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 strategies 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,5], 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,6,*], Tomás E. Gonzalo[a,5], 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], Sebastian Trojanowski[e,45,33], Roberto Trotta[9,48], Yue-Lin Sming Tsai[49,50], Jeriek Van den Abeele[a,13], Mauro Valli[d,51], Aaron C. Vincent[a,52,53,54], Georg Weiglein[b,41], Martin White[a,28], Peter Wienemann[c,8], Lei Wu[a,6], and Yang Zhang[a,5,55]

[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]School of Physics and Astronomy, Monash University, Melbourne, VIC 3800, Australia
[6]Department of Physics and Institute of Theoretical Physics, Nanjing Normal University, Nanjing, Jiangsu 210023, China
[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), Universidad Autónoma de Madrid Cantoblanco, E-28049 Madrid, Spain
[31]Institut für Astrophysik, 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]DESY, Notkestraße 85, D-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]SISSA International School for Advanced Studies, Via Bonomea 265, 34136, Trieste, Italy
[49]Key Laboratory of Dark Matter and Space Astronomy, Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210033, China
[50]Department of Physics, National Tsing Hua University, Hsinchu 300, Taiwan
[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]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 parameter estimation. Many models in particle physics, astrophysics and cosmology fall into one or both of these categories. These issues are often sidestepped with very simplistic and statistically unsound *ad hoc* methods, involving naive 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 rudimentary procedures, suggesting some 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 physicists with recommendations for reaching correct scientific conclusions, with only a modest increase in analysis burden.

Key points:

1. Models in particle physics, astrophysics and cosmology can have many free parameters, and assessing them often requires data from multiple experiments. As a result, simplistic methods widely used in the literature are not adequate and may lead to faulty and unreliable conclusions.

2. Overlaying confidence regions from multiple experiments is often a poor approximation of a combined confidence region and grossly underestimates error rates. We recommend constructing a composite likelihood function for all relevant experimental measurements and using it to find joint confidence or credible regions.

3. Uniform random and grid scans are not well suited for investigating high-dimensional parameter spaces. There are many dedicated algorithms available that are more efficient and yield more accurate results.

Website summary: Theories with many parameters and observable predictions present special challenges for rigorous scientific inference. We highlight problems with widely-used methods, and recommend easy-to-use alternatives for drawing reliable conclusions from data about novel or existing physical theories.

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 hypothesis testing and parameter estimation, respectively. Here we mainly focus on parameter estimation. This allows us to understand what a model could predict, and design future experiments to test it. On the theory side, inferring the parameters of a model allows us to construct theories that contain the model and naturally accommodate the observations.

Many analyses of 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. In this article, we point out why both of these methods give unreliable results, and give concrete recommendations for what should be done instead.

Our discussion covers both Bayesian methods,[1–6] in which one directly considers the plausibility of a model and regions of its parameter space, and frequentist methods,[7–10] 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, which tells us the probability of the observed data, assuming a particular model and a particular combination of numerical values for its unknown parameters.

2 Problems of overlaying exclusion limits

Experimental searches for new particles 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[11] 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 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.[12–14]

In the frequentist approach, if an experiment that measured a parameter were repeated over and over again, each repeat would lead to a slightly different confidence region for the measured parameter. The confidence level of the confidence region is the fraction of repeated experiments in which the resulting confidence region would contain the true parameter values.[15] For

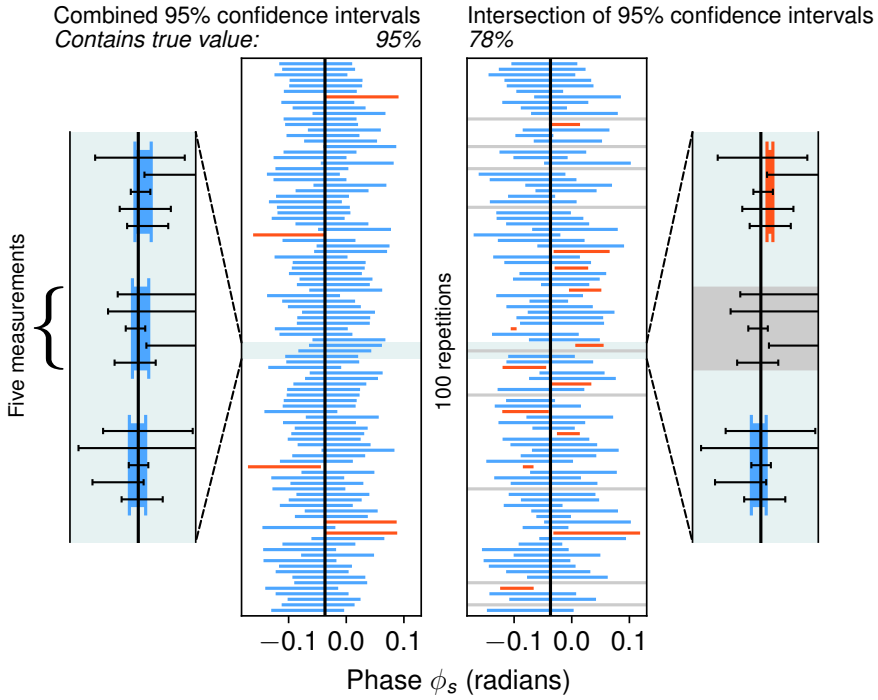

Figure 1. Confidence intervals in 100 pseudo-experiments, from the combination of five measurements (*left*) or from the intersection of five individual confidence intervals (*right*). We show the true value of ϕ_s with a vertical black line. Intervals that contain the true value are shown in blue; those 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.

example: a 95% confidence region contains 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%.* Confidence regions can often be found from the likelihood function alone. The analogous construct in Bayesian statistics is the credible region; a 95% credible region for a parameter contains the true value of the parameter with a probability of 95%.

When several experiments report measured 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. While this could be a reasonable rough estimate in some cases, it 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 e.g. a 5% error rate, giving an error rate of $1 - 0.95^n$.[17]

This issue is illustrated in fig. 1 using the B-physics observable ϕ_s, which is a well-measured phase characterising CP violation in B_s meson decays.[18] 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 $\phi_s = -0.0369$ 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 fig. 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 two-dimensional confidence regions. Clearly, rather than taking the intersection of reported results, one should combine

*Note though that confidence regions are sometimes defined to include the true parameter values in *at least* e.g. 95% of repeated experiments[16] and that in some cases the nominal confidence level may not hold in practice.

†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[18].

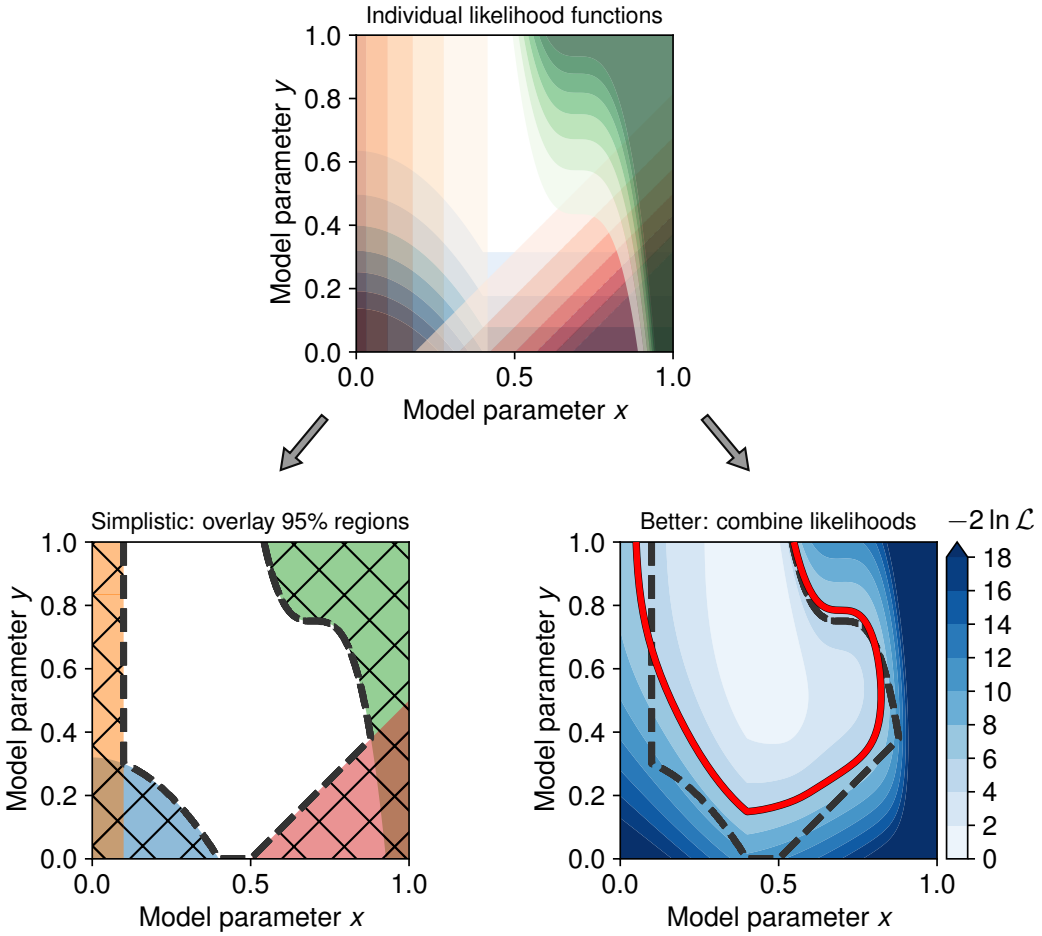

Figure 2. Starting from four individual likelihood functions (*top*), we compare overlaid 95% contours (*bottom left*) versus a combination of the likelihoods (*bottom right*; blue contours). The dashed black line in both bottom panels is the intersection of the limits from the individual likelihoods.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

likelihood functions from multiple experiments. Good examples can be found in the literature.[19–25]

In fig. 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 naive 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 of 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 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. In most cases this cannot be done analytically. The derivatives of the likelihood function are usually unavailable, and 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

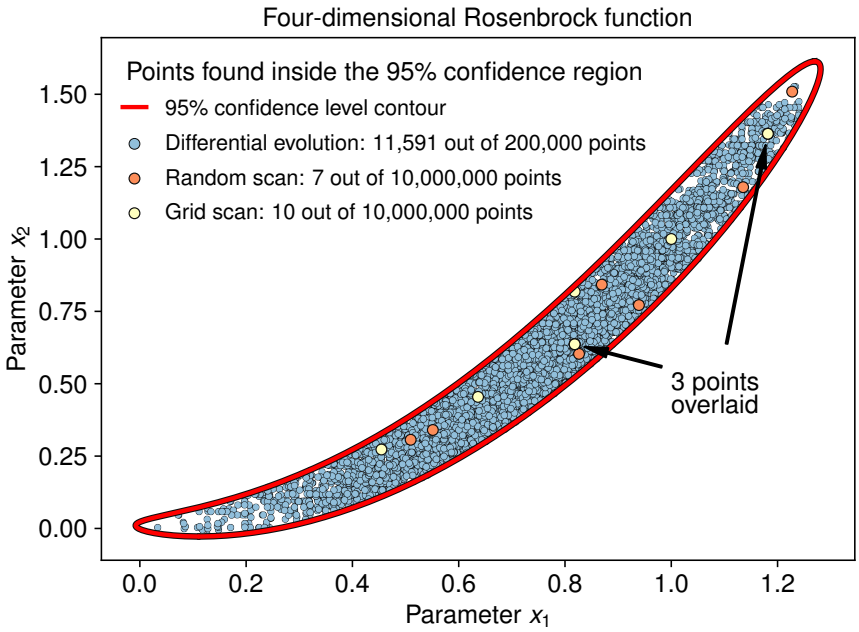

Figure 3. Points found inside the 95% confidence region of the likelihood function, in a two-dimensional plane of the four-dimensional Rosenbrock problem. Points are shown from scans using differential evolution (blue), random sampling (orange) and grid sampling (yellow). For reference, we also show 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.

space using a tractable number of evaluations of the likelihood function.

Random and grid scans are common 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 parameterisation of the model, which introduces a dependency on the choice of parameterisation. 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.

The efficiency of random and grid scans is also extremely poor. The "curse of dimensionality"[26] 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 ε of each dimension. A random scan would find points in that region with an efficiency of ε^D, i.e. random scans are exponentially inefficient.

These issues can be addressed by using more sophisticated sampling algorithms that adapt to the function that they are sampling, e.g., by preferentially moving to sample areas of the parameter space where the likelihood is larger. This is illustrated in fig. 3, where we assume that the logarithm of the likelihood function is given by a four-dimensional Rosenbrock function[27]

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

This is a challenging likelihood function with a global maximum at $x_i = 1$ ($i = 1,2,3,4$). We show samples found with $-2\ln\mathscr{L}(\boldsymbol{x}) \leq 5.99$. 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[28,29] operating inside the same limits (blue dots). With only 2×10^5 likelihood calls, the differential evolution scan finds more than 11 500 points in the high-likelihood region, 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 2×10^5 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 choice of algorithm should depend on your goal. Good examples for Bayesian analyses are Markov chain Monte Carlo[30] and nested sampling.[31] Good examples for maximizing the likelihood are simulated annealing,[32] differential evolution,[28] genetic algorithms[33] and local optimisers such as Nelder-Mead.[34] These are widely available in various public software packages.[29,35–41]

4 Discussion and Summary

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, misleading and prone to misinterpretation.

Testing and comparing individual models in a statistically defensible manner is challenging. 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. On the Bayesian side, one can perform Bayesian model comparison[42] to find any change brought about by data to the relative plausibility of two different models. Neither of these approaches is simple, and the task of model testing and comparison is in general full of subtleties.[43,44]

As first steps towards addressing the challenges posed by physical theories with many parameters and many testable predictions, we make two recommendations: *i*) construct a composite likelihood that combines constraints from individual experiments, and *ii*) use adaptive sampling algorithms (ones that target the interesting regions) to efficiently sample the parameter spaces. 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.[45–53]

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 big and complex data, as e.g. from the Large Hadron Collider, there are various recommended methods for achieving this goal.[54]

Our two recommendations can be taken separately when only one of the challenges exists, or where addressing both 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. D'Agostini, G. *Bayesian Reasoning In Data Analysis: A Critical Introduction* (World Scientific Publishing Company, 2003).
2. Gregory, P. *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge University Press, 2005).
3. Sivia, D. & Skilling, J. *Data Analysis: A Bayesian Tutorial* (Oxford University Press, 2006).
4. 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].
5. von der Linden, W., Dose, V. & von Toussaint, U. *Bayesian Probability Theory: Applications in the Physical Sciences* (Cambridge University Press, 2014).
6. Bailer-Jones, C. *Practical Bayesian Inference: A Primer for Physical Scientists* (Cambridge University Press, 2017).
7. Lyons, L. *Statistics for Nuclear and Particle Physicists* (Cambridge University Press, 1989).
8. Cowan, G. *Statistical Data Analysis* (Clarendon Press, 1998).
9. James, F. *Statistical Methods in Experimental Physics* (World Scientific, 2006).
10. Behnke, O., Kröninger, K., Schott, G. & Schörner-Sadenius, T. *Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods* (Wiley, 2013).

11. 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].

12. 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].

13. 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].

14. 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].

15. Neyman, J. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philos. Transactions Royal Soc. London. Ser. A, Math. Phys. Sci.* **236**, 333–380 (1937).

16. Zyla, P. *et al.* Review of Particle Physics. *PTEP* **2020**, 083C01, chap. 40.4.2, DOI: 10.1093/ptep/ptaa104 (2020).

17. Junk, T. R. & Lyons, L. Reproducibility and Replication of Experimental Particle Physics Results. *arXiv preprint* (2020). [arXiv:2009.06864].

18. Amhis, Y. S. *et al.* Averages of b-hadron, c-hadron, and τ-lepton properties as of 2018. *arXiv preprint* (2019). [arXiv:1909.12524].

19. 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].

20. 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].

21. 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].

22. 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].

23. 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].

24. 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].

25. 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].

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

27. 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).

28. 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).

29. 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).

30. 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).

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

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

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

34. 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).

35. 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].

36. 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].

37. 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].

38. 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].

39. 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].

40. 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).

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

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

43. 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).

44. 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].

45. 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].

46. 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].

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

48. 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].

49. 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].

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

51. 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].

52. 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 (2016). [arXiv:1601.00653].

53. 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].

54. 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].

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

56. GAMBIT Collaboration. Supplementary code: Simple and statistically sound strategies for analysing physical theories, DOI: 10.5281/zenodo.4322284 (2020).

Acknowledgments

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. 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. 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, 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.

Competing interests

The authors declare no competing interests.

Publisher's note

Code availability

The figures were originally prepared with `matplotlib`.[55] We have made the plotting scripts publicly available at `Zenodo`.[56]