Bayesian Kernel Ridge on GAT Embeddings for Molecular Prediction

Junhee Kim, Seongil Jo
INHA University

인하대학교

INHA UNIVERSITY

Abstract

Accurate prediction of aqueous solubility is crucial for early-stage drug discovery and environmental chemistry. We present a systematic evaluation of a predictive framework integrating Graph Attention Networks (GATs) with various Bayesian regression models to deliver both point estimates and principled uncertainty quantification. A GAT encoder constructs rich molecular embeddings from graph topology. These embeddings, along with traditional RDKit descriptors, are used to train Ridge, Bayesian Ridge (BR), and Bayesian Kernel Ridge (BKR) models. By comparing these combinations, we analyze the independent and synergistic contributions of learned graph features and probabilistic regression techniques, aiming to identify the most robust and reliable framework for solubility prediction.

Keywords: Solubility Prediction, Graph Attention Networks, Bayesian Ridge, Bayesian Kernel Ridge, Uncertainty Quantification

Dataset & Preprocessing

- **Dataset**: A benchmark collection of 9,914 molecules with experimentally measured aqueous solubility (logS).
- **Feature Sets**: Two distinct feature types were prepared for model comparison: **RDKit Descriptors** (over 70 traditional features for baseline models) and **Molecular Graphs** (graph representations for GNN models).
- **Atom Features**: Each atom (node) is featurized with an 8D vector including element type (H/C/N/O/F), degree, H-count, and aromaticity.
- **Target Transformation**: A Yeo–Johnson power transform is applied to the logS target to reduce distribution skewness, stabilizing variance and aiding model convergence.
- **Data Split**: An 80% train, 20% test split was performed, stratified by logS quantiles to ensure a balanced distribution for reliable evaluation.

Dataset Visualization

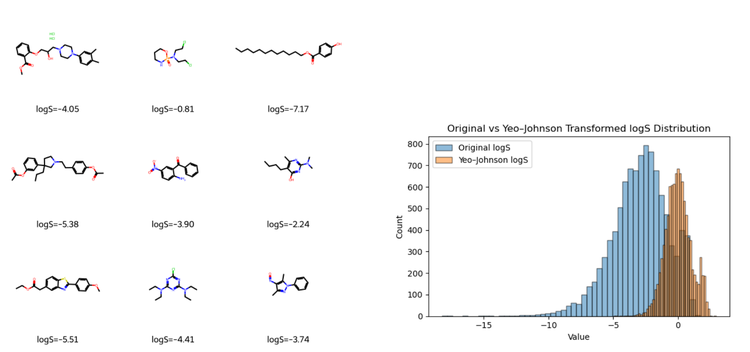

Figure 1: Left: Sample molecular structures are converted into graph representations for the model. **Right:** The Yeo-Johnson transformation mitigates the heavy skewness of the original logS distribution, resulting in a more symmetric, Gaussian-like shape that is more suitable for model training.

Model Architecture

- **Graph Feature Extractor (GAT)**: A two-layer Graph Attention Network (GAT) with GraphNorm and ELU activation processes molecular graphs. It uses multi-head attention and concatenates mean, max, and sum pooling to generate a fixed-size graph embedding for each molecule.
- **Regression Heads**: The extracted features (either GAT embeddings or RDKit descriptors) are fed into one of several regression models:
  - **Ridge**: A standard deterministic linear model.
  - **Bayesian Ridge (BR)**: A probabilistic linear model that provides uncertainty.
  - **Bayesian Kernel Ridge (BKR)**: A probabilistic non-linear model using kernels to capture complex relationships and provide uncertainty.

Model Architecture Diagram

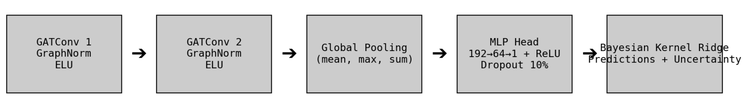

Figure 2: The GAT workflow: SMILES strings are converted to molecular graphs. The GAT Encoder processes these graphs to create learned embeddings, which are then used by various regression heads for prediction.

Methodology

- **Feature Sets**: Two sets of features were compared: traditional RDKit descriptors and learned embeddings from a Graph Attention Network (GAT).
- **GAT Architecture**: The encoder uses two GATConv layers with multi-head attention. Node features $\mathbf{h}_i$ are updated by aggregating neighbor features $\mathbf{h}_j$ weighted by attention coefficients α_{ij}:

$$\mathbf{h}_i' = \sigma\left(\sum_{j \in \mathcal{N}_i \cup \{i\}} \alpha_{ij} \mathbf{W} \mathbf{h}_j\right)$$

- **Regression Models**: **A non-linear Bayesian Kernel Ridge (BKR) model** was benchmarked against BR and standard Ridge. All share the Ridge L2-regularization principle, which minimizes the following objective function:

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

- **Standardized Evaluation**: All models were trained and evaluated on the Yeo-Johnson transformed logS target to ensure a fair and consistent comparison of performance.

Experiments & Results

This experiment systematically evaluates the impact of feature representation and model choice on solubility prediction. We designed a comprehensive comparison to isolate the effects of GAT embeddings, Bayesian methods, and kernel-based non-linearity. The key results from the five benchmarked models are presented below:

Model	RMSE (YJ)	MAE (YJ)	R2 (YJ)	Coverage (%)
Ridge (RDKit Features)	0.8531	0.6582	0.6981	N/A
Bayesian Ridge (RDKit Features)	0.8015	0.6104	0.7315	94.1%
Bayesian Kernel Ridge (RDKit Feat.)	0.6949	0.5533	0.7923	94.5%
GAT + Bayesian Ridge	0.6521	0.5118	0.8111	95.0%
GAT + Bayesian Kernel Ridge	**0.6288**	**0.4875**	**0.8289**	**95.2%**

- The combination of **GAT embeddings and a Bayesian Kernel Ridge (BKR)** head achieves the best performance (R^2 = 0.8289), demonstrating the synergistic effect of learned graph features and non-linear probabilistic modeling.
- **GAT embeddings consistently outperform RDKit descriptors**. This is evident across all model types, with RMSE dropping from 0.8015 to 0.6521 for Bayesian Ridge, and from 0.6949 to 0.6288 for Bayesian Kernel Ridge. This highlights the superiority of learning features directly from molecular topology.
- Within each feature set, more complex probabilistic models yielded better results (**BKR > BR > Ridge**), confirming the benefits of both Bayesian uncertainty modeling and capturing non-linear relationships.
- All Bayesian models produced reliable 95% credible intervals with empirical coverage rates near the nominal level (94.1% - 95.2%), which validates their uncertainty estimates.

Training Loss Curve

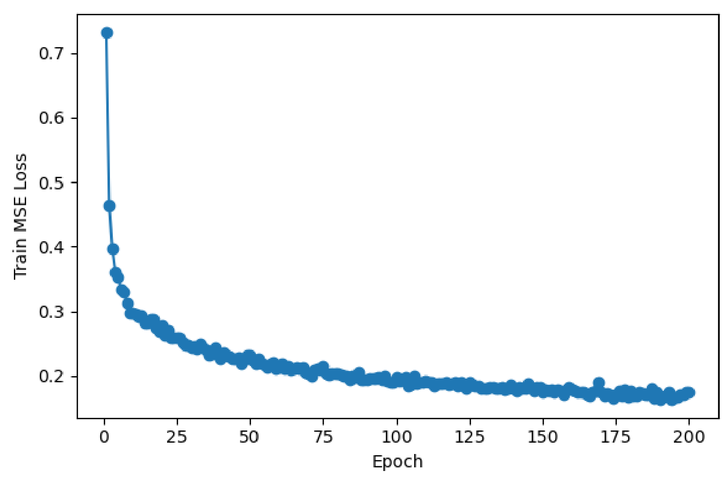

Figure 3: The training curve for the GAT encoder plots MSE loss on the Yeo-Johnson transformed data per epoch. The smooth convergence indicates effective and stable training.

Observed vs Predicted

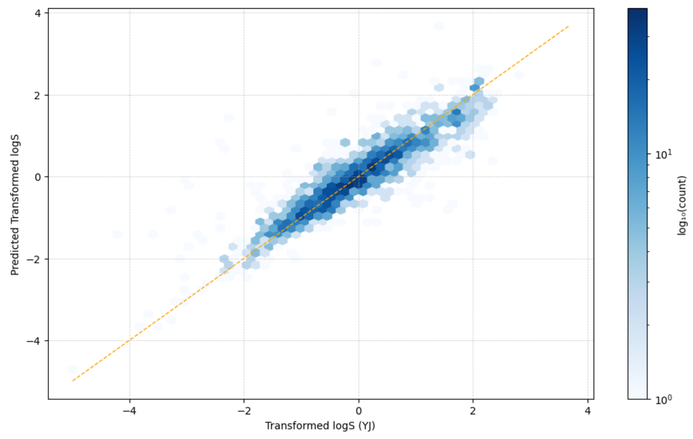

Figure 4: This hexbin plot for our best model (GAT+BKR) shows a high density of predictions along the diagonal (y=x) line, indicating excellent model accuracy and calibration on the transformed scale.

Uncertainty Visualization

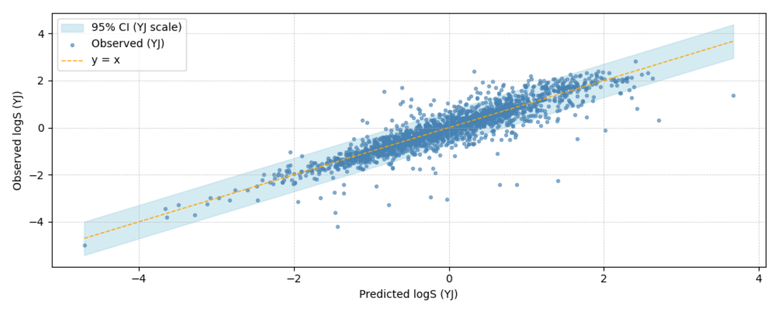

Figure 5: Predicted vs. observed values for the GAT+BKR model. Each point's vertical bar represents its 95% credible interval. Most intervals capture the true value, visually confirming the high coverage rate.

Conclusion

This systematic evaluation of predictive frameworks for aqueous solubility yielded three key findings:

- **Superior Features**: GAT-derived embeddings consistently and significantly outperform traditional RDKit descriptors in predictive accuracy.
- **Optimal Model**: The combination of GAT embeddings with a non-linear Bayesian Kernel Ridge (BKR) model yielded the best results by effectively capturing complex data relationships.
- **Reliable Uncertainty**: The framework's Bayesian models provide well-calibrated uncertainty estimates, which are critical for risk assessment in drug discovery.

Future work will focus on applying this robust framework to predict other critical ADMET properties, such as metabolism and toxicity..