

ABSTRACT

Title of Dissertation: Search for emerging-jet signatures in pp collisions at 13 TeV with CMS using a fully data-based method for background extraction

Yi-Mu Chen
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Alberto Belloni
Department of Physics

For this thesis, I present a search for emerging jets produced in proton-proton collisions at a center-of-mass energy of 13 TeV. This search examines a hypothetical dark QCD sector that couples to the standard model through a scalar mediator. The scalar mediator decays into a standard model quark and a dark sector quark. As the dark sector quark showers and hadronizes, it produces long-lived dark mesons that subsequently decay into SM particles, resulting in a jet with multiple displaced vertices, known as an emerging jet. This search extends the existing efforts of CMS by including the possibility of a flavored coupling between the standard model sector and dark sector, which results in emerging jets containing both long-lived and prompt decays. This search looks for pair production of the scalar mediators at the LHC, yielding events with two SM jets and two emerging jets.

As the detector signature left by such dark sector showers significantly deviates from the assumptions made when designing the detector, a fully data-based method is used for evaluating



the number of standard model events that can be mistaken as signal events. The search is carried out on data collected by the CMS experiment corresponding to an integrated luminosity of 138 fb^{-1} , and we observe no significant excess. The results are interpreted using two dark sector models and exclude mediator masses up to 1750 GeV for an unflavored dark QCD model and up to 2000 GeV for a flavored-aligned dark QCD model.

SEARCH FOR EMERGING-JET SIGNATURES IN PP COLLISIONS AT
13 TeV WITH CMS USING A FULLY DATA-BASED METHOD FOR
BACKGROUND EXTRACTION

by

Yi-Mu Chen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
20023

Advisory Committee:

Professor Alberto Belloni (Chair/Advisor)
Professor Sarah Eno (Co-chair)
Professor Christopher Palmer (Co-chair)
Professor Zackaria Chacko
Professor Manuel Franco Sevilla
Professor Massimo Ricotti

ACKNOWLEDGEMENT

I want to express my thanks to Alberto for his continued support, advice, and encouragement throughout the assigned projects while pursuing a PhD degree. I also want to thank Sarah for inviting me to join the CMS group at the University of Maryland and continuously pushing me to further research opportunities.

I want to thank all members of the emerging jets analysis team for the detailed discussion of the solutions and pitfalls presented in this thesis, which greatly advanced my understanding of many topics related to the analysis. I also want to thank all members of the CMS group at the University of Maryland for providing invaluable feedback during all stages of the analysis.

Finally, I want to thank all my friends and family who supported my decision to pursue a PhD degree abroad and continue to provide help, both tangible and intangible, throughout this five-year-long process.

Table of Contents

Acknowledgements	ii
Table of Contents	iii
List of Tables	vi
List of Figures	vii
Chapter 1: Introduction	1
1.1 The standard model of particle physics	2
1.1.1 The electroweak sector	7
1.1.2 Quantum chromodynamics	9
1.2 Dark sector extension and emerging jets	12
1.3 Overview of analysis strategy	19
Chapter 2: Experiment	22
2.1 The Large Hadron Collider	22
2.2 The Compact Muon Solenoid Detector	25
2.2.1 Coordinate system	25
2.2.2 Magnet configuration	28
2.2.3 Tracking subsystems	29
2.2.4 Calorimeter subsystems	33
2.2.5 Muon subsystems	36
2.2.6 Trigger	39
2.2.7 Luminosity measurements	42
Chapter 3: Physics object reconstruction	45
3.1 Particle flow elements	47
3.1.1 Track reconstruction	47
3.1.2 Vertex reconstruction	51
3.1.3 Impact parameters	54
3.1.4 Calorimeter clusters	56
3.2 Particle-flow linking and identification	57
3.3 Compound objects	60
3.3.1 Jets	60
3.3.2 Visible transverse momentum	64
3.3.3 Missing transverse momentum	64

3.4	Physics object quality filters and tagging	66
3.4.1	Primary vertex ordering	67
3.4.2	Photon identification	69
3.4.3	Jet identification	71
3.4.4	Jet flavor tagging	72
3.5	Event generation and simulation	78
3.5.1	Description of computational processes	79
3.5.2	Simulated object corrections	80
3.6	Data sets used for the EMJ analysis	88
3.6.1	Data collected by CMS	88
3.6.2	Signal model simulated events	89
3.6.3	Standard model simulated events	91
Chapter 4:	Emerging jet signature	95
4.1	Event-level selections	96
4.1.1	Trigger requirements	96
4.1.2	Primary vertex selection	97
4.1.3	Custom Jet-track association	99
4.2	Emerging jets tagging variables	102
4.2.1	Model-agnostic approach to EMJ tagging	104
4.2.2	Machine learning approach to EMJ tagging	110
4.3	Determining the cut values for signal extraction	115
4.4	Sources of displaced tracks in standard model processes	124
4.4.1	Long-lived standard model mesons	124
4.4.2	Pileup events	124
4.4.3	Material interaction	125
Chapter 5:	Data-based background estimation	129
5.1	Data-based mistag rate evaluation	132
5.1.1	Definition of the signal-free regions	132
5.1.2	Computing flavor dependence of mistag rate	134
5.1.3	Estimating b jet fraction	136
5.1.4	Signal free region mistag rate evaluation	137
5.2	The mistag rate scale factor	145
5.2.1	Calculating an event-level scale factor	145
5.2.2	Flavor assignment in SF computation	149
5.3	Determining mistag rate parameterization	154
5.4	Assessment of background estimation correctness	161
5.4.1	Validation test using MC events	161
5.4.2	Validation test using data	165
Chapter 6:	Uncertainties	168
6.1	Background estimation uncertainties	170
6.1.1	Propagation of statistical uncertainty	170
6.1.2	Uncertainties associated with the calculation assumption	173

6.2	Signal selection uncertainties	179
6.2.1	Reconstruction kinematic variable corrections	179
6.2.2	Event weighting corrections	183
6.2.3	Summary of uncertainties on signal samples	187
Chapter 7:	Results	189
7.1	The CL_s upper limit	189
7.2	Interpreting results of the different EMJ tagging methods	195
7.3	Conclusion and outlook	200
Chapter A:	Comparison with the previous EMJ search efforts	201
A.1	Removal of explicit p_T^{miss} cut for signal selection	201
A.2	Parameterization of the mistag rate using jet p_T	205
A.3	Discarding per-jet flavor assignment for background estimation	205
A.4	Removing a vertex quality cut	205
Appendix B:	Additional plots	208
B.1	Kinematic dependence of the mistag rates	208
B.2	Estimated b jet fraction of the signal-free region	215
B.3	Comparison of the mistag rates evaluated in different MC samples	244
B.4	Mistag rates evaluated in signal-free data stream	252
B.5	Estimated fraction of the b jet fraction of the control region	260
B.6	Limits results	276
Acronyms		287
Bibliography		289

List of Tables

1.1	Fermion in the standard model	7
1.2	Bosons in the standard model	7
3.1	Photon identification cut values	71
3.2	H_T -triggered CMS data sets used for analysis	89
3.3	γ -triggered CMS data sets use for analysis	90
3.4	Unflavored signal model parameters	90
3.5	Flavor-aligned signal model parameters	92
3.6	Simulated data set used for analysis	94
4.1	Cut values for model-agnostic EMJ tagging (unflavored)	116
4.2	Cut values for model-agnostic EMJ tagging (flavor-aligned)	117
4.3	Cut values for GNN tagging	117
4.4	Cut sets for signal event selection	118
5.1	Figure-of-merit comparison for mistag rate binning methods	160
5.2	MC background prediction and estimation using pixel Phase-0	162
5.3	MC background prediction and estimation using pixel Phase-1	164
5.4	Validation test of background estimation using data	167
6.1	Uncertainties of background estimation results with pixel Phase-0	177
6.2	Uncertainties of background estimation results with pixel Phase-1	178
6.3	Signal model acceptance uncertainties using model-agnostic EMJ tagging	187
6.4	Signal model acceptance uncertainties for GNN tagging	188
7.1	Events passing signal selection (pixel Phase-0 geometry)	190
7.2	Events passing signal selection (pixel Phase-1 geometry)	191

List of Figures

1.1	Example Feynman diagram	3
1.2	Screening correction Feynman diagrams	10
1.3	Hidden valley model diagram	13
1.4	Emerging jets production schematic	14
1.5	Flavored dark sector meson lifetime spectrum	17
1.6	Dark sector particle shower composition	18
1.7	Feynman diagram of production process of interest	20
1.8	Emerging jets analysis primary event topology	21
2.1	Accelerator complex at CERN	23
2.2	CMS detector overview schematic	26
2.3	Polar angle–pseudorapidity conversion	28
2.4	Pixel geometry comparison	31
2.5	Tracking subsystem layout	32
2.6	ECAL schematic layout	35
2.7	HCAL schematic layout	37
2.8	Muon system anode–cathode configurations	40
2.9	Muon system schematic layout	41
2.10	Integrated luminosity measurement	44
3.1	Overview of particle-flow elements	46
3.2	Track reconstruction efficiency measured with data	49
3.3	Track reconstruction efficiency measured with simulated events	50
3.4	Track reconstruction resolution	50
3.5	Weighting function used for vertex fitting	54
3.6	Vertex reconstruction resolution	55
3.7	Schematic for determining track impact parameter sign	56
3.8	Jet clustering method comparison	62
3.9	Jet energy correction stages	63
3.10	Jet energy correction example	64
3.11	Jet energy compositions	72
3.12	b tagging example diagram	75
3.13	DeepJet b tagging performance	78
3.14	Relative trigger efficiency of H_T -trigger	82
3.15	Pile up distribution measured at CMS	83
3.16	Vertex multiple with pileup reweighting corrections	84
3.17	DeepJet b tagging performance	87

3.18	DeepJet score correction factors	87
4.1	Trigger efficiency of H_T -trigger	98
4.2	Reconstruction and generated vertex separation	99
4.3	Distribution of event-level variables	103
4.4	Variables used for model-agnostic unflavored EMJ tagging	106
4.5	Comparison of dark shower distribution	107
4.6	Variables used for model-agnostic flavor-aligned EMJ tagging	109
4.7	Jet images for different underlying partons	112
4.8	GNN output score distribution	114
4.9	Selection efficiency of unflavored models using model-agnostic tagger	120
4.10	Selection efficiency of flavor-aligned models using model-agnostic tagger	121
4.11	Selection efficiency of unflavored models using GNN tagging	122
4.12	Selection efficiency of flavor-aligned models using GNN tagging	123
4.13	Photon conversion process	125
4.14	Distribution of track IP_{2D} - p_T product	126
4.15	Material interaction example structure	127
4.16	Material interaction vertex position and simulation distribution	128
5.1	Mistag rate of the “u-tag 1” tagger	130
5.2	Mistag rate of the “a-tag 1” tagger	130
5.3	Mistag rate of the “u-GNN tag 1” tagger	131
5.4	Mistag rate of the “a-GNN tag 1” tagger	131
5.5	Event selection flow of the signal free region	134
5.6	Mistag rate comparison between pixel geometry differences	138
5.7	Template fitting for b jet fraction estimation	139
5.8	Evaluated b jet fraction as a function of jet p_T	140
5.9	Comparison of mistag rates between data and MC	141
5.10	Comparison of mistag rate between γ +jets and QCD MC	143
5.11	QCD and γ +jets process differences	144
5.12	Distribution of tracks in H_T - and γ -triggered data sets	144
5.13	Event selection flow for SR/CR construction	146
5.14	Estimated b jet fraction of jets in the control region	150
5.15	DeepJet b tagging efficiency difference	154
5.16	Numerical experiment of sampling median	157
5.17	MC background prediction and estimation using pixel Phase-0	163
5.18	MC background prediction and estimation using pixel Phase-1	165
5.19	Validation test of background estimation using data	167
6.1	Mistag rate statistical error propagation example	171
6.2	Breakdown of background estimation assumptions	177
6.3	Estimation variations used for uncertainty evaluation	178
6.4	Differences in the distribution of track impact parameter in data and MC	181
6.5	Estimation of track reconstruction efficiency of very-displaced tracks	183
7.1	Comparison of events passing selection and background estimation	190

7.2	Example exclusion limit plot	195
7.3	Exclusion of unflavored model using model-agnostic EMJ tagging	196
7.4	Exclusion of unflavored model using GNN EMJ tagging	196
7.5	Exclusion of flavor-aligned model using model-agnostic EMJ tagging	197
7.6	Exclusion of flavor-aligned model using GNN EMJ tagging	198
A.1	Comparison of p_T^{miss} distribution in data and MC	202
A.2	Comparison of jet energy composition shifts with p_T^{miss} cuts	204
A.3	Selection efficiency as a function of vertex quality requirement	206
B.1	Mistag rate of “u-tag 1” tagger evaluated using QCD MC	208
B.2	Mistag rate of “u-tag 2” tagger evaluated using QCD MC	209
B.3	Mistag rate of “u-tag 3” tagger evaluated using QCD MC	209
B.4	Mistag rate of “u-tag 4” tagger evaluated using QCD MC	209
B.5	Mistag rate of “u-tag 5” tagger evaluated using QCD MC	210
B.6	Mistag rate of “validation u-tag” tagger evaluated using QCD MC	210
B.7	Mistag rate of “a-tag 1” tagger evaluated using QCD MC	210
B.8	Mistag rate of “a-tag 2” tagger evaluated using QCD MC	211
B.9	Mistag rate of “a-tag 3” tagger evaluated using QCD MC	211
B.10	Mistag rate of “a-tag 4” tagger evaluated using QCD MC	211
B.11	Mistag rate of “a-tag 5” tagger evaluated using QCD MC	212
B.12	Mistag rate of “validation a-tag” tagger evaluated using QCD MC	212
B.13	Mistag rate of “uGNN tag 1” tagger evaluated using QCD MC	212
B.14	Mistag rate of “uGNN tag 2” tagger evaluated using QCD MC	213
B.15	Mistag rate of “uGNN tag 3” tagger evaluated using QCD MC	213
B.16	Mistag rate of “uGNN validation tag” tagger evaluated using QCD MC	213
B.17	Mistag rate of “uGNN tag 1” tagger evaluated using QCD MC	214
B.18	Mistag rate of “uGNN tag 2” tagger evaluated using QCD MC	214
B.19	Mistag rate of “uGNN tag 3” tagger evaluated using QCD MC	214
B.20	Mistag rate of “uGNN validation tag” tagger evaluated using QCD MC	215
B.21	b jet fraction fit of γ +jets MC samples with pixel Phase-0 (part 1)	216
B.22	b jet fraction fit of γ +jets MC samples with pixel Phase-0 (part 2)	217
B.23	b jet fraction fit of γ +jets MC samples with pixel Phase-0 (part 3)	218
B.24	b jet fraction fit of γ +jets MC samples with pixel Phase-0 (part 4)	219
B.25	b fraction summary of γ +jets MC samples with pixel Phase-0	219
B.26	b jet fraction fit of γ -trigger data set with pixel Phase-0 (part 1)	220
B.27	b jet fraction fit of γ -trigger data set with pixel Phase-0 (part 2)	221
B.28	b jet fraction fit of γ -trigger data set with pixel Phase-0 (part 3)	222
B.29	b jet fraction fit of γ -trigger data set with pixel Phase-0 (part 4)	223
B.30	b fraction summary of γ -trigger data set with pixel Phase-0	223
B.31	b jet fraction fit of γ +jets MC samples with pixel Phase-1 (part 1)	224
B.32	b jet fraction fit of γ +jets MC samples with pixel Phase-1 (part 2)	225
B.33	b jet fraction fit of γ +jets MC samples with pixel Phase-1 (part 3)	226
B.34	b jet fraction fit of γ +jets MC samples with pixel Phase-1 (part 4)	227
B.35	b fraction summary of γ +jets MC samples with pixel Phase-1	227

B.36 b jet fraction fit of γ -trigger data set with pixel Phase-1 (part 1)	228
B.37 b jet fraction fit of γ -trigger data set with pixel Phase-1 (part 2)	229
B.38 b jet fraction fit of γ -trigger data set with pixel Phase-1 (part 3)	230
B.39 b jet fraction fit of γ -trigger data set with pixel Phase-1 (part 4)	231
B.40 b fraction summary of γ -trigger data set with pixel Phase-1	231
B.41 b jet fraction fit of γ +jets MC samples with pixel Phase-0 (part 1)	232
B.42 b jet fraction fit of γ +jets MC samples with pixel Phase-0 (part 2)	233
B.43 b jet fraction fit of γ +jets MC samples with pixel Phase-0 (part 3)	234
B.44 b fraction summary of γ +jets MC samples with pixel Phase-0	234
B.45 b jet fraction fit of γ -trigger data set with pixel Phase-0 (part 1)	235
B.46 b jet fraction fit of γ -trigger data set with pixel Phase-0 (part 2)	236
B.47 b jet fraction fit of γ -trigger data set with pixel Phase-0 (part 3)	237
B.48 b fraction summary of γ -trigger data set with pixel Phase-0	237
B.49 b jet fraction fit of γ +jets MC samples with pixel Phase-1 (part 1)	238
B.50 b jet fraction fit of γ +jets MC samples with pixel Phase-1 (part 2)	239
B.51 b jet fraction fit of γ +jets MC samples with pixel Phase-1 (part 3)	240
B.52 b fraction summary of γ +jets MC samples with pixel Phase-1	240
B.53 b jet fraction fit of γ -trigger data set with pixel Phase-1 (part 1)	241
B.54 b jet fraction fit of γ -trigger data set with pixel Phase-1 (part 2)	242
B.55 b jet fraction fit of γ -trigger data set with pixel Phase-1 (part 3)	243
B.56 b fraction summary of γ -trigger data set with pixel Phase-1	243
B.57 Mistag rate of GNN tagger for unflavored models in MC	244
B.58 Mistag rate of GNN tagger for unflavored models in MC	245
B.59 Mistag rate of GNN tagger for flavor-aligned models in MC	246
B.60 Mistag rate of GNN tagger for flavor-aligned models in MC	247
B.61 Mistag rate of model-agnostic tagger for unflavored models in MC	248
B.62 Mistag rate of model-agnostic tagger for unflavored models in MC	249
B.63 Mistag rate of model-agnostic tagger for flavor-aligned models in MC	250
B.64 Mistag rate of model-agnostic tagger for flavor-aligned models in MC	251
B.65 Mistag rate of GNN tagger for unflavored signal samples in data	252
B.66 Mistag rate of GNN tagger for unflavored signal samples in data	253
B.67 Mistag rate of GNN tagger for flavor-aligned models in data	254
B.68 Mistag rate of GNN tagger for flavor-aligned models in data	255
B.69 Mistag rate of model-agnostic tagger for unflavored models in data	256
B.70 Mistag rate of model-agnostic tagger for unflavored models in data	257
B.71 Mistag rate of model-agnostic tagger for flavor-aligned models in data	258
B.72 Mistag rate of model-agnostic tagger for flavor-aligned models in data	259
B.73 b jet fraction of control regions using GNN tagger for unflavored signal samples in MC (pixel Phase-0)	260
B.74 b jet fraction of control regions using GNN tagger for unflavored signal samples in data (pixel Phase-0)	261
B.75 b jet fraction of control regions using GNN tagger for unflavored signal samples in MC (pixel Phase-1)	262
B.76 b jet fraction of control regions using GNN tagger for unflavored signal samples in data (pixel Phase-1)	263

B.77 b jet fraction of control regions using GNN tagger for flavor-aligned models in MC (pixel Phase-0)	264
B.78 b jet fraction of control regions using GNN tagger for flavor-aligned models in data (pixel Phase-0)	265
B.79 b jet fraction of control regions using GNN tagger for flavor-aligned models in MC (pixel Phase-1)	266
B.80 b jet fraction of control regions using GNN tagger for flavor-aligned models in data (pixel Phase-1)	267
B.81 b jet fraction of control regions using model-agnostic tagger for unflavored models in MC (pixel Phase-0)	268
B.82 b jet fraction of control regions using model-agnostic tagger for unflavored models in data (pixel Phase-0)	269
B.83 b jet fraction of control regions using model-agnostic tagger for unflavored models in MC (pixel Phase-1)	270
B.84 b jet fraction of control regions using model-agnostic tagger for unflavored models in data (pixel Phase-1)	271
B.85 b jet fraction of control regions using model-agnostic tagger for flavor-aligned models in MC (pixel Phase-0)	272
B.86 b jet fraction of control regions using model-agnostic tagger for flavor-aligned models in data (pixel Phase-0)	273
B.87 b jet fraction of control regions using model-agnostic tagger for flavor-aligned models in MC (pixel Phase-1)	274
B.88 b jet fraction of control regions using model-agnostic tagger for flavor-aligned models in data (pixel Phase-1)	275
B.89 Limits of unflavored models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$ obtained using GNN-based tagging	277
B.90 Limits of unflavored models with $m_{\pi_{\text{dark}}} = 20 \text{ GeV}$ obtained using GNN-based tagging	278
B.91 Limits of flavor-aligned models with $m_{\pi_{\text{dark}}} = 6 \text{ GeV}$ obtained using GNN-based tagging	279
B.92 Limits of flavor-aligned models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$ obtained using GNN-based tagging	280
B.93 Limits of flavor-aligned models with $m_{\pi_{\text{dark}}} = 20 \text{ GeV}$ obtained using GNN-based tagging	281
B.94 Limits of unflavored models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$ obtained using model-agnostic tagging	282
B.95 Limits of unflavored models with $m_{\pi_{\text{dark}}} = 20 \text{ GeV}$ obtained using model-agnostic tagging	283
B.96 Limits of flavor-aligned models with $m_{\pi_{\text{dark}}} = 6 \text{ GeV}$ obtained using model-agnostic tagging	284
B.97 Limits of flavor-aligned models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$ obtained using model-agnostic tagging	285
B.98 Limits of flavor-aligned models with $m_{\pi_{\text{dark}}} = 20 \text{ GeV}$ obtained using model-agnostic tagging	286

Chapter 1: Introduction

The standard model (SM) of particle physics is the most successful theory in describing the interactions of subatomic particles. After the discovery of the Higgs boson [1, 2], all known terms of the standard model have been accounted for. However, it is also apparent that the standard model is not a complete description of physics, with many observations unaccounted for. In particular, though there are multiple astronomical [3, 4, 5] and cosmological [6, 7] indications for the existence of dark matter (DM), the particle nature of dark matter has yet to be uncovered. Multiple detector experiments have sought the simplest extension of the standard model, which includes a weakly interacting massive particle (WIMP) to account for dark matter [8]. Currently, no evidence has been found for such extensions [9, 10]. An alternate theory of dark matter proposes that dark matter is a class of stable particles that exist within a separate “dark sector” with new interactions within this new sector of physics [11, 12, 13, 14]. If a strongly self-interacting dark sector is coupled with standard model particles through some high-mass mediator, such models can also be probed at collider experiments, although the theory predicts drastically different detector signatures compared with WIMP models. One such signature is that of the **emerging jets** [15, 16], where fermions in the dark sector undergo a process similar to hadronization of standard-model quarks, generating a jet of semi-stable dark-sector particles before decaying back into standard model particles displaced from the original production point,

resulting in jets where their observed constituents “emerge” from multiple displaced vertices.

Since the formulation of the dark sector model borrows various concepts from the standard model, Section 1.1 will first give a brief overview of the standard model, highlighting points of interest used in the formulation of the dark sector model. Then, Section 1.2 will introduce the proposed dark sector theory, listing the constraints imposed by existing observations, additional beyond-standard-model problems the theory can potentially solve, and discuss the requirements and resulting detector signatures of the dark sector model that generate emerging jet signatures. Finally, Section 1.3 will provide an overview of the analysis strategy of how to look for emerging-jet signatures at collider experiments, which is the main topic of this thesis.

1.1 The standard model of particle physics

The standard model is built on **quantum field theory** (QFT), where particles are depicted as excitations in a quantum field, and the evolution of the fields between an initial state $\langle \Omega_i |$ and a final state $|\Omega_f\rangle$ of interest is described by solving the theory Lagrangian to obtain a **transition amplitude**. For Lagrangians containing interaction terms, the computation of this amplitude is performed by a perturbative expansion, commonly expressed as Feynman diagrams, as shown in Fig. 1.1, with external edges used to represent the initial and final state, and vertices to represent the allowed interactions between particles in the initial/final state and transient states. The total contribution to the scattering amplitude is calculated as the product of the interaction vertices, the internal edges joining the vertices contributing a factor of a propagator obtained from solving the interaction-free Lagrangian¹.

The perturbative expansion of higher-order diagrams containing loops, however, evaluates

¹Interaction-free Lagrangian contains only quadratic terms of a single field type.

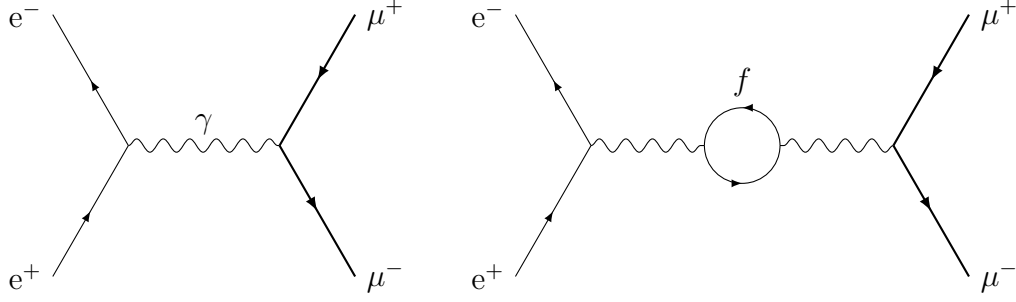


Figure 1.1: Example Feynman diagrams of the $e^+e^- \rightarrow \mu^+\mu^-$ process, the left is the leading order “tree-level” diagram and the right is a “next-to-leading order” diagram with a loop correction to the internal photon propagator.

to infinity if calculated directly. **Renormalization** [17] provides a framework to consistently reinterpret the results of diverging loop diagrams to be absorbed as corrections to the particle mass and coupling constants. The renormalized theory predicts that interaction strength will vary depending on the energy exchange between the particles of interest.

The construction of the various terms in the Lagrangian of the standard model are expected to preserve **Lorentz invariance** and **local gauge invariance** [18]. **Fermion fields** describe particles with spin-1/2 with a Lagrangian in the form of:

$$\mathcal{L}_{\text{ferm}} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi, \quad (1.1)$$

where ψ is the 4-component Dirac spinor field used to express the spin structure, γ^μ are the 4×4 Dirac matrices that obeys a **Clifford algebra**: $\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2g^{\mu\nu}$ with $g^{\mu\nu}$ being the Minkowski metric of special relativity, and m being the mass term of the fermion field. **Vector gauge fields** describe particles with spin-1 with a Lagrangian in the form of:

$$\mathcal{L}_{\text{vec}} = -\sum_a \frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu}; \quad F_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g \sum_{b,c} \mathcal{F}^{abc} G_\mu^b G_\nu^c; \quad (1.2)$$

where G_μ^a is the vector field (hence the Lorentz index μ) with the index a used to index the **generators** of the gauge group used to construct the Lagrangian, g is the vector field coupling constant, and $\mathcal{F}^{abc}$ is the structure constant of the gauge group: $T_{\alpha\beta}^a T_{\beta\gamma}^b - T_{\alpha\beta}^b T_{\beta\gamma}^a = i\mathcal{F}^{abc} T_{\alpha\gamma}^c$, where $T_{\alpha\beta}^c$ are the generators of the gauge group, with group elements indexed using α, β and γ . Interaction between fermion fields and vector gauge fields are expressed by the Lagrangian:

$$\mathcal{L}_i = \sum_{a,\alpha,\beta} g \bar{\psi}_\alpha \gamma^\mu G_\mu^a T_{\alpha\beta}^a \psi_\beta, \quad (1.3)$$

where the indices α and β are used to index the fermion fields that are part of the gauge group. A theory is guaranteed to be renormalizable [19] if the total Lagrangian is unchanged after a *local* gauge transformation $\theta^a(x)$:

$$\psi_\alpha \rightarrow \sum_{\beta} \exp(i\theta^a T_{\alpha\beta}^a) \psi_\beta ; \quad G_\mu^a \rightarrow G_\mu^a + \partial_\mu \theta^a. \quad (1.4)$$

For a more compact expression of including both the fermion and the boson along with their interaction, we can express the full Lagrangian as:

$$\mathcal{L}_{\text{ferm.}+\text{bos.}+i} = \sum_a \bar{\psi}_\alpha (i\gamma^\mu \partial_\mu + i\gamma^\mu G_\mu^a T_{\alpha\beta}^a) \psi_\beta - \frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu} - m \bar{\psi} \psi \quad (1.5)$$

$$= \bar{\psi} i\gamma^\mu D_\mu \psi - m \bar{\psi} \psi - \sum_a \frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu}, \quad (1.6)$$

where D_μ is the **covariant derivative operator** that include the gauge fields of interest.

To preserve local gauge invariance, all vector fields must be massless. While we can inject

a mass term into the Lagrangian of the vector boson (named the Proca Lagrangian):

$$\mathcal{L}_{\text{Proca.}} = - \sum_a \frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu} - \frac{1}{2} m_a^2 G^{a\mu} G_\mu^a \quad (1.7)$$

the explicit mass term of $G^\mu G_\mu$ is not invariant under local gauge transformation if the gauge group is non-abelian. The Brout–Englert–Higgs mechanism [20, 21] allows for the generation of gauge-preserving mass terms to be included in the Lagrangian via the inclusion of fields ϕ_α with that undergoes **spontaneous symmetry breaking**. The Lagrangian of the additional “**Higgs**” **field** is expressed with a negative mass term with a quartic self-interaction term:

$$\mathcal{L}_H = \frac{1}{2} \partial_\mu \phi_\alpha \partial^\mu \phi_\alpha + \frac{1}{2} \mu^2 \phi_\alpha^2 + \frac{\lambda}{4} \phi_\alpha^4. \quad (1.8)$$

The minimum energy of this field is evaluated at a non-zero field value, $\phi_{\text{VEV}}^\dagger \phi_{\text{VEV}} = \Phi_{\text{VEV}}^2 = \lambda/2\mu$, expanding the Lagrangian around this **vacuum expected value**: $\phi = \eta + \phi_{\text{VEV}}$, the components of the ϕ field will yield one massive particle field, and n massless **Nambu–Goldstone boson** fields [22], where n corresponds to the degree of symmetry broken by expanding around the vacuum expected value. Having the field interact with the vector boson gauge field via an interaction term:

$$\mathcal{L}_v = g'^2 \left(G^{a\mu} T_{\alpha\beta}^a \phi_\beta \right)^\dagger \left(G_\mu^{a'} T_{\alpha'\beta'}^{a'} \phi_{\beta'} \right) \quad (1.9)$$

Expansion of this interaction around the VEV of ϕ will have the Nambu-Goldstone be absorbed

as effective mass terms of G_a^μ :

$$\mathcal{L}_m = \Phi_{\text{VEV}} g'^2 G_\mu G^\mu \quad (1.10)$$

which preserves local gauge invariance if the additional Higgs field also transform under the local gauge transformation:

$$\phi_\alpha \rightarrow \sum_{a,\beta} \exp \left(i g' \theta^a T_{\alpha\beta}^a \right) \phi_\beta. \quad (1.11)$$

Having presented the theoretical foundations of the standard model, we can now list the components of the standard model as we currently know it. There are currently twelve **flavors** of fermions, four gauge bosons, and one Higgs field together in an $SU(3) \times SU(2)_L \times U(1)$ symmetry group. An overview of the various field types can be found in Tables 1.1 and 1.2. Of the twelve fermion flavors, six participate in the $SU(3)$ **strong interaction**, which we call **quarks**; the other 6 fermion flavors will be referred to as **leptons**. All particles are currently known to participate in the $SU(2)_L \times U(1)$ **electroweak interaction**. In the following sections, we will concretely introduce the various components, as well as highlight various phenomenological implications that arise from the theory.

Table 1.1: Fermions in the standard model with their electric charge and mass [23]. Neutrino mass upper limits are determined to a confidence level of 95%.

Generation	Particle	Leptons		Mass [MeV]	Particle	Quarks		Mass [GeV]
		Charge				Charge		
1	electron (e^-)	-1		0.511	down (d)	-1/3		4.67×10^{-3}
	neutrino (ν_e)	0	$< 0.8 \times 10^{-9}$		up (u)	+2/3		2.16×10^{-3}
2	muon (μ^-)	-1		106	strange (s)	-1/3		93.4×10^{-3}
	neutrino (ν_μ)	0	< 0.19		charm (c)	+2/3		1.27
2	tau (τ^-)	-1		1780	bottom (b)	-1/3		4.18
	neutrino (ν_τ)	0	< 18.2		top (t)	+2/3		173

Table 1.2: Bosons in the standard model with their electric charge and mass [23]. Other than the Higgs boson, all bosons have spin 1.

Force	Particle	Charge	Mass [GeV]
Strong	gluon (g)	0	0
Electromagnetic	photon (γ)	0	0
Weak	W bosons ($W^\pm$)	± 1	80.4
	Z bosons (Z)	0	91.2
–	Higgs boson (H)	0	125.2

1.1.1 The electroweak sector

The electroweak (EW) sector refers to the $SU(2)_L \times U(1)$ gauge symmetry interaction [24, 25, 26]. The twelve fermions are grouped into six **generations**:

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L ; \quad \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L ; \quad \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L ; \quad \begin{pmatrix} u \\ d \end{pmatrix}_L ; \quad \begin{pmatrix} c \\ s \end{pmatrix}_L ; \quad \begin{pmatrix} t \\ b \end{pmatrix}_L ; \quad (1.12)$$

where the subscript L indicates that we only take the left-hand **chiral projection** of the involved fermion fields: $\frac{1}{2} (1 - i\gamma^0\gamma^1\gamma^2\gamma^3) \psi$. As we can split all fermion fields into left- and right-handed components $\psi = \psi_L + \psi_R$, the covariant derivative of the $SU(2)_L$ interaction only acts on the ψ_L part of the fermions. The $W_\mu^\pm$ bosons are the **flavor-changing fields**, composed of the linear combination of the $SU(2)$ generator with off-diagonal components $W^\pm = W^1 \pm iW^2$. The

product between the $SU(2)$ and $U(1)$ indicates that there is mixing between the flavor-preserving W_μ^3 component of the $SU(2)$ interaction and the $U(1)$ interaction boson B_μ , expressed by the **Weinberg angle** θ_W [27]:

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \cos \theta_W & \sin \theta_W \\ -\sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} B_\mu \\ W_\mu^3 \end{pmatrix}, \quad (1.13)$$

with the resulting A_μ field being the massless photon of electromagnetic interaction, and Z_μ being the massive flavor-preserving weak-interaction boson. The value of θ_W can be measured from the mass ratio of the $W^\pm$ and Z bosons as $\cos \theta_W = m_W/m_Z$.

To account for the W and Z bosons being massive, it was predicted [22] that a scalar Higgs field composed of a complex doublet is required to generate the gauge-preserving mass terms of the electroweak bosons via spontaneous symmetry breaking, which will result in a massive Higgs boson. The Higgs boson was the last element of the standard model to be found [1, 2]. The Higgs field also generates the mass terms of the fermions via **Yukawa coupling terms**. Using the u and d doublet as an example, the mass term of the quarks can be expressed as:

$$\mathcal{L}_{\text{mass}} = -\kappa_u H_i \overline{\mathcal{D}}_i u - \kappa_d \epsilon^{ij} H_i^\dagger \overline{\mathcal{D}}_j d \quad (1.14)$$

where the H_i is the Higgs field complex doublet, $\mathcal{D}$ is the spinor doublet of the u and d quarks, and κ_u determines the magnitude of the mass.

If the mass eigenstates and the interaction eigenstates are aligned, bound states composed of particles of 2 different generations cannot decay. Experimentally, this is found to not be the case for quarks, where the mass eigenstates and weak-interaction eigenstates differ. This difference

generates the **Glashow–Iliopoulos–Maiani mechanism** [28] originally used to explain the rare decay of kaons, and is fully described using the unitary **Cabbibo–Kobayashi–Maskawa matrix** V_{CKM} [29]:

$$\begin{pmatrix} \mathcal{D}' \\ \mathcal{S}' \\ \mathcal{B}' \end{pmatrix} = V_{CKM} \begin{pmatrix} \mathcal{D} \\ \mathcal{S} \\ \mathcal{B} \end{pmatrix}, \quad (1.15)$$

where $\mathcal{D}$, $\mathcal{S}$, $\mathcal{B}$ are the spinor doublets of the quarks of the various generations, the primed elements notate the weak-interaction eigenstates, and the non-primed elements notate the mass eigenstates. For heavy-flavored bound states that contain a bottom and charm quarks (b mesons and c mesons²), the decays amplitude of such bound states are further suppressed by the smallness of the off-diagonal components of the CKM matrix, resulting in these bound states having long lifetimes of 10^{-3} ns [23]. Such lifetimes become an important feature in the study of b production processes at collider experiments [30, 31, 32].

1.1.2 Quantum chromodynamics

Quantum chromodynamics (QCD) is the $SU(3)$ gauge theory that describes the strong interaction [33]. Fermions participating in QCD contain an additional quantum number, QCD color charge, which can take on three values, and these fermions interact with the eight gluons. Unlike the electroweak interaction, QCD is non-chiral with no symmetry breaking. Because of this, fermions and gluons that differ only in color charge are effectively degenerate for interactions that are blind to color charge. Because of this, gluons are typically referred to as “one field”

²The definition of “mesons” will be formally given in Section 1.1.2.

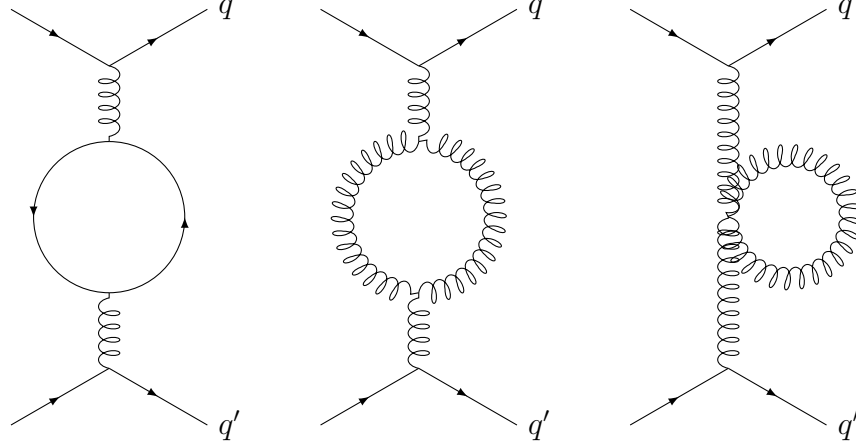


Figure 1.2: Example Feynman diagrams that contribute to screening corrections (fermion loops, left) and anti-screening corrections (boson loops, middle and right), for a quark-quark interaction. The key difference between the diagrams that give rise to asymptotic freedom is the sign of the contribution of anti-commuting fermion fields, and the commuting boson fields.

instead of eight degenerate fields for brevity outside the context of QCD calculations.

Particles interacting via strong interaction will have a “screening” contribution from fermion-loop corrections, and an “anti-screening” contribution from the boson-loop corrections. With eight gluon fields and the six flavors of quarks, the anti-screening effect dominates, resulting in the interaction strength decreasing as the interaction energy scale increases, a phenomenon named **asymptotic freedom** [34, 35]. The interaction strength as a function of the interaction energy can be expressed as:

$$\alpha_S(Q^2) = \frac{g_S^2(Q)}{4\pi} = \frac{12\pi}{(33 - 2N_f) \ln \left(\frac{Q^2}{\Lambda_S^2} \right)}, \quad (1.16)$$

where N_f is the number of fermions that carry color charge, and Λ_S is the energy scale under which the interaction strength diverges.

At low energies, the increase in interaction strength will confine color-charged particles into

color-singlet bound states called **hadrons**³. To form a color-neutral state, bound states created in hadronization processes are either a quark-antiquark pair called **mesons** or quark triplets with each quark being a different color called **baryons**. All mesons are unstable and decay to lighter hadrons or leptons. The computation of the bound state will require nonperturbative solutions to the QCD interaction and currently has no closed-form expression [36]. The interaction of hadrons with other particles in high-energy collisions will need to be expressed via a **parton distribution function** (PDF) [37, 38], $f_i(x, Q^2)$, which measures the probability of the collision involving the strong interacting particle of flavor i that carries momentum xp of the hadron momentum p given the interaction energy Q . While the values of the PDFs cannot be calculated from first principles and need to be evaluated using experimental observation, the evolution of $f_i(x, Q^2)$ with varying values of Q can be calculated using perturbative QCD [39]. The parton distribution function of the proton is particularly important for experiments at the LHC.

Color-charged partons emerging from hard-scattering processes will generate multiple color-charged particles as the energy scale drops, pulling fermion–antifermion pairs from the surrounding vacuum [40]. What is observed in the final state is a **jet** of color-singlet particles within an angular vicinity of the original color-charged parton. The **fragmentation functions** describe the kinematic distribution of particles in the hadronization shower and reside in the nonperturbative regime of QCD. The **Lund string model** [41] is commonly used to estimate the fragmentation functions for partons emerging from high-energy interactions.

³Though the resulting hadrons being color singlets is a well-established empirical result, this result is currently not possible to derive from first principles. For more information, see Ref. [36].

1.2 Dark sector extension and emerging jets

As stated in the opening paragraphs, the dark sector (DS) extension of the standard model postulates that there exists a separate sector of physics with its own set of gauge interactions separate from what is known as the existing standard model [12, 13]. Stable particles or stable states within the dark sector are potential dark-matter candidates. As astronomical and cosmological observations require that dark matter cannot have long-range self-interactions beyond gravity, new interactions within the dark sector will need some confining mechanism. For the remainder of this thesis, we shall assume that the dark sector exhibits an $SU(N_c^{\text{dark}} = 3)$, QCD-like interaction, with the number of dark sector fermions small enough that this interaction is also self-confining with asymptotic freedom. In such cases, dark fermions will be confined to dark mesons and dark baryons, with dark matter being the stable bound state that is protected from further decay by gauge symmetry requirements in the dark sector [15]. To allow such models to be accessible in collider experiments, a mediator particle X_{dark} that is charged in both the SM and dark sector needs to exist. We will work with the scenario where this mediator particle is charged in both the SM QCD and the dark sector QCD. In such cases, unstable dark mesons (π_{dark}) can decay back into QCD quarks, with the lifetime of dark mesons determined by the mediator particle's coupling strength to the two sectors. Given that this model describes the dark sector as being separated from the SM with interactions only possible through a high-mass mediator, this class of models is commonly referred to as the **hidden valley model**.

Another desirable feature of such extensions [13] is that they may give a natural explanation of the baryonic matter–dark matter energy density (Ω) differences observed in cosmological constraints [7]. The ratio of $\Omega_{\text{DM}} \sim 5\Omega_B$ can be explained by a process similar to that of baryo-

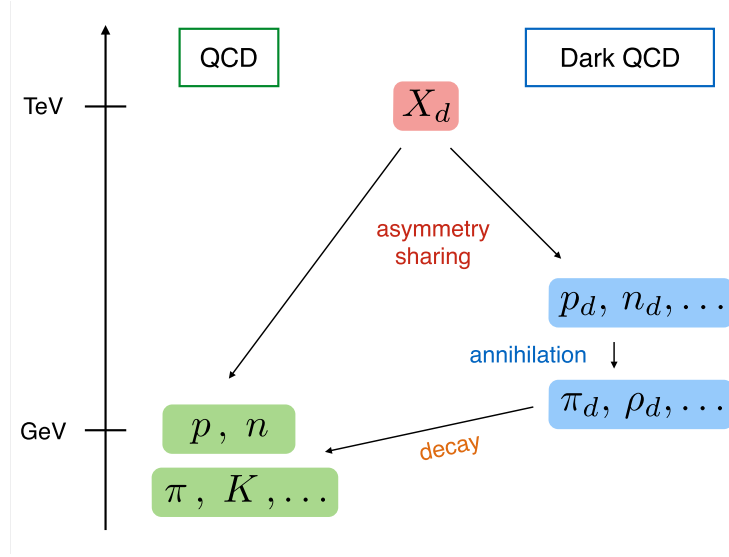


Figure 1.3: Schematic description of the energy scales of interest of the hidden valley model. Image taken from Ref. [15].

genesis [42] and early-universe nucleosynthesis [43], where the relative abundance of particles was determined by the thermal equilibrium of particle populations when the temperature of the early universe was much higher than the interaction energy threshold, then “frozen out” as the temperature of the universe dropped following its expansion.

As there are multiple possible scenarios of the dark-sector coupling within the constraints of $SU(N_c^{\text{dark}} = 3)$, we will focus on a subset of assumptions of the dark-sector interaction that gives rise to the phenomenological signature of interest, namely that the confinement scale of the dark sector Λ_{dark} is larger than the mass of most SM mesons, but still small compared with the energy scale of LHC collisions. In such a scenario, dark fermions will undergo a hadronization-like process within the dark sector, generating a shower of dark mesons. These dark mesons will then decay back into SM particles after, given the significant lifetime of the dark mesons, generating a second showering process [15]. For detectors that can only see SM particles, what is observed for the production of dark sector fermions is the production of a jet where the visible

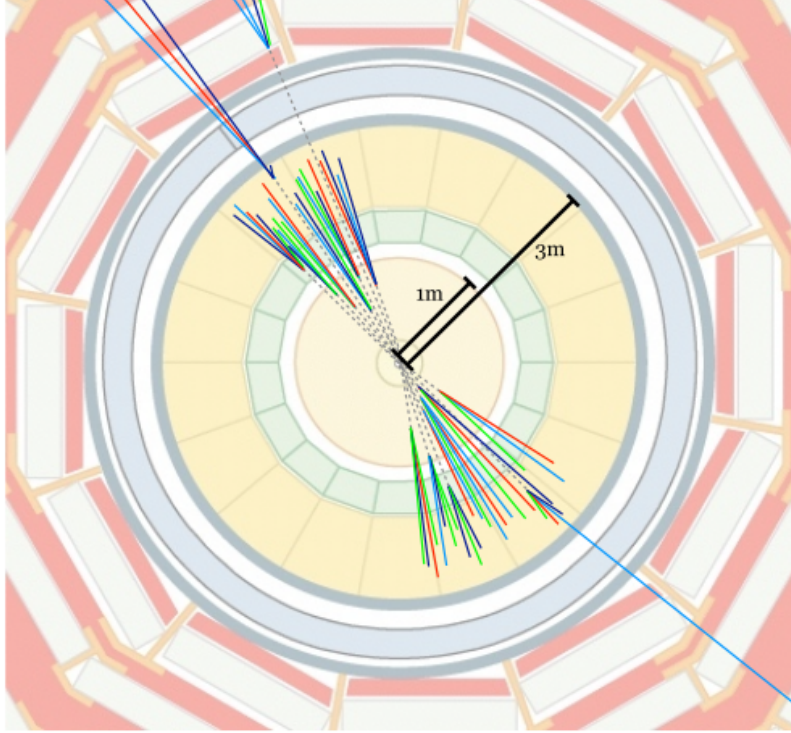


Figure 1.4: Schematic view of the back-to-back production of dark sector fermions, resulting in two emerging jets, as seen by detectors. Dark mesons generated in the dark sector hadronization shower are drawn as dashed lines and are invisible to the detector, while the SM particles are drawn in various colored lines. Figure taken from Ref. [15].

particles “emerge” from vertices a finite distance away from where the dark fermion is originally produced (Fig. 1.4). The search for such exotic jet signatures, which we call **emerging jets** (EMJs), will be the main topic of this analysis.

As dark-sector particles have not yet been observed, let us assume that the dark-sector fermions, and by extension, the dark mesons, are mass degenerate compared with the energy scales of interest. In such a case, the lifetime of dark mesons decaying back into SM particles is determined by the allowed coupling to the SM–DS mediator. For the simplest phenomenological case, the X_{dark} couple to only a single SM quark (using the SM down quark as an example), and

the decay amplitude of a dark sector meson can be expressed as [15]:

$$\Gamma_{\pi_{\text{dark}} \rightarrow d\bar{d}} = \frac{3\kappa^4 f_{\pi_{\text{dark}}}^2 m_d^2}{32\pi m_{X_{\text{dark}}}^4} m_{\pi_{\text{dark}}}, \quad (1.17)$$

where κ is the SM-DS mediator coupling strength, $f_{\pi_{\text{dark}}}$ is the dark meson decay constant, m_d is the mass of the SM down quark, $m_{\pi_{\text{dark}}}$ is the mass of the dark meson, and $m_{X_{\text{dark}}}$ is the mass of the mediator particle. All dark mesons have the same lifetime in such models, and the resulting EMJ signature has been sought by the CMS experiment using 2016 data [44].

A generalized description of how the SM and dark sector can be coupled by a portal mediator particle was explored in Ref. [45], and can be expressed via a Yukawa coupling:

$$\mathcal{L}_{\text{SM-DS}} = \kappa_{\rho i} X \bar{Q}_{\rho} q_i \quad (1.18)$$

where X is the SM-DS mediator field, $\bar{Q}_{\rho}$ is the dark sector fermion of flavor ρ , and q_i is the SM quark of flavor i . Using this expression, the decay amplitude of the dark meson composed of 2 dark sector fermions $Q_{\rho} \bar{Q}_{\sigma}$ decaying to SM quark pairs of flavor i and j can be expressed as:

$$\Gamma_{\rho\sigma, ij} = \frac{9N_{C_{\text{dark}}} m_{\pi_{\text{dark}}}}{8\pi m_{X_{\text{dark}}}^4} |\kappa_{\rho i} \kappa_{\sigma j}^*|^2 (m_i^2 + m_j^2) \sqrt{\left(1 - \frac{(m_i + m_j)^2}{m_{\pi_{\text{dark}}}^2}\right) \left(1 - \frac{(m_i - m_j)^2}{m_{\pi_{\text{dark}}}^2}\right)}, \quad (1.19)$$

with the $m_{X_{\text{dark}}}$ being the mass of mediator, m_i being the mass of the SM quark of interest and $m_{\pi_{\text{dark}}}$ being the mass of the dark meson. In this model, the dark sector mesons exhibit a lifetime spectrum driven by the mass spectrum of the SM quarks. To concisely distinguish the scenario where the mediator only couples to a single SM quark flavor, we will call models with

generalized coupling the **flavored scenario** (described using Eq. 1.19), and the models with the mono-flavored coupling the **unflavored scenario** (described using Eq. 1.17).

For the following discussion on the flavored scenario, we will limit the scope of the discussion to the case where there are exactly three dark sector fermion flavors, the $\kappa_{\alpha i}$ is a diagonal matrix with magnitude κ_0 (so $\kappa_{\alpha i} = \kappa_0 \delta_{\alpha i}$, with $\delta_{\alpha i}$ being the Kronecker delta matrix). For the theory to be consistent, this SM-DS couples to the three SM down-type quarks (quarks with electric charge -1/3). This class of dark sector models will be called the **flavor-aligned** scenario. In this case, the possible dark meson lifetimes are listed in Fig. 1.5: the significant disparity in the dark meson lifetime is driven by the mass gap between the SM bottom quark and the other two down-type quarks, while the “jumps” in dark meson lifetime as a function of dark meson mass are a result of allowed decay paths being kinematically impossible below the m_b threshold. This disparity in the dark meson lifetime will generate a mixed emerging jet signature in detectors: jets will be composed of both tracks consistent with prompt-like decays, and tracks from long-lived particle decays. At the same time, the dark mesons that decay prompt-like will contain b quarks and subsequently b mesons in their decay product, so the prompt-like signatures also contain long-lived signatures consistent with b meson decays. An example of the decay product composition for various signal samples under the flavor-aligned scenario can be found in Fig. 1.6.

Signatures for a flavored dark sector have yet to be studied and will require new methods of classification to study. In our studies, we limit dark meson masses to be greater than m_b . Phenomenologically, certain dark mesons become stable below this mass threshold, resulting in the particle shower generated by the dark sector fermions containing a significant fraction of particles invisible to typical particle detectors, requiring a different method of jet classifications (called “semi-visible” jets). Theoretically, such models are also excluded by the null observations

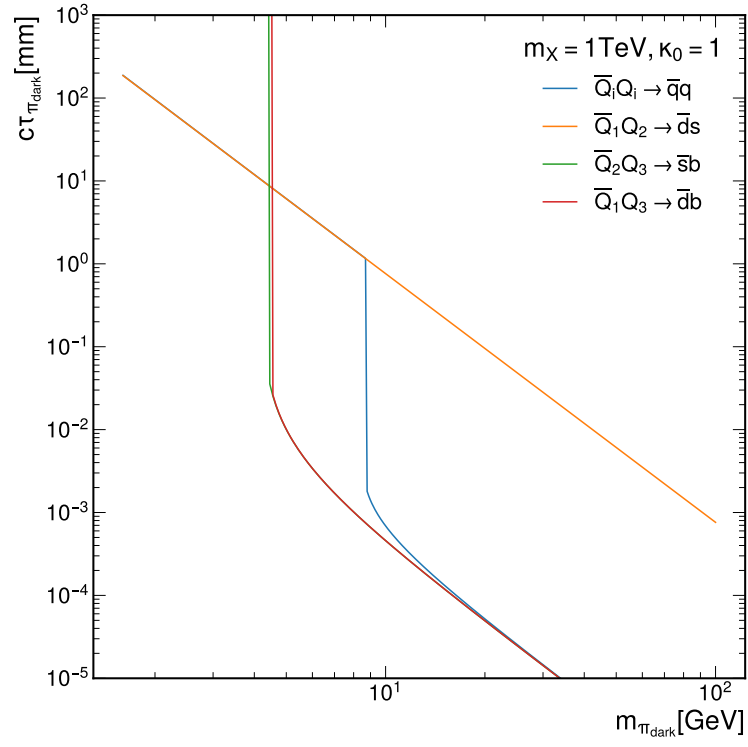


Figure 1.5: The calculated lifetime of dark mesons using Eq. 1.19 as a function of dark meson mass, assuming the flavor-aligned scenario with $\kappa_0 = 1$ and $m_{X_{\text{dark}}} = 1 \text{ TeV}$ as described in the main text.

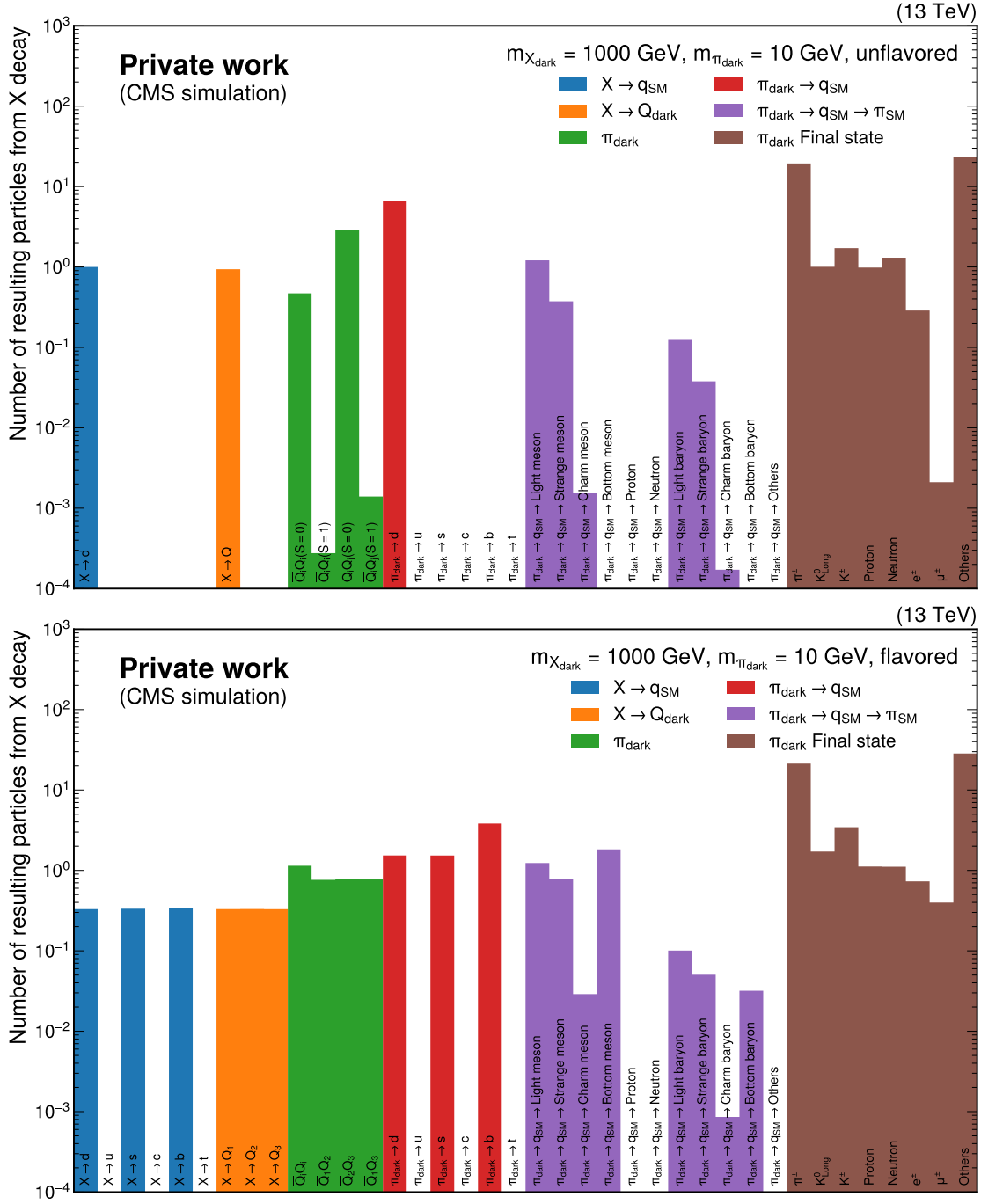


Figure 1.6: The decay products of the SM-DS mediator at various stages of the decay chain under the unflavored scenario (top) and the flavor-aligned scenario (bottom), with a dark meson mass of 10 GeV. The blue bars represent the immediate SM decay product of the X_{dark} , orange represents the immediate dark sector fermion decay products of X_{dark} , green shows π_{dark} of different compositions, red shows the immediate SM decay products of π_{dark} , purple shows the immediate SM meson decay products of π_{dark} , and brown shows the final state particles of π_{dark} decay. Dark quarks of different flavors are notated using the symbol Q_i .

of **flavor-changing neutral current** [46].

1.3 Overview of analysis strategy

Particle detectors are designed assuming the primary physics signature of interest is close to the hard-scattering interaction. Long-lived particle signatures, such as the production of dark meson showers, are then potentially discarded by the detector data acquisition systems as they appear to be “noise-like.” This analysis aims to present a robust strategy to search for emerging jets within the current constraints of the CMS detector.

For this search effort, we will focus on the pair production of the SM-DS mediator, with the SM-DS mediator then decaying to an SM quark and a dark sector fermion (see Fig. 1.7). The presence of the SM quark means that there will be clear SM signatures to reliably trigger the data acquisition of the detector. Since these production processes do not directly depend on the SM–DS mediator coupling, only on the SM strong coupling, the production cross sections of such processes are estimated using the available calculation of colored SUSY pairs [47] with an additional factor of 3 to account for the extra quantum numbers for the dark sector color charge. The final state of interest seen by the detector is the production of four high-energy jets, two of which should exhibit displaced particles from the long-lived dark mesons decaying back into SM particles (see Fig. 1.8). Given this constrain, both the unflavored and flavor-aligned scenario has 3 free parameters that can be adjusted: the mass of the SM–DS mediator $m_{X_{\text{dark}}}$, the mass of the dark meson $m_{\pi_{\text{dark}}}$, and the coupling strength of the SM–DS mediator that determined the lifetime of the dark mesons $c\tau_{\pi_{\text{dark}}}$. How this model parameters space is sampled will be discussed in Section 3.6.2.

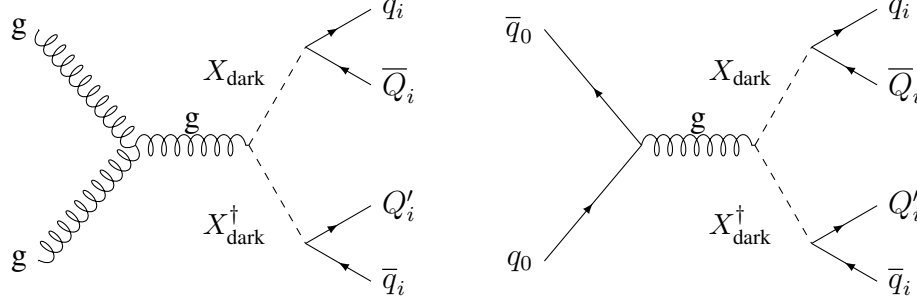


Figure 1.7: The Feynman diagrams of the production process of interest: the pair production of mediator particles via gluon-gluon fusion (left) and quark-annihilation (right) processes.

Having determined an event topology of interest, jet tagging can then be performed by checking the displacement of jet constituents to classify jets as “emerging-jet-like” to further isolate candidate signal events. The full description of how we perform jet-tagging for this analysis will be given in Chapter 4, as separate tagging observables best isolate the unflavored and the flavored dark sector scenarios. Since the jet tagging process can yield false positives, SM “background” events can still pass the event selection and jet-tagging criteria. As measuring the displacement of particles depends greatly on detector resolution, and is not necessarily modeled well by simulated events, the estimation of how many background events can pass the signal selection is best determined using data. How background estimation is performed using CMS data will be presented in Chapter 5.

The description of the CMS detector is given in Chapter 2, and the description of how detector signatures are processed into physics signatures is given in Chapter 3. Chapter 6 will present the various uncertainties, with the preliminary results of the analysis given in Chapter 7. The analysis presented in this thesis is fully documented and presented in Ref. [48].

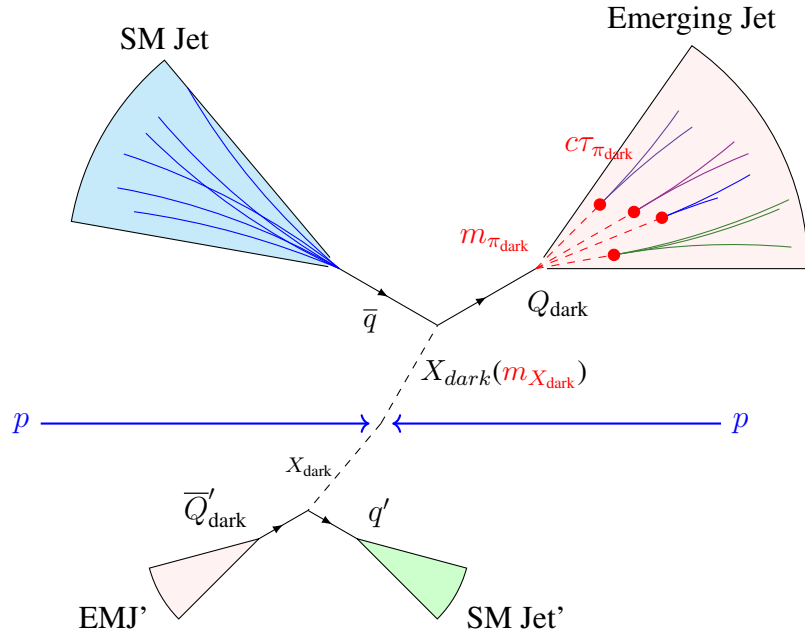


Figure 1.8: Primary event topology that will be sought by the analysis presented in this thesis. The curved lines in the jets represent the tracks of SM particles left in the detector system, and will be the primary signature that we will use to identify emerging jets. Parameters written in red are model parameters that we can explore.

Chapter 2: Experiment

This analysis will use the data collected by the CMS experiment for proton-proton collision delivered by the LHC with a center-of-mass energy of $\sqrt{s} = 13$ TeV during the Run-2 campaign of 2016–2018. This chapter will provide a broad overview of these two apparatus, highlighting the design and operation aspects that will be relevant to the analysis.

2.1 The Large Hadron Collider

The **Large Hadron Collider** (LHC) [49] is the world’s largest particle accelerator situated on Centre Européen pour la Recherche Nucléaire (CERN) campus crossing the France–Switzerland border with a circumference of 26.7 km. The LHC consists of 2 rings, each hosting a beam of ions circulating in opposite directions. Particle collisions occur at the designated **interaction points**, where the beam paths intersect. Two general-purpose experiments are built around the interaction points at LHC: “point 1” of the LHC houses the ATLAS experiment [50], while “point 5” on the opposite side of the accelerator ring houses the CMS experiment [51]. Both of these experiments are multi-purpose experiments aimed to cover as many physics signatures as possible. The asymmetric LHCb experiment [52] that is designed to measure b meson decay with high accuracy is located at “point 8”, while the ALICE experiment [53] designed to measure heavy-ion collisions is located at “point 2”.

The CERN accelerator complex *Complexe des accélérateurs du CERN*

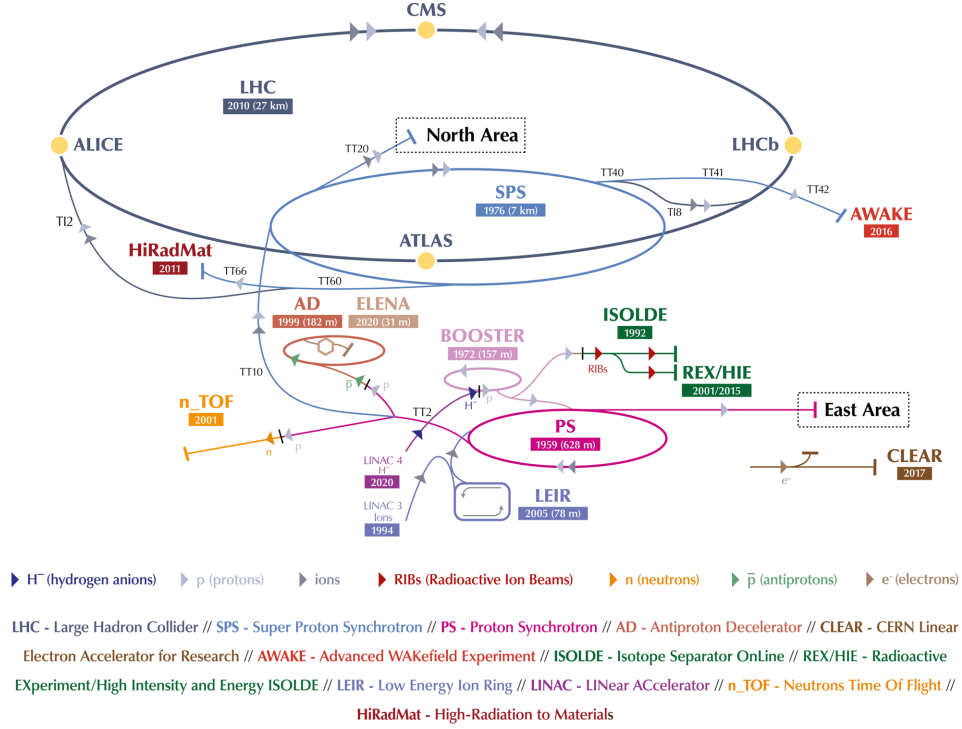


Figure 2.1: The accelerator complex at CERN. Figure taken from Ref. [54].

The LHC is the final stage in the acceleration complex at CERN, where the particles in the beam reach the energy of $6.5z$ TeV, with z being the atomic number of the nuclei used for the beam. Prior to the LHC, protons are accelerated by the Linear Accelerator 2, the Boost accelerator, the Proton Synchrotron, and the Super Proton Synchrotron, before being injected into the LHC rings at a particle energy of 450 GeV, before the final acceleration to 6.5 TeV is performed by the LHC. Figure 2.1 is the sketch of the acceleration complex at CERN.

When delivering proton-proton (pp) collisions, the protons in the LHC beams are grouped into **bunches**, each bunch containing around 10^{11} protons. A single bunch occupies ~ 1 mm in the transverse direction and ~ 1 cm along the beam direction. The minimum spacing between bunches is 25 ns (~ 7.5 m). This minimum spacing between bunches defines the maximum rate

at which collision can occur at the interaction points, and is used as the common clocking signal for data collection routines of the LHC experiment. Before entering the designated interaction point, the bunches are further focused to a transverse width of $\sim 50\mu\text{m}$ to increase the probability of high-energy pp collision. A collision **event** from the point of view of a detector is the record of all particles emerging from a single bunch-crossing. Within a single bunch-crossing event, multiple pp interactions can occur, with the highest energy transfer collision referred to as the “hard scattering” event, and other interactions are referred to as **pileup interactions**.

The scattering amplitude of high-energy processes directly corresponds to the number of events expected to occur throughout the operation of the LHC. The **instantaneous luminosity** of the LHC measures the rate at which processes can occur, which can be defined as:

$$\mathcal{L}_{\text{inst.}} = \frac{N_p^2 n_b f}{4\pi\sigma_x\sigma_y} F, \quad (2.1)$$

where N_p is the number of protons in a bunch, n_b is the total number of bunches in the beam, f is the circulation frequency of the LHC beam, σ_x, σ_y is the transverse size of each bunch, and F is a geometric correction factor for the bunch crossing configuration. The total number of events that can occur for some physics process of interest, N_i , can then be calculated as:

$$N_i = \sigma_i \mathcal{L}_{\text{int.}} = \sigma_i \int_T \mathcal{L}_{\text{inst.}} dt \quad (2.2)$$

where σ_i is the total cross section of the physics process of interest, $\mathcal{L}_{\text{int.}}$ is the **integrated luminosity**, and T is the time range over which data of interest is collected. To look for rare physics events within a reasonable time frame, the instantaneous luminosity needs to be high, so luminos-

ity is the key target for the operation of the LHC. In 2018, the LHC achieved a peak instantaneous luminosity at point 5 of 20 Hz/nb ($2 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$), delivering a total integrated luminosity of 138 fb^{-1} between 2016–2018 [55, 56, 57]. How the luminosity is measured for the condition at the point 5 interaction point will be further discussed in Section 2.2.7.

2.2 The Compact Muon Solenoid Detector

The **compact muon solenoid** (CMS) detector [51] is designed to be a general-purpose hermetic detector, aiming to measure the properties of all particles that emerge from the hard scatter with high precision by enclosing the interaction point with as much angular coverage as possible. CMS is designed around the superconducting solenoid to provide a uniform magnetic field in parallel with the beam line, which allows for high-precision momentum measurements of charged particles emerging from the hard-scattering process. The CMS detector is comprised of various subsystems, each targeting a specific physics signature of the high-energy particles. Most particles are expected to be captured in the compacted volume within the solenoid, with muon being the only type of particle that can leave a significant signature outside the solenoid.

This section will give a brief overview of the various subsystems in CMS and highlight key features that will be further discussed in the search for dark sector showers.

2.2.1 Coordinate system

The origin of all measurements is centered around the **beam spot**, the intended interaction point of the LHC beams. The direction $\hat{z}$ is defined along the beam direction in the clockwise rotation as seen from above the LHC ring, $\hat{x}$ point towards the center of the LHC ring, and $\hat{y}$ is

CMS DETECTOR

Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T

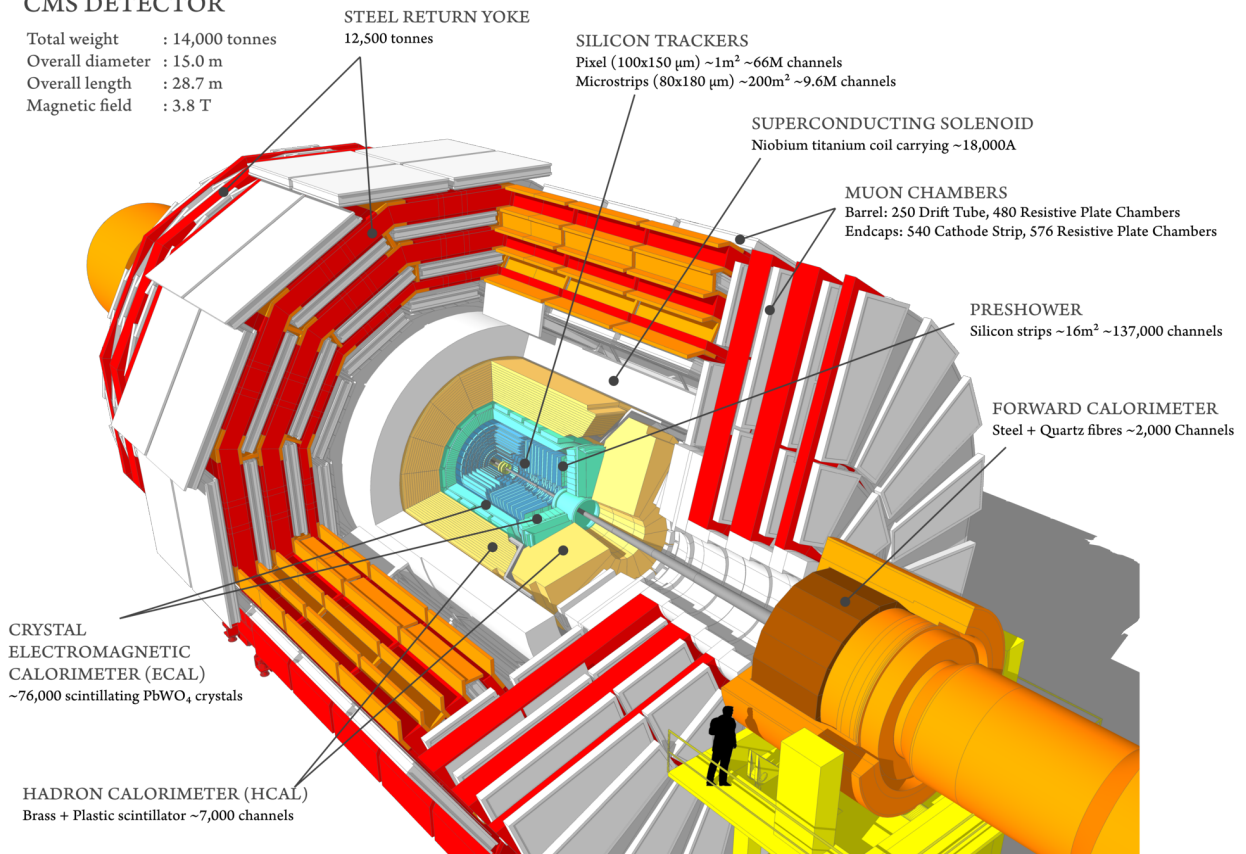


Figure 2.2: Schematic of the CMS detector and its various subsystems. Image taken from Ref. [51].

defined using the right-hand rule.

Given the layout of the detector around the beamline and the directionality of the pp collisions, measurements are most commonly expressed using a modified cylindrical coordinate system (r, η, ϕ) , where r is the magnitude of the measurement projected onto the x - y plane, ϕ is the usual azimuthal angle, and η is the **pseudorapidity** defined as the transformation of the polar angle θ :

$$\eta = -\ln \left(\tan \frac{\theta}{2} \right). \quad (2.3)$$

The reason for using η instead of θ is that since the hard-scattering process occurs between the proton fragments, the collision is expected to be boosted in the $\hat{z}$ direction, and θ is not a Lorentz invariant that should be used to compare objects within events. A Lorentz invariant angular measure that can be used for event and object selection is **rapidity** y :

$$y = -\ln \left(\frac{E + p_z}{E - p_z} \right), \quad (2.4)$$

though this relies on evaluating the full 4-momentum to be properly evaluated. One can show that in the high energy limit of $E \gg m$, the value of y will converge to the pure angular measurement η . When comparing angular separations between 2 physics objects, i and j , seen in the events, the most commonly used measure the separation in the (η, ϕ) space is:

$$\Delta R(i, j) = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}, \quad (2.5)$$

where the differences in ϕ take into account the wrap-around for ϕ .

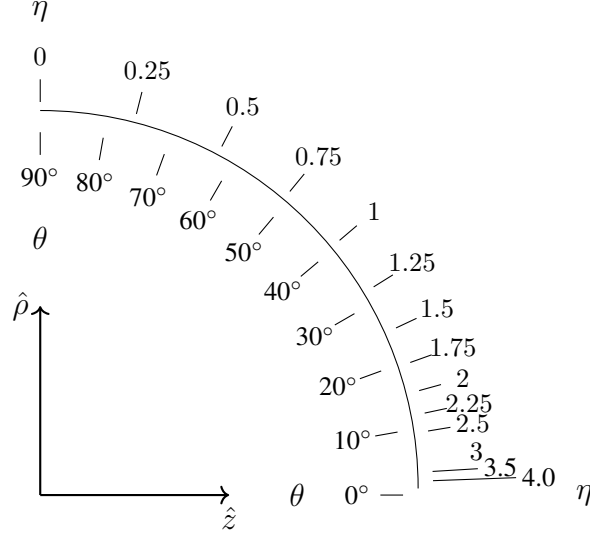


Figure 2.3: Example of converting from polar angular θ to pseudorapidity η defined as $\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right)$.

The momentum of particles will typically expressed using the components: (p_T, p_η, p_ϕ, E) , where p_T is the **transverse momentum**, the momentum component that is transverse to the beamline, p_η and p_ϕ are used to represent the angular direction of the momentum in (η, ϕ) space, and E is the total energy. In vectorial calculations, we will also use $\vec{p}_T$ to represent the projection of the momentum onto the x - y plane. While the technical unit of momentum should be GeV/c, for this thesis, we will follow the collaboration convention of simply GeV for brevity. Similarly, the unit for mass will also be abbreviated to simply GeV.

2.2.2 Magnet configuration

The solenoid magnet of CMS [58] uses over 2000 turns of niobium-titanium alloy wires cooled to superconducting temperatures, with each turn carrying over 18kA of current to generate a maximum 3.8 T magnetic field within the solenoid volume. The magnet has an inner diameter of 6 m and a length of 13 m. The 3.8 T magnetic allows for measuring particles with momentum

up to $p_T \sim 1$ TeV to a precision of 10% relative uncertainty. Additional steel return yokes are placed outside the solenoid to contain the magnetic fields, which also provides structural support to the muon detector systems.

2.2.3 Tracking subsystems

The tracking subsystems [59] attempt to record the spatial position of particles traversing the detector volume while perturbing the particle as little as possible. The path of the particles can then be reconstructed using the recorded information. These detector systems usually require the finest detector elements to achieve high spatial resolution, which makes such systems both expensive to manufacture and difficult to maintain, given the number of channels required to cover some detector volume. Because of this, the tracking systems are placed closest to the beamline, where spatial resolution is critical for reconstructing the position of the pp collisions and the required detector volume is the smallest. The resolution of the tracking subsystems is critical for the displaced signatures that characterize the various EMJ models that this analysis looks for.

Both the pixel system and the silicon strip system are silicon-based detectors and perform particle detection using a similar principle: a p-n junction is held in a high-voltage reverse bias, such that when a charged particle passes through the depletion region, the charged particle induces electron-hole production in the depletion region, generating a large measurable current. The spatial resolution of the detector systems is then determined by the geometrical configuration of the p-n junctions, which is distinct for the pixel and silicon strip detector, as will be discussed in the following section.

2.2.3.1 Pixel system

The pixel detector is a true 2-dimensional detector: for the silicon substrate, the active elements are arranged in a dense grid of **pixels**, each $100\mu\text{m} \times 150\mu\text{m}$ in size. A single pixel module contains the silicon substrate with an active area of $16.2\text{ mm} \times 64.8\text{ mm}$, along with the additional electronics to perform power delivery, amplification, digitization, and data readout. For the region closest to the interaction point, pixel modules are arranged into concentric cylinders (so-called “barrel pixels” or “BPIX”s), with each module overlapping with its neighboring modules to ensure total angular coverage. Further from the interaction point at the end of the BPIX cylinders, pixel modules are arranged in an overlapping fan-like configuration to provide coverage for the forward regions (so-called “forward pixels” or “FPIX”s). The modules are supported by a carbon fiber support structure, which also carries the wiring required for power delivery, data readout and cooling.

Between 2016 and 2017 [60], the pixel system was upgraded to include more layers, with the innermost layer placed significantly closer to the beamline, resulting in higher vertex reconstruction resolution. A comparison of the pixel geometry can be found in Fig. 2.4. For the rest of this thesis, these two configurations will be referred to as the pixel Phase-0 and Phase-1 geometries respectively.

2.2.3.2 Silicon strip system

For the silicon strip detectors, the p-n junctions are arranged as 1-dimensional strips along the silicon substrate, varying from $80\mu\text{m}$ to $205\mu\text{m}$ in width. This configuration allows for high-resolution measurements to be performed in the direction perpendicular to the strip direction,

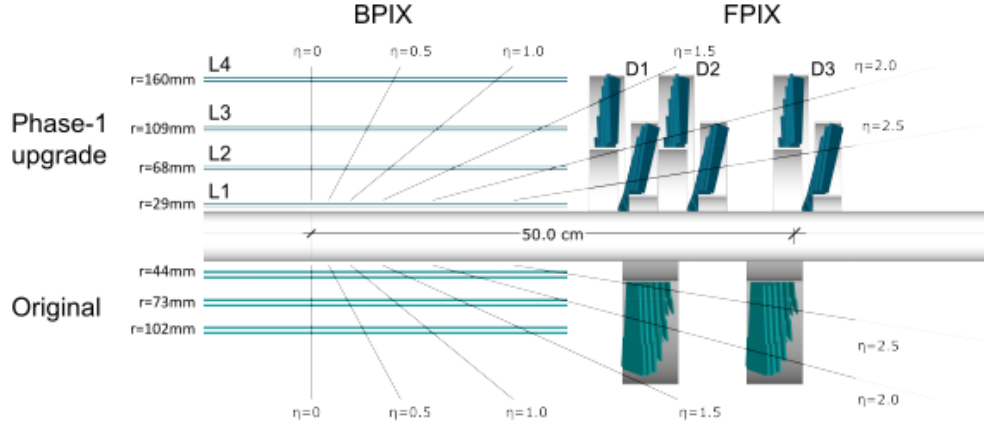


Figure 2.4: Comparison of the geometric layout differences between the pixel Phase-0 geometry (bottom) and the pixel Phase-1 geometry (top). The FPIX diagram also depicts the staggered fanning assembly used for the pixel modules. Figure is taken from Ref. [60].

using the readout of neighboring tracking systems modules to infer spatial information along the strip direction. Using strips allows for high-resolution measurements without overwhelming the data throughput with full two-dimensional (2D) spatial data, while also keeping the cost of the tracker volume down. As with the pixel detector, silicon strip detectors are composed of modules consisting of the silicon substrate housing the active elements with the supporting electronics, and are arranged into concentric cylinders and fanning endcaps, and supported using a carbon fiber support system. Special double-sided modules are installed closest to the interaction so that high-resolution pseudo-2D data can be measured.

The assembly of the silicon strip systems can be further separated according to their assembly geometry. The inner barrel (TIB) and inner disks (TID) have modules in each layer overlapping with their neighbors to ensure angular coverage. The outer barrel (TOB) and the outer endcap (TEC) have their module positions staggered between neighboring layers to provide angular coverage.

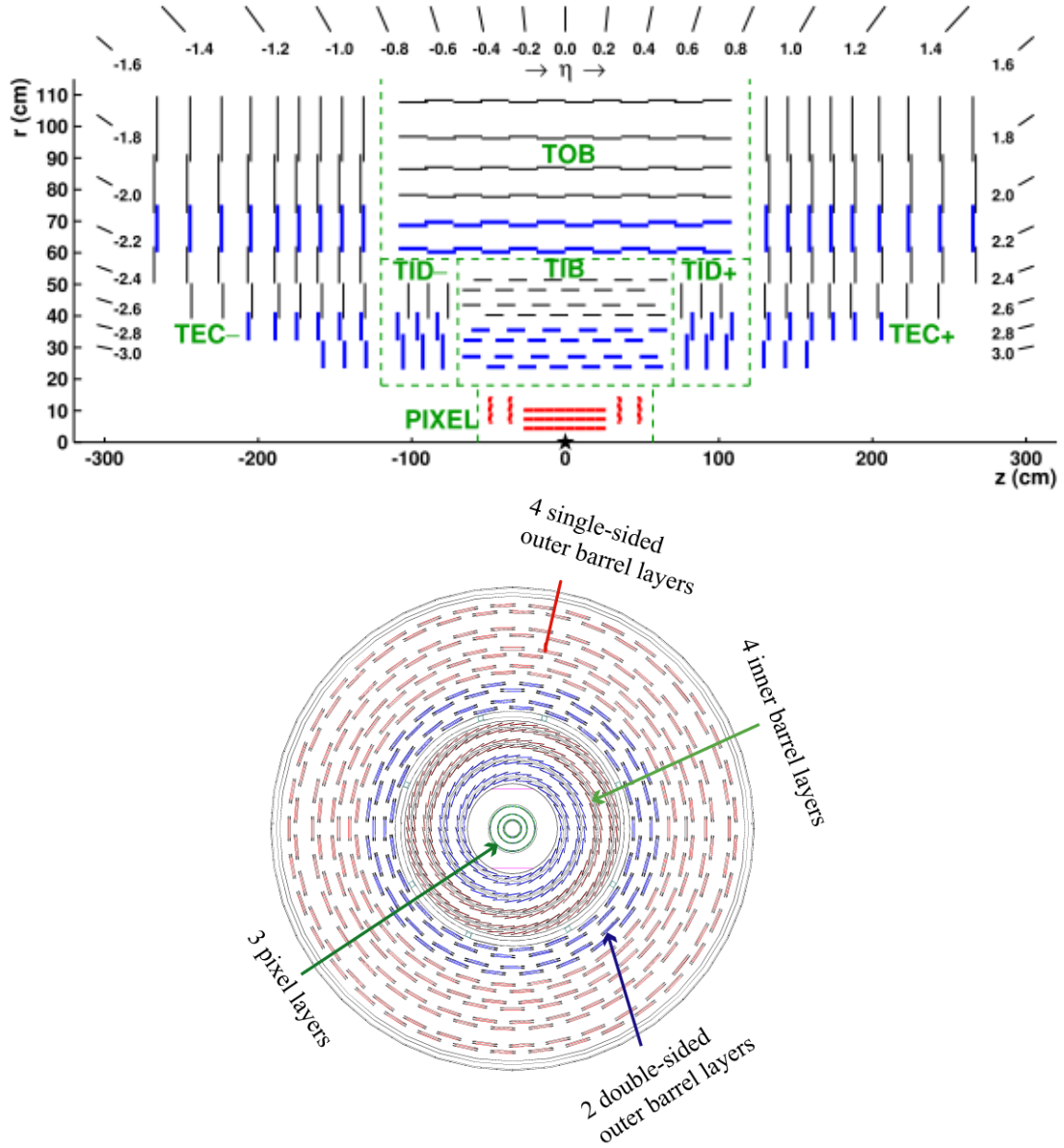


Figure 2.5: Layout of the pixel and tracker detector modules in the Phase-0 geometry, depicted in r - z space (top), and a slice in the barrel region (bottom). The staggered layout between modules in different layers ensures optimal angular coverage around the interaction point. Top image taken from Ref. [51]. Bottom image taken from Ref. [61].

An example of the geometric layout of the tracking system can be found in Fig. 2.5. Overall, the pixel system in the CMS detector extends to a radius of about 10.2 cm in the Phase-0 configuration, 16.0 cm for the Phase-1 configuration, and has a total surface area $\sim 1 \text{ m}^2$ covered by $\sim 66\text{M}$ readout channels; the silicon strip system extends to a radius of 110 cm, and covers a total surface area of $\sim 200 \text{ m}^2$ using $\sim 9.6\text{M}$ readout channels. Both systems cover the interaction point up to $|\eta| < 2.5$ compared with the beam spot.

2.2.4 Calorimeter subsystems

Past the CMS tracking systems, the calorimeter systems are designed to measure and stop all non-muon, non-neutrino particles generated in the hard-scattering processes. This is done by placing dense materials such that the high-energy particle entering the materials sheds energy by generating a shower of low-energy particles. The low-energy particles generate measurable signals proportional to their number, with the number proportional to the energy of the original incoming particle. Two distinct technologies to perform calorimetry in CMS.

2.2.4.1 Electromagnetic calorimeter

The **electromagnetic calorimeter (ECAL)** [62] is a **homogeneous calorimeter**, meaning that the material used for stopping the incoming particles and the material used for particle-shower energy measurement is the same. The ECAL uses lead-tungstate (PbWO_4) crystals to stop light particles that participate in electromagnetic interaction (electrons and photons) via the generation of electromagnetic showers (generation of breaking radiation photons, photons converting to electron–positron pairs). The secondary particles produced in the electromagnetic

shower generate scintillating light, which is then measured using avalanche photodiodes (APD) placed behind the crystal. The amount of scintillating light measured is used to infer the energy of the incident particle.

The design of the crystals and their assembly in ECAL is based on the following two electromagnetic showering properties of lead-tungstate [63]:

- **Radiation length** (0.89 cm). This is the distance a high-energy electron is expected to traverse in lead-tungstate before shedding $1/e$ of its energy.
- **Moliere radius** (2.2 cm). This is the expected range of the electromagnetic shower perpendicular to the incident particle.

Individual crystals produced in $2.2 \times 2.2 \times 23 \text{ cm}^3$ blocks in the low- η barrel region (ECAL barrel, EB), and in $3.0 \times 3.0 \times 22 \text{ cm}^3$ blocks in the high- η endcap region (ECAL endcap, EE), arranged such that the smaller face of each crystal faces the intended interaction point. This configuration roughly translates to an angular resolution of 0.0175×0.0175 in (η, ϕ) in the barrel region, and an angular resolution of 0.05×0.05 in the endcap region. An example of the geometric layout of the ECAL assembly is shown in Fig. 2.6. The high refractive index of lead-tungstate ($n=2.29$) means the generated scintillating light are guided to the avalanche photodiodes at the far side of each crystal via total internal reflection. The scintillating light has wavelengths between 380–550 nm, and over 90% of the light is expected to be released within 25 ns of the incident particles arriving at the crystal face.

An additional silicon-based preshower detector (SE) with a pitch size of 1.9 mm is installed in front of the ECAL endcap crystals to boost the angular resolution in the EE system. The greater resolution allows for the distinguishing of photons generated in the hard scattering processes from

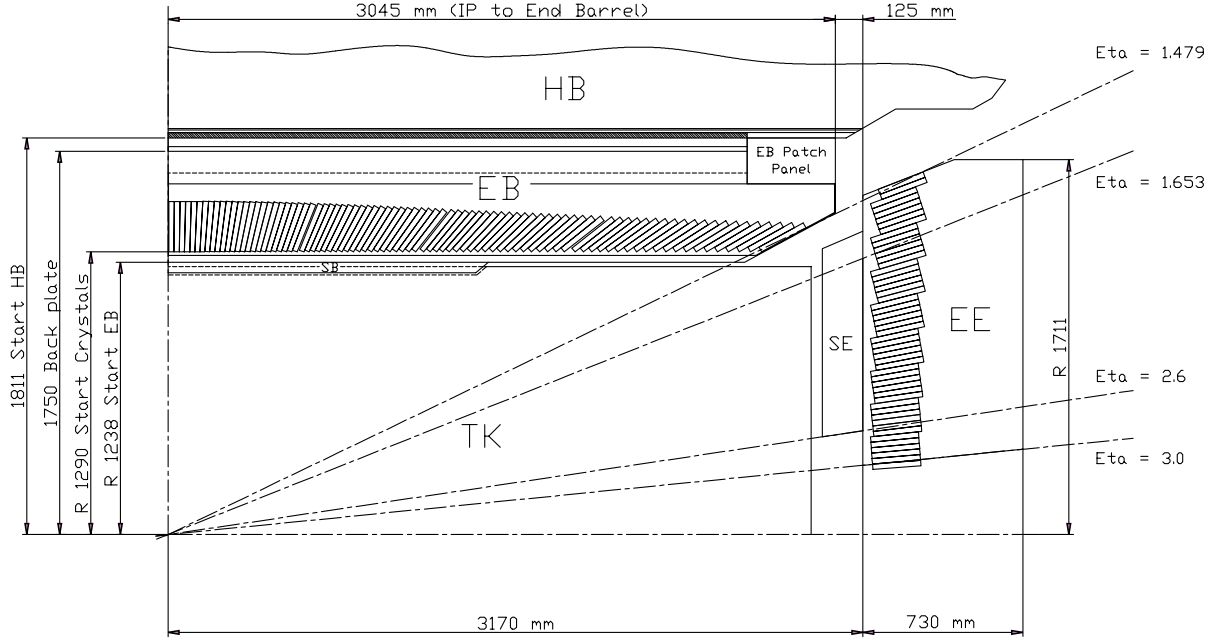


Figure 2.6: Schematic of the EB, EE, and the preshower assembly a quarter segment of CMS, projected in r - z space. Figure is taken from Ref. [62].

photons generated in in-flight particle decays (most commonly $\pi^0 \rightarrow \gamma\gamma$ processes).

Between the barrel–endcap geometry transition, there is a small gap at $\eta \in [1.4476, 1.5564]$ where crystals are absent. This region is typically excluded from analysis if precise energy resolution is required.

2.2.4.2 Hadronic Calorimeter

The **hadron calorimeter (HCAL)** [64] is a **sampling calorimeter**: copper and steel layers are used as the **absorber** material used to stop the high-energy incident particles, and scintillating plastic tiles are placed behind the metallic layers to measure the particle shower generated by the absorber material. Multiple layers are stacked to ensure the instance particle can be fully stopped in the detector material. Sampling calorimetry is a more cost-effective method for stopping high particles than indefinitely extending the ECAL crystal volume. Since the particle shower is spread

across multiple layers of material, the light generated in the plastic scintillators is collected via wavelength-shifting fibers, and the multiple fibers of tiles in the same **towers** are passed to the same silicon photomultiplier to be collectively measured. An example of the depth segmentation of the readout across HCAL towers can be found in Fig. 2.7. The geometrical configuration of the towers is distinct for the HCAL barrel (HB) and HCAL endcap (HE) systems. As high-energy particle showers may penetrate the magnet, additional HCAL layers are placed outside the magnet and are called the HCAL outer system (HO).

An additional forward HCAL system (HF) is installed in the region $|\eta| \in [3.0, 5.0]$. While it is also a sampling calorimeter, the sample technology for HB HE and HO cannot be used in this region due to the high radiation. The HF is composed of a steel absorber with longitudinal quartz fibers, and operates by measuring the Cerenkov radiation generated by the particle shower passing through the quartz fibers.

2.2.5 Muon subsystems

Due to their high mass and since they do not participate in the strong interaction, muons are expected to penetrate all detector material. Given this unique property, muons will generate the clearest signature for high-energy events. The muon subsystems [66] aim to perform precision measurements of muon trajectories outside the outer HCAL system and the solenoid magnet. Being the system that is furthest away from the collision point, the design of the muon detectors must be volumetrically cheap, while also being mechanically robust.

The sensor operating principle for the various muon subsystems is similar to the tracker systems: anodes and cathodes are maintained at high voltage differences, with high-energy muons

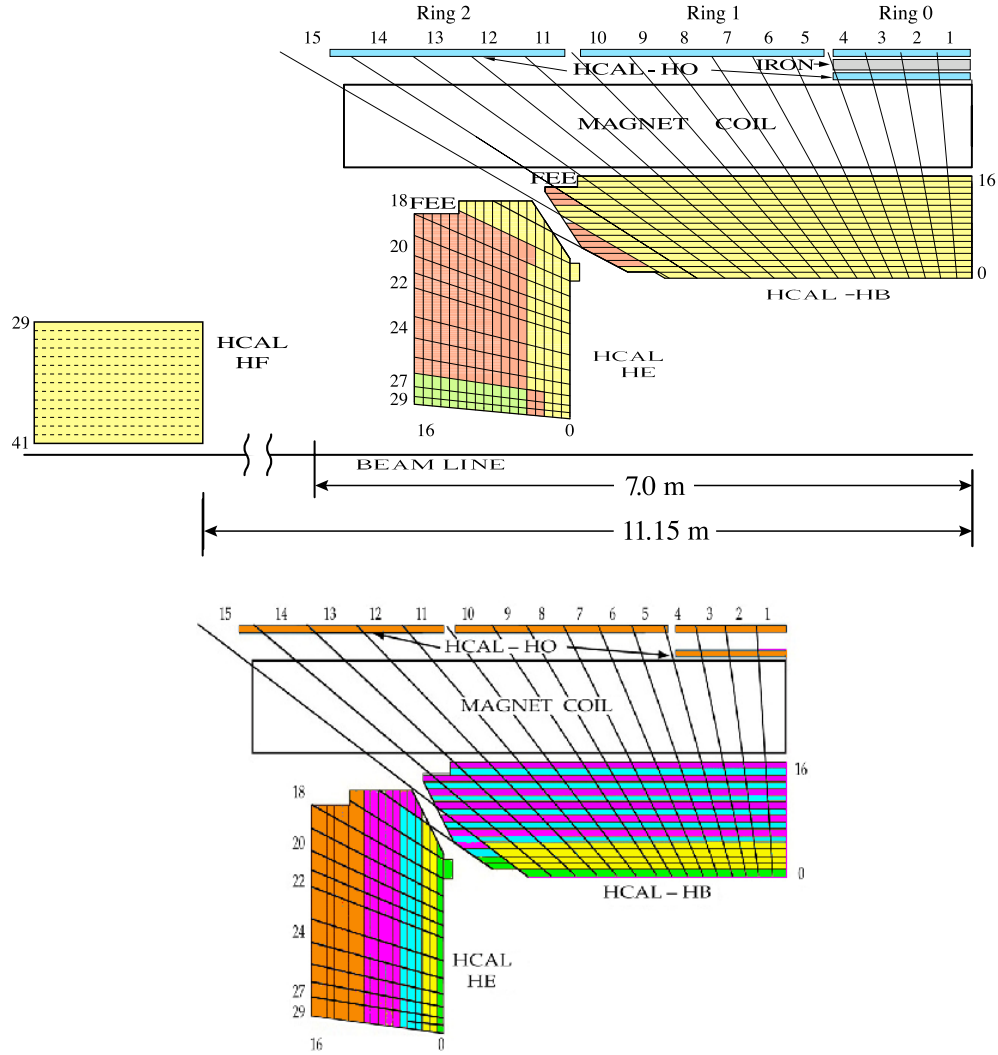


Figure 2.7: Schematic of the HCAL layout, with the different colors used to indicate the depth segmentation of the readout. The top figure shows the depth segmentation used for Phase-0, and the bottom figure shows the updated depth segmentation in the HB and HE systems after the Phase-1 upgrade. Figures taken from Ref. [65]

inducing a measurable current between the anode and cathode. Instead of silicon, the muon subsystems use conducting anodes and cathodes separated by some inert gas mixture. Muons passing through this sensor would ionize the gas, generating a drift current that can be measured. The different muon subsystems are distinct in the geometric layout and function of the electrode configuration (depicted in Fig 2.8), leading to differences in the operating voltage and the inert gas mixture used that could best utilize the electrode configurations.

Drift tubes (DT) are the primary technology used for the barrel region, where muon rates are expected to be lower. The anode is designed as a thin wire that runs along a chamber, with cathodes at a separation of ~ 2 mm from the wires. The thin wire amplifies the voltage gradient close to the wire, allowing for high detection efficiency at low working voltages. By facing the cathode–anode–cathode face direction towards the interaction point, the timing response when the currents are measured at the cathode and anode can extract spatial information of higher precision than the anode–cathode separation.

Cathode strip chambers (CSC) are the primary technology used for the endcap region. Cathodes are arranged in parallel strips running in the radial direction of the endcaps on the outer shell of the chamber. Anodes are wires suspended in the center of the chamber perpendicular to the strip direction running in the azimuthal direction. Spatial position can be extracted by comparing the current readout in the cathode and anode, meaning that the CSC system is not dependent on drift velocity to achieve high spatial resolution, and less susceptible to the changing of magnetic fields, making it suitable for the endcap region, where a higher muon rate is expected, and the magnetic field is less confined.

Resistive plate chambers (RPC) are used in both the barrel and endcap regions to provide high-resolution timing information to the DT and CSC systems by having the anode–cathode configuration in parallel plates with a small separation. As there is very little field variation compared with the wire-like anode configuration, the precise timing of when a muon passes through an RPC can be determined. Spatial information is inferred by further splitting the cathode into multiple strips. To prevent single muons from disrupting the field in the whole chamber, resistive plastic is placed in front of the anode and cathode pair. This configuration has the effect of the drift current temporarily disabling the field in the region where the muon passes through the chambers [67], allowing other regions of the chamber to continue operations nominally. Without the field amplification effects of a wired anode, the RPCs operate at a higher voltage difference than the DT and CSC systems.

The schematic of the muon system can be found in Fig. 2.9. RPC layers are placed in front of the DT and CSC segments to provide accurate timing information. The chambers are inter-joined with a steel return yoke, which serves as the material to regulate the magnetic field outside the superconducting solenoid ring and the structure support of the muon systems.

2.2.6 Trigger

At each LHC bunch crossing, the data volume required to record all detector responses takes up to 1 MB. The maximum LHC bunch crossing rate of 40 MHz corresponds to a data throughput of 40 TB/s, which is far too high to be reasonably stored. The **trigger system** is implemented to flag which events should be stored by comparing the detector responses with what is expected of a physics process of interest, and exclude the uninteresting “soft” scattering

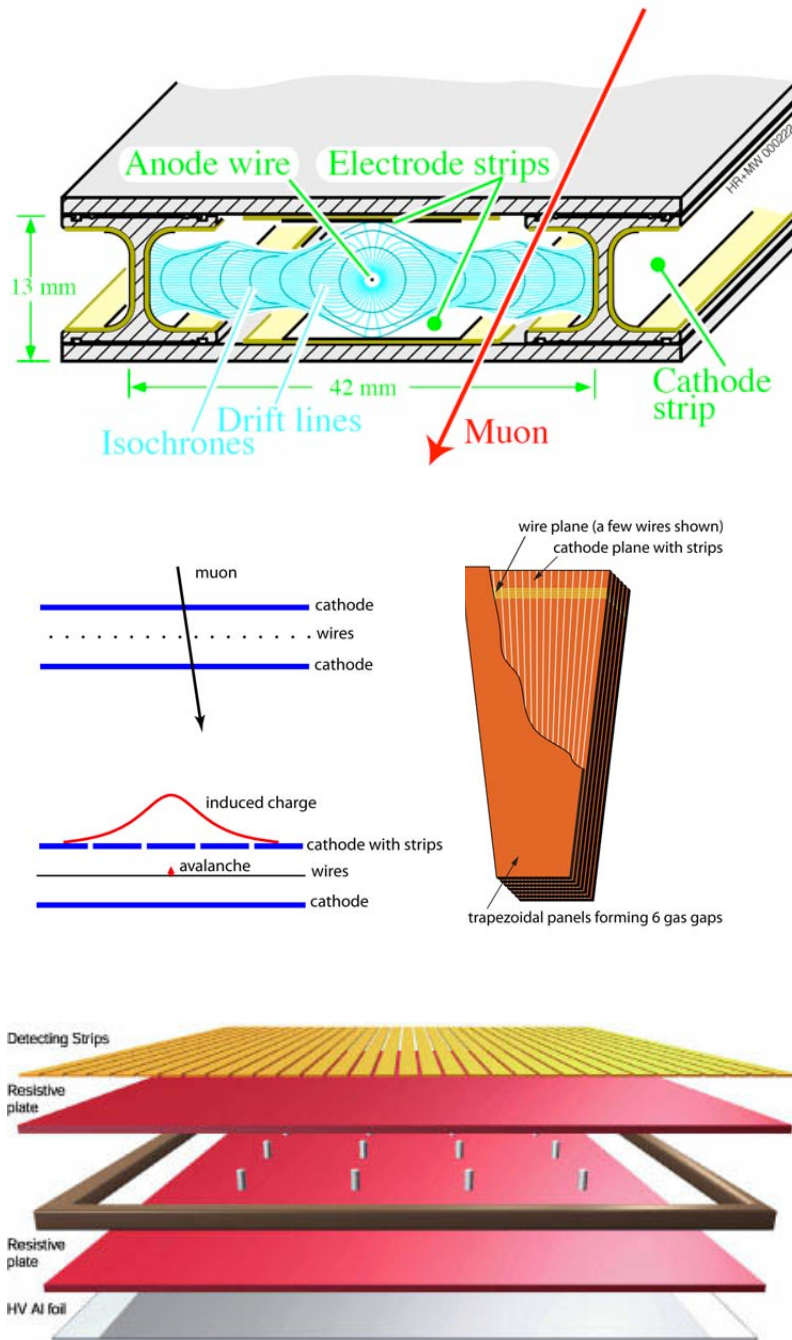


Figure 2.8: Sketches of the anode–cathode layout in the DT system (top), CSC system (middle) and the RPC system (bottom). Figures for the DR and CSC systems are taken from Ref. [51]. Figure for the RPC system is taken from Ref. [68].

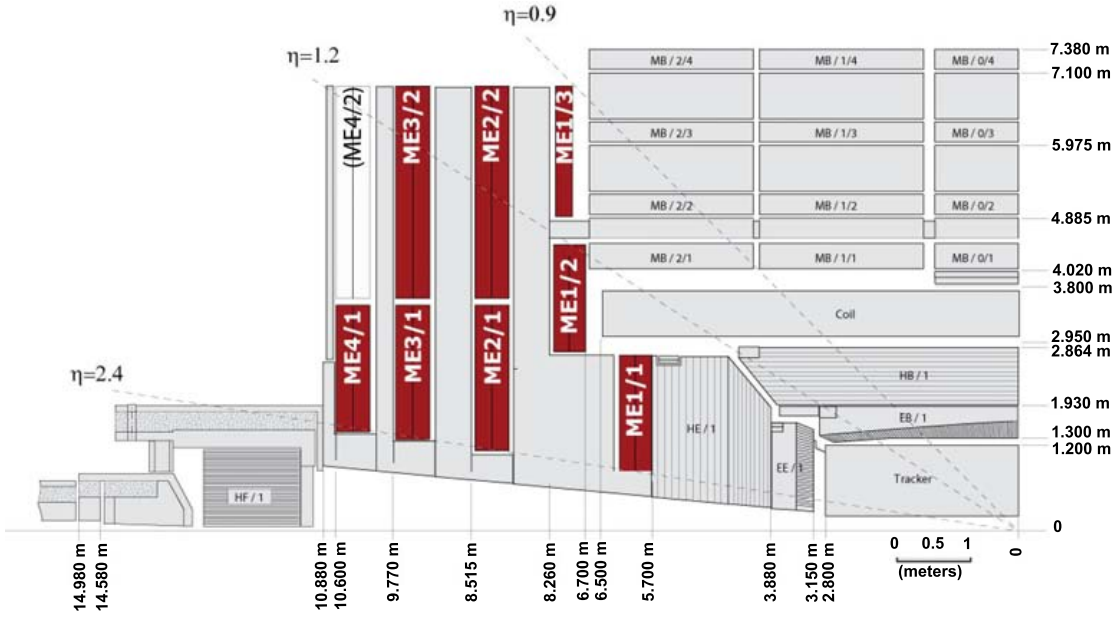


Figure 2.9: Schematic of the muon system geometry compared with the rest of CMS. The muon systems are located in the gray “MB” and the red “ME” blocks. Figure taken from Ref. [51].

events.

The **level-1 trigger** system (L1T) [69] is a hardware/firmware-based system: the readout systems of each subsystem module will compare its readout results with neighboring modules to generate keep-pass logic decisions, which are then passed to a global trigger system to compare logic from multiple subsystems to determine whether the event should be stored or not. The L1T system is only implemented in the calorimeter systems and the muon systems, as the data comparison in the tracker system is too complicated to perform on a hardware/firmware level. The event rate of events passing L1T is ~ 100 kHz.

The **high-level trigger** (HLT) [70] is a software-based system that operates on the entire detector response that passes the L1T. The reconstruction of physics objects passing through the detector is performed to check if there are physics objects of interest in the event and whether they should be stored (the full description of physics objects reconstruction is discussed in Section 3). The HLT system deploys a slightly simplified reconstruction algorithm to ensure the HLT

decisions can be made within 50 ms to avoid the data throughput being overwhelmed. The final event rate that passes the HLT is ~ 1 kHz.

While events passing the HLT trigger typically require some high- p_T objects to exist in the events, low- p_T objects are still used for the study of low-mass SM particles. To allow for studying low- p_T objects while keeping the data collection within reasonable bounds, low- p_T trigger decisions can be configured to be “prescaled” by some factor of N , meaning that such triggers only fire every N events.

2.2.7 Luminosity measurements

The luminosity of the pp collision at the CMS interaction point needs to be measured to ensure analyses can accurately scale the number of expected events to occur within the span of the detector operation period. Using instrumentation within the CMS, relative luminosity can be measured by comparing the relative detector activity, which is expected to be proportional to the collision rate.

The **pixel luminosity telescope** (PLT) [71], **fast beam condition monitor** (BCM1F) [72], and a dedicated readout of the HF system are used to measure a per-bunch crossing luminosity, regardless of what may happen to the various detector subsystems during data collection. The PLT uses the same pixel sensors as the main pixel modules but is assembled with specialized readout chips to operate at the full 40 MHz data rate. The BCM1F is a diamond-based semiconductor sensor. Both the PLT and the BCM1F are placed close to the beamline slightly beyond the pixel volume. Both detectors use zero-counting as the primary method of measuring luminosity: to avoid occupancy effects impacting the measured relative activity in the detector, it exploits

the feature that detector activity is expected to follow a Poisson distribution, and the probability of measuring 0 events is sufficient to calculate the Poisson mean μ . Similar techniques can be performed using the main CMS detector subsystems including the pixel and the DT systems, but because the HLT trigger system limits these detector readouts, only the integrated luminosity can be determined using these systems.

The detector-based methods all measure relative luminosity, as it effectively compares relative detector activities during operation to infer instantaneous collision rates. Absolute luminosity needs to be done by a **van-der Meer scan** [73], where the 2 LHC beams need to be shifted independently in the x and y axes, respectively. As the beam positions and beam profiles are monitored extensively for LHC operations, the shifting of the beam position changes the collision rate by a known amount, and can then be used to infer the total luminosity during nominal operations.

An example of the luminosity measured at CMS is displayed in Fig. 2.10, where the delivered luminosity is the aggregated luminosity measured by CMS luminometer systems, and the recorded luminosity is the fraction during which the CMS collected data.

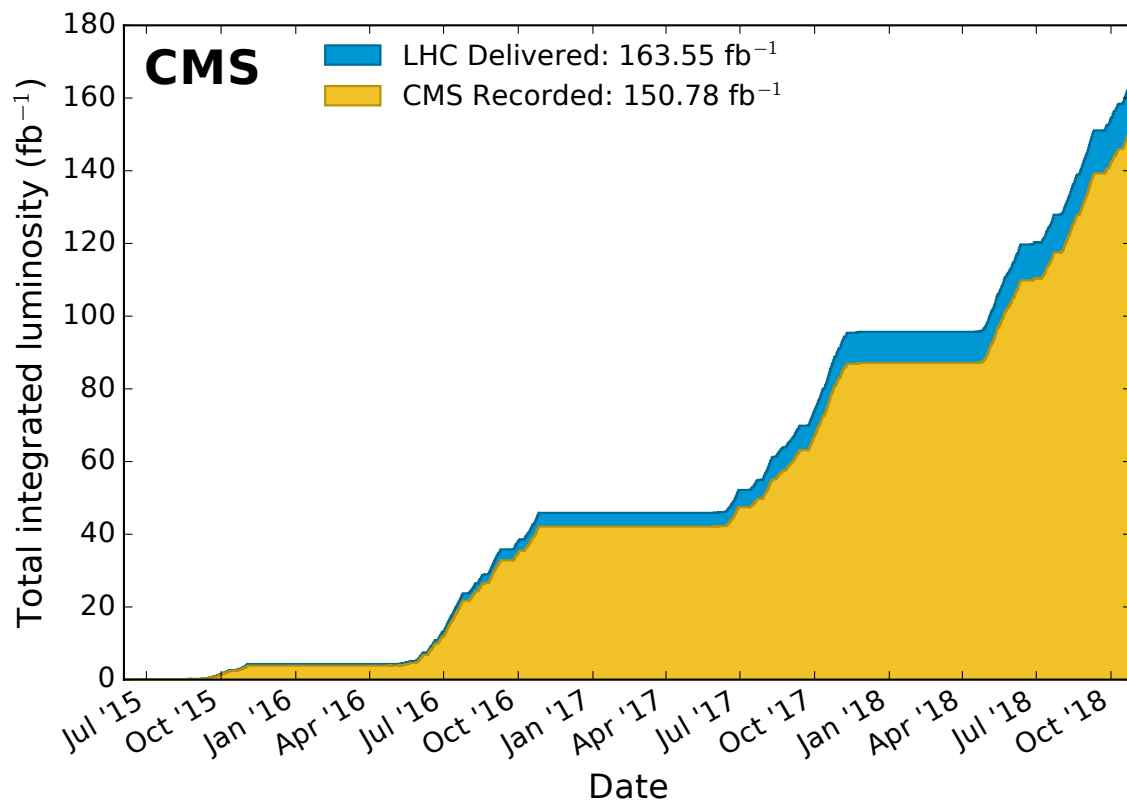


Figure 2.10: Integrated pp collision luminosity delivered by the LHC and recorded by CMS starting from 2015 till the end of 2018.

Chapter 3: Physics object reconstruction

The data recorded by the CMS detector consist of various electrical signals from the many detector channels across all subsystems described in Chapter 2. Physics object reconstruction refers to the computation process of taking these individual readouts together with the spatial information of the detector volume to get an image of particles generated at the collision point and passing through the detector volume. The particle-flow (PF) algorithm [74] first creates “elements” from subdetector responses, namely tracks of charged particles, vertices of multiple tracks, and calorimeter clusters. The algorithm then “links” the various elements together according to the expected detector responses for different particle types to form “PF candidates.” Reconstructed candidates aim to best represent particles passing through the detector for the responses recorded for a given event. A brief overview of the PF algorithm will be given in Sections 3.1–3.2.

Aside from the reconstruction of individual particles, multi-constituent objects are also helpful for representing the underlying physics. In particular, the clustering of PF candidates into jets, and subsequent correction needs are designed to allow the kinematic variables of the final jet to best represent the physics of the partons in the hard-scattering process. The definition of jets in the context of reconstructed objects will be discussed in Section 3.3. Other entities such as the missing transverse momentum ($\vec{p}_T^{\text{miss}}$) and the visible transverse energy (H_T) of the event

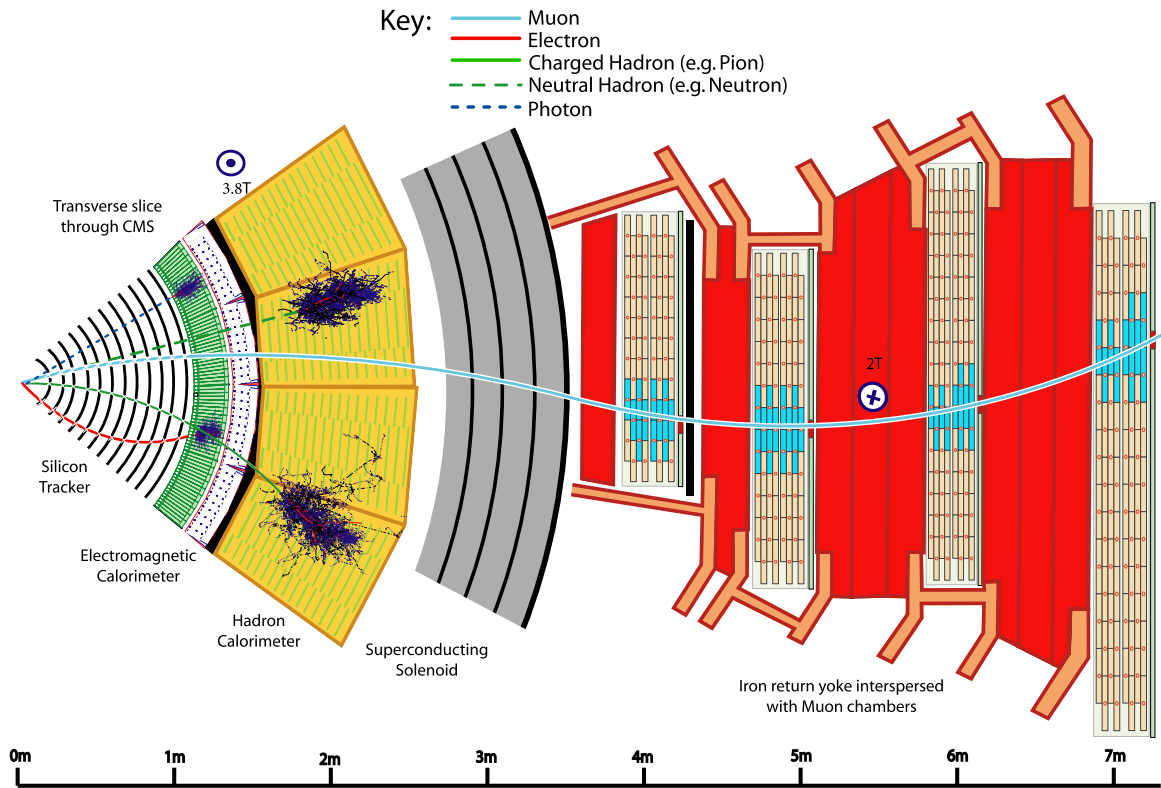


Figure 3.1: Diagram of various types of particle passing through a transverse slice of the CMS detector. The different detector responses serve as the basis for the particle flow algorithm. Image taken from Ref. [74].

will also be discussed in this section.

The PF algorithm makes minimal assumptions to ensure that the most interesting signatures can be captured. This may lead to the algorithm producing erroneous or ambiguous objects that need to be removed to have the objects used for physics analysis to accurately reflect the kinematics of the hard scattering process. Standard object quality filters are implemented by the CMS collaboration, and will be described in Section 3.4. The filters are flexible enough to accommodate a variety of working points, the most appropriate for this analysis will be given in Chapter 4.

The last part of this chapter discusses how various physics processes are included into CMS physics analyses. Section 3.5 overviews how events are generated and simulated given some physics process of interest. Finally, Section 3.6 lists all events used for this analysis, including events collected by the CMS detector, as well as the generated events.

3.1 Particle flow elements

3.1.1 Track reconstruction

Charged particles passing through the CMS pixel, silicon strip, and muon systems leave localized “hits” in each detector as they pass through their components. The spatial position of these hits can be fitted via the expected helix path to obtain a particle **track**.

Tracks are fitted by a Kalman filtering process [75]: **track seeds** are extracted by selecting a few hits in nearby detector layers that are sufficiently compatible with the expected particle trajectory. A trajectory is then extrapolated from the initial seeding algorithm, and the hits that match this trajectory are collected. Finally, all collected detector hits are fitted to extract the

track’s closest approach to the beam spot (referred to as the “reference point” of the track), the track’s transverse momentum and direction. The vectorial momentum evaluated at the reference point will be referred to as “the” track momentum, unless otherwise specified.

The track fitting is an iterative process to allow for high tracking efficiency for a wide range of track p_T and track reference point while keeping computation times and misreconstruction rates reasonably low, as discussed in Ref. [76]. The Kalman filtering routine is initially performed using the strictest seeding criteria, tracker hits that are compatible with tracks are removed, then the Kalman filtering routine is performed using a looser seeding criteria. The iterative fitting of CMS tracks starts with track seeds that require three consecutive hits in the pixel layer lying close to a straight line (a high p_T track), and subsequently loosened to seeding from silicon strips, seeding with intermediate missing layers, and seeding from the muon system.

The algorithm described above attempts to include as much of the detector responses as reasonably possible. Tracks can further be classified according to how many missing hits occur along the particle path, the uncertainty in the fitted track parameters, and the number of detector hits used to construct the track. **High-purity** tracks are the tracks constructed with low fitting uncertainties deemed suitable for general analysis use for track kinematic extraction [76], and will be the only tracks taken into consideration for the PF linking routine described in Section 3.2.

For tracks with hits in the muon system, the performance of the track reconstruction algorithm is measured in data for muon-compatible tracks using a tag-and-probe method. The CMS tracking algorithm was found to have effectively 100% track reconstruction efficiency for all tracks with $p_T > 0.8$ GeV that are compatible with a muon [77], as shown in Fig. 3.2. For tracks in jets, the reconstruction efficiency of high-purity tracks is measured in simulated $t\bar{t}$ production events, with the algorithm reaching full efficiency of $\sim 90\%$ for prompt tracks with p_T

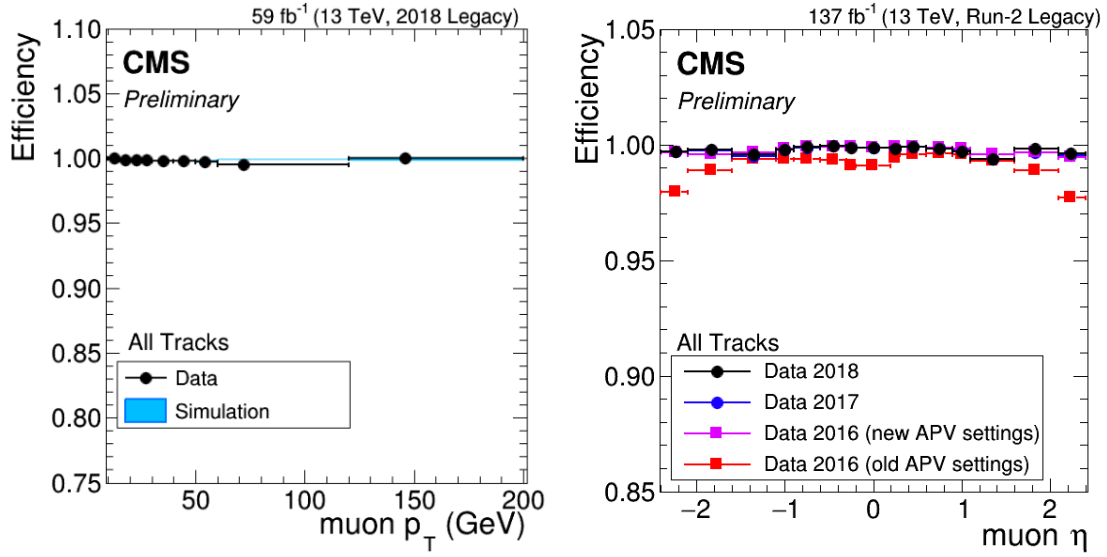


Figure 3.2: Reconstruction efficiency of muon tracks as a function of muon p_T (left) and η (right). The displayed efficiency of p_T plot is measured using 2018 data, while efficiency in the η plot compares the reconstruction efficiency from various data collection periods. Plots from Ref. [77].

> 0.8 GeV [78], as shown in Fig. 3.3. Figure 3.3 also demonstrates the importance of relaxed track-seeding requirements for the reconstruction of displaced tracks that may arise from long-lived particles. Reconstructed tracks are expected to have p_T resolution of $\sim 1.5\%$ for tracks with $p_T < 10$ GeV, and a spatial resolution of $\sim 10\mu\text{m}$ and $\sim 30\mu\text{m}$ in the transverse and longitudinal direction [76], as displayed in Fig. 3.4.

During the data collection of early 2016, the preamplifier of the APV25 readout chip of the silicon strips was misconfigured, resulting in a reduction of recorded hits [79]. A fix to the readout chip setting was deployed in mid-2016. While the tolerance parameters in tracking algorithm were later tuned to compensate for detector configuration differences, differences in the track efficiency were still observed, as demonstrated in Fig. 3.2. Because of these differences, we explicitly label the data collected in 2016 separately with “2016 pre-APV fix” and “2016 post-APV fix” respectively.

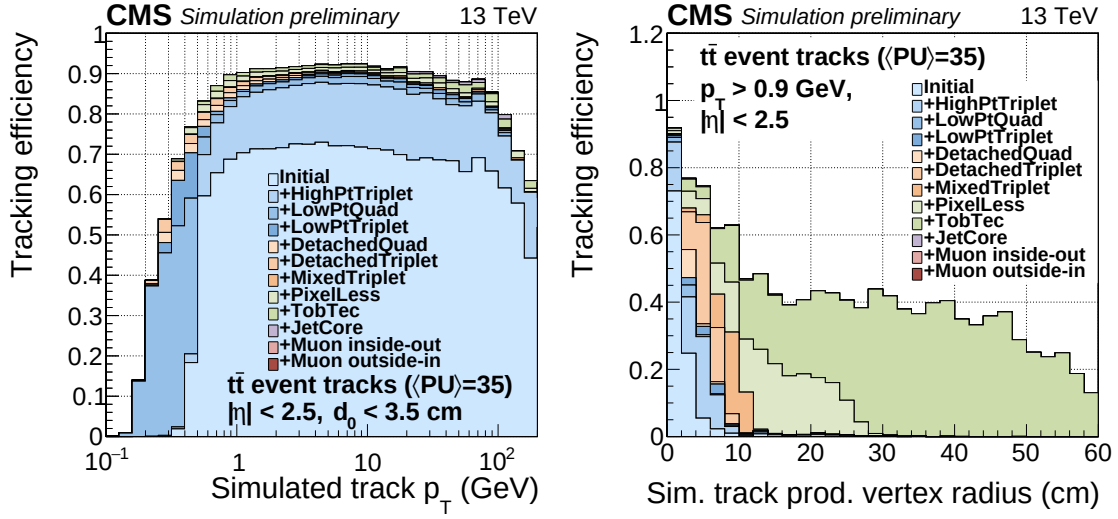


Figure 3.3: Reconstruction efficiency of tracks as a function of track p_T (left), and the radial position of the track production vertex (right), measured using simulated $t\bar{t}$ production with the detector conditions of 2017 data taking. Plots from Ref. [78]

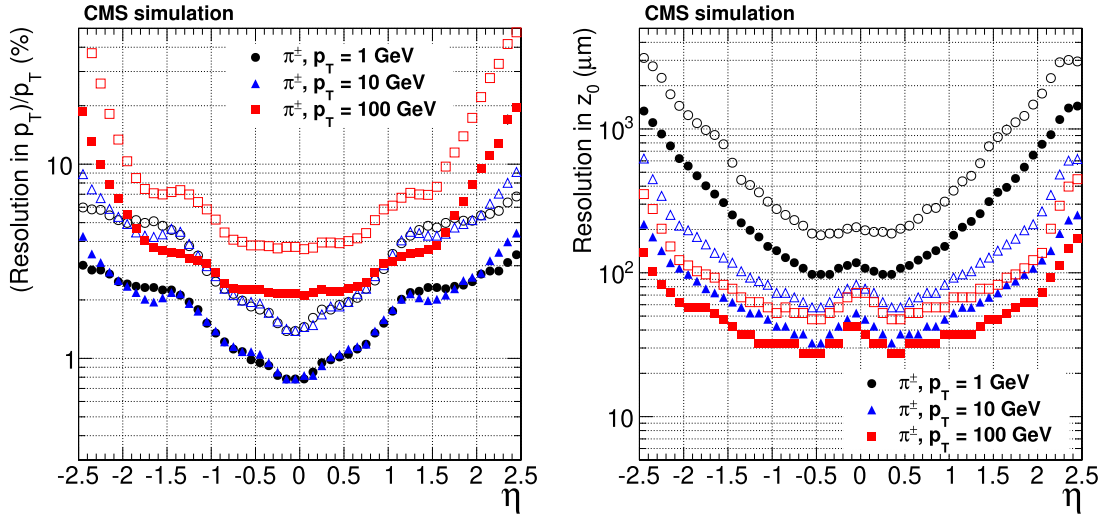


Figure 3.4: The measured track p_T resolution (left) and reference point longitudinal coordinate (right) as a function of track p_T in various bins of reconstructed track p_T , measured using simulated events. Plots from Ref. [76]

3.1.2 Vertex reconstruction

Conceptually, vertices refer to the points in space at which some physics interactions take place which results in multiple particles emerging from a single point in space. **Primary vertices** refer to the interaction points of the pp collisions, while **secondary vertices** (and sometimes tertiary vertices) refer to the point where long-lived particles generated from the pp collision decay some finite distance away from the primary vertex. In terms of detector signature, vertices leave no direct measurement. Instead, vertices will refer to a point in space where multiple tracks intersect. Vertex fitting (or sometimes simply **vertexing**) refers to finding the spatial points of interest given some set of tracks. In this section, the general description of the fitting algorithm will be given, along with the specifics on the determination of the primary vertex. The reconstruction of secondary vertices is discussed separately in Section [3.4.4](#).

As multiple pp collisions are presented within each event, **track clustering** is first performed to extract the number of vertices in the event, along with their estimated position. The clustering is performed using the set of z coordinates of all high-purity tracks within an event, along with their corresponding uncertainty. These sets of these numbers are labeled as $\{z_i^T\}$ and $\{\sigma_i\}$. Given a set of **proto-vertex** z coordinates $\{z_k^V\}$, we can then compute the **standardized z -distance** between a track and a proto-vertex as $\tilde{\chi}_{ik} = (z_i^T - z_k^V)/\sigma_i$. The value of $c\tilde{h}_{ik}$ is a simple measure to check how compatible a track is to a proposed vertex position. The **deterministic annealing (DA) algorithm** [\[80\]](#) equates the problem of clustering positions with random noise to the problem of minimizing a “free energy” measure. If we allow both the number of vertices in $\{z_k^V\}$ and their positions to be variable, the free energy measure F proposed for determining

the optimal proto-vertex positions is defined as:

$$F = -T \sum_i p_i \log \left(\sum_k \exp \left(-\frac{\tilde{\chi}_{ik}^2}{T} \right) \right), \quad (3.1)$$

Where p_i is the probability of the track i being compatible with the beam spot, and T is a temperature that roughly corresponds to the expected level of noise in the system. Because the number of proto-vertices can vary, we can always construct a solution that yields $\tilde{\chi}_{ik} = 0$. However, in the high- T limit, the function F heavily penalizes having multiple terms in the $\sum_k$ summation. So a DA algorithm working at high T prefers solutions with fewer proto-vertices. Conversely, for the low- T scenario, the extreme values of having large $\tilde{\chi}_{ik}$ values in the exponential function will push the DA algorithm to prefer solutions with more proto-vertices. The value of T can be adjusted to best match the expected noise level of the track parameters to obtain an estimate of the number of vertices and the vertex positions. As working with varying numbers of objects is computationally expensive, a “mass-constrain clustering” is performed by working with a fixed number of proto-vertices with varying weights ρ_k :

$$F = -T \sum_i p_i \log \left(\sum_k \rho_k \exp \left(-\frac{\tilde{\chi}_{ik}^2}{T} \right) \right) ; \quad \sum_k \rho_k = 1. \quad (3.2)$$

Proto-vertices that converge to the same coordinate value can be effectively pruned by adjusting the weight ρ_k . The DA algorithm deployed by CMS starts at a very high T value such that the number of proto-vertices is compatible with 1. Then, the value of T is lowered (“annealed”) by factors of 0.6 until it reaches $T = 4$. The number of proto-vertices is fixed using the results obtained at $T = 4$, then the value of T is further annealed to $T = 1$ to arrive at the final set of

proto-vertices. The values chosen are determined to maximize vertex reconstruction efficiency while minimizing the probability of constructing erroneous vertices by the presence of extreme outlier tracks.

Once the number and positions of proto-vertices are determined, the exact vertex position is refined by the **adaptive vertex fit** [81]. Working now with the standardized track-vertex distance in 3D space, χ_{ik} , and given vertex position $\vec{v}_k$, the optimal vertex position is determined by minimizing the measure $\mathcal{V}(\vec{v}_k)$ with the initial point determined by the proto-vertices obtained from the DA algorithm. The measure $\mathcal{V}(\vec{v}_k)$ is defined as:

$$\mathcal{V}(\vec{v}_k) = \sum_i \left(\frac{1}{2} (\chi_{ik}^2 - \psi_{ik}^2) \right) ; \quad \psi_{ik}^2 = T_a \log \left(\frac{1 + \exp \left(\frac{C^2}{2T_a} \right)}{\exp \left(\frac{\chi_{ik}^2}{2T_a} \right) + \exp \left(\frac{C^2}{2T_a} \right)} \right), \quad (3.3)$$

where T_a is again a “temperature” measure of how sensitive the overall sum is to χ_{ik} greater than critical measure C . Another way to look at this measure is that each track is weighted with the measure w_{ik} defined as:

$$w(\chi_{ik}, T, C) = \frac{\exp \left(-\frac{\chi_{ij}^2}{2T} \right)}{\exp \left(-\frac{\chi_{ij}^2}{2T} \right) + \exp \left(-\frac{C^2}{2T} \right)}, \quad (3.4)$$

and the minimization used to determine vertex position is performed on the sum $\sum_i w_{ik} \chi_{ik}^2$. Compared with the simple summation of $\sum_i \chi_{ik}^2$, the vertex measure given in Eq. 3.3 ensures that the vertex position can be set to be agnostic to extreme outliers, as seen in Fig. 3.5. The standard adaptive vertex fitter used by CMS uses the values of $T = 1$ and $C = 4$. The choice of $T = 4$ in the DA algorithm is chosen to match the choice of $C = 4$, effectively meaning that vertex splitting will not occur for separation smaller than 4 times the track position resolution.

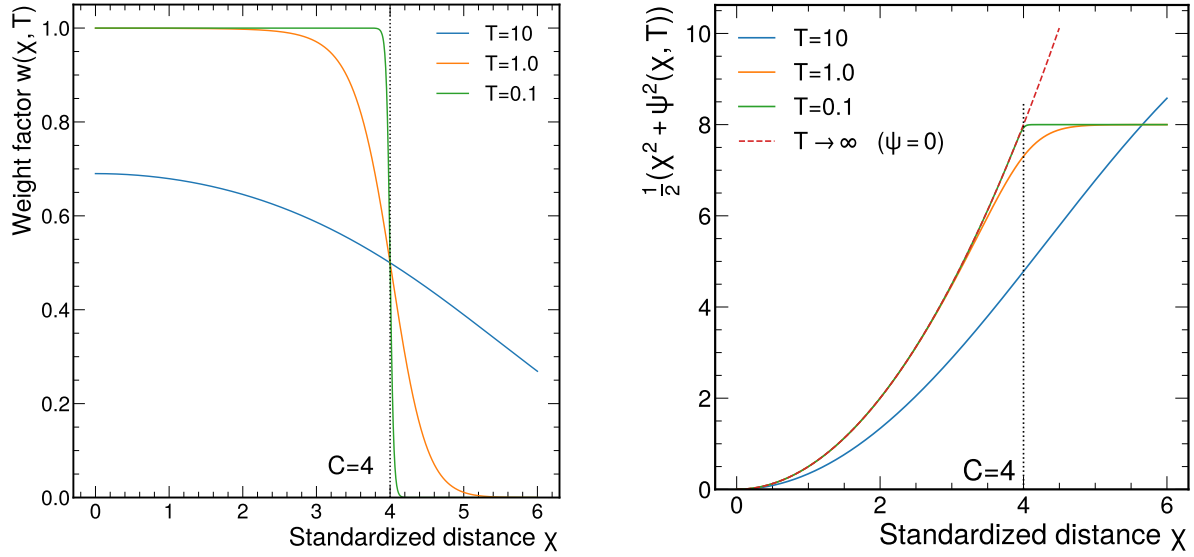


Figure 3.5: The weighting value of track w_{ik} used by the adaptive vertex fit (left), and the value of $\chi^2 + \psi^2$ (right) with varying values of T and a fixed value of $C = 4$.

The reconstruction efficiency of primary vertices was found to be effectively 100% for all vertices fitted with more than 2, with a reconstructed transverse and longitudinal resolution of $10 - 12\mu\text{m}$ respectively [76]. The vertex resolution is a strong function of the number of associated tracks, and displayed in Fig. 3.6.

3.1.3 Impact parameters

While impact parameters (IP) are *not* PF-elements, the distance between a track and a vertex of interest is used constantly throughout the analysis, so the formal definition will be given here.

The track trajectory is extrapolated to the point of closest approach relative to the vertex of interest, and IP is defined as the measured separation from the primary vertex to the point of closest approach. While this separation can be represented as a spatial vector $\overline{\text{IP}}$, we typically only work with the scalar components of this vector: the **transverse impact parameter** (labeled

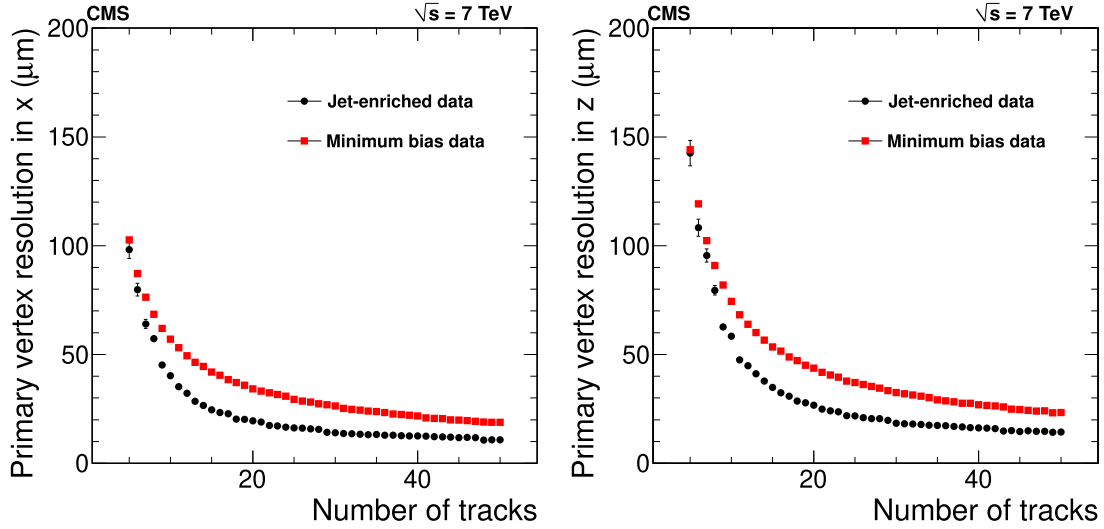


Figure 3.6: The transverse and longitudinal resolution of the reconstructed vertex position as a function of track multiplicity, measured using data collected with $\sqrt{s} = 7$ TeV. Plots are taken from [76]

IP_{2D}) is the radial component of the separation vector on the x - y plane; the **longitudinal impact parameter** (labeled IP_z) is the component of this vector in the $\hat{z}$ direction; the **(total) impact parameter** (labeled IP_{3D}) is the absolute value of this separation vector. The uncertainty of each of these parameters will be notated as σ_{IP_n} .

The three impact parameter scalars above are all signed. For IP_z , the sign is simply whether the point of closest approach or the PV is further along in the z direction, and carries no physical significance. For IP_{2D} and IP_{3D} , the sign represents the geometric relation between the vertex and the curvature of the helix path: positive if the vertex is outside the track's circular x - y trajectory, and negative if the vertex is inside. The sign can be calculated using the momentum information using the equation:

$$\text{sgn}(IP_{2D}) = q \cdot \text{sgn}\left((\vec{p}_T \times \vec{IP}_{2D}) \cdot \hat{z}\right), \quad (3.5)$$

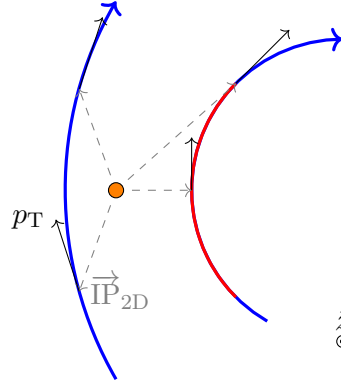


Figure 3.7: Evaluation of the sign function defined in Eq. 3.5, evaluated for a track with the vertex (orange circle) inside the circular path (left path) and outside the circular path (right path). The colors of the path are used to indicate the evaluated result, depending on which reference point is used for the evaluation of $\vec{IP}_{2D}$ (gray, dashed vectors) and $\vec{p}_T$ (black tangent vector): blue is negative, while red is positive. If evaluated far away from the primary vertex, tracks with positive IP_{2D} can evaluate the wrong value when using Eq. 3.5 (blue segments for right path). But this method is robust when the evaluation is performed close to the point of closest approach.

where $\vec{IP}_{2D}$ is the separation vector from the vertex to the point-of-closest-approach, $\vec{p}_T$ is the momentum evaluated at the point-of-closest-approach, and q is the integer charge of the particle (either +1 or -1 depending on whether the helix path propagates clockwise or counterclockwise compared with the direction of the CMS magnetic field). The geometric interpretation of Eq. 3.5 is demonstrated in Fig. 3.7.

3.1.4 Calorimeter clusters

Particles reaching the calorimeters are expected to leave localized energy deposits spread across multiple calorimeter cells, which we call “clusters”. The clustering algorithm is designed to isolate the energy deposits of individual particles as individual clusters, while also keeping energy deposits from in-flight soft radiation processes within the same cluster.

Cluster seeds are identified as being local maxima in the calorimeter cells that also have their energy reading above the noise threshold of the cell. These are taken as the starting point of

particle-flow clusters. A **topological clustering** is then carried out for all cells with energy reading above a threshold value, with cells associated with PF clusters when adjacent. At this point of the algorithm, a single cell can be associated with multiple PF clusters. Once all calorimeter cells has been assigned to a cluster via topological clustering, cluster position is calculated by an energy-weighted average position of all associated cells. To resolve the ambiguity of cells being shared by multiple PF clusters, the energy of shared cells is split into different clusters according to the cell–cluster distance. The cluster position is then updated according to the new cell energy contribution. This process is iteratively performed for all shared cells until all ambiguities are resolved.

The same clustering algorithm described above is run separately for ECAL endcap, ECAL barrel, HCAL endcap and HCAL barrel, with the cell energy threshold values being 80 MeV for ECAL barrel, 300 MeV for ECAL endcap and 800 MeV for HCAL respectively.

3.2 Particle-flow linking and identification

Most particles leave a mix of signatures as they pass through the various detector elements. PF linking attempts to group the ingredients of tracks and calorimeter clusters, referred to as **particle-flow blocks**: inner tracks trajectories are first extrapolated into the ECAL and HCAL to a depth of one interaction length. Tracks are then linked to PF clusters if the extrapolated position is within a PF cluster boundary up to one cell width away. To account for Bremsstrahlung photons being emitted from charged particles bending in the magnetic field, ECAL clusters that are compatible with the extrapolated tangent lines of a track are also linked to the track of interest, with the collection of ECAL clusters are referred to as **superclusters**. Clusters in ECAL and

HCAL are linked if the two clusters are mutually within each other's cluster boundaries. Finally, tracks in the inner tracker system and the muon system are linked if a global fit of the two tracks that are within three standard deviations of uncertainty.

After obtaining the PF blocks of an event, each block is identified as a physics objects of a particular type referred to as **particle-flow candidates** depending on the signature particles are expected to leave within the detector. The energy contributions of PF candidates are subtracted from overlapping clusters of the remaining PF blocks, and is iterated until all PF blocks have been resolved.

Muons are the first objects to be identified, as they are the only particles that leave a signature in both the inner and the muon tracking system. The mutually compatible muon and inner tracks are removed from any PF blocks, and clusters along the muon trajectory will have their energy subtracted according to the expected energy deposition of the muon passing through the systems (~ 3 GeV for ECAL, and ~ 0.5 GeV for HCAL). **Global muons** refer to muons tracks fitted with both inner tracker hits and muon system hits. **Standalone muons** refer to muons that only have tracks within the muon system. **Tracker muons** refer to muons with a track reconstructed with the inner tracker system and compatible muon system hits.

Electrons are prone to Bremsstrahlung radiation and energy losses in the tracker system due to their small mass. As a result, tracks generated by an electron may result in a low-quality fit with poorly fitted track parameters when fitted with the default Kalman filter. Tracks compatible with an ECAL supercluster are refitted using a Gaussian-sum filter [82] (a modified Kalman filter to account for the energy loss of the electrons at each layer of material), to accurately obtain electron

information.

Isolated photons are identified as ECAL superclusters with little overlap with HCAL clusters and no tracks in their vicinity.

Non-isolated photons, neutral hadrons and charged hadrons are identified once the PF elements assigned to the muons, electrons and isolated photons have been removed, and are assumed to be generated from hadronization showers. These particles are distinguished by whether a cluster has any associated tracks, and whether the tracks have compatible momenta compared with the energy of the clusters in the remaining track–cluster blocks. In the case that multiple tracks are linked to the same cluster, the vectorial sum of the momentum of all the associated tracks should be used for momentum–energy comparison. In the case that one track is linked to multiple clusters, only the ECAL cluster and HCAL cluster closest to the track position will be considered. Clusters with no associated tracks or no well-matched tracks are taken to as the photons (ECAL clusters) and neutral hadrons (HCAL clusters), and their clusters are removed from the remaining PF blocks. Charged hadrons are then the remaining tracks–cluster pairs with compatible momentum–energy matching. Because the ECAL crystals are only accurately calibrated to the expected energy deposit of electrons and photons, once an ECAL cluster has been associated with either a charged hadron or neutral hadron, the cluster energy is then re-calibrated to account for the expected energy deposits by hadrons in the ECAL crystals using test-beam data.

3.3 Compound objects

The PF algorithm described so far attempts to reconstruct individual particles traversing the detector volume to account for all signals measured within CMS. These PF candidates can be used directly, or be further grouped into compound objects that better reflect the physics of the underlying hard scattering process. The most common compound object is the **jet** composed of particles within an angular cone. Such a definition of a jet best captures the signature of particle showers generated from the hadronization process of QCD-colored particles leaving the interaction point.

3.3.1 Jets

3.3.1.1 Jet clustering

First, a generic clustering algorithm to cluster PF candidates within an angular vicinity that best matches the expected particle showering process can be defined. Given a list of p_T -ordered particles of interest in the event and an existing pseudo-jet (labeled using j), the algorithm can be described as an iterative process of the following steps:

1. For the leading particle in the particles list (labeled i), we calculate a distance measure between the particle and the existing pseudo-jet, d_{ij} , as well as a distance measure between the particle and the beam line d_i .
2. If the leading particle i is considered close to the pseudo-jet compared with the beamline (i.e. $d_{ij} < d_i$), we add this particle to the pseudo-jet, and repeat step 1 for the next particle in the list.

3. If the leading particle is not close to the pseudo-jet, the leading particle is discarded to a “remainder list”, and the step 1 calculation is performed with the next particle.
4. Steps 1–3 are repeated until all particles are either added to the pseudo-jet or are placed in the “remainder list.” At this point, this pseudo-jet is considered complete, and a new pseudo-jet is constructed using the largest p_T particle in the “remainder list.” The remainder list is used as the new input list of the algorithm, and the algorithm iterates steps 1–3 with the new inputs.

The algorithm in steps 1–4 will be repeated until no singular particles remain. The inputs to the algorithm will be the full list of particle-flow candidates obtained using the linking algorithm described in Section 3.2.

The distance measures d_{ij} and d_i are given as:

$$d_{ij} = \min(p_{T,i}^{2p}, p_{T,j}^{2p}) \cdot \frac{\Delta\bar{R}(i,j)^2}{R^2} ; \quad d_i = p_{T,i}^{2p}, \quad (3.6)$$

where $p_{T,i}$ is the p_T of the particle, $p_{T,j}$ is the p_T of the pseudo-jet (transverse components of the vectorial sum of all particles in the pseudo-jet particle collection), $\Delta\bar{R}(i,j)$ is the angular separation between the particle momentum and the pseudo-jet momentum in the (y, ϕ) space ($\Delta\bar{R}(i,j)^2 = \Delta y_{i,j}^2 + \Delta\phi_{i,j}^2$); finally, R and p parameters of the algorithm that can be modified: R defines the angular scale of the resulting jets, while is used to tune the definition of “closeness” in the jet clustering algorithm.

The anti- k_T algorithm [83] uses the seemingly counterintuitive parameter of $p = -1$ in the distance measure, allowing the jet clustering to be less sensitive to the soft radiation processes

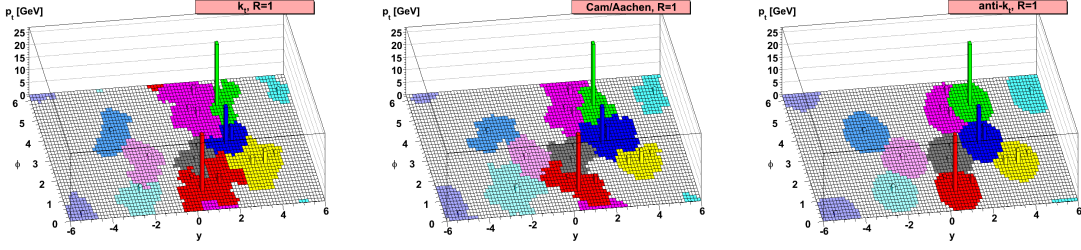


Figure 3.8: The jet clustering performed on the same events using the k_T algorithm ($p = 1$) on the left, the Cambridge-Aachen algorithm ($p = 0$) in the middle and the anti- k_T algorithm ($p = -1$) on the right, with random soft particles added into the event. Figures are taken from Ref. [83].

(see Fig. 3.8). The standard jets used at CMS are clustered using the anti- k_T algorithm with an angular parameter size of $R = 0.4$.

3.3.1.2 Jet corrections and resolution

Jets are typically used to represent the kinematics of colored partons in the hard scattering process. To have the jet kinematics best match the underlying parton, the constituents and the summed kinematics of jets are further corrected following the steps described in Ref. [84] and Fig. 3.9. While the correctness of the jet kinematics is not critical for tagging the EMJ signatures, corrections to jet kinematics will still impact the overall event selection efficiency when setting the jet kinematics thresholds. For the sake of completeness, we will include a brief description of the corrections applied to the standard CMS jets. The discussion of the impact of these corrections on the final analysis will be given in Section 6.1.2.

First, contributions from charged particles originating from pileup vertices are removed; this process is called the charged hadron subtraction (CHS). Contributions of neutral particles from pileup interactions should be removed. This is done by estimating the average energy per unit area (notated as ρ) of the event, and subtracting it from the total jet energy the expected

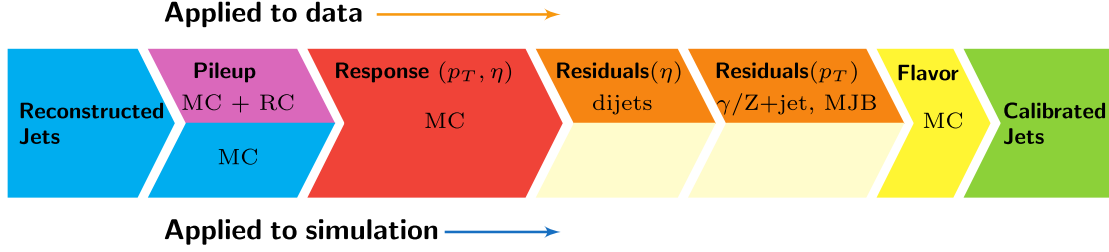


Figure 3.9: Consecutive stages of correction applied to jet in data and simulated events, to obtain jets that are suitable for analysis. Diagram taken from Ref. [84].

contribution estimated from the total jet area.

Next is a response correction, where we take the reconstructed jets in simulated events, and derive a correction by comparing the kinematics of reconstructed jets with the jets obtained using generator particles in the simulated events. This is expected to cover the effects of imperfect detector responses and the losses processes of the PF algorithm that can occur. For data, an additional residual correction can be obtained by comparing the jet kinematic spectrum of back-to-back production, such as dijet events and $Z \rightarrow \mu^+ \mu^- / e^+ e^- + \text{jet(s)}$ processes. To first order, the transverse kinematics of jets in such back-to-back processes are expected to be mirrored for jets in the same event. In particular, as the kinematic of the Z boson kinematics can be determined from leptons, the kinematics of the reconstructed Z boson experiences far fewer detector response ambiguities compared with jets.

After corrections to jet energies, CMS calculates the jet energy resolution as a function of the corrected jet p_T and η by comparing the random spread in the jet p_T when comparing back-to-back processes such as γ +Jet final sets.

An example of the jet energy response corrections calculated in MC and the final jet energy resolution can be found in Fig. 3.10 (from Ref. [84]). The final corrected momentum of the jets is expected to have an uncertainty of $\sim 3\%$ for jets in the barrel region with $p_T > 50$ GeV, with a

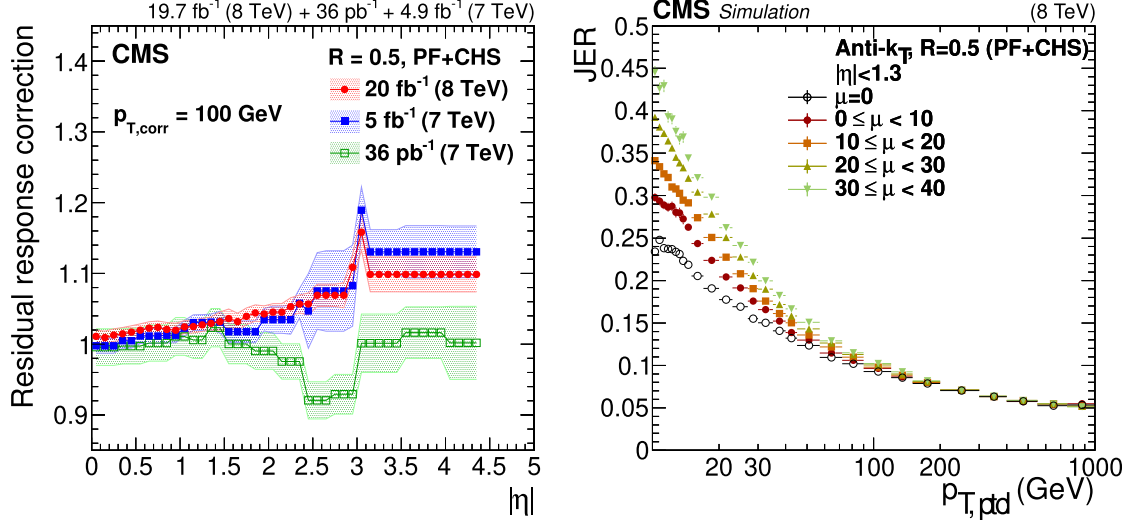


Figure 3.10: Example of the calculated jet energy response correction for jets with $p_T = 100$ GeV as a function of jet $|\eta|$ (left), and the final jet energy resolution as a function of jet p_T (right). Plots are taken from Ref. [84].

resolution of $\sim 10\%$.

3.3.2 Visible transverse momentum

While not strictly a “physics object”, the visible transverse energy of an event, H_T , is a useful variable that can be used to easily represent the overall energy scale of the event¹. The H_T is defined as the scalar p_T sum of all jets in the event that has $|\eta| < 5.0$.

3.3.3 Missing transverse momentum

Particles that neither carry EM charge nor QCD color are expected to escape the detector without leaving a signature. If such particles carry a significant fraction of the energy from the hard scattering process, this will result in an apparent imbalance of energy deposition in the

¹The reason for the label “ H_T ” is that, historically, the variable used for representing event energy was the total energy reading from hadronic calorimeters (hence the label H), as it was easy to calculate directly in hardware. Since “event energy” is not a well-defined variable of the hard scattering processes, the same symbol was borrowed throughout various experiments, each with a modified definition. What is presented here is the most typical definition of H_T used in CMS.

event. While the imbalance of the energy in the z direction is ill-defined, as the hard scattering process primarily involves partons carrying a random fraction of the proton energy, the transverse imbalance is well-defined, as the initial transverse momentum of the hard scattering parton is expected to be effectively 0.

The missing transverse momentum, $\vec{p}_T^{\text{miss}}$, is defined as the inverse of the vectorial $\vec{p}_T$ sum of all reconstructed PF-objects of interest:

$$\vec{p}_T^{\text{miss}} = - \sum_{i \in \text{PF cand}} \vec{p}_{Ti}. \quad (3.7)$$

Additional cleaning and corrections can be applied, similar to jet energy corrections described in Section 3.3.1.2, such that $\vec{p}_T^{\text{miss}}$ best reflects the underlying physics of the invisible particle.

This analysis will not directly use $\vec{p}_T^{\text{miss}}$, as the focus of this analysis is on EMJ signatures identifiable using the tracker volume. If the dark meson lifetime is so large that a large fraction of the dark shower energy escapes the detector, a new analysis targeting this dedicated signature should be used instead.

The correctness of $\vec{p}_T^{\text{miss}}$, however, is a very good indication that all subdetector components and reconstruction algorithms are operating nominally, as the definition of $\vec{p}_T^{\text{miss}}$ in Eq. 3.7 naturally takes into account all steps of reconstruction. For analysis that heavily relies on the correctness of $\vec{p}_T^{\text{miss}}$, CMS has developed a set of filters to remove events that may give erroneous $\vec{p}_T^{\text{miss}}$ due to excessive detector noise or sporadic beam effects [85]. These filters will be used for all events in the EMJ analysis, as this is found to reduce the chances of the SM jets being misidentified as EMJs.

3.4 Physics object quality filters and tagging

The PF algorithm described in Sections 3.1–3.2 attempts to impose minimal selection to reconstructed physics objects. This can result in false, or fake, physics objects being listed by the PF algorithm. For example: photons as described in Section 3.2 are simply ECAL energy without an associated track; such signatures can be generated without a true photon being generated in the hard scattering process: for example, an electron can leave an ECAL energy with its tracking information failing to be properly reconstructed as a track by the PF algorithm; or a neutral hadron can leave too little energy in HCAL, causing the clustering algorithm to classify the energy cluster as a photon rather than a neutral hadron. Additional selection cuts can be imposed on the spatial and energy information of the supposed photon energy cluster and the surrounding readout values to allow the analysis to be more confident that a photon reported by the PF algorithm is indeed a true photon produced from a hard scattering process, at the expense of potentially discarding true photons with excess noise in the detector during the event of interest. Such “object quality” (sometimes also referred to as “object identification”) selection cuts are not applied by default by the particle-flow algorithm, leaving the analyzer to determine which object quality cuts are reasonable for the analysis in question.

Aside from object quality, additional computations may be performed on the physics object of interest, which can be used to “tag” particular signatures of interest. One common use case for this would be to use the information of the jet constituents to estimate the underlying flavor of the particle that generated the jet.

While an analysis can implement its own quality selection and tagging variable depending on the physics signature of interest, standard recipes to select physics objects of various quality

levels are provided and studied by the CMS collaboration, which allows for a good starting point for analysis working with “typical” physics objects, while also providing a framework for evaluating the correctness of the analysis-specific object selection criteria. In the following sections, we will describe the physics object quality selections provided by the CMS collaboration that are used in this analysis. Quality selection and tagging variables specifically designed for the EMJ analysis will be given in Chapter 4.

3.4.1 Primary vertex ordering

Multiple primary vertices are reconstructed within an event from pileup interactions. Typically, only a single interaction is a high-energy collision of potential interest, with the rest being low- p_T interactions. Correctly identifying which primary vertex is responsible for the high-energy process of interest is important, especially for analyses that look at geometric relations between tracks, cluster and primary vertex for additional physical object identification or quality filtering.

While the simplest measure of the energy of a primary vertex can be given as the scale sum of p_T (or p_T^2) of the tracks associated with the vertex of interest (tracks that have a weighting factor ≥ 0.5 during the vertex fit, as described in Section 3.1.2), these simple measures were found to be problematic for analyses of the electroweak sector, where the hard scattering generates a few particles with potentially missing particles from neutrinos. To account for this, the primary vertices reconstructed in the event are ordered by the measured properties of their associated tracks according to the following algorithm [86]:

1. For all associated tracks, a pseudo-momentum p' is calculated as:

$$p' = p \left(1 - \frac{\sigma_{p_T}}{p_T} \right), \quad (3.8)$$

where p is the reconstructed momentum of the track, and σ_{p_T} is the uncertainty of the reconstructed tracks' p_T .

2. Tracks that are not associated with a particle-flow electron or muon are clustered into jets using their pseudo-momentum measure using the anti- k_T algorithm with an angular parameter of 0.4.
3. The electrons, muons and track-clustered jets are then used to compute the following measures:

$$S_T = \sum_{i \in e, \mu} p_{T,i}'^2 + \sum_{j \in \text{jets}} (0.8 p_{T,j}')^2 ; \quad (3.9)$$

$$S_E = \sum_{i \in e, \mu, \text{jets}} p_{T,i}' ; \quad (3.10)$$

$$S_v = \left| \sum_{i \in e, \mu, \text{jets}} \vec{p}_{T,i}' \right|. \quad (3.11)$$

Notice the difference in the scalar sum used for S_E and the vectorial sum used for S_v .

4. The final measure used to order the primary vertices in the events is defined as:

$$S = S_T + (\max(S_v - S_E, 0))^2 \quad (3.12)$$

The measure S attempts to take into account the energy scale of the particles used to reconstruct

the primary vertex, to pick the most energetic vertex in the event. The promotion of electrons and muons over the charged hadrons in jet-like clusters ensure that primary vertices from electroweak events, which are likely to include single high- p_T tracks from isolated leptons generation, are properly selected.

For the remainder of this thesis, when referring to the singular form of “the primary vertex” of an event, this statement will be referring to the leading primary vertex as defined by the measure listed above. Unless specified otherwise, all track displacement variables (defined in Section 3.1.3) are calculated relative to the leading primary vertex of the event.

3.4.2 Photon identification

As stated in the example given at the start of Section 3.4, fake photons can arise from an erroneous track-cluster assignment, or cluster assignments in the electron particle-flow reconstruction. Photon identification can be performed based on the shape of the ECAL energy cluster in the detector, and comparing the ECAL energy cluster with the energy of physics objects in the vicinity. CMS officially provides the following method of calculating photon identification [87]:

- The ratio of energy deposit in HCAL v.s. ECAL (notated as H/E) should be smaller than a cut-off value, $(H/E)_{\text{cut}}$.
- The spread in the η direction of the ECAL energy cluster should be small. The variable used to define the spread of an ECAL energy cluster is evaluated as the variable $\sigma_{i\eta i\eta}$, defined using the following formulas:

$$\sigma_{i\eta i\eta} = \sqrt{\frac{\sum_j^{5 \times 5} w_j (\eta_j + \bar{\eta}_{5 \times 5})^2}{\sum_j^{5 \times 5} w_j}}, \quad w_j = \max \left(0, 4.7 + \ln \left(\frac{E_j}{\sum_j^{5 \times 5} E_j} \right) \right), \quad (3.13)$$

where the summation $\sum_j^{5 \times 5}$ runs over the ECAL crystals in a 5×5 grid surrounding the most energetic crystal in the ECAL supercluster, $\eta_j (E_j)$ is the pseudorapidity (energy) of a single ECAL crystal, $\bar{\eta}_{5 \times 5}$ is the mean pseudorapidity of the 5×5 crystals of interest. The max function used for calculating the log-energy weight ensures we only count crystals that have energy greater than $\exp(-4.7) \sim 0.9\%$ of the total crystal energy in the 5×5 grid, suppressing the effects of noise from impacting the spread.

- Photons from hard scattering processes are expected to be isolated from other particles. An isolation variable is then defined by:

$$I_x = \max \left(\left(\sum_{\Delta R(i,\gamma) < 0.3} p_{T x, i} \right) - \rho A, 0 \right), \quad (3.14)$$

where x indexes a PF candidate type of interest, the summation runs over the PF candidates of types of interest within a ΔR cone of 0.3 relative to the photon of interest, ρ is the average energy density per unit area of the event (as defined for the L1 correction of jets, discussed in Section 3.3.1.1), and A is the effective angular area of the supercluster. The subtraction of ρA ensures the isolation measure is not impacted by pileup, and the max function ensures the isolation measure remains positive. The cut value of I_x will be a p_T -dependent function.

The CMS-recommended cut values for the photon identification are displayed in Table 3.1. The recommended cuts are estimated to have an 80% selection efficiency, with a $\sim 10\%$ false positive rate in the barrel region, and $\sim 18\%$ false positive rate in the endcap region. While photons are not directly used for the search for EMJs, the presence of photons will be used to

Table 3.1: Cut values used for photon identification used for the analysis. Values for the isolation variable are calculated using the p_T of the photon of interest in units GeV.

Cut variable	Cut values	
	Barrel photon ($ \eta < 1.4$)	Endcap photons ($ \eta > 1.5$)
H/E (j)	0.02197	0.326
$\sigma_{i\eta i\eta}(<)$	0.1015	0.0272
$I_{\text{charged had.}}$	1.141	1.051
$I_{\text{neutral had.}}$	$1.189 + 0.01512p_T + 0.00002259p_T^2$	$2.718 + 0.0117p_T + 0.000023p_T^2$
I_{photon}	$2.08 + 0.004017p_T$	$3.867 + 0.0037p_T$

construct a signal-free region for evaluating EMJ tagging rates. Details of how the presence of the photon is used will be given in Section 5.1.

3.4.3 Jet identification

While a jet is a generic term for particles clustered in angular space, and can be used to study any particle originating from the hard-scattering process, the use of jets in analyses is usually synonymous with the hadronization shower generated in the production of QCD-colored objects, rather than soft radiations from color singlet objects². Jet identification defines a set of selections such that jets passing the selection are expected to be from such showering processes [88]. The selection works by requiring the jet to be composed of multiple types of particles, with no extreme energy contributions from a single particle type, as demonstrated in Fig. 3.11.

- The neutral hadron energy fraction must be $< 90\%$.
- The neutral EM energy fraction must be $< 90\%$.
- The muon energy fraction must be $< 80\%$.
- There must be at least 1 charged constituent in the jet.

²An exception to this would be analyses involving τ leptons, as the large mass of τ (1.777 GeV) results in the large fraction of τ leptons decaying with hadronic products. For this analysis, we are not concerned with such cases

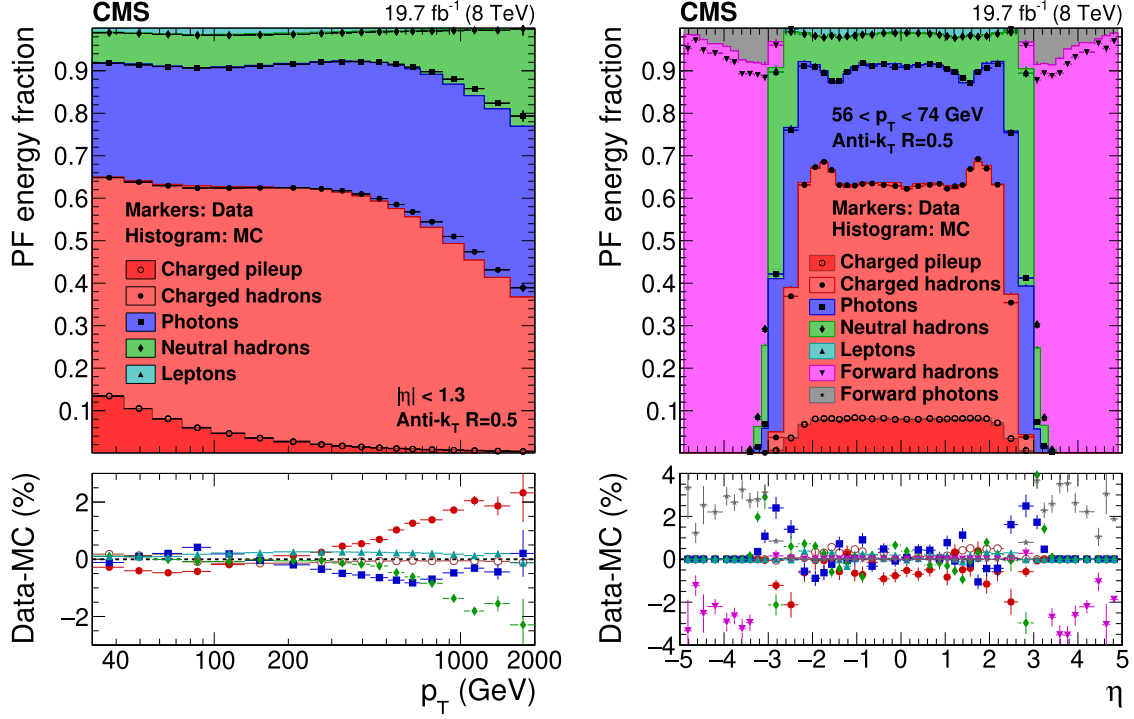


Figure 3.11: Energy fraction of jets constituents of various PF candidate types as a function of jet p_T (left), and jet η (right). Figures are taken from Ref. [84].

- The charged hadron energy fraction must be > 0 .
- The charged EM energy fraction must be $< 80\%$.

The selection listed above is expected to have a true positive rate of 98 – 99% and a false positive rate of electrons, muons, and photons consistently $< 2\%$ for jets with $|\eta| < 2.0$. While the production of EMJ signature doesn't just undergo SM hadronization processes, the jet identification selection is found to be highly compatible with the EMJ jets generated in the surveyed signal models. The jet selection presented above will be imposed on all jets used for the analysis.

3.4.4 Jet flavor tagging

Jet tagging refers to assigning a particle flavor to the jet using the reconstructed information of jet constituents to attempt to distinguish jets in the same event to different particles in the

hard-scattering process. This process is critical for the EMJ analysis, as the analysis looks to identify jets originating from the hadronic shower of dark quarks with long-lived dark mesons. The tagging of standard model particle flavors is also widely used for SM process analysis, as the production of heavy-flavored mesons is commonly used to assist in narrowing down the event selection to a specific sub-process of interest. For example, the top quarks have a 96% branching ratio of decaying to a bottom quark and an on-shell W boson [23], so selecting final states with bottom mesons can significantly increase the fraction of events containing top quarks in the hard-scattering process.

The tagging design for the EMJs will be given in full detail in Chapter 4, as the design and validation for the EMJ tagging scheme is the main focus of this analysis. This section will focus on the SM heavy-flavored jet tagging employed by CMS.

3.4.4.1 Truth jet flavor

Before, continuing with the design of jet flavor tagging, a concrete definition of a jet’s “true” flavor must be defined, assuming that complete information of all particles that originate from the hard-scattering process is available. While jets are typically associated with some parton of the hard scattering process, using the parton flavor as the definition of the jet flavor is not the most practical definition for jet flavor in the context of reconstruction features because there will never be reconstruction features of the parton that we can directly use. Instead, at most, jet tagging algorithms can use the features of the intermediate hadrons in the generated particle shower.

The standard definition of the hadron-based jet flavor used for simulated events in CMS

is the following: aside from the reconstructed PF candidates, additional intermediate simulated particles are included in the jet clustering algorithm. The simulated particles included for clustering are the heavy mesons (b mesons and D mesons) that decay into non-heavy meson particles. These simulated particles have their momentum scaled by 10^{-12} . The same jet clustering algorithm described in Section 3.3.1.1 is then run for the extended particle list. The extreme scaling of the simulated particles means that the clustering results of the PF candidates remain identical since the simulated particles do not modify the jet kinematics in any meaningful way. But, the simulated particles can be unambiguously associated with a jet. Jets are then given the unique flavor following this order:

1. If a jet contains a b meson when clustered using the extended particles list, then the jet is classified as a truth b jet.
2. If a jet does not contain a b meson but contain a D meson, then the jet is classified as a truth c jet.
3. If the jet neither contains a b meson nor a D meson, then the jet is classified as a truth light-flavored jet (or simply a light jet).

We perform a similar algorithm to select jets as a truth EMJ with the particle list extended with dark mesons. But we also require that the scalar sum of the unmodified dark meson p_T be at least 0.6 times the reconstructed jet p_T to account for the wider particle showers in the dark sector generating softer dark mesons outside the concentrated jet energy centers.

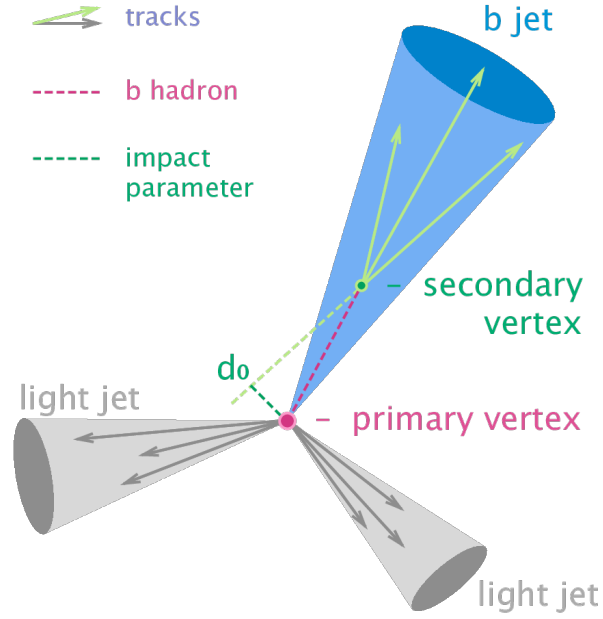


Figure 3.12: Diagram showing how secondary vertices can be used to identify jets containing heavy-flavored mesons in the particle shower. Diagram by Nazar Bartosik, distributed under the Creative Commons Attribution 4.0 International.

3.4.4.2 SM heavy-flavor jet tagging

The heavy-flavored bottom mesons and charm mesons in SM typically have longer lifetimes than other unstable SM mesons, with the decay paths determined by the CKM matrix's small, off-diagonal components [29]. With the b mesons and D mesons typically having a lifetime of $\sim 10^{-3} - 10^{-4}$ ns, and the typical energy of particles leaving the hard-scattering process being on the order 10–100 GeV, the decay of heavy-flavored mesons can be measured as a secondary vertex that is a finite distance away from the primary pp collision vertex. The kinematic properties of secondary vertices within a jet, along with its relation with the other jet constituents can then be used to tag the jet as being heavy-flavored.

Secondary vertices are constructed using the inclusive vertex finder algorithm [89], which is slightly different from the algorithm used for primary vertex finding as described in Section 3.1.2:

The vertex finder is computed using *all* tracks in the event with $p_T > 0.8 \text{ GeV}$ and $\text{IP}_z < 0.3 \text{ cm}$, labeling a special set of “seed tracks” that have $\text{IP}_{3D} > 50\mu\text{m}$ and $\text{IP}_{3D}/\sigma_{\text{IP}_{3D}} > 1.2$ when compared to the primary vertex. The remaining non-seed tracks are then passed through the following steps:

- **Track clustering:** non-seed tracks are clustered to the seed tracks by comparing if the point of closest approach to the seed tracks is smaller than the tracks’ closest approach to the primary vertex. Notice that this clustering is “inclusive,” meaning that a track can be clustered to more than one seed track.
- **Initial vertex fitting:** secondary vertices are then reconstructed using track clusters using the same adaptive vertex fitter described in Section 3.1.2. The fitted vertices will be discarded if the flight distance of the fitted vertex is too close to the primary vertex (typically a distance is smaller than ~ 2 time the position uncertainty). As duplicates vertices can arise from the inclusive clustering, secondary vertices that shared over 70% of the same tracks are also discarded.
- **Track arbitration:** tracks associated with some secondary vertex are then filtered if deemed more compatible with the primary vertex. If the track’s closest approach to the primary vertex is closer than that of the secondary vertex, the track will be discarded from the secondary vertex; if the angular separation between the track and the direction of flight of the secondary vertex is greater than 0.4, the track will also be discarded.
- **Vertex refitting and cleaning:** the secondary vertex is then refitted if it still has more than two tracks after track arbitration, and a track duplicate and distance filtering against the primary vertex is performed again.

- Reconstructed secondary vertices can then be associated with a jet via an angular matching of the jet momentum with the vertices' direction of flight.

Properties of the secondary vertex can then be used along with properties of various jet constituents to tag jets as heavy meson jets. Traditionally, this is done by defining cuts on the secondary vertex variables (such as vertex mass) and the expected angular distribution of jet constituents with the reconstruction direction of flight [90]. With the increase in the available computing power, machine-learning methods using exhaustive constituent-level information have become a viable method to perform jet tagging to improve the discriminating power. The DeepJet flavor tagger [91], which is the standard flavor tagger used by CMS for the LHC Run 2 campaign, takes the features of the jet constituents, the reconstructed secondary vertex within the jet cone, and jet-level reconstruction variables and event-level variables, and passed to a multi-layer, fully-connected deep neural network (DNN) classifier.

The DeepJet DNN outputs a score corresponding to how likely the jet is a b jet. A concrete tag requires a jet to pass some score threshold. A typical working point results in a b jet tagging efficiency of $\sim 80\%$ with a misidentification rate of $\sim 1\%$.

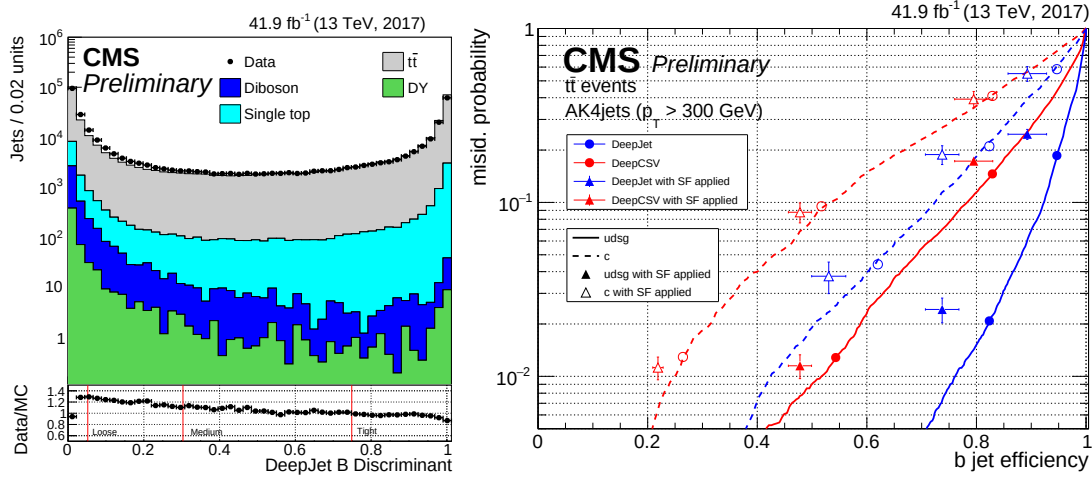


Figure 3.13: The distribution of the DeepJet DNN output score for jets of various selections using a $t\bar{t}$ -like event selection (left), and the calculated misidentification rate versus the b tagging efficiency (right), measured using 2017 data collected by CMS. The definition of “scale factor correction (SF applied)” will be discussed in Section 3.5.2.4. Plots taken from Ref. [92].

3.5 Event generation and simulation

Though it has been shown and discussed in passing, this thesis has yet to define what “simulated” events mean formally. A theoretical model would typically calculate some probability distribution of particle kinematics emerging from some hard scattering processes. Ideally, this probability distribution is propagated to the observables that we can obtain from the reconstruction algorithms to make direct comparisons with observed data. However, this propagation is impractical, given that the calculations involved require nonperturbative theories that are impossible to calculate analytically and empirical data on how particles interact with the dense matter in the detector volume. The solution is to use a Monte Carlo (MC)-based method to construct an empirical distribution of observables. Given some underlying process in the hard-scattering, many events are generated according to the theoretical model and propagated through the same computational machinery used for data collection and reconstruction to arrive at the distribution

of observables.

3.5.1 Description of computational processes

First, the **perturbative generation** of particles emerging from the hard-scattering process is calculated by sampling from the matrix element probability distributions derived from the theoretical model. Technically speaking, while the hard-scattering calculation is perturbative, the initial state of the hard scatter will need to be sampled from the *nonperturbative* parton distribution functions (Section 1.1.2) of the proton. Because this calculation is usually bundled under the same computation routine, for convenience, we will include this sampling of the parton distribution function as belonging to the perturbative generation step. The perturbative generation has both analysis-specific beyond-SM calculations (the details for this analysis will be discussed in Section 3.6.2), and common SM calculations that are expected to best represent events seen in data (discussed in Section 3.6.3).

Next, **nonperturbative generation** of color-charged particles hadronizing and undergoing secondary decays is performed. As this step heavily relies on empirical models, this set is usually handled by a computation routine separate from the perturbative generation. For this analysis, this step is handled exclusively by PYTHIA 8 [93], with the additional “CMS Pythia” parameter tuning [94] applied to account for the empirical jet distributions that are seen in CMS.

Finally, these particles are passed through **detector simulation** to derive the expected detector response. Since this step requires detailed knowledge of the CMS detector volume and operation condition, the CMS collaboration handles this step centrally to ensure all generated events see the same detector volume. The simulation of particles passing through the detector

volume is calculated using the GEANT4 [95]. While additional particles are generated when very high particles pass through the detector material, when referring to **final state particles** in analysis, this refers to the particles that are used as inputs to the detector simulation.

Given the description given above, the concept of generation and simulation is technically distinct. However, this thesis will interchangeably refer to “events created for the deriving empirical distributions via Monte Carlo methods” as generated events, simulated events, or simply MC events, given that these concepts will nearly always be used together, even if this is not the most technically accurate phrase to use. We will, however, draw the distinction between “generated” particles (particles produced in the perturbative and nonperturbative processes) and “simulated” particles (particles simulated to be generated from detector interaction). For analysis purposes, we will almost exclusively use generated particles, with exceptions being explicitly highlighted.

3.5.2 Simulated object corrections

While the same reconstruction algorithm is used for the generated events after detector response simulation, the final result of the reconstruction algorithm is known to be slightly different compared with what is obtained in data. This is because the modeling of the detector volume is imperfect. Correction factors are typically applied directly to the reconstructed physics object in the simulated samples so that the overall distribution to the object variables of interest seen in MC events better matches what is observed in data.

Which correction factors should be applied depends on the physics objects of interest in the analysis. While official corrections are provided for commonly used physics objects and their variables, corrections are not exhaustive. Physics object corrections that are uniquely determined

for the EMJ analysis will be discussed in Section 6.2. The following sections describe the official object correction factors that are used for this analysis.

3.5.2.1 Trigger efficiency scaling

As the trigger uses a modified version of the reconstruction algorithms to ensure data throughput during detector operation, triggers have a non-unity efficiency compared to the target physics variable computed using the full reconstruction algorithm. For example, when looking at a trigger which requires an event H_T threshold of 1050 GeV, it can be that reconstructed H_T is 1200 GeV but was discarded by the trigger system due to imperfect reconstructed during trigger evaluation.

To calculate the efficiency of a target trigger T , a trigger that operates on a separate physics object is used as the reference trigger R . The **relative trigger efficiency** of the target trigger can be defined as:

$$E(\vec{\theta}) = \frac{N_{T\&R}(\vec{\theta})}{N_R(\vec{\theta})}, \quad (3.15)$$

where $N_X(\vec{\theta})$ counts the number of events that pass trigger condition X that falls into some event variables bin according to $\vec{\theta}$. Due to simulation imperfections, the trigger efficiency is expected to be slightly different in data compared with simulated events. A trigger scale factor is then computed as:

$$F(\vec{\theta}) = \frac{E_{\text{data}}(\vec{\theta})}{E_{\text{MC}}(\vec{\theta})}, \quad (3.16)$$

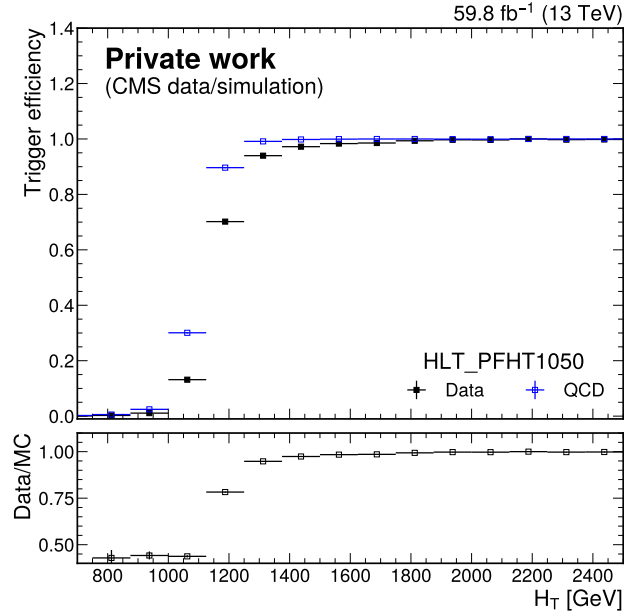


Figure 3.14: The relative trigger efficiency of HLT_PFHT1050 calculated with HLT_Mu50 as the reference trigger, evaluated using CMS data collected in 2018 and the corresponding QCD MC sample.

and should be applied to all MC events selected using the target trigger. An example of the calculation of trigger efficiencies and data–simulation correction scale factor can be found in Fig. 3.14, with the target trigger requiring an event $H_T \gtrsim 1050$ GeV, and a reference trigger requiring a muon with $p_T \gtrsim 50$ GeV. The region where the trigger efficiency grows from ~ 0 to ~ 1 is called the **trigger turn-on region**, and is typically avoided by most analyses, as it can contain significant data simulation mismatches that can result in excessively large correction factors.

This argument of applying corrections to selection efficiency mismatch between data and simulation can be applied to all selection criteria used throughout an analysis. Each analysis should evaluate whether such efficiency mismatches are expected to impact the final result. For the EMJ analysis, concerns of such mismatches are discussed in detail in Section 6.2.

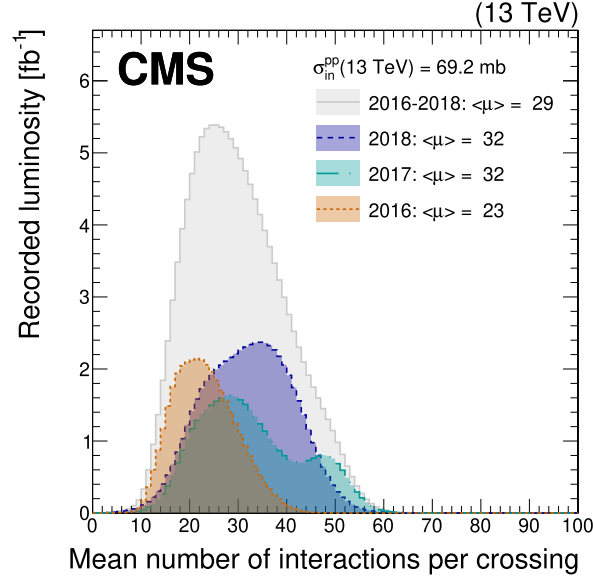


Figure 3.15: The pileup distribution of the data collected in CMS, calculated assuming $\sigma_{\text{min. bias}} = 69.2 \text{ mb}$. For 2016 (orange), 2017 (green), 2018 (blue) and the full Run 2 (gray). The plot is taken from Ref. [96]

3.5.2.2 Pileup reweighting

Pileup interactions are injected into the simulated events following some predetermined distribution of the expected pileup scenario. The true pileup distribution, however, is difficult to predict accurately, as it is determined by LHC beam conditions during machine operation (Fig. 3.15). As the number of pileup interactions directly impacts the performance of many reconstruction algorithms, ensuring that the simulated events have a pileup distribution that matches that of data is important when attempting to draw conclusions using MC events.

The pileup of data is calculated by comparing the measured instantaneous luminosity and the theoretical inclusive “minimum bias” pp scattering cross section $\sigma_{\text{min. bias}} = 69.2 \text{ mb}$. An event weight can then be derived by comparing the data pileup distribution with the distribution injected into simulation MC events.

The effectiveness of pileup reweighting can be validated by comparing the number of re-

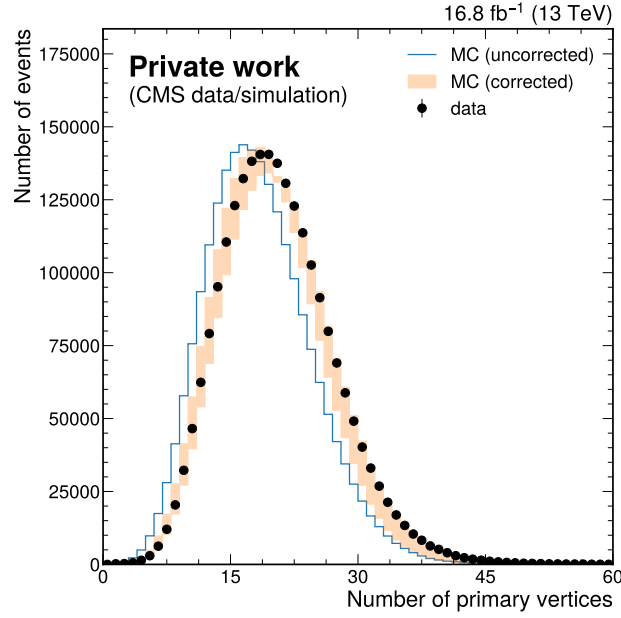


Figure 3.16: Distribution of the number of reconstructed primary vertices in H_T -triggered data collected in 2016 post-APV fix (points) and QCD MC with/without pileup reweighting (orange/blue). Both data and QCD MC events only require the H_T -trigger requirements as defined in Section 4.1.1.

constructed primary vertices of events in MC samples and data. An example of this distribution can be found in Fig. 3.16, and we can see how pileup reweighting better corrects the expected distribution of the vertex multiplicity when comparing data with MC.

3.5.2.3 Jet energy smearing

Those additional jet energy corrections are already applied to jets in data, as described in Section 3.3.1.2, the jet energy resolution of jets in simulation is typically too good compared with data. Momenta of simulated jets are artificially “smeared” by some scaling factor such that the randomness seen in simulated events better matches what is observed in data. Let us notated the average JER of jets in data and in MC, which are both functions of jet p_T as jet η , as $\sigma_{\text{JER}}^{\text{data}}(p_T, \eta)$ and $\sigma_{\text{JER}}^{\text{MC}}(p_T, \eta)$, respectively, and defined the relative factor $s_{\text{JER}}(p_T, \eta) =$

$\sigma_{\text{JER}}^{\text{data}}(p_T, \eta)/\sigma_{\text{JER}}^{\text{MC}}(p_T, \eta)$. Two smearing schemes are employed by CMS depending on the condition of the reconstructed jet in MC:

1. The reconstructed jet in MC is “well-reconstructed” if it can be matched with a jet clustered using final-state generated particles via momentum matching. Specifically, the criteria for well-reconstructed is defined jet as:

$$\Delta R(\text{reco.}, \text{gen.}) < 0.2 ; \quad \left| \frac{p_T^{\text{reco.}} - p_T^{\text{gen.}}}{p_T^{\text{reco.}}} \right|. \quad (3.17)$$

where $p_T^{\text{reco.}}$ is the momentum obtained from the clustering of reconstructed particles, $p_T^{\text{gen.}}$ is the momentum obtained from clustering generated particles. For well-reconstructed jets, the jet momentum is scaled by the factor:

$$1 + (s_{\text{JER}}(p_T^{\text{reco.}}, \eta^{\text{reco.}}) - 1) \left(\frac{p_T^{\text{reco.}} - p_T^{\text{gen.}}}{p_T^{\text{reco.}}} \right). \quad (3.18)$$

2. If the jet is not well-reconstructed, then a stochastic smearing is performed. The jet momentum is scaled by a factor of:

$$1 + x \sqrt{\max \{s_{\text{JER}}^2(p_T, \eta) - 1, 0\}}; \quad (3.19)$$

where x is a random variable sampled from a normal distribution with a standard deviation of $\sigma_{\text{JER}}^{\text{MC}}(p_T, \eta)$.

The scale momentum of jets will be used as the central value of jet kinematics when working with jets in MC.

3.5.2.4 DeepJet score corrections

The output of the DeepJet DNN is expected to differ in data and simulated events. The extent of this difference can be measured by imposed kinematic cuts on the event final set to construct regions with different b jet fractions and comparing the final DeepJet score distribution [97]. The inclusive sample of events requiring a single jet with p_T greater than some threshold can be used to construct a region dominated by light-flavored jets. To construct a region dominated by b jets or light jets, various selections can be imposed on the jet of interest: for example, we can require that at least one high- p_T muon exists within the jet cone, as heavy-flavored mesons have a large branching fraction to muons, and compare the relative transverse momentum of the muon relative to the jet axis (“PTRel” method); similarly, we can explicitly pick jets to include tracks with large negative impact parameter values, as such displaced vertices are typically a result of misreconstruction rather than a true displaced vertex (“Negative tag” method).

Figure 3.17 demonstrates the measurement of b jet tagging efficiency and light jet misidentification rates using data. The ratio between the rates evaluated in data and simulated events will be taken as the correction scale factor.

For this analysis, the use of DeepJet b tagging will be formally discussed in Chapter 5. The b tagging is not used to determine some absolute number of events passing a b tagging threshold, but uses the discriminator to determine the relative number of events. The correction factor used in this analysis interprets the data–simulation differences as a per-jet weighting factor that is a function of jet p_T , η , and the DeepJet score, such that the distribution of the discriminator in MC jets matches what is observed in data. An example of this correction scale factor is given in Fig. 3.18.

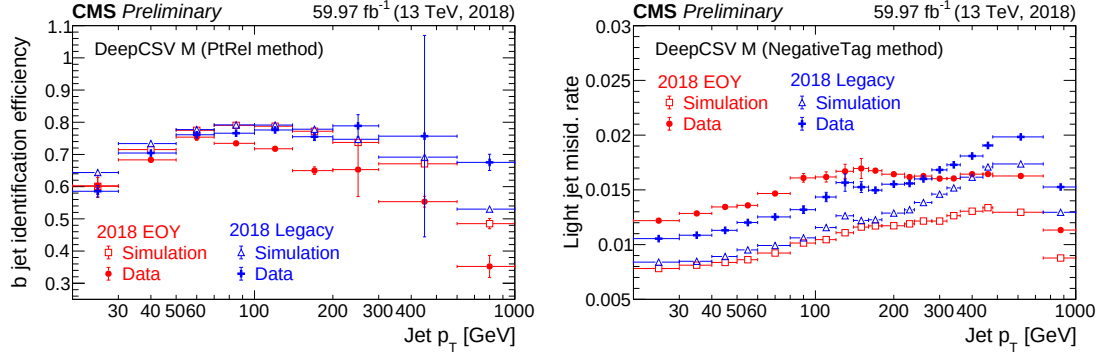


Figure 3.17: The comparison between the b jet tagging efficiency (left) and the light jet misidentification rate (right), measured using 2018 data using the (solid) corresponding MC events (open) as a function of jet p_T . For data, the b jet tagging efficiency is measured using the “PTRel” method, and for the light jet misidentification rate is measured using the “Negative tag” method described in Ref. [97]. The two colors correspond to different data processing campaigns with modified reconstruction parameter tuning. For this analysis, we use the results corresponding to the “legacy” processing. Plots are taken from Ref. [98]

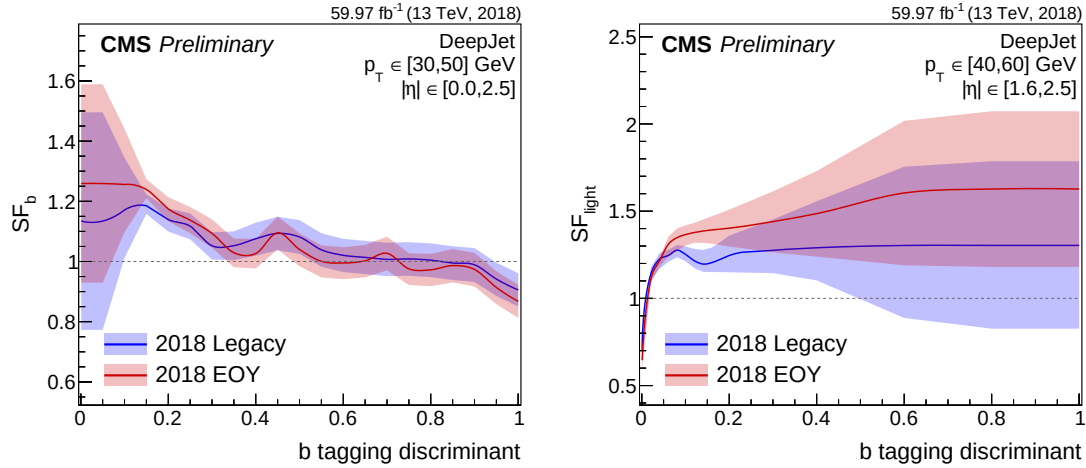


Figure 3.18: The calculated scale factor that should be applied to MC truth b jets (left) and MC truth light jets (right) to account for tagging efficiency differences for 2018. The two colors correspond to different data processing campaigns with modified reconstruction parameter tuning. For this analysis, we use the results corresponding to the “legacy” processing. Plots are taken from Ref. [98]

3.6 Data sets used for the EMJ analysis

3.6.1 Data collected by CMS

This analysis primarily uses events that pass some energetic jet triggers, labeled the H_T -triggered data set in this thesis. Details of what selections will be imposed for such events and how it will be used will be discussed in Section 4.1.1. CMS labels all events that pass some energetic jet triggers as the “JetHT data stream.”

This analysis will also use events that pass some energetic photon trigger, labeled the γ -triggered data set in this thesis. Details of what selection will be imposed for such events and why we will use these events will be discussed in Section 5.1.1. CMS labels all events that pass some energetic photon trigger as the “SinglePhoton data stream” in 2016–2017, and all events that pass some energetic ECAL cluster object as the “EGamma data stream” in 2018.

While the usage of the two data sets is typically kept orthogonal for this analysis, if ambiguity ever arises when comparing results from the two data sets in the same plot or equation, the shorthands “ H_T -data” and “ γ -data” will be used to index the data sets respectively.

The complete list of data collected by CMS used for this analysis is listed in Tables 3.2–3.3. In particular, the data sets of the data labeled as “2016-HIPM” correspond to data collected pre-APV fixes, and those labeled with simply “2016” correspond to data collected post-APV fixes (check Section 3.1.1).

Table 3.2: List of data collected by CMS using an energetic jet trigger that is used for the EMJ analysis. The labels in the first column are used to indicate the major and minor detector configuration changes deployed during CMS operation.

data set label	Integrated Luminosity
/JetHT/Run2016B-ver2_HIPM_UL2016_MiniAODv2-v2/ /JetHT/Run2016C-HIPM_UL2016_MiniAODv2-v2/ /JetHT/Run2016D-HIPM_UL2016_MiniAODv2-v2/ /JetHT/Run2016E-HIPM_UL2016_MiniAODv2-v2/ /JetHT/Run2016F-HIPM_UL2016_MiniAODv2-v2/	19.5 fb ⁻¹
/JetHT/Run2016F-UL2016_MiniAODv2-v2/ /JetHT/Run2016G-UL2016_MiniAODv2-v2/ /JetHT/Run2016H-UL2016_MiniAODv2-v2/	16.8 fb ⁻¹
/JetHT/Run2017B-UL2017_MiniAODv2-v1/ /JetHT/Run2017C-UL2017_MiniAODv2-v1/ /JetHT/Run2017D-UL2017_MiniAODv2-v1/ /JetHT/Run2017E-UL2017_MiniAODv2-v1/ /JetHT/Run2017F-UL2017_MiniAODv2-v1/	41.5 fb ⁻¹
/JetHT/Run2018A-UL2018_MiniAODv2-v1/ /JetHT/Run2018B-UL2018_MiniAODv2-v1/ /JetHT/Run2018C-UL2018_MiniAODv2-v1/ /JetHT/Run2018D-UL2018_MiniAODv2-v2/	59.8 fb ⁻¹

3.6.2 Signal model simulated events

The generation of the event topology described in Section 1.2, along with the hadronization process of the dark sector particles, are performed by the hidden-valley model [99, 100] implemented in PYTHIA 8 [93]. The generator configuration explicitly specifies the branching ratio of dark mesons to SM quarks. The decay of subsequent hadronization of the SM particles is also handled by PYTHIA.

For the unflavored criteria described in Section 1.2, we fix the number of dark sector fermions to be $N_f = 7$ with degenerate mass, the dark sector gauge interaction to be $SU(N_{\text{dark}} = 3)$, the dark confinement scale to be 2 times the dark meson mass, $\Lambda_{\text{dark}} = 2m_{\pi_{\text{dark}}}$. The dark mesons are explicitly set to decay only to SM d quarks. The unflavored scenario then has three remaining free parameters: The mass of the SM-DS mediator bifundamental $m_{X_{\text{dark}}}$, the mass of the dark meson $m_{\pi_{\text{dark}}}$, and the lifetime of the dark pion $c\tau_{\pi_{\text{dark}}}$. This analysis performs a survey of

Table 3.3: List of data collected by CMS using an energetic photon trigger (2016, 2017) and energetic ECAL objects (2018) that are used for the EMJ analysis. The labels in the first column are used to indicate major and minor detector configuration changes deployed during CMS operation.

data set label	Integrated Luminosity
/SinglePhoton/Run2016B-ver1_HIPM_UL2016_MiniAODv2-v2/	19.5 fb ⁻¹
/SinglePhoton/Run2016B-ver2_HIPM_UL2016_MiniAODv2-v2/	
/SinglePhoton/Run2016C-HIPM_UL2016_MiniAODv2-v2/	
/SinglePhoton/Run2016D-HIPM_UL2016_MiniAODv2-v2/	
/SinglePhoton/Run2016E-HIPM_UL2016_MiniAODv2-v2/	
/SinglePhoton/Run2016F-HIPM_UL2016_MiniAODv2-v2/	16.8 fb ⁻¹
/SinglePhoton/Run2016F-UL2016_MiniAODv2-v2/	
/SinglePhoton/Run2016G-UL2016_MiniAODv2-v2/	
/SinglePhoton/Run2016H-UL2016_MiniAODv2-v2/	
/SinglePhoton/Run2017B-UL2017_MiniAODv2-v1/	41.5 fb ⁻¹
/SinglePhoton/Run2017C-UL2017_MiniAODv2-v1/	
/SinglePhoton/Run2017D-UL2017_MiniAODv2-v1/	
/SinglePhoton/Run2017E-UL2017_MiniAODv2-v1/	
/SinglePhoton/Run2017F-UL2017_MiniAODv2-v1/	
/EGamma/Run2018A-UL2018_MiniAODv2-v1/	59.8 fb ⁻¹
/EGamma/Run2018B-UL2018_MiniAODv2-v1/	
/EGamma/Run2018C-UL2018_MiniAODv2-v1/	
/EGamma/Run2018D-UL2018_MiniAODv2-v2/	

Table 3.4: Table of model parameters in the unflavored scenario survey in this analysis. A total of 240 possible parameter configurations were surveyed for the unflavored scenario.

model parameter	values surveyed
$m_{X_{\text{dark}}}$ [GeV]	1000, 1200, 1400, 1500, 1600, 1800, 2000, 2200, 2400, 2500
$m_{\pi_{\text{dark}}}$ [GeV]	10, 20
$c\tau_{\pi_{\text{dark}}}$ [mm]	1, 2, 5, 25, 45, 60, 100, 150, 225, 300, 500, 1000

possible scenarios by a grid of parameter values listed in Table 3.4.

For the flavored-aligned scenario criteria described in Section 1.2, we fix the dark sector gauge interaction to be $SU(N_{\text{dark}} = 3)$, the dark confinement scale to be 2 times the dark meson mass, $\Lambda_{\text{dark}} = 2m_{\pi_{\text{dark}}}$, and the number of dark fermion flavors to be $N_f = 3$ such that it directly matches the number of SM quark generations. The flavored-aligned scenario then has three remaining free parameters: The mass of the SM-DS mediator bifundamental $m_{X_{\text{dark}}}$, the mass of the dark meson $m_{\pi_{\text{dark}}}$, and the overall magnitude of the SM-DS Yukawa coupling matrix κ_0 . As κ_0 is not an easily understandable value in terms of the expected detector signature, the

corresponding lifetime of the $\pi_{\text{dark}}(\overline{Q}_1 Q_2) \rightarrow d s$ as calculated using Eq. 1.19, which we label as $c\tau_{\pi_{\text{dark}}, \text{max}}$, will be used to label the data sets. This analysis surveys possible scenarios using a grid of parameter values in Table 3.5.

For each of the signal parameters listed in Tables 3.4 and 3.5, 5000 events are generated and simulated for each of the major detector configurations used to collect data in 2016 pre-APV fix, 2016 post-APV fix, 2017, and 2018.

3.6.3 Standard model simulated events

To match the data set collected by the energetic jets trigger, the dominant SM contribution to such events is expected to be from multi-jet events generated through QCD color interactions. A dedicated set of SM events generated by CMS is used to match this process. These simulated events use MADGRAPH [101] to calculate the matrix elements and PYTHIA to handle the hadronization particle shower. The shorthand label used for these samples in this thesis would typically be “QCD MC samples,” with events called “QCD MC events” to indicate that only SM QCD color interactions were used to generate events in such samples. In plots or equations, samples and events may simply be indexed as “QCD (MC)” or “QCD.”

To match the data set collected using an energetic photon trigger, the dominant SM contribution to such events will be multi-jet events with additional high-energy photons generated through initial-state radiation or final-state radiation. A dedicated set of SM events is generated by CMS using MADGRAPH to calculate the matrix elements and PYTHIA to handle the hadronization particle show. The shorthand label used for these samples in this thesis would typically be “ γ +Jets MC samples,” or in places where it may be confused with the γ -triggered data collected

Table 3.5: Full list of parameters used for the flavored-aligned scenario model. The lifetime listed in the column header is the maximum lifetime of the 4 dark pion flavors. The values under $c\tau_{\pi_{\text{dark}},\text{max}}$ are the choice of κ_0 in model generation.

$m_{X_{\text{dark}}}$ [GeV]	$m_{\pi_{\text{dark}}}$ [GeV]		$c\tau_{\pi_{\text{dark}},\text{max}}$ [mm]				
		–	5	25	45	100	500
			κ_0 value				
1000	6	1	0.92	0.61	0.53	0.43	0.29
	10	1	0.62	0.42	0.36	0.30	0.20
	20	1	0.37	0.25	0.21	0.18	0.12
1200	6	1	1.10	0.73	0.63	0.52	0.35
	10	1	0.75	0.50	0.43	0.35	0.24
	20	1	0.45	0.30	0.26	0.21	0.14
1400	6	1	1.28	0.86	0.74	0.61	0.41
	10	1	0.87	0.58	0.50	0.41	0.28
	20	1	0.52	0.35	0.30	0.25	0.16
1600	6	1	1.47	0.98	0.85	0.69	0.46
	10	1	1.00	0.67	0.58	0.47	0.32
	20	1	0.59	0.40	0.34	0.28	0.19
1800	6	1	1.65	1.10	0.95	0.78	0.52
	10	1	1.12	0.75	0.65	0.53	0.36
	20	1	0.67	0.45	0.39	0.32	0.21
2000	6	1	1.83	1.23	1.06	0.87	0.58
	10	1	1.25	0.84	0.72	0.59	0.40
	20	1	0.74	0.50	0.43	0.35	0.23
2200	6	1	2.02	1.35	1.16	0.95	0.64
	10	1	1.37	0.92	0.79	0.65	0.43
	20	1	0.82	0.55	0.47	0.39	0.26
2400	6	1	2.20	1.47	1.27	1.04	0.70
	10	1	1.50	1.00	0.87	0.71	0.47
	20	1	0.89	0.60	0.51	0.42	0.28
2500	6	1	2.29	1.53	1.32	1.08	0.72
	10	1	1.56	1.04	0.90	0.74	0.49
	20	1	0.93	0.62	0.54	0.44	0.29

by CMS, the label “GJets MC samples” may also be used. In plots or equations, samples and events can be simply indexed as “ γ +Jets (MC)” or “GJets.”

The QCD and γ +jets MC samples are binned with cuts on the generated H_T . This ensures the number of events is sufficiently large for energetic events. The GJets MC sample also requires that the generated photon has a ΔR separation of 0.4 with the leading jet. The complete list of SM events generated and simulated by CMS used in this analysis is shown in Table 3.6. The first column of the table corresponds to the label used to index the major detector configuration changes corresponding to what was deployed for CMS data taking. These correspond to the data collected in 2016 pre-APV fix, 2016 post-APV fix, 2017 and 2018.

Table 3.6: List of simulated SM samples used for the EMJ analysis. The effective luminosity (final column) is calculated by comparing the total number of events in the sample with the theoretical cross section of the process of interest.

data collection era configuration label	process label	cross section [pb]	Eff. $\mathcal{L}$ [fb^{-1}]
RunI Summer20UL16MiniAODv2-106X_mcRun2_asymptotic_preVFP_v11-v1	QCD_HT700to1000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	6403.0	2.47
	QCD_HT1000to1500_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	1117.0	12.4
	QCD_HT1500to2000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	108.4	92.6
	QCD_HT2000toInf_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	21.93	227.8
	GJets_DR-0p4_HT-100To200_TuneCP5_13TeV-madgraphMLM-pythia8	5030.0	1.01
	GJets_DR-0p4_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8	1128.0	13.0
	GJets_DR-0p4_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8	126.7	32.5
	GJets_DR-0p4_HT-600ToInf_TuneCP5_13TeV-madgraphMLM-pythia8	41.39	95.1
	QCD_HT700to1000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	6403.0	2.17
	QCD_HT1000to1500_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	1117.0	11.3
RunI Summer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v1	QCD_HT1500to2000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	108.4	86.5
	QCD_HT2000toInf_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	21.93	222.0
	GJets_DR-0p4_HT-100To200_TuneCP5_13TeV-madgraphMLM-pythia8	5030.0	0.878
	GJets_DR-0p4_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8	1128.0	12.2
	GJets_DR-0p4_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8	126.7	32.5
	GJets_DR-0p4_HT-600ToInf_TuneCP5_13TeV-madgraphMLM-pythia8	41.39	82.8
	QCD_HT700to1000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	6403.0	5.32
	QCD_HT1000to1500_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	1117.0	9.18
	QCD_HT1500to2000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	108.4	71.1
	QCD_HT2000toInf_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	21.93	187.5
RunI Summer20UL17MiniAODv2-106X_mc2017_realistic_v9-v1	GJets_DR-0p4_HT-100To200_TuneCP5_13TeV-madgraph-pythia8	5030.0	2.00
	GJets_DR-0p4_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8	1128.0	30.0
	GJets_DR-0p4_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8	126.7	71.2
	GJets_DR-0p4_HT-600ToInf_TuneCP5_13TeV-madgraphMLM-pythia8	41.39	201.3
	QCD_HT700to1000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	6403.0	7.58
	QCD_HT1000to1500_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	1117.0	13.0
	QCD_HT1500to2000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	108.4	100.3
	QCD_HT2000toInf_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	21.93	245.1
	GJets_DR-0p4_HT-100To200_TuneCP5_13TeV-madgraphMLM-pythia8	5030.0	2.84
	GJets_DR-0p4_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8	1128.0	42.1
RunI Summer20UL18MiniAODv2-106X_upgrade2018_realistic_v16_L1v1-v2	GJets_DR-0p4_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8	126.7	106.1
	GJets_DR-0p4_HT-600ToInf_TuneCP5_13TeV-madgraphMLM-pythia8	41.39	282.9
	QCD_HT700to1000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	6403.0	7.58
	QCD_HT1000to1500_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	1117.0	13.0
	QCD_HT1500to2000_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	108.4	100.3
	QCD_HT2000toInf_TuneCP5_PSWWeights_13TeV-madgraph-pythia8	21.93	245.1
	GJets_DR-0p4_HT-100To200_TuneCP5_13TeV-madgraphMLM-pythia8	5030.0	2.84
	GJets_DR-0p4_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8	1128.0	42.1
	GJets_DR-0p4_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8	126.7	106.1
	GJets_DR-0p4_HT-600ToInf_TuneCP5_13TeV-madgraphMLM-pythia8	41.39	282.9

Chapter 4: Emerging jet signature

The most critical component of this analysis would be the development and validation of a jet-level tagging scheme to distinguish jets arising from dark showers according to the emerging jet (EMJ) signature described in Section 1.2. Previously, the EMJ signature was described as jets where particles emerge from multiple vertices with a significant distance-of-flight from the hard scattering vertex. However, attempting to reconstruct such displaced vertices yields inconsistent results in signal selection. The tracking and vertexing algorithms described in Sections 3.1.1, 3.1.2, and 3.4.4 were designed for the most typical use cases for CMS analyses. The standard reconstruction algorithm aims to accommodate long-lived SM mesons, which generate displacement vertices well within the beam line. The reconstruction efficiency falls rapidly for tracks with large displacements [76]. Naively modifying the standard CMS algorithms can potentially generate erroneous physics objects, as the proper validation of the track reconstruction is a difficult task requiring dedicated studies.

To achieve reliable EMJ tagging without sacrificing the signal signature of interest, this search will work with the track displacement variables defined in Section 3.1.2 rather than attempting to reconstruct displaced vertices. Since the vertex reconstruction efficiency falls much faster than the track reconstruction efficiency, working with tracks was found to yield consistent results without modifications to the standard CMS reconstruction algorithms. Section 4.2

will describe how jet-level variables are derived from the displacement variables of associated tracks and how they will be used for tagging the EMJ signature; Section 4.3 will discuss how the cut-points of various variables are determined to best search for potential signal events in data; finally, Section 4.4 we will go over known sources of displacement tracks that can cause SM jets to be “mistagged” as EMJs.

4.1 Event-level selections

Before defining the EMJ signature, let us define the common selection criteria that will be used to define the baseline event signature of our analysis, as well as serve as a common working point for comparing data with simulated events.

4.1.1 Trigger requirements

First, let us define the trigger condition that is used for data collection. Given the event signature discussed in Section 1.3, the trigger employed for this analysis requires a reconstructed H_T greater than a defined threshold value. We will use the following trigger configuration in this analysis:

- For the 2016 data, the high-level trigger logic used is `HLT_PFHT900_v* | HLT_PFJet450_v*`, which requires events have a reconstructed $H_T > 900$ GeV, or a reconstructed jet with $p_T > 450$ GeV. The logical “or” of an additional jet-based trigger is to compensate for a misconfigured L1 setting for 2016 data collection.
- For the 2017 and 2018 data, the high-level trigger logic used is `HLT_PFHT1050_v*`, which requires event have a reconstructed $H_T > 1050$ GeV. The raised H_T threshold

was deployed to keep event rates manageable with the increased instantaneous luminosity compared to 2016.

Unless otherwise specified, we impose the minimal offline reconstructed $H_T > 1200$ GeV to avoid trigger turn-on region. The common offline H_T threshold also avoids biases of events obtained in 2016, 2017, and 2018 introduced by the different online trigger thresholds. An example of the trigger turn-on effect for intermediate event H_T , calculated in data and simulation can be found in Fig. 4.1.

4.1.2 Primary vertex selection

Unless otherwise specified by the analysis, we will use the singular “the primary vertex,” referring to the primary vertex that has the largest weighted- $\sum p_T^2$ of associated particles, as described in Section 3.4.1.

With the standard vertex quality imposed during the vertex reconstruction, there is still a chance that the vertex can be misidentified, as we can demonstrate by plotting the distance between the interaction vertex used for event generation and the reconstructed vertex position, as shown in Fig. 4.2. The additional structure at $\sim 10^{-2}$ cm matches the minimum separation of vertices as described in Section 3.1.2, indicating that this is likely driven by a misidentification of a pileup vertex as the hard-scattering vertex. Additional vertex qualities will, therefore, be imposed on the primary vertex to ensure event quality:

- The primary vertex must be constructed in the central region of the CMS detector volume: requiring that the z coordinate of the primary vertex satisfies $|z_{PV}| < 15$ cm.
- The fraction of tracks considered prompt in the z direction must be greater than 10%. The

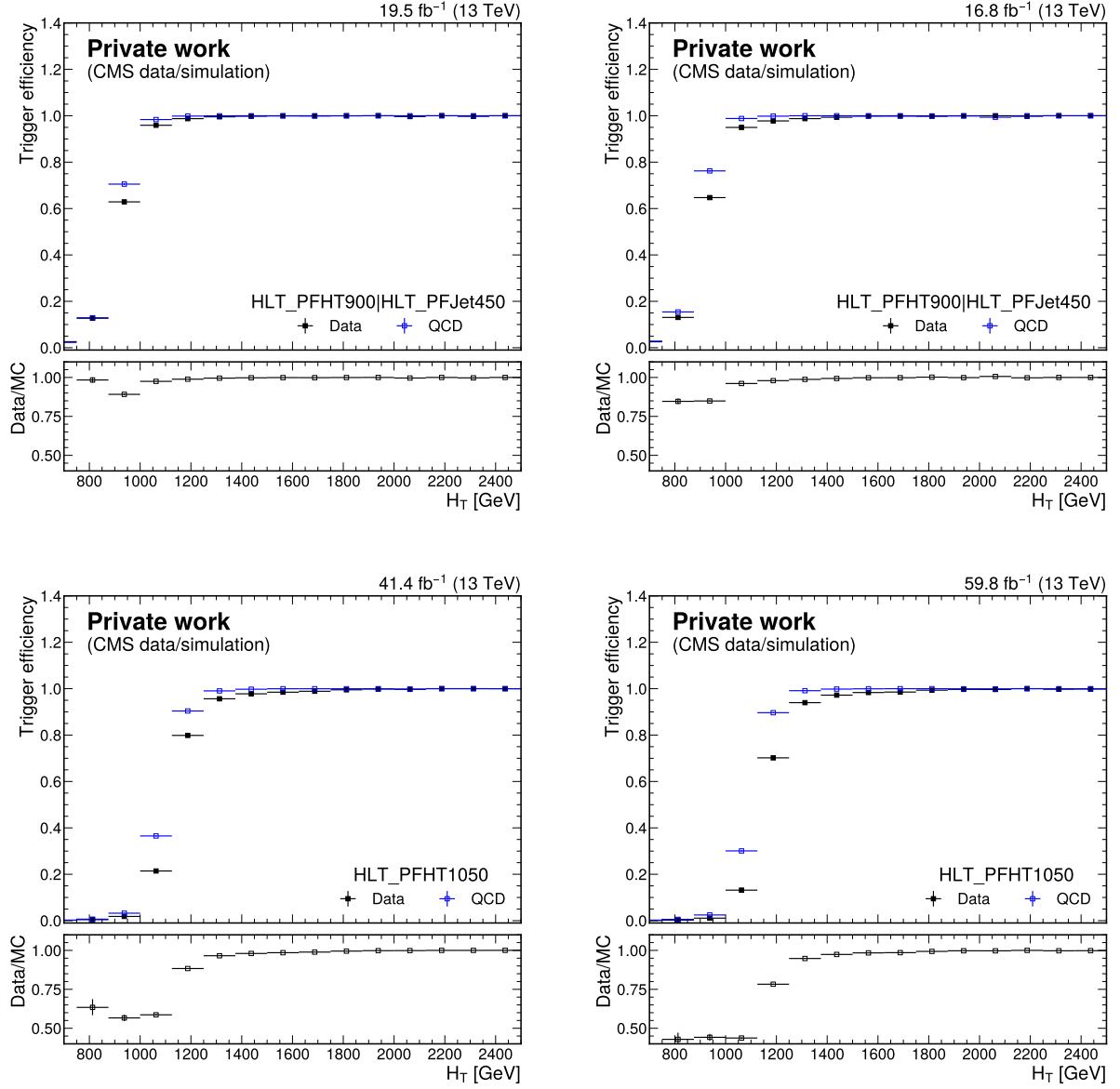


Figure 4.1: Trigger efficiency evaluated for the triggers used for the analysis for 2016 pre-APV fixes (top left), 2016 post-APV fixes (top right), 2017 (bottom left) and 2018 (bottom right), evaluated using data (black) and QCD MC samples (gray).

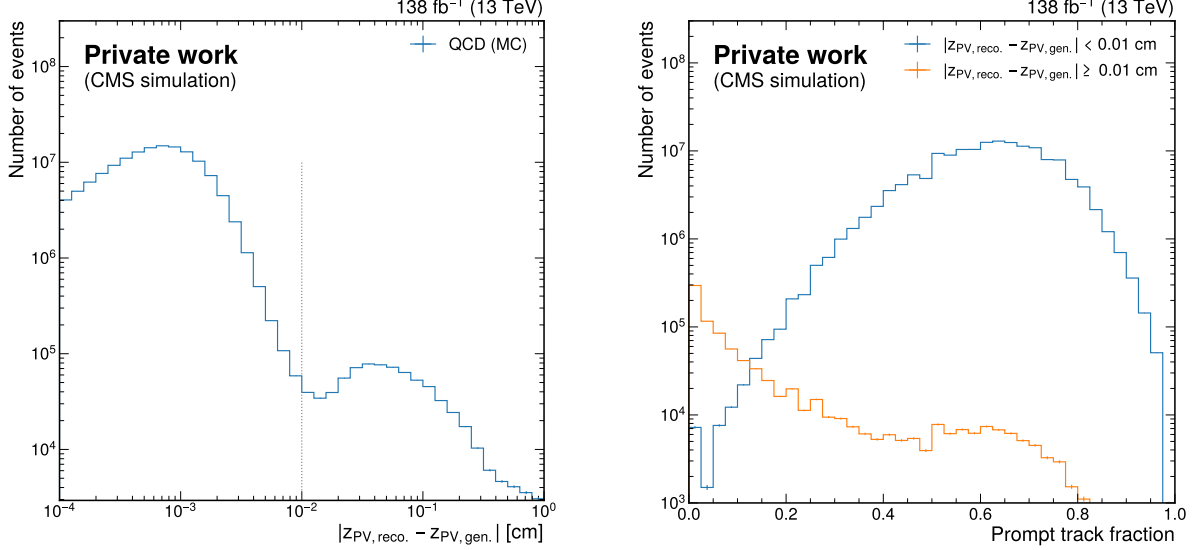


Figure 4.2: Plot on the left displays the distribution of separation between the hard scatter vertex used for event generation (“GEN”-vertex) in simulation and the reconstructed vertex obtained with the information available after the simulation of detector response (“RECO”-vertex). Plot on the right displays the prompt track fraction distribution (defined in Eq. 4.1) for events with “well-reconstructed” vertices (z separation of GEN-vertex and RECO-vertex is smaller than 0.01 cm) and poorly-reconstructed vertices in blue and orange, respectively. Plots are generated using events in the QCD MC samples.

prompt track fraction is defined as:

$$F = \frac{n_{\text{track}}(|IP_z| < 0.01 \text{ cm})}{n_{\text{track}}}, \quad (4.1)$$

where n_{track} count the number of tracks associated with selected jets per the criteria given in the parentheses. This vertex requirement was found to be an effective way of filtering out poorly reconstructed events while having a minimal impact on signal selection efficiency.

4.1.3 Custom Jet-track association

The EMJ signature is characterized by jets containing multiple tracks with large displacements. The standard jet clustering algorithm and post-processing routines used by CMS may be

inadequate, as the algorithms can potentially discard tracks of interest from the jets, such as with the CHS correction described in Section 3.3.1.2. However, several standard CMS jet variables are still required for the analysis, in particular, the well-studied b jet tagging tools.

For this analysis, the standard CMS jets (jets clustered with the anti- k_T algorithm with an angular parameter of 0.4 with all standard corrections described in Section 3.3.1.2 applied) will be used as the starting objects for event selection. The EMJ tagging methods, however, will look at tracks associated with jets determined by a custom association algorithm, effectively using the tracks in the vicinity of the determined energy centers to perform signature tagging rather than simply taking the jet constituents themselves. This is a compromise that minimizes the need to re-compute and re-validate variables used for physics object quality assurance while also improving signal efficiency.

Here let us list the common jet selection criteria, as well as the jet–track association scheme with a working angular association radius R_{cut} .

1. Jets must have $p_T > 100 \text{ GeV}$, $|\eta| < 2.0$ and pass the jet identification requirements listed in Section 3.4.3. The selection of jet η ensures that tracks associated with the jet will be contained within the track volume.
2. Jets within the same event selected by the previous kinematic restrictions must be mutually exclusive within an angular radius $\Delta R < R_{\text{cut}}$, to ensure that track features are not shared between jets, which may lead to correlated tagging results. If multiple jets are selected within $\Delta R < R_{\text{cut}}$ of each other, the jet with the highest p_T will be used, while other jets will be discarded.
3. Tracks are required to have at least $p_T > 1 \text{ GeV}$, and be flagged by the PF algorithm as

being high purity (as discussed in Section 3.1.1). Tracks are also required to pass IP_z cut of $|IP_z| < IP_{z,\text{cut}}$ to avoid pileup tracks from impacting the EMJ tagging results. The value of $IP_{z,\text{cut}}$ used for track selection will be discussed below.

4. Tracks are associated with the selected jets if it is within an angular range $\Delta R < R_{\text{cut}}$. This association is unique, so if a track can be associated with multiple selected jets, it will be assigned to the closest jet in ΔR space. Uniqueness ensures that displaced tracks are not shared between jets, which can result in tagging correlation. The value of ΔR_{cut} used for jet-track association will be discussed below.
5. Jets are further pruned by requiring at least one associated track, and none of the associate tracks can have $p_{T,\text{track}} > 0.6 \cdot p_{T,\text{jet}}$. While the momentum fraction selection was supposedly enforced in the jet identification selection, we will need to reintroduce this selection as we may include additional tracks in the jet-track association from the previous step.

Using the algorithm above, each selected jet will have a list of associated tracks we will use to compute jet-level variables to jets as EMJs. The algorithm currently has two adjustable variables: R_{cut} and $IP_{z,\text{cut}}$. Their determination will be discussed in Section 4.2. The following sections will describe how the list of tracks will be used to compute a jet-level variable for EMJ tagging.

The issue of “tagging correlation” is also mentioned multiple times in the jet-track association algorithm, and will be discussed in greater details in the next chapter. One can think that if EMJ tagging is correlated, the possibility of getting two EMJ-tagged jets in an event may be significantly enhanced, which in turn considerably impacts the signal sensitivity of the analysis. There is also the impact of correlations on the algorithm deployed for background estimation, the

details of which will be given in Section 5.2.

Using the selection criteria described in Section 4.1.1–4.1.3, we can impose the typical selection of a high- H_T event with a well-reconstructed primary vertex, and 4 high- p_T jets with associated jets that pass the selections defined above. These requirements match the basic event topology described in Section 1.3 while still being dominated by SM background processes. Unless otherwise specified, plots in this section will at least have these selections imposed. Additional H_T and jet p_T selection can be imposed to boost signal sensitivity further, as can be seen in Fig. 4.3. How this will be employed to better enhance signal sensitivity is discussed in Section 4.3. In the next section, we will describe how the EMJ signature can be defined to appropriately narrow event selection to a signal-rich region.

4.2 Emerging jets tagging variables

EMJ tagging follows the same principle as what is used for SM jet flavor tagging: an algorithm is deployed to reduce a list of “constituent variables” using some reduction formula to obtain a jet-level summary variable, and whether these summary variables for a jet passes some set of thresholds are used to label jets as either passing or failing the tagging criteria. Two methods will be described here: Section 4.2.1 describes a “model-agnostic” method, where a human-readable reduction formula that should give sufficient distinguishing power for all models that produce a similar signal signature; Section 4.2.2 describes a machine learning (ML) approach, where the reduction formula is performed by an artificial neural network. This approach represents the best signal-distinguishing power for the surveyed signal models achievable at CMS.

While it might be old-fashioned to keep a “model-agnostic” method with the popularity

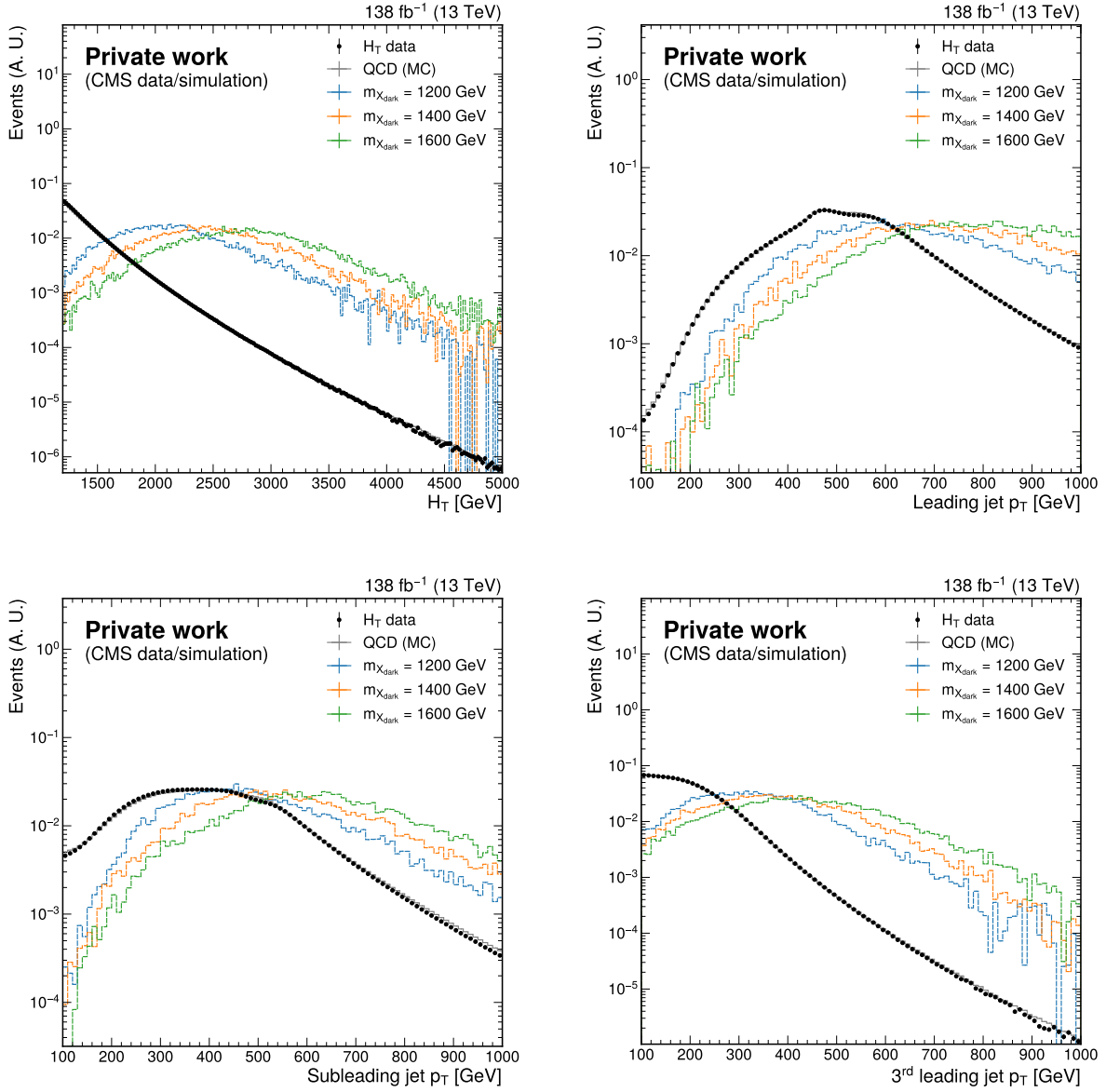


Figure 4.3: Distribution of the event H_T (top left), p_T of the leading jet (top right), p_T of the second leading jet p_T (bottom left), and p_T of the third leading jet (bottom right), for events that pass the H_T trigger requirement, the has at least 4 jets with $p_T > 100$ GeV in SM QCD sample (gray), various signal samples (colored), and data (black points). All distributions are normalized to unity to avoid guessing at excesses in the data samples.

of ML, the interpretability of ML results is still an open question and may not be the most suitable for adaption for future extension models that may have similar signatures with modified model parameters. More details of this problem will be given when discussing the final result in Chapter 7.

4.2.1 Model-agnostic approach to EMJ tagging

4.2.1.1 Variables used for tagging EMJs in the unflavored signal models

As this class of models has already been studied in the 2016 analysis performed by CMS [44], the same jet-level variables will also be used for this analysis. For this class of EMJ tagging, we employ the jet-track association algorithm is association radius $R_{\text{cut}} = 0.4$.

- The median of $|\text{IP}_{2\text{D}}|$ of the associated tracks, $\langle \text{IP}_{2\text{D}} \rangle$. The distinct peak in the $\langle \text{IP}_{2\text{D}} \rangle$ distribution can be seen when comparing jets in the signal samples to jets generated by QCD processes, as is seen in Fig. 4.4. Here we take the absolute value of $\text{IP}_{2\text{D}}$, as negative values would make it difficult to compare the displacement magnitude using simple statistical functions. The median function is chosen as it is more resilient to large fluctuations compared with the mean function.
- A p_{T} -weighted fraction of associated tracks that are considered prompt, $\alpha_{3\text{D}}$, defined using the formula:

$$\alpha_{3\text{D}} = \frac{\sum_{D_N < D_{N,\text{cut}}} p_{\text{T},i}}{\sum_i p_{\text{T},i}}, \quad (4.2)$$

where the index i runs over the tracks associated with the jet of interest, and the summation

in the numerator only runs over a subset of tracks that have D_N smaller than some cut-off value. The D_N is the normalized pseudo-significance of the track displacements, defined as:

$$D_N = \sqrt{\left(\frac{\text{IP}_z}{0.01 \text{ cm}}\right)^2 + \left(\frac{\text{IP}_{2\text{D}}}{\sigma_{\text{IP}_{2\text{D}}}}\right)^2}, \quad (4.3)$$

where $\sigma_{\text{IP}_{2\text{D}}}$ is the uncertainty of the impact parameter when considering the uncertainties of the track and vertex parameters. For the displacement in the z direction, it was found that comparing with the average vertex resolution gives better distinguishing power.

In addition, the value of $\text{IP}_{z,\text{cut}}$ will need to be adjusted depending on the displacement scale of interest, so it is considered part of the tagging algorithm. The cut values for track-level variables $\text{IP}_{z,\text{cut}}$, D_N , and the jet-level variables $\langle \text{IP}_{2\text{D}} \rangle$, α_{3D} will be determined via the method described in Section 4.3. An example of the variable distribution for SM jets and various signal model jets used for this class of EMJ tagging can be found in Fig. 4.4.

4.2.1.2 Variables used for tagging EMJs in the flavor-aligned signal models

For the flavor-aligned models, because the dark mesons decay to heavy-flavored SM quarks, the shower generated by the dark quarks is typically wider than the unflavored jets, as shown in Fig. 4.5. Because of this, tracks associated with a wider association radius with $R_{\text{cut}} = 0.8$ will be used to tag EMJs of this type.

Because the particles generated by the dark meson in flavor-aligned signal models are also a mixture of prompt and long-lived displaced particles, the variables used for the unflavored signal models were found to be ineffective in isolating the signal jets. A new set of dedicated variables

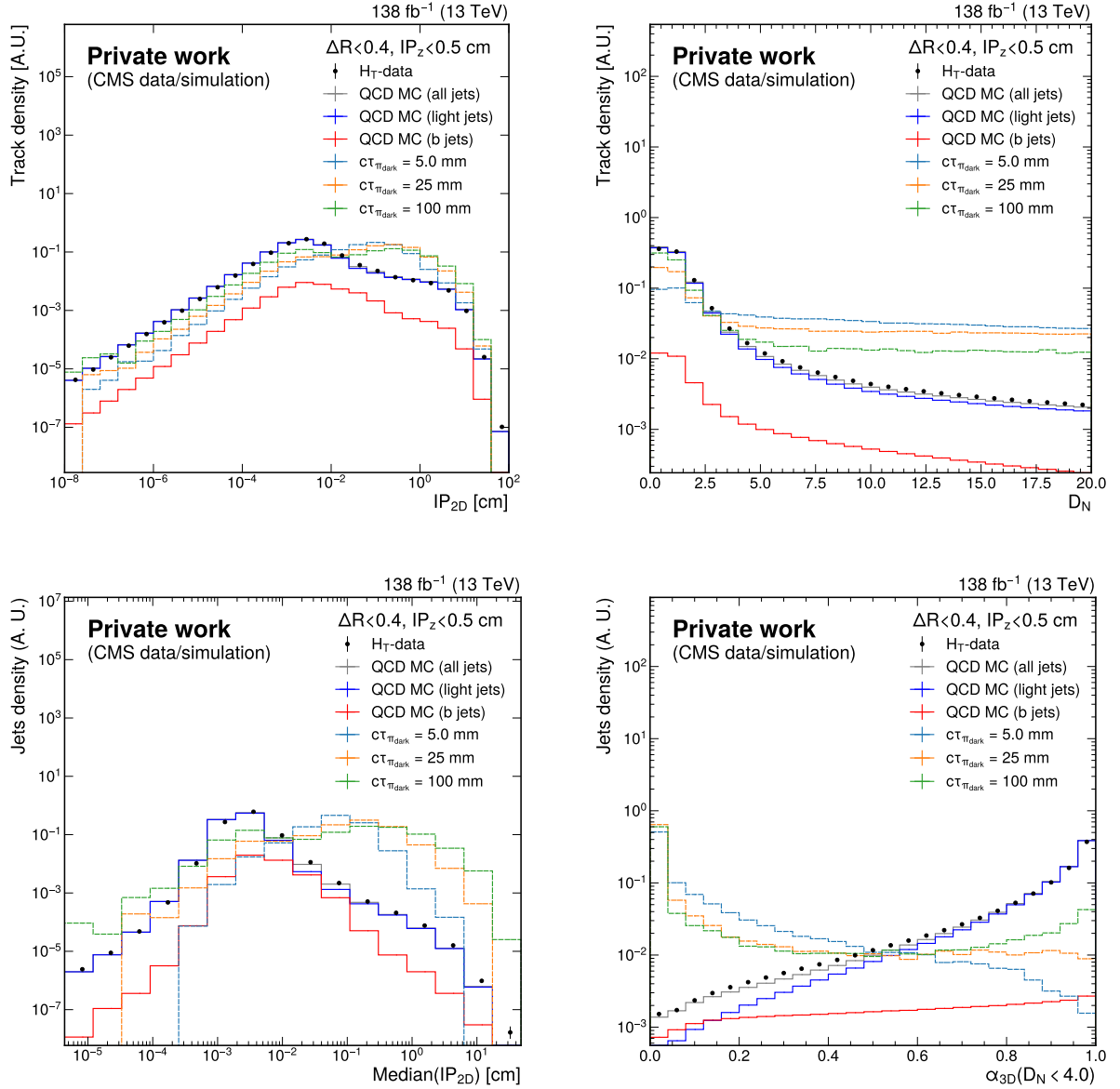


Figure 4.4: The distribution of $|IP_{2D}|$ of jet-associated tracks (top left), D_N of jet-associated tracks (top right), $\langle IP_{2D} \rangle$ of tracks associated with jets (bottom left), and α_{3D} of jets (bottom right), for SM light-flavored jets (blue), SM b jets (red), signal jets from various unflavored signal models (colored-dashed), and H_T -triggered data (black points). The plots are made with events that pass common event selection described in Section 4.1, and all distributions are normalized to unity to avoid prematurely guessing at the existence of excesses when looking at the distribution of data.

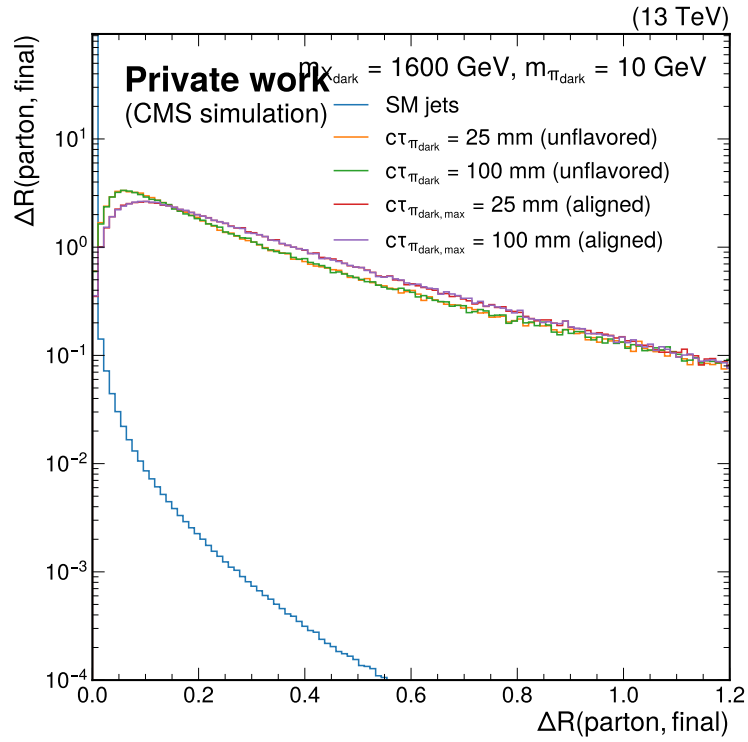


Figure 4.5: The ΔR separation between the final-state particles that and a hard-scattering parton for final-state particles produced in SM particle showers (blue), the dark shower of unflavored models (orange and green), and the dark shower from the flavored-aligned models (red and purple). Estimated using MC events.

will be used for tagging EMJs in the flavor-aligned models:

- The number of tracks with transverse displacements greater than a cutoff, $n_{\text{track}}^{|\text{IP}_{2\text{D}}| > \text{IP}_{2\text{D},\text{cut}}}$.

While the dark meson may decay promptly or be displaced, the promptly decaying dark mesons will also generate heavy-flavor SM mesons. An effective measure for tagging this mixed signature would be to tally the number of tracks with displacements on the order of b meson decay displacements ($|\text{IP}_{2\text{D}}| \sim 10^{-2.4} - 10^{-2.2} \text{ cm}$).

- The p_{T} -weighted ΔR average of associated tracks, g , otherwise named “track girth”, defined as:

$$g = \frac{\sum_i p_{\text{T},i} \cdot \Delta R_{i,\text{jet}}}{\sum_i p_{\text{T},i}}, \quad (4.4)$$

with the index i running over tracks of associated tracks. This variable characterizes the wider dark quark shower expected from the dark mesons containing heavy-flavored mesons in the final decay product.

The EMJ tagger will consist of track-level variables $\text{IP}_{2\text{D},\text{cut}}$, $\text{IP}_{z,\text{cut}}$, and jet-level variables $n_{\text{track}}^{|\text{IP}_{2\text{D}}| > \text{IP}_{2\text{D},\text{cut}}}$, g . The determination of these variables will be discussed in Section 4.3. An example of the distribution of these variables for SM and signal jets is shown in Fig. 4.6.

An additional variable $\tau_{2/1}$ is used for pileup jet contamination rejection. Since the jet-track association uses an angular value R_{cut} larger than the angular scale used for jet clustering, constituents from neighboring jet energy centers can potentially influence the tagging results. Notice that the issue with pileup jets is not just that neighboring jets can be simultaneously tagged as EMJs from shared tracks, but also that jets on the opposite side of the event can be

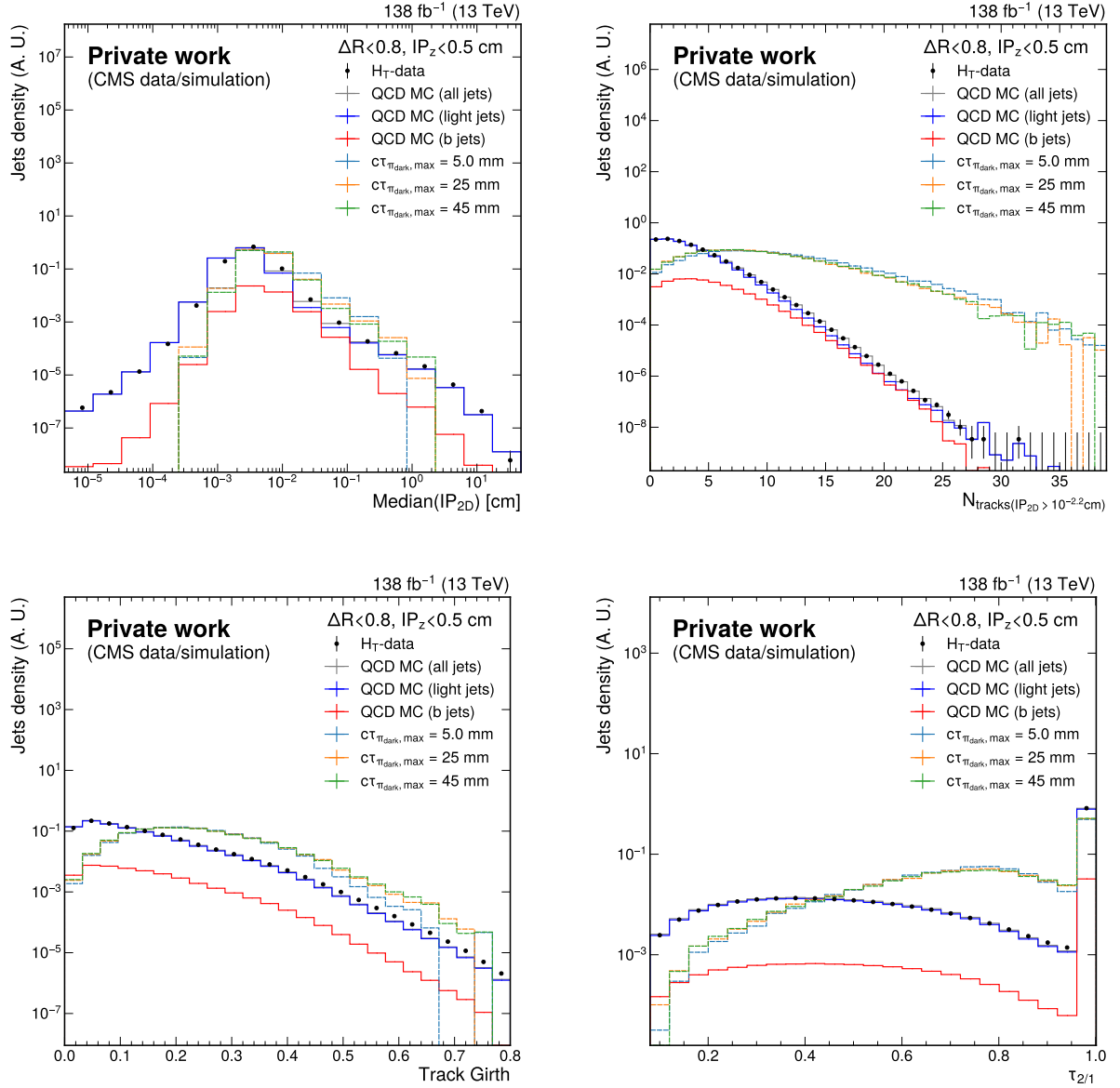


Figure 4.6: The distribution of $\langle IP_{2D} \rangle$ (top left), $n_{\text{track}}^{IP_{2D} > 10^{-2.2} \text{ cm}}$ (top right), track grith (bottom left), and $\tau_{2/1}$ (bottom right), for SM light-flavored jets (blue), SM b jets (red), signal jets from various flavored signal models (colored-dashed), and H_T -triggered data (black points). The plots are made with events that pass the common event selection described in Section 4.1. All distributions are normalized to unity to avoid prematurely guessing at the existence of excesses when looking at the distribution of data.

simultaneously tagged as EMJs, as pileup jets are most likely to be produced back-to-back. The variable $\tau_{2/1}$ is defined using track-based n-subjettiness measure, τ_n , based on the n-subjettiness measure [102] used by analyses in CMS:

1. For each of the selected jets, we additionally perform a “subjettiness”-association by taking all jets in the event, and matching each of them to selected jets within an angular radius of $\Delta R < 0.8$. Using this association, each jet will at least have one “subjettiness”, which is itself.
2. The value of τ_n for the selected jets is then calculated as:

$$\tau_n = \frac{\sum_i p_{T,i} \cdot \min_{j \leq n} \{\Delta R(i, j)\}}{\sum_i 0.8 p_{T,i}}, \quad (4.5)$$

where the summation runs over all tracks associated with the selected jets, and the index j is used to represent the subjettiness associated with the selected jets ordered by p_T . If the number of subjettiness associated with the selected jet is smaller than n , we take all subjettiness associated with the jets.

3. For pileup jet filtering, the variable of interest is $\tau_{2/1} = \tau_2/\tau_1$, a value which is expected to be close to unity for jets has a very small contribution from a secondary concentrated energy centers (such as from a pileup jet within the vicinity).

The determination of how $\tau_{2/1}$ will be used will also be discussed in Section 4.3.

4.2.2 Machine learning approach to EMJ tagging

Compared with manually defining how track-level variables should be reduced to some jet-level variables, ML methods can directly scan over the whole parameter space in search for the

“best performing” reduction. Instead of using a deep neural network, the ML technique employed as the standard b jet tagger used for the Run 2 data collected (Section 3.4.4), this analysis attempts to explore the signal discrimination power of a next-generation ML technique.

Deep neural networks (DNNs) have been demonstrated to be a robust classifier when given a dense matrix representing spatial data as an input, as can be seen in the popularity of using DNN in image recognition. A natural question would then be whether additional classification power can be extracted by providing a “jet image” to a DNN. While such spatial jet pattern information has served as the basis for jet-tagging, as shown in Fig. 4.7 and Ref. [102], the problem with naively extending this approach to ML-based methods is that jets are inherently spatially sparse objects consisting of a short list of constituent particles occupying a very small region in space. Direct digitization of jets into an image-like representation will result in lots of wasted memory, with most of the image storing null values (like the empty regions shown in Fig. 4.7). The excessive storing of null values will also result in a slow convergence time, as the ML training will also need to learn how to handle these null spaces. The development of graph neural networks (GNN) [103] is the inclusion of an edge convergence layer to smartly encode a sparse list of data points with coordinates values to be compatible with image-like representation for DNN, thus solving the memory and convergence issue. Studies for a GNN-based “ParticleNet” tagger for jet tagging [104] are currently under investigation by the CMS collaboration, and show promising performance improvements compared with the existing DNN-based DeepJet flavor tagger [105].

For the EMJ analysis, we attempt to set up a GNN to distinguish between the EMJs and SM jets using the tracks feature of the associated tracks. Track features that will be used as inputs to the GNN include:

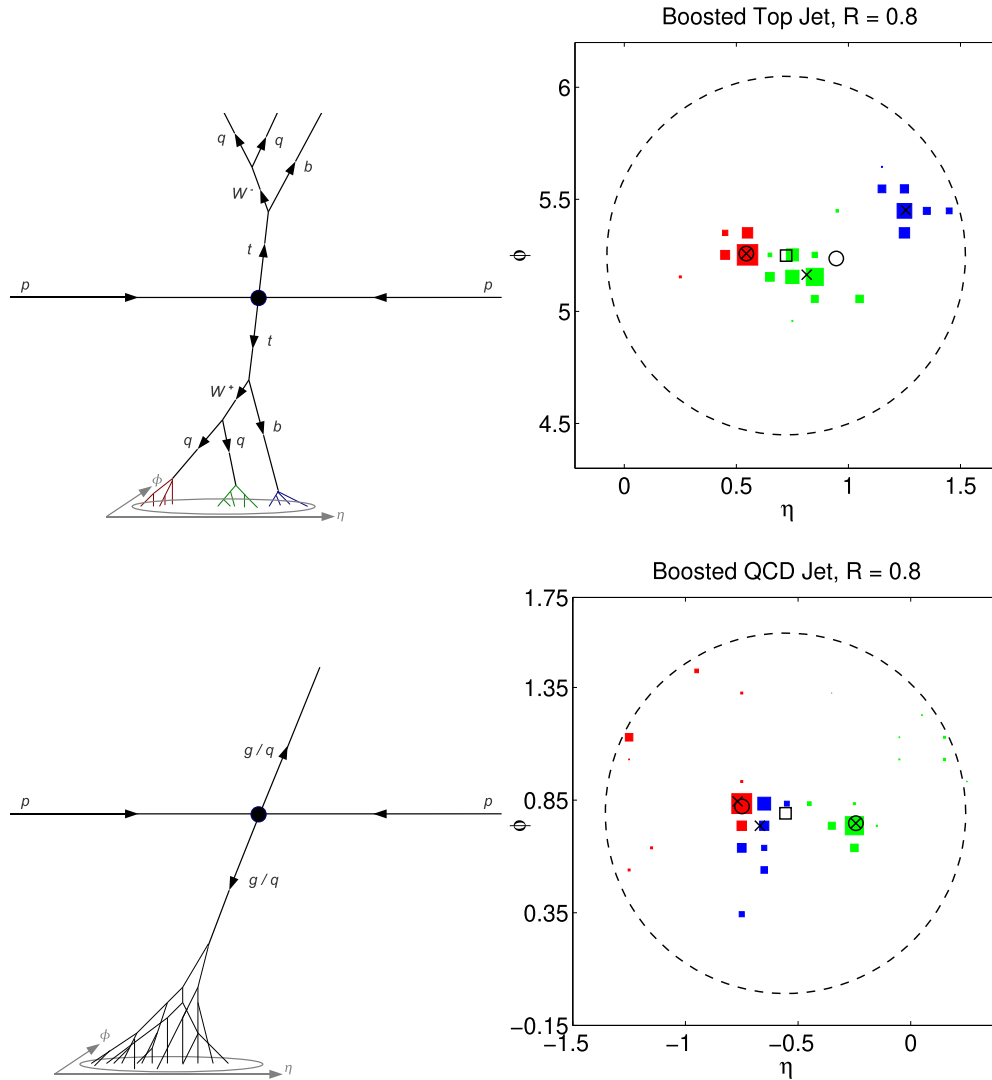


Figure 4.7: Example of how using jet images to observe jet substructure can be used to infer jet flavor information. The example uses jets generated in a $t\bar{t}$ production process (top), in which a jet generated from a top quark typically looks like a “3-prong” jet, while jets generated in QCD color interaction (bottom) typically looks structureless by comparison. These example figures are taken from Ref. [102].

- **Coordinate variables:** these will be used by the edge convergence layers to translate the discrete data points to the image-like representation. The variables $\Delta\eta$ and $\Delta\phi$ of the tracks relative to the jet energy center will be used as the coordinates variables for the GNN classifier.
- **Features variables:** these track variables will be the “pixel-values” that are used for the classification. 5 per-track features will be used for the GNN classifier:
 - ΔR of the tracks relative to the jet energy center.
 - $\ln(p_{T,\text{track}})$.
 - $\ln(p_{T,\text{track}} / (\sum_{\text{track}} p_{T,\text{track}}))$.
 - $T(\text{IP}_{2D})$ and $T(\text{IP}_z)$, where $T(x)$ is defined as:

$$T(x) = \text{sgn}(x) \cdot \ln \left(1 + \left| \frac{x}{x_0} \right| \right), \quad (4.6)$$

where x_0 is the unit measure of x used by the CMS collaboration. In the case of the distance measures for both IP_{2D} and IP_z , the value of x_0 is 1 cm.

The log functions for the latter variables help compress the unbounded variable that span many orders of magnitude into a smaller range to speed up training convergence.

While the variables ΔR appears redundant when $\Delta\eta$ and $\Delta\phi$ are used as the coordinates’ data, if it is already known that ΔR to be a useful feature for tagging EMJs (as seen in the model-agnostic EMJ tagging method), explicitly giving the GNN this variable can accelerate the training process. The momentum fraction is given along with the momentum for the same reason.

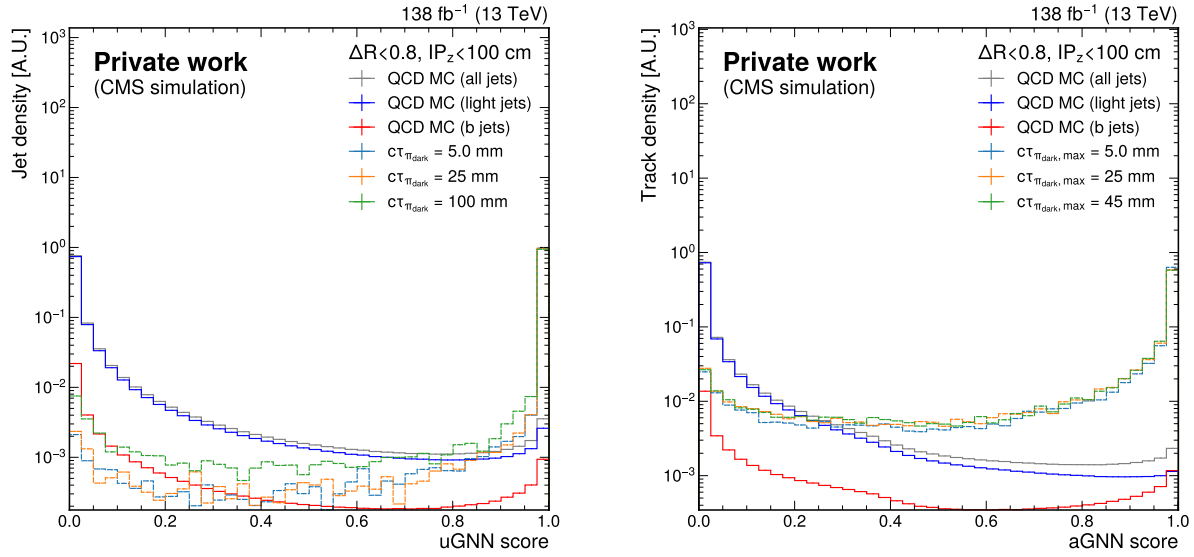


Figure 4.8: Example of the uGNN (aGNN) output score value on the left (right), for SM light jets (blue), SM b jets (red), and the various signal samples (colored dashed). Plots are made with events that pass common event selection described in Section 4.1. All distributions are normalized to unity. Data are omitted to avoid prematurely guessing at excesses in the distribution in the signal-rich regions of the distribution.

Two GNN classifiers are trained for this analysis, one for distinguishing between signal jets generated by the unflavored signal models from SM jets and the other for distinguishing between signal jets generated by the flavor-aligned signal models from SM jets. The two networks will be labeled as “uGNN” and “aGNN” respectively. During training, jets from different signal model MC events are randomized and trained against jets from QCD MC samples¹. An example of the final distribution of the GNN output score can be found in Fig. 4.8. EMJ tagging can then be done by requiring the GNN output score for the jet of interest to pass the threshold score. The determination of the thresholds will be discussed in Section 4.3.

¹As this is the work of another graduate student, details of how the GNN is designed, trained and validated will not be presented in this thesis.

4.3 Determining the cut values for signal extraction

To match the event topology described in Section 1.3, events are required to have four jets that pass the jet selection described in Section 4.1.3, with two of the selected jets passing a defined EMJ tagging criteria. The cut value used for the EMJ tagging parameters should be adjusted to give the best signal/background separation. We will vary the cut values used by the various object selections, count the number of signal and background events that are expected to pass the selection criteria using our existing MC samples, and select the set of cut values that give the maximum figure-of-merit $\mathcal{M}$:

$$\mathcal{M} = \frac{S}{\sqrt{S + B + 0.1B^2}}, \quad (4.7)$$

where S (B) is the number of signal (background) events that pass the defined selection values. The term $0.1B^2$ represents the rough estimate of the uncertainties expected in the background.

For the event selection, the following values we can use to achieve better signal/background separation:

- Event-level features: H_T , and additional p_T requirements of the 4 selected jets.
- The track and jet-level variables used to define EMJ tagging $\vec{v}_{\text{tag}}$:
 - For tagging jets generated from the unflavored model using the model-agnostic method, $\vec{v}_{\text{tag}}$ includes track-level variables $\text{IP}_{z,\text{cut}}$, D_N , and jet-level variables α_{3D} and $\langle \text{IP}_{2D} \rangle$.
 - For tagging jets generated from the flavor-aligned models using the model-agnostic method, $\vec{v}_{\text{tag}}$ includes track-level variables $\text{IP}_{z,\text{cut}}$, $\text{IP}_{2D,\text{cut}}$, and jet-level variables

Table 4.1: Cut values of jet and track variables used for tagging EMJs of the unflavored model with jet-track association radius $R_{\text{cut}} = 0.4$.

tag name	$ \text{IP}_z $ [cm] ($<$)	$\langle \text{IP}_{2D} \rangle$ [cm] ($>$)	D_N ($<$)	α_{3D} ($<$)
u-tag 1	0.5	$10^{-1.6}$	4	0.25
u-tag 2	1.0	$10^{-1.4}$	8	0.25
u-tag 3	5.0	$10^{-1.2}$	8	0.25
u-tag 4	5.0	$10^{-1.2}$	12	0.15
u-tag 5	5.0	$10^{-1.0}$	12	0.15
validation u-tag	0.5	$10^{-1.6}$	4	0.40

$$n_{\text{track}}^{|\text{IP}_{2D}| > \text{IP}_{2D, \text{cut}}}, g.$$

- For tagging jets using the ML-based method, $\vec{v}_{\text{tag}}$ is simply the output score of the GNN of interest.

A scan is performed across the various cut values in a discrete grid, and $\mathcal{M}$ is calculated for all surveyed signal samples across the scanning grid. For each signal sample, the set of cut values that yields the maximum $\mathcal{M}$ value is first determined and then grouped if alternate cut values from neighboring samples in the signal parameter space do not change the value of $\mathcal{M}$ by more than 10% from the evaluated maximum value. The cut values for the model-agnostic EMJ tagging method for the unflavored (flavor-aligned) scenario can be found in Table 4.1 (Table 4.2), the GNN output score cuts used for EMJ tagging are listed in Table 4.3, while the event-level variable selection used to accompany the EMJ tagger is listed in Table 4.4. The purpose of the additional “validation” tags and event selection cut sets will be discussed in Section 5.4.2.

One final selection that is not directly determined by the significance figure-of-merit is the $\tau_{2/1}$ used for pileup jet rejections, as the maximization of $\mathcal{M}$ does not consider correlated EMJ tagging from pileup jets. For determining a cut value for the $\tau_{2/1}$ value, the QCD MC events that pass the selection defined by the a-set series while having two EMJ-tagged jets appear back-to-

Table 4.2: Cut values of jet and track variables used for tagging EMJs of the flavor-aligned model with jet-track association radius $R_{\text{cut}} = 0.8$.

tag name	$ \text{IP}_z $ [cm] ($<$)	$\text{IP}_{2\text{D},\text{cut}}$ [cm]	$n_{\text{track}}^{ \text{IP}_{2\text{D}} > \text{IP}_{2\text{D},\text{cut}}} (>)$	girth ($>$)	$\tau_{2/1} (>)$
a-tag 1	0.5	$10^{-2.2}$	12	0.05	0.5
a-tag 2	0.5	$10^{-2.2}$	12	0.1	0.5
a-tag 3	0.5	$10^{-2.2}$	12	0.1	0.5
a-tag 4	0.5	$10^{-2.3}$	14	0.0	0.5
a-tag 5	0.5	$10^{-2.4}$	14	0.1	0.5
validation a-tag	0.5	$10^{-2.4}$	12	0.0	0.5

Table 4.3: Cut values using the output score of the GNNs used tagging EMJs in the unflavored and flavor-aligned signal models

tag name	Score range
uGNN tag 1	0.9997–1
uGNN tag 2	0.9998–1
uGNN tag 3	0.9996–1
uGNN validation tag	0.9980–0.9995
aGNN tag 1	0.9953–1
aGNN tag 2	0.9993–1
aGNN tag 3	0.9983–1
aGNN validation tag	0.9900–0.9950

Table 4.4: Event selection criteria used for signal detection in this analysis. All events are required to have at least two jets passing the EMJ tagging criteria assigned to the cut set of interest.

cut set name	H_T [GeV]	jet p_T [GeV] (>)				EMJ tagger
u-set 1	> 1600	275	250	250	150	u-tag 1
u-set 2	> 1600	200	200	150	150	u-tag 2
u-set 3	> 1600	200	150	100	100	u-tag 3
u-set 4	> 1500	200	150	100	100	u-tag 4
u-set 5	> 1200	200	150	100	100	u-tag 5
u-set validation	1000–1200	100	100	100	100	validation u-tag
a-set 1	> 1500	200	150	100	100	a-tag 1
a-set 2	> 1800	250	250	200	200	a-tag 2
a-set 3	> 1200	275	250	250	200	a-tag 3
a-set 4	> 1500	275	250	250	100	a-tag 4
a-set 5	> 1800	200	150	100	100	a-tag 5
a-set validation	1000–1200	100	100	100	100	validation a-tag
uGNN set 1	> 1350	170	120	120	100	uGNN tag 1
uGNN set 2	> 1750	300	260	250	250	uGNN tag 2
uGNN set 3	> 1800	240	180	180	100	uGNN tag 3
uGNN validation	> 1000	100	100	100	100	uGNN validation tag
aGNN set 1	> 1300	200	140	120	100	aGNN tag 1
aGNN set 2	> 1650	300	250	200	200	aGNN tag 2
aGNN set 3	> 1400	270	220	220	120	aGNN tag 3
aGNN validation	> 1000	100	100	100	100	aGNN validation tag

back ($|\Delta\phi| > 3.0$) are used. The maximum $\tau_{2/1}$ of these EMJ-tagged jets is taken and rounded up to the closest 0.1 value to arrive at the cut value of $\tau_{2/1}$ to be 0.5. An example of the final signal selection efficiency of the various signal models is shown in Fig. 4.11–4.12.

While comparing the signal selection efficiency, we typically see that the ML-based tagging method gives a higher signal selection efficiency when compared with the model-agnostic method. The details of how this result should be interpreted will be given in Chapter 7.

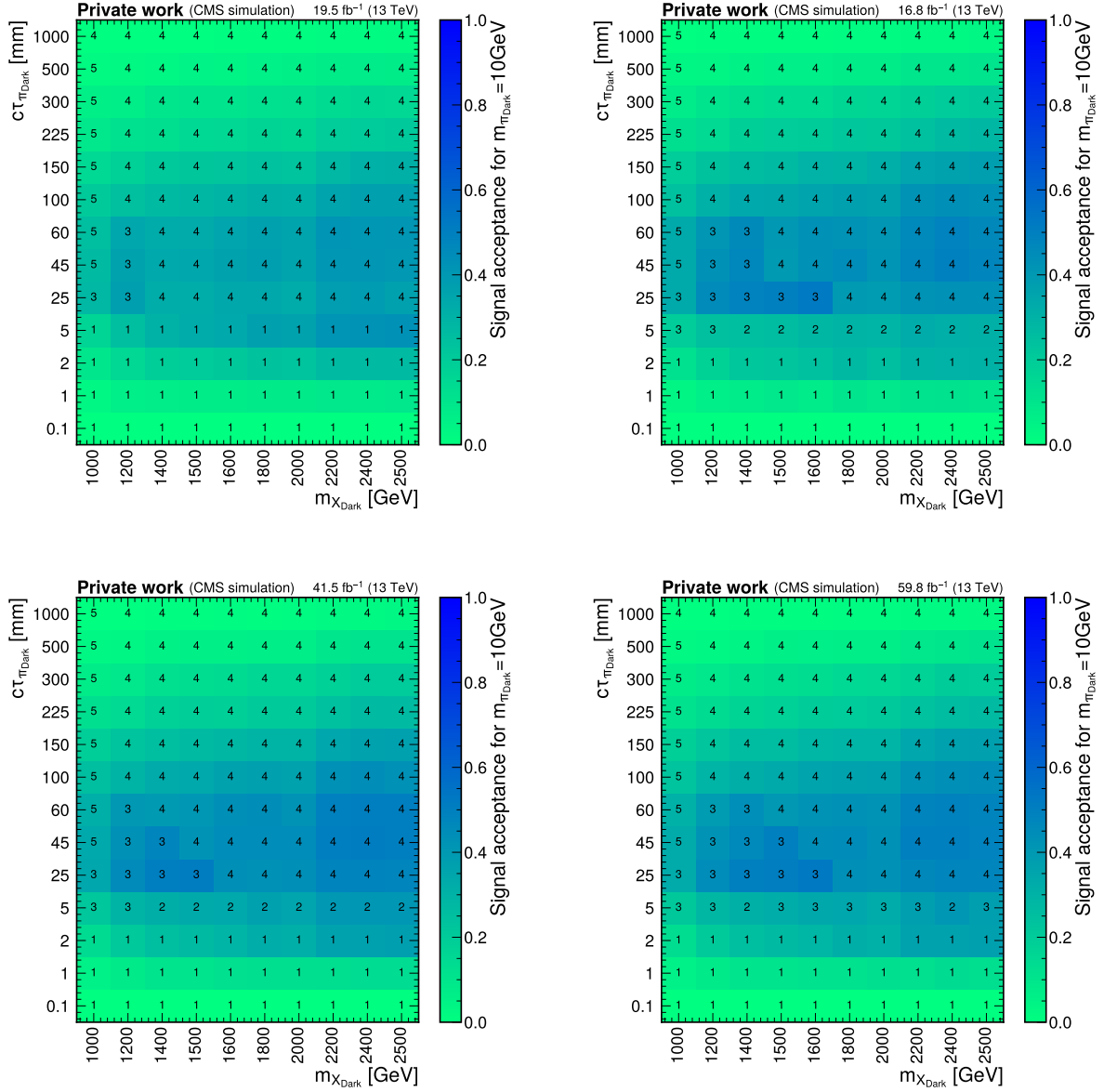


Figure 4.9: Signal selection efficiency using the model-agnostic tagging methods described in Section 4.2.1 for various unflavored signal models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$ as a function of $m_{X_{\text{dark}}}$ and $c\tau_{\pi_{\text{dark}}}$, evaluated for the 2016 pre-APV fix (top left), 2016 post-APV fix (top right), 2017 (bottom left), 2018 (bottom right) detector condition. The index in each of the cells indicates the cut set used for the signal sample of interest.

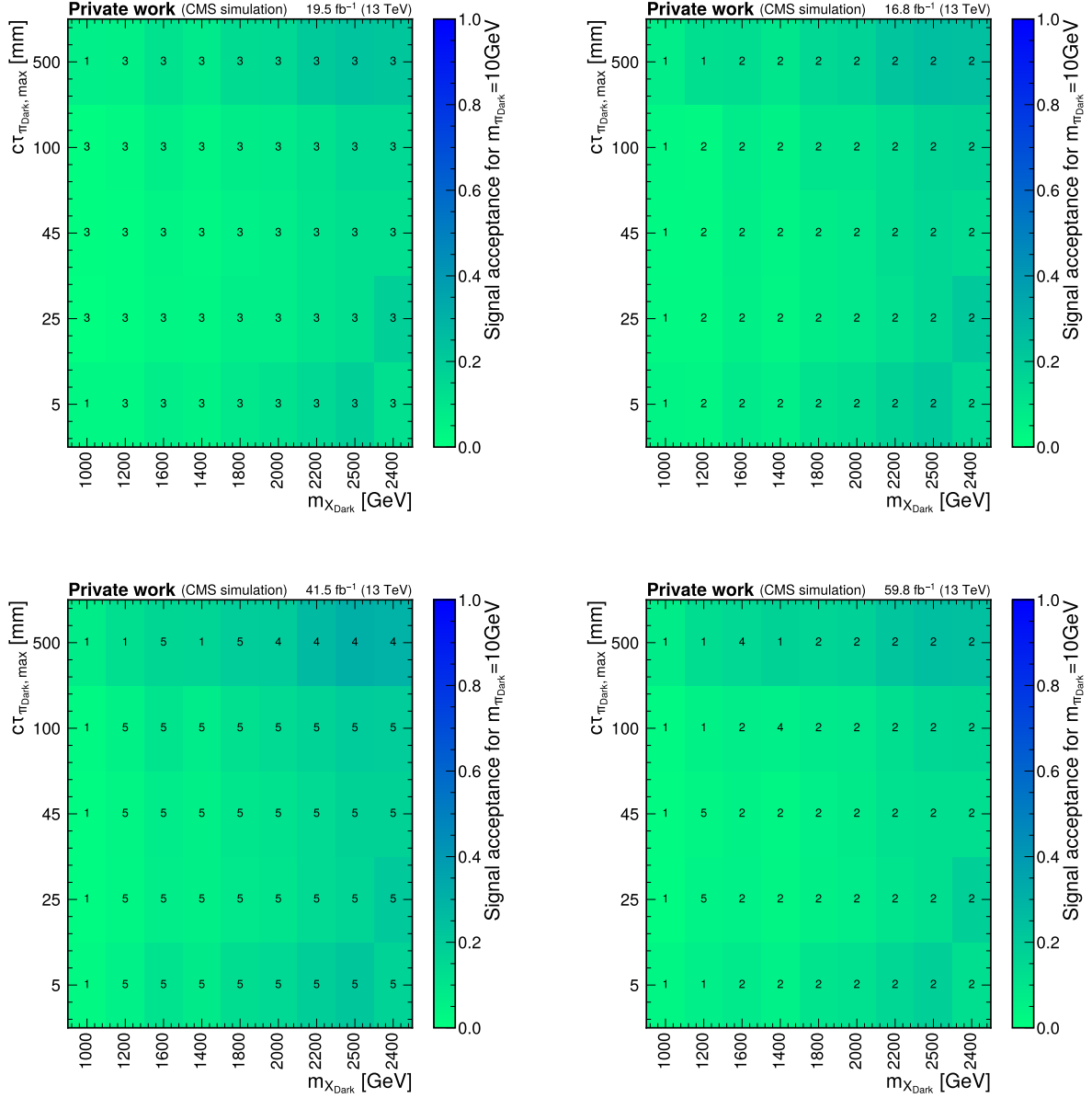


Figure 4.10: Signal selection efficiency using the model-agnostic tagging methods described in Section 4.2.1 for various flavor-aligned signal models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$ as a function of $m_{X_{\text{dark}}}$ and $c\tau_{\pi_{\text{dark}}, \text{max}}$, evaluated for the 2016 pre-APV fix (top left), 2016 post-APV fix (top right), 2017 (bottom left), 2018 (bottom right) detector condition. The index in each of the cells indicates the cut set used for the signal sample of interest.

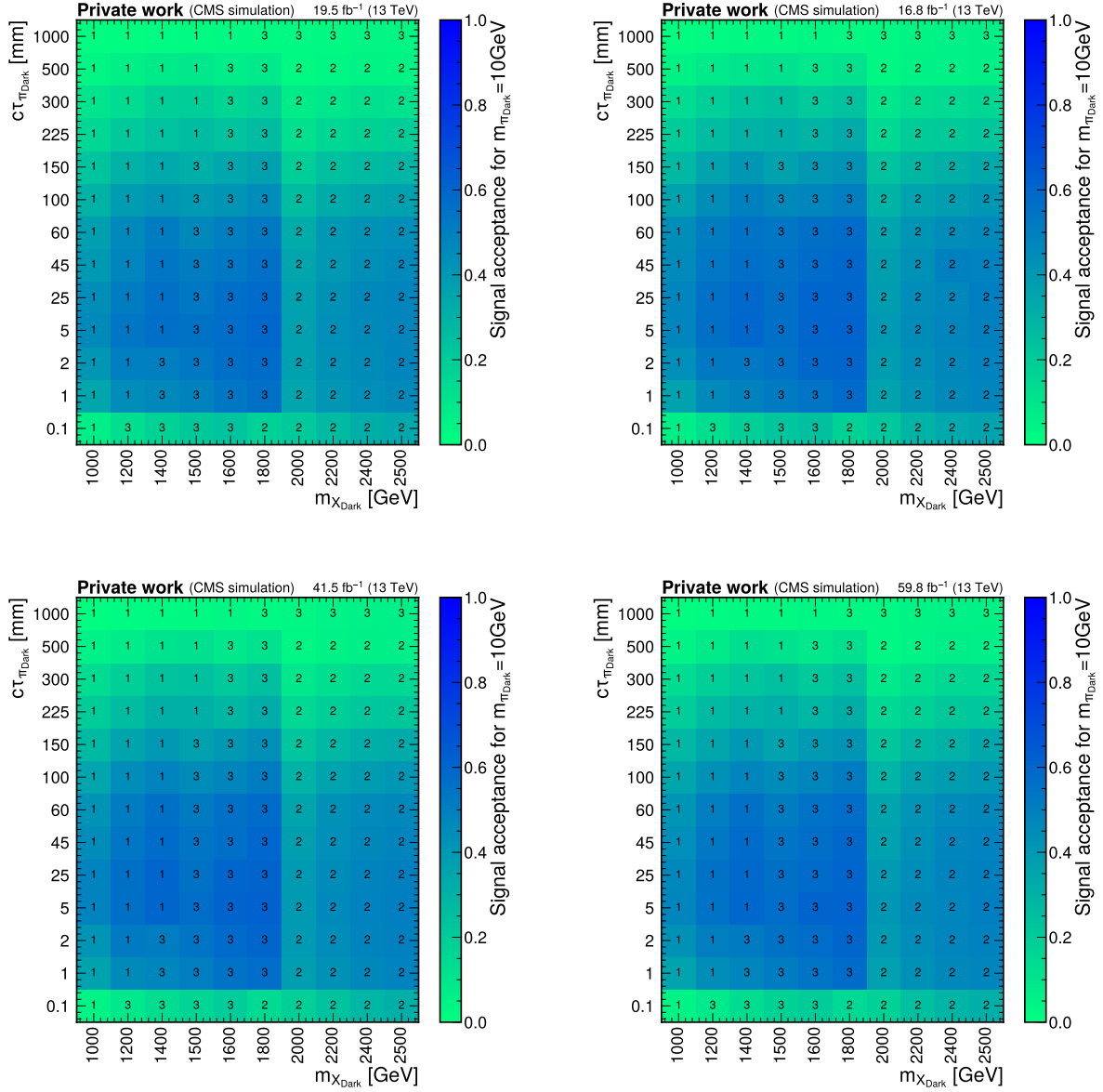


Figure 4.11: Signal selection efficiency using the ML tagging described in Section 4.2.2 for various unflavored signal models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$ as a function of $m_{X_{\text{dark}}}$ and $c\tau_{\pi_{\text{dark}}}$, evaluated for the 2016 pre-APV fix (top left), 2016 post-APV fix (top right), 2017 (bottom left), 2018 (bottom right) detector condition. The index in each of the cells indicate the cut set used for the signal sample of interest.

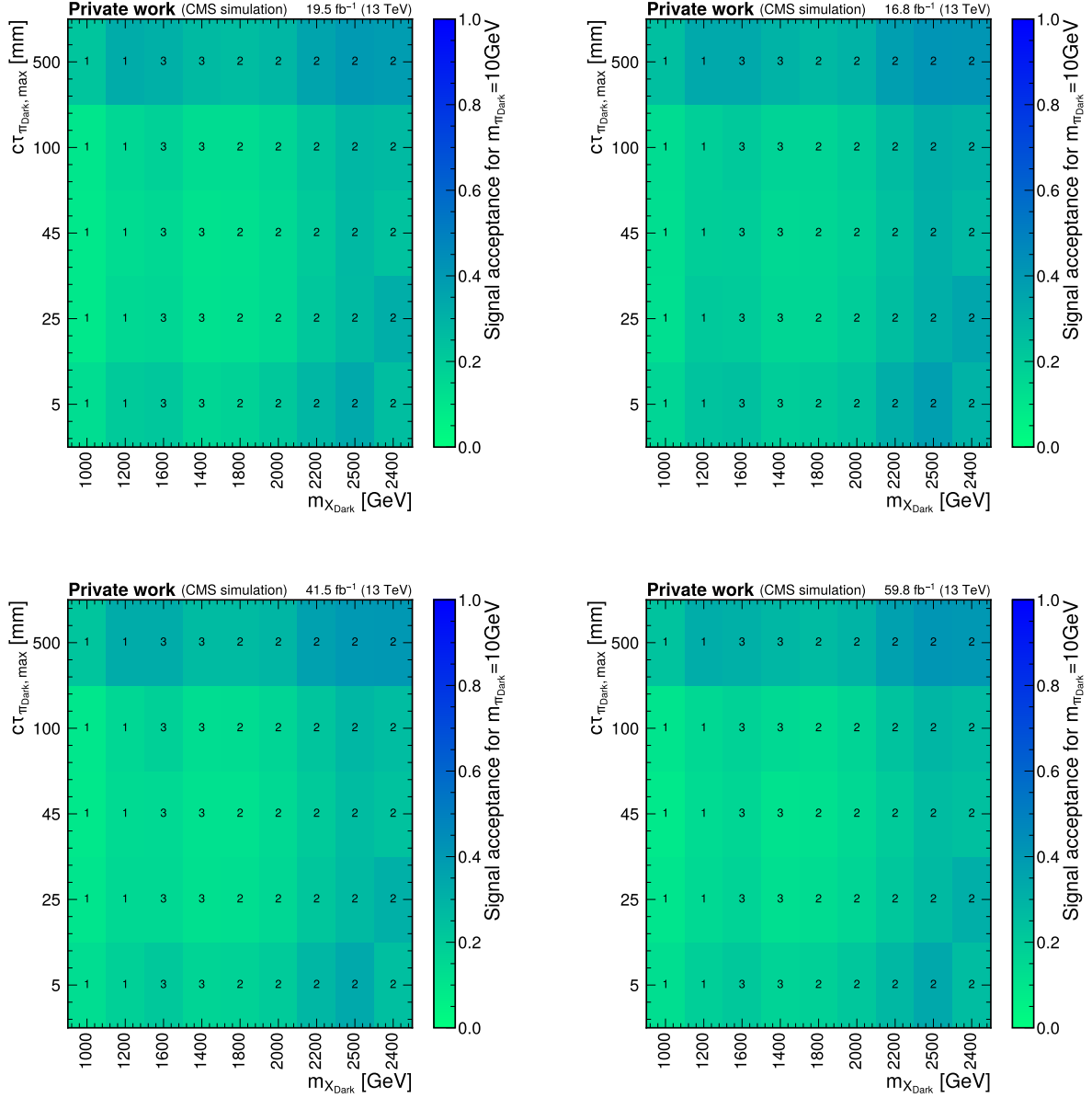


Figure 4.12: Signal selection efficiency using the ML tagging described in Section 4.2.2 for various flavor-aligned signal models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$ as a function of $m_{X_{\text{dark}}}$ and $c\tau_{\pi_{\text{dark}}, \text{max}}$, evaluated for the 2016 pre-APV fix (top left), 2016 post-APV fix (top right), 2017 (bottom left), 2018 (bottom right) detector condition. The index in each of the cells indicates the cut set used for the signal sample of interest.

4.4 Sources of displaced tracks in standard model processes

Standard model processes can also generate jets containing displaced tracks, and it is essential to understand how such tracks are generated so that they can accurately be accounted for when estimating the number of SM background events that enter our signal region. In this section, we will explore the possible sources of displaced tracks, while details of how these tracks will impact the method by which we evaluate the number of background events will be fully discussed in Chapter 5.

4.4.1 Long-lived standard model mesons

As discussed in Section 1.1.1 and 3.4.4, heavy-flavor SM mesons also generate the displaced vertices and tracks, resulting in the SM jets containing heavy-flavored mesons being disproportionately likely to be tagged as EMJs. This effect can also be shown in the distributions of the various tagging variables throughout this chapter. As the generation of quark flavors is likely to be correlated from the underlying hard scattering processes, standard model or not, we will attempt to evaluate the EMJ mistagging rate as a function of the underlying jet flavor. Details on how this will be evaluated will be given in Section 5.1.2.

4.4.2 Pileup events

Particles emerging from pileup vertices will naturally have a large displacement measure compared to the primary collision vertex. Because of this, the correct vertex must be chosen as the primary vertex when computing track displacement variables, while also rejecting pileup jets as being tagged as emerging jets. The attempt to reject pileup is the motivation behind the

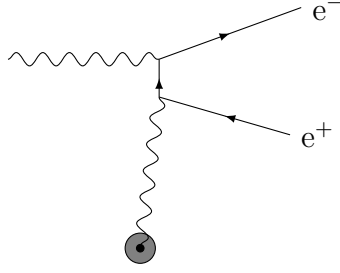


Figure 4.13: Feynman diagram describing the photon conversion process in the presence of a background field source (gray circle).

additional vertex quality cut selections discussed in Section 4.1.2, as well as the various track-level and jet-level selections discussed in Sections 4.2 and 4.3.

4.4.3 Material interaction

Whenever high-energy particles pass through the detector material, there is a finite chance that electron-positron pairs are generated from the primary high-energy particles through electromagnetic interactions. The most common process would be photon conversions (depicted in Fig. 4.13), where photons with energy larger than $2m_e$ can convert to e^+e^- pairs in the large electromagnetic field in the active volume of detector systems.

Tracks generated from material interaction display striking structures in their displacement and p_T distribution, as shown in Fig. 4.14. These structures can be shown to be directly tied to the innermost detector geometry of the CMS detector. Let us consider the case where a high-energy photon emerges from the center of the detector and travels to a detector element at radial distance L , where the photon converts to a pair of electrons. At the point of conversion, the electron-positron pair would have the highest probability of having momentum in parallel with the incident photon. For this configuration (depicted in Fig. 4.15), the track trajectory can be extrapolated back to the point of closest approach to the center of the detector. Within the approximation that

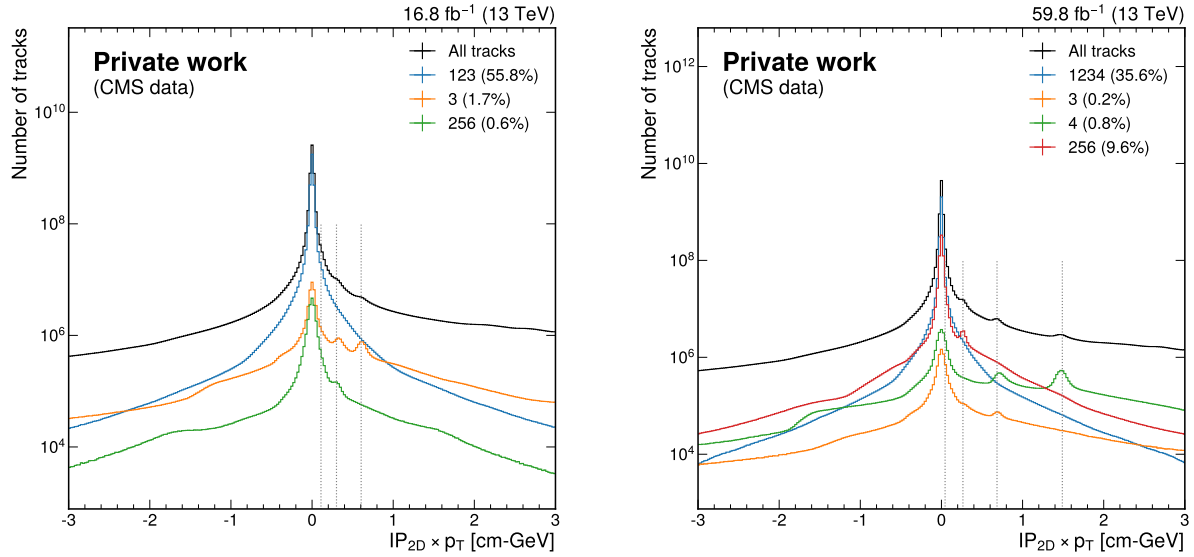


Figure 4.14: Distribution of the product of track IP_{2D} and track p_T of tracks collected in 2016 (left) and 2018 (right) of tracks that left hits in all pixel barrel layers (labeled “123” for 2016, and “1234” for 2018), and some subset of pixel barrel layers. The percentage signs behind the labels indicate the fraction of a particular hit pattern of all tracks reconstructed events.

the trajectory curvature R on the transverse plane is much larger than L , we get the relation:

$$IP_{2D} = \sqrt{R^2 - L^2} - R \xrightarrow{R \gg L} \frac{L^2}{2R}, \quad (4.8)$$

$$IP_{2D} \cdot p_T \sim \frac{L^2 q B}{2}. \quad (4.9)$$

Here we are using the helix path equation $p_T = qBR$, where p_T is the transverse momentum of the charged particle, q is the particle charge and B is the magnetic field strength. Inserting into the equation L with the values of the barrel pixel detectors radii of the pixel Phase-0 (1) geometry as with $L = 10.2, 7.3, 4.5$ cm (16.0, 10.85, 6.75, 2.80 cm) as described in Section 2.2.3.1, the CMS magnetic field strength of 3.8 T, we can estimate that the $IP_{2D} \cdot p_T$ value of tracks that originate from material interactions to be quantized to values at approximately 0.605, 0.310, 0.11 (1.49, 0.685, 0.265, 0.045 cm-GeV).

This quantized structure is observable in the reconstructed tracks in both data and back-

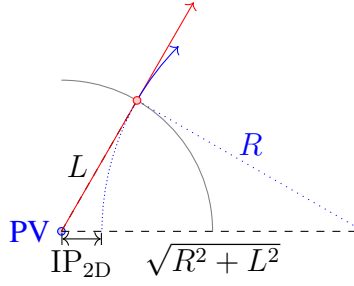


Figure 4.15: Diagram describing the geometric relation of low- p_T tracks originating from material interaction. The high- p_T photon (red) interacts with detector material at radius L from the primary vertex (PV), generating a low- p_T charged particle (blue) with curvature radius R . The path can be extrapolated back to the point of closest approach to the primary vertex (dotted blue) to calculate the IP_{2D} of the track in question.

ground MC samples. These peaks are small when looking at the inclusive distribution of all tracks, but can be enhanced by looking at tracks with specific pixel hit patterns: tracks generated from material interactions with the 3rd pixel barrel layer would be expected to only have pixel hits in at most 3rd and 4th barrel pixel layers. Examples of the predicted structure can be found in Fig. 4.14.

By plotting the transverse position of all displaced vertices reconstructed within CMS (Fig. 4.16), the vertex position shadows the material in the inner tracker volume, indicating that material interaction has a nontrivial contribution to displaced vertices outside the beam line. Also, in simulated events where the complete information of where and how charged particles are generated is available, such peaking structures can be further isolated to charged particles generated during detector interaction simulation.

While this source of displaced tracks exhibits structures, it is also random: having some material interaction within a jet cone neither increases nor decreases the changes of material interactions occurring in another jet cone. Because of this randomness, we can ensure that the method of estimating background contributions discussed in Chapter 5 is safe against these sources of

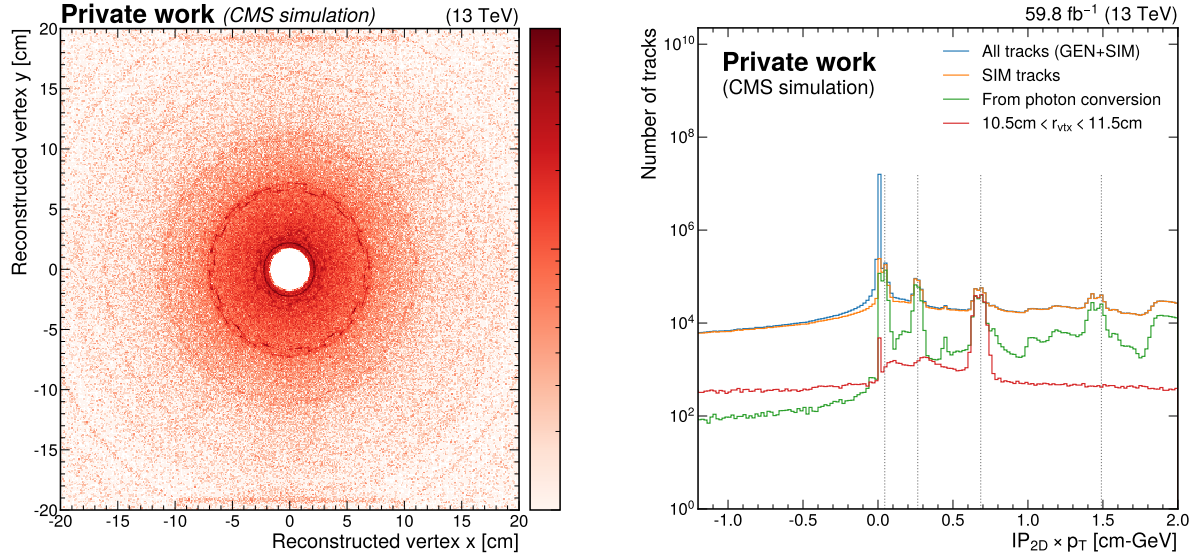


Figure 4.16: Left plot displays the position of all reconstructed vertices with position $r > 2$ cm, using the QCD MC events corresponding with the detector configuration of 2018. Right plot displays the distribution of $IP_{2D} \cdot p_T$ of at various stages of the generation and simulation process using tracks in QCD MC events.

displaced tracks. As a sanity check to ensure that our analysis sensitivity is not limited by material interaction, we can plot the track $IP_{2D} \cdot p_T$ distribution of tracks that pass the signal selection criteria defined in Table 4.4. If we look at the inclusive distribution of tracks in events in the QCD MC sample that pass signal selection, we do not see an apparent excess in the material structure values, which indicates that the current analysis is not limited in sensitivity by material interaction.

Chapter 5: Data-based background estimation

Regardless of the EMJ tagging method, jets originating from SM processes will have a finite chance of being “mistagged” as emerging jets (EMJ) due to the various sources described in Section 4.4. Using jets in QCD MC events, we can estimate the mistag rate of the different EMJ tagging algorithms defined in Section 4.2, as well as demonstrate the dependence of mistag rates on jet variables. Examples of the SM jet mistag rate can be found in Fig. B.1–5.4.

Most of the sources of displaced tracks in SM processes that may lead to SM jets being mistagged as EMJs, as discussed in Section 4.4, are effects of randomness in the reconstruction environment, a feature of data that is very difficult to represent correctly in simulation. Figures B.1–5.4 also demonstrate that the mistag rates strongly depend on jet variables not necessarily well-modeled in simulation. In particular, the mistag rate strongly depends on the track multiplicity, which, in turn, strongly depends on the modeling of nonperturbative QCD part of shower generation. For these reasons, relying solely on the correctness of simulated events is not desirable. Instead, the mistag rates of EMJ taggers are derived using CMS data in a signal-free region, which are then used to estimate a per-event scale factor that allows us to obtain the number of SM events that pass the signal region selection from a signal-poor control region. Section 5.1 will present how the mistag rates can be calculated in data. Section 5.2 describes performing the appropriate signal region/control region scaling. The validation of the calculation routines

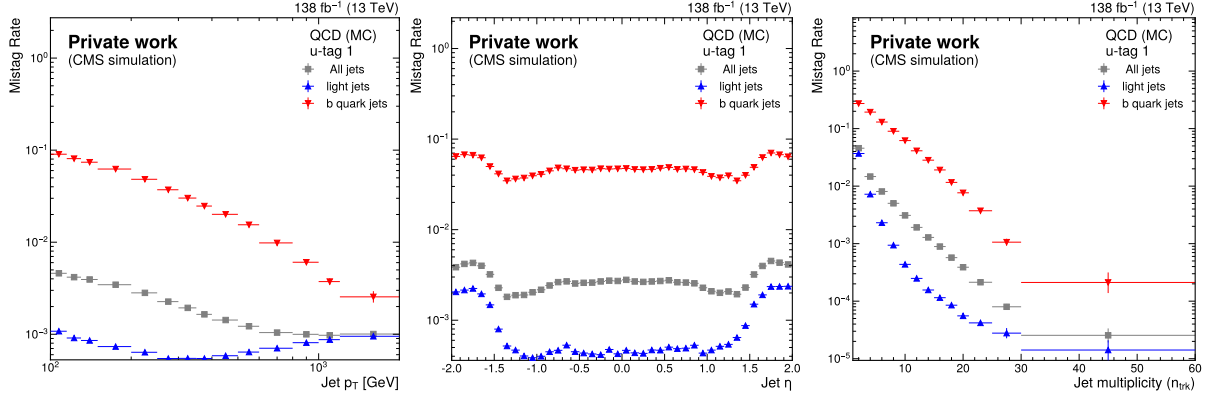


Figure 5.1: Example of the mistag rate of the “u-tag 1” tagger evaluated using jets in the QCD MC samples as a function of jet p_T (left), jet η (middle), and the number of associated tracks (right). The mistag rates of b jets are shown in red, non-b jets (light jets) in blue, while the combined mistag rate of all jet flavors is shown in gray.

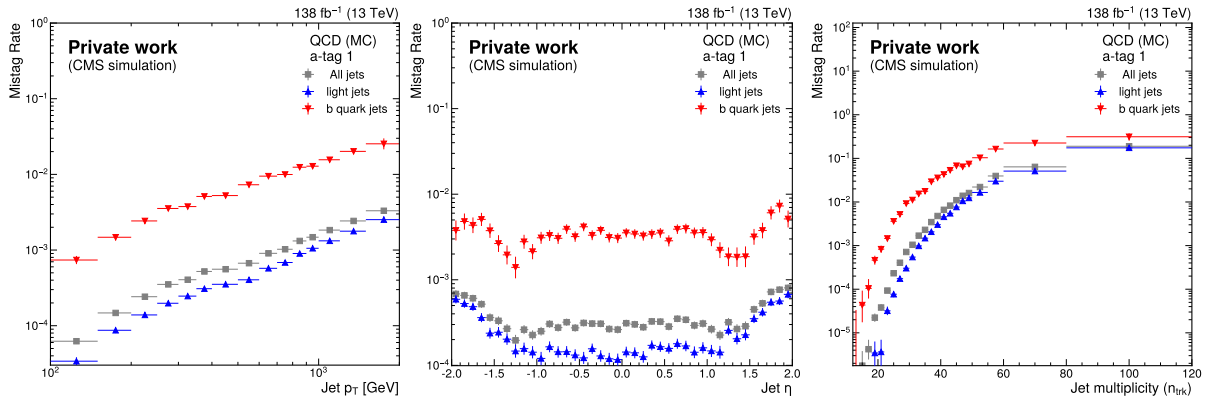


Figure 5.2: Example of the mistag rate of the “a-tag 1” tagger evaluated using jets in the QCD MC samples as a function of jet p_T (left), jet η (middle), and the number of associated tracks (right). The mistag rates of b jets are shown in red, non-b jets (light jets) in blue, while the combined mistag rate of all jet flavors is shown in gray.

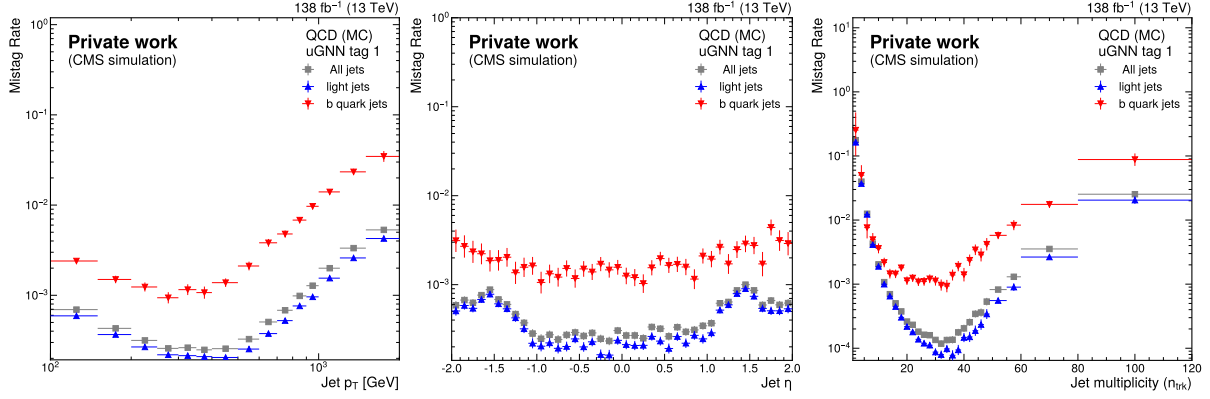


Figure 5.3: Example of the mistag rate of the “u-GNN tag 1” tagger evaluated using jets in the QCD MC samples as a function of jet p_T (left), jet η (middle), and the number of associated tracks (right). The mistag rates of b jets are shown in red, non-b jets (light jets) in blue, while the combined mistag rate of all jet flavors is shown in gray.

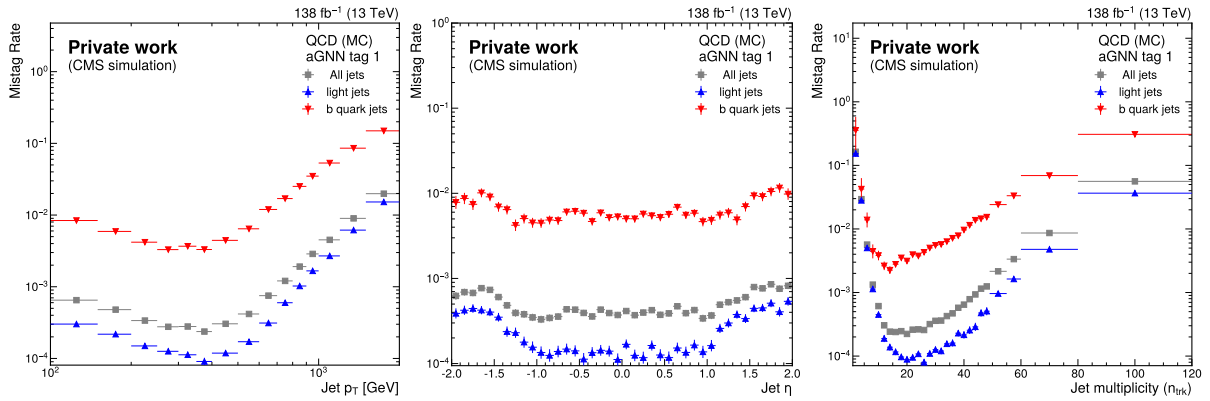


Figure 5.4: Example of the mistag rate of the “a-GNN tag 1” tagger evaluated using jets in the QCD MC samples as a function of jet p_T (left), jet η (middle), and the number of associated tracks (right). The mistag rates of b jets are shown in red, non-b jets (light jets) in blue, while the combined mistag rate of all jet flavors is shown in gray.

discussed will be presented in Section 5.4.

5.1 Data-based mistag rate evaluation

As stated at the start of the chapter, we want to evaluate the SM jet mistag rate $\epsilon(\vec{\theta})$ using data collected by CMS in a signal-free region (region F). Since signal samples are generated through QCD-color interactions, the number of signal events can be significantly suppressed by requiring a high- p_T color singlet to exist in the final state. Here, we construct a signal-free region F using the γ -triggered data stream acquired by CMS with a high- p_T photon. The formal construction of region F will be given in Section 5.1.1. Also, the flavor dependence of the mistag rate of the primary jets of interest can be derived without explicitly requiring flavor information by imposing b tagging information of an additional secondary jet within the same event. How this calculation is performed is given in Section 5.1.2.

5.1.1 Definition of the signal-free regions

Events of the primary signal-free region are required to pass the lowest p_T , unrescaled photon trigger available for the various data collection eras. For 2016, we use the trigger requiring a photon with $p_T > 165$ GeV in the central region of the ECAL detector (labeled `HLT_Photon165_HE10_v*`); for 2017 and 2018, we use the trigger requiring a photon with $p_T > 200$ GeV in the central region of the ECAL detector (labeled `HLT_Photon200_v*`). In the full offline reconstruction, the events are also required to have at least one well-reconstructed high- p_T photon in the event. The high- p_T photon is required to pass the following selection criteria:

- Have reconstructed $p_T > 200$ GeV and $|\eta| < 2.1$. These kinematic thresholds are chosen

to match the online trigger p_T requirement, to minimize effects of trigger turn-on. The p_T threshold is chosen to match the largest online p_T threshold, to minimize the bias in the evaluated mistag when comparing the results of data collected for different years.

- Have $|\eta|$ outside the ECAL barrel–endcap transition region $1.4442 < |\eta| < 1.566$.
- Does not have an associated pixel seed.
- Pass the cut-based photon isolation required at the medium working point, as described in Section 3.4.2.

No signal events are expected to pass these photon selection criteria, with an estimated signal acceptance upper bound of 10^{-6} . As jets in these events will have their associated tracks’ displacement measures passed to the tagging algorithms for mistag rate evaluation, the primary vertex requirements defined in Section 4.4.2 are also imposed. Events of interest have at least one or two jets pass the typical jet requirement defined in Section 4.1.3, with an additional requirement of having $\Delta R > 0.4$ with the high- p_T photon. These “primary” (selected) jets in events of interest will be used to evaluate mistag rates.

To extract the flavor dependence of the mistag rate, region F is further split into a “b-enhanced” region (region F^E or just region E) and a “b-reduced” region (region F^R or just region R) via the selection of an “extra” jet. This extra jet must not have been considered a primary jet and pass the following quality selection criteria:

- Have reconstructed $p_T > 50$ GeV.
- Have $|\eta| < 2.4$.

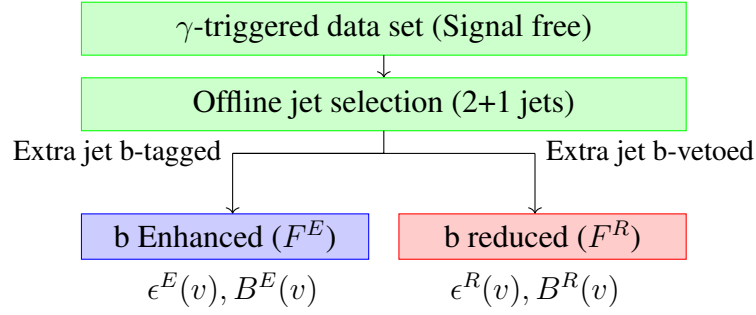


Figure 5.5: Selection flow of the signal-free region used for the evaluation of flavor-dependent mistag rate. The selection requirement for each block is described in Section 5.1.1. The calculation of the mistag rate $\epsilon^X(v)$ will be described in Section 5.1.2, and the calculation of the b jet fraction $B^X(v)$ will be described in Section 5.1.3.

- Pass the jet identification criteria as defined in Section 3.4.3.

Events where this extra jet is not found are discarded. Events are classified as belonging to region E if the extra jet passes the DeepJet b tagging requirements (described in Section 3.4.4) at the medium working point, and is classified as belonging to region R if the extra jet *fails* the DeepJet b tagging requirements at the loose working point. Because a significant fraction of b jets in events containing a high- p_T photon are generated from gluon splitting processes, an extra b jet in the event enhances the probability of the primary jets also being b jets. It should be noted that the extra jet is exclusively used for splitting the signal-free region, and will not be used for the mistag rate calculations.

5.1.2 Computing flavor dependence of mistag rate

Given the selected jets in the signal-free region X , the mistag rate binned in jet variable v , $\epsilon^X(v)$, can be calculated using the formula:

$$\epsilon^X(v) = \frac{N^{X,\text{tagged}}(v)}{N^{X,\text{all}}(v)} \quad (5.1)$$

where $N^{X,\text{all}}$ is the number of primary jets in region X that have jet variable in bin v , and $N^{X,\text{tagged}}$ is the subset of primary jets that also pass the EMJ tagging requirements.

The result of $\epsilon^E(v)$ and $\epsilon^R(v)$ will be different due to the strong flavor dependence of the mistag rate and the different b jet fractions of primary jets in the two regions. Assuming there exists a true underlying mistag rate for b jets, $\epsilon(b, v)$, and light jets, $\epsilon(l, v)$, the value of $\epsilon^X(v)$ can be expressed as:

$$\begin{cases} \epsilon^E(v) &= B^E(v)\epsilon(b, v) + (1 - B^E(v))\epsilon(l, v) \\ \epsilon^R(v) &= B^R(v)\epsilon(b, v) + (1 - B^R(v))\epsilon(l, v) \end{cases}, \quad (5.2)$$

where $B^X(v)$ is the estimated b jet fraction of primary jets in region X that fall into jet variable bin v . Once we have obtained $B^X(v)$, the relation can be inverted to calculate the flavor-dependent mistag rate from the regional mistag rate (omitting jet variable dependence for brevity):

$$\begin{cases} \epsilon(b, v) &= \frac{1-B^R}{B^E-B^R}\epsilon^E - \frac{1-B^E}{B^E-B^R}\epsilon^R \\ \epsilon(l, v) &= -\frac{B^R}{B^E-B^R}\epsilon^E + \frac{B^E}{B^E-B^R}\epsilon^R \end{cases}. \quad (5.3)$$

How $B^X(v)$ can be estimated will be detailed in Section 5.1.3.

Since the flavor dependence of this mistag rate is obtained via the “inversion” of a linear equation, let us label this mistag rate as $\epsilon_{\text{inv.}}(f, v)$. This notation is used to distinguish between mistag rates obtained using MC jets, where the flavor dependence can be calculated directly using truth information, which we label as $\epsilon_{\text{truth}}(f, v)$.

5.1.3 Estimating b jet fraction

The b jet fraction of a region of interest, $B^X(v)$, can be estimated by fitting the distribution of the DeepJet score to MC template distributions, where the flavor information can be obtained directly. For variable bin v and DeepJet score bin i , let us write the empirical distribution of the DeepJet score of jets that pass the primary jet selections of region X described in Section 5.1.1 obtained in data as $D_{v,i}^X$, where we normalize the distribution in each variable bin v such that $\sum_i D_{v,i}^X = 1$. Using the γ +jets MC sample that passes the same event and jet selection criteria and the available flavor truth information, the template score distribution can be written as $T_{v,i,f}^X$, where f can either be b or l . The template distributions are also normalized such that $\sum_i T_{v,i,f}^X = 1$. The b jet fraction, $B^X(v)$, can be estimated by minimizing the χ^2 measure:

$$\chi^2(B^X(v)) = \sum_i \frac{(D_{v,i}^X - (B^X(v)T_{v,i,b}^X + (1 - B^X(v))T_{v,i,l}^X))^2}{\sigma_{D_{v,i}^X}^2 + (B^X(v))^2 \sigma_{T_{v,i,b}^X}^2 + (1 - B^X(v))^2 \sigma_{T_{v,i,l}^X}^2}, \quad (5.4)$$

where σ_ξ is the uncertainty associated with the empirical distribution of interest given the bin value of ξ .

The uncertainties of the distribution for data σ_D only consider the counting uncertainties, while the template distribution will also take into consideration the recommended uncertainties from CMS. Discrepancies between the distribution of the DeepJet score seen in data and simulation have been studied extensively by the CMS collaboration as described in Sections 3.4.4 and 3.5.2.4, and per-jet scale factors have been developed to address the score distribution discrepancies seen. Alternate methods used to quantify this data/MC discrepancy manifests as additional uncertainties in the $T_{v,i,f}^X$ distribution, and are summed in quadrature with the counting

uncertainties when evaluating σ_T .

5.1.4 Signal free region mistag rate evaluation

In this section, we will compare the results of the calculation shown in Section 5.1, and discuss the considerations that were taken into account when choosing a final value to be used to calculate the final results. Unless otherwise specified, we will use the mistag rate parameterized in jet p_T (so $v = p_T$ in these following plots). The reason for this choice of parameterization scheme will be further discussed in Sections 5.2.1 and 5.3.

5.1.4.1 Comparison of pixel geometry upgrade

Because of the pixel Phase-1 upgrade of the CMS detector [60], the tracking and vertex resolution profile is noticeably different between data collected in 2016 and data collected in 2017 and 2018. Because of this, the results presented in later sections will typically display results separately for results evaluated using data collected with the pixel Phase-0 detector geometry, and data collected with the pixel Phase-1 detector geometry. The data collected with pixel Phase-0 geometry corresponds to an integrated luminosity of 36.3 fb^{-1} , while the data collected with pixel Phase-1 geometry corresponds to an integrated luminosity of 101.3 fb^{-1} .

An example of how the evaluated mistag rate differs with the pixel geometry can be found in Fig. 5.6. The impact is found to be largest for the “a-tag”, which is to be expected, as the tagging variables are most directly impacted by tracking resolution effects.

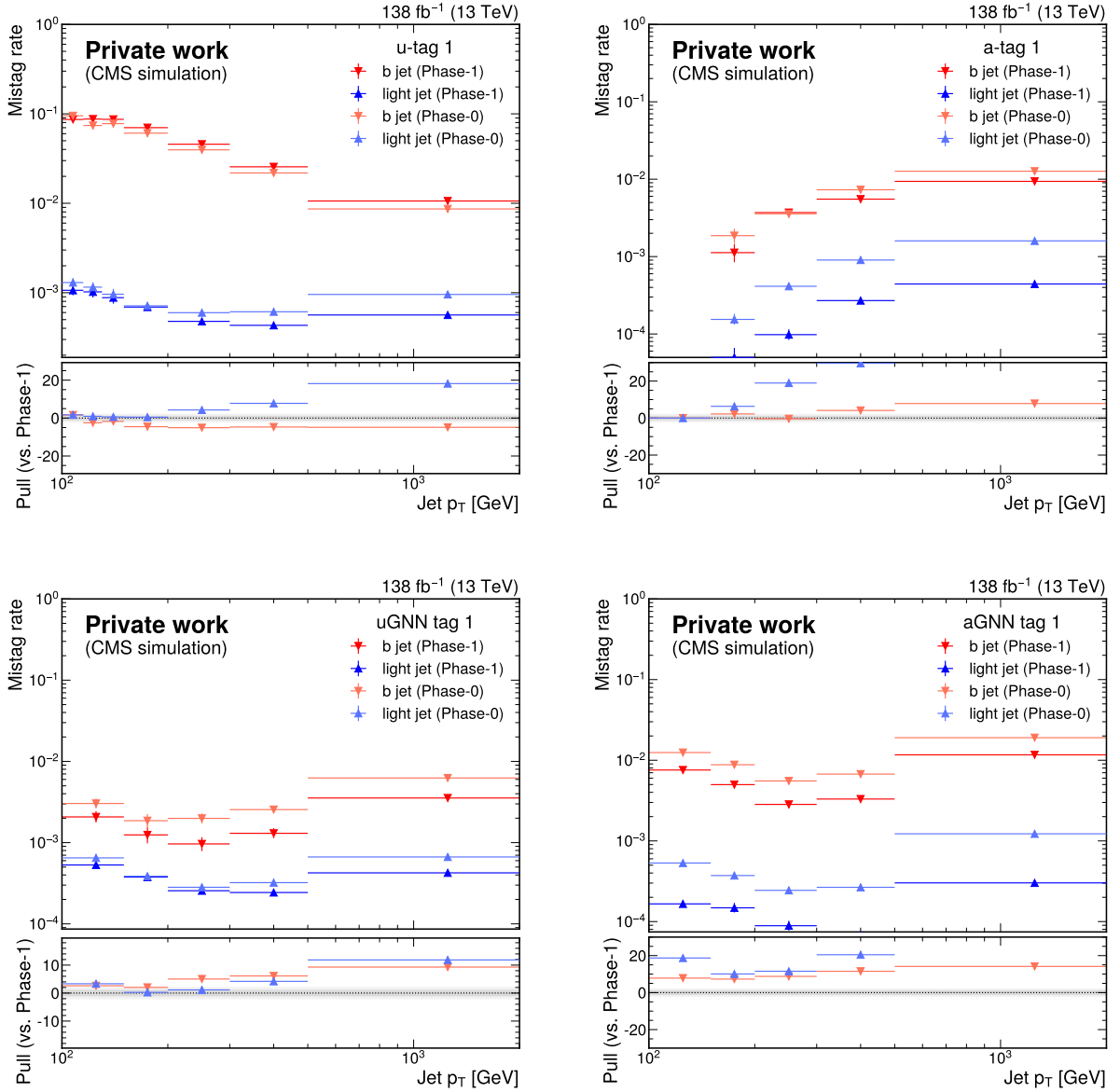


Figure 5.6: Comparison of the mistag rate of the u-tag 1 (top left), a-tag 1 (top right), uGNN tag 1 (bottom left) and uGNN tag 1 (bottom left) taggers, evaluated using SM light jets (blue) and SM b jets (red) collected with pixel Phase-1 geometry (solid) and pixel Phase-0 geometry (translucent) using the QCD MC samples. The jet flavor dependence is determined using MC truth information.

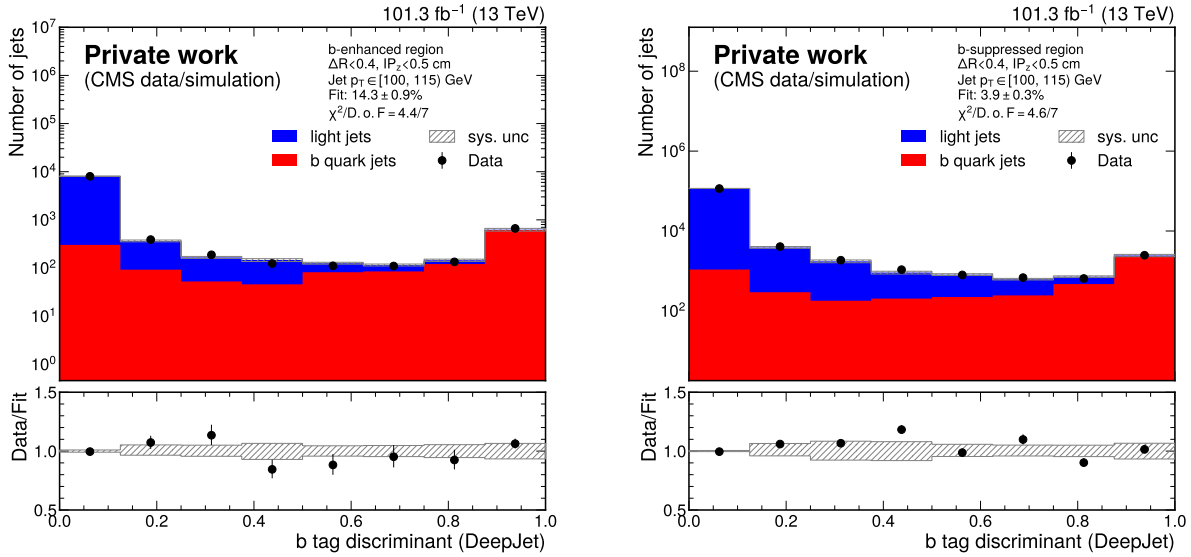


Figure 5.7: Example of the template fit running on the γ -triggered data set, as described in Section 5.1.3 running on jets with $p_T \in [100, 115)$ GeV in the b-enhanced (left) and b-suppressed (right) region. The jet-track association uses an angular radius of $R_{\text{cut}} = 0.4$.

5.1.4.2 Example of the b jet fraction fitting results

An example of the template fit running on the γ -triggered data set can be found in Fig. 5.7. Templates of the light jets and the b jets are generated using the γ +jets MC samples with jets passing the same selection criteria as jets used for data selection. Templates have the recommended per-jet scale factor applied to account for known DeepJet distribution discrepancies seen in data and MC. Here we are using the fit results using pixel Phase-1 geometry for an example.

The second example is the b jet fraction as a function of jet p_T , displayed in Fig. 5.8. The uncertainties here only include the uncertainties associated with the template fit as shown in Fig. 5.7.

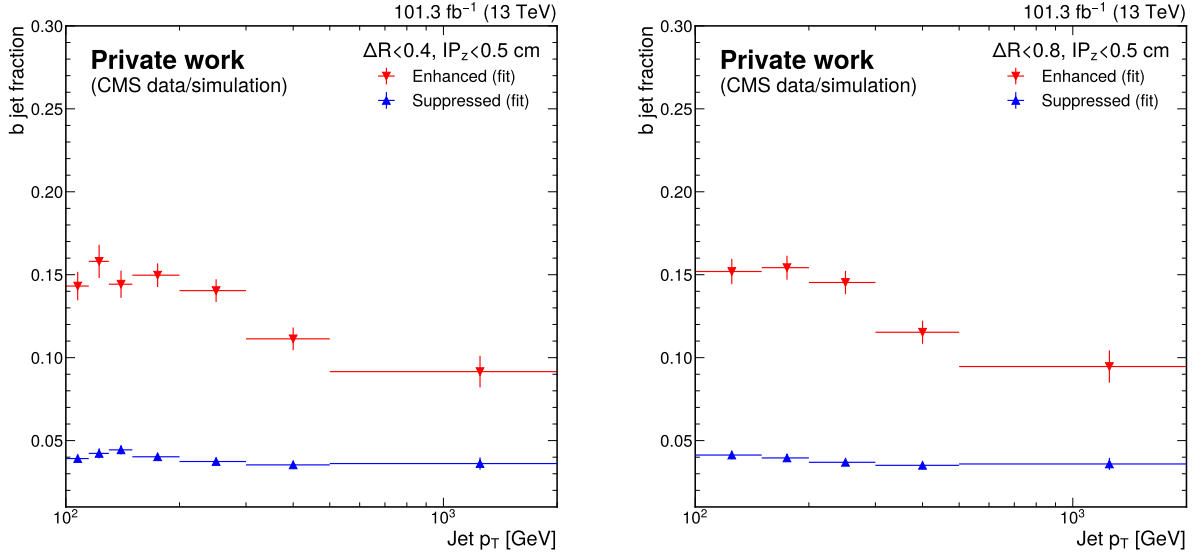


Figure 5.8: The fitted b jet fraction of jets in the b -enhanced region (red) and b -suppressed region (right) of jets in the γ -triggered data set with pixel Phase-1 geometry, with the results of using $R_{\text{cut}} = 0.4$ ($R_{\text{cut}} = 0.8$) jet-track association radius shown on the left (right).

5.1.4.3 Example of the mistag rate evaluated using the inversion formula

Finally, we can compare the mistag rate obtained via the inversion formula (Eq. 5.3). The uncertainties involved here include the statistical uncertainties involved in evaluating $\epsilon^X(v)$ and the uncertainty of the template fit. Details of how uncertainty is calculated and their impact on the analysis will be given in Section 6.1.1. An example of the evaluated mistag rate and comparison with direct calculation obtained in MC events can be found in Fig. 5.9.

While there are discrepancies between data and MC, this is not yet at the magnitude to cause alarms, as the difficulty in modeling data is the primary motivation for designing a fully data-based method. A proper method for evaluating the correctness of mistag rates calculated in data will be discussed in Section 5.4.2.

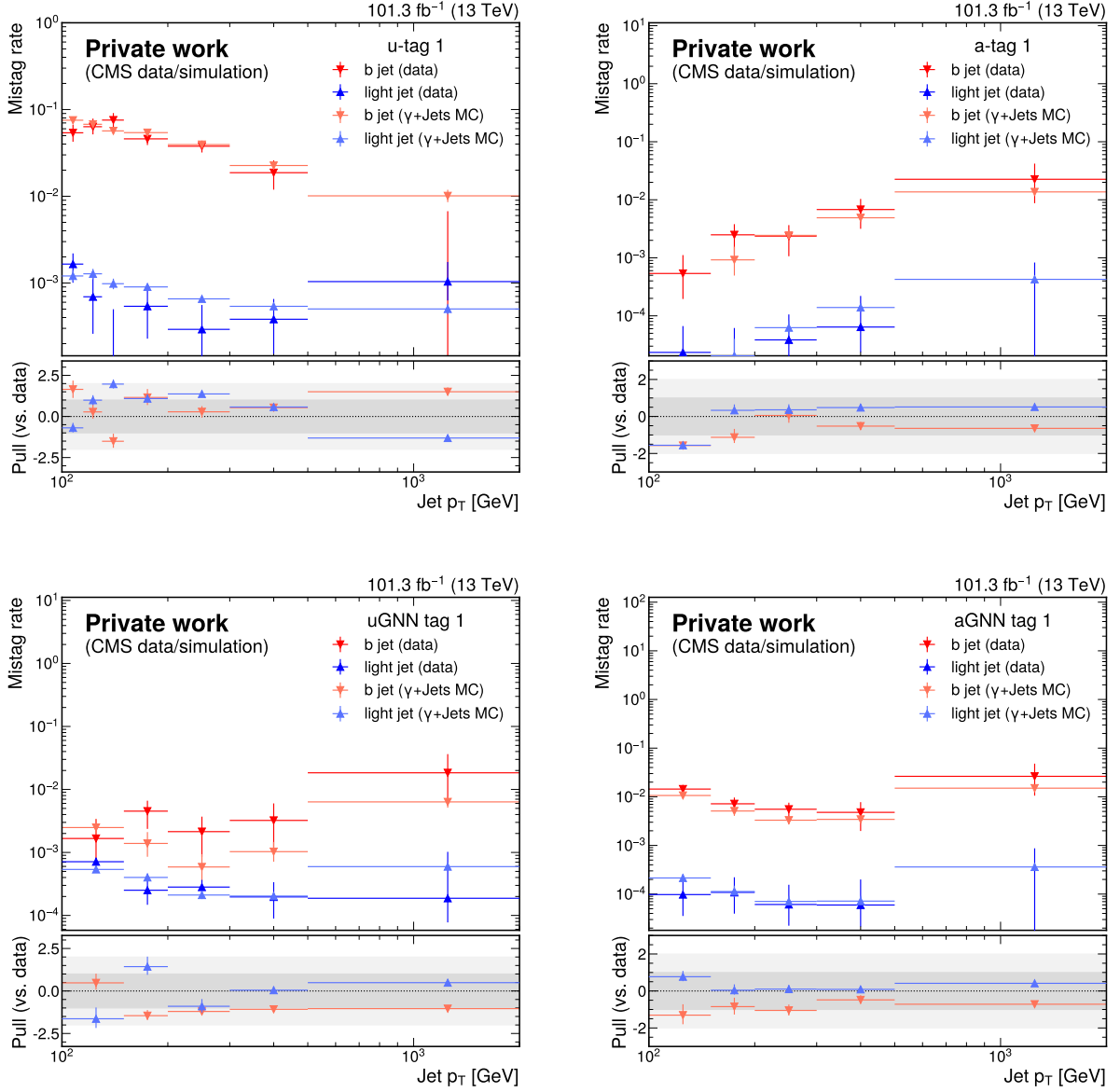


Figure 5.9: Comparison of the mistag rate of the u-tag 1 (top left), a-tag 1 (top right), uGNN tag 1 (bottom left) and uGNN tag 1 (bottom left) taggers, evaluated using SM light jets (blue) and SM b jets (red) collected with matrix inversion formula on data jets (solid) and truth information using γ +jets MC jets (translucent).

5.1.4.4 Comparison of mistag rates between evaluated between QCD and γ +jets MC events

The correctness of the mistag rate computation in the signal region assumes that jets in the signal-free region are similar in all kinematic distributions when compared with SM jets in the H_T -triggered data stream. As we can see in the example plot in Fig. 5.10, a comparison of mistag rate evaluated in QCD and γ +jets MC events, this is not strictly the case, with variations in the mistag rate being larger than the expected variations that can arise from statistical fluctuations.

Such discrepancies are not unique to the EMJ analysis, as other CMS analyses [106] see similar discrepancies in their jet variables of interest. This discrepancy is driven by the requirement of a high- p_T QCD-color singlet in the final state, which changes the probability of the underlying process that can generate jets in the event. An example of such discrepancies is demonstrated in the Feynman diagrams depicted in Fig. 5.11. An apparent discrepancy in jet kinematics of jets in the two regions can be seen in the track multiplicity as a function of jet p_T , as shown in Fig. 5.12 (which is also noted in Ref. [106]). While differences in mistag rates can be eliminated by parameterizing the mistag rate in multiple jet variables, this is not a practical approach, as there will not be enough jets in each variable bin to perform the calculations described in Section 5.1.2 reliably. We acknowledge that such differences exist and continue the calculations disregarding such differences for now. The impact of mistag rate differences on future calculations will be discussed in Section 5.4, while the calculation of appropriate uncertainties to cover such differences will be discussed in Section 6.1.2.

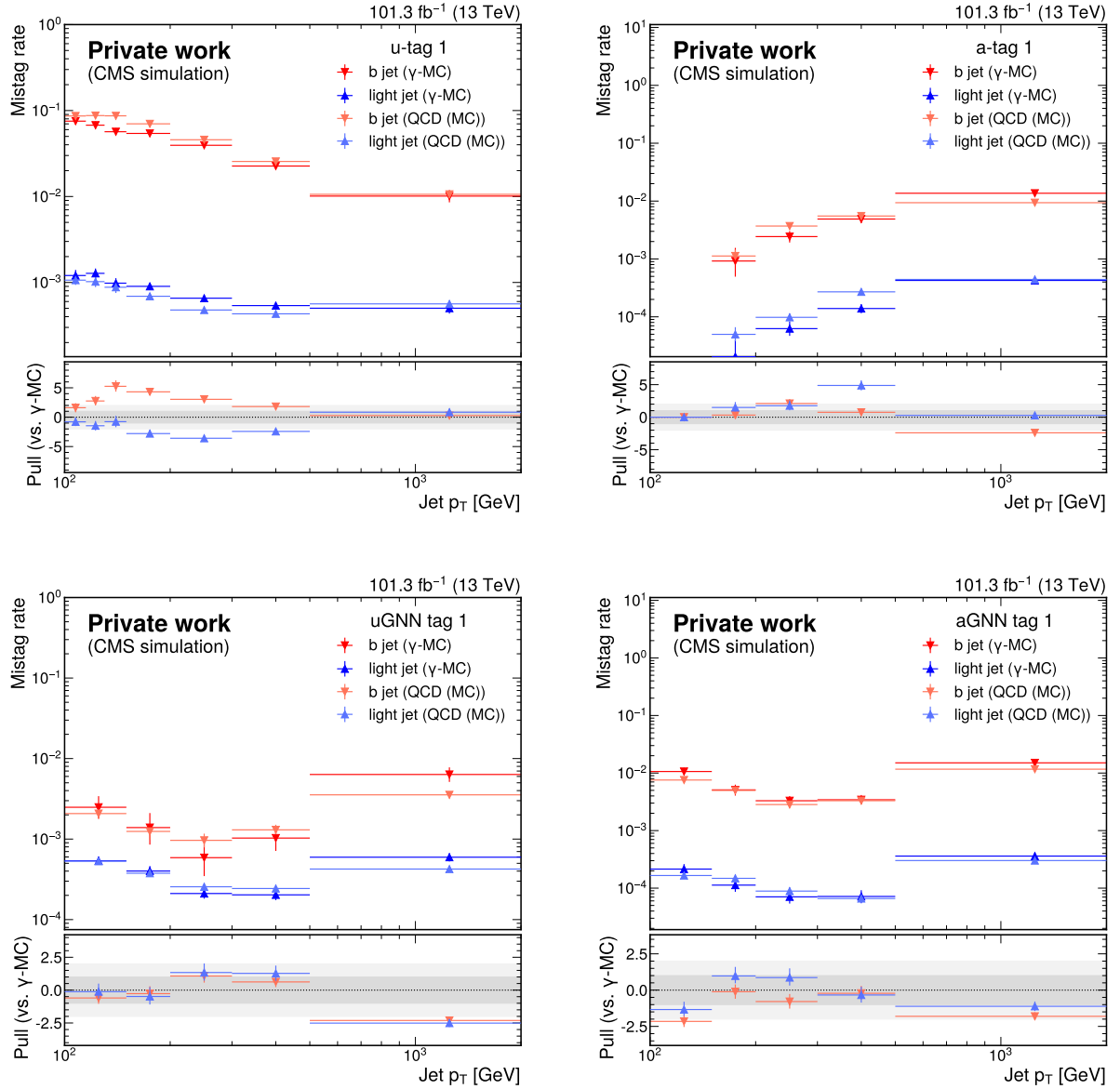


Figure 5.10: Comparison of the mistag rate of the u-tag 1 (top left), a-tag 1 (top right), uGNN tag 1 (bottom left) and uGNN tag 1 (bottom left) taggers, evaluated using SM light jets (blue) and SM b jets (red) collected with jets in the γ +jets MC sample (solid) and QCD MC sample (translucent).

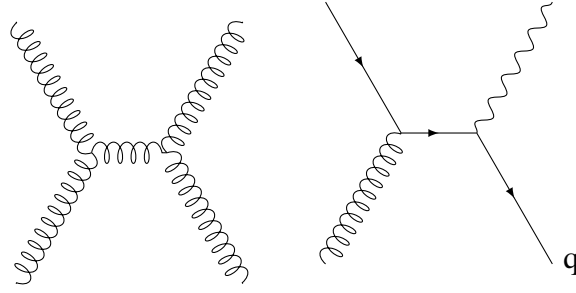


Figure 5.11: Example Feynman diagrams of the leading processes of generating jets in through QCD processes (left) and the γ +jets processes (right). The existence of a color singlet photon in the final state changes the composition of what can generate the jet in the final state.

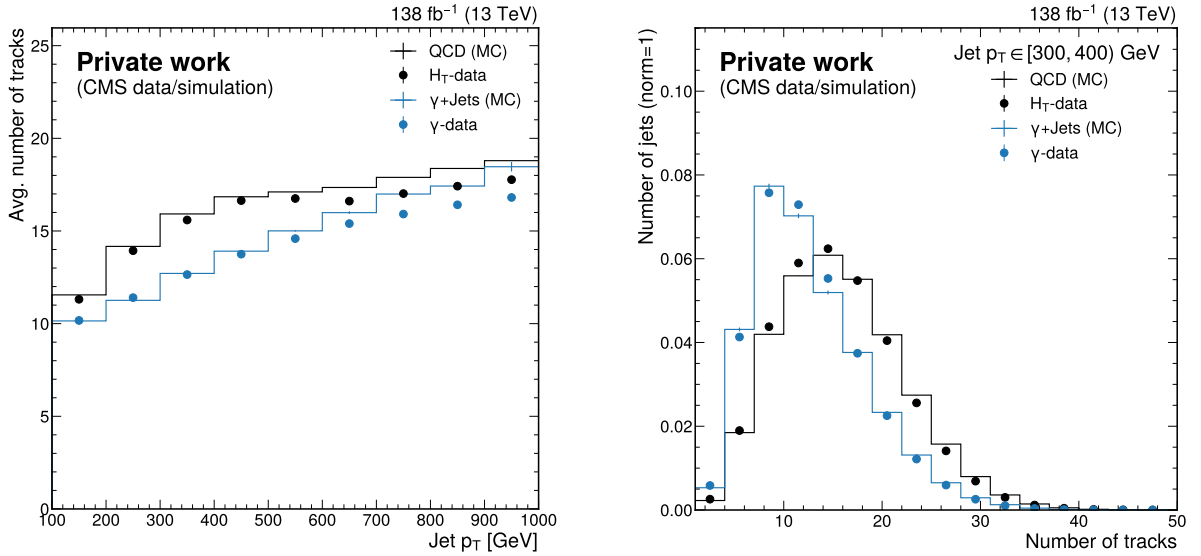


Figure 5.12: The number of tracks as a function of jet p_T (left) and the distribution of the number of tracks for jets with $p_T \in [300, 400) \text{ GeV}$ (right) evaluated for H_T -trigger data (black circle), γ -triggered data (gray diamonds), QCD MC samples (black lines), and the γ +jets MC samples (gray lines). For the H_T -triggered data and QCD MC samples, we only impose the trigger+4 jets requirements, as defined in Section 4.1.

5.2 The mistag rate scale factor

Events passing the defined signal region selection defined in Section 4.3 with ≥ 2 EMJ-tagged jets (region SR) comprise both SM and signal events. Let us label the number of SM and signal events that pass the signal region selection as S_{SM} and S_{sig} , respectively. Let us next define a control region (region CR) that requires the same event H_{T} and jet p_{T} selection as the signal region, and differs from the SR only by the number of EMJ-tagged jets in the events: exactly one EMJ-tagged jet be present in each event. We will label the number of SM and signal events that pass the control region selection as C_{SM} and C_{sig} , respectively. Since the mistag rate is typically of the order of $10^{-3} - 10^{-4}$, it is expected that $C_{\text{SM}} \gg C_{\text{sig}}$. Estimating S_{SM} on the total of events in the control region by some scale factor, SF , can be safely carried out:

$$S_{\text{SM}} = \sum_{i \in C_{\text{SM}}} SF_i(\{\epsilon_j\}) \xrightarrow{C_{\text{SM}} \gg C_{\text{sig}}} \sum_{i \in C} SF_i(\{\epsilon_j\}), \quad (5.5)$$

where i runs over events in the control region, and SF_i is some per-event SR/CR scale factor calculated using the mistag rate of jets within the event, $\{\epsilon_j\}$.

In the following sections, we will go over how the computation of SF_i can be performed, and the underlying assumptions required to ensure the correctness of the results.

5.2.1 Calculating an event-level scale factor

For the computation of a SR/CR scale factor, let us make the following assumptions:

1. The probability that an SM jet j is mistagged as an EMJ can be expressed as a function of jet variables $\vec{\theta}_j$, notated as $\epsilon(\vec{\theta}_j)$.

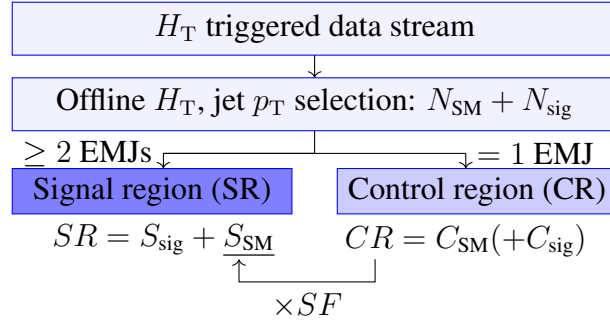


Figure 5.13: Selection flow that describes the construction of SR/CR pair defined by the primary SR selection criteria, and how it can be used to calculate the number of SM events that satisfy the SR selection criteria using events in CR.

2. For jets within the same event, the probability of each jet being mistagged is uncorrelated with each other, excluding the correlation of jet variables used to parameterize the mistag rate, $\vec{\theta}_j$.

Given these assumptions, starting from the global region of events and jets passing the event H_T and jet p_T selections, the probability of an event having n EMJ-tagged jets can be expressed as a function of the mistag rate of jets in the event, which we write as $G(n, \{\epsilon_j\})$, and can be calculated by listing the possible combinatorics. To simplify the calculation, let us only consider the leading 4 jets in the events (summation and product indices run from 0 to 3 unless otherwise specified). Systematically listing G given the possible tagging configuration, we get the following terms:

- $G(0, \{\epsilon_j\}) = \prod_i (1 - \epsilon_i)$.

The configuration where all four jets are correctly identified as not being an EMJ.

- $G(1, \{\epsilon_j\}) = \sum_i \epsilon_i \left(\prod_{j \neq i} (1 - \epsilon_j) \right)$.

This is the probability of any one of the jets within the event being mistagged as an EMJ, so the indices i and j run over all jets with the additional condition that $j \neq i$

- $G(2, \{\epsilon_j\}) = \frac{1}{2!} \sum_i \epsilon_i \left(\sum_{j \neq i} \epsilon_j \prod_{k \neq i, j} (1 - \epsilon_k) \right).$

This is the probability of any two jets in the event being misidentified as EMJs. The extra factor of $1/2!$ is needed when taking into account the second summation is unordered ($\sum_{j \neq i}$ rather than $\sum_{j > i}$). The reason for using an ordered sum allows for the simplification of further calculations presented in the next paragraph.

The following cases are very unlikely to occur given the number of events passing the selection criteria in the Run 2 dataset. We still compute the remaining terms for the sake of completeness:

- $G(3, \{\epsilon_j\}) = \frac{1}{3!} \sum_i \epsilon_i \left(\sum_{j \neq i} \epsilon_j \left(\sum_{k \neq i, j} \epsilon_k \left(\prod_{m \neq i, j, k} (1 - \epsilon_m) \right) \right) \right).$

This is a continuation of the computation used for $G(2, \{\epsilon_j\})$, except with three EMJ-tagged jets instead of two.

- $G(4, \{\epsilon_j\}) = \frac{1}{4!} \sum_i \epsilon_i \left(\sum_{j \neq i} \epsilon_j \left(\sum_{k \neq i, j} \epsilon_k \left(\sum_{m \neq i, j, k} \epsilon_m \right) \right) \right).$

This is the final continuation of the calculation corresponding to all four jets being tagged as EMJ. As a sanity check, one can verify that this term is equivalent to the more intuitive

form of $G(4, \{\epsilon_j\}) = \prod_i \epsilon_i$

Since both the control and signal regions require at least 1 EMJ to be mistagged, we should not use these probability results directly; instead these probabilities can be used to compute the fractional probability of observing n EMJ-tagged jets $Q(n, \{\epsilon_j\})$, given that we already know one of the jets is EMJ-tagged. Let us say that the t -th jet is known to be EMJ-tagged, we can list compute the various terms for Q to be:

- $Q(0, \{\epsilon_j\}) = 0$ by construction.

- $Q(1, \{\epsilon_j\}) = \epsilon_t \prod_{i \neq t} (1 - \epsilon_i).$

The form of $Q(1, \{\epsilon_j\})$ follows the same form as $G(1, \{\epsilon_j\})$, except that instead of summing over all possible values of the jet index i , we pick out the particular configurations of interest where the t -th jet is tagged as an EMJ.

- $Q(2, \{\epsilon_j\}) = \frac{1}{2!} \epsilon_t \left(\sum_{i \neq t} \epsilon_i \left(\prod_{j \neq t, i} (1 - \epsilon_j) \right) \right).$

The form of $Q(2, \{\epsilon_j\})$ follows the same argument used for $Q(1, \{\epsilon_j\})$. In addition, by choosing to use an unordered sum for index j , the summation over the second EMJ-tagged jet in the event can be kept in the same form as $G(2, \{\epsilon_j\})$. The summation would take a more complicated form if an ordered index were used for expressing $G(2, \{\epsilon_j\})$.

Again, for the sake of completeness, let us complete the following terms for $n = 3, 4$.

- $Q(3, \{\epsilon_j\}) = \frac{1}{3!} \epsilon_t \left(\sum_{i \neq t} \epsilon_i \left(\sum_{j \neq t, i} \epsilon_j \left(\prod_{k \neq t, i, j} (1 - \epsilon_k) \right) \right) \right).$

- $Q(4, \{\epsilon_j\}) = \frac{1}{4!} \epsilon_t \left(\sum_{i \neq t} \epsilon_i \left(\sum_{j \neq t, i} \epsilon_j \left(\sum_{k \neq t, i, j} \epsilon_k \right) \right) \right).$

Using Q , we can compute the SR/CR scale factor for each event in the control region to be:

$$\begin{aligned}
SF(\{\epsilon_j\}) &= \frac{Q(2, \{\epsilon_j\}) + Q(3, \{\epsilon_j\}) + Q(4, \{\epsilon_j\})}{Q(1, \{\epsilon_j\})} \\
&= \frac{\frac{1}{2} \sum_{i \neq t} \epsilon_i \left(\prod_{j \neq t, i} (1 - \epsilon_j) \right) + \frac{1}{6!} \sum_{\substack{i \neq t \\ j \neq t, i}} \epsilon_i \epsilon_j \left(\prod_{k \neq t, i, j} (1 - \epsilon_k) \right) + \frac{1}{4!} \sum_{t \neq i \neq j \dots} \epsilon_i \epsilon_j \epsilon_k}{\prod_{i \neq t} (1 - \epsilon_i)} \quad (5.6) \\
&\sim \frac{1}{2} \sum_{i \neq t} \epsilon_i.
\end{aligned}$$

The final approximation holds true under the assumption that $\epsilon_j \ll 1$.

5.2.2 Flavor assignment in SF computation

In the calculation above, $\epsilon_j = \epsilon(\vec{\theta}_j)$ assumes that all jet variables used for parameterizing the mistag rate can be perfectly assigned. For kinematic variables, the assignment of these variable is unambiguous. However, for jet flavor, some method of flavor assignment would need to be performed, as flavor information is not directly available in data. For ease of expression, let us explicitly separate $\vec{\theta}_j$ into the jet flavor and other kinematic variables $(f_j, \vec{v}_j)$. The value of SF can be directly calculated using flavor truth information for events in MC:

$$SF_{\text{truth}}(\{\epsilon_j\}) = SF(\{\epsilon(f_{j,\text{truth}}, \vec{v}_j)\}), \quad (5.7)$$

but this evaluation of SF_{truth} is not possible for data. The following section will discuss the two methods proposed for flavor assignment in the SF assignment in the original paper, showing the advantages and problems with each of the methods, as well as the final choice for this analysis.

5.2.2.1 Flavor averaging method

Instead of attempting to determine the flavor of each jet, we extract the b jet fraction of the non-EMJ tagged jets in the control region B^{CR} . A “flavor-averaged” scale factor, $SF_{\text{avg.}}$, can then be calculated as:

$$SF_{\text{avg.}}(\{\epsilon_j\}) = SF(\{B^{\text{CR}}\epsilon(b, \vec{v}_j) + (1 - B^{\text{CR}})\epsilon(l, \vec{v}_j)\}). \quad (5.8)$$

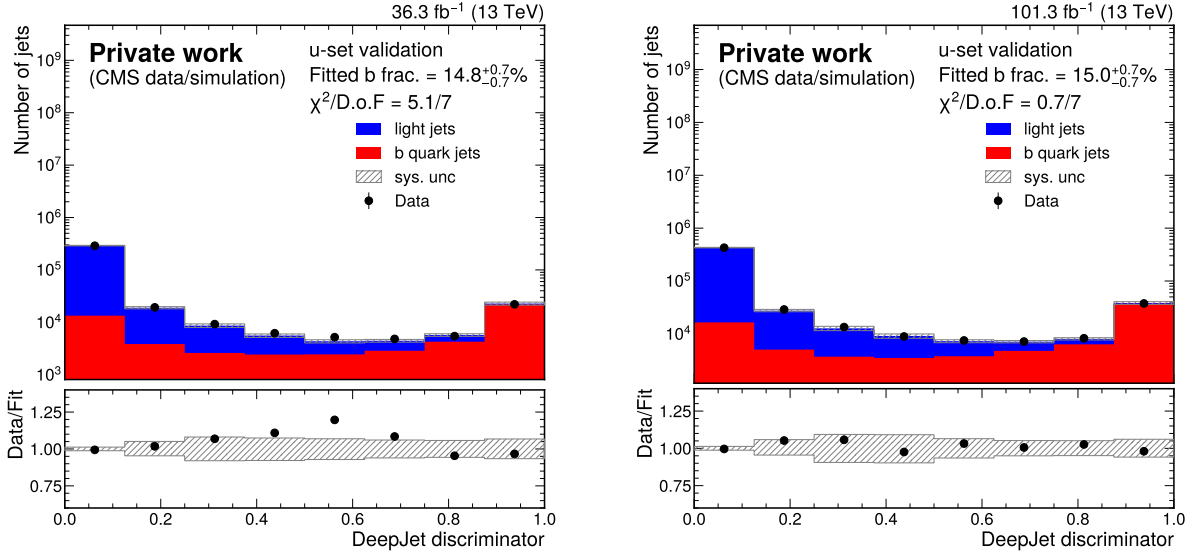


Figure 5.14: Example of the template fit of the non-EMJ-tagged jets in the control region defined by the validation u-set selection criteria (defined in Table 4.4) using pixel Phase-0 (1) data on the left (right).

The extraction of a flavor-dependent mistag rate $\epsilon(f, \vec{v})$ is still important here, as the b-jet fraction of the control region is not the same as the b jet fraction of the signal-free region.

The advantage of this method is that it is very straightforward to calculate: B^{CR} is calculated over numerous jets, and can be extracted reliably using the DeepJet distributions, as described in Section 5.1.3. An example of the b jet fraction evaluated using the events in CR using data can be found in Fig. 5.14.

The potential issue with this approach is that the results of this calculation do not strictly allow the assumptions listed at the start of Section 5.2 to hold: jet flavors of SM background events are correlated, as b jets generated in SM QCD processes are most probable to come from gluon splitting processes. While this region-averaging method is expected to converge to the correct value given a sufficiently large control region, and such correlations within events are also averaged out when we count the total number of events, this may cause problems with selection cuts with higher $H_{\text{T}}/p_{\text{T}}$ selection thresholds, where there are fewer jets in the control region.

Also, because b jets are more likely to be tagged as EMJs, the result of SF_{avg} for this method can only strictly be used for $n_{\text{CR}} = 1 \rightarrow n_{\text{SR}} = 2$ computations and not $n_{\text{CR}} = 0 \rightarrow n_{\text{SR}} = 2$, as simple fractional factors do not take into account the enhanced probability that two b jets exist in the event, and are more likely to be mistagged as EMJs.

Because of these caveats, both this analysis and its predecessor also outline a method to attempt a per-jet flavor assignment, which we will present in the following section.

5.2.2.2 Per-jet flavor assignment using b tagging information

While the underlying flavor of each jet is not directly observable using reconstructed variables, a good proxy for jet flavor available at the reconstruction level is the b tagging discriminant variable, where we can tag jets as being b jets by imposing a discriminant threshold. Using jets MC events, we can calculate the random “smearing” that occurs in the reconstruction process of the jet flavor, the fraction of which true jet flavor differs from the reconstructed flavor tagging results due to finite information available to the flavor tagging algorithm. The smearing information can then be “unfolded” to obtain a reasonable estimate of the true underlying jet flavor configuration.

For the four jets in the event, let us call the underlying flavor configuration $U = (f_0, f_1, f_2, f_3)$, where f_j can either be a b jet ($f_j = b$) or a non-b jet ($f_j = l$), and the tagged flavor configuration $T = (f_0', f_1', f_2', f_3')$. What we can calculate using MC is the smearing relation, $M(T, U)$: the probability of obtaining a tagging configuration T given some true underlying flavor configuration U . If we can somehow calculate the “inverse” of this relation, $N(U, T)$, the scale factor

listed in Eq. 5.6 can be weighted according to the estimated underlying flavor configuration:

$$SF_{\text{unfold}}(\{\epsilon_j\}) = \sum_{U=(f_0, f_1, f_2, f_3)} N(U, T) SF(\epsilon(f_0, \vec{v}_0), \epsilon(f_1, \vec{v}_1), \epsilon(f_2, \vec{v}_2), \epsilon(f_3, \vec{v}_3)). \quad (5.9)$$

In the 2016 analysis, $M(T, U)$ was proposed to be treated as a matrix M_{ut} , where u and t are the numerical indices representing the 16 possible jet flavor configurations, and $N(U, T)$ is evaluated by taking the matrix inverse of M_{ut} . The problem with this approach is that all elements in M_{ab} are, by construction, smaller than unity. The matrix inverse, M_{ab}^{-1} , will then necessarily contain elements greater than unity and smaller than zero. Such a weighting factor makes interpreting Eq. 5.9 difficult, as well as introducing the possibility that SF_{unfold} gives an unphysical value of either greater than unity or smaller than zero. Such results will be particularly problematic if only a few events are in the control region.

Reference [107] properly treats the relation described by $M(T, U)$ as a conditional probability: $M(T|U)$, and inverts this relation using Bayesian inference:

$$N(U|T) = \frac{M(T|U)P_0(U)}{\sum_{U'} M(T|U')P_0(U')}, \quad (5.10)$$

where $P_0(U)$ is some prior probability estimate of the underlying jet flavor configuration. The use of Bayesian inference ensures that $N(U|T)$ has all the properties expected of conditional probabilities, which is essential for the calculation of the weighted scale factor in Eq. 5.9 (all values strictly between zero and one, $\sum_U N(U|T) = 1$). Reference [107] further details how one may evaluate P_0 in a way that is less dependent on direct results obtained from simulated events.

While seemingly robust, this method is problematic when attempting to calculate $M(T|U)$

using reconstructed variables. The simplest method of estimating $M(T|U)$ is to use the b tagging efficiency e and mistag rates m of the DeepJet b tag discriminant and evaluate $M(T|U)$ as:

$$M(T = \{f_j\}|U = \{f'_j\}) = \prod_j \begin{cases} e_j & \text{if } f_j = f'_j = \text{b}; \\ (1 - m_j) & \text{if } f_j = f'_j = 1; \\ (1 - e_j) & \text{if } f_j \neq f'_j = \text{b}; \\ m_j & \text{if } f_j \neq f'_j = 1. \end{cases} \quad (5.11)$$

The problem is that despite being trained on reconstructed objects within the jet cluster, the DeepJet algorithm is sufficiently powerful to infer event-level features that are not directly passed to the training algorithm. This feature of the DeepJet algorithm dramatically changes the b tagging efficiency and mistag rate evaluated in a general region compared with jets in the control region, with the difference approaching an order of magnitude (Fig. 5.15). There are always uncertainties in the b tagging efficiency and mistag rate, and the CMS collaboration provides recommended correction factors and their associated uncertainties to account for differences in data and MC samples. However, the differences between the b tagging efficiencies after imposing a 1-EMJ requirement are much larger than the official uncertainties.

5.2.2.3 Decision

Due to the large discrepancy in the b tagging efficiencies between what is calculated and what can be verified by officially recommended methods, the analysis team decided to continue using the flavor-average method to account for the jet flavor information when calculating SF . The b tag discriminant templates used to determine the b jet fraction of non-EMJ-tagged jets

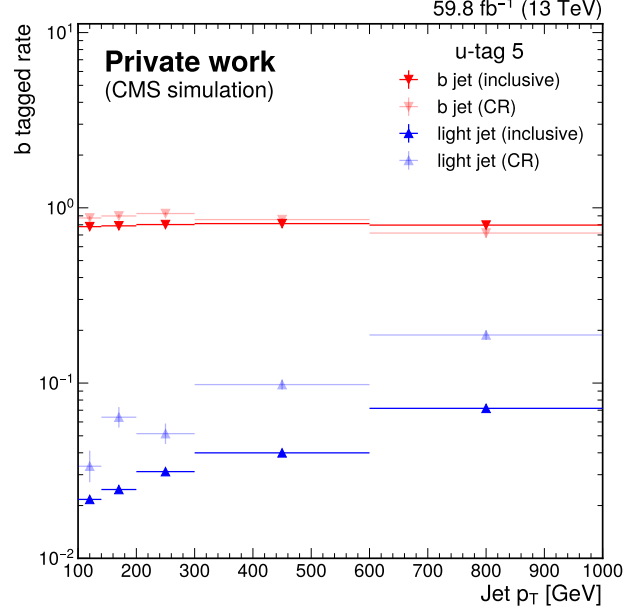


Figure 5.15: Difference in the DeepJet b tagging efficiency (red) and mistag rate (blue) as a function of jet p_T , with the CR 1-EMJ-tagged jet requirement (solid, labeled “CR”), and without the additional jet requirements (translucent, labeled “QCD”) evaluated using the QCD MC results.

in the control region are generated using the jets in the QCD MC sample with the same event selection criteria and the official uncertainties provided by the collaboration.

5.3 Determining mistag rate parameterization

While in the calculations above, calculations have all been carried out with mistag rate parameterized as $\epsilon(f, v)$, where v is a singular jet variable, the choice of this parameterization is intentionally left unspecified in the presentation of the calculations, as the choice of this parameter is not trivial. Here, we give the justification of how a concrete choice of v was determined.

When calculating the factors of $G(N, \{\epsilon_j\})$ in Section 5.2.1, simple combinations of ϵ_j were presented in the calculation. This use of simple combinatorics follows from the assumption at the start of Section 5.2.1 that jets being mistagged as EMJ are uncorrelated up to variables used to parameterize the mistag rate. As shown in Fig. B.1–5.4, EMJ mistag rates strongly depend

on jet flavor, jet p_T , and the number of associated tracks. Conceptually, for the calculation presented in Section 5.2.1 to be mathematically valid, we must parameterize the mistag rate with infinitely granular jet and event information to capture all potential mistag rate correlations. In other words, the calculations presented in Section 5.2.1 are only valid using:

$$\epsilon_j = \epsilon(f, p_T, n_{\text{track}}, \dots) = \epsilon(\vec{\theta}^\infty). \quad (5.12)$$

In practice, given the finite number of jets available for calculating mistag rate values, we can only parameterize the mistag rates in a finite number of variables, so we will need to pick variables that best capture potential EMJ tagging correlations.

The choice to always parameterize in jet flavor is due to flavor correlation being fully physics-driven: it is not a secondary result caused by particularities in the physic object reconstruction algorithm or detector resolution effect, and thus irreducible. Strong flavor dependence of the mistag rate is also seen for all surveyed jet kinematics and EMJ tagging schemes. After the additional jet population separation required for flavor dependence evaluation, as described in Section 5.1.2, we can only practically parameterize the mistag rate in another jet variable to keep the calculation statistically robust. The choice of variable, then, takes into consideration the following issues:

- The parameterization needs to capture EMJ tagging features common in both the H_T -triggered data set and the γ -triggered data set, as SM jet populations can have a nontrivial shift when imposing event-level selection requirements. Additional information on why and how such population shifts occur is further discussed in Sections 5.1.4.4 and 6.1.

- The choice of parameterization should attempt to capture all physical correlations across jets in the event, such that the assumption described at the start of Section 5.2 is best preserved.
- Mistag rate dependencies are not always physics-driven, but may present a component driven by the statistical algorithm employed for the EMJ tagging computations. For example: the sampling median used for the model-agnostic taggers for unflavored signal models is expected to show very strong dependence on the sampling number towards high track multiplicity, regardless of any track displacement physics. This dependence on sampling number is an analog of the central limit theorem of sampling means, with the sampling median $\tilde{m}$ of a probability distribution f with sampling number N reaching a limiting distribution $P(\tilde{m})$ of:

$$\lim_{N \rightarrow \infty} P(\tilde{m}; f, N) \rightarrow \mathcal{N}\left(\tilde{m}; m, \frac{1}{2f(m)\sqrt{N-1}}\right) \quad (5.13)$$

where m is the distribution median of the target distribution f , and $\mathcal{N}$ is the normal distribution. The width of $P(\tilde{m})$ has a sharp dependence on N , indicating that when we are using outliers of $\tilde{m}$ as a method for tagging, the sharp dependence on N can be an algorithmic-driven consequence of using sampling medians as the test statistic, not necessarily an indication that there are fundamental shifts in the underlying distribution f . This sharp dependence on the N can be demonstrated with a simple numerical experiment in Fig. 5.16. All tracks are generated to have $\text{IP}_{2\text{D}}$ following a fixed exponential distribution with varying tracks in each collection. Here, we can see how the distribution of the sampling median $\langle \text{IP}_{2\text{D}} \rangle$ narrows and the fraction of outliers falls rapidly as N increases.

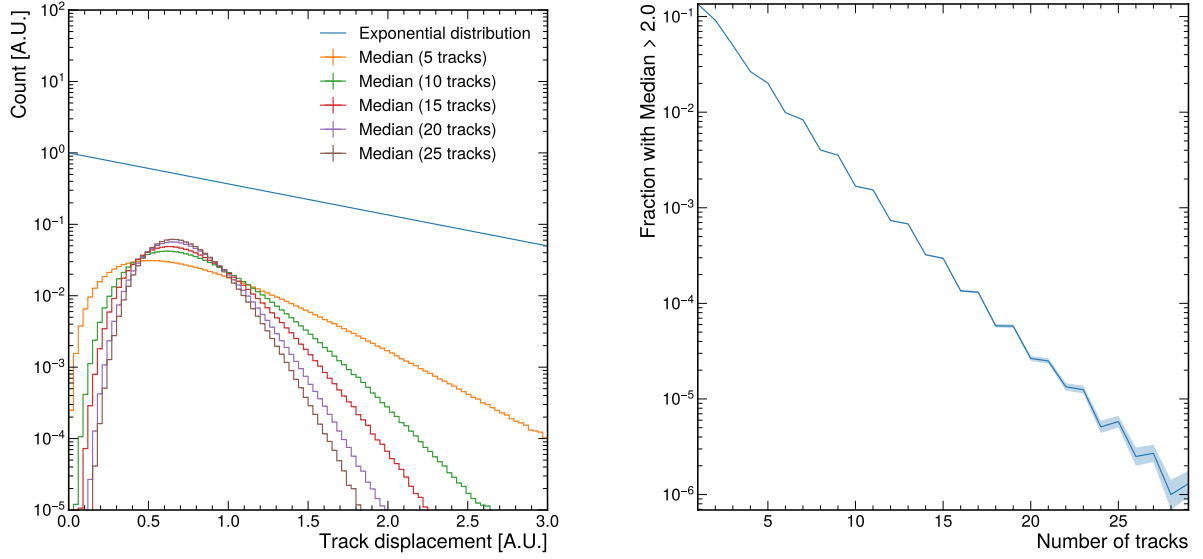


Figure 5.16: Example of the median distribution obtained from sampling from a exponential distribution of various sample sizes (left), and the fraction of median “outliers” as a function of sampling size (right).

Because of the reasons given above, the parameterization is tested by comparing the results obtained by swapping out the ingredients for the background estimation, and comparing which method of parameterization yields the most consistent background estimation results.

As the following notation will be useful for further discussion, let us take time to clarify the terminology:

- The *predicted* number of background events is obtained by counting the number of MC events that pass the signal selection criteria. This value is typically labeled as “Pred.” in tables and figures.
- The *estimated* number of background events is obtained by applying the SF calculation to a control region, and counting the weighted sum. This value is typically labeled as “Est.” in tables and figures. As many components go into the calculation of SF , we will use the

following notation to specify which estimation result we are comparing:

$$\text{Est}_a^R(\epsilon_d^F(v)), \quad (5.14)$$

where:

- R is used to indicate which data set is used to generate the CR. The index R can either be the QCD MC sample or the H_T -triggered data (notated simply as H_T).
- a is used to indicate the flavor assignment technique used for SF calculation. The index a can be either truth flavor assignment for the QCD MC CR (labeled “truth”, using Eq. 5.7), or a flavor averaged result (labeled “avg.”, using Eq. 5.8).
- F is used to indicate which signal-free data set is used to evaluate the EMJ mistag rate of SM jets. The index F can be the QCD MC samples, the γ +jets MC sample (label “GJets”), or the γ -triggered data set (labeled simply as γ).
- d is used to indicate how the flavor dependence of the mistag rate is determined. The index d can either be using flavor truth information for the MC samples (labeled “truth”) or the linear relation inversion method (labeled “inv.”, using Eq. 5.3).
- v is used to indicate which jet variable is used to parameterize EMJ mistag rates.

For the comparison of consistency, we will compare the estimation results of $E_1(v) = \text{Est}_{\text{truth}}^{\text{QCD}}(\epsilon_{\text{truth}}^{\text{QCD}}(v))$, $E_2(v) = \text{Est}_{\text{truth}}^{\text{QCD}}(\epsilon_{\text{truth}}^{\text{GJets}}(v))$, and $E_3(v) = \text{Est}_{\text{avg.}}^{\text{QCD}}(\epsilon_{\text{inv.}}^{\text{GJets}}(v))$ with the predicted values using different parameterization variables v , and find which one gives the smallest figure-

of-merit, $\xi(v)$:

$$\xi(v) = \Delta(E_1(v)) + \Delta(E_2(v)) + \Delta(E_3(v)) + \Xi(E_1, E_2) + \Xi(E_2, E_3) ; \quad (5.15)$$

$$\Delta(x) = \left(\frac{x - \text{Pred}}{\sigma_{\text{Pred}}} \right)^2 ; \quad (5.16)$$

$$\Xi(x, y) = \left(\frac{x - y}{\sigma_{\text{Pred}}} \right)^2 . \quad (5.17)$$

The terms $\Delta(E_j(v))$ are chosen to compare the correctness of the estimation results with the direct selection predictions. The terms $\Xi(E_i(v), E_j(v))$ are used to compare the consistency of the estimation results when changing ingredients used in the calculation of the background estimation. The comparison will always be made relative to the statistical uncertainty of the direct prediction results σ_{Pred} , which is expected to be the most significant uncertainty. The choices of $E_i(v)$ are further discussed when defining systematic uncertainties in Section 6.1.

Since the EMJ mistag rate shows the steepest dependence in jet p_T and n_{track} , we compare the results of $\xi(p_T)$ and $\xi(n_{\text{track}})$. To determine which parameterization scheme is “superior,” we compare the results of $\xi(v)$ averaged across cut sets that use similar EMJ tagging schemes. To avoid ambiguity in the σ_{Pred} for 0-count cut sets, the calculation of $\xi(v)$ will use the sum over the full Run 2 MC samples available. The evaluated results of $\xi(v)$ are given in Table 5.1. Comparing the results, we determine that p_T gives the most consistent results across the EMJ tagging schemes employed, and hence the primary figures in the leading sections mostly feature mistag rates parameterized in jet p_T .

Table 5.1: Computation of the figure-of-merit defined in Eq. 5.15, used to determine the binning variable for mistag rate parameterization for the various cut sets defined in Table 4.4. The values listed in columns 3 and 5 are the average of the figure-of-merit cross each group of cut sets.

Cut set	$\xi(p_T)$		$\xi(n_{\text{track}})$	
u-set 1	1.520		19.845	
u-set 2	6.496		3.298	
u-set 3	0.999	5.325	3.484	12.27
u-set 4	2.539		4.118	
u-set 5	0.121		1.388	
validation u-set	20.266		44.230	
a-set1	5.311		3.512	
a-set2	17.356		26.857	
a-set3	1.671	6.072	1.320	7.474
a-set4	0.129		0.522	
a-set5	10.501		11.201	
validation a-set	1.466		1.430	
uGNN set 1	1.032		2.174	
uGNN set 2	1.475	5.152	1.140	5.391
uGNN set 3	1.520		9.717	
uGNN validation set	16.582		8.534	
aGNN set 1	6.707		12.443	
aGNN set 2	0.361	3.971	0.669	6.081
aGNN set 3	5.021		8.518	
aGNN validation set	3.793		2.695	

5.4 Assessment of background estimation correctness

Given the complexity of the background estimation method described across Sections 5.1–5.3, we will want to validate and assess the potential systematic issues in the estimation scheme before using this to conclude whether there is an excess of signal-like events in data. Following the notation introduced in Eq. 5.14, we want to ensure the estimation $\text{Est}_{\text{avg.}}^{H_T}(\epsilon_{\text{inv.}}^\gamma(p_T))$ is consistent with the actual number of SM events passing the SR selection criteria. If given a signal-free data set, a correct estimation calculation will yield the same number (within uncertainty) when counting the number of events passing the SR selection and weighting the number of events in the accompanying control region by the assigned scale factor. Two validation checks of this type will be presented below.

5.4.1 Validation test using MC events

Using QCD MC samples as the signal-free stand-in for the H_T -triggered data set, and the γ +jets MC samples as a stand-in for the γ -triggered data stream, we can compare the results of the MC prediction with the estimation results $\text{Est}_{\text{avg.}}^{\text{QCD}}(\epsilon_{\text{inv.}}^{\text{GJets}}(p_T))$ for the various cut sets. Since only simulated SM events are used, the calculation is guaranteed to be signal free, and would be the most straightforward check to see whether there are any errors in the algorithm.

The numerical values of the prediction and estimation results using simulated events with the pixel Phase-0 (1) geometry can be found in Table 5.2 (5.3), with a graphical equivalent of the tables displayed in Fig. 5.17 (5.18). In these tables, the predicted results only contain the counting uncertainties. The estimation results include additional systematic uncertainties that will be discussed in Section 6.1. In the plots, the systematic uncertainties are summed in quadrature

with the counting uncertainties. We can see that for nearly all defined cut sets, prediction and estimation results close within one standard deviation of the uncertainty, with outlying results typically either having zero events passing the full signal selection criteria, or only a few events with very large weighting. This result indicates that our background estimation calculation is correct and robust after considering systematic uncertainties.

Table 5.2: Comparison of the predicted and estimated results using the SR/CR pair constructed using QCD MC samples with pixel Phase-0 geometry. The number of SM events that pass the signal region selection (with statistical uncertainties) is listed in column 2. The estimation results with scale factors calculated using mistag rates evaluated using the QCD MC samples (with statistical and systematic uncertainties) are listed in column 3. The estimation result with scale factors calculated mistag rates evaluated using the γ +jets MC samples (with statistical and systematic uncertainties) is listed in column 4. The computation of uncertainties is further discussed in Section 6.1.2. Definition of the additional validation cut sets will be given in Section 5.4.2.

Cut set	Pred.	Est _{truth} ^{QCD} ($\epsilon_{\text{truth}}^{\text{QCD}}(p_T)$)	Est _{avg.} ^{QCD} ($\epsilon_{\text{inv.}}^{\text{GJets}}(p_T)$)
u-set 1	13.7 ^{+1.8} _{-1.6}	14.4 ^{+0.3} _{-0.3} ± 4.2	17.9 ^{+4.4} _{-2.2} ± 7.0
u-set 2	4.7 ^{+2.6} _{-1.8}	2.96 ^{+0.09} _{-0.09} ± 0.87	4.1 ^{+2.0} _{-0.9} ± 1.6
u-set 3	5.5 ^{+2.6} _{-1.9}	3.71 ^{+0.08} _{-0.08} ± 0.56	5.2 ^{+2.3} _{-0.9} ± 2.3
u-set 4	1.02 ^{+0.67} _{-0.43}	1.18 ^{+0.06} _{-0.06} ± 0.23	1.57 ^{+0.42} _{-0.36} ± 0.56
u-set 5	6.6 ^{+4.8} _{-3.0}	1.64 ^{+0.06} _{-0.06} ± 0.23	2.04 ^{+0.53} _{-0.34} ± 0.76
validation u-set	879.1 ^{+67.4} _{-62.7}	786.1 ^{+13.6} _{-13.7} ± 114.9	739.5 ^{+63.5} _{-64.9} ± 123.1
a-set 1	5.9 ^{+2.4} _{-1.8}	3.92 ^{+0.08} _{-0.08} ± 0.34	3.63 ^{+0.72} _{-0.31} ± 1.00
a-set 2	0.30 ^{+0.24} _{-0.15}	0.612 ^{+0.021} _{-0.021} ± 0.027	0.52 ^{+0.10} _{-0.05} ± 0.14
a-set 3	1.7 ^{+2.9} _{-1.2}	0.78 ^{+0.04} _{-0.04} ± 0.12	0.83 ^{+0.17} _{-0.09} ± 0.18
a-set 4	1.09 ^{+0.41} _{-0.31}	1.40 ^{+0.04} _{-0.04} ± 0.18	1.48 ^{+0.31} _{-0.14} ± 0.28
a-set 5	8.9 ^{+1.2} _{-1.1}	9.22 ^{+0.12} _{-0.12} ± 0.07	7.0 ^{+0.9} _{-0.5} ± 2.4
validation a-set	62.3 ^{+18.1} _{-14.3}	56.7 ^{+1.4} _{-1.3} ± 7.8	67.2 ^{+3.0} _{-2.9} ± 13.6
uGNN set 1	2.18 ^{+0.69} _{-0.54}	2.19 ^{+0.06} _{-0.06} ± 0.26	2.61 ^{+0.64} _{-0.28} ± 0.47
uGNN set 2	0.15 ^{+0.21} _{-0.10}	0.115 ^{+0.007} _{-0.007} ± 0.011	0.148 ^{+0.065} _{-0.028} ± 0.051
uGNN set 3	1.46 ^{+0.48} _{-0.37}	1.46 ^{+0.03} _{-0.03} ± 0.16	1.33 ^{+0.39} _{-0.17} ± 0.38
uGNN validation set	4.0 ^{+1.6} _{-1.2}	4.04 ^{+0.14} _{-0.13} ± 0.36	6.3 ^{+1.1} _{-0.6} ± 2.2
aGNN set 1	9.5 ^{+1.2} _{-1.0}	7.7 ^{+0.2} _{-0.2} ± 2.4	9.8 ^{+1.9} _{-1.1} ± 3.5
aGNN set 2	0.0 ^{+0.0} _{+0.0}	0.131 ^{+0.008} _{-0.008} ± 0.067	0.153 ^{+0.094} _{-0.030} ± 0.089
aGNN set 3	1.44 ^{+0.45} _{-0.35}	0.83 ^{+0.03} _{-0.03} ± 0.42	1.14 ^{+0.48} _{-0.17} ± 0.63
aGNN validation set	6.9 ^{+3.1} _{-2.3}	6.91 ^{+0.25} _{-0.24} ± 0.55	7.8 ^{+1.6} _{-0.8} ± 1.1

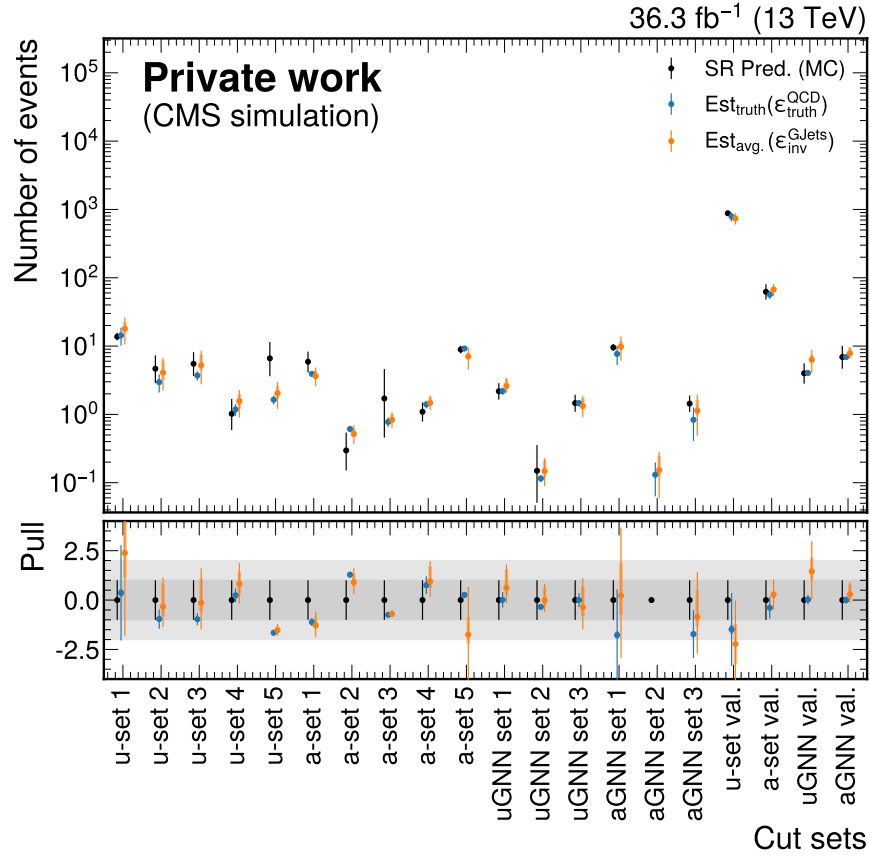


Figure 5.17: Graphical summary of the MC background estimation validation for all cut sets evaluated using QCD MC samples with pixel Phase-0 geometry. The thick error bars of each data point correspond to the statistical uncertainties. The thin error bars correspond to the total uncertainties, obtained by summing in quadrature all uncertainty sources. The entries in this plot correspond to the values found in Table 5.2.

Table 5.3: Comparison of the predicted and estimated results using the SR/CR pair constructed using QCD MC samples with pixel Phase-1 geometry. The number of SM events that pass the signal region selection (with statistical uncertainties) is listed in column 2. The estimation results with scale factors calculated using mistag rates evaluated using the QCD MC samples (with statistical and systematic uncertainties) are listed in column 3. The estimation result with scale factors calculated mistag rates evaluated using the γ +jets MC samples (with statistical and systematic uncertainties) is listed in column 4. The computation of uncertainties is further discussed in Section 6.1.2. Definition of the additional validation cut sets will be given in Section 5.4.2.

Cut set	Pred.	Est _{truth} ^{QCD} ($\epsilon_{\text{truth}}^{\text{QCD}}(p_T)$)	Est _{avg.} ^{QCD} ($\epsilon_{\text{inv.}}^{\text{GJets}}(p_T)$)
u-set 1	48.9 ^{+8.8} _{-7.5}	48.0 ^{+1.3} _{-1.3} ± 15.7	41.3 ^{+3.6} _{-3.2} ± 14.4
u-set 2	21.1 ^{+7.0} _{-5.4}	15.2 ^{+0.5} _{-0.5} ± 5.0	13.2 ^{+1.6} _{-1.4} ± 4.6
u-set 3	14.0 ^{+3.2} _{-2.7}	14.5 ^{+0.3} _{-0.3} ± 2.5	14.5 ^{+1.2} _{-0.9} ± 2.6
u-set 4	7.3 ^{+2.5} _{-1.9}	5.9 ^{+0.2} _{-0.2} ± 1.4	6.1 ^{+1.0} _{-1.0} ± 1.7
u-set 5	1.6 ^{+1.5} _{-0.9}	7.0 ^{+0.2} _{-0.2} ± 1.1	6.7 ^{+0.9} _{-0.8} ± 1.0
validation u-set	765.3 ^{+68.3} _{-62.9}	725.8 ^{+14.0} _{-14.2} ± 107.2	630.0 ^{+32.2} _{-32.9} ± 112.8
a-set 1	2.4 ^{+1.9} _{-1.1}	1.94 ^{+0.08} _{-0.08} ± 0.22	2.02 ^{+0.26} _{-0.19} ± 0.31
a-set 2	0.0 ^{+0.0} _{+0.0}	0.329 ^{+0.021} _{-0.021} ± 0.044	0.338 ^{+0.066} _{-0.040} ± 0.068
a-set 3	0.6 ^{+1.4} _{-0.5}	0.427 ^{+0.036} _{-0.035} ± 0.028	0.510 ^{+0.084} _{-0.059} ± 0.073
a-set 4	0.8 ^{+1.2} _{-0.6}	0.530 ^{+0.029} _{-0.028} ± 0.021	0.513 ^{+0.096} _{-0.053} ± 0.069
a-set 5	2.3 ^{+1.3} _{-0.9}	3.41 ^{+0.09} _{-0.09} ± 0.27	2.82 ^{+0.37} _{-0.22} ± 0.85
validation a-set	21.8 ^{+18.1} _{-10.8}	12.45 ^{+0.56} _{-0.52} ± 0.23	15.0 ^{+0.7} _{-0.7} ± 2.5
uGNN set 1	6.5 ^{+8.9} _{-4.3}	3.25 ^{+0.10} _{-0.10} ± 0.33	4.7 ^{+0.8} _{-0.5} ± 1.4
uGNN set 2	0.0 ^{+0.0} _{+0.0}	0.136 ^{+0.010} _{-0.009} ± 0.022	0.186 ^{+0.058} _{-0.029} ± 0.071
uGNN set 3	1.40 ^{+0.84} _{-0.56}	1.69 ^{+0.04} _{-0.04} ± 0.20	1.89 ^{+0.41} _{-0.21} ± 0.60
uGNN validation set	1.91 ^{+0.95} _{-0.66}	3.65 ^{+0.12} _{-0.11} ± 0.33	4.55 ^{+0.41} _{-0.27} ± 0.97
aGNN set 1	5.7 ^{+1.7} _{-1.3}	3.9 ^{+0.2} _{-0.2} ± 1.2	4.9 ^{+0.6} _{-0.5} ± 1.7
aGNN set 2	0.26 ^{+0.60} _{-0.22}	0.053 ^{+0.007} _{-0.006} ± 0.033	0.063 ^{+0.029} _{-0.015} ± 0.041
aGNN set 3	0.72 ^{+0.71} _{-0.40}	0.46 ^{+0.04} _{-0.04} ± 0.20	0.57 ^{+0.19} _{-0.09} ± 0.26
aGNN validation set	1.41 ^{+0.85} _{-0.56}	3.29 ^{+0.17} _{-0.16} ± 0.29	5.0 ^{+1.0} _{-0.7} ± 1.6

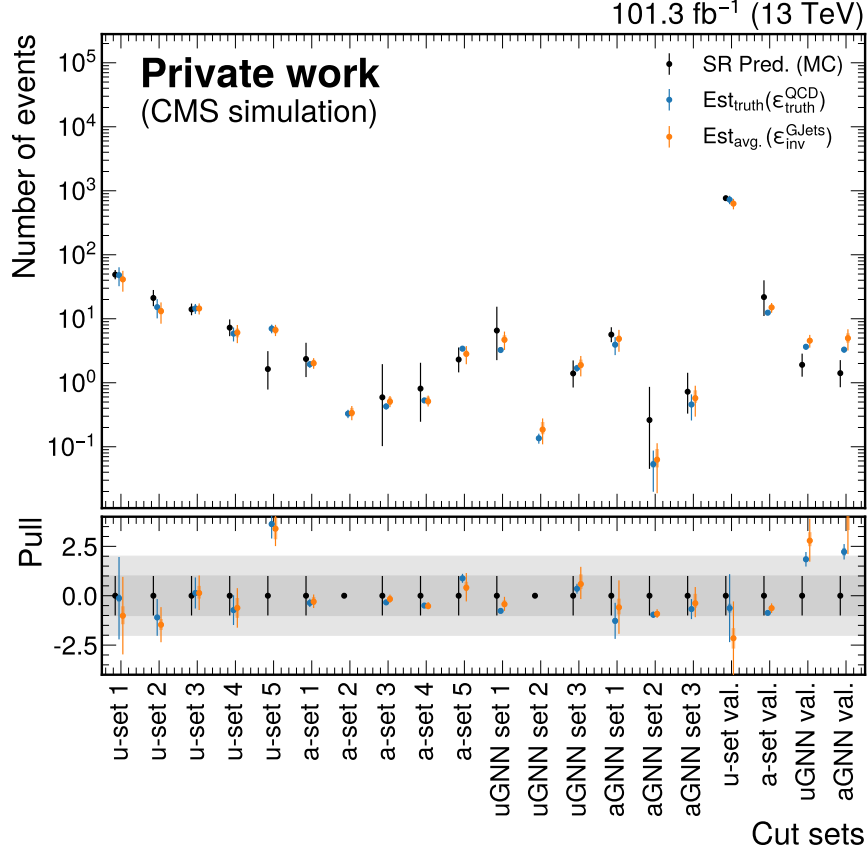


Figure 5.18: Graphical summary of the MC background estimation validation for all cut sets evaluated using QCD MC samples with pixel Phase-1 geometry. The thick error bars of each data point correspond to the statistical uncertainties. The thin error bars correspond to the total uncertainties, obtained by summing in quadrature all uncertainty sources. The entries in this plot correspond to the values found in Table 5.3.

5.4.2 Validation test using data

To check that the validity of the SR/CR comparison obtained in MC is not a result of the MC events being “too ideal” compared to data, we devise a method to perform the same validation test in data in an SR-like selection criteria region that is expected to be sufficiently signal-free.

For the model-agnostic EMJ tagging scheme, we can suppress signal events by inverting the event H_T selection requirements, requiring events to have H_T *below* a threshold value instead of above one. This exploits the fact that low mediator mass signal models have already been

excluded by the 2016 analysis. To further dilute signal events that can contaminate the results, we take the EMJ tagging designed for the most prompt-like signal model, which is expected to have the highest mistag rate, and further relax one of the selection cuts. This ensures that events that pass the SR-like selection criteria defined by the 2-tagged jet requirements are overwhelmingly occupied by SM events. The exact event and jet cut values for performing validation test on data are given in Tables 4.1–4.4. Using QCD MC samples for background yield estimation, signal events can at most account for $\sim 0.2\%$ of all events that enter the SR-like region defined for validation tests.

For the GNN-based EMJ tagging scheme, we can suppress signal events by defining a jet tagger that selects jets with output score in the side-band region. Signal samples can further be diluted by fully relaxing event H_T and jet p_T selection requirements. The exact event and jet cut values for performing validation test on data are given in Tables 4.3–4.4. Using QCD MC samples for background yield estimation, signal events can at most account for $\sim 0.3\%$ of all events that enter the SR-like region defined for validation tests.

Having evaluated the possibility of signal contamination of the validation selection criteria using MC events, we can now safely run the validation test selection with the accompanying background estimation calculation using an SR/CR-like pair constructed using the H_T -triggered data set and mistag rate evaluated in the γ -triggered data stream. The numerical results are summarized in Table 5.4 and the graphical results are summarized in Fig. 5.19. Here, we also see that the number of events passing the SR-like selection criteria is consistent with the estimation calculation, indicating that the background estimation is correct and robust when used on data.

Table 5.4: Number of events that pass the selection requirement defined by the validation cut sets, the number of events estimated using the scale factor computation using the H_T -triggered data stream with mistag rates evaluated in the γ -triggered data stream. For the background estimation, the relative uncertainties are evaluated using the uncertainties obtained for Tables 5.2–5.3. The uncertainties will be further discussed in Section 6.1.2.

	SR events	$\text{Est}_{\text{avg.}}^{H_T}(\epsilon_{\text{inv.}}^\gamma(p_T))$
validation u-set (Phase 0)	945	$798.3^{+57.7}_{-58.8} \pm 132.8$
validation a-set (Phase 0)	97	$60.9^{+3.5}_{-3.2} \pm 12.3$
uGNN validation set (Phase 0)	12	$16.2^{+5.7}_{-2.8} \pm 5.6$
aGNN validation set (Phase 0)	21	$16.3^{+3.1}_{-1.3} \pm 2.2$
validation u-set (Phase I)	539	$420.7^{+20.7}_{-18.8} \pm 75.3$
validation a-set (Phase I)	21	$18.8^{+1.0}_{-0.9} \pm 3.1$
uGNN validation set (Phase I)	9	$13.1^{+2.1}_{-1.1} \pm 2.8$
aGNN validation set (Phase I)	16	$12.8^{+3.7}_{-1.8} \pm 4.2$

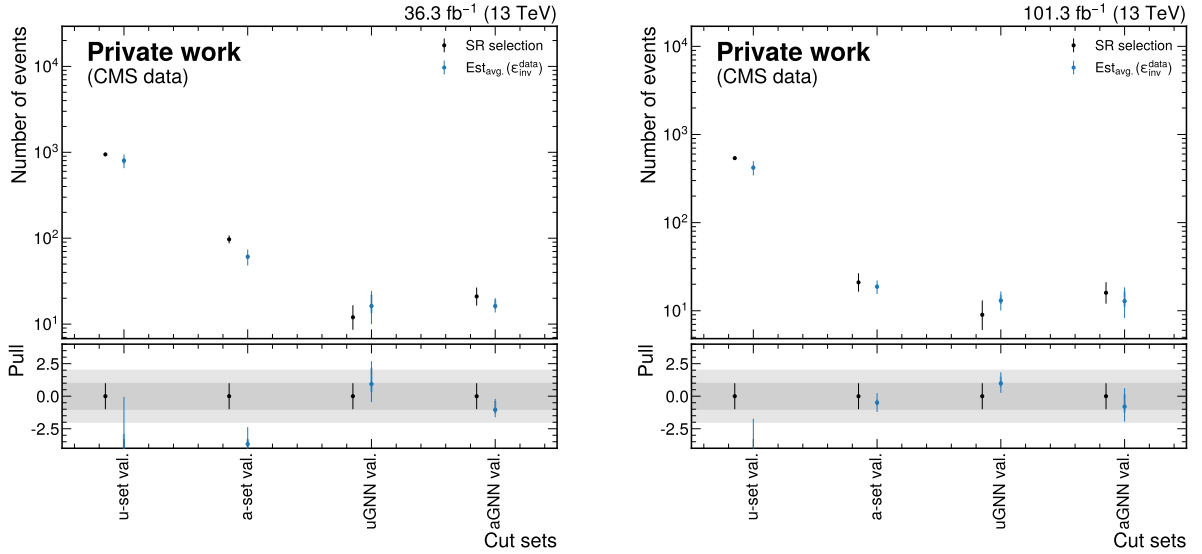


Figure 5.19: Results of the validation test cut sets calculations evaluated using H_T -triggered data set and mistag rates evaluated in γ -triggered data evaluated for the pixel Phase-0 (1) geometry on the left (right) with the validation cut sets. The numerical values used to generate these plots are given in Table 5.4.

Chapter 6: Uncertainties

Before we can draw any conclusions from the data, we must first evaluate potential systematic errors that may be introduced in the various calculation steps. We must also assess the uncertainties in the results due to finite information available in data and simulation. In this chapter, we will detail how to assign uncertainties to the routines used for standard model (SM) background estimations and uncertainties in the estimated signal acceptance.

Throughout this analysis, we have been using simulated samples as guiding tools to design and validate analysis strategies. However, we need to scrutinize the correctness of the simulation carefully to correctly identify sources of uncertainty. In this chapter, we outline the strategy for studying data–simulation discrepancies relevant to the analysis and assign an appropriate uncertainty in the final estimation results.

We can broadly split the source of data–simulation discrepancies into the following two categories:

1. **Mismodeling of the underlying physics.** This includes the mismodeling of the hard-scatter process, such as imperfect values of the various physics parameters involved, and not fully considering all possible physics processes that can enter the signal selection process. As fully modeling the underlying physics is effectively impossible, we will need to assess the possibility of whether the data/simulation discrepancies that arise from im-

perfect physics modeling can introduce biases or unexpected correlations in the primary observables used in the analysis.

2. **Mismodeling of the detector geometry and response.** The detector configuration used for simulated events is idealized when compared with data, resulting in discrepancies in the distributions of final reconstructed objects. If such effects are expected to impact the primary observables used for the analysis, appropriate corrections or additional randomness within their associated uncertainties is applied to the observable in simulated events to account for such discrepancies.

The difficulty of obtaining a perfect data–simulation agreement for all relevant physics objects was the primary motivation for using the fully data-based calculation, as the EMJ signatures heavily rely on the correctness of the CMS tracker system. The expected primary SM background, jet production through SM QCD processes, is also arduous to model correctly. The mapping of data to data through the scale-factor method to estimate background, as presented in Chapter 5, allows many data–simulation discrepancies to be omitted as mismodeling effects never enter the calculation. The tradeoff for using data-based methods is that underlying assumptions of physics similarities between the signal region (SR) and the multiple control regions (CR) must also be properly assessed for potential biases. Section 6.1 will discuss the biases that may enter the SM background estimation calculation and a method for assigning an appropriate uncertainty value.

Regardless of how robust the data-based estimation is, we will still need to carefully assess the correctness of CMS simulation and how it can potentially impact signal acceptance for the surveyed signal models. Section 6.2 will discuss how we evaluate simulation correctness as well

as how the uncertainties in the final signal yield are estimated.

6.1 Background estimation uncertainties

6.1.1 Propagation of statistical uncertainty

All elements used in calculating the SR/CR scale factor in Eq. 5.8 are calculated using a finite jet population and thus have statistical uncertainties. In particular, the calculation of the flavor dependence of mistag rates using a linear inversion (Eq. 5.3) also introduces correlations in the mistag rates used in Eq. 5.8. This analysis uses a Monte Carlo-based random sampling approach to propagate the statistical uncertainties through the calculation routines.

First, we observe that all numerical elements that enter Eq. 5.8, $\epsilon^X(p_T)$ and $B^X(p_T)$, are evaluated using statistically independent jet populations, and thus can be assumed to be perfectly uncorrelated. The evaluation of $\epsilon^X(p_T)$ will have a binomial uncertainty, summarized using the Clopper–Pearson interval [108], while the evaluation of $B^X(p_T)$ will have uncertainties evaluated from the minimum- χ^2 fit described in Eq. 5.4. For each evaluation of $\epsilon^X(p_T)$ and $B^X(p_T)$, we can generate a set of randomized values that follow the bifurcated-normal distribution, $\mathcal{B}(x; \mu, \sigma^\pm)$:

$$\mathcal{B}(x; \mu, \sigma^+, \sigma^-) = \begin{cases} \mathcal{N}(x; \mu, \sigma^+) & x \geq \mu \\ \mathcal{N}(x; \mu, \sigma^-) & x < \mu \end{cases}, \quad (6.1)$$

with the asymmetric Gaussian widths $\sigma^\pm$ following the evaluated uncertainties of the statistical calculations. Let us index the set of randomized values using the index λ , so now our expressions now involve the sets $\{\epsilon_\lambda^X(p_T)\}$ and $\{B_\lambda^X(p_T)\}$. We will reserve $\lambda = 0$ for the “central value,”

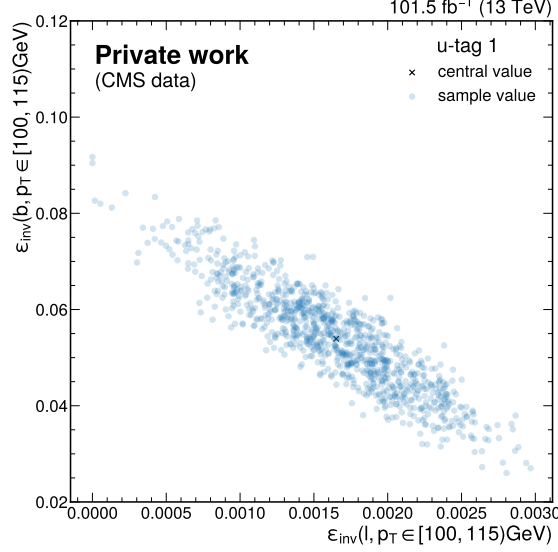


Figure 6.1: The 2-dimensional distribution of the calculated $\epsilon(l, p_T)$ and $\epsilon(b, p_T)$ values using the common set of $\{\epsilon_\lambda^X(p_T)\}$ and $\{B_\lambda^X(p_T)\}$ obtained using jets in the γ -triggered data set with pixel Phase-1 geometry, for the “u-tag 1” EMJ tagger with jet $p_T \in [100, 115)$ GeV.

which is calculated without any randomization. Using this, we evaluate the flavor-dependent mistag rate using Eq. 5.3 simply by expanding the results with the additional λ index. The obtained set of mistag rates, $\{\epsilon_\lambda(b, p_T)\}$ and $\{\epsilon_\lambda(l, p_T)\}$, track the statistical correlation by the results of the λ index. An example of the statistical correlation between $\epsilon(b, p_T)$ and $\epsilon(l, p_T)$ observed when using this random sampling method can be found in Fig. 6.1.

The set of values $\{\epsilon_\lambda(f, p_T)\}$ can then be used to calculate the scale factor using Eq. 5.8, where we also get a set of values $\{SF_\lambda\}$ for each event, as well as a set of estimation results $\{E_\lambda = \sum_{\text{event} \in \text{CR}} SF_\lambda\}$. Since B^{CR} is obtained from yet another statistically independent set of jets, we can also generate a set of randomized $\{B_\lambda^{CR}\}$ that follow distribution $\mathcal{B}$ with the fit uncertainty in the evaluation of $\{SF_{\text{avg}, \lambda}\}$ for the flavor-averaged method of flavor assignment. We can then take the one-sided standard deviation of $\{E_\lambda\}$ compared with E_0 to be the statistical

uncertainty associated with the mistag rate computation routine:

$$\bar{\sigma}_E^+ = \frac{1}{N(\{E_\lambda | E_\lambda > E_0\})} \sqrt{\sum_{E_\lambda > E_0} (E_\lambda - E_0)^2} ; \quad (6.2)$$

$$\bar{\sigma}_E^- = \frac{1}{N(\{E_\lambda | E_\lambda < E_0\})} \sqrt{\sum_{E_\lambda < E_0} (E_\lambda - E_0)^2} ; \quad (6.3)$$

where $N(\{x_\lambda\})$ counts the number of elements in set $\{x_i\}$. The results of $\bar{\sigma}_E^\pm$ are then summed in quadrature to the scaled Garwood interval [109] to account for the Poisson uncertainty associated with the finite number of events in the CR used to perform the sum to obtain E . This final result is what is presented in Tables 5.2–5.4 as the asymmetric statistical uncertainty of the background estimation.

Let us consider some of the properties of the statistical uncertainty obtained using the random sampling approach described above:

- As λ is a separate index from the binning variables p_T , these calculations assume that the statistical uncertainty is uncorrelated between different p_T bins. This is a required feature, as the value of $\epsilon_\lambda^X(p_T)$ is calculated using statistically independent jet populations for different p_T bins.
- The evaluation of each SF_λ and E_λ simply performs the scale factor calculation multiple times for different λ values. This means that the calculations assume that the statistical uncertainties are correlated in the p_T bin across all events. This is also a required feature, as the mistag rate calculated with jets in some p_T bin in the signal-free region is used for *all* jets in the CR that fall in the same p_T bin.
- Regarding statistical correlations across eras, notice that the computation of $\epsilon^X(p_T)$ and by

extension $\epsilon_\lambda^X(p_T)$ requires a defined parent population to proceed. Since ϵ_λ is generated by a random number generator, this necessarily means that the statistical uncertainty of mistag rates calculated between different data collection eras are uncorrelated, regardless of how many eras the final sum runs over.

The points above indicate that the random sampling approach for treating the statistical uncertainty contains all the correlation properties expected. As a final note of implementation details, the index λ runs from 0 to 200, which is sufficient to give a smaller than 10% precision of the statistical uncertainty.

6.1.2 Uncertainties associated with the calculation assumption

Borrowing the notation described in Eq. 5.14 that is used to present the various background estimations results by swapping out calculation ingredients, let us take a look at which computation ingredients should be used in the estimation calculation to yield the mathematically correct background estimation result. The “true” background estimation result that should be used for background extraction should ideally be:

$$\mathcal{E}_0 = \text{Est}_{\text{truth}}^{H_T}(\epsilon_{\text{truth}}^{H_T|\text{SM}}(\vec{\theta}_\infty)), \quad (6.4)$$

where the estimation is carried out over CR constructed using the H_T -triggered data set (Est^{H_T}), the mistag rate is evaluated on SM jets in the same data set ($\epsilon^{H_T|\text{SM}}$), the mistag rate is binned with arbitrarily fine variable dependence ($\vec{\theta}_\infty$), and the flavor evaluation uses the “true” jet flavor. This calculation, however, is clearly not possible to perform. What we can actually calculate in data will need multiple approximations:

- $\mathcal{E}_1 = \text{Est}_{\text{truth}}^{H_T}(\epsilon_{\text{truth}}^{H_T|\text{SM}}(p_T))$: Given that mistag rates are calculated using a finite number of jets, we can only parameterized the mistag rate in some finite number of jet variables. Following the choice from Section 5.3, we will concretely choose the estimation result obtained when binning the mistag rate in jet p_T .
- $\mathcal{E}_2 = \text{Est}_{\text{truth}}^{H_T}(\epsilon_{\text{truth}}^\gamma(p_T))$: Since there is no reliable way to extract just the SM jets in the data set, the signal-free region constructed using the γ -triggered data set is used for mistag rate evaluation.
- $\mathcal{E}_3 = \text{Est}_{\text{avg.}}^{H_T}(\epsilon_{\text{inv.}}^\gamma(p_T))$: As flavor information is not directly accessible in data, the flavor-blind calculation routines introduced in Sections 5.1.2 and 5.2.2 need to be employed.

The value $\mathcal{E}_3$ is the only method that can actually be calculated using data, and each step from $\mathcal{E}_i$ to $\mathcal{E}_{i+1}$ potentially introduces systematic errors into the calculation:

- $\mathcal{E}_0 \rightarrow \mathcal{E}_1$ assumes that p_T captures all the mistag rate correlation that occurs for jets within the same event. While the arguments for p_T being the variable that best captures mistagging correlations are already presented in Section 5.3, some residual correlations are expected to exist and needs to be accounted for.
- $\mathcal{E}_1 \rightarrow \mathcal{E}_2$ assumes that kinematics of jets in the H_T -triggered data set and the γ -triggered data set are the same. The magnitude of mistag rate differences driven by jet population differences has already been demonstrated in Section 5.1.4.4, and needs to be accounted for in the final estimation calculation.
- $\mathcal{E}_2 \rightarrow \mathcal{E}_3$ assumes that the flavor-blind calculations are correct, i.e., the underlying mistag rate is identical between the signal-free regions regardless of the selection imposed by the

extra jet in the event.

Each of these assumptions should have an associated uncertainty to cover systematic error that can enter the calculation.

While we can only calculate $\mathcal{E}_3$ using data, we can calculate an SM MC counterpart E_i to each $\mathcal{E}_i$ by swapping out the H_T -triggered and γ -triggered data sets with the most dominant SM MC counterpart samples: QCD and γ +Jets. Effectively, E_i is obtained by swapping out the sample indices of $\mathcal{E}_i$:

$$E_i \xleftarrow{H_T \rightarrow \text{QCD}, \gamma \rightarrow \text{GJets}} \mathcal{E}_i. \quad (6.5)$$

Using E_i , let us define the relative uncertainty measure:

$$U_i = u(E_i, E_{i+1}) = \frac{2|E_i - E_{i+1}|}{E_i + E_{i+1}}, \quad (6.6)$$

then apply the relative uncertainties U_i to the final calculated results $\mathcal{E}_3$. One value that is not calculable, even using MC, is $E_0 = \text{Est}_{\text{truth}}^{\text{QCD}}(\epsilon_{\text{truth}}^{\text{QCD}}(\vec{\theta}_{\infty}))$, as we only have a finite number of jets available for calculation. Instead of calculating U_0 directly, the value E_1 can be compared with $E'_1 = \text{Est}_{\text{truth}}^{\text{QCD}}(\epsilon_{\text{truth}}^{\text{QCD}}(p_{T,\text{fine}}))$, which also uses mistag rates evaluated with finer p_T binning, and $E''_1 = \text{Est}_{\text{truth}}^{\text{QCD}}(\epsilon_{\text{truth}}^{\text{QCD}}(n_{\text{tracks}}))$, which compares the primary calculation results with an alternate binning variable, n_{tracks} . The uncertainty term U_0 can then be estimated using the extra terms:

$$U_0 \sim \sqrt{U_0^2 + U_0''^2} = \sqrt{u(E_1, E'_1)^2 + u(E_1, E''_1)}. \quad (6.7)$$

For better readability, let us give the uncertainty terms U_i a readable index according to the assumption taken to approach the data-compatible calculation:

- U_0 corresponds to the uncertainty associated with the mistag rate parameterization, so U_0 with re-labeled using the explicit sub-index $U_0 \Rightarrow U_{\text{param.}}$. The term $U_{\text{param.}}$ can be subdivided into the bin width component $U'_0 \Rightarrow U_{\text{width}}$ and the variable choice component $U''_0 \Rightarrow U_{\text{var.}}$.
- U_1 corresponds to the assumption that jets variable distributions in the γ -triggered data set are sufficiently similar to the jets in the H_T -triggered data set, given the differences in the final state objects. This uncertainty term will be re-labeled as $U_1 \Rightarrow U_{\text{fin.}}$.
- U_2 corresponds to the assumptions related to flavor-related calculations. This uncertainty term will be relabeled as $U_2 \Rightarrow U_{\text{flav.}}$.

A diagram summarizing the steps in assumptions required to arrive at the data-compatible calculation $\mathcal{E}_3$, as well as the various uncertainty components using MC estimation counterparts can be found in Fig. 6.2.

As the mistag rate is calculated separately for Phase-0 and Phase-1 pixel geometry, the uncertainties listed above will also be calculated separately for the two pixel geometries. The values of the uncertainties can be found in Tables 6.1–6.2. Visually the scatter of the various estimation results calculated using MC with various background estimation ingredients can be found in Fig. 6.3.

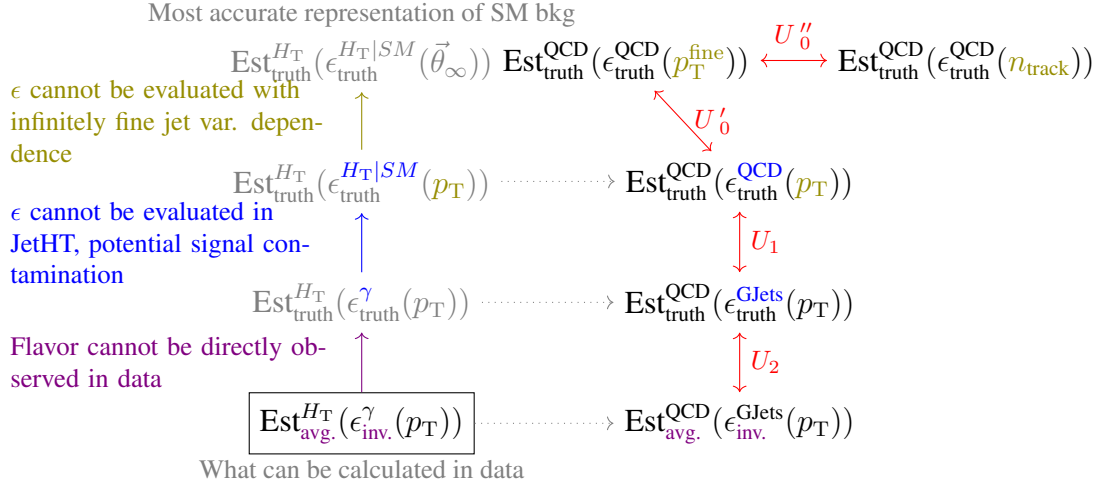


Figure 6.2: Summary of the various assumptions that need to take place for the final data-driven calculation results ($\mathcal{E}_3 = \text{Est}_{\text{avg.}}^{H_T}(\epsilon_{\text{inv.}}^{\gamma}(p_T))$) to be mapped to a mathematically correct calculations routine, and the MC counterparts used to evaluate the uncertainty associated with the assumptions made for alternate calculations.

Table 6.1: Comparison of the various background estimation uncertainties evaluated using the control region events in QCD MC samples with pixel Phase-0 geometry (in %).

Cut set	$U_{\text{var.}}$	U_{width}	$U_{\text{fin.}}$	$U_{\text{flav.}}$
u-set 1	29.50	0.35	3.63	25.61
u-set 2	29.37	0.20	5.15	26.73
u-set 3	15.01	0.04	7.28	41.15
u-set 4	19.36	1.35	2.12	30.26
u-set 5	14.25	0.45	11.24	32.71
u-set validation	14.62	< 0.01	7.77	1.66
a-set 1	8.02	3.48	21.85	14.16
a-set 2	4.41	0.88	24.96	8.60
a-set 3	14.85	1.85	6.47	13.54
a-set 4	11.50	5.46	6.95	12.34
a-set 5	0.54	0.48	33.60	6.78
a-set validation	13.79	0.33	2.49	14.49
uGNN set 1	2.83	11.60	5.23	12.30
uGNN set 2	6.12	7.21	7.49	32.42
uGNN set 3	3.14	10.52	22.93	13.56
uGNN validation set	5.65	7.04	29.99	14.59
aGNN set 1	2.72	31.10	13.39	10.82
aGNN set 2	11.13	50.26	9.43	24.83
aGNN set 3	1.54	50.92	15.06	16.04
aGNN validation set	3.57	7.15	11.15	1.25

Table 6.2: Comparison of the various background estimation uncertainties evaluated using the control region events in QCD MC samples with pixel Phase-1 geometry (in %).

Cut set	$U_{\text{var.}}$	U_{width}	$U_{\text{fin.}}$	$U_{\text{flav.}}$
u-set 1	32.67	0.36	11.85	3.20
u-set 2	33.05	0.46	9.88	4.30
u-set 3	17.50	0.09	3.59	3.55
u-set 4	24.44	0.15	6.62	10.28
u-set 5	15.14	0.27	2.73	2.27
u-set validation	14.77	0.01	6.10	8.06
a-set 1	11.02	2.93	5.26	8.89
a-set 2	13.44	0.14	8.94	11.74
a-set 3	6.48	1.56	7.42	10.18
a-set 4	0.52	3.96	10.55	7.30
a-set 5	7.73	1.83	27.59	8.85
a-set validation	1.84	0.34	2.34	16.46
uGNN set 1	5.21	8.67	12.55	24.14
uGNN set 2	4.62	15.62	3.39	34.28
uGNN set 3	2.05	11.88	14.55	25.52
uGNN validation set	0.34	9.01	19.20	2.62
aGNN set 1	6.58	30.21	18.09	2.93
aGNN set 2	17.94	59.24	5.37	21.60
aGNN set 3	7.39	42.38	11.43	11.33
aGNN validation set	2.59	8.34	12.20	28.83

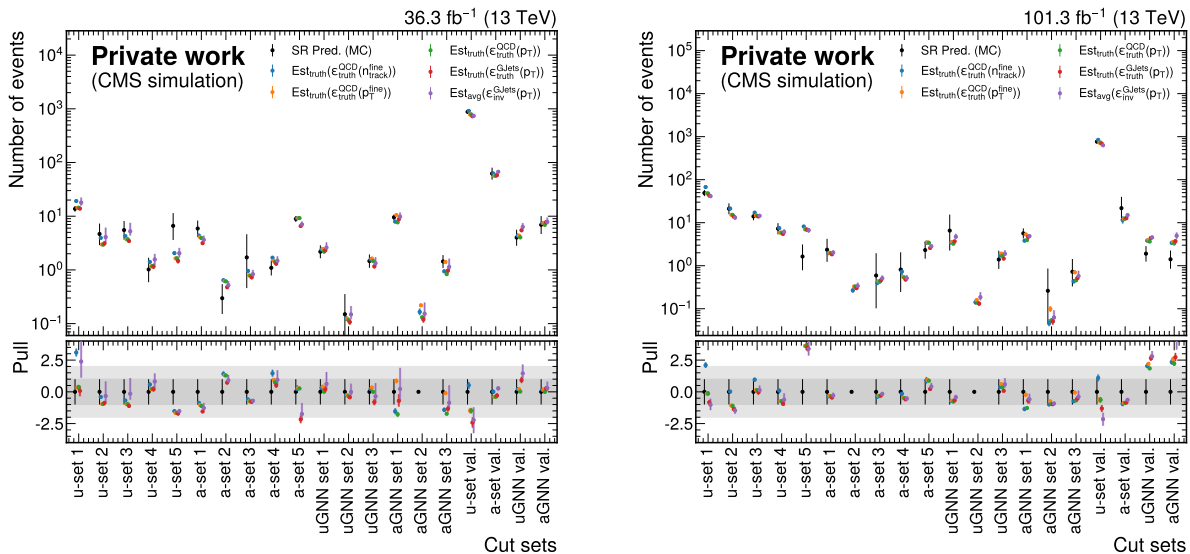


Figure 6.3: Scatter of different background estimation results using different mistag rates sources, parameterization, and flavor information used to estimate the various sources of background estimation uncertainties evaluated using QCD MC samples with pixel Phase-0 (Phase-1) geometry on the left (right). Only the statistical uncertainties are displayed in these plots.

6.2 Signal selection uncertainties

As excess events that agree with a signal hypothesis have yet to be observed, the uncertainties associated with signal samples primarily consist of understanding how data/MC discrepancies of the used standard physics objects may impact the signal selection, how corrections may be performed as well as assigning an appropriate uncertainty to these corrections. Broadly speaking, such corrections can be of two types, which will be discussed separately in Section [6.2.1](#) and [6.2.2](#) respectively.

6.2.1 Reconstruction kinematic variable corrections

These are corrections applied directly to kinematic variables of physics objects directly used for event selection. Discrepancies between data and MC are corrected on a per-object basis, with some uncertainties in the correction factors. These uncertainties are then propagated to the signal acceptance by varying the kinematic variables within the correction uncertainties, re-running the event selection chain, and comparing the total number of signal events expected to enter the signal region. Two corrections of this type are taken into consideration:

6.2.1.1 Jet energy correction and jet energy resolution

The p_T , η dependent jet energy correction (JEC) and resolution (JER), which we have described in Section [3.3.1.2](#). To estimate the impact of JEC and JER on signal acceptance, we scale the jet p_T and energy, and by extension the H_T of the event within one standard deviation of the uncertainties of JEC and by one standard deviation in JER before any selection criteria are imposed, then rerun the event selection chain with the shifted jet kinematics. For simplicity,

we assume that JEC/JER shifts are perfectly correlated between each jet, while JEC and JER are perfectly uncorrelated.

6.2.1.2 Track modelling uncertainties

The performance of the tracking and vertexing algorithm is very sensitive to “randomness” in the tracker system, typically resulting in an overestimation of tracking performance in MC events compared with data. The primary MC tracks variables, the displacement measures, IP_{2D} and IP_z , should then be corrected to account for this. This is the only non-standard correction that is evaluated using recipes specific to this analysis. Two effects will be discussed here:

- **Displacement resolution:** When discussing prompt-like tracks with $IP_{2D} < 10^{-1}$ cm, we will assume that the track reconstruction efficiency for tracks of interest is effectively 100% for both data and MC. The discrepancy between the distribution of the track displacement measures seen in data and MC is then primarily driven by the resolution of the tracking and vertexing algorithm performing differently between data and MC. The correction that we will apply consists of the addition of a random “smearing” value to the track displacement measures. The difference between the selection results with and without displacements smearing will be taken as the uncertainty associated with track resolution effects.

The smearing factor s is obtained by fitting the distribution of the track displacement measures within $|IP_{2D}| < 50\mu\text{m}$ and $|IP_z| < 50\mu\text{m}$ in data and MC to a normal distribution $\mathcal{N}(0, \sigma)$. The smearing factor is then determined for the track displacement variable as $s = (\sigma_{\text{data}}^2 - \sigma_{\text{MC}}^2)^{1/2}$. The value of s is determined separately for 2016 pre-APV fixes, 2016 post-APV fixes, 2017 and 2018 to account for major geometry and detector config-

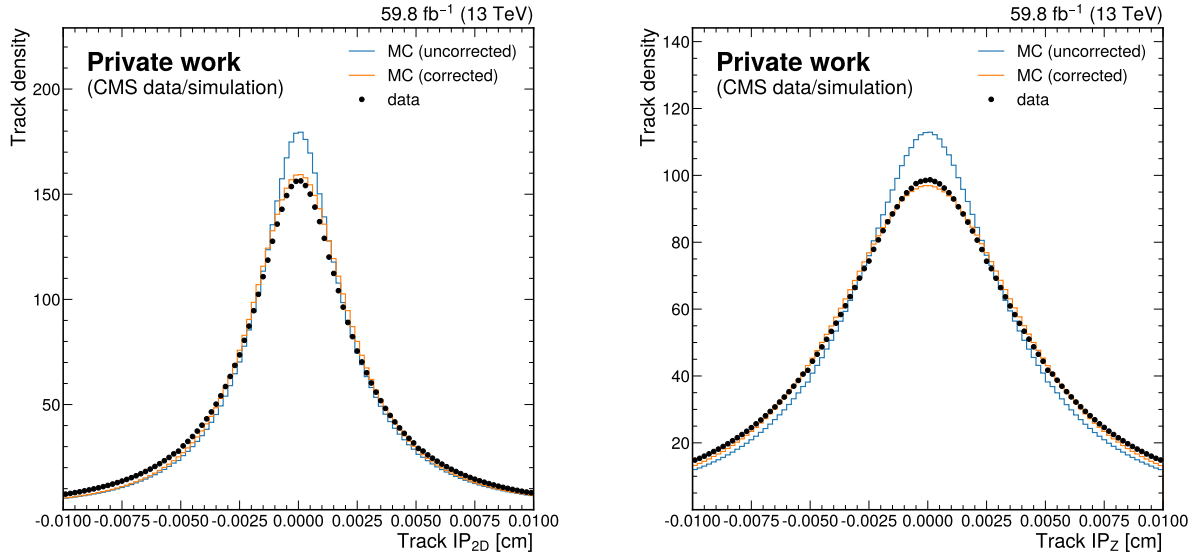


Figure 6.4: Distribution of the track IP_{2D} (IP_z) distribution on the left (right) plot using tracks in H_T -triggered data collected in 2018 (points), and MC with/without the smearing correction (orange/blue).

uration changes. The data and MC samples used to obtain the displacement distribution consist of the H_T -triggered data stream and QCD MC samples with only the H_T trigger requirements imposed. An example of how this smearing correction match the track displacement measured in data and MC can be found in Fig. 6.4

- **Reconstruction efficiency of displaced tracks:** The reconstruction efficiency of tracks falls off fast for tracks originating from vertices outside the beamline at $\rho \sim 2$ cm. Exactly how fast this efficiency falls off may be different between data and MC, and may impact signal selection efficiency, given that the primary signature is a search for displaced tracks. Tracking efficiency, however, is very difficult to evaluate in data, an example of evaluating tracking efficiency can be seen in Ref. [76], and is not plausible to be fully evaluated for the events of interest. Also, the EMJ tagging algorithm is designed to work with b-like displacement measures to avoid the signal selection be overly dependent on extremely outlying tracks that are known to be difficult to model. For this type of data/MC discrepancy,

we only check if differences in reconstruction efficiency might cause catastrophic changes in the final signal acceptance.

To obtain an upper limit of the track reconstruction discrepancies without looking at the primary signal region, we will compare the distribution of track displacement measures of tracks in the γ -triggered data set and the γ +Jets MC samples. We make the assumption that tracks with displacement measure $|\text{IP}_{2\text{D}}| < 10^{-1}$ cm have effectively 100% reconstruction efficiency for both data and MC. Using this assumption, we normalized the number of tracks with $|\text{IP}_{2\text{D}}| < 10^{-1}$ cm to the same value for data and MC. After this, we assume that all differences in the track displacement measure seen for $|\text{IP}_{2\text{D}}| > 10^{-1}$ cm are entirely driven by differences in track reconstruction efficiency. An example of the calculated track reconstruction efficiency differences using these assumptions can be found in Fig. 6.5. Notice that the evaluated efficiency differences are expected to be much larger than the true difference between reconstruction efficiencies, as we are explicitly ignoring effects such as missing hard-scattering physics. We also see that the evaluated efficiency differences are much larger than the typical discrepancies seen in Ref. [76].

Tracks can then be weighted according to the reconstruction efficiency difference, and weighted versions of the track variables will be used to calculate the jet-tagging variables described in Section 4.2. The weighted jet-tagging variables are then used for event selection, and the selection efficiency with and without the weighting can be compared to arrive at the potential impact of track reconstruction efficiency differences on signal acceptance. The estimated maximum difference in signal selection efficiency using the weighted tagging variables was found to be $\sim 2\%$, with the average difference being $< 0.3\%$, much

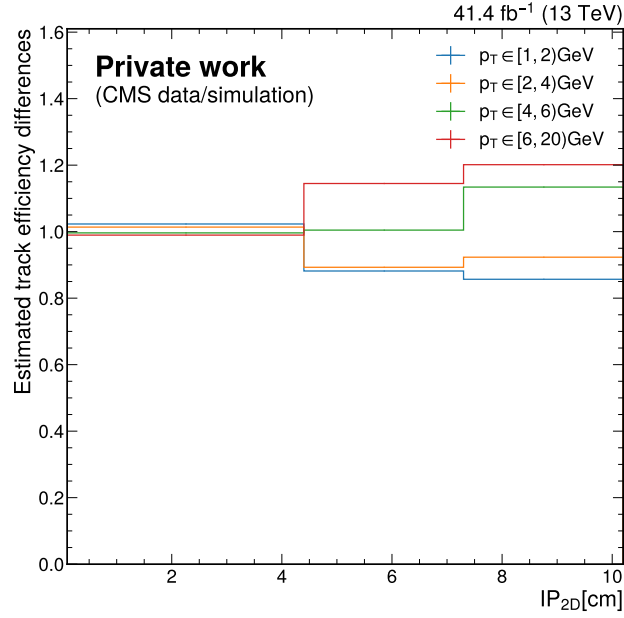


Figure 6.5: Crudely estimated track reconstruction efficiency differences evaluated as a function of track $|IP_{2D}|$ and p_T in data and MC events evaluated using the γ -triggered data set collected in 2017 and the corresponding γ +Jets MC samples.

smaller than the expected difference caused by the track resolution effects (see Tables 6.3 and 6.4).

As the described method of obtaining data/MC track reconstruction efficiency difference is excessively overestimated, with even the worst-case scenario having a small impact on signal acceptance, no corrections will be applied to address potential data/MC discrepancies arising from reconstruction efficiency.

6.2.2 Event weighting corrections

The remaining corrections to the data/MC discrepancies are resolved by weighting the MC events such that high-level features of MC simulation are corrected to match the data distribution. Event weights calculated for such corrections also include uncertainties, either from finite populations or alternate MC models, so uncertainties are associated with the weights applied. Below

are the weight corrections applied to the signal events with their associated uncertainties.

6.2.2.1 Luminosity

For the calculation of the expected event yields, all events in the MC events of a signal model have been weighted with the factor $(\mathcal{L}\sigma)/N$, where N is the total number of events available for the signal model of interest, σ is the theoretical cross section of the process of interest, and $\mathcal{L}$ is the integrated luminosity of the CMS data set corresponding to the sample of interest. The luminosity and its uncertainty are evaluated for each data collection campaign [55, 56, 57], and the uncertainty in the final event yield is calculated by varying the value of $\mathcal{L}$ within the quoted uncertainty. The uncertainty is 1.2% for 2016, 2.3% for 2017, and 2.5% for 2018.

6.2.2.2 Trigger efficiency

The trigger efficiency scale factor discussed in Section 3.5.2.1 also has its uncertainties. This analysis only considers the statistical uncertainty associated with the finite number of events used to evaluate Eq. 3.15–3.16. The impact on the signal selection efficiency is estimated by varying the trigger scale factor by one standard deviation within its statistical uncertainty by comparing the number of signal events passing the selection. Following the evaluation of the H_T trigger efficiencies given in Section 4.1.1, the uncertainty is considered uncorrelated for different data collection eras.

6.2.2.3 Pileup reweighting

The pileup reweighting scheme described in Section 3.5.2.2 depends on the inclusive pp scattering cross section $\sigma_{\text{min. bias}}$ used to calculate the pileup distribution in data. This cross section is only known within an uncertainty of $\pm 4.6\%$, so alternate pileup weights are computed by varying $\sigma_{\text{min. bias}}$ within the one standard deviation of uncertainty, and calculate the expected number of events with altered pileup weighting schemes.

6.2.2.4 Parton distribution function variation and matrix element scaling

The generation of the signal process relies on some assumption of theoretical parameters, most notably the parton distribution function (PDF) of the proton used to generate the hard scattering partons, the scale of the QCD interaction strength α_S , and the fragmentation scale used to calculate the parton shower α_F . While the generator executes event generation assuming some concrete theoretical parameter values, it also includes various event weights representing the relative probability of the identical event occurring assuming alternate theoretical parameter values. Let us label these alternate weights w_{ij} , where the index i is used to index the event, and j is used to index the variable alterations.

The relative probability will also include the impact of alternate theoretical values on the theoretical cross section of the process of interest. Effects on the theoretical cross section should be factored out, as such effects are expected to be independent of the overall event selection flow.

To factor out the such events, we calculated a normalized weight $\bar{w}_{ij}$, defined as:

$$\bar{w}_{ij} = w_{ij} \frac{\sum_{i \in \text{all events}} w_{i0}}{\sum_{i \in \text{all events}} w_{ij}}, \quad (6.8)$$

where the summation runs over all events before any selection is performed. The definition of $\bar{w}_{ij}$ ensures that the relation $\sum_{i \in \text{all events}} \bar{w}_{ij} = \sum_{i \in \text{all events}} w_{i0}$ for all possible variables j when compared with some references theoretical parameter variation $j = 0$, effectively factoring out the impacts of shifting parameter values on the cross section of the process of interest. The uncertainty of theoretical parameters on the signal acceptance given our event selection choices can then be estimated by comparing the value of:

$$W_j = \sum_{i \in SR} \bar{w}_{ij}, \quad (6.9)$$

with the signal acceptance evaluated using the reference theoretical parameters W_0 .

For PDF uncertainties, the event generator Pythia calculates the relative probabilities of the generated events by comparing the PDF value within the NNPDF family [110]. The standard variation of $\{W_j\}$ is the conventional way of representing the uncertainty of the signal acceptance related to PDF uncertainties.

For the QCD scale and factorization uncertainty, Pythia calculates the relative probability of the generated events with α_S and α_F scaled up and down by a factor of 2 compared with the relative value, giving a total of 9 variations. The envelope of 7 variations (excluding the variation where α_S is scaled up by 2 and α_F is scaled down by 2, as well as the variation where α_S is scaled down by 2 and α_F is scaled up by 2) is used as the conventional way of representing the

Table 6.3: Mean and maximum relative uncertainty of the number of events passing the selection criteria using model-agnostic EMJ taggers for various surveyed signal (uncertainty value in %).

Uncertainty source	Unflavored		Flavor-aligned	
	mean	max.	mean	max.
MC track modeling	0.2	3.7	1.4	16.2
JEC	1.0	19.8	0.8	7.0
JER	0.3	7.6	0.3	3.7
Luminosity	1.8	2.5	1.8	2.5
Pileup re-weighting	1.6	8.7	1.4	10.4
Trigger efficiency	0.3	0.9	0.3	0.9
PDF Variation	< 0.1	< 0.1	< 0.1	< 0.1
Matrix element scale	< 0.1	< 0.1	< 0.1	< 0.1

uncertainty related with matrix element evaluation.

6.2.3 Summary of uncertainties on signal samples

As there are many signal samples, we will not exhaustively list all the uncertainties of each signal sample here, but we will provide a summary of the mean and maximum value of the relative uncertainties of various sources in Tables 6.3 and 6.4. Uncertainties for various sources are typically smaller than 5%, with cut sets with tighter selection thresholds raising the uncertainty to 20%. While the PDF and matrix element uncertainties are significantly smaller than all other uncertainty sources, they will still be included in the analysis, as this is a standard uncertainty for all analyses using MC events.

Table 6.4: Mean and maximum relative uncertainty of the number of events passing the selection criteria using GNN-based EMJ taggers for various surveyed signal (uncertainty value in %).

Uncertainty source	Unflavored sample		Flavored sample	
	mean	max.	mean	max.
MC track modeling	0.3	9.3	0.5	3.6
JEC	1.3	10.0	0.7	4.7
JER	0.2	6.6	0.2	0.8
Luminosity	1.8	2.5	1.8	2.5
Pileup re-weighting	0.9	6.9	1.0	3.8
Trigger efficiency	0.3	0.9	0.3	0.9
PDF Variation	< 0.1	< 0.1	< 0.1	< 0.1
Matrix element scale	< 0.1	< 0.1	< 0.1	< 0.1

Chapter 7: Results

Having carefully assessed the correctness of the background estimation methods in Chapter 5 and the potential uncertainties in the calculation in Chapter 6, we can now safely compare the number of events passing the signal region selections defined in Table 4.4 and the corresponding background estimation to see if there are excesses of events in the signal region. The results can be found in Tables 7.1 and 7.2, with the graphical results presented in Fig. 7.1. No significant excess in the number of events were observed in any signal detection selection criteria. This result is interpreted as an upper limit of the production cross section of the surveyed signal models.

7.1 The CL_s upper limit

The upper limit of the production cross section of a signal model is determined by comparing the likelihood of the number of observed events occurring given a background-only hypothesis $\mathcal{H}_b$ and a signal-plus-background hypothesis $\mathcal{H}_{s+b}$, with the expected number of events under the background-only hypothesis determined from the background estimation results.

The expected number of events that pass a signal selection criteria, n , can be written as the sum of the expected background events, b , and the expected signal events, s :

$$n = b(\vec{\nu}_{\text{bkg.}}) + \mu s(\vec{\nu}_{\text{sig.}}) \quad (7.1)$$

Table 7.1: The number of events that pass the signal selection criteria defined in Table 4.4, and the estimated number of events using the method discussed in Chapter 5 applied to the events in the control region of the H_T -triggered data set. Results obtained using 2016 data (pixel Phase-0 geometry) with $\mathcal{L} = 36.3 \text{ fb}^{-1}$.

Cut set	Observed Yield	Estimated background
u-set 1	21	$18.1^{+5.9}_{-2.7} \pm 7.1$
u-set 2	7	$6.0^{+3.7}_{-1.6} \pm 2.4$
u-set 3	6	$6.6^{+7.7}_{-1.6} \pm 2.9$
u-set 4	3	$1.95^{+0.89}_{-0.63} \pm 0.70$
u-set 5	6	$3.0^{+1.3}_{-1.1} \pm 1.1$
a-set 1	10	$3.9^{+1.5}_{-0.5} \pm 1.1$
a-set 2	2	$0.59^{+0.27}_{-0.10} \pm 0.16$
a-set 3	1	$0.83^{+0.32}_{-0.12} \pm 0.18$
a-set 4	1	$1.31^{+0.60}_{-0.24} \pm 0.25$
a-set 5	8	$4.6^{+1.4}_{-0.6} \pm 1.6$
uGNN set 1	6	$4.7^{+2.1}_{-0.6} \pm 0.8$
uGNN set 2	0	$0.20^{+0.17}_{-0.05} \pm 0.07$
uGNN set 3	2	$2.2^{+1.4}_{-0.4} \pm 0.6$
aGNN set 1	34	$25.7^{+11.9}_{-4.8} \pm 9.1$
aGNN set 2	1	$0.19^{+0.18}_{-0.05} \pm 0.11$
aGNN set 3	4	$2.2^{+1.6}_{-0.5} \pm 1.2$

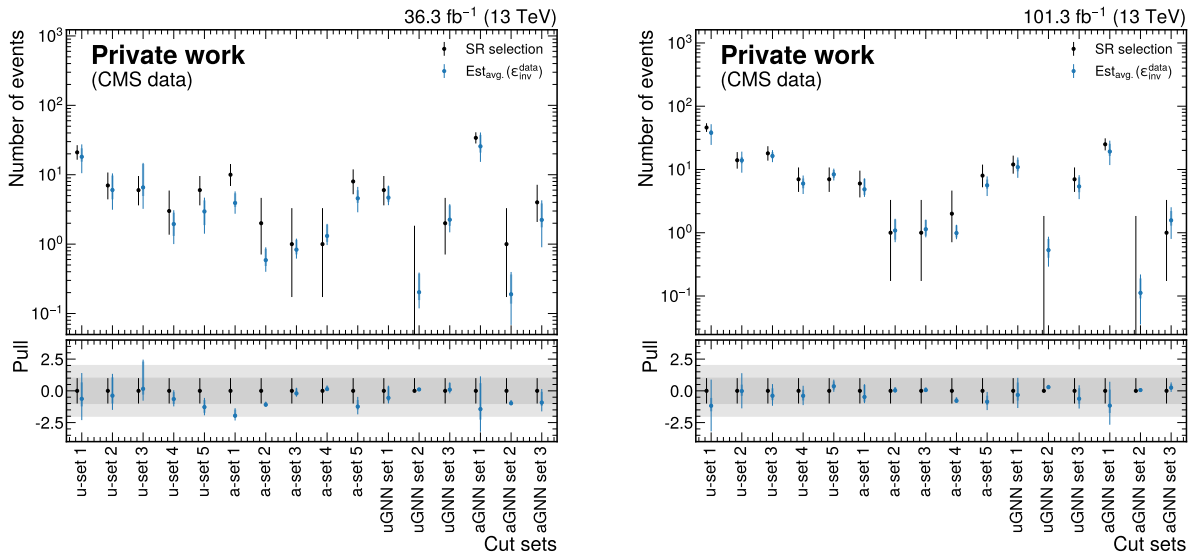


Figure 7.1: Graphical representation of the number of events passing the SR selection criteria defined in Table 4.4 compared with the background estimation calculation defined in Chapter 5. Data collected in 2016 (2017+2018) are shown on the left (right).

Table 7.2: The number of events that pass the signal selection criteria defined in Table 4.4, and the estimated number of events using the method discussed in Chapter 5 applied to the events in the control region of the H_T -triggered data set. Results obtained using 2016 data (pixel Phase-1 geometry) with $\mathcal{L} = 101.3 \text{ fb}^{-1}$.

Cut set	Observed Yield	Estimated background	
u-set 1	46	$38.0^{+4.2}_{-2.9}$	± 13.3
u-set 2	14	$13.9^{+1.9}_{-1.4}$	± 4.8
u-set 3	18	$16.4^{+2.5}_{-1.5}$	± 3.0
u-set 4	7	$6.0^{+1.2}_{-0.9}$	± 1.6
u-set 5	7	$8.4^{+1.4}_{-1.0}$	± 1.3
a-set 1	6	$4.9^{+2.3}_{-0.9}$	± 0.7
a-set 2	1	$1.08^{+0.54}_{-0.31}$	± 0.22
a-set 3	1	$1.14^{+0.45}_{-0.24}$	± 0.16
a-set 4	2	$0.99^{+0.32}_{-0.15}$	± 0.13
a-set 5	8	$5.6^{+1.3}_{-0.7}$	± 1.7
uGNN set 1	12	$10.9^{+3.3}_{-1.5}$	± 3.2
uGNN set 2	0	$0.53^{+0.27}_{-0.13}$	± 0.20
uGNN set 3	7	$5.4^{+2.1}_{-1.0}$	± 1.7
aGNN set 1	25	$19.2^{+6.3}_{-2.9}$	± 6.9
aGNN set 2	0	$0.112^{+0.078}_{-0.021}$	± 0.074
aGNN set 3	1	$1.57^{+0.63}_{-0.28}$	± 0.72

where μ is the **signal strength** parameter used to express the both $\mathcal{H}_s$ and $\mathcal{H}_{s+b}$ in a single expression, with $\mu = 0$ corresponding to $\mathcal{H}_b$ and $\mu = 1$ corresponding to $\mathcal{H}_{s+b}$. Both the expected number of signal events and background events have additional **nuisance parameters** $\vec{\nu}_x$ that modify the number of expected events, which are used to model the uncertainties described in Chapter 6. The likelihood that we observe a fixed number of events given a hypothesis can then be expressed as:

$$L(n|\mu, \vec{\nu}_{\text{bkg.}}, \vec{\nu}_{\text{sig.}}) = P\left(n|b(\vec{\theta}_{\text{bkg.}}) + \mu s(\vec{\nu}_{\text{sig.}})\right) \cdot \mathcal{W}(\vec{\nu}_{\text{bkg.}}, \vec{\nu}_{\text{sig.}}), \quad (7.2)$$

where P is a Poisson distribution, and $\mathcal{W}$ is the weighting factor used to model the uncertainty indexed i . For the multiplicative relative uncertainties that are uncorrelated, like the uncertainties listed in Chapter 6, the convention of modeling uncertainty is defined as [111]:

$$s(\vec{\nu}) = s_0 \prod_i (1 + \kappa_i)^{\nu_i} ; \quad W(\vec{\nu}) = \prod_i \mathcal{N}(\nu_i; 0, 1), \quad (7.3)$$

where s_0 is the central value of the number of expected number of events, κ_i is the relative uncertainty of the source i (as listed in Tables 6.1–6.4), ν_i is the parameter used to model uncertainty, and $\mathcal{N}$ is a normal distribution. For brevity, the various uncertainty terms in the likelihood definition will be expressed using the single $\vec{\theta}$.

Having defined the likelihood function, the next step is to define the test-statistic for performing a hypothesis testing. The standard test-statistic used at LHC experiments for comparing

hypotheses is the modified negative log-likelihood ratio [112, 112]:

$$q(n|\mu) = \begin{cases} \lambda(n|\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} ; \quad \lambda(n|\mu) = -2 \ln \left(\frac{L(n|\mu, \tilde{\nu}(\mu))}{L(n|\hat{\mu}, \hat{\nu})} \right) \quad (7.4)$$

where $\tilde{\nu}(\mu)$ is the maximum likelihood estimate for $\vec{\nu}$ for a fixed value of μ , while $\hat{\mu}, \hat{\nu}$ are the unconditional maximum likelihood estimates for μ and $\vec{\nu}$, respectively. Using this test statistic we can formally define the upper limit $q_{\text{lim.}}$ and subsequently $\mu_{\text{lim.}}$ using the Frequentist p -value approach by solving:

$$p = \int_{q_{\text{lim.}}}^{\infty} \mathcal{P}(q(n|\mu)|\mathcal{H})dq, \quad (7.5)$$

where p is typically chosen to be 5% such that the results of $q_{\text{lim.}}$ corresponds to the 95% confidence level, and $\mathcal{P}$ is the probability distribution function of observing test statistic q under the hypothesis $\mathcal{H}$, given the observation of n events occurring. The modification to q from the simple log-likelihood ratio λ is to ensure that over-fluctuation of the observed signal yield is not used to exclude $\mathcal{H}_{s+b}$. For the results of an experiment, we arrive at a singular observed value n_{obs} which is used to define the **observed upper limit** (or exclusion limit) μ_{obs} . We can also generate a set of $\{n\}$ values under $\mathcal{H}_b$, and use $\{n\}$ calculated a set of **expected upper limits** (or exclusion limit) $\{\mu_{\text{exp}}\}$. The consistency between μ_{obs} and $\{\mu_{\text{exp}}\}$ is also used as a validation for the statement of “no significant excessive observed”.

Finally, to avoid excluding $\mathcal{H}_{s+b}$ when the sensitivity is low, the **CL_s upper limit** (or

exclusion limit) is determined using p -value ratios crossing the 5% threshold:

$$\text{CL}_s = \frac{p_{s+b}}{p_b}, \quad (7.6)$$

where p_b is calculated using Eq. 7.5 except with μ fixed at 0. While we can sample the probability distribution of $\mathcal{P}$ by following the distributions defined in Eq. 7.2 and 7.3, Ref. [113] provides an asymptotic limit of the expected distribution to speed up the calculations of μ_{obs} and $\{\mu_{\text{exp}}\}$. All the calculations described above are performed using the `combine` software package [114] developed for the Higgs discovery analysis.

The calculations of μ_{obs} and $\{\mu_{\text{exp}}\}$ are performed separately for all surveyed signal samples, comparing the total number of signal events expected to pass the designated selection criteria listed in Table 4.4 with the expected number of background events using the calculation defined in Chapter 5 and displayed in Tables 7.1 and 7.2. The results of μ_{obs} and $\{\mu_{\text{exp}}\}$ are scaled by their theoretical cross section for comparison, and are typically presented using the “Brazilian flag” plot, as shown in Fig. 7.2: colored bands are used to summarize the spread of $\{\mu_{\text{exp}}\}$, the black line is used to represent μ_{obs} . The consistency of the black line and the colored bands represent the claim of “no significant excess”. For the theoretical phase space where the exclusion limit is calculated to be smaller than the theoretical cross section, we can claim that such models are excluded by experimental observation at CL_s of 95%.

To represent the results of the surveyed signal modes, we further compact the results as 2D plots displayed in Fig. 7.3–7.6. The median value of the expected upper limit is displayed as the color gradients; the intersection of the median of the expected upper limit and the $\pm 1\sigma$ variations with the theoretical cross section are displayed using a red line and red bands, respectively; the

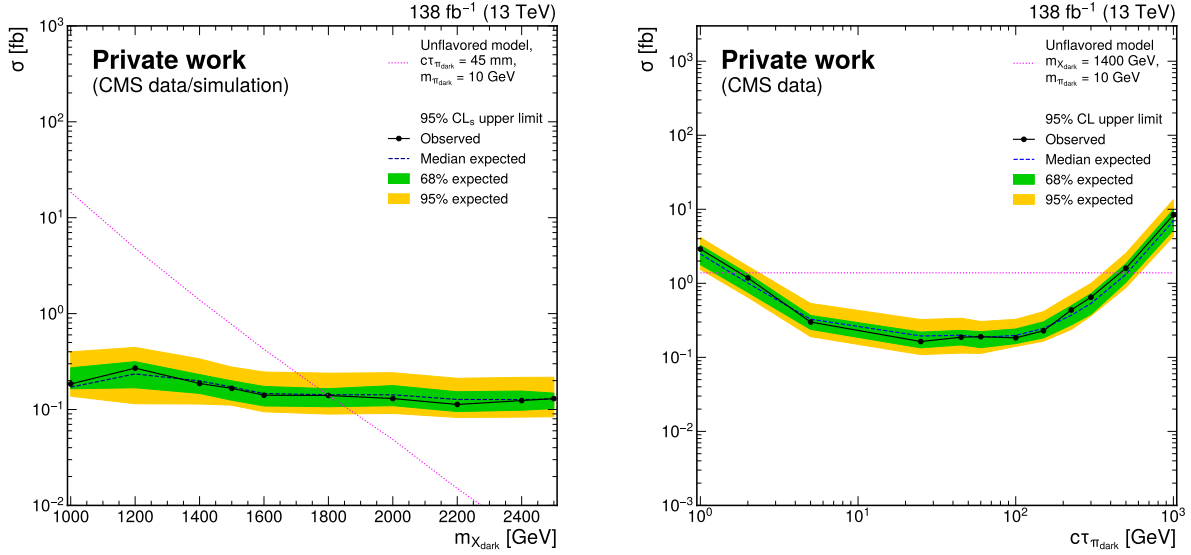


Figure 7.2: The observed (black line) and expected (red, dashed line and color bands) exclusion limit compared with the theoretical cross section (blue, dotted line) for various signal samples calculated using the asymptotic CL_s method. The left plot contains unflavored signal models with $m_{\pi_{\text{dark}}} = 10$ GeV and $c\tau_{\pi_{\text{dark}}} = 100$ GeV, while the right plot contains unflavored signal models with $m_{\pi_{\text{dark}}} = 10$ GeV and $m_{X_{\text{dark}}} = 1400$ GeV.

intersection of the observed upper limit with the theoretical cross section displayed using a black line. As the theoretical cross section of the signal model monotonically decreases with $m_{X_{\text{dark}}}$, the signal models left of the curves shown in the figures are excluded by our experimental observations at CL_s of 95%. The results of the previous CMS analysis (Ref. [44]) are also portrayed for the signal models in the unflavored scenario with $m_{\pi_{\text{dark}}} = 10$ GeV, and are displayed in the figures as blue lines.

7.2 Interpreting results of the different EMJ tagging methods

For the unflavored scenario, as all dark mesons have the same lifetime, the model parameters uniformly shift all displaced track signatures within the signal jet. Since the model-agnostic method directly looks for track displacement signature, as presented in Chapter 4, this method of tagging is expected to lose sensitivity when the scale of $c\tau_{\pi_{\text{dark}}}$ reaches the maximum scale of

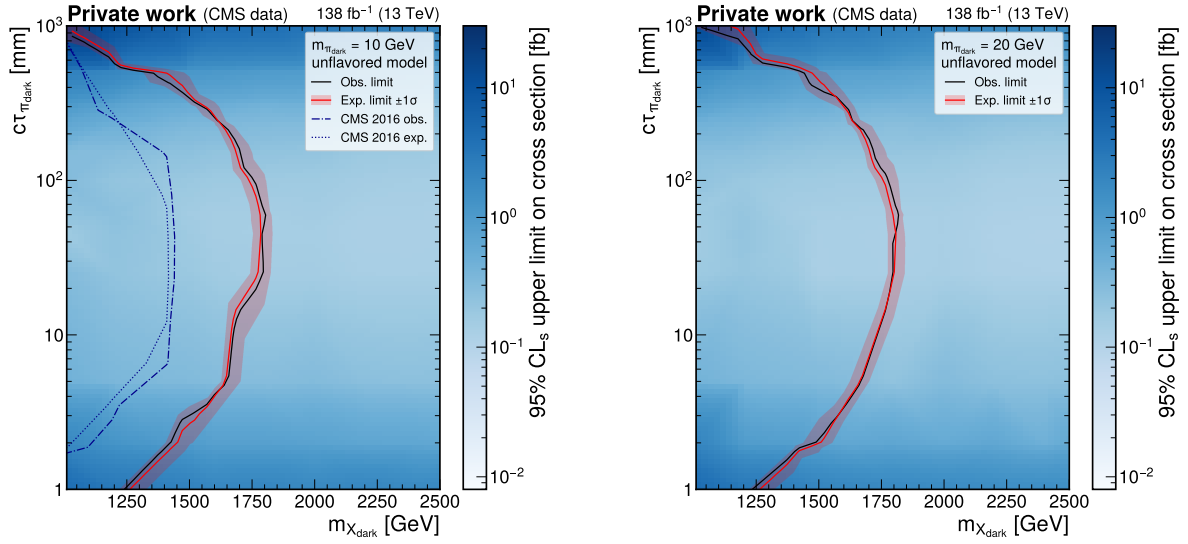


Figure 7.3: The computed limits of the production cross section calculated as a function of $m_{X_{\text{dark}}}$ and $c\tau_{\pi_{\text{dark}}}$ and the model-agnostic EMJ tagging method for signal models in the unflavored scenario with $m_{\pi_{\text{dark}}} = 10$ GeV (left) and $m_{\pi_{\text{dark}}} = 20$ GeV (right). Details of elements are described in the main text.

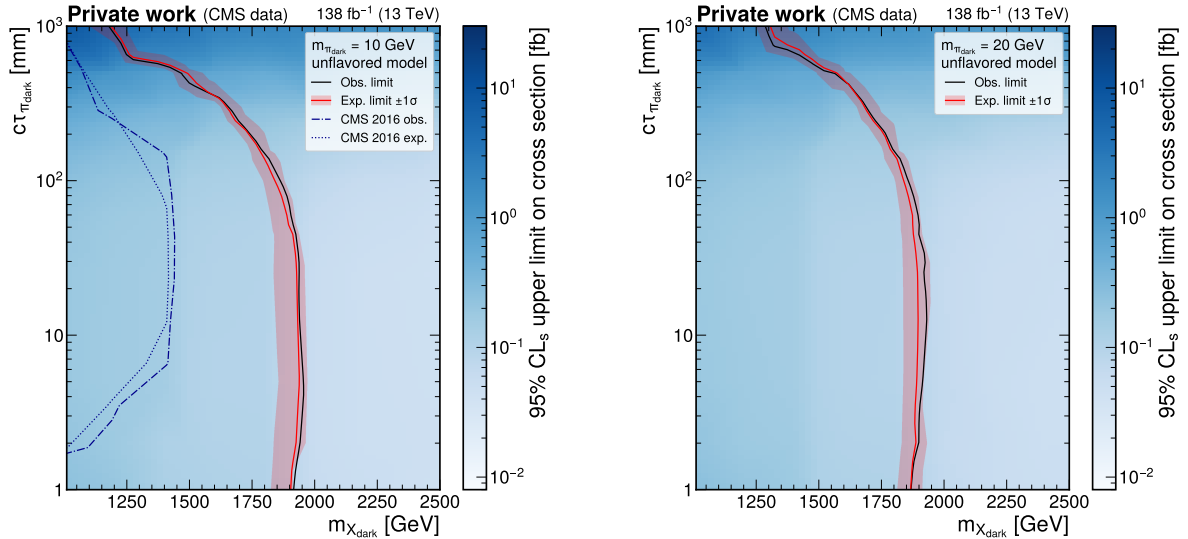


Figure 7.4: The computed limits of the production cross section calculated as a function of $m_{X_{\text{dark}}}$ and $c\tau_{\pi_{\text{dark}}}$ and the GNN EMJ tagging method for signal models in the unflavored scenario with $m_{\pi_{\text{dark}}} = 10$ GeV (left) and $m_{\pi_{\text{dark}}} = 20$ GeV (right). Details of the plot elements are described in main the text.

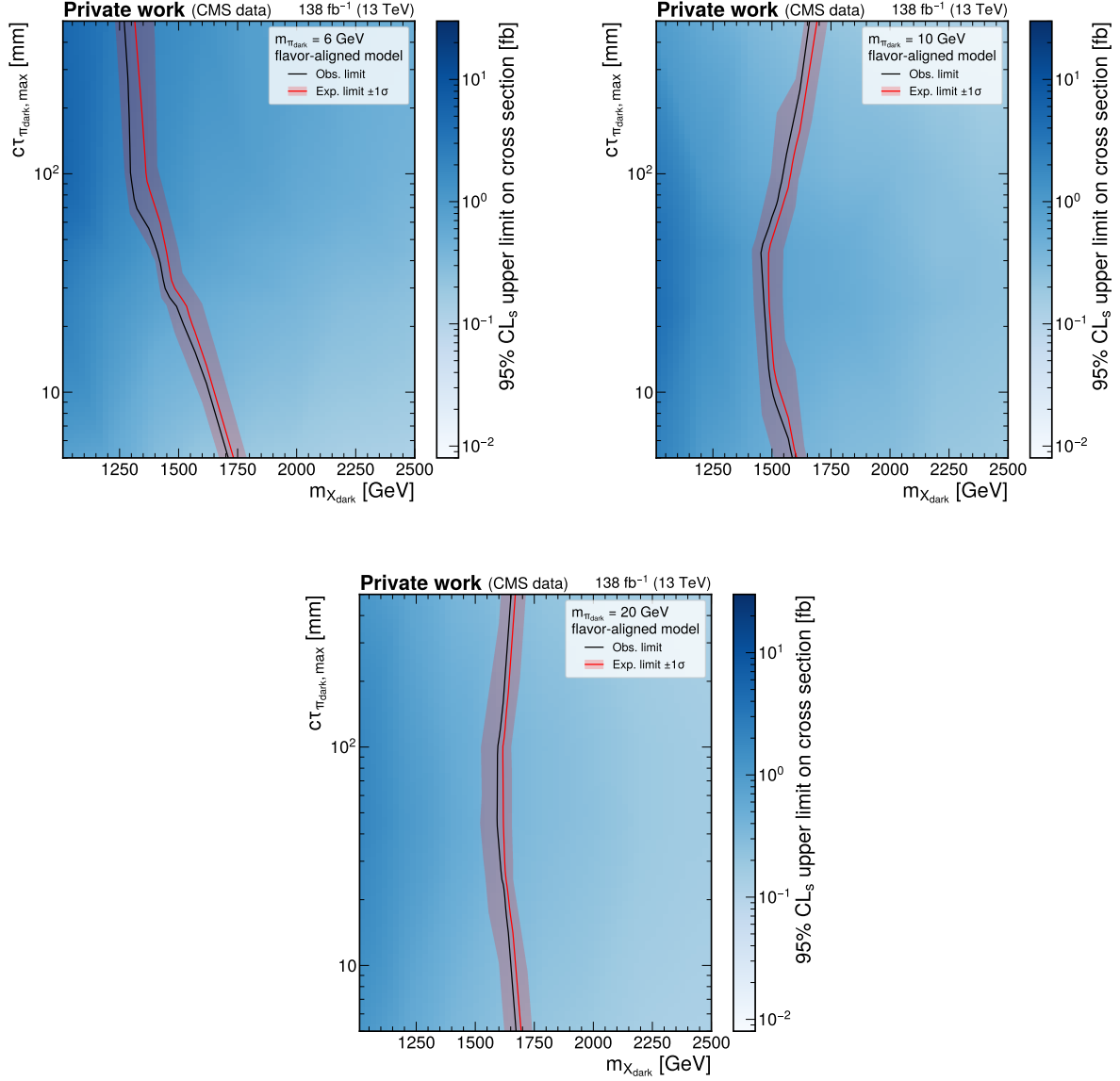


Figure 7.5: The expected limit of the production cross section calculated as a function of $m_{X_{\text{dark}}}$ and $c\tau_{\pi_{\text{dark}}}$ and the model-agnostic EMJ tagging method for signal models in the flavor-aligned scenario with $m_{\pi_{\text{dark}}} = 6$ GeV (top left), $m_{\pi_{\text{dark}}} = 10$ GeV (top right) and $m_{\pi_{\text{dark}}} = 20$ GeV (bottom). Details of plot elements are given in the main text.

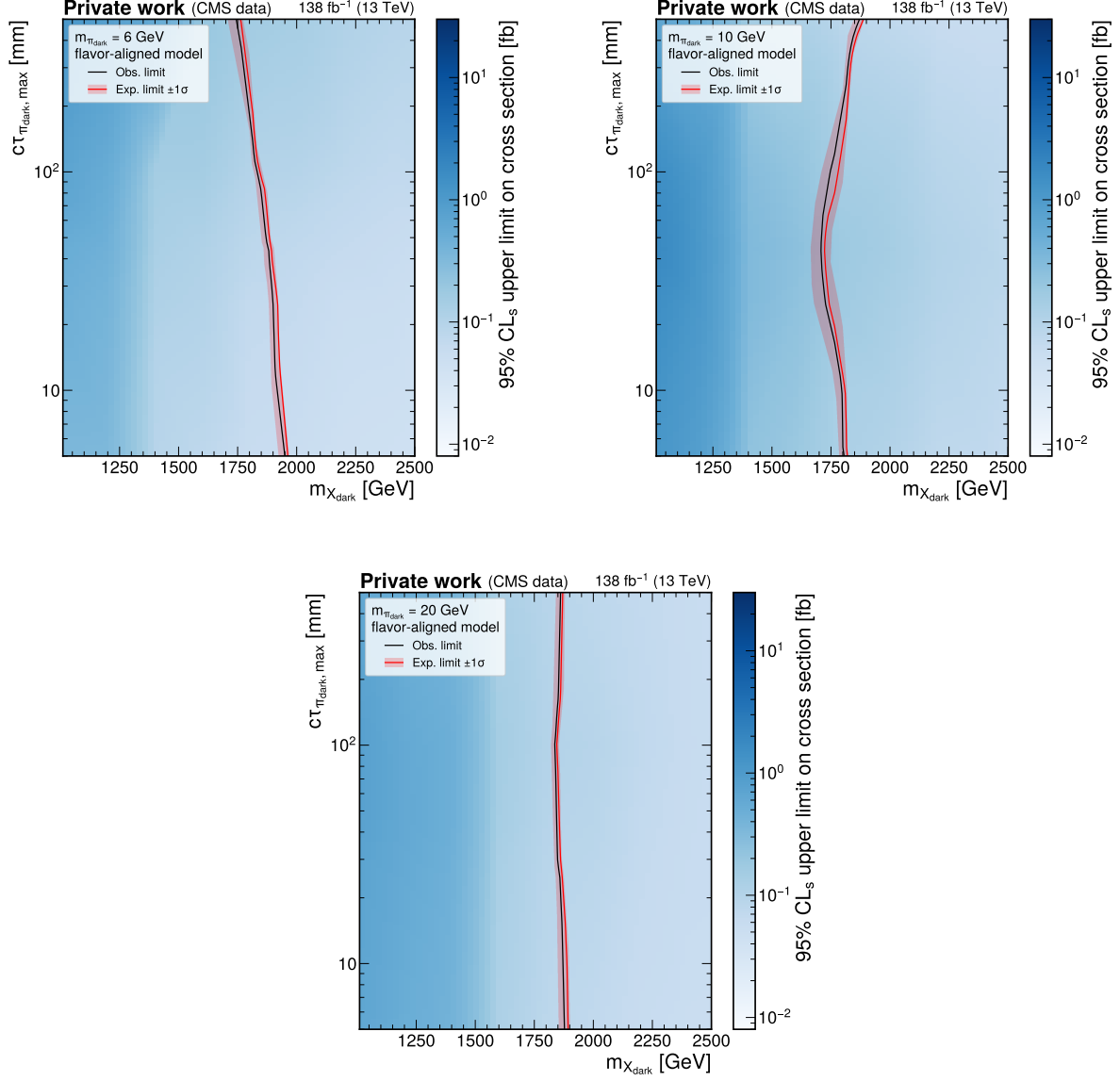


Figure 7.6: The expected limit of the production cross section calculated as a function of $m_{X_{\text{dark}}}$ and $c\tau_{\pi_{\text{dark}}}$ and the GNN-based EMJ tagging method for signal models in the flavor-aligned scenario with $m_{\pi_{\text{dark}}} = 6 \text{ GeV}$ (top left), $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$ (top right), and $m_{\pi_{\text{dark}}} = 20 \text{ GeV}$ (bottom). Details of plot elements are given in the main text.

tracker volume, or when $c\tau_{\pi_{\text{dark}}}$ is very small such that all tracks appear prompt-like. This limitation of sensitivity can be seen in the loss of sensitivity for the long-lifetime signal models seen in Fig. 7.3.

For the flavor-aligned scenario, with the dark mesons having mixed lifetime signatures, the model-agnostic method primarily operates by detecting the b meson production that is expected to occur in the dark meson decay. Because of this, the overall sensitivity to signal events is less dependent on the lifetime parameter κ_0 , as seen in Fig. 7.5.

When comparing the results of the GNN-based EMJ tagging methods with the results of the model-agnostic EMJ tagging, the GNN-based method consistently performs “better,” as seen in the extended area of the model parameters space that is excluded by the GNN-based tagging method in Fig. 7.4 and 7.6. However, we also note that the sensitivity of the GNN-based tagging method exhibits a very different behavior in the unflavored scenario for the short-lifetime signal models. The sensitivity roughly remains constant below $c\tau_{\pi_{\text{dark}}} \sim 45$ mm. As short-lifetime EMJ showers should not be easily distinguishable from SM showers looking at displacement, this is an indication that the GNN is probing features not directly related to the displaced signature of an EMJ shower, but something specific about the dark sector shower of the simulated events. This difference in sensitivity is potentially problematic, as it indicates that the GNN can potentially misinterpret the results of dark sector showering if the dark sector deviates from the models defined in Section 3.6.2. Given this potential issue, the GNN-based approach should only be used to draw conclusions regarding the particular signal models surveyed. In contrast, the model-agnostic approach can be readily interpreted to set limits on alternate EMJ-generating models that might be proposed in the future, at the expense of limited sensitivity compared with the GNN for the classes of signal models surveyed in this analysis.

7.3 Conclusion and outlook

This thesis presents a search for emerging jets arising from a strongly interacting dark sector produced in pp collisions at the LHC, using data collected by the CMS detector corresponding to a total luminosity of 138 fb^{-1} at $\sqrt{s} = 13 \text{ GeV}$. Both flavor-aligned and unflavored couplings between the SM quarks and the dark quarks are examined in the search. With the model-specific search method using GNN-based jet tagging methods, the result excludes the unflavored model parameter ranges of up to $m_{X_{\text{dark}}} < 1750 \text{ GeV}$ for $c\tau_{\pi_{\text{dark}}} \sim 100 \text{ mm}$ and $m_{\pi_{\text{dark}}} \sim 10 \text{ GeV}$, and excludes the flavored model parameter range up to $m_{X_{\text{dark}}} < 2000 \text{ GeV}$ at $c\tau_{\pi_{\text{dark}}} \sim 500 \text{ mm}$ for $m_{\pi_{\text{dark}}} \sim 10 \text{ GeV}$. This result surpasses the previous search for emerging jets in the unflavored scenario, increasing the theoretical limit of the dark mediator particle by $\sim 550 \text{ GeV}$ to set the most stringent limits to date, and providing the first direct exclusion of the flavored scenario.

This analysis primarily focuses on the search for the dark sector signature using tracker signatures. CMS has performed complimentary phenomenological searches such as the search for “track-less” jets [115] (a search for long-lived particles with calorimeter signature) and the search for muon clusters [116] (a search for long-lived particles with muon system signatures) that can potentially be reinterpreted to further extend the search space. There are also searches for other dark sector scenarios, such as the search for semi-visible jets [117, 118]. While the search for emerging jets presented in this thesis is an important milestone in dark sector searches at the LHC, many more interesting directions can be taken for this field.

Chapter A: Comparison with the previous EMJ search efforts

The search for emerging jets has previously been conducted by CMS using the data collected in 2016 post-APV fixes [44]. While the analysis presented in this thesis is based on the previous search effort, aside from the addition of a new signal model and the design of the corresponding EMJ tagging technique, there are also multiple nontrivial changes to the analysis strategy, after closer investigations into the validity of the assumptions that are required for the background estimation method to be robust. While some of these investigations are provided by the main text, some are difficult to properly introduce along with the other content. In this appendix, we will list all the changes to the analysis since the previous EMJ analysis, including arguments for the final decision made for the analysis

A.1 Removal of explicit p_T^{miss} cut for signal selection

Signal samples with very long lifetime dark mesons are expected to have larger p_T^{miss} as a larger fraction of the dark shower constituents leave the detector undetected. While requiring that the signal region pass an explicit event p_T^{miss} threshold may provide additional signal background separation power (as seen in Fig. A.1), it was observed that event-level p_T^{miss} requirements will inevitably shift the constituents of jets entering the signal region. An example can be demonstrated in Fig. A.2 by looking at the leptonic energy fraction of the jet in events with and without explicit

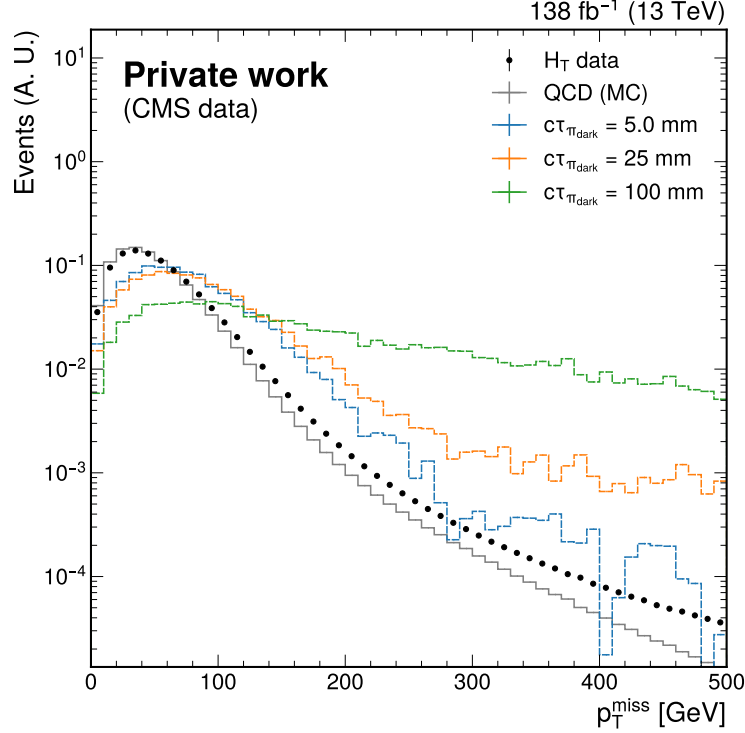


Figure A.1: Comparison between the p_T^{miss} distribution of the H_T -triggered data set, the QCD MC samples distribution, and various signal samples, corresponding to the full run 2 luminosity. All samples have the total number of events normalized to 1. The signal samples displayed here are the unflavored samples with $m_{X_{\text{dark}}} = 1400 \text{ GeV}$ and $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$.

p_T^{miss} requirements. Such shifts are expected, as requiring event-level p_T^{miss} thresholds will favor events with jets containing leptonic decays that generate neutrinos. Shifts in the jet population will ultimately shift the high-level EMJ tagging variables, as seen in the right plot Fig. A.2.

This is particularly problematic when we attempt to use the mistag rate evaluated in a region without a p_T^{miss} requirement to evaluate the background estimation scale factor applied to control region with an explicit p_T^{miss} requirement. When examining the result in Ref. [44], the cut sets with explicit p_T^{miss} requirements demonstrated the largest uncertainty, indicating that this was already an implicit issue in the previous uncertainty. Now knowing that p_T^{miss} has a significant impact on the displacement measure used to tag EMJs, the following strategies have been considered:

1. Performing a p_T^{miss} cut on the signal-free region for the EMJ taggers that are paired with

an explicit p_T^{miss} cut. This resulted in too few jets in the signal-free region to perform a statistically stable mistag rate computation.

2. Correcting the mistag rate obtained in the signal-free region by the differences in the mistag rate obtained in the QCD MC and the γ +jets MC samples. This method however fails to consider that events in the H_T -triggered data set has a clear excess of large p_T^{miss} events when compared with QCD MC samples, as seen in Fig. A.1. This excess is generated from processes other than QCD interactions, such as multi-boson generation, $t\bar{t}$ processes. Simply correcting the mistag rate by comparing QCD MC samples and γ +jets MC samples will not correctly reflect the contributions of such events in the background estimation.
3. Attempt to measure mistag rate in a signal-free region with the H_T -triggered data stream. This alternative was also sought as a solution to solve the differences in jet constituents in the H_T -triggered data set and the γ -triggered data set as discussed in Section 5.1.4. However, we were unable to find a reliable way to construct a region that is guaranteed to be signal-free using the H_T -triggered data set.
4. Remove any explicit p_T^{miss} cuts in signal region selection. This ensures that the mistag rates obtained in the signal-free region are not skewed by potential jet constituent drifts induced by the p_T^{miss} cut. The main concern for this method is the potential loss in signal significance for the long-lifetime signal processes.

Ultimately, removing the explicit p_T^{miss} cuts was the simplest path forward without needing to reevaluate many of the analysis strategies, and this is what was performed in the main text.

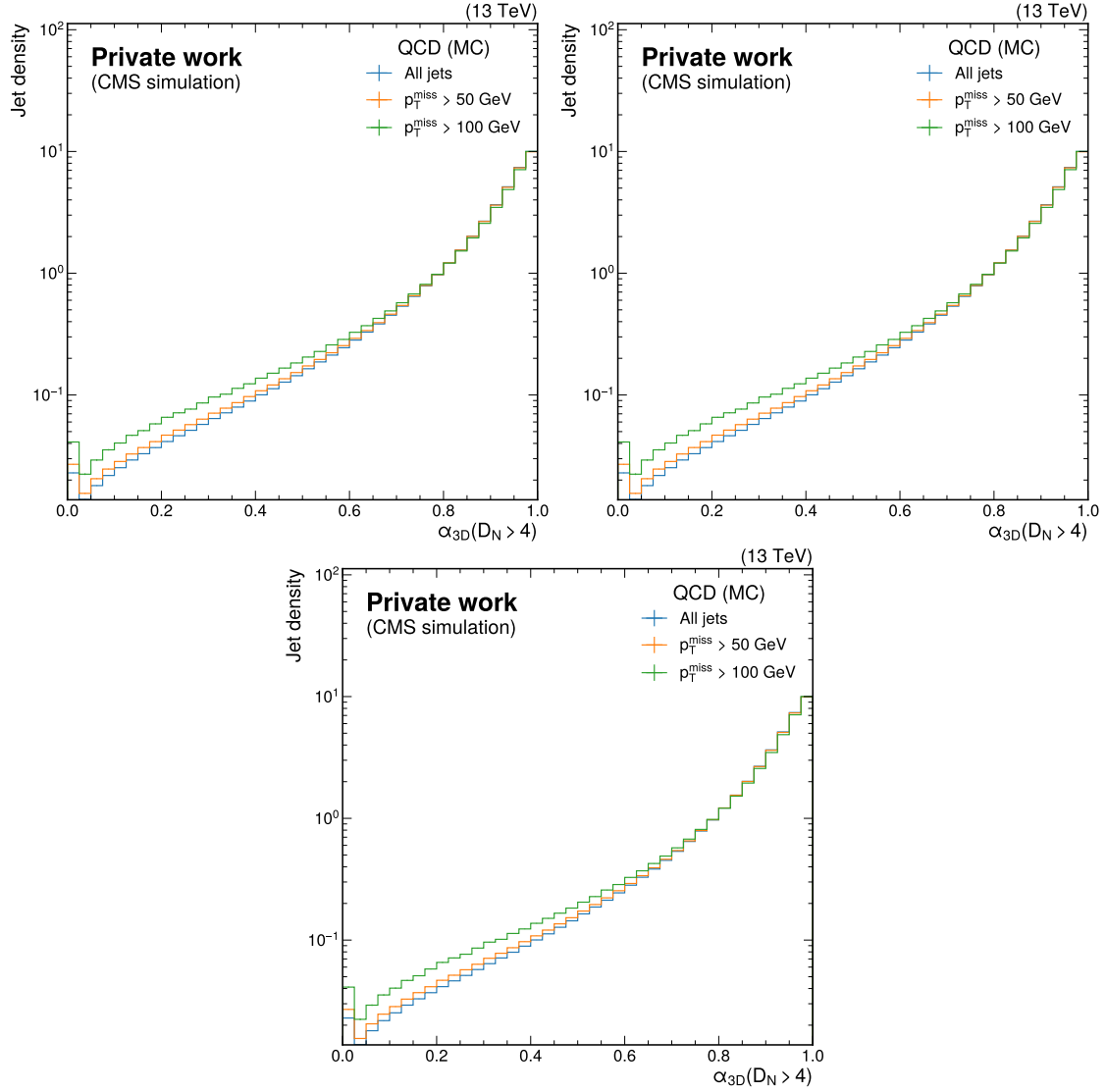


Figure A.2: Shifts in electron energy fraction (top left), muon energy fraction (top right), and the α_{3D} defined with $D_N > 4$ (bottom) of jets in the QCD MC samples, with varying event-level p_T^{miss} cuts.

A.2 Parameterization of the mistag rate using jet p_T

Previously, the analysis parameterized mistag rates using track multiplicity, as mistag rates appear to have the strongest dependence on track multiplicity compared with other jet variables. The correctness of this assessment has been discussed extensively in Section 5.3, and it was found that parameterizing the mistag rate in track multiplicity does not properly reflect the underlying physics of mistag rate variations among the jets of interest. Because of the arguments presented in the main text, EMJ mistag rates will be parameterized in jet p_T .

A.3 Discarding per-jet flavor assignment for background estimation

In the previous analysis, the per-jet flavor assignment was presented as a backup to the primary flavor-averaging method for calculating the SR/CR scale factor presented in Section 5.2.2. For this analysis, we even further completed the argument of per-jet flavor assignment by expanding the argument to be fully per-jet compared with jet counting presented in the previous analysis, and presenting a numerically stable Bayesian interpretation to the calculation. However, as presented in Section 5.2.2, it was found that the EMJ tagging requirements dramatically shift the b jet discriminant distribution. Because of this, we ultimately discarded the per-jet flavor assignment for this analysis.

A.4 Removing a vertex quality cut

In the previous analysis, aside from the leading primary vertex being ordered by the weighted $\sum p_T^2$ as defined in Section 3.4.1, it also required the leading primary vertex to have the largest

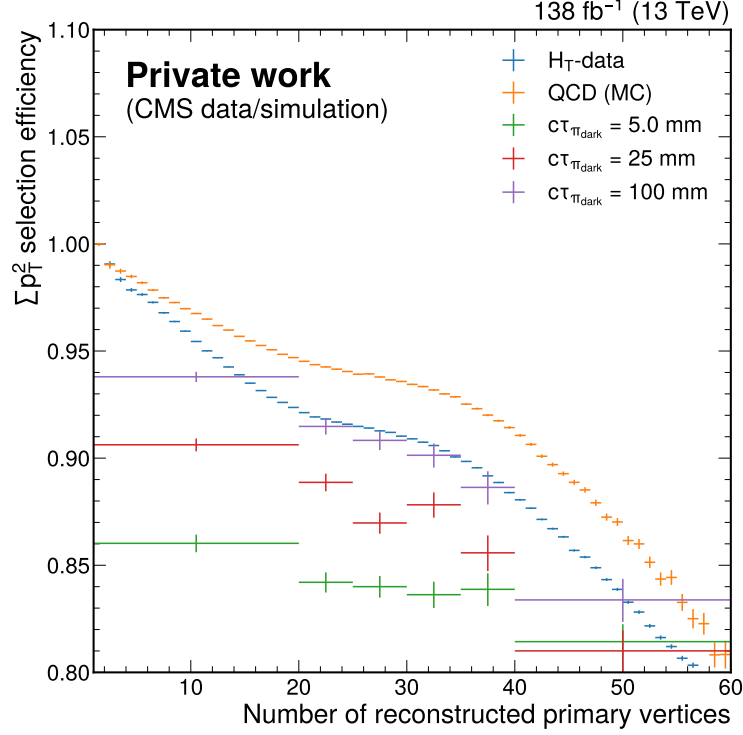


Figure A.3: Event selection efficiency of requiring the leading PV to have the maximum unweighted Σp_T^2 for various events in samples passing the trigger requirement. The only selection imposed on the various samples is the H_T -trigger used for the analysis. Signal samples of the same lifetimes are summed together to reduce the statistical uncertainty in the plot.

unweighted Σp_T^2 , as it was found that this additional quality cut was found to remove erroneous vertices which resulted in all jets being tagged as displaced.

In this analysis, we performed an additional check to compare how well this quality cut is modeled in MC by comparing QCD MC with the H_T -triggered data set. For this comparison, only the trigger requirement used by the analysis selection where imposed on the data sets of interest. The results are demonstrated in Fig. A.3. The quality cut selection efficiency is not well modeled when comparing the results of QCD MC events and H_T -triggered events, and the discrepancies are found to be not covered by pileup uncertainties. It was also found that the event selection efficiency for signal samples is lower than expected, at around $\sim 85\%$.

In light of this discrepancy and the poor signal selection efficiency, we also performed

a check whether this quality cut is still required, as there had been adjustments to the official reconstruction algorithm since the previous analysis. It was found the reconstruction failure mode of all 4 jets in the event being tagged as EMJs never occurred in the QCD MC events using the updated reconstruction algorithm. Relaxing this requirement also didn't cause issues with background closure estimation when evaluated in MC. The analysis decided to remove this vertex quality cut from the analysis to signal sensitivity.

Chapter B: Additional plots

The following is the exhaustive display of plots of results that are used for the analysis, but clutters the reading of the main text too much.

B.1 Kinematic dependence of the mistag rates

The plots in the following section display the kinematic dependence of the various EMJ taggers defined in Table 4.1– 4.3. For all plots in this section, the mistag rates are evaluated using jets in the QCD MC samples as a function of jet p_T , jet η , and the number of associated tracks. The mistag rates of b jets are shown in red, non-b jets (labeled light jets) in blue, while the combined mistag rate of all jet flavors is shown in gray.

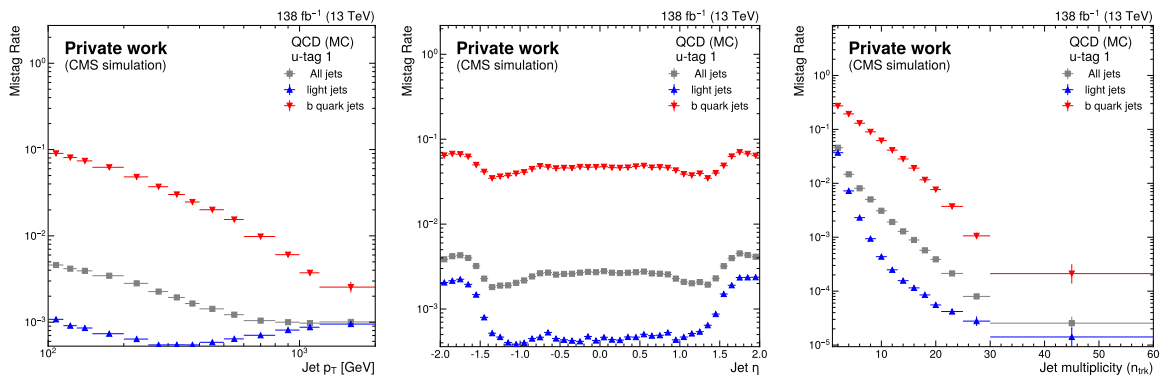


Figure B.1: Mistag rate of the “u-tag 1” tagger evaluated using QCD MC samples.

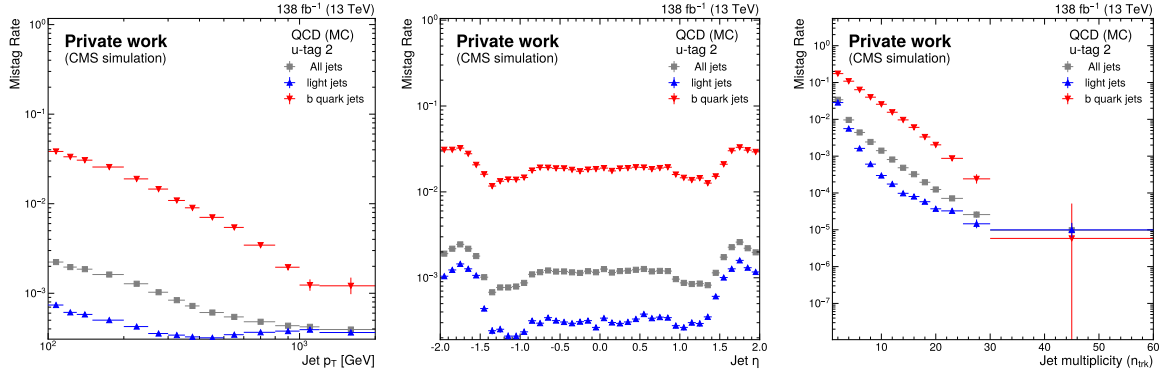


Figure B.2: Mistag rate of the “u-tag 2” tagger evaluated using QCD MC samples.

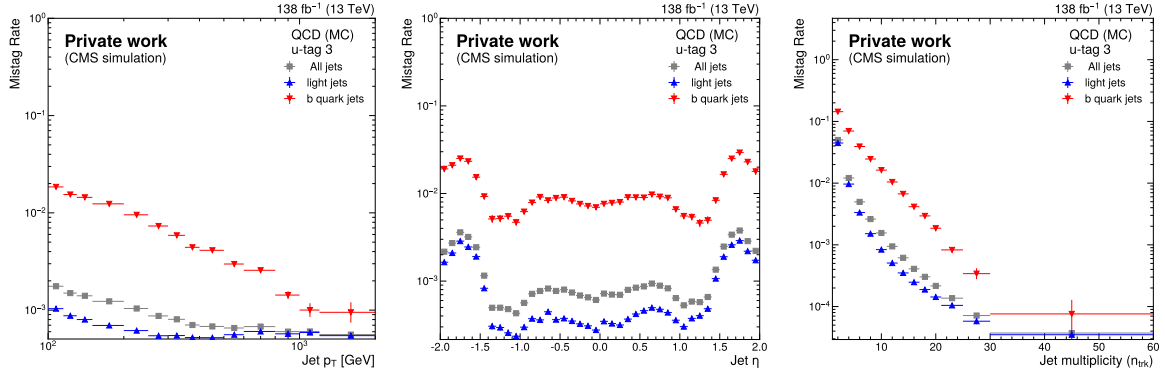


Figure B.3: Mistag rate of the “u-tag 3” tagger evaluated using QCD MC samples.

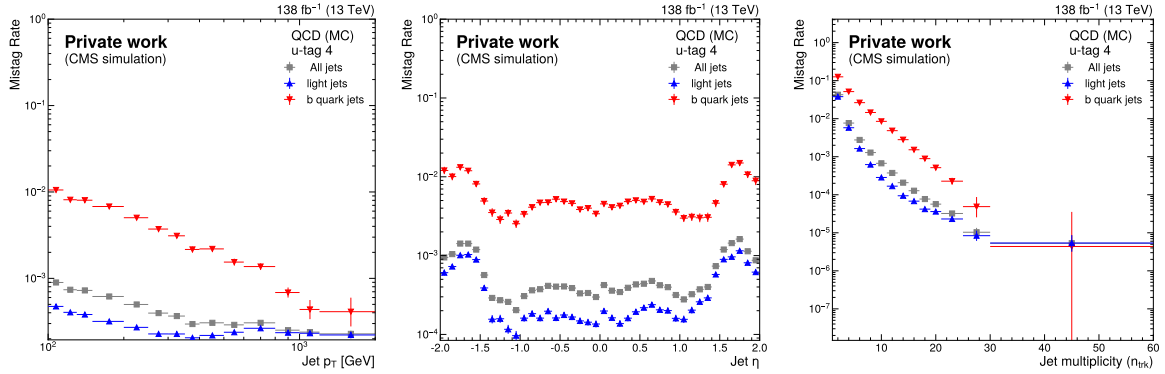


Figure B.4: Mistag rate of the “u-tag 4” tagger evaluated using QCD MC samples.

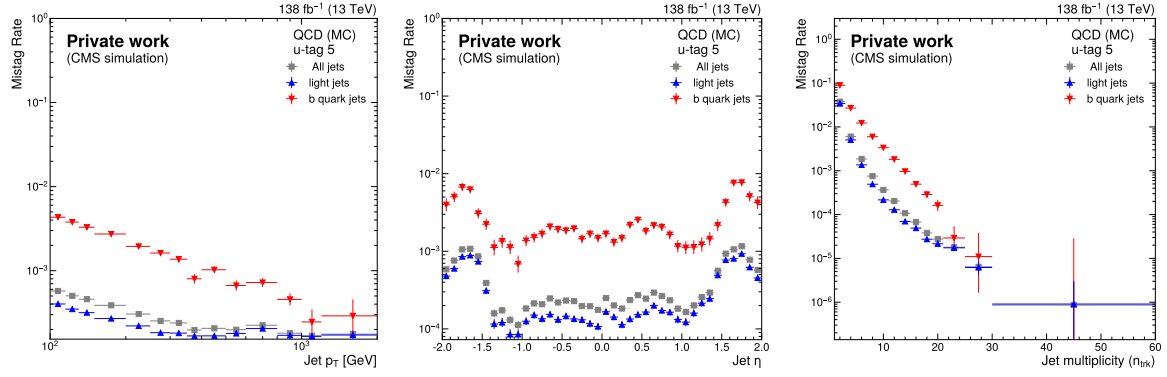


Figure B.5: Mistag rate of the “u-tag 5” tagger evaluated using QCD MC samples.

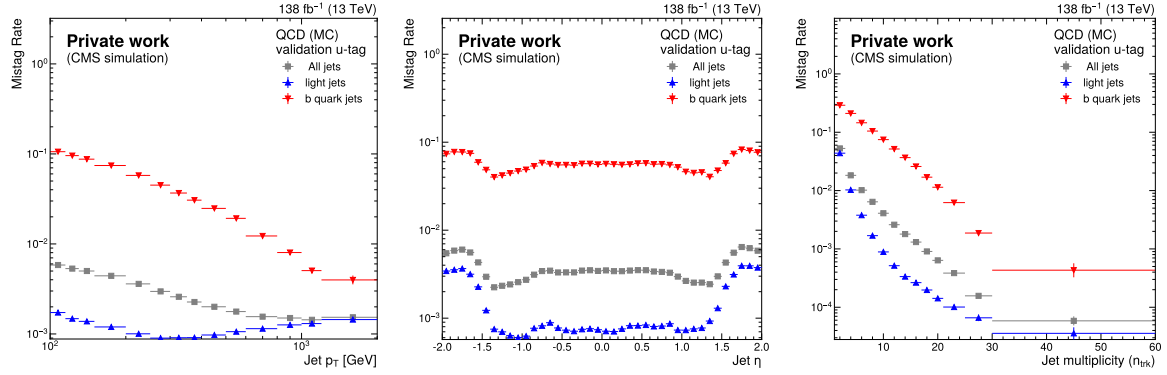


Figure B.6: Mistag rate of the “validation u-tag” tagger evaluated using QCD MC samples.

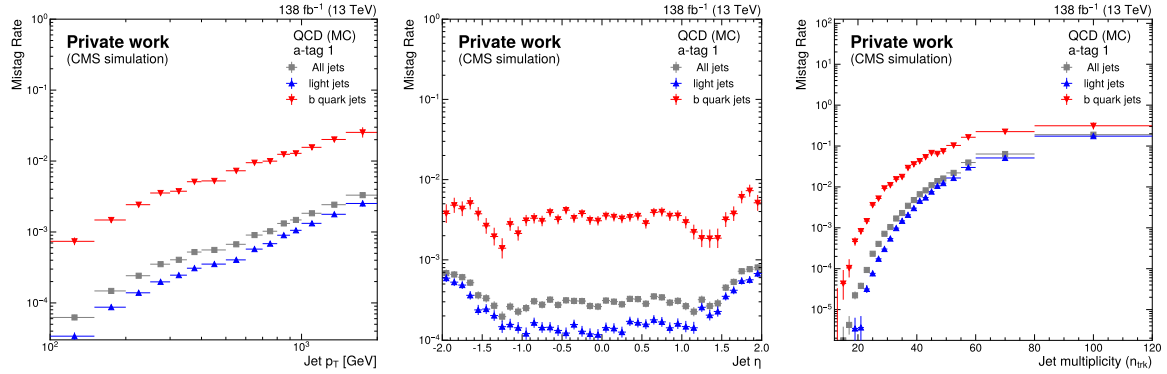


Figure B.7: Mistag rate of the “a-tag 1” tagger evaluated using QCD MC samples.

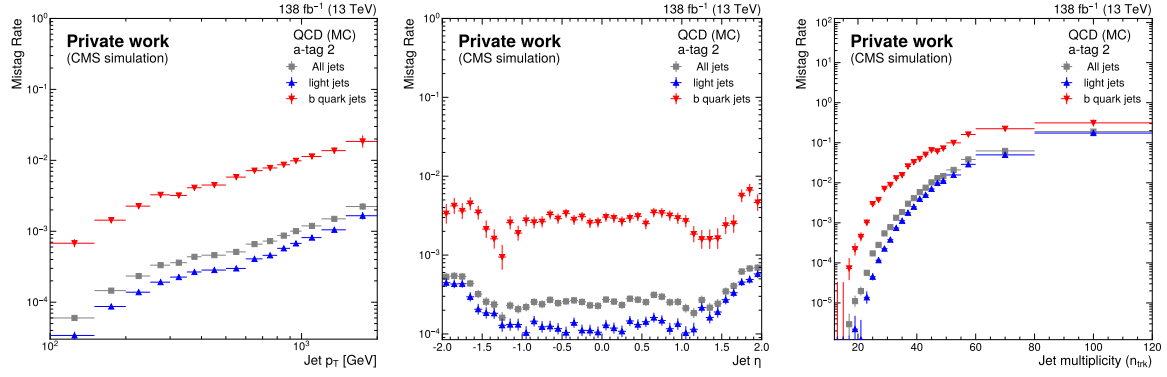


Figure B.8: Mistag rate of the “a-tag 2” tagger evaluated using QCD MC samples.

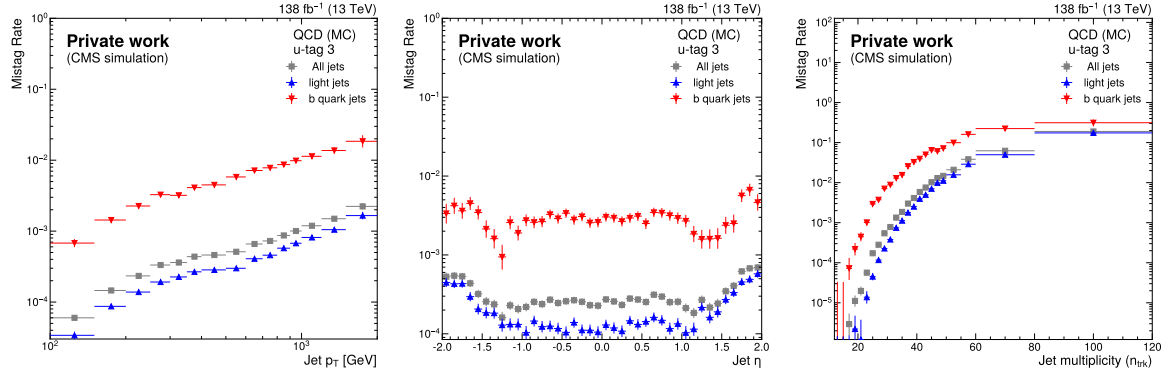


Figure B.9: Mistag rate of the “a-tag 3” tagger evaluated using QCD MC samples.

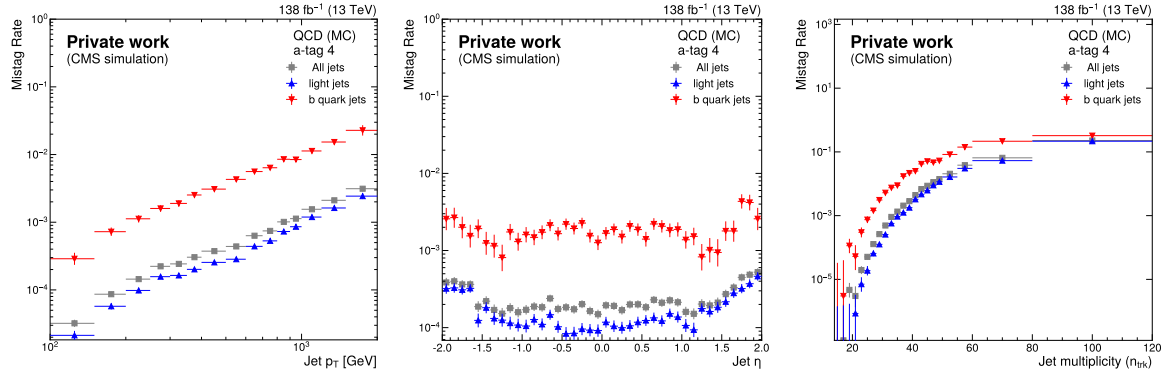


Figure B.10: Mistag rate of the “a-tag 4” tagger evaluated using QCD MC samples.

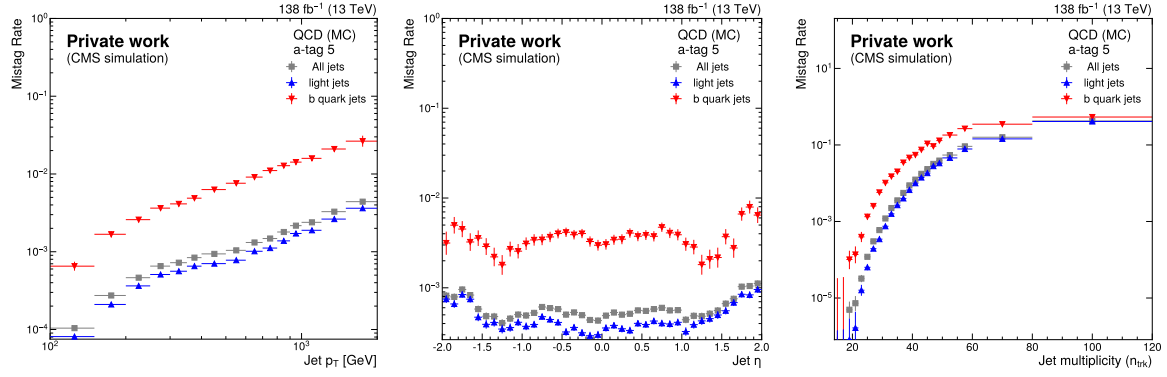


Figure B.11: Mistag rate of the “a-tag 5” tagger evaluated using QCD MC samples.

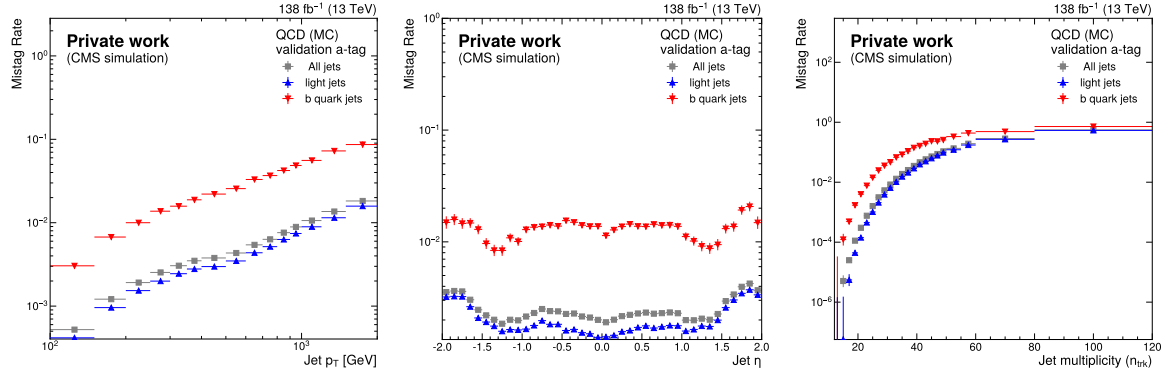


Figure B.12: Mistag rate of the “validation a-tag” tagger evaluated using QCD MC samples.

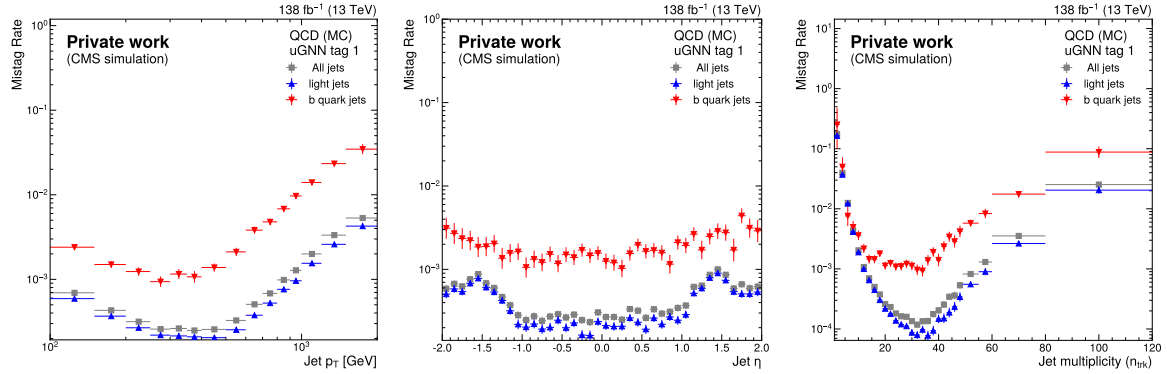


Figure B.13: Mistag rate of the “uGNN tag 1” tagger evaluated using QCD MC samples.

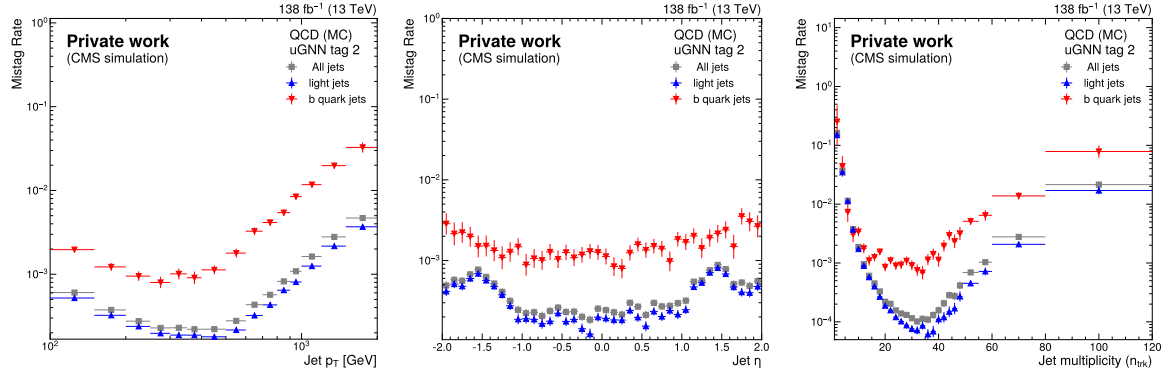


Figure B.14: Mistag rate of the “uGNN tag 2” tagger evaluated using QCD MC samples.

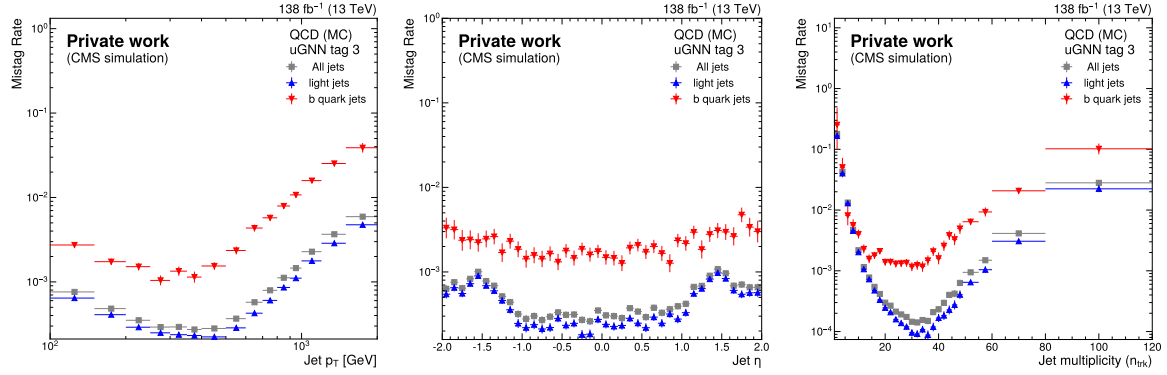


Figure B.15: Mistag rate of the “uGNN tag 3” tagger evaluated using QCD MC samples.

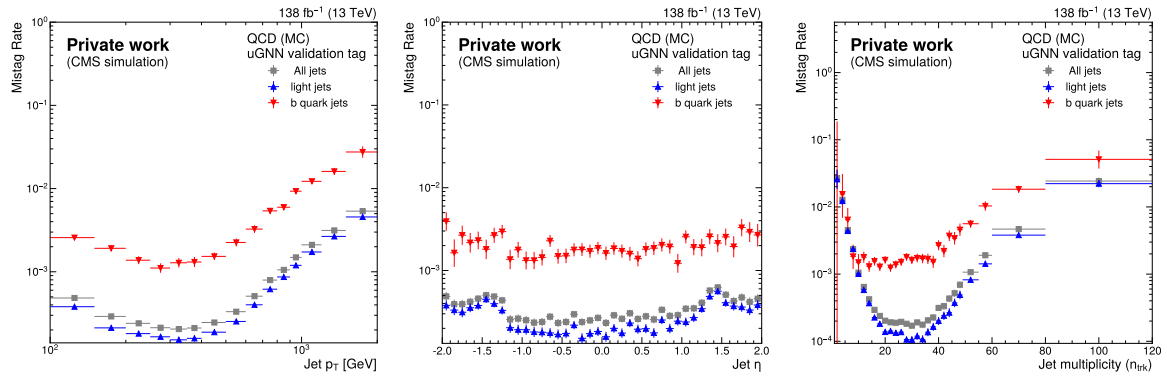


Figure B.16: Mistag rate of the “uGNN validation tag” tagger evaluated using QCD MC samples.

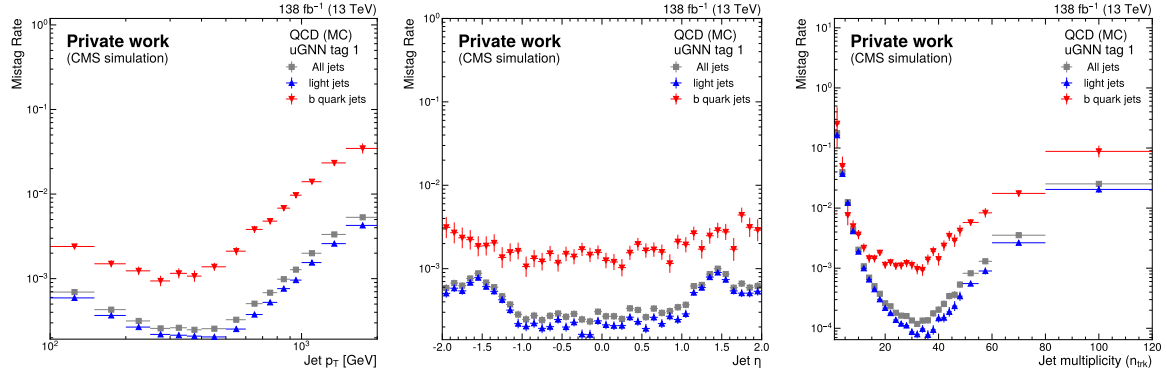


Figure B.17: Mistag rate of the “uGNN tag 1” tagger evaluated using QCD MC samples.

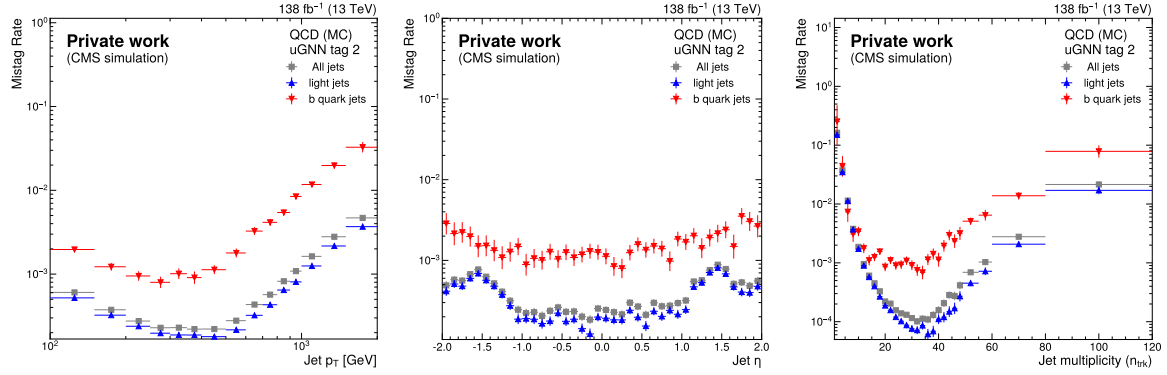


Figure B.18: Mistag rate of the “uGNN tag 2” tagger evaluated using QCD MC samples.

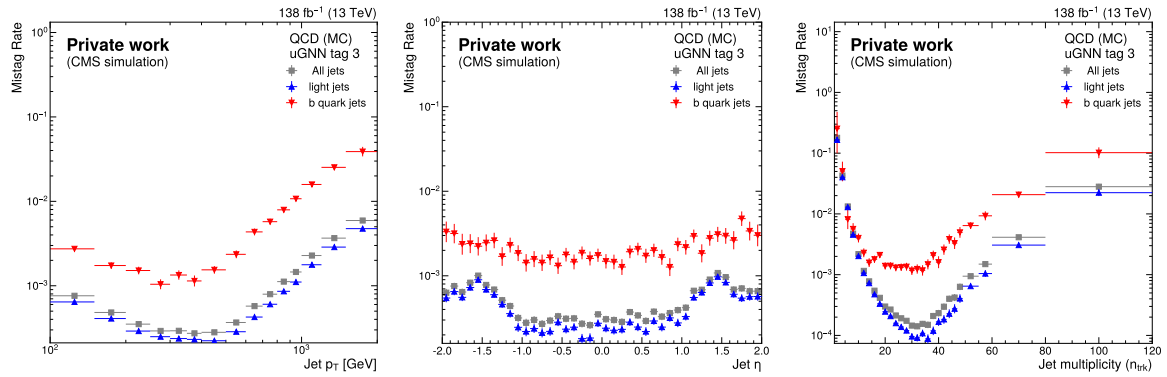


Figure B.19: Mistag rate of the “uGNN tag 3” tagger evaluated using QCD MC samples.

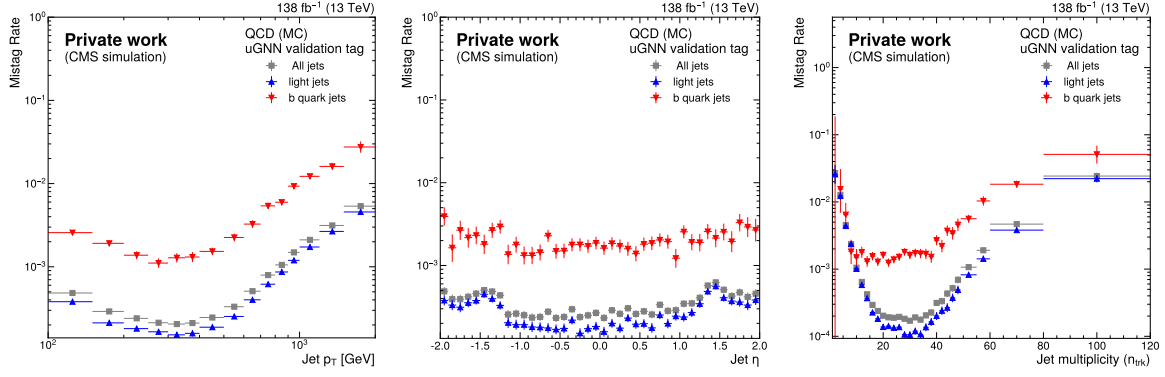


Figure B.20: Mistag rate of the “uGNN validation tag” tagger evaluated using QCD MC samples.

B.2 Estimated b jet fraction of the signal-free region

Here we display the estimated b jet fraction evaluated for the signal-free region using the template fit method described in Eq. 5.4. While the jet population is expected to change when we modify the R_{cut} and $\text{IP}_{z,\text{cut}}$ values used for jet–track association and track selection, The changes in jet population by adjusting the $\text{IP}_{z,\text{cut}}$ was so small that the values are well within the evaluated uncertainties. For this reason, we will only show the evaluation of the template fit for $R_{\text{cut}} = 0.4$ and $R_{\text{cut}} = 0.8$. We will also compare the template fit difference evaluated using γ +jets MC samples and using the γ -triggered data set. For the following plots in this section, we will keep the fit performed on the b-enhanced region on the left and the b-suppressed region on the right.

In addition to the individual fits, we will also show the estimated b jet fraction as a function of jet p_T for the various jet–track association schemes and data sets.

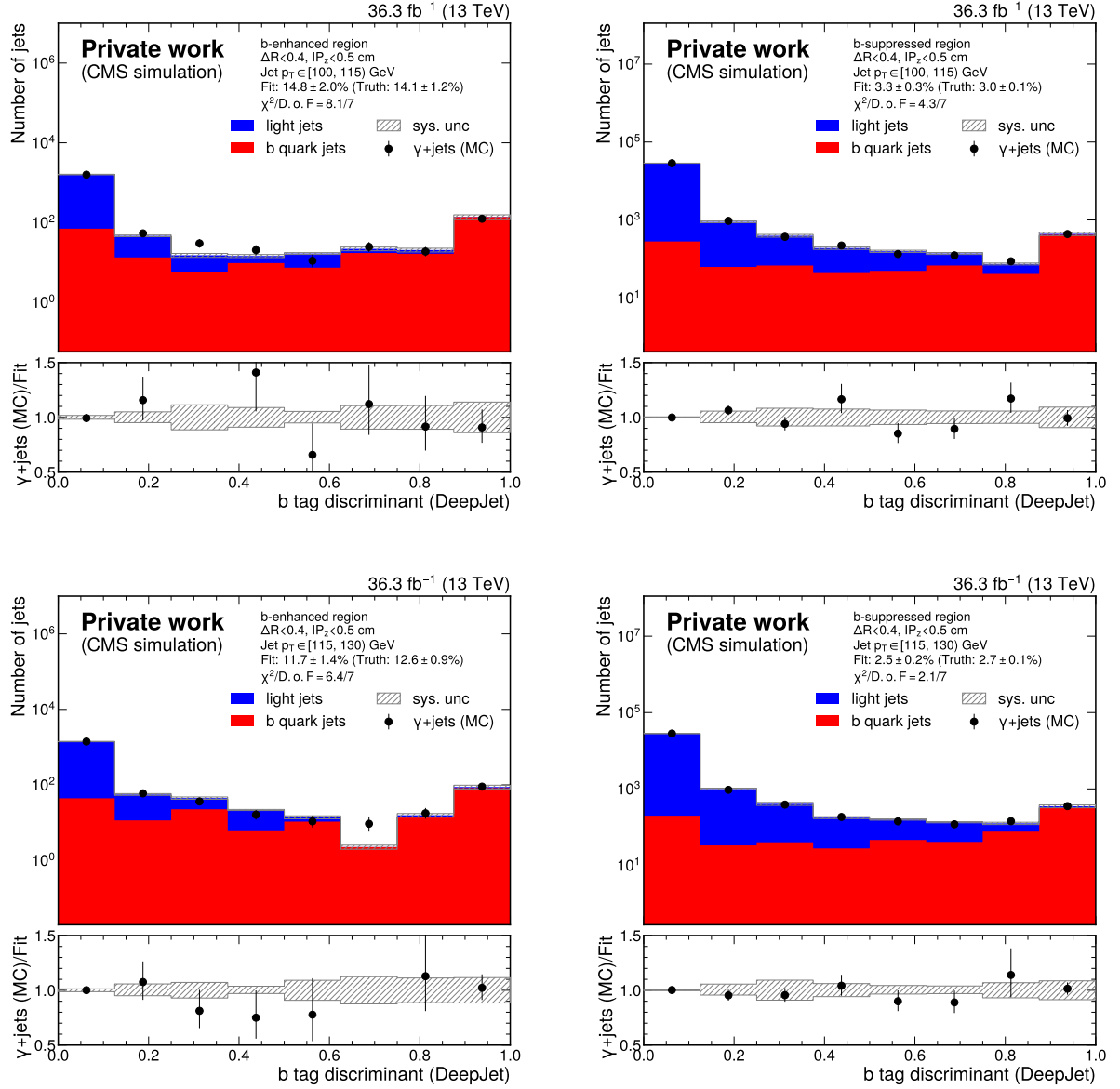


Figure B.21: The b jet fraction estimated for γ + jets MC samples (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

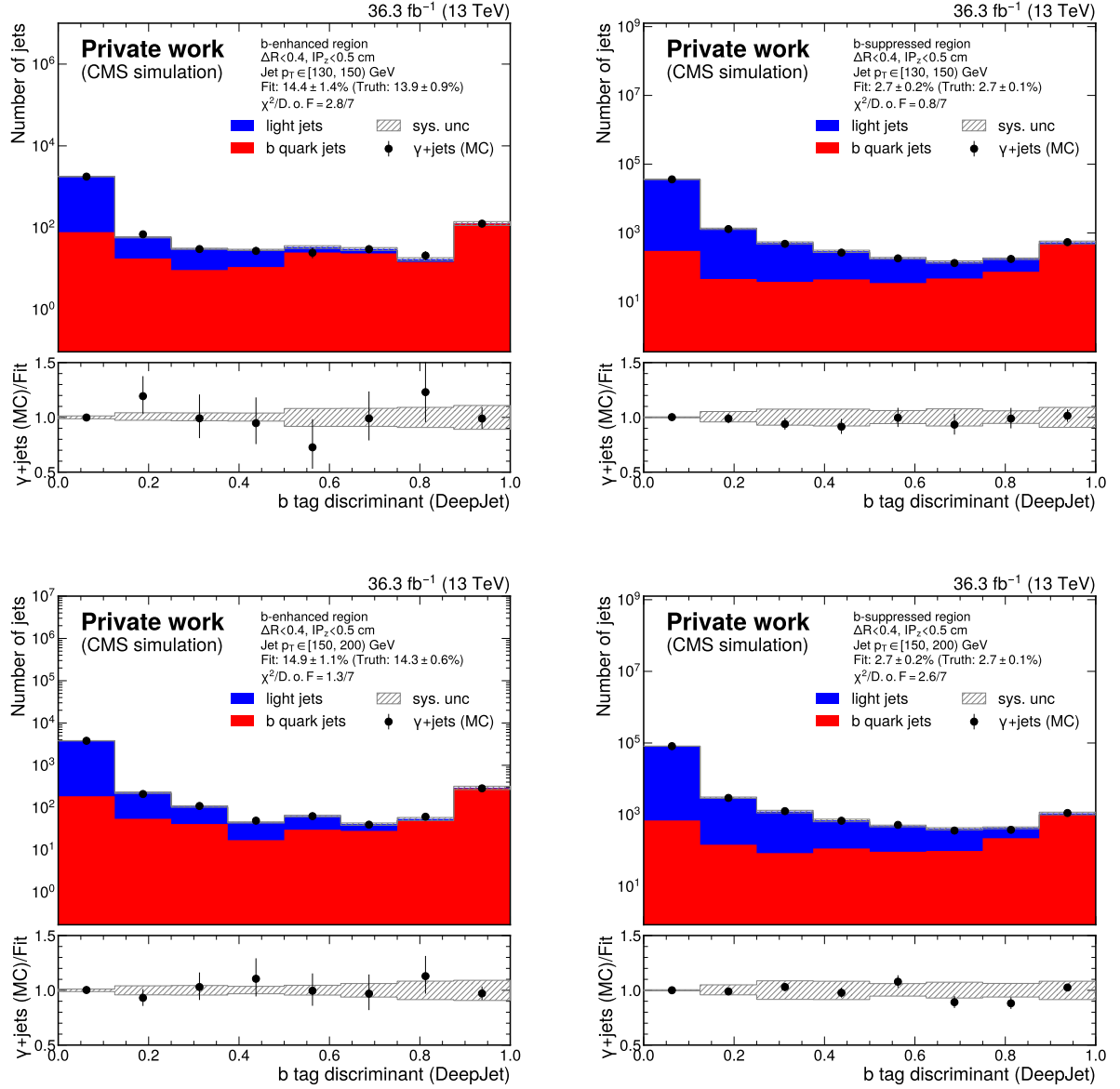


Figure B.22: The b jet fraction estimated for γ +jets MC samples (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

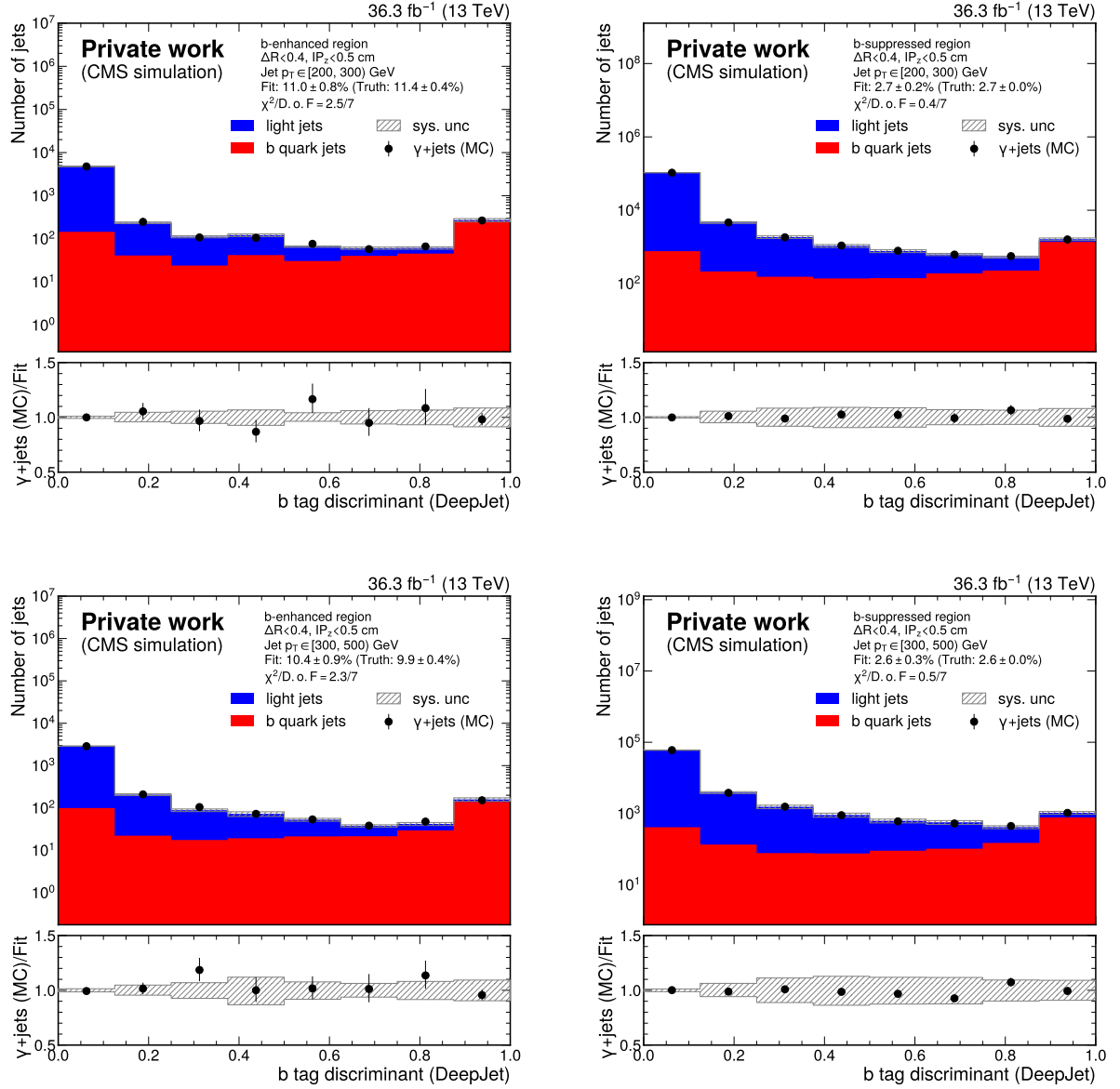


Figure B.23: The b jet fraction estimated for γ +jets MC samples (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

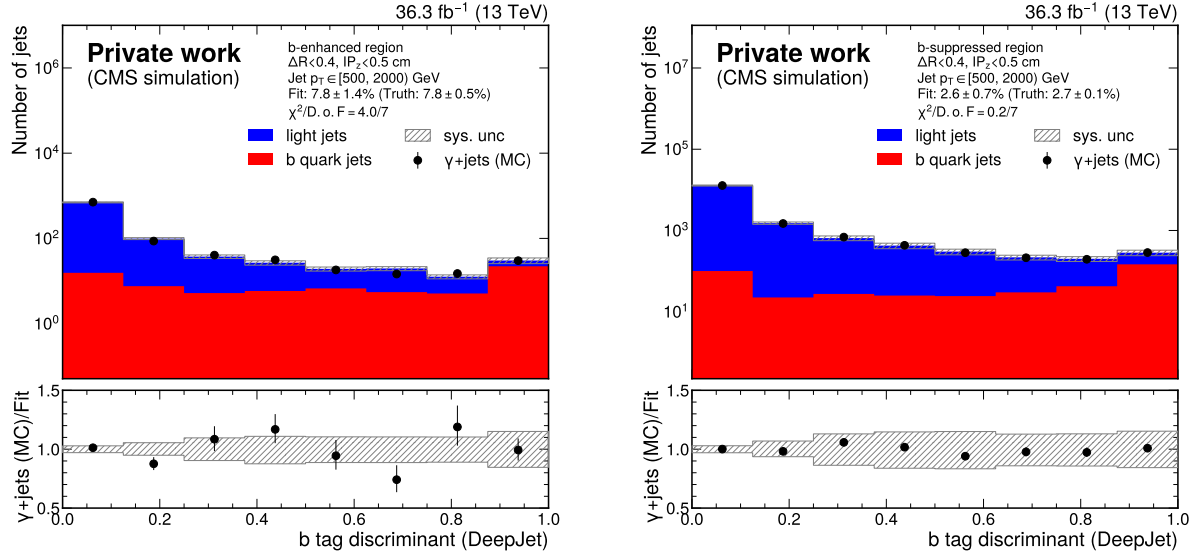


Figure B.24: The b jet fraction estimated for γ +jets MC samples (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

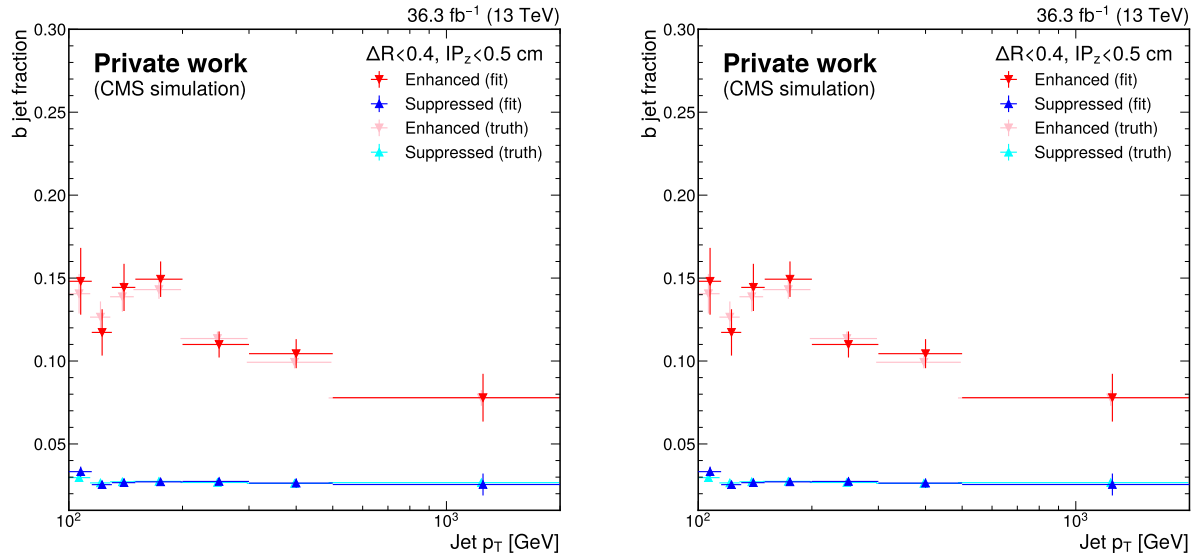


Figure B.25: The b jet fraction evaluated as a function of jet p_T estimated for γ +jets MC samples (pixel Phase-0) with a jet-track association of $R_{\text{cut}} = 0.4$.

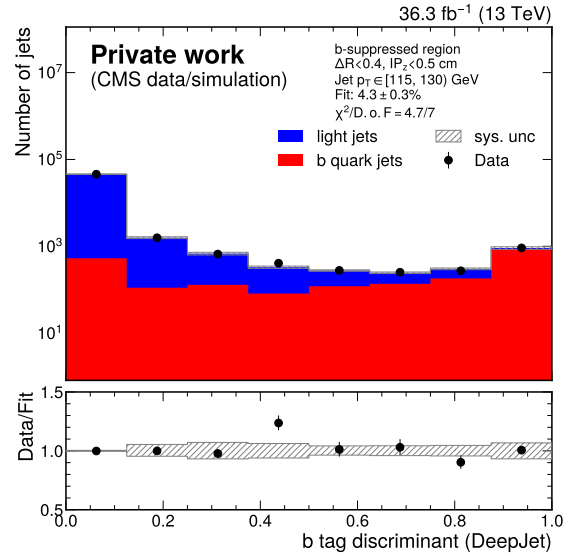
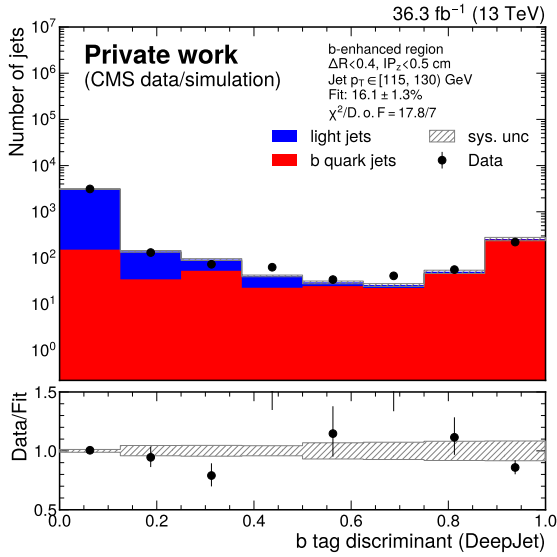
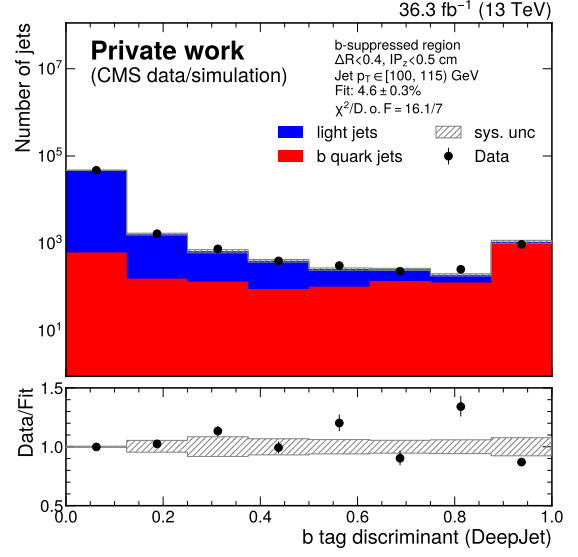
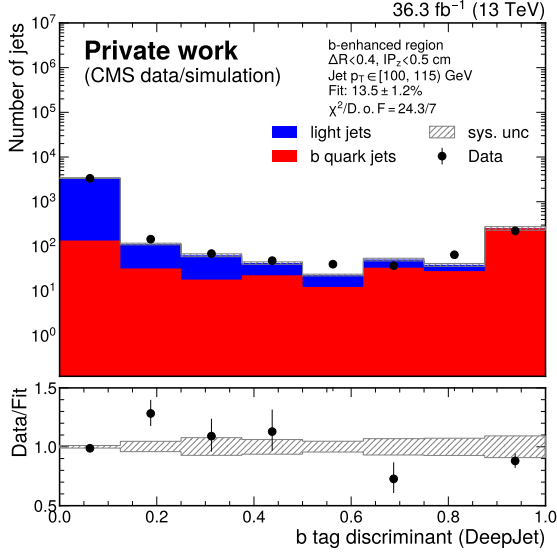


Figure B.26: The b jet fraction estimated for γ -trigger data set (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

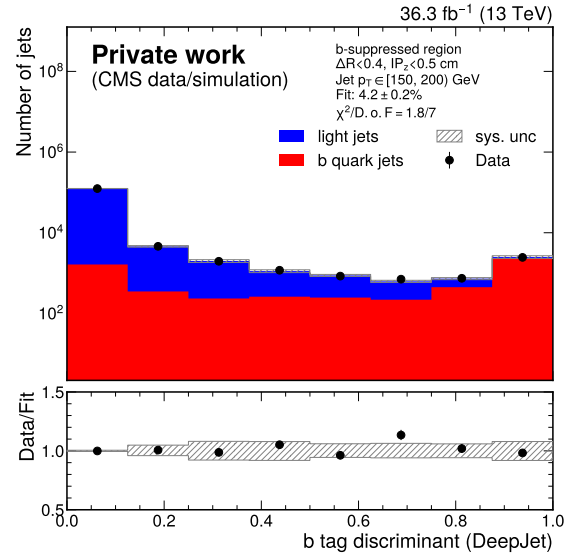
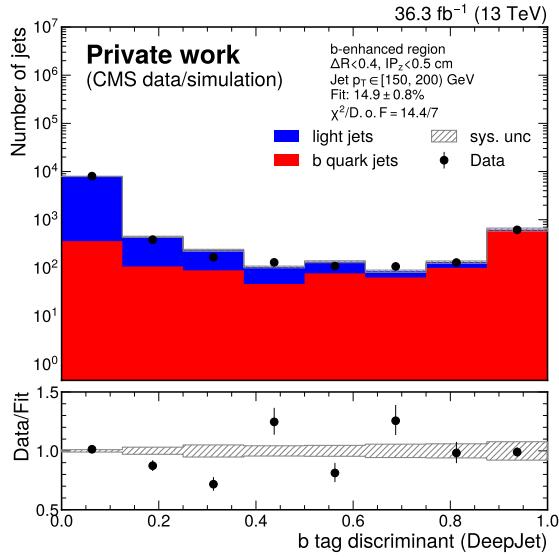
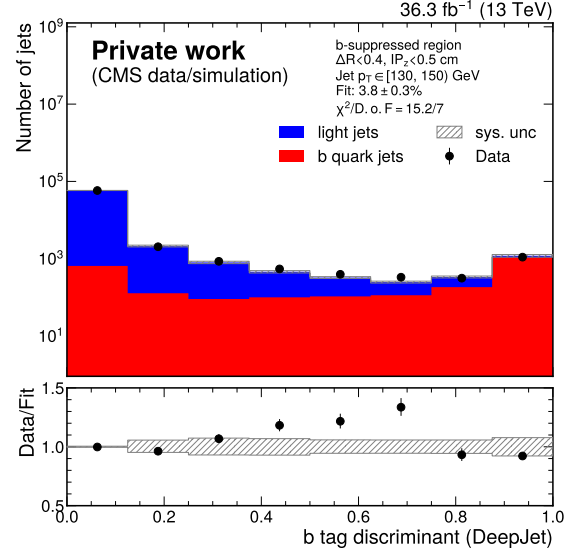
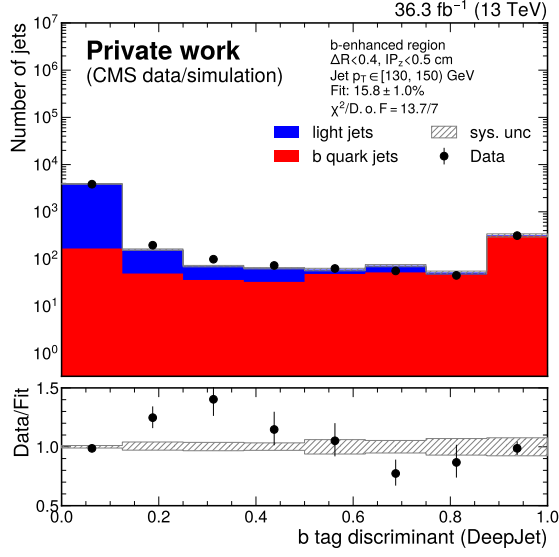


Figure B.27: The b jet fraction estimated for γ -trigger data set (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

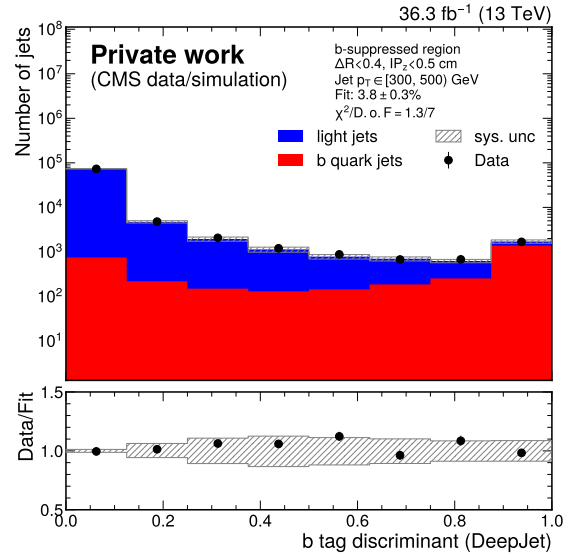
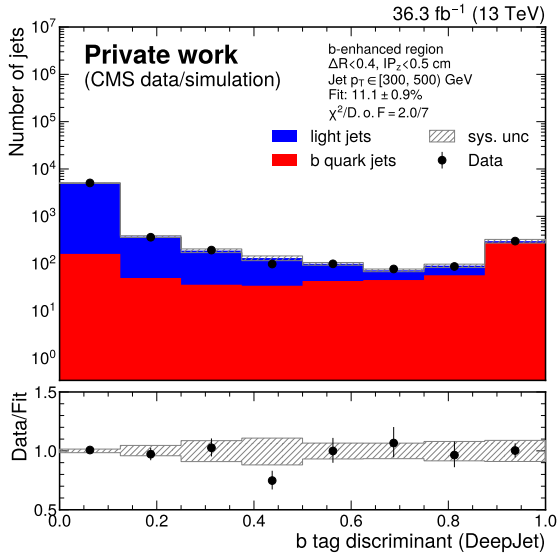
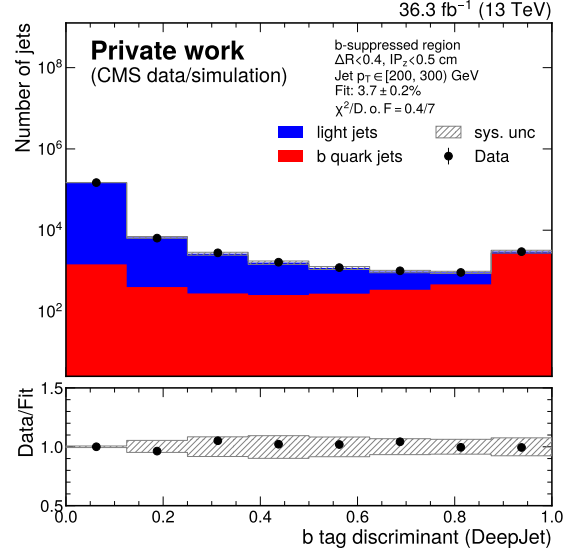
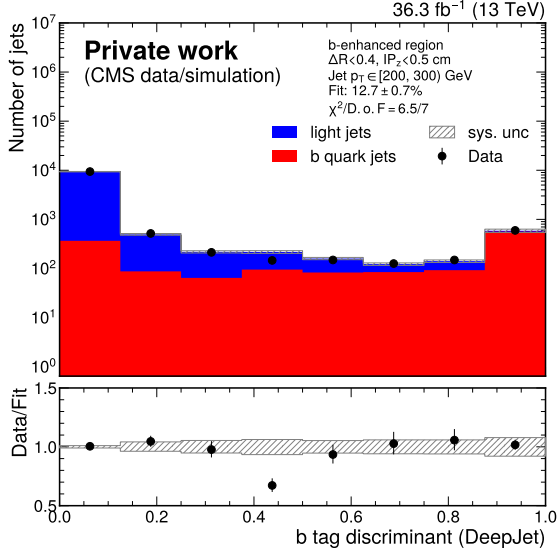


Figure B.28: The b jet fraction estimated for γ -trigger data set (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

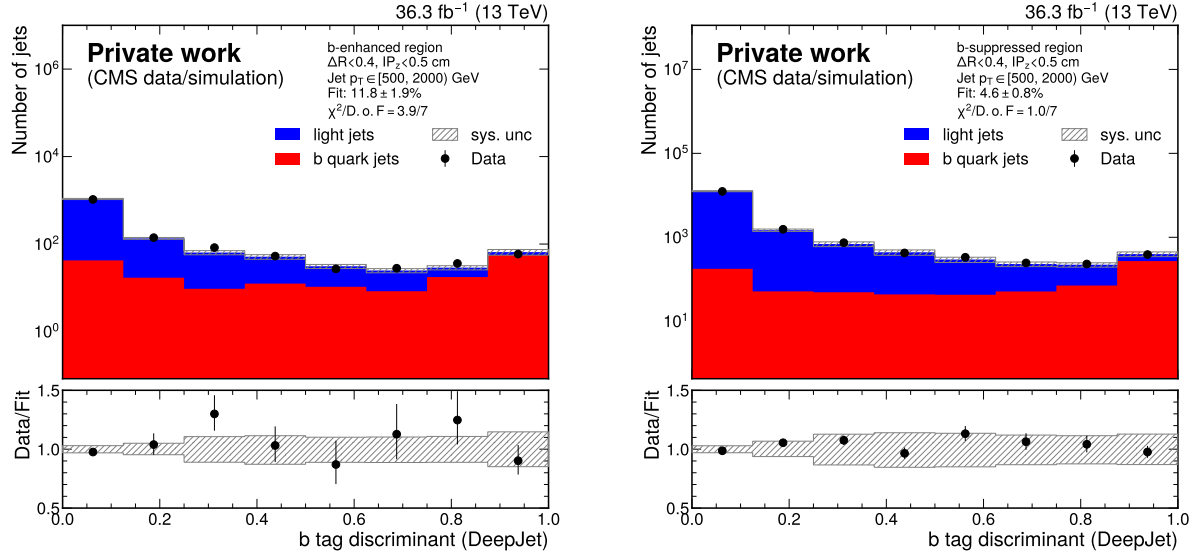


Figure B.29: The b jet fraction estimated for γ -trigger data set (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

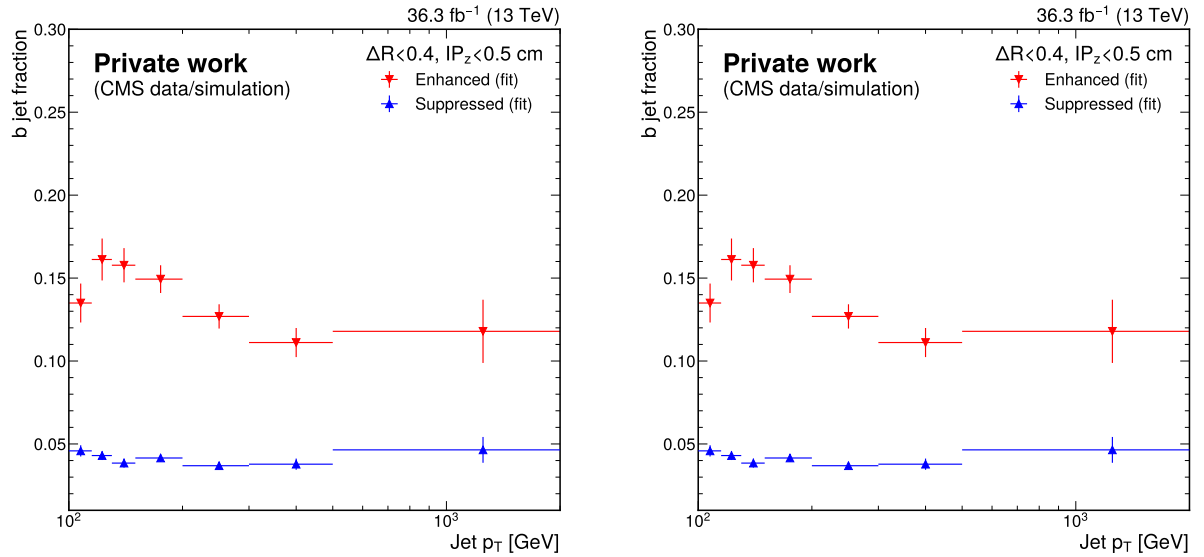


Figure B.30: The b jet fraction evaluated as a function of jet p_T estimated for γ -trigger data set (pixel Phase-0) with a jet-track association of $R_{\text{cut}} = 0.4$.

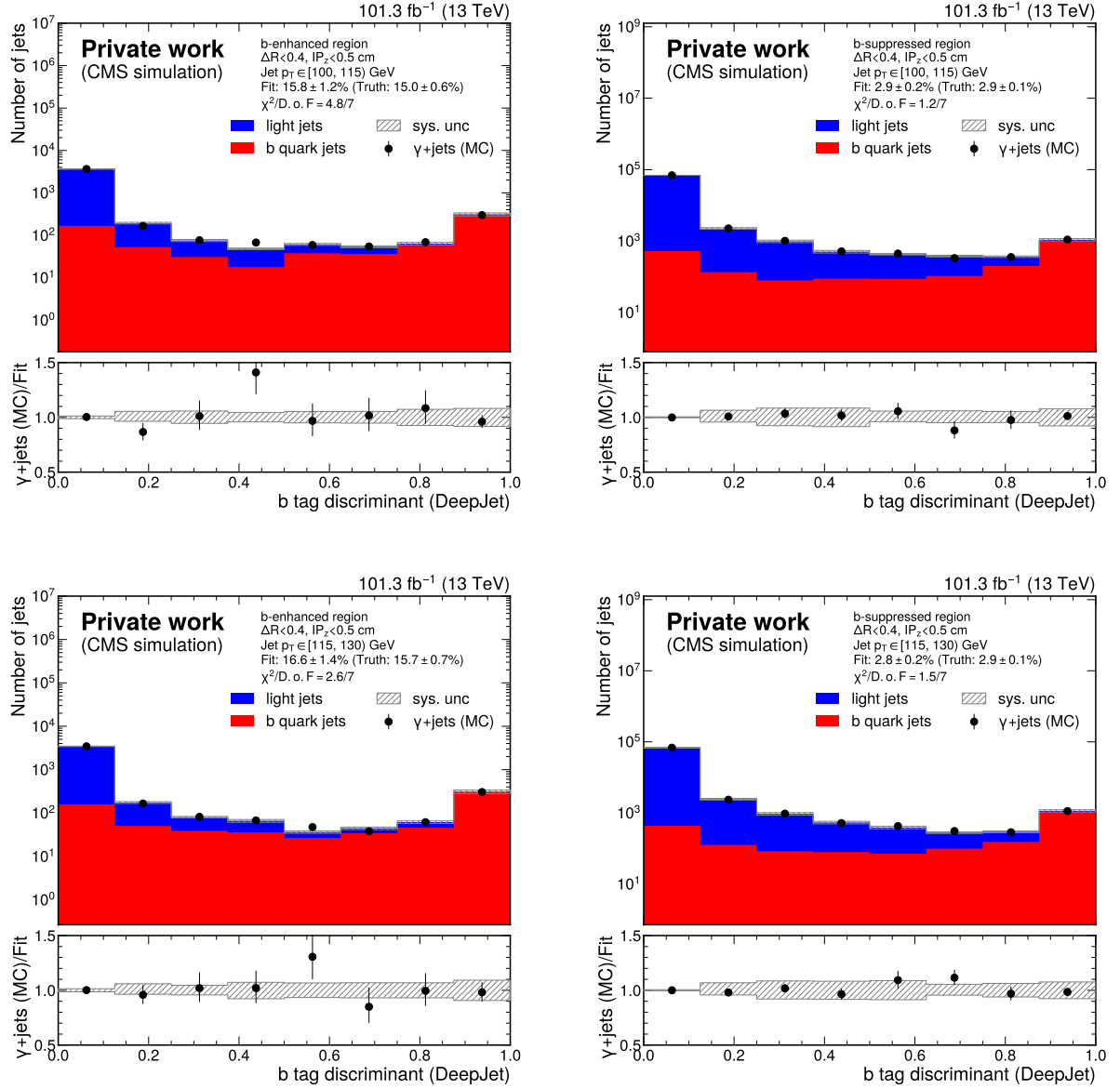


Figure B.31: The b jet fraction estimated for γ +jets MC samples (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

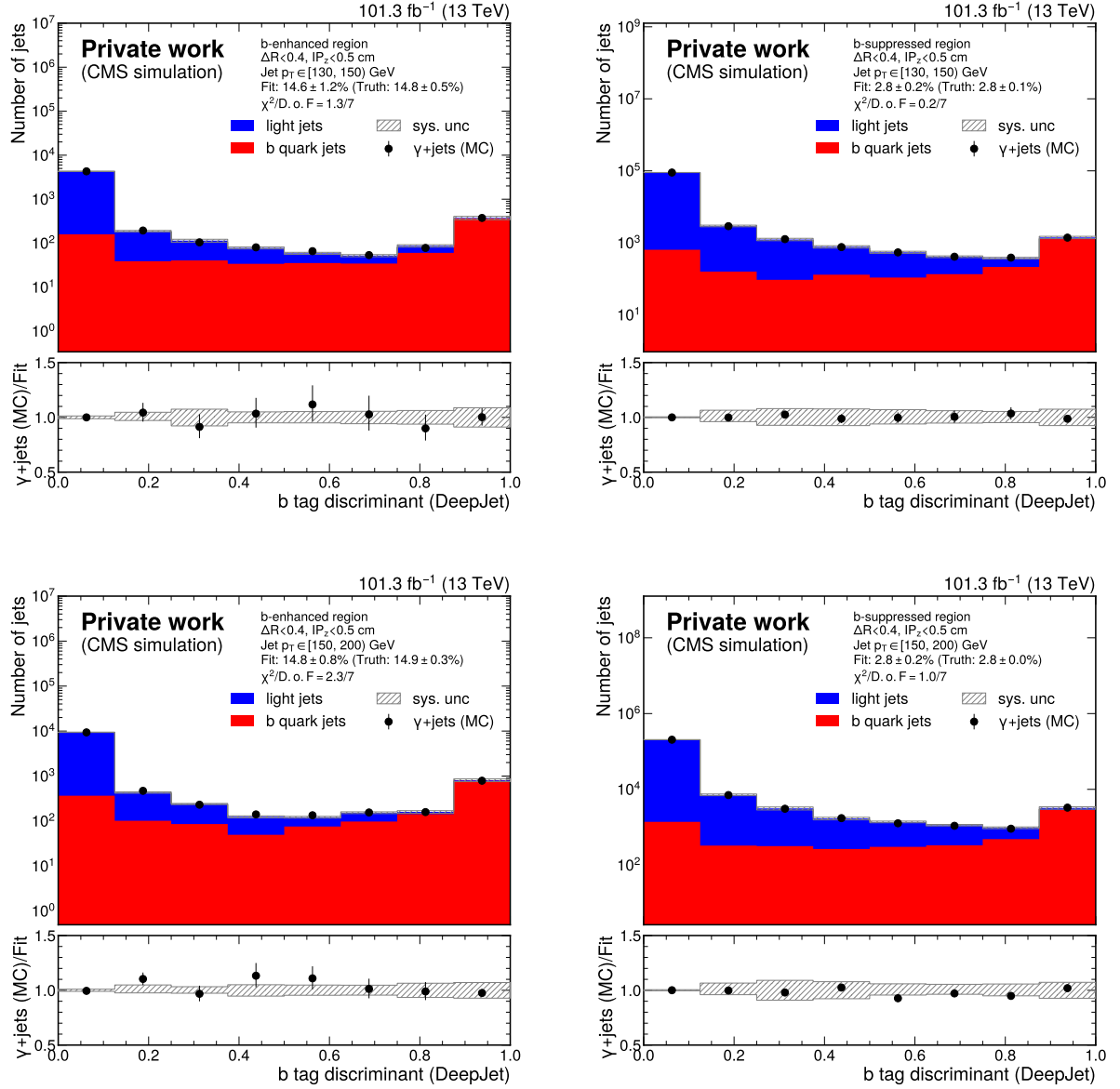


Figure B.32: The b jet fraction estimated for γ +jets MC samples (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

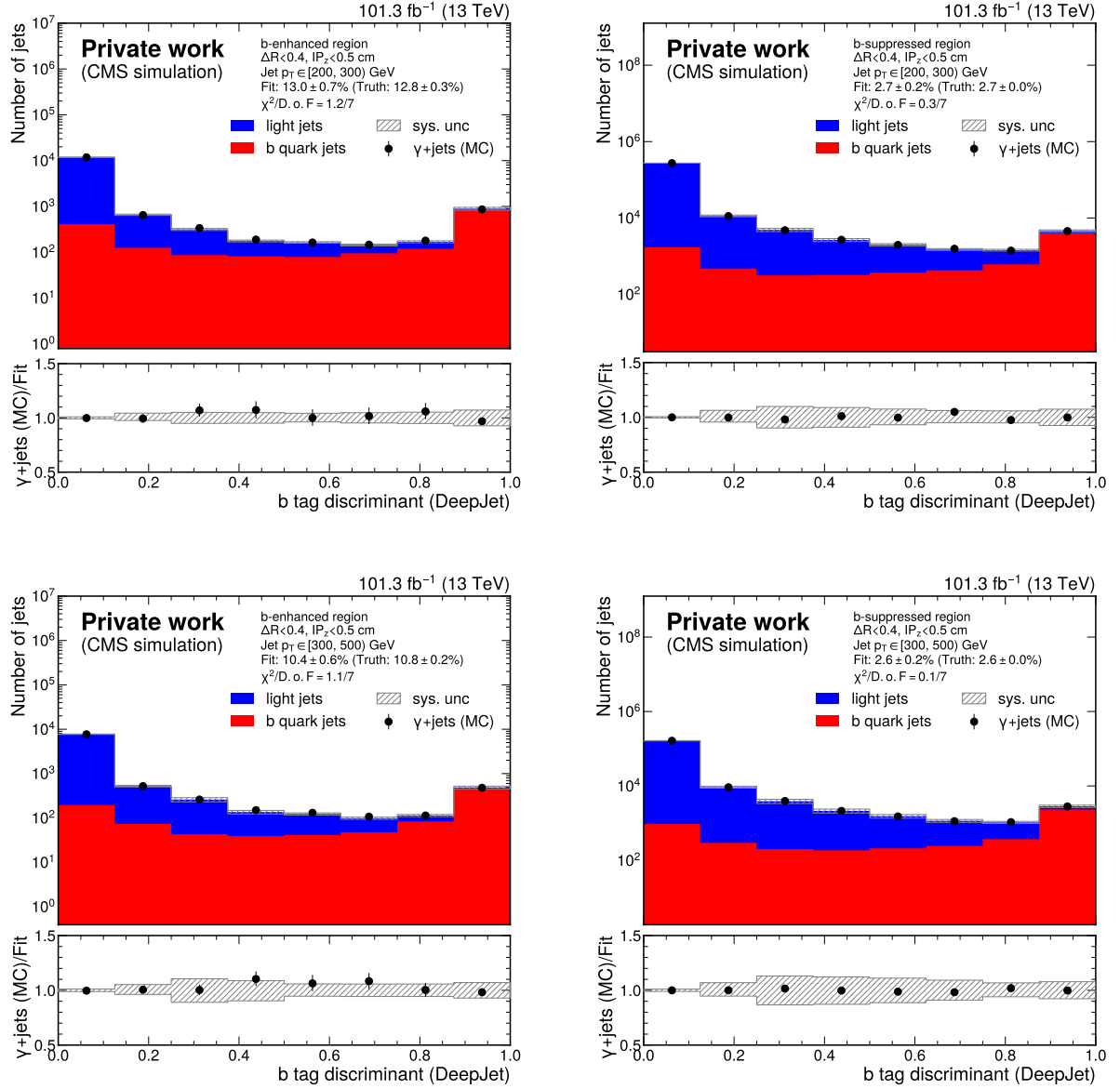


Figure B.33: The b jet fraction estimated for γ +jets MC samples (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

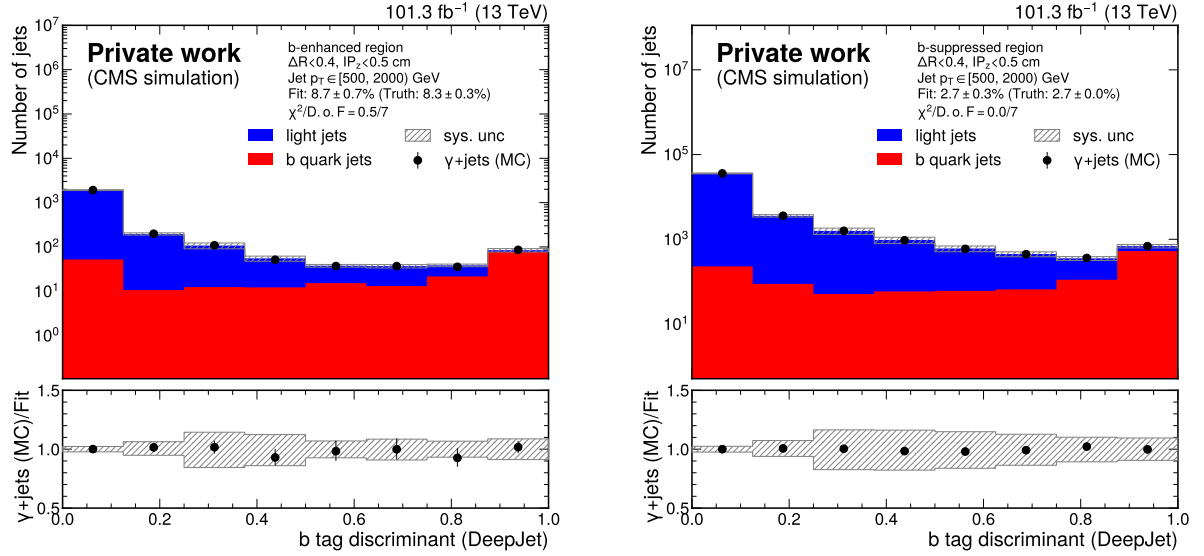


Figure B.34: The b jet fraction estimated for γ +jets MC samples (pixel Phase-1) in various bins of jet p_T with a jet–track association of $R_{\text{cut}} = 0.4$

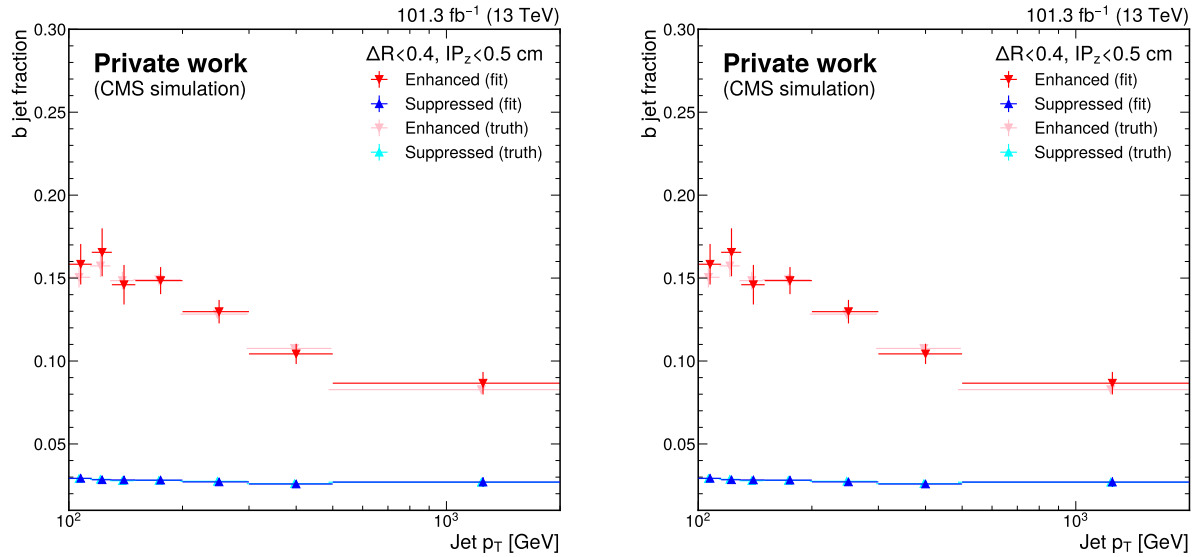


Figure B.35: The b jet fraction evaluated as a function of jet p_T estimated for γ +jets MC samples (pixel Phase-1) with a jet–track association of $R_{\text{cut}} = 0.4$.

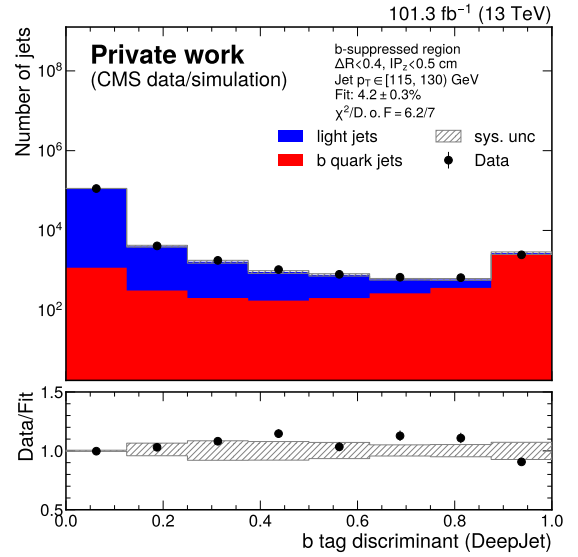
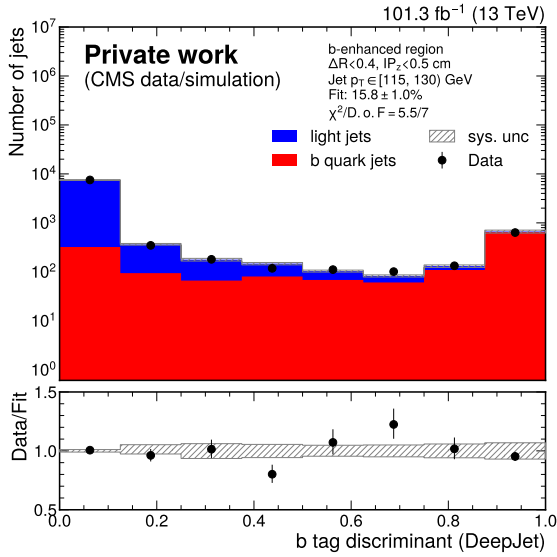
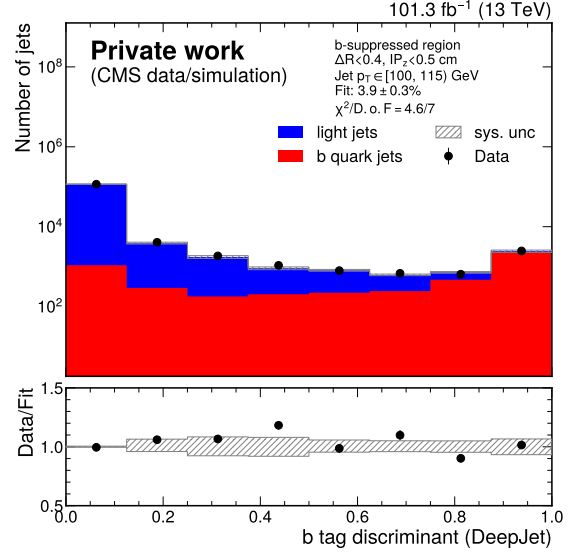
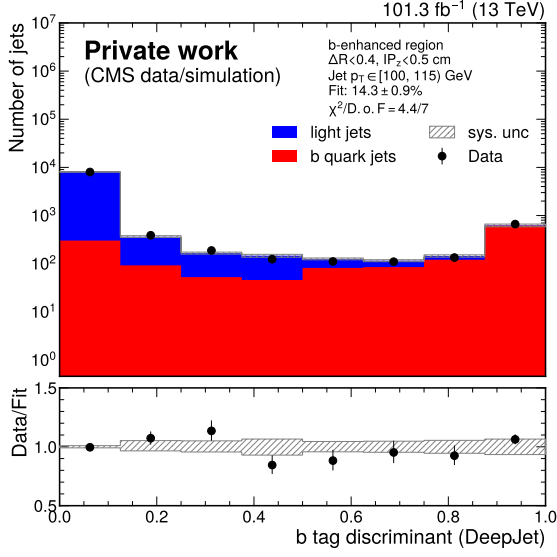


Figure B.36: The b jet fraction estimated for γ -trigger data set (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

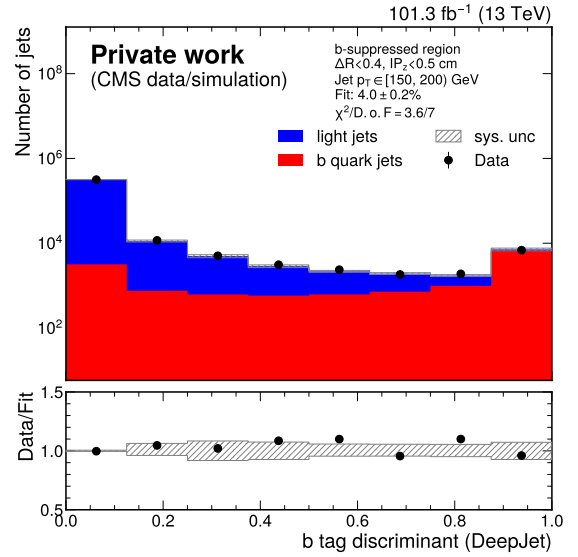
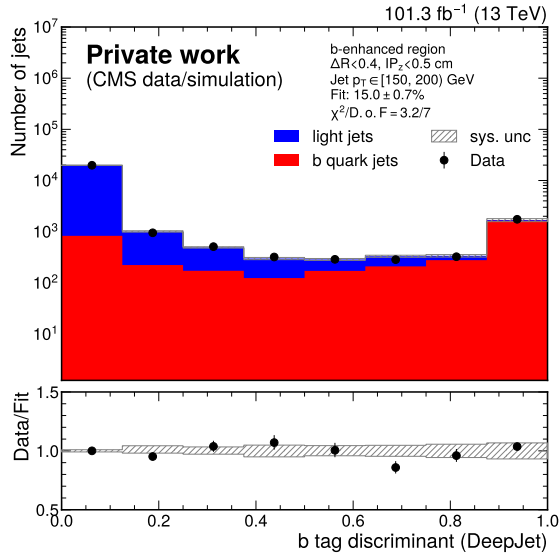
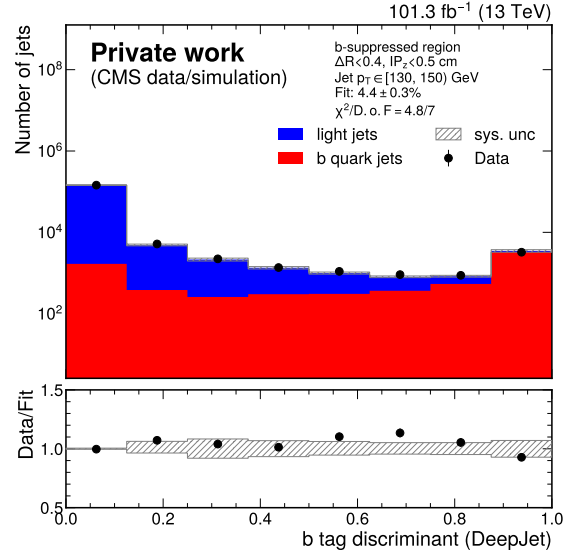
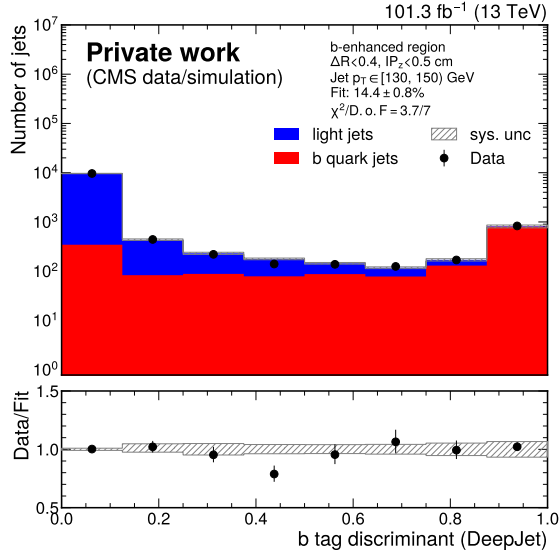


Figure B.37: The b jet fraction estimated for γ -trigger data set (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

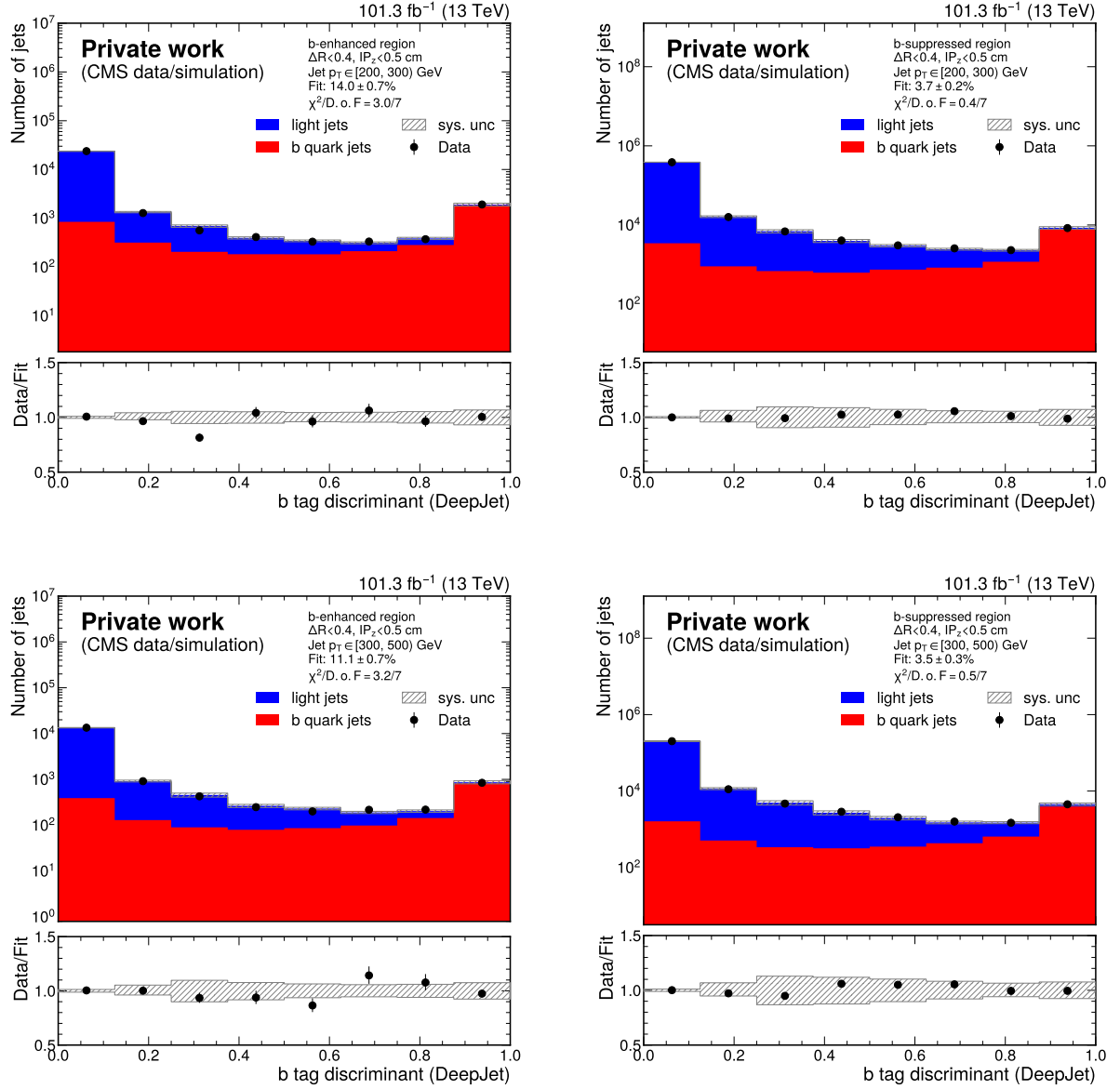


Figure B.38: The b jet fraction estimated for γ -trigger data set (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.4$

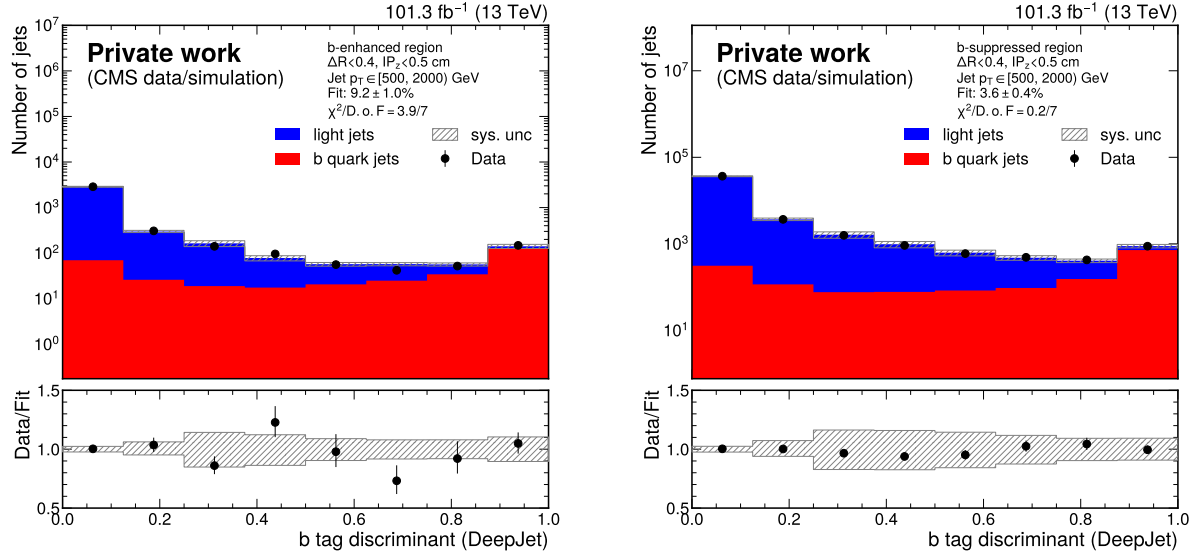


Figure B.39: The b jet fraction estimated for γ -trigger data set (pixel Phase-1) in various bins of jet p_T with a jet–track association of $R_{\text{cut}} = 0.4$

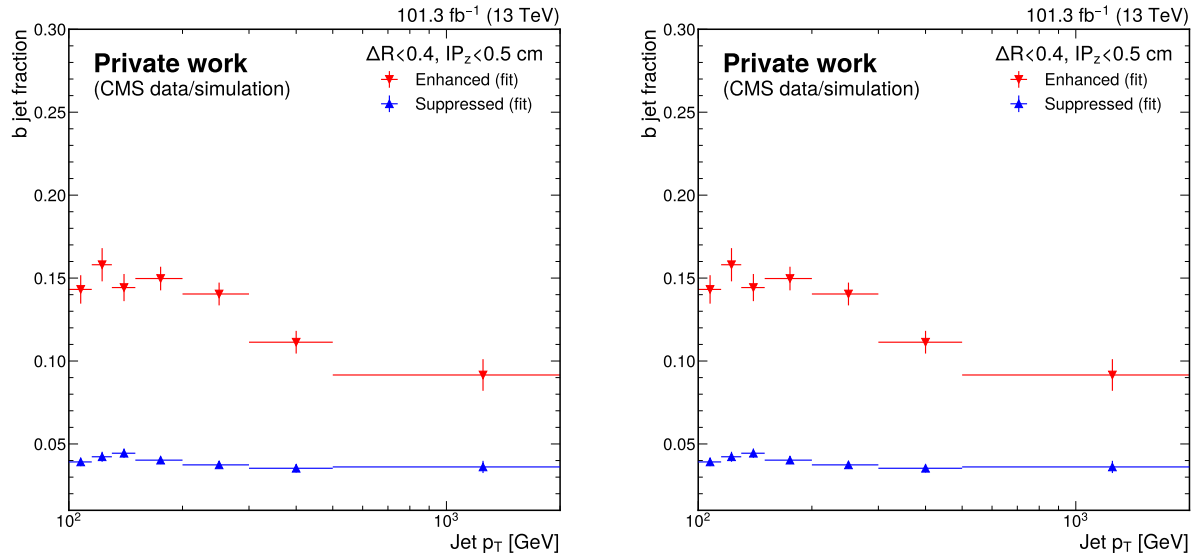


Figure B.40: The b jet fraction evaluated as a function of jet p_T estimated for γ -trigger data set (pixel Phase-1) with a jet–track association of $R_{\text{cut}} = 0.4$.

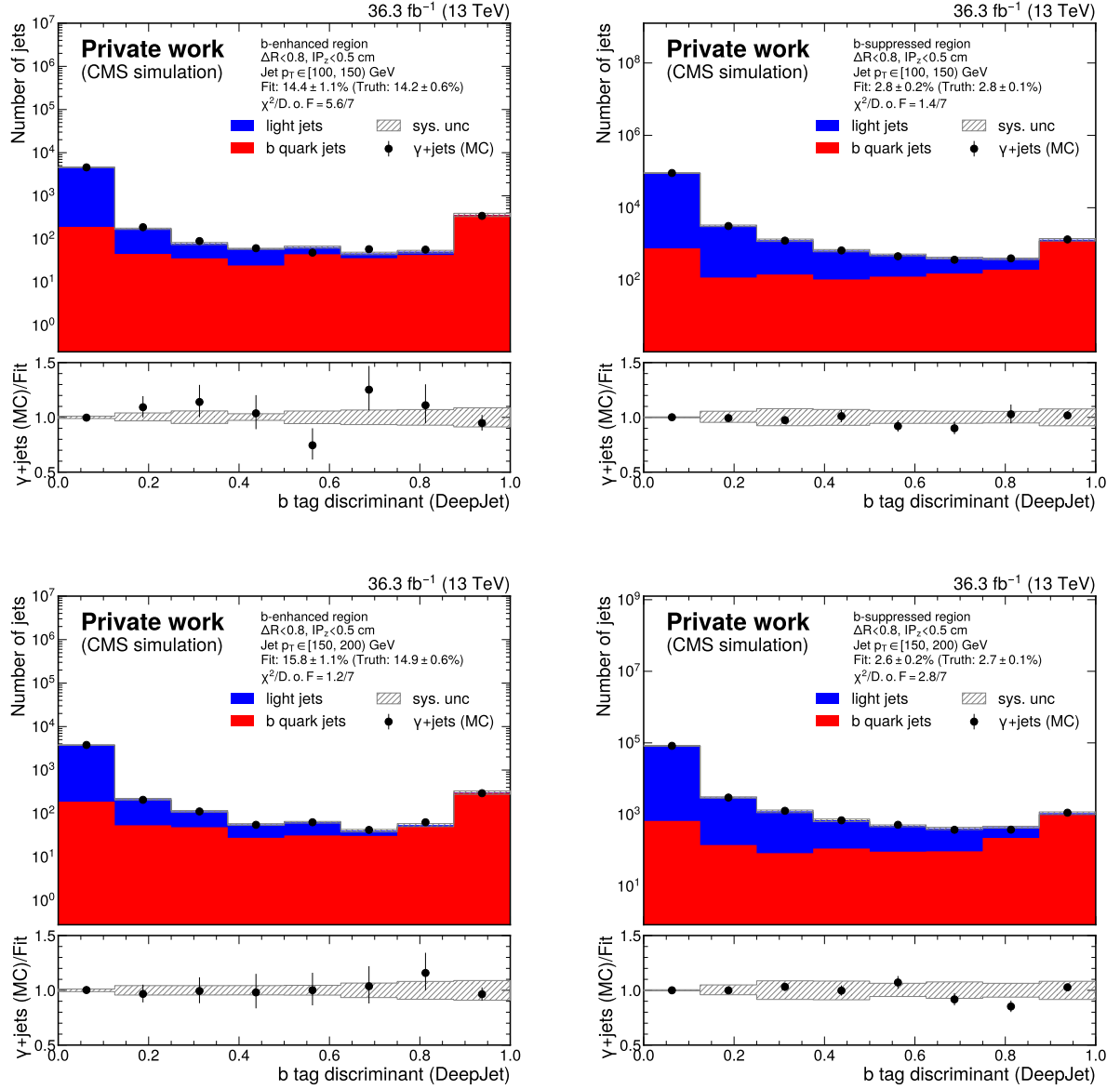


Figure B.41: The b jet fraction estimated for γ +jets MC samples (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

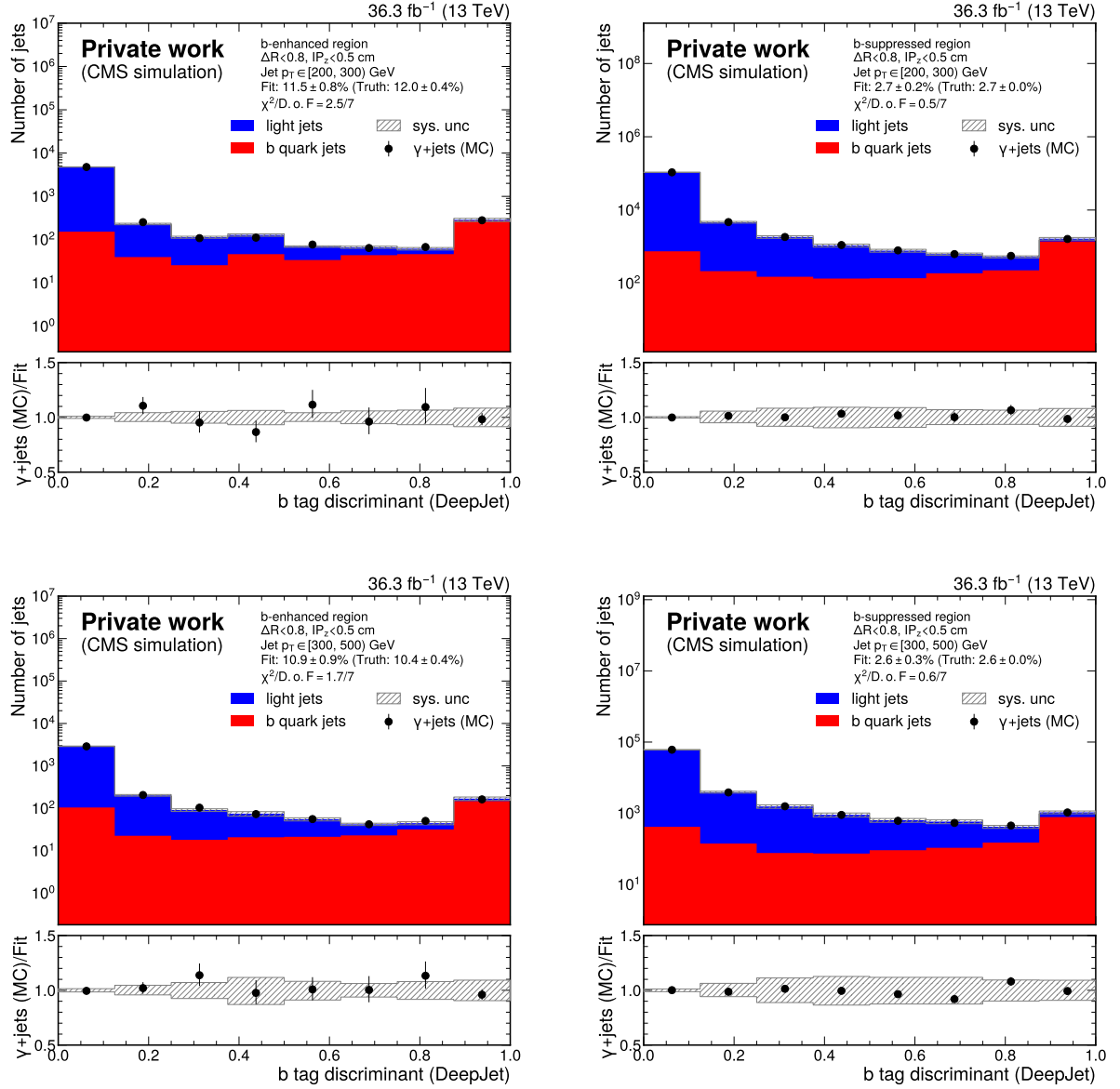


Figure B.42: The b jet fraction estimated for γ +jets MC samples (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

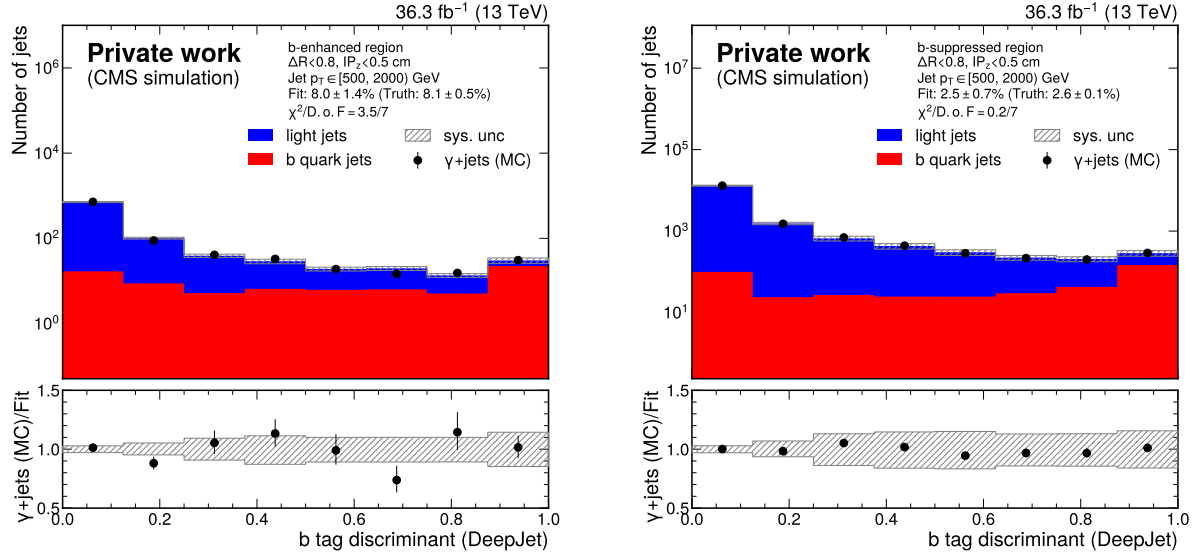


Figure B.43: The b jet fraction estimated for γ +jets MC samples (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

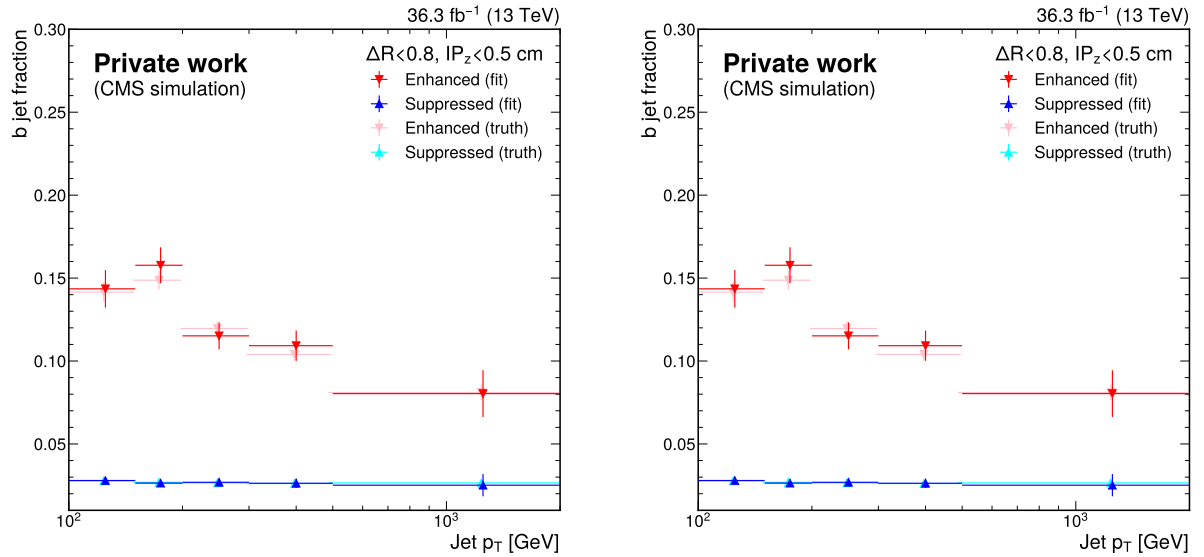


Figure B.44: The b jet fraction evaluated as a function of jet p_T estimated for γ +jets MC samples (pixel Phase-0) with a jet-track association of $R_{\text{cut}} = 0.8$.

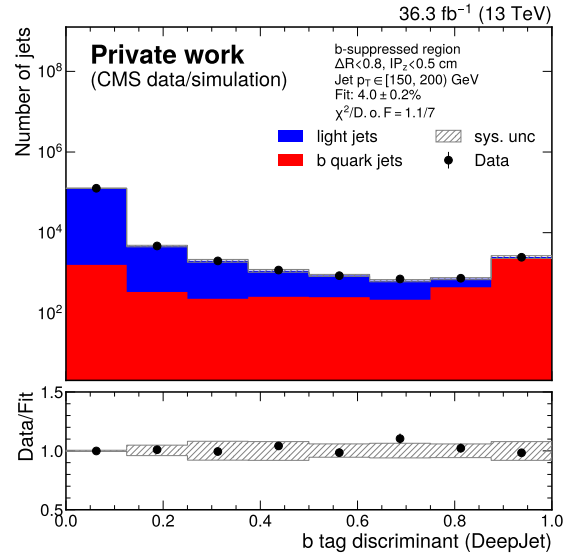
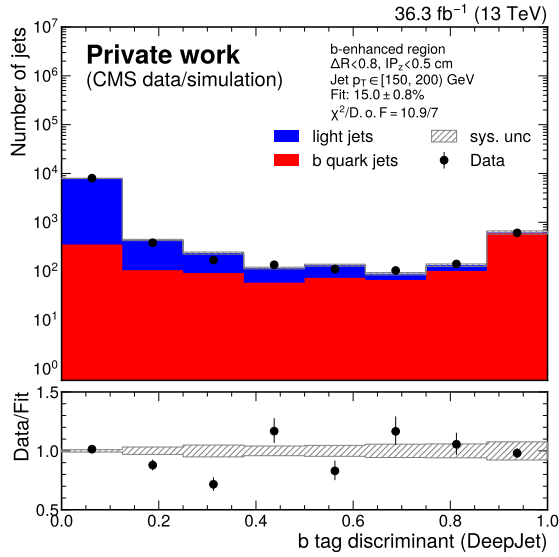
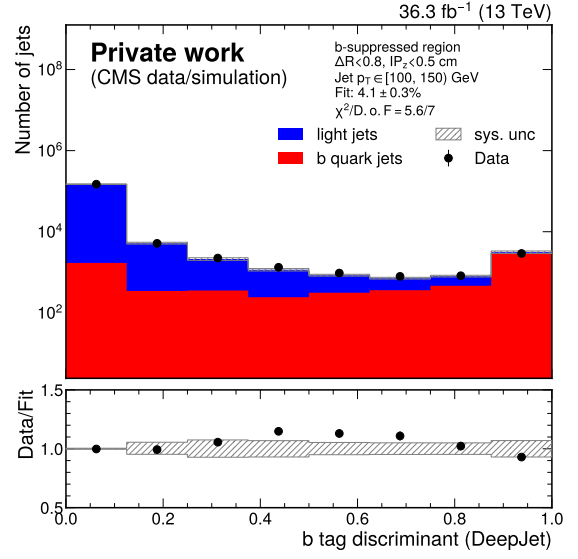
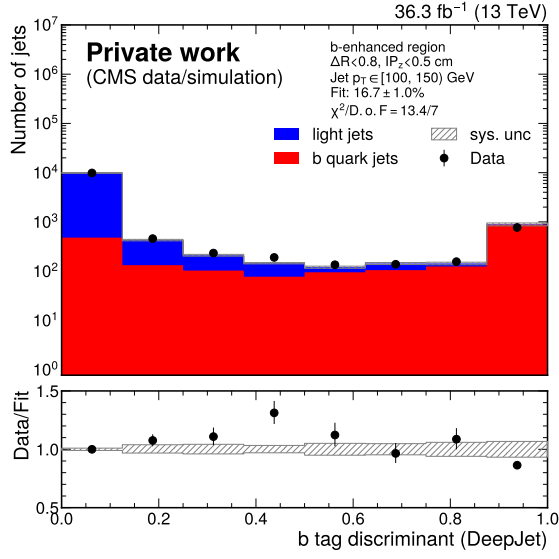


Figure B.45: The b jet fraction estimated for γ -trigger data set (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

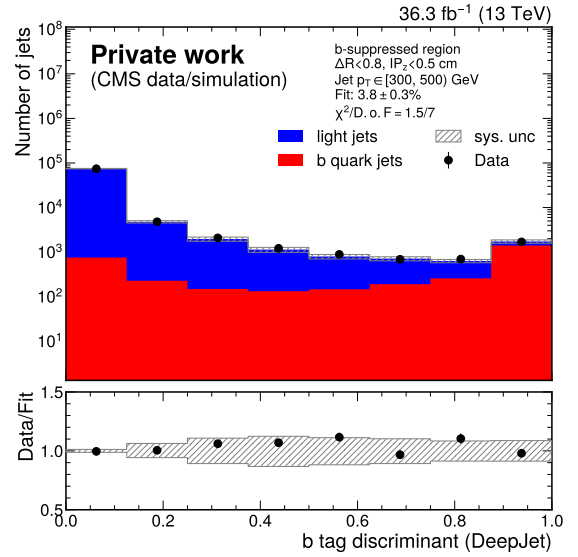
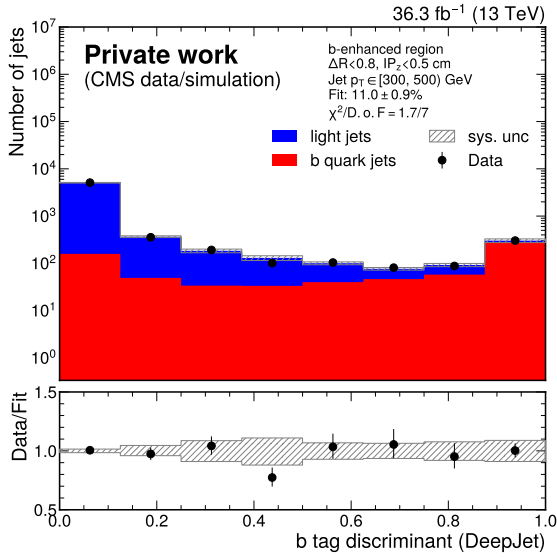
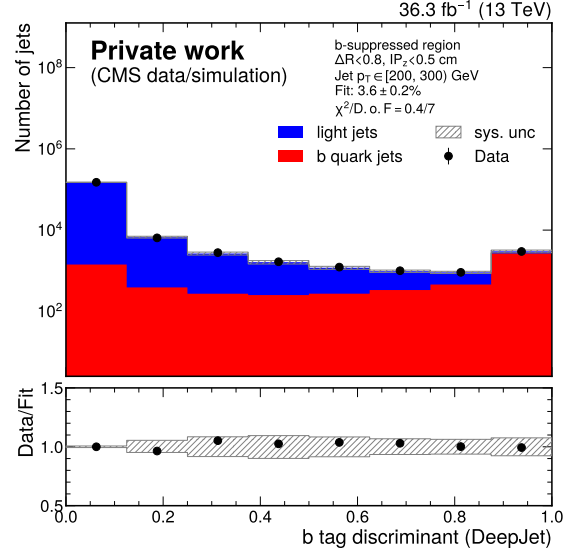
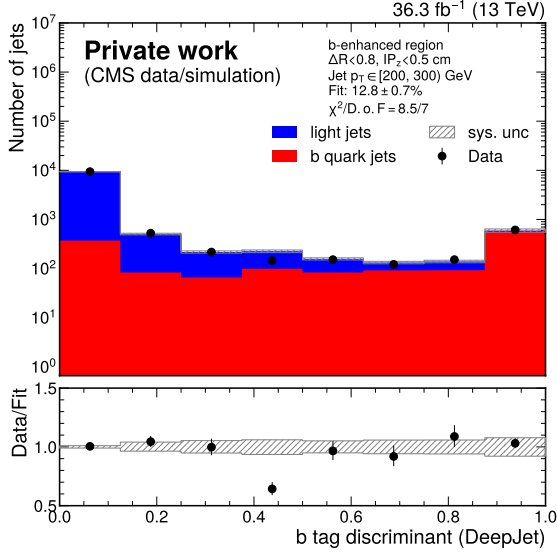


Figure B.46: The b jet fraction estimated for γ -trigger data set (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

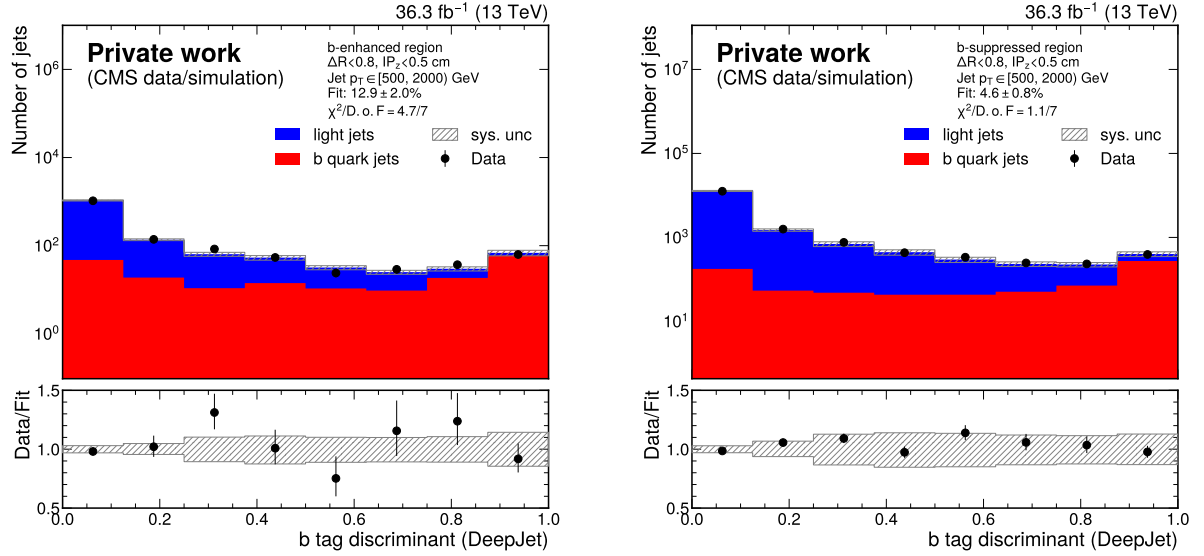


Figure B.47: The b jet fraction estimated for γ -trigger data set (pixel Phase-0) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

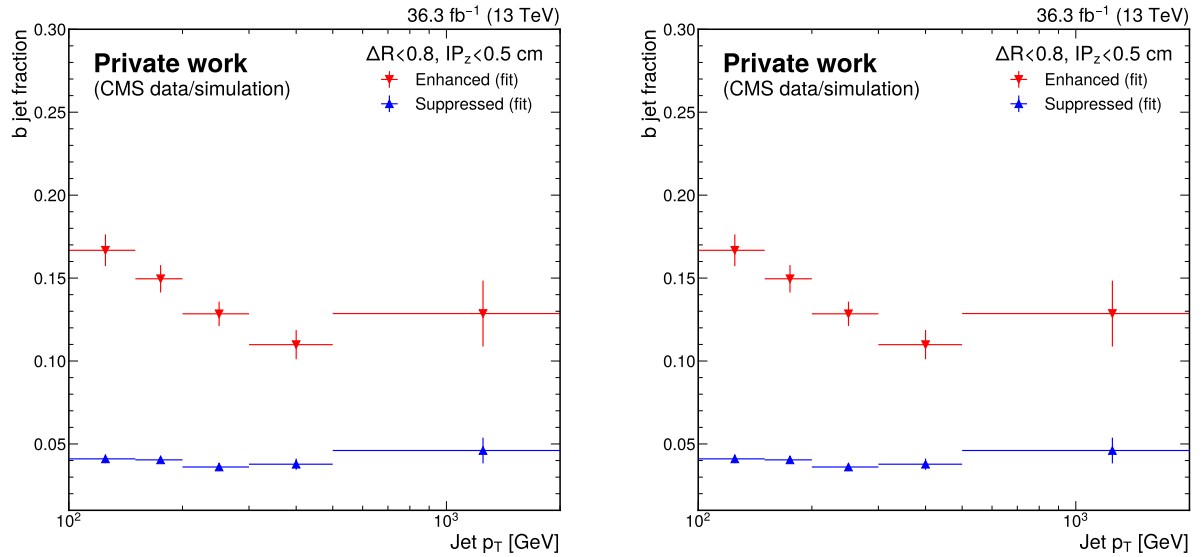


Figure B.48: The b jet fraction evaluated as a function of jet p_T estimated for γ -trigger data set (pixel Phase-0) with a jet-track association of $R_{\text{cut}} = 0.8$.

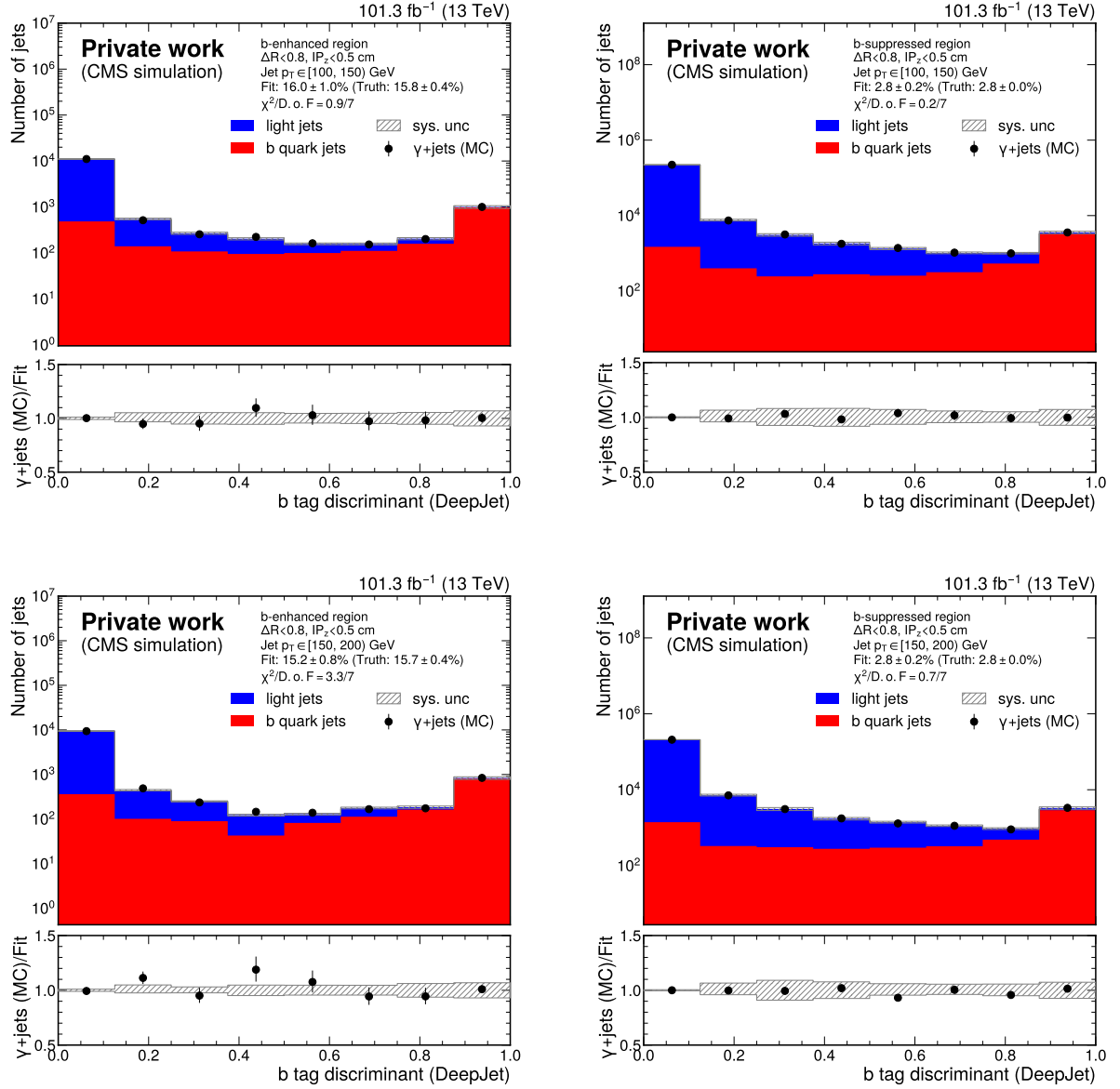


Figure B.49: The b jet fraction estimated for γ +jets MC samples (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

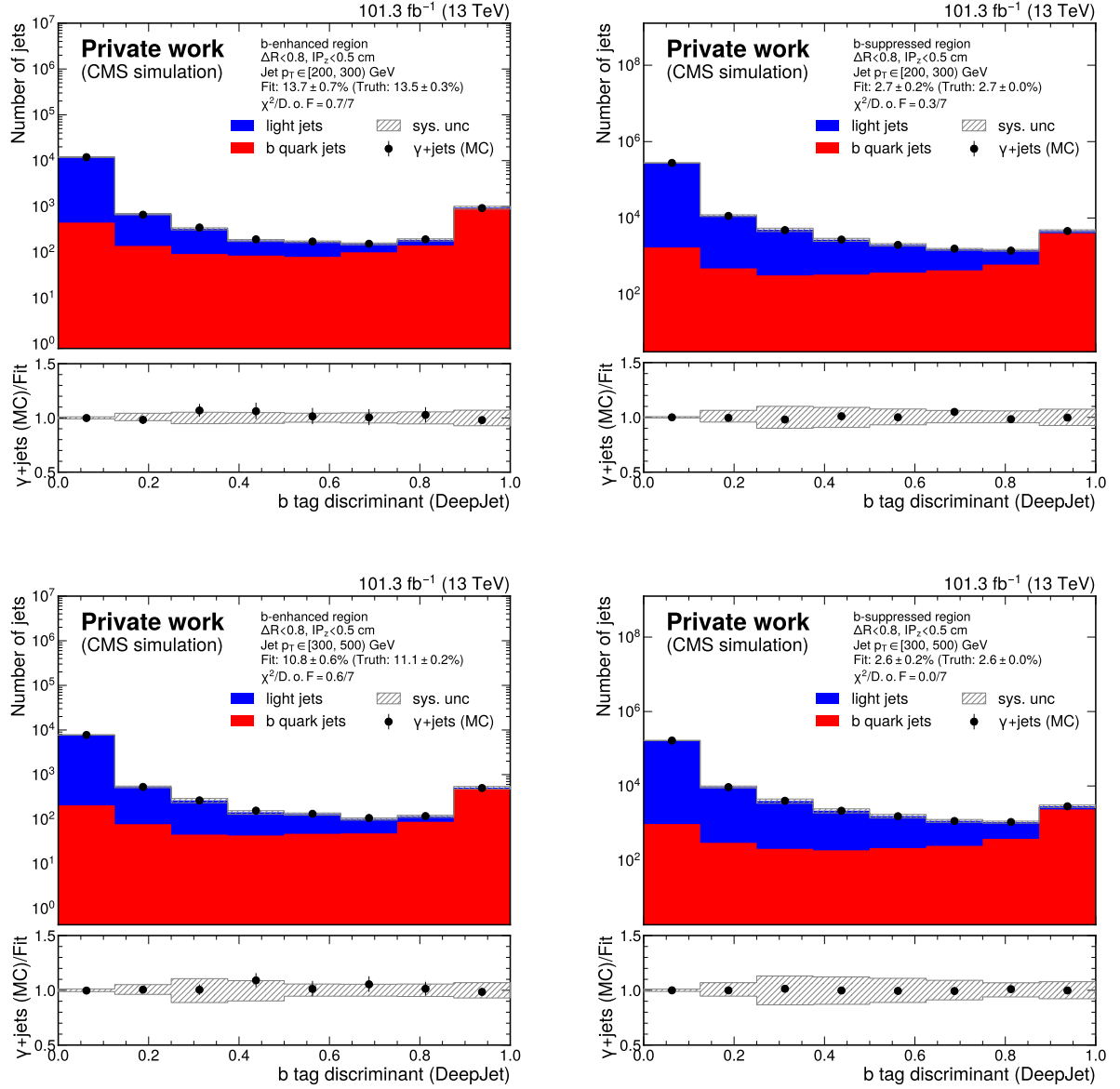


Figure B.50: The b jet fraction estimated for γ +jets MC samples (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

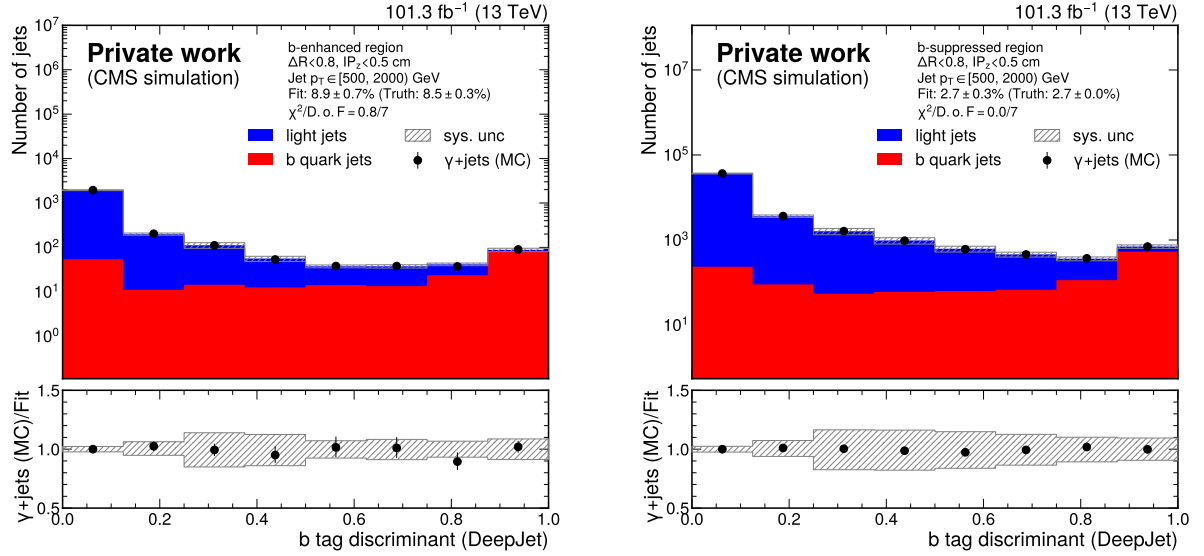


Figure B.51: The b jet fraction estimated for γ +jets MC samples (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

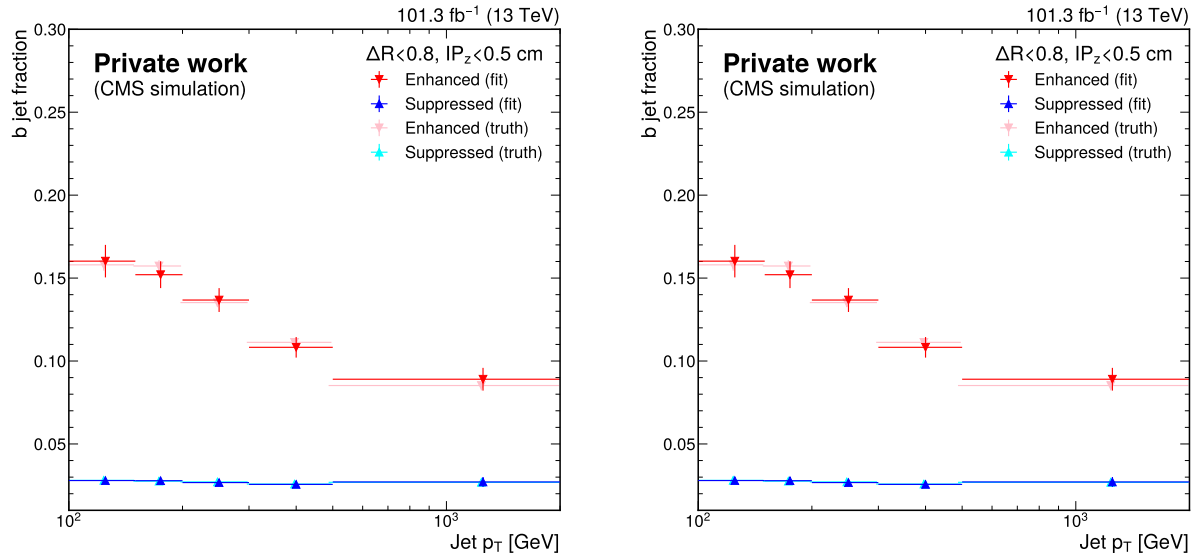


Figure B.52: The b jet fraction evaluated as a function of jet p_T estimated for γ +jets MC samples (pixel Phase-1) with a jet-track association of $R_{\text{cut}} = 0.8$.

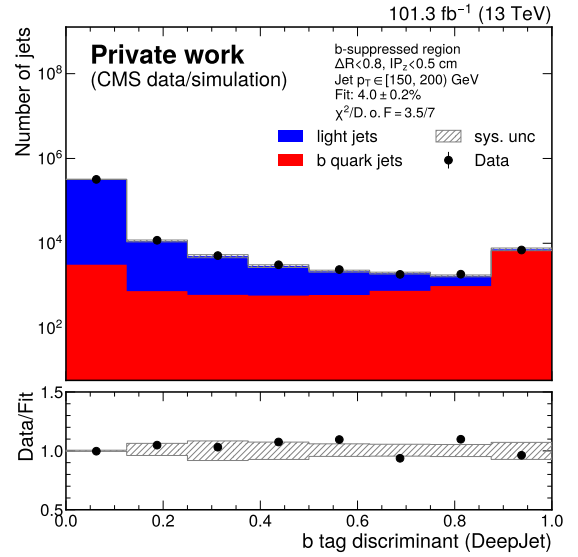
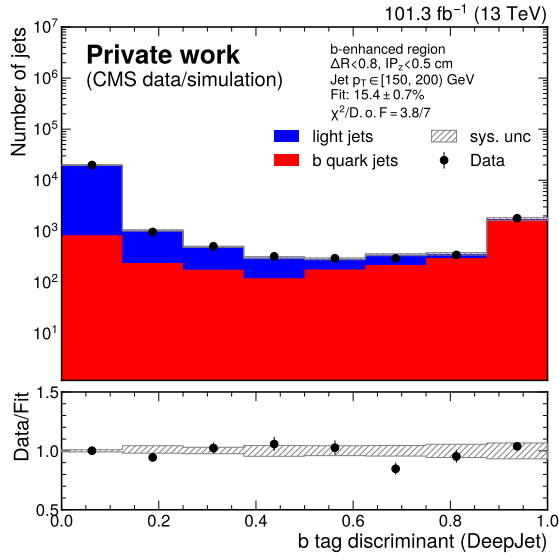
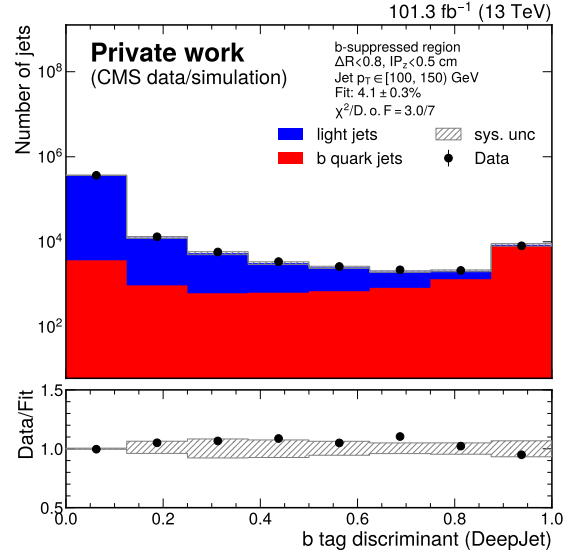
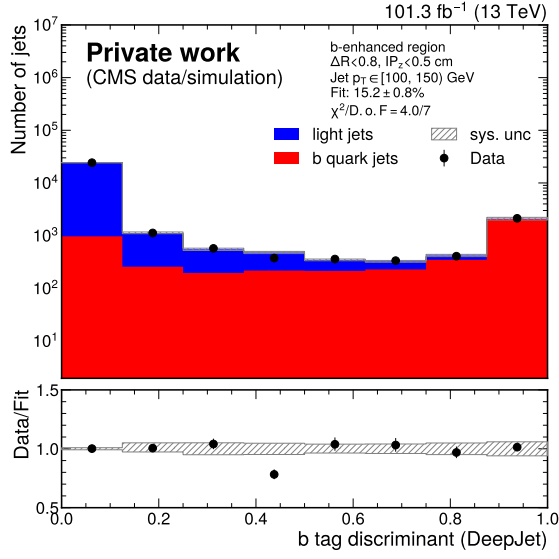


Figure B.53: The b jet fraction estimated for γ -trigger data set (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

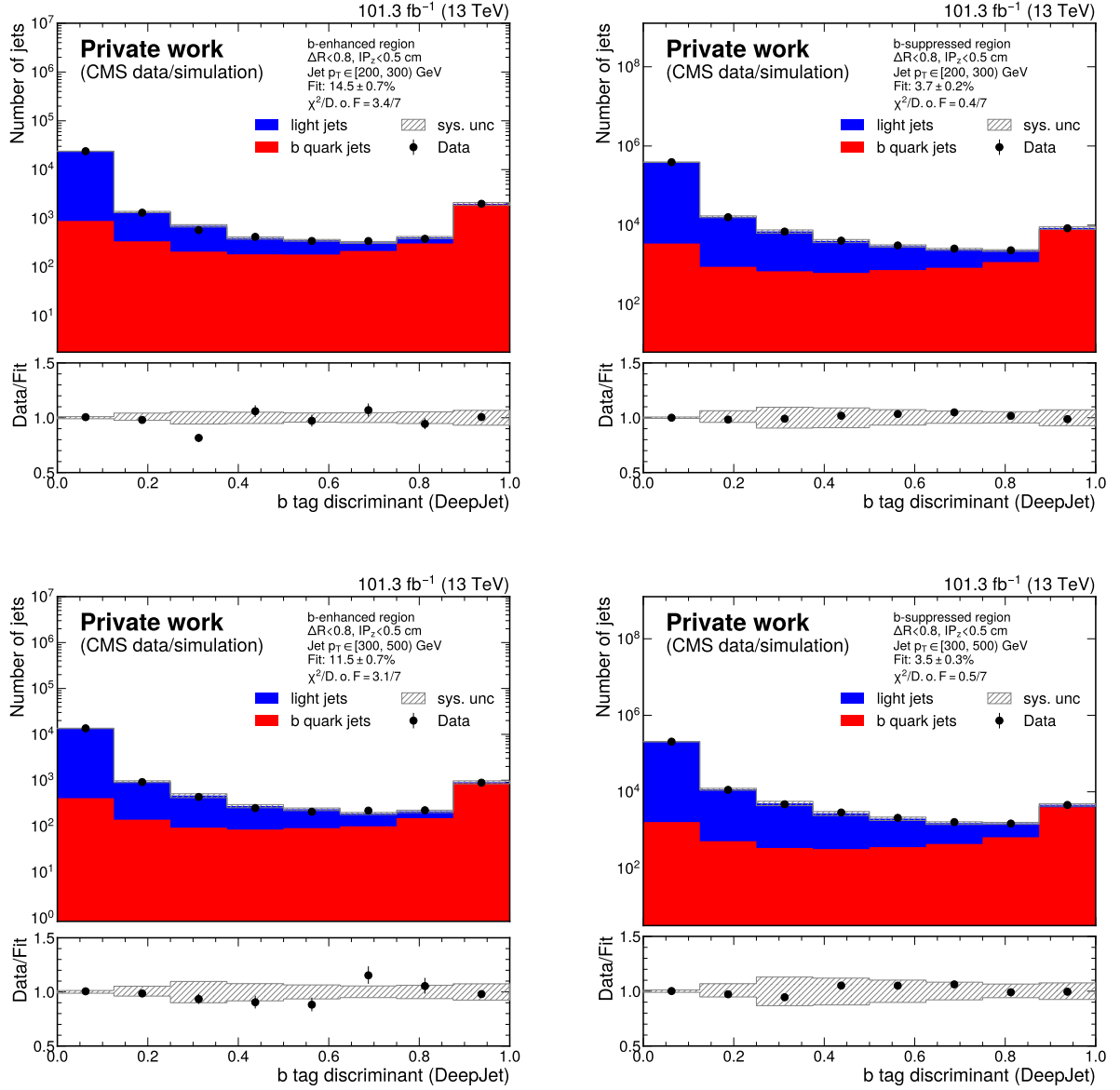


Figure B.54: The b jet fraction estimated for γ -trigger data set (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

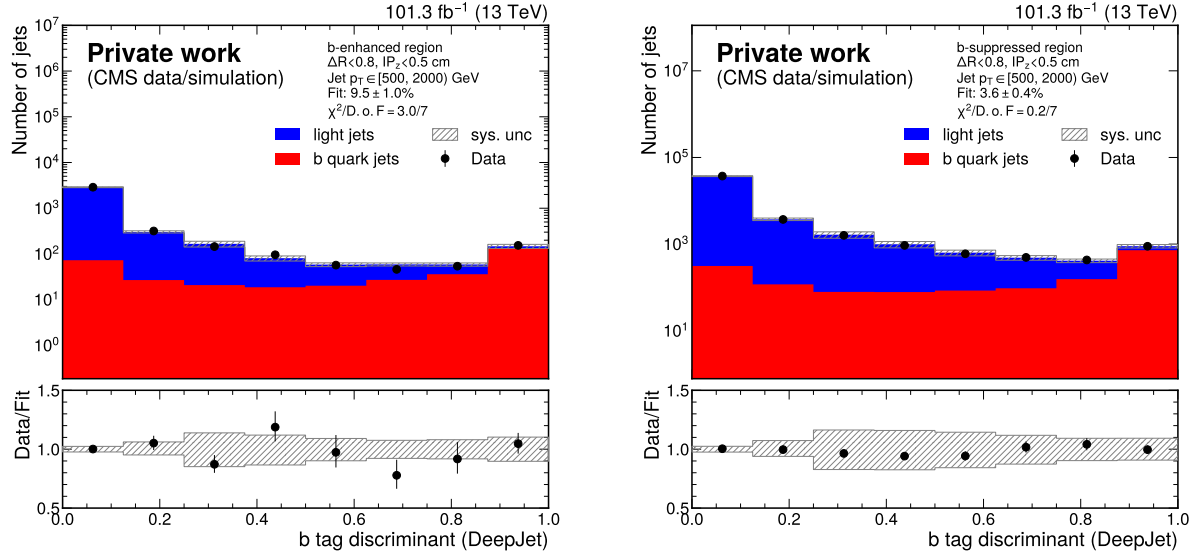


Figure B.55: The b jet fraction estimated for γ -trigger data set (pixel Phase-1) in various bins of jet p_T with a jet-track association of $R_{\text{cut}} = 0.8$

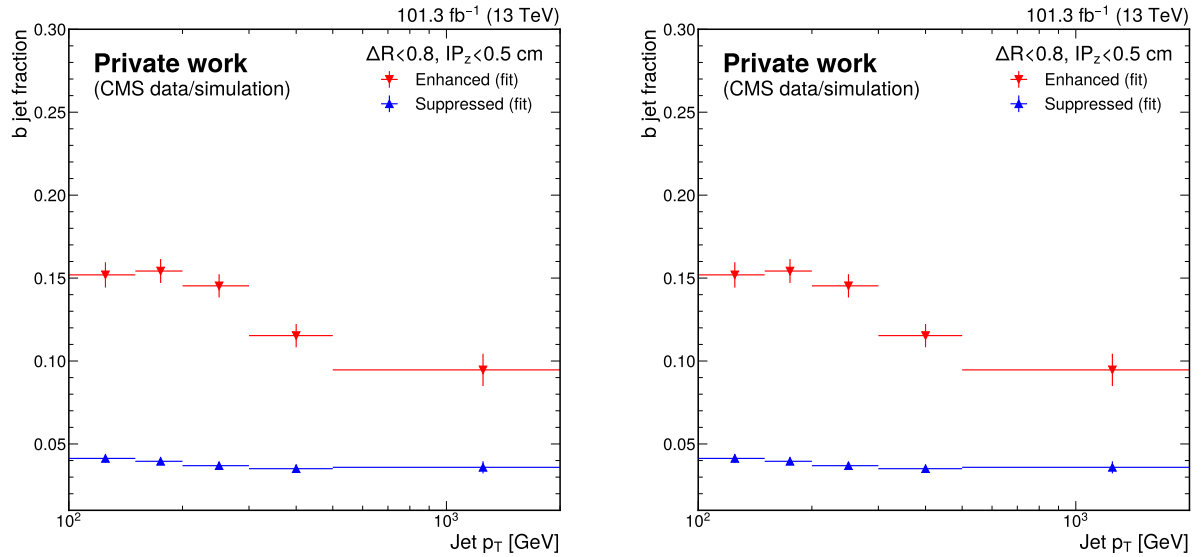


Figure B.56: The b jet fraction evaluated as a function of jet p_T estimated for γ -trigger data set (pixel Phase-1) with a jet-track association of $R_{\text{cut}} = 0.8$.

B.3 Comparison of the mistag rates evaluated in different MC samples

Here we display the mistag rates using the QCD MC samples and the γ +jets MC samples to display potential differences between the H_T - and γ -triggered data sets. For all of these plots, results evaluated using γ +jets MC samples will be displayed using solid blue and red colors and the results evaluated using QCD MC samples will be displayed in the light blue and orange data point.

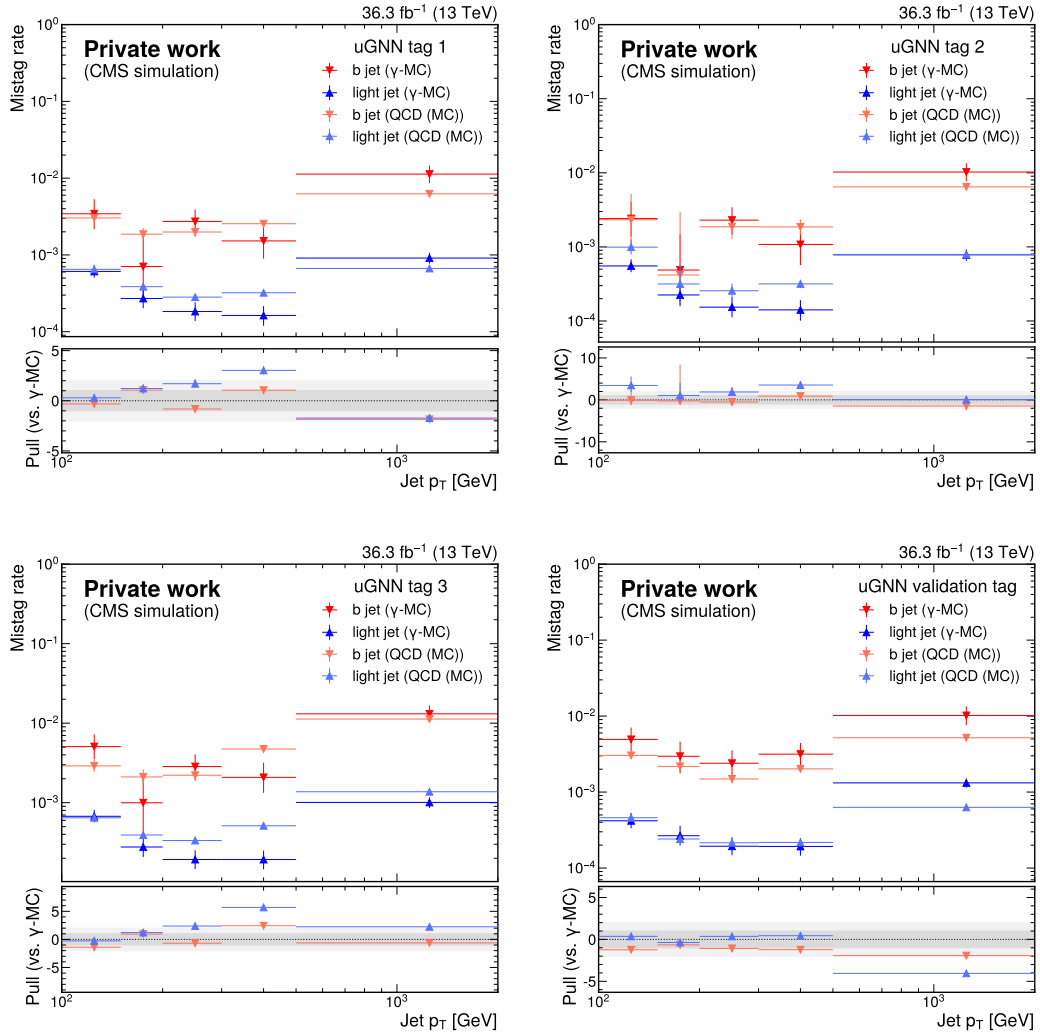


Figure B.57: Comparison of mistag rates of GNN tagger for unflavored models evaluated using the γ +jets MC samples (blue, red), and the QCD MC samples (light blue, orange) with pixel Phase-0 geometry.

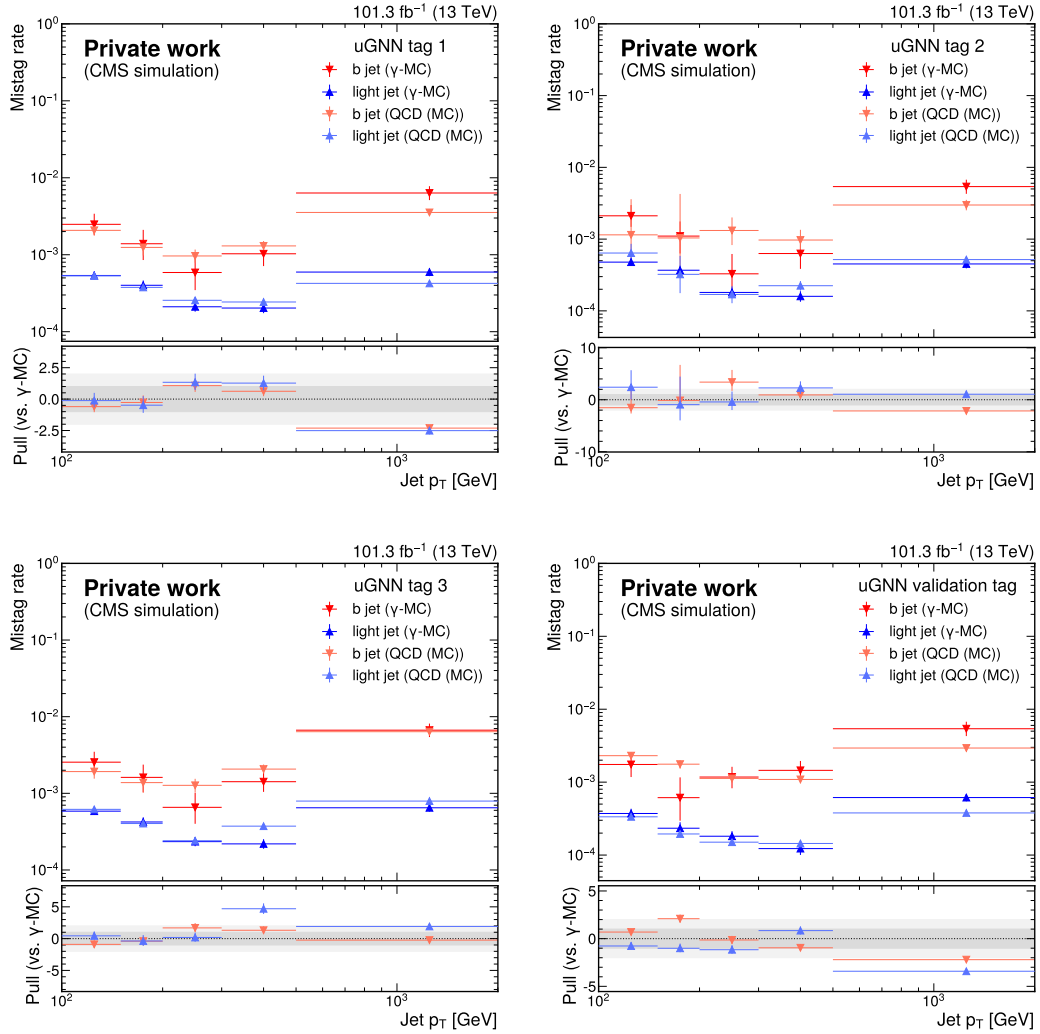


Figure B.58: Comparison of mistag rates of GNN tagger for unflavored models evaluated using the γ +jets MC samples (blue, red), and the QCD MC samples (light blue, orange) with pixel Phase-1 geometry.

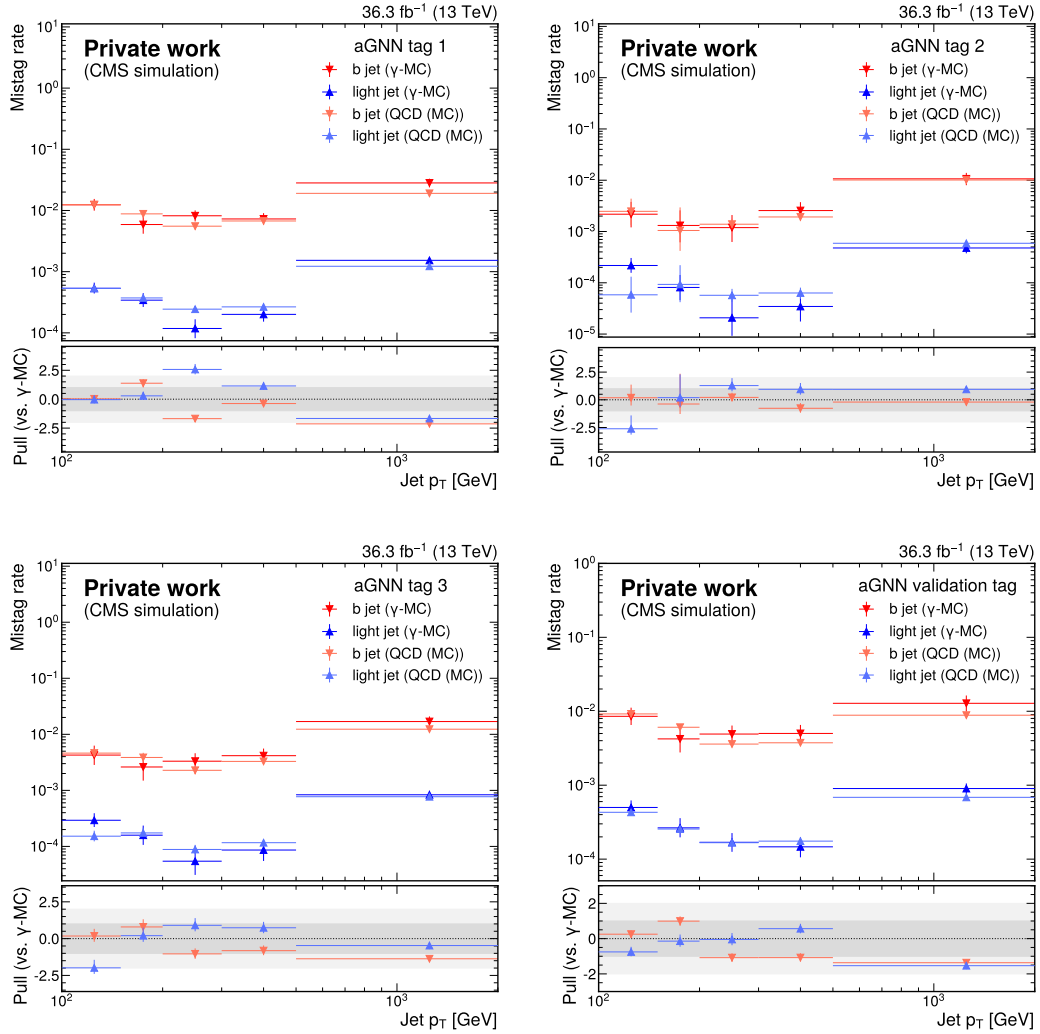


Figure B.59: Comparison of mistag rates of GNN tagger for flavor-aligned models evaluated using the γ +jets MC samples (blue, red), and the QCD MC samples (light blue, orange) with pixel Phase-0 geometry.

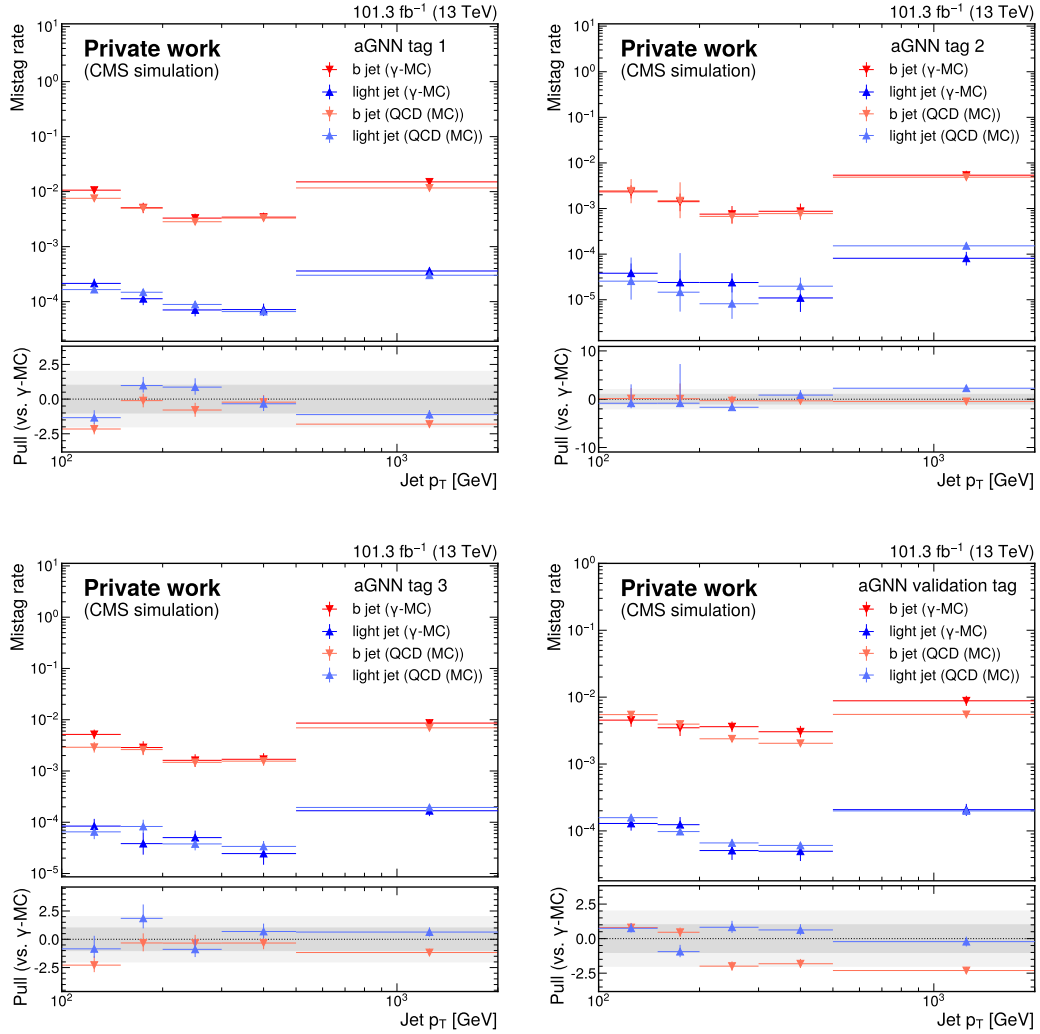


Figure B.60: Comparison of mistag rates of GNN tagger for flavor-aligned models evaluated using the γ +jets MC samples (blue, red), and the QCD MC samples (light blue, orange) with pixel Phase-1 geometry.

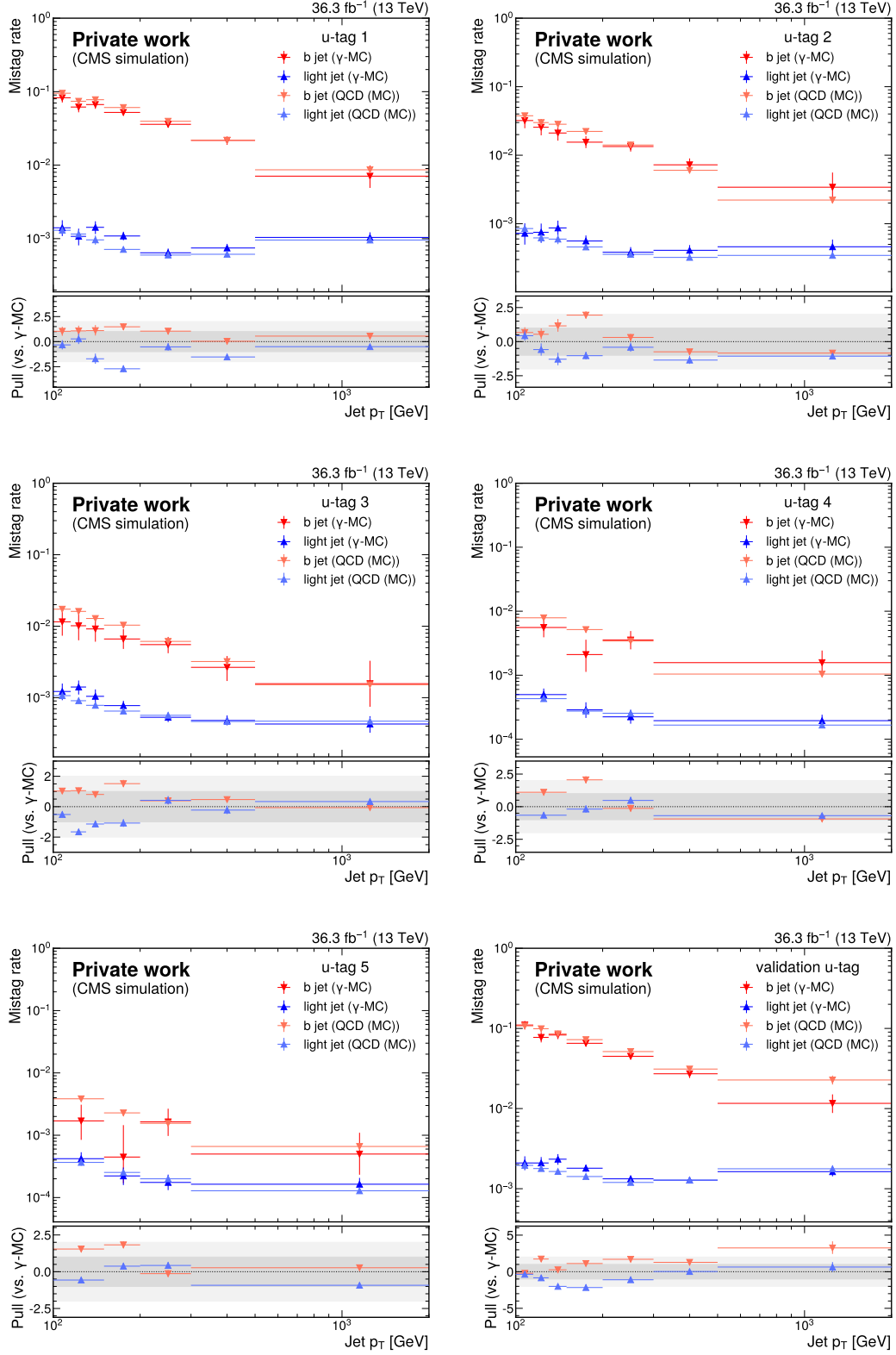


Figure B.61: Comparison of mistag rates of model-agnostic tagger for unflavored models evaluated using the γ +jets MC samples (blue, red), and the QCD MC samples (light blue, orange) with pixel Phase-0 geometry.

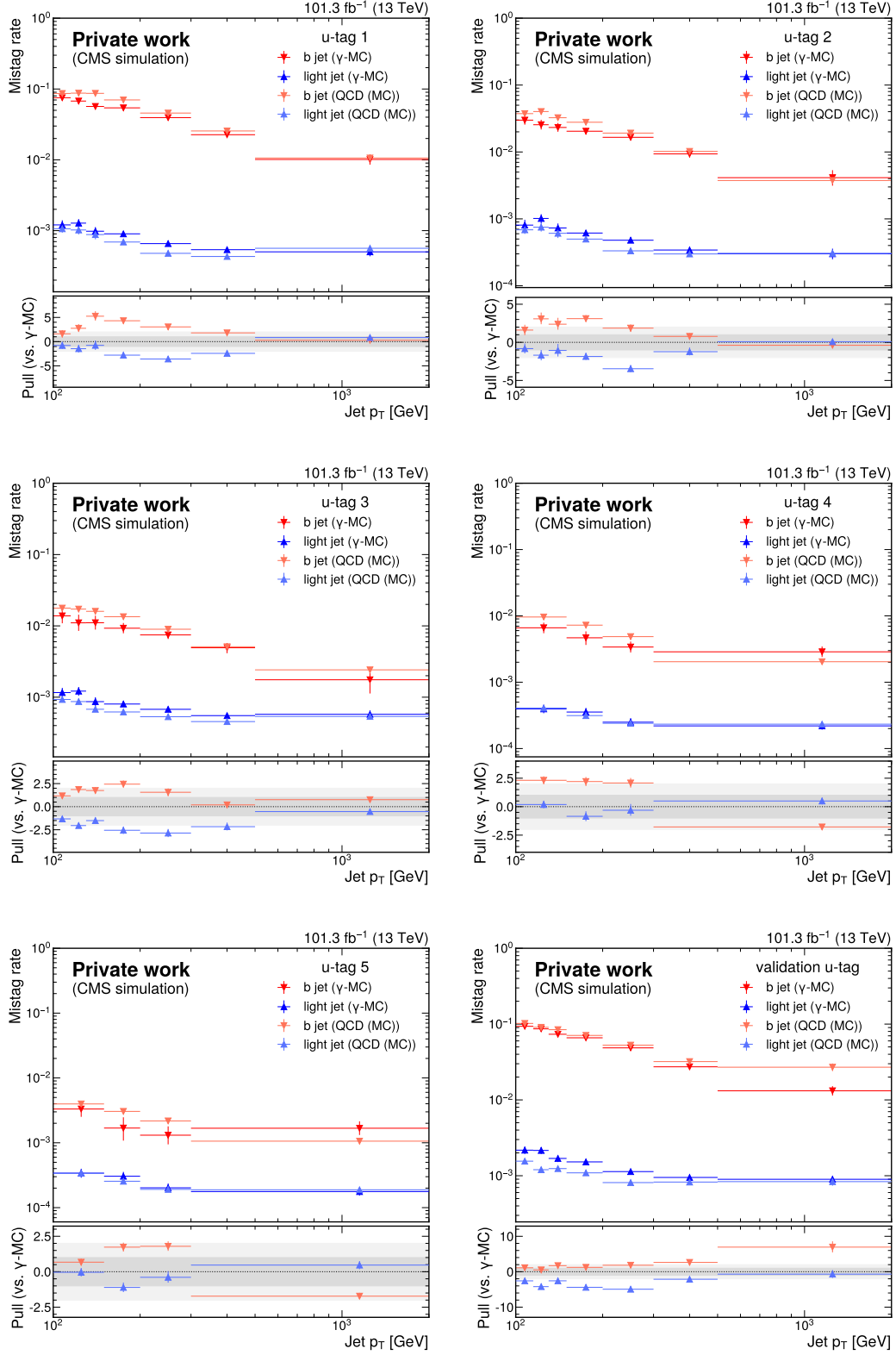


Figure B.62: Comparison of mistag rates of model-agnostic tagger for unflavored models evaluated using the γ +jets MC samples (blue, red), and the QCD MC samples (light blue, orange) with pixel Phase-1 geometry.

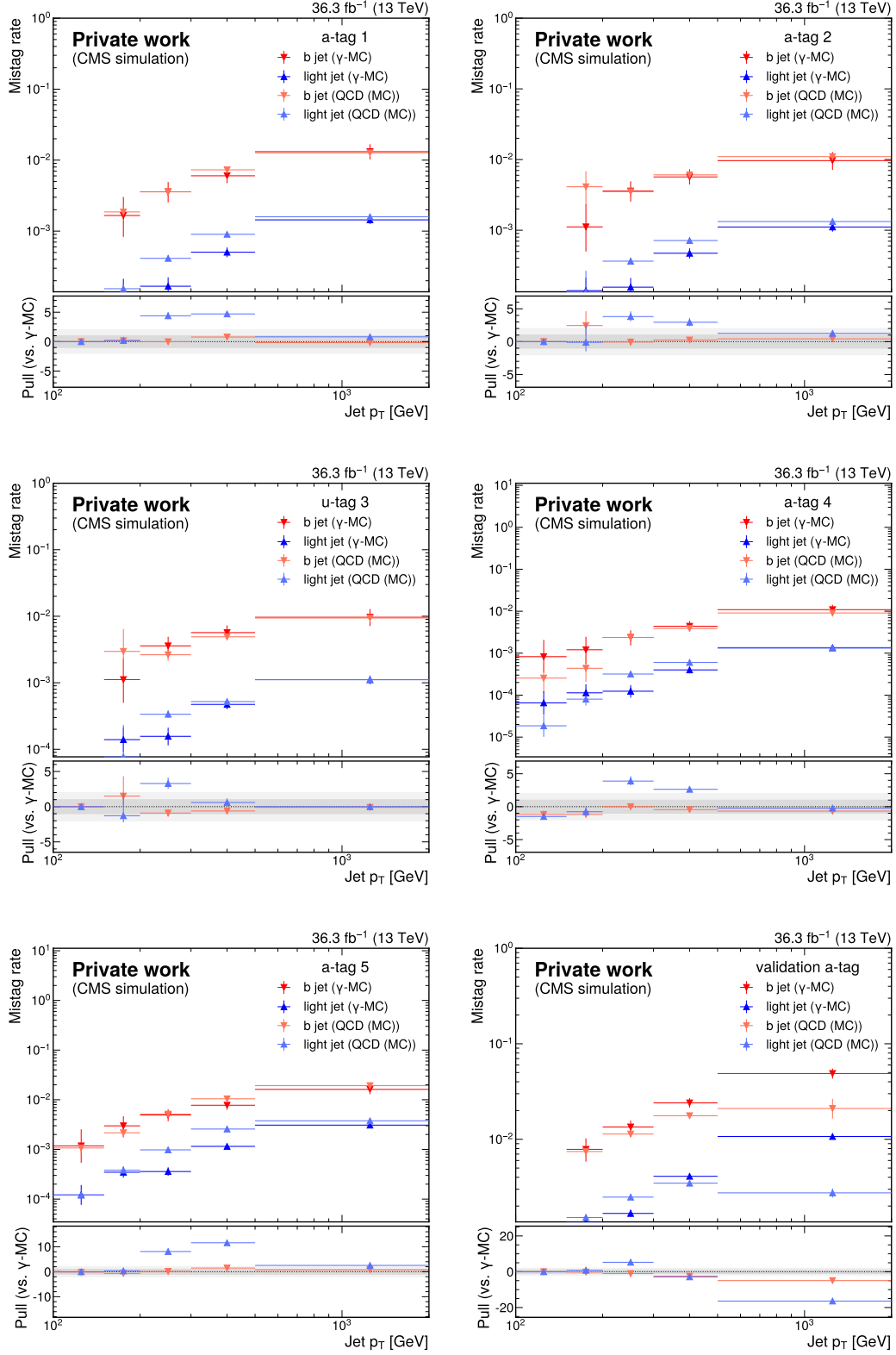


Figure B.63: Comparison of mistag rates of model-agnostic tagger for flavor-aligned models evaluated using the γ +jets MC samples (blue, red), and the QCD MC samples (light blue, orange) with pixel Phase-0 geometry.

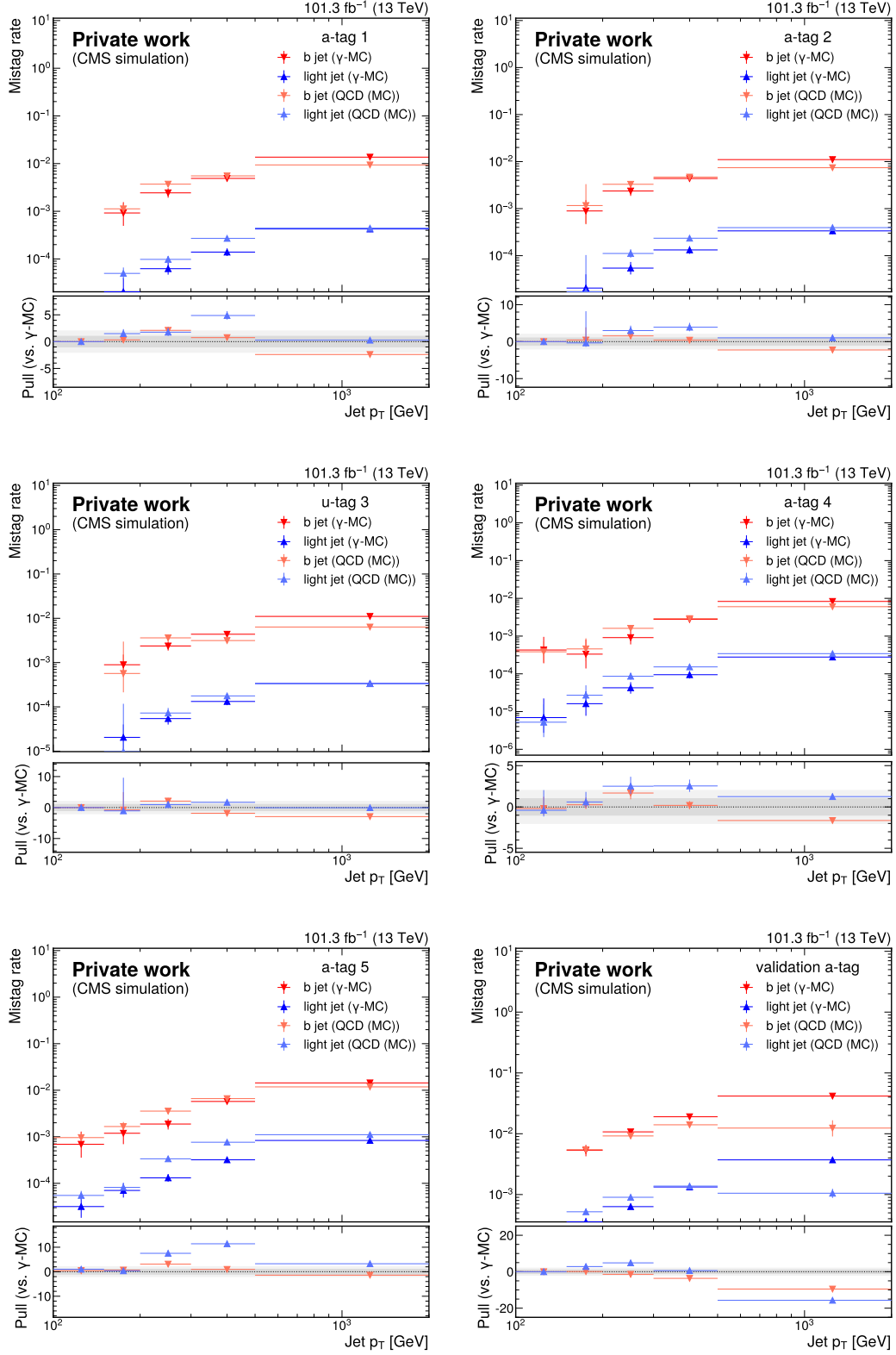


Figure B.64: Comparison of mistag rates of model-agnostic tagger for flavor-aligned models evaluated using the γ +jets MC samples (blue, red), and the QCD MC samples (light blue, orange) with pixel Phase-1 geometry.

B.4 Mistag rates evaluated in signal-free data stream

In this section, we display the mistag rate evaluated in the γ -triggered data set, shown in solid blue and red colors, compared with the γ +jets MC samples, shown in light blue and orange points.

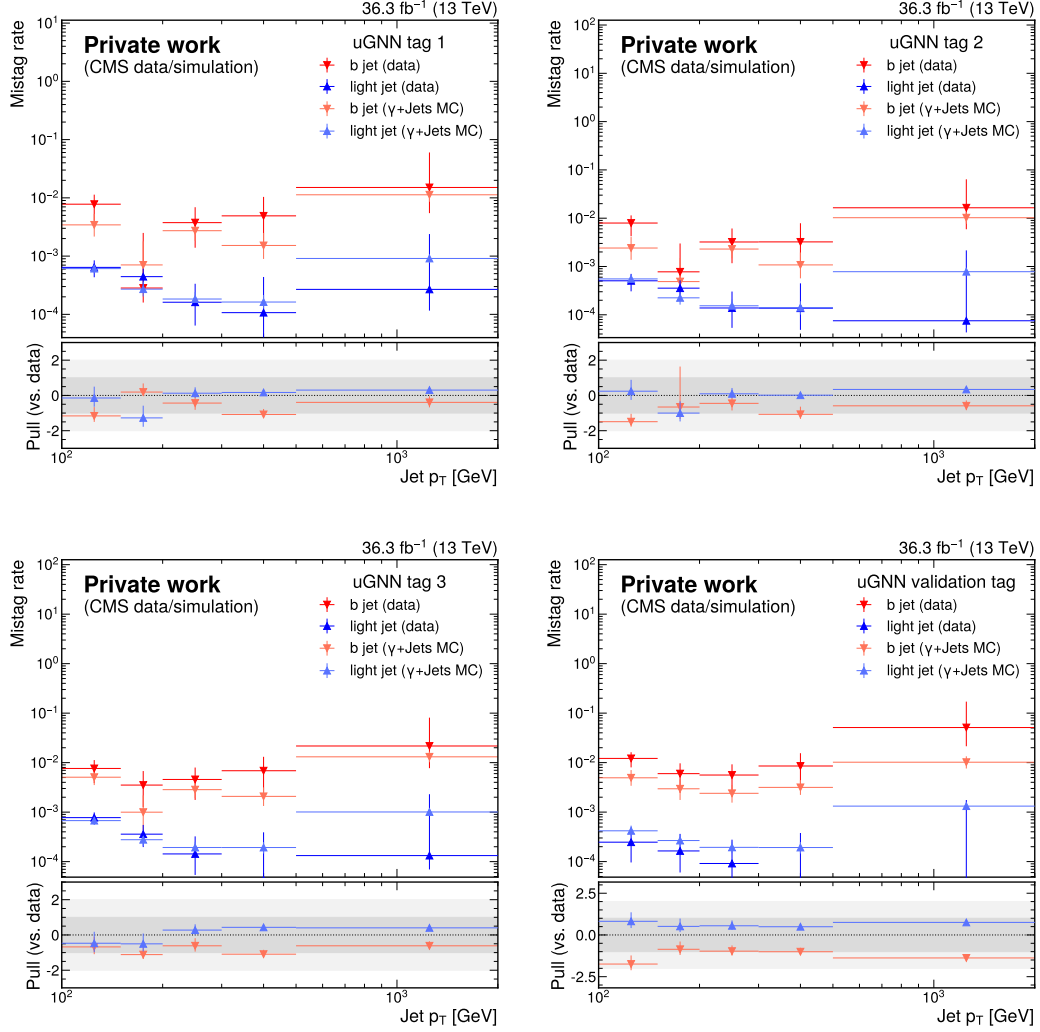


Figure B.65: Comparison of mistag rates of GNN tagger for unflavored signal samples evaluated using the γ +triggered data collected by CMS (blue, red), and the γ +jets MC samples (light blue, orange) with pixel Phase-0 geometry.

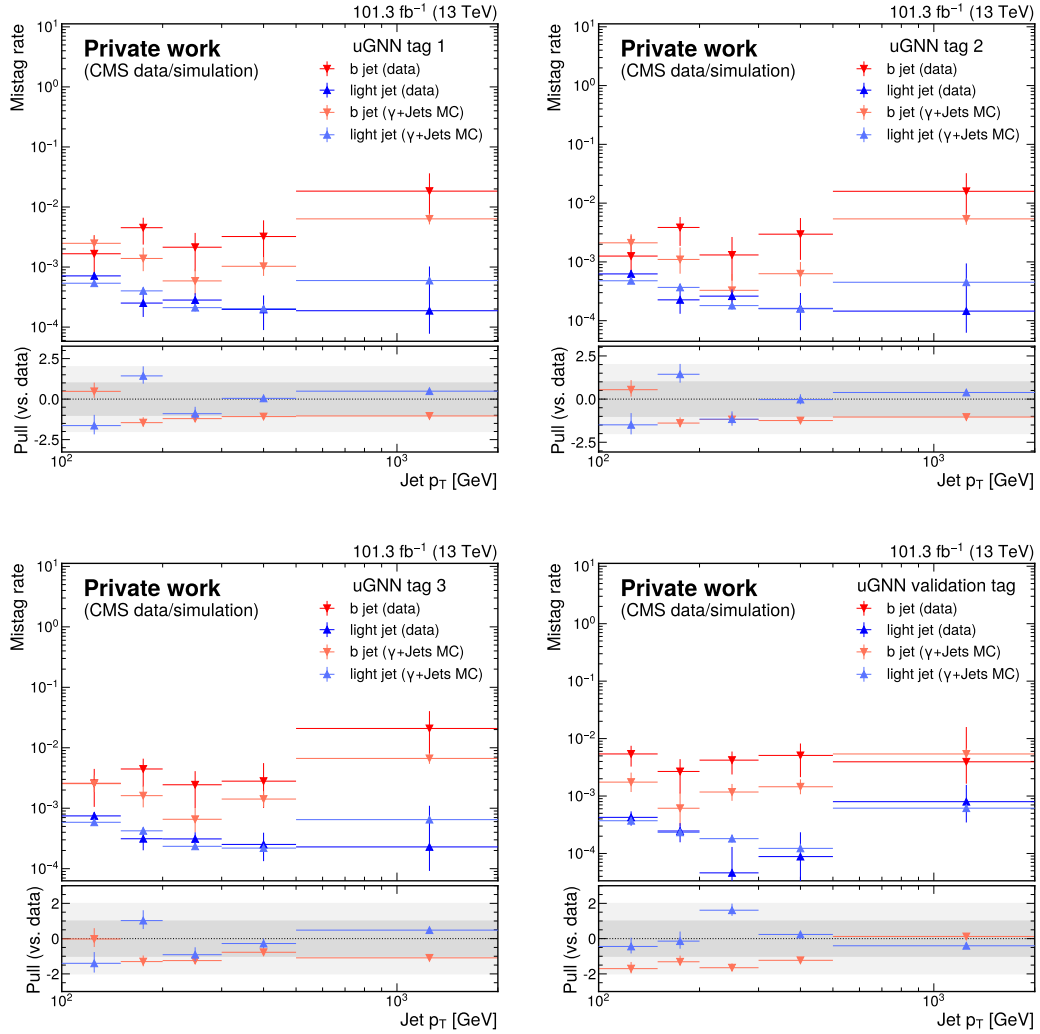


Figure B.66: Comparison of mistag rates of GNN tagger for unflavored signal samples evaluated using the γ +triggered data collected by CMS (blue, red), and the γ +jets MC samples (light blue, orange) with pixel Phase-1 geometry.

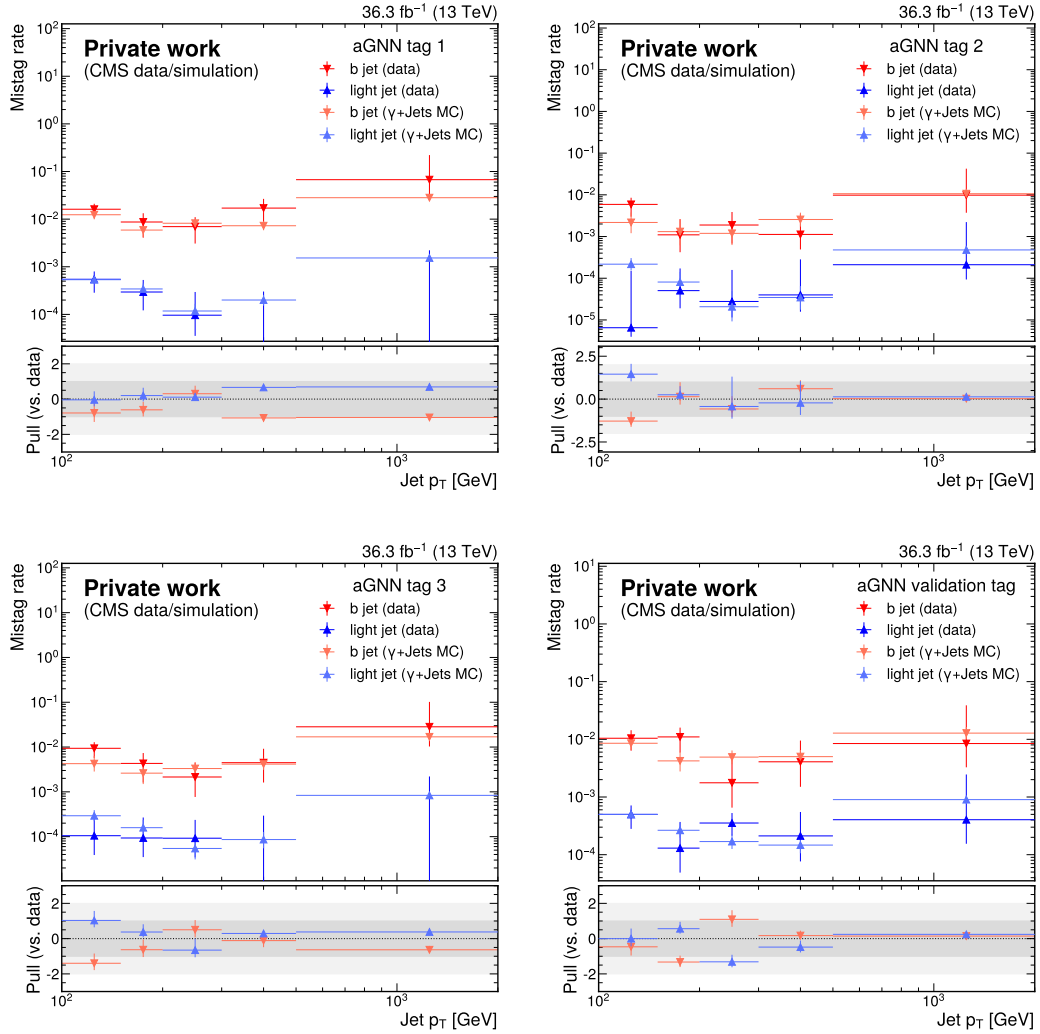


Figure B.67: Comparison of mistag rates of GNN tagger for flavor-aligned models evaluated using the γ +triggered data collected by CMS (blue, red), and the γ +jets MC samples (light blue, orange) with pixel Phase-0 geometry.

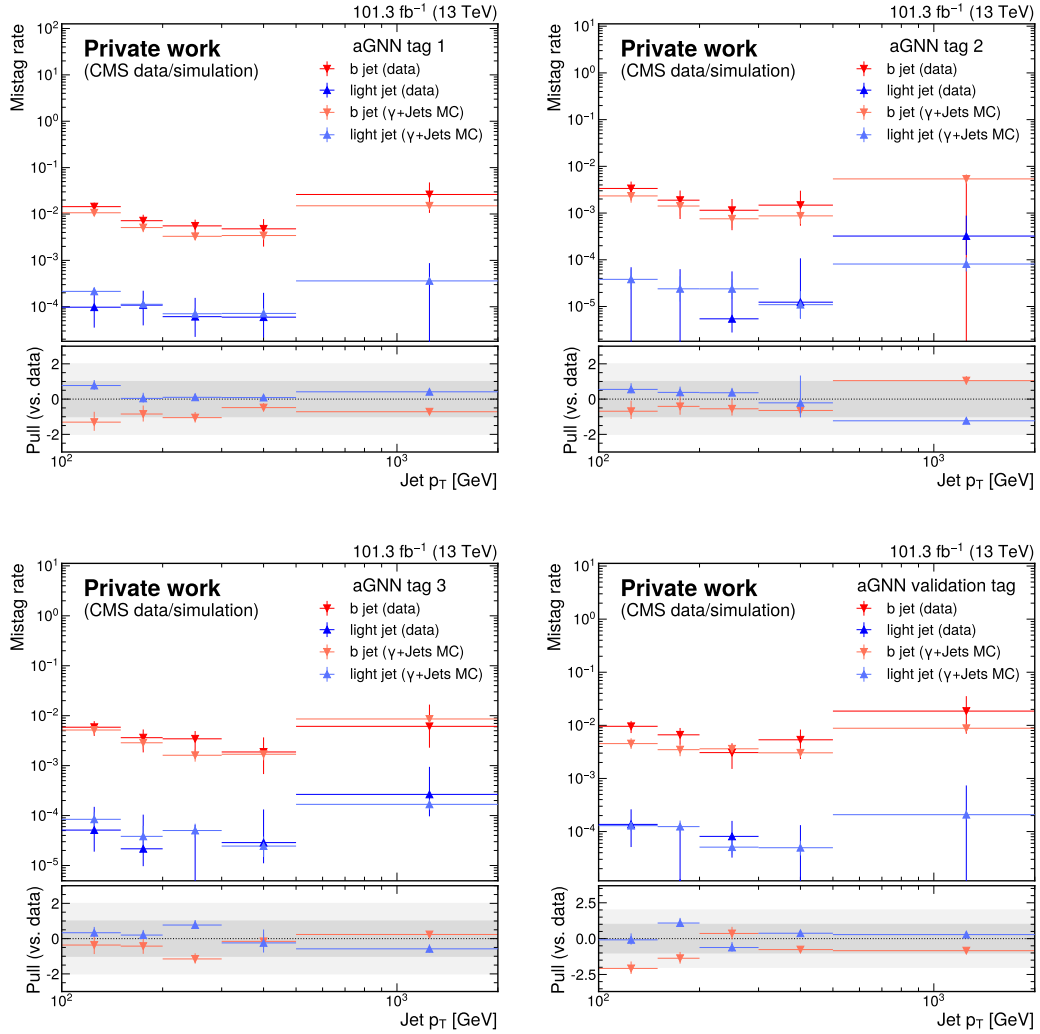


Figure B.68: Comparison of mistag rates of GNN tagger for flavor-aligned models evaluated using the γ -triggered data collected by CMS (blue, red), and the γ -jets MC samples (light blue, orange) with pixel Phase-1 geometry.

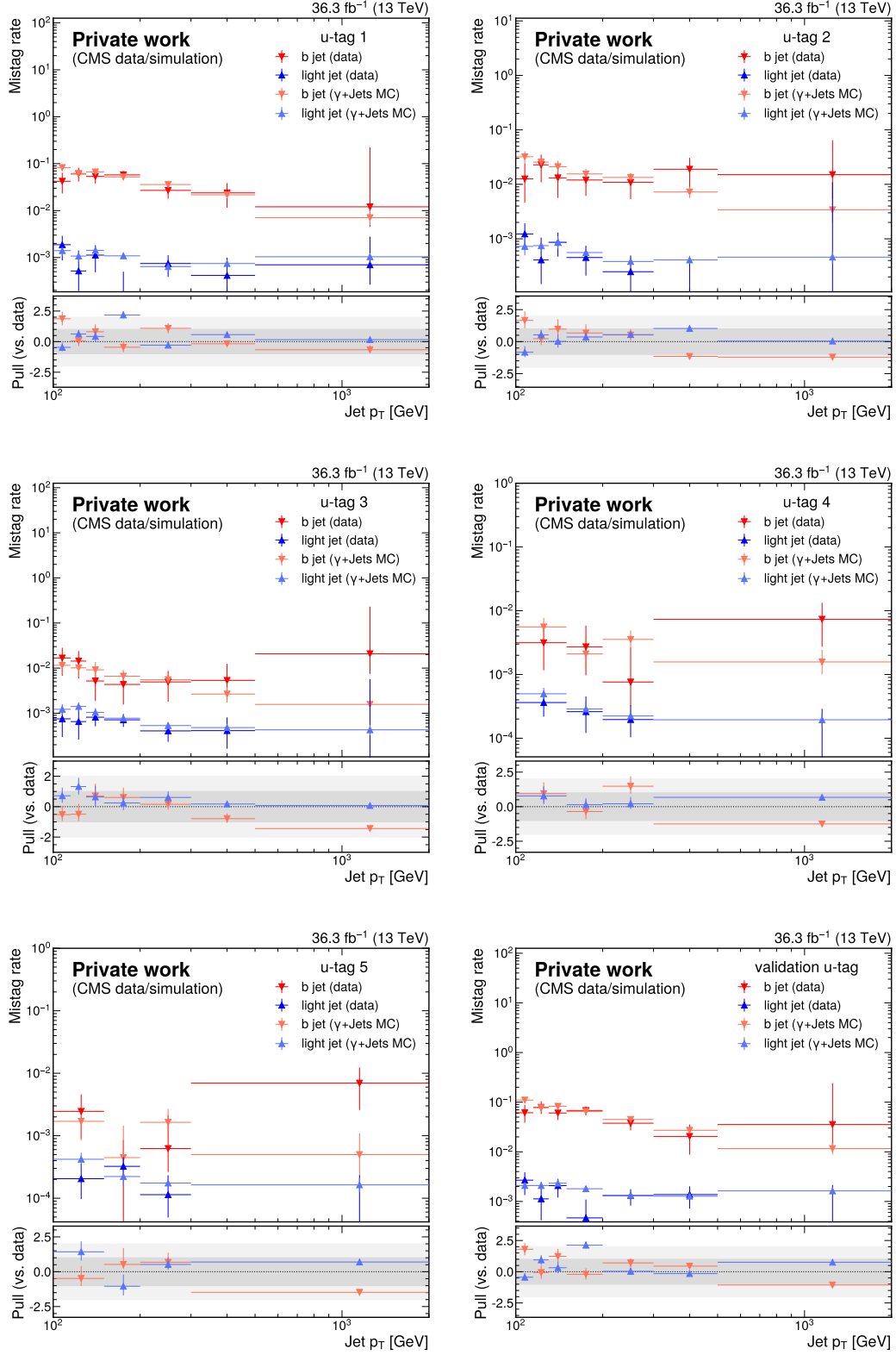


Figure B.69: Comparison of mistag rates of model-agnostic tagger for unflavored models evaluated using the γ +triggered data collected by CMS (blue, red), and the γ +jets MC samples (light blue, orange) with pixel Phase-0 geometry.

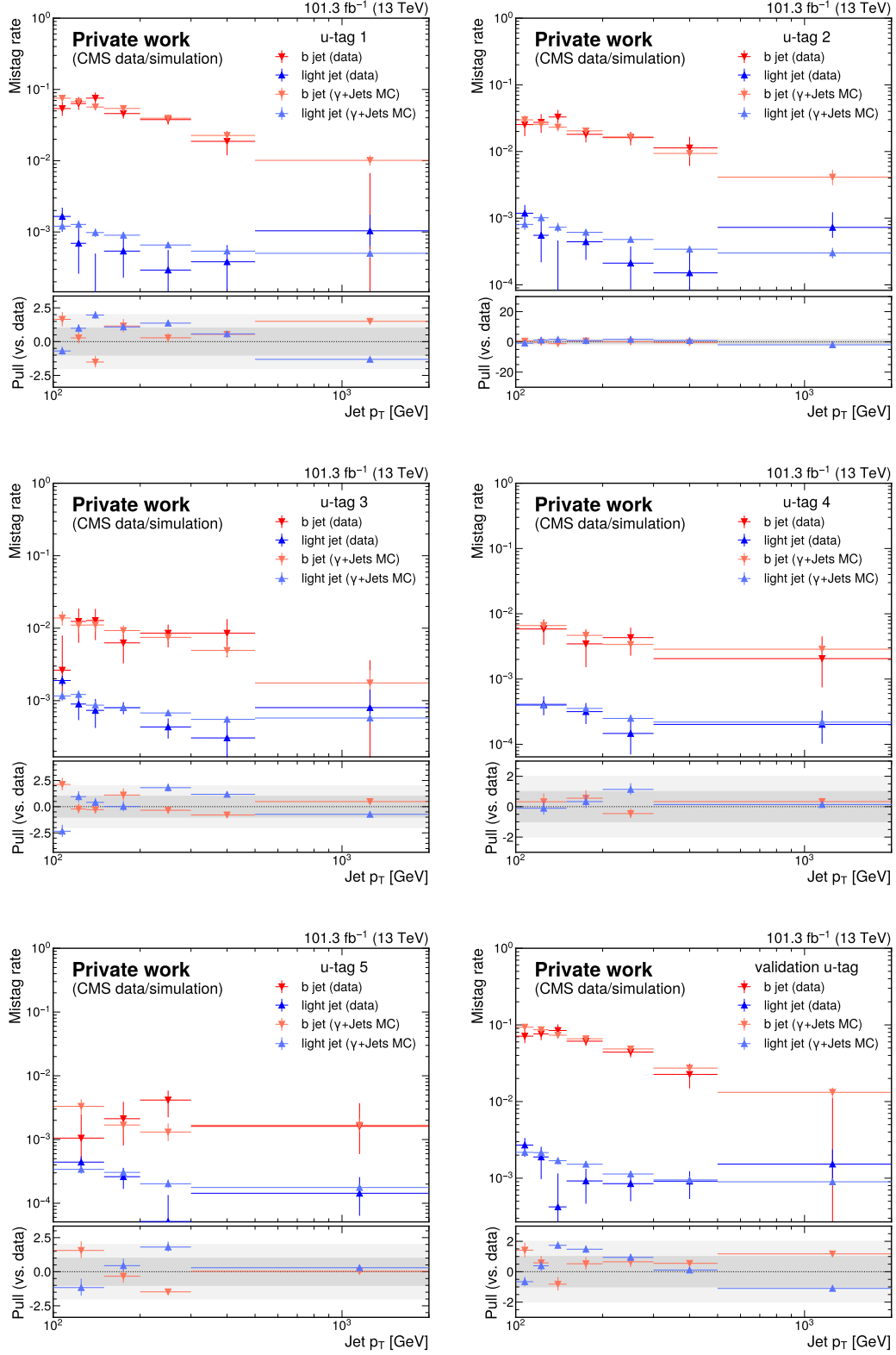


Figure B.70: Comparison of mistag rates of model-agnostic tagger for unflavored models evaluated using the γ +triggered data collected by CMS (blue, red), and the γ +jets MC samples (light blue, orange) with pixel Phase-1 geometry.

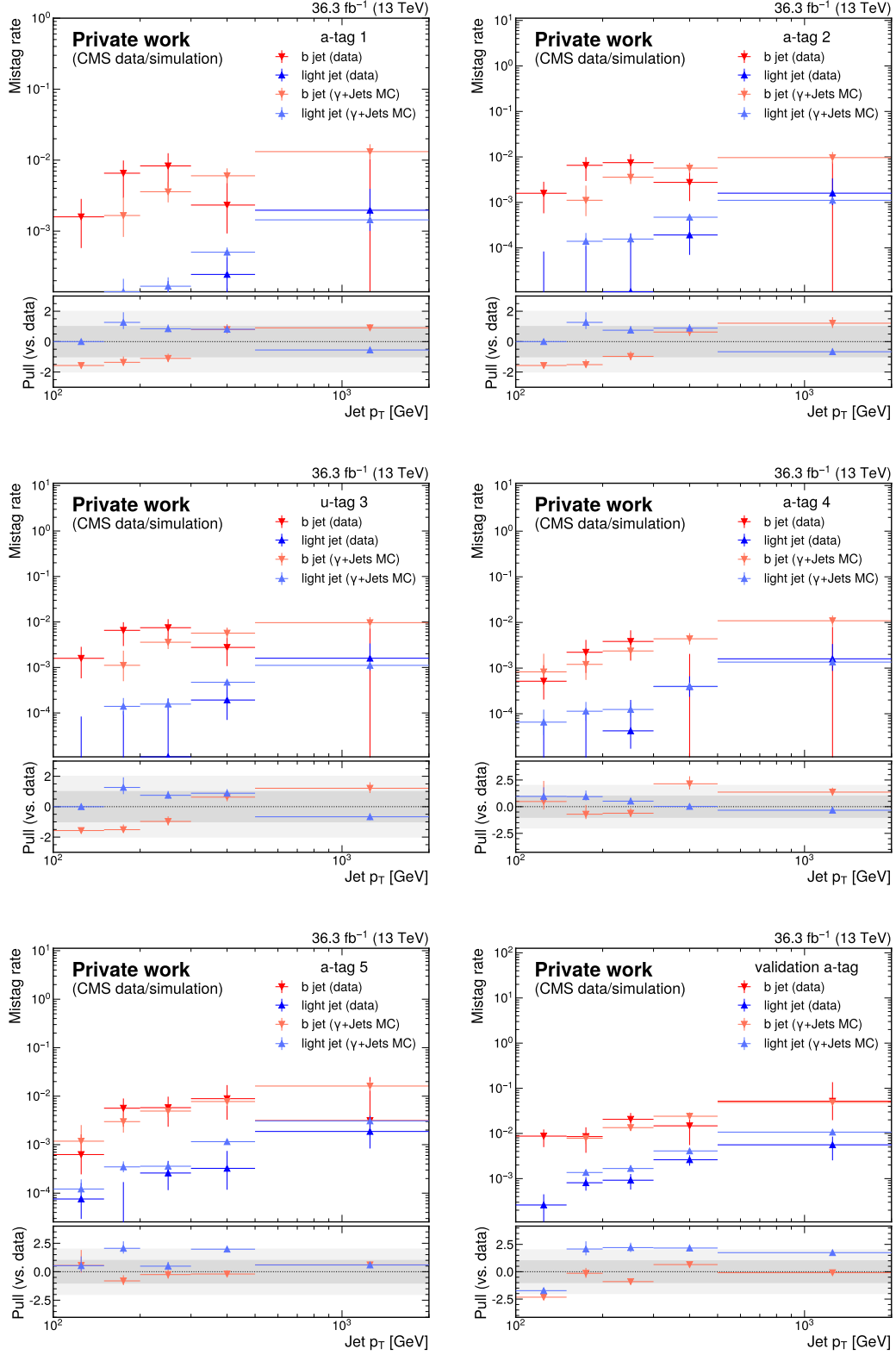


Figure B.71: Comparison of mistag rates of model-agnostic tagger for flavor-aligned models evaluated using the γ +triggered data collected by CMS (blue, red), and the γ +jets MC samples (light blue, orange) with pixel Phase-0 geometry.

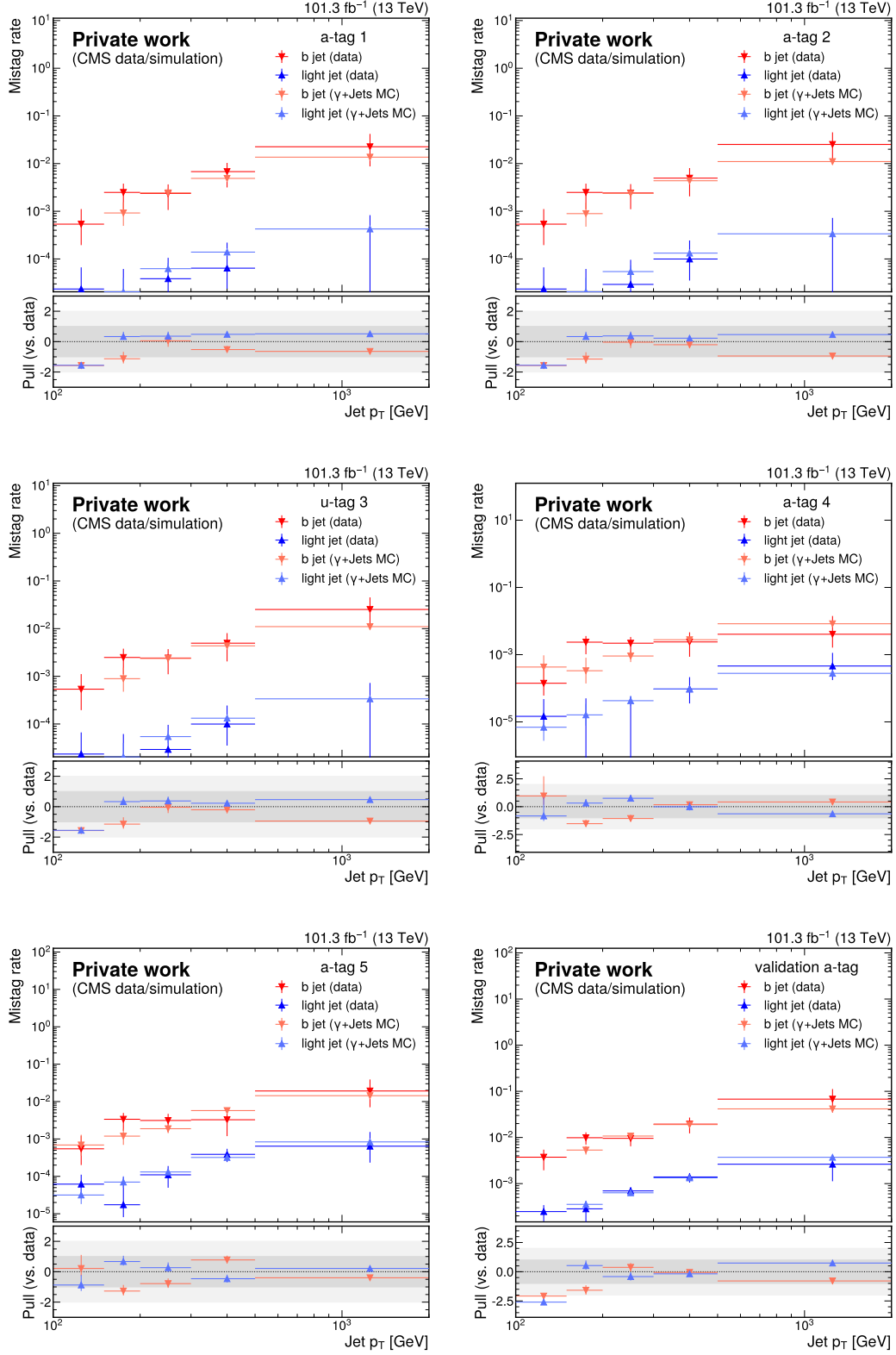


Figure B.72: Comparison of mistag rates of model-agnostic tagger for flavor-aligned models evaluated using the γ +triggered data collected by CMS (blue, red), and the γ +jets MC samples (light blue, orange) with pixel Phase-1 geometry.

B.5 Estimated fraction of the b jet fraction of the control region

This section displays the b jet fraction evaluated using the template fit method for the non-EMJ tagged jets in the control region defined using the selection criteria define in Table 4.4. Here, we display the fit results evaluated in QCD MC samples, and the H_T -triggered data sets.

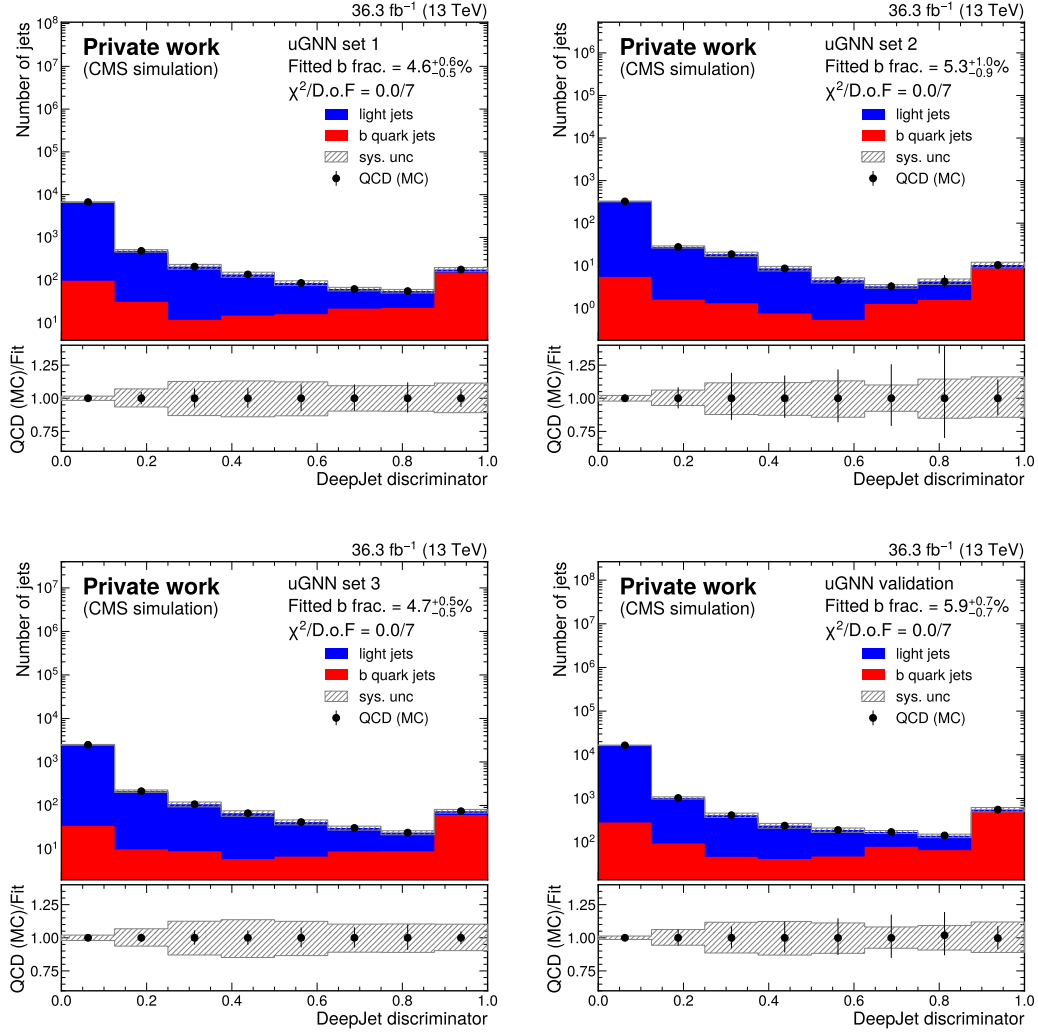


Figure B.73: Estimation of the b jet fraction of control regions defined using the GNN tagger for unflavored signal samples evaluated using QCD MC samples with pixel Phase-0 geometry.

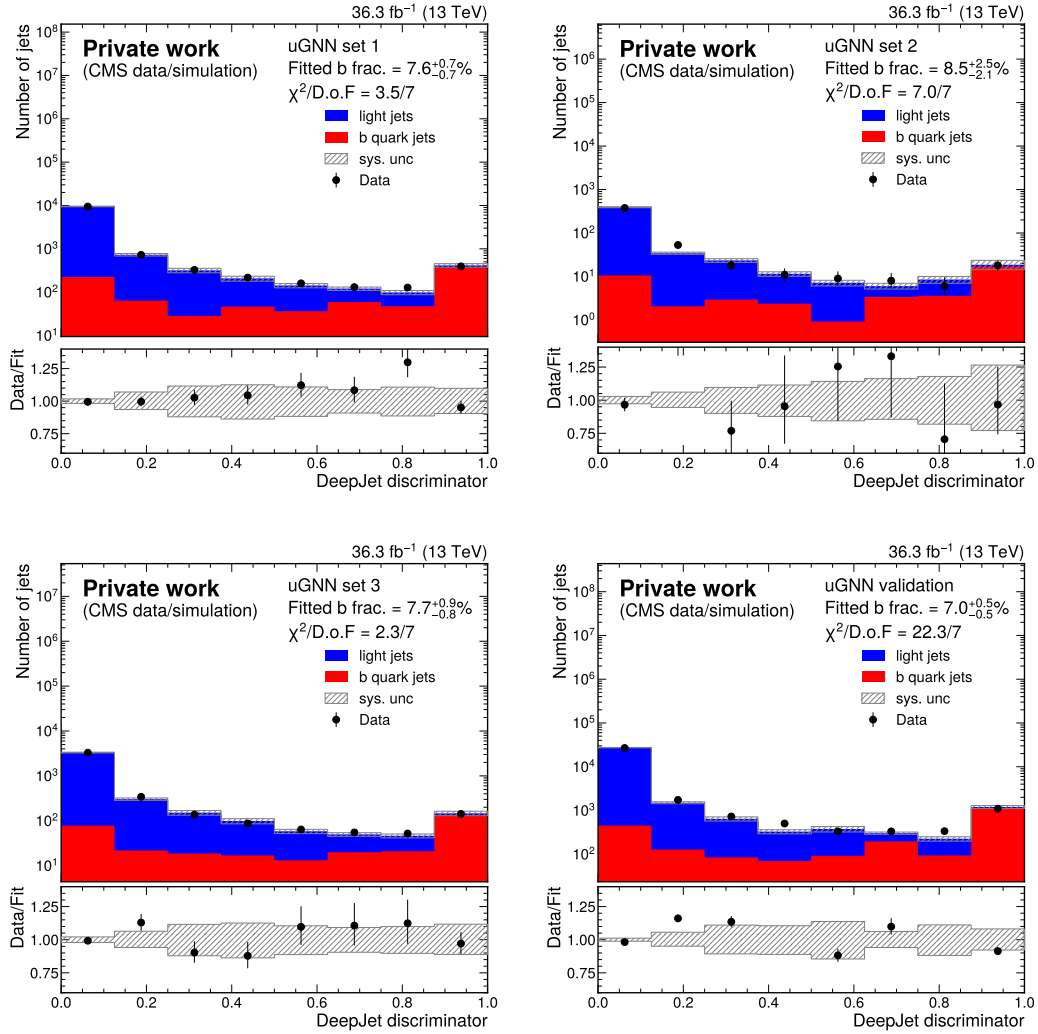


Figure B.74: Estimation of the b jet fraction of control regions defined using the GNN tagger for unflavored signal samples evaluated using H_T -triggered data with pixel Phase-0 geometry.

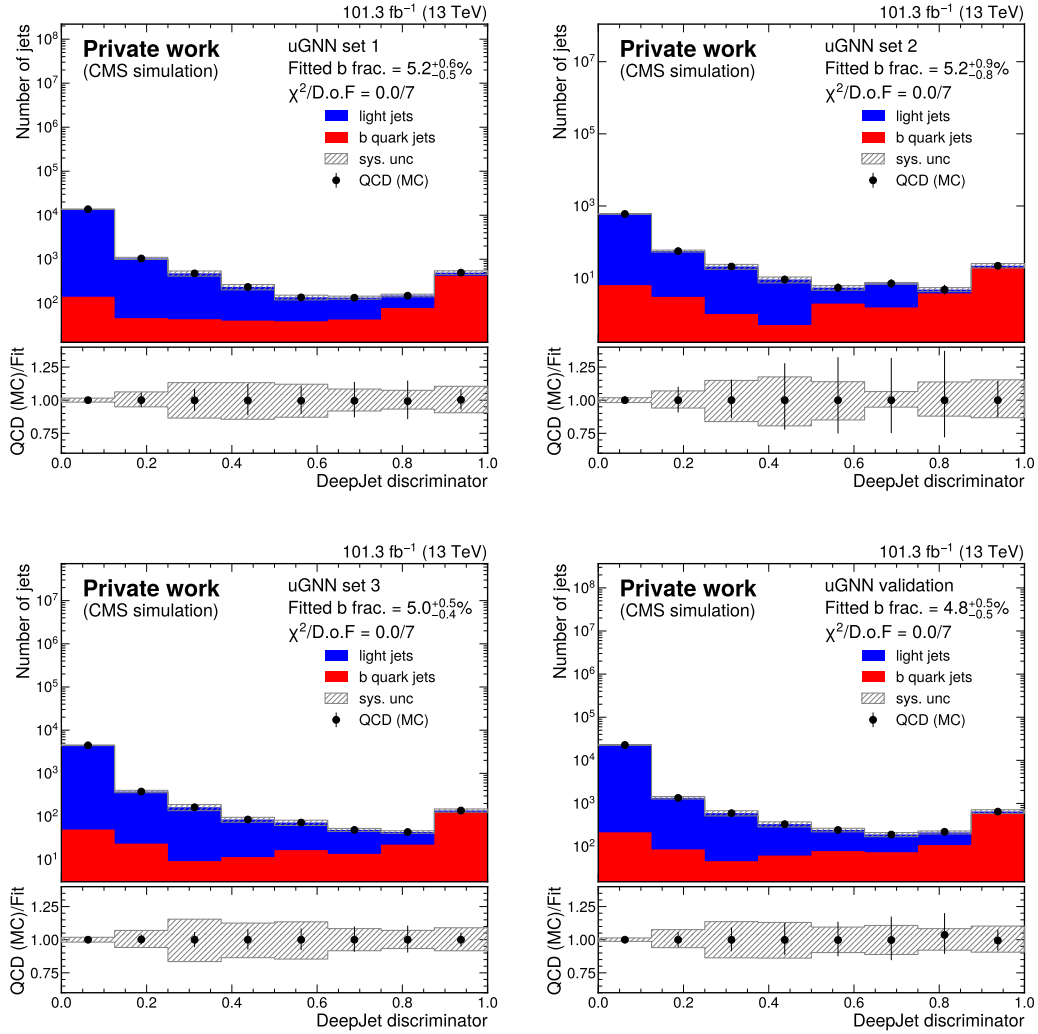


Figure B.75: Estimation of the b jet fraction of control regions defined using the GNN tagger for unflavored signal samples evaluated using QCD MC samples with pixel Phase-1 geometry.

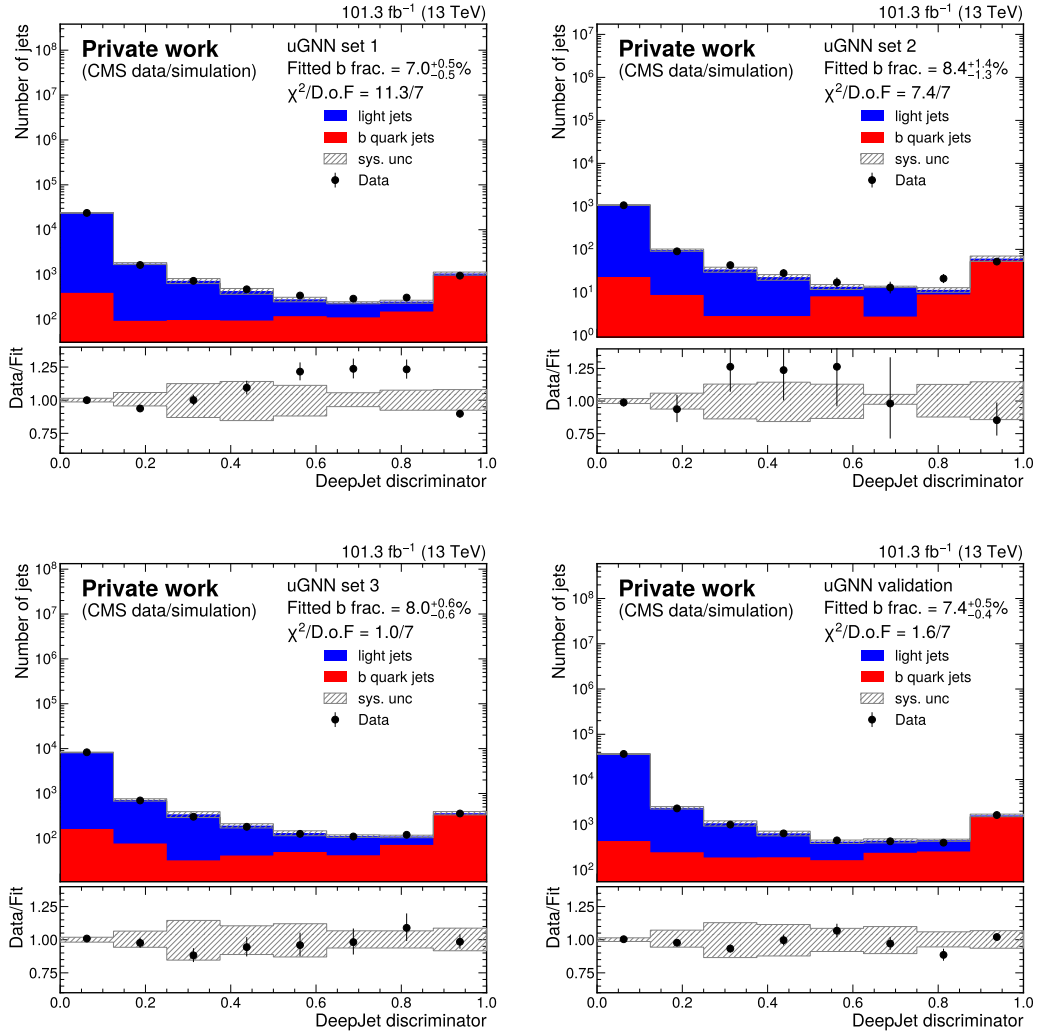


Figure B.76: Estimation of the b jet fraction of control regions defined using the GNN tagger for unflavored signal samples evaluated using H_T -triggered data with pixel Phase-1 geometry.

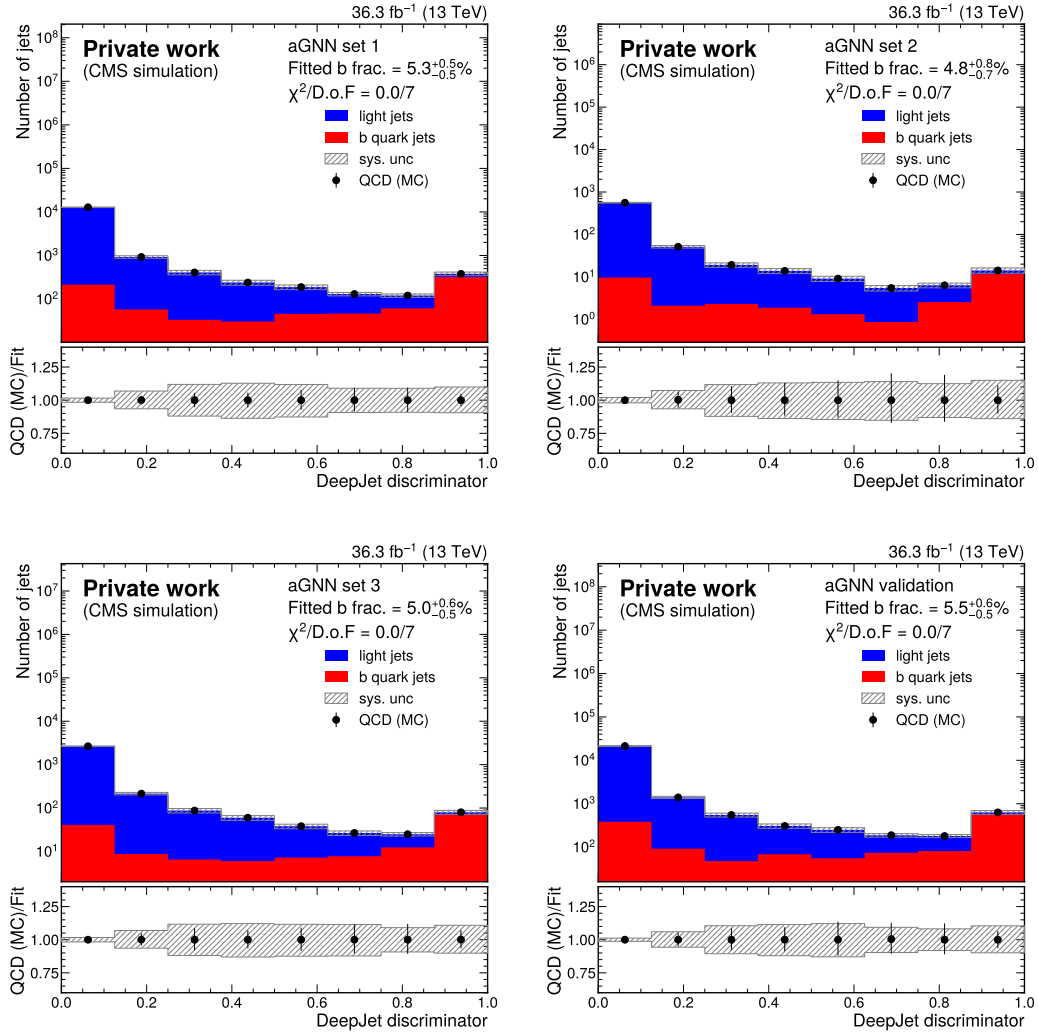


Figure B.77: Estimation of the b jet fraction of control regions defined using the GNN tagger for flavor-aligned models evaluated using QCD MC samples with pixel Phase-0 geometry.

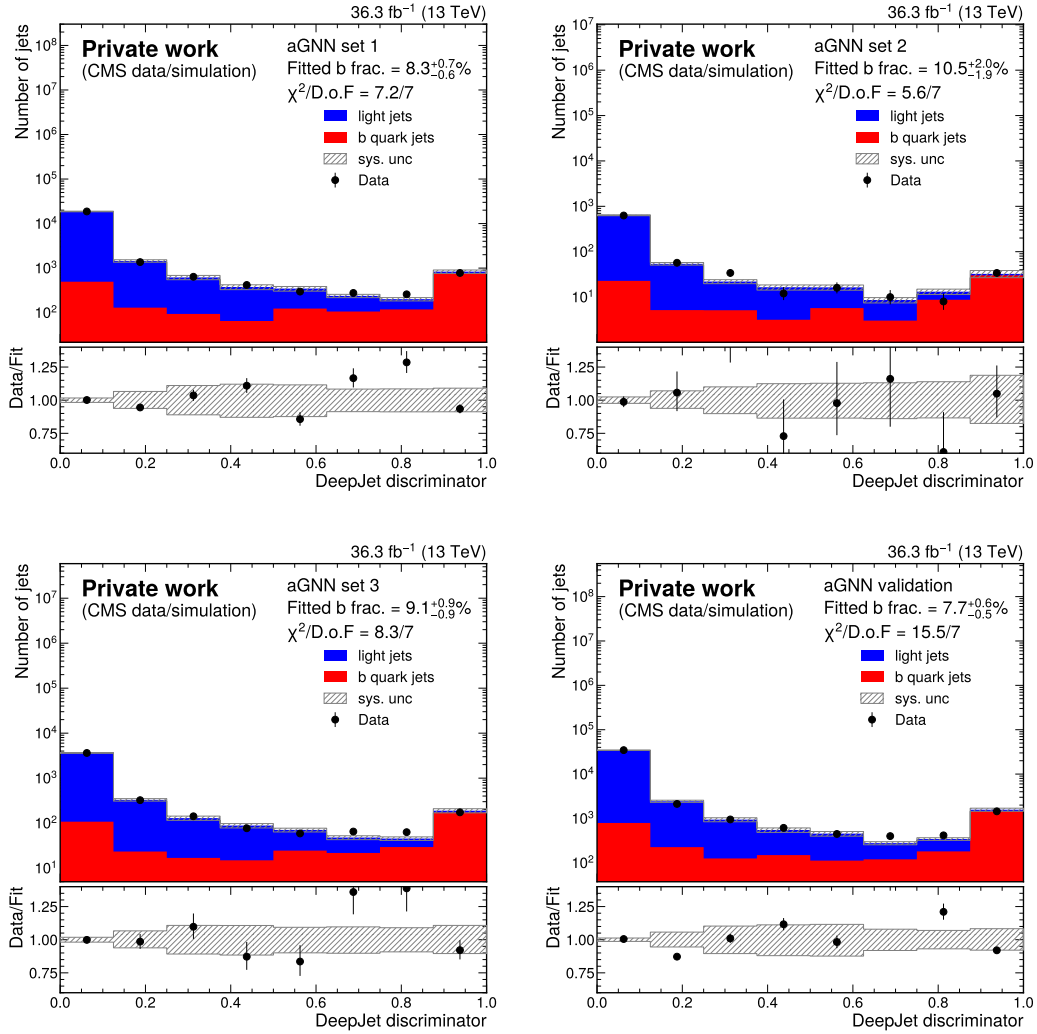


Figure B.78: Estimation of the b jet fraction of control regions defined using the GNN tagger for flavor-aligned models evaluated using H_T -triggered data with pixel Phase-0 geometry.

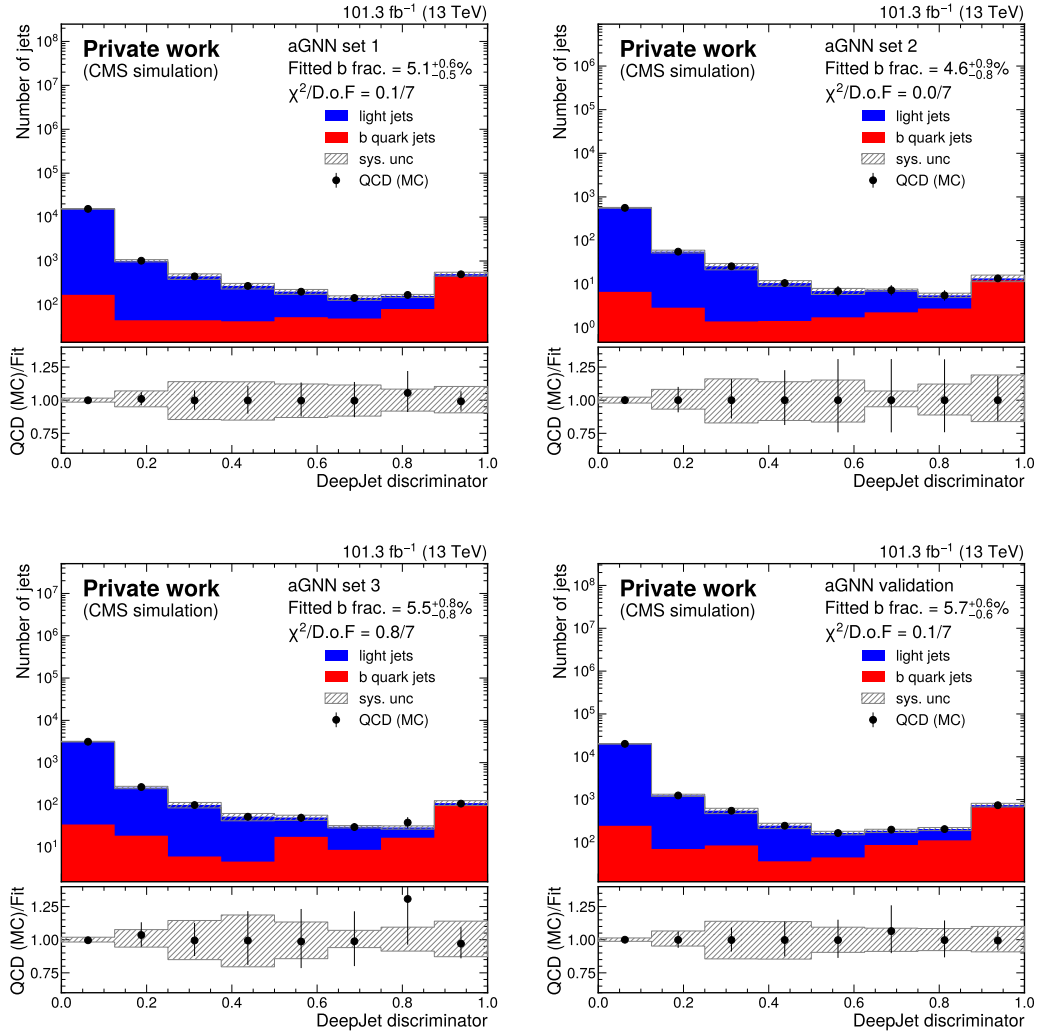


Figure B.79: Estimation of the b jet fraction of control regions defined using the GNN tagger for flavor-aligned models evaluated using QCD MC samples with pixel Phase-1 geometry.

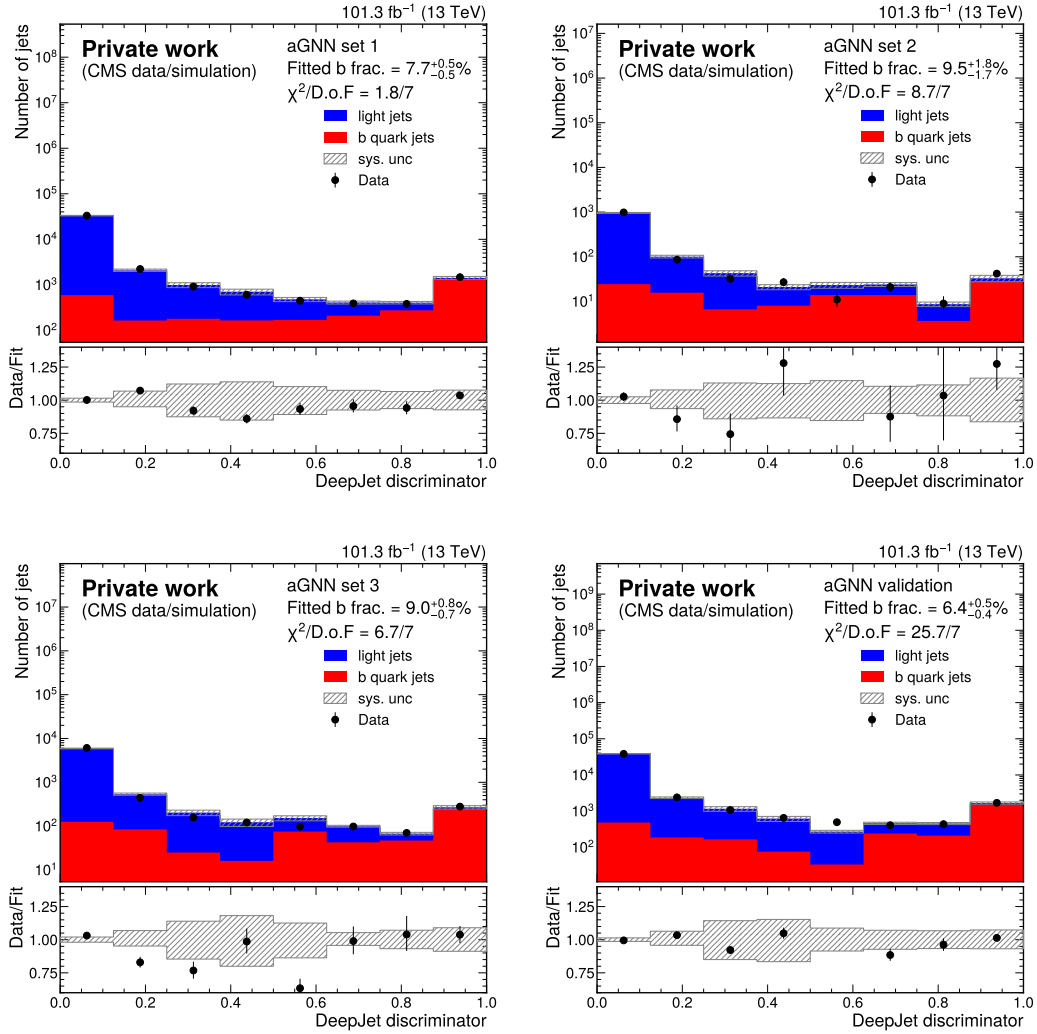


Figure B.80: Estimation of the b jet fraction of control regions defined using the GNN tagger for flavor-aligned models evaluated using H_T -triggered data with pixel Phase-1 geometry.

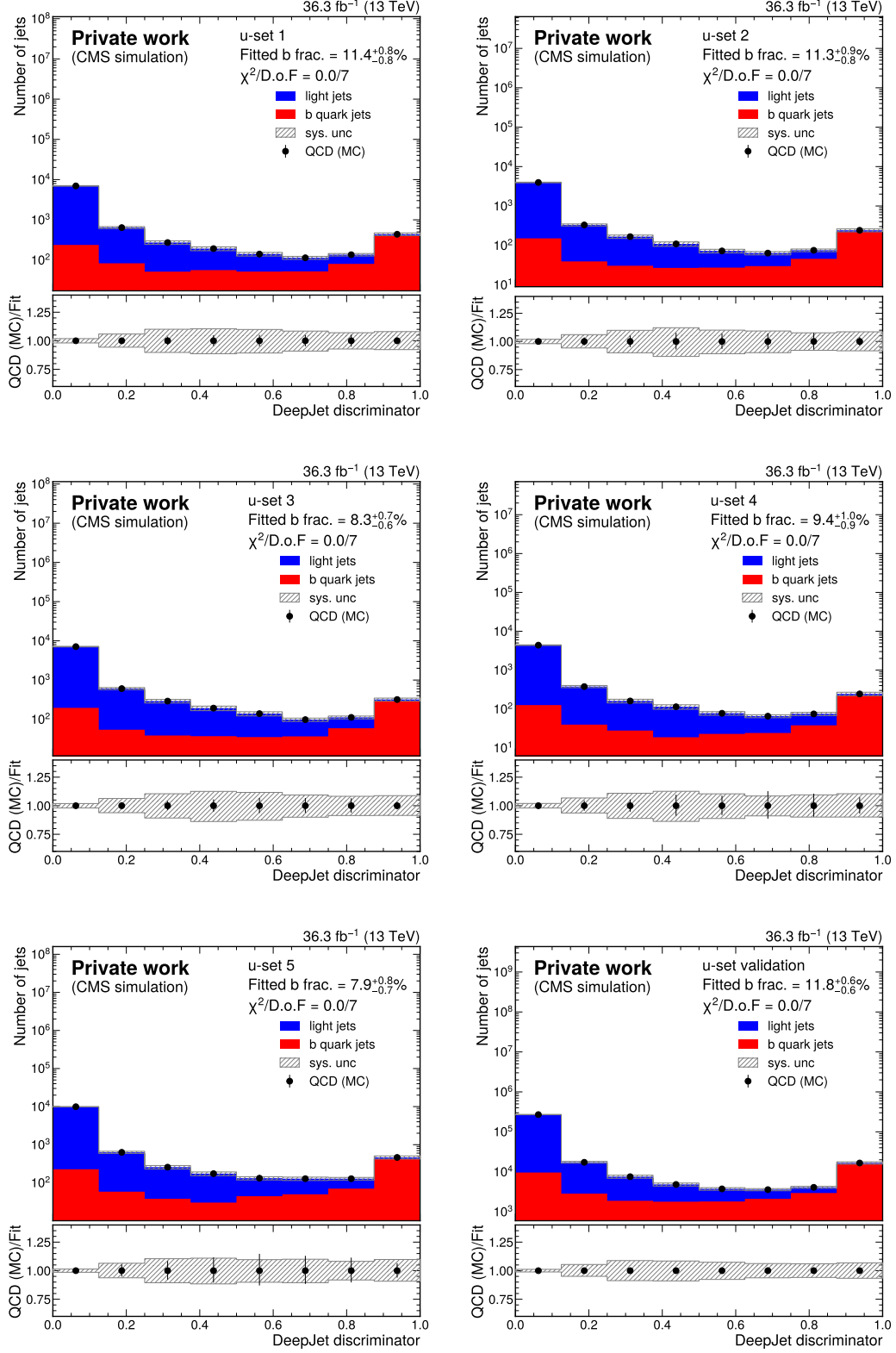


Figure B.81: Estimation of the b jet fraction of control regions defined using the model-agnostic tagger for unflavored models evaluated using QCD MC samples with pixel Phase-0 geometry.

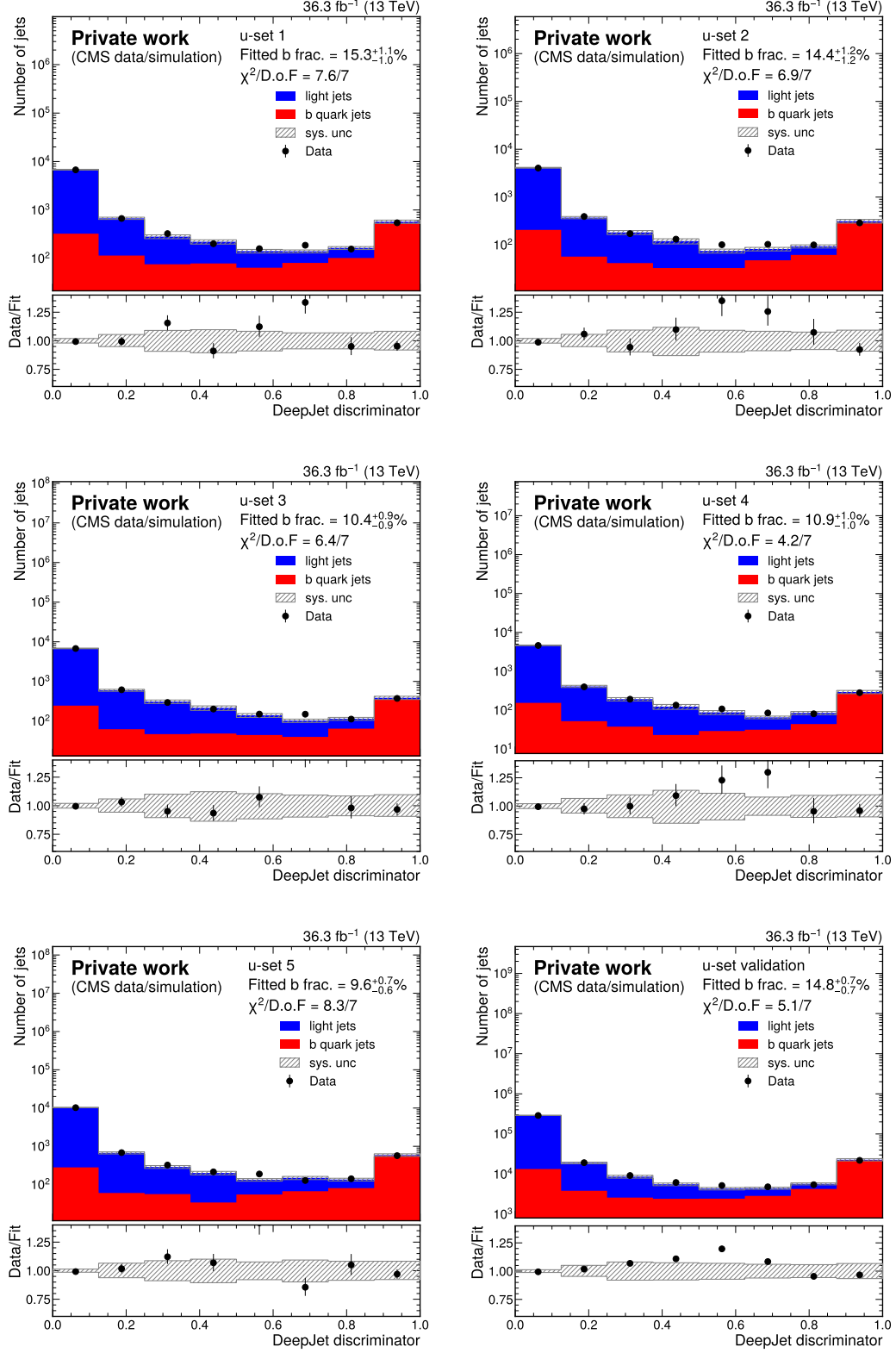


Figure B.82: Estimation of the b jet fraction of control regions defined using the model-agnostic tagger for unflavored models evaluated using H_T -triggered data with pixel Phase-0 geometry.

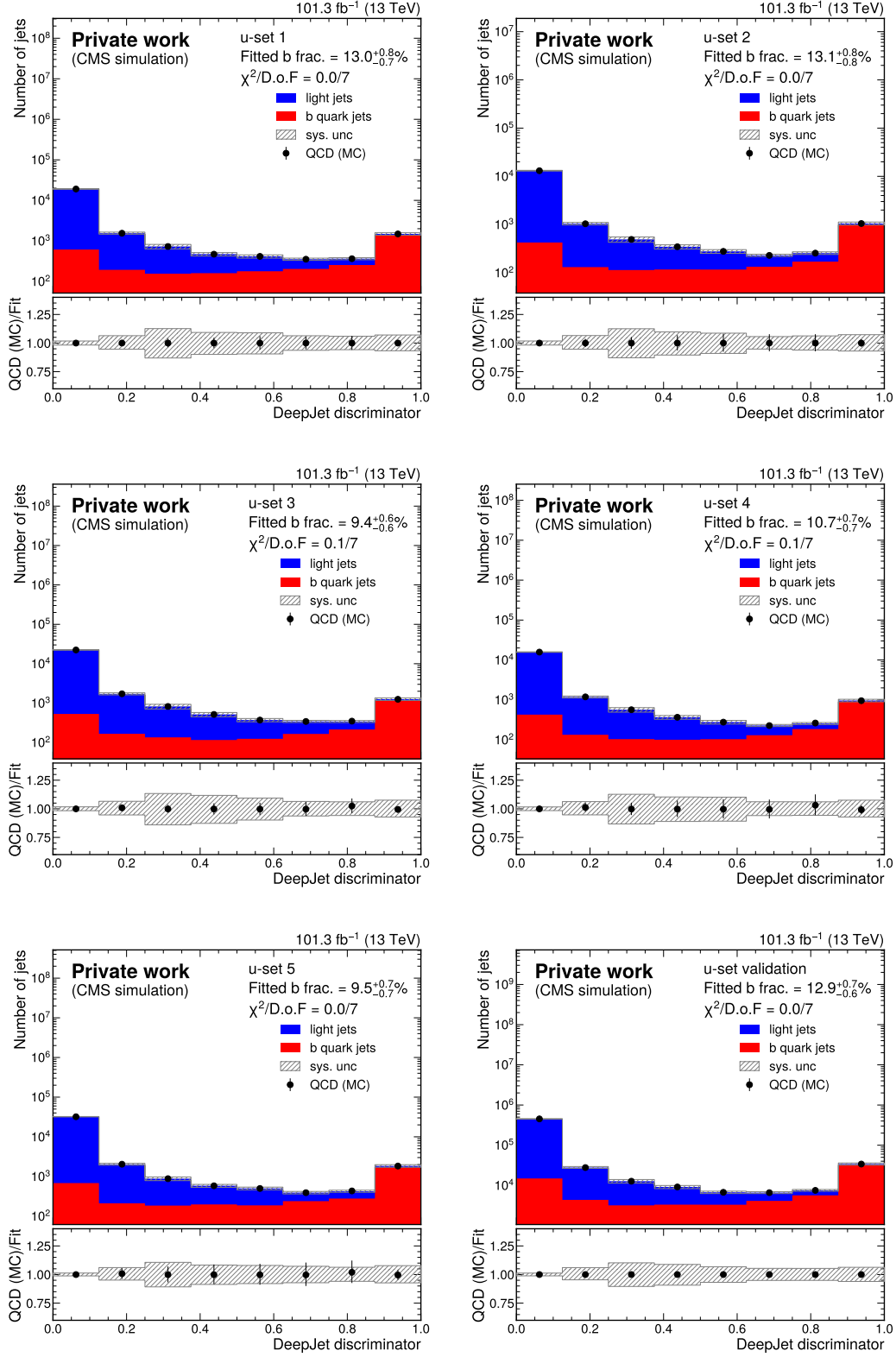


Figure B.83: Estimation of the b jet fraction of control regions defined using the model-agnostic tagger for unflavored models evaluated using QCD MC samples with pixel Phase-1 geometry.

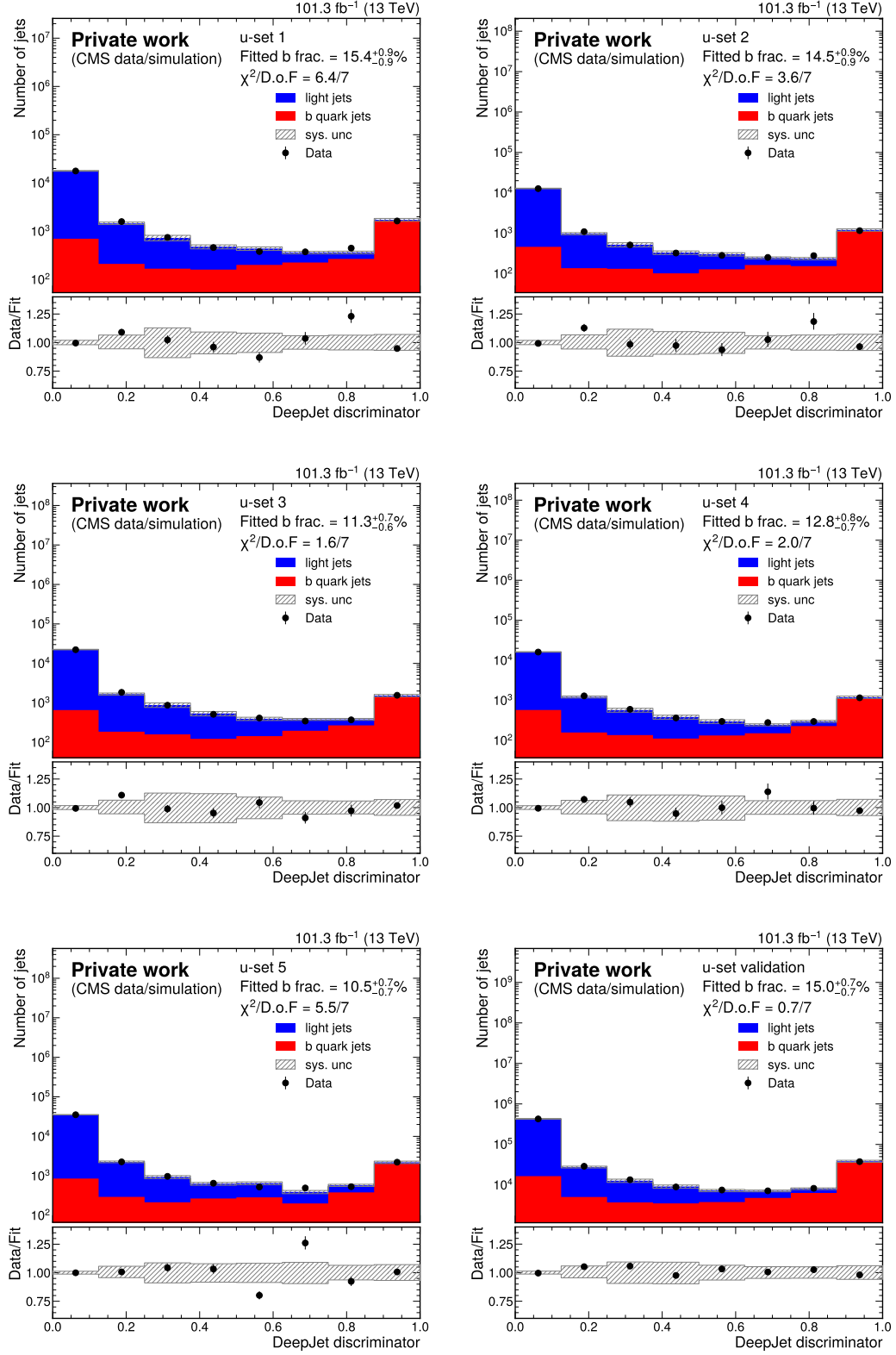


Figure B.84: Estimation of the b jet fraction of control regions defined using the model-agnostic tagger for unflavored models evaluated using H_T -triggered data with pixel Phase-1 geometry.

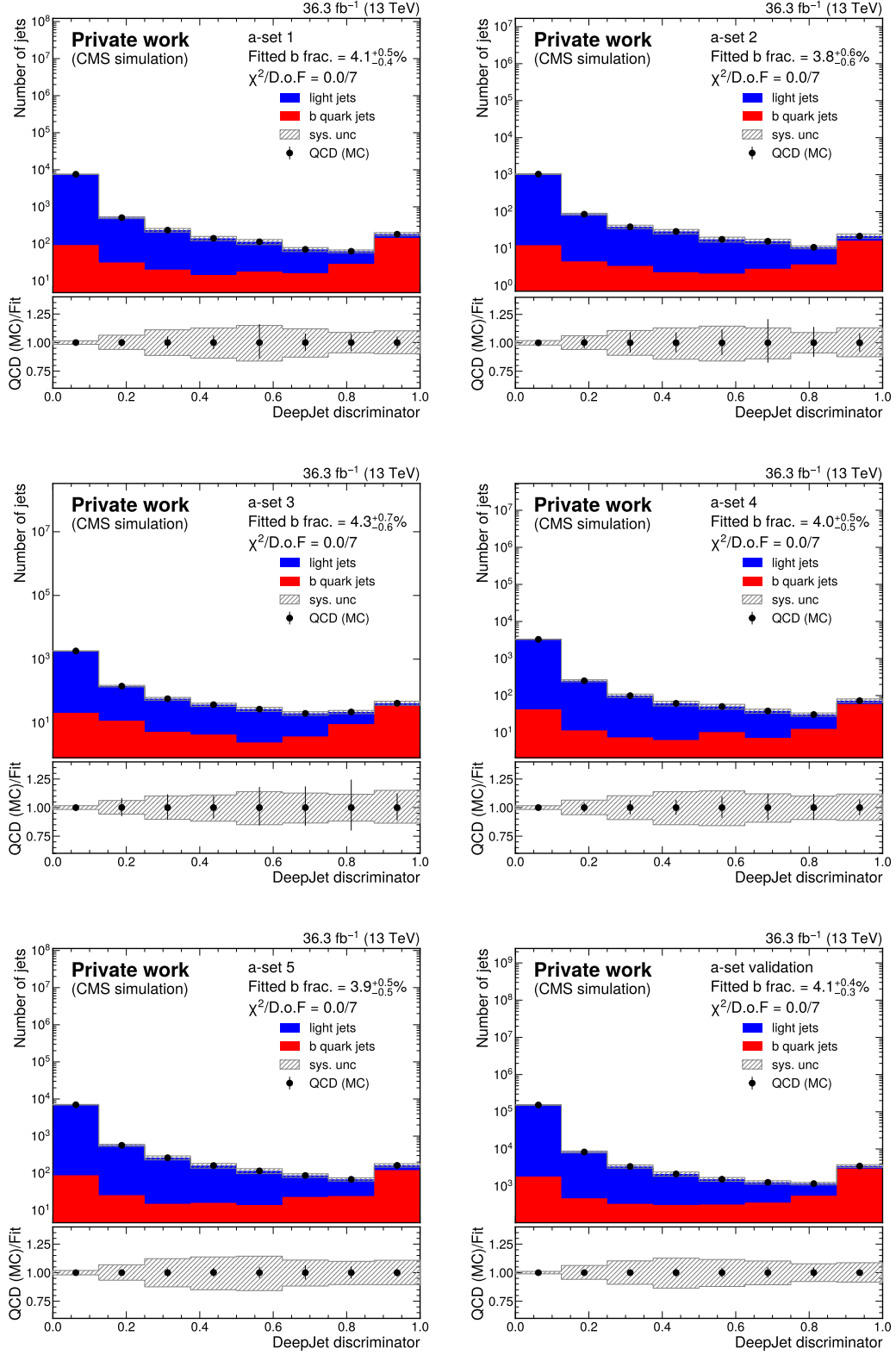


Figure B.85: Estimation of the b jet fraction of control regions defined using the model-agnostic tagger for flavor-aligned models evaluated using QCD MC samples with pixel Phase-0 geometry.

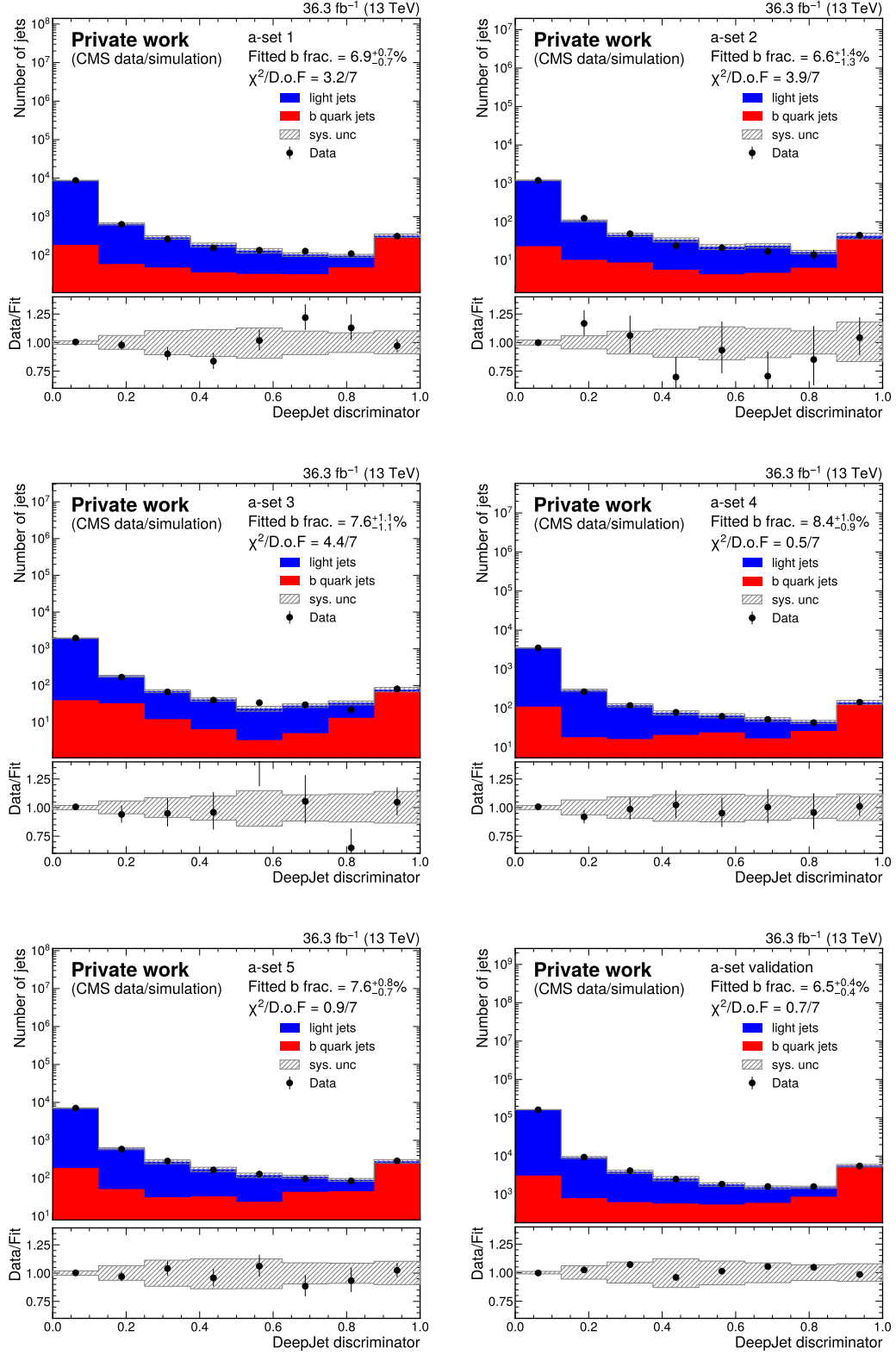


Figure B.86: Estimation of the b jet fraction of control regions defined using the model-agnostic tagger for flavor-aligned models evaluated using H_T -triggered data with pixel Phase-0 geometry.

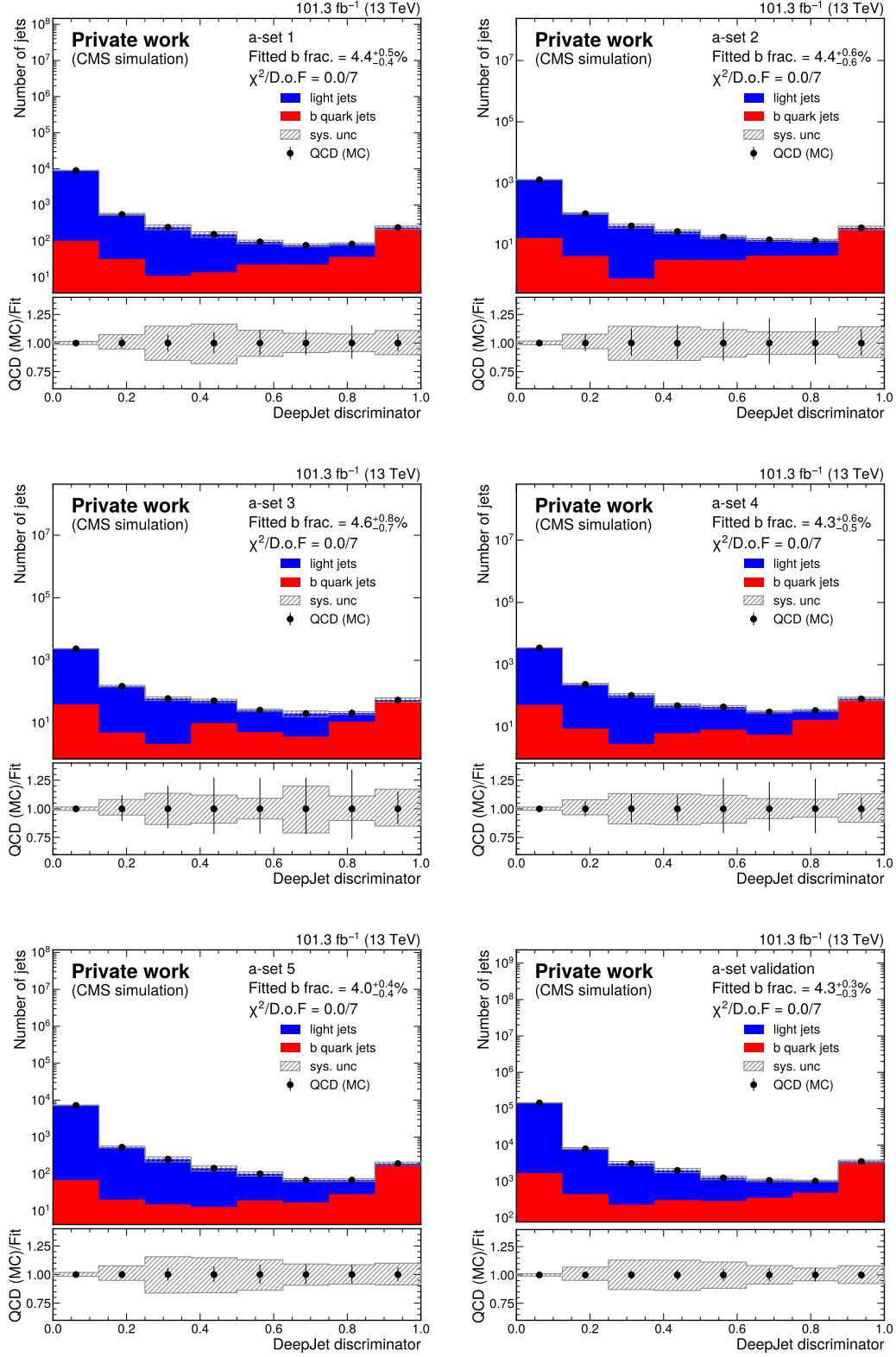


Figure B.87: Estimation of the b jet fraction of control regions defined using the model-agnostic tagger for flavor-aligned models evaluated using QCD MC samples with pixel Phase-1 geometry.

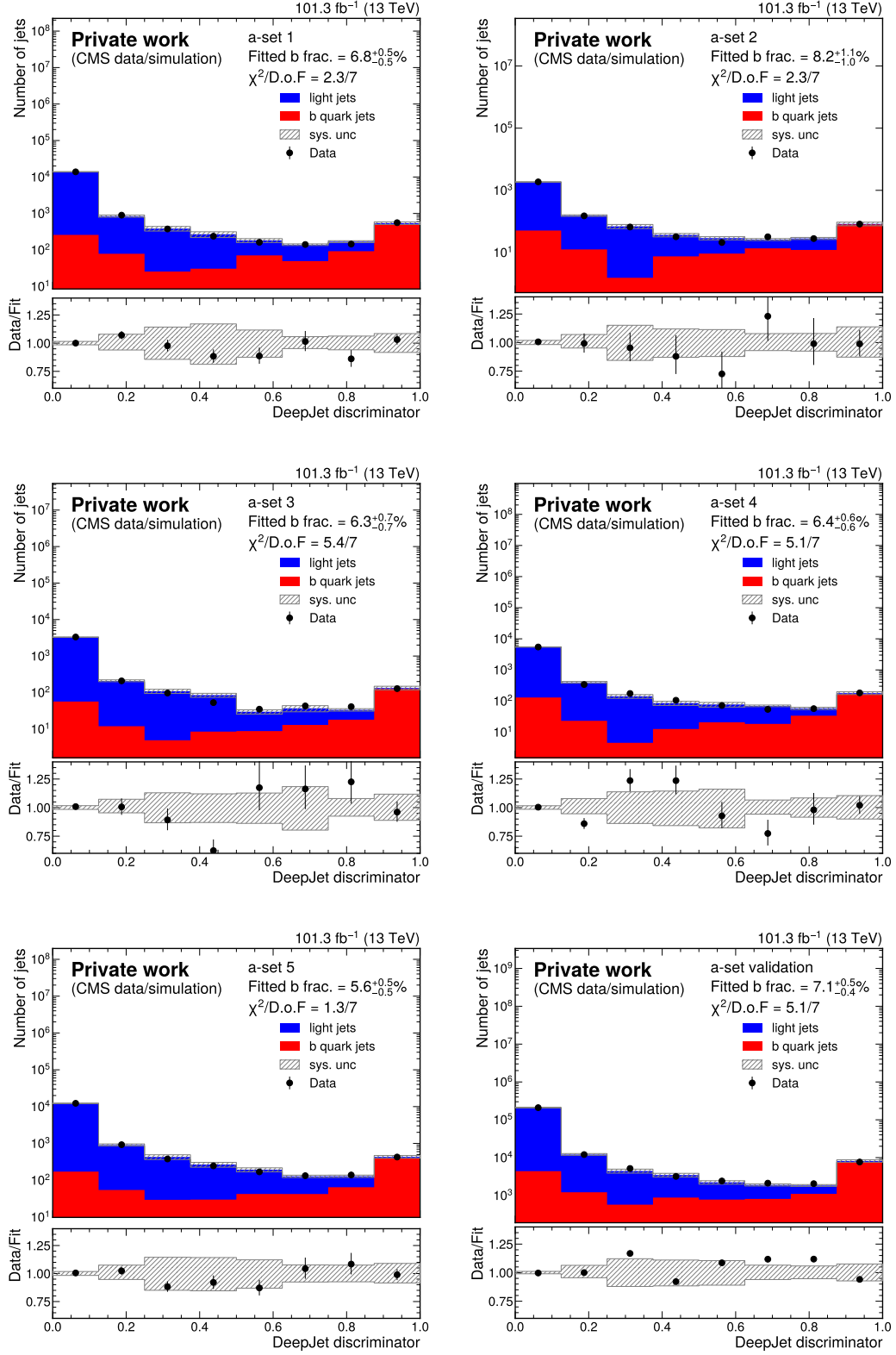


Figure B.88: Estimation of the b jet fraction of control regions defined using the model-agnostic tagger for flavor-aligned models evaluated using H_T -triggered data with pixel Phase-1 geometry.

B.6 Limits results

This section displays the expected and observed limits of various signal models, separated using the tagging method used. For these plots, the brazilian flag layout will be used.

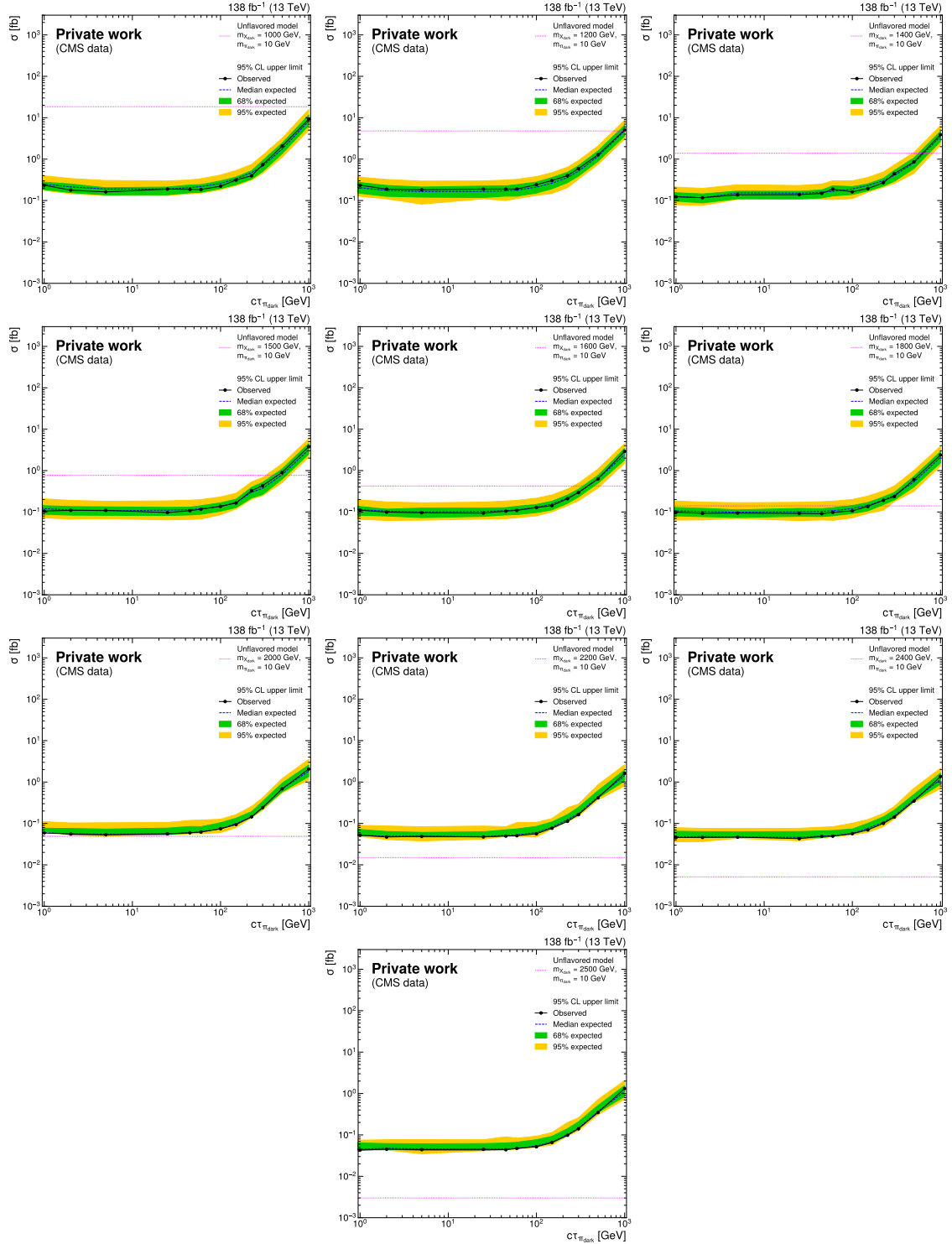


Figure B.89: Expected and observed limits of unflavored signal models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$, obtained using GNN-based tagging.

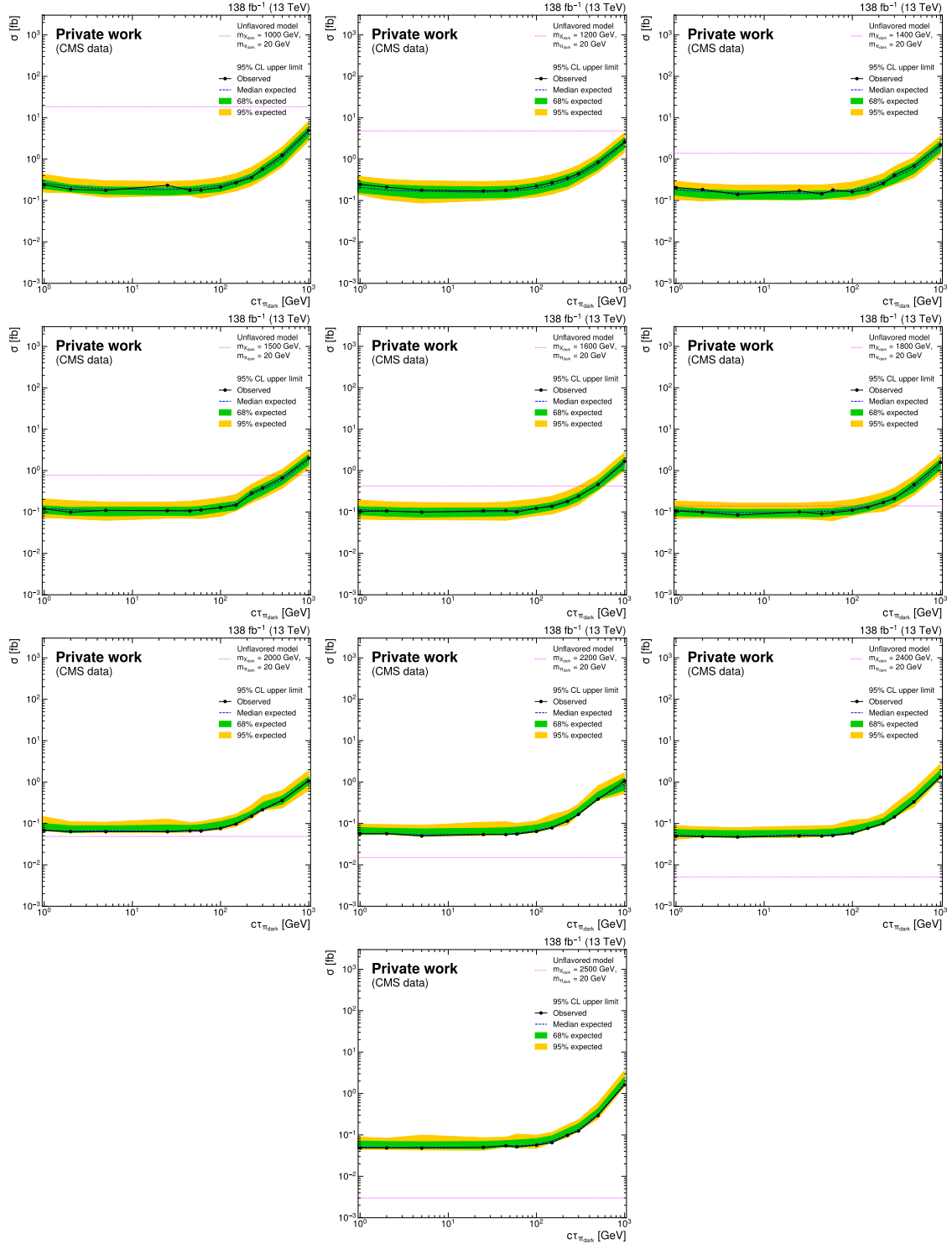


Figure B.90: Expected and observed limits of unflavored signal models with $m_{\pi_{\text{dark}}} = 20 \text{ GeV}$, obtained using GNN-based tagging.

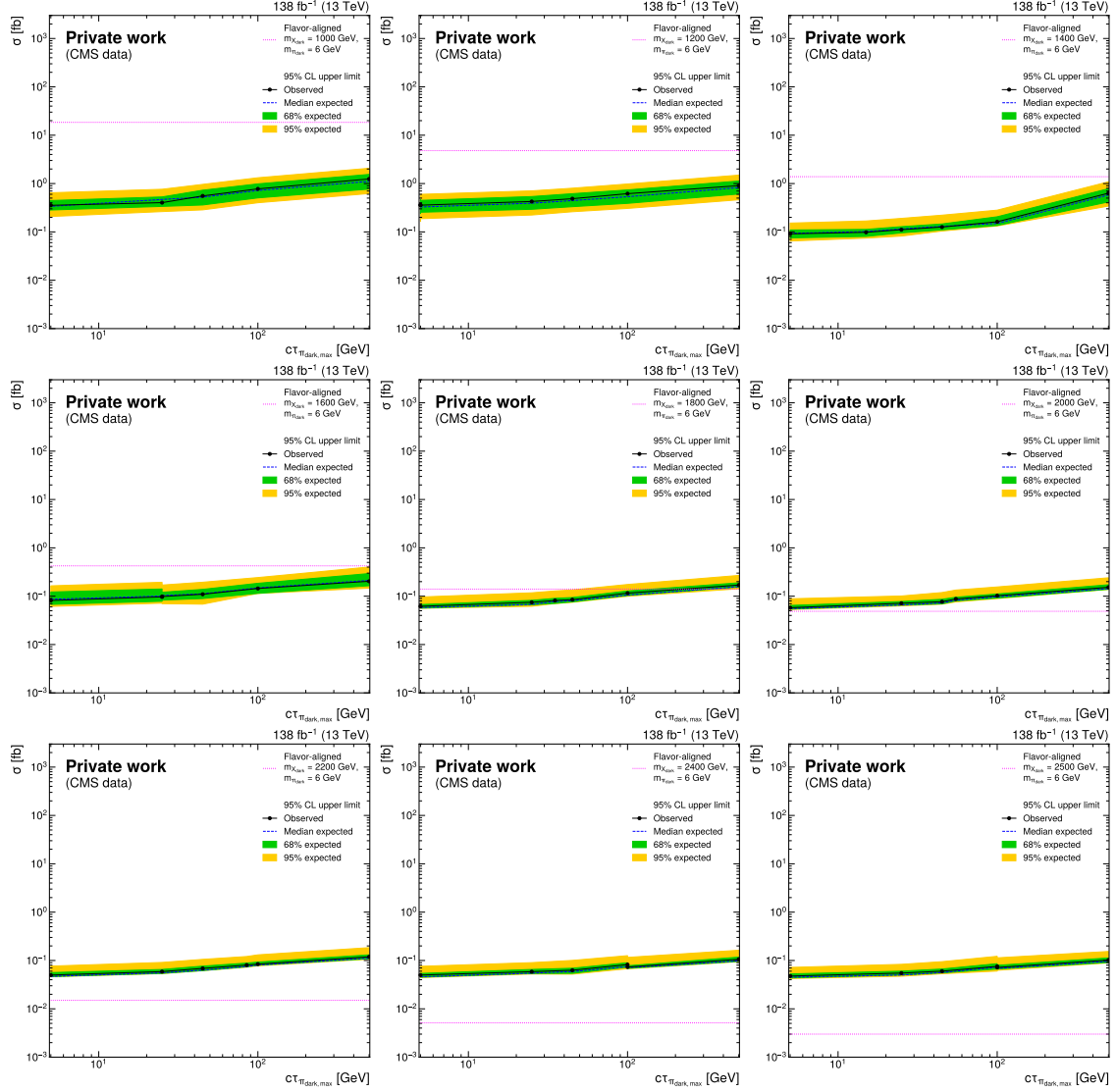


Figure B.91: Expected and observed limits of flavor-aligned signal models with $m_{\pi_{\text{dark}}} = 6 \text{ GeV}$, obtained using GNN-based tagging.

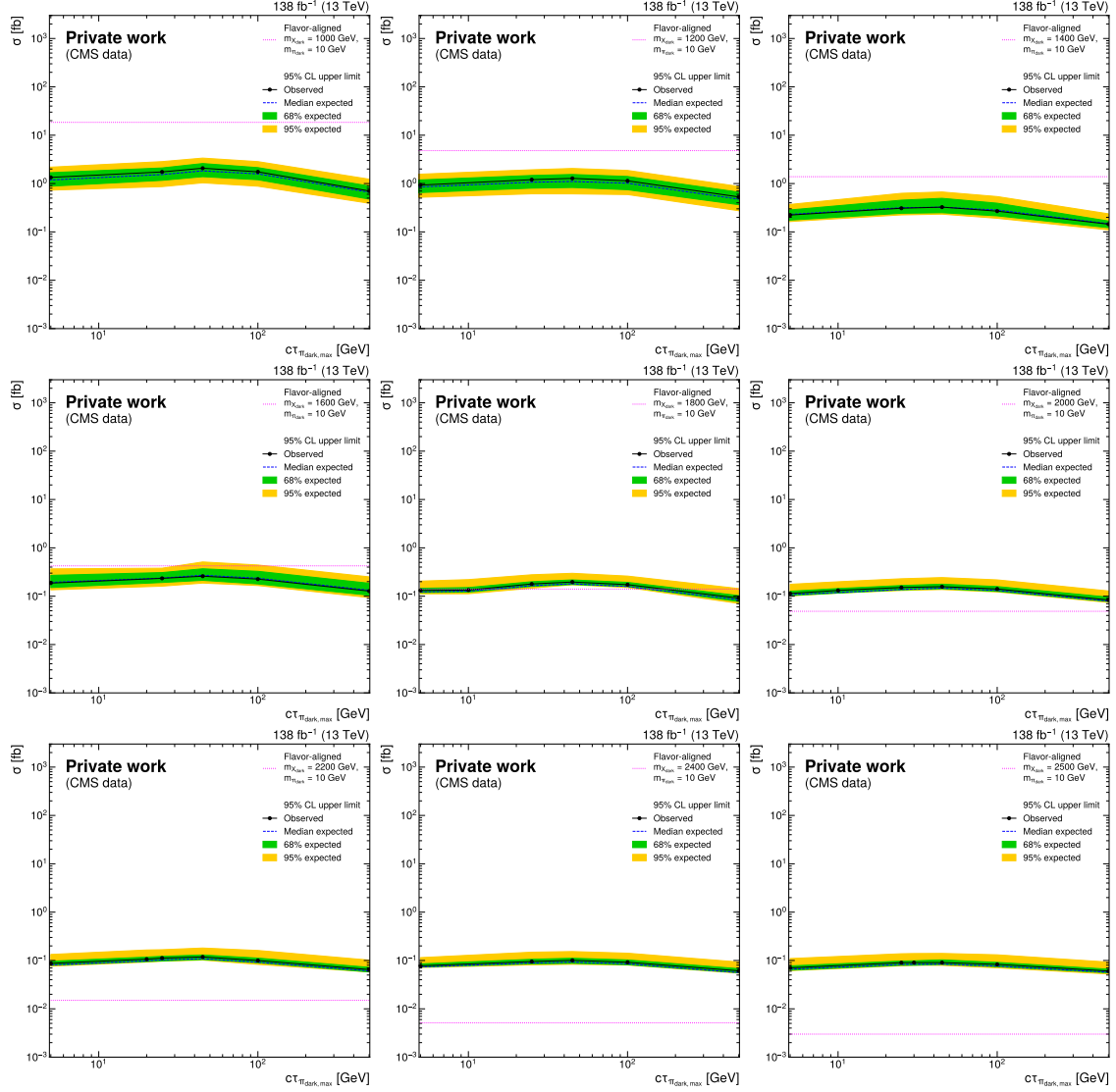


Figure B.92: Expected and observed limits of flavor-aligned signal models with $m_{\pi_{\text{dark}}} = 10$ GeV, obtained using GNN-based tagging.

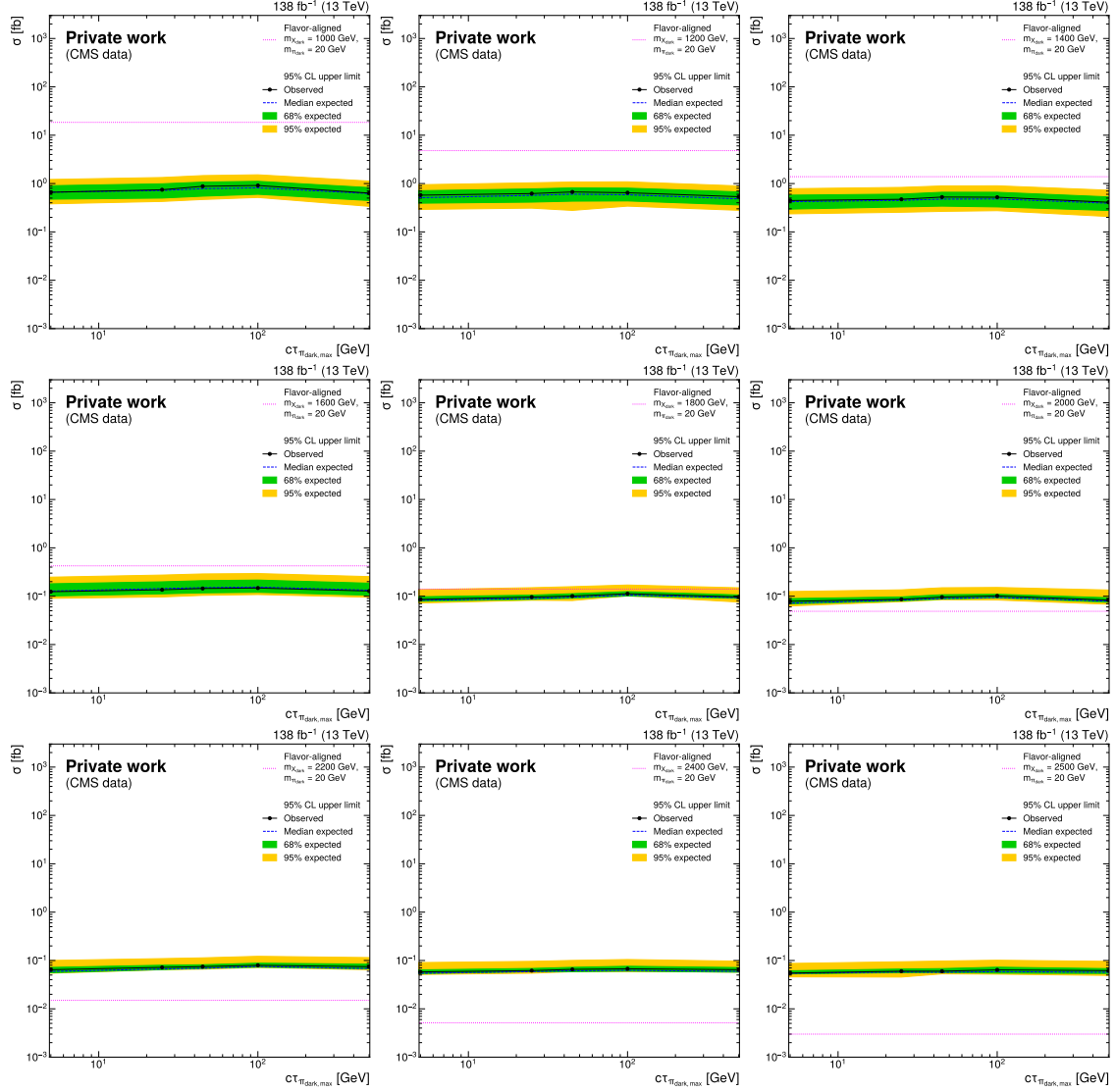


Figure B.93: Expected and observed limits of flavor-aligned signal models with $m_{\pi_{\text{dark}}} = 20$ GeV, obtained using GNN-based tagging.

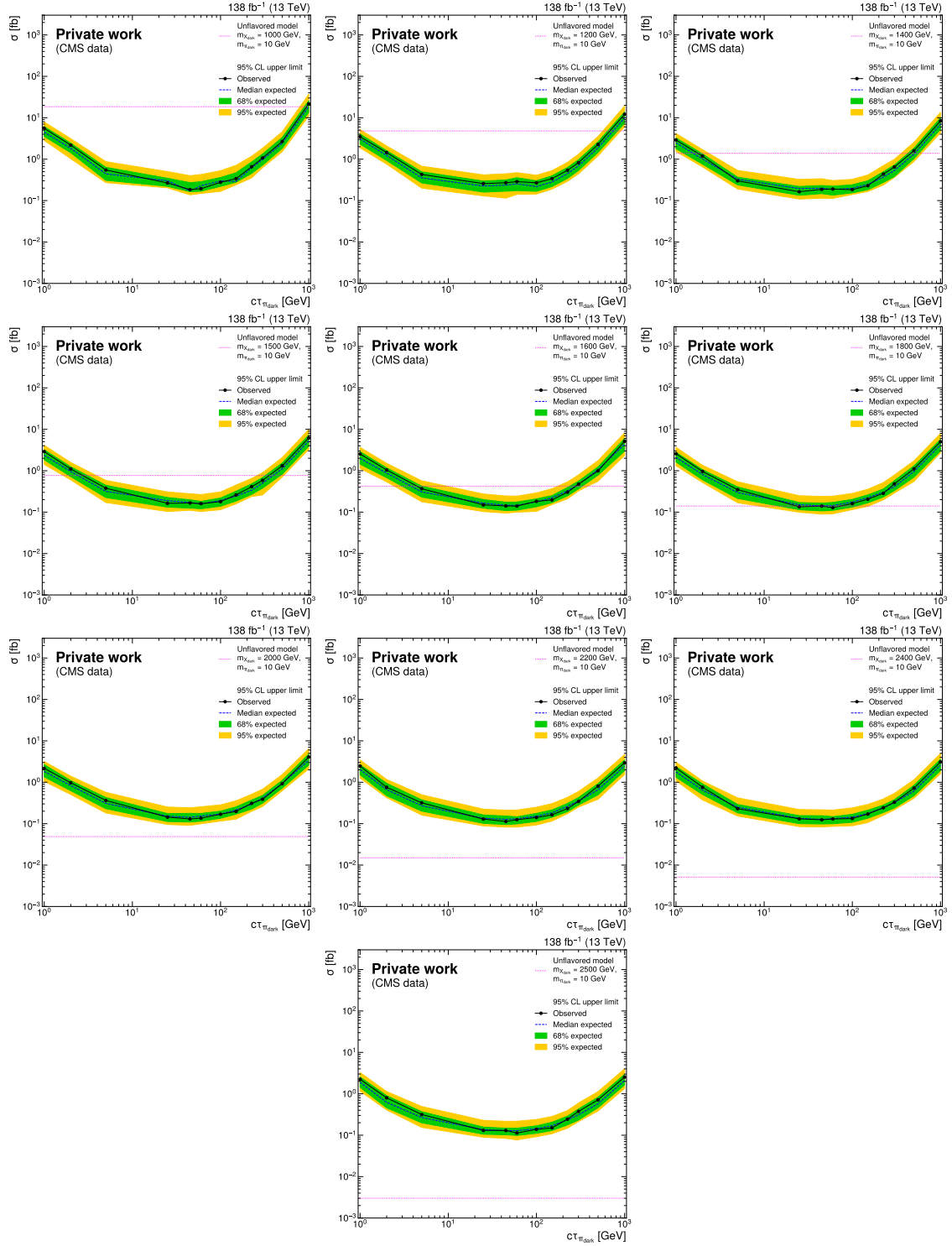


Figure B.94: Expected and observed limits of unflavored signal models with $m_{\pi_{\text{dark}}} = 10 \text{ GeV}$, obtained using model-agnostic tagging.

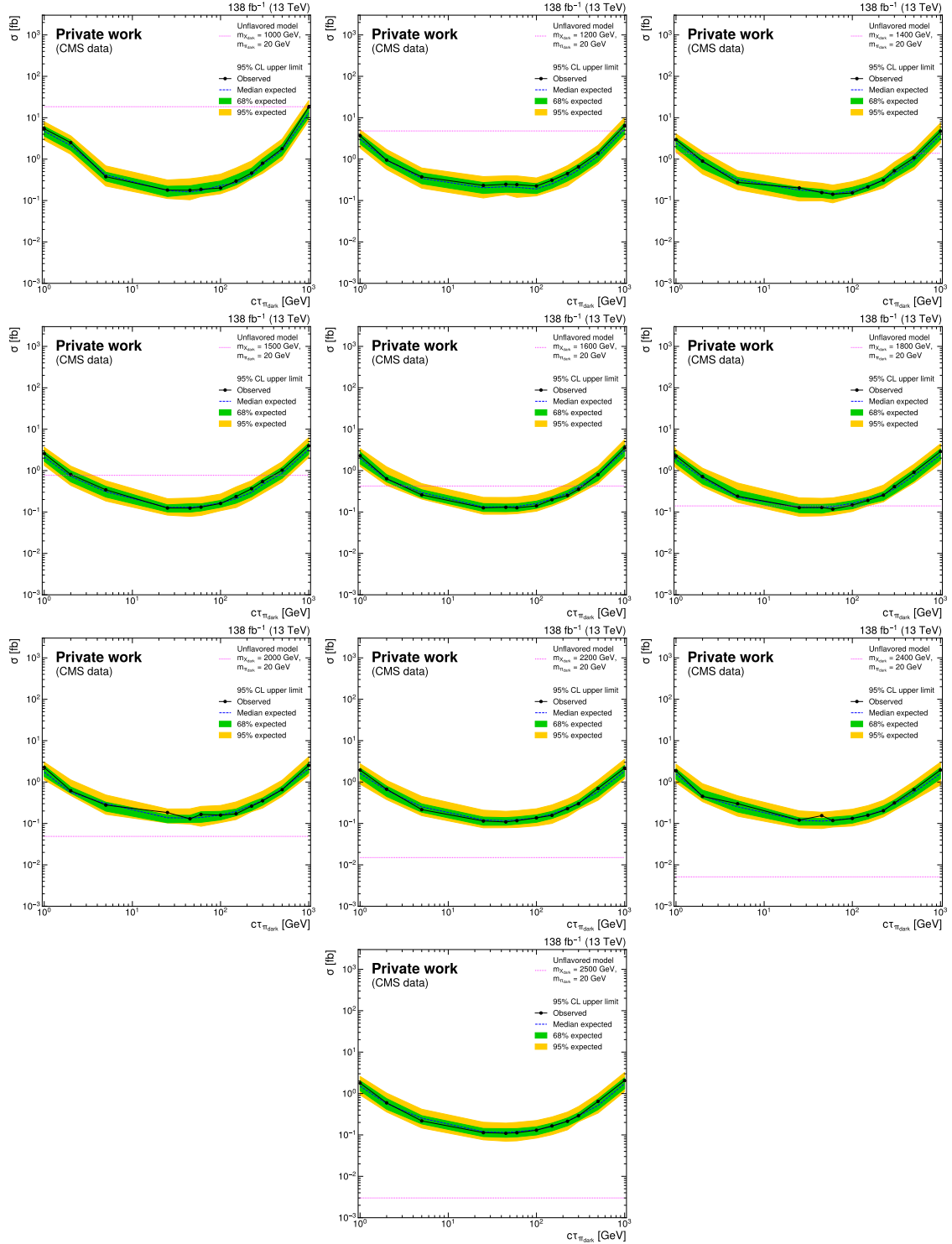


Figure B.95: Expected and observed limits of unflavored signal models with $m_{\pi_{\text{dark}}} = 20 \text{ GeV}$, obtained using model-agnostic tagging.

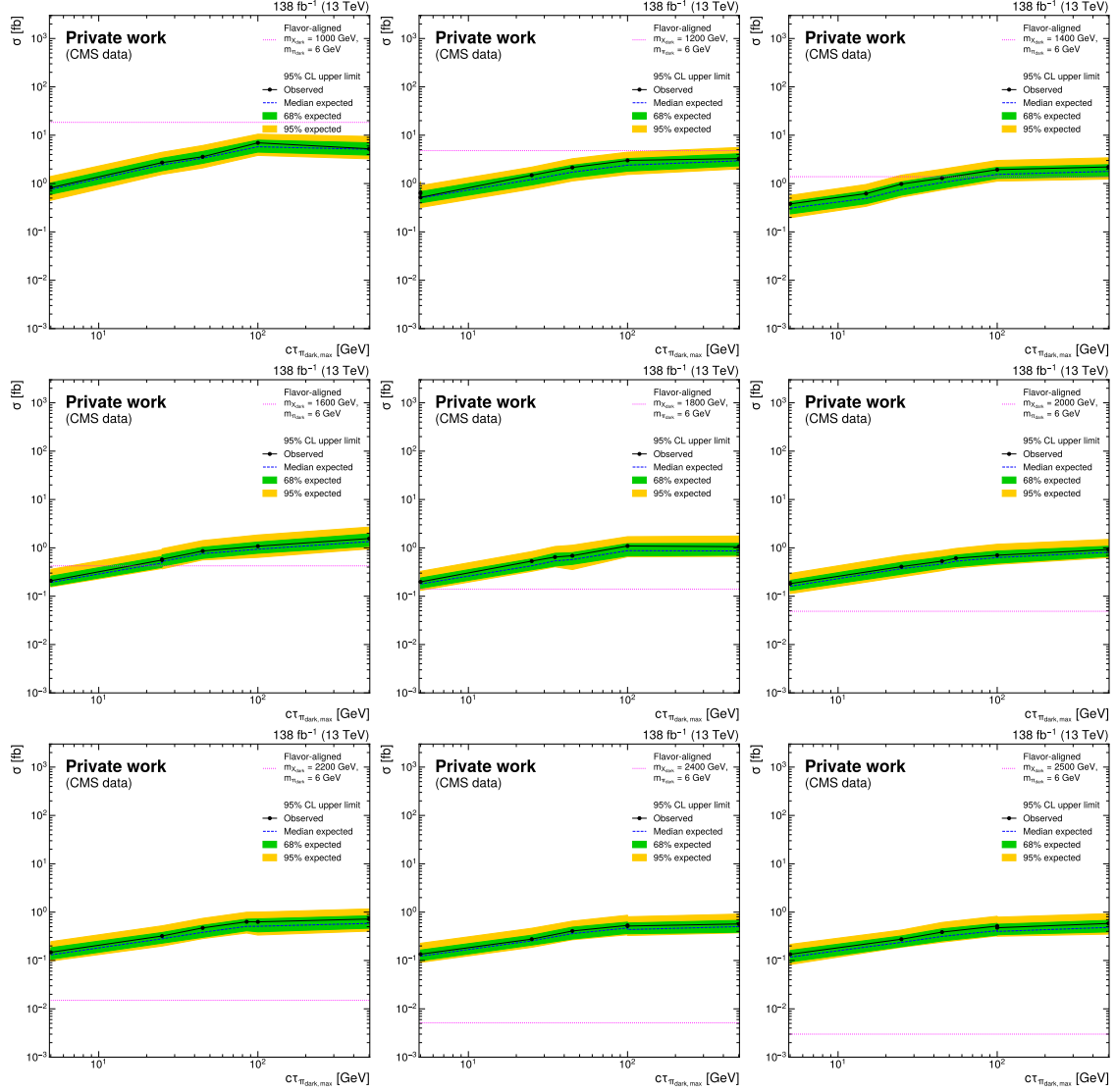


Figure B.96: Expected and observed limits of flavor-aligned signal models with $m_{\pi_{\text{dark}}} = 6 \text{ GeV}$, obtained using model-agnostic tagging.

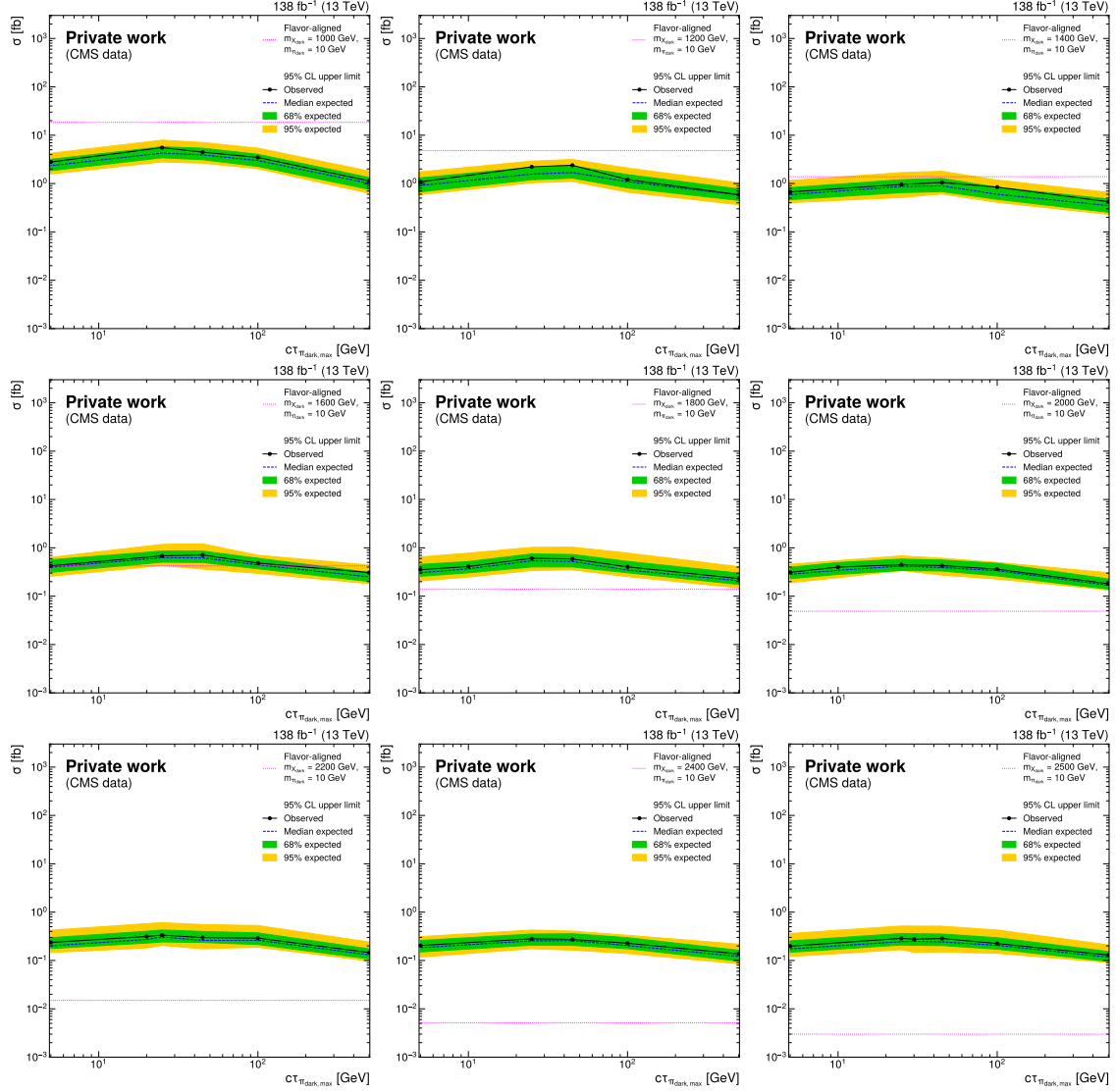


Figure B.97: Expected and observed limits of flavor-aligned signal models with $m_{\pi_{\text{dark}}} = 10$ GeV, obtained using model-agnostic tagging.

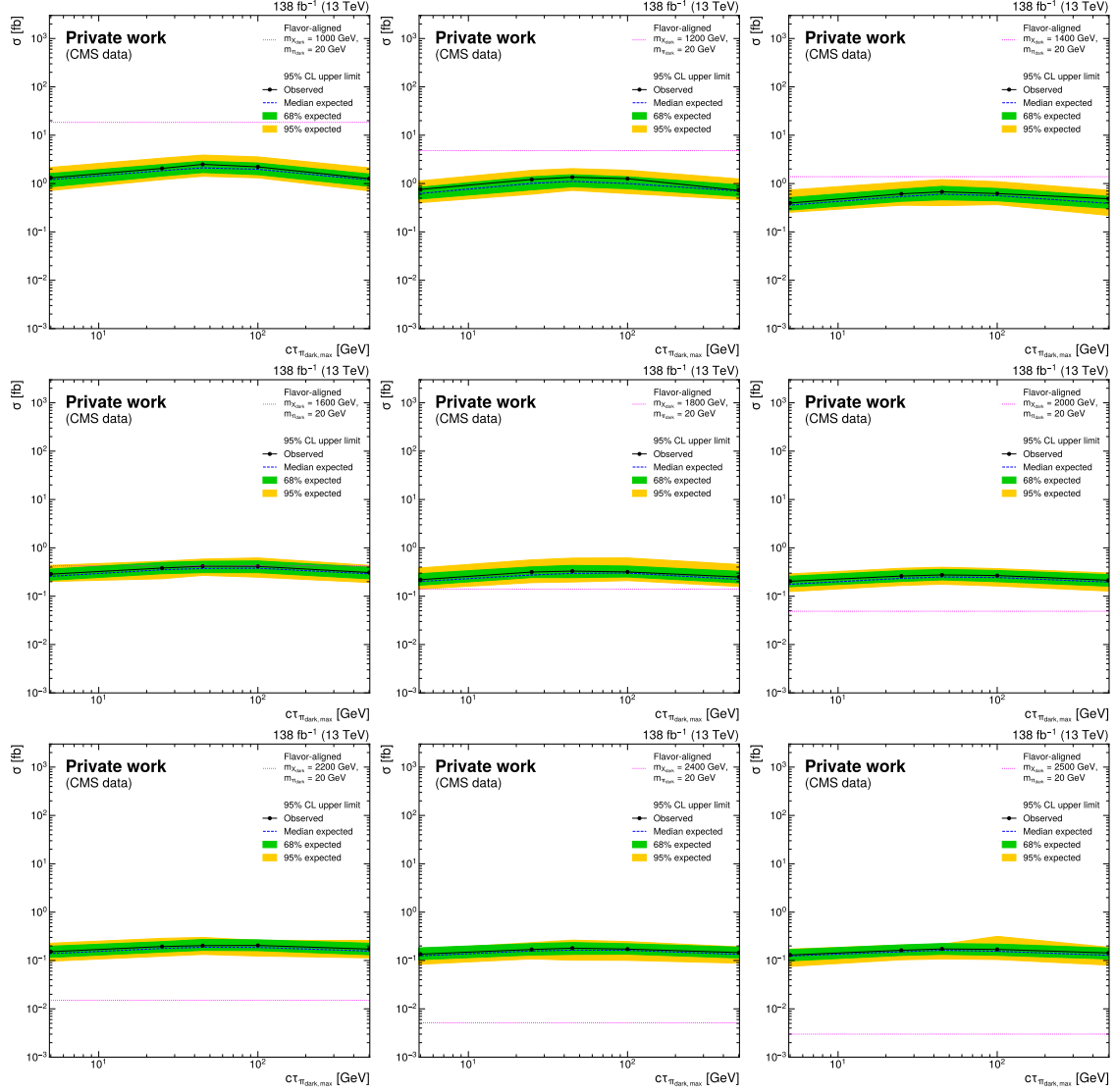


Figure B.98: Expected and observed limits of flavor-aligned signal models with $m_{\pi_{\text{dark}}} = 20 \text{ GeV}$, obtained using model-agnostic tagging.

Acronyms

2D	two-dimensional.	31
APD	avalanche photodiodes.	34
BCM1F	fast beam condition monitor.	43
BPIX	barrel pixel.	30
CERN	Centre Européen pour la Recherche Nucléaire.	22
CHS	charged hadron subtraction.	63
CMS	compact muon solenoid.	25
CR	control region.	147
CSC	Cathod strip chamber.	38
DA	deterministic annealing.	52
DM	dark matter.	1
DNN	deep neural network.	78
DT	drift tube.	38
EB	ECAL barrel.	34
ECAL	electromagnetic calorimeter.	33
EE	ECAL endcap.	34
EMJ	emerging jet.	14
EW	electroweak.	7
FPIX	forward pixel.	30
GNN	graph neural network.	113
HB	HCAL barrel.	36
HCAL	hadron calorimeter.	35

HE HCAL endcap. 36

HF HCAL forwards. 36

HLT high-level trigger. 42

HO HCAL outer. 36

IP impact parameter. 55

JEC jet energy correction. 184

JER jet energy resolution. 184

L1T level-1 trigger. 41

LHC large hadron collider. 22

MC Monte Carlo. 79

ML machine learning. 106

PDF parton distribution function. 11

PF particle-flow. 45

PLT pixel luminosity telescope. 43

pp proton-proton. 23

QCD quantum chromodynamics. 9

QFT quantum field theory. 2

RPC Resistive plate chamber. 39

SE ECAL preshower. 34

SM standard model. 1

SR signal region. 147

TEC tracker endcap. 31

TIB tracker inner barrel. 31

TID tracker inner disk. 31

TOB tracker outer barrel. 31

WIMP weakly interacting massive particle. 1

Bibliography

- [1] ATLAS Collaboration, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”, *Phys. Lett. B* **716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.
- [2] CMS Collaboration, “Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC”, *Phys. Lett. B* **716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.
- [3] V. C. Rubin, N. Thonnard, and W. K. Ford, Jr., “Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 ($R = 4$ kpc) to UGC 2885 ($R = 122$ kpc)”, *Astrophys. J.* **238** (1980) 471, doi:10.1086/158003.
- [4] M. Persic, P. Salucci, and F. Stel, “The universal rotation curve of spiral galaxies: I. The dark matter connection”, *Mon. Not. Roy. Astron. Soc.* **281** (1996) 27, doi:10.1093/mnras/278.1.27, arXiv:astro-ph/9506004.
- [5] D. Clowe et al., “A direct empirical proof of the existence of dark matter”, *Astrophys. J.* **648** (2006) L109, doi:10.1086/508162, arXiv:astro-ph/0608407.
- [6] DES Collaboration, “Dark Energy Survey year 1 results: curved-sky weak lensing mass map”, *Mon. Not. Roy. Astron. Soc.* **475** (2018) 3165, doi:10.1093/mnras/stx3363, arXiv:1708.01535.
- [7] Planck Collaboration, “Planck 2018 results. VI. Cosmological parameters”, *Astron. Astrophys.* **641** (2020) A6, doi:10.1051/0004-6361/201833910, arXiv:1807.06209.
- [8] G. Jungman, M. Kamionkowski, and K. Griest, “Supersymmetric dark matter”, *Phys. Rept.* **267** (1996) 195–373, doi:10.1016/0370-1573(95)00058-5, arXiv:hep-ph/9506380.
- [9] XENON Collaboration, “First Dark Matter Search with Nuclear Recoils from the XENONnT Experiment”, *Phys. Rev. Lett.* **131** (2023), no. 4, 041003, doi:10.1103/PhysRevLett.131.041003, arXiv:2303.14729.
- [10] LZ Collaboration, “First Dark Matter Search Results from the LUX-ZEPLIN (LZ) Experiment”, *Phys. Rev. Lett.* **131** (2023), no. 4, 041002, doi:10.1103/PhysRevLett.131.041002, arXiv:2207.03764.

- [11] M. J. Strassler and K. M. Zurek, “Echoes of a hidden valley at hadron colliders”, *Phys. Lett. B* **651** (2007) 374, doi:10.1016/j.physletb.2007.06.055, arXiv:hep-ph/0604261.
- [12] K. Petraki and R. R. Volkas, “Review of asymmetric dark matter”, *Int. J. Mod. Phys. A* **28** (2013) 1330028, doi:10.1142/S0217751X13300287, arXiv:1305.4939.
- [13] S. J. Lonsdale, M. Schroor, and R. R. Volkas, “Asymmetric Dark Matter and the hadronic spectra of hidden QCD”, *Phys. Rev. D* **96** (2017), no. 5, 055027, doi:10.1103/PhysRevD.96.055027, arXiv:1704.05213.
- [14] H. Beauchesne, E. Bertuzzo, and G. Grilli di Cortona, “Dark matter in Hidden Valley models with stable and unstable light dark mesons”, *JHEP* **04** (2019) 118, doi:10.1007/JHEP04(2019)118, arXiv:1809.10152.
- [15] P. Schwaller, D. Stolarski, and A. Weiler, “Emerging Jets”, *JHEP* **05** (2015) 059, doi:10.1007/JHEP05(2015)059, arXiv:1502.05409.
- [16] T. Cohen, M. Lisanti, and H. K. Lou, “Semivisible Jets: Dark Matter Undercover at the LHC”, *Phys. Rev. Lett.* **115** (2015), no. 17, 171804, doi:10.1103/PhysRevLett.115.171804, arXiv:1503.00009.
- [17] F. J. Dyson, “The radiation theories of tomonaga, schwinger, and feynman”, *Phys. Rev.* **75** (Feb, 1949) 486–502, doi:10.1103/PhysRev.75.486.
- [18] C. N. Yang and R. L. Mills, “Conservation of isotopic spin and isotopic gauge invariance”, *Phys. Rev.* **96** (Oct, 1954) 191–195, doi:10.1103/PhysRev.96.191.
- [19] G. Hooft, “Renormalizable lagrangians for massive yang-mills fields”, *Nuclear Physics B* **35** (1971), no. 1, 167–188, doi:10.1016/0550-3213(71)90139-8.
- [20] F. Englert and R. Brout, “Broken symmetry and the mass of gauge vector mesons”, *Phys. Rev. Lett.* **13** (Aug, 1964) 321–323, doi:10.1103/PhysRevLett.13.321.
- [21] P. W. Higgs, “Broken symmetries and the masses of gauge bosons”, *Phys. Rev. Lett.* **13** (Oct, 1964) 508–509, doi:10.1103/PhysRevLett.13.508.
- [22] J. Goldstone, A. Salam, and S. Weinberg, “Broken symmetries”, *Phys. Rev.* **127** (Aug, 1962) 965–970, doi:10.1103/PhysRev.127.965.
- [23] Particle Data Group Collaboration, “Review of Particle Physics”, *PTEP* **2022** (2022) 083C01, doi:10.1093/ptep/ptac097.
- [24] S. L. Glashow, “The renormalizability of vector meson interactions”, *Nuclear Physics* **10** (1959) 107–117, doi:10.1016/0029-5582(59)90196-8.
- [25] A. Salam, “Weak and Electromagnetic Interactions”, *Conf. Proc. C* **680519** (1968) 367–377, doi:10.1142/9789812795915_0034.

- [26] S. Weinberg, “A model of leptons”, *Phys. Rev. Lett.* **19** (Nov, 1967) 1264–1266, doi:10.1103/PhysRevLett.19.1264.
- [27] S. L. Glashow, “Partial-symmetries of weak interactions”, *Nuclear Physics* **22** (1961), no. 4, 579–588, doi:10.1016/0029-5582(61)90469-2.
- [28] S. L. Glashow, J. Iliopoulos, and L. Maiani, “Weak interactions with lepton-hadron symmetry”, *Phys. Rev. D* **2** (Oct, 1970) 1285–1292, doi:10.1103/PhysRevD.2.1285.
- [29] M. Kobayashi and T. Maskawa, “CP-Violation in the Renormalizable Theory of Weak Interaction”, *Progress of Theoretical Physics* **49** (02, 1973) 652–657, doi:10.1143/PTP.49.652, arXiv:https://academic.oup.com/ptp/article-pdf/49/2/652/5257692/49-2-6
- [30] BaBar Collaboration, “The BABAR detector”, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **479** (feb, 2002) 1–116, doi:10.1016/S0168-9002(01)02012-5.
- [31] Belle Collaboration, “Belle II technical design report”, 2010.
- [32] BaBar, Belle Collaboration, “The Physics of the B Factories”, *Eur. Phys. J. C* **74** (2014) 3026, doi:10.1140/epjc/s10052-014-3026-9, arXiv:1406.6311.
- [33] H. Fritzsch, M. Gell-Mann, and H. Leutwyler, “Advantages of the color octet gluon picture”, *Physics Letters B* **47** (1973), no. 4, 365–368, doi:10.1016/0370-2693(73)90625-4.
- [34] D. J. Gross and F. Wilczek, “Ultraviolet behavior of non-abelian gauge theories”, *Phys. Rev. Lett.* **30** (Jun, 1973) 1343–1346, doi:10.1103/PhysRevLett.30.1343.
- [35] H. D. Politzer, “Reliable perturbative results for strong interactions?”, *Phys. Rev. Lett.* **30** (Jun, 1973) 1346–1349, doi:10.1103/PhysRevLett.30.1346.
- [36] J. Greensite, “An Introduction to the Confinement Problem”. Springer Berlin, Heidelberg, 1616-6361. doi:10.1007/978-3-642-14382-3.
- [37] G. Altarelli and G. Parisi, “Asymptotic freedom in parton language”, *Nuclear Physics B* **126** (1977), no. 2, 298–318, doi:10.1016/0550-3213(77)90384-4.
- [38] S. D. Drell and T.-M. Yan, “Massive lepton-pair production in hadron-hadron collisions at high energies”, *Phys. Rev. Lett.* **25** (Aug, 1970) 316–320, doi:10.1103/PhysRevLett.25.316.
- [39] S. I. Alekhin et al., “DGLAP evolution and parton fits”, doi:10.5170/CERN-2005-014.119.
- [40] J. C. Collins and D. E. Soper, “Parton distribution and decay functions”, *Nuclear Physics B* **194** (1982), no. 3, 445–492, doi:10.1016/0550-3213(82)90021-9.

- [41] S. Ferreres-Solé and T. Sjöstrand, “The space–time structure of hadronization in the Lund model”, *Eur. Phys. J. C* **78** (2018), no. 11, 983,
doi:10.1140/epjc/s10052-018-6459-8, arXiv:1808.04619.
- [42] J. D. Barrow and M. S. Turner, “Baryosynthesis and the origin of galaxies”, *Nature* **291** (Jun, 1981) 469–472, doi:10.1038/291469a0.
- [43] S. A. Bludman, “Baryonic mass fraction in rich clusters and the total mass density in the cosmos”, *Astrophys. J.* **508** (1998) 535, doi:10.1086/306412,
arXiv:astro-ph/9706047.
- [44] CMS Collaboration, “Search for new particles decaying to a jet and an emerging jet”, *JHEP* **02** (2019) 179, doi:10.1007/JHEP02(2019)179, arXiv:1810.10069.
- [45] S. Renner and P. Schwaller, “A flavoured dark sector”, *JHEP* **08** (2018) 052,
doi:10.1007/JHEP08(2018)052, arXiv:1803.08080.
- [46] P. Agrawal, M. Blanke, and K. Gemmler, “Flavored dark matter beyond Minimal Flavor Violation”, *JHEP* **10** (2014) 072, doi:10.1007/JHEP10(2014)072,
arXiv:1405.6709.
- [47] C. Borschensky et al., “Squark and gluino production cross sections in pp collisions at $\sqrt{s} = 13, 14, 33$ and 100 TeV”, *Eur. Phys. J. C* **74** (2014), no. 12, 3174,
doi:10.1140/epjc/s10052-014-3174-y, arXiv:1407.5066.
- [48] CMS Collaboration, “Search for Emerging Jets with full Run 2 data”, CMS Analysis note CMS-AN-21-112, CMS-EXO-22-025, CERN, 2022.
- [49] O. S. Brüning et al., “LHC Design Report”. CERN Yellow Reports: Monographs. CERN, Geneva, 2004. doi:10.5170/CERN-2004-003-V-1.
- [50] ATLAS Collaboration, “The ATLAS Experiment at the CERN Large Hadron Collider”, *JINST* **3** (2008) S08003, doi:10.1088/1748-0221/3/08/S08003. Also published by CERN Geneva in 2010.
- [51] CMS Collaboration, “The CMS experiment at the CERN LHC. The Compact Muon Solenoid experiment”, *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004. Also published by CERN Geneva in 2010.
- [52] LHCb Collaboration, “The LHCb Detector at the LHC”, *JINST* **3** (2008) S08005, doi:10.1088/1748-0221/3/08/S08005. Also published by CERN Geneva in 2010.
- [53] ALICE Collaboration, “The ALICE experiment at the CERN LHC. A Large Ion Collider Experiment”, *JINST* **3** (2008) S08002, doi:10.1088/1748-0221/3/08/S08002. Also published by CERN Geneva in 2010.

- [54] E. Mobs, “The CERN accelerator complex in 2019”,. General Photo.
- [55] CMS Collaboration, “Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS”, *Eur. Phys. J. C* **81** (2021) 800, doi:10.1140/epjc/s10052-021-09538-2, arXiv:2104.01927.
- [56] CMS Collaboration, “CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV”, CMS Physics Analysis Summary CMS-PAS-LUM-17-004, 2018.
- [57] CMS Collaboration, “CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV”, CMS Physics Analysis Summary CMS-PAS-LUM-18-002, 2019.
- [58] V. Klyukhin, “Design and Description of the CMS Magnetic System Model”, *Symmetry* **13** (2021), no. 6, 1052, doi:10.3390/sym13061052, arXiv:2201.10478.
- [59] CMS Collaboration V. Karimäki, et al., “The CMS tracker system project: Technical Design Report”. Technical design report. CMS. CERN, Geneva, 1997.
- [60] CMS Tracker Group Collaboration, “The CMS Phase-1 Pixel Detector Upgrade”, *JINST* **16** (2021), no. 02, P02027, doi:10.1088/1748-0221/16/02/P02027, arXiv:2012.14304.
- [61] “Silicon Strips”,. <https://cms.cern/detector/identifying-tracks/silicon-strips>.
- [62] CMS Collaboration, “The CMS electromagnetic calorimeter project: Technical Design Report”. Technical design report. CMS. CERN, Geneva, 1997.
- [63] P. Lecoq et al., “Lead tungstate (PbWO_4) scintillators for LHC EM-calorimetry”, *Nucl. Instrum. Methods Phys. Res., A* **365** (1995) 291–298, doi:10.1016/0168-9002(95)00589-7.
- [64] CMS Collaboration, “The CMS hadron calorimeter project: Technical Design Report”. Technical design report. CMS. CERN, Geneva, 1997.
- [65] J. Manset al., “CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter”, technical report, 2012.
- [66] CMS Collaboration, J. G. Layter, “The CMS muon project: Technical Design Report”. Technical design report. CMS. CERN, Geneva, 1997.
- [67] R. Santonico and R. Cardarelli, “Development of resistive plate counters”, *Nuclear Instruments and Methods in Physics Research* **187** (1981), no. 2, 377–380, doi:10.1016/0029-554X(81)90363-3.
- [68] “Resistive Plate Chambers”,. <https://cms.cern/detector/detecting-muons/resistive-plate-chambers>.
- [69] CMS Collaboration, “CMS Technical Design Report for the Level-1 Trigger Upgrade”, technical report, 2013.

- [70] CMS Trigger, Data Acquisition Group Collaboration, “The CMS high level trigger”, *Eur. Phys. J. C* **46** (2006) 605–667, doi:10.1140/epjc/s2006-02495-8, arXiv:hep-ex/0512077.
- [71] E. Bartz et al., “The plt: A luminosity monitor for cms based on single-crystal diamond pixel sensors”, *Nuclear Physics B - Proceedings Supplements* **197** (2009), no. 1, 171–174, doi:10.1016/j.nuclphysbps.2009.10.060. 11th Topical Seminar on Innovative Particle and Radiation Detectors (IPRD08).
- [72] A. Bell et al., “Fast beam conditions monitor BCM1F for the CMS experiment”, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **614** (mar, 2010) 433–438, doi:10.1016/j.nima.2009.12.056.
- [73] S. van der Meer, “Calibration of the effective beam height in the ISR”, technical report, CERN, Geneva, 1968.
- [74] CMS Collaboration, “Particle-flow reconstruction and global event description with the CMS detector”, *JINST* **12** (2017), no. 10, P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:1706.04965.
- [75] W. Adam, B. Mangano, T. Speer, and T. Todorov, “Track Reconstruction in the CMS tracker”, technical report, CERN, Geneva, 2006.
- [76] CMS Collaboration, “Description and performance of track and primary-vertex reconstruction with the CMS tracker”, *JINST* **9** (2014), no. 10, P10009, doi:10.1088/1748-0221/9/10/P10009, arXiv:1405.6569.
- [77] CMS Collaboration, “Muon tracking performance in the CMS Run-2 Legacy data using the tag-and-probe technique”, CMS Detector Performance Summary CERN-CMS-DP-2020-035, 2020.
- [78] “CMS Tracking POG Performance Plots For 2017 with PhaseI pixel detector”, <https://twiki.cern.ch/twiki/bin/view/CMSPublic/TrackingPOGPerformance2017MC>.
- [79] CMS Collaboration, “Simulation of the Silicon Strip Tracker pre-amplifier in early 2016 data”, CMS Detector Performance Summary CMS-DP-2020-045, 2020.
- [80] K. Rose, “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems”, *Proceedings of the IEEE* **86** (1998), no. 11, 2210–2239, doi:10.1109/5.726788.
- [81] R. Frühwirth, W. Waltenberger, and P. Vanlaer, “Adaptive Vertex Fitting”, technical report, CERN, Geneva, 2007.
- [82] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov, “Reconstruction of electrons with the gaussian-sum filter in the CMS tracker at the LHC”, *Journal of Physics G: Nuclear and Particle Physics* **31** (jul, 2005) N9–N20, doi:10.1088/0954-3899/31/9/n01.

- [83] M. Cacciari, G. P. Salam, and G. Soyez, “The anti- k_t jet clustering algorithm”, *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.
- [84] CMS Collaboration, “Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV”, *JINST* **12** (2017), no. 02, P02014, doi:10.1088/1748-0221/12/02/P02014, arXiv:1607.03663.
- [85] CMS Collaboration, “Mitigation of anomalous missing transverse momentum measurements in data collected by CMS at $\sqrt{s} = 13$ TeV during the LHC Run 2”, CMS Detector Performance Summary CMS-DP-2020-18, 2020.
- [86] “Primary vertex sorting in CMSSW”.
<https://github.com/cms-sw/cmssw/blob/master/CommonTools/RecoAlgos/src/PrimaryVertexSorting.cc>.
- [87] CMS Collaboration, “Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC”, *JINST* **16** (2021), no. 05, P05014, doi:10.1088/1748-0221/16/05/P05014, arXiv:2012.06888.
- [88] CMS Collaboration, “Jet algorithms performance in 13 TeV data”, CMS Physics Analysis Summary CMS-PAS-JME-16-003, CERN, Geneva, 2017.
- [89] CMS Collaboration, “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”, *JINST* **13** (2018), no. 05, P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.
- [90] CMS Collaboration, “Identification of b quark jets at the CMS Experiment in the LHC Run 2”, CMS Physics Analysis Summary CMS-PAS-BTV-15-001, CERN, Geneva, 2016.
- [91] E. Bols et al., “Jet Flavour Classification Using DeepJet”, *JINST* **15** (2020), no. 12, P12012, doi:10.1088/1748-0221/15/12/P12012, arXiv:2008.10519.
- [92] CMS Collaboration, “Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13TeV with Phase 1 CMS detector”, CMS Detector Performance Summary CMS-DP-2018-058, 2018.
- [93] C. Bierlich et al., “A comprehensive guide to the physics and usage of PYTHIA 8.3”, doi:10.21468/SciPostPhysCodeb.8, arXiv:2203.11601.
- [94] CMS Collaboration, “Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements”, *Eur. Phys. J. C* **80** (2020), no. 1, 4, doi:10.1140/epjc/s10052-019-7499-4, arXiv:1903.12179.
- [95] GEANT4 Collaboration, “GEANT4—a simulation toolkit”, *Nucl. Instrum. Meth. A* **506** (2003) 250–303, doi:10.1016/S0168-9002(03)01368-8.
- [96] CMS Collaboration, “Pileup mitigation at CMS in 13 TeV data”, *JINST* **15** (2020), no. 09, P09018, doi:10.1088/1748-0221/15/09/P09018, arXiv:2003.00503.

- [97] CMS Collaboration, “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”, *JINST* **13** (2018), no. 05, P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.
- [98] CMS Collaboration, “Performance of b tagging algorithms in proton-proton collisions at 13 TeV with Phase 1 CMS detector”, CMS Detector Performance Summary CMS-DP-2018-033, 2018.
- [99] L. Carloni and T. Sjostrand, “Visible Effects of Invisible Hidden Valley Radiation”, *JHEP* **09** (2010) 105, doi:10.1007/JHEP09(2010)105, arXiv:1006.2911.
- [100] L. Carloni, J. Rathsman, and T. Sjostrand, “Discerning Secluded Sector gauge structures”, *JHEP* **04** (2011) 091, doi:10.1007/JHEP04(2011)091, arXiv:1102.3795.
- [101] J. Alwall et al., “MadGraph 5 : Going Beyond”, *JHEP* **06** (2011) 128, doi:10.1007/JHEP06(2011)128, arXiv:1106.0522.
- [102] J. Thaler and K. Van Tilburg, “Identifying Boosted Objects with N-subjettiness”, *JHEP* **03** (2011) 015, doi:10.1007/JHEP03(2011)015, arXiv:1011.2268.
- [103] Y. Wang et al., “Dynamic graph cnn for learning on point clouds”, *ACM Trans. Graph.* **38** (oct, 2019) doi:10.1145/3326362.
- [104] H. Qu and L. Gouskos, “ParticleNet: Jet Tagging via Particle Clouds”, *Phys. Rev. D* **101** (2020), no. 5, 056019, doi:10.1103/PhysRevD.101.056019, arXiv:1902.08570.
- [105] CMS Collaboration, “Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques”, CMS Detector Performance Summary CMS-DP-2020-002, 2020.
- [106] CMS Collaboration, “Study of quark and gluon jet substructure in Z+jet and dijet events from pp collisions”, *JHEP* **01** (2022) 188, doi:10.1007/JHEP01(2022)188, arXiv:2109.03340.
- [107] G. D’Agostini, “A multidimensional unfolding method based on Bayes’ theorem”, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **362** (1995), no. 2, 487–498, doi:10.1016/0168-9002(95)00274-X.
- [108] C. J. Clopper and E. S. Pearson, “The use of confidence or fiducial limits illustrated in the case of the binomial”, *Biometrika* **26** (1934), no. 4, 404–413.
- [109] V. V. Patil and H. V. Kulkarni, “Comparison of confidence intervals for the poisson mean: Some new aspects”, *Revstat - Statistical Journal* **10** (06, 2012) 211–227.
- [110] NNPDF Collaboration, “The path to proton structure at 1% accuracy”, *Eur. Phys. J. C* **82** (2022), no. 5, 428, doi:10.1140/epjc/s10052-022-10328-7, arXiv:2109.02653.

- [111] TheATLAS, TheCMS, TheLHCHiggsCombinationGroup Collaboration, “Procedure for the LHC Higgs boson search combination in Summer 2011”, technical report, CERN, Geneva, 2011.
- [112] T. Junk, “Confidence level computation for combining searches with small statistics”, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **434** (1999), no. 2, 435–443, doi:10.1016/S0168-9002(99)00498-2.
- [113] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, “Asymptotic formulae for likelihood-based tests of new physics”, *Eur. Phys. J. C* **71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727. [Erratum: Eur.Phys.J.C 73, 2501 (2013)].
- [114] “Higgs Analysis Combined Limit”,
<https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/>.
- [115] CMS Collaboration, “Search for strongly interacting massive particles generating trackless jets in proton–proton collisions at $\sqrt{s} = 13$ TeV”, *Eur. Phys. J. C* **82** (2022), no. 3, 213, doi:10.1140/epjc/s10052-022-10095-5, arXiv:2105.09178.
- [116] CMS Collaboration, “Search for Long-Lived Particles Decaying in the CMS End Cap Muon Detectors in Proton-Proton Collisions at $\sqrt{s}=13$ TeV”, *Phys. Rev. Lett.* **127** (2021), no. 26, 261804, doi:10.1103/PhysRevLett.127.261804, arXiv:2107.04838.
- [117] CMS Collaboration, “Search for resonant production of strongly coupled dark matter in proton-proton collisions at 13 TeV”, *JHEP* **06** (2022) 156, doi:10.1007/JHEP06(2022)156, arXiv:2112.11125.
- [118] ATLAS Collaboration, “Search for non-resonant production of semi-visible jets using Run~2 data in ATLAS”, arXiv:2305.18037.