VAESA: Learning A Continuous and Reconstructible Latent Space for Hardware Accelerator Design

Qijing Jenny Huang*, Charles Hong,
John Wawrzynek, Mahesh Subedar[†], Yakun Sophia Shao

jennyhuang@nvidia.com
*NVIDIA, UC Berkeley, [†]Intel

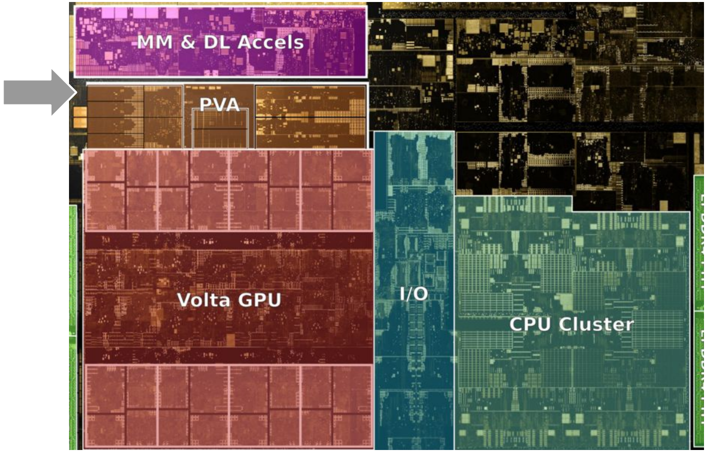
Hardware acceleration is everywhere

Hardware acceleration is the driving force for many innovations.

Robots

Drones

Autonomous Vehicles



NVIDIA Drive Xavier SoC

Augmented Reality

Mobile phones

Genomics

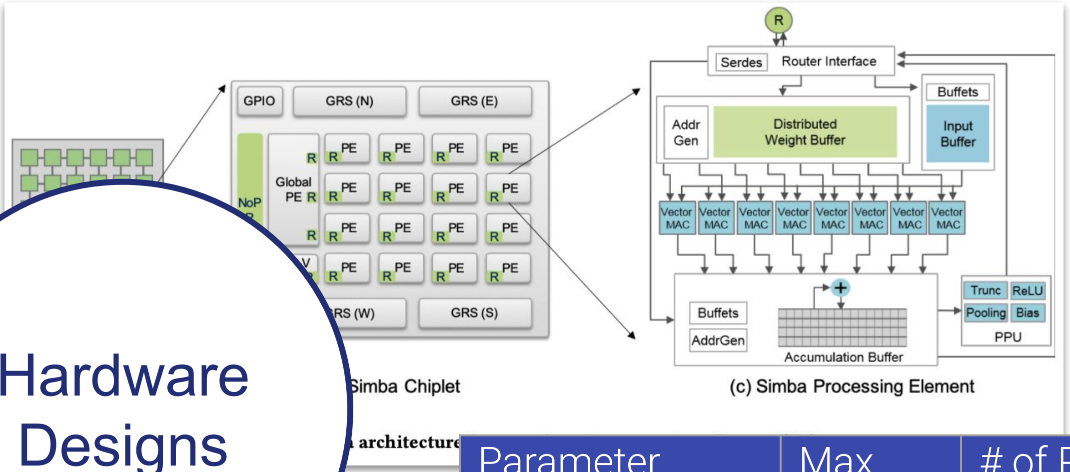
*NVIDIA Press

Designing accelerators is challenging

Hardware design space exploration (DSE) challenges:

1. High-dimensional and discrete
2. Multi-objective and non-convex
3. Costly

Challenge #1: High-dimensional and discrete



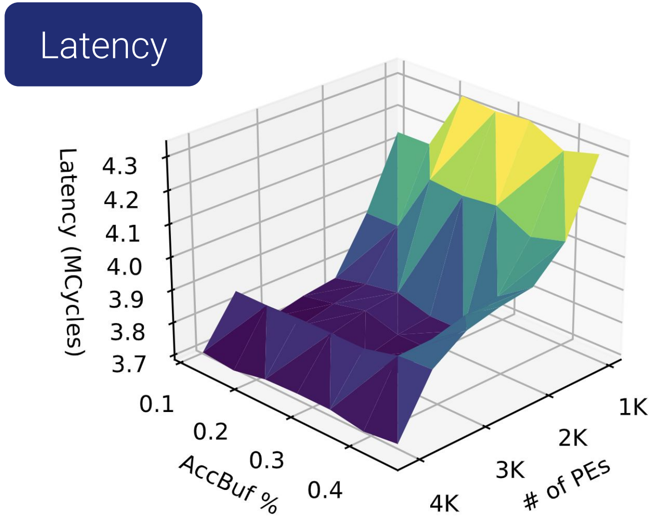
Hardware Designs

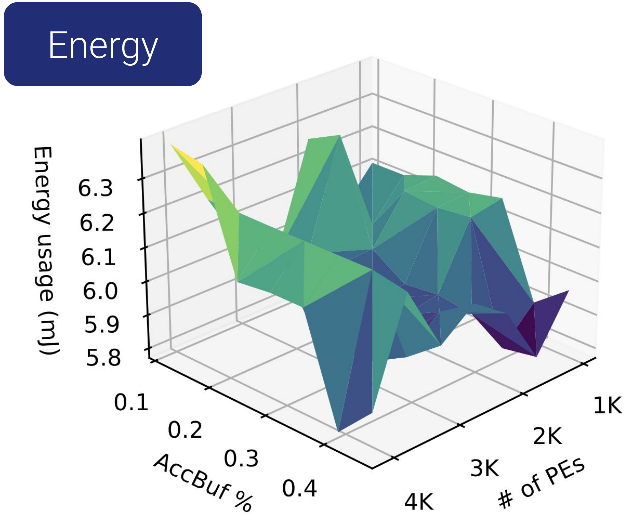
~10^{17}

Parameter	Max	# of Possible Values
# of PEs	64	5
# of MAC units	4096	64
Accum. buffer size	96 KB	128
Weight buffer size	8 MB	32768
Input buffer size	256 KB	2048
Global buffer size	256 KB	131072

From “Simba: Scaling Deep-Learning Inference with Multi-Chip-Module-Based Architecture.” Shao et al., 2019.

Challenge #2: Multi-objective and non-convex





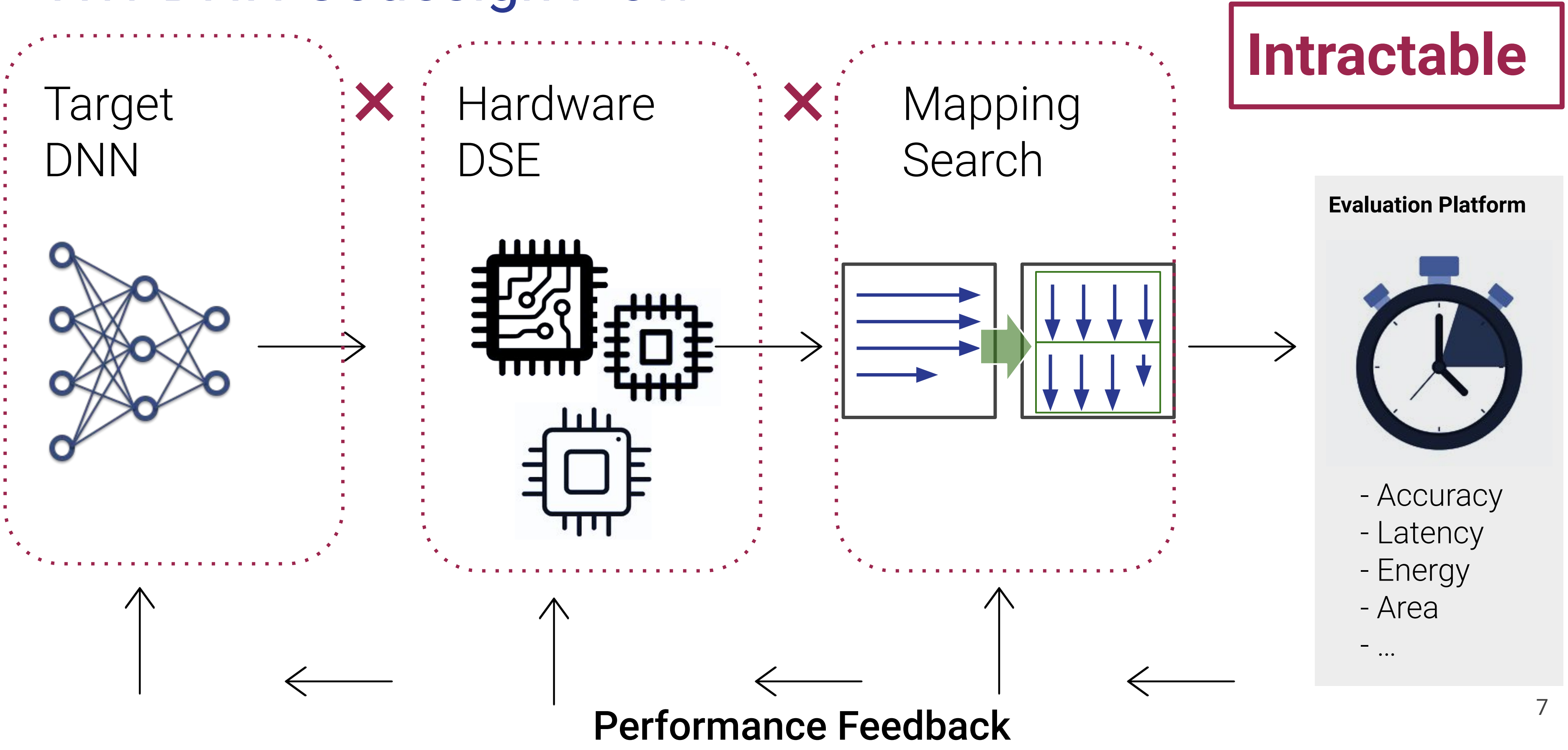
Performance of ResNet-50 as # of PEs and accumulation buffer size change

Challenge #3: Costly

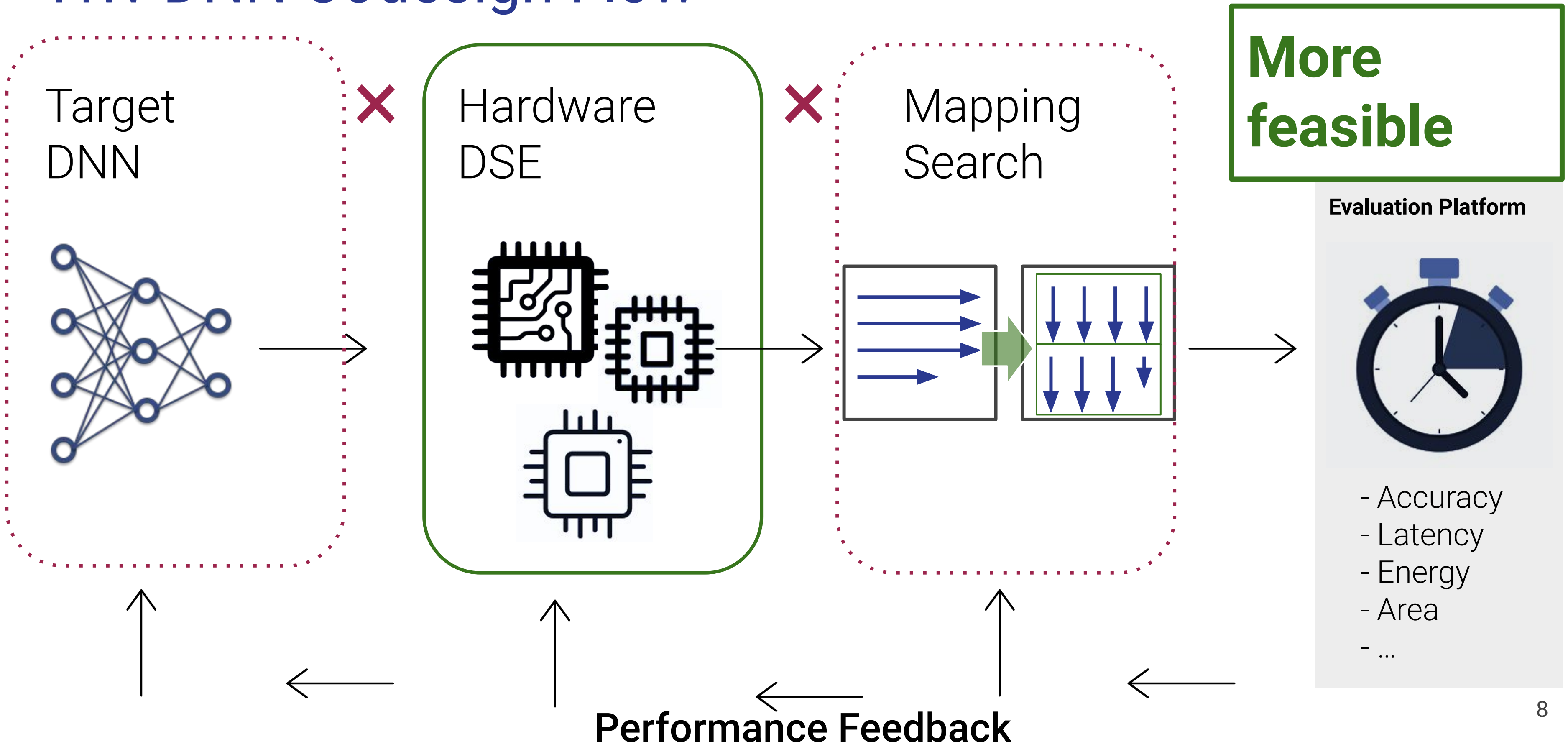
Evaluation Time × Hardware Designs $\sim 10^{17}$ = **>> 32M years**

Platform	Evaluation Time
Timeloop	0.01s
VCS	10 mins
FPGA	2 mins

HW-DNN Codesign Flow



HW-DNN Codesign Flow



Problem Statement

How can we efficiently navigate the accelerator design space for deep learning algorithms?

Prior HW DSE work: Search strategy oriented

	Heuristic-Driven	Black-box Optimization	Gradient-based Optimization
Original Space	Interstellar	Bayesian Opt Apollo NAAS AutoSA	EDD DiffTune Prime

Prior HW DSE work: Search strategy oriented

	Heuristic-Driven	Black-box Optimization	Gradient-based Optimization
Original Space	Interstellar	Bayesian Opt Apollo NAAS AutoSA	EDD DiffTune Prime

Focus on developing **search strategies** to explore the original design space

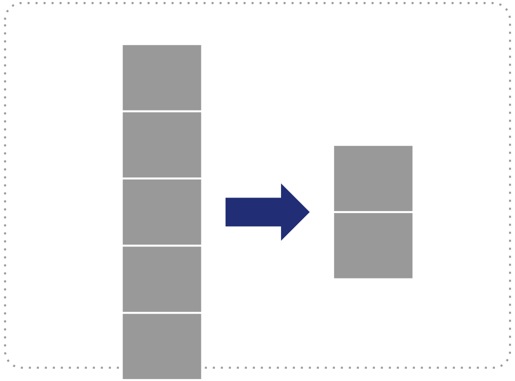
Prior HW DSE work: Search strategy oriented

	Heuristic-Driven	Black-box Optimization	Gradient-based Optimization
Original Space	Interstellar	Bayesian Opt Apollo NAAS AutoSA	EDD DiffTune Prime
New Design Space			

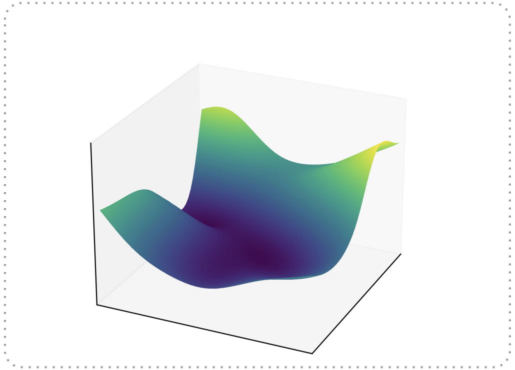
Can we transform the design space instead?

Desirable hardware design space properties

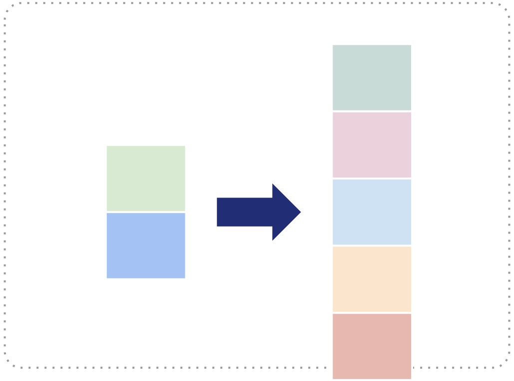
1. Reduced dimensionality



2. Smooth surface



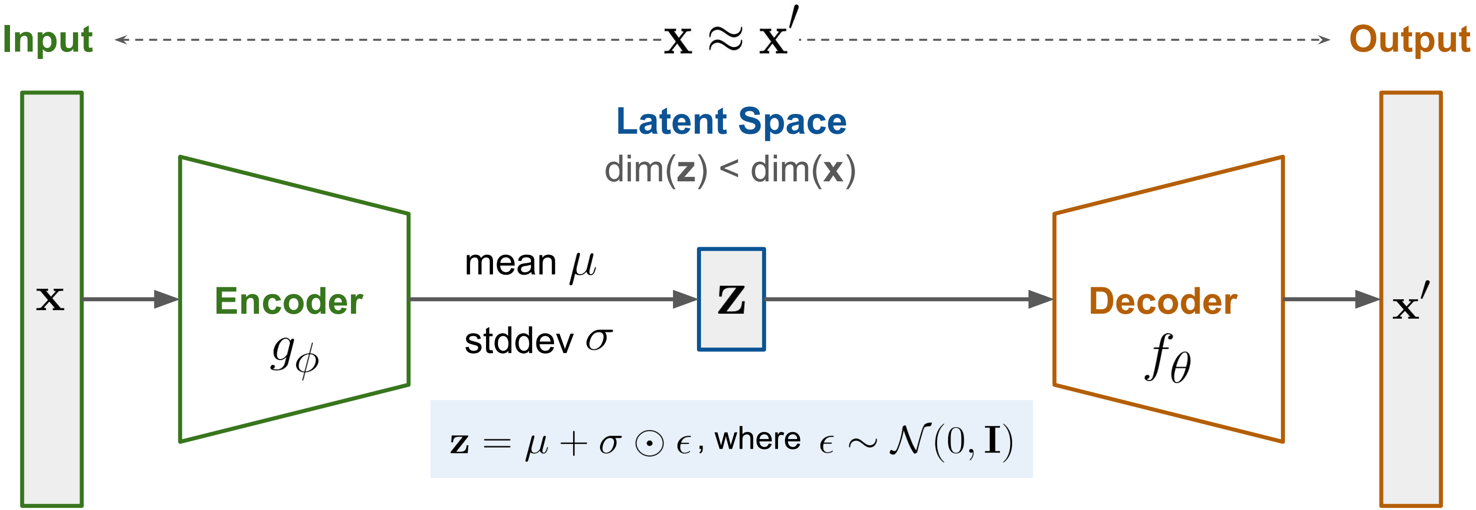
3. Reconstructible



Variational Autoencoder (VAE)

Background: Variational Autoencoder (VAE)

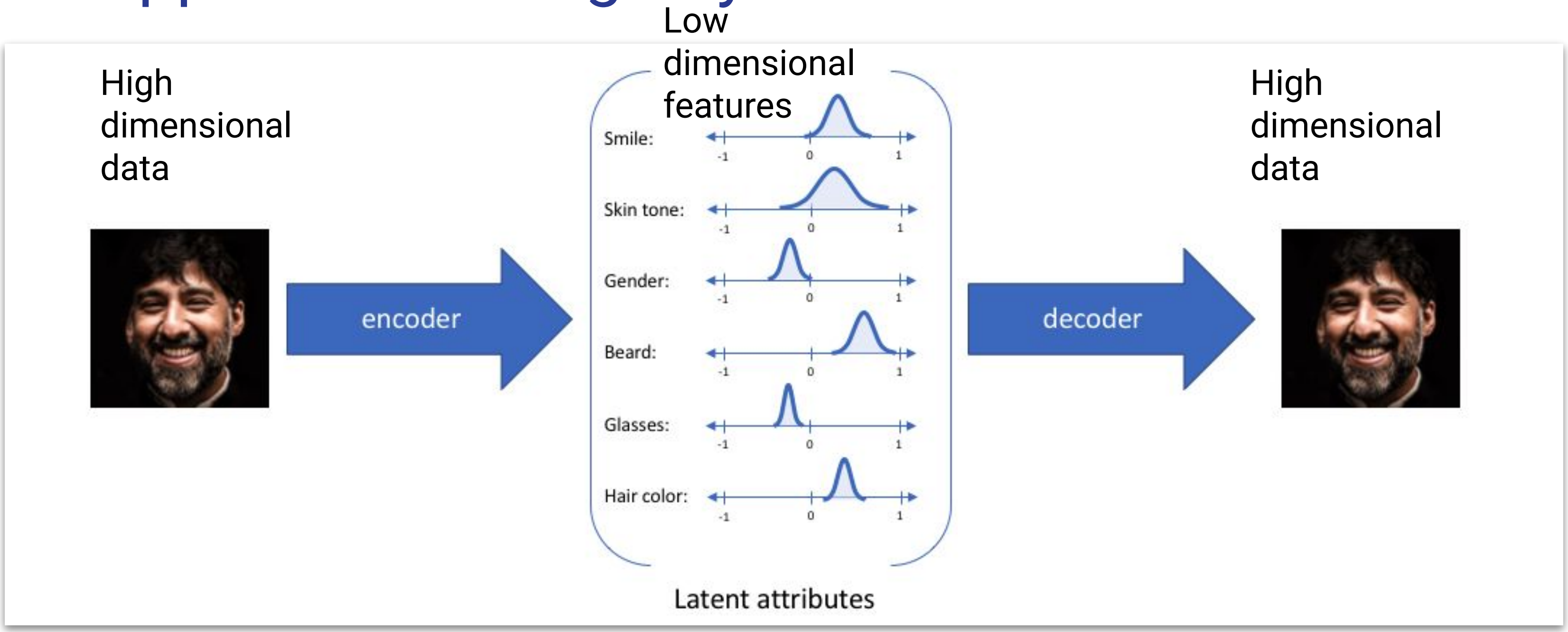
A **model** that learns **a compressed representation** $\mathbf{z}$ of input data $\mathbf{x}$



- The feed-forward model predicts x’ from x through a bottleneck layer
- Training minimizes mean-squared error between x and x’

“Auto-Encoding Variational Bayes”. Diederik P Kingma, Max Welling, ICLR 2014.

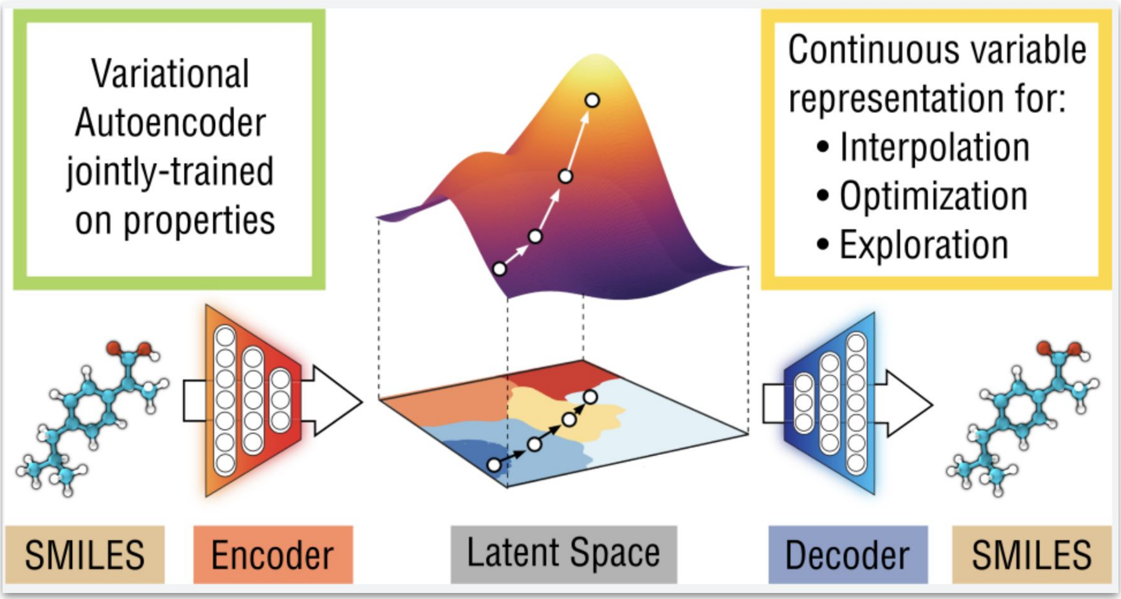
VAE Application: Image Synthesis



- VAE learns latent features by identifying structure in data
- Varying the latent space features generates different faces

“Variational Autoencoders.” Jeremy Jordan, 2018.

VAE Application: Chemical Design



- Training a classifier jointly assigns categorical meaning to the latent space
- Molecules with desired properties can be generated by sampling the latent space

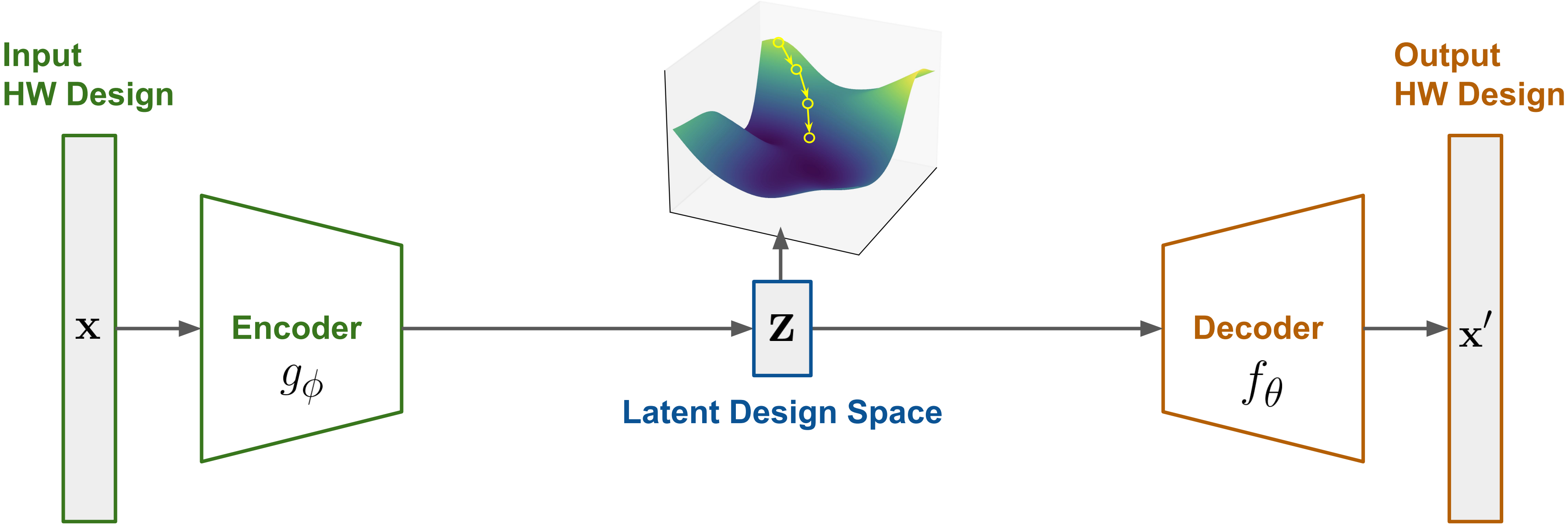
"Automatic chemical design using a data-driven continuous representation of molecules." Gómez-Bombarelli et al. *ACS Central Science*, 2018.

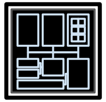
Our work: Search space oriented

	Heuristic-Driven	Black-box Optimization	Gradient-based Optimization
Original Space	Interstellar	Bayesian Opt Apollo NAAS	EDD DiffTune Prime
Latent Space	**VAE** for **S**patial **A**ccelerator Design (**VAESA**)		

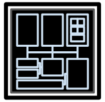
NEW

Our Framework - VAESA

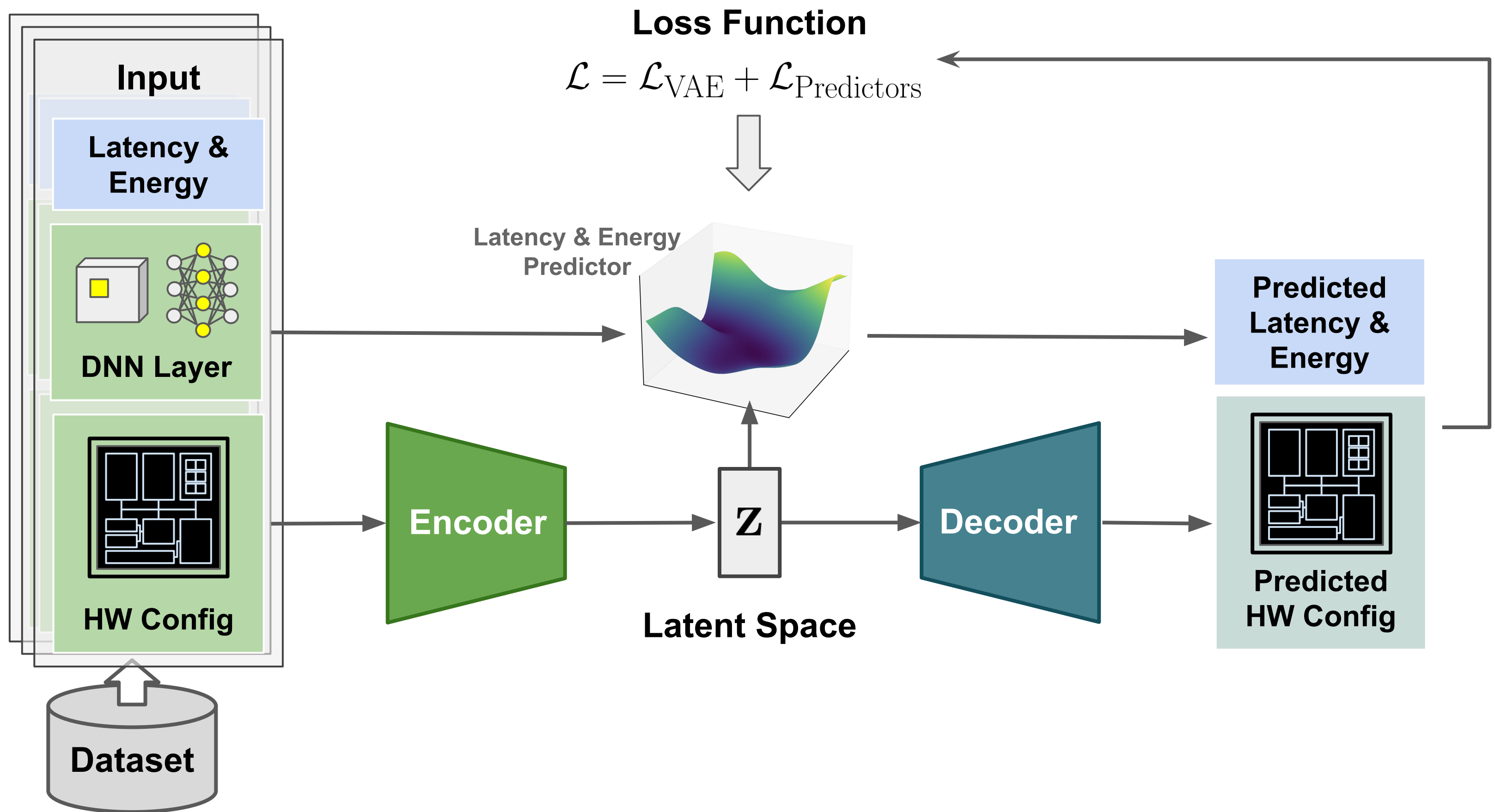




- VAESA provides a transformed search space for HW DSE

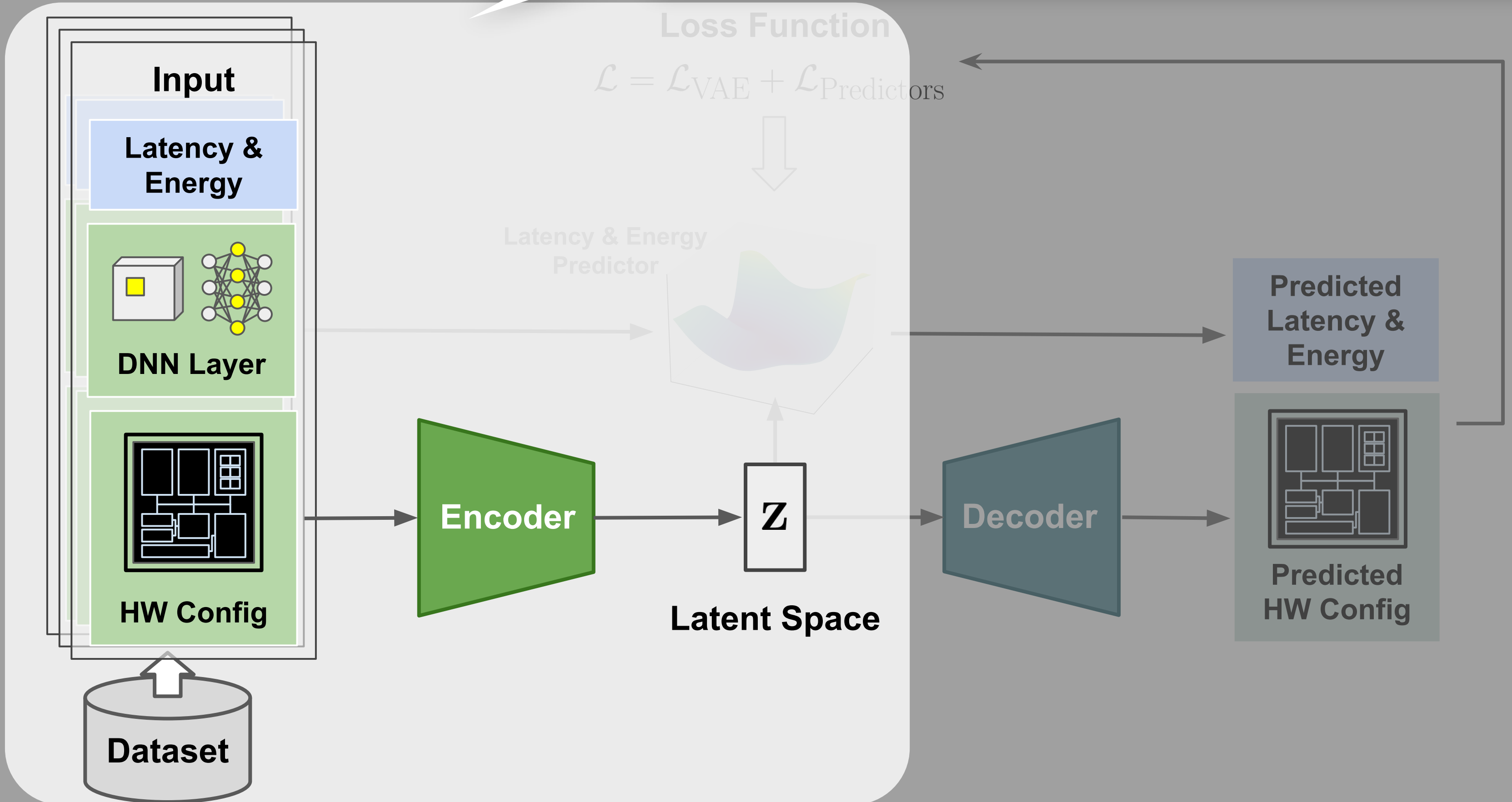


VAESA Training



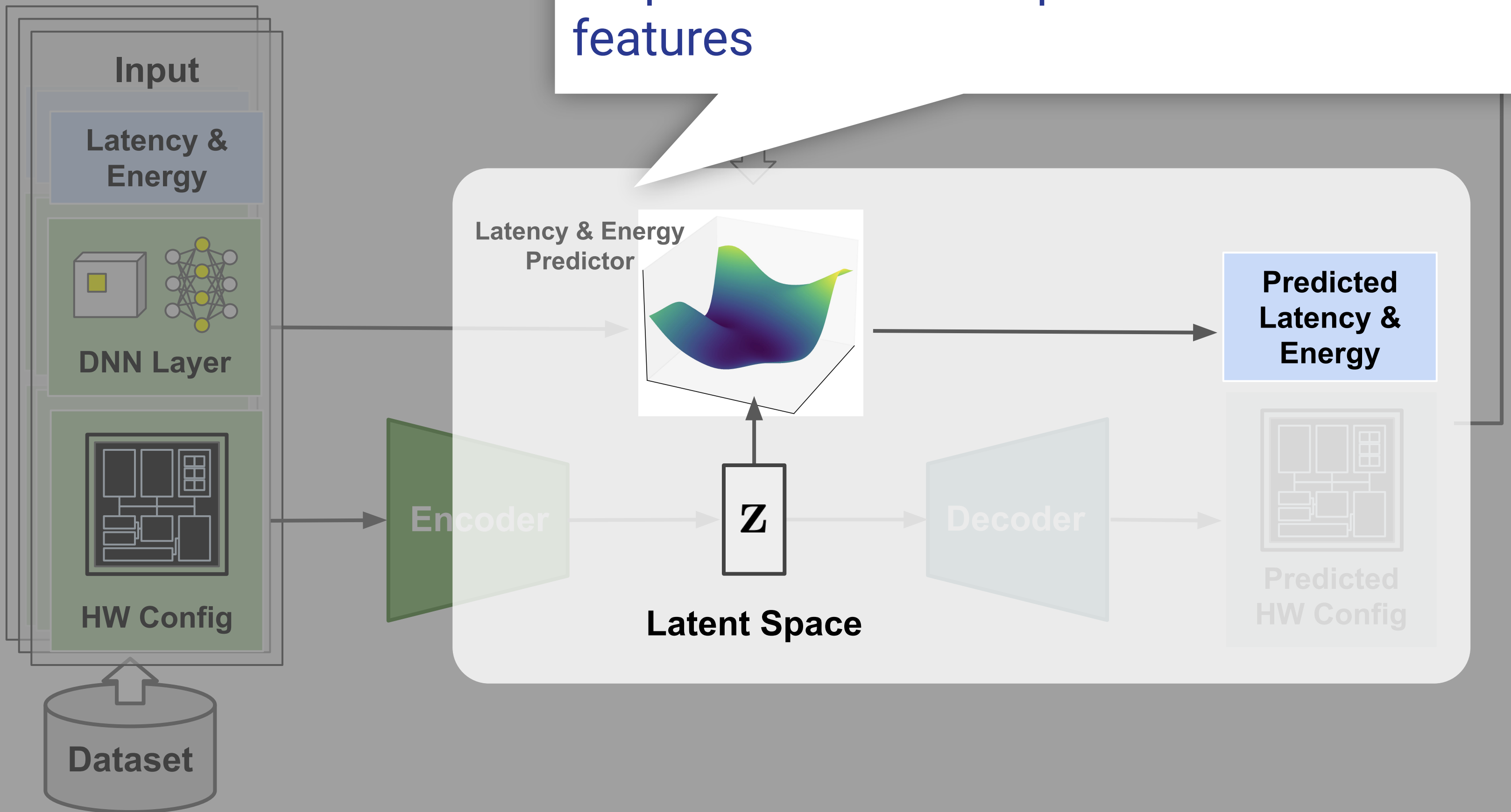
VAESA Training

Step 1: Encode HW designs to a compact, continuous latent space



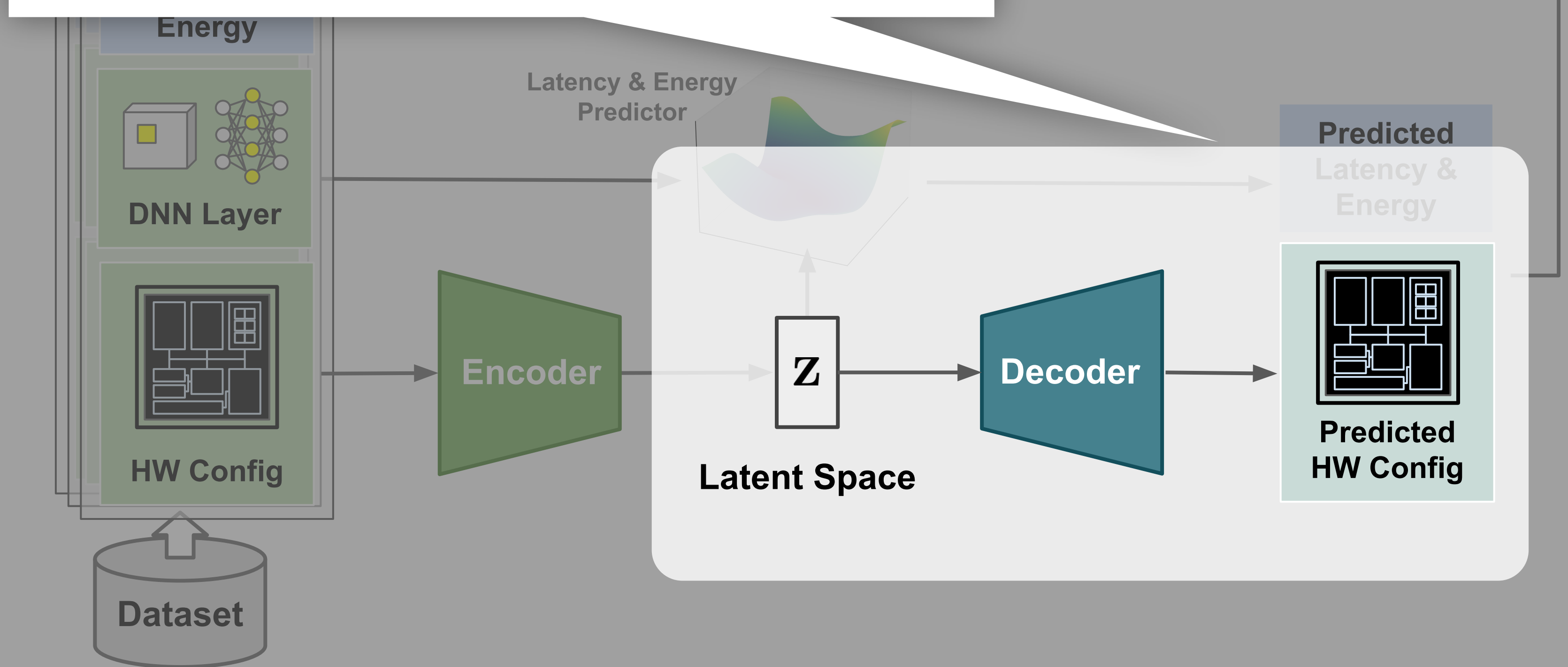
VAESA Training

Step 2: Performance prediction from latent features

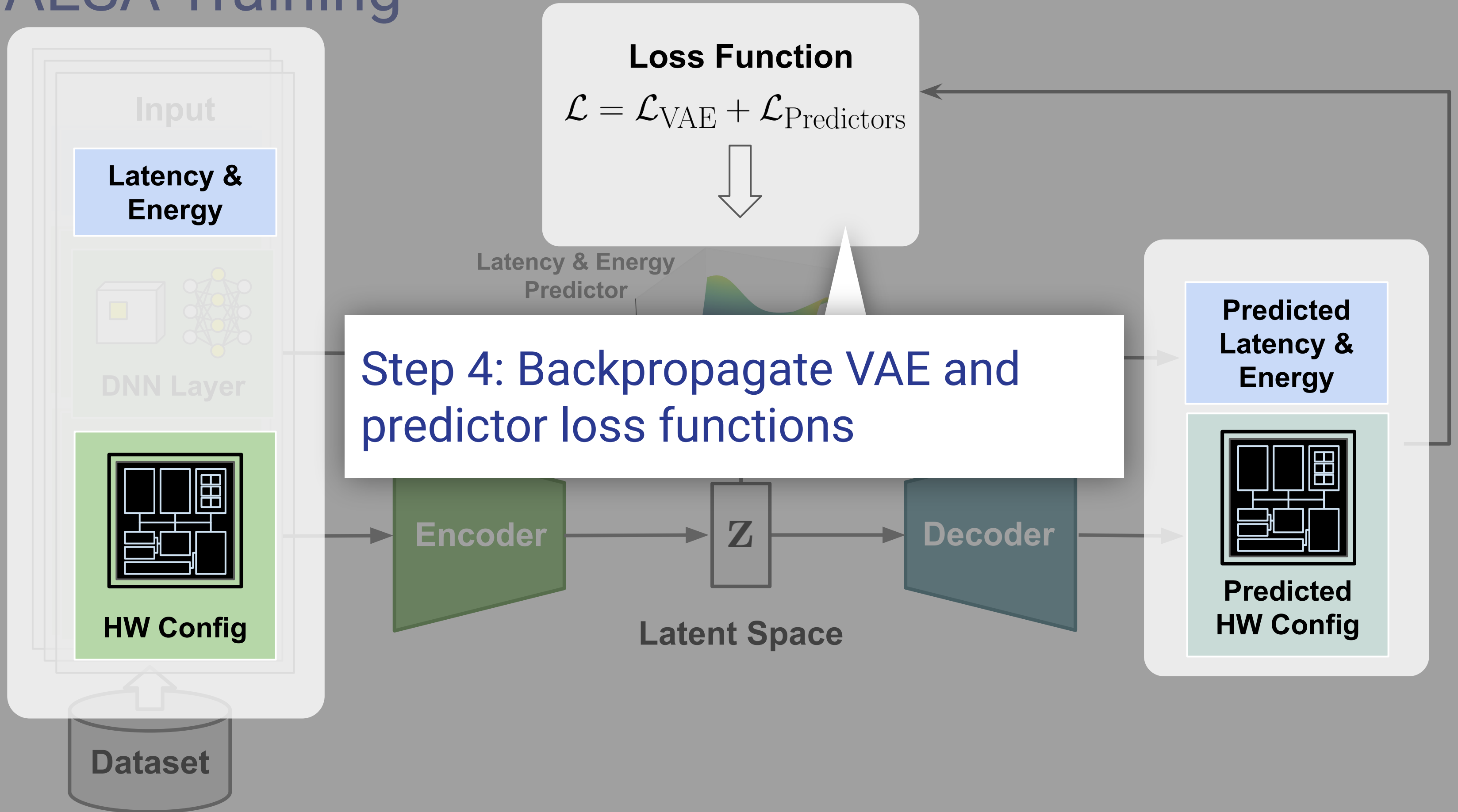


VAESA Training

Step 3: Reconstruct actual hardware configurations



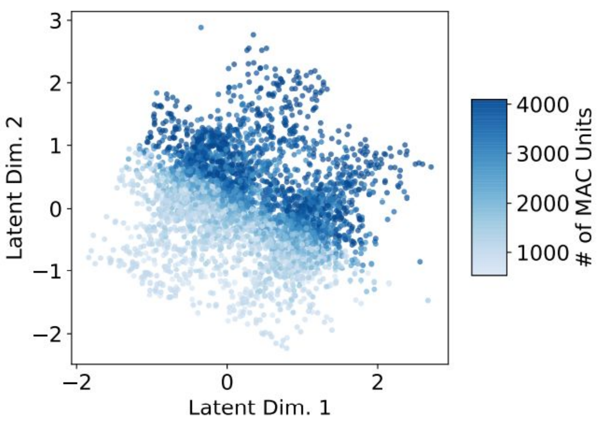
VAESA Training



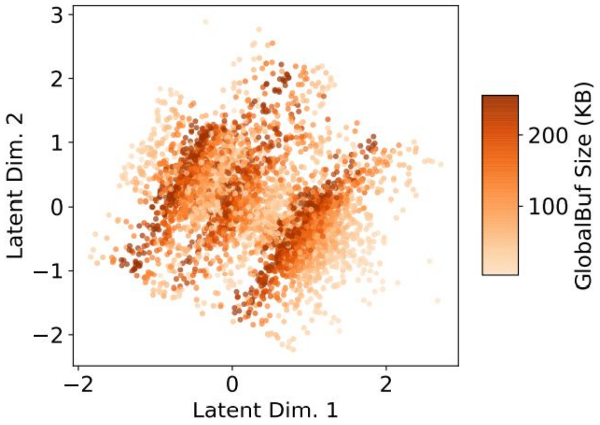
VAESA Visualization (2D)

Learned latent space

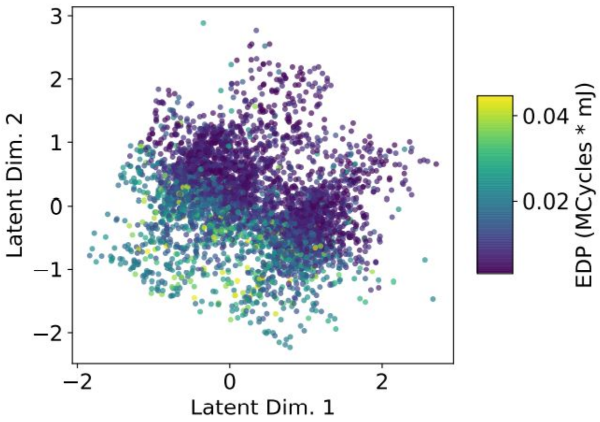
a) Number of MAC units



b) Global buffer size



c) Energy-delay product

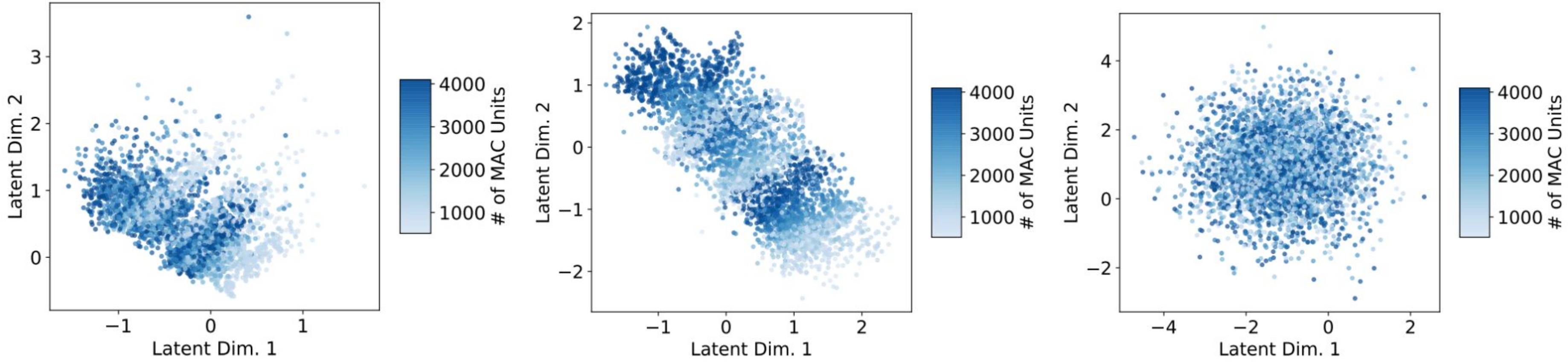


- Good clustering and structures are observed in the latent space designs

VAE Hyperparameter Tuning

Weighting KL divergence

- Coefficient adjusts weight of KLD (closeness of a given point's mean+variance encoding to the standard normal) relative to reconstruction loss

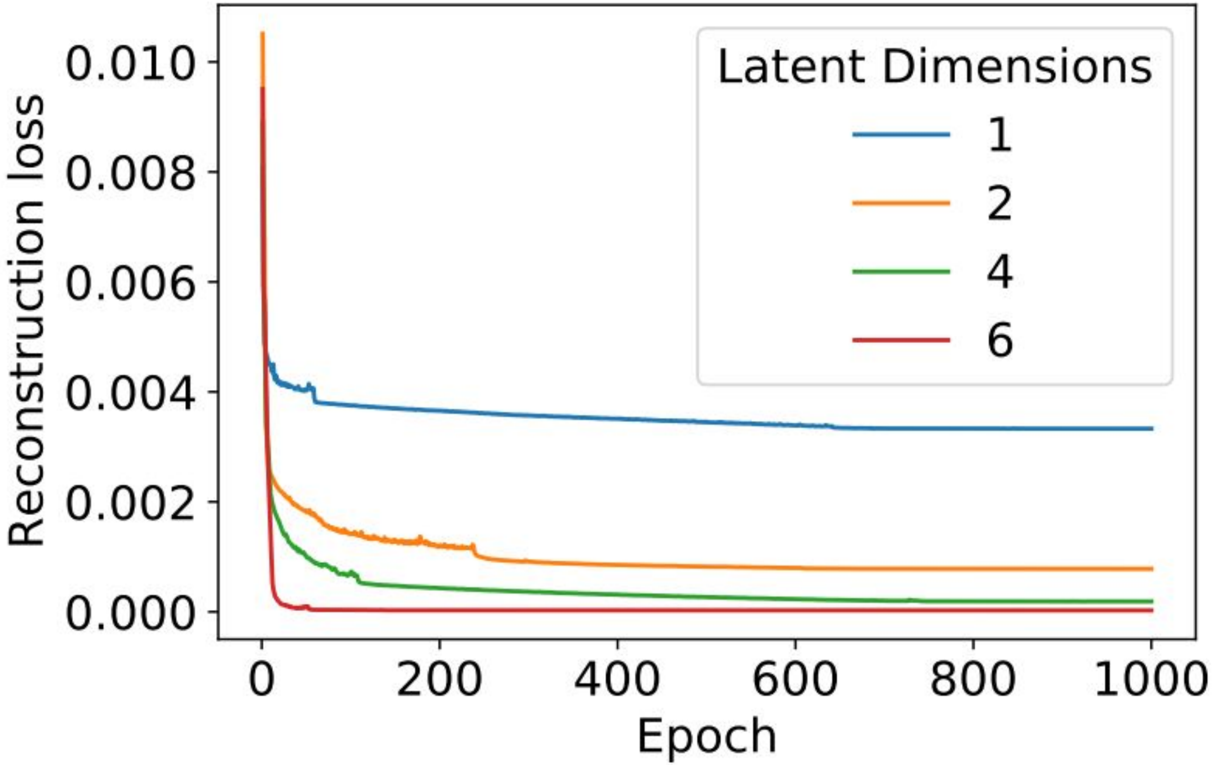


(a) $\alpha = 0$ (b) $\alpha = 0.0001$ (c) $\alpha = 0.01$

$$L_{\text{VAE}} = L_{\text{recon}} + \boxed{\alpha} L_{\text{kld}}$$

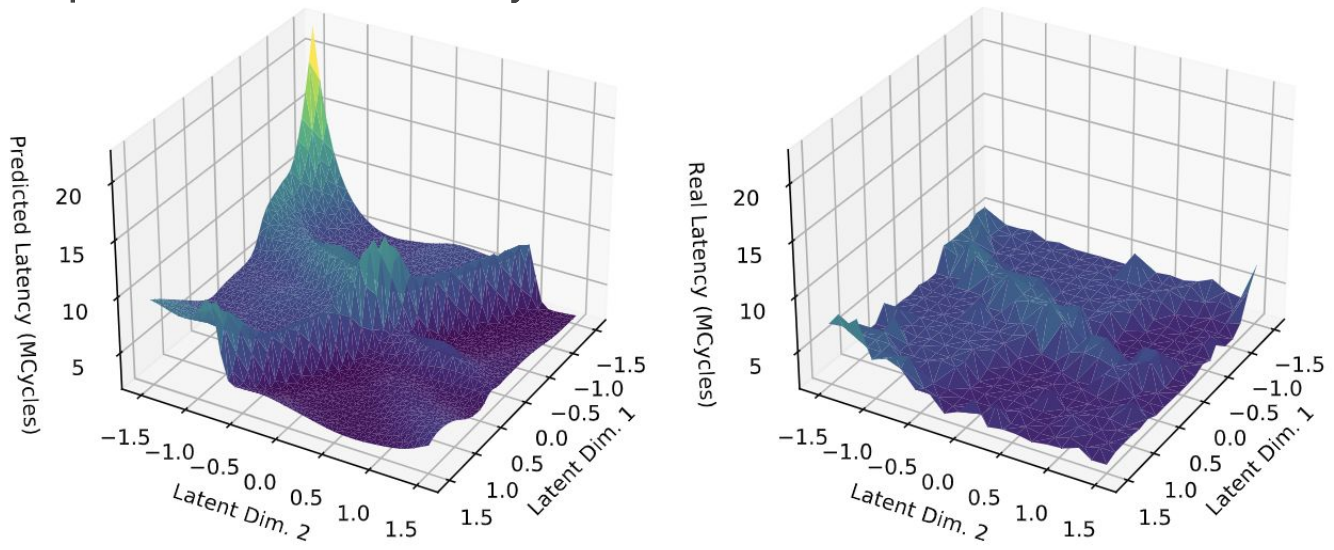
VAE Hyperparameter Tuning

Latent space dimensionality



VAESA Visualization (2D)

Predicted performance: Latency



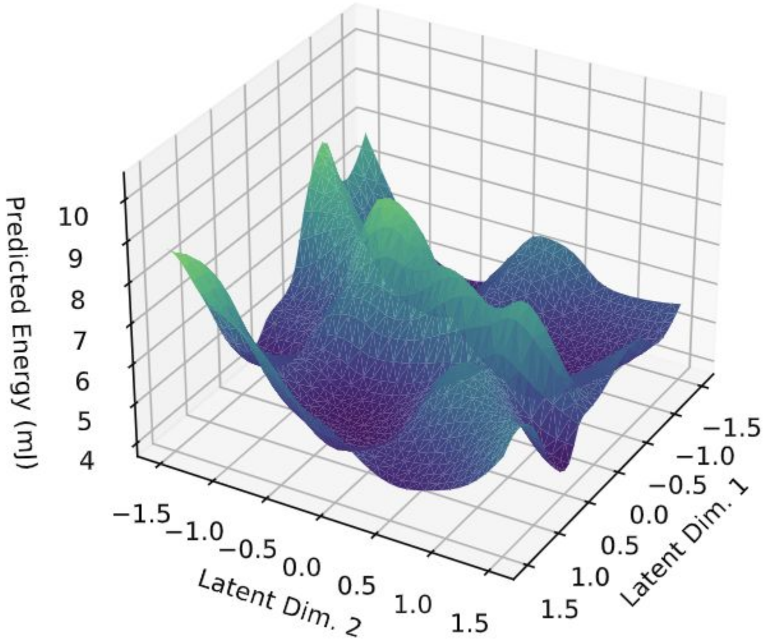
(a) Predicted latency

(b) Real latency of decoded accelerator

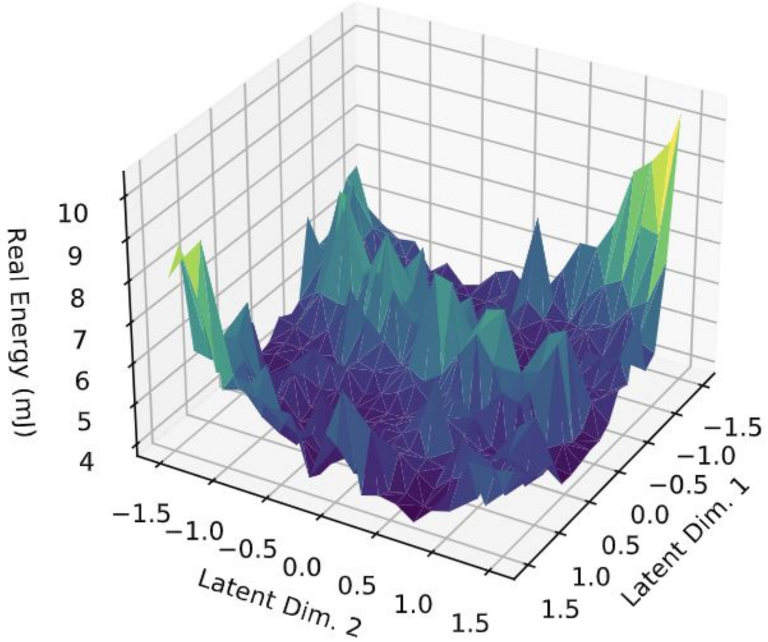
- Good clustering and structures are observed in the latent space designs

VAESA Visualization (2D)

Predicted performance: Energy

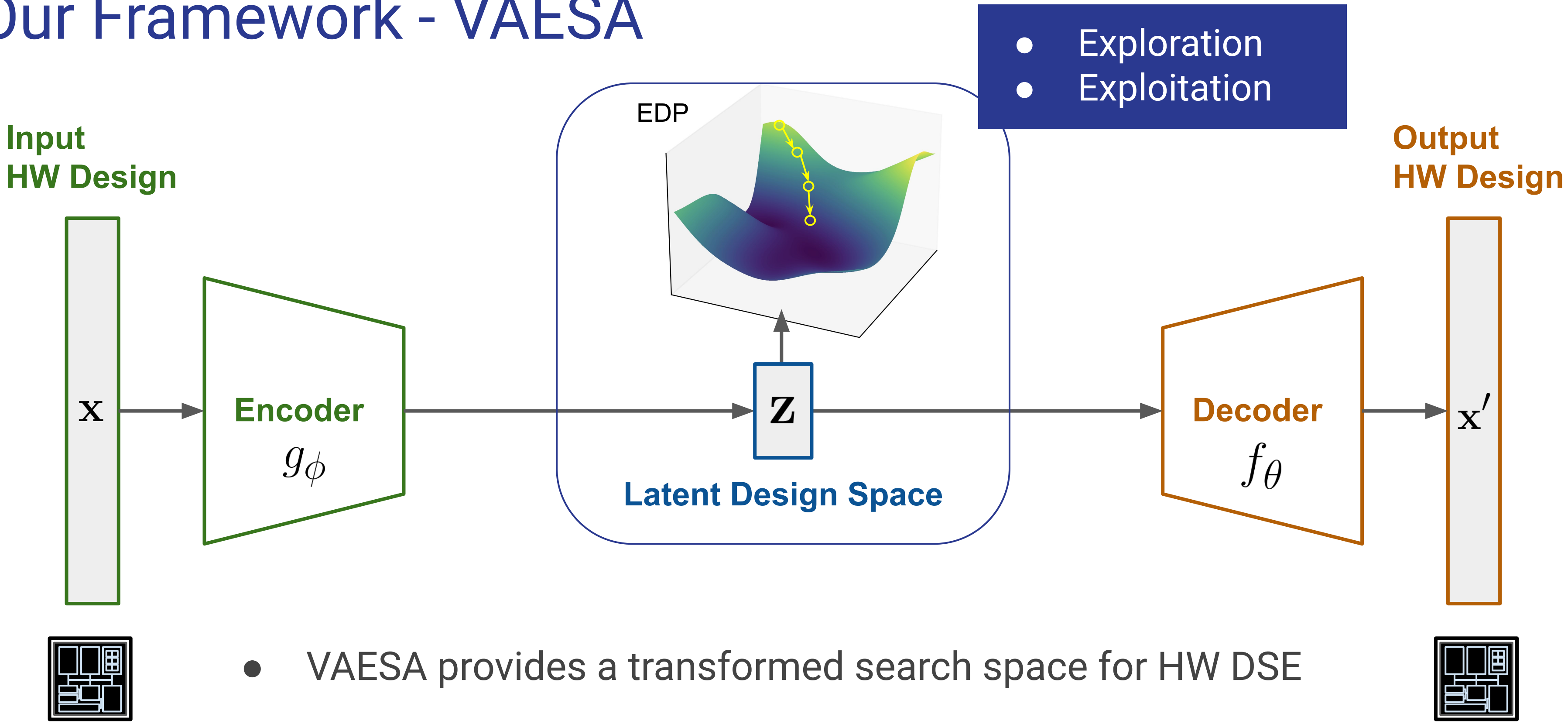


(c) Predicted energy usage



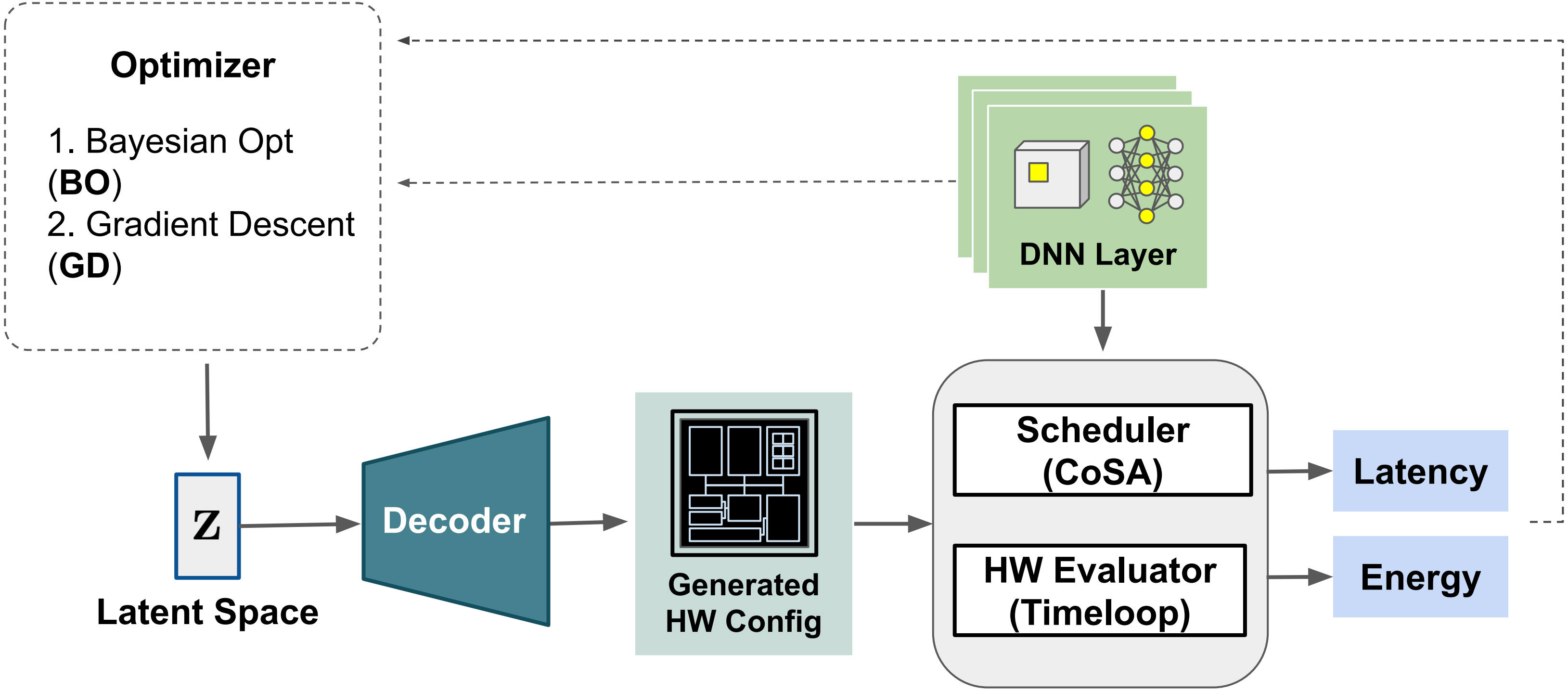
(d) Real energy usage of decoded accelerator

Our Framework - VAESA



- VAESA provides a transformed search space for HW DSE

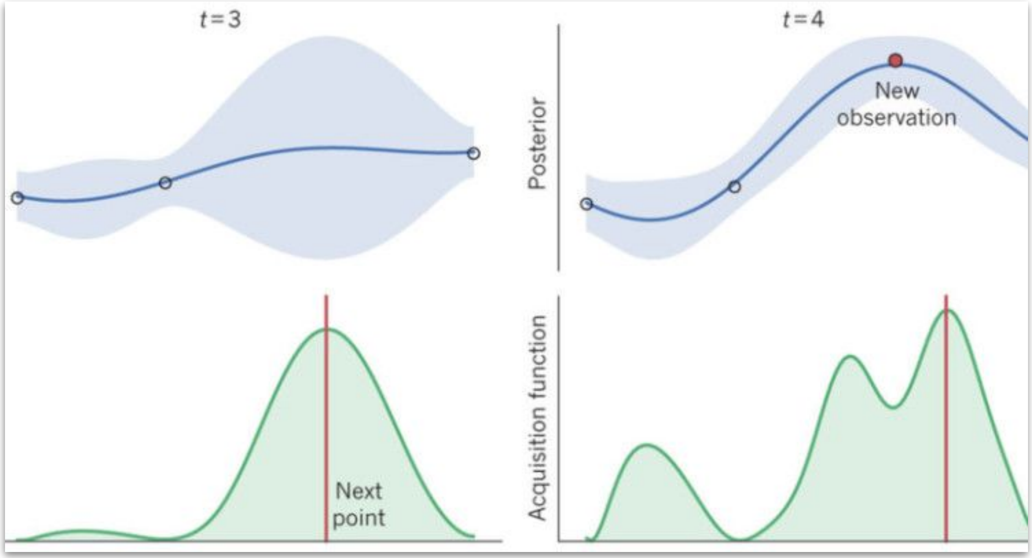
VAESA Inference



VAESA Inference

Bayesian Optimization (BO)

- BO iteratively updates **a statistical model** to approximate the unknown objective function and uses **an acquisition function** to decide which input to sample next.

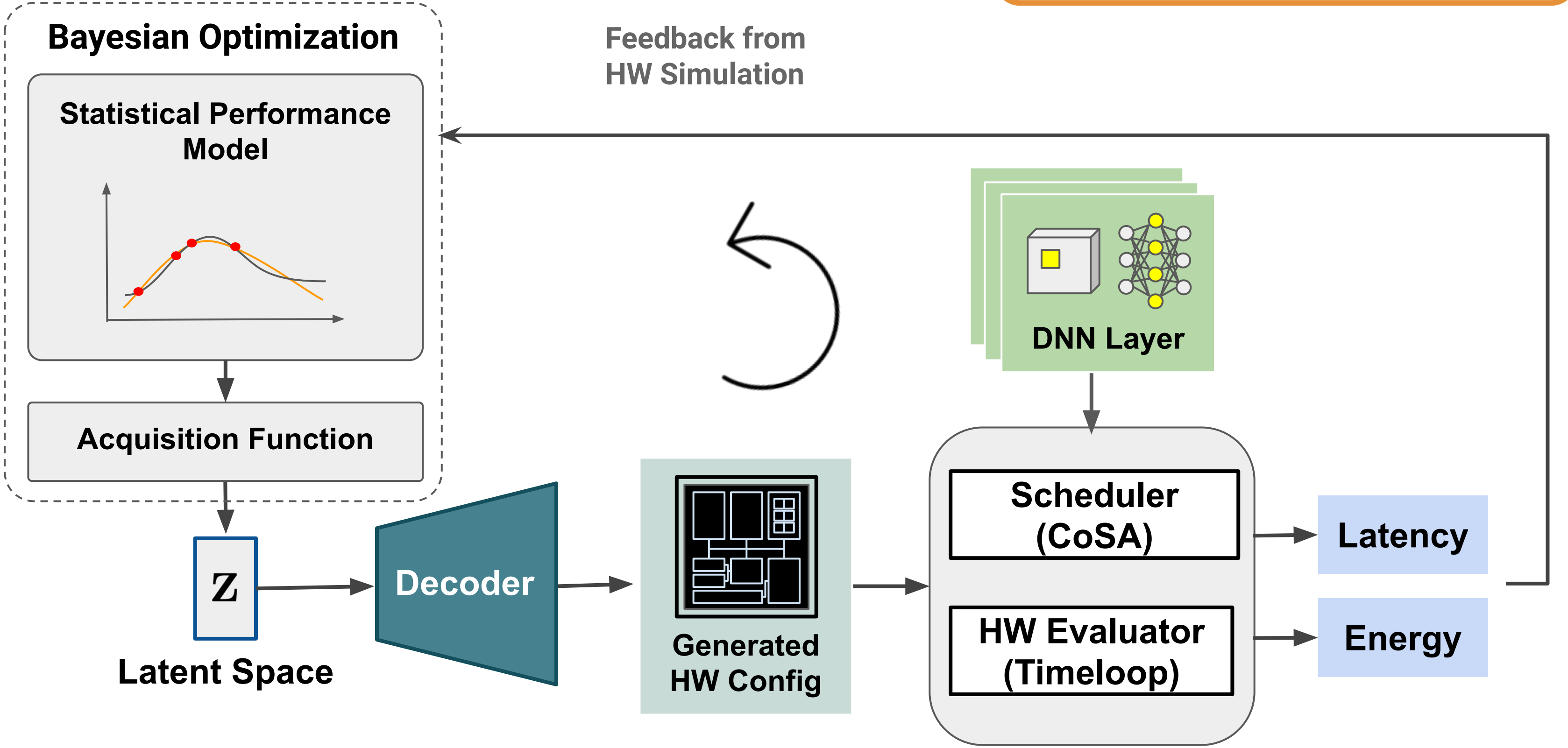


* Ghahramani, Zoubin. "Probabilistic machine learning and artificial intelligence." Nature 521.7553 (2015): 452-459.

VAESA Inference

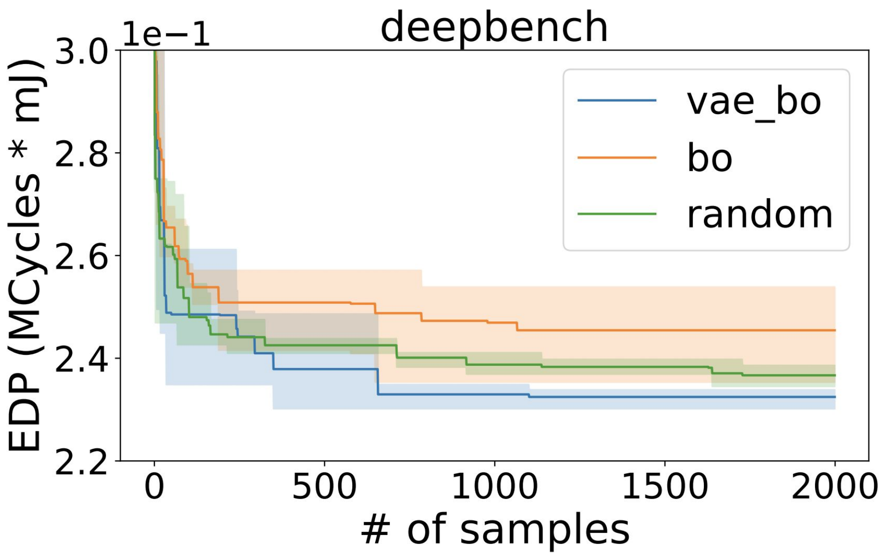
VAESA+BO

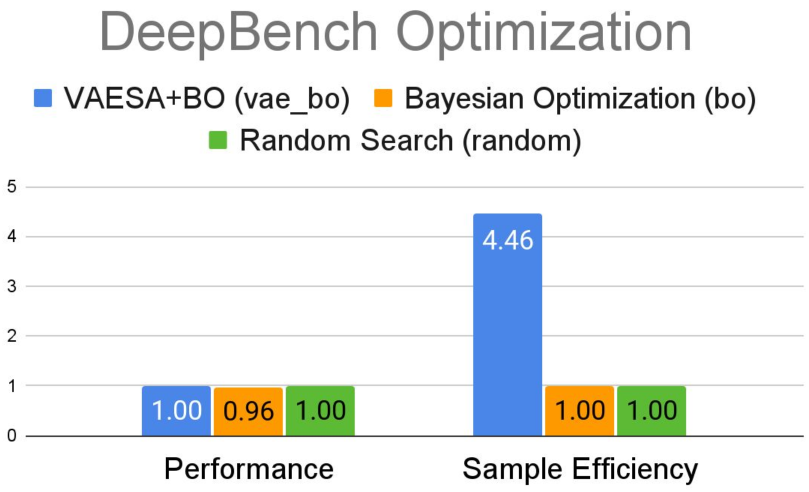
Black-box optimization on the latent space



Results

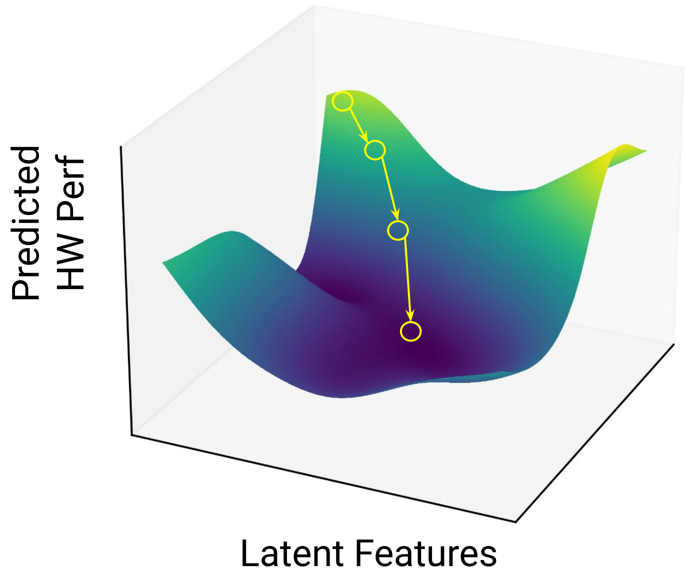
VAESA+BO Comparison





VAESA+BO improves the sample efficiency of BO and finds the best accelerator design

Gradient Descent (GD) for VAESA Inference

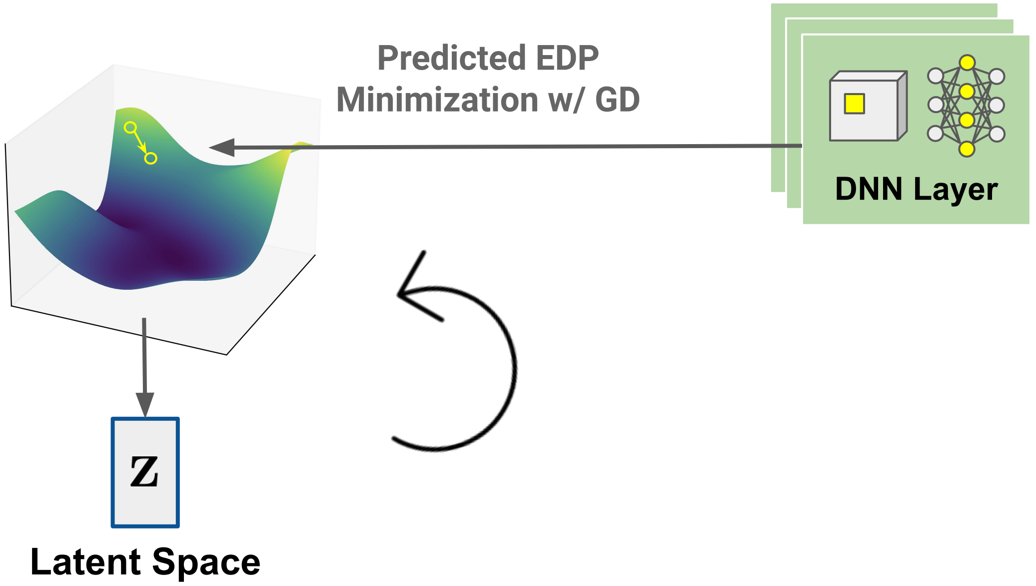


- GD is an iterative method for optimizing an objective function with suitable smoothness properties by take repeated steps **in the opposite direction of the gradient** of the function at the current point.

VAESA Inference

VAESA+GD

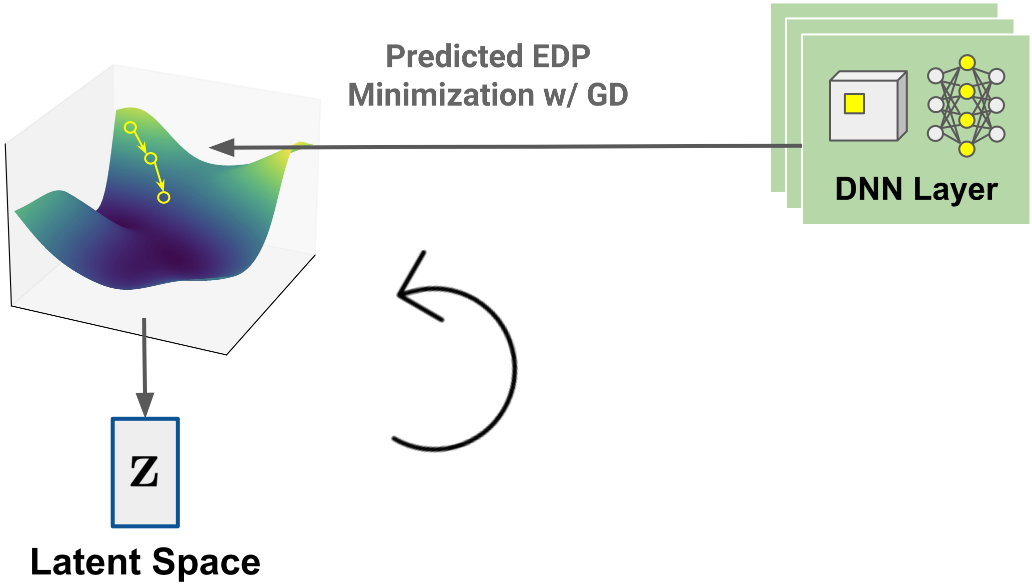
Predictor-based search on the latent space



VAESA Inference

VAESA+GD

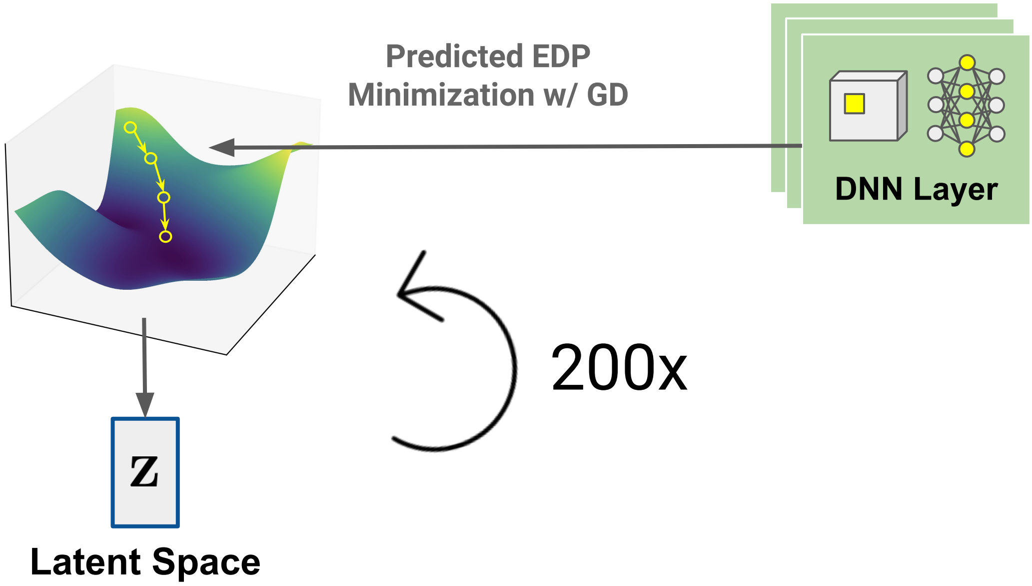
Predictor-based search on the latent space



VAESA Inference

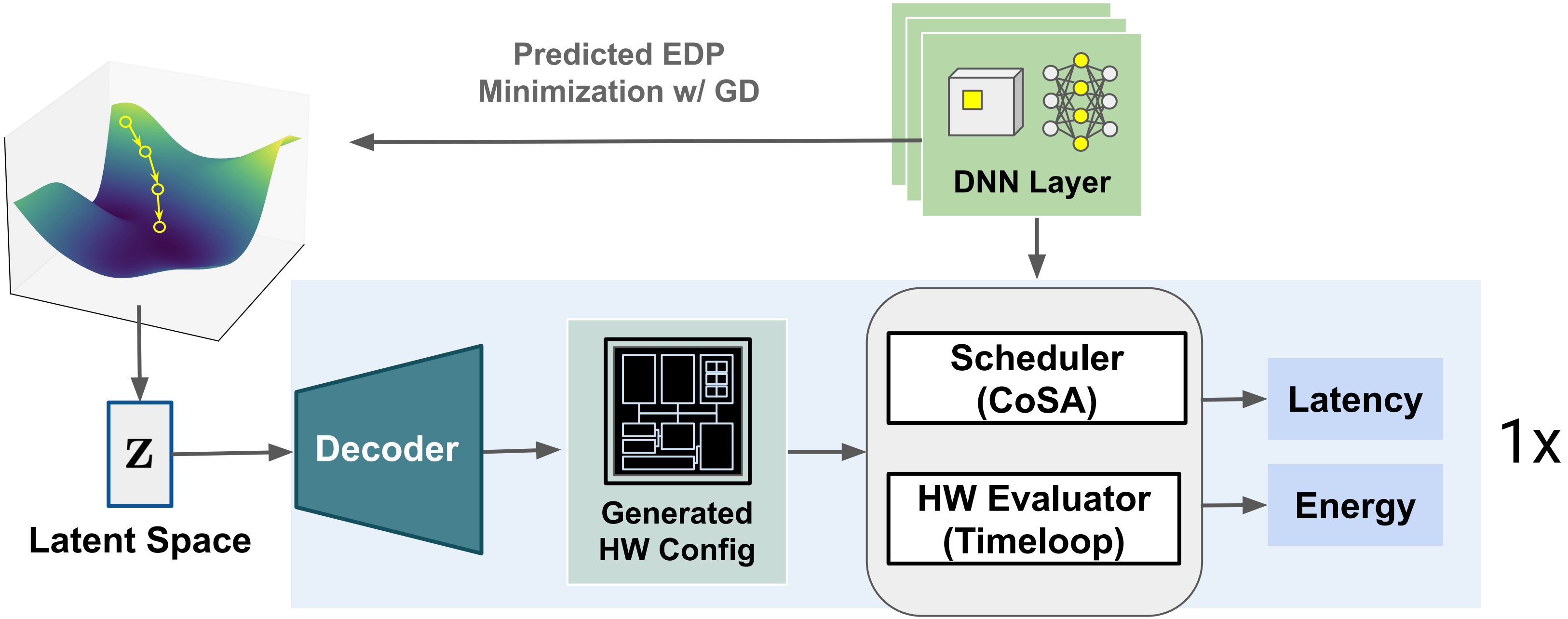
VAESA+GD

Predictor-based search on the latent space



VAESA Inference
VAESA+GD

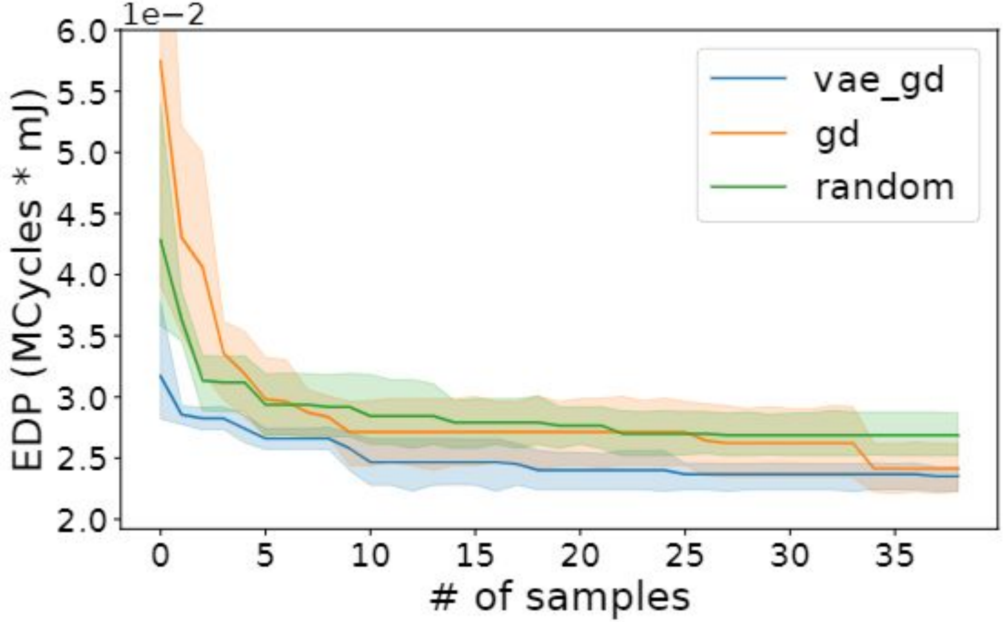
Predictor-based search on the latent space



Results

VAESA+GD Comparison

Average EDP over 12 test DNN layers



GD on the latent space achieves better design points faster than GD on the original space.

Conclusion

In VAESA,

- We introduce an DSE framework where the search is performed on a **continuous** and **reconstructible** latent space
- We show that using learned latent design space enhances two state-of-the-art search algorithms: *BO* and *GD*

Email: jennyhuang@nvidia.com, charleshong@berkeley.edu

Git: https://github.com/ucb-bar/vaesa.git