Методология применения глубоких нейронных сетей в поиске "новой физики" на коллайдерах и статистическая интерпретация ожидаемых результатов

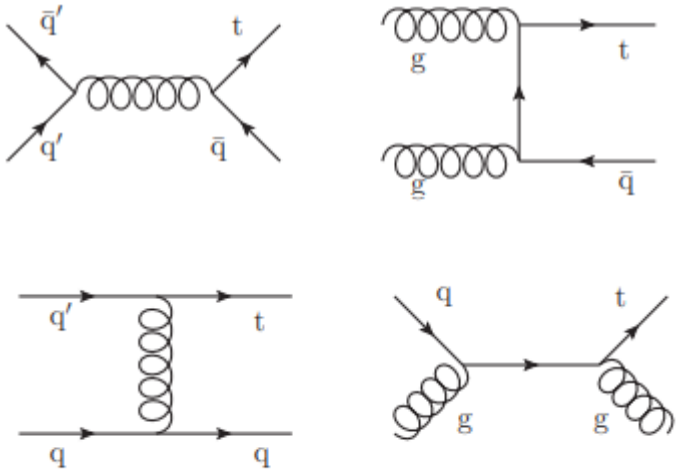
Э.Э. Абасов, Г.А. Воротников, П.В. Волков, Л.В. Дудко, М.А. Перфилов, А.Д. Заборенко,
Е.С Сивакова, М.И. Белоброва
(НИИЯФ МГУ)

План доклада

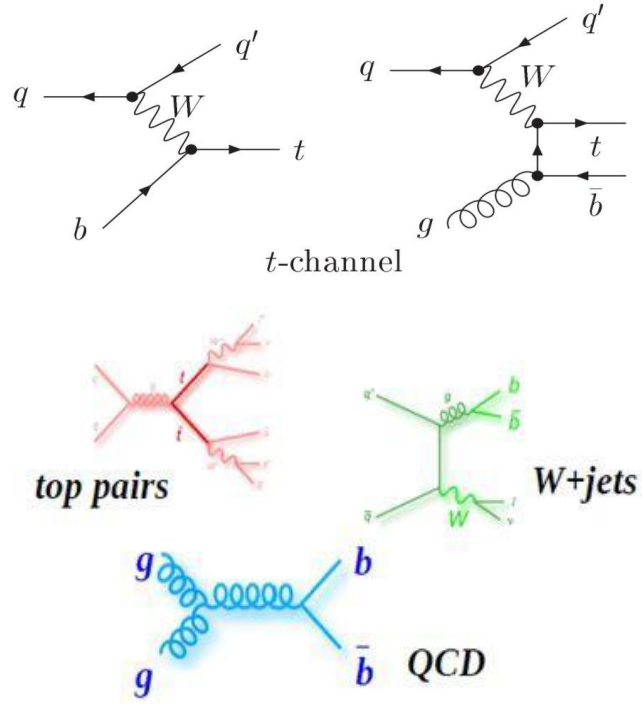
- Физическая задача
- Автоматизация отбора входных переменных
- Применение нейронных сетей
  - Оптимизация гиперпараметров глубокой нейросети
  - Каскады глубоких нейросетей
- Статистический анализ в пакете theta
  - Построение моделей
  - Добавление систематических неопределенностей
  - Байесовский метод подсчета
- Результаты (на примере одиночного рождения t-кварка)

Физическая задача: поиск FCNC

Основной задачей данного анализа является постановка ограничений на бренчинги tug и tcg взаимодействий как проявления «Новой физики» по данным эксперимента CMS(LHC).



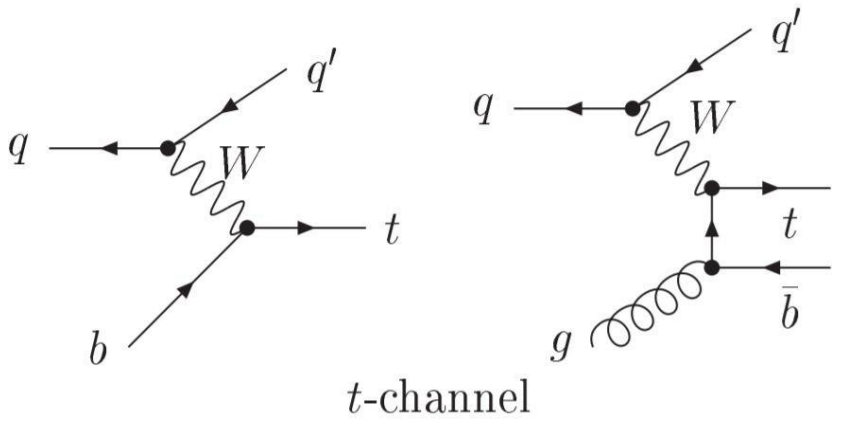
Фейнмановские диаграммы FCNC tqg процессов



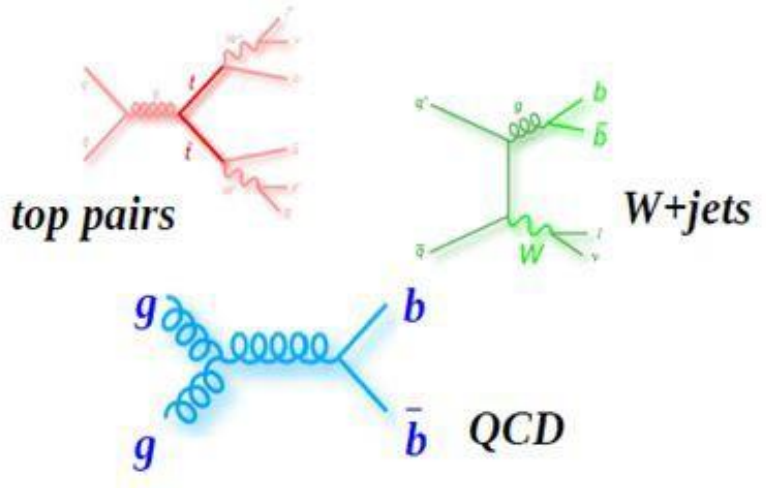
Фейнмановские диаграммы фоновых процессов

Физическая задача: бенчмарк

В текущем анализе группы в качестве бенчмарка используется измерение сечения одиночного рождения топ-кварка.



Фейнмановские диаграммы одиночного рождения топ-кварка



Фейнмановские диаграммы фоновых процессов

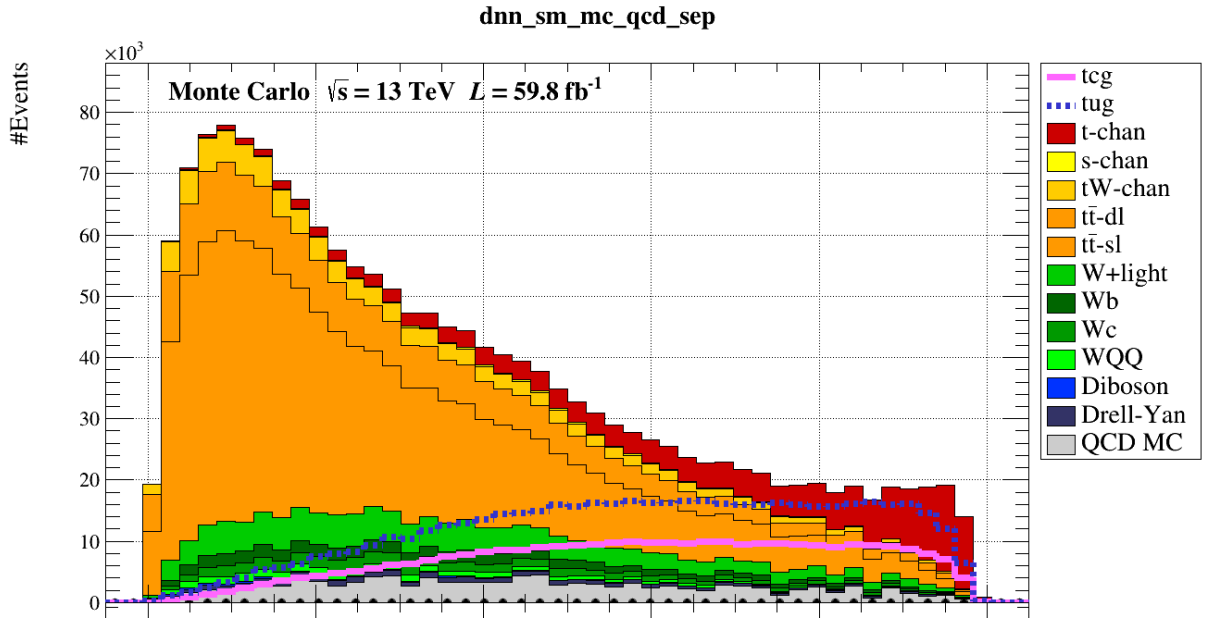
Физическая задача: подавление и разделение процессов

В текущем анализе с помощью глубоких нейросетей (DNN) решаются две основные задачи:

1. Подавление фона многоструйных КХД событий.
2. Разделение процессов. Анализ форм распределений выхода DNN.

Нейронная сеть тренируется так, чтобы выход DNN был в интервале [0,1], и распределение сигнальных событий группировалось ближе к 1, а событий фона -- к 0.

Используя обрезание по дискриминатору, можно получить набор событий с высокой долей сигнальных событий. Анализируя форму распределений, можно измерить сечения разных процессов.



Пример распределения событий для одной из DNN в физическом анализе

Физическая задача с точки зрения машинного обучения

- Датасет из нескольких десятков числовых переменных
- Каждое событие в датасете имеет свой вес, обусловленный физикой исследуемых процессов
- В задаче важна эффективность классификации, а не время вычислений
- В датасете имеется достаточное количество событий для классификации

	log(MtW)	log(MET)	log(Pt_Lep)	DPhi_LepNu	DPhi_WNu	J1_BTag	J2_BTag	LJ_BTag	J1_GTag	J2_GTag	...	log(S_LepJ2)	log(S_LepBJ1)	log(S_LepLJ)	log(S_NuJ1)	log(S_NuJ2)	log(S_NuBJ1)	log(S_NuLJ)	log(S_J1J2)	log(S_LJTop)	label
0	-0.158017	0.471724	1.424701	-1.097130	-0.678345	1.126241	-0.729301	0.273487	-0.634198	1.054127	...	-0.373750	2.240603	-0.676210	0.762591	0.429506	1.239701	0.072749	1.451033	1.136899	1.0
1	0.907071	1.969453	2.026695	-0.835542	-0.692116	0.731135	-1.002372	-0.482982	0.311643	-0.385726	...	1.606672	1.845200	1.097137	1.475462	1.726856	1.985753	1.280465	1.076546	1.662873	1.0
2	-0.700078	-3.498501	1.593446	-0.267253	0.586437	-1.064300	1.226518	-0.508284	0.434454	-0.734026	...	0.961636	0.992050	0.872574	-1.465286	-2.498035	-2.639637	-1.285374	0.305218	-0.143560	1.0
3	-0.123182	-0.294286	-1.153883	-0.220428	-0.304978	0.821729	-0.974507	-0.405791	-0.169487	-0.253373	...	-1.454158	0.462707	-1.643650	-0.811467	0.751113	-0.407621	0.372136	0.117872	0.212915	1.0
4	0.179653	-0.112133	-0.299896	-0.149657	-0.128932	-0.563185	-0.814614	0.037152	0.950013	1.516735	...	0.523618	-0.014386	0.127328	0.012216	-1.288597	-0.342761	-1.526650	0.498253	-0.871705	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
182816	-1.748436	-0.260889	-0.977265	-1.402976	-0.953263	-1.055078	1.219687	-0.473089	-0.333425	-0.708669	...	-1.040228	-1.157145	0.552928	1.622404	1.325665	1.387337	1.576067	2.650428	1.274662	0.0
182817	-0.108051	-1.467500	0.093458	0.079289	0.664688	1.145743	-0.989863	-0.445685	-0.696072	-0.149611	...	0.111343	1.032728	-0.244086	0.037291	0.823232	0.483152	0.438850	0.678071	0.581852	0.0
182818	0.638017	-0.137699	-0.259352	1.486933	1.281930	-1.030288	1.013197	-0.403290	-0.118881	-0.116235	...	-1.644836	-1.806644	0.483828	-0.082090	-0.239609	-0.256398	-0.009695	1.231487	-0.298308	0.0
182819	-1.371875	0.690908	-1.158380	-1.384540	-1.003338	-1.030251	1.100544	-0.403185	1.152087	-0.369942	...	-0.116359	-0.164680	-1.085661	-0.044813	0.361127	0.374450	0.024986	-1.197866	-0.718219	0.0
182820	-1.937246	-2.537960	-1.094366	-1.161556	-0.597909	1.144940	-0.972643	-0.397793	-0.693653	-0.250767	...	1.358322	1.823278	0.873849	-2.053396	-1.883606	-1.687182	-2.108618	-0.598085	0.178820	0.0

182821 rows × 72 columns

Выбор входных переменных

Ранее группой был предложен универсальный метод для формирования списка наблюдаемых для нейронных сетей в задачах физики высоких энергий на основе анализа фейнмановских диаграмм. [International Journal of Modern Physics A Vol. 35, No. 21 (2020) 2050119]

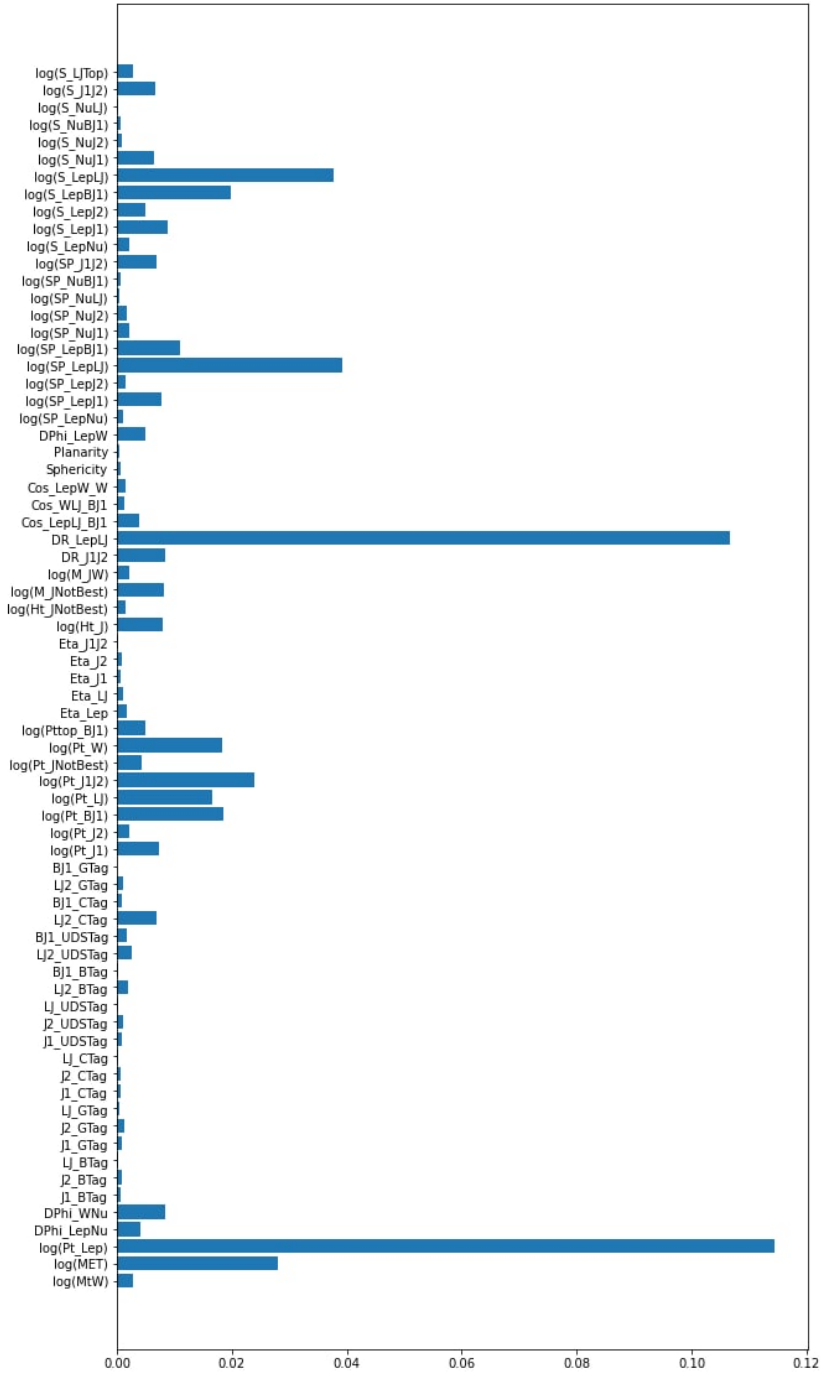
Входные переменные следует нормализовать и применить к некоторым логарифмическую трансформацию.

Следующим этапом подготовки пространства входных переменных может стать метод permutation_importance, реализованный в пакете sklearn. Этот метод можно использовать в работе с любыми табличными датасетами и предсказательными моделями.

Выбор входных переменных

В рамках этого метода каждая из переменных “зашумляется”, а затем результат работы модели сравнивается с эталонным. Чем сильнее разница в результатах для отдельно взятой переменной, тем важнее она для текущей классификации.

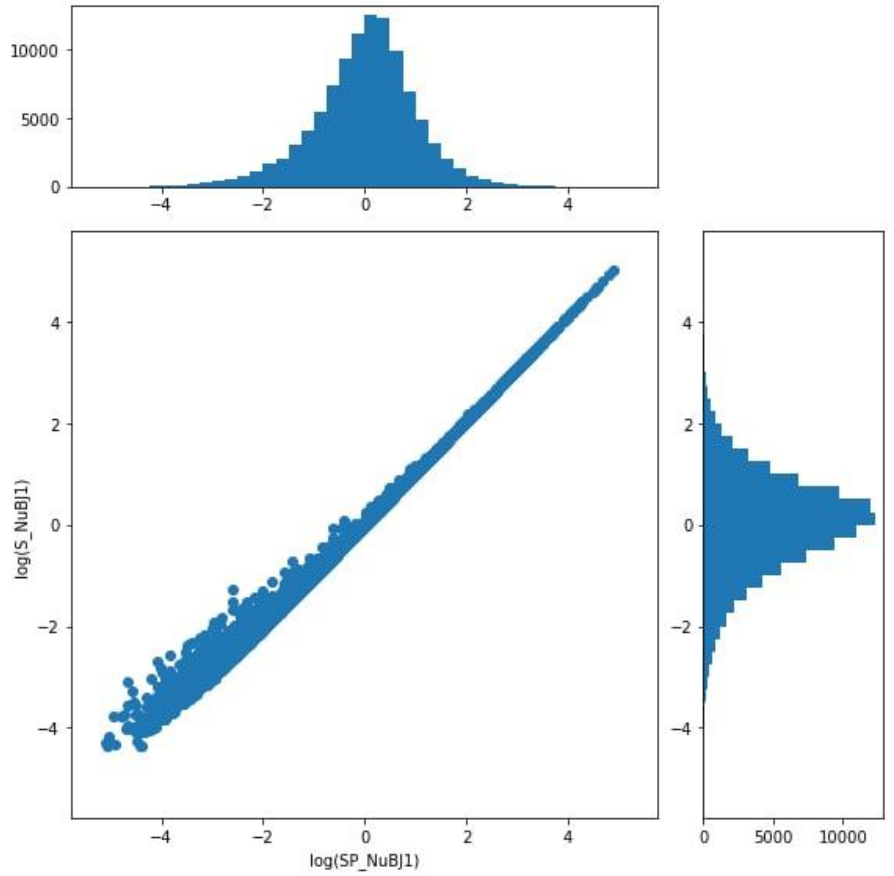
Затем данная операция повторяется несколько раз для всего датасета для уменьшения случайных флуктуаций.



Выбор входных переменных: корреляция

Минус данного метода в том, что если пара переменных скоррелирована и варьируется только одна, то модель может извлечь информацию из незашумленной переменной, уменьшая значимость и той, и другой переменной.

Решением такой проблемы может стать кластеризация переменных на основе их взаимной корреляции.



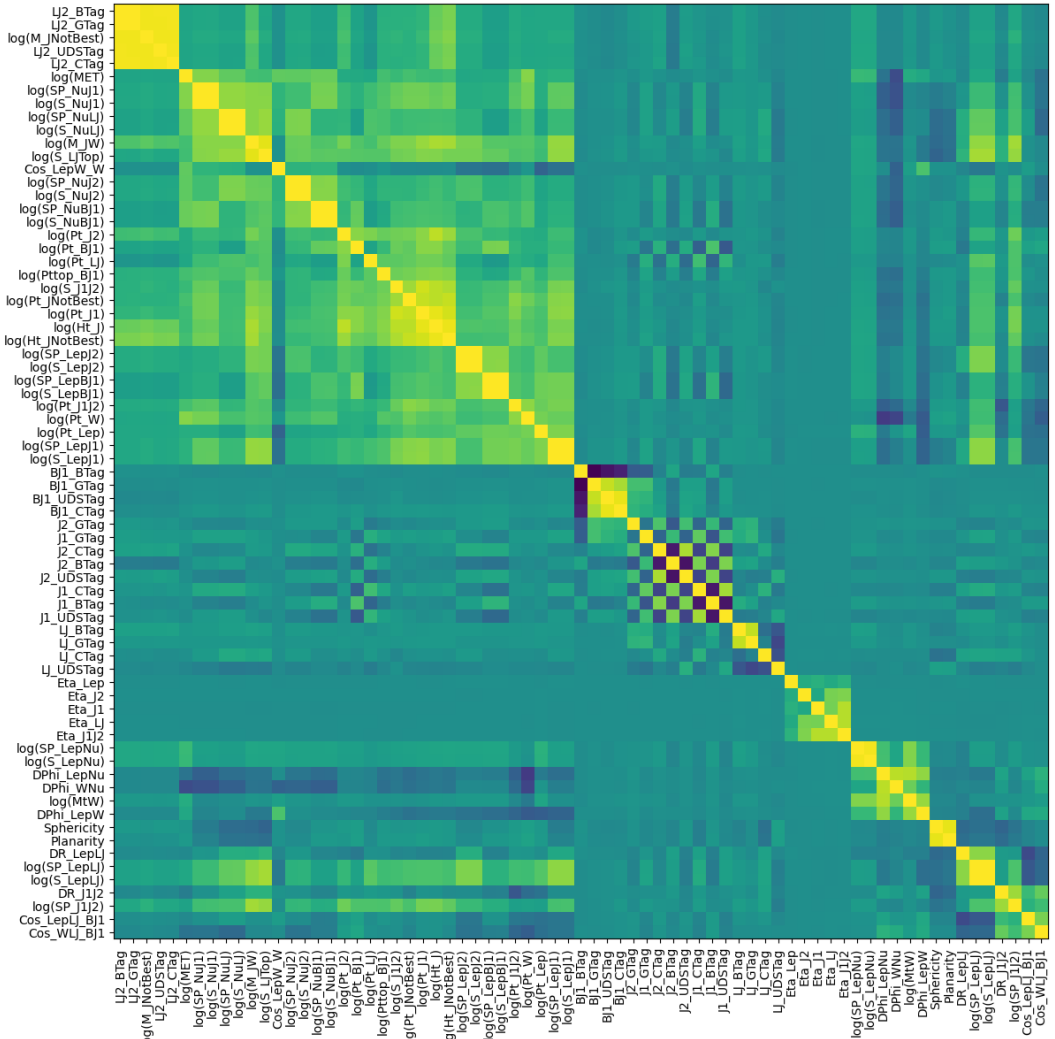
Пример двух сильно скоррелированных переменных. Для классификации выбирается только одна переменная.

Отбор переменных: кластеризация

Корреляционная матрица для всех элементов датасета (справа) трансформируется в матрицу расстояний.

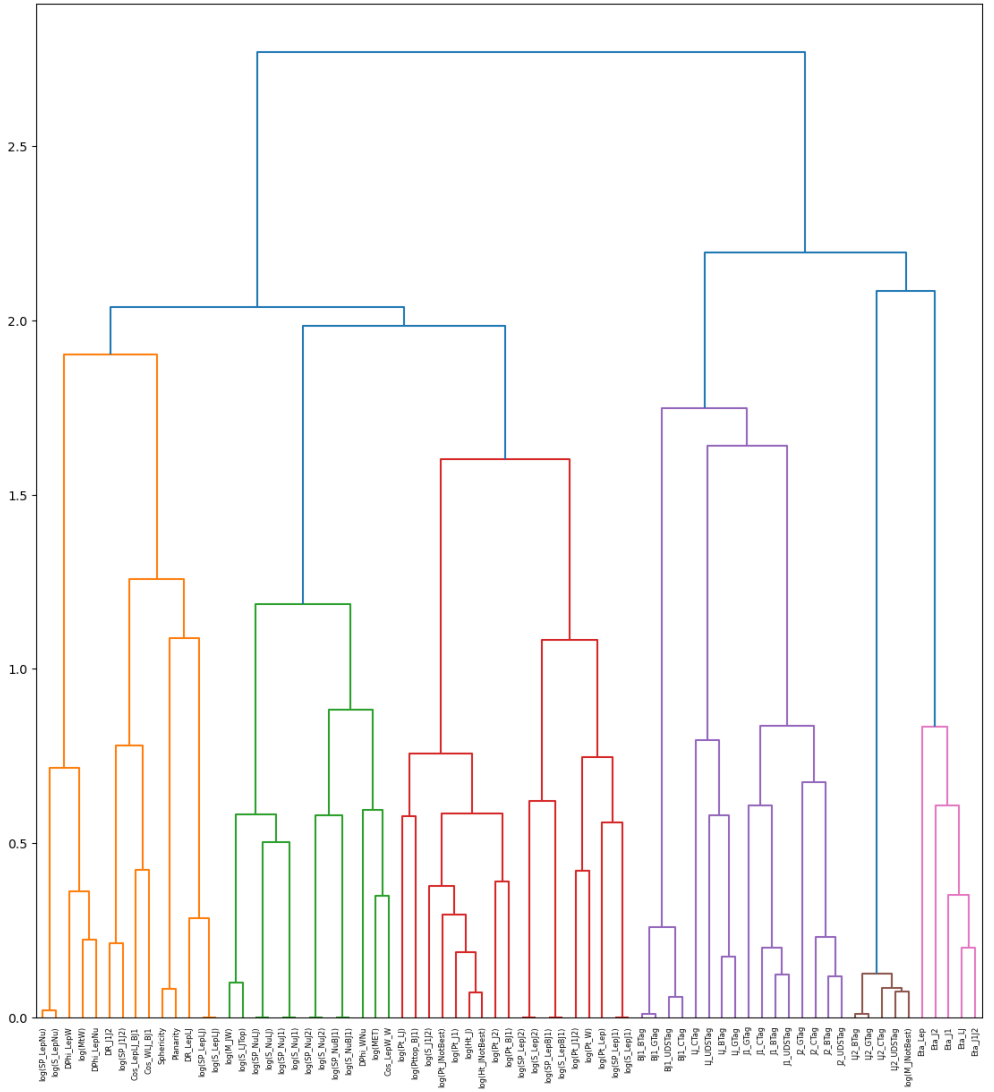
Затем эта матрица визуализируется в виде дендрограммы, в которой переменные формируют кластеры в зависимости от их взаимной корреляции.

Более подробно эти трансформации описаны в документации sklearn.

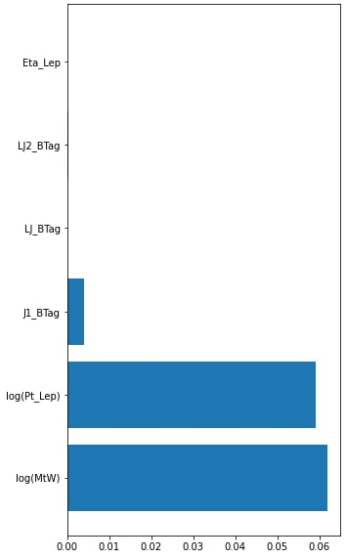


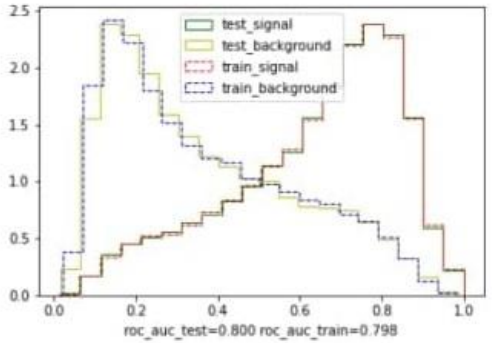
Отбор переменных: обрезание по дендрограмме

Различные обрезания по дендрограмме выделяют кластеры переменных, из которых выбирается одна. После этих обрезаний можно провести еще один этап permutation feature importance анализа для получения степени важности уже менее скоррелированных переменных.

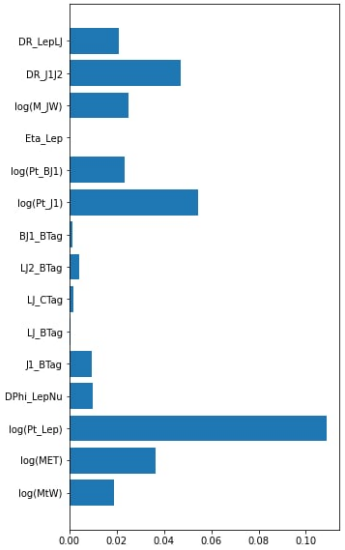


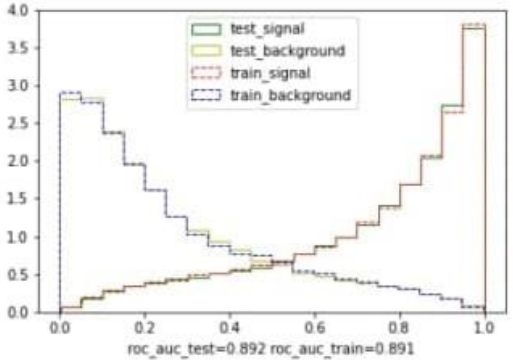
Отбор переменных: построение моделей



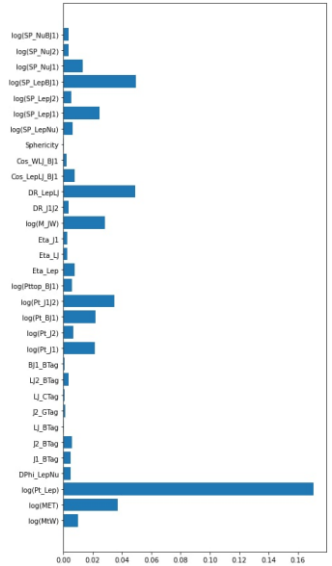


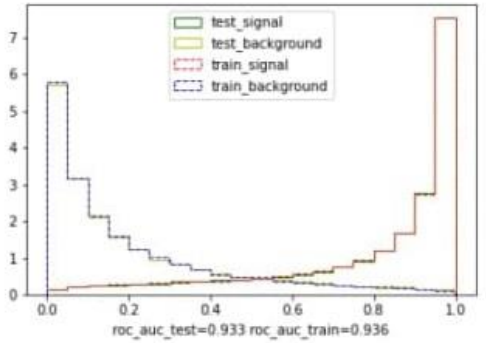
Обрезание степени корреляции 2.0



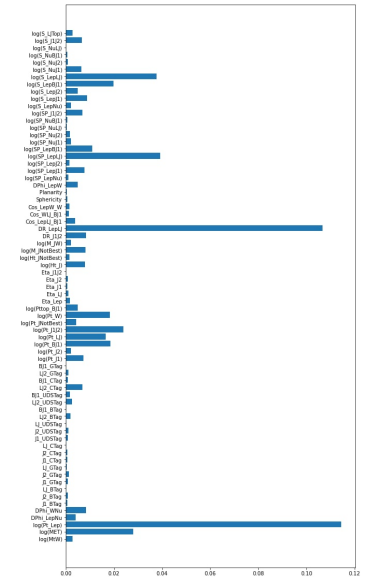


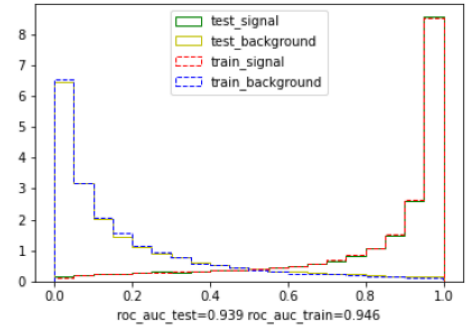
Обрезание 1.0





Обрезание 0.5





Без обрезания

Отбор переменных: итоги

Составление оптимального набора переменных для задачи машинного обучения может быть нетривиальной задачей, требующей глубокого понимания датасета. Используемые методы permutation feature importance и кластеризация по корреляции могут помочь исследователю определить, какие переменные являются наиболее информативными для текущей модели.

Именно на эти входные переменные нужно обращать внимание в первую очередь в процессе оптимизации. В текущем анализе описанные методы были применены для формирования конечного набора наблюдаемых.

Пространство гиперпараметров тренировки глубокой сети

Веса глубоких сетей меняются в процессе тренировки, но на эффективность модели влияют и другие параметры, задающие ее конфигурацию. Эти параметры не меняются в процессе градиентного спуска и называются гиперпараметрами.

Физическая задача позволяет использовать полносвязные DNN с прямым распространением сигнала, их основные гиперпараметры следующие:

- Число скрытых слоев
- Ширина скрытых слоев
- Величина дропаута
- Функция активации
- Learning rate
- Константа регуляризации

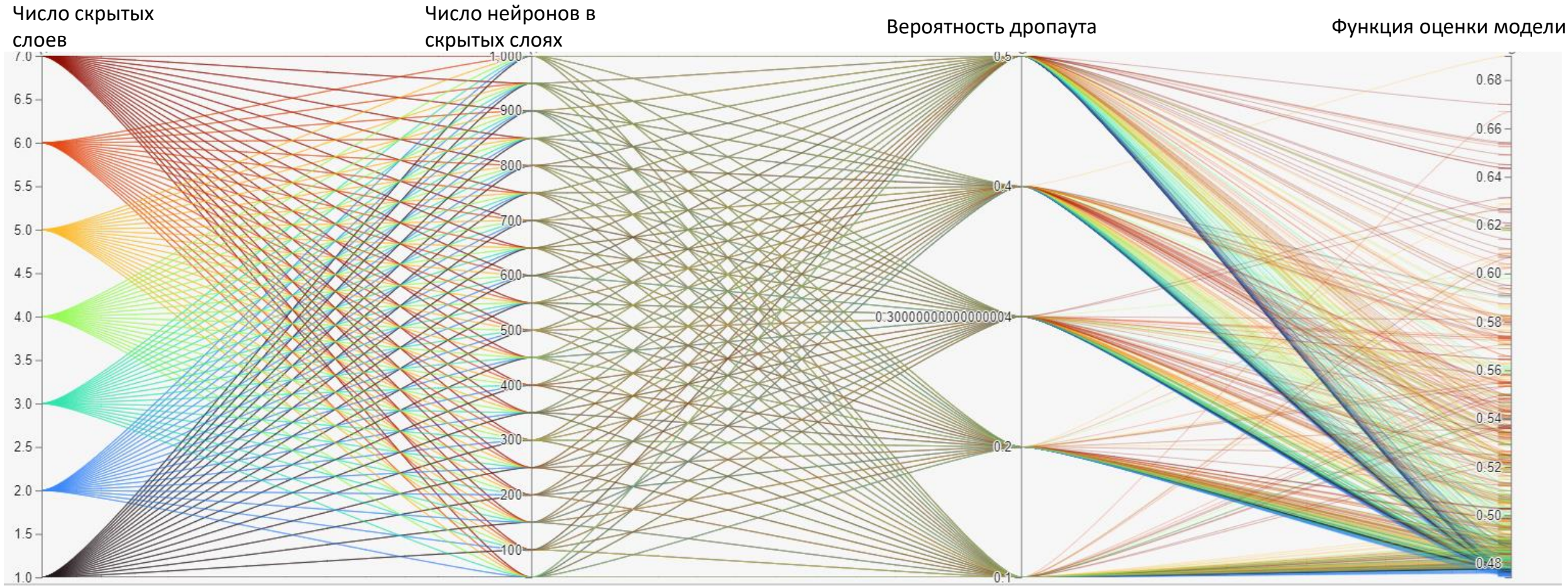
Для контроля переобучения для всех конфигураций используется критерий ранней остановки: сравнение функции ошибки для тренировочного и тестового набора данных.

Для оценки эффективности сравнивается функция ошибки (score), форма распределений выхода сети, ROC AUC.

Оптимизация в пространстве гиперпараметров

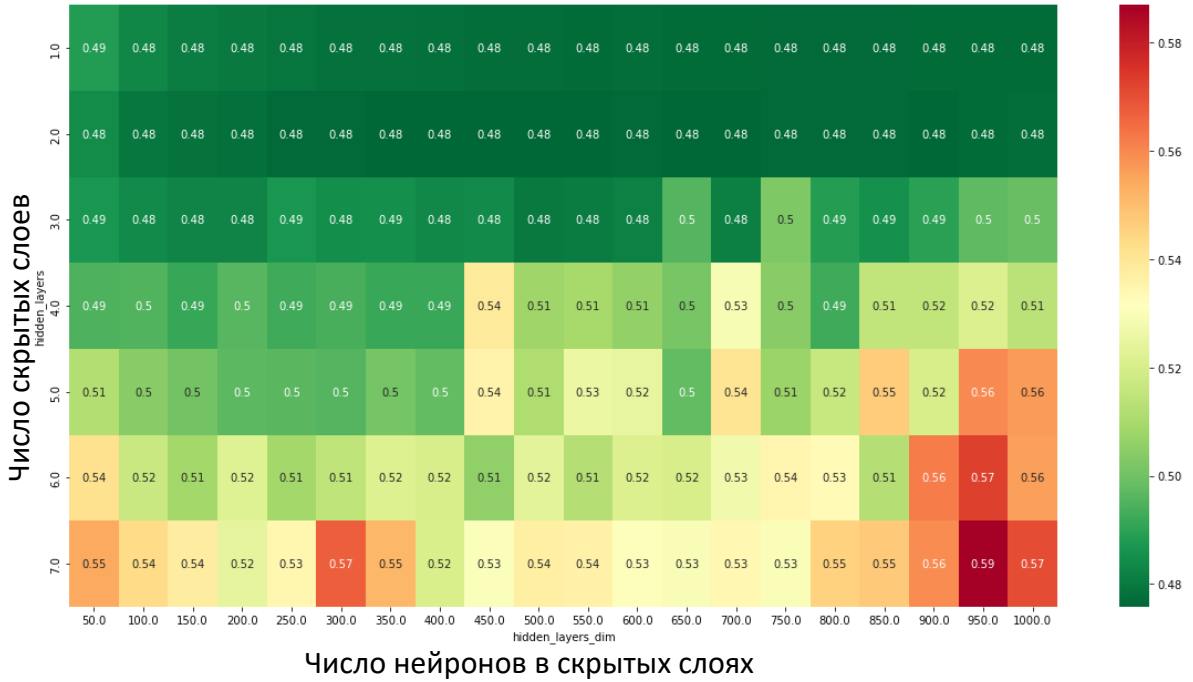
- Grid Search: перебор всех комбинаций подряд, ресурсоемко
- Random Search: случайный перебор комбинаций гиперпараметров, адекватная комбинация находится значительно быстрее, чем перебором
- Hyperband: тренировка различных комбинаций параметров идет небольшое количество эпох, затем результаты после этих эпох сравниваются, и дальнейшее обучение проходят только лучшие комбинации
- Генетические алгоритмы
- Bayesian Optimization: быстрая оптимизация к локальному минимуму
- В принципе, любой оптимизатор на основе машинного обучения

Результаты исследований гиперпространства

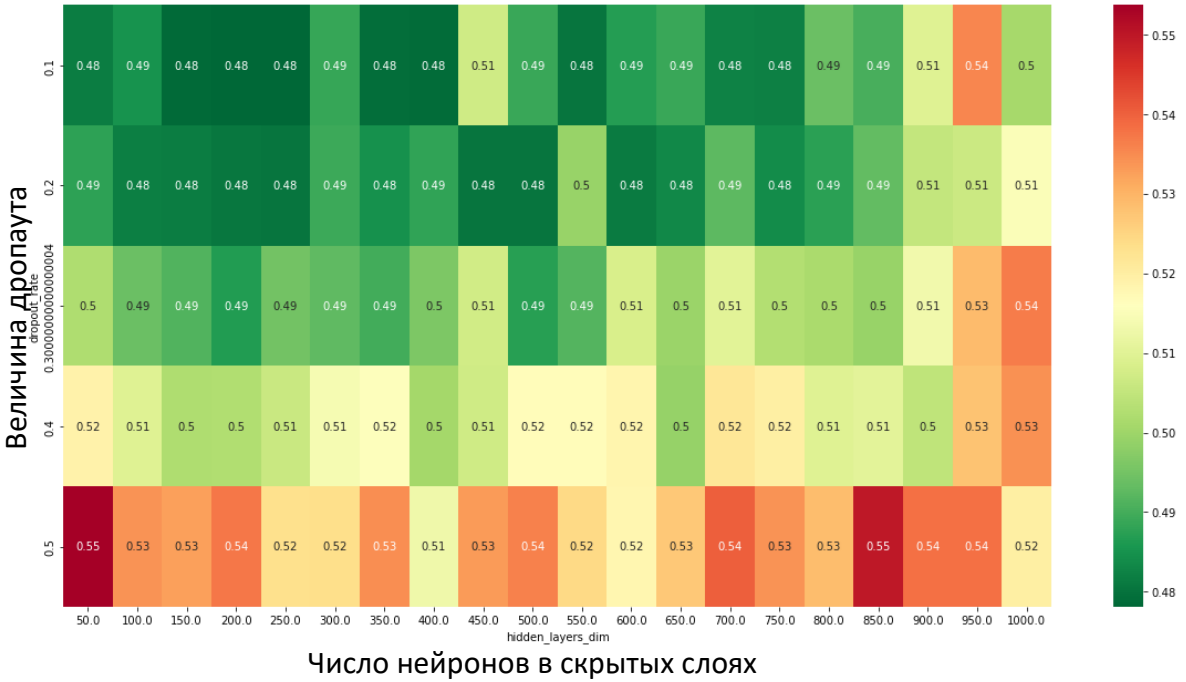


Одновременная визуализация трех переменных гиперпространства. Большинство конфигураций показывают стабильную работу: 50% всех комбинаций лежат на расстоянии в 2% от лучшей.

Результаты исследований гиперпространства



Число нейронов в скрытых слоях

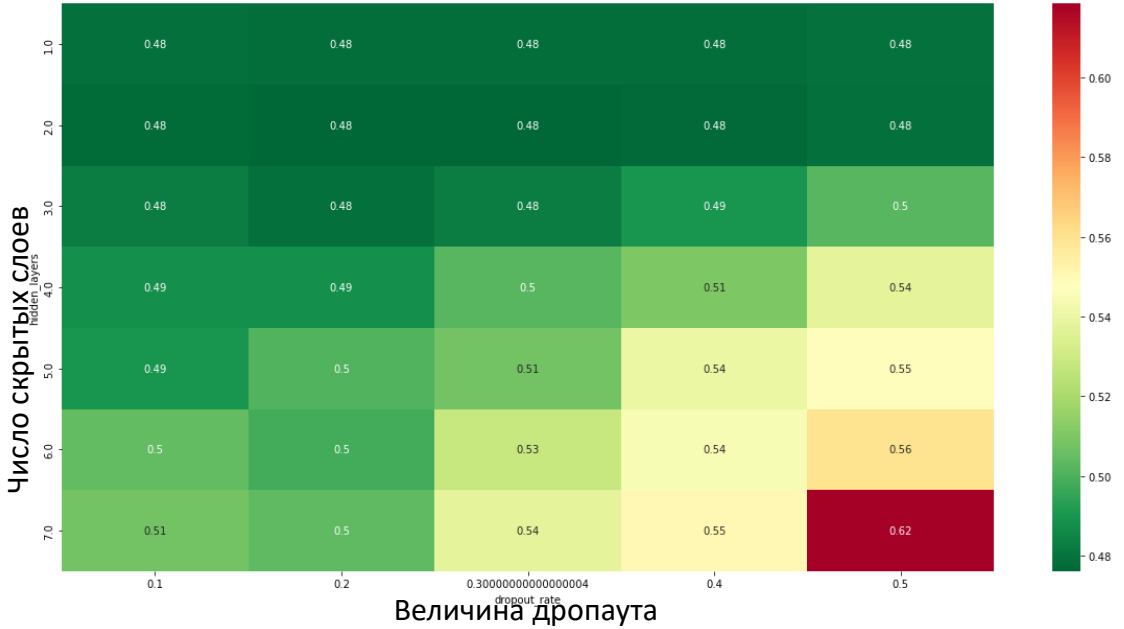


Число нейронов в скрытых слоях

Визуализация двумерных зависимостей в пространстве гиперпараметров. Цветом показана средняя разделяющая способность моделей с текущей конфигурацией -- функция бинарной кроссэнтропии (чем меньше, тем лучше).

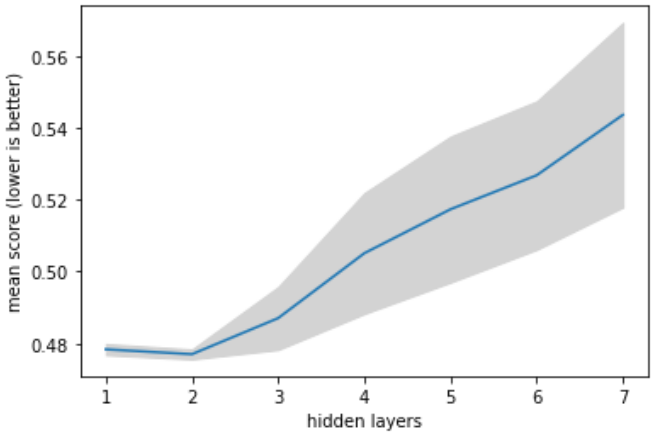
Существует широкая область стабильности для 1-2 слоев и 0.1-0.3 величины дропаута.

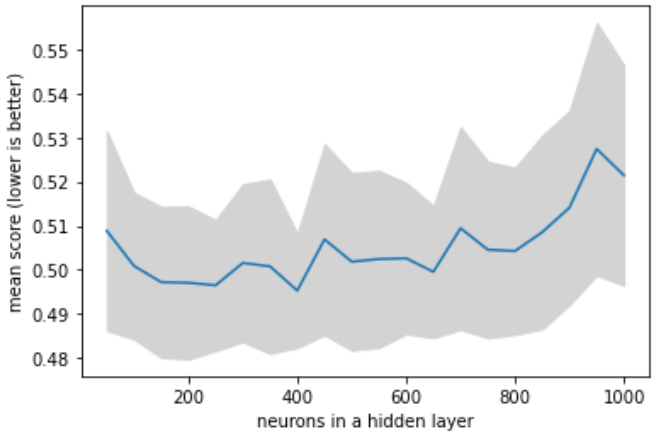
Результаты исследований гиперпространства

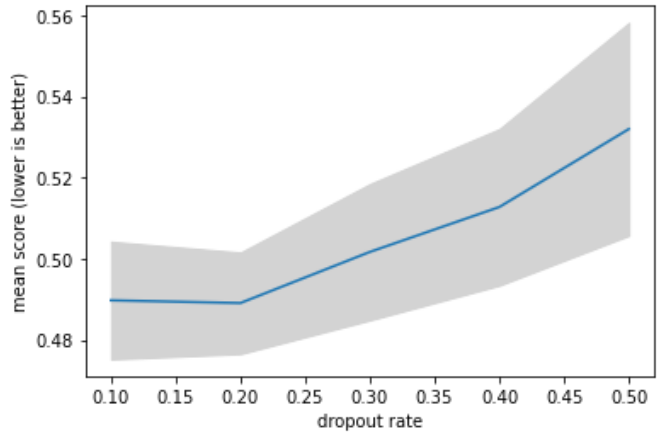


Слева: зависимость качества модели от числа скрытых слоев и вероятности дропаута. Для 1-2 слоев все величины дропаута приводят к хорошей производительности, а для более глубоких сетей следует уменьшить дропаут до 0.1-0.2.

Внизу: одномерные зависимости качества модели от параметров гиперпространства. По остальным параметрам берется среднее.







Результаты исследований гиперпространства

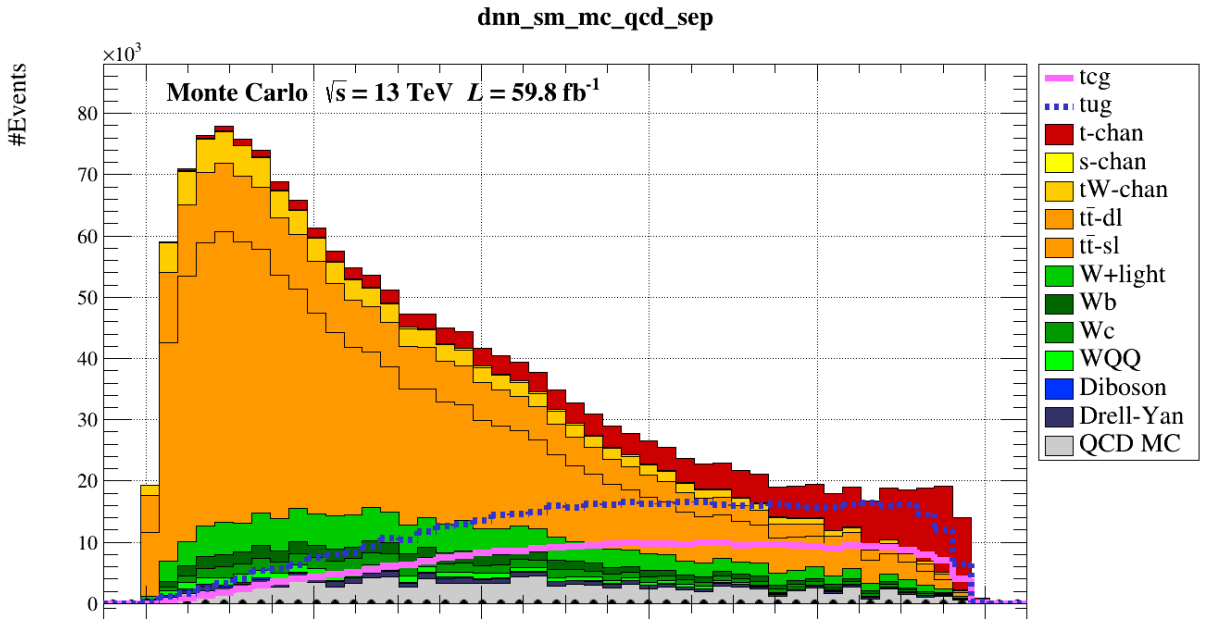
Итоги:

- Были описаны распространенные методы оптимизации в пространстве гиперпараметров
- Была проведена визуализация пространства гиперпараметров
- Общие выводы для типичной задачи анализа данных коллайдерных экспериментов:
  - модель нейронной сети устойчива ко многим комбинациям гиперпараметров (50% всех комбинаций лежат на расстоянии в 2% от лучшей)
  - для множества вариаций сетей в анализе оптимальный регион гиперпараметров похож: это 1-2 скрытых слоя, 400-800 нейронов в скрытых слоях, 0.2-0.3 вероятность дропаута

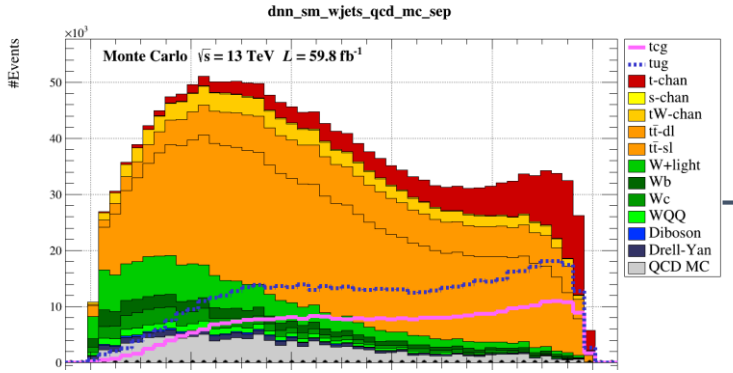
Каскады глубоких нейросетей: идея

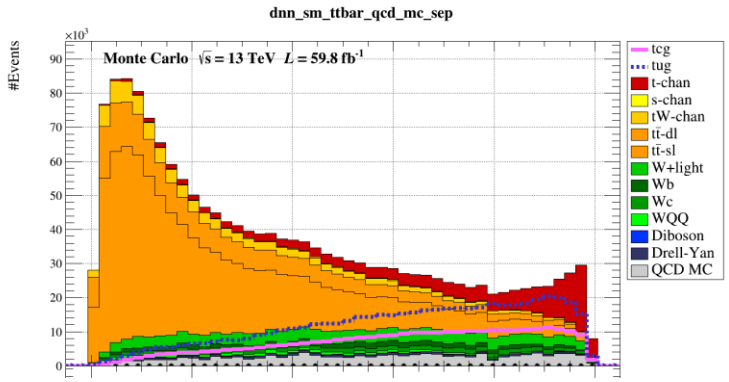
В предыдущих итерациях анализа использовалась единая сеть для отделения т-канального процесса от процессов W+jets и t-tbar.

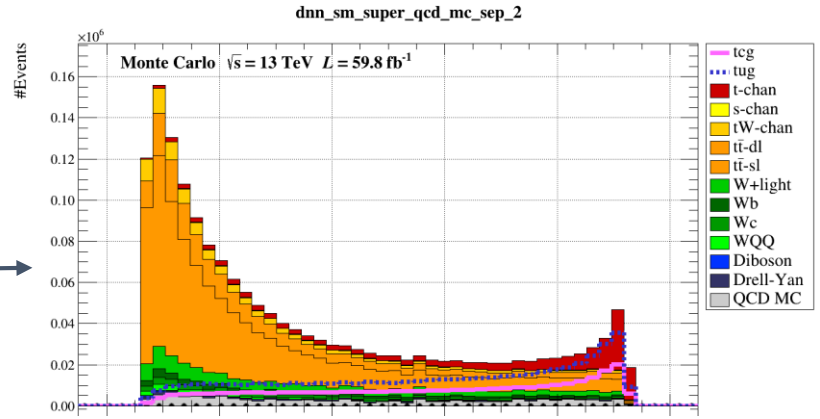
Одна из возможностей дальнейшей оптимизации -- использование каскада нескольких нейросетей: сначала тренируются отдельные сети для разделения т-канала от W+jets и т-канала от t-tbar (так как W+jets и t-tbar сильно отличаются между собой). Затем выходы этих сетей объединяются суперсетью, разделяющей т-канал от W+jets и t-tbar одновременно.

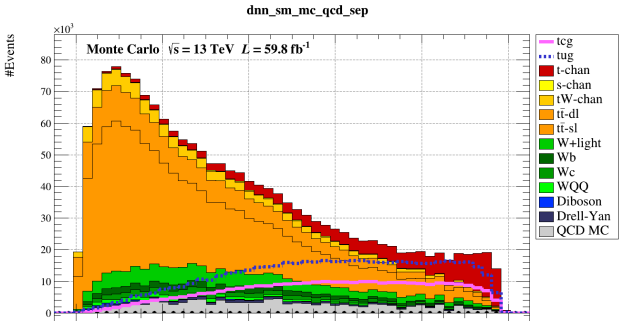


Каскады глубоких нейросетей: применение к СМ





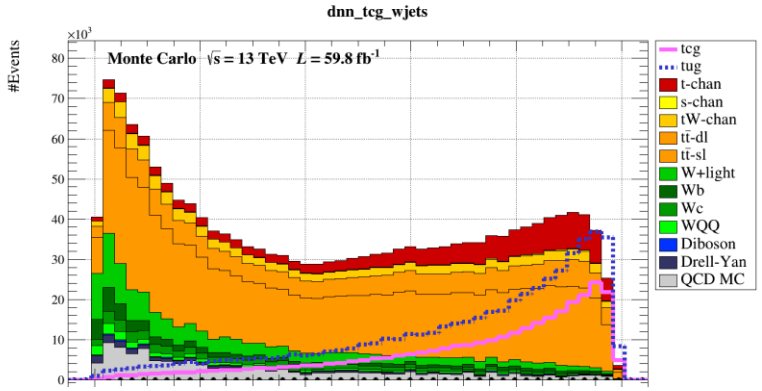




Предыдущая итерация DNN

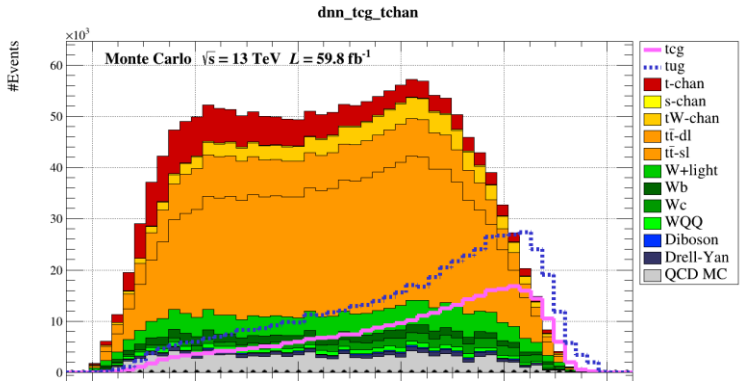
Дискриминатор полученной суперсети отличается от предыдущей СМ сети лучшим группированием сигнальных событий в сигнальной области и более плоской смешанной областью (0.4 - 0.7)

Каскады глубоких нейросетей: поиск “Новой физики”

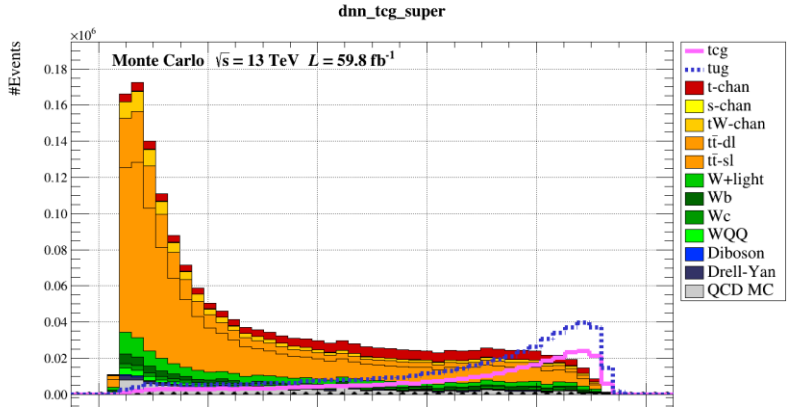


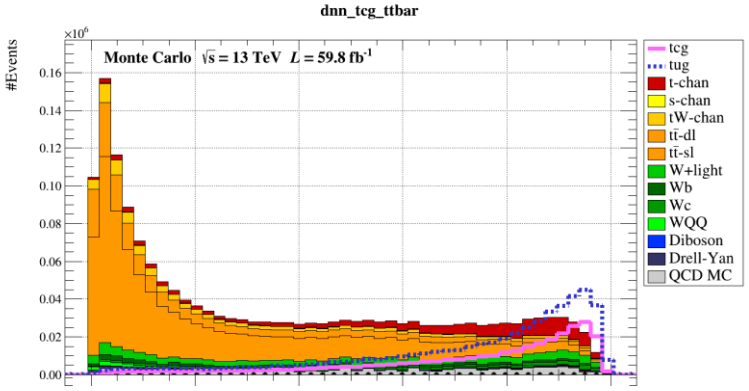
w+jets vs FCNC

+



t-chan vs FCNC





t-tbar vs FCNC

Для отделения FCNC-tug от СМ событий необходимо объединить три сети в одну:

- w+jets vs FCNC
- t-tbar vs FCNC
- t-chan vs FCNC

В результате из-за применения физических закономерностей к обучению нейронных сетей (тренировка против одного физически схожего процесса) было улучшено разделение СМ и FCNC событий.

Статистический анализ

Дальнейшим этапом является статистический анализ формы распределений полученных нейронных сетей. В текущем анализе для измерений и постановки ограничений применяется пакет theta. Основные этапы анализа с использованием данного пакета включают:

- Построение статистической модели (путем создания конфигурационного файла)
- Выбор метода анализа
- Интерпретация результатов: квантилей для частотных методов и апостериорных распределений для байесовских.

https://github.com/emil2001/MSU_top_stat

http://www-ekp.physik.uni-karlsruhe.de/~ott/theta/testing/html/index.html

Theta: Построение модели

В наиболее общем случае вектор p задает параметры модели (например, теоретические сечения процессов). Для i-ой наблюдаемой

$$m_i(\vec{p}) = \sum_{k=1}^{M_i} c_{i,k}(\vec{p}) t_{i,k}(\vec{p})$$

Здесь k – индекс бина для гистограммы i-й наблюдаемой

Вероятность получить “данные” d при условии параметров p тогда задаётся как:

$$p_m(d|\vec{p}) = \prod_{i=1}^{N} \prod_{l=1}^{b_i} \mathrm{Poisson}(d_{i,l}|m_{i,l}(\vec{p}))$$,где $$\mathrm{Poisson}(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

i – индекс наблюдаемой
l – индекс бина

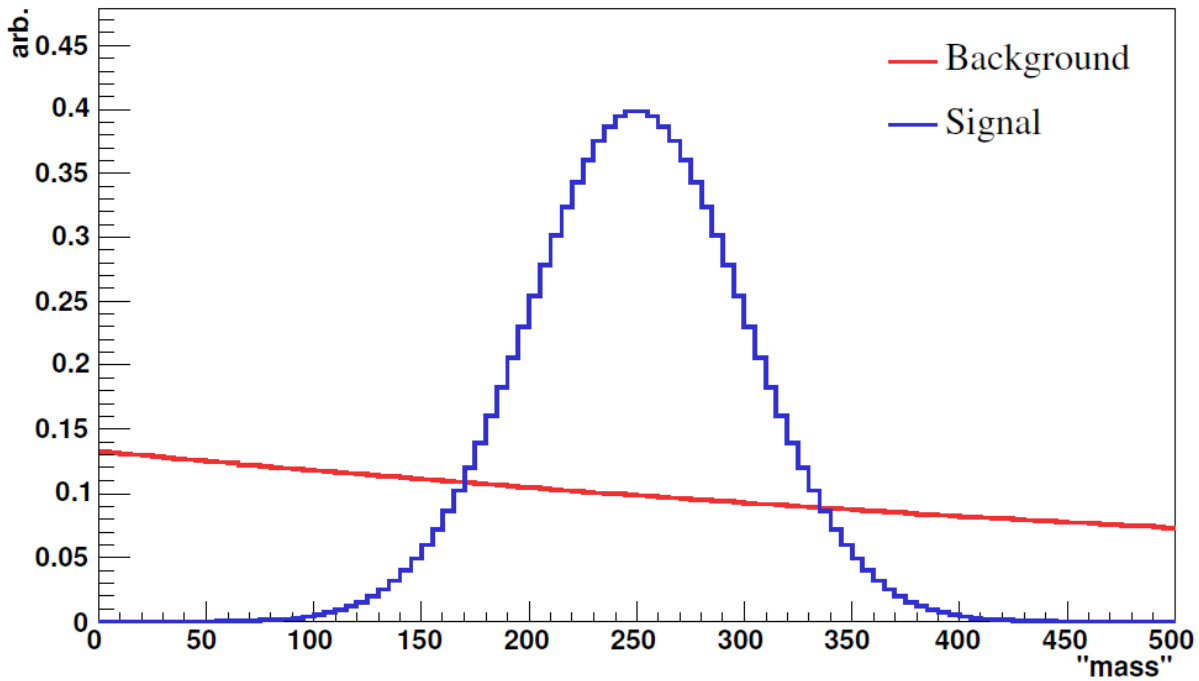
Построение функции правдоподобия:

$$L(\vec{p}|d) = p_m(d|\vec{p}) D(\vec{p}).$$

D(p) – дополнительный априорный множитель. (например, при измерениях с разными моделями, может включать полученные распределения из предыдущих измерений)

Theta: Пример простейшей модели

Здесь приведен пример модели, построенной по вышеописанному методу.
Параметры - сигнал и фон с их средним и дисперсией, других систематик нет



$$m_b(\mu_b) = \mu_b * t_b$$
$$m_s(\mu_s, \mu_b) = \mu_b * t_b + \mu_s * t_s$$

Построено две модели: только фон и сигнал+фон. Для каждой модели задается набор шаблонов формы t и нормировочные коэффициенты μ

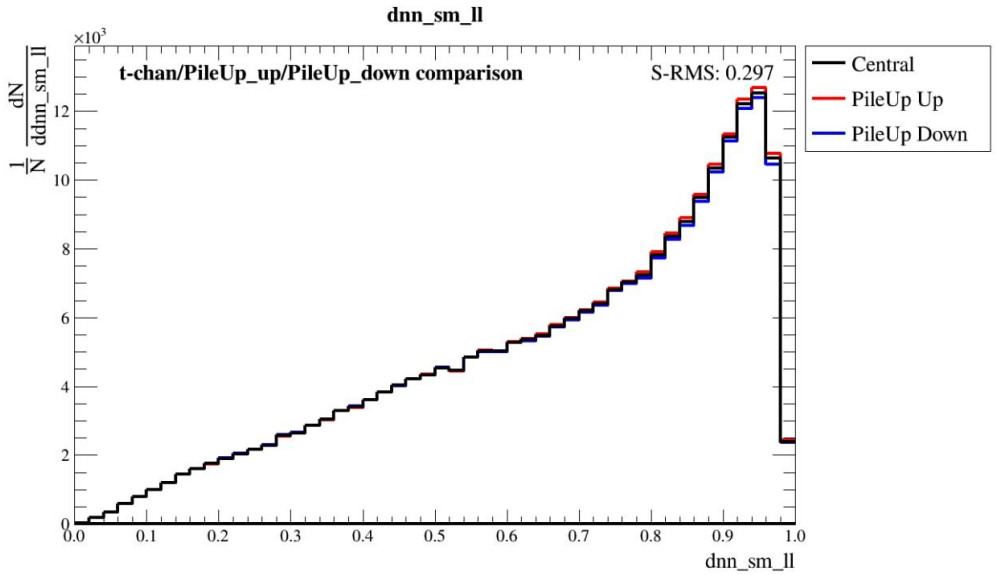
Добавление систематик. Неопределенности нормировки

Основные типы систематических неопределенностей можно разделить на неопределенности формы и неопределенности нормировки.

Для неопределенностей нормировки (например, сечений процессов) процесс добавления в модель заключается в задании нового параметра τ, влияющего на исходные функции распределения:

$$m_b(\mu_b) = \tau * \mu_b * t_b$$
$$m_s(\mu_s, \mu_b) = \mu_b * t_b + \mu_s * t_s$$

В качестве априоров для сечений σ фоновых процессов используется лог-нормальное распределение (таким образом можно избавиться от отрицательной области). Для сигнального процесса априор берется плоским.



Добавление систематик. Неопределенности формы

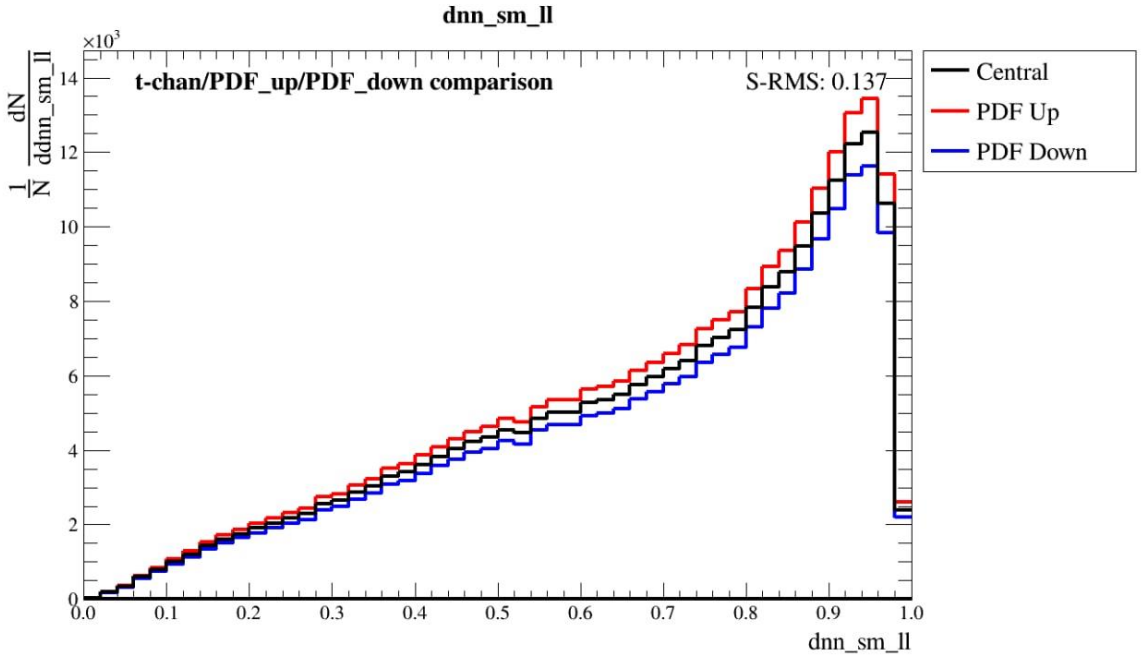
Для задания неопределенностей формы используются смещенные гистограммы, задаваемые как:

$$t(0) = t$$
$$t\,(1) = t_+$$
$$t\,(-1) = t_-$$

Здесь выполнены смещения на 1σ.

Для интерполяции результатов на произвольное значение в theta используется выражение

$$t(\delta) = t * \left(\frac{t_{sign(\delta)}}{t}\right)^{|\delta|}$$



Условия на функцию интерполяции:

- Положительные значения во всех бинах при любом δ
- Непрерывная дифференцируемость на всей числовой прямой

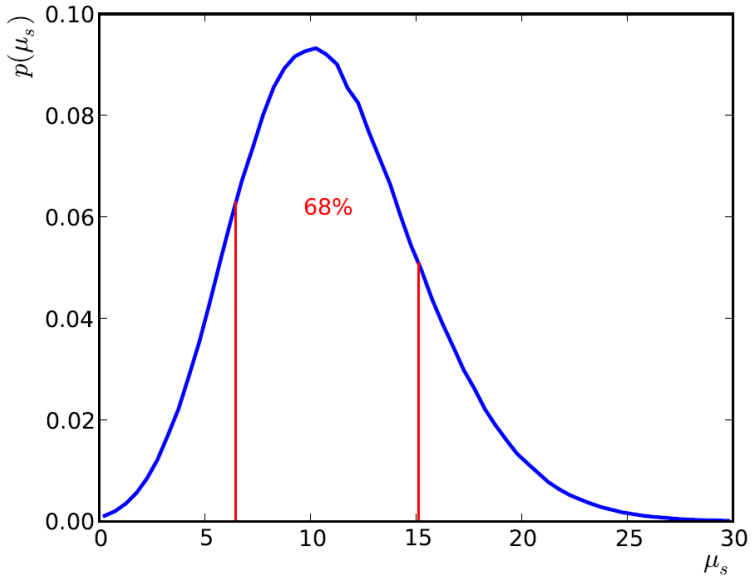
Байесовский метод подсчета

В Байесовской статистике апостериорное распределение можно считать результатом работы theta. Для всех параметров модели оно задается как

$$p(\mu_s, \mu_b|d) = p(d|\mu_s, \mu_b) * \frac{\pi(\mu_s, \mu_b)}{\pi(d)}$$ (теорема Байеса)

где π - априорные распределения. π(d) здесь является нормировочным множителем, и не обязан быть задан априорно. Для выделения распределения параметра интереса необходимо проинтегрировать по всем другим случайным параметрам:

$$p(\mu_s|d) = \int p(\mu_s, \mu_b|d) d\mu_b.$$



Алгоритм Markov-Chain Monte-Carlo (MCMC)

Для подсчета интеграла в пакете Theta реализован алгоритм Метрополиса-Гастингса. Он основан на создании цепи Маркова, то есть на каждом шаге новое выбранное значение x^{t+1} зависит только от предыдущего x^t. Используется вспомогательная функция распределения $Q(x' | x^t)$, зависящая от x^t (x' - случайное число). Затем с вероятностью

$$u = \frac{P(x^t) \cdot Q(x^t | x')}{P(x') \cdot Q(x' | x^t)}$$

выбранное значение принимается как новое, иначе оставляется старое.

Для такого алгоритма важен “холостой прогон” (burn-in) в течение нескольких первых шагов для получения независимости от стартового значения. Анализ зависимости результата от размера burn-in является одной из важнейших частей анализа с использованием метода MCMC.

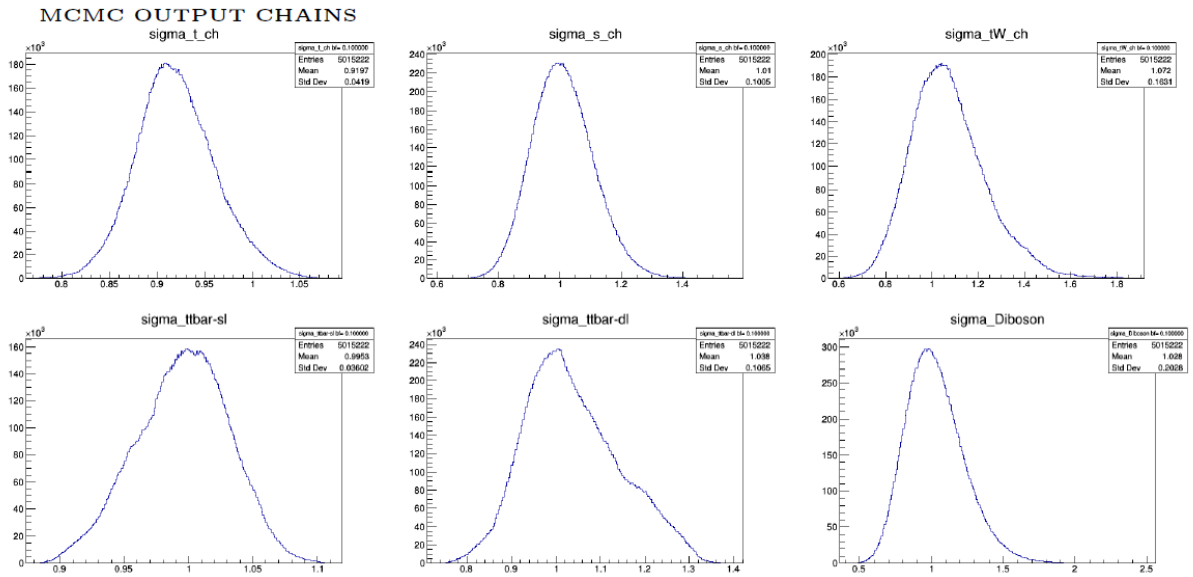
Результаты

На выходе статистического анализа требуется получение многомерной плотности вероятности, затем производится интегрирование по всем параметрам, кроме параметра интереса.
Отсюда можно вычислить:

- Среднее значение, моду, медиану, дисперсию
- Доверительные интервалы всех систематических параметров (в том числе и сигнального)

Для удобства анализа также приводятся апостериорные распределения, а также сравнение данных и моделирования

parameter	$-\sigma$	central	$+\sigma$
sigma_t_ch	0.866	0.913	0.958
sigma_s_ch	0.911	1.01	1.11
sigma_tW_ch	0.906	1.06	1.23
sigma_ttbar-sl	0.903	0.959	1.01
sigma_ttbar-dl	0.96	1.08	1.22
sigma_Diboson	0.832	1.01	1.23
sigma_DY	0.805	0.975	1.19
sigma_WQQ	0.877	1.17	1.54
sigma_Wc	0.828	1.08	1.41
sigma_Wb	1.19	1.61	2.06
sigma_Wother	0.904	1.12	1.56
sigma_Wlight	0.775	1.02	1.34
sigma_QCD	0.491	0.644	0.799
lumi	0.989	1.01	1.03



Статистический анализ нескольких сетей

В качестве альтернативы анализа выхода одной сети, натренированной на разделение t-канального рождения t-кварка (t-ch) от всех остальных процессов, проводится анализ выходов двух сетей: t-ch/ttbar и t-ch/wjets (отделяются отдельно разные по топологии фоновые процессы). Такой анализ практически не отличается от одномерного: исходное двумерное распределение проходит процедуру побинной развёртки, после чего анализируется так же, как и выход одной сети. Это возможно в силу независимости анализа от порядка бинов в theta.

parameter	$-\sigma$	central	$+\sigma$
sigma_t_ch	0.912	0.946	0.982
sigma_s_ch	0.936	1.04	1.15
sigma_tW_ch	1.02	1.16	1.34
sigma_ttbar-sl	0.936	0.969	1
sigma_ttbar-dl	1.08	1.15	1.2
sigma_Diboson	0.854	1.04	1.28
sigma_DY	0.839	1.01	1.22
sigma_WQQ	0.926	1.21	1.53
sigma_Wc	1.1	1.34	1.58
sigma_Wb	1.57	1.82	2.09
sigma_Wother	1.1	1.28	1.49
sigma_Wlight	0.631	0.802	1.03
sigma_QCD	0.635	0.698	0.76
lumi	0.98	0.992	1

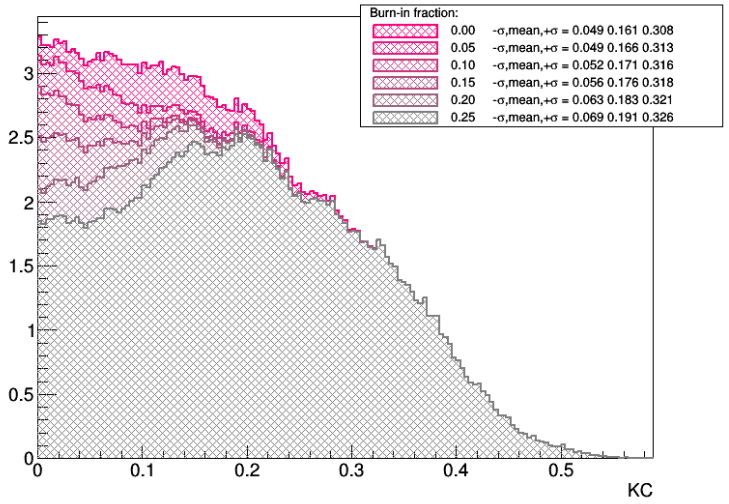
Результаты 2D-анализа

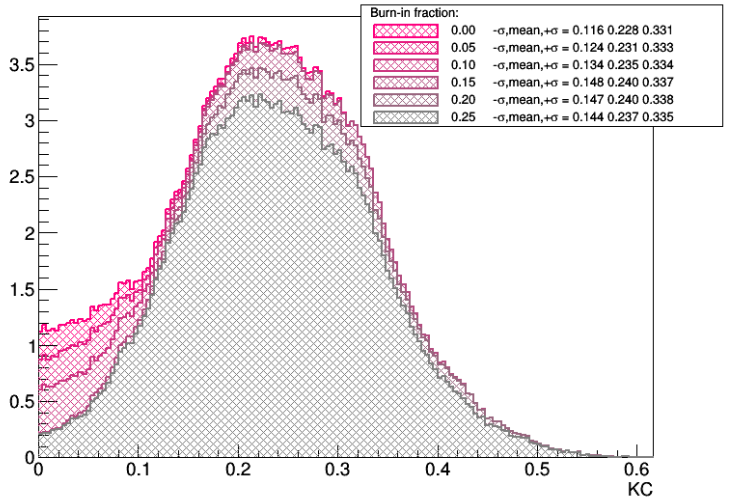
parameter	$-\sigma$	central	$+\sigma$
sigma_t_ch	0.866	0.913	0.958
sigma_s_ch	0.911	1.01	1.11
sigma_tW_ch	0.906	1.06	1.23
sigma_ttbar-sl	0.903	0.959	1.01
sigma_ttbar-dl	0.96	1.08	1.22
sigma_Diboson	0.832	1.01	1.23
sigma_DY	0.805	0.975	1.19
sigma_WQQ	0.877	1.17	1.54
sigma_Wc	0.828	1.08	1.41
sigma_Wb	1.19	1.61	2.06
sigma_Wother	0.904	1.12	1.56
sigma_Wlight	0.775	1.02	1.34
sigma_QCD	0.491	0.644	0.799
lumi	0.989	1.01	1.03

Результаты 1D-анализа

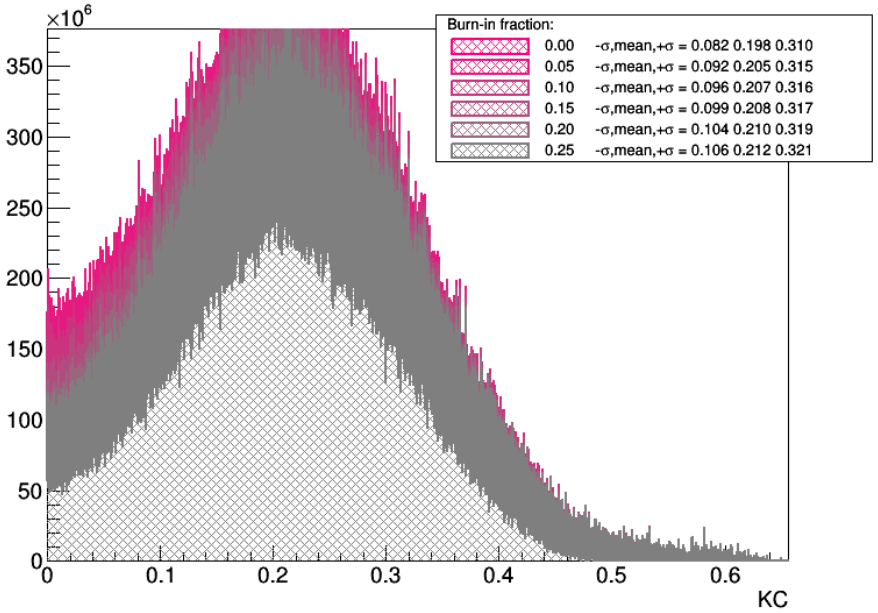
Объединение цепей

В силу сильно возросшего числа систематик в анализе стала наблюдаться нестабильность в апостериорных распределениях. В качестве одного из решений данной проблемы анализ проводится несколько раз при одинаковых условиях, а затем апостериорные распределения суммируются. Такой метод позволяет избежать предвзятости при выборе конкретной итерации анализа.





Распределения после одной цепи



Объединенное распределение

Сравнение пакетов для анализа

Для дальнейшей проверки результатов проводится сравнение анализа, проведенного при помощи пакета theta, с полностью аналогичным анализом в CombinedLimit (разработан в коллаборации CMS). Совпадение результатов в пределах 1σ говорит о независимости результата от методов конкретного пакета.

parameter	$-\sigma$	central	$+\sigma$
sigma_t_ch	0.852	0.888	0.936
sigma_s_ch	0.915	1	1.11
sigma_tW_ch	0.942	1.05	1.15
sigma_ttbar-sl	0.942	0.968	0.981
sigma_ttbar-dl	0.992	1.05	1.11
sigma_Diboson	0.845	1.01	1.21
sigma_DY	0.82	0.988	1.18
sigma_WQQ	0.943	1.2	1.55
sigma_Wc	0.916	1.17	1.45
sigma_Wb	1.1	1.34	1.75
sigma_Wother	1	1.21	1.38
sigma_Wlight	0.755	0.959	1.21
sigma_QCD	0.534	0.71	0.855
lumi	0.986	1	1.02

CombinedLimit

parameter	$-\sigma$	central	$+\sigma$
sigma_t_ch	0.866	0.913	0.958
sigma_s_ch	0.911	1.01	1.11
sigma_tW_ch	0.906	1.06	1.23
sigma_ttbar-sl	0.903	0.959	1.01
sigma_ttbar-dl	0.96	1.08	1.22
sigma_Diboson	0.832	1.01	1.23
sigma_DY	0.805	0.975	1.19
sigma_WQQ	0.877	1.17	1.54
sigma_Wc	0.828	1.08	1.41
sigma_Wb	1.19	1.61	2.06
sigma_Wother	0.904	1.12	1.56
sigma_Wlight	0.775	1.02	1.34
sigma_QCD	0.491	0.644	0.799
lumi	0.989	1.01	1.03

theta

https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/

Заключение

- Отбор переменных: один из методов автоматического отбора переменных; анализ важности переменных для модели, анализ корреляций
- Гиперпространство параметров DNN: оптимизация, визуализация, рекомендации
- Каскады нейронных сетей: улучшение классификации с помощью применения физических закономерностей к созданию нейросетей
- Описан общий рецепт статистического анализа данных коллайдерных экспериментов на примере применения нейронных сетей в анализе одиночного рождения топ-кварка
- На примере t-канального рождения t-кварка получены результаты и приведены разные конфигурации статистического анализа