Spectral clustering for jets

September 18, 2020

1 Introduction

The investigations presented here straddle a number of topics, machine learning, high energy experimental physics and high energy phenomenology. The focus is tools that form and classify jets in challenging topologies. Such topologies represent hard to access areas of many predictions, such as those of the Two Higgs Doublet Model (2HDM). Tools that were able to utilise more of the information held in the data generated by high energy experiments, such as those at the LHC, might be able to exclude or confirm the present of multiple Higgs like particles.

A particle jet is a particular clustering of observed particles in a detector. The clustering is designed to reflect the origin of these particles and it provides vital structure in the processing of the overall event. The default choice for jet clustering tends to be one of three algorithms; the anti-kt algorithm [9], the Cambridge-Aachen algorithm [24] and the kt algorithm [13]. They have been the default choice for some time because they have a number of desirable properties. They are infrared safe, excellent implementations of them are available (see `FastJet` [8]) and they are flexible enough to capture many signals with minimal parameter change. These algorithms are recursive and agglomerative. A recursive algorithm is well suited to clustering objects when the number of groups is not known at outset. Agglomerative algorithms are easier to design in a manner that is infrared safe, as they can recombine soft and collinear emissions in early steps.

Once the jets have been formed it then remains to identify the decayed tag particle they represent. The tag particle is the particle that leaves the hard interaction, or is immediately descendant of the proton beam, that decays to form the shower. There are a number of factors effecting the difficulties of identifying this particle. One of these would be how well the shower has been isolated by the jet. Sometimes two showers overlap strongly, so one jet in fact contains the combination of two showers, identifying both originating particles together is especially challenging. This is often the case when a very high energy particle decays into lighter particles, and the descendants have high kinetic energy, which sends them in a boosted configuration as a shallow angle to the beam line.

A signal of particular interest is found in the extended Higgs sector. Since the discovery of the Higgs Boson in 2012, it's couplings have been seen to be in agreement with the Standard Model (SM), however additional Higgs particles remain possible. One of the simplest extensions to the Higgs sector is the two Higgs doublet model (2HDM) [7]. The second doublet of the 2HDM allows a further 5 particles; two CP even (h and H, with, conventionally, $m_h < m_H$), one CP odd (A) and a pair of charged ($H^{\pm}$) Higgs bosons.

This is relevant to jet physics because the additional Higgs particles most commonly to decay to b-quarks which shower to form jets. These jets may have a highly boosted configuration due to the mass difference between the b-quark and the heavy Higgs. As such the 2HDM is a prime example of the types of signals that might benefit from advances in jet formation and classification.

2 Problem statement

This section will offer an overview of the key physics for the benefit of readers from outside of experimental particle physics.

Particle colliders such as the LHC grant insight into physics by creating heavy short lived particles that do not occur without a high energy collision and only persist for a fraction of a second before decaying into lighter particles. There is then a sequence of decays, known as a shower, before something stable and long live enough to reach the detectors is created. From the remnants of the decay that reach the detector we seek to characterise the heavy particle created in the collision. This is the raw data available to understand the event with. In theory this is quite difficult and in practice it is also not easy.

The data pipeline can be seen as a whole in [4, 20]. Evidence of the shower is gathered from multiple concentric detectors about the collision point, the primary components being the silicon tracker for charged particles and the calorimeter for neutral particles. These can be seen in figure 1. This data is essentially a series of energy deposits

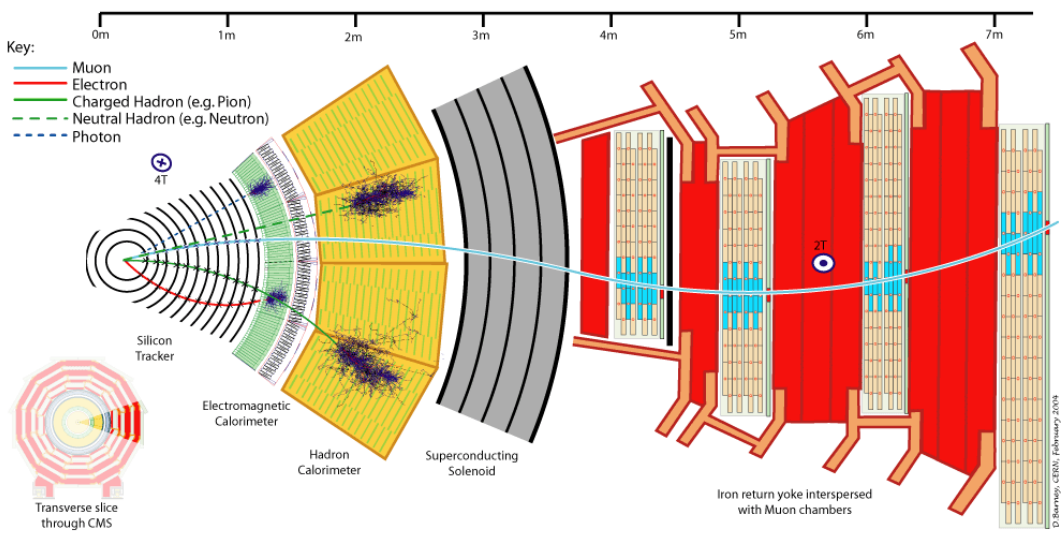

Figure 1: The CMS detector has a concentric structure with each layer sensitive to a subset of possible particles. [4]

associated with three dimensional coordinates, there is no time stamp available because the interaction occurs faster than the electronics can read out. At this point the Particle Flow algorithm is used to reconstruct four vector tracks from the readings [5].

There are often a great many particles in the shower by the time it reaches the detector, attempting to cluster them by common origin, into jets, is a key step in the reconstitution process. It is particularly important for the identification and characterisation of quarks, which will decay before the detector and have readily identified properties. The majority of jets, however, come from gluons which radiate from the beam. This clustering process will be considered in depth in later sections.

A good jet clustering algorithm needs to be infra-red and collinear safe [18, p. 10] and it seeks to group the descendants of the quarks into exclusive clusters. Thus the clustered jets can be used to estimate the kinematics of the quarks. Jet clustering via novel methods is investigated in section 4.

Trimming may be used to remove 'soft' radiation, that is products of beam radiation and lower energy interactions that occurred alongside the signal process. Before the final stage considered here (tagging the jet) various preprocessing techniques may be applied to the jet and the tracks clustered in it. Finally an algorithm will tag the jet, by estimating the identity of the particle that created it.

2.1 Phenomonology

In order to best develop tools to identify evidence of new physics it is important to establish what a new physics is likely to appear. There are many proposed, and as yet unobserved, ideas for new physics. As mentioned in the introduction the 2HDM is the model considered throughout this work.

In its most general form the 2HDM adds 6 real parameters and 4 complex parameters. These parameters can be further reduced by symmetry considerations. Of the remaining parameter space the values of interest are constrained by two factors; firstly the parameter choices must be such that it is reasonable that the particles created by the 2HDM have not already been observed, secondly the parameter choice need to be such that it is plausible we could differentiate this model from the null hypothesis (the SM) using detectors available to us in the near future.

Applying these constraints refines the form of the signal that we design searches for to one that is of greater interest. This problem is addressed in more depth in section 3.

2.2 Simulation

When developing tools, it is a great help to use data produced by Monte Carlo (MC) simulation. Although this may be an imperfect representation of real particle behaviour it is a great advantage to be able to relate an observation to the high energy interaction that created it. This MC truth information is very valuable in assessing the success of both jet clustering and jet formation techniques.

There are a variety of mature and well tested packages available for generating this data. Specifically `GEANT 4` [1] is used to simulate interactions between particles and the detector material. `POWHEG 2.0` [2] and `MadGraph5_aMC@NLO 2.2.2` [3] are used to generate signal events. `PYTHIA 8.2` [22] is used to simulate the background, parton showering and hadronizsation.

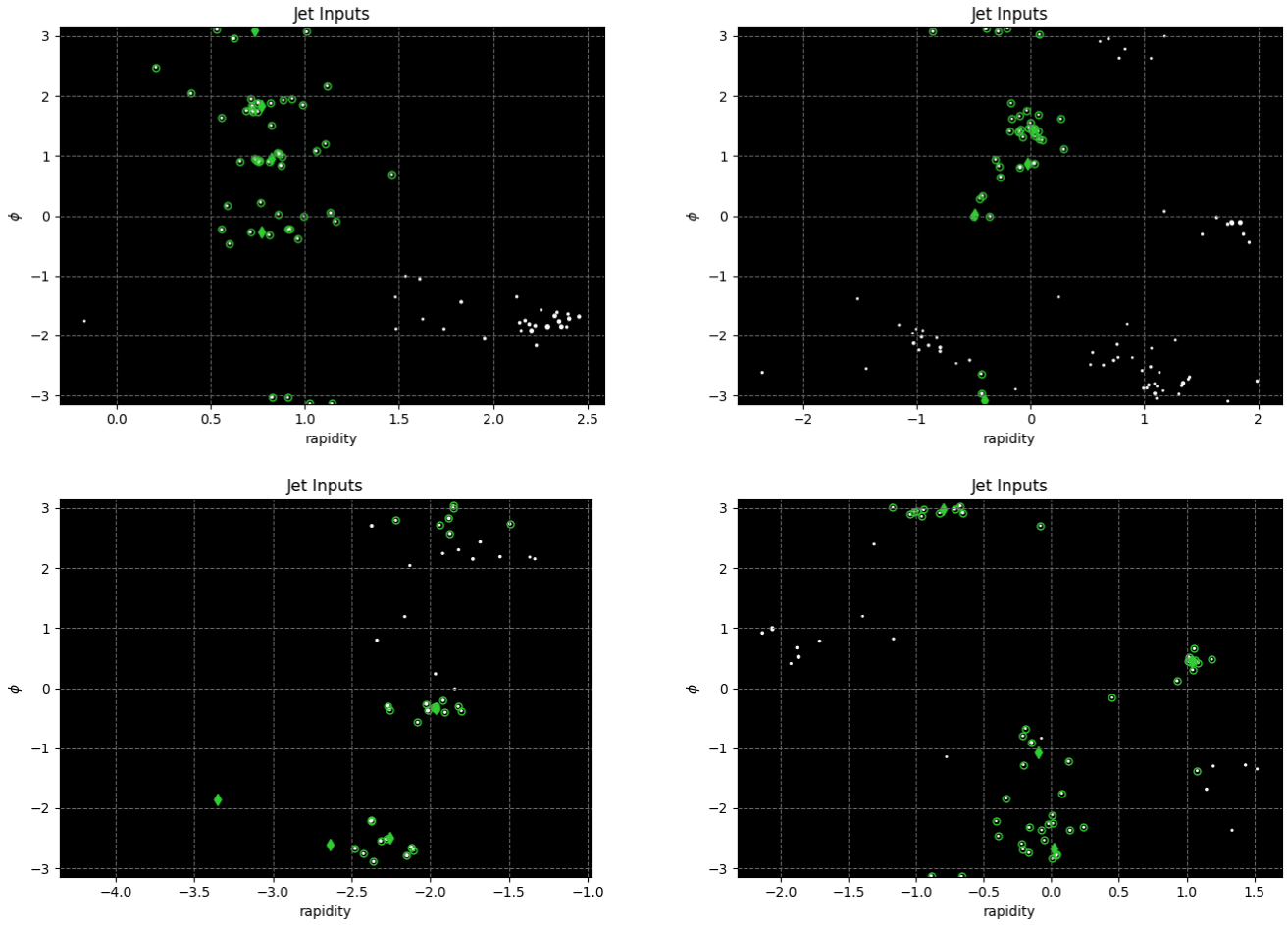

Figure 2: Some examples of events after cuts. Each white dot is a particle that passes the cuts and can be used for jet clustering. A green diamond indicates the location of the b-quark (not an input to clustering). A green circle indicates a descendant of the b-quark.

For the purposes of jet physics, the next step is to identify a particle in the MC truth as a tag particle. In this study this tag particle is a b-quark. The tag will decay before reaching the detector and jet physics aims to reconstruct it from its decay products. The stable decay products that exist in the final state are termed here as 'descendants'. If these descendants have sufficient energy to be detectable and exist at the right angle to meet the detector they are used as inputs for jet formation. These are known as track cuts, and they stand in for the sensitivity of the detector.

Mixed in with the descendants are a number of other particles that have come from objects that are not tags decaying. In this study many of these are gluons radiated from the beam. They form the background of the event and make jet formation and identification more difficult. Four examples of such events can be seen in fig 2.

These simulation tools are used throughout this work.

3 Two Higgs Doublet Model Parameters

This section summarises the work done in collaboration with Rachid Benbrik and Souad Semlali, for which my supervisor Stefano Moretti is also co-author [6]. The work extends existing scan of the 2HDM model, to locate parameter space of interest now, and after a planned luminosity and energy upgrade to the LHC.

Five new particles are offered in the 2HDM, two CP even (h and H, with, conventionally, $m_h < m_H$), one CP

odd (A) and a pair of charged ($H^\pm$) Higgs bosons. These come about from two complex scalar doublets $\Phi_{1,2}$ from $SU(2)_L$ with the most general gauge invariant renormalisable scalar potential of the 2HDM given by:

$$\begin{aligned} V(\Phi_1,\Phi_2) =& m_{11}^2\Phi_1^\dagger\Phi_1 + m_{22}^2\Phi_2^\dagger\Phi_2 - [m_{12}^2\Phi_1^\dagger\Phi_2 + \text{h.c.}] \\ &+\frac{\lambda_1}{2}(\Phi_1^\dagger\Phi_1)^2 + \frac{\lambda_2}{2}(\Phi_2^\dagger\Phi_2)^2 + \lambda_3(\Phi_1^\dagger\Phi_1)(\Phi_2^\dagger\Phi_2) \\ &+\lambda_4(\Phi_1^\dagger\Phi_2)(\Phi_2^\dagger\Phi_1) + \frac{1}{2}[\lambda_5\ (\Phi_1^\dagger\Phi_2)^2 + \ \text{h.c.}] \\ &+\left[(\lambda_6(\Phi_1^\dagger\Phi_1) + \lambda_7(\Phi_2^\dagger\Phi_2))\Phi_1^\dagger\Phi_2 + \text{h.c.}\right]. \end{aligned} \tag{1}$$

Following the hermiticity of the scalar potential, m_{11}^2, m_{22}^2 and $\lambda_{1,\dots 4}$ are real parameters whereas m_{12}^2, $\lambda_{5,6,7}$ can be complex. Assuming the CP-conserving version of the 2HDM, m_{12}^2, $\lambda_{5,6,7}$ and the VEVs of the fields Φ_i are real parameters. As a consequence of extending the discrete Z_2 symmetry to the Yukawa sector in order to avoid Flavour Changing Neutral Currents (FCNCs) at tree level, $\lambda_{6,7} = 0$, whereas the mass term m_{12}^2 breaks the symmetry in a soft way.

Amongst the many signals that these additional Higgs states could produce, of particular relevance are those involving their cascade decays, wherein a heavier Higgs state decays in a pair of lighter ones or else into a light Higgs state and a gauge boson. This is the case as the former process gives access to the shape of the Higgs potential of the enlarged Higgs sector while the latter channel is intimately related to the underlying gauge structure, which may well be larger than the SM one. Parameter areas of interest for the process $pp \to H \to ZA \to l^+l^-b\bar{b}$ and its mirror image $pp \to A \to ZH \to l^+l^-b\bar{b}$ can be established by eliminating areas ruled out by theory constraints, previous observations and experimental sensitivity requirements.

The pattern of Branching Ratios (BRs) of the two decays $A \to ZH$ and $H \to ZA$ was first discussed in Refs. [16] and [11] (albeit in a Supersymmetric version of the 2HDM) and more recently implemented in Refs. [**Djouadi:1997yw**, **Krause:2018wmo**] in the 2HDM. As for production channels, the by far most relevant one is gluon-gluon fusion, i.e., $gg \to A$ or H, with an occasional competing contribution from $b\bar{b} \to A$ or H, respectively.

LHC searches for the complete channels $gg, b\bar{b} \to A \to ZH$ and $gg, b\bar{b} \to H \to ZA$ have been carried out at both ATLAS [**Aaboud:2018eoy**] and CMS [**Khachatryan:2016are**, **Sirunyan:2019wrn**], by exploiting leptonic decays of the gauge boson, $Z \to l^+l^-$ ($l = e, \mu$), and hadronic decays of the accompanying neutral Higgs state, in particular, H or $A \to b\bar{b}$ or $\tau^+\tau^-$. Based on this approach, current experimental data exclude heavy neutral Higgses with masses up to about 600–700 GeV, depending on the BSM Higgs spectrum and the value of $\tan(\beta)$, the ratio of the Vacuum Expectation Values (VEVs) of the aforementioned two Higgs doublets. These findings are broadly in line with previous phenomenological results obtained in Ref. [**Coleppa:2014hxa**], which had forecast the LHC scope in accessing both $A \to ZH$ and $H \to ZA$ decays in a variety of final states.

Here, we consider the final state $l^+l^-b\bar{b}$ and start from the results of [**Aaboud:2018eoy**] for the $A \to ZH$ decay in order to obtain the corresponding ones for the complementary channel $H \to ZA$, and extend the analysis to higher energy and luminosity from planned detector upgrades.

The different transformations of the quark fields under the Z_2 symmetry lead to four structures of Higgs-fermions interactions: in Type-I only one doublet couples to all fermions; in Type-II one of the doublets couples to the up quarks while the other doublets couples to the down quark; in Type-X (or Lepton specific) one of the doublets couples to all quarks and the other couples to all leptons; in Type-Y (or Flipped) one of the doublet couples to up-type quarks and to leptons and the other couples to down-type quarks.

3.1 Scan

In this study, we identify the lightest CP-even Higgs boson of the 2HDM as the observed Higgs state at the LHC, with $m_h = 125$ GeV, and assume $\sin(\beta - \alpha) = 1$.

We scan over the following parameter range:

$$\begin{gathered} m_h = 125 \text{ GeV},\ \sin(\beta-\alpha) = 1, 0 < m_{12}^2 < 2\times 10^5 \text{ GeV}, \\ 130\text{GeV} < m_X < 700 \text{ GeV},\ m_X \geq m_Y + 100 \text{ GeV}, \\ m_X, m_Y \text{ chosen at 10 GeV intervals.} \\ \tan(\beta) \in \begin{cases} \{1,\ 2,\ 3\}, & \text{if Lepton Specific} \\ \{1,\ 5,\ 10,\ 20\}, & \text{otherwise} \end{cases} \end{gathered} \tag{2}$$

The set of values chosen for $\tan(\beta)$, and the masses, align with the choices in [**Aaboud:2018eoy**].

- For the process mediated by $A \to ZH$, we choose $m_X = m_A$, $m_Y = m_H$ and $m_{H^\pm} = m_A$. (Note that this choice is consistent with Ref. [**Aaboud:2018eoy**].)
- For the process mediated by $H \to ZA$, we choose $m_X = m_H$, $m_Y = m_A$ and $m_{H^\pm} = m_H$. (Note this choice is specular to that in Ref. [**Aaboud:2018eoy**].)

While an evident symmetry exists between the two cases, neither the constraints affecting the two processes nor their sensitivity reaches should expected to be. On the one hand, the role played by the heavy CP-even and CP-odd Higgs states of the 2HDM in both theoretical and experimental limits is different, owing to their different quantum numbers (and hence couplings). On the other hand, their production and decay rates at the LHC are different despite leading to the same final states, including residual differences due to width effects entering their normalisation (but, as mentioned, not their kinematics), since, e.g., the A state does not decay to W^+W^- and ZZ pairs while the H state does and, conversely, the A state decays to Zh while the H state does not. However, in the alignment limit used here these decay channels are closed.

3.1.1 Theoretical constraints

Within these ranges there are several theoretical and experimental constraints for the parameter points of the 2HDM to pass, discussed below.

- Unitarity: various scattering processes require that unitarity is conserved at the tree-level at high energy. The unitarity requirements in the 2HDM have been studied in [**Kanemura:1993hm**, **Akeroyd:2000wc**, **Arhrib:2000is**]. Sets of eigenvalues e_i $(i-1,...12)$ for the scattering matrix of all Higgs and Goldstone bosons of the 2HDM are obtained as follows:

$$\begin{aligned} &e_{1,2} = \lambda_3 + 2\lambda_4 \pm 3|\lambda_5|, \qquad e_{3,4} = \lambda_3 \pm \lambda_4, \quad e_{5,6} = \lambda_3 \pm |\lambda_5|, \\ &e_{7,8} = 3(\lambda_1+\lambda_2) \pm \sqrt{9(\lambda_1-\lambda_2)^2 + 4(2\lambda_3+\lambda_4|)^2}, \\ &e_{9,10} = \lambda_1+\lambda_2 \pm \sqrt{(\lambda_1-\lambda_2)^2 + 4|\lambda_5|^2}, \\ &e_{11,12} = \lambda_1+\lambda_2 \pm \sqrt{(\lambda_1-\lambda_2)^2 + 4|\lambda_5|^2}. \end{aligned} \tag{3}$$

  We require all e_i's to be less than 16π for each $i = 1, ...12$.

- Perturbativity constraints [**Kanemura:1993hm**, **Branco:2011iw**] implies that all that the quartic couplings of the scalar potential satisfy the condition $|\lambda_i| \leqslant 8\pi$ for each $i = 1, ...5$.

- Vacuum stability requires the scalar potential to be bounded from below [**Gunion:2002zf**] by satisfying the following inequalities:

$$\lambda_{1,2} > 0,\ \lambda_3 > -\sqrt{\lambda_1\lambda_2},\ \lambda_3 + \lambda_4 - |\lambda_5| > -\sqrt{\lambda_1\lambda_2}. \tag{4}$$

In practice the theoretical constraints are automatically checked by the program 2HDMC [**Eriksson:2009ws**]. The program will tell us if a selected parameter combination is valid, however owing to the range of possible values in m_{12}^2 some level of curve fitting to the valid points is required to augment the MC sampling. Points that satisfy the most constraints are fitted to a polynomial and values of m_{12}^2 near the surface of the polynomial is samples further. The objective is to locate values that are permitted by all theory constraints.

3.1.2 Experimental constraints

In addition to theoretical constraints experimental constrains need to be accounted for.

- EW Precision Observables (EWPOs) [**Haller:2018nnx**], such as the oblique parameters S and T [**Peskin:1991sw**, **Grimus:2008nb**], require a level of degeneracy between the charged Higgs boson state and one of the heavier neutral Higgs bosons. Here, we assume $m_{H^\pm} = m_A$ or m_H, as appropriate (see below), so that the T parameter exactly vanishes in the alignment limit.
- Exclusion limits at 95% Confidence Level (CL) from Higgs searches at colliders (LEP, Tevatron and LHC) via HiggsBounds, version 5.3.2 [**Bechtle:2008jh**, **Bechtle:2011sb**, **Bechtle:2013wla**] are enforced. Furthermore, the ATLAS Collaboration has set an upper limit at 95% CL on the production cross section σ of the A state times its decay BR into $ZH \to l^+l^-b\bar{b}$, i.e., $\sigma(A) \times \mathrm{BR}(A \to ZH \to l^+l^-b\bar{b})$ [**Aaboud:2018eoy**], that is not included in this tool, hence we have accounted for it separately.

tan(β)	1	5	10	20
Type-I	Flavour constraints	Some masses	Many masses	Low sensitivity
Type-II	Flavour constraints	Some masses after upgrade	Some masses after upgrade	Theory constraints
Flipped	Flavour constraints	Some masses after upgrade	Some masses after upgrade	Theory constraints
tan(β)	1	2	3	
Lepton specific	Flavour constraints	Excluded by HiggsBounds	Excluded by HiggsBounds	

Table 1: Table summarising the findings in Figs. 3 to 6. An overview of the possibility of each Yukawa type and value of tan(β) is given. Entries in red indicate that the combination has little or no mass combinations that are not forbidden while those in blue represent available parameter space accessible presently at Run 2 or after the upgrade of Run 3.

- Constraints from the Higgs boson signal strength measurements are automatically satisfied as we assume $\sin(\beta-\alpha)=1$.
- Constraints of flavour physics observables, namely, $B\to X_s\gamma$, $B_{s,d}\to\mu^+\mu^-$ and $\Delta m_{s,d}$ [**Haller:2018nnx**].

Comparing to observed and expected exclusion limits requires obtaining branching ratios and production cross sections. The relevant branching ratios, $A\to ZH$, $H\to ZA$, $A\to b\bar{b}$ and $H\to b\bar{b}$ are calculated with 2HDMC [**Eriksson:2009ws**]. The production cross sections of the heavy CP-even (H) and CP-odd (A) Higgs bosons, at Next-to-Next-to-Leading Order (NNLO) in QCD, for both $gg\to H,A$ and $b\bar{b}\to A,H$, at the Centre-of-Mass (CM) energies of 13 TeV and 14 TeV, are calculated using SusHi [**Harlander:2012pb**, **Harlander:2016hcx**, **Harlander:2002wh**, **Harlander:2003ai**].

2HDMC code also includes an interface to HiggsBounds, which is used to apply the aforementioned exclusion limits at 95% CL from Higgs searches at LEP, Tevatron and LHC.

The choice of $m_{12}^2=m_A^2\tan(\beta)/(1+\tan(\beta))^2$ enables us to reconstruct the exclusion limits at 95% CL given in Ref [**Aaboud:2018eoy**]. However, this choice does not actually allow to satisfy theoretical constraints in all four types of 2HDM.

The first part of this study deals with the two production and decay processes $pp\to H(A)\to ZA(H)\to b\bar{b}l^-l^+$. The observed and expected confidence limits for all four types of Yukawa couplings in the 2HDM are produced at $\sqrt{s}=13$ TeV, with an integrated luminosity, L, of 36.1 fb^{-1}, by combining our calculations with the data from Ref. [**Aaboud:2018eoy**]. In the second part, we rescale the expected exclusion limit to the CM energy of $\sqrt{s}=14$ TeV, with an integrated luminosity of 300 fb^{-1}, by calculating the so called 'upgrade factor' for both signals and backgrounds, while retaining the acceptance and selection efficiencies of the analysis at the lower $\sqrt{s}$ value. The change in energy will naturally affect signals and backgrounds differently. We treat the former by using SusHi (as intimated) and the latter by using MadGraph5, version 2.6.4 [**Alwall:2011uj**]. (For completeness, the background is considered to be any reducible or irreducible SM process that creates a pair of b-jets plus a pair of electrons or muons, as in Ref. [**Aaboud:2018eoy**].)

3.2 Numerical results

After performing a scan over the parameter space delimited by Eq. (2), we compare the prediction of the model with the observed and expected limits given in Ref. [**Aaboud:2018eoy**]. If the prediction exceeds the observed limit, then the parameter combination is excluded. When the prediction exceeds the expected limit, we anticipate that the signal would be visible above background given the energies and luminosities available, hence, the experiment is sensitive to these parameters.

The choice of $m_{12}^2=m_A^2\tan(\beta)/(1+\tan(\beta))^2$ enables us to reconstruct the exclusion limits at 95% CL given in Ref [**Aaboud:2018eoy**]. However, this choice does not actually allow to satisfy theoretical constraints in all four types of 2HDM. Therefore, we have dismissed it in our analysis. In contrast, our choice of m_{12}^2 above aims to simultaneously satisfy as many theoretical constraints as possible while affording one with significant parameter space amenable to experimental investigation. Indeed, this is achieved by randomly sampling values of m_{12}^2 between 0 and 2×10^5 GeV for each point of the scan and selecting the one that passes most theoretical checks.

Figs. 3 to 6 illustrate the outcome the scan for each Yukawa type, tan(β) and mass combination (m_H,m_A). Each figure provides results for one choice of Yukawa couplings and each frame in each figure provides results at one

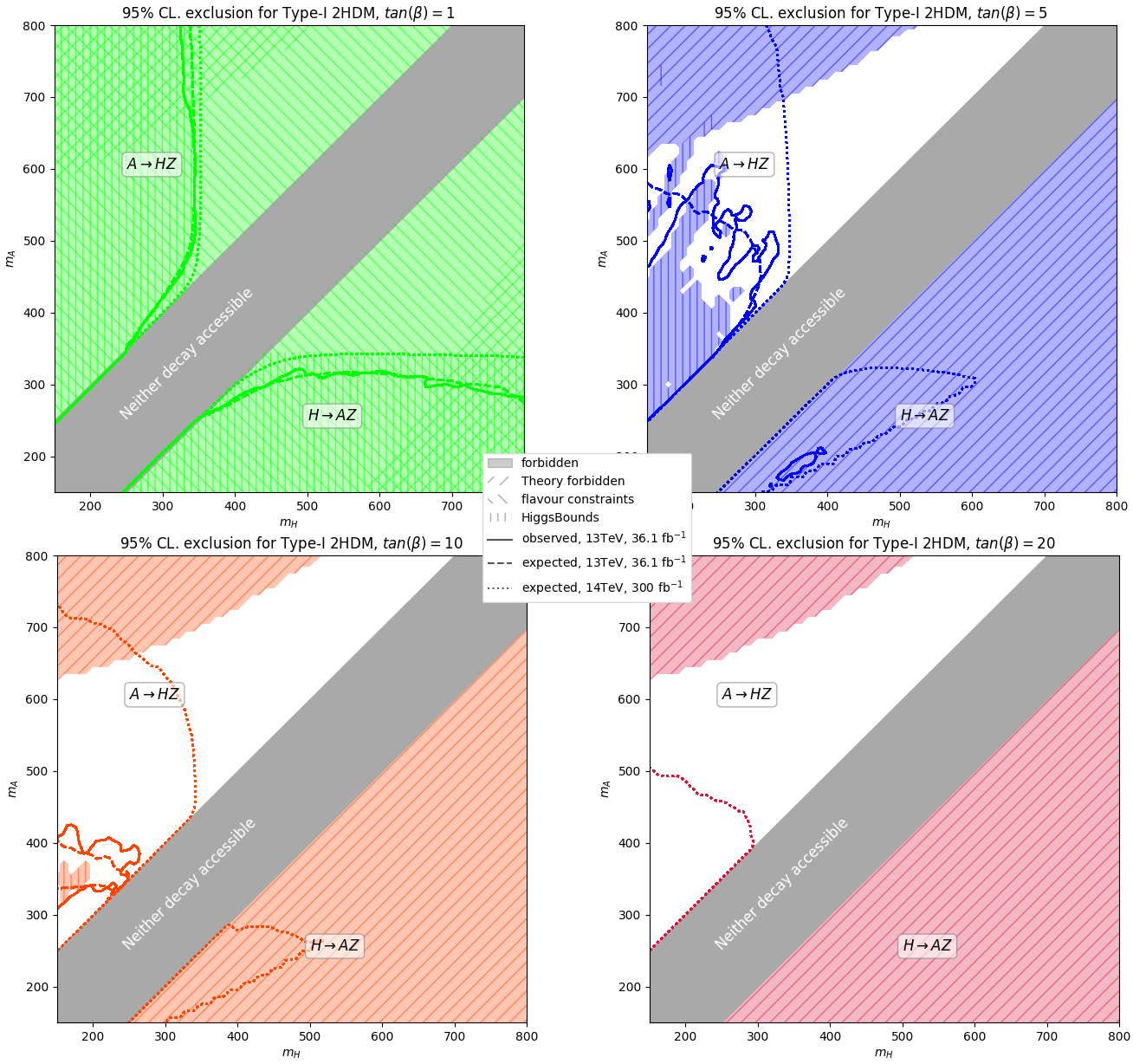

Figure 3: Exclusion limits at 95% CL in Type-I. The lines denoting expected and observed exclusion limits do not appear at all on some plots when the prediction never exceeds the expected or observed limit.

value of $\tan(\beta)$. In the top left of each plot, where $m_A > m_H + 100$ GeV, the decay $A \to ZH$ is considered while in the bottom right of each plot, where $m_H > m_A + 100$ GeV, the decay $H \to ZA$ is considered. The corridor along the diagonal between these regions is coloured grey to indicate that neither decay is accessible. If a combination of parameters is forbidden by theory, HiggsBounds or flavour constraints then the corresponding area is filled with solid colour, conversely, white areas pass all these checks and so are of interest. The hatching over the solid colour is used to indicate which of the checks causes the corresponding parameter combination to fail. There are three boundary lines drawn over the plots: these are the observed and expected 95% CLs for the ATLAS detector in its present state, 13 TeV and 36.1 fb^{-1}, plus the expected 95% CL for an upgraded LHC and ATLAS detector at 14 TeV and 300 fb^{-1}[1]. The model predictions exceed the 95% CL inside the curve.

In Fig. 3 the parameter space with Type-I Yukawa couplings is shown. The upper left plot shows that $\tan(\beta) = 1$ is always forbidden by flavour constraints. The upper right plot shows that there are many mass combinations that do not prevent the decay $A \to ZH$ for $\tan(\beta) = 5$, but theory constraints forbid all mass combinations relevant

[1]We neglect here to consider the case of $\sqrt{s} = 13$ TeV and $L \approx 140$ fb^{-1}, as it only improves marginally the present situation yet it would be make the plots far too crowded.

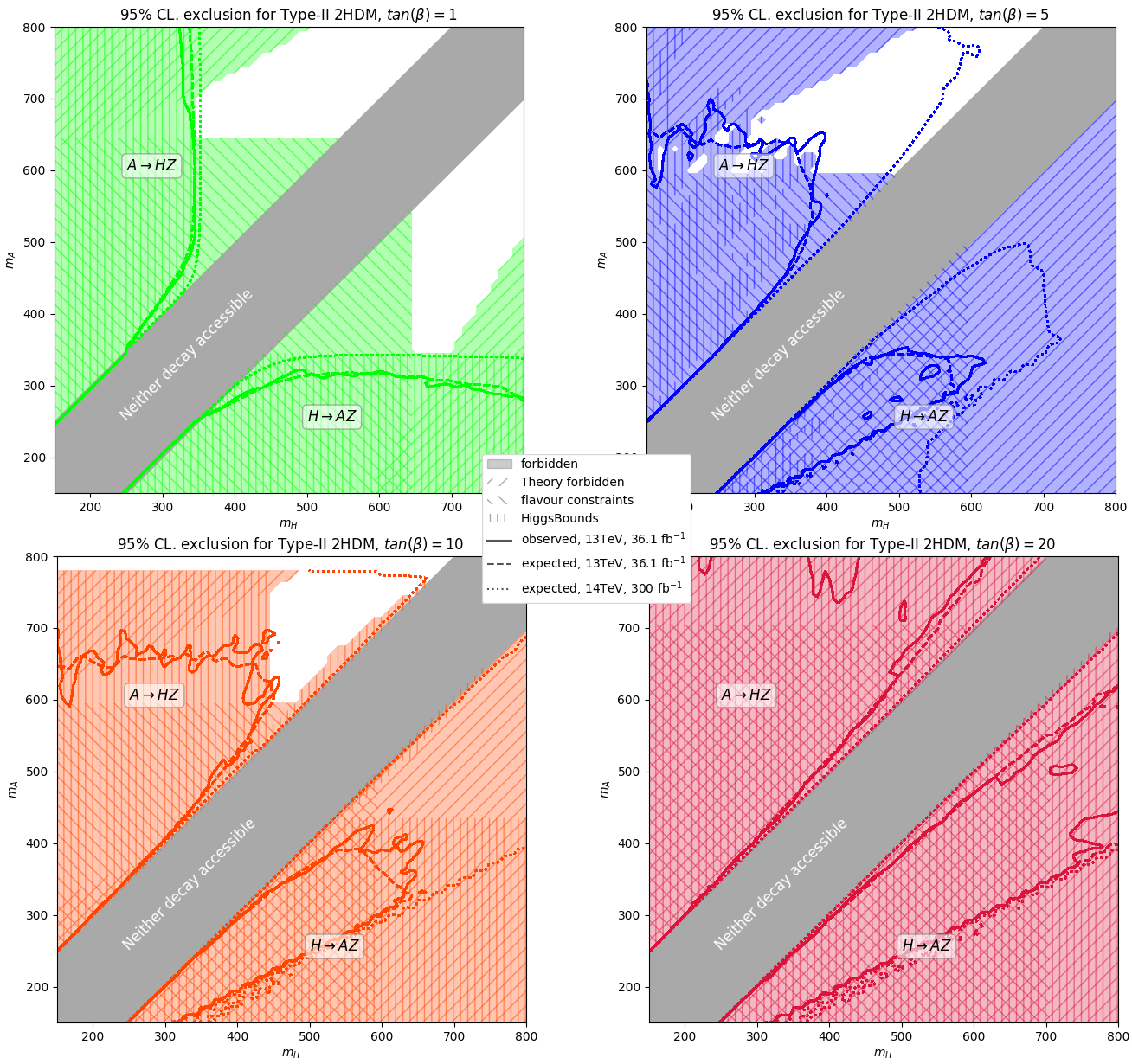

Figure 4: Like in Fig. 3 but for Type-II.

to $H \to ZA$. At $\tan(\beta) = 5$ for 13 TeV (and 36.1 fb^{-1}) the area of sensitivity (inside the expected curve) that is not excluded by observation (inside the observed curve) is very limited. It is also seen that the $H \to AZ$ signal has reduced sensitivity when $\tan(\beta)$ is 5 or more. This is due to $H \to AA$ competing with $H \to AZ$, as shown in figure 7. The branching ratio $H \to AA$ becomes significant because of the enhancement of the trilinear coupling λ_{HAA} at large $\tan(\beta)$. At 14 TeV and 300 fb^{-1}, however, we expect many mass combinations to be testable that have not yet been excluded. The lower left plot shows the behaviour at $\tan(\beta) = 10$ to be similar to $\tan(\beta) = 5$, i.e., everything is forbidden for $H \to ZA$ by theory while for $A \to ZH$ most combinations for which there is sensitivity have been excluded at 13 TeV but 14 TeV offers even more possible parameter space than seen at $\tan(\beta) = 5$. Finally, in the lower right frame of Fig. 3, the parameter space for $\tan(\beta) = 20$ is shown. The state of $H \to ZA$ is unchanged, but now $A \to ZH$ has no expected or observed exclusion at 13 TeV, i.e., these parameters are harder to probe. With the upgrade to 14 TeV and 300 fb^{-1} there is some sensitivity to $A \to ZH$ at $\tan(\beta) = 20$.

As might be expected, the behaviour of Type-II, shown in Fig. 4 and Type-Y, shown in Fig. 5, is remarkably similar. The upper left plot shows that $\tan(\beta) = 1$ is forbidden by flavour constraints in all areas where there is sensitivity. At 13 TeV and 36.1 fb^{-1} the upper right plot shows that the same can be said for $\tan(\beta) = 5$, however, after Run 3, at 14 TeV and 300 fb^{-1}, there are many permitted mass combinations for $A \to ZH$. However,

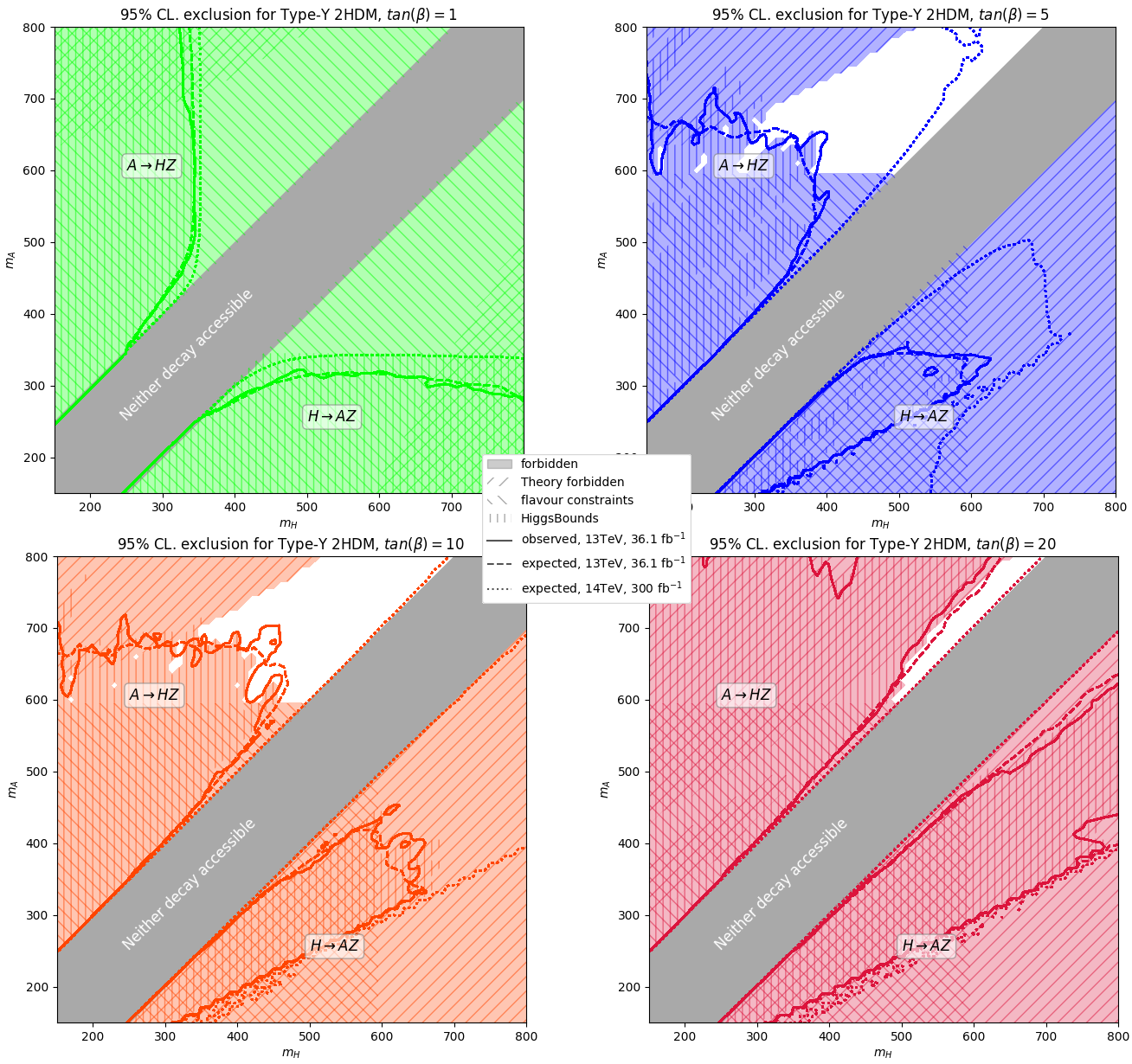

Figure 5: Like in Fig. 3 but for Type-Y (Flipped).

$H \to ZA$ is excluded by theory. The behaviour at $\tan(\beta) = 10$, shown in the lower left plot, is much the same as for $\tan(\beta) = 5$, except more of the exclusion at 13 TeV and 36.1 fb^{-1} is from observations provided by HiggsBounds. Finally, in the lower right plot, $\tan(\beta) = 20$ is shown to be excluded for almost all mass choices, by multiple constraints.

In Fig. 6 the behaviour of the Type-X 2HDM is shown, at a set of $\tan(\beta)$ values that differs from those previously considered. The change is made because the parameter space in Type-X shrinks more rapidly with increasing $\tan(\beta)$ compared to the other Yukawa types. For these choices HiggsBounds excludes all areas inside the expected limits. This remains true even after the end of Run 3.

Finally, Tab. 1 summarises our findings, highlighting that sensitivity only really exists for $5 < \tan(\beta) < 10$ and limitedly to the 2HDM Type-I, both at Run 2 and 3, and -II and -Y (or Flipped), but only at Run 3. The case of Type-X (or Lepton specific) is never accessible.

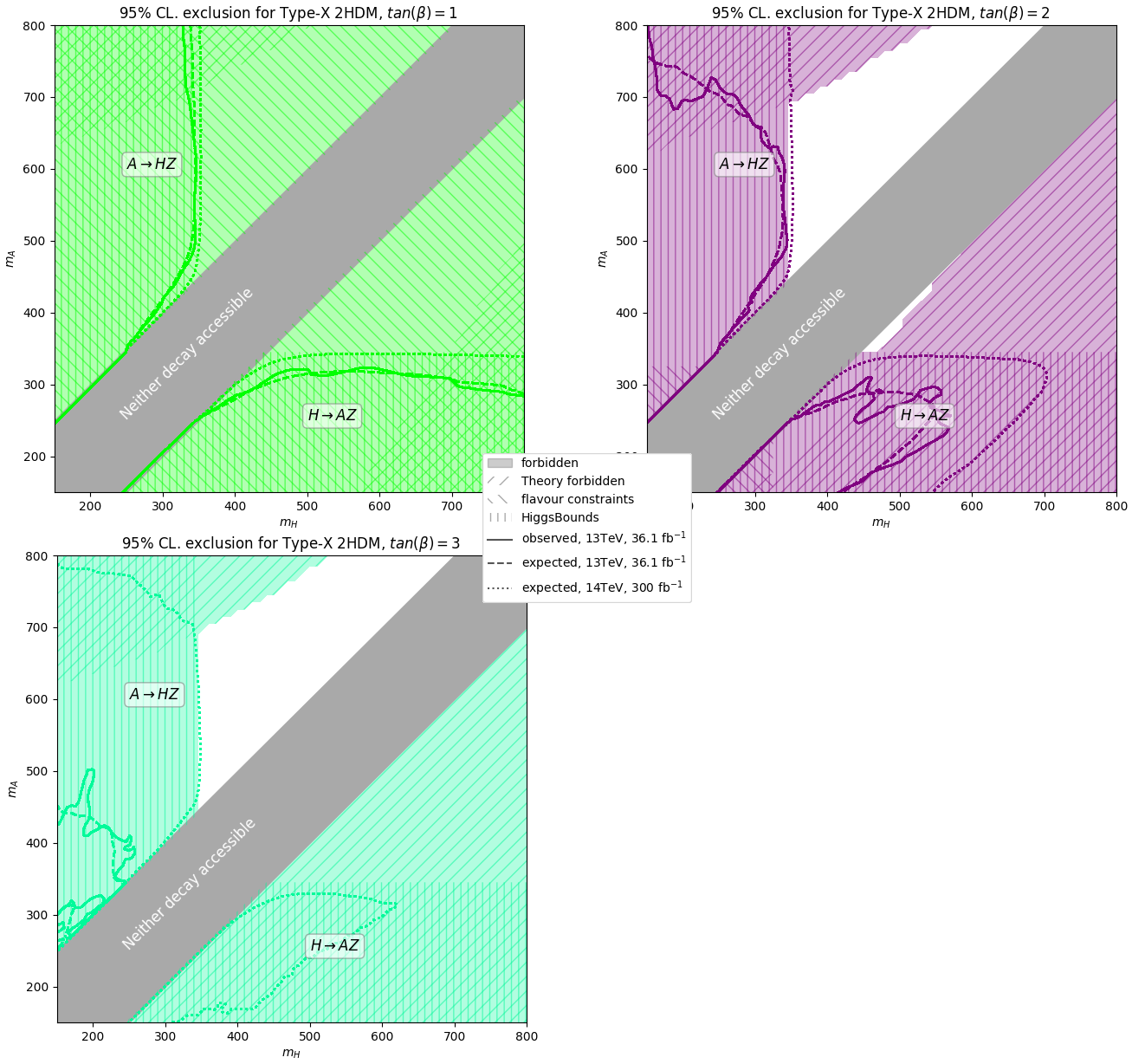

Figure 6: Like in Fig. 3 but for Type-X (Lepton specific).

3.3 Conclusions

In summary, we have revisited an experimental analysis of the ATLAS Collaboration of the production and decay process $gg, b\bar{b} \to A \to ZH \to l^+l^-b\bar{b}$ performed at Run 2 with 36.1 fb^{-1} of luminosity, which had been interpreted in terms of exclusion limits over the parameter space of the four types of the 2HDM, wherein the lightest Higgs state is identified with the SM-like Higgs boson discovered during Run 1 at the LHC with mass 125 GeV. Upon validating the ATLAS interpretation in our framework, though, we have discovered that their (fixed) choice of m_{12}, a mass parameter in the 2HDM Lagrangian that softly breaks an underlying Z_2 symmetry of the 2HDM to avoid FCNCs, yields parameter space configurations which are ruled out by theoretical requirements of model consistency. Hence, we have allowed this parameter to vary freely and subject the ensuing parameter space configurations to both the aforementioned theoretical constraints as well as those emerging from past and present experiments, thereby redrawing the actual sensitivity of such an experimental search to all four Yukawa types of the 2HDM, as a function of tan(β). In doing so, we have have also forecast the potential sensitivity of this channel to the 2HDM parameter space at the end of Run 3, assuming increased energy to 14 TeV and luminosity to 300 fb^{-1}. This revealed some extended coverage of the 2HDM Type-I, -II and -Y (but not -X), especially for intermediate tan(β) values (say,

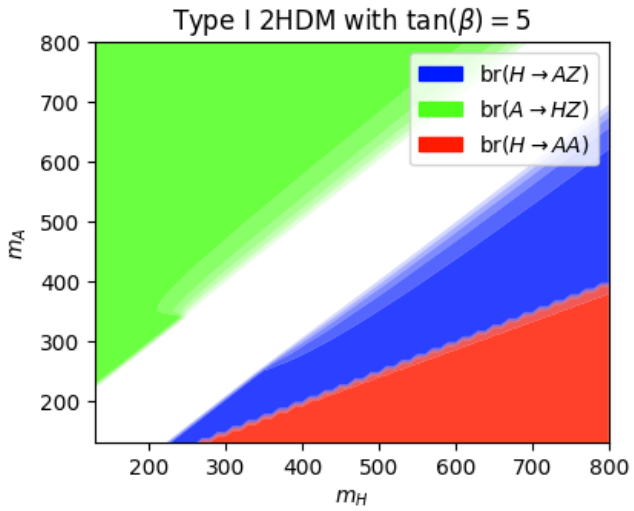

Figure 7: The branching ratio $H \to AZ$ is suppressed by the branching ratio $H \to AA$. This effect occurs for all types, but does not occur at small $\tan(\beta)$.

between 5 and 10), with m_A up to 800 GeV and m_H up to 700 GeV. This is somewhat beyond what is presently covered, i.e., up to 150 GeV or so in mass of either Higgs state, so as to justify further searches for this signature at the next stage of the LHC. Finally, we have recast the sensitivity of this analysis onto that of the channel $gg, b\bar{b} \to H \to ZA \to l^+l^-b\bar{b}$. However, we have found that the complementary parameter space accessible this way (i.e., $m_H \geq m_A + m_Z$) is actually entirely excluded already by existing theoretical and/or experimental constraints, so as to conclude that it is not warranted to pursue further this channel at the LHC, at least, not with a view to interpret it in the context of the standard four Yukawa types of the 2HDM[2].

4 Jet Clustering

This section exhibits a possible alternative method for jet clustering.

The default choice for jet clustering tends to be one of there algorithms; the anti-kt algorithm [9], the Cambridge-Aachen algorithm [24] and the kt algorithm [13]. They have been the default choice for some time because they have a number of desirable properties. They are infrared safe, excellent implementations of them are available (see `FastJet` [8]) and it is flexible enough to capture many signals with minimal parameter change. These algorithms are recursive and agglomerative. A recursive algorithm is well suited to clustering objects when the number of groups is not known at outset. Agglomerative algorithms are easier to design in a manner that is infrared safe, as they can recombine soft and collinear emissions in early steps.

Finding a clustering method that compares favourably to these algorithms is challenging. Spectral clustering is a candidate that has had considerable success in other studies. In fluid dynamics spectral clustering has been used to identify the motion of vortices [14], finding that it is possible to successfully identify the vortex structures in cases with less data available. It was also seen that spectral clustering was proficient at determining the correct number of clusters to be found in the fluid. To reduce the risk of blackouts, power grids may be subdivided into 'islands'. The ideal allocation is found by minimising power flow between islands, and it was shown in [19] that spectral clustering can produce a good solution in less time than other algorithms commonly used for this problem.

To the authors knowledge this clustering algorithm has not yet been applied to jet physics, however, given its recursive, agglomerative form it could be a good fit.

[2]We finally note that analyses similar to Ref. [**Aaboud:2018eoy**] performed by the CMS Collaboration exist [**Khachatryan:2016are**, **Sirunyan:2019wrn**]. We have not used these for two reasons. On the one hand, they did not convey all the information necessary to make extrapolations to higher energies. On the other hand, they did not afford one with significantly different sensitivity to the 2HDM at present energies than what achieved by the ATLAS analysis [**Aaboud:2018eoy**] that we have adopted as benchmark.

5 Theory of spectral clustering

Add a short outline of how spectral clustering should work

Spectral clustering is the result of relaxing the criteria that would precisely to identify infinitely separated groups to estimate the members groups that may have some connection. This is best described in [15], a short summary is given here.

Let us imagine a graph, disconnected in n clusters. The initial aim is to identify which of the disconnected components each point belongs to. The objective used for a good clustering will be the NCut objective;

$$\text{NCut} = \sum_k \frac{W(A_k, \bar{A_k})}{|A_k|} \tag{5}$$

Where $W(A_k, \bar{A_k})$ measures the strength of the connections that must be broken in order to separate the cluster A_k from the graph and $|A_k|$ corrisponds to the number of elements in A_k.

Membership of cluster A_k will be determined by the indicator vector h_k;

$$h_{i,k} = \begin{cases} 1/\sqrt{|A_k|}, & \text{if point } i \in A_k \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $|A_k|$ is the number of points in set A_k. The edge between two points is assigned an affinity, $a_{i,j} = a_{j,i}$, of either 1 for connected i and j or 0 if points i and j are unconnected.

The graph is represented by the graph Laplacien, a square matrix with as many rows and columns as there are points. On ith diagonal it holds the value $\sum_j a_{i,j}$, which is the degree of the vertex i. Off the diagonal it has the negative affinity; $-a_{i,j}$. Notice that this is a real symmetric matrix, and therefore all it's eigenvalues are real.

Considering just one cluster we find that when the Laplacien is multiplied by two indicator vectors the result is the function that NCut seeks to minimise for that cluster.

$$h_k' L h_k = \frac{1}{|A_k|} \sum_{i \in A_k, j \in A_k} \left(\delta_{i,j} \sum_l a_{l,i} - a_{i,j} \right) = \frac{W(A_k, \bar{A_k})}{|A_k|} \tag{7}$$

Stack the indicator vectors into a matrix;

$$h_k' L h_k = (H' L H)_{kk} \tag{8}$$

and the NCut aim described earlier becomes the trace;

$$\text{NCut}(A_1, A_2, \dots A_n) \equiv \frac{1}{2} \sum_{i=1}^{n} \frac{W(A_i, \bar{A_i})}{|A_i|} = \text{Tr}(H' L H) \tag{9}$$

Where $H'H = I$. Trace minimisation in this form is done by finding the eigenvectors of L with smallest eigenvalues.

Generalising this to a graph that is not disconnected only requires relaxing the requirements on the form of the indicator vectors; h_k.

5.0.1 Impact of p_T factors

In the case of reconstructed particles, a possible input to the affinity measure is the p_T of the particles. A format commonly used for the distance is;

$$d_{i,j} = \min(p_{T,i}^{2x}, p_{T,j}^{2x}) \Delta R'_{i,j} \tag{10}$$

where $\Delta R'_{i,j}$ is an angular distance. This is discussed from a practical standpoint in section **??**, from a theoretical standpoint it is important to investigate the impact of the p_T factor on the eigenvalue equations.

There are a number of different ways of converting a distance measure into an affinity, they are discussed in section **??**. For now only one will be considered $a_{i,j} = 1/d_{i,j}$.

Starting with the **unnormalised laplacien** a relation between the size of the eigenvalues and the affinities in each group can be uncovered. The unnormalised laplacien has the form $L = D - A$, as we are already using d to mean distance, let us take $z_i = \sum_{j \neq i} a_{i,j}$ where it is understood that by default a sum goes over all node indices. From this point onwards it is also assumed that all $a_{i,i} = 0$, so such sums can omit the $i \neq j$. This gives

$$L = \begin{pmatrix} z_0 & -a_{0,1} & -a_{0,2} & \dots \\ -a_{0,1} & z_1 & -a_{1,2} & \\ -a_{0,2} & -a_{1,2} & z_2 & \\ \vdots & & & \ddots \end{pmatrix} \tag{11}$$

Now imagine a group that includes the first q nodes, it's indicator vector approximately contains two values; $-Q_0$ for nodes 0 to q inclusive and Q_1 for nodes $q+1$ to n inclusive, where $Q_{0/1} > 0$. As the ordering of the nodes is arbitrary (the rows and columns can have any order) conclusions drawn for this hold for any group containing q nodes. This must be approximate because we do not require that the q nodes are perfectly separated, there for the eigenvalue equation is a relaxation.

The eigenvalue equation has the form

$$\begin{pmatrix} z_0 & -a_{0,1} & -a_{0,2} & \dots \\ -a_{0,1} & z_1 & -a_{1,2} & \\ -a_{0,2} & -a_{1,2} & z_2 & \\ \vdots & & & \ddots \end{pmatrix} \begin{pmatrix} -Q_0 \\ -Q_0 \\ \vdots \\ Q_1 \end{pmatrix} \approx \lambda \begin{pmatrix} -Q_0 \\ -Q_0 \\ \vdots \\ Q_1 \end{pmatrix} \tag{12}$$

.

The simultaneous equations that come from this can be simplified into two forms. The first q, where $i \leq q$, take the form

$$-Q_0\lambda \approx Q_0 \sum_{j \leq q} a_{i,j} - Q_0 z_i - Q_1 \sum_{j>q} a_{i,j} \tag{13}$$

$$\lambda \approx \left(1 + \frac{Q_1}{Q_0}\right) \sum_{j>q} a_{i,j} \tag{14}$$

. In a similar manner the remaining simultaneous equations, where $k > q$ have the form

$$\lambda \approx \left(1 + \frac{Q_0}{Q_1}\right) \sum_{j \leq q} a_{k,j} \tag{15}$$

. In both cases the sum is over affinities that attach to the node associated with the row and cross the cluster boundary.

Clearly the size of the eigenvalue will be related to the size of the affinities that cross the boundary. More subtly, if the number of elements in each group differs strongly then the values of Q_0 and Q_1 will need to differ more to compensate. Approximating all the affinities to the same value

$$\left(1 + \frac{Q_1}{Q_0}\right) qa = \left(1 + \frac{Q_0}{Q_1}\right) (n-q)a \tag{16}$$

$$\left(1 + \frac{Q_1}{Q_0}\right) = \left(1 + \left(\frac{Q_1}{Q_0}\right)^{-1}\right) \frac{q-n}{q} \tag{17}$$

it can be seen that as $\frac{q-n}{q}$ deviates from 1 the solution grows the left side faster than it shrinks the right. This will also tend to increase the value of λ.

So essentially a split associated with a small eigenvalue will minimise the value of the crossing affinities and create groups with equal number of members. Returning to the idea of the influence of p_T, this will have impact on the former condition but not the latter. Taking $a_{i,j} \sim 1/\min(p_{T,i}^{2x}, p_{T,j}^{2x})$ with $x > 0$ affinities involving a low p_T particle will be larger, so the placement of soft emissions will be important. Alternatively, with $x < 0$ affinities involving one high p_T particle will be larger, so the placement of hard emissions will be most important.

This has found the role of p_T in an unnormalised laplacien, the same steps can be taken for the **symmetric laplacien**; $L = D^{-1/2}(D-A)D^{-1/2}$.

$$L_{\text{symm}} = \begin{pmatrix} 1 & -a_{0,1}(z_0 z_1)^{-1/2} & -a_{0,2}(z_0 z_2)^{-1/2} & \dots \\ -a_{0,1}(z_0 z_1)^{-1/2} & 1 & -a_{1,2}(z_1 z_2)^{-1/2} & \\ -a_{0,2}(z_0 z_2)^{-1/2} & -a_{1,2}(z_1 z_2)^{-1/2} & 1 & \\ \vdots & & & \ddots \end{pmatrix} \tag{18}$$

Considering the same group of the first q rows the forms of the simultaneous equations are; for $i \leq q$

$$\lambda \approx 1 + z_i^{-1/2} \left(\frac{Q_1}{Q_0} \sum_{j>q} a_{i,j} z_j^{-1/2} - \sum_{j \leq q} a_{i,j} z_j^{-1/2} \right). \tag{19}$$

Notice that the first sum is over affinities that cross out of the group, divided by the root of the connectedness of the outside node, (z_j can be seen as the connectedness of node j), the second sum is over affinities contained within the first q nodes, also divided by the root of their connectedness.

Then for $k > q$

$$\lambda \approx 1 + z_k^{-1/2} \left(\frac{Q_0}{Q_1} \sum_{j \leq q} a_{k,j} z_j^{-1/2} - \sum_{j>q} a_{k,j} z_j^{-1/2} \right). \tag{20}$$

Again the first sum is over affinities that cross the group, and now the second sum is over affinities not in the first q nodes. Now finding the impact of p_T looks rather more complicated, so it will be addressed in stages. Firstly consider $z_j = \sum_i a_{i,j} \sim \sum_i 1/\min(p_{Ti}^{2x}, p_{Tj}^{2x})$, it's behaviour splits into 4 categories. Consider $N_j(p_T)$ to be roughly the p_T of a node that node j is connected to (Neighbour). Also, let nodes on average have the equivalent of c connections, where $0 \leq c \leq n$.

	$p_{Tj} < N_j(p_T)$	$p_{Tj} > N_j(p_T)$
$x > 0$	$z_j \sim c p_{Tj}^{-2x}$ therefore $z_j^{-1/2} \sim c^{-1/2} p_{Tj}^{x}$	$z_j \sim c N_j(p_T)^{-2x}$ therefore $z_j^{-1/2} \sim c^{-1/2} N_j(p_T)^{x}$
$x < 0$	$z_j \sim c N_j(p_T)^{-2x}$ therefore $z_j^{-1/2} \sim c^{-1/2} N_j(p_T)^{x}$	$z_j \sim c p_{Tj}^{-2x}$ therefore $z_j^{-1/2} \sim c^{-1/2} p_{Tj}^{x}$

The first sum in equation 19 is $\sum_{j>q} a_{i,j} z_j^{-1/2}$, where i is in the first q nodes and j is always outside the first q nodes. Let $w_i = \sum_{j>q} a_{i,j} z_j^{-1/2}$. If connections were randomly allocated after clustering then the probability of a connection crossing between clusters would be $q(n-q)/n(n-q)$. However as clustering is supposed to avoid breaking connections the probability number of crossing connections will be lower than this. Call this χ, so that the sum $\sum_{j>q} a_{i,j}$ has on average χc non zero terms. A similar table is constructed for w_i;

	For most j; $p_{Tj} < p_{Ti}$	For most j; $p_{Tj} > p_{Ti}$
$x > 0$	$w_i \sim \chi c^{1/2} p_{Tj}^{-x}$	$w_i \sim \chi c^{1/2} N_j(p_T)^{x} p_{Ti}^{-2x}$
$x < 0$	$w_i \sim \chi c^{1/2} N_j(p_T)^{x} p_{Ti}^{-2x}$	$w_i \sim \chi c^{1/2} p_{Tj}^{-x}$

Note that from the perspective of node i connected to node j $N_j(p_T)$ relates to the p_T two steps away in the direction of j.

The other sums in equation 19 have similar form, the distinction being is indices i and j belong to the same or opposing clusters. This changes the prefactor, χ, as it is no longer a sum over crossing vertices, but otherwise the algebra will be the same as there are no assumptions about group membership in it.

The table is recast for those two cases, and the entries are described in terms of the influence of p_T values on the magnitude of the sum. Firstly, when both indices belong to the same cluster the sum is subtracted from the value of the eigenvalue. As such larger sums are favoured.

	Node i is high p_T so mostly $p_{Tj} < p_{Ti}$	Node i is low p_T so mostly $p_{Tj} > p_{Ti}$
$x > 0$	The smaller the p_T of a node connected to i the greater its contribution to the sum. The clustering is favoured when i connects to many soft nodes. The harder p_{Ti} is not used directly.	The larger the neighbourhood p_T for a node connected to i in comparison to soft p_{Ti} the greater its contribution to the sum. The clustering is favoured when nodes connecting to i sit in a higher p_T neighbourhood. The harder p_{Tj} is not used directly.
$x < 0$	The larger the neighbourhood p_T for a node connected to i in comparison to hard p_{Ti} the greater its contribution to the sum. The clustering is favoured when nodes connecting to i sit in a higher p_T neighbourhood. The softer p_{Tj} is not used directly.	The smaller the p_T of a node connected to i the greater its contribution to the sum. The clustering is favoured when i connects to many soft nodes. The softer p_{Ti} is not used directly.

The difference between $x > 0$ and $x < 0$ can be summarised as $x > 0$ indicates that things of differing scales should be connected and the highest p_Ts should not reduce any connectedness. On the other hand $x < 0$ indicates that things of the same scale should be clustered together and the lowest p_Ts should not reduce connectedness.

The sums that cross between groups are added to the value of the eigenvalue, so increasing the value of the sum disfavours the clustering. This symmetry means that the implications for $x > 0$ or $x < 0$ will be the same as summarised above.

5.0.2 Distances in Physical space

The distances in physical space will be input values for affinities, which will be clustered subject to the aims of RatioCut, or NCut. As a reminder, RatioCut aims to minimise

$$\text{RatioCut}(A_1 \dots A_k) = \frac{1}{2}\sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|} \tag{21}$$

. And NCut aims to minimise

$$\text{NCut}(A_1 \dots A_k) = \frac{1}{2}\sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{\text{Vol}(A_i)} \tag{22}$$

. Where A_i is a point inside the cluster, and $\bar{A}_i$ is a point outside the cluster.

As the distance grows the affinity shrinks, so to have these minimise on the right groupings the distances between clusters should be large compared to the distances inside them.

Three measures of this have been tried. See figure **??**. The first would be the ROC-AUC (Receiver Operating Characteristic - Area Under Curve), for the distance measure between two particles and the particles being descendent from the same b-quark. This can be thought of as a measure of how good that distance measure is at separating quarks from the same b-quark from other quark pairs.

The second would be the Spearman's rank of the measured distance against the Shower distance. Shower distance is defined by taking the simulated shower as a graph, where nodes are particles and vertices are interactions, and finding the least number of vertices that must be transvered in order to get between the particles in the pair.

These two are in relative agreement about which distance measures work best. The third is the normed distances; in each event the difference between the mean of distances that would be cut and the mean of distances between particle of the same b-quark is divided by the mean distance for that event. The mean of this measure over all event sis then taken. There seems to be some issue with this measure as it does not match well with the other two. It seems to prefer lower overall distances.

There are 10 distance measures that were tried in Physical space.

- Euclidean; $d = \sqrt{\sum_i \delta x_i^2}$
- L3; $d = \left(\sum_i \delta x_i^3\right)^{1/3}$
- L4; $d = \left(\sum_i \delta x_i^4\right)^{1/4}$
- Taxicab; $d = \sum_i |\delta x_i|$
- Braycutis; $d = \sum_i |x_{1,i} - x_{2,i}| / \sum_i |x_{1,i} + x_{2,i}|$
- Canberra; $d = \sum_i |x_{1,i} - x_{2,i}| / \sum_i |x_{1,i}| + |x_{2,i}|$
- Min; $d = \min_i |\delta x_i|$
- Max; $d = \max_i |\delta x_i|$
- Correlation; $d = 1 - \frac{\sum_i (x_{1,i} - \bar{x}_{1,i})(x_{2,i} - \bar{x}_{2,i})}{\sqrt{\sum_i (x_{1,i} - \bar{x}_{1,i})^2}\sqrt{\sum_i (x_{2,i} - \bar{x}_{2,i})^2}}$
- Cosine; $d = 1 - \frac{\sum_i x_{1,i} x_{2,i}}{\sqrt{\sum_i x_{1,i}^2}\sqrt{\sum_i x_{2,i}^2}}$

A comparison of the two scoring criteria over two thousand events can be seen in figure **??**.

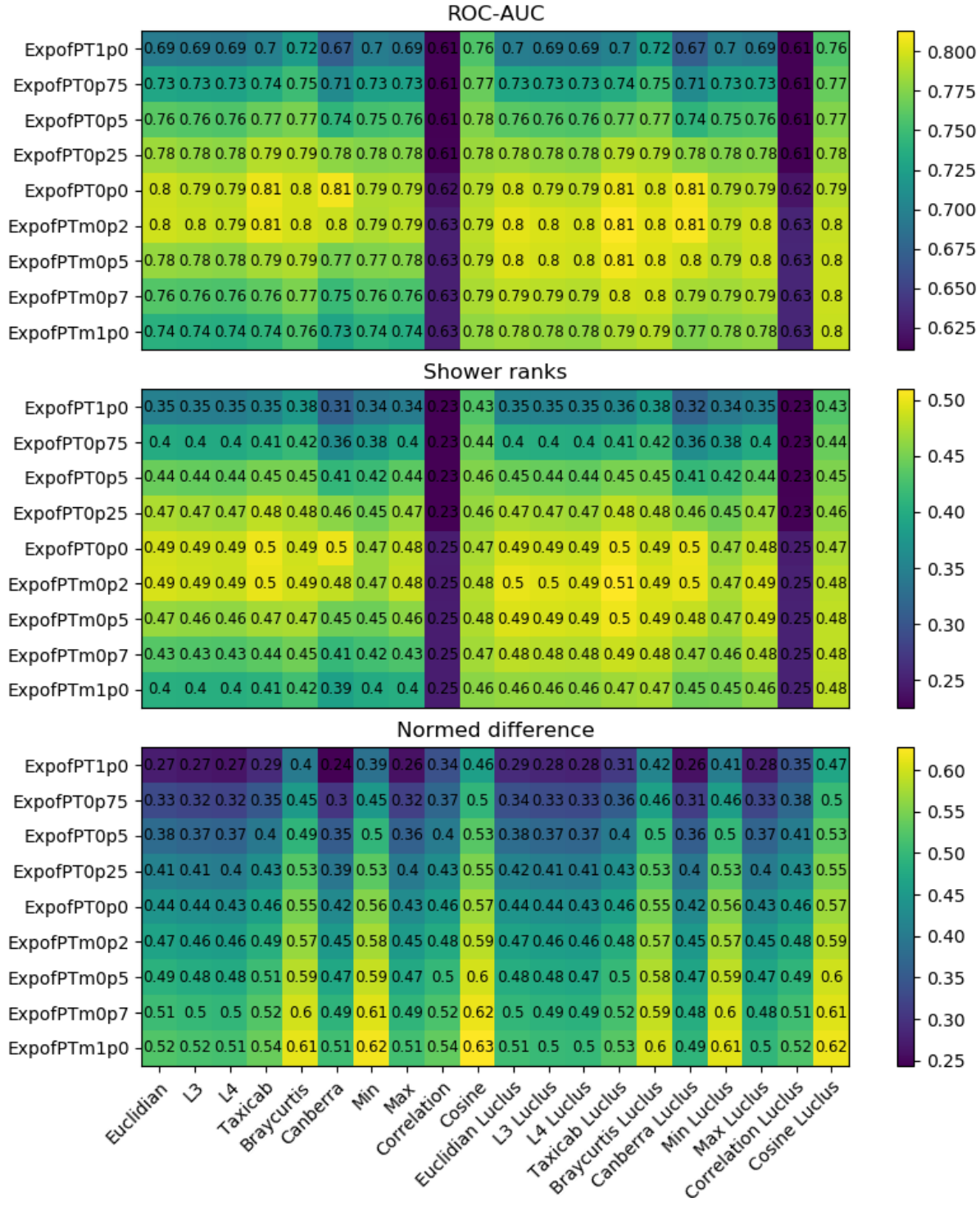

Figure 8: Along the y axis there are jets with varying p_T exponents, going from kt jets at the top, Cambridge Aachen in the middle row and anti-kt on the last row. On the x axis are various distance measures. The third plot seems anomalous, but ignoring this it seems that a Luclus PT factor is preferable. The distance measures Min, Max, Correlation and Cosine do not perform well, with the exception of Cosine-Luclus, which works will for anti-kt jets. Taxicab performs best overall, but there is not a great distinction.

5.0.3 Distances in Eigenspace

In order to test the best way to measure distance in Eigenspace there are two simple comparisons that could be made to a measured distance.

The first would be the ROC-AUC (Receiver Operating Characteristic - Area Under Curve), for the distance measure between two particles and the particles being descendent from the same b-quark. This can be thought of as a measure of how good that distance measure is at separating quarks from the same b-quark from other quark pairs.

The second would be the Spearman's rank of the measured distance against the Shower distance. Shower distance is defined by taking the simulated shower as a graph, where nodes are particles and vertices are interactions, and finding the least number of vertices that must be transvered in order to get between the particles in the pair.

These the first measure is more closely related to our goals, the second measure has a simpler physical interpretation. All particles from the same b-quark have a small shower distance but not all particles with a small shower distance are from the same b-quark. This can be seen in figure **??**. For a visualisation of these comparisons in the first event see figure **??**.

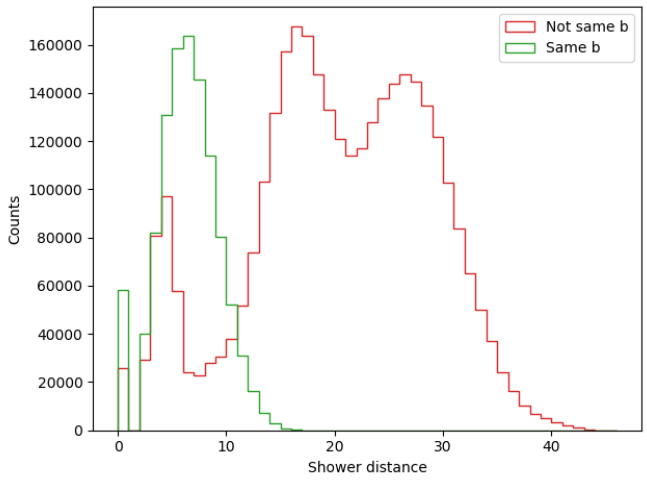

Figure 9: Comparison of shower distance (see sec 5.0.3) for pairs that are in the same b-quark and those that are not.

There are 10 distance measures that were tried in Eigenspace.

- Euclidean; $d = \sqrt{\sum_i \delta x_i^2}$
- L3; $d = \left(\sum_i \delta x_i^3\right)^{1/3}$
- L4; $d = \left(\sum_i \delta x_i^4\right)^{1/4}$
- Taxicab; $d = \sum_i |\delta x_i|$
- Braycutis; $d = \sum_i |x_{1,i} - x_{2,i}| / \sum_i |x_{1,i} + x_{2,i}|$
- Canberra; $d = \sum_i |x_{1,i} - x_{2,i}| / \sum_i |x_{1,i}| + |x_{2,i}|$
- Min; $d = \min_i |\delta x_i|$
- Max; $d = \max_i |\delta x_i|$
- Correlation; $d = 1 - \frac{\sum_i (x_{1,i} - \bar{x}_{1,i})(x_{2,i} - \bar{x}_{2,i})}{\sqrt{\sum_i (x_{1,i} - \bar{x}_{1,i})^2} \sqrt{\sum_i (x_{2,i} - \bar{x}_{2,i})^2}}$
- Cosine; $d = 1 - \frac{\sum_i x_{1,i} x_{2,i}}{\sqrt{\sum_i x_{1,i}^2} \sqrt{\sum_i x_{2,i}^2}}$

And each of these was reattempted after the eigenspace had been "normed" by dividing each eigenvector by it's eigenvalue. This creates 20 possible ways to measure distance in Eigenspace.

A comparison of the two scoring criteria over two thousand events can be seen in figure **??**.

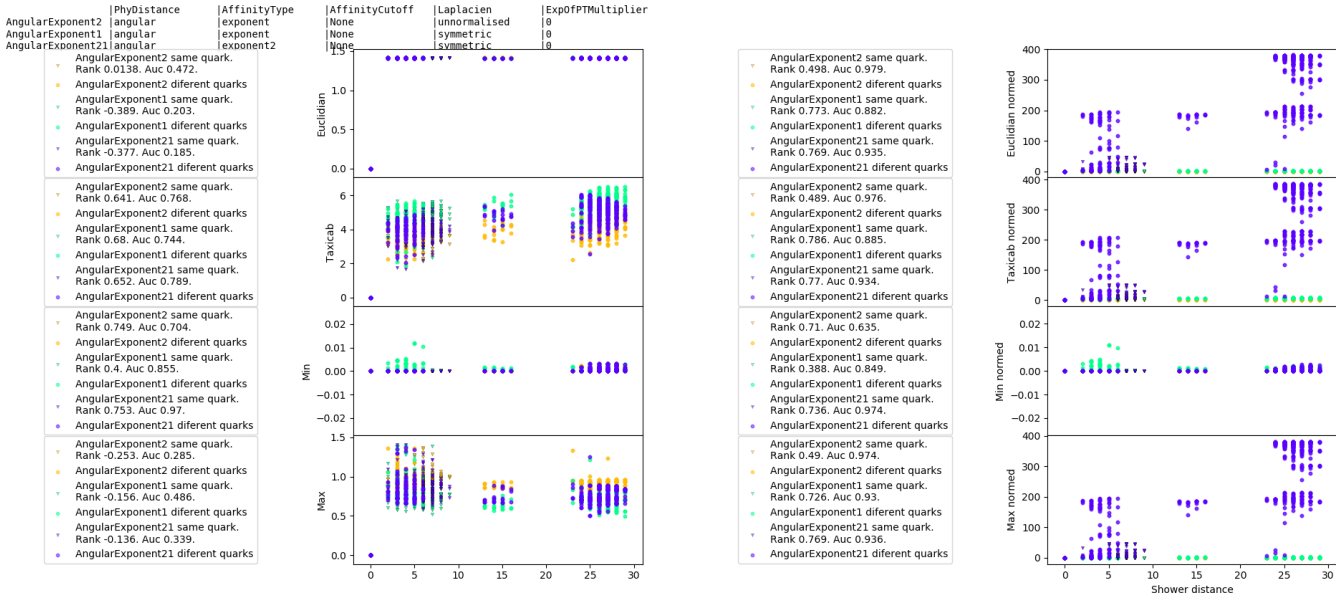

Figure 10: Each plot is a comparison of shower distance (see sec 5.0.3) against a distance measure applied in the Eigenspace created at the first stage of jet formation. Each point is on a plot represents the relationship between a pair of particles that are used to form jets in the first event.

	PhyDistance	AffinityType	AffinityCutoff	Laplacien	ExpOfPTMultiplier
AngularExponent1	angular	exponent	None	symmetric	0
AngularExponent21	angular	exponent2	None	symmetric	0
AngularExponent22	angular	exponent2	None	unnormalised	0
AngularExponent23	angular	exponent2	('distance', 2)	unnormalised	0
AngularExponent24	angular	exponent2	('distance', 2)	symmetric	0
AngularExponent2	angular	exponent	None	unnormalised	0
AngularExponent3	angular	exponent	('distance', 2)	unnormalised	0
AngularExponent4	angular	exponent	('distance', 2)	symmetric	0
LuclusExponent1	Luclus	exponent	None	symmetric	0
LuclusExponent21	Luclus	exponent2	None	symmetric	0
LuclusExponent22	Luclus	exponent2	None	unnormalised	0
LuclusExponent23	Luclus	exponent2	('distance', 2)	unnormalised	0
LuclusExponent24	Luclus	exponent2	('distance', 2)	symmetric	0
LuclusExponent2	Luclus	exponent	None	unnormalised	0
LuclusExponent3	Luclus	exponent	('distance', 2)	unnormalised	0
LuclusExponent4	Luclus	exponent	('distance', 2)	symmetric	0

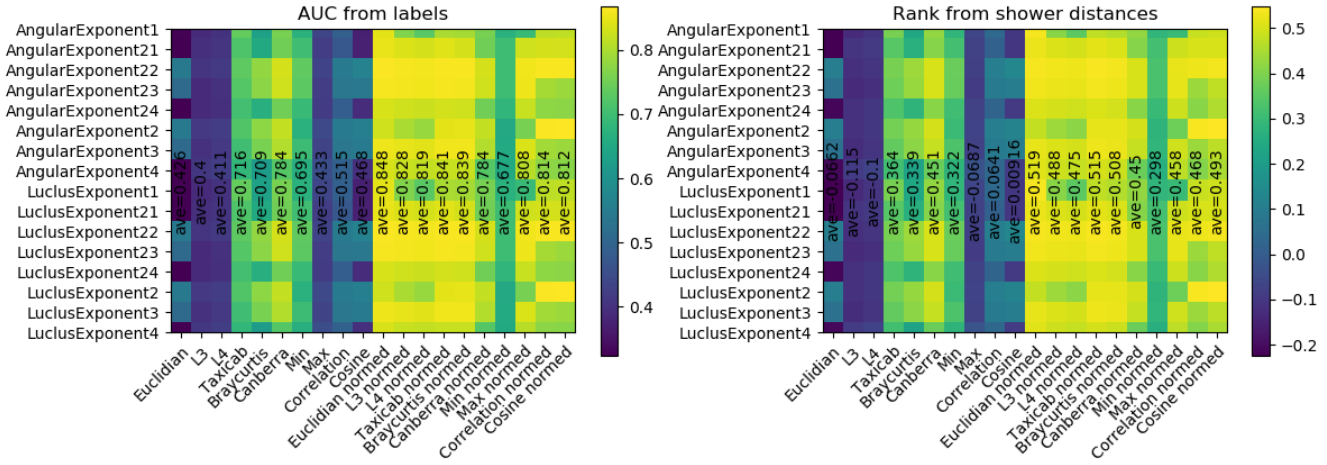

Figure 11: The plots are heatmaps with various jet clustering approaches on the y axis and different ways to measure the distance in eigenspace on the x axis (see sec 5.0.3). The left and plot is the ROC-AUC (measure closet to our aim) the right hand plot is the shower distance (measure easiest to interpret physically). Both plots are in strong agreement. It is clearly important to norm the eigenstate, and either Euclidean or Taxicab distance are equally acceptable with Euclidean results being marginally stronger.

6 Method

6.1 Particle data

Describe the production of the input data The dataset used for the majority of this work is a simulated Higgs cascade decay. One Standard Model (SM) Higgs at 125GeV decays to two light higgs at 40GeV, which in turn decay to b-$\bar{b}$ quark pairs. That is $H_{125\text{GeV}} \to h_{40\text{GeV}} h_{40\text{GeV}} \to b\bar{b}b\bar{b}$.

This dataset has the desirable property of creating b-jets with a range of geometries owing to the boost provided by large mass of the SM Higgs and the high chance of overlap with 4 b-quarks in the event.

Other radiation from the protons is also included.

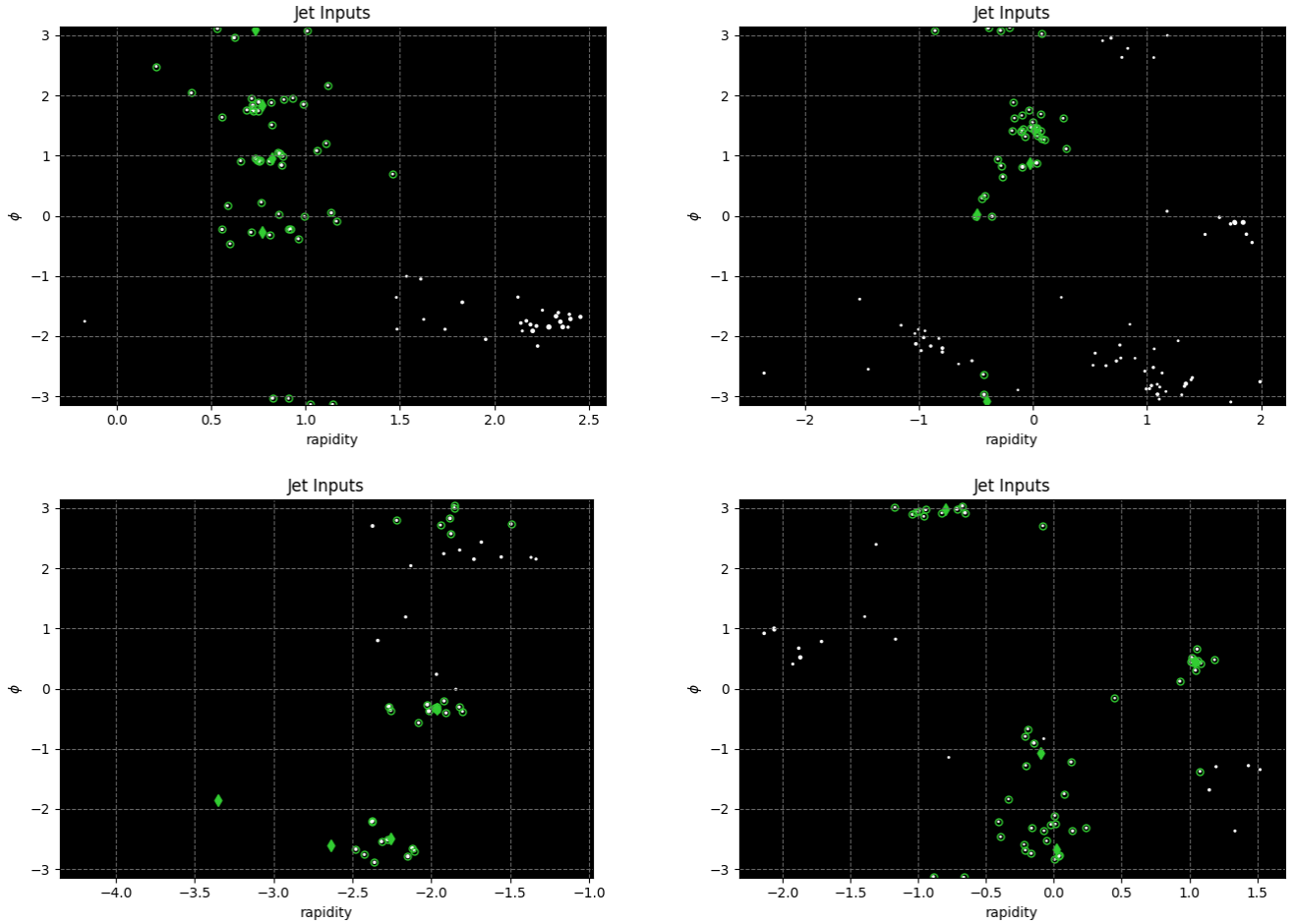

Figure 12: Some examples of events after cuts. Each white dot is a particle that passes the cuts and can be used for jet clustering. A green diamond indicates the location of the b-quark (not an input to clustering). A green circle indicates a descendant of the b-quark.

Before any evaluation can be performed on the final state of the simulation particles that ended outside the range of the silicon tracker ($|\eta| > 2.5$) or particles with low transverse momentum ($p_T < 0.5$ GeV) are cut. This is to mimic restrictions from reconstruction accuracy.

After cuts, 72% of events have at least 5 b-descendants and 5 non b-descendants available.

6.2 Clustering algorithm

Describe the spectral algorithm in detail Now the implementation of the theory described in section 5 will be specified. For every simulated event this process if used to select a clustering.

1. The particles are to be used to form the nodes of a graph, the edges of which will be weighted by some measure of proximity between the particles known as affinity. To obtain an affinity, first a distance is obtained; $d_{i,j} = \sqrt{(y_i - y_j)^2 + (\phi_i - \phi_j)^2}$ where y_i is the rapidity of particle i and ϕ_i is the barrel angle is particle i.

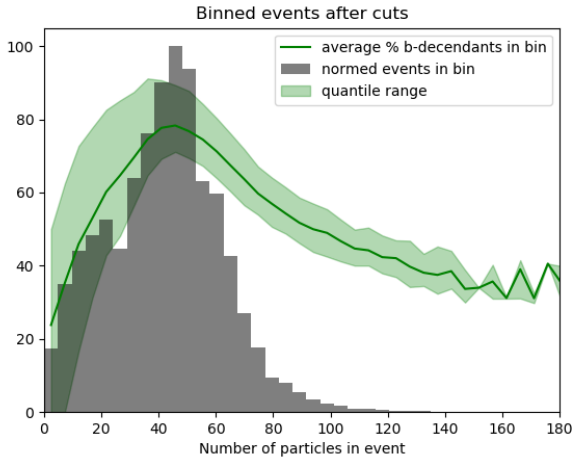

Figure 13: The end state particle in the simulated events are filtered with the standard cuts, $p_T > 0.5$, $|\eta| < 2.5$. The events are binned according to how many particles remain after the cuts. The percentage of b-descendants in the remaining event after the cuts is averaged for each bin and plotted on the same axis. After the cuts have been applied most events are left with around 50 particles. The percentage of particles that are descendant from a b-quark varies, it is at it's highest in events with 50 particles, and most variable in events will small multiplicity.

2. The distance shrinks as particles become similar, to obtain an affinity this must be transformed so that the value grows with increasing similarity. This is done by taking an inverse so that $a_{i,j} = 1/d_{i,j}$, as done in [14].

3. These affinities allow the construction of a Laplacien. The Laplacien used is the unnormalised Laplacien, which has $\sum_j a_{i,j}$ in the ith diagonal entry (also known as the degree of node i) and $-a_{i,j}$ of the diagonal in column i row j.

4. From the Laplacien a predetermined number of eigenvectors are calculated to create the embedding space. The eigenvectors have as many elements as there are particles, and the coordinates of the ith particle in the embedding space is the ith element of each eigenvector.

5. The first clustering can be done based on a measure of distance derived from the particles p_T and it's position in the embedding space. The p_T is used as in the Luclus [17], with a variable exponent, q;

$$p_T \text{ factor} = \left(\frac{p_{T\,i} p_{T\,j}}{s(p_{T\,i} + p_{T\,j})} \right)^q$$

. Where s is the invariant mass of all observed particles in the event, it is used to make the p_T factor unitless. The embedded distance is euclidean distance in the embedding space multiplied by this factor.

6. The two object that have the smallest embedding distance are combined. In physical space the combined object is created by summing the respective four momenta, in the embedding space two methods for locating the combined object are tried.

   (a) In a `SpectralMeanJet` clustering the location of the combined object is the geometric mean of the inputs. The clustering then continues to combine things in this manner.

   (b) In a `SpectralFullJet` once two object have been combined in physical space the embedding space is recalculate from step 1.

7. When the closest object to a particle in the embedding space is further away than ΔR then the combined object is considered to be a complete jet and removed from future clustering.

To provide a basis for comparison the results of clustering with an anti-kt algorithm (as in [9]) is also shown.

7 Results

Describe the evaluation process and show plots Three choices of clustering algorithm are considered here;

- A standard anti-KT algorithm with $\Delta R = 0.6$
- A `SpectralMeanJet` algorithm with an inverse affinity $\Delta R = 0.033$, and $q = -0.34$subject to change
- A `SpectralFullJet` algorithm with an linear affinity $\Delta R = 0.037$, and $q = -0.71$subject to change

As the data is simulated it is possible to compare the performance of clustering algorithms to Monte Carlo truth. Each event contains 4 b-quarks and for each of them it is possible to identify the particle into which they decayed, as a subset of the particles in the final state. Henceforth the detectable decay products of the b-quark will be called the descendants of the b-quark. For two reasons it is not possible for this clustering algorithm to gather all the descendants of each b-quark into one jet: firstly not all the descendants make the p_T and η cuts, so some are discarded before clustering; secondly the descendants of the b-quarks in an event are not mutually exclusive, due to interactions during hadronisation the quarks share descendants, and our clustering algorithm does produce exclusive clusters.

Knowing the parts of the final state that are descended from each b-quark creates a clear allocation of jets to quarks. For each quark, the jet that contains the greatest mass in descendent particles is tagged to represent that quark.

Mass peaks can the be constructed from the tagged jets, using all the particles in the jet, both descendant of the quarks and background. In figure 14 the masses of all jets that have been b-tagged are plotted. In figure 15 three selections are plotted; firstly only events where some trace of all 4 b-quarks is found are plotted with the mass of the descendants of the heavy Higgs in the background. Then the two light Higgs in each event are sorted by the mass of their descendants, in effect they are ranked by how well they were picked up by the detector. The mass of the light descendants is plotting in the background and over that the mass of the associated jets in each event is shown.

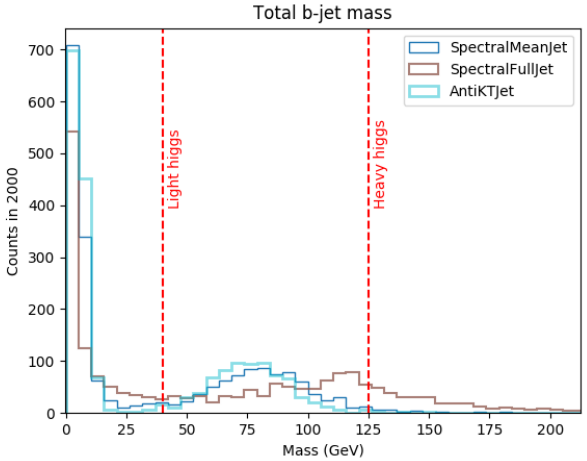

Figure 14: All tagged jets in each event are plotted. A peak just shy of 125GeV is desired as this is the mass of the decaying heavy Higgs. Another peak at 40GeV could be observed is only one light Higgs was found in the detector volume, `SpectralFullJet` gets closest, peaking close to 115 GeV and showing some contamination from background with the bins > 125 GeV. Anti-KT and `SpectralMeanJet` both give similar behaviour, falling somewhat short. The Anti-kt jet used a ΔR of 0.63.

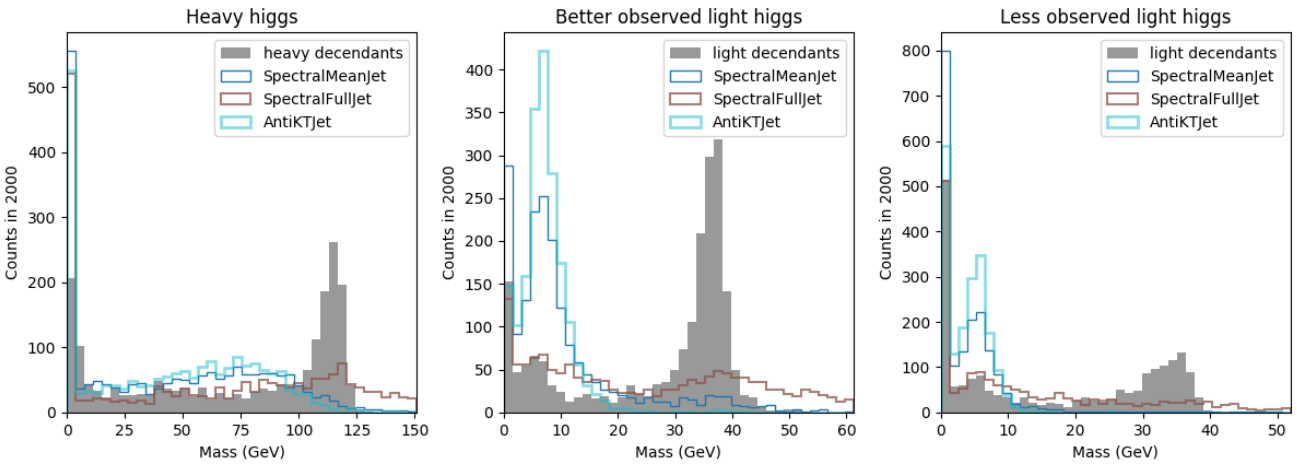

Figure 15: Starting with the plot on the far left, the jet mass of events where some descendants from all 4 b-quarks were found is plotted over the total mass of all descendants. The plot in the centre takes the light Higgs whose descendants have the greatest mass in each event. The mass of these descendants is plotted in grey, and over this the masses of the jets in each event that correspond to this better observed Higgs are shown. That is, the tags in each event that have been produced by the better observed light Higgs are allocated to jets, and only the mass of these jets is tabulated - thus a good cluster will have obtained the mass of the light Higgs descendants. Finally on the left the light Higgs who's descendants have less mass is shown along with the corresponding jets. These three plots make it clear that only `SpectralFullJet` is obtaining the majority of the descendants, but it is combining them with some background as the tail of the `SpectralFullJet` distribution continues past the descendants distribution out of the frame. The behaviour of `SpectralMeanJet` does not significantly differ from Anti-KT

8 Conclusions

Compare spectral clustering to standard clustering.

8.1 Recursive Neural Tensor Networks

There are many possible forms for a neural network that classifies jets;

- A feed forward Deep Neural Network (DNN). This is simple to train but requires a fixed length input. Obtaining a fixed length input is a challenge because the number of tracks in a jet is not fixed, expert features must be used, and often some form of interpolation or padding is required. This doesn't make a DNN a very natural fit for the problem.
- A Convolutional Neural Network (CNN). This can be run on an image generated by the calorimeter output. It has the advantage of interpretability because the filters it developed to scan the image can be analysed and some of it's process can be estimated. Removing symmetries from the images, however, is difficult to achieve without distorting their physical content. Rotations in particular are frequently done improperly, altering the jet mass. Further, the images are sparse, and a CNN focuses on local correlations so cannot perform optimally on spares images.
- A Recurrent Neural Network (RNN). This can read in a sequence, so if we could order the tracks it would be able to used all there data. Finding a natural ordering for the tracks is difficult, however. Previously p_T has been used, the Lund Plane is certainly a natural ordering, but it doesn't capture all the tracks.
- A Recurrent Neural Tensor Network (RTRN). RTRNs follow a tree shape, being originally developed for parsing syntax trees. This happens to coincide exactly with the shape of a jet clustering algorithm, so we could apply them to this.

To begin with the starting point considered was DeepCSV. It is a DNN, so it's foible is a fixed length input vector. As each input is one jet long, two types of variable are considered, variables that describe an aspect of the jet as a whole (p_T of the jet for instance) or variables that describe an aspect of a single track (such as p_T of that track). The first kind have a fixed length and provided all values are successfully reconstructed (which is not always true) will require no 'padding'. Padding is the manner in which missing values are filled. The second kind, variables

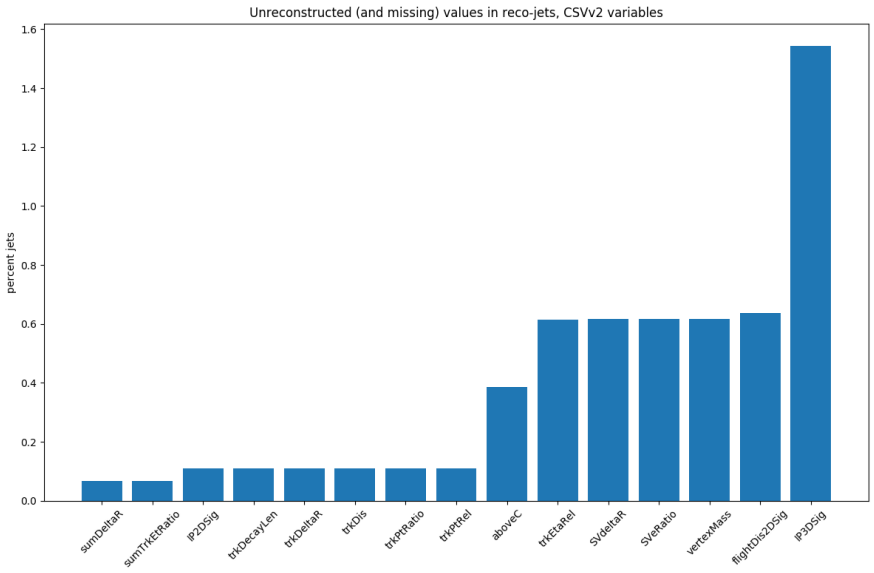

Figure 16: Both DeepCSV and CSVv2 used zero padding to fill in values that were not available from the data. In my data sample the percentage of each variable that must be zero padded for CSVv2 is shown here.

per track, will not have a fixed length even if the reconstruction is perfect because the number of tracks in a jet varies. Unless the plan is to train a separate DNN for each size of jet (which would be statistically wasteful and computationally expensive) some form of padding will have to be chosen.

The amount of padding required for a restricted set of tracks in my dataset is shown in figure 16.

The padding decision taken in [21] was called zero padding, but as the distributions are all centred on zero before padding this is equivalent to padding with the average value of each distribution. Thus we can imagine that we are inserting "average" tracks into each event until there are sufficient tracks to fill the fixed length input. This make very little physical sense; the momentum of the additional tracks will contradict the momentum of the jet, and the angle of the tracks will have no relation to the angle of the jet at all.

This was deemed to be an undesirable approach, so new methods that didn't require fixed length input were considered.

8.1.1 Picture from the Monte Carlo

In order to better envision a good neural network some time was spend graphing the behaviour of a selection of Monte Carlo showers and their jets. A shower from a heavy Higgs decay can be seen in figure 17. Here showers are considered to start with the particles that leave the hard event, and any child of the proton beam besides the hard interaction. These are dubbed originating particles. Any descendant of an originating particle is in the originating particle's shower. Due to the requirements of colour confinement, if any originating particle is colour charged it must share descendants with another shower so that the final descendants are colour neutral. This happens in hadronisation.

There were a few casual observations made at this point. In most events, the graph of all showers is fully connected, that is to say that the hadronisation links all showers to all other showers in most cases. Information in one part of an event might be expected to be strongly dependant on information in the rest of the event. This favours dealing with the event as a whole instead of jet by jet.

It also indicates that a non exclusive jet formula might make more sense - if tracks can belong to multiple showers, perhaps they should be able to belong to multiple jets. This overlap between showers is shown for one event in figure 18. One promising feature of this plot is that shared descendants between soft gluon showers and hard b quark showers is minimal, so exclusive jets will merge b showers, but it looks possible to avoid excessive merging of b quark showers and soft showers.

Another picture of the distinction between the signal and the soft jets can be seen in figure 19. It can be seen that the hard showers fall into two jets in the bottom right of the heatmap, and the soft showers are spread over many jets. This is strikingly different behaviour and one might imagine that the structure of the jet is sensitive to the nature of the shower it is built on.

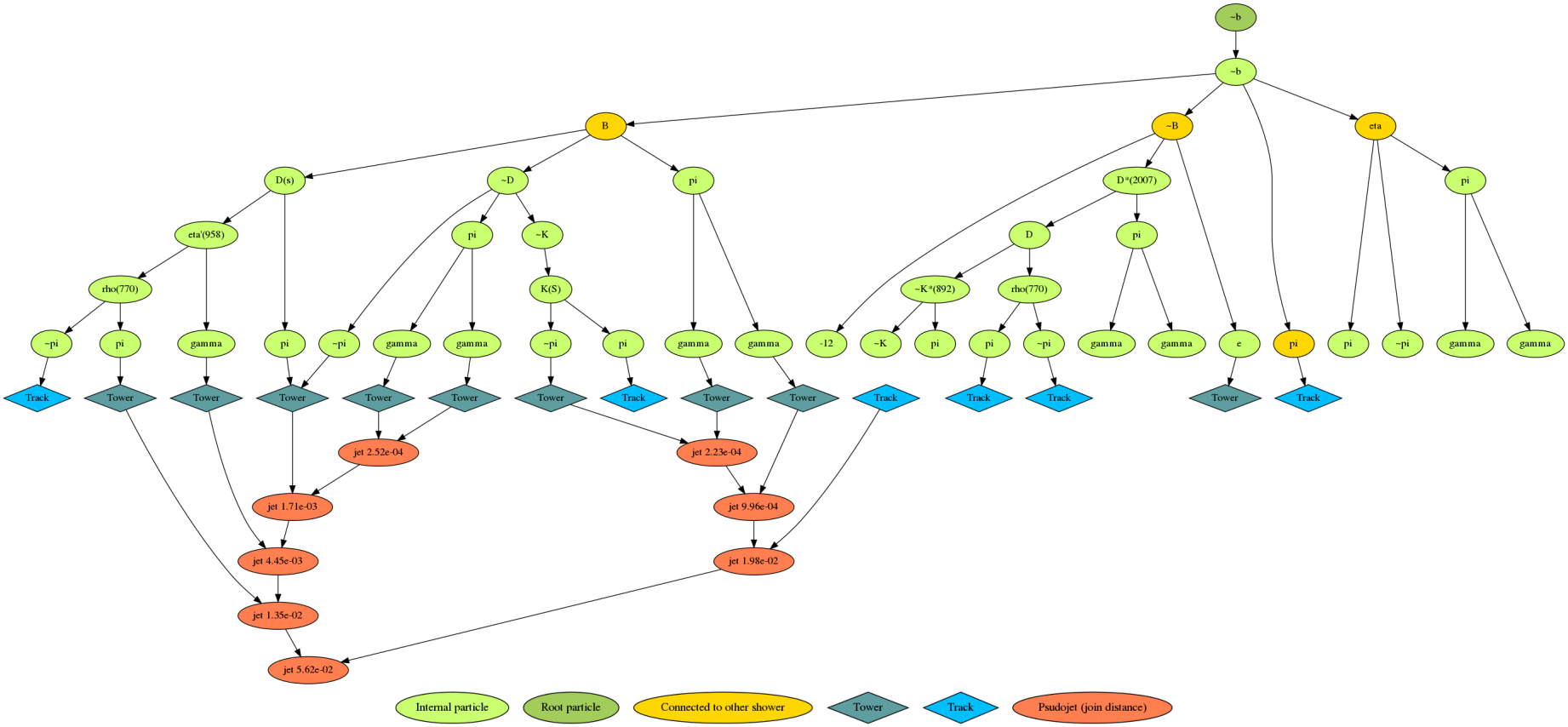

Figure 17: A jet attempts to capture the observables let by one shower. Here we see an example shower generated in Monte Carlo. It was generated with aid of Madgraph [3], Pythia [22] and Delphes [23]. The shower, and it's Monte Carlo truth is shown at the top, at the bottom the process of the jet clustering algorithm is seen. The jet clustering algorithm captures most but not all of the shower, and it captures some tracks from other parts of the event.

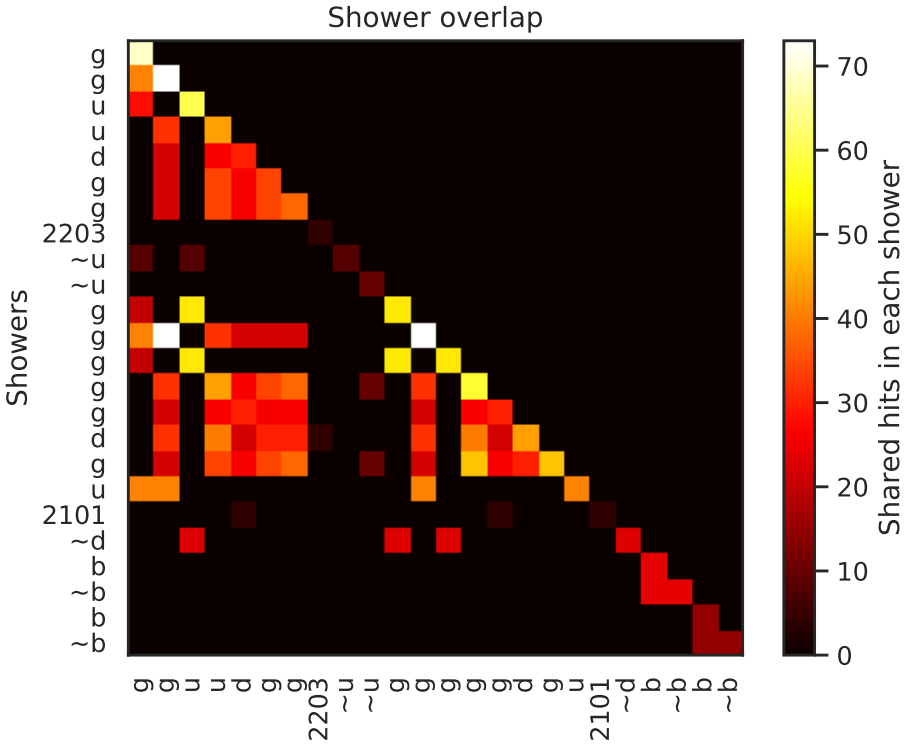

Figure 18: The overlap between showers in one event. A shower is defined here as all the descendants of a particle who's parents are of the hard interaction or the proton beams. The axis labels identify the originating particle of the shower. Particles in one shower will interact with particles in another shower and then produce common descendants. For a shower from a colour charged this is required to form colour neutral end products.

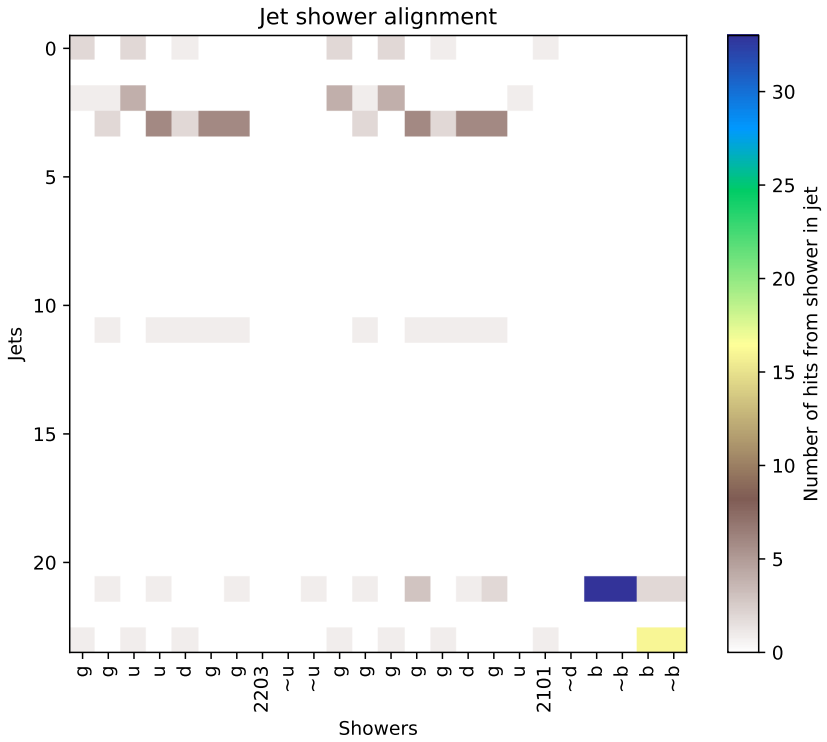

Figure 19: The alignment between showers and jets for one event. The shower axis identifies the originating particle of the shower, the jets themselves are ordered to make the plot as close to diagonal as possible. Ideally each jet would capture exactly one shower if this occurred the plot would be diagonal. What is seen in this event is that many shower have be incorrectly split between many jets.

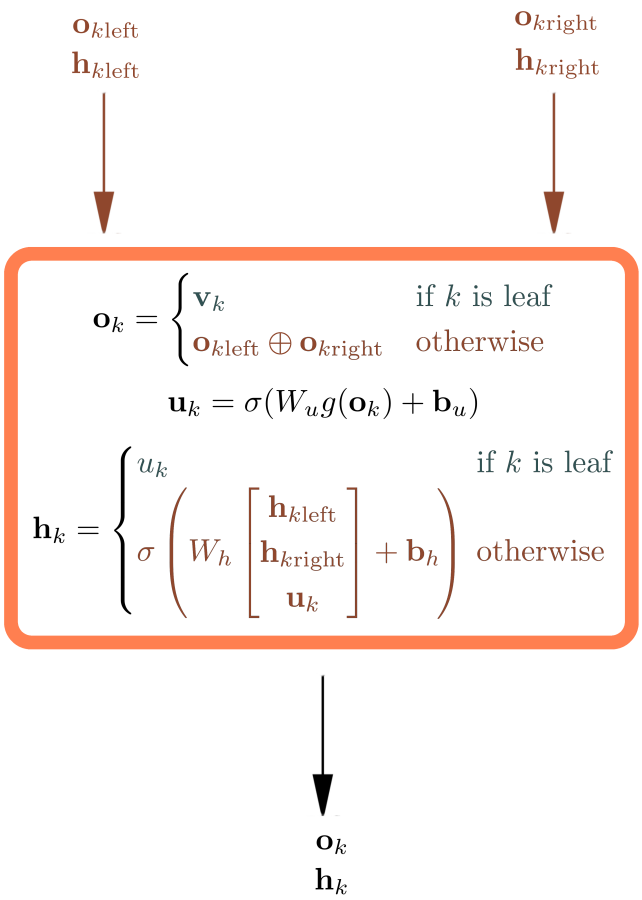

Figure 20: A single node of an RNTN. The vector $\mathbf{o}_k$ is the vector of observable properties of the k^{th} node. If the k^{th} node is a leaf then these would be the track properties, otherwise they are a combination of the properties of the tracks below k. The vector $\mathbf{u}_k$ is the embedding of $\mathbf{o}_k$ into the latent space of the net. The vector $\mathbf{k}_k$ is the 'state' of the k^{th} node, structural information about the tree propagates in the state of the nodes.

It can be seen in figure 17 that the jet structure is not a mirror of the shower. So the relationship between jet structure and physics is not immediately obvious, but perhaps it would be a useful structure for a classifier to combine information. This is the inspiration for the plan to use RNTNs.

8.2 Architectural direction

The network used by [10] was originally developed for parsing syntax trees. It the case of jet clustering it generates a state vector ($\mathbf{h}_k$) for every pseudojet. That state vector encodes the useful information about the pseudojet structure and kinematics.

The equations involved can be seen in figure 20. Qualitatively the first pseudojets being the tracks or towers must make a state vector out of the observables alone. These are known as leaf nodes in the net, and they learn a process for encoding observables into a new hidden state. When two pseudojets merge into a new pseudojet there are three kinds of input to the state vector; the left and right state vectors of the child nodes and the combined kinematics of the pseudojet. A separate procedure is learnt to combine these three things. These two processes combine all the tracks into one root state vector.

The root state vector, or final state vector, will be passed on to a DNN, along with global event information which tags the jet.

8.2.1 Subsequent steps

Once this net has been replicated and seen to perform acceptably there are a number of forward direction open;

1. Most RNNs and RNTNs have a target available at every node. This greatly improves the stability of the

training process. A physically inspired intermediate target might be identified of increase the stability of the net as it trains.

2. RNTNs have been implemented with Long Short Term Memory (LSTM) and jet taggers with LSTM have been implemented [12]. To my knowledge, however, they have yet to be combined. This would be a very natural extension.

3. Such a net will be sensitive to the subtleties of hadronisation, there would be a number of possibility's for testing their robsutness to Monte Calro errors. Unsupervised training on known signal/background regions in data could be compared to equivalent MC trained nets. Different MC generators could be compared.

4. It would be interesting to find an interpretation of the net's latent space. Perhaps t-SNE would be suitable for this.

Points 1 and 2 should certainly be undertaken, if the structure show promise then points 3 and 4 would be worth attempting.

References

[1] S. Agostinelli et al. "Geant4 a simulation toolkit". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8. URL: http://www.sciencedirect.com/science/article/pii/S0168900203013688.

[2] Simone Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *Journal of High Energy Physics* 2010.6 (June 2010), p. 43. ISSN: 1029-8479. DOI: 10.1007/JHEP06(2010)043. URL: https://doi.org/10.1007/JHEP06(2010)043.

[3] Johan Alwall et al. "MadGraph 5: going beyond". In: *Journal of High Energy Physics* 2011.6 (June 2011), p. 128. ISSN: 1029-8479. DOI: 10.1007/JHEP06(2011)128. URL: https://doi.org/10.1007/JHEP06(2011)128.

[4] Markus Stoye and. "Deep learning in jet reconstruction at CMS". In: *Journal of Physics: Conference Series* 1085 (Sept. 2018), p. 042029. DOI: 10.1088/1742-6596/1085/4/042029. URL: https://doi.org/10.1088%2F1742-6596%2F1085%2F4%2F042029.

[5] Florian Beaudette. "The CMS Particle Flow Algorithm". In: *Proceedings, International Conference on Calorimetry for the High Energy Frontier (CHEF 2013): Paris, France, April 22-25, 2013*. 2013, pp. 295–304. arXiv: 1401.8155 [hep-ex].

[6] Rachid Benbrik et al. *Mapping $pp \to A \to ZH \to l^+l^-b\bar{b}$ and $pp \to H \to ZA \to l^+l^-b\bar{b}$ Current and Future Searches onto 2HDM Parameter Spaces*. 2020. arXiv: 2006.05177 [hep-ph].

[7] G.C. Branco et al. "Theory and phenomenology of two-Higgs-doublet models". In: *Physics Reports* 516.1-2 (July 2012), pp. 1–102. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2012.02.002. URL: http://dx.doi.org/10.1016/j.physrep.2012.02.002.

[8] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "FastJet User Manual". In: *Eur. Phys. J. C* 72 (2012), p. 1896. DOI: 10.1140/epjc/s10052-012-1896-2. arXiv: 1111.6097 [hep-ph].

[9] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[10] Taoli Cheng. "Recursive Neural Networks in Quark/Gluon Tagging". In: *Computing and Software for Big Science* 2.1 (Nov. 2018). INSPEC:18015611, 3 (13 pp.) ISSN: 2510-2036. DOI: 10.1007/s41781-018-0007-y.

[11] A. Djouadi, J. Kalinowski, and P.M. Zerwas. "Two and three-body decay modes of SUSY Higgs particles". In: *Z. Phys. C* 70 (1996), pp. 435–448. DOI: 10.1007/s002880050121. arXiv: hep-ph/9511342.

[12] Shannon Egan et al. "Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC". In: (2017). arXiv: 1711.09059 [hep-ex].

[13] Stephen D. Ellis and Davison E. Soper. "Successive combination jet algorithm for hadron collisions". In: *Phys. Rev.* D48 (1993), pp. 3160–3166. DOI: 10.1103/PhysRevD.48.3160. arXiv: hep-ph/9305266 [hep-ph].

[14] Alireza Hadjighasem et al. "Spectral-clustering approach to Lagrangian vortex detection". In: *Phys. Rev. E* 93 (6 June 2016), p. 063107. DOI: 10.1103/PhysRevE.93.063107. URL: https://link.aps.org/doi/10.1103/PhysRevE.93.063107.

[15] Ulrike von Luxburg. *A Tutorial on Spectral Clustering*. 2007. arXiv: 0711.0189 [cs.DS].

[16] S. Moretti and W.James Stirling. "Contributions of below threshold decays to MSSM Higgs branching ratios". In: *Phys. Lett. B* 347 (1995). [Erratum: Phys.Lett.B 366, 451 (1996)], pp. 291–299. DOI: 10.1016/0370-2693(95)00088-3. arXiv: hep-ph/9412209.

[17] Stefano Moretti, Leif Lönnblad, and Torbjörn Sjöstrand. "New and old jet clustering algorithms for electron-positron events". In: *Journal of High Energy Physics* 1998.08 (1998), p. 001.

[18] Gavin P. Salam. "Towards jetography". In: *The European Physical Journal C* 67.3-4 (May 2010), pp. 637–686. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-010-1314-6. URL: http://dx.doi.org/10.1140/epjc/s10052-010-1314-6.

[19] R. J. Sánchez-García et al. "Hierarchical Spectral Clustering of Power Grids". In: *IEEE Transactions on Power Systems* 29.5 (2014), pp. 2229–2237.

[20] Steven Schramm. *ATLAS Jet Reconstruction, Calibration, and Tagging of Lorentz-boosted Objects*. Tech. rep. ATL-PHYS-PROC-2017-236. Geneva: CERN, Nov. 2017. URL: https://cds.cern.ch/record/2291608.

[21] A. M. Sirunyan et al. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *JINST* 13.05 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158 [physics.ins-det].

[22] Torbjorn Sjostrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (2015), pp. 159–177. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2015.01.024. URL: http://www.sciencedirect.com/science/article/pii/S0010465515000442.

[23] The DELPHES 3 collaboration et al. "DELPHES 3: a modular framework for fast simulation of a generic collider experiment". In: *Journal of High Energy Physics* 2014.2 (Feb. 2014), p. 57. ISSN: 1029-8479. DOI: 10.1007/JHEP02(2014)057. URL: https://doi.org/10.1007/JHEP02(2014)057.

[24] M. Wobisch and T. Wengler. "Hadronization corrections to jet cross-sections in deep inelastic scattering". In: *Monte Carlo generators for HERA physics. Proceedings, Workshop, Hamburg, Germany, 1998-1999*. 1998, pp. 270–279. arXiv: hep-ph/9907280 [hep-ph].