TabPFGen - Tabular Data Generation with TabPFN

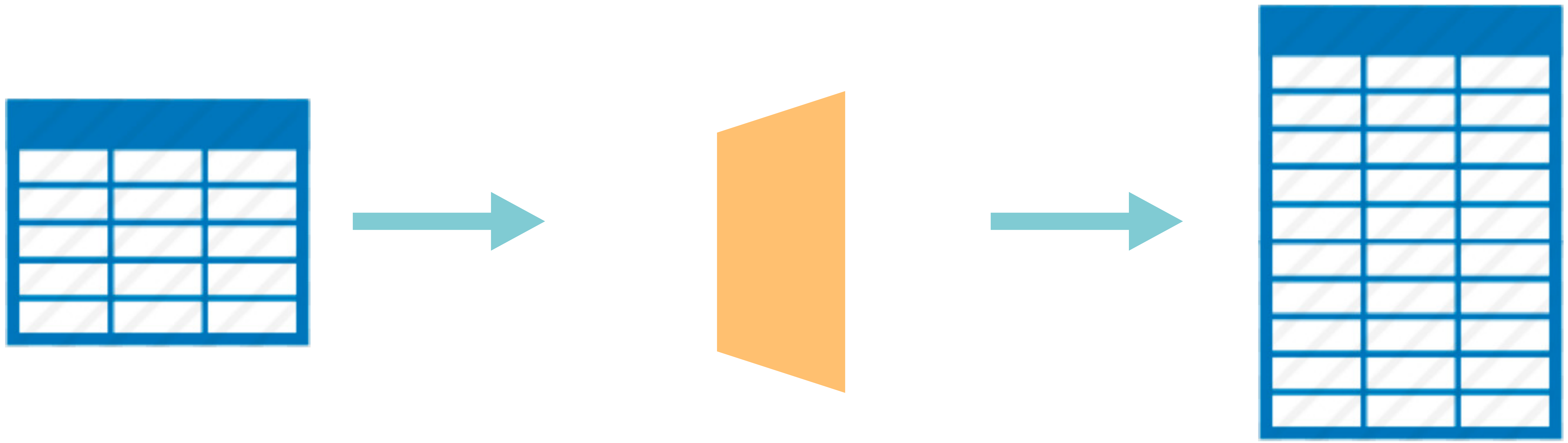
Junwei Ma, Apoorv Dankar, George Stein, Guangwei Yu, Anthony Caterini

15 Dec, 2023

layer 6

Tabular Data Generation

- The task is to generate more data according to the underlying data generation process



Tabular Data | Generative Model | Augmented Tabular Data

layer6

Goals of Tabular Data Generation

- **Augmentation** - for performance
- **Replacement** - for privacy
- **Class Balancing** - for imbalanced data
- **Imputation** - for missing data
- **Data Summarization** - for redundant data

Challenges

- **Generalization across datasets**
  - Past works usually involve unstable and time-consuming training and hyper-parameter tuning
  - The problem exacerbates when the data size is small
- **Inductive bias**
  - The inductive bias in tabular data is not clear especially with small size data
  - This is in contrast to images and text where the inductive bias is commonly exploited

Motivation - TabPFN

- TabPFN (Hollmann et al.) outperforms other tree-based and deep learning methods according to independent study (McElfresh et al.)
  - Can we harness this high-performing discriminative model for generation?
- TabPFN is exposed to a large number of data generation processes and inductive biases
  - Can we utilize the discriminative power of TabPFN for generation?

layer6

Harnessing TabPFN - Energy-based Models

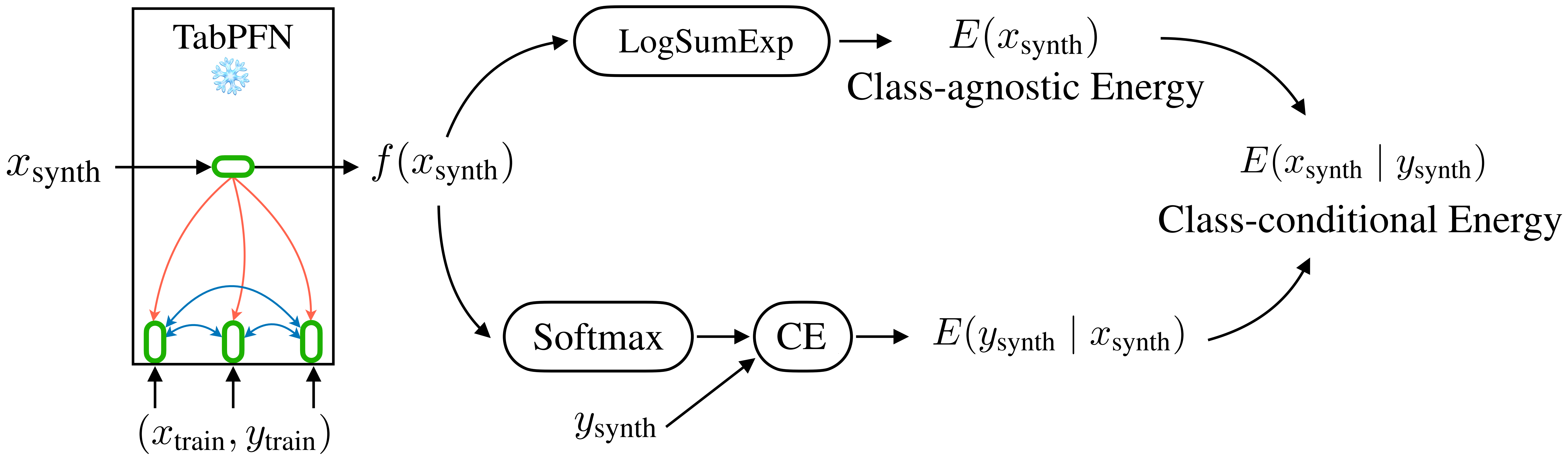
- Energy-based models (EBMs) background
  - EBMs parameterize a density using its unnornamlized log-density function
  - $p_\theta(x) = \frac{exp(-E_\theta(x))}{Z_\theta}$, where $Z_\theta = \int_x exp(-E_\theta(x))dx$
  - E_θ can be any function or network
- How do we define the energy function for generation?

Harnessing TabPFN - the Energy Function

- Grathwohl et al. proposed to use classifier outputs as the energy function
  - $$E_\theta(x) = -\log \sum_y exp(f_\theta(x)[y]),$$
  - where $f_\theta(x)$ is the classifier logit and $[]$ indicates indexing function
- To sample from this EBM, we can use Stochastic Gradient Langevin Dynamics (SGLD)

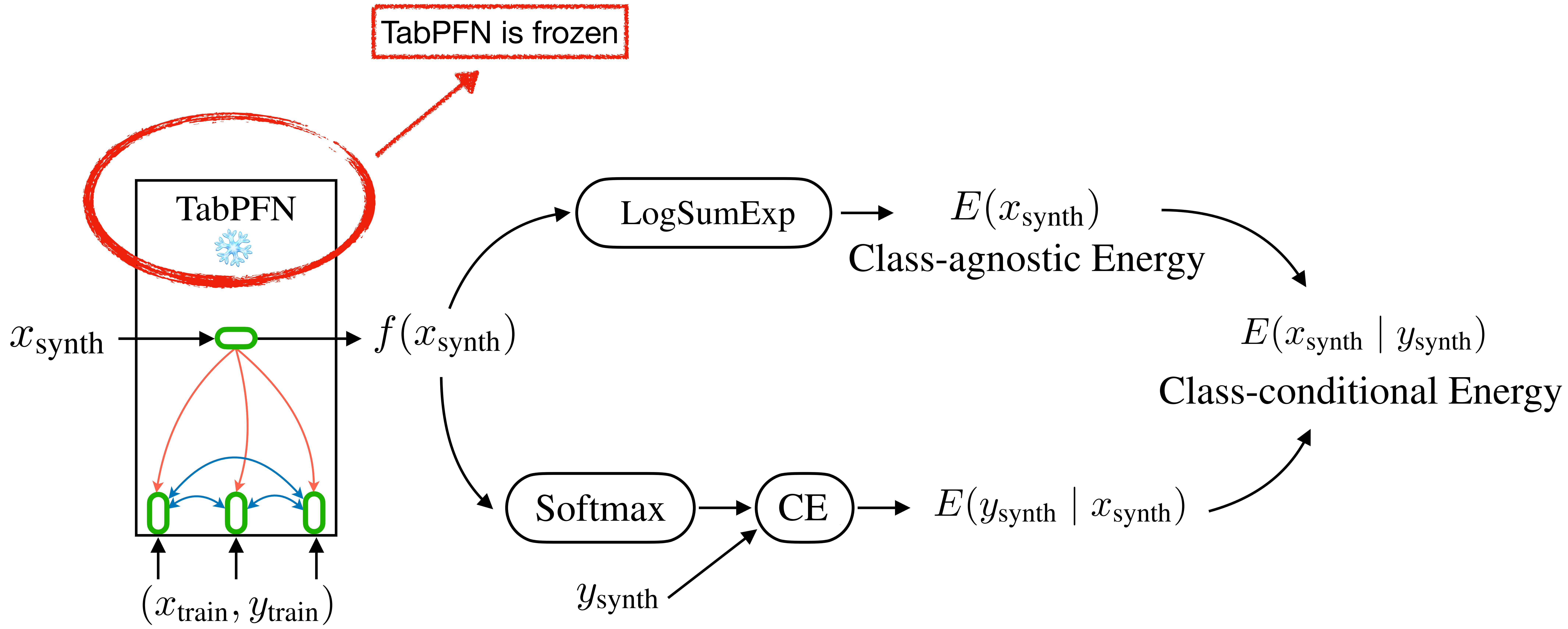
layer6

TabPFGen



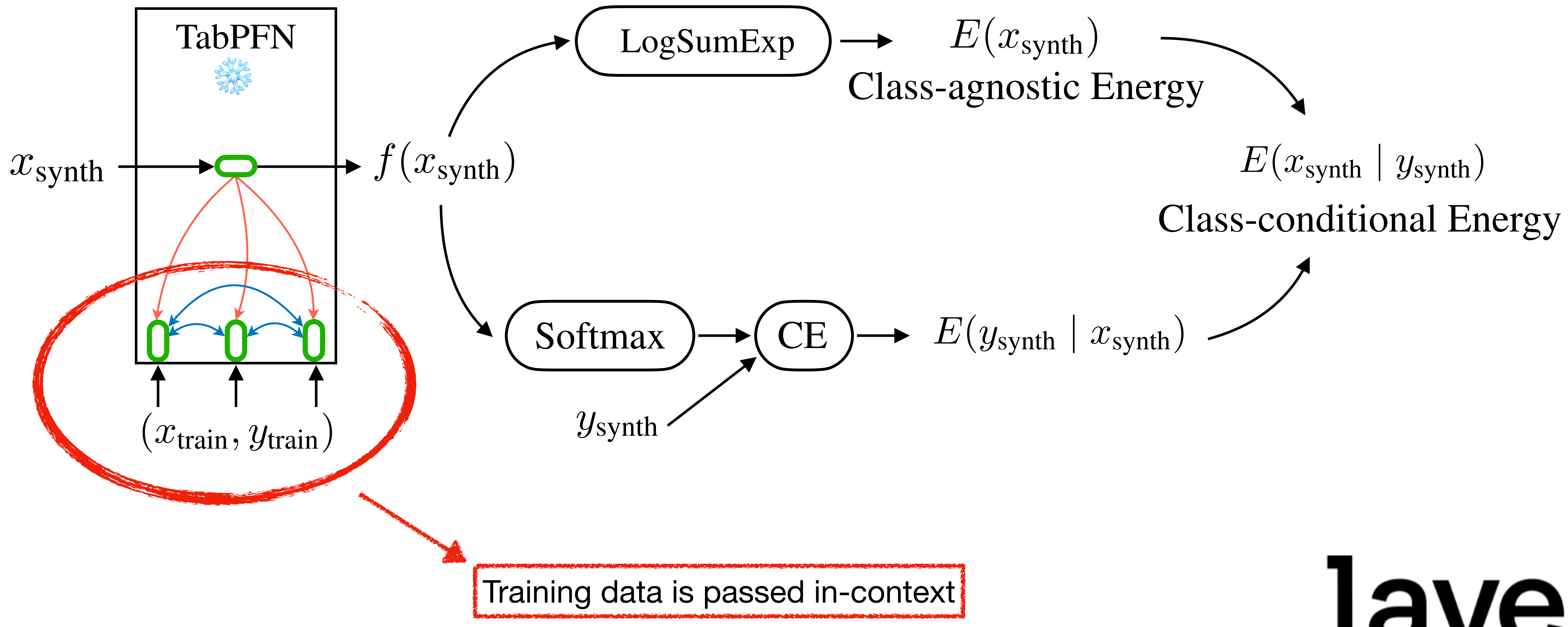
layer 6

TabPFGen



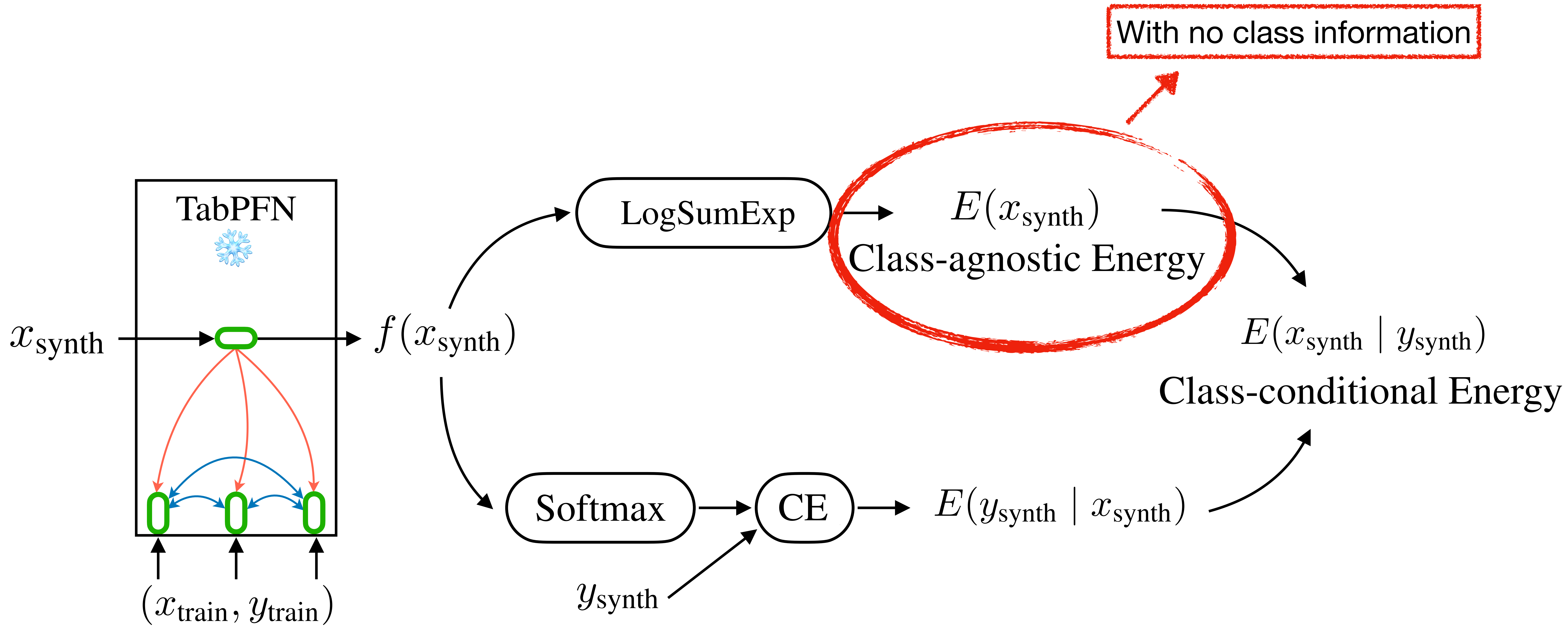
TabPFN is frozen
TabPFN
x_{synth}
$f(x_{\text{synth}})$
$(x_{\text{train}}, y_{\text{train}})$
LogSumExp
$E(x_{\text{synth}})$
Class-agnostic Energy
$E(x_{\text{synth}} \mid y_{\text{synth}})$
Class-conditional Energy
Softmax
CE
$E(y_{\text{synth}} \mid x_{\text{synth}})$
y_{synth}

TabPFGen



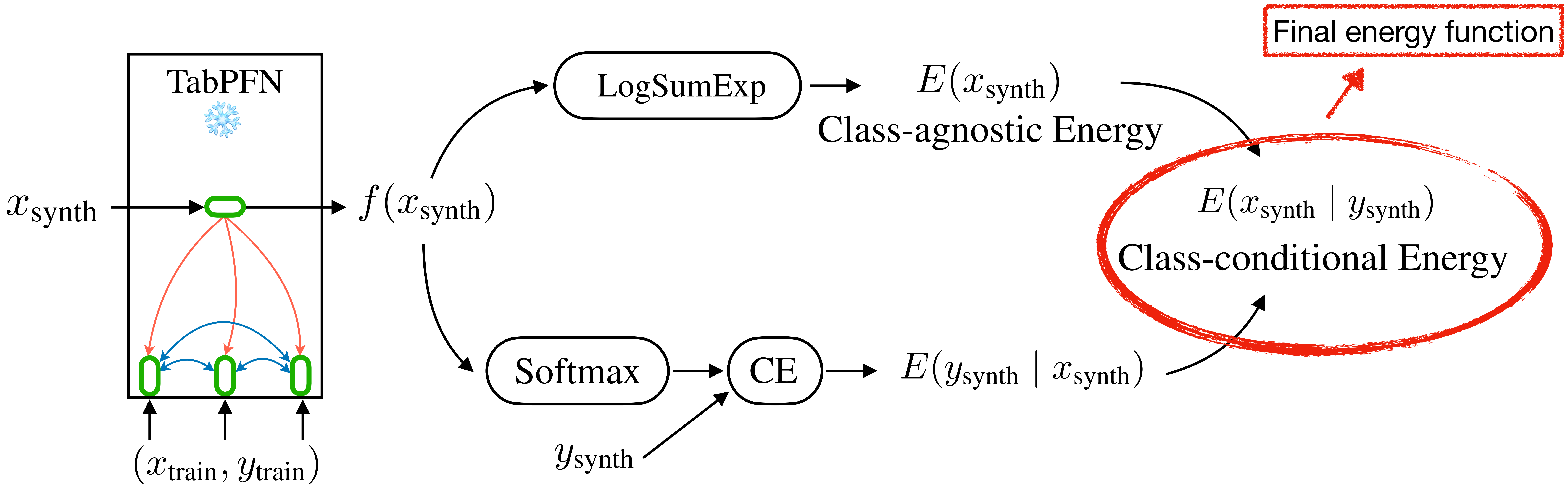
layer6

TabPFGen



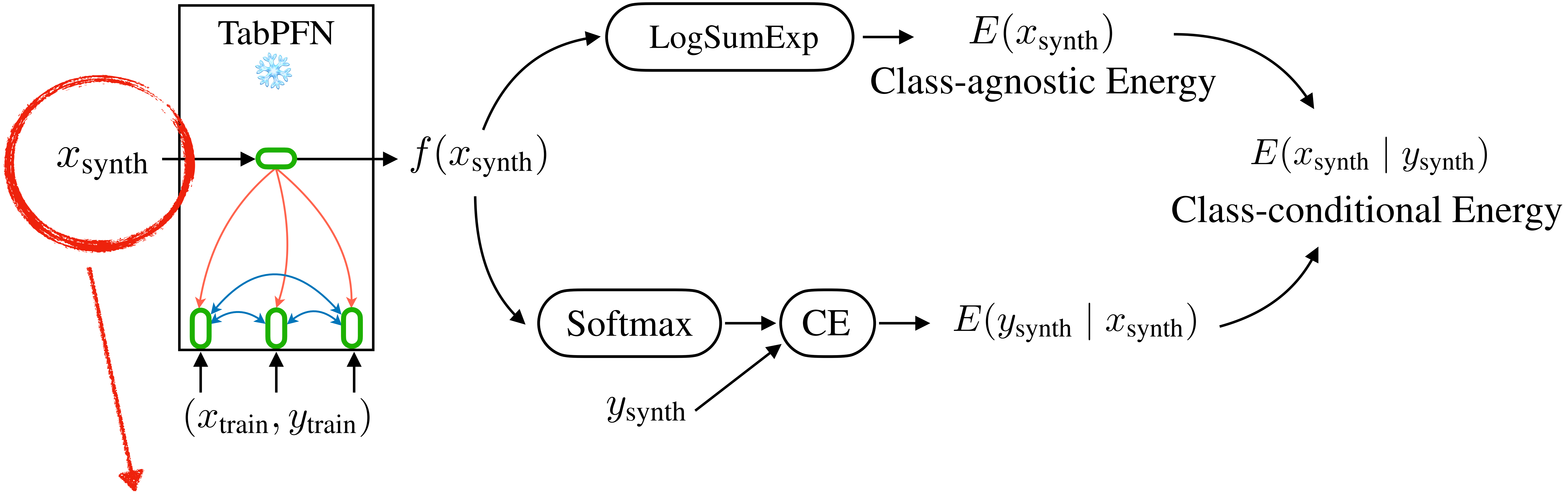
layer 6

TabPFGen



layer6

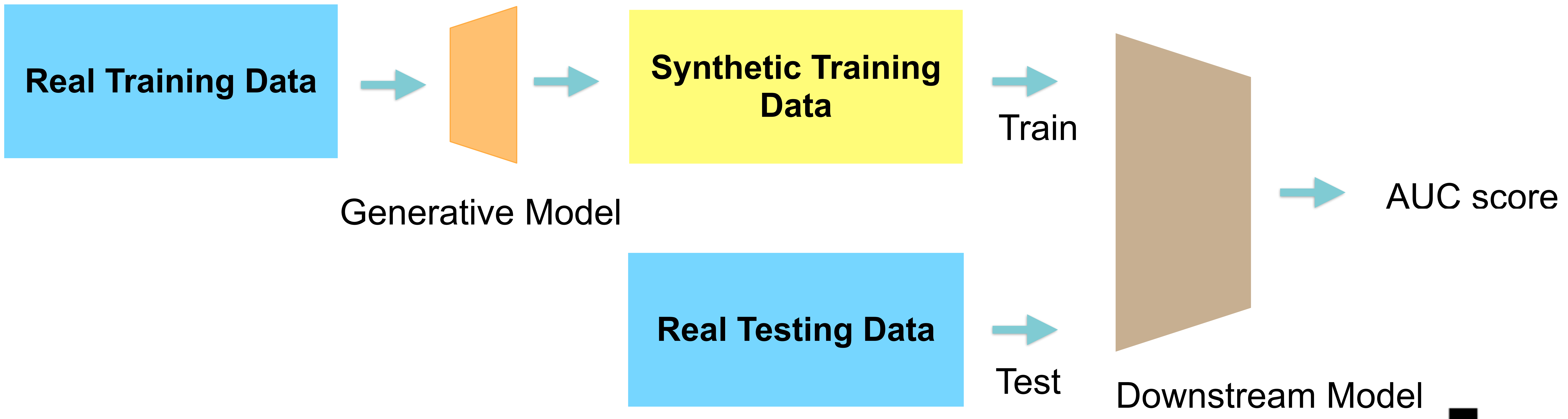
TabPFGen



In-context learning generation with SGLD:
- Frozen TabPFN's discriminative power is inverted for generation
- No additional training or hyper-parameter tuning required

Experimental Setup

- Use synthetic data to augment, replace or class balance
- Test set is used to evaluate downstream models performance with synthetic data
- Downstream models include: XGB, RF, LR and TabPFN



layer6

Results - Augmentation and Replacement

Augmentation

Model	Original	SMOTE	CTGAN	TVAE	NF	RTVAE	TabDDPM	TabPFGen
XGB	0.924±3e−4	0.926±3e−4	0.912±2e−4	0.914±7e−4	0.912±4e−4	0.917±3e−4	0.927±3e−4	**0.934**±3e−4
RF	0.906±3e−4	0.906±2e−3	0.898±1e−3	0.904±1e−3	0.894±3e−4	0.907±2e−3	0.911±7e−4	**0.912**±4e−4
LR	0.920±7e−4	0.914±3e−3	0.904±3e−3	0.909±6e−3	0.901±9e−4	0.906±8e−3	0.885±3e−4	**0.921**±2e−4
TabPFN	0.934±2e−3	0.927±1e−3	0.930±1e−3	0.931±1e−3	0.928±3e−4	0.932±1e−3	0.929±5e−4	**0.935**±3e−4
XGB	N/A	0.907±4e−4	0.842±8e−4	0.858±2e−3	0.700±6e−4	0.795±9e−4	0.812±3e−4	**0.927**±3e−4
RF	N/A	0.894±1e−3	0.837±6e−4	0.844±5e−4	0.676±2e−3	0.774±3e−4	0.814±9e−4	**0.906**±6e−4
LR	N/A	0.893±2e−3	0.843±6e−4	0.873±1e−3	0.722±3e−3	0.854±7e−4	0.876±3e−4	**0.920**±1e−3
TabPFN	N/A	0.920±8e−4	0.888±4e−4	0.887±3e−4	0.705±2e−3	0.862±1e−3	0.894±7e−4	**0.934**±2e−4

layer 6

Results - Augmentation and Replacement

Model	Original	SMOTE	CTGAN	TVAE	NF	RTVAE	TabDDPM	TabPFGen
XGB	0.924±3e−4	0.926±3e−4	0.912±2e−4	0.914±7e−4	0.912±4e−4	0.917±3e−4	0.927±3e−4	**0.934**±3e−4
RF	0.906±3e−4	0.906±2e−3	0.898±1e−3	0.904±1e−3	0.894±3e−4	0.907±2e−3	0.911±7e−4	**0.912**±4e−4
LR	0.920±7e−4	0.914±3e−3	0.904±3e−3	0.909±6e−3	0.901±9e−4	0.906±8e−3	0.885±3e−4	**0.921**±2e−4
TabPFN	0.934±2e−3	0.927±1e−3	0.930±1e−3	0.931±1e−3	0.928±3e−4	0.932±1e−3	0.929±5e−4	**0.935**±3e−4
XGB	N/A	0.907±4e−4	0.842±8e−4	0.858±2e−3	0.700±6e−4	0.795±9e−4	0.812±3e−4	**0.927**±3e−4
RF	N/A	0.894±1e−3	0.837±6e−4	0.844±5e−4	0.676±2e−3	0.774±3e−4	0.814±9e−4	**0.906**±6e−4
LR	N/A	0.893±2e−3	0.843±6e−4	0.873±1e−3	0.722±3e−3	0.854±7e−4	0.876±3e−4	**0.920**±1e−3
TabPFN	N/A	0.920±8e−4	0.888±4e−4	0.887±3e−4	0.705±2e−3	0.862±1e−3	0.894±7e−4	**0.934**±2e−4

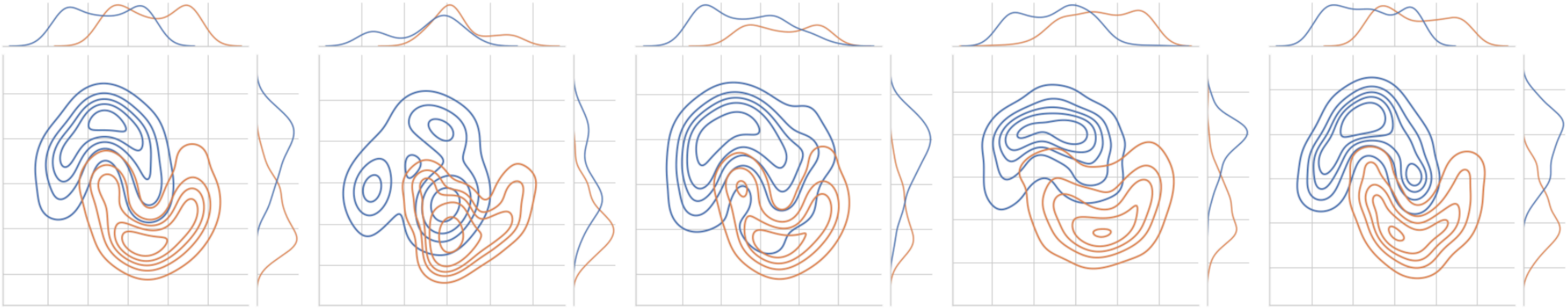
Replacement

layer 6

Results - Class Balancing

Dataset	Original	Sampling	SMOTE	CTGAN	TVAE	NF	RTVAE	TabDDPM	TabPFGen
KC	0.823	0.875	0.872	0.859	0.848	0.862	0.866	0.805	**0.877**
PC	0.824	0.811	0.836	0.827	0.835	0.825	**0.841**	0.825	**0.841**
BL	0.731	0.757	0.756	0.743	0.714	0.755	0.706	**0.774**	0.767
CL	0.925	0.935	0.949	0.793	0.771	0.795	0.909	0.915	**0.955**
DI	0.837	0.832	0.832	0.831	0.813	0.832	0.837	0.843	**0.844**

Results - Qualitative



(a) Original (b) CTGAN (c) NF (d) TabDDPM (e) TabPFGen

Thank you!

jeremy@layer6.ai