CAFE: Causally-Guided Automated Feature Engineering with Multi-Agent Reinforcement Learning

Arun Vignesh Malarkkan[1,*], Wangyang Ying[1], Hongyu Cao[2], Dongjie Wang[3], Yanjie Fu[1]

[1]Arizona State University [2]Clemson University [3]Northeast Normal University
*arun.malarkkan@asu.edu

github.com/arunvignesh28/CAFE

NEURAL INFORMATION PROCESSING SYSTEMS

1 How CAFE works

Search for the feature set F* with the best validation utility; a learned causal graph biases the search, but not as a hard constraint.

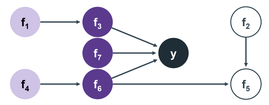
PHASE I **Causal prior from a learned DAG** — fit on outer-training data only

Training data
$X \in \mathbb{R}^{N \times K}$, y
features F, target y

Sparse DAG learning (NOTEARS-Lasso)

$$\min_{W \in \mathbb{R}^{d \times d}} \frac{1}{2N}\|X - XW\|_F^2 + \lambda\|W\|_1$$
$$\text{s.t. } h(W) = \operatorname{tr}\left(e^{W \circ W}\right) - d = 0$$

Learned graph G = (V, E), V = F ∪ {y}



Soft prior: 3 groups
- C^*_{direct}
- $C^*_{indirect}$
- C^*_{other}

Screening
keep the top-k features of each group by mutual information with y

Causal role of each feature

$$\mathcal{M}(f) = \begin{cases} \text{direct} & f \to y \\ \text{indirect} & f \rightsquigarrow y, \text{ path length} \geq 2 \\ \text{other} & \text{otherwise} \end{cases}$$

PHASE II **Cascaded multi-agent DQN** — repeat until validation gain stalls

Exploration mixes causal-hierarchical pairing, top-k mutual information, and diversity sampling

π_1 Group agent
picks a causal group $C^{(1)}$
sees statistics of all features and of each group

π_o Operator agent
picks an operator o (11 unary, 4 binary)
also sees the chosen group

π_2 Partner agent
picks a partner group $C^{(2)}$ for binary operators
also sees the chosen operator

Generate
apply π_o to features from the chosen groups, with numerical guards

Prune
drop near-constant features, keep the top-k by mutual information

Evaluate
fit XGBoost, score on inner-validation

Causally shaped reward, then a TD update of all three agents

$$R_t = R_{\text{perf},t}\left(1 + \alpha\,\Psi_{\text{causal},t}\right) + \lambda_{\text{div}}\,H(\pi_t) - \lambda_{\text{comp}}\,C(\mathcal{F}_t^g)$$

$$\Psi_{\text{causal},t} = \frac{1}{|\mathcal{F}_t^g|}\sum_{f \in \mathcal{F}_t^g} w_{\mathcal{M}(f)}\,\text{Usage}(f)$$

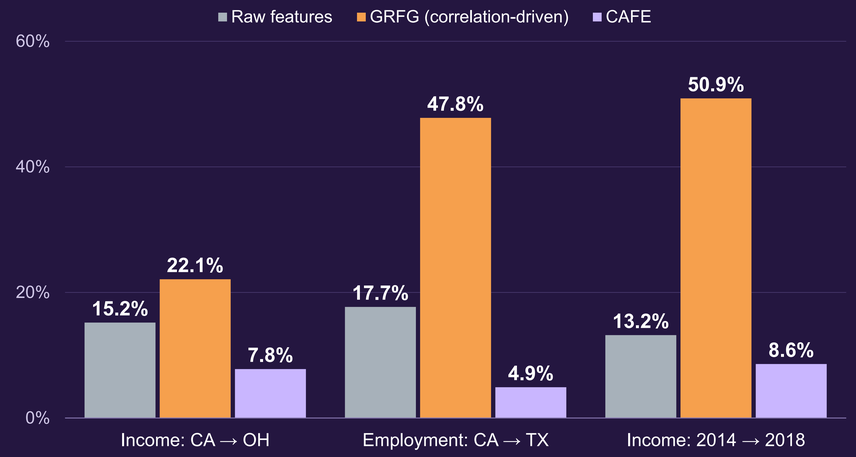
$w_{direct} > w_{indirect} > w_{other} \geq 0$: validation gains built on causal features count more. H: policy entropy. C: feature-set complexity.

PROTOCOL **Strictly inductive nested evaluation** — 5 outer folds × 5 seeds

outer-training fold: Phase I (DAG, roles, MI screening) fit here		outer test
search-train: feature construction	inner-val: reward, selection	frozen F*, scored once

2 MAIN RESULT · NATURAL DISTRIBUTION SHIFT (FOLKTABLES ACS)

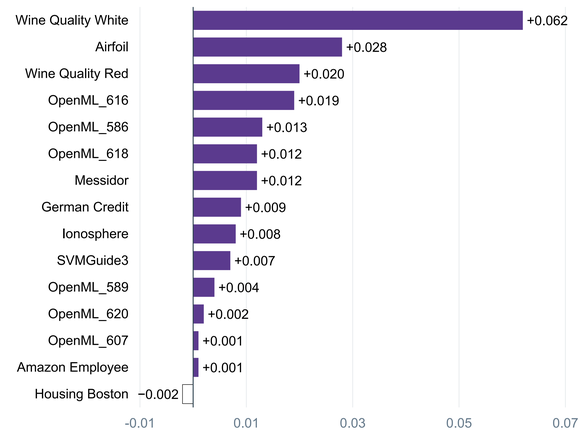
Under real distribution shift, CAFE features lose **4.9 to 8.6%** of F1. Correlation-driven search loses **22.1 to 50.9%** and falls below raw features on two of three shifts.



Relative F1 drop from training distribution to shifted test. CAFE keeps 91.4 to 95.2% of its in-distribution F1; GRFG keeps 49.0 to 77.9%.

3 Consistent utility gains

CAFE minus the best of 9 baselines



Macro-F1 or 1-RAE; nested 5-fold CV, 5 seeds, XGBoost. Positive on 14 of 15. Wilcoxon vs GRFG p < 0.001 (r = 0.68), vs ELLM-FT p = 0.003 (r = 0.51). Against OpenFE (appendix), CAFE wins 13 of 15.

4 Causal, or just grouping?

Gain over random equal-size groups

Correlation groups	+2 to 6%
Causal groups (CAFE)	**+5 to 12%**

Same agents, reward, and group sizes; only the grouping changes.

5 Controlled shift

Performance drop (%) by shift intensity

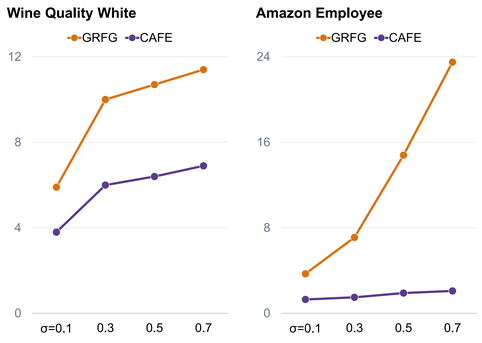
	CAFE low	CAFE med	CAFE high	GRFG low	GRFG med	GRFG high
OpenML_586	−0.3	−0.1	6.6	14.7	49.1	38.6
OpenML_589	0.4	30.3	17.6	10.7	37.1	25.1
Wine Red	3.2	0.3	3.4	11.2	11.2	74.0
Wine White	22.2	0.1	1.2	86.7	10.3	10.4

0 — 87% drop

Mean in-distribution → shifted score: CAFE 0.773 → 0.718 (−7.1%), GRFG 0.734 → 0.528 (−28.1%), ELLM-FT 0.741 → 0.503 (−32.1%).

6 Stable attributions

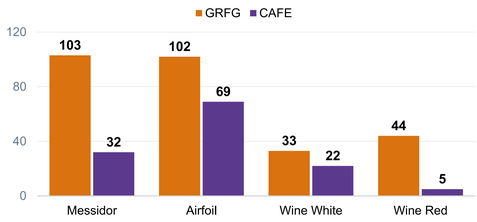
SHAP variance ($\times 10^{-3}$) under input noise



Lower is better. Mean reduction vs GRFG is 58.6% across 4 datasets, up to 90.8% (Amazon Employee, σ = 0.7).

7 Efficiency and scope

Steps to convergence: 32 to 89% fewer



- **Runtime:** Phase I takes 5 to 15% of wall-clock time
- **Scale:** at d = 272 (Arrhythmia), F1 0.823 vs 0.808 for GRFG
- **Scope:** static tabular data; works with any causal-discovery backend that returns a sparse DAG (NOTEARS-Lasso, NOTEARS-MLP, DAGMA tested)

References: GRFG, Wang et al., KDD 2022. ELLM-FT, Gong et al., AAAI 2025. NOTEARS, Zheng et al., NeurIPS 2018. Folktables, Ding et al., 2022. OpenFE, Zhang et al., ICML 2023. This work is supported in part by NSF grants 2426340, 2416727, 2421864, 2421865, 2421803 and NAE Grainger Foundation Frontiers of Engineering Grants.