GRAFT：結合 LUNAR 與 T2G-Former 之 GRAPE 擴充

邱聖佐 (Sheng-Tso Chiu)
指導教授：李政德 (Cheng-Te Li)
National Cheng Kung University

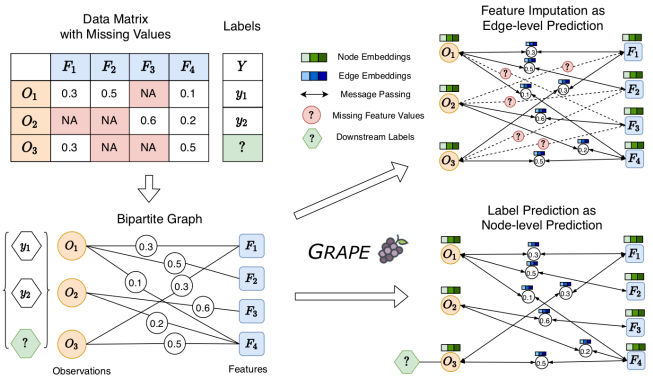


Fig. 1: Overview of the proposed GRAPE framework

***Abstract*—本研究以 GRAPE [1] 為基底，設計一條可組合 LUNAR（資料異常列過濾） [2] 與 T2G-Former（特徵關聯圖評分 / 軟式特徵選擇） [3] 的剪枝管線（稱為 GRAFT），透過將各階段輸出的遮罩與邊保留集合化，最後交由 GRAPE 進行標籤預測訓練。實驗涵蓋 UCI 迴歸資料集與 T2G 原文的大型資料（另有抽樣設定）。總結來看：LUNAR+GRAPE 在 UCI 場景最為穩健；T2G 在低特徵數資料上增益有限；在高特徵 / 大樣本設定下，T2G 的邊數量 $(n \times d)$ 使記憶體成本顯著上升，但在稀疏觀測下能有效提純訊號。**

***Index Terms*—GRAPE, LUNAR, T2G-Former, 缺失值, 特徵選擇, 圖學習, 縮圖**

I. 研究動機與貢獻

A. 動機

GRAPE 能有效處理表格資料的缺失值；然而在實務中，計算資源、延遲與資料治理常要求我們盡量縮小圖的規模（更少的樣本與特徵），同時不顯著犧牲模型穩定性與效能。本研究以 **GRAFT** 管線結合 LUNAR（於列層級減樣）與 T2G-Former（由 FR-Graph 於欄層級做關聯導引的特徵選擇），系統性探索：在不顯著影響 Label 表現與穩定性的前提下，圖可以縮小到什麼程度。

B. 貢獻

(1) 系統化接入 LUNAR（row keep 遮罩）與 T2G-Former（FR-Graph 導出的欄遮罩），並與 GRAPE 無縫串接；(2) 建立可重現的中介資料介面，支援列 / 欄雙向縮圖與合成；(3) 在 UCI 與 Year（高特徵數）上實證「縮圖—效能—穩定性」的關係與方法先後次序的影響。

II. 方法

A. 遮罩合成與邊保留

依 GRAPE 架構將表格資料建成二部圖：列（observations）與欄（features）分別為兩類節點；當第 i 列第 j 欄為可觀測值時，即在兩節點之間連一條邊。在進入 GRAPE 訓練前，我們先對行與列進行「保留 / 移除」的前置篩選，以縮減圖規模並穩定訓練。整合既有的可見性遮罩後，於行、列兩個層級決定是否保留對應節點；被移除的行 / 列及其對應邊在訓練中不予考慮。

B. *LUNAR* 階段（*Row-level Keep*）

依資料集給定鄰居數 k 與保留比例 $keep$ 的預設；因 LUNAR 對 k 不敏感，目的在於減少超參數調整，本實驗亦不針對 k 做額外微調（UCI 之 k 依資料量做線性縮放）。對每一列計算「異常分數」，將高分者視為異常列予以剔除、低分者保留；被判定為異常的列，其對應到所有特徵的觀測邊一併去除；保留下來的列則維持原始觀測。

C. *T2G* 階段（*Feature-level Soft Prune*）

依 T2G 的設計，先估計特徵之間的關聯強度，得到一個能反映「哪些特徵更具代表性 / 與其他特徵互動更關鍵」的分數。再依此分數挑選要保留的特徵（可設定保留比例）；分數較低者視為冗餘或影響力較小予以移除。被移除的特徵，其在圖中的所有相關邊同步刪除；保留的特徵則完整保留其觀測關係。

D. *GRAPE* 訓練階段

最終的訓練邊集合由「被保留的列」與「被保留的欄」共同決定：只有同時通過兩道篩選的列與欄之間的觀測，才會留在圖中。這樣的合成方式，能在不顯著犧牲預測與補值表現的前提下，有效降低邊數、記憶體占用與訓練時間，並讓模型聚焦在更乾淨、關聯性更高的資料子集。T2G 和 GRAPE 都可做標籤預測；本研究以 GRAPE 為最終訓練，保留其 end-to-end 架構。

III. 實驗設計

A. *Baseline* 中介輸出

既有實作多為端到端、缺乏階段性輸出。為了可重現、可疊代與可做消融比較，在不改動原演算法邏輯前提下，我們將前處理與評估所需的中介資料標準化輸出，作為各階段共同介面：資料標準化結果、缺失與觀測遮罩、訓練 / 驗證 / 測試分割、補值評估用測試觀測索引、二部圖列–欄對應關係，以及任務（標籤預測 / 特徵補值）的預測與指標。

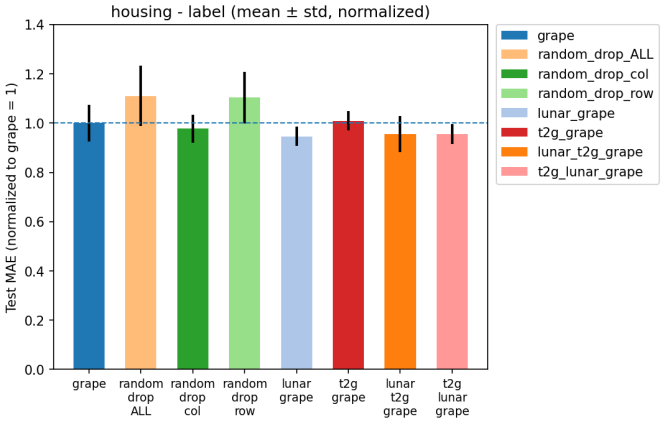


Fig. 2: UCI–Housing：各方法之標準化 MAE（越低越好）。

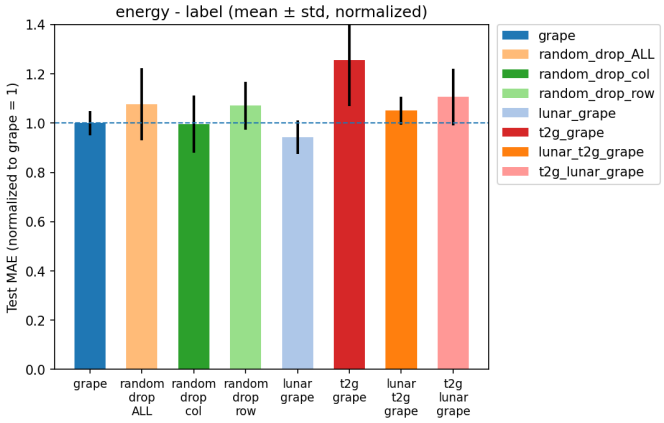


Fig. 3: UCI–Energy：各方法之標準化 MAE。

B. *Pack Domain* 支援

將上述中介資料視為統一資料介面，由訓練流程直接讀取，不受限於特定資料讀取器或原生資料庫結構。當特徵維度偏大或資料來自非原生環境（例如非 GRAPE 原生的 UCI 版本），可透過打包域快速接入既有流程，並以最小度數等衛生檢查避免過度稀疏。

C. 任務與方法

聚焦於**標籤預測**，主要評估指標為測試集 MAE。比較四種方法序列：**LUNAR→GRAPE**、**T2G→GRAPE**、**LUNAR→T2G→GRAPE**、**T2G→LUNAR→GRAPE**，藉此觀察列與欄兩層級縮圖時，不同先後次序對效能與穩定性的影響；同時設計四種**隨機剪枝基線**並搭配 GRAPE，作為不帶任何引導的縮圖對照。

D. 超參與設定

UCI：GRAPE 採原生預設超參，以利與基線對齊；LUNAR 依資料集大小與稀疏程度調整保留比例（heuristic），避免過度刪除；T2G 默認採輕量監督（約 50 epoch）。

Year（大圖，小圖驗證）：樣本與特徵同時偏大時，二部圖邊數近似 $n \times d$，訓練時間與顯存需求快速上升。我們按原訓練 / 驗證 / 測試比例抽樣至約 10k 列，在此小圖上先進行 T2G 的特徵關聯學習（輕量監督），依分數保留約 50% 欄；GRAPE 則以小型嵌入（8 維）與極低可見邊比例（約 1%）訓練，用以檢測 T2G 在小圖的邊際貢獻。

IV. 結果與討論

A. *UCI* 總覽（小特徵數）

多數資料集（*housing, energy, kin8nm, concrete, yacht* 等）中，**LUNAR+GRAPE** 對標籤預測（MAE）大多帶來穩健的改善；相對地，**T2G** 在小特徵數（約 6–16 維）上多數情況不如或僅略等同於純 GRAPE，而與 LUNAR 串接亦未必優於 LUNAR 單獨。

B. 對「縮圖」目標的意涵

列層級（LUNAR）是首選的穩健縮圖工具：希望先縮小圖而不犧牲表現時，宜優先採列層級降噪。

欄層級（T2G）效益依賴特徵互動結構：在小特徵數場景，T2G 的欄層級互動受限於可用關係結構——此趨勢與 T2G 原文回歸結果一致（如 CA、HO、FB、YE）。

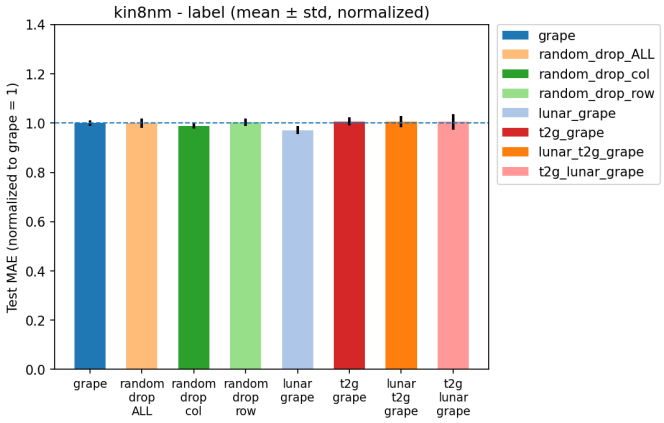


Fig. 4: UCI–Kin8nm：各方法之標準化 MAE。

方法次序需納入設計空間：建議同時嘗試 LUNAR→T2G 與 T2G→LUNAR，並以驗證集 MAE / 方差共同決策。

V. 失敗案例與分析

在小圖設定下（列 10k、已知邊 1%），先以 T2G 進行特徵關聯學習（約 50 epoch），依分數保留約 50% 欄，再以小型嵌入（8 維）進行 GRAPE 訓練。此設定同時具備兩個特點：其一，列數縮減使跨列的特徵互動訊號較弱；其二，極低可見邊比例進一步壓縮可學關係。綜合而言，這種「互動訊號不足、訓練邊稀薄」的小圖環境，解釋了為何 T2G 與純 GRAPE 的差距在此不易被放大（但在高維、非小圖設定時，T2G 的好處更明顯）。

A. *Year*（高特徵數）的小圖設定與觀察

在縮小版 *year* 上，僅比較 (a)GRAPE 與 (b)T2G+GRAPE：列數縮減使跨列互動訊號較弱，極低可見邊（1%）進一步壓縮可學關係。，以直接觀察 T2G 在小圖上的邊際貢獻。

VI. Ablation

A. 僅調整 *LUNAR* 的特徵保留率（*keep*）

我們聚焦於 **T2G→LUNAR→GRAPE** 的組合做單因子消融；其餘元件與訓練流程固定。先前觀察顯示：

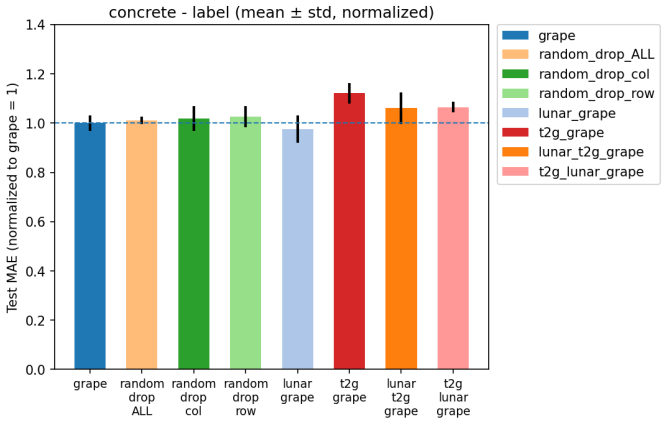


Fig. 5: UCI–Concrete：各方法之標準化 MAE。

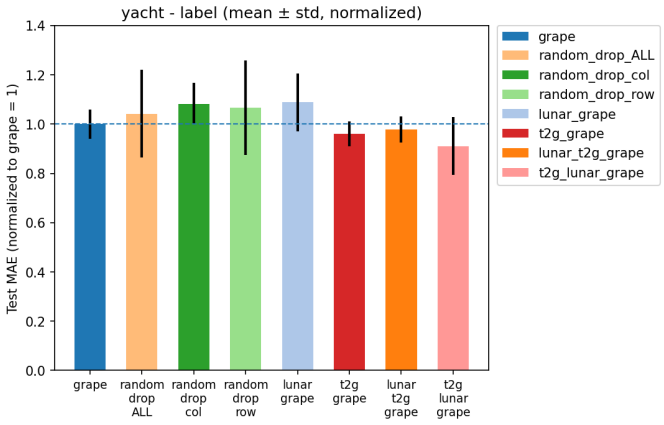


Fig. 6: UCI–Yacht：各方法之標準化 MAE（先 T2G 後 LUNAR 為例外的最佳序）。

在低維特徵情境下，T2G（tabular-to-graph）於欄端的刪除邊際效益有限，且 **T2G→LUNAR→GRAPE** 與 **LUNAR→T2G→GRAPE** 表現相近，顯示方法順序對整體效果影響有限。因此本節僅針對 **LUNAR 的 row 端 keep 比例**做掃描，以觀察「降噪 vs. 訊息量」的權衡。

設定：固定 column 端保留率為 0.90；列端 keep $\in \{0.70, 0.80, 0.90, 0.95\}$。任務為 label（MAE）。資料集採 *yacht*、*energy*、*housing*、*wine*；k 依樣本數 N 設上限，且不超過 100。所有前處理與訓練細節保持一致以排除非目標因素。逐資料集的最優 keep 與趨勢，請見附錄圖 **Figs.7-Figs.11** 的圖說。

VII. 未來工作

LUNAR *keep_ratio* 自動化（以 GRAPE valid MAE 擇優）；Column-level LUNAR：以 $X^\top$ 偵測「壞特徵」，產生列 / 欄雙遮罩；T2G 權重學習消融：在小步監督（epochs>0）與層權重組合上，檢驗對 Label / MDI 的差異。

VIII. 結論

本研究以嚴格可重現的 GRAFT 管線整合 LUNAR 與 T2G-Former 對 GRAPE 進行增強。實驗顯示 LUNAR+GRAPE 在多數 UCI 迴歸資料上最為穩健，而 T2G

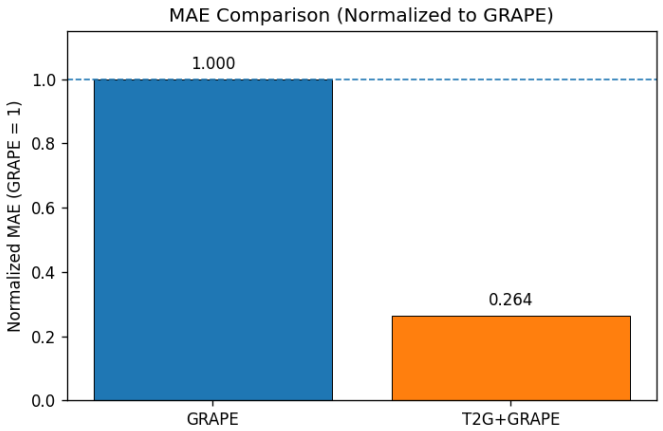


Fig. 7: Year（高特徵數、列 ∼10k）：T2G+GRAPE 相對 GRAPE 的標準化 MAE（GRAPE=1）。

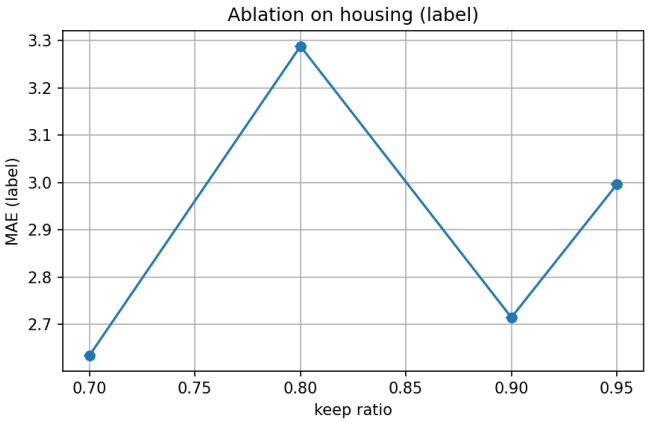


Fig. 8: **Housing**：**keep=0.90 最佳**；0.80 反而最差，顯示關鍵特徵較分散，需保留較高比例以維持訊息量。

的貢獻隨特徵數上升而放大；在高特徵 / 大量樣本場景下，需權衡其資源成本與實際可用資源。

References

[1] J. You, X. Ma, D. Y. Ding, M. J. Kochenderfer, and J. Leskovec, "Handling missing data with graph representation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] A. Goodge, B. Hooi, S. K. Ng, and W. S. Ng, "Lunar: Unifying local outlier detection methods via graph neural networks," 2022.

[3] J. Yan, J. Chen, Y. Wu, D. Z. Chen, and J. Wu, "T2g-former: Organizing tabular features into relation graphs promotes heterogeneous feature interaction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. [Online]. Available: https://arxiv.org/abs/2211.16887

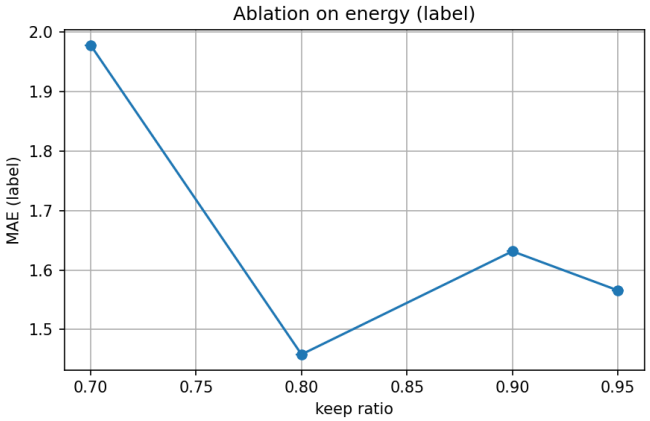


Fig. 9: **Energy**：趨勢與 *yacht* 類似但幅度較小，**keep=0.80 最佳**、0.95 次佳、0.70 最差。

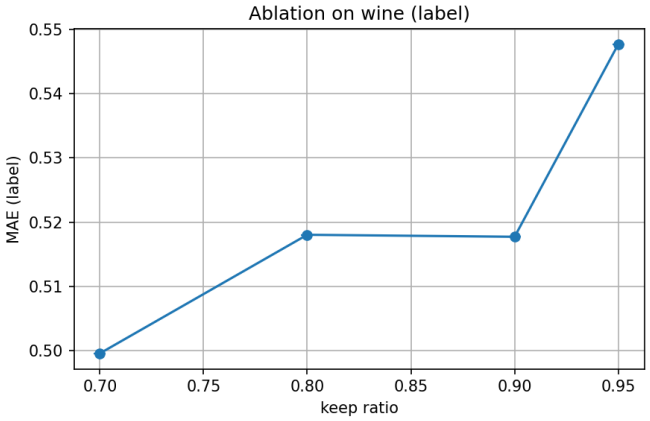


Fig. 10: **Wine**：近似單調上升，**keep=0.70 最佳**、0.95 最差，顯示特徵冗餘較高，積極裁剪更能抑制噪聲與過擬合。

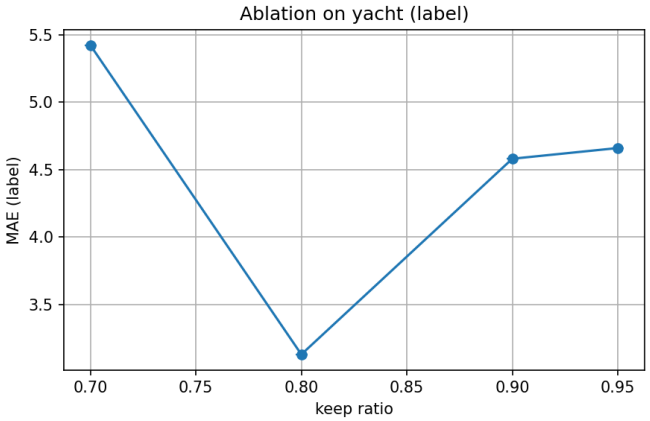
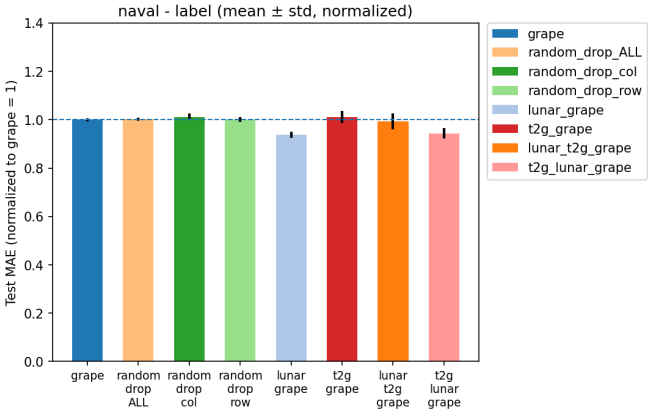


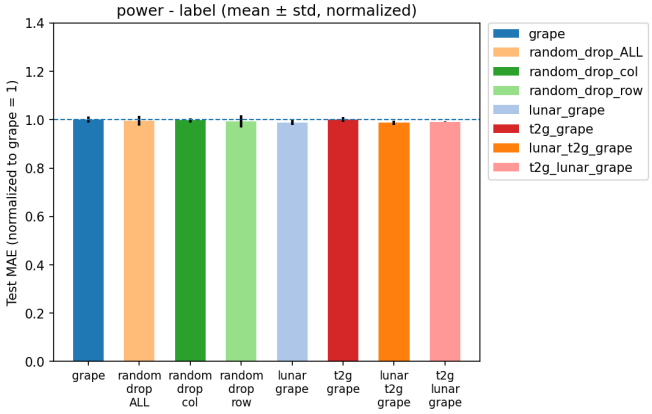
Fig. 11: **Yacht**：呈明顯 *V* 形，**keep=0.80 最佳**；0.70 過度刪除導致 MAE 上升，0.90/0.95 夾帶噪聲亦使 MAE 增加。

Appendix

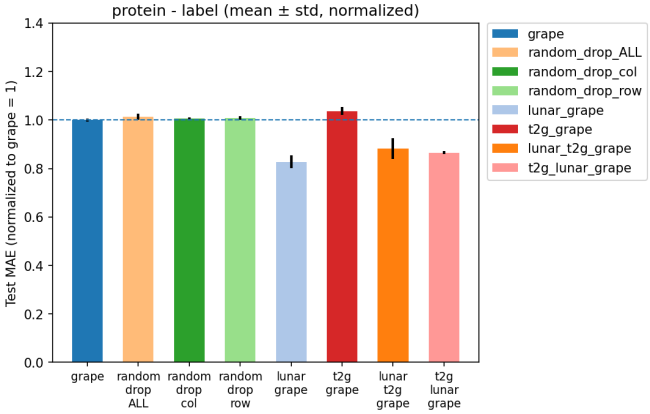
Appendix —Supplementary Figures



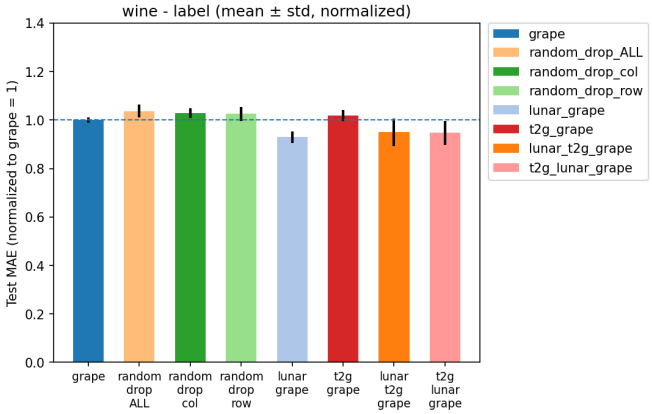
(a) UCI–Naval



(b) UCI–Power



(c) UCI–Protein



(d) UCI–Wine

Fig. 12: 補充圖：各資料集之標籤預測（標準化 MAE；越低越好）。黑色誤差棒為平均 ± 標準差。