ICML | 2023
Thirty-ninth International Conference on Machine Learning

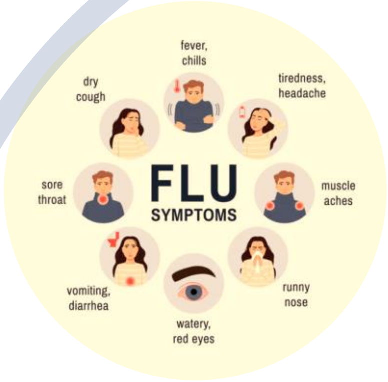
Stable Estimation of Heterogeneous Treatment Effects

Anpeng Wu[1✉], Kun Kuang[1*], Ruoxuan Xiong[2], Bo Li[3], Fei Wu[145]

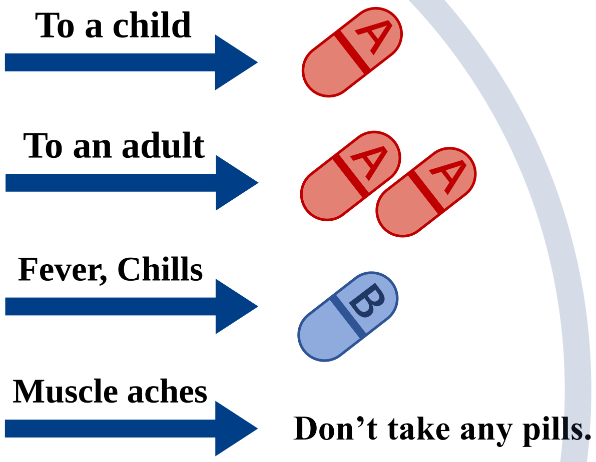
1 Department of Computer Science and Technology, Zhejiang University, Hangzhou, China
2 Department of Quantitative Theory and Methods, Emory University, Atlanta, USA
3 School of Economics and Management, Tsinghua University, Peking, China
4 Shanghai Institute for Advanced Study, Zhejiang University, Shanghai, China
5 Shanghai AI Laboratory, Shanghai, China
✉anpwu@zju.edu.cn
*Corresponding author

A healthcare example.

Personalized Treatment

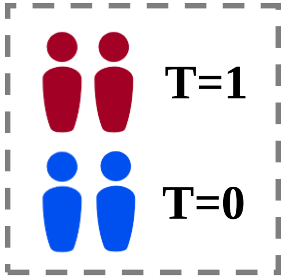


Individualized Symptoms
Coughs and Headaches

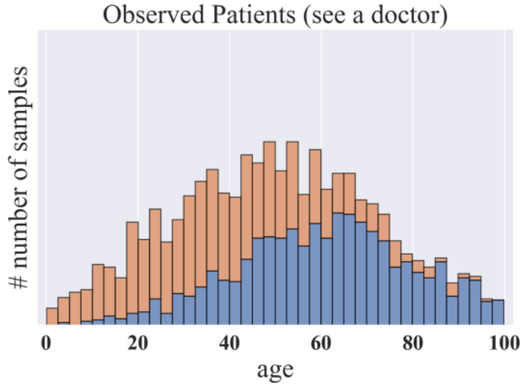


Heterogeneous Treatment Effect:
The treatment outcomes can vary among individuals or subgroups due to factors like individual characteristics, genetics, or environment, emphasizing that interventions may not have the same impact on everyone.

- **Confounding Bias from Imbalanced Treatment Assignment**



Population

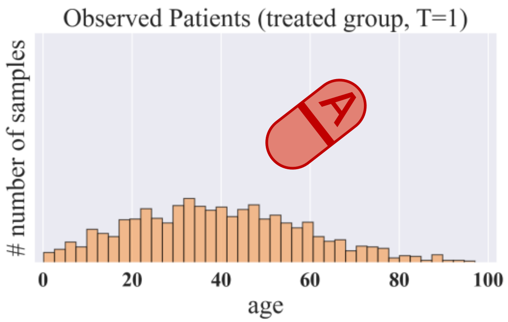


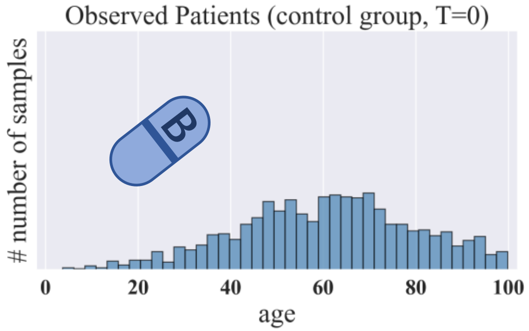
The treatment is a choice, not a random assignment.

X: The distribution of observed patients on the age attribute.

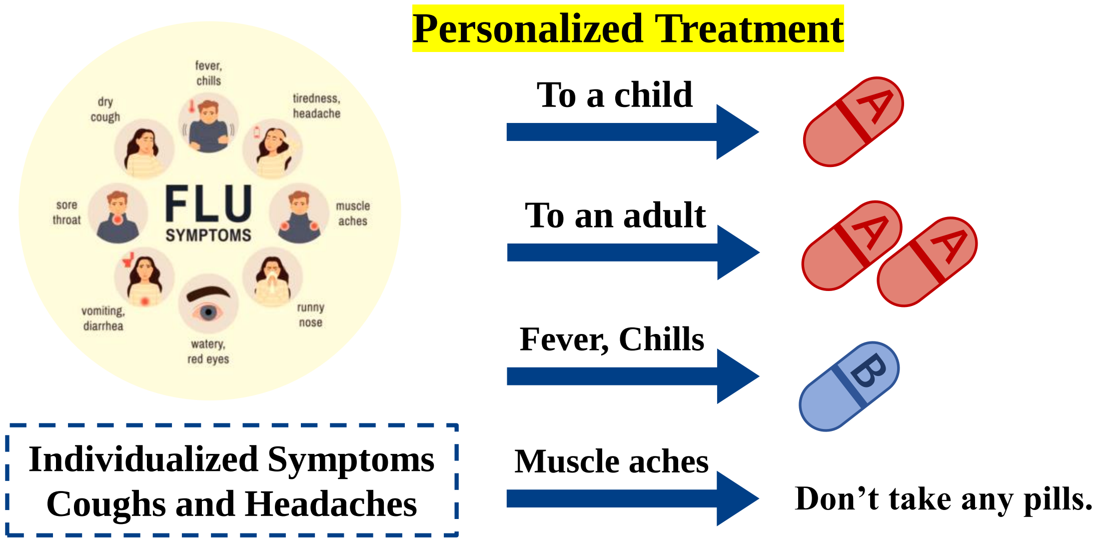
Treatment Choice

1. Cost: I cannot afford it.
2. I don't like the bitter taste.
3. Concern about side effects.





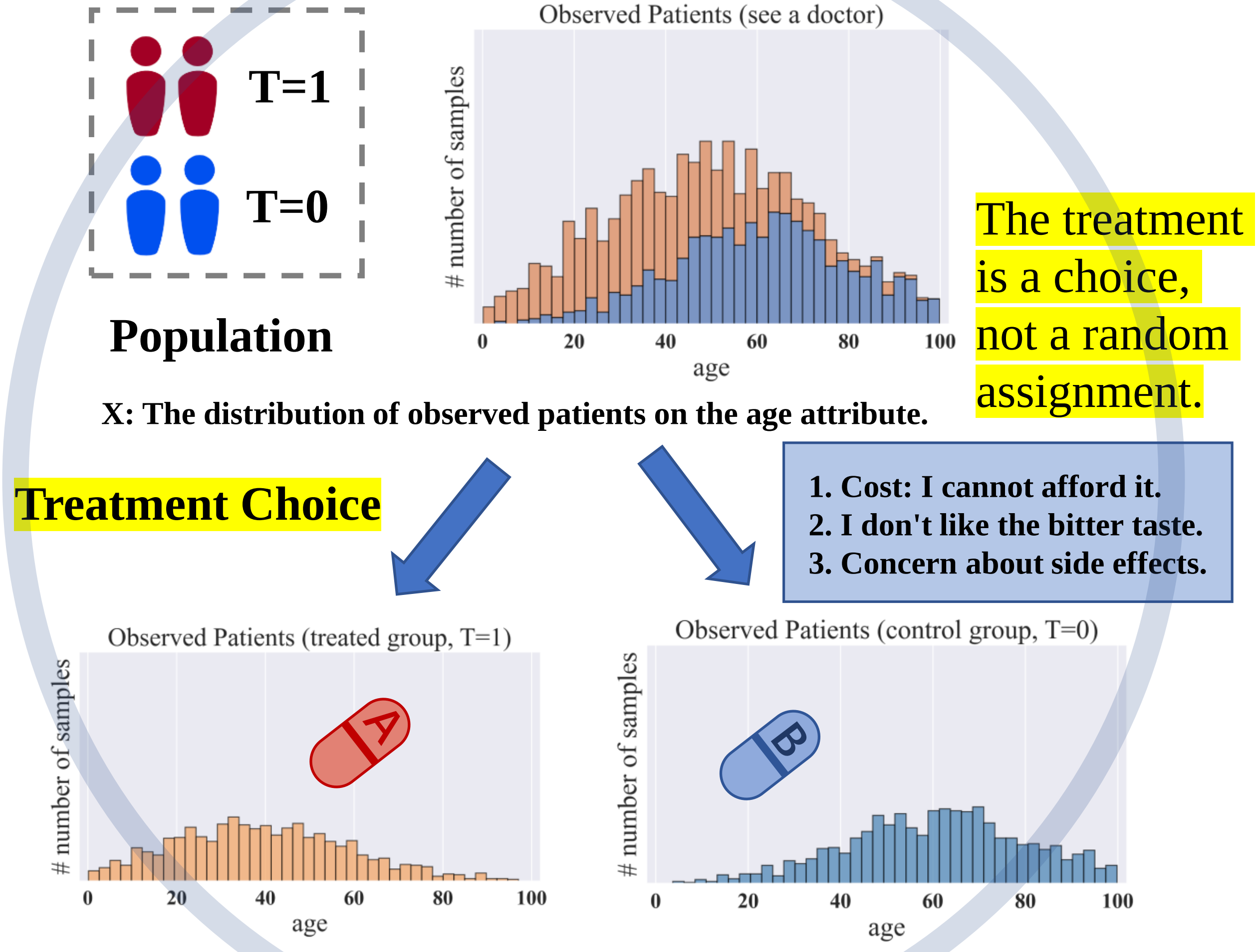
A healthcare example.

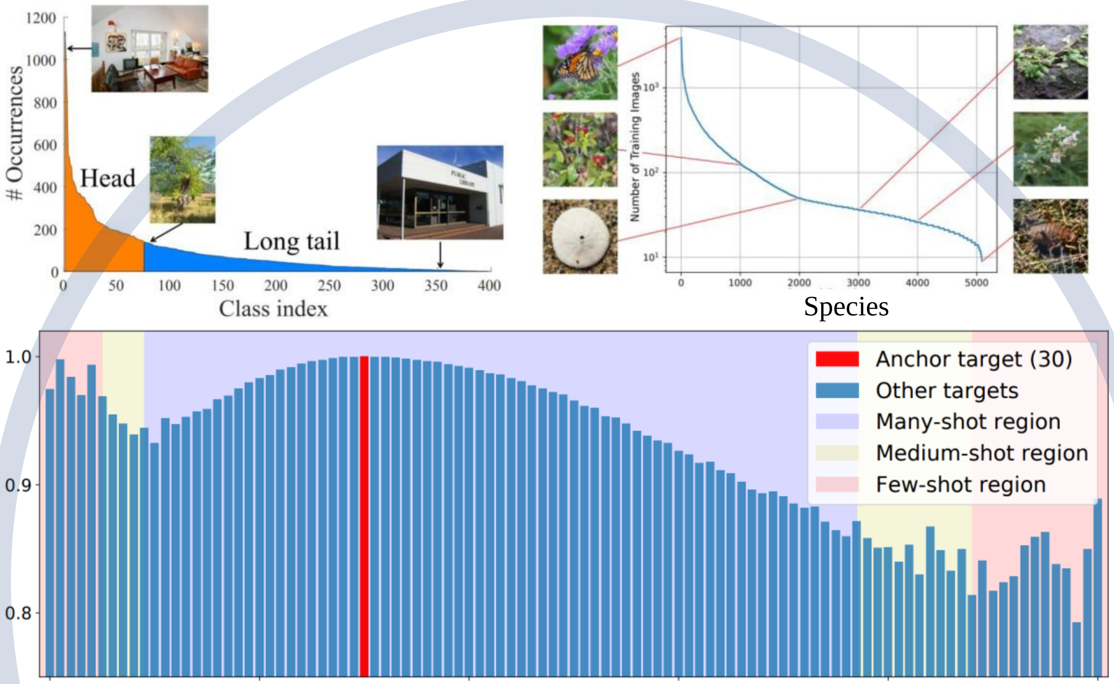


Heterogeneous Treatment Effect:

The treatment outcomes can vary among individuals or subgroups due to factors like individual characteristics, genetics, or environment, emphasizing that interventions may not have the same impact on everyone.

- **Confounding Bias from Imbalanced Treatment Assignment**

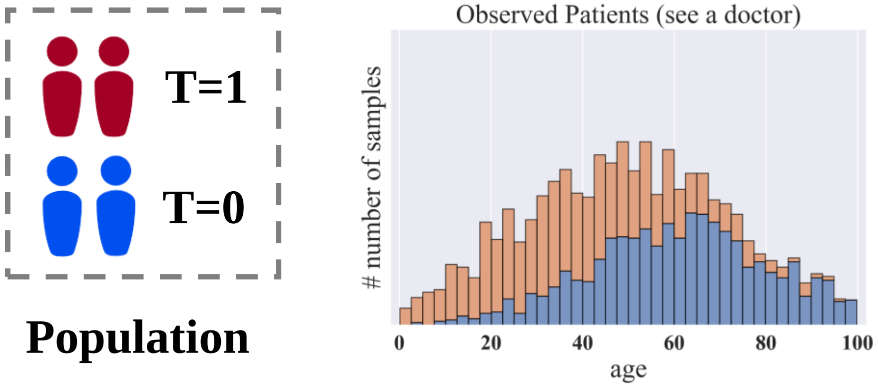




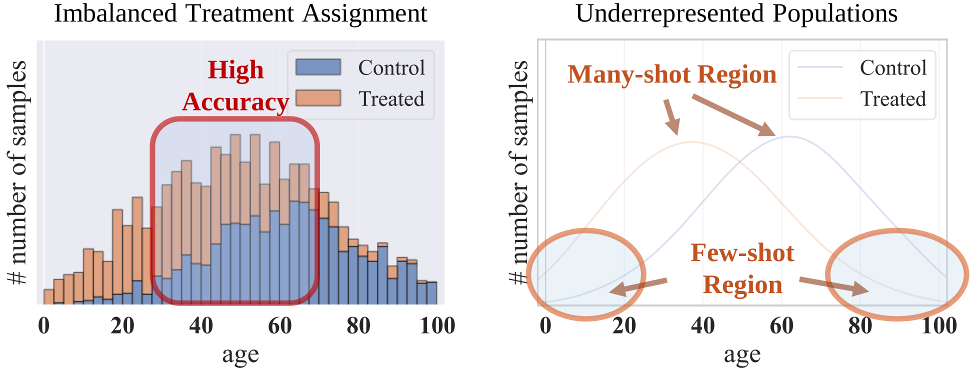
Figures in [Yang et al., 2021]

- **Data imbalance is ubiquitous and prevalent in many real-life applications.**
- In biomedical applications, e-markets, and social media, observational datasets are typically constructed by pooling from multiple sources or from certain time periods. This raises concerns about the sample representativeness in some sample spaces.

- **Confounding Bias from Imbalanced Treatment Assignment**

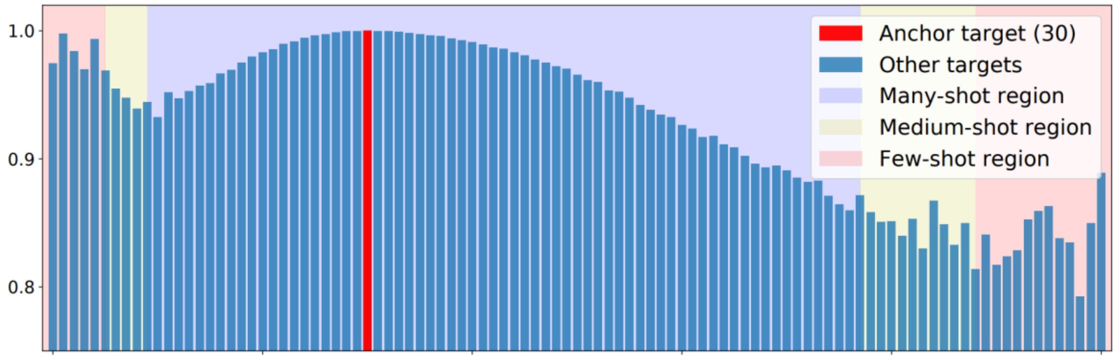


X: The distribution of observed patients on the age attribute.



Observational Samples from Underrepresented Populations

[Yang et al., 2021] Yang, Yuzhe, et al. "Delving into deep imbalanced regression." International Conference on Machine Learning. PMLR, 2021.



Figures in [Yang et al., 2021]

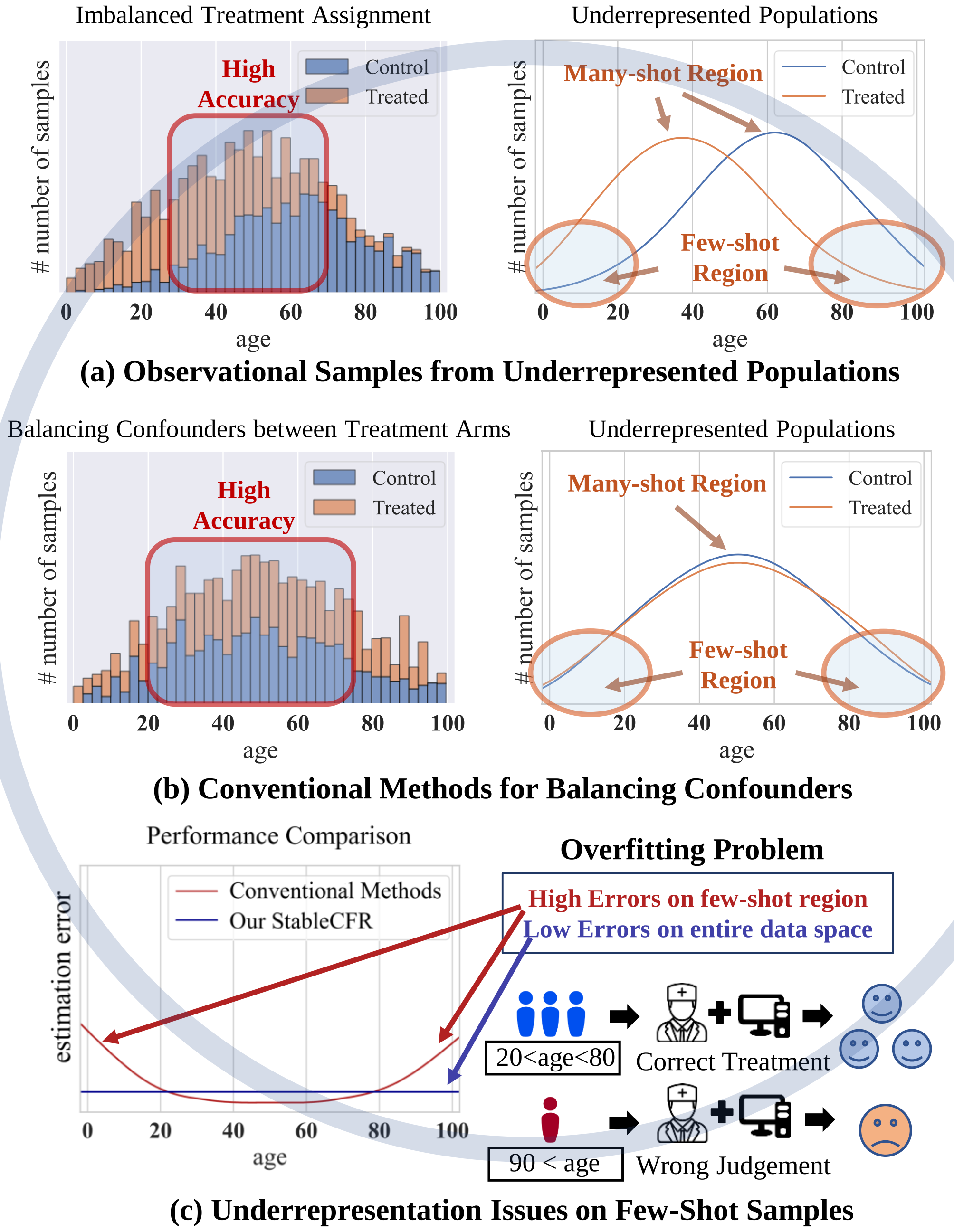
- Previous regression models prioritize improving the average HTE performance by minimizing *mean square error (MSE)*:

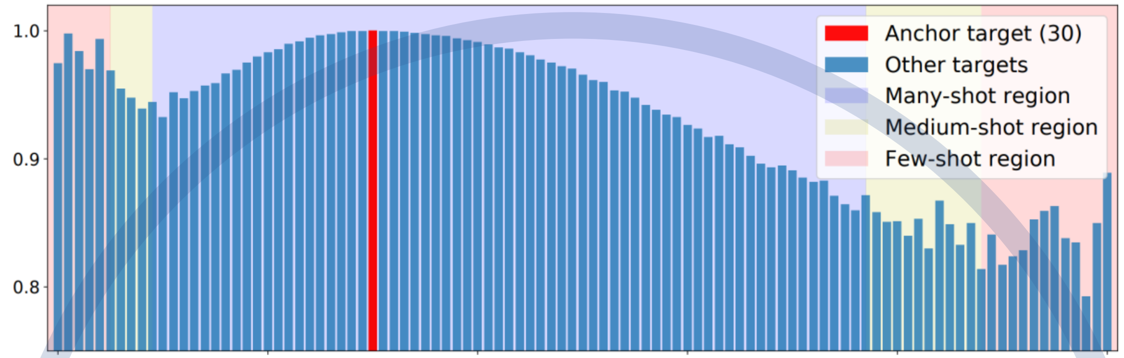
$$\min \int_{\boldsymbol{x}\in\mathcal{X}} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2 dF_X(\boldsymbol{x}), \qquad (1)$$

- Stable HTE estimator that re-samples the underrepresented data using uniform sampling to *assign weights proportional to the Lebesgue measure of the support of each subpopulation*:

$$\min \int_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{|\mathcal{X}|} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2 d\boldsymbol{x}, \qquad (2)$$

[Yang et al., 2021] Yang, Yuzhe, et al. "Delving into deep imbalanced regression." International Conference on Machine Learning. PMLR, 2021.





Figures in [Yang et al., 2021]

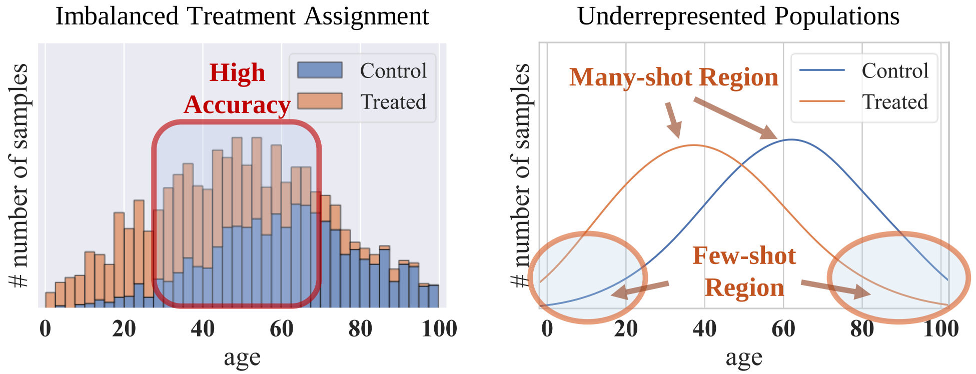
- Previous regression models prioritize improving the average HTE performance by minimizing *mean square error (MSE):*

$$\min \int_{\boldsymbol{x}\in\mathcal{X}} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2 dF_X(\boldsymbol{x}), \quad (1)$$

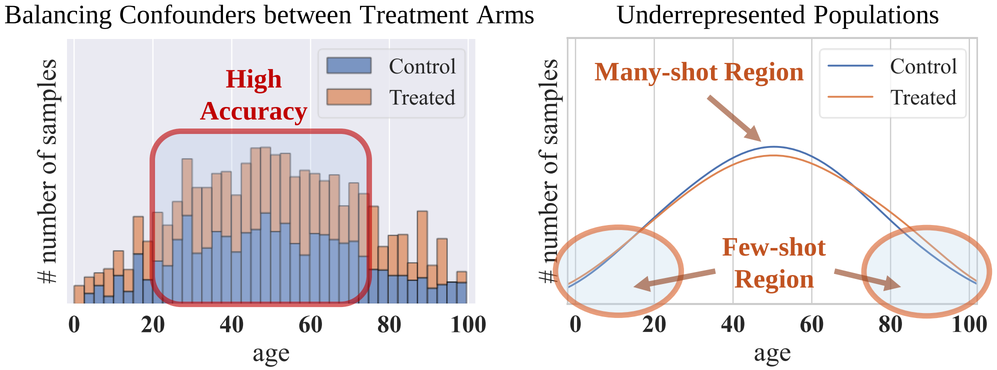
- Stable HTE estimator that re-samples the underrepresented data using uniform sampling to *assign weights proportional to the Lebesgue measure of the support of each subpopulation:*

$$\min \int_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{|\mathcal{X}|} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2 d\boldsymbol{x}, \quad (2)$$

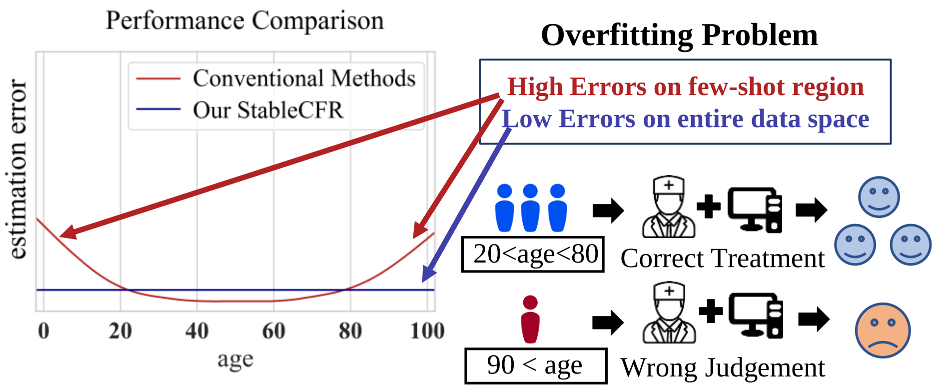
[Yang et al., 2021] Yang, Yuzhe, et al. "Delving into deep imbalanced regression." International Conference on Machine Learning. PMLR, 2021.



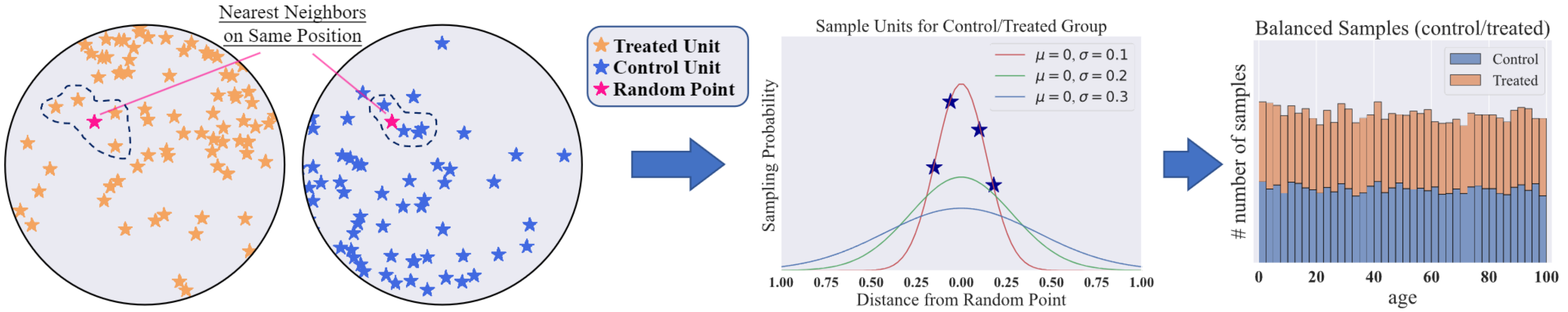
(a) Observational Samples from Underrepresented Populations



(b) Conventional Methods for Balancing Confounders



(c) Underrepresentation Issues on Few-Shot Samples



(a) Sample Point from an Ideal Uniform Distribution **(b) Sample Units from Nearest Neighbors** **(c) Deep Balance Data**

Figure 2. Stable CounterFacutal Regression Architecture.

Sample Points from an Ideal Uniform Distribution.

$$\boldsymbol{Q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \cdots, \boldsymbol{q_m}\} \sim Unif(\boldsymbol{Q}_{min}, \boldsymbol{Q}_{max}), \quad (3)$$
$$\boldsymbol{Q}_{min} = \min(\boldsymbol{X}) - 0.01\text{range}(\boldsymbol{X}), \quad (4)$$
$$\boldsymbol{Q}_{max} = \max(\boldsymbol{X}) + 0.01\text{range}(\boldsymbol{X}), \quad (5)$$

where $\min(\boldsymbol{X})$ and $\max(\boldsymbol{X})$ represent the minimum and maximum values of the covariate $\boldsymbol{X}$ in the training sample, respectively. $\text{range}(\boldsymbol{X})$ represents the range size of the covariates, i.e., $\text{range}(\boldsymbol{X}) = \max(\boldsymbol{X}) - \min(\boldsymbol{X})$.

Match Pairs from Nearest Neighbors.

$$I_j^{t=0} = \arg\min_{i:t_i=0} \|\boldsymbol{x_i} - \boldsymbol{q_j}\|_2^2, \quad (6)$$
$$I_j^{t=1} = \arg\min_{i:t_i=1} \|\boldsymbol{x_i} - \boldsymbol{q_j}\|_2^2, \quad (7)$$

Epsilon-Greedy Match with Distance-based Sampling. Although pure uniform sampling with nearest neighbor samples has balanced the treated and control groups, the greedy strategy of searching for nearest neighbors for matching may lead to frequent sampling of sparse samples in under-represented populations, hurting the model's performance. Therefore, we select the top-K nearest neighbors and use a hyperparameter ϵ to trade-off exploration (distance-based sampling, with the receptive field controlled by hyperparameter σ) and exploitation (top-K nearest neighbor sampling).

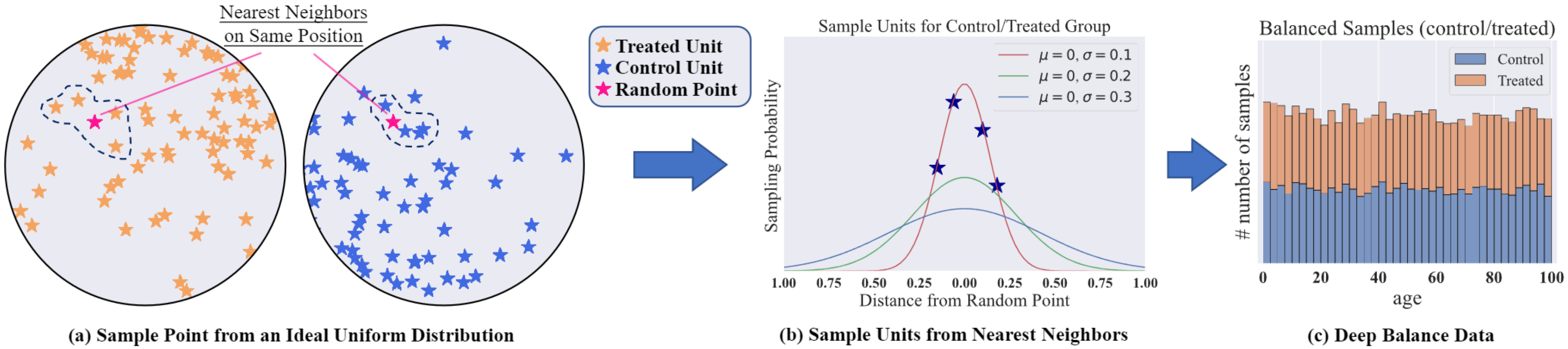


Figure 2. Stable CounterFacutal Regression Architecture.

Sample Points from an Ideal Uniform Distribution.

$$\boldsymbol{Q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \cdots, \boldsymbol{q_m}\} \sim Unif(\boldsymbol{Q}_{min}, \boldsymbol{Q}_{max}), \quad (3)$$
$$\boldsymbol{Q}_{min} = \min(\boldsymbol{X}) - 0.01\text{range}(\boldsymbol{X}), \quad (4)$$
$$\boldsymbol{Q}_{max} = \max(\boldsymbol{X}) + 0.01\text{range}(\boldsymbol{X}), \quad (5)$$

where $\min(\boldsymbol{X})$ and $\max(\boldsymbol{X})$ represent the minimum and maximum values of the covariate $\boldsymbol{X}$ in the training sample, respectively. $\text{range}(\boldsymbol{X})$ represents the range size of the covariates, i.e., $\text{range}(\boldsymbol{X}) = \max(\boldsymbol{X}) - \min(\boldsymbol{X})$.

Match Pairs from Nearest Neighbors.

$$I_j^{t=0} = \arg\min_{i:t_i=0} \|\boldsymbol{x_i} - \boldsymbol{q_j}\|_2^2, \quad (6)$$
$$I_j^{t=1} = \arg\min_{i:t_i=1} \|\boldsymbol{x_i} - \boldsymbol{q_j}\|_2^2, \quad (7)$$

Epsilon-Greedy Match with Distance-based Sampling. Although pure uniform sampling with nearest neighbor samples has balanced the treated and control groups, the greedy strategy of searching for nearest neighbors for matching may lead to frequent sampling of sparse samples in underrepresented populations, hurting the model's performance. Therefore, we select the top-K nearest neighbors and use a hyperparameter ϵ to trade-off exploration (distance-based sampling, with the receptive field controlled by hyperparameter σ) and exploitation (top-K nearest neighbor sampling).

Simulation Data

Syn-γ ($\gamma = 0.5, 0.8, 1.0$): Under the unconfoundedness assumption, we sample 3,000 units with an 80/20 proportion of training/validation splits, and use skewed distributions with locally undersampled data to generate the covariates $\boldsymbol{X} = \{X_1, X_2, X_3\}$ where X_i denotes i-th variable in $\boldsymbol{X}$:

$$X_1 \sim \mathcal{N}(-2.0, 2.0 \mid 0, \gamma), \quad (8)$$
$$X_2 \sim \mathcal{N}(-0.1, 2.0 \mid 0, \gamma), \quad (9)$$
$$X_3 \sim \mathcal{N}(-2.0, 0.1 \mid 0, \gamma), \quad (10)$$
$$Pr(T \mid \boldsymbol{X}) = \frac{1}{1 + \exp\left(X_1 + X_2 + X_3\right)/3},$$
$$T \sim Bernoulli(Pr(T \mid \boldsymbol{X})). \quad (11)$$

The outcome variable $Y(T)$ is generated as follows:

$$Y(T) = TY(1) + (1 - T)Y(0), \quad (12)$$
$$Y(0) = |X_2^2 - X_3^2| + 2\cos\left(X_1 - X_2 + X_3\right) - s(\boldsymbol{X}),$$
$$Y(1) = |X_2^2 - X_3^2| + 2\sin\left(X_1 - X_2 + X_3\right) + s(\boldsymbol{X}),$$
$$s(X) = X_1 X_2 X_3 + X_1^2 + (1 - X_2 + X_3)^2.$$

Real-Wrold Data

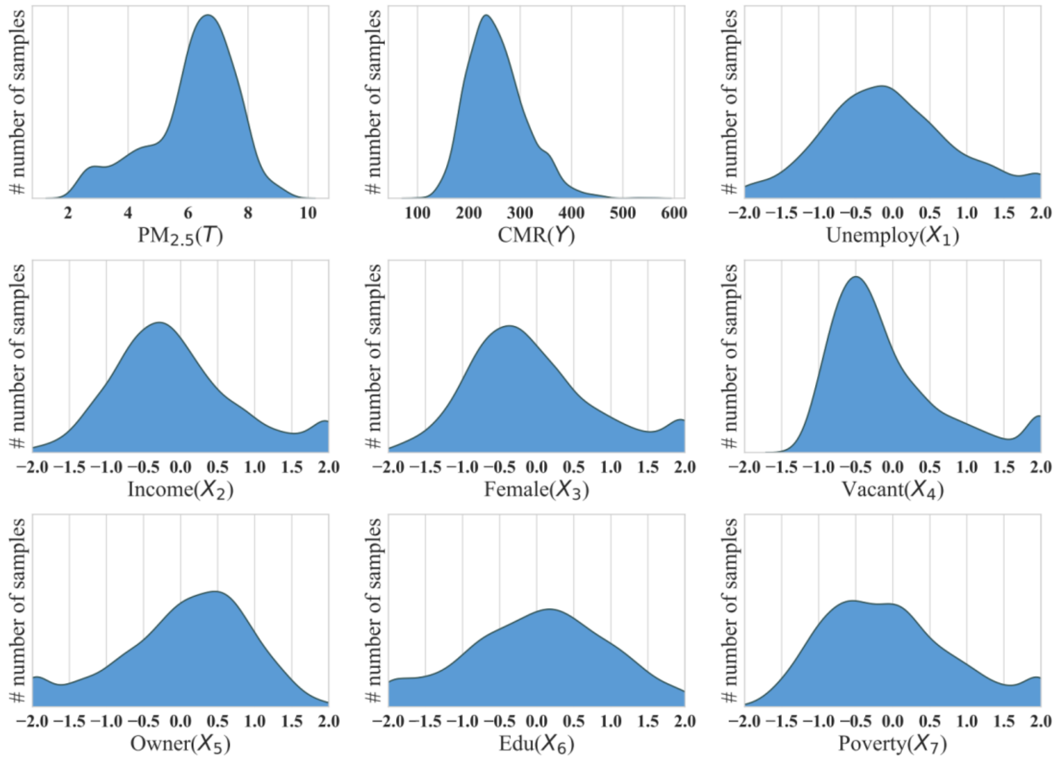


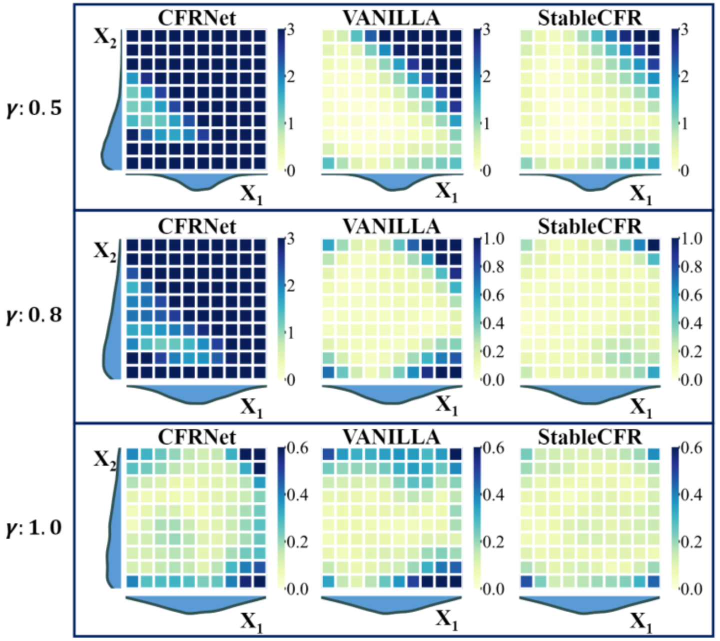
Figure 5. The Data Distribution on PM-CMR.

$$\min \int_{\boldsymbol{x} \in \mathcal{X}} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2 dF_X(\boldsymbol{x}), \quad (1)$$

Target: $$\min \int_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2 d\boldsymbol{x}, \quad (2)$$

Table 1. The Generalization Experiments for Heterogeneous Treatment Effect on Syn-0.8 dataset.

	ID data (training distribution)				OOD data (uniform distribution)					
Method	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MAXE	$P_{\downarrow 0.3}$
VANILLA	**0.006(0.002)**	**0.005(0.004)**	**0.089(0.020)**	**0.003(0.019)**	0.139(0.043)	0.069(0.036)	0.500(0.113)	0.047(0.056)	10.044(1.745)	84.8%
Reweight	0.026(0.007)	0.353(0.246)	0.609(0.170)	0.055(0.069)	0.197(0.071)	1.008(0.319)	1.205(0.140)	0.155(0.114)	14.508(1.455)	31.4%
BMC	1.263(0.174)	2.246(0.322)	1.782(0.102)	0.499(0.108)	1.565(0.252)	0.816(0.146)	1.574(0.109)	0.361(0.123)	10.128(1.225)	8.1%
GAI	0.315(0.028)	0.245(0.045)	0.919(0.045)	0.032(0.020)	1.140(0.203)	1.462(0.513)	2.136(0.208)	1.323(0.148)	6.329(0.917)	19.1%
DIRNet	0.025(0.003)	0.073(0.012)	0.294(0.023)	0.003(0.023)	0.188(0.051)	0.465(0.119)	0.860(0.105)	0.025(0.042)	12.698(1.322)	52.5%
IPWNet	0.007(0.002)	0.009(0.004)	0.105(0.018)	0.003(0.021)	0.140(0.053)	0.095(0.044)	0.546(0.093)	0.031(0.089)	10.220(1.561)	80.5%
CFRNet	2.517(1.653)	2.634(1.731)	2.370(1.482)	0.124(0.103)	10.955(7.129)	16.377(10.66)	5.990(3.702)	3.582(2.362)	28.930(16.81)	26.9%
SITE	3.491(1.769)	1.933(1.093)	2.741(0.999)	0.058(0.060)	19.208(8.076)	11.341(6.235)	7.008(2.078)	4.086(1.399)	36.735(5.464)	7.4%
CFRISW	0.953(1.337)	0.875(1.359)	1.070(1.142)	0.047(0.069)	3.950(5.892)	5.400(9.204)	2.560(2.996)	1.212(2.010)	16.136(12.82)	37.6%
DRCFR	0.077(0.024)	0.065(0.012)	0.249(0.027)	0.008(0.039)	0.446(0.121)	0.315(0.091)	0.683(0.146)	0.007(0.083)	8.306(2.345)	60.4%
DERCFR	0.021(0.013)	0.020(0.007)	0.192(0.037)	0.007(0.065)	0.200(0.076)	0.223(0.103)	0.634(0.114)	0.128(0.144)	7.628(1.908)	64.4%
StableCFR	0.008(0.003)	0.006(0.003)	0.099(0.017)	0.009(0.017)	**0.046(0.020)**	**0.037(0.017)**	**0.299(0.054)**	**0.004(0.042)**	**5.889(1.577)**	**90.8%**



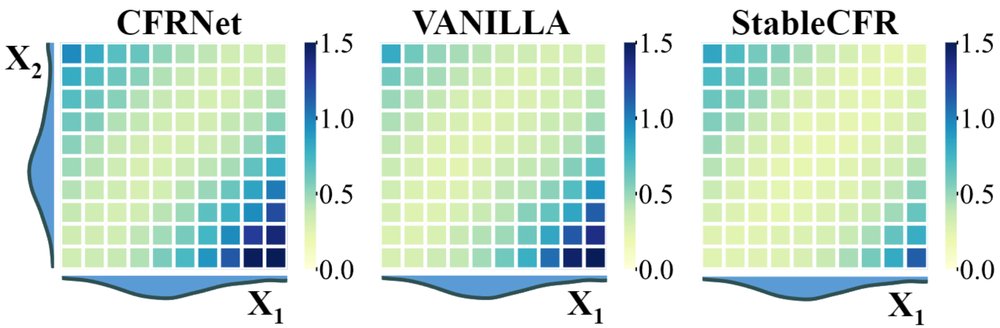


Figure 4. The PEHE Results on Semi-PM-CMR Dataset.

A lighter color represents a lower estimation error.

Thanks

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (62006207, 62037001, U20A20387), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Zhejiang Province Natural Science Foundation (LQ21F020020), Project by Shanghai AI Laboratory (P22KS00111), Program of Zhejiang Province Science and Technology (2022C01044), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010), and the Fundamental Research Funds for the Central Universities (226-2022-00142, 226-2022-00051). Bo Li's research was supported by the National Natural Science Foundation of China (No.72171131, 72133002); the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403.