

Improving Data Efficiency for Severe Accident Surrogate Training with Active Learning

Seok Ho Song^a, Seunghyeok Yang^b, Taewoo Kim^c, Mi Ro Seo^c, Jeong Ik Lee^{a*}

^aDepartment of Nuclear and Quantum Engineering N7-1 KAIST 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea 305-338, 1812wow@kaist.ac.kr

^bFNC Technology CO., LTD, Heungdeok IT Valley Bldg. 32F, 13, Heungdeok 1-ro, Giheung-gu, Yongin-si, Gyeonggi-do, 16954, Republic of Korea

^cKorea Hydro and Nuclear Power Central Research Institute, 70, Yuseong-daero 1312beon-gil, Yuseong-gu, Daejeon, Republic of Korea

*Corresponding author: jeongiklee@kaist.ac.kr

***Keywords : Severe accident, Active learning, BADGE, Data efficiency, Surrogate model**

1. Introduction

The progression of severe accidents in nuclear power plants (NPPs) involves highly complex, nonlinear, and tightly coupled thermal-hydraulic phenomena. Accurately predicting the evolution of these dynamics—especially parameters such as primary system pressure, containment pressure, and reactor coolant system temperatures—is of paramount importance. These predictions form the technical basis for operator decision-making and the timely execution of accident mitigation procedures outlined in the Severe Accident Management Guidelines (SAMG). Traditionally, these analyses are performed using integral system codes like the Modular Accident Analysis Program (MAAP). While MAAP is highly reliable and provides detailed physical insights, it requires substantial computational resources. The significant computation time strictly limits its applicability in scenarios requiring real-time optimal decision-making during an active emergency, or in large-scale Probabilistic Safety Assessments (PSA) where millions of iterative calculations are necessary.

To address this computational bottleneck, data-driven surrogate models employing advanced deep learning architectures have emerged as highly effective alternatives. Once trained, these surrogate models offer nearly instantaneous inference capabilities, allowing for real-time monitoring and predictive diagnostics without sacrificing essential accuracy. However, developing robust artificial neural networks (ANNs) for safety-critical applications introduces a new challenge: the 'Data Efficiency Bottleneck.' Training highly accurate AI models requires massive datasets encompassing a wide variety of accident scenarios. Paradoxically, generating this Big Data requires running the same computationally expensive MAAP system codes that the surrogate model aims to bypass. Relying on traditional Monte Carlo or random sampling of the parameter space exacerbates this problem, as it often wastes the limited data generation budget on redundant, 'easy-to-predict' regions of the operational space while missing critical edge cases.

To overcome the data efficiency bottleneck, this study introduces a Pool-Based Active Learning (AL)(figure1) framework tailored specifically for nuclear severe accident data. Unlike traditional passive learning where

data is randomly collected beforehand, AL is a dynamic process where the AI iteratively analyzes a vast pool of unlabeled scenarios and explicitly queries only the most informative data points for simulation. This fundamentally maximizes the model's learning rate and predictive performance under a strictly fixed data generation budget. Specifically, this study adapts the state-of-the-art Batch Active learning by Diverse Gradient Embeddings (BADGE) algorithm. Originally designed for classification tasks, BADGE uniquely synthesizes both model uncertainty and spatial data diversity to select optimal batches. By adapting BADGE for continuous physical regression and integrating a strict Generalization Loss (GL) based early stopping mechanism, the proposed methodology seeks to drastically reduce the required volume of expensive physical simulations while ensuring the highest level of surrogate model fidelity.

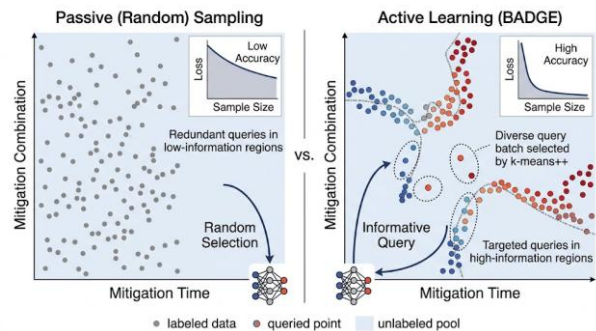


Figure 1 Conceptual Illustration of Active Learning

2. Methods

2.1 Accident Scenario and Data Generation

The foundational dataset for this research was generated by simulating Total Loss of Feedwater (TLOFW) accidents in a reference APR1400 nuclear power plant utilizing the MAAP system code. The TLOFW scenario was selected due to its extreme conservatism; it assumes a worst-case condition where all engineering safety features fail simultaneously. This results in a rapid accumulation of decay heat and complex transient behaviors that vigorously test the capabilities of the predictive surrogate models.

To ensure comprehensive coverage of the operational parameter space and to prevent data bias, the Maximin Sampling technique was applied, generating an extensive pool of 5,000 unique scenarios. In order to realistically model the impact of operator interventions, five distinct SAMG mitigation strategies were incorporated: (1) Reactor Cooling System (RCS) depressurization, (2) Steam Generator (SG) external injection, (3) RCS external injection, and (4) Containment Spray Pump (CSP) activation. These mitigation actions were programmed to actuate at randomized intervals ranging between 1 and 24 hours after the initiating event. All input and target features were normalized using a Min-Max scaling technique, mapped to a bounded range of [0.1, 0.9] to prevent numerical saturation during deep neural network training.

The surrogate model is designed to map a 4-dimensional input space of operator actions to a 10-dimensional output space of critical thermal-hydraulic responses (Figure 2). The input parameter space consists of the continuous actuation times (ranging from 1 to 24 hours) of four specific SAMG mitigation strategies: RCS depressurization, SG external injection, RCS external injection, and CSP activation. For each unique 4-dimensional input vector sampled via the Maximin technique, the MAAP code calculates a comprehensive transient response spanning 500 discrete time steps. The model's prediction target encompasses 10 distinct thermal-hydraulic variables, carefully selected to represent the integral safety state of the plant, including primary and secondary side pressures, coolant temperatures, and mass inventories. This 4 to 10-dimensional mapping over a temporal sequence forms the fundamental architecture of the surrogate model's predictive task.

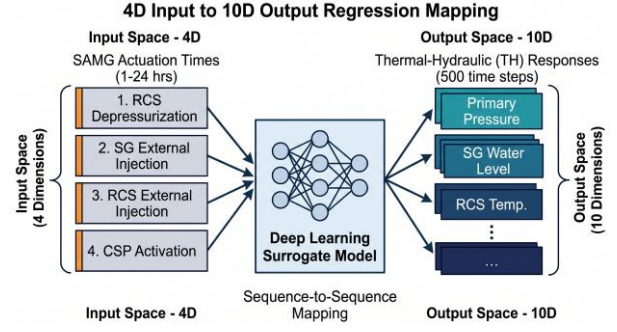


Figure 2 4D to 10D Input-Output Mapping Architecture

2.2 Active Learning via BADGE for Regression

The core innovation of the data acquisition strategy lies in overcoming the historical limitations of active learning. Existing AL paradigms often suffer from the mutually exclusive nature of exploitation and exploration. 'Exploitation' typically involves uncertainty sampling, which can lead to the 'Mode Collapse Trap,' where the model repeatedly queries highly similar points. Conversely, 'Exploration' emphasizes diversity sampling but ignores the model's internal confidence, wasting valuable labeling budget on regions the model already understands perfectly.

The BADGE algorithm resolves this dilemma by transforming the query selection problem from the raw input feature space into a latent gradient embedding space derived from the neural network's final layer parameters. Figure 3 illustrates the overall workflow of the BADGE algorithm applied in this study.

Since, BADGE was originally designed for image classification tasks, it fundamentally relies on Soft-max probabilities and cross-entropy loss. To bridge this classification-to-regression gap for continuous physical variables, this study integrates Monte Carlo (MC) Dropout during the inference phase. By applying dropout

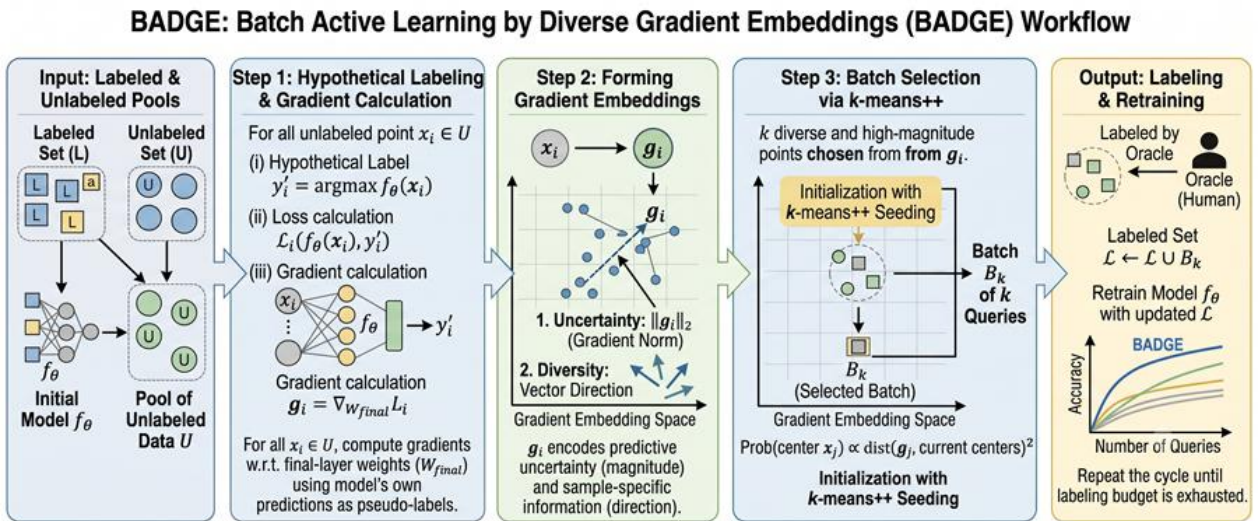


Figure 3 Workflow of the BADGE Active Learning Algorithm

at test time, the network acts as a Bayesian Neural Network, outputting a distribution of predictions. The variance across these MC iterations serves as a robust proxy for epistemic uncertainty. The gradients of the network parameters with respect to this variance are extracted to form the gradient embeddings. In this transformed space, the magnitude of the gradient vector inherently represents the model's uncertainty, while the direction of the vector represents the specific features of the data point. A k-means++ initialization algorithm is then applied to these gradient embeddings, mathematically guaranteeing the selection of a batch that is both highly uncertain (large magnitude) and structurally diverse (distant in vector direction).

2.3 Validation-Based Early Stopping (GL)

To ensure the surrogate model captures fundamental physical laws rather than merely memorizing the training dataset (overfitting), a strict Validation-Based Early Stopping mechanism was enforced based on Generalization Loss (GL). In this traditional paradigm, a statistically representative portion of the generated data is held out strictly as a validation set. During training, the model's performance on this unseen validation set is continuously monitored. The algorithm utilizes a specific patience hyperparameter and a minimum delta threshold. If the validation loss fails to decrease by at least the minimum delta over the patience period, training is immediately terminated. This guarantees that the neural network retains its generalizability across the vast accident state space.

$$GL(t) = 100 \cdot \left(\frac{E_{va}(t)}{E_{opt}(t)} - 1 \right) \quad eq. (1)$$

3. Results & Discussions

3.1 Active Learning Performance: AULC Analysis

The overarching goal of the Active Learning framework is to achieve maximum predictive accuracy using the minimum amount of MAAP simulation data. To comprehensively evaluate the performance of the regression-adapted BADGE algorithm, it was directly compared against a standard random sampling baseline. The models were iteratively trained and evaluated on progressively increasing fractions of the total dataset, specifically analyzing data ratios ranging from 2% up to 20%.

The primary metric utilized for this evaluation is the Area Under the Learning Curve (AULC) (Figure 4). By calculating the definite integral of the model's test error (Loss) as a function of the available data volume, the AULC quantifies the holistic data efficiency across the entire training lifecycle. A lower AULC indicates that the model rapidly descended the error curve, achieving

higher accuracy significantly faster and with fewer data points.

$$AULC = \int \mathcal{L}(r) dr \approx \sum_{i=1}^{N-1} \frac{\mathcal{L}(r_i) + \mathcal{L}(r_{i+1})}{2} (r_{i+1} - r_i) \quad eq. (2)$$

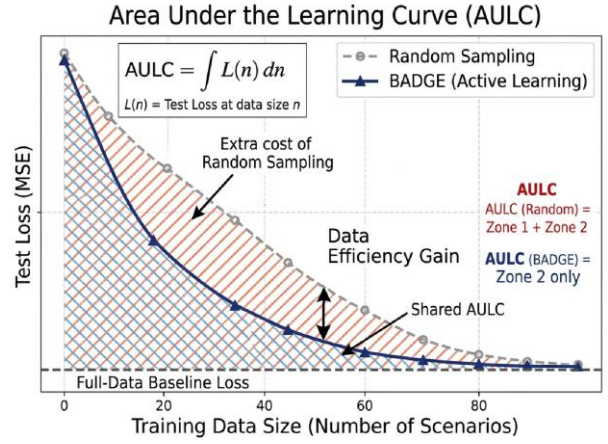


Figure 4 Conceptual Illustration of AULC

The empirical results demonstrate a clear superiority of the active learning approach. The Random sampling baseline achieved an overall AULC of 0.000253, whereas the BADGE framework achieved a distinctly lower AULC of 0.000240, translating to an absolute improvement of 5.07%. To contextualize this gain further, the results were compared against the theoretical optimal AULC (calculated as 0.000108), which represents the performance limit if the model had access to the entire dataset instantaneously. Measured against this theoretical limit, BADGE achieved a normalized AULC improvement (eq. 3) of 8.83%.

Norm. AULC Improvement (%)

$$= 100 \times \left(1 - \frac{AULC_{Baseline} - AULC_{BADGE}}{AULC_{Baseline} - AULC_{Random}} \right) \quad eq. (3)$$

To appreciate the practical significance of this result, consider the following. Each MAAP simulation of a single 72-hour TLOFW scenario requires approximately 2.56 hours of computation time [1]. With 5,000 unique scenarios generated in this study, the complete dataset represents a cumulative cost of approximately 12,800 computation-hours. The 5.07% absolute reduction in AULC—from 0.000253 (Random) to 0.000240 (BADGE)—effectively eliminates the need for approximately 250 redundant simulation runs, yielding a direct saving on the order of 640 computation-hours within the studied data range alone.

More critically, these gains compound as the target dataset scales. When extended to the full 30,000-scenario corpus encompassing TLOFW, LLOCA, and SBO scenarios—as planned for future research—the same efficiency ratio implies a potential reduction of over 1,500 redundant simulation runs. At this scale, the

difference between random and active sampling transitions from a modest percentage improvement to a qualitative shift: one that can determine whether a comprehensive training dataset is computationally feasible within a research program's resource constraints. As clearly illustrated in the learning curve progression (Figure 5), the BADGE algorithm rapidly converges toward the target threshold defined by the full-data model, whereas the random selection baseline suffers from inefficient, staggered learning.

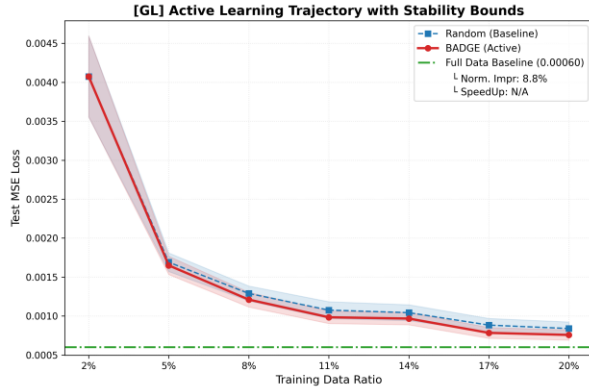


Figure 5 Learning Curve Progression:
BADGE vs. Random Sampling

Figure 6 further visualizes this dynamic, showing the normalized AULC area differences. The BADGE algorithm systematically avoided querying redundant data, instead focusing its computational budget on the complex, highly non-linear regions of the TLOFW transient, resulting in a noticeably steeper initial descent in the error curve.

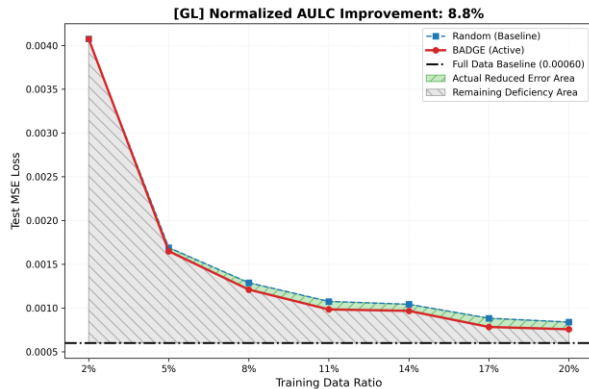


Figure 6 Normalized AULC Area Comparison

3.2 Statistical Significance Analysis

In deep learning, performance metrics can occasionally be skewed by favorable random initialization of network weights or fortuitous early data splits. To mathematically validate that the performance gains of the BADGE framework represent a consistent, structural advantage rather than statistical noise, the Mann-Whitney U Test was conducted. This robust non-parametric test compared the terminal error distributions between the

BADGE and Random sampling methodologies across 50 completely independent random seeds.

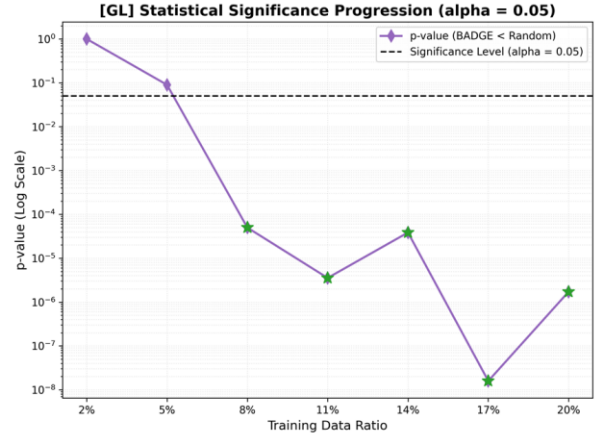


Figure 7 P-Value Progression over Data Ratios
(Alpha = 0.05)

The statistical progression reveals a critical transition point in data efficiency. In the extremely data-scarce regimes (specifically the 2% and 5% ratios), the p-values were calculated at 1.0000 and 0.0899 respectively. At these very early stages, the models are severely undetermined, meaning that neither strategy has accumulated enough foundational data to reliably map the physical environment; thus, the performance difference is not yet statistically significant. However, as the available data ratio reaches 8%, a distinct tipping point occurs. The BADGE algorithm demonstrates a decisive and statistically significant advantage, yielding a p-value of 4.98×10^{-5} (which is far below the standard alpha threshold of 0.05). At this 8%-mark, BADGE has successfully established a foundational understanding of the physics and leverages its diverse gradient embeddings to selectively target and map the complex boundary conditions of the SAMG mitigation effects.

This statistical significance is not only maintained but continuously strengthened throughout the subsequent data ratios, reaching an overwhelming minimum p-value of 1.58×10^{-8} at the 17% ratio. The visual progression of these p-values (Figure 7) effectively illustrates how BADGE reliably dominates as the dataset grows.

4. Conclusions & Further Works

This study successfully designed, implemented, and rigorously evaluated an extreme data-efficient active learning framework tailored for nuclear severe accident surrogate modeling. By effectively modifying the classification-based BADGE algorithm to handle continuous physical regression via Monte Carlo Dropout, and integrating a robust Generalization Loss (GL) based early stopping criterion, the proposed framework elegantly resolved the traditional exploitation-exploration dilemma in data sampling. The extensive empirical results confirmed that BADGE fundamentally outperforms traditional random sampling

across all metrics. BADGE achieved an 8.83% normalized improvement in Area Under the Learning Curve (AULC), rapidly driving down prediction errors with minimal data. Furthermore, extensive Mann-Whitney U testing across 50 independent seeds mathematically validated that the BADGE method achieves robust, undeniable statistical significance ($p < 0.05$) starting from just an 8% data availability ratio, maintaining superior reliability thereafter.

While the GL-based early stopping approach proved highly effective in preventing overfitting, its reliance on a held-out validation set exposes a fundamental contradiction in active learning known as the Validation Paradox. Discarding highly expensive, computationally intensive physical simulation runs purely to monitor validation loss contradicts the ultimate goal of maximizing data efficiency. Future research will focus heavily on dismantling this paradox by establishing a completely validation-free early stopping framework. This will involve formulating an Epistemic Uncertainty Proxy (EUP) based on the saturation of network disagreement, allowing 100% of the generated data to be used for active training. Additionally, the expanded framework will be stress-tested against other severe accident classes featuring fundamentally different statistical distributions, such as Large Break Loss of Coolant Accidents (LLOCA) and Station Blackout (SBO) scenarios.

ACKNOWLEDGMENT

This work was supported by the K-Cloud project of KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2024-Tech-08)

REFERENCES

- [1] Lee, Yeonha, et al. "Surrogate model for predicting severe accident progression in nuclear power plant using deep learning methods and Rolling-Window forecast." *Annals of Nuclear Energy* 208 (2024): 110816.
- [2] Joo, Semin. et al "Improving accuracy of a machine learning approach for forecasting severe accidents." 14th International Topical Meeting on Nuclear Reactor Thermal-Hydraulics, Operation, and Safety, 28 Aug.
- [3] Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., & Agarwal, A. (2019). Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671.