Dimensionality Reduction with Variational Autoencoders for Multiobjective RF Cavity Optimization in Ring Cyclotron

A. H. Shali*,[1], M. Fukuda, T. Yorita, H. Kanda, Y. Matsuda
T. H. Chong, H. Zhao, S. Matsui, T. Imura, N. Itakura, S. Ishihata, T. Tsujisaka
Research Center for Nuclear Physics Osaka University, Osaka, Japan
[1]also at National Research and Innovation Agency, Jakarta, Indonesia

Abstract

The study of a variational autoencoder-assisted multi-objective optimization scheme for the design of RF cavities in a ring cyclotron is presented. The cavity geometry is parameterized using Non-Uniform Rational B-Splines (NURBS), with control point positions and weights acting as the design space. A variational autoencoder is trained to learn low-dimensional latent representation of NURBS parameters, allowing optimization processes to run efficiently. Multi-objective optimization is performed in the latent space using differential evolution, with objective functions evaluated using an ensemble of neural network surrogate models trained on eigenmode simulations from Ansys HFSS. Both the surrogate model and the variational autoencoder are periodically retrained during the optimization process using new simulation data, improving the accuracy. This approach shows the possibility of combining dimensional reduction with surrogate-assisted optimization for complex RF cavity design of a particle accelerator.

INTRODUCTION

Variational autoencoders (VAEs) are a type of artificial neural network architecture capable of learning a compact and continuous latent representation of high-dimensionality data [1]. Although VAEs were originally developed for unsupervised learning and data generation, they have also found applications in design optimization by means of efficient complex parameter space exploration through latent space manipulation [2]. VAEs can serve as dimensionality reduction tools that preserve design diversity while reducing the decision space. This is useful for engineering design problems with costly numerical simulation, such as the optimization of radio frequency (RF) cavities in particle accelerators [3].

RF cavity design, except for the very high frequency ones [4], generally involves adjusting the shape of the cavity [5]. For a given cavity material, some quantity of interests, such as the resonant frequency, shunt impedance, quality factor, depend significantly on the geometrical shape of the cavity. Optimizing the shape of the cavity is therefore an important step to achieve an efficient and high performing cavity under various constraint limits.

In this paper, we propose a constrained multi-objective evolutionary algorithm with dimensionality reduction, named Latent-DE-NSGA-II. As the name implies, the mutation and crossover operations are performed using differential evolution [6] while non-dominated sorting (NSGA-II) is used as the selection algorithm [7]. The optimization framework was extended from our previous study on the optimization of an RF cavity of a ring cyclotron with ensemble neural network surrogate model as the objective function evaluator [8]. The focus of this paper is the incorporation of dimensionality reduction using VAE and evaluating its performance compared to the standard DE-NSGA-II scheme.

METHODS

The optimization of RF cavity using Latent-DE-NSGA-II presented in this paper was carried out in five stages. They are cavity design parameterization, data generation using numerical codes, neural network surrogate training, variational autoencoder training, and optimization process using neural network. Each of the stages will be explained in this section.

Design Parameterization and Data Generation

Electromagnetic field in a closed perfectly conducting surface is governed by Maxwell equation without source. Time dependence of the field is assumed to be harmonic, that is $\psi(\vec{x},t) = \psi(\vec{x})\exp(-i\omega t)$ with ψ representing the components of electric and magnetic field and ω is the angular frequency of the oscillation. This will result in Helmholtz equation given in Eq. (1) below [9]

$$\left(\nabla^2 + \mu\varepsilon\omega^2\right)\psi = 0. \qquad (1)$$

For perfectly conducting surface, the electric field and magnetic field inside the cavity must be perpendicular and parallel to the surface of the cavity, respectively. Specifically $\hat{n}\times\vec{E} = 0$ and $\hat{n}\cdot\vec{H} = 0$, with $\hat{n}$ is the normal vector of the cavity surface. Assuming that the cavity is filled entirely with vacuum, the field distribution and properties will entirely be defined by the shape of the cavity (the boundary conditions). Therefore, optimizing an RF cavity in this sense would correspond to optimizing the shape of the RF cavity.

In this paper, the shape of the cavity is parameterized using *non-uniform rational b-splines* (NURBS) method, commonly used in aeronautical and electromagnetic engineering [10]. In this case, the use of NURBS parameter allows the cavity to have a smooth and rich design variation while having discrete inputs. The discrete inputs are important because it can be used as the input of both neural network surrogate (for objective function evaluator), and the evolutionary algorithm for optimization. A curve parameterized

[*] ahsahafi@rcnp.osaka-u.ac.jp

by NURBS are defined by several parameters, they are the number of control points, the position and weight of each control point, the knot vector, and the degree of the NURBS. The details of how NURBS works is avaliable at [11].

In this research, a fourth order NURBS curve is only defined on one quadrant of the cavity cross section (on an x-y plane). The full cross section is obtained by first mirroring the curve with respect to y-axis continued by second mirroring by x-axis. The full cavity shape is then obtained by extruding the cross section along the z-axis. The full NURBS setup, including control point configuration and parameter ranges, is not shown here for brevity and can be found in our previous work [8]. An example of NURBS curve representing cavity shape is given in Fig. 1.

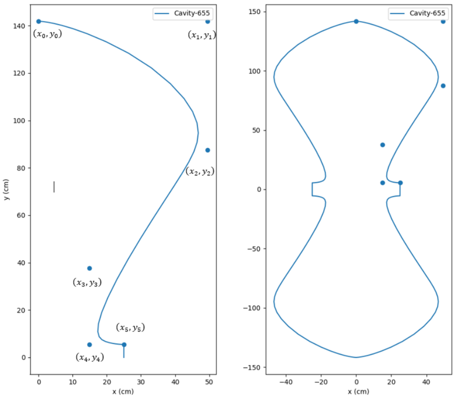

Figure 1: Cross sectional cavity shape generated using NURBS.

The cavity shape obtained using NURBS parameter is then rewritten in a format understood by Ansys HFSS software. 20000 cavity variations are sampled using latin hypercube sampling [12] then numerical calculation to solve the Helmholtz equation with the cavity walls as the corresponding boundary condition is performed. Several quantity of interests are extracted from the simulation, they are the lowest resonant frequency (f), the maximum electric field divided by accelerating gradient (MaxE/Vacc), and the shunt impedance (R). Both of the shunt impedance and accelerating gradient are calculated using the same integral line, which stretches from the midpoint of one accelerating gap to the midpoint of the other accelerating gap. All of the input parameters and output parameters are then normalized for training data of the neural network. For shunt impedance, inversion operation is performed first (resulting in $1/R$) since it is easier to do minimization during the optimization procedure.

Neural Network Surrogate Model and Variational Autoencoder for Dimensionality Reduction

A standard artificial neural network can be thought of a nonlinear curve fitting algorithm capable of approximating any function given the enough neural network parameter [1]. Given enough data, it can be trained to replace a conventional and more computationally expensive conventional solver, thus the name surrogate model [13]. Training in this context means adjusting the model parameters (called the connection weights) so that the neural network prediction matches the data [1]. However, there is a problem of data overfitting where the network could predict the training data correctly but fail to generalize. When the network has too many parameters for a given dataset, there are many combination of neural network parameter that could give correct output prediction. Thus, it is desirable to know the uncertainty of neural network prediction.

In this paper, the uncertainty of neural network prediction is obtained by using ensemble neural network. There are several types of ensemble neural network, such as bagging, boosting, or stacking [14]. The one used in this paper is bagging, specifically by training fifty identically structured neural network with different connection weight initialization. This way, the network will still give a similar prediction for the output dataset, but the difference of them on the data-scarce region can be used to quantify the uncertainty. Specifically, ensemble neural network calculates the average and variance of each output parameter. The optimization algorithm can be modified to prioritize the prediction with low uncertainty, making the optimized result more reliable. The network architecture used in this research is shown in Fig. 2.

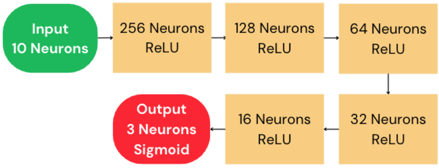

Figure 2: Neural network surrogate model trained on Ansys HFSS data.

An autoencoder is a type of artificial neural network aimed for dimensionality reduction or feature learning [1]. The main idea of an autoencoder is that the network is trained to reconstruct the input. The middle hidden layer part of the network can be made to have smaller dimension (smaller number of neurons). This bottleneck is called the latent space, while the layers before and after the latent space layer is called the encoder and decoder, respectively. For dimensionality reduction of optimization algorithm search space, the encoder part can be discarded and the optimization is run entirely on the latent space, called latent space optimization (LSO) [2]. The latent space input is first decoded and the result is then passed to objective function evaluator. Note that the autoencoder does not need to be able to perfectly reconstruct the input. As long as there is some degree of consistency between the latent space and the reconstructed space, and that the variation of latent space is diverse enough, optimization can be performed.

There is a weakness to autoencoder when it is to be applied for dimensionality reduction. The fact that the latent space is not regularized means that it is possible that there are pockets of empty areas between cluster of encoded data in the latent space, which will return an unfeasible result when decoded. Variational autoencoder solve this problem by enforcing the latent space to have multivariate gaussian distribution centered around zeros and variance equal to one [15]. It is done by setting the encoder to output both the means $\vec{\mu}$ and variances $\vec{\sigma^2}$ with the dimension equal to the dimension of the latent space. After that, the latent space is randomly sampled with Gaussian distribution, as shown in Fig. 3. Kullback-Leibler (KL) divergence is added along with reconstruction loss to enforce $\vec{\mu} = \vec{0}$ and $\vec{\sigma^2} = \vec{1}$.

VAE is used to reduce the dimensionality of search space from 10 to 4. The encoder consists of three fully connected layers, with 256, 128, and 64 neurons, respectively. All of the layers are activated using rectified linear unit (ReLU) function. The decoder is identical to the encoder, but the order of the layers is reversed (64, 128, and 256). A sigmoid function is applied at the output to enforce the reconstructed input values within [0,1] range.

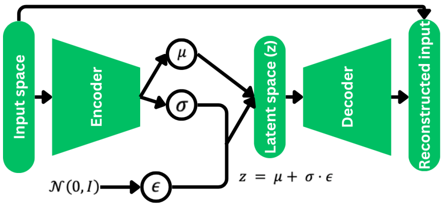

Figure 3: Standard VAE architecture.

Optimization Algorithm

In a multi-objective algorithm, the aim is to find the Pareto front, a subspace of decision space where all of the samples cannot be said to be better compared with each other [7]. When optimization is constrained, an additional condition is imposed where constraint-violating samples (infeasible) are discarded. One of popular algorithm class for constrained multi-objective problem is the evolutionary algorithm.

In this research, differential evolution based non-dominated sorting algorithm (DE-NSGA-II) is used to perform optimization. Differential evolution is used as the means for generating new samples through mutation and crossover operations. Specifically, DE/rand/1/bin is used because of its capability to avoid convergence to local minima. Parent and offspring samples are then pooled together and then selected using NSGA-II, maintaining the population size. Constraints are handled using constraint-dominance scheme, where constraint satisfying samples are prioritized over good performing but infeasible samples.

The objective of the optimization is to minimize both $1/R$ and MaxE/Vacc simultaneously, while keeping the resonant frequency constrained to some value. Here, the frequency is constrained to 68.9 MHz, with a tolerance of 0.35 MHz. Population size is fixed to 100, with the mutation rate and the crossover rate set to 0.5 and 0.9 respectively. Additionally, the result of Latent-DE-NSGA-II is compared to standard DE-NSGA-II (without dimensionality reduction) at population generation equal to 20 and 500. The first one is used to check how quickly the algorithms converge, while the second one is used for final convergence comparison.

The objective functions are evaluated using the combination of mean and variance from the ensemble neural network [16], as shown in Eq. (2). The constant η indicates how much the uncertainty is prioritized. As the optimization algorithm will try to minimize both mean and variance, a small η value means that prediction with a small uncertainty is not prioritized, and vice versa.

$$\mathrm{Obj}(x) = \mathrm{mean}(x) + \eta\sqrt{\mathrm{var}(x)} \tag{2}$$

RESULTS

Desktop workstation with an Intel Core i9-14900KF processor was used to generate the data. Simulating 20000 cavity samples using eigenmode solver on Ansys HFSS took about 4 days and 15 hours. NVIDIA T4 GPU on Google Colab was used for both ensemble neural networks and VAE, which took 75 and 60 minutes, respectively.

The results related to data generation and ensemble neural network training are available in the previous publication [8] and will not be repeated here. The variational auto-encoder was trained for 4000 epochs, with the use of KL divergence loss annealing done up to 2000 epochs. 50000 data points randomly sampled between $x = [0, 1]$ were trained with fixed learning rate $l_r = 0.0005$. Total training loss were shown in Fig. 4. Histogram shown in Fig. 5 shows how two points in latent space are related to two points in reconstructed space. A small spread of data indicates that two close latent space points are also close in the reconstructed space, meaning that the search space is smooth. However, zero spread in the distribution is not possible due to VAE compression.

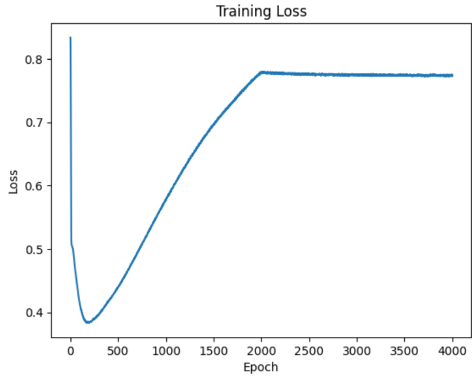

Figure 4: VAE training loss vs epoch with KL-loss annealing.

Active learning mechanism is adopted to improve the surrogate model predictive performance around the given frequency constraint. Optimization with 200 samples and 100 generation are performed 50 times. To encourage exploration, frequency tolerance is initially set higher (3.5 MHz) and slowly decreased to 0.35 MHz. The uncertainty constant is set low $\eta = 1$ to prioritize mean prediction rather the uncertainty. Eigenmode simulation using Ansys HFSS is then

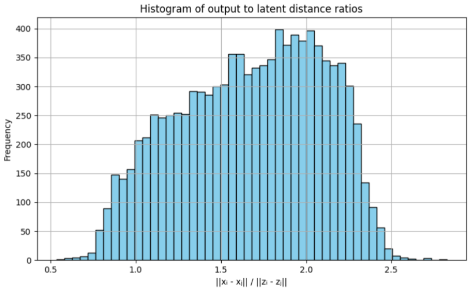

Figure 5: Histogram of output to latent distance ratio $||x_i - x_j||/||z_i - z_j||$.

performed on 542 samples belonging to the combined Pareto front from each run. The network is then retrained using newly obtained HFSS data. Figure 6 shows the comparison plot of neural network and HFSS predictions. MaxE/Vacc prediction by the surrogate model is less accurate compared to inverse shunt impedance, which is in accordance to the result obtained in [8]. However, the network tends to underestimate the inverse shunt impedance value for small inverse shunt impedance. Among all of the samples, inverse shunt impedance by HFSS did not go below about $1.3 \times 10^{-7}\Omega^{-1}$.

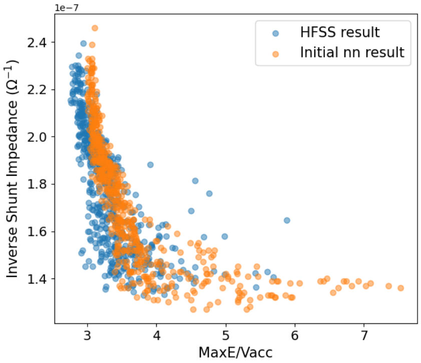

Figure 6: Scatter plot of surrogate model predictions vs Ansys HFSS simulation results for initial training dataset.

Optimization comparison between compressed and uncompressed input dimensionality was performed after ensemble neural network retraining. As previously mentioned, two cases are considered here: low function calls (100 samples for 20 generations, which equals to 4000 function calls) and high function calls (100 samples for 500 generations, which equals to 100000 function calls). For this case, the uncertainty constant is set to be $\eta = 2$ which decrease the tendency of the optimizer to take samples with uncertain prediction. Figure 7 shows the objective space plot after 50 generations for Latent-DE-NSGA-II and standard DE-NSGA-II scheme. It can be seen that due to tight frequency constraint imposed from the very beginning of the optimization, it is difficult for both schemes to achieve convergence quickly. However, it is quite clear that the objective space of Latent-DE-NSGA-II scheme is closer to the Pareto front compared compared to the standard DE-NSGA-II. However,

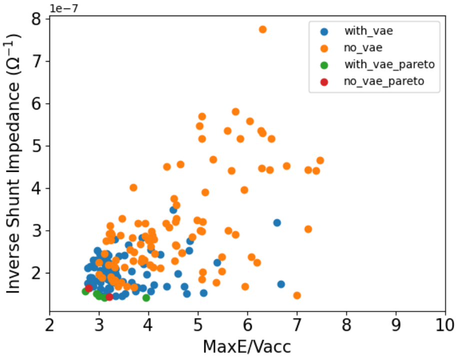

Figure 7: Scatter plot of objective space for low function calls of 100 population size and 20 generations.

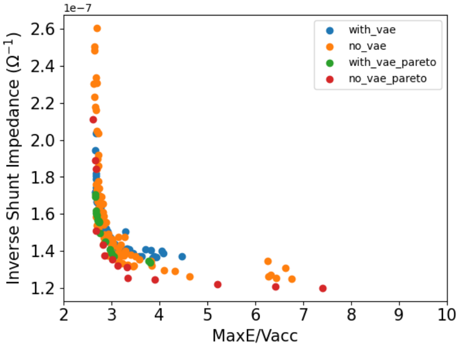

Figure 8: Scatter plot of objective space for high function calls of 100 population size and 500 generations.

after 500 iterations, Latent-DE-NSGA-II reached similar pareto front compared to standard DE-NSGA-II, indicating that Latent-DE-NSGA-II could also reach the correct Pareto front (with respect to the surrogate model). However, the region of low inverse shunt impedance is dominated by samples from standard DE-NSGA-II. This might be caused by the inability of the VAE to reconstruct a certain combination of input. This can be mitigated by retraining the VAE using the input of the best performing samples, and not using randomly sampled data [17].

The cavity cross-sectional shape belonging to Latent-DE-NSGA-II Pareto front are shown in Fig. 9. Cavity number 13 has a smaller MaxE/Vacc of 2.68 with slightly higher inverse shunt impedance of $1.6 \times 10^{-7}\Omega$, while cavity number 14 has a larger MaxE/Vacc of 3.045 with lower inverse shunt impedance of $1.37 \times 10^{-7}\Omega$. Both design satisfy the frequency constraint with Cavity 13 and 14 has a resonant frequency of 69.1 MHz and 68.6 MHz, respectively. It is evident that cavity with tighter curvature around the cone will produce higher shunt impedance but also higher MaxE/Vacc. The performance comparison mentioned above shows that Latent-DE-NSGA-II will converge faster compared to the standard-DE-NSGA-II, even though the pareto front is not fully covered especially for low inverse shunt impedance region. However, this scheme is advantageous for expensive to evaluate function, such as simulation without surrogate.

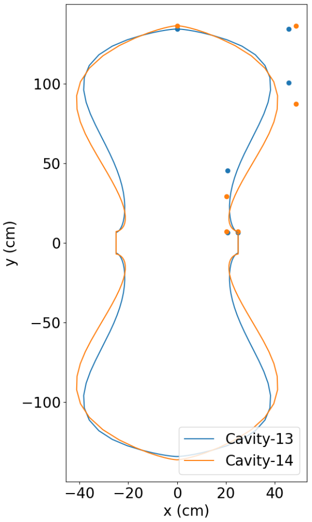

Figure 9: Two cavities belonging Latent-DE-NSGA-II Pareto front with opposing qualities.

After several iterations, the Latent-DE-NSGA-II result can be relayed to the standard scheme. Further investigation is still required to check if similar results are also obtained using different mutation and crossover method such as by using SHADE [18].

CONCLUSION

We have investigated a constrained multiobjective latent space optimization using Latent-DE-NSGA-II and applied it to the design of RF cavities for a ring cyclotron. The results show that dimensionality reduction via a variational autoencoder accelerates convergence speed towards the Pareto front. However, the final Pareto front of Latent-DE-NSGA-II is partially dominated by that of standard DE-NSGA-II. Therefore, the scheme is useful for expensive to evaluate problems, and its solution after some generation can be relayed to the standard solver for further refinement if desired.

REFERENCES

[1] C. M. Bishop and H. Bishop, *Deep learning: Foundations and concepts*. Springer Nature, 2023.

[2] A. Tripp, E. Daxberger, and J. M. Hernández-Lobato, “Sample-efficient optimization in the latent space of deep generative models via weighted retraining,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 259–11 272, 2020.

[3] S. Smith, M. Southerby, S. Setiniyaz, R. Apsimon, and G. Burt, “Multiobjective optimization and pareto front visualization techniques applied to normal conducting rf accelerating structures,” *Physical Review Accelerators and Beams*, vol. 25, no. 6, p. 062 002, 2022.

[4] Ö. Apsimon, G. Burt, R. B. Appleby, R. J. Apsimon, D. M. Graham, and S. P. Jamison, “Six-dimensional phase space preservation in a terahertz-driven multistage dielectric-lined rectangular waveguide accelerator,” *Physical Review Accelerators and Beams*, vol. 24, no. 12, p. 121 303, 2021.

[5] M. Kranjčević, S. Gorgi Zadeh, A. Adelmann, P. Arbenz, and U. Van Rienen, “Constrained multiobjective shape optimization of superconducting rf cavities considering robustness against geometric perturbations,” *Physical Review Accelerators and Beams*, vol. 22, no. 12, p. 122 001, 2019.

[6] R. Storn and K. Price, “Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces,” *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[7] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, “A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii,” in *International conference on parallel problem solving from nature*, Springer, 2000, pp. 849–858.

[8] A. Shali, M. Fukuda, T. Yorita, and H. Kanda, “Multi-objective optimization of ring cyclotron rf cavity using neural network ensembles with uncertainty quantification,” in *International Particle Accelerator Conference*, 2025. `doi:10.18429/JACoW-IPAC25-WEPS066`

[9] J. D. Jackson, *Classical electrodynamics*. John Wiley & Sons, 2021.

[10] R. Sevilla, S. Fernández-Méndez, and A. Huerta, “Nurbs-enhanced finite element method (nefem),” *International Journal for Numerical Methods in Engineering*, vol. 76, no. 1, pp. 56–83, 2008.

[11] L. Piegl and W. Tiller, *The NURBS book*. Springer Science & Business Media, 2012.

[12] M. D. McKay, R. J. Beckman, and W. J. Conover, “A comparison of three methods for selecting values of input variables in the analysis of output from a computer code,” *Technometrics*, vol. 42, no. 1, pp. 55–61, 2000.

[13] A. Edelen, N. Neveu, M. Frey, Y. Huber, C. Mayes, and A. Adelmann, “Machine learning for orders of magnitude speedup in multiobjective optimization of particle accelerator systems,” *Physical Review Accelerators and Beams*, vol. 23, no. 4, p. 044 601, 2020.

[14] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, “Ensemble deep learning: A review,” *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105 151, 2022.

[15] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[16] P. Putek, S. G. Zadeh, and U. van Rienen, “Multi-objective shape optimization of tesla-like cavities: Addressing stochastic maxwell’s eigenproblem constraints,” *Journal of Computational Physics*, vol. 513, p. 113 125, 2024.

[17] P. J. Bentley, S. L. Lim, A. Gaier, and L. Tran, “Evolving through the looking glass: Learning improved search spaces with variational autoencoders,” in *International Conference on Parallel Problem Solving from Nature*, Springer, 2022, pp. 371–384.

[18] R. Tanabe and A. Fukunaga, “Success-history based parameter adaptation for differential evolution,” in *2013 IEEE congress on evolutionary computation*, IEEE, 2013, pp. 71–78.