Applications of Standard Model Effective Field Theory to 2D differential distributions of top pair production

Alexander Veltman
Advisor: Dr. James Keaveney

Department of Physics,
University of Cape Town

October 18, 2021

Abstract

1 Introduction

- Standard model is good
- Cannot explain some phenomena
- New frame work called SMEFT

2 LHC, ATLAS and Top Physics

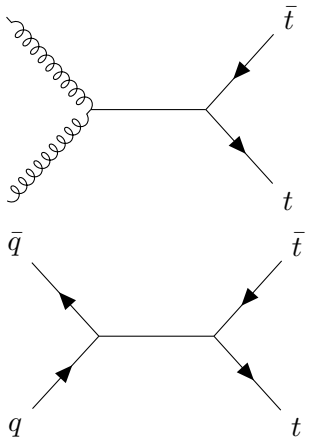

Figure 1: Tree level Feynman diagrams for top quark pair production

3 Standard Model Effective field theory

Standard Model effective field theory (SMEFT) is a model-independent framework for identifying and constraining deviations from Standard Model predictions. This is done by considering that some higher mass particles or higher energy reactions may exist and seeing the imprints on regular Standard Model cross-sections and interactions. The framework introduces a set of dimension-six terms into the Standard Model Lagrangian which only contains operators of dimension-four. These includes 59 independent operators (according to what is known as the Warsaw basis) which are built from Standard Model fields and follow the gauge symmetries of the Standard Model [1].

With these additional operators, the SMEFT Lagrangian is

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \frac{1}{\Lambda^2}\sum_i C_i O_i + \mathcal{O}\left(\frac{1}{\Lambda^3}\right) \tag{1}$$

where O_i is a dimension-six operator and C_i is an associated dimensionless coupling constant known as a *Wilson Coefficient*. The operators are reduced by the energy scale Λ of the BSM physics. These effects manifest themselves in observable cross sectional data [2] as

$$\sigma = \sigma_{\text{SM}} + \sum_i \frac{1}{\Lambda^2} C_i \sigma_i + \sum_{j,k} \frac{1}{\Lambda^4} C_j C_k \sigma_{jk} \tag{2}$$

where σ is an integrated cross section. This can also be extended to differential cross sections which is commonly obtained in high energy physics experiments. It is important to note that when $C_i = 0$ for all operators, the SMEFT Lagrangians simplifies to the SM Lagrangian. This implies that a sufficient deviation from a zero measurement may imply affects of new physics.

The effects on differential cross section with respect to an observable X are similar,

$$\frac{d\sigma}{dX} = \frac{d\sigma_{\text{SM}}}{dX} + \sum_i \frac{1}{\Lambda^2} C_i \frac{d\sigma_i}{dX} + \sum_{j,k} \frac{1}{\Lambda^4} C_j C_k \frac{d\sigma_{jk}}{dX} \tag{3}$$

in which these differentials are presented as binned measurements in data.

Using (3), the influences of SMEFT can be identified within differential cross section measurements obtained through modern collider experiments. Typically, this is done using global fits to many differential cross sectional measurements with respect to different observables. This allows different operators, which may be coupled to some observables more than others, to be more effectively constrained.

There has been interest in looking at the influences of SMEFT within the study of top quarks [2–4] due to the possibility of the top quark as an area for possible BSM physics. This report will investigate top pair production whose cross section at lowest order is only impacted by limited sets of the dimension-6 operators. For the $q\bar{q} \to t\bar{t}$ process, the only relevant operator is O_{tg}. The $gg \to t\bar{t}$ process is affected by O_{tg} as well as a set of 8 operators known as four-fermion operators. This report will only look at the four-fermion operator O_{tq}^8.

maybe add more

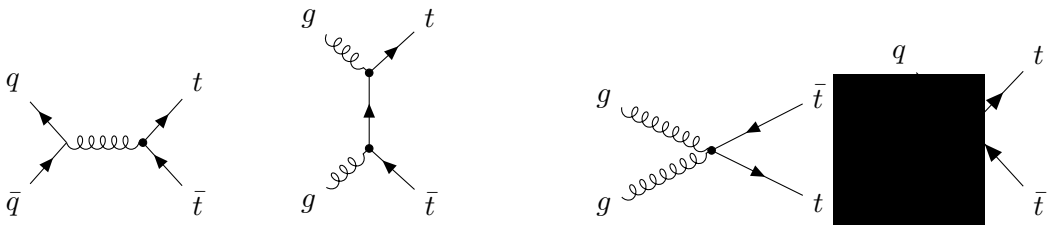

Figure 2: Examples of leading order diagrams which contribute to top pair production in SMEFT

4 dEFT

dEFT, or differential Effective Field Theory tool, is an Python package created by Dr. James Keaveney [5] to allow for predictions of SMEFT effects using differential cross section measurements. For this report, the repository was forked and further development was performed. The version used for the analysis contained in this report is available through Github [6] or the PyPI repositories [7]. Using a single configuration file containing both data and Monte Carlo predictions, dEFT can build a predictive morphing model which is used to estimate a posterior distribution for the relevant Wilson coefficients.

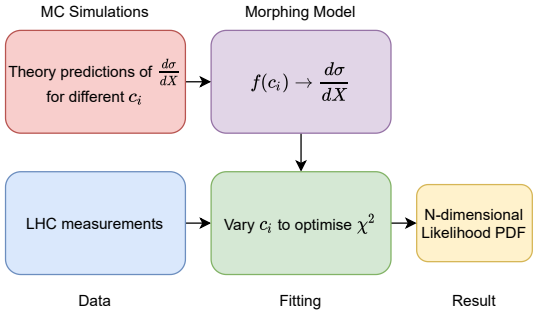

Figure 3: Overview of dEFT workflow

4.1 Model building

dEFT creates predictions for the observables of interest for varying values of Wilson coefficients by constructing a morphing model. A morphing model is a linear regression model which allows for the interpolation between different templates. These templates are Monte Carlo predictions which are generated around the region of parameter space which would be relevant for some dataset. For dEFT's application, cross sectional SMEFT predictions which are generated using a event generation framework are treated as templates. These predictions must describe the relevant set of operators O_i with varying Wilson Coefficients in order to produce a reliable model. Using (2), a linear model is constructed using these templates and produces a predictive model $\hat{\sigma}(C_i)$. This model now allows for the prediction of some cross sectional observable for any values of Wilson coefficients pertaining to the relevant set of operators. Due to the reliance on MC samples, the model is dependent on the ability to produce high quality SMEFT differential cross section predictions.

The model must be validated to ensure sensible predictions are being performed and the model predicitons can be considered reliable. Additional Monte Carlo samples are generated at coefficient values between the templates used to create the model. Theses differential cross section samples are then compared to the model predictions to ensure the predictions agree within a suitable error comparable to the statistical error on the Monte Carlo samples. Unfortunately, due to validating using Monte Carlo samples, this method of validation is still vulnerable to issues which may arise from the modelling of the samples themselves.

4.2 Fitting Method

Once a model has been built, a fit to data is possible. dEFT performs fitting using Monte Carlo Markov Chain (MCMC) methods which allow for estimation of the likelihood distributions of the Wilson Coefficients using prior assumptions about their possible values. The fitting procedure uses *emcee* [8] as its MCMC implementation. MCMC requires an estimation for the likelihood function $P(y|C_i)$ which represents the probability of obtaining some data y given a set of model parameters C_i.

A common log likelihood definition for binned data with Gaussian errors with the associated model f is

$$P(y|C_i) \propto \ln \mathcal{L}(y|C_i) = -\sum_n (y_n - f_n(C_i))V^{-1}(y_n - f_n(C_i)) \tag{4}$$

where y_n is the binned cross sectional data, V is the associated covariance matrix and $f_n(C_i)$ is the morphing model prediction. In order for an estimation for the posterior likelihood distribution $P(C_i|y)$ to be made, a prior distribution $P(C_i)$ is required. This takes the form of uniform distributions defined by some minimum and maximum for each C_i parameter. MCMC will systematically sample thoughout C_i space building an estimation for the posterior distribution $P(C_i|y)$. Properties regarding C_i can then be inferred.

Since MCMC methods are used, an approximations for the C_i distributions is obtained rather than a single value with an associated uncertainty which is common from other likelihood maximisation methods. This avoids the issue of finding a local maximisation which can be common due to the quadratic nature of the SMEFT model.

The estimation for the coefficient is extracted from the likelihood distributions by considering percentiles of the discrete MCMC sampler prediction of the marginalised distribution of each operator. The 50^{th} percentile is attributed as the estimation for the coefficient with the 16^{th} and 84^{th} percentile forming a 68% confidence interval about the estimate.

talk about Smefit methods

5 Analysis

This analysis will examine the possibility of using double differential cross section measurements with respect to multiple in a SMEFT analysis. The results will be compared to outcomes when considering a single differential cross section. The main aim of a SMEFT analysis is to place constraints onto Wilson coefficients of SMEFT operators allowing us to investigate the potential occurrences of new interactions or modifications to SM interaction. Double differential cross sections are of interest due to

the possibility of simultaneously constraining multiple operators which may present as different modifications to observable cross section distributions.

This report will use differential cross section data of top pair production from the ATLAS experiment [9] at the CERN Large Hadron Collider. This data was produced from pp collisions performed at a centre-of-mass energy $\sqrt{s} = 13\text{TeV}$ over the course of 2015 and 2016 with an integrated luminosity of 36.1fb^{-1}. The $t\bar{t}$ final states are extracted from the ℓ+jets channel in the resolved topology. This channel is characterised by the manner in which the two W bosons produced by the tops decay. This channel requires one of the W bosons to decay into a lepton and an associated anti-neutrino and the other W boson must decay into an quark-antiquark pair. The tops are then classified as decaying leptonically or hadronically by how the W decayed. Resolved topology implies that the decay products of the hadronically decaying top quark are angularly well separated.

The double differential cross section observable considered were the differential cross section as a function of the invariant mass of the $t\bar{t}$ system $m_{t\bar{t}}$ and the transverse momentum of the hadronically decaying top quark p_T^t. For the comparison with a single observable, the differential cross section as a function of just $m_{t\bar{t}}$ was examined.

talk about error on data

The only SMEFT operators considered were O_{tG} and O_{tq}^8 with corresponding Wilson coefficients C_{tG} and C_{tq}^8.

This analysis will begin with the details of the Monte Carlo event generation needed to create the morphing models for performing the constraints on the Wilson coefficients. This will move into applying this model and obtaining estimations for the distribution of the coefficients for both the single observable and the double observable.

ask james about multiple smeft operators

5.1 Monte Carlo event generation

In order to build the morphing model required to generate cross sectional predictions, simulated samples are required throughout the space of Wilson coefficients of the operators of interest. These samples were generate using the MadGraph5_aMC@NLO [10] framework which allows for the simulation of processes for a user-defined Lagrangian. The SMEFTatNLO [11] FEYNRULES model implements SMEFT tree level and one loop processes into MadGraph5. Though there is the capacity to perform predictions at next-to-leading order, these calculations are very recent and greatly increase the processing time required to produce the Monte Carlo predictions. The simulations are performed to fixed order where only the desired observables of $m_{t\bar{t}}$ and p_T^t are calculated and binned in the same binning arrangement as the ATLAS dataset of interest.

Justify why LO better

Separate sets of MC signal was required for the single observable and the double observables analysis. Events were generated with values of C_{tg} as -4, -3, -2.5, -2, -1, -0.5, 0, 0.5, 1, 2, 2.5, 3 and 4 and values of C_{tq}^8 as -4, -3, -2.5, -2, -1, -0.5, 0, 0.5, 1, 2, 2.5, 3 and 4. All permutations of these values were used resulting in a total number of 169 samples to build the model. The model validation involved generating 100 validation samples at points different to those used for the building of the model. This simulation step imposes a great challenge when wanting to expand into using more SMEFT operators in the model. The number of required MC samples to ensure consistent description of the parameter space increases exponentially with the number of operators. This is a major factor in the decision to not include all 4-fermion operators in this report and would require more time and computational capacity.

maybe rephrase this

mention this fact in conclusion

5.1.1 Uncertainty due to simulation

The simulation calculations were required to obtain a statistical accuracy of 1% for the prediction of the integrated cross section of top-antitop production. This level was considered a reliable since this error was minimal in comparison to the total uncertainty of the cross sectional data but still able to be performed in a reasonable time frame. An accuracy requirement on each bin was unavailable and would have allowed for a clearer comparison to the error associated with the data.

talk about error compared to validation value

5.1.2 k-factor determination

Since the generated Monte Carlo samples only included LO processes, theses predictions needed to be scaled to be comparable to the necessary data sets. It can be considered fairly accurate to compare NNLO predictions to actual cross sectional data so a method of scaling the current predictions to this level is required. Due to the difficulty in calculating the k-factor for different combinations of Wilson coefficients values, the k-factor for the Standard Model prediction was used across the various SMEFT predictions.

Maybe add plot comparing stat to mc error

A flat k-factor across the differential cross sections was attempted to bring the LO predictions to the scale of NNLO predictions using the proportions between tree level calculations of total cross section [10] and measurement of total cross section using the ATLAS detector [9]. This method failed due to some regions of the differential cross sections being poorly described at LO which caused a flat scaling to incorrectly describe the shape of the distributions at NNLO. This is exemplified by the low $m_{t\bar{t}}$ and high p_T^t region in our two observable data set, as seen in Figure 4. This is remedied by requiring a per-bin k-factor when scaling from LO to NLO but still using the flat factor to build up to NNLO. The per-bin k-factor was found by comparing the Standard Model predictions of MadGraph5 at LO and NLO and applying this ratio to the Monte Carlo signal. The theoretical error associated with the calculated k-factor is fairly difficult to propagate through the model construction and should be dominated by the error associated with the data. This may have consequences on bins which are not well described at LO though the scale variances in these regions should dominate the theoretical uncertainties.

ask james about explaining scale variance

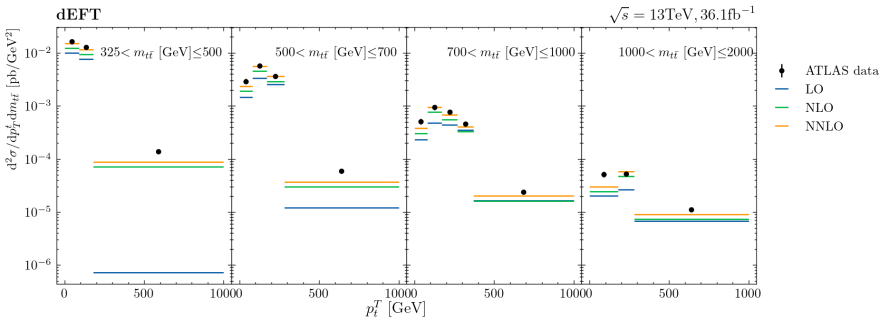

Figure 4: Comparison of approximations for processes contributing to absolute double differential $t\bar{t}$ cross section with respect to $m_{t\bar{t}}$ and p_T^t. The scaling from an LO approximation to an NLO approximation was determined using a per-bin method and the NNLO prediction was determined using a flat factor.

5.2 $m_{t\bar{t}}$ differential cross section

Probably add an intro to the data set

maybe add hep-

5.2.1 Model Validation

The results of the validation testing can be seen in Figure 5 where deviations between the model predictions and the validation samples are shown by the average relative residuals. This deviation is expected to be comparable to the 1% required accuracy for integrated cross section imposed on the event generation. These tests appears to follow an accuracy of 1.09% which is in-line with the expectation from the event generation.

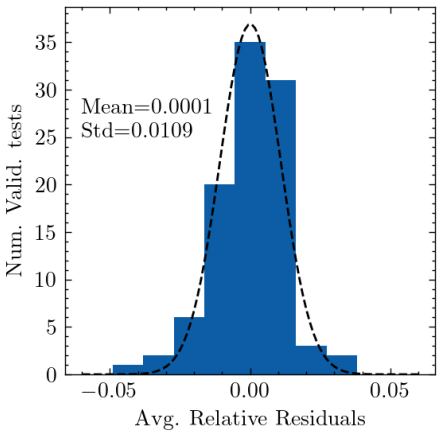

Figure 5: Distribution of average relative residuals between model predictions at some set of coefficient values and the corresponding MC validation sample for the single observable model. This includes the validation tests over 100 different validation points. The dotted line represents a Gaussian fit to the histogram through χ^2 minimisation with the results of the fit shown.

maybe include example validation plots

5.2.2 Results

The estimations for the likelihood distributions for the Wilson coefficients are shown in Figure 6. These distributions provide estimates $C_{tG} = 0.24^{+0.11}_{-0.11}$ and $C^8_{tq} = 1.56^{+0.25}_{-0.31}$ with theses estimates only considering statistical deviation of the distribution. A notable features of the distribution is the second less prominent peak. This is a product of the quadratic nature of the SMEFT model and shows that an alternative solution could still describe the data well. The major peak in the distribution appears to show a slight correlation in the coefficients.

The model created using the upper bound of the scale variance of the MC signal produced an estimation of $C_{tG} = 0.22^{+0.12}_{+0.11}$ and $C_{tq} = 1.39^{+0.26}_{-0.35}$. The lower bound of the scale variance of the MC signal produced an estimation of $C_{tG} = 0.26^{+0.11}_{0.11}$ and $C_{tq} = 1.73^{+0.24}_{-0.28}$. This implies a resulting systematic uncertainties introduced due to the scale variance of the simulated data of ± 0.02 for C_{tG} and ± 0.17 for C^8_{tq}.

The single observable analysis obtained an estimate for the Wilson coefficients of $C_{tg} = 0.24 \pm 0.11(stat.) \pm 0.02(sys.)$ and $C^8_{tq} = 1.56^{+0.25}_{-0.31}(stat.) \pm 0.17(sys.)$. The model prediction for this estimation can be seen in Figure 7. The predictions from the model at the optimised coefficients appears to better describe the data compared to using the model with all coefficients force to zero. The optimised prediction corresponded to a χ^2 per degree of freedom of 0.54 with respect to the data while the all zero prediction corresponded to a χ^2 per degree of freedom of 3.16 indicating that the optimised prediction

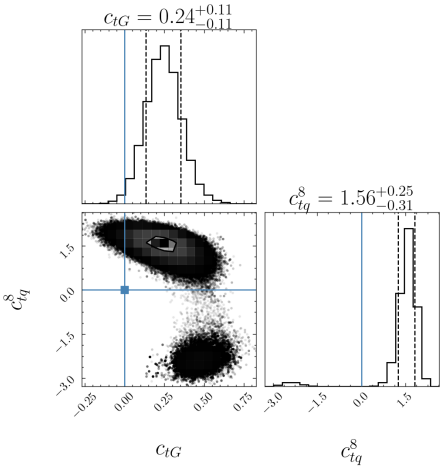

Figure 6: Estimation for the 2-dimensional likelihood distribution and 1-dimensional marginal distributions for the Wilson coefficient C_{tG} and C_{tq}^{8}. The blue lines represent SM predictions. The dotted lines on the marginal distributions represent a 68% confidence interval. Uncertainty estimates for the Wilson coefficients are the discussed in Section 4.2 and only represent statistical error.

better describes the data.

talk to jame: is it close to zero

The values obtained for the Wilson coefficients C_{tG} and C_{tq}^{8} seem to deviate from their Standard Model value of zero but still agree within 2 or 3 times the 68% confidence interval. This does provide some interest in possibly applying constraints beyond the Standard Model but there are some faults with this consideration. The systematic uncertainty associated with the model creation and prediction was not accounted for in the uncertainty estimation. Since these influences will equally be present when considering the double differential cross section data, this discussion will be deferred for later and these results will be used as a point of comparison between single and double observables analyses.

do the later

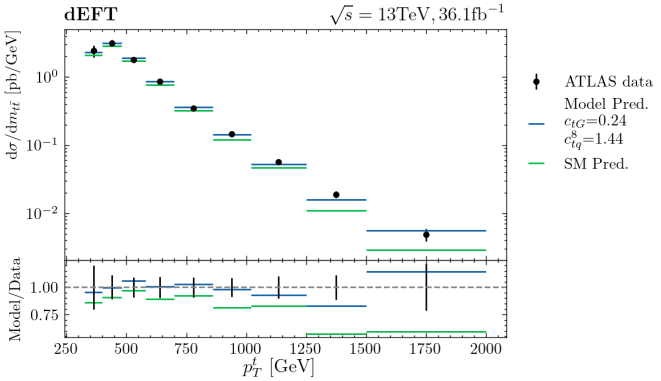

Figure 7: Comparison of differential cross section of $t\bar{t}$ production as a function of $m_{t\bar{t}}$ between various morphing model predictions and ATLAS data. This includes model predictions for the optimised Wilson coefficients and for the all zero coefficient case (labelled as SM pred.). The ratio between each model prediction and the data is shown underneath. The errors shown represent statistical and systematic uncertainties of the data.

5.3 Double differential cross section

5.3.1 Model Validation

The outcome of the validation tests of the double observable model are shown in Figure 8. The average relative residuals appear to be distributed by a normal distribution with a deviation of 1.68%. This result is larger than the validation deviation of the single obervable but is still comparable to the statistical accuracy of the MC samples.

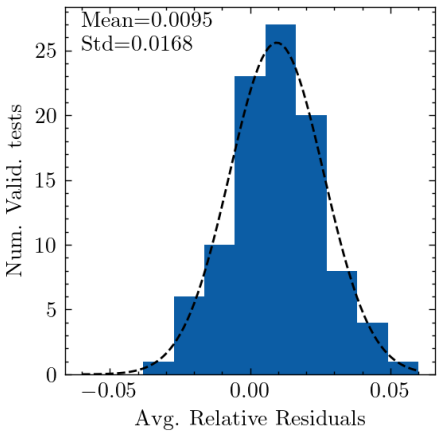

Figure 8: Distribution of average relative residuals between model predictions at some set of coefficient values and the corresponding MC validation sample for the double observable model. This includes the validation tests over 100 different validation points. The dotted line represents a Gaussian fit to the histogram through χ^2 minimisation with the results of the fit shown.

5.3.2 Results

The estimations for the likelihood distributions for the Wilson coefficients are shown in Figure 9. The distributions produced throught the MCMC method have estimates of the Wilson coefficient values of $C_{tG} = 0.49^{+0.06}_{-0.07}$ and $C^8_{tq} = -0.51^{+0.69}_{-0.70}$. The estimated systematic uncertainty associated with the scale variance of the simulated samples is $^{+0.11}_{-0.13}$ for C_{tG} and $^{+0.21}_{-0.87}$ for C^8_{tq}. The final estimation for the Wilson coefficients are $C_{tG} = 0.49^{+0.06}_{-0.07}(stat)^{+0.11}_{-0.13}(sys.)$ and $C^8_{tq} = -0.51^{+0.69}_{-0.70}(stat.)^{+0.21}_{-0.87}(sys.)$. The model prediction for this estimation can be seen in Figure 7.

maybe change to table

The shape of the distribution shows a more prominent double peak structure compared to what was seen in the single operator analysis. This makes interpreting the value of the coefficients fairly difficult.

more detail

6 Conclusion

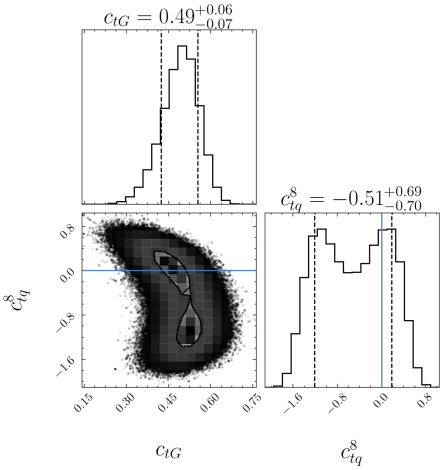

Figure 9: Estimation for the 2-dimensional likelihood distribution and 1-dimensional marginal distributions for the Wilson coefficient C_{tG} and C_{tq}^{8}. The blue lines represent SM predictions. The dotted lines on the marginal distributions represent a 68% confidence interval. Uncertainty estimates for the Wilson coefficients are the discussed in Section 4.2 and only represent statistical error.

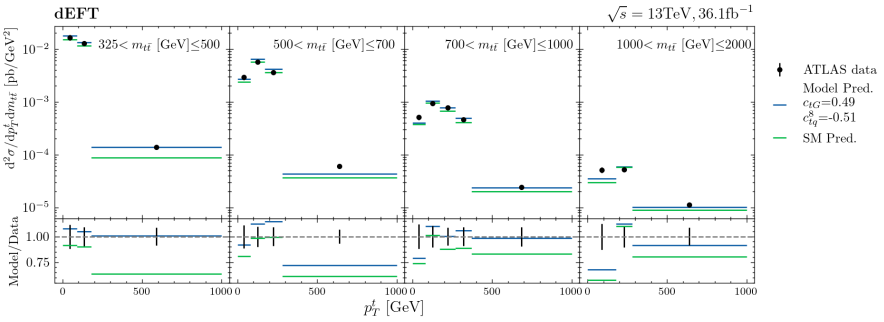

Figure 10: Comparison of differential cross section of $t\bar{t}$ production as a function of $m_{t\bar{t}}$ and p_T^t between various morphing model predictions and ATLAS data. This includes model predictions for the optimised Wilson coefficients and for the all zero coefficient case (labelled as SM pred.). The ratio between each model prediction and the data is shown underneath. The errors shown represent statistical and systematic uncertainties of the data.

References

[1] B. Grzadkowski et al. "Dimension-six terms in the Standard Model Lagrangian". In: *Journal of High Energy Physics* 2010.10 (Oct. 2010). ISSN: 1029-8479. DOI: 10.1007/jhep10(2010)085. URL: http://dx.doi.org/10.1007/JHEP10(2010)085.

[2] Nathan P. Hartland et al. "A Monte Carlo global analysis of the Standard Model Effective Field Theory: the top quark sector". In: *Journal of High Energy Physics* 2019.4 (Apr. 2019). ISSN: 1029-8479. DOI: 10.1007/jhep04(2019)100. URL: http://dx.doi.org/10.1007/JHEP04(2019)100.

[3] Andy Buckley et al. "Global fit of top quark effective theory to data". In: *Physical Review D* 92.9 (Nov. 2015). ISSN: 1550-2368. DOI: 10.1103/physrevd.92.091501. URL: http://dx.doi.org/10.1103/PhysRevD.92.091501.

[4] Ilaria Brivio et al. "O new physics, where art thou? A global search in the top sector". In: *Journal of High Energy Physics* 2020.2 (Feb. 2020). ISSN: 1029-8479. DOI: 10.1007/jhep02(2020)131. URL: http://dx.doi.org/10.1007/JHEP02(2020)131.

[5] J Keaveney. *dEFT - differential Effective Field Theory*. https://github.com/keaveney/deft. 2021.

[6] codecalec. *dEFT - differential Effective Field Theory*. https://github.com/codecalec/deft. 2021.

[7] Alexander Veltman. *deft-hep · PyPI*. https://pypi.org/project/deft-hep/. 2021.

[8] Daniel Foreman-Mackey et al. "emcee: The MCMC Hammer". In: *Publications of the Astronomical Society of the Pacific* 125.925 (Mar. 2013), pp. 306–312. ISSN: 1538-3873. DOI: 10.1086/670067. URL: http://dx.doi.org/10.1086/670067.

[9] Georges Aad et al. "Measurements of top-quark pair differential and double-differential cross-sections in the ℓ+jets channel with pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector". In: *Eur. Phys. J. C* 79.12 (2019). [Erratum: Eur.Phys.J.C 80, 1092 (2020)], p. 1028. DOI: 10.1140/epjc/s10052-019-7525-6. arXiv: 1908.07305 [hep-ex].

[10] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *Journal of High Energy Physics* 2014.7 (July 2014). ISSN: 1029-8479. DOI: 10.1007/jhep07(2014)079. URL: http://dx.doi.org/10.1007/JHEP07(2014)079.

[11] Céline Degrande et al. *Automated one-loop computations in the SMEFT*. 2020. arXiv: 2008.11743 [hep-ph].