Bridging the Simulation-to-Experiment Gap with Generative Models using Adversarial Distribution Alignment

Kai Tyrus Nelson[1*] **Tobias Kreiman**[1*] **Sergey Levine**[1] **Aditi S. Krishnapriyan**[1,2]
[1]University of California, Berkeley [2]LBNL
{kai_nelson, tkreiman, sergey.levine, aditik1}@berkeley.edu

Abstract

A fundamental challenge in science and engineering is the simulation-to-experiment gap. While we often possess prior knowledge of physical laws, these physical laws can be too difficult to solve exactly for complex systems. Such systems are commonly modeled using simulators, which impose computational approximations. Meanwhile, experimental measurements more faithfully represent the real-world, but experimental data typically consists of *observations* that only partially reflect the system's full underlying state. We propose a data-driven distribution alignment framework that bridges this simulation-to-experiment gap by pre-training a generative model on fully observed (but imperfect) simulation data, then aligning it with partial (but real) observations of experimental data. While our method is domain-agnostic, we ground our approach in the physical sciences by introducing Adversarial Distribution Alignment (ADA). This method aligns a generative model of atomic positions—initially trained on a simulated Boltzmann distribution—with a distribution of experimental observations. We prove that our method recovers the target observable distribution, even with multiple, potentially correlated observables. We also empirically validate our framework on synthetic, molecular, and experimental protein data, demonstrating that it can align generative models with diverse observables. Our code is available at https://kaityrusnelson.com/ada/.

1 Introduction

Physical laws precisely describe the behavior of matter, but solving them is often prohibitively expensive when applied to real-world systems. A central challenge in computational science is thus the development of accurate yet tractable models. For example, numerical simulation methods in atomistic modeling trade computational cost for accuracy by introducing various approximations from the classical to the quantum level [2, 22, 47, 26, 7]. Despite their simplifying assumptions, by predicting physical properties *in silico*, accurate computational models can reduce reliance on expensive laboratory experiments.

Meanwhile, direct experimental measurements better reflect the real world, but exact measurements of the full underlying state of a system can be costly or completely inaccessible. Consequently, experimental data often only reflects *partial observations* of the real world. Although latent variable models can be used to reason about partially observed systems [40, 3], such models can be data-intensive and challenging to train in practice, especially in experimental settings where measurements are expensive. As a result, this leaves a discrepancy between abundant, fully-observed but approximate simulation data and scarce, partially-observed but more accurate experimental measurements: a discrepancy which we refer to as the *simulation-to-experiment gap*.

To address this gap, we propose **Adversarial Distribution Alignment from Partial Observations (ADA)**, an algorithm that first learns a generative model of fully observed simulated data, and

*Denotes equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

then aligns it with partially observed experimental measurements (see Figure 1). ADA aligns the model with experimental observable *distributions* via a min–max objective. Our algorithm alternates between learning to distinguish observable samples generated by the model from samples drawn from the experimental dataset, and updating the generative model accordingly. By leveraging real-world data as a scalable source of supervision, ADA helps bridge the simulation-to-experiment gap when training machine learning models of the physical world.

Importantly, these real-world data distributions correspond to the true targets of interest in downstream applications such as materials and drug discovery, where accurate property prediction enables screening of promising candidates. For example, ADA can align a generative model of protein structures—pre-trained on classical molecular dynamics simulations as the initial physical prior—with high-dimensional, noisy cryo-EM observations to better capture the real-world distribution of protein states (amino acid positions). We preview such an application in Section 4.3.

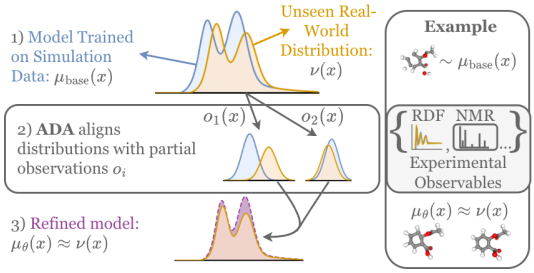

Figure 1: **ADA aligns generative models trained on approximate simulation data with the real world by leveraging partial experimental observations.** (1) ADA starts from a base generative model $\mu_{\text{base}}(x)$ trained on simulation data, such as molecular dynamics simulations using a classical force field. (2) ADA aligns the base model with *multiple, potentially correlated* partial experimental observations (e.g., radial distribution functions (RDFs) or nuclear magnetic resonance (NMR) measurements). (3) ADA yields an aligned model $\mu_\theta(x)$ that approximates the true underlying real-world distribution $\nu(x)$.

Our main contribution is the introduction of ADA as a framework that leverages physical knowledge through simulated data, while mitigating the approximation errors of purely simulation-based approaches by aligning to real-world distributions. We prove that ADA recovers the target observable distribution, even in the presence of multiple, potentially correlated observables. We also show empirically that ADA can align generative models to diverse molecular and protein observables, with alignment to experimental distributions improving as additional observables are incorporated.

2 Distribution Matching with Observables

Our goal is to train a generative model $\mu_\theta(x)$ that matches the distribution of observables measured from an underlying system $\nu(x)$. In this context, we define observables to be any function of the state of the system x:

$$o^{(i)} : X \to O^{(i)} \subseteq \mathbb{R}^{k_i},$$

where k_i is the dimensionality of observable i, and $i \in \{1, \ldots, m\} = I$ indexes the m observables. We only assume access to the observable quantities drawn from our underlying system, and presume no direct access to the distribution of true states $\nu(x)$. We additionally note that our m observables may be arbitrarily correlated.

In general, modeling $\nu(x)$ is challenging because the amount of data for the full system might be limited or impossible to access, leaving observables that only offer a *partial* view of the underlying system. Therefore, to make it more feasible to model $\nu(x)$, we leverage a model, $\mu_{\text{base}}(x)$, pre-trained on *fully observed* data from an imperfect simulator of the experimental distribution, e.g., data generated from a numerical simulation. Our goal is to shift the base distribution to produce a learned distribution $\mu_\theta(x)$, which satisfies the constraint:

$$o^{(i)}_{\#}\mu_\theta = o^{(i)}_{\#}\nu, \quad \forall i, \tag{1}$$

where $o^{(i)}_{\#}\nu$ denotes the *pushforward* of the distribution ν through the observable $o^{(i)}$, that is, the distribution over observable i obtained by sampling $x \sim \nu(x)$ and mapping it to $o^{(i)}(x)$. Succinctly, we wish to learn a distribution μ_θ whose observables match the experimental observable distribution.

The objective in Equation (1) is generally underconstrained, since the observable is only a partial observation of the underlying system, and many states could map to the same observable value. Therefore, we regularize the solution by enforcing closeness to the base distribution via a KL divergence. This acts as an informed prior, supplementing sparse and lossy real world data with reasonably accurate simulator data. If we define $\mathcal{M}_o(\mathcal{X}) = \{\mu : o^{(i)}_{\#}\mu = o^{(i)}_{\#}\nu; \forall i \in I\}$ to be the set of all probability measures satisfying Equation (1), our full objective becomes:

$$\begin{aligned} \arg\min_{\mu_\theta} \quad & D_{\mathrm{KL}}(\mu_\theta \,\|\, \mu_{\text{base}}) \\ \text{s.t.} \quad & \mu_\theta \in \mathcal{M}_o(\mathcal{X}). \end{aligned} \tag{2}$$

From an information-theoretic perspective, Equation (2) corresponds to an information projection of the base distribution onto the set of distributions satisfying the observable constraints, with the KL term acting as an entropic regularizer (see Figure 2). We note the connection between this objective and inverse reinforcement learning (IRL), where the goal is to learn a reward function explaining expert behavior [19, 46]. In this framework, the learned reward implicitly distinguishes expert behavior from that induced by the current policy, and the policy is iteratively improved by optimizing this reward. In our setting, we have experimental observables instead of expert trajectories, while the policy corresponds to a generative model. In contrast to prior work that matches only *expectations* of reference features, our formulation seeks to match the *full distribution* of multiple, potentially correlated observables.

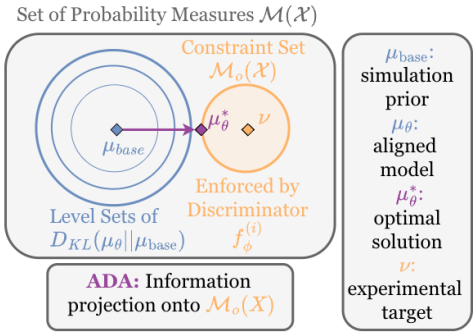

Figure 2: **ADA finds the closest distribution μ^*_θ to a base distribution μ_{base} that satisfies the observable constraints $o^{(i)}_{\#}\mu_\theta = o^{(i)}_{\#}\nu$ for all observables $i \in \{1, \ldots, m\}$.** We denote the resulting feasible set by $\mathcal{M}_o(\mathcal{X})$ (orange). The KL term in Equation (2) acts as an entropic regularizer, reflecting the fact that the constraints in Equation (1) do not uniquely specify the underlying distribution, since many states can map to the same observable value. The discriminator $f^{(i)}_\phi$ learns the observable Wasserstein distance to align the model with the distributional observable constraint.

2.1 Comparison of our Problem Setting to Conditional Generation and Expectation Alignment

To build intuition about our problem setting, we first contrast it with other related formulations. We discuss conditional generative modeling and expectation alignment (EA) approaches [34, 5, 36], which align distributions by matching observable moments, to motivate our algorithm in Section 2.2.

Comparison with conditional generative modeling. We first consider conditional generative modeling, which explicitly factorizes the target distribution into an observable prior and a generative model conditioned on an observable value. If one had access to a paired dataset $\mathcal{D} = (x_j, o^{(i)}_j = o^{(i)}(x_j))$, one could attempt to satisfy Equation (2) by training a conditional model $\mu_{\text{base}}(x \mid o^{(i)}_j)$ and sampling from: $\mu_\theta(x) = \int \mu_{\text{base}}(x \mid o^{(i)}) d\nu(o^{(i)})$.

While this approach is viable when paired samples are available for a *single* observable, it fails in the multi-observable setting when we have access only to marginal distributions $\{\nu(o^{(i)})\}_{i=1}^m$. Having access only to marginals prevents conditioning on values sampled from the joint distribution of observables, making it impossible to simultaneously satisfy the observable and KL constraints in Equation (2).

Comparison to expectation alignment methods. We additionally contrast our problem setting, which matches full distributions, with methods that align only to observable expectations. Throughout this paper, we refer to this class of methods as *expectation alignment (EA)*. In this formulation, one seeks a distribution μ_θ that minimizes a KL divergence to a base distribution while matching the expected values of a prescribed set of observables:

$$\arg\min_{\mu_\theta} \quad D_{\mathrm{KL}}(\mu_\theta \,\|\, \mu_{\text{base}}) \quad \text{s.t.} \quad \mathbb{E}_{x\sim\mu_\theta}[o^{(i)}(x)] = \mathbb{E}_{x\sim\nu}[o^{(i)}(x)], \quad \forall i \in I. \tag{3}$$

In theory, under mild regularity assumptions, EA can recover the full distribution-matching objective. This follows by augmenting the observable set with all polynomial moments, since enforcing Equation (3) over this family corresponds to matching all mixed moments of the observable distributions. In the limit of infinitely many moments, this uniquely identifies the observable marginals and recovers the solution to Equation (2).

In practice, however, this approach is infeasible: the number of moments grows combinatorially with the observable dimension and degree, and exact recovery of the target distribution generally requires matching infinitely many moments (see Section 4.1 for an empirical example). From this perspective, matching the full observable distribution amounts to enforcing equality over a sufficiently rich class of test functions of the observables. We instead propose parameterizing a discriminator as a neural network that directly learns functions of the observables, enabling full distributional alignment beyond matching a finite number of moments.

2.2 Adversarial Distribution Alignment with Partial Observations (ADA)

We now present our algorithm, Adversarial Distribution Alignment with Partial Observations (ADA), which performs full distributional alignment using only partial observations. To realize Equation (2), we re-frame the objective as a dual optimization problem, learning the Wasserstein distance between generated and real experimental observables in the same way as Arjovsky et al. [1]. Together with the known connections between GANs and inverse reinforcement learning [9, 16], this motivates an adversarial optimization procedure for Equation (2).

Specifically, to enforce the observable constraint in Equation (1), we use a Wasserstein distance over the space of observables to learn a pseudo-metric over state distributions. We will refer to this pseudo-metric as the observable Wasserstein distance. For each observable, we define the observable Wasserstein distance between the learned and target distributions as follows:

$$d^{(i)}(\mu,\nu) = \sup_{f^{(i)}\in \mathrm{Lips}(1)} \left(\mathbb{E}_{o^{(i)}_{\#}\mu}[f^{(i)}] - \mathbb{E}_{o^{(i)}_{\#}\nu}[f^{(i)}]\right), \tag{4}$$

where $\mathrm{Lips}(1)$ is the class of all Lipschitz-1 functions. In practice, this Wasserstein distance is approximated by learning a discriminator $f^{(i)}_\theta(o^{(i)}(x))$ with a local Lipschitz constant enforced by a gradient penalty [1].

We then parameterize the soft version of our constraint objective (Equation (2)) as:

$$\arg\max_{\mu_\theta} \quad -D_{\mathrm{KL}}(\mu_\theta \,\|\, \mu_{\text{base}}) - \beta \sum_{i\in I} d^{(i)}(\mu_\theta,\nu), \tag{5}$$

which corresponds to the mini-max objective:

$$\begin{aligned} &\max_{\mu_\theta} \min_{(f^{(i)})\in \mathrm{Lip}(1)^{|I|}} \mathcal{L}\big(\mu_\theta,(f^{(i)}),\beta\big) \qquad \text{where} \\ &\mathcal{L}\big(\mu_\theta,(f^{(i)}),\beta\big) = -D_{\mathrm{KL}}\big(\mu_\theta \,\|\, \mu_{base}\big) + \beta \sum_{i\in I} \left(\mathbb{E}_{(o^{(i)})_{\#}\mu_\theta}[f^{(i)}] - \mathbb{E}_{(o^{(i)})_{\#}\nu}[f^{(i)}]\right). \end{aligned} \tag{6}$$

We propose to alternate between learning our observable Wasserstein distance and improving our generative model to find the distribution closest to our base model that matches the target observable distribution (Figure 2).

Algorithm summary. We summarize our overall algorithm. We first initialize our generative model as $\mu_\theta = \mu_{\text{base}}$, where μ_{base} is trained on the simulated data distribution. Starting from this initialization, our algorithm iterates the following three steps to align μ_θ with the experimental data. (see Alg. 1). First, we sample from the current iteration of the aligned generative model $\mu_\theta(x)$. We then evaluate observables from the generated samples and compute the Lagrangian of our objective (Equation (6)). We finally update the generative model and the discriminators using stochastic gradient descent on the Lagrangian. We also emphasize that each discriminator may be trained independently per observable. As we will show later in Section 2.3, this is sufficient for satisfying the desired observable constraints, even in the case of multiple, correlated observables.

Table 1 summarizes the capability of ADA and compares it with conditional generative modeling and EA approaches discussed in Section 2.1. In short, while conditional generation can handle single observables and EA methods can align with a finite set of moments, ADA can recover the full distribution of multiple, potentially correlated observables.[2]

Practical implementation details. We describe how we optimize the Lagrangian Equation (6), which balances the observable Wasserstein distance with the prior. We note that the Lagrangian does not need to be explicitly calculated as long as we can optimize μ and $(f^{(i)})$ with respect to it. Many methods which are capable of optimizing a generative model with respect to a reward function could work in practice. In this work, we parameterize μ_θ and μ_{base} as diffusion models [35] and we use adjoint matching [8] to obtain unbiased gradient estimates of the Lagrangian with respect to μ without requiring backpropagation through the sampling process.

We note that adjoint matching requires observables to be differentiable. Although this may seem restrictive, many structural observables and energy functions can be computed efficiently in a differentiable manner, and entropy regularized reinforcement learning could be used for non-differentiable observables. [41, 33].

Algorithm 1 Observable Alignment from a Pretrained Generative Model

Input:
  Pretrained base generator $\mu_{\text{base}}(x)$
  Experimental observable datasets $\mathcal{D}^{(i)} = \{ o_j^{(i)} \sim (o^{(i)})_{\#}\nu \}$ for each observable $i \in I$
Initialize:
  Generator parameters θ with $\mu_\theta \leftarrow \mu_{\text{base}}$
  Observable critics $f^{(i)}_{\phi^{(i)}} : O_i \to \mathbb{R}$ for all $i \in I$

while not converged **do**
  Sample model configurations $\{x_j\}_{j=1}^N \sim \mu_\theta$
  for each observable $i \in I$ **do**
    Compute model observables $u_j^{(i)} \leftarrow o^{(i)}(x_j)$
    Sample reference observables $\{\tilde{o}_j^{(i)}\}_{j=1}^N \sim \mathcal{D}^{(i)}$
  end for
  Construct empirical push forward measures:

$$(o^{(i)})_{\#}\mu_\theta \approx \frac{1}{N}\sum_{j=1}^N \delta_{u_j^{(i)}}, \quad (o^{(i)})_{\#}\nu \approx \frac{1}{N}\sum_{j=1}^N \delta_{\tilde{o}_j^{(i)}}$$

$$\mathcal{L}(\theta,\{\phi^{(i)}\}) = -D_{\text{KL}}\big(\mu_\theta \,\|\, \mu_{\text{base}}\big) + \beta \sum_{i\in I} \Big(\mathbb{E}_{(o^{(i)})_{\#}\mu_\theta}[f^{(i)}_{\phi^{(i)}}] - \mathbb{E}_{(o^{(i)})_{\#}\nu}[f^{(i)}_{\phi^{(i)}}] \Big).$$

  Update generator (ascent, use Adjoint Matching to compute $\nabla_\theta \mathcal{L}$):

$$\theta \leftarrow \theta + \eta_\theta \nabla_\theta \mathcal{L}$$

  for each observable $i \in I$ **do**
    Update critic (descent):

$$\phi^{(i)} \leftarrow \phi^{(i)} - \eta_\phi \nabla_{\phi^{(i)}} \mathcal{L}$$

  end for
end while

2.3 Theoretical Convergence Analysis

We now address the theoretical guarantees of our objective. Our goal with the theoretical analysis is to show that the objective in Equation (6), (1) has a unique solution and (2) converges to the desired reference observable distribution $o^{(i)}_{\#}\nu$ as we increase the weight β on the observable Wasserstein distance. We first build intuition by analyzing the scenario where there is only a single observable, and then provide theorems for the general case of multiple, potentially correlated observables.

Single observable case. For the case of a single observable, we can directly solve Equation (2) to find the optimal distribution. Following previous work [46, 19, 24], we consider the Lagrangian of the hard constrained optimization problem. To delineate it from the Lagrangian of Equation (5), we denote it by $\mathcal{Z}$:

$$\mathcal{Z}(\mu_\theta, \lambda, \alpha) = \int dx\, \mu_\theta(x) \log \frac{\mu_\theta(x)}{\mu_{\text{base}}(x)} + \int do\, \lambda(o)\big(\nu(o) - \mu_\theta(o)\big) + \alpha \int dx\, (\mu_\theta(x) - 1) \quad (7)$$

[2] When we only have access to a distributional mean, ADA can subsume EA approaches by restricting the function class to a linear discriminator. This is well studied in the inverse reinforcement learning literature (e.g. Ziebart et al. [46])

where $\mu_\theta(o) = \int dx \delta(o - o(x))\mu_\theta(x)$ using δ as the Dirac delta function. We also note that there is a Lagrange multiplier for each value the observable could take on ($\lambda(o)$).

If we differentiate to find the optimal μ_θ, we find that: $\mu_\theta(x) = \mu_{\text{base}}(x)\frac{e^{\lambda(o(x))}}{Z}$. To solve for $\lambda(o(x))$, we note that by definition of the Dirac delta function:

$$\mu_\theta(o) = \int dx\, \delta(o - o(x))\mu_\theta(x) = \mu_{\text{base}}(o)\frac{e^{\lambda(o)}}{Z}. \tag{8}$$

Considering our constraint on the observable distribution: $\nu(o) = \mu_\theta(o) = \mu_{\text{base}}(o)e^{\lambda(o)}/Z$, we see that: $\lambda(o) = \log(Z\ \nu(o)/\mu_{\text{base}}(o))$, yielding a final distribution:

$$\mu^*_\theta(x) = \mu_{\text{base}}(x)\frac{\nu(o)}{\mu_{\text{base}}(o)} \tag{9}$$

that satisfies $\mu_\theta(o) = \nu(o)$. In short, for a single observable, Equation (2) has a solution that recovers the target observable distribution by tilting the base distribution towards regions where it underrepresents the observable density. Although solving for the closed-form tilted distribution is possible for a single observable— since we can estimate $\nu(o)$ from samples—the same derivation in the multi-observable setting would require access to the joint distribution of experimental observables (see Section 2.1 for details). Since we assume access only to marginal observable distributions, we instead provide a different analysis to handle the multi-observable case in the following section.

General case. Analyzing the general case with multiple observables requires more care. We present the following guarantees:

Theorem 2.1 (Existence and uniqueness of the saddle point)**.** *Assume X is a compact Polish space and that μ_{base} has full support on X. Additionally, allow $o^{(i)}$ to be continuous for $i \in I$. For any $\beta \in \mathbb{R}$, Equation (6) admits a saddle point $(\mu^*, (f^{(i)})^*)$, where μ^* is unique. Consequently, the recovered distribution is well defined. We denote this solution by $\mu^* \in \operatorname{sol}\mathcal{L}(\beta)$.*

Theorem 2.2 (Asymptotic convergence in Wasserstein)**.** *Assume $D_{\text{KL}}(\nu\|\mu_{base}) < \infty$. Let $\mu^*_\beta \in \operatorname{sol}\mathcal{L}(\beta)$ for $\beta \in \mathbb{R}^+$. Then, as $\beta \to \infty$, we have $d^{(i)}(\mu^*_\beta, \nu) \to 0$ for all $i \in I$.*

Theorem 2.3 (Asymptotic convergence to the constraint set)**.** *Let $(\mu^*_\beta)_{\beta>0}$ be the sequence of saddle-point solutions from Theorem B.2. Suppose that $d^{(i)}(\mu^*_\beta, \nu) \to 0$ for all $i \in I$. Then any weak limit point $\bar{\mu}$ of (μ^*_β) satisfies $(o^{(i)})_\#\bar{\mu} = (o^{(i)})_\#\nu$ for all $i \in I$, that is, $\bar{\mu} \in \mathcal{M}_o$.*

To summarize, our algorithm in theory converges and will be arbitrarily close to satisfying the constraint set for a sufficiently large β. The associated proofs can be found in Appendix B. Although the KL constraint on Theorem B.2 may seem overly restrictive, we note that this constraint is satisfied for many models of interest. For instance, any set of particles in a closed and bounded box with positive density are sufficient. We also note that we have imposed no unnecessary restrictions on the push forward of our observables, only that they continuously map to real valued vectors. We have not required that they be independent or uncorrelated. In fact, as long as the experimental dataset can be realized by a real measure with finite KL divergence from the base measure, we are able to get arbitrarily close to matching the observable measures in Wasserstein distance.

3 Related Work

Generative models. ADA builds on previous generative modeling frameworks. We parameterize μ_{base} and μ_θ as diffusion models [35, 17], which learn to reverse a stochastic process that corrupts a data distribution into noise. Generative adversarial networks (GANs) [13] are also closely related to our work. GANs learn a target distribution by alternating between learning a discriminator, which distinguishes real samples from generated ones, and a generator trained to fool the discriminator. ADA builds upon the GAN framework to learn distributions in the partially observed setting.

Maximum entropy inverse reinforcement learning. Adversarial training also plays a central role in inverse reinforcement learning (IRL) [46, 10], where the objective is to infer a reward function that explains observed behavior and to extract a policy consistent with that reward. The connection between GANs and IRL has been noted in prior work [9, 16], where the learned reward function acts as a discriminator and the policy serves as the generator. The maximum entropy IRL formulation

[19, 46], which augments reward learning with an entropic regularizer, closely aligns with our problem setting. However, rather than inferring a policy from demonstrations, we seek to align a generative model with experimental observations while leveraging simulated data as a prior. Importantly, we only leverage partial observations without requiring demonstrations of the full underlying state.

Learning from partial observations in atomistic generative modeling. In the context of computational chemistry and biology, Iterative Boltzmann Inversion (IBI) [36, 32, 15, 29] and related conditional modeling approaches (see Section 2.1) have leveraged a single observable, such as cryo-EM images [38, 37], to improve models. Simpler parametrized model classes [43] have also been used to learn from partial observations. More recent simulation-in-the-loop approaches [31, 11, 14, 12] use automatic differentiation to improve a machine learning force field using observable quantities. Most similar to our approach is concurrent work that aligns the outputs of a generative model using a linear reward [34].

However, we emphasize that these prior methods that leverage *multiple* observables typically frame their objective as an *expectation alignment (EA) problem* [5, 36, 31, 27, 23, 30, 44, 18, 21]. For instance, guidance methods [23, 27, 30] modify the score of a diffusion model, often with a trained classifier, to produce samples that match the moments of observables. Similarly, Smith et al. [34] fix a linear reward signal that they use to fine-tune a generative model to match the expectation of an observable. While in theory these EA approaches could learn a full distribution by enforcing infinitely many moments of the target distribution, this is in practice infeasible and modeling only a finite number of moments leads to discrepancies, even in low dimensions (see Section 2.1 and Section 4.1).

Our problem setting differs in that we aim to match the full distribution of *multiple* experimental observables rather than only low-order statistics. For example, in computational biology, matching a single, expected protein conformation is insufficient; modeling the full experimental distribution is necessary to study folding and unfolding events, compute free energies, and estimate transition rates. Furthermore, instead of relying on a single reward signal as in the IRL and conditional generative modeling formulations, we consider multiple, potentially correlated experimental observables that jointly constrain the target distribution. ADA extends the EA and IRL formulations to handle full distributional constraints over multiple observables.

4 Experiments

We evaluate ADA's ability to align a generative model, pretrained on simulated data, with experimental distributions using only partial observations. We first compare ADA to EA approaches similar to Smith et al. [34] to highlight the difference between full distributional alignment and moment matching (Section 4.1). Next, to obtain controlled evaluation metrics, we use simulated data at increasing levels of fidelity to quantify how alignment improves as additional observables are provided (Section 4.2). Finally, we assess ADA's ability to bridge the gap between a generative model trained on classical force field simulations and experimental measurements from the Protein Data Bank using cryo-EM observables [4] (Section 4.3). We provide experimental details, including further discussion of the chosen observables, in Appendix A. While our experiments rely on simulating the observable forward model, their goal is to understand ADA's behavior under controlled conditions that enable quantitative evaluation in multi-observable and high-dimensional settings.

4.1 Synthetic Data

To illustrate the limitations of EA methods, we consider a synthetic three-dimensional mixture-of-Gaussians benchmark (see Figure 6). We define the base distribution μ_{base} as an isotropic mixture with eight components located at the vertices of a cube, with uniform mixture weights. We define the target distribution ν by perturbing the variances and mixture weights for the Gaussians (see Appendix A for more details).

We choose observables $o^{(i)}$ as the three pairwise coordinate projections: $\pi_{X_i,X_j}(x_1, x_2, x_3) = (x_i, x_j)$. Despite their simplicity, these observables induce correlated, multimodal marginals.

For evaluation, we report two metrics: (1) the Wasserstein-1 distance between energy histograms of generated samples, where $E(x) = -\log \nu(x)$, and (2) the cluster assignment accuracy via the KL divergence $D_{\text{KL}}(\mu_\theta(z) \,\|\, \nu(z))$, where $\mu_\theta(z) = \int \mu_\theta(x)\nu(z \mid x)\, dx$ denotes the induced mixture

weights of generated samples. Intuitively, these metrics quantify discrepancies in global energy structure and mass allocation across modes, respectively.

Results. In Table 3 and Figure 3, we find that ADA successfully recovers the full target distribution using the correlated projection observables. In contrast, EA fails even when enforcing moments up to fourth order. Recovering a multimodal distribution from finitely many moments is not possible in general, and the number of moments that need to be enforced scales polynomially with the dimension of the problem, highlighting the fundamental limitations of expectation-based alignment.

4.2 Correlated Small Molecule Observables

We next evaluate whether ADA can align molecular conformational distributions using multiple correlated structural observables, and whether incorporating observables improves alignment on both constrained and held-out quantities. As a controlled proxy for sim-to-real transfer, we consider two simulator fidelities on the MD17 small-molecule benchmark [6], aligning a low-cost GFN2-xTB potential [2] to a higher-fidelity density functional theory (DFT) reference using only observable-level supervision.

We progressively add structural observables to the alignment procedure—including the mean interatomic distance (MD), radius of gyration (R_g), bond lengths, hydrogen-bond distance, and COOH–ester distance (details in Appendix A.2)—to study ADA 's ability to reconcile multiple correlated marginals while preserving underlying molecular structure.

For evaluation, we report two complementary metrics. First, for each observable we compute the Wasserstein-1 distance (W_1) between histograms of generated and target samples, measuring marginal distributional alignment. Beyond marginal comparisons, we evaluate the model's ability to recover free energy surfaces (FES) over pairs of physically meaningful observables. For each observable we bin both the reference and generated samples into a joint 2D histogram, normalize each to a discrete probability distribution, and compute the Jensen–Shannon divergence. We report this over three observable pairs on aspirin, capturing hydrogen-bonding geometry, molecular compactness, and functional group arrangement.

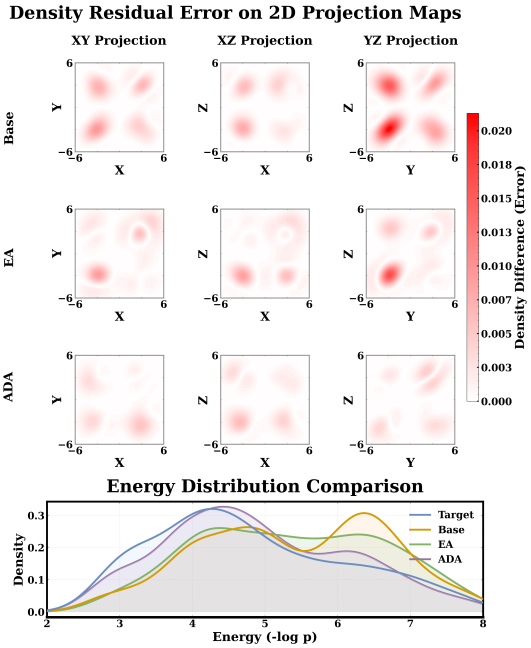

Figure 3: **Comparison between expectation alignment (EA) and ADA on synthetic mixture-of-Gaussians benchmark.** We align a base distribution, defined as a mixture of Gaussians centered at the corners of a cube, with a target distribution with perturbed variances and mixture weights. For the observables, we use pairwise coordinate projections, yielding correlated, multimodal marginals. We report the ℓ_1 pdf residual between target and generated distributions before and after alignment. ADA consistently achieves larger reductions in distributional error on multidimensional observables, even when EA uses fourth moments. This is taken from a kernel density estimate of 2000 samples.

Results. Figure 7 and Table 4 show that ADA consistently improves alignment to the DFT target as additional observables are incorporated. Across all observables, ADA achieves substantially lower W_1 as compared to EA while simultaneously reducing FES JSD, indicating that it preserves molecular correlations while enforcing accurate marginal constraints.

4.3 Noisy and High-Dimensional Cryo-EM Observable with Experimental Data

Finally, we use the coarse-grained fast-folding proteins [28, 25] as a test bed for aligning to experimental data using measurements from the Protein Data Bank (PDB) [4]. We first train a base model to match the Boltzmann distribution induced by molecular dynamics simulations using a classical force

field [28, 25]. Starting from this base model, we then use experimental structures of the Trp-cage (ID 2JOF) and BBL (ID 2WXC) proteins from the PDB as our target distribution [42].

For this experiment, ADA uses cryo-EM images as observables. We simulate the image formation from the reference experimental structures [45], yielding noisy, high-dimensional observables (128×128 pixels; Figure 4 and Figure 9). We choose this evaluation setting since it provides a challenging benchmark that measures ADA's ability to improve a prior derived from classical simulations using a small set (< 30 structures) of real-world structures. Simulating the forward cryo-EM imaging enables us to control the level of noise of the observable and allows us to provide precise evaluation metrics by comparing to the reference experimental structures.

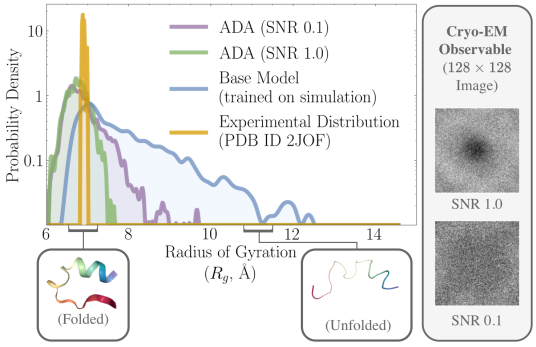

Figure 4: **ADA shifts the simulated distribution to align with the experimentally measured Trp-cage structures by using only noisy, high-dimensional cryo-EM images.** The base generative model is trained on MD simulations using a classical force field, which samples unfolded states more often than the experimental ensemble (2JOF PDB). ADA fine-tunes the base model using the cryo-EM images at SNR 1.0 and 0.1. We evaluate on a held-out observable, the radius of gyration, which is not used during alignment. At both noise levels, ADA removes most of the unfolded tail and shifts the density toward the experimental structures.

For evaluation, we report the W_1 distance to other structural observables to see how ADA improves alignment to held-out observables. To measure ADA's ability to recover the underlying full state distribution from the cryo-EM images, we also measure the maximum Kabsch-aligned RMSD to the experimental amino acid positions. We provide further details in Appendix A.4.

Results. Figure 4 and Table 5 show how ADA aligns the base generative model to the distribution of experimental protein structures using the cryo-EM image observable. As in the molecular experiments, ADA decreases the Wasserstein distance on held-out observables by up to 86%. Beyond observables, ADA also reduces the maximum Kabsch-aligned RMSD to the experimental structures, indicating fewer extreme deviations in the underlying amino acid positions. Even when we decrease the signal-to-noise ratio (SNR) for the cryo-EM observable, ADA improves the base distribution's alignment with the experimental data.

This experiment shows the importance of ADA's ability to perform full *distributional* alignment. At low SNR, matching only the expectation of the cryo-EM observable would just match the average noise of the images, and does not improve alignment (see Table 5). Similarly, imposing a sufficient number of moment constraints for the 128×128 dimensional distribution would be impractical, as shown by the synthetic experiment in Section 4.1. Appropriately scaled, this experiment suggests the potential of ADA to leverage a large dataset of noisy experimental measurements to bridge the simulation-to-experiment gap.

Inferring unknown parameters of the observable forward model. In a real-world imaging setup, the exact parameters of the cryo-EM imaging model might be unknown. We conduct an additional experiment using Trp-cage where ADA must recover part of the observable (forward) model itself, instead of knowing the exact image formation process. We create each cryo-EM image at a signal-to-noise ratio (SNR) drawn from $\mathcal{N}(1,\, 0.2)$. The model is provided only with an informed prior over the SNR, $\mathcal{N}(0.8,\, 0.4)$, since the typical SNR of a cryo-EM measurement is often known but can vary slightly across measurements.

In addition to the ADA objective, the model now variationally infers the SNR distribution, learning a variational posterior over the SNR (regularized toward the prior via a KL term) while aligning to the observable. We formalize this in the ADA framework by considering the underlying states to be tuples of the protein conformation and noise level of the measurement device. On held-out observables, ADA reduces the Wasserstein distance to the experimental radius-of-gyration distribution from 1.07 to 0.25, and from 1.26 to 0.26 for the mean interatomic distance distribution, while the inferred SNR concentrates near its true value (learned $\mathcal{N}(0.98, 0.19)$). We visualize this procedure in Figure 5.

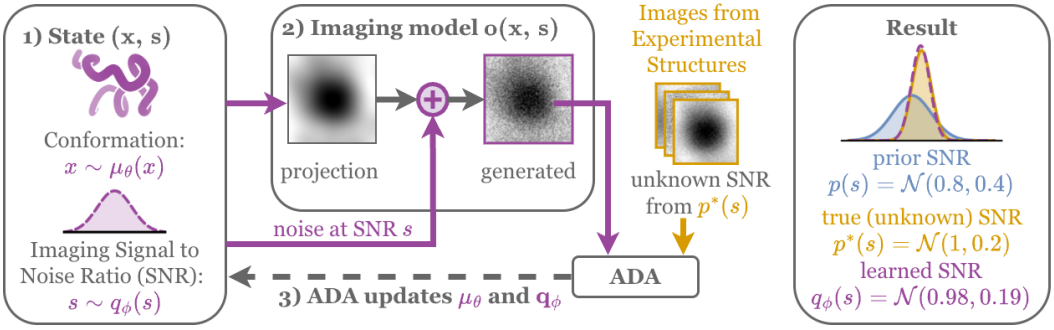

Figure 5: **ADA jointly infer parameters of the observable forward model while aligning to the experimental distribution.** In a more realistic imaging scenario, the parameters of the cryo-EM forward model may not be known exactly. We run an experiment where ADA must infer the SNR of the image formation process while also aligning the distribution of generated conformations with the experimental structures. ADA now receives only an informed prior over the possible SNRs, and we sample a new SNR for each generated image. ADA recovers the correct SNR distribution while still improving the alignment of the generated structures.

5 Conclusion

In this work, we introduced ADA, an iterative algorithm for bridging the gap between simulation and experiment that alternates between distinguishing low- and high-fidelity samples and refining a generative model accordingly. We provided both theoretical guarantees and empirical evidence that ADA can align a classical force field with simulated experimental measurements, and that its performance improves as additional observables are incorporated, even when these observables are correlated.

While our experiments focus on observables derived from a single state, the proposed objective naturally extends to dynamic observables, such as an autocorrelation function, presenting an important direction for future work. Moreover, although we evaluate ADA on widely used benchmarks, these datasets remain a simplified proxy for the scale and complexity encountered at the frontier of biomolecular research. Real experimental data are also often noisy and expensive to obtain, posing additional challenges that must be addressed to enable broader deployment.

At the same time, the observation that ADA improves with an increasing number of observables suggests that scaling ADA with more computational resources and experimental datasets could further improve performance. We also hope that the development of algorithms capable of leveraging experimental measurements at scale will motivate the creation of more standardized and widely accessible experimental datasets. Finally, while our experiments are grounded in computational chemistry, we emphasize that ADA is not domain-specific and provides a general framework for learning from partial observations. By combining structured simulated priors with real-world measurements, ADA represents a step toward models that more faithfully align with the real world.

Acknowledgments. We thank James Bowden and Sanjeev Raja for feedback on the manuscript. We thank Nithin Chalapathi for feedback on the cryo-EM experiments and Michael Psenka for feedback on the proofs. This research was initially supported by Laboratory Directed Research and Development (LDRD) funding under the Department of Energy Contract Number DE-AC02-05CH11231 and later supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under contract number DE-SC0026116, DARPA TIAMAT, and the Toyota Research Institute as part of the Synthesis Advanced Research Challenge.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. URL https://arxiv.org/abs/1701.07875.

[2] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, February 2019. ISSN 1549-9626. doi: 10.1021/acs.jctc.8b01176. URL http://dx.doi.org/10.1021/acs.jctc.8b01176.

[3] David Bartholomew, Martin Knott, and Irini Moustaki. *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley, June 2011. ISBN 9781119970583. doi: 10.1002/9781119970583. URL http://dx.doi.org/10.1002/9781119970583.

[4] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28 (1):235–242, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

[5] Andrea Cesari, Sabine Reißer, and Giovanni Bussi. Using the maximum entropy principle to combine simulations and solution experiments. *Computation*, 6(1):15, February 2018. ISSN 2079-3197. doi: 10.3390/computation6010015. URL http://dx.doi.org/10.3390/computation6010015.

[6] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, 2017. doi: 10.1126/sciadv.1603015. URL https://www.science.org/doi/abs/10.1126/sciadv.1603015.

[7] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, May 1995. ISSN 1520-5126. doi: 10.1021/ja00124a002. URL http://dx.doi.org/10.1021/ja00124a002.

[8] Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control, 2025. URL https://arxiv.org/abs/2409.08861.

[9] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models, 2016. URL https://arxiv.org/abs/1611.03852.

[10] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization, 2016. URL https://arxiv.org/abs/1603.00448.

[11] Paul Fuchs, Stephan Thaler, Sebastien Röcken, and Julija Zavadlav. chemtrain: Learning deep potential models via automatic differentiation and statistical physics, May 2025. ISSN 0010-4655. URL http://dx.doi.org/10.1016/j.cpc.2025.109512.

[12] Abhijeet Sadashiv Gangan, Ekin Dogus Cubuk, Samuel S Schoenholz, Mathieu Bauchy, and NM Anoop Krishnan. Force-field optimization by end-to-end differentiable atomistic simulation, 2025.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

[14] Bin Han and Kuang Yu. Refining potential energy surface through dynamical properties via differentiable molecular simulation, January 2025. ISSN 2041-1723. URL http://dx.doi.org/10.1038/s41467-025-56061-z.

[15] Martin Hanke. Well-posedness of the iterative boltzmann inversion. *Journal of Statistical Physics*, 170(3):536–553, December 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1944-2. URL http://dx.doi.org/10.1007/s10955-017-1944-2.

[16] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning, 2016. URL https://arxiv.org/abs/1606.03476.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

[18] Gerhard Hummer and Jürgen Köfinger. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of Chemical Physics*, 143(24), December 2015. ISSN 1089-7690. doi: 10.1063/1.4937786. URL http://dx.doi.org/10.1063/1.4937786.

[19] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, May 1957. ISSN 0031-899X. doi: 10.1103/physrev.106.620. URL http://dx.doi.org/10.1103/PhysRev.106.620.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[21] Jürgen Köfinger, Lukas S. Stelzl, Klaus Reuter, Cesar Allande, Katrin Reichel, and Gerhard Hummer. Efficient ensemble refinement by reweighting, April 2019. ISSN 1549-9626. URL http://dx.doi.org/10.1021/acs.jctc.8b01231.

[22] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

[23] Christopher Kolloff, Tobias Höppe, Emmanouil Angelis, Mathias Jacob Schreiner, Stefan Bauer, Andrea Dittadi, and Simon Olsson. Minimum-excess-work guidance, 2025. URL https://arxiv.org/abs/2505.13375.

[24] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018. URL https://arxiv.org/abs/1805.00909.

[25] Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, October 2011. ISSN 1095-9203. doi: 10.1126/science.1208351. URL http://dx.doi.org/10.1126/science.1208351.

[26] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, April 1998. ISSN 1520-5207. doi: 10.1021/jp973084f. URL http://dx.doi.org/10.1021/jp973084f.

[27] Sai Advaith Maddipatla, Nadav Bojan, Meital Bojan, Sanketh Vedula, Paul Schanda, Ailie Marx, and Alexander Bronstein. Inverse problems with experiment-guided alphafold. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qzM37nOy3N.

[28] Maciej Majewski, Adrià Pérez, Philipp Thölke, Stefan Doerr, Nicholas E. Charron, Toni Giorgino, Brooke E. Husic, Cecilia Clementi, Frank Noé, and Gianni De Fabritiis. Machine learning coarse-grained potentials of protein thermodynamics. *Nature Communications*, 14 (1), September 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-41343-1. URL http://dx.doi.org/10.1038/s41467-023-41343-1.

[29] Sakib Matin, Alice E. A. Allen, Justin Smith, Nicholas Lubbers, Ryan B. Jadrich, Richard Messerly, Benjamin Nebgen, Ying Wai Li, Sergei Tretiak, and Kipton Barros. Machine learning potentials with the iterative boltzmann inversion: Training to experiment, February 2024. ISSN 1549-9626. URL http://dx.doi.org/10.1021/acs.jctc.3c01051.

[30] Rishwanth Raghu, Axel Levy, Gordon Wetzstein, and Ellen D. Zhong. Multiscale guidance of protein structure prediction with heterogeneous cryo-em data, 2025. URL https://arxiv.org/abs/2506.04490.

[31] Sanjeev Raja, Ishan Amin, Fabian Pedregosa, and Aditi S. Krishnapriyan. Stability-aware training of machine learning force fields with differentiable boltzmann estimators. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=ZckLMGO0s0.

[32] Dirk Reith, Mathias Pütz, and Florian Müller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *Journal of Computational Chemistry*, 24(13):1624–1636, August 2003. ISSN 1096-987X. doi: 10.1002/jcc.10307. URL http://dx.doi.org/10.1002/jcc.10307.

[33] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[34] Henry D. Smith, Nathaniel L. Diamant, and Brian L. Trippe. Calibrating generative models to distributional constraints, 2026. URL https://arxiv.org/abs/2510.10020.

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

[36] A.K. Soper. Empirical potential monte carlo simulation of fluid structure. *Chemical Physics*, 202(2–3):295–306, January 1996. ISSN 0301-0104. doi: 10.1016/0301-0104(95)00357-6. URL http://dx.doi.org/10.1016/0301-0104(95)00357-6.

[37] Wai Shing Tang, David Silva-Sánchez, Julian Giraldo-Barreto, Bob Carpenter, Sonya M. Hanson, Alex H. Barnett, Erik H. Thiede, and Pilar Cossio. Ensemble reweighting using cryo-em particle images. *The Journal of Physical Chemistry B*, 127(24):5410–5421, 2023. ISSN 1520-5207. doi: 10.1021/acs.jpcb.3c01087. URL http://dx.doi.org/10.1021/acs.jpcb.3c01087.

[38] Wai Shing Tang, Ellen D. Zhong, Sonya M. Hanson, Erik H. Thiede, and Pilar Cossio. Conformational heterogeneity and probability distributions from single-particle cryo-electron microscopy, August 2023. ISSN 0959-440X. URL http://dx.doi.org/10.1016/j.sbi.2023.102626.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

[40] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, December 2008. ISSN 1935-8245. doi: 10.1561/2200000001. URL http://dx.doi.org/10.1561/2200000001.

[41] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/bf00992696. URL http://dx.doi.org/10.1007/BF00992696.

[42] Kurt Wüthrich. Nmr with proteins and nucleic acids. *Europhysics News*, 17(1):11–13, 1986. ISSN 1432-1092. doi: 10.1051/epn/19861701011. URL http://dx.doi.org/10.1051/epn/19861701011.

[43] Kailai Xu and Eric Darve. Solving inverse problems in stochastic models using deep neural networks and adversarial training. *Computer Methods in Applied Mechanics and Engineering*, 384:113976, October 2021. ISSN 0045-7825. doi: 10.1016/j.cma.2021.113976. URL http://dx.doi.org/10.1016/j.cma.2021.113976.

[44] Tigany Zarrouk and Miguel A. Caro. Linear-scaling calculation of experimental observables for molecular augmented dynamics simulations, 2025. URL https://arxiv.org/abs/2509.22388.

[45] Ellen D. Zhong, Tristan Bepler, Bonnie Berger, and Joseph H. Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature Methods*, 18(2):176–185, February 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01049-4. URL http://dx.doi.org/10.1038/s41592-020-01049-4.

[46] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1433–1438. AAAI Press, 2008. ISBN 9781577353683.

[47] Jiří Čížek. On the correlation problem in atomic and molecular systems. calculation of wavefunction components in ursell-type expansion using quantum-field theoretical methods. *The Journal of Chemical Physics*, 45(11):4256–4266, December 1966. ISSN 1089-7690. doi: 10.1063/1.1727484. URL http://dx.doi.org/10.1063/1.1727484.

A Experimental Details

Table 1: **Observable Alignment naturally handles multiple observables.** Our method learns the marginal distribution of multiple, potentially correlated observables directly from data via a discriminator. In contrast, guidance-based approaches typically target a single observable and offer no guarantees for aligning multiple (potentially correlated) observables. To match the distribution of a single observable, guidance would need to rely on conditioning on samples from the target observable distribution, whereas Observable Alignment learns the distribution directly from data.

Method	Match Expectation		Match Distribution		Requires Training
	Single	Multiple	Single	Multiple	
Conditional Generative Model	✓	✗	✓	✗	classifier (or CFG)
Expectation Alignment	✓	✓	✗	✗	generator
ADA (ours)	✓	✓	✓	✓	generator + discriminator

We provide further details on our experiments. Unless otherwise specified, we use a Transformer encoder [39] for the generative model and an MLP for our discriminators. We apply rotation augmentation in all our experiments. We used Adam [20] for all of our experiments and normalized both the molecular and observable inputs. We also found that using a gradient penalty to handle the Lipschitz constraint to be more flexible than fixing a spectral norm. We discuss individual experiments below.

A.1 Synthetic Mixture-of-Gaussians

We visualize the synthetic mixture-of-Gaussians benchmark in Figure 6. Although this is a simple distribution in three dimensions, EA struggles to align to the target distribution (see Section 4.1 and Table 3).

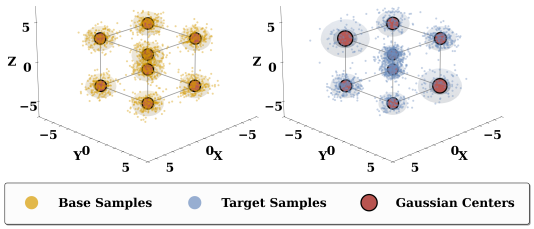

Figure 6: **Synthetic mixture-of-Gaussians benchmark.** The base distribution (left) is an isotropic mixture-of-Gaussians with eight components located at the vertices of a cube, with uniform mixture weights. We define the target distribution (right) by perturbing the variances and mixture weights for the Gaussians (for example, larger variance on top-left to bottom-right diagonal).

Dataset Parameters The synthetic benchmark consists of two three-dimensional mixture-of-Gaussians distributions, both with eight components centered at the vertices of a cube with side length 6.0 Å. We summarize the key parameters in Table 2.

A.2 Small Molecules

In this work, we consider the aspirin system ($C_9H_8O_4$) composed of 21 atoms, with DFT data taken from the MD17 dataset [6]. To characterize the conformational landscape and hydrogen bonding dynamics of aspirin, we compute the following set of observables:

Radius of Gyration The radius of gyration R_g quantifies the spatial extent of the molecular structure and is defined as:

$$R_g^2 = \frac{1}{M} \sum_{i=1}^{N} m_i \|\mathbf{r}_i - \mathbf{r}_{\text{COM}}\|^2 \tag{10}$$

where m_i and $\mathbf{r}_i$ are the mass and position of atom i, $N = 21$ is the total number of atoms, $M = \sum_{i=1}^{N} m_i$ is the total mass, and $\mathbf{r}_{\text{COM}}$ is the center of mass. This observable captures the overall

Table 2: **Synthetic mixture-of-Gaussians parameters.** Both distributions have eight Gaussian components positioned at the cube vertices $\{(\pm 3, \pm 3, \pm 3)\}$ Å. The base distribution is uniform, while the target distribution has heterogeneous variances and mixture weights.

Parameter	**Base (Uniform)**	**Target (Reweighted)**
Cube Side Length	6.0 Å	6.0 Å
Number of Components	8	8
Component Positions	Cube vertices	Cube vertices
Mixture Weights:		
Diagonal Pair 1	0.25	0.40
Diagonal Pair 2	0.25	0.30
Diagonal Pair 3	0.25	0.10
Diagonal Pair 4	0.25	0.20
Variances (σ^2, Å^2):		
Vertex 1: $(+3, +3, +3)$	0.5	0.3
Vertex 2: $(+3, +3, -3)$	0.5	0.5
Vertex 3: $(+3, -3, +3)$	0.5	1.0
Vertex 4: $(+3, -3, -3)$	0.5	0.4
Vertex 5: $(-3, +3, +3)$	0.5	0.6
Vertex 6: $(-3, +3, -3)$	0.5	1.2
Vertex 7: $(-3, -3, +3)$	0.5	0.5
Vertex 8: $(-3, -3, -3)$	0.5	0.3

Table 3: **Distributional alignment metrics on mixture of Gaussians.** We compare alignment using ADA against EA style approaches using an observable set of 2-dimensional projection maps. Up until the fourth order, moment-based matching is unable to recover the distribution empirically.

Method	Energy W2 ↓	Cluster KL ↓
Base	1.10	0.12
ADA (ours)	**0.05**	**0.01**
EA (1st)	1.10	0.12
EA (2nd)	0.90	0.02
EA (3rd)	0.90	0.02
EA (4th)	0.70	0.05

compactness of the molecule, with lower values indicating more compact conformations and higher values indicating more extended structures. Units: .

Mean Interatomic Distance The mean interatomic distance $\langle d \rangle$ provides a global measure of the molecular size:

$$\langle d \rangle = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|\mathbf{r}_i - \mathbf{r}_j\| \tag{11}$$

This quantity averages over all pairwise atomic distances and serves as a complementary measure to R_g, being sensitive to the overall distribution of atoms rather than their distance from the center of mass. Units: Å.

Bond Lengths We monitor all covalent bond lengths d_{ij} between bonded atoms i and j:

$$d_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\| \tag{12}$$

These observables track local geometric distortions and vibrational modes within the molecular framework. For aspirin, we focus particularly on the bonds within the carboxylic acid (COOH) and ester functional groups. Units: Å.

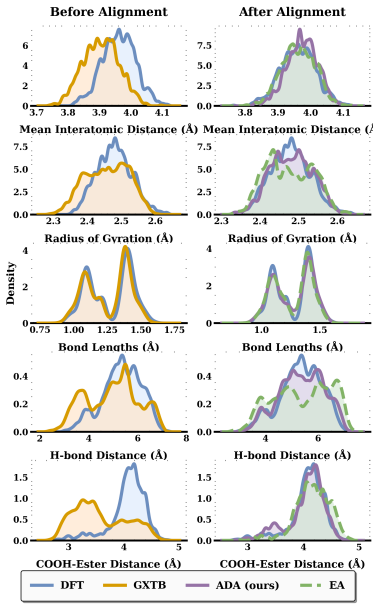

Figure 7: **Comparison of expectation alignment (EA) and ADA across all evaluated observables on the MD17 aspirin system.** While EA aligns first-order statistics, ADA more accurately captures higher-order structure, yielding closer agreement with the full target observable distributions.

Hydrogen Bond Distance The intramolecular hydrogen bond distance d_{HB} is defined as the distance between the hydroxyl hydrogen of the carboxylic acid group and the carbonyl oxygen of the ester group:

$$d_{\text{HB}} = \|\mathbf{r}_{\text{H(COOH)}} - \mathbf{r}_{\text{O(ester)}}\| \tag{13}$$

This observable is crucial for characterizing the intramolecular hydrogen bonding that stabilizes certain aspirin conformers. Shorter distances indicate stronger hydrogen bonding interactions. Units: Å.

COOH-Ester Distance The COOH-ester distance d_{CE} measures the separation between the center of mass of the carboxylic acid group and the center of mass of the ester group:

$$d_{\text{CE}} = \|\mathbf{r}_{\text{COM}}^{\text{COOH}} - \mathbf{r}_{\text{COM}}^{\text{ester}}\| \tag{14}$$

This collective variable captures the relative orientation and proximity of the two functional groups, which is directly related to the molecule's ability to form the intramolecular hydrogen bond. Units: Å.

A.3 Free Energy Surfaces

To explore the conformational landscape of aspirin, we compute two-dimensional free energy surfaces (FES) as functions of pairs of these observables:

- **FES A**: $F(d_{\text{HB}}, R_g^2)$ – This surface relates hydrogen bonding strength to molecular compactness, revealing how intramolecular hydrogen bonds affect the overall molecular geometry.
- **FES B**: $F(d_{\text{CE}}, d_{\text{HB}})$ – This surface directly connects the functional group separation to the hydrogen bond distance, capturing the geometric requirements for hydrogen bond formation.
- **FES C**: $F(\xi_1, \xi_2)$ – This surface explores the functional group geometry using appropriate collective variables that characterize the relative orientation and configuration of the COOH and ester moieties.

We take the Jensen-Shannon divergence of the free energy surfaces by a standard binning procedure of the 2-dimensional distributions.

Table 4: **ADA improves in alignment to the target distribution when given access to an increasing number of observables.** We train a base model on constant-temperature molecular dynamics simulations using the semi-empirical GFN2-xTB potential. This model is then aligned with the distribution induced by a more accurate DFT potential using structural observables. Each column reports the distance to the reference DFT observable distribution. Moving down the rows of the table adds observables to the alignment procedure, with the bottom row representing aligning with the full observable set: mean interatomic distance (MD), radius of gyration (R_g), bond lengths, hydrogen-bond distance, and COOH–ester distance. The upper-triangular part of the left side of the table represents improvements on held-out observables (highlighted in purple). We additionally report improvement on Jensen-Shannon divergences on three fully held-out free energy surfaces: (MD, Rg), (bond length, H-bond distance), and (H-bond distance, COOH–ester distance). These surfaces are not used during training (indicated by the purple highlight) and therefore measure preservation of joint molecular structure beyond the enforced observable marginals. Compared to Expectation Alignment (EA), ADA better aligns to the target distribution (indicated with bold), especially with additional observables.

	Observable marginals $W_1 \downarrow$										Held-out FES JSD $\downarrow$					
Constraint	MD		R_g		Bond		Hbond		COOH		A		B		C	
	Ours	EA	Ours	EA	Ours	EA	Ours	EA	Ours	EA	Ours	EA	Ours	EA	Ours	EA
Base	0.06		0.02		0.018		0.40		0.50		0.22		0.44		0.38	
+ MD	**0.008**	0.012	0.020	**0.009**	**0.007**	0.010	**0.25**	0.41	**0.37**	0.39	0.23	**0.20**	**0.40**	0.41	**0.35**	**0.35**
+ RoG	**0.009**	0.012	0.010	**0.009**	**0.010**	0.011	**0.30**	0.41	**0.43**	0.48	0.20	**0.19**	0.42	**0.40**	0.35	**0.34**
+ Bond	**0.004**	0.019	**0.005**	0.009	**0.004**	0.009	**0.21**	0.36	**0.46**	0.48	**0.15**	0.20	**0.41**	0.43	**0.36**	**0.36**
+ Hbond	**0.007**	0.017	**0.009**	0.010	**0.006**	0.011	**0.06**	0.19	**0.39**	**0.39**	**0.13**	0.36	**0.39**	0.43	**0.33**	0.36
+ COOH	**0.009**	**0.009**	**0.002**	0.012	**0.006**	0.011	**0.06**	0.31	**0.04**	0.06	**0.08**	0.17	**0.13**	0.20	**0.09**	0.19

Table 5: **ADA aligns a generative model trained on simulated data from a classical force field with experimental protein structures from the PDB using, noisy, high-dimensional cryo-EM images.** We pre-train a generative model on samples of fast-folding proteins obtained from molecular dynamics simulations using a classical force field [28]. Using only the noisy cryo-EM images, ADA improves alignment with the underlying experimental distribution of structures, as measured by the Wasserstein distance to other, held-out observables. The maximum Kabsch-aligned RMSD to the experimental structures also decreases, indicating better distributional alignment at level of amino acid positions.

	W_1 **to Experiment (↓)**		**Max RMSD**
	MeanDist	$\mathbf{R_g}$	**(nm, ↓)**
Trp-Cage Base	1.26	1.07	1.18
EA (SNR 1.0, 1st order moments)	1.28	1.02	1.17
EA (SNR 1.0, 1st+2nd order moments)	1.25	1.08	1.17
ADA (SNR 1.0)	0.17	0.21	0.91
ADA (SNR 0.1)	0.28	0.36	0.96
BBL Base	1.36	1.08	1.76
EA (SNR 1.0, 1st order moments)	1.23	0.98	1.73
EA (SNR 1.0, 1st+2nd order moments)	1.26	1.01	1.71
ADA (SNR 1.0)	0.21	0.18	1.38
ADA (SNR 0.1)	0.27	0.33	1.36

A.4 Proteins

In this work, we considered two proteins: Trp-Cage and BBL with 20 and 47 amino acids, respectively. We used the same experimental structures, derived by NMR, used by Majewski et al. [28] with the following PDB IDs: 2JOF for Trp-cage (29 structures) and 2WXC for BBL (21 structures). We first pre-trained the base model using 100,000 samples of each of the proteins obtained by running MD simulations with a classical force field [28]. We then used ADA individually on each protein to match the experimental distribution of structures by using cryo-EM images as observables. We note that the

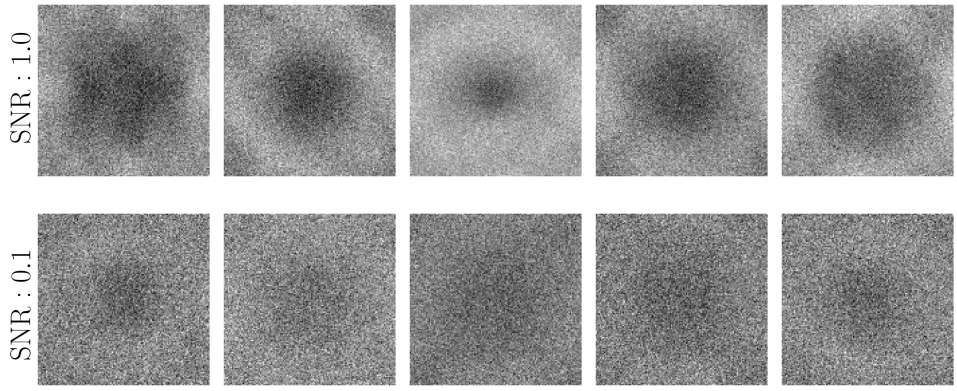

Figure 9: **Example cryo-EM images of Trp-cage used by ADA to align with the experimental distribution of protein structures.**

simulation data was coarse-grained using the alpha-carbon [28], so we applied the same alpha-carbon coarse-graining to the experimental data.

We simulate the cryo-EM images to provide precise control over the image formation, allowing us to evaluate ADA's robustness to noise. We build off the code provided by Zhong et al. [45] to generate 128×128 cryo-EM images. We use a standard deviation of 2.5 for the Gaussian kernels for volume generation, an amplitude of 0.1, a defocus between 0.8 and 2.5 μm , a b-factor between 1 and 100Å, and a pixel size of 0.3Å. We also run experiments with a signal-to-noise ratio (SNR) of 1.0 and 0.1. We generate a new cryo-EM image each batch, mimicking a real cryo-EM imaging setup where there might be multiple different images of similar underlying structures. See Figure 9 for example images.

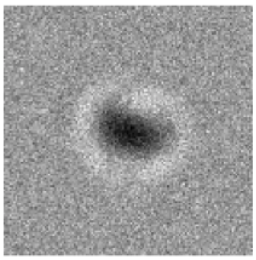

Figure 8: **Since we worked with small proteins, we ran an additional experiment with a defocus between 0.2 and 0.5 μm and a pixel size of 0.5Å to emulate different imaging scenarios.**

We also run an additional experiment with a defocus between 0.2 and 0.5 μm and a pixel size of 0.5Å since experimental images for small proteins can require different simulation parameters than larger ones (see Figure 8). We see very similar results to the parameters described above, with a decrease in Wasserstein distance of more than 80% to held out observables compared to the base model.

Note that for this experiment, we use a convolutional neural network (CNN) for the discriminator. The convolutional encoder consists of 3 convolutional blocks, each containing a convolution (kernel size 3, padding 1) and max pooling (stride 2). The channels are $1, 32, 64, 128$. An adaptive average pooling layer reduces spatial dimensions to a fixed size of 2×2, followed by a linear projection. We report further hyperparameters in Table 6.

A.5 Computational Resources

All experiments were run on a single GPU. We used both A6000s and H100s to run our experiments. Training the base models and applying ADA took less than 24 GPU hours across all of our experiments. Interestingly, performance kept improving with ADA as the models kept training longer, especially with more observables and more complicated systems. This suggests that ADA could benefit and take advantage of further computation on larger experimental datasets.

B Proofs

We provide the full proofs of the theorems presented in the main text.

Table 6: **Experiment Hyperparameters.** Note that discriminator steps refers to the number of gradient steps taken between each update of the generative model, and the discriminator weight refers to the weighting of the discriminator term relative to the KL term in the generative model update step.

Hyperparameter	Value		
	Synthetic	Small Molecules	Proteins
Base Model:			
Layers	MLP(512, 512, 512, 512)	8	8
Heads	N/A	12	12
Hidden Dimension	N/A	384	384
Number of Euler Steps (sampling)	N/A	45	35
Discriminator Model:			
Hidden Dimension	512	512	(CNN, see A.4)
Layers	2	2	(CNN, see A.4)
ADA:			
Discriminator Weight	128.0	4096.0	512
Discriminator Steps	1	1	3
Gradient Penalty Weight	1000	1000	1×10^6
Training:			
Generative Model LR (Pre-training, fine-tuning)	$1 \times 10^{-3}, 1 \times 10^{-5}$	$1 \times 10^{-3}, 1 \times 10^{-5}$	$1 \times 10^{-3}, 1 \times 10^{-5}$
Discriminator LR	1×10^{-3}	1×10^{-3}	1×10^{-3}
Weight Decay (Pre-training, fine-tuning)	$1 \times 10^{-4}, 1 \times 10^{-8}$	$1 \times 10^{-4}, 1 \times 10^{-8}$	$1 \times 10^{-4}, 1 \times 10^{-8}$
# Gradient Steps (Pre-training, fine-tuning)	10,000, 10,000	500,000, 10,000	300,000, 1,000
Batch Size	1024	64	128

B.1 Proof of Theorem B.1

Theorem B.1 (Existence and uniqueness of the saddle point)**.** *Assume X is a compact Polish space and that μ_{base} has full support on X. Additionally, allow $o^{(i)}$ to be continuous for $i \in I$. For any $\beta \in \mathbb{R}$, Equation (6) admits a saddle point $(\mu^*, (f^{(i)})^*)$, where μ^* is unique. Consequently, the recovered distribution is well defined. We denote this solution by $\mu^* \in \operatorname{sol} \mathcal{L}(\beta)$.*

Proof. The proof is straightforward Sion's theorem. We start by showing that the space of probability measures is compact with respect to the weak topology. We then note that Lipschitz functions are a convex subset of the continuous functions, equipped with the supremum norm. With this, both our probability measures (generators) and functions (discriminators) live on convex subsets of linear topological vectors spaces. We then use basic facts about the continuity of our Lagrangian to justify the existence of a unique saddle point μ^*.*

We first equip the space of probability measures with the weak topology; allowing $\mu \mapsto \langle \mu, f \rangle = \int f \, d\mu : \mathcal{P} \to \mathbb{R}$ to be continuous; resulting in $\mathcal{P}_1$ being a compact convex subset of a linear topological vector space with respect to the weak* topology.*

Let $\mathcal{P} = \mathcal{M}(X)$ denote the space of Radon measures on $(X, \mathcal{F})$, and let $\mathcal{P}_1 \subset \mathcal{P}$ be the subset of probability measures.

Allow $C_0(X)$ to be the set of continuous functions on X vanishing at infinity. As X is compact we have that $C_0(X) = C(X)$; the set of all continuous functions.

We equip $\mathcal{P}$ with the weak* topology induced by $C(X)^*$, i.e., the coarsest topology such that for all $f \in C(X)$,

$$\mu \mapsto \langle \mu, f \rangle := \int f \, d\mu,$$

is continuous. We have that $\mathcal{P}$ is a linear topological vector space, and $\mathcal{P}_1$ is a convex subset of $\mathcal{P}$.

We note that probability measures are known to be compact with respect to the weak*. $C_0(X)$ is a Banach space, thus the closed TV unit ball $B_1 = \{\mu : \|\mu\|_{TV} \leq 1\}$ is compact in weak*. As probability measures are closed in weak*, and $\mathcal{P}_1 \subset B_1$ we get compactness of $\mathcal{P}_1$ in weak*.

We further define the set of our rewards $\mathcal{F}^m$ such that it is a convex subset of a linear topological vector space.

Allow $f^{(i)} \in \operatorname{LIP}_1(O^i) : O^{(i)} \to \mathbb{R}$ to be the discriminator associated with observable i. By considering the set of continuous functions $C(O^{(i)})$ under the supremum norm, we find that $\mathcal{F}_i \subset C(O^{(i)})$ is a convex subset of a linear topological vector space.

In our setup, we consider a vector of finite discriminators, one per observable. We denote this by: $\mathcal{F}^m = \prod_{i\in I}(F_i)$, which is additionally a convex subspace of $\prod_{i\in I} C(O^{(i)})$.

Thus $\mathcal{F}^m \subset \prod_{i\in I} C(O^{(i)})$ is a convex subset linear topological vector space.

We now finish, by showing that we can invoke Sion's theorem using the compactness of $\mathcal{P}_1$, yielding the existence of a saddle point which is unique in μ^*.

The Lagrangian is given by

$$\mathcal{L}(\mu, (f^{(i)})) = \beta \sum_{i\in I} \mathbb{E}_{o^{(i)}_{\#}((\mu-\nu))}[f^{(i)}] - D_{\mathrm{KL}}(\mu\|\mu_{base}).$$

First, we demonstrate that $\mu \mapsto \mathcal{L}(\mu, (f^{(i)}))$ is upper semi-continuous in μ, and strictly concave.

The term $-D_{\mathrm{KL}}(\mu\|\mu_{\text{base}})$ is upper semi-continuous in the weak* topology and strictly concave in μ.

Likewise, for fixed $(f^{(i)})$, the mapping $\mu \mapsto \sum_{i\in I} \mathbb{E}_{o^{(i)}_{\#}((\mu-\nu))}[f^{(i)}]$ is continuous in the weak* topology, as we have a finite sum of continuous functions.

Next, we demonstrate that $(f^{(i)}) \mapsto \mathcal{L}(\mu, (f^{(i)}))$ is lower semi-continuous in $(f^{(i)})$, and convex.

For fixed $(\mu - \nu)$:

$$(f^{(i)}) \mapsto \sum_{i\in I} \mathbb{E}_{o^{(i)}_{\#}(\mu-\nu)}[f^{(i)}]$$

is continuous on $\mathcal{F}^m$ with respect to the product topology.

This is because $\pi_i : (f(i)) \mapsto f^{(i)}$ is continuous and integration with respect to a finite measure $f(i) \mapsto \langle \gamma, f(i)\rangle$ is continuous, making $\sum_{i\in I} \mathbb{E}_{o^{(i)}_{\#}((\mu-\nu))}[f^{(i)}] = [\sum_{i\in I} \langle o^{(i)}_{\#}((\mu-\nu)), \pi_i(\cdot)\rangle](f^{(i)})$ continuous in $(f^{(i)})$.

Hence, we achieve lower semi-continuity in $(f^{(i)})$ trivially.

Convexity comes naturally from linearity.

We now invoke Sion's theorem, as we have proved the constituent lemmas.

Therefore, $\mathcal{L}$ is concave and upper semi-continuous in μ, and convex and lower semi-continuous in $(f^{(i)})$.

Sion's theorem applies, yielding

$$\max_{\mu} \inf_{(f^{(i)})} \mathcal{L}(\mu, (f^{(i)})) = \inf_{(f^{(i)})} \max_{\mu} \mathcal{L}(\mu, (f^{(i)})).$$

As $\mathcal{P}_1$ is compact, we get attainment of a maximizer μ^*. Uniqueness of μ^* follows from the strict concavity of $-D_{\mathrm{KL}}(\mu\|\mu_{\text{base}})$ in μ.

□

B.2 Asymptotic Convergence in Wasserstein

Theorem B.2 (Asymptotic convergence in Wasserstein)**.** *Assume $D_{\mathrm{KL}}(\nu\|\mu_{base}) < \infty$. Let $(\mu^*_\beta, (f^{(i)}_\beta)^*) \in \operatorname{sol}\mathcal{L}\big(\beta\big)$ for $\beta \in \mathbb{R}^+$. Then, as $\beta \to \infty$, we have*

$$d^{(i)}(\mu^*_\beta, \nu) \to 0 \quad \textit{for all } i \in I.$$

Proof. As $D_{\mathrm{KL}}(\nu\|\mu_{base}) < \infty$ there exists a constant $B \in \mathbb{R}$ such that

$$D_{\mathrm{KL}}(\nu\|\mu_{base}) \le B.$$

Since ν is feasible for the maximization problem, optimality of μ^*_β implies

$$\begin{aligned} \mathcal{L}(\mu^*_\beta, (f^{(i)}_\beta)^*; \beta) &\ge \mathcal{L}(\nu, (f^{(i)}_\beta)^*; \beta) \\ &= -D_{\mathrm{KL}}(\nu\|\mu_{base}) \ge -B. \end{aligned} \tag{15}$$

On the other hand, by definition of the saddle point,

$$\begin{aligned}\mathcal{L}(\mu_\beta^*, (f_\beta^{(i)})^*; \beta) &= -D_{\mathrm{KL}}(\mu_\beta^* \| \mu_{base}) - \beta \sum_{i \in I} d^{(i)}(\mu_\beta^*, \nu) \\ &\leq -\beta \sum_{i \in I} d^{(i)}(\mu_\beta^*, \nu),\end{aligned} \tag{16}$$

since the KL term is nonnegative.

Combining the two inequalities yields

$$\beta \sum_{i \in I} d^{(i)}(\mu_\beta^*, \nu) \leq B.$$

In particular, for each $i \in I$,

$$d^{(i)}(\mu_\beta^*, \nu) \leq \frac{B}{\beta},$$

which implies

$$d^{(i)}(\mu_\beta^*, \nu) \to 0 \quad \text{as } \beta \to \infty.$$

□

B.3 Weak Limit Converges in Constraint Set

Theorem B.3 (Asymptotic convergence to the constraint set)**.** *Let $(\mu_\beta^*)_{\beta>0}$ be the sequence of saddle-point solutions from Theorem B.2. Suppose that $d^{(i)}(\mu_\beta^*, \nu) \to 0$ for all $i \in I$. Then any weak limit point $\bar{\mu}$ of (μ_β^*) satisfies*

$$(o^{(i)})_\# \bar{\mu} = (o^{(i)})_\# \nu \quad \textit{for all } i \in I,$$

that is, $\bar{\mu} \in \mathcal{M}_o$.

Proof. Since $d^{(i)}(\mu_\beta^*, \nu) \to 0$, we have

$$W_1\big((o^{(i)})_\# \mu_\beta^*, (o^{(i)})_\# \nu\big) \to 0$$

for each i. Because W_1 metrizes weak convergence on probability measures over Polish spaces with finite first moment, it follows that

$$(o^{(i)})_\# \mu_\beta^* \Rightarrow (o^{(i)})_\# \nu.$$

By continuity of pushforward under weak convergence, any weak limit point $\bar{\mu}$ therefore satisfies $(o^{(i)})_\# \bar{\mu} = (o^{(i)})_\# \nu$ for all i, completing the proof. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the introduction and abstract are backed up by experimental and theoretical evidence.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limiatations throughout the text and in the conclusions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We clearly state assumptions for proofs and detail the full proofs in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The algorithm and implementation details are sufficiently provided in Section 2, Appendix A, and Appendix A.4, describing the architecture, data generation, hyperparameters, and practical implementation tricks. The algorithm could reasonably and faithfully reproduced from the provided descriptions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use publicly available datasets and will release the full code upon the publication of the work.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe experimental details throughout the text and in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where appropriate, we run experiments on multiple systems and settings to ensure validity of our results. We do not have the computational resources at the moment to run multiple seeds for all experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the computational resources used in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We fully acknowledge and comply with the Neurips code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work's purpose it to generally advance algorithms for scientific ML. We describe the relevant impact on the field throughout the work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not train large models and use publicly available data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the appropriate creators of the used datasets and models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We document our experiments throughout the paper and will release code to rerun experiments upon publication.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.