Journal of Data-centric Machine Learning Research (2023) Submitted 02/02/2024; Published -

Denoising Drug Discovery ADMET Data for Improved Regression Task Performance

Matthew Adrian MATTHEW.ADRIAN@MERCK.COM
Modeling and Informatics
Merck & Co., Inc. South San Francisco, California 94080, USA

Yunsie Chung YUNSIE.CHUNG@MERCK.COM
Modeling and Informatics
Merck & Co., Inc. South San Francisco, California 94080, USA

Alan Cheng ALAN.CHENG@MERCK.COM
Modeling and Informatics
Merck & Co., Inc. South San Francisco, California 94080, USA

Abstract

Predicting ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of small molecules is a key task in drug discovery. A major challenge in building better ADMET models is the experimental error inherent in the data. Here, we develop denoising schemes based on deep learning to address this. The most significant performance increase occurs when the original model is finetuned with the denoised data using training error as the noise detection metric. Our denoising scheme outperforms other literature schemes for ADMET data and has implications for improving models for experimental assay data in general.

1 Introduction

Predicting ADMET properties is a crucial task in the optimization of small molecules during drug discovery (Ferreira and Andricopulo, 2019; Cáceres et al., 2020; Beckers et al., 2023). ADMET assays, in practice, have experimental error even when using validated procedures (Wenlock and Carlsson, 2015). In previous studies, ensemble-based, split-based, and forgotten events methods were used to detect noise, but mostly in classification tasks (Gupta and Gupta, 2019; Nguyen et al., 2020; Toniato et al., 2021; Yuan et al., 2022; Kim et al., 2023). ADMET predictors are typically regression tasks due to the continuous nature of the data, making it difficult to apply existing denoising schemes. In this study, we survey several noise detection metrics and devise denoising schemes for ADMET assay data. The results show that finetuning the model with the data denoised based on the training error gives the best performance improvement. To our knowledge, this is the first study to present a denoising scheme for assay data that improves predictive performance on regression tasks.

2 Methods

Artificial gaussian noise was added to varying percentages of the data at varying magnitudes as shown in Figure 1. Graph convolutional network models were then trained on noisy data using Chemprop (Heid et al., 2024). The top 10% of 'noisy' data were removed using four

different noise detection metrics, as shown in Figure 2: (1) absolute training error (TE), (2) number of forgotten events, (3) ensemble variance, and (4) split variance. For the top performing TE metric, four denoising schemes were tested: (1) filtering out the top 10% noisy data as the baseline (TE Filter); (2) finetuning the model after filtering (TE Finetune); (3) replacing the data with the average of the model prediction and dataset value (Mean Correction); and (4) replacing the data with the model prediction (Prediction Correction). The TE Filter was benchmarked against Ground Truth and Random Filters, which filter out the true top noisy data and randomly selected data, respectively. The models were tested on a 20% held-out set without artificial noise, and across four ADMET datasets for LogD, $F_{u,p}$, P_{app}, and hERG binding (see Table 1 for details). Further Methods details are available in Appendix A.

3 Results and Discussions

Metrics from the literature—split variance, ensemble variance, and forgotten event metrics—have no correlation with added noise in the data, as shown in Figures 3-6. Therefore, these are not suitable as noise detection metrics for ADMET data. This inconsistency with prior studies (Toniato et al., 2021; Yuan et al., 2022) may be due to differences in modeling of regression over classification data. The TE metric we introduced here shows a strong correlation with data noise, and this correlation improves as the magnitude of noise increases. The TE does not correlate with either the number of rare atoms or similarity to train molecules, indicating its effectiveness in filtering out noisy data without excluding important data (further details in Appendix B).

Our model trained after TE Finetuning has improved performance relative to the original model built without de-noising, especially in the medium noise regions, with average R^2 gains of 0.01-0.05 (Figures 7-11). Considering that the model with the Ground Truth Filter improves R^2 by 0.03-0.09 for these regions, the performance boost from our denoising scheme is not negligible. In the low noise and high noise regions, the performance is similar to the original model. Our other denoising schemes that use TE are similarly improved in the medium noise regions and had similar or worse performance to the original model in low noise regions. In the high-noise regime where noise is added to 100% of data, de-noising methods other than TE Finetuning cause the performance to worsen. The TE Finetuning Model has the largest average performance increase in the medium noise region relative to the base model (Figure 11). Preliminary results also indicate that R^2 improvement up to 0.07 can be achieved using an adaptive threshold finetuning approach (Figure 11). Finetuning is advantageous over filtering because there is no loss of information from filtering.

4 Conclusion

We propose a denoising scheme that uses training error as a noise detection metric and we use finetuning to improve the performance of models after denoising the dataset. We are currently assessing our improved predictive approach on larger datasets and prospective use cases.

Broader Impact Statement

Improving the performance of ADMET predictors will enable drug discovery teams to discover better therapeutics at a faster rate. We have developed new denoising schemes that improve predictive ADMET models, with broader implications for modeling any experimental assay data.

Acknowledgments and Disclosure of Funding

The authors have no conflicts of interest to declare. This work was funded by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA.

References

Ignacio Aliagas, Alberto Gobbi, Man Ling Lee, and Benjamin D. Sellers. Comparison of logP and logD correction models trained with public and proprietary data sets. *Journal of Computer-Aided Molecular Design*, 36(3):253–262, 2022. doi: 10.1007/s10822-022-00450-9.

Maximilian Beckers, Noé Sturm, Finton Sirockin, Nikolas Fechner, and Nikolaus Stiefl. Prediction of small-molecule developability using large-scale in silico admet models. *Journal of Medicinal Chemistry*, 66(20):14047–14060, 2023. doi: 10.1021/acs.jmedchem.3c01083.

Rodolpho C. Braga, Vinicius M. Alves, Meryck F.B. Silva, Eugene Muratov, Denis Fourches, Luciano M. Lião, Alexander Tropsha, and Carolina H. Andrade. Pred-hERG: A Novel web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Molecular Informatics*, 34(10):698–701, 2015. doi: 10.1002/minf.201500040.

Elena L. Cáceres, Matthew Tudor, and Alan C. Cheng. Deep learning approaches in predicting admet properties. *Future Medicinal Chemistry*, 12(22):1995–1999, 2020. doi: 10.4155/fmc-2020-0259.

Gabriela Falcón-Cano, Christophe Molina, and Miguel Ángel Cabrera-Pérez. Reliable Prediction of Caco-2 Permeability by Supervised Recursive Machine Learning Approaches. *Pharmaceutics*, 14(10), 2022. doi: 10.3390/pharmaceutics14101998.

Leonardo L.G. Ferreira and Adriano D. Andricopulo. Admet modeling approaches in drug discovery. *Drug Discovery Today*, 24:1157–1165, 2019. doi: 10.1016/j.drudis.2019.03.015.

Shivani Gupta and Atul Gupta. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science*, 161:466–474, 2019. doi: 10.1016/j.procs.2019.11.146.

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64:9–17, 2024. doi: 10.1021/acs.jcim.3c01250.

Hiroaki Iwata, Tatsuru Matsuo, Hideaki Mamada, Takahisa Motomura, Mayumi Matsushita, Takeshi Fujiwara, Kazuya Maeda, and Koichi Handa. Predicting Total Drug Clearance and Volumes of Distribution Using the Machine Learning-Mediated Multimodal Method through the Imputation of Various Nonclinical Data. *Journal of Chemical Information and Modeling*, 62(17):4057–4065, 2022. doi: 10.1021/acs.jcim.2c00318.

Jihye Kim, Aristide Baratin, Yan Zhang, and Simon Lacoste-Julien. Crosssplit: Mitigating label noise memorization through data splitting. *International Conference on Machine Learning*, pages 16377–16392, 2023. doi: 10.48550/arXiv.2212.01674.

Chuang Li and Zhizhong Mao. A label noise filtering method for regression based on adaptive threshold and noise score. *Expert Systems with Applications*, 228:120422, 2023. doi: 10.1016/j.eswa.2023.120422.

Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. *8th International Conference on Learning Representations, ICLR 2020*, pages 1–15, 2020.

Alessandra Toniato, Philippe Schwaller, Antonio Cardinale, Joppe Geluykens, and Teodoro Laino. Unassisted noise reduction of chemical reaction datasets. *Nature Machine Intelligence*, 3:485–494, 2021. doi: 10.1038/s42256-021-00319-w.

Reiko Watanabe, Tsuyoshi Esaki, Hitoshi Kawashima, Yayoi Natsume-Kitatani, Chioko Nagao, Rikiya Ohashi, and Kenji Mizuguchi. Predicting Fraction Unbound in Human Plasma from Chemical Structure: Improved Accuracy in the Low Value Ranges. *Molecular Pharmaceutics*, 15(11):5302–5311, 11 2018. doi: 10.1021/acs.molpharmaceut.8b00785.

Mark C. Wenlock and Lars A. Carlsson. How experimental errors influence drug metabolism and pharmacokinetic qsar/qspr models. *Journal of Chemical Information and Modeling*, 55:125–134, 2015. doi: 10.1021/ci500535s.

Jan Wenzel, Hans Matter, and Friedemann Schmidt. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling*, 59(3):1253–1268, 2019. doi: 10.1021/acs.jcim.8b00785.

Bo Yuan, Farhad Hormozdiari, Cory Y. McLean, and Justin Cosentino. An empirical study of ml-based phenotyping and denoising for improved genomic discovery. *bioRxiv*, 2022. doi: 10.1101/2022.11.17.516907.

Hang Zhou, Jonas Mueller, Mayank Kumar, Jane-Ling Wang, and Jing Lei. Detecting errors in numerical data via any regression model. *ICML Workshop on Data-centric Machine Learning Research*, 2023. doi: 10.48550/arXiv.2305.16583.

Appendix A. Further details on Methods

A.1 Noise addition procedure

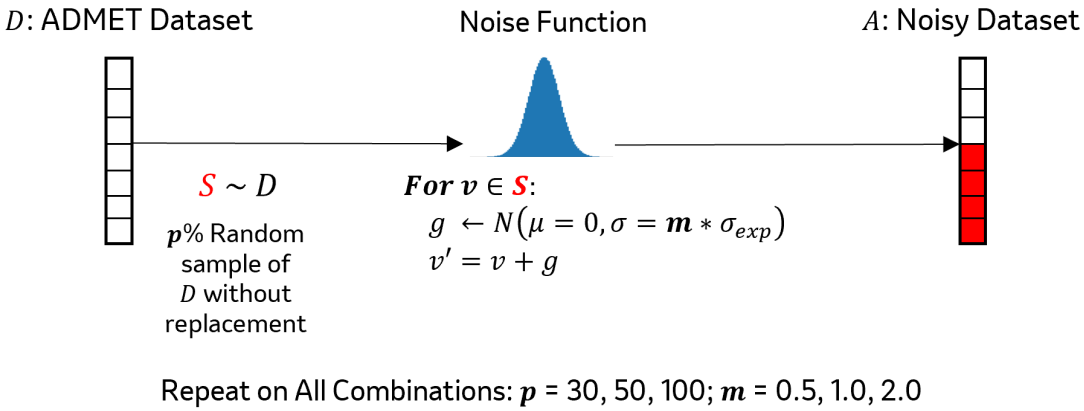

Figure 1: Overview of noisy dataset creation. For each dataset, this creates nine unique noisy training sets.

Artificial random gaussian noise was added to varying percents of the data ($p \subset 30, 50, 100$) and at varying standard deviations ($\sigma = m{*}\sigma_{\text{expt}}$ where $m \subset 0.5, 1, 2$) as described in Figure 1. The standard deviation of the original dataset distribution, σ_{expt}, normalizes the magnitude of noise being added to each unique dataset. For the purposes of quantifying noise in each datapoint, we defined the original ADMET data as clean ("no noise") reference data and measured our denoising schemes against the artificial added noise. Combining all sampled percentages and magnitudes results in 10 different noise combinations for each dataset, including the original data set which has no artificial noise added. The performance of the models was evaluated on a 20% held-out test set without artificial noise.

A.2 Noise detection metrics

Four noise detection metrics were surveyed: (1) Training error, (2) number of forgotten events, (3) ensemble variance, and (4) split variance. A forgotten event at epoch n occurs when the training error of epoch n is greater than the training error of epoch $n-1$:

$$\text{Forgotten Event} = \begin{cases} 1, & \text{if TrainingError}_{\text{epoch } n} > \text{TrainingError}_{\text{epoch } n-1}, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The total number of forgotten events is calculated by the summation of forgotten events across all training epochs. Ensemble variance is calculated by the variance of predictions of four different models trained on the same dataset: two with varying initial parameters and two with varying random validation sets. Split variance is calculated by the variance of predictions of three different models each trained on equal-sized, non-overlapping splits of the data.

A.3 Denoising schemes and baselines

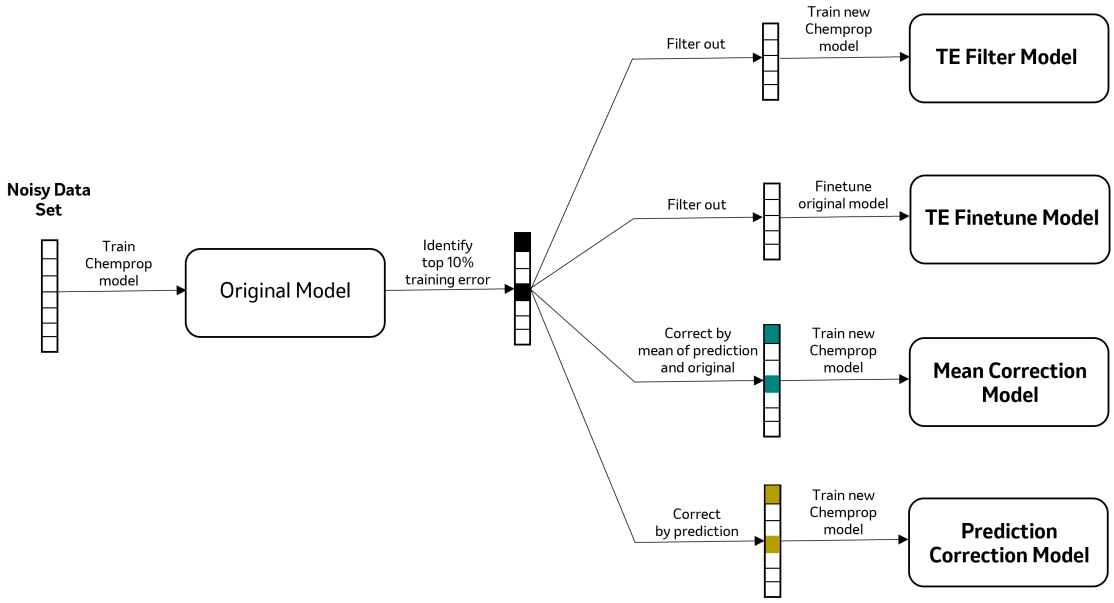

Figure 2: Overview of our four denoising schemes. In this visualization, absolute training error is being used as a metric to detect noise. The noise detection metric is interchangeable for each denoising scheme.

Four main denoising schemes were tested as visualized in Figure 2. The TE Filter Model uses a filtering of the top 10% molecules with the highest training error. The filtered dataset is then fed into a new Chemprop model to train. The TE Finetune Model uses the same filter as the TE Filter Model but instead, the final model comes from finetuning the original model on the filtered dataset. The Mean Correction Model differs from the two filter models as it replaces corresponding values of the top 10% molecules with the mean of the predicted and original values. Similarly, the Prediction Correction Model replaces these values with just the prediction. These four denoising schemes were tested using training error as the metric. The other metrics were tested using a denoising scheme analogous to the TE Filter Model. These models are referred to as the FE Filter Model, EV Filter Model, and SV Filter Model accordingly.

We also assessed performance against models built using two baseline schemes that are analogous to the TE Filter Model. The Ground Truth Filter Model is an oracle which filters the 10% data with the true highest noise. The Random Filter Model filters out 10% data randomly.

A.4 Details on datasets

Table 1: ADMET assay datasets used in this study.

Assay	Data count	Description	Data source
LogD	4190	Distribution coefficient between octanol and water at pH 7.4 in $\log_{10}$ unit	Aliagas et al. (2022)
Fraction of unbound plasma in human ($F_{u,p}$)	2717	In $\log_{10}$(fraction unbound)	Watanabe et al. (2018), Iwata et al. (2022)
Apparent permeability (P_{app})	6457	Caco-2 apparent permeability in $\log_{10}(10^{-6}$ cm/s)	Wenzel et al. (2019), Falcón-Cano et al. (2022)
hERG binding	5108	Binding affinity (IC50) to human hERG potassium (K+) channel in log10(nMolar IC50)	Braga et al. (2015)

A.5 Model details

All ADMET models were built on a directed message passing neural network (D-MPNN) based architecture as implemented in Chemprop (Heid et al., 2024). The hyperparameters were chosen based off optimization done previously for this data. Chemprop's default hyperparameters were used for model parameters not specified in each table.

Table 2: Chemprop model parameters

Hyperparameter	Value
MPN depth	3
MPN hidden size	600
FFN number of layers	3
FFN hidden size	1200
Dropout	0
Aggregation	Norm
Number of folds (training/validation split seed)	2
Ensemble size (parameter initialization seed)	2
Epochs	80 (15 for finetuning)

Appendix B. Additional details on results

B.1 Correlation between the noise and noise metrics

To evaluate the suitability of each noise detection metric, we compared the Spearman R correlation of that metric to absolute artificial noise added. A higher correlation to noise indicates a better ability to detect noise.

Both split and ensemble variance metrics yield an uncorrelated parity plot with small Spearman R correlation values as shown in Figure 3 e-f. In addition, their stacked bar distribution is uniform across all percentiles. Similarly, the forgotten events metric has a low correlation with noise added and does not identify high noise data particularly well at either extreme (Figure 3 d). This suggests that split variance, ensemble variance, and forgotten events are not suitable metrics for detecting noise. These findings are contrary to that of prior studies using forgotten events or ensembling as a noise detection metric (Gupta and Gupta, 2019; Nguyen et al., 2020; Toniato et al., 2021; Yuan et al., 2022; Kim et al., 2023). Our conflicting results may be attributed to differences in the model when performing regression rather than classification.

The training error metric correlates with data noise considerably more compared to the three literature metrics at the same noise scale (Figure 3 b). In addition, the stacked bar plot shows that 85% of the molecules in the highest 10% of training error have high noise. This suggests that the training error metric is suitable for noise detection. Additionally, the correlation becomes more pronounced when the dataset contains more noise as presented in Figure 3 a-c.

In addition to its strong noise detection capabilities, training error is simple and quick to calculate for each datapoint. This metric is thus extremely applicable in practice compared to more complex noise detection methods especially concerning ADMET data where datasets are large. Figures 3-6 show that the results are consistent across all four datasets tested.

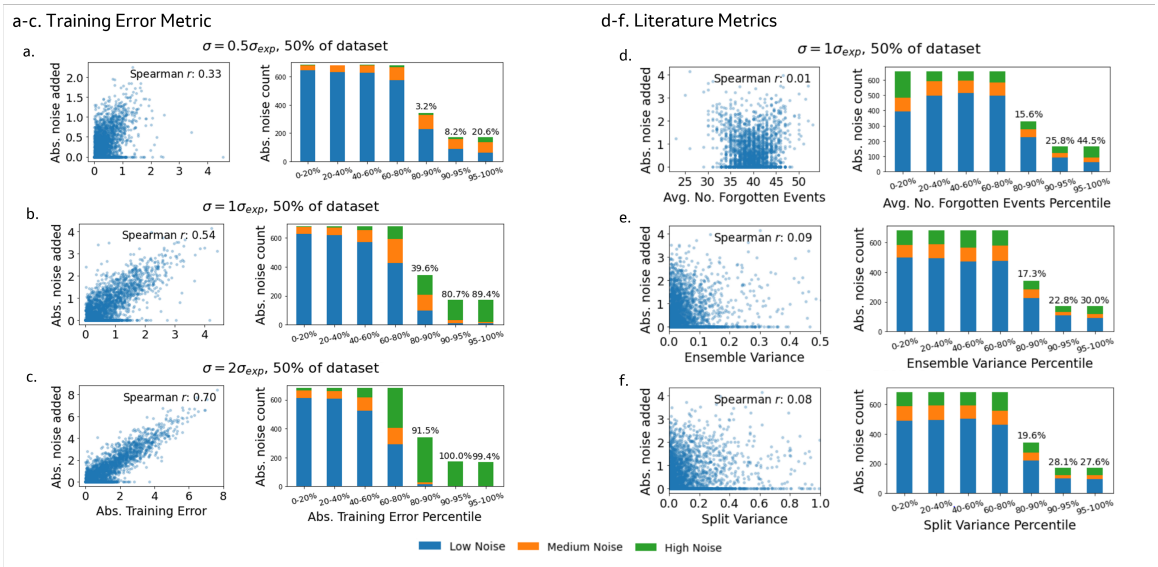

Figure 3: Correlation between the artificial noise and various noise detection metrics for the logD dataset. Low noise molecules are defined as those that have less than $0.5\sigma_{\text{expt}}$. Medium noise molecules are defined as those that have between $0.5\sigma_{\text{expt}}$. and $1\sigma_{\text{expt}}$.. High noise molecules are defined as those that have over $0.5\sigma_{\text{expt}}$. **a, b, c**: Absolute training error correlation to artificial noise with increasing amounts of noise added. **d, e, f**: Three literature noise detection metrics. Subfigures b and d-f have the same noise combination.

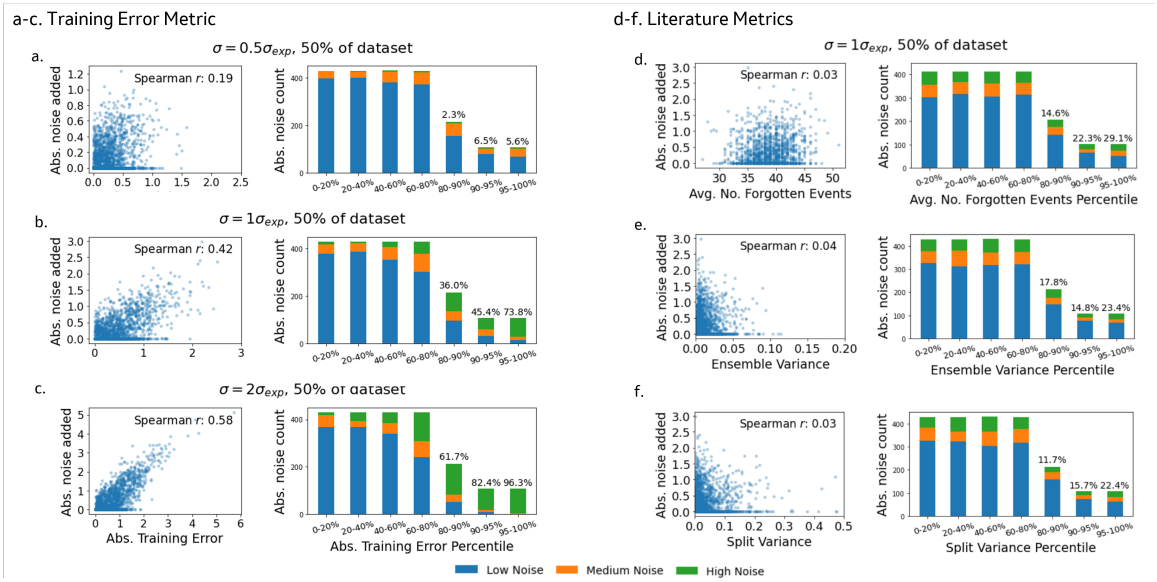

Figure 4: Correlation between the artificial noise and various noise detection metrics for the human fraction of unbound plasma ($F_{u,p}$) dataset.

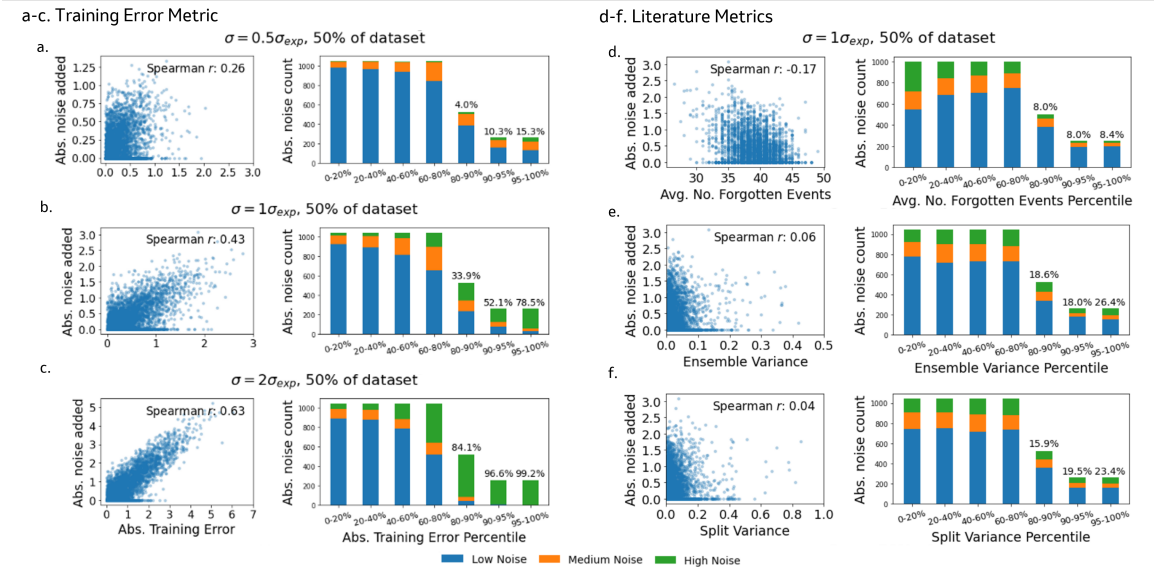

Figure 5: Correlation between the artificial noise and various noise detection metrics for the apparent permeability (P_{app}) dataset.

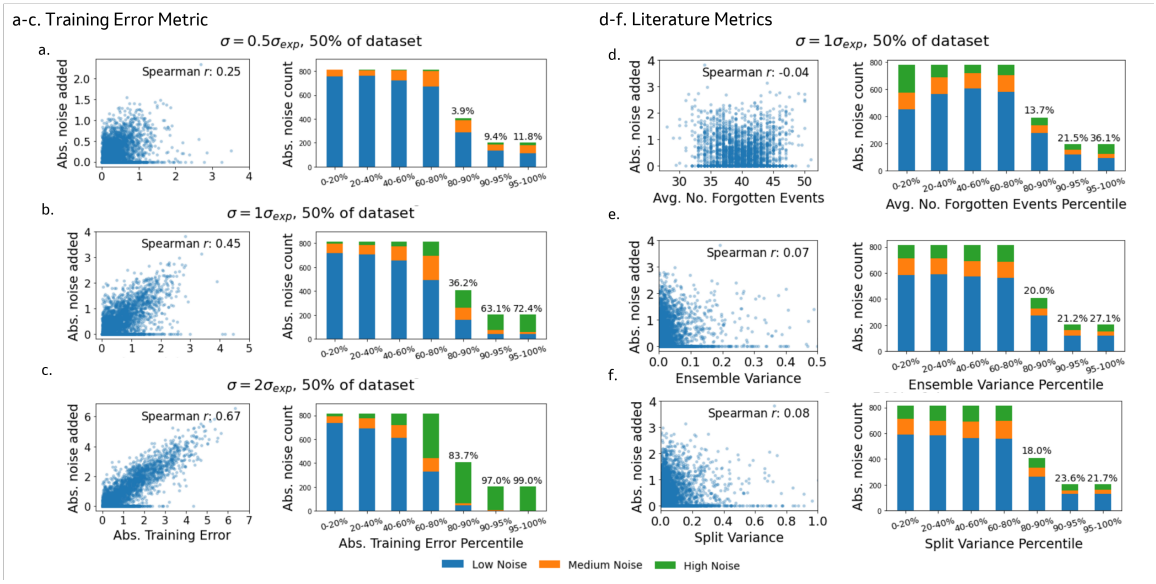

Figure 6: Correlation between the artificial noise and various noise detection metrics for the hERG binding dataset.

B.2 Visualizing difficult or scarce datapoints

Other studies using training error as a noise detection metric attempted to use more complex noise detection schemes in order to combat the mislabeling of noisy examples due to imbalance in the dataset (Li and Mao, 2023; Zhou et al., 2023). Additional analysis was conducted on our datasets to determine whether a more complex noise detection scheme would benefit the performance of our models or if simply using training error is sufficient. To determine underrepresented molecule types in the training set, we looked at the Tanimoto similarity of molecules, molecules with unique atom types, and molecules in sparsely filled clusters. The Tanimoto similarity of each molecule was calculated against each of the molecules in the rest of the training set and averaged across the set. A lower Tanimoto similarity suggests that it is more underrepresented as it is less similar to the training set. However, our findings show that there is no correlation between the Tanimoto similarity score and training error in all noise combinations across all datasets. This suggests that less similar molecules are not being filtered out at a higher rate.

Furthermore, there is often imbalance in the dataset when there are molecules with atom types that are less typical in drug-like molecules. In the original dataset, the molecules containing the less represented atom types in the dataset do not generally produce a higher error than other molecules which do not contain these atom types. Moreoever, these more unique molecules follow the general trend with respect to noise correlation and are not filtered out at a higher rate than molecules with higher representation, even when these molecules have low-noise. This further suggests that less represented molecules are not filtered at a higher rate.

Additionally, the molecules were k-means clustered by their Morgan/circular fingerprint, a typical vectorized representation used to determine the similarity between molecules. Similar results are observed as the molecules in smaller clusters were not filtered out at a higher rate and do not lead to a higher training error in the original data set. Based on the identical findings of each method to identify less represented tasks, it can be concluded that for our chemical datasets and model architecture, these tasks do not cause an inflation of training error in these samples. The GCNN model that we use is likely able to generalize better to these unique training tasks. Therefore, for ADMET predictors using the Chemprop architecture, training error itself is likely a sufficient noise detection metric as informative/less represented tasks are not more likely to labeled as noisy.

B.3 Denoising performance summary

The performances of each denoising scheme for each noise combination is reported in Figures 7-10. In most noise combinations, our methods (blue numbers) performed better on the test set compared to the other literature metric filter methods (gold numbers). This corroborates our earlier findings of training error as the best metric for noise identification and proves its utility in a denoising scheme.

Our models yield varying performance changes relative to the original non-denoised model in varying noise regions. We identify the low noise region as all combinations where $\sigma = 0.5 * \sigma_{\text{expt}}$, the medium noise region as all combinations where $\sigma = 1 * \sigma_{\text{expt}}$ or $2 * \sigma_{\text{expt}}$ and noise is added to 30% or 50% of the data, and the high noise region as all combinations where $\sigma = 1 * \sigma_{\text{expt}}$ or $2 * \sigma_{\text{expt}}$ and noise is added to 100% of the data. The results in these separated regions are empirically similar across all datasets.

The performance of our denoising schemes in the low noise region are similar to that of the un-denoised model as shown in Figures 7-10 b-d. As observed before, there is a weaker signal when using training error to detect noise when less noise is added to the dataset. There are thus less true noisy values being identified in these schemes.

In the high noise region, the TE Finetune Model did not have any significant performance decrease whereas the other models did in Figures 7-10 g,j. The base TE Filter Model is a naive approach to denoising. Simply removing data limits the total examples the algorithm can learn from. Because the remaining training set still has high noise and fewer examples after filtering, the model performance is worse as it is more likely to overfit to the high noise examples. The Mean Correction and Prediction Correction models are dependent on the accuracy of the base model. In the high noise region, the base model is much less accurate and is likely adding more noise into the dataset. The TE Finetuning Model has less loss of information due to pretraining with the unfiltered data set and does not add additional noise to the dataset explaining its superiority to the other models.

Our models have improved performance compared to the baseline un-denoised model in the medium noise region as presented in Figures 7-10 e-f, h-i. Similar to the other noise regions, the denoising scheme which has the best performance in the medium noise region using the 10% training error cutoff is the TE Finetune Model (Figures 7-10).

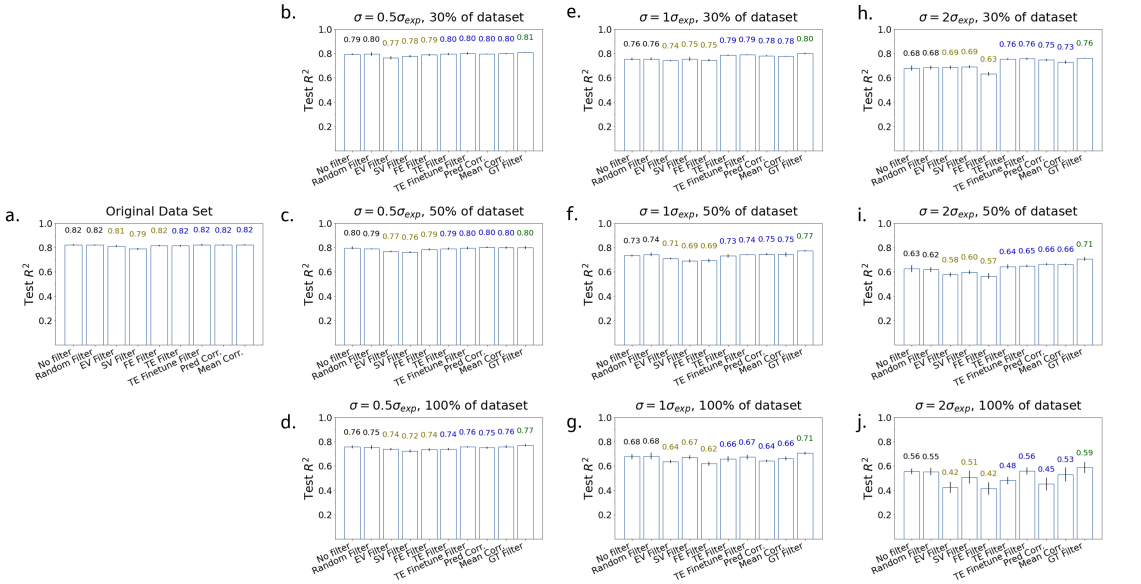

Figure 7: Results summary for all noise combinations tested and for all denoising schemes on the logD dataset. Black numbered columns represent our baselines, gold numbered columns represent denoising schemes using literature metrics, blue numbered columns represent our denoising schemes, and the green numbered column is the ground truth baseline. EV, SV, FE, TE, and GT stand for ensemble variance, split variance, forgotten events, training error, and ground truth, respectively. Each model is tested on a clean, held-out test set. Error bars are calculated using the standard deviation of the R^2 (coefficient of determination) for each of the four training ensembles.

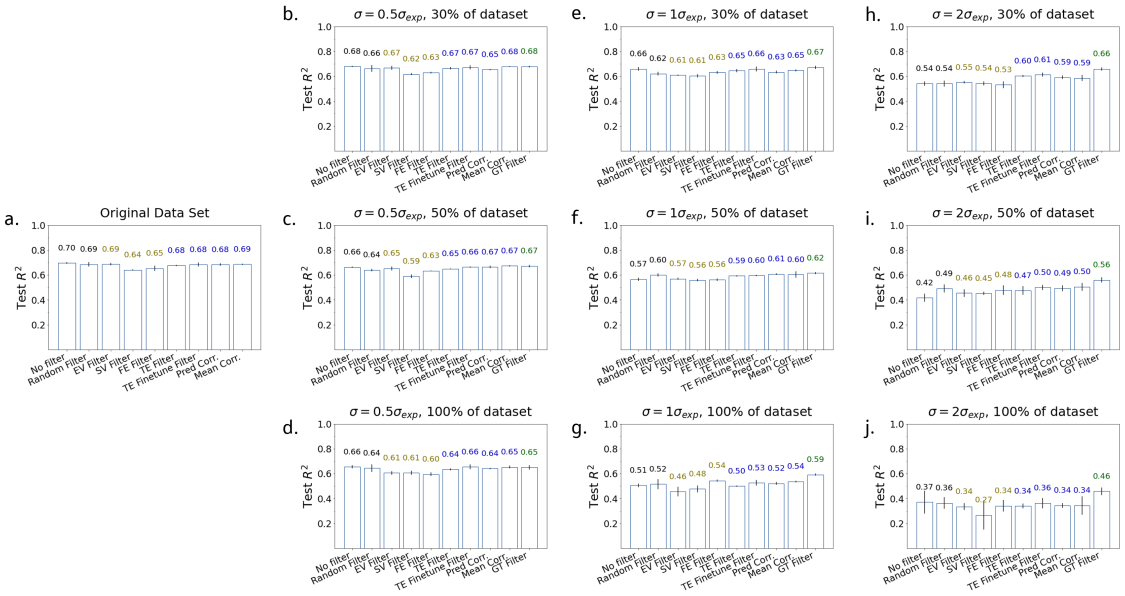

Figure 8: Results summary for all noise combinations tested and for all denoising schemes on the human fraction of unbound plasma ($F_{u,p}$) dataset.

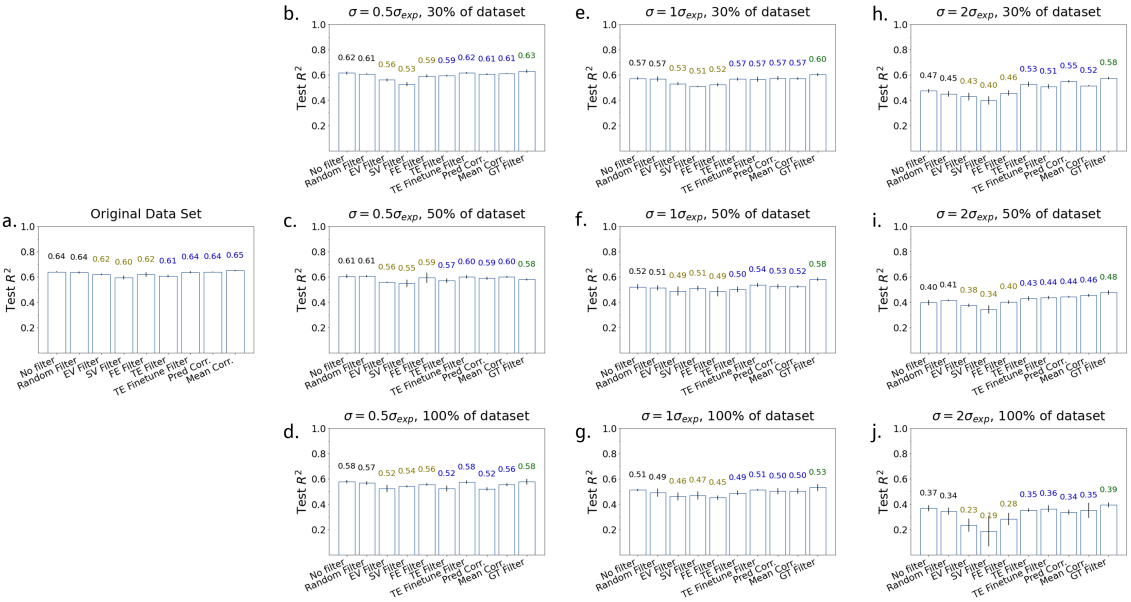

Figure 9: Results summary for all noise combinations tested and for all denoising schemes on the apparent permeability (P_{app}) dataset.

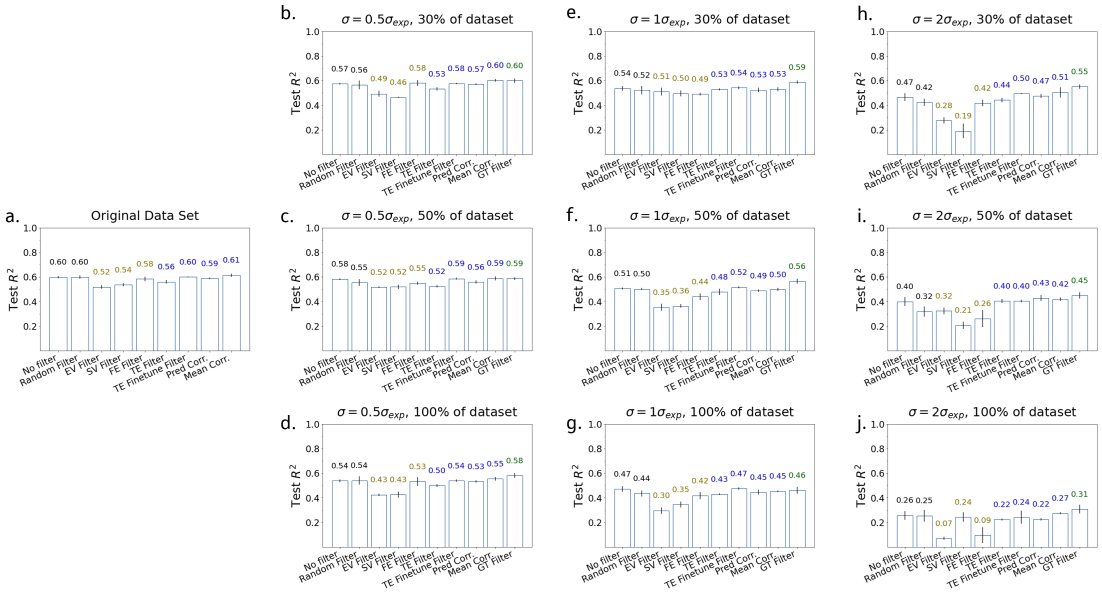

Figure 10: Results summary for all noise combinations tested and for all denoising schemes on the hERG binding dataset.

To give a final recommendation, we quantified the average performance change in each model over all datasets in each noise combination. Overall, the TE Finetune Model is the best performing denoising scheme for models using the 10% training error cutoff. Figure 11 shows a summary of the performance of each model relative to base, non-denoised model for all dataset types. This further reinforces that the TE Finetune Model does not exacerbate the predictive performance in the low and high noise regions and enhances the performance in the medium noise region. Comparing our model with the ground truth baseline further demonstrates its utility. The ground truth baseline improves the R^2 value by 0.03 – 0.09 on average in the medium noise region while our TE Finetune Model improves the R^2 by 0.01 – 0.05 on average. These improvements are on the same order which is impressive given that the ground truth gives a theoretical upper bound to the denoising methods. Additionally, finetuning is less computationally intensive and quicker relative to training a deep-learning model from scratch.

Furthermore, an adaptive threshold was used to detect noisy data rather than a simple 10% cutoff. This noise detection method was applied to the TE Finetuning denoising scheme. The adaptive threshold was based on the standard deviation of the training data that we are denoising. Thus, the number of samples that are filtered out changes based on the amount of noise in the data. This method gives a preliminary improvement in R^2 of 0.01 – 0.07 compared to the base model (Figure 11). We are planning to investigate this method further in hopes to attain further performance improvement.

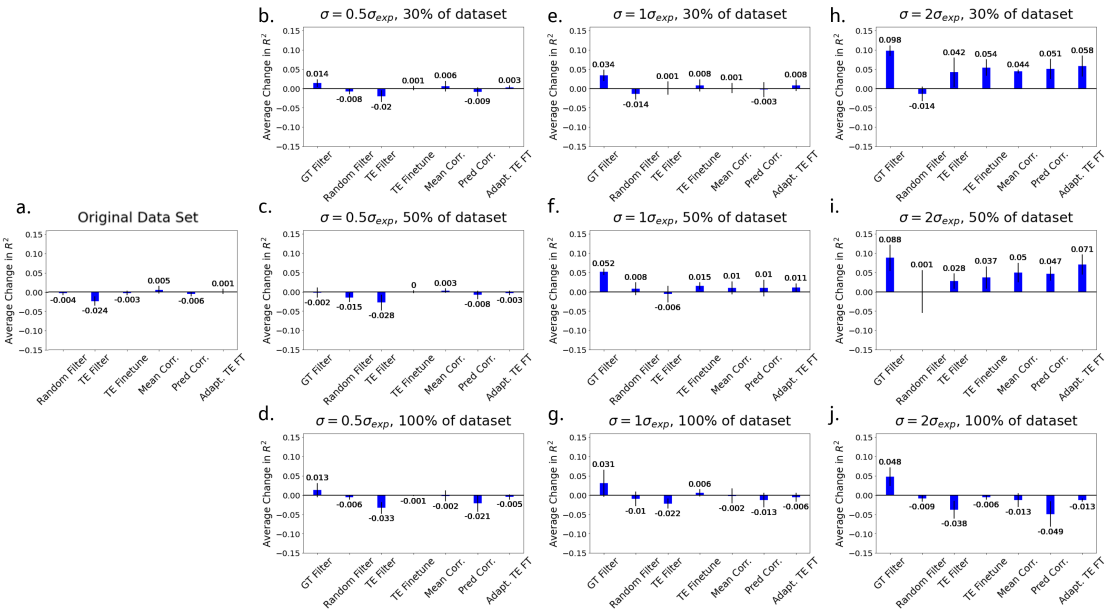

Figure 11: Performance of each denoising scheme relative to the un-denoised model across all four public datasets for each noise combination individually. Each bar is the average R^2 change over four ADMET assays tested. The Adapt. TE FT label stands for the adaptive training error finetuning model.