Appendix for BMT-EpiPred: Bayesian Multitask Learning for Reliable Prediction of Small-Molecule Epigenetic Modulators

APPENDX A

ALGORITHMS

A.1 Algorithms of Optimizing BMT-EpiPred model

Algorithm 1 delineates the optimization procedure of our BMT-EpiPred model, where it should be noted that the encoder and the distribution parameters of stochastic task heads are updated synchronously during this process.

Algorithm 1 Optimizing BMT-EpiPred model

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1\sim N}$, learning rate lr

Output: Shared representation encoder $\boldsymbol{\theta}$ and distribution parameters $\mathbf{m}_t$ and $\mathbf{S}_t$ of task specific head $t(t = 1, \dots, \mathrm{T})$

Optimizing:

1: Randomly initialize shared encoder and variational distribution parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_t = \{\mathbf{m}_t, \mathbf{S}_t\}_{t=1,\dots,\mathrm{T}}$, set task-specific prior $p(\mathbf{W}_t) = \mathcal{N}(\mathbf{0}, \mathbf{I})(t = 1, \dots, \mathrm{T})$
2: **for** $i = 1, \dots, K$ **do**
3: Sample mini-batch of data $\mathcal{D}_i \sim \mathcal{D}$ containing B samples
4: Calculate hidden representation $\mathbf{h}_i = \boldsymbol{\theta}^{\top}\mathbf{x}_i$
5: **for** $t = 1, \dots, \mathrm{T}$ **do**
6: Calculate task loss

$$\mathcal{L}_t(\boldsymbol{\theta}, \boldsymbol{\phi}_t, \boldsymbol{\Sigma}) = \frac{1}{\gamma}\sum_{i=1}^{B}\left(\mathbf{y}_i^{\top}\mathbf{m}_t\mathbf{h}_i - LSE_j\left(\mathbf{m}_t^{(j)^{\top}}\mathbf{h}_i + \frac{1}{2}\left(\mathbf{h}_i^{\top}\mathbf{S}_t^{(j)}\mathbf{h}_i + \sigma_t^{(j)^2}\right)\right)\right) - \frac{1}{\gamma}\sum_{i=1}^{N_{\mathbf{y}}} KL\left(q(\mathbf{W}_t|\mathbf{m}_t, \mathbf{S}_t)||p(\mathbf{W}_t)\right)$$

7: **end for**
8: Calculate the total loss of all tasks

$$\mathcal{L} = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\mathcal{L}_t$$

where $\mathcal{L}_t = \begin{cases} -\frac{N_i}{N_a}\sum_{n=1}^{N}\mathcal{L}_t^{(n)}(\boldsymbol{\theta}, \boldsymbol{\phi}_t, \boldsymbol{\Sigma}), active \\ -\sum_{n=1}^{N}\mathcal{L}_t^{(n)}(\boldsymbol{\theta}, \boldsymbol{\phi}_t, \boldsymbol{\Sigma}), inactive \end{cases}$

9: Optimize parameters $\boldsymbol{\theta} = \boldsymbol{\theta} - lr \times \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$ and $\boldsymbol{\phi}_t = \boldsymbol{\phi}_t - lr \times \frac{\partial \mathcal{L}}{\partial \boldsymbol{\phi}_t}(t = 1, \dots, \mathrm{T})$
10: **end for**

A.2 Algorithms of Optimizing BMT-EpiPred model

Algorithm 2 delineates the optimization procedure of optimizing Bayesian models by marginal likelihood, where it should be noted that the updates of the encoder and the distribution parameters of the stochastic task head are not synchronized during this optimization process. Instead, the encoder and the distribution parameters of the stochastic task head are alternately updated in each training epoch until the completion of training.

Algorithm 2 Optimizing by Margin Likelihood

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1\sim N}$, learning rate lr

Output: Shared representation encoder $\boldsymbol{\theta}$ and distribution parameters $\mathbf{m}_t$ and $\mathbf{S}_t$ of task specific head $t(t = 1, \dots, \mathrm{T})$

Optimizing:

1: Randomly initialize shared encoder and variational distribution parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_t = \{\mathbf{m}_t, \mathbf{S}_t\}_{t=1,\dots,\mathrm{T}}$, set task-specific prior $p(\mathbf{W}_t) = \mathcal{N}(\mathbf{0}, \mathbf{I})(t = 1, \dots, \mathrm{T})$

2: **for** $k = 1, \dots, K$ **do**

3: Sample mini-batch of data $\mathcal{D}_k \sim \mathcal{D}$ containing B samples

4: Optimize shared encoder parameters

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}_{\boldsymbol{\theta}}$$

where $\mathcal{L}_{\boldsymbol{\theta}} = \sum_{t=1}^{\mathrm{T}} \mathcal{L}_t$ and

$$\mathcal{L}_t = \begin{cases} -\dfrac{N_i}{N_a} \displaystyle\sum_{n=1}^{N} \left[\frac{1}{M} \sum_{m=1}^{M} \log p\left(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, \boldsymbol{\theta}, \mathbf{W}_t^{(m)}\right) \right], \textit{active data point} \\ -\displaystyle\sum_{n=1}^{N} \left[\frac{1}{M} \sum_{m=1}^{M} \log p\left(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, \boldsymbol{\theta}, \mathbf{W}_t^{(m)}\right) \right], \textit{inactive data point} \end{cases}$$

$\mathbf{W}_t^{(m)}$ is mth sample of $q(\mathbf{W}_t; \mathbf{m}_t, \mathbf{S}_t)$.

5: Set $\boldsymbol{\theta} = \boldsymbol{\theta}^*$

6: Optimize task-specific heads' parameters

$$\mathbf{m}_t^*, \mathbf{S}_t^* = \underset{\mathbf{m}_t, \mathbf{S}_t}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}, \mathbf{m}_t, \mathbf{S}_t)$$

where $\mathcal{L}(\boldsymbol{\theta}^*, \mathbf{m}_t, \mathbf{S}_t) = \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}[\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \mathbf{W}_t)] - \mathrm{KL}[q(\mathbf{W}_t; \mathbf{m}_t, \mathbf{S}_t) || p(\mathbf{W}_t)]$

7: Set $\mathbf{m}_t = \mathbf{m}_t^*$ and $\mathbf{S}_t = \mathbf{S}_t^*$

8: **end for**

Note: We still use stochastic gradient descent for the $\operatorname{argmax}(\cdot)$ operation

APPENDIX B

EXPERIMENTAL SETTING

B.1 Dataset Preparation

The experimental dataset used in this study adheres to the curation protocol established by Zhang et al. [19], comprising epigenetic target-compound activity data extracted from the ChEMBL database. The dataset specifically focuses on six major epigenetic protein classes: histone methyltransferases (HMTs), histone acetyltransferases, histone demethylases, histone deacetylases (HDACs), DNA methyltransferases, and epigenetic reader proteins.

The initial dataset was compiled from the ChEMBL database and comprised bioactivity records for 110 epigenetic targets involving 100,412 distinct compound-target interactions. Following the data processing protocol established by Zhang et al. [19], we implemented a rigorous multi-step curation procedure to ensure data quality and consistency. Specifically, we retained only those records containing standardized activity measurements, including inhibition values, half-maximal inhibitory concentrations (IC_{50}), half-maximal effective concentrations (EC_{50}), inhibition constants (K_i), or dissociation constants (Kd). For quantitative measurements, compounds demonstrating potency values $\leqslant 10$ μM were classified as active, while those with values >10 μM were designated as inactive. Qualitative inhibition data were carefully evaluated based on the concentration parameters specified in the corresponding assay descriptions to determine activity status. To maintain data integrity, we excluded ambiguous records exhibiting intermediate activity (e.g., IC50 >5 μM) or showing substantial but subthreshold inhibition (>70% at 30 μM). Additional quality control measures involved removing entries lacking valid SMILES representations, resolving conflicting activity annotations for the same compound through majority voting, and eliminating records with irreconcilable discrepancies. This comprehensive data standardization approach yielded a high-quality dataset with well-defined activity classifications, providing a reliable foundation for subsequent computational modeling efforts while maintaining methodological consistency with established benchmarks in the field.

B.2 Feature Representations of Molecules

We conducte a comprehensive comparative analysis of two widely used molecular representation methods, Extended-Connectivity Fingerprints (ECFP) and molecular graph representations, to evaluate their respective impacts on model prediction performance in the context of epigenetic target prediction. For the ECFP based approach, we implemented the ECFP4 variant (with a diameter of 4) using a 1024-bit fingerprint length, maintaining consistency with the parameters reported by Zhang et al. [19] . This circular fingerprint method captures local atomic environments and has become a standard tool in cheminformatics for molecular similarity searching and property prediction. Regarding molecular graph representations, we employed a directed graph formalism where atoms serve as nodes (characterized by 133-dimensional feature vectors) and bonds as edges (represented by 14-dimensional features), with all graph constructions automatically generated from SMILES strings using the Chemprop library.

B.3 Base Line Model Construction

In our experiments, we compared multiple baseline models to demonstrate the advantages of our proposed method. The experimental configurations for each baseline model are described below:

SVM [25]: We implemented SVM models for each target separately using the SVM class from the scikit-learn library. The hyperparameters were set as follows: C=1, kernel="rbf", gamma="scale", shrinking=True, tolerance=1e-3.

LR: We constructed logistic regression models for each target using the LogisticRegression class from scikit-learn, with the following parameter settings: maximum iterations max_iter=1000, solver="lbfgs", and L2 regularization (penalty="l2").

NN: We implemented a neural network architecture identical to MT-EpiPred [19], consisting of a shared feature encoder with MLP layers of dimensions 1024×1500 and 1500×1000, task-specific output layers of dimension 1000×2 and ReLU activation functions for all non-linear transformations.

For CNN [26], VMTL [27], PreGNN [24] and MT-EpiPred [19] models, we faithfully reproduced the original architecture as reported in their respective works.

B.4 Model Evaluation

TABLE S1. CALCULATION OF MODEL EVALUATION METRICS.

evaluation metric	equation
accuracy	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
precision	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
recall	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
F1 score	$2\times\frac{\mathrm{presicion}\times\mathrm{recall}}{\mathrm{presicion}+\mathrm{recall}}$
BA	$0.5\times\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}+\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}\right)$
MCC	$\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}\times\mathrm{FP})(\mathrm{TP}\times\mathrm{FN})(\mathrm{TN}\times\mathrm{FP})(\mathrm{TN}\times\mathrm{FN})}}$
ECE	$\sum_{m=1}^{M}\frac{\lvert B_m\rvert}{N}\lvert\mathrm{acc}(B_m)-\mathrm{conf}(B_m)\rvert$
OOD-AUROC	$\int_0^1 \mathrm{TPR}(f)\,df$

TP (True Positive): Correctly classified positive samples (predicted positive and actually positive); TN (True Negative): Correctly classified negative samples; FP (False Positive): Incorrectly classified positive samples; FN (False Negative): Incorrectly classified negative samples; M: Total number of confidence bins; B_m: Set of samples in the m-th confidence bin; $|B_m|$: Number of samples in the m-th confidence bin; N: Total number of samples; $\mathrm{acc}(B_m)$: Classification accuracy of samples within the m-th bin; $\mathrm{conf}(B_m)$: Mean predicted confidence score of samples within the m-th bin; $\mathrm{TPR}(f)$: representing the probability of correctly identifying in-distribution samples within complete data space f, is computed following Liu et al. [28]. The confidence scores are derived from the maximum softmax probabilities as described in reference [22], enabling effective discrimination between in-distribution and out-of-distribution samples.

The evaluation framework for our study incorporates comprehensive metrics to assess both predictive performance and uncertainty quantification capabilities across all four implemented methods. For comparative analysis with MT-EpiPred [19], we employed standard predictive accuracy metrics including accuracy, precision, recall, F1 score, balanced accuracy (BA), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (auROC), and area under the precision-recall curve (AUPR). To specifically evaluate uncertainty calibration and out-of-distribution detection performance, we additionally computed expected calibration error (ECE) and out-of-distribution area under the receiver operating characteristic curve (OOD-AUROC), with detailed calculation methods provided in Table S1.

B.5 Experimental Setup

For model training and evaluation, we implemented a standardized 8:1:1 split ratio to partition the data into training, validation, and test sets. The OOD detection capability was quantified using the OOD-AUROC metric, with all experimental results obtained through triplicate runs under different random seeds to ensure statistical robustness. Final reported metrics represent the averaged performance across 3 random seeds at the epoch demonstrating peak validation accuracy during the training process.

Margin Likelihood with ECFP Encoding. To establish a fair benchmarking baseline against MT-EpiPred [19], we carefully reproduced its neural architecture while incorporating our Bayesian enhancements. The model architecture consists of a multilayer perceptron feature encoder with sequential linear transformations (1024×1500 and 1500×1000 dimensions) employing ReLU activation functions, followed by a Bayesian output layer (1000×2 dimensions) operating under the independent and identically distributed (i.i.d.) assumption. This probabilistic output layer is parameterized by two learnable components: a 1000×2 mean vector and a corresponding 1000×2 diagonal covariance matrix. Following established practices from Zhang et al. [19], we implemented task-specific loss scaling to address multitask learning challenges.

The training strategy employed an alternating optimization method: initially fixing the distribution parameters while training the encoder through marginal likelihood maximization across the complete dataset, then fixing the encoder parameters while optimizing the distribution parameters through maximizing ELBO. This iterative two-phase optimization procedure was executed for 1000 complete epochs, with each epoch comprising both optimization stages to ensure proper convergence of all model parameters.

BMT-EpiPred with ECFP Encoding. This model employs the same architecture as the marginal likelihood model, directly utilizing the variational Bayesian layer library contributed by Harrison et al. [22]. The variational Bayesian classification serves as the task head for our BMT-EpiPred model, which similarly incorporates 1000×2 mean vectors and 1000×2 diagonal covariance matrices as learnable parameters. We also apply scaling processing to the loss values generated by each task. In BMT-EpiPred approach, the loss values are directly utilized for backpropagation to simultaneously update all learnable parameters, with the training process likewise conducted for 1000 epochs.

Across all BMT-EpiPred implementations, we adopt the method proposed by Kwon et al. [30], employing (S1) to comprehensively account for both data uncertainty and model uncertainty. In this process, $\mathbf{W}_m$ is obtained by sampling from $q(\mathbf{W}; \mathbf{m}, \mathbf{S})$, with a total of $M = 25$ samples drawn.

$$Var_{q(\mathbf{y}|\mathbf{x})}(\mathbf{y}) \approx \frac{1}{M}\sum_{m=1}^{M}\left[diag\left(p(\mathbf{y}|\mathbf{x}, \mathbf{W}_m)\right) - p(\mathbf{y}|\mathbf{x}, \mathbf{W}_m)^2\right] + \frac{1}{M}\sum_{m=1}^{M}\left[p(\mathbf{y}|\mathbf{x}, \mathbf{W}_m) - \frac{1}{M}\sum_{m=1}^{M} p(\mathbf{y}|\mathbf{x}, \mathbf{W}_m)\right]^2 \quad \text{(S1)}$$

BMT-EpiPred with Graph. To investigate the impact of different initial feature encodings on model predictive performance, we constructed a graph-based BMT-EpiPred model. We employed MPNN [23] as the initial feature encoder for the model, with all MPNN configurations consistent with those used by Heid et al. [23]. Specifically, the input node and edge feature dimensions of the MPNN were set to 133 and 14, respectively, with ReLU serving as the activation function. The architecture comprised three message-passing layers, each utilizing a MLP as the message function. The molecular features were computed via global average pooling, with the dimension set to 300. The remaining experimental settings were identical to those of the BMT-EpiPred model using ECFP encoding.

APPENDIX C

PROOF OF THEOREM

C.1 Proof of Equation (2)

Let $\mathbf{W}_t$ denotes parameters of head t, which follow a Gaussian distribution $q(\mathbf{W}_t; \mathbf{m}_t, \mathbf{S}_t) = \mathcal{N}(\mathbf{W}_t; \mathbf{m}_t, \mathbf{S}_t)$, with $\mathbf{m}_t$ representing the mean vector and $\mathbf{S}_t$ the covariance matrix. Then, Equation (5) holds with

$$\mathcal{L}_t(\boldsymbol{\theta}, \boldsymbol{\phi}_t, \boldsymbol{\Sigma}) = \frac{1}{\gamma}\sum_{i=1}^{N}\left(\mathbf{y}_i^\top \mathbf{m}_t \mathbf{h}_i - \mathrm{LSE}_j\left(\mathbf{m}_t^{(j)\top}\mathbf{h}_i + \frac{1}{2}\left(\mathbf{h}_i^\top \mathbf{S}_t^{(j)} \mathbf{h}_i + \sigma_t^{(j)^2}\right)\right)\right) - \frac{1}{\gamma}\sum_{i=1}^{N}\mathrm{KL}\big(q(\mathbf{W}_t|\mathbf{m}_t, \mathbf{S}_t)||p(\mathbf{W}_t)\big)$$

where $\boldsymbol{\theta}$ denotes the parameters of the feature encoder, $\boldsymbol{\phi}_t = \{\mathbf{m}_t, \mathbf{S}_t\}$ represents the trainable parameters of the variational Bayesian output layer, $\boldsymbol{\Sigma}_t$ corresponds to task t noise covariance matrix of observed activity values, $\sigma_t^{(j)}$ is the jth dimension of $\boldsymbol{\Sigma}_t$, γ serves as a scaling factor, N indicates the data size, $\mathbf{h}_i$ signifies the encoded features of sample i from the feature encoder, and $p(\mathbf{W}_t)$ denotes the prior distribution of model parameters.

Proof. Firstly,

$$\begin{aligned}
\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) &= \log \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \mathbf{W}) p(\mathbf{W}) d\mathbf{W} \\
&= \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}\left[\log \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \mathbf{W}) p(\mathbf{W})}{q(\mathbf{W})}\right] \\
&\geq \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}[\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \mathbf{W})] - \mathrm{KL}[q(\mathbf{W}|\boldsymbol{\phi})||p(\mathbf{W})] \\
&= \sum_{b=1}^{N_\mathbf{y}} \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}\left[\mathbf{y}_b^\top \log \mathrm{softmax}_\mathbf{y}(\log p(\mathbf{x}_b, \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi}_t))\right] - \mathrm{KL}[q(\mathbf{W}|\boldsymbol{\phi})||p(\mathbf{W})]
\end{aligned}$$

And expand log-softmax term, we have

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}\left[\mathbf{y}_b^\top \log \mathrm{softmax}_\mathbf{y}(\log p(\mathbf{x}_b, \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi}_t))\right] &= \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}[\mathbf{y}_b^\top \log p(\mathbf{x}_b, \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi}_t)] - \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}\left[\mathrm{LSE}_j(\log p(\mathbf{x}_b, \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi}_t))\right] \\
&= \mathbf{y}_b^\top \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}[\mathbf{W}_t]\mathbf{h}_{\boldsymbol{b}} - \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}\left[\mathrm{LSE}_j(\log p(\mathbf{x}_b, \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi}_t))\right] \\
&= \mathbf{y}_b^\top \mathbf{m}_t \mathbf{h}_{\boldsymbol{b}} - \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}\left[\mathrm{LSE}_j(\log p(\mathbf{x}_b, \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi}_t))\right]
\end{aligned}$$

Via Jensen's inequality, we have

$$\begin{aligned}
\mathbf{y}_b^\top \mathbf{m}_t \mathbf{h}_{\boldsymbol{b}} - \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}\left[\mathrm{LSE}_j(\log p(\mathbf{x}_b, \mathbf{y}_b|\boldsymbol{\theta}, \boldsymbol{\phi}_t))\right] &\geq \mathbf{y}_b^\top \mathbf{m}_t \mathbf{h}_{\boldsymbol{b}} - \log \sum_{j=1}^{N_y} \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\phi})}\left[\exp\left(\log p\left(\mathbf{x}_b, \mathbf{y}_b^{(j)}|\boldsymbol{\theta}, \boldsymbol{\phi}_t\right)\right)\right] \\
&= \mathbf{y}_b^\top \mathbf{m}_t \mathbf{h}_{\boldsymbol{b}} - \log \textstyle\sum_{j=1}^{N_y} \exp\left(\mathbf{m}_t^{(j)\top}\mathbf{h}_b + \frac{1}{2}\left(\mathbf{h}_b^\top \mathbf{S}_t^{(j)} \mathbf{h}_b + \sigma_t^{(j)^2}\right)\right)
\end{aligned}$$

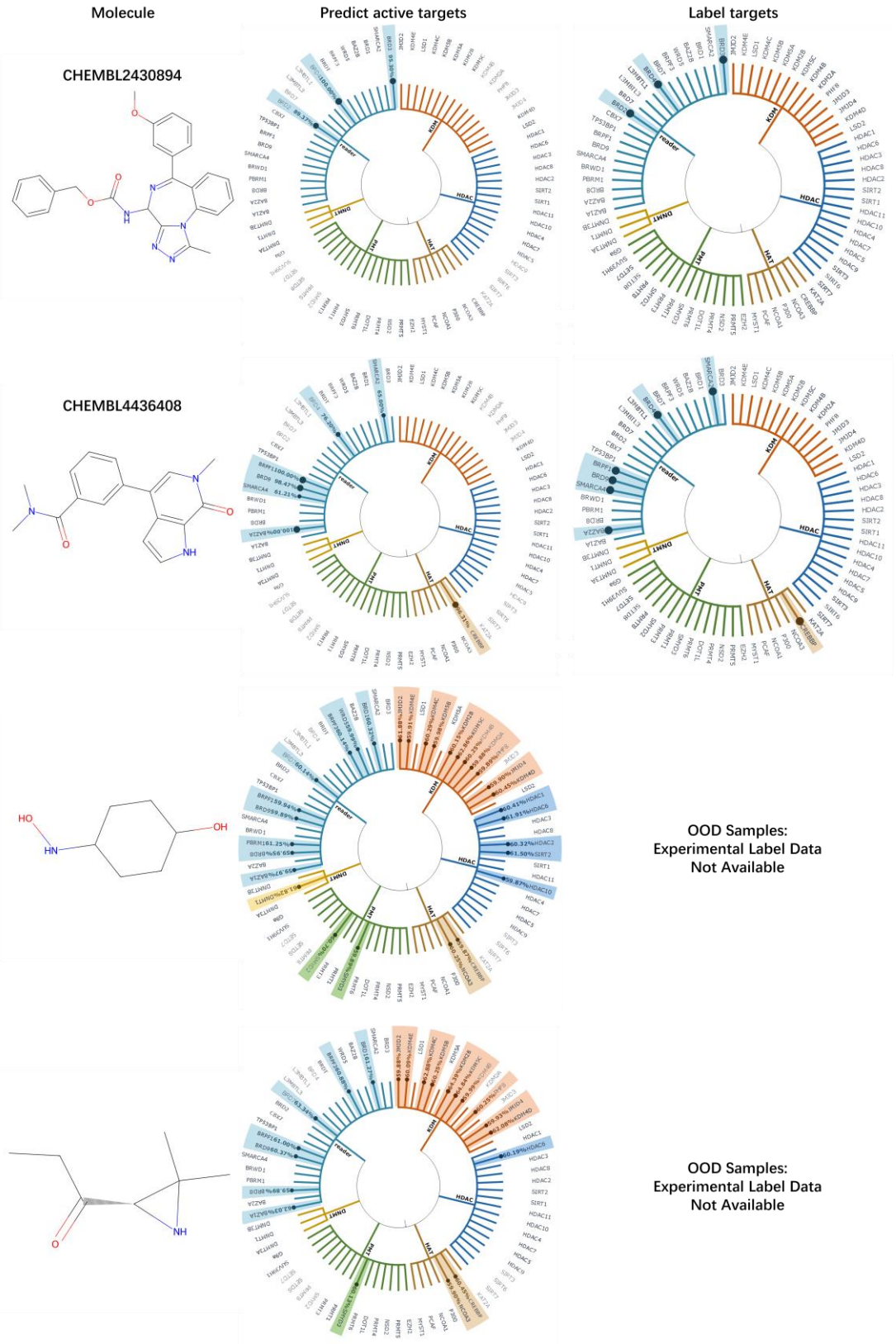
So we can prove

$$\begin{aligned}
\mathcal{L}_t(\boldsymbol{\theta}, \boldsymbol{\phi}_t, \boldsymbol{\Sigma}) &= \frac{1}{\gamma}\sum_{b=1}^{N_\mathbf{y}}\left(\mathbf{y}_b^\top \mathbf{m}_t \mathbf{h}_{\boldsymbol{b}} - \log \sum_{j=1}^{N_y} \exp\left(\mathbf{m}_t^{(j)\top}\mathbf{h}_b + \frac{1}{2}\left(\mathbf{h}_b^\top \mathbf{S}_t^{(j)} \mathbf{h}_b + \sigma_t^{(j)^2}\right)\right)\right) - \frac{1}{\gamma}\sum_{k=1}^{N_\mathbf{y}}\mathrm{KL}\big(q(\mathbf{W}_t|\mathbf{m}_t, \mathbf{S}_t)||p(\mathbf{W}_t)\big) \\
&= \boldsymbol{y}_b^\top \boldsymbol{m}_t \boldsymbol{h}_{\boldsymbol{b}} - log \sum_{j=1}^{N_y} exp\left(\boldsymbol{m}_t^{(j)\top}\boldsymbol{h}_b + \frac{1}{2}\left(\boldsymbol{h}_b^\top \boldsymbol{S}_t^{(j)} \boldsymbol{h}_b + \sigma_t^{(j)^2}\right)\right) - \frac{1}{\gamma}\sum_{k=1}^{N_\mathbf{y}} KL\big(q(\boldsymbol{W}_t|\boldsymbol{m}_t, \boldsymbol{S}_t)||p(\boldsymbol{W}_t)\big)
\end{aligned}$$

APPENDX D

APPLICATION PERFORMANCE TEST

D.1 Application Performance Test



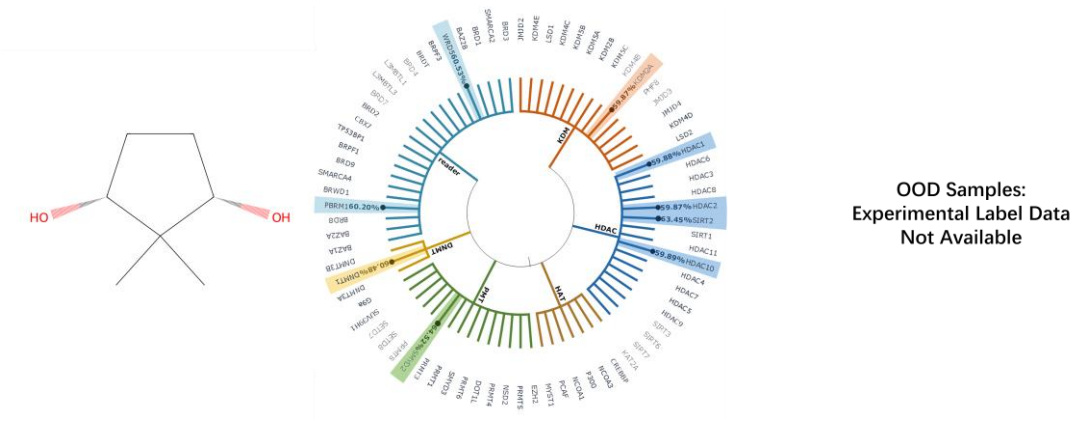


Fig S1. Comparison of BMT-EpiPred model predictions with experimental activity data for five compounds. The left panels display molecular structures. The center panels show predicted active targets (defined as predicted distribution's mean > 0.5) with shaded backgrounds indicating activity. The point sizes and percentage values correspond to the predicted confidence levels, while the right panel displays the ground truth labels. The first two rows display in-distribution test results (data unseen during model training), while the last three rows shows OOD testing results.

Figure S1 demonstrates the predictive performance of our model for five drug molecules across 78 target activities. As illustrated in the Figure S1, the BMT-EpiPred model demonstrates high accuracy in predicting active compounds for IND data, accompanied by relatively high confidence scores (averaging above 0.8) for correct predictions. In contrast, for OOD data, which may not have been encountered during training, the model fails to provide confident judgments, consequently assigning lower prediction confidence scores (averaging below 0.65) to such instances.