Generative, energy-based models for diagnostic reconstruction and analysis

Phil Travis[1], Troy Carter[2]

1. Ergodic LLC
2. Oak Ridge National Laboratory

phil@physicistphil.com

1 Motivation: learn relationships from data

Trend inference can be done using machine learning

- Long-term goal: self-optimizing fusion reactors or other plasma devices
- Insight can be gained from models learned on data

Generative, energy-based models are useful and flexible for plasma science

- EBMs **provide insight** by learning exploitable structure
- The model can be sampled to **reconstruct any diagnostic**
- Uncertainty is intrinsic to the model
- Learned models can be composed after training

Previous work:
trend inference in and optimization of LAPD mirrors

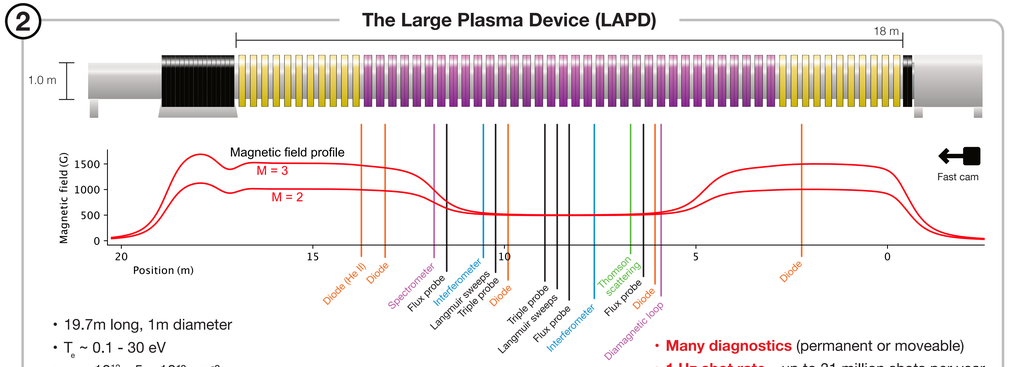
Publication
https://doi.org/10.1063/5.0270755

GitHub repository
https://github.com/physicistphil/lapd-isat-predict

This work:
generative modeling of LAPD mirrors

Dissertation (chapter 6 is on EBMs)
https://physicistphil.com/PhD/

2 The Large Plasma Device (LAPD)

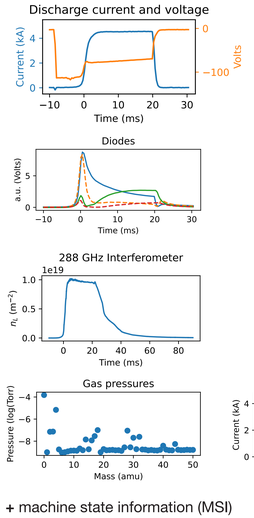


- 19.7m long, 1m diameter
- T_e ~ 0.1 - 30 eV
- n_e ~ 10^{12} - 5 x 10^{13} cm^{-3}
- Flexible magnetic geometry — good for mirror studies

- **Many diagnostics** (permanent or moveable)
- **1 Hz shot rate** – up to 31 million shots per year
- **Data-rich environment** — ideal for ML

3 Collecting a comprehensive datasets

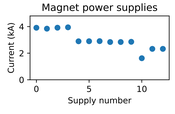
Signals

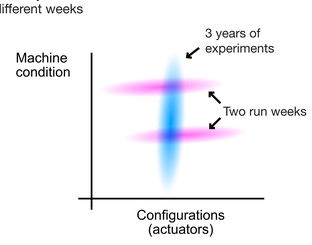


+ machine state information (MSI)

Datasets

Timeframe	Two run weeks	Three years continuous
Dates	Feb 2023, Apr 2024	Sep 2021 to Oct 2024
Number of shots	130 thousand	**29 million** subsampled to 250 thousand
Probe data	**Yes**	No
MSI + passive diagnostics	Yes	Yes
Configuration diversity	**Yes, ~60 randomly sampled**	No, standard configurations
Condition diversity	No, only two different weeks	**Yes**





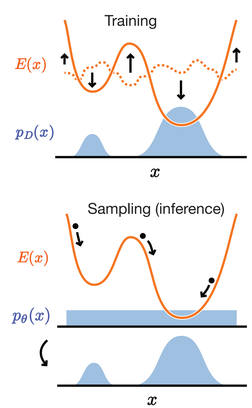
4 Learning with Energy-Based Models (EBMs)

EBM basics

$$p_\theta(x) \propto e^{-E_\theta(x)}$$

$$E_\theta(x) = \mathrm{NN}_\theta(x)$$

Learned parameters θ



Algorithm 1 EBM training algorithm

Require: Training samples x_i^+, training data distribution p_D, energy function E_θ, replay buffer $\mathcal{B}$, step size ϵ, MCMC steps L, KL MCMC steps K, energy regularization strength α, stop gradient operator $\Omega(\cdot)$, replay fraction $f_\mathcal{B}$, batch size M

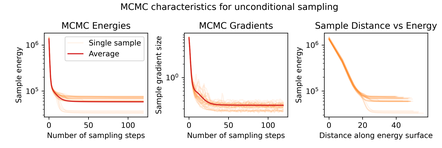
$\mathcal{B} \leftarrow \mathcal{U}(-1,1)$ ▷ Fill buffer from uniform distribution
while not converged **do**
  $x_i^+ \sim p_D$
  $\tilde{x}_i^0 \sim \mathcal{B}$ sample $Mf_\mathcal{B}$ negative examples, $\mathcal{U}(-1,1)$ otherwise
  $X^- \sim \mathcal{B}$ nearest-neighbor samples such that $X \cap \tilde{x}_i^0 = \varnothing$

  for sample step $\ell = 1$ to L **do** ▷ Run Langevin dynamics
    $\tilde{x}_i^\ell \leftarrow \tilde{x}_i^{\ell-1} - \frac{\epsilon^2}{2}\nabla_x E_\theta(\tilde{x}_i^{\ell-1}) + \epsilon\mathcal{N}(0,1)$
  end for
  $\tilde{x}_i^\ell = \Omega(\tilde{x}_i^\ell)$

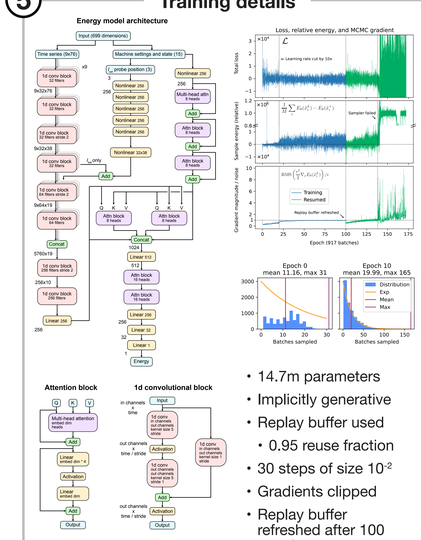
  $\tilde{x}_i^0 = \tilde{x}_i^\ell$ where $\ell = L - K$ ▷ Run Langevin dynamics for KL loss
  for KL sample step $k = 1$ to K **do**
    $\tilde{x}_i^k \leftarrow \tilde{x}_i^{k-1} - \frac{\epsilon^2}{2}\nabla_x E_\theta(\tilde{x}_i^{k-1}) + \epsilon\mathcal{N}(0,1)$
  end for

  $\mathcal{L}_{\mathrm{CD}} = \frac{1}{M}\sum_i E_\theta(x_i^+) - E_\theta(\tilde{x}_i^\ell)$
  $\mathcal{L}_{\mathrm{KL}} = \frac{1}{M}\sum_i E_{\Omega(\theta)}(E_\theta(\tilde{x}_i^K) - \mathrm{NN}(X, \tilde{x}_i^K)$ ▷ Has gradients through MCMC
  $\mathcal{L}_{\mathrm{reg}} = \frac{1}{M}\sum_i E_\theta(x_i^+)^2 + E_\theta(\tilde{x}_i^\ell)^2$
  $\mathcal{L} = \mathcal{L}_{\mathrm{CD}} + \mathcal{L}_{\mathrm{KL}} + \alpha\mathcal{L}_{\mathrm{reg}}$

  Apply $\nabla_\theta\mathcal{L}$ to θ via the Adam optimizer
  $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{U}(-1,1)$ and remove samples to maintain buffer size
end while



5 Training details



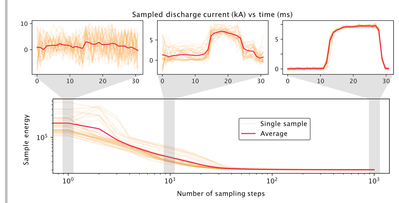
- 14.7m parameters
- Implicitly generative
- Replay buffer used
  - 0.95 reuse fraction
- 30 steps of size 10^{-2}
- Gradients clipped
- Replay buffer refreshed after 100 epochs

6 Unconditional sampling: learning all modes of the distribution

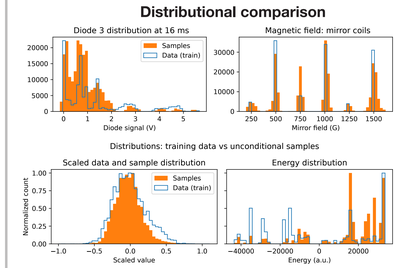
- Sample from the learned distribution:
  $\tilde{x} \sim p_\theta(x)$
- Samples are initialized on uniform noise
- Use Langevin dynamics over the energy surface

$$\tilde{x}_i^0 \sim \mathcal{U}(-1,1)$$

$$\tilde{x}_i^\ell \leftarrow \tilde{x}_i^{\ell-1} - \frac{\epsilon^2}{2}\nabla_x E_\theta(\tilde{x}_i^{\ell-1}) + \epsilon\mathcal{N}(0,1)$$

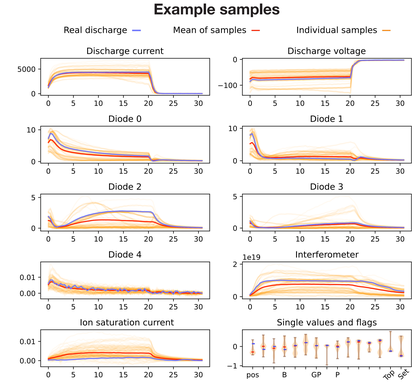


Distributional comparison

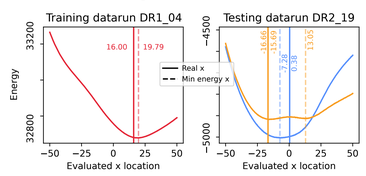


- **Model learned the vast majority of modes**
- Probability of samples and data are comparable

Example samples



- Symmetry in probe position is evident (at times)
- **Information is embedded in the energy surface**
- Relationships need not be invertible



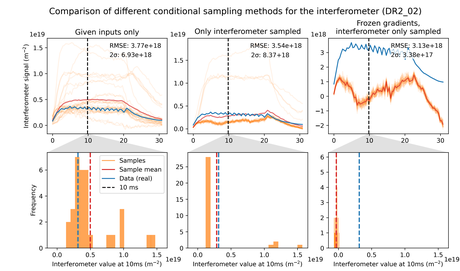
7 Conditional sampling: diagnostic reconstruction

- Conditionally sample a diagnostic from the learned distribution:

$$\text{ifo} \sim p_\theta(\text{ifo}|\text{other signals})$$

- Freeze gradients (traditional way in literature), or
- **Use an auxiliary energy function (novel)**

$$E_{\mathrm{cond}}(\tilde{x}) = E(\tilde{x}) + F(\tilde{x}), \quad F(\tilde{x}) = \left(\frac{\tilde{x} - x_i}{2\epsilon}\right)^2 \rightarrow \text{Gaussian constraint}$$



- Frozen gradients lead to **unphysical solutions**
- Additional diagnostics improve interferometer reconstruction
- **Can reconstruct any diagnostic using an EBM**
- **EBMs intrinsically have uncertainty**

RMSE and 2σ of conditional interferometer samples for test set and DR2_02

Given:	LAPD settings only	All signals	Frozen gradients
RMSE (test set)	4.12×10^{18}	2.91×10^{18}	3.13×10^{18}
RMSE (DR2_02)	3.77×10^{18}	3.54×10^{18}	2.51×10^{18}
$2\ \sigma$ (DR2_02)	6.93×10^{18}	8.37×10^{18}	3.38×10^{17}
Training RMSE	4.40×10^{17}		

8 Summary and future work

- Energy-based models (EBMs) are an incredibly flexible way of modeling data
- An EBM was trained on diverse LAPD data
  - Learned nearly all modes of the distribution
  - Insight is gained directly from the energy function
  - Can reconstruct any diagnostic with or or without other signals
- Next step: train model on 29m-shot dataset
- Final step: compose the two models to extend results to different machine conditions
  - Add energy functions → joint distribution

Acknowledgments

The authors would like to thank the following individuals for diagnostics setup, run support, and mentorship:

- Jia Han
- Tom Look
- Lukas Rovige
- Chris Niemann
- Shreekrishna Tripathi
- Steve Vincena
- Yhoshua Wug
- Zoltan Lucky
- Pat Pribyl
- Tony Peebles
- Ryan Buckley
- Walter Gekelman
- Jacob Bortnik

Useful references

- Generative modeling:
  - Ari Seff, Wenhan Cao, Ryan Prescott Adams, Yilun Du, Yang Song, Mengjiao Yang, Zhisheng Xiao, Tian Han, Jascha Sohl-Dickstein. Denoising Diffusion Probabilistic Models. June 2020. arXiv:2006.11239 [cs, stat].
  - Kevin P. Murphy. Probabilistic Machine Learning: Advanced Topics. MIT Press, 2023. URL: http://probml.github.io/book2.
  - Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. April 2022. arXiv:2204.06125 [cs].
  - Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. April 2022. arXiv:2112.10752 [cs].
  - Jiaming Song, Chenlin Meng, Stefano Ermon. Denoising Diffusion Implicit Models. November 2020. arXiv:2010.02502 [cs, stat].
  - Yang Song, Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. July 2020. arXiv:1907.05600 [cs, stat].
  - Dustin Tran, Keyon Vafa, Kumar Agrawal, Laurent Dinh, Ben Poole. Discrete Flows: Invertible Generative Models of Discrete Data. May 2019. arXiv:1905.10347 [cs, stat].
  - Lilian Weng. What are diffusion models? July 2021.
- EBMs:
  - Ruiqi Gao, Erik Nijkamp, Diederik P. Kingma, Zhen Xu, Andrew M. Dai, Ying Nian Wu. Training Discriminative Models to Evaluate Sample Quality. May 2023. arXiv:2302.02544 [cs, stat].
  - Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, Chris Maddison. Oops I Took A Gradient: Scalable Sampling for Discrete Distributions. August 2021. arXiv:2102.05379 [cs, stat].
  - Erik Hoogeboom, Vikor Garcia Satorras, Clément Vignac, Max Welling. Equivariant Diffusion for Molecule Generation in 3D. June 2022. arXiv:2203.17003 [cs, stat].
  - Erik Hoogeboom, Vincent Tom Peluchetti, Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. July 2021. arXiv:2102.05379 [cs, stat].
  - Erik Nijkamp, Mitch Hill, Song-Chun Zhu, Ying Nian Wu. On the Anatomy of MCMC-Based Maximum Likelihood Learning of Energy-Based Models. April 2020. arXiv:1903.12370 [cs, stat].
  - Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, Ying Nian Wu. On the Anatomy of MCMC-Based Maximum Likelihood Learning of Energy-Based Models. In AAAI Conference on Artificial Intelligence, 34(04):5272–5280, April 2020. ISSN 2374-3468, 2159-5399. doi:10.1609/aaai.34i04.5973.
  - Gradient clipping: Prem Seetharaman, Gordon Wichern, Bryan Pardo, and Jonathan Le Roux. AutoClip: Adaptive Gradient Clipping for Source Separation Networks. July 2020. arXiv:2007.14469 [cs, eess, stat].

Work was performed at the Basic Plasma Science Facility (BaPSF), UCLA, supported by DOE FES