

# Digital twins for engineering and scientific discovery

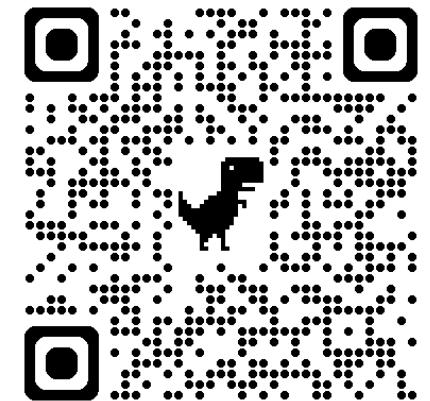
## Multimodality, structure-preservation and UQ

Nat Trask

University of Pennsylvania



Sandia  
National  
Laboratories



# What comes next once forward simulation is “free”?

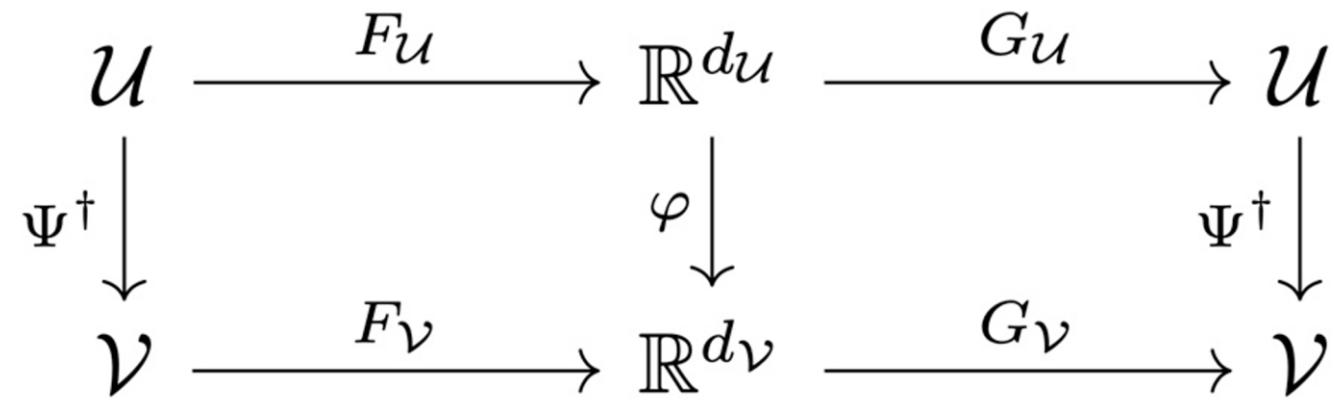
**Many operator regression frameworks realize  
1000x speedup vs traditional methods**

1. Neural operators, DeepONets, foundation models
2. Projection based ROMs

**Key algebraic structures**

1. Latent finite dimensional representations  
of domain  $U$  and range  $V$
2. Mapping  $\phi$  in reduced space

**Limited Physics! Limited guarantees!**



**Key Idea**  
Generate examples of function-to-function  
map that can be used to construct fast  
data-driven models or surrogates in  
reduced dimension space

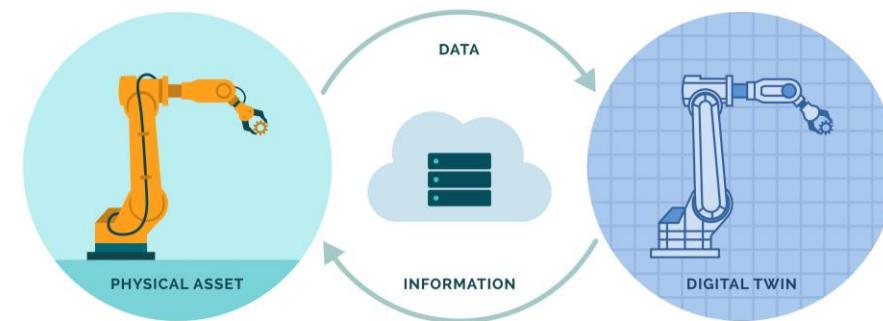
(top) Stuart et al.

# Goal 1: Digital Twins of Systems-of-Systems

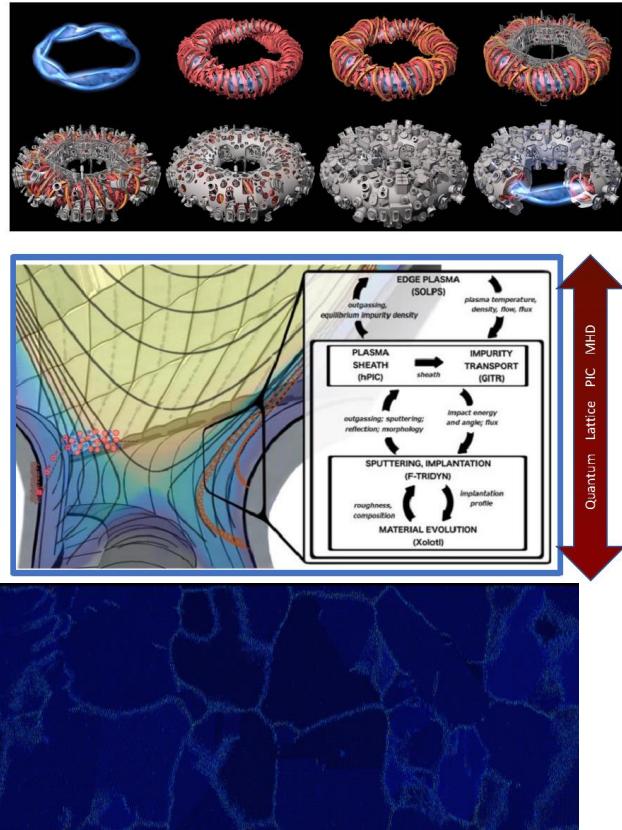
**Learn fast FEM models with parametric dependencies on unknown physics to perform real-time prediction and UQ**

- Preserve physical structure + realizability
- Guarantee numerical stability under extrapolation
- Disentanglement of aleatoric/epistemic uncertainty
- Support real time inference and data assimilation
  - Support causal/mechanistic inference

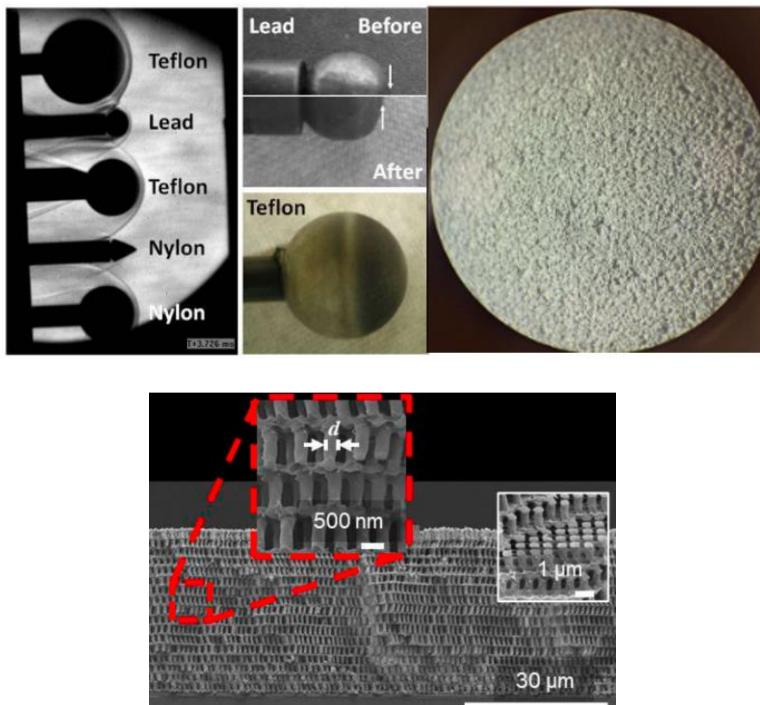
“The size and complexity of many systems being built for government, industry, and the military have reached a threshold where customary methods of analysis, design, implementation, and operation are no longer sufficiently reliable. Many of these large systems are properly described as “systems-of-systems” in that they are composed of many systems” (Dvorak 2005)



# Structure preservation requirements for digital twins



**Stellerator shape optimization**  
Inner loop acceleration to design  
MCF reactors and ICF targets  
*Need exact handoff of conserved  
fluxes, Hugoniot relations,  
Gauge symmetries*



**Multimodal hypersonic metamaterials**  
Non-equilibrium physics in multiscale  
ablation process with indirect  
measurements  
*Need exact mass transfer, non-equilibrium  
chemistry, fluctuation/dissipation balance*

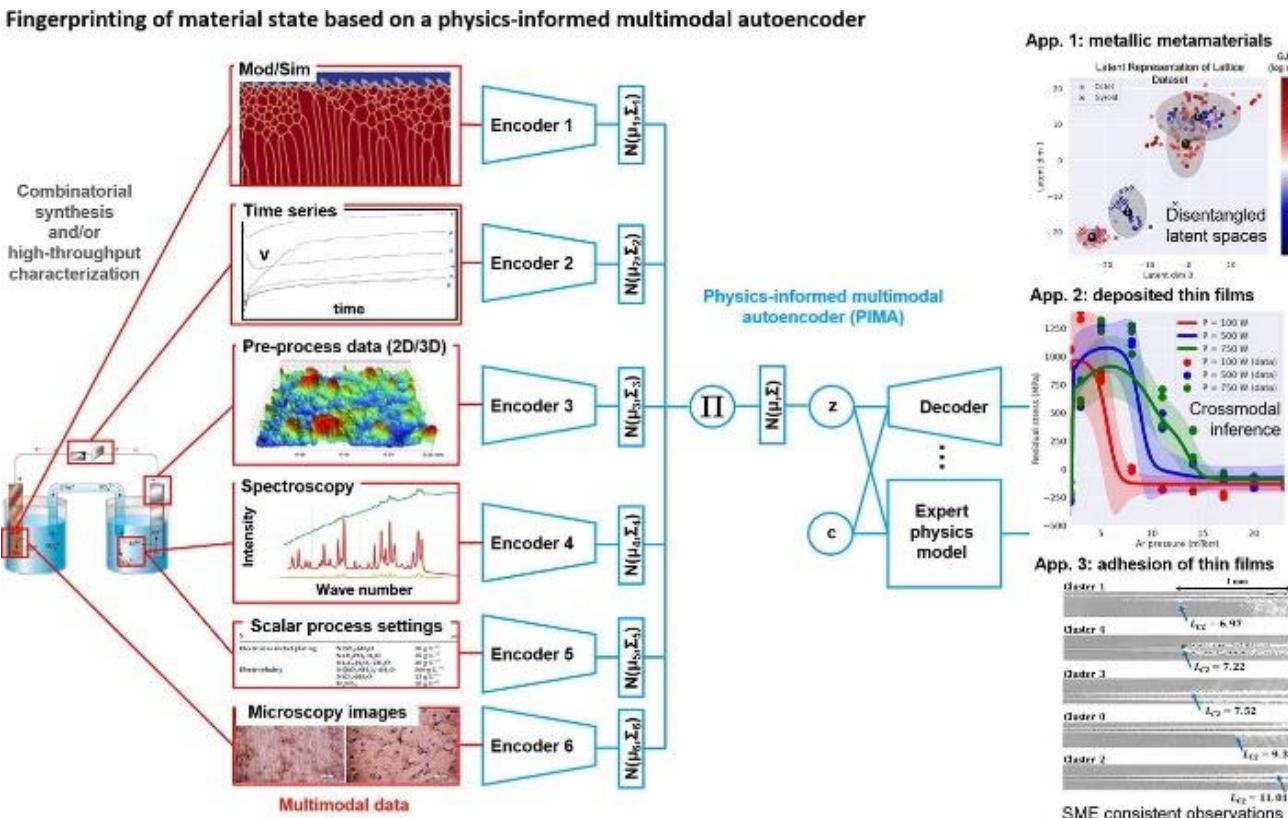


**Data assimilation in climate**  
Causal Fingerprinting in climate  
attribution for legacy DOE codes  
*Detect long range “telecasting” across  
climate subsystems, stably surrogate  
climate subsystems*

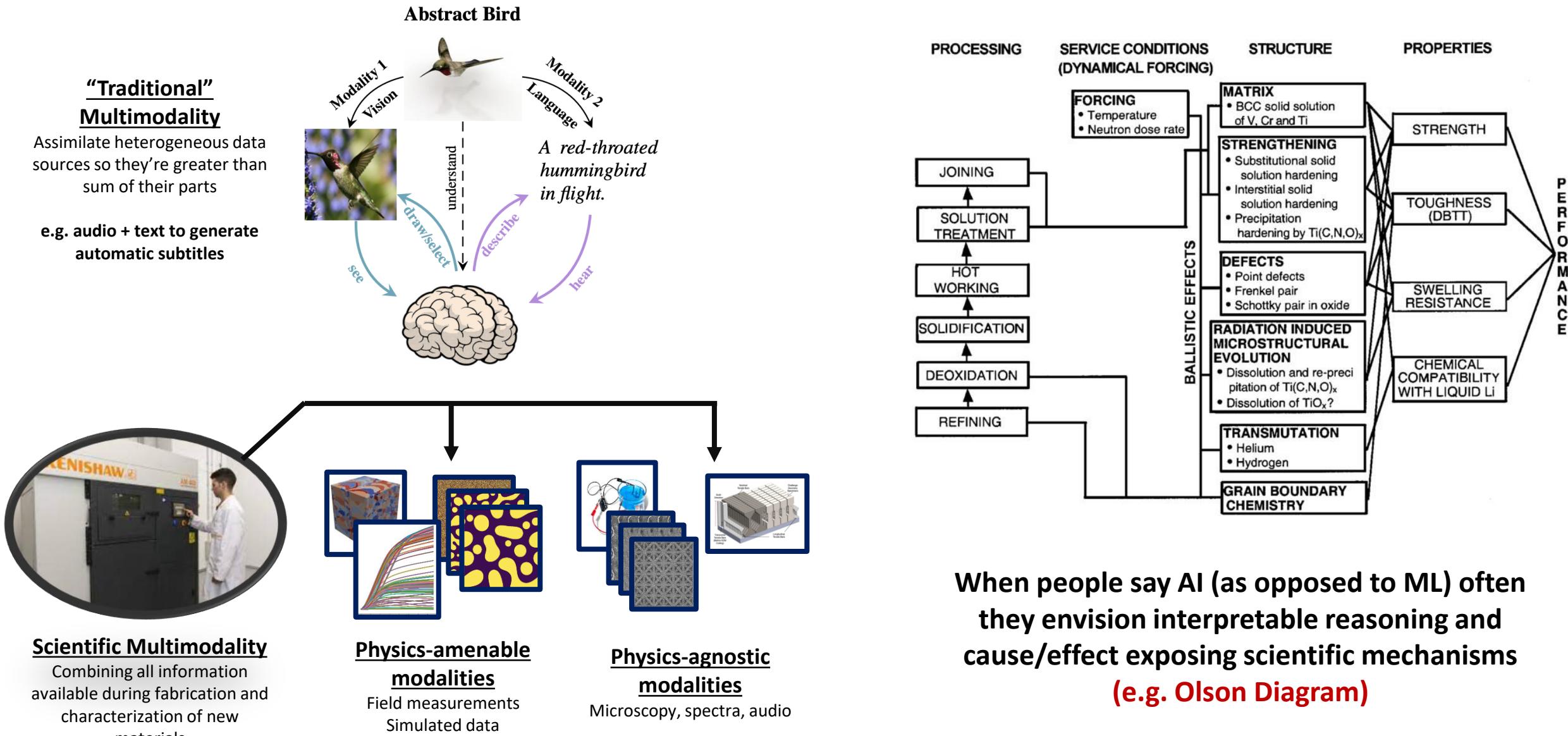
# Goal 2: Integrated physics in AI-driven scientific discovery

Construct foundation models for material systems able to fuse multimodal data, embed physics and autonomously propose experiments

- Handle sparse, multimodal data
- Integrate physical models encoding mechanistic behavior
- Support Bayesian inference to drive optimal experimental design



# Multimodality and mechanistic causality in materials discovery



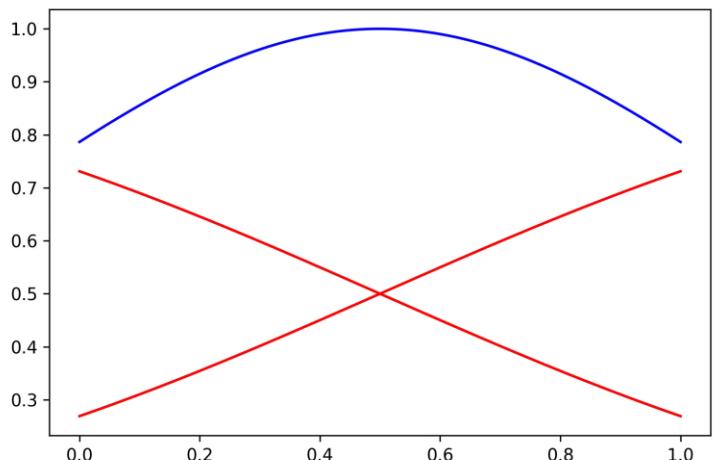
# **Goal 1:**

## **Digital Twins of**

## **Systems-of-Systems**

**Learn fast FEM models with parametric dependencies on  
unknown physics to perform real-time prediction and UQ**

# Structure preserving neural operators via finite element exterior calculus



$$\mathcal{W}_i = \lambda_i$$

$$\mathcal{W}_{ij} = \lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i$$

**Neural Whitney forms**  
 Differentiable architecture  
 parameterizing control  
 volumes and their  
 boundaries

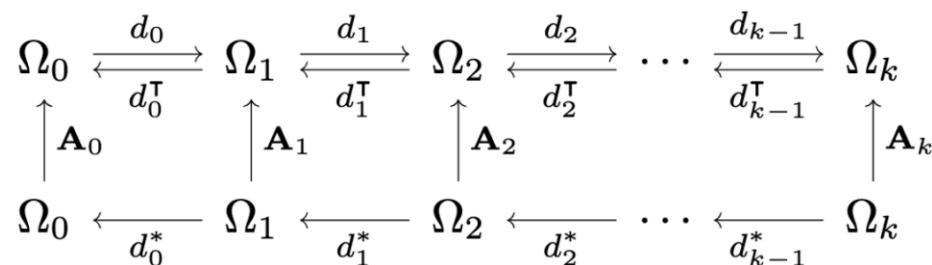
$$d_{k-1} d_{k-1}^* \mathbf{u}_k + d_k^* \mathbf{w}_{k+1} = \mathbf{f}_k$$

**Conservation balance**  
 Exact physics treatment

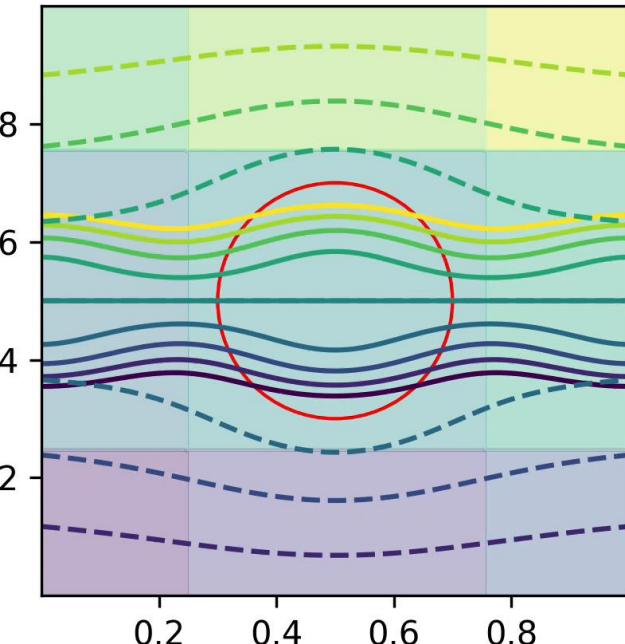
$$\mathbf{w}_{k+1} = d_k \mathbf{u}_k + \mathcal{N}[d_k \mathbf{u}_k; \theta]$$

**Black-box generalized fluxes**

Diffusion stabilized nonlinearity w/ uncertainty

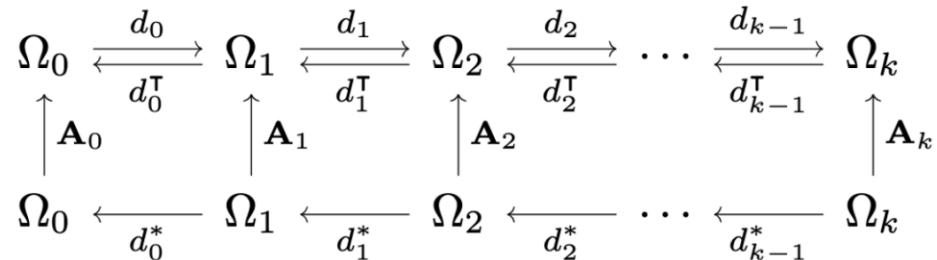


**Exact structure preservation**  
 Machine learnable de Rham  
 complex which provides  
 downstream stability theory



**Data-driven FEM**  
 Simultaneously identify  
 control volumes and integral  
 balance laws whose solution  
 matches (sparse) data

# Result: Combinatorial Hodge + Lax Milgram theory for elliptic operators



**Combinatorial  
Hodge Laplacian**

$$\Delta_k = d_{k-1}d_{k-1}^* + d_k^*d_k$$

Obtain standard results from  
traditional finite element analysis:

- Preserve exact sequence property
- Hodge decomposition
- Poincaré inequality
- Lax-Milgram stability theory
- Conservation structure

Trask, N., Huang, A. and Hu, X., 2022. Enforcing exact physics in scientific machine learning: a data-driven exterior calculus on graphs. *Journal of Computational Physics*, 456, p.110969.

**Theorem 3.1.** *The discrete derivatives  $d_k$  in (11) form an exact sequence if the simplicial complex is exact, and in particular  $d_{k+1} \circ d_k = 0$ . In  $\mathbb{R}^3$ , we have  $CURL_h \circ GRAD_h = DIV_h \circ CURL_h = 0$ .*

**Theorem 3.2.** *The discrete derivatives  $d_k^*$  in (11) form an exact sequence of the simplicial complex is exact, and in particular  $d_k^* \circ d_{k+1}^* = 0$ . In  $\mathbb{R}^3$ ,  $DIV_h^* \circ CURL_h^* = CURL_h^* \circ GRAD_h^* = 0$ .*

**Theorem 3.3** (Hodge Decomposition). *For  $C^k$ , the following decomposition holds*

$$C^k = \text{im}(d_{k-1}) \bigoplus_k \ker(\Delta_k) \bigoplus_k \text{im}(d_k^*), \quad (17)$$

where  $\bigoplus_k$  means the orthogonality with respect to the  $(\cdot, \cdot)_{D_k B_k^{-1}}$ -inner product.

**Theorem 3.4** (Poincaré inequality). *For each  $k$ , there exists a constant  $c_{P,k}$  such that*

$$\|\mathbf{z}_k\|_{D_k B_k^{-1}} \leq c_{P,k} \|d_k \mathbf{z}_k\|_{D_{k+1} B_{k+1}^{-1}}, \quad \mathbf{z}_k \in \text{im}(d_k^*),$$

and another constant  $c_{P,k}^*$  such that

$$\|\mathbf{z}_k\|_{D_k B_k^{-1}} \leq c_{P,k}^* \|d_{k-1}^* \mathbf{z}_k\|_{D_{k-1} B_{k-1}^{-1}}, \quad \mathbf{z}_k \in \text{im}(d_{k-1}).$$

Thus, for  $\mathbf{u}_k \in C^k$ , we have

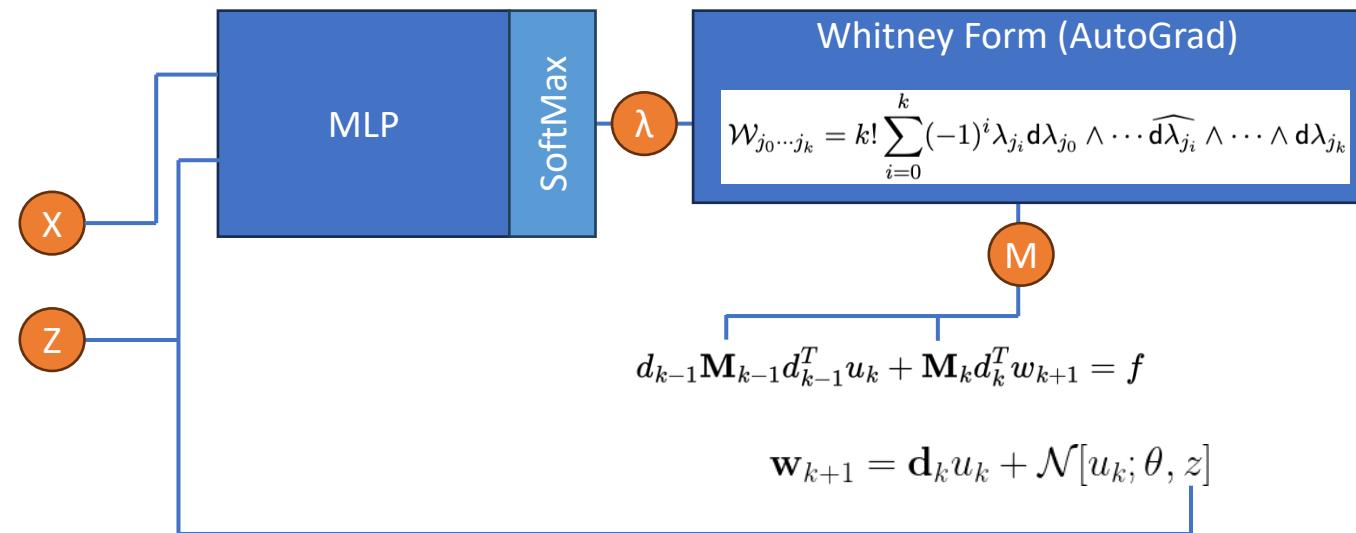
$$\inf_{\mathbf{h}_k \in \ker(\Delta_k)} \|\mathbf{u}_k - \mathbf{h}_k\|_{D_k B_k^{-1}} \leq C \left( \|d_k \mathbf{u}_k\|_{D_{k+1} B_{k+1}^{-1}} + \|d_{k-1}^* \mathbf{u}_k\|_{D_{k-1} B_{k-1}^{-1}} \right),$$

where constant  $C > 0$  only depends on  $c_{P,k}$  and  $c_{P,k}^*$ .

**Theorem 3.5** (Invertibility of Hodge Laplacian). *The  $k^{th}$ -order Hodge Laplacian  $\Delta_k$  is positive-semidefinite, with the dimension of its null-space equal to the dimension of the corresponding homology  $H^k = \ker(d_k)/\text{im}(d_{k-1})$ .*

# Real-time digital twins via conditional neural Whitney forms

Paper to be released end of Jan



Massive strides in **conditional generative modeling**, generating images conditioned on a prompt (above)

We extend the idea to sample from the space of finite element models conditioned on an input Z (sensor readings, parameterized geometry, or a latent variable)  
1000x faster than standard FEM model

$$\underset{\mathbf{A}, \theta}{\operatorname{argmin}} \|\mathbf{u} - \mathbf{u}_{data}\|^2 + \epsilon^2 \|\mathbf{w} - \mathbf{w}_{data}\|^2$$

such that  $a(\mathbf{u}, \mathbf{v}; \mathbf{A}) + N_{\mathbf{v}}[\mathbf{u}; \theta] = b(\mathbf{v}) \quad \forall \mathbf{v}$

Actor, J.A., Hu, X., Huang, A., Roberts, S.A. and Trask, N., 2024. Data-driven Whitney forms for structure-preserving control volume analysis. *Journal of Computational Physics*, 496, p.112520.

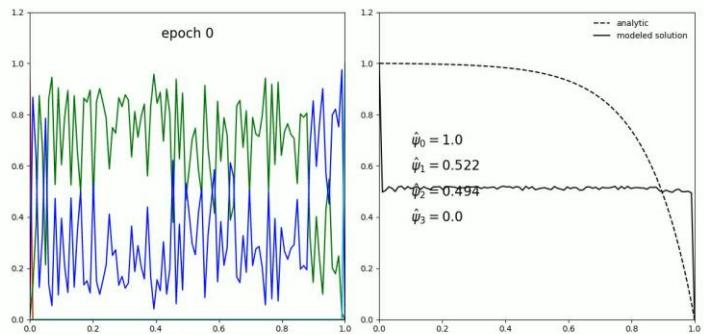
# Supervised conditioning: parametric families of data-driven FEM models

Singularly perturbed advection-diffusion

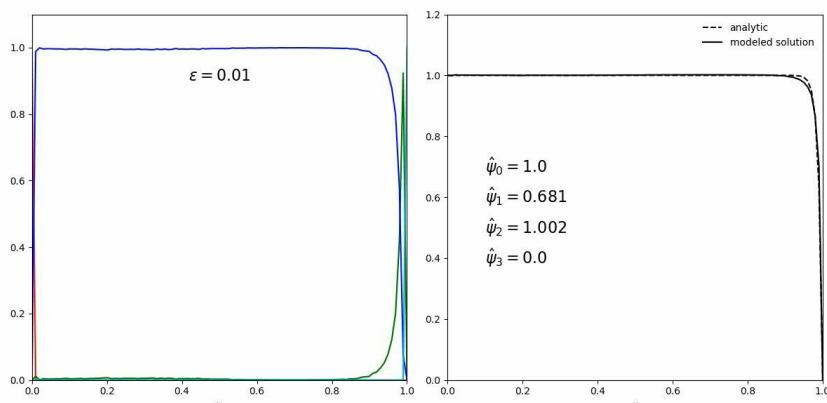
Condition on Peclet number

$$\partial_x u + \frac{1}{Pe} \partial_{xx} u = f$$

$$Pe = \frac{\text{advective transport rate}}{\text{diffusive transport rate}}$$



Unsupervised fitting of subdomains to boundary layers

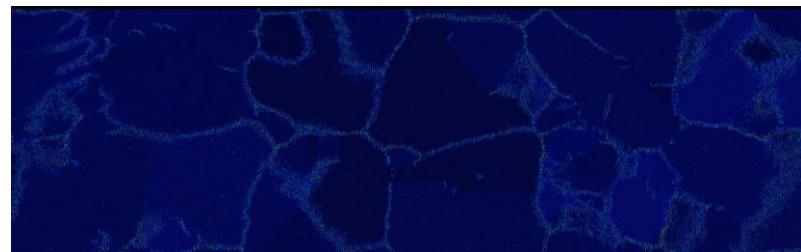
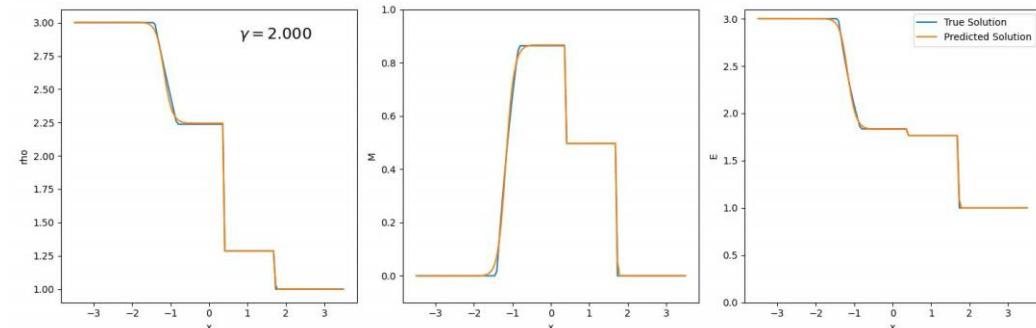
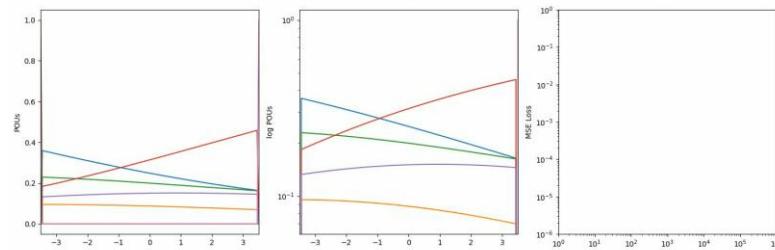
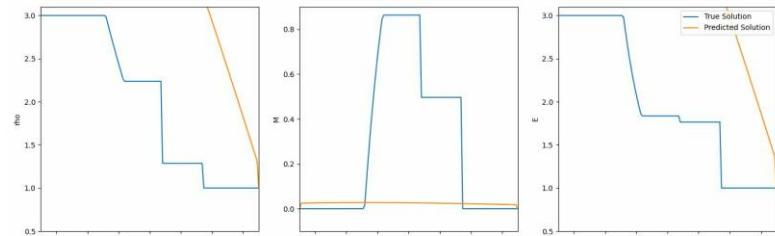


Conditioning of subdomains on Peclet #

Unsupervised fitting of control volumes to strong and weak shock boundaries in spacetime

Conditioning on constitutive equation yields a family of shock models for different materials

Support fitting of data-driven Riemann solvers to upscale MD data

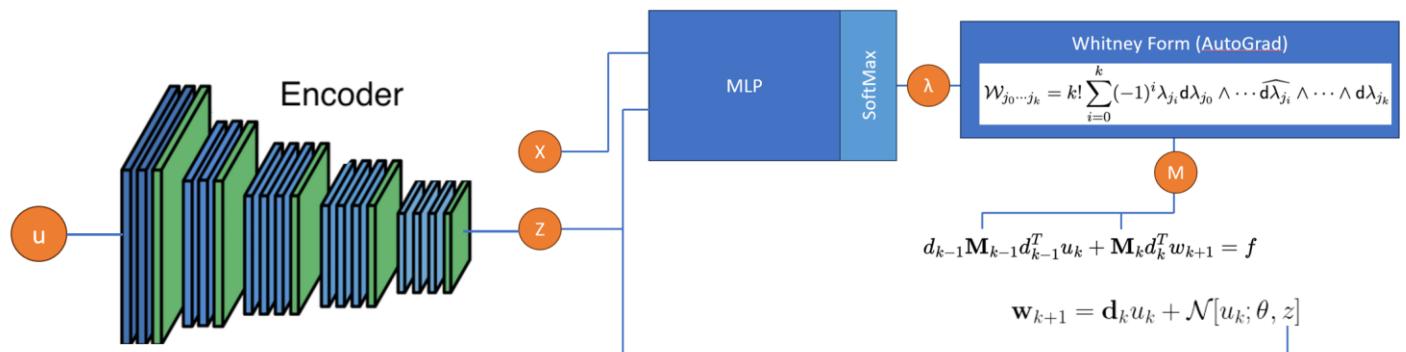


# Unsupervised conditioning: identification of exogenous physics

$$\partial_x u + \frac{1}{Pe} \partial_{xx} u = f$$

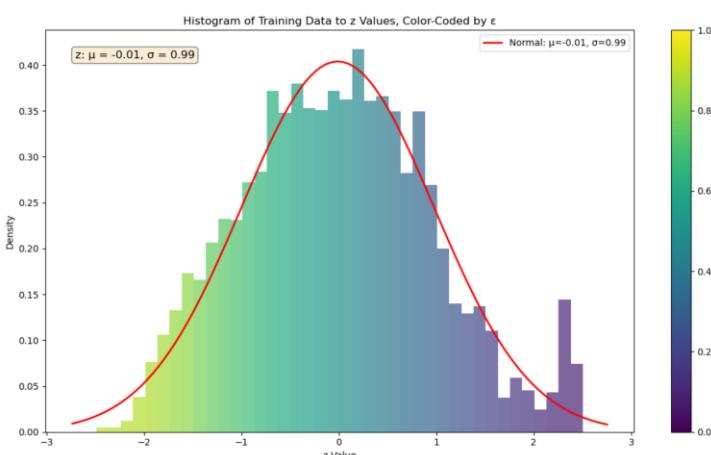
$$Pe = \frac{\text{advective transport rate}}{\text{diffusive transport rate}}$$

$$Pe \sim \mathcal{N}(\mu, \sigma^2)$$

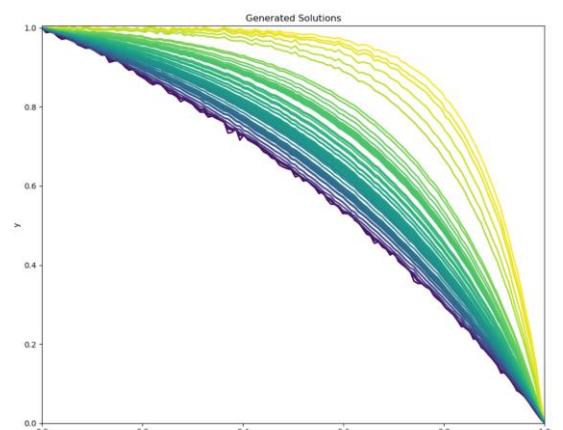


$$\mathcal{L} = -\mathbb{E}[\log p(u|x, z)] + \mathcal{KL}(p(z)||q(z|u, x)) + \lambda^\top (\mathbf{d}_0^\top \mathbf{M}_1 \mathbf{d}_0 u + \mathbf{d}_0^\top \mathbf{M}_1 \mathcal{N}(u; z, \theta) - b)$$

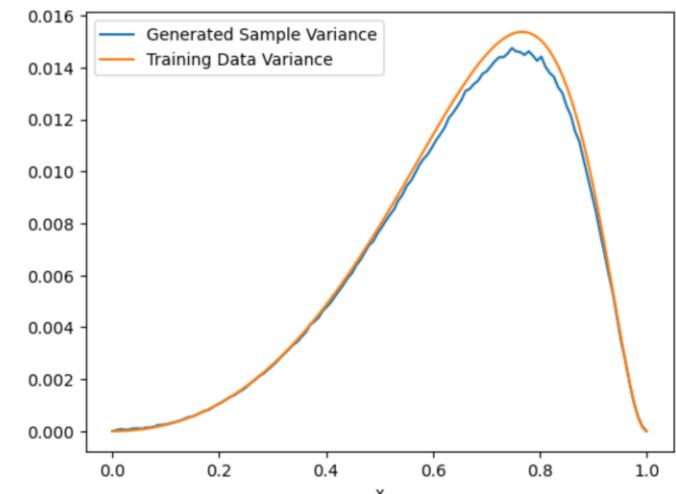
**Reconstruction**      **Prior Penalty**      **Physics constraint**



Latent variable maps directly to Peclet number

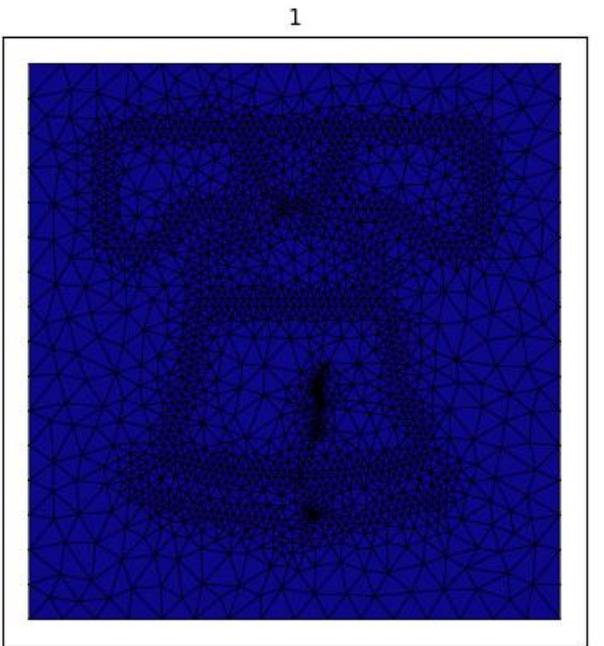
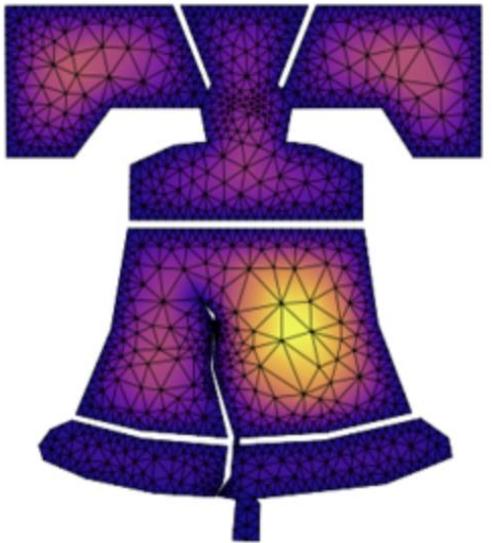


Generative model guaranteed to preserve physical structure



Uncertainty from sampling prior accurately matches data distribution

# Exact structure preservation for complex geometries

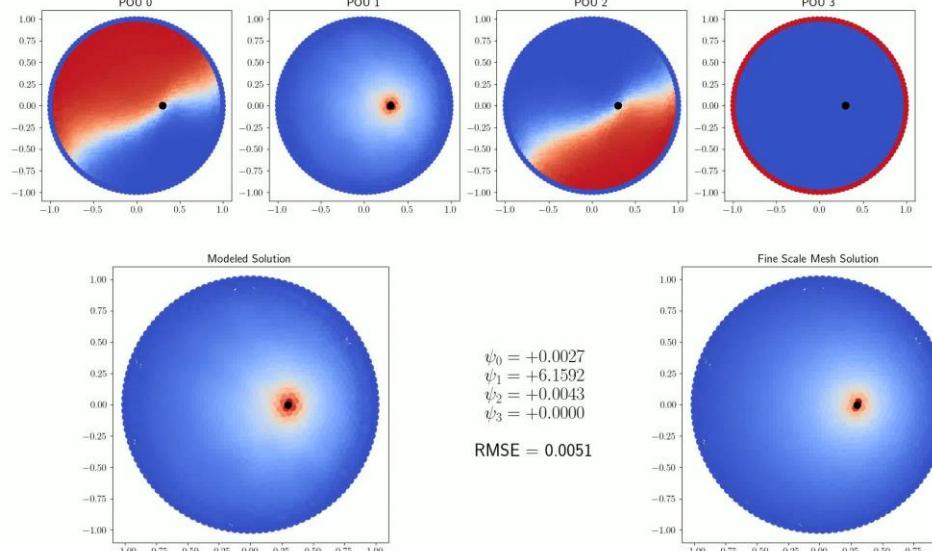


Conditioning on  
material properties

$$\nabla \cdot E = \delta(x - \mathbf{z})$$

Conditioning on  
point charge location

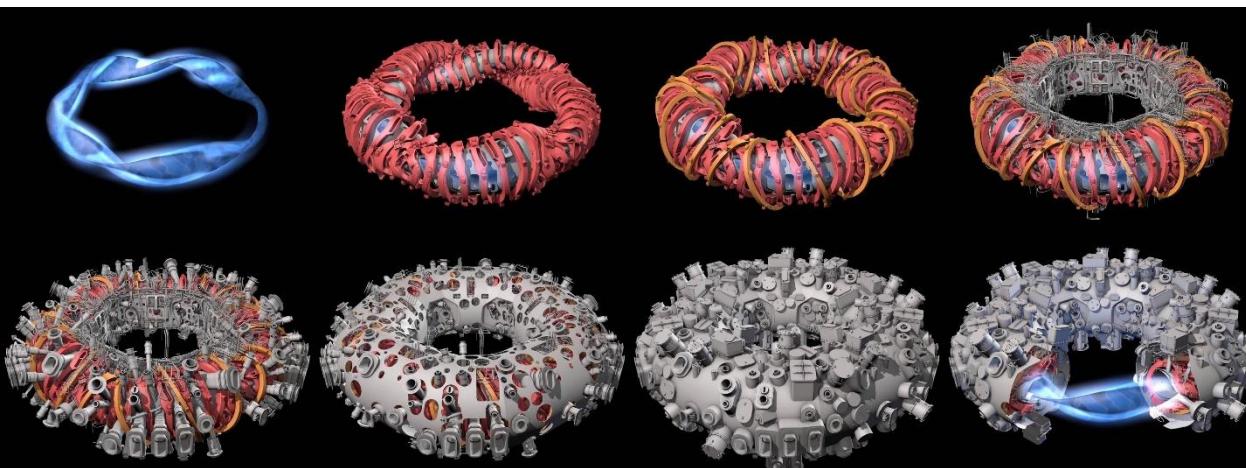
$$E + \nabla \phi = 0$$



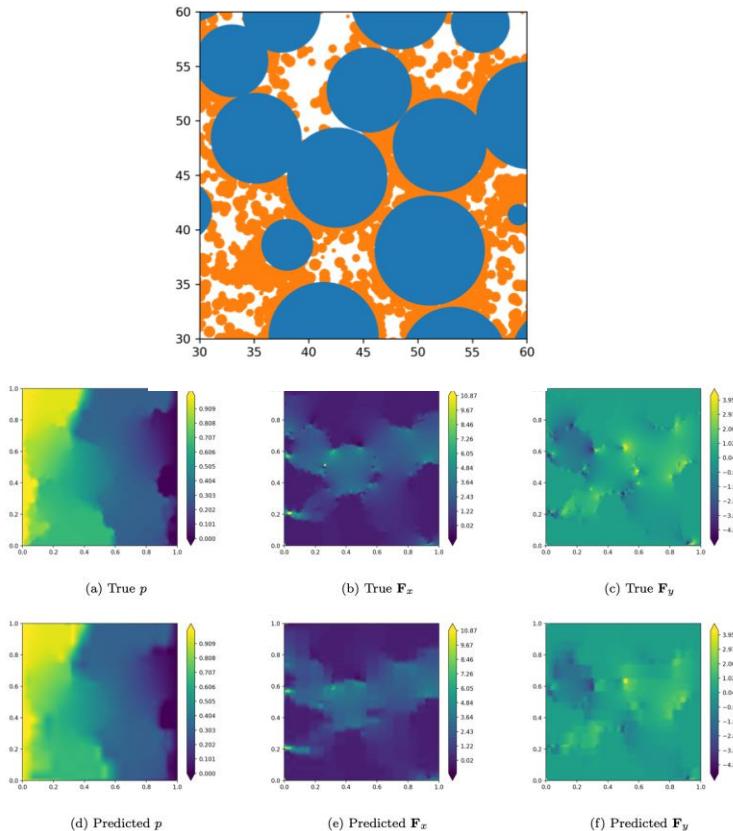
**Complex geometries:** implementations in FENICS to treat models on arbitrary polyhedral meshes without invasive code modifications

**Structure-preserving source identification:** identify sources compatible with exact charge balance

**Extensions to physics-critical systems:** Digital twins of stellarators and inertial confinement fusion targets require exact handoffs of



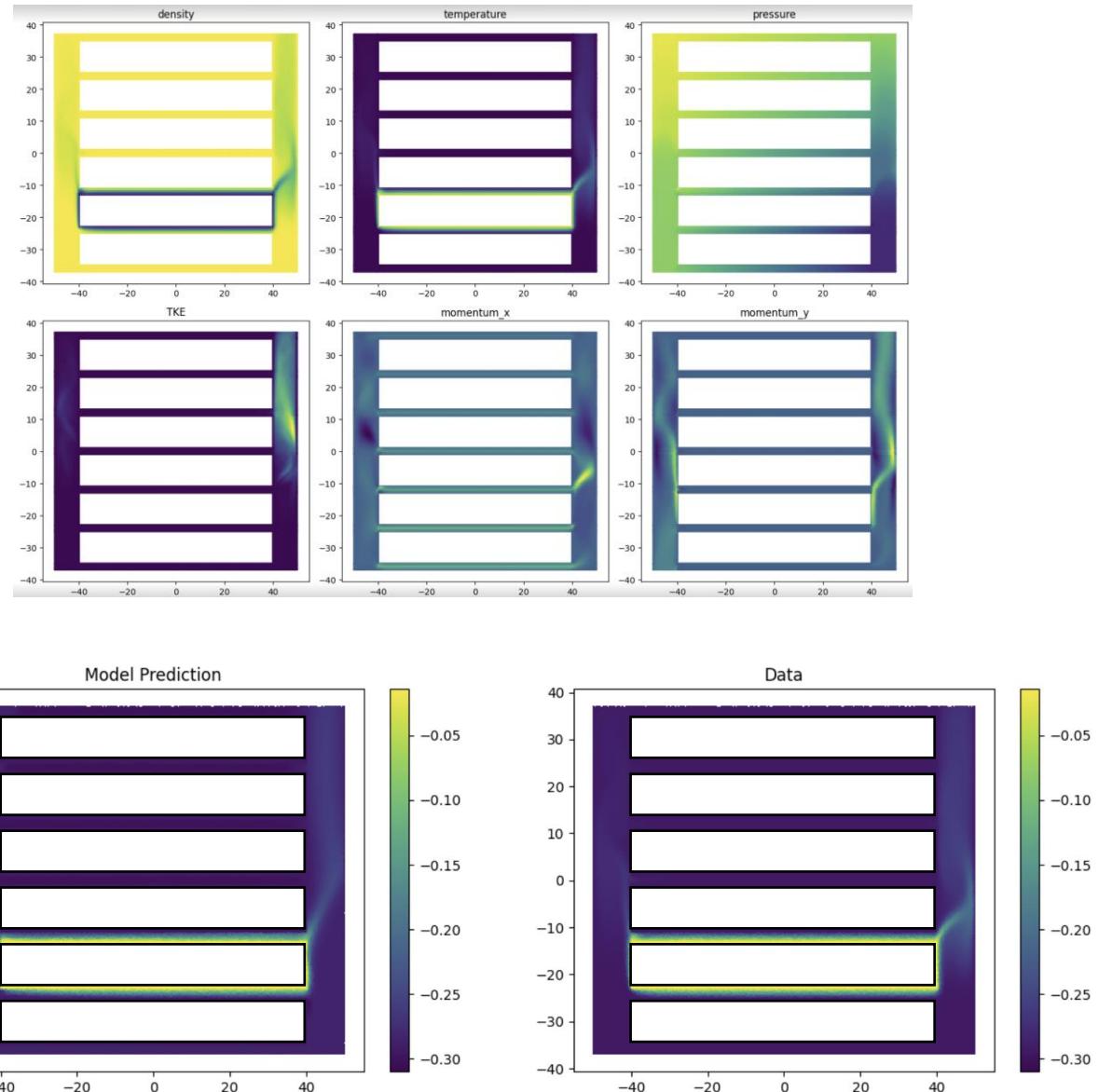
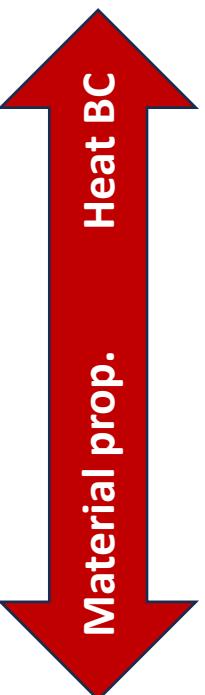
# Li-ion battery digital twin bridging material to engineering scale



**Microstructure:** Replace a ~6M finite element simulation of as-built geometry with 8 data-driven elements w/ ~0.1% error

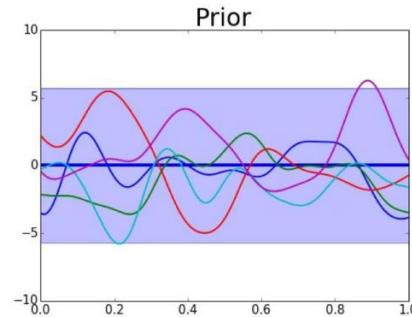
**Battery pack:** Replace ~1M element LES w/ 6 elements reproducing Reynolds average field quantities and providing exact heat flux

Structure preservation allows bi-directional and modular coupling across scales from mesoscale (microns) to server rooms macroscale (meters)

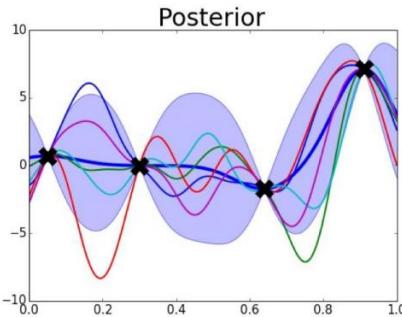


# Beyond BayesOpt – Active learning with DTs

$$y = f(\mathbf{x}) + \epsilon$$



$$f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'; \theta))$$



Training via maximizing the marginal likelihood

$$\log p(\mathbf{y}|\mathbf{X}, \theta) = -\frac{1}{2} \log |\mathbf{K} + \sigma_\epsilon^2 \mathbf{I}| - \frac{1}{2} \mathbf{y}^T (\mathbf{K} + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathbf{y} - \frac{N}{2} \log 2\pi$$

Prediction via conditioning on available data

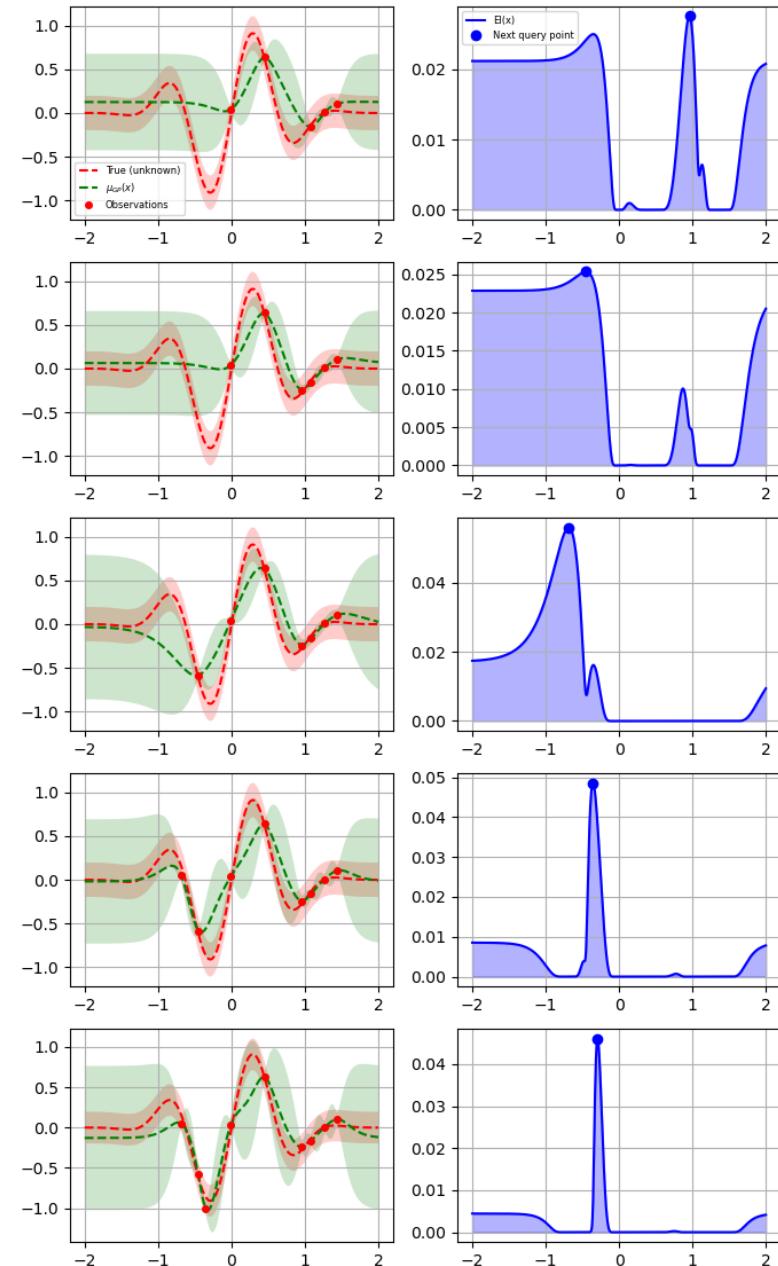
$$p(f_* | \mathbf{y}, \mathbf{X}, \mathbf{x}_*) = \mathcal{N}(f_* | \mu_*, \sigma_*^2),$$

$$\mu_*(\mathbf{x}_*) = \mathbf{k}_{*N} (\mathbf{K} + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathbf{y},$$

$$\sigma_*^2(\mathbf{x}_*) = \mathbf{k}_{**} - \mathbf{k}_{*N} (\mathbf{K} + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathbf{k}_{N*},$$

**The good:** closed form expressions for posterior distributions that can easily be used to design acquisition functions in active learning.

**The bad:** Poor scaling leads to typical tractability only for moderately low-dimension scalar to scalar maps, with limited ability to embed physics.



# Enforcing physics via the optimal recovery problem [Owhadi 2022]

Let  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a symmetric positive definite bivariate kernel, and let  $\mathcal{H}_K = \text{span}\{K(\cdot, x_i)\}$  be the induced RKHS space with accompanying RKHS norm  $\|\cdot\|_K$ . The **optimal recovery problem** consists of finding  $f \in \mathcal{H}_K : \mathbf{X} \subset \mathcal{X} \mapsto \mathbf{Y} \subset \mathbb{R}$ .

$$\min_{f \in \mathcal{H}_K} \|f\|_K^2 + \frac{1}{\epsilon} \|f(\mathbf{X}) - \mathbf{Y}\|_2^2.$$

Expanding in terms of linear algebra

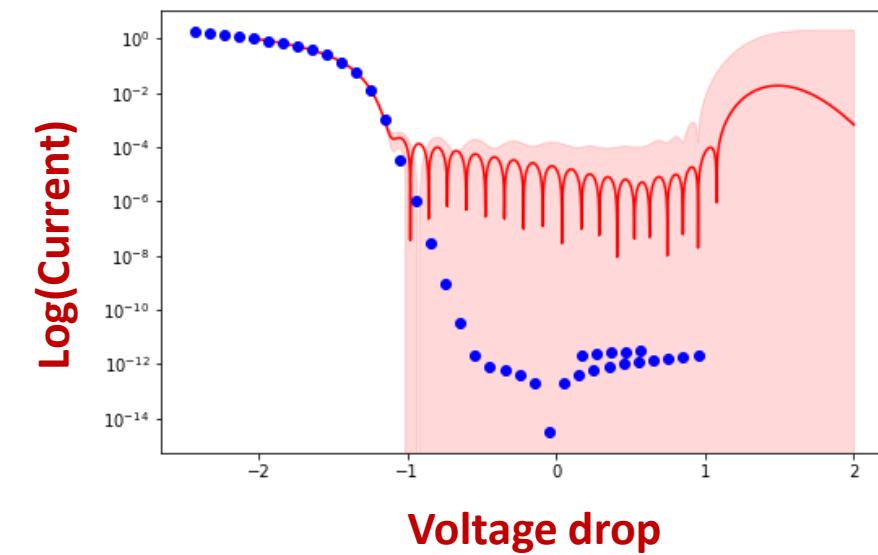
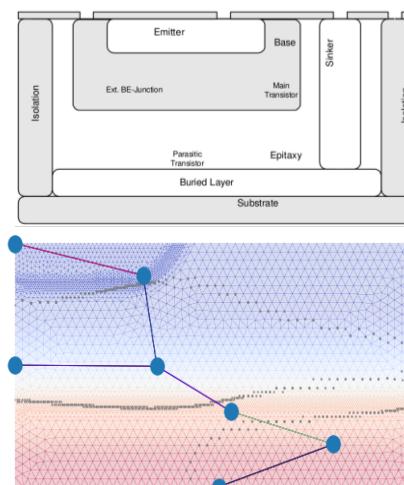
$$V^* = \arg \min_{V \in \mathbb{R}^N} V^T K(\mathbf{X}, \mathbf{X}) V + \frac{1}{\epsilon} \|K(\mathbf{X}, \mathbf{X}) V - \mathbf{Y}\|_2^2.$$

We recover in expectation the traditional Gaussian process posterior

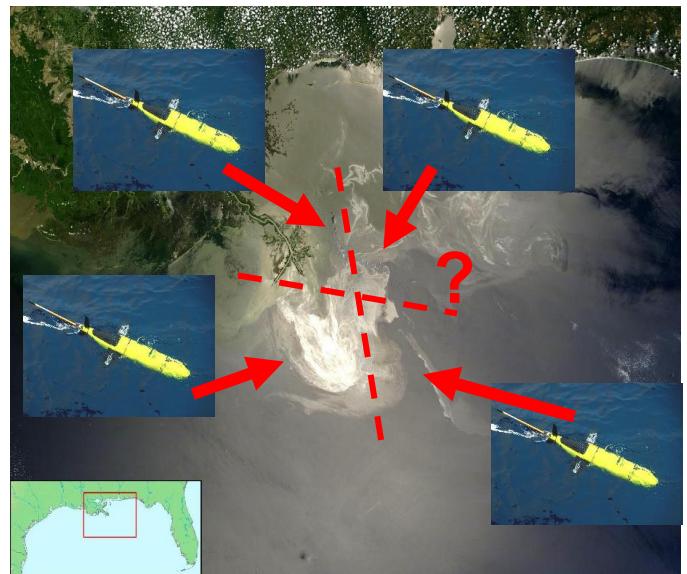
$$f^*(\cdot) = K(\cdot, \mathbf{X})(\mathbf{K} + \epsilon I)^{-1}\mathbf{Y}.$$

**Reframe control volume fluxes as constrained GPs**

$$\begin{aligned} \min_{\theta, \mathbf{u}_{\text{un}}} \min_{\mathbf{F}_{\text{un}}} & \sum_{e \in \mathcal{E}} \mathbf{F}_e^T (K_e(\delta_0 \mathbf{u}_e, \delta_0 \mathbf{u}_e) + \epsilon I)^{-1} \mathbf{F}_e \\ & + \log \det(K_e(\delta_0 \mathbf{u}_e, \delta_0 \mathbf{u}_e) + \epsilon I) \\ \text{s.t. } & \mathbf{d}_k^\top \mathbf{F} = 0 \end{aligned}$$



# Feeding the DT beast: real time data assimilation and the optimal coverage problem



[Deepwater Horizon, 2010](#)  
[NASA/GSFC, MODIS](#)

Minimize: Sensor degradation

$$\mathcal{H}(P, \mathcal{W}) = \sum_{i=1}^n \int_{W_i} f(\|q - p_i\|) \phi(q) dq \quad (1)$$

Sensor location      Dominance region      Importance

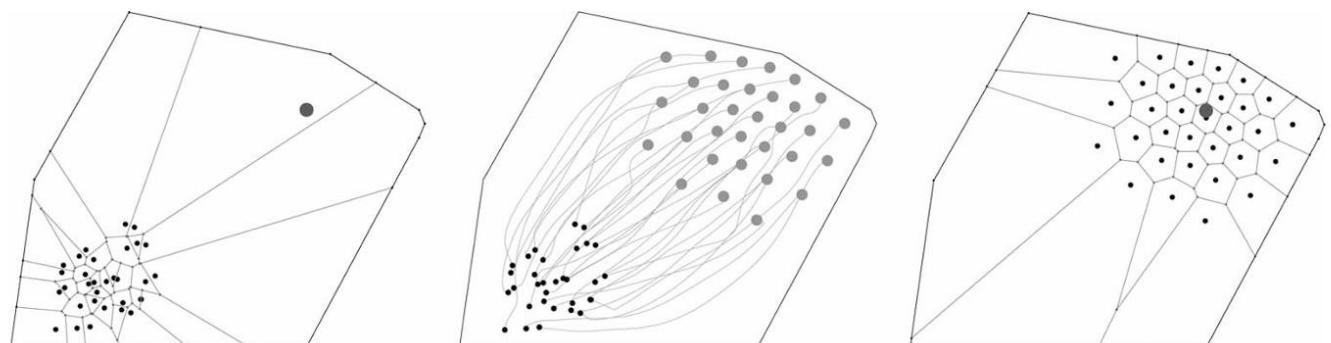
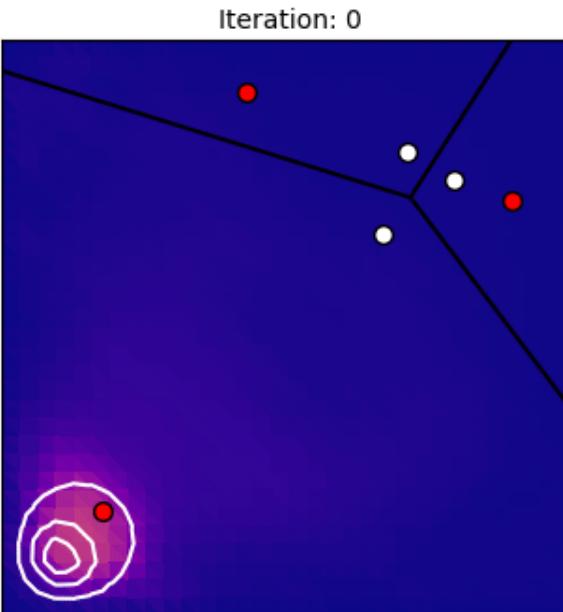
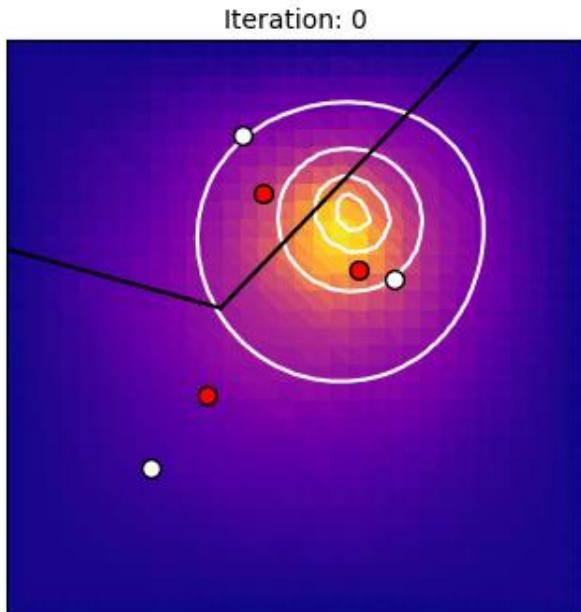


Fig. 3. Lloyd continuous-time algorithm for 32 agents on a convex polygonal environment, with the Gaussian density function of Fig. 1. The control gain in (6) is  $k_{\text{prop}} = 1$  for all the vehicles. The left (respectively, right) figure illustrates the initial (respectively, final) locations and Voronoi partition. The central figure illustrates the gradient descent flow.

Can lightweight digital twins be embedded on distributed sensor platforms to guide collection of data in near-real time?

# Feeding the DT beast: real time data assimilation and the optimal coverage problem



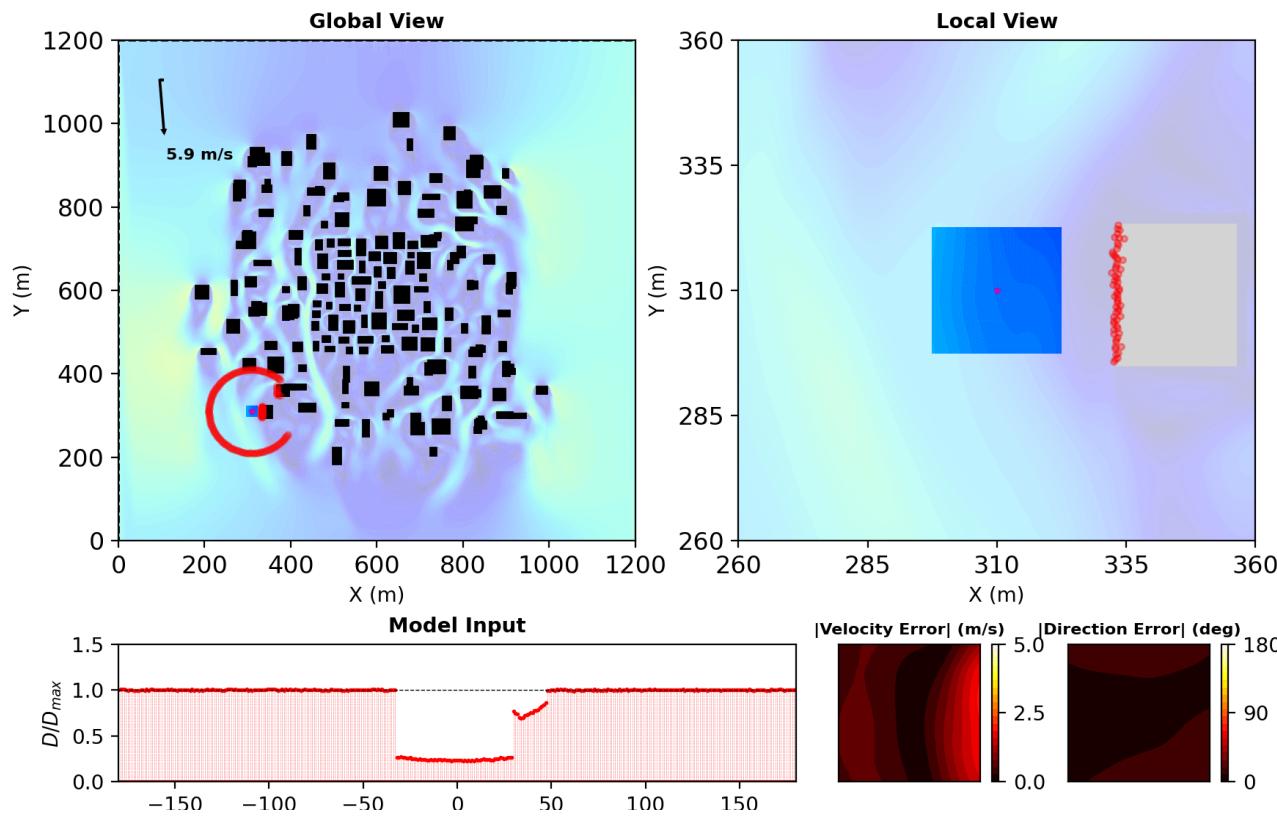
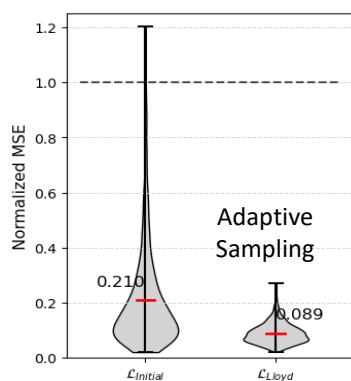
**Left:** Static source  
**Right:** Moving source

**White points:** sensors

**Red points:** weighted centroid

**Heatmap:** CNWF solution based on sensor data

**Contour:** true FEM solution



**Left:** Control scheme to identify physical model for source location for a stationary (*left*) and moving (*right*) source.

**Right:** Collaboration with robotics group at Upenn (Hsieh, Kumar) to sense physics based digital twins in an urban environment

## **Goal 2: Integrating physics into AI-driven scientific discovery**

**Construct foundation models for material systems able to fuse  
multimodal data, embed physics and autonomously propose  
experiments**

# Sparse data in material discovery: high-throughput experiments and multimodality

## AI-enabled thin film design:

Laser powderbed fusion  
Physical vapor deposition  
Electroplating



**Even with high-throughput, data is small!**

Conventional campaigns target a handful of bespoke experiments

High-throughput collect hundreds – far from the billions of datapoints in LLMS!

**Can we compensate with multimodality?**

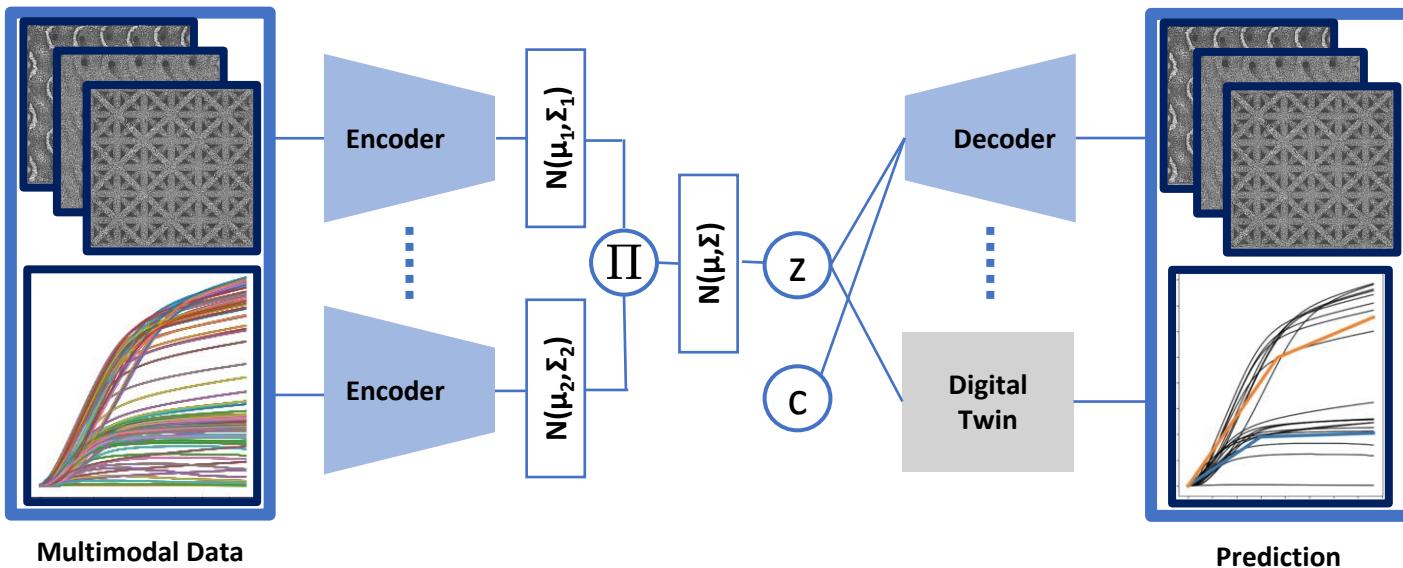


**Process parameters:** laser power, speed, path, powder composition

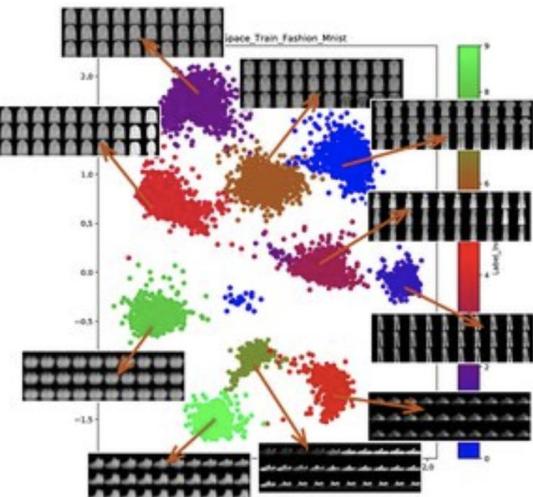
**High-fidelity characterization:** microscopy, XRF, XRD, TEM, SEM

**Low-fidelity signals:** light, sound, images, profilometry

# Physics-informed multimodal autoencoder (PIMA)



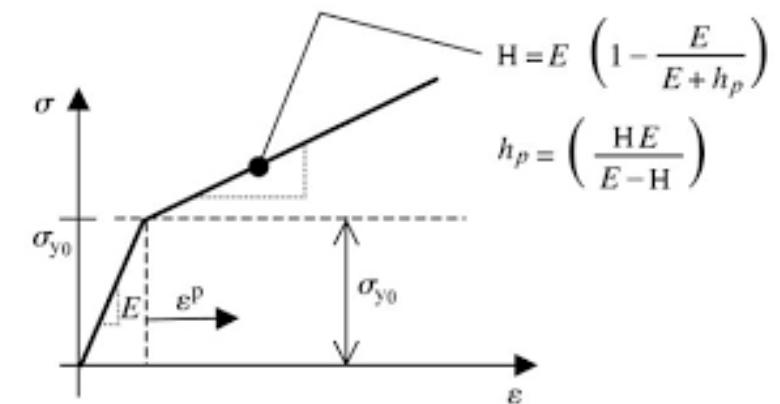
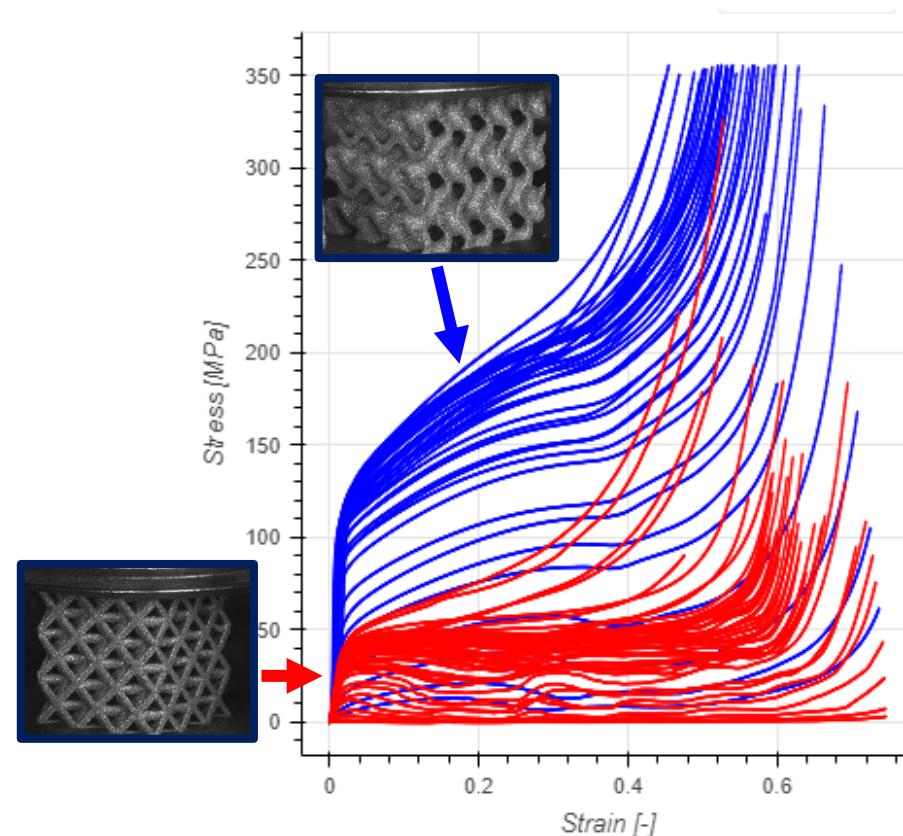
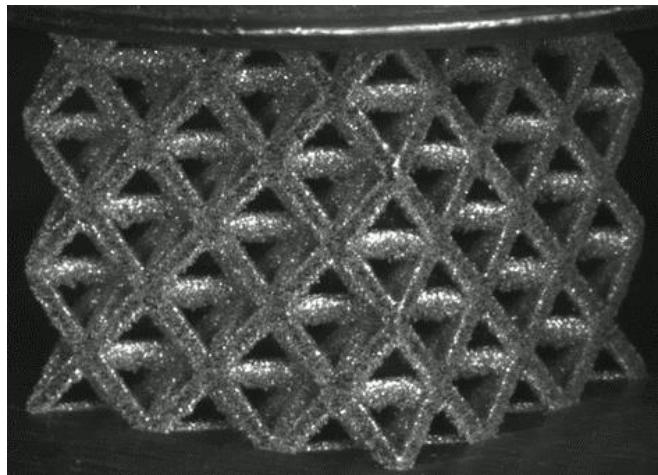
**Informal Idea:**  
We discover a shared  
latent representation  
of data providing a  
Rosetta stone  
across modalities  
w/ uncertainty  
estimation



## Details:

- Gaussian product distribution gives deep posterior embedding for each modality
- Gaussian mixture prior in latent space identifies populations in data across modalities
  - *Closed form expressions* for ELBO – no Monte Carlo
- Supports Bayesian inference across modalities
  - Fast training with EM

# A simple multimodal dataset: images+uniaxial tension of metal additive lattices



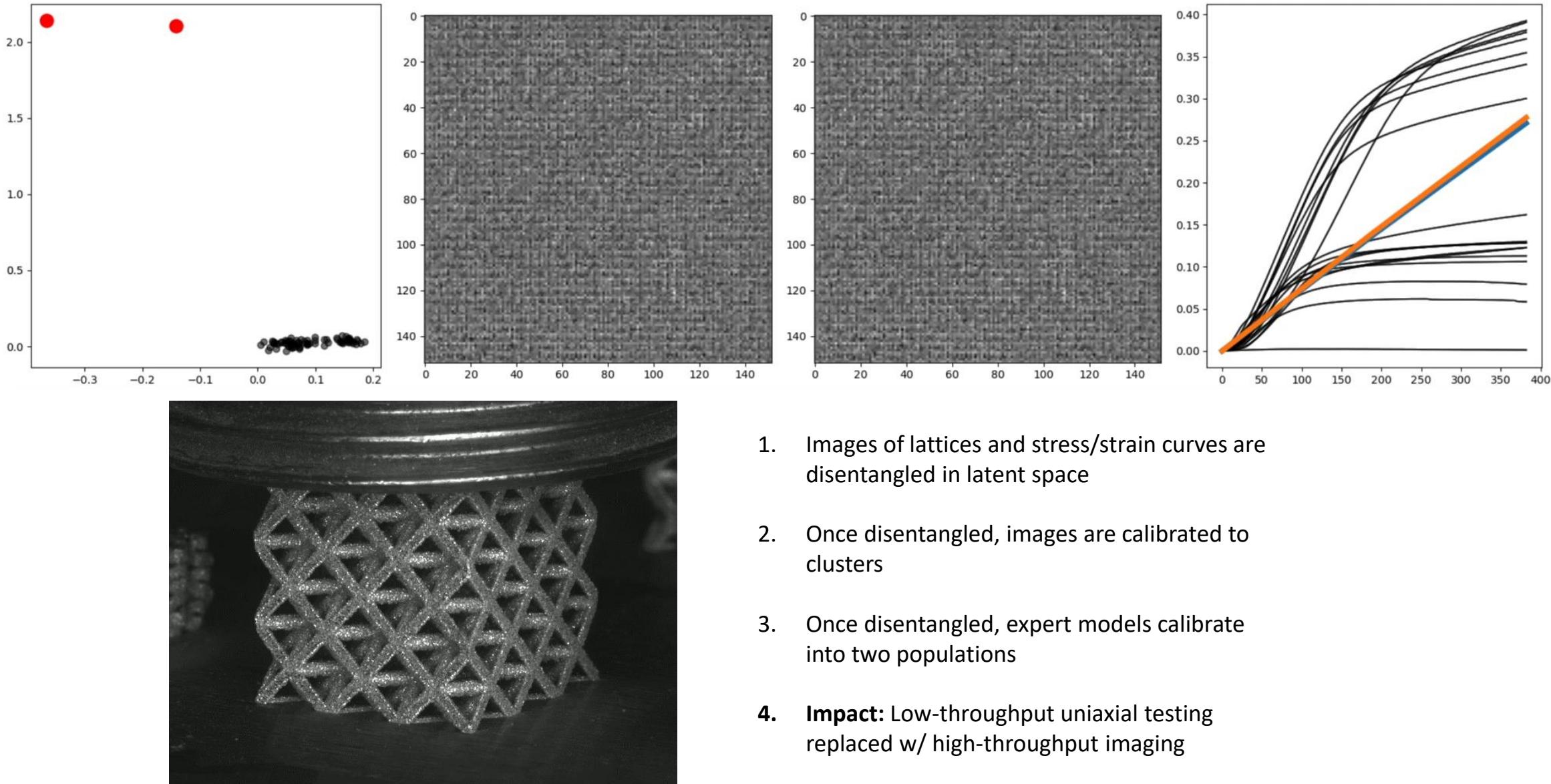
For historical lattice compression dataset:

Modality 1: Images of microstructure prior to deformation

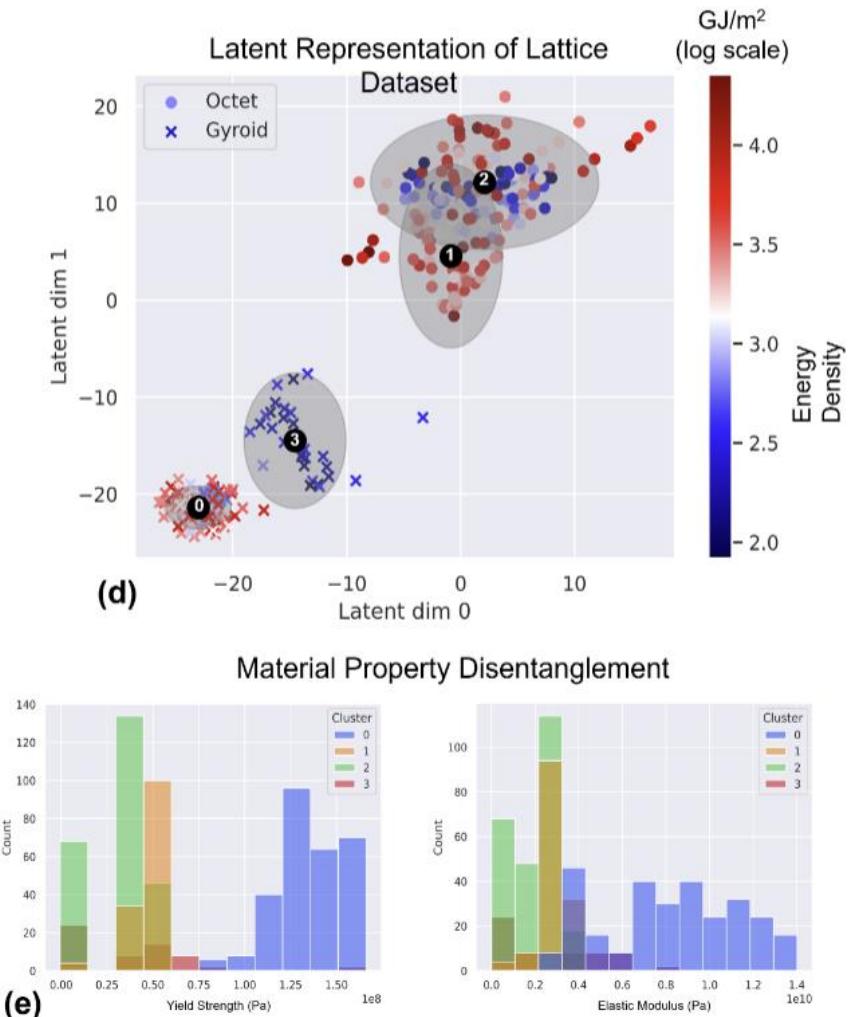
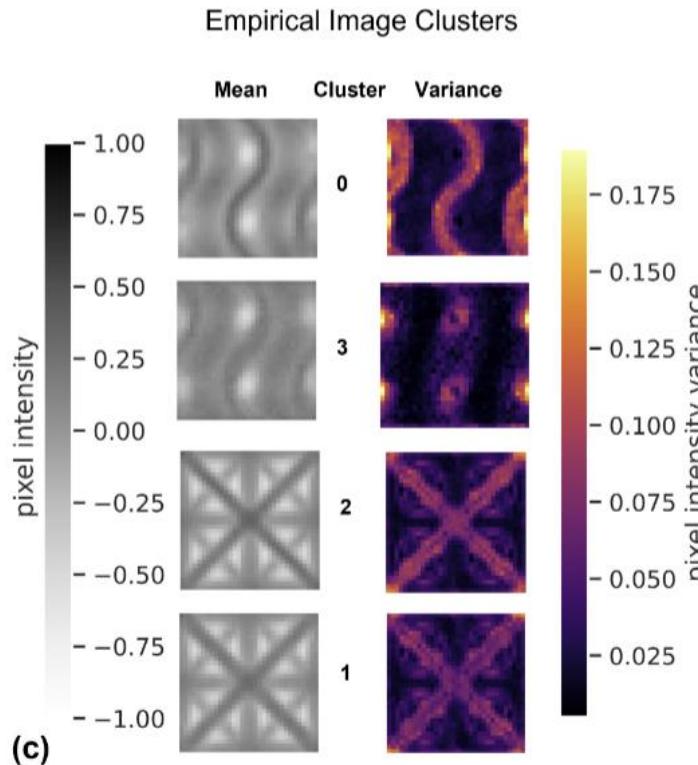
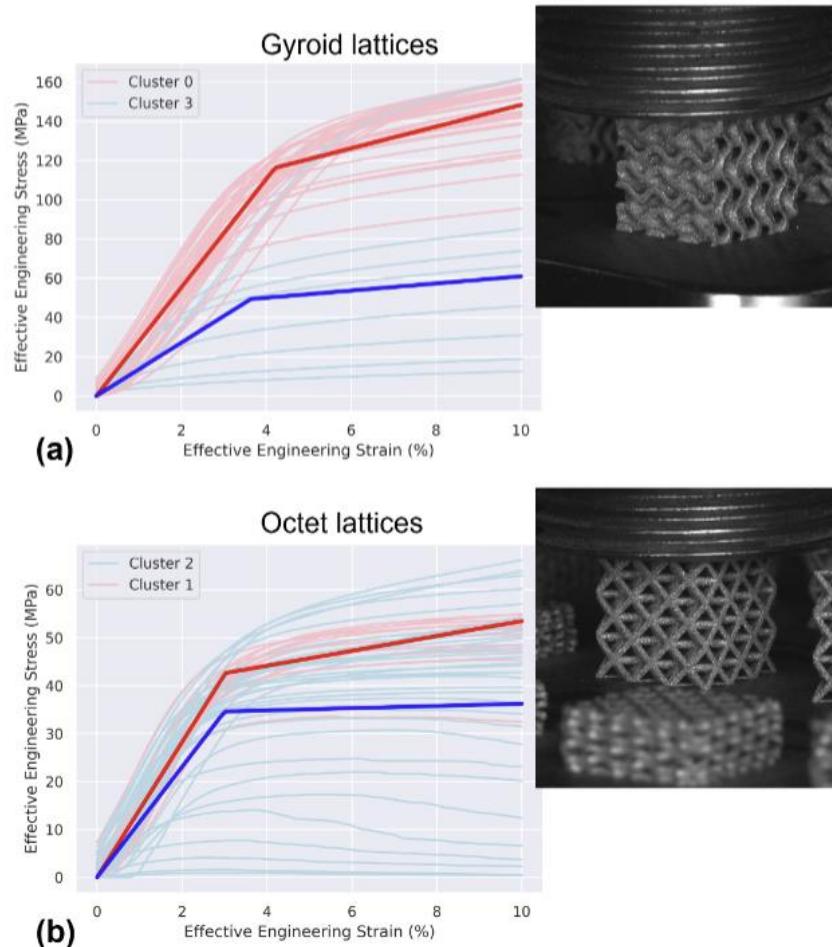
Modality 2: Uniaxial stress strain curve

**Expert physics model:** simple linear strain hardening model

## Example 1: Disentangling latent representations with physics



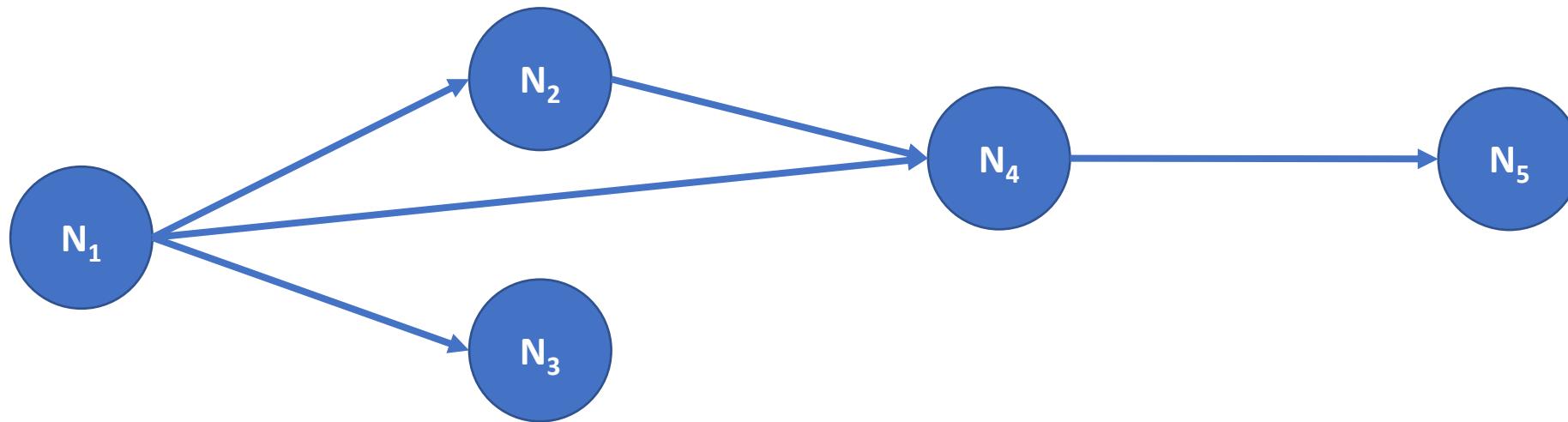
# Disentangling latent representations with physics



**Physical model reveals qualitative regimes of performance, other modalities facilitate hypothesis generation for explainability**

Trask, N., Martinez, C., Shilt, T., Walker, E., Lee, K., Garland, A., ... & Boyce, B. L. (2024). Unsupervised physics-informed disentanglement of multimodal materials data. *Materials Today*, 80, 286-296.

# Directed Acyclic Graphs encode causal relationships



## Why DAGs?

- Preclude cyclic reasoning
- Suggest dependency structure
- Have causal interpretation

## Structural Causal Decomposition

$$p(N) = \prod_{\ell} p(N_{\ell} | Pa(N_{\ell}))$$

Idea: replace categorical distribution Gaussian mixture prior with SCD

# Causality: can use DAG-structured priors to reveal causal relationships?

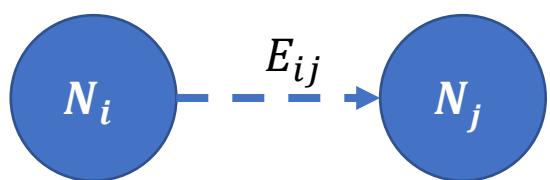
$$E_{ij} = \text{ReLU} \left( \tanh \frac{(B \cdot \mathcal{G}\xi)_{ij}}{\beta} \right)$$

$E$ : directed adjacency matrix.  $E_{ij} = p(N_i \rightarrow N_j)$

$B$ : trainable metric (nonnegative)

$\mathcal{G}$ : graph gradient operator

$\xi$ : trainable node scores



**Theorem:**  $\lim_{\beta \rightarrow 0} E$  defines a DAG and can recover any DAG.

*Proof sketch:*  $\text{Curl} \circ \text{Grad} = 0$ , so  $E$  is curl-free and induces a DAG. Graph gradient recovers complete DAGs, and  $B$  allows recovery for any sub-DAG.

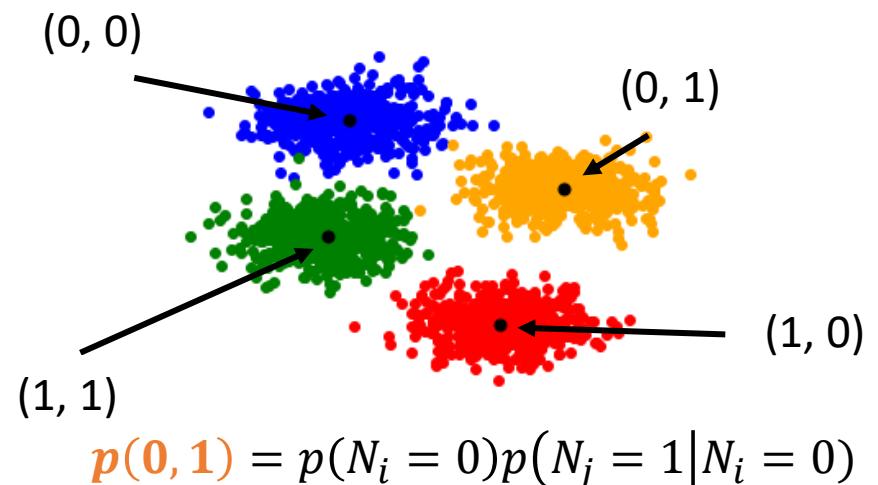
$$\text{ELBO} = \mathbb{E}_{q(Z,C|X)} \left[ \log \frac{p(X, Z, C)}{q(Z, C|X)} \right]$$

$X$ : input data

$Z$ : latent space

$C$ : Categorical for Gaussian Mixture

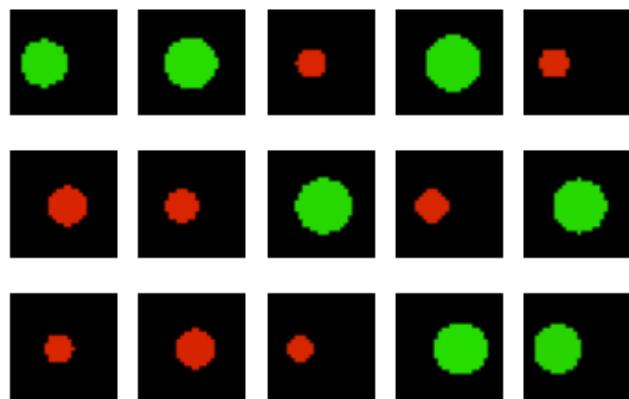
$$p(C) = p(N) = \prod_{\ell} p(N_{\ell} | Pa(N_{\ell}))$$



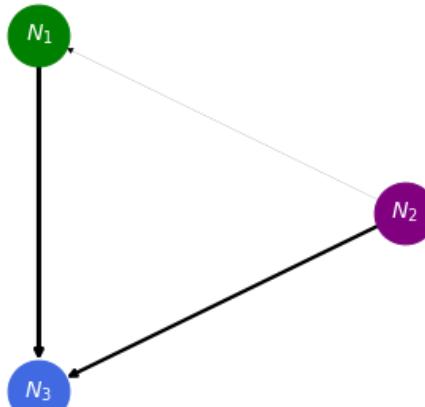
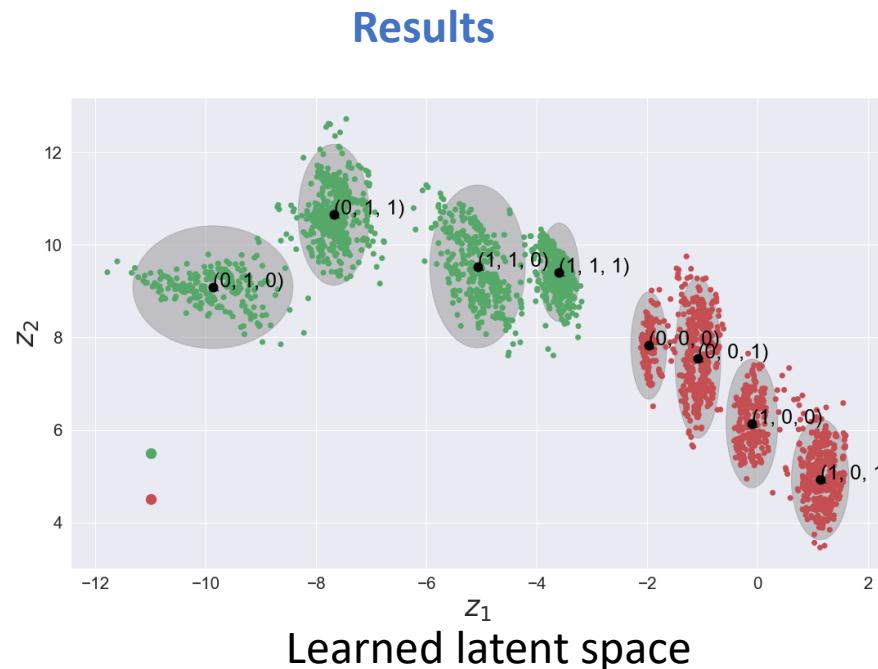
# Causality: some synthetic toy results

## Synthetic Circles Dataset

features: color, size, and shift



Obtain an embedding that shows that circle size (N1) and circle color (N2) determine the position of the circle (N3), without establishing a priori that N1-3 are features



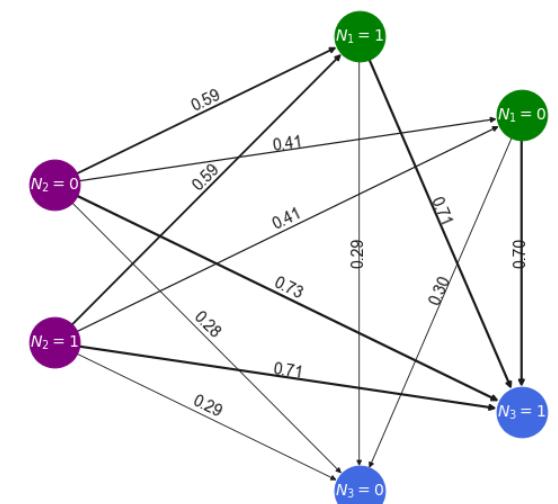
## Analysis

$N_1$ : Size

$N_2$ : Color

$N_3$ : Shift

$$p(N) = p(N_2)p(N_1|N_2)p(N_3|N_1, N_2)$$

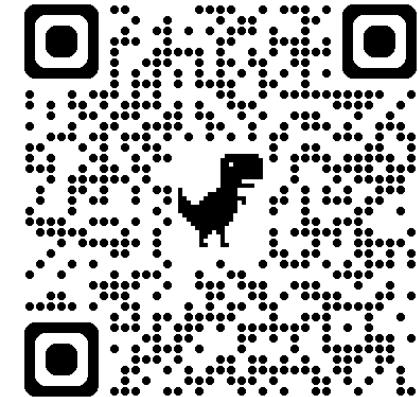


Edge weights are  
 $p(N_i = n | N_j = m)$

# Acknowledgements & References

## Structure preservation in digital twins

- Gruber, Anthony, Kookjin Lee, and Nathaniel Trask. "Reversible and irreversible bracket-based dynamics for deep graph neural networks." *Advances in Neural Information Processing Systems* 36 (2024).
- Jiang, S., Actor, J., Roberts, S. and Trask, N., 2024. A structure-preserving domain decomposition method for data-driven modeling. *arXiv preprint arXiv:2406.05571*. (accepted to special issue)
- Actor, J.A., Hu, X., Huang, A., Roberts, S.A. and Trask, N., 2024. Data-driven Whitney forms for structure-preserving control volume analysis. *Journal of Computational Physics*, 496, p.112520.
- Trask, N., Huang, A. and Hu, X., 2022. Enforcing exact physics in scientific machine learning: a data-driven exterior calculus on graphs. *Journal of Computational Physics*, 456, p.110969.



## Multimodal and causal scientific discovery

- Trask, N., Martinez, C., Shilt, T., Walker, E., Lee, K., Garland, A., Adams, D.P., Curry, J.F., Dugger, M.T., Larson, S.R. and Boyce, B.L., 2024. Unsupervised physics-informed disentanglement of multimodal materials data. *Materials Today*, 80, pp.286-296.
- Walker, Elise, Jonas A. Actor, Carianne Martinez, and Nathaniel Trask. "Causal disentanglement of multimodal data." *arXiv preprint arXiv:2310.18471* (2023).
- Walker, Elise, et al. "Flow-based parameterization for DAG and feature discovery in scientific multimodal data." *Frontiers in Mechanical Engineering* 10 (2024): 1408649.
- Walker, E., Trask, N., Martinez, C., Lee, K., Actor, J.A., Saha, S., Shilt, T., Vizoso, D., Dingreville, R. and Boyce, B.L., 2025. Unsupervised physics-informed disentanglement of multimodal data. *Foundations of Data Science*, 7(1), pp.418-445.
- Boyce, Brad, et al. *BeyondFingerprinting: AI-guided discovery of robust materials & processes*. No. SAND2024-11708. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2024.

## Collaborators

- **UPenn:** Brooks Kinch, Paris Perdikaris, Ben Schaffer, Ani Hsieh, Vijay Kumar, Spencer Folk
- **Tufts:** Xiaozhe Hu
- **SNL:** Andy Huang, Jonas Actor, Marshall Jiang, Anthony Gruber, Eric Cyr
- **PNNL:** Panos Stinis
- **PPPL:** Michael Churchill, Nate Ferraro
- **Caltech:** Houman Owhadi
- **Stanford:** Daniel Tartakovsky, Adrienne Propp
- **Boston University:** Emma Lejeune
- **USF:** Ivan Oleynik



U.S. DEPARTMENT OF  
**ENERGY**

Office of Science



**SEA-CROGS**

<https://www.pnnl.gov/projects/sea-crogs>