Supporting Information for: Linear Graphlet Models for Accurate and Interpretable Cheminformatics

Michael Tynes,[*,†,‡,¶,§] Michael G. Taylor,[†] Jan Janssen,[†] Daniel J. Burrill,[†,‡] Danny Perez,[†] Ping Yang,[*,†] and Nicholas Lubbers[*,‖]

[†] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

[‡] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

[¶] *Analytics, Intelligence, and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA (Current address)*

[§] *Department of Computer Science, The University of Chicago, Chicago, IL 60637, USA (Current address)*

[‖] *Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

E-mail: mtynes@uchicago.edu; pyang@lanl.gov; nlubbers@lanl.gov

S1 QM9

S1.1 L2 hyperparameter optimization

Following preliminary experiments, we searched over the below ranges for the linear regression regularization α. For the nonhierarchical linear models, we searched for α in a range of 10^{-2} to 10^{-1} for models base don Morgan fingerprints and over a range from 10^{-1} to 10^{2} for models built on graphlet and RDKit fingerprints. The ranges were motivated by slower runtimes for small values of α for the latter two types of fingerprint. For similar reasons, for hierarchical linear models we searched the range 10^{-2} to 10^{2} for fragment sizes $s \leq 4$, 10^{0} to 10^{3} for fragment sizes $s \leq 8$ and from 10^{1} to 10^{4} for fragment size $s = 9$.

S1.2 Fragment Counts by fingerprint type

Table S1: Number of fragments identified during training for each fingerprint type, for each maximum fragment size. For Graphlet fingerprints, size is the maximum number of atoms included in a fragment corresponding to a fingerprint element. For RDKit fingerprints, it is the maximum number of bonds. For Morgan, it is the Morgan radius.

	Morgan		RDKit		Graphlet	
Fragment Size s	# Fragments	(Cumulative)	# Fragments	(Cumulative)	# Fragments	(Cumulative)
0	25	(25)	-	-	-	-
1	1,068	(1,093)	29	(29)	9	(9)
2	68,669	(69,762)	106	(135)	40	(49)
3	329,419	(399,181)	518	(653)	198	(247)
4	99,133	(498,314)	2,530	(3,183)	992	(1,239)
5	2,162	(500,476)	11,998	(15,181)	4,536	(5,775)
6	0	-	50,041	(65,222)	19,524	(25,299)
7	-	-	178,608	(243,830)	77,599	(102,898)
8	-	-	514,229	(758,059)	270,792	(373,690)
9	-	-	1,145,734	(1,903,793)	776,315	(1,150,005)

S1.3 Learning curves

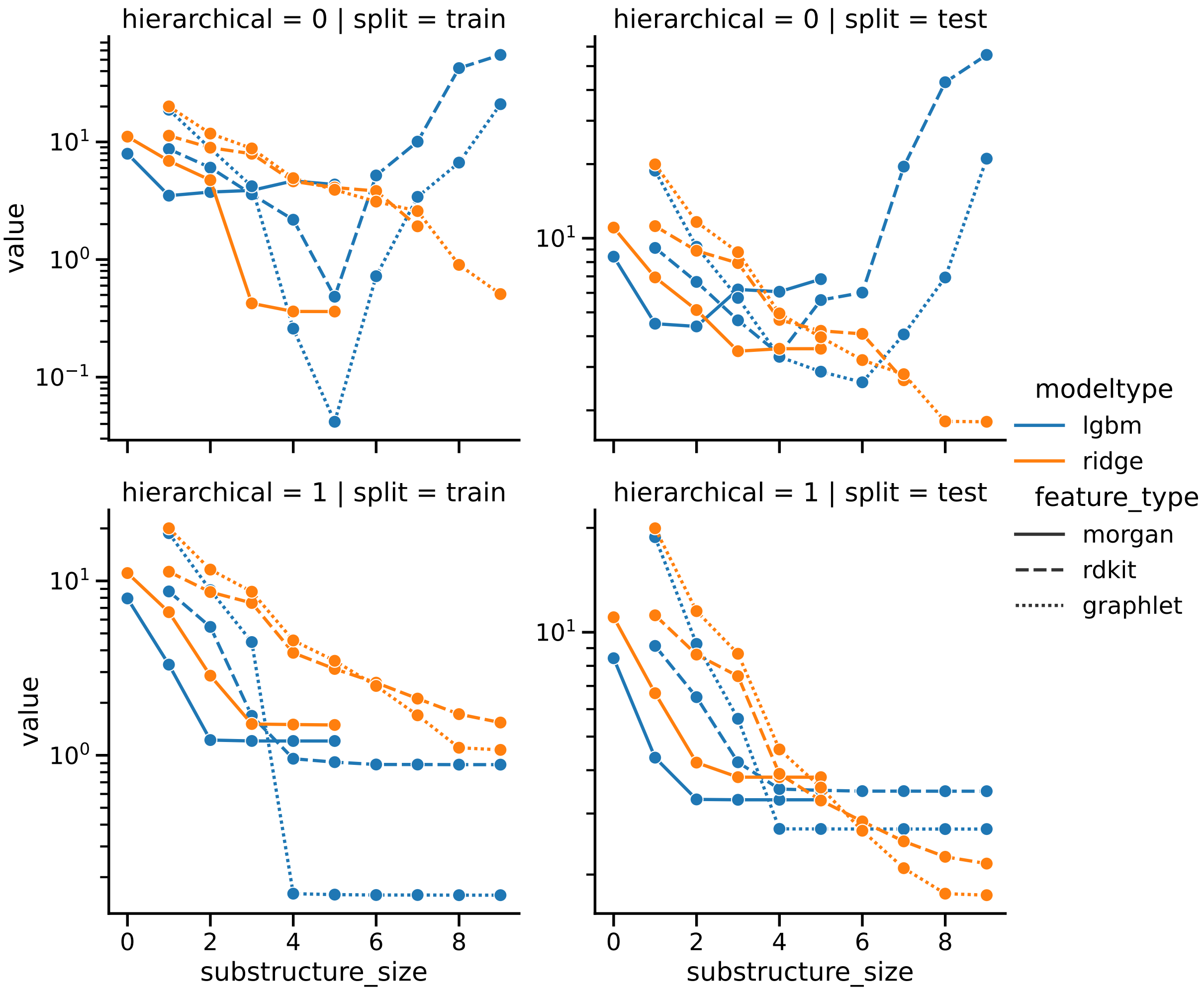

Figure S1: Learning curves for all model and fingerprint types by maximum size (substructure_size) for each fingerprint type. For Graphlet fingerprints (ours), size is the maximum number of atoms included in a fragment corresponding to a fingerprint element. For RDKit fingerprints, it is the maximum number of bonds. For Morgan, it is the Morgan radius.

S2 Bond Dissociation Energies

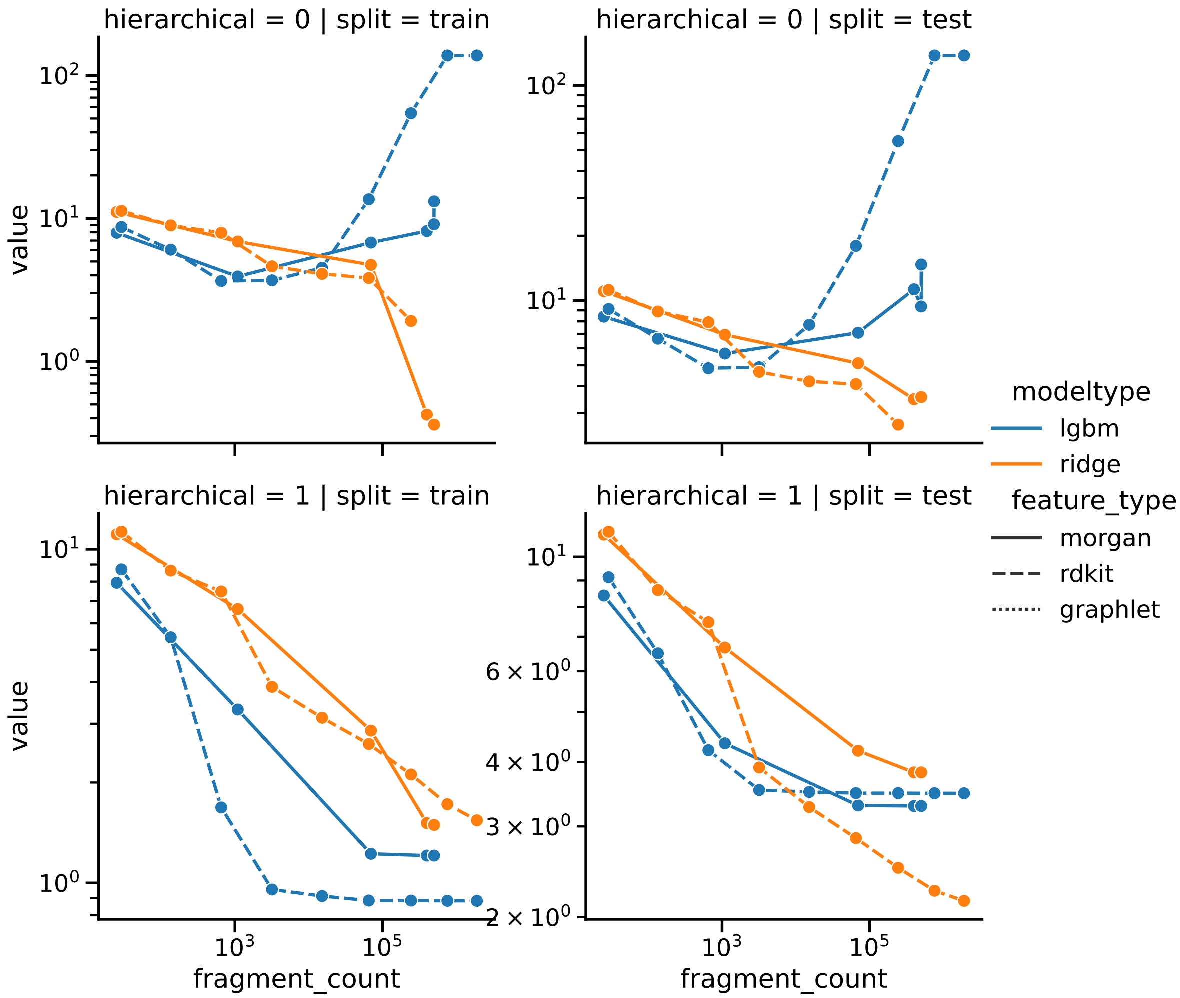

Figure S2: Learning curves for all model and fingerprint types by number of fragments identified during training. Each point corresponds to a choice of maximum size in figure S1, but here the horizontal axis has the same meaning for each curve.

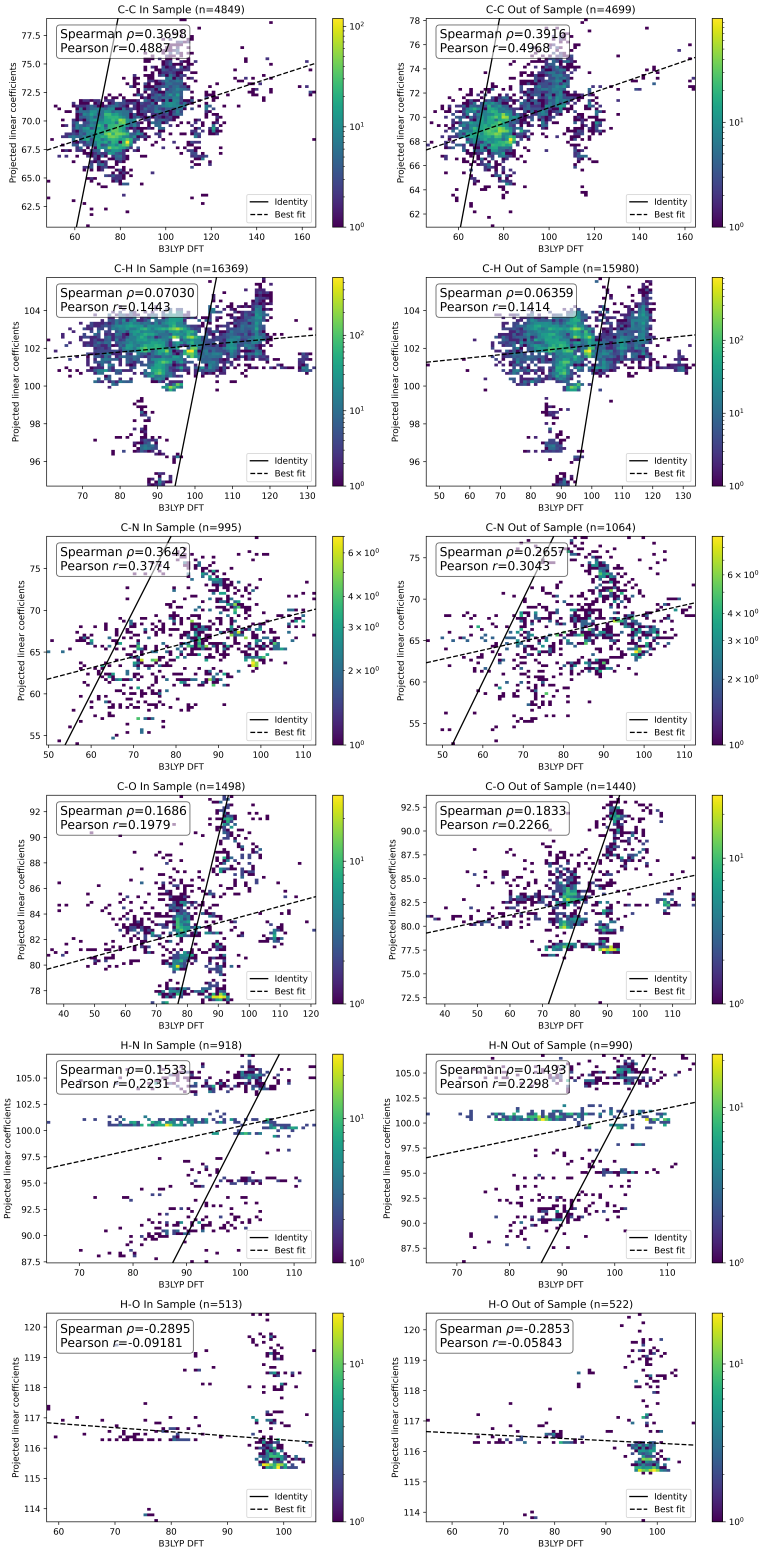

Figure S3: Relationships between bond-projected coefficients $\vec{\chi}^2$ and bond dissociation energies by bond type and dataset split.

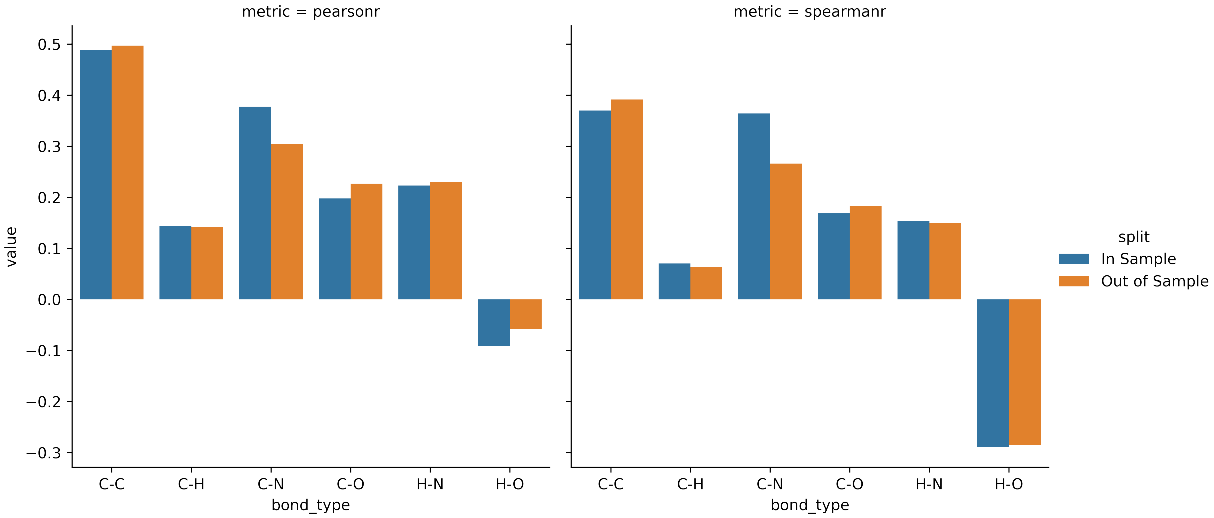

Figure S4: Correlation coefficients between bond-projected coefficients $\vec{\chi}^2$ and bond dissociation energies by bond type and dataset split.

S3 Graphlet Fingerprint Fragment Counts for Solubility Tasks

Table S2: Graphlet fingerprint feature counts for solubility tasks

				Dataset			
Fragment Size	Acetone	Benzene	Ethanol	Water	Water-wide	Water-WN	Average
1	25	22	26	24	24	24	24
2	116	101	127	118	125	125	118
3	451	354	514	515	570	570	495
4	0	1222	1932	0	0	0	525
Total	592	1699	2599	657	719	719	1164

S4 Solubility interpretation

Figure S5 shows atom-level coefficient projections constructed following Section 3.3.1 for five example molecules selected for structural diversity in acetone and water. These projected coefficients sum to the model prediction on each molecule, and can be thought of as the contributions of individual atoms to the model prediction after taking account its context in the molecular graph. By comparing the same molecule under different solvents, one can examine how structural motifs contribute to solubility in the different solvents. In this regard, the projections presented in Figure S5 largely agree with chemical intuition, e.g., carbon rings (molecules a and b) and chains (molecule d) explicitly lead to lower predictions of solubility in water than they do in acetone.

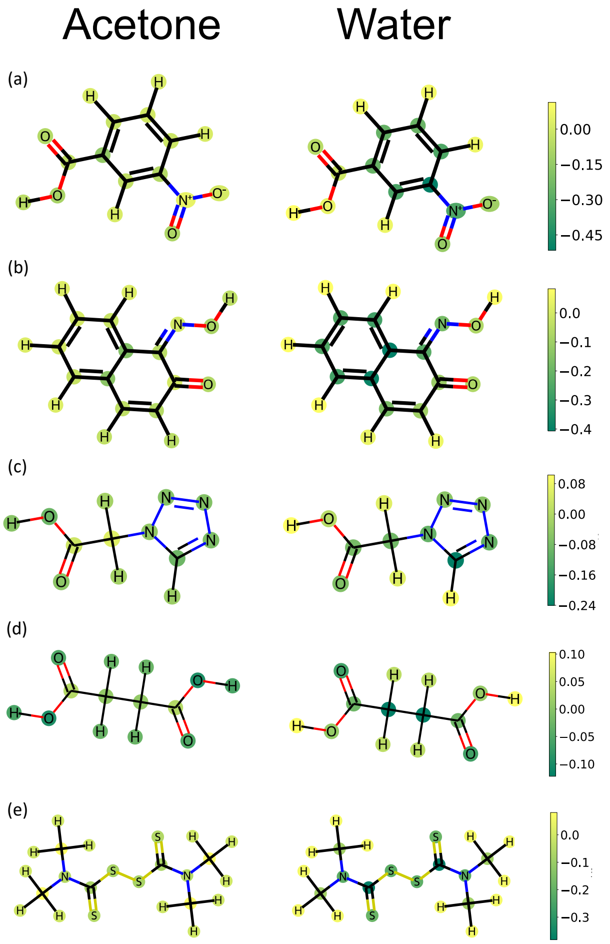

Figure S5: Linear model contributions projected to the atom level for molecules from the acetone and water sets presented in Ref 1. Colors show the contribution to the predicted log solubility (measured in molarity). Each column is based on predictions from the model fit on the corresponding solvent. Atoms in each row are colored using the same color map.

S5 TDC ADMET Leaderboard

S5.1 TDC ADMET Tasks

Our graphlet models were tested on Regression tasks in the Therapeutic Data Commons ADMET benchmark group. We give a brief description of the tasks, as well as the size of the dataset n and whether TDC scoring is performed using Mean Absolute Error (MAE), where lower is better, or Spearman's rank-rank coefficient ρ, where higher is better. The tasks are as follows:

1. **AqSol** ($n = 9,982$; MAE): Aqueous Solubility. A quantification of the ability for a molecule to dissolve in water.

2. **Caco2** ($n = 906$; MAE): Cell Effective Permeability. A cell-line-based experimental proxy for the rate at which a molecule may permeate the intestinal wall tissue.

3. **LD50** ($n = 7,385$; MAE): Acute Toxicity LD50. A measurement of the dosage at which a molecule is expected to be lethal.

4. **Lipo** ($n = 4,200$; MAE): Lipophilicity. A quantification of the ability for a molecule to dissolve in a lipid environment.

5. **PPBR** ($n = 1,614$; MAE): Plasma Protein Binding Rate. The fraction of a molecule which will be found bound to blood plasma proteins, as opposed to unbound molecules, which are far more pharmacologically active.

6. **CL-Hepa** ($n = 1,020$; ρ): Hepatocyte Clearance. A measurement of the volumetric rate of plasma which can be cleared of the molecule via hepatocytes.

7. **Cl-Micro** ($n = 1,102$; ρ): Microsome Clearance. A measurement of the volumetric rate of plasma which can be cleared of the molecule via microsomes.

8. **Half Life** (n = 667; ρ): Half Life. A measurement of the overall rate at which a molecule is eliminated from the body.

9. **VDss** (n = 1,130; ρ): Volume of Distribution at steady state. A measurement of the relative concentration of the molecule in the body overall in comparison to the concentration of the molecule in the plasma, given in units of volume.

S5.2 ADMET Results

Here we show performance results for our models compared to those on the existing TDC leaderboards[2] as bar plots.

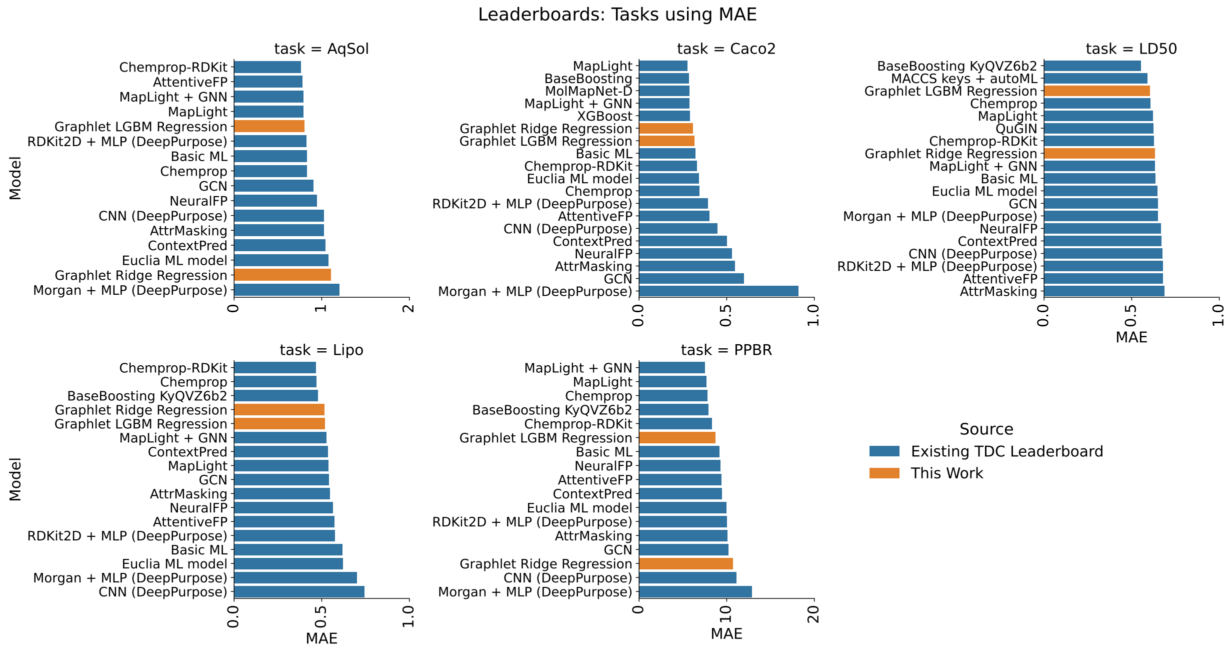

Figure S6: ADMET Performance bar plots for leaderboards where task performance is measured in mean absolute error (MAE) (lower is better)

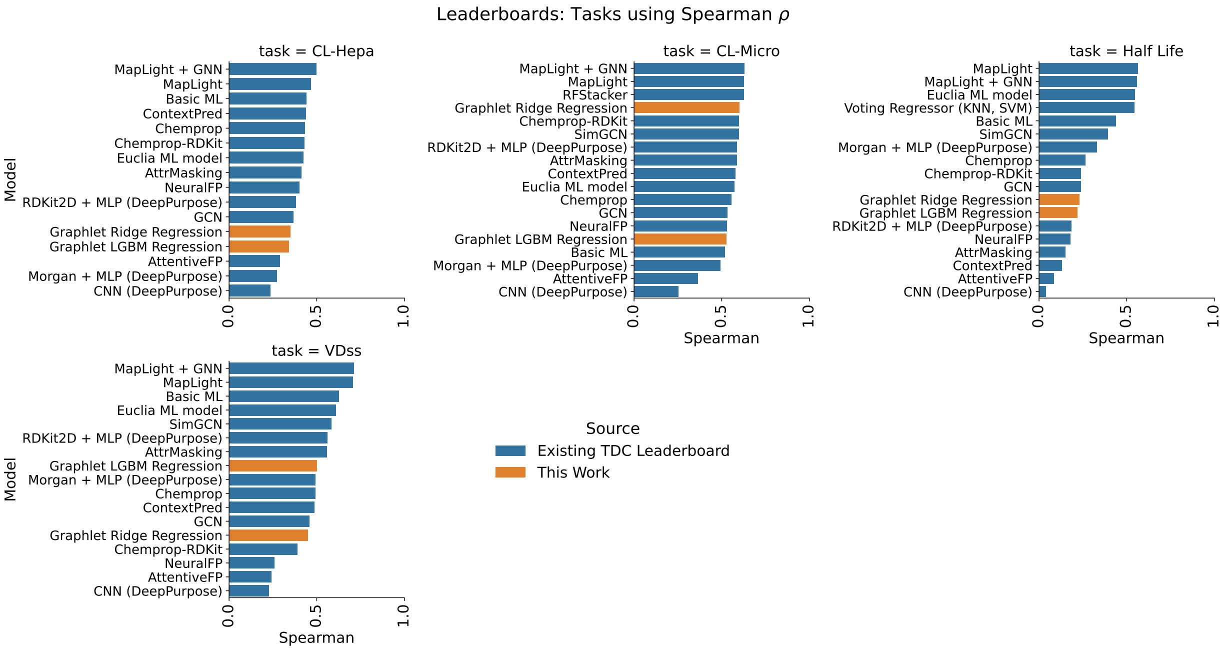

Figure S7: ADMET Performance bar plots for leaderboards where task performance is measured using Spearman's rank-rank correlation coefficient ρ (higher is better)

S6 Holdout fragments for adjustment experiments

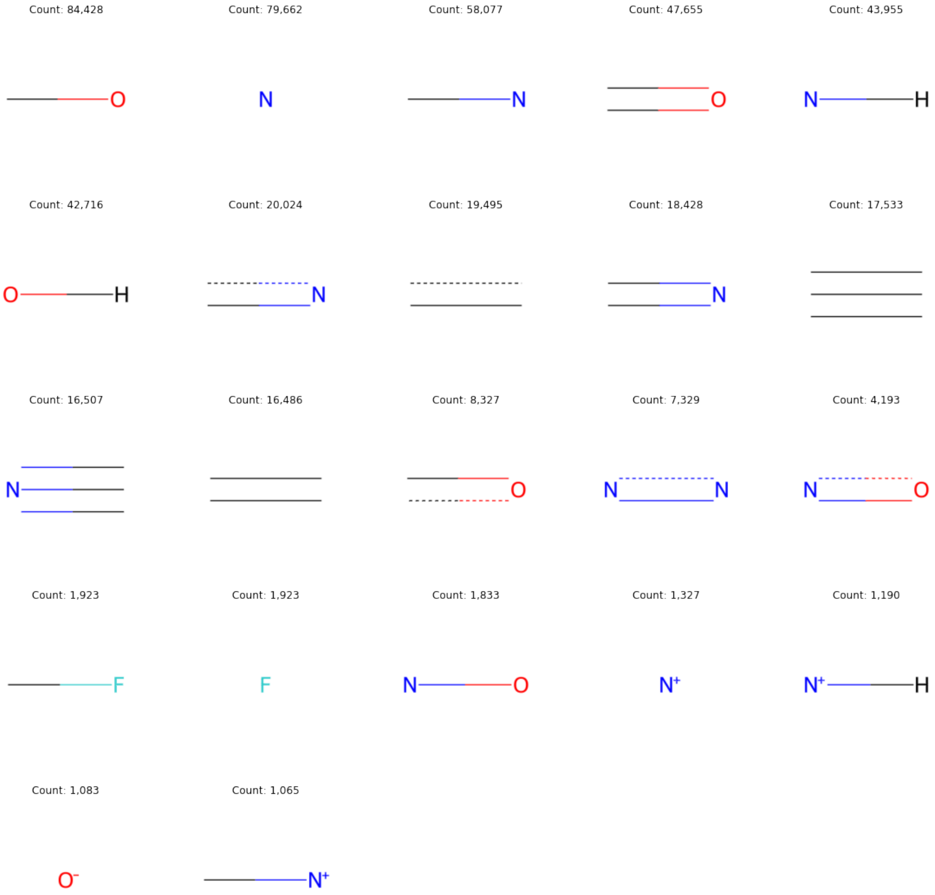

Figure S8: Molecular graphlets and their corresponding counts in QM9 that were used to construct fragment holdout sets for experiments in section 3.5

S7 Uncertainty Quantification

S7.1 UQ regression coefficients

Table S3: Error model coefficients for the calibrated uncertainty model.

Fragment Size	Error Model Coefficient (eV)
1	0
2	0
3	18.7
4	0
5	0.441
6	1.52
7	0

References

(1) Boobier, S.; Hose, D. R.; Blacker, A. J.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature communications* **2020**, *11*, 1–10.

(2) Therapeutics Data Commons: ADMET Leaderboards. `https://tdcommons.ai/benchmark/admet_group/overview/`, Accessed: 2023-07-24.