



UNIWERSYTET
WARSZAWSKI

Wydział Fizyki

**WYDZIAŁ
FIZYKI**

Warszawa, 10.10.2024

dr hab. Artur Kalinowski, prof. UW
Wydział Fizyki, Uniwersytet Warszawski
ul. Pasteura 5
02-093 Warszawa

**Recenzja rozprawy doktorskiej mgr. inż. Michała Mazurka
zatytułowanej “New simulation software and machine learning
technologies in the LHCb experiment to evaluate physics
performance of Run 3”**

Rozprawa jest napisana w języku angielskim i zamyka się na 130 stronach. Tekst jest podzielony na pięć rozdziałów i cztery dodatki. Rozprawa jest poświęcona opisowi nowej wersji środowiska programistycznego, ang. *framework* eksperymentu LHCb, oraz zastosowaniu generatywnej sztucznej inteligencji do parametryzacji odpowiedzi kalorymetru elektromagnetycznego eksperymentu LHCb, oraz modelu YOLO do identyfikacji klastrów w tym kalorymetrze. Jakość parametryzacji została zbadana z użyciem ważnych kanałów rozpadów mezonów B w których występują elektrony, fotony lub miony w stanie końcowym. Rozprawa nie zawiera wyniku fizycznego, typu wartości lub zakresów wykluczeń parametrów fizycznych. Pomimo tego stanowi ważny wkład do rozbudowy eksperymentu LHCb, jako, że tworzenie i rozwój oprogramowania współczesnych eksperymentów jest równie ważne jak budowa i kalibracja detektorów.

Opis rozprawy i komentarze

Rozdział pierwszy wprowadza Czytelnika w zagadnienie i potrzebę zaawansowanych symulacji komputerowych w dziedzinie fizyki cząstek elementarnych. Rozdział ten zawiera też bardzo krótkie przypomnienie struktury Modelu Standardowego.

Uwagi: nie jest prawdą, że „if (...) objects are very heavy, then general relativity is needed to describe their motion” – mechanika Newtonowska była i nadal jest z powodzeniem stosowana do opisu ruchu gwiazd, galaktyk ich gromad. Rysunek 1.1 ilustrujący zakresy stosowalności różnych teorii fizycznych jest niespójny: „Fast” ma jednostki „1/c”, czyli duże wartości na tej osi oznaczają chyba małe prędkości. Pogodnie oś podpisana „small” na jednostki „ \hbar ”, czyli

małe wielkości są na początku osi, ale ten obszar opisano jako reżym mechaniki klasycznej... W tekście napisano że „Universe is made of fermions, half-odd-integer spin ($\frac{1}{2}$, $\frac{3}{2}$,...)”, ale przecież wszystkie elementarne fermiony mają spin $\frac{1}{2}$. W opisie grup cechowania Modelu Standardowego brakuje indeksów „L” i „Y”. W podrozdziale 1.2.2 napisano, że detektory krzemowe są wyjątkowe – nie bardzo wiem na czym polega ich wyjątkowość. Dalej napisano, że kąt rozwarcia stożka Czerenkowa jest proporcjonalny do prędkości cząstki, co nie jest prawdą. Rysunek 1.4 jest opisany jako „trigger rate” ale jednostki na osi Y wskazują że jest to pasmo przepustowości układu akwizycji danych. Niektóre punkty na rysunku są opisane nazwami eksperymentów, np. „LHCb”, inne zderzaczy, np. „LEP”. Opis tła w danych jest co najmniej niewystarczający, jeśli nie mylący – tło pochodzące z błędnej identyfikacji czy szumu elektroniki często stanowi zaniebdywany wkład względem standardowego źródła, czyli procesów o tym samym, co poszukiwany, stanie końcowym. O symulacjach komputerowych napisano „The power of Monte Carlo simulations resides in repeated sampling from the same probability distribution function many times” – uważam że nie jest to prawdą. Jeżeli pojedynczy przykład wylosowano z poprawnego rozkładu prawdopodobieństwa to jest on dobrze wysymulowany – nie ma potrzeby wielokrotnych symulacji. Na koniec niepewność statystyczna typu $1/\sqrt{n}$ nie powinna być nazywana błędem – ang. error.

Rozdział drugi opisuje detektor LHCb. W rozdziale tym opisano krótko podzespoły detektora w jego stanie obecnym – z roku 2024, oraz w przeszłości w czasie pierwszego i drugiego etapu działania LHC – Run1 i Run2. W rozdziale tym opisano też program fizyczny eksperymentu LHCb poświęcony głównie badaniu rozpadów mezonów pięknych.

Uwagi: w opisie mechanizmu produkcji kwarków b w LHC napisano „ratio of incoming parton momenta is strongly asymmetric,„ - nie wiadomo jednak co oznacza „asymetryczny stosunek”. W innym miejscu napisano, że pary bb są produkowane w LHCb pod innymi kątami niż w detektorach ATLAS i CMS - Autorowi chodziło zapewne o różnice w akceptancji tych detektorów. Zderzenia są w nich takie same. W opisie konfiguracji detektora dla Run3 napisano „luminosity of the experiment” – powinno być „luminosity of the accelerator”. Brakuje odnośników do dokumentacji programów ALIEN oraz MOORE.

Rozdział trzeci zawiera opis pierwszego z dwu osiągnięć własnych Autora. Przedstawiono tu architekturę nowego środowiska programistycznego służącego do obsługi symulacji odpowiedzi detektora LHCb – Gauss-on-Gaussino. Tekst zawiera schemat dotychczasowej architektury oprogramowania – Gauss, i nowej - Gauss-on-Gaussino. Opisano dosyć dokładnie oba kroki symulacji: generację listy cząstek, oraz ich propagację przez materiał detektora. Na uwagę zwraca fakt że nowe środowisko – Gaussino – zostało zaprojektowane jako ogólne narzędzie które może być użyte do konfiguracji symulacji innych detektorów niż LHCb. Przedstawiono też tutaj wyniki wizualizacji z użyciem nowej biblioteki DD4HEP. Rozdział zamyka się wzmianką o dokumentacji środowisk Gauss i Gauss-on-Gaussino.

Uwagi: w pewnym miejscu napisano “The age of some parts of the code, which evolved from a version created almost 20 years ago, required pruning to make it easier to extend.” Wydaje się że to nie wiek, ale jakoś i struktura kodu decydują o potrzebie jego refaktoryzacji. Czy do pisania kodu napisanego 20 lat temu użyto innego języka niż dziś? Rzut oka na repozytorium projektu Gaussino pokazuje że 75.6% kodu to nadal C++. W pewnym miejscu napisano, że szybka symulacja kalorymetru elektromagnetycznego używa “point library” – nie jest jasne o jaki typ biblioteki tu chodzi. Po inspekcji odnośnika okazuje się że “point” to nazwa własna. Ponieważ Autor w kolejnym rozdziale opisuje swoje podejście do szybkiej symulacji ECAL

przydałoby się trochę więcej informacji o tej bibliotece. Na stronie 28 „events” zgubiło początkowe „e” w opisie „klasy SignalPlain. Rysunek 3.17 – nie rozumiem co to znaczy „draw trajectories by particle ID, momentum”.

Rozdział czwarty opisuje próbę zastosowania technik uczenia maszynowego do szybkiej symulacji odpowiedzi kalorymetru elektromagnetycznego. Pierwsze podrozdziały opisują krótko implementację interfejsu łączącego środowiska Gaussino i PyTorch oraz ONNXRuntime. W kolejnych częściach rozdziału opisano procedurę treningu i testów generatywnego modelu uczenia maszynowego – kontrolowanego wariacyjnego autoenkodera. Z architektury wynika, że był to kontrolowany autoekoder - CVAE, gdzie parametrami kontrolnymi był czteropęd cząstki (i przypuszczalnie jej typ), a nie VAE jak napisał Autor. Autor atakuje też inne zagadnienie – klastryzację depozytów w kalorymetrze. Zagadnienie to jest rozwiązywane z użyciem ogólnodostępnego modelu YOLO w wersji 3, który był „dotrenowany” przez Autora. Autor opisuje strukturę modelu YOLOv3, oraz wynik porównania jego wydajności z algorytmem automatu komórkowego używanym obecnie.

Uwagi: interfejs łączący środowisko Gaussino i świat narzędzi ML jest ograniczony do PyTorch i ONNXRuntime. Czy rozważono implementację dla interfejsu Keras, który reklamuje się jako narzędzie mogące obsługiwać różne silniki ML? Brakuje dyskusji rysunków 4.1 i 4.2 porównujących wydajność środowisk PyTorch i ONNXRuntime. O ile dobrze je czytam, to PyTorch zużywa mniej pamięci i jest szybsze dla dużej części przestrzeni liczby wątków. Pytanie czy środowisko użyte do testowania – pojedyncza maszyna z 10 rdzeniami MT, jest odpowiednia do takiego testu? Spodziewałbym się, że zrównoleglenie wewnątrz operacyjne nie działa dobrze w takim ustawieniu. Jak to wyglądałoby na maszynie wyposażonej w kartę graficzną z dużą liczbą rdzeni obliczeniowych? Poza tym prosty test polegający na pomnożeniu dwu macierzy nie wygląda na zbyt realistyczny – myślę, że za mało tu komunikacji między wątkami.

W części poświęconej modelowi CVAE brakuje wielu detali:

- wyjaśnienia struktury danych treningowych przedstawionych CaloChallenge, w szczególności struktur widocznych na rysunku 4.6. Z rysunku wynika, że znakomita większość przykładów ma wartość $E=0$, a jedynie niewielki ułamek przykładów ma inne wartości energii. To wskazuje na ogromny poziom niezrównoważenia w zbiorze treningowym. Poniżej piku z zerze znajduje się górnica dla około $E=-0.2$ która pewnie jest odzwierciedleniem depozytów cząstek pochodzących z rozpadów. Nota bene na rysunku 4.9 pokazującym analogiczny rozkład brakuje piku w zerze.
- w opisie próbki użytej do treningu CVAE brak informacji o rozkładzie prawdopodobieństwa energii użytym do generacji próbek.
- próbka walidacyjna miała w sumie 14 tysięcy przypadków, co nie jest dużą liczbą. Dlaczego użyto tak małej próbki w sytuacji gdy do treningu można było przygotować 2 miliony przypadków
- szczegółów architektury, treningu i walidacji modelu VAE
- wyjaśnienia na czym dokładnie polega dodanie profili. W jaki sposób te profile zostały zaimplementowane?
- definicji entropii krzyżowej i rozbieżności KL (czyli Kullbacka–Leiblera)
- wyjaśnienia w jaki sposób modele wytrenowane na ogólnych zbiorach testowych mogą być zaadaptowane do specyfiki poszczególnych eksperymentów. O ile rozumiem Autor ma tu na myśli adaptację w sensie wyników modeli a nie technicznego „spięcia” ich z oprogramowaniem danego eksperymentu.
- rysunki 4.12, pokazujące porównanie CVAE wymagają dyskusji:

- wizualna zgodność rekonstruowanej energii w wariantach Geant4 (G4) i CVAE może być tylko efektem użycia skali logarytmicznej
- ilość czasu użytego przez CVAE zależy od energii cząstki – czy to jest efekt narzutu przez elementy symulacji obsługujące informację z CVAE? Czas wykonania CVAE nie powinien zależeć od energii cząstki – przecież to „tylko” parametryzacja.
- brakuje pogłębionej analizy wydajności CVAE: rozdzielczości energii dla fotonów i elektronów oddzielnie, rozdzielczości energii w funkcji położenia
- brakuje też porównania z obecną szybką symulacją zakodowaną w bibliotece „point”

Uwagi do części poświęconej YOLOv3:

- dlaczego użyto wersji 3 modelu? Obecna wersja to 11, a rysunek reklamujący model pokazuje że ta wersja daje dużo lepsze wyniki niż poprzednie (wersji 3 nawet na nim nie ma)
- w opisie modelu YOLO pojawiają się określenie „skip connections”, ale brakuje jego wyjaśnienia
- brakuje odnośnika do opisu modelu Darknet-53
- w warunkach selekcji zrekonstruowanych klastrów napisano, że różnica energii nie może być większa niż rząd wielkości, ale nie wiadomo wielkości czego
- rysunki 4.15 lewy i prawy różnią się względem liczby kropek, czyli prawdziwych klastrów – po lewej stronie rysunku z automatu komórkowego są dwa klastry, po prawej jeden
- w komentarzu do wydajności algorytmu YOLOv3 napisano, że cechował się dużą liczbą pominiętych klastrów, ale w tabeli 4.1 podano, że oba algorytmy – komórkowy i YOLOv3 mają podobny ułamek zrekonstruowanych klastrów – ponad 97%. Tabela ta też nie jest spójna z rysunkiem 4.17 gdzie pokazano ułamek pominiętych klastrów w funkcji energii klastra
- podobnie jak w części poświęconej parametryzacji odpowiedzi kalorymetru w części poświęconej rekonstrukcji klastrów brakuje pogłębionej analizy.

Rozdział piąty przedstawia analizę rozkładu masy niezmienniczej zrekonstruowanych cząstek w stanie końcowym dla kilku typowych procesów analizowanych w LHCb, czyli rozpadów mezonów B^+ i B^0_s . Dane użyte w tym rozdziale były symulowane z użyciem algorytmu CVAE opisanego w rozdziale czwartym. Rozwożono procesy w których pojawiają się obiekty rejestrowane przez kalorymetr elektromagnetyczny: elektrony i/lub fotony. Przedstawione rysunki rozkładu masy wyraźnie wskazują że symulacja z użyciem CVAE prowadzi do niedoszacowania energii. Autor stosuje procedurę kalibracji próbek uzyskanych przez model CVAE. Procedura ta poprawia wartości średnie, ale szerokości rozkładów masy nadal nie są zgodne. Autor zauważa że może to być wynikiem prostoty modelu CVAE. *Uwagi:* na początku rozdziału pojawia się skrót „CaloML”, który nie jest wyjaśniony. Czym różni się CaloML od CVAE z rozdziału czwartego? Brak etykiety na wstawce z ilorazem na rysunkach 5.1 do 5.10. W części o kalibracji próbek pochodzących z CVAE Autor odwołuje się do procedury „postprocessingu”, która ma być opisana w podrozdziale 4.2.1. Niestety w tym podrozdziale omawiana procedura jest jedynie wymieniona z nazwy, ale brakuje jej opisu: „The total energy deposit e^{\max} and total number of hits n^{\max} are also added in order to

stabilize the postprocessing step, in which the energy deposits are reconstructed from the profiles., Niestety to oznacza, że procedura kalibracji próbek wyprodukowanych z użyciem CVAE jest dla mnie całkowicie niezrozumiała. Podobnie jak w poprzednich rozdziałach brak pogłębionej analizy. Autor nie zauważa, że efekt przesunięcia piku masy niezmienniczej jest dużo silniejszy dla elektronów niż fotonów, np. widać to na rysunkach 5.1 i 5.3. Autor nie próbuje zbadać problemu głębiej. Na przykład dlaczego rozkłady pędu poprzecznego nie wykazują efektu wyraźnego przesunięcia ku niższym energiom dla CVAE, rys. 5.8, gdy masa niezmiennicza jest bardzo przesunięta, rys. 5.6. Myślę, że zamiast (niejasnej) procedury kalibracji należało lepiej zbadać wydajność modelu CVAE. Brak jest porównania z obecną szybką symulacją. Na koniec pytanie jak model CVAE działa dla różnego rodzaju przypadków tła.

Rozdział ostatni, którego nie ma. *Uwagi:* brakuje podsumowania. Praca urywa się nagle, bez podsumowania i planów na przyszłość - czy i jak prace nad modelami uczenia maszynowego będą kontynuowane. W Pracy brakuje też deklaracji wkładu własnego Autora. Nie znalazłem miejsca gdzie Autor opisuje swój wkład do środowiska Gaussino. Dopiero analiza wpisów w repozytorium projektu – gitlab.cern.ch/Gaussino pokazuje że od grudnia 2020 Autor jest głównym programistą pakietu Gaussino.

Podsumowanie

Rozprawa przedstawiona przez mgr. inż. Michała Mazurka nie jest typową pracą w której przedstawiono wynik jakiejś analizy fizycznej zakończony oszacowaniem wartości, lub zakresów wykluczenia pewnej obserwabli fizycznej. Praca Pana Mazurka przedstawia solidną pracę fizyka doświadczalnego na polu które niestety często nie jest doceniane – czyli rozwoju oprogramowania. Oprogramowanie nowoczesnych układów doświadczalnych jest wykonywane rękoma fizyków i tylko przez fizyków może być tworzone. To oznacza że praca nad oprogramowaniem jest taką samą pracą doświadczalnika jak praca nad budową czy kalibracją nowego detektora. Liczne niedociągnięcia edycyjne i braki wyjaśnień nie umniejszają osiągnięć Pana Mazurka. Jako główne, ale trudne do opisanie, osiągnięcie przyjmuję wkład w rozwój środowiska Gaussino. Części poświęcone uczeniu maszynowemu wymagają jednak znacznego dopracowania. Myślę, że byłoby lepiej skupić się na jednym modelu – np. CVAE i dopracować go, niż rozpoczynać nowe zagadnienie jakim było użycie modelu YOLOv3.

Uważam jednak że praca w pełni spełnia wymagania stawiane pracom doktorskim i stawiam wniosek o dopuszczenie mgr. inż. Michała Mazurka do dalszych etapów przewodu doktorskiego.

Artur Kalinowski