RESEARCH ARTICLE | FEBRUARY 10 2026

An information-matching approach to optimal experimental design and active learning FREE

Yonatan Kurniawan ; Tracianne B. Neilsen ; Benjamin L. Francis ; Alex M. Stankovic ; Mingjian Wen ; Ilia Nikiforov ; Ellad B. Tadmor ; Vasily V. Bulatov ; Vincenzo Lordi ; Mark K. Transtrum

Appl. Phys. Lett. 128, 064104 (2026)
https://doi.org/10.1063/5.0296026
CHORUS

Applied Physics Letters ARTICLE pubs.aip.org/aip/apl

An information-matching approach to optimal experimental design and active learning

Cite as: Appl. Phys. Lett. **128**, 064104 (2026); doi: 10.1063/5.0296026
Submitted: 12 August 2025 · Accepted: 19 January 2026 ·
Published Online: 10 February 2026

Yonatan Kurniawan,[1] Tracianne B. Neilsen,[1] Benjamin L. Francis,[2] Alex M. Stankovic,[3] Mingjian Wen,[4] Ilia Nikiforov,[5] Ellad B. Tadmor,[5] Vasily V. Bulatov,[6] Vincenzo Lordi,[6] and Mark K. Transtrum[1,2,3,a)]

AFFILIATIONS

[1]Brigham Young University, Provo, Utah 84602, USA
[2]Achilles Heel Technologies, Orem, Utah 84097, USA
[3]SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA
[4]University of Electronic Science and Technology of China, Chengdu 611731, China
[5]University of Minnesota, Minneapolis, Minnesota 55455, USA
[6]Lawrence Livermore National Laboratory, Livermore, California 94550, USA

[a)]**Author to whom correspondence should be addressed:** mktranstrum@byu.edu

ABSTRACT

The efficacy of mathematical models heavily depends on the quality of the training data, yet collecting sufficient data is often expensive and challenging. Many modeling applications require inferring parameters only as a means to predict other quantities of interest (QoI). Because models often contain many unidentifiable (sloppy) parameters, QoIs often depend on a relatively small number of parameter combinations. Therefore, we introduce an information-matching criterion based on the Fisher information matrix to select the most informative training data from a candidate pool. This method ensures that the selected data contain sufficient information to learn only those parameters that are needed to constrain downstream QoIs. It is formulated as a convex optimization problem, making it scalable to large models and datasets. We demonstrate the effectiveness of this approach across various modeling problems in diverse scientific fields, including power systems and underwater acoustics. Finally, we use information-matching as a query function within an active learning (AL) loop for materials science applications. In all these applications, we find that a relatively small set of optimal training data can provide the necessary information for achieving precise predictions. These results are encouraging for diverse future applications, particularly AL in large machine-learning models.

A model's predictive performance depends strongly on the quality and quantity of data available for training. Curating comprehensive datasets, however, often confronts practical constraints, including instrumentation, available resources, and cost. Optimal experimental design (OED)[1] and active learning (AL)[2] emerge as practical data collection strategies. Intentionally designing maximally informative experiments guarantees that data are most informative relative to the underlying phenomena of interest, minimizes costs, and meets operational requirements. These methodologies have broad applications across scientific domains, including sensor placement problems in power systems[3–8] and underwater acoustics,[9–13] the development of accurate interatomic potentials in materials science,[14–17] and many other scientific fields.[18–21]

Many OED criteria utilize the Fisher information matrix (FIM), whose inverse establishes a lower bound on parameter covariance, known as the Cramér–Rao bound.[22–24] Common approaches optimize parameter precision through the FIM, for example, by minimizing its trace (A-optimality),[25–27] maximizing its determinant (D-optimality),[16,27–29] or maximizing the smallest eigenvalues (E-optimality).[29–31]

However, many applications of predictive models do not require precise parameter estimates *per se* but accurate predictions for key quantities of interest (QoIs).[18,32–35] This distinction is well illustrated by *sloppy models*, where many parameter combinations are *practically unidentifiable* yet still yield precise predictions.[36–39] Such models typically have only a few stiff, identifiable directions, while the rest are sloppy and poorly constrained. Furthermore, the parameters that are identifiable from the training data may not align with those most relevant for predicting QoIs. Precise predictions are still possible when these QoI-relevant directions fall within the identifiable subspace; however, if they do not, the resulting uncertainty may be large or even divergent, regardless of how well some parameters are constrained.[40]

Furthermore, sloppy models exhibit a characteristic information spectrum, where FIM eigenvalues are nearly uniformly spaced on a log scale over many orders of magnitude. Many eigenvalues are smaller than the model's evaluation precision, rendering the aforementioned OED criteria sensitive to numerical noise.

Motivated by these considerations, we develop an information-matching method for OED that prioritizes the precision of predictions for the QoIs. This approach uses a model parameterized by $\boldsymbol{\theta}$ in two key scenarios: training, $\boldsymbol{f}$, and prediction, $\boldsymbol{g}$ (see Fig. 1). Given a dataset of M independent inputs $\{\mathbf{x}_m\}_{m=1}^M$ and their corresponding ground truth labels $\{\mathbf{p}_m\}_{m=1}^M$, we first use $\boldsymbol{f}(\boldsymbol{\theta};\mathbf{x}_m)$ to train the parameters against the data $\{\mathbf{x}_m,\mathbf{p}_m\}$. Then, we use the trained parameters to predict the QoIs $\mathbf{q}$ corresponding to the input $\mathbf{y}$ through the mapping $\boldsymbol{g}(\boldsymbol{\theta};\mathbf{y})$. In this scenario, $\mathbf{y}$ may act as an input, similar to $\mathbf{x}_m$, as a control parameter for the QoIs, or as a discrete index to distinguish between different QoI values. The information-matching method leverages both scenarios to identify the minimal subset of training data that contains the information necessary to precisely constrain the parameters relevant to QoIs. Given a target precision for QoIs, our strategy is to align the FIM for the training data with that of the QoIs, ensuring the training data carry the information needed to constrain the predictions precisely. Thus, only the parameter combinations that need to be identified are trained, bypassing numerical stability issues in cases where the FIMs are ill-conditioned.

The FIM is defined as the expectation value of the Hessian of the log-likelihood over the probability of the labels. For weighted least squares, by far the most common regression scenario, the negative log-likelihood is (up to an additive constant)

$$\ell(\boldsymbol{\theta}) = \frac{1}{2}\sum_{m=1}^{M} w_m ||\mathbf{p}_m - \boldsymbol{f}(\boldsymbol{\theta};\mathbf{x}_m)||_2^2, \tag{1}$$

where the weight w_m is the inverse variance of the label $\mathbf{p}_m$, i.e., $w_m = 1/\sigma_m^2$. As shown in the supplementary material, the FIM for the training scenario is given by

$$\mathcal{I}(\boldsymbol{\theta}) = \sum_{m=1}^{M} w_m \mathcal{I}_m(\boldsymbol{\theta}), \quad \mathcal{I}_m(\boldsymbol{\theta}) = J_f^T(\boldsymbol{\theta};\mathbf{x}_m) J_f(\boldsymbol{\theta};\mathbf{x}_m), \tag{2}$$

where $\mathcal{I}_m$ and $J_f(\boldsymbol{\theta};\mathbf{x}_m)$ are the FIM and the Jacobian matrix of $\boldsymbol{f}(\boldsymbol{\theta};\mathbf{x}_m)$ with respect to the parameters $\boldsymbol{\theta}$ corresponding to the mth datum, respectively.[36] This equation highlights a generic, fundamental property of the FIM: that the expected information in the entire training data is the sum of information from each independent datum.

Notably, the FIM denotes the expected information over the probability of the labels and does not depend on the observed label value $\mathbf{p}_m$. This property means the FIM can be evaluated for any model predictions, including downstream applications for which ground truth labels are not available, and accounts for its broad appeal in OED. The FIM then quantifies how much information about the model parameters is required to achieve a target precision. We denote the target precision of the QoIs $\mathbf{q}$ by the covariance matrix $\boldsymbol{\Sigma}$; the FIM for the QoIs is given as

$$\mathcal{J}(\boldsymbol{\theta}) = J_g^T(\boldsymbol{\theta})\boldsymbol{\Sigma}^{-1} J_g(\boldsymbol{\theta}), \tag{3}$$

where $J_g(\boldsymbol{\theta})$ is the Jacobian matrix of the proxy $\boldsymbol{g}(\boldsymbol{\theta};\mathbf{y})$ with respect to the parameters $\boldsymbol{\theta}$.

The information-matching method leverages the FIMs of both the training data and QoIs [Eqs. (2) and (3), respectively] to guarantee that the information in the training data is sufficient to achieve the target precision in the QoIs. To select the minimal set of training data that achieves this minimal information bound, we solve the following convex problem for the weight vector $\mathbf{w} = [\,w_1 \quad w_2 \quad \ldots \quad w_M\,]^T$:

$$\begin{aligned} &\text{minimize} && ||\mathbf{w}||_1 \\ &\text{subject to} && w_m \geq 0, \\ & && \mathcal{I} = \sum_{m=1}^{M} w_m \mathcal{I}_m \succcurlyeq \mathcal{J}. \end{aligned} \tag{4}$$

We conjecture that for many practical problems, the key information required for precise predictions is contained in a few key data points; therefore, we design the objective function to minimize the ℓ_1-norm of the weight vector to encourage sparse solutions. The nonzero weights identify the most important data and the precision with which the labels must be measured to ensure the target precision in the QoIs.

The matrix inequality constraint in Eq. (4) is crucial for ensuring the target precision on the QoIs. Formally, it means that the difference

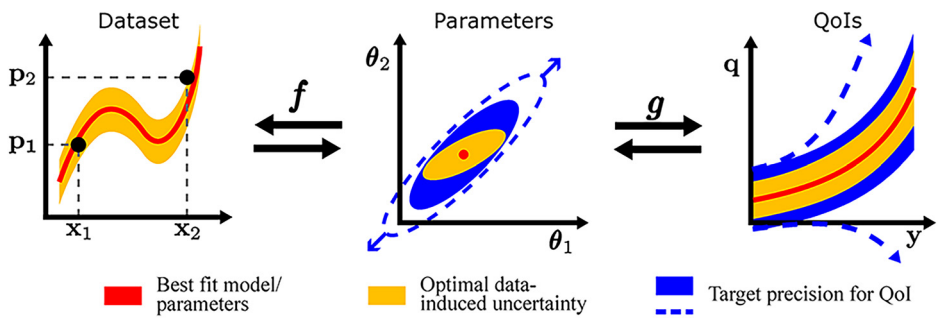

FIG. 1. Relationship between training data (left), model parameters (middle), and QoIs (right) in the information-matching framework. One first selects the target precision for the QoIs (blue envelope in the right panel). This QoI precision induces a minimal confidence region in parameter space (blue ellipse in the middle panel). The information-matching criterion selects training data and target precision (orange envelope in the left panel) such that the resulting parameter uncertainty (orange ellipse in the middle panel) is more restrictive than that induced by the QoIs. Propagating the parameter uncertainty to the QoIs gives predictions that are at least as precise as the original target (orange envelope in the right panel). This relationship holds even if the target uncertainties were divergent for certain QoIs (dashed blue curves in the right panel), resulting in the target parameter confidence diverging for some parameter combinations (dashed blue ellipse in the middle panel, extending in some directions).

$\mathcal{I} - \mathcal{J}$ is positive semidefinite. Intuitively, it indicates that fitting the down-selected data results in smaller parameter variance compared to fitting the QoIs directly, as illustrated in Fig. 1. Theorem 1 formalizes this statement (proof in the supplementary material).

Theorem 1. *Let $\boldsymbol{g}(\boldsymbol{\theta}; \mathbf{y})$ denote a mapping from the model parameters $\boldsymbol{\theta}$ to the QoIs for input $\mathbf{y}$ that is analytic at $\boldsymbol{\theta}_0 = \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\theta}}$, where $\langle \cdot \rangle_{\boldsymbol{\theta}}$ denotes an expectation value over the distribution of parameters. Consider parameters of the form $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \epsilon \delta \boldsymbol{\theta}$. If the constraints in Eq. (4) are satisfied, then*

$$\mathrm{Cov}(\boldsymbol{g}) \preceq \boldsymbol{\Sigma} + \mathcal{O}(\epsilon^3), \tag{5}$$

where $\boldsymbol{\Sigma}$ is the target covariance of the QoIs.

Theorem 1 states that the uncertainties of the QoIs propagated from the optimal training data [$\mathrm{Cov}(\boldsymbol{g})$] are within the predefined target uncertainties ($\boldsymbol{\Sigma}$), up to third order in ϵ. The information-matching method is unique in that it simultaneously aims to minimize data usage while ensuring adequate information for precise predictions. In contrast, the previously mentioned OED criteria (A-, D-, and E-optimality) only prioritize reducing some measure of parameter variance. By minimizing the number of data points, the information-matching approach not only enhances efficiency but also improves model interpretability by focusing the analysis on only the most critical training data.

We first demonstrate the information-matching method for optimally placing sensors (Phasor Measurement Unit, or PMU) in a power system network. The goal is to use a few strategically placed PMUs to infer the complex-valued voltages (system states) at every bus. A PMU placed on a bus measures the bus voltage and currents in adjoining branches, synchronized with GPS time stamps. By measuring voltages and currents, it is possible to achieve full-state observability without requiring PMUs at every bus.

We use the IEEE 39-bus system,[41] represented graphically in Fig. 2, as a benchmark. Nodes denote buses, and edges represent transmission lines and transformers.[3] The parameters $\boldsymbol{\theta}$ to be inferred are the system states (voltage magnitude and angle), and $\boldsymbol{f}$ represents PMU measurements. Full parameter identifiability implies that the QoIs are the state variables themselves [$\boldsymbol{g}(\boldsymbol{\theta}; \mathbf{y}) = \boldsymbol{\theta}$] and the QoI FIM is nonsingular. We enforce this condition by setting $\mathcal{J} = \lambda I$ for some small $\lambda > 0$. The information-matching condition implies that $\mathcal{I}$ is also nonsingular, and all states are observable. The optimal PMU placements for this problem are represented as the orange highlighted buses in Fig. 2. On this initial benchmark test, the information-matching method naturally selects the same buses identified in previous studies,[6,7] even without preassigning PMU locations. Furthermore, this result is broadly consistent with prior findings.[3–5]

In practice, the power system analysts model reduced portions of the full network, focusing on regions under their direct control. However, these areas are influenced by states outside the target areas, and identifying an appropriate reduced area equivalent is a challenging task.[42,43] We next partition the IEEE 39-bus system into non-overlapping regions (indicated by red, green, and blue in Fig. 2) and seek a minimal set of sensors to achieve observability within each area,[44] without regard for states outside it. We implement this by setting the diagonal elements of $\mathcal{J}$ corresponding to external states to zero, allowing infinite uncertainties for those states. The optimal PMU locations for each area are shown in Fig. 2 as buses are highlighted in different colors based on their respective areas. Notably, there are

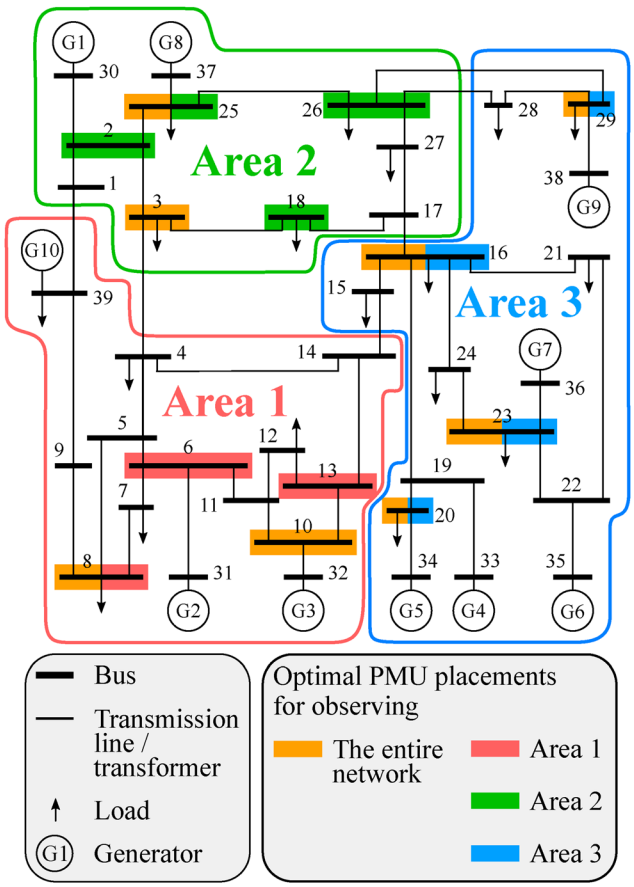

FIG. 2. The IEEE 39-bus system. Buses are represented by thick black lines, while transmission lines and transformers are shown as thin lines connecting them. Loads are depicted as black arrows pointing outward from the buses, and the circled labels G1 through G10 indicate the generators. Buses highlighted in orange denote the optimal PMU placements for full observability of the entire network. Buses highlighted in other colors (red, green, and blue) represent the optimal PMU placements for partial observability in the corresponding area. Many buses are double-highlighted with orange and another color, showing overlaps between full and partial observability. Non-overlapping optimal buses result from unobserved branches.

overlaps (double-highlighted buses) between optimal PMU placements for observing the full network and (the union of) three subnetworks, while non-overlapping locations are a consequence of enforcing observability for each of the subnetworks separately.

Next, we consider an optimal sensor placement problem in passive acoustic source localization in the ocean. The objective is to determine the optimal sensor (sound receiver) locations to infer the location of sound sources in a shallow ocean. This problem is difficult because sound propagation depends in complicated ways on unknown properties of the ocean environment, including the water temperature and the sediments in the seabed. Traditionally, source localization requires first estimating the ocean environmental parameters, as is done in matched-field processing.[45] In contrast, information-matching allows us to learn only those combinations of acoustic parameters that

are necessary to infer the source location, avoiding full environmental inversion when it is not required.

We aim to localize two vertically separated sound sources at depths of 8 and 16 m (red speakers) within ±2.5 m vertically and ±100 m horizontally (Fig. 3). Candidate receivers are arranged in a rectangular grid (small dots), motivated by common practices of using vertical and horizontal line arrays for ocean sound measurements.[9–12,46] We use a range-independent normal mode model, called ORCA,[47] to simulate the sound propagation in the ocean and compute the transmission loss at 200 Hz for each candidate location illustrated as small dots in Fig. 3 (details in the supplementary material). The model parameters $\boldsymbol{\theta}$ for simulating sound propagation include the source and receiver locations, as well as the parameterization of the ocean environment. The ocean environment is modeled with 75 m-deep water above an ocean floor consisting of a sandy sediment layer on a half-space basement layer.

To localize the two sources using Eq. (4), the QoI FIM $\mathcal{J}$ is a diagonal matrix, where diagonal elements that correspond to source locations are set to their inverse target precision, and all other elements are zero, indicating that only the relevant environmental parameters for localizing the sources are constrained as needed. The optimal receiver locations are shown as the large dots in Fig. 3, which account for only 5% of the candidate locations.

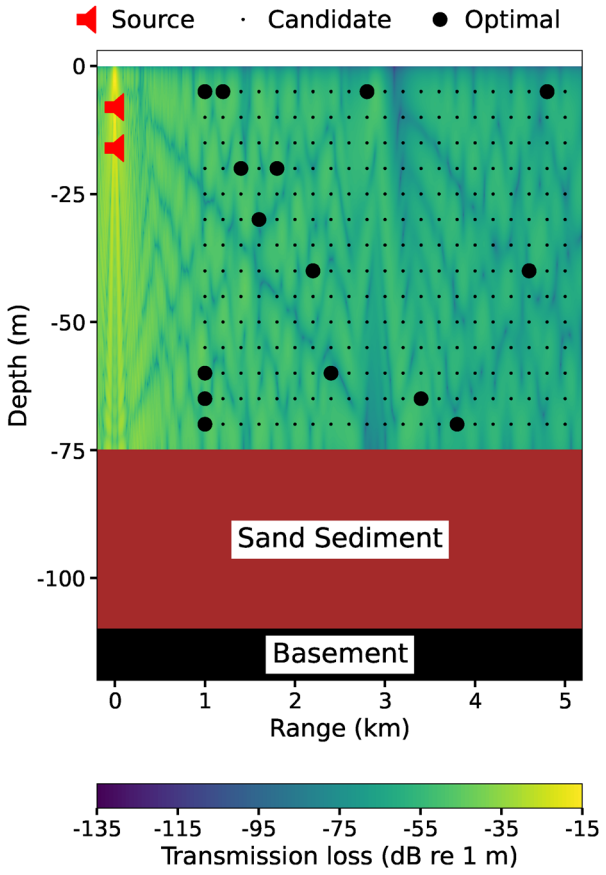

FIG. 3. Source localization in a shallow ocean. Optimal receiver locations for localizing two sound sources (red speakers) in a shallow ocean with a sandy seabed using transmission loss data at 200 Hz. Small dots indicate candidate sites; large dots are the optimal receiver locations.

The FIM is a local quantity that can vary (sometimes significantly) for different parameter values. In the preceding examples, we assumed a reasonable prior estimate of the models' parameters (e.g., the bus voltage phasor and the ocean environmental parameters), eliminating the need to recalculate the FIM after fitting with the optimal data. However, parameters often exhibit significant variability in response to new data, so we need to optimize the parameters alongside the data. To address this, we extend the OED problem to an active learning (AL) strategy and use Eq. (4) as a data query function.

ALGORITHM 1 Active learning via information-matching.

Initialize:
  $\mathbf{X} \leftarrow \{\mathbf{x}_m\}_{m=1}^{M}$ ▷ Candidate input data
  $\mathbf{P} \leftarrow \text{empty}(M)$ ▷ To store labels
  $\mathbf{w}^{\text{opt}} \leftarrow \text{zeros}(M)$ ▷ To store optimal weights
  $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}$ ▷ Initial parameters
while True **do**
  Compute $\mathcal{J}(\boldsymbol{\theta}_0)$ using Eq. (3)
  for $m = 1 : M$ **do**
  Compute $\mathcal{I}_m(\boldsymbol{\theta}_0)$ using Eq. (2)
  end for
  $\mathbf{w} \leftarrow$ Solve Eq. (4)
  $\mathbf{w}^{\text{opt}} \leftarrow \{\max(w_m^{\text{opt}}, w_m), \forall m = 1 : M\}$
  if $\mathbf{w}^{\text{opt}}$ converge **then**
  break
  else
  for all $\{m \mid w_m^{\text{opt}} > 0\}$ **do**
  $\mathbf{p}_m \leftarrow$ Generate label for $\mathbf{x}_m$
  end for
  $\boldsymbol{\theta}_0 \leftarrow \arg\min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ ▷ Update parameters
  end if
end while

Algorithm 1 outlines the iterative AL process based on information-matching. We begin by preparing a pool of candidate inputs $\mathbf{X} = \{\mathbf{x}_m\}_{m=1}^{M}$ and initializing the parameters to the *a priori* best estimate $\boldsymbol{\theta}_0$. The procedure starts by evaluating $\mathcal{J}$ and $\mathcal{I}_m$ at $\boldsymbol{\theta}_0$ for all m inputs and solving Eq. (4) for the weights $\mathbf{w}$. For each datum, we update the optimal weight by comparing the new optimal value with the current one, retaining the larger one. This step ensures that the amount of information in the training data in subsequent iterations is non-decreasing. Convergence occurs when the change in optimal weights between subsequent iterations is below some chosen threshold. If not converged, we generate labels for data with nonzero weights and update the parameters by minimizing Eq. (1), then iterate.

As an example, we apply this AL algorithm to the development of interatomic potentials in materials science. Interatomic potentials are crucial in atomistic simulations, as they approximate the interaction energy between atoms.[48] These potentials are typically trained on small-scale quantities, e.g., energy and atomic forces obtained from computationally expensive first-principles calculations,[33,49] then used in larger-scale simulations to predict material properties. Despite the scale discrepancy between training and prediction, the dynamics of

atoms primarily depend on their local neighborhoods. Thus, our objective is to identify training data (atomic configurations) that are maximally informative about the atomic neighborhoods for precise material predictions.

We apply Algorithm 1 to develop an optimal 15-parameter Stillinger–Weber (SW) potential for molybdenum disulfide (MoS_2) to precisely predict the energy (E) as a function of lattice parameter (a) under uniform in-plane strain.[33,40,50,51] The predictions are shifted by E_c (the energy at the equilibrium lattice constant a_0) to align the minimum with the origin, effectively showing strain-induced energy changes. The candidate dataset comprises 2000 atomic configuration snapshots from an *ab initio* molecular dynamics trajectory at 750 K, each with 96 Mo and 192 S atoms.[33,52] While this dataset contains force labels, in practice, the candidate dataset does not need to include labels; Algorithm 1 generates them on demand for the optimal configurations. We choose the target precision to be 10% of the values predicted by the potential trained on the full dataset.[33,52] In Fig. 4, we compare the uncertainties of the QoI obtained from the optimal configurations (red envelope) with the target uncertainty (blue envelope). Our findings indicate that seven atomic configurations suffice to constrain the parameters and achieve the target precision.

While some details of the calculation—the selected configurations, optimal weights, and optimal parameters—depend on the choice of initial parameters, these differences do not affect the validity of the final estimated uncertainty. As long as the problem is feasible, the resulting prediction uncertainty estimated by the FIM is guaranteed to be within the target uncertainty. This dependence on initialization is examined in more detail in the supplementary material, where we provide numerical examples based on the SW MoS_2 model. Additional results for other SW potentials for silicon are also provided in the supplementary material.

In summary, we have introduced an information-matching method to identify a minimal set of informative data points to meet target precision requirements for downstream QoIs. We have demonstrated the versatility of the approach across diverse domains,

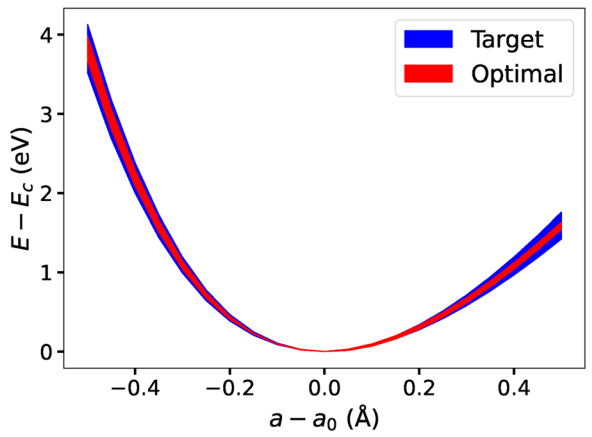

FIG. 4. Uncertainty in the energy (E) of monolayer MoS_2 vs in-plane lattice parameter (a). Predictions are shifted by the energy E_c at the equilibrium lattice constant a_0, aligning the minimum with the origin. The blue envelope is the target uncertainty (10% of the values predicted by the potential trained on the full dataset). In contrast, the red envelope shows the uncertainty propagated from the seven optimal training atomic configurations. Notice that the optimal propagated uncertainty is smaller than the target uncertainty.

including power system networks, underwater acoustics, and AL for interatomic potential development. Unlike classical FIM-based OED criteria that aim at reducing global parameter uncertainty, information-matching optimizes a fundamentally different objective: it matches the information content of the data to the precision required for the QoIs. As a result, only those parameters that influence the QoI are constrained—and only to the extent needed to satisfy the target QoI precision—while others may remain unconstrained if they are irrelevant. This distinction highlights that comparisons with traditional OED methods must be interpreted carefully, as the two frameworks prioritize different aspects of the inference problem and therefore tend to yield different designs. The accompanying theorem additionally formalizes this perspective and shows that QoI precision can be controlled without explicitly estimating the full parameter uncertainty. This approach provides opportunities in settings where full parameter identifiability is unnecessary or infeasible, such as PMU placement for partially observable networks or localization of ocean acoustic sources without environmental inversion. Broader applications are likely in areas such as biology, neuroscience, geology, and atmospheric science, where models often contain many weakly identifiable parameters but well-defined QoIs.[39]

Despite these advantages, successfully solving Eq. (4) depends on the richness of the initial candidate dataset; if any parameter that is relevant for the QoIs cannot be constrained by any of the candidate datasets, the optimization is infeasible. Because Eq. (4) is convex, algorithms will detect when a problem is infeasible. In such a case, it would be desirable to augment the candidate datasets with more informative candidates. Identifying such additional data points to achieve feasibility remains an open, problem-dependent question. Our current practical approach is to generate large and diverse candidate datasets. This can often be done quickly and cheaply. For example, in interatomic potential development, many atomic configurations with varied lattice parameters can be generated, while in sensor placement problems, one may define a fine grid of potential sensor locations. Because Eq. (4) is convex and scales favorably with problem size, this approach is practical in many scenarios.

Future work may extend the method to larger models and machine-learning applications, including machine-learned interatomic potentials. Investigating the feasibility and potential advantages of such integration could unlock broader applications and insights. Finally, a deeper theoretical analysis of the limiting behavior of the optimization may offer valuable guidance on robustness and further broaden applicability across diverse scientific domains.

See the supplementary material for the mathematical formulation of the least squares problem and the associated uncertainty quantification framework. It also includes a proof of the information-matching theorem, detailed descriptions of the models used in the example test cases, and additional results for those models.

Y.K., M.K.T., I.N., E.B.T., and A.M.S. acknowledge partial support through NSF under Grants Nos. 1834251, 1834332, and 2223986. Y.K., I.N., E.B.T., V.V.B., V.L., and M.K.T. acknowledge funding support from the Laboratory Directed Research and Development program (project code 23-SI-006) and a special computational time allocation on the Lassen supercomputer from the Computational Grand Challenge program at Lawrence

Livermore National Laboratory. Portions of this work were performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344. T.B.N. and M.K.T. acknowledge that this work is related to the Department of the Navy Award N00014-24-12566. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research. Some of the calculations were conducted using computational facilities provided by the Brigham Young University Office of Research Computing.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Yonatan Kurniawan: Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Writing – original draft (equal). **Tracianne B. Neilsen:** Data curation (equal); Writing – original draft (equal). **Benjamin L. Francis:** Data curation (equal); Writing – original draft (equal). **Alex M. Stankovic:** Writing – review & editing (equal). **Mingjian Wen:** Data curation (equal); Writing – review & editing (equal). **Ilia Nikiforov:** Writing – review & editing (supporting). **Ellad B. Tadmor:** Writing – review & editing (equal). **Vasily V. Bulatov:** Writing – review & editing (equal). **Vincenzo Lordi:** Writing – review & editing (equal). **Mark K. Transtrum:** Conceptualization (equal); Formal analysis (equal); Methodology (equal); Writing – original draft (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in Figshare at http://doi.org/10.6084/m9.figshare.27274029.v1, Ref. 53 and GitHub at https://github.com/yonatank93/information-matching, Ref. 54.

REFERENCES

[1]R. Leardi, "Experimental design in chemistry: A tutorial," Anal. Chim. Acta **652**(1), 161–172 (2009).
[2]A. Alizadeh, P. Tavallali, M. R. Khosravi, and M. Singhal, "Survey on recent active learning methods for deep learning," in *Advances in Parallel & Distributed Processing, and Applications*, edited by H. R. Arabnia, L. Deligiannidis, M. R. Grimaila, D. D. Hodson, K. Joe, M. Sekijima, and F. G. Tinetti (Springer International Publishing, Cham, 2021), pp. 609–617.
[3]W. Yuill, A. Edwards, S. Chowdhury, and S. P. Chowdhury, "Optimal PMU placement: A comprehensive literature review," in *2011 IEEE Power and Energy Society General Meeting, July 2011* (IEEE, 2011), pp. 1–8.
[4]J. Peng, Y. Sun, and H. F. Wang, "Optimal PMU placement for full network observability using Tabu search algorithm," Int. J. Electr. Power Energy Syst. **28**(4), 223–231 (2006).
[5]F. Aminifar, A. Khodaei, M. Fotuhi-Firuzabad, and M. Shahidehpour, "Contingency-constrained PMU placement in power networks," IEEE Trans. Power Syst. **25**(1), 516–523 (2010).
[6]B. Milosevic and M. Begovic, "Nondominated sorting genetic algorithm for optimal phasor measurement placement," IEEE Trans. Power Syst. **18**(1), 69–75 (2003).
[7]S. Chakrabarti and E. Kyriakides, "Optimal placement of phasor measurement units for power system observability," IEEE Trans. Power Syst. **23**(3), 1433–1440 (2008).
[8]F. Soudi and K. Tomsovic, "Optimal distribution protection design: Quality of solution and computational analysis," Int. J. Electr. Power Energy Syst. **21**(5), 327–335 (1999).
[9]D. A. Wood and D. J. Allwright, "Optimisation of hydrophone placement: A dynamical systems approach," Eur. J. Appl. Math. **14**(4), 369–386 (2003).
[10]S. E. Dosso and B. J. Sotirin, "Optimal array element localization," J. Acoust. Soc. Am. **106**(6), 3445–3459 (1999).
[11]M. Barlee, S. Dosso, and P. Schey, "Array element localization of a bottom moored hydrophone array," Can. Acoust. **30**(4), 3–14 (2002).
[12]S. E. Dosso and G. R. Ebbeson, "Array element localization accuracy and survey design," Can. Acoust. **34**(4), 3–13 (2006).
[13]M. D. Tidwell and J. R. Buck, "Designing linear FM active sonar waveforms for continuous line source transducers to maximize the Fisher information at a desired bearing," in *2019 Sensor Signal Processing for Defence Conference (SSPD), May 2019* (IEEE, 2019), pp. 1–5.
[14]G. Csányi, T. Albaret, M. C. Payne, and A. De Vita, "'Learn on the fly': A hybrid classical and quantum-mechanical molecular dynamics simulation," Phys. Rev. Lett. **93**(17), 175503 (2004).
[15]N. Artrith and J. Behler, "High-dimensional neural network potentials for metal surfaces: A prototype study for copper," Phys. Rev. B **85**(4), 045439 (2012).
[16]E. V. Podryabinkin and A. V. Shapeev, "Active learning of linearly parametrized interatomic potentials," Comput. Mater. Sci. **140**, 171–180 (2017).
[17]K. Gubaev, E. V. Podryabinkin, G. L. W. Hart, and A. V. Shapeev, "Accelerating high-throughput searches for new alloys with active learning of interatomic potentials," Comput. Mater. Sci. **156**, 148–156 (2019).
[18]F. P. Casey, R. N. Gutenkunst, C. R. Myers, D. Baird, K. S. Brown, J. J. Waterfall, Q. Feng, R. A. Cerione, and J. P. Sethna, "Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model," IET Syst. Biol. **1**(3), 190–202 (2007).
[19]J. E. Jeong, Q. Zhuang, M. K. Transtrum, E. Zhou, and P. Qiu, "Experimental design and model reduction in systems biology," Quant. Biol. **6**(4), 287–306 (2018).
[20]Z. Wang and A. Brenning, "Active-learning approaches for landslide mapping using support vector machines," Remote Sens. **13**(13), 2588 (2021).
[21]Y.-J. Kim and W.-T. Kim, "Uncertainty assessment-based active learning for reliable fire detection systems," IEEE Access **10**, 74722–74732 (2022).
[22]H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, 1999), Vol. 26.
[23]D. L. Streiner and G. R. Norman, "'Precision' and 'accuracy': Two terms that are neither," J. Clin. Epidemiol. **59**(4), 327–330 (2006).
[24]A. Van den Bos, *Parameter Estimation for Scientists and Engineers* (John Wiley & Sons, 2007).
[25]M. Jacroux, "The A-optimality of block designs for comparing test treatments with a control," J. Am. Stat. Assoc. **84**(405), 310–317 (1989).
[26]D. G. Butler, J. A. Eccleston, and B. R. Cullis, "On an approximate optimality criterion for the design of field experiments under spatial dependence," Aust. N. Z. J. Stat. **50**(4), 295–307 (2008).
[27]B. Jones, K. Allen-Moyer, and P. Goos, "A-optimal versus D-optimal design of screening experiments," J. Qual. Technol. **53**(4), 369–382 (2021).
[28]J. Andere-Rendon, D. C. Montgomery, and D. A. Rollier, "Design of mixture experiments using Bayesian D-optimality," J. Qual. Technol. **29**(4), 451–463 (1997).
[29]E. Balsa-Canto, A. Alonso, and J. Banga, "Computing optimal dynamic experiments for model calibration in predictive microbiology," J. Food Process Eng. **31**(2), 186–206 (2008).
[30]K. Leung Chow, "E-optimality for regression designs of supplementary experiments," J. Math. Anal. Appl. **214**(1), 207–218 (1997).
[31]J. P. Morgan and X. Wang, "E-optimality in treatment versus control experiments," J. Stat. Theory Pract. **5**(1), 99–107 (2011).
[32]M. K. Transtrum and P. Qiu, "Optimal experiment selection for parameter estimation in biological differential equation models," BMC Bioinf. **13**(1), 181 (2012).
[33]M. Wen, S. N. Shirodkar, P. Plecháč, E. Kaxiras, R. S. Elliott, and E. B. Tadmor, "A force-matching Stillinger-Weber potential for MoS_2: Parameterization and Fisher information theory based sensitivity analysis," J. Appl. Phys. **122**(24), 244301 (2017).

[34]G. Shmueli and O. R. Koppius, "Predictive analytics in information systems research," MIS Q. **35**(3), 553–572 (2011).
[35]F. Tavazza, B. DeCost, and K. Choudhary, "Uncertainty prediction for machine learning models of material properties," ACS Omega **6**(48), 32431–32440 (2021).
[36]M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Geometry of nonlinear least squares with applications to sloppy models and optimization," Phys. Rev. E **83**(3), 036701 (2011).
[37]M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna, "Perspective: Sloppiness and emergent theories in physics, biology, and beyond," J. Chem. Phys. **143**(1), 010901 (2015).
[38]A. F. Brouwer and M. C. Eisenberg, "The underlying connections between identifiability, active subspaces, and parameter space dimension reduction," arXiv:1802.05641 (2018).
[39]K. N. Quinn, M. C. Abbott, M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Information geometry for multiparameter models: New perspectives on the origin of simplicity," Rep. Prog. Phys. **86**(3), 035901 (2022).
[40]Y. Kurniawan, C. L. Petrie, K. J. Williams, M. K. Transtrum, E. B. Tadmor, R. S. Elliott, D. S. Karls, and M. Wen, "Bayesian, frequentist, and information geometric approaches to parametric uncertainty quantification of classical empirical interatomic potentials," J. Chem. Phys. **156**(21), 214103 (2022).
[41]Illinois Center for a Smarter Electric Grid, see https://icseg.iti.illinois.edu/ieee-39-bus-system/ for "IEEE 39-Bus System" (accessed 6 August 2024).
[42]A. T. Sarić, M. T. Transtrum, and A. M. Stanković, "Data-driven dynamic equivalents for power system areas from boundary measurements," IEEE Trans. Power Syst. **34**(1), 360–370 (2018).
[43]C. Zhao, J. Zhao, C. Wu, X. Wang, F. Xue, and S. Lu, "Power grid partitioning based on functional community structure," IEEE Access **7**, 152624–152634 (2019).
[44]See https://matpower.org/docs/ref/matpower5.0/case39.html for "Power flow data for 39 bus New England system" (accessed June 25, 2024).
[45]A. Tolstoy, *Matched Field Processing for Underwater Acoustics* (World Scientific, 1993).
[46]J. R. Buck, "Information theoretic bounds on source localization performance," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings, August 2002* (IEEE, 2002), pp. 184–188.
[47]E. K. Westwood, C. T. Tindle, and N. R. Chapman, "A normal mode model for acousto-elastic ocean environments," J. Acoust. Soc. Am. **100**(6), 3631–3645 (1996).
[48]E. B. Tadmor and R. E. Miller, *Modeling Materials: Continuum, Atomistic and Multiscale Techniques* (Cambridge University Press, 2011).
[49]F. Ercolessi and J. B. Adams, "Interatomic potentials from first-principles calculations: The force-matching method," Europhys. Lett. **26**(8), 583 (1994).
[50]M. Wen, *Stillinger-Weber Model Driver for Monolayer MX2 Systems v001* (OpenKIM, 2018).
[51]M. Wen, *Modified Stillinger-Weber Potential (MX2) for Monolayer MoS_2 Developed by Wen et al. (2017) v001* (OpenKIM, 2018).
[52]M. Wen, *Dataset of MoS_2 Monolayer from AIMD Trajectory* (Zenodo, 2024).
[53]Y. Kurniawan, see http://doi.org/10.6084/m9.figshare.27274029.v1 for "Information_matching_examples_dataset.tar.gz" (2024).
[54]Y. Kurniawan, see https://github.com/yonatank93/information-matching for "Information-matching Github Repository" (2024).