A THEORETICAL SUPPLEMENTS

A.1 PROOF OF LEMMA 1

For completeness, we start by re-stating Lemma 1 in more detail.

We consider a multi-class classification problem, with $\mathcal{K} := \{1, 2, \dots, K\}$ the set of K labels and $\hat{p}(x) := (\hat{p}_1(x), \dots, \hat{p}_K(x))$ the predicted probabilities for x that $y = k$ for all $k \in \mathcal{K}$. We denote y_i the true label for a sample x_i, $\hat{y}_i := \arg\max_k \hat{p}_k(x_i)$ the predicted class by the model and $\hat{v}_i := \max_{k \in \mathcal{K}} \hat{p}_k(x_i)$ the predicted probability value for that class. In the context of sequential label acquisition, we denote $\mathcal{X}_{\text{inf}}^{t-1} := \{x_i\}_{i=1}^{N}$ the inference set containing N samples at acquisition step $t-1$, and $\mathcal{D}_{\text{inf}}^{t-1} := \{(x_i, y_i)\}_{i=1}^{N}$ the same set with the associated labels y_i. Stepping forward, at time t, an acquisition batch of N_b samples and their labels $\mathcal{D}_b^t := \{(x_i, y_i)\}_{i=1}^{N_b}$ are selected from the inference dataset of the previous step $\mathcal{D}_{\text{inf}}^{t-1}$.

Following the least-confidence acquisition function, at time-step t, the samples in $\mathcal{D}_{\text{inf}}^{t-1}$ are ordered by confidence where x_1 has the predicted class with the lowest confidence and x_N has the predicted class with the highest confidence:

$$\underbrace{\hat{v}_1 \leq \hat{v}_2 \leq \cdots \leq \hat{v}_{N_b}}_{\text{first } N_b \text{ samples}} \leq \underbrace{\hat{v}_{N_b+1} \leq \cdots \leq \hat{v}_N}_{\text{last } N - N_b \text{ samples}} \tag{6}$$

and the selected batch consists of the first N_b samples. The remaining inference set is composed of the remaining $N - N_b$ samples, i.e. $\mathcal{D}_{\text{inf}}^t := \mathcal{D}_{\text{inf}}^{t-1} \setminus \mathcal{D}_b^t = \{(x_i, y_i)\}_{i=1+N_b}^{N}$.

The practical utility of this ordering depends on the calibration of the model. Calibration reflects the correctness of the model's confidence compared to its true performance. A perfectly calibrated model $\hat{p}$ would output confidence levels that perfectly match the true probability of correct classification. We extend the notation of Guo et al. (2017) to be dataset-specific, and denote $P_{\mathcal{D}}[Y = \hat{Y}]$ the probability of $\hat{p}$ producing the correct class prediction for any sample from a dataset $\mathcal{D}$, which can equivalently be thought of as $\hat{p}$'s accuracy on that dataset:

$$P_{\mathcal{D}}[Y = \hat{Y}] := E_{(x_i, y_i) \sim \mathcal{D}}[\mathbf{1}(y_i = \hat{y}_i)] \tag{7}$$

Similarly, $P_{\mathcal{D}}[Y = \hat{Y} | \tilde{v}]$ denotes the *confidence-conditional* accuracy of the model for $\tilde{v}$, a specific confidence level[2]:

$$P_{\mathcal{D}}[Y = \hat{Y} | \tilde{v}] := E_{(x_i, y_i) \sim \mathcal{D}}[\mathbf{1}(y_i = \hat{y}_i) | \hat{v}_i = \tilde{v}] \tag{8}$$

Perfect calibration on the inference set would thus imply that for any confidence level $\tilde{v} \in [0, 1]$, we have $P_{\mathcal{D}_{\text{inf}}^{t-1}}[Y = \hat{Y} | \tilde{v}] = \tilde{v}$. Instead, in Lemma 1, we only assume *weak calibration* of the model on the predicted class, i.e. that the confidence *ordering* of the model's top predictions reflects the ordering of probabilities of correct classification. Formally, weak calibration on $\mathcal{D}_{\text{inf}}^{t-1}$ implies that for any pair of samples x_i and x_j in our dataset $\mathcal{D}_{\text{inf}}^{t-1}$, we have:

$$\hat{v}_i \leq \hat{v}_j \implies P_{\mathcal{D}_{\text{inf}}^{t-1}}[Y = \hat{Y} | \hat{v}_i] \leq P_{\mathcal{D}_{\text{inf}}^{t-1}}[Y = \hat{Y} | \hat{v}_j] \tag{9}$$

We aim to prove that, assuming weak calibration on the inference set, the expected accuracy on the least-confidence batch is bounded by the expected accuracy over the remaining inference set:

$$P_{\mathcal{D}_b^t}[Y = \hat{Y}] \leq P_{\mathcal{D}_{\text{inf}}^t}[Y = \hat{Y}] \tag{10}$$

Proof. The expected accuracy from Equation 7 can be rewritten in terms of its confidence-conditional form in Equation 8:

$$P_{\mathcal{D}}[Y = \hat{Y}] = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} P_{\mathcal{D}}[Y = \hat{Y} | \hat{v}_i] \tag{11}$$

[2] In practice such confidence levels would be binned together such that $P_{\mathcal{D}}[Y = \hat{Y} | \tilde{v}]$ represents the accuracy over all samples in $\mathcal{D}$ for which the model's confidence is in $[\tilde{v} - \epsilon, \tilde{v} + \epsilon]$, see (Guo et al., 2017).

From the weak calibration assumption in Equation 9, it follows that the model's confidence-conditional accuracy is a non-decreasing function of its confidence. Combined with the acquisition mechanism of Equation 6, we obtain an ordering of the probability that the model obtains the correct label along the entire inference set at step $t-1$:

$$\underbrace{P_{\mathcal{D}_{\text{inf}}^{t-1}}[Y=\hat{Y}|\hat{v}_1] \leq \cdots \leq P_{\mathcal{D}_{\text{inf}}^{t-1}}[Y=\hat{Y}|\hat{v}_{N_b}]}_{N_b \text{ first terms}} \leq \underbrace{P_{\mathcal{D}_{\text{inf}}^{t-1}}[Y=\hat{Y}|\hat{v}_{N_b+1}] \leq \cdots \leq P_{\mathcal{D}_{\text{inf}}^{t-1}}[Y=\hat{Y}|\hat{v}_N]}_{N-N_b \text{ remaining terms}} \tag{12}$$

The empirical average of a set of numbers that are inferior or equal to those of a second set has to be inferior or equal to the empirical average of the second set:

$$\frac{1}{N_b}\sum_{i=1}^{N_b} P_{\mathcal{D}_{\text{inf}}^{t-1}}[Y=\hat{Y}|\hat{v}_i] \leq \frac{1}{N-N_b}\sum_{i=1+N_b}^{N} P_{\mathcal{D}_{\text{inf}}^{t-1}}[Y=\hat{Y}|\hat{v}_i] \tag{13}$$

By definition, $\mathcal{D}_{\text{inf}}^{t} := \mathcal{D}_{\text{inf}}^{t-1} \setminus \mathcal{D}_{b}^{t}$, the left-hand side from Equation 13 thus captures all the terms from $\mathcal{D}_{b}^{t}$ and the right-hand side those from $\mathcal{D}_{\text{inf}}^{t}$. Rewriting each side using Equation 11, we obtain:

$$P_{\mathcal{D}_{b}^{t}}[Y=\hat{Y}] \leq P_{\mathcal{D}_{\text{inf}}^{t}}[Y=\hat{Y}] \tag{14}$$

This concludes the proof. □

A.2 Bound derivation

Here we derive the bound presented in Equation 5. The Chernoff bound (Chernoff, 1952) for Bernouilli random variables provides an exponential tail bound for the true mean μ of a sequence of Bernouilli distributed random variables $Z_1, Z_2, \ldots, Z_n$. It can be expressed using the Kullback-Leibler Divergence (KL) for Bernouilli distributions (Lattimore & Szepesvári, 2020, page 135, Corrollary 10.4):

$$P(\mu \leq a) \leq \exp\big(-n \cdot \text{KL}(\hat{\mu}||a)\big) \quad \forall \quad a \in [0, \hat{\mu}] \tag{15}$$

$$\text{with} \quad \text{KL}(\hat{\mu}||a) := \hat{\mu}\log\frac{\hat{\mu}}{a} + (1-\hat{\mu})\log\frac{1-\hat{\mu}}{1-a} \tag{16}$$

It quantifies the probability that the true mean μ is smaller or equal to some bound a given the observed sample mean $\hat{\mu}$ and sample size n. For us, the value $\{0,1\}$ of a variable Z_i indicates whether a particular sample x_i was correctly classified i.e. $Z_i := \mathbf{1}(y_i = \hat{y}_i)$. In the context of inference set design and following the notation from Section 2, $\hat{\mu}_b^t$ represents the observed accuracy on the acquired batch at time-step t, with acquisition batch-size N_b, and μ_{inf}^t represents the (unknown) accuracy of the model on the remaining inference set. Note that the Chernoff bound usually assumes that $\hat{\mu}_b^t$ is an unbiased estimator of μ_{inf}^t. In our case, $\hat{\mu}_b^t$ is a biased estimator due to the active selection mechanism of the batch. However, in Lemma 1, we show that this estimator is actually a conservative estimate of the true accuracy, and thus in turn contributes to making the bound presented in Equation 15 even more conservative, and preserves its validity.

To establish a confidence level on that bound, we can lower-bound the right-hand side itself to the desired bound-failure probability δ, which after rearranging yields:

$$\text{KL}(\hat{\mu}||a) \geq \frac{\log(\frac{1}{\delta})}{N_b} \tag{17}$$

At time t, we thus seek the maximum bound value $a = \alpha_t$ for μ_{inf}^t such that the inequality on Equation 17 holds by finding the value of a that satisfies the following condition:

$$\alpha_t = \min_a \left\{ a \in [0, \hat{\mu}_b^t] : \text{KL}(\hat{\mu}||a) \leq \frac{\log(\frac{1}{\delta})}{N_b} \right\} \tag{18}$$

Since this is a scalar optimization problem from closed-form expressions, computing α_t can be done easily and efficiently using a grid-search. With this choice for a, we obtain the desired probabilistic bound $P(\mu_{\text{inf}}^t < \alpha_t) \leq \delta$ summarized in Equation 5.

B ALGORITHM

Algorithm 1 Hybrid Screen using Inference Set Design

Input: Acquisition batch-size N_b, threshold γ, margin δ
Initialize Step t=0, observation set $\mathcal{X}_{\text{obs}}^{t=0} \leftarrow \emptyset$, inference set $\mathcal{X}_{\text{inf}}^{t=0} \leftarrow \mathcal{X}_{\text{target}}$, predictor $\hat{p}$.
repeat
  Train the predictor $\hat{p}$ on $(\mathcal{X}, \mathcal{Y})_{\text{obs}}^{t}$ (if not empty)
  Obtain the predictions on the inference set $\hat{\mathcal{P}}_{\text{inf}}^{t} \leftarrow \{\hat{p}(x_i)\ \forall x_i \in \mathcal{X}_{\text{inf}}\}$
  Run acquisition function to obtain scores $\mathcal{S}_{\text{inf}}^{t} \leftarrow g\big(\hat{\mathcal{P}}_{\text{inf}}^{t}\big)$
  Select a batch of N_b inputs with the highest scores $\mathcal{S}_{\text{inf}}^{t}$ to form $\mathcal{X}_b^t$
  Remove the acquired batch from the inference set $\mathcal{X}_{\text{inf}}^{t} \leftarrow \mathcal{X}_{\text{inf}}^{t-1} \setminus \mathcal{X}_b^t$
  Obtain the true labels $\mathcal{Y}_b^t$ for the acquisition batch
  Append the acquired batch to the observation set $(\mathcal{X}, \mathcal{Y})_{\text{obs}}^{t} \leftarrow (\mathcal{X}, \mathcal{Y})_{\text{obs}}^{t-1} \cup (\mathcal{X}, \mathcal{Y})_b^t$
  Compute α on $(\mathcal{X}, \mathcal{Y})_b^t$ from Equation 5
until

$$\frac{|\mathcal{X}_{\text{obs}}^{t}| + \alpha|\mathcal{X}_{\text{inf}}^{t}|}{|\mathcal{X}_{\text{target}}|} > \gamma \quad \text{or} \quad \mathcal{X}_{\text{inf}}^{t} = \emptyset$$

Return hybrid screen readouts: $(\mathcal{X}, \mathcal{Y})_{\text{obs}}^{t=\tau} \cup (\mathcal{X}, \hat{\mathcal{Y}})_{\text{inf}}^{t=\tau}$

C ADDITIONAL DETAILS ON DATASETS AND PREPROCESSING

C.1 MOLECULAR DATASET PREPROCESSING

In many practical applications exact geometries of screened molecules are unknown as they require computationally expensive DFT calculations. As a first data processing step, we use RDKit library (Landrum et al., 2024) to convert molecular structures into SMILES strings and compute their Extended Connectivity Fingerprints (ECFPs). The SMILES representation provides complete information about molecule's composition and atomic connectivity, however, it removes all information about 3D atomic positions. Using SMILES representation is a common solution that simplifies generation of candidate molecules for screening but makes property prediction a more challenging task as many properties vary depending on specific 3D conformation of a molecule.

Both molecular datasets are cleaned by removing duplicated SMILES and fingerprints as well as single-atom structures. For total energies in Molecules3D dataset we use reference correction technique where atomic energies are calculated using linear model fitted to the counts of atoms in a molecule of each element present in the dataset (obtained atomic energies are presented in Table 1). For reference correction a randomly selected sample of 100k molecules is used. The atomic energies are then subtracted from the total energies of all molecules in the dataset. The obtained referenced-corrected energies are normally distributed with mean around 0 eV. A small number of outliers with reference-corrected energy values above 10 standard deviations are removed from the dataset as well as 100k samples that were used for reference correction to avoid data leakage.

The final QM9 and Molecule3D datasets contain $133,885$ and $3,453,538$ molecules respectively. Both datasets are split into inference, validation, and test sets with 80%, 5%, 15% fractions. The QM9 HOMO-LUMO gap values are discretized into 2 balanced classes using median as a boundary condition to explore agents' performance on classification task.

Atomic Number	Energy (eV)	Atomic Number	Energy (eV)
1	-26.765	14	-7922.180
5	-673.550	15	-9313.610
6	-1054.411	16	-10810.329
7	-1483.001	17	-12474.680
8	-2034.056	32	-56538.647
9	-2687.455	33	-60828.588
12	-5367.759	34	-65296.349
13	-6657.476	35	-69980.303

Table 1: Atomic energies used for reference correction on Molecules3D dataset.

C.2 RxRx3 dataset preprocessing

The RxRx3 dataset contains one embedding for each well. Each perturbation type (gene-guide pair or compound-concentration pair) has several replicates across wells, plates and experiments. Each plate also contains unperturbed control cells which are used to keep track of and eliminate a portion of the batch effects (Sypetkowski et al., 2023). These raw embeddings thus need to be aligned and aggregated. We align them by centering and scaling each perturbation embedding to the embeddings of the experiment-level unperturbed control wells. The embeddings are then aggregated through a multi-stage averaging procedure, across wells, plates, experiments and guides (for CRISPR perturbations), which yields an average embedding for each gene-perturbation and each compound-concentration perturbation. We then use the obtained embeddings to compute cosine similarities between gene and compound perturbations in the RxRx3 dataset.

D Hyperparameters and Implementation details

For all presented experiments in this work we use MLP models with residual connections (Touvron et al., 2021). All experiments were repeated with 3 different random seeds. Hyperparameters for each experiments are summarized in Table 2.

Hyperparemeter name	MNIST	QM9	Molecules3D	RxRx3	Proprietary data
Acquisition batch size	1,000	250	10,000	10	1000
Number of hidden layers	2	3	2	2	2
Hidden layer size	512	512	512	512	1024
Learning rate	0.001	0.001	0.001	0.001	0.001
gradient norm clip	1.0	1.0	1.0	1.0	1.0
Dropout	0.1	0.1	0.1	0.1	0.1
Train epochs	1,000	1,000	30	30	1000
Train batch size	1,024	1,024	32,768	1,024	1,024
Early stop patience	50	50	15	25	25
Number of ensemble members	None	None	5	None	None

Table 2: Hyperparameters for experiments.

E ADDITIONAL RESULTS

E.1 REGRESSION TASK ON MOLECULES3D

To evaluate the inference set design paradigm on a regression task we use a large dataset Molecules3D (Xu et al., 2021). Molecules3D contains structures and DFT-computed properties of approximately 4 million molecules. In our experiments we aim to predict the HOMO-LUMO gap and total energy. For inputs, we convert the SMILES strings molecular representations into their Extended Connectivity Fingerprints (ECFPs).

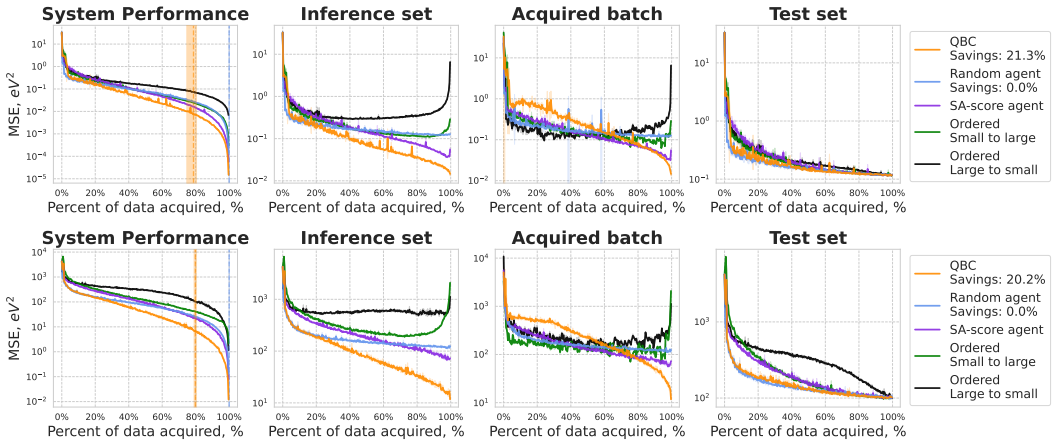

Figure 10: Performance of QBC, random, and heuristic agents for HOMO-LUMO gap prediction (top) and energy prediction (bottom) on Molecules3D dataset. The dashed vertical shows average stopping time across random seeds, surrounded by mix to max intervals (left to right). The patience parameter $p = 10$ was used in all experiments with the threshold MSE $t_{MSE} = 0.1eV^2$ for HOMO-LUMO gap and $t_{MSE} = 100eV^2$ for Energy predictions. QBC agent satisfied stopping condition after acquiring $\approx 80\%$ of the data.

For this regression task, we use a query-by-committee (QBC) active learning approach that computes variance across the predictions of an ensemble. To determine the stopping time we use a criterion with two parameters: MSE threshold t_{MSE} on the acquired batch and patience p. The stopping time is reached if the acquired batch MSE is lower than t_{MSE} for p steps. Like for our QM9 experiments, in addition to active and random agents, we also evaluate the performance of heuristic-based acquisition orderings (molecules ordered by size, sorted by SA-score, etc.). QBC achieves an approximately five times lower MSE compared to the random agent or heuristic-based orderings (see Figure 10). This shows that inference set design approach is not limited to classification tasks and can be applied to regression problems.

Although, in many applications predicting properties of larger molecules present a more challenging task, our experiments on the Molecules3D dataset demonstrate that acquiring molecules ordered from large to small may harm the predictions on inference set and overall system performance (see Figure 10). One of the reasons is the distribution of chemical elements across molecules in the dataset. When acquiring molecules from small to large, all unique chemical elements of the Molecules3D dataset are present in the training set after acquiring just the first 1,000 samples. However, when acquiring molecules from large to small, some chemical elements remain only in the inference set until the very end of the experiment which is especially detrimental for the total energy predictions (see Figure 11). This result demonstrates that using heuristic rules such as ordering molecules by size for data acquisition does not guarantee optimal acquisition or generalization of heuristic rules to new data.

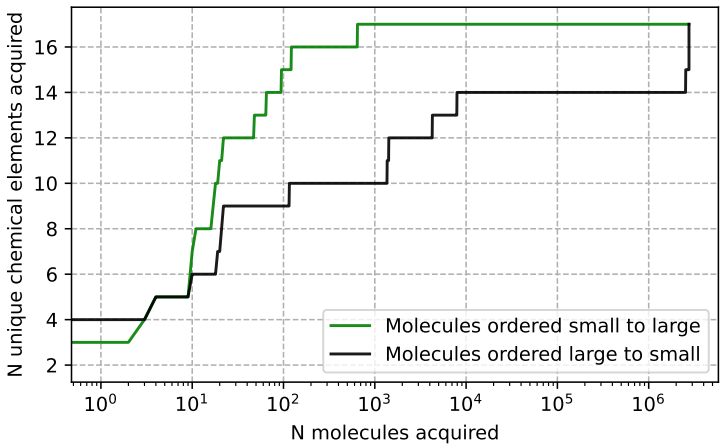

Figure 11: Number of unique chemical elements in the training set when acquiring samples ordered by molecule size. Acquiring molecules from large to small leaves several chemical elements out of the training set until the end of the dataset.

E.2 Classification task on RxRx3

Results on the RxRx3 dataset demonstrate minor improvements from the active agents (LC and BALD) compared to random selection. The accuracy on inference and test set remains low throughout the experiment regardless of the acquisition method. This is expected considering the extreme difficulty of predicting biological relationships in a low data regime. The result suggests that in the setting with low predictive power, even when the majority of the data is acquired, the ability of inference set design to provide significant budget reductions is limited.

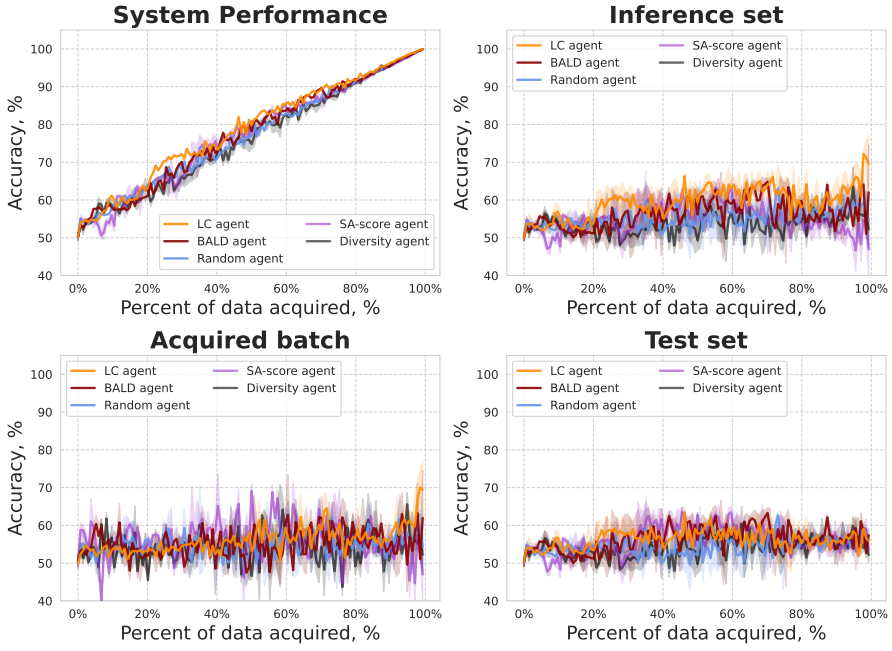

Figure 12: Performance of agents on pheno-similarity classification task on RxRx3.

F ADDITIONAL ANALYSIS

F.1 EMPIRICAL VALIDATION OF WEAK CALIBRATION ASSUMPTION

In Lemma 1, we show that when using a least-confidence acquisition function, at any time-step t, the measured accuracy $\hat{\mu}_b^t$ on the acquisition batch is a lower-bound on the unobserved accuracy on the remaining inference set $\hat{\mu}_{\text{inf}}^t$, assuming that the model $\hat{p}$ is weakly calibrated in the infence set. A weakly calibrated model is such that an increased confidence translates to a higher likelihood of correct prediction (higher accuracy). In Figure 13, we can see that at several points throughout the active learning loop, the least-confidence based agent is *not* perfectly calibrated. Indeed, there is a substantial gap between its confidence levels and the true accuracy (seen when comparing the colored bars to the identity function shown in gray). However, the model is generally *weakly* calibrated (the colored bars are always increasing). These observations support the empirical validity of our assumption for Lemma 1.

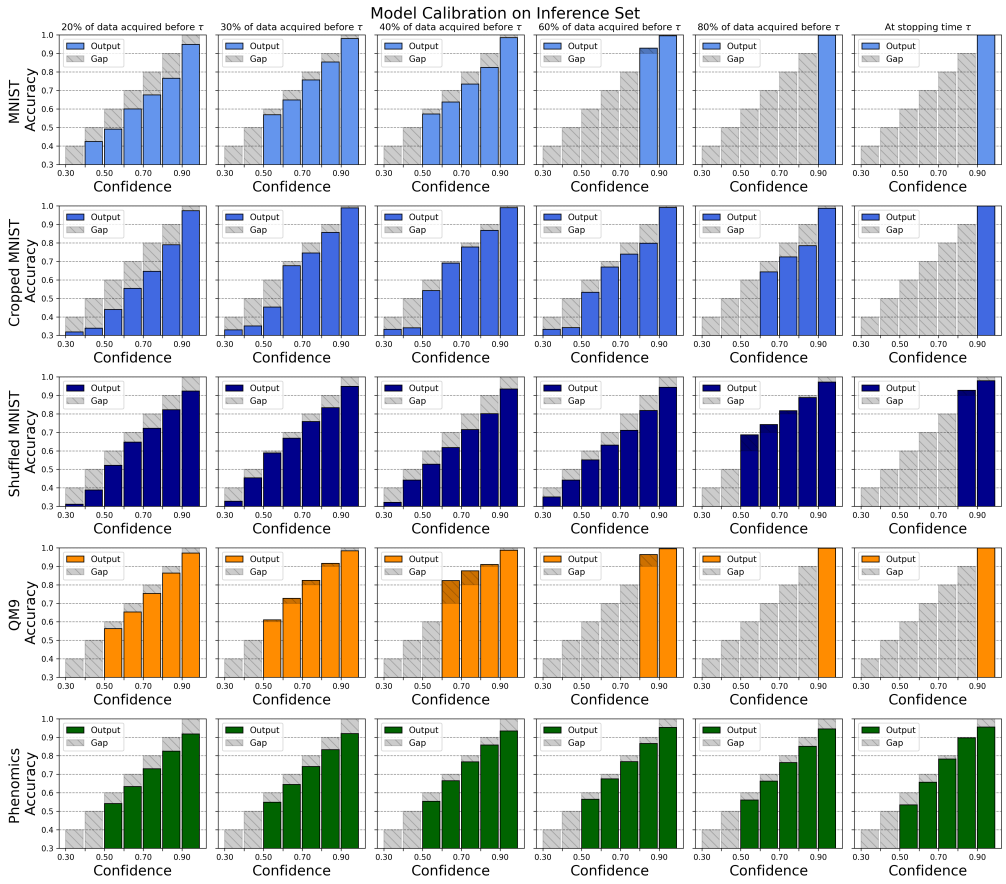

Figure 13: Model calibration analysis for a run of the least confidence agent at different active learning steps (columns) for all of our classification experiments (rows). The colored bars represent the accuracy of predictions binned w.r.t their confidence level, and the gray bars show the identity function illustrating what perfect calibration would look like. The models satisfy the condition of being weakly calibrated since the accuracy of the model's predictions increases monotonically with their confidence. The confidence distribution shifting to the right as t increases indicates that the growing confidence of the model in correctly predicting the labels of the remaining examples in the inference set.

F.2 EMPIRICAL VALIDATION OF STOPPING CRITERION

In Section 2.2, we present a stopping criterion based on the fact that the accuracy measured on the acquisition batch can be used as a proxy for the accuracy on the remaining, non-acquired samples

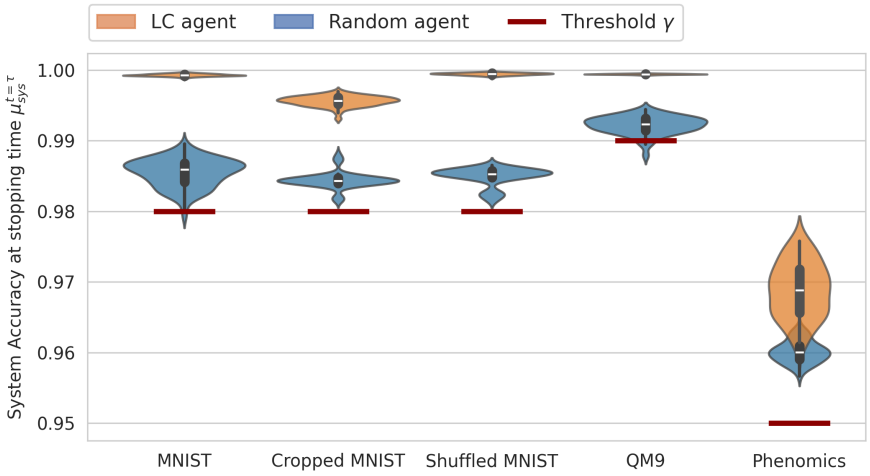

Figure 14: System accuracy at stopping time observed across 100 seeds for the MNIST, Cropped MNIST, Shuffled MNIST and QM9 experiments, and 50 seeds for the Phenomics experiments. In all experiments, we use a bound-failure probability of $\delta = 0.05$. For the LC agent, all trials lead to a stopping time $t = \tau$ at which the true system accuracy $\mu_{\text{sys}}^{t=\tau}$ higher than the threshold γ. For the Random Agent, we observed the same thing for Cropped MNIST, Shuffled MNIST and Phenomics. We also observed 1% of bound failure for (regular) MNIST, and 6% for QM9.

(inference set). The criterion is simple, once the estimated system accuracy $\hat{\mu}_{\text{sys}}^{t}$ at time t (see Equation 4) is above a user-defined targeted threshold γ, the acquisition is stopped. This criterion makes use of the bound of Equation 5, and both a random sampler and a least-confidence based agent can use this bound in a principled way. The random agent can validly use it because the accuracy of its acquisition batches, uniformly drawn from the inference set, are unbiased estimates of the true accuracy on the entire inference set, which is typically required for such bounds. The least-confidence agent can use it validly because we show in Lemma 1 that, assuming that the model is weakly calibrated (which we empirically validate in Appendix F.1), because of its confidence-based acquisition function, the accuracy of its acquisition batches represent a *lower-bound* on the true inference set accuracy, making the bound of Equation 5 even more conservative.

To empirically validate this bound (Equation 5), we run a large number of trials for both the least-confidence (LC) and random agents across the classification datasets used in the experiments section. The results show that for the LC agent, all trials ended at a stopping time $t = \tau$ at which the true system accuracy μ_{sys}^{t} was above the threshold γ, which is in accordance with the fact that the LC agent uses a lower-bound estimate in place of an unbiased estimate for the inference set accuracy, resulting in a looser bound for Equation 5 and an actual failure-probability lower than δ. For the random agent, we observed 1% of the trials end with a system accuracy below the threshold in one of the experiment, 6% in another, and 0% on the three remaining datasets. These results are in accordance with the theory presented in Section 2.2 and Appendix A.2. On QM9, 6 out of 100 experiments resulting in slight bound failure is statistically in accordance with the theoretical bound failure probability of $\delta = 5\%$. For the other datasets, the bound was looser, which is also a possibility, and showcases that the observed gap between δ and the true bound-failure probability can be problem-dependent.