Based on the following publications:
doi: 10.1103/PhysRevD.97.072006; 10.1007/JHEP03(2017)162; cds.cern.ch/record/2256875 (CERN public document);
cds.cern.ch/record/2668755 (submitted to EPJC) .

Ph.D. thesis summary: A Novel Multidimensional Search for Diboson Resonances in the Boosted Dijet Final State and Encoding Jet Substructure with a Deep Neural Network

Thea Klaeboe Aarrestad
University of Zurich

Abstract

In this thesis I seek answers to some of the most fundamental open questions in particle physics: why is the Higgs boson mass we measured at 125 GeV so much lighter than what we would have predicted when taking into account loop corrections from energy scales up to the Planck scale? And why is gravity so weak compared to the other known forces arising from electroweak and QCD interactions? These are both aspects of the same problem, referred to as the "hierarchy problem". I look for solutions to these questions in the form of extensions of the Standard Model of particle physics, by comparing the physics of high-energy collisions with such alternate models. Specifically, I look for new particles predicted by Composite Higgs models, offering an answer to the first question, and warped extra dimensions theories, attempting to explain both the former and the latter. This thesis presents three different searches for new heavy resonances that decay to a pair of vector bosons in the all-hadronic final state with the CMS detector at the Large Hadron Collider (LHC). The diboson final states under consideration are challenging to resolve due to the bosons being highly energetic ("boosted"), resulting in the two quarks from the decay being collimated and merging into a single jet. This leads to a dijet final state topology where each jet displays substructure in its spatial distribution of energy. I present the results of three different searches I have done, with a number of improvements in techniques that have benefited my analyses, as well as dozens of other measurements done at CMS, ending with the development of a novel multi-dimensional search method yielding a significant improvement in analysis sensitivity and introducing a new way of doing model-independent searches at LHC.

1 FIRST RESULTS WITH 13 TEV DATA

In my first few months a PhD student, I helped finalize the first Higgs-boson tagger at CMS. As a masters student, I had demonstrated the very first dedicated algorithm for the identification of a Higgs boson decaying into two b-quarks, where the Higgs boson was sufficiently energetic so that the two b-quarks were very close together and merged into one single jet (a cluster of particles stemming from the decay and hadronisation of a quark) [1]. After the discovery of the Higgs boson in 2012, it became an established standard model (SM) particle, and we could now use it as a probe for non-SM massive particles that could decay into Higgs bosons. These Higgs bosons would be of much higher transverse momentum (p_{T}) than expected by the SM, due to the high mass of the resonance, and therefore more likely to be boosted. After I established and documented the approach, it was adopted by CMS, and I helped finalize the method by providing crucial studies on how to de-correlate tagging performance from jet p_{T} and pseudo rapidity, as well as performing dedicated trainings against different backgrounds [2]. The Higgs(bb) tagger resulted in several CMS publications, as cited in my publication list.

Through my early work on Higgs-jet tagging, I became interested in jets and jet-substructure algorithms. In June 2015, a day before the LHC was to collide protons with a center-of-mass energy of 13 TeV for the very first time, a paper by the ATLAS collaboration was published regarding the search for heavy resonances decaying to vector bosons in the all-hadronic state, based on the full data set collected at a center-of-mass energy of 8 TeV [3]. The analysis documented a 3.4σ excess consistent with a heavy resonance decaying to WZ with a mass of around 2 TeV. Several interesting extensions to the Standard Model, offering explanations for some of the important remaining questions in physics, predict such heavy resonances decaying to vector bosons, making the observation an extremely exciting one. Did physics beyond the Standard Model exist and did we have the correct alternative models at hand to describe it? New data was being collected at twice the collision energy at the time of publication, and the increase in partonic luminosities would lead to the same expected signal sensitivity at 13 TeV with a data set only a fraction of the size ($\sim 1/7$) of the full 8 TeV data set. Excitement over a possible new particle and the feasibility to confirm it within a short time of 13 TeV data taking led me to my first analysis and my thesis topic: Searches for heavy resonances decaying to dibosons in the all-hadronic final state.

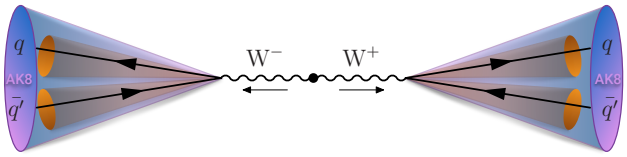

Figure 1. If a heavy (> 1 TeV) resonance decays into vector bosons, the transverse momentum of each boson will be large and its decay products are merged into one single large jet.

When a resonance X with a mass above 1 TeV decays into a vector-boson pair, the bosons have a very high energy and Lorentz boost and are referred to as boosted. The decay products of a hadronically decaying boosted vector boson will therefore not appear as back-to-back in the lab frame, but rather be collimated. This results in a final state with two high-p_{T}, large-radius jets that fully contain the two quarks coming from the vector boson decay. This is illustrated in Fig. 1. The two jets are each expected to have a mass around the W or Z boson mass, and some intrinsic substructure stemming from their two-pronged decay. The invariant mass of the dijet system, m_{jj}, should be roughly equal to the resonance mass m_X. This dijet system is the final state under scrutiny and the dijet invariant mass is the parameter of interest. The main background for such an analysis is QCD multijet events. Quark/gluon jets can obtain a high mass due to diffuse radiation, and QCD processes have such a large cross section that the number of QCD jets with a mass compatible with the W mass can be large. In order to discriminate between the two, we take advantage of three properties. First, the true mass of signal and background jets should be very different. Second, signal jets should appear two-prong like, as opposed to quark/gluon jets. And third, the dijet invariant mass for the signal process should peak around the resonance mass while the QCD spectrum is predicted to be a smoothly falling distribution. The strategy therefore consists of performing a smoothness test on m_{jj} of the observed data; a so-called "bump-hunt", by assuming that the signal will appear as a bump on top of a smooth distribution. The benefit of such a method is that there is no need for a simulation of the background, a necessary advantage as Monte Carlo with $\sqrt{s} = 13$ TeV and had never before been validated.

I was the sole analyst running a full analysis of the first 13 TeV data collected in 2015, and managed to bring the search to a published results [4] within 5 months of data taking, as one of the most anticipated results with the new dataset due to the previous excess. It became one of the first 13 TeV results ever published, and one of the two first "boosted" searches to be performed at 13 TeV at CMS. The search set the most stringent limits on the signal hypotheses under scrutiny to date.

2 DEVELOPING A PILEUP-RESISTANT AND PERTURBATIVELY-ROBUST VECTOR-BOSON TAGGER

After publishing the analysis of the full 2015 data set, I looked towards 2016. In this year, CMS would collect 10 times more data, but at the cost of a higher mean number of interactions per proton bunch crossing, causing additional interaction vertices per event (pileup) whose effect would need to be mitigated in order to maintain analysis sensitivity. I had become very much involved with jet algorithms in CMS and wanted to explore better algorithms intended to distinguish hadronically decaying vector bosons from jets originating from quark and gluons; and one promising direction was exploring a new pileup-removal algorithm called Pileup per particle identification (PUPPI) [5]. PUPPI had proven itself far superior to the current pileup-removal algorithm in terms of jet observables, like the jet mass, for large radius jets. Another promising new algorithm to explore was Soft Drop [6], intending to improve the vector-boson mass resolution and better discriminate it against quark/gluon jets by removing soft radiation from the jet. Algorithms intended to improve the jet mass resolution by removing radiation from the jet are referred to as "jet grooming algorithms", and the benefit of softdrop was that it had certain favorable theoretical qualities: the softdrop algorithm, in addition to removing sensitivity to the soft divergences of QCD as all grooming algorithms do, eliminates all correlated soft emissions in the jet, leading to no non-global logarithmic terms (NGLs) in the jet mass [7]. NGLs arise from configurations where, for instance, a soft gluon is radiated into the jet cone, as illustrated in Figure 3. The benefit of being NGL-free, is that one can calculate

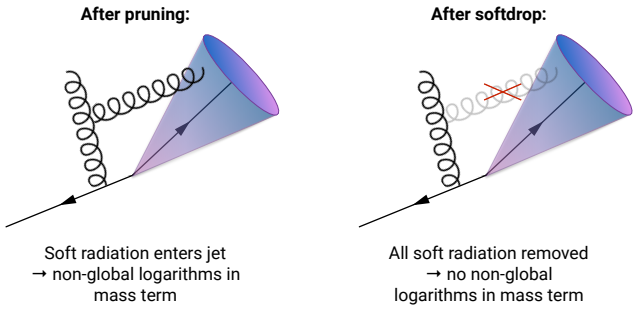

Figure 2. The pruning algorithm does not remove all soft emission and therefore has non-global logarithmic terms in the jet mass. The softdrop algorithm removes all soft emissions and is free of non-global logarithms.

the softdrop jet mass to a significantly higher precision than what is possible for other grooming algorithms or for the plain jet mass (NGLs are the main reason a full resummation of the plain jet mass beyond NLL accuracy does not exist). There were therefore theoretically well-motivated reasons for wanting the baseline CMS

vector-boson tagger to be softdrop-based.

I therefore optimized, commissioned and validated a new vector-boson tagging algorithm, with the goal of a more robust and theoretically well motivated tagger for all future analyses in CMS using vector-boson tagging [8]. This included the derivation of dedicated jet-mass corrections for this new type of jets (jets after PUPPI and softdrop have been applied) to account for detector effects that can affect the observed jet mass, and depend on the jet p_{T} and pseudo rapidity η. The W-jet mass mean as a function of jet p_{T} and η before (left) and after (right) corrections, is shown in Fig. 3. I published

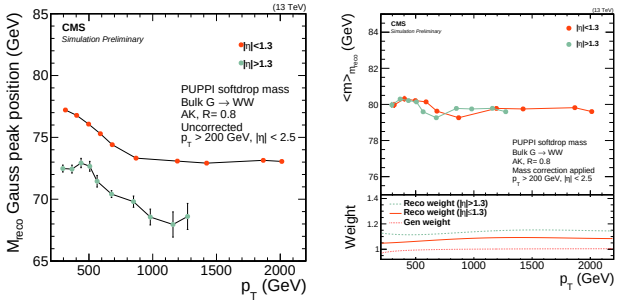

Figure 3. The Gaussian mean of the fitted jet softdrop mass for W-jets before (left) and after (right) dedicated jet mass corrections have been applied as a function of jet transverse momentum.

an analysis of the data collected in 2016 as the first analysis to take advantage of this new vector-boson tagging algorithm that I developed [9, 10], an algorithm (with corresponding jet mass corrections) that afterwards became the default vector-boson tagging algorithm in CMS. More than 20 subsequent CMS analysis take advantage of the tagger and jet mass corrections (these are listed in my attached publication list). In parallel, I developed efficiency scale factors to account for any mismodeling in simulation that affect the tagging efficiency, both for the newly developed tagger, and for the previous default tagger. These are the CMS recommended scale factors for any analysis using vector-boson tagging, and I evaluated these for data collected throughout years 2015, 2016, and 2017 (full analyses in their own right requiring months of dedicated work). These scale factors are crucial for a correct signal yield estimation. The published analysis, in addition to introducing a novel tagging algorithm, also included a brand new search: a search for excited quarks decaying to a quark and a vector boson. This was an extremely interesting search attempting to probe whether quarks truly are fundamental particles or not, by checking whether a collision could excite a quark to a higher energy state, leaving its constituents the same, but with a higher mass. It was the very first time such a search in this channel was performed at 13 TeV [9]. No significant deviations from the Standard Model prediction were observed in either of the searches, leading us to exclude non-SM resonances decaying to VV (where V=W/Z) or qV up to very high masses, ~ 3 TeV for VV signals and ~ 5 TeV for qV signals.

3 A NOVEL MULTI-DIMENSIONAL SIGNAL EXTRACTION FRAMEWORK

No excesses had been observed in analyses of the 2015 and 2016 data sets described above. For my next project I therefore wanted to look beyond standard signal hypotheses by establishing a method to efficiently probe a large range of different signal hypotheses. There could still be signal present in our data, but it might look slightly different than what we had assumed up until now. For instance, new signals could exist where the observed jet mass is slightly different than that of a W/Z boson mass. Further, these could be 4-pronged objects rather than 2-prong, which would cause the excess to vary in size depending on the analysis-specific, vector-boson tagger in use. This is illustrated in Figure 4. In

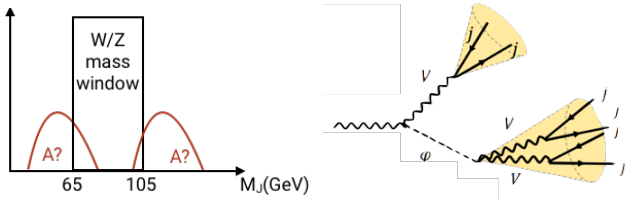

Figure 4. New signals that peak at a mass slightly different than the W/Z mass (left) and have a slightly different substructure (right) could be present in our data.

order to efficiently search for any signal peaking in the jet mass, we decided to build a generic framework that would allow searching for peaks anywhere in the jet mass and dijet invariant mass spectrum. Rather than selecting jets with a jet mass compatible with the W/Z boson mass and searching for resonances peaking in the dijet invariant mass, we would attempt to look for resonances peaking anywhere in the hyperplane formed by the mass of each jet and their dijet invariant mass, scanning the full mass spectrum in a single analysis, as illustrated in Fig. 5. We would first demonstrate the new method in the context of the diboson all-hadronic search, which would allow for a straight forward comparison of the obtained results. The benefits of doing a multi-dimensional fit is that we can search for resonances decaying to VV (V=W/Z), VH (H=Higgs), HH, VX, VH, XX, or XY, where X and Y are new hypothetical bosons, in the same analysis. Additionally, a jet-mass selection is no longer needed as we fit the full jet mass line-shape to extract the signal. This effectively increases the signal statistics since a large fraction of the W/Z signal falls outside the mass window (20%). Fitting the groomed-jet mass and resonance mass together also allows for the addition of nuisance parameters that simultaneously affect both in order to fully account for the correlation between the variables. Finally, we would model the background

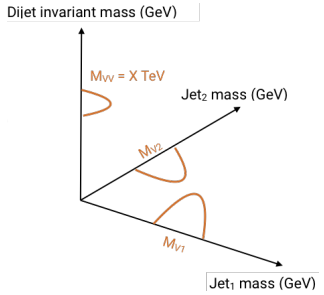

Figure 5. With the new multi-dimensional fit method, the signal is extracted from the m_{jet1}-m_{jet2}-m_{jj} hyperplane, where it peaks in all three dimensions.

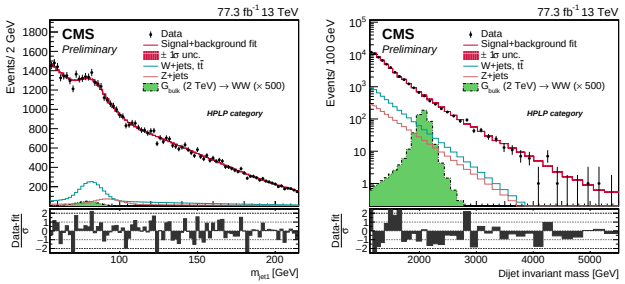

Figure 6. Comparison between the fitted result and data distributions for the jet mass of one of the jets (left) and the dijet invariant mass (right). An example of a signal distribution is overlaid, using an arbitrary normalisation. Two beautiful peaks from Z(qq)+jets and W(qq)+jets are measured for the first time in diboson analyses.

starting from simulation, rather than from a dijet fit to data. This allows the background shape to assume non-smooth distributions, and could allow the search to probe lower dijet masses. Replacing the parametric fit by a simulation-based model would also reduce the fit sensitivity to background fluctuations in the extreme tails of the dijet invariant mass spectrum. However, this technique would be challenging to implement, and require new techniques, such as Gaussian kernels with a mean and width obtained through forward-folding rather than single points in order to model the backgrounds, to be incorporated. We decided to re-analyze the 2016 data set in order to directly compare the sensitivity between the previous method and the multi-dimensional technique, as well as for the first time analyze the data collected in 2017.

For this search, I additionally optimized and commissioned a new vector-boson tagging algorithm intended to remove correlations between the tagger and jet p_{T}/mass (crucial for an analysis covering a large jet mass and dijet invariant mass range). It also provided a significantly higher signal sensitivity for our phase space.

Fig. 6 shows the final fitted result and the data distributions projected onto two of the three dimensions of interest: the jet mass of one of the jets and the dijet invariant mass. Two beautiful peaks from the Standard Model Z(qq)+jets and W(qq)+jets background are observed. This is the first time these SM backgrounds have ever been measured in a diboson analysis. Their extracted cross sections are found to be compatible with the SM expectation [11], a measurement made possible due to the nature of the fit and the optimized vector-boson tagging algorithm. The two peaks also allow us to constrain uncertainties affecting the signal yield, leading to a better analysis sensitivity. The obtained expected upper limits using the multi-dimensional fit method introduced here, can be compared to those obtained in the previous search described above using the same data set, Ref. [10], in order to estimate whether there is a sensitivity gain in using the new method. Figure 7 shows the expected limits based on analyses of the data collected in 2016, either using the fit method presented here, or using the previous one-dimensional method. We obtain a 20-30% improvement in sensitivity when using the multi-dimensional fit method, and about a 35-40% improvement when combining the two data sets with respect to the individual results. These results can also

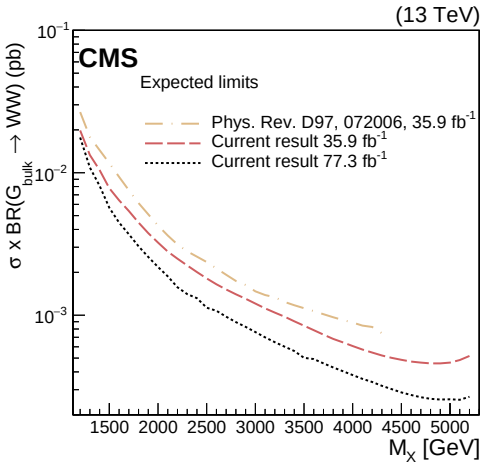

Figure 7. Expected limits for a Bulk $G \to WW$ signal obtained using the multi-dimensional fit method presented here (pink line), compared to the result obtained using previous methods (beige line) [10]. The final limit obtained when combining data collected in 2016 and 2017 is also shown (black dotted line).

be compared to those obtained by the ATLAS collaboration in a similar search presented in Ref. [12] analyzing the same data set, where we also observe an improvement in sensitivity using the method presented here, by up to 35% for comparable signal models.

4 ENCODING JET SUBSTRUCTURE IN A DEEP NEURAL NETWORK

In an effort to go beyond current techniques to the next level, I have also developed a novel vector-boson tagger that uses all of the Lorentz vectors of the particles in the jet, and assembles the information using a deep neural

network (DNN) in order to distinguish W and Z jets from quark and gluon jets [13]. This algorithm yields twice the signal efficiency for a given mis-tagging rate with respect to current methods, and could significantly improve the sensitivity of future searches. Due to its special architecture, intended to encode jet substructure algorithms, it could also be trained as a signal-independent tagger (separating quark/gluon jets from "any" potential signal jet). Together with the multi-dimensional framework, this could be used to design a truly model-independent search.

5 SUMMARY

As a Ph.D. student I have become a key contributor to the study of jet algorithms in the high-energy domain and provided essential studies for the improvement of jet tagging algorithms used in CMS. My work has allowed for the publication of dozens CMS analyses and has driven the field forward with original contributions. I have introduced a novel multi-dimensional search framework that, in addition to improving the search sensitivity by up to 35%, allows us to easily incorporate alternate signal hypotheses, a key ingredient for future searches as no deviation has remained in the present data in searches for new resonances decaying into vector bosons. We are starting an era at LHC where improvements from adding more data are not where large sensitivity gains can be made anymore, but where real improvements in physics reach is obtained through improved methods. This new fit method, together with my work on improving vector-boson tagging, therefore makes this thesis highly relevant and important for future searches. My achievements include: the first demonstration of a Higgs(bb) tagging algorithm in CMS, bringing a highly anticipated result to publication as one of the first 13 TeV analyses, developing a new boson-tagging algorithm that is now default in CMS, providing kinematic-dependent corrections for vector-boson tagging between simulation and data, developing a brand new signal extraction method with a significantly improved sensitivity with respect to current methods, and demonstrating the first measurement of the SM V(qq)+jets background in diboson searches. I have also originated a deep neural network W/Z tagging algorithm, achieving a 50% higher signal efficiency at a given mistag rate compared to current methods. In addition to my own work, I have had the pleasure and responsibility of supervising a CERN Summer Student on developing a tool to discriminate between transversally and longitudinally polarized vector bosons, and a bachelor student in the development of a b-tagging algorithm based only on number of hits in the pixel detector (see CV). Finally, I have been responsible for all pixel gain calibrations in 2017.

REFERENCES

[1] T. Aarrestad, ""A neural network based dedicated boosted Higgs b-tagging algorithm in CMS"". `thaarres.web.cern.ch/thaarres/MasterThesis_TAarrestad.pdf`, Oct 2014.

[2] CMS Collaboration, "Identification of double-b quark jets in boosted event topologies", Technical Report CMS-PAS-BTV-15-002, CERN, Geneva, 2016.

[3] The ATLAS collaboration, "Search for high-mass diboson resonances with boson-tagged jets in proton-proton collisions at s=8 TeV with the ATLAS detector", *JHEP* **2015** (Dec, 2015) .

[4] The CMS collaboration, "Search for massive resonances decaying into WW, WZ or ZZ bosons in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **2017** (Mar, 2017) .

[5] D. e. a. Bertolini, "Pileup Per Particle Identification", *JHEP* **10** (2014) , `arXiv:1407.6013`.

[6] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft Drop", *JHEP* **05** (2014) 146, , `arXiv:1402.2657`.

[7] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, "Towards an understanding of jet substructure", *JHEP* **09** (2013) 029, , `arXiv:1307.0007`.

[8] CMS Collaboration, "Jet algorithms performance in 13 TeV data", Technical Report CMS-PAS-JME-16-003, CERN, Geneva, 2017.

[9] CMS Collaboration, "Search for massive resonances decaying into WW, WZ, ZZ, qW and qZ in the dijet final state at $\sqrt{s} = 13$ TeV using 2016 data", Technical Report CMS-PAS-B2G-16-021, CERN, Geneva, 2016.

[10] CMS Collaboration, "Search for massive resonances decaying into WW, WZ, ZZ, qW, and qZ with dijet final states at $\sqrt{s} = 13$ TeV", *Phys. Rev. D.* **97** (2018) , `arXiv:1708.05379`.

[11] CMS Collaboration, "Search for heavy resonances in the all-hadronic vector-boson pair final state with a multi-dimensional fit", Technical Report CMS-PAS-B2G-18-002, CERN, Geneva, 2019.

[12] ATLAS Collaboration, "Search for diboson resonances in hadronic final states in 79.8 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", ATLAS Conference Note ATLAS-CONF-2018-016, 2018.

[13] T. Aarrestad, ""LoLa: Lorentz Invariance Based DNN for heavy-resonance tagging"". `https://www.physik.uzh.ch/dam/jcr:1463636d-1b8b-45fc-960c-b2155f9ddd93/poster_5.pdf`, Nov 2017.