19 June 2026, IACM Journal Club

Feature Selection and Knowledge Discovery in Automated Machine Learning and Automated Causal Discovery

Ioannis Tsamardinos

(Yianni)

Professor, Computer Science Department, University of Crete

Affiliated Researcher, Institute of Applied and Computational Mathematics, FORTH

CEO and founder, JADBio Gnosis DA S.A

Locations and Affiliations

- University of Crete and FORTH

Heraklion, Crete

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

FORTH
FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS

JADBio®

- University of Crete Spin Off company
- Advanced ML and AI products and services
- Founded in 2013
- Greek, Danish, German partners
- Spin Off of the Year 2021 Award by Greek Ministry

People

- Mens Ex Machina academic group (MXM) www.mensxmachina.org
- JADBio - Gnosis DA SA www.jadbio.com

Research:
Feature Selection, Causal Discovery, Automated Machine Learning, Automated Causal Discovery, Bioinformatics, Biomedical Informatics

2016

Ioannis Tsamardinos
Prof. Computer Science Department, University of Crete
Affiliated Faculty, Institute of Applied and Computational Mathematics, FORTH
CEO and co-founder, JADBio

Nikolaos Gkorgkolis
Post-doc, Institute of Applied and Computational Mathematics, FORTH

Konstantina Biza
PhD student, Computer Science Department, University of Crete
Affiliated Faculty, Institute of Applied and Computational Mathematics, FORTH

Antonis Ntroumpogiannis
PhD student, Computer Science Department, University of Crete
Affiliated Faculty, Institute of Applied and Computational Mathematics, FORTH

Nikolaos Modestos Kougioulis
Master's student, Computer Science Department, University of Crete
Affiliated Faculty, Institute of Applied and Computational Mathematics, FORTH

Vassillis Christophides
Prof.École Nationale Supérieure de l'Électronique et de ses Applications (ENSEA)

Giorgos Papoutsoglou
Chief Operations Officer, JADBio

Etienne Vareille
PhD student, École Nationale Supérieure de l'Électronique et de ses Applications (ENSEA)

Conflict of Interest Declaration

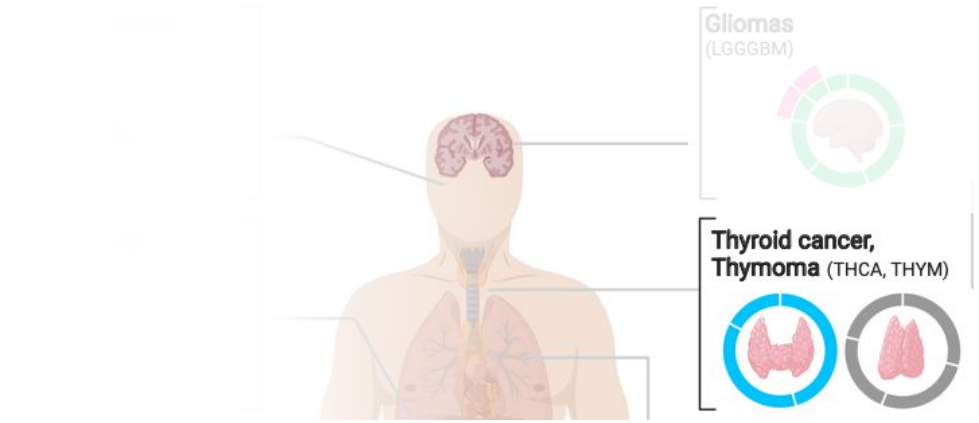
Some of the research and algorithmic results are commercially exploited by JADBio - Gnosis DA S.A.

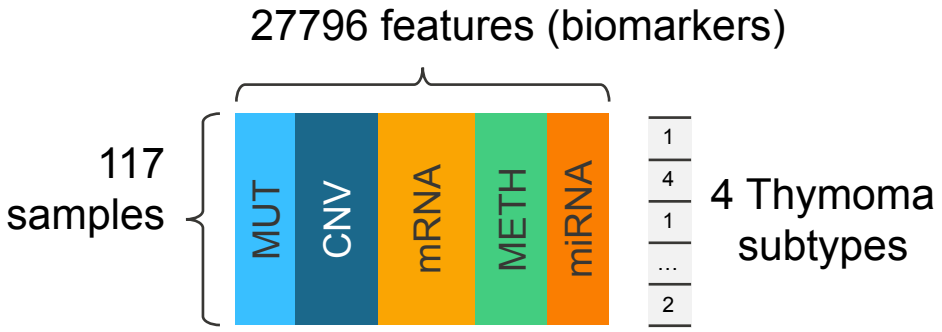
What Automated Machine Learning Should Deliver

For Predictive Modeling and Knowledge Discovery

Thymoma subtyping the TCGA data

- Can I **diagnose among 4 types of** Thymic Epithelial Tumors? How well?
- What are the relevant **oncogenes (features, biomarkers)**?



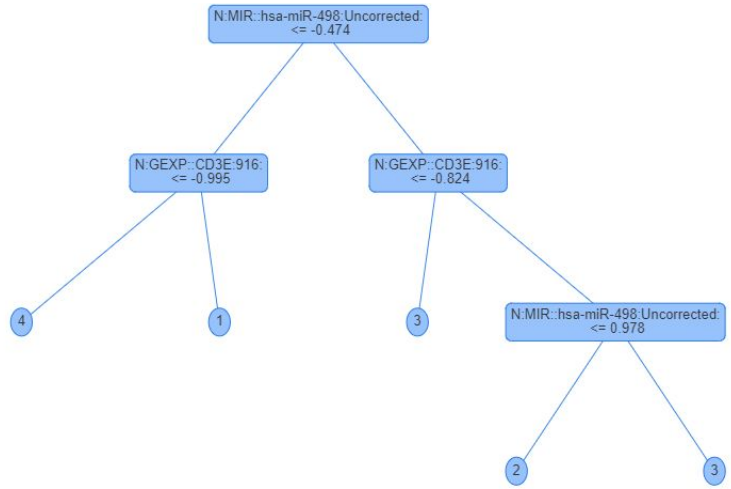


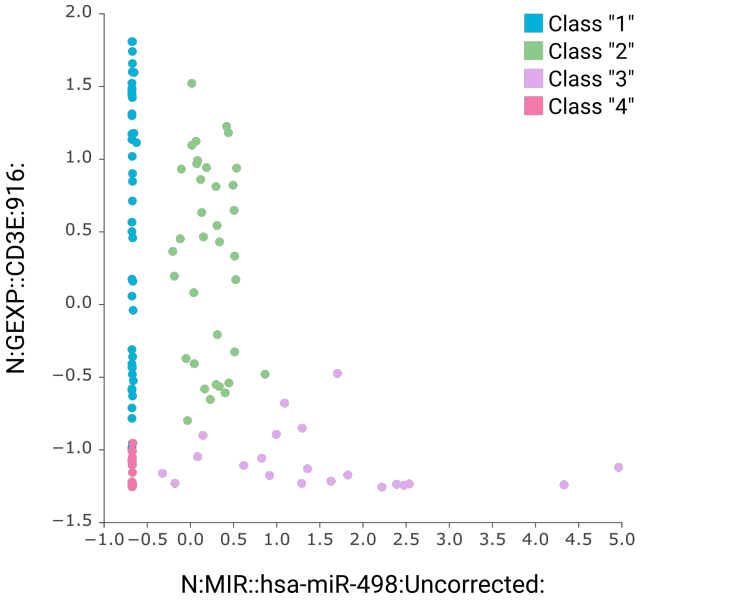
Data from **The Cancer Genome Atlas**

Thymoma subtyping: Biomarker discovery

- **Best model performance**: 0.976 AUC
  - Over hundreds of ML pipelines.
  - Surrogate interpretable model (Decision Tree) with few biomarkers
  - Is this performance estimate correct?

- **Knowledge Discovery**: Only 2 out of 27796 features are selected
  - First feature: lowest p-value. // Second feature: ranked 190.
    - **Examining biomarkers in order of correlation fails!**
  - But also! 45 pairs of 2 biomarkers are equally predictive
    - **Multiple signatures provide**:
      - **Choices** for designing a diagnostic assay
      - Alternative **causal** explanations





Mission Impossible

1. **Predictive Modeling**: Given past examples of **profiles** and their actual **outcome of interest**, learn a **predictive (diagnostic, prognostic) model** for new, unseen, profiles
2. **Feature Selection**: Identify what are the features that are jointly predictive
3. **Estimation**: How predictive is the model?

Profiles:
- Clinical and Medical Quantities
- Environmental, exposure, treatment factors
- (Multi) Omics
- Genetic
- **Images**
- Text (doctor's notes or publications)

Combinations of the above

Outcome of interest:
- Disease status (diagnosis)
- Sub-type of disease
- Response to treatment
- Phenotypic trait
- Time to death, relapse, complication
- Properties of a document

Mission Impossible

1. **Predictive Modeling**: Given past examples of **profiles** and their actual **outcome of interest**, learn a **predictive (diagnostic, prognostic) model** for new, unseen, profiles
2. **Feature Selection**: Identify what are the features that are jointly predictive
3. **Estimation**: How predictive is the model?

Profiles:
- Client characteristics
- Contract / subscription characteristics
- Transaction features
- Stock features
- Images
-

Outcome of interest:
- Fraudulent transaction or not?
- Default on bank loan or not?
- Will have a traffic accident or not?
- Will the stock value increase or not?
- Is this a picture of a human?
- …

```
names(Xbeta10yrsDr) <- row.names(Anewth)
a1 <- sort(Xbeta10yrsDr1, decreasing = TRUE)
head(a1)
a1[1668]#the cutoff risk threshold in order to select the 1668 highest risk
persons;2.225
i = which(Xbeta10yrsDr1 >= a1[[1668]])
sum(Anewth1$event[i],na.rm = TRUE)#50


i = which(Xbeta10yrsDr1 >= a1[[1350]])
sum(Anewth1$event[i],na.rm = TRUE)#47
i = which(Xbeta10yrsDr1 >= a1[[1300]])
sum(Anewth1$event[i],na.rm = TRUE)#47

#35 events as NLST
i = which(Xbeta10yrsDr1 >= a1[[709]])
sum(Anewth1$event[i],na.rm = TRUE)#

a <- sort(Xbeta10yrsDr, decreasing = TRUE)
head(a)
a[1870]#the cutoff risk threshold in order to select the 1870 highest risk
persons;2.804
a[3100]
i = which(Xbeta10yrsDr >= a[[3100]])
sum(Anewth$event[i],na.rm = TRUE)
1-0.9978^(exp(a[3100]))#2.24% vs 2.7% from quantile of events
1-0.9978^(exp(a[1870]))#3.57% risk within 10 years
aaax <- names(a)[1:1870]
i = which(Xbeta10yrsDr >= a[[1870]])#1870
sum(Anewth$event[i])#101/149!!!!
a1 <- sort(Xbeta10yrsDr1, decreasing = TRUE)
head(a1)
a1[918]#the cutoff risk threshold in order to select the 1870 highest risk
persons;2.8023
1-0.9978^(exp(a1[918]))#3.565% risk within 10 years
aaax <- names(a1)[1:918]
i = which(Xbeta10yrsDr1 >= a1[[918]])#918
sum(Anewth1$event[i])#40/52

i = which(Xbeta10yrsDr1 >= a[[1870]])#917
sum(Anewth1$event[i],na.rm = TRUE)#40
sum(Anewth1$event)#52
40/52#0.76923
#Better model performance in Rgroup == 1
#because only in this group quit time and BMI significantly reduce lung cancer risk
```

Current Practice: Scripting

Manual Analysis and Scripting …

- Requires expertise
- It is time-consuming
- It is prone to errors
- Critical components are not available (performance estimation, multiple feature selection)

Automated Machine Learning Challenges in Biomedicine Data Analysis

AutoML: automate predictive modeling is a relatively **new emerging paradigm**
(thymoma results shown have been achieved **automatically**)

Challenges in Biomedicine

- **Small sample sizes** and/or highly **imbalanced** classes
  - Cannot afford to “lose samples to estimation”
  - Need to account for rare classes
- The primary goal of the analysis is often **knowledge discovery**
- **Robustness:** AutoML needs to run on all outcome types, data sizes, data types, user preferences

*Tsamardinos, “Don’t lose samples to estimation”, Patterns, 2022, Link
*Tsamardinos, et. al, “Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation”, Machine Learning Journal, 2018, Link

JADBio (Just Add Data Bio)

Purpose-built AutoML platform for high-dimensional, low-sample biomedical data knowledge discovery

Builds and **deploys** accurate and explainable predictive models with speed and ease.

Discovers the features and measured quantities that are predictive among millions of measurements

https://jadbio.com

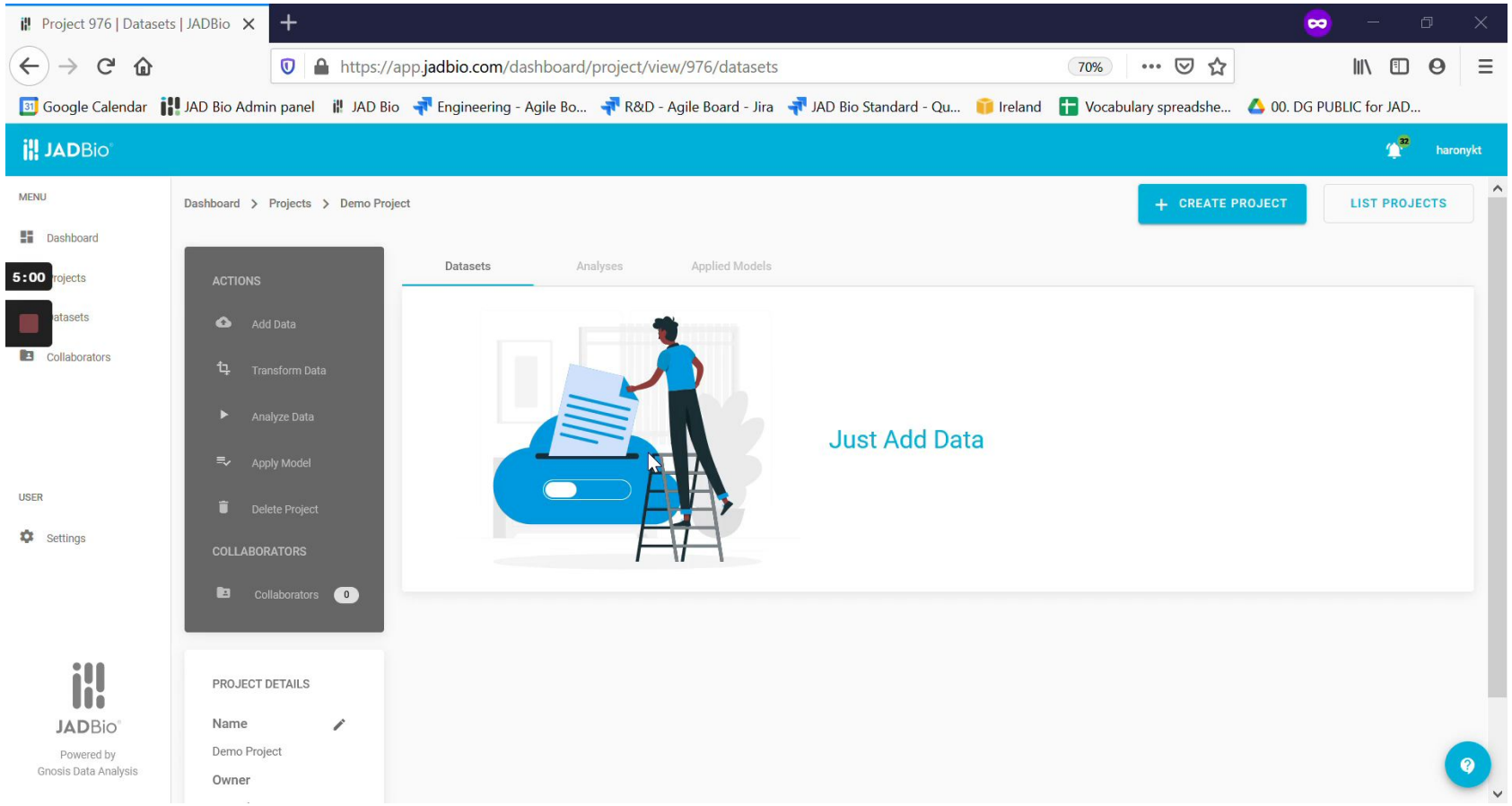
…Model interpretation, explanation, visualization, and deployment

Tsamardinos, *et al.* “Just Add Data: automated predictive modeling for knowledge discovery and feature selection.” *npj Precis. Onc.* **6**, 38 (2022). Link

npj | precision oncology

JADBio Video Demo

https://www.youtube.com/watch?v=u9hhofSidfM&ab_channel=JADBio



NATIONAL CANCER INSTITUTE
Center for Cancer Genomics

Kyle Ellrott, Christopher K. Wong, Christina Yau, Mauro A. A. Castro, Jordan A. Lee, Brian J. Karlberg, Jasleen K. Grewal, **Vincenzo Lagani**, Bahar Tercan, Verena Friedl, Toshinori Hinoue, Vladislav Uzunangelov, Lindsay Westlake, Xavier Loinaz, Ina Felau, Peggy I. Wang, Anab Kemal, Samantha J. Caesar-Johnson, Ilya Shmulevich, Alexander J. Lazar, **Ioannis Tsamardinos**, Katherine A. Hoadley, The Cancer Genome Atlas Analysis Network, A. Gordon Robertson, Theo A. Knijnenburg, Christopher C. Benz, Joshua M. Stuart, Jean C. Zenklusen, Andrew D. Cherniack, Peter W. Laird

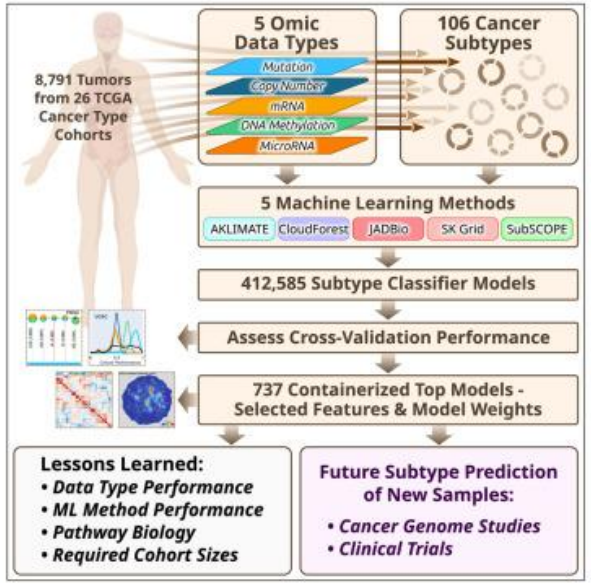
Cancer (sub)types Classification using multi-omics

Joint work with the Tumor Molecular Pathology (TMP) Analysis Working Group (AWG) of the US National Institute of Health (NIH) Center for Cancer Genomics (CCG)

Cancer Subtyping Project

TCGA

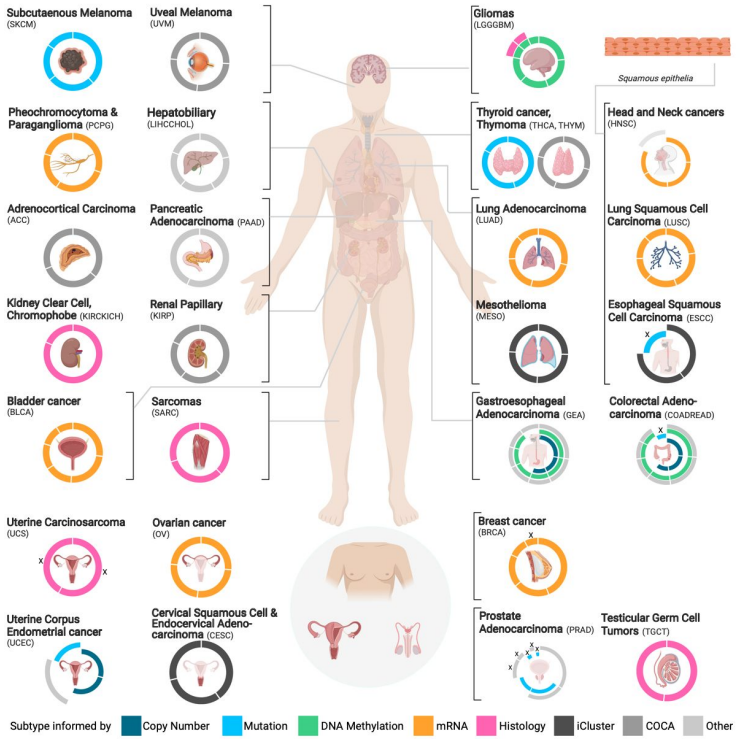
- The Thymoma analysis was part of a larger project, not a contrived example
- **Goal**: Create cancer subtype classification models with as few molecular features as possible; construct a pan-cancer subtyping assay
- **Data**: 8,791 samples (26 primary tumor types, 106 subtypes total), 5 molecular modalities with over 30000 features

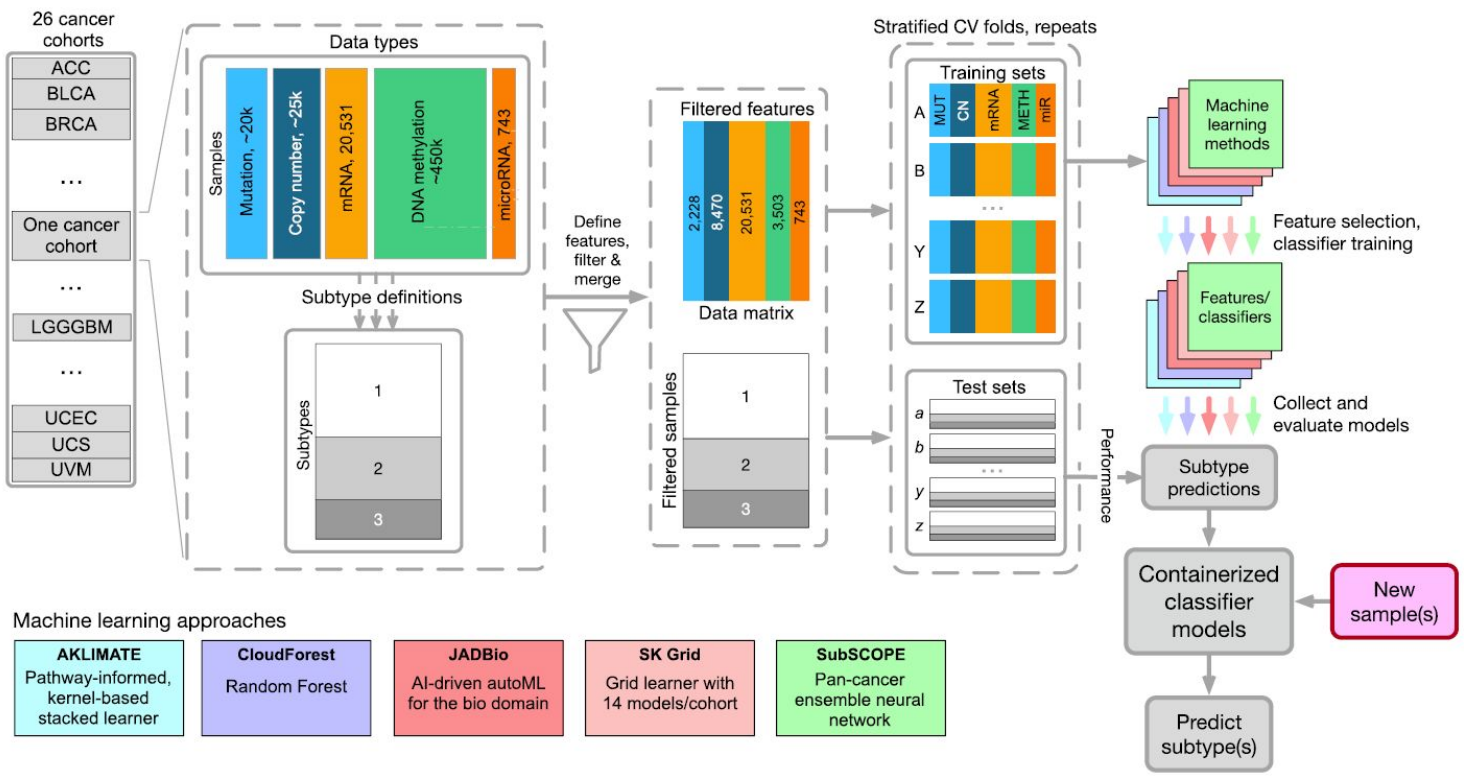


Cancer Cell

Data Analysis of TCGA Cohorts : The setup

Ellrott, et al., Cancer Cell, 2025, link



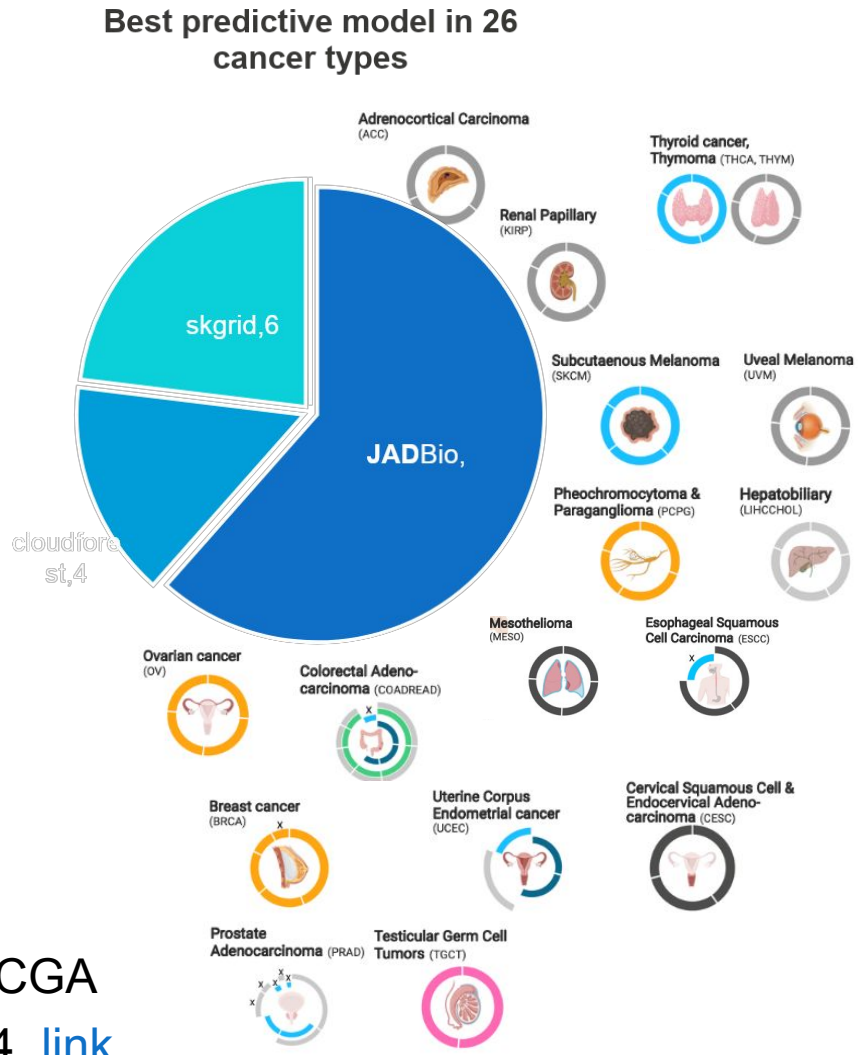


Ellrott, et al., Cancer Cell, 2025, link

Predictive Performance and Biomarker discovery

- Strong predictive performance by **JAD**Bio in 96% of cases (25/26)
  - highest performance in **16 out of 26** (62%) of cancer types
  - tied best performance in 1 cancer type (Prostate, PRAD)
  - within 1 s.d. of the best in 30% (8/26) of the remaining ones.

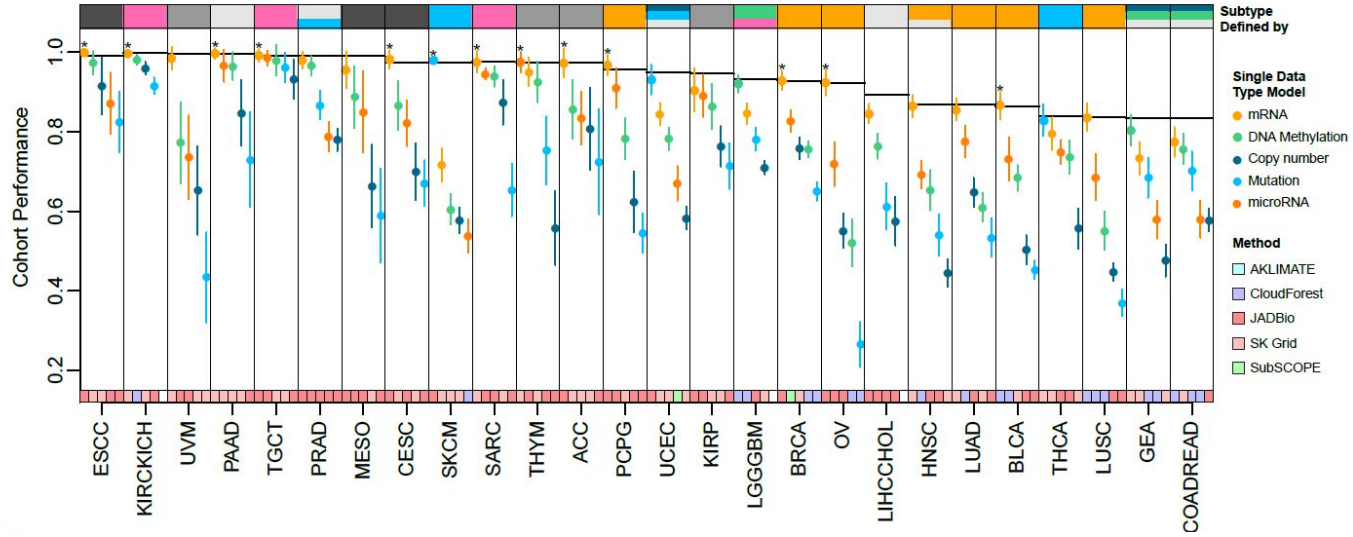
- Our feature selection algorithms select the **fewest biomarkers**
- The second-best method used the public version of our feature selection algorithms (MXM R package)

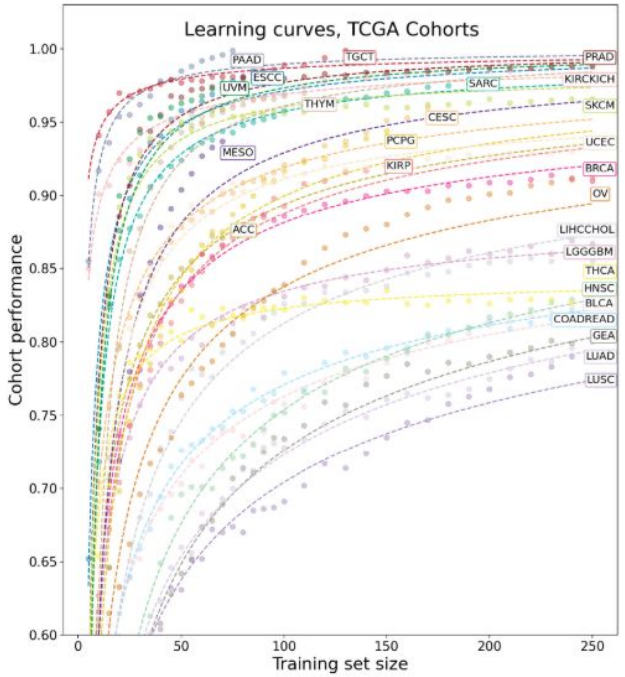


*Kyle Ellrott, et al, “Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets”, **Cancer Cell**, 2024, link

Summary of subtyping diagnostic performances

- When all molecular types are considered, feature selection employs only mRNA features in 19 out of the 26 cancer types
- **mRNA profiles achieve the highest performances** when considered in isolation
- Between **70 to 150 samples are required for cancer subtype classification**





Building an AutoML system

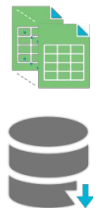
For Predictive Modeling and Knowledge Discovery

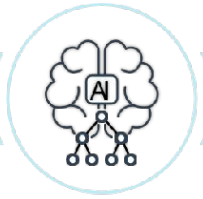
AutoML Architecture and Challenges

Tabular*, Images, or Signals **& their combinations**

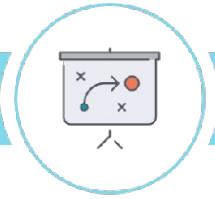
CASH: **C**ombined **a**lgorithm **s**election and **h**yper-parameter optimization

Data



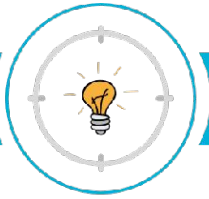


AI expert system that encodes the tacit knowledge of a data scientis



It produces an **analysis plan** with thousands of machine learning pipelines.

Pipelines executed in parallel on the **cloud**



Results are visualized, explained, and interpreted for decision support and **Knowledge Discovery**

Deploy

Decide which algorithms and machine learning pipelines are appropriate; decide which estimation protocol to run

l failures to i

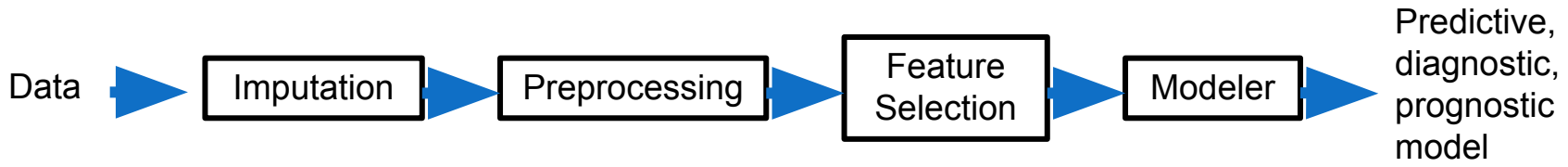
Estimate the predictive performance of the final model
Facilitate and optimize the

Self-improvement using meta-level learning

*multi-omics (e.g. transcriptomics, proteomics, methylomics, genomics), clinica

Step 1: Determine the set of analysis algorithms' options

Choices for algorithms



Which algorithms should I try for each step?

Which hyper-parameter values should I try for each

- Hyper-parameters affect the **sensitivity** of identifying patterns and correlations
- Instantiation of all choices for algorithms and their hyper-parameter is called a **configuration** (data --> model instance, a **machine learning pipeline**)

Solution 1: Represent existing knowledge

Good, Old-fashion AI

Ontologies describe the concepts and their properties

Rules describe the reasoning to perform

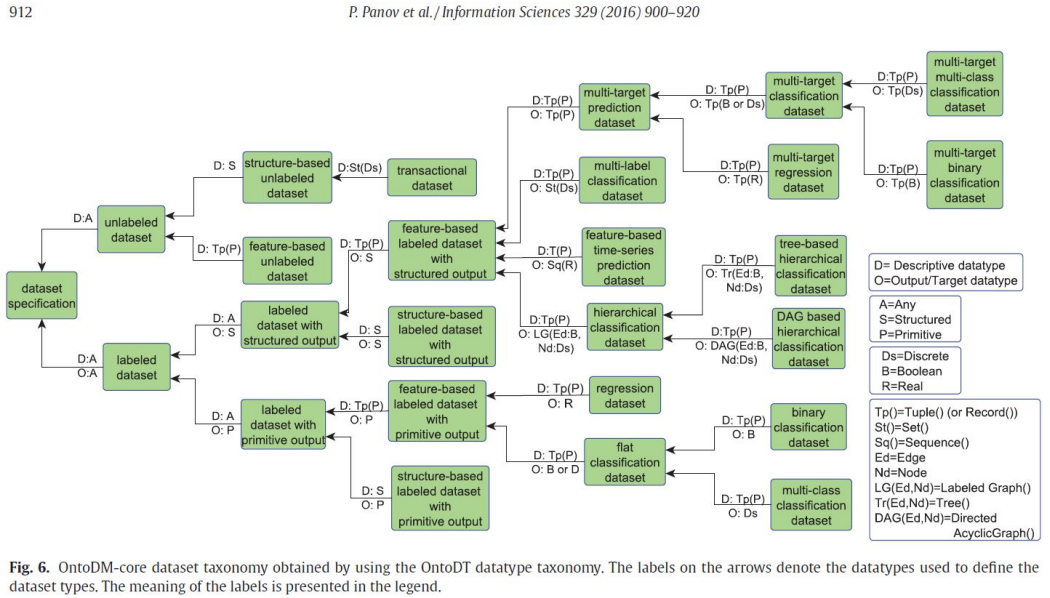


Fig. 6. OntoDM-core dataset taxonomy obtained by using the OntoDT datatype taxonomy. The labels on the arrows denote the datatypes used to define the dataset types. The meaning of the labels is presented in the legend.

```
IF task=classification THEN algorithm=RandomForest
```

Solution 2: Learn new knowledge

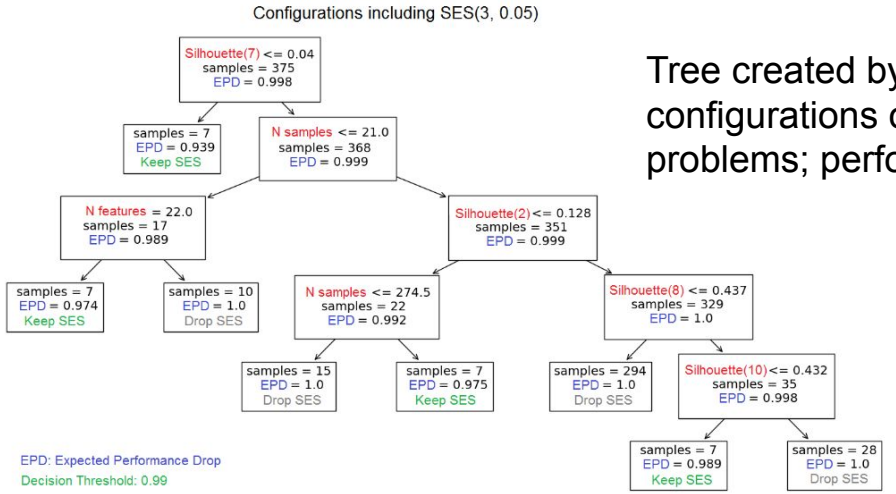
Meta-Level learning or Meta-learning: apply ML to predict which method/pipeline will lead to the best model

Meta-Level Learning makes an AutoML system self-improving

	Info about dataset		Info about the analysis			Info about performance	
Task	# samples	# features	Pre-processing	Feature selection	Classifier	AUC	Accuracy
1	12879	452	None	No	SVM	0.91	0.89
2	235	12000	Normalize	Yes	RF	0.87	0.67
...	...	...	...	...	...	...	...

ML Engine

Predict which choices can be removed without a drop in performance



Tree created by executing 8633 configurations on 375 regression problems; performance metric is R^2

Samples: the number of samples in the node supporting the decision

*Borboudakis “A Meta-Level Learning Algorithm for Sequential Hyper-Parameter Space Reduction in AutoML”, et. al. 2024 Link

Current **JAD**Bio Solution to Defining the Configuration Space

- AI Decision Support System: Ontologies + Rules
- Meta-Level Learning : Sequential Hyper-Parameter Space Reduction algorithm to remove non promising choices

Step 2:
Tuning and Optimization of Configurations

Challenge: Identify the optimal configuration

1. How do we estimate how “predictive” is a configuration?

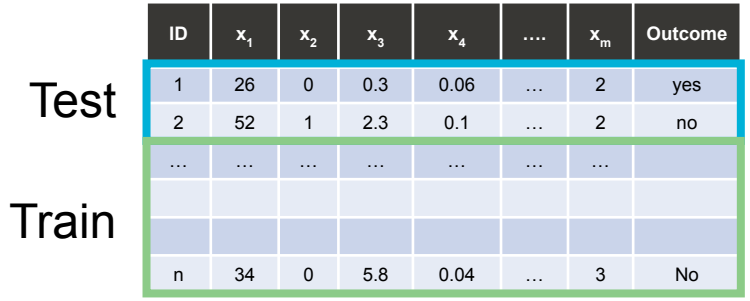
2. How do we intelligently search in the space of configurations?

1. Performance evaluation for single configurations

- Typical **out-of-sample** protocols
  1. **Hold-out**: do it once
  2. **Cross validation (CV)**: partition data to multiple, non-overlapping Test sets (folds), train on all but one-fold, test on held out fold, repeat, and average
  3. **Repeated Cross-validation**: repeat CV for multiple partitions to folds
- Proper protocol selection is **paramount** in delivering the optimal model

JADBio considers the outcome imbalance, total sample, and user preferences to **automatically select the protocol.**

Hold-out

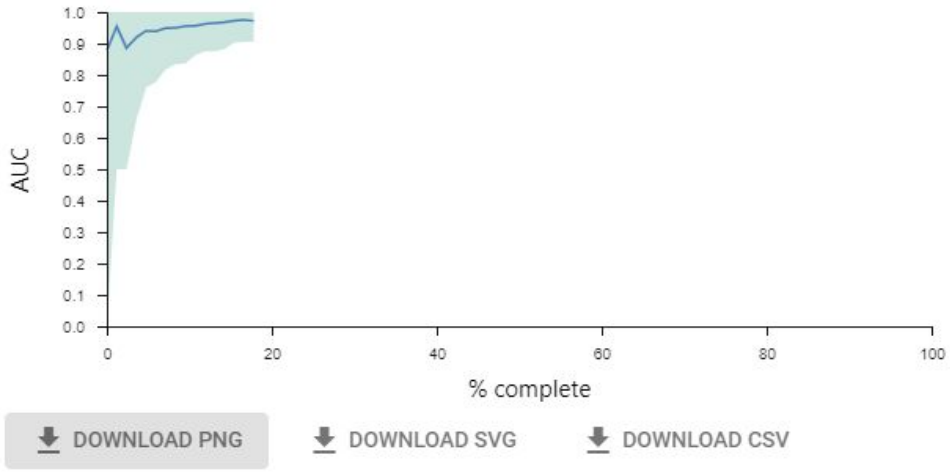


Cross Valid.

2. Searching for optimal configuration

- **Simple grid search**
  - Evaluate all pre-selected configurations
  - Continuous hyper-parameters are discretized
  - Other more intelligent search methods are available (e.g., Bayesian Sequential Optimization)

- **Dynamic heuristic rules** to stop cross-validation efforts
  - Stop cross-validation **for a given configuration**, when it is deemed inferior to the best current configuration
    - Early dropping heuristic when sample size is “enough”
    - Statistical test determines when to drop
  - Stop repeating cross-validation **overall** when there is no progress in shrinking the confidence intervals

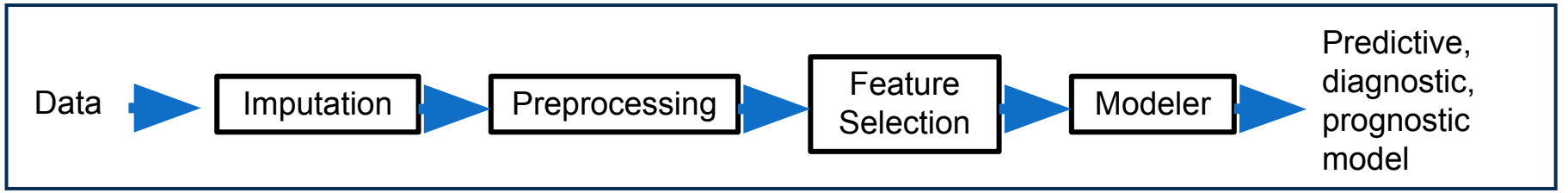


Stage: Analyzing...
Status: 21229 / 119200 models have been trained
Best configuration so far:
Preprocessing: Constant Removal, Standardization
Feature selection: Test-Budgeted Statistically Equivalent Signature (SES) algorithm with hyper-parameters: maxK = 2, alpha = 0.05 and budget = 3 * nvars
Model: Support Vector Machines (SVM) of type C-SVC with Polynomial Kernel and hyper-parameters: cost = 1.0, gamma = 0.1, degree = 2

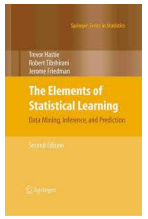
SHOW RESULTS

Note! Configurations are atoms

CV as
<u>one</u> step



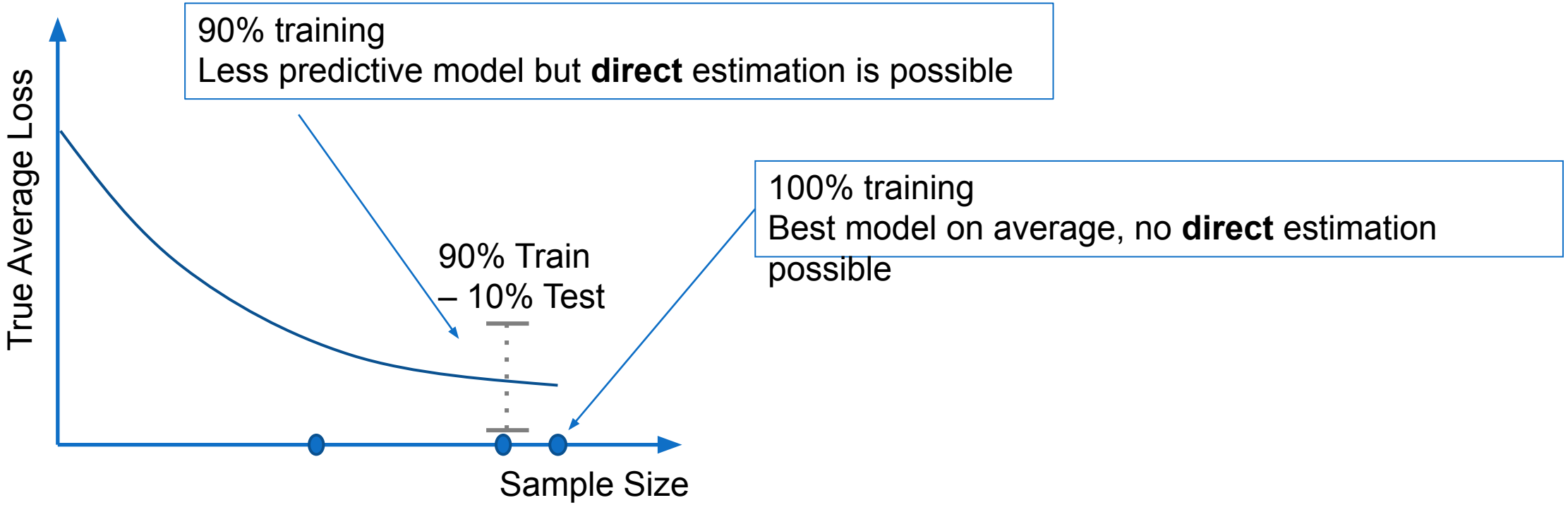
- **All steps of the analysis are cross-validated as one**
- Or be doomed to overfit! One of the most common mistakes in applied data analysis.
- Eye-opening example in page 245 of Hastie et al. book
  - True performance: 50% accuracy
  - Incorrect estimation of performance: 93% accuracy



*Hastie, T., Tibshirani, R. & Friedman, J. H. The elements of statistical learning : data mining, inference, and prediction. (Springer, 2016)

Step 3:
Produce a final model

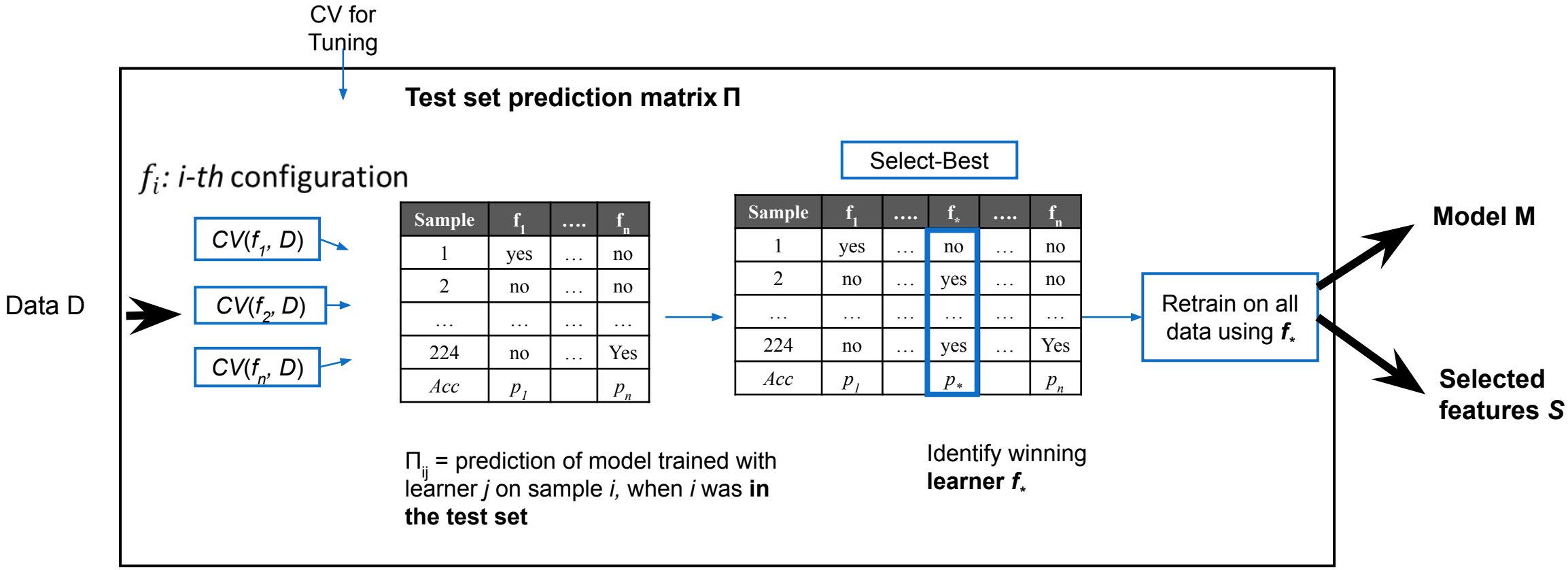
Learning Curve of a Configuration



Return best possible model (on average) using all data as training

Use the estimation of some other model from the same configuration as a proxy

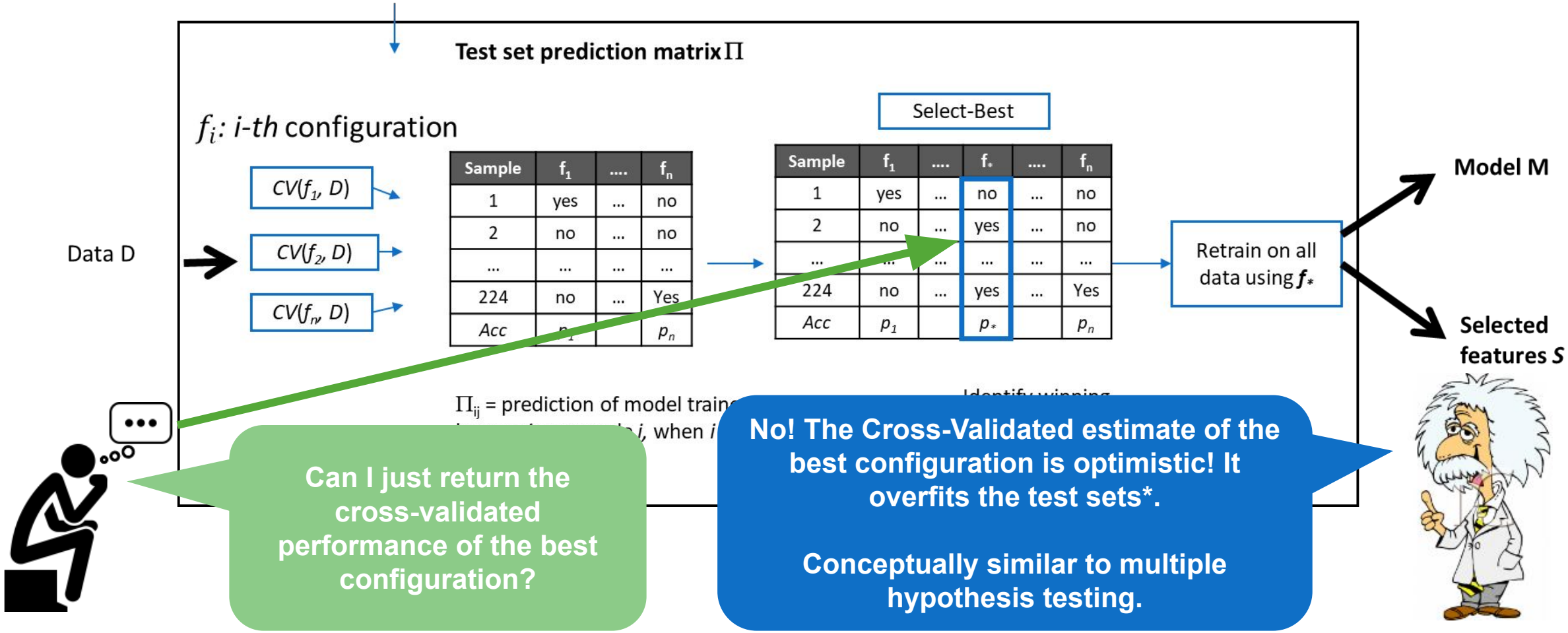
Cross-Validation with Tuning (CVT)



- Use the **winning configuration** on **all available data** to for the **final model instance** and the **final selection of features**
- On average learns the best possible model; but we have **no samples left** to estimate its performance!

Step 4:
Estimate Performance of the final model

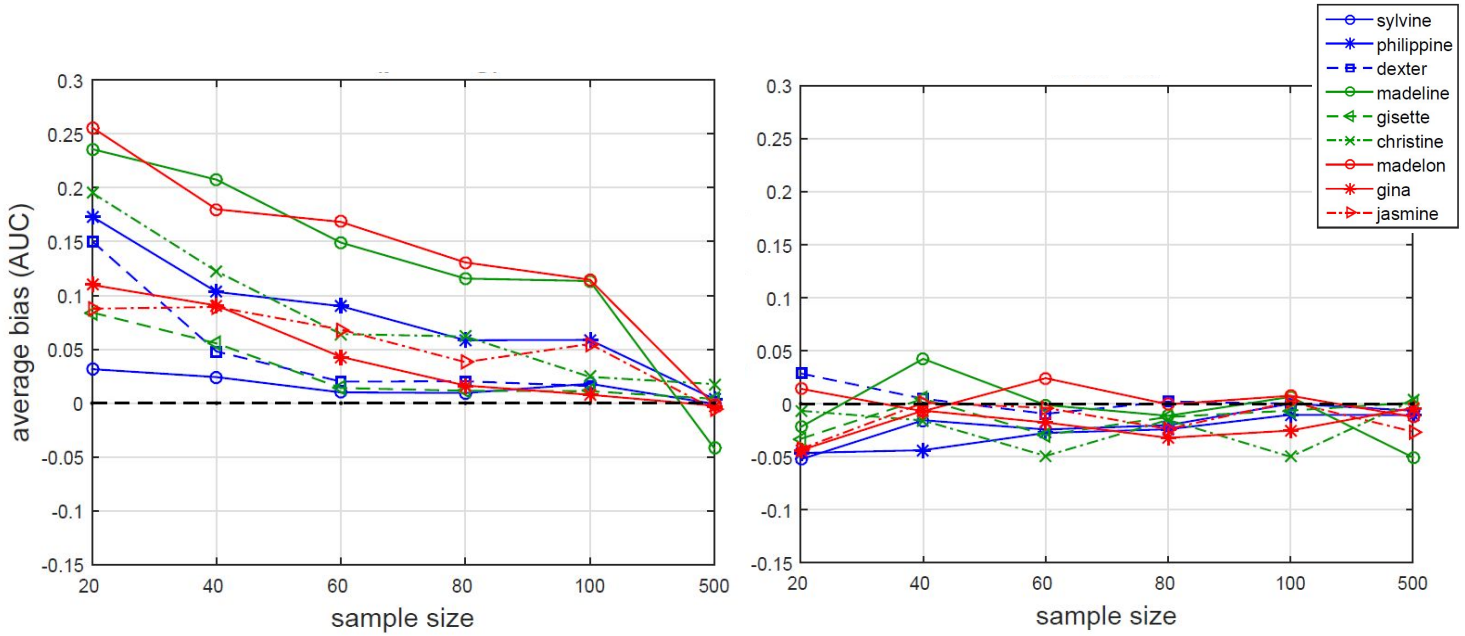
The “winner’s curse”



*This is called the multiple comparison in induction algorithms problem in machine learning [D. Jensen 1992, Ph.D.] and the “winner’s curse” in statistics.

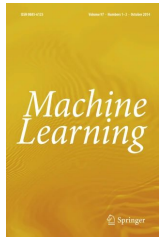
Bootstrap-Bias Correction removes estimation bias from the best model performance

- **Remove the bias of estimation due to multiple tries** (adjusted performances)
- Also computes **Confidence Intervals**
- **Minimal computational cost**
- Allows machine learning with very small sample sizes (< 100)
- Does need a separate hold out! **<u>Does not lose any samples to estimation</u>**



Cross Validation With Tuning - **CVT**

BBC-CV



*Tsamardinos, et. al, “Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation”, **Machine Learning Journal**, 2018, Link

Feature Selection for Knowledge Discovery

Challenge: Incorporate feature selection

Feature Selection Problem(s)

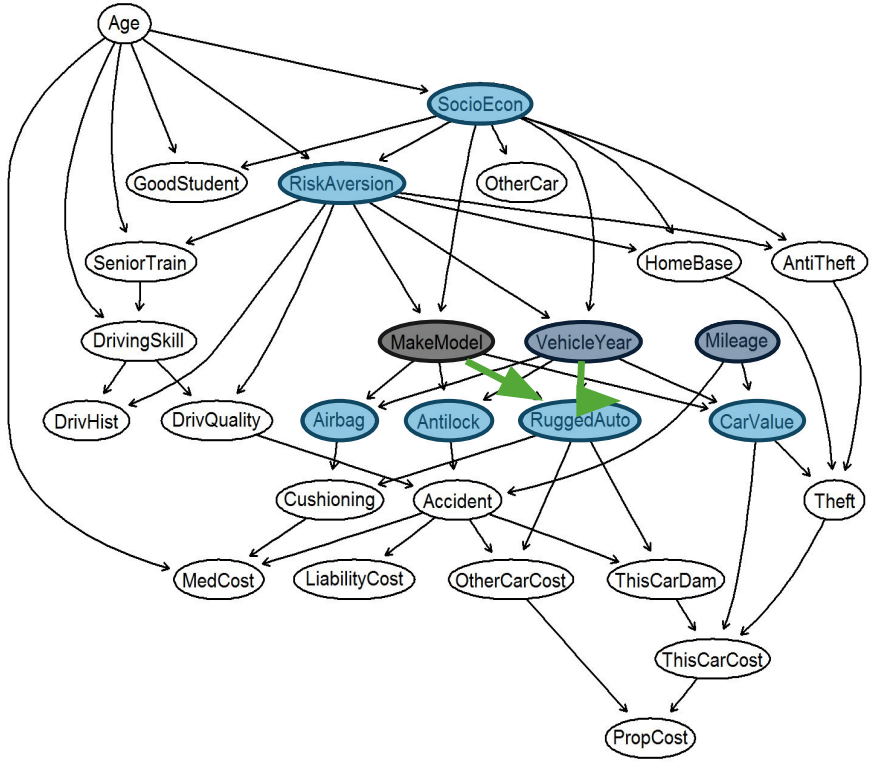
- **Single Feature Selection**: Find **one** **minimal** set of features that **collectively** carry all the information for **optimal** prediction
  - **Minimal**: Throw away irrelevant or superfluous features
  - **Collectively**: May need to consider interactions
  - **Optimal**: Requires constructing a classification / regression model and estimating its performance
- Feature selection **removes** both **irrelevant** but also **redundant** features/biomarkers
- **Multiple Feature Selection**: Find **all** such sets

{AFFX-BloB-5_at, AFFX-Blob-3_at}
{AFFX-BloC-5_at, Affx-Bloc-3_at}

	Genes / Probe Sets							
Sample	AFFX-BloB-5_at	AFFX-BloB-M_at	AFFX-Blob-3_at	AFFX-BloC-5_at	…	Affx-Bloc-3_at	AFFX-BloDn-5_at	Metastatic?
1	123.00	1.00	2,3	12.00		23.00	34.00	Yes
2	323.00	23.00	4,54	2.00		21.00	65.00	No
								No
								No
N	~~232.00~~	4,5	~~23.00~~	0,55		~~75.00~~	343.00	Yes
	Expression Values							

Knowledge Discovery = Feature Selection

- The **primary task** in many applications
- Deeply related to the **causal mechanisms** of the outcome
  - The solution to the Feature Selection is the **neighbors** (direct causes and direct effects) of T and the **spouses** of T in the (unknown) causal network
    - Spouses = nodes with common direct effects
- Feature selection discovers a local causal neighborhood of the outcome



[Tsamardinos and Aliferis, 2003, Aliferis et al. 2010]

Multiple Feature Selection

- **Single** FS can be **misleading**
  - Informing a biologist that only genes in S are “important” when other genes S′ could replace them!
  - Cost-aware feature selection: when features have measurement cost, give options what to measure

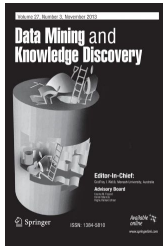
- Single Feature selection suffices for predictive purposes; multiple feature selection is required for Knowledge Discovery

- Much less studied problem
  - See KIAMB, [Peña et al., 2007)], TIE* [Statnikov et al., 2013], SES [Tsamardinos et al., 2012], [Lagani et al., 2017], [Borboudakis Tsamardinos, DAMI 2021], ChronoEpilogi [Vareille, et. al. NeurIPS 2024]

Replaceable Features and Informationally Equivalent Features

- If biomarker *A* and *B* can replace each other in an optimal model for *Y*, they are **informationally equivalent** w.r.t. *Y*

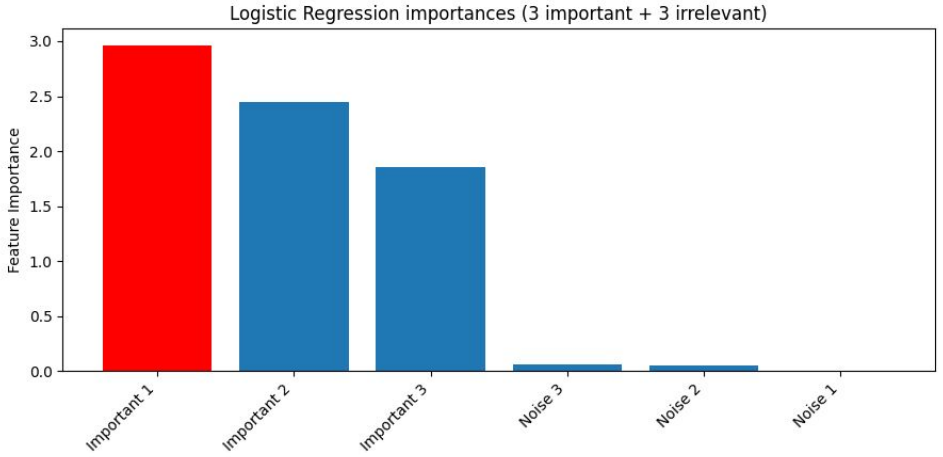
- Higher-order equivalences are possible: set {Body-Mass-Index} is often informationally equivalent with {Mass, Height}

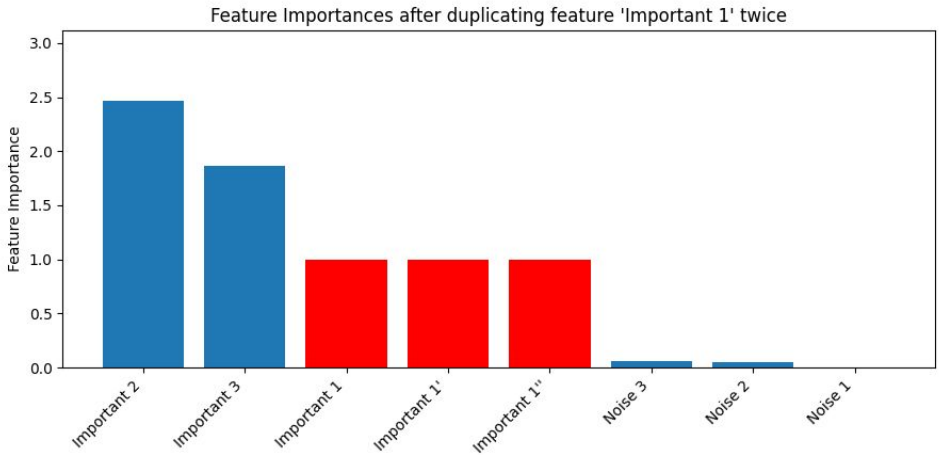


*Borboudakis, G., Tsamardinos, I. “Extending greedy feature selection algorithms to multiple solutions”. **Data Mining and Knowledge Discovery** (2021), Springer Nature Link

Feature Importance is Misleading

- Feature importance computes the contribution of each feature to a predictive model's performance.
  - **Commonly** used technique to explain models.
- **Feature "importance" breaks down with informationally equivalent features**
  - Selecting the top *k* more stable or more "important" features is **wrong**!
- **We need to consider informational equivalence before computing importance**





Advances in Feature Selection

- **Scalability**
  - **UP** to **tens of millions of features** and/or hundreds of millions of samples (Big Data) [1]
  - **Down** to **very few samples** (20-30) [2]
- **Generality**: Consider **any outcome**
  - Nominal, ordinal, continuous, count, time-to-event, zero-inflated continuous, time-course repeated measurement
- Identify **multiple, statistically equivalent solutions** [2,3]
- **Theoretical guarantees** and characterized properties
- Performance on par with Lasso for the same number of selected features.

- **Typically, return just a handful of biomarkers for most omics problems**

Algorithm	Outcome	#Samples	#Features	#Solutions	Comments	Guarantees	Ref.
SES	Small sample, hundreds of thousands features						[2,4]
εpilogi	Medium Sample, tens of millions of features						[6,7]
FBED	Medium Sample, tens of millions of features						[5]
PFBP	Infinite Sample, tens of millions of features						[1]
ChronoEpilogi	Timeseries Selection. Medium time-horizon, hundreds of thousands of timeseries						[8]

Mature feature selection technology exists for any analysis scenario

Publicly available: R Package MXM

1. Tsamardinos et al., A greedy feature selection algorithm for Big Data of high dimensionality, Mach Learning (2019).
2. Lagani et al., Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets, Journal of statistical software, (2017)
3. Borboudakis and Tsamardinos, Extending greedy feature selection algorithms to multiple solutions, Data Min Knowl Disc (2021)
4. Tsamardinos et al., Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations, KDD (2003)
5. Borboudakis et al., Forward-Backward Selection with Early Dropping, Journal of Machine Learning Research (2019)
6. Tsagris, et al., The γ-OMP algorithm for feature selection with application to gene expression data, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2022)
7. Lakiotaki, et al., Automated machine learning for Genome Wide Association Studies, Bioinformatics, (2023)
8. Vareille, et al., ChronoEpilogi: Scalable Time Series Selection with Multiple Solutions, NeurIPS (2024)

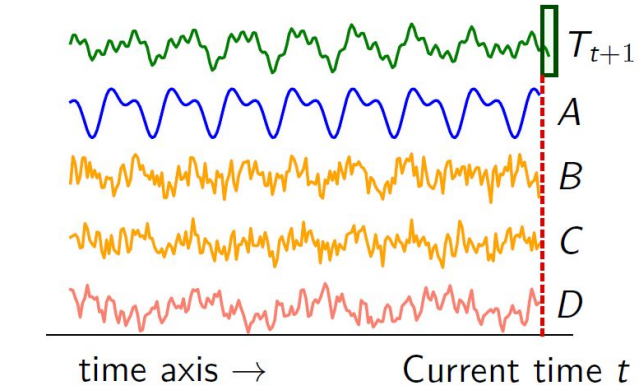
ChronoEpilogi for Timeseries Selection

Problem:

- given multivariate time series data to forecast **timeseries *T***
- select **all minimal-size subsets of timeseries** that lead to **optimal forecasting** of ***T***
- Suitable for Energy and Environmental timeseries

ChronoEpilogi: greedy, efficient, and general algorithm with theoretical guarantees

- on par forecasting performance with GroupLasso, but more efficient
- guaranteed to return all equivalent solutions under broad distributional conditions.
- easily adaptable to different data types
- scales to hundreds of thousands of timeseries



Optimal sets: $\{T, A, B\}$ and $\{T, A, C\}$

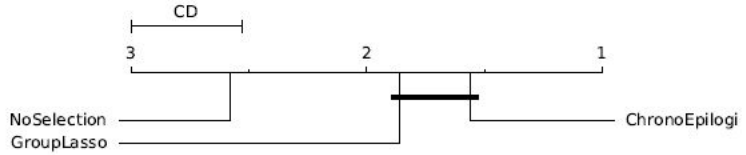


Figure: Critical difference diagram of predictive performances (R^2) of a downstream forecasting model

*Vareille, et al. "ChronoEpilogi: Scalable Time Series Selection with Multiple Solutions", **NeurIPS**, (2024)
Link

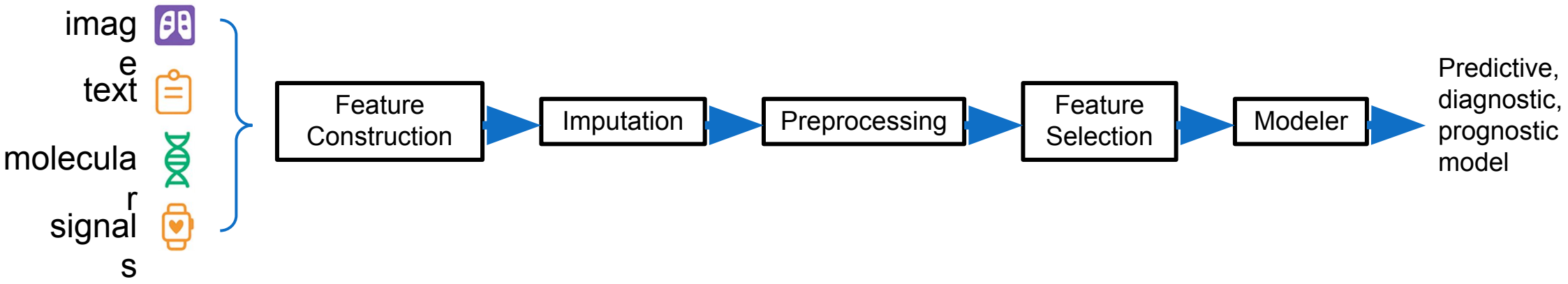
Dealing with Complex Data Types

Complex Data Types: Images, Text, Graphs, Chemical Compounds

Complex Data Types

- **Standard Approach: Use Deep Learning**
  - Requires large sample size, deep expertise, easy to overfit, time consuming, computationally intensive
  - **+ high quality results when done properly**
- **Add feature construction as a step in AutoML configurations**
  - Use pre-trained networks for feature construction (**transfer learning**)
  - Optimize the feature construction step, just like all other choices
  - **JAD**Bio optimizes over 3 publicly available foundational **image** models
  - **ChemX** (QSAR AutoML product from Denmark) featurizes chemical compounds then uses **JAD**Bio
  - **+ Fast and fully automatic**
  - **+ If you can use the mouse, you can do it**

Compare with fine tuning a DNN vs JADBio+Foundational Model



Key Ingredients for Performance and Correctness

- Cross-validating the whole pipeline as an atom avoids overestimation of predictive performance
- Final model produced on all available data
- Extensive cross-validation (stratified, repeated cross-validation) for small samples/rare classes improves model selection
- Properly dealing with imbalanced/rare classes
- Carefully crafted knowledge that determines the configuration space
- Good-quality, scalable, and multiple feature selection
- Optimize over feature constructors.
- Self-improvement using meta-level learning

Selected Use Cases

Validated results of **JAD**Bio in the lab

- **Prof. Nikos Tapinos,** Brown University, Neuroscience & Neuro-Oncology
  - Detected miRNA blood biomarkers of **glioblastoma**
  - **Secured funding** for developing cost-effective screening test

ROBERT J. & NANCY D. CARNEY
INSTITUTE FOR BRAIN SCIENCE
BROWN UNIVERSITY

- **Prof. Ekaterini Chatzaki,** Democritus University, Faculty of Medicine
  - Innovative liquid biopsy method for **diabetes detection**.
  - Founded a **spinoff**, **Patent** submitted, Secured MIT-based angel **funding**

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

ABCURED
Connecting the genetic dots with AI

- AIDA oncology
  - Accurate **epirubicin response signature** for breast cancer patients

AIDA

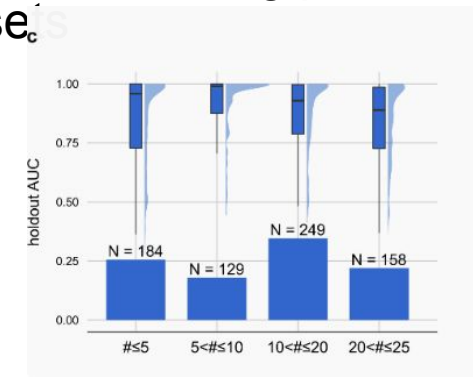
- Dr. Theodora Katsila
  - Drug repurposing based on multi-omic biomarkers (Patent rights acquired by a US Inc.)

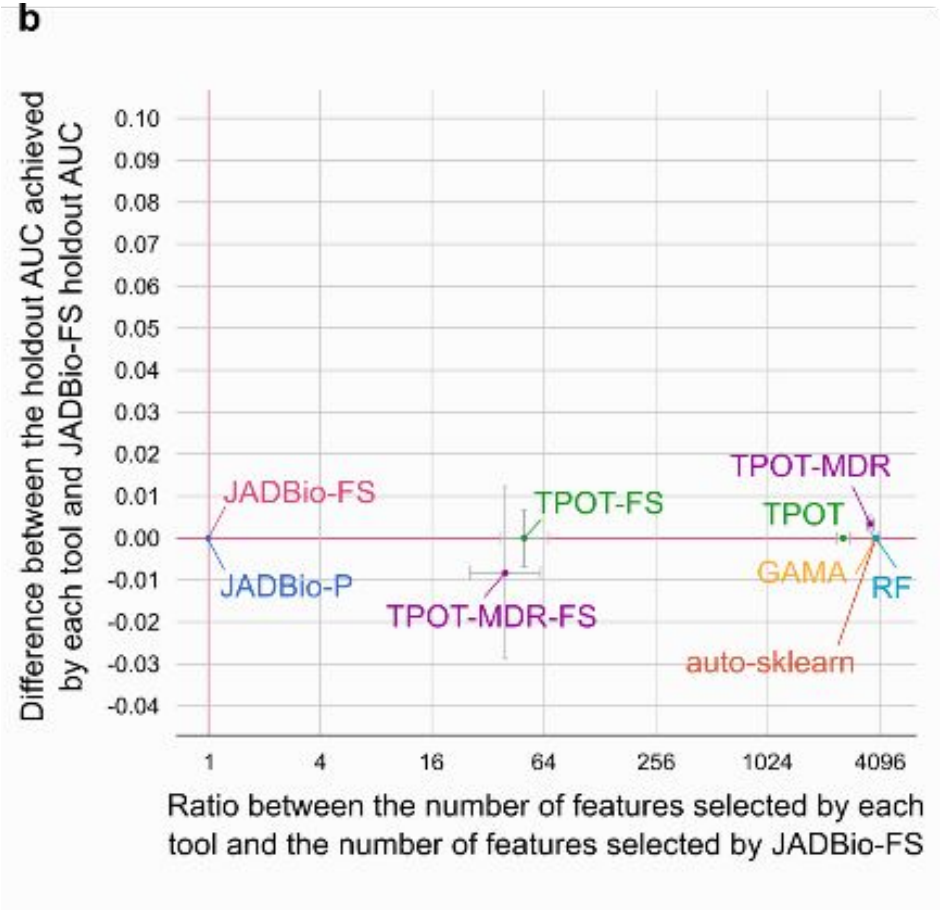
ΕΘΝΙΚΟ ΙΔΡΥΜΑ ΕΡΕΥΝΩΝ
National Hellenic Research Foundation

npj | precision oncology

JADBio comparative evaluation on high-dimensional omics data

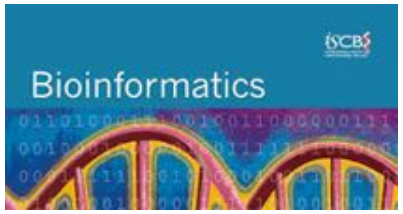
- **Question**: How **JADBio** performs against the state-of-the-art in AutoML for omics?
- **Data**: 360 unique, public, datasets of genes and proteins (#features ranging from 15,000-50,000)
- **Result**: JADBio reduces the number of Biomarkers by a factor of 4,000 (to fewer than 20) maintaining performance quality.
- **Conclusion**: possible to reduce the number of features without losing performance in omics datasets





Tsamardinos, I., Charonyktakis, P., Papoutsoglou, G. *et al. npj* ***Precision Oncology***. **6**, 38 (2022). https://doi.org/10.1038/s41698-022-00274-8

Case study: GWAS using **JAD**Bio



Lakiotaki, et al., Automated machine learning for Genome Wide Association Studies, **Bioinformatics**, (2023), Link

- **Problem**: Discover causally-related, predictive variants (SNPs)

Alcrowd challenge
(height prediction)

- Samples: 784 individuals, separate test set
- Features: 6,854,199 genetic variants

JADBio	SOTA
R^2=0.495 selected 50 SNPs (max allowed)	R^2=0.53 (using prior knowledge)

On disease-related datasets

	JADBio	Standard GWAS
Disease	**Predictive Performance – AUC [CI_{AUC}]**	
Ankylosing Spondylitis	0.89 [0.86 – 0.91]	0.61 [0.57 – 0.65]
Multiple Sclerosis	0.82 [0.80 - 0.85]	0.59 [0.55 - 0.66]
Parkinson's	0.76 [0.73 – 0.79]	0.57 [0.53 – 0.61]

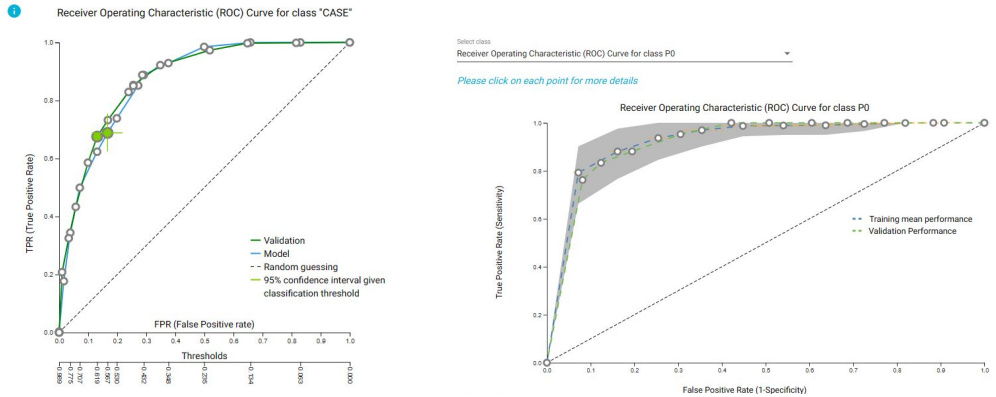
- Against SOTA: JADBio **scales** to millions of molecular measurements **without overfitting**.
- Biomarker discovery: **Discovered multiple equivalent variants that are more predictive** than standard GWAS variant selection

(Large) Microbiome data analysis

STOOL TEST

University of Ljubljana

- **Question**: Can we diagnose Non-communicable diseases (NCDs) from microbiome data?
- **Data**: 12,121,728 species - 3,480 samples (training), 1,489 samples (testing) - 17 classes
- **Result**
  - Binary classification: AUC 0.871 (train), 0.878 (test) – 100 features
  - Multiclass classification: mean AUC 0.844 (train), 0.86 (test) – 100 features
- **Conclusion**: FS scales to millions of features without overfitting



Performance on test data

Sample name	mgshot_S5084Nr1.x_taxonomic_profile
Prob (class = Healthy)	**0.709692835**
Prob (class = IGT imparied glucose tolerance)	0.069851338
Prob (class = Obesity)	0.061362275
Prob (class = Overweight)	0.05091827
Prob (class = CRC colorectal cancer)	0.034201082
Prob (class = T2D type two diabetis)	0.030716836
Prob (class = advanced adenoma)	0.022314261
Prob (class = Ulcerative colitis)	0.00702411
Prob (class = Symptomatic atherosclerosis)	0.005448599
Prob (class = Underweight)	0.005245175
Prob (class = Rheumatoid Arthritis)	0.001948884
Prob (class = Crohns disease)	0.001085616
Prob (class = ACVD Atherosclerotic cardiovascular disease)	1.91E-04

Predicted probabilities for class:

Class	Probability
Bladder	0.000
Breast	0.004
CNS	0.114
Colorectal	0.081
Leukemia	0.149
Lung	0.049
Lymphoma	0.106
Melanoma	0.000
Mesothelioma	0.030
Ovary	0.120
Pancreas	0.067
Prostate	0.142
Renal	0.076
Uterus_Adeno	0.062

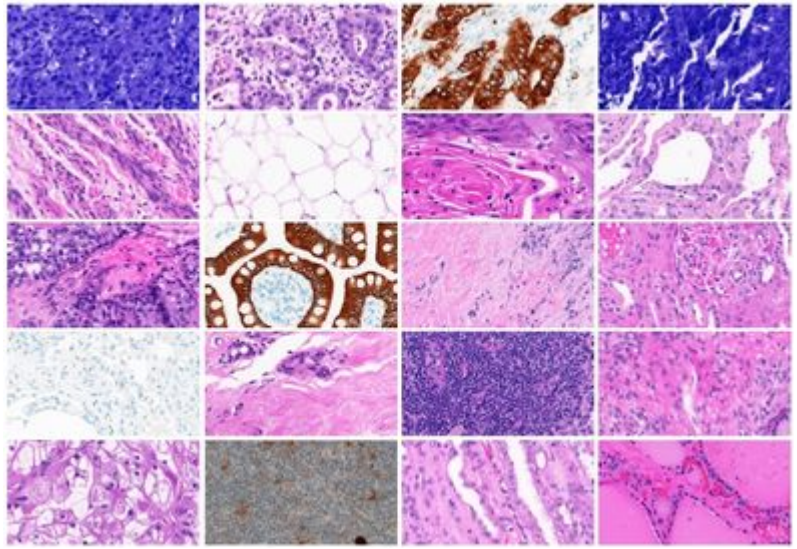
Output on new samples

Histopathology dataset: Kimia Path960

Samples:
960 whole slide images (WSIs) of muscle, epithelial and connective tissue

Potential predictors:
>200 automatically constructed features (transfer learning)

Task:
Classify 20 balanced classes (20 different textures/patterns)

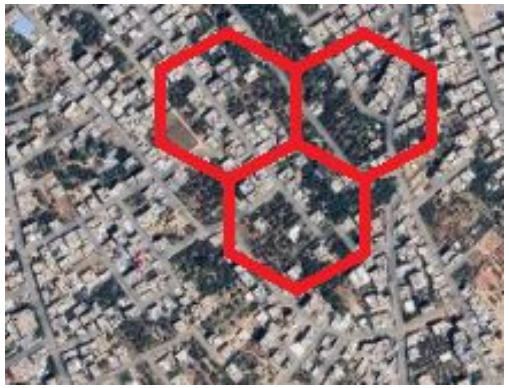


JADBio outperforms Alhindi et al.* in terms of Accuracy achieving 0.995 versus 0.8114

*Alhindi, Taha J., et al. "Comparing LBP, HOG and deep features for classification of histopathology images." *2018 international joint conference on neural networks (IJCNN)*. IEEE, 2018.

Damage forecasting in Gaza and Lebanon

- Collaboration with the **United Nations Development Program**
- What is the impact of war on city infrastructure - streets and buildings? Important for preparing humanitarian aid.
- Data: Individual incidents aggregated in time periods and geographical hexagons.



Predict future destruction intensity, given the characteristics and the evolution of the strikes so far.

Model performance: R^2_{oos} = 15% (r = 38%)

R^2_{oos}: out of sample coefficient of determination

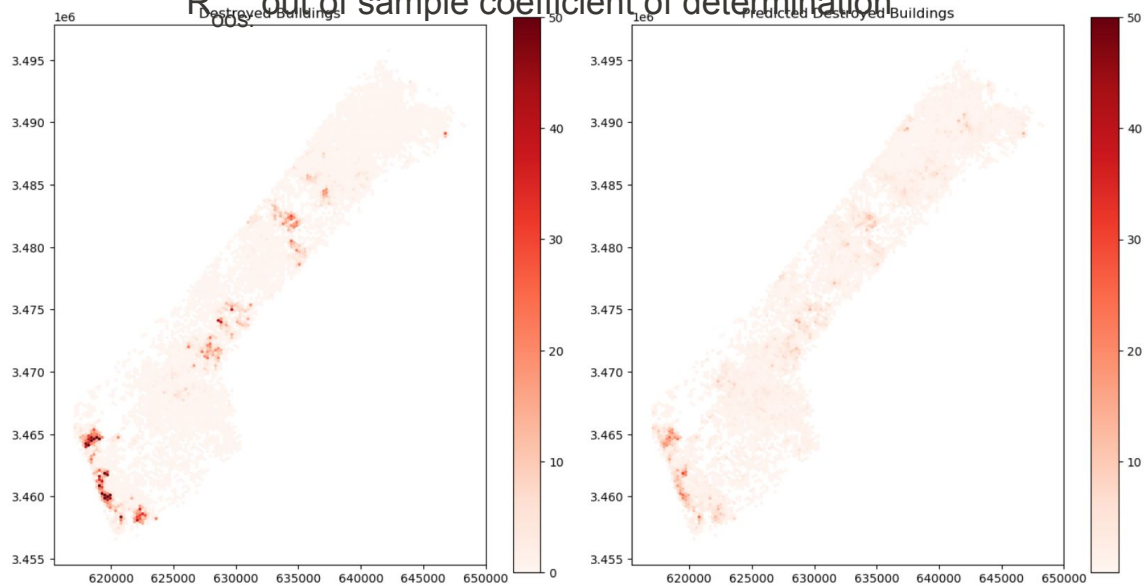


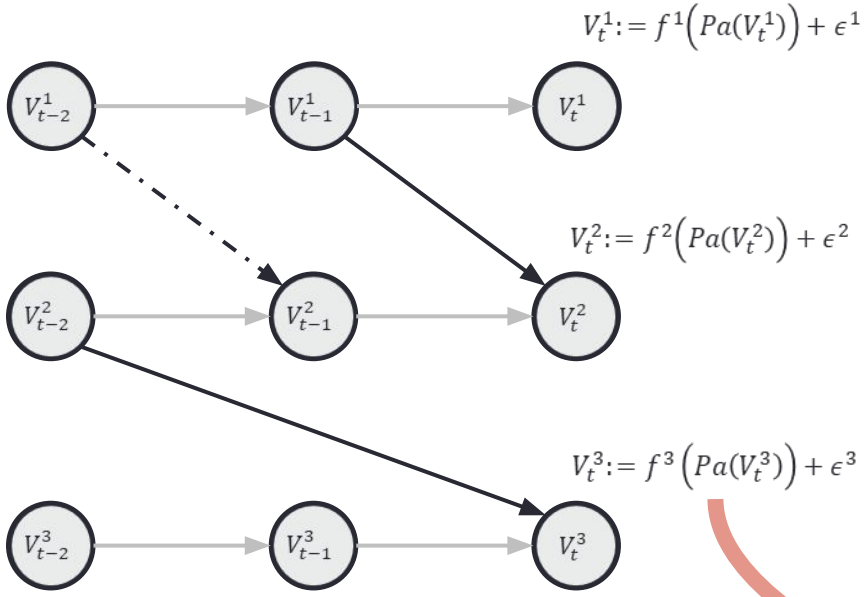
image prepared by Dr. Hamed Sambo and Gaia Rigodanza

JADBio and Feature Selection Summary

- Fully automated, robust, and general
- Results have been validated in the lab
- In massive benchmarking against other AutoML tools shows on par performance …
- but it requires only a handful of markers for typical omics data
- Returns multiple sets of features (signatures)
- Works with images, chemical compounds, (and medical signals)
- Scales up to tens of millions of features, scales down to few samples

Causal Discovery and Automated Causal Discovery

Temporal Structural Causal Models (TSCM)



- Edges = **direct** causal relations
  - no other observed quantity mediates the causation
- Directed paths = causal relations

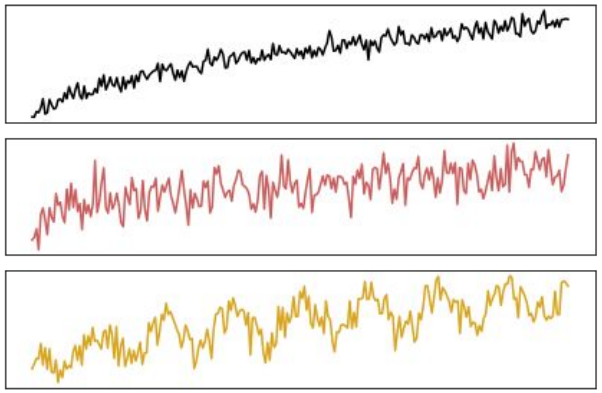
- Temporal Structural Causal Model $TSCM(G, F, \mathrm{E})$
  - G: causal graph (qualitative part)
  - $F = \{f^1, \dots, f^n)$ causal functional dependencies
  - $\mathrm{E} = \{\epsilon^1, \dots, \epsilon^n\}$ noise distributions

} quantitative part

V_t^i : value of quantity *i* at time-point *t*
Pa(V) : parents (direct causal effects) of V in the graph
L : maximum time-points (lags) of the model

Assumptions:
Causal stationarity : the causal structure or functional dependencies do not change over time
Causal sufficiency : no latent confounders
No contemporaneous causation

Ancestral Sampling



Causal Models vs. Predictive Models

Query	Causal	Predictive
Prediction: Predict/Diagnose Y given X	✔	✔
What if Scenario Exploration: What-if I set X_1=5	✔	☐
Optimal decision making (related to counterfactual explanation): What is the optimal value to set X_1 to increase Y given $X_2 = 6$?	✔	☐
Explanation/Interpretation: What is the importance of X_1 in affecting Y ? What are the features that affect Y ?	✔	
Root-Cause-Analysis: What was the root cause of a failure	✔	☐
Causal Counterfactual Reasoning: A client *i* has $Y=3$, when the GUI color was X_3 = *"yellow"* i. What would be her Y if the GUI color was X_3 = *"green"* (counterfactual reasoning)	✔	

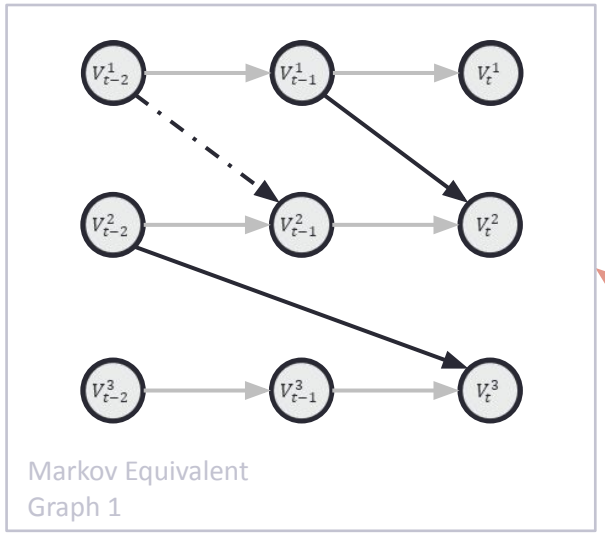
LLMs and Causality

*"Without an understanding of causality, LLMs may produce outputs that are contextually relevant but **not logically sound**, leading to potential issues such as **hallucinations**, **biased outputs**, and an **inability** to perform well on decision-making tasks that depend on causal relationships. Incorporating causality into LLMs is essential for several reasons."*

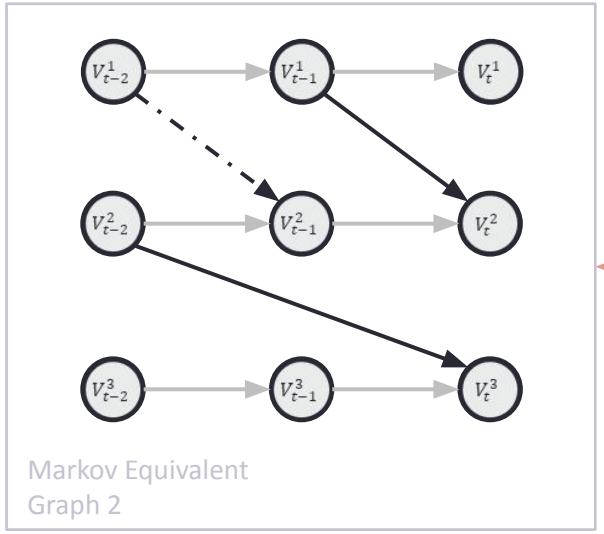
[Wu, Kuang, et. al. Causality for Large Language Models, arxiv, 2024]

Causal modeling is necessary for decision making, root-cause analysis, counterfactual reasoning, and LLM training.

Causal Discovery

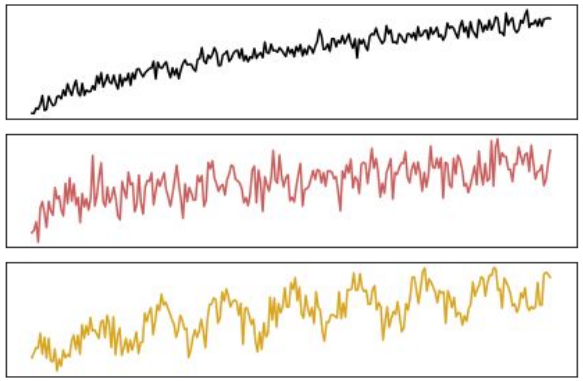


- **Numerous Causal Discovery algorithms** in the literature
  - learn causal models from data
- **Markov Equivalent causal models**
  - fit the data equally well
  - form a Markov Equivalent Class (MEC)



Causal Discovery Algorithm

Measured Data



Correlation-based decision making is broken

- Decisions, decisions, decisions: “How to optimally set the configuration parameters of my 5G network?”
- Data analytics and predictive models rely on **correlation**, not **causation**
- Leading to **wrong conclusions** and **costly decisions**

Real-world failures:

Predictive models estimated a **+1400% ROI** from advertising, while causal analysis revealed the **true ROI was –63%** [1]

“*using naive correlation, our price targets would have been **50% to 250% higher than the true value.** Causal analysis eliminated the guesswork …*”[2]

[1] Beyond prediction: Using big data for policy problems | Science
[2] The Importance of Being Causal · Issue 2.3, Summer 2020

Vision for Automated Causal Discovery: ETIA

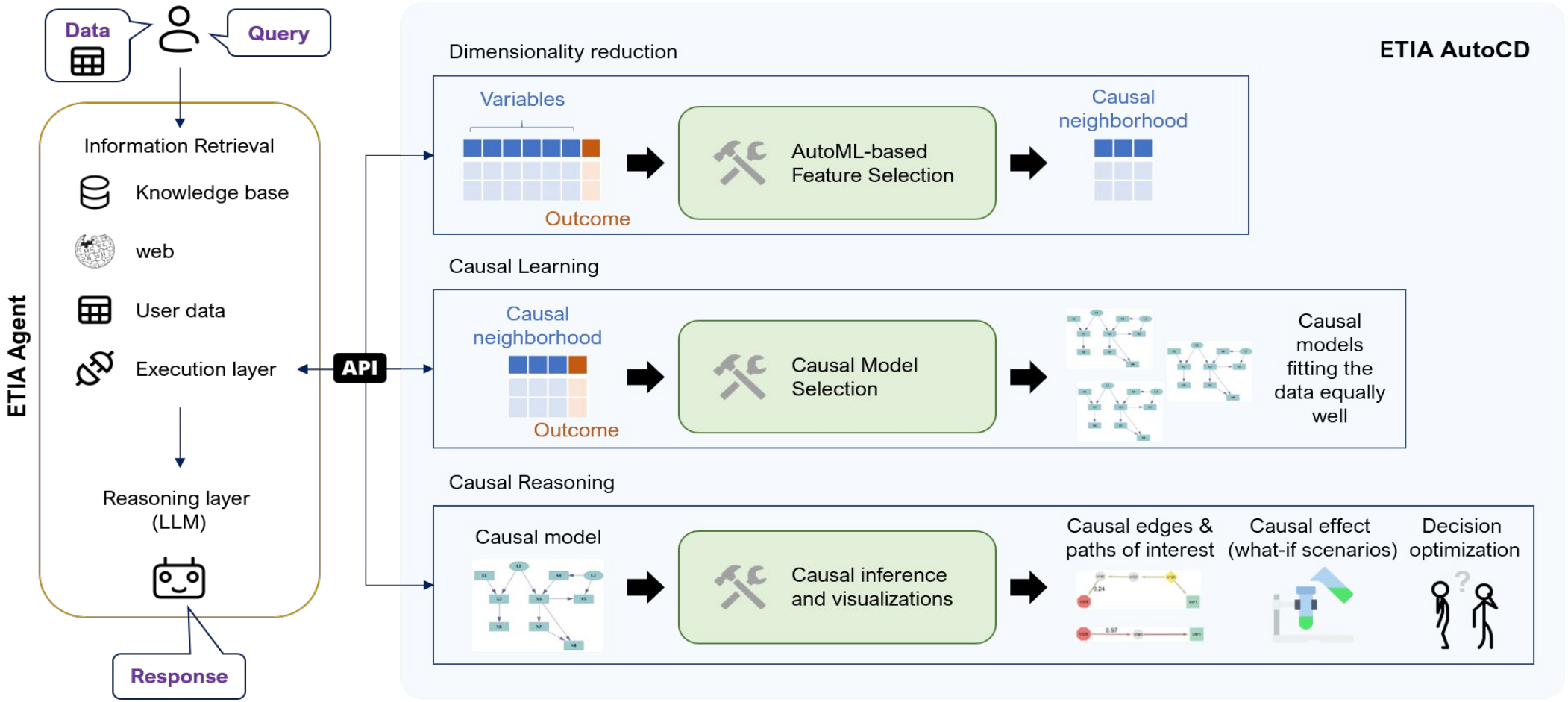
- ETIA: “cause” in Greek (female form)
- **Automated Causal AI** platform
  - Deliver **rigorous causal discovery from data** and **inference** (**what-if scenarios**, **optimal decisions**, **root-cause-analysis**, **counterfactuals**)
  - Produce **correct**, **trustworthy** insights
  - Offer **Agentic AI** for intuitive, explainable interaction

- **Just awarded a European Innovation Counsil Transition grant (starting July 1st 2026)**
- **Early prototype available**

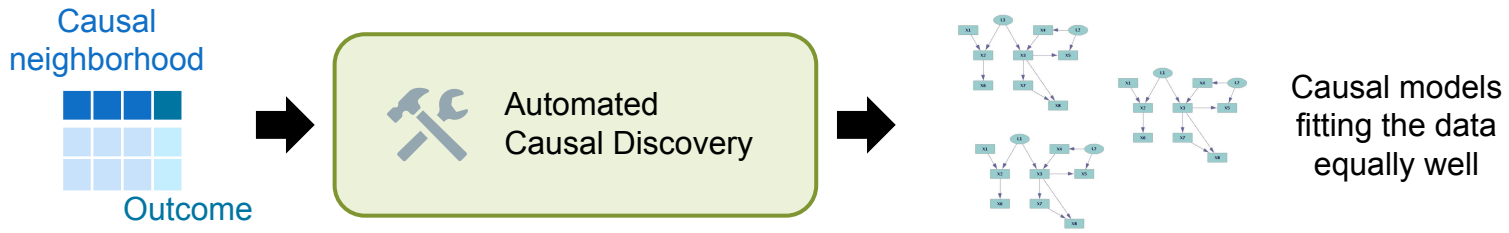
erc
European Research Council
Established by the European Commission

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

ETIA architecture



Challenge 1: Optimize the Causal Model



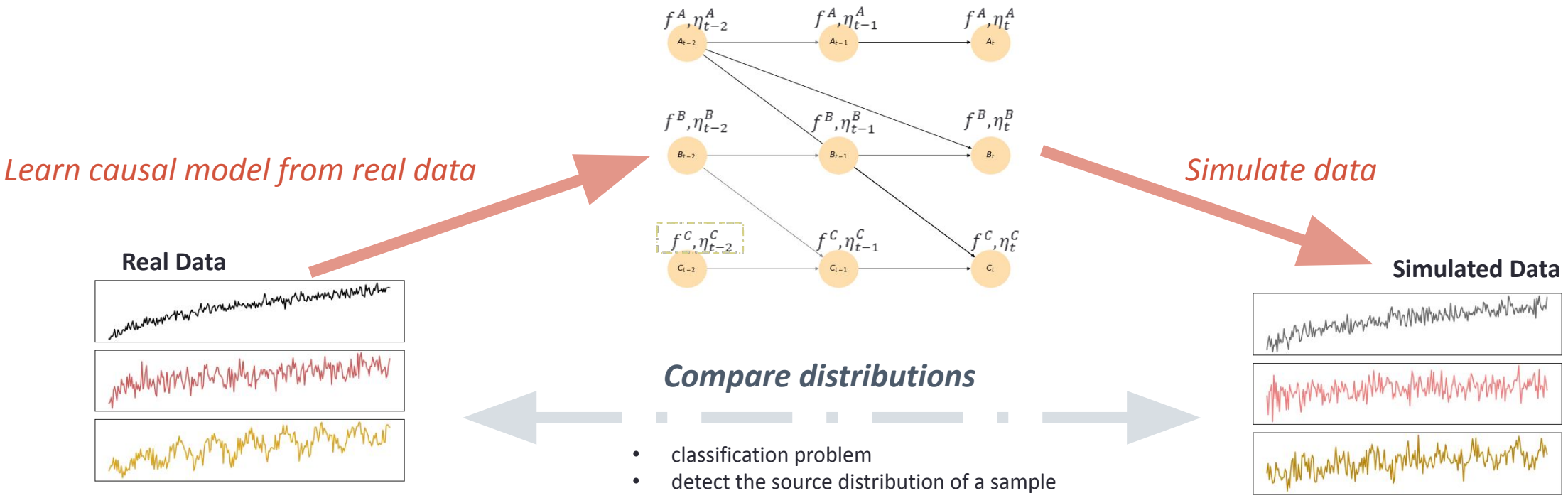
- **Problem**: No algorithm is best on every problem (depends on the data); Need tuning
- **Challenge**: Optimal algorithm selection is an **unsupervised** problem (AutoCD)
- **Idea**: Select the Causal Model that results in the most predictive local neighborhood
  - **Out-of-Sample Causal Tuning (OCT) :** *Best student paper award from the University of Crete for the academic year 2022-2023*

*Biza, et al., “Out-of-Sample Tuning for Causal Discovery”, IEEE TNNLS, 2022, Link

IEEE TRANSACTIONS ON
NEURAL NETWORKS AND
LEARNING SYSTEMS

Challenges: How to Select the Optimal Causal Model?

- **<u>Idea</u>: select the causal model that generates realistic data**



Challenges:

Which algorithm (and hyper-parameters) to use to learn the causal model? Optimize

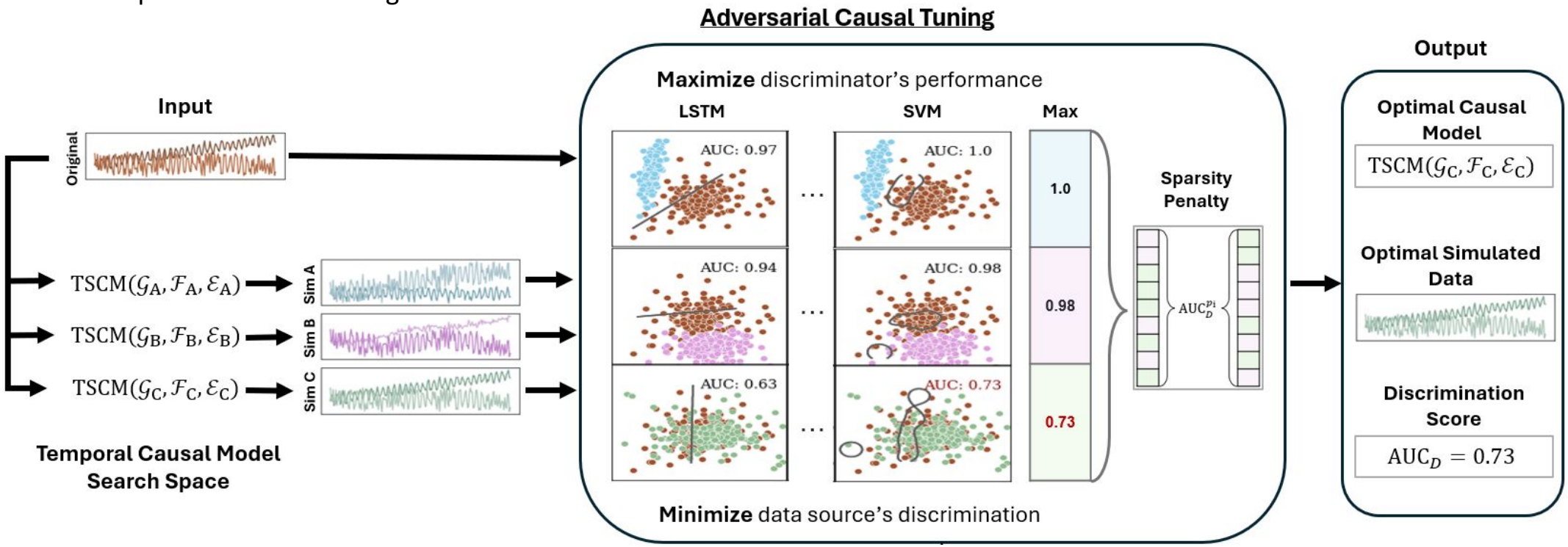
Which algorithm (and hyper-parameters) to use to discriminate between the real and the simulated data? Optimize

Adversarial Causal Tuning

*Gkorgkolis, et. al. "Temporal Causal-based Simulation for Realistic Time-series Generation" 2025 Preprint link

- New, unpublished work
- Identify as **best performing** the causal model that generates **simulated data** which **despite our best efforts** are **indistinguishable from the real training data**
- Select the **simplest** causal model whose performance is **indistinguishable** from the best (using permutation-based hypothesis testing)

- Can be coupled with AutoML to generate discriminators



TSCM and Discriminator Search Spaces

TSCM Search Space			
TCS Phase	**Method**	**Params**	**Values**
Phase 1: Causal Discovery	**CP**	model version	{LCM_1, LCM_2}
	PCMCI		{1, 2, 3}
			{10, 50}
	DYNOTEARS		{1, 2, 3}
Phase 2: Functional Dependencies Estimation	**Random Forests**		{100, 500}
	TCDF	kernel size	{2, 3}
			{2, 3}
	Gradient Boosting		{100, 500}
	TimesFM	model version	timesfm-1.0-200m
Phase 3: Noise Distribution Estimation	**Uniform**		empirically found
			empirically found
	Gaussian		empirically found
			empirically found
	Spline		
	RealNVP		

Discriminator Search space

Method	Params	Values
SVM	C	{1.0, 0.75, 0.5, 0.25}
	Kernel	{linear, poly, rbf}
	Degree	{3}
	Gamma	{auto, scale}
LSTM	Batch Size	{32, 64}
	Hidden	{128, 256}
	Layers	{2, 3}
	Dropout	{0.05, 0.1}
	Seq. Len.	{10, 20}
	Epochs	{10, 50}
	Lr	{1e-4, 1e-3}

In total:

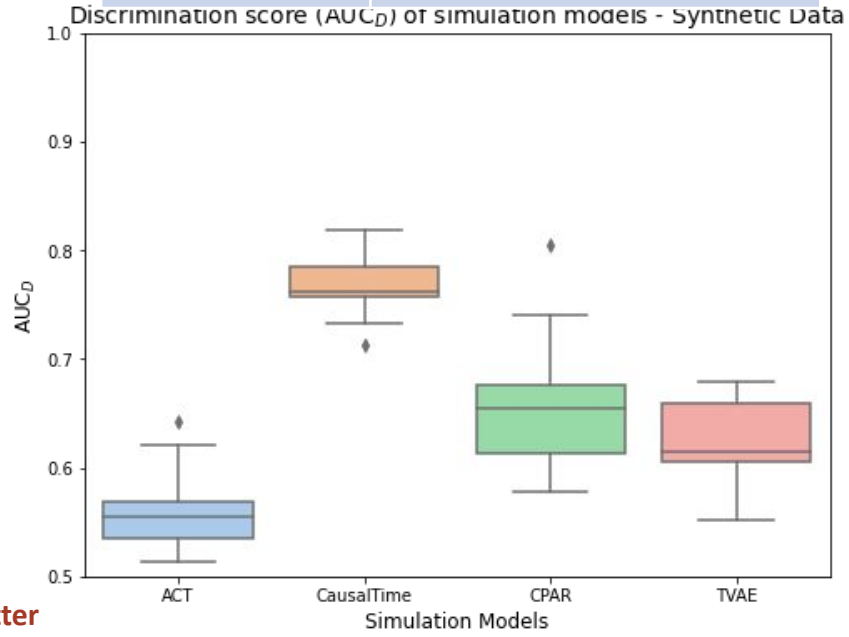
- ~700 causal configurations used during the **TSCM Generative Module**
- ~150 discriminators applied during the **Adversarial Module**

How Good is ACT at Simulating Data on Synthetic Distributions?

Experimental setup

- Compute AUC_D of SOTA simulation methods on **synthetic** datasets
  - **20** synthetic datasets
    - generated randomly
    - non-linear functional dependencies and
    - Gaussian or uniform noise.
  - **3** SOTA simulation baselines considered:
    - **CausalTime** (does not learn lagged causal model),
    - **CPAR** (non-causal),
    - **TVAE** (non-causal)

TSCM Random Synthetic Data Generation : 20 Datasets	
# Variables	
# Lag	
Edge Probability	
Functional Dependencies	
Noise Distributions	



Closer to 0.5 is better

ACT successfully simulates synthetic data, surpassing the existing SOTAs.

How Good is ACT at Simulating Data from Real Distributions?

Experimental setup

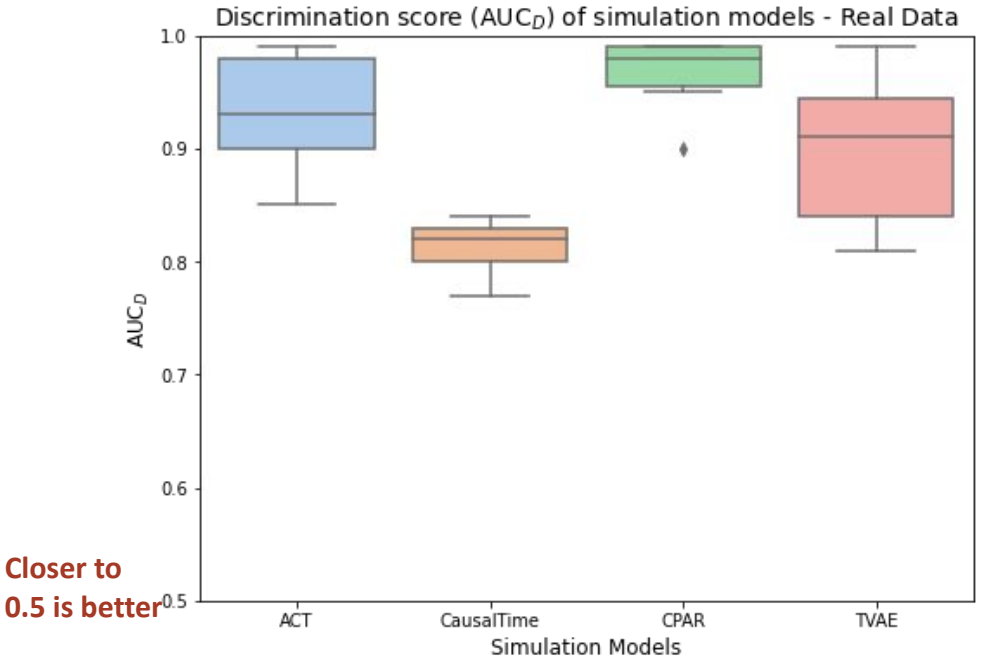
- Compute AUC_D of SOTA simulation methods on **real** datasets (closer to 0.5 means indistinguishable distributions)
  - **7** real datasets
    - Bike usage (**Yannik et al** [1]), Air quality (**Cheng et al** [2])
    - WTH, ETTh1, ETTm1 (**MvTS** [3])
  - **3** SOTA simulation baselines considered:
    - **CausalTime** (does not learn lagged causal model),
    - **CPAR** (non-causal),
    - **TVAE** (non-causal)

***ACT** is trying to fit the distributions with a causal model, which is a more difficult task **ACT** informs the user if the fitting is not successful, unlike baselines*

All methods struggle with real data, a previously unknown fact.

Why?

- Causal process not stationary
- Noise is not additive
- Poor fitting of noise distribution and/or functional dependencies
- Causal discovery algorithms are not accurate



[1] : Hahn Yannik and Langer, Tristan et. al. (2023). Time Series Dataset Survey for Forecasting with Deep Learning,
[2]: Yuxiao Chang et al, CausalTime: Realistically Generated Time-series for Benchmarking of Causal Discovery,
[3] : Junchen Ye et al, MvTS-library: An open library for deep multivariate time series forecasting,

Challenge 2: High Quality Causal Discovery

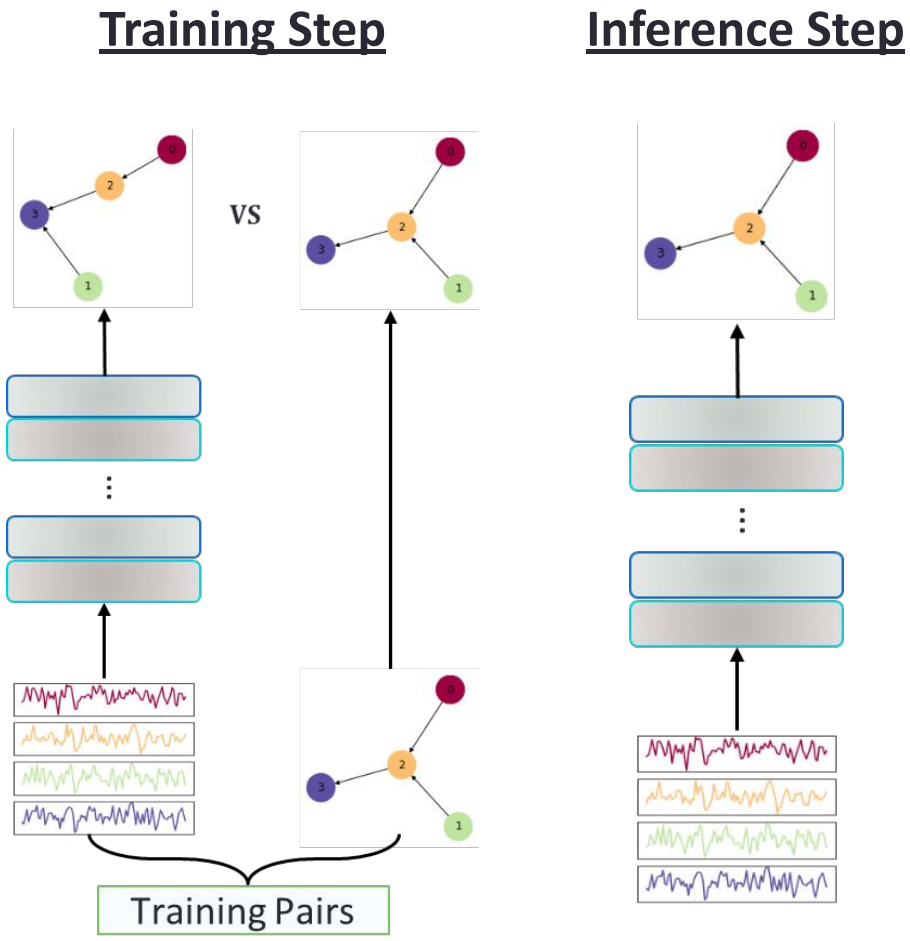
- **Standard approach**: a **human** designs algorithms based on our **expectations** about real data [1]

- **Idea**: Train a **Large Causal Model**; **learn** the Causal Discovery algorithm

↑ Improves quality

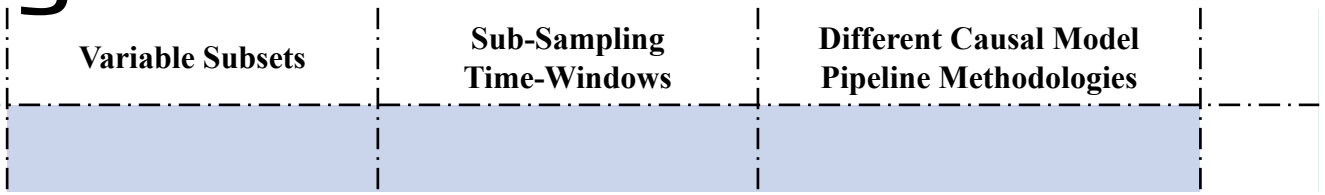
🕑 Constant-time inference step

↓ Cannot handle large number of variables/lags



[1] Stein, G., Shadaydeh, M. and Denzler, J., 2024. “Embracing the black box: Heading towards foundation models for causal discovery from time series data” AAAI'24 Workshop (AI4TS)

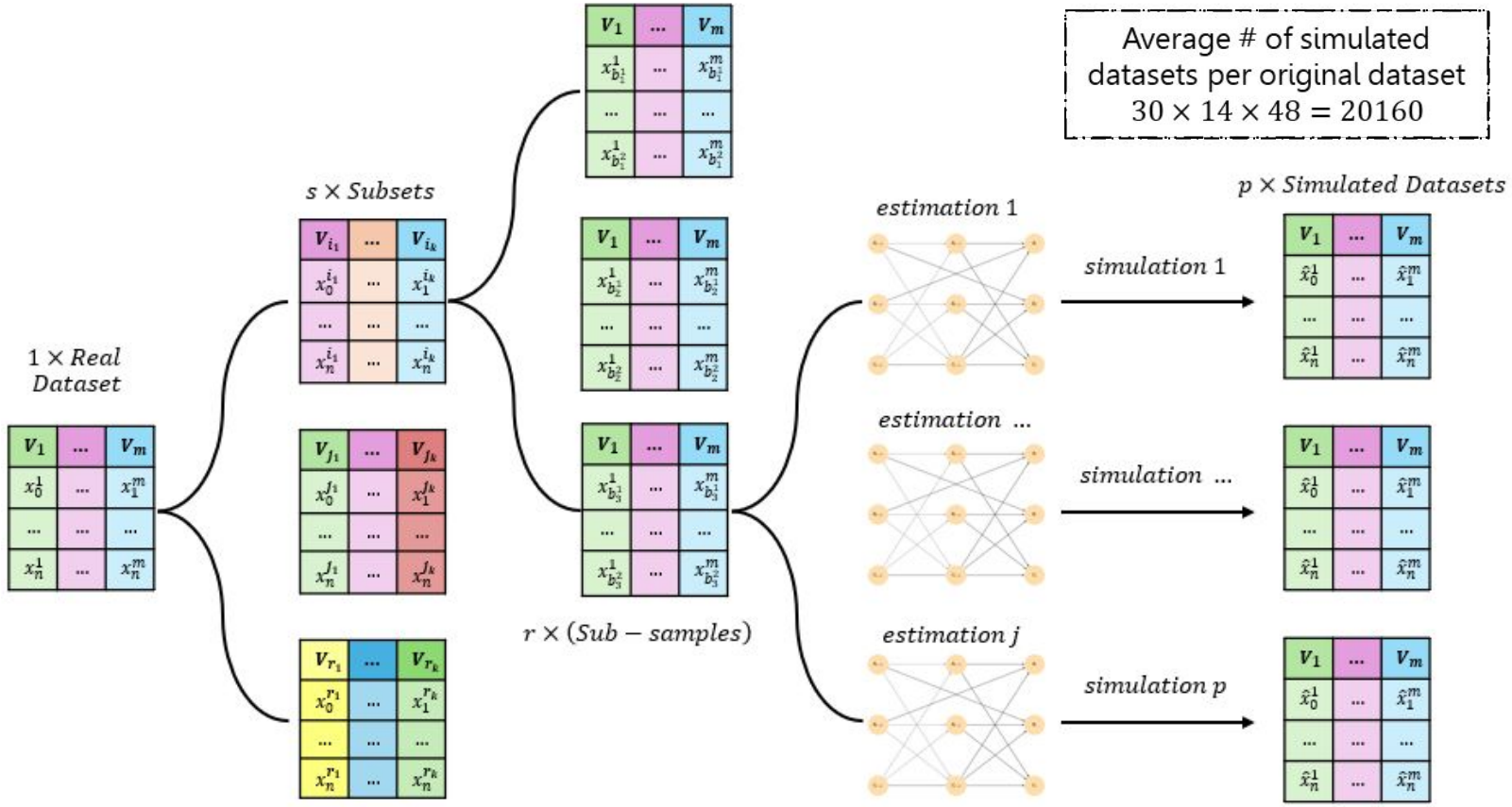
Realistic Training: Simulated Data for Training LCMs



Publicly available datasets are **limited**; we need hundreds of thousands of training instances

Idea: causal models fit on real data should be more realistic than synthetic causal models

Given a single original dataset, produce ~20000 pairs of **realistic** causal models and corresponding datasets.



Realistic Training Data Improve Learning Quality

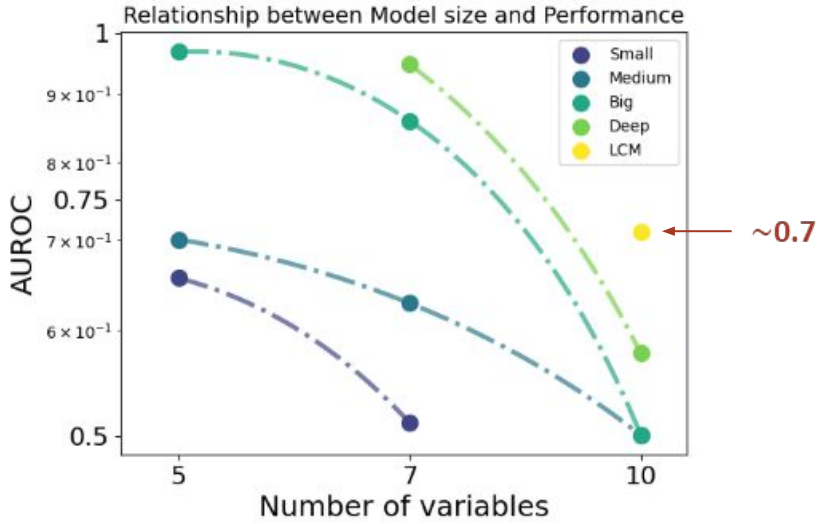
Model LCM_{synth}:
- Large-sized model $(\sim 350M)$
- Trained on strictly synthetic data of $220k$ training instances.

Model $\mathrm{LCM}_{synth+realistic}$:
- Large-sized model $(\sim 350M)$
- Trained on mixture of **80-20** synthetic & realistic data of $290k$ training instances.

Baselines: PCMCI+ [1], DYNOTEARS [2]

LCM Baseline: not publicly available LCM in [3] (top-right figure), tested on in-distribution data, smaller scope of tasks

Test set:
- (synthetic) 300 out-of-sample synthetic dataset from unseen (during training) mechanisms *
- (semi-synthetic) 26 semi-synthetic datasets on fMRI data



Conclusion: enriched training dataset improves performance on synthetic and semi-synthetic tasks ✔

Results		
Model \ Test Corpus	Semi-synthetic	Synthetic
	0.73	0.50
	0.70	0.60
	0.97	0.69
	0.98	**0.83**

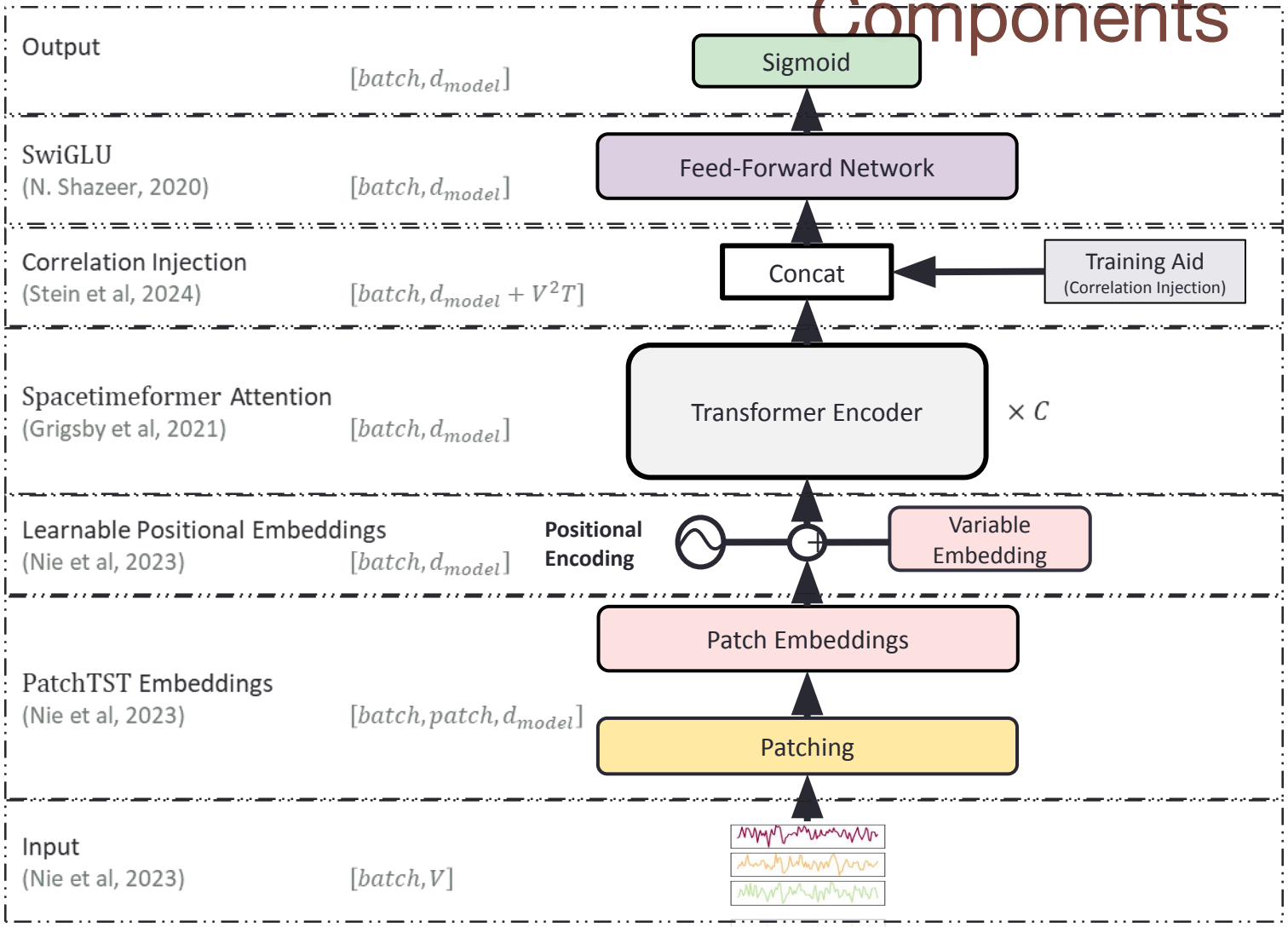
[1]: Runge, Jakob. "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets." Conference on uncertainty in artificial intelligence. Pmlr, 2020.
[2]: Pamfil, Roxana, et al. "Dynotears: Structure learning from time-series data." International Conference on Artificial Intelligence and Statistics. Pmlr, 2020.
[3]: Stein, G., Shadaydeh, M. and Denzler, J., 2024. "Embracing the black box: Heading towards foundation models for causal discovery from time series data" AAAI'24 Workshop (AI4TS)

[*]: Bounded Neural Networks

Improved Architecture: Ablation Study of LCM Components

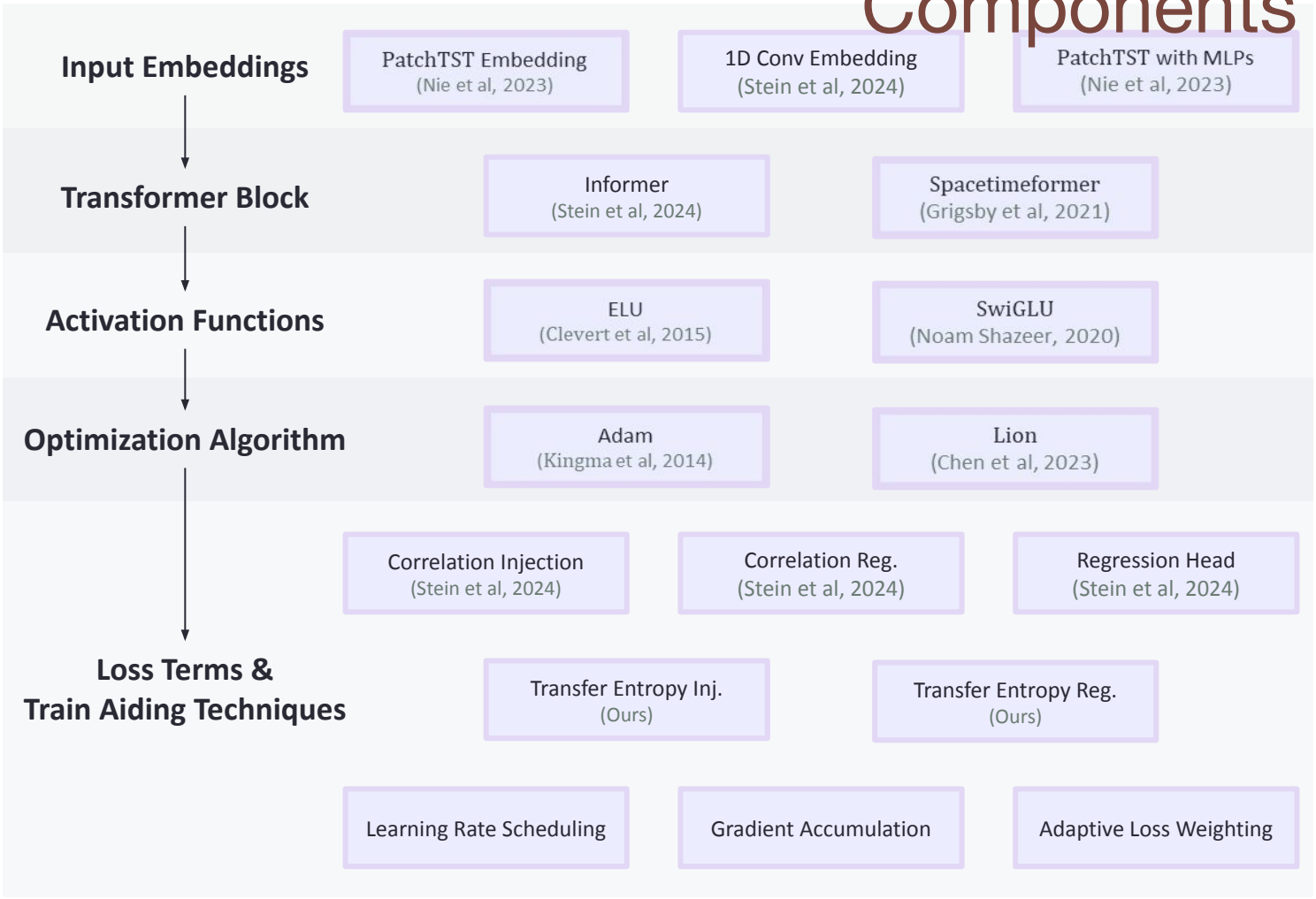
LCM Proposed Improved Architecture

Notation:
- T: # lags
- V: # variables
- $batch$: batch size
- $patch$: patching size
- C: # encoder blocks
- d_{model}: transformer internal dimension



Improved Architecture: Ablation Study of LCM Components

- Optimize over several architectural choices for medium scale models $(\sim 10M)$ for computational efficiency
- Ablation studies with ~60 of LCM pipelines
- Selected the pink modifications to the architecture



Improved Large Causal Model (LCM)

Improvements

- Training with realistic causal models produced by running ACT on real datasets
- Architectural improvements through ablation studies.
- In preparation for submission

Conclusion: Our LCM for timeseries data outperforms state of the art classical and prior LCM causal discovery algorithms (publication under preparation)

Results (*semi-synthetic* test set)		
Model	Semi-synthetic	Synthetic
	0.73	0.50
	0.70	0.60
	0.98	**0.79**

$\boldsymbol{AUC_C}$: AUC of discovering the true causal edges

Model $\text{LCM}_{synth+realistic+arch}$:

- Large-sized model $(\sim 10M)$
- Trained on mixture of **80-20** synthetic & realistic data of $290k$ training instances.

Baselines: PCMCI+ [1], DYNOTEARS [2]

Test set:

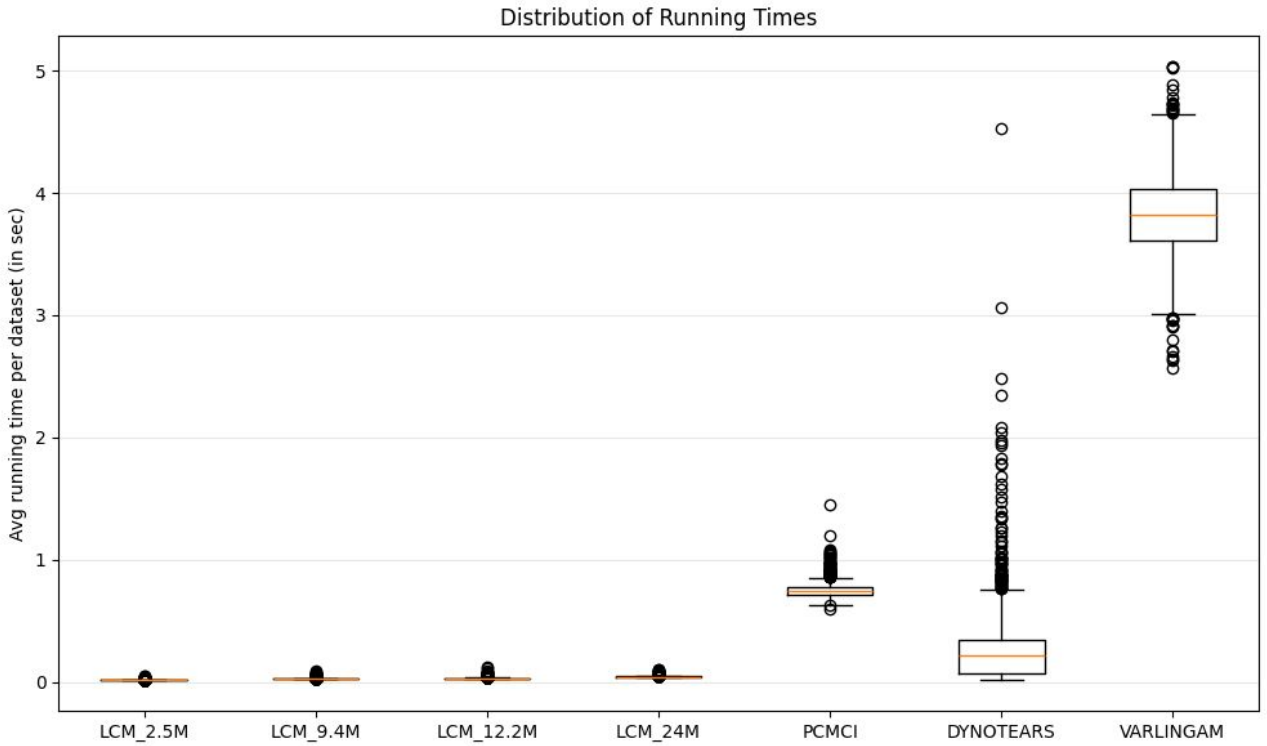
- (synthetic) 300 out-of-sample synthetic dataset from unseen (during training) mechanisms
- (semi-synthetic) 26 semi-synthetic datasets on fMRI data

[1]: Runge, Jakob. "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets." Conference on uncertainty in artificial intelligence. Pmlr, 2020.
[2]: Pamfil, Roxana, et al. "Dynotears: Structure learning from time-series data." International Conference on Artificial Intelligence and Statistics. Pmlr, 2020.
[3]: Stein, G., Shadaydeh, M. and Denzler, J., 2024. “Embracing the black box: Heading towards foundation models for causal discovery from time series data” AAAI'24 Workshop (AI4TS)

Execution time of LCM vs. Standard Methods

- **Baselines**:
  PCMCI[1], DYNOTEARS[2], VARLINGAM[3]

Models		
	0.014±.001	(0.012, 0.047)
	0.027±.002	(0.023, 0.086)
	0.03±.003	(0.026, 0.124)
	0.041±.004	(0.036, 0.1)
PCMCI	0.749±.056	(0.596, 1.453)
DYNOTEARS	0.264±.274	(0.02, 4.532)
VARLINGAM	3.825±.327	(2.567, 5.032)



Running times comparison in synthetic in-distribution holdout test set.

Conclusion: Deep LCMs are from **6 to over 90 times faster** compared the state-of-the-art no-DL algorithms

Paper "Large Causal Models for Temporal Causal Discovery" to appear in ECML / PKKD 2026

[1] Runge, Jakob. "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets." Conference on uncertainty in artificial intelligence. Pmlr, 2020.
[2] Pamfil, Roxana, et al. "Dynotears: Structure learning from time-series data." International Conference on Artificial Intelligence and Statistics. Pmlr, 2020.
[3] Hyvärinen, Aapo, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. "Estimation of a structural vector autoregression model using non-Gaussianity." Journal of Machine Learning Research 11, no. 5 (2010).

ETIA Use-Cases and Pilots

- **Looking for collaborations** of interesting use-cases to apply Automated Causal Discovery
  - Medical, real-world data
  - Energy
  - Environmental
  - Telecommunications
  - Bioinformatics

- **Looking to hire** for all positions for the EIC Transition grant
  - Developers, front-end, back-end, full stack, ML researchers, data scientists, AI Agent builders, etc.
  - Content creators, marketeers, project managers, business development, sales

Summary and Conclusions

- **Multiple** feature selection that is general (all data types), scalable, good quality is necessary for knowledge discovery. It is a mature technology.
- Automated Machine Learning for biomedical data increases productivity, reduces errors, leads to clinically validated results.
- Causal discovery can find causal models suitable for optimization of decisions
- Automated Causal Discovery and Causal AI Agent is an emerging technology

Thank you!

- Contact tsamard.it@gmail.com

- Mens Ex Machina academic group (MXM) www.mensxmachina.org
- JADBio - Gnosis DA SA www.jadbio.com

- Looking for interesting applications and collaborations for new exciting causal grant.

APPENDIX

Novel Scientific Results Obtained with JADBio

- Makrina Karaglani, Krystallia Gourlia, Ioannis Tsamardinos, Ekaterini Chatzaki, “Accurate blood-based diagnostic biosignatures for Alzheimer’s disease via Automated machine learning”, ***Journal of Clinical Medicine***, 2020, 9(9), https://doi.org/10.3390/jcm9093016
- Karstoft KI, Tsamardinos I, Eskelund K, Andersen SB, Nissen LR, “Applicability of an Automated Model and Parameter Selection in the Prediction of Screening-Level PTSD in Danish Soldiers Following Deployment: Development Study of Transferable Predictive Models Using Automated Machine Learning”, ***JMIR Med Inform*** 2020;8(7):e17119
- Maria Loos, Reshmi Ramakrishnan, Wim Vranken, Alexandra Tsirigotaki, Evrydiki-Pandora Tsare, Valentina Zorzini, Jozefien De Geyter, Biao Yuan, Ioannis Tsamardinos, Maria I Klapa, Joost Schymkowitz, Frederic Rousseau, Spyridoula Karamanou, Tassos Economou, “Structural basis of the sub-cellular topology landscape of Escherichia coli”, ***Frontiers in Microbiology*** 10, 2019, pp. 675, doi: 10.3389/fmicb.2019.01670
- Alberto Montesanto, Patrizia D’Aquila, Vincenzo Lagani, Ersilia Paparazzo, Silvana Geracitano, Laura Formentini, Robertina Giacconi, Maurizio Cardelli, Mauro Provinciali, Dina Bellizzi, Giuseppe Passarino, “A New Robust Epigenetic Model for Forensic Age Prediction”, Journal of Forensic Sciences, 65(5) 2020, **https://doi.org/10.1111/1556-4029.14460**

Novel Scientific Results Obtained with JADBio

- Dimitrios Kyriakis, Alexandros Kanterakis, Tereza Manousaki, Alexandros Tsakogiannis, Michalis Tsagris, Ioannis Tsamardinos, Leonidas Papaharisis, DImitris Chatziplis, George Potamias, Costas S Tsigenopoulos, “Scanning of genetic variants and genetic mapping of phenotypic traits in gilthead seabream through ddRAD sequencing”, ***Frontiers in Genetics***, vol. 10, p. 675, 2019
- Maria Panagopoulou, Makrina Karaglani, Ioanna Balgkouranidou, Eirini Biziota, Triantafillia Koukaki, Evaggelos Karamitrousis, Evangelia Nena, Ioannis Tsamardinos, George Kolios, Evi Lianidou, Stylianos Kakolyris, Ekaterini Chatzaki, "Circulating cell free DNA in Breast cancer: size profiling, levels and methylation patterns lead to prognostic and predictive classifiers", ***Oncogene*** 2019, 38(18):3387-3401. doi: 10.1038/s41388-018-0660-y.
- Simantiraki, O., Charonyktakis, P., Pampouchidou, A., Tsiknakis, M., and Cooke, M. “Glottal source features for automatic speech-based depression assessment.”, in Proceedings of the 18th Conference of the International Speech Communication Association **INTERSPEECH** 2700–2704, 2017

Novel Scientific Results Obtained with JADBio

- Marios Adamou, Grigoris Antoniou, Elissavet Greasidou, Vincenzo Lagani, Paulos Charonyktakis, Ioannis Tsamardinos, Michael Doyle, “Towards Automatic Risk Assessment to Support Suicide Prevention”, **Crisis** 2018,
- Marios Adamou, Grigoris Antoniou, Elissavet Greassidou, Vincenzo Lagani, Paulos Charonyktakis, Ioannis Tsamardinos, “Mining Free-Text Medical Notes for Suicide Risk Assessment”, **SETN** 2018
- George Froudakis, George Borboudakis, Taxiarchis Stergiannakos, Maria Frysali, Emanuel Klontzas, and Ioannis Tsamardinos, “Chemically-intuited, large-scale screening of MOFs by machine learning techniques”, **NPJ Computational Materials**, (2017) 3:40, doi:10.1038/s41524-017-0045-8

Novel Scientific Results Obtained with JADBio

- Christina Chatzipantsiou, Vincenzo Lagani, Maria Markaki, OT Duc Nguyen, I Tsamardinos, HF Kvitvang, A Nordborg, R Mjelle, OD Røe, "Metabolomics signature in serum months to years before thoracic cancer: A HUNT study", **ECCB 2018** 17th European Conference on Computational Biology, September 8-12 2018, Athens Greece
- V. Danilatou y, D. Antonakaki, C. Tzagkarakis, A. Kanterakis, V. Katos , and T. Kostoulas, "Automated Mortality Prediction in Critically-ill Patients with Thrombosis using Machine Learning", **BIBE**, 2020
- Georgia Orfanoudaki, Maria Markaki, Katerina Chatzi, Ioannis Tsamardinos, and Anastassios Economou, "MatureP: prediction of secreted proteins with exclusive information from their mature regions", **Scientific Reports** 7, 2017, doi:10.1038/s41598-017-03557-4

Selected Algorithmic and Methodological Publications

- Iordanis Xanthopoulos, Ioannis Tsamardinos, Vassilis Christophides, Eric Simon, Alejandro Salinger, “Putting the Human Back in the AutoML Loop”, **EDBT/ICDT** Workshops 2020
- Ioannis Tsamardinos, Giorgos Borboudakis, Pavlos Katsogridakis, Polyvios Pratikakis, Vassilis Christophides, “A Greedy Feature Selection Algorithm for Big Data of high dimensionality”, **Machine Learning** 108(2), pp 149-202, 2018, https://doi.org/10.1007/s10994-018-5714-4
- Ioannis Tsamardinos, Elissavet Greassidou, Giorgos Borboudakis, “Bootstrapping the Out-of-sample Predictions for Efficient and Accurate Cross-Validation”, **Machine Learning Journal** (2018) 107(12), pp 1895–1922, https://doi.org/10.1007/s10994-018-5714-4
- Vincenzo Lagani, Giorgos Athineou, Alessio Farcomeni, Michail Tsagris, Ioannis Tsamardinos (2016): “Feature Selection with the R Package MXM: Discovering Multiple, Statistically-Equivalent, Predictive Feature Subsets”, **Journal of Statistical Software** v80(7), 2017 doi: 10.18637/jss.v080.i07
- I. Tsamardinos, C. Aliferis, and A. Statnikov, "Time and sample efficient discovery of Markov Boundarys and direct causal relations," *The Ninth ACM **SIGKDD International Conference on Knowledge Discovery and Data Mining** (KDD 2003)*, 2003. http://dl.acm.org/citation.cfm?id=956838
- G. Borboudakis and I. Tsamardinos, "Forward-Backward Selection with Early Dropping," ***Journal of Machine Learning Research***, vol. 20, iss. 8, pp. 1-39, 2019. http://jmlr.org/papers/volume20/17-334/17-334.pdf
- M. Tsagris, Z. Papadovasilakis, K. Lakiotaki, and I. Tsamardinos, “The γ-OMP algorithm for feature selection with application to gene expression data**,” IEEE/ACM Transactions on Computational Biology and Bioinformatics** , 2020. doi:10.1109/TCBB.2020.3029952
- Ioannis Tsamardinos, Giorgos Borboudakis, Pavlos Katsogridakis, Polyvios Pratikakis, Vassilis Christophides, “Massively-Parallel Feature Selection for Big Data”, https://arxiv.org/abs/1708.07178
- V. Lagani, G. Athineou, A. Farcomeni, M. Tsagris, and I. Tsamardinos, "Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets," **Journal of Statistical Software**, vol. 80, iss. 7, 2017. doi:10.18637/jss.v080.i07
- Yannis Pantazis, Vincenzo Lagani, Paulos Charonyktakis, Ioannis Tsamardinos, “Multiple Equivalent Solutions for the Lasso”, https://arxiv.org/abs/1710.04995]