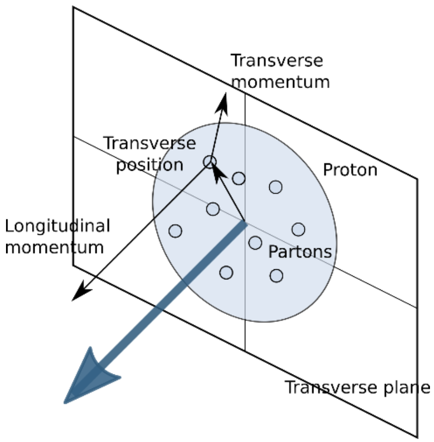
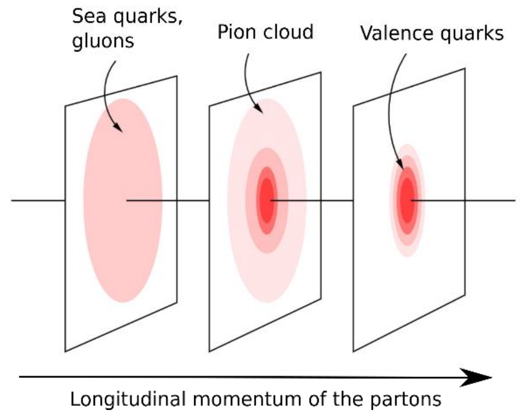
MACHINE LEARNING FOR CLAS12 DATA ANALYSIS WITH GENERALIZED ADDITIVE MODELS

IN2P3/IRFU workshop | Noëlie Cherrier

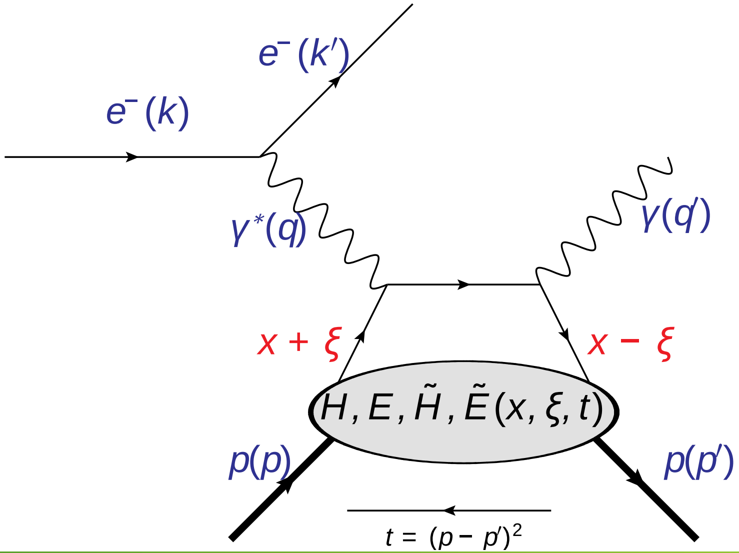
INTRODUCTION

- Physics objective: tomography of the nucleon through **Generalized Parton Distributions** (GPDs)
  - ⟶ Correlation between longitudinal momentum and transverse position of the partons in the nucleon





- Accessed through exclusive inelastic processes including **Deeply Virtual Compton Scattering** (DVCS)

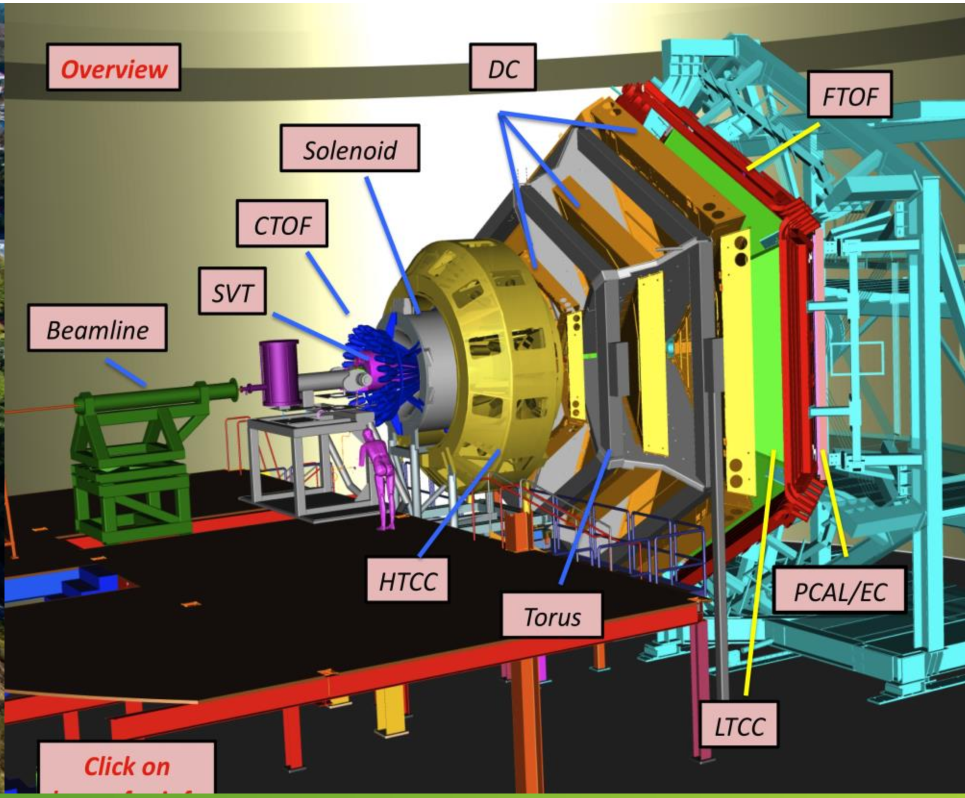


INTRODUCTION

- Jefferson Lab: 10.6 GeV electron beam
- CLAS12 data taking since 2018: hydrogen target

Event classification task: isolate DVCS events ($ep \rightarrow ep\gamma$)

Machine learning approach to be compared to classical approach



INTERPRETABLE / TRANSPARENT / INTELLIGIBLE MACHINE LEARNING

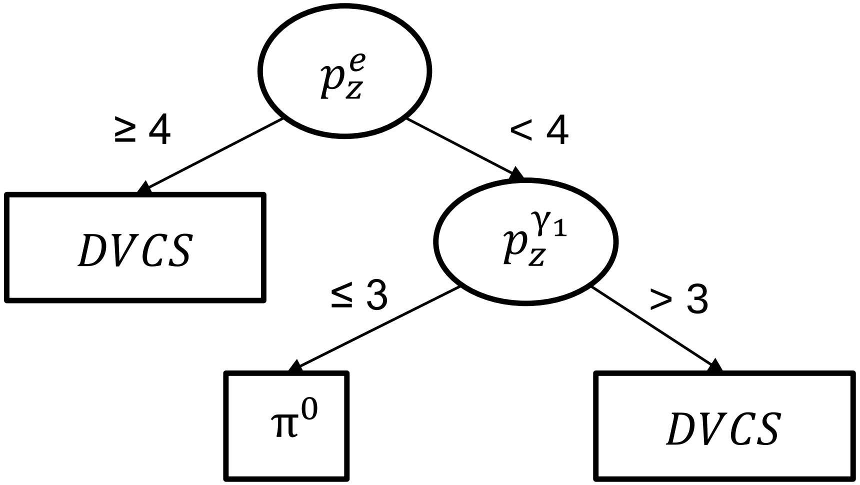
- Interpretability: it is defined as the ability to explain or to provide the meaning in understandable terms to a human
- Transparency: a model is considered to be transparent if by itself it is understandable. A model can feature different degrees of understandability
- Intelligibility (or understandability) denotes the characteristic of a model to make a human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally

⚠ The lack of interpretability is controversial

Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* (2019).

INTERPRETABLE / TRANSPARENT / INTELLIGIBLE MACHINE LEARNING

Models for which post-hoc analysis is not needed



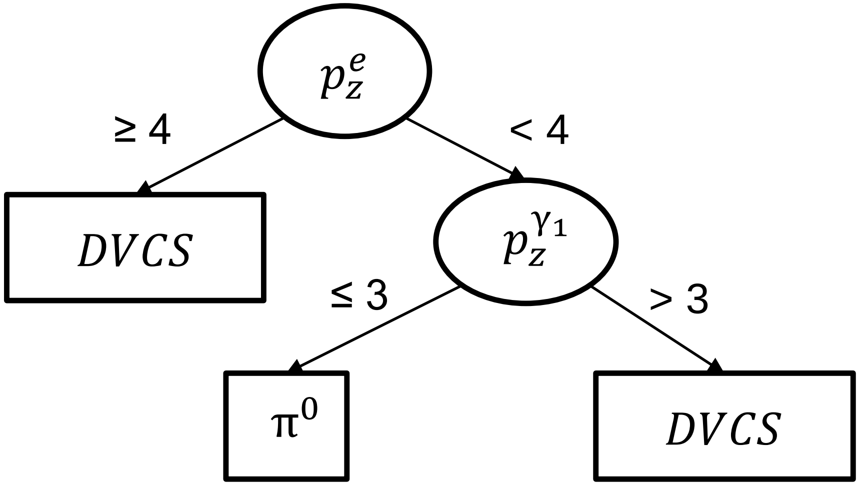
Decision trees

Rule bases

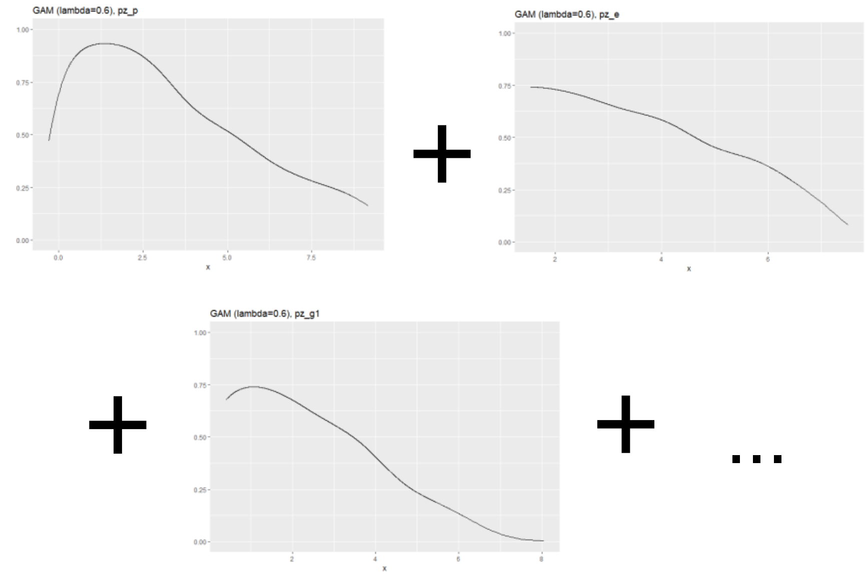
```
(inv_masss_g1g2 in [-inf, -inf, 0.665977, 0.666042]) and (inv_masss_g1g2 in [0.007705, 0.007706, inf, inf]) => Class=DVMP (CF = 0.8)
(energy_g1 in [-inf, -inf, 2.209962, 2.21012]) and (cone_angle_g1 in [-inf, -inf, 16.272992, 16.275288]) => Class=DVMP (CF = 0.76)
(energy_g1 in [-inf, -inf, 3.100969, 3.101338]) and (MM_eg1 in [0.525376, 0.525439, inf, inf]) => Class=DVMP (CF = 0.65)
(energy_g1 in [-inf, -inf, 1.735166, 2.66702]) and (MM_eg1 in [-1.85998, -1.857006, inf, inf]) => Class=DVMP (CF = 0.61)
(MM_eg1 in [1.298545, 1.304201, inf, inf]) and (energy_g1 in [-inf, -inf, 4.182, 4.182101]) => Class=DVMP (CF = 0.66)
(energy_g1 in [3.333313, 3.333823, inf, inf]) and (MM_eg1 in [-inf, -inf, 0.96117, 0.961204]) => Class=DVCS (CF = 0.82)
(energy_g1 in [3.100909, 3.101237, inf, inf]) and (MM_eg1 in [-inf, -inf, 1.084021, 1.084045]) => Class=DVCS (CF = 0.8)
(MM_eg1 in [-inf, -inf, 0.852413, 0.852521]) and (energy_g1 in [2.103109, 2.103411, inf, inf]) => Class=DVCS (CF = 0.76)
(cone_angle_g1 in [16.137178, 21.604087, inf, inf]) and (MM_epg1 in [-inf, -inf, -0.538689, -0.537701]) => Class=DVCS (CF = 0.56)
```

INTERPRETABLE / TRANSPARENT / INTELLIGIBLE MACHINE LEARNING

Models for which post-hoc analysis is not needed



Decision trees



$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + \ldots + f_m(x_m)$$

Generalized Additive Models (GAM)

Rule bases

```
(inv_masss_g1g2 in [-inf, -inf, 0.665977, 0.666042]) and (inv_masss_g1g2 in [0.007705, 0.007706, inf, inf]) => Class=DVMP (CF = 0.8)
(energy_g1 in [-inf, -inf, 2.209962, 2.21012]) and (cone_angle_g1 in [-inf, -inf, 16.272992, 16.275288]) => Class=DVMP (CF = 0.76)
(energy_g1 in [-inf, -inf, 3.100969, 3.101338]) and (MM_eg1 in [0.525376, 0.525439, inf, inf]) => Class=DVMP (CF = 0.65)
(energy_g1 in [-inf, -inf, 1.735166, 2.66702]) and (MM_eg1 in [-1.85998, -1.857006, inf, inf]) => Class=DVMP (CF = 0.61)
(MM_eg1 in [1.298545, 1.304201, inf, inf]) and (energy_g1 in [-inf, -inf, 4.182, 4.182101]) => Class=DVMP (CF = 0.66)
(energy_g1 in [3.333313, 3.333823, inf, inf]) and (MM_eg1 in [-inf, -inf, 0.96117, 0.961204]) => Class=DVCS (CF = 0.82)
(energy_g1 in [3.100909, 3.101237, inf, inf]) and (MM_eg1 in [-inf, -inf, 1.084021, 1.084045]) => Class=DVCS (CF = 0.8)
(MM_eg1 in [-inf, -inf, 0.852413, 0.852521]) and (energy_g1 in [2.103109, 2.103411, inf, inf]) => Class=DVCS (CF = 0.76)
(cone_angle_g1 in [16.137178, 21.604087, inf, inf]) and (MM_epg1 in [-inf, -inf, -0.538689, -0.537701]) => Class=DVCS (CF = 0.56)
```

GENERALIZED ADDITIVE MODELS (GAM)

Generalized Linear Models (GLM) :

$$g(\hat{y}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$g(\hat{y}) = \hat{y}$ for regression, $g(\hat{y}) = \ln(\frac{\hat{y}}{1-\hat{y}})$ for classification

Hastie, T. J. (1986). Generalized additive models. In *Statistical models in S* (pp. 249-307). Routledge.
Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013, August). Accurate intelligible models with pairwise interactions. *ACM SIGKDD 2013.*

GENERALIZED ADDITIVE MODELS (GAM)

Generalized Linear Models (GLM) :

$$g(\hat{y}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$g(\hat{y}) = \hat{y}$ for regression, $g(\hat{y}) = \ln(\frac{\hat{y}}{1-\hat{y}})$ for classification

Generalized Additive Models (GAM) :

$$g(\hat{y}) = \beta_0 + f_1(x_1) + \ldots + f_d(x_d)$$

Hastie, T. J. (1986). Generalized additive models. In *Statistical models in S* (pp. 249-307). Routledge.
Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013, August). Accurate intelligible models with pairwise interactions. *ACM SIGKDD 2013.*

GENERALIZED ADDITIVE MODELS (GAM)

Generalized Linear Models (GLM) :

$$g(\hat{y}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$g(\hat{y}) = \hat{y}$ for regression, $g(\hat{y}) = \ln(\frac{\hat{y}}{1-\hat{y}})$ for classification

Generalized Additive Models (GAM) :

$$g(\hat{y}) = \beta_0 + f_1(x_1) + \ldots + f_d(x_d)$$

Generalized Additive Models with pairwise interactions (GA2M) :

$$g(\hat{y}) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j)$$

Hastie, T. J. (1986). Generalized additive models. In *Statistical models in S* (pp. 249-307). Routledge.
Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013, August). Accurate intelligible models with pairwise interactions. *ACM SIGKDD 2013.*

GENERALIZED ADDITIVE MODELS (GAM)

Generalized Linear Models (GLM) :

$$g(\hat{y}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$g(\hat{y}) = \hat{y}$ for regression, $g(\hat{y}) = \ln(\frac{\hat{y}}{1-\hat{y}})$ for classification

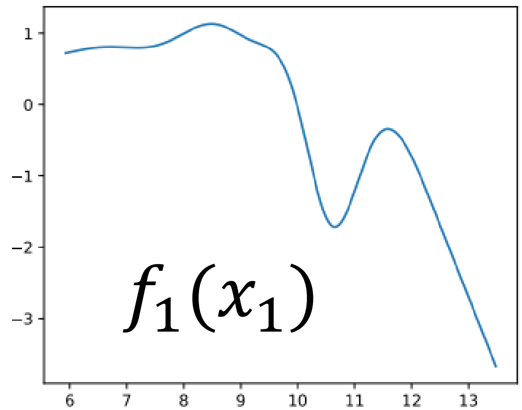
Generalized Additive Models (GAM) :

$$g(\hat{y}) = \beta_0 + f_1(x_1) + \ldots + f_d(x_d)$$

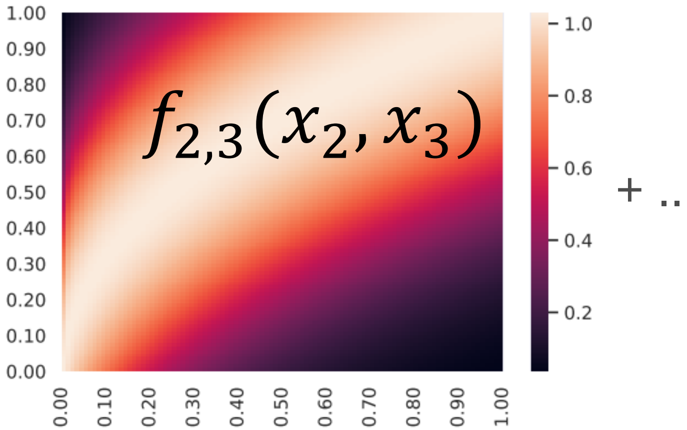
Generalized Additive Models with pairwise interactions (GA2M) :

$$g(\hat{y}) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j)$$

$g(\hat{y})$ = -8,42 +



+



+ ...

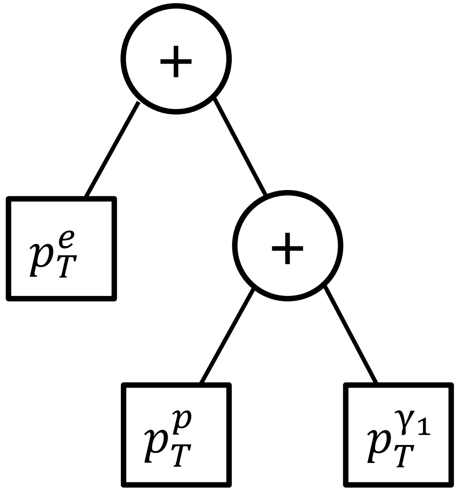
Hastie, T. J. (1986). Generalized additive models. In *Statistical models in S* (pp. 249-307). Routledge.
Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013, August). Accurate intelligible models with pairwise interactions. *ACM SIGKDD 2013.*

HOW TO USE PHYSICS KNOWLEDGE?

1. Feature construction

→ Motivation: these models do not build a sufficiently complex internal representation of the data

Constrained Genetic Programming: evolve a population of high-level feature candidates



Feature candidate example
→ Nodes are mathematical operators
→ Leaves are base variables

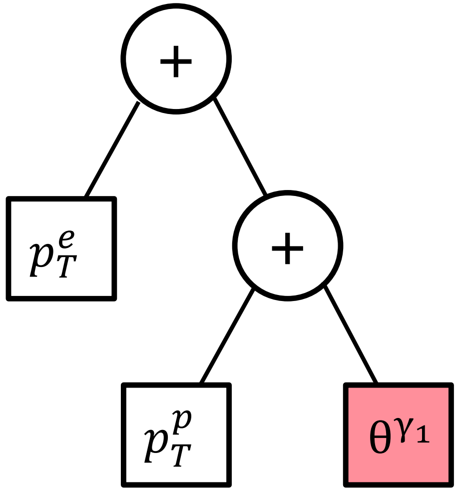
Cherrier, N., Poli, J. P., Defurne, M., & Sabatié, F. (2019, June). Consistent Feature Construction with Constrained Genetic Programming for Experimental Physics. In *2019 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1650-1658). IEEE.

HOW TO USE PHYSICS KNOWLEDGE?

1. Feature construction

→ Motivation: these models do not build a sufficiently complex internal representation of the data

Constrained Genetic Programming: evolve a population of high-level feature candidates



Feature candidate example
→ Nodes are mathematical operators
→ Leaves are base variables

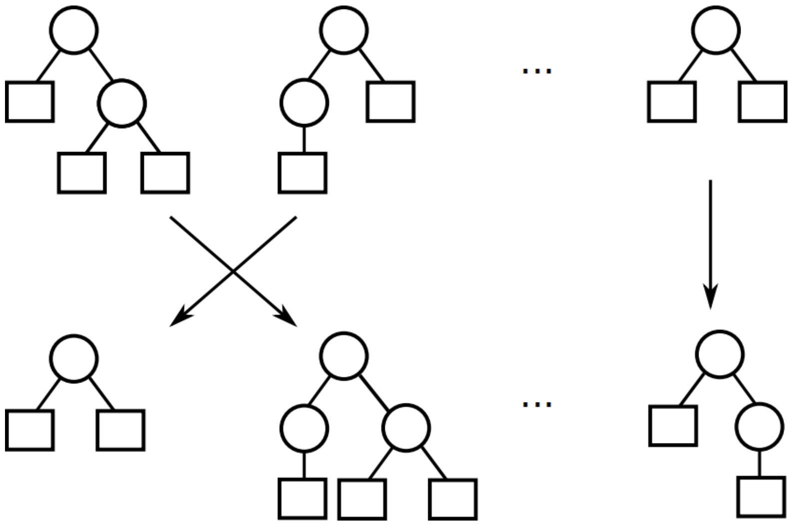
Cherrier, N., Poli, J. P., Defurne, M., & Sabatié, F. (2019, June). Consistent Feature Construction with Constrained Genetic Programming for Experimental Physics. In *2019 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1650-1658). IEEE.

HOW TO USE PHYSICS KNOWLEDGE?

1. Feature construction

   ⟶ Motivation: these models do not build a sufficiently complex internal representation of the data

Constrained Genetic Programming: evolve a population of high-level feature candidates

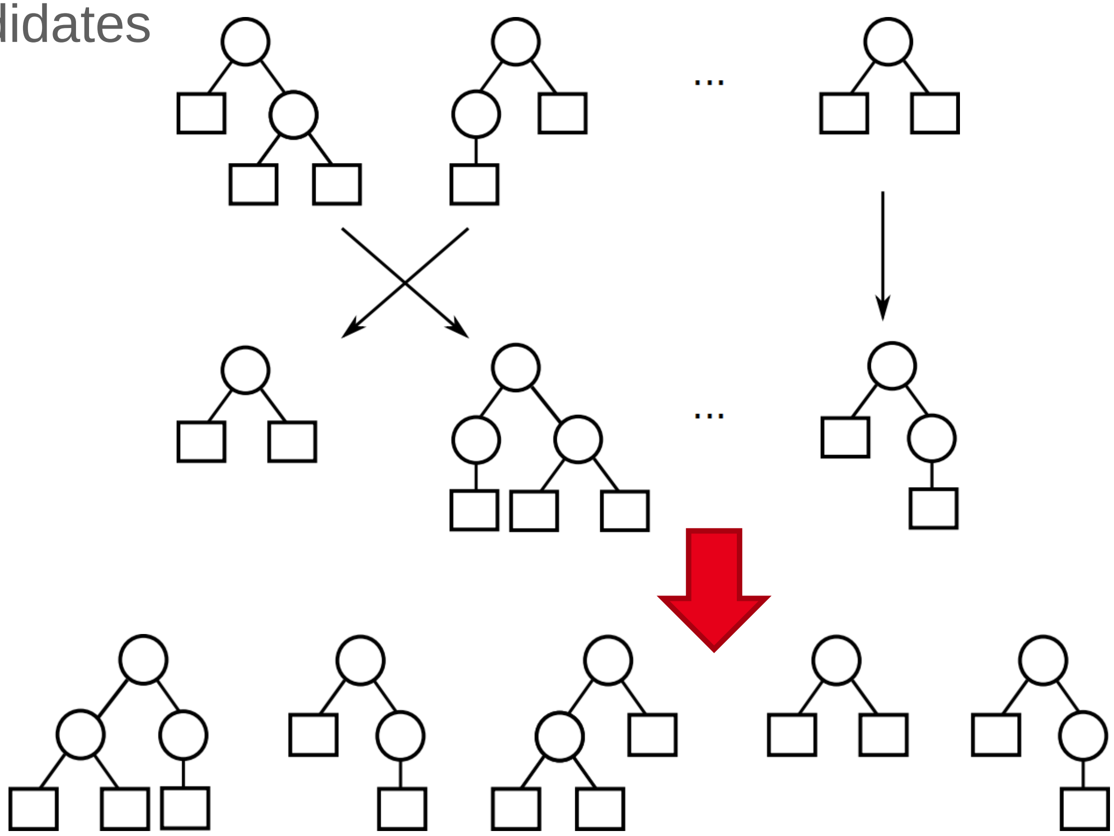


HOW TO USE PHYSICS KNOWLEDGE?

1. Feature construction

⟶ Motivation: these models do not build a sufficiently complex internal representation of the data

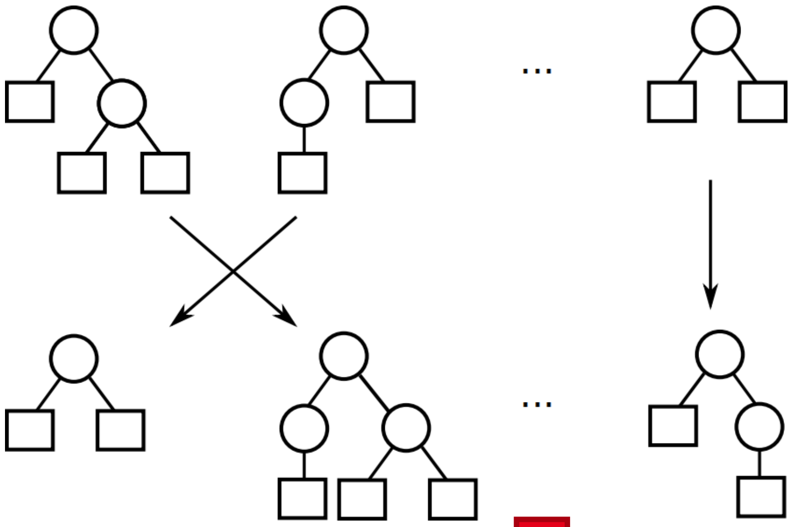
Constrained Genetic Programming: evolve a population of high-level feature candidates



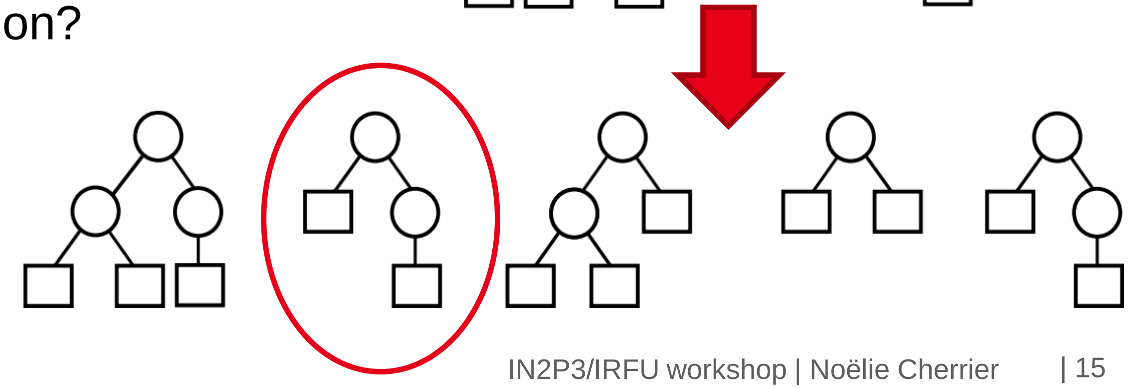
HOW TO USE PHYSICS KNOWLEDGE?

1. Feature construction
   → Motivation: these models do not build a sufficiently complex internal representation of the data

Constrained Genetic Programming: evolve a population of high-level feature candidates



Evaluation function?



FEATURE CONSTRUCTION IN GA2M

Idea: build one feature at a time, associated with one term of the GAM

$\longrightarrow$ **gradient boosting**

FEATURE CONSTRUCTION IN GA2M

Idea: build one feature at a time, associated with one term of the GAM

$\longrightarrow$ **gradient boosting**

Objective function: minimize the cross entropy $-y\ln(\hat{y}) - (1-y)\ln(1-\hat{y})$

FEATURE CONSTRUCTION IN GA2M

Idea: build one feature at a time, associated with one term of the GAM

$\longrightarrow$ **gradient boosting**

Objective function: minimize the cross entropy $-y\ln(\hat{y}) - (1-y)\ln(1-\hat{y})$

1) Compute $\beta_0 = \ln\left(\frac{p_0}{1-p_0}\right)$ to form the 1st model $g(\hat{y}) = \beta_0$.
The residual is $r = y - \hat{y} = y - p_0$ (p_0 proportion of the majority class)

FEATURE CONSTRUCTION IN GA2M

Idea: build one feature at a time, associated with one term of the GAM

$\longrightarrow$ **gradient boosting**

Objective function: minimize the cross entropy $-y\ln(\hat{y}) - (1-y)\ln(1-\hat{y})$

1) Compute $\beta_0 = \ln\left(\frac{p_0}{1-p_0}\right)$ to form the 1st model $g(\hat{y}) = \beta_0$.
The residual is $r = y - \hat{y} = y - p_0$ (p_0 proportion of the majority class)

2) Build one feature x_1 or a pair of features (x_1, x_2) discriminative wrt the residual (see next slide)

FEATURE CONSTRUCTION IN GA2M

Idea: build one feature at a time, associated with one term of the GAM

$\longrightarrow$ **gradient boosting**

Objective function: minimize the cross entropy $-y\ln(\hat{y}) - (1-y)\ln(1-\hat{y})$

1) Compute $\beta_0 = \ln\left(\frac{p_0}{1-p_0}\right)$ to form the 1st model $g(\hat{y}) = \beta_0$.
   The residual is $r = y - \hat{y} = y - p_0$ (p_0 proportion of the majority class)

2) Build one feature x_1 or a pair of features (x_1, x_2) discriminative wrt the residual (see next slide)

3) Fit a shape function $f_1(x_1)$ (or $f_{1,2}(x_1, x_2)$) to the residual

FEATURE CONSTRUCTION IN GA2M

Idea: build one feature at a time, associated with one term of the GAM

$\longrightarrow$ **gradient boosting**

Objective function: minimize the cross entropy $-y\ln(\hat{y}) - (1-y)\ln(1-\hat{y})$

1) Compute $\beta_0 = \ln\left(\frac{p_0}{1-p_0}\right)$ to form the 1st model $g(\hat{y}) = \beta_0$.
   The residual is $r = y - \hat{y} = y - p_0$ (p_0 proportion of the majority class)

2) Build one feature x_1 or a pair of features (x_1, x_2) discriminative wrt the residual (see next slide)

3) Fit a shape function $f_1(x_1)$ (or $f_{1,2}(x_1, x_2)$) to the residual

4) Compute the new model: $g(\hat{y}) = g(\hat{y}) + f_1(x_1)$ (or $g(\hat{y}) + f_{1,2}(x_1, x_2)$) and the new residual $r = y - \hat{y}$, and go back to step 2

FEATURE CONSTRUCTION IN GA2M

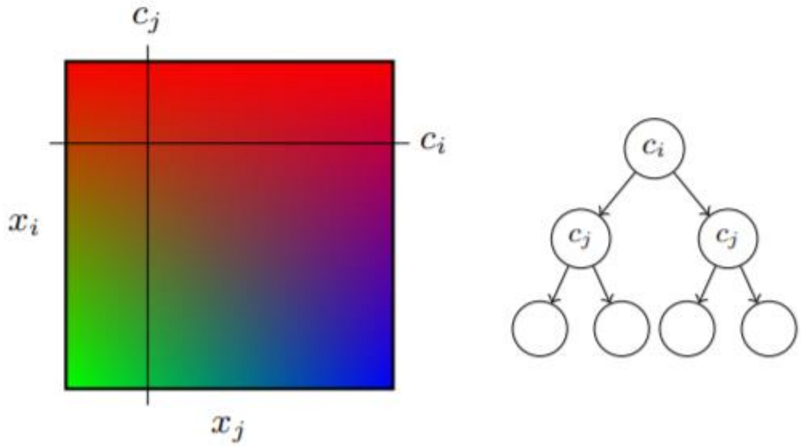
Fitness function for the Genetic Programming algorithm:

Single feature case

Shallow tree (maximum 4 leaves)
Feature fitness: RMS error of the inducted tree with the residual $y - \hat{y}$

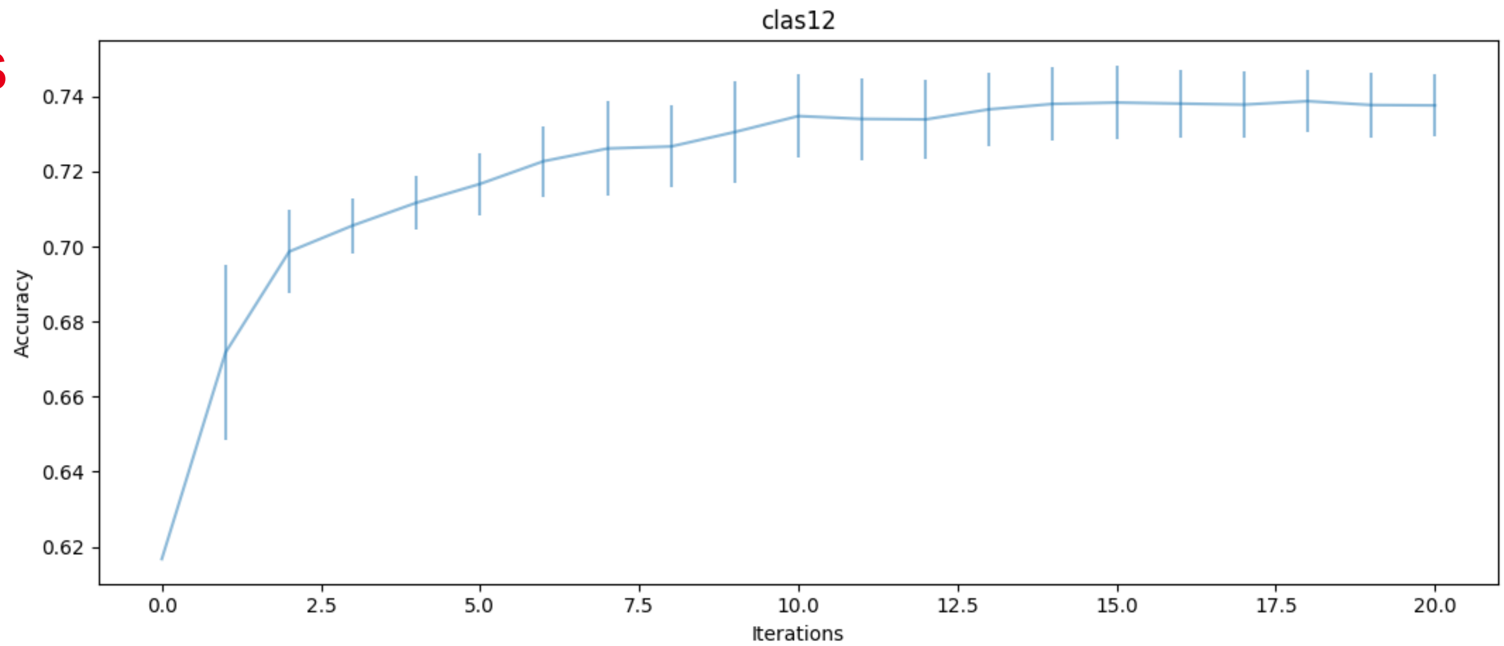
Feature pair case

FAST algorithm, the target being the residual $y - \hat{y}$



Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013, August). Accurate intelligible models with pairwise interactions. *ACM SIGKDD 2013.*

RESULTS

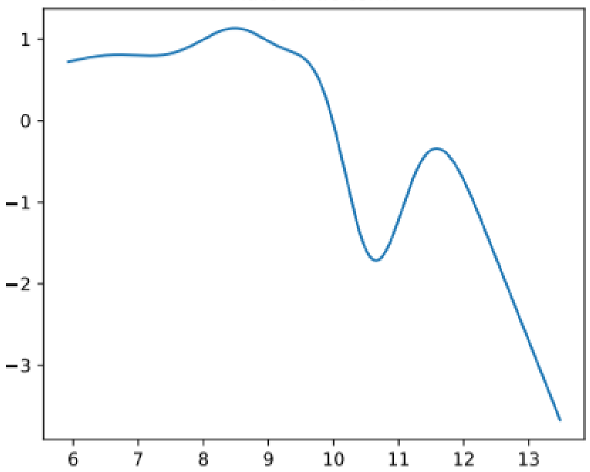


	Neural network	0.7012 ± 0,0062
	Linear SVM	0.6911
Baselines:	C4.5 with feature construction	0.7266 ± 0,0086 (15 nodes using feature construction)
	AdaBoost with feature construction	0.7280 ± 0.0063 (50 trees of 1 node each with feature construction)
	Gradient Boosting with feature construction	0.7446 ± 0.0071 (100 trees of 7 nodes each with feature construction)

RESULTS

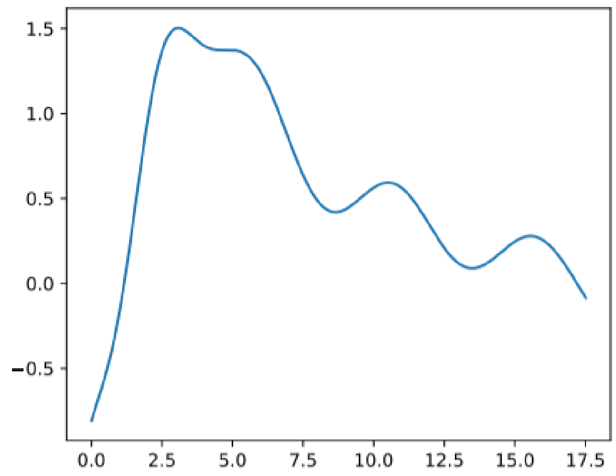
Example of a model (the lower the y value, the higher the probability to have a DVCS event):

$$p_z^e + p_z^p + p_z^{\gamma_1}$$



$+$

$$angle(p^{\gamma_2}, p^{\gamma_1} + p^{\gamma_2})$$



$+ \quad \ldots$

HOW TO USE PHYSICS KNOWLEDGE?

1. Feature construction
2. Using assumption on variable distributions to guide GAM/GA2M fitting

HOW TO USE PHYSICS KNOWLEDGE?

1. Feature construction
2. Using assumption on variable distributions to guide GAM/GA2M fitting

Some works use the a priori monotonicity of the input variables w.r.t. the target

Kotłowski, W., & Słowiński, R. (2009, June). Rule learning with monotonicity constraints. *ICML 2009.*

Fard, M. M., Canini, K., Cotter, A., Pfeifer, J., & Gupta, M. (2016). Fast and flexible monotonic functions with ensembles of lattices. *NIPS 2016*.

HOW TO USE PHYSICS KNOWLEDGE?

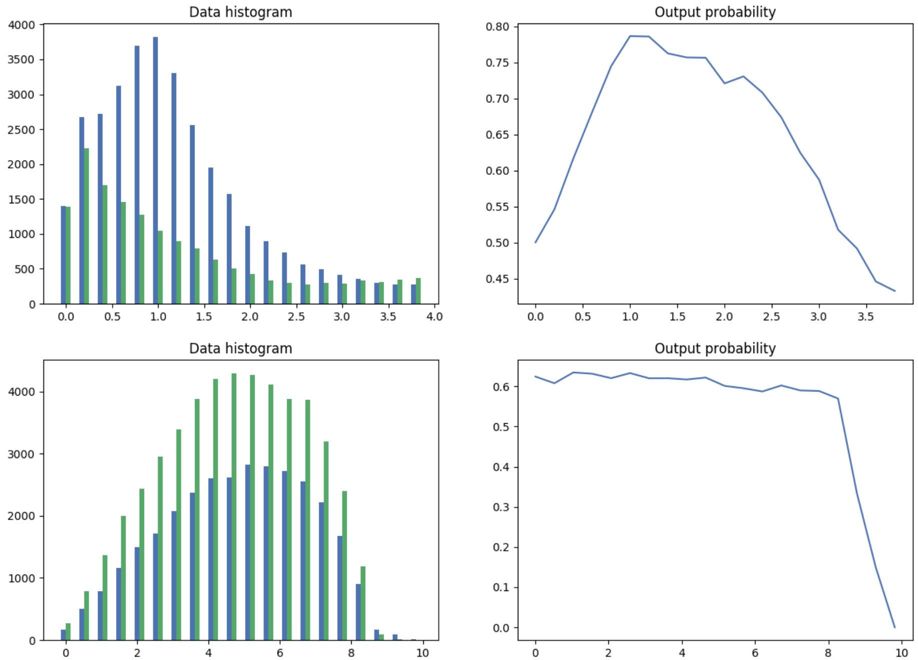
1. Feature construction
2. Using assumption on variable distributions to guide GAM/GA2M fitting

Some works use the a priori monotonicity of the input variables w.r.t. the target

Kotłowski, W., & Słowiński, R. (2009, June). Rule learning with monotonicity constraints. *ICML 2009.*

Fard, M. M., Canini, K., Cotter, A., Pfeifer, J., & Gupta, M. (2016). Fast and flexible monotonic functions with ensembles of lattices. *NIPS 2016*.

Monotonicity in physics?



Angle between hypothetical π^0 and photon

$$angle(p^{\gamma_1}, p^{\gamma_1} + p^{\gamma_2})$$

Missing mass $ep \rightarrow e\gamma$

$$\sqrt{\left(\|p^e\| + \|p^{\gamma_1}\| - 10{,}6 - M_p\right)^2 - \|p^e + p^{\gamma_1}\|^2}$$

HOW TO USE PHYSICS KNOWLEDGE?

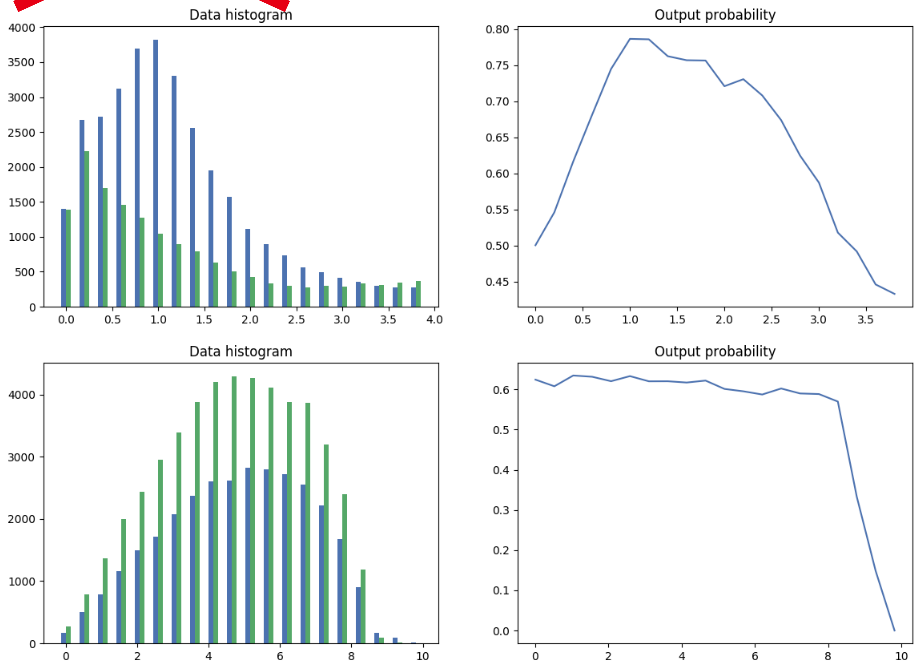
1. Feature construction
2. Using assumption on variable distributions to guide GAM/GA2M fitting

Some works use the a priori monotonicity of the input variables w.r.t. the target

Kotłowski, W., & Słowiński, R. (2009, June). Rule learning with monotonicity constraints. *ICML 2009.*

Fard, M. M., Canini, K., Cotter, A., Pfeifer, J., & Gupta, M. (2016). Fast and flexible monotonic functions with ensembles of lattices. *NIPS 2016*.

~~Monotonicity~~ in physics? **Bitonicity**



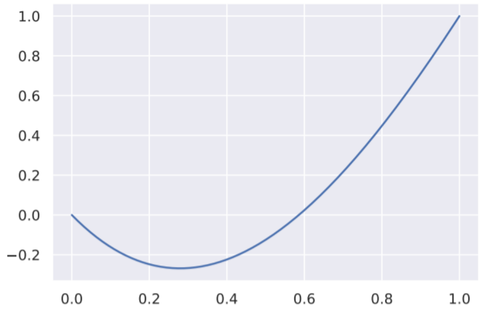
Angle between hypothetical π^0 and photon

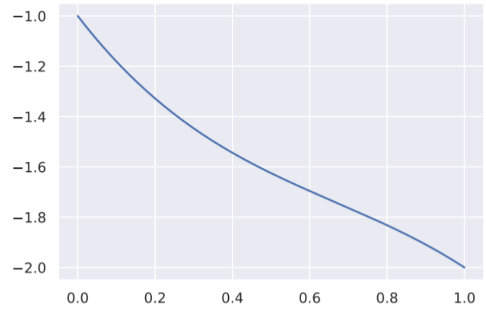
$$angle(p^{\gamma_1}, p^{\gamma_1} + p^{\gamma_2})$$

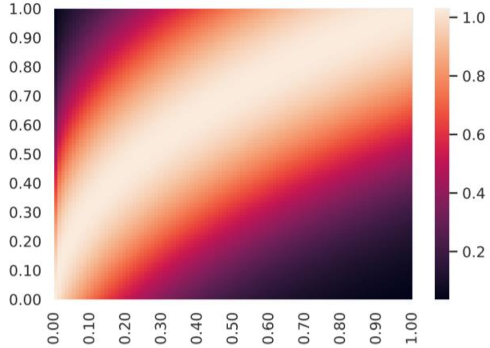
Missing mass $ep \rightarrow e\gamma$

$$\sqrt{\left(\|p^e\| + \|p^{\gamma_1}\| - 10{,}6 - M_p\right)^2 - \|p^e + p^{\gamma_1}\|^2}$$

BITONICITY





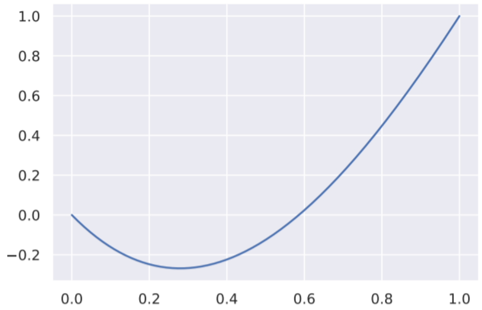


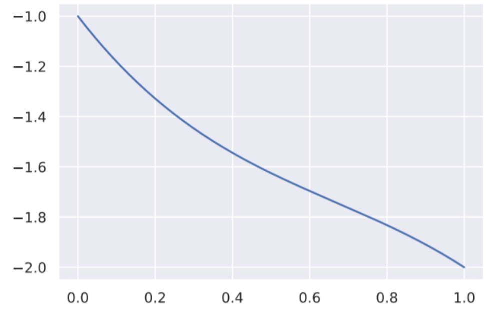
Bitonicity: either monotonic, or increasing then decreasing, or decreasing then increasing (i.e. unimodal)

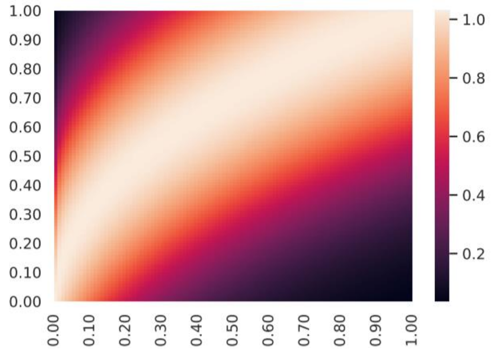
Bitonicity criteria:

difference between the function and its cumulative maximum/minimum

BITONICITY







Bitonicity: either monotonic, or increasing then decreasing, or decreasing then increasing (i.e. unimodal)

Bitonicity criteria:

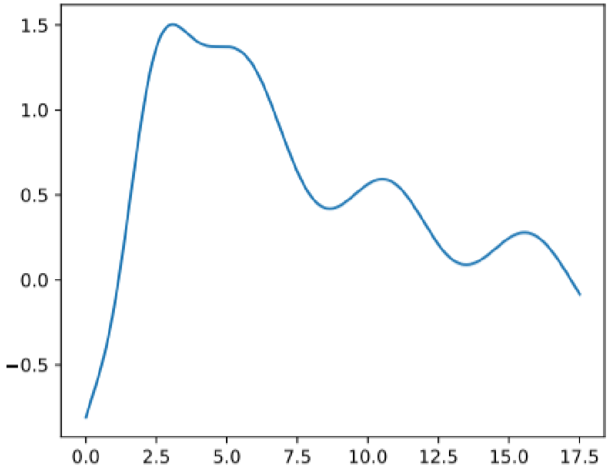
difference between the function and its cumulative maximum/minimum

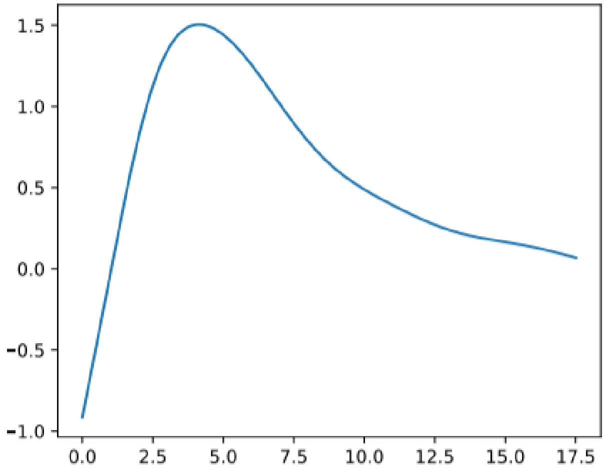
Penalization:

- in feature construction: fitness = $s - \lambda b$
- in shape functions with regularization in spline fitting

RESULTS

$$angle(p^{\gamma_2}, p^{\gamma_1} + p^{\gamma_2})$$



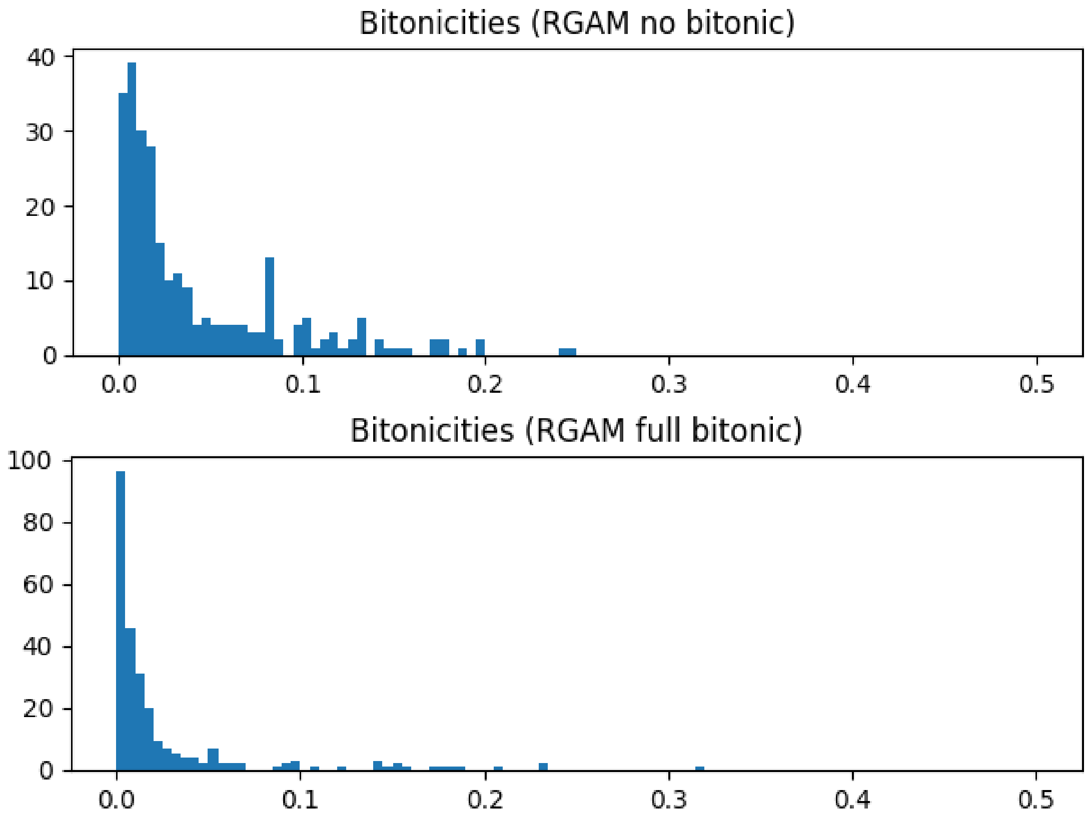


	Accuracy	Bitonicity score (penalty)
Without bitonicity constraint	0.738 ± 0.008	0.041 ± 0.048
With bitonicity constraint	0.735 ± 0.006	0.025 ± 0.046

RESULTS

	Accuracy	Bitonicity score (penalty)
Without bitonicity constraint	0.738 ± 0.008	0.041 ± 0.048
With bitonicity constraint	0.735 ± 0.006	0.025 ± 0.046

Bitonicity penalties distributions:



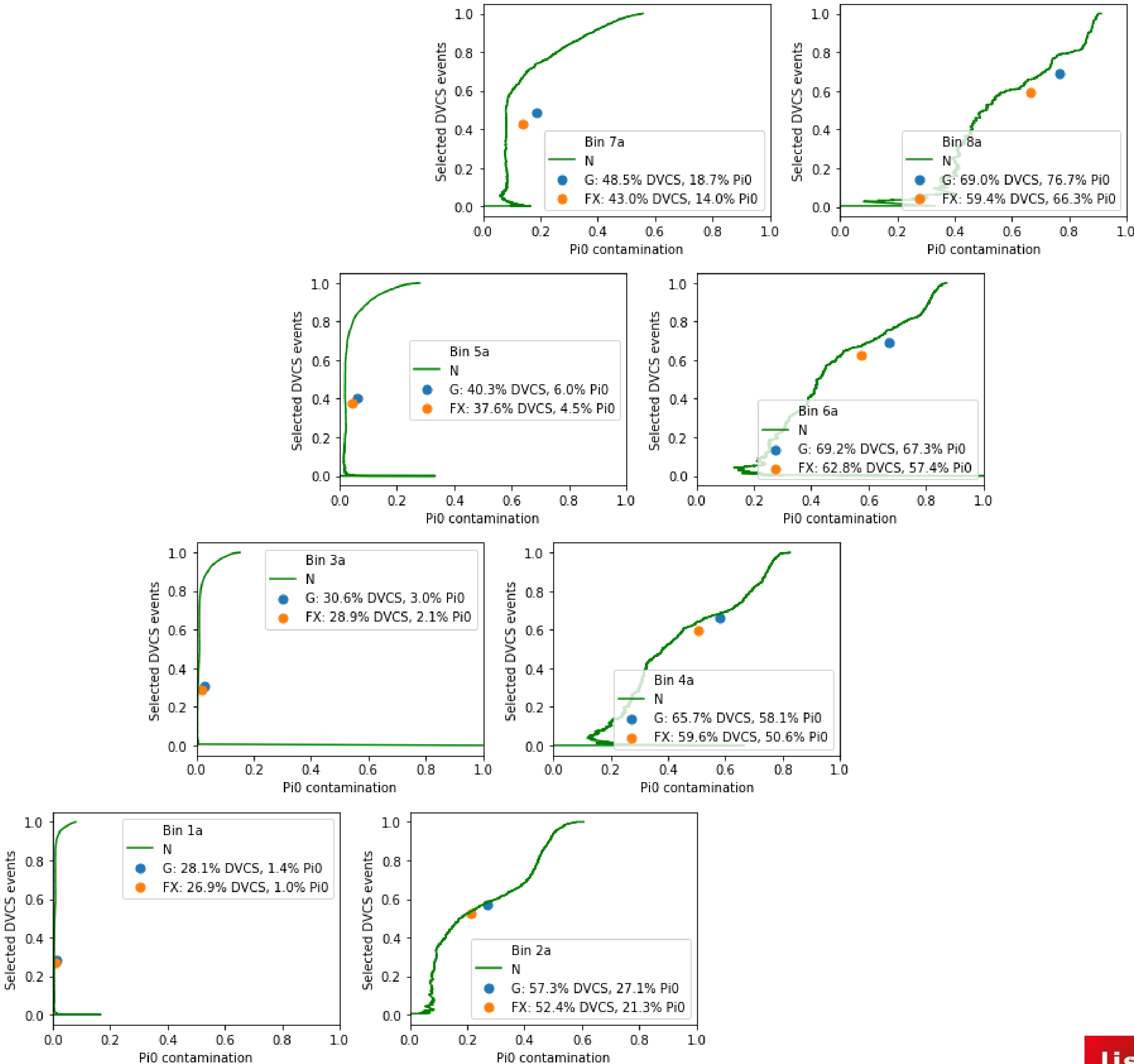
CONCLUSION

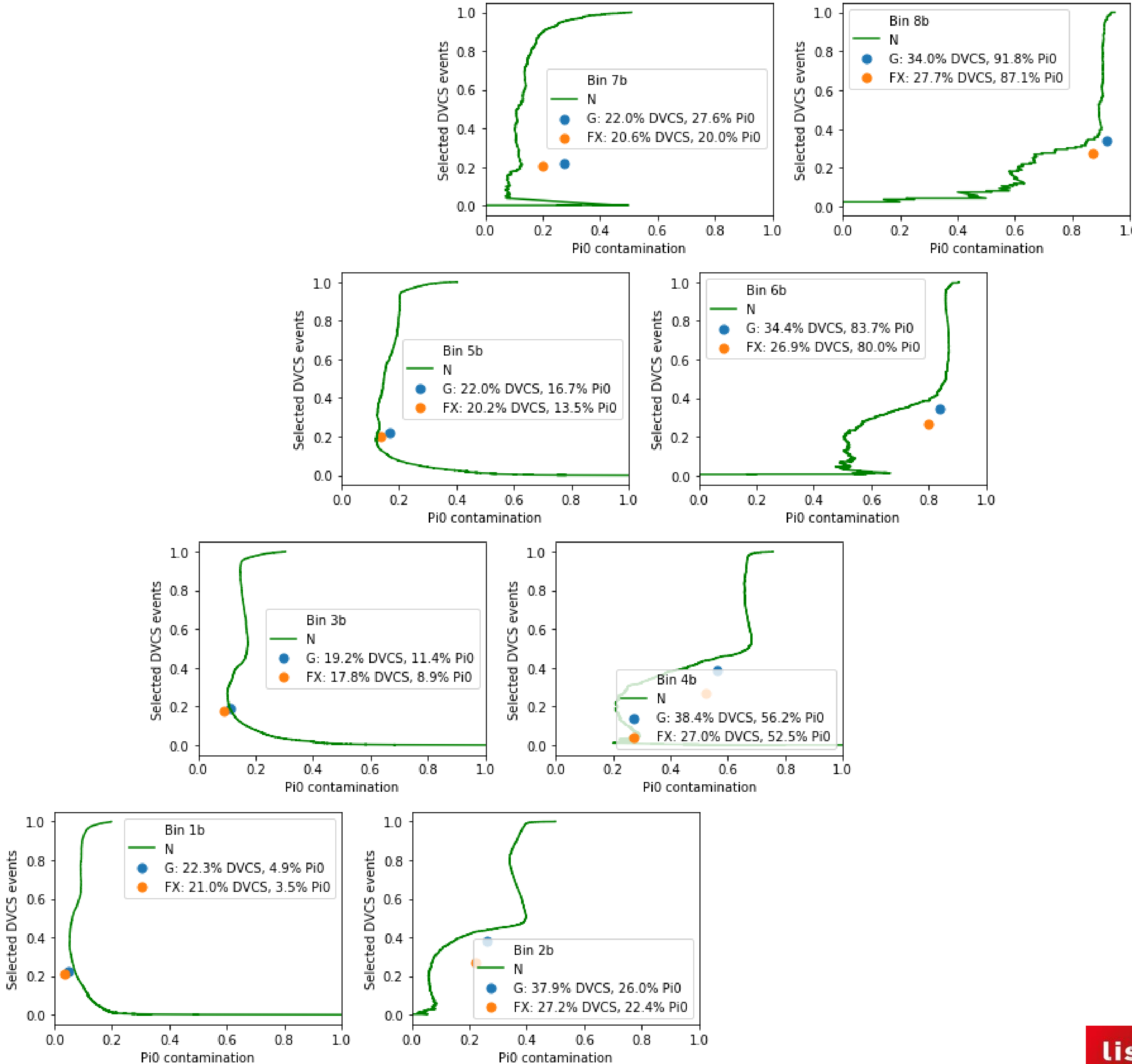
- GAM and GA2M: intelligible models, not perfectly transparent but more flexible than a rule base
- Gives good results on CLAS12 data particularly when exploiting feature construction
- Prior knowledge to include: bitonicity of the most discriminative variables
- Using this prior knowledge leads to simpler models that remain efficient
  - ⟶ Enforcing bitonicity is equivalent to increasing the regularization parameter
  - ⟶ The model is more understandable when it matches prior knowledge on the input variables

CONCLUSION

- GAM and GA2M: intelligible models, not perfectly transparent but more flexible than a rule base
- Gives good results on CLAS12 data particularly when exploiting feature construction
- Prior knowledge to include: bitonicity of the most discriminative variables
- Using this prior knowledge leads to simpler models that remain efficient
  - ⟶ Enforcing bitonicity is equivalent to increasing the regularization parameter
  - ⟶ The model is more understandable when it matches prior knowledge on the input variables

Thank you for listening!


Bin 7a
N
G: 48.5% DVCS, 18.7% Pi0
FX: 43.0% DVCS, 14.0% Pi0
Bin 8a
N
G: 69.0% DVCS, 76.7% Pi0
FX: 59.4% DVCS, 66.3% Pi0
Bin 5a
N
G: 40.3% DVCS, 6.0% Pi0
FX: 37.6% DVCS, 4.5% Pi0
Bin 6a
N
G: 69.2% DVCS, 67.3% Pi0
FX: 62.8% DVCS, 57.4% Pi0
Bin 3a
N
G: 30.6% DVCS, 3.0% Pi0
FX: 28.9% DVCS, 2.1% Pi0
Bin 4a
N
G: 65.7% DVCS, 58.1% Pi0
FX: 59.6% DVCS, 50.6% Pi0
Bin 1a
N
G: 28.1% DVCS, 1.4% Pi0
FX: 26.9% DVCS, 1.0% Pi0
Bin 2a
N
G: 57.3% DVCS, 27.1% Pi0
FX: 52.4% DVCS, 21.3% Pi0
Selected DVCS events
Pi0 contamination
0.0
0.2
0.4
0.6
0.8
1.0


Bin 7b
N
G: 22.0% DVCS, 27.6% Pi0
FX: 20.6% DVCS, 20.0% Pi0
Bin 8b
N
G: 34.0% DVCS, 91.8% Pi0
FX: 27.7% DVCS, 87.1% Pi0
Bin 5b
N
G: 22.0% DVCS, 16.7% Pi0
FX: 20.2% DVCS, 13.5% Pi0
Bin 6b
N
G: 34.4% DVCS, 83.7% Pi0
FX: 26.9% DVCS, 80.0% Pi0
Bin 3b
N
G: 19.2% DVCS, 11.4% Pi0
FX: 17.8% DVCS, 8.9% Pi0
Bin 4b
N
G: 38.4% DVCS, 56.2% Pi0
FX: 27.0% DVCS, 52.5% Pi0
Bin 1b
N
G: 22.3% DVCS, 4.9% Pi0
FX: 21.0% DVCS, 3.5% Pi0
Bin 2b
N
G: 37.9% DVCS, 26.0% Pi0
FX: 27.2% DVCS, 22.4% Pi0
Selected DVCS events
Pi0 contamination