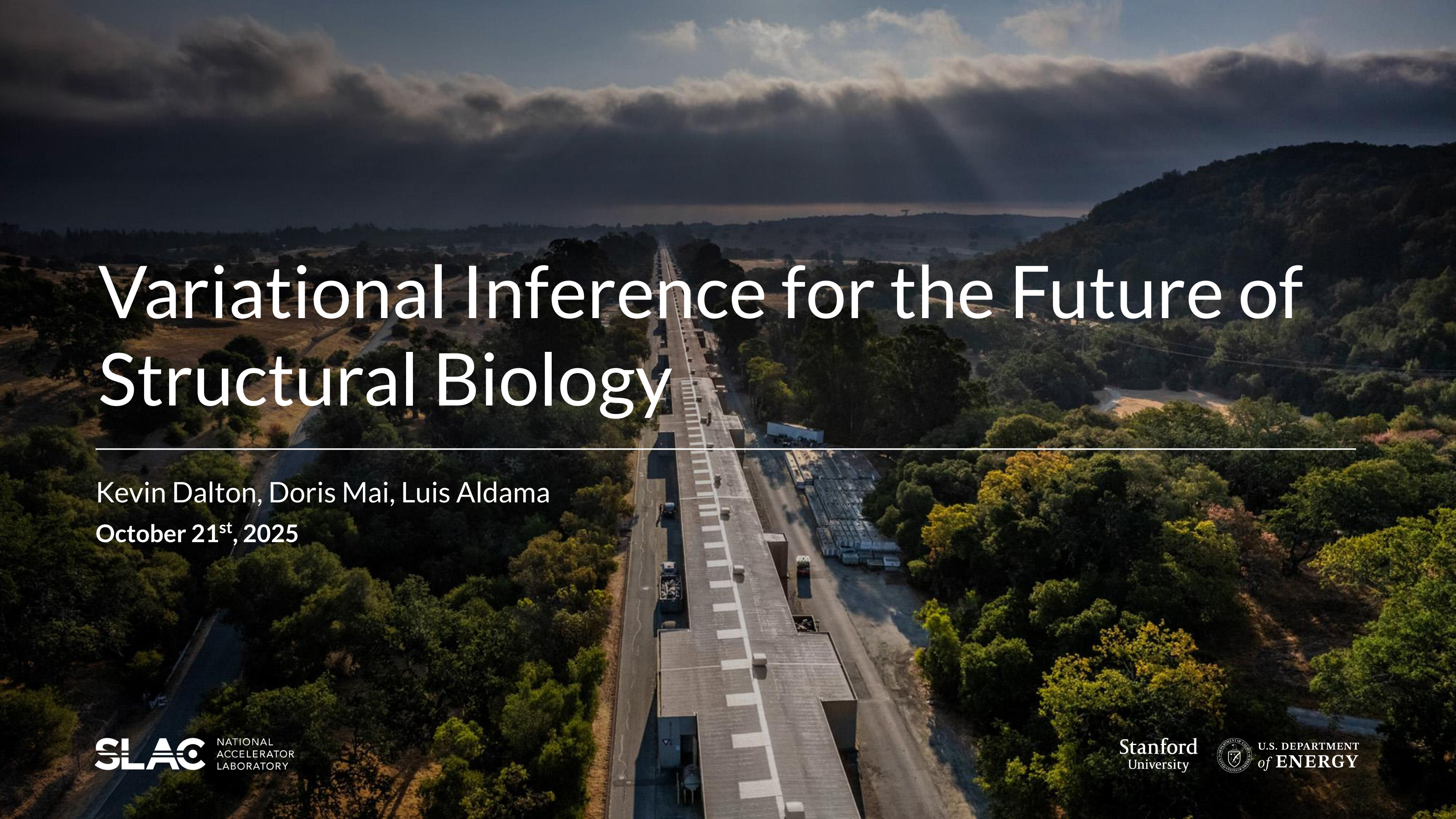


Variational Inference for the Future of Structural Biology

Kevin Dalton, Doris Mai, Luis Aldama

October 21st, 2025

SLAC NATIONAL ACCELERATOR LABORATORY

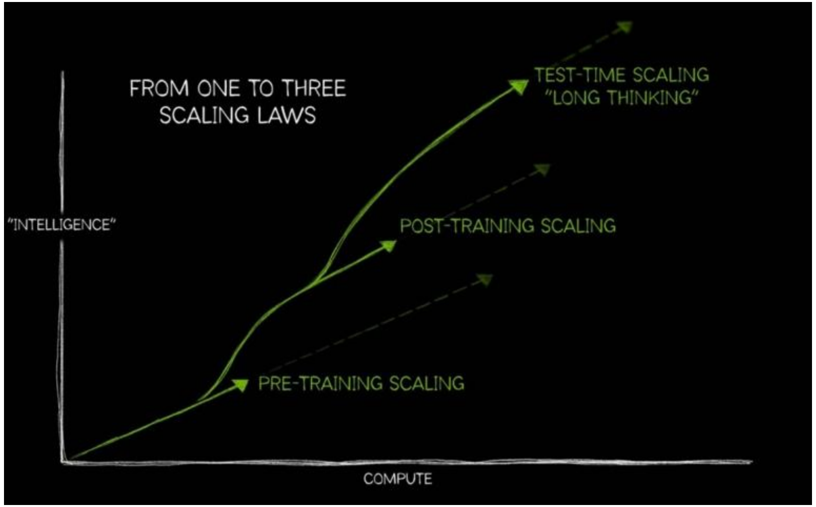
Stanford University

U.S. DEPARTMENT of ENERGY

Contemporary Machine Learning

Stargate Data Center (*OpenAI)

- 10 gigawatts
- $500 billion

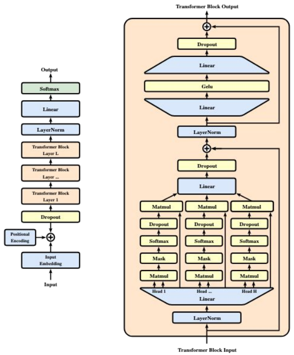


Inference Time Compute (*NVIDIA)

- Generative models
- Frozen weights

Comparison of Model Philosophies

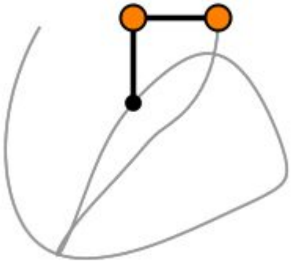
Pre-Trained



Foundation Models

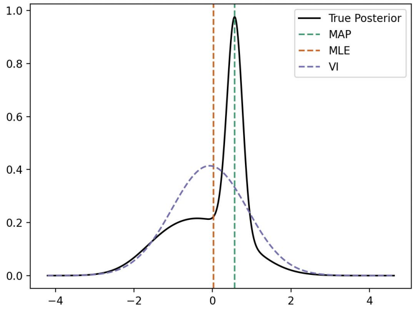
- Large ML models are
- Scalable at inference time
- Data hungry
- Not interpretable

Re-Trained



Physical Models (*Jousef Murad)

- Very interpretable
- May include local parameters
- May require resources to scale



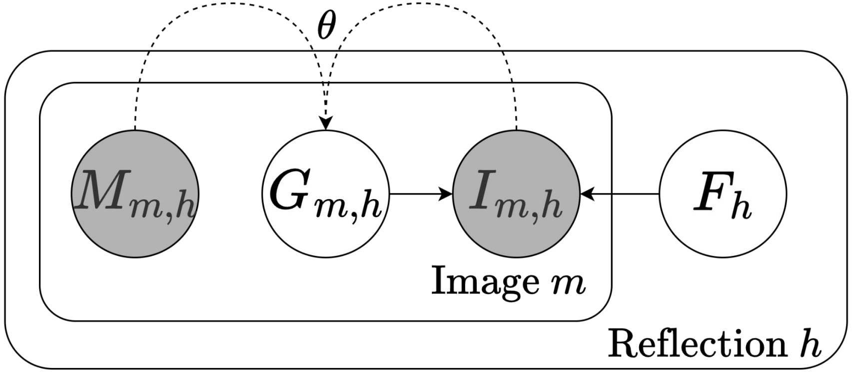
Stochastic Variational Inference

- Use DNNs and Bayesian inference
- Scalable and
- Interpretable
- Can integrate physical models

Conventional Physical Models

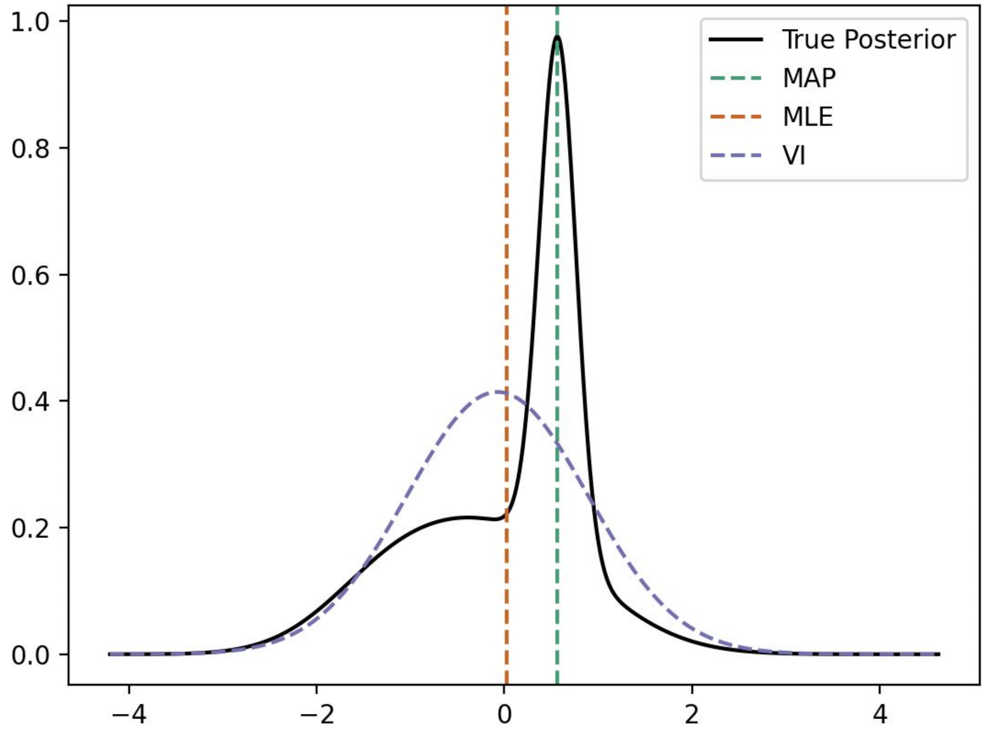
- Physical models often built on inherently local parameters that prevents scalability
- Large ML models are scalable at inference time but data hungry
- ML parameterized VI enables both physical interpretability and scalability

Variational Inference: Interpretable Model of Data Generation



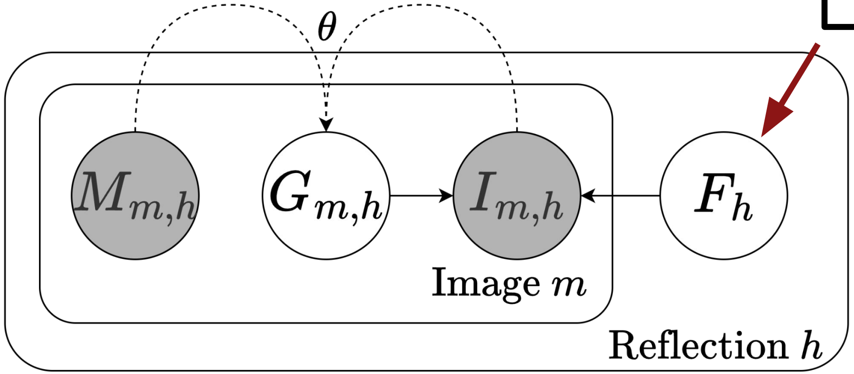
- **I**: observed scattering intensities
- **F**: Fourier coefficients of the electron density
- **G**: Systematic error in measurements
- **G** is the output of a neural network parameterized by θ
- **M**: is the metadata about each reflection observation, **I**
- **F** and θ are jointly estimated by optimizing the **ELBO**

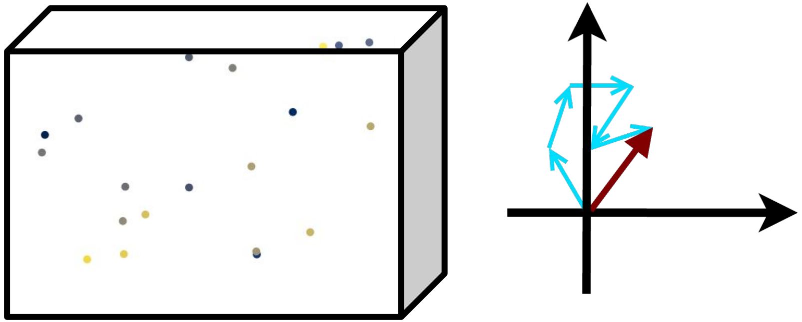
Variational Inference: Rigorous Uncertainty Estimates



Variational Inference: Natural Ways to Include Prior Info

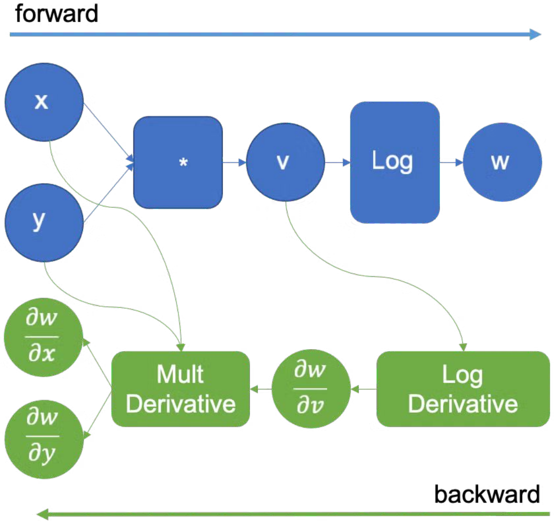
- Diffraction intensities
- Atomic structures
- Molecular sequences
- …

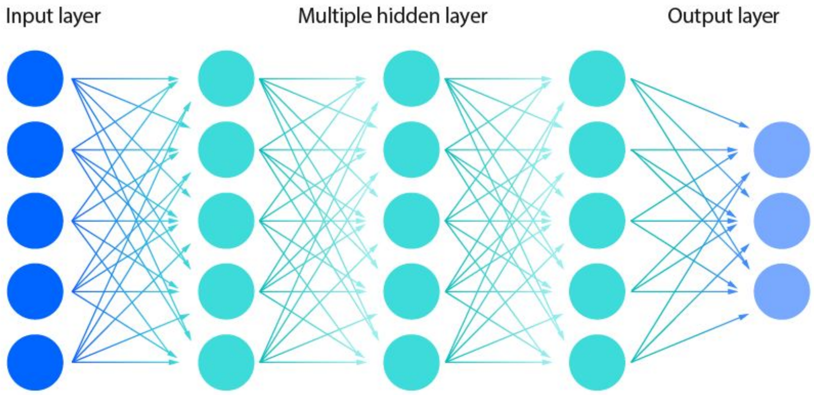




Random atom model
(Wilson distribution)

Variational Inference: Compatible with Modern ML





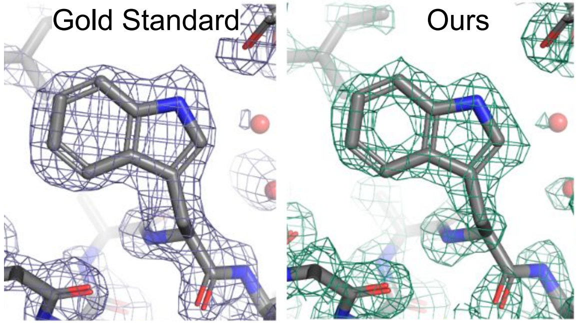
Autograd (*PyTorch)

- Computes gradients of distributions
- Easily add physics-based functions

Deep Neural Networks (*IBM)

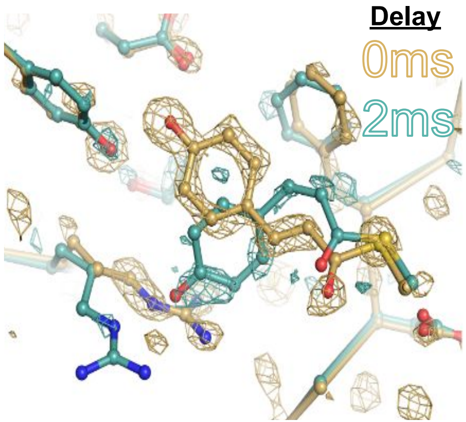
- Black box learnable functions
- Implicit representations
- Many architectures CNNs, MLPs, Transformers, etc

Variational Inference: Flexibly Process Diverse Data Sources



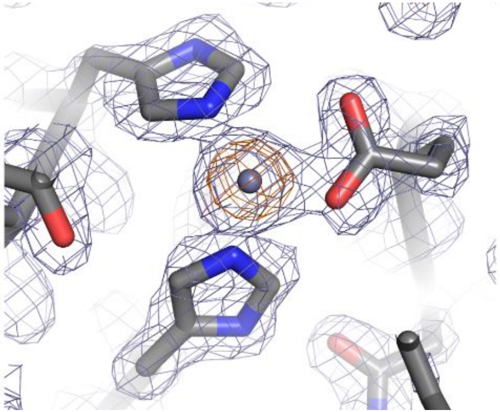
Conventional Diffraction

Ab initio phasing of hen egg white lysozyme from native sulfur



Time-Resolved Laue

Time-resolved, polychromatic diffraction of photoactive yellow protein

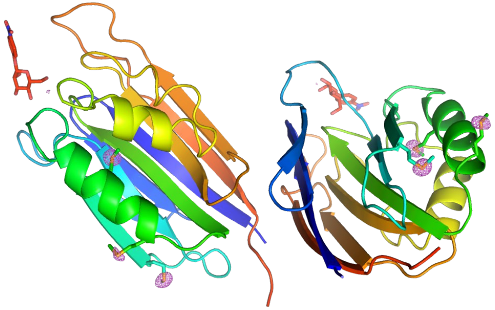


Serial-Femtosecond

Serial crystallography of a zinc metalloprotease from LCLS

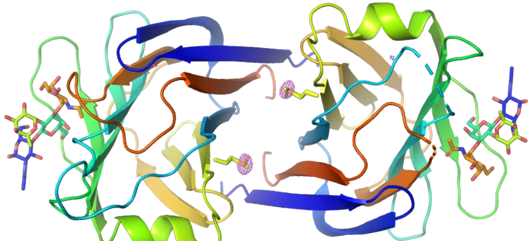
*Dalton KM, Greisman JB, Hekstra DR. Careless: A Variational Bayesian Model for Merging X-ray Diffraction Data. bioRxiv. 2021 Jan 7;2021.01.05.425510.

Variational Inference: Scale to Large Data Sets



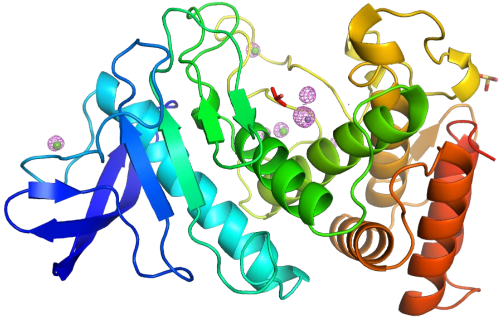
CXIDB-61

26,583 Images
1.4 Å Resolution Cutoff
SACLA



CXIDB-62

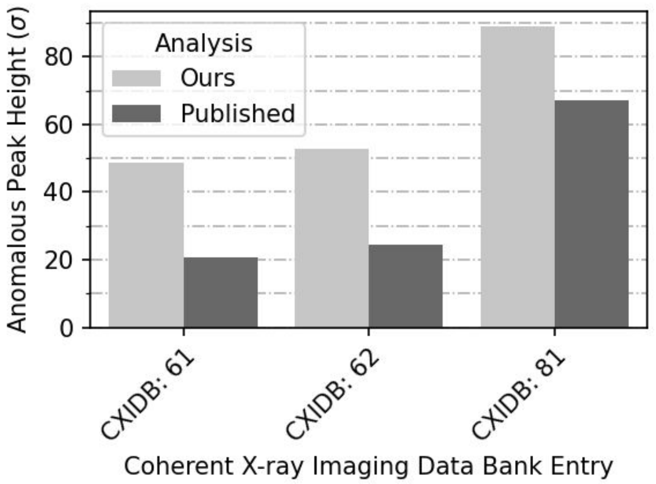
133,242 Images
1.5 Å Resolution Cutoff
SACLA

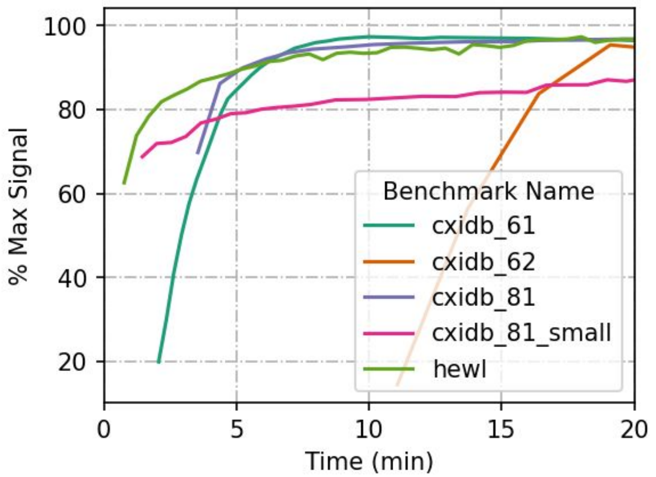


CXIDB-81

164,639 Images
1.8 Å Resolution Cutoff
LCLS

Variational Inference: Scale to Large Data Sets





Anomalous Peak Heights

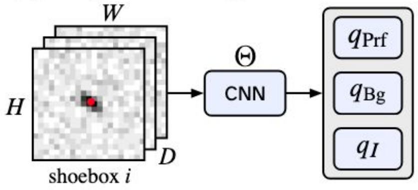
- State of the art results
- No hyperparameter tuning

Training Time

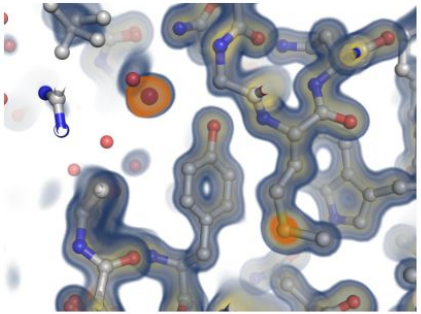
- Training converges in < 20 min for all datasets
- Single A100 GPU

Variational Inference: Easy to Extend

(a) Diagram of integration network



(b) electron density map produced by differentiable integrator



(c) Iodide anomalous peak heights (σ)

Epoch	204 IOD	205 IOD	206 IOD
epoch 1	30.23	13.19	29.15
epoch 3	31.82	13.90	30.24
epoch 5	33.16	14.58	30.65
epoch 7	33.77	15.10	31.14
epoch 9	32.33	14.42	30.03
epoch 11	**34.09**	**15.26**	**30.88**
epoch 13	32.02	14.02	30.10
Ref. (DIALS)	**32.68**	**14.75**	**29.69**

Luis Aldama

Doeke Hekstra

HARVARD UNIVERSITY

Estimating Photon Flux with Variational Inference

- Use VI to estimate photons scattered to Bragg peaks
- Amortized intensity, background, and profile
- SOTA performance on hen egg white lysozyme dataset

Acknowledgements

U.S. DEPARTMENT OF ENERGY

SLAC

LCLS

Jana Thayer

Frédéric Poitevin

Doris Mai

Hekstra Lab

Minhuan Li

Luis Aldama

Flavia Giehr

Variational Inference

VI Algorithm

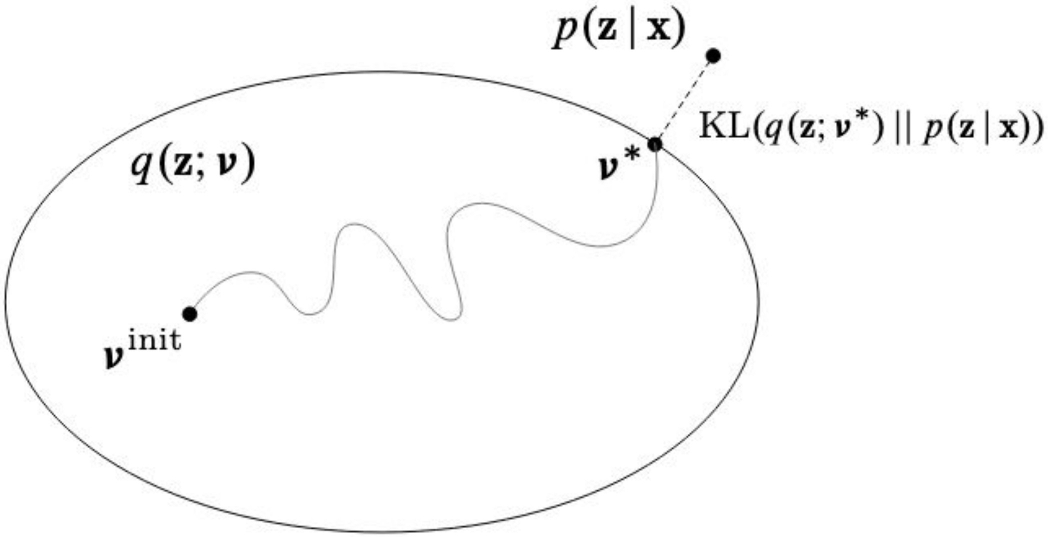
- VI turns inference into **optimization**.

- Posit a **variational family** of distributions over the latent variables,
  q(z; v)

- Fit the **variational parameters,** v, to be close (in KL) to the exact posterior.

- Provides an interpretable statistical model of data generation.
- Natural ways to incorporate prior information.
- Rigorous uncertainty estimates.
- Scalable to large datasets using stochastic training.
- Compatible with DNNs and AutoDiff.

Variational Inference

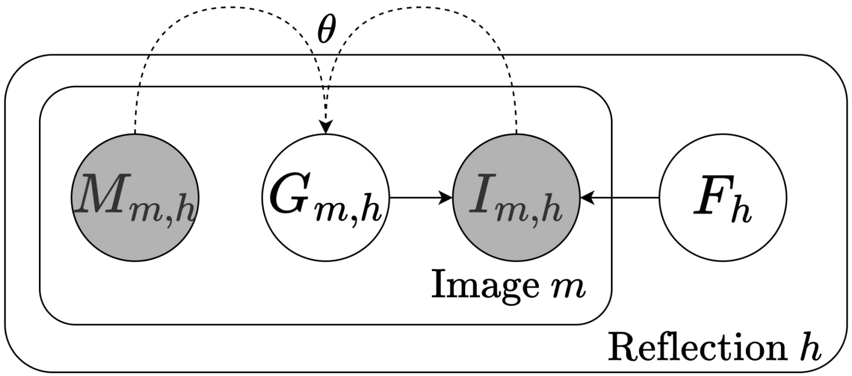


VI Algorithm

- VI turns inference into **optimization**.
- Posit a **variational family** of distributions over the latent variables,
  q(z; v)
- Fit the **variational parameters**, v, to be close (in KL) to the exact posterior.

*David Blei, Rajesh Ranganath, Shakir Mohamed. NeurIPS 2016 Tutorial.

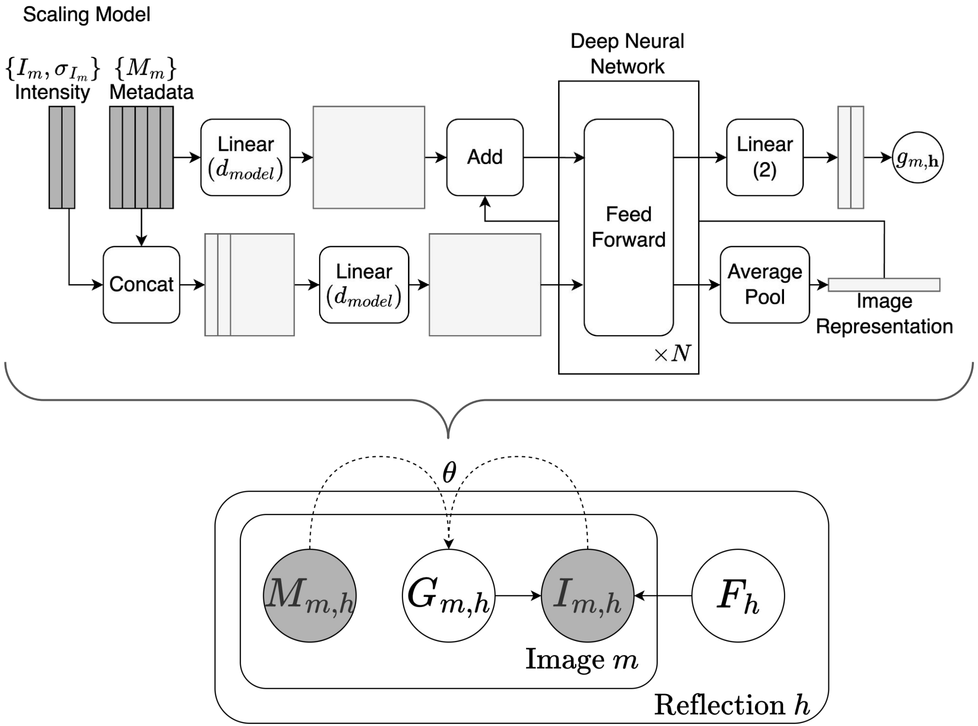
A Statistical Model of Diffraction



Algorithm

- Estimate random variables
  - **F**: Fourier coefficients of the electron density
  - **G**: Systematic error in measurements
- **G** is amortized by θ, the parameters of a neural network
  - Predict systematic errors from metadata
- Learn **F** and θ to maximize the evidence lower bound (ELBO)

VI Can Scale to Large Data Sets



Algorithm

- Amortize systematic error, **G**, using a simple CNN
  - Permutation invariant
  - Residual feed forward network