± 0.040	0.790 ± 0.049	0.808 ± 0.028	0.826 ± 0.024	0.777 ± 0.049
SR-ATAD5	0.857 ± 0.051	0.828 ± 0.066	0.738 ± 0.079	0.839 ± 0.037	0.823 ± 0.041	0.841 ± 0.032	0.837 ± 0.041	0.811 ± 0.022
SR-HSE	0.793 ± 0.028	0.764 ± 0.043	0.731 ± 0.034	0.774 ± 0.036	0.771 ± 0.036	0.773 ± 0.049	0.786 ± 0.065	0.750 ± 0.063
SR-MMP	0.886 ± 0.019	0.879 ± 0.026	0.856 ± 0.027	0.888 ± 0.018	0.886 ± 0.022	0.895 ± 0.017	0.909 ± 0.017	0.865 ± 0.024
SR-p53	0.849 ± 0.033	0.823 ± 0.057	0.759 ± 0.034	0.840 ± 0.053	0.813 ± 0.065	0.833 ± 0.033	0.843 ± 0.054	0.805 ± 0.036
average	0.830 ± 0.046	0.811 ± 0.048	0.771 ± 0.047	0.821 ± 0.045	0.811 ± 0.039	0.822 ± 0.045	0.831 ± 0.046	0.803 ± 0.045

- Mean and standard deviation on 5-fold cross validation.
- Top three results are **bolded**.
- MP graph is short for Message-passing graph.

Figure A.1: Tukey's Test on Tox21

- The Tukey's Test on test ROC.
- XGB on N-gram graph and RF on Morgan Fingerprints are the top two models.

Table 5: Generalization performance: Train and test gap on AUC[ROC]

	Morgan RF	Morgan XGB	Morgan DNN	MP graph GCNN	MP graph Weave	N-gram graph RF	N-gram graph XGB	N-gram graph DNN
NR-AR	0.213 ± 0.065	0.209 ± 0.063	0.243 ± 0.057	0.125 ± 0.085	0.115 ± 0.082	0.198 ± 0.082	0.210 ± 0.073	0.093 ± 0.072
NR-AR-LBD	0.136 ± 0.056	0.144 ± 0.051	0.180 ± 0.051	0.115 ± 0.040	0.156 ± 0.036	0.156 ± 0.039	0.142 ± 0.027	0.112 ± 0.049
NR-AhR	0.097 ± 0.028	0.091 ± 0.023	0.146 ± 0.036	0.054 ± 0.020	0.059 ± 0.045	0.110 ± 0.020	0.102 ± 0.020	0.048 ± 0.025
NR-Aromatase	0.172 ± 0.070	0.190 ± 0.064	0.258 ± 0.099	0.112 ± 0.057	0.099 ± 0.064	0.155 ± 0.066	0.148 ± 0.053	0.066 ± 0.077
NR-ER	0.274 ± 0.019	0.238 ± 0.021	0.307 ± 0.018	0.129 ± 0.033	0.129 ± 0.036	0.273 ± 0.043	0.267 ± 0.035	0.032 ± 0.026
NR-ER-LBD	0.184 ± 0.051	0.205 ± 0.056	0.228 ± 0.019	0.134 ± 0.027	0.119 ± 0.061	0.189 ± 0.062	0.181 ± 0.036	0.105 ± 0.069
NR-PPAR- gamma	0.161 ± 0.042	0.199 ± 0.094	0.244 ± 0.037	0.142 ± 0.110	0.147 ± 0.055	0.197 ± 0.105	0.175 ± 0.104	0.147 ± 0.129
SR-ARE	0.181 ± 0.039	0.166 ± 0.041	0.219 ± 0.046	0.118 ± 0.053	0.099 ± 0.064	0.192 ± 0.028	0.174 ± 0.024	0.075 ± 0.080
SR-ATAD5	0.143 ± 0.051	0.167 ± 0.066	0.262 ± 0.079	0.125 ± 0.042	0.129 ± 0.043	0.159 ± 0.032	0.163 ± 0.041	0.123 ± 0.032
SR-HSE	0.206 ± 0.028	0.222 ± 0.047	0.269 ± 0.034	0.155 ± 0.052	0.155 ± 0.037	0.225 ± 0.048	0.214 ± 0.065	0.095 ± 0.074
SR-MMP	0.114 ± 0.019	0.109 ± 0.026	0.144 ± 0.027	0.069 ± 0.029	0.063 ± 0.024	0.105 ± 0.017	0.091 ± 0.017	0.047 ± 0.037
SR-p53	0.151 ± 0.033	0.170 ± 0.058	0.241 ± 0.034	0.107 ± 0.058	0.112 ± 0.069	0.167 ± 0.033	0.157 ± 0.054	0.064 ± 0.050
average	0.169 ± 0.046	0.176 ± 0.042	0.229 ± 0.047	0.115 ± 0.027	0.115 ± 0.030	0.177 ± 0.044	0.169 ± 0.046	0.084 ± 0.033

- Mean and standard deviation on 5-fold cross validation.
- Top three results are **bolded**.
- MP graph is short for Message-passing graph.

Figure A.2: Tukey's Test on Generalization Gap

- The generalization gap between train ROC and test ROC.

Table A.1: Performance of Different Vector Embeddings

Original feature presented in baseline papers, and new feature presented in N-gram graph

	GCNN original feature	GCNN new feature	Weave original feature	Weave new feature
NR-AR	0.793 \pm 0.068	0.802 \pm 0.068	0.796 \pm 0.057	0.792 \pm 0.076
NR-AR-LBD	0.858 \pm 0.037	0.852 \pm 0.042	0.816 \pm 0.047	0.830 \pm 0.041
NR-AhR	0.896 \pm 0.018	0.898 \pm 0.018	0.869 \pm 0.036	0.880 \pm 0.034
NR-Aromatase	0.824 \pm 0.052	0.819 \pm 0.047	0.830 \pm 0.050	0.834 \pm 0.066
NR-ER	0.734 \pm 0.036	0.735 \pm 0.023	0.729 \pm 0.025	0.734 \pm 0.020
NR-ER-LBD	0.805 \pm 0.024	0.803 \pm 0.017	0.804 \pm 0.030	0.809 \pm 0.013
NR-PPAR-gamma	0.821 \pm 0.105	0.799 \pm 0.093	0.803 \pm 0.064	0.801 \pm 0.083
SR-ARE	0.782 \pm 0.040	0.793 \pm 0.046	0.790 \pm 0.049	0.779 \pm 0.022
SR-ATAD5	0.839 \pm 0.037	0.838 \pm 0.029	0.823 \pm 0.041	0.800 \pm 0.038
SR-HSE	0.774 \pm 0.036	0.786 \pm 0.030	0.771 \pm 0.036	0.772 \pm 0.038
SR-MMP	0.888 \pm 0.018	0.887 \pm 0.027	0.886 \pm 0.022	0.885 \pm 0.015
SR-p53	0.840 \pm 0.053	0.836 \pm 0.051	0.813 \pm 0.065	0.813 \pm 0.052
average	0.821 \pm 0.045	0.821 \pm 0.043	0.811 \pm 0.039	0.811 \pm 0.041

- We compare original GCNN and Weave and ones with feature vector in Eq (1)
- Here's the result for Tukey's Test:

Table A.2: Tukey's Test on Models with Different Vector Embeddings

Group 1	Group 2	mean diff	reject
GCNN new feature	GCNN original feature	-0.0012	False
Weave new feature	Weave original feature	0.0008	False

- Null hypothesis is that both means are same. If reject=False, then we accept the null hypothesis: two means are same.
- This result shows that two vector embeddings (original paper and Eq (1) in the submitted version) have similar effects.

