Denoising Drug Discovery Data for Improved ADMET Property Prediction

Matthew Adrian[1], Yunsie Chung[1*], Alan C. Cheng[1*]

[1]Modeling and Informatics, Merck & Co., Inc., South San Francisco, California 94080, United States

* Corresponding authors

Keywords

Denoising, noise filter, noise reduction, denoising for regression, ADMET prediction, drug discovery, deep learning

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. Furthermore, ADMET predictors are typically regression tasks due to the continuous nature of the data. This makes it difficult to apply existing methods as most focus on classification tasks. Here, we develop denoising schemes based on deep learning to address this. We find that the training error can be used to identify the noise in regression tasks while ensemble-based and forgotten event-based metrics fail to detect the noise. 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 method has the ability to improve models with medium noise and does not degrade the performance of models with noise outside this range. To our knowledge, our denoising scheme is the first to improve model performance 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.[1-3] In the early 2000's, it was found that about 50% of the attrition in drug candidates came from poor pharmacokinetics and toxicity profiles.[4] Due to advances in high-throughput screening (HTS) and machine learning, we are now able to more accurately predict the ADMET profile of drug candidates *in silico* during early stages of drug discovery to reduce the likelihood of late-stage attrition.[2,5] An ADMET predictor is also an integral tool in the lead optimization of candidate molecules as it can help make informed decisions and prioritize the most promising compound for synthesis.[6-8]

ADMET assays, in practice, have experimental errors even when using validated procedures.[9] Erroneous measurements stem from multiple factors including procedural changes between different sources or overtime,[9] calibration error of instruments,[10] impurity or degradation in starting material,[10] and human error. In addition, compounds with assay values beyond the

measurement limits can lead to incorrect values, resulting in a high peak near the measurement threshold in the data distribution. The error can be reduced or characterized by performing multiple measurements; however, in many cases this is impractical due to factors such as cost or time.[9] It is challenging to directly reduce or identify noisy measurements within ADMET datasets, especially those that are large. Studies have shown that these experimental errors worsen the predictive performance of ADMET models[11] and pose challenges in accurately evaluating the model's true performance.[12] It is thus crucial to devise a denoising scheme that can effectively reduce the noise in the ADMET data and recover the model performance.

Denoising research has been primarily focused on imaging datasets to recover image quality and improve image classification tasks with corrupted labels.[13-15] For other data types, including those related to health care, chemistry, and bioassays, denoising and noise detection have mainly been studied for class noise rather than regression noise.[16-19] For example, two works in 2000 and 2016 looked at denoising for improving the predictive performance of machine learning classifiers for rheumatic disease diagnosis and breast cancer diagnosis based on clinically-measured descriptors.[16, 17] We explore existing denoising studies in more detail in the related work section, below. ADMET predictors are typically regression-based due to the continuous nature of the data, making it difficult to apply existing classification-based denoising schemes. Denoising studies conducted on regression tasks are very limited, and those leveraging chemical or drug discovery datasets are even more scarce leaving the topic largely unexplored.

In this study, we survey several noise detection metrics, including ensembling, forgotten events, and training error, and devise deep learning based denoising schemes for ADMET assay data. Machine learning models are trained on several ADMET endpoints with artificial noise added and tested on a held-out set to evaluate the performance improvement after denoising. The results show that finetuning the model with the data denoised based on a deep learning model training metric, the training error, gives the best performance improvement. In addition, we present thorough analysis on the impact of data imbalance and dataset size on the denoising schemes. We also investigate how experimental errors in the test set affect the perceived model performance. Lastly, the effect of noise on multi-task models is examined to determine if the noise in one task propagates and affects performance on other tasks in the model. To our knowledge, this is the first study to present a denoising scheme for drug discovery ADMET data that improves predictive performance on regression tasks.

2. Related Work

2.1 Ensemble-Based Methods

Many studies use ensembling of multiple submodels as a way to detect and filter noisy labels. Nguyen et. al. used ensembles of both models and epochs in their denoising framework to progressively filter noisy labels in image classification tasks.[20] Yuan et. al. employed an ensemble of models trained on separate splits of data and filtered out data based on disagreement for a regression task related to glaucoma diagnosis.[21] Both studies demonstrated that the ensembling-based approach can identify noise and improve predictive performance of machine learning models. However, Heid et al. found that ensemble variance as a metric is a quantification of the prediction uncertainty based on model variance rather than model bias or error due to noise.[12] This is a concern when using ensemble-based metrics for noise-detection.

2.2 Forgetting Events

Catastrophic forgetting is a phenomenon where neural networks forget previously learnt information.[22-23] Toneva et. al defined a *forgetting event* as an event where an example is classified correctly at time t in neural network training but subsequently misclassified at time t' where $t' > t$.[24] This study showed that the samples in image classification datasets that exhibit a higher number of forgetting events tend to have noisy labels. However, in datasets with no noise, a higher number of forgetting events indicates that the sample may be more valuable to model performance. Toniato et. al. adopted this forgetting events strategy to both identify and denoise chemical reaction data for forward and retrosynthesis predictions.[25] They found an increase in model accuracy and confidence with a decrease in bias when using this denoising strategy.

2.3 Training Error

Li and Mao conducted a study on denoising regression tasks from publicly available datasets covering various types of data including finance, real estate, social media, and the environment.[26] For this, they used training error as a noise detection metric. To combat the inhomogeneity of example types in a dataset, they used an iterative adaptive threshold for filtering noisy samples. Recently, Zhou et al. employed scaled versions of training error such that they account for varying epistemic and aleatoric uncertainties in the dataset.[27] In a study from another group, training error was used to detect noise in bioactivity and toxicity datasets for both classification and regression tasks but the authors found that their filtering scheme did not improve, and was in fact detrimental to, prediction accuracy.[28] To our knowledge, the training error metric has been utilized much less than aforementioned metrics and has not been shown to be successful in denoising schemes for chemical data.

3. Methods

3.1 Datasets

Four ADMET assay datasets were collected from the literature[29-34] as listed in Table 1: logD, human $F_{u,p}$, P_{app}, and hERG binding. These assays were chosen because they are important endpoints that are commonly measured in small molecule drug discovery that span a fairly accurate and diverse range of achieved model predictive performances. Each of these assays has inherent experimental error within the dataset that has not been reported and is difficult to quantify. To address this, we additionally used two quantum chemical property tasks from the QM9 dataset,[35,36] namely HOMO-LUMO gap and H298, as clean data for this study. These are synthetic data computed using density functional theory (DFT) at the B3LYP/6-31G(2df,p) level of theory. As these quantum chemical data contain no experimental noise, they serve as great alternative references for validating our denoising approaches. Each public dataset was split randomly such that the training set contained 80% of the molecules and the other 20% of molecules were used as a held-out test set. Our methods were additionally tested on internal data collected from Merck & Co., Inc. (Rahway, NJ, USA) as shown in Table 2. We used seven key ADMET endpoints from our drug discovery programs with varying levels of experimental error. These endpoints include logD, human $F_{u,p}$, rat $F_{u,p}$, P_{app}, hERG binding, kinetic FaSSIF solubility, and kinetic solubility at pH 7. The internal ADMET datasets were split to create clean test sets in the following way: only

the datapoints with multiple measurements and whose standard deviation among multiple measurements is less than 0.2 times the standard deviation of the entire data distribution are placed in the test set. The rest of the data was used in the training set.

Table 1. Public datasets used in this study

Dataset	Data Count	Description	Data source
LogD	4190	Distribution coefficient between octanol and water at pH 7.4, $\log_{10}$ transformed	(29)
Fraction of unbound plasma in human (human $F_{u,p}$)	2717	Fraction unbound, $\log_{10}$ transformed	(30-31)
Apparent Permeability (P_{app})	6457	Apparent permeability in Caco-2 cells, in 10^{-6} cm/s, $\log_{10}$ transformed	(32-33)
hERG binding	5108	Binding affinity (IC50) to human hERG potassium (K+) channel, in nM, $\log_{10}$ transformed	(34)
QM9 HOMO-LUMO gap	133802	Synthetic dataset from MoleculeNet of organic compounds with up to 9 heavy atoms. Computed at the B3LYP/6-31G(2df,p) level. HOMO-LUMO gap in Hartree (duplicate SMILES deleted)	(35-36)
QM9 H298	133802	Synthetic dataset from MoleculeNet of organic compounds with up to 9 heavy atoms. Computed at the B3LYP/6-31G(2df,p) level. Enthalpy at 298 K in kcal/mol (duplicate SMILES deleted)	(35-36)

Table 2. Internal ADMET datasets used in this study, from Merck & Co., Inc. (Rahway, NJ, USA)

Dataset	Data Count	Description
LogD	608053	Lipophilicity measure at pH 7 where logD is the ratio of a compound's concentration between organic solvent (octanol) and water, $\log_{10}$ transformed
Fraction of unbound plasma in human (human $F_{u,p}$)	23369	Fraction of the unbound (free) drug in human plasma, $\log_{10}$ transformed
Fraction of unbound plasma in rat (rat $F_{u,p}$)	56572	Fraction of the unbound (free) drug in rat plasma, $\log_{10}$ transformed
Apparent Permeability (P_{app})	55807	Apparent permeability through a cell monolayer. In-house P_{app} measurements in LLC-PK1 and MDCKII cell lines in 10^{-6} cm/sec, $\log_{10}$ transformed
hERG MK499	370148	Binding to the HERG channel through the displacement of MK-499 in Molar (M), $-\log_{10}$ transformed
FaSSIF Solubility	307341	Kinetic (high-throughput) solubility in FaSSIF (Fasted State Simulated Intestinal Fluid) solution in Molar (M), $\log_{10}$ transformed
Solubility at pH 7 (SOLY 7)	454288	Kinetic (high-throughput) solubility at pH 7 in Molar (M), $\log_{10}$ transformed

3.2 Model Details

All ADMET models were built on a directed message passing neural network (D-MPNN) based architecture as implemented in Chemprop v1.6.1.[37,38] The model takes the SMILES strings of chemical compounds as input and generates graph-based structures of compounds with initial atom and bond features. The graph representations pass through a message passing neural network and convert into molecular latent representations, which are fed into a feed-forward neural network to provide property predictions. A single-task model was constructed for an individual dataset. Hyperparameter optimization was performed on the datasets, and an optimal set of hyperparameters that provides reasonable accuracy across multiple datasets was selected as listed in Table 3. Chemprop's default hyperparameters were used for model parameters not specified in the table. An ensemble of four submodels was used for all models discussed in this work unless otherwise noted. The ensemble was generated by using two different random training/validation set splits and subsequently using two different parameter initializations for each training/validation split, resulting in a total of four submodels. The average of the ensemble predictions was used to assess the model performance on a held-out test set. Unless further specified, the error bars in each figure correspond to the standard deviation among the four submodels of the stated metric in the figure.

Table 3. Chemprop model parameters

Hyperparameter	Value
MPN Depth	3 (4 for internal)
MPN hidden size	600
FFN number of layers	3 (4 for internal)
FFN hidden size	1200 (1300 for internal)
Dropout	0
Aggregation	Norm
Number of folds (training/validation split seed)	2
Ensemble size (parameter initialization seed)	2
Epochs	80 (60 for internal; 15 for finetuning)

3.3 Adding artificial noise

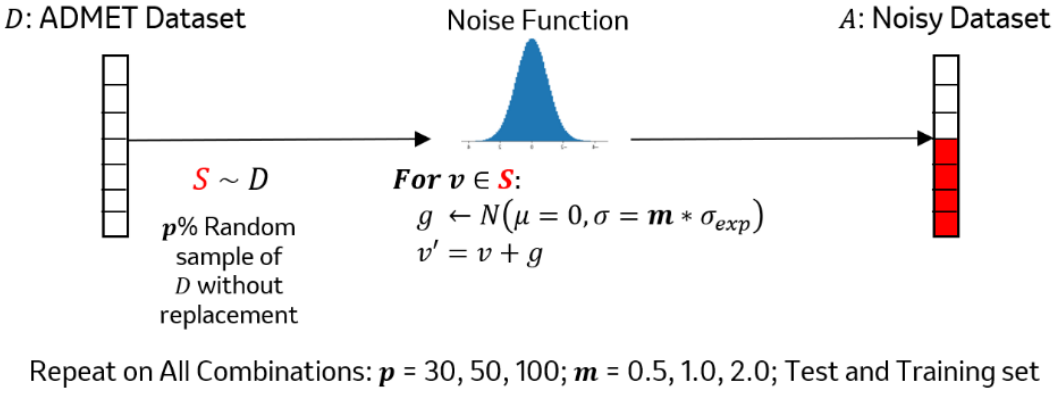

Figure 1. Overview of noisy dataset creation. This creates nine unique noisy training and test sets for each dataset.

Artificial noise was added on the aforementioned datasets as described in Figure 1. For the purpose 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 added artificial noise. We randomly sampled a subset of the data from the original dataset with a percentage $p \in \{30, 50, 100\}$. Noise was added to each datapoint of this sampled subset. For each datapoint, the value of noise, g, was determined by randomly sampling a gaussian distribution with magnitude scaling factor $m \in \{0.5, 1.0, 2.0\}$ as shown in Equation 1.

$$g \leftarrow Normal(\mu = 0, \sigma = m * \sigma_{exp}) \quad (1)$$

The standard deviation of the full original dataset, σ_{exp}, normalizes the magnitude of noise added to each unique dataset. The datapoints with the artificial noise were then combined with the remaining datapoints without noise to generate a noisy dataset. Combining all sampled percentages and magnitudes resulted in 10 different noise combinations for each dataset, including the original data set which has no artificial noise added. This procedure is done both on the training and testing sets separately. The performance of the models was evaluated on a 20% held-out test set both with

and without artificial noise. The training set was randomly split into around 90% training and 10% validation sets with different split seeds as described in Section 3.2.

3.4 Noise Detection Methods

Four noise detection metrics were applied: (1) Training error, (2) number of forgotten events, (3) ensemble variance, and (4) split variance. The procedure below outlines each noise detection metric.

3.4.1 Training Error (TE)

A model was trained on the training set. This model was used to predict values of the molecules in the same training set. The absolute training error was calculated on a per molecule basis.

3.4.2 Number of Forgotten Events (FE)

A forgotten event at epoch *n* occurs when the training error of epoch *n* is greater than the training error of epoch *n* – 1 (Equation 2).

$$\mathrm{FE} = \begin{cases} 1 \text{ if Training Error}_{\text{Epocb } n} > \text{ Training Error}_{\text{Epocb } n-1} \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

The number of forgotten events is calculated for an individual datapoint by the summation of forgotten events across all epochs used for training.

3.4.3 Ensemble Variance (EV)

Each chemprop model was trained using an ensemble of four submodels as described in Section 3.2. The variance among the four model predictions was calculated on a per molecule basis.

3.4.4 Split Variance (SV)

Three chemprop models were trained on three non-overlapping splits of the training set. Each model gave predictions on the entire training set. The variance among the three model predictions was calculated on a per molecule basis.

3.5 Denoising schemes

Four main denoising schemes were tested as visualized in Figure 2**.** The TE Filter Model filters out the top 10% molecules with the highest training error. The remaining 90% filtered dataset was then used to train a new chemprop model. The TE Finetune Model uses the same filter as the TE Filter Model, however, the final model comes from finetuning the original model on the filtered dataset rather than training a new model from scratch. In this approach, the model parameters from the original model trained on the entire dataset were used to initialize the second model, which was subsequently fine-tuned on the filtered dataset. The Mean Correction Model differs from the two filter models as it replaces corresponding values of the top 10% molecules rather than filtering.

This model corrects by taking the mean of the predicted and original values. Similarly, the Prediction Correction Model replaces these values with just the prediction. 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.

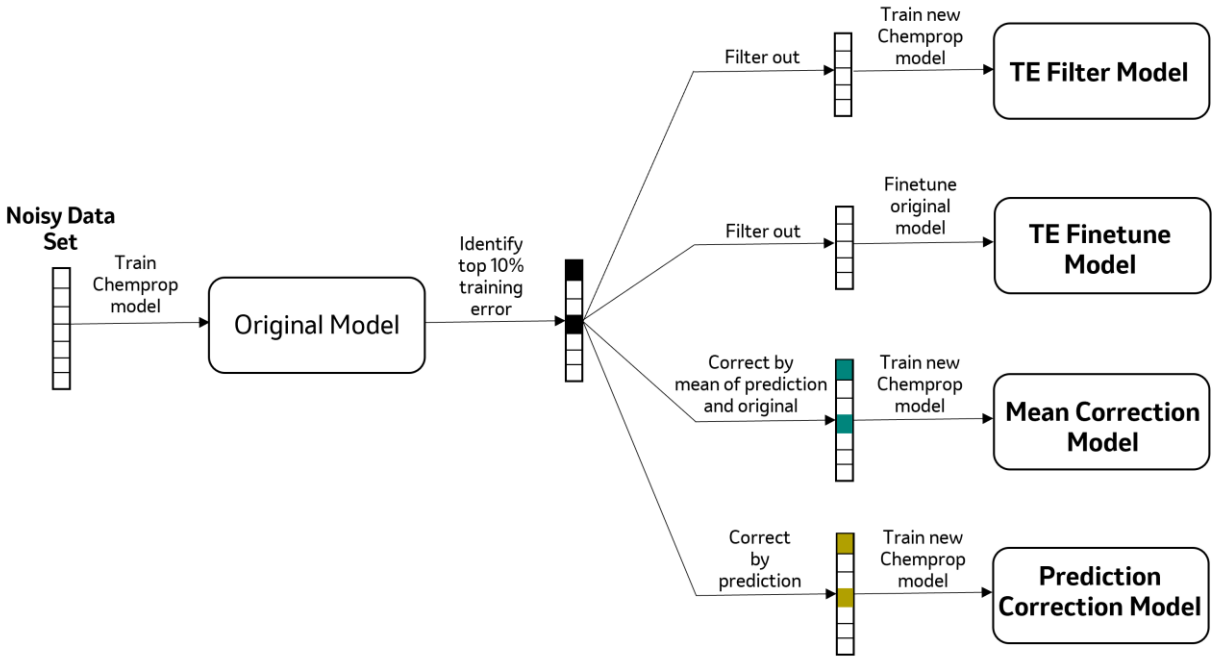

Figure 2. Overview of the four denoising schemes tested. In this visualization, absolute training error is being used as a metric to detect noise. The noise detection metric is interchangeable for each denoising scheme.

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

The denoising schemes were evaluated on the same test set and compared using the coefficient of determination (R^2) and mean absolute error (MAE) as the main performance metrics. It should be noted that the data filter or correction was performed only on the training set, and the test set remains intact and untouched throughout the analysis.

4. Results and Discussion

4.1 Noise Detection

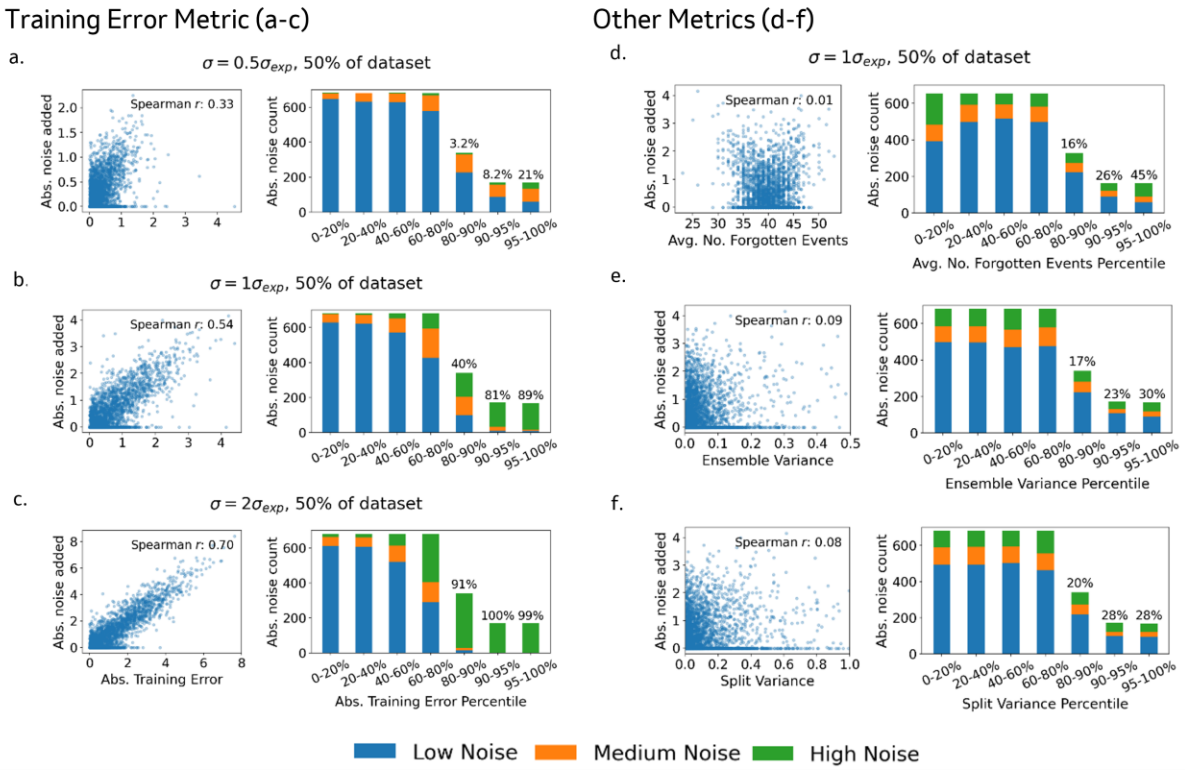

Figure 3. Correlation with noise for each noise detection metric. The results presented are from the public logD dataset. (a-c): Absolute training error correlation to artificial noise with increasing amounts of noise added. (d-f): Other three noise detection metrics. Subfigures b and d-f have the same noise combination. We define low noise data points as those with artificial noise less than $0.5 * \sigma_{exp}$, medium noise data points as those with artificial noise between $0.5 * \sigma_{exp}$ to $1.0 * \sigma_{exp}$, and high noise data points as those with artificial noise above $1.0 * \sigma_{exp}$.

To evaluate the suitability of each noise detection metric, we compared the Spearman rank correlation coefficient *r* between the chosen metric and the absolute artificial noise added. A higher absolute value of the correlation coefficient indicates a better ability to detect noise. The results on the public logD dataset are displayed in Figure 3, and the corresponding figures on the other ADMET datasets can be found in the Supporting Information Section S1.

Both split and ensemble variance metrics yield an uncorrelated scatter plot and small Spearman *r* correlation values as shown in Figure 3e-f. In addition, their stacked bar plots of the noisy distribution are uniform across all percentiles of their corresponding error metrics. Similarly, the forgotten events metric has a low correlation with the added noise and does not identify high noise data particularly well at either extreme (Figure 3d). This suggests that split variance, ensemble variance, and forgotten events are not suitable metrics for detecting noise. These findings are contrary to findings from prior studies using forgotten events or ensembling as a noise detection

metric,[14,18,20-21,24-25] and this could 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 other metrics at the same noise scale (Figure 3b). The correlation becomes more pronounced when the dataset contains more noise (Figure 3a-c). 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.

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 when noisy data needs to be quickly identified for hundreds of thousands of compounds for ADMET datasets. These results are observed across all datasets (see Supporting Information Section S1).

4.2 Effects of Noise in Test Sets

Prior studies indicated that the perceived performance of a predictive model is heavily dampened by a noisy test set.[12,39] When testing on noisy data, metrics will show how well the model predicts on values with error rather than the population mean of the measurement.

The same phenomena is observed in our results as shown in Figure 4. In every noise combination, the true performance of each model on the clean test set (blue bar) is much higher than the perceived performance on the noisy test set (orange bar). Overall, the un-denoised model (denoted as "No Filter" in Figure 4) demonstrates robustness to small magnitude of noise ($0.5\sigma_{exp}$) when evaluated on the clean test set while the perceived performance underestimates the robustness with R^2 decrease of 0.1 or more. The deviation between the true and perceived performances becomes more pronounced when the magnitude of noise becomes high ($2\sigma_{exp}$). On the clean test set, the performance of the un-denoised model drops by R^2 of 0.14 to 0.26 with high amounts of noise added whereas on the noisy test set, the drop is more substantial, ranging from R^2 decrease of 0.5 to 0.7. Additionally, the perceived increase in performance from our denoising schemes is heavily dampened (Figure 4i). These trends are consistent across all ADMET assays tested (Supporting Information Section S2).

This highlights the necessity of utilizing clean test sets when assessing the accuracy of ADMET models and comparing the change in performance between different models. This is especially important when determining the efficacy of a denoising scheme as it is difficult to tell the true performance improvement after denoising when the test set has noise. Thus, in this study, we compare the performance of the models on the clean test set without artificial noise to determine the true performances of the models.

4.3 Denoising Scheme Performance on ADMET Data

The performances of each denoising scheme for each noise combination is reported in Figure 4. In most noise combinations, the methods using the TE metric (blue numbers) perform better on the test set compared to the other metric filter methods (gold numbers). This corroborates our earlier findings that identify training error as the best metric for noise identification and proves its utility in a denoising scheme.

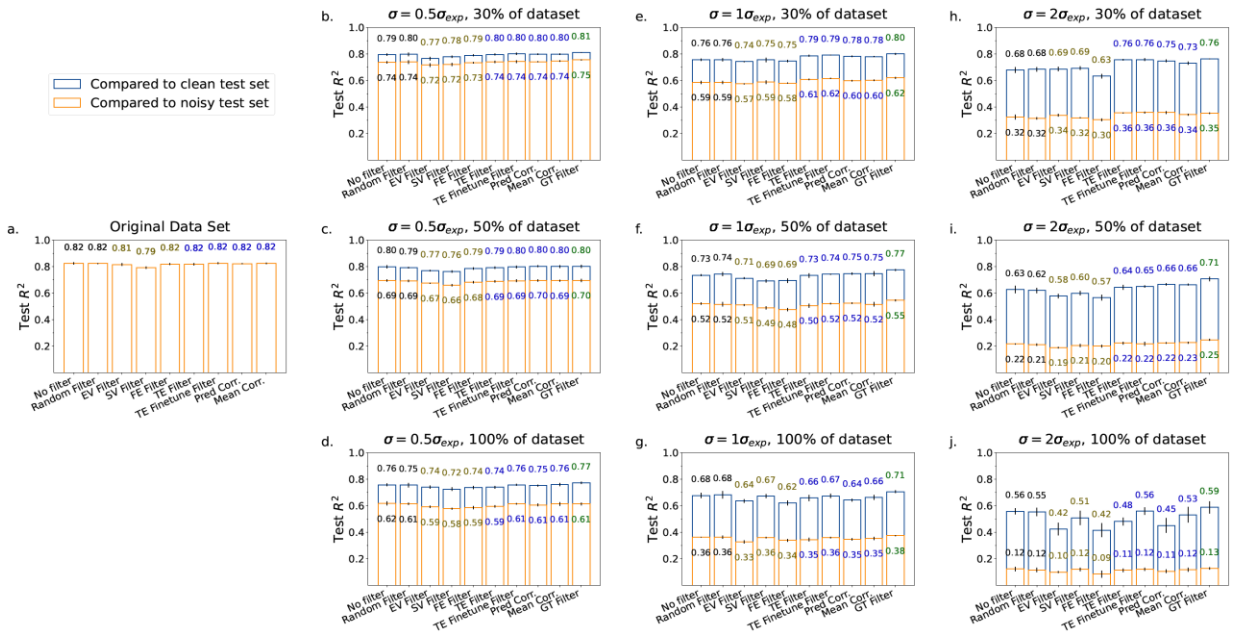

Figure 4. (a-j): Results summary for all noise combinations and all denoising schemes on the public logD dataset. Black numbered columns represent the baselines, gold numbered columns represent the denoising schemes using EV, SV, and FE, blue numbered columns represent the denoising schemes using TE, 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. The clean test set is the original test set with no noise added while the noisy test set has the same combination of noise added as the training set, denoted in the title of each figure. Error bars are calculated using the standard deviation of the R^2 among the four ensembles.

Our models yield varying performance changes relative to the original un-denoised model across different noise regions. For analysis throughout the paper, we identify the "low noise region" as all combinations where $\sigma = 0.5 * \sigma_{exp}$ and noise is added to 30% or 50% of the data, the "medium noise region" as all combinations where $\sigma = 1 * \sigma_{exp}$ or $2 * \sigma_{exp}$ and noise is added to 30% or 50% of the data, and the "high noise region" as all combinations where noise is added to 100% of the data. The noise combinations are categorized in these three separate regions because similar results are observed empirically within each region for all public datasets.

In the original dataset with no artificial noise added, it appears that our denoising methods yield the same performance as the un-denoised model (Figure 4a). However, as discussed earlier, this may be because the test set has inherent noise which dampens the perceived performance increase of our models.

The performance of our denoising schemes in the low noise region are similar to that of the un-denoised model (Figure 4b-c). As observed previously in Figure 3, there is a weaker signal between training error and added noise when the dataset has smaller/less noise. The true noisy values are thus identified less by the proposed denoising schemes in this region. In addition, because the magnitude of noise is low, the inherent noise in the original dataset, which was assumed to be clean, could have a larger effect on the perceived performance of the models, resulting in an underestimated true performance increase. However, considering that the performance of the original model only decreases by 0.02-0.03 R^2 when the noise is added for the low noise region

($\sigma = 0.5 * \sigma_{exp}$ to 30%, 50% data), it is also possible that the room for model improvement is marginal for these cases. In fact, the ground truth (GT) filter yields only 0.01 or no R^2 enhancement for the low noise region, which indicates that the benefit of denoising the training set is minimal for the low/less noise combinations. In the high noise region, the TE Finetune Model does not display any significant performance decrease whereas the other models have worse performance than the original model without denoising (No filter) as shown in Figure 4c,g,j. The 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 heavily dependent on the accuracy of the un-denoised model. In the high noise region, the un-denoised model is less accurate and likely adds more noise into the dataset when being used to correct data values. In contrast, the TE Finetuning Model has less loss of information due to pretraining with the unfiltered dataset and does not introduce additional noise to the dataset, explaining its superiority to the other denoising approaches.

Our methods increase the performance of the models in the medium noise region (Figure 4e-f, h-i). The TE Finetune method and the two correction methods are the best performing denoising schemes in the medium noise region. Very similar results are seen with all public ADMET datasets we tested (Supporting Information Section S2).

Overall, the TE Finetune Model is the best performing denoising scheme for models using the 10% training error cutoff. Figure 5 shows a summary of the performance of each model relative to un-denoised model averaged over all four ADMET assays. 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 – 0.1 on average in the medium noise region while our TE Finetune Model improves the R^2 by 0.01 – 0.05 on average (Figures 5e, 5f, 5h, 5i). These improvements are on the same order, which is impressive given that the ground truth marks a theoretical upper bound to the denoising methods. Furthermore, it is evident that the performance increase of the TE Finetune denoising scheme is not due to randomness. Figure 5 shows that in most noise cases, the random filter baseline deteriorates the performance of the model on average while our TE Finetune denoising method performs significantly better than the random baseline. Additionally, finetuning is less computationally intensive and quicker relative to training a new deep learning model from scratch. These findings support and extend a previous study which used training error to denoise chemical datasets.[28] This study found that simply dropping the noisy samples labeled by training error deteriorated the performance of the model due to overfitting to the remaining, smaller training set. However, this was tested where noise was added to 100% of the data and with relatively smaller datasets (most of them with less than 1000 data points). For similar cases (high noise, p=100%), our TE Filter method yielded similar results to their study, leading to lower model accuracy. On the contrary, we find that the TE Filter can improve the model performance in different noise combinations such as the medium noise cases displayed in Figures 5 e,f,h,i. A more extensive set of noise combinations and denoising schemes is investigated in our study, which demonstrates that regression-based ADMET data can be denoised for model improvement, especially when using the TE Finetune method.

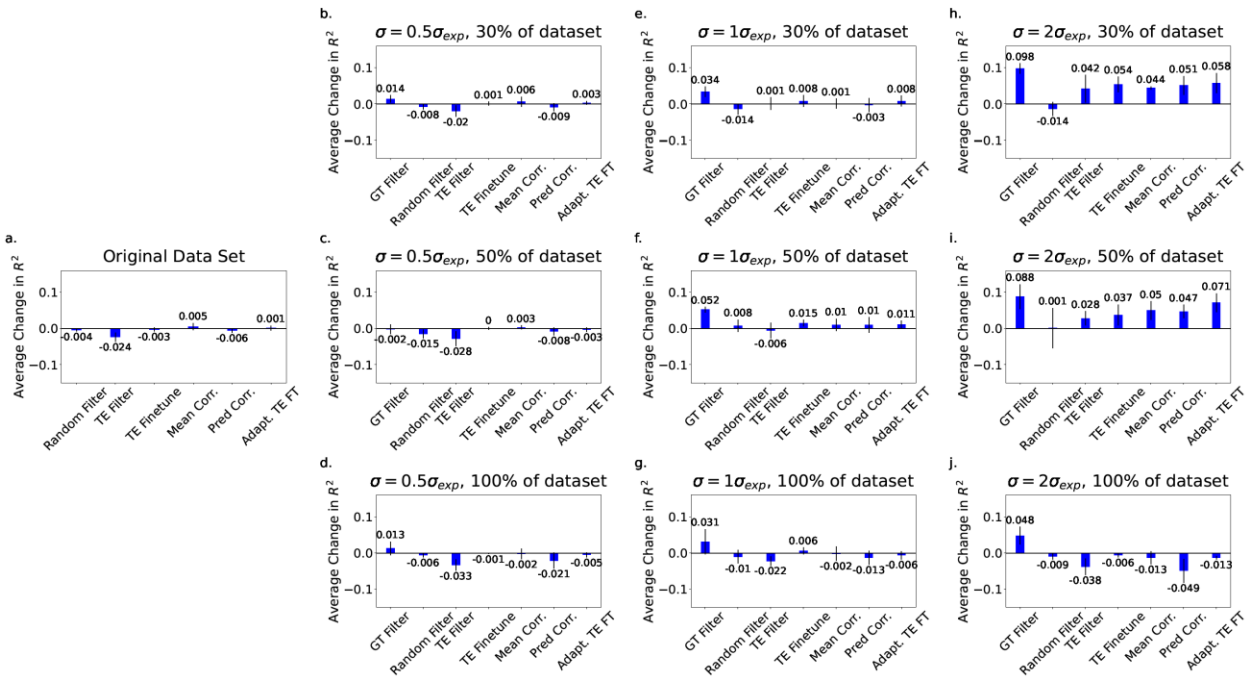

Figure 5. (a-j): Performance of each denoising scheme relative to the un-denoised model across all four public datasets for each noise combination. Each bar corresponds to the average R^2 change over four ADMET assays tested. "Adapt. TE FT" indicates the adaptive training error finetuning model.

4.4 Determining an Adaptive Threshold for Denoising

The TE Finetuning scheme is further tested using various threshold values to refine the amount of data removed in the finetuning step. So far, our denoising schemes have dropped or corrected a fixed 10% of the training data for all noise combinations. However, this approach is inefficient when the dataset contains much more or less noisy samples than the 10%. We therefore tested several thresholds that adaptively filter data depending on the amount of noise present in the dataset. From the analysis, it is found that using one standard deviation of the entire training data distribution (σ_{tr}) as the threshold typically optimizes the performance of the TE Finetuning method (refer to Supporting Information Section S3). This threshold naturally filters out more or less data for higher or lower noise cases, respectively.

The TE Finetuning denoising scheme is tested on the same public ADMET datasets with the chosen threshold value of σ_{tr}. This approach yields up to 0.07 R^2 performance increase on average in the medium noise range compared to the un-denoised model, performing better than simply finetuning the model with 10% of the data removed ("Adapt. TE FT" in Figure 5). However, in the high noise cases, too much data are removed and as a result the model performance deteriorates and underperforms compared to the simple filtering of the 10% data. Although high noise combinations show slight deterioration in performance, this may not be an issue as these noise ranges in which noise is added to 100% data are not typically practical for real ADMET datasets. Figure S7 is a figure corresponding to Figure 5 comparing the TE Finetune methods with the un-denoised model using mean-absolute error (MAE) as an evaluation metric, where similar results

are observed. We further evaluate the utility of the adaptive threshold on our internal datasets as discussed in Section 4.8.

4.5 Denoising Scheme Performance on QM9 Data

In order to benchmark our findings on truly clean data, we applied the TE Finetuning method using the 10% training error filter on two tasks in the QM9 dataset: (1) HOMO-LUMO gap and (2) H298. Since the QM9 dataset is formed from calculated data based on DFT, it can be assumed to be free of noise. The results on these tasks are presented in the Supporting Information Section S4. We observed no significant change in performance across all noise combinations when the TE Finetune method is used to denoise the dataset (Figures S12 and S13). Also, the effect of artificial noise was found to be minimal for the QM9 dataset as the R^2 decreases by only 0.006 - 0.05 even for the highest noise cases. It appears that predicting these quantum chemical properties is a relatively easy task for the model, and the un-denoised models are extremely robust to noise, achieving R^2 of 0.92 - 0.99 for all noise cases. Since the model is already robust to noise, our denoising scheme has little impact on the model performance.

Generally, it is harder for the model to learn generalizable information when the training set size is smaller. Due to this, we postulated that decreasing the dataset size would make the models less robust to noise leading to more room for model improvement by denoising when artificial noise is added. This loss in robustness is discussed further in our internal datasets in Section 4.7. To test this, we created randomly sampled training sets from size 500 to 10000 where each smaller dataset is a subset of the larger ones. Five folds of random training sets for each size were used, which reduces uncertainty due to randomness especially for smaller data sets. The average performance of models trained on each of the five folds separately for each size is reported in Figure 6 for the noise combination where noise is added to 30% of the data at $\sigma = 0.5 * \sigma_{exp}$. We used the MAE metric to enable comparison of prediction performance in terms of experimental units.

For smaller training sets, we observed MAE decreases of up to 3.24 kcal/mol and 1.03 kcal/mol in the H298 and HOMO-LUMO gap datasets, respectively, with the TE Finetune scheme. In the H298 dataset, the MAE change increases as the data size becomes smaller up until data size 2000 where the MAE decrease is maximized. At smaller data sizes, the model performance worsens, which likely causes the training error metric to be less accurate. At larger data sizes, the model becomes more robust to noise, reaching the point where denoising has little effect on the performance. This differs in the HOMO-LUMO gap dataset as the MAE improvement is maximal at data size 1000. The data size which the model has the most improvement from our method varies depending on the endpoint because each task has varying learning difficulties. In summary, the TE Finetune method increases the performance of the models built on data with purely artificial noise when the models are less robust to noise. Similar results are observed in the other noise combinations, especially in the medium noise region where performance increase was observed in the public ADMET data, as can be seen in Figures S14 and S15. Figures S16 and S17 show the corresponding results using the R^2 evaluation metric.

Additionally, it is interesting to note that the perceived performance of the model is substantially underestimated compared to the true performance (see Figures S12 and S13). When the noisy test set is used, the R^2 drops to 0.17 - 0.19 for the highest noise case while the true performance on the clean test set remains around 0.92 - 0.99. This finding further emphasizes the necessity of employing a clean test set for precise model assessment.

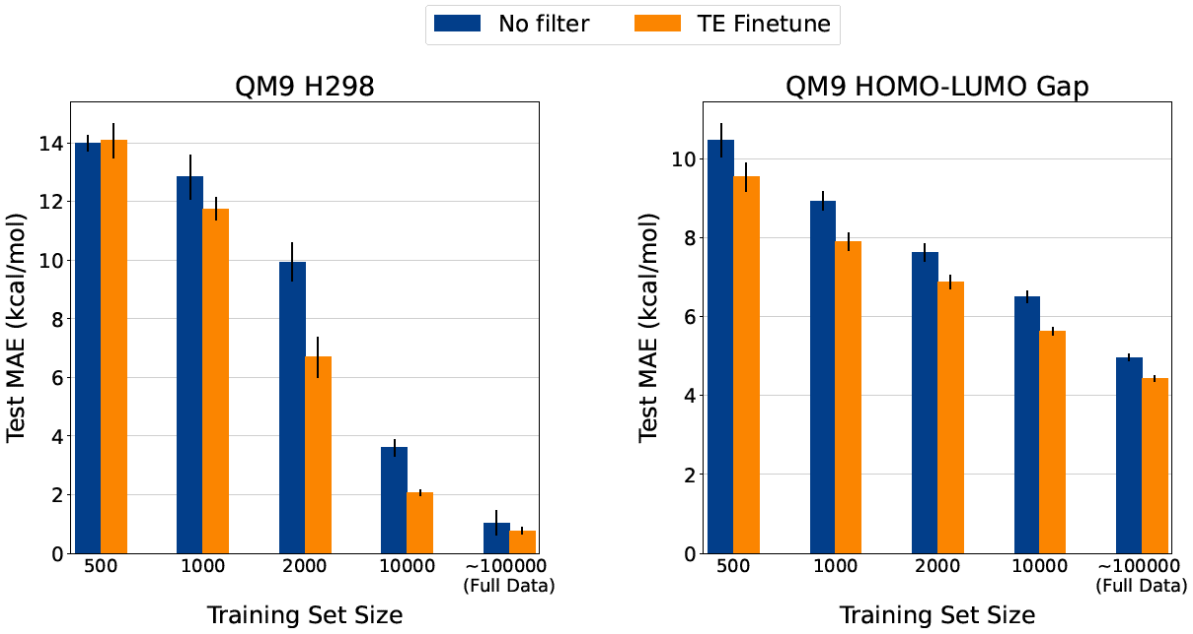

Figure 6. Performance of the TE finetuning denoising scheme compared to the base un-denoised model at varying training set sizes. Results shown are for the noise combination where noise is added to 30% of the data at $\sigma = 0.5 * \sigma_{exp}$. The ~100000 bar cluster refers to the models trained on the full training set, which did not use five folds of random training sets. Error bars in all other data sizes represent the standard deviation among the five folds. In the full data, the error bars are the standard deviation among the four submodels as mentioned in Section 3.2. Results shown are tested on the same clean, held-out test set for each task. HOMO-LUMO gap MAE was converted to kcal/mol from Hartree for better comparison.

4.6 Effect of Sample Imbalance and Prediction Difficulty on Training Error Metric

Although the majority of the compounds that are identified as noisy samples by the TE metric are indeed the compounds with high added noise, some clean samples with low or no added noise have high TE and are mislabeled as noisy, particularly in low noise regimes (Figure 7, Figures S22 – S24). Such mislabeling causes our denoising scheme to underperform compared to the Ground Truth model, and therefore we investigated several potential approaches to avoid filtering out clean samples. Other studies using training error as a noise detection metric attempted to use more complex schemes in order to mitigate the mislabeling of clean samples as noisy. In one study, a filter was designed that varied the filtering threshold with respect to a sample's k-nearest neighbors (kNN) to account for varying densities of similar sample types and varying fitting difficulties in datasets.[26] This was applied iteratively to label samples as potentially noisy initially, gradually increasing the confidence that a sample is truly noisy over iterations. A final threshold filter is created for each iteration based on the mean squared error of the regressor. Additionally, Zhou et al. claimed that samples with high uncertainties are more likely to be mislabeled as noisy when using the training error metric.[27] Prediction uncertainty inherently stems from data scarcity and randomness in the relationship between the covariates and target values, making targets more difficult to predict.[27] To combat this, they used a noise detection metric referred to as a veracity score which scales the training error by the prediction uncertainty.

Based on these studies, additional analysis was conducted on ADMET datasets to determine if a more complex noise detection scheme would improve the performance of our models or if training error itself is sufficient. Namely, we investigated the effects of underrepresented samples and difficult tasks on the training error metric. To identify underrepresented samples and difficult tasks in the training set, we looked at the following measures: (1) Tanimoto similarity of molecules, (2) molecules with unique atom types, (3) molecules in sparsely filled clusters, (4) activity cliff (AC) molecules, and (5) prediction uncertainty. We used ensemble variance to estimate prediction uncertainty in samples.

The Tanimoto similarity of each molecule was calculated against all other molecules in the training set and averaged across the set.[40] A lower Tanimoto similarity suggests the molecule is underrepresented as it is less similar to the training set. We observed that there is no correlation between the Tanimoto similarity score and training error in all noise combinations across all datasets (Supporting Information S5.1). This suggests that less similar molecules are not being mislabeled as high noise when using training error as the noise detection metric.

Furthermore, there can be imbalance in the dataset when there are molecules containing atom types that are less typical in drug-like molecules. In the public ADMET datasets used in this study, the molecules containing Br, I, P, B, and Si are relatively scarce compared to other atom types. Yet, we found that the molecules containing the underrepresented atom types generally do not produce higher training errors than other molecules as shown in Figure 7. This finding is consistent across all public ADMET assays (Supporting Information Section S5.2). Moreover, these underrepresented molecules follow the general trend of the correlation between the noise and training error and are not filtered out at a higher rate than the molecules with more highly represented atom types. This further indicates that less represented molecules are not being mislabeled as noisy.

Additionally, the molecules were k-means clustered by a PCA-reduced form of their Morgan/circular fingerprint, a typical vectorized representation used to determine the similarity between molecules.[41] Similar results are observed as the molecules in the smallest clusters are not being filtered out at a higher rate and do not lead to higher training errors (see Supporting Information Section S5.3). Additionally, the molecules that have no added noise but high training error were visualized throughout the clusters in the 2-dimensional PCA map (Figures S29 – S32) to investigate whether the compounds in certain clusters or locations are mislabeled at a higher rate. However, no apparent trend was observed for these molecules within the clusters.

We also adopted the method developed by Walter et al.[42] to determine whether the compounds that are falsely identified as noisy by the training error metric are also an activity cliff (AC) compound. AC compounds are those that have vastly different activity values compared to other chemically similar compounds. The properties of AC compounds have been reported to be difficult to predict when using machine learning models.[11,42,43] Walter et el. defined the structure activity landscape index (SALI) as shown in Equation 3 and labeled a compound as AC if the median SALI exceeds 1.[42]

$$SALI = \frac{\Delta activity}{1 - Tanimoto} \quad (3)$$

Since the SALI value is dependent on the activity/assay value for the compound and its neighbors, we visualized the AC compounds identified using both clean and noisy activities as presented in Supporting Information Section S5.4. In the following discussion, “clean AC”

compounds refer to the AC compounds with SALI values calculated using data with no artificial noise and "noisy AC" compounds refer to the AC compounds with SALI values calculated using data with artificial noise. In the P_{app} dataset, we observed that when the noise is low, the "clean AC" compounds with the top 10% training error were frequently mislabeled as noisy samples (Figure S35). This observed trend is not very apparent in the other datasets tested (Figures S33, S34, S36). Though this shows the possibility to detect mislabeled samples in specific cases, we are not able to attain the "clean AC" labels when dealing with noisy data. In a realistic setting in which the assay data have nonnegligible noise, only the "noisy AC" labeled compounds can be used in a denoising scheme as the true, clean activity values are unknown. We observe that many of the "noisy AC" compounds with the top 10% training error indeed have high added noise and are correctly labeled as noisy samples, especially as the amount of noise in the dataset increases. Therefore, AC compounds which are more difficult to predict cannot be used to identify mislabeled samples.

Further analysis was conducted by using ensemble variance to quantify the uncertainty in each sample. Figures S37 – S40 show that in the low noise cases molecules with high ensemble variance tend to have high training error even if added noise is small and are thus mislabeled as noisy at a higher frequency. This trend is most apparent in the P_{app} dataset (Figure S39). Additionally, in the original training set without artificial noise, many of the data points with high ensemble variance are further from the parity line indicating that high uncertainty samples typically have high training error (Figures S37 – S40 a). Because slight correlation between ensemble variance and mislabeled samples was observed in the low noise regime, we tested whether preventing high ensemble variance molecules in the training set from being filtered out would improve the performance of the TE Finetuning method. Yet, we found that this approach generally has no effect or even deteriorated the model performance compared to simply denoising based on training error (see Figures S41 – S44). In the low noise regime where the correlation of uncertainty with mislabeled compounds is most apparent, the performance remains similar to the TE Finetuning denoising scheme showing that retaining high uncertainty samples in the training set does not greatly affect model performance. In the other noise regimes where this correlation is much less apparent or non-existent, the model performance remains the same or deteriorates due to noise being reintroduced into the training set.

For our chemical datasets and model architecture, underrepresented or hard to predict sample types (e.g. unique atom types, sparse clusters, low Tanimoto similarity) do not appear to cause an inflation of training error. The Chemprop model that we use is likely able to generalize better to these unique training compounds. Although clean AC compounds inflate training error in specific cases within the low noise regime, retaining these compounds is infeasible as it is not possible to identify clean AC compounds when the assay values are noisy. Additionally, it is observed that compounds with high prediction uncertainty calculated using ensemble variance affects the performance of the training error metric in the low noise regime, however, retaining these compounds did not improve the model performance. Therefore, for the ADMET predictors based on Chemprop architecture, training error itself is likely a sufficient noise detection metric.

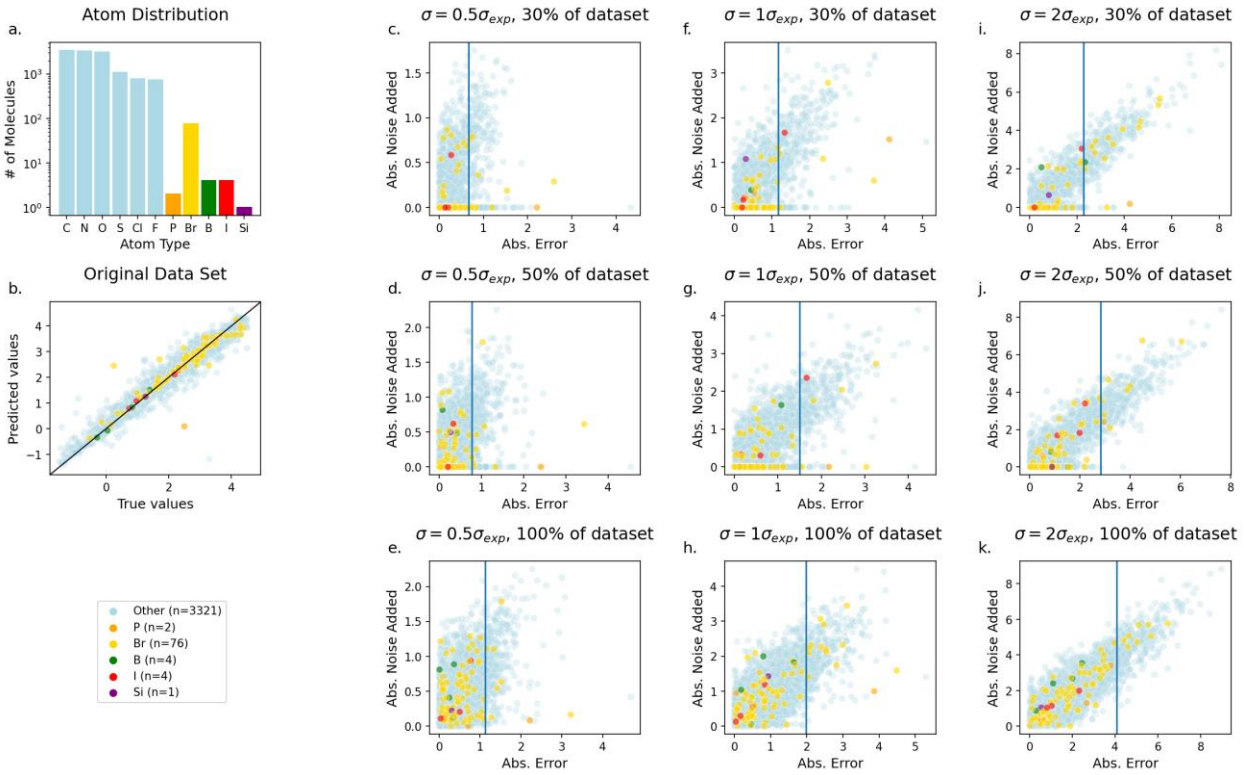

Figure 7. Effects of underrepresented samples in the form of unique atoms on the training error. (a): The atom distribution across all training set molecules. (b): The parity plot on the data set with no artificial noise added. (c-k): The correlation between training error and artificial noise added. The vertical blue line is the top 10% training error threshold. This figure was formed based on the logD public dataset.

4.7 Effects of Dataset Size on Denoising

To determine the effects of dataset size on the TE Finetuning method, we tested the denoising method on our internal logD, P_{app}, and hERG datasets with varying sizes. The datasets were prepared using 7 different sizes, ranging from 500 to 600,000 datapoints depending on total dataset size, where each smaller dataset is a complete subset of the larger ones. All models are tested on an identical cleaner test set with lower experimental uncertainty as mentioned in Section 3.1. The results are provided in the Supporting Information Figures S45-S47. In these figures, we observed multiple cases where performance improvement increases as the data size decreases. However, the relationship is not as clear in some cases compared to the results from the QM9 data in Section 4.5, likely because the internal data have inherent noise. Furthermore, we examined the effect of the dataset size on the model's robustness against noise as depicted in Figures S48-S50. The results show that both un-denoised and TE Finetune models become more robust against the added noise as the dataset size increases. This suggests that using more data in model training can help combat the loss of performance due to noise. Our finding demonstrates that even if new data exhibit similar percentage and magnitude of noise as the existing data, having more data benefits the model. We observed the positive impact of having larger data against noise up to the highest dataset size of 600k, but we anticipate this effect to plateau out as the training data begins to saturate.

4.8 Results on the Internal ADMET Data

Table 4. Summary of performances on each internal dataset on a held-out clean test set evaluated in MAE in original assay units. Bolded numbers indicate the model with the best performance, and the columns without a bolded number indicate all models have the same performance. Underlined numbers indicate the model with the second best performance.

	P_{app} (10^{-6} cm/s)	LogD (-)	Rat $F_{u,p}$ (Fraction unbound)	Human $F_{u,p}$ (Fraction Unbound)	hERG binding (μM)	SOLY 7 (μM)	FaSSIF Solubility (μM)
Data Size	55807	608053	56572	23369	370148	454288	307341
No Filter	<u>3.270 ± 0.006</u>	0.108 ± 0.003	0.022 ± 0.009	0.03 ± 0.03	9.73 ± 0.02	26.794 ± 0.008	32.064 ± 0.009
TE Finetune 10% Filter	**3.246 ± 0.002**	**0.0889 ± 0.0005**	0.022 ± 0.004	0.03 ± 0.01	**8.714 ± 0.005**	**21.21 ± 0.01**	**24.381 ± 0.007**
TE Finetune Adapt. Thres.	3.279 ± 0.005	<u>0.098 ± 0.005</u>	0.02 ± 0.03	0.03 ± 0.03	<u>9.15 ± 0.01</u>	<u>24.035 ± 0.005</u>	<u>26.78 ± 0.01</u>

We tested the TE Finetune method using both a 10% training error filter and an adaptive threshold filter on seven internal ADMET endpoints from Merck & Co., Inc. (Rahway, NJ, USA) to determine whether these workflows show utility in an industrial setting. The models were trained on the original datasets without artificial noise and tested on the cleaner held-out test sets prepared as described in Section S3.1. Since no artificial noise is added, the performance improvements from our denoising schemes, shown in Table 4, are only from denoising the experimental error in these datasets.

We report the MAE of the original un-denoised model compared to MAE of the TE Finetune model using both 10% and adaptive filters in Table 4. The MAE is compared because chemists typically prefer to view model performance using MAE reported in the original units of the assay. The TE Finetune method, regardless of filter choice, outperforms the no filter baseline in most endpoints tested. Denoising had no effect on two endpoints (rat and human $F_{u,p}$), where all model performances are equivalent. The TE Finetune method using a 10% training error filter consistently outperforms the TE Finetune method using an adaptive filter in assays where denoising has an effect on the model performance. More tuning and optimization is likely needed for the adaptive filter. The amount of experimental uncertainty in each endpoint is estimated by the correlation between repeat measurements in Figure S54. We observe that the improvement from denoising is most significant in assays with more experimental uncertainty, namely solubility at pH 7 and FaSSIF solubility. The improvements in MAE for the SOLY7 and FaSSIF solubility assays are 5.58 μM and 7.683 μM, respectively. This corroborates our findings on the public ADMET data that our method has the ability to improve models with medium noise but does not degrade the performance of models with noise outside this range. Figure S55 shows similar results

using R^2 as the model evaluation metric. We additionally observe that models with lower initial performance have a higher increase from denoising, similar to the QM9 data. We recommend the use of the TE Finetune method using a 10% training error filter in data pre-processing pipelines to improve the performance of ADMET models for drug development, especially for datasets with a significant amount of noise and with models that are low performing.

4.9 Effects of Noise on Multitask Learning Models

Multitask learning is a paradigm where a single model is trained to predict multiple related tasks simultaneously. It encourages learning between related tasks during training to improve the predictions of an assay that is either difficult to predict or has a small amount of data. Multitask models have been increasingly used to predict ADMET endpoints.[42,44] It is unknown, however, how these models react to noise. Namely, (1) if noise from one task propagates to other tasks and (2) if multitask learning is more or less robust to noise compared to single-task models.

To test this, we added noise to a single endpoint and visualized how this affects the performance of the other endpoints within the same multitask learning model. For this analysis, we trained the models on the data collected before 2023 and tested on the data collected during 01/01/2023 - 09/19/2023 to evaluate the performance on a temporal split. The multitask model was constructed using 29 ADMET assays collected within Merck & Co., Inc., Rahway, NJ, USA.[44] Six different multitask models were built, each with noise added to logD, human $F_{u,p}$, P_{app}, hERG binding, SOLY7, and CYP3A4 inhibition assays, respectively, while keeping remaining 28 assay data intact. Figure 8 shows that although the noise deteriorates the performance and increases model uncertainty for the task with noise added (green label), it does not affect the other tasks, including the tasks that are most correlated with the noisy task (black label). This observation is consistent among all internal ADMET assays tested (refer to Supporting Information Figure S51), indicating that the noise in one assay does not propagate to other related assays in multitask models. In addition, we compared the performances of the single task and multitask models with noise added to a single endpoint. While the multitask model has better predictive accuracy, there is no significant difference in the robustness against noise between single-task and multitask models (see Supporting Information Section S7).

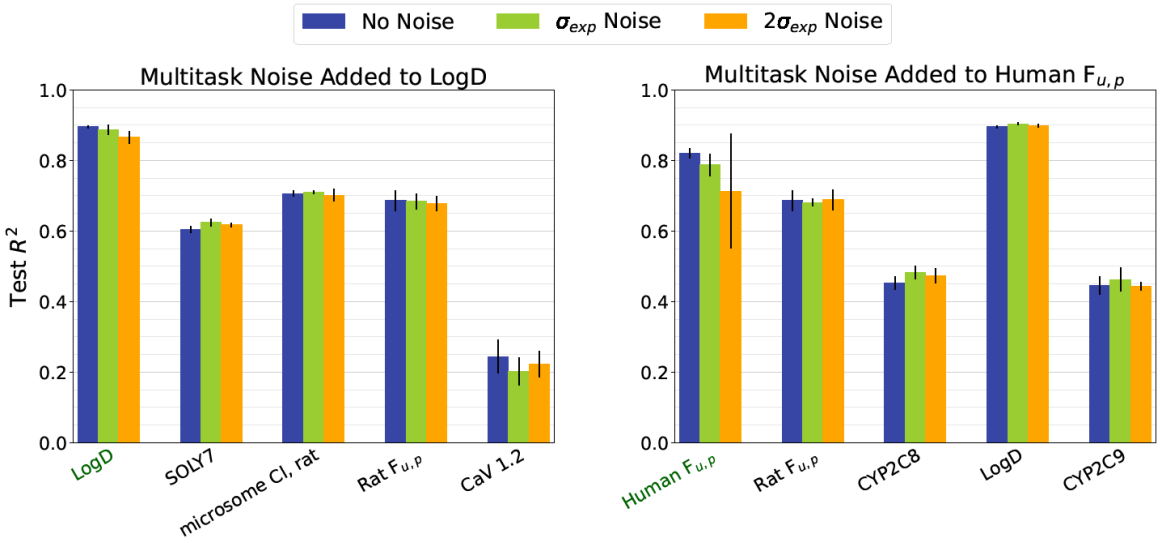

Figure 8. Visualizing noise propagation to the 4 properties with highest Spearman *r* correlation to the property with added noise (green label) in multitask models. The property colored with green text is the only task with noise added to it. This property is logD on the left figure and human $F_{u,p}$ on the right figure. Noise was added to 50% of the training data. The models were constructed using the internal dataset of 29 different ADMET endpoints.

5. Conclusions

Here, we propose a novel deep-learning based denoising scheme for regression-based ADMET datasets. We show that ensemble-based and forgotten events-based noise detection metrics do not work across the four public ADMET regression tasks used in this study. In contrast, training error is found to be highly correlated with the data noise, and the correlation strengthens as the magnitude of the noise and the fraction of noisy data increases. Our proposed TE Finetune denoising scheme, in which a model is pre-trained on the original dataset and fine-tuned on the denoised dataset with the top 10% training error data filtered out, is deemed most effective across multiple ADMET assays. This method provides performance improvement in a similar order of magnitude as the ground truth baseline when medium to high magnitude noise is present in 30% to 50% of the dataset. In other noise cases where the entire dataset is noisy, the proposed method does not decrease performance. Additionally, its positive performance on internal datasets proves its utility in industrial data pre-processing pipelines for more efficient drug discovery and development. It requires the least computational time and resources compared to the other methods investigated, and its simple framework allows easy application to other datasets. To further improve the method, we are investigating more tuning and optimization of the adaptive threshold.

Through an exhaustive analysis of various types of underrepresented samples within chemical datasets, we found that underrepresented samples do not affect the ability of the training error metric to identify noise. While more uncertain samples with higher ensemble variance affect the accuracy of the training error metric, retaining the samples with high ensemble variance in the training set does not enhance model performance. We also found that adding more data improves the model performance up to a certain degree, even if the new data are noisy. Finally, we show that noise in one task does not propagate to other related tasks in multitask models and that multitask models have similar robustness to noise compared to their single task counterparts.

6. Data and Software Availability

All public datasets and the noisy datasets generated in this work are provided as Supporting Information. The data splits used in this work are also provided in the Supporting Information. All models are constructed using the open-source software Chemprop (https://github.com/chemprop/chemprop).[37,38] This work also leverages proprietary data sets from Merck & Co. (Rahway, NJ, USA) to provide higher confidence conclusions.

Acknowledgements

This work was funded by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. The authors thank Dr. Bo Yuan for helpful suggestions and discussions and Tavleen Bhatia, Joseph Chung, and Amy Kim for their contributions.

Abbreviations

ADMET, absorption, distribution, metabolism, excretion, and toxicity; VCDR, vertical cup to disk ratio; HOMO, highest occupied molecular orbital; LUMO, lowest unoccupied molecular orbital; H298, enthalpy at 298 Kelvin; D-MPNN, directed message passing neural network; MPN, message passing network; FFN, feed-forward network; TE, training error; FE, forgotten event; EV, ensemble variance; SV, split variance; GT, ground truth; FT, finetuning; k-NN k-nearest neighbors; PCA, principal component analysis; AC, activity cliff; GCNN, graph convolutional neural network

References

(1) Ferreira, L. L. G.; Andricopulo, A. D. ADMET Modeling Approaches in Drug Discovery. *Drug Discov. Today* **2019**, *24* (5), 1157–1165. DOI: 10.1016/j.drudis.2019.03.015.

(2) Cáceres, E. L.; Tudor, M.; Cheng, A. C. Deep Learning Approaches in Predicting ADMET Properties. *Future Med. Chem.* **2020**, *12* (22), 1995–1999. DOI: 10.4155/fmc-2020-0259.

(3) Beckers, M.; Sturm, N.; Sirockin, F.; Fechner, N.; Stiefl, N. Prediction of Small-Molecule Developability Using Large-Scale In Silico ADMET Models. *J. Med. Chem.* **2023**, *66* (20), 14047–14060. DOI: 10.1021/acs.jmedchem.3c01083.

(4) van de Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discov.* **2003**, *2* (3), 192–204. DOI: 10.1038/nrd1032.

(5) Moroy, G.; Martiny, V. Y.; Vayer, P.; Villoutreix, B. O.; Miteva, M. A. Toward in Silico Structure-Based ADMET Prediction in Drug Discovery. *Drug Discov. Today* **2012**, *17* (1–2), 44–55. DOI: 10.1016/j.drudis.2011.10.023.

(6) Yang, H.; Lou, C.; Sun, L.; Li, J.; Cai, Y.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. AdmetSAR 2.0: Web-Service for Prediction and Optimization of Chemical ADMET Properties. *Bioinformatics* **2019**, *35* (6), 1067–1069. DOI: 10.1093/bioinformatics/bty707.

(7) Fralish, Z.; Chen, A.; Skaluba, P.; Reker, D. DeepDelta: Predicting ADMET Improvements of Molecular Derivatives with Deep Learning. *J. Cheminform.* **2023**, *15,* 101. DOI: 10.1186/s13321-023-00769-x.

(8) Wei, Y.; Li, S.; Li, Z.; Wan, Z.; Lin, J. Interpretable-ADMET: A Web Service for ADMET Prediction and Optimization Based on Deep Neural Representation. *Bioinformatics* **2022**, *38* (10), 2863–2871. DOI: 10.1093/bioinformatics/btac192.

(9) Wenlock, M. C.; Carlsson, L. A. How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models. *J. Chem. Inf. Model.* **2015**, *55* (1), 125–134. DOI: 10.1021/ci500535s.

(10) Hanson, S. M.; Ekins, S.; Chodera, J. D. Modeling Error in Experimental Assays Using the Bootstrap Principle: Understanding Discrepancies between Assays Using Different Dispensing Technologies. *J. Comput. Aided. Mol. Des.* **2016**, *29* (12), 1073–1086. DOI: 10.1007/s10822-015-9888-6.

(11) Sheridan, R. P.; Karnachi, P.; Tudor, M.; Xu, Y.; Liaw, A.; Shah, F.; Cheng, A. C.; Joshi, E.; Glick, M.; Alvarez, J. Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure−Activity Relationship Models? *J. Chem. Inf. Model.* **2020**, 60 (4), 1969-1982. DOI: 10.1021/acs.jcim.9b01067.

(12) Heid, E.; McGill, C. J.; Vermeire, F. H.; Green, W. H. Characterizing Uncertainty in Machine Learning for Chemistry. *J. Chem. Inf. Model.* **2023**, *63* (13), 4012–4029. DOI: 10.1021/acs.jcim.3c00373.

(13) Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; Lin, L. Unsupervised Image Super-Resolution Using Cycle-in-Cycle Generative Adversarial Networks. In *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work*, Salt Lake City, UT**,** June 18-23, 2018; pp 814–823. DOI: 10.1109/CVPRW.2018.00113.

(14) Kim, J.; Baratin, A.; Zhang, Y.; Lacoste-Julien, S. CrossSplit: Mitigating Label Noise Memorization through Data Splitting. In *International Conference on Machine Learning*; *Proceedings of Machine Learning Research*, Honolulu, HI, July 23-29, 2023; pp 16377–16392. DOI: 10.48550/arXiv.2212.01674.

(15) Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; Sugiyama, M. Co-Teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *Adv. Neural Inf. Process. Syst.,* Montreal, CA, December 2-8, **2018**; pp 8527–8537. DOI: 10.48550/arXiv.1804.06872.

(16) Gamberger, D.; Lavrac, N.; Dzeroski, S. Noise Detection and Elimination in Data Preprocessing: Experiments in Medical Domains. *Appl. Artif. Intell.* **2000**, *14* (2), 205–223. DOI: 10.1080/088395100117124.

(17) Ekambaram, R.; Fefilatyev, S.; Shreve, M.; Kramer, K.; Hall, L. O.; Goldgof, D. B.; Kasturi, R. Active Cleaning of Label Noise. *Pattern Recognit.* **2016**, *51*, 463–480. DOI: 10.1016/j.patcog.2015.09.020.

(18) Gupta, S.; Gupta, A. Dealing with Noise Problem in Machine Learning Data-Sets: A Systematic Review. *Procedia Comput. Sci.* **2019**, *161*, 466–474. DOI: 10.1016/j.procs.2019.11.146.

(19) Yuan, W.; Guan, D.; Ma, T.; Khattak, A. M. Classification with Class Noises through Probabilistic Sampling. *Inf. Fusion* **2018**, *41*, 57–67. DOI: 10.1016/j.inffus.2017.08.007.

(20) Nguyen, D. T.; Mummadi, C. K.; Nhung Ngo, T. P.; Phuong Nguyen, T. H.; Beggel, L.; Brox, T. Self: Learning To Filter Noisy Labels With Self-Ensembling. *arXiv* **2019**. DOI: 10.48550/arXiv.1910.01842.

(21) Yuan, B.; Mclean, C. Y. An Empirical Study of ML-Based Phenotyping and Denoising for Improved Genomic Discovery. *bioRxiv* **2022**. DOI: 10.1101/2022.11.17.516907.

(22) Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; Hadsell, R. Overcoming Catastrophic Forgetting in Neural Networks. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (13), 3521–3526. DOI: 10.1073/pnas.1611835114.

(23) Ritter, H.; Botev, A.; Barber, D. Online Structured Laplace Approximations for Overcoming Catastrophic Forgetting. *arXiv* **2018**. DOI: 10.48550/arXiv.1805.07810.

(24) Toneva, M.; Trischler, A.; Sordoni, A.; Bengio, Y.; Des Combes, R. T.; Gordon, G. J. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *7th Int. Conf. Learn. Represent.,* New Orleans, LA, May 6-9, 2019; DOI: 10.48550/arXiv.1812.05159.

(25) Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. Unassisted Noise Reduction of Chemical Reaction Datasets. *Nat. Mach. Intell.* **2021**, *3* (6), 485–494. DOI: 10.1038/s42256-021-00319-w.

(26) Li, C.; Mao, Z. A Label Noise Filtering Method for Regression Based on Adaptive Threshold and Noise Score. *Expert Syst. Appl.* **2023**, *228,* 120422. DOI: 10.1016/j.eswa.2023.120422.

(27) Zhou, H.; Mueller, J.; Kumar, M.; Wang, J.-L.; Lei, J. Detecting Errors in Numerical Data via Any Regression Model. *arXiv* **2023**. DOI: 10.48550/arXiv.2305.16583.

(28) Zhao, L.; Wang, W.; Sedykh, A.; Zhu, H. Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do. *ACS Omega* **2017**, *2* (6), 2805–2812. DOI: 10.1021/acsomega.7b00274.

(29) Aliagas, I.; Gobbi, A.; Lee, M. L.; Sellers, B. D. Comparison of LogP and LogD Correction Models Trained with Public and Proprietary Data Sets. *J. Comput. Aided. Mol. Des.* **2022**, *36* (3), 253–262. DOI: 10.1007/s10822-022-00450-9.

(30) Watanabe, R.; Esaki, T.; Kawashima, H.; Natsume-Kitatani, Y.; Nagao, C.; Ohashi, R.; Mizuguchi, K. Predicting Fraction Unbound in Human Plasma from Chemical Structure:

Improved Accuracy in the Low Value Ranges. *Mol. Pharm.* **2018**, *15* (11), 5302–5311. DOI: 10.1021/acs.molpharmaceut.8b00785.

(31) Iwata, H.; Matsuo, T.; Mamada, H.; Motomura, T.; Matsushita, M.; Fujiwara, T.; Maeda, K.; Handa, K. Predicting Total Drug Clearance and Volumes of Distribution Using the Machine Learning-Mediated Multimodal Method through the Imputation of Various Nonclinical Data. *J. Chem. Inf. Model.* **2022**, *62* (17), 4057-4065. DOI: 10.1021/acs.jcim.2c00318.

(32) Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59* (3), 1253–1268. DOI: 10.1021/acs.jcim.8b00785.

(33) Falcón-Cano, G.; Molina, C.; Cabrera-Pérez, M. Á. Reliable Prediction of Caco-2 Permeability by Supervised Recursive Machine Learning Approaches. *Pharmaceutics* **2022**, *14* (10), 1998. DOI: 10.3390/pharmaceutics14101998.

(34) Braga, R. C.; Alves, V. M.; Silva, M. F. B.; Muratov, E.; Fourches, D.; Lião, L. M.; Tropsha, A.; Andrade, C. H. Pred-HERG: A Novel Web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Mol. Inform.* **2015**, *34* (10), 698–701. DOI: 10.1002/minf.201500040.

(35) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9* (2), 513–530. DOI: 10.1039/c7sc02664a.

(36) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022. DOI: 10.1038/sdata.2014.22.

(37) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64* (1), 9–17. DOI: 10.1021/acs.jcim.3c01250.

(38) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3353–3600. DOI: 10.1021/acs.jcim.9b00237.

(39) Kolmar, S. S.; Grulke, C. M. The Effect of Noise on the Predictive Limit of QSAR Models. *J. Cheminform.* **2021**, *13,* 92. DOI: 10.1186/s13321-021-00571-7.

(40) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7*, 20. DOI: 10.1186/s13321-015-0069-3.

(41) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures — A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113. DOI: 10.1021/c160017a018.

(42) Walter, M.; Borghardt, J. M.; Humbeck, L.; Skalic, M. Multi-Task ADME / PK Prediction at Industrial Scale : Leveraging Large and Diverse Experimental Datasets. *ChemRxiv* **2024**. DOI: 10.26434/chemrxiv-2024-pf4w9.

(43) Sheridan, R. P.; Culberson, J. C.; Joshi, E.; Tudor, M.; Karnachi, P. Prediction Accuracy of Production ADMET Models as a Function of Version: Activity Cliffs Rule. *J. Chem. Inf. Model.* **2022**, *62* (14), 3275–3476. DOI: 10.1021/acs.jcim.2c00699.

(44) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, *63* (16), 8835–8848. DOI: 10.1021/acs.jmedchem.9b02187.