

Learning Set Functions Under the Optimal Subset Oracle via Equivariant Variational Inference

Zijing Ou, Tingyang Xu, Qinliang Su, Yingzhen Li, Peilin Zhao, Yatao Bian 

Tencent AI Lab

Sun Yat—sen University

Imperial College London

2022-03-11



Database



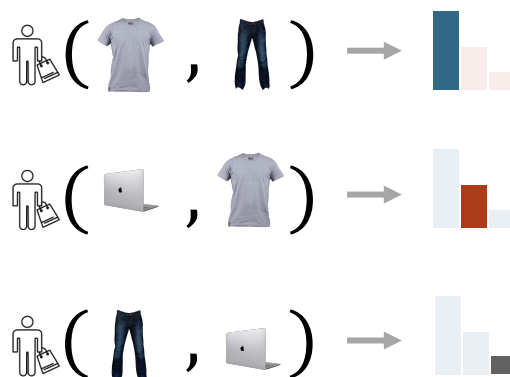
Customer



Shopping cart



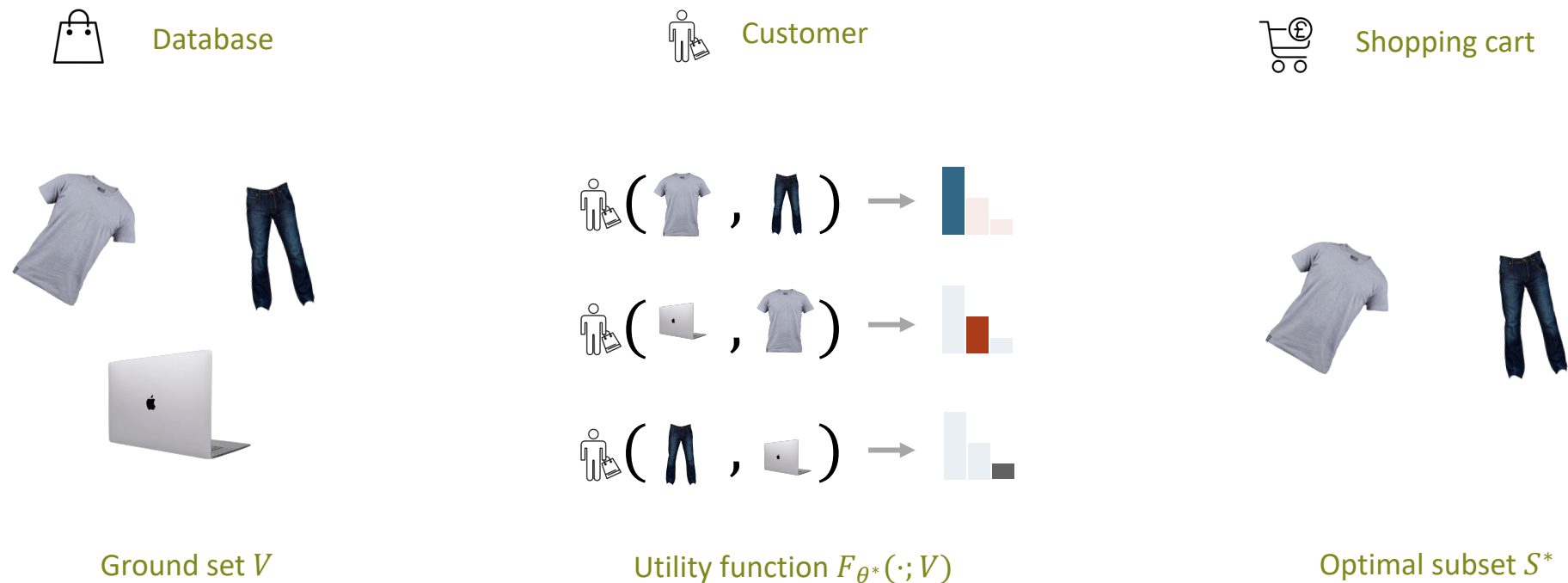
Ground set V



Utility function $F_{\theta^*}(\cdot; V)$



Optimal subset S^*



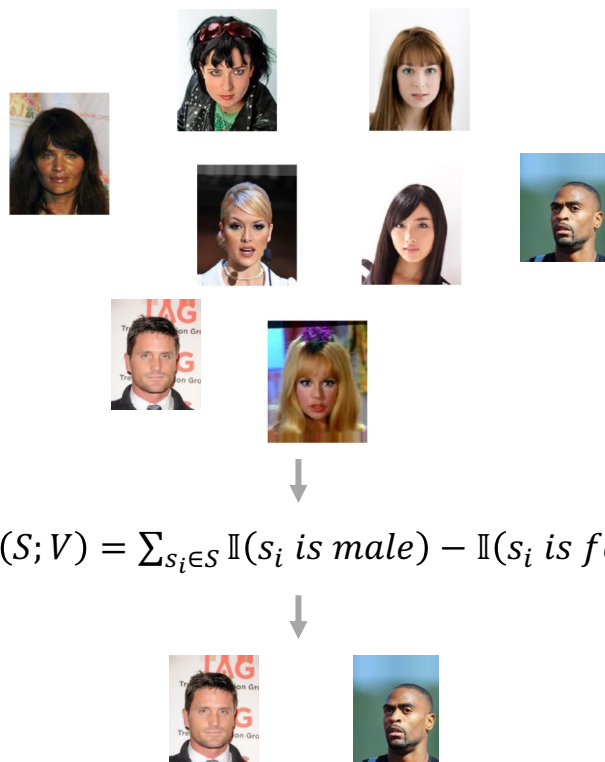
Data Generation Process:

$$S^* = \operatorname{argmax}_{S \in 2^V} F_{\theta^*}(S; V)$$

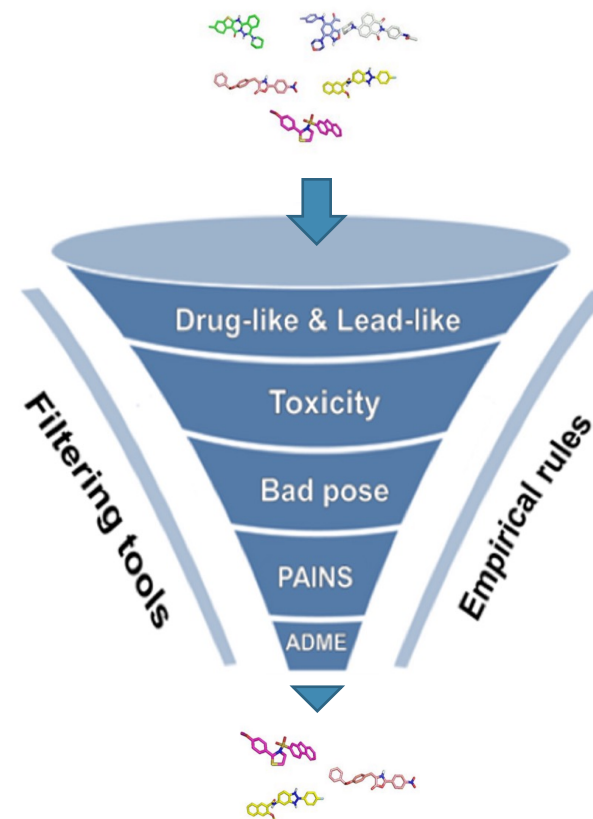
$$\sim \mathbb{p}(S, V) =: \delta_{S=S^*|V}$$

$$S^* = \operatorname{argmax}_{S \in 2^V} F_{\theta^*}(S; V)$$

Set anomaly detection



Compound selection



$$S^* = \operatorname{argmax}_{S \in 2^V} F_{\theta^*}(S; V)$$

Goal: Learn a surrogate F_{θ} to approximate the oracle utility function F_{θ^*} .

$$S^* = \operatorname{argmax}_{S \in 2^V} F_{\theta^*}(S; V)$$

Goal: Learn a surrogate F_{θ} to approximate the oracle utility function F_{θ^*} .

Setting 1, FV oracle:



}}

Matching via empirical risk minimization



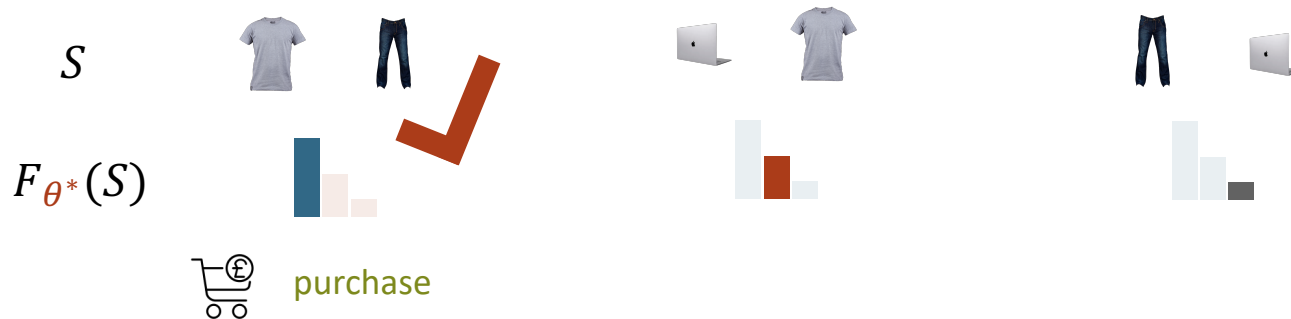
$$\theta^* = \min_{\theta} \sum_i L(F_{\theta}(S_i; V); F_{\theta^*}(S_i; V))$$

$$S^* = \operatorname{argmax}_{S \in 2^V} F_{\theta^*}(S; V)$$

Goal: Learn a surrogate F_{θ} to approximate the oracle utility function F_{θ^*} .

Setting 1, FV oracle:

$$\theta^* = \min_{\theta} \sum_i L(F_{\theta}(S_i; V); F_{\theta^*}(S_i; V))$$



$$S^* = \operatorname{argmax}_{S \in 2^V} F_{\theta^*}(S; V)$$

Goal: Learn a surrogate F_{θ} to approximate the oracle utility function F_{θ^*} .

Setting 1, FV oracle:

$$\theta^* = \min_{\theta} \sum_i L(F_{\theta}(S_i; V); F_{\theta^*}(S_i; V))$$

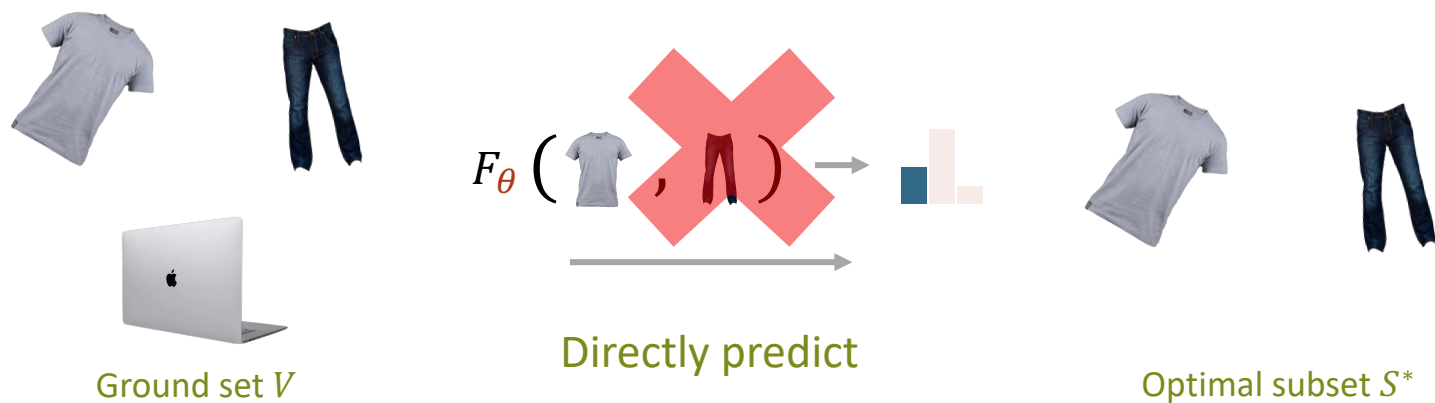
Curse of amounts of supervision signals 🤔

⇒ Training data in form of $\{(S_i, F_{\theta^*}(S_i; V))\}$ for each V

$$S^* = \operatorname{argmax}_{S \in 2^V} F_{\theta^*}(S; V)$$

Goal: Learn a surrogate F_{θ} to approximate the oracle utility function F_{θ^*} .

Setting 2, OS oracle:



\Rightarrow Training data in form of $\{(S^*, V)\}$



$$\begin{aligned} & \text{Empirical distribution} \\ & \downarrow \\ & \operatorname{argmax}_{\theta} \mathbb{E}_{\mathbb{P}(S^*, V)} [\log p_{\theta}(S^* | V)] \\ & s.t. p_{\theta}(S | V) \propto F_{\theta}(S; V), \forall S \in 2^V \\ & \uparrow \\ & \text{Monotonically grows with the utility function} \end{aligned}$$

$$\begin{aligned} & \text{Empirical distribution} \\ & \downarrow \\ & \operatorname{argmax}_{\theta} \mathbb{E}_{\mathbb{P}(S^*, V)} [\log p_{\theta}(S^* | V)] \\ & s.t. p_{\theta}(S | V) \propto F_{\theta}(S; V), \forall S \in 2^V \\ & \uparrow \\ & \text{Monotonically grows with the utility function} \end{aligned}$$

How to construct a proper set mass function $p_{\theta}(S|V)$?

$$\begin{aligned} & \operatorname{argmax}_{\theta} \mathbb{E}_{\mathbb{P}(S^*, V)} [\log p_{\theta}(S^* | V)] \\ & s.t. p_{\theta}(S | V) \propto F_{\theta}(S; V), \forall S \in 2^V \end{aligned}$$

Desiderata:

Permutation invariance

$$F_{\theta} \left(\text{👕}, \text{👖} \right) = F_{\theta} \left(\text{👖}, \text{👕} \right)$$

$$\begin{aligned} & \operatorname{argmax}_{\theta} \mathbb{E}_{\mathbb{P}(S^*, V)} [\log p_{\theta}(S^* | V)] \\ & s.t. p_{\theta}(S | V) \propto F_{\theta}(S; V), \forall S \in 2^V \end{aligned}$$

Desiderata:

Permutation invariance

$$F_{\theta}(\text{👕}, \text{👖}) = F_{\theta}(\text{👖}, \text{👕})$$

Varying ground set

$$F_{\theta}(\text{👕}; \text{👕} \text{ 👖} \text{ 📱}) \rightarrow \text{■} \qquad F_{\theta}(\text{👕}; \text{👕} \text{ 👖} \text{ 📱} \text{ 📱}) \rightarrow \text{■}$$

$$\begin{aligned} & \operatorname{argmax}_{\theta} \mathbb{E}_{\mathbb{P}(S^*, V)} [\log p_{\theta}(S^* | V)] \\ & s.t. p_{\theta}(S | V) \propto F_{\theta}(S; V), \forall S \in 2^V \end{aligned}$$

Desiderata:

Permutation invariance

$$F_{\theta}(\text{👕}, \text{👖}) = F_{\theta}(\text{👖}, \text{👕})$$

Varying ground set

$$F_{\theta}(\text{👕}; \text{👕} \text{👖} \text{💻}) \rightarrow \text{■} \qquad F_{\theta}(\text{👕}; \text{👕} \text{👖} \text{💻} \text{📱}) \rightarrow \text{■}$$

Differentiability; Minimum prior & Scalability

$$\operatorname{argmax}_{\theta} \mathbb{E}_{\mathbb{P}(S^*, V)} [\log p_{\theta}(S^* | V)]$$

$$s.t. p_{\theta}(S | V) \propto F_{\theta}(S; V), \forall S \in 2^V$$

Probabilistic greedy model (PGM):

$$p_{\theta}(S | V) = \sum_{\pi \in \Pi^S} p_{\theta}(\pi | V)$$

Permutation $\pi = \{s_1, s_2, \dots, s_{|S|}\}$
↓

$$p_{\theta}(\pi | V) = \prod_{j=0}^{|S|-1} \frac{\exp(F_{\theta}(s_{j+1} \cup \mathbf{s}_j) / \gamma)}{\sum_{s \in V \setminus S_j} \exp(F_{\theta}(s \cup \mathbf{s}_j) / \gamma)}$$

The first j chosen elements: $S_j = \{s_1, s_2, \dots, s_j\}$
↓

$$\begin{aligned} & \operatorname{argmax}_{\theta} \mathbb{E}_{\mathbb{P}(S^*, V)} [\log p_{\theta}(S^* | V)] \\ & s.t. p_{\theta}(S | V) \propto F_{\theta}(S; V), \forall S \in 2^V \end{aligned}$$

Probabilistic greedy model (PGM):

$$\begin{aligned} & \text{Permutation } \pi = \{s_1, s_2, \dots, s_{|S|}\} \\ & \downarrow \\ & p_{\theta}(S | V) = \sum_{\pi \in \Pi^S} p_{\theta}(\pi | V) \\ & \text{The first } j \text{ chosen elements: } S_j = \{s_1, s_2, \dots, s_j\} \\ & \downarrow \\ & p_{\theta}(\pi | V) = \prod_{j=0}^{|S|-1} \frac{\exp(F_{\theta}(s_{j+1} \cup S_j) / \gamma)}{\sum_{s \in V \setminus S_j} \exp(F_{\theta}(s \cup S_j) / \gamma)} \end{aligned}$$

Training:

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{(V, S^*)} \log \sum_{\pi \in \Pi^{S^*}} p_{\theta}(\pi | V)$$

Log-likelihood: approximate with sampling

DSF:

$$F_{\theta}(S; V) = \sum_i \overset{\text{Concave function } \rho_i: \mathbb{R}_+ \rightarrow \mathbb{R}_+}{\rho_i} \left(\sum_{s \in S} \overset{\text{Transform } \kappa_i: V \rightarrow \mathbb{R}_+}{\kappa_i}(s) \right)$$

PGM:

$$p_{\theta}(S|V) = \sum_{\pi \in \Pi^S} p_{\theta}(\pi|V)$$

$$p_{\theta}(\pi|V) = \prod_{j=0}^{|S|-1} \frac{\exp(F_{\theta}(s_{j+1} \cup S_j)/\gamma)}{\sum_{s \in V \setminus S_j} \exp(F_{\theta}(s \cup S_j)/\gamma)}$$

	Permutation Invariance	Varying Ground Set	Differentiability	Minimum Prior	Scalability	FV/OS Oracle
DSF	✓	✗	✓	-	-	FV
PGM	✓	✗	✓	✗	✗	OS
EquiVSet (Ours)	✓	✓	✓	✓	✓	OS

$$p_{\theta}(S|V) = \frac{\exp(F_{\theta}(S; V))}{Z}$$

$Z \leftarrow$ Partition function $Z := \sum_{S \subseteq V} \exp(F_{\theta}(S; V))$

$$p_{\theta}(S|V) = \frac{\exp(F_{\theta}(S; V))}{Z}$$

$Z \leftarrow$ Partition function $Z := \sum_{S \subseteq V} \exp(F_{\theta}(S; V))$

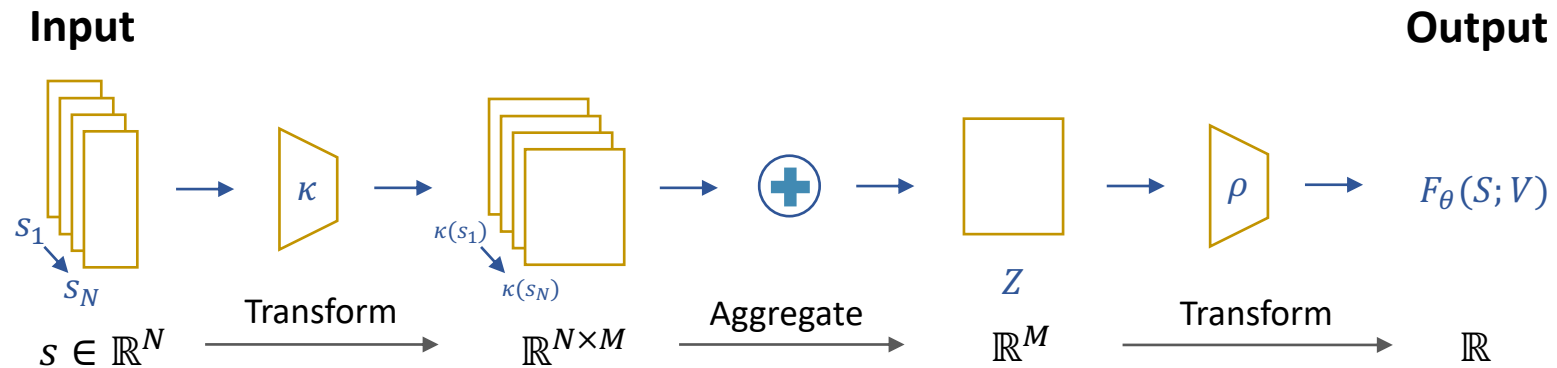
EBMs for Minimum Prior:

Energy-based modeling has **maximum entropy**

$$p_{\theta}(S|V) = \frac{\exp(F_{\theta}(S; V))}{Z}$$

$Z \leftarrow$ Partition function $Z := \sum_{S \subseteq V} \exp(F_{\theta}(S; V))$

DeepSet for Permutation Invariance:



$$p_{\theta}(S|V) = \frac{\exp(F_{\theta}(S; V))}{Z} \quad Z \leftarrow \text{Partition function } Z := \sum_{S \subseteq V} \exp(F_{\theta}(S; V))$$

Training Discrete EBM:

Contrastive Divergence \Rightarrow Hard to converge

Score Matching \Rightarrow NonDifferentiable

Ratio Matching \Rightarrow Unstable

Separate training and inference procedure



$$p_{\theta}(S|V) = \frac{\exp(F_{\theta}(S; V))}{Z} \quad \leftarrow \text{Partition function } Z := \sum_{S \subseteq V} \exp(F_{\theta}(S; V))$$

Marginal-based Loss:

$\psi \in [0,1]^{|V|}$: odds that $s \in V$ shall be selected in the OS S^*

\downarrow

$$\psi^* = \underset{\psi}{\operatorname{argmax}} D(q(S; \psi) \| p_{\theta}(S))$$
$$L(\theta; \psi^*) = \sum_{i=1}^N [-\sum_{j \in S_i^*} \log \psi_j^* - \sum_{j \in V_i \setminus S_i^*} \log(1 - \psi_j^*)]$$

Cohesive training and inference procedure 😁

$$p_{\theta}(S|V) = \frac{\exp(F_{\theta}(S; V))}{Z} \quad \leftarrow \text{Partition function } Z := \sum_{S \subseteq V} \exp(F_{\theta}(S; V))$$

Marginal-based Loss:

$\psi \in [0,1]^{|V|}$: odds that $s \in V$ shall be selected in the OS S^*

$$\psi^* = \underset{\psi}{\operatorname{argmax}} D(q(S; \psi) \| p_{\theta}(S))$$

$$L(\theta; \psi^*) = \sum_{i=1}^N [-\sum_{j \in S_i^*} \log \psi_j^* - \sum_{j \in V_i \setminus S_i^*} \log(1 - \psi_j^*)]$$

↑

Require ψ^* is differentiable w.r.t. θ



Variational distribution $q(S; \psi) = \prod_{i \in S} \psi_i \prod_{j \notin S} (1 - \psi_j), \psi \in [0, 1]^{|V|}$



$$\min_{\psi} \text{KL}(q(S; \psi) \| p_{\theta}(S))$$

$$\Leftrightarrow \max_{\psi} f_{mt}^{F_{\theta}}(\psi) + \mathbb{H}(q(S; \psi)) =: \text{ELBO}$$



multilinear extension $f_{mt}^{F_{\theta}}(\psi) := \sum_{S \subseteq V} F_{\theta}(S) \prod_{i \in S} \psi_i \prod_{j \notin S} (1 - \psi_j)$

Variational distribution $q(S; \psi) = \prod_{i \in S} \psi_i \prod_{j \notin S} (1 - \psi_j), \psi \in [0, 1]^{|V|}$



$$\min_{\psi} \text{KL}(q(S; \psi) \| p_{\theta}(S))$$

$$\Leftrightarrow \max_{\psi} f_{mt}^{F_{\theta}}(\psi) + \mathbb{H}(q(S; \psi)) =: \text{ELBO}$$



multilinear extension $f_{mt}^{F_{\theta}}(\psi) := \sum_{S \subseteq V} F_{\theta}(S) \prod_{i \in S} \psi_i \prod_{j \notin S} (1 - \psi_j)$

RNN-like fixed-point iteration:

$$\psi^{(0)} \leftarrow \text{Initialize in } [0, 1]^{|V|}$$

$$\psi_{\theta}^{(k)} \leftarrow \left(1 + \exp \left(-\nabla_{\psi^{(k-1)}} f_{mt}^{F_{\theta}}(\psi^{(k-1)}) \right) \right)^{-1} \quad \left. \vphantom{\psi_{\theta}^{(k)}} \right\} \text{MFVI}(\psi^{(0)}, V, K)$$

$$\psi_{\theta}^* \leftarrow \psi_{\theta}^{(K)}$$

$\psi_{\theta}^* = \text{MFVI}(\psi^{(0)}, V, K)$ is **differentiable** w.r.t. θ 😊

$$L(\theta; \psi^*) = \sum_{i=1}^N [-\sum_{j \in S_i^*} \log \psi_j^* - \sum_{j \in V_i \setminus S_i^*} \log(1 - \psi_j^*)]$$

Algorithm DiffMF(V, S^*):

Initialize variational parameter ψ

$$\psi^{(0)} \leftarrow 0.5 * \mathbf{1}$$

Compute the variational marginals

$$\psi^* \leftarrow \text{MFVI}(\psi^{(0)}, V, K)$$

Update parameter θ

$$\theta \leftarrow \theta - \eta \nabla_{\theta} L(\theta; \psi^*)$$

$$L(\theta; \psi^*) = \sum_{i=1}^N [-\sum_{j \in S_i^*} \log \psi_j^* - \sum_{j \in V_i \setminus S_i^*} \log(1 - \psi_j^*)]$$

Training:

Initialize variational parameter ψ

$$\psi^{(0)} \leftarrow 0.5 * \mathbf{1}$$

Compute the variational marginals

$$\psi^* \leftarrow \text{MFVI}(\psi^{(0)}, V, K)$$

Update parameter θ

$$\theta \leftarrow \theta - \eta \nabla_{\theta} L(\theta; \psi^*)$$

Inference:

$$S = \text{topN}(\psi^*)$$

$$\psi^* = \text{MFVI}(\psi^{(0)}, V, K)$$

- Expensive computation complexity

$$\psi_{\theta}^{(k)} \leftarrow \left(1 + \exp \left(-\nabla_{\psi^{(k-1)}} f_{mt}^{F_{\theta}}(\psi^{(k-1)}) \right) \right)^{-1}$$

↑
expensive sampling loop per data point

- Discard interaction pattern

$$q(S; \psi) = \prod_{i \in S} \psi_i \prod_{j \notin S} (1 - \psi_j), \psi \in [0, 1]^{|V|}$$

↑
Independent assumption

$$\psi_{\theta}^{(k)} \leftarrow \left(1 + \exp \left(-\nabla_{\psi^{(k-1)}} f_{mt}^{F_{\theta}}(\psi^{(k-1)}) \right) \right)^{-1}$$

↑
expensive sampling loop per data point

Approximate with neural networks:

neural network $q_{\phi}: \mathbb{R}^{|V|} \rightarrow \mathbb{R}^{|V|}$
↓

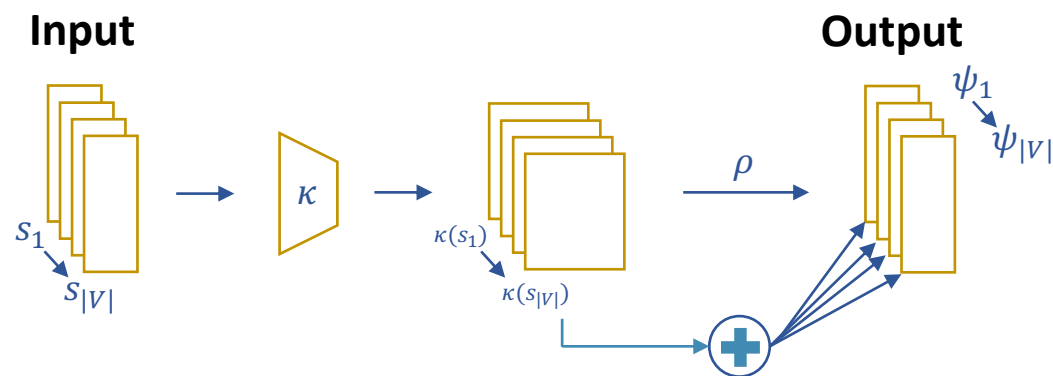
$$\begin{aligned} L &= \text{KL}(q_{\phi}(S; \psi) \| p_{\theta}(S)) \\ &= f_{mt}^{F_{\theta}}(\psi) + \mathbb{H}(q_{\phi}(S; \psi)) + \text{const} \end{aligned}$$

$q_{\phi}(S; \psi)$ should satisfy **equivariant**

$$\psi_{\theta}^{(k)} \leftarrow \left(1 + \exp \left(-\nabla_{\psi^{(k-1)}} f_{mt}^{F_{\theta}}(\psi^{(k-1)}) \right) \right)^{-1}$$

↑
expensive sampling loop per data point

EquiNet($V; \phi$) with permutation equivariance



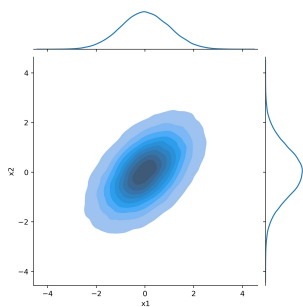
$$q(S; \psi) = \prod_{i \in S} \psi_i \prod_{j \notin S} (1 - \psi_j), \psi \in [0, 1]^{|V|}$$



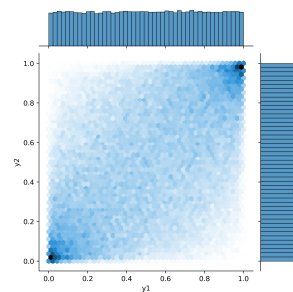
Independent assumption

Correlation-aware inference:

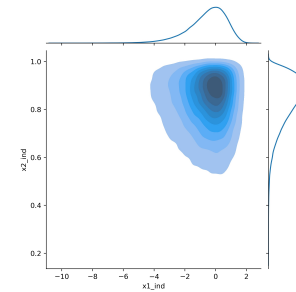
$$q(s_1, s_2, \dots, s_{|V|}) = \underset{\substack{\text{Copula density} \\ \downarrow}}{c} \left(\underset{\substack{\text{Cumulative distribution function} \\ \uparrow}}{Q_1(s_1)}, Q_2(s_2), \dots, Q_{|V|}(s_{|V|}) \right) \prod_{i=1}^{|V|} \underset{\substack{\text{Marginal distribution} \\ \downarrow}}{q_i(s_i)}$$



$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

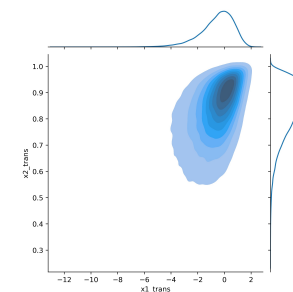


$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim c(Q_1(x_1), Q_2(x_2))$$



$$\begin{aligned} x_{1trans} &\sim \text{Gumbel} \\ x_{2trans} &\sim \text{Beta} \end{aligned}$$

Copula



$$q(s_1, s_2, \dots, s_{|V|}) = \underset{\uparrow}{\mathbf{c}} \left(Q_1(s_1), Q_2(s_2), \dots, Q_{|V|}(s_{|V|}) \right) \prod_{i=1}^{|V|} q_i(s_i)$$

Apply Gaussian copula here

Induce Gaussian copula:

Sample an auxiliary noise

$$g \sim N(0, \Sigma) \rightarrow \text{Covariance matrix, parameterized by neural network}$$

Apply element-wise Gaussian CDF

$$u = \Phi_{\text{diag}(\Sigma)}(g)$$

Obtain binary sample

$$s = \mathbb{I}(\psi \leq u) \rightarrow \text{Original logits of Bernoulli}$$

$$\phi^* = \operatorname{argmax}_{\phi} f_{mt}^{F_{\theta}}(\psi) + \mathbb{H}(q_{\phi}(S; \psi)) := \text{ELBO}$$

$$\theta^* = \operatorname{argmin}_{\theta} \sum_{i=1}^N [-\sum_{j \in S_i^*} \log \psi_j^* - \sum_{j \in V_i \setminus S_i^*} \log(1 - \psi_j^*)]$$

\uparrow
 $\psi^* = \text{MFVI}(\psi^{(0)}, V, K)$

Algorithm EquiVSet(V, S^*):

Update parameter ϕ

$$\phi \leftarrow \phi + \eta \nabla_{\phi} \text{ELBO}(\phi) \quad \rightarrow \text{Optimize } \phi$$

Initialize variational parameter

$$\psi^{(0)} \leftarrow \text{EquiNet}(V; \phi)$$

One step fixed point iteration

$$\psi^* \leftarrow \text{MFVI}(\psi^{(0)}, V, K = 1)$$

} Mean-field
Iteration

Update parameter θ

$$\theta \leftarrow \theta - \eta \nabla_{\theta} L(\theta; \psi^*) \quad \rightarrow \text{Optimize } \theta$$

EquiVSet is a cooperative training framework

$$\theta^* = \operatorname{argmin}_{\theta} \sum_{i=1}^N [-\sum_{j \in S_i^*} \log \psi_j^* - \sum_{j \in V_i \setminus S_i^*} \log(1 - \psi_j^*)]$$

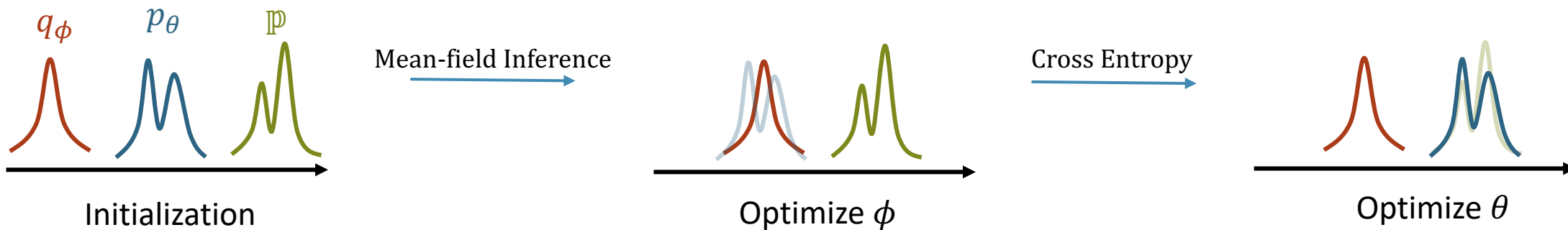
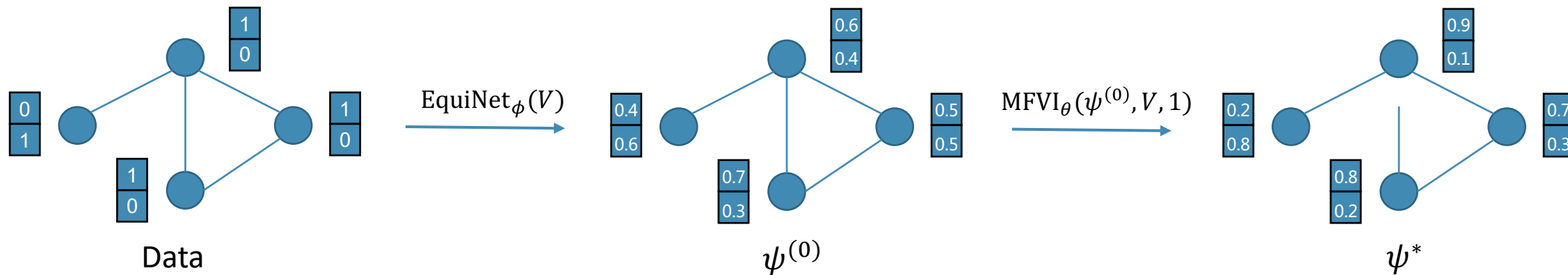




Table 2: Product recommendation results in the MJC metric on the Amazon dataset.

Categories	Random	PGM	DeepSet (NoSetFn)	DiffMF (ours)	EquiVSet _{ind} (ours)	EquiVSet _{copula} (ours)
Toys	0.0832	0.4414 ± 0.0036	0.4287 ± 0.0047	0.6147 ± 0.0102	0.6491 ± 0.0152	0.6762 ± 0.0221
Furniture	0.0651	0.1746 ± 0.0069	0.1758 ± 0.0072	0.1744 ± 0.0121	0.1775 ± 0.0108	0.1724 ± 0.0091
Gear	0.0771	0.4712 ± 0.0037	0.3806 ± 0.0019	0.5622 ± 0.0171	0.6103 ± 0.0193	0.6973 ± 0.0119
Carseats	0.0659	0.2330 ± 0.0115	0.2121 ± 0.0096	0.2229 ± 0.0104	0.2141 ± 0.0073	0.2149 ± 0.0123
Bath	0.0763	0.5638 ± 0.0077	0.4241 ± 0.0058	0.6901 ± 0.0061	0.6457 ± 0.0200	0.7567 ± 0.0095
Health	0.0758	0.4493 ± 0.0024	0.4481 ± 0.0041	0.5650 ± 0.0092	0.6315 ± 0.0153	0.7003 ± 0.0159
Diaper	0.0839	0.5802 ± 0.0092	0.4572 ± 0.0050	0.7011 ± 0.0112	0.7344 ± 0.0199	0.8275 ± 0.0136
Bedding	0.0791	0.4799 ± 0.0061	0.4824 ± 0.0081	0.6408 ± 0.0093	0.6287 ± 0.0195	0.7688 ± 0.0121
Safety	0.0648	0.2495 ± 0.0060	0.2211 ± 0.0044	0.2007 ± 0.0527	0.2250 ± 0.0287	0.2524 ± 0.0285
Feeding	0.0925	0.5596 ± 0.0081	0.4295 ± 0.0021	0.7496 ± 0.0114	0.6955 ± 0.0063	0.8101 ± 0.0074
Apparel	0.0918	0.5333 ± 0.0050	0.5074 ± 0.0036	0.6708 ± 0.0225	0.6465 ± 0.0150	0.7521 ± 0.0114
Media	0.0944	0.4406 ± 0.0092	0.4241 ± 0.0105	0.5145 ± 0.0105	0.5506 ± 0.0072	0.5694 ± 0.0105

$$\text{Metric: MJC} := \frac{1}{|\mathcal{D}_t|} \sum_{(V, S^*) \in \mathcal{D}_t} \frac{|S^* \cap S|}{|S^* \cup S|}$$

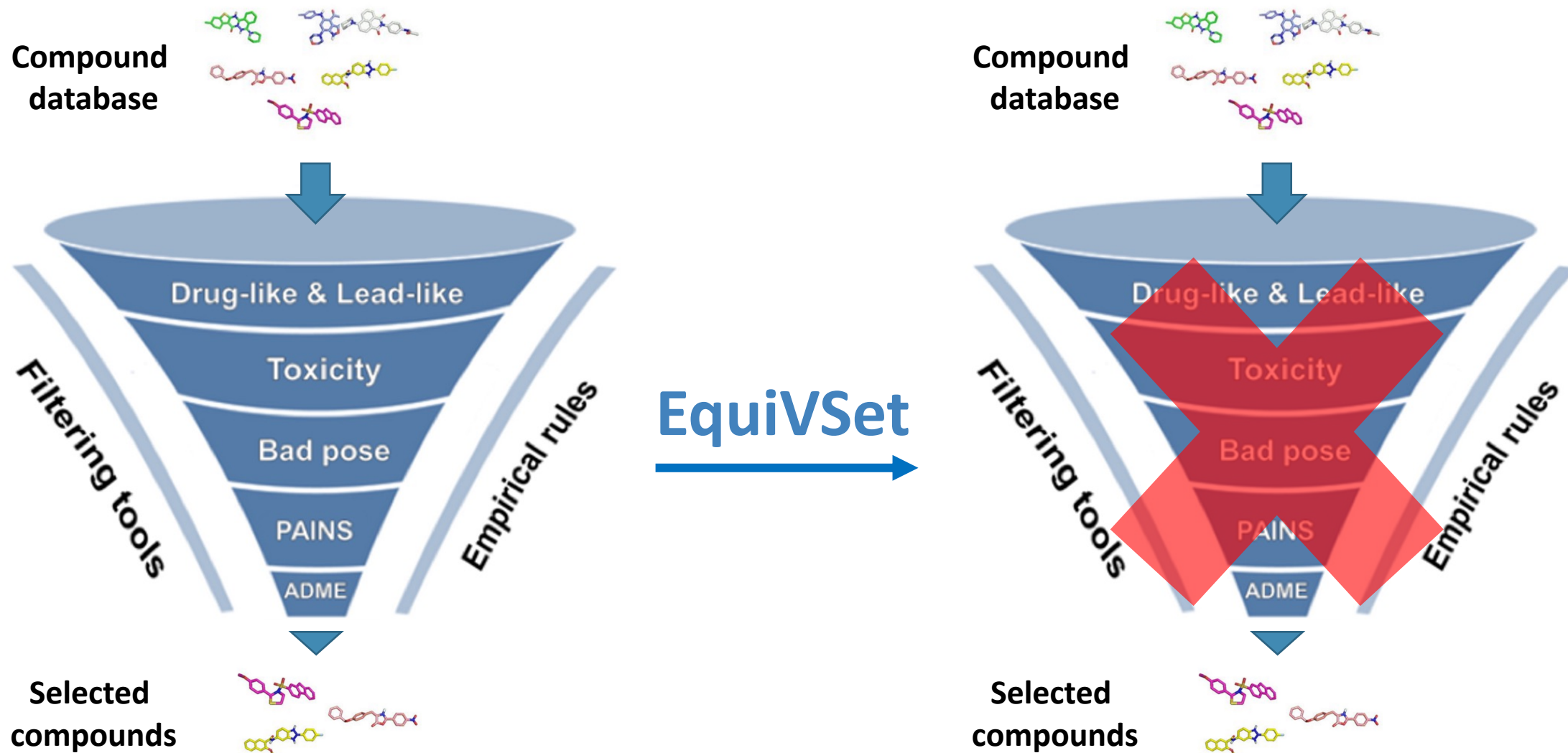


Table 3: Set anomaly detection results in the MJC metric.

Method	Double MNIST	CelebA
Random	0.0816	0.2187
PGM	0.3031 ± 0.0118	0.4812 ± 0.0064
DeepSet (NoSetFn)	0.1108 ± 0.0031	0.3915 ± 0.0133
DiffMF (ours)	0.6064 ± 0.0133	0.5455 ± 0.0079
EquiVSet _{ind} (ours)	0.4054 ± 0.0122	0.5310 ± 0.0123
EquiVSet _{copula} (ours)	0.5878 ± 0.0068	0.5549 ± 0.0053



Application: Compound Selection



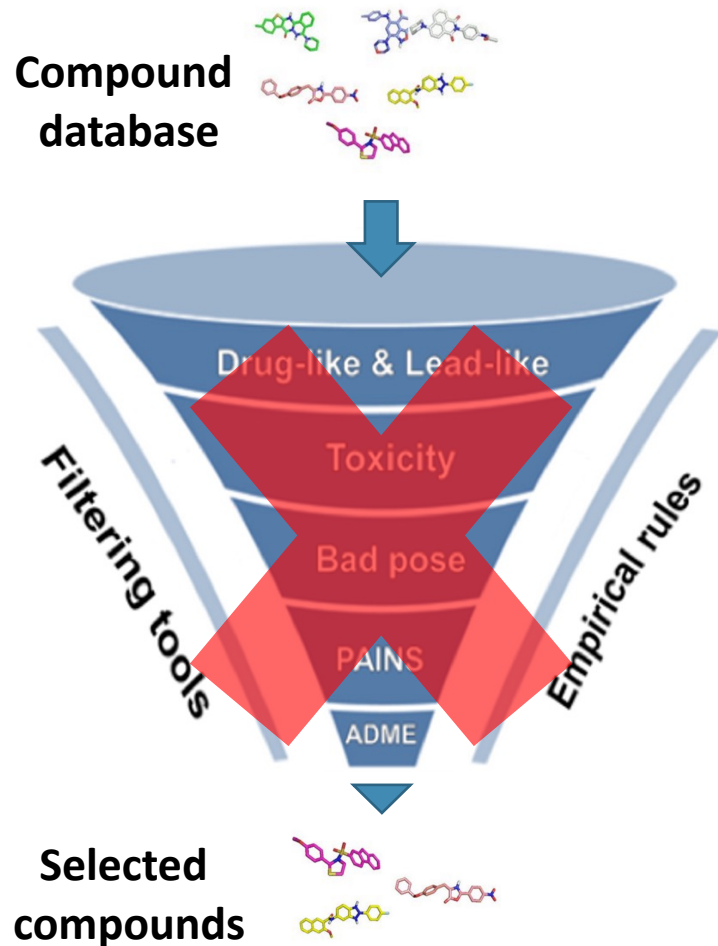


Table 4: Compound selection results in the MJC metric.

Method	PDBBind	BindingDB
Random	0.0725	0.0267
PGM	0.3499 ± 0.0087	0.1760 ± 0.0055
DeepSet (NoSetFn)	0.3189 ± 0.0034	0.1615 ± 0.0074
DiffMF (ours)	0.3534 ± 0.0143	0.1894 ± 0.0021
EquivSet _{ind} (ours)	0.3553 ± 0.0049	0.1904 ± 0.0034
EquivSet _{copula} (ours)	0.3536 ± 0.0083	0.1875 ± 0.0032

Thank you!