Supporting Information for "When Machine Learning Models Learn Chemistry I: Quantifying Explainability with Matched Molecular Pairs"

Kerrin Janssen[1], Jan M. Wollschläger[2], Jonny Proppe[1,*] and Andreas H. Göller[3,*]

November 14, 2025

[1]TU Braunschweig
Institute of Physical and Theoretical Chemistry
Gauss Str 17, 38106 Braunschweig, Germany

[2]Bayer AG
Pharmaceuticals, R&D, Machine Learning Research
13353 Berlin, Germany

[3]Bayer AG
Pharmaceuticals, R&D, Computational Molecular Design
42096 Wuppertal, Germany

1 MMP Generation Analysis

To illustrate the impact of different MMP generation settings, we applied them to the Crippen log P `chemprop` model, with the results summarized in Table SI-1.

Table SI-1: Explainability performance under different MMP generation settings using the Crippen log P `chemprop` model.

MMP Settings	Training Set					Test Set				
	r^2 Whole Molecule to Pred.	r^2 Variable Part to Pred.	r^2 Whole Molecule to POI	std Constant Part	Accuracy Variable Part to Pred.	r^2 Whole Molecule to Pred.	r^2 Variable Part to Pred.	r^2 Whole Molecule to POI	std Constant Part	Accuracy Variable Part to Pred.
`mmpdb` default + `num_cuts` = 1 + `max_variable_ratio` = 0.2	0.73	0.79	0.66	2.85	0.90	0.73	0.77	0.64	3.10	0.93
`mmpdb` default + `num_cuts` = 1	0.39	0.56	0.35	2.75	0.81	0.58	0.73	0.51	2.83	0.85

2 Grid Search

The results of the best performing models from the grid search with a five-fold cross-validation are shown in Table SI-2.

*E-mail: j.proppe@tu-braunschweig.de, andreas.goeller@bayer.com

Table SI-2: Mean performance on the five-fold cross-validation on the training set with the standard deviation displayed in brackets.

Property of Interest	**Model Type**	r^2	R^2	**MAE**	**RMSE**
Crippen log P	MolGraph; Chemprop	0.92 (0.016)	0.92 (0.016)	0.25 (0.024)	0.37 (0.049)
exp log P	MolGraph; Chemprop	0.65 (0.097)	0.60 (0.17)	0.50 (0.030)	0.74 (0.15)
Solubility	MolGraph; Chemprop	0.88 (0.016)	0.87 (0.016)	0.54 (0.031)	0.73 (0.034)
Crippen log P	Morgan Fingerprint; Bayesian Ridge	0.71 (0.0098)	0.70 (0.0093)	0.51 (0.019)	0.71 (0.036)
exp log P	MACCS Fingerprint; SVR	0.51 (0.042)	0.51 (0.043)	0.63 (0.045)	0.83 (0.050)
Solubility	MACCS Fingerprint; Gradient Boosting	0.76 (0.025)	0.76 (0.022)	0.76 (0.034)	1.01 (0.030)

3 Significance Test

A pairwise t-test was conducted using the `ttest_rel` python function to assess the significance of performance differences between models on the absolute error on the test set (Table 2). The significance level was set to 0.01. The null hypothesis (H_0) stated that both models trained on the same dataset exhibit equal error rates.

Table SI-3: Pairwise p-values on the prediction errors on the test sets between the different model types on the same datasets.

Model 1	**Model 2**	**p-value**
Crippen log P Chemprop	Crippen log P Bayesian Ridge	$2.72 \cdot 10^{-47}$
exp log P Chemprop	exp log P SVR	$3.96 \cdot 10^{-22}$
Solubility Chemprop	Solubility Gradient Boosting	$3.48 \cdot 10^{-5}$