Joint cosmological parameter inference and initial condition reconstruction with Stochastic Interpolants

Carolina Cuesta-Lazaro
IAIFI
Massachusetts Institute of Tenchology
Cambridge, MA 02139, USA
cuestalz@mit.edu

Adrian E. Bayer
Department of Astrophysical Sciences
Princeton University
Princeton, NJ 08544, USA
abayer@princeton.edu

Michael S. Albergo
Society of Fellows
Harvard University
Cambridge, MA 02138, USA
malbergo@fas.harvard.edu

Siddharth Mishra-Sharma
IAIFI
Massachusetts Institute of Tenchology
Cambridge, MA 02139, USA
smsharma@mit.edu

Chirag Modi
The Flatiron Institute
162 5th Ave, New York, NY, 10010, USA
cmodi@flatironinstitute.org

Daniel J. Eisenstein
Center for Astrophysics | Harvard & Smithsonian
60 Garden St., Cambridge MA 02138 USA
deisenstein@cfa.harvard.edu

Abstract

In this work, we present a unified approach to cosmological parameter inference and initial condition reconstruction using Stochastic Interpolants and normalizing flows. We apply this method to jointly reconstruct simulations of non-linear dark matter fields and infer simulator parameters, demonstrating its accuracy and scalability with dataset size. We show that the amortized learned distribution reproduces the posterior obtained with Hamiltonian Monte Carlo without the need for a differentiable forward model or explicit likelihood. Additionally, we introduce a flexible framework for controllable simulators that impose partial constraints, showcasing its potential in generating tailored simulations. This work provides a scalable and accurate approach for reconstructing initial conditions in cosmological simulations, with broad implications for upcoming galaxy surveys such as DESI.

1 Introduction

Simulations are fundamental in scientific research, offering insights into the causal structure of complex problems that may be too costly or impossible to replicate experimentally. Unlike fields where controlled laboratory experiments are feasible, astrophysics relies on the Universe itself as a natural experiment, making simulations essential as its primary laboratory. A key challenge is simulating structures that either replicate a given observation or produce a desired outcome. In this paper, we address the problem of jointly inferring the initial conditions and parameters of scientific simulators, from either an observation or a desired constraint, that lead to simulations that are consistent with an observation.

In cosmology, where the goal is to simulate the Universe on its largest scales, inferring the initial conditions that give rise to an observable requires navigating a parameter space with over 10^8 dimensions. Traditionally, [1–4] combined differentiable simulators [1, 5, 6] with Hamiltonian

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

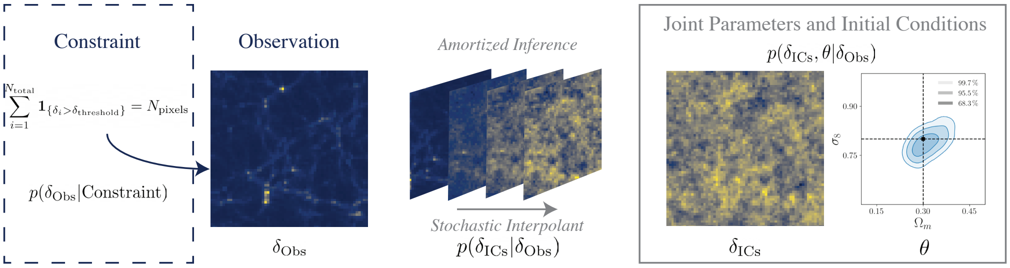

Constraint
$\sum_{i=1}^{N_{\text{total}}} \mathbf{1}_{\{\delta_i > \delta_{\text{threshold}}\}} = N_{\text{pixels}}$
$p(\delta_{\text{Obs}}|\text{Constraint})$
Observation
δ_{Obs}
Amortized Inference
Stochastic Interpolant
$p(\delta_{\text{ICs}}|\delta_{\text{Obs}})$
Joint Parameters and Initial Conditions
$p(\delta_{\text{ICs}}, \theta|\delta_{\text{Obs}})$
99.7%
95.5%
68.3%
0.90
0.75
σ_8
0.15
0.30
0.45
Ω_m
δ_{ICs}
θ

Figure 1: Example initial conditions reconstruction, δ_{ICs}, and parameter inference, θ, starting from a constraint (here, number of pixels above a certain threshold, $\delta_{\text{threshold}}$), or an observation δ_{Obs}.

Monte Carlo (HMC) samplers to tackle this issue. More recently, [7] demonstrated that diffusion models can also approximate the distribution of initial conditions given an observation. Crucially, generative models that directly learn the posterior do not require a differentiable simulator or an explicit likelihood. Here, we extend this line of work by, i) replacing the diffusion model with a Stochastic Interpolant (SI) approach, ii) jointly constraining both the initial conditions and the simulator parameters, and iii) comparing the inferred posteriors with those obtained using HMC, particularly as the number of simulations used for training is varied.

Additionally, in hydrodynamical simulations of galaxy formation, precise control over initial conditions is vital for understanding the formation of different structures. For instance, simulating a Milky Way-like galaxy often involves imposing equilibrium constraints or using zoom-in simulations to find a galaxy that meets approximate criteria. However, these methods lack true controllability. We demonstrate how an amortized initial condition sampler can be combined with conditional generators to control the outcomes of simulations when only partial constraints are available, rather than a full observation.

2 Conditional sampling with Stochastic Interpolants

We aim to constrain the joint distribution of initial conditions, δ_{ICs}, and simulator parameters, θ, given an observation, δ_{Obs}, or a constraint, $\mathcal{C}$, that we wish our observation to satisfy. Let us first focus on the former problem, we will delve into the latter in Section 3.3. Here, we decompose the joint distribution over initial conditions and parameters into its conditionals:

$$p(\delta_{\text{ICs}}, \theta|\delta_{\text{Obs}}) = p(\delta_{\text{ICs}}|\delta_{\text{Obs}})p(\theta|\delta_{\text{ICs}}, \delta_{\text{Obs}}), \tag{1}$$

with a focus on cases where the initial conditions have the same dimensionality as the observations, while the simulator parameters are relatively lower-dimensional. The problem setup and example variables are summarized in Figure 1.

Our approach differs from that of Gibbs sampling [8]. While the method proposed by [8] offers the advantage of explicitly coupling two conditional models into a single diffusion model, this comes at the cost of increased computational expense due to the need for running HMCs over the diffusion model likelihood. In contrast, our inference process is fully amortized. Through experiments, we demonstrate that training the two models independently still yields reliable inference.

In this framework, $p(\theta|\delta_{\text{ICs}}, \delta_{\text{Obs}})$ can be effectively modeled using a Masked Autoregressive Flow (MAF) [9]. Details of the MAF architecture are provided in Appendix A. For $p(\delta_{\text{ICs}}|\delta_{\text{Obs}})$, we note that in cosmological simulations the evolved fields are typically more clustered versions of the initial conditions, with objects displaced by approximately $10\,\text{Mpc}\,h^{-1}$ due to non-linear gravitational evolution. Hence, it is natural to model $p(\delta_{\text{ICs}}|\delta_{\text{Obs}})$ as a mapping between the distributions of initial and final densities, rather than mapping the initial conditions from a normal prior, as done in standard diffusion models [7].

To do this, we employ the formulation of Stochastic Interpolants described in [10, 11] that allows to think about predicting the initial conditions as a problem of probabilistic forecasting. Given an observation δ_{Obs}, we seek to generate a forecast, or probabilistic ensemble, of possible δ_{ICs} associated to it. In particular we build an interpolant, I_s,

$$I_s = \alpha_s x_0 + \beta_s x_1 + \sigma_s W_s, \tag{2}$$

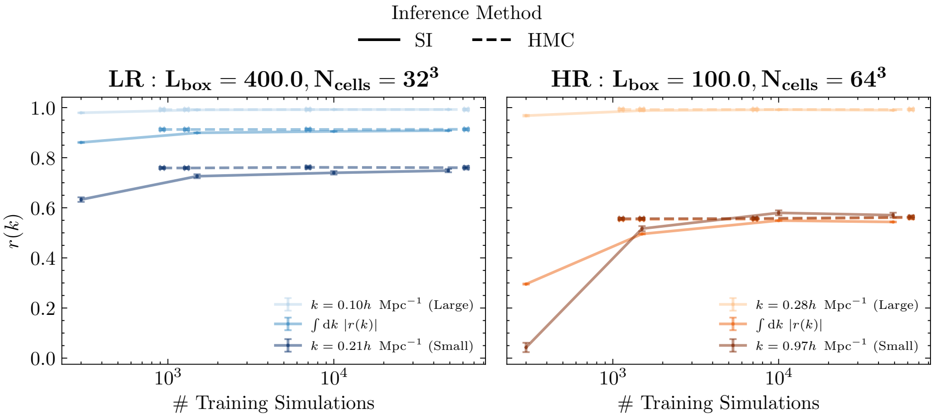


Figure 2: Cross correlation coefficient between reconstructed initial conditions and true initial conditions as a function of the number of simulations used for training, for two different box sizes and resolutions. The dashed lines show the mean and standard deviation of posterior samples obtained with HMC. We show the cross correlation coefficient evaluated at large and small scales , together with the integral over the entire scale range.

where W_s is a Wiener process, that can be sampled by $W_s = \sqrt{s}z$ with $z \sim \mathcal{N}(0, I)$. The interpolant maps a point mass measure at x_0 to $p(x_s|x_0)$ as s varies from 0 to 1. The interpolant boundary conditions are $\alpha_0 = \beta_1 = 1$, and $\alpha_1 = \beta_0 = \sigma_1 = 0$. Here, we also chose $\alpha_s = \sigma_s = 1 - s$, and $\beta_s = s^2$. These boundary conditions imply $x_0 = \delta_{\mathrm{Obs}}$ and $x_1 = \delta_{\mathrm{ICs}}$.

The conditional distribution defined by the interpolant, $p(I_s|x_0)$, is also the law of the solution to a SDE that be used as a generative model [11],

$$dX_s = b_s(x_s, x_0)ds + \sigma_s dW_s,\; X_{s=0} = x_0, \tag{3}$$

whose drift b_s we solve for parametrizing a neural network ansatz $\hat{b}_s(x, x_0)$ and minimizing the objective

$$L_b\left[\hat{b}_s\right] = \int_0^1 ds\mathbb{E}\left[|\hat{b}_s(I_s, x_0, s) - R_s|^2\right], \tag{4}$$

and R_s is determined by the interpolant, $R_s = \dot{\alpha}_s x_0 + \dot{\beta}_s x_1 + \dot{\sigma}_s W_s$. We can then sample the interpolant in Equation 2, and train a neural network to predict the drift used for the generative SDE.

3 Experiments

We simulate non-linear dark matter fields using `pmwd` [6], a differentiable particle-mesh (PM) N-body code. Although the method presented here does not rely on differentiating the forward model, we validate our approach by recovering initial conditions with HMC. To assess the scaling of the algorithm, we generate a large dataset of 50,000 simulations at two different resolutions:

- Low Resolution (LR): Box size $L = 400\,\mathrm{Mpc}\,h^{-1}$ with 32^3 voxels, giving a voxel resolution of $12.5\,\mathrm{Mpc}\,h^{-1}$. Cosmological parameters are either fixed ($\Omega_m = 0.3$, $\sigma_8 = 0.8$) or sampled from uniform priors when performing joint inference.
- High Resolution (HR): $L = 100\,\mathrm{Mpc}\,h^{-1}$ with 64^3 voxels ($1.56\,\mathrm{Mpc}\,h^{-1}$ voxel resolution).

All simulations start from 2LPT with five PM steps from $z = 9$ to $z = 0$, and Gaussian noise of scale 0.1, for fixed cosmology experiments, and 1, for joint inference (in units of the particle shot noise) is added to the output fields.

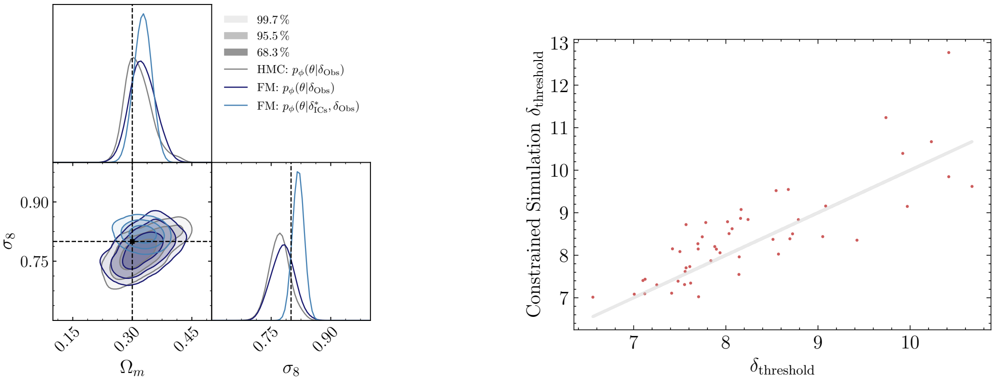
3.1 Validating Initial Conditions sampling with Hamiltonian Monte Carlo

We first compare the SI and HMC posteriors over initial conditions at fixed cosmological parameters; details of the setup can be found in App. B. In Fig. 2, we show the mean and variance of the cross correlation between true initial conditions and those sampled from the learned posterior, as a function of simulations ran. We show results for small and large scales, together with the integrated area under the curve over the entire scale range. In dashed lines, we show the results from explicit inference with HMC. We find that the interpolant samples converge to the HMC posterior when a large number of training simulations is used ($\mathcal{O}(10^4)$). As expected, the convergence with number of simulations is slower for small scales and higher resolutions.

3.2 Jointly sampling initial conditions and simulator parameters

In this section, we validate our joint modelling of initial conditions and simulation parameters, as described in Equation 1, where the model for $p(\delta_{\mathrm{ICs}}|\delta_{\mathrm{Obs}})$ is the same interpolant model presented in the previous section, this time trained by marginalizing over cosmology, and $p(\theta|\delta_{\mathrm{ICs}}, \delta_{\mathrm{Obs}})$ is a Masked Autoregressive Flow (MAF) [9]. Note that the weights of the two models are independent from each other and therefore are trained separately but sampled jointly.

In Figure 3a, we show the agreement of the cosmological parameters inferred with our method (SI) and those from HMC, for a random LR simulation in the test set. Moreover, we include a comparison to a posterior with the initial conditions fixed to the true ones, $p_\phi(\theta|\delta^*_{\mathrm{ICs}}, \delta_{\mathrm{Obs}})$. The true value of the cosmological parameters is shown with a dashed vertical line.



(a) Posteriors over cosmological parameters obtained through Stochastic Interpolants and HMC in a random test simulation. The true value is highlighted with dashed lines.

(b) $\delta_{\mathrm{threshold}}$ in the sampled constrained simulations versus the desired constraint. If the samples perfectly followed the constraints they would lie on the gray line.

3.3 Controllable simulators

We showcase the versatility of our amortized initial condition generator by applying it to produce observations that fulfill certain partial constraints. For example, one could constrain a simulation based on the number of halos with a specific mass and satellite distribution to produce a Milky-Way-like galaxy. Here, we demonstrate this capability with a toy example, constraining the number of pixels, N_{pixels}, in the final density field that exceed a density threshold, $\delta_{\mathrm{threshold}}$.

We approach this problem by first generating a plausible observation, δ_{Obs}, that satisfies the constraint by training an additional conditional model $p(\delta_{\mathrm{Obs}}|N_{\mathrm{pixels}}, \delta_{\mathrm{threshold}})$ through Flow Matching [12] from a Gaussian distribution to the training set density fields. Note that one could directly target the distribution of $p(\delta_{\mathrm{ICs}}|N_{\mathrm{pixels}}, \delta_{\mathrm{threshold}})$ directly, but this is a much harder problem to solve due to the effect of non-linear evolution in the desired condition, whereas the condition can be directly checked in δ_{Obs}.

During training, we uniformly sample N_{pixels} between 1 and 20 and estimate $\delta_{\mathrm{threshold}}$ for each example. The amortized sampler introduced in Section 3.1 then generates initial conditions for the constraint observations. In Figure 3b, we show the compliance of the resulting simulated fields with our desired constraint, when we vary the value of $\delta_{\mathrm{threshold}}$. We find that the simulated density fields only approximately satisfy the input condition and that errors in reproducing the constraints primarily arise from $p(\delta_{\mathrm{Obs}}|N_{\mathrm{pixels}}, \delta_{\mathrm{threshold}})$ only approximately enforcing the conditions, an issue noted in the literature [13]. Future work will focus on improving this conditioning in more scientifically relevant scenarios.

4 Conclusions

We have demonstrated that Stochastic Interpolants efficiently estimate the joint distribution of initial conditions and simulator parameters in cosmological simulations. Compared to HMC, our SI approach is amortized, and does not rely on a differentiable forward model or explicit likelihood. This makes the approach well-suited for current and upcoming galaxy surveys, where developing accurate, differentiable, and fast forward models remains a challenge. Additionally, we showed how the method enables controllable simulators through flexible constraints.

In future work, we plan to extend the method for application to galaxy surveys such as DESI, by training the model on high-resolution, forward-modeled galaxy fields that account for survey masks and systematic effects. Additionally, we aim to enhance the capabilities of our controllable simulators by incorporating more realistic constraints, paving the way for more flexible and accurate simulations. Ultimately, we believe this approach has the potential to significantly improve our ability to reconstruct initial conditions and explore the complex parameter spaces of large-scale cosmological simulations.

Acknowledgments and Disclosure of Funding

We thank Xinyi Chen and Francois Lanusse for useful discussions. This work was performed in part at the Aspen Center for Physics, which is supported by a grant from the Simons Foundation (1161654, Troyer). This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, `http://iaifi.org/`). This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics of U.S. Department of Energy under grant Contract Number DE-SC0012567. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. The computations reported in this paper were in part performed using resources made available by the Flatiron Institute.

References

[1] Jens Jasche and Benjamin D. Wandelt. Bayesian physical reconstruction of initial conditions from large-scale structure surveys. *Monthly Notices of the Royal Astronomical Society*, 432(2): 894–913, April 2013. ISSN 1365-2966. doi: 10.1093/mnras/stt449. URL `http://dx.doi.org/10.1093/mnras/stt449`.

[2] Ludvig Doeser, Drew Jamieson, Stephen Stopyra, Guilhem Lavaux, Florent Leclercq, and Jens Jasche. Bayesian inference of initial conditions from non-linear cosmic structures using field-level emulators, 2023. URL `https://arxiv.org/abs/2312.09271`.

[3] Chirag Modi, Martin White, Anže Slosar, and Emanuele Castorina. Reconstructing large-scale structure with neutral hydrogen surveys. *Journal of Cosmology and Astroparticle Physics*, 2019 (11):023–023, November 2019. ISSN 1475-7516. doi: 10.1088/1475-7516/2019/11/023. URL `http://dx.doi.org/10.1088/1475-7516/2019/11/023`.

[4] Chirag Modi, Martin White, Emanuele Castorina, and Anže Slosar. Mind the gap: the power of combining photometric surveys with intensity mapping. *Journal of Cosmology and Astroparticle Physics*, 2021(10):056, October 2021. ISSN 1475-7516. doi: 10.1088/1475-7516/2021/10/056. URL `http://dx.doi.org/10.1088/1475-7516/2021/10/056`.

[5] Chirag Modi, Francois Lanusse, and Uros Seljak. Flowpm: Distributed tensorflow implementation of the fastpm cosmological n-body solver, 2020. URL `https://arxiv.org/abs/2010.11847`.

[6] Yin Li, Libin Lu, Chirag Modi, Drew Jamieson, Yucheng Zhang, Yu Feng, Wenda Zhou, Ngai Pok Kwan, François Lanusse, and Leslie Greengard. pmwd: A differentiable cosmological particle-mesh n-body library, 2022. URL `https://arxiv.org/abs/2211.09958`.

[7] Ronan Legin, Matthew Ho, Pablo Lemos, Laurence Perreault-Levasseur, Shirley Ho, Yashar Hezaveh, and Benjamin Wandelt. Posterior sampling of the initial conditions of the universe from non-linear large scale structures using score-based generative models, 2023. URL https://arxiv.org/abs/2304.03788.

[8] David Heurtel-Depeiges, Charles C. Margossian, Ruben Ohana, and Bruno Régaldo-Saint Blancard. Listening to the noise: Blind denoising with gibbs diffusion, 2024. URL https://arxiv.org/abs/2402.19455.

[9] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation, 2018. URL https://arxiv.org/abs/1705.07057.

[10] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2023. URL https://arxiv.org/abs/2303.08797.

[11] Yifan Chen, Mark Goldstein, Mengjian Hua, Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and föllmer processes, 2024. URL https://arxiv.org/abs/2403.13724.

[12] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

[13] Jacob K Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models, 2024. URL https://arxiv.org/abs/2402.03559.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[16] Adrian E. Bayer, Uros Seljak, and Chirag Modi. Field-level inference with microcanonical langevin monte carlo, 2023. URL https://arxiv.org/abs/2307.09504.

[17] Adrian E. Bayer, Chirag Modi, and Simone Ferraro. Joint velocity and density reconstruction of the universe with nonlinear differentiable forward modeling. *Journal of Cosmology and Astroparticle Physics*, 2023(06):046, June 2023. ISSN 1475-7516. doi: 10.1088/1475-7516/2023/06/046. URL http://dx.doi.org/10.1088/1475-7516/2023/06/046.

A Model's architectures and training details

A.1 Stochastic Interpolant

The drift term in the Stochastic Interpolant is modelled with a U-Net [14] like architecture with ResNet blocks [15], with $32, 64, 64$ channels per convolution for the HR set, and $64, 128, 128$ for the LR dataset.

All models are trained with a batch size of 16, a starting learning rate of 3×10^{-4} and a learning rate scheduler reducing the learning rate by a factor of 10 when the validation loss plateaus, for 50000 gradient steps. We pick the checkpoint with lowest validation loss.

A.2 Normalizing flow

The normalizing flow consists of a summarizer, a ResNet [15] architecture of the same number of channels as the Stochastic Interpolant described above and a summary dimension of 128, and Masked Autoregressive Flow, consisting of five transforms with three MLP layers each, and a hidden dimensionality of 128.

B Inference with Hamiltonian Monte Carlo

To perform inference with HMC we explicitly sample the posterior,

$$-2\log P(z,\theta|\delta_{\mathrm{Obs}}) = \sum_{\vec{k}} \left[\frac{|f_{\vec{k}}(z,\theta) - \delta_{\mathrm{Obs},\vec{k}}|^2}{N} + |z_{\vec{k}}|^2 \right], \tag{5}$$

where z are the standard normal Fourier modes of the initial overdensity field δ_{IC}, θ are the forward model parameters, and N is the variance of the injected noise. The first term is the likelihood and is evaluated using the same forward model f as was used to generate the data. The second term is the prior, resembling the standard normal nature of the initial Fourier modes. The sum is performed over all voxels in k-space.

A conservative 1000 warm-up steps were taken during which the step size and mass matrix for the cosmological parameters were tuned. The mass matrix for the initial conditions was set analytically as in [16]. Annealing was applied during NUTS warm-up to first converge on large scales and then on progressively smaller scales, as in [17]. After warm-up, NUTS was run for 10,000 iterations, with a maximum tree depth of 5 for initial condition reconstruction at fixed cosmology, and 10 when jointly inferring cosmological parameters, as this was required to achieve convergence. This was repeated 5 times from different starting points to produce 5 chains of samples. It would be interesting further work to benchmark against microcanonical samplers [16].