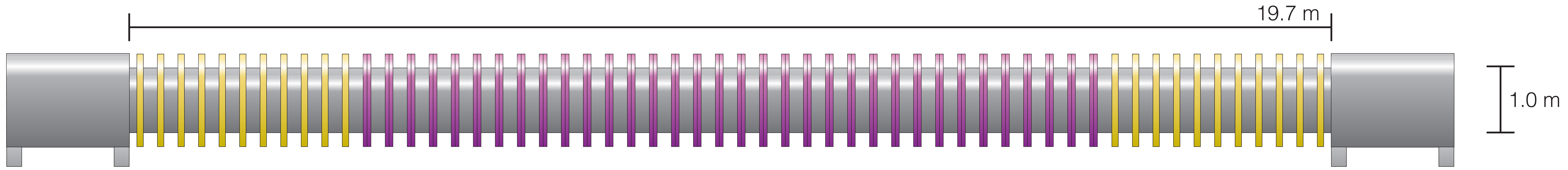
Upgrading LAPD diagnostic pipelines for training generative ML models

Phil Travis (phil@physics.ucla.edu), Steve Vincena
UCLA Basic Plasma Science Facility



The Large Plasma Device (LAPD)

- 19.7m long, 1m diameter
- T_e ~ 5-10 eV
- n_e up to ~10^{13} cm^{-3}
- 1 Hz shot rate – up to 31 million shots per year
- Hundreds of diagnostics ports
- **Data-rich environment**

1 LAPD data sources

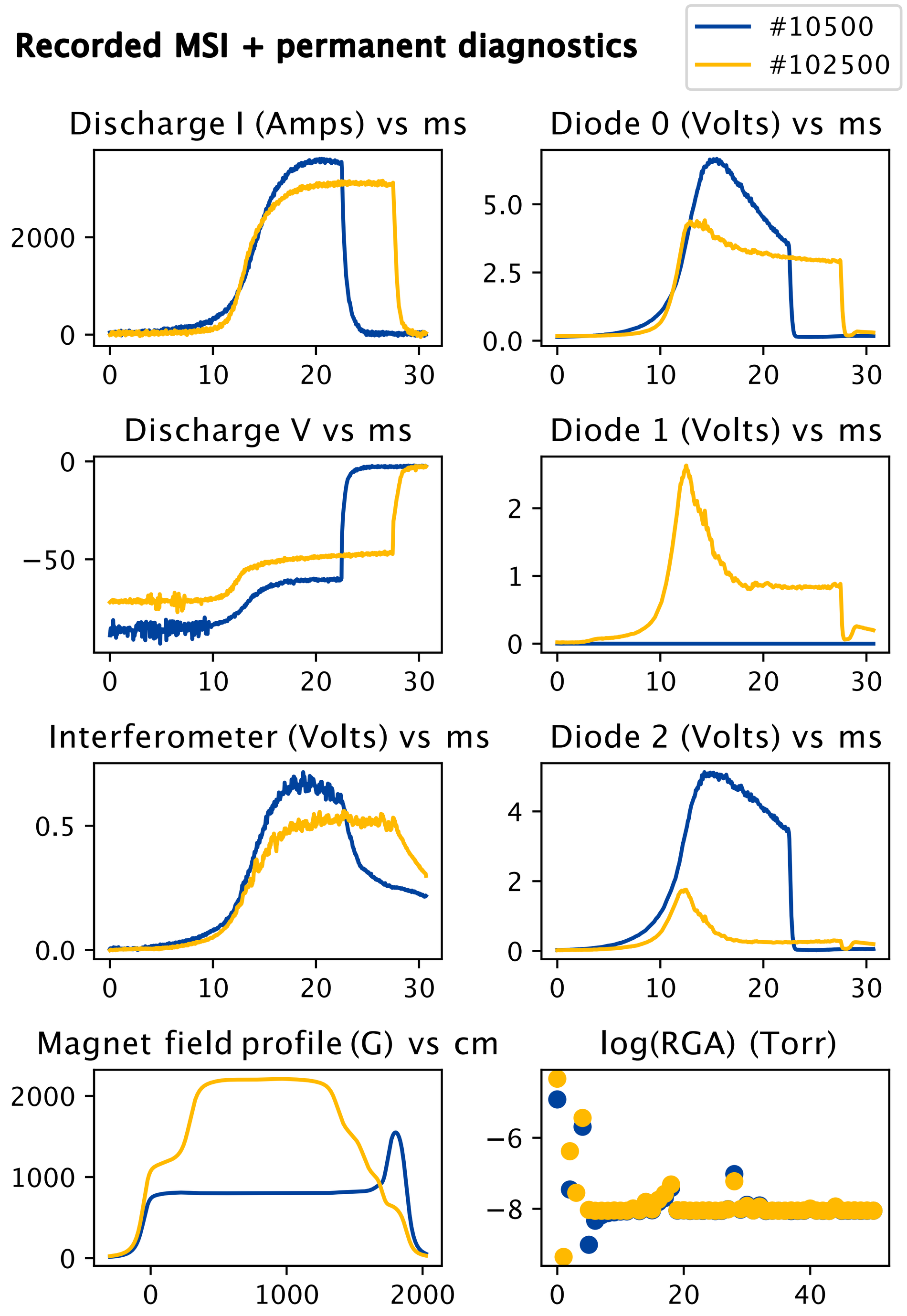
Fixed diagnostics
- Interferometers (x1)
- Visible light diodes (x3)
- Fast framing camera

Probe diagnostics (mobile)
- Langmuir probes
  - Ion saturation current
  - Floating potential
  - Langmuir sweeps (Te)
- Hairpin resonator density measurements
- Magnetic fluctuations

Machine state information (MSI)
- Discharge current
- Discharge voltage
- Gas pressure
- RGA partial pressures
- Axial magnetic field

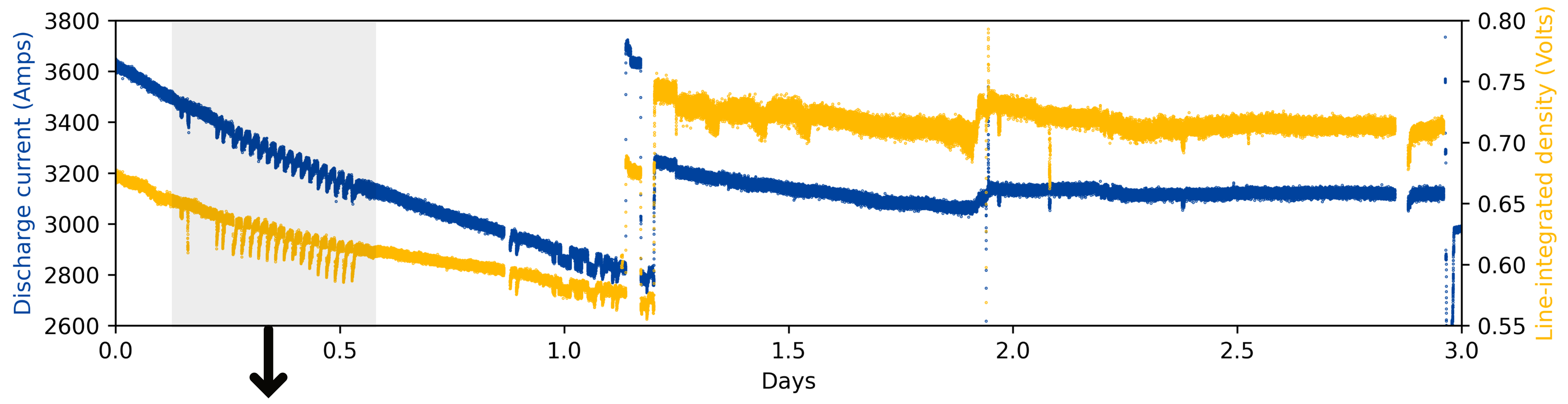
This work
- Digitization of MSI and fixed diagnostics
- Creation of auxiliary data pipeline
- Preliminary analysis of MSI
- **Training of energy-based model on MSI and fixed diagnostics**

Recorded MSI + permanent diagnostics



2 Preliminary analysis: long-timescale trends observed

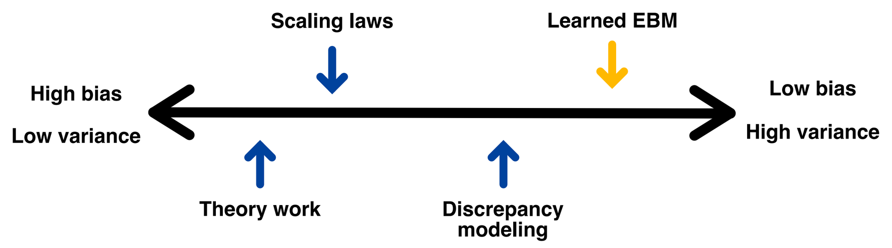
Flattop discharge current and interferometer measurements



- **Changes in discharge current likely in-part caused by probe positions**
- Current assumption: plasmas do not change significantly shot-to-shot
  - **May be able to relax this assumption using ML**

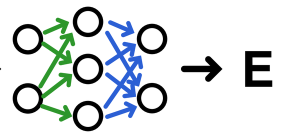
High-variance plasma modeling

- Discharge signals may contain information that is difficult to exploit
- In a high-variance (learned) approach:
  - All effects accounted for in prediction
  - Model has few preconceived notions
- Learning patterns permits automated exploration

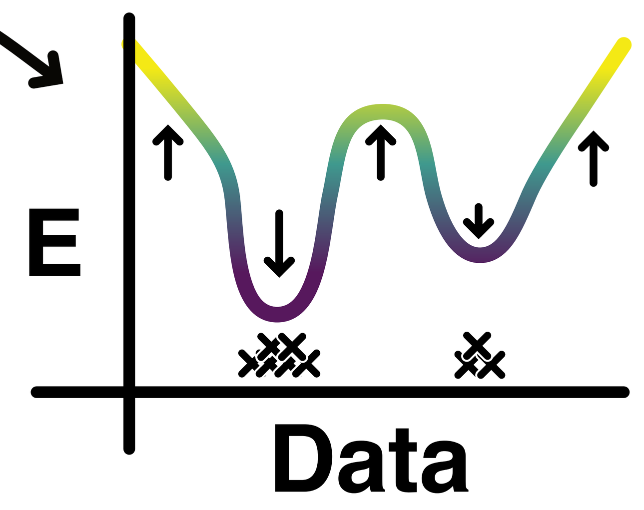


3 Introduction to energy based models (EBMs)

$p(x) \sim e^{-\beta E(x)}$

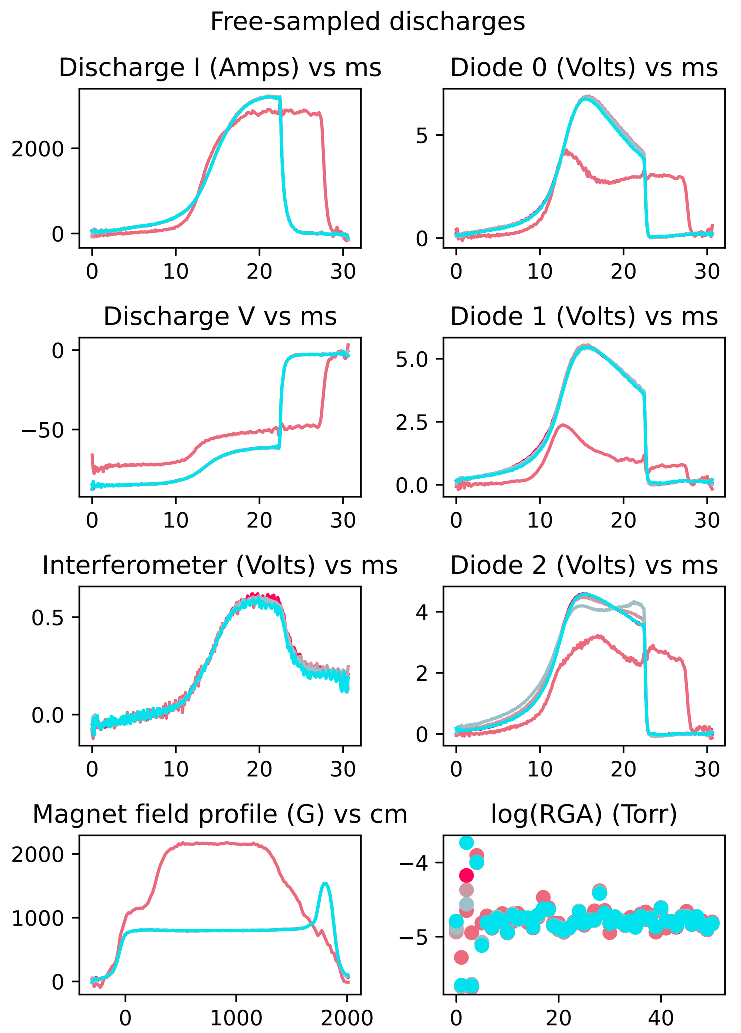


- Assigns an energy value to configurations of input variables — generative
- Trained by pushing energy down on data, up on samples (contrastive divergence)
- Learns the relationship between all input variables – predict anything from anything
  - Conditional sampling is easy
  - Solution to inverse problems are built-in
  - Can fill in missing data
- Energies are additive: can easily combine models



4 Training and sampling from an EBM

Training specifications
- 130756 shots (training set)
  - Downsampled from 25 kHz to 8.33 kHz
- Convolutional + dense nets
- 2,068,497 parameters



- Trained using contrastive divergence
- Based on methodology outlined in:
  Du & Mordatch (2020) arXiv:1903.08689v6
  Du et al. (2021) arXiv:2012.01316v4
- Sample from models: Langevin dynamics

$$\ddot{x} = -\nabla E(x) + \sqrt{T}\mathcal{N}(0,1)$$

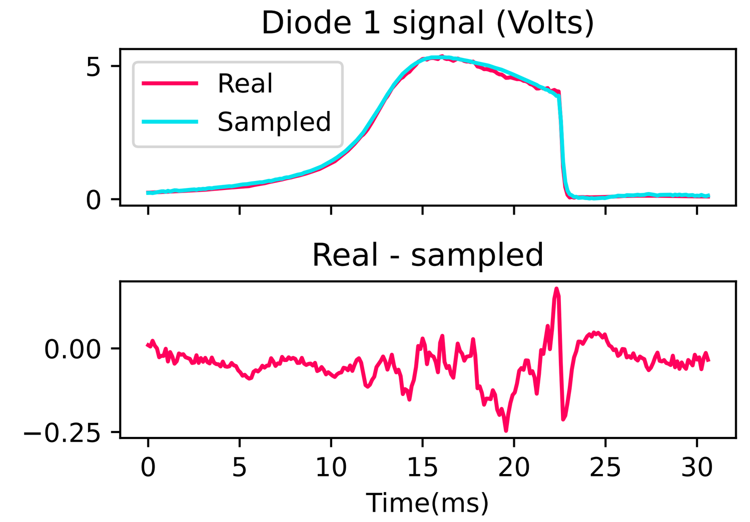
Energy surface ↑ Gaussian process ↑

- Conditionally sample: fill in missing data

$$y \sim p(y \mid x)$$

Missing signal ↑ Existing signals ↑

Example: reconstructing diode signal (from validation dataset)



Summary

- Data pipeline was constructed for machine state information (MSI) and fixed diagnostics (an interferometer and diodes)
- Preliminary data exploration undertaken: long-term trends observed in plasma discharge parameters
- Energy-based models (EBMs) learn a probability distribution by assigning an energy value to each input configuration
- EBMs can be freely sampled to generate synthetic discharges
- EBMs can be conditionally sampled to fill in missing signals

Future work

- Integrate more diagnostics into pipeline: interferometers (x3-7), visible light diodes (x3-6), a magnetic pickup coil, permanent Langmuir probes, spectrometers (x1-3), a spectral line ratio diagnostic
- Include probe diagnostics in training
- Uncover trends by sampling conditionally from energy-based models
  - Characterize new plasma source
  - Infer density profiles from a few shots or less

Work supported by the US DOE and NSF, and performed at the Basic Plasma Science Facility (BaPSF), UCLA.