Flow-based networks and their benefits for us in high-energy physics

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

Based on arXiv:2008.05825

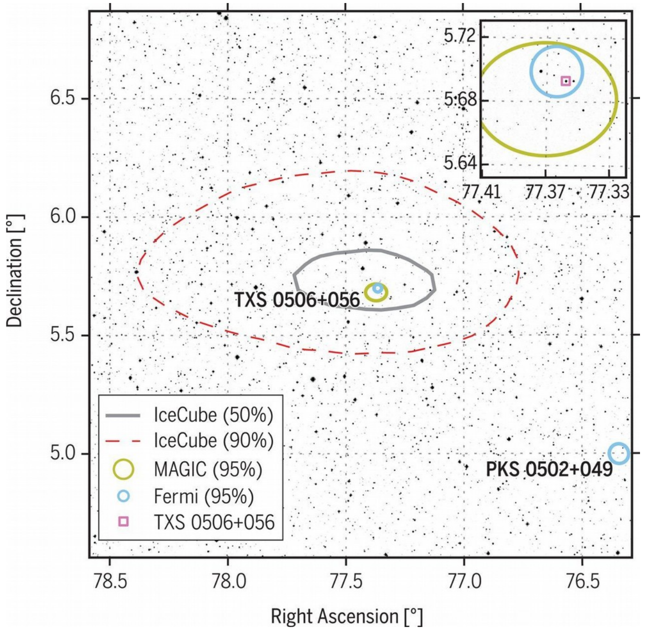
Thorsten Glüsenkamp, ErUM-Data meeting, Sep. 22nd, 2020

Overview

- Motivation
- Joint KL-divergence
- Flows generalize MSE
- Coverage
- Systematics
- Goodness of fit

Motivation (here astronomical Posteriors)

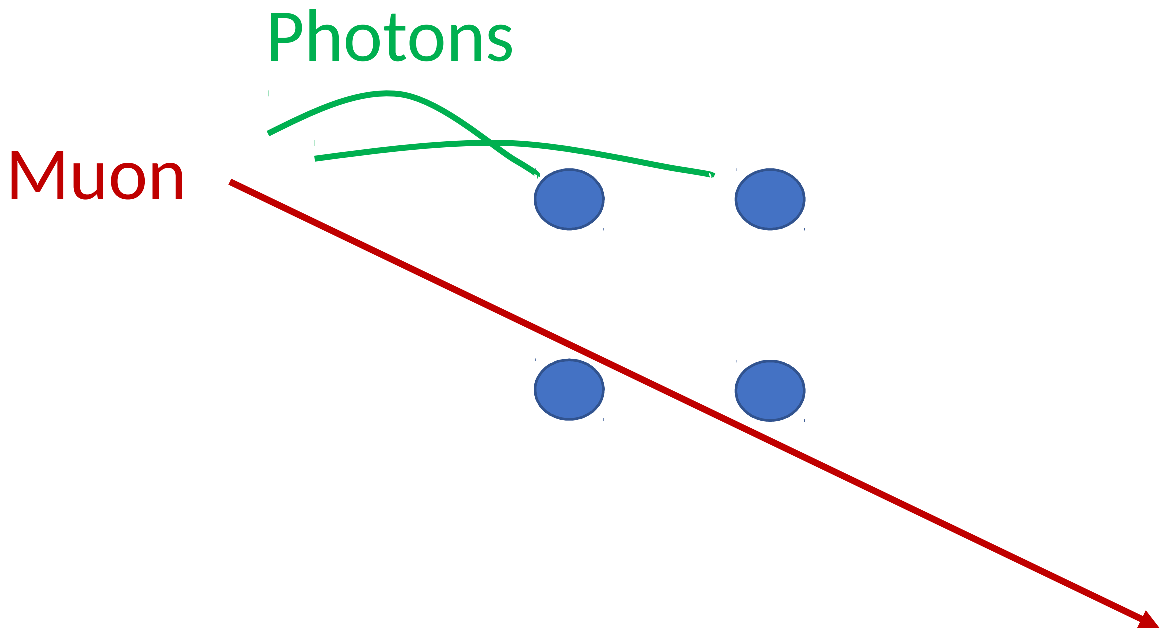
- Standard recos: Coverage + correct systematics big issue (since years)
  (what about goodness-of-fit !?)



- Can we maybe solve all issues with neural networks?
  Indeed, just using a simple upgrade of our existing networks, using so called „normalizing flows“.

What is a Monte Carlo simulation?

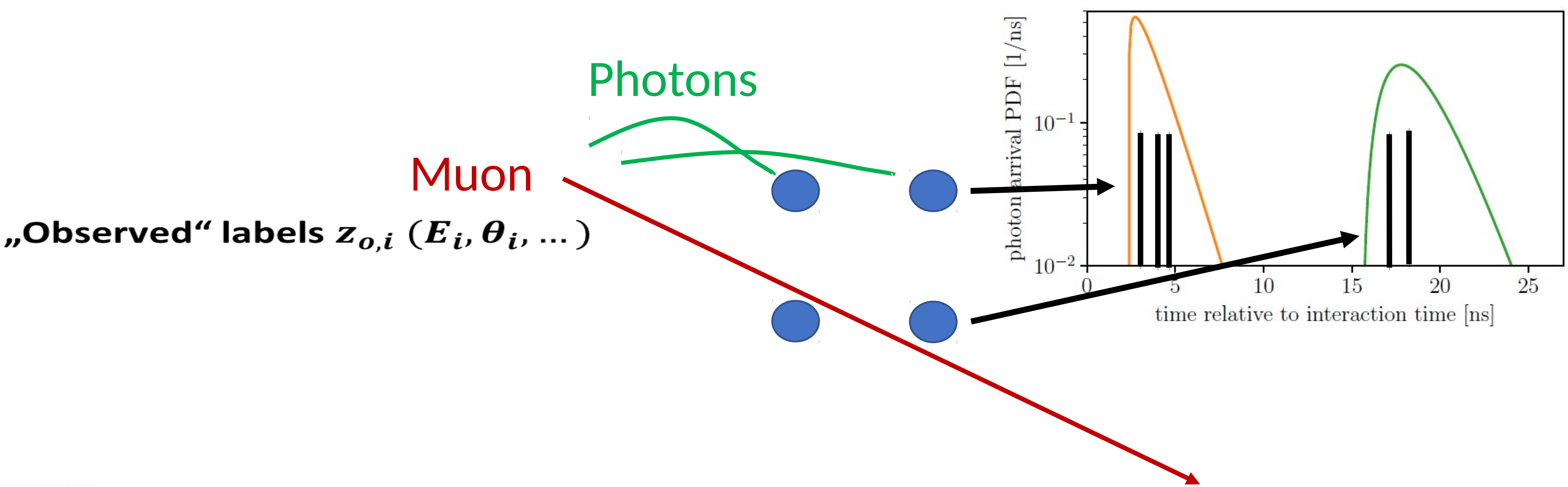
An answer: **Samples $x_i, z_{o,i}$ from a „true" (intractable) joint distribution** $\mathcal{P}_t(x, z_o)$



ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

What is a Monte Carlo simulation?

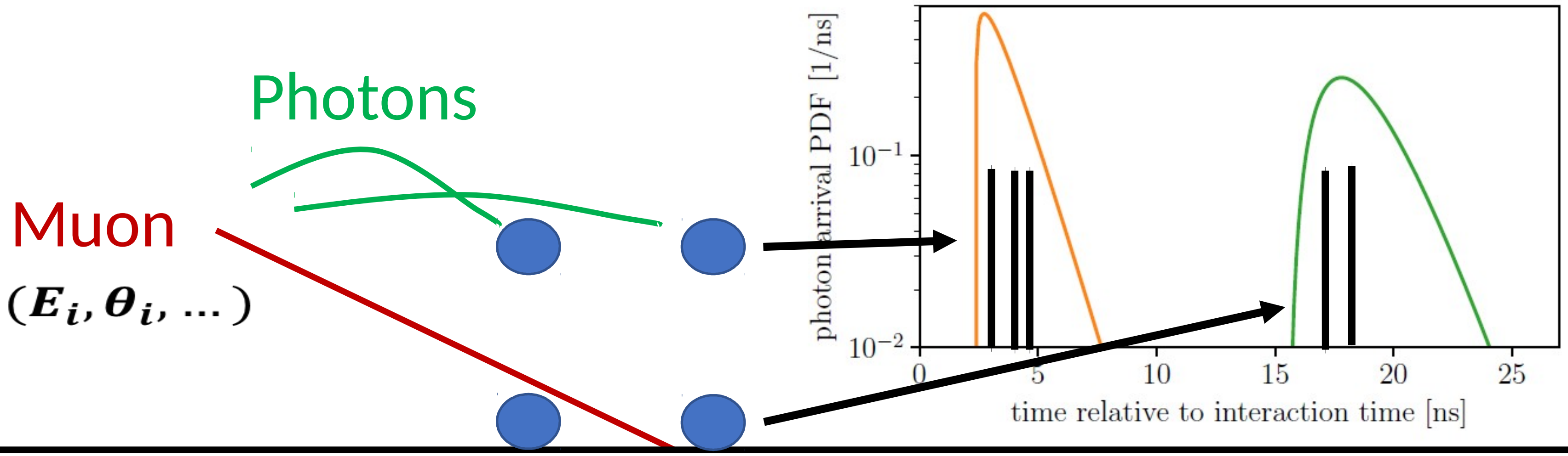
- An answer: **Samples $x_i, z_{o,i}$ from a „true“ (intractable) joint distribution** $\mathcal{P}_t(x, z_o)$



What is a Monte Carlo simulation?

- An answer: **Samples $x_i, z_{o,i}$ from a „true“ (intractable) joint distribution** $\mathcal{P}_t(x, z_o)$

data x_i



„Observed“ labels $z_{o,i}$ $(E_i, \theta_i, \ldots)$

→ **Can evaluate expectation of arbitrary function $f(x, z_o)$ via samples!**

$$E_{x,z_o}[f(x, z_o)] = \int_{x,z_o} \mathcal{P}_t(x, z_o) f(x, z_o)\, dx dz_o \approx \frac{1}{N} \sum_{x_i, z_{o,i}} f(x_i, z_{o,i})$$

Supervised learning loss

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z_o}) = \ln \frac{P_t(x, z_o)}{q(x, z_o)}$$

➡ $$E_{x,z_o}[f(x, z_o)] = D_{\mathrm{KL,joint(x,z_o)}}(\mathcal{P}_t; q) = \int_x \int_{z_o} \mathcal{P}_t(z_o, x) \cdot \ln \frac{\mathcal{P}_t(z_o, x)}{q(z_o, x)} dz_o dx$$

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

Supervised learning loss

- A particular choice of $\boldsymbol{f(x, z_o)}$ yields the loss fuction in supervised learning!

$$\boldsymbol{f(x, z_o)} = \ln \frac{P_t(x, z_o)}{q(x, z_o)}$$

$$\Rightarrow \quad E_{x,z_o}[f(x, z_o)] = D_{\mathrm{KL,joint(x,z_o)}}(\mathcal{P}_t; q) = \int_x \int_{z_o} \mathcal{P}_t(z_o, x) \cdot \ln \frac{\mathcal{P}_t(z_o, x)}{q(z_o, x)} dz_o dx$$

Use samples, parametrize $\boldsymbol{q(z_o; x)}$ with neural network as $\boldsymbol{q_\phi}$, minimize result over $\boldsymbol{\phi}$

$$\Rightarrow \quad \arg\min_\phi \hat{D}_{\mathrm{KL,joint(x,z_o)}}(\mathcal{P}_t; q_\phi) = \ldots = \arg\min_\phi \frac{1}{N} \sum_{S \in x_i, z_{o,i}} -\ln\left(q_\phi(z_{o,i}; x_i)\right)$$

Supervised learning loss

$$\boldsymbol{f(x, z_o)} = \ln \frac{P_t(x, z_o)}{q(x, z_o)}$$

$$\Rightarrow \quad E_{x,z_o}[f(x, z_o)] = D_{\mathrm{KL,joint(x,z_o)}}(\mathcal{P}_t; q) = \int_x \int_{z_o} \mathcal{P}_t(z_o, x) \cdot \ln \frac{\mathcal{P}_t(z_o, x)}{q(z_o, x)} dz_o dx$$

Use samples, parametrize $q(z_o; x)$ with neural network as q_ϕ, minimize result over ϕ

$$\Rightarrow \quad \arg\min_\phi \hat{D}_{\mathrm{KL,joint(x,z_o)}}(\mathcal{P}_t; q_\phi) = \ldots = \arg\min_\phi \frac{1}{N} \sum_{S \in x_i, z_{o,i}} -\ln\left(q_\phi(z_{o,i}; x_i)\right)$$

MSE-Loss: $\sum \left(\boldsymbol{\mu_\phi(x_i) - z_{o,i}}\right)^2$ ← $q_\phi = N(\mu; 1)$ (standard Normal)

Meaning of the KL-divergence viewpoint

$$\arg\min_{\phi} \hat{D}_{\mathrm{KL,joint(x,z_o)}}(\mathcal{P}_t; q_\phi) = \arg\min_{\phi} \frac{1}{N} \sum_{S \in x_i, z_{o,i}} \ln\left(\frac{\mathcal{P}_t(z_{o,i}; x_i)}{q_\phi(z_{o,i}; x_i)}\right) + \ln\left(\frac{\mathcal{P}_t(x_i)}{q(x_i)}\right)$$

$$\ldots = \arg\min_{\phi} \frac{1}{N} \sum_{S \in x_i, z_{o,i}} -\ln\left(q_\phi(z_{o,i}; x_i)\right)$$

Meaning of the KL-divergence viewpoint

$$\arg\min_{\phi} \hat{D}_{\mathrm{KL,joint(x,z_o)}}(\mathcal{P}_t; q_\phi) = \arg\min_{\phi} \frac{1}{N} \sum_{S \in x_i, z_{o,i}} \ln\left(\frac{\mathcal{P}_t(z_{o,i}; x_i)}{q_\phi(z_{o,i}; x_i)}\right) + \ln\left(\frac{\mathcal{P}_t(x_i)}{q(x_i)}\right)$$

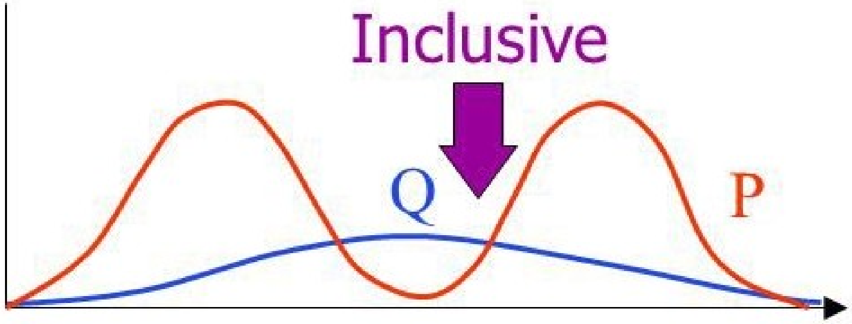
$$\ldots = \arg\min_{\phi} \frac{1}{N} \sum_{S \in x_i, z_{o,i}} -\ln\left(q_\phi(z_{o,i}; x_i)\right)$$

„Minimizing KL-divergence between **True Posterior P_t** and **approximate Posterior q_ϕ over ϕ** = minimizing supervised learning loss“

Minimising

KL$(P||Q)$

$$= \sum_H P(H|V) \ln \frac{P(H|V)}{Q(H)}$$



Meaning of the KL-divergence viewpoint

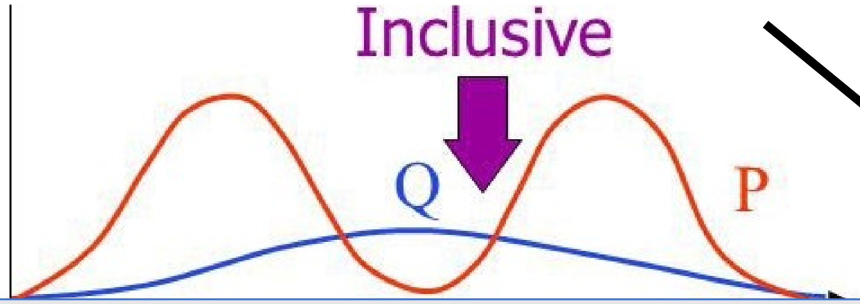
$$\arg\min_{\phi} \hat{D}_{\mathrm{KL,joint(x,z_o)}}(\mathcal{P}_t; q_\phi) = \arg\min_{\phi} \frac{1}{N} \sum_{S \in x_i, z_{o,i}} \ln\left(\frac{\mathcal{P}_t(z_{o,i}; x_i)}{q_\phi(z_{o,i}; x_i)}\right) + \ln\left(\frac{\mathcal{P}_t(x_i)}{q(x_i)}\right)$$

$$\ldots = \arg\min_{\phi} \frac{1}{N} \sum_{S \in x_i, z_{o,i}} -\ln\left(q_\phi(z_{o,i}; x_i)\right)$$

„Minimizing KL-divergence between **True Posterior P_t** and **approximate Posterior q_ϕ over ϕ** = minimizing supervised learning loss“

Minimising KL($P||Q$)

$$= \sum_H P(H|V) \ln \frac{P(H|V)}{Q(H)}$$



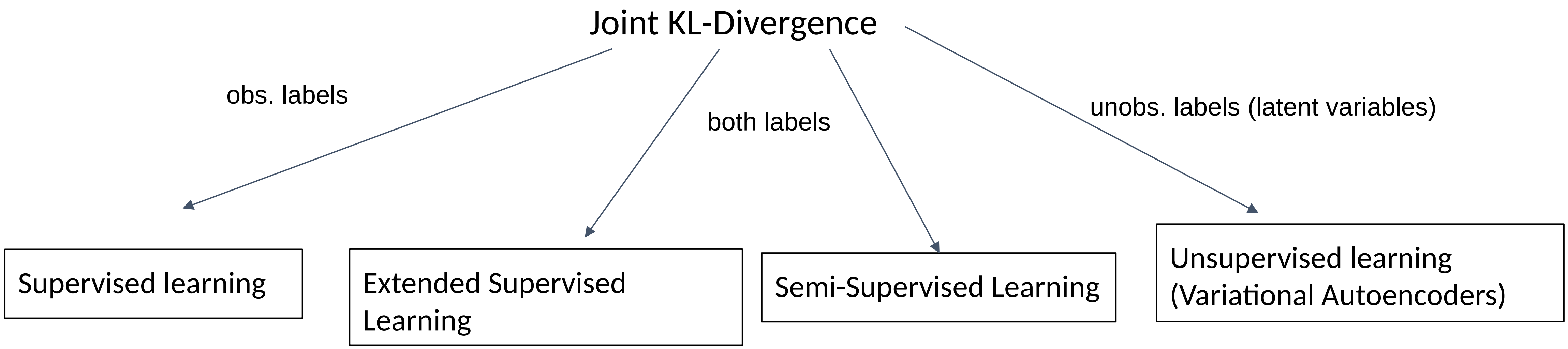
! This is a VERY useful viewpoint for us in physics !

Standard supervised learning is performing approximate Likelihood-free inference

„Neural networks learn to approximate the true posterior“

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

This viewpoint unifies various approaches!



This viewpoint unifies various approaches!

Joint KL-Divergence

obs. labels

both labels

unobs. labels (latent variables)

Supervised learning

Extended Supervised Learning

Semi-Supervised Learning

Unsupervised learning (Variational Autoencoders)

classification

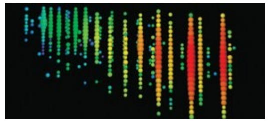
regression

Systematics

Goodness-of-Fit

Normalizing Flows

Coverage of posterior regions

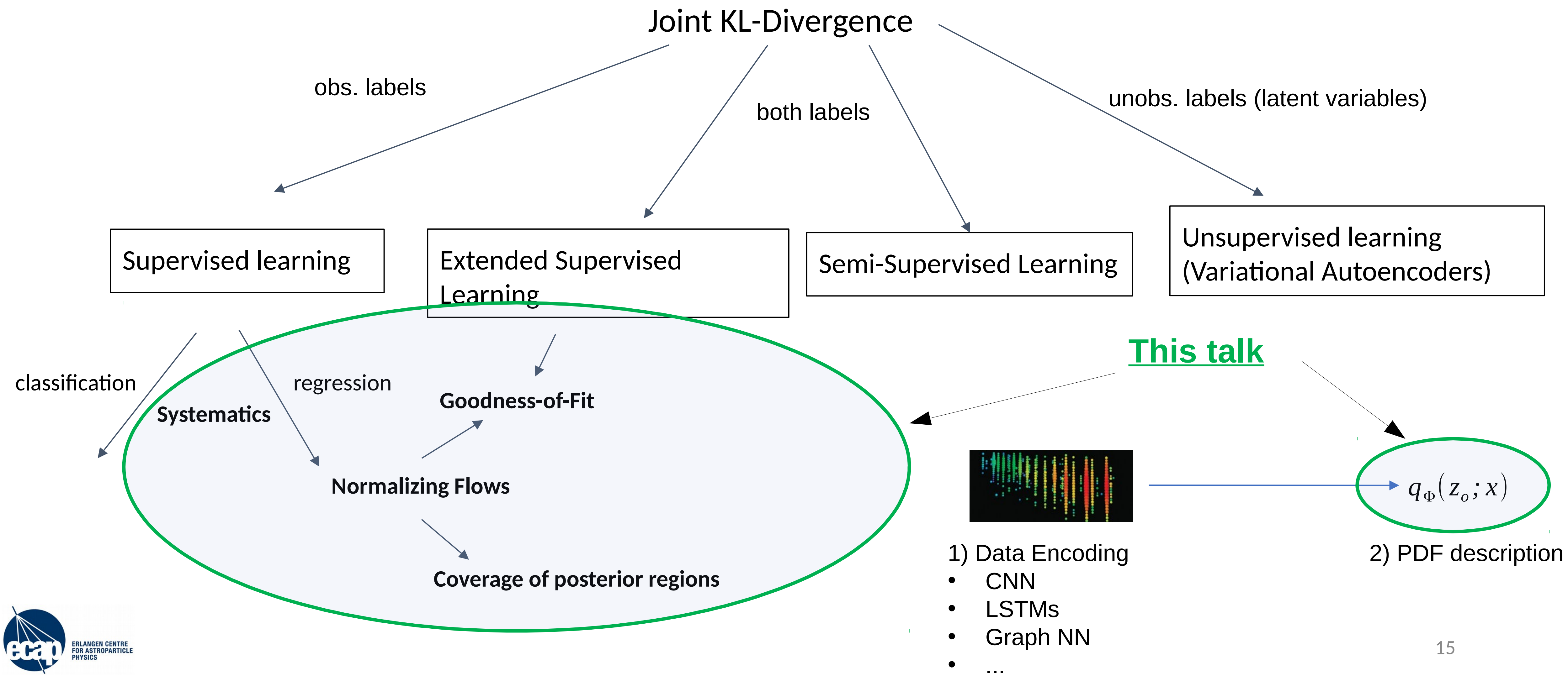


$q_{\Phi}(z_o; x)$

1) Data Encoding
- CNN
- LSTMs
- Graph NN
- ...

2) PDF description

This viewpoint unifies various approaches!



What are good PDF Aproximators q_ϕ ?

Predicting parameters of any complex distribution? E.g. a sum of gaussians? Possible, but there is something better …

Normalizing flows: (1912.02762 for a recent review)

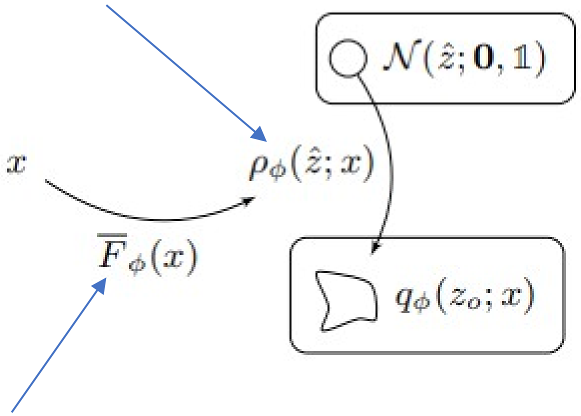
What are good PDF Aproximators q_ϕ ?

Predicting parameters of any complex distribution? E.g. a sum of gaussians? Possible, but there is something better …

Normalizing flows: (1912.02762 for a recent review)

<u>Generic Flow:</u>

Invertible mapping (inverse required for density evaluation)



Base distribution
(typically standard normal)

Final distribution

Parameters of mapping
are output of Neural Network (here conditional on x)

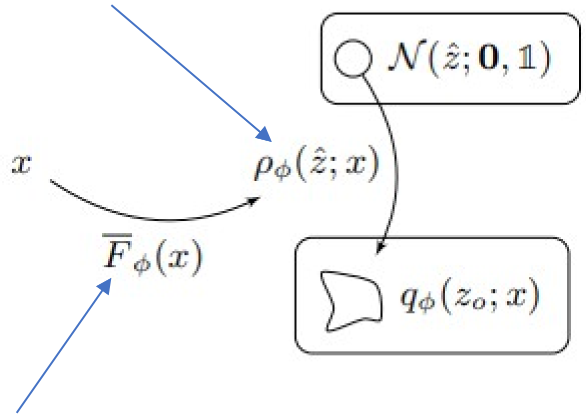
What are good PDF Aproximators q_ϕ ?

Predicting parameters of any complex distribution? E.g. a sum of gaussians? Possible, but there is something better …

Normalizing flows: (1912.02762 for a recent review)

Generic Flow:

Invertible mapping (inverse required for density evaluation)



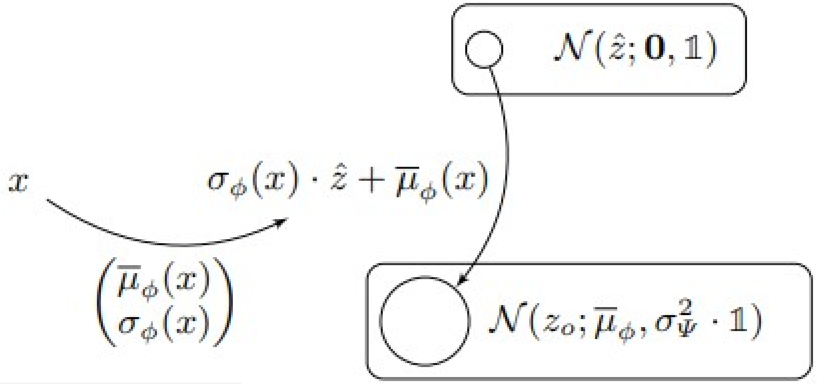
Base distribution (typically standard normal)

Final distribution

Parameters of mapping are output of Neural Network (here conditional on x)

A particular Flow:

A Gaussian PDF is an affine normalizing flow:

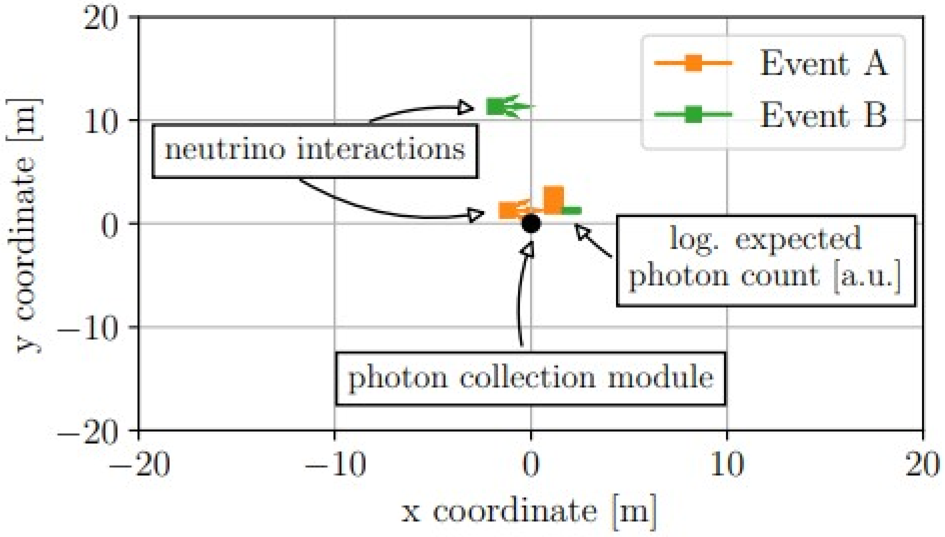


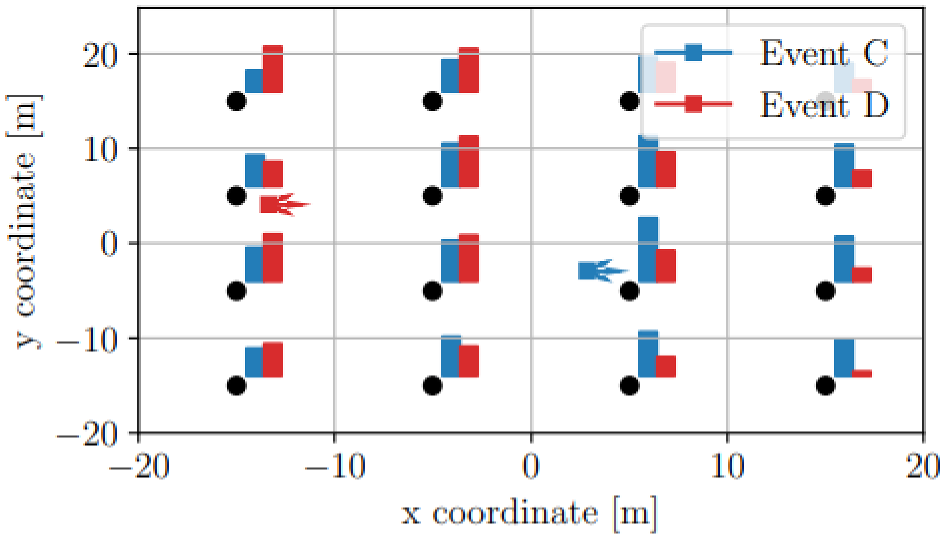
Final Gaussian distribution

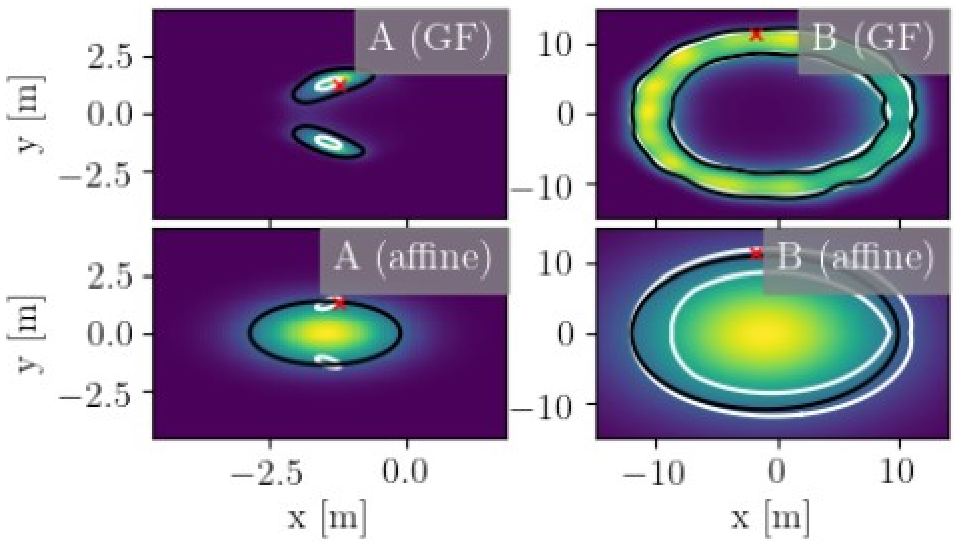
Generalizes MSE loss …
MSE loss corresponds to an affine flow with no scaling!

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

Example Posteriors

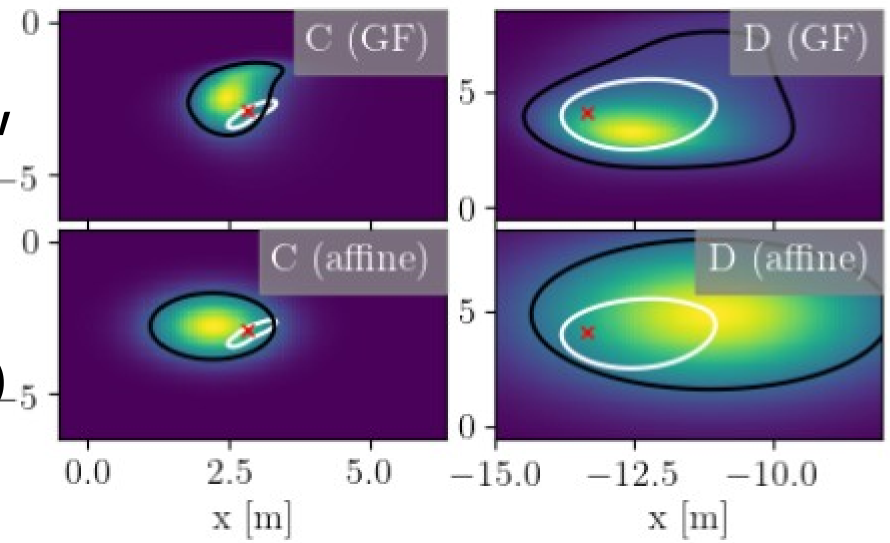




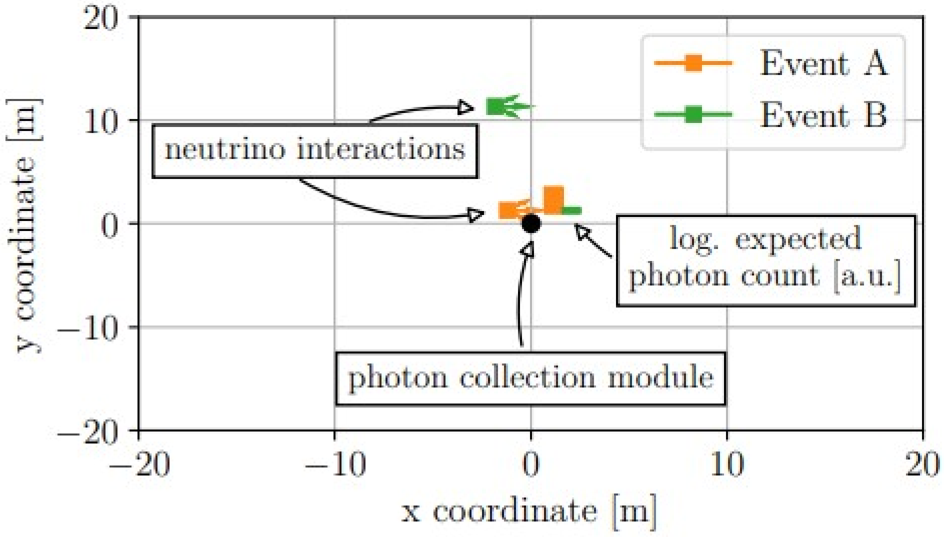


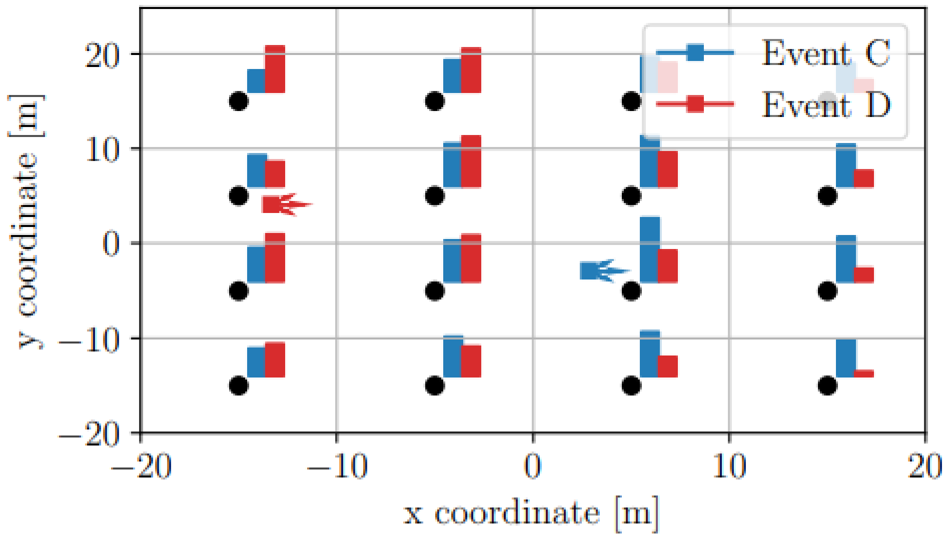
Complex flow

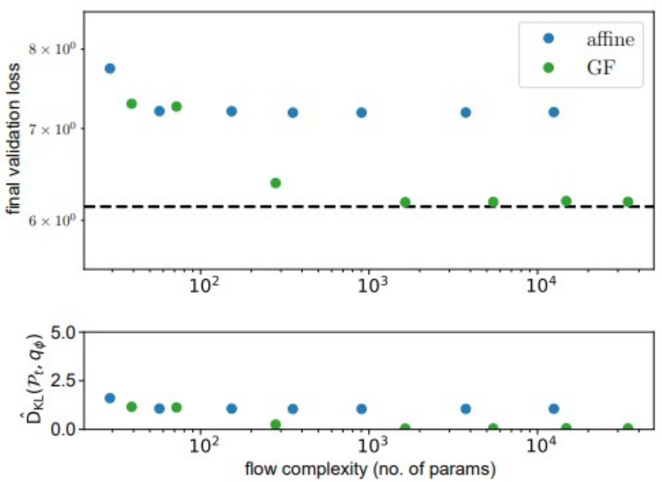
Affine flow (MSE + sigma)

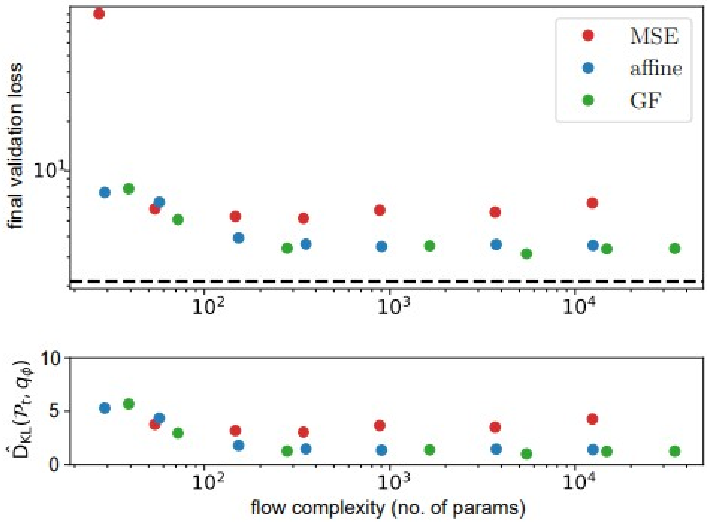


Example Posteriors



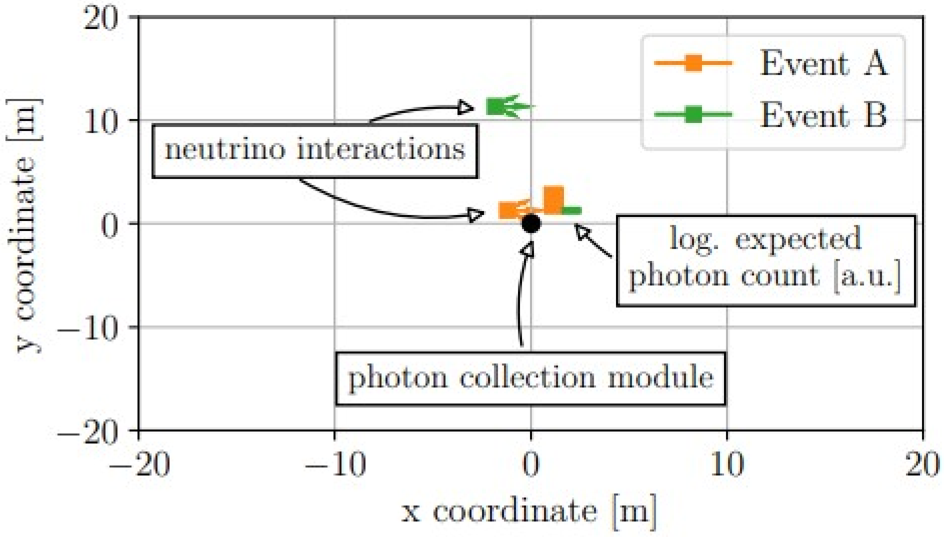


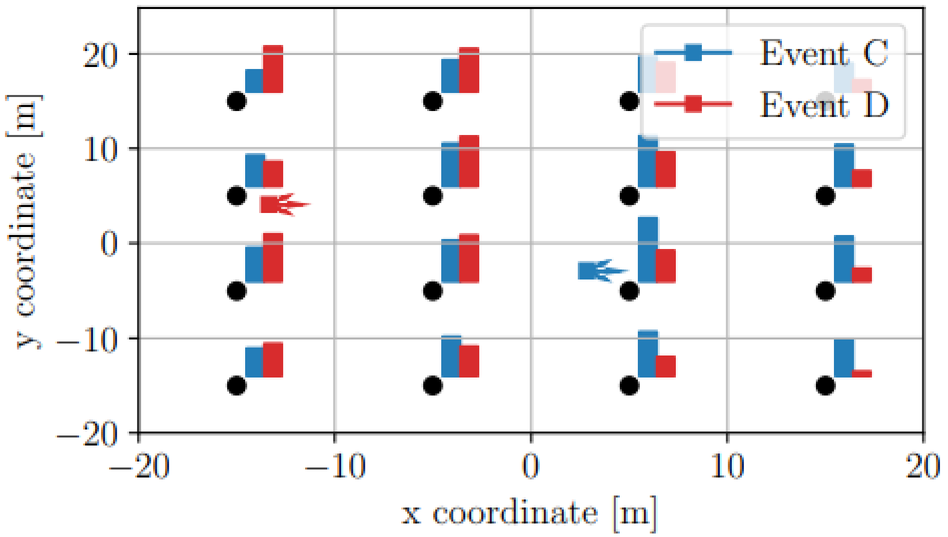


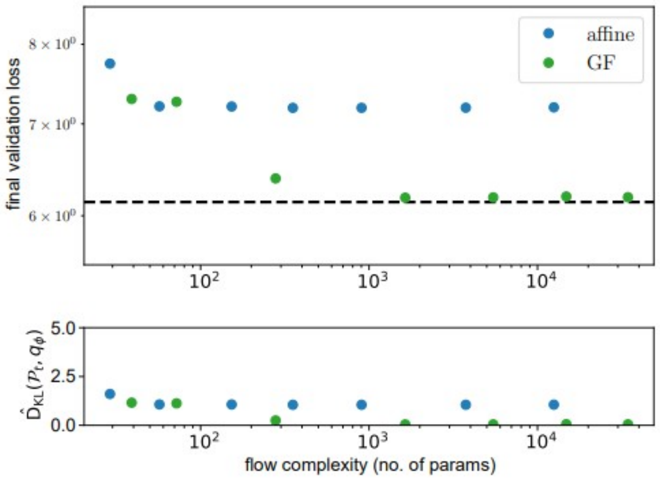


ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

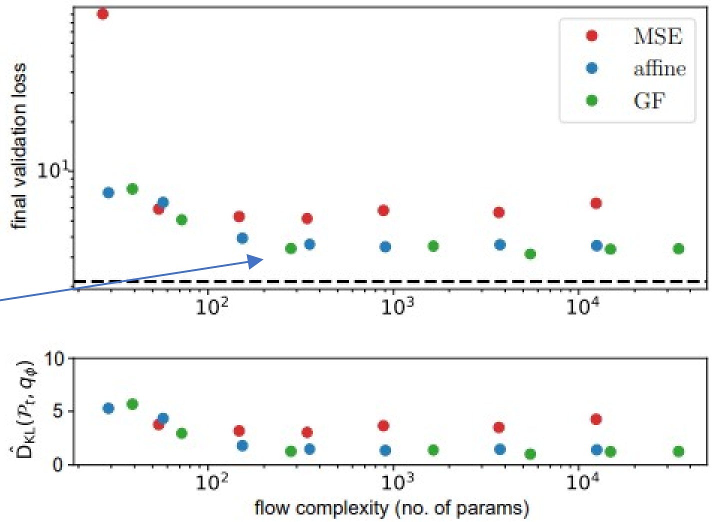
Example Posteriors







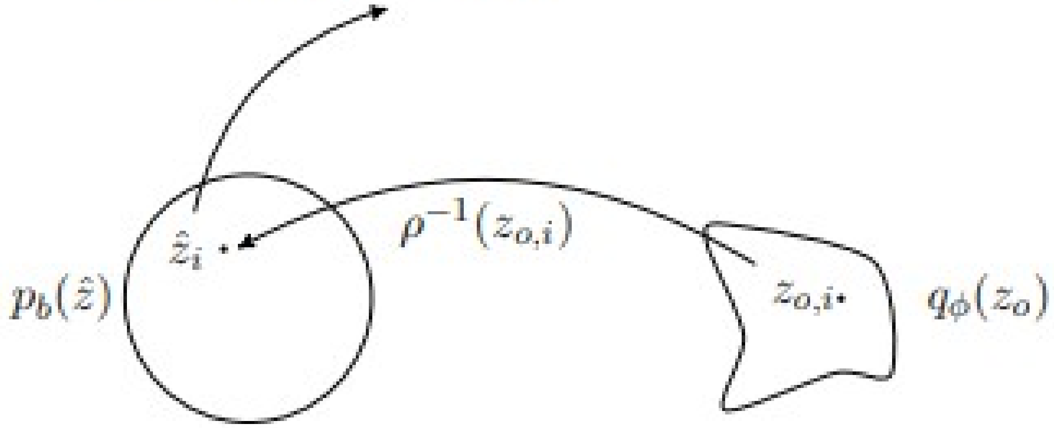
Data encoding is bottleneck, not the flow!

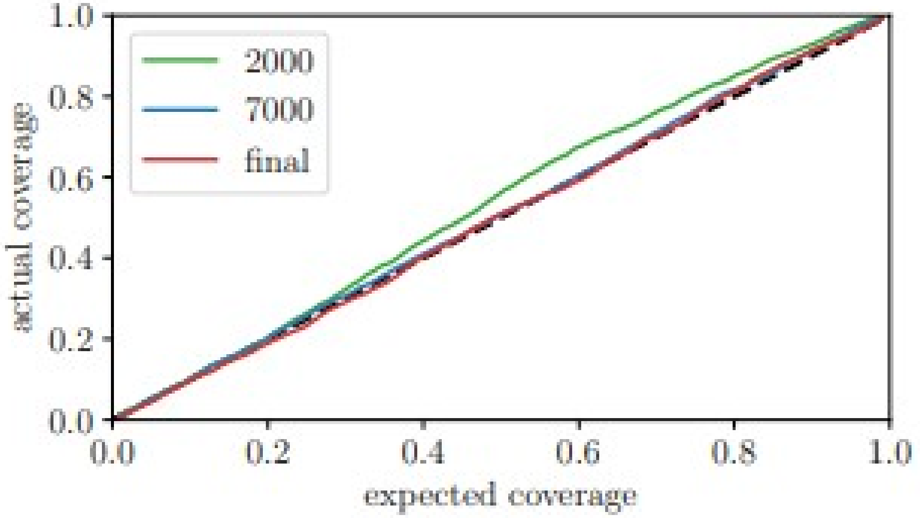


Coverage

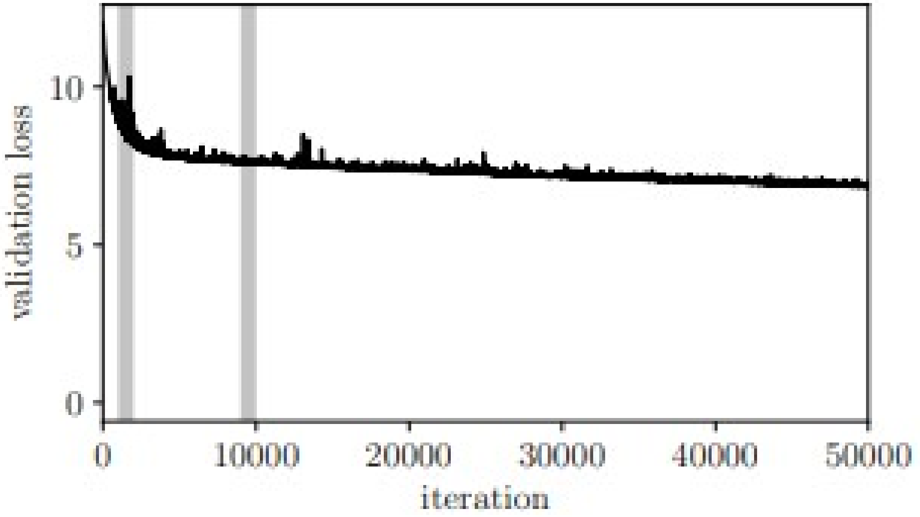
- Can calculate coverage of arbitrary PDF at the base using standard χ^2- test

$$-2 \cdot (\ln p_b(\hat{z}_i) - \ln p_b(0)) \sim \chi^2$$





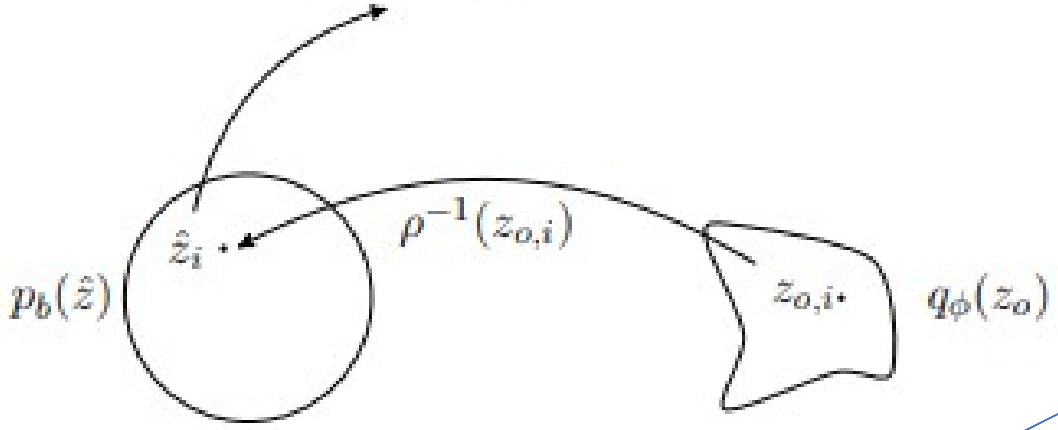
(a) Coverage of 3-d posteriors using dataset 3 for different stages of training.

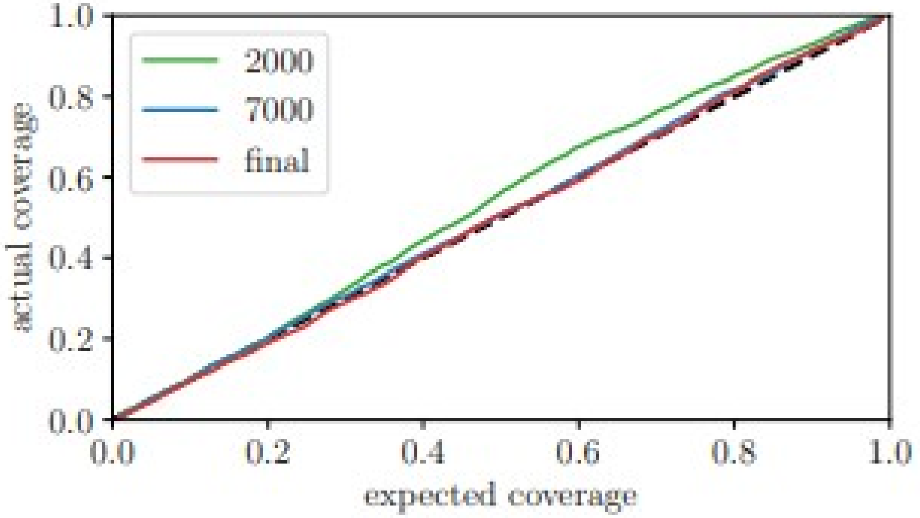


Coverage

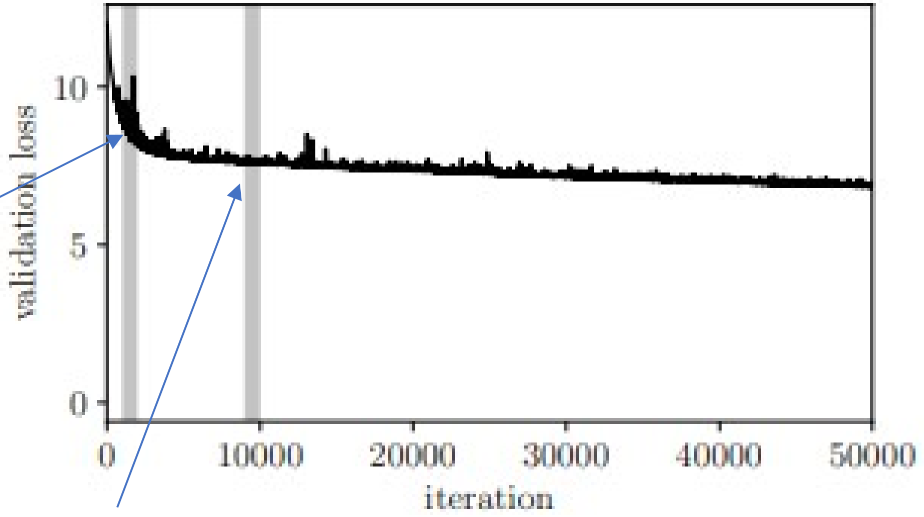
- Can calculate coverage of arbitrary PDF at the base using standard χ^2- test

$$-2\cdot(\ln p_b(\hat{z}_i)-\ln p_b(0))\sim\chi^2$$





(a) Coverage of 3-d posteriors using dataset 3 for different stages of training.

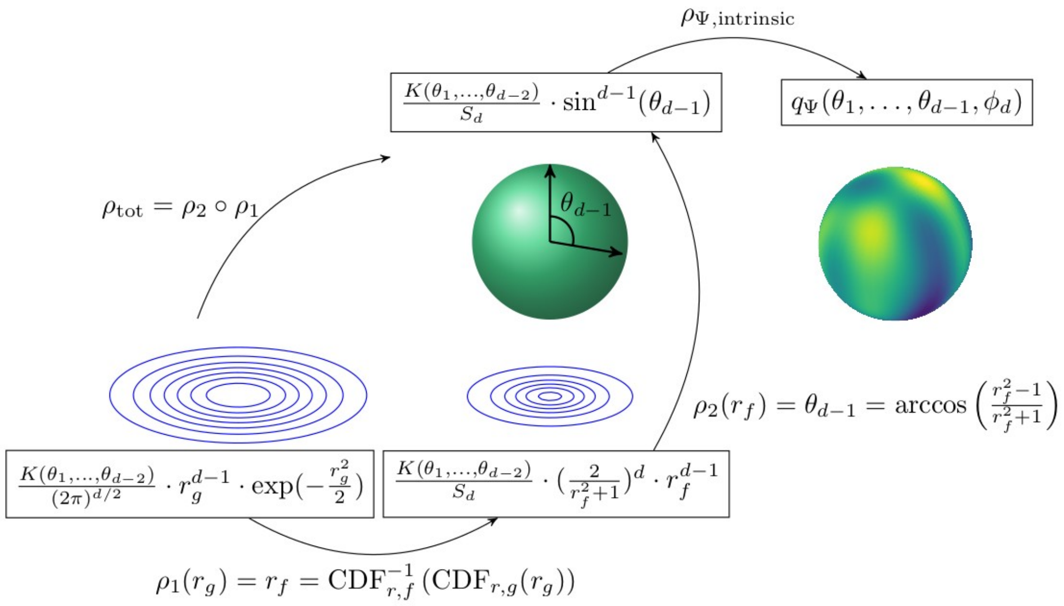


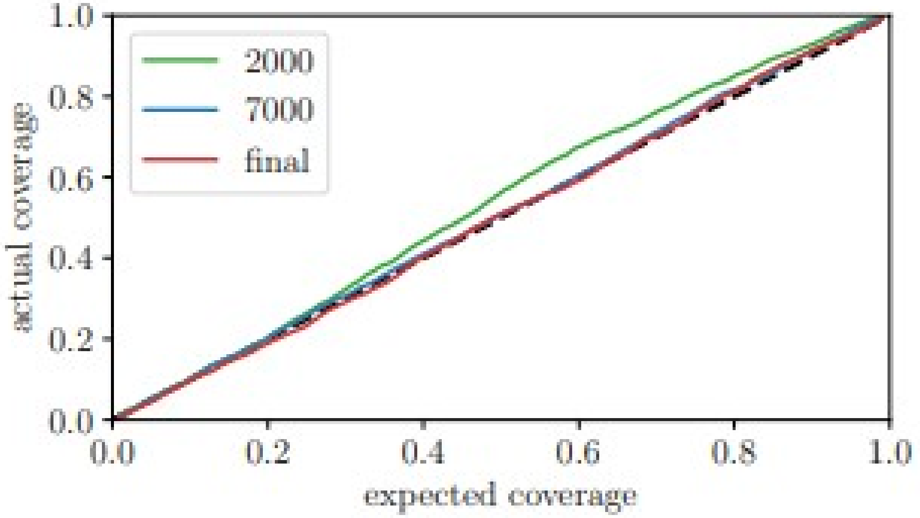
Rapid initial training phase to obtain coverage

Slow training phase (diffusive phase) shrinks posterior regions while maintaining coverage!

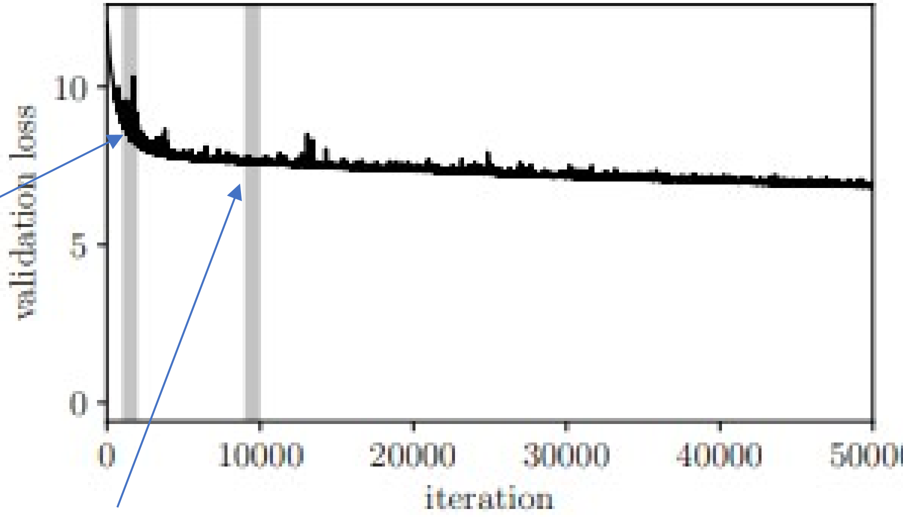
Coverage

Also works for arb. posteriors of directions (on spheres)





(a) Coverage of 3-d posteriors using dataset 3 for different stages of training.



Rapid initial training phase to obtain coverage

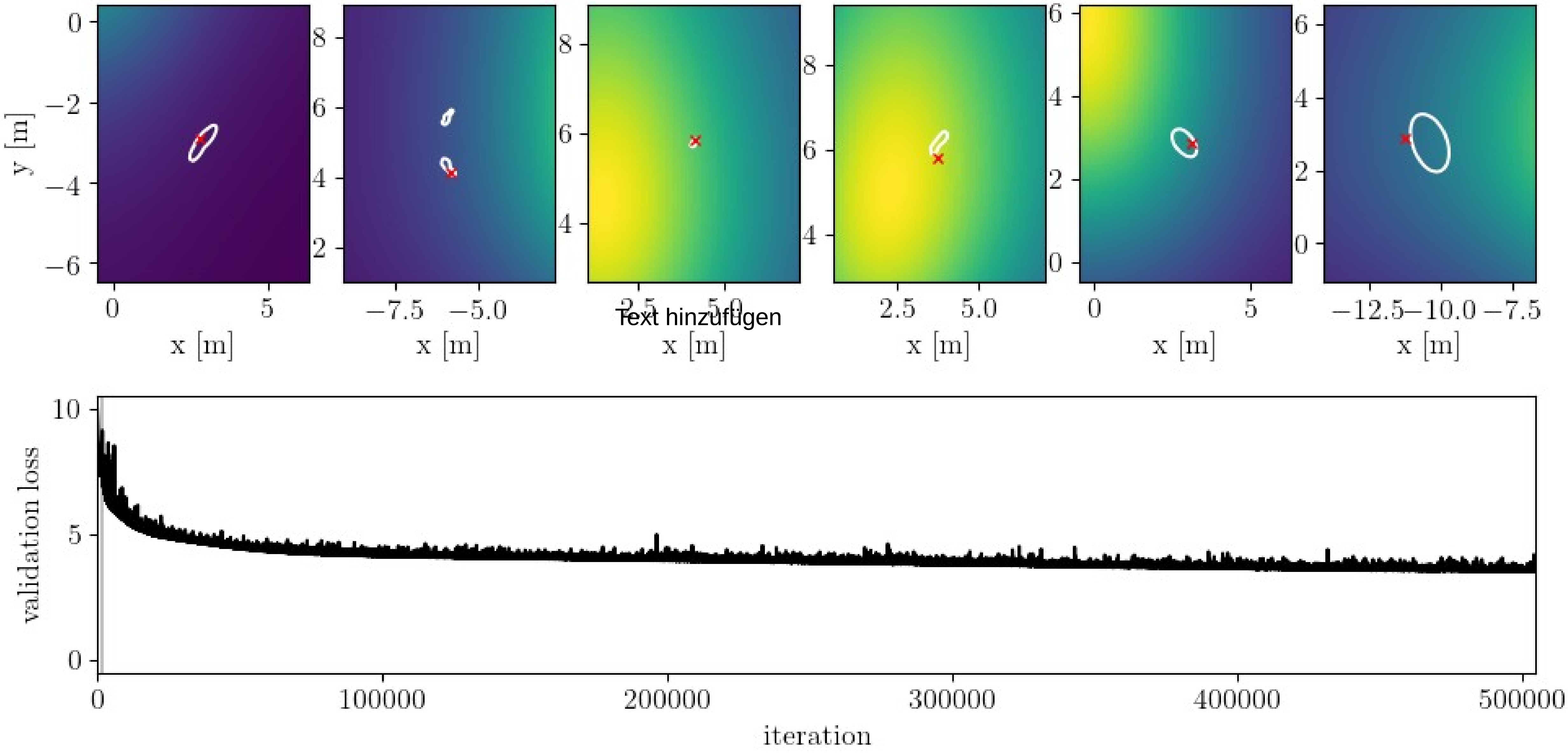
Slow training phase (diffusive phase) shrinks posterior regions while maintaining coverage!

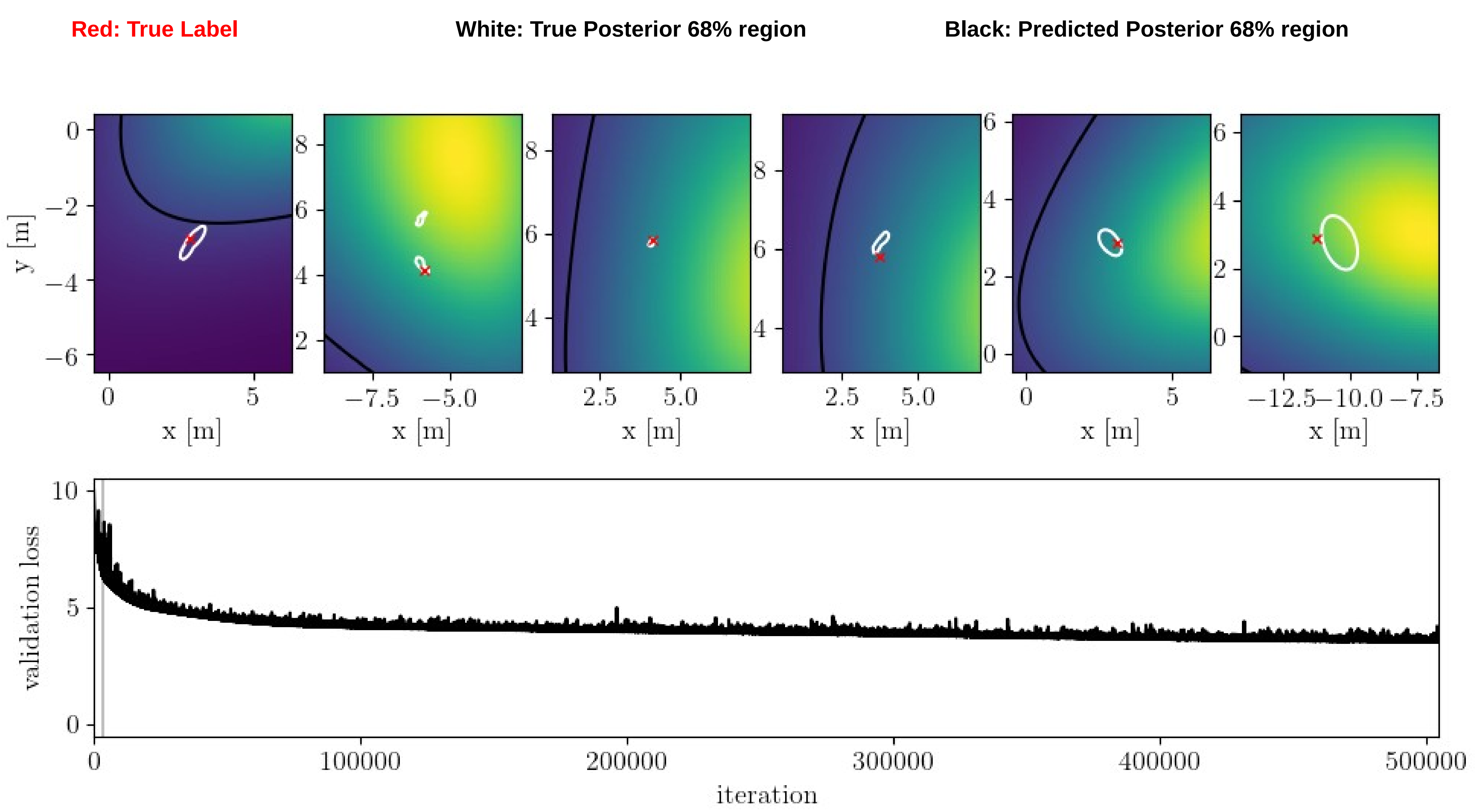
ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

Red: True Label

White: True Posterior 68% region

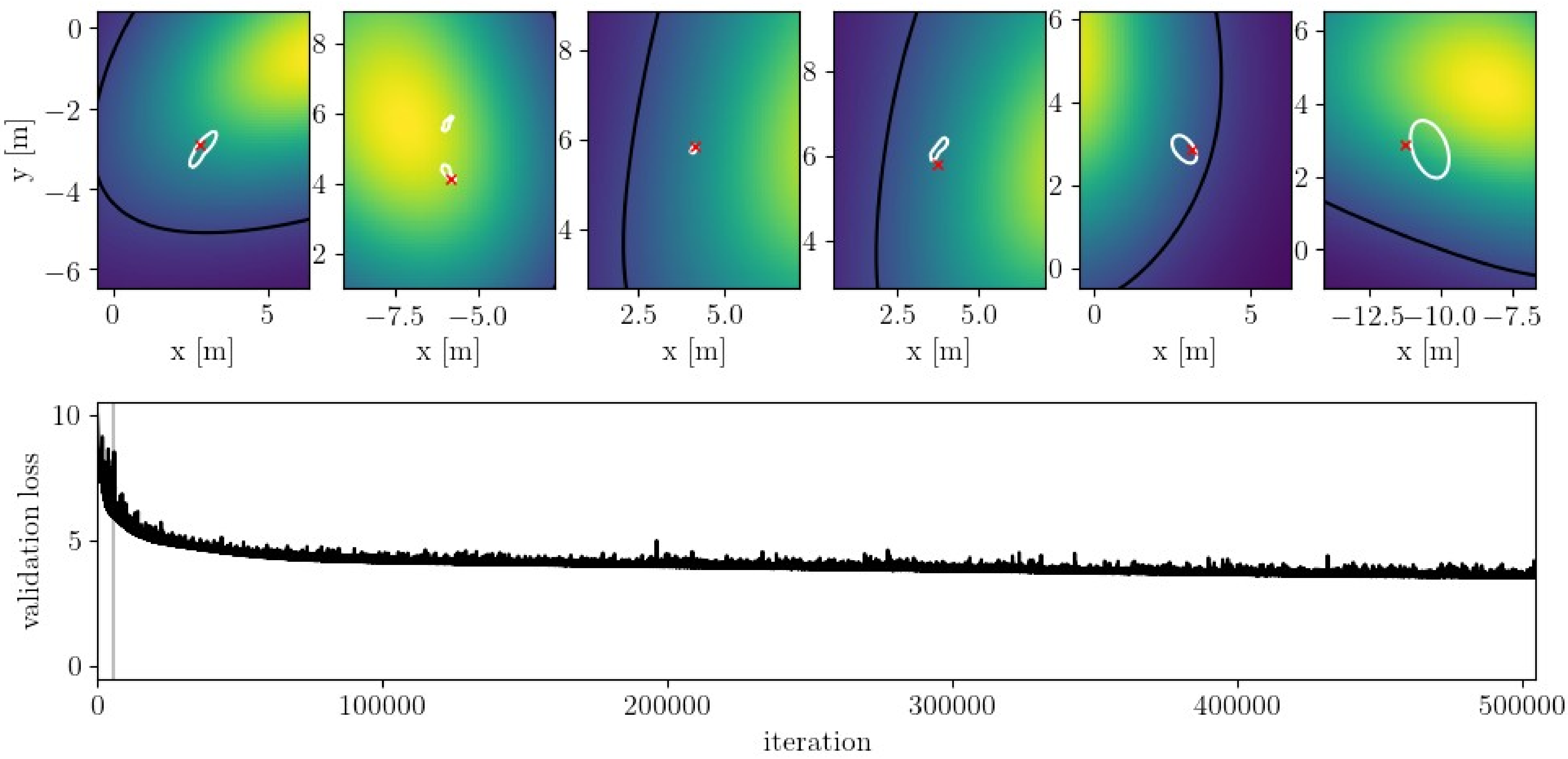
Black: Predicted Posterior 68% region





Red: True Label
White: True Posterior 68% region
Black: Predicted Posterior 68% region
y [m]
x [m]
validation loss
iteration

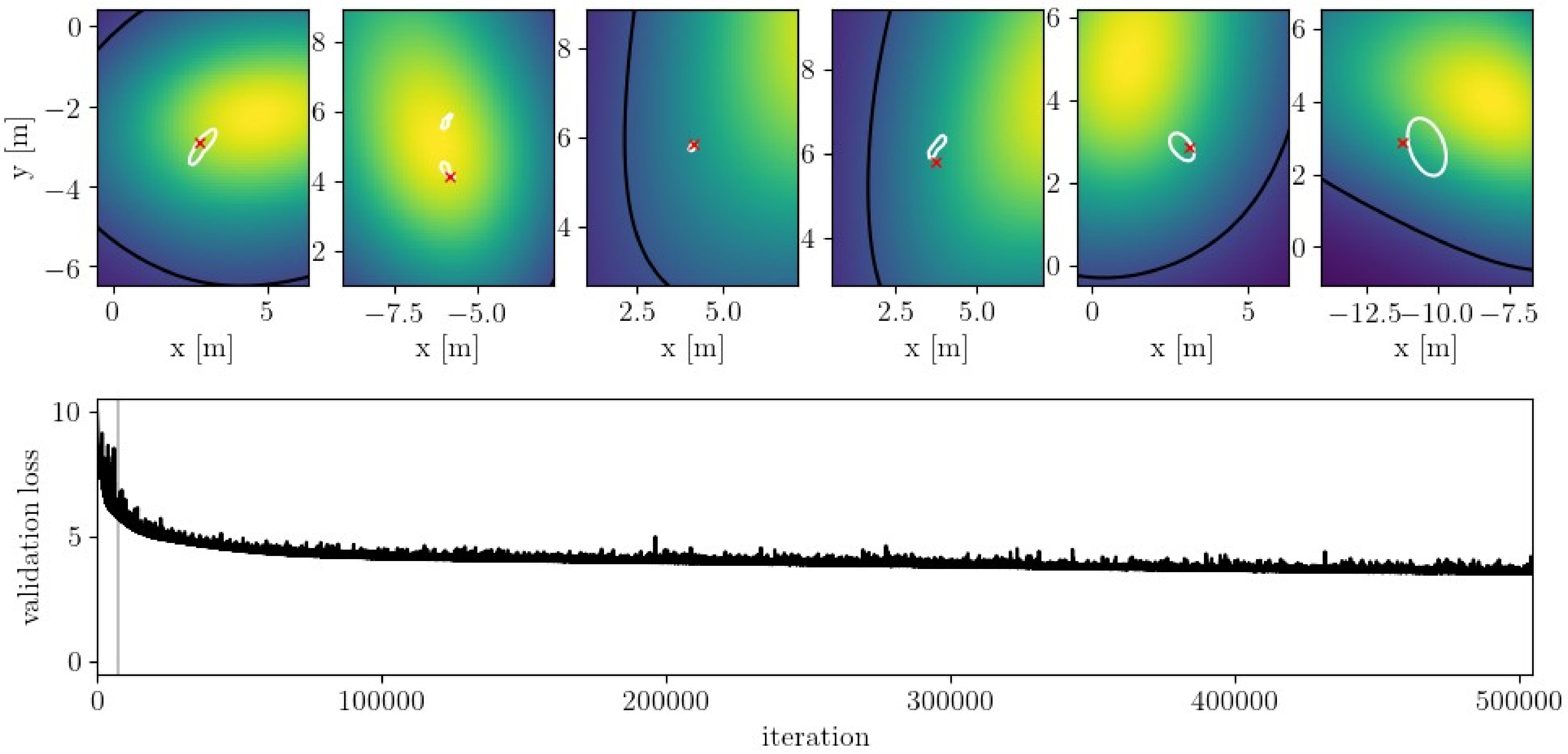
Red: True Label   White: True Posterior 68% region   Black: Predicted Posterior 68% region



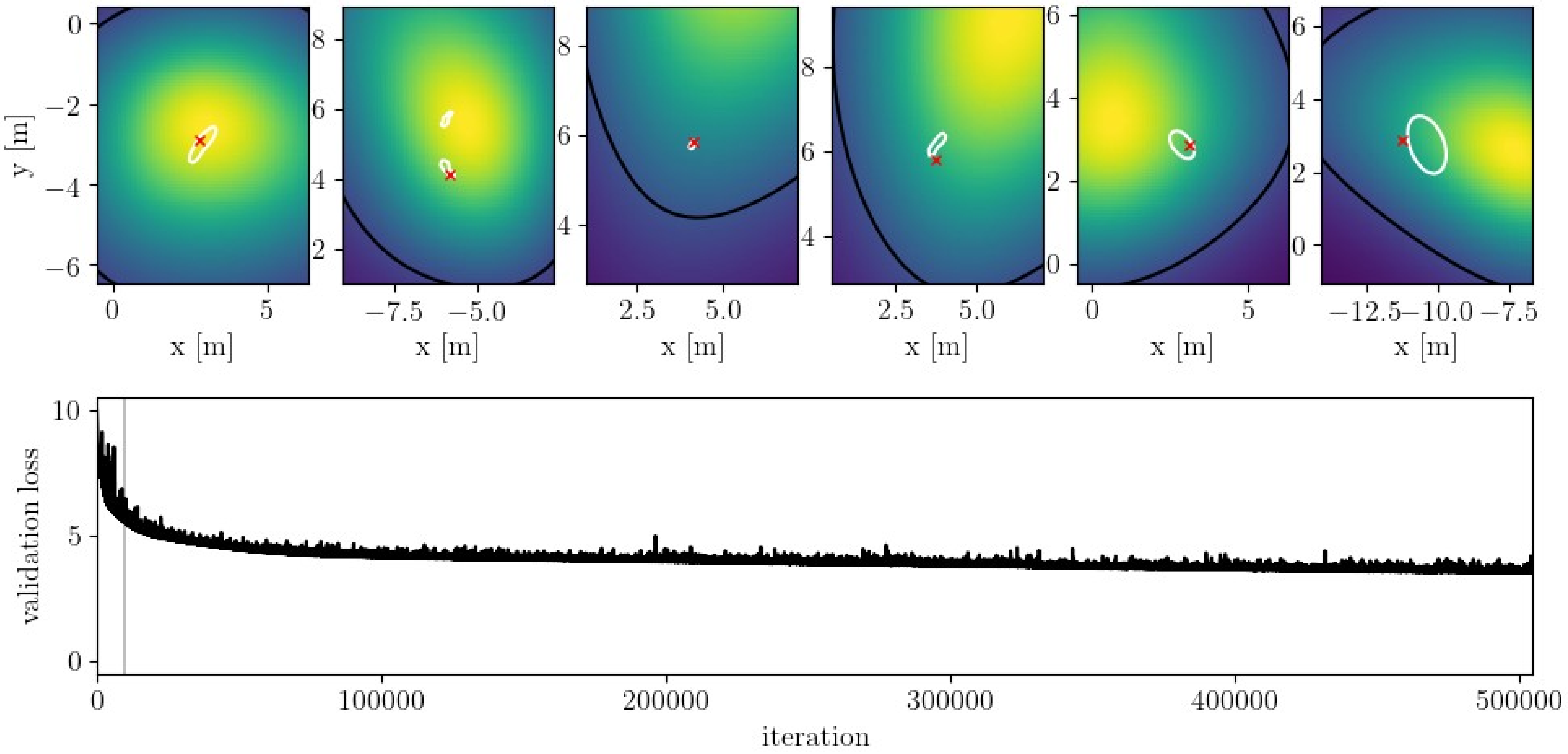
Red: True Label

White: True Posterior 68% region

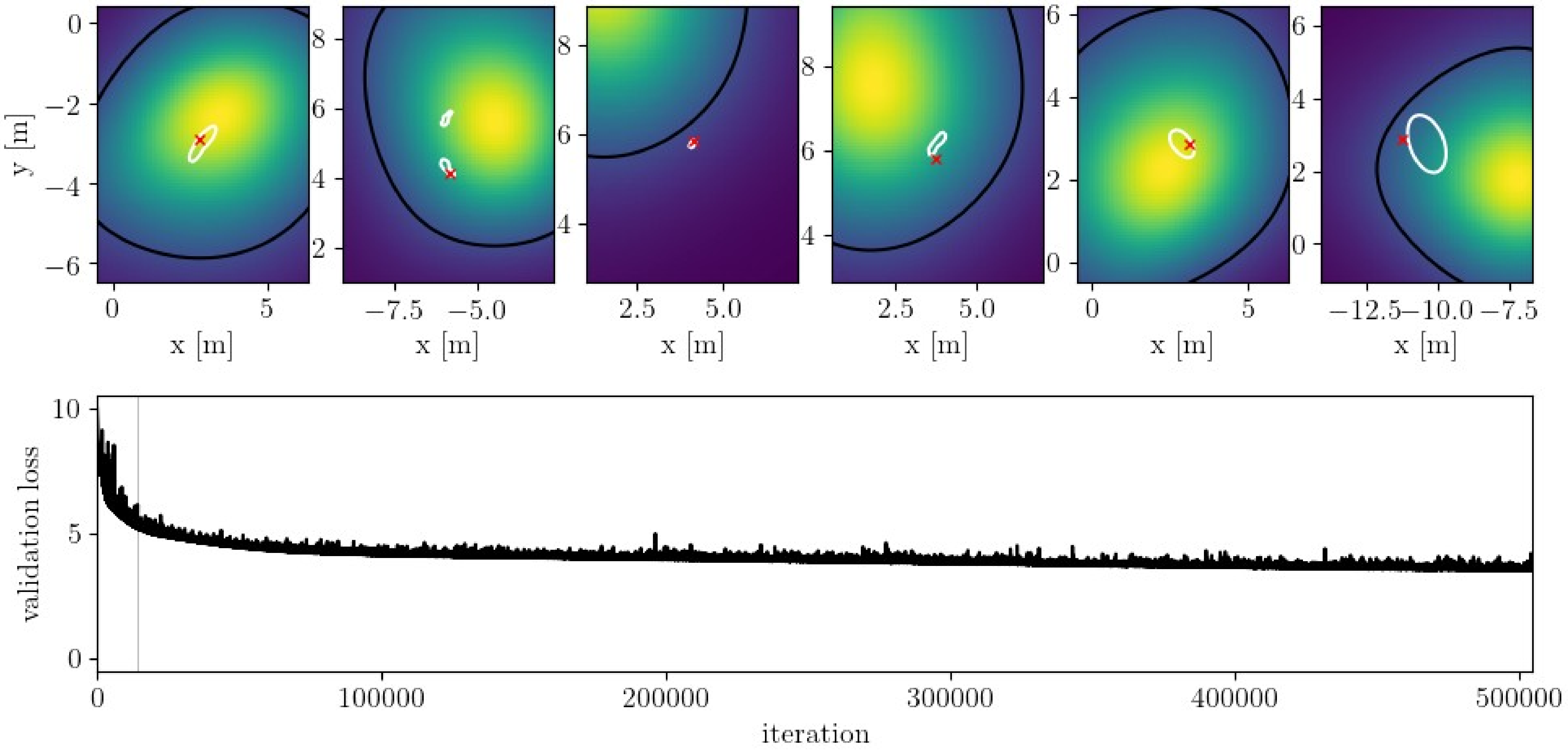
Black: Predicted Posterior 68% region

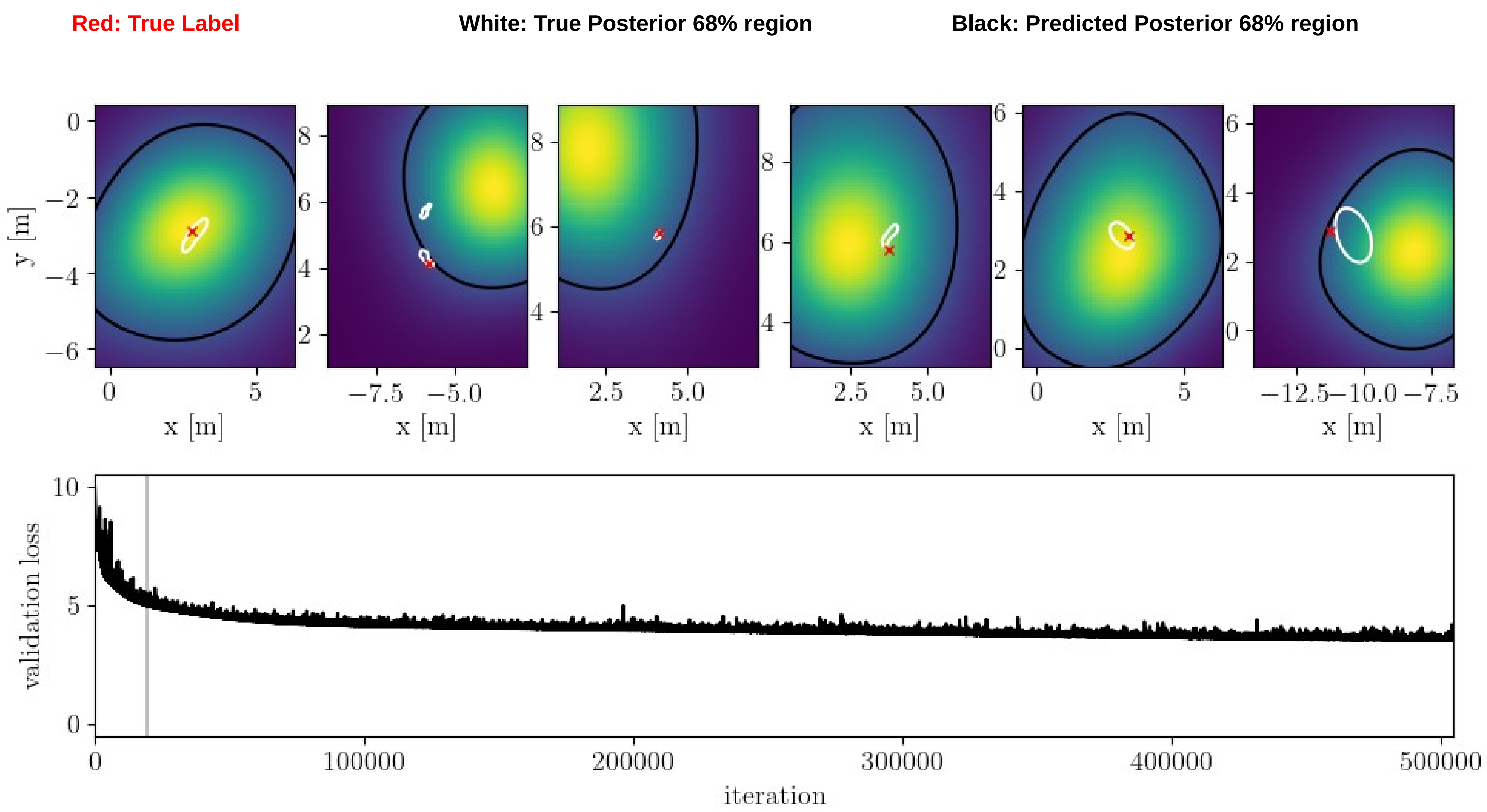


Red: True Label **White: True Posterior 68% region** **Black: Predicted Posterior 68% region**

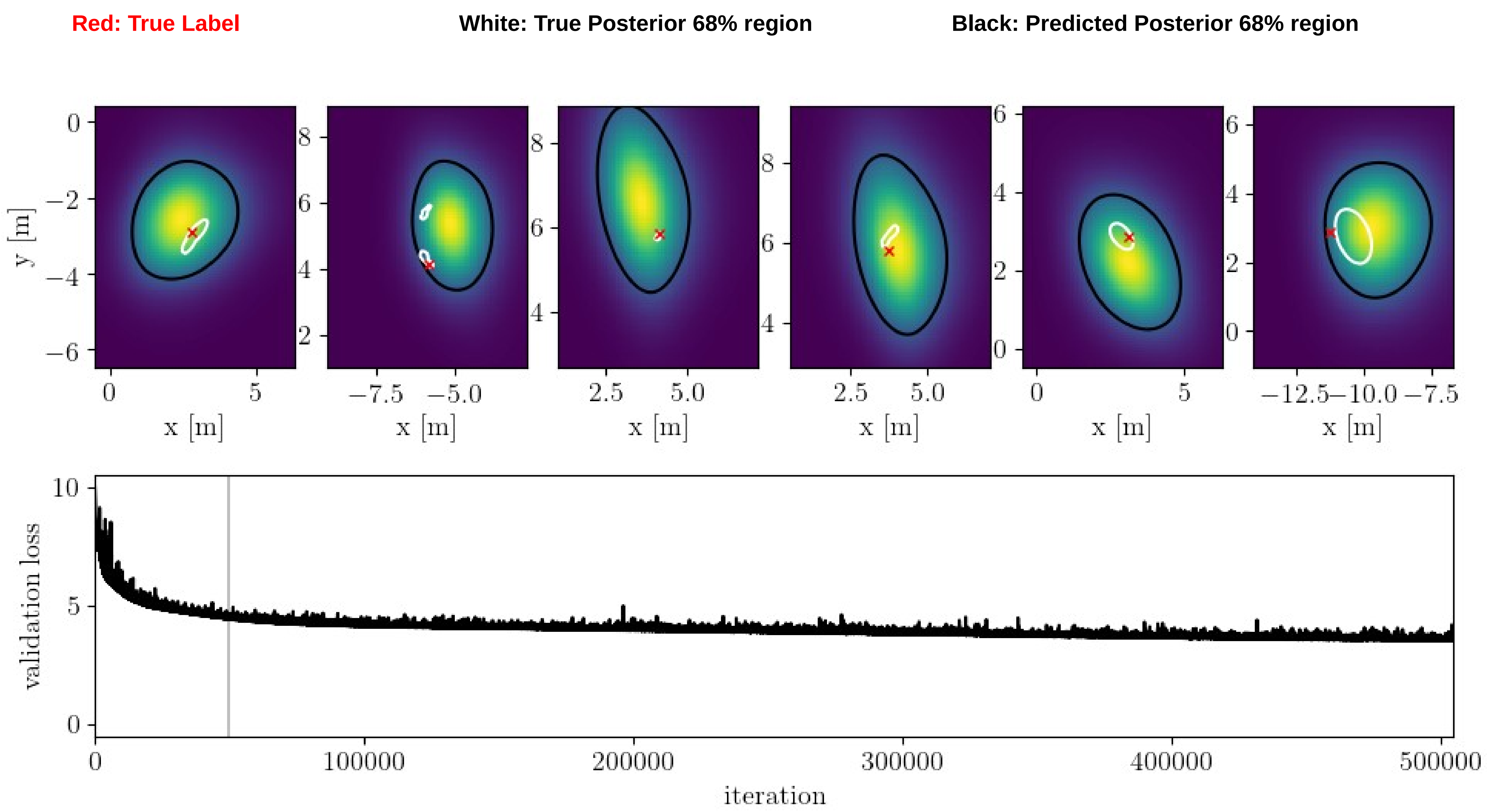


Red: True Label   White: True Posterior 68% region   Black: Predicted Posterior 68% region



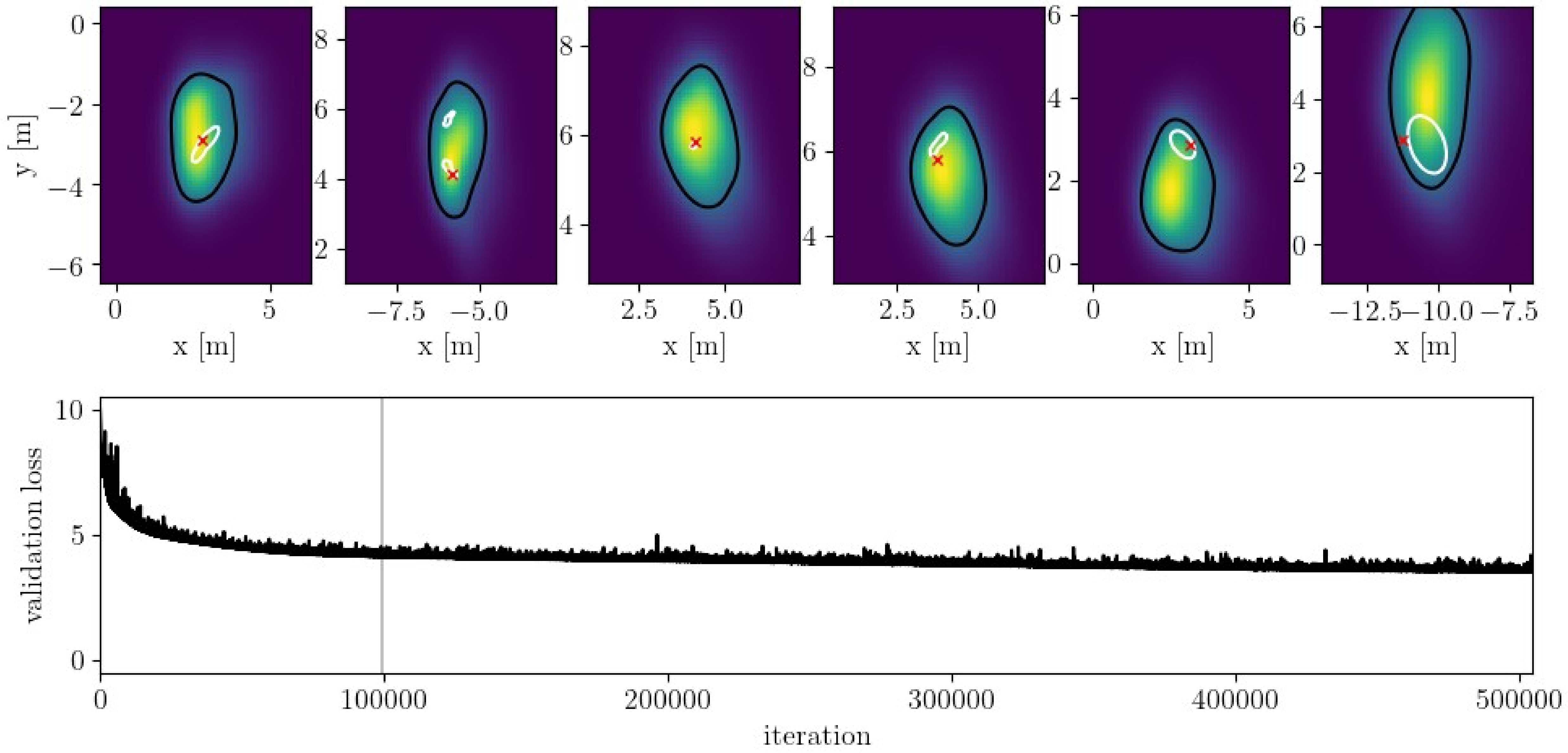


Red: True Label
White: True Posterior 68% region
Black: Predicted Posterior 68% region
y [m]
x [m]
validation loss
iteration

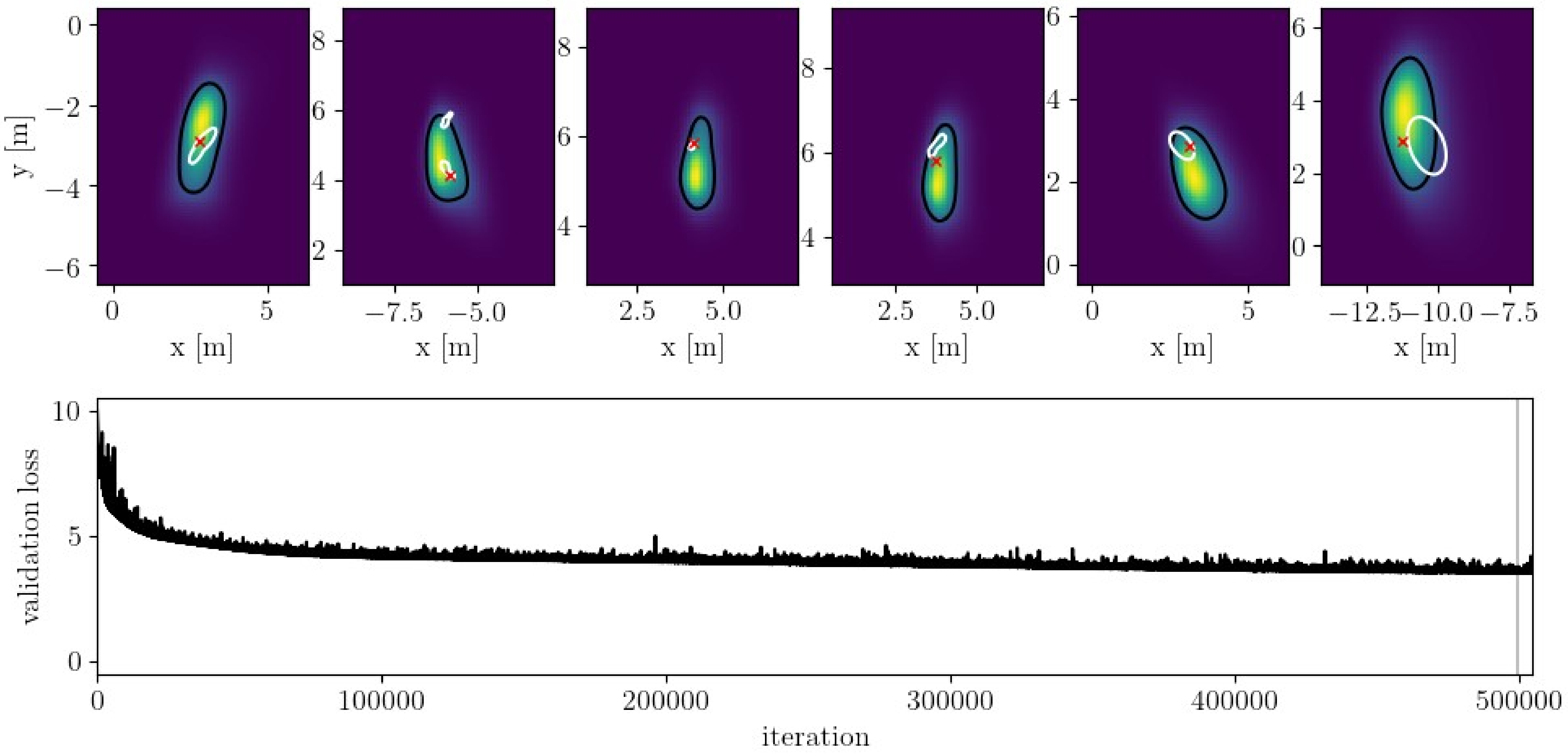


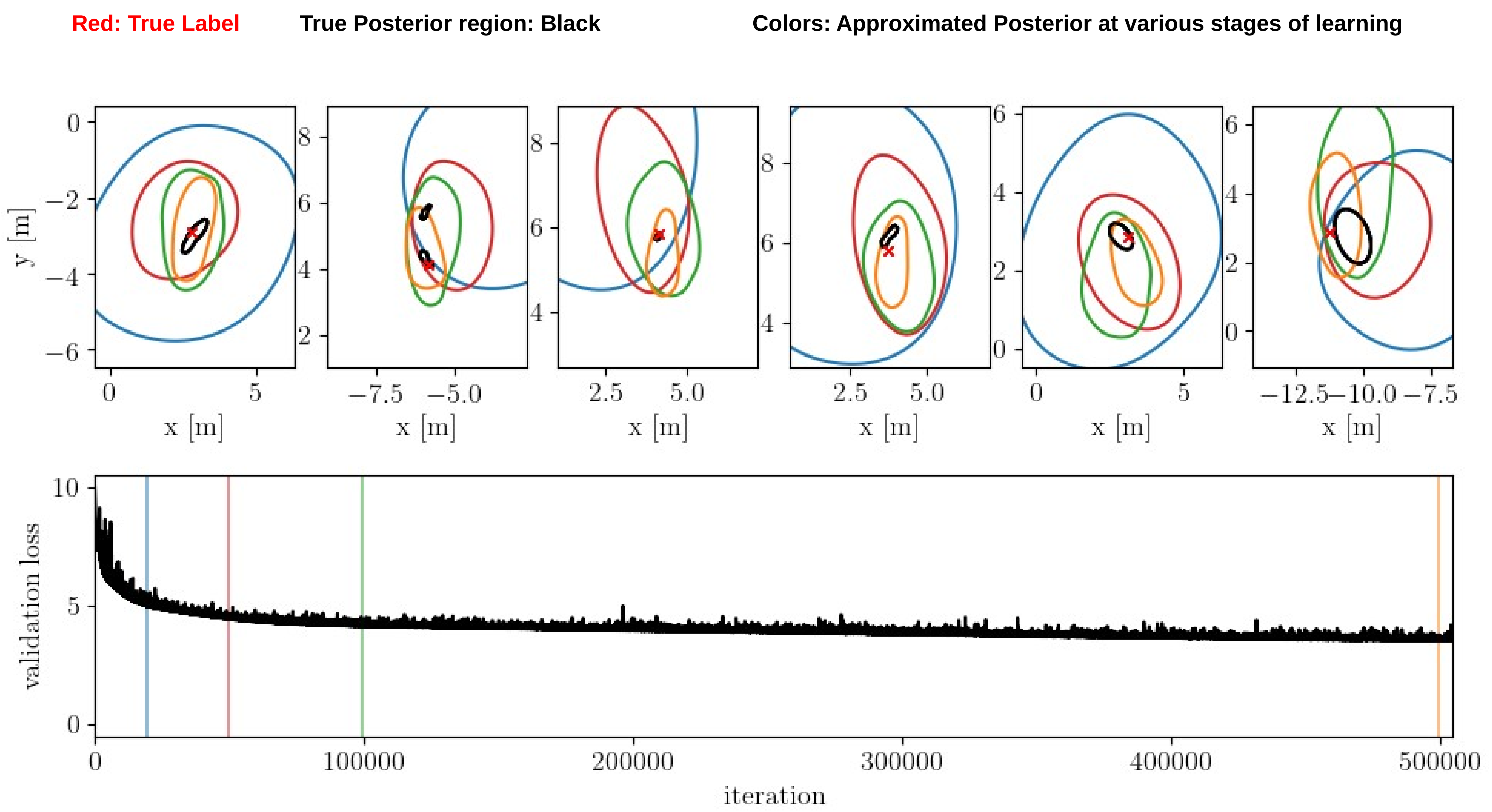
Red: True Label
White: True Posterior 68% region
Black: Predicted Posterior 68% region
y [m]
x [m]
validation loss
iteration

Red: True Label White: True Posterior 68% region Black: Predicted Posterior 68% region



Red: True Label
White: True Posterior 68% region
Black: Predicted Posterior 68% region


y [m]
x [m]
validation loss
iteration


Red: True Label
True Posterior region: Black
Colors: Approximated Posterior at various stages of learning
y [m]
x [m]
validation loss
iteration
0
100000
200000
300000
400000
500000

Systematics

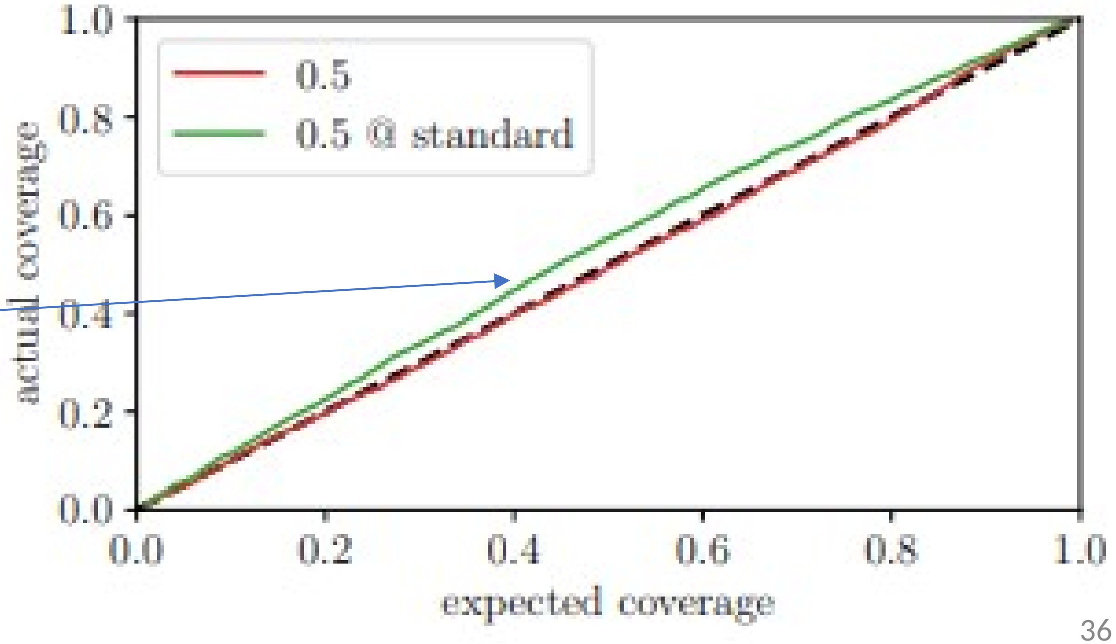
Draw events from prior $p(\nu)$ before event generation

This will effectively produce samples from the marginalized True distribution

$$\mathcal{P}_{t,M}(\theta;x) = \int \mathcal{P}_t(\theta;x,\nu) \cdot p(\nu) d\nu$$

, which again is approximated by
the neural network

Coverage for a distribution fitted with systematics (red) applied to a standard dataset with a fixed systematic value (green)



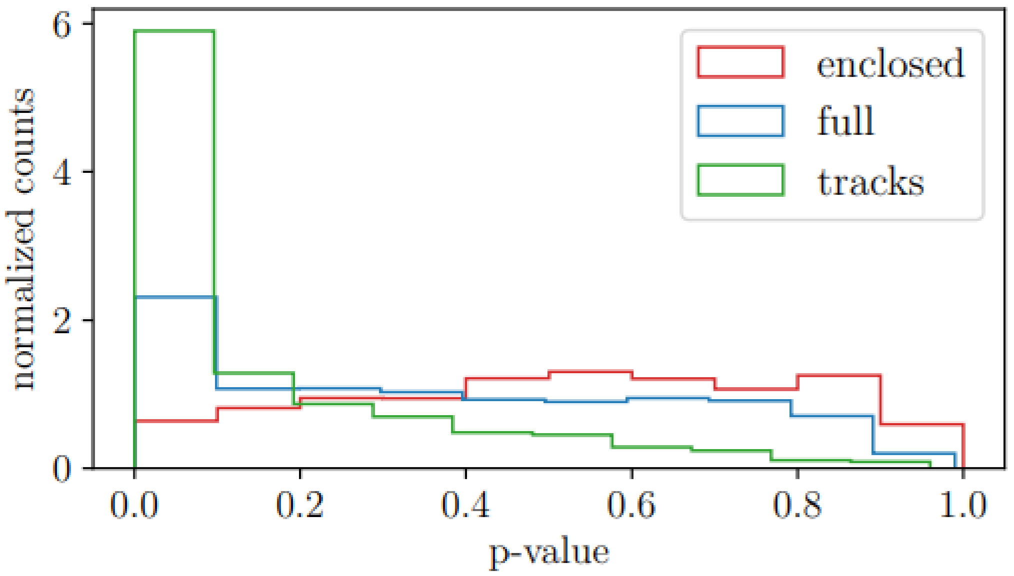
Overcoverage
is desired when including
systematics

Goodness-of-fit

- There is a well-motivated construction that upgrades the supervised loss to an extended loss , also fitting a likelihood p_θ (generative model) and some other terms (see 2008.05825 for details)
- Having both a generative model and the posterior allows to calculate a Bayesian p-value

$$p = \int_{x,z} \boldsymbol{I}_{T(x,z)>T(x_{\mathrm{obs}},z)} p_\theta(x;z) q_\phi(z;x_{\mathrm{obs}}) dx dz$$

$$T(x,z) = \ln p_\theta(x;z)/N_d$$



Any event sufficiently dissimilar to training data has a low p-value …

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

Summary

- **Supervised Learning learns to approximate the true posterior with a conditional PDF**
  -> **Bayesian systematics**: approximates marginalized Posterior
- Normalizing flows (NFs) can be designed to make this conditional PDF as precise as possible
  -> **supervised learning** can be „upgraded“ to behave as usable **likelihood-free inference** (standard MSE loss is not good enough for that)

Summary

- **Supervised Learning learns to approximate the true posterior with a conditional PDF** -> **Bayesian systematics**: approximates marginalized Posterior
- Normalizing flows (NFs) can be designed to make this conditional PDF as precise as possible -> **supervised learning** can be „upgraded“ to behave as usable **likelihood-free inference** (standard MSE loss is not good enough for that)
- Furthermore, normalizing flows allow to
  - **Calculate exact coverage** of the approximate posterior of ANY shape
    - Coverage is obtained very quickly in the training, long before it finishes!
  - **Calculate a goodness-of-fit** that can potentially be used in event selection
- We can upgrade our existing supervised learning models (CNN/LSTM) with NFs, **get improved performance (most for non-gaussian Posteriors**), and get coverage/g-o-f (incl. systematics)

More info: arXiv:2008.05825

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS