Amortized inverse modeling across tasks from incomplete, unpaired data

Conlain Kelly (conlain.kelly@nrel.gov), Juli Mueller (Juliane.Mueller@nrel.gov)
Computational Science Center, NREL

Topics 2-4: Probabilistic approaches; noisy, incomplete data; uncertainty-aware hybrid models

Challenges A great number of methods for solving inverse problems are predicated on either knowledge of the underlying forward model or availability of curated pairs of inputs and outputs for a system of interest. More concretely, we can interpret a forward model as a causal conditional distribution $p(y|x;\eta)$ relating parameters of interest x to observations y under varying conditions η (e.g., material feedstock, operating environments, performance constraints, or processing pipelines). In this setting, inverse modeling in DOE-relevant applications faces three fundamental barriers: task-specificity of existing models, limited access to complete or paired observations, and a lack of principled methods for fusing information across tasks and information sources.

For example, consider the task of designing and manufacturing materials with bespoke micro- and macro-scopic properties. For a given set of target properties one can run simulations and experiments to find an optimal set (or distribution) of manufacturing parameters w.r.t. the laboratory conditions. However, advancing through technology readiness levels **necessarily induces a shift in scale and associated operating conditions** η, resulting in deviations from the estimated performance. Most modern inversion techniques (e.g., MCMC, variational inference, regularized reconstruction) are inherently tailored to individual tasks η_i or observations y_i. Instead, **inverse models must be developed which generalize across varying operational regimes, as well as techniques for "fingerprinting" new problems and assessing similarity to known ones**.

Likewise, many input/output pairs are simply unobservable due to practical limitations (such as internal deformations within a deployed mechanical part) or safety considerations (such as experimental reactor responses in nearly-unstable regimes). When evaluating a novel manufacturing process, the destructive nature of sample acquisition (e.g., load testing, focused ion beam milling, surface indentation) prohibits us from capturing the material's complete performance profile with a single sample. **Efficient and reliable inference in these situations requires models that are capable of handling incomplete, corrupted, and multimodal observations.**

Finally, the material and labor costs of experimentation requires carefully selecting which measurements to collect and maximizing the information gained from each experiment. When only limited information is available about the true underlying forward model, surrogates can be used to interpolate observed/simulated responses y and impute missing observations. However, standard learning-based surrogates often degrade dramatically outside of their training distribution, especially in rare, extreme, or previously-unobserved regimes. This limits our ability to transfer/fine-tune previous surrogate models and necessitates a new set of *ab nihilo* experiments or high-fidelity simulations. We argue that a central challenge for solving modern inverse problems is efficiently querying the full **process manifold** — the joint space of inputs, outputs, conditions, and dynamics that define the complex system's behavior across tasks.

Opportunities Recent breakthroughs in generative deep learning [1, 2] now enable us to train probabilistic models on partially-paired data approximating either the posterior [3] or the joint distribution and arbitrary conditionals [4]. Likewise, incorporating physical domain knowledge via loss functions, encodings, and architectures has produced more stable and extrapolable surrogate models with reduced training data requirements [5]. Integrated strategically, these approaches could allow us to **infer joint probabilistic latent spaces** (for both parameters *and* observations) across tasks, as well as **task-specific stochastic linkages** [6]. However, deep generative surrogates still require large amounts of training data.

Long-term investments in experimental and HPC facilities have positioned DOE to generate large datasets for problems of interest. Using these resources efficiently necessitates developments in uncertainty-aware, multi-task batch active learning. Especially when the true forward model is unknown, balancing forward and inverse uncertainty is crucial to efficiently explore the manifold of $x \leftrightarrow y$ linkages (both forward [7] and inverse [8]) and refine predictions at critical scenarios of interest. Unifying the criteria of exploration, calibration, and information maximization in a sequential decision-making process would enable autonomous exploration of complex inverse linkages with **reduced experimental and simulation time, human labor, and capital costs**.

Finally, given the myriad simulation and learning-based surrogates available today, there is both opportunity and need to construct a *spectrum* of models for a given task. Rather than selecting simulators or experiments based on a hierarchy known or assumed fidelities, we propose treating different experiments as probabilistic projections of the underlying process manifold. This necessitates a means of estimating how much *information* each data source can provide for a given objective, motivating a probabilistic, Bayesian perspective [9]. This approach would allow for principled, real-time integration of diverse information sources – from high-fidelity simulation to low-cost experimental data – within a unified probabilistic framework. Fusing data sources with adaptive, data-efficient strategies could accelerate scientific discovery across DOE-relevant domains.

Innovation The described opportunities in amortized, multi-task inverse modeling leveraging several sources of incomplete and noisy data will allow us to significantly accelerate scientific discovery while minimizing the resources required to collect data (simulation or experiment) by maximizing information exploitation with impact across DOE offices, including materials discovery and manufacturing, bioreactors, and reconstruction in high-energy science. Exploring the full process manifold for such tasks would require a massive-scale collaboration across laboratories, as well as the knowledge-management infrastructure to support it. Once constructed, however, an *inverse foundation model* could enable a virtuous cycle wherein each new task requires progressively fewer measurements to perform actionable and reliable inference.

References

[1] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, February 2023. arXiv:2210.02747 [cs].

[2] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models, October 2022. arXiv:2206.00364 [cs].

[3] Adam P. Generale, Andreas E. Robertson, and Surya R. Kalidindi. Conditional Variable Flow Matching: Transforming Conditional Densities with Amortized Conditional Optimal Transport, April 2025. arXiv:2411.08314 [cs].

[4] Manuel Gloeckler, Michael Deistler, Christian Weilbach, Frank Wood, and Jakob H. Macke. All-in-one simulation-based inference, July 2024. arXiv:2404.09636 [cs].

[5] Conlain Kelly and Surya R. Kalidindi. Thermodynamically-Informed Iterative Neural Operators for heterogeneous elastic localization. *Computer Methods in Applied Mechanics and Engineering*, 441:117939, June 2025.

[6] Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, July 1997.

[7] Andreas Kirsch and Yarin Gal. Unifying Approaches in Active Learning and Active Sampling via Fisher Information and Information-Theoretic Quantities, November 2022. arXiv:2208.00549 [cs].

[8] Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational Bayesian Optimal Experimental Design, January 2020. arXiv:1903.05480 [stat].

[9] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, May 2010.