Data-Efficient Drug Design with Pretrained-BERT and Bayesian Active Learning

Muhammad Arslan Masood[1], *Samuel Kaski*[1,2], *Tainyu Cui*[1,3],

[1] *Department of Computer Science, Aalto University, Finland*
[2] *University of Manchester, United Kingdom*
[3] *Imperial College London, United Kingdom*
arslan.masood@aalto.fi

Aalto University
School of Science

Abstract

In drug discovery, selecting the right compounds to test is critical. Traditional active learning relies only on labeled data, ignoring the rich information in unlabeled molecules. We change that by integrating a BERT model pretrained on 1.26 million unlabeled compounds into the active learning pipeline, enabling robust molecular representations. This significantly improves uncertainty estimation, allowing us to identify toxic compounds with 50% fewer iterations—faster, smarter, and more efficient.

Research Questions

- How to perform active learning with limited datasets?
- How to disentangle representation learning from uncertainity estimation?
- How to perform target domain aware active learning?

SS-Active learning

- Molecular Representations of SMILES s_i of both labeled $\mathcal{D}_{\text{labeled}} = \{(s_i^{\text{labeled}}, y_i)\}_{i=1}^{N}$ and unlabeled set $\mathcal{D}_u = \{s_i^u\}_{i=1}^{N_u}$ are generated by using pretrained MolBERT model:

$$\mathbf{x}_i^{\text{labeled}} = \text{MolBERT}(s_i^{\text{labeled}}; \Theta^*_{\text{pretrain}})$$
$$\mathbf{x}_i^{u} = \text{MolBERT}(s_i^{u}; \Theta^*_{\text{pretrain}})$$
$$\mathbf{x}_i^{\text{labeled}}, \mathbf{x}_i^{u} \in \mathbb{R}^d$$

- The labelled set $\mathcal{D}_{\text{labeled}} = \{(\mathbf{x}_i^{\text{labeled}}, y_i)\}_{i=1}^{N}$ is used to train a probabilistic model $f(\mathbf{x}_i^{\text{labeled}}; \phi)$ with parameters ϕ

$$p(\phi|\mathcal{D}_{\text{labeled}}) \propto p(\mathcal{D}_{\text{labeled}}|\phi)p(\phi)$$

- A new data point is selected from the unlabeled set $\mathcal{D}_u = \{(\mathbf{x}_i^u)\}_{i=1}^{N_u}$ by maximizing the acquisition function:

$$\mathbf{x}_s^u = \arg\max_{\mathbf{x} \in \mathcal{D}_u} a(\mathbf{x})$$

- The label y_s is acquired for $\mathbf{x}_s^u$ and incorporated into the training set

$$\mathcal{D}_{\text{updated}} = \mathcal{D} \bigcup \{(\mathbf{x}_s^u, y_s)\}$$

- The posterior of probablisitc model is updated by using $\mathcal{D}_{\text{updated}}$

$$p(\phi|\mathcal{D}_{\text{updated}}) \propto p(\mathcal{D}_{\text{updated}}|\phi)p(\phi)$$

Acquisition functions

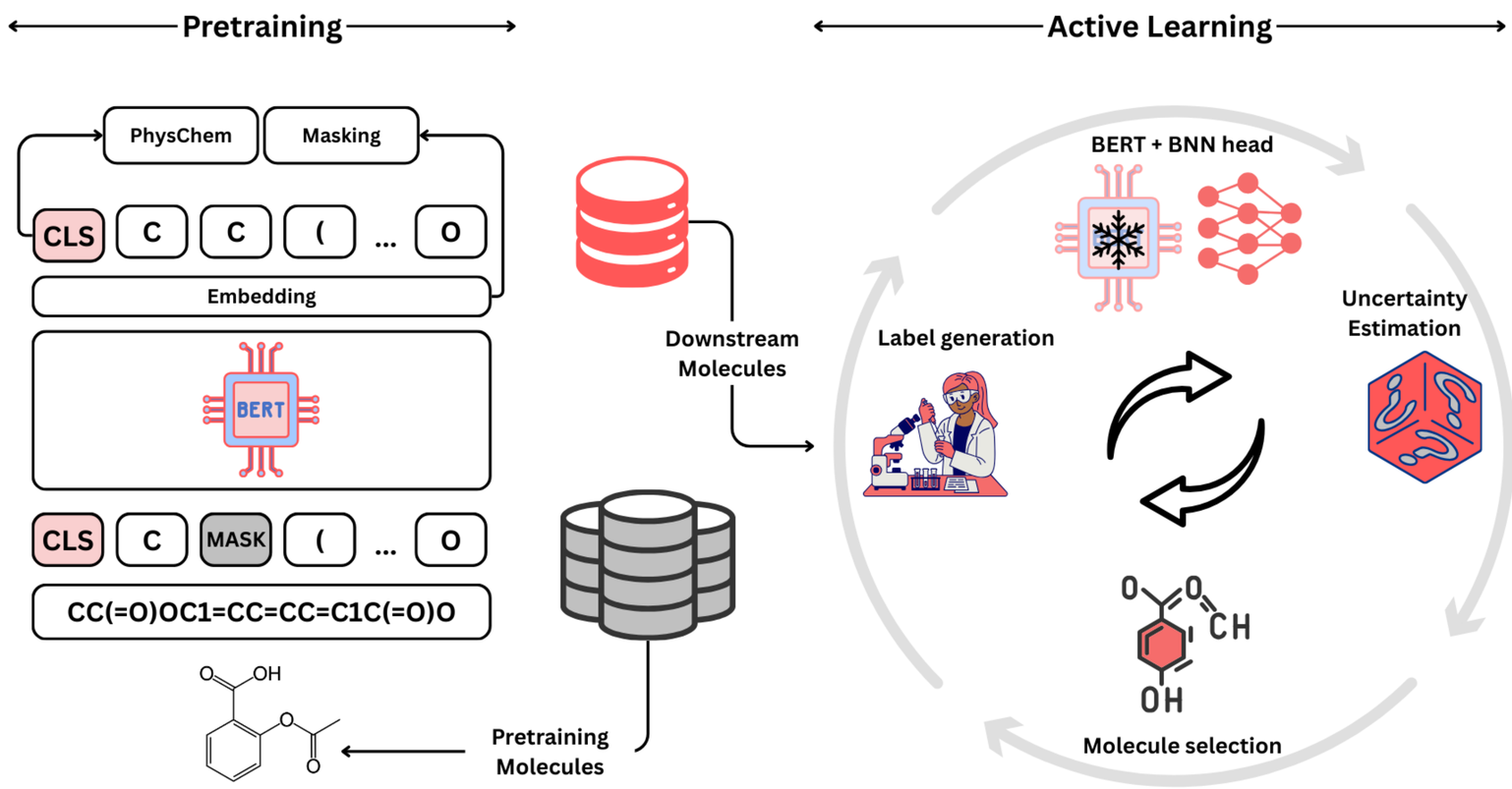
- Random Acquisition $= \mathbf{x}_s^u \sim \text{Uniform}(\mathcal{D}_u)$
- BALD

$$= \mathbb{E}_{y \sim p(y|\mathbf{x}, \mathcal{D})} \left[\text{H}[\phi|\mathcal{D}] - \text{H}[\phi|\mathbf{x}, y, \mathcal{D}] \right]$$
$$= \text{H}[y|\mathbf{x}, \mathcal{D}] - \mathbb{E}_{\phi \sim p(\phi|\mathcal{D})} \left[\text{H}[y|\mathbf{x}, \phi] \right]$$

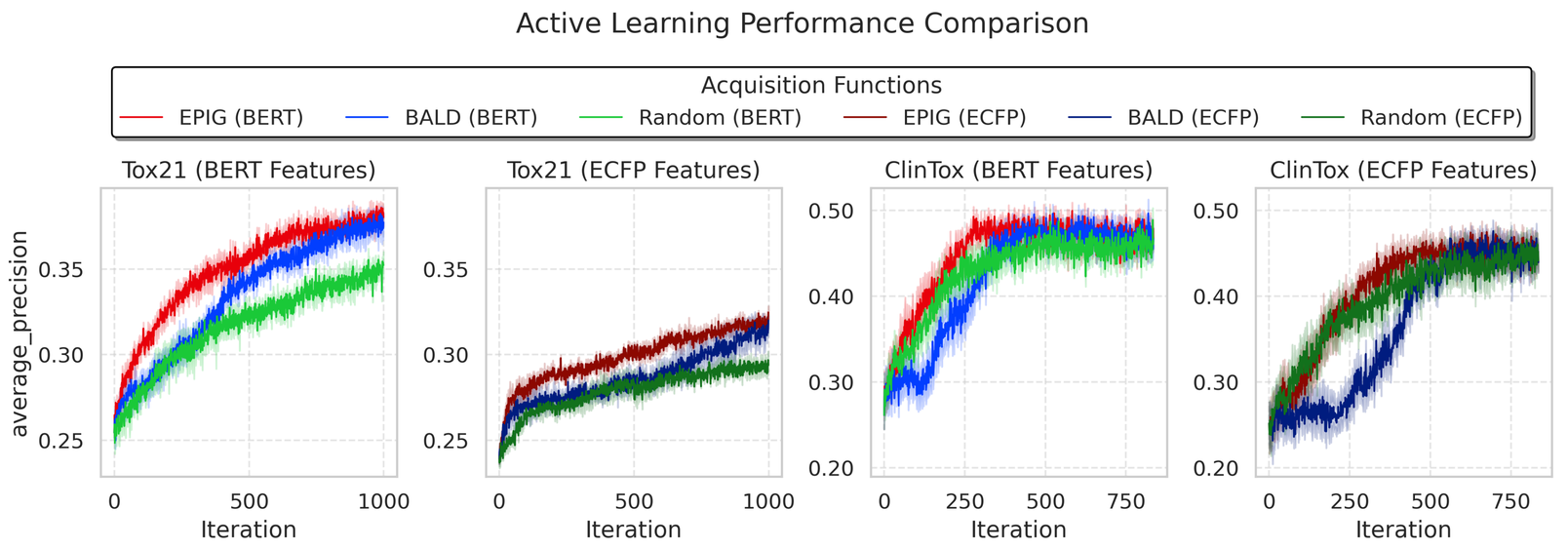
- EPIG

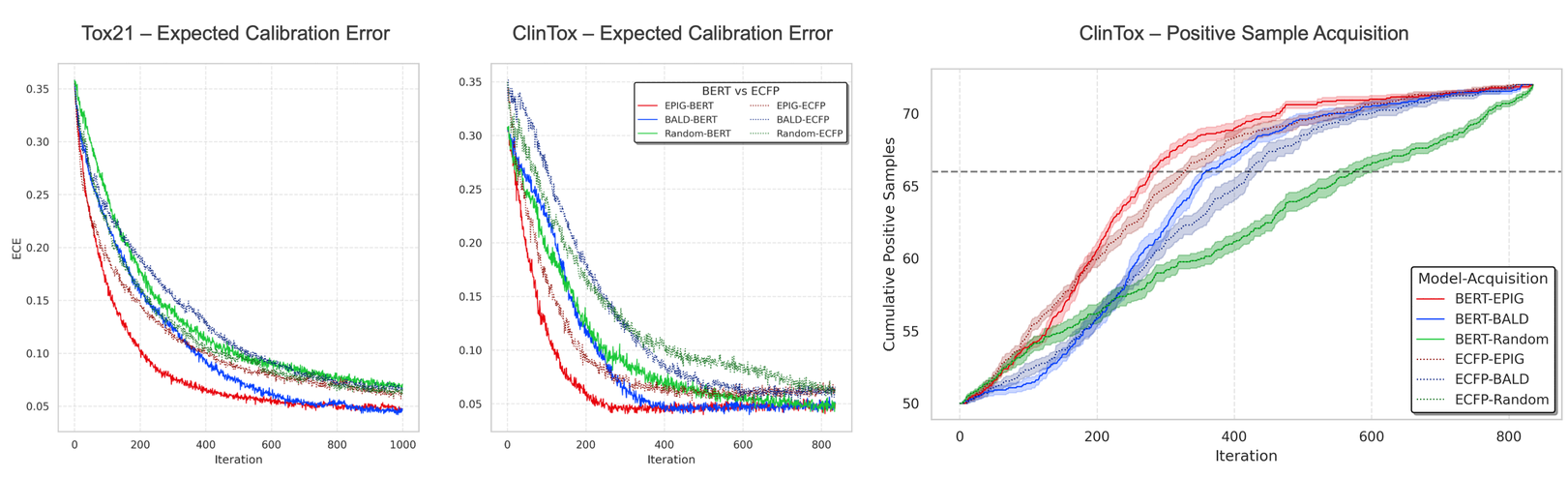
$$= \mathbb{E}_{p(\mathbf{x}_*)} \left[\text{H}[y_*|\mathbf{x}_*, \mathcal{D}] - \mathbb{E}_{p(y|\mathbf{x}, \mathcal{D})} \left[\text{H}[y_*|\mathbf{x}_*, y, \mathbf{x}] \right] \right]$$

Active Learning



Results





Acknowledgements

This project is funded by Horizon2020 research and innovation programme under the MSC-ITN actions, grant agreement No. 956832.