Simple and statistically sound strategies for analysing physical theories

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

\textsuperscript{a}The GAMBIT Community
\textsuperscript{b}The MasterCode Collaboration
\textsuperscript{c}The Fittino Collaboration
\textsuperscript{d}HEPfit
\textsuperscript{e}BayesFits Group

\textsuperscript{1}Department of Physics, Shahid Beheshti University, Tehran, Iran
\textsuperscript{2}Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge, CB3 0HE, UK
\textsuperscript{3}Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge, CB3 0HA, UK
\textsuperscript{4}DAMTP, University of Cambridge, Cambridge, CB3 0WA, UK
\textsuperscript{5}School of Physics and Astronomy, Monash University, Melbourne, VIC 3800, Australia
\textsuperscript{6}Department of Physics and Institute of Theoretical Physics, Nanjing Normal University, Nanjing, Jiangsu 210023, China
\textsuperscript{7}Paul Scherrer Institut, CH-5232 Villigen, Switzerland
\textsuperscript{8}University of Bonn, Physikalisches Institut, Nussallee 12, D-53115 Bonn, Germany
\textsuperscript{9}Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK
\textsuperscript{10}Centre for Cosmology, Particle Physics and Phenomenology (CP3), Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium
\textsuperscript{11}Institute of Nuclear Physics, Polish Academy of Sciences, Krakow, Poland
\textsuperscript{12}School of Mathematics and Physics, The University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia
\textsuperscript{13}Department of Physics, University of Oslo, Box 1048, Blindern, N-0316 Oslo, Norway
\textsuperscript{14}School of Physics and Astronomy, University of Glasgow, University Place, Glasgow, G12 8QQ, UK
\textsuperscript{15}Institut für Theoretische Physik, Universität Heidelberg, Germany
\textsuperscript{16}Department of Physics and Astronomy, Uppsala University, Box 516, SE-751 20 Uppsala, Sweden
\textsuperscript{17}Oskar Klein Centre for Cosmoparticle Physics, AlbaNova University Centre, SE-10691 Stockholm, Sweden
\textsuperscript{18}Department of Physics, Weber State University, 1415 Edvalson St., Dept. 2508, Ogden, UT 84408, USA

\textsuperscript{1}Department of Physics, Shahid Beheshti University, Tehran, Iran
\textsuperscript{2}Department of Physics, University of Cambridge, JJ Thomson Avenue, Cambridge, CB3 0HE, UK
\textsuperscript{3}Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge, CB3 0HA, UK
\textsuperscript{4}DAMTP, University of Cambridge, Cambridge, CB3 0WA, UK
\textsuperscript{5}School of Physics and Astronomy, Monash University, Melbourne, VIC 3800, Australia
\textsuperscript{6}Department of Physics and Institute of Theoretical Physics, Nanjing Normal University, Nanjing, Jiangsu 210023, China
\textsuperscript{7}Paul Scherrer Institut, CH-5232 Villigen, Switzerland
\textsuperscript{8}University of Bonn, Physikalisches Institut, Nussallee 12, D-53115 Bonn, Germany
\textsuperscript{9}Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK
\textsuperscript{10}Centre for Cosmology, Particle Physics and Phenomenology (CP3), Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium
\textsuperscript{11}Institute of Nuclear Physics, Polish Academy of Sciences, Krakow, Poland
\textsuperscript{12}School of Mathematics and Physics, The University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia
\textsuperscript{13}Department of Physics, University of Oslo, Box 1048, Blindern, N-0316 Oslo, Norway
\textsuperscript{14}School of Physics and Astronomy, University of Glasgow, University Place, Glasgow, G12 8QQ, UK
\textsuperscript{15}Institut für Theoretische Physik, Universität Heidelberg, Germany
\textsuperscript{16}Department of Physics and Astronomy, Uppsala University, Box 516, SE-751 20 Uppsala, Sweden
\textsuperscript{17}Oskar Klein Centre for Cosmoparticle Physics, AlbaNova University Centre, SE-10691 Stockholm, Sweden
\textsuperscript{18}Department of Physics, Weber State University, 1415 Edvalson St., Dept. 2508, Ogden, UT 84408, USA
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
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%.\footnote{Note though that confidence regions are sometimes defined to include the true parameter values in at least e.g. 95% of repeated experiments\cite{16} and that in some cases the nominal confidence level may not hold in practice.} 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 $1 - 0.95^n$.\footnote{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.}

This issue is illustrated in fig. 1 using the $B$-physics observable $\phi_s$, which is a well-measured phase characterising CP violation in $B_s$ meson decays.\footnote{See eq. 91 and table 22 in\cite{18}.} We perform 10000 pseudo-experiments.\footnote{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...} 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.\footnote{We perform 10000 pseudo-experiments.} 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...
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” 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 points in that region with an efficiency of \( \epsilon^D \), 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\(^{27}\)

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

This is a challenging likelihood function with a global maximum at \( x_i = 1 \) (\( i = 1, 2, 3, 4 \)). We show samples found with \(-2 \ln \mathcal{L}(\mathbf{x}) \leq 5.99\). 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 \(2 \times 10^5\) 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 $2 \times 10^5$ 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 choice of algorithm should depend on your goal. Good examples for Bayesian analyses are Markov chain Monte Carlo and nested sampling. Good examples for maximizing the likelihood are simulated annealing, differential evolution, genetic algorithms and local optimisers such as Nelder-Mead. These are widely available in various public software packages.

**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 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.

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.

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.

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**


**Acknowledgments**

BCA has been partially supported by the UK Science and Technology Facilities Council (STFC) Consolidated HEP theory grants ST/P000681/1 and ST/T00694/1. 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 CE200100080.
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
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Code availability
The figures were originally prepared with matplotlib. We have made the plotting scripts publicly available at Zenodo.