MOLECULAR PROPERTY PREDICTION USING PRETRAINED-BERT AND BAYESIAN ACTIVE LEARNING: A DATA-EFFICIENT APPROACH TO DRUG DESIGN

Muhammad Arslan Masood
Aalto University, Finland
{arslan.masood}@aalto.fi

Samuel Kaski
Aalto University, Finland
University of Manchester, United Kingdom
{samuel.kaski}@aalto.fi

Tianyu Cui
Aalto University, Finland
Imperial College London, United Kingdom
{tianyu.cui}@aalto.fi

ABSTRACT

In drug discovery, prioritizing compounds for experimental testing is a critical task that can be optimized through active learning by strategically selecting informative molecules. Active learning typically trains models on labeled examples alone, while unlabeled data is only used for acquisition. This fully supervised approach neglects valuable information present in unlabeled molecular data, impairing both predictive performance and the molecule selection process. We address this limitation by integrating a transformer-based BERT model, pretrained on 1.26 million compounds, into the active learning pipeline. This effectively disentangles representation learning and uncertainty estimation, leading to more reliable molecule selection. Experiments on Tox21 and ClinTox datasets demonstrate that our approach achieves equivalent toxic compound identification with 50% fewer iterations compared to conventional active learning. Analysis reveals that pretrained BERT representations generate a structured embedding space enabling reliable uncertainty estimation despite limited labeled data, confirmed through Expected Calibration Error measurements. This work establishes that combining pretrained molecular representations with active learning significantly improves both model performance and acquisition efficiency in drug discovery, providing a scalable framework for compound prioritization.

1 INTRODUCTION

Active learning (AL) is a semi-supervised machine learning approach that selects new data points to be labeled in an iterative process. Starting with a small initial dataset, the model strategically identifies and requests labels for the most informative samples from a larger unlabeled pool. These newly labeled points are then incorporated into the training set, and the model is retrained, progressively improving its predictive accuracy through each iteration(Cohn et al., 1994). This iterative approach enables efficient model development with minimal labeled data, making it particularly valuable when labeling is expensive or time-consuming.

The effectiveness of active learning critically depends on accurate uncertainty estimation to guide the selection of informative training samples. In predictive modeling, two fundamental uncertainty types exist: epistemic uncertainty, arising from insufficient data coverage in chemical space, and aleatoric uncertainty, stemming from experimental measurement noise (Hüllermeier & Waegeman, 2021). Traditional approaches to uncertainty quantification in drug discovery like distance metrics and ensemble variance only capture epistemic uncertainty (Liu & Wallqvist, 2019; Lakshminarayanan et al., 2017), and auxiliary neural networks are needed for aleatoric uncertainty estimation (Hirschfeld et al., 2020). However, Bayesian frameworks offer a unified approach to capture both

uncertainty types in a principled manner (Kendall & Gal, 2017b; Zhang & Lee, 2019). Bayesian experimental design (BeD) formalizes the selection process by modeling uncertainties in predictions and using them to guide experimental choices. Several Bayesian acquisition functions have been developed to optimize the selection process. Bayesian Active Learning by Disagreement (BALD) selects samples that maximize information gain about model parameters (Houlsby et al., 2011), while Expected Predictive Information Gain (EPIG) prioritizes samples expected to most improve predictive performance (Smith et al., 2023).

However, the success of these Bayesian approaches fundamentally depends on the quality of molecular representations. Traditional quantitative structure-property relationships (QSPR) methods rely on handcrafted molecular descriptors (Cherkasov et al., 2014), which fail to capture complex chemical patterns, leading to poorly calibrated uncertainty estimates that compromise the effectiveness of even sophisticated Bayesian acquisition functions. While modern deep learning approaches, particularly graph neural networks and transformer-based architectures, overcome this through end-to-end representation learning (Cherkasov et al., 2014; Heid et al., 2024), they require large training datasets and are impractical in active learning scenarios that begin with limited data ($\approx$100 molecules).

To address these limitation, we leverage MolBERT (Fabian et al., 2020), a BERT-based model pretrained on 1.26 million compounds, by using its learned molecular representations as features in our active learning pipeline. This integration enables robust uncertainty estimation with limited labeled data, bridging the gap between deep learning capabilities and active learning constraints in drug discovery.

2 Materials and methods

2.1 Bayesian experimental design and active learning

Bayesian experimental design provides a principled framework for optimizing experiment utility (Rainforth et al., 2024). Let $\xi \in \Xi$ be the design in space Ξ and y be the experimental output with likelihood $p(y|\xi)$. The optimal design $\xi^\star$ is obtained by maximizing the expected utility $U(\xi, y)$:

$$\xi^\star = \arg\max_{\xi \in \Xi} \mathbb{E}_{y \sim p(y|\xi)} \left[U(\xi, y) \right], \tag{1}$$

where the expectation accounts for unobserved outcomes.

In active learning, we apply this framework to optimize the labeling process, starting with a small initial labeled set ($\approx$100 samples) and a large pool of unlabeled data. Consider a probabilistic model $f(\boldsymbol{x}; \phi)$ with likelihood $p(y|\boldsymbol{x}, \phi)$ for predicting molecular properties y from molecules $\boldsymbol{x}$, where ϕ has prior $p(\phi)$. Given a labeled dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, the posterior is $p(\phi|\mathcal{D}) \propto \prod_i^N p(y_i|\boldsymbol{x}_i, \phi) p(\phi)$. From the unlabeled pool $\mathcal{D}_u = \{(\boldsymbol{x}_i^u)\}_{i=1}^{N_u}$ (analogous to Ξ), we select the most informative sample $\boldsymbol{x}_s$ (analogous to ξ) to label, updating the posterior with the new observation $(\boldsymbol{x}_s^u, y_s)$.

The informativeness of unlabeled data points is defined by the acquisition function explained below:

Uniform (Random) Acquisition Function: The uniform(random) acquisition function randomly selects unlabeled data points with equal probability, serving as a baseline strategy. Specifically, for any unlabeled input $\boldsymbol{x} \in \mathcal{D}$, the uniform acquisition function is defined as:

$$\text{UNIFORM}(\boldsymbol{x}) = \frac{1}{|\mathcal{D}|}, \tag{2}$$

where $|\mathcal{D}|$ is the size of the pool dataset. While simple, this strategy provides an important baseline for comparing more sophisticated acquisition functions like BALD and EPIG, as it helps quantify the benefits of active learning over random sampling.

BALD Acquisition Function: Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011), which is the expected information gain, measured by the reduction in Shannon entropy of the model parameter ϕ from labeling $\boldsymbol{x}$ across all possible realizations of its label y given by $p(y|\boldsymbol{x}, \mathcal{D})$. Specifically, we have $\text{BALD}(\boldsymbol{x}) = \mathbb{E}_{y \sim p(y|\boldsymbol{x}, \mathcal{D})} \left[\text{H}[\phi|\mathcal{D}] - \text{H}[\phi|\boldsymbol{x}, y, \mathcal{D}] \right]$, which is usually intractable due to the high-dimensional posterior over the parameters. By observing the

equivalence between BALD and the conditional mutual information between the parameter and the unknown output, $\mathrm{I}[\phi, y|\boldsymbol{x}, \mathcal{D}]$, BALD can be rearranged to compute the information in the output space:

$$\mathrm{BALD}(\boldsymbol{x}) = \mathrm{I}[\phi, y|\boldsymbol{x}, \mathcal{D}] = \mathrm{H}[y|\boldsymbol{x}, \mathcal{D}] - \mathbb{E}_{\phi \sim p(\phi|\mathcal{D})} \left[\mathrm{H}[y|\boldsymbol{x}, \phi]\right] \tag{3}$$

with the optimal design $\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x}} \mathrm{BALD}(\boldsymbol{x})$. The first term in BALD measures the total uncertainty on the output y for its input $\boldsymbol{x}$ while the second term measures its aleatoric uncertainty, i.e., the irreducible uncertainty from observational noise. Therefore, BALD selects $\boldsymbol{x}$ with the highest epistemic uncertainty, i.e., the reducible uncertainty from the lack of data (Kendall & Gal, 2017a).

EPIG Acquisition Function: BALD targets global uncertainty reduction on the parameter space ϕ. However, in most supervised learning tasks, users are interested in improving the model accuracy on a target set $p(\boldsymbol{x}_*)$, e.g., the test set. Therefore, recent work (Smith et al., 2023) claimed that as acquisition function, Expected Predictive Information Gain (EPIG), explicitly reducing the model output uncertainty on random samples from $p(\boldsymbol{x}_*)$ is more effective than BALD in improving the model performance, defined as:

$$\mathrm{EPIG}(\boldsymbol{x}) = \mathbb{E}_{p(\boldsymbol{x}_*)} \left[\mathrm{H}[y_*|\boldsymbol{x}_*, \mathcal{D}] - \mathbb{E}_{p(y|\boldsymbol{x}, \mathcal{D})} \left[\mathrm{H}[y_*|\boldsymbol{x}_*, y, \boldsymbol{x}]\right]\right] \tag{4}$$

is *expected* reduction of the "expected predictive uncertainty" over the *target input distribution* $p(\boldsymbol{x}_*)$ by observing the label of $\boldsymbol{x}$. Intuitively, compared with BALD which reduces the parameter uncertainty globally, EPIG only reduces the parameter uncertainty that reduces model output uncertainty on $p(\boldsymbol{x}_*)$.

2.2 Semi-supervised active learning (SSAL)

Traditional supervised learning relies solely on labeled data, which is inefficient for active learning given the initially limited dataset. Learning a good input manifold for uncertainty estimation becomes particularly challenging in the complex chemical space (Zhou et al., 2019). While semi-supervised active learning approaches (Zhang et al., 2019; Hao et al., 2020) attempt to address this by leveraging both initial labeled-set and unlabeled pool-set, most public molecular datasets remain too small for effective representation learning.

In this paper, we propose to use molecular representations from a pretrained self-supervised learning model. Specifically, we encoded the molecular SMILES sequences into corresponding embeddings, utilizing a large transformer model MolBERT, pretrained on 1.26 million SMILES via masking, alongside physicochemical properties (Fabian et al., 2020). The embedding of each SMILES sequence is a pooled output from the pretrained MolBERT with dimension 764. We employed these embeddings from MolBERT to train a fully connected (i.e., MLP) head. This strategy allowed us to leverage a significant volume of molecule data, offering particular benefits for conducting active learning on relatively small datasets.

2.3 Baselines

We consider three acquisition functions, random, BALD, and EPIG (Section 2.1), and two learning paradigms, supervised active learning (AL) and semi-supervised active learning (SSAL). In SSAL, we use the BERT features pretrained on 1.26 million SMILES, and in AL, we use ECFP, or Extended-Connectivity Fingerprints, directly. ECFP is a method used in cheminformatics to represent molecular structures as binary fingerprints, capturing structural information by encoding the presence or absence of substructural features within a specified radius around each atom. Through iterative traversal of the molecular structure, unique substructural fragments are identified and hashed into a fixed-length bit vector, generating a binary fingerprint where each bit indicates the presence or absence of a specific substructural fragment. We encoded each molecule into a fixed 1024-dimensional binary vector using a radius of 4.

3 Results and discussion

For the Tox21 dataset, the impact of feature quality on active learning efficiency manifests distinctly across acquisition functions. BERT-EPIG demonstrates superior learning dynamics with a steeper improvement slope compared to ECFP-EPIG, indicating more efficient sample selection

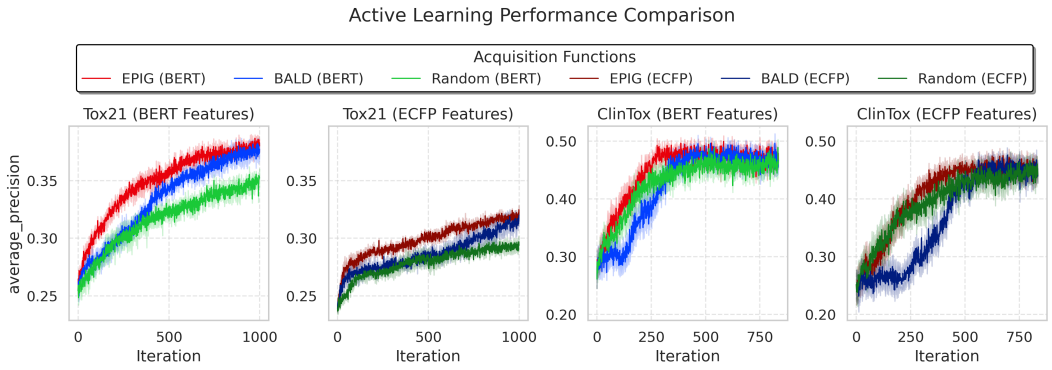

Figure 1: Active learning performance comparison on Tox21 and ClinTox datasets using BERT and ECFP molecular representations. BERT features consistently outperform ECFP, with EPIG showing superior sample selection over BALD and uniform sampling. Lines show mean performance (averaged across 12 tasks and 3 seeds for Tox21; 10 seeds for ClinTox) with 95% confidence intervals (shaded regions). Evaluation metric is average precision.

per iteration. The timing of separation from random baseline reveals feature quality's influence on uncertainty estimation - BERT-BALD diverges from random sampling at 400 iterations, while ECFP-BALD requires 600 iterations, underscoring how better features enable earlier identification of informative samples.

The ClinTox results further emphasize this pattern while revealing task-specific behaviors. BERT-EPIG achieves convergence significantly earlier (300 iterations) compared to ECFP-EPIG (600 iterations), demonstrating how high-quality representations accelerate learning. Notably, BALD underperforms random sampling in both feature spaces, aligning with previous findings about BALD's potential limitations in certain scenarios. These observations, combined with our UMAP visualization showing BERT's more structured embedding space, strongly support our hypothesis that effective active learning fundamentally depends on the quality of molecular representations enabling reliable uncertainty estimation.

3.1 Better Features Enable Better Uncertainty Estimation

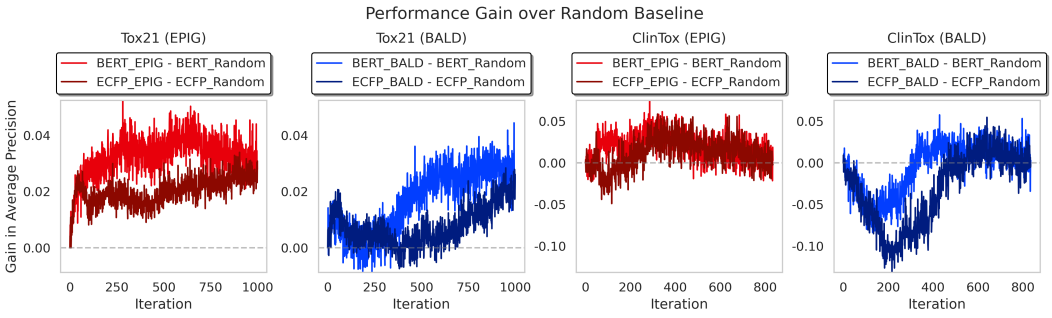

Figure 2: Performance gains of EPIG and BALD compared to random sampling baseline for Tox21 and ClinTox. BERT features (light colors) show consistently higher gains than ECFP (dark colors), with EPIG demonstrating more stable improvements than BALD across iterations. The y-axis shows the difference in average precision between each acquisition function and its corresponding random baseline (averaged across 12 tasks and 3 seeds for Tox21; 10 seeds for ClinTox).

Our experimental results reveal two key aspects of active learning performance: absolute gains from feature representations and relative gains from acquisition functions. Comparing absolute performance (Figure 1), BERT features consistently outperform ECFP, with BERT-EPIG achieving the highest average precision (0.38 for Tox21, 0.50 for ClinTox). While this superior performance

could stem from better feature quality, we demonstrate it primarily arises from improved uncertainty estimation.

To disentangle these factors, we analyzed relative gains over random sampling baselines (Figure 2). The steeper slope of BERT-EPIG's gain curve in early iterations (0-200) indicates more accurate uncertainty estimation, leading to efficient sample acquisition. BERT-EPIG achieves a maximum gain of 0.05 over BERT-Random in Tox21, compared to ECFP-EPIG's 0.02 gain over ECFP-Random. This disparity suggests BERT features not only provide better base performance but also enable more reliable uncertainty estimation for superior sample selection.

The acquisition function comparison further reveals EPIG's advantages over BALD. While BALD shows positive gains after 400-600 iterations, EPIG maintains consistent improvements from early stages. This difference is most pronounced in ClinTox, where ECFP-BALD initially degrades performance (-0.125) before recovery, while EPIG maintains stable gains. These findings demonstrate that successful molecular property prediction requires both high-quality representations and well-calibrated uncertainty estimation, with BERT-EPIG optimally combining both aspects.

3.2 Further Analysis

BERT features exhibit a more structured organization with distinct clusters of positive samples, enabling better-informed predictions about unlabeled samples, while ECFP representations display a scattered distribution with significant overlap between positive and negative regions, as demonstrated in Figure A.1. This structural difference manifests in sample acquisition efficiency, where BERT-EPIG identifies approximately 70% of toxic compounds within 400 iterations, compared to 600 iterations for BERT-BALD and 800 iterations for BERT-Uniform (Figure A.2, panel 3). The superiority of BERT-based approaches is further established through Expected Calibration Error (ECE) analysis, where BERT features demonstrate consistently better uncertainty calibration, particularly in early learning stages (Figure A.2, panels 1 and 2). While EPIG with BERT features achieves the fastest reduction in ECE, ECFP-based methods maintain elevated ECE values for longer periods and require substantially more labeled data to achieve comparable calibration. This comprehensive analysis reveals that the effectiveness of Bayesian acquisition functions fundamentally depends on well-structured molecular representations that enable reliable uncertainty estimation from limited training data, thus explaining the significant performance advantage of BERT-based approaches over ECFP in active learning scenarios.

4 Conclusion

Our study demonstrates that the success of active learning in molecular property prediction depends critically on the synergy between feature representations and acquisition functions. BERT features enable more effective uncertainty estimation compared to ECFP, as evidenced by faster ECE convergence and steeper learning curves. EPIG consistently outperforms BALD, maintaining stable improvements from early iterations across both datasets. The superior performance of BERT-EPIG stems from two key factors: (1) BERT's structured representation space, which clusters chemically similar compounds, facilitating reliable uncertainty estimation from limited data, and (2) EPIG's ability to leverage this structure for efficient sample acquisition, particularly in identifying rare positive samples. These findings highlight that successful active learning requires both high-quality molecular representations and well-calibrated uncertainty estimation. Future work should incorporate biological context during pretraining, leveraging protein-ligand interaction data and pathway information to enhance the model's ability to capture biologically relevant molecular features, leading to more robust uncertainty estimation that is crucial for active learning applications.

ACKNOWLEDGMENTS

This study was partially funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 956832 "Advanced Machine Learning for Innovative Drug Discovery". Further, this work was supported by the Academy of Finland Flagship program: the Finnish Center for Artifcial Intelligence FCAI. Samuel Kaski was supported by the UKRI Turing AI World-Leading Researcher Fellowship, [EP/W002973/1]

REFERENCES

Guy W. Bemis and Mark A. Murcko. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, January 1996. ISSN 0022-2623. doi: 10.1021/jm9602928. URL https://doi.org/10.1021/jm9602928. Publisher: American Chemical Society.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry*, 57(12):4977–5010, June 2014. ISSN 0022-2623. doi: 10.1021/jm4004285. URL https://doi.org/10.1021/jm4004285. Publisher: American Chemical Society.

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, May 1994. ISSN 1573-0565. doi: 10.1007/BF00993277. URL https://doi.org/10.1007/BF00993277.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks, November 2020. URL http://arxiv.org/abs/2011.13230. arXiv:2011.13230 [cs].

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1050–1059, 2016.

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

Kaitlyn M. Gayvert, Neel S. Madhukar, and Olivier Elemento. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. *Cell Chemical Biology*, 23(10):1294–1301, October 2016. ISSN 24519456. doi: 10.1016/j.chembiol.2016.07.023. URL https://linkinghub.elsevier.com/retrieve/pii/S2451945616302914.

Zhongkai Hao, Chengqiang Lu, Zhenya Huang, Hao Wang, Zheyuan Hu, Qi Liu, Enhong Chen, and Cheekong Lee. Asgn: An active semi-supervised graph neural network for molecular property prediction. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 731–752, 2020.

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A Machine Learning Package for Chemical Property Prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17, January 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01250. URL https://doi.org/10.1021/acs.jcim.3c01250. Publisher: American Chemical Society.

Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W. Coley. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, August 2020. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.0c00502. URL https://pubs.acs.org/doi/10.1021/acs.jcim.0c00502.

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

Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-021-05946-3. URL http://arxiv.org/abs/1910.09457. arXiv:1910.09457 [cs].

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017a.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, October 2017b. URL http://arxiv.org/abs/1703.04977. arXiv:1703.04977 [cs].

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, November 2017. URL http://arxiv.org/abs/1612.01474. arXiv:1612.01474 [stat].

Ruifeng Liu and Anders Wallqvist. Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds. *Journal of Chemical Information and Modeling*, 59(1):181–189, January 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00597. URL https://doi.org/10.1021/acs.jcim.8b00597. Publisher: American Chemical Society.

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

Vineeth Rakesh and Swayambhoo Jain. Efficacy of bayesian neural networks in active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2601–2609, 2021.

Ann M. Richard, Ruili Huang, Suramya Waidyanatha, Paul Shinn, Bradley J. Collins, Inthirany Thillainadarajah, Christopher M. Grulke, Antony J. Williams, Ryan R. Lougee, Richard S. Judson, Keith A. Houck, Mahmoud Shobair, Chihae Yang, James F. Rathman, Adam Yasgar, Suzanne C. Fitzpatrick, Anton Simeonov, Russell S. Thomas, Kevin M. Crofton, Richard S. Paules, John R. Bucher, Christopher P. Austin, Robert J. Kavlock, and Raymond R. Tice. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chemical Research in Toxicology*, 34(2):189–216, February 2021. ISSN 0893-228X, 1520-5010. doi: 10.1021/acs.chemrestox.0c00264. URL https://pubs.acs.org/doi/10.1021/acs.chemrestox.0c00264.

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*, pp. 7331–7348. PMLR, 2023.

Yao Zhang and Alpha A. Lee. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical Science*, 10(35):8154–8163, 2019. ISSN 2041-6520, 2041-6539. doi: 10.1039/C9SC00616H. URL http://xlink.rsc.org/?DOI=C9SC00616H.

Yao Zhang et al. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical science*, 10(35):8154–8163, 2019.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):10752, 2019.

A APPENDIX

A.1 DATASETS

Tox21: The Tox21 dataset, or Toxicology in the 21st Century dataset, is a publicly available dataset used in the field of computational toxicology (Richard et al., 2021). The Tox21 dataset consists of a large collection of chemical compounds, each of which is associated with various types of toxicity outcomes. These outcomes are typically measured using high-throughput screening assays to evaluate the potential toxic effects of the compounds. The dataset provides a quantitative assessment (in form of binary labels) of toxicity of $\approx$ 8000 compounds in 12 different toxicity pathways. The Tox21 dataset is widely used as a benchmark in the development of in silico toxicology models. In this dataset, 6.24% measurements are active (ranges from 2% to 12%), 73% are inactive, while 20.56% are missing values.

ClinTox: The ClinTox dataset (Gayvert et al., 2016) combines data from two distinct sources: FDA-approved drugs and drugs that failed clinical trials due to toxicity. It contains information for 1,484 compounds with binary labels. The dataset provides valuable insights into the relationship between chemical structures and drug safety profiles in human clinical trials.

A.2 DATA SPLITTING

Test, Train set: For the better of evaluation of generalization, we employed scaffold splitting with 80:20 ratio to create distinct training and testing sets. Scaffold splitting partitions a molecular dataset according to core structural motifs identified by the Bemis-Murcko scaffold representation (Bemis & Murcko, 1996), prioritizing larger groups while ensuring that the train and test sets do not share identical scaffolds. The test set is identical for all the experiments.

Initial and Pool Sets: A balanced initial set was constructed by randomly selecting 100 molecules from the training set, with equal representation of positive and negative instances. Subsequently, a pool set was generated by excluding the initial set from the training set.

A.3 PRACTICAL BAYESIAN NEURAL NETWORKS

In this work, we use a Bayesian neural network to account for the model uncertainty. Previous studies on dropout variational inference (Gal & Ghahramani, 2016) suggest that a practical Bayesian neural network for a wide variety of architectures can be obtained by simply training a neural network with dropout (MC dropout), and interpreting this as being equivalent to variational inference (Blei et al., 2017). The uncertainty is then estimated by using multiple forward-passes with different dropout masks. Although the uncertainty from MC dropout is often underestimated, it has been a popular choice for Bayesian active learning with neural networks and shows promise on real-world datasets (Gal et al., 2017; Rakesh & Jain, 2021).

This neural network uses $\boldsymbol{x}_0$ initialized as the input features $\boldsymbol{x}$, which can be either BERT features (in the semi-supervised AL) or binary fingerprints (in the supervised AL). We utilize dropout for uncertainty estimation, batch normalization for training stability, and the rectified linear unit (ReLU) activation function as the default activation. Additionally, the network incorporates a skip connection, merging the input and output of the hidden layer, enhancing information flow. Finally, the output layer generates logits, which can be transformed into probabilities by passing through a sigmoidal activation function.

$$
\begin{aligned}
&\boldsymbol{x}_0 = \boldsymbol{x} \quad \texttt{BERT features or ECFP} \\
&\boldsymbol{x}_\ell = \text{Dropout}(\text{ReLU}(\text{BatchNorm}(W_\ell \boldsymbol{x}_0 + \mathbf{b}_\ell))) \\
&\tilde{\boldsymbol{x}}_{\ell+1} = \text{BatchNorm}(W_{\ell+1} \boldsymbol{x}_\ell + \mathbf{b}_{\ell+1}) \\
&\boldsymbol{x}_{\ell+1} = \text{Dropout}(\text{ReLU}(\boldsymbol{x}_\ell + \tilde{\boldsymbol{x}}_{\ell+1})) \\
&x_{out} = W_{\ell+2} \boldsymbol{x}_{\ell+1} + \mathbf{b}_{\ell+1}
\end{aligned}
\tag{5}
$$

The hyper-parameters of this model are given in table 1.

Approximating acquisition functions: In practice, the posterior $p(\phi|\mathcal{D})$ is intractable, but we can approximate each of the acquisition functions using an approximated distribution $q(\phi)$, such as the dropout distribution (Gal & Ghahramani, 2016) used in Section A.3. Specifically, for BALD, the acquisition function can be rewritten as:

$$\begin{aligned}&\mathrm{BALD}(\boldsymbol{x}) = \mathrm{H}[y|\boldsymbol{x},\mathcal{D}] - \mathbb{E}_{\phi\sim p(\phi|\mathcal{D})}\left[\mathrm{H}[y|\boldsymbol{x},\phi]\right] \\ &= -\sum_{c\in\{0,1\}} p(y=c|\boldsymbol{x},\mathcal{D})\log p(y=c|\boldsymbol{x},\mathcal{D}) + \mathbb{E}_{q(\phi)}\left[\sum_{c\in\{0,1\}} p(y=c|\boldsymbol{x},\phi)\log p(y=c|\boldsymbol{x},\phi)\right],\end{aligned} \tag{6}$$

where c is the class label that y can take and $p(y=c|\boldsymbol{x},\mathcal{D}) \approx \mathbb{E}_{q(\phi)}\left[p(y=c|\boldsymbol{x},\phi)\right]$.

For EPIG (Smith et al., 2023), first we observe

$$\mathrm{EPIG}(\boldsymbol{x}) = \mathbb{E}_{p(\boldsymbol{x}_*)}\left[\mathrm{KL}\left[p(y,y_*|\boldsymbol{x},\boldsymbol{x}_*,\mathcal{D})|p(y|\boldsymbol{x},\mathcal{D})p(y_*|\boldsymbol{x}_*,\mathcal{D})\right]\right], \tag{7}$$

where $p(y|\boldsymbol{x},\mathcal{D}) \approx \mathbb{E}_{q(\phi)}\left[p(y|\boldsymbol{x},\phi)\right]$ and $p(y,y_*|\boldsymbol{x},\boldsymbol{x}_*,\mathcal{D}) \approx \mathbb{E}_{q(\phi)}\left[p(y|\boldsymbol{x},\phi)p(y_*|\boldsymbol{x}_*,\phi)\right]$.

All expectations in above acquisition functions can be approximated with Monte Carlo sampling. For example, with T samples from $q(\phi)$:

$$\mathbb{E}_{q(\phi)}\left[p(y|\boldsymbol{x},\phi)\right] \approx \frac{1}{T}\sum_{t=1}^{T} p(y|\boldsymbol{x},\phi^{(t)}), \tag{8}$$

where $\phi^{(t)} \sim q(\phi)$.

A.4 Analysis of Learned Representations

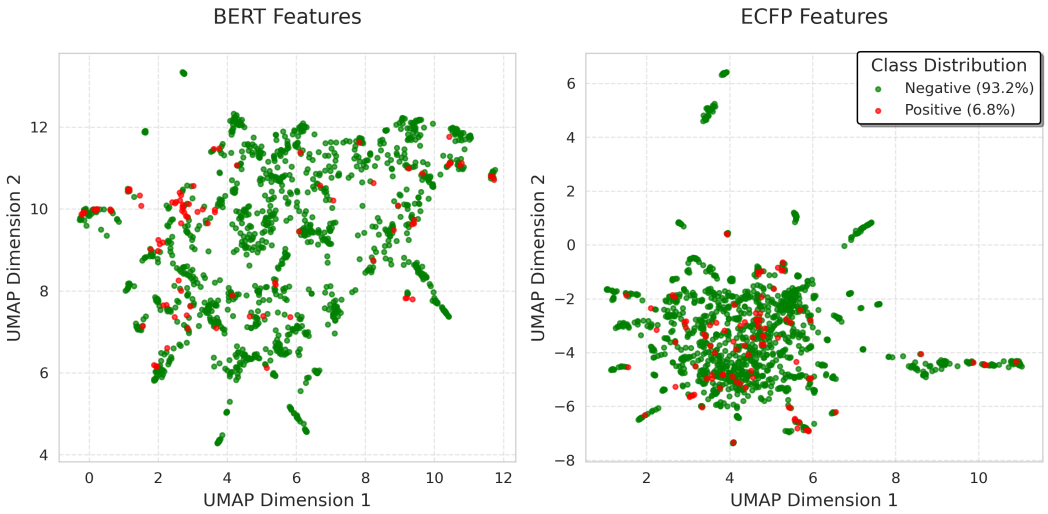

Figure A.1: UMAP visualization of molecular features projected into 2D space, BERT (left) and ECFP (right). The points represent individual molecules colored by their class labels (red for positive, green for negative)

To understand why BERT-based approaches significantly outperform ECFP in active learning, we visualized both representation spaces using UMAP dimensionality reduction (Figure A.1). The BERT features exhibit more structured organization, where positive samples (red points, 6.8% of dataset) are distributed in distinct clusters, indicating that semantically similar molecules are mapped to nearby regions. This structured manifold enables the model to make better-informed predictions about unlabeled samples based on their proximity to labeled examples, even with limited initial training data.

In contrast, ECFP representations show a more scattered distribution with significant overlap between positive and negative regions, making it difficult for the model to learn meaningful patterns

from small initial labeled sets. This poorly structured space leads to unreliable uncertainty estimates, explaining why ECFP-based Bayesian acquisition functions (BALD, EPIG) show only marginal improvement over uniform sampling. The visualization results support our finding that the effectiveness of uncertainty-based active learning methods critically depends on having well-structured molecular representations that enable reliable uncertainty estimation from limited training data.

A.5 Analysis of Sample Acquisition Patterns

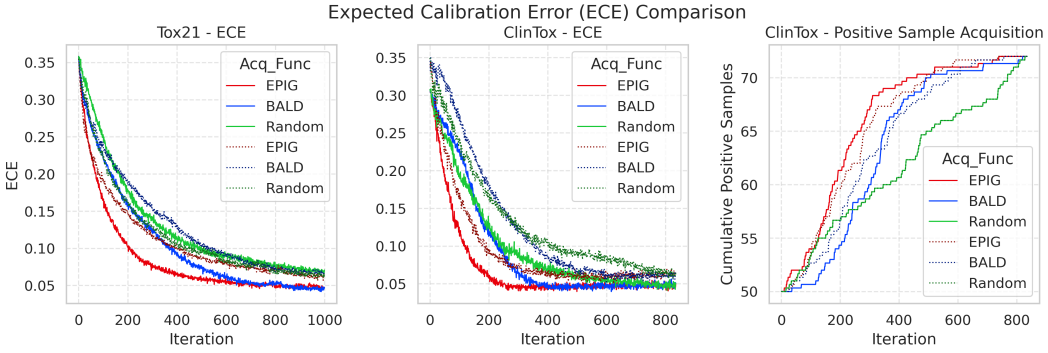

Figure A.2: (Panel 1) Evolution of Expected Calibration Error (ECE) for Tox21 (averaged across 12 tasks and 3 seeds)
(Panel 2) Evolution of ECE for ClinTox (averaged across 10 seeds). Lower ECE indicates better-calibrated uncertainty estimates. EPIG with BERT features (solid red) achieves the fastest convergence to low ECE values, demonstrating superior uncertainty calibration.
(Panel 3) Comparison of positive sample acquisition rates across different feature representations and acquisition functions on the ClinTox dataset. The plot shows cumulative toxic compound identification starting from a balanced initial set (50 positive, 50 negative). BERT representations with uncertainty-based acquisition (EPIG, BALD) identify positive samples more efficiently compared to uniform sampling and ECFP-based approaches, demonstrating better exploration of the chemical space when starting with limited labeled data.

To further understand why BERT representations enable more effective active learning, we analyzed the cumulative acquisition of positive samples (toxic compounds) across iterations (Figure A.2, panel 3). Starting from a balanced initial set (50 positive, 50 negative samples), the acquisition patterns reveal key differences between BERT and ECFP approaches in handling the significant class imbalance present in the pool set (22 positive out of 835 samples).

BERT-EPIG shows the most efficient acquisition rate for positive samples, identifying approximately 70% of toxic compounds within 400 iterations, compared to 600 iterations for BERT-BALD and 800 iterations for BERT-Uniform. This accelerated discovery of minority class samples aligns with the structured representation space observed in UMAP visualization, where BERT features organize molecules into meaningful clusters that facilitate identification of informative toxic compounds.

Interestingly, while ECFP-EPIG initially shows comparable acquisition rates to BERT-EPIG, its performance plateaus earlier, suggesting that the scattered representation space limits its ability to make reliable uncertainty estimates as learning progresses. ECFP-BALD exhibits similar limitations, highlighting that even sophisticated Bayesian acquisition functions struggle when the underlying representation space lacks clear structure for learning from limited initial data.

To further investigate why Bayesian acquisition functions might underperform with ECFP features, we analyzed the Expected Calibration Error (ECE) throughout the active learning process. ECE measures the difference between model confidence and actual accuracy, with lower values indicating better-calibrated uncertainty estimates. Figure A.2 (panel 1 and 2) shows the evolution of ECE across different feature types and acquisition functions.

The results reveal a clear relationship between feature quality and uncertainty estimation. All methods initially exhibit high ECE (0.30-0.38), indicating poor calibration due to limited training data.

However, the BERT-based approaches demonstrate consistently lower ECE compared to their ECFP counterparts throughout the early stages of active learning (iterations 0-200). This aligns with our previous observation that ECFP features lead to less reliable uncertainty estimates, which in turn compromises the effectiveness of Bayesian acquisition functions like BALD and EPIG.

Particularly noteworthy is the EPIG acquisition function with BERT features, which achieves the fastest reduction in ECE (solid red line), suggesting it learns well-calibrated uncertainties more efficiently. This explains its superior performance in the main task, as shown in Figure 1. In contrast, ECFP-based methods maintain higher ECE for a longer period, indicating persistent struggles in uncertainty estimation despite sophisticated acquisition strategies.

While all methods eventually converge to well-calibrated uncertainties (ECE ¡ 0.1) after 600-800 iterations, the path to achieving good calibration is markedly different. ECFP-based approaches require substantially more labeled data to achieve comparable calibration, which is particularly problematic in the active learning setting where labeled data is initially scarce. This finding reinforces our hypothesis that the success of Bayesian acquisition functions is fundamentally limited by the quality of input representations and their ability to enable reliable uncertainty estimation from limited training data.

	Hyperparameter	Values
BNN	Activation	[ReLU]
	Batch normalization	[True]
	Skip connection	[True]
	Input layer	[768, 1024]
	hidden layer dim	[128]
	Number of hidden layers	[1]
	Dropout probability	[0.3]
Training	Optimizer	[Adam]
	Learning rate	$[10^{-3}]$
	Weight decay	[1e-2]
	Scheduler	[CosineAnnealingLR]
	T-max (LR cycle)	[10]
	Batch size	[16]
	Epochs	[110]
	num. Forward pass	[20]

Table 1: Hyperparameters used of BNN and training