SAND2021-16130C

Sandia National Laboratories

Exceptional service in the national interest

Efficient Sampling Methods for Machine Learning Error Models with Application to Surrogates of Steady Hypersonic Flows

Elizabeth Krath, David Ching, Patrick Blonigan

January 5, 2022

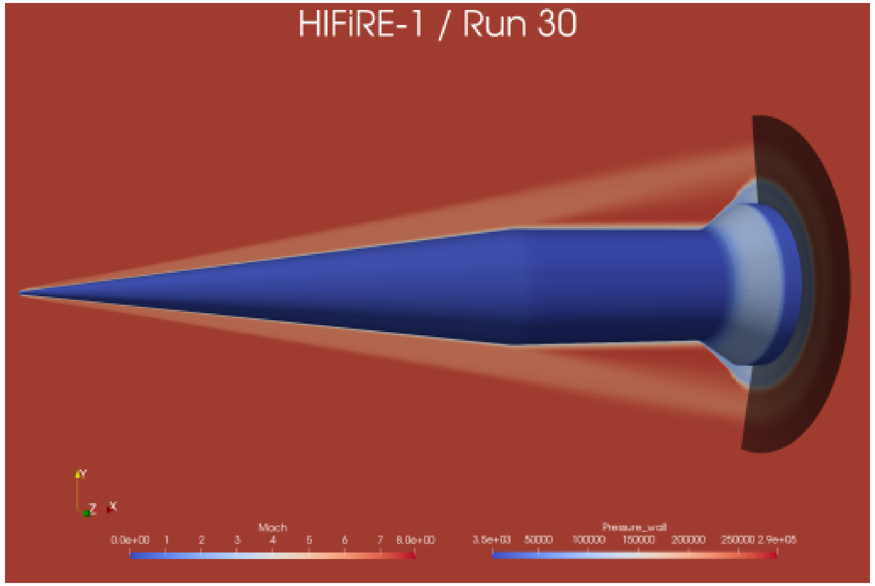
AIAA SciTech Forum, San Diego, CA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Motivation

- Full-order model (FOM) simulations of hypersonic vehicles is **expensive**
- **Mitigate expense** by using reduced-order models (ROM)
- Using a ROM can lead to **reduction in accuracy**
- Would like to quantify reduction in accuracy
  - Don’t want to run blind
- Utilize error models to **quantify and predict error** between ROM and FOM
  - Able to make better ROM predictions
  - Possibly able to update ROM basis with “knowledgeable” training points



How do we create the error models without adding extra cost?

Goal: Develop Error Model Efficiently

QOI Error

$$\delta_s(\mu) := s(\boldsymbol{\mu}) - \tilde{s}(\boldsymbol{\mu})$$

Develop Error Model Efficiently by focusing on Sampling Methods

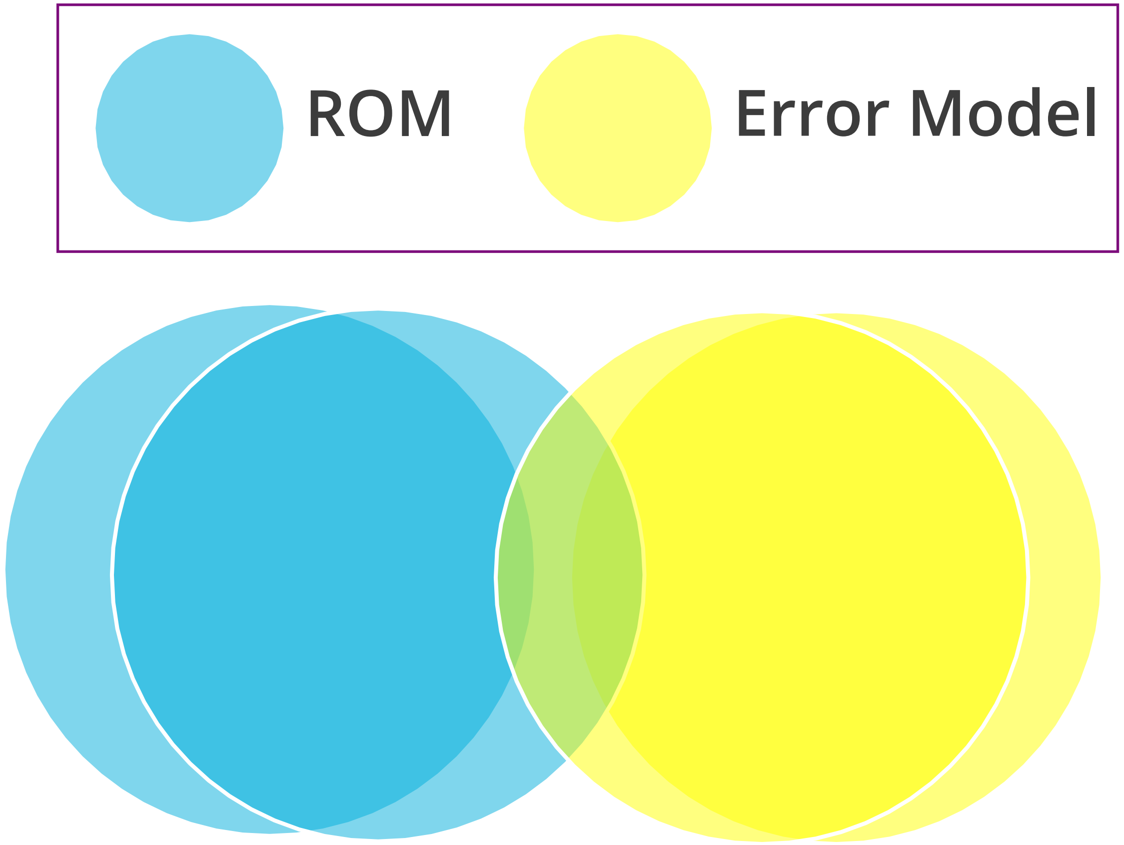
- Problem: Developing the error model adds to the computational expense
  - FOM is computationally expensive, but…
  - Necessary to use FOM for training points for ROM and error model
- Why not train ROM with more FOM points than develop error model?
  - Unpredictability
  - Nonlinear domain
  - Flying blind
- Solution:
  1. Sampling types
  2. Sampling strategies

Sampling Types

1. Latin Hypercube Sampling (LHS)
2. LHS with maximin criterion
    - Adds constraint on distance between sampling points
3. D-Optimal design
    - Maximizes determinant of information matrix
        - Reduces variance in results
    - Contains replicates not useful for computational experiments
        - Replace replicates with random LHS points
        - End result may not be a true D-Optimal design

Sampling Strategies

- Distinct training set
- Augmented training set
- Single training set



Summary of Sampling Methods

Sampling Types

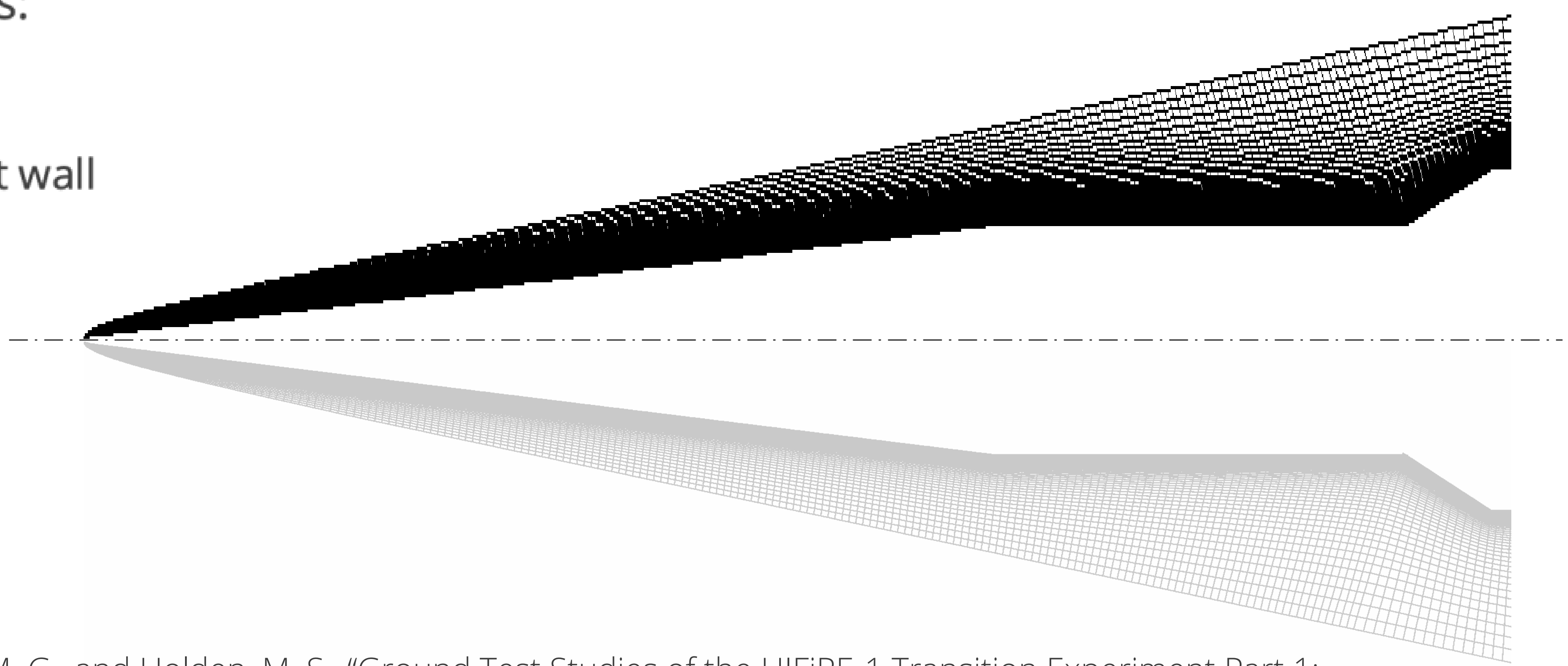
1. LHS
2. LHS with Maximin Constraint
3. D-Optimal

Sampling Strategies

1. Distinct Training Sets
2. Single Training Set
3. Augmented Training Sets

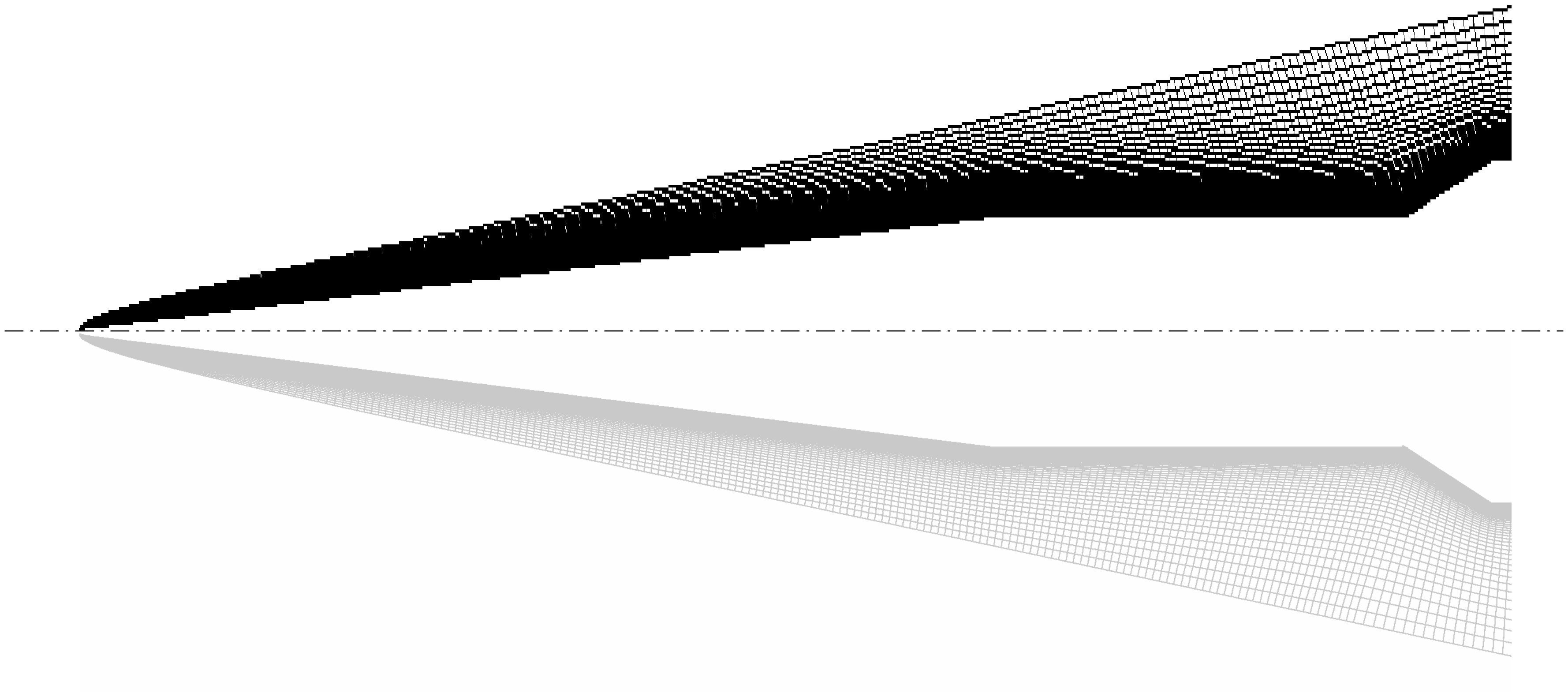
HIFiRE-1

- Run 30 of CALSPAN University of Buffalo HIFiRE-1 wind tunnel tests [3]
- $N = 32{,}768$ cells
- Steady-state
- Boundary conditions:
  - Supersonic inlet
  - Supersonic outlet
  - No-slip enforced at wall
  - Fixed temperature



[3] Wadhams, T. P., Mundy, E., MacLean, M. G., and Holden, M. S., "Ground Test Studies of the HIFiRE-1 Transition Experiment Part 1: Experimental Results," *Journal of Spacecraft and Rockets*, Vol. 45, No. 6, 2008, pp. 1134–1148. https://doi.org/10.2514/1.38338.

Modeling the HIFiRE-1



Full-Order Model → Reduced-Order Model → Error Model

Modeling the HIFiRE-1

Full-Order Model → Reduced-Order Model → Error Model

- Compressible RANS Equations

  $f(q; \mu) = 0$

  μ – vector of system parameters

  $q = [\rho, \rho v, \rho E, \rho \phi]$ – state vector
- Turbulence modeled using SA
- CFD Solver: SPARC [1]
  - Primarily for NNSA's nuclear security programs
  - Uses cell-centered finite volume scheme
  - Developed for transonic and hypersonic flows

[1] Howard, M., Bradley, A., Bova, S. W., Overfelt, J., Wagnild, R., Dinzl, D., Hoemmen, M., and Klinvex, A., "Towards Performance Portability in a Compressible CFD Code," *23rd AIAA Computational Fluid Dynamics Conference*, Vol. 1, AIAA, Denver, CO, 2017. https://doi.org/10.2514/6.2017-4407.

Modeling the HIFiRE-1

Full-Order Model → Reduced-Order Model → Error Model

- Reduced-order model developed using POD with LSPG
  - Clipping function to enforce $\rho > 0$ and $T > 0$

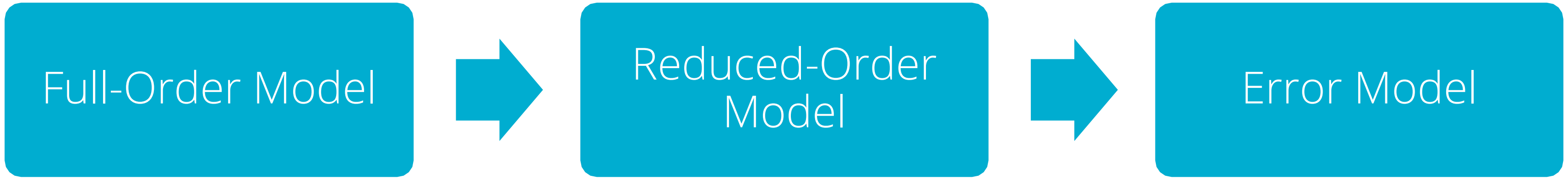
$$\boldsymbol{\Psi}^T \boldsymbol{A}\boldsymbol{f}\left(\boldsymbol{h}(\boldsymbol{q}^0(\boldsymbol{\mu}) + \boldsymbol{\Phi}\hat{\boldsymbol{q}}; \boldsymbol{\mu})\right) = \boldsymbol{0}$$

$$\boldsymbol{\Psi} = \mathbf{A}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{q}}|_{\boldsymbol{h}(\tilde{\boldsymbol{q}})}\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{q}}|_{\tilde{\boldsymbol{q}}}\boldsymbol{\Phi}$$

- ROM solver: Pressio*
  - Open source framework for ROMs
  - Solves large-scale nonlinear dynamical systems
  - Uses generic programming

*https://github.com/Pressio

Modeling the HIFiRE-1



- Predict quantity of interest error

$$\delta_s(\mu) := s(\boldsymbol{\mu}) - \tilde{s}(\boldsymbol{\mu})$$

- Seven regression models [2]
  - MLP, SVR with RBF, SVR with linear kernel, k-NN, linear OLS, quadratic OLS, RF
  - Features include parameters and residual principal components

$$\boldsymbol{q}(\boldsymbol{\mu}) = [\boldsymbol{\mu}; \hat{\boldsymbol{r}}(\boldsymbol{\mu})]$$

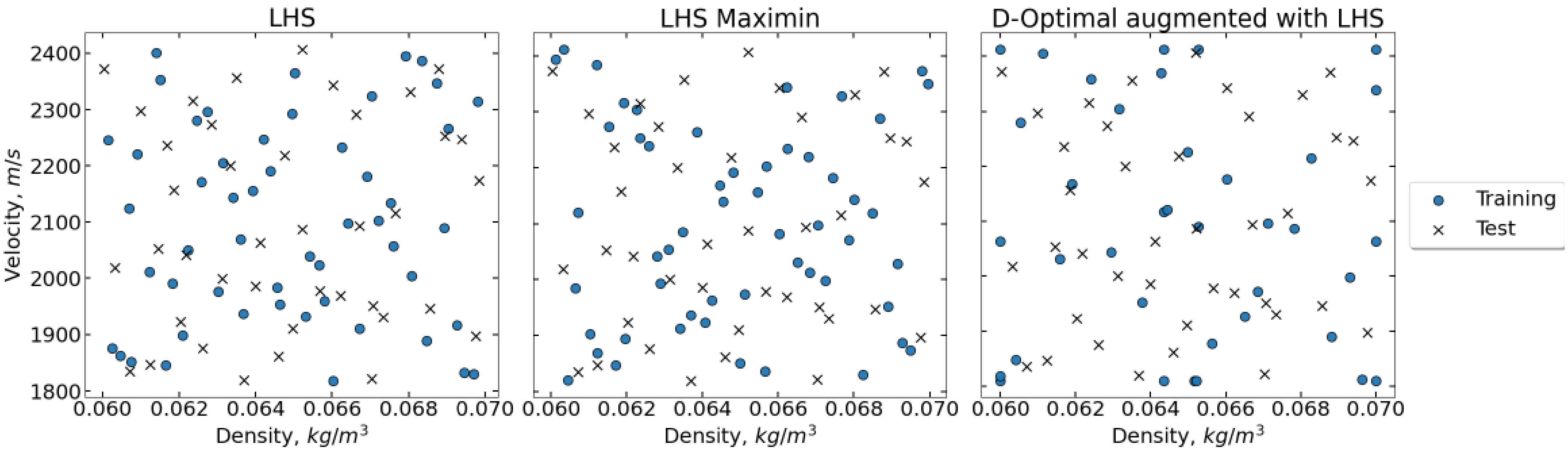
  - $k = 5$ cross-validation to select best hyperparameters

[2] Freno, B. A., and Carlberg, K. T., "Machine-learning error models for approximate solutions to parameterized systems of nonlinear equations," *Computer Methods in Applied Mechanics and Engineering*, Vol. 348, 2019, pp. 250–296. https://doi.org/https://doi.org/10.1016/j.cma.2019.01.024.

Distinct Training Sets

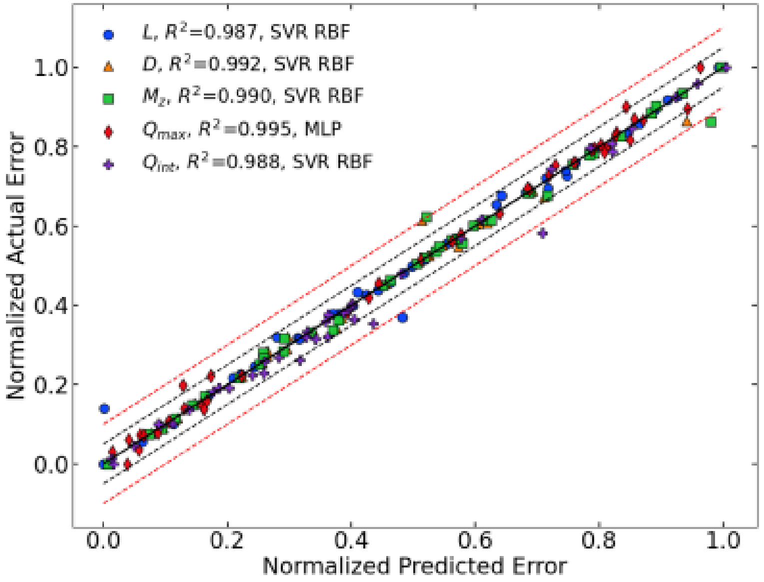
- Reduced-order model trained with 50 points
- Error model trained with 3 sampling types
  - Number of points determined by achieving statistical power of 80%
  - Statistical power: ability to detect whether test deviates from null hypothesis

Sampling Type	[illegible]
LHS	52
LHS Maximin	53
D-Optimal	37

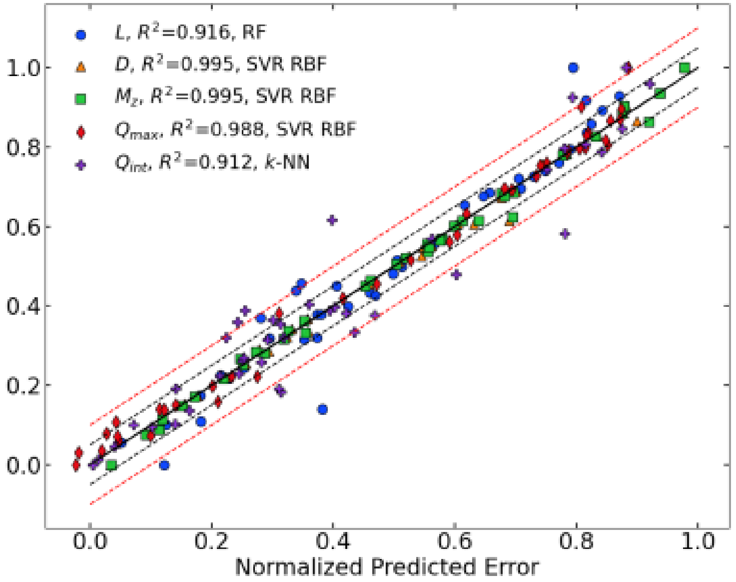


Distinct Training Set Normalized Error

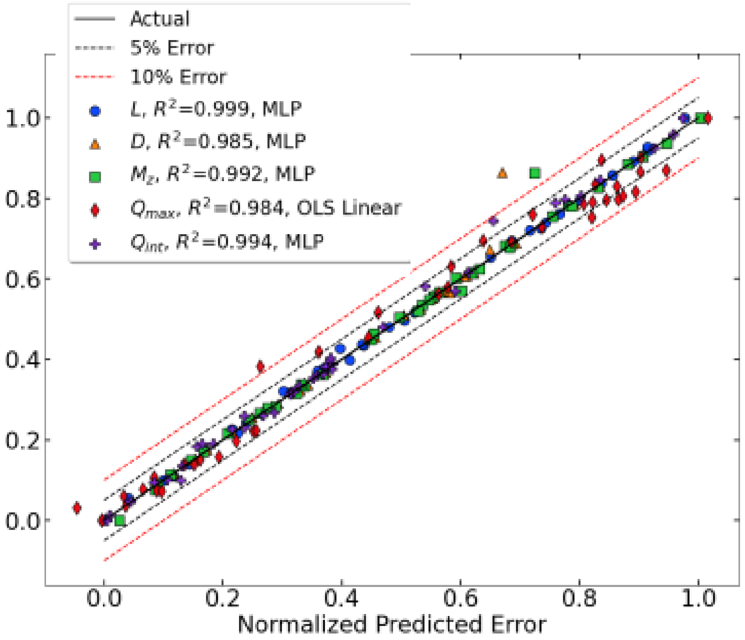
- D-Optimal and LHS sampling types produced best results
  - D-Optimal little bit better than LHS ($\bar{R}_D^2 = 0.991 > \bar{R}_{LHS}^2 = 0.990$)



LHS



LHS Maximin



D-Optimal

Use D-Optimal sampling type for rest of strategies

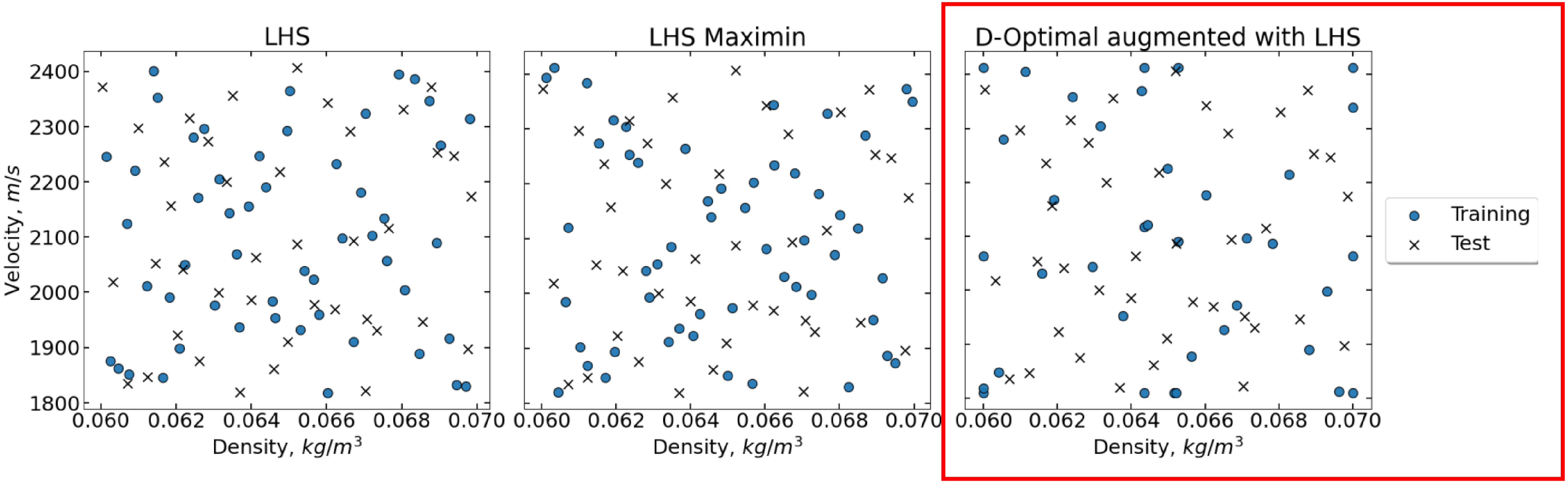
Single Training Set

- Leave-one-out cross validation (LOOCV)
  - Same training set for reduced-order model and error model

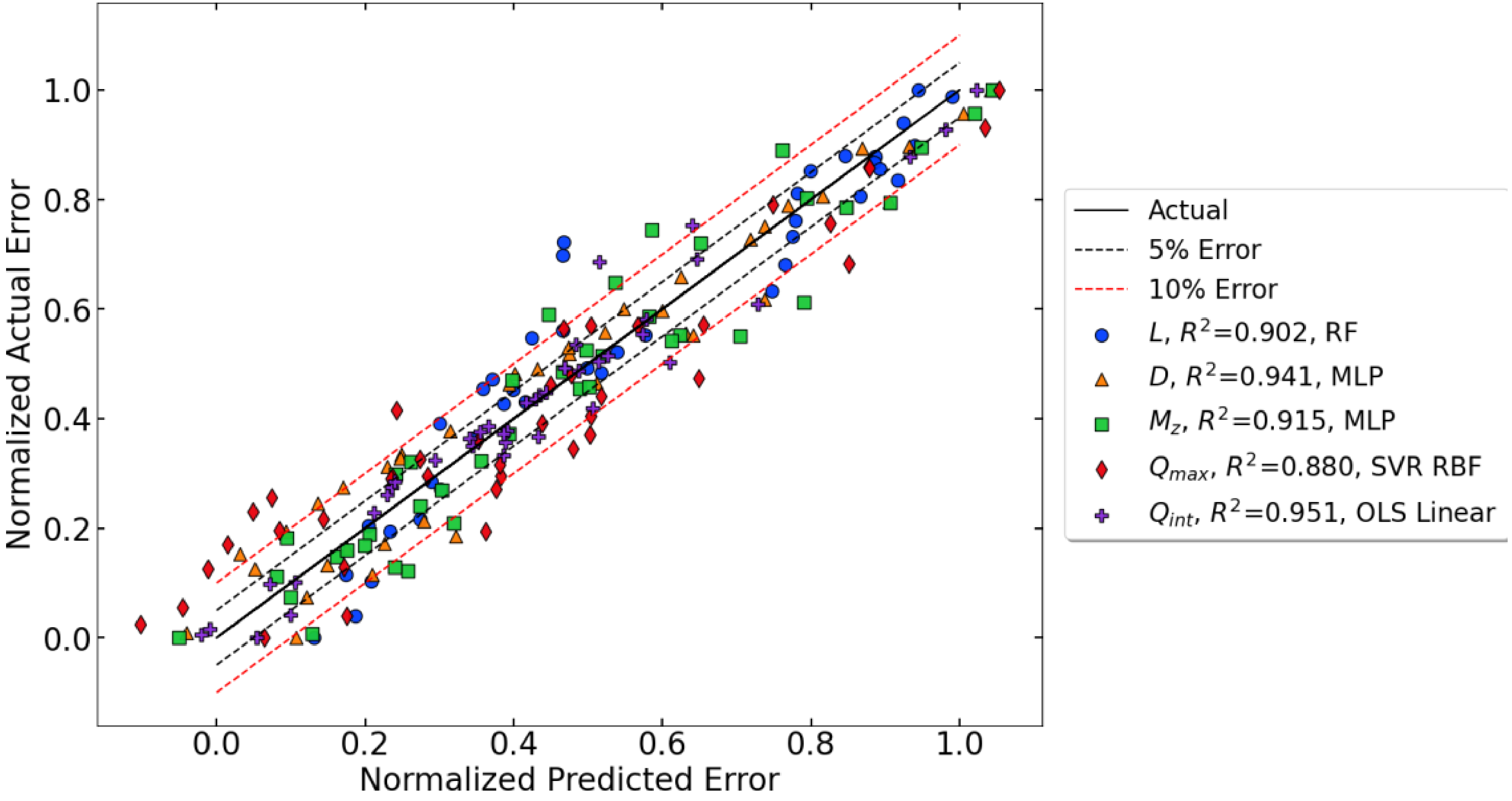
Algorithm

- Set of training points for ROM and Error Model: $\boldsymbol{A} = \{\alpha_i\}_{i=1}^{N_{\text{train}}}$
- Iterate over training points in $\boldsymbol{A}$
  - Select training point $\alpha^* \subset \boldsymbol{A}$
  - Create new training set, $B = \{\alpha_i\}_{i=1}^{N_{\text{train}}-1}$, where $\alpha^* \not\subset B$
  - Train ROM with $\boldsymbol{B}$ and predict at α^*
- Calculate δ_s at training points in $\boldsymbol{A}$
- Train error model using training points in $\boldsymbol{A}$

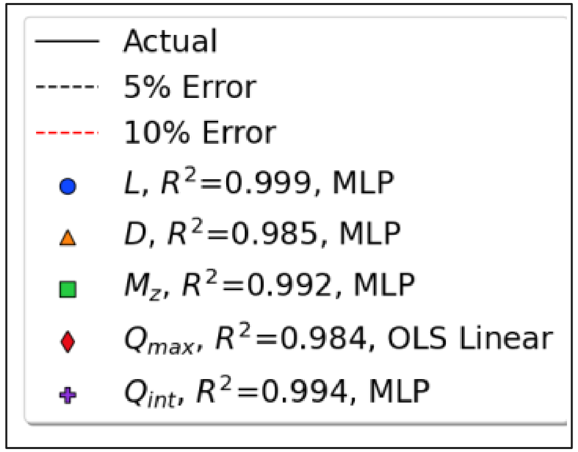
Single Training Set Points



Single Training Set Normalized Error

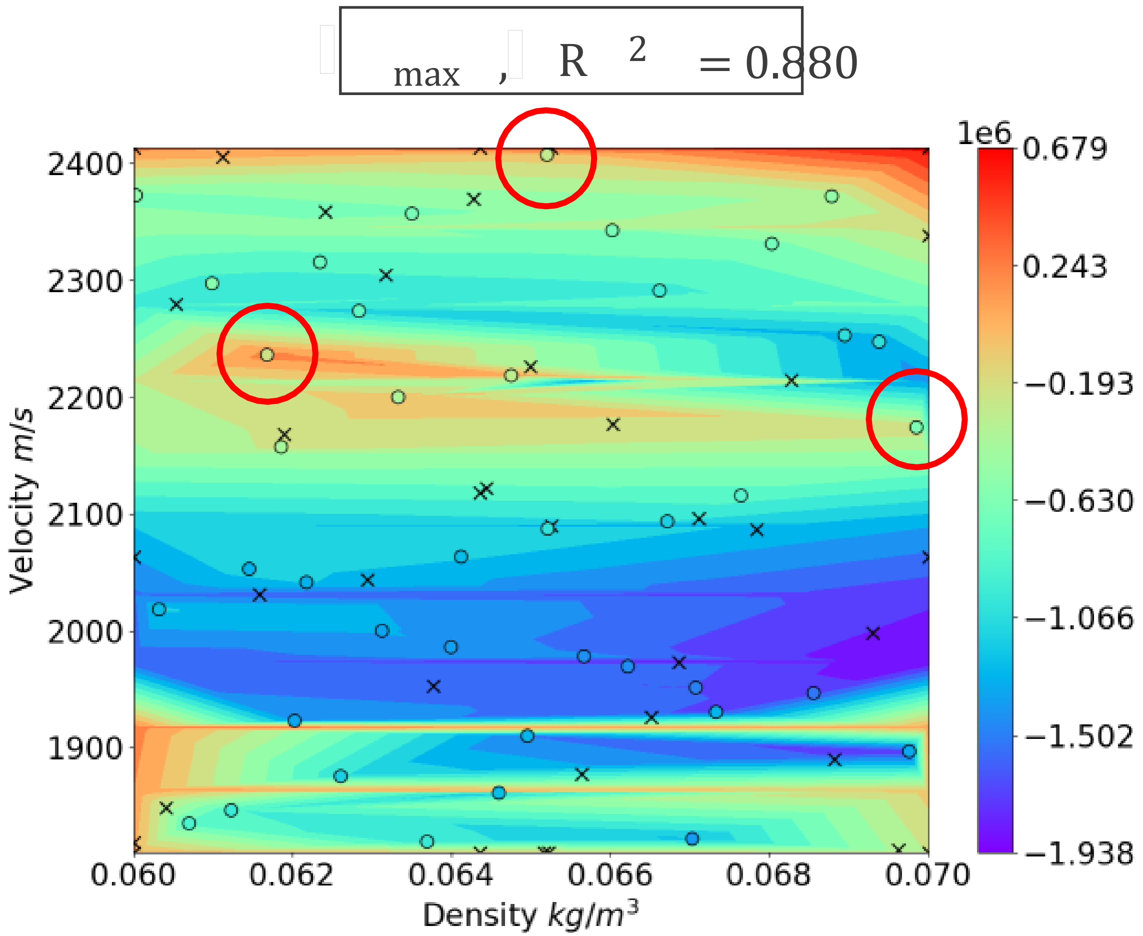


Distinct Training Set



Single Training Set Error Contours

- Highly nonlinear error surface
- Large errors occur where
  - Few training points placed
  - Highly nonlinear areas
- Improve error model by improving spread of training points
  - D-Optimal design augmented with LHS
  - No distance constraint on augmented LHS points

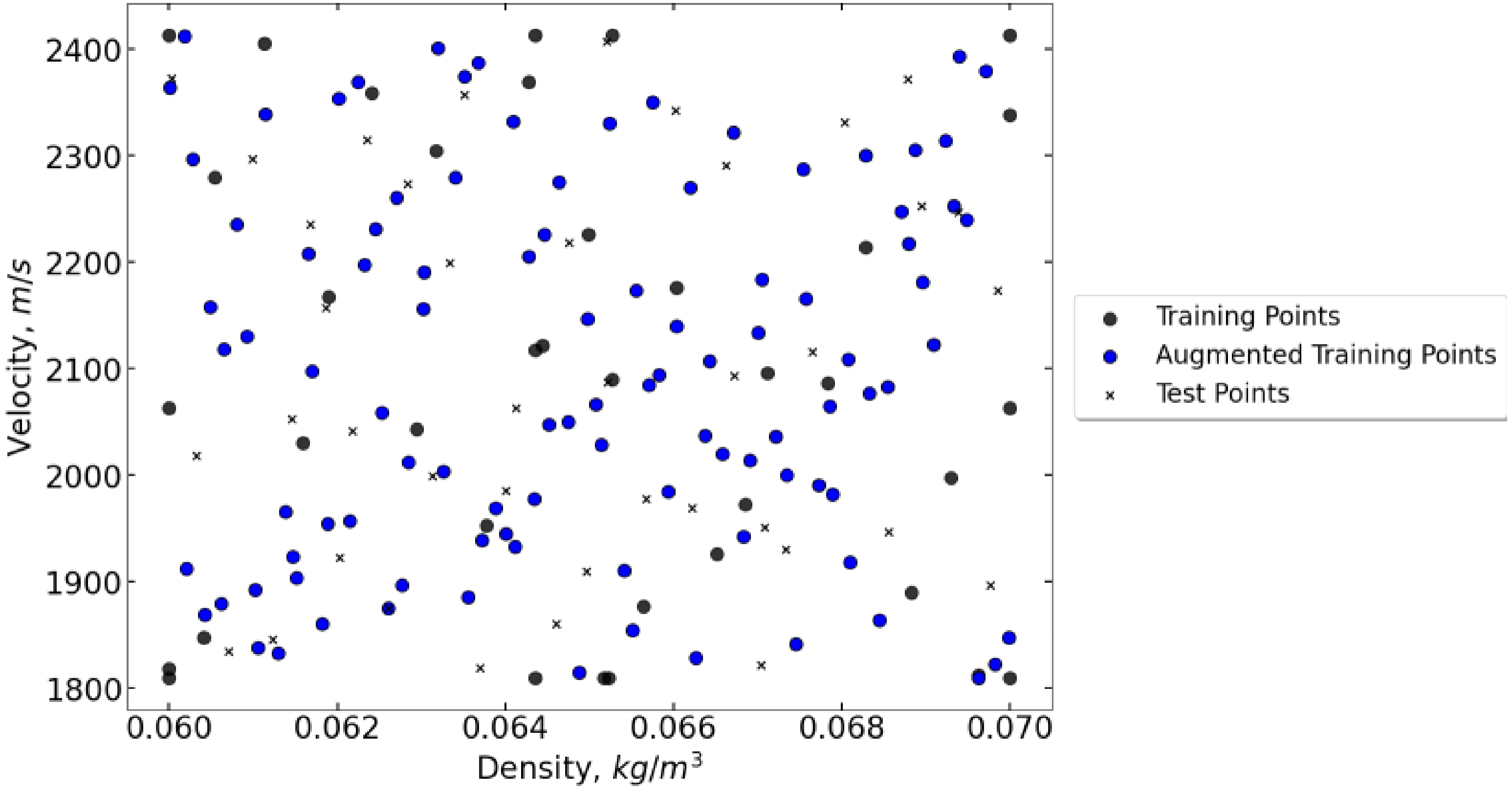


Placement of training points matters!

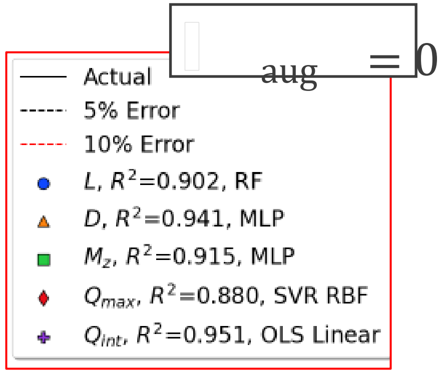
Augmented Training Set

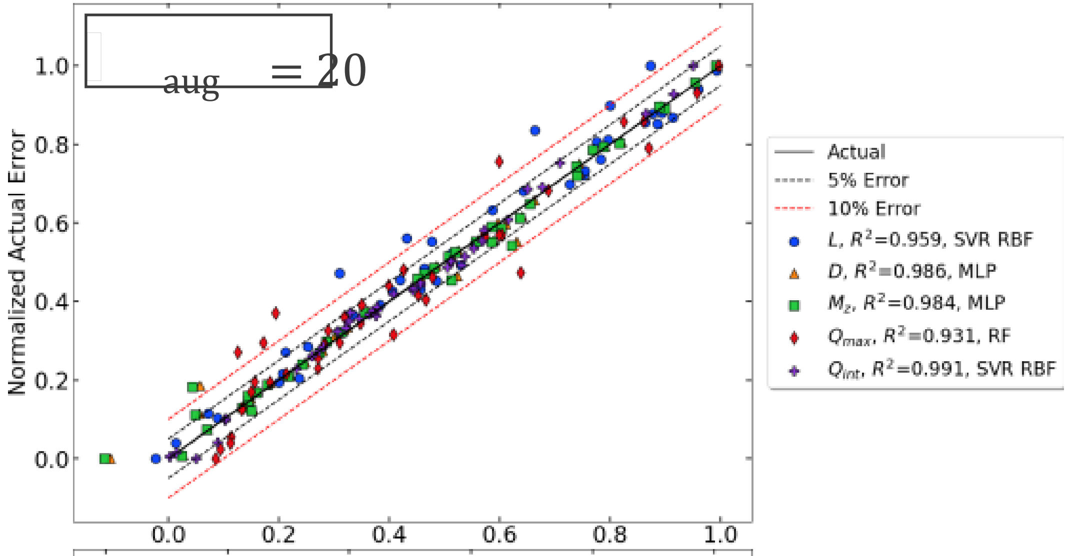
- Improve error model by augmenting with additional points

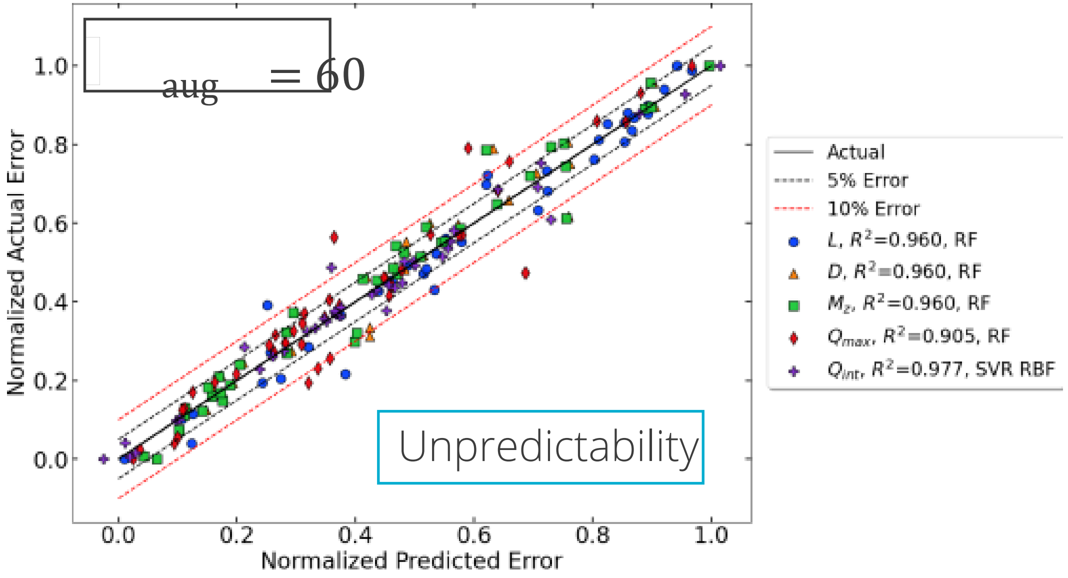
100 Augment Points

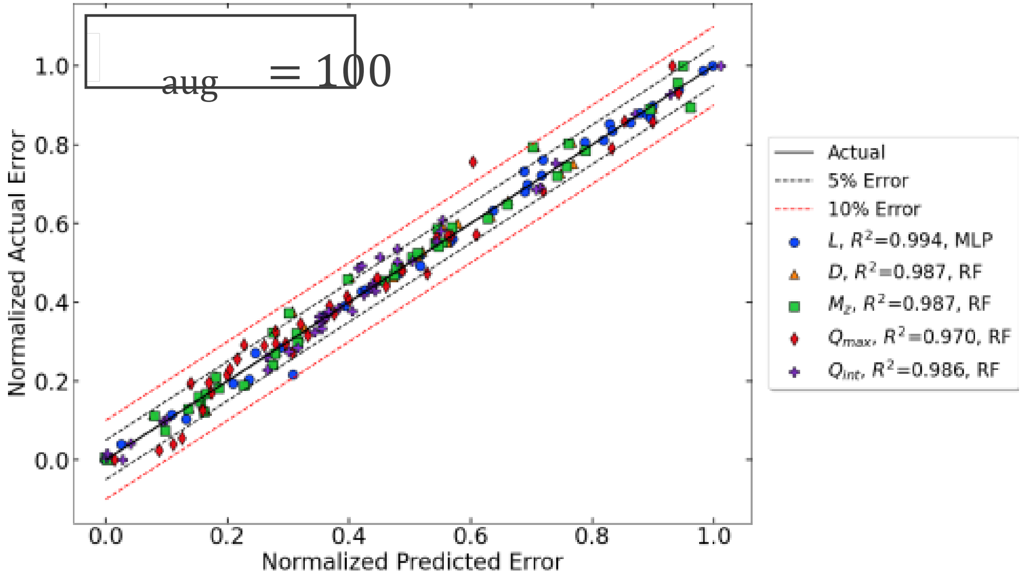


Augmented Training Set Normalized Error









Computational Runtime

- D-Optimal with LOOCV cuts computational expense by 64% compared to Distinct LHS
  - Design with 20 augmented points cuts computational expense by 44%

Category	Sampling Type	N_{train}	Total Time [s] $\times 10^6$	Relative Time to Distinct LHS	$\bar{R}^2_{\text{Error}}$
Distinct	LHS	102	1.43	1.00	0.990
	LHS Maximin	103	1.44	1.01	0.961
	D-Optimal	87	1.22	0.85	0.991
Single Training Set	D-Optimal	37	0.52	0.36	0.918
Augmented Training Set	D-Optimal	57	0.80	0.56	0.970
		77	1.08	0.75	0.954
		97	1.36	0.95	0.952
		117	1.64	1.15	0.962
		137	1.91	1.34	0.985

Conclusions

- D-Optimal design reduces development cost of error model by 15%
  - Reduced total number of training points from 102 to 87
- LOOCV with D-Optimal design reduces development cost of error model by 64%
  - Adding 20 augment points improves accuracy from $\bar{R}^2 = 0.92$ to 0.97
  - Using 20 augment points reduces development cost by 44%
- May improve overall cost reduction by improving POD updates
  - Recalculating POD takes up 30% of overall cost with LOOCV
  - Possibility to use rank-1 updates to POD basis [4]

- The authors would like to thank Eric Parish for his feedback.

[4] Brand, M., "Fast low-rank modifications of the thin singular value decomposition," *Linear Algebra and its Applications*, Vol. 415, No. 1, 2006, pp. 20–30. https://doi.org/https://doi.org/10.1016/j.laa.2005.07.021, special Issue on Large Scale Linear and Nonlinear Eigenvalue Problems.