PSI-PR-20-23; BONN-TH-2020-11; CP3-20-59; KCL-PH-TH/2020-75; P3H-20-080; TTP20-044;
TUM-HEP-1310/20; IFT-UAM/CSIC-20-180; TTK-20-47; CERN-TH-2020-215; FTPI-MINN-20-36;
UMN-TH-4005/20; HU-EP-20/37; DESY 20-222; ADP-20-33/T1143; Imperial/TP/2020/RT/04; UCI-TR-2020-19



# Simple and statistically sound recommendations for analysing physical theories

**Shehu S. AbdusSalam[1], Fruzsina J. Agocs[a,2,3], Benjamin C. Allanach[4], Peter Athron[a,5,6], Csaba Balázs[a,6], Emanuele Bagnaschi[b,7], Philip Bechtle[c,8], Oliver Buchmueller[b,9], Ankit Beniwal[a,10], Jihyun Bhom[a,11], Sanjay Bloor[a,9,12], Torsten Bringmann[a,13], Andy Buckley[a,14], Anja Butter[15], José Eliel Camargo-Molina[a,16], Marcin Chrzaszcz[a,11], Jan Conrad[a,17], Jonathan M. Cornell[a,18], Matthias Danninger[a,19], Jorge de Blas[d,20], Albert De Roeck[b,21], Klaus Desch[c,8], Matthew Dolan[b,22], Herbert Dreiner[c,8], Otto Eberhardt[d,23], John Ellis[b,24], Ben Farmer[a,9,25], Marco Fedele[d,26], Henning Flächer[b,27], Andrew Fowlie[a,5,*], Tomás E. Gonzalo[a,6], Philip Grace[a,28], Matthias Hamer[c,8], Will Handley[a,2,3], Julia Harz[a,29], Sven Heinemeyer[b,30], Sebastian Hoof[a,31], Selim Hotinli[a,9], Paul Jackson[a,28], Felix Kahlhoefer[a,32], Kamila Kowalska[e,33], Michael Krämer[c,32], Anders Kvellestad[a,13], Miriam Lucio Martinez[b,34], Farvah Mahmoudi[a,35,36], Diego Martinez Santos[b,37], Gregory D. Martinez[a,38], Satoshi Mishima[d,39], Keith Olive[b,40], Ayan Paul[d,41,42], Markus Tobias Prim[a,8], Werner Porod[c,43], Are Raklev[a,13], Janina J. Renk[a,9,12,17], Christopher Rogan[a,44], Leszek Roszkowski[e,45,33], Roberto Ruiz de Austri[a,30], Kazuki Sakurai[b,46], Andre Scaffidi[a,47], Pat Scott[a,9,12], Enrico Maria Sessolo[e,33], Tim Stefaniak[c,41], Patrick Stöcker[a,32], Wei Su[a,28,48], Sebastian Trojanowski[e,45,33], Roberto Trotta[9,49], Yue-Lin Sming Tsai[50], Jeriek Van den Abeele[a,13], Mauro Valli[d,51], Aaron C. Vincent[a,52,53,54], Georg Weiglein[b,41,55], Martin White[a,28], Peter Wienemann[c,8], Lei Wu[a,5], and Yang Zhang[a,6,56]**

[a]The GAMBIT Community
[b]The MasterCode Collaboration
[c]The Fittino Collaboration
[d]HEPfit
[e]BayesFits Group

[1]Department of Physics, Shahid Beheshti University, Tehran, Iran
[2]Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge, CB3 0HE, UK
[3]Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge, CB3 0HA, UK
[4]DAMTP, University of Cambridge, Cambridge, CB3 0WA, UK
[5]Department of Physics and Institute of Theoretical Physics, Nanjing Normal University, Nanjing, Jiangsu 210023, China
[6]School of Physics and Astronomy, Monash University, Melbourne, VIC 3800, Australia
[7]Paul Scherrer Institut, CH-5232 Villigen, Switzerland
[8]University of Bonn, Physikalisches Institut, Nussallee 12, D-53115 Bonn, Germany
[9]Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK
[10]Centre for Cosmology, Particle Physics and Phenomenology (CP3), Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium
[11]Institute of Nuclear Physics, Polish Academy of Sciences, Krakow, Poland
[12]School of Mathematics and Physics, The University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia
[13]Department of Physics, University of Oslo, Box 1048, Blindern, N-0316 Oslo, Norway
[14]School of Physics and Astronomy, University of Glasgow, University Place, Glasgow, G12 8QQ, UK
[15]Institut für Theoretische Physik, Universität Heidelberg, Germany
[16]Department of Physics and Astronomy, Uppsala University, Box 516, SE-751 20 Uppsala, Sweden
[17]Oskar Klein Centre for Cosmoparticle Physics, AlbaNova University Centre, SE-10691 Stockholm, Sweden
[18]Department of Physics, Weber State University, 1415 Edvalson St., Dept. 2508, Ogden, UT 84408, USA
[19]Department of Physics, Simon Fraser University, 8888 University Drive, Burnaby B.C., Canada
[20]Institute of Particle Physics Phenomenology, Durham University, Durham DH1 3LE, UK
[21]Experimental Physics Department, CERN, CH–1211 Geneva 23, Switzerland
[22]ARC Centre of Excellence for Dark Matter Particle Physics, School of Physics, The University of Melbourne, Victoria 3010, Australia
[23]Instituto de Física Corpuscular, IFIC-UV/CSIC, Apt. Correus 22085, E-46071, Valencia, Spain

[24]Theoretical Particle Physics and Cosmology Group, Department of Physics, King's College London, London WC2R 2LS, UK
[25]Bureau of Meteorology, Melbourne, VIC 3001, Australia
[26]Institut für Theoretische Teilchenphysik, Karlsruhe Institute of Technology, D-76131 Karlsruhe, Germany
[27]H. H. Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK
[28]ARC Centre for Dark Matter Particle Physics, Department of Physics, University of Adelaide, Adelaide, SA 5005, Australia
[29]Physik Department T70, James-Franck-Straße, Technische Universität München, D-85748 Garching, Germany
[30]Instituto de Física Teórica UAM-CSIC, Cantoblanco, 28049, Madrid, Spain
[31]Institut für Astrophysik und Geophysik, Georg-August-Universität Göttingen, Friedrich-Hund-Platz 1, D-37077 Göttingen, Germany
[32]Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University, Sommerfeldstraße 14, D-52056 Aachen, Germany
[33]National Centre for Nuclear Research, ul. Pasteura 7, PL-02-093 Warsaw, Poland
[34]Nikhef National Institute for Subatomic Physics, Amsterdam, Netherlands
[35]Université de Lyon, Université Claude Bernard Lyon 1, CNRS/IN2P3, Institut de Physique des 2 Infinis de Lyon, UMR 5822, F-69622, Villeurbanne, France
[36]Theoretical Physics Department, CERN, CH-1211 Geneva 23, Switzerland
[37]Instituto Galego de Física de Altas Enerxías, Universidade de Santiago de Compostela, Spain
[38]Physics and Astronomy Department, University of California, Los Angeles, CA 90095, USA
[39]Theory Center, IPNS, KEK, Tsukuba, Ibaraki 305-0801, Japan
[40]William I. Fine Theoretical Physics Institute, School of Physics and Astronomy, University of Minnesota, Minneapolis, MN 55455, USA
[41]Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany
[42]Institut für Physik, Humboldt-Universität zu Berlin, D-12489 Berlin, Germany
[43]University of Würzburg, Emil-Hilb-Weg 22, D-97074 Würzburg, Germany
[44]Department of Physics and Astronomy, University of Kansas, Lawrence, KS 66045, USA
[45]Astrocent, Nicolaus Copernicus Astronomical Center Polish Academy of Sciences, Bartycka 18, PL-00-716 Warsaw, Poland
[46]Institute of Theoretical Physics, Faculty of Physics, University of Warsaw, ul. Pasteura 5, PL-02-093 Warsaw, Poland
[47]Istituto Nazionale di Fisica Nucleare, Sezione di Torino, via P. Giuria 1, I–10125 Torino, Italy
[48]Korea Institute for Advanced Study, Seoul 02455, Korea
[49]SISSA International School for Advanced Studies, Via Bonomea 265, 34136, Trieste, Italy
[50]Key Laboratory of Dark Matter and Space Astronomy, Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210033, China
[51]Department of Physics and Astronomy, University of California, Irvine, California 92697, USA
[52]Department of Physics, Engineering Physics and Astronomy, Queen's University, Kingston ON K7L 3N6, Canada
[53]Arthur B. McDonald Canadian Astroparticle Physics Research Institute, Kingston ON K7L 3N6, Canada
[54]Perimeter Institute for Theoretical Physics, Waterloo ON N2L 2Y5, Canada
[55]Institut für Theoretische Physik, Universität Hamburg,Luruper Chaussee 149, 22761 Hamburg, Germany
[56]School of Physics, Zhengzhou University, ZhengZhou 450001, China
[*]E-mail: andrew.j.fowlie@njnu.edu.cn

## Abstract

Physical theories that depend on many parameters or are tested against data from many different experiments pose unique challenges to statistical inference. Many models in particle physics, astrophysics and cosmology fall into one or both of these categories. These issues are often sidestepped with statistically unsound *ad hoc* methods, involving intersection of parameter intervals estimated by multiple experiments, and random or grid sampling of model parameters. Whilst these methods are easy to apply, they exhibit pathologies even in low-dimensional parameter spaces, and quickly become problematic to use and interpret in higher dimensions. In this article we give clear guidance for going beyond these procedures, suggesting where possible simple methods for performing statistically sound inference, and recommendations of readily-available software tools and standards that can assist in doing so. Our aim is to provide any physicists lacking comprehensive statistical training with recommendations for reaching correct scientific conclusions, with only a modest increase in analysis burden. Our examples can be reproduced with the code publicly available at Zenodo.

## 1 Introduction

The search for new particles is underway in a wide range of high-energy, astrophysical and precision experiments. These searches are made harder by the fact that theories for physics beyond the Standard Model almost always contain unknown parameters that cannot be uniquely derived from the theory itself. For example, in particle physics models of dark matter, these would be the dark matter mass and its couplings. Models usually make a range of different experimental predictions depending on the assumed values of their unknown parameters. Despite an ever-increasing wealth of experimental data, evidence for specific physics beyond the Standard Model has not yet emerged, leading to the proposal of increasingly complicated models. This increases the number of unknown parameters in the models, leading to high-dimensional parameter spaces. This problem is compounded by additional calibration and nuisance parameters that are required as experiments become more complicated. Unfortunately, high-dimensional parameter spaces, and the availability of relevant constraints from an increasing number of experiments, expose flaws in the simplistic methods sometimes employed in phenomenology to assess models. In this article, we recommend alternatives suitable for today's models and data, consistent with established statistical principles.

When assessing a model in light of data, physicists typically want answers to two questions: *a*) Is the model favoured or allowed by the data? *b*) What values of the unknown parameters are favoured or allowed by the data? In statistical language, these questions concern model testing and parameter estimation, respectively. Parameter estimation allows us to understand what a model could predict, and design future experiments to test it. On the theory side, it allows us to construct theories that contain the model and naturally accommodate the observations. Model testing, on the other hand, allows us to test whether data indicate the presence of a new particle or new phenomena.

Many analyses of particle physics models suffer from two key deficiencies. First, they overlay exclusion curves from experiments and, second, they perform a random or grid scan of a high-dimensional parameter space. These techniques are often combined to perform a crude hypothesis test. In this article, we recapitulate relevant statistical principles, point out why both of these methods give unreliable results, and give concrete recommendations for what should be done instead. Despite the prevalence of these problems, we stress that there is diversity in the depth of statistical training in the physics community. Physicists contributed to major developments in statistical theory[1,2] and there are many statistically rigorous works in particle physics and related fields, including the famous Higgs discovery,[3,4] and global fits of electroweak data.[5] Our goal is to make clear recommendations that would help lift all analyses closer to those standards, though we urge particular caution when testing hypotheses as unfortunately there are no simple recipes. The examples that we use to illustrate our recommendations can be reproduced with the code publicly available through Zenodo.[6]

Our discussion covers both Bayesian methods,[7–12] in which one directly considers the plausibility of a model and regions of its parameter space, and frequentist methods,[13–16] in which one compares the observed data to data that could have been observed in identical repeated experiments.* Our recommendations are agnostic about the relative merits of the two sets of methods, and apply whether one is an adherent of either form, or neither. Both approaches usually involve the so-called likelihood function,[17] which tells us the probability of the observed data, assuming a particular model and a particular combination of numerical values for its unknown parameters.

In the following discussions, we assume that a likelihood is available and consider inferences based on it. In general, though, the likelihood alone is not enough in frequentist inference (as well as for reference priors and some methods in Bayesian statistics that use simulation). One requires the so-called sampling distribution; this is similar to the likelihood function, except that the data is not fixed to the observed data (see the likelihood principle[18] for further discussion). There are, furthermore, situations in which the likelihood is intractable. In such cases, likelihood-free techniques may be possible.[19] In fact, in realistic applications in physics, the complete likelihood is almost always intractable. Typically, however, we create summaries of the data by e.g. binning collider events into histograms.

## 2 Problems of overlaying exclusion limits

Experimental searches for new phenomena are usually summarised by confidence regions, either for a particular model's parameters or for model-independent quantities more closely related to the experiment that can be interpreted in any model. For example, experiments performing direct searches for dark matter[20] publish confidence regions for the mass and scattering cross section of the dark matter particle, rather than for any parameters included in the Lagrangian of a specific dark matter model. To apply those results to a given dark matter model, the confidence

*We cite here introductory textbooks about statistics by and for scientists. Refs. 9, 15 are particularly concise.

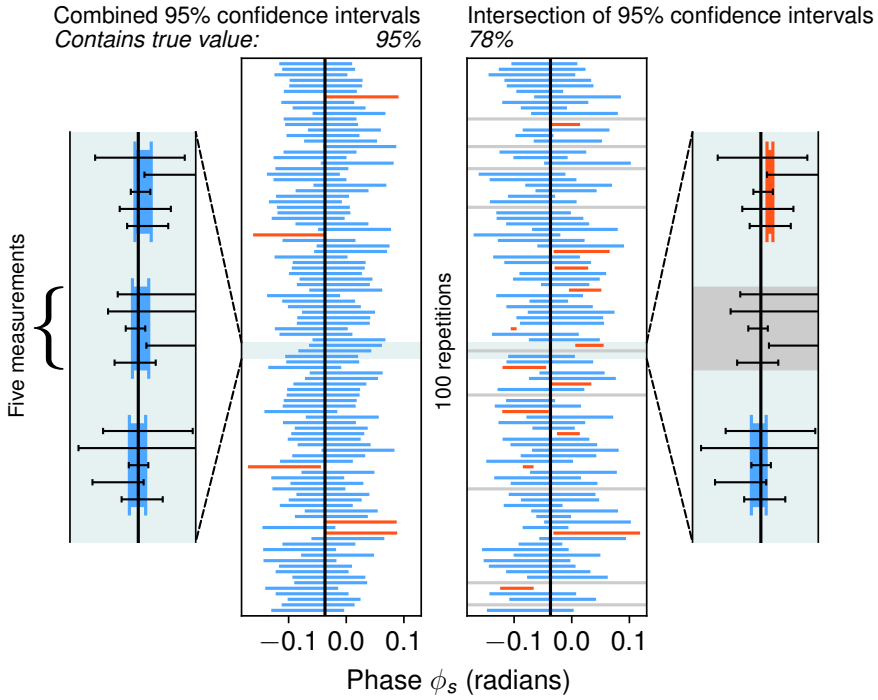


**Figure 1.** Confidence intervals in 100 pseudo-experiments, from the combination of five measurements (*left*) or from the intersection of five individual confidence intervals (*right*). We show the true value of $\phi_s$ with a vertical black line. Intervals that contain the true value are shown in blue; those that do not are shown in red. On the right-hand side, grey bands indicate cases where no value can be found where the 95% intervals from all five measurements overlap. Each bar originates from five pseudo-measurements, as shown zoomed-in to the side for a few points.

regions must be transformed to the parameter space of the specific model of interest. This can sometimes modify the statistical properties of the confidence regions, so care must be taken in performing the transformation.[21–23]

In the frequentist approach, if an experiment that measured a parameter were repeated over and over again, each repeat would lead to a different confidence region for the measured parameter. The coverage is the fraction of repeated experiments in which the resulting confidence region would contain the true parameter values.[24] The confidence level of a confidence region is the desired coverage.[†] For example: a 95% confidence region should contain the true values in 95% of repeated experiments, and the rate at which we would wrongly exclude the true parameter values is controlled to be 5%. Approximate confidence regions can often be found from the likelihood function alone using asymptotic assumptions about the sampling distribution, e.g., Wilks' theorem.[29] However, it is important to check carefully that the required assumptions hold.[30]

Confidence intervals may be constructed to be one- or two-tailed. By construction, in the absence of a new effect, a 95% upper limit would exclude all effect sizes, including zero, at a rate of 5%. The fact that confidence intervals may exclude effect sizes that the experiment had no power to discover was considered a problem in particle physics and lead to the creation of $CL_s$ intervals.[31] By construction, these intervals cannot exclude negligible effect sizes, and thus over-cover.

The analogous construct in Bayesian statistics is the credible region. First, prior information about the parameters and information from the observed data contained in the likelihood function are combined into the posterior using Bayes' theorem. Second, parameters that are not of interest are integrated over, resulting in a marginal posterior distribution. A 95% credible region for the remaining parameters of interest is found from the marginal posterior by defining a region containing 95% of the posterior probability. In general, credible regions only guarantee average coverage: suppose we re-sampled model parameters and pseudo-data from the model and constructed 95% credible regions. In 95% of such trials, the credible region would contain the sampled model parameters.[15,32] Whilst credible

[†]Note that for discrete observations[25] or in the presence of nuisance parameters,[26,27] confidence regions are often defined to include the true parameter values in *at least* e.g. 95% of repeated experiments,[28] and that in some cases the nominal confidence level may not hold in practice.

regions and confidence intervals are identical in some cases (e.g. in normal linear models), the fact that they in general lead to different inferences remains a point of contention.[33] For both credible regions and confidence intervals, the level only stipulates the size of the region. One requires an ordering rule to decide which region of that size is selected. For example, the Feldman-Cousins construction[34] for confidence regions and the highest-posterior density ordering rule for credible regions naturally switch from a one- to a two-tailed result.

When several experiments report confidence regions, requiring that the true value must lie within all of those regions amounts to approximating the combined confidence region by the intersection of regions from the individual experiments. This quickly loses accuracy as more experiments are applied in sequence, and leads to much greater than nominal error rates. This is because by taking an intersection of $n$ independent 95% confidence regions, a parameter point has $n$ chances to be excluded at a 5% error rate, giving an error rate of $1 - 0.95^n$.[35]

This issue is illustrated in Figure 1 using the $B$-physics observable $\phi_s$, which is a well-measured phase characterising CP-violation in $B_s$ meson decays.[36] We perform 10,000 pseudo-experiments.[‡] Each pseudo-experiment consists of a set of five independent Gaussian measurements of an assumed true Standard Model value of $\phi_s = -0.037$ with statistical errors 0.078, 0.097, 0.037, 0.285, and 0.17, which are taken from real ATLAS, CMS and LHCb measurements.[§] We can then obtain the 95% confidence interval from the combination of the five measurements in each experiment,[¶] and compare it to the interval resulting from taking the intersection of the five 95% confidence intervals from the individual measurements. We show the first 100 pseudo-experiments in Figure 1. As expected, the 95% confidence interval from the combination contains the true value in 95% of simulated experiments. The intersection of five individual 95% confidence intervals, on the other hand, contains the true value in only 78% of simulations. Thus, overlaying regions leads to inflated error rates and can create a misleading impression about the viable parameter space. Whilst this is a one-dimensional illustration, an identical issue would arise for the intersection of higher-dimensional confidence regions. Clearly, rather than taking the intersection of reported results, one should combine likelihood functions from multiple experiments. Good examples can be found in the literature.[38–44]

In Figure 2 we again show the dangers of simply overlaying confidence regions. We construct several toy two-dimensional likelihood functions (top), and find their 95% confidence regions (bottom left). In the bottom right panel, we show the contours of the combined likelihood function (blue) and a combined 95% confidence region (red contour). We see that the intersection of confidence regions (dashed black curve) can both exclude points that are allowed by the combined confidence region, and allow points that should be excluded. It is often useful to plot both the contours of the combined likelihood (bottom right panel) and the contours from the individual likelihoods (bottom left panel), in order to better understand how each measurement or constraint contributes to the final combined confidence region.

*Recommendation:* Rather than overlaying confidence regions, combine likelihood functions. Derive a likelihood function for all the experimental data (this may be as simple as multiplying likelihood functions from independent experiments), and use it to compute approximate joint confidence or credible regions in the native parameter space of the model.

## 3 Problems of uniform random sampling and grid scanning

Parameter estimation generally involves integration of a posterior or maximisation of a likelihood function. This is required to go from the full high-dimensional model to the one or two dimensions of interest or to compare different models. In most cases this cannot be done analytically. The likelihood function, furthermore, may be problematic in realistic settings. In particle physics,[45] it is usually moderately high-dimensional, and often contains distinct modes corresponding to different physical solutions, degeneracies in which several parameters can be adjusted simultaneously without impacting the fit, and plateaus in which the model is unphysical and the likelihood is zero. On top of that, only noisy estimates of the likelihood may be available, such as from Monte Carlo simulations of collider searches for new particles, and derivatives of the likelihood function are usually unavailable.[46] As even single evaluations of the likelihood function can be computationally expensive, the challenge is then to perform integration or maximisation in a high-dimensional parameter space using a tractable number of evaluations of the likelihood function.

[‡] In a pseudo-experiment, we simulate the random nature of a real experimental measurement using a pseudo-random number generator on a computer. Pseudo-experiments may be used to learn about the expected distributions of repeated measurements.

[§] See Eq. (91) and Table 22 in Ref. 36.

[¶] We used the standard weighted-mean approach to combine the results.[37]

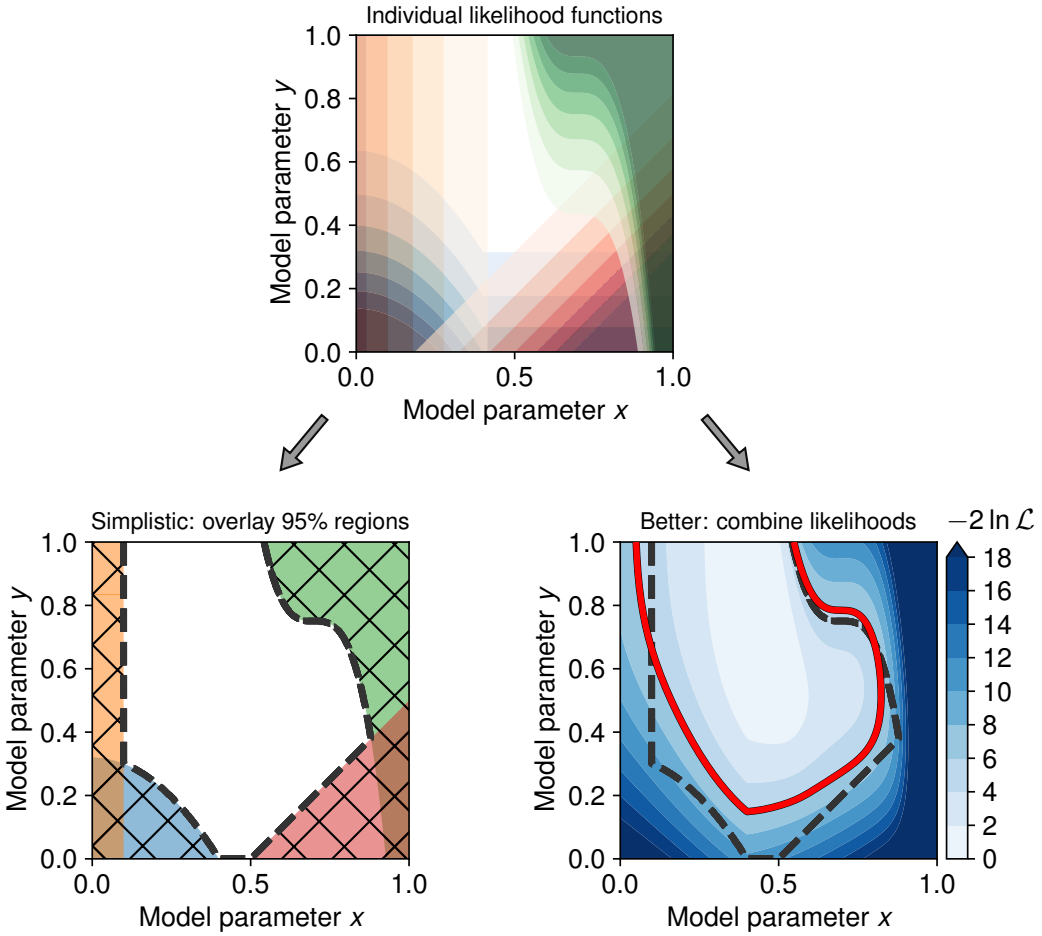


**Figure 2.** Starting from four individual likelihood functions (*top*; orange, blue, red and green, where lighter shades indicate greater likelihood), we compare overlaid 95% contours (*bottom left*) versus a combination of the likelihoods (*bottom right*; blue contours). The dashed black line in both bottom panels is the intersection of the limits from the individual likelihoods. The red line in the bottom right panel is the resulting 95% contour of the product of all likelihoods.

Random and grid scans are common strategies in the high-energy phenomenology literature. In random scans, one evaluates the likelihood function at a number of randomly-chosen parameter points. Typically the parameters are drawn from a uniform distribution in each parameter in a particular parametrisation of the model, which introduces a dependency on the choice of parametrisation. In grid scans, one evaluates the likelihoods on a uniformly spaced grid with a fixed number of points per dimension. It is then tempting to attribute statistical meaning to the number or density of samples found by random or grid scans. However, such an interpretation is very problematic, in particular when the scan is combined with the crude method described in Section 2, i.e. keeping only points that make predictions that lie within the confidence regions reported by every single experiment. It is worth noting that random scans often outperform grid scans: consider 100 likelihood evaluations in a two-parameter model where the likelihood function depends much more strongly on the first parameter than on the second. A random scan would try 100 different parameter values of the important parameter, whereas the grid scan would try just 10. In a similar vein, quasi-random samples that cover the space more evenly than truly random samples can out-perform truly random sampling.[47] This is illustrated in Figure 3 with 256 samples in two-dimensions.

However, random, quasi-random and grid scans are all extremely inefficient in cases with even a few parameters. The "curse of dimensionality"[48] is one of the well-known problems: the number of samples required for a fixed resolution per dimension scales exponentially with dimension $D$: just 10 samples per dimension requires $10^D$ samples. This quickly becomes an impossible task in high-dimensional problems. Similarly, consider a $D$-dimensional model in which the interesting or best-fitting region occupies a fraction $\epsilon$ of each dimension. A random scan would find

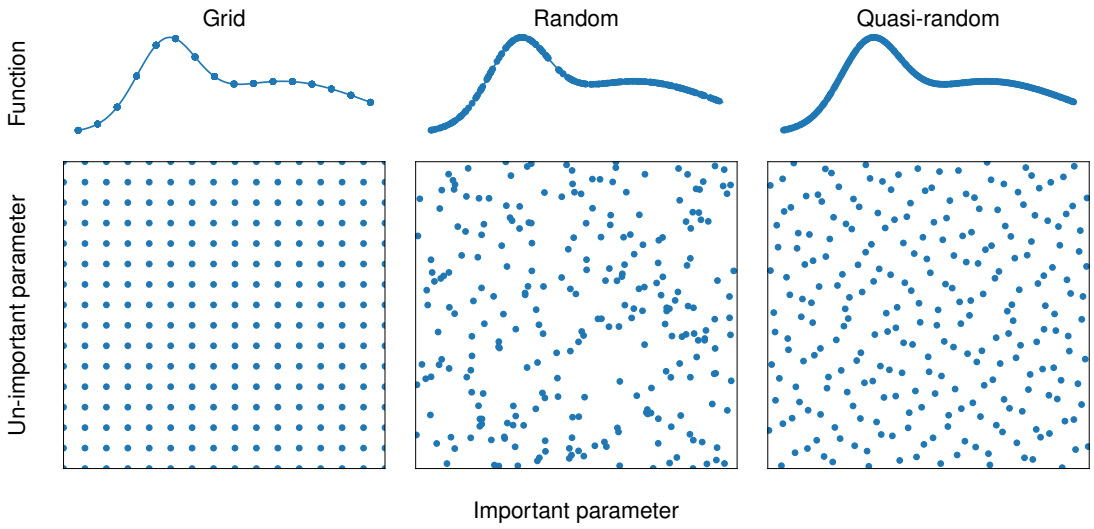


**Figure 3.** Grid, random and quasi-random sampling with 256 samples in two dimensions when the likelihood function is approximately one-dimensional. When the number of important parameters increases these methods perform poorly, as shown in Figure 4.

points in that region with an efficiency of $\epsilon^D$, i.e. random scans are exponentially inefficient. See Ref. 49 for further discussion and examples.

These issues can be addressed by using more sophisticated algorithms that, for example, preferentially explore areas of the parameter space where the likelihood is larger. Which algorithm is best suited for a given study depends on the goal of the analysis. For Bayesian inference, it is common to draw samples from the posterior distribution or compute an integral over the model's parameter space, relevant for Bayesian model selection. See Ref. 50 for a review of Bayesian computation. For frequentist inference, one might want to determine the global optimum and obtain samples from any regions in which the likelihood function was moderate. This can be more challenging than Bayesian computation. In particular, algorithms for Bayesian computation might not be appropriate optimizers. For example, Markov chain Monte Carlo methods draw from the posterior. In high-dimensions, the bulk of the posterior probability (the typical set) often lies well away from the maximum likelihood. This is another manifestation of the curse of dimensionality.

In Figure 4 we illustrate one such algorithm that overcomes the deficiencies of random and grid sampling and is suitable for frequentist inference. Here we assume that the logarithm of the likelihood function is given by a four-dimensional Rosenbrock function[51]

$$-2\ln\mathcal{L}(\boldsymbol{x}) = 2\sum_{i=1}^{3} f(x_i, x_{i+1}), \quad \text{where } f(a,b) = (1-a)^2 + 100\,(b-a^2)^2\,. \tag{1}$$

This is a challenging likelihood function with a global maximum at $x_i = 1$ ($i = 1,2,3,4$). We show samples found with $-2\ln\mathcal{L}(\boldsymbol{x}) \leq 5.99$. This constraint corresponds to the two-dimensional 95% confidence region, which in the $(x_1, x_2)$ plane has a banana-like shape (red contour). We find the points using uniform random sampling from $-5$ to 5 for each parameter (orange dots), using a grid scan (yellow dots), and using an implementation of the differential evolution algorithm[52,53] operating inside the same limits (blue dots). With only $2 \times 10^5$ likelihood calls, the differential evolution scan finds more than 11,500 points in the high-likelihood region,[11] whereas in $10^7$ tries the random scan finds only 7 high-likelihood samples, and the grid scan just 10. The random and grid scans would need over $10^{10}$ likelihood calls to obtain a similar number of high-likelihood points as obtained by differential evolution in just $2 \times 10^5$ evaluations. If likelihood calls are expensive and dominate the run-time, this could make differential evolution about $10^5$ times faster.

*Recommendation:* Use efficient algorithms to analyse parameter spaces, rather than grid or random scans. The

[11] We used a population size of 50 and stopped once the coefficient of variation of the fitness of the population dropped below 1%. See the associated code for the complete settings.[6]

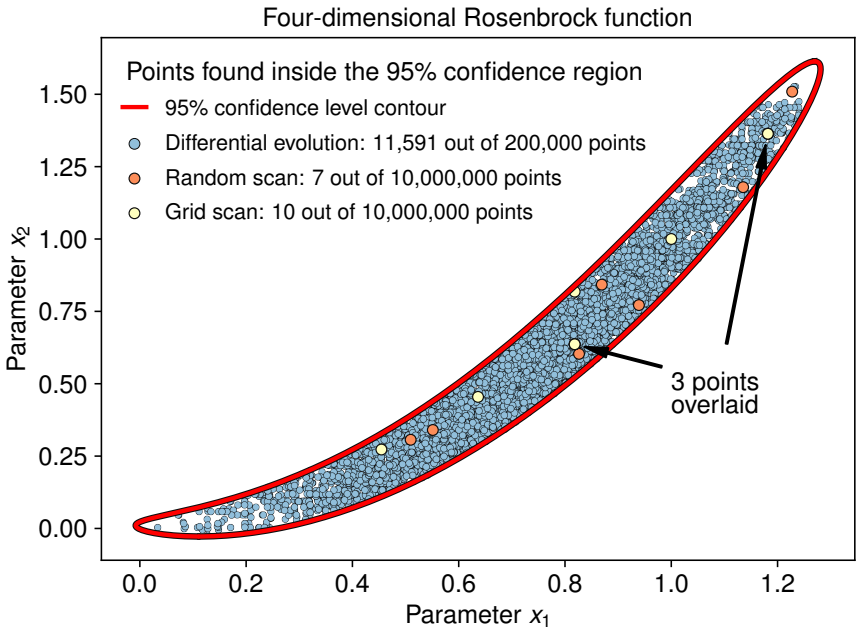


**Figure 4.** Points found inside the 95% confidence region of the likelihood function, in a two-dimensional plane of the four-dimensional Rosenbrock problem. Points are shown from scans using differential evolution (blue), random sampling (orange) and grid sampling (yellow). For reference, we also show the actual 95% confidence level contour of the likelihood function (red). Note that due to the projection of the four-dimensional space down to just two dimensions, two of the points shown from the grid sampler actually consist of three points each in the full four-dimensional space.

choice of algorithm should depend on the goal. Good examples for Bayesian analyses are Markov chain Monte Carlo[54,55] and nested sampling.[56] Good examples for maximizing and exploring the likelihood are simulated annealing,[57] differential evolution,[52] genetic algorithms[58] and local optimizers such as Nelder-Mead.[59] These are widely available in various public software packages.[53,60–66]

## 4 Problems with model testing

Overlaying confidence regions and performing random scans are straightforward methods for "hypothesis tests" of physical theories with many parameters or testable predictions. For example, it is tempting to say that a model is excluded if a uniform random or grid scan finds no samples for which the experimental predictions lie inside every 95% confidence region. This procedure is, however, prone to misinterpretation: just as in Section 2, it severely under-estimates error rates, and, just as in Section 3, it easily misses solutions.

Testing and comparing individual models in a statistically defensible manner is challenging and contentious. On the frequentist side, one can calculate a global $p$-value: the probability of obtaining data as extreme or more extreme than observed, if the model in question is true. The $p$-value features in two distinct statistical approaches:[67] first, the $p$-value may be interpreted as a measure of evidence against a model.[68] See Refs. 69–73 for discussion of this approach. Second, we may use the $p$-value to control the rate at which we would wrongly reject the model when it was true.[74] If we reject when $p < \alpha$, we would wrongly reject at a rate $\alpha$. In particle physics, we adopt the $5\sigma$ threshold, corresponding to $\alpha \simeq 10^{-7}$.[75] When we compute $p$-values, we should take into account all the tests that we might have performed. In the context of searches for new particles, this is known as the look-elsewhere effect. Whilst calculations can be greatly simplified by using asymptotic formulae,[76,77] bear in mind that they may not apply.[30] Also, care must be taken to avoid common misinterpretations of the $p$-value.[78,79] For example, the $p$-value is not the probability of the null hypothesis, or the probability that the observed data were produced by chance alone, or the probability of the observed data given the null hypothesis, or the rate at which we would wrongly reject the null hypothesis when it was true.

On the Bayesian side, one can perform Bayesian model comparison[1,80] to find any change brought about by data

to the relative plausibility of two different models. The factor that updates the relative plausibility of two models is called a Bayes factor. The Bayes factor is a ratio of integrals that may be challenging to compute in high-dimensional models. Just as in Bayesian parameter inference, this requires constructing priors for the parameters of the two models, permitting one to coherently incorporate prior information. In this setting, however, inferences may be strongly prior dependent, even in cases with large data sets and where seemingly uninformative priors are used.[81,82] This sensitivity can be particularly problematic in high-dimensional models. Unfortunately, there is no unique notion of an uninformative prior representing a state of indifference about a parameter,[83] though in special cases symmetry considerations may help.[84]

Neither of these approaches is simple, either philosophically or computationally, and the task of model testing and comparison is in general full of subtleties. For example, they depend differently on the amount of data collected which leads to somewhat paradoxical differences between them.[1,85,86] See Refs. 87–91 for recent discussions in other scientific settings. It is worth noting that there are connections between model testing and parameter inference in the case of nested models, i.e. when a model can be viewed as a subset of the parameter space of some larger, "full" model. A hypothesis test of a nested model can be equivalent to whether it lies inside a confidence region in the full model.[92,93] Similarly, the Bayes factor between nested models can be found from parameter inference in the full model alone through the Savage-Dickey ratio.[94] There are, furthermore, approaches beyond Bayesian model comparison and frequentist model testing that we do not discuss here.

*Recommendation:* In Bayesian analyses, carefully consider the choice of priors, their potential impact particularly in high-dimensions and check the prior sensitivity. In frequentist analyses, consider the look-elsewhere effect, check the validity of any asymptotic formulae and take care to avoid common misinterpretations of the $p$-value. If investigation of such subtleties fall outside the scope of the analysis, refrain from making strong statements on the overall validity of the theory under study.

## 5 Summary

As first steps towards addressing the challenges posed by physical theories with many parameters and many testable predictions, we make three recommendations: *i*) construct a composite likelihood that combines constraints from individual experiments, *ii*) use adaptive sampling algorithms (ones that target the interesting regions) to efficiently sample the parameter spaces, and *iii*) avoid strong statements on the viability of a theory unless a proper model test has been performed. The second recommendation can be easily achieved through the use of any one of a multitude of publicly-available implementations of efficient sampling algorithms (for examples see Section 3). For the first recommendation, composite likelihoods are often relatively simple to construct, and can be as straightforward as a product of Gaussians for multiple independent measurements. Even for cases where constructing the composite likelihood is more complicated, software implementations are often publicly available already.[95–103]

Given the central role of the likelihood function in analysing experimental data, it is in the interest of experimental collaborations to make their likelihood functions (or a reasonable approximation) publicly available to truly harness the full potential of their results when confronted with new theories. Even for large and complex datasets, e.g. those from the Large Hadron Collider, there exist various recommended methods for achieving this goal.[104–106]

Our recommendations can be taken separately when only one of the challenges exists, or where addressing them all is impractical. However, when confronted with both high-dimensional models and a multitude of relevant experimental constraints, we recommend that they are used together to maximise the validity and efficiency of analyses.

## References


1. Jeffreys, H. *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences (Oxford University Press, 1939).
2. Robert, C. & Casella, G. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Stat. Sci.* **26**, 102 – 115, doi:10.1214/10-STS351 (2011). [arXiv:0808.2902].
3. Chatrchyan, S. *et al.* Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. *Phys. Lett. B* **716**, 30–61, doi:10.1016/j.physletb.2012.08.021 (2012). [arXiv:1207.7235].
4. Aad, G. *et al.* Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716**, 1–29, doi:10.1016/j.physletb.2012.08.020 (2012). [arXiv:1207.7214].

5. Baak, M. *et al.* The global electroweak fit at NNLO and prospects for the LHC and ILC. *Eur. Phys. J. C* **74**, 3046, doi:10.1140/epjc/s10052-014-3046-5 (2014). [arXiv:1407.3792].
6. GAMBIT Collaboration. Supplementary code: Simple and statistically sound recommendations for analysing physical theories, doi:10.5281/zenodo.4322283. This DOI represents all versions, and will always resolve to the latest one.
7. D'Agostini, G. *Bayesian Reasoning In Data Analysis: A Critical Introduction* (World Scientific Publishing Company, 2003).
8. Gregory, P. *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge University Press, 2005).
9. Sivia, D. & Skilling, J. *Data Analysis: A Bayesian Tutorial* (Oxford University Press, 2006).
10. Trotta, R. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemp. Phys.* **49**, 71–104, doi:10.1080/00107510802066753 (2008). [arXiv:0803.4089].
11. von der Linden, W., Dose, V. & von Toussaint, U. *Bayesian Probability Theory: Applications in the Physical Sciences* (Cambridge University Press, 2014).
12. Bailer-Jones, C. *Practical Bayesian Inference: A Primer for Physical Scientists* (Cambridge University Press, 2017).
13. Lyons, L. *Statistics for Nuclear and Particle Physicists* (Cambridge University Press, 1989).
14. Cowan, G. *Statistical Data Analysis* (Clarendon Press, 1998).
15. James, F. *Statistical Methods in Experimental Physics* (World Scientific, 2006).
16. Behnke, O., Kröninger, K., Schott, G. & Schörner-Sadenius, T. *Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods* (Wiley, 2013).
17. Cousins, R. D. What is the likelihood function, and how is it used in particle physics? *arXiv preprint* (2020). CERN EP Newsletter, [arXiv:2010.00356].
18. Berger, J. & Wolpert, R. *The Likelihood Principle*, vol. 6 of *Lecture notes – monographs series* (Institute of Mathematical Statistics, 1988), second edn.
19. Brehmer, J. & Cranmer, K. Simulation-based inference methods for particle physics. *arXiv preprint* (2020). [arXiv:2010.06439].
20. Marrodán Undagoitia, T. & Rauch, L. Dark matter direct-detection experiments. *J. Phys. G* **43**, 013001, doi:10.1088/0954-3899/43/1/013001 (2016). [arXiv:1509.08767].
21. Bridges, M. *et al.* A Coverage Study of the CMSSM Based on ATLAS Sensitivity Using Fast Neural Networks Techniques. *JHEP* **03**, 012, doi:10.1007/JHEP03(2011)012 (2011). [arXiv:1011.4306].
22. Akrami, Y., Savage, C., Scott, P., Conrad, J. & Edsjö, J. Statistical coverage for supersymmetric parameter estimation: a case study with direct detection of dark matter. *JCAP* **7**, 2, doi:10.1088/1475-7516/2011/07/002 (2011). [arXiv:1011.4297].
23. Strege, C., Trotta, R., Bertone, G., Peter, A. H. G. & Scott, P. Fundamental statistical limitations of future dark matter direct detection experiments. *Phys. Rev. D* **86**, 023507, doi:10.1103/PhysRevD.86.023507 (2012). [arXiv:1201.3631].
24. Neyman, J. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philos. Trans. Roy. Soc. Lond. Ser. A* **236**, 333–380, doi:10.1098/rsta.1937.0005 (1937).
25. Cousins, R. D., Hymes, K. E. & Tucker, J. Frequentist evaluation of intervals estimated for a binomial parameter and for the ratio of Poisson means. *Nucl. Instruments Methods Phys. Res. A* **612**, 388–398, doi:10.1016/j.nima.2009.10.156 (2010). [arXiv:0905.3831].
26. Rolke, W. A., Lopez, A. M. & Conrad, J. Limits and confidence intervals in the presence of nuisance parameters. *Nucl. Instrum. Meth. A* **551**, 493–503, doi:10.1016/j.nima.2005.05.068 (2005). [arXiv:physics/0403059].
27. Punzi, G. Ordering algorithms and confidence intervals in the presence of nuisance parameters. In *Statistical Problems in Particle Physics, Astrophysics and Cosmology*, doi:10.1142/9781860948985_0019 (2005). [arXiv:physics/0511202].
28. Zyla, P. *et al.* Review of Particle Physics. *PTEP* **2020**, 083C01, chap. 40.4.2, doi:10.1093/ptep/ptaa104 (2020).
29. Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**, 60–62, doi:10.1214/aoms/1177732360 (1938).

**30.** Algeri, S., Aalbers, J., Dundas Morå, K. & Conrad, J. Searching for new physics with profile likelihoods: Wilks and beyond. *Nat. Rev. Phys* **2**, 245–252, doi:10.1038/s42254-020-0169-5 (2020). [arXiv:1911.10237].

**31.** Read, A. L. Presentation of search results: The CL(s) technique. *J. Phys. G* **28**, 2693–2704, doi:10.1088/0954-3899/28/10/313 (2002).

**32.** Rubin, D. B. & Schenker, N. Efficiently simulating the coverage properties of interval estimates. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **35**, 159–167, doi:10.2307/2347266 (1986).

**33.** Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D. & Wagenmakers, E.-J. The fallacy of placing confidence in confidence intervals. *Psychon. Bull. & Rev.* **23**, 103–123, doi:10.3758/s13423-015-0947-8 (2016).

**34.** Feldman, G. J. & Cousins, R. D. A Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D* **57**, 3873–3889, doi:10.1103/PhysRevD.57.3873 (1998). [arXiv:physics/9711021].

**35.** Junk, T. R. & Lyons, L. Reproducibility and Replication of Experimental Particle Physics Results. *Harv. Data Sci. Rev.* **2**, doi:10.1162/99608f92.250f995b (2020). [arXiv:2009.06864].

**36.** Amhis, Y. S. *et al.* Averages of $b$-hadron, $c$-hadron, and $\tau$-lepton properties as of 2018. *Eur. Phys. J. C* **81**, 226, doi:10.1140/epjc/s10052-020-8156-7 (2021). [arXiv:1909.12524].

**37.** Zyla, P. *et al.* Review of Particle Physics. *PTEP* **2020**, 083C01, chap. 40.2.1, doi:10.1093/ptep/ptaa104 (2020).

**38.** Ciuchini, M. *et al.* 2000 CKM triangle analysis: A Critical review with updated experimental inputs and theoretical parameters. *JHEP* **07**, 013, doi:10.1088/1126-6708/2001/07/013 (2001). [arXiv:hep-ph/0012308].

**39.** Ruiz de Austri, R., Trotta, R. & Roszkowski, L. A Markov chain Monte Carlo analysis of the CMSSM. *JHEP* **05**, 002, doi:10.1088/1126-6708/2006/05/002 (2006). [arXiv:hep-ph/0602028].

**40.** Allanach, B. C., Cranmer, K., Lester, C. G. & Weber, A. M. Natural priors, CMSSM fits and LHC weather forecasts. *JHEP* **08**, 023, doi:10.1088/1126-6708/2007/08/023 (2007). [arXiv:0705.0487].

**41.** Buchmueller, O. *et al.* Higgs and Supersymmetry. *Eur. Phys. J. C* **72**, 2020, doi:10.1140/epjc/s10052-012-2020-3 (2012). [arXiv:1112.3564].

**42.** Bechtle, P. *et al.* Constrained Supersymmetry after two years of LHC data: a global view with Fittino. *JHEP* **06**, 098, doi:10.1007/JHEP06(2012)098 (2012). [arXiv:1204.4199].

**43.** Fowlie, A. *et al.* The CMSSM Favoring New Territories: The Impact of New LHC Limits and a 125 GeV Higgs. *Phys. Rev. D* **86**, 075010, doi:10.1103/PhysRevD.86.075010 (2012). [arXiv:1206.0264].

**44.** Athron, P. *et al.* Global fits of GUT-scale SUSY models with GAMBIT. *Eur. Phys. J. C* **77**, 824, doi:10.1140/epjc/s10052-017-5167-0 (2017). [arXiv:1705.07935].

**45.** Balázs, C. *et al.* A comparison of optimisation algorithms for high-dimensional particle and astrophysics applications. *JHEP* **05**, 108, doi:10.1007/JHEP05(2021)108 (2021). [arXiv:2101.04525].

**46.** Balázs, C. *et al.* ColliderBit: a GAMBIT module for the calculation of high-energy collider observables and likelihoods. *Eur. Phys. J. C* **77**, 795, doi:10.1140/epjc/s10052-017-5285-8 (2017). [arXiv:1705.07919].

**47.** Bergstra, J. & Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).

**48.** Bellman, R. *Adaptive Control Processes: A Guided Tour*. Princeton Legacy Library (Princeton University Press, 1961).

**49.** Blum, A., Hopcroft, J. & Kannan, R. *Foundations of data science* (Cambridge University Press, 2020). Chap. 2. High-Dimensional Space.

**50.** Martin, G. M., Frazier, D. T. & Robert, C. P. Computing Bayes: Bayesian Computation from 1763 to the 21st Century. *arXiv e-prints* (2020). [arXiv:2004.06425].

**51.** Rosenbrock, H. H. An Automatic Method for Finding the Greatest or Least Value of a Function. *The Comput. J.* **3**, 175–184, doi:10.1093/comjnl/3.3.175 (1960).

**52.** Storn, R. & Price, K. Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **11**, 341–359, doi:10.1023/A:1008202821328 (1997).

**53.** Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272, doi:10.1038/s41592-019-0686-2 (2020).

**54.** Hogg, D. W. & Foreman-Mackey, D. Data analysis recipes: Using Markov Chain Monte Carlo. *Astrophys. J. Suppl.* **236**, 11, doi:10.3847/1538-4365/aab76e (2018). [arXiv:1710.06068].

**55.** Brooks, S., Gelman, A., Jones, G. & Meng, X. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods (CRC Press, 2011).

**56.** Skilling, J. Nested sampling for general Bayesian computation. *Bayesian Analysis* **1**, 833–859, doi:10.1214/06-BA127 (2006).

**57.** Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680, doi:10.1126/science.220.4598.671 (1983).

**58.** Charbonneau, P. Genetic Algorithms in Astronomy and Astrophysics. *ApJS* **101**, 309, doi:10.1086/192242 (1995).

**59.** Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *The Comput. J.* **7**, 308–313, doi:10.1093/comjnl/7.4.308 (1965).

**60.** Speagle, J. S. dynesty: A Dynamic Nested Sampling Package for Estimating Bayesian Posteriors and Evidences. *Mon. Not. Roy. Astron. Soc.* doi:10.1093/mnras/staa278 (2020). [arXiv:1904.02180].

**61.** Feroz, F., Hobson, M. P. & Bridges, M. MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Mon. Not. Roy. Astron. Soc.* **398**, 1601–1614, doi:10.1111/j.1365-2966.2009.14548.x (2009). [arXiv:0809.3437].

**62.** Handley, W. J., Hobson, M. P. & Lasenby, A. N. PolyChord: nested sampling for cosmology. *Mon. Not. Roy. Astron. Soc.* **450**, L61–L65, doi:10.1093/mnrasl/slv047 (2015). [arXiv:1502.01856].

**63.** Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: The MCMC Hammer. *Publ. Astron. Soc. Pac.* **125**, 306–312, doi:10.1086/670067 (2013). [arXiv:1202.3665].

**64.** Martinez, G. D. *et al.* Comparison of statistical sampling methods with ScannerBit, the GAMBIT scanning module. *Eur. Phys. J. C* **77**, 761, doi:10.1140/epjc/s10052-017-5274-y (2017). [arXiv:1705.07959].

**65.** James, F. & Roos, M. Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations. *Comput. Phys. Commun.* **10**, 343–367, doi:10.1016/0010-4655(75)90039-9 (1975).

**66.** Dembinski, H. *et al.* scikit-hep/iminuit: v1.4.9, doi:10.5281/zenodo.3951328 (2020).

**67.** Hubbard, R. & Bayarri, M. J. Confusion over measures of evidence ($p$'s) versus errors ($\alpha$'s) in classical statistical testing. *Am. Stat.* **57**, 171–178, doi:10.1198/0003130031856 (2003).

**68.** Fisher, R. A. *Statistical Methods for Research Workers* (Oliver and Boyd, 1925).

**69.** Hubbard, R. & Lindsay, R. M. Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychol.* **18**, 69–88, doi:10.1177/0959354307086923 (2008).

**70.** Schervish, M. J. P values: What they are and what they are not. *Am. Stat.* **50**, 203–206, doi:10.1080/00031305.1996.10474380 (1996).

**71.** Berger, J. O. & Sellke, T. Testing a point null hypothesis: The irreconcilability of p values and evidence. *J. Am. Stat. Assoc.* **82**, 112–122, doi:10.1080/01621459.1987.10478397 (1987).

**72.** Senn, S. Two cheers for p-values? *J. Epidemiol. Biostat.* **6**, 193–204, doi:10.1080/135952201753172953 (2001).

**73.** Murtaugh, P. A. In defense of P values. *Ecology* **95**, 611–617, doi:10.1890/13-0590.1 (2014).

**74.** Neyman, J. & Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. Lond. Ser. A* **231**, 289–337, doi:10.1098/rsta.1933.0009 (1933).

**75.** Lyons, L. Discovering the Significance of 5 sigma. *arXiv preprint* (2013). [arXiv:1310.1284].

**76.** Cowan, G., Cranmer, K., Gross, E. & Vitells, O. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C* **71**, 1554, doi:10.1140/epjc/s10052-011-1554-0 (2011). [Erratum: *Eur. Phys. J. C* **73**, 2501 (2013), doi:10.1140/epjc/s10052-011-1554-0], [arXiv:1007.1727].

**77.** Gross, E. & Vitells, O. Trial factors for the look elsewhere effect in high energy physics. *Eur. Phys. J. C* **70**, 525–530, doi:10.1140/epjc/s10052-010-1470-8 (2010). [arXiv:1005.1891].

**78.** Goodman, S. A dirty dozen: Twelve p-value misconceptions. *Semin. Hematol.* **45**, 135–140, doi:10.1053/j.seminhematol.2008.04.003 (2008).

**79.** Greenland, S. *et al.* Statistical tests, *p* values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* **31**, 337–350, doi:10.1007/s10654-016-0149-3 (2016).

**80.** Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795, doi:10.1080/01621459.1995.10476572 (1995).

**81.** Berger, J. O. & Pericchi, L. R. Objective Bayesian methods for model selection: Introduction and comparison. *IMS Lecture Notes – Monograph Series* **38**, 135–207, doi:10.1214/lnms/1215540968 (2001).

**82.** Cousins, R. D. Comment on 'Bayesian Analysis of Pentaquark Signals from CLAS Data', with Response to the Reply by Ireland and Protopopsecu. *Phys. Rev. Lett.* **101**, 029101, doi:10.1103/PhysRevLett.101.029101 (2008). [arXiv:0807.1330].

**83.** Kass, R. E. & Wasserman, L. The Selection of Prior Distributions by Formal Rules. *J. Am. Stat. Assoc.* **91**, 1343–1370, doi:10.1080/01621459.1996.10477003 (1996).

**84.** Jaynes, E. T. Prior probabilities. *IEEE Transactions on Syst. Sci. Cybern.* **4**, 227–241, doi:10.1109/TSSC.1968.300117 (1968).

**85.** Lindley, D. V. A statistical paradox. *Biometrika* **44**, 187–192, doi:10.1093/biomet/44.1-2.187 (1957).

**86.** Cousins, R. D. The Jeffreys-Lindley paradox and discovery criteria in high energy physics. *Synthese* **194**, 395–432, doi:10.1007/s11229-014-0525-z, 10.1007/s11229-015-0687-3 (2017). [arXiv:1310.3791].

**87.** Wagenmakers, E.-J. A practical solution to the pervasive problems of *p* values. *Psychon. Bull. & Rev.* **14**, 779–804, doi:10.3758/BF03194105 (2007).

**88.** Lakens, D. The practical alternative to the *p* value is the correctly used *p* value. *Perspectives on Psychol. Sci.* doi:10.1177/1745691620958012 (2021).

**89.** Benjamin, D. J. *et al.* Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10, doi:10.1038/s41562-017-0189-z (2018).

**90.** McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. Abandon statistical significance. *The Am. Stat.* **73**, 235–245, doi:10.1080/00031305.2018.1527253 (2019).

**91.** Lakens, D. *et al.* Justify your alpha. *Nat. Hum. Behav.* **2**, 168–171, doi:10.1038/s41562-018-0311-x (2018).

**92.** Kendall, M., Stuart, A., Ord, J. & Arnold, S. *Kendall's Advanced Theory of Statistics*, vol. 2A, chap. 21 (Wiley, 2009), sixth edn.

**93.** Cousins, R. D. Lectures on Statistics in Theory: Prelude to Statistics in Practice. *arXiv e-prints* (2018). See Sec. 7.4, [arXiv:1807.05996].

**94.** Dickey, J. M. The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals Math. Stat.* **42**, 204–223 (1971).

**95.** Athron, P. *et al.* GAMBIT: The Global and Modular Beyond-the-Standard-Model Inference Tool. *Eur. Phys. J. C* **77**, 784, doi:10.1140/epjc/s10052-017-5321-8 (2017). [Addendum: *Eur. Phys. J. C* **78**, 98 (2018), doi:10.1140/epjc/s10052-017-5513-2], [arXiv:1705.07908].

**96.** De Blas, J. *et al.* `HEPfit`: a Code for the Combination of Indirect and Direct Constraints on High Energy Physics Models. *Eur. Phys. J. C* **80**, 456, doi:10.1140/epjc/s10052-020-7904-z (2019). [arXiv:1910.14012].

**97.** Brinckmann, T. & Lesgourgues, J. MontePython 3: boosted MCMC sampler and other features, doi:10.1016/j.dark.2018.100260 (2019). [arXiv:1804.07261].

**98.** Bhom, J. & Chrzaszcz, M. HEPLike: an open source framework for experimental likelihood evaluation. *Comput. Phys. Commun.* **254**, 107235, doi:10.1016/j.cpc.2020.107235 (2020). [arXiv:2003.03956].

**99.** Huang, X., Tsai, Y.-L. S. & Yuan, Q. LikeDM: likelihood calculator of dark matter detection. *Comput. Phys. Commun.* **213**, 252–263, doi:10.1016/j.cpc.2016.12.015 (2017). [arXiv:1603.07119].

**100.** Simplified likelihood for the re-interpretation of public CMS results. Tech. Rep. CMS-NOTE-2017-001, CERN, Geneva (2017).

**101.** Aghanim, N. *et al.* Planck 2018 results. V. CMB power spectra and likelihoods. *Astronomy and Astrophysics* **641**, A5, doi:10.1051/0004-6361/201936386 (2020). [arXiv:1907.12875].

**102.** Aartsen, M. G. *et al.* Improved limits on dark matter annihilation in the Sun with the 79-string IceCube detector and implications for supersymmetry. *JCAP* **04**, 022, doi:10.1088/1475-7516/2016/04/022 (2016). [arXiv:1601.00653].

**103.** Scott, P., Savage, C., Edsjö, J. & the IceCube Collaboration: R. Abbasi et al. Use of event-level neutrino telescope data in global fits for theories of new physics. *JCAP* **11**, 57, doi:10.1088/1475-7516/2012/11/057 (2012). [arXiv:1207.0810].

**104.** Cousins, R. D. Comments on methods for setting confidence limits. *Workshop on Confidence Limits* doi:10.5170/CERN-2000-005.49 (2000). See point 5, p57.

**105.** Vischia, P. Reporting results in High Energy Physics publications: A manifesto. *Rev. Phys.* **5**, 100046, doi:10.1016/j.revip.2020.100046 (2020). [arXiv:1904.11718].

**106.** Abdallah, W. *et al.* Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2. *SciPost Phys.* **9**, 022, doi:10.21468/SciPostPhys.9.2.022 (2020). [arXiv:2003.07868].

**107.** Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. Sci. & Eng.* **9**, 90–95, doi:10.1109/MCSE.2007.55 (2007).

## Acknowledgements

BCA has been partially supported by the UK Science and Technology Facilities Council (STFC) Consolidated HEP theory grants ST/P000681/1 and ST/T000694/1. PA is supported by Australian Research Council (ARC) Future Fellowship FT160100274, and PS by FT190100814. PA, CB, TEG and MW are supported by ARC Discovery Project DP180102209. CB and YZ are supported by ARC Centre of Excellence CE110001104 (Particle Physics at the Tera-scale) and WS and MW by CE200100008 (Dark Matter Particle Physics). ABe is supported by F.N.R.S. through the F.6001.19 convention. ABuc is supported by the Royal Society grant UF160548. JECM is supported by the Carl Trygger Foundation grant no. CTS 17:139. JdB acknowledges support by STFC under grant ST/P001246/1. JE was supported in part by the STFC (UK) and by the Estonian Research Council. BF was supported by EU MSCA-IF project 752162 – DarkGAMBIT. MF and FK are supported by the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center TRR 257 "Particle Physics Phenomenology after the Higgs Discovery" under Grant 396021762 – TRR 257 and FK also under the Emmy Noether Grant No. KA 4662/1-1. AF is supported by an NSFC Research Fund for International Young Scientists grant 11950410509. SHe was supported in part by the MEINCOP (Spain) under contract PID2019-110058GB-C21 and in part by the Spanish Agencia Estatal de Investigación (AEI) through the grant IFT Centro de Excelencia Severo Ochoa SEV-2016-0597. SHoof is supported by the Alexander von Humboldt Foundation. SHoof and MTP are supported by the Federal Ministry of Education and Research of Germany (BMBF). KK is supported in part by the National Science Centre (Poland) under research Grant No. 2017/26/E/ST2/00470, LR under No. 2015/18/A/ST2/00748, and EMS under No. 2017/26/D/ST2/00490. LR and ST are supported by grant AstroCeNT: Particle Astrophysics Science and Technology Centre, carried out within the International Research Agendas programme of the Foundation for Polish Science financed by the European Union under the European Regional Development Fund. MLM acknowledges support from NWO (Netherlands). SM is supported by JSPS KAKENHI Grant Number 17K05429. The work of K.A.O. was supported in part by DOE grant DE-SC0011842 at the University of Minnesota. JJR is supported by the Swedish Research Council, contract 638-2013-8993. KS was partially supported by the National Science Centre, Poland, under research grants 2017/26/E/ST2/00135 and the Beethoven grants DEC-2016/23/G/ST2/04301. AS is supported by MIUR research grant No. 2017X7X85K and INFN. WS is supported by KIAS Individual Grant (PG084201) at Korea Institute for Advanced Study. ST is partially supported by the Polish Ministry of Science and Higher Education through its scholarship for young and outstanding scientists (decision no. 1190/E-78/STYP/14/2019). RT was partially supported by STFC under grant number ST/T000791/1. The work of MV is supported by the NSF Grant No. PHY-1915005. ACV is supported by the Arthur B. McDonald Canadian Astroparticle Physics Research Institute. Research at Perimeter Institute is supported by the Government of Canada through the Department of Innovation, Science, and Economic Development, and by the Province of Ontario through MEDJCT. LW is supported by the National Natural Science Foundation of China (NNSFC) under grant Nos. 117050934, by Jiangsu Specially Appointed Professor Program.

## Author contributions

The project was led by AF and in preliminary stages by BF and FK. ABe, AF, SHoof, AK, PSc and WS contributed to creating the figures. PA, CB, TB, ABe, ABuc, AF, TEG, SHoof, AK, JECM, MTP, AR, PSc, ACV and YZ contributed to writing. WH and FK performed official internal reviews of the article. All authors read, endorsed and discussed the content and recommendations.

## Code availability

The figures were prepared with `matplotlib`.[107] We have made all scripts publicly available at Zenodo.[6]