Probabilistic Inference and Learning *under Constraints*

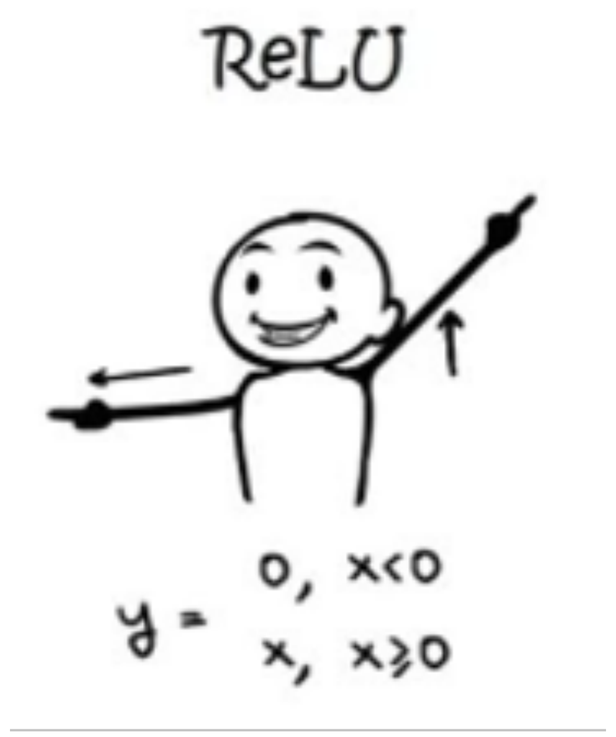
Zhe Zeng

University of California, Los Angeles

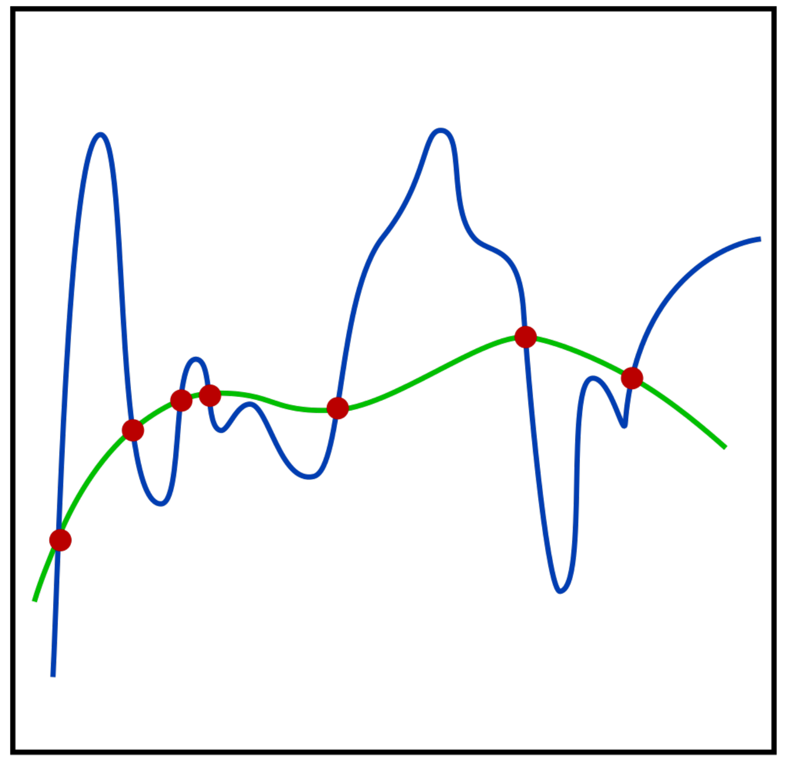
July 21st, 2023 - Amazon

Where are the *constraints* from?

Properties



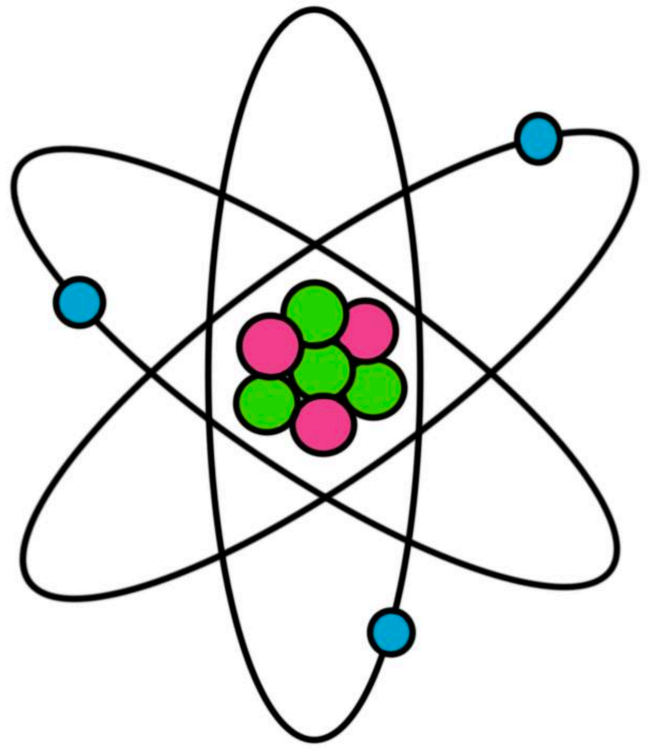
Architecture



Regularization

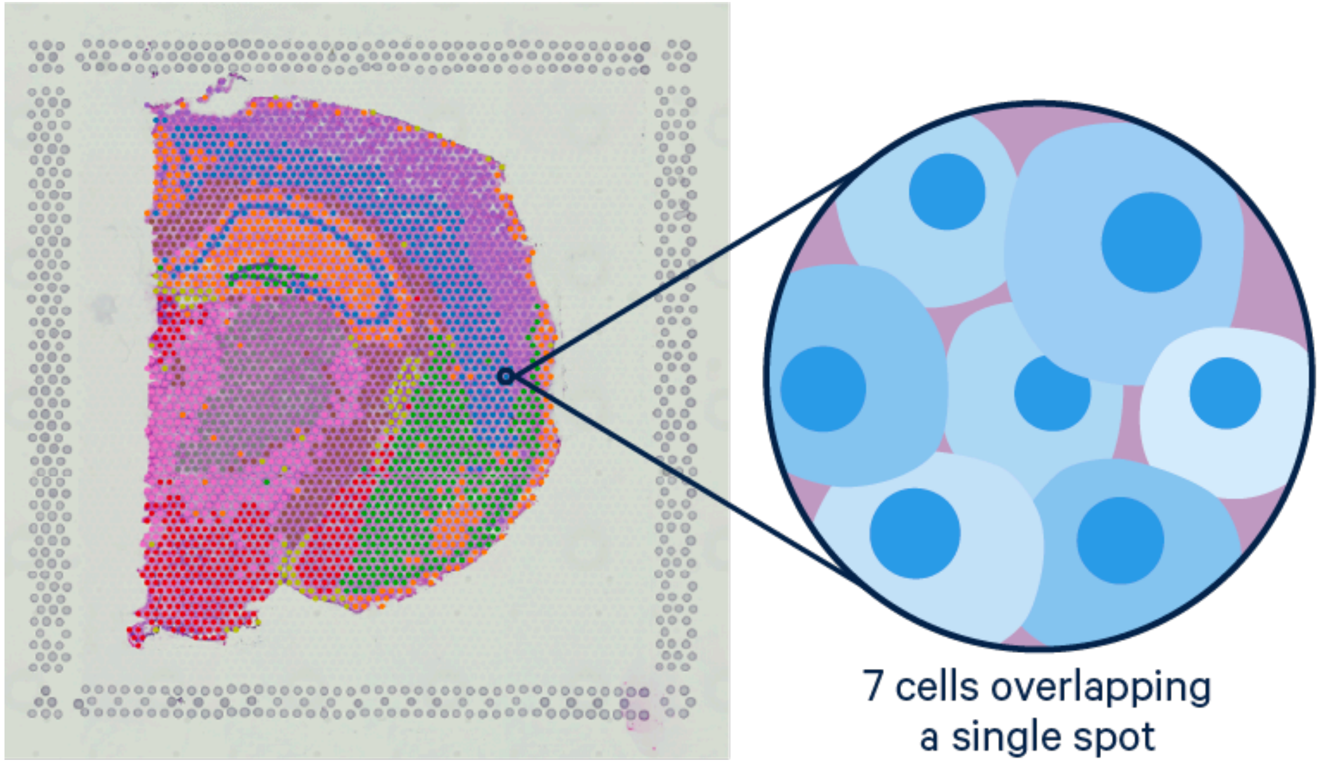
Explainability

Domain Knowledge



Physical Laws

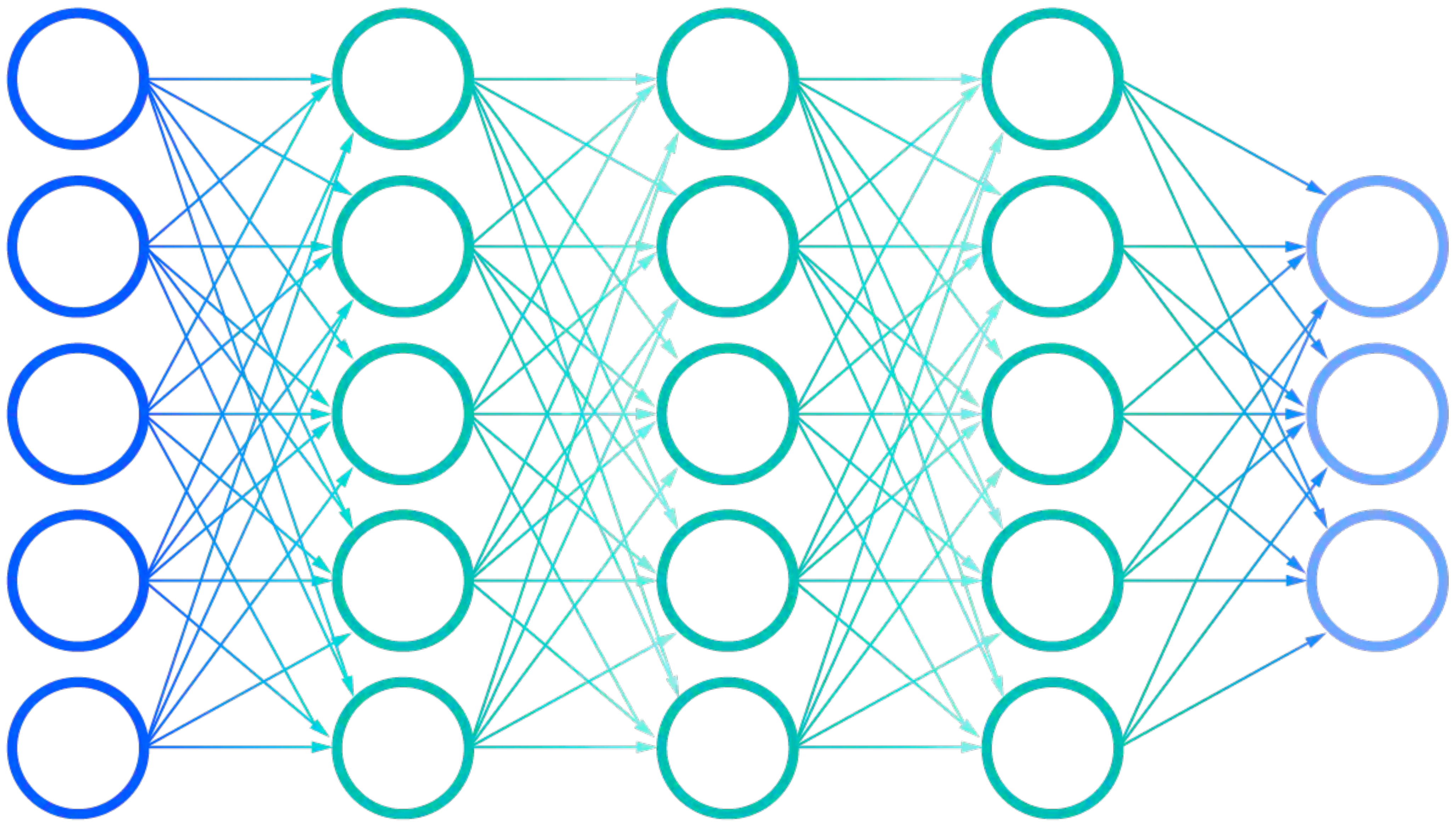
Molecular Structure

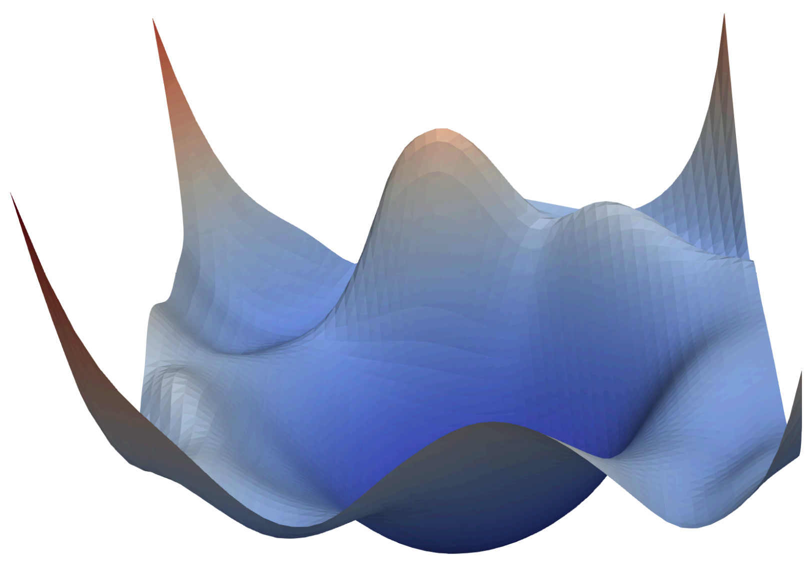


Gene Expression

Where to integrate *constraints*?

Input Latent Space Output Loss

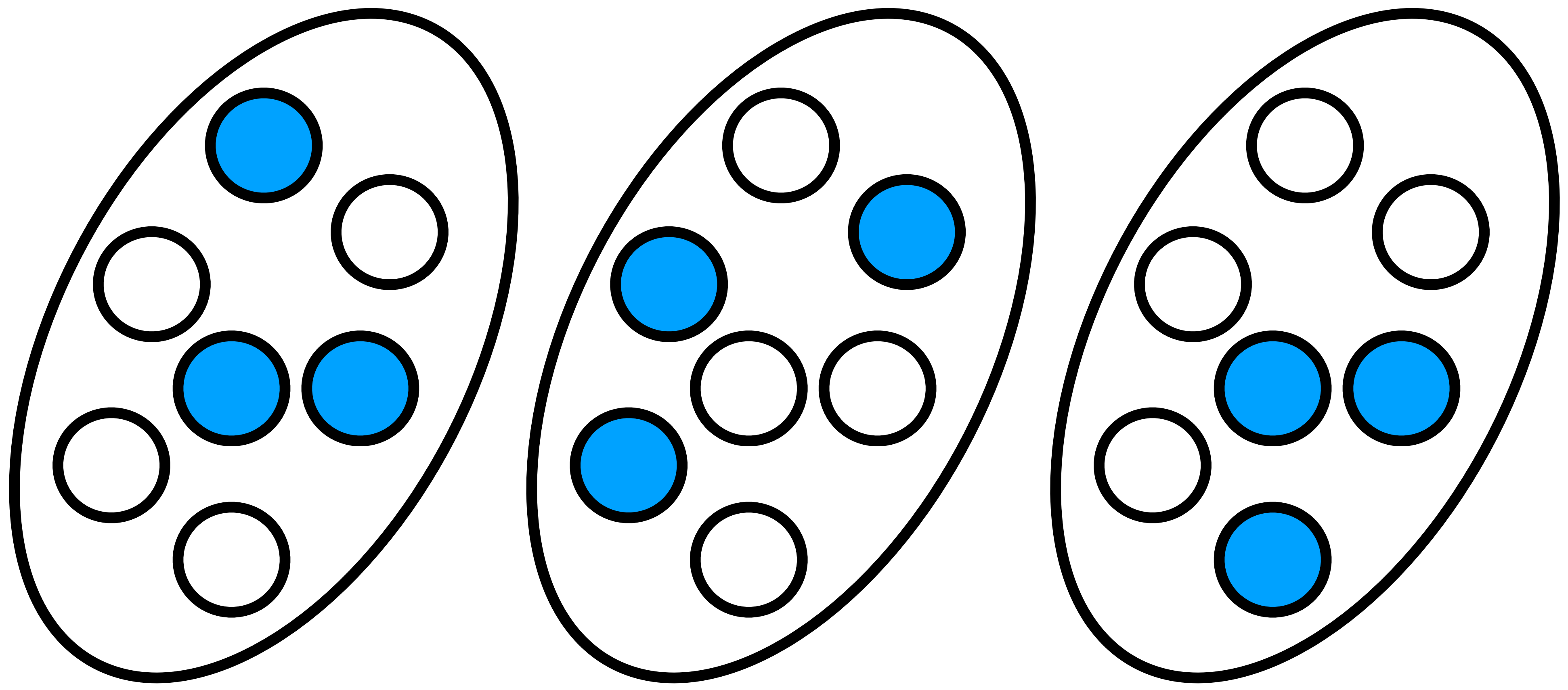




Challenges in *constraint integration*

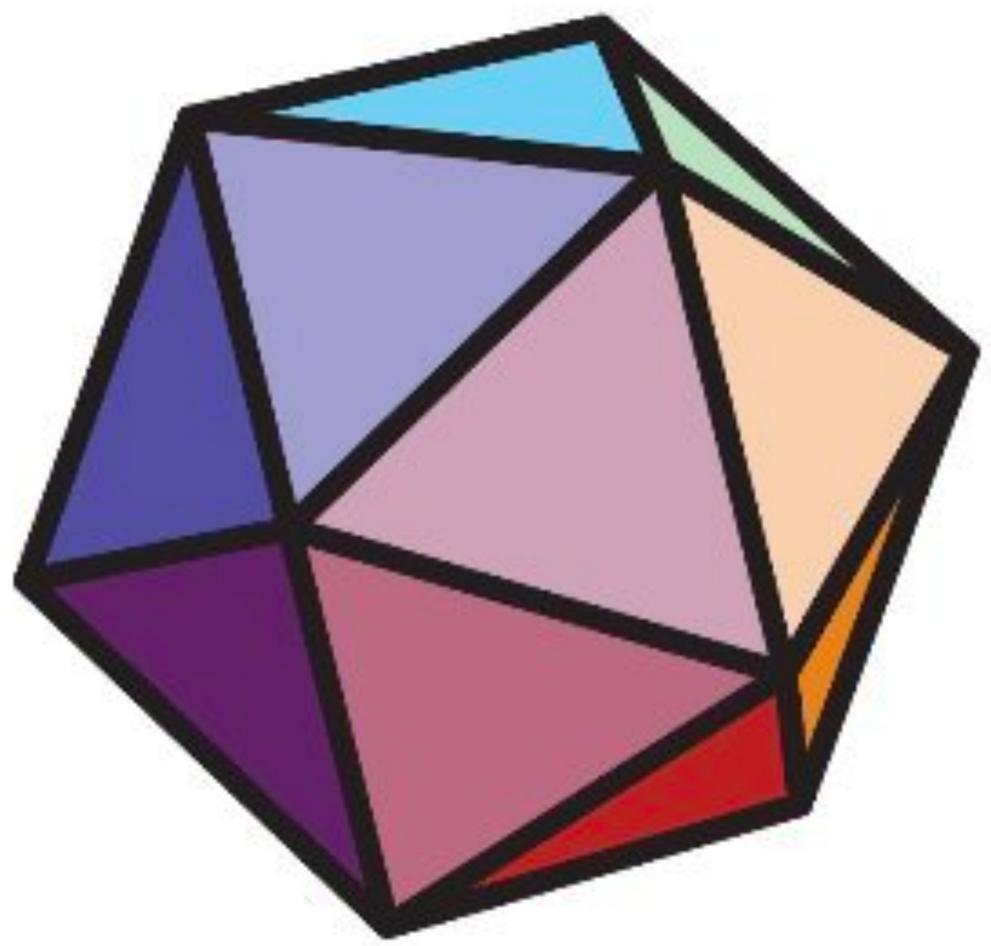
Non-differentiability

Discrete nature



Intractability

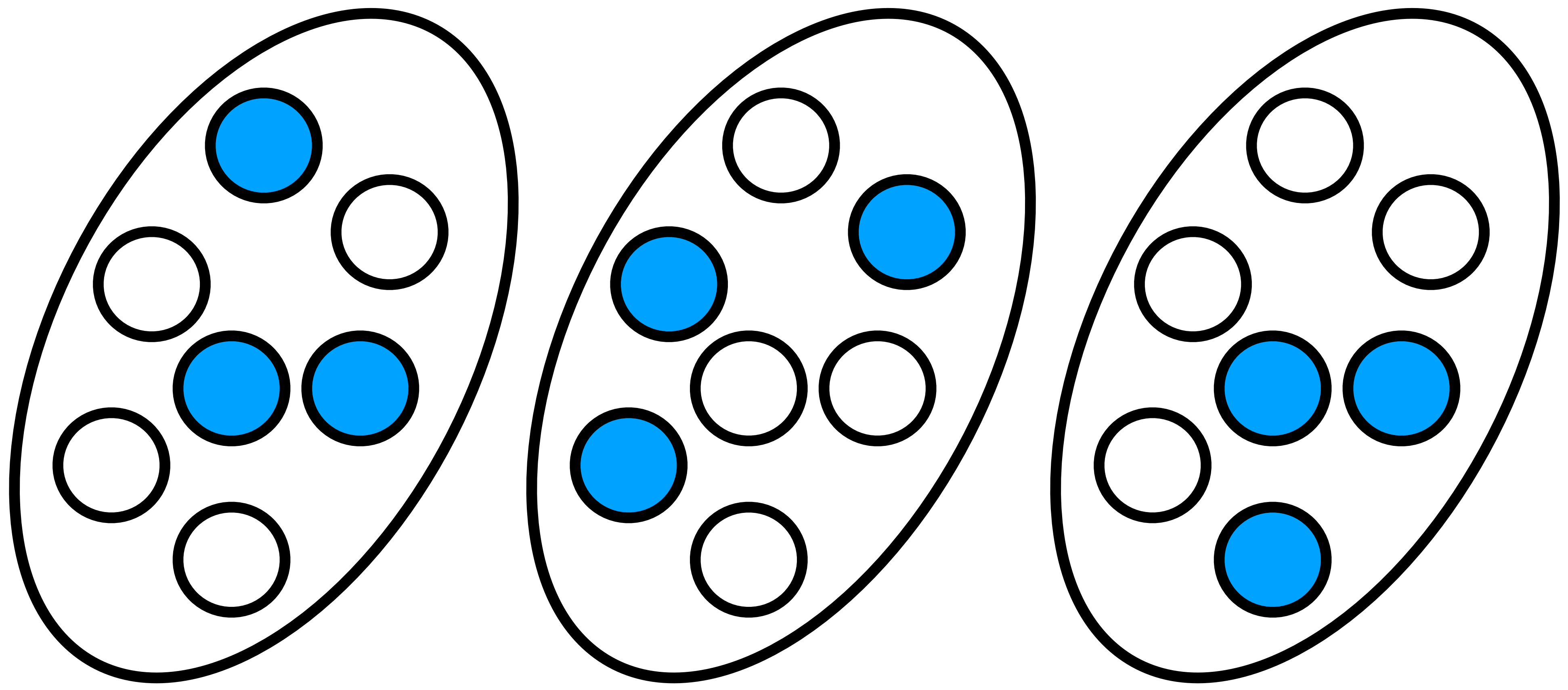
#P-hard



How to integrate *diverse constraints*?

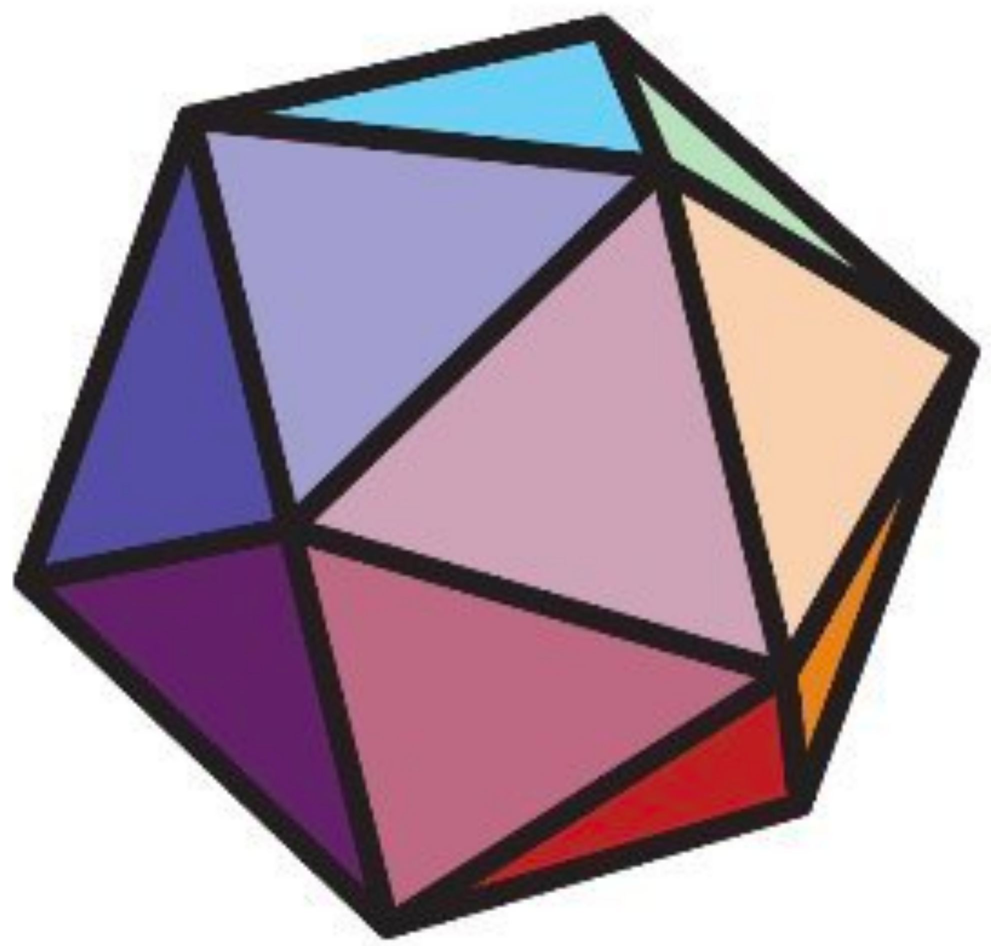
Non-differentiability

Discrete nature



Intractability

#P-hard



How to integrate *diverse constraints?*

Outline

- Differentiable learning under constraints
- Constrained probabilistic inference

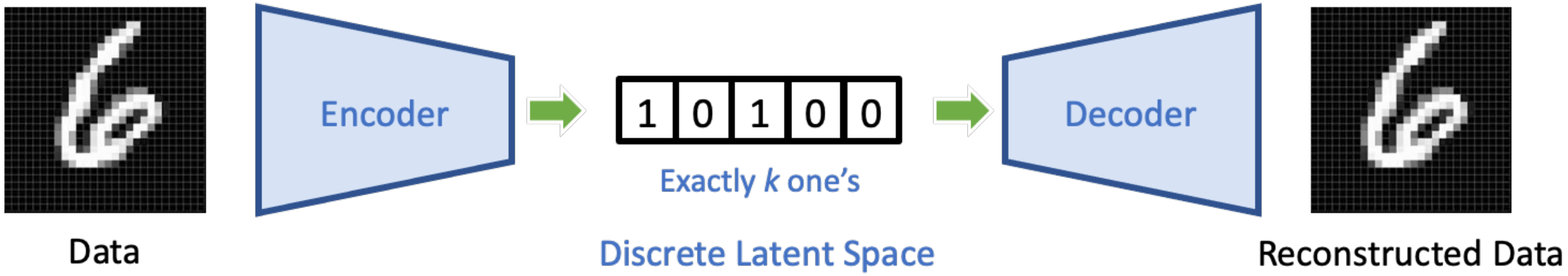
How to integrate *diverse constraints?*

Outline

- Differentiable learning under constraints
- Constrained probabilistic inference

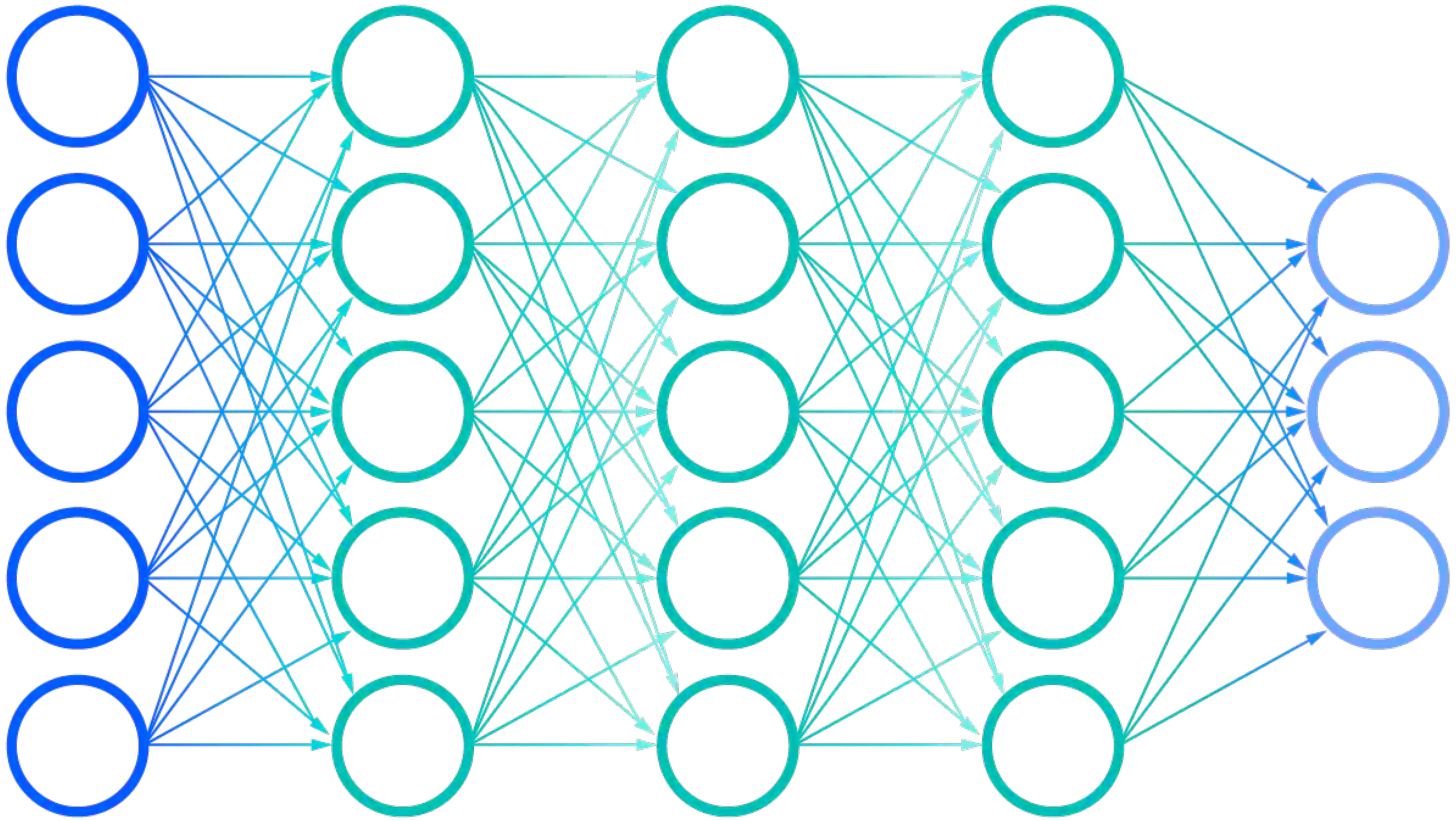
Why *k-subset constraint?*

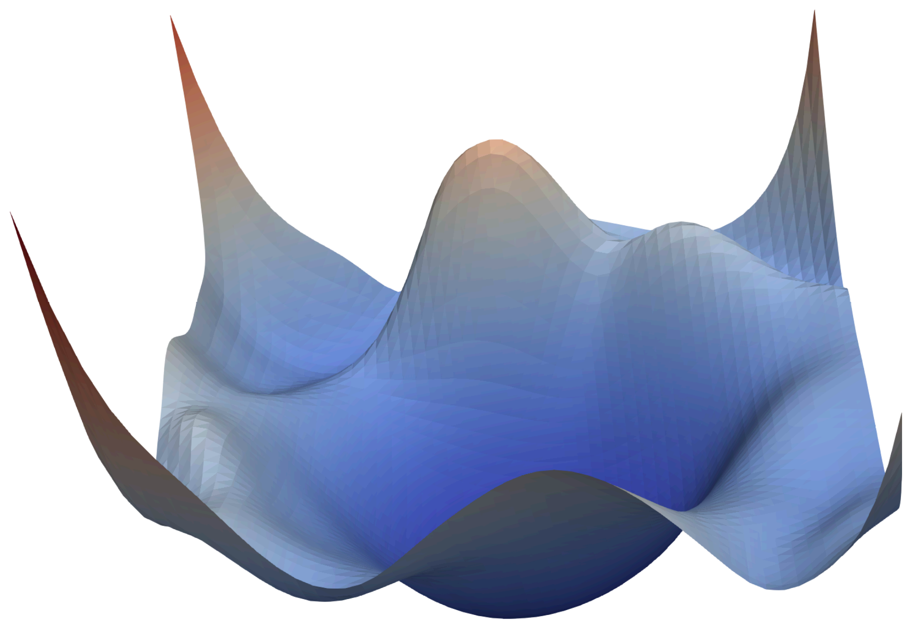
Discrete VAE



Differentiable learning under *k-subset* constraints

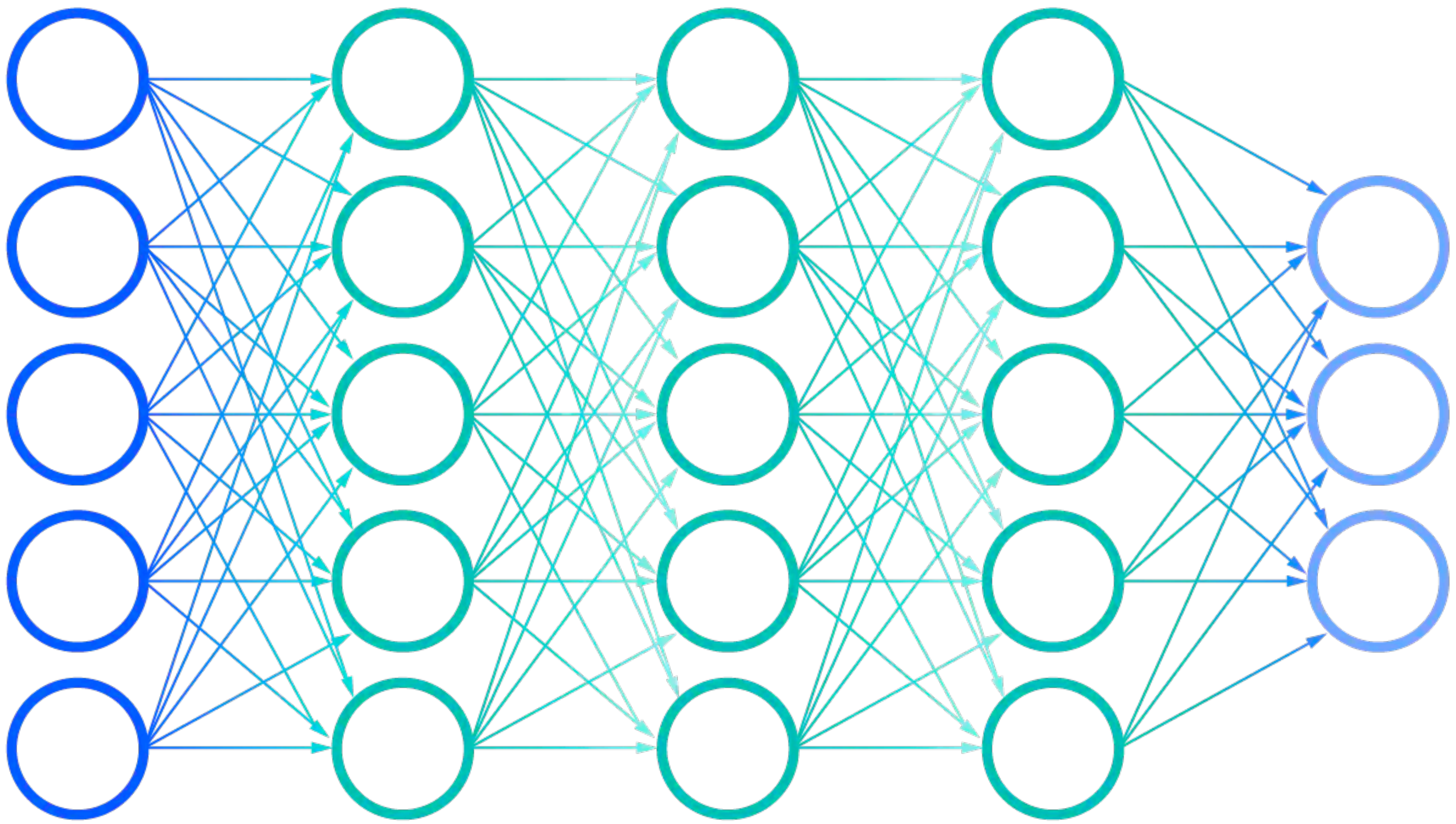
Input	Latent Space	Output	Loss
Learn to Explain	DiscreteVAE	AI for Science	Weakly Supervised

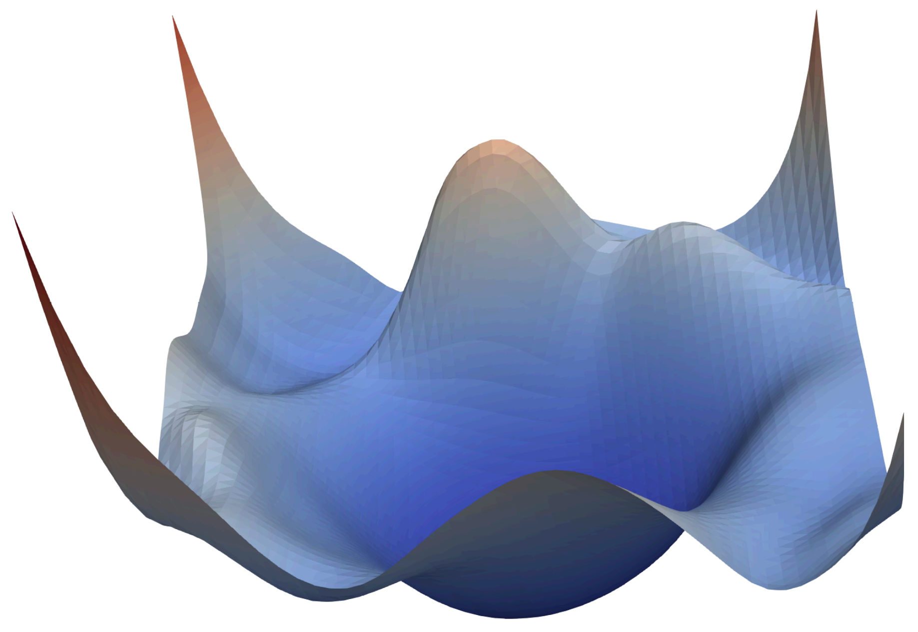




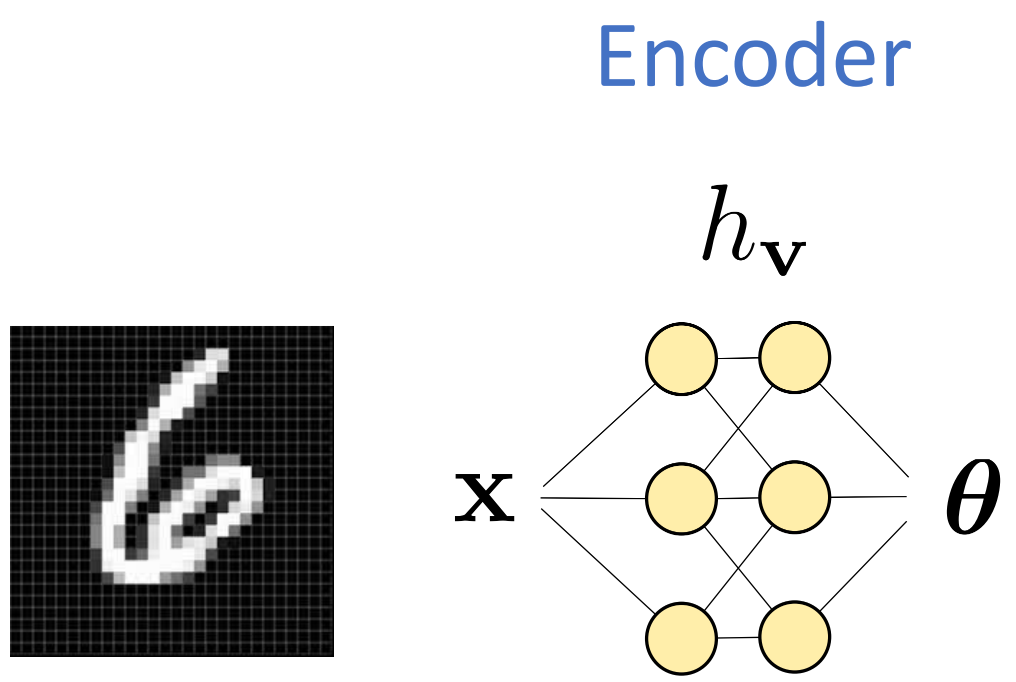
Differentiable learning under *k-subset* constraints

Input	**Latent Space**	Output	Loss
Learn to Explain	**DiscreteVAE**	AI for Science	Weakly Supervised





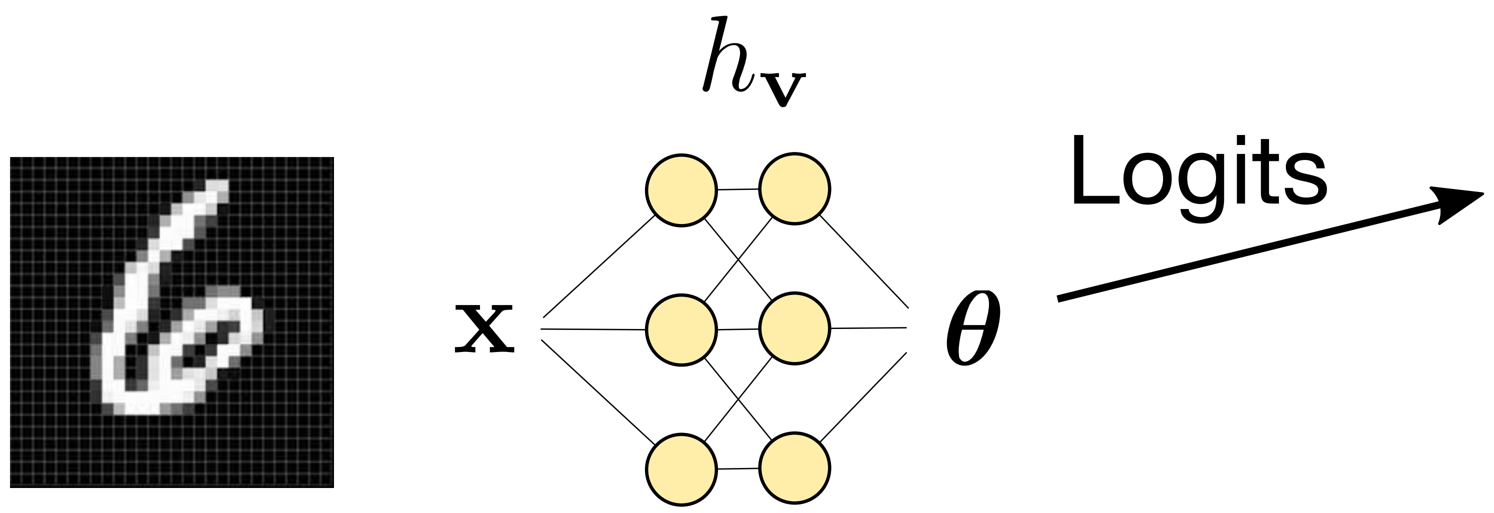
SIMPLE: Gradient Estimator for k-Subset Sampling [1]



SIMPLE: Gradient Estimator for k-Subset Sampling [1]

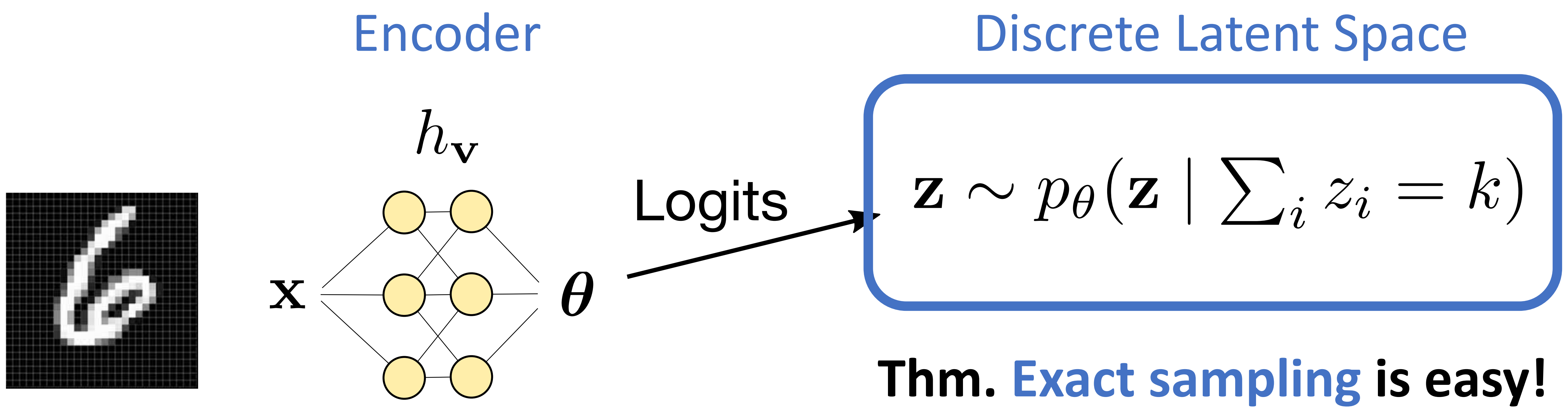
Encoder

Discrete Latent Space



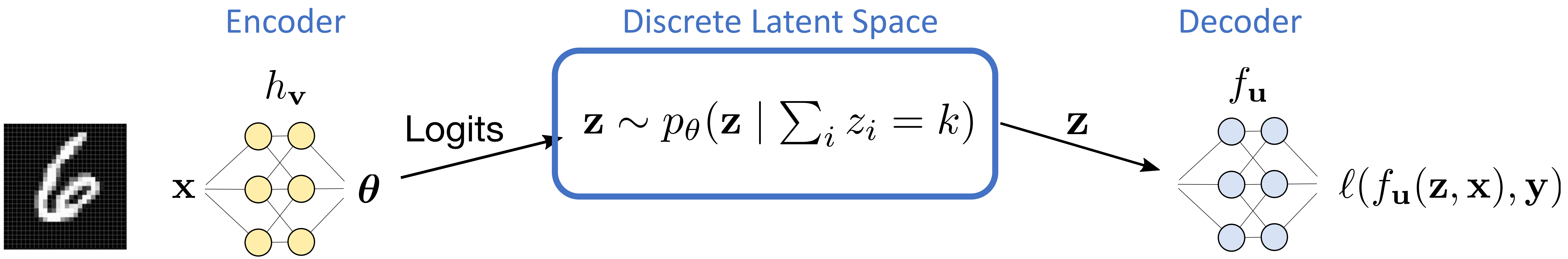
$$p_{\boldsymbol{\theta}}(\mathbf{z} \mid \sum_i z_i = k)$$

SIMPLE: Gradient Estimator for k-Subset Sampling [1]

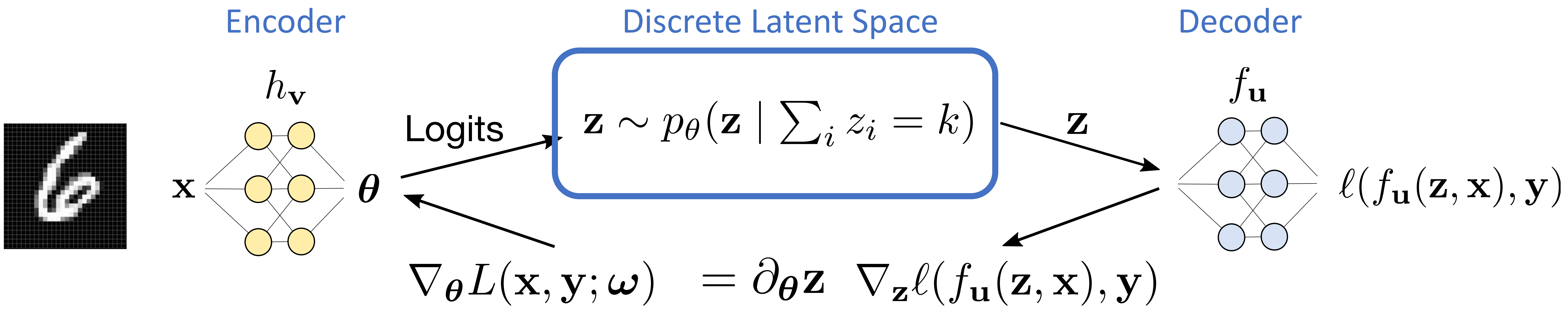


Thm. Exact sampling is easy!

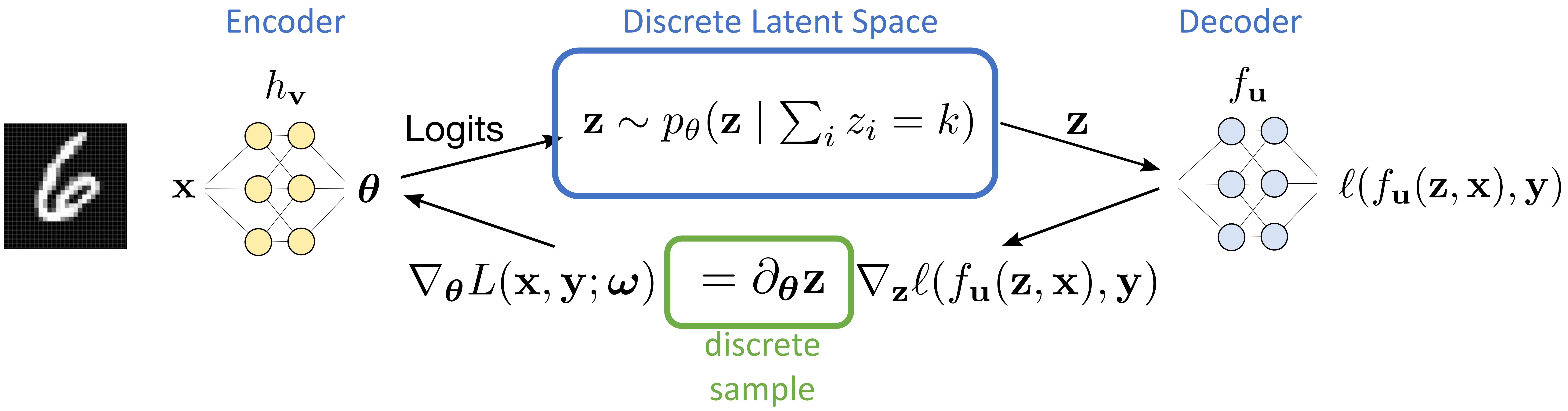
SIMPLE: Gradient Estimator for k-Subset Sampling [1]



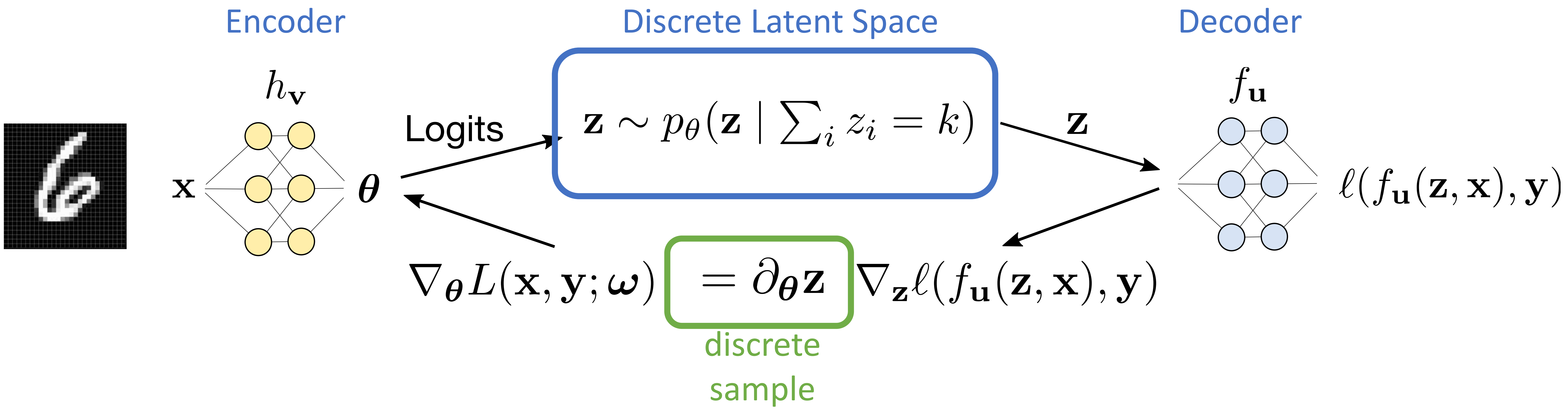
SIMPLE: Gradient Estimator for k-Subset Sampling [1]



SIMPLE: Gradient Estimator for k-Subset Sampling [1]



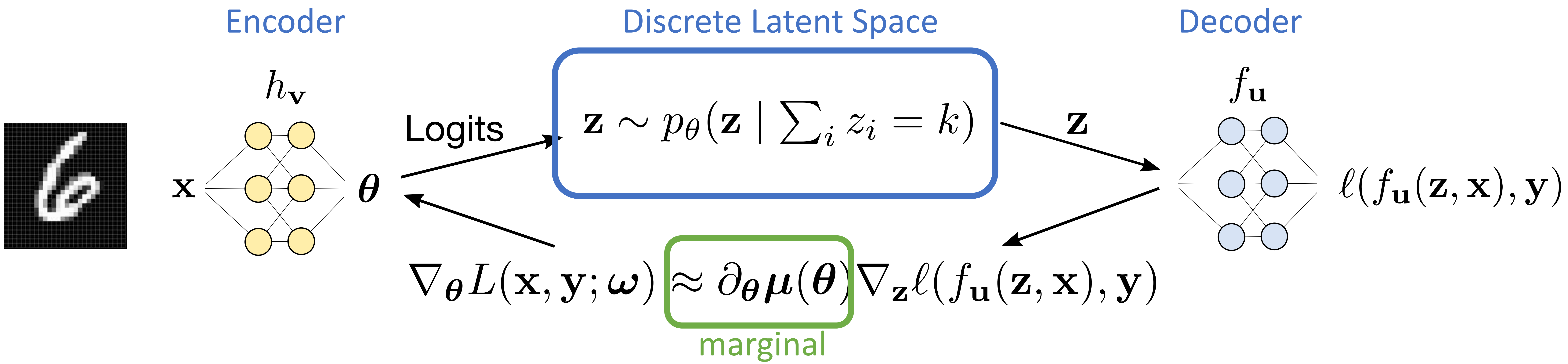
SIMPLE: Gradient Estimator for k-Subset Sampling [1]



Intuition: update θ such that

z_1	z_2	z_3	$p(z_1 \mid \sum_i z_i = k)$	$p(z_2 \mid \sum_i z_i = k)$	$p(z_3 \mid \sum_i z_i = k)$
1	1	0	0.7	0.8	0.2
	↓	↓		↓	↓
1	0	1	0.7	0.3	0.75

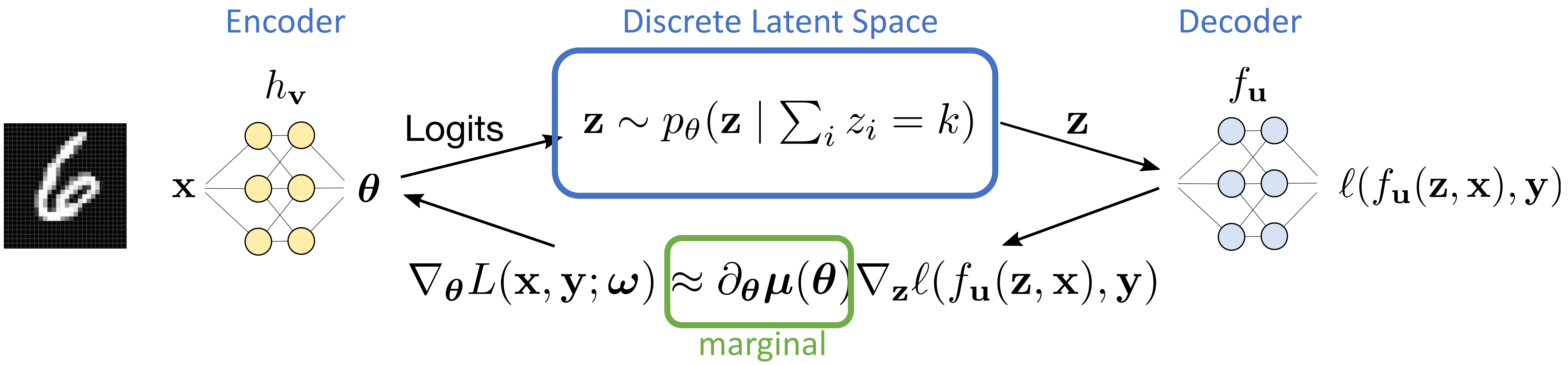
SIMPLE: Gradient Estimator for k-Subset Sampling [1]



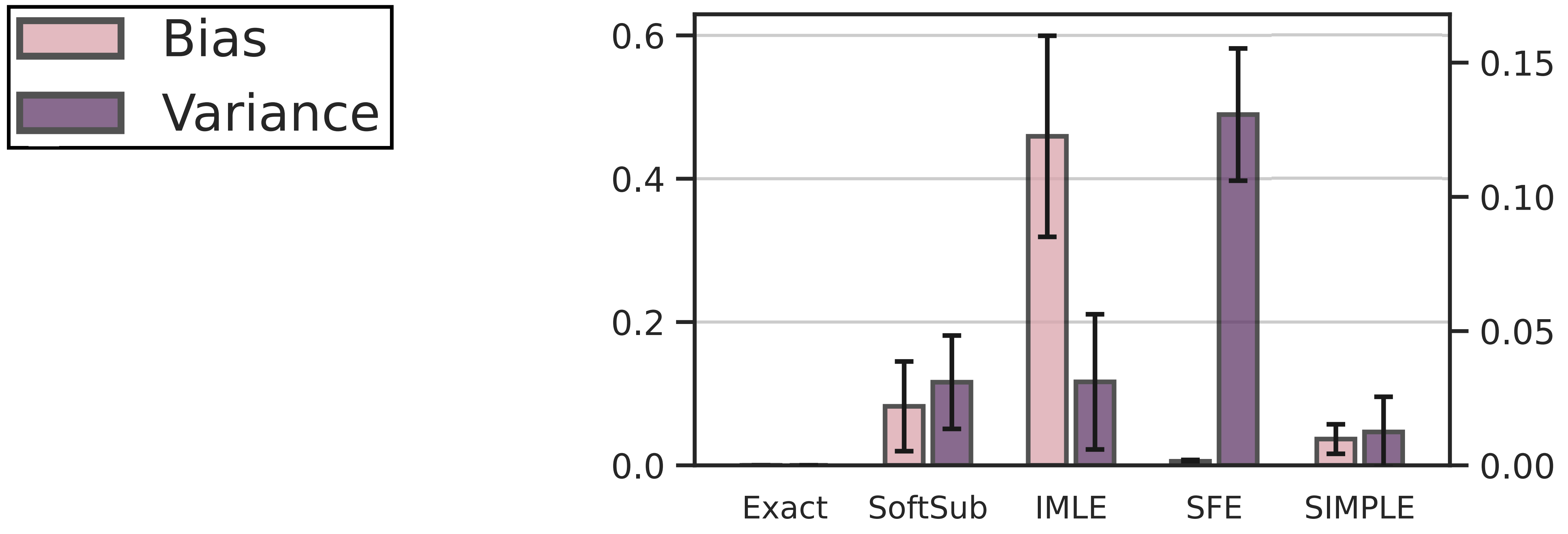
Prop. conditional marginals can be obtained by

$$\frac{\partial}{\partial \theta_i} \log p_{\boldsymbol{\theta}}(\textstyle\sum_j z_j = k) = p_{\boldsymbol{\theta}}\left(z_i \mid \textstyle\sum_j z_j = k\right) = \mu(\boldsymbol{\theta})$$

SIMPLE: Gradient Estimator for k-Subset Sampling [1]



We achieve ***lower bias and variance*** by exact, discrete samples and exact derivative of conditional marginals.



Ablation Study

Why constraint probability helps?

Perturb-and-map (PAM)

PAM Sampling

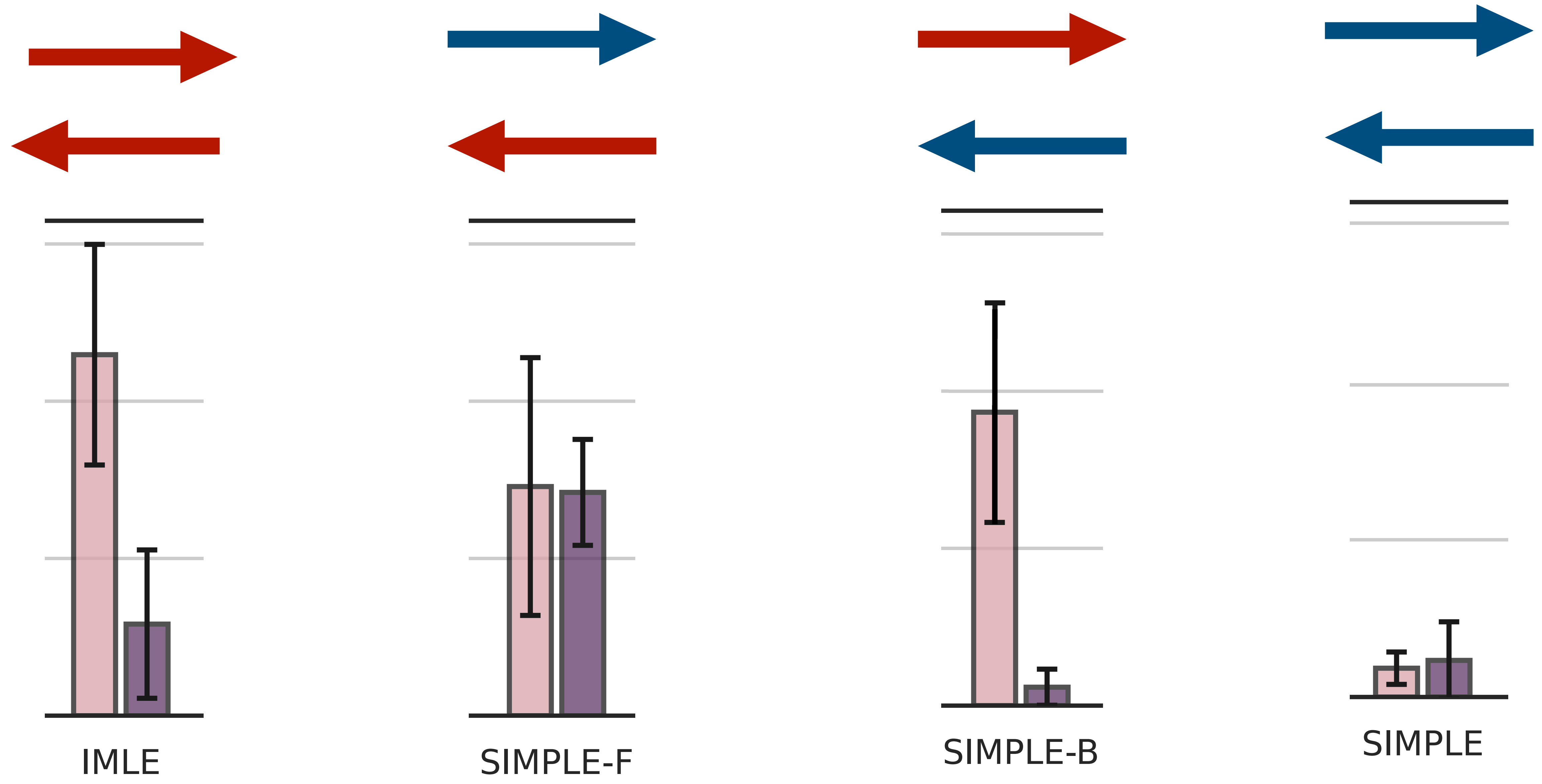
PAM Marginal

Exact computation by SIMPLE

Exact Sampling $\mathbf{z} \sim p_\theta(\mathbf{z} \mid \sum_i z_i = k)$

Exact Marginal $\mu(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}\left(z_i \mid \sum_j z_j = k\right)$

Bias
Variance



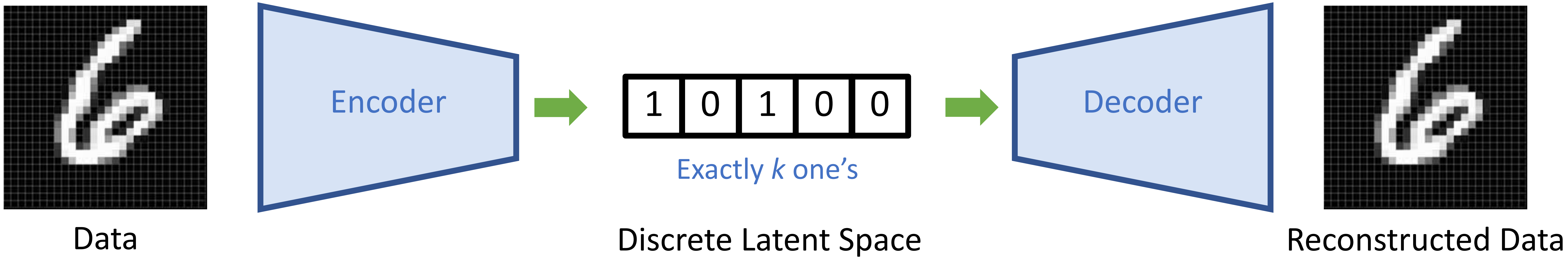
Baseline

Exact sampling helps reduce bias!

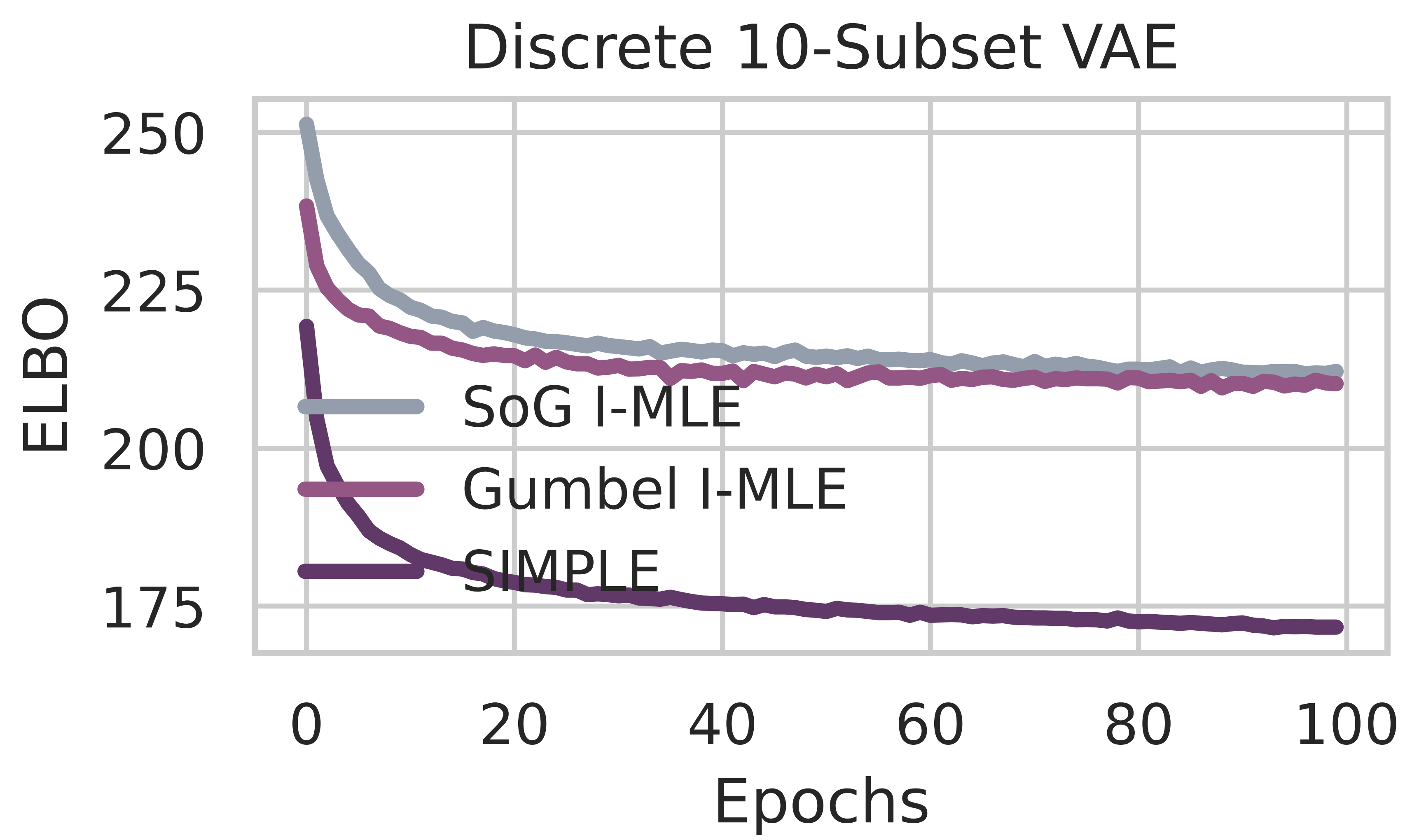
Exact marginal helps reduce variance!

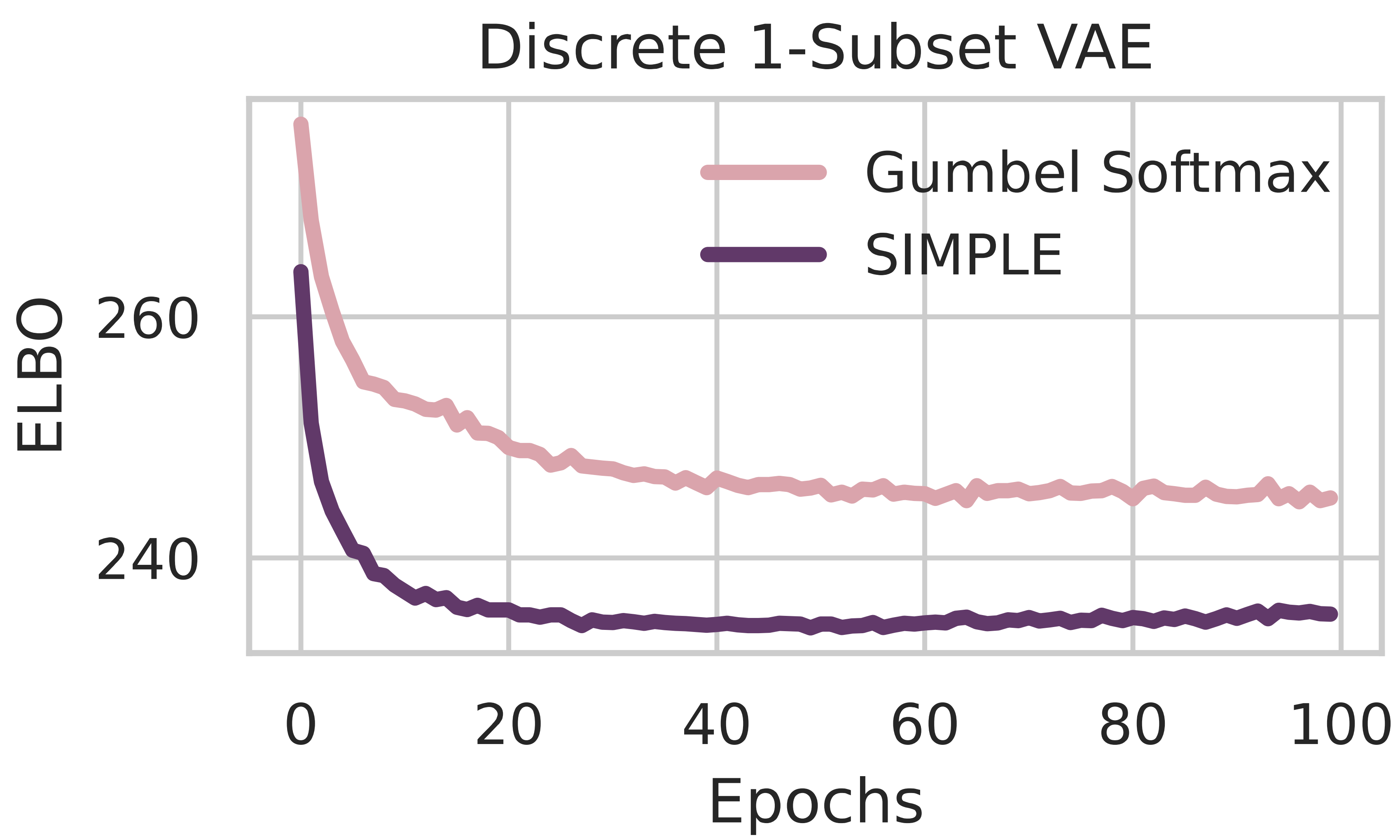
Both are reduced!

Experiment



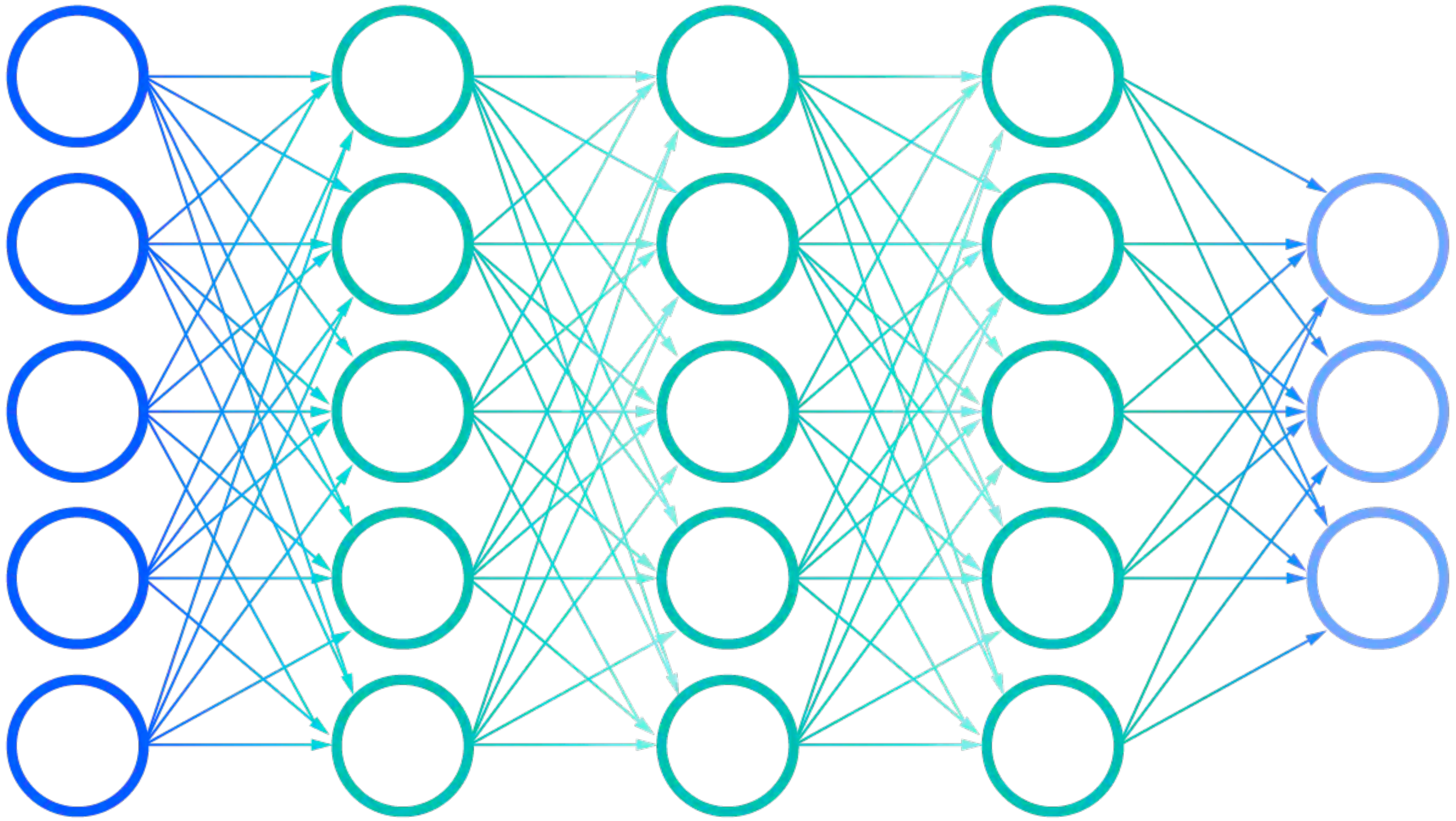
Metric: *exact* ELBO

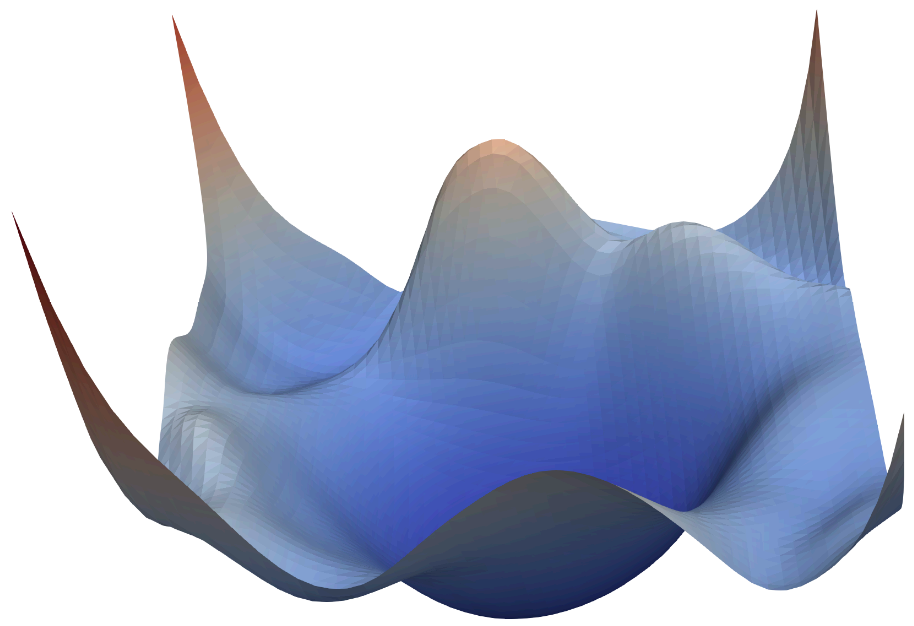




Differentiable learning under *k-subset* constraints

Input	Latent Space	Output	Loss
Learn to Explain	DiscreteVAE	AI for Science	Weakly Supervised





Learn to Explain (L2X)[1]

Input: **Key words (*k* = 10)**	**Output:** **Taste Score**
a lite bodied beer with a pleasant taste. was like a reddish color. a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly. not hard to down, though a bit harder then one would expect given the taste.	0.7

Learn to Explain (L2X)

Input: **Key words (k = 10)**	**Output:** **Taste Score**
a lite bodied beer with a pleasant taste. was like a reddish color. a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly. not hard to down, though a bit harder then one would expect given the taste.	0.7

Results for three aspects with k = 10

Method	Appearance		Palate		Taste	
	Test MSE	Precision	Test MSE	Precision	Test MSE	Precision
SIMPLE (Ours)	**2.35 ± 0.28**	**66.81 ± 7.56**	**2.68 ± 0.06**	**44.78 ± 2.75**	**2.11 ± 0.02**	**42.31 ± 0.61**
L2X (t = 0.1)	10.70 ± 4.82	30.02 ± 15.82	6.70 ± 0.63	**50.39 ± 13.58**	6.92 ± 1.61	32.23 ± 4.92
SoftSub (t = 0.5)	**2.48 ± 0.10**	52.86 ± 7.08	2.94 ± 0.08	39.17 ± 3.17	2.18 ± 0.10	**41.98 ± 1.42**
I-MLE (τ = 30)	**2.51 ± 0.05**	**65.47 ± 4.95**	2.96 ± 0.04	40.73 ± 3.15	2.38 ± 0.04	**41.38 ± 1.55**

Results for aspect Aroma, for k in {5, 10, 15}

Method	$k = 5$		$k = 10$		$k = 15$	
	Test MSE	Precision	Test MSE	Precision	Test MSE	Precision
SIMPLE (Ours)	**2.27 ± 0.05**	**57.30 ± 3.04**	**2.23 ± 0.03**	**47.17 ± 2.11**	3.20 ± 0.04	**53.18 ± 1.09**
L2X (t = 0.1)	5.75 ± 0.30	33.63 ± 6.91	6.68 ± 1.08	26.65 ± 9.39	7.71 ± 0.64	23.49 ± 10.93
SoftSub (t = 0.5)	2.57 ± 0.12	**54.06 ± 6.29**	2.67 ± 0.14	44.44 ± 2.27	**2.52 ± 0.07**	37.78 ± 1.71
I-MLE (τ = 30)	2.62 ± 0.05	**54.76 ± 2.50**	2.71 ± 0.10	**47.98 ± 2.26**	2.91 ± 0.18	39.56 ± 2.07

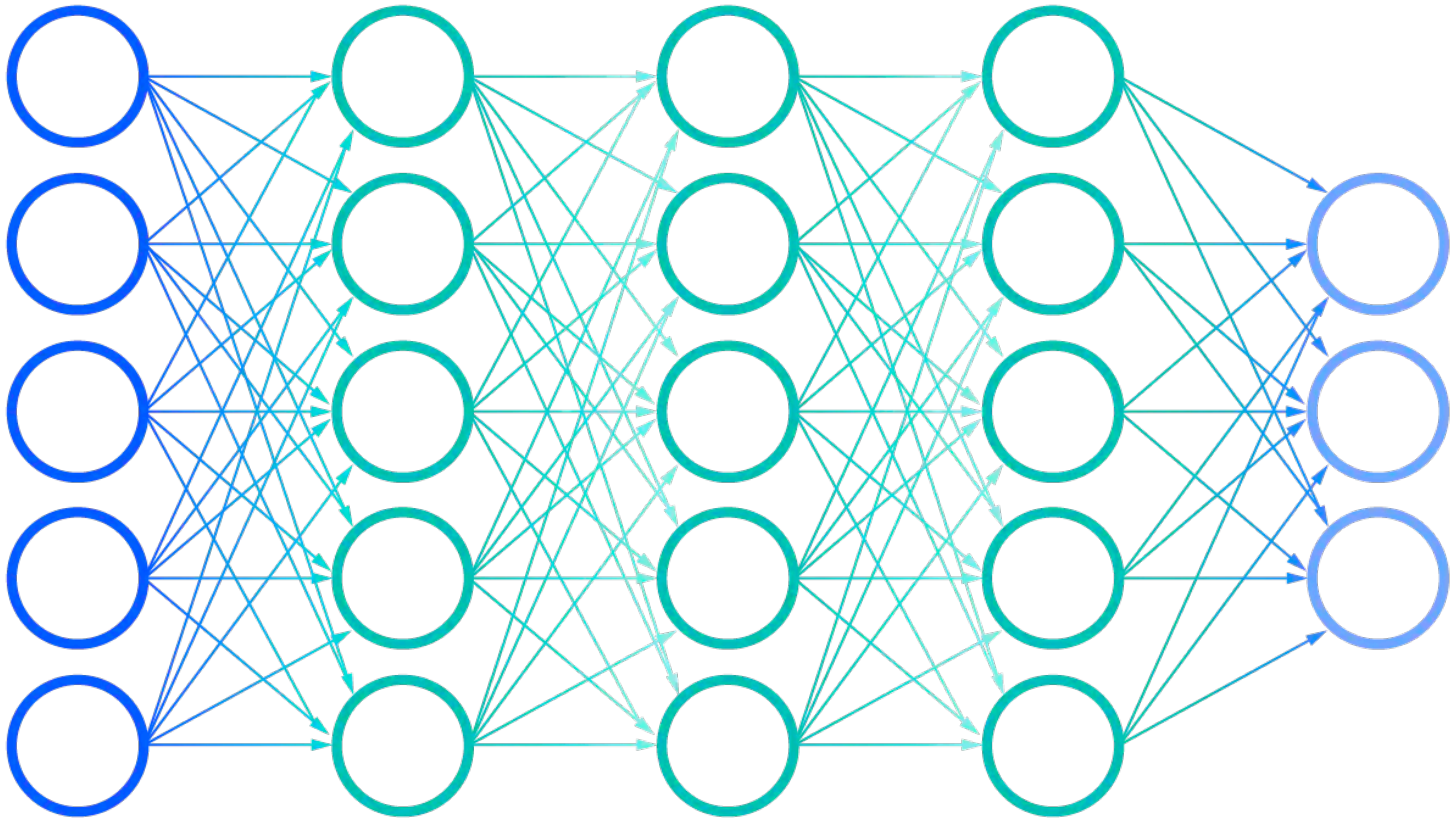
Differentiable learning under *k-subset* constraints

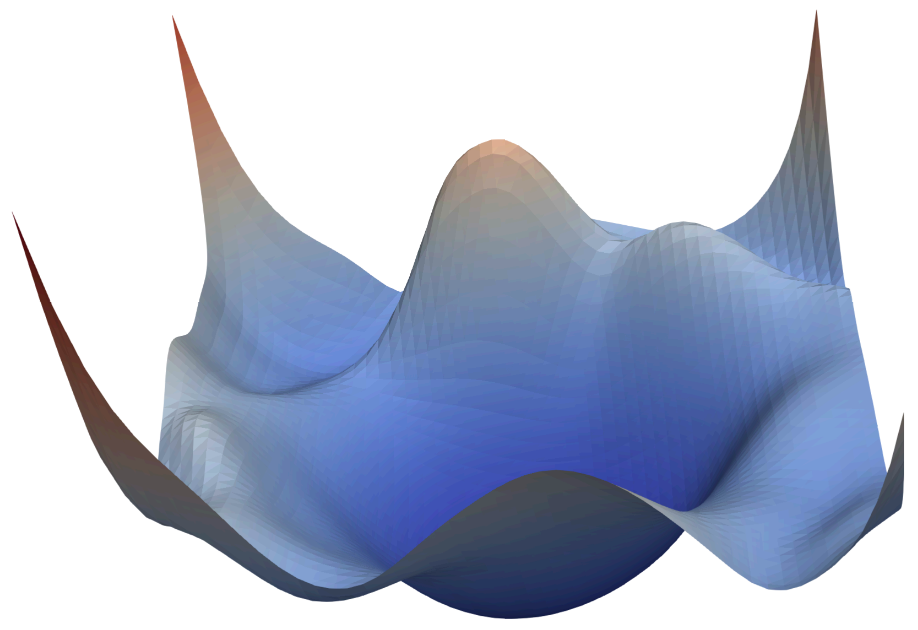
Input
Learn to Explain

Latent Space
DiscreteVAE

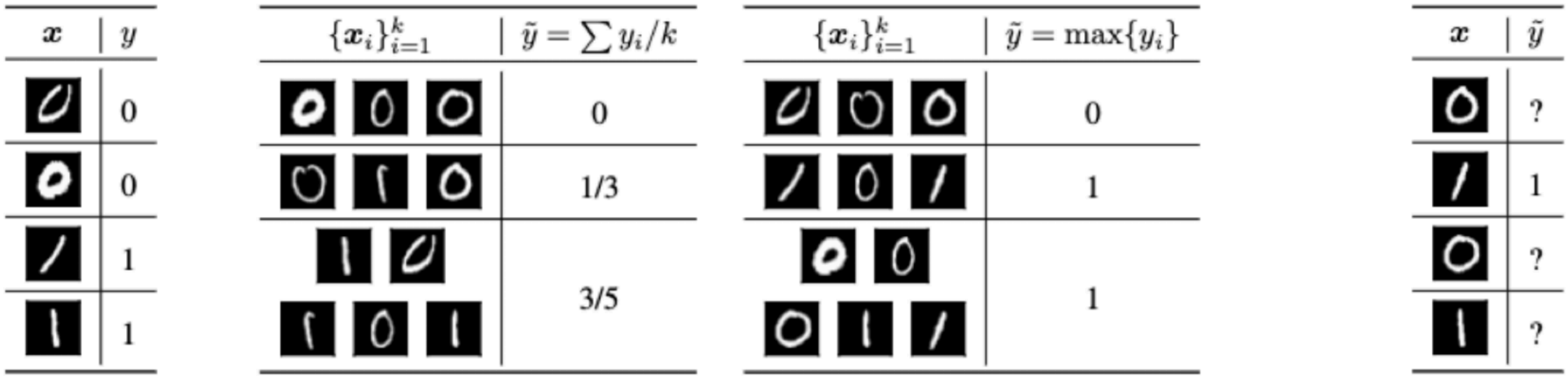
Output
AI for Science

Loss
Weakly Supervised





A Unified Approach to Count-Based Weakly-Supervised Learning[2]



Classical

Learning from Label Proportions

Multiple Instance Learning

Learning from Positive & Unlabeled

Objective: To maximize the probability of weak supervisions, i.e., constraints on label counts

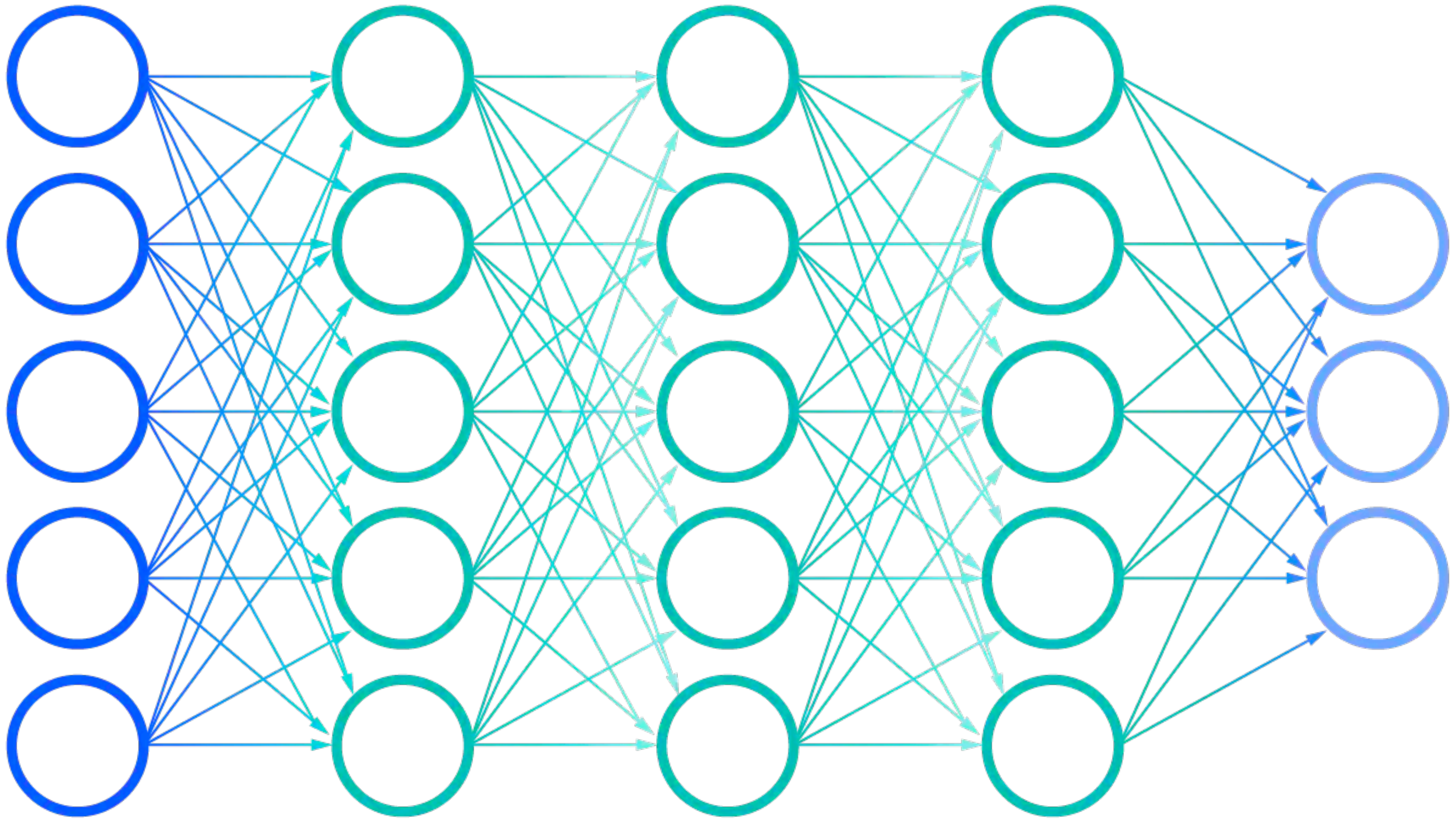
Differentiable learning under *k-subset* constraints

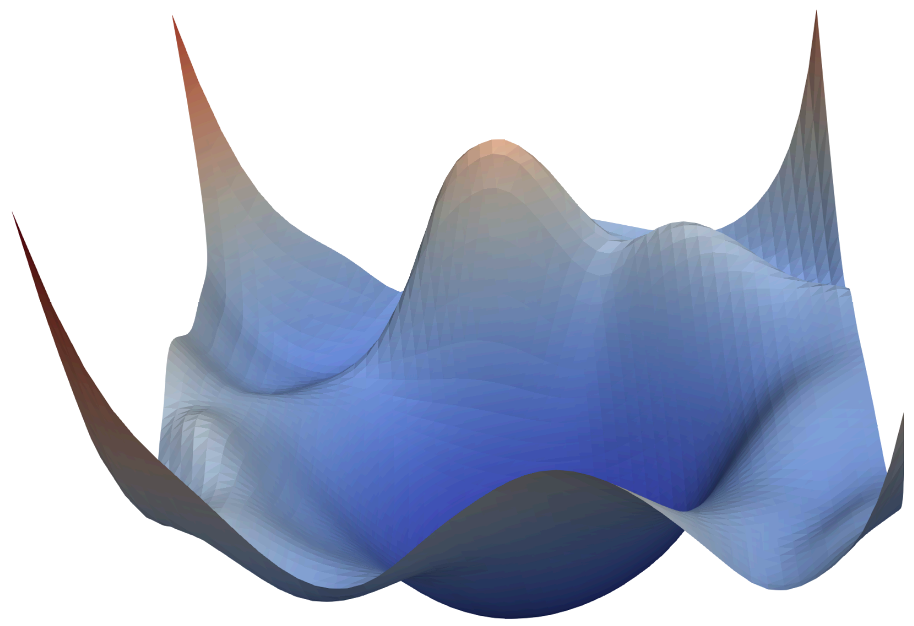
Input
Learn to Explain

Latent Space
DiscreteVAE

Output
AI for Science

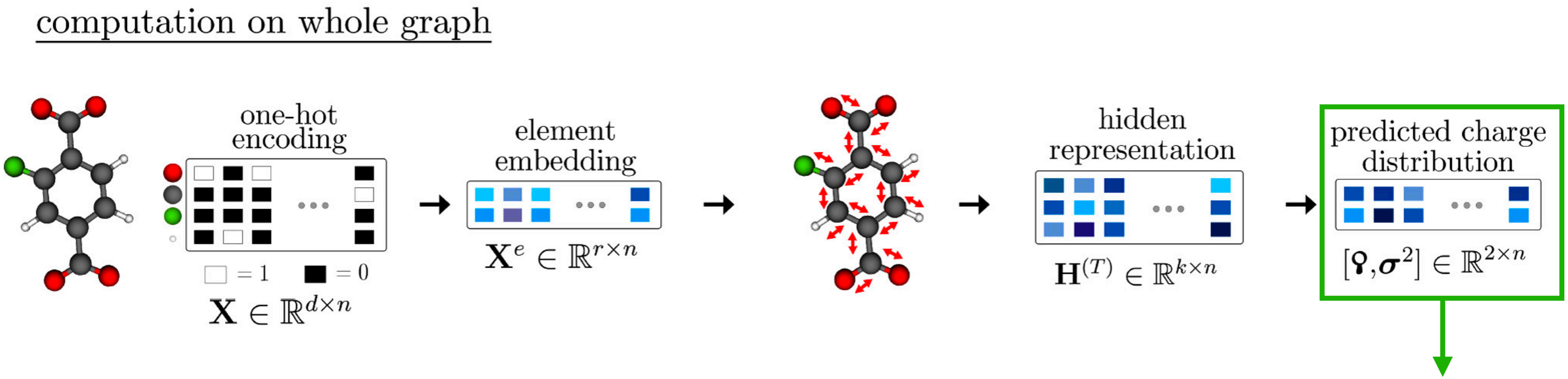
Loss
Weakly Supervised





Partial Charge Assignment to Metal–Organic Frameworks[3]

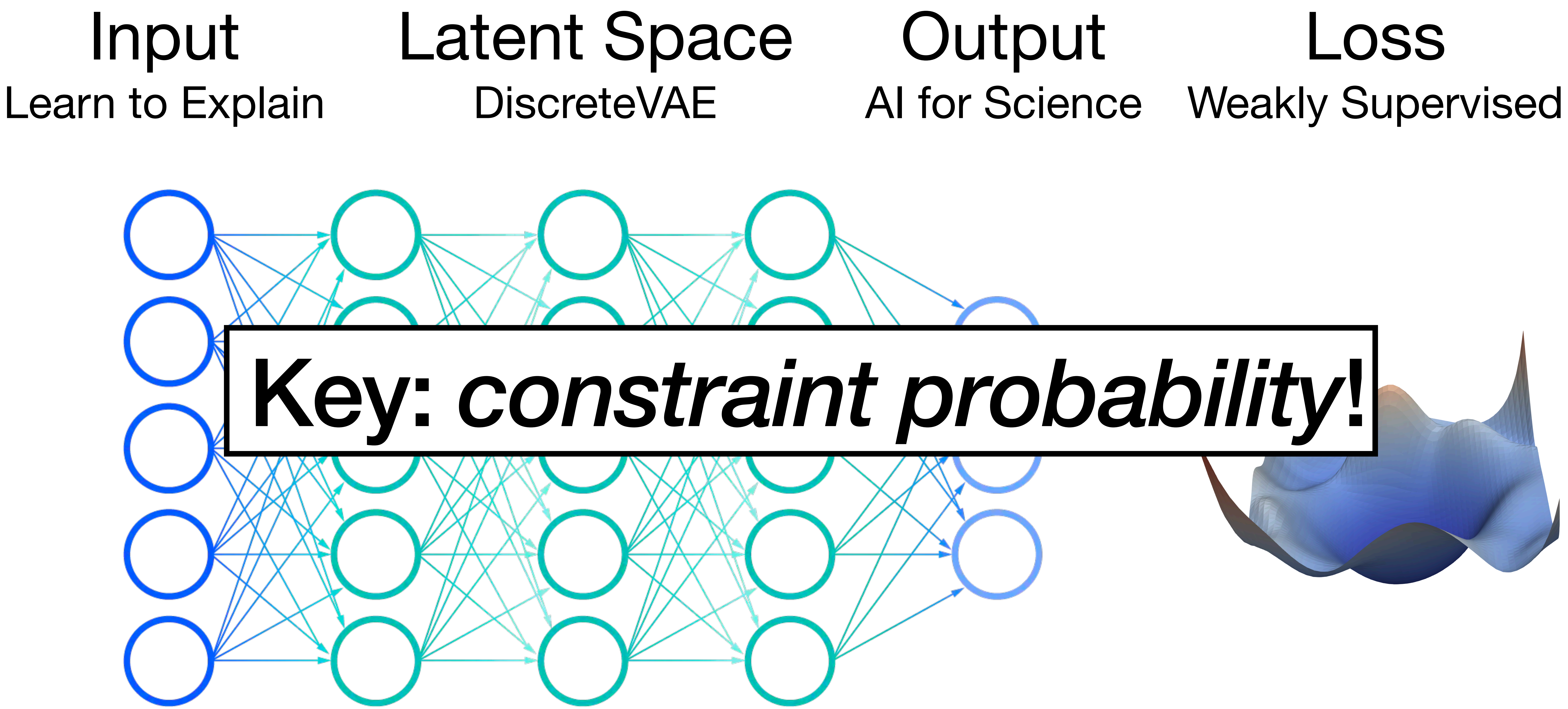
Application in Computational Chemistry



Neutral Charge Constraint

$\sum_i ♀_i = 0$

Differentiable learning under *k-subset* constraints



How to integrate *diverse constraints?*

Outline

- Differentiable learning under constraints
    - Key: constraint probability!
- Constrained probabilistic inference

How to integrate *diverse constraints?*

Outline

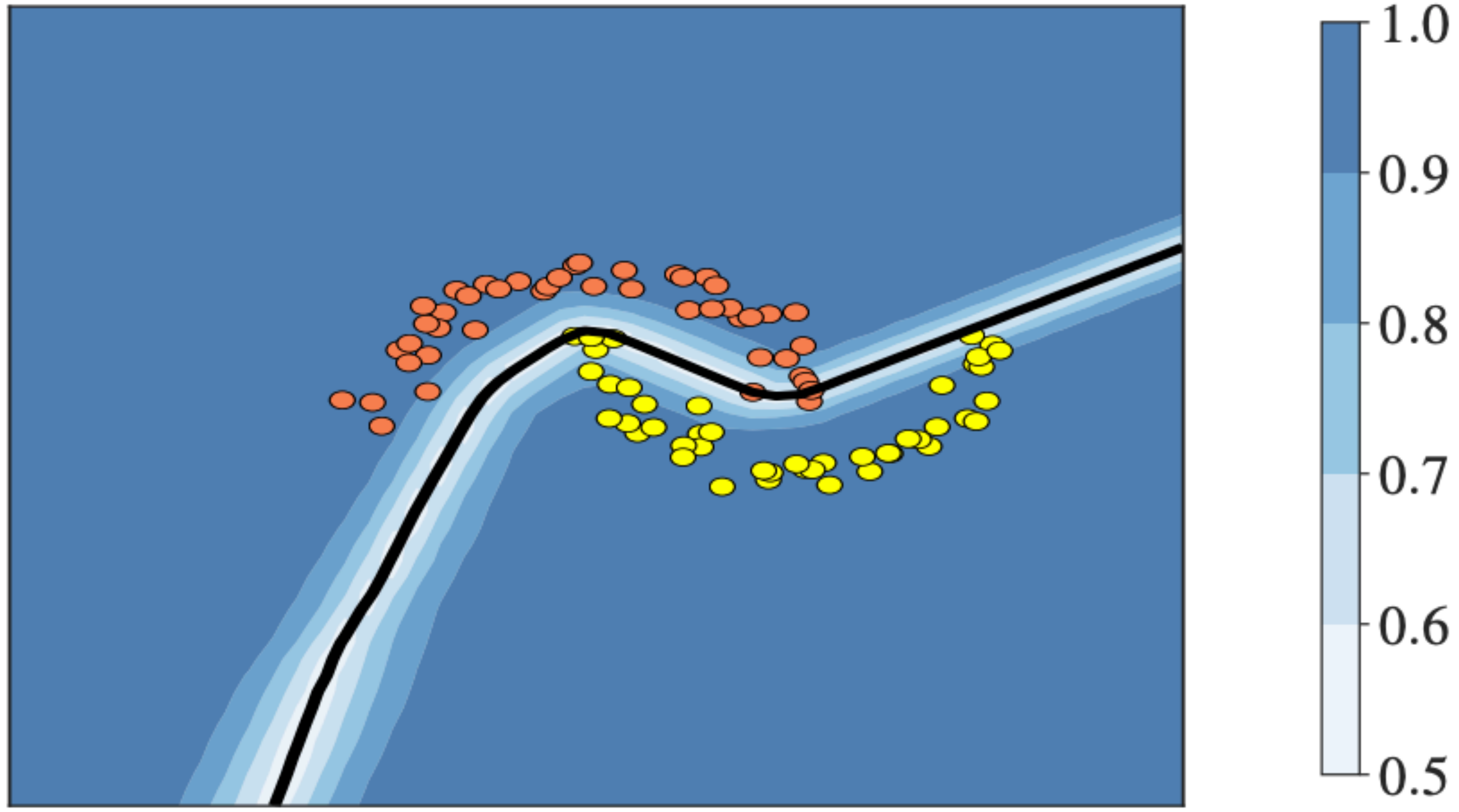
- Differentiable learning under constraints
    - Key: constraint probability!
- Constrained probabilistic inference

Collapsed inference for Bayesian deep learning

Constrained probabilistic inference

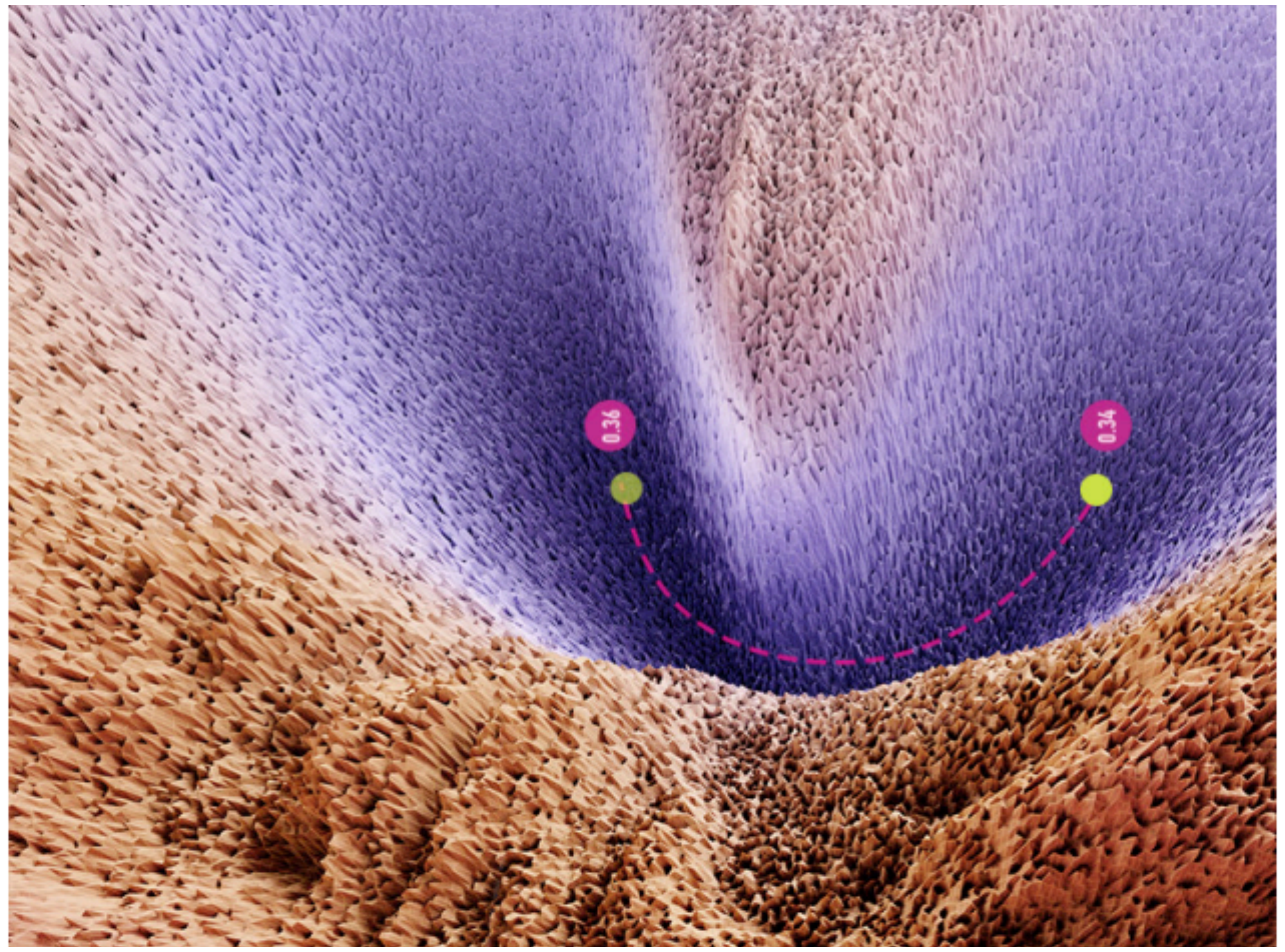
Motivation

Bad Uncertainty Estimation



Confidence by a ReLU neural network [6]

Risky Point Estimation

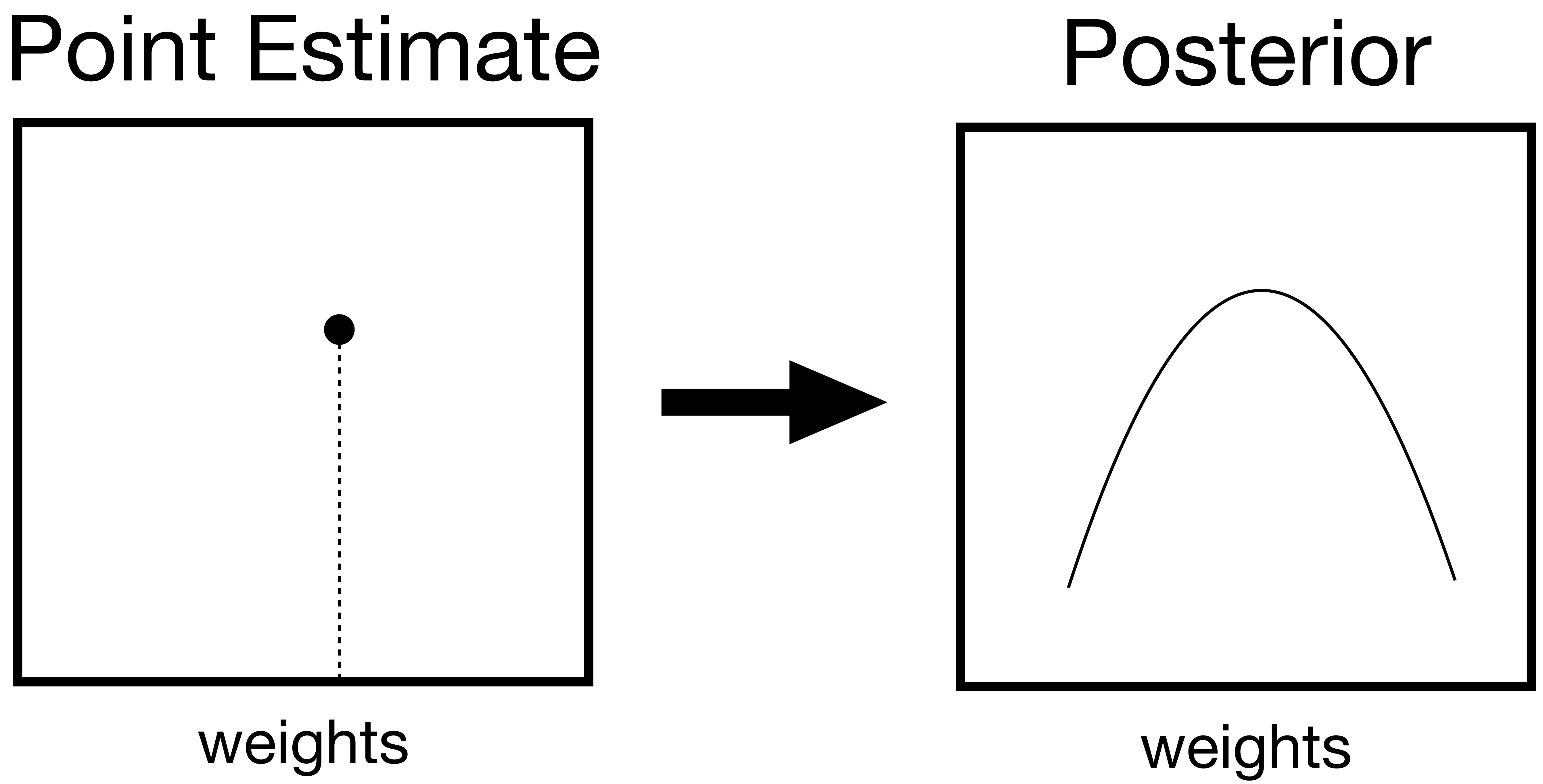


Loss surface [7]

➡ **Bayesian Deep Learning** for *robust* and *reliable* predictions

Bayesian Model Average (BMA)

Key idea



$$p(y \mid x, w) \qquad\qquad p(y \mid x) = \int p(y \mid x, w) p(w) \, dw$$

Motivation

- **Goal:** Bayesian model average

  Predictive posterior $p(y \mid x) = \int p(y \mid x, w) p(w)\, dw$

  Expected prediction $\mathbb{E}[y] = \int y \; p(y \mid x)\, dy$

- **Challenge:** DNNs are too big!

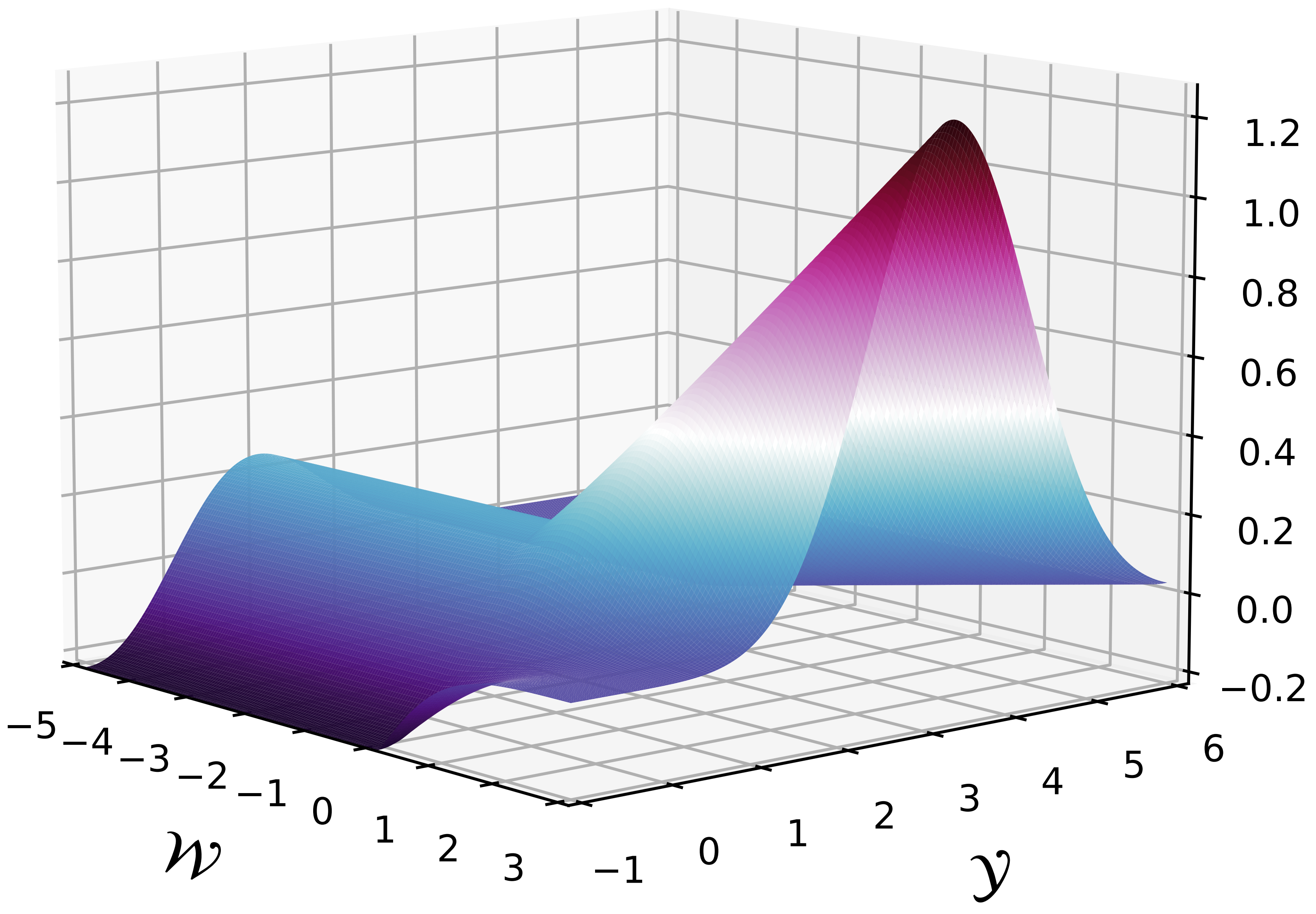
  ➡ *Costly* to maintain too many samples

  ➡ *Low sample efficiency* given the complex integrand

How complex? 🤔

How *complex* is the integrand?

Expected prediction $$\mathbb{E}[y] = \int y \, \underbrace{p(w \mid D)}_{\substack{\text{Weight posterior}\\ \text{1d Uniform}}} \, \underbrace{p(y \mid \underbrace{f(x)}_{\substack{\text{NN model}\\ f(x) = \texttt{ReLU}(wx)}}, w)}_{\substack{\text{Predictive}\\ \text{1d Gaussian}}} \, dw \, dy$$



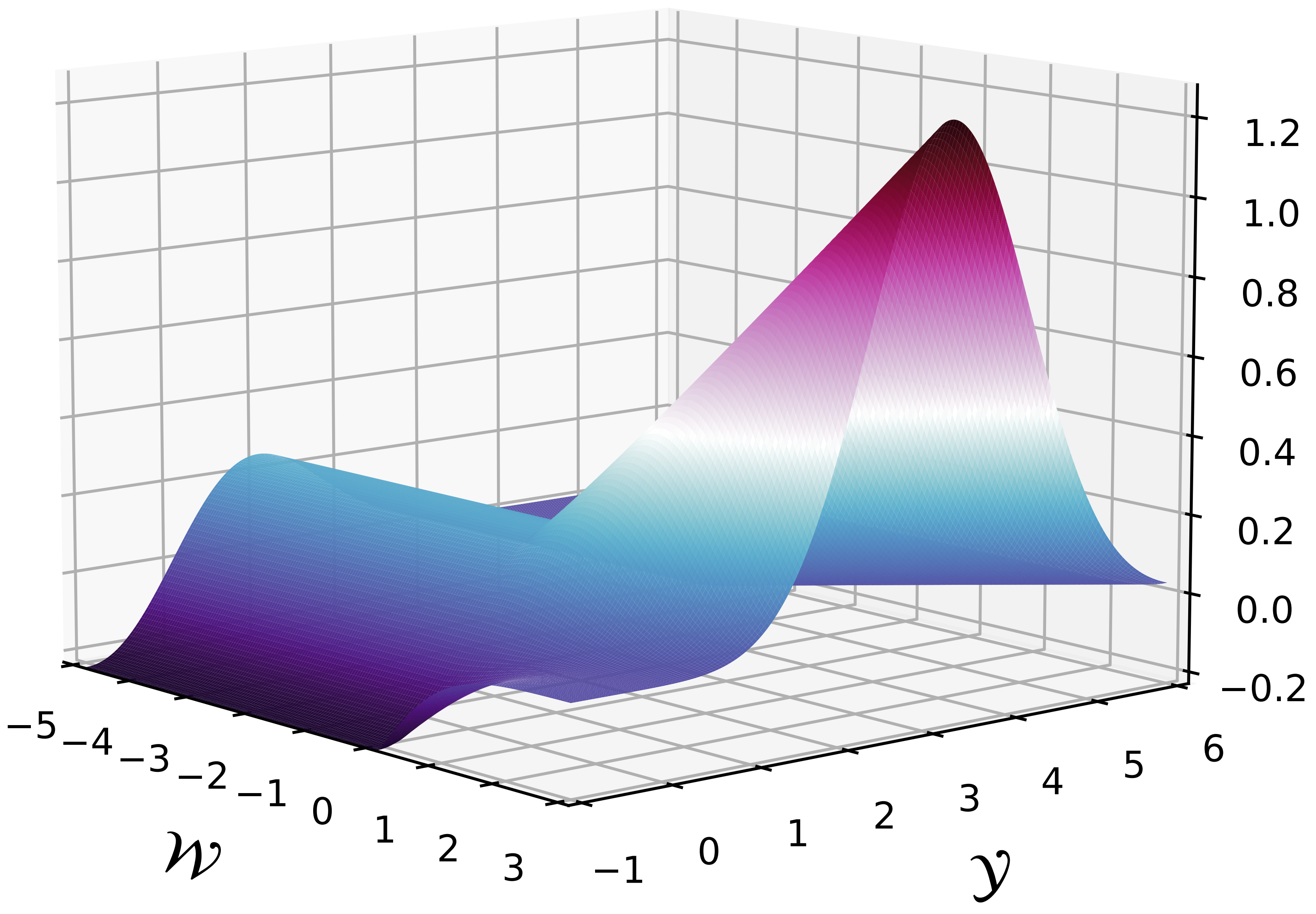
Non-convex, multi-modal, no closed form 🤯

Motivation

- **Goal:** Bayesian model average

Predictive posterior $p(y \mid x) = \int p(y \mid x, w) p(w) \, dw$

Expected prediction $\mathbb{E}[y] = \int y \; p(y \mid x) \, dy$



Is there a *better* way to estimate the integral than *sampling*?

Yes! 🤩

Idea

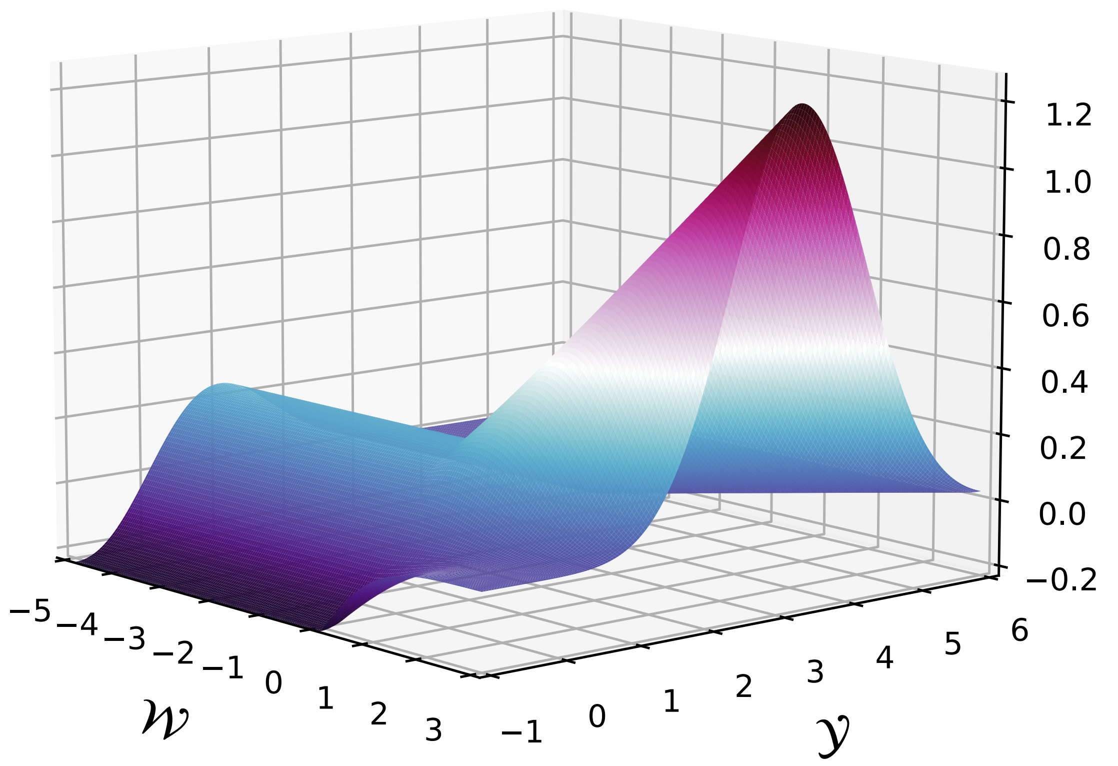
A reduction from BMA to WMI

- Weighted Model Integration (WMI)[4]
  - A class of weighted volume computation problems
  - Definition:
    - *Region*: SMT formula (a logical combination of arithmetic constraints)
    - *Weight function* $\phi : \quad \to \mathbb{R}$
  - Existing WMI solvers are able to give ***exact marginalization*** results
    - for (piecewise) polynomial weights

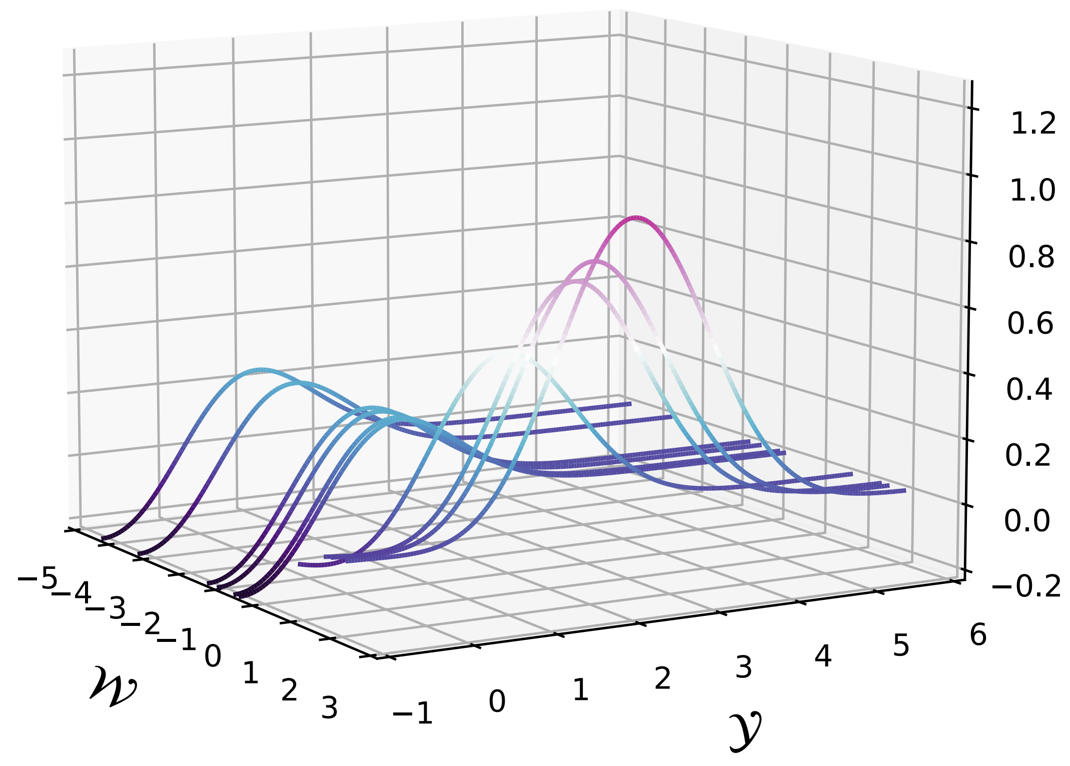
Idea

A reduction from BMA to WMI

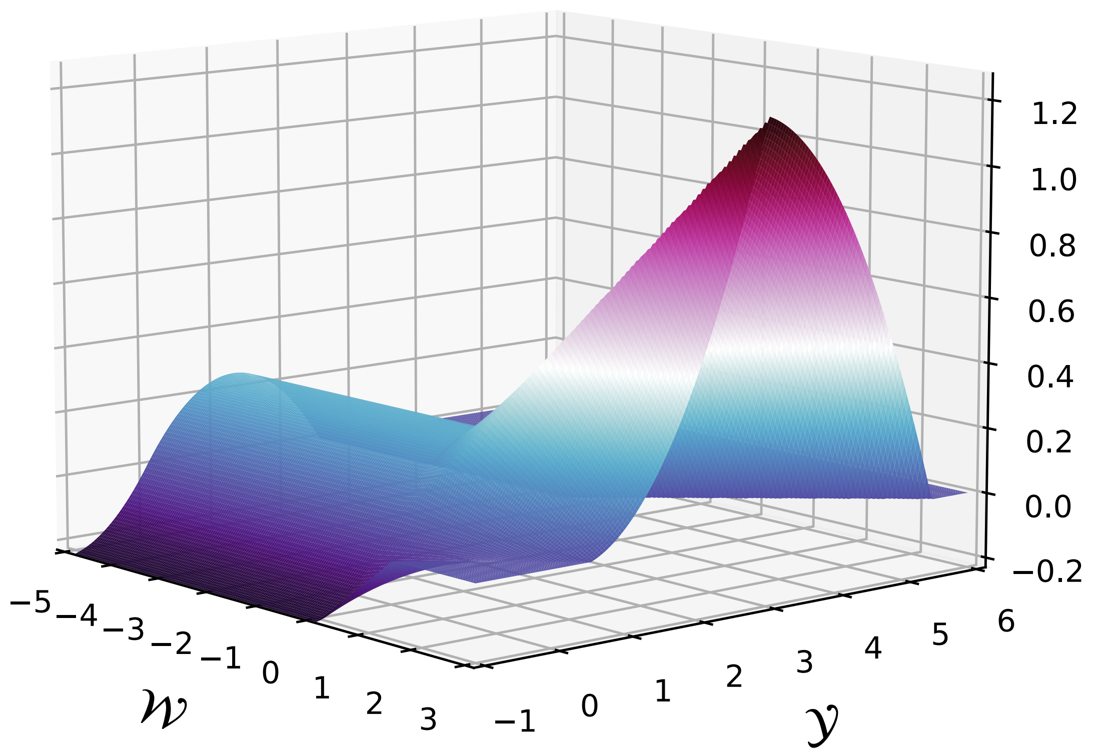
$$\begin{tikzcd}
\text{BMA} \arrow[r, dashed, "\text{Sampling}"] \arrow[d, "\text{Approximated by}"'] & \text{BMA integral} \\
\text{WMI} \arrow[r, "\text{WMI solvers}"] & \text{WMI integral} \arrow[u, "\text{Approximate}"']
\end{tikzcd}$$



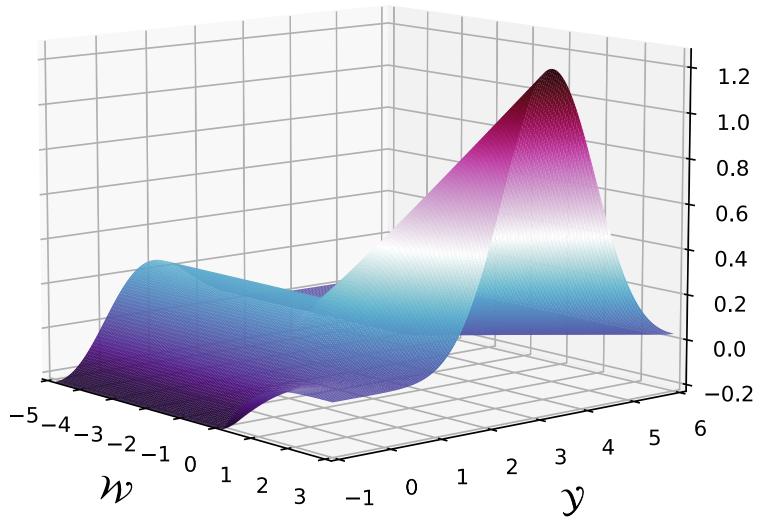
Ground truth BMA



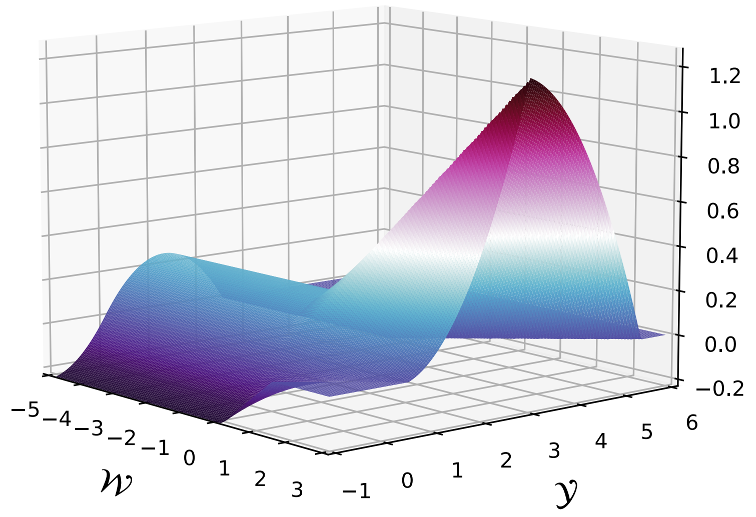
BMA by sampling



BMA by WMI



$\approx$



Accurate approximation!

… but scalability?

Limitations

	Sampling	BMA via WMI
Accuracy	❌	✅
Flexibility	✅	❌*
Scalability	✅	❌**

* Limited to fully connected layers

** Integration over polytopes in arbitrarily high dimensions is #P-hard

How to combine good from both worlds? 🤔

Limitations

	Sampling	BMA via WMI	Collapsed Inference
Accuracy	❌	✅	✅
Flexibility	✅	❌*	✅
Scalability	✅	❌**	✅

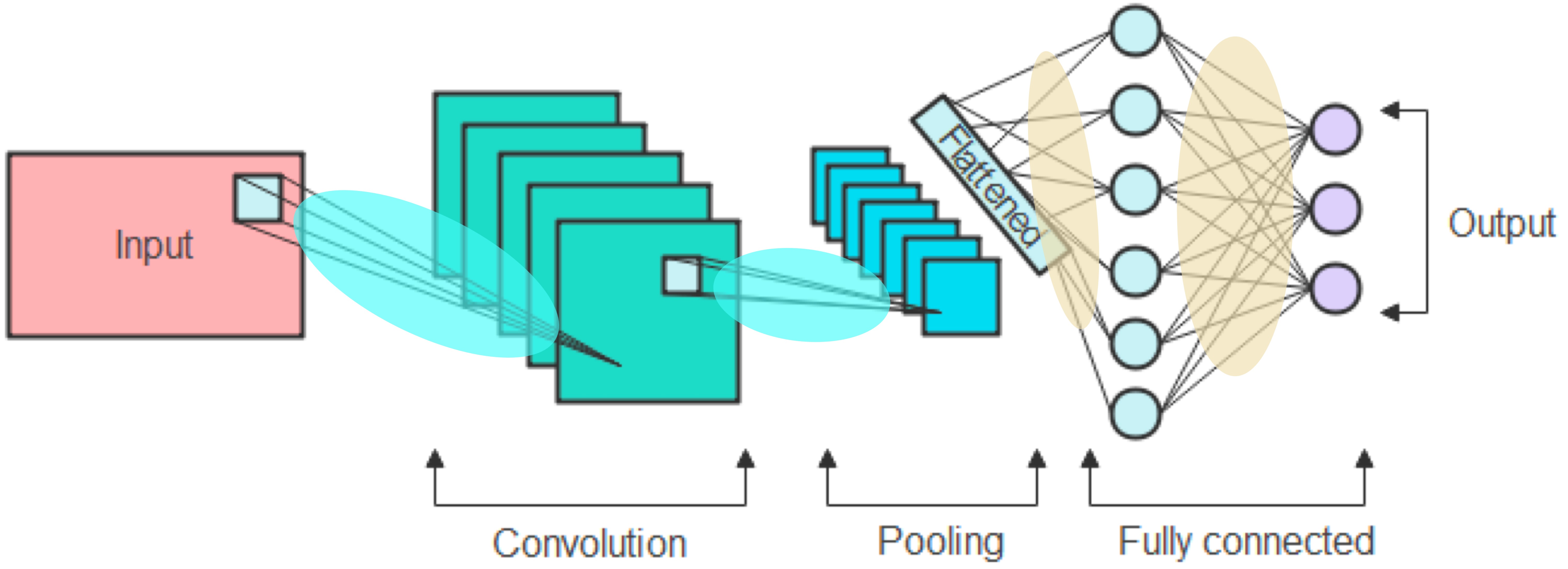
* Limited to fully connected layers

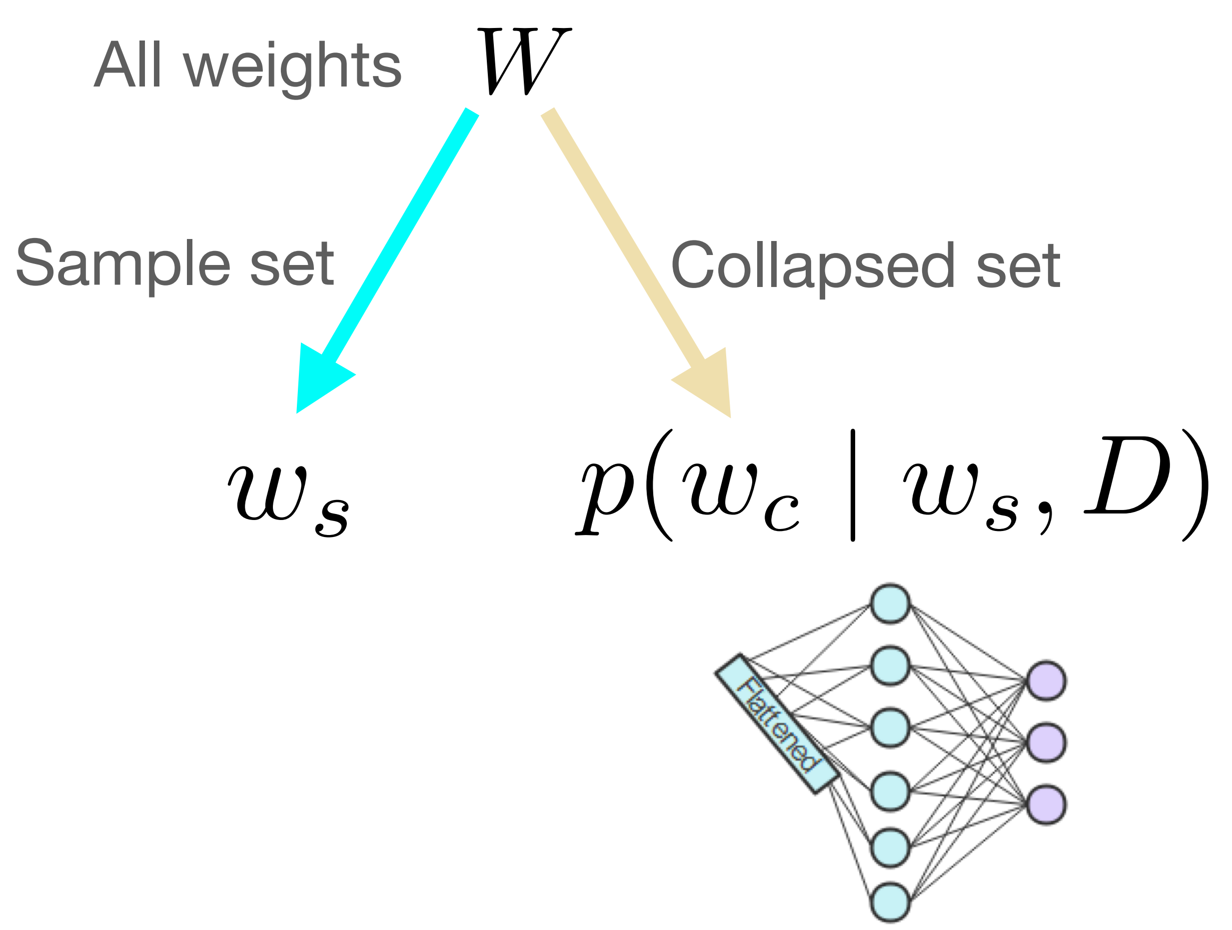
** Integration over polytopes in arbitrarily high dimensions is #P-hard

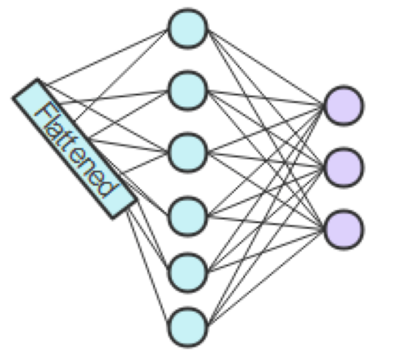
How to combine good from both worlds? 🤔

➡ *Collapsed inference scheme!* 💪

Collapsed Inference[5]





Expected prediction in BMA $\mathbb{E}[y] = \frac{1}{n} \sum_{w_s} \texttt{WMI}($  $)$

Accuracy + Flexibility, Scalability! 🥳

Experiment: UCI Regression

	Boston	Concrete	Yacht	Naval	Energy
CIBER (second)	**-2.471 ± 0.140**	-2.975 ± 0.102	-0.678 ± 0.301	7.276 ± 0.532	**-0.716 ± 0.211**
CIBER (last)	**-2.471 ± 0.140**	**-2.959 ± 0.109**	-0.687 ± 0.301	**7.482 ± 0.188**	**-0.716 ± 0.211**
SWAG	-2.761 ± 0.132	-3.013 ± 0.086	-0.404 ± 0.418	6.708 ± 0.105	-1.679 ± 1.488
PCA+ESS (SI)	-2.719 ± 0.132	-3.007 ± 0.086	**-0.225 ± 0.400**	6.541 ± 0.095	-1.563 ± 1.243
PCA+VI (SI)	-2.716 ± 0.133	-2.994 ± 0.095	-0.396 ± 0.419	6.708 ± 0.105	-1.715 ± 1.588
SGD	-2.752 ± 0.132	-3.178 ± 0.198	-0.418 ± 0.426	6.567 ± 0.185	-1.736 ± 1.613
DVI	-2.410 ± 0.020	-3.060 ± 0.010	-0.470 ± 0.030	6.290 ± 0.040	-1.010 ± 0.060
DGP	-2.330 ± 0.060	-3.130 ± 0.030	-1.390 ± 0.140	3.600 ± 0.330	-1.320 ± 0.030
VI	-2.430 ± 0.030	-3.040 ± 0.020	-1.680 ± 0.040	5.870 ± 0.290	-2.380 ± 0.020
MCD	-2.400 ± 0.040	-2.970 ± 0.020	-1.380 ± 0.010	4.760 ± 0.010	-1.720 ± 0.010
VSD	-2.350 ± 0.050	-2.970 ± 0.020	-1.140 ± 0.020	4.830 ± 0.010	-1.060 ± 0.010

CIBER Wins on 7/11!

	Elevators	KeggD	KeggU	Protein	Skillcraft	Pol
CIBER (second)	-0.378 ± 0.026	**1.245 ± 0.090**	**1.125 ± 0.269**	-0.720 ± 0.036	-1.003 ± 0.035	**2.555 ± 0.115**
CIBER (last)	-0.371 ± 0.023	1.178 ± 0.088	0.964 ± 0.231	-0.720 ± 0.036	**-1.001 ± 0.032**	2.506 ± 0.150
SWAG	-0.374 ± 0.021	1.080 ± 0.035	0.749 ± 0.029	**-0.700 ± 0.051**	-1.180 ± 0.033	1.533 ± 1.084
PCA+ESS (SI)	-0.351 ± 0.030	1.074 ± 0.034	0.752 ± 0.025	-0.734 ± 0.063	-1.181 ± 0.033	-0.185 ± 2.779
PCA+VI (SI)	**-0.325 ± 0.019**	1.085 ± 0.031	0.757 ± 0.028	-0.712 ± 0.057	-1.179 ± 0.033	1.764 ± 0.271
SGD	-0.538 ± 0.108	1.012 ± 0.154	0.602 ± 0.224	-0.854 ± 0.085	-1.162 ± 0.032	1.073 ± 0.858
OrthVGP	-0.448	1.022	0.701	-0.914	—	0.159
NL	-0.698 ± 0.039	0.935 ± 0.265	0.670 ± 0.038	-0.884 ± 0.025	-1.002 ± 0.050	-2.840 ± 0.226

Experiment: Image Classification

Metric	NLL		ACC		ECE	
Dataset	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100
CIBER	$\mathbf{0.1927 \pm 0.0029}$	$\mathbf{0.9193 \pm 0.0027}$	$\mathbf{93.64 \pm 0.09}$	$\mathbf{74.71 \pm 0.18}$	0.0130 ± 0.0011	$\mathbf{0.0168 \pm 0.0025}$
SWAG	0.2503 ± 0.0081	1.2785 ± 0.0031	93.59 ± 0.14	73.85 ± 0.25	0.0391 ± 0.0020	0.1535 ± 0.0015
SGD	0.3285 ± 0.0139	1.7308 ± 0.0137	93.17 ± 0.14	73.15 ± 0.11	0.0483 ± 0.0022	0.1870 ± 0.0014
SWA	0.2621 ± 0.0104	1.2780 ± 0.0051	93.61 ± 0.11	74.30 ± 0.22	0.0408 ± 0.0019	0.1514 ± 0.0032
SGLD	0.2001 ± 0.0059	0.9699 ± 0.0057	93.55 ± 0.15	74.02 ± 0.30	$\mathbf{0.0082 \pm 0.0012}$	0.0424 ± 0.0029
KFAC	0.2252 ± 0.0032	1.1915 ± 0.0199	92.65 ± 0.20	72.38 ± 0.23	0.0094 ± 0.0005	0.0778 ± 0.0054

- achieves accurate estimation of uncertainty
- applicable to large NNs
- boosts predictive performance

How to integrate *diverse constraints?*

Outline

- Differentiable learning under constraints
  - Key: constraint probability!
- Constrained probabilistic inference
  - Key: constraint solvers + statistical ML!

Future Work

How to integrate *diverse constraints?*

- Differentiable learning under constraints
  - Key: constraint probability!
- Constrained probabilistic inference
  - Key: constraint solvers + statistical ML!

What more constraints are tractable?

How to deal with intractable ones? …

What inference amenable to the reduction?

How to deliver reliable & scalable inference? …

Thanks!

References

[1] Kareem Ahmed*,Zhe Zeng*, Mathias Niepert, and Guy Van den Broeck. SIMPLE: A gradient estimator for k-subset sampling, ICLR, 2023.

[2] Vinay Shukla, Zhe Zeng*, Kareem Ahmed*, and Guy Van den Broeck. A unified approach to count-based weakly-supervised learning. In ICML 2023 Workshop on Differentiable Almost Everything, 2023.

[3] Raza, Ali, et al. "Message passing neural networks for partial charge assignment to metal–organic frameworks." *The Journal of Physical Chemistry C* 124.35 (2020): 19070-19082.

[4] V. Belle, A. Passerini, and G. Van den Broeck. Probabilistic inference in hybrid domains by weighted model integration. In Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI), pages 2770–2776, 2015.

[5] Zhe Zeng and Guy Van den Broeck. Collapsed inference for Bayesian deep learning. In ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling, 2023.

[6] Kristiadi, Agustinus, Matthias Hein, and Philipp Hennig. "Being bayesian, even just a bit, fixes overconfidence in relu networks." *International conference on machine learning*. PMLR, 2020.

[7] Garipov, Timur, et al. "Loss surfaces, mode connectivity, and fast ensembling of dnns." *Advances in neural information processing systems* 31 (2018).