Supporting Information

Analyzing Learned Molecular Representations for Property Prediction

Kevin Yang,[*,†] Kyle Swanson,[*,†] Wengong Jin,[†] Connor Coley,[‡] Philipp Eiden,[¶] Hua Gao,[§] Angel Guzman-Perez,[§] Timothy Hopper,[§] Brian Kelley,[‖] Miriam Mathea,[¶] Andrew Palmer,[¶] Volker Settels,[¶] Tommi Jaakkola,[†] Klavs Jensen,[‡] and Regina Barzilay[†]

†*Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, United States*

‡*Department of Chemical Engineering, MIT, Cambridge, MA 02139, United States*

¶*BASF SE, Ludwigshafen 67063, Germany*

§*Amgen Inc., Cambridge, MA 02141, United States*

‖*Novartis Institutes for BioMedical Research, Cambridge, MA 02139, United States*

E-mail: yangk@mit.edu; swansonk@mit.edu

Code

Our code is publicly available at `https://github.com/swansonk14/chemprop`, which also includes a web interface that supports non-programmatic training and predicting with our model. Code for computing the RDKit features is available at `https://github.com/bp-kelley/descriptastorus`. A public web demonstration of our model's prediction capability on public datasets is available at `http://chemprop.csail.mit.edu` (see Figure S1 for a screenshot of the demonstration).

In addition, we ran the baseline from Mayr et al.[1] using the code at `https://github.com/yangkevin2/lsc_experiments`.

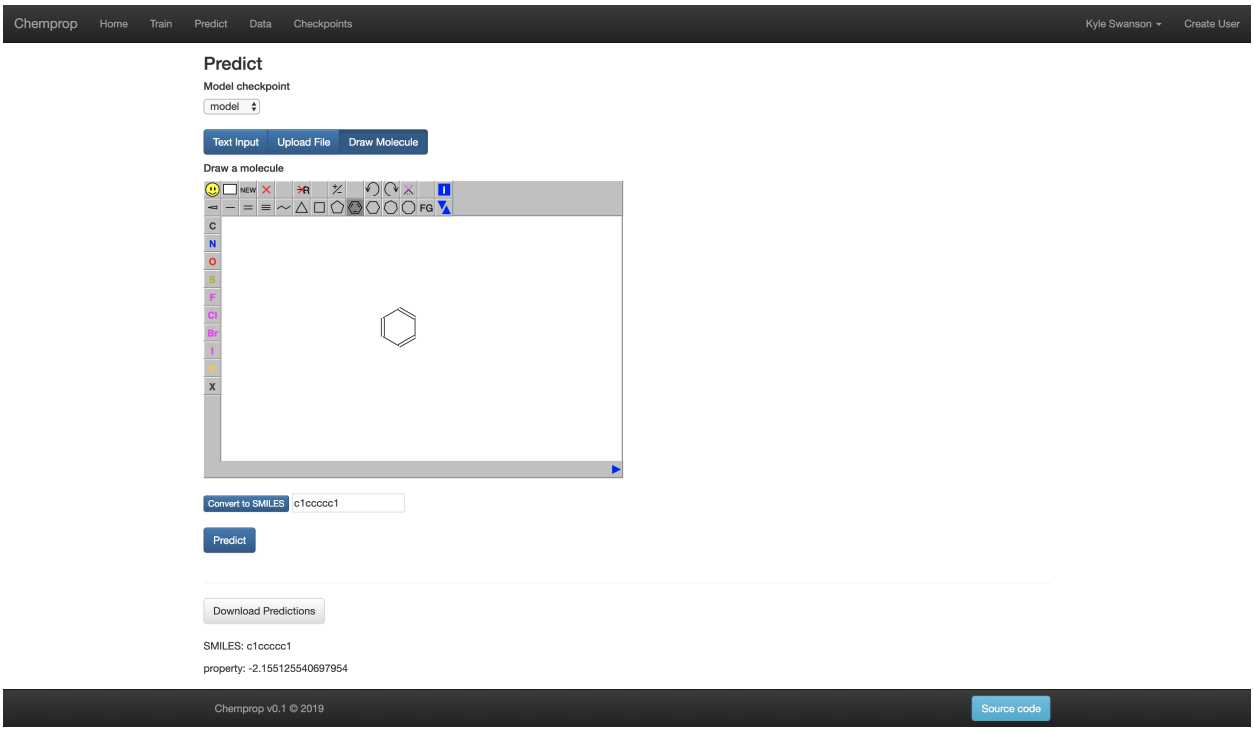


Figure S1: Screenshot of the prediction page of our web interface.

Additional Dataset Statistics

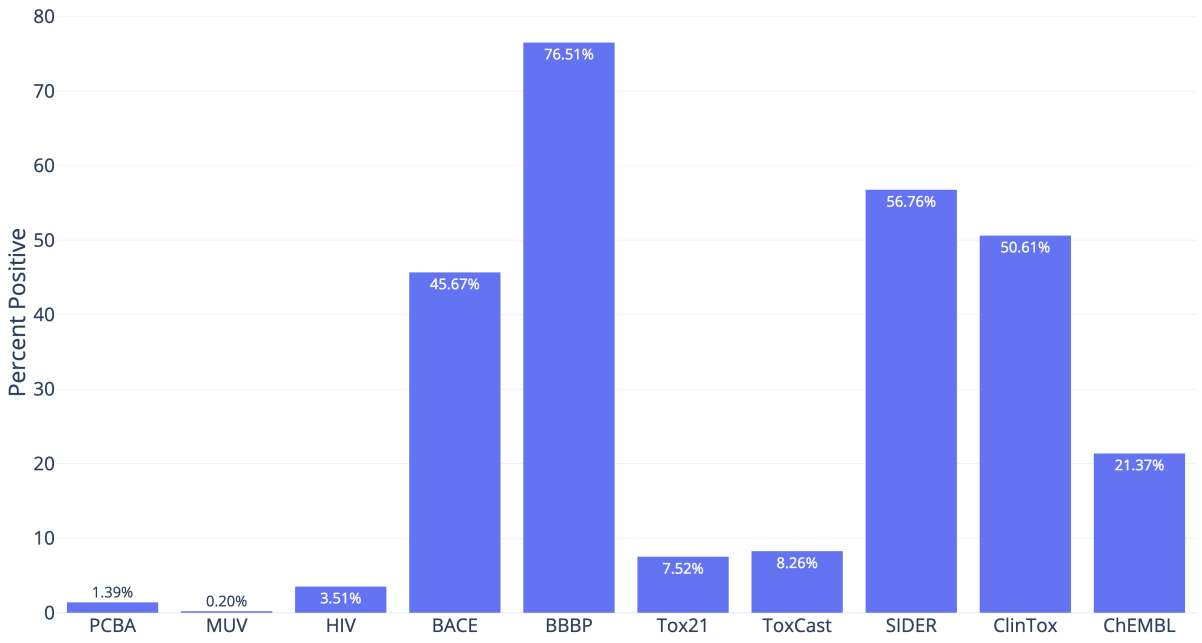


Figure S2: Class balance on the publicly available classification datasets. The Y-axis is the average percent of positives in the tasks in a dataset, weighted by the number of molecules with known values for each task.

In addition to the class balance statistics in Figure S2, we also analyze the class imbalances introduced in both random and scaffold splits, quantifying the imbalance using the following metric. Let r be the fraction of the less common class (0 or 1) in the full dataset, and let r_t be the same fraction for a particular test fold for one of our splits. Then we measure imbalance for a particular property and test fold by $\max\left(\frac{r}{r_t}, \frac{r_t}{r}\right)$ (that is, the ratio of the larger over the smaller) and average across all properties and test folds for each dataset. Thus, a higher metric indicates greater class imbalance introduced by data splitting. On rare occasions for both the random and scaffold split, r_t is 0 due to the sparsity and imbalance of some properties in datasets that contain a large number of properties. We omit these cases from the average, and for each dataset we denote the number of property-test fold pairs for which this occurred. (Such properties are omitted when calculating the average AUC for that particular test fold.) Overall, as indicated in Figure S3 and Tables S1 and S2, the scaffold

split is more imbalanced than the random split, but the numbers are comparable. On no dataset does the scaffold split test set's class balance differ (on average) from the full dataset's class balance by more than a factor of 2.

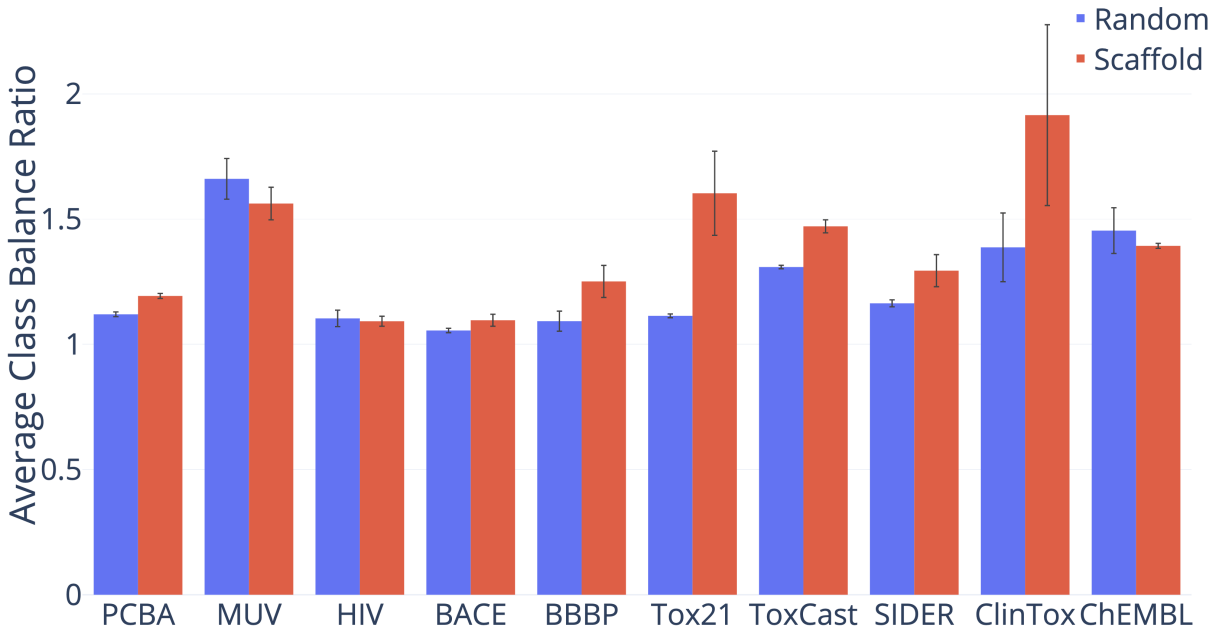


Figure S3: Ratio of class balance between the full dataset and the test set on the publicly available classification datasets (see text for definition of the ratio).

Table S1: Class Balance Ratio (Random Split).

Dataset	Class Balance Ratio	# Failed Property-Fold Pairs	# Property-Fold Pairs
PCBA	1.120 ± 0.017	0	384
MUV	1.661 ± 0.141	2	51
HIV	1.103 ± 0.057	0	3
BACE	1.055 ± 0.028	0	10
BBBP	1.092 ± 0.127	0	10
Tox21	1.114 ± 0.024	0	120
ToxCast	1.309 ± 0.022	103	6170
SIDER	1.164 ± 0.044	2	270
ClinTox	1.387 ± 0.434	0	20
ChEMBL	1.454 ± 0.158	230	3930

Table S2: Class Balance Ratio (Scaffold Split).

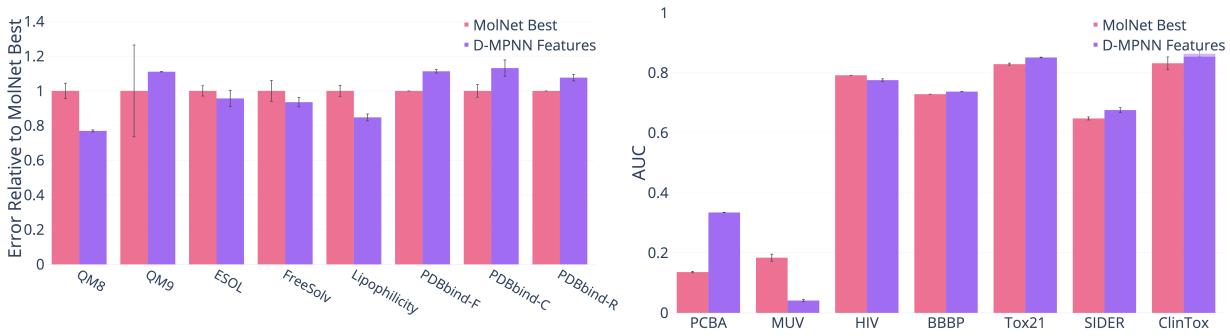
Dataset	Class Balance Ratio	# Failed Property-Fold Pairs	# Property-Fold Pairs
PCBA	1.193 ± 0.018	0	384
MUV	1.562 ± 0.112	3	51
HIV	1.092 ± 0.035	0	3
BACE	1.096 ± 0.076	0	10
BBBP	1.251 ± 0.203	0	10
Tox21	1.603 ± 0.530	0	120
ToxCast	1.471 ± 0.082	180	6170
SIDER	1.294 ± 0.204	1	270
ClinTox	1.915 ± 1.142	0	20
ChEMBL	1.393 ± 0.017	284	3930

Comparison to Baselines

Note: All p-values for comparisons involving MoleculeNet,[2] Mayr et al.,[1] and different dataset split types are from a one-sided Welch's t-test. All remaining p-values are from a one-sided Wilcoxon signed-rank test. See the main text for more details.

Comparison to MoleculeNet

Comparison between our D-MPNN with RDKit features and the best model from MoleculeNet using the splits from MoleculeNet.[2] We were unable to reproduce the splits from MoleculeNet on QM7, BACE, and ToxCast, so we leave out those datasets. The QM8, QM9, and PDBbind datasets include 3D coordinates that our model does not use but some MoleculeNet models may use.



(a) Regression Datasets (lower = better). (b) Classification Datasets (higher = better).

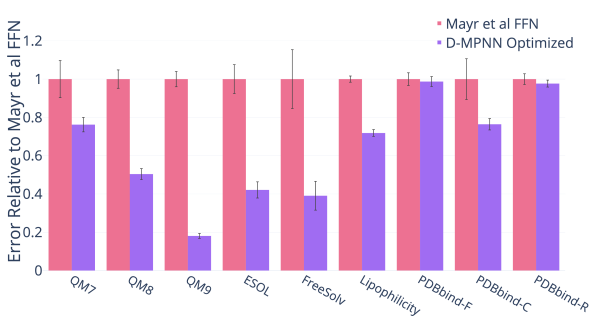
Figure S4: Comparison to MoleculeNet models on Wu et al.'s[2] original splits.

Table S3: Comparison to MoleculeNet models on Wu et al.'s[2] original splits.

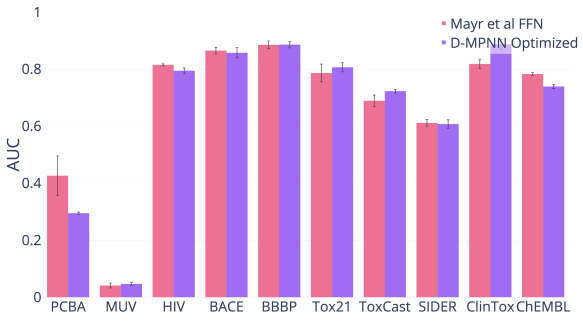
Dataset	Metric	Split Type	MolNet[2] Best	D-MPNN Features
QM8	MAE	Random	0.014 ± 0.001	0.011 ± 0.000 (-23.03% p=0.02)
QM9	MAE	Random	2.400 ± 1.100	2.666 ± 0.006 (+11.07% p=0.38)
ESOL	RMSE	Random	0.580 ± 0.030	0.555 ± 0.047 (-4.33% p=0.28)
FreeSolv	RMSE	Random	1.150 ± 0.120	1.075 ± 0.054 (-6.48% p=0.24)
Lipophilicity	RMSE	Random	0.655 ± 0.036	0.555 ± 0.023 (-15.26% p=0.02)
PDBbind-F	RMSE	Time	1.250 ± 0.000	1.391 ± 0.012 (+11.31% p=0.00)
PDBbind-C	RMSE	Time	1.920 ± 0.070	2.173 ± 0.090 (+13.15% p=0.02)
PDBbind-R	RMSE	Time	1.380 ± 0.000	1.486 ± 0.026 (+7.65% p=0.01)
PCBA	PRC-AUC	Random	0.136 ± 0.004	0.335 ± 0.001 (+146.05% p=0.00)
MUV	PRC-AUC	Random	0.1840 ± 0.0200	0.041 ± 0.007 (-77.66% p=0.00)
HIV	ROC-AUC	Scaffold	0.792 ± 0.000	0.776 ± 0.008 (-2.05% p=0.05)
BBBP	ROC-AUC	Scaffold	0.729 ± 0.000	0.738 ± 0.001 (+1.17% p=0.00)
Tox21	ROC-AUC	Random	0.829 ± 0.006	0.851 ± 0.002 (+2.69% p=0.01)
SIDER	ROC-AUC	Random	0.648 ± 0.009	0.676 ± 0.014 (+4.32% p=0.04)
ClinTox	ROC-AUC	Random	0.832 ± 0.037	0.864 ± 0.017 (+3.83% p=0.18)

Comparison to Mayr et al.[1]

Comparison between our best single model (i.e. optimized hyperparameters and optionally RDKit features but without ensembling) and the feed-forward network (FFN) architecture of Mayr et al.[1] using their best descriptor set. We ran this comparison only on scaffold split due to computational cost; as the model of Mayr et al.[1] uses hyperparameter optimization spanning 40 different parameter settings for 300 epochs each, their hyperparameter optimization is actually more expensive than that of our D-MPNN.



(a) Regression Datasets (Scaffold Split, lower = better).



(b) Classification Datasets (Scaffold Split, higher = better).

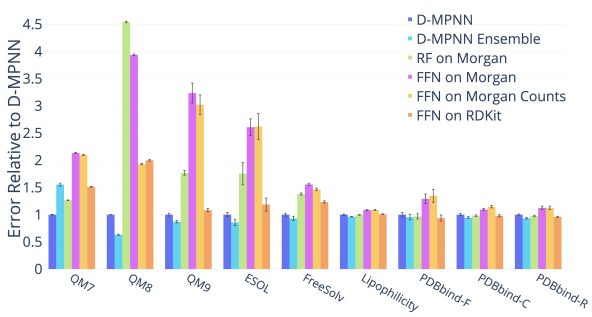
Figure S5: Comparison to Mayr et al..[1]

Table S4: Comparison to Mayr et al.[1] (Scaffold Split).

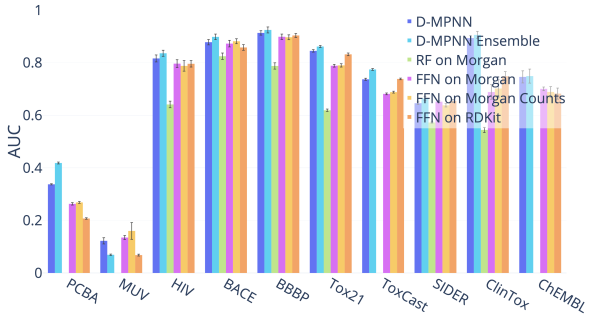
Dataset	Metric	Mayr et al.[1] FFN	D-MPNN Optimized
QM7	MAE	119.220 ± 36.490	90.869 ± 14.199 (-23.78% p=0.03)
QM8	MAE	0.024 ± 0.004	0.012 ± 0.002 (-49.62% p=0.00)
QM9	MAE	13.117 ± 0.907	2.370 ± 0.294 (-81.93% p=0.00)
ESOL	RMSE	2.344 ± 0.561	0.987 ± 0.314 (-57.89% p=0.00)
FreeSolv	RMSE	4.513 ± 2.196	1.763 ± 1.075 (-60.93% p=0.00)
Lipophilicity	RMSE	0.856 ± 0.044	0.615 ± 0.048 (-28.17% p=0.00)
PDBbind-F	RMSE	1.415 ± 0.148	1.397 ± 0.117 (-1.28% p=0.39)
PDBbind-C	RMSE	2.507 ± 0.845	1.916 ± 0.236 (-23.59% p=0.03)
PDBbind-R	RMSE	1.514 ± 0.135	1.479 ± 0.087 (-2.34% p=0.26)
PCBA	PRC-AUC	0.426 ± 0.120	0.295 ± 0.006 (-30.87% p=0.13)
MUV	PRC-AUC	0.041 ± 0.015	0.047 ± 0.009 (+13.48% p=0.34)
HIV	ROC-AUC	0.815 ± 0.006	0.794 ± 0.017 (-2.53% p=0.11)
BACE	ROC-AUC	0.865 ± 0.037	0.857 ± 0.057 (-0.89% p=0.37)
BBBP	ROC-AUC	0.885 ± 0.043	0.886 ± 0.036 (+0.10% p=0.48)
Tox21	ROC-AUC	0.786 ± 0.099	0.806 ± 0.050 (+2.55% p=0.30)
ToxCast	ROC-AUC	0.689 ± 0.063	0.723 ± 0.020 (+4.86% p=0.08)
SIDER	ROC-AUC	0.611 ± 0.038	0.607 ± 0.047 (-0.67% p=0.42)
ClinTox	ROC-AUC	0.818 ± 0.050	0.887 ± 0.058 (+8.44% p=0.01)
ChEMBL	ROC-AUC	0.783 ± 0.008	0.739 ± 0.012 (-5.56% p=0.01)

Comparison to Other Baselines

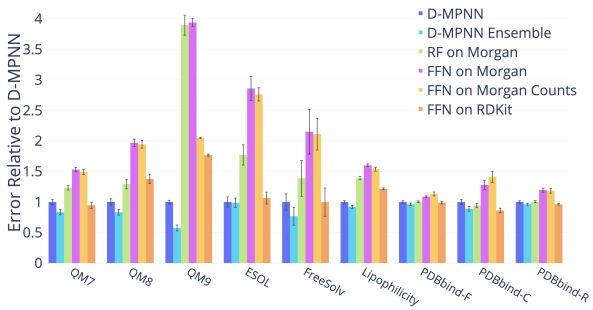
Comparison to several baselines using feed-forward neural networks on molecular fingerprints or descriptors.



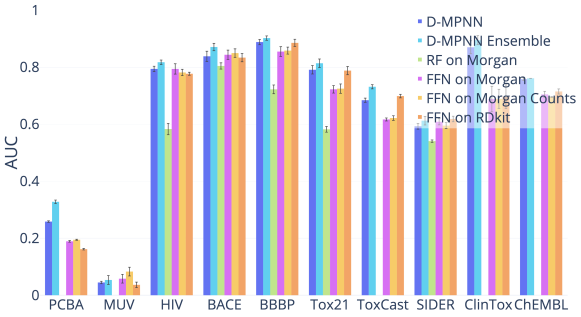
(a) Regression Datasets (Random Split, lower = better).



(b) Classification Datasets (Random Split, higher = better).



(c) Regression Datasets (Scaffold Split, lower = better).



(d) Classification Datasets (Scaffold Split, higher = better).

Figure S6: Comparison to Baselines.

Table S5: Comparison to Baselines, Part I (Random Split).

Dataset	**Metric**	**D-MPNN**	**D-MPNN Ensemble**
QM7	MAE	66.475 ± 2.088	59.379 ± 2.315 (-10.67% p=0.00)
QM8	MAE	0.011 ± 0.000	0.008 ± 0.000 (-25.01% p=0.00)
QM9	MAE	3.101 ± 0.010	1.959 ± 0.066 (-36.82% p=0.00)
ESOL	RMSE	0.665 ± 0.052	0.578 ± 0.046 (-13.08% p=0.00)
FreeSolv	RMSE	1.167 ± 0.150	0.998 ± 0.207 (-14.43% p=0.00)
Lipophilicity	RMSE	0.596 ± 0.050	0.555 ± 0.067 (-6.83% p=0.00)
PDBbind-F	RMSE	1.311 ± 0.034	1.262 ± 0.031 (-3.75% p=0.00)
PDBbind-C	RMSE	2.151 ± 0.285	2.057 ± 0.353 (-4.39% p=0.15)
PDBbind-R	RMSE	1.395 ± 0.087	1.322 ± 0.077 (-5.22% p=0.00)
PCBA	PRC-AUC	0.337 ± 0.004	0.418 ± 0.006 (+24.03% p=0.00)
MUV	PRC-AUC	0.122 ± 0.020	0.069 ± 0.005 (-43.38% p=1.00)
HIV	ROC-AUC	0.816 ± 0.023	0.836 ± 0.020 (+2.40% p=0.01)
BACE	ROC-AUC	0.878 ± 0.032	0.898 ± 0.034 (+2.31% p=0.00)
BBBP	ROC-AUC	0.913 ± 0.026	0.925 ± 0.036 (+1.23% p=0.01)
Tox21	ROC-AUC	0.845 ± 0.015	0.861 ± 0.012 (+1.95% p=0.00)
ToxCast	ROC-AUC	0.737 ± 0.013	0.774 ± 0.011 (+5.09% p=0.00)
SIDER	ROC-AUC	0.646 ± 0.016	0.664 ± 0.021 (+2.79% p=0.01)
ClinTox	ROC-AUC	0.894 ± 0.027	0.906 ± 0.043 (+1.33% p=0.05)
ChEMBL	ROC-AUC	0.746 ± 0.040	0.749 ± 0.046 (+0.41% p=0.00)

Table S6: Comparison to Baselines, Part II (Random Split).

Dataset	**RF on Morgan**	**FFN on Morgan**
QM7	124.667 ± 3.928 (+87.54% p=0.00)	134.720 ± 3.724 (+102.66% p=0.00)
QM8	0.014 ± 0.000 (+26.49% p=0.00)	0.024 ± 0.000 (+113.72% p=0.00)
QM9	14.089 ± 0.079 (+354.34% p=0.00)	12.215 ± 0.076 (+293.92% p=0.00)
ESOL	1.176 ± 0.086 (+76.86% p=0.00)	2.152 ± 0.386 (+223.59% p=0.00)
FreeSolv	2.048 ± 0.769 (+75.50% p=0.00)	3.043 ± 0.567 (+160.82% p=0.00)
Lipophilicity	0.823 ± 0.035 (+38.06% p=0.00)	0.928 ± 0.044 (+55.65% p=0.00)
PDBbind-F	1.309 ± 0.035 (-0.15% p=1.00)	1.423 ± 0.023 (+8.54% p=0.00)
PDBbind-C	2.083 ± 0.324 (-3.20% p=0.72)	2.778 ± 0.599 (+29.12% p=0.00)
PDBbind-R	1.369 ± 0.064 (-1.90% p=1.00)	1.528 ± 0.093 (+9.50% p=0.00)
PCBA	—	0.263 ± 0.008 (-22.03% p=0.00)
MUV	—	0.135 ± 0.013 (+10.25% p=0.36)
HIV	0.641 ± 0.022 (-21.45% p=0.00)	0.796 ± 0.026 (-2.42% p=0.80)
BACE	0.825 ± 0.039 (-6.08% p=0.00)	0.873 ± 0.040 (-0.61% p=0.11)
BBBP	0.788 ± 0.038 (-13.77% p=0.00)	0.899 ± 0.033 (-1.61% p=0.05)
Tox21	0.619 ± 0.015 (-26.75% p=0.00)	0.788 ± 0.017 (-6.70% p=0.00)
ToxCast	—	0.681 ± 0.011 (-7.52% p=0.00)
SIDER	0.572 ± 0.007 (-11.38% p=0.00)	0.652 ± 0.010 (+0.89% p=0.80)
ClinTox	0.544 ± 0.031 (-39.13% p=0.00)	0.688 ± 0.088 (-22.99% p=0.00)
ChEMBL	—	0.700 ± 0.012 (-6.10% p=0.00)

Table S7: Comparison to Baselines, Part III (Random Split).

Dataset	FFN on Morgan Counts	FFN on RDKit
QM7	123.314 ± 3.936 (+85.50% p=0.00)	75.857 ± 2.447 (+14.11% p=0.00)
QM8	0.023 ± 0.000 (+109.94% p=0.00)	0.017 ± 0.000 (+51.26% p=0.00)
QM9	5.984 ± 0.076 (+92.96% p=0.00)	6.201 ± 0.074 (+99.97% p=0.00)
ESOL	2.009 ± 0.379 (+202.09% p=0.00)	0.721 ± 0.068 (+8.37% p=0.00)
FreeSolv	3.057 ± 0.881 (+161.99% p=0.00)	1.384 ± 0.440 (+18.66% p=0.11)
Lipophilicity	0.874 ± 0.043 (+46.58% p=0.00)	0.735 ± 0.039 (+23.36% p=0.00)
PDBbind-F	1.424 ± 0.032 (+8.64% p=0.00)	1.321 ± 0.029 (+0.76% p=0.00)
PDBbind-C	2.901 ± 0.812 (+34.84% p=0.03)	2.020 ± 0.376 (-6.10% p=0.88)
PDBbind-R	1.599 ± 0.093 (+14.62% p=0.00)	1.367 ± 0.089 (-2.04% p=0.75)
PCBA	0.268 ± 0.006 (-20.43% p=0.00)	0.207 ± 0.005 (-38.59% p=0.00)
MUV	0.160 ± 0.055 (+30.81% p=0.38)	0.068 ± 0.006 (-44.37% p=0.06)
HIV	0.788 ± 0.035 (-3.47% p=0.43)	0.796 ± 0.021 (-2.49% p=0.66)
BACE	0.882 ± 0.030 (+0.41% p=0.41)	0.858 ± 0.034 (-2.28% p=0.02)
BBBP	0.897 ± 0.029 (-1.82% p=0.01)	0.904 ± 0.024 (-1.07% p=0.16)
Tox21	0.790 ± 0.020 (-6.54% p=0.00)	0.832 ± 0.016 (-1.57% p=0.00)
ToxCast	0.688 ± 0.011 (-6.63% p=0.00)	0.738 ± 0.010 (+0.20% p=0.21)
SIDER	0.638 ± 0.020 (-1.21% p=0.07)	0.654 ± 0.019 (+1.33% p=0.90)
ClinTox	0.702 ± 0.105 (-21.49% p=0.00)	0.749 ± 0.055 (-16.26% p=0.00)
ChEMBL	0.689 ± 0.035 (-7.67% p=0.00)	0.682 ± 0.037 (-8.52% p=0.00)

Table S8: Comparison to Baselines, Part I (Scaffold Split).

Dataset	**Metric**	**D-MPNN**	**D-MPNN Ensemble**
QM7	MAE	105.775 ± 13.202	88.201 ± 13.899 (-16.61% p=0.00)
QM8	MAE	0.014 ± 0.002	0.012 ± 0.002 (-16.69% p=0.00)
QM9	MAE	3.451 ± 0.174	1.983 ± 0.289 (-42.53% p=0.00)
ESOL	RMSE	0.980 ± 0.258	0.968 ± 0.237 (-1.21% p=0.00)
FreeSolv	RMSE	2.177 ± 0.914	1.670 ± 1.008 (-23.27% p=0.00)
Lipophilicity	RMSE	0.653 ± 0.046	0.600 ± 0.049 (-8.04% p=0.00)
PDBbind-F	RMSE	1.419 ± 0.089	1.365 ± 0.092 (-3.79% p=0.00)
PDBbind-C	RMSE	2.138 ± 0.278	1.900 ± 0.262 (-11.12% p=0.00)
PDBbind-R	RMSE	1.507 ± 0.095	1.453 ± 0.080 (-3.60% p=0.00)
PCBA	PRC-AUC	0.258 ± 0.005	0.328 ± 0.011 (+26.88% p=0.00)
MUV	PRC-AUC	0.045 ± 0.007	0.053 ± 0.027 (+19.59% p=0.12)
HIV	ROC-AUC	0.794 ± 0.016	0.817 ± 0.013 (+2.94% p=0.00)
BACE	ROC-AUC	0.838 ± 0.056	0.871 ± 0.041 (+3.89% p=0.00)
BBBP	ROC-AUC	0.888 ± 0.029	0.902 ± 0.024 (+1.56% p=0.01)
Tox21	ROC-AUC	0.791 ± 0.047	0.814 ± 0.047 (+2.89% p=0.00)
ToxCast	ROC-AUC	0.684 ± 0.023	0.731 ± 0.023 (+6.89% p=0.00)
SIDER	ROC-AUC	0.593 ± 0.032	0.612 ± 0.047 (+3.31% p=0.03)
ClinTox	ROC-AUC	0.870 ± 0.072	0.895 ± 0.050 (+2.86% p=0.01)
ChEMBL	ROC-AUC	0.758 ± 0.008	0.761 ± 0.000 (+0.39% p=0.00)

Table S9: Comparison to Baselines, Part II (Scaffold Split).

Dataset	**RF on Morgan**	**FFN on Morgan**
QM7	130.146 ± 12.179 (+23.04% p=0.00)	161.956 ± 12.556 (+53.11% p=0.00)
QM8	0.019 ± 0.004 (+29.25% p=0.00)	0.028 ± 0.003 (+96.90% p=0.00)
QM9	13.441 ± 0.980 (+289.49% p=0.00)	13.591 ± 0.386 (+293.86% p=0.00)
ESOL	1.734 ± 0.512 (+76.99% p=0.00)	2.801 ± 0.610 (+185.88% p=0.00)
FreeSolv	3.019 ± 2.021 (+38.72% p=0.00)	4.683 ± 2.518 (+115.12% p=0.00)
Lipophilicity	0.908 ± 0.052 (+38.99% p=0.00)	1.045 ± 0.042 (+60.11% p=0.00)
PDBbind-F	1.425 ± 0.060 (+0.44% p=0.00)	1.544 ± 0.054 (+8.85% p=0.00)
PDBbind-C	2.011 ± 0.240 (-5.93% p=0.95)	2.737 ± 0.518 (+28.06% p=0.00)
PDBbind-R	1.514 ± 0.079 (+0.44% p=0.14)	1.802 ± 0.157 (+19.53% p=0.00)
PCBA	—	0.189 ± 0.005 (-26.83% p=0.00)
MUV	—	0.058 ± 0.027 (+29.71% p=0.31)
HIV	0.583 ± 0.034 (-26.59% p=0.00)	0.794 ± 0.031 (-0.02% p=0.62)
BACE	0.804 ± 0.035 (-4.04% p=0.01)	0.843 ± 0.052 (+0.64% p=0.69)
BBBP	0.722 ± 0.049 (-18.68% p=0.00)	0.855 ± 0.054 (-3.76% p=0.06)
Tox21	0.582 ± 0.031 (-26.42% p=0.00)	0.722 ± 0.041 (-8.67% p=0.00)
ToxCast	—	0.616 ± 0.017 (-9.87% p=0.00)
SIDER	0.540 ± 0.013 (-8.79% p=0.00)	0.608 ± 0.035 (+2.69% p=0.90)
ClinTox	—	0.688 ± 0.142 (-20.88% p=0.00)
ChEMBL	—	0.705 ± 0.018 (-7.03% p=0.00)

Table S10: Comparison to Baselines, Part III (Scaffold Split).

Dataset	FFN on Morgan Counts	FFN on RDKit
QM7	157.856 ± 14.847 (+49.24% p=0.00)	100.180 ± 16.776 (-5.29% p=0.00)
QM8	0.028 ± 0.003 (+94.49% p=0.00)	0.020 ± 0.004 (+37.75% p=0.00)
QM9	7.074 ± 0.066 (+104.99% p=0.00)	6.099 ± 0.099 (+76.75% p=0.00)
ESOL	2.706 ± 0.334 (+176.14% p=0.00)	1.043 ± 0.304 (+6.48% p=0.00)
FreeSolv	4.596 ± 1.790 (+111.16% p=0.00)	2.172 ± 1.593 (-0.21% p=0.88)
Lipophilicity	1.003 ± 0.068 (+53.67% p=0.00)	0.792 ± 0.032 (+21.34% p=0.00)
PDBbind-F	1.605 ± 0.143 (+13.13% p=0.00)	1.402 ± 0.079 (-1.20% p=0.21)
PDBbind-C	3.015 ± 0.636 (+41.06% p=0.00)	1.842 ± 0.252 (-13.80% p=1.00)
PDBbind-R	1.781 ± 0.175 (+18.18% p=0.00)	1.451 ± 0.054 (-3.73% p=1.00)
PCBA	0.195 ± 0.003 (-24.69% p=0.00)	0.161 ± 0.005 (-37.49% p=0.00)
MUV	0.083 ± 0.027 (+85.65% p=0.21)	0.036 ± 0.016 (-18.44% p=0.33)
HIV	0.781 ± 0.019 (-1.58% p=0.17)	0.777 ± 0.009 (-2.15% p=0.56)
BACE	0.849 ± 0.047 (+1.34% p=0.87)	0.833 ± 0.046 (-0.55% p=0.10)
BBBP	0.858 ± 0.039 (-3.38% p=0.03)	0.885 ± 0.040 (-0.32% p=0.35)
Tox21	0.725 ± 0.052 (-8.37% p=0.00)	0.788 ± 0.046 (-0.37% p=0.30)
ToxCast	0.621 ± 0.025 (-9.16% p=0.00)	0.699 ± 0.020 (+2.12% p=0.98)
SIDER	0.595 ± 0.033 (+0.41% p=0.30)	0.618 ± 0.031 (+4.32% p=1.00)
ClinTox	0.689 ± 0.106 (-20.80% p=0.00)	0.697 ± 0.131 (-19.91% p=0.00)
ChEMBL	0.695 ± 0.002 (-8.25% p=0.00)	0.715 ± 0.015 (-5.70% p=0.00)

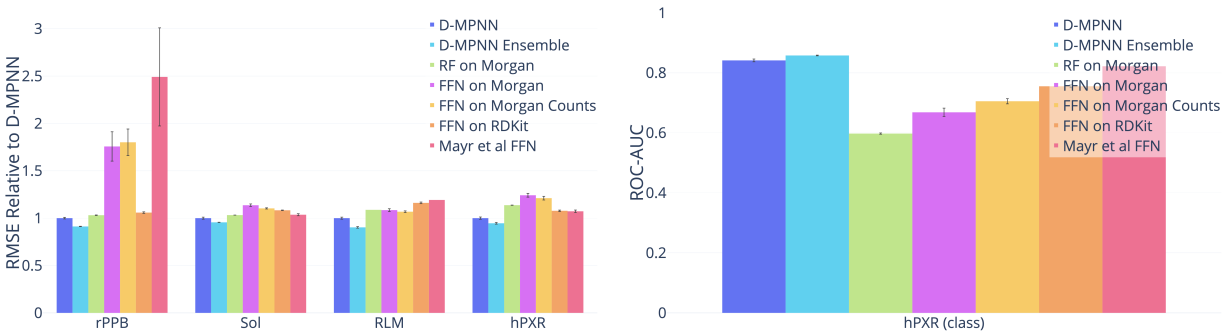
Proprietary Datasets

Amgen

Comparison of our D-MPNN in both its unoptimized and optimized form against baseline models on Amgen internal datasets using a time split of the data. Note that rPPB is in logit while Sol and RLM are in $\log_{10}$.

Table S11: Comparison to Baselines on Amgen, Part I. Note: The metric for hPXR (class) is ROC-AUC; all others are RMSE. *Only one run.

Dataset	D-MPNN	D-MPNN Ensemble	RF on Morgan
rPPB	1.057 ± 0.026	0.964 ± 0.007 (-8.78%)	1.089 ± 0.009 (+3.11%)
Sol	0.706 ± 0.013	0.675 ± 0.001 (-4.42%)	0.729 ± 0.000 (+3.22%)
RLM	0.331 ± 0.004	0.298 ± 0.003 (-9.72%)	0.360* (+8.78%)
hPXR	36.584 ± 0.751	34.604 ± 0.568 (-5.41%)	41.600 ± 0.070 (+13.71%)
hPXR (class)	0.842 ± 0.008	0.858 ± 0.002 (+1.95%)	0.598 ± 0.004 (-28.98%)



(a) Regression Datasets (lower = better). (b) Classification Datasets (higher = better).

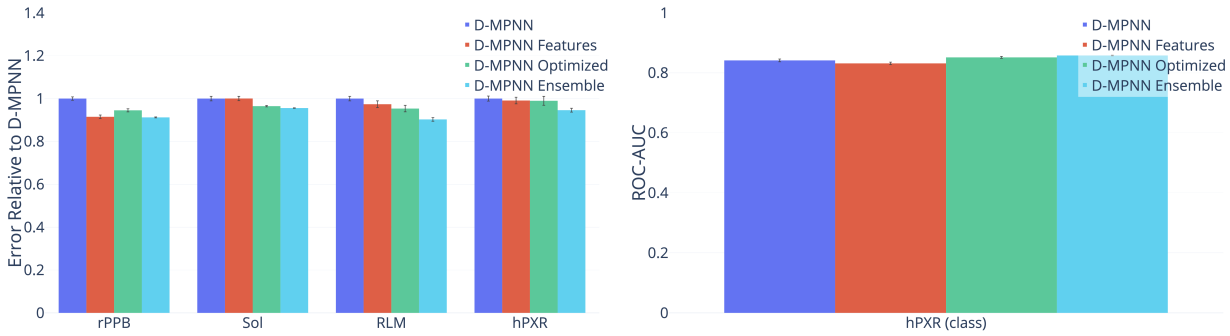
Figure S7: Comparison to Baselines on Amgen.

Table S12: Comparison to Baselines on Amgen, Part II. Note: The metric for hPXR (class) is ROC-AUC; all others are MSE.

Dataset	FFN on Morgan	FFN on Morgan Counts
rPPB	1.856 ± 0.517 (+75.69%)	1.903 ± 0.468 (+80.09%)
Sol	0.802 ± 0.017 (+13.64%)	0.779 ± 0.008 (+10.31%)
RLM	0.359 ± 0.005 (+8.48%)	0.353 ± 0.003 (+6.81%)
hPXR	45.428 ± 1.255 (+24.17%)	44.305 ± 1.226 (+21.10%)
hPXR (class)	0.669 ± 0.024 (-20.56%)	0.705 ± 0.015 (-16.18%)

Table S13: Comparison to Baselines on Amgen, Part III. Note: The metric for hPXR (class) is ROC-AUC; all others are MSE. *Only one run.

Dataset	FFN on RDKit	Mayr et al.[1]
rPPB	1.119 ± 0.027 (+5.87%)	2.632 ± 1.730 (+149.12%)
Sol	0.765 ± 0.003 (+8.30%)	0.732 ± 0.013 (+3.62%)
RLM	0.384 ± 0.003 (+16.14%)	0.394* (+19.20%)
hPXR	39.426 ± 0.465 (+7.77%)	39.230 ± 0.836 (+7.23%)
hPXR (class)	0.755 ± 0.008 (-10.25%)	0.822 ± 0.004 (-2.32%)



(a) Regression Datasets (lower = better). (b) Classification Datasets (higher = better).

Figure S8: Optimizations on Amgen.

Amgen Model Optimizations

Table S14: Optimizations on Amgen, Part I. Note: The metric for hPXR (class) is ROC-AUC; all others are RMSE.

Dataset	D-MPNN	D-MPNN Features
rPPB	1.057 ± 0.026	0.967 ± 0.028 (-8.50%)
Sol	0.706 ± 0.013	0.706 ± 0.013 (+0.04%)
RLM	0.331 ± 0.004	0.322 ± 0.005 (-2.64%)
hPXR	36.584 ± 0.751	36.252 ± 0.980 (-0.91%)
hPXR (class)	0.842 ± 0.008	0.832 ± 0.007 (-1.18%)

Table S15: Optimizations on Amgen, Part II. Note: The metric for hPXR (class) is ROC-AUC; all others are RMSE.

Dataset	D-MPNN Optimized	D-MPNN Ensemble
rPPB	0.999 ± 0.024 (-5.48%)	0.964 ± 0.007 (-8.78%)
Sol	0.681 ± 0.004 (-3.54%)	0.675 ± 0.001 (-4.42%)
RLM	0.315 ± 0.005 (-4.66%)	0.298 ± 0.003 (-9.72%)
hPXR	36.206 ± 1.326 (-1.03%)	34.604 ± 0.568 (-5.41%)
hPXR (class)	0.852 ± 0.006 (+1.20%)	0.858 ± 0.002 (+1.95%)

BASF

Comparison of our D-MPNN in both its unoptimized and optimized form against baseline models on BASF internal datasets using a scaffold split of the data.

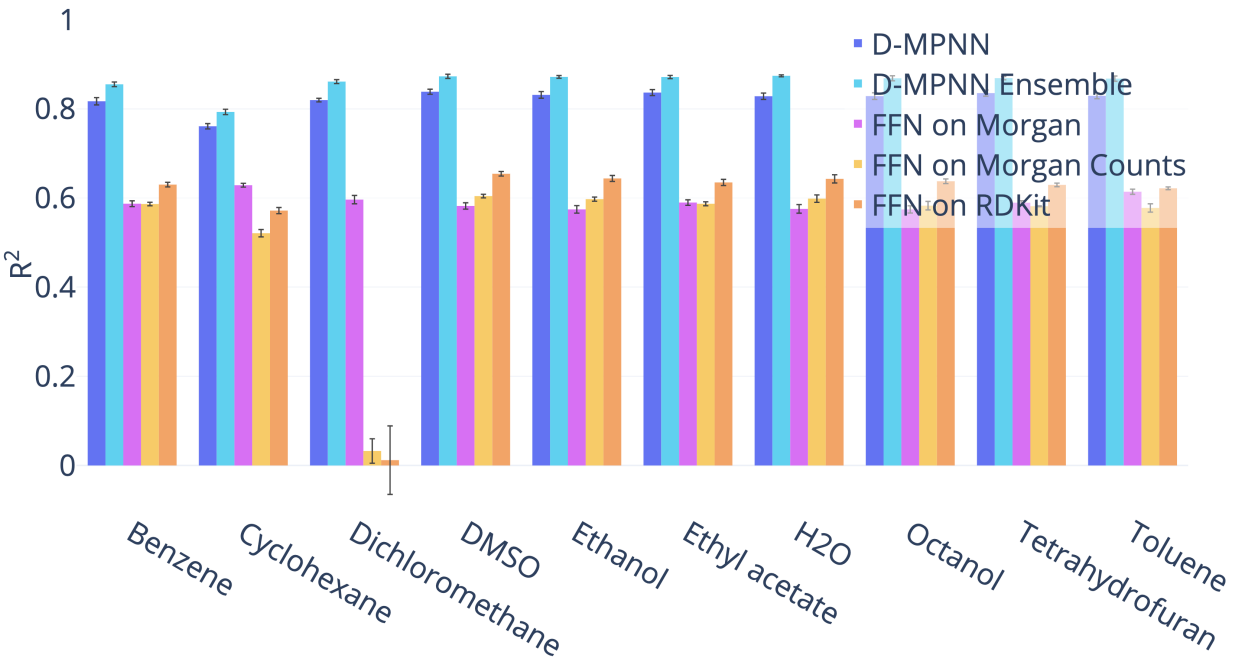


Figure S9: Comparison to Baselines on BASF (higher = better).

Table S16: Comparison to Baselines on BASF, Part I. Note: All numbers are R^2.

Dataset	D-MPNN	D-MPNN Ensemble	RF on Morgan
Benzene	0.817 ± 0.014	0.855 ± 0.009 (+4.65%)	0.587 ± 0.011 (-28.13%)
Cyclohexane	0.761 ± 0.010	0.793 ± 0.011 (+4.22%)	0.629 ± 0.007 (-17.38%)
Dichloromethane	0.820 ± 0.007	0.861 ± 0.008 (+5.06%)	0.596 ± 0.016 (-27.23%)
DMSO	0.838 ± 0.010	0.873 ± 0.008 (+4.12%)	0.582 ± 0.012 (-30.57%)
Ethanol	0.831 ± 0.013	0.872 ± 0.005 (+4.85%)	0.574 ± 0.015 (-30.90%)
Ethyl acetate	0.837 ± 0.012	0.871 ± 0.006 (+4.17%)	0.590 ± 0.011 (-29.49%)
H2O	0.828 ± 0.012	0.874 ± 0.004 (+5.53%)	0.576 ± 0.017 (-30.51%)
Octanol	0.829 ± 0.013	0.869 ± 0.010 (+4.82%)	0.574 ± 0.014 (-30.68%)
Tetrahydrofuran	0.835 ± 0.012	0.869 ± 0.006 (+4.01%)	0.589 ± 0.008 (-29.44%)
Toluene	0.829 ± 0.011	0.868 ± 0.010 (+4.65%)	0.614 ± 0.010 (-25.95%)

Table S17: Comparison to Baselines on BASF, Part II. Note: All numbers are R^2.

Dataset	FFN on Morgan	FFN on Morgan Counts	FFN on RDKit
Benzene	0.587 ± 0.007 (-28.21%)	0.630 ± 0.009 (-22.89%)	0.742 ± 0.007 (-9.23%)
Cyclohexane	0.521 ± 0.014 (-31.53%)	0.572 ± 0.012 (-24.86%)	0.682 ± 0.007 (-10.32%)
Dichloromethane	0.032 ± 0.047 (-96.04%)	0.012 ± 0.133 (-98.55%)	0.695 ± 0.014 (-15.15%)
DMSO	0.604 ± 0.007 (-27.92%)	0.654 ± 0.009 (-21.95%)	0.755 ± 0.007 (-9.92%)
Ethanol	0.597 ± 0.008 (-28.13%)	0.644 ± 0.012 (-22.54%)	0.755 ± 0.005 (-9.21%)
Ethyl acetate	0.587 ± 0.008 (-29.82%)	0.635 ± 0.012 (-24.10%)	0.748 ± 0.009 (-10.58%)
H2O	0.599 ± 0.014 (-27.73%)	0.643 ± 0.016 (-22.38%)	0.754 ± 0.002 (-8.99%)
Octanol	0.583 ± 0.017 (-29.66%)	0.638 ± 0.009 (-23.05%)	0.749 ± 0.007 (-9.60%)
Tetrahydrofuran	0.582 ± 0.002 (-30.38%)	0.629 ± 0.007 (-24.65%)	0.747 ± 0.013 (-10.55%)
Toluene	0.578 ± 0.016 (-30.34%)	0.622 ± 0.005 (-25.02%)	0.756 ± 0.005 (-8.82%)

BASF Model Optimization

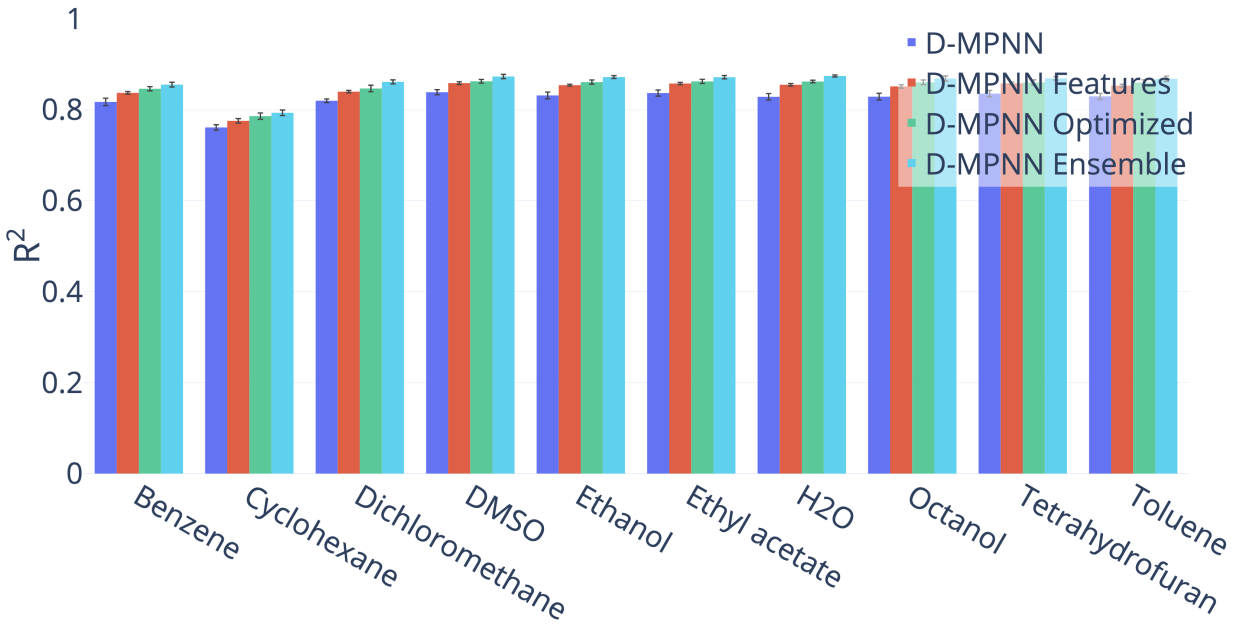


Figure S10: Optimizations on BASF (higher = better).

Table S18: Optimizations on BASF, Part I. Note: All numbers are R^2.

Dataset	D-MPNN	D-MPNN Features
Benzene	0.817 ± 0.014	0.837 ± 0.005 (+2.43%)
Cyclohexane	0.761 ± 0.010	0.775 ± 0.009 (+1.90%)
Dichloromethane	0.820 ± 0.007	0.840 ± 0.005 (+2.44%)
DMSO	0.838 ± 0.010	0.858 ± 0.005 (+2.38%)
Ethanol	0.831 ± 0.013	0.854 ± 0.004 (+2.72%)
Ethyl acetate	0.837 ± 0.012	0.858 ± 0.004 (+2.51%)
H2O	0.828 ± 0.012	0.855 ± 0.005 (+3.20%)
Octanol	0.829 ± 0.013	0.851 ± 0.007 (+2.72%)
Tetrahydrofuran	0.835 ± 0.012	0.858 ± 0.005 (+2.71%)
Toluene	0.829 ± 0.011	0.853 ± 0.005 (+2.85%)

Table S19: Optimizations on BASF, Part II. Note: All numbers are R^2.

Dataset	D-MPNN Optimized	D-MPNN Ensemble
Benzene	0.846 ± 0.008 (+3.54%)	0.855 ± 0.009 (+4.65%)
Cyclohexane	0.786 ± 0.012 (+3.27%)	0.793 ± 0.011 (+4.22%)
Dichloromethane	0.847 ± 0.013 (+3.28%)	0.861 ± 0.008 (+5.06%)
DMSO	0.862 ± 0.007 (+2.86%)	0.873 ± 0.008 (+4.12%)
Ethanol	0.861 ± 0.008 (+3.53%)	0.872 ± 0.005 (+4.85%)
Ethyl acetate	0.862 ± 0.008 (+3.06%)	0.871 ± 0.006 (+4.17%)
H2O	0.862 ± 0.005 (+4.06%)	0.874 ± 0.004 (+5.53%)
Octanol	0.860 ± 0.009 (+3.81%)	0.869 ± 0.010 (+4.82%)
Tetrahydrofuran	0.860 ± 0.010 (+2.95%)	0.869 ± 0.006 (+4.01%)
Toluene	0.860 ± 0.010 (+3.66%)	0.868 ± 0.010 (+4.65%)

Novartis

Comparison of our D-MPNN in both its unoptimized and optimized form against baseline models on a Novartis internal dataset using a time split of the data.

Table S20: Comparison to Baselines on Novartis, Part I. Note: All numbers are RMSE.

Dataset	D-MPNN	D-MPNN Ensemble	FFN on Morgan
LogP	0.692 ± 0.017	0.595 ± 0.004 (-14.02%)	0.915 ± 0.020 (+32.23%)

Table S21: Comparison to Baselines on Novartis, Part II. Note: All numbers are RMSE.

Dataset	FFN on Morgan Counts	FFN on RDKit	Mayr et al.[1]
LogP	0.838 ± 0.019 (+21.10%)	0.753 ± 0.013 (+8.82%)	1.583 ± 0.207 (+128.76%)

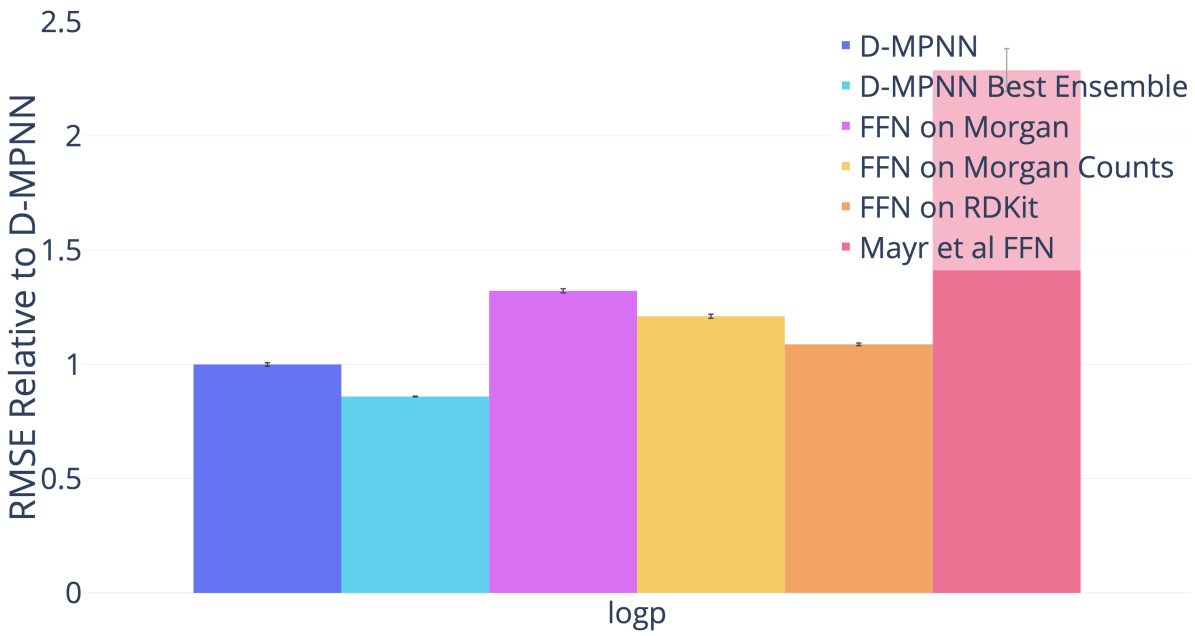
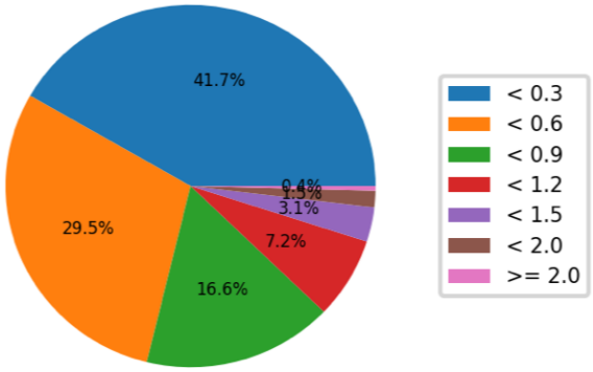


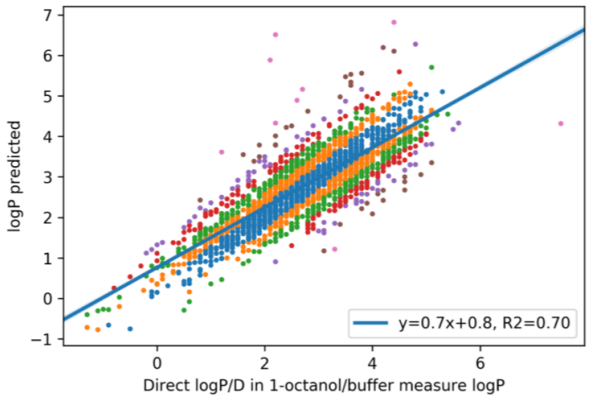
Figure S11: Comparison to Baselines on Novartis (lower = better).

Additional Novartis Results

We provide additional results for different versions of our D-MPNN model as well as for a Lasso regression model[3] on the Novartis dataset in Figures S12, S13, and S14. These results showcase the importance of proper normalization for the additional RDKit features. Our base D-MPNN predicts a logP within 0.3 of the ground truth on 47.4% of the test set, comparable to the best Lasso baseline. However, augmenting our model with features normalized using a Gaussian distribution assumption (simply subtracting the mean and dividing by the standard deviation for each feature) results in only 43.0% of test set predictions within 0.3 of the ground truth. But using properly normalized CDFs drastically improves this number to 51.2%. Note that the Lasso baseline runs on Morgan fingerprints as well as RDKit features using properly normalized CDFs.

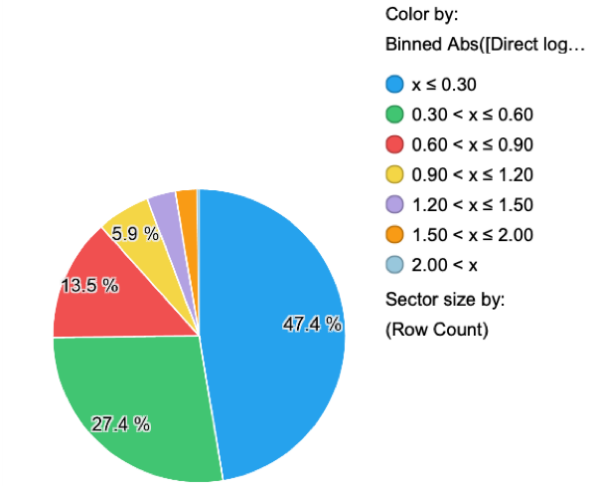


(a) Binned distribution of errors for Lasso baseline.

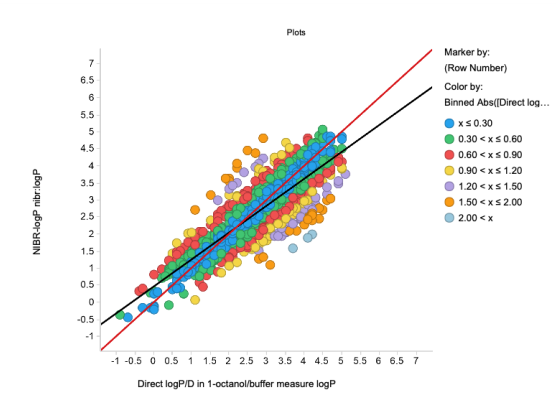


(b) Scatterplot of Lasso baseline predictions vs. ground truth.

Figure S12: Performance of Lasso models on the proprietary Novartis logP dataset. For each model, a pie chart shows the binned distribution of errors on the test set, and a scatterplot shows the predictions vs. ground truth for individual data points.

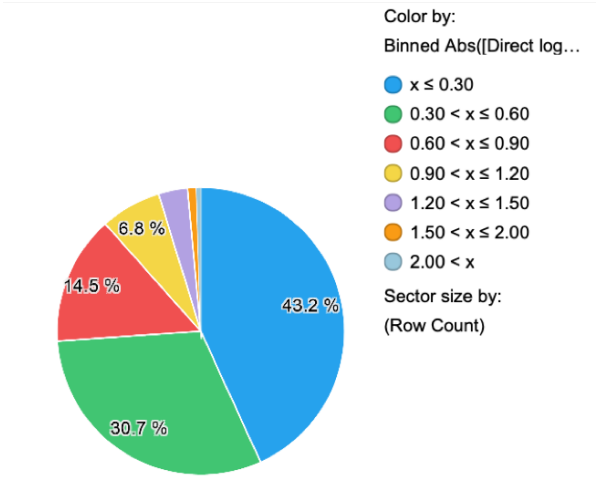


(a) Binned distribution of errors for D-MPNN.

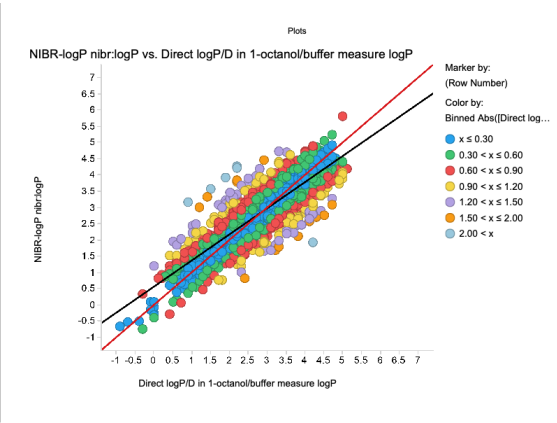


(b) Scatterplot of predictions vs. ground truth for D-MPNN.

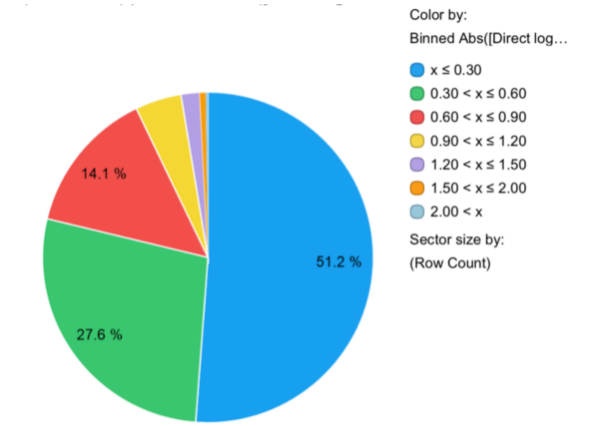
Figure S13: Performance of base D-MPNN model on the proprietary Novartis logP dataset. A pie chart shows the binned distribution of errors on the test set, and a scatterplot shows the predictions vs. ground truth for individual data points.



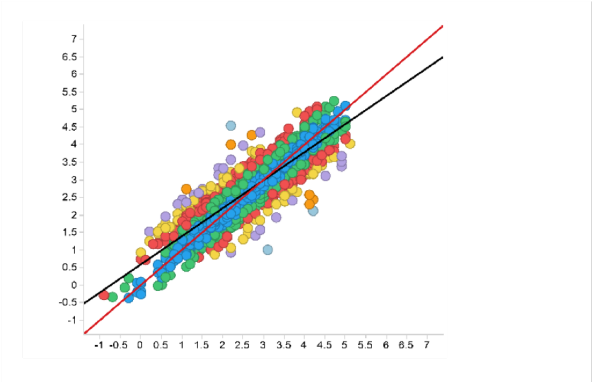
(a) Binned distribution of errors for D-MPNN with features normalized according to a Gaussian distribution assumption.



(b) Scatterplot of predictions vs. ground truth for D-MPNN with features normalized according to a Gaussian distribution assumption.



(c) Binned distribution of errors for D-MPNN with features normalized using CDFs.

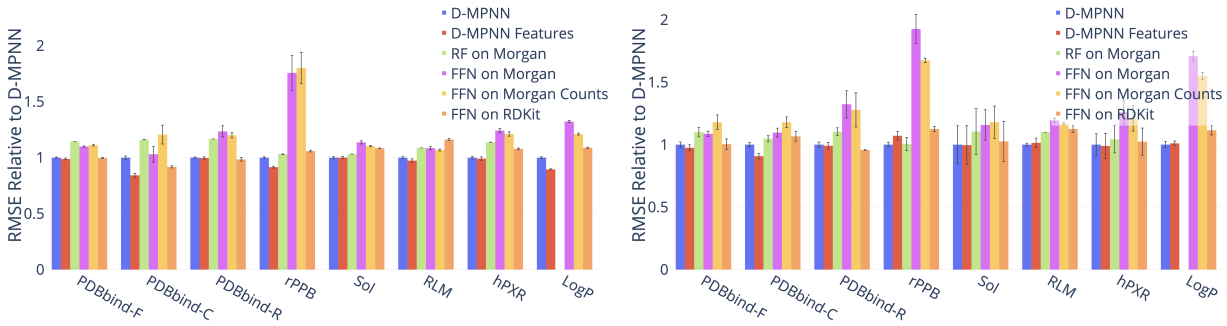


(d) Scatterplot of predictions vs. ground truth for D-MPNN with features normalized using CDFs.

Figure S14: Performance of D-MPNN models with features on proprietary Novartis logP dataset. For each model, a pie chart shows the binned distribution of errors on the test set, and a scatterplot shows the predictions vs. ground truth for individual data points.

Sliding Time Window Splits

We additionally evaluated our model on sliding time window splits where chronological data splits were available. For each dataset we divided it chronologically into 14 equally sized chunks. For each contiguous group of 5 chunks, we used the first 3 as training, the fourth as validation, and the fifth as test, for a total of 10 3:1:1 splits. Due to constraints of computational cost, we only evaluated on 3 of the splits for the Amgen datasets RLM, Sol, and hPXR. Overall, the time window split results are very noisy due to the smaller dataset size, so it is hard to make many strong conclusions, but overall the relative ranking of model architectures stays approximately stable compared to the full time splits.



(a) Time Split (lower = better).

(b) Sliding Time Window Split (higher = better).

Figure S15: Comparison to Baselines Using Time and Sliding Time Window Splits.

Table S22: Comparison to Baselines Using Time Split, Part I. *Only one run.

Dataset	D-MPNN	D-MPNN Features	RF on Morgan
PDBbind-F	2.187 ± 0.041	2.161 ± 0.055 (-1.20%)	2.500 ± 0.002 (+14.31%)
PDBbind-C	3.632 ± 0.170	3.058 ± 0.207 (-15.81%)	4.218 ± 0.014 (+16.12%)
PDBbind-R	2.424 ± 0.046	2.417 ± 0.084 (-0.28%)	2.830 ± 0.001 (+16.72%)
rPPB	1.057 ± 0.026	0.967 ± 0.028 (-8.50%)	1.089 ± 0.009 (+3.11%)
Sol	0.706 ± 0.013	0.706 ± 0.013 (+0.04%)	0.729 ± 0.000 (+3.22%)
RLM	0.331 ± 0.004	0.322 ± 0.005 (-2.64%)	0.360* (+8.78%)
hPXR	36.584 ± 0.751	36.252 ± 0.980 (-0.91%)	41.600 ± 0.070 (+13.71%)
LogP	0.692 ± 0.017	0.620 ± 0.011 (-10.48%)	—

Table S23: Comparison to Baselines Using Time Split, Part II.

Dataset	FFN on Morgan	FFN on Morgan Counts	FFN on RDKit
PDBbind-F	2.403 ± 0.044 (+9.88%)	2.431 ± 0.055 (+11.16%)	2.180 ± 0.033 (-0.34%)
PDBbind-C	3.743 ± 0.808 (+3.05%)	4.380 ± 0.964 (+20.57%)	3.334 ± 0.114 (-8.21%)
PDBbind-R	2.993 ± 0.380 (+23.45%)	2.908 ± 0.185 (+19.95%)	2.387 ± 0.126 (-1.56%)
rPPB	1.856 ± 0.517 (+75.69%)	1.903 ± 0.468 (+80.09%)	1.119 ± 0.027 (+5.87%)
Sol	0.802 ± 0.017 (+13.64%)	0.779 ± 0.008 (+10.31%)	0.765 ± 0.003 (+8.30%)
RLM	0.359 ± 0.005 (+8.48%)	0.353 ± 0.003 (+6.81%)	0.384 ± 0.003 (+16.14%)
hPXR	45.428 ± 1.255 (+24.17%)	44.305 ± 1.226 (+21.10%)	39.426 ± 0.465 (+7.77%)
LogP	0.915 ± 0.020 (+32.23%)	0.838 ± 0.019 (+21.10%)	0.753 ± 0.013 (+8.82%)

Table S24: Comparison to Baselines Using Sliding Time Window Split, Part I. *Only one run.

Dataset	D-MPNN	D-MPNN Features	RF on Morgan
PDBbind-F	1.259 ± 0.085	1.227 ± 0.096 (-2.50%)	1.385 ± 0.149 (+9.99%)
PDBbind-C	1.548 ± 0.081	1.405 ± 0.114 (-9.22%)	1.622 ± 0.125 (+4.77%)
PDBbind-R	1.347 ± 0.086	1.335 ± 0.118 (-0.89%)	1.486 ± 0.136 (+10.32%)
rPPB	1.310 ± 0.078	1.404 ± 0.140 (+7.12%)	1.318 ± 0.217 (+0.57%)
Sol	0.992 ± 0.260	0.988 ± 0.267 (-0.42%)	1.095 ± 0.316 (+10.44%)
RLM	0.395 ± 0.005	0.401 ± 0.015 (+1.44%)	0.434* (+9.97%)
hPXR	47.812 ± 7.330	47.316 ± 8.259 (-1.04%)	49.868 ± 9.046 (+4.30%)
LogP	0.726 ± 0.055	0.734 ± 0.042 (+1.05%)	—

Table S25: Comparison to Baselines Using Sliding Time Window Split, Part II.

Dataset	FFN on Morgan	FFN on Morgan Counts	FFN on RDKit
PDBbind-F	1.368 ± 0.083 (+8.67%)	1.485 ± 0.235 (+17.95%)	1.264 ± 0.169 (+0.41%)
PDBbind-C	1.697 ± 0.175 (+9.63%)	1.825 ± 0.211 (+17.88%)	1.651 ± 0.194 (+6.67%)
PDBbind-R	1.783 ± 0.460 (+32.33%)	1.721 ± 0.584 (+27.80%)	1.289 ± 0.010 (-4.28%)
rPPB	2.524 ± 0.483 (+92.64%)	2.194 ± 0.075 (+67.39%)	1.475 ± 0.084 (+12.57%)
Sol	1.148 ± 0.209 (+15.74%)	1.170 ± 0.223 (+17.96%)	1.018 ± 0.276 (+2.61%)
RLM	0.472 ± 0.007 (+19.44%)	0.461 ± 0.004 (+16.70%)	0.444 ± 0.011 (+12.48%)
hPXR	60.244 ± 8.015 (+26.00%)	57.859 ± 8.343 (+21.02%)	48.961 ± 9.012 (+2.40%)
LogP	1.240 ± 0.091 (+70.80%)	1.126 ± 0.064 (+55.10%)	0.810 ± 0.089 (+11.56%)

Experimental Error

Comparison of Amgen's internal model and our D-MPNN (evaluated on a chronological split) to experimental error. Note that the experimental error is not evaluated on the exact

same time split as the two models since it can only be measured on molecules which were tested more than once, but even so the difference in performance is striking.

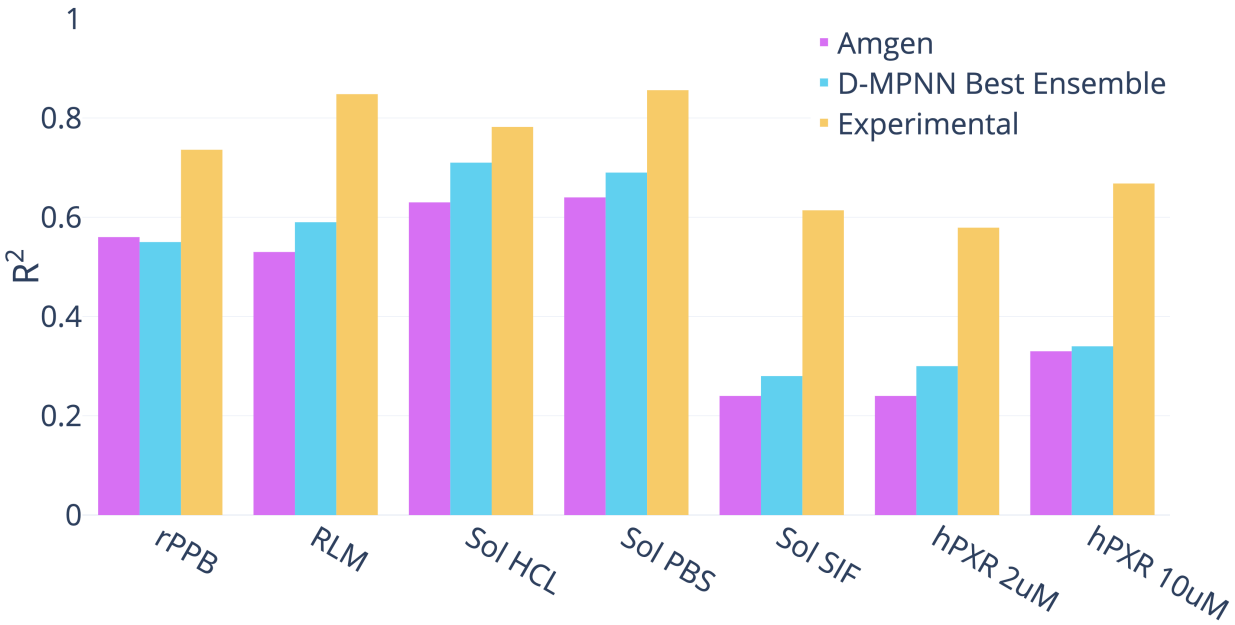


Figure S16: Experimental Error on Amgen (higher = better).

Table S26: Experimental Error on Amgen. Note: All numbers are R^2.

Dataset	Amgen	D-MPNN Optimized	Experimental
rPPB	0.56	0.55	0.736
RLM	0.53	0.59	0.848
Sol HCL	0.63	0.71	0.782
Sol PBS	0.64	0.69	0.856
Sol SIF	0.24	0.28	0.614
hPXR 2uM	0.24	0.3	0.579
hPXR 10uM	0.33	0.34	0.668

Analysis of Split Type

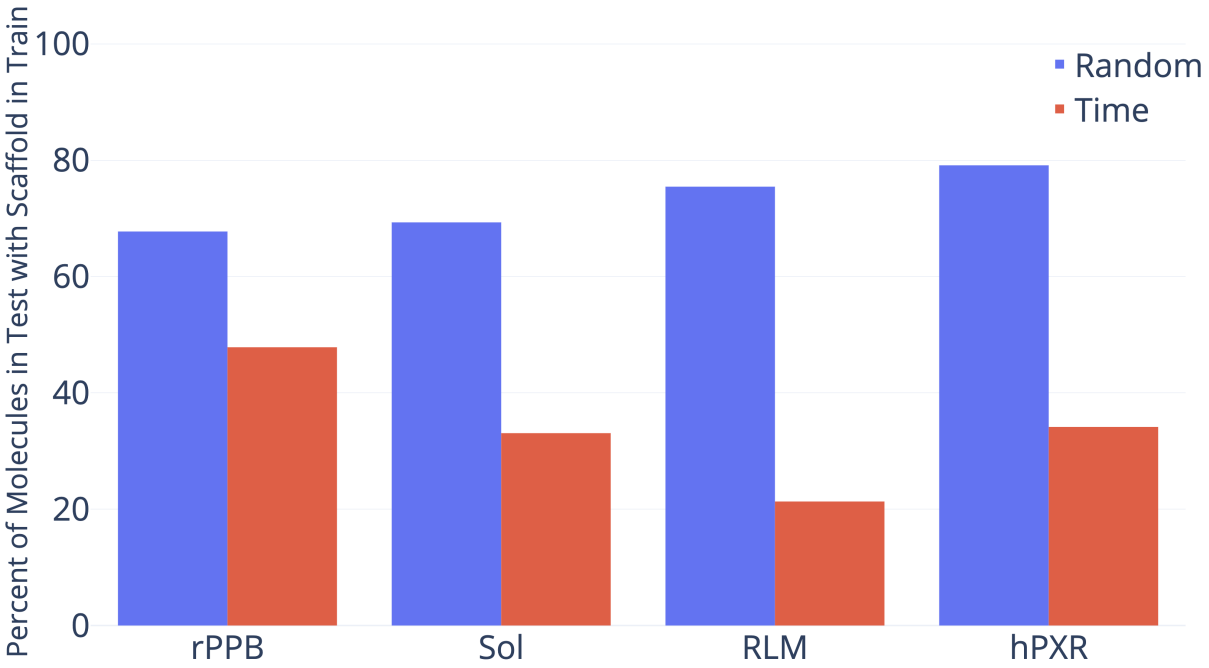


Figure S17: Overlap of molecular scaffolds between the train and test sets for a random or chronological split of four Amgen regression datasets. Overlap is defined as the percent of molecules in the test set which share a scaffold with a molecule in the train set.

Table S27: Overlap of molecular scaffolds between the train and test sets for a random or chronological split of four Amgen regression datasets. Overlap is defined as the percent of molecules in the test set which share a scaffold with a molecule in the train set.

Dataset	Random	Time
rPPB	67.74%	47.84%
Sol	69.31%	33.07%
RLM	75.45%	21.32%
hPXR	79.12%	34.14%

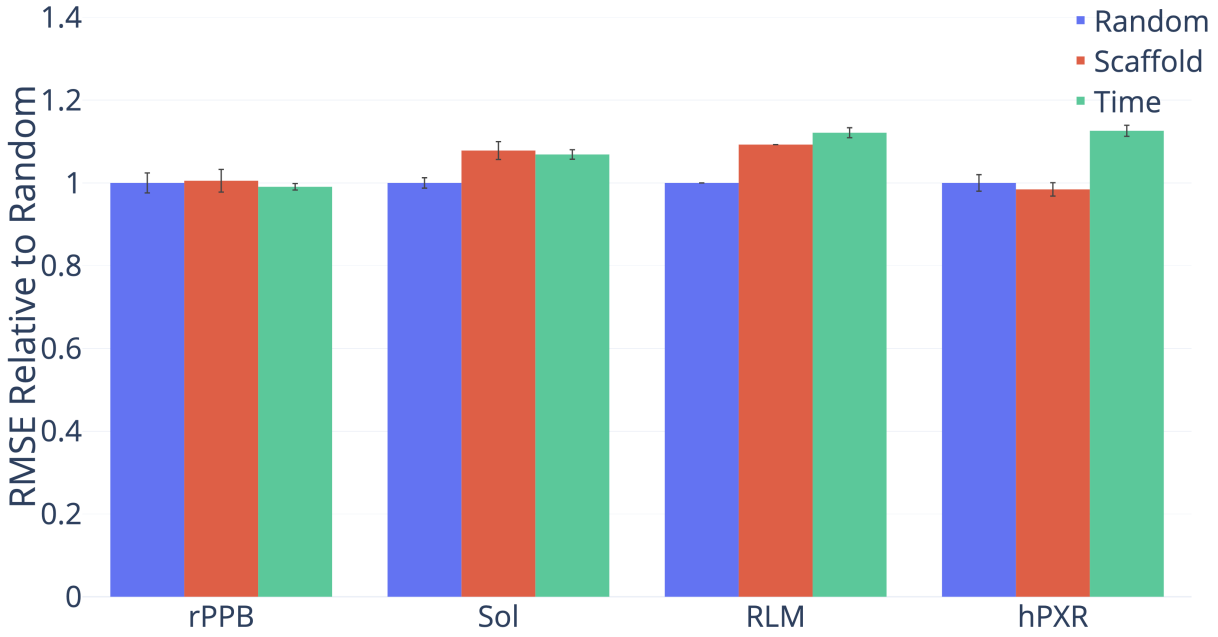


Figure S18: D-MPNN Performance by Split Type on Amgen datasets (lower = better).

Table S28: D-MPNN Performance by Split Type on Amgen datasets. Note: All numbers are RMSE. *Only one run.

Dataset	Random	Scaffold	Time
rPPB	1.067 ± 0.081	1.072 ± 0.093 (+0.53% p=0.45)	1.057 ± 0.026 (-0.94% p=0.37)
Sol	0.661 ± 0.014	0.712 ± 0.025 (+7.82% p=0.04)	0.706 ± 0.013 (+6.88% p=0.02)
RLM	0.295*	0.322* (+9.26%)	0.331 ± 0.004 (+12.13%)
hPXR	32.490 ± 1.124	31.984 ± 0.908 (-1.56% p=0.32)	36.584 ± 0.751 (+12.60% p=0.01)

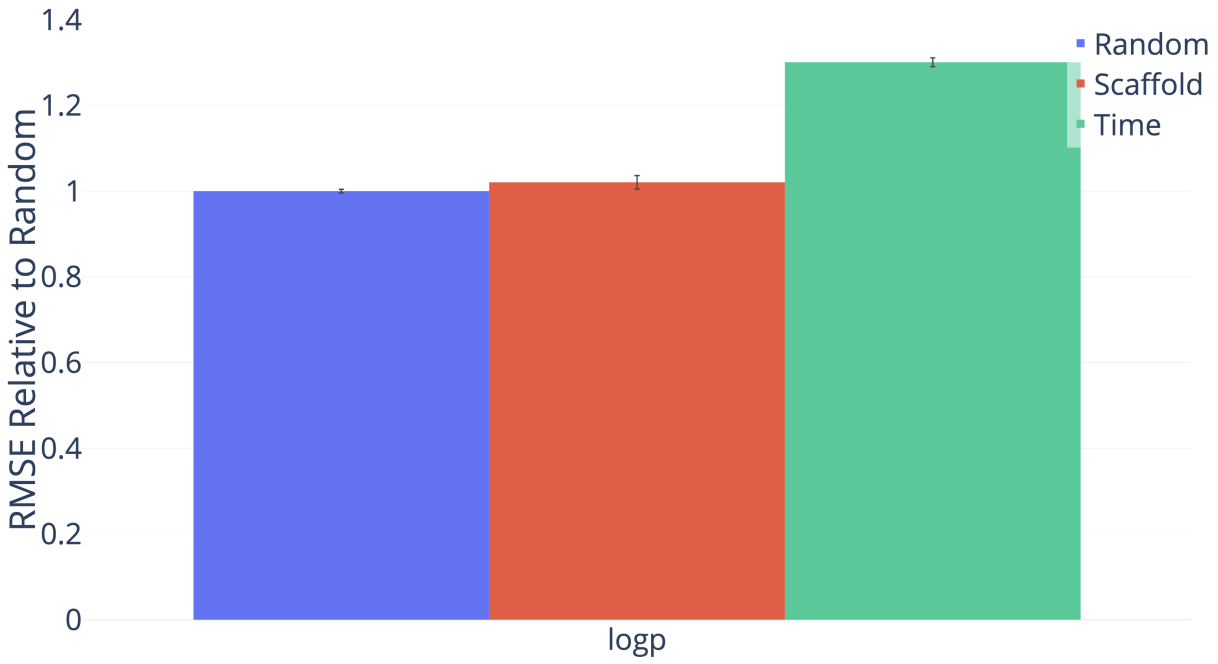


Figure S19: D-MPNN Performance by Split Type on Novartis datasets (lower = better).

Table S29: Performance of D-MPNN on different data splits on Novartis datasets. Note: All numbers are RMSE.

Dataset	**Random**	**Scaffold**	**Time**
LogP	0.532 ± 0.007	0.543 ± 0.027 (+2.07% p=0.13)	0.692 ± 0.017 (+30.08% p=0.00)

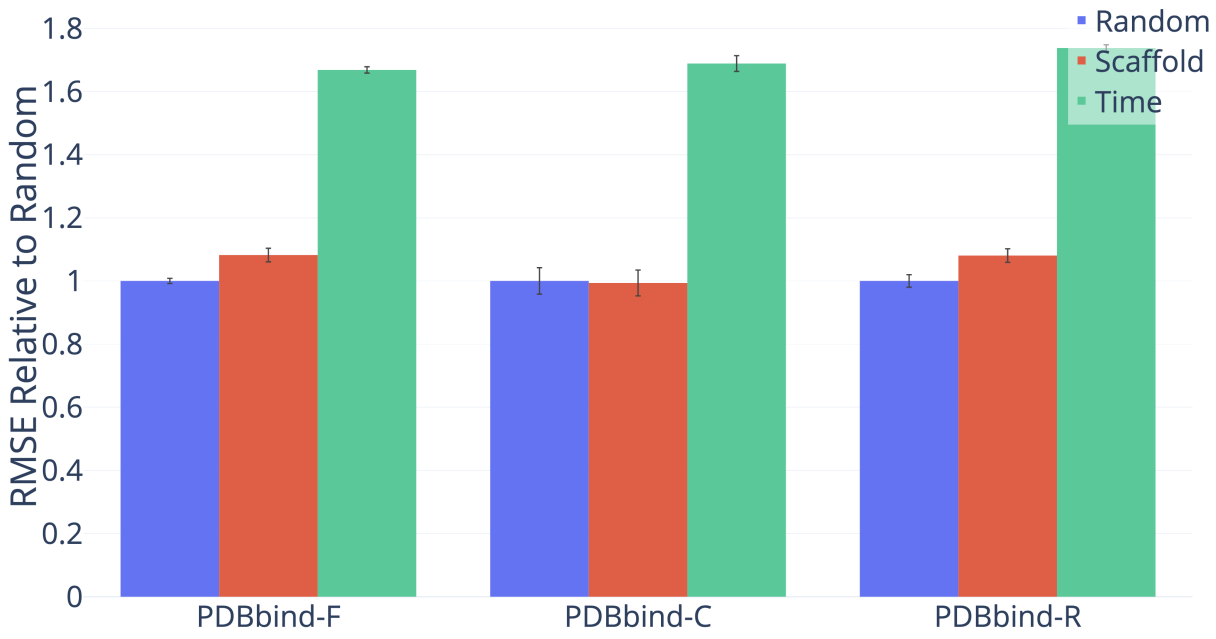
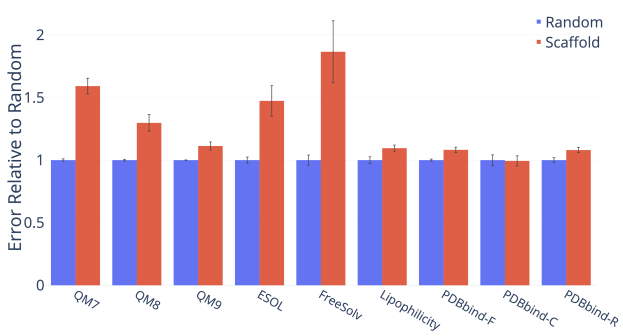


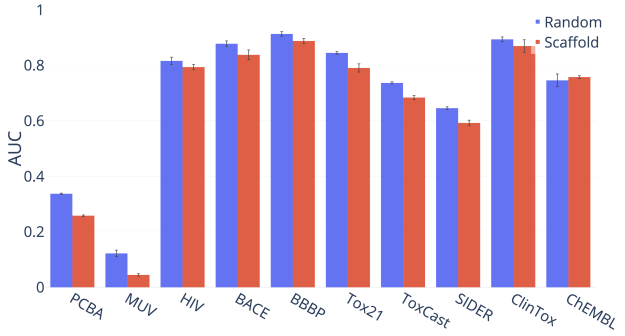
Figure S20: D-MPNN Performance by Split Type on PDBBind (lower = better).

Table S30: D-MPNN Performance by Split Type on PDBBind. Note: All numbers are RMSE.

Dataset	Random	Scaffold	Time
PDBbind-F	1.311 ± 0.034	1.419 ± 0.089 (+8.20% p=0.00)	2.187 ± 0.041 (+66.82% p=0.00)
PDBbind-C	2.151 ± 0.285	2.138 ± 0.278 (-0.64% p=0.46)	3.632 ± 0.170 (+68.84% p=0.00)
PDBbind-R	1.395 ± 0.087	1.507 ± 0.095 (+8.04% p=0.01)	2.424 ± 0.046 (+73.76% p=0.00)



(a) Regression Datasets (lower = better).



(b) Classification Datasets (higher = better).

Figure S21: D-MPNN Performance by Split Type on Public Datasets.

Table S31: D-MPNN Performance by Split Type on Public Datasets.

Dataset	Metric	Random	Scaffold
QM7	MAE	66.475 ± 2.088	105.775 ± 13.202 (+59.12% p=0.00)
QM8	MAE	0.011 ± 0.000	0.014 ± 0.002 (+29.75% p=0.00)
QM9	MAE	3.101 ± 0.010	3.451 ± 0.174 (+11.28% p=0.05)
ESOL	RMSE	0.665 ± 0.052	0.980 ± 0.258 (+47.31% p=0.00)
FreeSolv	RMSE	1.167 ± 0.150	2.177 ± 0.914 (+86.56% p=0.00)
Lipophilicity	RMSE	0.596 ± 0.050	0.653 ± 0.046 (+9.54% p=0.01)
PDBbind-F	RMSE	1.311 ± 0.034	1.419 ± 0.089 (+8.20% p=0.00)
PDBbind-C	RMSE	2.151 ± 0.285	2.138 ± 0.278 (-0.64% p=0.46)
PDBbind-R	RMSE	1.395 ± 0.087	1.507 ± 0.095 (+8.04% p=0.01)
PCBA	PRC-AUC	0.337 ± 0.004	0.258 ± 0.005 (-23.43% p=0.00)
MUV	PRC-AUC	0.122 ± 0.020	0.045 ± 0.007 (-63.44% p=0.01)
HIV	ROC-AUC	0.816 ± 0.023	0.794 ± 0.016 (-2.73% p=0.17)
BACE	ROC-AUC	0.878 ± 0.032	0.838 ± 0.056 (-4.55% p=0.04)
BBBP	ROC-AUC	0.913 ± 0.026	0.888 ± 0.029 (-2.78% p=0.04)
Tox21	ROC-AUC	0.845 ± 0.015	0.791 ± 0.047 (-6.42% p=0.00)
ToxCast	ROC-AUC	0.737 ± 0.013	0.684 ± 0.023 (-7.16% p=0.00)
SIDER	ROC-AUC	0.646 ± 0.016	0.593 ± 0.032 (-8.25% p=0.00)
ClinTox	ROC-AUC	0.894 ± 0.027	0.870 ± 0.072 (-2.70% p=0.18)
ChEMBL	ROC-AUC	0.746 ± 0.040	0.758 ± 0.008 (+1.60% p=0.36)

Ablations

Message Type

Here we describe the implementation and performance of our atom-based and undirected bond-based messages. For the most direct comparison, we implemented these as options in our model; the changes are only a few lines of code in each case. Therefore, in each case, we simply detail the differences from our directed bond-based messages.

Atom Messages

We initialize messages based on atom features rather than bond features, according to $h_v^0 = \tau(W_i x_v)$ rather than $h_{vw}^0 = \tau(W_i \, \texttt{cat}(x_v, e_{vw}))$, with matrix dimensions adjusted accordingly.

During message passing, each atom receives messages according to $m_v^{t+1} = \sum_{k \in \{N(v)\}} h_k^t$.

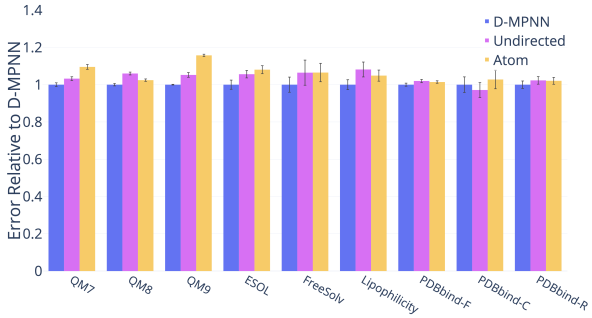
Finally, m_v is the sum of all of the atom hidden states at the end of message passing.

Undirected Bond Messages

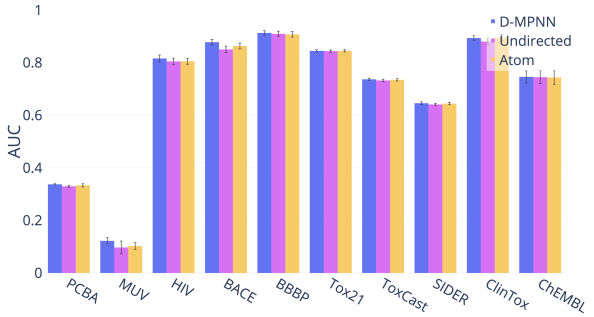
The only difference between undirected bonds and our D-MPNN is that before each message passing step, for each pair of bonded atoms v and w, we set h_{vw}^t and h_{wv}^t to each be equal to their average. Consequently, the hidden state for each directed bond is always equal to the hidden state of its reverse bond, resulting in message passing on undirected bonds.

Comparison of Different Message Types

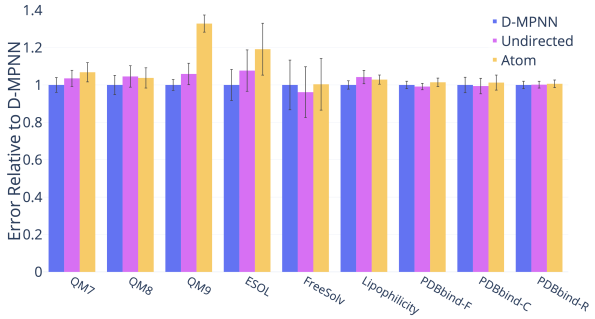
Comparison of performance using different message passing paradigms. Our D-MPNN uses directed messages.



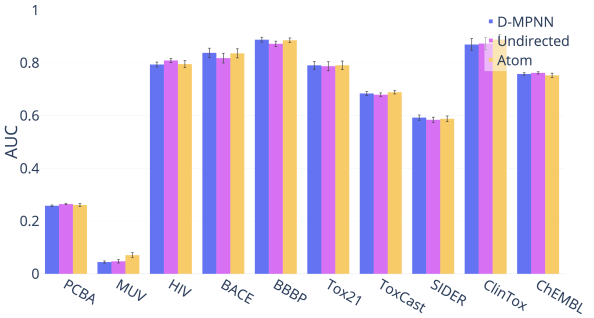
(a) Regression Datasets (Random Split, lower = better).



(b) Classification Datasets (Random Split, higher = better).



(c) Regression Datasets (Scaffold Split, lower = better).



(d) Classification Datasets (Scaffold Split, higher = better).

Figure S22: Message Type.

Table S32: Message Type (Random Split).

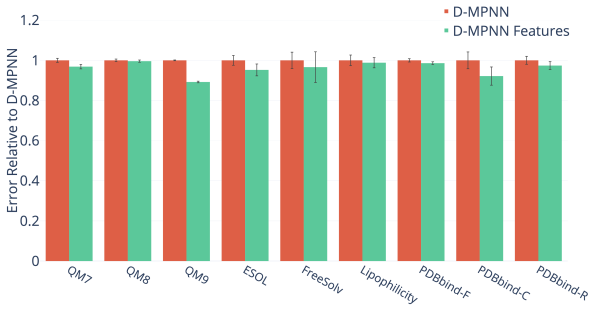
Dataset	Metric	D-MPNN	Undirected	Atom
QM7	MAE	66.475 ± 2.088	68.628 ± 2.177 (+3.24% p=0.01)	72.811 ± 2.737 (+9.53% p=0.00)
QM8	MAE	0.0110 ± 0.0002	0.012 ± 0.000 (+5.99% p=0.00)	0.011 ± 0.000 (+2.43% p=0.00)
QM9	MAE	3.101 ± 0.010	3.263 ± 0.069 (+5.23% p=0.00)	3.589 ± 0.033 (+15.73% p=0.00)
ESOL	RMSE	0.665 ± 0.052	0.702 ± 0.042 (+5.62% p=0.00)	0.719 ± 0.045 (+8.05% p=0.00)
FreeSolv	RMSE	1.167 ± 0.150	1.242 ± 0.249 (+6.44% p=0.00)	1.243 ± 0.182 (+6.50% p=0.06)
Lipophilicity	RMSE	0.596 ± 0.050	0.645 ± 0.075 (+8.15% p=0.00)	0.625 ± 0.056 (+4.87% p=0.00)
PDBbind-F	RMSE	1.311 ± 0.034	1.337 ± 0.036 (+1.98% p=0.00)	1.330 ± 0.027 (+1.42% p=0.00)
PDBbind-C	RMSE	2.151 ± 0.285	2.090 ± 0.270 (-2.86% p=0.98)	2.211 ± 0.339 (+2.79% p=0.17)
PDBbind-R	RMSE	1.395 ± 0.087	1.427 ± 0.090 (+2.30% p=0.00)	1.424 ± 0.082 (+2.07% p=0.00)
PCBA	PRC-AUC	0.337 ± 0.004	0.330 ± 0.007 (-2.23% p=0.01)	0.333 ± 0.010 (-1.15% p=0.03)
MUV	PRC-AUC	0.1222 ± 0.0204	0.097 ± 0.042 (-20.66% p=0.00)	0.103 ± 0.022 (-16.05% p=0.08)
HIV	ROC-AUC	0.816 ± 0.023	0.805 ± 0.022 (-1.40% p=0.92)	0.805 ± 0.019 (-1.33% p=0.80)
BACE	ROC-AUC	0.878 ± 0.032	0.850 ± 0.039 (-3.13% p=0.00)	0.864 ± 0.035 (-1.63% p=0.00)
BBBP	ROC-AUC	0.913 ± 0.026	0.910 ± 0.032 (-0.40% p=0.12)	0.908 ± 0.033 (-0.63% p=0.03)
Tox21	ROC-AUC	0.845 ± 0.015	0.844 ± 0.014 (-0.14% p=0.17)	0.845 ± 0.014 (+0.04% p=0.29)
ToxCast	ROC-AUC	0.737 ± 0.013	0.732 ± 0.015 (-0.61% p=0.00)	0.735 ± 0.014 (-0.27% p=0.25)
SIDER	ROC-AUC	0.646 ± 0.016	0.641 ± 0.014 (-0.73% p=0.34)	0.644 ± 0.014 (-0.23% p=0.50)
ClinTox	ROC-AUC	0.894 ± 0.027	0.881 ± 0.037 (-1.49% p=0.03)	0.896 ± 0.037 (+0.22% p=0.62)
ChEMBL	ROC-AUC	0.746 ± 0.040	0.745 ± 0.043 (-0.14% p=0.02)	0.744 ± 0.045 (-0.31% p=0.01)

Table S33: Message Type (Scaffold Split).

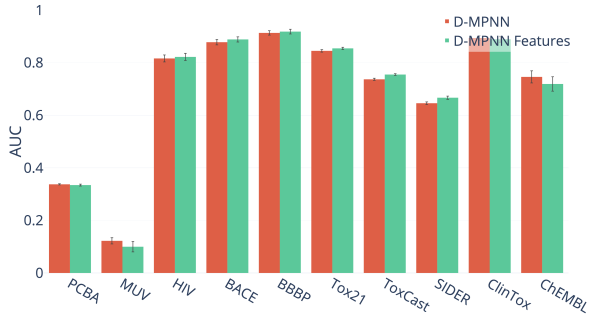
Dataset	Metric	D-MPNN	Undirected	Atom
QM7	MAE	105.775 ± 13.202	109.494 ± 14.420 (+3.52% p=0.89)	112.960 ± 17.211 (+6.79% p=0.00)
QM8	MAE	0.0143 ± 0.0023	0.015 ± 0.003 (+4.54% p=0.00)	0.015 ± 0.002 (+3.76% p=0.00)
QM9	MAE	3.451 ± 0.174	3.654 ± 0.343 (+5.89% p=0.00)	4.583 ± 0.274 (+32.82% p=0.00)
ESOL	RMSE	0.980 ± 0.258	1.055 ± 0.343 (+7.68% p=0.00)	1.167 ± 0.430 (+19.13% p=0.00)
FreeSolv	RMSE	2.177 ± 0.914	2.093 ± 0.936 (-3.86% p=0.11)	2.185 ± 0.952 (+0.37% p=0.00)
Lipophilicity	RMSE	0.653 ± 0.046	0.681 ± 0.074 (+4.22% p=0.00)	0.672 ± 0.051 (+2.86% p=0.00)
PDBbind-F	RMSE	1.419 ± 0.089	1.407 ± 0.078 (-0.85% p=0.57)	1.439 ± 0.100 (+1.43% p=0.00)
PDBbind-C	RMSE	2.138 ± 0.278	2.125 ± 0.280 (-0.59% p=0.33)	2.165 ± 0.272 (+1.26% p=0.09)
PDBbind-R	RMSE	1.507 ± 0.095	1.510 ± 0.086 (+0.15% p=0.03)	1.517 ± 0.097 (+0.62% p=0.02)
PCBA	PRC-AUC	0.258 ± 0.005	0.264 ± 0.004 (+2.36% p=0.89)	0.261 ± 0.010 (+1.12% p=0.86)
MUV	PRC-AUC	0.0447 ± 0.0074	0.047 ± 0.011 (+6.26% p=0.72)	0.071 ± 0.016 (+58.69% p=0.36)
HIV	ROC-AUC	0.794 ± 0.016	0.809 ± 0.014 (+1.94% p=1.00)	0.795 ± 0.023 (+0.20% p=0.84)
BACE	ROC-AUC	0.838 ± 0.056	0.818 ± 0.059 (-2.39% p=0.00)	0.836 ± 0.055 (-0.21% p=0.19)
BBBP	ROC-AUC	0.888 ± 0.029	0.872 ± 0.032 (-1.75% p=0.07)	0.886 ± 0.028 (-0.25% p=0.85)
Tox21	ROC-AUC	0.791 ± 0.047	0.787 ± 0.054 (-0.43% p=0.13)	0.791 ± 0.051 (+0.02% p=0.12)
ToxCast	ROC-AUC	0.684 ± 0.023	0.679 ± 0.022 (-0.69% p=0.02)	0.689 ± 0.021 (+0.68% p=0.95)
SIDER	ROC-AUC	0.593 ± 0.032	0.584 ± 0.031 (-1.40% p=0.03)	0.588 ± 0.034 (-0.76% p=0.10)
ClinTox	ROC-AUC	0.870 ± 0.072	0.873 ± 0.073 (+0.41% p=0.47)	0.888 ± 0.064 (+2.14% p=0.98)
ChEMBL	ROC-AUC	0.758 ± 0.008	0.762 ± 0.007 (+0.57% p=0.41)	0.753 ± 0.014 (-0.67% p=0.04)

RDKit Features

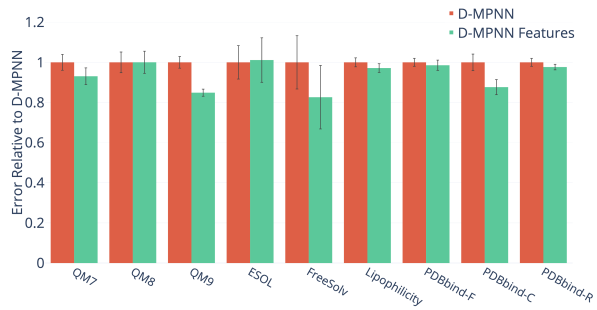
Effect of adding RDKit features to our optimized D-MPNN.



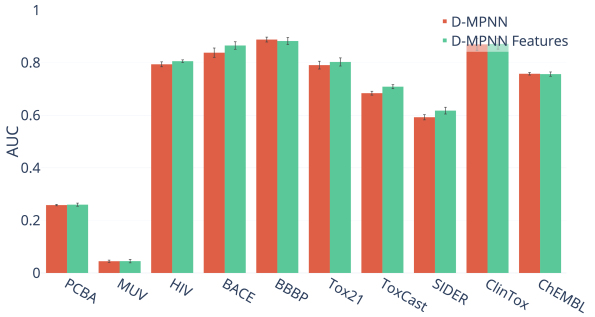
(a) Regression Datasets (Random Split, lower = better).



(b) Classification Datasets (Random Split, lower = better).



(c) Regression Datasets (Scaffold Split, lower = better).



(d) Classification Datasets (Scaffold Split, higher = better).

Figure S23: RDKit Features.

Table S34: RDKit Features (Random Split).

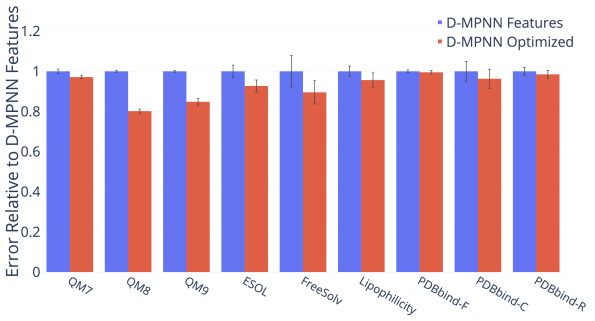
Dataset	Metric	D-MPNN	D-MPNN Features
QM7	MAE	66.475 ± 2.088	64.390 ± 2.361 (-3.14% p=0.14)
QM8	MAE	0.011 ± 0.000	0.011 ± 0.000 (-0.42% p=0.00)
QM9	MAE	3.101 ± 0.010	2.766 ± 0.022 (-10.79% p=0.00)
ESOL	RMSE	0.665 ± 0.052	0.633 ± 0.062 (-4.77% p=0.00)
FreeSolv	RMSE	1.167 ± 0.150	1.127 ± 0.282 (-3.39% p=0.00)
Lipophilicity	RMSE	0.596 ± 0.050	0.589 ± 0.048 (-1.18% p=0.00)
PDBbind-F	RMSE	1.311 ± 0.034	1.293 ± 0.028 (-1.41% p=0.00)
PDBbind-C	RMSE	2.151 ± 0.285	1.983 ± 0.309 (-7.84% p=0.04)
PDBbind-R	RMSE	1.395 ± 0.087	1.359 ± 0.086 (-2.61% p=0.03)
PCBA	PRC-AUC	0.337 ± 0.004	0.334 ± 0.006 (-0.86% p=0.91)
MUV	PRC-AUC	0.122 ± 0.020	0.100 ± 0.034 (-18.34% p=0.85)
HIV	ROC-AUC	0.816 ± 0.023	0.822 ± 0.024 (+0.72% p=0.20)
BACE	ROC-AUC	0.878 ± 0.032	0.888 ± 0.031 (+1.20% p=0.02)
BBBP	ROC-AUC	0.913 ± 0.026	0.918 ± 0.028 (+0.54% p=0.17)
Tox21	ROC-AUC	0.845 ± 0.015	0.854 ± 0.013 (+1.12% p=0.00)
ToxCast	ROC-AUC	0.737 ± 0.013	0.755 ± 0.010 (+2.46% p=0.00)
SIDER	ROC-AUC	0.646 ± 0.016	0.667 ± 0.019 (+3.25% p=0.00)
ClinTox	ROC-AUC	0.894 ± 0.027	0.889 ± 0.036 (-0.51% p=0.57)
ChEMBL	ROC-AUC	0.746 ± 0.040	0.719 ± 0.047 (-3.61% p=1.00)

Table S35: RDKit Features (Scaffold Split).

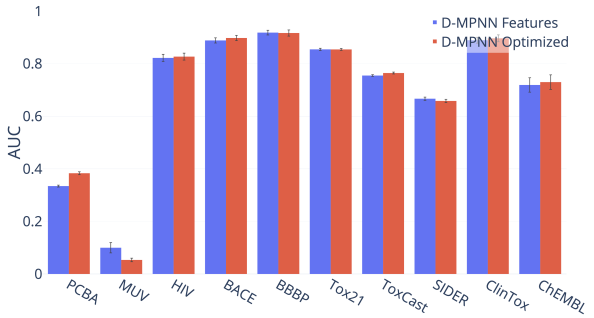
Dataset	Metric	D-MPNN	D-MPNN Features
QM7	MAE	105.775 ± 13.202	98.442 ± 13.936 (-6.93% p=0.00)
QM8	MAE	0.014 ± 0.002	0.014 ± 0.003 (+0.02% p=0.01)
QM9	MAE	3.451 ± 0.174	2.929 ± 0.106 (-15.12% p=0.00)
ESOL	RMSE	0.980 ± 0.258	0.991 ± 0.343 (+1.14% p=0.98)
FreeSolv	RMSE	2.177 ± 0.914	1.799 ± 1.088 (-17.37% p=0.00)
Lipophilicity	RMSE	0.653 ± 0.046	0.634 ± 0.045 (-2.85% p=0.00)
PDBbind-F	RMSE	1.419 ± 0.089	1.398 ± 0.115 (-1.45% p=0.01)
PDBbind-C	RMSE	2.138 ± 0.278	1.874 ± 0.253 (-12.35% p=0.00)
PDBbind-R	RMSE	1.507 ± 0.095	1.472 ± 0.066 (-2.35% p=0.01)
PCBA	PRC-AUC	0.258 ± 0.005	0.260 ± 0.010 (+0.62% p=0.41)
MUV	PRC-AUC	0.045 ± 0.007	0.045 ± 0.011 (+0.73% p=0.42)
HIV	ROC-AUC	0.794 ± 0.016	0.806 ± 0.009 (+1.49% p=0.07)
BACE	ROC-AUC	0.838 ± 0.056	0.865 ± 0.045 (+3.26% p=0.01)
BBBP	ROC-AUC	0.888 ± 0.029	0.882 ± 0.043 (-0.62% p=0.35)
Tox21	ROC-AUC	0.791 ± 0.047	0.803 ± 0.049 (+1.52% p=0.00)
ToxCast	ROC-AUC	0.684 ± 0.023	0.709 ± 0.024 (+3.61% p=0.00)
SIDER	ROC-AUC	0.593 ± 0.032	0.618 ± 0.041 (+4.24% p=0.00)
ClinTox	ROC-AUC	0.870 ± 0.072	0.872 ± 0.063 (+0.21% p=0.65)
ChEMBL	ROC-AUC	0.758 ± 0.008	0.757 ± 0.014 (-0.17% p=0.82)

Hyperparameter Optimization

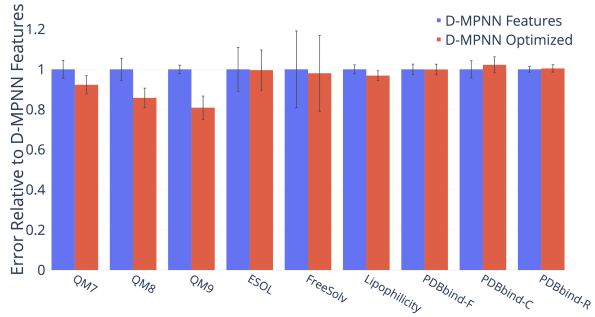
Effect of performing Bayesian hyperparameter optimization on the depth, hidden size, number of fully connected layers, and dropout of our model. Optimization was done on random splits and then the optimized model was applied to both random and scaffold splits.



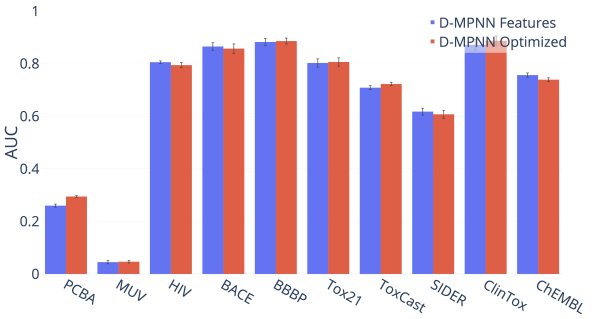
(a) Regression Datasets (Random Split, lower = better).



(b) Classification Datasets (Random Split, higher = better).



(c) Regression Datasets (Scaffold Split, lower = better).



(d) Classification Datasets (Scaffold Split, higher = better).

Figure S24: Hyperparameter Optimization.

Table S36: Hyperparameter Optimization (Random Split).

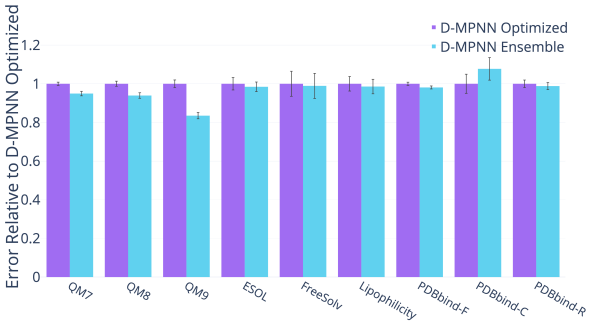
Dataset	Metric	D-MPNN Features	D-MPNN Optimized
QM7	MAE	64.390 ± 2.361 (-3.14%)	62.542 ± 1.649 (-2.87% p=0.00)
QM8	MAE	0.011 ± 0.000 (-0.42%)	0.009 ± 0.000 (-19.82% p=0.00)
QM9	MAE	2.766 ± 0.022 (-10.79%)	2.346 ± 0.080 (-15.19% p=0.00)
ESOL	RMSE	0.633 ± 0.062 (-4.77%)	0.587 ± 0.060 (-7.28% p=0.00)
FreeSolv	RMSE	1.127 ± 0.282 (-3.39%)	1.009 ± 0.207 (-10.46% p=0.00)
Lipophilicity	RMSE	0.589 ± 0.048 (-1.18%)	0.563 ± 0.067 (-4.35% p=0.00)
PDBbind-F	RMSE	1.293 ± 0.028 (-1.41%)	1.286 ± 0.033 (-0.48% p=0.06)
PDBbind-C	RMSE	1.983 ± 0.309 (-7.84%)	1.910 ± 0.299 (-3.69% p=0.02)
PDBbind-R	RMSE	1.359 ± 0.086 (-2.61%)	1.338 ± 0.082 (-1.51% p=0.00)
PCBA	PRC-AUC	0.334 ± 0.006 (-0.86%)	0.383 ± 0.009 (+14.62% p=0.00)
MUV	PRC-AUC	0.100 ± 0.034 (-18.34%)	0.053 ± 0.012 (-46.50% p=0.89)
HIV	ROC-AUC	0.822 ± 0.024 (+0.72%)	0.827 ± 0.023 (+0.59% p=0.55)
BACE	ROC-AUC	0.888 ± 0.031 (+1.20%)	0.898 ± 0.031 (+1.05% p=0.04)
BBBP	ROC-AUC	0.918 ± 0.028 (+0.54%)	0.917 ± 0.037 (-0.17% p=0.55)
Tox21	ROC-AUC	0.854 ± 0.013 (+1.12%)	0.854 ± 0.012 (-0.03% p=0.42)
ToxCast	ROC-AUC	0.755 ± 0.010 (+2.46%)	0.764 ± 0.011 (+1.27% p=0.00)
SIDER	ROC-AUC	0.667 ± 0.019 (+3.25%)	0.658 ± 0.020 (-1.28% p=0.88)
ClinTox	ROC-AUC	0.889 ± 0.036 (-0.51%)	0.897 ± 0.042 (+0.81% p=0.37)
ChEMBL	ROC-AUC	0.719 ± 0.047 (-3.61%)	0.730 ± 0.048 (+1.52% p=0.01)

Table S37: Hyperparameter Optimization (Scaffold Split).

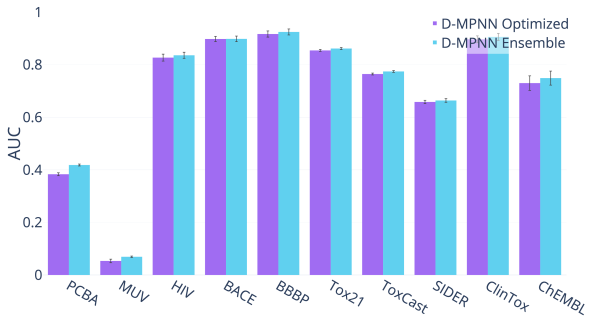
Dataset	Metric	D-MPNN Features	D-MPNN Optimized
QM7	MAE	98.442 ± 13.936 (-6.93%)	90.869 ± 14.199 (-7.69% p=0.00)
QM8	MAE	0.014 ± 0.003 (+0.02%)	0.012 ± 0.002 (-14.21% p=0.00)
QM9	MAE	2.929 ± 0.106 (-15.12%)	2.370 ± 0.294 (-19.09% p=0.00)
ESOL	RMSE	0.991 ± 0.343 (+1.14%)	0.987 ± 0.314 (-0.39% p=0.00)
FreeSolv	RMSE	1.799 ± 1.088 (-17.37%)	1.763 ± 1.075 (-1.95% p=0.00)
Lipophilicity	RMSE	0.634 ± 0.045 (-2.85%)	0.615 ± 0.048 (-3.12% p=0.00)
PDBbind-F	RMSE	1.398 ± 0.115 (-1.45%)	1.397 ± 0.117 (-0.08% p=0.01)
PDBbind-C	RMSE	1.874 ± 0.253 (-12.35%)	1.916 ± 0.236 (+2.25% p=0.90)
PDBbind-R	RMSE	1.472 ± 0.066 (-2.35%)	1.479 ± 0.087 (+0.45% p=0.72)
PCBA	PRC-AUC	0.260 ± 0.010 (+0.62%)	0.295 ± 0.006 (+13.39% p=0.00)
MUV	PRC-AUC	0.045 ± 0.011 (+0.73%)	0.047 ± 0.009 (+3.61% p=0.25)
HIV	ROC-AUC	0.806 ± 0.009 (+1.49%)	0.794 ± 0.017 (-1.40% p=0.87)
BACE	ROC-AUC	0.865 ± 0.045 (+3.26%)	0.857 ± 0.057 (-0.93% p=0.77)
BBBP	ROC-AUC	0.882 ± 0.043 (-0.62%)	0.886 ± 0.036 (+0.41% p=0.28)
Tox21	ROC-AUC	0.803 ± 0.049 (+1.52%)	0.806 ± 0.050 (+0.45% p=0.14)
ToxCast	ROC-AUC	0.709 ± 0.024 (+3.61%)	0.723 ± 0.020 (+1.96% p=0.00)
SIDER	ROC-AUC	0.618 ± 0.041 (+4.24%)	0.607 ± 0.047 (-1.68% p=1.00)
ClinTox	ROC-AUC	0.872 ± 0.063 (+0.21%)	0.887 ± 0.058 (+1.78% p=0.01)
ChEMBL	ROC-AUC	0.757 ± 0.014 (-0.17%)	0.739 ± 0.012 (-2.30% p=0.58)

Ensembling

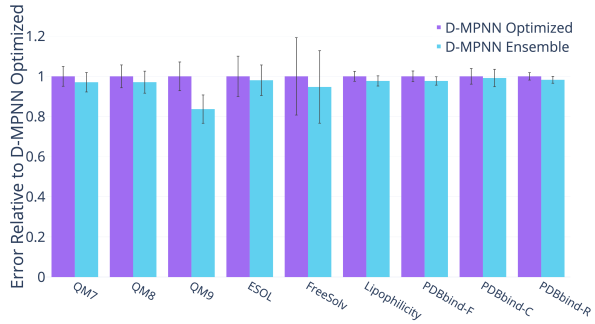
Benefit of ensembling five models instead of a single model. All results are using our best model settings (i.e. optimized hyperparameters and RDKit features, if they improved performance in the single model setting).



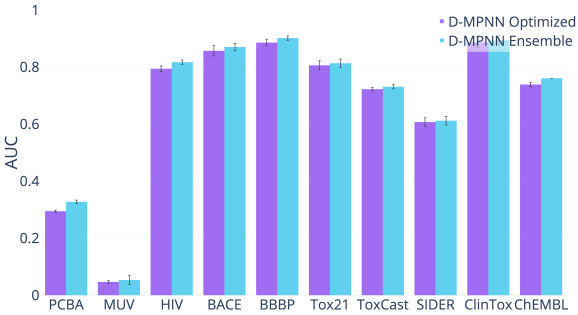
(a) Regression Datasets (Random Split, lower = better).



(b) Classification Datasets (Random Split, higher = better).



(c) Regression Datasets (Scaffold Split, lower = better).



(d) Classification Datasets (Scaffold Split, higher = better).

Figure S25: Ensembling.

Table S38: Ensembling (Random Split).

Dataset	Metric	D-MPNN Optimized	D-MPNN Ensemble
QM7	MAE	62.542 ± 1.649 (-5.92%)	59.379 ± 2.315 (-5.06% p=0.00)
QM8	MAE	0.009 ± 0.000 (-20.16%)	0.008 ± 0.000 (-6.08% p=0.00)
QM9	MAE	2.346 ± 0.080 (-24.34%)	1.959 ± 0.066 (-16.50% p=0.00)
ESOL	RMSE	0.587 ± 0.060 (-11.70%)	0.578 ± 0.046 (-1.56% p=0.02)
FreeSolv	RMSE	1.009 ± 0.207 (-13.49%)	0.998 ± 0.207 (-1.08% p=0.22)
Lipophilicity	RMSE	0.563 ± 0.067 (-5.48%)	0.555 ± 0.067 (-1.43% p=0.00)
PDBbind-F	RMSE	1.286 ± 0.033 (-1.88%)	1.262 ± 0.031 (-1.90% p=0.00)
PDBbind-C	RMSE	1.910 ± 0.299 (-11.24%)	2.057 ± 0.353 (+7.72% p=1.00)
PDBbind-R	RMSE	1.338 ± 0.082 (-4.08%)	1.322 ± 0.077 (-1.19% p=0.01)
PCBA	PRC-AUC	0.383 ± 0.009 (+13.63%)	0.418 ± 0.006 (+9.16% p=0.00)
MUV	PRC-AUC	0.053 ± 0.012 (-56.31%)	0.069 ± 0.005 (+29.60% p=0.92)
HIV	ROC-AUC	0.827 ± 0.023 (+1.31%)	0.836 ± 0.020 (+1.08% p=0.00)
BACE	ROC-AUC	0.898 ± 0.031 (+2.26%)	0.898 ± 0.034 (+0.05% p=0.43)
BBBP	ROC-AUC	0.917 ± 0.037 (+0.37%)	0.925 ± 0.036 (+0.85% p=0.00)
Tox21	ROC-AUC	0.854 ± 0.012 (+1.09%)	0.861 ± 0.012 (+0.85% p=0.00)
ToxCast	ROC-AUC	0.764 ± 0.011 (+3.77%)	0.774 ± 0.011 (+1.28% p=0.00)
SIDER	ROC-AUC	0.658 ± 0.020 (+1.93%)	0.664 ± 0.021 (+0.84% p=0.01)
ClinTox	ROC-AUC	0.897 ± 0.042 (+0.30%)	0.906 ± 0.043 (+1.03% p=0.01)
ChEMBL	ROC-AUC	0.730 ± 0.048 (-2.15%)	0.749 ± 0.046 (+2.61% p=0.00)

Table S39: Ensembling (Scaffold Split).

Dataset	Metric	D-MPNN Optimized	D-MPNN Ensemble
QM7	MAE	90.869 ± 14.199 (-14.09%)	88.201 ± 13.899 (-2.94% p=0.06)
QM8	MAE	0.012 ± 0.002 (-14.20%)	0.012 ± 0.002 (-2.90% p=0.00)
QM9	MAE	2.370 ± 0.294 (-31.32%)	1.983 ± 0.289 (-16.32% p=0.00)
ESOL	RMSE	0.987 ± 0.314 (+0.75%)	0.968 ± 0.237 (-1.94% p=0.00)
FreeSolv	RMSE	1.763 ± 1.075 (-18.99%)	1.670 ± 1.008 (-5.29% p=0.05)
Lipophilicity	RMSE	0.615 ± 0.048 (-5.88%)	0.600 ± 0.049 (-2.29% p=0.00)
PDBbind-F	RMSE	1.397 ± 0.117 (-1.53%)	1.365 ± 0.092 (-2.30% p=0.00)
PDBbind-C	RMSE	1.916 ± 0.236 (-10.38%)	1.900 ± 0.262 (-0.83% p=0.52)
PDBbind-R	RMSE	1.479 ± 0.087 (-1.91%)	1.453 ± 0.080 (-1.73% p=0.00)
PCBA	PRC-AUC	0.295 ± 0.006 (+14.10%)	0.328 ± 0.011 (+11.20% p=0.00)
MUV	PRC-AUC	0.047 ± 0.009 (+4.37%)	0.053 ± 0.027 (+14.58% p=0.39)
HIV	ROC-AUC	0.794 ± 0.017 (+0.07%)	0.817 ± 0.013 (+2.87% p=0.00)
BACE	ROC-AUC	0.857 ± 0.057 (+2.30%)	0.871 ± 0.041 (+1.55% p=0.01)
BBBP	ROC-AUC	0.886 ± 0.036 (-0.21%)	0.902 ± 0.024 (+1.78% p=0.10)
Tox21	ROC-AUC	0.806 ± 0.050 (+1.98%)	0.814 ± 0.047 (+0.90% p=0.00)
ToxCast	ROC-AUC	0.723 ± 0.020 (+5.65%)	0.731 ± 0.023 (+1.18% p=0.03)
SIDER	ROC-AUC	0.607 ± 0.047 (+2.49%)	0.612 ± 0.047 (+0.80% p=0.17)
ClinTox	ROC-AUC	0.887 ± 0.058 (+1.99%)	0.895 ± 0.050 (+0.85% p=0.17)
ChEMBL	ROC-AUC	0.739 ± 0.012 (-2.47%)	0.761 ± 0.000 (+2.94% p=0.00)

Effect of Data Size

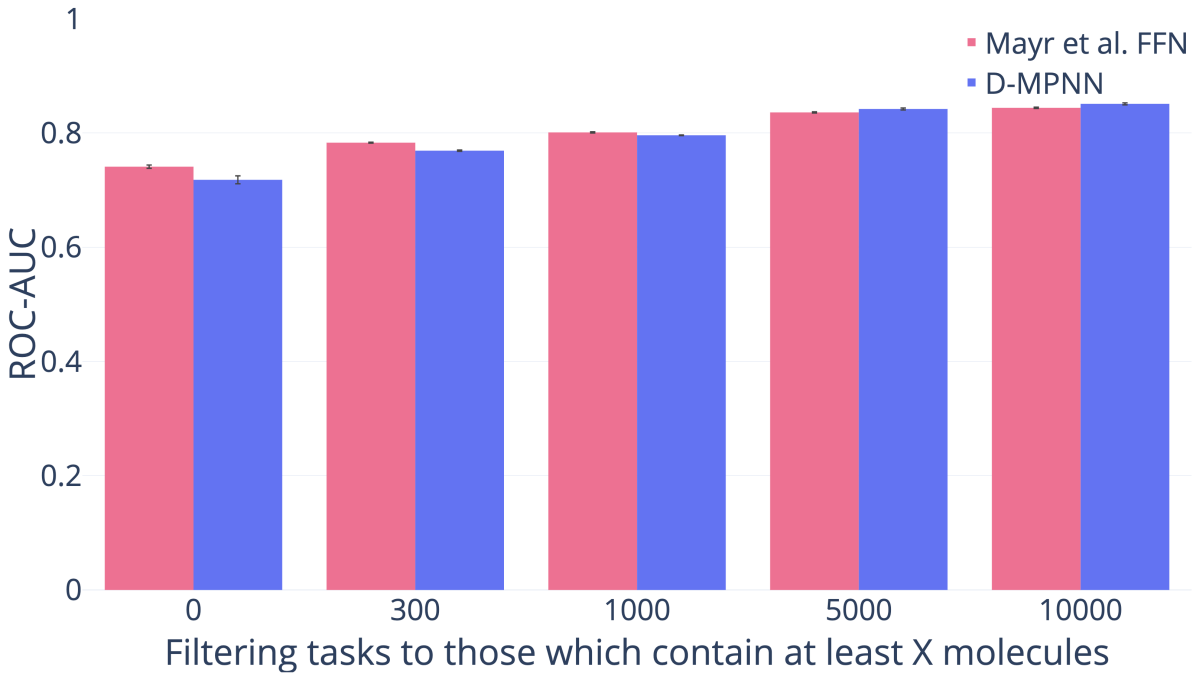


Figure S26: Effect of Data Size on ChEMBL (higher = better).

Table S40: Effect of Data Size on ChEMBL. All numbers are ROC-AUC.

Min # of Compounds	Mayr et al.[1] FFN	D-MPNN
0	0.741 ± 0.005	0.718 ± 0.012 (-3.10% p=0.05)
300	0.783 ± 0.002	0.769 ± 0.002 (-1.79% p=0.00)
1,000	0.801 ± 0.002	0.796 ± 0.001 (-0.62% p=0.03)
5,000	0.836 ± 0.002	0.842 ± 0.003 (+0.72% p=0.04)
10,000	0.844 ± 0.002	0.851 ± 0.003 (+0.83% p=0.03)

RDKit-Calculated Features

We used the following list of 200 RDKit functions to calculate the RDKit features used by our model.

BalabanJ	BertzCT	Chi0
Chi0n	Chi0v	Chi1

Chi1n	Chi1v	Chi2n
Chi2v	Chi3n	Chi3v
Chi4n	Chi4v	EState_VSA1
EState_VSA10	EState_VSA11	EState_VSA2
EState_VSA3	EState_VSA4	EState_VSA5
EState_VSA6	EState_VSA7	EState_VSA8
EState_VSA9	ExactMolWt	FpDensityMorgan1
FpDensityMorgan2	FpDensityMorgan3	FractionCSP3
HallKierAlpha	HeavyAtomCount	HeavyAtomMolWt
Ipc	Kappa1	Kappa2
Kappa3	LabuteASA	MaxAbsEStateIndex
MaxAbsPartialCharge	MaxEStateIndex	MaxPartialCharge
MinAbsEStateIndex	MinAbsPartialCharge	MinEStateIndex
MinPartialCharge	MolLogP	MolMR
MolWt	NHOHCount	NOCount
NumAliphaticCarbocycles	NumAliphaticHeterocycles	NumAliphaticRings
NumAromaticCarbocycles	NumAromaticHeterocycles	NumAromaticRings
NumHAcceptors	NumHDonors	NumHeteroatoms
NumRadicalElectrons	NumRotatableBonds	NumSaturatedCarbocycles
NumSaturatedHeterocycles	NumSaturatedRings	NumValenceElectrons
PEOE_VSA1	PEOE_VSA10	PEOE_VSA11
PEOE_VSA12	PEOE_VSA13	PEOE_VSA14
PEOE_VSA2	PEOE_VSA3	PEOE_VSA4
PEOE_VSA5	PEOE_VSA6	PEOE_VSA7
PEOE_VSA8	PEOE_VSA9	RingCount
SMR_VSA1	SMR_VSA10	SMR_VSA2
SMR_VSA3	SMR_VSA4	SMR_VSA5

SMR_VSA6	SMR_VSA7	SMR_VSA8
SMR_VSA9	SlogP_VSA1	SlogP_VSA10
SlogP_VSA11	SlogP_VSA12	SlogP_VSA2
SlogP_VSA3	SlogP_VSA4	SlogP_VSA5
SlogP_VSA6	SlogP_VSA7	SlogP_VSA8
SlogP_VSA9	TPSA	VSA_EState1
VSA_EState10	VSA_EState2	VSA_EState3
VSA_EState4	VSA_EState5	VSA_EState6
VSA_EState7	VSA_EState8	VSA_EState9
fr_Al_COO	fr_Al_OH	fr_Al_OH_noTert
fr_ArN	fr_Ar_COO	fr_Ar_N
fr_Ar_NH	fr_Ar_OH	fr_COO
fr_COO2	fr_C_O	fr_C_O_noCOO
fr_C_S	fr_HOCCN	fr_Imine
fr_NH0	fr_NH1	fr_NH2
fr_N_O	fr_Ndealkylation1	fr_Ndealkylation2
fr_Nhpyrrole	fr_SH	fr_aldehyde
fr_alkyl_carbamate	fr_alkyl_halide	fr_allylic_oxid
fr_amide	fr_amidine	fr_aniline
fr_aryl_methyl	fr_azide	fr_azo
fr_barbitur	fr_benzene	fr_benzodiazepine
fr_bicyclic	fr_diazo	fr_dihydropyridine
fr_epoxide	fr_ester	fr_ether
fr_furan	fr_guanido	fr_halogen
fr_hdrzine	fr_hdrzone	fr_imidazole
fr_imide	fr_isocyan	fr_isothiocyan
fr_ketone	fr_ketone_Topliss	fr_lactam

fr_lactone	fr_methoxy	fr_morpholine
fr_nitrile	fr_nitro	fr_nitro_arom
fr_nitro_arom_nonortho	fr_nitroso	fr_oxazole
fr_oxime	fr_para_hydroxylation	fr_phenol
fr_phenol_noOrthoHbond	fr_phos_acid	fr_phos_ester
fr_piperdine	fr_piperzine	fr_priamide
fr_prisulfonamd	fr_pyridine	fr_quatN
fr_sulfide	fr_sulfonamd	fr_sulfone
fr_term_acetylene	fr_tetrazole	fr_thiazole
fr_thiocyan	fr_thiophene	fr_unbrch_alkane
fr_urea	qed	

References

(1) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chemical Science* **2018**, *9*, 5441–5451.

(2) Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chemical Science* **2018**, *9*, 513530.

(3) Hans, C. Bayesian lasso regression. *Biometrika* **2009**, *96*, 835–845.