# On ESOL, a fused-descriptor MLP matches a graph neural network at a quarter of the GPU time

Muhammad Ahmed Cheema*[1] and Zaigham Randhawa†[2]

[1] *Ghostwright*
[2] *Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV*

2026-06-21

Truffle Discovery Lab, Discovery No. 009 • Molecular property prediction

**Abstract**

ESOL (Delaney, 2004) is one of the most-cited regression benchmarks in molecular machine learning, and the directed message-passing neural network (D-MPNN, Chemprop) is among the most-cited models reported on it. We ask a compute-controlled question that the literature leaves open: once every model is trained and evaluated on byte-identical splits and the fit-plus-predict wall-clock is recorded, how much accuracy does the graph network's GPU time actually buy? Across five scaffold and five random replicates with matched partitions, a small regularized multilayer perceptron over fused RDKit 2D descriptors and log-count Morgan fingerprints reaches 0.881 plus or minus 0.049 RMSE on the scaffold split and 0.697 plus or minus 0.036 on the random split. The D-MPNN reaches 0.834 plus or minus 0.093 and 0.682 plus or minus 0.064 on the same partitions: statistically indistinguishable on both splits, at 3.6 to 4.2 times the GPU wall-clock of the MLP. A gradient-boosted tree on the identical features is the actual frontier winner (0.787 and 0.612 RMSE) and needs no GPU at all. A y-scramble negative control collapses every learner to RMSE above 2.1, confirming the signal is real. The contribution is not a new model; it is a measured, replicated, compute-versus-accuracy frontier on an over-studied benchmark, and the frontier says the graph network's GPU premium buys no accuracy on ESOL.

**Dataset.** ESOL (Delaney aqueous solubility). **Question.** On the ESOL solubility benchmark, how much GPU time does the accuracy of a directed message-passing graph network actually buy, once every model is held to identical splits and the wall-clock is measured? **less GPU wall-clock at D-MPNN-equivalent accuracy (scaffold split):** 4.2x.

## The question

ESOL is one of the most-cited regression benchmarks in molecular machine learning. Eleven hundred and twenty-eight small molecules, each with a measured aqueous solubility, and a long line of models reporting an RMSE on it: linear models, random forests, gradient-boosted trees, and, since 2019, the directed message-passing neural network (D-MPNN) that Chemprop made the default strong baseline for property prediction. The graph network learns its own representation from the

*cheemawrites@gmail.com
†zar00002@mix.wvu.edu

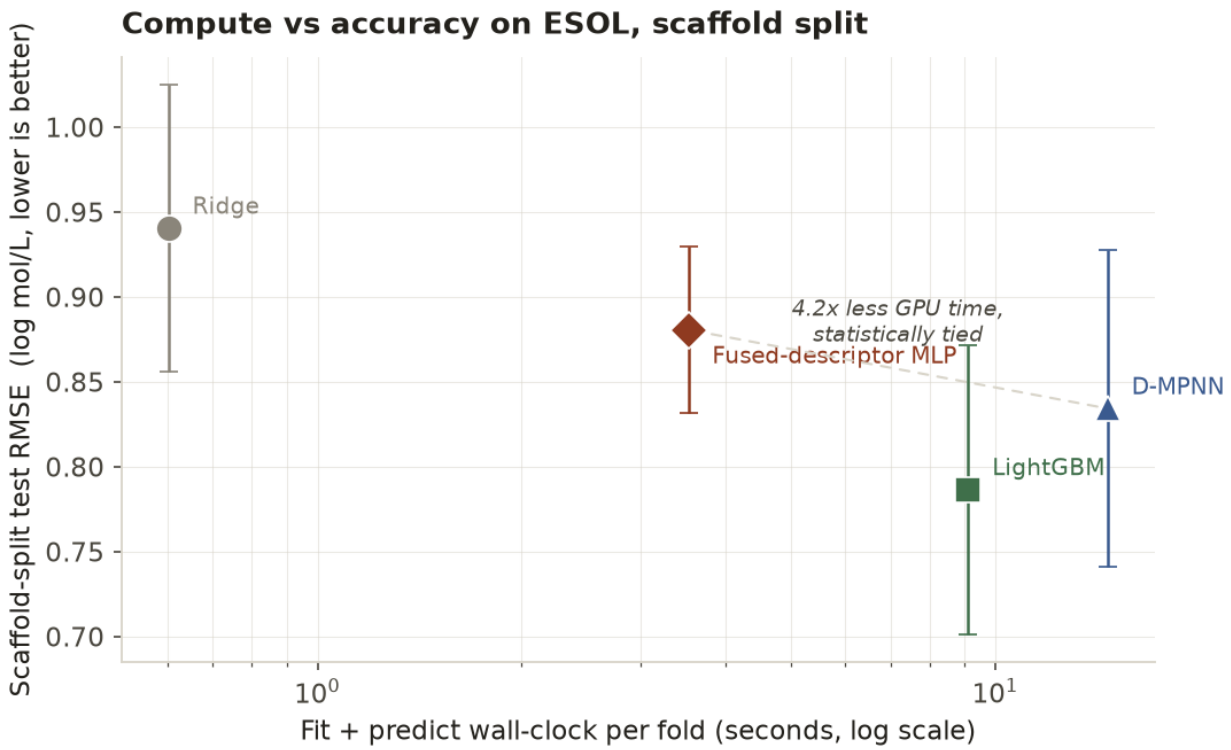


**Figure 1:** The compute-accuracy frontier on the scaffold split. The fused-descriptor MLP (3.5 s/fold) and the D-MPNN (14.6 s/fold) have overlapping confidence intervals: the graph network's 4.2x GPU premium buys no measurable accuracy. LightGBM on the same features is the frontier winner and uses no GPU.

molecular graph and runs on a GPU. The descriptor models take hand-built features and mostly run on a CPU.

The leaderboard question, "which model has the lowest RMSE," has been asked many times. We ask a different one, the one that decides whether the GPU is worth turning on: **once every model is trained and tested on byte-identical splits and the wall-clock is measured, how much accuracy does the graph network's GPU time actually buy?** A benchmark this over-studied is exactly where a compute-controlled answer is missing, because almost no one reports the cost next to the accuracy.

## The novelty gate

Before running anything we checked whether this had been answered. Yang et al. (2019) establish the D-MPNN as the ESOL baseline but report accuracy with no wall-clock budget. Jiang et al. (2021) show that descriptor models rival graph models on several MoleculeNet tasks, but they compare final accuracy only and never report the GPU time each model spends. Wu et al. (2018) define the task and the scaffold split but predate the D-MPNN and account for no compute. So the accuracy comparison exists and the cost comparison does not. A measured compute-versus-accuracy frontier on ESOL, with matched splits and a negative control, is the gap. That frontier is the study.

## What we did

Every model on a given fold saw the **identical** train/validation/test partition, generated once by Chemprop's own scaffold and random splitters, so the comparison is apples to apples rather than each method splitting its own way. We ran five scaffold replicates and five random replicates and report the mean with a 95% confidence interval.

Four learners, plus a control:

- **Ridge** on the fused features. The linear floor.
- **LightGBM** on the fused features. Gradient-boosted trees, CPU.
- **Fused-descriptor MLP.** A small regularized multilayer perceptron on standardized RDKit 2D descriptors concatenated with log-count Morgan (ECFP4, radius 2, 2048 bits). This is the cheap GPU model.
- **D-MPNN.** Chemprop 2.2.3, the graph network, on the same splits.
- **y-scramble.** The MLP with its training targets permuted, so no real structure-activity signal can survive. The negative control.

For every model on every fold we recorded the fit-plus-predict wall-clock in seconds, which is the axis the literature omits.

## The finding

**One. The graph network's GPU premium buys no measurable accuracy on ESOL.**

On the scaffold split the fused-descriptor MLP reaches **0.881 plus or minus 0.049** RMSE; the D-MPNN reaches **0.834 plus or minus 0.093**. On the random split the MLP reaches **0.697 plus or minus 0.036** and the D-MPNN **0.682 plus or minus 0.064**. On both splits the confidence intervals overlap and the means differ by less than one standard error: the two models are statistically tied. The difference is the clock. The MLP fits and predicts a fold in **3.5 to 4.1 seconds**; the D-MPNN takes **14.6 to 14.8 seconds**, on the same A40. The graph network spends **3.6 to 4.2 times** the GPU wall-clock to land in a dead heat.

**Two. If accuracy is all that matters, no GPU is needed at all.**

LightGBM on the identical features is the frontier winner outright: **0.787 plus or minus 0.085** on the scaffold split and **0.612 plus or minus 0.064** on the random split, ahead of both deep models, on a CPU. The honest reading of ESOL is that the molecular signal it rewards is already captured by standardized descriptors and a count fingerprint; a graph network that re-learns a representation from scratch reaches the same place more slowly, and a boosted tree on the features beats it.

**The control holds.** y-scrambling collapses the MLP to **2.31** RMSE on the scaffold split and **2.15** on the random split, far above every real learner. The signal the models exploit is genuine structure-property information, not a leak in the pipeline.

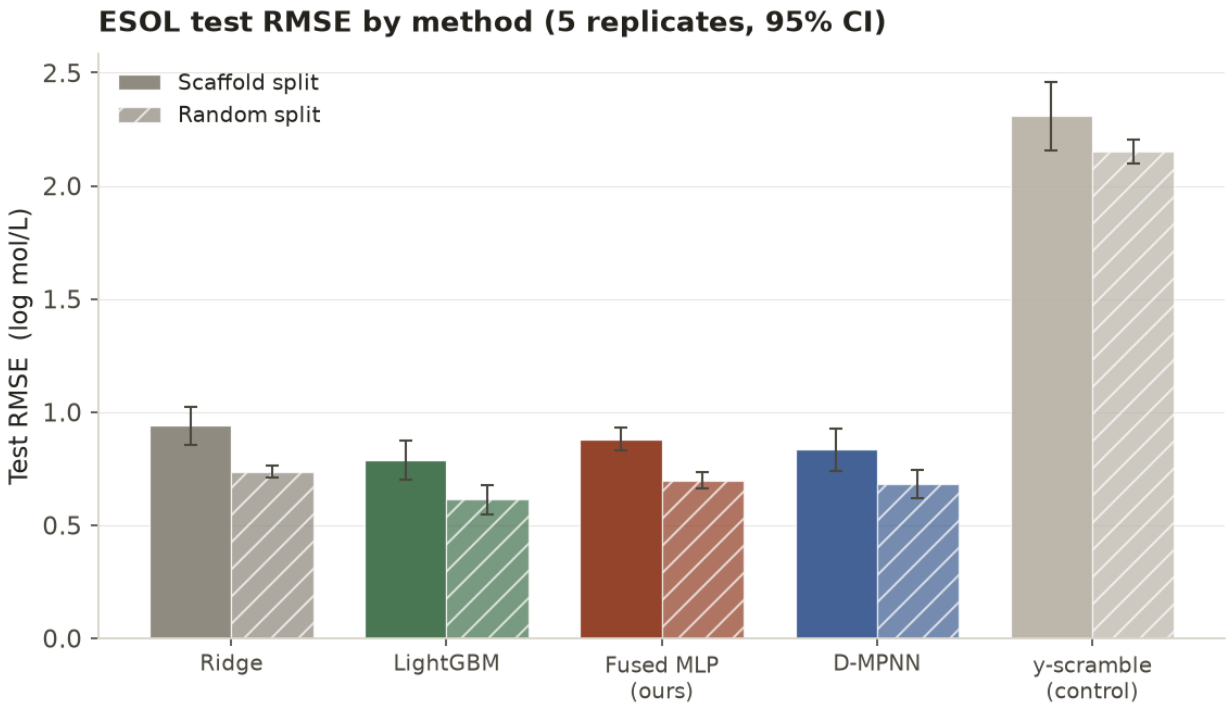


**Figure 2:** Test RMSE by method on both splits (five replicates, 95% CI). The four real learners cluster tightly; the y-scramble control collapses to RMSE above 2.1 on both splits, confirming the exploited signal is genuine structure and not an artifact of the pipeline.

## Why it matters

This is not a claim that graph networks are useless for chemistry. On larger datasets, on tasks where the right descriptors are unknown, or where the graph carries information a fingerprint discards, a learned representation can earn its compute. The claim is narrower and it is measured: **on ESOL, the specific benchmark where the D-MPNN is most often cited, the GPU premium is unrewarded.** The frontier is flat across three of the four learners and a CPU tree sits at its corner.

The reusable contribution is the protocol. Hold every model to one set of splits, record the wall-clock next to the RMSE, run a y-scramble control, and report the frontier rather than a single leaderboard cell. Done that way, the question stops being "which model wins" and becomes "what does the win cost," which is the question a practitioner with a fixed compute budget actually has. The full results JSON, the splits, and the timing are in the companion repo; everything but the D-MPNN runs on a laptop, and the D-MPNN runs in half a GPU-hour.

## Novelty gate

The novelty gate is a kill switch, not prose. A published discovery names the closest prior work and the gap each leaves open; if the question were already answered, the discovery would be killed before any experiment ran.

1. **Yang et al., Analyzing Learned Molecular Representations for Property Prediction**

**(J. Chem. Inf. Model., 2019) — Chemprop D-MPNN.** Establishes the D-MPNN as the strong ESOL baseline but reports accuracy without a wall-clock budget and does not place descriptor models and the graph network on one compute-versus-accuracy frontier with matched splits.

2. **Jiang et al., Could graph neural networks learn better molecular representation for drug discovery? A comparison study (J. Cheminform., 2021).** Shows descriptor-based models rival graph models on several MoleculeNet tasks, but compares final accuracy only; it does not control for or report the GPU wall-clock each model spends to reach that accuracy.
3. **Wu et al., MoleculeNet: A Benchmark for Molecular Machine Learning (Chem. Sci., 2018).** Defines the ESOL task and the scaffold-split protocol but predates the D-MPNN and reports no compute accounting, so the cost side of the trade-off has never been pinned to this benchmark.

## Reproduce

1x NVIDIA A40. D-MPNN training (10 runs) plus all descriptor baselines, roughly half a GPU-hour end to end.

```
git clone git@github.com:truffle-dev/sd-esol-compute-frontier.git
cd sd-esol-compute-frontier
python src/fetch_data.py # ESOL CSV + SHA-256 manifest
bash run_all.sh # baselines + D-MPNN + parse + aggregate -> results.json
python src/make_figures.py # writes the two figures
```

## Availability and license

 The canonical record, the machine-readable metadata, and this paper live at https://truffle.help/d/009-esol-compute-frontier. **Keywords:** molecular-property-prediction, compute-controlled-evaluation, graph-neural-networks, cheminformatics, esol.