High-dimensional maximum-entropy phase space tomography using normalizing flows

Austin Hoover*
Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, USA

Jonathan C. Wong
Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou 730000, China
(Dated: July 24, 2024)

Particle accelerators generate charged particle beams with tailored distributions in six-dimensional position-momentum space (phase space). Knowledge of the phase space distribution enables model-based beam optimization and control. In the absence of direct measurements, the distribution must be tomographically reconstructed from its projections. In this paper, we highlight that such problems can be severely underdetermined and that entropy maximization is the most conservative solution strategy. We leverage *normalizing flows*—invertible generative models—to extend maximum-entropy tomography to six-dimensional phase space and perform numerical experiments to validate the model's performance. Our numerical experiments demonstrate consistency with exact two-dimensional maximum-entropy solutions and the ability to fit complicated six-dimensional distributions to large measurement sets in reasonable time.

I. INTRODUCTION

Particle accelerators generate charged particle beams with tailored distributions in position-momentum space (phase space). Measuring the phase space distribution in the accelerator enables model-based beam optimization and control and provides a valuable benchmark for simulation codes. In the absence of direct measurements [1–3], the distribution must be reconstructed from its projections.[1] Fig. 1 illustrates a generic setup in which the beam is measured under varying accelerator conditions and reconstructed at a location upstream of the measurement device.

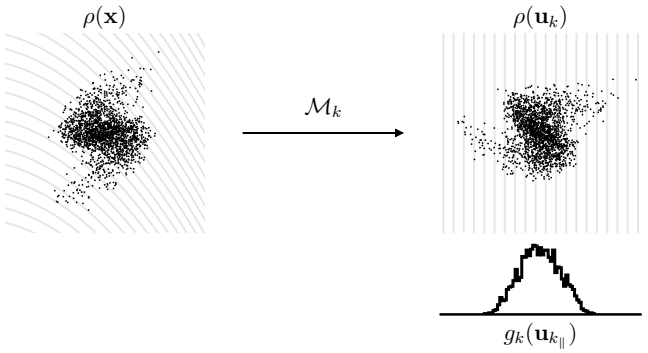

FIG. 1. Generic phase space tomography setup. An initial phase space distribution $\rho(\mathbf{x})$ travels through an accelerator segment represented by the symplectic transformation $\mathbf{u}_k = \mathcal{M}_k(\mathbf{x})$ for measurement index k. Each projection $g_k(\mathbf{u}_{k\parallel})$ of the transformed distribution is a different low-dimensional view of the initial distribution.

If the accelerator linearly transforms the phase space coordinates and does not couple the three planes of motion, one can reconstruct the 2D phase space distribution using conventional tomography algorithms. It is more challenging to reconstruct the 4D or 6D phase space distribution. Many conventional algorithms represent the distribution on a grid and face massive storage requirements as the phase space dimension scales [5]. Several authors have developed new algorithms and diagnostics to sidestep this issue and fit 4D phase space distributions to 2D projections [5–10]. There has also been one extension to 5D phase space [11], and one extension to 6D phase space [12].

An additional challenge is that high-dimensional reconstructions may be ill-posed; since the measured dimension is fixed, the set of feasible distributions (those consistent with the measurements) may proliferate with the phase space dimension. It is usually infeasible to compensate by exponentially increasing the number of measurements, as one is typically limited to tens of views because of slow diagnostic devices and limited beam time. Additionally, it is not yet clear how to derive the information-maximizing set of high-dimensional phase space transformations under given measurement conditions—and in any case, accelerator constraints place many transformations out of reach.

To select a single solution from the feasible set, our strategy is to define a prior probability distribution over the phase space coordinates and update the prior to a posterior by incorporating the information in the measurements. Our information comes in the form of constraints, and we perform the update by maximizing a convex functional subject to these constraints. Under basic self-consistency requirements, the functional must be the relative entropy [13–16]. Entropy maximization ensures that the posterior does not deviate from the prior

* hooveram@ornl.gov

[1] The 1D beam density can be measured by recording the secondary electron emission from a wire swept across the beam. Scintillating screens provide 2D projections of electron beams or low-intensity, low-energy hadron beams. 2D projections of higher energy hadron beams are only available from specialized diagnostics such as laser wires [4].

unless forced to by the data. This is a conservative strategy that eliminates all spurious features from the reconstructed distribution.

Entropy maximization is not always feasible, especially in high dimensions, because it entails a highly nonlinear constrained optimization. Although a reliable exact maximum-entropy algorithm exists for 2D tomography, its computational complexity scales exponentially with the phase space dimension, rendering its extension to 6D prohibitively expensive at this time. In this paper, we leverage *normalizing flows*—invertible generative models—to find approximate 6D maximum-entropy solutions. Our approach is a straightforward extension of two previous studies. Loaiza-Ganem, Gao, and Cunningham [17] first proposed the use of normalizing flows for entropy maximization subject to statistical moment constraints; we incorporate projection constraints using the differentiable physics simulations and projected density estimation proposed by Roussel et al. [10] in the Generative Phase Space Reconstruction (GPSR) framework. We refer to the resulting approach as MENT-Flow.

We begin by deriving the form of the n-dimensional maximum-entropy distribution subject to m-dimensional projection constraints, following the analysis in [18]. We then discuss the shortcomings of existing maximum-entropy tomography algorithms when $n = 6$ and describe the flow-based solution. Finally, we perform numerical experiments to validate the model's reliability in 2D settings and examine the effects of entropic regularization in 6D tomography.

II. MAXIMUM ENTROPY TOMOGRAPHY

Let $\rho_*(\mathbf{x})$ be a prior probability distribution over the phase space coordinates $\mathbf{x} \in \mathbb{R}^n$. We wish to update the prior to a posterior $\rho(\mathbf{x})$ by maximizing the relative entropy

$$H[\rho(\mathbf{x}), \rho_*(\mathbf{x})] = -\int \rho(\mathbf{x}) \log\left(\frac{\rho(\mathbf{x})}{\rho_*(\mathbf{x})}\right) d\mathbf{x} \tag{1}$$

while enforcing consistency with a set of m-dimensional projections. We will refer to this problem as an n:m reconstruction.

We assume the kth measurement occurs after a symplectic transformation $\mathcal{M}_k : \mathbb{R}^n \to \mathbb{R}^n$. By splitting the transformed coordinates

$$\mathbf{u}_k = \mathcal{M}_k(\mathbf{x}) \tag{2}$$

into a projection axis $\mathbf{u}_{k_\parallel} \in \mathbb{R}^m$ and orthogonal integration axis $\mathbf{u}_{k_\perp} \in \mathbb{R}^{n-m}$, we can write the constraints as

$$G_k\left[\rho(\mathbf{x})\right] = g_k(\mathbf{u}_{k_\parallel}) - \tilde{g}_k(\mathbf{u}_{k_\parallel}) = 0, \tag{3}$$

where $g_k(\mathbf{u}_{k_\parallel})$ are the *measured* projections and

$$\tilde{g}_k(\mathbf{u}_{k_\parallel}) = \int \rho(\mathbf{x}(\mathbf{u}_k)) d\mathbf{u}_{k_\perp} \tag{4}$$

are the *simulated* projections. The form of the maximum-entropy posterior distribution can be derived from a new functional

$$\Psi = H[\rho(\mathbf{x}), \rho_*(\mathbf{x})] + \sum_k \int \lambda_k(\mathbf{u}_{k_\parallel}) G_k\left[\rho(\mathbf{x})\right] d\mathbf{u}_{k_\parallel}, \tag{5}$$

where $\lambda_k(\mathbf{u}_{k_\parallel})$ are Lagrange multipliers [19]. Enforcing zero variation of Ψ with respect to $\rho(\mathbf{x})$ and $\lambda_k(\mathbf{u}_{k_\parallel})$ gives

$$\begin{aligned}\rho(\mathbf{x}) &= \rho_*(\mathbf{x}) \prod_k \exp\left(\lambda_k(\mathbf{u}_{k_\parallel}(\mathbf{x}))\right) \\ &= \rho_*(\mathbf{x}) \prod_k h_k(\mathbf{u}_{k_\parallel}(\mathbf{x})).\end{aligned} \tag{6}$$

where we have defined $h_k(\mathbf{u}_k) = \exp(\lambda_k(\mathbf{u}_k))$. Substituting Eq. (6) into Eq. (3) generates a set of coupled nonlinear integral equations from which h_k are to be solved.

A. MENT

The MENT algorithm [18–21] leverages a Gauss-Seidel relaxation method to optimize the Lagrange functions in Eq. (6). After initializing the distribution to the prior within the measurement boundaries:

$$h_k(\mathbf{u}_{k_\parallel}) = \begin{cases} 1, & \text{if } g_k(\mathbf{u}_{k_\parallel}) > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$

the Lagrange functions are updated as

$$h_k(\mathbf{u}_{k_\parallel}) \leftarrow h_k(\mathbf{u}_{k_\parallel}) \left(1 + \omega\left(\frac{g_k(\mathbf{u}_{k_\parallel})}{\tilde{g}_k(\mathbf{u}_{k_\parallel})} - 1\right)\right) \tag{8}$$

where

$$\tilde{g}_k(\mathbf{u}_{k_\parallel}) = \int \rho_*(\mathbf{x}(\mathbf{u}_k)) \prod_j h_j(\mathbf{u}_{j_\parallel}(\mathbf{u}_k)) d\mathbf{u}_{k_\perp} \tag{9}$$

are the simulated projections and $0 < \omega \leq 1$ is a learning rate [19]. The updates are performed in order ($k = 1, 2, 3, \ldots$), and each updated h_k is immediately used to simulate the next projection. One epoch is completed when all functions are updated. The iterations in Eq. (8) converge [20, 21].

MENT maximizes entropy by design: fitting the data generates an exact solution to the constrained optimization problem. MENT is also efficient: it stores the exact number of parameters needed to define the maximum-entropy distribution and typically converges in a few epochs. Finally, MENT is essentially free of hyperparameters.

The MENT formulation above is valid for n:m tomography, but the integrals in Eq. (9) limit the value of n in practice. Ongoing work aims to demonstrate efficient implementations when $n = 4$ [18, 22]. Extension to $n = 6$ may be possible, but it has yet to be demonstrated, and the runtime would likely be quite long if there were many high-resolution measurements. Even if the algorithm converged, sampling particles from the posterior (Eq. (6)) would be a nontrivial extra step.

B. MENT-Flow

In the absence of a method to directly optimize the Lagrange functions in Eq. (6), we may try to minimize the loss function

$$L = -H[\rho(\mathbf{x}), \rho_*(\mathbf{x})] + \mu \sum_k D[g_k(\mathbf{u}_{k_\parallel}), \tilde{g}_k(\mathbf{u}_{k_\parallel})] \quad (10)$$

for an increasing sequence of penalty parameters μ. Here, $D[g_k(\mathbf{u}_{k_\parallel}), \tilde{g}_k(\mathbf{u}_{k_\parallel})]$ is a non-negative number quantifying the discrepancy between the measured and simulated projections, which we choose to be the Kullback-Leibler (KL) divergence. Exact solutions may require $\mu \to \infty$, but approximate solutions obtained with finite μ are often sufficient.

The above approach requires us to represent the distribution using a finite set of parameters, $\boldsymbol{\theta}$. Grid-based representations become expensive when $n \geq 4$. An attractive alternative is to directly predict the value of $\rho(\mathbf{x})$ up to a normalization constant; however, computing the distribution's entropy (Eq. (1)) and projections (Eq. (4)) would require expensive integration or Monte Carlo sampling. We might instead define the distribution indirectly via the transformation

$$\mathbf{x} = \mathcal{F}(\mathbf{z}; \boldsymbol{\theta}), \quad (11)$$

where $\mathcal{F} : \mathbb{R}^{n'} \to \mathbb{R}^n$ is a map parameterized by $\boldsymbol{\theta}$, and $\mathbf{z} \in \mathbb{R}^{n'}$ is a random variable drawn from a base distribution $\rho_0(\mathbf{z})$ defined in a "normalized" or "latent" space. The base distribution is typically a Gaussian. Sampling from $\rho(\mathbf{x})$ reduces to sampling from $\rho_0(\mathbf{z})$ and applying the unnormalizing transformation in Eq. (11). Thus, Eq. (11) defines a *generative model*. In most generative models, $\mathcal{F}$ is a neural network trained to learn an unknown distribution from data samples [23].

Roussel et al. [10] showed that generative models can also be trained to match *projections* of the unknown distribution. To train the model via gradient descent, the transformations from the base distribution to the measurement locations must be differentiable:

$$\mathbf{u}_k = \mathcal{M}_k(\mathcal{F}(\mathbf{z}; \boldsymbol{\theta})). \quad (12)$$

This is possible using a differentiable beam physics simulation [24] to represent $\mathcal{M}_k$. The calculation of the projected density $\tilde{g}_k(\mathbf{u}_k)$ in Eq. (4) must also be differentiable. This is possible using 1D or 2D kernel density estimation. It is, however, difficult to maximize the entropy without access to the density $\rho(\mathbf{x})$.[2]

A *normalizing flow*, or simply *flow*, follows the same paradigm but provides access to the probability density. A normalizing flow is a differentiable map $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^n$ with a differentiable inverse $\mathcal{F}^{-1}$ [26]. These properties ensure we can compute the change in probability density under the transformation in Eq. (11):

$$\log \rho(\mathbf{x}) = \log \rho_0(\mathbf{z}) - \log |\det J_{\mathcal{F}}(\mathbf{z})|, \quad (13)$$

where

$$J_{\mathcal{F}}(\mathbf{z}) = \frac{d\mathcal{F}}{d\mathbf{z}} = \begin{bmatrix} \frac{\partial \mathcal{F}_1}{\partial z_1} & \cdots & \frac{\partial \mathcal{F}_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{F}_n}{\partial z_1} & \cdots & \frac{\partial \mathcal{F}_n}{\partial z_n} \end{bmatrix}. \quad (14)$$

is the $n \times n$ Jacobian matrix of $\mathcal{F}$, accounting for volume change, and $\mathbf{z} = [z_1, \ldots, z_n]^T$. To compute the probability density at $\mathbf{x}$, we flow backward and multiply the base distribution at $\mathbf{z}$ by the absolute value of the Jacobian matrix determinant. To generate samples $\{\mathbf{x}_i\}$, we sample points $\{\mathbf{z}_i\}$ from the base distribution and unnormalize them by flowing forward. We also obtain the probability density $\{\rho(\mathbf{x}_i)\}$ at each sampled point by tracking the Jacobian matrix determinant during this forward pass.

The ability to generate particles *and* evaluate the probability density at each particle is useful for computing expected values. Given N samples $\{\mathbf{x}_i\}$ from $\rho(\mathbf{x})$, the following expression is an unbiased estimate of the expected value of a functional $Q[\rho(\mathbf{x})]$:

$$\mathbb{E}\left[Q[\rho(\mathbf{x})]\right] = \int \rho(\mathbf{x}) Q[\rho(\mathbf{x})] d\mathbf{x} \approx \frac{1}{N} \sum_{i=1}^{N} Q[\rho(\mathbf{x}_i)]. \quad (15)$$

Since the entropy is the expected value of $\log(\rho(\mathbf{x})/\rho_*(\mathbf{x}))$, the following expression is an unbiased estimate of the entropy [17]:

$$H[\rho(\mathbf{x}), \rho_*(\mathbf{x})] \approx -\frac{1}{N} \sum_{i=1}^{N} \log(\rho(\mathbf{x}_i)/\rho_*(\mathbf{x}_i)) \quad (16)$$

Since the estimate is differentiable, it can be maximized via stochastic gradient descent [17]. Thus, our approach is to use the Generative Phase Space Reconstruction (GPSR) method [10] with a normalizing flow instead of a conventional neural network. We call this approach MENT-Flow, in reference to the MENT algorithm.

---

[2] Roussel et al. [10] proposed to maximize the *emittance*, or root-mean-square (rms) volume, $\varepsilon = |\boldsymbol{\Sigma}|^{1/2}$, where $\boldsymbol{\Sigma} = \langle \mathbf{x}\mathbf{x}^T \rangle$ is the $n \times n$ covariance matrix of second order moments, as a proxy for the entropy. For certain distributions, the logarithm of the emittance is proportional to the entropy, but this is not true in general. Like entropy maximization, emittance maximization removes unnecessary linear correlations from the reconstructed distribution. However, it cannot remove nonlinear correlations, as the emittance depends only on second-order moments. Furthermore, the maximum-emittance distribution is not unique. In linear systems, the covariance matrix is typically overdetermined by the tomographic measurements, i.e., all distributions that fit the data have the same emittance. Particle-based entropy estimates based on k nearest neighbors [25] may perform better.

It is not immediately obvious whether normalizing flows can learn complex 6D distributions from projections in reasonable time. Flows preserve the topological features of the base distribution; for example, flows cannot perfectly represent disconnected modes if the base distribution has a single mode [27]. Thus, building complex flows requires layering transformations, either as a series of maps (discrete flows) or a system of differential equations (continuous flows), often leading to large models and expensive training.[3]

We found that neural spline flows (NSF) [29] provide a sufficient blend of speed and power. In this model, the unnormalizing transformation $\mathcal{F}$ has the following autoregressive form:

$$x_i = \tau(z_i; c_i(z_1, \ldots, z_{i-1})), \tag{17}$$

where $\mathbf{x} = [x_1, \ldots, x_n]^T$, $\mathbf{z} = [z_1, \ldots, z_n]^T$, and τ is an invertible function parameterized by $c_i(z_1, \ldots, z_{i-1})$. The transformation is invertible for any c_i, and since c_i depends only on the first $i-1$ dimensions, the transformation has a triangular Jacobian matrix whose determinant can be computed efficiently. In the NSF model, the 1D transformer (τ) is a monotonic rational-quadratic spline [29]. The spline is defined by the locations of K different knots and the derivative at each knot. These parameters are provided by the conditioner (c), a masked neural network [30] in which connections between nodes in a regular feedforward neural network are removed to produce the triangular Jacobian matrix.

The model's representational power increases with the number of parameters in the masked neural network and the number of knots in the rational-quadratic splines. We can also define more than one flow layer. For the composition of T layers

$$\mathcal{F} = \mathcal{F}_T \circ \mathcal{F}_{T-1} \circ \cdots \circ \mathcal{F}_2 \circ \mathcal{F}_1, \tag{18}$$

and transformed coordinates

$$\mathbf{z}_t = \mathcal{F}_t(\mathbf{z}_{t-1}), \tag{19}$$

the Jacobian determinant is available from

$$|\det J_{\mathcal{F}}(\mathbf{z}_0)| = \prod_{t=1}^{T} |\det J_{\mathcal{F}_t}(\mathbf{z}_{t-1})|\,. \tag{20}$$

Compared to MENT, MENT-Flow increases the reconstruction model complexity and does not guarantee an exact entropy maximum. However, MENT-Flow scales straightforwardly to n-dimensional phase space and immediately generates independent and identically distributed samples from the reconstructed distribution function.

[3] A relevant example comes from Green, Ting, and Kamdar [28], who used continuous flows for 6D phase space density estimation from measured stellar phase space coordinates. Training times ranged from hours to days on a GPU, depending on the distribution complexity, with approximately 10^4 particles per batch.

III. NUMERICAL EXPERIMENTS

The following numerical experiments demonstrate that MENT-Flow solutions approach MENT solutions in 2D phase space. Subsequent experiments demonstrate that MENT-Flow can fit complicated 6D phase space distributions to large measurement sets in reasonable time and that entropic regularization keeps the reconstruction close to the prior. To simplify the examples, we focused on linear phase space transformations rather than more realistic accelerator models. We also tended to use ground-truth distributions without linear interplane correlations, highlighting nonlinear features.[4] We chose to maximize the entropy relative to a Gaussian prior.[5] The flow's base distribution is also a Gaussian, so the entropy penalty pushes the flow toward an identity or scaling transformation.

Our normalizing flow architecture is described in the previous section. The flow consists of five layers. Each layer is an autoregressive transformation, where the 1D transformation along each dimension is a rational-quadratic spline with 20 knots; the function values and derivatives at the knots are parameterized by a masked neural network with 3 hidden layers of 64 hidden units. Note that increasing the model size should *not* lead to overfitting since we train via maximum entropy, not maximum likelihood.

We compare MENT-Flow to MENT. Our MENT implementation uses linear interpolation to evaluate the Lagrange functions at any location on the projection axes, and we simulate the projections by numerical integration. We also compare to an unregularized neural network (NN) whose only aim is to fit the data. The NN is a standard fully connected feedforward network with 3 hidden layers of 32 hidden units and tanh activation functions.

We used 2×10^4 samples to estimate the entropy and projections. During each epoch, we trained the normalizing flow using 400 iterations of the Adam optimizer. After each epoch, we multiplied the penalty parameter

[4] Equivalently, we assume we know the covariance matrix $\mathbf{\Sigma} = \langle \mathbf{x}\mathbf{x}^T \rangle = \mathbf{V}\mathbf{V}^T$, where $\mathbf{V}$ is a symplectic matrix, and reconstruct the distribution in normalized coordinates $\mathbf{x}_n = \mathbf{V}^{-1}\mathbf{x}$ by setting $\mathcal{M}_k \to \mathcal{M}_k\mathbf{V}$. The covariance matrix is usually overdetermined by the measurements—for example, three measurements determine the 2×2 covariance matrix—so that all distributions that fit the data share the same covariance matrix. In these cases, it is reasonable to fit the covariance matrix first.

[5] A Gaussian prior may be a reasonable choice for accelerator applications: (i) the prior has no interplane dependence and can expand to approximate a uniform distribution; (ii) any known elements of the $n \times n$ covariance matrix can be used to define the Gaussian prior; (iii) beams are typically clustered in phase space and approximately Gaussian at equilibrium; (iv) a Gaussian prior can be used to limit the beam size in dimensions that are weakly constrained by the data.

by 1.5. We stopped training when

$$\langle D \rangle = \frac{1}{K} \sum_{k=1}^{K} D[g_k(\mathbf{u}_{k_\parallel}), \tilde{g}_k(\mathbf{u}_{k_\parallel})] < \epsilon, \tag{21}$$

where $\langle D \rangle$ is the average divergence between the simulated and measured projections, K is the number of measurements, and $\epsilon \approx 10^{-4}$ is a threshold. Other hyperparameter values are found in [31], which contains the code to reproduce the figures in this paper.

A. 2D reconstructions from 1D projections

Our first experiment tests the model performance in 2:1 phase space tomography. We assume an accelerator composed of drifts and quadrupole magnets, such that a symplectic transfer matrix $\mathbf{M}$ approximates the dynamics. The transfer matrix can be decomposed as

$$\mathbf{M} = \mathbf{V}(\alpha_2, \beta_2)\mathbf{R}(\mu)\mathbf{V}(\alpha_1, \beta_1)^{-1}, \tag{22}$$

where

$$\mathbf{V}(\alpha, \beta) = \begin{bmatrix} \sqrt{\beta} & 0 \\ -\frac{\alpha}{\sqrt{\beta}} & \frac{1}{\sqrt{\beta}} \end{bmatrix} \tag{23}$$

is a normalization matrix, parameterized by α and β, and

$$\mathbf{R}(\mu) = \begin{bmatrix} \cos\mu & \sin\mu \\ \sin\mu & \cos\mu \end{bmatrix} \tag{24}$$

is a rotation by the phase advance μ. The projection angle, and hence the reconstruction quality, depends only on the phase advance. Various constraints can limit the projection angle range, but we assume the projection angles are evenly spaced over the maximum 180-degree range.

Fig. 2 shows reconstructions from a varying number of projections, comparing MENT, MENT-Flow, and the unregularized neural network (NN). It is clear that maximizing the stochastic estimate in Eq. (16) pushes the distribution's entropy close to its constrained maximum. (Recall that MENT maximizes entropy by construction). Although the MENT solutions are of higher quality, the differences are not visible from afar.

Fig. 2 illustrates that entropy maximization is a conservative approach to the reconstruction problem. All reconstructed features are implied by the data. In contrast, the distributions in the bottom rows fit the data but are unnecessarily complex. Of course, reconstructions from one or two projections are bound to fail if the prior is uninformative, but these cases are still useful because they demonstrate MENT's logical consistency: given only the marginal distributions and an uncorrelated prior, the posterior is the product of the marginals. On the other extreme, with enough data, the feasible distributions differ only in minor details. MENT shines in intermediate cases where the measurements contain just enough information to constrain the distribution's primary features. For example, the continuous spiral structure develops rapidly with the number of views in Fig. 2.

Fig. 2 also illustrates the flow's capacity to represent complicated distributions despite the restriction to invertible transformations. This example focuses on spiral patterns, which are characteristic of nonlinear dynamics. (Additional examples are included in the supplemental material.) It is important to note that, while our analysis focuses on the beam core, low-density regions can also impact accelerator performance [32]. Flows can struggle to model distribution tails [33]. Our ground-truth distribution does not have significant halo and we do not report the agreement at this level; however, preliminary studies indicate the Kullback-Leibler (KL) divergence may enhance dynamic range relative to, i.e., the mean absolute error when fitting data.

Fig. 3 plots the entropy and data mismatch terms during training. The end of each epoch is clear from the sharp jumps in the loss curves when the penalty parameter μ increases. We aimed to keep the penalty parameter updates as small as possible. More aggressive update schedules did not lead to dramatically different results, but we did not explore this in detail. We did not see significant improvements using more sophisticated Augmented Lagrangian (AL) methods [17, 34]. A more important choice seems to be the stopping criteria, as increasing μ can eventually cause ill-conditioning. Our stopping condition ($\langle D \rangle \leq 10^{-4}$ in Eq. (21)) was chosen based on visual comparison of the simulated and true projections. We are not sure if the ideal stopping condition can be determined automatically.

B. 6D reconstructions from 1D projections

It is more difficult to design and evaluate high-dimensional numerical experiments. *First*, establishing reconstruction accuracy requires high-dimensional visualization or statistical distance metrics. We selected ground-truth distributions with clear high-dimensional structure and leveraged complete sets of pairwise projections and limited sets of partial projections (projections of slices) to aid the visualization.

Second, we cannot determine the distance from the reconstructed distribution to the true maximum-entropy distribution without an analytic solution. We point to Fig. 2 as evidence that the entropy penalty can push the MENT-Flow solution close to the exact solution. We also continued to train an unregularized neural network on the same data to show that additional solutions can exist far from the prior.

Third, for a given beamline and a fixed number of measurements, we do not yet know how to find the information-maximizing set of 6D phase space transformations. In 2:1 tomography, if the transformations are linear, the reconstruction quality is tied to a single parameter (the projection angle). There is no such connec-

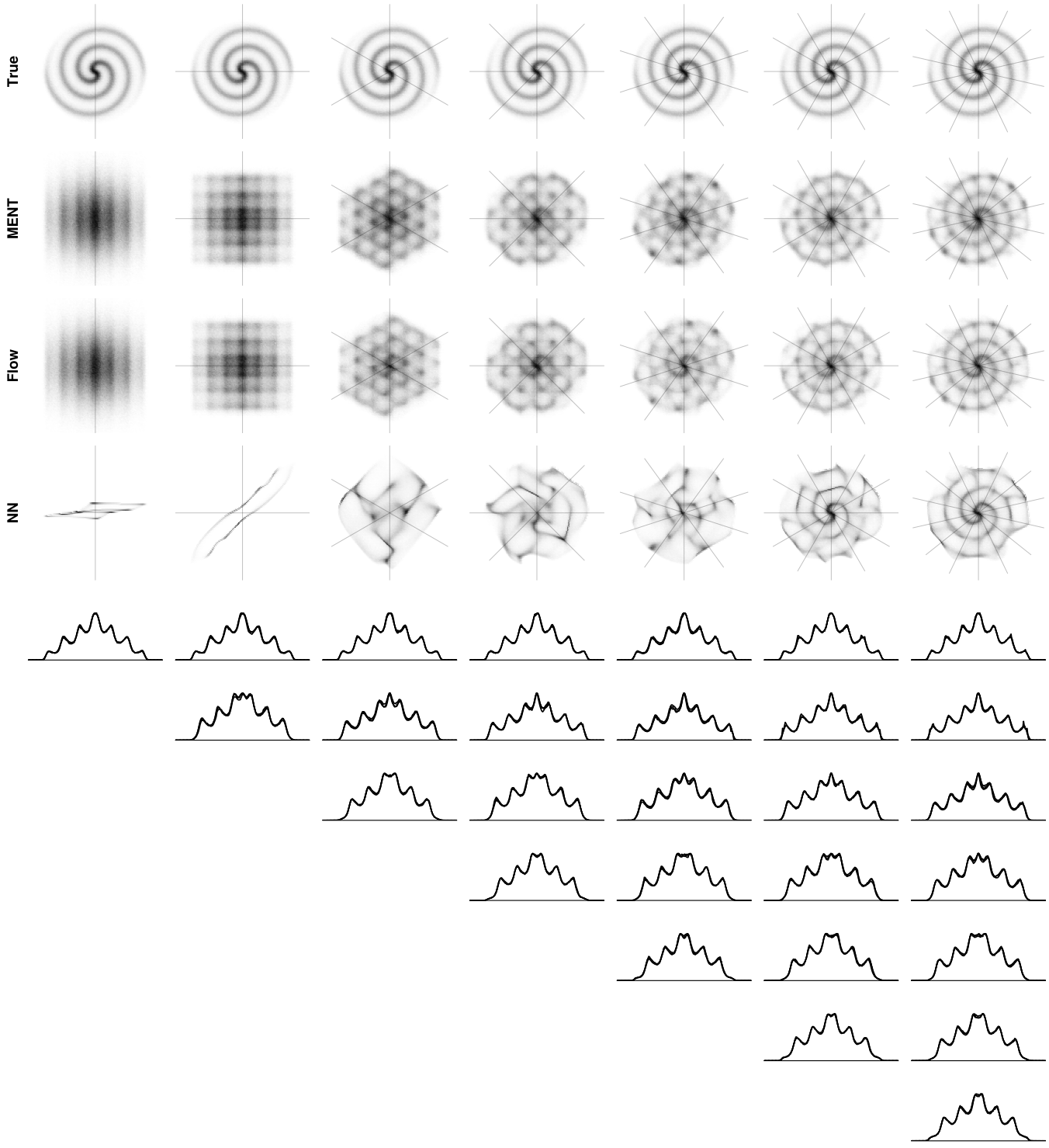

FIG. 2. 2D reconstructions from evenly spaced 1D projections. The top four rows plot samples from the true distribution, MENT reconstruction, MENT-Flow reconstruction, and NN reconstruction. Faint lines show the evenly spaced projection angles, increasing from 1 in the left column to 7 in the right column. In the bottom rows, the distributions are projected onto the measurement axes. (The four profiles overlap in most cases.)

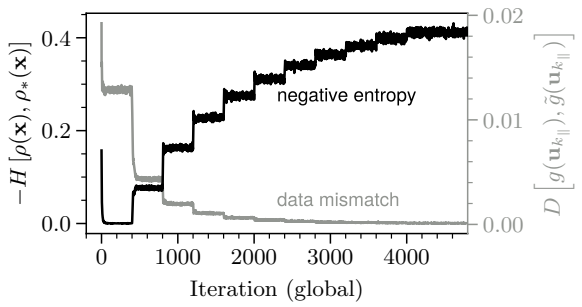

FIG. 3. Entropy and data mismatch during training.

tion in n:2 tomography when $n > 3$, as there is no obvious analog of the projection angle in these cases. Here, to demonstrate the method, we instead restrict our attention to 1D projections. A 1D projection axis can be specified by a point on the unit sphere; if the distribution is spherically symmetric, we hypothesize that the optimal projection axes are uniformly spaced on the sphere. In 2D, this leads to evenly spaced projection angles between 0 and π radians. In our numerical experiments, we approximated this condition by randomly sampling points from a uniform distribution on the sphere. The points will *not* be uniformly spaced, but in the limit of many projections, the reconstruction should converge to the true distribution [35].

Our first high-dimensional experiment, shown in Figs. 4-5, reconstructs a seven-mode Gaussian mixture distribution (a superposition of seven Gaussian distributions, each with a random mean and variance) from random 1D projections. Fig. 4 uses 25 projections and Fig. 5 uses 100 projections. This reconstruction used the same flow architecture as the 2D experiments. The NN architecture was changed to 2 layers of 50 units, still with tanh activation functions. We draw the following conclusions. (i) Normalizing flows can represent complicated 6D distributions far from the unimodal base distribution. All simulated measurements match the training data. Charged particle beams are often smooth and unimodal, so this example represents a challenging case. Therefore, flow-based models are likely sufficient for many applications in accelerator physics. (ii) MENT-Flow can simultaneously fit a large number of measurements. (iii) The entropy penalty works as intended. The entropy-regularized solution fits the data just as well as the NN solution but eliminates high-frequency terms in the distribution function. The MENT-Flow solution is much closer to the smooth prior.

The Gaussian mixture distribution has little overlap between modes, so mismatch between the true and reconstructed distribution is obvious from low-dimensional views. Hollow structures in high-dimensional phase space are not always evident from low-dimensional views. As an example, measurements at the Spallation Neutron Source (SNS) Beam Test Facility (BTF) show space-charge-driven hollowing in 3D and 5D projections of the 6D phase space distribution [1–3]. This motivates us to consider distributions with hidden internal structure. To this end, an n-dimensional "rings" distribution serves as the ground truth in Fig. 6-7; particles populate two concentric n-spheres with radii $r_2 = 2r_1$, and the radii are perturbed with Gaussian noise to generate a smooth density.

The entropy-regularized solution maintains the spherical symmetry of the Gaussian prior, flattening and eventually inverting its radial density profile to fit the data.[6] The sliced views reveal an internal structure—a dense core surrounded by a low-density cloud—that MENT-Flow better approximates when measurements are scarce. In addition to injecting unnecessary correlations between planes, the unregularized solution ejects all particles from the core. Surprisingly, adding additional measurements does not solve the problem and generates two distinct modes in the reconstructed density. Using a different random seed to define the measurement axes can generate different patterns, but the hollowing and splitting just described are typical. Note that this internal structure is not obvious from the full 2D projections in the left column of Figs. 6-7.

IV. CONCLUSION AND EXTENSIONS

In conclusion, MENT-Flow is a promising approach to high-dimensional phase space tomography. Numerical experiments demonstrate consistency with known 2D maximum-entropy solutions and the ability to fit complex 6D distributions to large measurement sets. In the 6D tests, although there are no available benchmarks, we found that entropic regularization pulls the solution closer to the prior. Thus, MENT-Flow is an effective way to incorporate prior information in high-dimensional reconstructions. Our numerical experiments also emphasize the potential importance of uncertainty quantification in high-dimensional tomography, as we found that some distributions can only be reconstructed from large numbers of 1D measurements. Future work should apply MENT-Flow to more realistic distributions, accelerator

[6] The one-dimensional projections in Figs. 6-7 are nearly Gaussian. Klartag [36] proved that almost all m-dimensional projections of an isotropic (no linear correlations) n-dimensional log-concave distribution function are nearly Gaussian when $n \gg m$. Many distributions commonly used in accelerator modeling are log-concave, such as the n-dimensional Gaussian, Waterbag (uniformly filled ball), and KV (uniformly filled sphere) distributions. A practical implication of this theorem is that small fluctuations in the m-dimensional projections have a greater impact on the n-dimensional reconstructed distribution as $n-m$ increases—for instance, completely inverting the density profile from peaked to hollow. Thus, we found that later training epochs can significantly change the distribution while only slightly decreasing the loss function. It follows that, for certain distributions, there may be some value of $n-m$ for which n:m tomography is practically impossible.

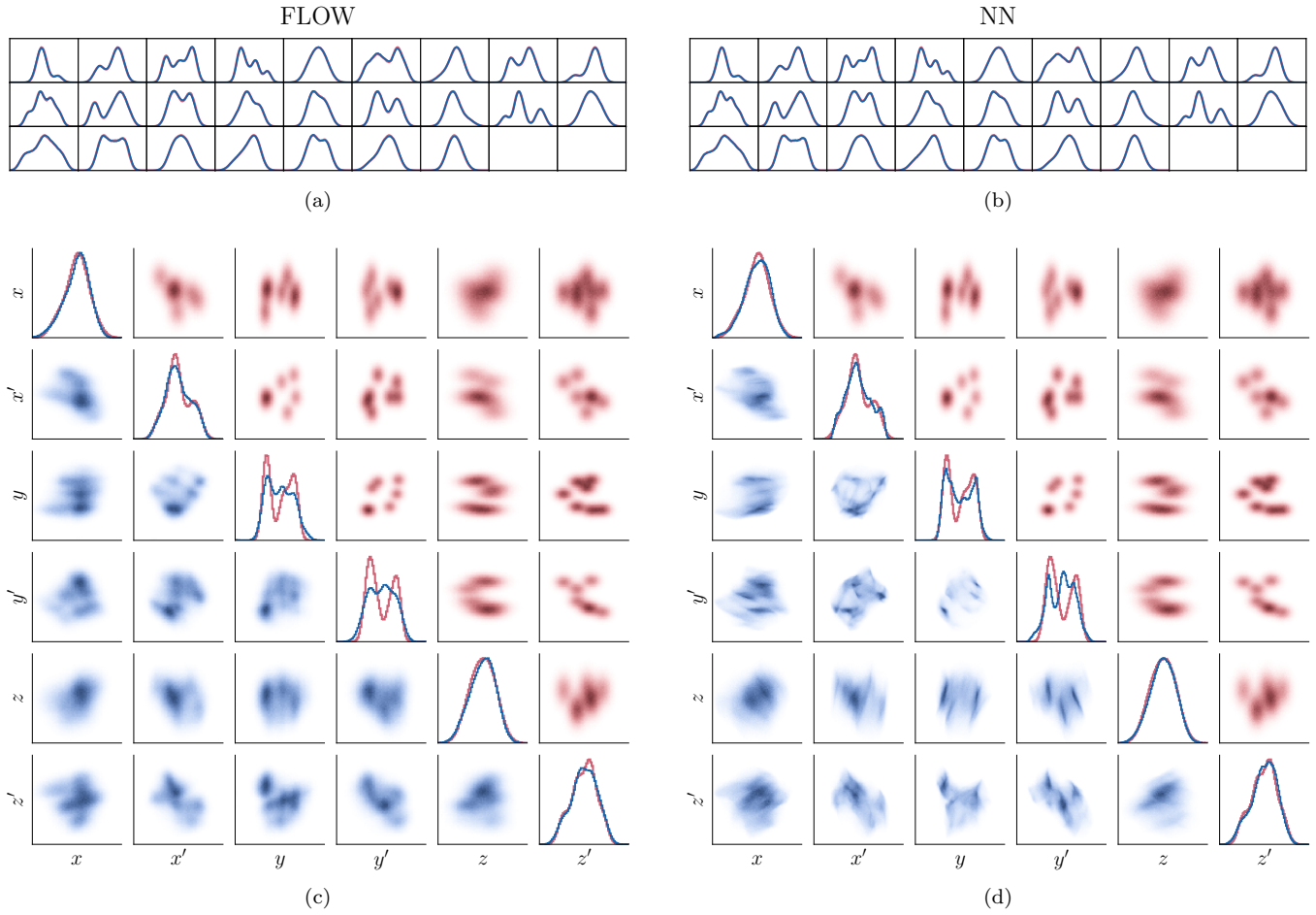

FIG. 4. Reconstruction of a 6D Gaussian mixture distribution from 25 random 1D projections. The MENT-Flow reconstruction on the left is compared to the NN reconstruction on the right. (a-b) Simulated projections (blue) vs. measured projections (red). (c-d) Low-dimensional views of the reconstructed distribution (blue) and the ground-truth distribution (red). 1D profiles are plotted on the diagonal subplots. 2D projections are plotted on the off-diagonal subplots.

models, and diagnostics, especially 2D projections. Future work should also aim to extend MENT to higher dimensions to serve as a benchmark.

MENT-Flow has several limitations. First, particle sampling is over 50% slower than a conventional neural network, and the total runtime is inflated by the need to solve multiple subproblems to approach the maximum-entropy distribution from below. This motivates the search for more efficient flows and sample-based entropy estimates. Note that our training times ranged from 5 to 20 minutes on a single GPU, depending on the number of projections, phase space dimension, batch size, and penalty parameter updates. Second, MENT-Flow maximizes the entropy using a penalty method that does not generate exact solutions and requires a hand-tuned penalty parameter schedule to avoid ill-conditioning. It is unclear whether this process can be automated or whether alternative strategies can better prevent ill-conditioning. Third, MENT-Flow does not attach uncertainty to its output.

We now discuss possible extensions to new problems. First, it may be possible to fit n-dimensional distributions to m-dimensional projections when $m > 2$. This problem is of theoretical interest but also has some practical relevance. 3D and 4D projections can be measured relatively quickly using slit-screen-dipole measurement systems in low-energy hadron accelerators [1–3]. We propose to draw samples from the measured projections and minimize a differentiable statistical distance between these samples and samples from the normalizing flow.

Second, an interesting application of maximum-entropy tomography is to intense hadron beams, in which particles respond to both applied and self-generated electromagnetic fields. Phase space tomography is a significant challenge for such beams because the forward process depends on the unknown initial distribution. We do not know if the maximum-entropy distribution is unique in this case. Including space charge in the GPSR forward process may be possible using differentiable space charge solvers [37].

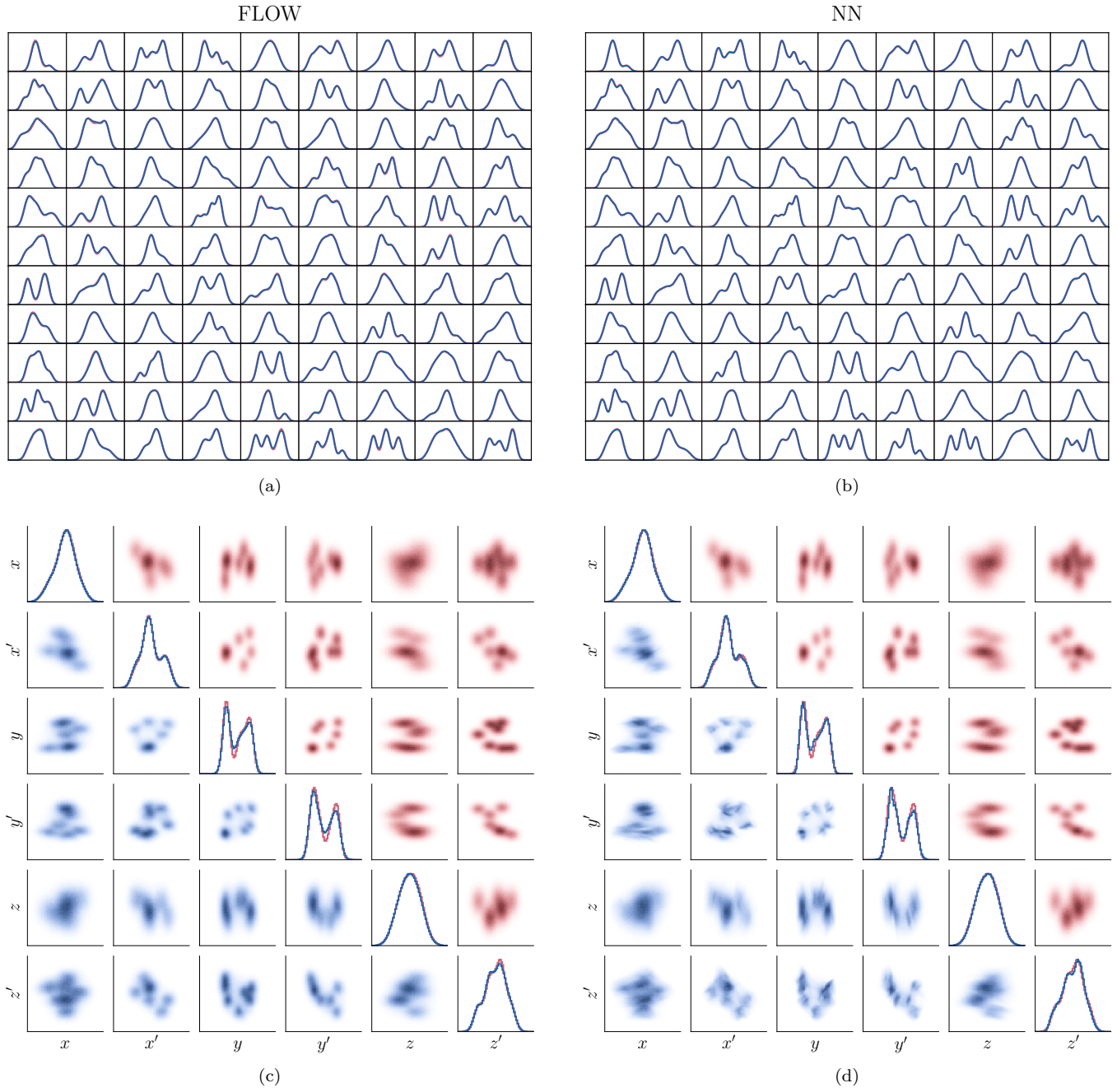

FIG. 5. Reconstruction of a 6D Gaussian mixture distribution from 100 random 1D projections. The MENT-Flow reconstruction on the left is compared to the NN reconstruction on the right. (a-b) Simulated projections (blue) vs. measured projections (red). (c-d) Low-dimensional views of the reconstructed distribution (blue) and the ground-truth distribution (red). 1D profiles are plotted on the diagonal subplots. 2D projections are plotted on the off-diagonal subplots.

Finally, quantifying reconstruction uncertainty is a crucial step for the operational use of phase space tomography. In this paper, we defined a prior probability distribution $\rho(\mathbf{x})$ over the phase space coordinates $\mathbf{x}$. We may also define a prior over the space of distribution functions. In practice, the phase space distribution is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^N$, so we may write the prior as $\mathcal{P}(\boldsymbol{\theta})$. The set of discretized measurements, i.e., histograms, can be expressed as a another parameter vector $\mathbf{d} \in \mathbb{R}^M$. Bayesian inference provides the update from prior to posterior:

$$\mathcal{P}(\boldsymbol{\theta}|\mathbf{d}) = \frac{\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})\mathcal{P}(\boldsymbol{\theta})}{\mathcal{P}(\mathbf{d})}, \tag{25}$$

where $\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})$ is the *likelihood*, encoding the forward model, and $\mathcal{P}(\mathbf{d})$ is a normalizing constant. The principle

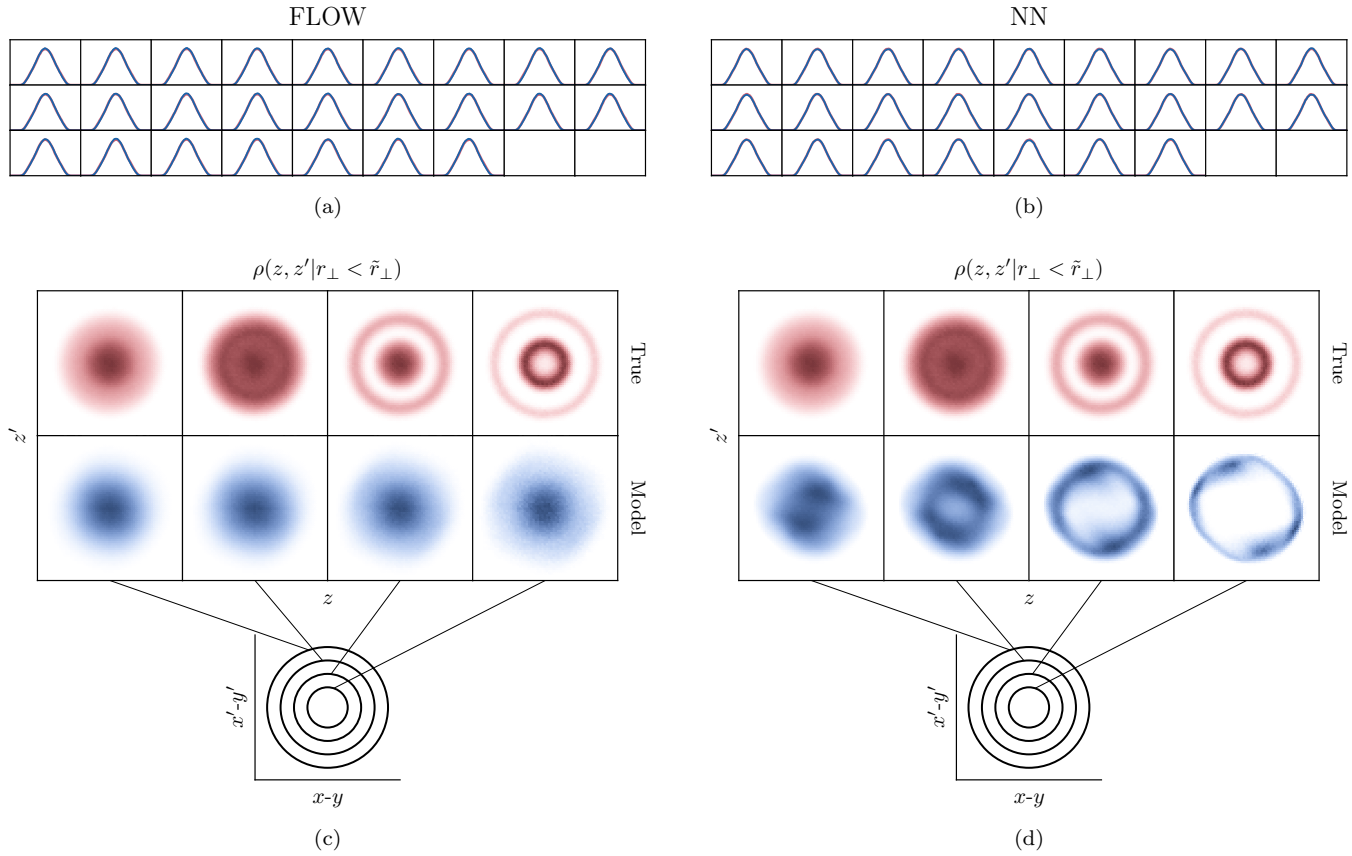

FIG. 6. Reconstruction of 6D "rings" distribution from 25 random 1D projections. The MENT-Flow reconstruction on the left is compared to the NN reconstruction on the right. (a-b) Simulated projections (blue) vs. measured projections (red). (c-d) 2D projections of the 6D distribution $\rho(x, x', y, y', z, z')$ within a shrinking 4D ball in the x-x'-y-y' plane. We define a ball of radius $\tilde{r}_\perp$ by $r_\perp \leq \tilde{r}_\perp$, where $r_\perp = \sqrt{x^2 + x'^2 + y^2 + y'^2}$. Therefore, we write the projected density as $\rho(z, z'|r_\perp \leq \tilde{r}_\perp)$. The ball shrinks from left to right. The largest radius (on the left) selects nearly all particles, while the smallest radius (on the right) selects particles near the core.

of maximum entropy implies that we should prefer higher entropy phase space distributions. Gull and Skilling [14] addressed this problem for image reconstruction and argued that the prior $\mathcal{P}(\boldsymbol{\theta})$ should be proportional to the exponential of the entropy $H(\boldsymbol{\theta})$. Given this prior, the strategy in this paper finds the maximum, or mode, of the posterior $\mathcal{P}(\boldsymbol{\theta}|\mathbf{d})$. The full posterior maps the entire solution space, encoding the reconstruction uncertainty. Future work could attempt to sample from a Gaussian approximation of the posterior at its maximum [14]. Alternatively, it may be possible to sample from a more accurate approximation of the posterior using Markov Chain Monte Carlo (MCMC) or machine learning methods [38].

V. ACKNOWLEDGEMENTS

We are grateful to Ryan Roussel (SLAC National Accelerator Laboratory), Juan Pablo Gonzalez-Aguilera (University of Chicago), and Auralee Edelen (SLAC National Accelerator Laboratory) for discussions that seeded the idea for this work and for sharing their differentiable kernel density estimation code.

This manuscript has been authored by UT Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

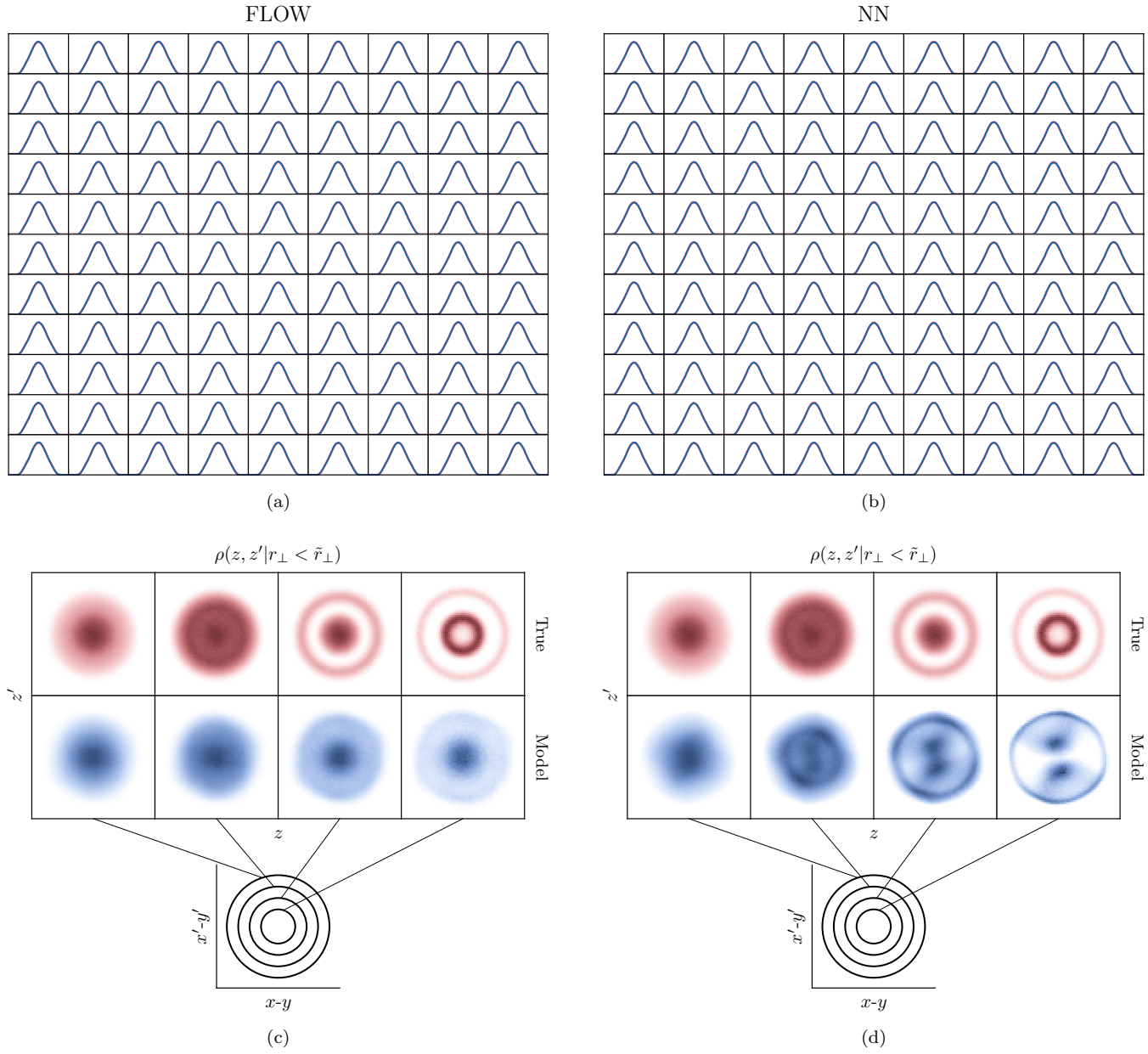

FIG. 7. Reconstruction of 6D "rings" distribution from 100 random 1D projections. The MENT-Flow reconstruction on the left is compared to the NN reconstruction on the right. (a-b) Simulated projections (blue) vs. measured projections (red). (c-d) 2D projections of the 6D distribution $\rho(x, x', y, y', z, z')$ within a shrinking 4D ball in the x-x'-y-y' plane. We define a ball of radius $\tilde{r}_\perp$ by $r_\perp \leq \tilde{r}_\perp$, where $r_\perp = \sqrt{x^2 + x'^2 + y^2 + y'^2}$. Therefore, we write the projected density as $\rho(z, z'|r_\perp \leq \tilde{r}_\perp)$. The ball shrinks from left to right. The largest radius (on the left) selects nearly all particles, while the smallest radius (on the right) selects particles near the core.

[1] B. Cathey, S. Cousineau, A. Aleksandrov, and A. Zhukov, First six dimensional phase space measurement of an accelerator beam, Phys. Rev. Lett. **121**, 064804 (2018).

[2] K. Ruisard, A. Aleksandrov, S. Cousineau, V. Tzoganis, and A. Zhukov, High dimensional characterization of the longitudinal phase space formed in a radio frequency quadrupole, Phys. Rev. Accel. Beams **23**, 124201 (2020).

[3] A. Hoover, K. Ruisard, A. Aleksandrov, A. Zhukov, and S. Cousineau, Analysis of a hadron beam in five-dimensional phase space, Phys. Rev. Accel. Beams **26**, 064202 (2023).

[4] Y. Liu, C. Long, and A. Aleksandrov, Nonintrusive mea-

surement of time-resolved emittances of 1-gev operational hydrogen ion beam using a laser comb, Phys. Rev. Accel. Beams **23**, 102806 (2020).
[5] A. Wolski, D. C. Christie, B. L. Militsyn, D. J. Scott, and H. Kockelbergh, Transverse phase space characterization in an accelerator test facility, Phys. Rev. Accel. Beams **23**, 032804 (2020).
[6] K. Hock and A. Wolski, Tomographic reconstruction of the full 4D transverse phase space, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **726**, 8 (2013).
[7] M. Wang, Z. Wang, D. Wang, W. Liu, B. Wang, M. Wang, M. Qiu, X. Guan, X. Wang, W. Huang, and S. Zheng, Four-dimensional phase space measurement using multiple two-dimensional profiles, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **943**, 162438 (2019).
[8] B. Marchetti, A. Grudiev, P. Craievich, R. Assmann, H.-H. Braun, N. Catalan Lasheras, F. Christie, R. D'Arcy, R. Fortunati, R. Ganter, *et al.*, Experimental demonstration of novel beam characterization using a polarizable x-band transverse deflection structure, Scientific reports **11**, 3560 (2021).
[9] A. Wolski, M. A. Johnson, M. King, B. L. Militsyn, and P. H. Williams, Transverse phase space tomography in an accelerator test facility using image compression and machine learning, Phys. Rev. Accel. Beams **25**, 122803 (2022).
[10] R. Roussel, A. Edelen, C. Mayes, D. Ratner, J. P. Gonzalez-Aguilera, S. Kim, E. Wisniewski, and J. Power, Phase space reconstruction from accelerator beam measurements using neural networks and differentiable simulations, Physical Review Letters **130**, 145001 (2023).
[11] S. Jaster-Merz, R. W. Assmann, R. Brinkmann, F. Burkart, W. Hillert, M. Stanitzki, and T. Vinatier, 5D tomographic phase-space reconstruction of particle bunches, Phys. Rev. Accel. Beams **27**, 072801 (2024).
[12] R. Roussel, J. P. Gonzalez-Aguilera, A. Edelen, E. Wisniewski, A. Ody, W. Liu, Y.-K. Kim, and J. Power, Efficient 6-dimensional phase space reconstruction from experimental measurements using generative machine learning (2024).
[13] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Principles of maximum entropy and maximum caliber in statistical physics, Rev. Mod. Phys. **85**, 1115 (2013).
[14] J. Skilling and S. F. Gull, Bayesian maximum entropy image reconstruction, Lecture Notes-Monograph Series , 341 (1991).
[15] R. D. Rosenkrantz, *ET Jaynes: Papers on probability, statistics and statistical physics*, Vol. 158 (Springer Science & Business Media, 2012).
[16] A. Giffin, *Maximum Entropy: The Universal Method for Inference*, Ph.D. thesis, University at Albany, State University of New York, Albany, NY, USA (2008).
[17] G. Loaiza-Ganem, Y. Gao, and J. P. Cunningham, Maximum entropy flow networks, in *International Conference on Learning Representations* (2016).
[18] J. C. Wong, A. Shishlo, A. Aleksandrov, Y. Liu, and C. Long, 4D transverse phase space tomography of an operational hydrogen ion beam via noninvasive 2d measurements using laser wires, Physical Review Accelerators and Beams **25**, 10.1103/PhysRevAccelBeams.25.042801 (2022).
[19] C. Mottershead, Maximum entropy tomography, in *Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods, Santa Fe, New Mexico, USA* (Springer, 1996) pp. 425–430.
[20] G. Minerbo, MENT: A maximum entropy algorithm for reconstructing a source from projection data, Computer Graphics and Image Processing **10**, 48 (1979).
[21] N. J. Dusaussoy and I. E. Abdou, The extended MENT algorithm: a maximum entropy type algorithm using prior knowledge for computerized tomography, IEEE Transactions on Signal Processing **39**, 1164 (1991).
[22] A. Tran and Y. Hao, Beam tomography with coupling using maximum entropy technique, in *Proc. 14th International Particle Accelerator Conference*, 14 (JACoW Publishing, Geneva, Switzerland, 2023) pp. 3944–3947.
[23] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models, IEEE transactions on pattern analysis and machine intelligence **44**, 7327 (2021).
[24] J. Kaiser, C. Xu, A. Eichler, and A. Santamaria Garcia, Bridging the gap between machine learning and particle accelerator physics with high-speed, differentiable simulations, Phys. Rev. Accel. Beams **27**, 054601 (2024).
[25] Z. Ao and J. Li, Entropy estimation via normalizing flow, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36 (2022) pp. 9990–9998.
[26] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, The Journal of Machine Learning Research **22**, 2617 (2021).
[27] V. Stimper, B. Schölkopf, and J. M. Hernández-Lobato, Resampling base distributions of normalizing flows, in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Proceedings of Machine Learning Research, Vol. 151 (PMLR, 2022) pp. 4915–4936.
[28] G. M. Green, Y.-S. Ting, and H. Kamdar, Deep potential: Recovering the gravitational potential from a snapshot of phase space, The Astrophysical Journal **942**, 26 (2023).
[29] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, Advances in neural information processing systems **32** (2019).
[30] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, Advances in neural information processing systems **30** (2017).
[31] A. Hoover, MENT-Flow: maximum-entropy phase space tomography using normalizing flows, 10.5281/zenodo.11110801 (2024).
[32] A. Aleksandrov, S. Cousineau, and K. Ruisard, Understanding beam distributions in hadron linacs in the presence of space charge, Journal of Instrumentation **15** (7), P07025.
[33] M. Laszkiewicz, J. Lederer, and A. Fischer, Marginal tail-adaptive normalizing flows, in *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 162, edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (PMLR, 2022) pp. 12020–12048.
[34] S. Basir and I. Senocak, An adaptive augmented lagrangian method for training physics and equality constrained artificial neural networks, arXiv preprint arXiv:2306.04904 (2023).

[35] B. Dai and U. Seljak, Sliced iterative normalizing flows, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021) pp. 2352–2364.

[36] B. Klartag, A central limit theorem for convex sets, Inventiones mathematicae **168**, 91 (2007).

[37] J. Qiang, Differentiable self-consistent space-charge simulation for accelerator design, Phys. Rev. Accel. Beams **26**, 024601 (2023).

[38] M. Mardani, J. Song, J. Kautz, and A. Vahdat, A variational perspective on solving inverse problems with diffusion models, arXiv preprint arXiv:2305.04391 (2023).