

2026/10/19

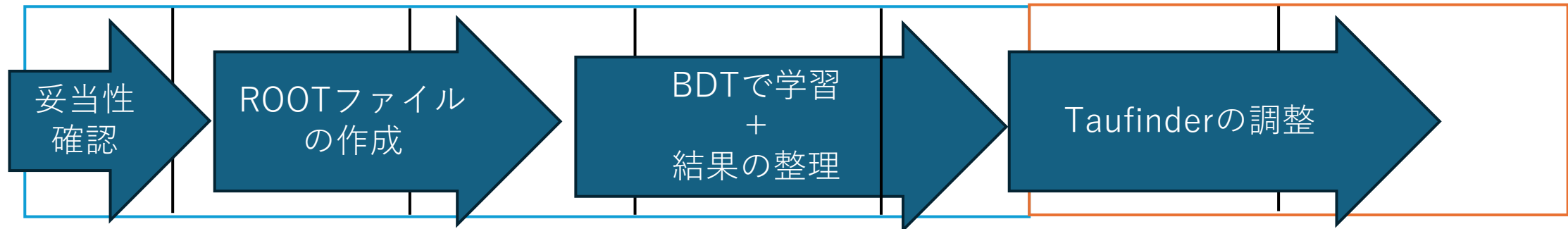
進捗報告

活動内容

- BDTに学習させるROOTファイルの生成
 - ファイルの生成
 - 重み付け
- BDTを用いた多変量解析の実行
 - モデルの学習
 - 全データへの適用最適カット値の探索
 - 結果の評価・グラフ化

10月

11月



BDTに学習させるROOTファイルの生成

• 十分な統計量の確保

- BDTの学習および性能評価のため、信号・背景事象ともに大規模なシミュレーションデータが必要になる。

• Marlinによるデータ抽出

- 物理事象の lcio ファイルから解析に必要な運動学的変数群を抽出し、生データ（ROOTファイル）を生成。

• バッチジョブによる並列処理

- 全データの処理はランタイム上限を超過する問題があったため、KEKCCのジョブシステム（bsub）を活用して分割・並列処理を行い、効率的に生成を完遂した。

生成された各事象に対する生イベント数

プロセス	ファイル名	生イベント数
$ZH \rightarrow \nu\bar{\nu}s\bar{s}$	nnss	200000
$ZH \rightarrow \nu\bar{\nu}b\bar{b}$	nnbb	200000
$ZH \rightarrow \nu\bar{\nu}c\bar{c}$	nncc	200000
$ZH \rightarrow \nu\bar{\nu}g\bar{g}$	nngg	200000
$ZH \rightarrow \nu\bar{\nu}ZZ^*$	hzz	1000000
$ZH \rightarrow \nu\bar{\nu}WW^*$	hww	1000000
$ZH \rightarrow \nu\bar{\nu}\tau^+\tau^-$	htautau	1000000
$ZZ \rightarrow q\bar{q}l^+l^-/q\bar{q}\nu\bar{\nu}$	4fZZsemileptonic	3099600
$\nu\bar{\nu}Z \rightarrow \nu\bar{\nu}q\bar{q}$	4fZnusemileptonic	2298400
$WW \rightarrow lvq\bar{q}$	4fWWsemileptonic	36086800
$WW \rightarrow lvq\bar{q}$	4fWsemileptonic	45357600
$WW/ZZ \rightarrow lvq\bar{q}$	4fmixleptonic	8400000

重み付けとは

- **シミュレーションと現実のギャップを埋める操作**

- シミュレーション（MC）では、解析に必要な統計精度を確保するため、実際の発生確率（断面積）に関わらず、意図的に大量の事象（例：数十万～数千万件）を生成している。

- **本来の物理的な発生スケールへの変換**

- 生成したそのままの事象数で解析を行うと、現実と全く異なるS/B比（シグナルと背景の比率）になってしまう。
- そこで、実際のILC実験の条件（積分ルミノシティ 900 fb^{-1} など）を想定し、各事象が「現実の検出器で何回観測されるか」という本来の期待値にスケールを合わせる必要がある。

- **重み付けのメリット・目的**

- 現実の物理環境（正しい事象数分布）を正確に再現できる。
- これにより、シグナルの有意性（Significance：どれだけ背景に埋もれず発見できるか）の正しい評価や、最適なカット条件の探索が可能になる。

weight

重み

$$Weight = \frac{\mathcal{L} \times \sigma}{N_{gen}} \times W_{pol} \times BR$$

$\mathcal{L}$ (Luminosity): 積分ルミノシティ 900 fb^{-1}

σ (Cross Section): 事象の断面積

N_{gen} (Total Generated Events): ある事象についての全生成事象数

W_{pol} : 偏極の重み

BR (Branching Ratio): 崩壊分岐比

重み付けしたイベント数

シミュレーションで生成されたすべての事象の重みを足し合わせた合計値

$$\begin{aligned} N_{weighted} &= \sum_i^{N_{gen}} Weight_i \\ &= N_{gen} \times \left(\frac{\mathcal{L} \times \sigma}{N_{gen}} \times W_{pol} \times BR \right) \\ &= \mathcal{L} \times \sigma \times W_{pol} \times BR \end{aligned}$$

想定ルミノシティと（偏極や分岐比を考慮した）物理的な断面積を掛け合わせた値になる
つまり、実験を行った際に、検出器で実際に観測され
と思われる、本来の事象数

*BRを考慮する必要があるのは今回はZH随伴生成のブランチであるものに対してだけで、その他の事象に対してはBR=1

重み付けの結果

生成された生イベント数と重み付けされたイベント数

プロセス	ファイル名	生イベント数	重み付けされたイベント数
$ZH \rightarrow \nu\bar{\nu}s\bar{s}$	nnss	200000	16.5
$ZH \rightarrow \nu\bar{\nu}b\bar{b}$	nnbb	200000	38721.6
$ZH \rightarrow \nu\bar{\nu}c\bar{c}$	nncc	200000	1952.9
$ZH \rightarrow \nu\bar{\nu}g\bar{g}$	nngg	200000	5751.2
$ZH \rightarrow \nu\bar{\nu}ZZ^*$	hzz	1000000	1771.7
$ZH \rightarrow \nu\bar{\nu}WW^*$	hww	1000000	14428.3
$ZH \rightarrow \nu\bar{\nu}\tau^+\tau^-$	htautau	1000000	4241.3
$ZZ \rightarrow q\bar{q}l^+l^-/q\bar{q}\nu\bar{\nu}$	4fZZsemileptonic	3099600	321101.2
$\nu\bar{\nu}Z \rightarrow \nu\bar{\nu}q\bar{q}$	4fZnusemileptonic	2298400	238962.4
$WW \rightarrow l\nu q\bar{q}$	4fWWsemileptonic	36086800	9887219.3
$WW \rightarrow l\nu q\bar{q}$	4fWsemileptonic	45357600	5404003.8
$WW/ZZ \rightarrow l\nu q\bar{q}$	4fmixleptonic	8400000	861910.4

•BDTを用いた多変量解析の実行

タスクの内容

- **モデルの学習**: 軽量化したデータを用いたTMVAによる学習とモデル（重みファイル）の構築
- **全データへの適用**: 完成したモデルを用いて、約1億件の全事象に対するBDTスコアを計算・付与
- **最適カット値の探索**: BDTスコアを走査し、シグナルの有意性（Significance）が最大となる条件を決定
- **結果の評価・グラフ化**: 最終的な事象分布の比較と物理結果の可視化

モデルの学習

目的: 多変量解析ツール (TMVA) による識別器の構築

- 44個の運動学的変数（質量、運動量、フレーバータグなど）を入力とし、シグナルと背景を最適に分離する決定木 (BDT) を学習させる。

課題 メモリ制限による学習のクラッシュ

- 約1億件に及ぶ全データをそのままTMVAに入力すると、計算ノードのメモリ上限 (4GB) を超過し、プログラムが強制終了してしまう問題に直面した。

工夫と解決策: 物理分布を保ったデータの軽量化

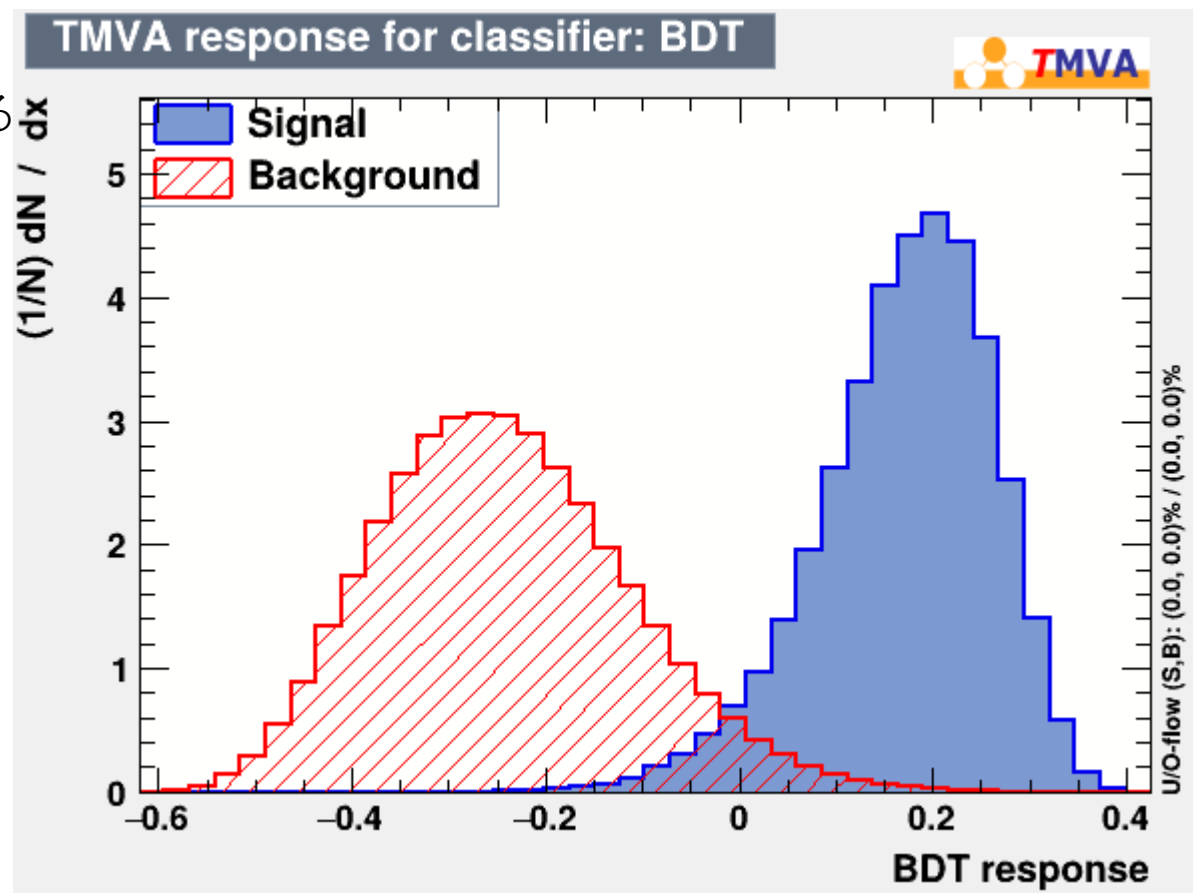
- 各背景事象の比率（物理的な断面積の比）を維持したまま、事象数を一様に 1/50 に間引いた軽量版データ（約200万件）を生成して学習を実行。
- 加えて、学習の妨げとなる極端な異常値 (NaNや無限大) を持つ事象をプレカットで排除。
⇒メモリ溢れを回避しつつ、重みファイルの構築に成功した

モデルの学習の結果①

BDTスコアに対する
事象の割合

- SignalとBackground
がわかれていることを
確認できる
- 両者の分布の重なりが
少ない、適切なスコアで
カット（足切り）を行
えば、背景だけを排除
できることがわかる

オーバーラーニングの
チェックは？



シグナルっぽくない

シグナルらしい

今後1週間の予定

BDTを用いた多変量解析の実行における以下のタスクを終わらせる

- 全データへの適用
- 最適カット値の探索
- 結果の評価・グラフ化

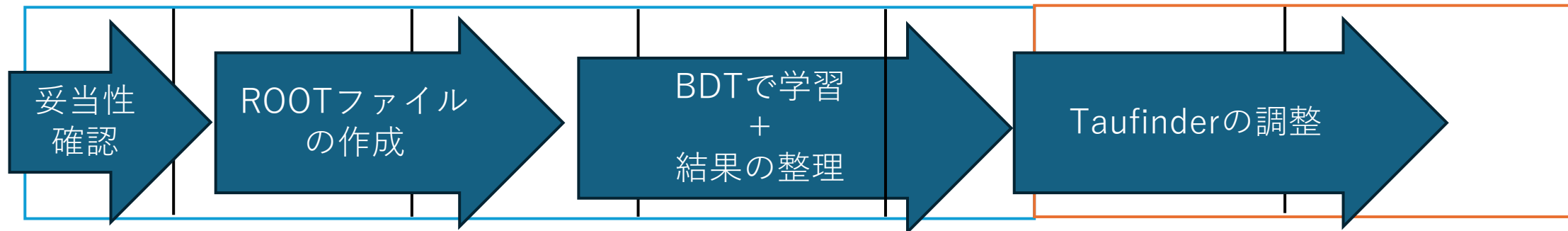
$WW \rightarrow lvq\bar{q}$ の4fWWsemileptonic、4fWsemileptonic、4fmixleptonic、の3つについて、モンテカルロからtauの情報を持ってこないといけない？

モンテカルロから1事象の中のtauの個数は持ってきているけど、これだけだと、tau自体が悪さをしているのか、isoleptonfinderに引っ掛からなかったものなのかわからないので、問題の切り分けが出来ないかも？

2026/10/19
末原先生より

10月

11月



バックアップ

precut

- Missingmass<140GeV,
- 100GeV<dijetmass<200GeV,
- 5GeV<Tmomentum<100GeV,
- P_{max} <80GeV,
- $9 < N_{charged}$,
- isolepton=0.

File Name	Original	After Cut	Efficiency
4fWWsemilep	9.88722e+06	3.50232e+06	35.42 %
4fWsl	5.404e+06	218446	4.04 %
4fZZsemilep	321101	132905	41.39 %
4fZnusemilep	238962	96582.4	40.42 %
4fmixlep	861910	36533.5	4.24 %
htautau	4241.26	341.019	8.04 %
hww	14428.3	7099.68	49.21 %
hzz	1771.67	947.553	53.48 %
nnbb	38721.6	36201	93.49 %
nncc	1952.86	1907.32	97.67 %
nngg	5751.2	5687.56	98.89 %
Total Background (B)	1.67801e+07	4.03897e+06	24.07 %
Signal (nnss) (S)	16.5087	16.1884	98.06 %

Original Significance : 0.004

Pre-cut Significance : 0.0081

重み付けの結果を先輩と比べて

表 5.6: ニュートリノチャンネル解析に用いたデータセット

プロセス	$P = (-100\%, +100\%)$		$P = (+100\%, -100\%)$	
	Generated	CrossSection[fb]	Generated	CrossSection[fb]
$ZH \rightarrow \nu\bar{\nu}s\bar{s}$	199474	0.025492	199472	0.012878
$ZH \rightarrow \nu\bar{\nu}b\bar{b}$	198888	74.1827	198930	37.4762
$ZH \rightarrow \nu\bar{\nu}c\bar{c}$	199532	13.68364	199582	1.86093
$ZH \rightarrow \nu\bar{\nu}gg$	199611	10.4391	199646	5.27371
$ZH \rightarrow \nu\bar{\nu}\tau^+\tau^-$	821053	7.99184	821765	4.03738
$ZH \rightarrow \nu\bar{\nu}WW^*$	632822	27.2385	632719	13.7606
$ZH \rightarrow \nu\bar{\nu}ZZ^*$	857587	3.33949	857685	1.68707
$WW/ZZ \rightarrow \nu q\bar{q}$	3.38×10^6	1637.06	145182	55.4225
$WW \rightarrow \nu q\bar{q}$	4.38×10^7	18779.1	464214	173.468
$ZZ \rightarrow q\bar{q}l^+l^-/q\bar{q}\nu\bar{\nu}$	3.18×10^6	609.879	1.45×10^6	261.567
$\nu\bar{\nu}Z \rightarrow \nu\bar{\nu}q\bar{q}$	2.39×10^6	453.87	717198	131.22

プロセス	ファイル名	生イベント数	重み付けされたイベント数
$ZH \rightarrow \nu\bar{\nu}s\bar{s}$	nnss	200000	16.5
$ZH \rightarrow \nu\bar{\nu}b\bar{b}$	nnbb	200000	38721.6
$ZH \rightarrow \nu\bar{\nu}c\bar{c}$	nncc	200000	1952.9
$ZH \rightarrow \nu\bar{\nu}gg$	nngg	200000	5751.2
$ZH \rightarrow \nu\bar{\nu}ZZ^*$	hzz	1000000	1771.7
$ZH \rightarrow \nu\bar{\nu}WW^*$	hww	1000000	14428.3
$ZH \rightarrow \nu\bar{\nu}\tau^+\tau^-$	htautau	1000000	4241.3
$ZZ \rightarrow q\bar{q}l^+l^-/q\bar{q}\nu\bar{\nu}$	4fZZsemileptonic	3099600	321101.2
$\nu\bar{\nu}Z \rightarrow \nu\bar{\nu}q\bar{q}$	4fZnusemileptonic	2298400	238962.4
$WW \rightarrow \nu q\bar{q}$	4fWWsemileptonic	36086800	9887219.3
$WW \rightarrow \nu q\bar{q}$	4fWsemileptonic	45357600	5404003.8
$WW/ZZ \rightarrow \nu q\bar{q}$	4fmixleptonic	8400000	861910.4

自分の重み付けの結果が清野さんの結果より明らかに大きくなっているように見られるが、自分のものはルミノシティの分の900をかけているのでそうになっているだけで問題なし