Performance of a Vision Transformer Surrogate in SA-Based APR-1400 Loading Pattern Search

Hyungjoon Baek [a], Dongmin Yun [a], Deokjung Lee [a*]
[a] *Department of Nuclear Engineering, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Ulsan, 44919, Republic of Korea*
[*] *Corresponding author: deokjung@unist.ac.kr*

***Keywords: Loading pattern search, Simulated annealing, Vision Transformer, Surrogate model, Screening**

1. Introduction

Simulated annealing (SA) has been used for nuclear fuel loading pattern (LP) optimization, where a large number of possible fuel assembly arrangements must be explored. Yun et al. [1] applied SA to the OPR-1000 equilibrium-cycle LP problem using STREAM/RAST-K [2] for candidate evaluation. Neural-network models have also been introduced into SA-based LP optimization. Jang et al. [3] used an artificial neural network for in-core fuel management optimization, and a convolutional neural network (CNN) was later applied to LP optimization by SA [4].

In the neural-network approaches, surrogate predictions replace repeated core calculations. Screening methods retain the reference core calculation and instead reduce the number of candidates submitted to it.

Park et al. [5] introduced a screening technique into SA using a lower-cost calculation before the final core evaluation. Machine-learning models have since been used for the same purpose, including a binary classifier for rejecting inappropriate patterns during SA [6], a pin-wise Vision Transformer (ViT) for OPR-1000 LP screening [7], and CNN and ViT surrogates compared within the same screening-based SA framework [8]. These machine-learning screening studies [6-8] were performed for OPR-1000.

In this work, we evaluate a ViT surrogate in an SA-based APR-1400 loading-pattern search. The model predicts MaxFXY and MaxPinBU and screens candidate patterns before RAST-K evaluation. We report prediction error on 54,600 core calculations in the test split, retrospective screening results for 605,483 candidate LPs from a separate SA campaign, and the corresponding estimated computational cost. Because the training labels are generated by core calculations, we also examine the effects of model configuration and training-data size using more than 40 ViT configurations and training sets ranging from 740,000 to 10,000 samples.

2. AI-Based Screening for Loading Pattern Search

2.1 Simulated Annealing and Surrogate Screening

Core calculations are performed with STREAM/RAST-K [2], which provides the 3-D power distribution and the prediction targets MaxFXY and MaxPinBU. SA [12] explores the LP space using assembly swap and rotation operators and evaluates an objective J combining cycle length with penalties for MaxFXY and MaxPinBU margin violations. A candidate that improves J is accepted; otherwise, it is accepted with probability $P = exp\left(-\frac{\Delta J}{T}\right)$ under the cooling schedule. Candidates that pass the surrogate screen are evaluated by RAST-K, and the SA acceptance test is then applied to the RAST-K objective.

2.2 Vision Transformer for Loading Pattern Evaluation

The ViT [9] represents an input as a sequence of patch tokens and uses self-attention [10] to model interactions across the sequence. For LP evaluation, the two-dimensional core map is divided into non-overlapping patches, projected into D-dimensional tokens, and combined with positional information and a CLS token. The classification head is replaced by two regression heads for MaxFXY and MaxPinBU.

3. Surrogate Model and Dataset

3.1 APR-1400 Dataset and Input Representation

The APR-1400 quarter-core LP is represented by a 9x9 assembly map. For the categorical inputs, each assembly entry is expanded to a 2x2 block, while the four burnup values are assigned to the corresponding node positions. Assembly type, shuffling history, and rotation are encoded using embedding dimensions of 32, 16, and 4, respectively, and combined with the burnup input to form an 18x18x53 tensor. The categorical inputs contain 16 assembly types, 37 shuffling states, and 6 rotation states. Two SA campaigns yielded 894,600 labeled samples after removing outliers and duplicates, split into 740,000 training, 100,000 validation, and 54,600 test samples.

3.2 CNN Baseline and Reference ViT

The CNN baseline consists of three convolutional blocks (64, 128, 256 channels, 3x3 kernels, no pooling), followed by two fully connected layers (512 and 256 units) and two output heads. The reference ViT (Fig. 1) uses a 6x6 patch size, token dimension D=256, six encoder layers, eight attention heads, and a feed-forward dimension of 512. It uses a Pre-Norm residual structure [13] and is trained with AdamW [11], cosine annealing, and early stopping. The CNN baseline contains 43.0

million trainable parameters, of which 42.5 million belong to the first fully connected layer because the 18x18x256 feature map is flattened without pooling; the reference ViT contains 3.7 million. The ViT therefore attains lower error with about one twelfth of the parameters. The CNN is the prediction baseline, and the reference ViT is used for the screening results in Section 4. Section 5 varies the ViT architecture and training-data size.

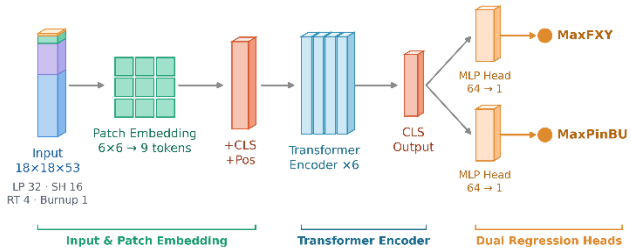


Fig. 1. Architecture of the reference ViT surrogate model

3.3 Training, Evaluation, and Screening Criterion

Architecture comparisons use MaxFXY mean relative error (MRE), Eq. (1):

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{y_i} \times 100 \qquad (1)$$

Training minimized a weighted MaxFXY/MaxPinBU MSE loss (1.0:0.05). MaxFXY MRE was used for early stopping and model comparison. All surrogate-model training and ablation experiments were performed on a single NVIDIA RTX A5000 GPU.

For screening, a two-sided uncertainty range is defined for each target:

$$I_j = \left[\hat{y}_j - 1.96\sigma_j,\ \hat{y}_j + 1.96\sigma_j\right], \qquad (2)$$

$$\hat{y}_j - 1.96\sigma_j \le L_j, \quad \forall j \in \{\text{MaxFXY}, \text{MaxPinBU}\} \qquad (3)$$

where $\hat{y}_j$, σ_j and L_j denote the surrogate prediction, residual standard deviation used for screening, and screening limit for target j, respectively. The two-sided uncertainty range in Eq. (2) is defined using $1.96\sigma_j$ under the assumption of normally distributed residuals. A candidate is passed to RAST-K when the lower bound of the uncertainty range does not exceed the screening limit for both targets, as defined in Eq. (3). Otherwise, the candidate is rejected. The retrospective screening analysis used screening limits of $L_j = 1.8$ for MaxFXY and 61 GWd/MTU for MaxPinBU.

4. Surrogate Performance in Loading Pattern Search

4.1 Prediction Performance

The CNN baseline and reference ViT were evaluated on the same 54,600 sample test set. The reference ViT achieved MaxFXY and MaxPinBU MREs of 0.63% and 0.17%, compared with 1.81% and 0.45% for the CNN baseline. Table I also reports MAE and residual statistics. Residual means are close to zero for both models, while the ViT shows smaller residual standard deviations.

Two structural differences are consistent with the observed gap. The difference is not one of capacity: the ViT uses about one twelfth of the parameters. Three 3x3 convolutions produce features from a 7x7 local receptive field before flattening, whereas self-attention relates all patch tokens within each encoder layer. The CNN also retains spatial position by flattening, which is what produces the 42.5 million parameter first layer, while the ViT encodes position through learned embeddings and shares projection weights across patches. These differences in spatial interaction may contribute to the lower ViT error.

The comparison is not tuning-matched. The sweep in Section 5 was run for the ViT only; both models use the same split, loss weighting and early-stopping rule, but the CNN was not tuned to the same extent. The result shows that the ViT reaches lower error at this parameter count, not that a tuned CNN could not.

Table I. Prediction metrics on the 54,600-sample test set.

Metric	CNN	ViT
MaxFXY MRE (%)	1.81	0.63
MaxPinBU MRE (%)	0.45	0.17
MaxFXY MAE	0.0374	0.0131
MaxFXY resid. (mean±std)	−0.0041±0.0518	+0.0014±0.0308
MaxPinBU MAE (GWd/MTU)	0.2848	0.1053
MaxPinBU resid. (mean±std, GWd/MTU)	−0.0127±0.3822	−0.0098±0.1454

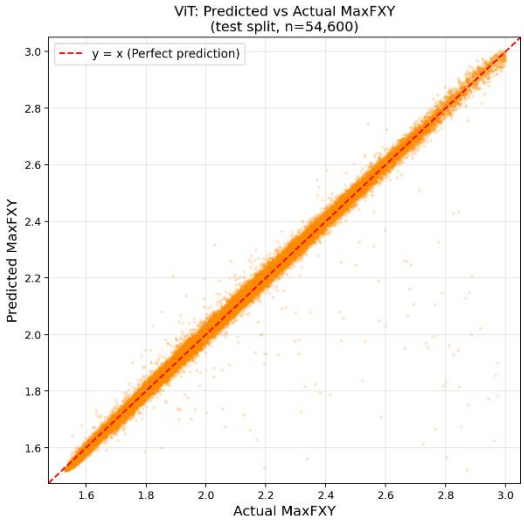


Fig. 2. ViT predicted vs. actual MaxFXY.

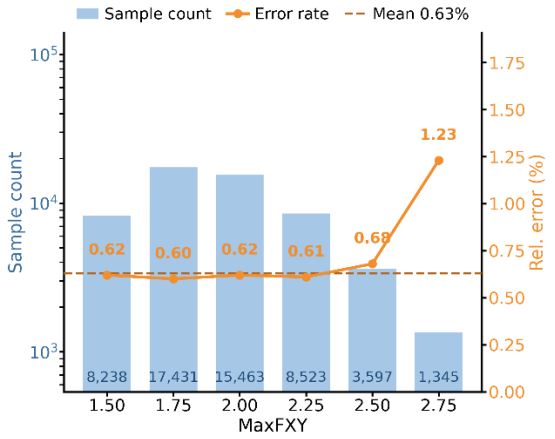


Fig. 3. MaxFXY MRE by actual-value range for the reference ViT (test set, n = 54,600).

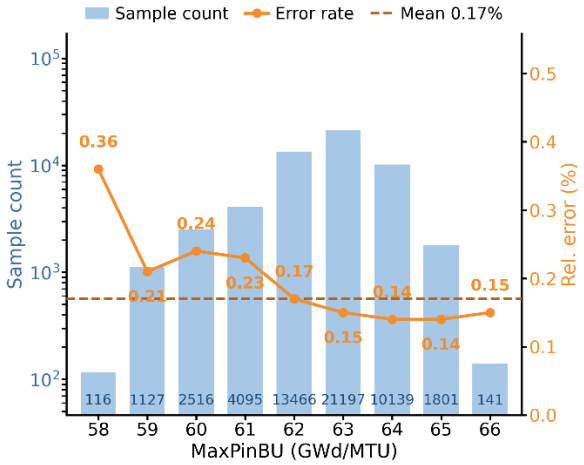


Fig. 4. MaxPinBU MRE by actual-value range for the reference ViT (test set, n = 54,600).

4.2 Screening Behavior

A separate SA campaign generated 605,483 candidate LPs using limits of 2.0 for MaxFXY and 63 GWd/MTU for MaxPinBU. For the present analysis, the screening limits were set to 1.8 and 61 GWd/MTU with k = 1.96, and the criterion in Eq. (3) was applied retrospectively to these candidates.

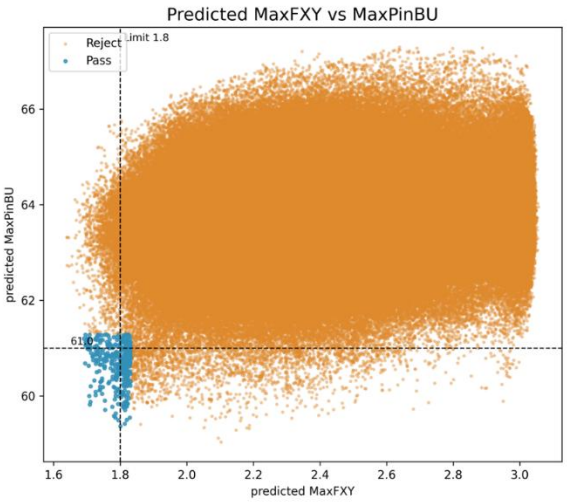


Fig. 5. Predicted MaxFXY and MaxPinBU in the retrospective screening analysis. Pass/reject decisions are based on Eqs. (2) and (3).

Under the revised screening criterion, 605,041 candidates (99.93%) would have been rejected and 442 (0.07%) would have been passed to RAST-K. Of all candidates, 596,419 (98.50%) failed both screening criteria, 3,940 (0.65%) failed only the MaxFXY criterion, and 4,682 (0.77%) failed only the MaxPinBU criterion.

Of the 605,041 candidates rejected under the revised criterion, 586,592 were also rejected by the screen applied during the campaign and were never submitted to RAST-K. The remaining 18,449 had already been evaluated during the campaign and were excluded from the additional audit. The audit therefore focused on the 586,592 candidates for which no core calculation existed. The audit below therefore covers the 586,592 candidates for which no core calculation exists. A stratified sample of 750 was re-evaluated with RAST-K, stratified by the margin by which the screened lower bound exceeds the limit on the MaxFXY scale, with sampling concentrated near the boundary. Of these, 715 completed; 35 diverged during depletion and were excluded. No false rejects were observed among the 715. In the stratum nearest the boundary, which contains 9,793 candidates within 0.08 of the limit, 0 of 381 were feasible, giving a one-sided 95% upper bound of 0.79%, or 77 candidates (Table II).

Table II. Stratified re-evaluation of the 586,592 candidates rejected without a core calculation.

Margin from boundary	Population	Evaluated	Feasible
0 - 0.08	9,793	381	0
0.08 - 0.12	15,319	140	0
0.12 - 0.20	41,065	97	0
0.20 - 0.40	168,221	49	0
> 0.40	352,194	48	0
Total	586,592	715	0

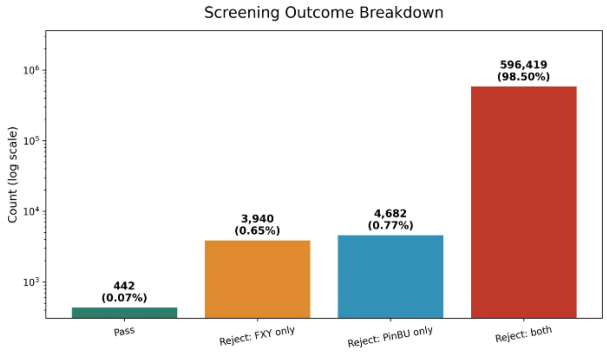


Fig. 6. Retrospective screening outcomes for 605,483 candidate LPs.

4.3 Computational Cost

Surrogate inference averaged 5.64 ms per candidate (p95: 6.89 ms). One RAST-K evaluation on an AMD EPYC 7452 32-core processor (2.35 GHz) takes approximately 280 s. Assuming 40 parallel RAST-K evaluations, the computational cost for all 605,483 candidates is estimated at 49.1 days. Based on the retrospective screening result, the corresponding cost is estimated at 1.81 hours, including surrogate inference over all candidates.

5. Sensitivity to Model Architecture and Training Data

5.1 Architecture Sensitivity

Starting from the reference ViT, more than 40 configurations were evaluated under the fixed 740,000/100,000/54,600 split. The sweep varied patch size (2, 3, 6, 9), encoder depth (4-10), attention heads (4-16), feed-forward dimension (256-1024), embedding dimensions, dropout, weight decay, loss weighting, positional encoding, and SH/RT input features. Selected combinations were also tested at full and reduced data scales.

At 740K samples, reducing dropout from 0.1 to 0.05 and increasing patch size from 6x6 to 9x9 gave the lowest validation MRE, 0.546%, compared with 0.626% for the reference configuration. Increasing depth to 10 layers (0.573%) and removing the RT channel (0.570%) also improved the single-split result. A five-factor combination reached 0.612% and did not improve on the best one- or two-factor changes.

Weight decay, loss weighting, and fixed sinusoidal positional encoding gave MREs of 0.606-0.622%, 0.598-0.624%, and 0.675%, respectively. Table III lists selected results at 740K and 100K.

Table III(a). Selected validation MaxFXY MRE results at 740K training samples.

Configuration	Error (%)	vs. Ref.
Reference	**0.626**	—
Layers10 + Heads16	0.587	↓
Heads 8→16	0.580	↓
Layers 6→10	0.573	↓
Remove rotation channel	0.570	↓
Patch 6→9	0.554	↓
Dropout 0.1→0.05	0.550	↓
Dropout 0.05 + Patch 9	0.546	↓

Table III(b). Selected validation MaxFXY MRE results at 100K training samples.

Configuration	Error (%)	vs. Ref.
Remove SH + RT	1.253	↑
Remove SH channel	1.113	↑
Reference	**0.963**	—
Dropout 0.05 + Patch9	0.831	↓
Heads 8→16	0.803	↓
Dropout 0.1→0.05	0.763	↓
Shrunk width (0.85M params)	0.745	↓
Layers 6→10	0.741	↓

5.2 Effect of Training-Data Scale

For the reference ViT, MaxFXY MRE increased monotonically as the training set was reduced from 740K to 10K, with the steepest degradation below 100K (Table IV).

Table IV. Reference ViT validation MaxFXY MRE vs. training-set size.

Samples	740K	500K	250K	100K	50K	10K
Error (%)	0.626	0.660	0.812	0.963	1.442	3.068

Selected full-data and 100K configurations were then examined by five-fold cross-validation using common test partitions within each fold (Table V).

Table V. Validation MRE from the single split and test MRE from five fold cross validation for selected configurations (740K and 615,680 training samples, respectively).

Data scale	Configuration	Single-split (val)	5-fold test mean±std
Full data	Reference (patch6)	0.626%	0.658%±0.035%
Full data	Dropout .05+Patch9	0.546%	0.613%±0.014%
100K	Reference (patch6)	0.963%	0.997%±0.108%
100K	Layers 6→10	0.741%	0.822%±0.046%
100K	Shrunk (0.85M)	0.745%	0.968%±0.062%

Five-fold cross-validation retained the gain from increased depth at 100K (0.997% to 0.822%), whereas the apparent gain from width reduction was much smaller (0.997% to 0.968%). The single split models used 740,000 training samples, whereas each five fold split used 615,680 training, 100,000 validation, and 178,920 test samples.

Feature effects also depended on the tested configuration. Removing RT from the 740K reference improved MRE from 0.626% to 0.570%, whereas removing RT from the 100K dropout+patch9 configuration worsened it from 0.831% to 0.949%. Because the base architectures differ, this contrast cannot be attributed to data scale alone. Removing SH was detrimental in the tested reduced-data configurations (1.113-1.366%).

Combining the 100K narrow-width and depth-10 changes produced 1.065%, worse than either change alone (0.745% and 0.741%). The same combination improved on the reference at 50K (1.350% vs. 1.442%) and 250K (0.730% vs. 0.812%).

The current random split can place samples from the

same SA trajectory in different partitions, so trajectory-level independence is not ensured. Most configurations were trained once, and five-fold cross-validation was applied only to configurations chosen from the single-split results.

6. Conclusion and Future Work

The reference ViT achieved MaxFXY and MaxPinBU MREs of 0.63% and 0.17% on the test set of 54,600 samples. In a retrospective analysis of 605,483 candidate LPs, 442 candidates (0.07%) satisfied the revised screening criterion. A stratified re-evaluation of 715 rejected candidates produced no false rejects, with a one-sided 95% upper bound of 0.79% in the stratum nearest the boundary. The corresponding computational cost was estimated at 1.81 hours with 40 parallel RAST-K processes, compared with 49.1 days for evaluating all candidates by RAST-K. The relative ranking of the tested configurations changed between the full-data and 100K cases: increasing encoder depth remained beneficial at 100K under five-fold cross-validation, while the single-split gain from width reduction was much smaller.

Future work will evaluate the model using trajectory-independent LP data, extend the architecture and data-scale study to the CNN baseline under a matched tuning budget, and compare screened SA with conventional SA. Model applicability outside the current training domain will also be examined.

ACKNOWLEDGMENT

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Climate, Energy & Environment (MCEE) of the Republic of Korea (No. RS-2025-16063033).

REFERENCES

[1] D. Yun, D.-Y. Kim, H. Kim, C. Kong, and D. Lee, "Fuel Loading Pattern Optimization for OPR-1000 Equilibrium Cycle by Simulated Annealing Algorithm," Transactions of the Korean Nuclear Society Virtual Spring Meeting, 2020.

[2] J. Park, J. Jang, H. Kim, J. Choe, D. Yun, P. Zhang, A. Cherezov, and D. Lee, "RAST-K v2-Three-dimensional nodal diffusion code for pressurized water reactor core analysis," Energies, Vol. 13, p. 6324, 2020.

[3] C. S. Jang, H. J. Shim, and C. H. Kim, "Optimization Layer by Layer Networks for In-core Fuel Management Optimization Computations in PWRs," Annals of Nuclear Energy, Vol. 28, pp. 1115-1132, 2001.

[4] H. Jang, H. C. Shin, D.-Y. Kim, and H. C. Lee, "Application of Convolutional Neural Network to Fuel Loading Pattern Optimization by Simulated Annealing," Transactions of the Korean Nuclear Society Virtual Autumn Meeting, 2020.

[5] T. K. Park, H. G. Joo, C. H. Kim, and H. C. Lee, "Multiobjective Loading Pattern Optimization by Simulated Annealing Employing Discontinuous Penalty Function and Screening Technique," Nuclear Science and Engineering, Vol. 162, pp. 134-147, 2009.

[6] G.-H. Sim, M.-G. Park, G.-R. Bae, and J.-U. Sohn, "Loading Pattern Optimization Using Simulated Annealing and Binary Machine Learning Pre-screening," Nuclear Engineering and Technology, Vol. 56, pp. 1672-1678, 2024.

[7] S. Jeong and H. C. Lee, "Loading Pattern Optimization for OPR-1000 by Simulated Annealing with a Screening Technique using Pin-wise Vision Transformer Model," Transactions of the Korean Nuclear Society Spring Meeting, 2025.

[8] S. Jeong and H. C. Lee, "Effect of Surrogate Model on Screening-based Simulated Annealing for OPR-1000 Loading Pattern Optimization," Transactions of the Korean Nuclear Society Spring Meeting, 2026.

[9] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021.

[10] A. Vaswani et al., "Attention is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.

[11] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," International Conference on Learning Representations (ICLR), 2019.

[12] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," Science, Vol. 220, No. 4598, pp. 671-680, 1983.

[13] R. Xiong et al., "On Layer Normalization in the Transformer Architecture," International Conference on Machine Learning (ICML), 2020.