



---

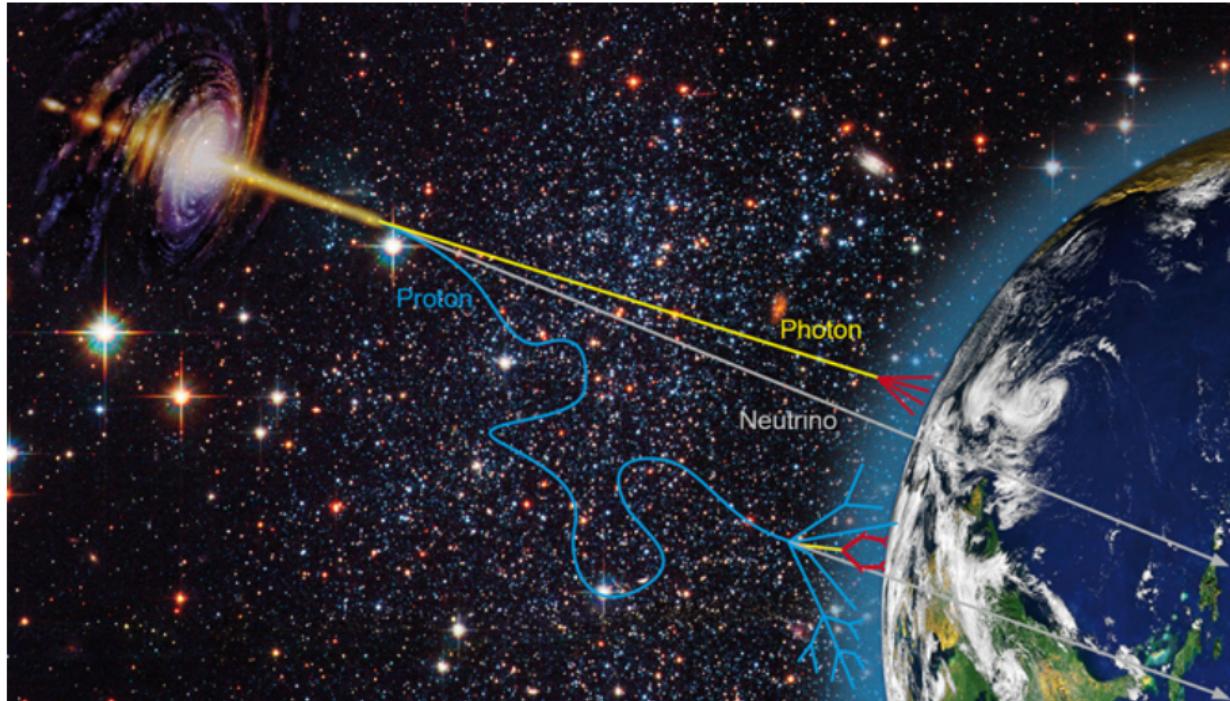
# Gamma/Hadron-Separation mit gemessenen Untergrunddaten bei FACT

---

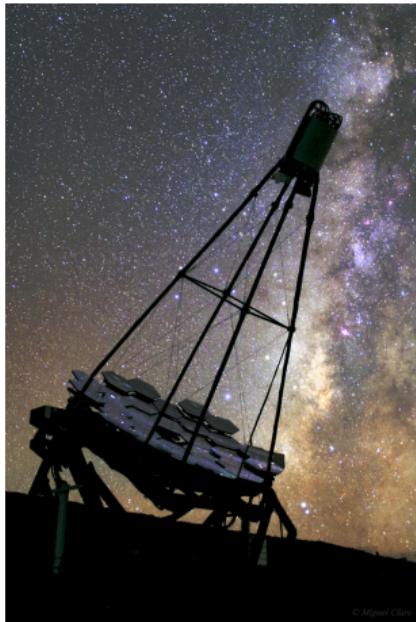
Maximilian Sackel

**19. Oktober 2017**

Experimentelle Physik 5b

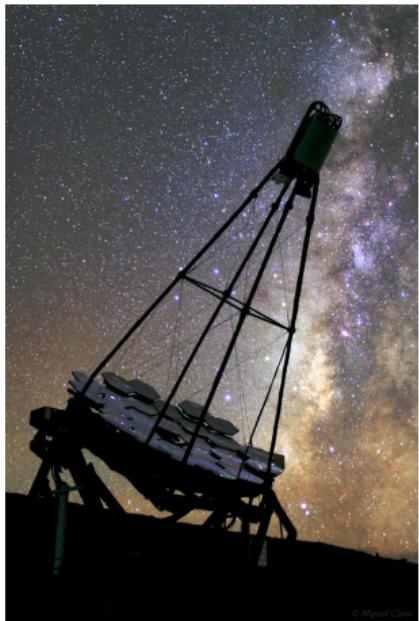


Quelle: DESY, Astroparticle Physik



Quelle: Carlo, FACT Cherenkov Telescope in a Milky Way Backlight

## First G-APD Cherenkov Telescope



Standort

La Palma, Roque de  
los Muchachos, 2200 m

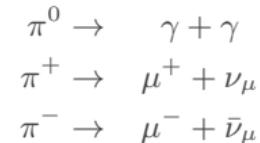
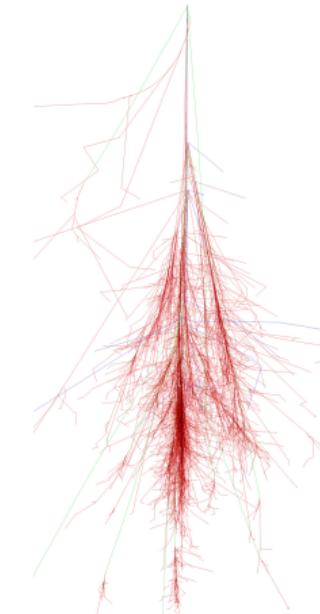
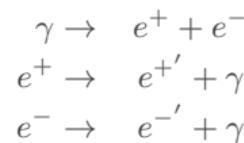
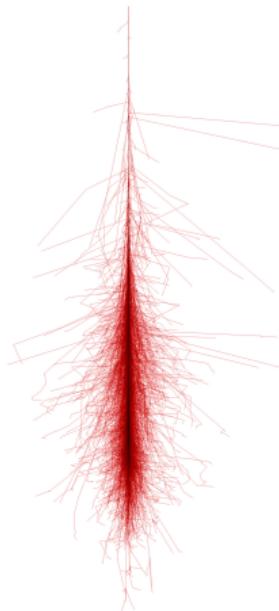
Reflektor

30 Spiegel, 9.5 m<sup>2</sup>  
Spiegelfläche

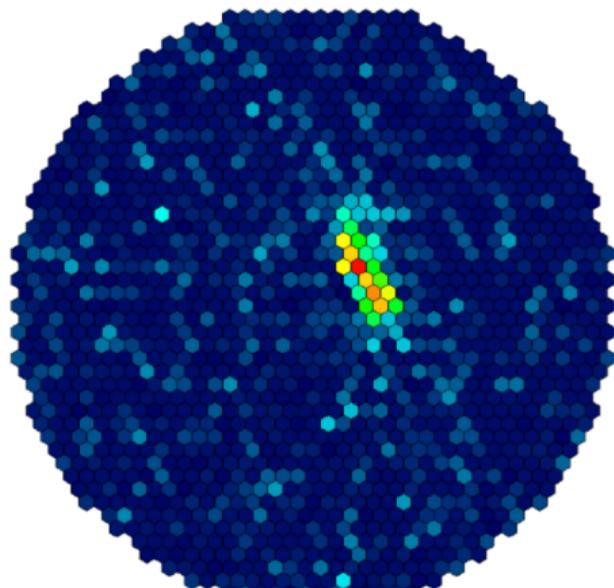
Kamera

1440 SiPMs,  
robust und sensitiv

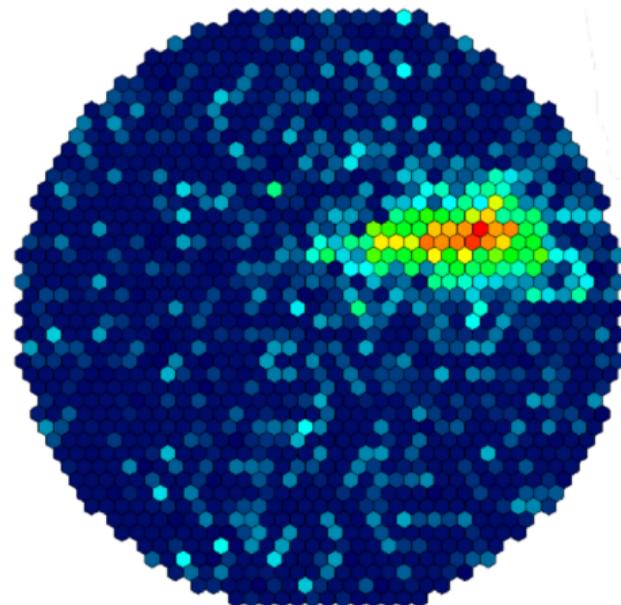
Quelle: Carlo, FACT Cherenkov Telescope in a Milky Way Backlight

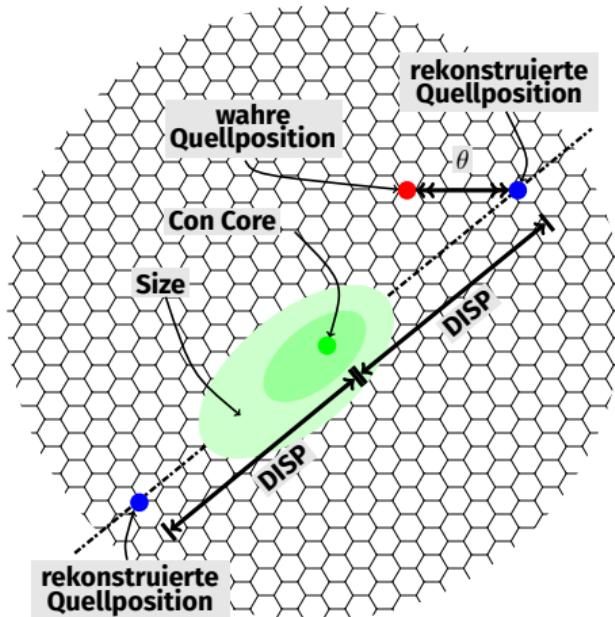


Quelle: Schmidt, CORSIKA – an Air Shower Simulation Program

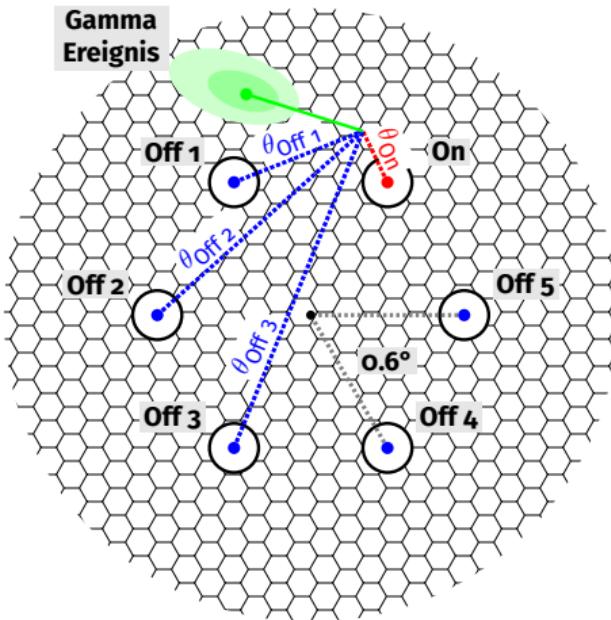


Quelle: Carlo, FACT Cherenkov Telescope in a Milky Way Backlight



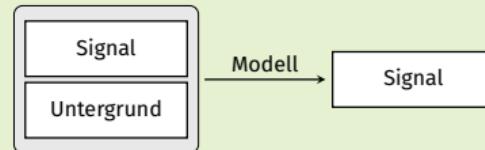


- berechne Bildparameter (Hillas Parameter) des Kamerabildes
- Bildparameter werden zum Klassifizieren benötigt
- Richtung des Schauers nicht eindeutig

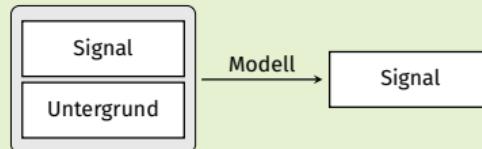


- FACT nimmt keine OFF-Daten
- Daten werden im Wobble-Modus aufgenommen

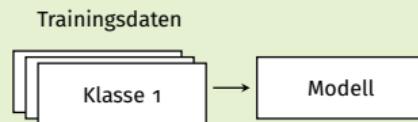
## Separation



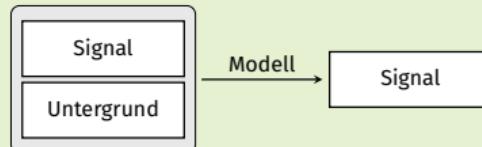
## Separation



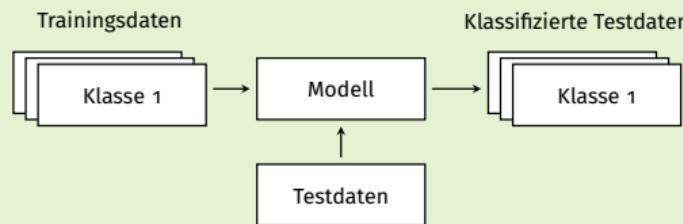
## Überwachtes Lernen



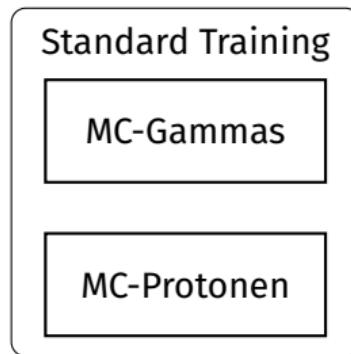
## Separation



## Überwachtes Lernen

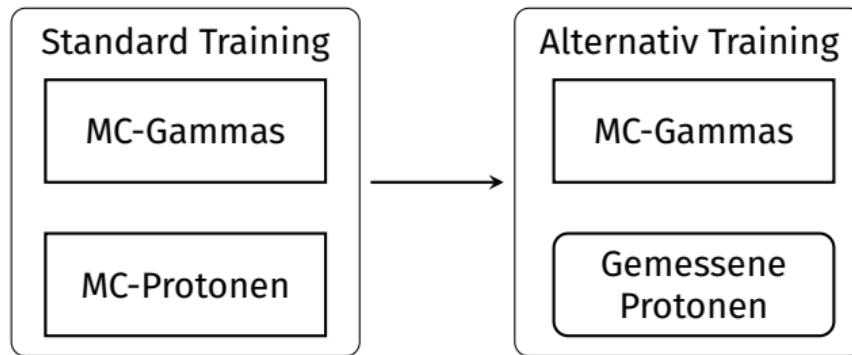


## Trainingsdaten



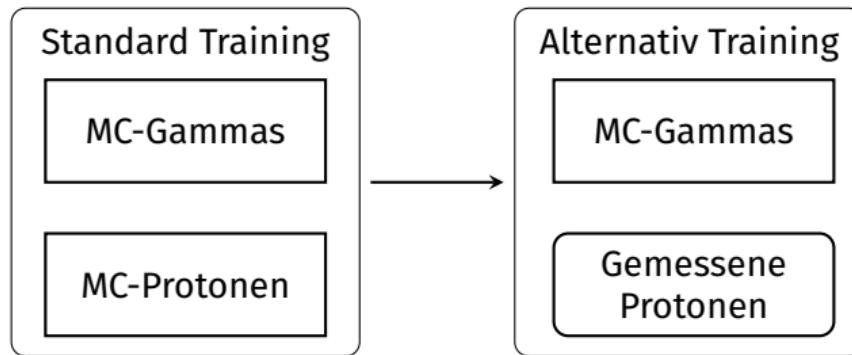
- Trainingsdaten simuliert mit CORSIKA, sowie Detektorsimulationen

## Trainingsdaten



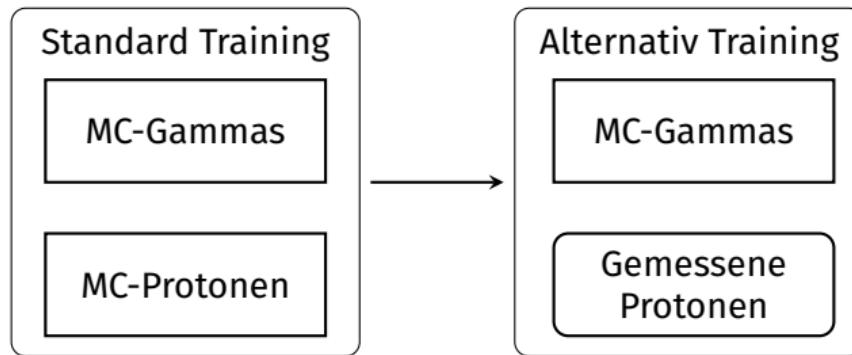
- Trainingsdaten simuliert mit CORSIKA, sowie Detektorsimulationen

## Trainingsdaten



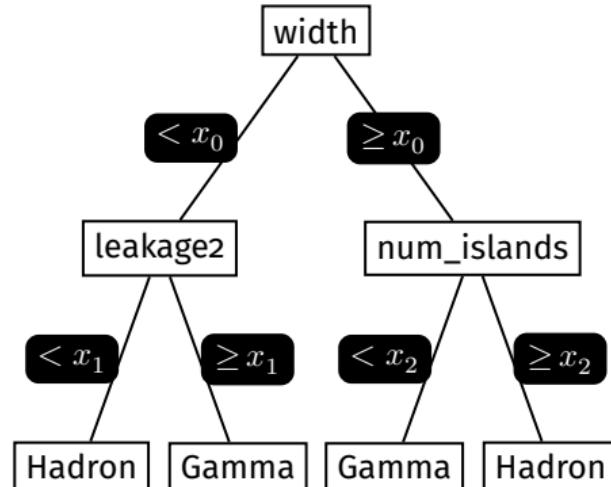
- Trainingsdaten simuliert mit CORSIKA, sowie Detektorsimulationen
- gemessener Untergrund folgt der wahren Verteilung
  - Verbesserung der Separation

## Trainingsdaten



- Trainingsdaten simuliert mit CORSIKA, sowie Detektorsimulationen
- gemessener Untergrund folgt der wahren Verteilung
  - Verbesserung der Separation
- Protonen-Simulation kann eingestellt werden

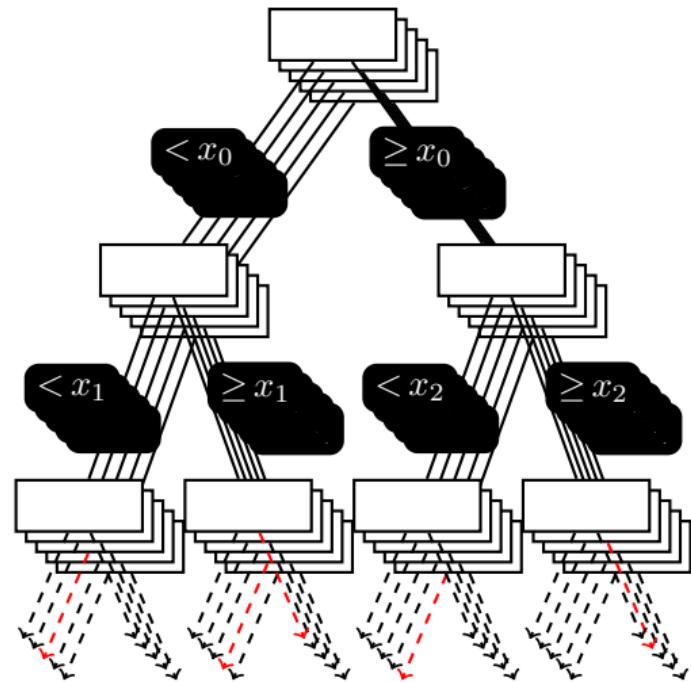
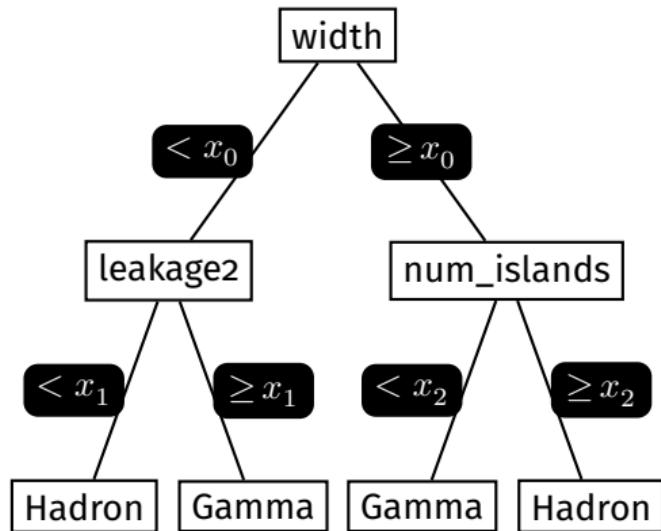
## Entscheidungsbaum



- Verknüpfte Abfragen
- Loss-function
- Beschränkung der Komplexität

Ereignis	width	leakage2	num_islands	...	Konfi.
1	4.2	0.4	3	...	0.12
2	3.8	0.0	2	...	0.56
3	15.3	0.8	1	...	0.08
4	7.7	0.1	1	...	0.43
5	6.2	0.0	1	...	0.85

## Random Forest



Modelle

## Random Forest

**Entscheidungsbaum**

Ereignis	Konfi.
1	0.12
2	0.56
3	0.08
4	0.43
5	0.85

**Random Forest**

Ereignis	Konf <sub>1</sub>	Konf <sub>2</sub>	Konf <sub>3</sub>	...	Konf
1	0.12	0.01	0.08	...	0.06
2	0.40	0.66	0.53	...	0.56
3	0.02	0.17	0.10	...	0.08
4	0.41	0.42	0.42	...	0.43
5	0.96	0.81	0.85	...	0.85

## Boosted Trees



- additives Training
- höhere Gewichtung von Fehlklassifizierungen
- ausgeglichenere Vorhersage
- lässt sich nicht parallelisieren
- Modelle mit geringerer Komplexität

## Boosted Trees



- additives Training
- höhere Gewichtung von Fehlklassifizierungen
- ausgeglichenere Vorhersage
- lässt sich nicht parallelisieren
- Modelle mit geringerer Komplexität

## Boosted Trees

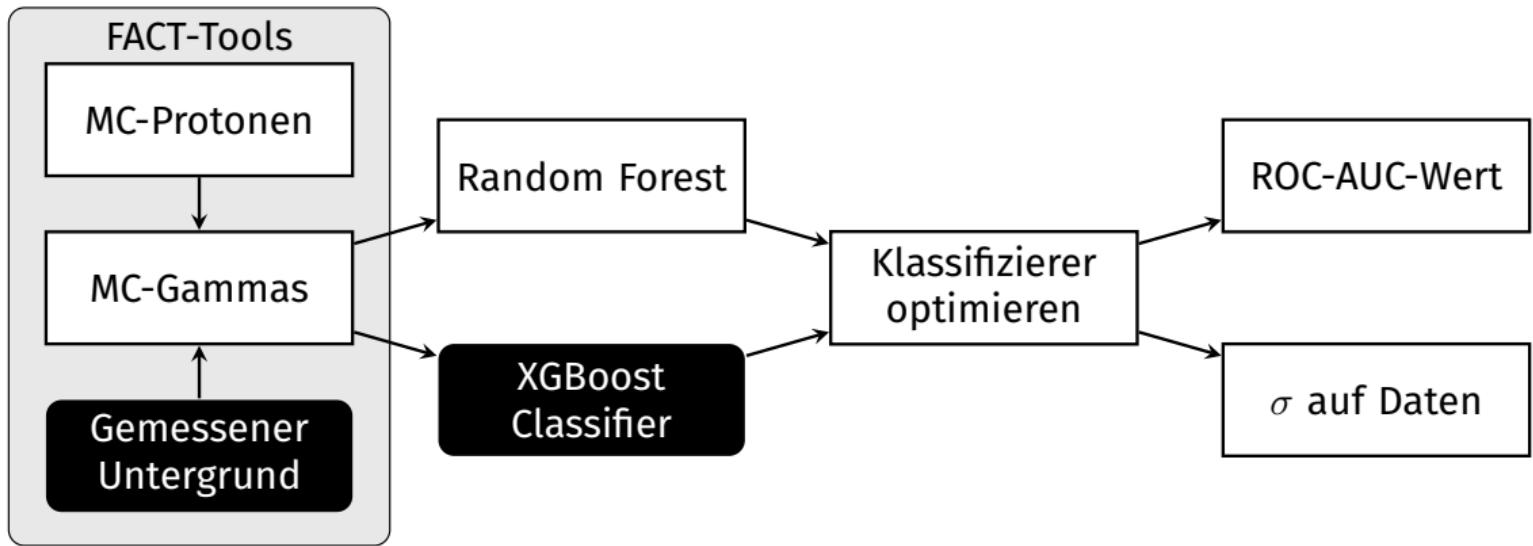


- additives Training
- höhere Gewichtung von Fehlklassifizierungen
- ausgeglichenere Vorhersage
- lässt sich nicht parallelisieren
- Modelle mit geringerer Komplexität

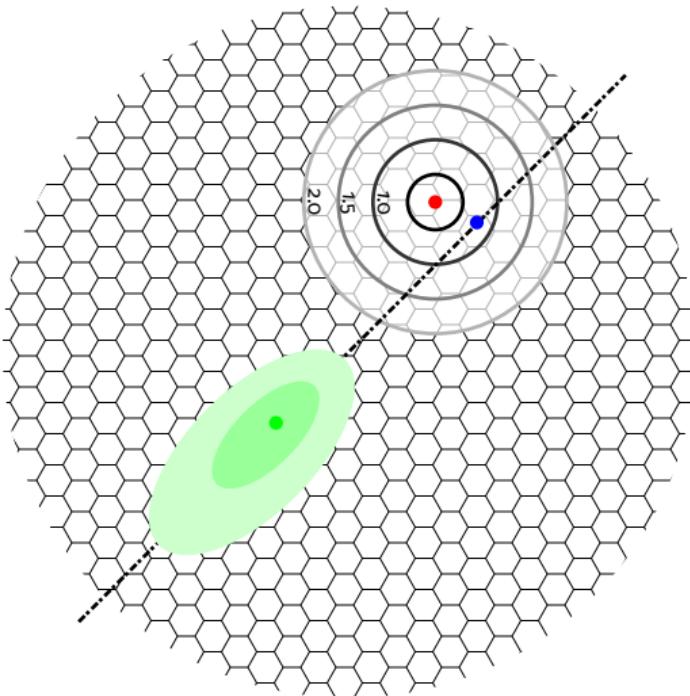
## Boosted Trees



- additives Training
- höhere Gewichtung von Fehlklassifizierungen
- ausgeglichenere Vorhersage
- lässt sich nicht parallelisieren
- Modelle mit geringerer Komplexität



## Erstellen des Trainingsdatensatzes

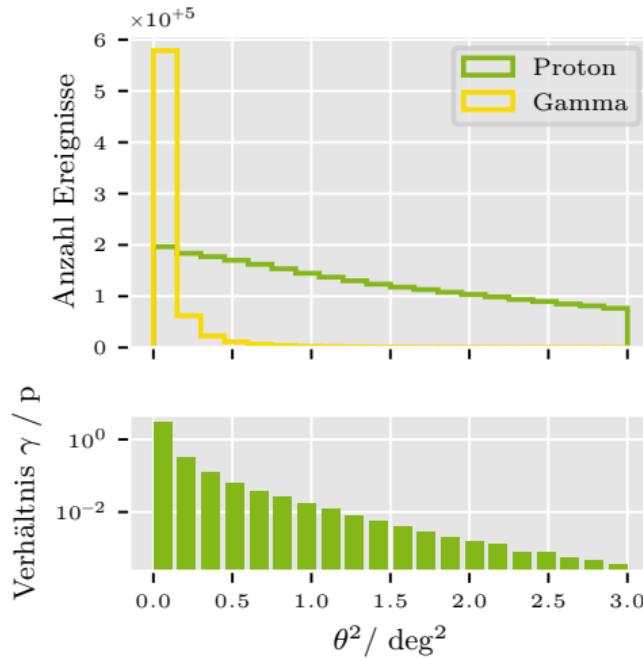


Training

Protonen werden aus Daten extrahiert

- ohne Eingang des Daten-Monte Carlo-Mismatches
- möglichst reiner Datensatz
- wenige diffuse Gamma lassen sich physikalisch motivieren

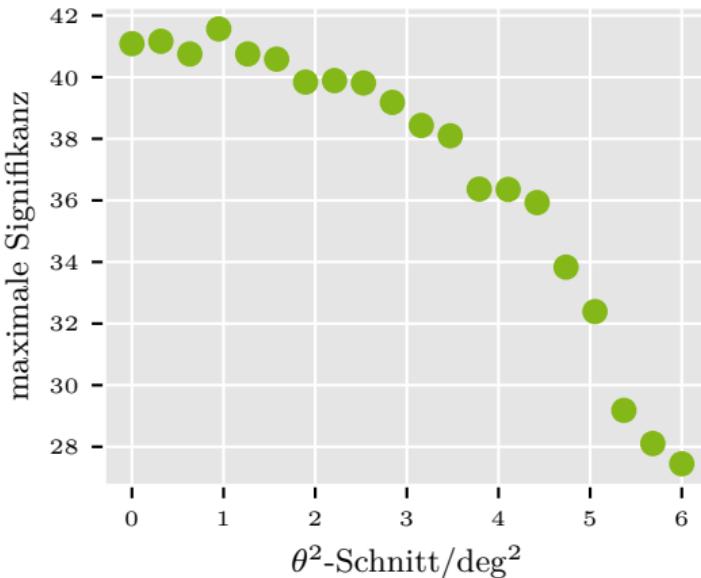
## Erstellen des Trainingsdatensatzes



Protonen werden aus Daten extrahiert

- ohne Eingang des Daten-Monte Carlo-Mismatches
- möglichst reiner Datensatz
- wenige diffuse Gamma lassen sich physikalisch motivieren

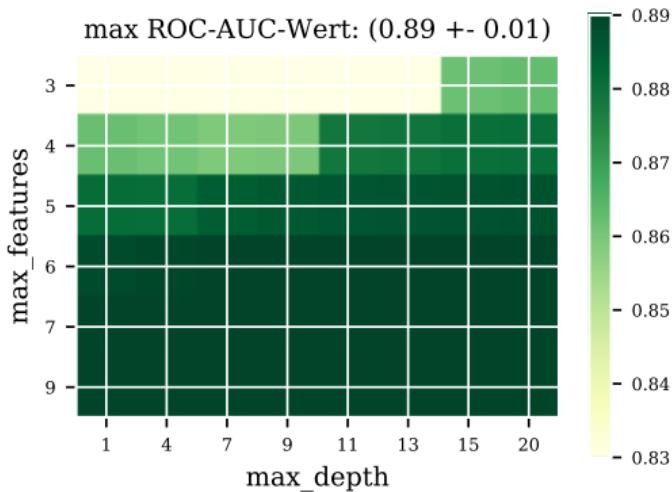
## Überprüfen des Trainingsdatensatzes



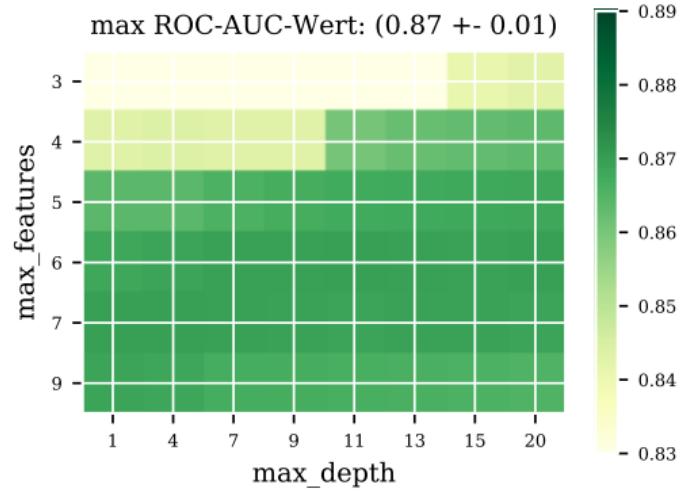
- simulierte Proton-Datensets werden durch verschiedene  $\theta^2$ -Schnitte erstellt
  - Modelle werden mit verschiedenen Proton-Datensets trainiert
1. Detektoreigenschaften für große  $\theta^2$  nicht zu vernachlässigen
  2. Abwegen zwischen Signifikanz und Reinheit des Datensatzes

## Optimieren der Modelle

### Gemessener Untergrund

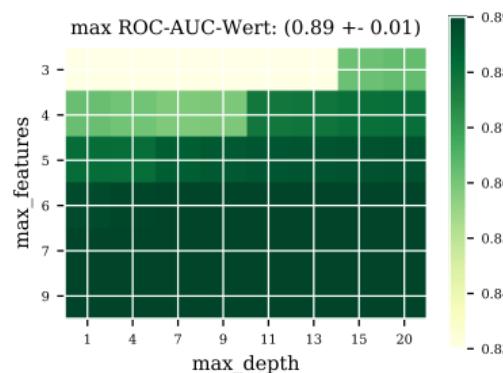


### Monte Carlo-Protonen

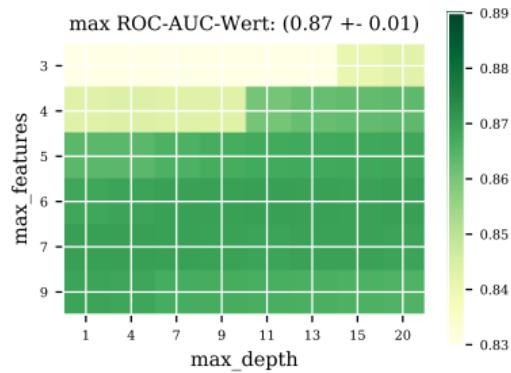


## Optimieren der Modelle

### Gemessener Untergrund



### Monte Carlo Protonen

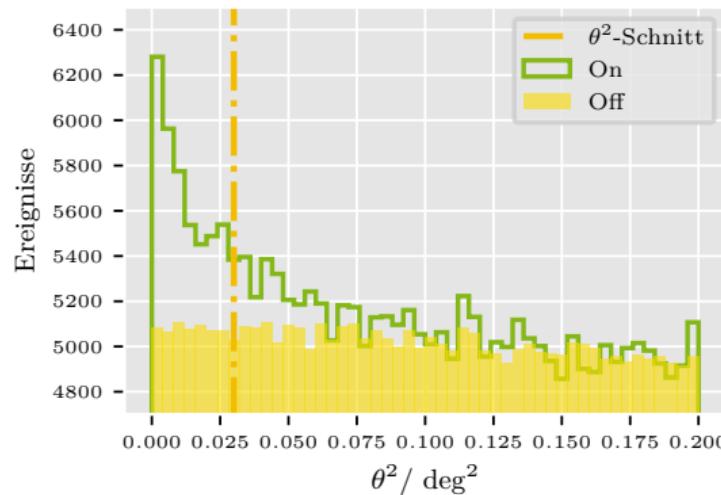
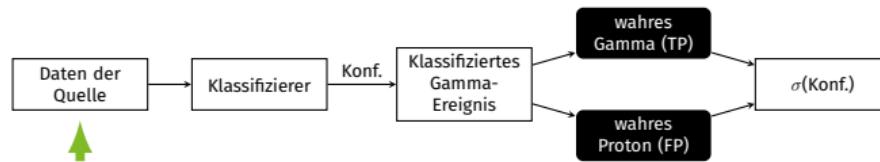


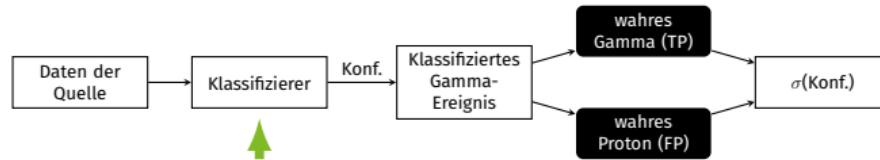
	Messdaten	MC – Daten
XGBoost (Tiefe 1)	0.869(5)	0.86(2)
Random Forest	0.89(1)	0.87(1)

## Validieren auf echten Daten

### Li und Ma Signifikanz

$$S(N_{\text{on}}, N_{\text{off}}, \alpha) = \sqrt{2} \left( N_{\text{on}} \ln \left[ \frac{1+\alpha}{\alpha} \left( \frac{N_{\text{on}}}{N_{\text{on}} + N_{\text{off}}} \right) \right] + N_{\text{off}} \ln \left[ (1+\alpha) \left( \frac{N_{\text{off}}}{N_{\text{on}} + N_{\text{off}}} \right) \right] \right)^{1/2}$$



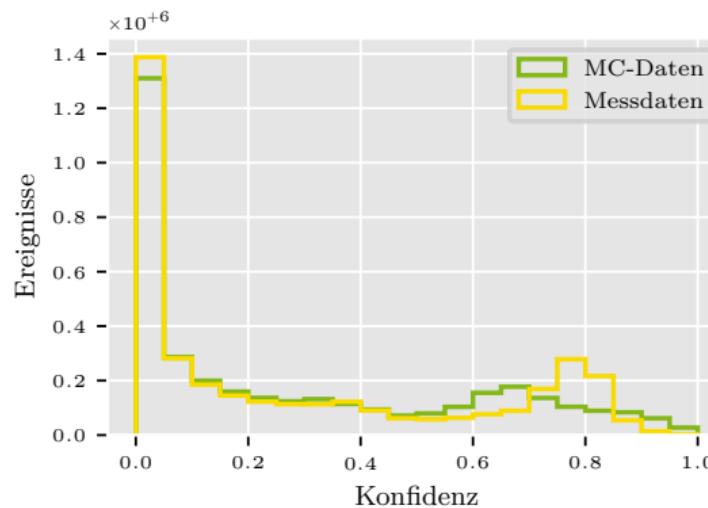
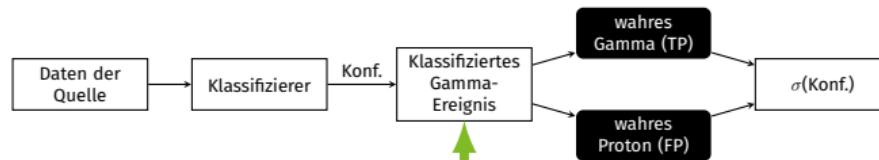


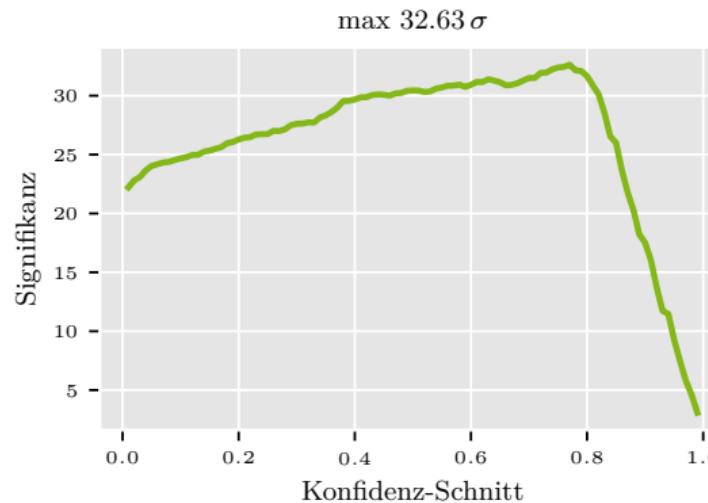
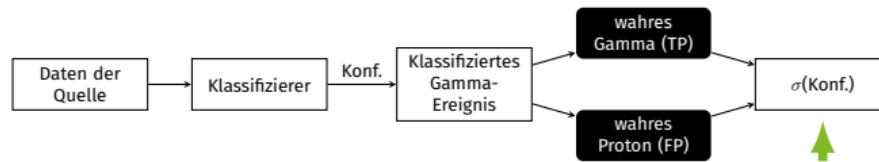
## Random Forest

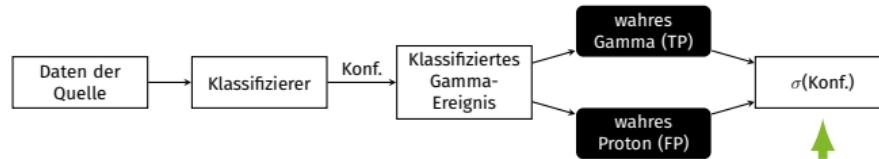
- Komplexität
- viele Spezialisten auf ihrem Gebiet
- sensitiv auf Daten-Monte Carlo-Matches

## XGBoost Classifier (Tiefe 1)

- geringe Komplexität
- resistent gegen Mismatches







	Krebsnebel		Markarian 501	
	Random Forest	XGBoost (Tiefe = 1)	Random Forest	XGBoost (Tiefe = 1)
unklassifizierte Daten		21.4 $\sigma$		17.1 $\sigma$
MC-Protonen	41.9 $\sigma$	41.3 $\sigma$	35.5 $\sigma$	35.6 $\sigma$
gemessene Protonen	32.9 $\sigma$	37.8 $\sigma$	23.6 $\sigma$	35.2 $\sigma$

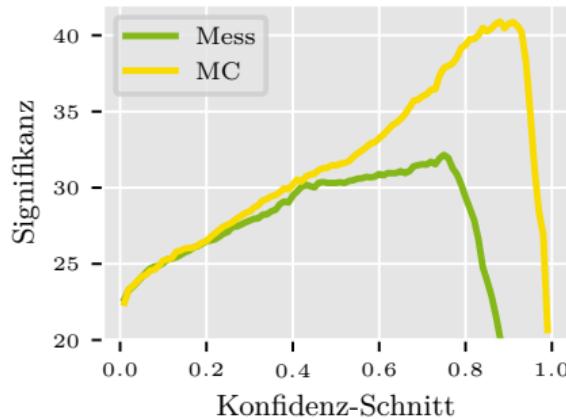
## Thesen

- Modelle trainieren Unterschied zwischen echten und simulierten Daten
  - klassifizierte Datensätze weisen bei komplexeren Modellen niedrigere Signifikanzen auf

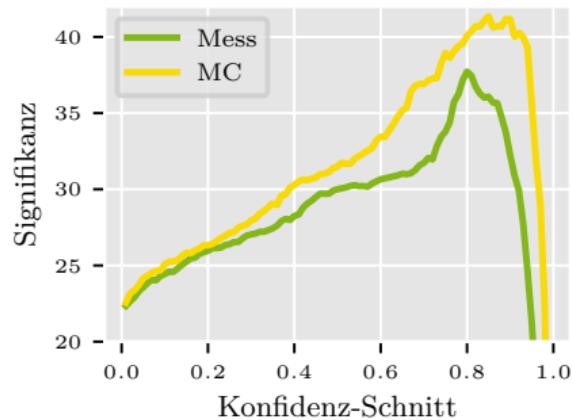
## Thesen

- Modelle trainieren Unterschied zwischen echten und simulierten Daten
  - klassifizierte Datensätze weisen bei komplexeren Modellen niedrigere Signifikanzen auf
- Reduzierung von schlecht simulierten Attributen
  - Erhöhung der Signifikanz durch Reduzierung von Mismatches

## Random Forest

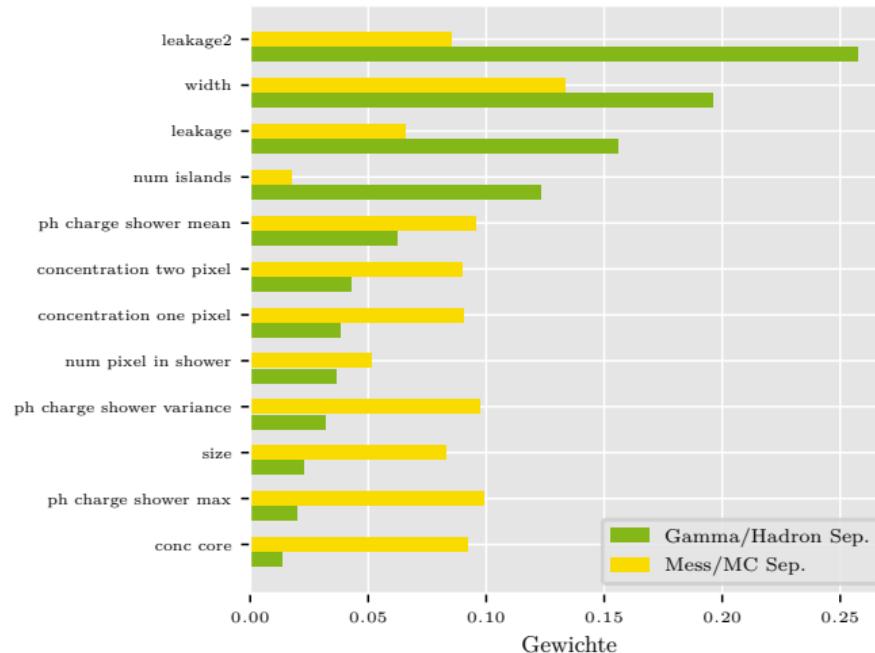


## XGBoost Classifier (Tiefe 1)



- Konfidenzverteilung nicht direkt vergleichbar
- beide Bäume nach demselbem Kriterium gebaut

## Rekursive Feature Elimination



	ohne Feature Elimination	mit Feature Elimination
ROC-AUC-Wert	0.64	0.61
Li und Ma Signifikanz	$32.9\sigma$	$34.4\sigma$

- ROC-AUC-Wert: Separation zwischen Monte-Carlo und gemessenem Untergrund
- Signifikanz des Datensatz vor und nach dem Entfernen der Attribute

- momentan keine Verbesserung
- Verbesserung der Monte Carlo-Simulationen
- Datennahme von OFF-Daten
- Mismatch-unempfindliches Modell