Energy Matching: Unifying Flow Matching and Energy-Based Models for Generative Modeling

Michal Balcerak
University of Zurich
michal.balcerak@uzh.ch

Tamaz Amiranashvili
University of Zurich
Technical University of Munich

Antonio Terpin
ETH Zurich

Suprosanna Shit
University of Zurich

Lea Bogensperger
University of Zurich

Sebastian Kaltenbach
Harvard University

Petros Koumoutsakos
Harvard University

Bjoern Menze
University of Zurich

Abstract

Current state-of-the-art generative models map noise to data distributions by matching flows or scores. A key limitation of these models is their inability to readily integrate available partial observations and additional priors. In contrast, energy-based models (EBMs) address this by incorporating corresponding scalar energy terms. Here, we propose *Energy Matching*, a framework that endows flow-based approaches with the flexibility of EBMs. Far from the data manifold, samples move from noise to data along irrotational, optimal transport paths. As they approach the data manifold, an entropic energy term guides the system into a Boltzmann equilibrium distribution, explicitly capturing the underlying likelihood structure of the data. We parameterize these dynamics with a single time-independent scalar field, which serves as both a powerful generator and a flexible prior for effective regularization of inverse problems. The present method substantially outperforms existing EBMs on CIFAR-10 and ImageNet generation in terms of fidelity, while retaining simulation-free training of transport-based approaches away from the data manifold. Furthermore, we leverage the flexibility of the method to introduce an interaction energy that supports the exploration of diverse modes, which we demonstrate in a controlled protein generation setting. This approach learns a scalar potential energy, without time conditioning, auxiliary generators, or additional networks, marking a significant departure from recent EBM methods. We believe this simplified yet rigorous formulation significantly advances EBMs capabilities and paves the way for their wider adoption in generative modeling in diverse domains.

1 Introduction

Generative models learn to map from a simple, easy-to-sample distribution, such as a Gaussian, to a desired data distribution. They do so by approximating the optimal transport (OT) map—such as in flow matching [Lipman et al., 2023, Liu et al., 2023, Albergo and Vanden-Eijnden, 2023]—or through iterative noising and denoising schemes, such as in diffusion models [Ho et al., 2020, Song et al., 2021]. In addition to being highly effective in sample generation, diffusion- and flow-based models have also been used as priors to regularize poorly posed inverse problems [Chung et al., 2023, Mardani et al., 2024, Ben-Hamu et al., 2024]. However, these models do not explicitly capture the unconditional data score and instead model the score of smoothed manifolds at different noise levels. The measurement likelihood, on the other hand, is not tractable on these noised manifolds. As a

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

https://doi.org/10.52202/085713-0291

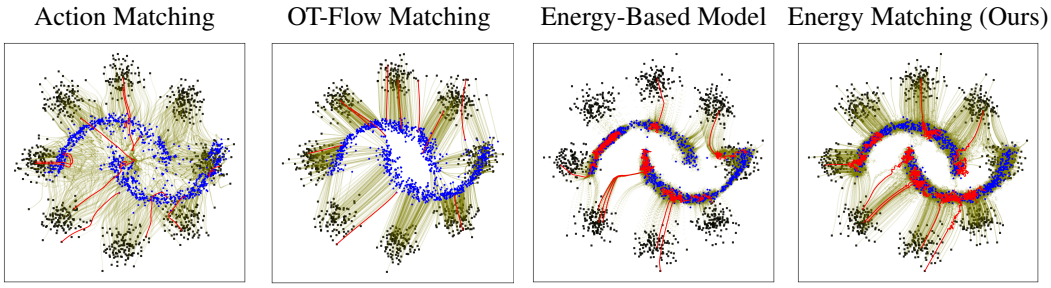


Figure 1: Trajectories (green lines) of samples traveling from a noise distribution (black dots; here, a Gaussian mixture model) to a data distribution (blue dots; here, two moons as in [Tong et al., 2023]) under four different methods: Action Matching [Neklyudov et al., 2023], Flow Matching (OT-CFM) [Tong et al., 2023], EBMs trained via contrastive divergence [Hinton, 2002], and our proposed Energy Matching. We highlight several individual trajectories in red to illustrate their distinct behaviors. Both Action Matching and Flow Matching learn time-dependent transports and are not trained for traversing the data manifold. Conversely, EBMs and Energy Matching are driven by time-independent fields that can be iterated indefinitely, allowing trajectories to navigate across modes. While samples from EBMs often require additional steps to equilibrate (see, e.g., the visible mode collapses that slow down sampling from the data manifold), Energy Matching directs samples toward the data distribution in "straight" paths, without hindering the exploration of the data manifold.

result, existing approaches repeatedly shuttle between noised and data distributions, leading to crude approximations of complex, intractable terms Daras et al. [2024]. For example, DPS [Chung et al., 2023] approximates an intractable integral using a single sample. More recently, D-Flow [Ben-Hamu et al., 2024] optimizes initial noise by differentiating through the simulated trajectory. To the best of our knowledge, these models lack a direct way to navigate the data manifold in search of the optimal solution without repeatedly transitioning between noised and data distributions.

EBMs [Hopfield, 1982, Hinton, 2002, LeCun et al., 2006] provide an alternative approach for approximating the data distribution by learning a scalar-valued function $E(x)$ that specifies an *unnormalized* density $p(x) \propto \exp(-E(x))$. Rather than explicitly mapping noise samples onto the data manifold, EBMs assign low energies to regions of high data concentration and high energy elsewhere. This defines a Boltzmann distribution from which one can sample, for example, via *Langevin sampling*. In doing so, EBMs explicitly retain the likelihood information in $E(x)$. This likelihood information can then be used in conditional generation (e.g., to solve inverse problems), possibly together with additional priors simply by adding their energy terms [Du and Mordatch, 2019]. Moreover, direct examination of local curvature on the data manifold—allows the computation of local intrinsic dimension (LID) (an important proxy for data complexity)—whereas diffusion models can only approximate such curvature in the proximity of noise samples.

Despite the theoretical elegance of using a *single*, time-independent scalar energy, practical EBMs have historically suffered from poor generation quality, falling short of the performance of diffusion or flow matching models. Traditional methods [Song and Kingma, 2021] for training EBMs, such as contrastive divergence via Markov chain Monte Carlo (MCMC) or local score-based approaches [Song and Ermon, 2019], often fail to adequately explore the energy landscape in high-dimensional spaces, leading to instabilities and mode collapse. Consequently, many methods resort to time-conditioned ensembles [Gao et al., 2021], hierarchical latent ensembles [Cui and Han, 2024], or combine EBMs with separate generator networks trained in cooperation [Guo et al., 2023, Zhang et al., 2024, Yoon et al., 2024], thereby requiring significantly higher parameter counts and training complexity.

https://doi.org/10.52202/085713-0291

Contributions. In this work, we propose *Energy Matching*, a two-regime training strategy that combines the strengths of EBMs and flow matching; see Figure 1.

When samples lie far from the data manifold, they are efficiently transported toward the data. Once near the data manifold, the flow transitions into Langevin steps governed by an internal energy component, enabling precise exploration of the Boltzmann-like density well around the data distribution. This straightforward approach produces a *time-independent* scalar energy field whose gradient both accelerates sampling and shapes the final density well—via a contrastive objective that directly learns the score at the data manifold—yet remains efficient and stable to train. Empirically, our method significantly outperforms existing EBMs on both CIFAR-10 and ImageNet generation in terms of fidelity, and compares favorably to flow-matching and diffusion models—without auxiliary generators or time-dependent EBM ensembles.

Our proposed process complements the advantages of flow matching with an explicit likelihood modeling, enabling traversal of the data manifold without repeatedly shuffling between noise and data distributions. This simplifies both inverse problem solving and controlled generations under a prior. In addition, to encourage diverse exploration of the data distribution, we showcase how repulsive interaction energies can be easily and effectively incorporated, with an application to conditional protein generation. Finally, we also showcase how analyzing the learned energy reveals insight on the LID of the data with fewer approximations than diffusion models.[a]

[a]Code repository: `https://github.com/m1balcerak/EnergyMatching`

2 Energy matching

In this section, we show how a scalar potential $V(x)$ can simultaneously provide an optimal-transport-like flow from noise to data while also yielding a Boltzmann distribution that explicitly captures the unnormalized log-likelihood of the data.

The Jordan–Kinderlehrer–Otto (JKO) scheme. The starting point of our approach is the JKO scheme [Jordan et al., 1998], which is the basis of the success of numerous recent generative models [Xu et al., 2023, Terpin et al., 2024, Choi et al., 2024]. The JKO scheme describes the discrete-time evolution of a probability distribution ρ_t along energy-minimizing trajectories in the Wasserstein space,

$$\rho_{t+\Delta t} = \arg\min_{\rho} \underbrace{\frac{W_2^2(\rho, \rho_t)}{2\Delta t}}_{\text{Transport Cost}} + \underbrace{\int V_\theta(x) \mathrm{d}\rho(x)}_{\text{Potential Energy}} + \underbrace{\varepsilon(t) \int \rho(x) \log \rho(x) \mathrm{d}x}_{\text{Internal Energy (-Entropy)}}. \tag{1}$$

Here, θ denotes the learnable parameters of the scalar potential $V_\theta(x)$, and $\varepsilon(t)$ is a temperature-like parameter tuning the entropic term. The transport cost is given by the Wasserstein distance,

$$W_2^2(\rho, \rho_t) = \min_{\gamma \in \Gamma(\rho, \rho_t)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\gamma(x, y), \tag{2}$$

where $\Gamma(\rho, \rho_t)$ is the set of couplings between ρ and ρ_t, i.e., the set of probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ρ and ρ_t. Here, d is the dimensionality of the data. Henceforth, we call OT coupling any γ_t that yields the minimum in (2). When $\gamma_t = (\mathrm{id}, T)_\# \rho$, i.e., it is the *pushforward* of the map $x \mapsto (x, T(x))$ for some function T, we say that T is an OT map from ρ to ρ_t.

Differently from most literature, we consider $\varepsilon(t)$ to be dependent on time and study the behavior of Equation (1) as $t \to \infty$. To fix the ideas, consider, for instance, a linear scheduling:

$$\varepsilon(t) = \begin{cases} 0, & 0 \le t < \tau^*, \\ \varepsilon_{\max} \frac{t - \tau^*}{1 - \tau^*}, & \tau^* \le t < 1, \\ \varepsilon_{\max}, & t \ge 1. \end{cases} \tag{3}$$

First-order optimality conditions. Following Terpin et al. [2024], we analyze (1) at each time t via its first-order optimality conditions [Lanzetti et al., 2024, 2025]. These conditions characterize the properties of the desired solution and thus represent the optimization goal:

https://doi.org/10.52202/085713-0291

$$\frac{1}{\Delta t}(x-y)+\nabla_x V_\theta(x)+\varepsilon(t)\nabla_x \log \rho_{t+\Delta t}(x)=0, \qquad (x,y)\in \mathrm{supp}(\gamma_t) \tag{4}$$

where γ_t is an OT plan between the distributions $\rho_{t+\Delta t}$ and ρ_t and $\mathrm{supp}(\gamma_t)$ is the support of γ_t. That is, this condition has to hold for all pairs of points in the support of $\rho_{t+\Delta t}$ and ρ_t that are coupled by OT. Intuitively, analyzing (4) provides us with two key insights:

1. For times $t<\tau^*$, $\varepsilon(t)=0$ and (4) becomes
$$\frac{1}{\Delta t}(x-y)+\nabla_x V_\theta(x)=0 \qquad (x,y)\in \mathrm{supp}(\gamma_t). \tag{5}$$
That is, the system is in an OT, flow-like, regime.
2. Near the data manifold, which we aim at modeling with the equilibrium distribution ρ_{eq} of (1), $\rho_{t+\Delta t}\approx\rho_{\text{eq}}$ and, thus, for $t\gg 1$, $x\approx y$ for all $(x,y)\in\mathrm{supp}(\gamma_t)$. Then, we can simplify (4) as
$$\varepsilon_{\max}\nabla_x \log\rho_{\text{eq}}(x)=-\nabla_x V_\theta(x) \quad\Longrightarrow\quad \rho_{\text{eq}}(x)\propto\exp\left(-\tfrac{V_\theta(x)}{\varepsilon_{\max}}\right).$$
Thus, the equilibrium distribution is described by an EBM, $\exp(-E(x))$, with $E(x)=\frac{V_\theta(x)}{\varepsilon_{\max}}$.

Our approach in a nutshell. Combining the two insights above, we propose a generative framework that combines OT and EBMs to learn a *time-independent* scalar potential $V_\theta(x)$ whose Boltzmann distribution,

$$\rho_{\text{eq}}(x)\propto\exp\left(-\frac{V_\theta(x)}{\varepsilon_{\max}}\right), \tag{6}$$

matches ρ_{data}. To transport samples efficiently from noise ρ_0 to $\rho_{\text{eq}}\approx\rho_{\text{data}}$, we use two regimes:

- *Away from the data manifold*: $\varepsilon\approx 0$. The flow is deterministic and OT-like, allowing rapid movement across large distances in sample space.
- *Near the data manifold*: $\varepsilon\approx\varepsilon_{\max}$. Samples diffuse into a stable Boltzmann distribution, properly covering all data modes.

By combining the long-range transport capability of flows with the local density modeling flexibility of EBMs, we achieve tractable sampling and explicitly encode the unnormalized log-likelihood $-V_\theta(x)/\varepsilon_{\max}$ of the underlying data distribution; see Figure 1.

2.1 Training objectives

In practice, we balance the two objectives by initially training V_θ exclusively with the optimal-transport-like objective ($\varepsilon=0$, see Section 2.1.1), ensuring a stable and consistent generation of high-quality negative samples for the contrastive phase. Subsequently, we jointly optimize both the transport-based and contrastive divergence objectives, progressively increasing the effective temperature to $\varepsilon=\varepsilon_{\max}$ as samples approach the data manifold (i.e., the equilibrium distribution); see Section 2.1.2.

2.1.1 Flow-like objective $\mathcal{L}_{\text{OT}}$

We begin by constructing a global velocity field $-\nabla_x V_\theta(x)$ that carries noise samples $\{x_0\}$ to data samples $\{x_{\text{data}}\}$ with minimal detours. For this, we consider geodesics in the Wasserstein space [Ambrosio et al., 2008]. Practically, we compute the OT coupling γ^* between two uniform empirical probability distributions, one supported on a mini-batch of the data, and one supported on a set of noise samples with the same cardinality. These samples are drawn from an easy-to-sample distribution; in our case, a Gaussian. Since the probability distributions are uniform and empirical with the same number of samples, a transport map T is guaranteed to exist [Ambrosio et al., 2008].

Remark 2.1 (OT solver)**.** *Depending on the method used to compute the OT coupling, an explicit OT map may or may not be obtained. Similarly, if the number of noise samples differs from the mini-batch size B, the resulting OT coupling generally will not correspond to a map. In this case, one can adapt the algorithm by defining a threshold π_{th} and considering all pairs (x_{data},x_0) for which the coupling value satisfies $\gamma^*(x_{\text{data}},x_0)>\pi_{\text{th}}$. In our experiments, we used the* `POT` *solver [Flamary et al., 2021] and did not observe benefits from using a sample size different from B, consistent with previous approaches [Tong et al., 2023].*

Then, for each data point x_{data} we define the *interpolation* $x_t = (1-t)T(x_{\text{data}}) + tx_{\text{data}}$, which is a point along the geodesic. The *velocity* of each x_t is $x_{\text{data}} - T(x_{\text{data}})$ (i.e., the samples move from the noise to the data distribution at constant speed) and, in this regime, we would like to have $-\nabla_x V_\theta(x_t) \approx x_{\text{data}} - T(x_{\text{data}})$. For this, we define the loss:

$$\mathcal{L}_{\text{OT}} = \mathbb{E}_{t\sim U(0,\tau^*)}^{x_{\text{data}}\in\mathcal{D}} \left[\left\| \nabla_x V_\theta(x_t) + x_{\text{data}} - T(x_{\text{data}}) \right\|^2 \right].$$

This objective can be interpreted as a flow-matching formulation under the assumption that the velocity field is both time-independent and given by the gradient of a scalar potential, thereby imposing an irrotational condition. This aligns naturally with OT, which also yields an irrotational velocity field—any rotational component would add unnecessary distance to the flow and thus inflate the transport cost without benefit. Our experimental evidence adds to the recent study by [Sun et al., 2025], in which the authors observed that time-independent velocity fields can, under certain conditions, outperform time-dependent noise-conditioned fields in sample generation.

Algorithm 1 Phase 1 (warm-up).

Initialize model parameters θ
for iteration $n = 0, 1, \ldots$ **do**
  Sample mini-batch $\{x_{\text{data},b}\}_{b=1}^B \sim \mathcal{D}$ ▷ Data samples
  Sample mini-batch $\{x_{0,b}\}_{b=1}^B \sim \mathcal{N}(0, I)$ ▷ Random Gaussian samples
  $T \leftarrow \text{OTsolver}(\{x_{\text{data},b}\}, \{x_{0,b}\})$ ▷ Compute OT map
  Sample $\{t_b\}_{b=1}^B \sim U(0, \tau^*)$ ▷ Typically $\tau^* = 1$ for the warm-up
  Set interpolations $x_{t_b} \leftarrow (1-t_b)\, T(x_{\text{data},b}) + t_b\, x_{\text{data},b}$ ▷ Interpolation along geodesics
  $\mathcal{L}_{\text{OT}}(\theta) \leftarrow \sum_{b=1}^B \|\nabla_x V_\theta(x_{t_b}) + x_{\text{data},b} - T(x_{\text{data},b})\|^2$ ▷ Loss function
  $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{OT}}(\theta)$ ▷ Gradient update with learning rate α
end for
return θ ▷ Trained θ

2.1.2 Contrastive objective $\mathcal{L}_{\text{CD}}$

Near the data manifold, $V_\theta(x)$ is refined so that $\rho_{\text{eq}}(x) \propto \exp\left(-V_\theta(x)/\varepsilon_{\max}\right)$ matches the data distribution. We adopt the contrastive divergence loss described in EBMs [Hinton, 2002],

$$\mathcal{L}_{\text{CD}} = \mathbb{E}_{x\sim p_{\text{data}}} \left[\frac{V_\theta(x)}{\varepsilon_{\max}} \right] - \mathbb{E}_{\tilde{x}\sim \text{sg}(p_{\text{eq}})} \left[\frac{V_\theta(\tilde{x})}{\varepsilon_{\max}} \right],$$

where $\tilde{x}$ are "negative" samples of the equilibrium distribution induced by V_θ. We approximate these samples using an MCMC Langevin chain [Welling and Teh, 2011]. We split the initialization for negative samples: half begin at real data, and half begin at the noise distribution. This way, $V_\theta(x)$ forms well-defined basins around high-density regions while also shaping regions away from the manifold, correcting the generation. The $\text{sg}(\cdot)$ indicates a *stop-gradient* operator, which ensures gradients do not back-propagate through the sampling procedure.

2.1.3 Dual objective and implementation notes

To balance the deterministic flow-like regime (where $\varepsilon \approx 0$) away from the data manifold and the stochastic Boltzmann regime (where $\varepsilon \approx \varepsilon_{\max}$) near equilibrium, we adopt the linear temperature schedule described in (3). We introduce a dataset-specific hyperparameter λ_{CD} to stabilize the contrastive objective by appropriately weighting $\mathcal{L}_{\text{CD}}$ relative to $\mathcal{L}_{\text{OT}}$. The resulting algorithm is described in detail in Algorithm 1 and Algorithm 2. Since Algorithm 2 benefits from high-quality negatives, we begin with Algorithm 1 (and, thus, with $\mathcal{L}_{\text{OT}}$ only) to ensure sufficient mixing of noise-initialized negatives.

Given the trained models, we define a sampling time τ_{s}. Although convergence to the equilibrium distribution is guaranteed only as $\tau_{\text{s}} \to \infty$, we empirically observe that sample quality (measured with Fréchet inception distance (FID)) plateaus by $\tau_{\text{s}} = 3.25$ on CIFAR-10; see Section A.2. The sampling procedure, which optionally includes conditional and interaction terms, is detailed in Algorithm 3. In practice, we implement training using explicit Euler–Maruyama updates and sampling with an Euler–Heun predictor-corrector scheme, while for simplicity the algorithms illustrate only explicit updates. Additionally, the constant factor $1/\varepsilon_{\max}$ in $\mathcal{L}_{\text{CD}}$ is absorbed into λ_{CD}.

Section A.1 discusses how the landscape of V_θ evolves across these two phases. Hyperparameters for each dataset, along with intuitions to guide their selection for new datasets, are provided in Section D.

https://doi.org/10.52202/085713-0291

Algorithm 2 Phase 2 (main training).

$\theta \leftarrow \theta_{\text{pretrained}}$ ▷ Initialize from Algorithm 1
for iteration $n = 0, 1, \ldots$ **do**
  $\mathcal{L}_{\text{OT}} \leftarrow$ Use lines 3–8 from Algorithm 1
  Initialize negative samples $\{x^{(0)}_{\text{neg},b}\}_{b=1}^{B}$ from noise and/or data ▷ Negative samples
  for $m = 0, 1, \ldots, M_{\text{Langevin}} - 1$ **do**
    for $b = 1$ to B **do**
      $\varepsilon^{(m)} \leftarrow \begin{cases} \varepsilon_{\max}, & \text{if initialized from data (Optional)} \\ \varepsilon(m\Delta t) \text{ from (3)}, & \text{otherwise} \end{cases}$
      Sample $\eta_b \sim \mathcal{N}(0, I)$
      $x^{(m+1)}_{\text{neg},b} \leftarrow x^{(m)}_{\text{neg},b} - \Delta t \nabla_x V_{\text{sg}(\theta)}(x^{(m)}_{\text{neg},b}) + \sqrt{2\Delta t \varepsilon^{(m)}}\, \eta_b$ ▷ Langevin dynamics step
    end for
  end for
  $\mathcal{L}_{\text{CD}} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \left[V_\theta(x_{\text{data},b}) - V_\theta(x^{(M_{\text{Langevin}})}_{\text{neg},b}) \right]$ ▷ Contrastive divergence loss
  $\mathcal{L}(\theta) \leftarrow \mathcal{L}_{\text{OT}} + \lambda_{\text{CD}}\, \mathcal{L}_{\text{CD}}$
  Update $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta)$ ▷ Gradient descent step
end for
return θ ▷ Trained θ

Table 1: FID↓ score comparison for unconditional CIFAR-10 generation (lower is better). Unless otherwise specified, we use results for solvers that most closely match our setup (325 fixed-step Euler–Heun [Butcher, 2016]). * indicates reproduced methods, while all other entries reflect the best reported results. EGC in its unconditional version has been reported in [Zhu et al., 2024]

.

Learning Unnormalized Data Likelihood		**Learning Transport/Score Along Noised Trajectories**	
Ensembles: Diffusion + (one or many) EBMs		**Diffusion Models**	
Hierarchical EBM Diffusion [Cui and Han, 2024]	8.93	DDPM* [Ho et al., 2020]	6.45
EGC [Guo et al., 2023]	5.36	DDPM++ *(62M params, 1000 steps)* [Kim et al., 2021]	3.45
Cooperative DRL *(40M params)* [Zhu et al., 2024]	4.31	NCSN++ *(107M params, 1000 steps)* [Song et al., 2021]	2.45
Cooperative DRL-large *(145M params)* [Zhu et al., 2024]	3.68		
Energy-based Models		**Flow-based Models**	
ImprovedCD [Du et al., 2021]	25.1	Action Matching [Neklyudov et al., 2023]	10.07
CLEL-large *(32M params)* [Lee et al., 2023]	8.61	Flow-matching [Lipman et al., 2023]	6.35
Energy Matching *(50M params, **Ours**)*	3.34	OT-CFM* *(37M params)* [Tong et al., 2023]	4.04

3 Applications

In this section, we demonstrate the effectiveness and versatility of our proposed Energy Matching approach across three applications: (i) unconditional generation (ii) inverse problems, and (iii) LID estimation. The model architecture and all the training details are reported in Section D.

3.1 Unconditional generation

We compare four classes of generative models: (1) Diffusion models, which deliver state-of-the-art quality but typically require many sampling steps; (2) Flow-based methods, which learn OT paths for more efficient sampling with fewer steps; (3) EBMs, which directly model the log-density as a scalar field, offering flexibility for inverse problems and constraints but sometimes at the expense of sample quality; and (4) Ensembles (Diffusion with one or many EBMs), which combine diffusion's robust sampling with elements of EBM flexibility but can become large and complex to train. Our approach, Energy Matching, offers a simple (a single time-independent scalar field) yet powerful EBM-based framework. We evaluate our approach on CIFAR-10 [Krizhevsky and Hinton, 2009] and ImageNet32x32 [Deng et al., 2009, Chrabaszcz et al., 2017] datasets, reporting FID scores in Table 1 and Table 2, respectively. Our method outperforms state-of-the-art EBMs, reducing the FID score by more than 50%.

Table 2: FID↓ score comparison for unconditional ImageNet 32x32 generation (lower is better). Unless otherwise specified, we use results for solvers that most closely match our setup (300 fixed-step Euler–Heun [Butcher, 2016]).

Learning Unnormalized Data Likelihood		**Learning Transport/Score Along Noised Trajectories**	
Ensembles: Diffusion + (one or many) EBMs		**Diffusion Models**	
Cooperative DRL *(40M params)* [Zhu et al., 2024]	9.35	DDPM++ *(62M params, 1000 steps)* [Kim et al., 2021]	8.42
Energy-based Models		**Flow-based Models**	
ImprovedCD [Du et al., 2021]	32.48	Flow-matching [Lipman et al., 2023] *(196M params)*	5.02
CLEL-base [Lee et al., 2023] *(7M params)*	22.16		
CLEL-large [Lee et al., 2023] *(32M params)*	15.47		
Energy Matching *(50M params,* ***Ours****)*	6.64		

3.2 Inverse problems

In many practical applications, we are interested in recovering some data x from noisy measurements y generated by an operator A, $y = A(x) + w$, where $w \sim \mathcal{N}\big(0, \sqrt{2}\,\zeta I\big)$. In this setting, the posterior distribution of x given y is

$$p(x|y) \propto \underbrace{\exp\left(-\frac{1}{\zeta^2}\|y - A(x)\|^2\right)}_{\propto\, p(y|x)} \underbrace{\exp\left(-E_\theta(x)\right)}_{\propto\, p(x)}, \tag{7}$$

where $E_\theta(x)$ is an energy function which one can learn from the data, an EBM. Because we want to sample x given a measurement y, this reconstruction task is often referred to as an *inverse problem*. Here, $\|y - Ax\|^2$ encodes the measurement fidelity with ζ controlling the balance between this fidelity term and the prior. We obtain the prior term $E_\theta(x) = \frac{V_\theta(x)}{\varepsilon_{\max}}$ by training $V_\theta(x)$ via Energy Matching. Samples from this posterior can be drawn by starting from a random sample $x^{(0)} \sim \mathcal{N}(0, I)$ and following a Langevin update. We detail the algorithm for generating solutions to inverse problems in Algorithm 3 (which also incorporates additional interaction energy $W(x, x')$ between generated samples). We demonstrate our model's capabilities qualitatively through a controlled inpainting task and quantitatively via a protein inverse design benchmark. Specific hyperparameters are detailed in Section D.

Algorithm 3 Unconditional/conditional sampling with optional interaction energy

for $m = 1$ to M **do**
  Initialize $x_m^{(0)}$ from noise and/or data ▷ Initialize each chain
end for
$N \leftarrow \lfloor \tau_{\mathrm{s}} / \Delta t \rfloor$ ▷ Number of Langevin steps for sampling time τ_{s}
for $n = 0, 1, \ldots, N-1$ **do**
  for $m = 1, 2, \ldots, M$ **do** ▷ Prior + data fidelity + interaction
    $\varepsilon^{(n)} \leftarrow \begin{cases} \varepsilon_{\max}, & \text{if initialized from data (Optional)} \\ \varepsilon(n\Delta t) \text{ from (3)}, & \text{otherwise} \end{cases}$
    $U_\theta\big(x_m^{(n)}\big) \leftarrow V_\theta\big(x_m^{(n)}\big) + \varepsilon^{(n)} \big\|y - A\big(x_m^{(n)}\big)\big\|^2 / \zeta^2 + \varepsilon^{(n)} \sum_{k \neq m} W\big(x_m^{(n)}, x_k^{(n)}\big)$
    Sample $\eta_m^{(n)} \sim \mathcal{N}(0, I)$
    $x_m^{(n+1)} \leftarrow x_m^{(n)} - \Delta t\, \nabla_x U_\theta\big(x_m^{(n)}\big) + \sqrt{2\varepsilon^{(n)}\Delta t}\; \eta_m^{(n)}$ ▷ Langevin dynamics step
  end for
end for
return $\{x_m^{(N)}\}_{m=1}^M$ ▷ Final samples

Controlled inpainting. Suppose we want to recover two images from a masked image while encouraging diverse reconstructions. EBMs allow this by introducing an additional interaction energy, $W(x_1, x_2) = -\frac{\|B(x_1 - x_2)\|^2}{\sigma^2}$, where B has ones in the region of interest (focusing diversity there) and zeros elsewhere, and σ is a hyperparameter controlling the interaction's strength. Specifically, we define $p(x_1, x_2 \mid y) \propto p(x_1 \mid y)\, p(x_2 \mid y)\, \exp\big(-W(x_1, x_2)\big)$, which gives high probability to pairs (x_1, x_2) that lie far apart in the specified region B. This encourages exploring the edges of the

https://doi.org/10.52202/085713-0291

posterior rather than just its modes, and with suitable W, samples shift toward rare events without needing many draws. To illustrate the interaction term's advantages for diverse reconstruction, we apply our method to a CelebA [Liu et al., 2015] 64×64 inpainting task. As shown in Figure 2, we start from a partially observed (masked) face and aim to reconstruct two distinct high-fidelity completions.

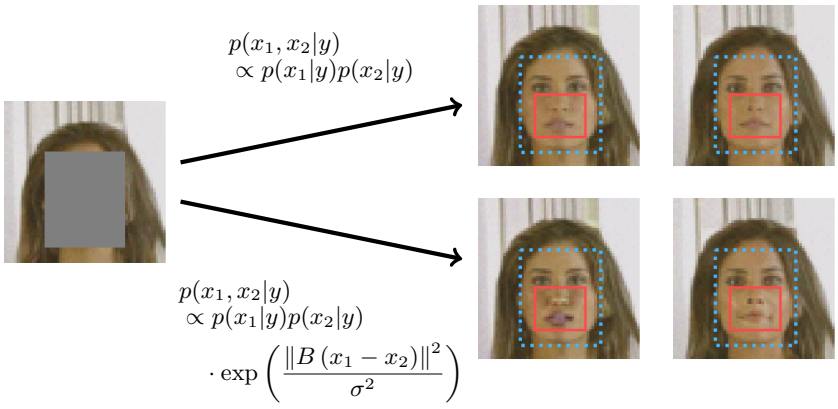


Figure 2: Controlled inpainting for diverse reconstructions. On the left is the masked face. On the right are two reconstructions: the top pair without the interaction term and the bottom pair with it. The interaction term applies in the solid red square (where B has ones), and the measurement matrix A is the dotted blue square (zeros inside, ones outside). By encouraging x_1 and x_2 to differ in the target region, the interaction yields a wider range of completions while preserving fidelity.

Protein inverse design. In Figure 3, we demonstrate our method's performance on the inverse design problem of generating Adeno-Associated Virus (AAV) capsid protein segments [Bryant et al., 2021]. Given a desired functional property (fitness)—here defined as the predicted viral packaging efficiency normalized between 0 and 1—the goal is to design novel protein sequences satisfying this target condition. Beyond achieving high fitness, practical inverse design requires generating diverse candidate sequences to ensure robustness in the downstream experimental validation [Jain et al., 2022]. We evaluate on two benchmark splits (*medium* and *hard*), which correspond to subsets of the original AAV dataset differing in baseline fitness distributions and required mutational distance from known high-performing variants [Kirjner et al., 2024]. Leveraging the latent-space representation of VLGPO [Bogensperger et al., 2025], we employ our Energy-Matching Langevin sampler with an inference-time tunable repulsion term, allowing explicit control over the diversity of the designed proteins. This enables a flexible trade-off between fitness and diversity, resulting in high fitness scores alongside substantially improved sequence diversity. See Section B for experimental details and dataset descriptions.

3.3 Local intrinsic dimension estimation

Real-world datasets, despite displaying a high number of variables, can often be represented by lower-dimensional manifolds—a concept referred to as the *manifold hypothesis* [Fefferman et al., 2016]. The dimension of such a manifold is called the intrinsic dimension. Estimating the LID at a given point reveals its effective degrees of freedom or *directions of variation*, offering insight into data complexity and adversarial vulnerabilities. We defer the precise definition to Section C.

Diffusion-based approaches. Recent work leverages pretrained *diffusion models* to estimate the LID [Kamkari et al., 2024, Stanczuk et al., 2024] by examining the learned score function. However, since these models do not learn the score at the data manifold ($t = 1$), their estimates become unreliable there. Consequently, current methods rely on approximations, for instance by evaluating the score in the proximity of the data manifold ($t = 1 - t_0$), where computations remain sufficiently reliable.

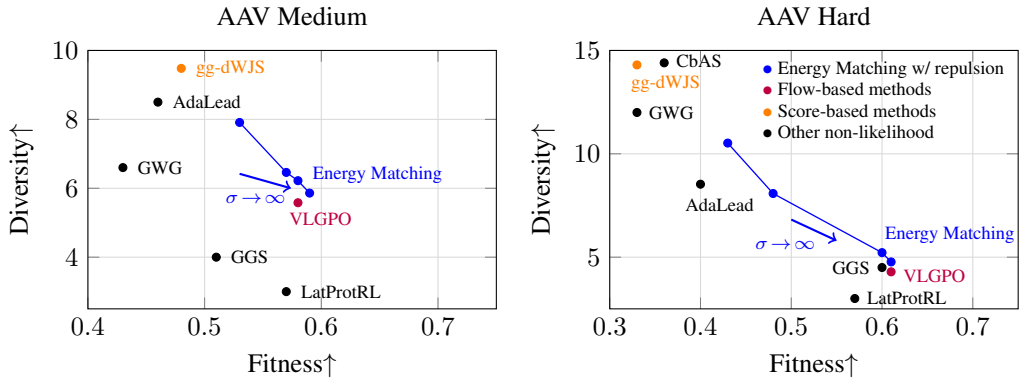


Figure 3: Fitness–diversity trade-off for protein inverse design on the AAV Medium (left) and Hard (right) benchmarks. We compare our Energy Matching method (blue), with diversity explicitly controlled by a repulsion strength parameter ($\propto \frac{1}{\sigma^2}$), against leading flow-based (purple), score-based (orange), and other non-likelihood methods (black). Fitness measures how well generated sequences satisfy the target property (predicted viral packaging efficiency), while diversity quantifies the average Levenshtein distance between sequences in each generated batch.

Spearman's correlation ↑	MNIST	CIFAR-10
ESS [Johnsson et al., 2014]	0.444	0.326
FLIPD [Kamkari et al., 2024]	0.837	0.819
NB [Stanczuk et al., 2024]	0.864	0.894
Energy Matching (Ours)	0.877	0.901

Table 3: Spearman's correlation coefficients of LID estimates with PNG compression rate. Benchmarks results reported in [Kamkari et al., 2024].

Hessian-based LID Estimation. Unlike diffusion models, EBMs explicitly parametrize the relative data likelihood. This explicit parametrization enables efficient analysis of the curvature of the underlying data manifold – in this example, estimating the LID. To this end, we compute the Hessian matrix $\nabla_x^2 V(x_{\text{data}})$ at a given data point and perform its spectral decomposition. We define near-zero eigenvalues as those whose absolute values lie within a small threshold τ (in our experiments, we set $\tau = 3$ for MNIST [Deng, 2012] and $\tau = 2$ for CIFAR-10). The count of near-zero eigenvalues reflects the number of *flat* directions and thus reveals the local dimension. As shown in Table 3, the LID estimates we obtain exhibit stronger correlations with PNG compression size[1] (evaluated on 4096 images) using Spearman's correlation. Figure 4 offers qualitative illustrations. Our EBM-based approach compares favorably to diffusion-based methods, as it relies on fewer approximations by performing computations exactly on the data manifold rather than merely in its vicinity.

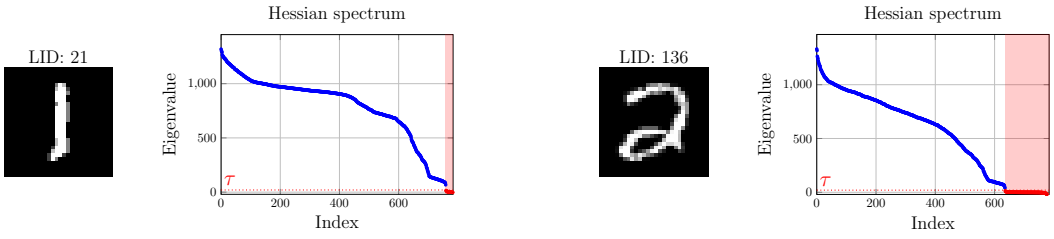


Figure 4: Qualitative results for LID estimation using the Hessian spectrum of $V_\theta(x)$. Left: Spectrum for a low-LID image. Right: Spectrum for a high-LID image. The eigenvalues quantify curvature along principal directions (eigenvectors). A degenerate spectrum (many near-zero eigenvalues, marked in red) indicates locally "flat" regions, revealing the LID. Intuitively, higher image complexity often corresponds to a higher LID.

[1]PNG is a lossless compression scheme specialized for images and can provide useful guidance when no LID ground truth is available Kamkari et al. [2024].

https://doi.org/10.52202/085713-0291

4 Conclusion and limitations

Contributions. We introduced a generative framework, *Energy Matching*, that reconciles the advantages of EBMs and OT flow matching models for simulation-free likelihood estimation and efficient high-fidelity generation. Specifically, it:

- Learns a *time-independent scalar potential energy* whose gradient drives rapid high-fidelity sampling–surpassing state-of-the-art energy-based models–while also forming a Boltzmann-like density well suitable for controlled generation. All without auxiliary generators.
- Offers efficient sampling from target data distributions on par with the state-of-the-art, while learning the score at the data manifold with manageable trainable parameters overhead.
- Offers a simulation-free, principled likelihood estimation framework for solving inverse problems—where additional priors can be easily introduced—and enables the estimation of a data point's LID with fewer approximations than score-based methods.

Limitations. First, our method requires an additional gradient computation with respect to the input, which can increase GPU memory usage (e.g., by 20–40%), particularly during training. Second, when estimating the LID (Section 3.3) for very high-dimensional datasets, computing the full Hessian spectrum may be impractical due to its computational complexity of $O(d^3)$; in such cases, partial-spectrum methods such as random projections or iterative solvers can be employed instead.

Outlook. Contrary to widespread belief, we demonstrated that time-independent irrotational methods for generative flows are highly effective and offer an exciting direction for future research. Our Energy Matching approach has the potential to yield novel insights into controlled generation and inverse problems for cancer research Weidner et al. [2024], Balcerak et al. [2024], molecules and proteins Wu et al. [2022], Bilodeau et al. [2022], computational fluid dynamics Gao et al. [2024], Shysheya et al. [2024], Molinaro et al. [2024], and other fields where precise control over generated samples and effective integration of priors or constraints are crucial. Moreover, Energy Matching aligns naturally with recent generative AI trends toward scaling inference for new capabilities [Zhang et al., 2025, Ma et al., 2025], further broadening its potential impact across scientific and engineering domains.

Acknowledgments and Disclosure of Funding

This research was supported by the Helmut Horten Foundation and the European Cooperation in Science and Technology (COST).

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory (ICDT)*, 2001.

Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations (ICLR)*, 2023.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Basel, 2008.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2024.

Michal Balcerak, Tamaz Amiranashvili, Andreas Wagner, Jonas Weidner, Petr Karnakov, Johannes C Paetzold, Ivan Ezhov, Petros Koumoutsakos, Benedikt Wiestler, et al. Physics-regularized multimodal image assimilation for brain tumor localization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-flow: differentiating through flows for controlled generation. In *International Conference on Machine Learning (ICML)*, 2024.

Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

Lea Bogensperger, Dominik Narnhofer, Ahmed Allam, Konrad Schindler, and Michael Krauthammer. A variational perspective on generative protein fitness optimization. In *International Conference on Machine Learning (ICML)*, 2025.

David Brookes, Hahnbeom Park, and Jennifer Listgarten. Conditioning by adaptive sampling for robust design. In *International conference on machine learning (ICML)*, 2019.

Drew H. Bryant et al. Deep diversification of an aav capsid protein by machine learning. *Nature Biotechnology*, 39:691–696, 2021.

John C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, 3rd edition, 2016.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable wasserstein gradient flow for generative modeling through unbalanced optimal transport. In *International Conference on Machine Learning (ICML)*, 2024.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiali Cui and Tian Han. Learning latent space hierarchical ebm diffusion models. In *International Conference on Machine Learning (ICML)*, 2024.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICML)*, 2020.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. In *International Conference on Machine Learning (ICML)*, 2021.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

 https://doi.org/10.52202/085713-0291

Han Gao, Sebastian Kaltenbach, and Petros Koumoutsakos. Generative learning for forecasting the dynamics of high-dimensional complex systems. *Nature Communications*, 15(1):8904, 2024.

Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. In *International Conference on Learning Representations (ICLR)*, 2021.

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. In *International Conference on Machine Learning (ICML)*, 2021.

Qiushan Guo, Chuofan Ma, Yi Jiang, Zehuan Yuan, Yizhou Yu, and Ping Luo. Egc: Image generation and classification via a diffusion energy-based model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

Aapo Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued data. *Neural Computation*, 18(8):1527–1550, 2006.

Zarif Ikram, Dianbo Liu, and M Saifur Rahman. Antibody sequence optimization with gradient-guided discrete walk-jump sampling. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024.

Moksh Jain et al. Biological sequence design with gflownets. In *International Conference on Machine Learning (ICML)*, 2022.

Kerstin Johnsson, Charlotte Soneson, and Magnus Fontes. Low bias local intrinsic dimension estimation from expected simplex skewness. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):196–202, 2014.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Hamid Kamkari, Brendan Ross, Rasa Hosseinzadeh, Jesse Cresswell, and Gabriel Loaiza-Ganem. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning (ICML)*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Andrew Kirjner et al. Improving protein optimization with smoothed fitness landscapes. In *International Conference on Learning Representations (ICLR)*, 2024.

Alex Krizhevsky and Geoffrey E. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Department of Computer Science, Toronto, Ontario, Canada, 2009.

Nicolas Lanzetti, Antonio Terpin, and Florian Dörfler. Variational analysis in the wasserstein space. *arXiv preprint arXiv:2406.10676*, 2024.

Nicolas Lanzetti, Saverio Bolognani, and Florian Dörfler. First-order conditions for optimization in the wasserstein space. *SIAM Journal on Mathematics of Data Science*, 7(1):274–300, 2025.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.

Hankook Lee, Jongheon Jeong, Sejun Park, and Jinwoo Shin. Guiding energy-based models via contrastive latent variables. *arXiv preprint arXiv:2303.03023*, 2023.

Minji Lee, Luiz Felipe Vecchietti, Hyunkyu Jung, Hyun Joo Ro, Meeyoung Cha, and Ho Min Kim. Robust optimization in protein fitness landscapes using reinforcement learning in latent space. In *International Conference on Machine Learning (ICML)*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint*, arXiv:2501.09732, 2025.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Roberto Molinaro, Samuel Lanthaler, Bogdan Raonić, Tobias Rohner, Victor Armegioiu, Stephan Simonis, Dana Grund, Yannick Ramic, Zhong Yi Wan, Fei Sha, et al. Generative ai for fast and accurate statistical computation of fluids. *arXiv preprint arXiv:2409.18359*, 2024.

Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *International Conference on Machine Learning (ICML)*, 2023.

Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, José Miguel Hernández-Lobato, Richard Turner, and Emile Mathieu. On conditional diffusion models for pde simulations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sam Sinai, Richard Wang, Alexander Whatley, Stewart Slocum, Elina Locane, and Eric D Kelsic. Adalead: A simple and robust adaptive greedy search algorithm for sequence design. *arXiv preprint arXiv:2010.02141*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in neural information processing systems (NeurIPS)*, 2019.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *International Conference on Machine Learning (ICML)*, 2024.

Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models? In *International Conference on Machine Learning (ICML)*, 2025.

Antonio Terpin, Nicolas Lanzetti, Martín Gadea, and Florian Dorfler. Learning diffusion at lightspeed. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning (ICML)*, 2008.

 https://doi.org/10.52202/085713-0291

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2023.

Jonas Weidner, Ivan Ezhov, Michal Balcerak, Marie-Christin Metz, Sergey Litvinov, Sebastian Kaltenbach, Leonhard Feiner, Laurin Lux, Florian Kofler, Jana Lipkova, et al. A learnable prior improves inverse tumor growth modeling. *IEEE Transactions on Medical Imaging*, 2024.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International conference on machine learning (ICML)*, 2011.

Lemeng Wu, Chengyue Gong, Xingchao Liu, Mao Ye, and Qiang Liu. Diffusion-based molecule generation with informative prior bridges. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by jko scheme. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Sangwoong Yoon, Himchan Hwang, Dohyun Kwon, Yung-Kyun Noh, and Frank Park. Maximum entropy inverse reinforcement learning of diffusion models with energy-based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yasi Zhang, Peiyu Yu, Yaxuan Zhu, Yingshan Chang, Feng Gao, Ying Nian Wu, and Oscar Leong. Flow priors for linear inverse problems via iterative corrupted trajectory matching. *arXiv preprint arXiv:2405.18816*, 2024.

Yaxuan Zhu, Jianwen Xie, Ying Nian Wu, and Ruiqi Gao. Learning energy-based models by cooperative diffusion recovery likelihood. In *International Conference on Learning Representations (ICLR)*, 2024.

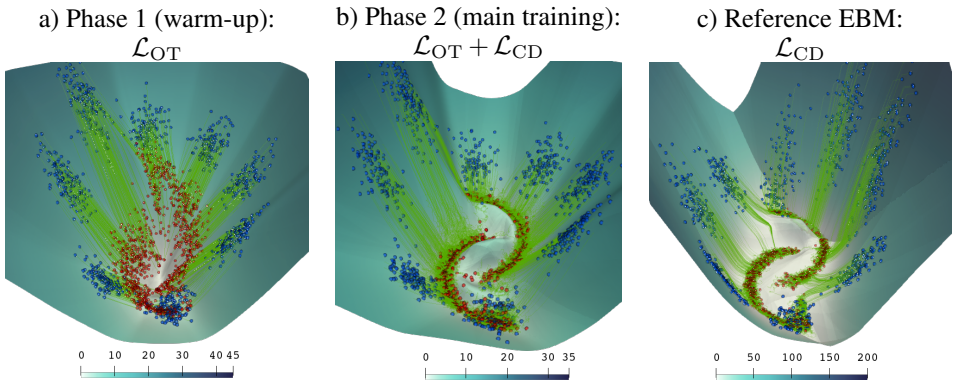


Figure 5: Visualization of the energy $V_\theta(x)$ landscapes driving the samples from eight Gaussians to two moons. See Figure 1 for the 2D perspective. (a) The OT flow loss enforces zero curvature in $V_\theta(x)$ along the trajectories to the target. (b) Around the 2 Moons, the curvature of $V_\theta(x)$ is adjusted to approximate $\log p_{\text{moons}}(x) \propto V_\theta(x)$ while remaining close to the pretrained landscape elsewhere. Combining these objectives yields a potential energy landscape that is both efficient for sampling and representative of the underlying target data distribution. (c) An EBM is shown for comparison, trained using contrastive divergence loss. Visible mode collapse that slows down the equilibration. Less regular landscape away from the data as it needs many simulations to explore it.

A Additional details on Energy Matching

In this section, we provide additional studies and visualizations on our method.

A.1 Energy landscape during training

In Figure 5, we visualize how the potential $V_\theta(x)$ transitions from a flow-like regime, where the OT loss enforces nearly zero curvature away from the data manifold (a), to an EBM-like regime, where the curvature around the new data geometry (here, two moons) is adaptively increased to approximate $\log p_{\text{data}}(x)$ (b). This two-stage design yields a well-shaped landscape that is both efficient to sample (thanks to a mostly flat potential between clusters) and accurate for density estimation near the data modes. For comparison, (c) shows an EBM trained solely with contrastive divergence, exhibiting sharper but less globally consistent basins.

A.2 Ablation on the sampling time

Here, we provide ablation studies on CIFAR-10 unconditional generation. Specifically, we first pretrain using $\mathcal{L}_{\mathrm{OT}}$, and then fine-tune with $(\mathcal{L}_{\mathrm{OT}} + \mathcal{L}_{\mathrm{CD}})$, producing a stable Boltzmann distribution from which one can sample. Figure 6 illustrates the FID as a function of sampling time τ_{s} for models trained under these different regimes. In the case of pure $\mathcal{L}_{\mathrm{OT}}$, the quality measure drops (FID increases) sharply when sampling at $\tau_{\mathrm{s}} > 1$; this occurs because, once the samples move close to the data manifold, there is no Boltzmann-like potential well to keep them from drifting away. Because the fidelity slope near the data manifold is steep with respect to sampling time, methods lacking explicit time-conditioning can easily overshoot or undershoot, significantly impacting fidelity. This behavior might explain why some models degrade in performance when made time-independent, as recently reported by Sun et al. [2025].

In Figure 6 we also report results for different values of the temperature-switching parameter τ^*, which influences the sampling along the paths towards the data manifold (see Equation (3)).

A.3 Ablations on the OT Solver

We evaluate the computational overhead and sensitivity to solver choice of the OT solver employed in our experiments.

https://doi.org/10.52202/085713-0291

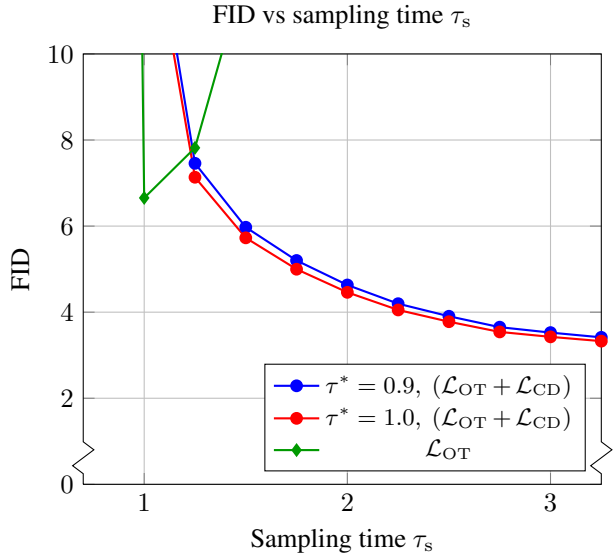


Figure 6: CIFAR-10 unconditional generation FID vs. sampling time τ_s when sampling from models trained under different scenarios: pure $\mathcal{L}_{OT}$ and combined $(\mathcal{L}_{OT} + \mathcal{L}_{CD})$, with temperature regime switching parameter $\tau^* \in \{0.9, 1.0\}$ during sampling. Lower FID indicates better generative quality. All results for Euler-Heun with $\Delta t = 0.01$.

Computational overhead. In the CIFAR-10 experiment, the OT solver accounts for roughly 1.5% of the training iteration time during Phase 1 Algorithm 1. This overhead decreases to a negligible level (approximately 0.01%) in Phase 2 Algorithm 2, where computational costs are predominantly dominated by the generation of negative samples.

Impact of solver accuracy and complexity. Let us (over-idealise) and model a standardised CIFAR-10 image as a vector $x \in \mathbb{R}^d$, drawn from $\mathcal{N}(0, I_d)$, with dimensionality $d = 32 \times 32 \times 3 = 3072$, and paired with an independent Gaussian-noise vector $z \sim \mathcal{N}(0, I_d)$. Each coordinate of the difference $(z - x)$ thus follows $\mathcal{N}(0, 2)$, and the squared Euclidean distance distribution is $|z - x|_2^2 \sim 2\chi_d^2$, which has mean $2d$, standard deviation $\sqrt{8d}$, and relative spread $\frac{\sqrt{8d}}{2d} \approx 0.025$. This demonstrates the "thin-shell" phenomenon, implying that all entries of the cost matrix $C_{ij} = |z_i - x_j|_2^2$ concentrate around nearly identical values, consistent with the distance-concentration effect observed by [Aggarwal et al., 2001]. Consequently, the choice among exact linear programming (LP), entropic regularisation (Sinkhorn), or even random matching should yield nearly identical cumulative optimal-transport costs, despite their different complexities: $O(n^3 \log n)$ for LP, $O(\frac{n^2}{\kappa})$ for Sinkhorn (with regularisation strength κ), and $O(n)$ for random matching.

Empirical evidence supporting this is summarised below:

- Our CIFAR-10 runs: FID degrades from 3.34 (LP) to 3.37 (random).
- [Tong et al., 2023] Table 5: FID scores of 4.44 (LP) vs 4.46 (random) at 100 steps.
- [Tong et al., 2023] Fig. D.2: solver accuracy saturates beyond batch size 16 in the 2D task.
- [Terpin et al., 2024] App. C.2: LP and Sinkhorn methods produce indistinguishable results unless regularisation is extreme.

We employ the LP approach for robustness, negligible cost, and no additional hyperparameters.

A.4 Sampling Time Analysis

Computing the gradient $\nabla_x V_\theta(x)$ via automatic differentiation (`autograd` [Ansel et al., 2024]) introduces additional computational overhead compared to directly evaluating $V_\theta(x)$. Specifically, on the CIFAR-10 network architecture (see Figure 7), gradient evaluation is approximately $2.15\times$

slower, as it requires both forward and backward passes. In contrast, flow matching and diffusion models directly parameterize the velocity field, thus only requiring forward computations during sampling.

Nevertheless, despite this per-step computational cost, our method achieves competitive overall sampling efficiency due to a reduced number of integration steps needed for high-quality generation. As demonstrated in Table 4, Energy Matching achieves a lower FID (3.34) in 173 seconds per batch, outperforming OT-FM (FID 3.74 in 136 seconds per batch) and DDPM++ (FID 3.45 in 183 seconds per batch). Our results thus indicate a favorable balance between computational overhead per step and total sampling runtime.

Table 4: Comparison of sampling efficiency and quality on CIFAR-10 (batch size 128, NVIDIA R6000 48GB GPU). Despite gradient computation overhead ($\nabla_x V_\theta(x)$ via backward pass), Energy Matching achieves superior FID scores with competitive wall-clock sampling time.

Method	Params	Steps	Sampling Time [s]↓	FID↓
Flow-/Diffusion-based Models				
OT-FM [Tong et al., 2023]	37M	1000	**136**	3.74
DDPM++ [Kim et al., 2021]	62M	1000	183	3.45
Energy-based Models				
Energy Matching *(Ours)*	50M	325	173	**3.34**

B Details on AAV inverse design protein generation

We optimize protein fitness for adeno-associated virus (AAV) sequences in the *medium* and *hard* data regimes proposed by [Kirjner et al., 2024], using latent encodings and backbone architectures from [Bogensperger et al., 2025]. Conditional sampling employs classifier guidance via learned predictor networks g_ϕ or $\tilde{g}_\phi$ to steer samples toward high-fitness regions. The CNN-based fitness predictors from [Kirjner et al., 2024] are trained only on the limited training data for each regime.

Training follows Algorithm 1 and Algorithm 2. We sample 128 sequences using Algorithm 3, keeping the same batch size across all baselines; detailed hyperparameters are given in Section D. Generated sequences are evaluated for fitness using the learned oracle from [Kirjner et al., 2024], and further assessed for both intra-set diversity and novelty relative to the training sequences [Jain et al., 2022]. Our approach achieves state-of-the-art fitness while improving diversity (see Table 5). Incorporating interaction energy in Algorithm 3 further enhances diversity with manageable impact on fitness.

Table 5: AAV optimization results. For VLGPO (flow-based) and Energy Matching, medium difficulty uses g_ϕ, hard difficulty uses $\tilde{g}_\phi$. Metrics (*Fitness*↑, *Diversity*↑, *Novelty*↑). Reported uncertainty of Fitness is expressed as standard deviation.

	AAV medium			AAV hard		
Method	**Fitness↑**	**Diversity↑**	**Novelty↑**	**Fitness↑**	**Diversity↑**	**Novelty↑**
Learning Unnormalized Data Likelihood						
Energy-based Models						
Energy Matching ***(Ours)***	0.59 (0.0)	5.86	5.0	0.61 (0.0)	4.77	6.7
Energy Matching (+repulsion) ***(Ours)***	0.58 (0.0)	6.22	5.0	0.60 (0.0)	5.22	6.6
Learning Transport/Score Along Noised Trajectories						
Flow-based Models						
VLGPO [Bogensperger et al., 2025]	0.58 (0.0)	5.58	5.0	0.61 (0.0)	4.29	6.2
Diffusion Models						
gg-dWJS [Ikram et al., 2024]	0.48 (0.0)	9.48	4.2	0.33 (0.0)	14.3	5.3
Other Methods						
LatProtRL [Lee et al., 2024]	0.57 (0.0)	3.00	5.0	0.57 (0.0)	3.00	5.0
GGS [Kirjner et al., 2024]	0.51 (0.0)	4.00	5.4	0.60 (0.0)	4.50	7.0
AdaLead [Sinai et al., 2020]	0.46 (0.0)	8.50	2.8	0.40 (0.0)	8.53	3.4
CbAS [Brookes et al., 2019]	0.43 (0.0)	12.70	7.2	0.36 (0.0)	14.4	8.6
GWG [Grathwohl et al., 2021]	0.43 (0.1)	6.60	7.7	0.33 (0.0)	12.0	12.2
GFN-AL [Jain et al., 2022]	0.20 (0.1)	9.60	19.4	0.10 (0.1)	11.6	19.6

https://doi.org/10.52202/085713-0291

C Details on LID estimation

Definition. To start, we need to introduce the concept of *local mass*, defined as

$$M(r) \;=\; \int_{B(x_{\text{data}};r)} p(x)\,dx,$$

where $p(x)$ is the local density and $B(x_{\text{data}}, r)$ is a ball of radius r around x_{data}, i.e. $B(x_{\text{data}}, r) = \{x \in \mathbb{R}^d : \|x - x_{\text{data}}\| \le r\}$. The LID is then given by:

$$\text{LID}(x_{\text{data}}) = d - \lim_{r\to 0} \frac{\log\big(M(r)\big)}{\log(r)}.$$

Intuitively, $M(r)$ measures how much probability mass is concentrated in a ball of radius r around x_{data}. As we shrink this ball, the growth rate of $M(r)$ in terms of r reveals the local dimensional structure of the data.

Assumptions. In the context of contrastive divergence, we assume that data points x_{data} lie in well-like regions [Hyvärinen, 2006], i.e.:

$$\nabla V(x_{\text{data}}) \approx 0 \quad \text{and} \quad \nabla^2 V(x_{\text{data}}) \text{ is positive semidefinite (or nearly so).}$$

Conceptually, $V(x)$ can be thought of as an energy function; points where $\nabla V(x_{\text{data}}) = 0$ are near local minima of this energy, and the Hessian $\nabla^2 V(x_{\text{data}})$ provides information about local curvature (see Figure 4 for a qualitative illustration).

Energy-based density. We define an energy-based density

$$p(x) \;\propto\; \exp\Big(-\tfrac{V(x)}{\varepsilon}\Big),$$

where ε is a temperature parameter. Near a data point x_{data} satisfying $\nabla_x V(x_{\text{data}}) = 0$, we can approximate $V(x)$ by its second-order Taylor expansion:

$$V(x) \approx V(x_{\text{data}}) + \frac{1}{2}(x - x_{\text{data}})^\top \nabla_x^2 V(x_{\text{data}})(x - x_{\text{data}}).$$

Consequently, in view of the assumptions above,

$$p(x) \propto \exp\left(-\frac{1}{2\varepsilon}(x - x_{\text{data}})^\top \nabla_x^2 V(x_{\text{data}})(x - x_{\text{data}})\right).$$

Local mass derivation and the rank of the energy Hessian. Substituting the local quadratic form of $p(x)$ near x_{data} into the definition of the local mass $M(r)$, we obtain:

$$M(r) = \int_{B(x_{\text{data}},r)} p(x)dx \propto \int_{B(x_{\text{data}},r)} \exp\left(-\tfrac{1}{2\varepsilon}(x - x_{\text{data}})^\top \nabla_x^2 V(x_{\text{data}})(x - x_{\text{data}})\right) dx.$$

For small r, the dominant contribution depends on the rank of the Hessian $\nabla_x^2 V(x_{\text{data}})$. Let $k = \text{rank}(\nabla_x^2 V(x_{\text{data}}))$. Then, as $r \to 0$, one can show that $M(r) = Cr^k$, where C does not depend on r. We take the logarithm on both sides and divide by $\log(r)$ to get

$$\frac{\log(M(r))}{\log(r)} = \frac{\log(C) + k\log(r)}{\log(r)} = k + \frac{\log(C)}{\log(r)},$$

and the second term vanishes as $r \to 0$. Hence,

$$\text{LID}(x_{\text{data}}) = d - k.$$

Practical estimation. In practice, the LID at a data point x_{data} can be estimated through the following procedure:

1. Train $V(x)$ with Energy Matching.
2. Compute the Hessian $H = \nabla_x^2 V(x_{\text{data}})$.
3. Perform an eigenvalue decomposition on H.

Then the estimated local data-manifold dimension corresponds to the number of directions with negligible curvature (smaller magnitude than some τ).

D Training details

Below, we detail the training configurations for CIFAR-10, ImageNet 32x32, CelebA, MNIST, and AAV. Additionally, we provide intuitions for practical hyperparameter choices to facilitate effective training across additional datasets. We recommend using SiLU activation functions wherever possible, as they smooth out the energy landscape and improve the numerical stability of the $\nabla_x V(x)$ computation. The gradient of the potential, $\nabla_x V(x)$, is computed using automatic differentiation via PyTorch's `autograd` [Ansel et al., 2024]. We optimize all models using the Adam optimizer [Kingma and Ba, 2014] and maintain an exponential moving average (EMA) of the model weights.

While we specifically adopt (i) a one-sided trimmed mean of negative sample energies and (ii) clamping of the contrastive loss for stability, any commonly used EBM technique (e.g., persistent contrastive divergence [Tieleman, 2008], replay buffers, multi-scale negative sampling) could be readily employed.

In our approach, we introduce two hyperparameters, α and β, to control these stabilizing techniques:

$\alpha =$ fraction of negative energies discarded to remove outliers that skew the mean (e.g., top 10%),

$\beta =$ clamp threshold for $\mathcal{L}_{\text{CD}}$ (i.e., we clamp $\mathcal{L}_{\text{CD}}$ to be $\geq -\beta$).

CIFAR-10: The architecture is shown in Figure 7. We use the same UNet from [Tong et al., 2023] (with fixed $t = 0.0$, making it effectively time-independent) followed by a small vision transformer (ViT) [Dosovitskiy et al., 2020] to obtain a scalar output. Hyperparameters are: $\tau_s = 3.25$, $\tau^* = 1.0$, $\Delta t = 0.01$, $M_{\text{Langevin}} = 200$. We train for 145k iterations using Algorithm 1 with EMA 0.9999 and then 2k more with Algorithm 2 and EMA 0.99 on 4xA100. The batch size is 128, learning rate is 1.2×10^{-3}, $\varepsilon_{\max} = 0.01$, $\lambda_{\text{CD}} = 1 \times 10^{-3}$, $\alpha = 0.1$, and $\beta = 0.02$. Negatives initialized on the data manifold follow the same temperature schedule as those initialized from the noise.

ImageNet 32x32: The architecture is shown in Figure 7 (same as for CIFAR-10). Hyperparameters are: $\tau_s = 2.5$, $\tau^* = 1.0$, $\Delta t = 0.01$, $M_{\text{Langevin}} = 200$. We train for 640k iterations using Algorithm 1 with EMA 0.9999 and then 1k more with Algorithm 2 and EMA 0.99 on 7xA100. The batch size is 128, learning rate is 6×10^{-4}, $\varepsilon_{\max} = 0.01$, $\lambda_{\text{CD}} = 1 \times 10^{-3}$, $\alpha = 0.1$, and $\beta = 0.02$. Negatives initialized on the data manifold follow the same temperature schedule as those initialized from the noise.

CelebA: We scale the CIFAR-10 model by $\sim 2\times$; see Figure 8. We set $\tau_s = 2.0$, $\tau^* = 1.0$, $\Delta t = 0.01$, $M_{\text{Langevin}} = 200$, and train for 250k iterations using Algorithm 1 with EMA 0.9999 then 4k with Algorithm 2 and EMA 0.99 on 4xA100. The batch size is 32, learning rate is 1×10^{-4}, $\varepsilon_{\max} = 0.05$, $\lambda_{\text{CD}} = 1 \times 10^{-4}$.

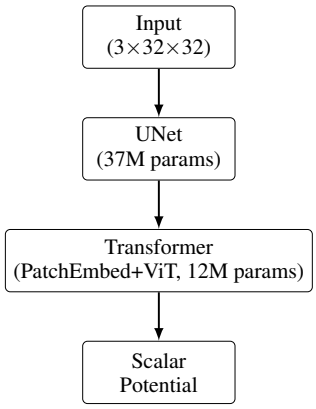
MNIST: We downscale the CIFAR-10 model (Figure 7) to 2M parameters by reducing the UNet base width to 32 channels, using channel multipliers [1, 2, 2], setting the number of attention heads in the UNet to 2, simplifying the Transformer head to an embedding dimension of 128, 2 Transformer layers, 2 attention heads, and adjusting the output scale to 100.0. We set $\tau_s = 2.0$, $\tau^* = 1.0$, $\Delta t = 0.025$, $M_{\text{Langevin}} = 75$, and train for 50k iterations using Algorithm 1 with EMA 0.999 then 3.3k with Algorithm 2 and EMA 0.99 on a single A100. The batch size is 128, learning rate is 1×10^{-4}, $\varepsilon_{\max} = 0.1$, $\lambda_{\text{CD}} = 1 \times 10^{-3}$, $\alpha = 0.0$, and $\beta = 0.05$. Negatives initialized on the data manifold follow the same temperature schedule as those initialized from the noise.

AAV: We adopt the one-dimensional CNN architecture as used in [Bogensperger et al., 2025], summing the final-layer activations to obtain the potential. We train for 10k iterations using Algorithm 1 and for 1k iterations using Algorithm 2 on a single A100. The batch size is 128, learning rate is 1×10^{-4}, $\varepsilon_{\max} = 0.1$, $M_{\text{Langevin}} = 200$, $\Delta t = 0.01$, and $\lambda_{\text{CD}} = 1 \times 10^{-4}$. For Algorithm 3 we use $\tau_s = 1.7$ for AAV medium and $\tau_s = 1.3$ for AAV hard, $\tau^* = 0.9$, $\zeta = 0.01$ for AAV medium and $\zeta = 0.009$ for AAV hard. We set the target fitness to $y = 1$ to aim for the maximum fitness in the generated sequences.

Intuition for other datasets: It is essential for negative samples to reach the equilibrium distribution induced by the model or at least the proximity of the data manifold. The condition $M_{\text{Langevin}} \times \Delta t \gg 1$ is critical to achieving this, with Δt small enough to ensure that negative samples remain of sufficient

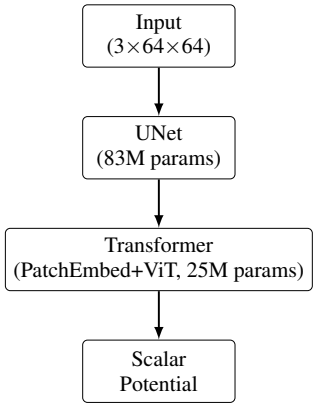
https://doi.org/10.52202/085713-0291

quality—typically the same Δt as used in flow matching for the given generation task. In practice, we set $M_{\text{Langevin}} \times \Delta t = 2$ across most experiments. We recommend starting with $\tau^* = 1.0$ to ensure optimal transport regularization near the data manifold, thereby enhancing training stability. If additional conditions are required during sampling, exploring lower values ($\tau^* < 1.0$) may be beneficial, as this parameter does not need to remain consistent between training and sampling (as shown in Figure 6). Training with $\tau^* < 1.0$ is advised only in special cases, such as extremely low-dimensional problems like that shown in Figure 1, where it is possible to simultaneously be far from the data manifold (from the perspective of the target mode) and close to it (from the perspective of another mode). The parameter $\varepsilon_{\max}$ controls how extensively negative samples explore the space. For unconditional generation, we use $\varepsilon_{\max} = 0.01$, but for inverse problems or design tasks, higher values (e.g., $\varepsilon_{\max} = 0.05$) can improve robustness. The parameter τ_s significantly depends on the task (unconditional or conditional) and thus must be tuned accordingly post-training. Without explicit tuning, selecting $\tau_s = M_{\text{Langevin}} \times \Delta t$ is a reasonable default. Finally, parameters λ_{CD}, α, and β influence the stability of contrastive training and must be empirically determined alongside appropriate early stopping.



Hyperparameter	Value
Image size	3×32×32
Base channels (UNet)	128
ResBlocks	2
channel_mult	[1, 2, 2, 2]
Attention resolution	16
Attention heads (UNet)	4
Head channels (UNet)	64
Dropout	0.1
Transformer (ViT) Head	
Patch size	4
Embedding dim	384
Transformer layers	8
Transformer heads	4
Output scale	1000.0

Figure 7: Diagram of our UNet+Transformer EBM for CIFAR-10 and ImageNet 32x32. A UNet (37M params) processes a 3×32×32 image; its output is fed into a Transformer head (PatchEmbed + 8-layer ViT, 12M params) that produces a scalar potential. Here we employ the identical UNet architecture as in [Tong et al., 2023], but with the time parameter fixed at $t = 0$ to render the model time-independent.



Hyperparameter	Value
Image size	3×64×64
Base channels (UNet)	128
ResBlocks	2
channel_mult	[1, 2, 3, 4]
Attention resolution	16
Attention heads (UNet)	4
Head channels (UNet)	64
Dropout	0.1
Transformer (ViT) Head	
Patch size	4
Embedding dim	512
Transformer layers	8
Transformer heads	8
Output scale	1000.0

Figure 8: Diagram of our UNet+Transformer EBM for CelebA. A UNet (83M params) processes a 3×64×64 image; its output is fed into a Transformer head (PatchEmbed+8-layer ViT, 25M params) that produces a scalar potential.

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: Every substantive claim in the abstract and introduction is substantiated by the theory, experiments, and ablation studies reported in the main text; limitations are stated explicitly. The "Outlook" (Section 4) contains clearly flagged aspirational goals that are presented as future directions rather than accomplished results, ensuring the scope is not overstated.

   Guidelines:

   - The answer NA 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 NA 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: A dedicated "Limitations" paragraph (Section 4) clearly details two concrete constraints of the method. Hence the paper meets the requirement to openly discuss its limitations.

   Guidelines:

   - The answer NA 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.

https://doi.org/10.52202/085713-0291

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: Each theorem lists its full assumptions and is accompanied by a complete proof—intuitive sketches in the main text and formal details in the appendix. When we rely on known results, we cite the originals and restate the necessary arguments for self-containment.

   Guidelines:

   - The answer NA 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 manuscript specifies architectural details, hyper-parameter, and evaluation methods needed for an independent researcher to replicate the reported results.

   Guidelines:

   - The answer NA 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: The experiments rely only on publicly available datasets, and an anonymized code bundle illustrating the full training + evaluation pipeline is attached to the submission; the repository and all ancillary assets needed for reproduction will be made openly available upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All key choices—data splits, hyper-parameters, optimizers, and selection criteria—are documented in the "Training Details" (Section D), while remaining low-level implementation specifics are available in the accompanying code. We follow standard procedures as much as possible.

   Guidelines:

   - The answer NA 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?

https://doi.org/10.52202/085713-0291

Answer: [No]

Justification: Due to high computational costs involved in training and evaluating generative models, metrics like FID are typically computed only a few times, and thus standard deviations are usually not reported. Nevertheless, we do provide standard deviations for the Fitness results obtained on the AAV dataset, where feasible.

Guidelines:

- The answer NA 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: The hardware specifications and number of training iterations (compute resources) are detailed in the "Training Details" (Section D).

   Guidelines:

   - The answer NA 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: The study uses only publicly licensed datasets, releases anonymized code for full reproducibility, involves no human subjects or sensitive personal data, and discloses compute resources and limitations. These practices align with the NeurIPS Code of Ethics on transparency, privacy, attribution, and social responsibility; no deviations are required.

   Guidelines:

   - The answer NA 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: [NA]

    Justification: The method augments an existing class of generative models with a tunable flexibility parameter aimed at inverse problems for science (e.g. protein design) rather than at highest-fidelity large-scale content generation. It (i) trains solely on low-resolution widely used public benchmarks, (ii) releases no new data, and (iii) mirrors the well-studied risk profile of prior open-source generators. Hence it poses no incremental societal risks beyond those already introduced by e.g. flow matching, warranting [NA] .

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The work trains and evaluates its model exclusively on widely used public datasets (MNIST, CIFAR-10, ImageNet32×32, CelebA) and does not release any new data. Because many comparable generative models for these benchmarks are already available, the paper introduces no incremental misuse risk and thus requires no additional safeguards.

    Guidelines:

    - The answer NA 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.

 https://doi.org/10.52202/085713-0291

- 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: All external assets used in this work are standard, publicly-available benchmarks. For each dataset we cite the original creators in the manuscript. Below we summarise the licence associated with every dataset we use:

    - **MNIST** – Creative Commons Attribution–ShareAlike 3.0 (CC BY-SA 3.0).[2]
    - **CIFAR-10** – Custom licence from the authors: free use for research and educational purposes provided the dataset is cited.[3]
    - **ImageNet32×32** –
      - Images inherit the original ImageNet non-commercial research licence.[4]
      - Down-sampling scripts released under the MIT License.[5]
    - **CelebA** – Custom non-commercial research agreement (no redistribution or commercial use; see dataset webpage).[6]
    - **AAV** – Public-domain molecular data (NCBI places no restrictions provided the dataset is cited)[7]

    Guidelines:

    - The answer NA 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: The newly introduced assets are fully documented: detailed algorithm descriptions and the complete implementation are bundled in an anonymized ZIP archive that accompanies this submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

[2] `https://keras.io/api/datasets/mnist/`
[3] `https://ar5iv.org/html/2111.02374`
[4] `https://image-net.org/`
[5] `https://github.com/PatrykChrabaszcz/Imagenet32_Scripts`
[6] `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`
[7] `https://www.ncbi.nlm.nih.gov/bioproject/PRJNA673640/`

- 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: [NA]

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

    Guidelines:

    - The answer NA 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: [NA]

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

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs only used for editing.

    Guidelines:

    - The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-0291