ML Lab Research Proposal

Non-Sinusoidal DTLMUW Waveform Discovery via Variational Autoencoder and Bayesian Optimization for Turbulent Drag Reduction

Team: Adhiraj Mann (solo)

Date: May 24th, 2026

Project: VAE-BO Waveform Optimizer for DTLMUW Drag Reduction

Target Compute: 1x AMD Instinct MI300X GPU (of available 8), approx. 2-week window

---

1. PROBLEM STATEMENT

Dolphin-skin-inspired Downstream-Traveling Longitudinal Micro-Ultrasonic Waves (DTLMUWs) have recently demonstrated drag reductions of up to 90% on airfoil surfaces by inducing a dynamic Stokes boundary layer that disrupts near-wall turbulent structures. However, virtually all existing DTLMUW research assumes the wall oscillates in a purely sinusoidal waveform—a mathematically convenient but physically arbitrary choice.

Prior DNS studies of spanwise wall oscillations have shown that non-sinusoidal waveforms can outperform the sinusoid when actuator constraints prevent reaching optimal operating conditions (Cimarelli et al., 2013), yet no work has applied machine learning to discover optimal non-sinusoidal shapes for out-of-plane (DTLMUW-type) oscillations. The combinatorial space of possible waveform shapes is far too large to explore exhaustively with expensive DNS simulations. This project addresses that gap: we propose a generative ML pipeline that learns the structure of the waveform shape space from existing DNS data and proposes novel, physically plausible non-sinusoidal DTLMUW waveform candidates predicted to outperform the sinusoidal baseline—reducing the number of expensive DNS evaluations required by orders of magnitude.

2. PROPOSED APPROACH

I propose a two-stage pipeline combining a Variational Autoencoder (VAE) with a Bayesian Optimization (BO) surrogate in latent space.

Stage 1 - VAE: Learning the Waveform Shape Space

A VAE is trained on a combined dataset of (a) real DNS waveforms from published literature with known drag reduction labels, and (b) synthetically generated waveforms produced via Fourier series superposition with randomized harmonic amplitudes and phase shifts. The VAE encoder compresses each 64-point wall velocity time series into a low-dimensional latent vector (8-16 dimensions); the decoder reconstructs the waveform from those coordinates. The forced continuity of the latent space ensures that coordinates between known waveforms decode into physically plausible interpolated shapes.

Stage 2 - Surrogate Predictor + Bayesian Optimization

A Gaussian Process Regression (GPR) surrogate is trained to map latent coordinates to predicted drag reduction % using only the labeled DNS subset. Bayesian optimization then navigates the learned latent space using an Expected Improvement acquisition function, identifying coordinates predicted to maximize drag reduction per unit actuator power. Promising coordinates are decoded back into waveform shapes—novel non-sinusoidal profiles never explicitly present in the training data—which constitute the paper's proposed candidates for future DNS validation.

Why This Approach Over Alternatives

- **Plain Neural Network Regressor:** Could predict DR% for known waveforms but cannot generate new ones.
- **GAN / Diffusion Model:** Could generate novel waveforms but requires far more training data than DNS studies provide and is harder to interpret.
- **VAE + BO Stack:** Uniquely suited to small, labeled datasets (tens to low hundreds of DNS samples), produces interpretable latent representations, and has strong precedent in engineering design optimization. Applying it to out-of-plane DTLMUW oscillations—rather than the in-plane spanwise oscillations covered by all prior ML work—is the primary novelty of this paper.

3. DATASET

Subset	Source	Approx. Size	Labels?	Role
Labeled DNS non-sinusoidal	Cimarelli et al. (2013), Physics of Fluids	~90-150 waveforms (9 families x amplitude/period grid)	Yes (DR% and net power saving%)	Train predictor + VAE
Labeled DNS sinusoidal	Gatti & Quadrio (2016) parametric study	~200-500 samples (subset extractable from paper)	Yes (DR%)	Train predictor + VAE
Synthetic unlabeled	Generated via Fourier superposition (randomized harmonics/phases)	~10,000 waveforms	No	VAE shape-space training only

Raw size for all subsets is essentially negligible, only a couple of MB, since the data consists of 1D physics simulations rather than raw sensor recordings.

One Row of the Dataset

Each labeled row contains: 64 normalized wall velocity values ($w_1 ... w_{64}$) sampled across one oscillation cycle; physical parameters amplitude A^+, period T^+, Reynolds number Re; waveform family label; source reference; and output targets DR% and net power saving%. Unlabeled synthetic rows contain only the 64 velocity values and physical parameters; DR% fields are null.

Preprocessing

- Normalize each waveform so max |w| = 1.0 to ensure amplitude-invariant shape encoding.
- Filter synthetic waveforms exceeding realistic actuator acceleration limits.
- Train/validation/test split of 70/15/15 applied to labeled rows only; synthetic waveforms used exclusively for VAE training.

4. EVALUATION METRICS

Category	Metric	Target / Interpretation
VAE reconstruction	Mean Absolute Error (MAE) on held-out waveforms	Low MAE confirms decoder accurately reconstructs unseen shapes
VAE reconstruction	Latent space smoothness (nearest-neighbor interpolation quality)	Interpolated waveforms should be physically coherent, not noisy

DR% predictor	MAE and R^2 on test set	$R^2 > 0.85$ required before BO is trusted
DR% predictor	95% confidence interval calibration	GPR uncertainty estimates should be well-calibrated for BO to work
Generated waveforms	Predicted DR% vs. sinusoidal baseline	Primary claim: at least one novel waveform exceeds best sinusoidal DR%
Generated waveforms	Net power saving% of proposed candidates	Proposed waveforms must have positive net power saving to be practically relevant
Ablation	VAE latent dimension sensitivity (4, 8, 16, 32)	Identifies optimal compression level for this dataset size

Note: Because proposed waveforms are not validated with new DNS in this study, the headline scientific claim is framed as 'predicted to outperform' rather than 'demonstrated to outperform'. DNS validation is explicitly scoped as future work.

5. COMPUTE REQUIREMENTS

Phase	Hardware	Est. Wall-Clock	Notes
Synthetic data generation (10,000 Fourier waveforms)	CPU only	< 1 minute	Pure NumPy - no GPU needed
VAE training (~10,200 waveforms, 64-point series, latent dim 8-16)	1x AMD Instinct MI300X	Minutes to ~1 hour	Very lightweight 1-D time series VAE; 192 GB HBM3 far exceeds requirements
GPR surrogate training (~200-500 labeled samples)	1x AMD Instinct MI300X	< 5 minutes	GPR on small tabular data is computationally trivial
Bayesian optimization loop (~500-1,000 acquisition evaluations)	1x AMD Instinct MI300X	< 1 hour	Each evaluation queries the surrogate, not DNS; extremely fast
Latent dimension ablations (4 runs, 5 seeds)	1x AMD Instinct MI300X	2-4 hours total	Parallelizable across seeds
Total	**1x AMD Instinct MI300X**	**~1 wall-clock day**	**Well within 2-week window; remaining 7 GPUs available for extensions**

Software Dependencies

- ROCm 6.x + PyTorch ROCm wheel
- PyTorch, scikit-learn (GPR), NumPy, SciPy
- No CUDA-only kernels - all operations use standard torch and numpy
- No Docker required; pure Python venv

The AMD Instinct MI300X's 192 GB HBM3 is substantially more than required for this workload. The extra headroom allows exploration of larger VAE architectures or higher latent dimensions as a stretch goal without memory constraints.

6. EXPECTED DELIVERABLES

- **Trained VAE Model Checkpoint:** Encoder/decoder weights, latent space visualization (2-D PCA projection of all waveforms colored by DR%), and reconstruction error table across waveform families.
- **Trained GPR Surrogate:** Complete weights with calibration plot and R^2/MAE on held-out test set.
- **Novel Waveform Proposals:** A set of 10-20 candidate waveforms identified by the BO optimizer, reported as 64-point time series with predicted DR% and net power saving%, compared against the best sinusoidal baseline.
- **Ablation Results Table:** Latent dimension sensitivity and impact of synthetic data augmentation on VAE reconstruction quality.
- **Research Paper Draft:** Written in standard ML conference format, covering methodology, results, and discussion of proposed waveforms as candidates for DNS validation.
- **Public Code Repository:** Fully reproducible, with data generation scripts, training scripts, and a README with MI300X/ROCm setup instructions.

Intentionally not promised: DNS validation of proposed waveforms (future work) and SOTA drag reduction claims. The goal is a clean, reproducible proof-of-concept demonstrating the pipeline's ability to propose physically grounded novel candidates.

7. RISK & MITIGATION

Risk	Likelihood	Mitigation
Labeled dataset too small (~200 samples) for GPR to generalize reliably	Medium	Add physics-informed features (acceleration parameter A+/T+, Stokes penetration depth) as additional input dimensions to the predictor. These are known to correlate with DR% from theory, giving the GPR structure to lean on beyond raw waveform shape.
VAE latent space loses smoothness with small dataset - novel coordinates decode to noise	Medium	Increase synthetic waveform augmentation to 50,000 samples. Apply beta-VAE regularization (increase KL divergence weight) to enforce smoother latent structure at the cost of slightly higher reconstruction error.
All BO-proposed waveforms cluster near the sinusoidal optimum - no genuine novelty	Low-Medium	Add diversity constraint to BO acquisition function penalizing proposals too close to known waveforms in latent space. Reframe paper as methodology demonstration rather than novel waveform discovery if necessary.
Cimarelli data primarily covers in-plane (spanwise) oscillations, not out-of-plane	High (known limitation)	Explicitly acknowledge in paper that training data is from spanwise oscillations and frame as a cross-domain transfer study. This is a limitation but also part of the paper's novelty argument.
Compute overrun or environment setup issues on AMD Instinct MI300X	Low	Entire pipeline (excluding DNS) runs comfortably on CPU in 1 day. MI300X is used for speed, not necessity. ROCm compatibility verified via standard PyTorch ROCm wheel with no CUDA-specific dependencies.