Toward a Well-Calibrated Discrimination through Survival Outcome-Aware Contrastive Learning

Dongjoon Lee*
Chung-Ang University
dongzza97@cau.ac.kr

Hyeryn Park*
Chung-Ang University
hyeryn2000@cau.ac.kr

Changhee Lee
Korea University
changheelee@korea.ac.kr

Outline

- Introduction to Survival Analysis
- Consideration
- Objective
- Challenges and Motivation
- Proposed Method
- Experiments

What is survival analysis?

- A very common outcome in medical studies is the time until an event occurs:
  - The time until a patient dies
  - The time until a patient suffers a heart attack
  - The time until a liver transplant patient needs a new liver
  - The time until the recurrence of cancer following treatment

- Data involving such an outcome is often called “time-to-event” data or “failure-time data” or “survival” data, and the branch of statistics that deals with analyzing these data is called survival analysis

Survival data

- Survival (a.k.a. time-to-event) data

$$\mathcal{D} = \{(\mathbf{x}_i, \tau_i, \delta_i)\}_{i=1}^{N}$$

  - $\mathbf{x}$: Observed features (covariates)
  - τ: Time-to-event or time-to-censoring elapsed since the baseline (e.g., the entry to a clinical trial)
  - δ: Label indicating whether event the event or the censoring occurred

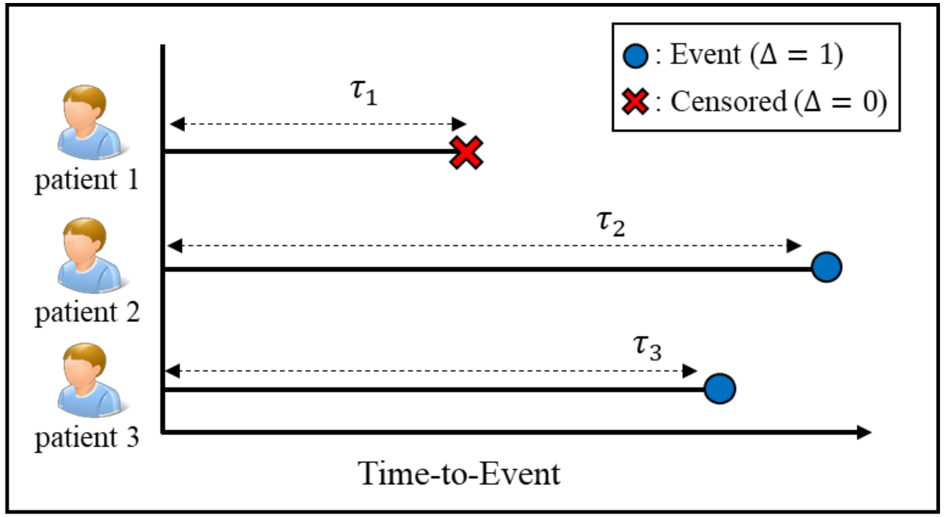


Figure. An illustration of survival data

Survival data

- Survival (a.k.a. time-to-event) data

$$\mathcal{D} = \{(\mathbf{x}_i, \tau_i, \delta_i)\}_{i=1}^{N}$$

  - $\mathbf{x}$: Observed features (covariates)
  - τ: Time-to-event or time-to-censoring elapsed since the baseline (e.g., the entry to a clinical trial)
  - δ: Label indicating whether event the event or the censoring occurred

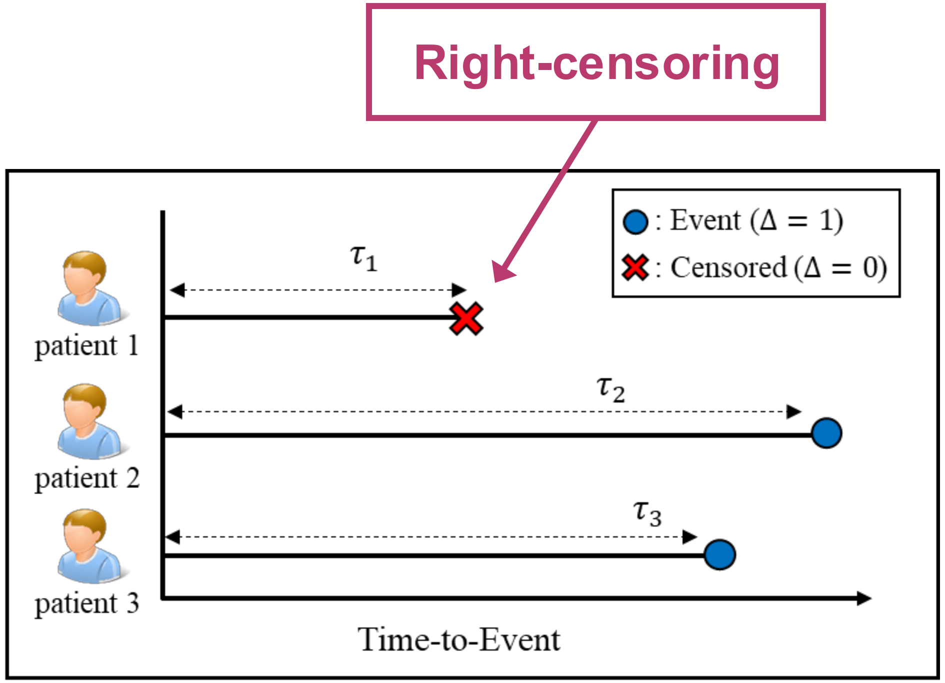


Figure. An illustration of survival data

- Distinct Characteristics: **Right-censoring**

Survival data

- Notations
  - $T \in \mathbb{R}_+$ be the random variable for time-to-event
  - $C \in \mathbb{R}_+$ be the random variables for time-to-censoring

- Right-censoring indicates when censoring occurs before the event of interest is observed. Denoting t and c be the realizations of r.v.s T and C, we have

$$\delta = \mathbb{I}(t \leq c) \qquad \tau = \min(t, c)$$

  - Often assume "independent censoring", i.e., $P(T, C|X = \mathbf{x}) = P(T|X = \mathbf{x})P(C|X = \mathbf{x})$

Solution: Survival Analysis

- **Our goal**
  - Provides the probability an event occurring as a function of time and patient features
  - Provides understanding of interactions between features and the time-to-event outcomes

- We want to use partial information from the right-censored samples:
  - Censoring implies that the event will occur after the censoring time

Important quantities : Survival / Risk function

- Formally, we want to estimate the survival function given $\mathbf{x}$

$$S(t|\mathbf{x}) = \mathbb{P}(T > t|\mathbf{x})$$

probability an event occurring after time t

  - $T \in \mathbb{R}_+$: Random variable for the time-to-event
  - $\mathbf{x}$: Patient input feature

- Or equivalently, we want to estimate the risk function given $\mathbf{x}$

$$R(t|\mathbf{x}) = 1 - S(t|\mathbf{x}) = \mathbb{P}(T \leq t|\mathbf{x})$$

probability an event occurring before time t

Consideration

- Discriminate patients' risks of having an event of interest

Patient A

- *Age : 55*
- *Stage : 2B*
- *ER+ / PR +*

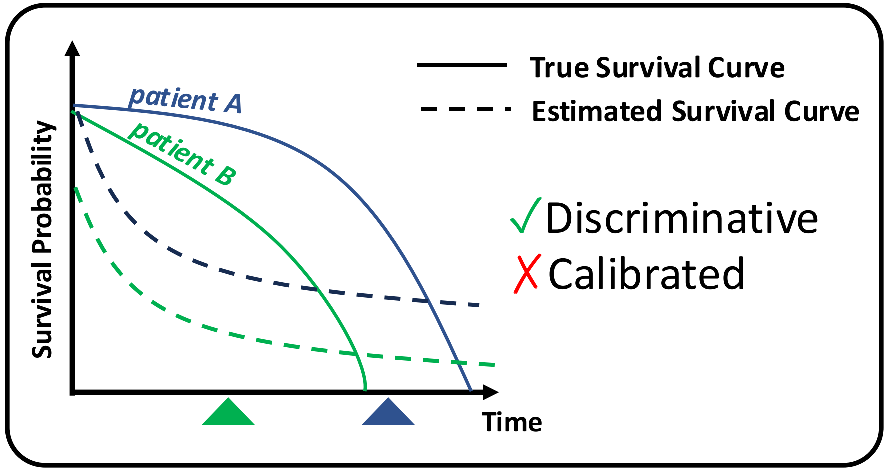
  ⋮
- *Event time : 32 days*

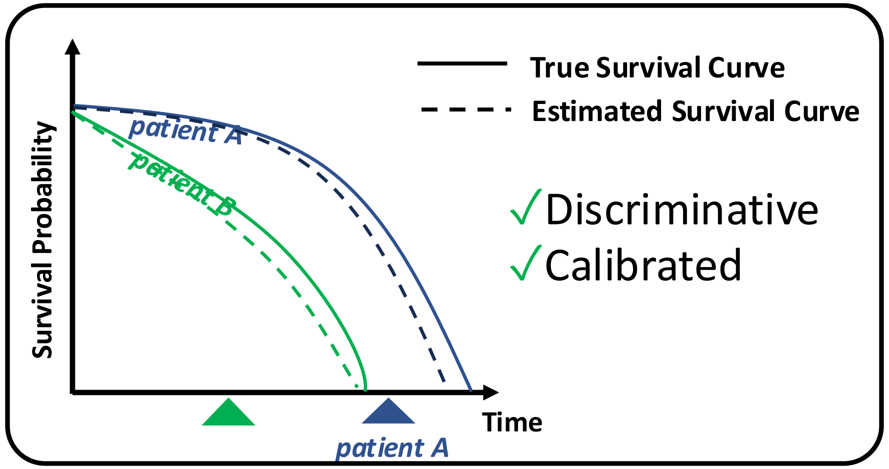
Patient B

- *Age : 45*
- *Stage : 3A*
- *ER- / PR -*

  ⋮
- *Event time : 14 days*

Time-to-Event Models





Objective : Negative log-likelihood loss

- The log-likelihood of the time-to-events for survival dataset → unbiased
  - Event is observed (i.e., $\delta_i = 1$), knowing that the event occurred at time τ_i
  - Event is not observed (i.e., $\delta_i = 0$), knowing that the event will occur after time τ_i

$$\mathcal{L}_{NLL} = -\log \prod_{i=1}^{N} \left[\hat{p}(\tau_i|\mathbf{x}_i)^{\delta_i} \cdot \hat{S}(\tau_i|\mathbf{x}_i)^{(1-\delta_i)} \right]$$

$$= -\sum_{i=1}^{N} \left[\underbrace{\delta_i \log \hat{p}(\tau_i|\mathbf{x}_i)}_{\text{for uncensored}} + \underbrace{(1-\delta_i) \log \hat{S}(\tau_i|\mathbf{x}_i)}_{\text{for censored}} \right]$$

Objective : Ranking loss

- Often augmented with the NLL loss to enhance the discriminative power
- Aim to maximize a relaxed proxy of the concordance index
  - Well-established metric for evaluating the quality of patient rankings based on the risk predictions of survival model

$$\mathcal{L}_{Rank} = \sum_{i \neq j} A_{i,j} \cdot \eta\left(\hat{R}(\tau_i|\mathbf{x}_i), \hat{R}\left(\tau_i\middle|\mathbf{x}_j\right)\right)$$

Objective : Ranking loss

$$\mathcal{L}_{Rank} = \sum_{i \neq j} A_{i,j} \cdot \eta\left(\hat{R}(\tau_i|\mathbf{x}_i), \hat{R}(\tau_i|\mathbf{x}_j)\right)$$

acceptable pairs ($A_{i,j}$)

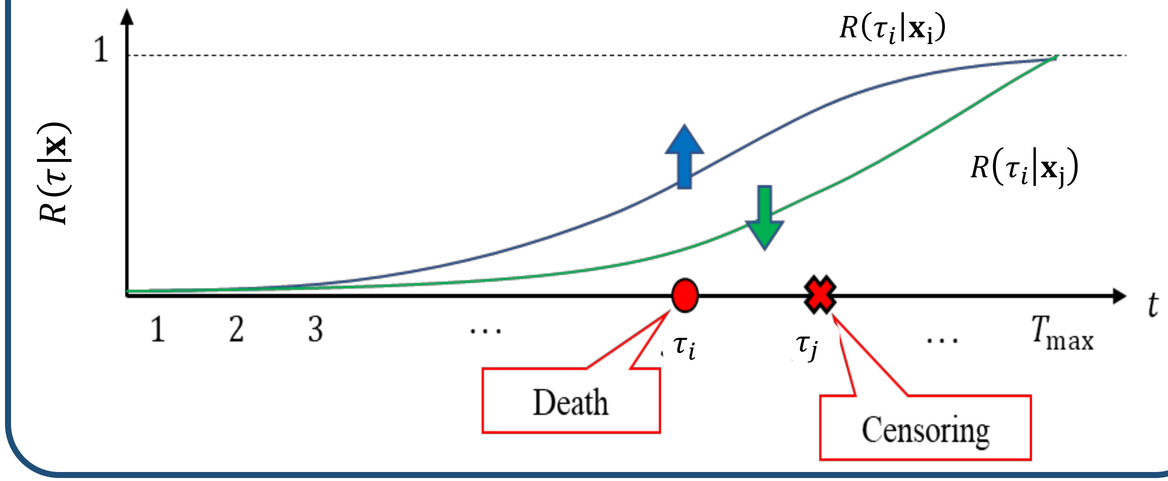
where $A_{i,j} = \mathbb{I}(\delta_i = 1, \tau_i < \tau_j)$

and $\eta(x, y) = \exp\left(\frac{-(x-y)}{\sigma}\right)$

Case 1: Correctly ordered pairs

- $\hat{R}(\tau_i|\mathbf{x}_i) > \hat{R}(\tau_i|\mathbf{x}_j)$ (O)
- Rewards the estimated Risk Function

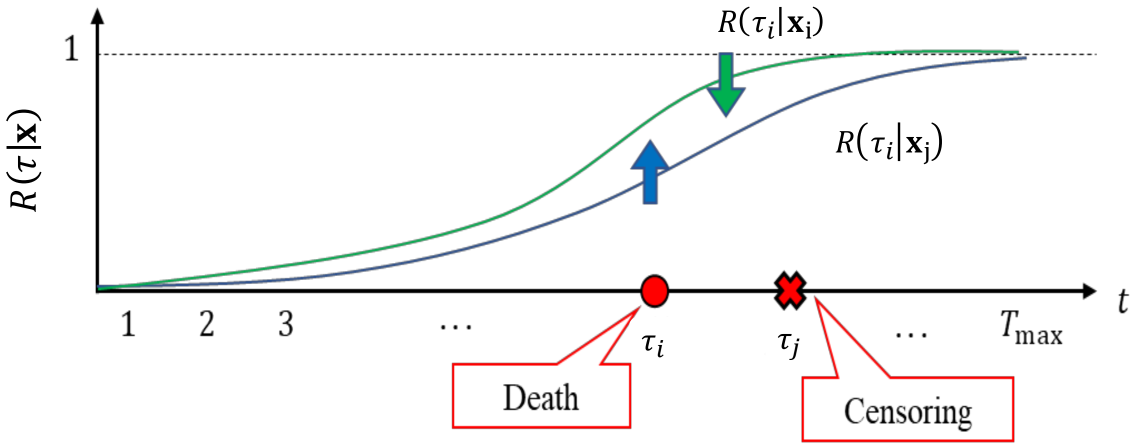
$\hat{R}(\tau_i|\mathbf{x}_i)$ ↑ $\hat{R}(\tau_i|\mathbf{x}_j)$ ↓



Case 2: Wrongly ordered pairs

- $\hat{R}(\tau_i|\mathbf{x}_i) < \hat{R}(\tau_i|\mathbf{x}_j)$ (X)
- Penalizes the estimated Risk Function

$\hat{R}(\tau_i|\mathbf{x}_i)$ ↑ $\hat{R}(\tau_i|\mathbf{x}_j)$ ↓



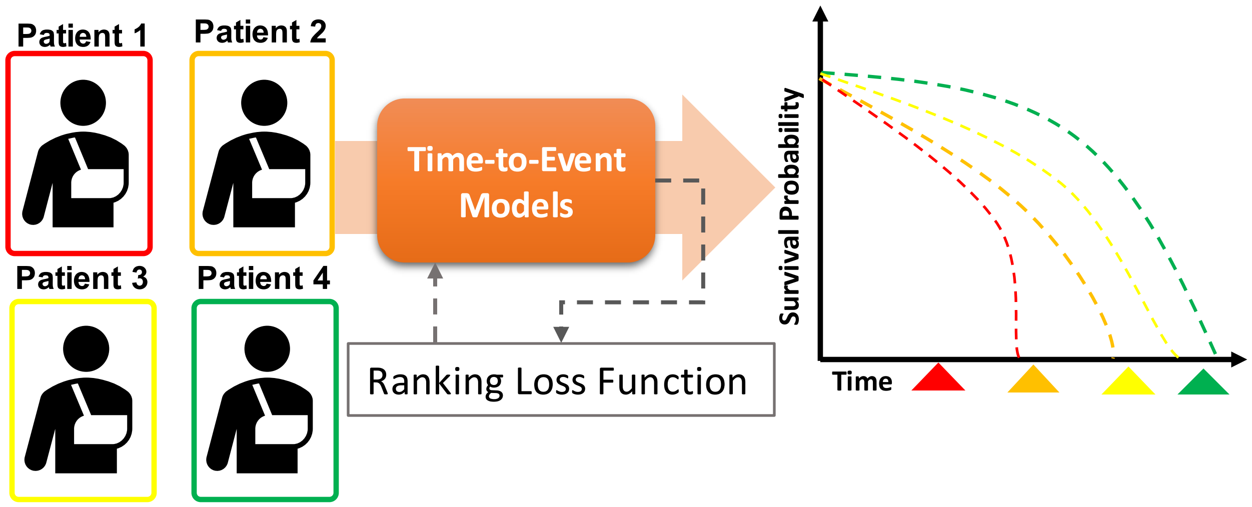
Challenges

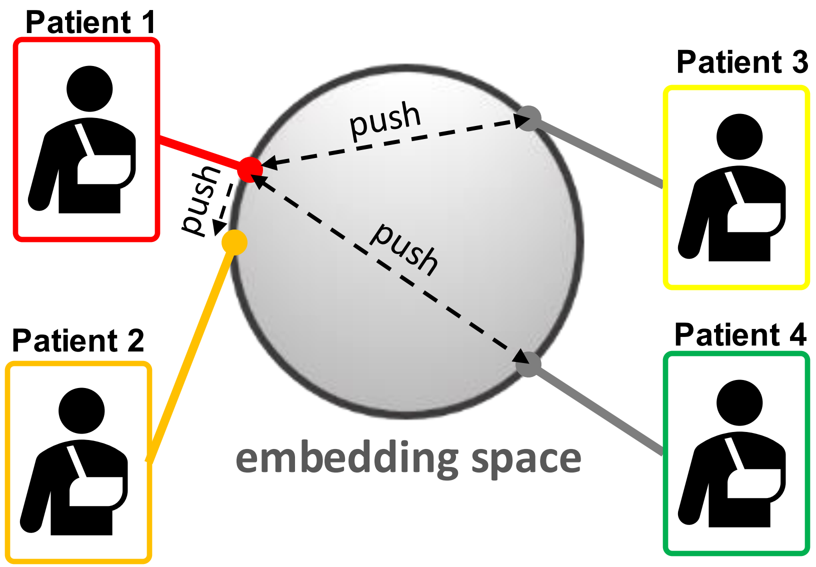
- Combining NLL with ranking loss enhances discrimination but compromises calibration, harming the clinical utility of predicted survival outcomes.
- Ranking loss directly modifies model outputs, potentially leading to misalignment with the actual risk distribution.
  - Typically based on exponential, log-sigmoid, or linear functions

Model	Ranking Loss
DeepHit	$\exp(-(\hat{R}(\tau_i \mid \mathbf{x_i}) - \hat{R}(\tau_i \mid \mathbf{x}_j)/\boldsymbol{\kappa})$
DCS	$\exp(-(\hat{S}(\tau_i \mid \mathbf{x_j}) - \hat{S}(\tau_i \mid \mathbf{x}_i)/\boldsymbol{\kappa})$
LowerCI	$\log \sigma(\hat{R}(\tau_i \mid \mathbf{x_i}) - \hat{R}(\tau_i \mid \mathbf{x}_j)$
SSMTL	$\hat{h}(\tau_i \mid \mathbf{x}_j) - \hat{h}(\tau_i \mid \mathbf{x_i})$

Motivation

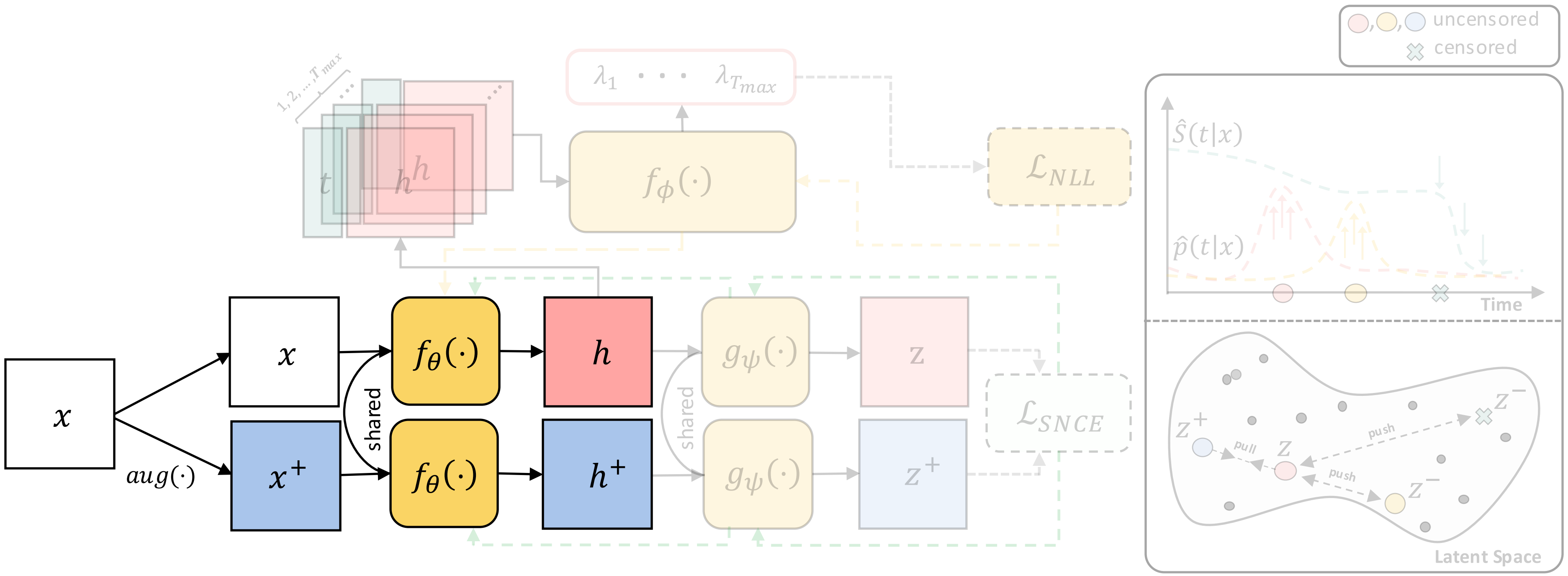
- Propose a novel contrastive learning approach for deep survival model
  - Differentiate each sample by their survival outcome, leveraging contrastive learning framework
  - Overcomes ranking loss limitations from directly comparing model outcome in the form of risk/survival function.





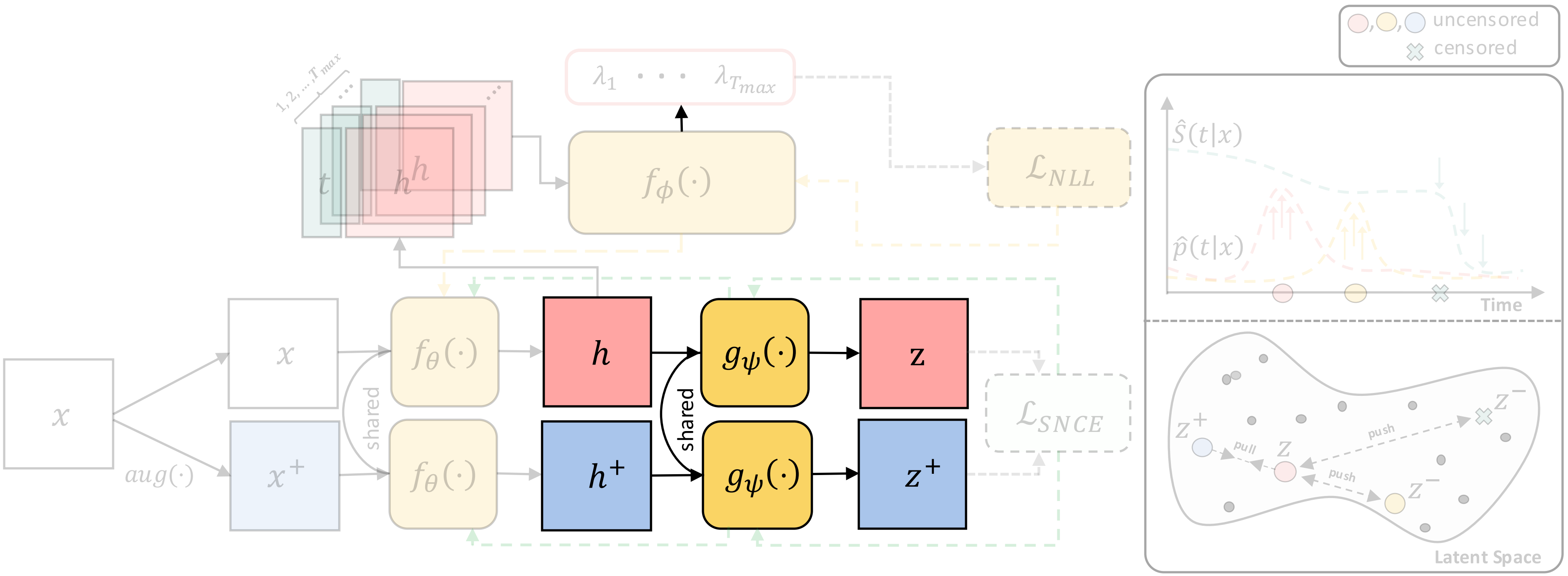
Proposed Method : Network components

- **The encoder**, $f_\theta : \mathcal{X} \rightarrow \mathcal{H}$, takes features $\mathbf{x} \in \mathcal{X}$ as input and outputs latent representation, i.e., $\mathbf{h} = f_\theta(\mathbf{x})$.



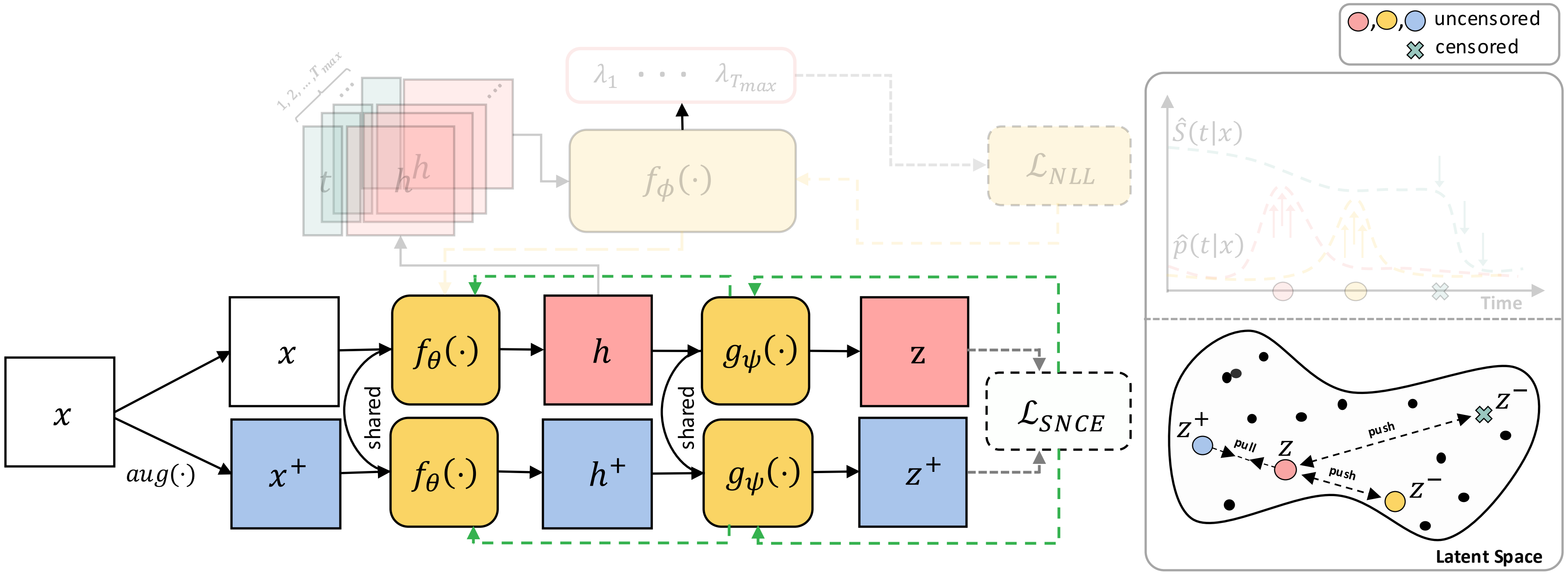
Proposed Method : Network components

- **The projection head.**, $g_\psi : \mathcal{H} \rightarrow \mathbb{R}^d$, maps latent representation $\mathbf{h}$ to the embedding space where contrastive learning is applied, i.e., $\boldsymbol{z} = f_\theta(\mathbf{h})$.



Proposed Method : Network components

- **Contrastive Learning Network**
  - By passing the original, positive, and negative samples through $f = g_\psi \circ f_\theta$, computing our survival outcome-based contrastive learning loss function $\mathcal{L}_{SNCE}$



Proposed Method : Contrastive Learning for SA

- Goal : Aligns with our inductive bias that patients with similar survival outcomes should share similar clinical status, which manifests through similar representations.
- Noise Contrastive Estimation (NCE)
  - To learn mapping $f = g_\psi \circ f_\theta$ utilizing a positive sample $\mathbf{x}^+ \sim p_{X^+}$, and negative samples $\mathbf{x}^- \sim q$

$$\mathbb{E}_{\substack{\mathbf{x} \sim p_X \\ \mathbf{x}^+ \sim p_{X^+}}} \left[-\log \frac{e^{s(\mathbf{x}, \mathbf{x}^+)}}{M \cdot \mathbb{E}_{\mathbf{x}^- \sim q}\left[e^{s(\mathbf{x}, \mathbf{x}^-)}\right]} \right]$$

  - M : scaling term which is set to the batch size, $s(\mathbf{x}, \mathbf{x}') = \frac{f(\mathbf{x})^T f(\mathbf{x}')}{||f(\mathbf{x})|| \cdot ||f(\mathbf{x}')||}$
  - omit the corresponding temperature ν and write $e^{s(\mathbf{x}, \mathbf{x}^-)}$ to denote $e^{s(\mathbf{x}, \mathbf{x}^-)/\nu}$

Proposed Method : Contrastive Learning for SA

- Key aspect of NCE : selecting negative samples to differentiate the anchor sample

- To reflect the difference in the time-to-events in the embedding space, we design a novel distribution q by utilizing the available information from survival outcomes.

Proposed Method : Contrastive Learning for SA

- To accurately distinguish patients based on their time-to-event outcomes, we fully utilize the time-to-event information
- Hence, given an anchor $(\mathbf{x}, \tau)$ and a negative $(\mathbf{x}^{-}; \tau^{-})$, we define the weight function, $\sigma > 0$ is a temperature coefficient.
  – This function is a variant of the Laplacian Kernel, which assigns larger weights to samples with large differences in time-to-event outcomes, and smaller weights to samples with small differences

$$w(\tau^{-}; \tau) = 1 - e^{\,|\tau - \tau^{-}|\,/\sigma}$$

Proposed Method : Contrastive Learning for SA

- Designing q based on the following inductive bias : similar patients are more likely to experience the event at similar time points than the ones who are not.

- We will slightly abuse the notation $w(\mathbf{x}^{-};\mathbf{x})$ to denote $w(\tau^{-};\tau)$

$$q(\mathbf{x}^{-};\mathbf{x}) = \frac{1}{Z} w(\mathbf{x}^{-};\mathbf{x}) p(\mathbf{x}^{-})$$

normalizing constant $Z = \frac{1}{M}\sum_{j=1}^{M} w(x_j^{-};\ x)$

Proposed Method : Contrastive Learning for SA

- Importance sampling using survival outcomes

$$E_{x^- \sim q}\left[e^{s(x,x^-)}\right] = E_{x^- \sim p}\left[\left(\frac{q(x^-;x)}{p(x^-)}\right) \cdot e^{s(x,x^-)}\right]$$

$$= E_{x^- \sim p}\left[\left(\frac{w(x^-;x)}{Z}\right) \cdot e^{s(x,x^-)}\right]$$

$$\approx \frac{1}{Z \cdot M} \sum_{j=1}^{M} w\left(x_j^-;\ x\right) \cdot e^{s\left(x,x_j^-\right)}$$

normalizing constant $Z = \frac{1}{M}\sum_{j=1}^{M} w\left(x_j^-;\ x\right)$

Proposed Method : Contrastive Learning for SA

- Survival outcome-aware NCE (SNCE) loss

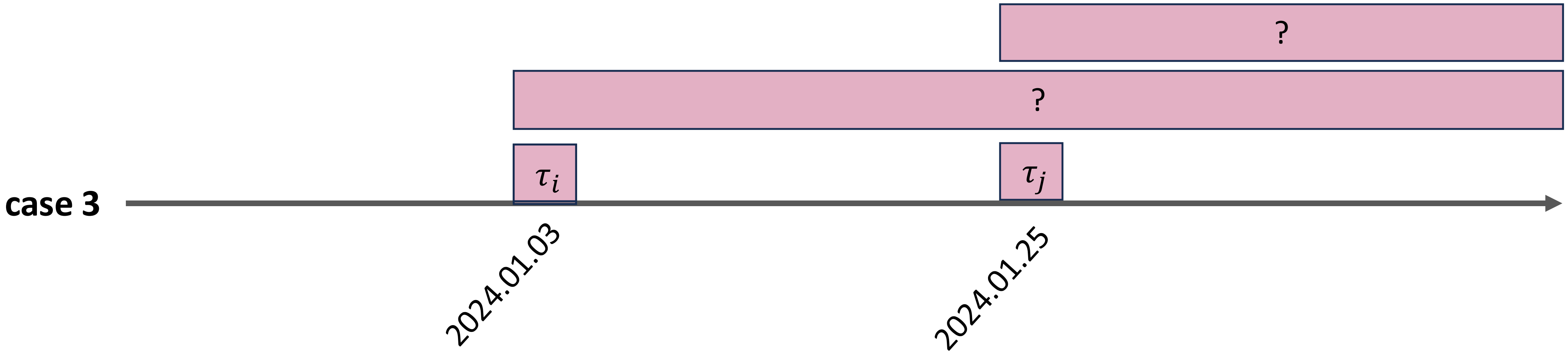
$$\mathcal{L}_{SNCE} = \sum_{i=1}^{N} \left[-\log \left(\frac{e^{s(x_i\,,\,x_i^+)}}{\frac{1}{Z} \sum_{j=1}^{M} w\left(x_j^-; x_i\right) \cdot e^{s\left(x_i\,,\,x_j^-\right)}} \right) \right]$$

Proposed Method : Handling Right-Censoring

- case 1 : Both samples are uncensored(i.e., have observed events)
- case 2 : Both samples are censored
- case 3 : One is uncensored and the other is censored.

Proposed Method : Handling Right-Censoring

- case 1 : Both samples are uncensored(i.e., have observed events)
- case 2 : Both samples are censored
- case 3 : One is uncensored and the other is censored.

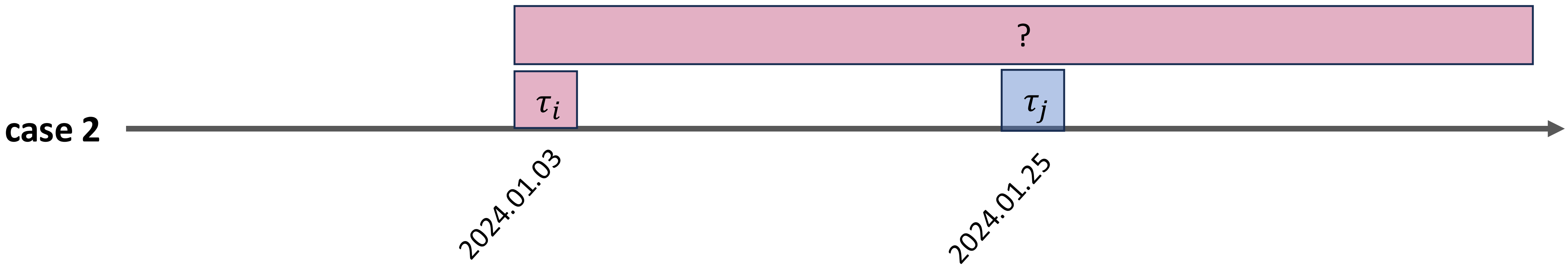


Proposed Method : Handling Right-Censoring

- case 1 : Both samples are uncensored(i.e., have observed events)
- case 2 : Both samples are censored
- case 3 : One is uncensored and the other is censored.

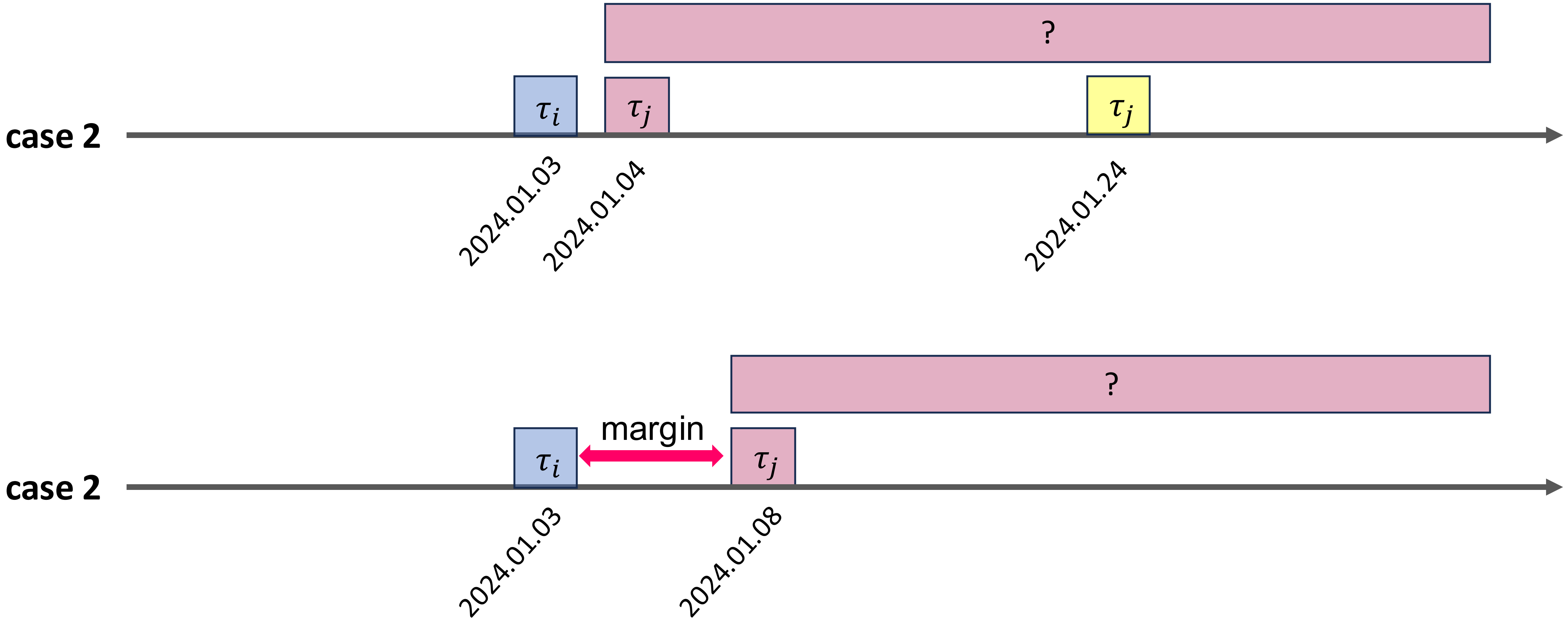
Proposed Method : Handling Right-Censoring

- case 1 : Both samples are uncensored(i.e., have observed events)
- case 2 : Both samples are censored
- case 3 : One is uncensored and the other is censored.



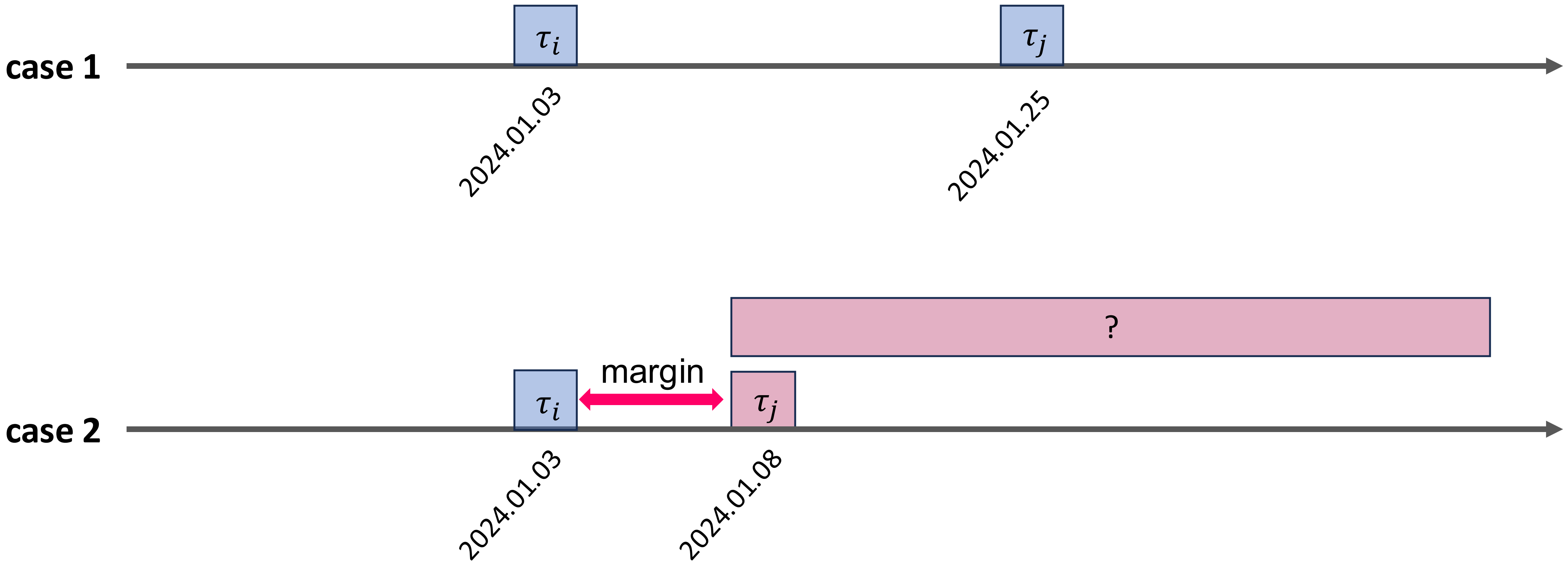
Proposed Method : Handling Right-Censoring

- case 3 : One is uncensored and the other is censored.



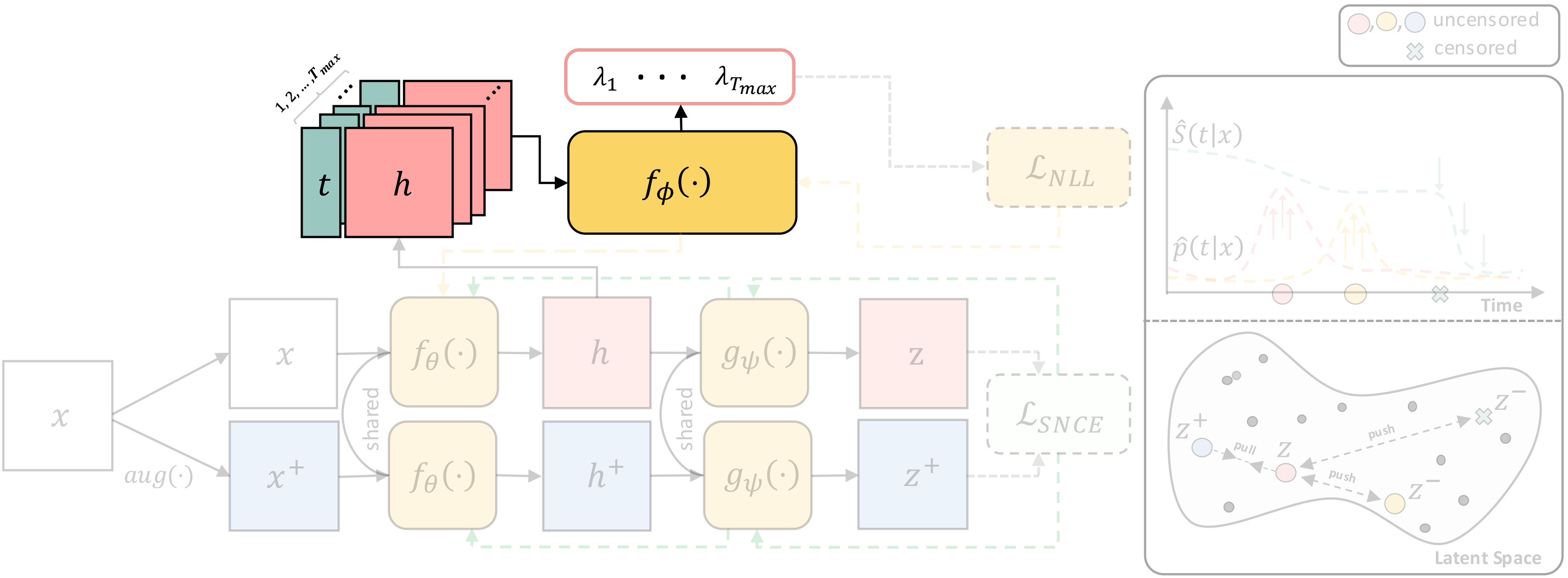
Proposed Method : Handling Right-Censoring

- Redefine the weight function considering the right-censoring as



Proposed Method : Network components

- **The hazard network.**, $f_\phi : \mathcal{H} \times \mathcal{T} \rightarrow [0,1]$, predicts the hazard rate at each time point $t \in \mathcal{T}$ given input latent representation $\mathbf{h}$, i.e., $\hat{\lambda}(t|\mathbf{x}) = f_\phi(\mathbf{h}, t) = f_\phi(f_\theta(\mathbf{x}), t)$



Important quantities : Hazard function

- The hazard function, $\lambda(t)$, is the instantaneous rate of failure at time t, given that an individual has survived until at least time t:

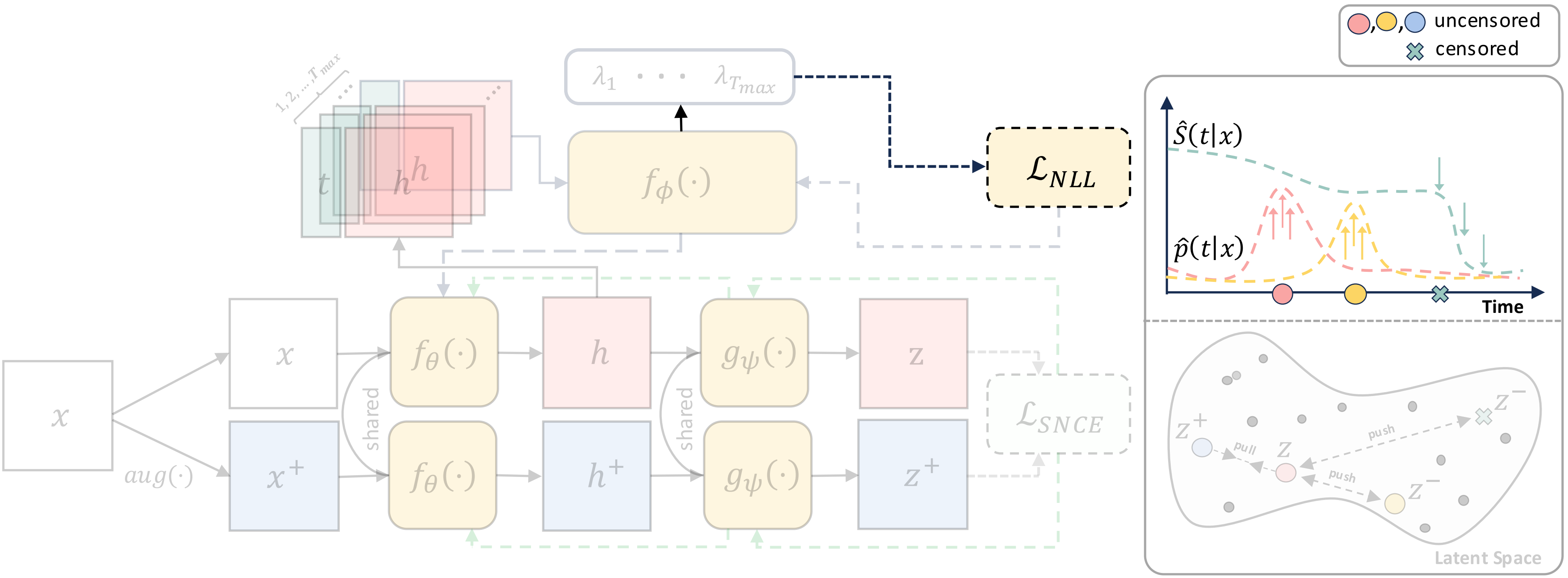
$$\lambda(t|\mathbf{x}) = P(T = t|T \geq t, \mathbf{x}) \text{ for } t \in \{1,2,\dots\}$$

- There is an important relationship between the survival and hazard functions:

$$\begin{aligned} S(t|\mathbf{x}) &= P(T > \mathrm{t} \mid \mathbf{x}) \\ &= P(T \neq 1 \mid \mathbf{x}) \cdot P(T \neq 2 \mid T > 1, \mathbf{x}) \cdots P(T = t \mid T > t-1, \mathbf{x}) \\ &= P(1 - \lambda(1 \mid \mathbf{x})) \cdot P(1 - \lambda(2 \mid \mathbf{x})) \cdot \cdots \cdot P(1 - \lambda(t \mid \mathbf{x})) \\ &= \prod_{t' \leq t} \left(1 - \lambda(t'|\mathbf{x})\right) \end{aligned}$$

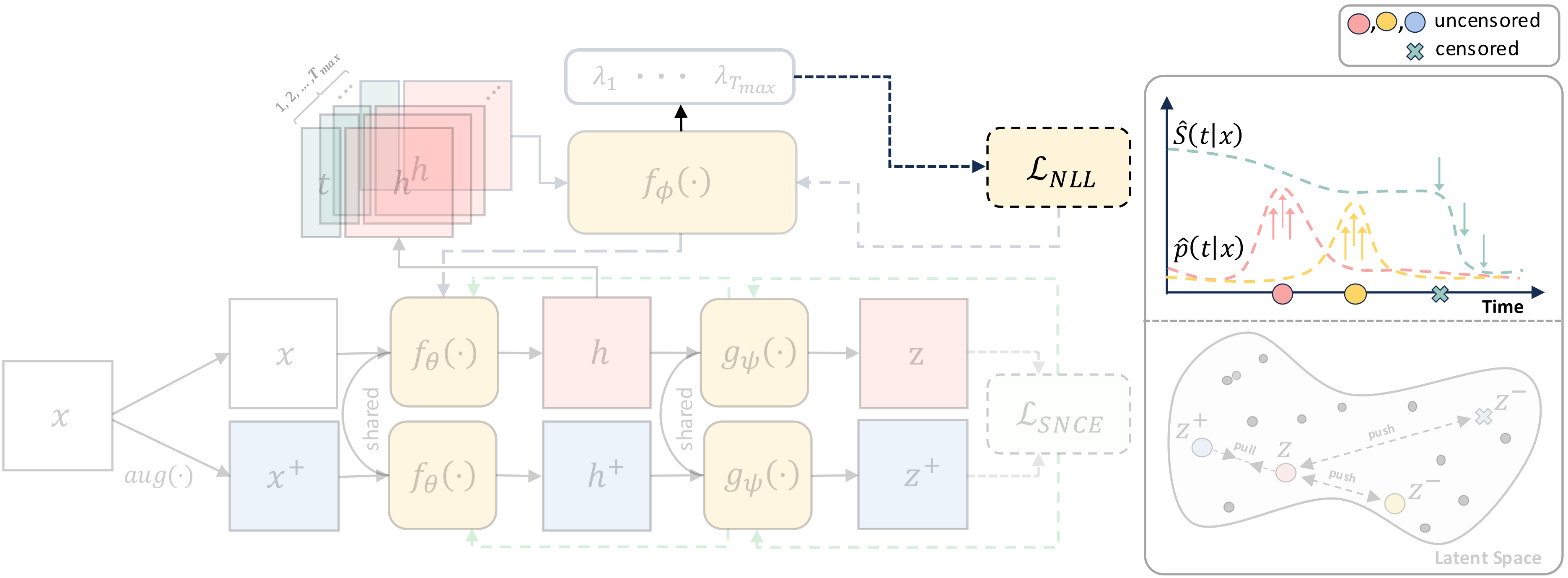
Proposed Method : Network components

- **Negative Log-likelihood**
  - Hazard estimate is defined as a function of time given an input feature, we can naturally model the time-varying effect of input features on risk/survival functions.



Proposed Method : Network components

- **Negative Log-likelihood**
  - Then, compute $\mathcal{L}_{NLL}^{\theta,\phi}$ by plugging in $f_\phi(f_\theta(\mathbf{x}), t)$ into $\hat{p}$ and $\hat{S}$
  - $\hat{p}(\tau|\mathbf{x}) = f_\phi(f_\theta(\mathbf{x}), \tau) \prod_{t' \le \tau - 1} \left(1 - f_\phi(f_\theta(\mathbf{x}), t')\right)$, $\hat{S}(\tau|\mathbf{x}) = \prod_{t' \le \tau} \left(1 - f_\phi(f_\theta(\mathbf{x}), t')\right)$



Proposed Method : Network components

- Overall, we can estimate the hazard function by training ConSurv with a loss function that combines the NLL loss and the SNCE loss, where β is a balancing coefficient

$$\mathcal{L}_{Total}^{\theta,\phi,\psi} = \mathcal{L}_{NLL}^{\theta,\phi} + \beta\mathcal{L}_{SNCE}^{\theta,\psi}$$

Experiments Setup : Datasets & Benchmarks & Metrics

○ Datasets

Dataset	No. Uncensored	No. Censored	No. Features (Real, Binary, Category)
METABRIC	888 (55.2%)	1093 (44.8%)	21 (6, 0, 15)
NWTCO	571 (14.2%)	3457 (85.5%)	6 (1, 4, 1)
GBSG	1267 (56.8%)	965 (43.2%)	7 (4, 2, 1)
FLCHAIN	4562 (69.9%)	1962 (30.3%)	8 (4, 2, 2)
SUPPORT	6036 (68.1%)	2837 (31.9%)	14 (8, 3, 3)
SEER	604 (1.11%)	53940 (98.9%)	12 (4, 5, 3)

○ Benchmarks

Loss Function	Type	Model
Partial Log-likelihood	ML	CoxPH
	DL	DeepSurv
Ranking Loss	DL	DeepHit
	DL	DRSA
Calibration Loss	DL	DCS
	DL	X-CAL

○ Metrics

Evaluation Metric	Type	Range	
Concordance Index (CI)	Discrimination	0.000 ~ 1.000	↑
Integrated Brier Score (IBS)	Calibration	0.000 ~ 1.000	↓
Distribution Divergence for Calibration (DDC)	Calibration	0.000 ~ 1.000	↓
D-calibration (D-CAL)	Calibration	P-value>0.05	×

Experiments : Quantitative Analysis

Method	METABRIC CI ↑	METABRIC IBS ↓	METABRIC DDC ↓	METABRIC D-CAL	NWTCO CI ↑	NWTCO IBS ↓	NWTCO DDC ↓	NWTCO D-CAL
CoxPH	$0.645_{\pm 0.019}$	$\mathbf{0.175}_{\pm \mathbf{0.028}}$	$0.111_{\pm 0.024}$	**25**	$0.716_{\pm 0.025}$	$0.108_{\pm 0.008}$	$0.515_{\pm 0.022}$	**25**
DeepSurv	$0.625_{\pm 0.025}$	$0.183_{\pm 0.029}$	$0.103_{\pm 0.026}$	**25**	$0.640_{\pm 0.080}$	$0.117_{\pm 0.011}$	$0.792_{\pm 0.011}$	24
DeepHit	$0.604_{\pm 0.019}$	$0.204_{\pm 0.018}$	$0.292_{\pm 0.017}$	0	$0.717_{\pm 0.028}$	$0.143_{\pm 0.024}$	$0.657_{\pm 0.024}$	12
DRSA	$0.604_{\pm 0.032}$	$0.249_{\pm 0.038}$	$0.178_{\pm 0.060}$	0	$0.709_{\pm 0.019}$	$0.281_{\pm 0.041}$	$0.218_{\pm 0.065}$	0
DCS	$0.612_{\pm 0.029}$	$0.206_{\pm 0.043}$	$\mathbf{0.054}_{\pm \mathbf{0.039}}$	2	$0.642_{\pm 0.036}$	$0.119_{\pm 0.018}$	$0.209_{\pm 0.043}$	19
X-CAL	$0.632_{\pm 0.027}$	$0.182_{\pm 0.023}$	$0.065_{\pm 0.037}$	2	$0.622_{\pm 0.037}$	$0.128_{\pm 0.025}$	$\mathbf{0.191}_{\pm \mathbf{0.079}}$	12
$\mathcal{L}_{NLL}$	$0.642_{\pm 0.022}$	$0.197_{\pm 0.030}$	$0.077_{\pm 0.020}$	13	$0.707_{\pm 0.024}$	$0.109_{\pm 0.008}$	$0.556_{\pm 0.041}$	23
$\mathcal{L}_{NLL}$ & $\mathcal{L}_{NCE}$	$0.659_{\pm 0.020}$	$0.193_{\pm 0.029}$	$0.080_{\pm 0.022}$	21	$0.715_{\pm 0.024}$	$0.108_{\pm 0.009}$	$0.563_{\pm 0.054}$	22
$\mathcal{L}_{NLL}$ & $\mathcal{L}_{Rank}$	$0.652_{\pm 0.022}$	$0.247_{\pm 0.030}$	$0.177_{\pm 0.020}$	0	$0.717_{\pm 0.027}$	$0.137_{\pm 0.008}$	$0.653_{\pm 0.050}$	0
ConSurv	$\mathbf{0.665}_{\pm \mathbf{0.023}}$	$0.186_{\pm 0.021}$	$0.110_{\pm 0.024}$	23	$\mathbf{0.718}_{\pm \mathbf{0.025}}$	$\mathbf{0.107}_{\pm \mathbf{0.008}}$	$0.554_{\pm 0.045}$	24

Method	SUPPORT CI ↑	SUPPORT IBS ↓	SUPPORT DDC ↓	SUPPORT D-CAL	SEER CI ↑	SEER IBS ↓	SEER DDC ↓	SEER D-CAL
CoxPH	$0.604_{\pm 0.006}$	$0.191_{\pm 0.005}$	$0.262_{\pm 0.013}$	0	$0.858_{\pm 0.018}$	$\mathbf{0.009}_{\pm \mathbf{0.005}}$	$0.966_{\pm 0.003}$	**25**
DeepSurv	$0.603_{\pm 0.090}$	$0.192_{\pm 0.007}$	$0.245_{\pm 0.036}$	0	$0.814_{\pm 0.020}$	$0.010_{\pm 0.000}$	$1.000_{\pm 0.000}$	**25**
DeepHit	$0.503_{\pm 0.009}$	$0.272_{\pm 0.003}$	$0.337_{\pm 0.006}$	0	$0.840_{\pm 0.033}$	$0.020_{\pm 0.001}$	$0.836_{\pm 0.003}$	0
DRSA	$0.570_{\pm 0.009}$	$0.259_{\pm 0.015}$	$0.486_{\pm 0.084}$	0	$0.834_{\pm 0.078}$	$0.021_{\pm 0.015}$	$\mathbf{0.671}_{\pm \mathbf{0.135}}$	0
DCS	$0.598_{\pm 0.008}$	$0.207_{\pm 0.012}$	$0.175_{\pm 0.032}$	0	$0.860_{\pm 0.020}$	$0.010_{\pm 0.001}$	$0.911_{\pm 0.044}$	21
X-CAL	$0.603_{\pm 0.007}$	$0.204_{\pm 0.012}$	$0.181_{\pm 0.025}$	0	$0.837_{\pm 0.040}$	$0.015_{\pm 0.006}$	$0.900_{\pm 0.049}$	18
$\mathcal{L}_{NLL}$	$0.606_{\pm 0.006}$	$0.193_{\pm 0.008}$	$0.123_{\pm 0.019}$	0	$0.854_{\pm 0.016}$	$0.009_{\pm 0.001}$	$0.867_{\pm 0.009}$	25
$\mathcal{L}_{NLL}$ & $\mathcal{L}_{NCE}$	$0.608_{\pm 0.007}$	$0.192_{\pm 0.007}$	$0.127_{\pm 0.021}$	0	$0.859_{\pm 0.017}$	$0.012_{\pm 0.003}$	$0.964_{\pm 0.006}$	25
$\mathcal{L}_{NLL}$ & $\mathcal{L}_{Rank}$	$0.617_{\pm 0.007}$	$0.173_{\pm 0.008}$	$0.231_{\pm 0.024}$	0	$0.862_{\pm 0.017}$	$0.139_{\pm 0.001}$	$1.000_{\pm 0.000}$	0
ConSurv	$\mathbf{0.616}_{\pm \mathbf{0.005}}$	$\mathbf{0.190}_{\pm \mathbf{0.006}}$	$\mathbf{0.148}_{\pm \mathbf{0.023}}$	0	$\mathbf{0.864}_{\pm \mathbf{0.016}}$	$\mathbf{0.009}_{\pm \mathbf{0.001}}$	$0.863_{\pm 0.006}$	**25**

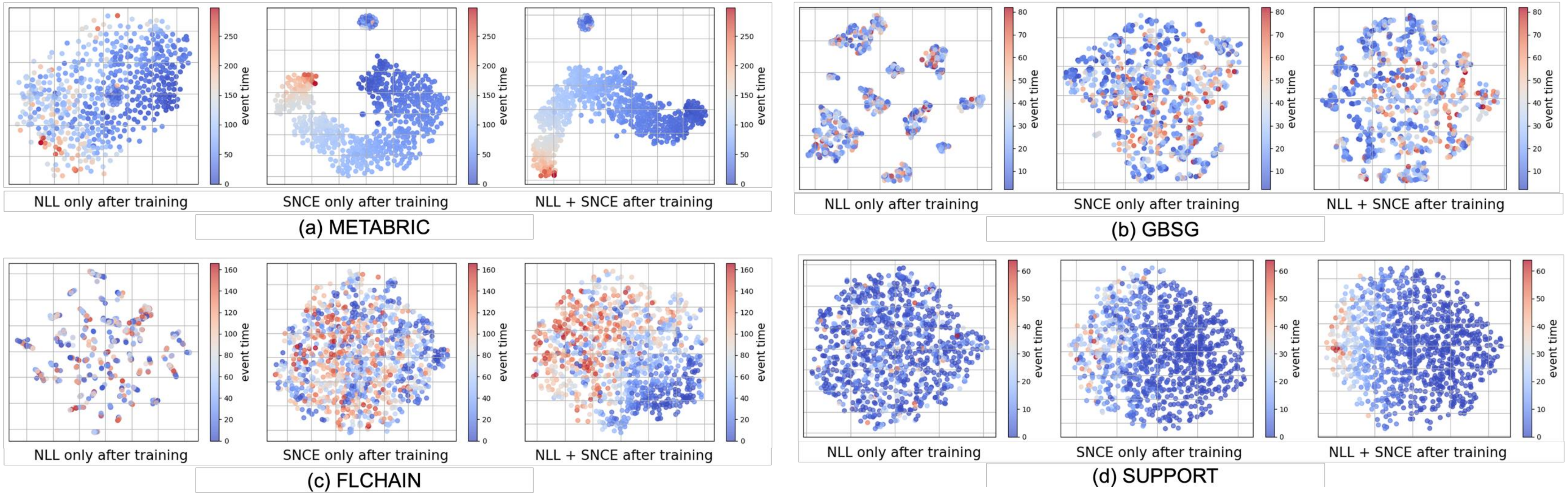
Experiments : Quantitative Analysis

METHOD	METABRIC CI ↑	METABRIC IBS ↓	METABRIC DDC ↓	METABRIC D-CAL	NWTCO CI ↑	NWTCO IBS ↓	NWTCO DDC ↓	NWTCO D-CAL
COXPH	$0.645_{\pm 0.019}$	$\mathbf{0.175}_{\pm \mathbf{0.028}}$	$0.111_{\pm 0.024}$	**25**	$0.716_{\pm 0.025}$	$0.108_{\pm 0.008}$	$0.515_{\pm 0.022}$	**25**
DEEPSURV	$0.625_{\pm 0.025}$	$0.183_{\pm 0.029}$	$0.103_{\pm 0.026}$	**25**	$0.640_{\pm 0.080}$	$0.117_{\pm 0.011}$	$0.792_{\pm 0.011}$	24
DEEPHIT	$0.604_{\pm 0.019}$	$0.204_{\pm 0.018}$	$0.292_{\pm 0.017}$	0	$0.717_{\pm 0.028}$	$0.143_{\pm 0.024}$	$0.657_{\pm 0.024}$	12
DRSA	$0.604_{\pm 0.032}$	$0.249_{\pm 0.038}$	$0.178_{\pm 0.060}$	0	$0.709_{\pm 0.019}$	$0.281_{\pm 0.041}$	$0.218_{\pm 0.065}$	0
DCS	$0.612_{\pm 0.029}$	$0.206_{\pm 0.043}$	$\mathbf{0.054}_{\pm \mathbf{0.039}}$	2	$0.642_{\pm 0.036}$	$0.119_{\pm 0.018}$	$0.209_{\pm 0.043}$	19
X-CAL	$0.632_{\pm 0.027}$	$0.182_{\pm 0.023}$	$0.065_{\pm 0.037}$	2	$0.622_{\pm 0.037}$	$0.128_{\pm 0.025}$	$\mathbf{0.191}_{\pm \mathbf{0.079}}$	12
$\mathcal{L}_{NLL}$	$0.642_{\pm 0.022}$	$0.197_{\pm 0.030}$	$0.077_{\pm 0.020}$	13	$0.707_{\pm 0.024}$	$0.109_{\pm 0.008}$	$0.556_{\pm 0.041}$	23
$\mathcal{L}_{NLL}$ & $\mathcal{L}_{NCE}$	$0.659_{\pm 0.020}$	$0.193_{\pm 0.029}$	$0.080_{\pm 0.022}$	21	$0.715_{\pm 0.024}$	$0.108_{\pm 0.009}$	$0.563_{\pm 0.054}$	22
$\mathcal{L}_{NLL}$ & $\mathcal{L}_{Rank}$	$0.652_{\pm 0.022}$	$0.247_{\pm 0.030}$	$0.177_{\pm 0.020}$	0	$0.717_{\pm 0.027}$	$0.137_{\pm 0.008}$	$0.653_{\pm 0.050}$	0
CONSURV	$\mathbf{0.665}_{\pm \mathbf{0.023}}$	$0.186_{\pm 0.021}$	$0.110_{\pm 0.024}$	23	$\mathbf{0.718}_{\pm \mathbf{0.025}}$	$\mathbf{0.107}_{\pm \mathbf{0.008}}$	$0.554_{\pm 0.045}$	24

METHOD	SUPPORT CI ↑	SUPPORT IBS ↓	SUPPORT DDC ↓	SUPPORT D-CAL	SEER CI ↑	SEER IBS ↓	SEER DDC ↓	SEER D-CAL
COXPH	$0.604_{\pm 0.006}$	$0.191_{\pm 0.005}$	$0.262_{\pm 0.013}$	0	$0.858_{\pm 0.018}$	$\mathbf{0.009}_{\pm \mathbf{0.005}}$	$0.966_{\pm 0.003}$	**25**
DEEPSURV	$0.603_{\pm 0.090}$	$0.192_{\pm 0.007}$	$0.245_{\pm 0.036}$	0	$0.814_{\pm 0.020}$	$0.010_{\pm 0.000}$	$1.000_{\pm 0.000}$	**25**
DEEPHIT	$0.503_{\pm 0.009}$	$0.272_{\pm 0.003}$	$0.337_{\pm 0.006}$	0	$0.840_{\pm 0.033}$	$0.020_{\pm 0.001}$	$0.836_{\pm 0.003}$	0
DRSA	$0.570_{\pm 0.009}$	$0.259_{\pm 0.015}$	$0.486_{\pm 0.084}$	0	$0.834_{\pm 0.078}$	$0.021_{\pm 0.015}$	$\mathbf{0.671}_{\pm \mathbf{0.135}}$	0
DCS	$0.598_{\pm 0.008}$	$0.207_{\pm 0.012}$	$0.175_{\pm 0.032}$	0	$0.860_{\pm 0.020}$	$0.010_{\pm 0.001}$	$0.911_{\pm 0.044}$	21
X-CAL	$0.603_{\pm 0.007}$	$0.204_{\pm 0.012}$	$0.181_{\pm 0.025}$	0	$0.837_{\pm 0.040}$	$0.015_{\pm 0.006}$	$0.900_{\pm 0.049}$	18
$\mathcal{L}_{NLL}$	$0.606_{\pm 0.006}$	$0.193_{\pm 0.008}$	$0.123_{\pm 0.019}$	0	$0.854_{\pm 0.016}$	$0.009_{\pm 0.001}$	$0.867_{\pm 0.009}$	25
$\mathcal{L}_{NLL}$ & $\mathcal{L}_{NCE}$	$0.608_{\pm 0.007}$	$0.192_{\pm 0.007}$	$0.127_{\pm 0.021}$	0	$0.859_{\pm 0.017}$	$0.012_{\pm 0.003}$	$0.964_{\pm 0.006}$	25
$\mathcal{L}_{NLL}$ & $\mathcal{L}_{Rank}$	$0.617_{\pm 0.007}$	$0.173_{\pm 0.008}$	$0.231_{\pm 0.024}$	0	$0.862_{\pm 0.017}$	$0.139_{\pm 0.001}$	$1.000_{\pm 0.000}$	0
CONSURV	$\mathbf{0.616}_{\pm \mathbf{0.005}}$	$\mathbf{0.190}_{\pm \mathbf{0.006}}$	$\mathbf{0.148}_{\pm \mathbf{0.023}}$	0	$\mathbf{0.864}_{\pm \mathbf{0.016}}$	$\mathbf{0.009}_{\pm \mathbf{0.001}}$	$0.863_{\pm 0.006}$	**25**

Experiments : Qualitative Analysis

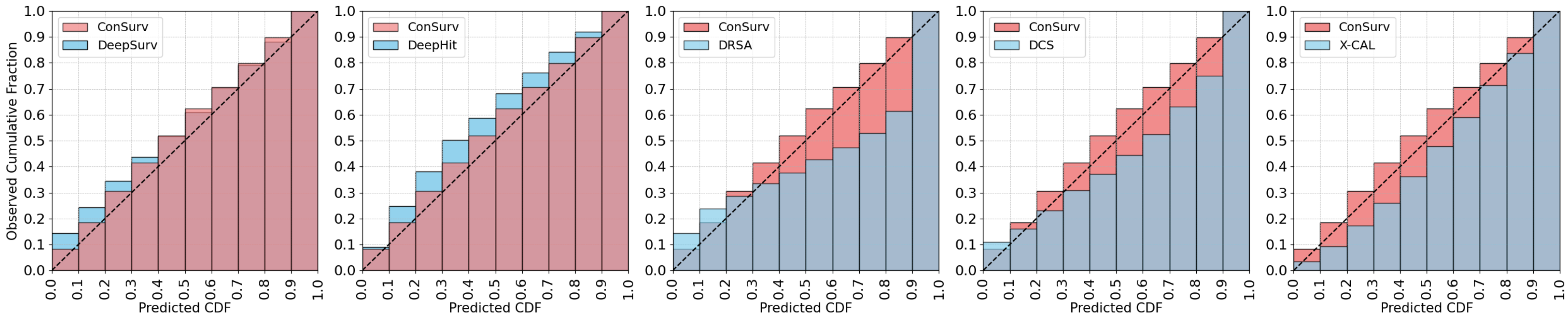
- Effect of Contrastive Learning
  - $\mathcal{L}_{NLL}$ only, $\mathcal{L}_{SNCE}$ only, and ConSurv (i.e., $\mathcal{L}_{NLL}$ & $\mathcal{L}_{SNCE}$)
  - significantly improves the alignment of representations with event time information



(a) METABRIC (b) GBSG (c) FLCHAIN (d) SUPPORT

Experiments : Qualitative Analysis

- Comparing calibration plot of ConSurv with the DL-based survival models
  - The x=y line represents the ideal state where predicted probabilities perfectly match the observed outcome



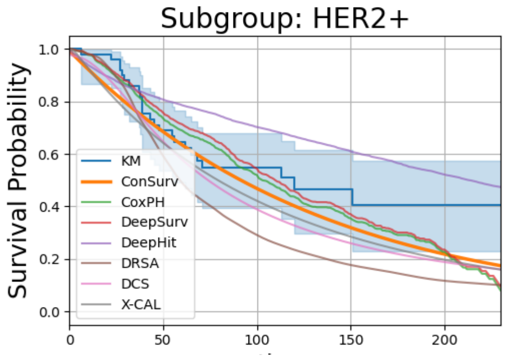
(a) vs DeepSurv (b) vs DeepHit (c) vs DRSA (d) vs DCS (e) vs X-CAL

Experiments : Qualitative Analysis

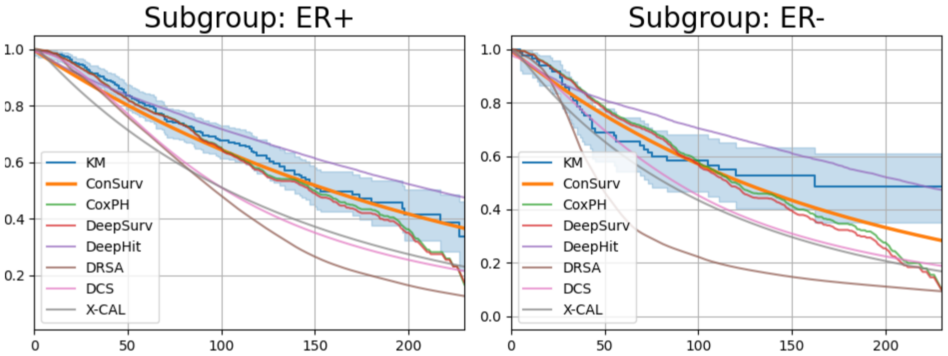
- **Subgroup Analysis**
  - To confirm the calibration performance of survival models, compare their survival plots with the Kaplan-Meier (KM) curve
  - KM curve provides a non-parametric estimate of survival function at population level
  - Examine three binary hormone receptor status in the METABRIC dataset: estrogen receptor (ER) ,human epidermal growth factor receptor 2 (HER2), and progesterone receptor (PR)

Experiments : Qualitative Analysis

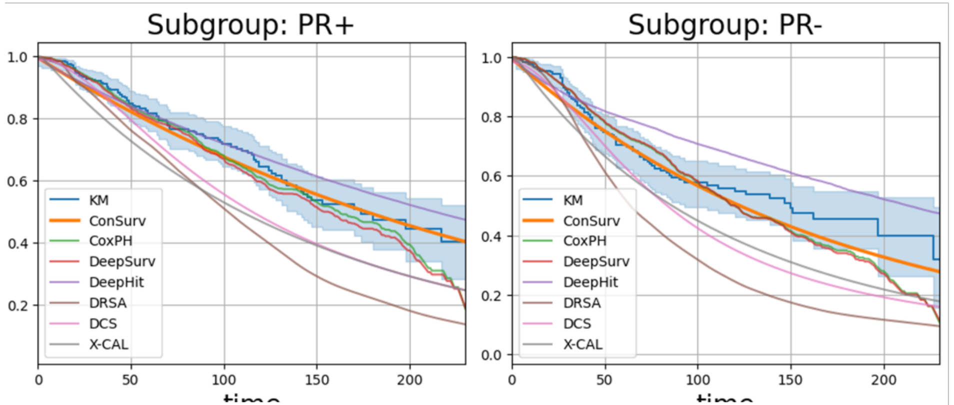
- **Subgroup Analysis**



(a) HER2



(b) ER



(c) PR

Experiments : Qualitative Analysis

- **Subgroup Analysis**
  - To quantitatively assess calibration performance, compare the survival predictions of each model with the KM curves for each subgroup using the Wasserstein distance

Subgroup	ER		HER2		Size		PR	
	+	-	+	-	+	-	+	-
CoxPH	0.030	0.108	0.063	**0.089**	0.049	0.067	0.076	0.044
DeepSurv	0.033	0.115	0.066	0.101	0.054	0.074	0.105	0.089
DeepHit	0.063	0.082	0.146	0.156	**0.043**	0.102	0.233	0.033
DRSA	0.181	0.293	0.233	0.328	0.205	0.276	0.118	0.234
DCS	0.130	0.146	0.087	0.178	0.154	0.126	0.091	0.124
X-CAL	0.136	0.165	0.105	0.180	0.159	0.148	**0.060**	0.176
ConSurv	**0.024**	**0.077**	**0.044**	**0.089**	**0.043**	**0.042**	**0.060**	**0.025**

Discussion & Future works

- Survival data often lacks clear event times complicating learning due to censored data.
  - Potential avenues include modifying models to account for uncertainties or developing alternative learning approaches
  - Need for new evaluation metrics that consider the characteristics
- Limited augmentation techniques in tabular data reduce model robustness
  - Explore augmentation methods suitable for survival datasets to enhance contrastive learning performance

Thank you