

## Chapter 10

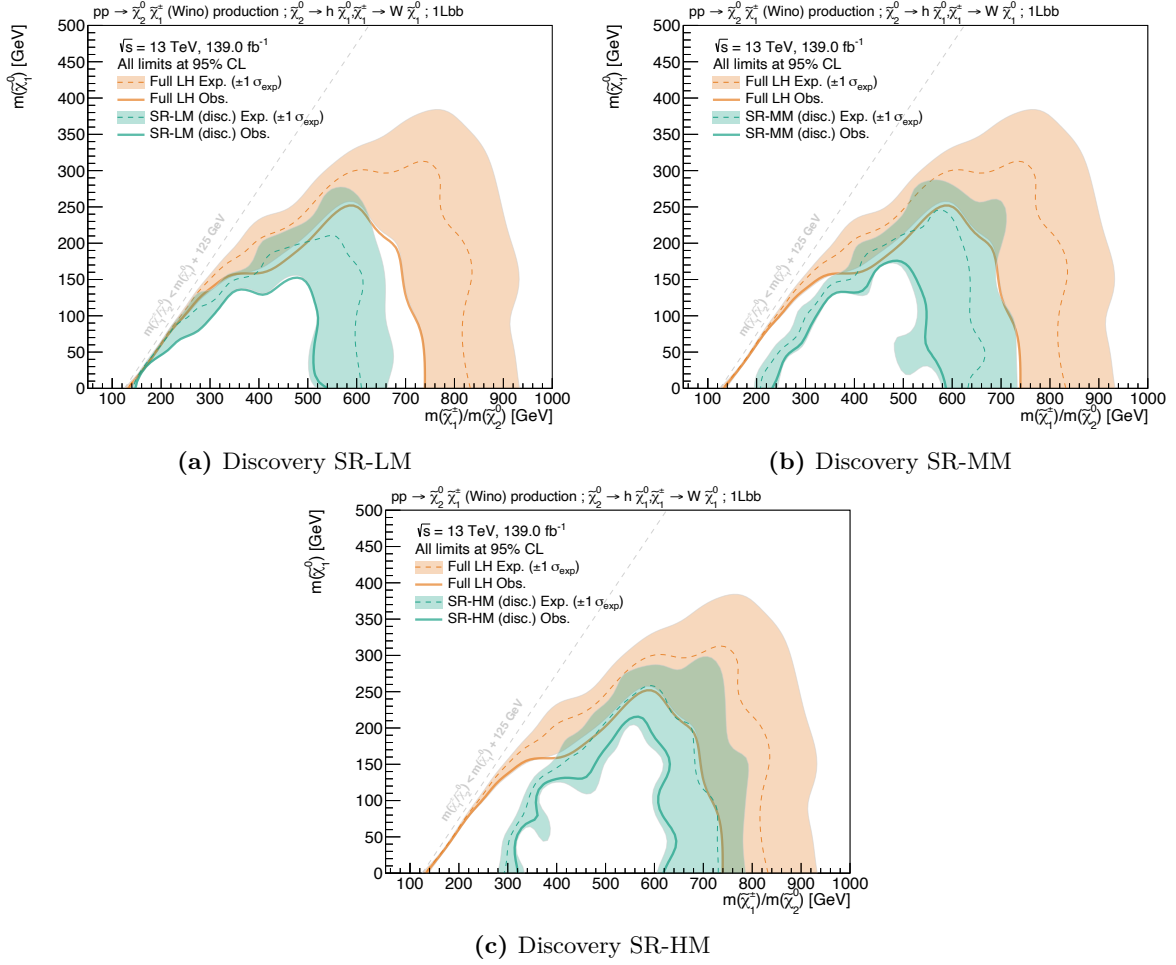
# Simplified likelihoods

In the previous chapter, the concept of preserving an analysis for the purpose of reinterpretations was introduced, and a truth-level version of the signal pipeline was discussed. In large-scale reinterpretations involving a large number of SUSY models to be tested against, not only the signal pipeline but also the statistical inference requires significant computational effort. This chapter introduces the concept of *simplified likelihoods*, an approach approximating the statistical model of an analysis using the HISTFACTORY template in order to achieve more efficient statistical inference.

### 10.1 Motivation

Reinterpretations of ATLAS searches for SUSY in more complete and realistic SUSY scenarios (as opposed to simplified models) typically involves high-dimensional parameter spaces that are computationally extremely challenging to sample and compare to ATLAS data in an exhaustive way. Large-scale reinterpretations of this type have already been performed in ATLAS after the Run 1 data-taking period in both the 19-dimensional pMSSM [76] as well as a 5-dimensional representation of the pMSSM [75]. Due to the complexity of the statistical models of today’s ATLAS searches for SUSY, originating from the large number of channels and the large set of nuisance parameters typically considered, the wall time needed for the statistical inference is usually far from negligible. In a typical large-scale reinterpretation involving  $\mathcal{O}(10^5\text{--}10^6)$  sampled models, an optimistic estimation of the wall time needed for the statistical inference per model of  $\mathcal{O}(10\text{ s--}10^2\text{ s})$  is too computationally expensive, especially when more than just a single ATLAS search for SUSY is included. It is thus crucial to reduce the number of models that need to be evaluated using the searches’ full statistical model.

One approach of alleviating this computational problem is to approximate the SUSY searches through their model-independent limits published in conjunction with the model-dependent exclusion limits. By construction, the model-independent limits are derived using cut-and-count signal regions without relying on multi-bin or shape-fit setups, thus making minimal model assumptions. In the case of searches using dedicated multi-bin signal regions to derive their exclusion limits, this approach—while computationally very efficient—naturally underestimates the true exclusion power the respective search due to the fact that model-dependent properties



**Figure 10.1:** Comparison of exclusion limits obtained using a likelihood built from all nine exclusion signal regions (orange), and the discovery signal regions (green). As discussed in section 5.3, the discovery signal regions are simple cut-and-count regions with minimal model assumptions. They are not mutually exclusive, they cannot be fitted together, thus resulting in three separate exclusion contours. All statistical and systematic uncertainties on the background and the signal event rates are included.

are not exploited. Figure 10.1 illustrates this approach with the example of the  $1\ell$  search. The exclusion contours obtained with exclusion signal regions implementing a two-dimensional shape-fit are compared to the exclusion contours obtained using the discovery signal regions, defined in table 5.3. As the discovery signal regions are not mutually exclusive, they cannot be statistically combined and thus three separate observed and expected contours need to be drawn. It can be observed that even a best-expected combination of the three discovery signal regions does not reach the sensitivity achieved using the two-dimensional shape-fit setup resulting from statistical combination of the nine exclusion signal regions. In the past, such approaches were used nonetheless in large-scale scans of the pMSSM using ATLAS data from Run 1 [76, 75], yielding conservative results and thus leaving substantial room for improvement.

Hence the motivation to introduce a method for approximating ATLAS searches for SUSY without disregarding their elaborate use of multi-bin signal regions exploiting the varying shapes of signal and SM background distributions. The method introduced in the following targets ATLAS searches for SUSY using likelihoods built according to the HISTFACTORY template.

## 10.2 Building simplified likelihoods

In order to retain the full statistical combination of multiple signal region bins implemented in many SUSY searches, while still being able to achieve a sufficiently fast approximation, the statistical treatment of the background model including its uncertainties needs to be simplified. In the procedure presented in the following, this is achieved by first performing a background-only fit to data in all SRs and CRs, in order to determine the best-fit values of all the model parameters  $\phi$ . The post-fit total background estimate as well as the total uncertainty on the estimate in every bin are subsequently computed and used to construct a simplified likelihood.

As the full likelihood in JSON format defines the full statistical model used for the statistical inference, the above background-only fit can be performed using `pyhf` and the preserved full likelihood of the analysis. With the full likelihoods starting to become available on HEPDATA (see e.g. Ref. [262]) this procedure can rely on public information only and is therefore widely accessible to the HEP community. The simplified likelihoods introduced in the following, follow the same JSON specification introduced for the full likelihoods in Ref. [147]. The following description highlights the specification details relevant to the simplified likelihood.

### Background model

In the simplified likelihood, the background model is approximated with a single background sample, called `total_bkg` in listing 10.1 and representing the total SM background estimate in the different analysis channels. The pre-fit sample rate of the total background sample is set to the total post-fit background estimate obtained in the background-only fit using the full likelihood (in listing 10.1 set to be 10.0). Additionally, the complete set of nuisance parameters in the original full likelihood is reduced to a single constrained parameter  $\alpha$  with up and down variations corresponding to the post-fit uncertainties on the total SM background estimates in each bin. In listing 10.1, the single nuisance parameter is called `total_error` and implemented as a *rate modifier* (introduce in section 3.1). It is constrained by a Gaussian  $\text{Gaus}(a = 0|\alpha, \sigma = 1)$

and is correlated over all bins in each channel. Although the final uncertainty is constrained by a simple Gaussian, the full treatment of the uncertainties in the background-only fit using the full likelihood ensures that non-Gaussian effects are included to some extent.

---

```
{
  "name": "total_bkg",
  "data": [10.0],
  "modifiers": [{"data": {"hi_data": [12.0], "lo_data": [8.0]}, "name": "total_error", "type": "histosys"}]
}
```

---

**Listing 10.1:** Example of a total background sample with sample rate and total uncertainty as derived from a previous fit in the SRs and CRs. The

### Analysis channels

Each channel in the full likelihood with the original number of bins is also entering the simplified likelihood, and each contains the total background sample as specified above. Apart from the total background sample, one additional sample is needed—the signal sample. An example of a signal sample is shown in listing 10.2. It introduces the unconstrained signal strength parameter  $\mu$  as second parameter of the statistical model. For simplicity, the example shown in listing 10.2 does not introduce any additional uncertainties on the signal rates, thereby assuming them to be negligible. Depending on the BSM scenario, signal uncertainties can however be introduced through additional rate modifiers.

---

```
{
  "name": "signal",
  "data": [7.0],
  "modifiers": [{"data": null, "name": "mu", "type": "normfactor"}]
}
```

---

**Listing 10.2:** Example of a signal sample with sample rate and unconstrained normalisation parameter that will be used as POI.

### Observations and measurements

According to the JSON specification defined in Ref. [147], the data observed by the analysis in each channel (and each bin) is introduced by means of an *observation*. In the case of the simplified likelihood, this is taken directly from the full likelihood and, by construction, does not need to be modified. An example of an observation including several channels and bins is shown in listing 10.3.

---

```
{
  "observations": {
    {"name": "channel_A" : "data": [25.0]},
    {"name": "channel_B" : "data": [20.0]},
    {"name": "channel_C" : "data": [11.0, 13.0]}
  }
}
```

---

**Listing 10.3:** Example of an observation in the simplified likelihood. It can be directly taken from the corresponding full likelihood. This example implements three channels, two with one bin, and one with three bins.

The only part of the JSON specification left to be defined is the *measurement*, specifying the name of the parameter of interest as well as parameter set configurations not already covered in the channel definitions. For the simplified likelihood, it is straightforward to write down, as can be seen in listing 10.4.

---

```
{
  "measurements": {
    "name": "myMeasurement",
    "config": { "poi": "mu", "parameters": []}
  }
}
```

---

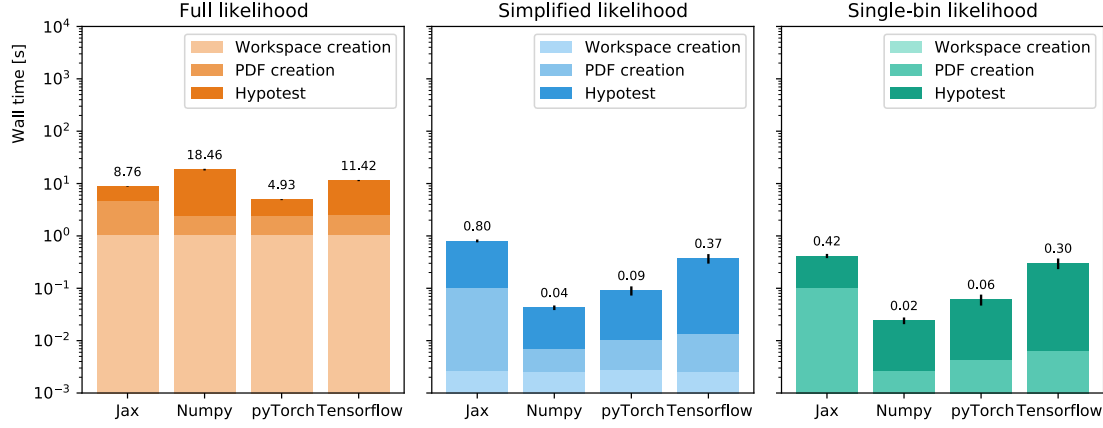
**Listing 10.4:** Example of a measurement in the simplified likelihood. The signal strength is the parameter of interest, no additional parameters need further configuration.

Put together, the above pieces result in a simplified likelihood for a given signal model, using a background model obtained from an initial background-only fit using the full likelihood considering the full treatment of systematic uncertainties. Replacing the signal sample by the means of JSON patches allows for systematic reinterpretations of any signal model for which the expected rates in the analysis regions are known.

## 10.3 Computational performance

One of the main figures of merit of an analysis approximation naturally is the reduction in wall time compared to the full analysis. Figure 10.2 shows a benchmark for different likelihood configurations in the context of the  $1\ell$  search. The wall time of the full analysis likelihood is compared with that of the simplified likelihood constructed following the previously introduced prescription. In addition, the wall time of the single-bin likelihood using the discovery SRs already used in fig. 10.1, is shown. For each likelihood, different computational backends are used for the tensor algebra operations in `pyhf`. All benchmarks have been performed on an Intel i7-4790 CPU with a nominal clock speed of 3.60 GHz, 4 cores and 8 threads. The CPU was not isolated but under minimal load. The original 125 signal points of the  $1\ell$  search were used in each configuration.

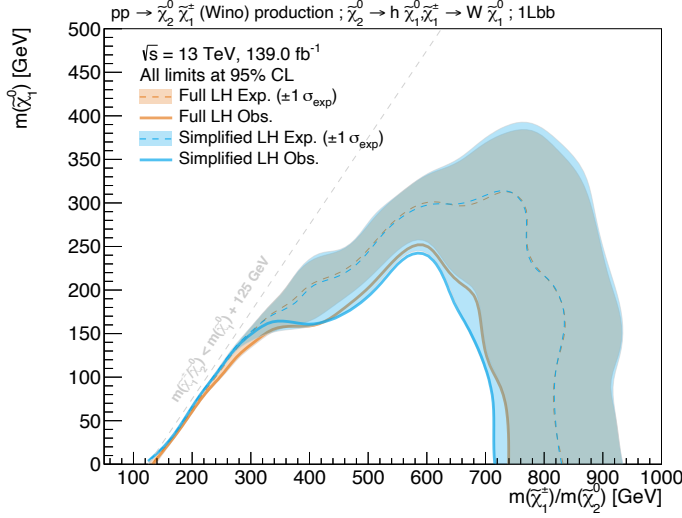
The use of automatic differentiation of the full likelihood gradient enabled by some of the tensor algebra backends to `pyhf` offers an efficient minimisation of the negative log-likelihood, resulting in fast hypothesis tests of  $\mathcal{O}(5\text{ s})$  for the full analysis. In large-scale reinterpretations, this is however still too computationally expensive. The simplified likelihood, on the other hand, results in a wall time for hypothesis tests as fast as 0.04 s per signal model. Thus, the



**Figure 10.2:** Benchmarks of the CPU-time necessary for hypothesis testing using different likelihood and `pyhf` configurations in the context of the  $1\ell$  search. Benchmark details are given in the text. The full likelihood (left) includes the full statistical implementation of the original analysis, the simplified likelihood (center) represents the simplified likelihood approach presented in this document, and the single-bin likelihood (right) represents the single-bin approximation using the discovery signal regions. The uncertainties represent the standard deviation of the benchmark test sample.

**Table 10.1:** Benchmarks of the wall times needed for computing the  $CL_s$  value for a single signal model using the full and the simplified likelihoods. The signal models used for the benchmarks include all signal models originally considered in the respective search. The uncertainty corresponds to the standard deviation of the benchmark sample. The performance gain is stated as ratio between the wall times. The PYTORCH (NUMPY) backend of `pyhf` is used for the full (simplified) likelihood, in conjunction with the SCIPY optimiser. Searches without reference quoted are not yet public.

| Analysis                       | Full likelihood [s] | Simplified likelihood [s] | Improvement |
|--------------------------------|---------------------|---------------------------|-------------|
| ATLAS compressed search [272]  | $16.49 \pm 3.16$    | $0.073 \pm 0.012$         | $236\times$ |
| ATLAS $3\ell$ search           | $40.41 \pm 15.7$    | $0.082 \pm 0.021$         | $495\times$ |
| ATLAS $2\ell$ search [246]     | $5.93 \pm 0.16$     | $0.079 \pm 0.0082$        | $75\times$  |
| ATLAS $1\ell$ search [170]     | $4.93 \pm 0.11$     | $0.040 \pm 0.0057$        | $123\times$ |
| ATLAS direct stau search [273] | $1.91 \pm 0.090$    | $0.039 \pm 0.0055$        | $49\times$  |
| ATLAS sbottom search [274]     | $1.36 \pm 0.067$    | $0.038 \pm 0.0046$        | $36\times$  |
| ATLAS stop search              | $2.27 \pm 0.062$    | $0.044 \pm 0.011$         | $51\times$  |



**Figure 10.3:** Comparison of the simplified likelihood (blue contours) and full likelihood (orange contours) results for the search for electroweakinos presented previously. The observed contours are shown as solid lines, while the expected contours are shown as dashed lines. The uncertainty band includes all MC statistical and systematic uncertainties in the case of the full likelihood, and only the simplified uncertainties in the case of the simplified likelihood.

computational performance of the simplified likelihood is in the same order of magnitude as that of the naive single-bin approach<sup>†</sup>, but offers by construction a significantly better approximation of the true analysis exclusion power.

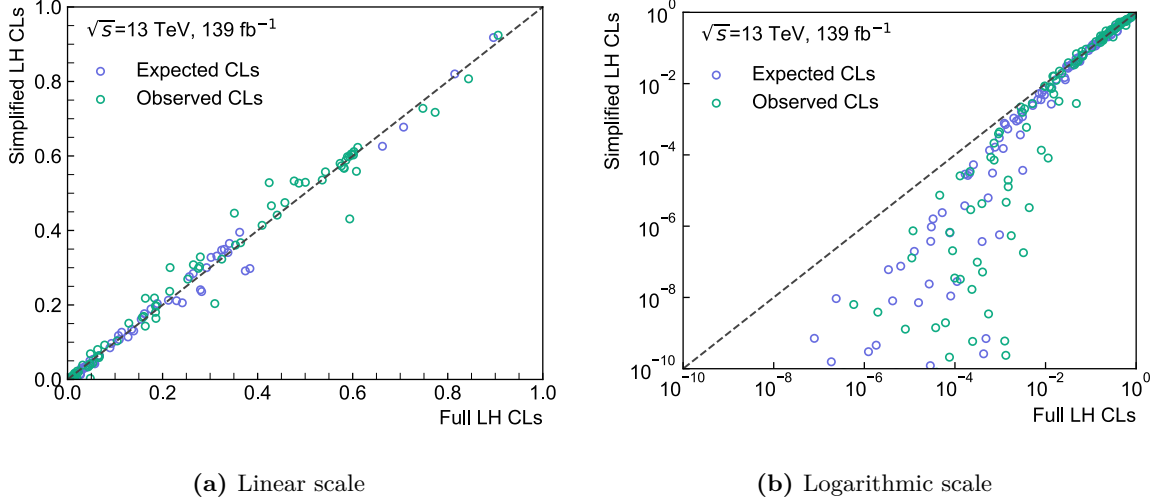
Interestingly, the wall time of the simplified likelihood does not benefit from the usage of features like automatic differentiation offered by e.g. `PYTORCH`. This is due to the extreme simplicity of the simplified likelihood function, causing the computational benefits from features like automatic differentiation to not outweigh the sizeable overhead of libraries like `PYTORCH`.

In addition to the search for electroweakinos presented herein, the simplified likelihood approach has also been applied on a number of other ATLAS searches for SUSY. Table 10.1 summarises the mean wall time of all ATLAS searches investigated using the simplified likelihood approach. In all cases, `PYTORCH` offers the fastest backend for the full likelihood while `NUMPY` performs best for the simplified likelihood. The performance improvement of roughly two orders of magnitude obtained in the  $1\ell$  search is confirmed in the other ATLAS SUSY searches investigated. The wall time of the simplified likelihood appears to be bound from below at  $\mathcal{O}(10^{-2} \text{ s})$ , limiting the performance gain for some of the faster analyses whose full likelihoods are already relatively simple to begin with.

## 10.4 Physics performance

A comparison of the exclusion contours obtained with the full and simplified likelihoods in the context of the  $1\ell$  search is shown in fig. 10.3. The results obtained using the simplified likelihood are shown in blue, while the results obtained using the full likelihood are presented in orange. Both the observed (without the usual theoretical up and down variations on the signal cross section) and expected exclusion limits including the uncertainty band are shown. In the case of the full likelihood, the complete set of MC statistical and systematic uncertainties introduced

<sup>†</sup> In fact, the simplified likelihood is actually even faster than the single-bin approach, as the latter needs to be executed separately for each discovery SR and thus the numbers quoted need to be multiplied by the number of discovery SRs used in the analysis (three in the case of the  $1\ell$  search).



**Figure 10.4:** Scatter plots comparing the observed and expected  $CL_s$  values obtained using the simplified and the full likelihoods for the same set of signal models considered in the search for electroweakinos. Both linear and logarithmic scale representations are shown.

in chapter 7 are taken into account. As discussed in section 10.2, the uncertainty band on the simplified likelihood contour results from the single nuisance parameter built through reduction of the original nuisance parameters.

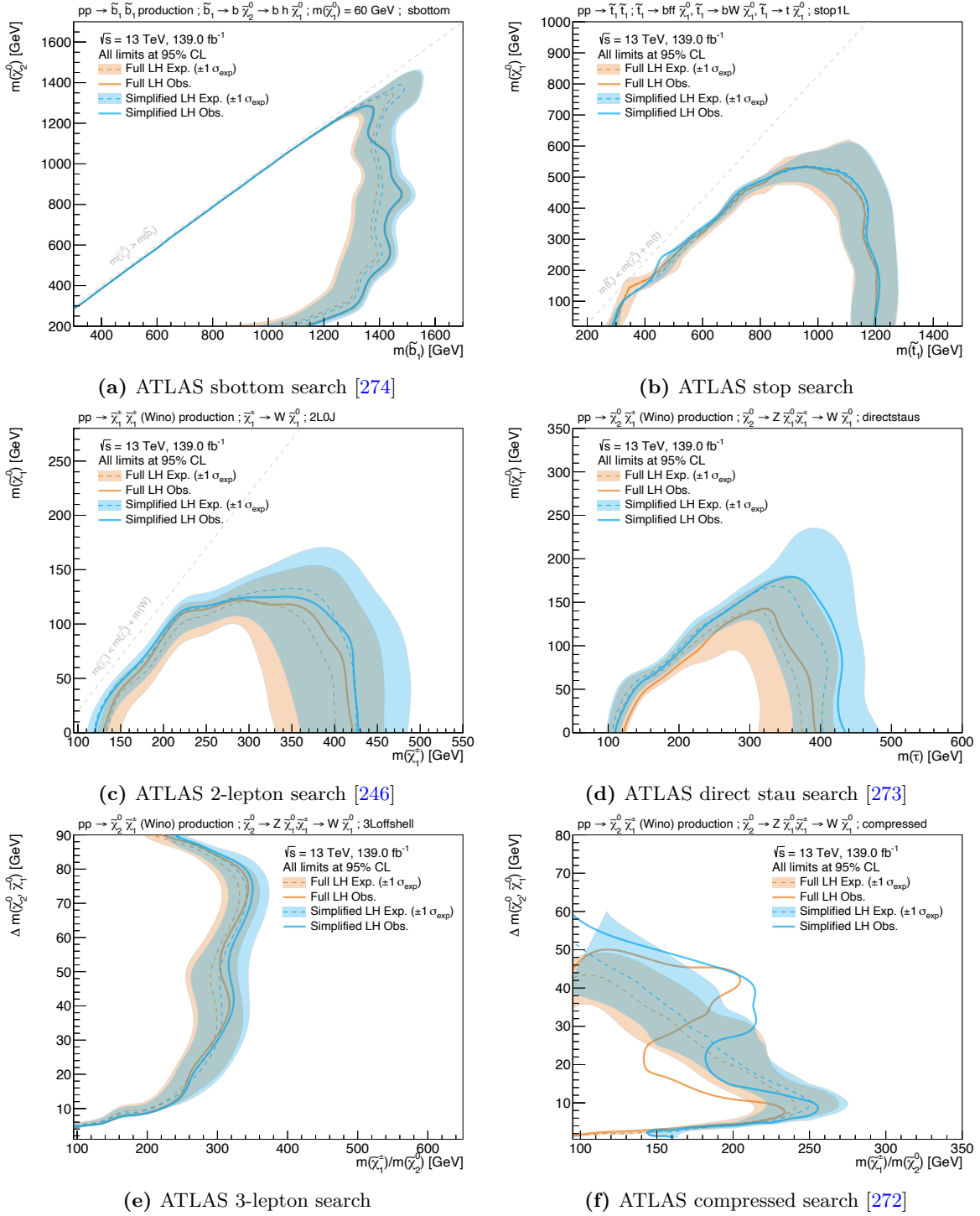
The observed and expected  $CL_s$  values obtained using both likelihoods are shown in fig. 10.4. As expected from the exclusion contour, both the simplified and the full likelihood agree reasonably well across the majority of the shown range in  $CL_s$ . For signal models well within exclusion with the full likelihood, i.e.  $CL_s \ll 0.05$ , the simplified likelihood of the  $1\ell$  search tends to result in slightly lower  $CL_s$  values than the full likelihood, thus giving a slightly too optimistic sensitivity estimate. In the range relevant to the exclusion contour at 95% CL, the results from the simplified likelihood agree however well with those from the full likelihood.

In addition to the  $1\ell$  search, the simplified likelihood approach has been tested on the ATLAS SUSY searches listed in table 10.1. An overview of the results can be seen in fig. 10.5, comparing the exclusion contours obtained with the simplified likelihood against the full analysis results. In some analyses, e.g. the ATLAS sbottom search as well as the ATLAS  $3\ell$  search, the simplified likelihoods show excellent agreements. In other analyses like e.g. the ATLAS direct stau search, the agreement is less good but overall still acceptable, demonstrating that this method can, in many cases, offer a fast and reliable approximation of ATLAS searches for SUSY.

## 10.5 Limitations

Building a well-performing simplified likelihood is not always as straightforward as described in section 10.2 and some analyses require special care when approximated. For example, in the case of the ATLAS compressed search [272] shown in fig. 10.5(f), only a subset of the original analysis signal regions are entering the simplified likelihood, in order to improve the general agreement. The straightforward structure of the simplified likelihood is, in this case, not able





**Figure 10.5:** Simplified likelihood results for the different ATLAS searches studied in this document. The results from the simplified likelihood (blue) are compared with the results of the full analysis likelihood (orange). The coloured numbers represent the observed  $CL_s$  numbers obtained with both likelihoods.

to reproduce the statistical behaviour of the background model in the full likelihood. As the omitted channels only add limited sensitivity to the search, their removal in the simplified likelihood yields an overall improvement in agreement.

The reason for this is that the simplified likelihood assumes that the background model can be described by a single sample with a single systematic uncertainty constrained by a Gaussian correlated over all bins, with background event rates and uncertainties obtained from a background-only fit in all CRs and SRs using the full likelihood. This in particular assumes that the background model is sufficiently constrained by the large statistics in the CRs and that the introduction of signal contributions—especially in the SRs—does not significantly change the background model in a way that cannot be replicated with a single background sample where the event rates only depend on a single nuisance parameter. While it can be argued that such a log-likelihood fit configuration, where CRs are no longer sufficiently constraining the background, should be avoided in an analysis, such a configuration is especially problematic for the simplified likelihood, where the background model is assumed to be fixed up to a single constrained nuisance parameter.

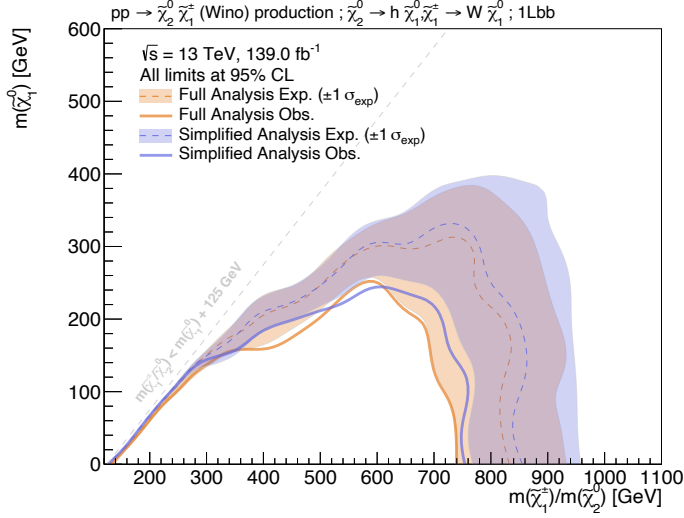
An additional limitation arises in cases of significant signal contamination in the CRs. In the full likelihood, significant signal contamination in the CRs generally leads to smaller background estimates in the SRs, which, in turn, results in conservative exclusion limits given the observed data. In the simplified likelihood, even with the CRs included, the single constrained nuisance parameter might in such cases not offer the log-likelihood fit enough freedom to scale down the background model enough in the  $\mu = 1$  hypothesis, resulting in *fake* sensitivity in the CRs. Although it is generally important to limit signal contamination in the CRs for the sake of healthy statistical log-likelihood fits, this is especially true in the case of very simplified likelihoods as introduced herein. In the case of the ATLAS stop search shown in fig. 10.5(b), significant signal contamination of more than 30% appears for many signal models with  $m(\tilde{t}_1) < m(\tilde{\chi}_1^0) + m(t)^\dagger$ , which can thus not be evaluated with the simplified likelihood.

## 10.6 Outlook and future prospects

The simplified likelihoods introduced in this chapter can offer precise and extremely efficient approximations of ATLAS SUSY searches for which the full likelihood in JSON format is available (either internally but preferably publicly). A proof-of-concept python tool has been developed for generic conversion of any full likelihood into the simplified format introduced here.

The procedure of approximating the statistical model of a search is orthogonal to the truth-level analysis discussed in section 9.4 in the sense that both approximations target a different part of the analysis workflow shown in fig. 9.1. As such, both approaches can be combined into a *simplified analysis* that runs a truth-level analysis in order to determine an estimate for signal event rates, followed by a simplified statistical inference using the simplified likelihood instead of the full statistical model. Figure 10.6 compares the expected and observed exclusion contours obtained in the full  $1\ell$  search with those obtained with a simplified version of the same search, using truth-level analysis and a simplified likelihood. A good agreement can be observed between the exclusion contours obtained by both the full and simplified analyses, especially considering the sizeable scope of the two-fold approximation applied in the simplified

<sup>†</sup> This is a kinematic region that the analysis is not designed to be sensitive in.



**Figure 10.6:** Expected and observed exclusion contours obtained with the full likelihood and reconstruction-level inputs (orange) and the simplified likelihood and smeared truth-level inputs (purple). Uncertainties include all statistical and systematic uncertainties on the background and signal for the reconstruction-level contours determined using the full likelihood, but only statistical and systematic uncertainties on the background for the simplified likelihood contours (through the process of constructing the simplified likelihood).

analysis. This simplified analysis will be used in the following chapter in a computationally efficient reinterpretation of the  $1\ell$  search in the pMSSM

As the full likelihood defines the full statistical model given the observed data in an analysis, other forms of likelihood simplifications can be thought of. One possible approach worth investigating is the construction of likelihood simplifications using a variable number of nuisance parameters, as opposed to reducing the full set of nuisance parameters to a single one. In such an approach, a principal components analysis could project the full  $N$ -dimensional nuisance parameter space onto a number  $n \leq N$  principal components maximising the variance of the projected space, i.e. resulting in minimal loss in correlation information. The  $n$  principal components can then be kept separate, while the  $N - n$  remaining components can be combined in quadrature into a *residual* term. A similar approach was already introduced in chapter 7 where the large number of nuisance parameters connected to the JER and JES uncertainties in the  $1\ell$  search were reduced to a more manageable set of *effective* nuisance parameters with minimal loss in bin-by-bin correlation information.

Up until very recently, the only way for physicists outside the collaboration to re-use the statistical inference of ATLAS searches for SUSY involved building approximations of their statistical models based on lossy projections of the full likelihood. With ATLAS' recent push to publish full analysis likelihoods, new approaches for approximation of the statistical models are becoming available. In principle, the full likelihood contains all information necessary for generating a simplified likelihood with a certain compromise between statistical precision and computational efficiency, allowing to find an ideal approximation given constraints on available computing resources of the specific use-case.

