Discovering Association Rules in High-Dimensional Small Tabular Data

Erkan Karabulut[1] (e.karabulut@uva.nl), Daniel Daza[2], Paul Groth[1], Victoria Degeler[1]
[1]University of Amsterdam, [2]Amsterdam University Medical Center

1 Learning Rules?

Knowledge discovery: Reveal associations between data features, e.g., columns of a given table (ARM).
Interpretable inference: Draw conclusions using learned rules instead of black box models, such as classification rules.

Formalization: Table with k features $F = \{f_1, ..., f_k\}$, each with classes $f_i^1, ..., f_i^{c_i}$.
Define the item universe $I = \left\{ f_i^j \mid 1 \le i \le k, 1 \le j \le c_i \right\}$.
Each row (transaction) $T \subset I$ satisfies $\forall i \in \{1, ..., k\}, \exists! j \in \{1, ..., c_i\}, f_i^j \in T$
An association rule is $X \to Y$ with $X, Y \subset I$, $X \cap Y = \emptyset$, $|Y| = 1$.
Logical form: $X \to Y \equiv (\neg \wedge_{x \in X}\, x) \vee y$ (Horn clause in CNF).

Gene expression datasets:

gene1	gene2	gene3	...	gene18107
normal	normal	normal		normal
normal	normal	high	...	high
normal	normal	normal		low

Gao et al. 2015.

Gene2(high) ∧ Gene29(high) → Gene14(low)

Rule Explosion:

Even small datasets can generate millions of redundant or trivial rules --- slow, hard to interpret, and unscalable.

$$X \to Y, \quad |X| + |Y| \le a$$

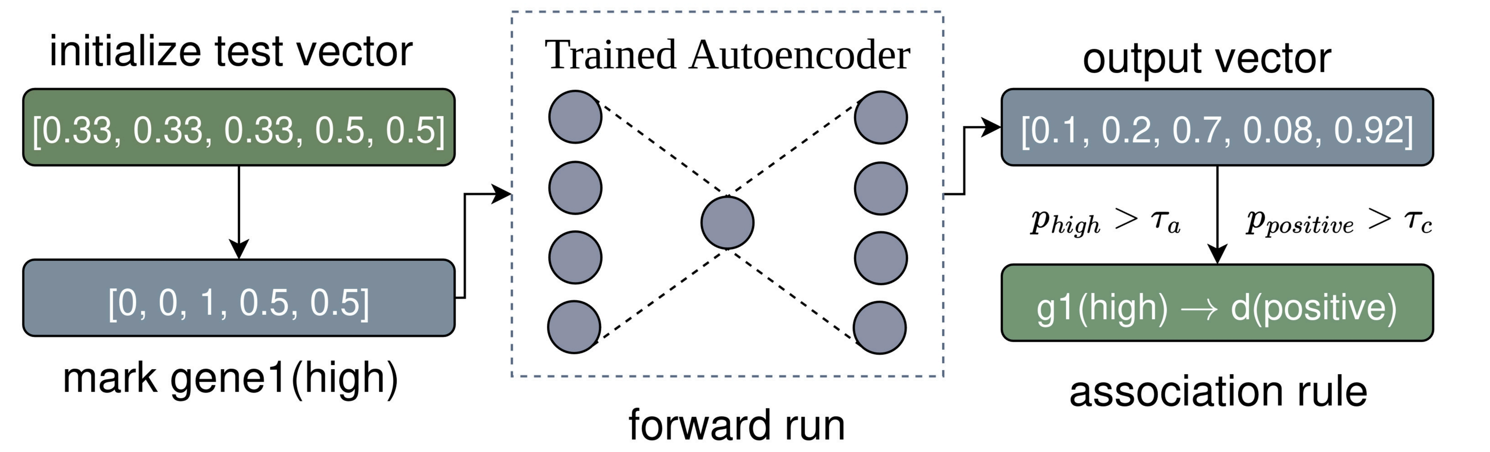
Example: 20 features, 4 values each, $a = 4 \Rightarrow \mathbf{5,186,240}$ rules!

$$\#\,\text{Rules} = \sum_{p=1}^{a} c_p! \left(\prod_{i \ne p} (c_i + 1) - 1 \right)$$

2 Neurosymbolic ARM

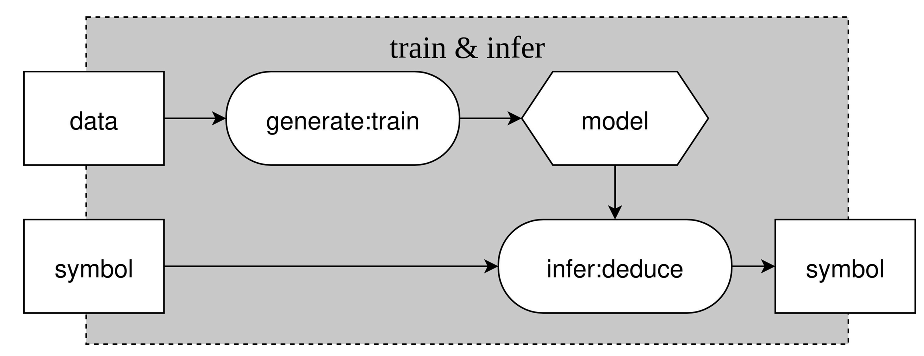
Intuition: Autoencoders capture feature associations via reconstruction. If, after training, a forward pass with marked categories A reconstructs categories C with high probability, then $A \to C \setminus A$ (no self-implication).

gene1 = {low, normal, high}, disease = {negative, positive} τ_a = 0.5, τ_c = 0.8



(Aerial+), Karabulut et al. 2025

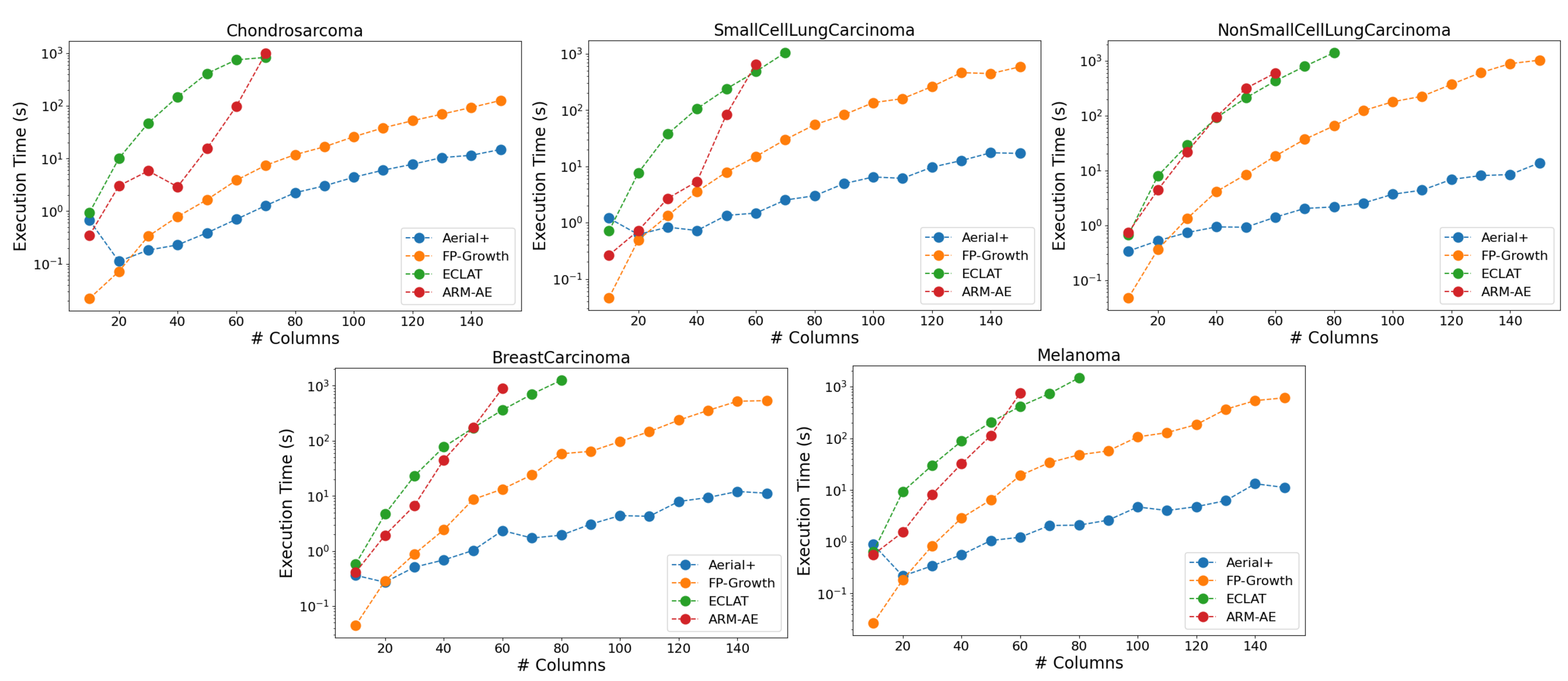
Observation:
The reconstruction objective of Aerial+ prioritizes significant patterns rather than redundant patterns!



(Boxology), van Bekkum et al. 2021

3 Empirical Analysis Neurosymbolic ARM is Scalable!

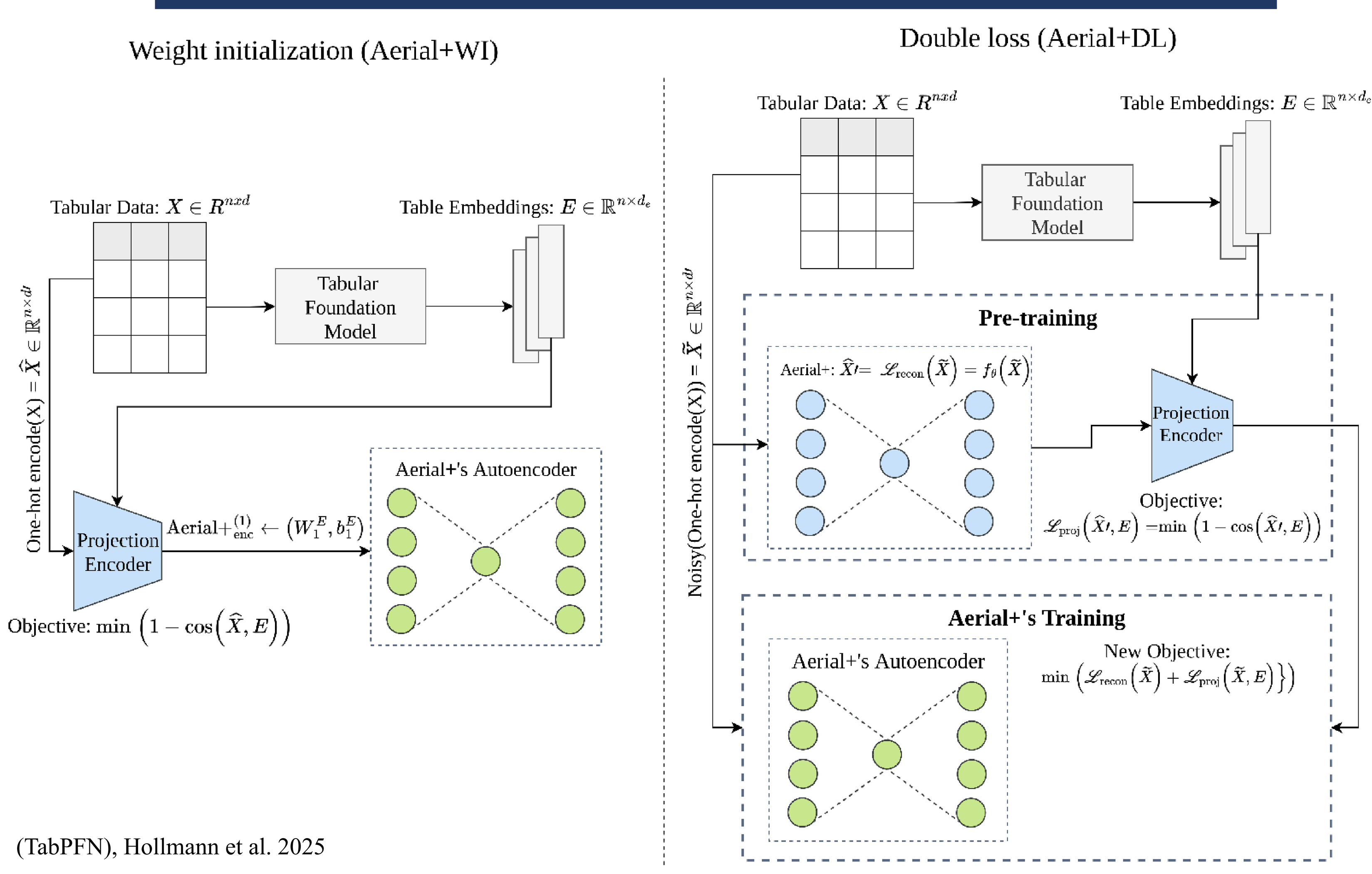
Empirical result: NeSy method scales 1-2 orders of magnitude faster, despite low data!



Dataset	# Columns	# Rows
Chondrosarcoma	18006	6
SmallCellLungCarcinoma	18237	60
NonSmallCellLungCarcinoma	18108	86
BreastCarcinoma	18061	51
Melanoma	17902	55

Algorithm	Type	Parameters
Aerial+	Neurosymbolic	$a = 2, \tau_a = 0.5, \tau_c = 0.8$
ARM-AE	Neurosymbolic	M=2, N=$\|R\|/\|C\|$, L=0.5
FP-Growth	Algorithmic	(both) antecedents = 2, min_conf=0.8,
ECLAT	Algorithmic	min_support=0.5 $*\ \mathbb{E}[\text{support}(R)]$

4 How to Boost Knowledge Discovery in a Low-Data Regime?



(TabPFN), Hollmann et al. 2025

Weight initialization from tabular foundation models (**Aerial+WI**) and semantic alignment of table embeddings with code layers using a dual loss function (**Aerial+-DL**)

5 Validation, Interpration and Future Research

Empirical result: A concise set of higher-quality association rules!

Approach	# Rules	~Rule Coverage	~Support	~Confidence	Data Coverage	~Zhang's Metric	Exec. Time (s)
Chondrosarcoma							
Aerial+	200	**0.23**	0.21	0.921	0.533	0.784	2.25
Aerial+WI	75	0.217	0.206	0.945	0.524	0.813	5.80
Aerial+DL	75	0.235	0.219	**0.947**	**0.536**	**0.828**	5.36
SmallCellLungCarcinoma							
Aerial+	1576	0.068	0.041	0.579	**0.835**	0.476	10.58
Aerial+WI	664	**0.076**	0.052	**0.633**	0.715	**0.577**	13.48
Aerial+DL	1338	0.070	0.044	0.597	0.816	0.513	18.23
NonSmallCellLungCarcinoma							
Aerial+	1620	0.059	0.035	0.584	0.823	0.554	18.03
Aerial+WI	978	**0.078**	0.057	**0.663**	0.698	**0.639**	28.67
Aerial+DL	1453	0.053	0.028	0.547	**0.849**	0.501	24.27
BreastCarcinoma							
Aerial+	1017	0.072	0.046	0.641	**0.816**	0.575	9.64
Aerial+WI	590	0.077	0.052	**0.686**	0.686	**0.644**	12.09
Aerial+DL	535	**0.078**	0.050	0.652	0.761	0.590	15.31
Melanoma							
Aerial+	1220	0.067	0.035	0.545	**0.888**	0.440	13.09
Aerial+WI	773	0.070	0.038	**0.575**	0.772	**0.496**	13.19
Aerial+DL	859	**0.071**	0.038	0.566	0.860	0.461	16.49

Metrics:
$\text{Support}(X \to Y) = |\{T : X \cup Y \subseteq T\}| \,/\, n$
$\text{Confidence}(X \to Y) = |\{T : X \cup Y \subseteq T\}| \,/\, |\{T : X \subseteq T\}|$
$\text{Coverage}(X \to Y) = |\{T : X \subseteq T\}| \,/\, n$

$$\text{Zhang}(X \to Y) = \frac{\text{conf}(X \to Y) - \text{conf}(X' \to Y)}{\max\left(\text{conf}(X \to Y), \text{conf}(X' \to Y)\right)}$$

- **Dataset dependency:** Algorithmic methods' execution time increases with data density (many high-support itemsets), while Aerial+ maintains constant polynomial-time extraction regardless of density.
- **Validation scope:** Testing on more diverse domains and datasets with higher instance-to-feature ratios (n ≫ d) needed to assess convergence and rule quality
- **Foundation model constraint:** Current approach limited to TabPFN (only available tabular foundation model with table embedding interface); designed for classification/regression rather than column associations
- **Background knowledge in knowledge discovery:** What other types of knowledge, e.g., structured or bayesian, can be utilized in knowledge discovery?

Code

Paper

University of Amsterdam INDE lab Amsterdam UMC University Medical Centers

Funded by the Dutch Research Council (NWO), in the scope of Digital Twin for Evolutionary Changes in water networks (DiTEC) project, file number 19454.