Workshop at the 7th Symposium on Advances in Approximate Bayesian Inference (non-archival), 2025 1–14

Semi–Supervised Bayesian Active Learning with Task–Driven Representations

Kianoosh Ashouritaklimi KIANOOSH.ASHOURITAKLIMI@STATS.OX.AC.UK
Tom Rainforth RAINFORTH@STATS.OX.AC.UK
Department of Statistics, University of Oxford

Abstract

Current strategies for semi-supervised Bayesian active learning are generally based on learning unsupervised representations and then performing active learning on the resulting latent space with a supervised model. We find that this approach can break down with messy, uncurated pools as the representations fail to capture the right similarities between our inputs. To address this, we propose the use of task-driven representations that are periodically updated during the active learning process. Initial empirical results suggest our approach leads to more effective acquisitions and enhances model performance.

1. Introduction

Bayesian active learning (MacKay, 1992; Houlsby et al., 2011; Gal et al., 2017) is a framework for selecting the most informative data points to label during the training of a probabilistic model. It achieves this by estimating how the uncertainty of the model will change once updated with new data, then choosing labels to minimize this.

While the field has traditionally used fully supervised models (Houlsby et al., 2011; Chitta et al., 2018; Kirsch, 2023), there has been a growing line of work highlighting the benefits of using semi-supervised models (Burkhardt et al., 2018; Hacohen et al., 2022; Seo et al., 2022b; Mittal et al., 2023; Bickford Smith et al., 2024). By incorporating the rich information available in the unlabeled data, semi-supervised approaches are not only able to improve the immediate predictions of the model, but also the effectiveness of the acquisitions by improving reducible uncertainty estimation (Osband et al., 2022b; Bickford Smith et al., 2024).

Current semi-supervised approaches typically learn unsupervised representations using the unlabeled data upfront, then perform active learning on top of the learnt representations using a fully supervised prediction head (Emam et al., 2021; Osband et al., 2022a; Seo et al., 2022b; Bickford Smith et al., 2024). As we can expect these representations to capture much of the important information in our data, this allows for the use of more lightweight supervised prediction heads, which in turns improves both the computational efficiency and consistency in uncertainty across updates (Bickford Smith et al., 2024). Moreover, by compressing our inputs to a lower-dimensional space, it also allows our model to make better similarity judgments between inputs and consequently produce more appropriate predictive correlations, which are crucial for effective acquisitions (Wang et al., 2021; Osband et al., 2022d).

Our key insight is that this approach can fail in the presence of messy pools—that is pools where information about the target predictive task is heavily diluted by information not relevant to the task—which is precisely the scenario where active learning can be most impactful (Sun et al., 2017; Emam et al., 2021; Citovsky et al., 2021; Zhang et al., 2022). Indeed, we show that the task-agnostic nature of unsupervised representation learning can lead to representations that fail to capture all the information relevant to our task. As a result, the representations can fail to capture the right notion of similarity between inputs for our task, leading to inaccurate predictive correlations and, ultimately, suboptimal acquisitions.

We suggest to address this issue by introducing *task-driven* representations. Namely, we argue that updating our representations throughout the active learning process using semi-supervised learning techniques allows us to guide these representations towards capturing task-relevant information. This, in turn, should enable our model to better learn the relevant similarities between inputs, improve reducible uncertainty estimation, and make better acquisition decisions. Initial experiments find that this leads to improved performance.

2. Background

We consider probabilistic models $p_\phi(y|x)$ for inputs $x \in \mathcal{X}$ and outputs $y \in \mathcal{Y}$. We assume that $p_\phi(y|x)$ takes the form $p_\phi(y|x) = \mathbb{E}_{p_\phi(\theta)}\left[p_\phi(y|x,\theta)\right]$, where θ are stochastic model parameters and ϕ indicates learnable aspects of the model, such that model updates are reflected through changes to ϕ. Data is treated to be i.i.d. conditional on θ. For ease of exposition, we will further assume a classification setting, such that $\mathcal{Y} = \{1, \ldots, K\}$, but note that all the ideas introduced apply more generally.

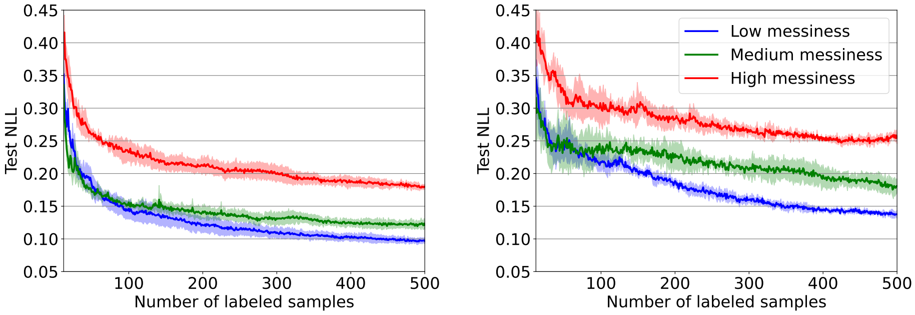
Bayesian active learning Deriving information–theoretic acquisition functions for active learning using ideas from the framework of Bayesian experimental design (Lindley, 1956; Rainforth et al., 2024) leads to what is known as Bayesian active learning (BAL, MacKay (1992)). These acquisition functions target data that would maximally reduce our model's uncertainty through a hypothetical Bayesian update to $p_\phi(\theta)$. For example, the EPIG acquisition function (Bickford Smith et al., 2023) aims to reduce uncertainty in future hypothetical predictions, while the BALD score (Houlsby et al., 2011) aims to reduce uncertainty in θ itself: $\mathrm{BALD}(x) = \mathbb{E}_{p_\phi(y|x)}[\mathrm{H}[p_\phi(\theta)] - \mathrm{H}[p_\phi(\theta|x,y)]]$.

Semi–supervised active learning By using the rich information available from unlabeled data points, semi-supervised active learning approaches (Burkhardt et al., 2018; Hacohen et al., 2022; Seo et al., 2022b; Mittal et al., 2023; Bickford Smith et al., 2024) are not only able to offer immediate gains in predictive performance, but are also able to better capture predictive correlations and reducible uncertainty, allowing for more effective acquisition.

Current semi-supervised BAL approaches are typically based on splitting the predictive model $p_\phi(y|x)$ into a fixed deterministic encoder, $g : X \to \mathbb{R}^d$, and stochastic prediction head, $p_\phi(y|z,\theta_h)$, where $z = g(x)$ is the representation output by the encoder, $\theta_h \sim p_\phi(\theta_h)$ are the stochastic parameters of our prediction head, and our overall predictive model is $p_\phi(y|x) = \mathbb{E}_{p(\theta_h)}\left[p_\phi(y|g(x),\theta_h)\right]$ (Bickford Smith et al., 2024). By fixing the encoder, we can leverage large, pretrained unsupervised encoders that capture much of the information needed for our downstream task in a lower-dimensional space (Chen et al., 2020b,a). This allows the use of smaller prediction heads which improves the computational efficiency of the active learning and the quality of the updates (Bickford Smith et al., 2024).

3. Shortfalls of unsupervised representation learning in BAL

While unsupervised representation learning followed by active learning on the latent space has proven effective for both vision and natural-language processing tasks (Emam et al.,



(*a*) EPIG (*b*) BALD

Figure 1: Test NLL for EPIG and BALD with unsupervised representations on **F+MNIST** (setup as per Table 1 in Appendix A) under increasing levels of pool "messiness", namely decreasing the number of pool samples which are of the classes of interest.

2021; Osband et al., 2022a; Bickford Smith et al., 2024), we now explain how this can break down when applied to messy data pools, for which active learning is often most needed.

Real-world active learning often involves messy pools that can be characterized by two key features: **class imbalance** and **redundant information** (Kothawade et al., 2021; Emam et al., 2021; Zhang et al., 2022). Redundant information can either be presented in the form of *redundant classes* (e.g. images of cats when our task is classifying dog breeds) or *redundant features* (e.g. pictures of cars when our task is to classifying number plate digits).

A weakness of unsupervised representations here is that as our data becomes increasingly messy, the representations fail to capture all the information relevant for our task (see Figure 4 in the Appendix). This has been observed outside of the active learning context for various unsupervised representation learning methods (Caron et al., 2019; Tian et al., 2021; Shi et al., 2022). At a high level, it comes from unsupervised representations being *task-agnostic*: as the pool becomes messier, the *task-specific* information becomes smaller compared to the task-irrelevant information, and the representation increasingly focuses on the latter.

A direct consequence of this is that it hurts the performance of BAL algorithms, as shown in Figure 1. This is expected since selecting the most informative data points relies on the model's ability to make accurate similarity judgments in the latent space (Bickford Smith et al., 2024). Capturing these similarities is essential for establishing the predictive correlations that drive effective exploration and exploitation in active learning (Wang et al., 2021; Osband et al., 2022c,d). However, these similarities are task-dependent and break down with messier pools as our representations fail to include relevant task-specific information.

4. Using Semi–Supervised Representations in Bayesian Active Learning

Motivated by the issues discussed in Section 3, we propose to instead use *task-driven* representations in semi-supervised BAL. Our suggested approach builds on the semi-supervised approach of Bickford Smith et al. (2024) described in Section 2. However, instead of using a fixed unsupervised encoder, we regularly update it as we acquire more labels using a semi-supervised representation learning technique. That is, our predictive model is given by $p_\phi(y|x) = \mathbb{E}_{p(\theta_h)}\left[p_\phi(y|g(x), \theta_h)\right]$, where $p_\phi(y|z, \theta_h)$ is our prediction head with stochastic parameters $\theta_h \sim p_\phi(\theta_h)$ and $z = g(x)$ are our representations as before, but $g : X \to \mathbb{R}^d$ is now a semi-supervised encoder that utilizes both the unlabeled data *and* acquired labels.

There are a variety of different methods one could use to learn this task-driven semi-supervised encoder (Kingma et al., 2014; Narayanaswamy et al., 2017; Chen et al., 2020b; Assran et al., 2021; Mo et al., 2023). Something they generally have in common is that they utilise a "guidance classifier", $c : \mathbb{R}^d \to [0, 1]^{|\mathcal{Y}|}$, that maps representations to class probabilities. This classifier will be learned alongside the encoder itself, typically by maximising an objective that accounts for both fidelity of the representation across all the data and the performance of the classifier on the labelled data. The aim of this is to guide the representations to be task–driven, such that they retain the information required for both effective downstream prediction and label acquisition. The guidance classifier can simply be taken to be the prediction head, but as we explain later it will typically be beneficial for it to be distinct.

The best setup to use for training the encoder will inevitably vary between problems, but we now outline one possible setup—inspired by the CCVAE approach of Joy et al. (2021)—which is carefully curated to our needs for effective active learning.

4.1. A Task–Driven Semi–Supervised Active Learning Approach

The characteristic capturing variational auto-encoder (CCVAE) approach of Joy et al. (2021) is a VAE-based (Kingma et al., 2013), semi-supervised representation learning method that aims to capture label–specific information in the representations it learns. This is achieved by partitioning the representations as $z = z_c \cup z_{\setminus c}$, where only z_c is taken as input to the guidance classifier(s), while the whole z is used in reconstruction. This encourages a disentanglement of the information in the representation, with z_c (hopefully) containing all the information relevant for classification. Unlike the original CCVAE approach, we will focus on the single output setting with no further partitioning of z_c.

This split representation perspective is attractive for our purposes because it first allows for relatively strong pressure to be applied to z_c to be highly predictive of y. This means that we can use a relatively simple prediction head in our active learning loop that will hopefully have reliable reducible uncertainty estimates and be quick to update. Second, by also having an explicit representation for ostensibly non-label-relevant information, in the form of $z_{\setminus c}$, we are well placed to perform diagnostic checks for needing to update the encoder, e.g. by comparing the accuracy of the prediction head to a classifier trained with the full z. Finally, we found this to empirically give better downstream predictions than approaches where the classifier is used to guide the entire representation, e.g. Kingma et al. (2014).

We now describe other key algorithmic decisions, full details are provided in Appendix A.

Encoder training Unlike in the original CCVAE, we have no need to perform generations or interventions with our representation. We therefore eschew the introduction of an additional conditional generative model on $z_c|y$ and directly train the encoder and downstream classifier in an end-to-end manner using both the labeled and unlabeled data. Specifically, we maximize the following objective, corresponding to Equation (2) in Joy et al. (2021),

$$\mathcal{J}(\lambda, \psi, \omega) = \sum_{x \in \mathcal{D}_{\text{pool}}} \mathcal{L}(\lambda, \psi; x) + \sum_{(x,y) \in \mathcal{D}_{\text{labelled}}} \mathcal{L}(\lambda, \psi; x) + \alpha \mathbb{E}_{q_\lambda(z|x)} \left[\{c_\omega(z_c)\}_y \right] \tag{1}$$

where $\mathcal{L}(\lambda, \psi; x) = \mathbb{E}_{q_\lambda(z|x)} \left[\log \left(p_\psi(x \mid z) p(z) / q_\lambda(z \mid x) \right) \right]$ is the standard VAE objective, $q_\lambda(z \mid x)$ is the VAE encoder with parameters λ (and we take $g(x) = \mathbb{E}_{q_\lambda(z|x)}[z]$), $p_\psi(x \mid z)$ is the VAE decoder with parameters ψ, $p(z)$ is a fixed isotropic Gaussian prior, c_ω is the downstream classifier with parameters ω, $\mathcal{D}_{\text{pool}}$ is the unlabelled pool data, $\mathcal{D}_{\text{labelled}}$ is the labelled data gathered thusfar, and α is a hyperparameter controlling the label pressure on z_c.

Following Joy et al. (2021); Kingma et al. (2014), we perform the optimization using stochastic gradient ascent with minibatching, where updates with the labelled and unlabelled data are conducted in separate batches. As semi-supervised encoders typically struggle with class imbalance and the low-data regimes considered in active learning (Oliver et al., 2018; Yu et al., 2020; Guo et al., 2020), we further perform simple data augmentations on our labelled set and upsample minority classes. To deal with the redundant classes in our pool, we follow Bickford Smith et al. (2023, 2024) by labeling them as a single "redundant" category and retaining them in our labeled set, noting that these labels still contain useful information for future acquisition by marking points as not being one of the target classes.

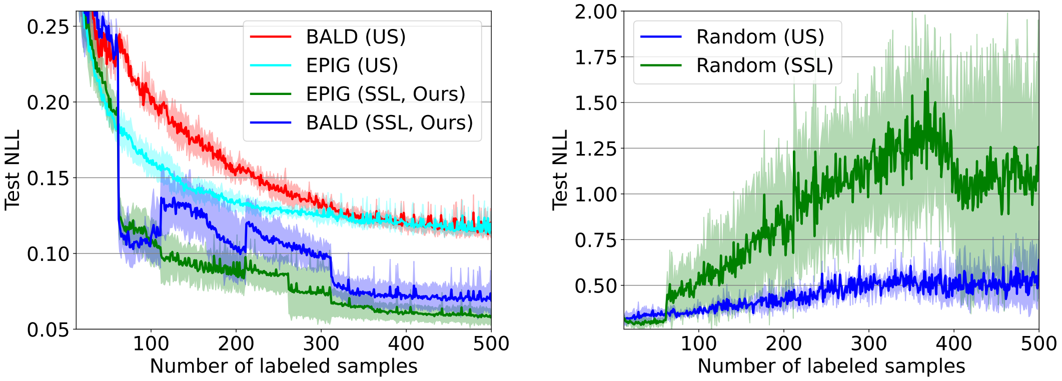
Classifier and prediction head While on the face of it the guidance classifier, c_ω, and the prediction head, $p_\phi(y|z, \theta_h)$, are both simply predictors for the output given there representation, their roles in our pipeline differ significantly. As such, their desirable characteristics also significantly differ and we generally recommend that they are chosen separately. The guidance classifier need not be probabilistic but must be differentiable. It is typically beneficial for it to have limited capacity and be smoothly varying in its inputs, as this forces the encoder to learn a z_c from which it is easy to make predictions. In our experiments, we use a simple neural network with one hidden layer of 128 units.

The prediction head, on the other hand, needs to be probabilistic with well-calibrated reducible uncertainty estimates. It will be updated at every iteration so it should ideally be cheap to train/update, and it should not require careful hyperparamter tuning or access to validation data. In our experiments, we use Random Forests (Breiman, 2001), due to their fast training and strong "out-of-the-box" performance.

Encoder retraining For simplicity, we retrain our encoder regularly after every k acquired labels. We recommend using larger values of k ($>= 50$) as this naturally suits many semi-supervised methods which only make significant gains once labels on the order of 10^2 have been acquired (Sohn et al., 2020; Joy et al., 2021; Chen et al., 2020b), while also keeping computational costs low. We also note that very small choices of k, and in particular taking $k = 1$, could in principle harm performance, by creating a disconnect between the update strategy assumed by the acquisition function (which is based only on the prediction head) and the actual updates performed (with the encoder also updated every kth step).

5. Experiments

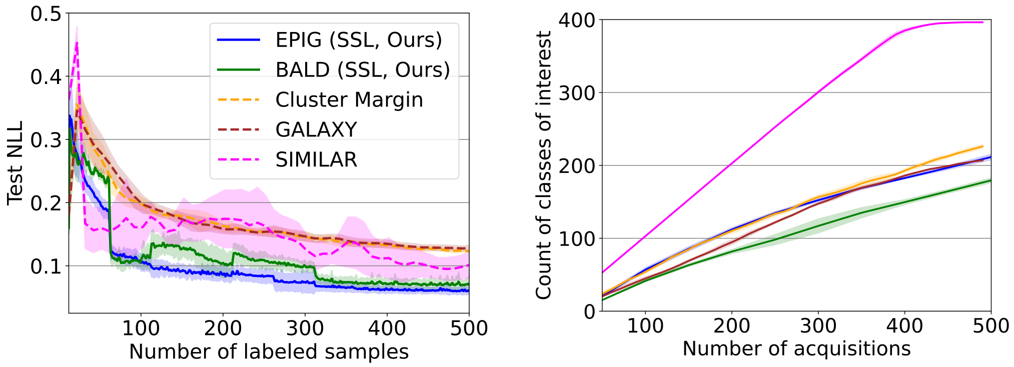
To validate our approach (**SSL**), we compare our approach to using unsupervised representations (**US**, based on a VAE encoder with matching architecture) on an adaptation of the MNIST dataset (Deng, 2012) with the BAL acquisition strategies BALD (Houlsby et al., 2011) and EPIG (Bickford Smith et al., 2023). We focus on messy pools which we create by introducing redundant labels and class imbalance. We choose our classes of interest as "5" and "6", introduce redundant labels by including the FashionMNIST dataset in our pool (Xiao et al., 2017), and use an extreme imbalance ratio of 300. We refer to this dataset as **F+MNIST**. We set $k = 50$ and use a budget of 500 labels. We run all experiments for 4 seeds. Full details of our experimental setup can be found in Appendix A.



(a) BAL

(b) Random

Figure 2: Test NLL for EPIG, BALD and random acquisition for **US**, **SSL** on **F+MNIST**.



(a) Test NLL

(b) Label counts

Figure 3: Test NLL for our **SSL** approach and baselines and counts for classes of interest.

Semi-supervised retraining improves both BALD and EPIG In Figure 2, we see that **SSL** improves active learning performance compared to **US** for both BALD and EPIG. In particular, we see a significant boost in model performance after the first retraining step, with more gains following in the remainder of the active learning. We also note that semi-supervised retraining with random acquisition performs terribly and does not result in gains after model retraining. This makes sense as random acquisition leads to mostly redundant labels, undermining the encoder's ability to incorporate task-relevant information.

Comparison with other methods We also compare **SSL** with baselines that have been designed specifically to deal with messy pools: **SIMILAR** (Kothawade et al., 2021), **Cluster Margin** (Citovsky et al., 2021), and **GALAXY** (Zhang et al., 2022). Figure 3 shows that our **SSL** approach comfortably outperforms them. In particular, our approach with the EPIG acquisition function performs the best, which is not surprising as EPIG makes explicit use of the predictive correlations in our model. Furthermore, examining the acquisition counts of the classes of interest, we find that the best performing approaches do not have the highest number of acquisitions for the classes of interest. This suggests that actively selecting our classes of interest, as done in approaches such as **SIMILAR** or **Cluster Margin**, does not always lead to improved active learning performance. This observation supports the idea that acquiring from redundant classes can, in fact, enhance classifier performance.

Acknowledgments

The authors are grateful to Freddie Bickford-Smith for helpful discussions. KA is supported by the EPSRC CDT in Modern Statistics and Statistical Machine Learning (EP/S023151/1) and TR is supported by the UK EPSRC grant EP/Y037200/1.

References

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8443–8452, 2021.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 7331–7348. PMLR, 2023.

Freddie Bickford Smith, Adam Foster, and Tom Rainforth. Making better use of unlabelled data in Bayesian active learning. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 847–855. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/bickford-smith24a.html.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Sophie Burkhardt, Julia Siekiera, and Stefan Kramer. Semi-supervised bayesian active learning for text classification. 2018. URL https://api.semanticscholar.org/CorpusID:202591938.

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2959–2968, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Kashyap Chitta, Jose M Alvarez, and Adam Lesnikowski. Large-scale visual active learning with deep probabilistic ensembles. *arXiv preprint arXiv:1811.03575*, 2018.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:236635499.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Zeyad Ali Sami Emam, Hong-Min Chu, Ping-Yeh Chiang, Wojciech Czaja, Richard Leapman, Micah Goldblum, and Tom Goldstein. Active learning at the imagenet scale. *arXiv preprint arXiv:2111.12880*, 2021.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*, 2022.

Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3897–3906. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/guo20i.html.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Tom Joy, Sebastian Schmon, Philip Torr, Siddharth N, and Tom Rainforth. Capturing label characteristics in {vae}s. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=wQRlSUZ5V7B.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *ArXiv*, abs/1406.5298, 2014. URL https://api.semanticscholar.org/CorpusID:6377199.

Andreas Kirsch. Black-box batch active learning for regression. *arXiv preprint arXiv:2302.08981*, 2023.

Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34:18685–18697, 2021.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 07 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.4.590. URL https://doi.org/10.1162/neco.1992.4.4.590.

Sudhanshu Mittal, Joshua Niemeijer, Jörg P Schäfer, and Thomas Brox. Best practices in active learning for semantic segmentation. In *DAGM German Conference on Pattern Recognition*, pages 427–442. Springer, 2023.

Sangwoo Mo, Jong-Chyi Su, Chih-Yao Ma, Mido Assran, Ishan Misra, Licheng Yu, and Sean Bell. Ropaws: Robust semi-supervised representation learning from uncurated data. *arXiv preprint arXiv:2302.14483*, 2023.

Siddharth Narayanaswamy, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank D. Wood, and Philip H. S. Torr. Learning disentangled representations with semi-supervised deep generative models. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, pages 5925–5935, 2017. URL http://dblp.uni-trier.de/db/conf/nips/nips2017.html#NarayanaswamyPM17.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

Ian Osband, Seyed Mohammad Asghari, Benjamin Van Roy, Nat McAleese, John Aslanides, and Geoffrey Irving. Fine-tuning language models via epistemic neural networks. *arXiv preprint arXiv:2211.01568*, 2022a.

Ian Osband, Seyed Mohammad Asghari, Benjamin Van Roy, Nat McAleese, John Aslanides, and Geoffrey Irving. Fine-tuning language models via epistemic neural networks. *arXiv preprint arXiv:2211.01568*, 2022b.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Xiuyuan Lu, Morteza Ibrahimi, Dieterich Lawson, Botao Hao, Brendan O'Donoghue, and Benjamin Van Roy. The neural testbed: Evaluating joint predictions. *Advances in Neural Information Processing Systems*, 35:12554–12565, 2022c.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Xiuyuan Lu, and Benjamin Van Roy. Evaluating high-order predictive distributions in deep learning. In *Uncertainty in Artificial Intelligence*, pages 1552–1560. PMLR, 2022d.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703, 2019. URL https://api.semanticscholar.org/CorpusID:202786778.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Seungmin Seo, Donghyun Kim, Youbin Ahn, and Kyong-Ho Lee. Active learning on pre-trained language model with task-independent triplet loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11276–11284, 2022a.

Seungmin Seo, Donghyun Kim, Youbin Ahn, and Kyong-Ho Lee. Active learning on pre-trained language model with task-independent triplet loss. In *AAAI Conference on Artificial Intelligence*, 2022b. URL https://api.semanticscholar.org/CorpusID:250291238.

Yuge Shi, Imant Daunhawer, Julia E Vogt, Philip HS Torr, and Amartya Sanyal. How robust is unsupervised representation learning to distribution shift? *arXiv preprint arXiv:2206.08871*, 2022.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

Yonglong Tian, Olivier J Henaff, and Aäron Van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10063–10074, 2021.

Chaoqi Wang, Shengyang Sun, and Roger Grosse. Beyond marginal uncertainty: How accurately can bayesian regression models estimate posterior predictive correlations? In *International Conference on Artificial Intelligence and Statistics*, pages 2476–2484. PMLR, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Tian Xie, Jifan Zhang, Haoyue Bai, and Robert Nowak. Deep active learning in the open world. *arXiv preprint arXiv:2411.06353*, 2024.

Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 438–454. Springer, 2020.

Jifan Zhang, Julian Katz-Samuels, and Robert Nowak. Galaxy: Graph-based active learning at the extreme. In *International Conference on Machine Learning*, pages 26223–26238. PMLR, 2022.

Appendix A. Experimental setup

A.1. Datasets

Figures 1, 4 To show that unsupervised representations break down with messy pools (Figure 4), we pretrained a VAE with the setup in Appendix A.2 on the **F+MNIST** dataset with increasing levels of "messiness" in the pool. This was done by creating an increasing amount of imbalance for our classes of interest ("5" and "6") vs. our redundant class. Table 1 shows the imbalance ratios we used for different levels of messiness, where the imbalance ratio is defined as

$$\text{Imbalance Ratio} = \frac{\max_{c \in \mathcal{C}} N_c}{\min_{c \in \mathcal{C}} N_c}$$

for N_c being the number of samples belonging to class c, and $\mathcal{C}$ the set of all classes in the dataset. We evaluated the quality of these representations by training a linear classifier on 5000 samples from classes "5" and "6", and evaluating it on 1000 samples from the same classes.

Messiness Level	Imbalance Ratio
Low	14
Medium	300
High	4000

Table 1: Messiness levels for **F+MNIST** and their corresponding imbalance ratios.

To show that the representations from increasingly messy pools lead to increasingly worse active learning performance (Figure 1) we took the representations pretrained on the messiness levels in Table 1 and used them for active learning with the setup described in Sections 4.1 and 5. For the active learning we used an imbalance ratio of 14 in the pool to make the comparison between the different representations fair and started with off with a labeled set of 12 samples, with 4 samples per class. We again evaluated on 1000 samples for "5" and "6".

Figures 2-3 For Figures 2-3 we used the "medium" messiness level from Table 1 for our pool. We started off with a labeled set of 12 samples, with 4 samples per class, and again evaluated on 1000 samples for "5" and "6".

A.2. Models

Prediction heads For our prediction heads, we used a random forest with 200 trees for all the experiments. During training, we trained on both the classes of interest and redundant classes and at inference we discarded the redundant classes and only focused on the predictions for the classes of interest.

Unsupervised encoders For our unsupervised encoders, we trained a VAE (Kingma et al., 2013) on all samples (labeled and unlabeled) before starting the active learning. We used a batch size of 64, the Adam optimizer and a learning rate of 0.005. We trained the encoders for 300 epochs.

Semi-supervised encoders For our semi-supervised encoders we used the model described in Section 4. We used a batch size of 64 for the unlabeled data and a batch size of 8 for the labeled data. We set $\alpha = 80$ to balance the supervised and unsupervised components of the loss. We again used the Adam optimizer, a learning rate of 0.005 and trained for 300 epochs. We used a 1-layer neural network with 128 hidden units as our guidance classifier and trained it using both the classes of interest and redundant classes. Moreover, when training the semi-supervised encoder, we extended our labeled set by using random rotations in $[0, \frac{\pi}{4}]$ and random scaling of the images between 60%-100% of their original size. We also upsampled the minority classes.

Following Joy et al. (2021), we partitioned our representation as $z = z_c \cup z_c = z_{i_1}, \ldots, z_{i_l} \cup z_{i_{l+1}}, \ldots, z_{i_d}$, where z_c denotes the subset used for classification, and the full z is used for reconstruction. By using only a subset for classification, we are able to more effectively compress the labeled information into z_c and achieve a better disentanglement between labeled and unlabeled information in z (since the loss no longer has to jointly optimize for both reconstruction and labeled performance). Additionally, this setup enables the possibility of more targeted and efficient active learning updates as downstream predictions can rely solely on the z_c component—though we do not explore this in the present work. For all our experiments, we set $|z_c| = 3$.

A.3. Active learning

For active learning, we compared the BALD Houlsby et al. (2011) and EPIG Bickford Smith et al. (2023) acquisition functions for a budget of 500 labels:

$$\begin{aligned}
\text{BALD}(x) &= \mathbb{E}_{p_\phi(y|x)}[\text{H}[p_\phi(\theta)] - \text{H}[p_\phi(\theta|x, y)]] \\
\text{EPIG}(x) &= \mathbb{E}_{p_*(x_*)p_\phi(y|x)}\left[\text{H}[p_\phi(y_*|x_*)] - \text{H}[p_\phi(y_*|x_*, x, y)]\right]
\end{aligned}$$

For EPIG, we use a target set of 500 points for each class of interest. For our approach, we retrained our semi-supervised encoder every 50 labels. We used a small validation set (90 labels) to evaluate the quality of the retrained encoder and decided to update the encoder if its loss on the validation set had improved. All experiments were ran for 4 seeds.

A.4. Baselines

For our baselines in Figure 3, we compared with other approaches in the literature that deal with the setting of messy pools. We focused on **SIMILAR** (Kothawade et al., 2021), **GALAXY** (Zhang et al., 2022), **Cluster Margin** (Citovsky et al., 2021). Similar to Zhang et al. (2022), we use the FLQMI relaxation of the submodular mutual information (SMI) for **SIMILAR**.

A.5. Implementation and compute resources

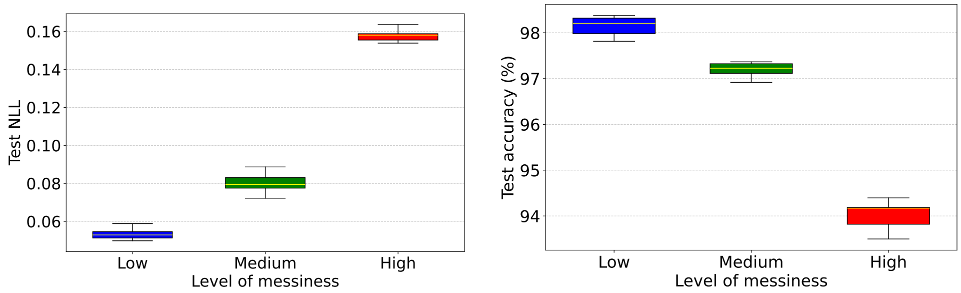
We ran all of our experiments on an NVIDIA H100 80GB GPU and used PyTorch (Paszke et al., 2019) for our implementations.

Appendix B. Related work

Previous work on semi-supervised BAL has focused mainly on the use of fixed unsupervised representations (Seo et al., 2022a; Osband et al., 2022a; Gleave and Irving, 2022; Hacohen et al., 2022; Mittal et al., 2023; Bickford Smith et al., 2024). Although some of these approaches, in particular those for natural-language-processing tasks, have been found to provide gains over random acquisition, their main focus has not been on the messy, uncurated pool setting that is common for real-world active learning. Hacohen et al. (2022) noted this and showed that their approach can fail with pools with class imbalance. At the same time, approaches that specifically consider the messy uncurated pool setting typically use pre-trained, unsupervised representations Citovsky et al. (2021); Emam et al. (2021); Zhang et al. (2022).

Moreover, there exists a lack of methods that look at incorporating both unlabeled and actively acquired labeled data into the model. Closest to our approach is Burkhardt et al. (2018) which uses a semi-supervised VAE trained with both labeled and unlabeled data. However, they do not consider re-training the encoder as more data is acquired and also focus only on well-curated datasets. Xie et al. (2024) uses a pretrained encoder and finetunes it after every acquisition step. They, however, work in the different setting of open-world active learning and do not consider the benefit of semi-supervised representations more broadly as we do in this work.

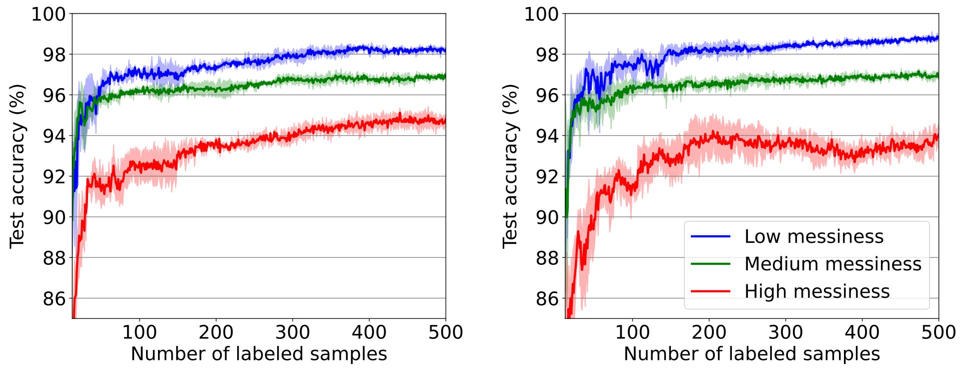
Appendix C. Additional plots



(a) NLL

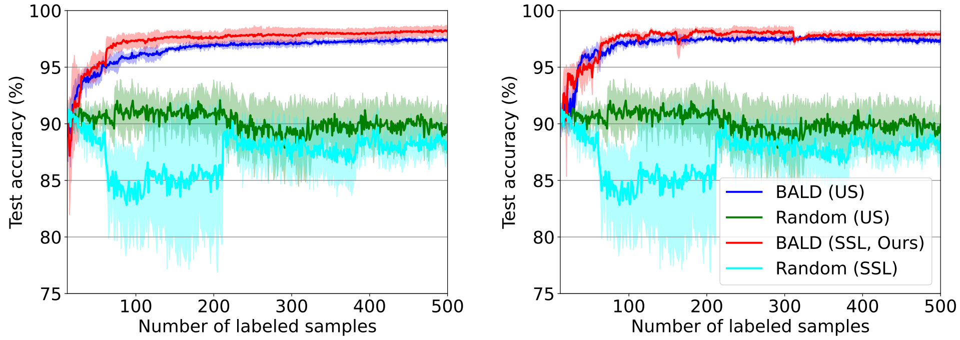
(b) Accuracy

Figure 4: Test NLL and accuracy of a linear classifier trained on pretrained representations, with pretraining performed on progressively messier **F+MNIST**. Messiness here refers to the amount of imbalance for the classes of interest vs. the redundant classes (see Appendix A).



(a) EPIG (b) BALD

Figure 5: Test accuracy on ”5” and ”6” for EPIG and BALD on F+MNIST where the pool becomes progressively messier by increasing the amount of imbalance for the classes of interest vs. redundant classes.



(a) EPIG (b) BALD

Figure 6: Test accuracy for EPIG, BALD and random acquisition for **US**, **SSL** on **F+MNIST**.

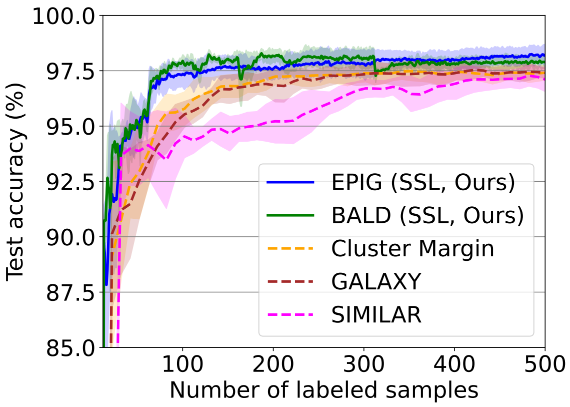


Figure 7: Test accuracy for our **SSL** approach and baselines.