



## City Research Online

### City, University of London Institutional Repository

---

**Citation:** He, Y-H., Jejjala, V. & Pontiggia, L. (2017). Patterns in Calabi-Yau Distributions. *Communications in Mathematical Physics*, 354(2), pp. 477-524. doi: 10.1007/s00220-017-2907-9

This is the published version of the paper.

This version of the publication may differ from the final published version.

---

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/18200/>

**Link to published version:** <https://doi.org/10.1007/s00220-017-2907-9>

**Copyright:** City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

**Reuse:** Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

---

---

---

City Research Online:

<http://openaccess.city.ac.uk/>

[publications@city.ac.uk](mailto:publications@city.ac.uk)

---



# Patterns in Calabi–Yau Distributions

Yang-Hui He<sup>1,2,3</sup>, Vishnu Jejjala<sup>4</sup>, Luca Pontiggia<sup>4</sup>

<sup>1</sup> School of Physics, NanKai University, Tianjin 300071, People’s Republic of China

<sup>2</sup> Department of Mathematics, City University, London EC1V 0HB, UK

<sup>3</sup> Merton College, University of Oxford, Oxford OX1 4JD, UK. E-mail: hey@maths.ox.ac.uk

<sup>4</sup> NiTheP, School of Physics, and Mandelstam Institute for Theoretical Physics, University of the Witwatersrand, 1 Jan Smuts Avenue, Johannesburg 2050, South Africa. E-mail: vishnu@neo.phys.wits.ac.za; lucaPontiggia@gmail.com

Received: 18 December 2015 / Accepted: 17 April 2017

Published online: 30 May 2017 – © The Author(s) 2017. This article is an open access publication

**Abstract:** We explore the distribution of topological numbers in Calabi–Yau manifolds, using the Kreuzer–Skarke dataset of hypersurfaces in toric varieties as a testing ground. While the Hodge numbers are well-known to exhibit mirror symmetry, patterns in frequencies of combination thereof exhibit striking new patterns. We find pseudo-Voigt and Planckian distributions with high confidence and exact fit for many substructures. The patterns indicate typicality within the landscape of Calabi–Yau manifolds of various dimension.

## Contents

1. Introduction	478
2. Calabi–Yau Threefolds	480
2.1 Analysis of $h^{1,1} - h^{1,2}$	481
2.1.1 A pseudo-Voigt fit	483
2.2 Analysis of $h^{1,1} + h^{1,2}$	488
2.2.1 A Planckian fit	488
2.3 The distribution of the Euler number	495
2.4 Goodness-of-fit	496
2.5 Implications for physics	501
3. Calabi–Yau Twofolds: K3 Surfaces	502
4. Calabi–Yau Fourfolds	502
5. Conclusions and Outlook	505
A. Appendix	507
A.1 Supplementary plots for the $h^{1,1} - h^{1,2}$ distribution	507
A.1.1 Plots for the odd distribution as counterparts to the even ones	507
A.1.2 Comparative plots	507
A.1.3 A first approximation to the data	510
A.1.4 Table of parameter values and statistics	510

A.2 Supplementary plots for the  $h^{1,1} + h^{1,2}$  distribution . . . . . 511  
 A.2.1 Plots for the odd distribution as counterparts to the even ones . . . . . 511  
 A.2.2 Table of parameter values, coefficient values and statistics . . . . . 512  
 A.3 Supplementary plots for the fourfold data . . . . . 515

**1. Introduction**

A Calabi–Yau  $n$ -fold is a Kähler manifold of  $n$  complex dimensions with a trivial canonical bundle. In superstring theory, it serves as a compactification manifold wherein a ten dimensional theory at high energies reduces to an effective theory in four spacetime dimensions. In particular, global  $SU(n)$  holonomy ensures that  $2^{1-n}$  of the original supersymmetry is preserved. Thus, confronted by the vacuum selection problem, Calabi–Yau compactifications present an avenue for Standard Model building, especially in the context of the heterotic string [1–4]. Indeed, the basis of the landscape is to consider flux compactifications on these geometries [5,6].

To facilitate this approach to a low-energy phenomenology derived from string theory, mathematicians and physicists have constructed large datasets of Calabi–Yau threefolds [7,9–22] as well as various refined analyses of properties thereof [28–35]. By far the largest database was constructed in a *tour de force* of algebraic geometry, combinatorics, physics, and computer algorithms by Kreuzer and Skarke based on the theorems of Batyrev and Borisov [9–14,36,37]. In short, these Calabi–Yau  $n$ -manifolds  $X_n$  are realized as a smooth hypersurface embedded in a toric variety  $A_{n+1}$  of complex dimension  $n + 1$ ; the Calabi–Yau condition simply translates to the requirement that the polytope defining  $A_{n+1}$  be **reflexive**. We will henceforth consider only such Calabi–Yau manifolds, of which there are a plethora.

Let us briefly recollect what all this means. The (possibly singular) toric variety  $A_{n+1}$  is specified by an integer polytope  $\Delta$  in  $\mathbb{R}^{n+1}$ , which is a collection of vertices (dimension 0) each of which is an  $(n + 1)$ -vector with integer entries and such that each pair of neighboring vertices defines an edge (dimension 1), each pair of edges defines a face (dimension 2), etc., all the way up to a facet (dimension  $n$ ). Alternatively,  $\Delta$  can be defined by a set of integer linear inequalities, each of which slices a facet. The polytope is then the convex body in  $\mathbb{R}^{n+1}$  enclosed by these facets. We will always include the origin as being contained in  $\Delta$ . Using the usual dot product  $\langle \cdot, \cdot \rangle$  inherited from  $\mathbb{R}^{n+1}$ , the dual polytope is defined by

$$\Delta^\circ := \left\{ v \in \mathbb{R}^{n+1} \mid \langle m, v \rangle \geq -1, \forall m \in \Delta \right\}. \tag{1.1}$$

The polytope  $\Delta$  is *reflexive* if all the vertices of  $\Delta^\circ$  are integer vectors. In this case, we can define the Calabi–Yau hypersurface  $X_n$  explicitly as the polynomial equation

$$\sum_{m \in \Delta} c_m \prod_{r=1}^k x_r^{\langle m, v_r \rangle + 1} = 0, \tag{1.2}$$

where  $v_{r=1, \dots, k}$  are the vertices of  $\Delta^\circ$  with  $k$  being the number of vertices of  $\Delta^\circ$  (or equivalently the number of facets of  $\Delta$ ),  $x_r$  are the coordinates of  $A_{n+1}$ , and  $c_m$  are numerical coefficients parameterizing the complex structure of  $X_n$ . Indeed, the reflexivity of  $\Delta$  ensures that the exponents are integral thereby making the hypersurface polynomial as required.

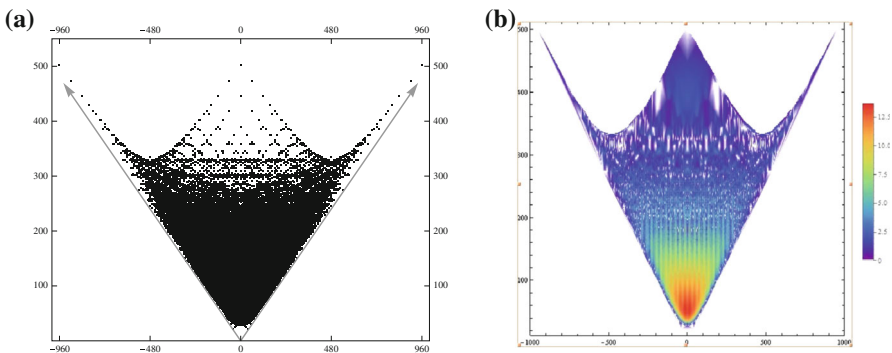


The classification of these Calabi–Yau manifolds thus amounts to that of reflexive polytopes in various dimensions, and the intense computer work of Kreuzer and Skarke was to combinatorially find such polytopes. For  $n = 1$ , there are 16 such polytopes in  $\mathbb{R}^2$ , and we have Calabi–Yau onefolds, or elliptic curves. For  $n = 2$ , there are 4319 such polytopes in  $\mathbb{R}^3$ , and we have Calabi–Yau twofolds, or K3 surfaces. For  $n = 3$ , there are 473, 800, 776 such polytopes (which was a formidable computer task!), and we have the Calabi–Yau threefolds. This sequence

$$\{1, 16, 4319, 473800776, \dots\} \tag{1.3}$$

of remarkable growth rate can be found in the Online Encyclopedia of Integer Sequences [38]. The numbers in higher dimension are still not known, nor has there been an asymptotic analysis of their growth. It should be emphasized that generically a reflexive polytope corresponds to a *singular* toric variety even though the hypersurface is chosen (by generic coefficients  $c_m$ ) to miss the singularities and hence ensuring the smoothness of the Calabi–Yau  $X_n$ . For example, of the some half-billion reflexive polytopes in  $\mathbb{R}^4$ , only 136  $A_4$  are in fact smooth [39]. As we desingularize the toric variety by various star-triangulations of  $\Delta$ , we are led to potentially *inequivalent* Calabi–Yau manifolds. In principle, the *same* Calabi–Yau geometry can arise from different reflexive polytopes or triangulations of a given reflexive polytope. Whereas K3 is essentially unique, we do not know how many Calabi–Yau threefolds there are. A systematic study to classify the desingularizations, to compute the necessary topological data, and to build an interactive online database [19] is under way. The moral is that there are almost certainly far more than half a billion Calabi–Yau threefolds!

Luckily, the Hodge numbers depend only on the polytope and not on the choice of desingularization. (The intersection numbers, however, do depend on the choice.) For Calabi–Yau threefolds, the pair of Hodge numbers  $(h^{1,1}, h^{1,2})$  is a famous quantity. Indeed, the plot in Part (a) of Fig. 1 has become iconic. Here, the sum  $h^{1,1} + h^{1,2}$  is plotted against the Euler number  $\chi = 2(h^{1,1} - h^{1,2})$ , and the left-right symmetry supplies “experimental evidence” for *mirror symmetry*. There is enormous redundancy in this data: of the some half a billion reflexive polytopes, there are only 30, 108 distinct pairs of Hodge numbers and the pair  $(27, 27)$  dominates the multiplicity, totaling almost one million. In Part (b) of Fig. 1 we have attempted to visualize the distribution of the



**Fig. 1.** **a** The cumulative plot of  $\chi = 2(h^{1,1} - h^{1,2})$  on the abscissa versus  $h^{1,1} + h^{1,2}$  on the ordinate for Calabi–Yau threefolds as hypersurfaces in toric fourfolds; **b** marking also the natural logarithm of the multiplicity of the Hodge pair with a *color* grading (color figure online)

multiplicity by having a color density plot of the logarithm of the number over each Hodge pair.

Understanding this multiplicity forms the inspiration for the present work. While there have been analyses on the *shape* of the funnel-like plot [28,33,35], there has not been much work on its *density*, i.e., the distribution of the multiplicity of Hodge data for the Calabi–Yau manifolds of various dimension. Of course, fundamentally, this is entirely due to the combinatorics of reflexive polytopes and might in principle be analytically determined. However, given the complexity of the problem it is expedient to analyze the available data which have been compiled over the years, observe intriguing patterns, and draw statistical inferences before turning to analytic treatments. This is what we achieve in this work.

The organization of the paper is as follows. We perform a detailed analysis on the structure and behavior of the threefold data in Sect. 2. This is motivated by looking for an exact function describing the relationship of the distribution of the Hodge pairs  $(h^{1,1}, h^{1,2})$  with frequency.

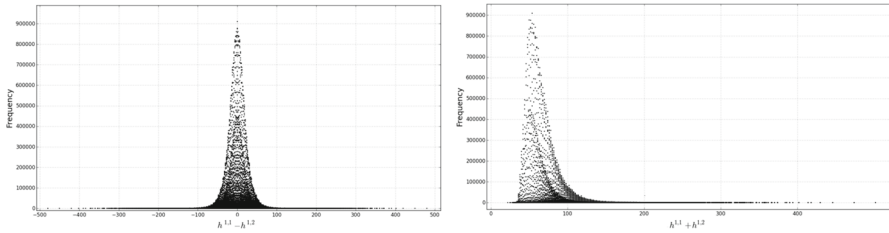
In Sect. 2.1, we study the distribution of  $(h^{1,1} - h^{1,2}, f)$ . We find that this distribution is composed of a family of curves, for which each curve can be described using a modified pseudo-Voigt model. Although an approximation, the model is able to describe the general trend of the data, as well as some additional fine structure within each individual data point. Performing an analysis on the parameter relationships shows that three out of the five parameters can be expressed as a single variable, but we conclude that additional modifications need to be introduced in the model to overcome certain shortfalls.

Subsequently, Sect. 2.2 performs an analysis on the structure of  $(h^{1,1} + h^{1,2}, f)$ . Similarly, this distribution is composed of a family of curves for which each curve can be described using a Planckian profile. Combining the regression analysis for each curve within the distribution, we construct a single function able to approximately model the entire distribution of  $(h^{1,1} + h^{1,2}, f)$  with only two variables. Section 2.3 uses the model developed in Sect. 2.1 to describe the distribution of the Euler number  $\chi$ .

Section 2.4 is dedicated to the description of model validation in our context, as the usual statistical tests are inadequate. Section 2.5 discusses possible implications to physics by referencing recent advancements in F theory and further investigations of structures within the Kreuzer–Skarke database. In Sects. 3 and 4, we perform primary analyses of Calabi–Yau twofolds (Picard number and multiplicity) and Calabi–Yau fourfolds. Due to the lack of a complete data set, we are unable to provide a thorough analysis of the fourfolds as with threefolds. Finally, the Appendix presents many supplementary plots and figures for the various sections. We conclude with a summary and outlook in Sect. 5.

## 2. Calabi–Yau Threefolds

As advertised in the Introduction, we will begin with the analysis of threefolds and identify patterns within this rich distribution of Hodge numbers and their frequency as plotted in Fig. 1. It turns out striking patterns do exist, pointing to a definite structure within the threefold data, which consists of the triple  $(h^{1,1}, h^{1,2}, f)$ , where  $f$  is the number of reflexive polytopes in the Kreuzer–Skarke database with the given Hodge pair. Here,  $h^{1,1}$  and  $h^{1,2}$  respectively count the Kähler and complex structure moduli of the Calabi–Yau obtained from the reflexive polytope. More precisely [8], we have that



**Fig. 2.** **a** Frequency  $f$  plotted against  $\frac{1}{2}\chi = h^{1,1} - h^{1,2}$ ; **b** frequency  $f$  plotted against the sum of Hodge numbers  $h^{1,1} + h^{1,2}$

$$\begin{aligned}
 h^{1,1}(X) &= \ell(\Delta^*) - \sum_{\text{codim}\theta^*=1} \ell^*(\theta^*) + \sum_{\text{codim}\theta^*=2} \ell^*(\theta^*)\ell^*(\theta) - 5; \\
 h^{1,2}(X) &= \ell(\Delta) - \sum_{\text{codim}\theta=1} \ell^*(\theta) + \sum_{\text{codim}\theta=2} \ell^*(\theta)\ell^*(\theta^*) - 5.
 \end{aligned}
 \tag{2.1}$$

In the above,  $\Delta$  is the defining polytope for the Calabi–Yau threefold  $X$  and  $\Delta^*$  is its dual. Moreover,  $\theta$  and  $\theta^*$  are the faces of specified codimension of these polytopes respectively;  $\ell(\cdot)$  is the number of integer points of the polytope while  $\ell^*(\cdot)$  is the number of interior integer points. Indeed, our analysis of the distribution of Hodge numbers ultimately reduces to counting these integer points.

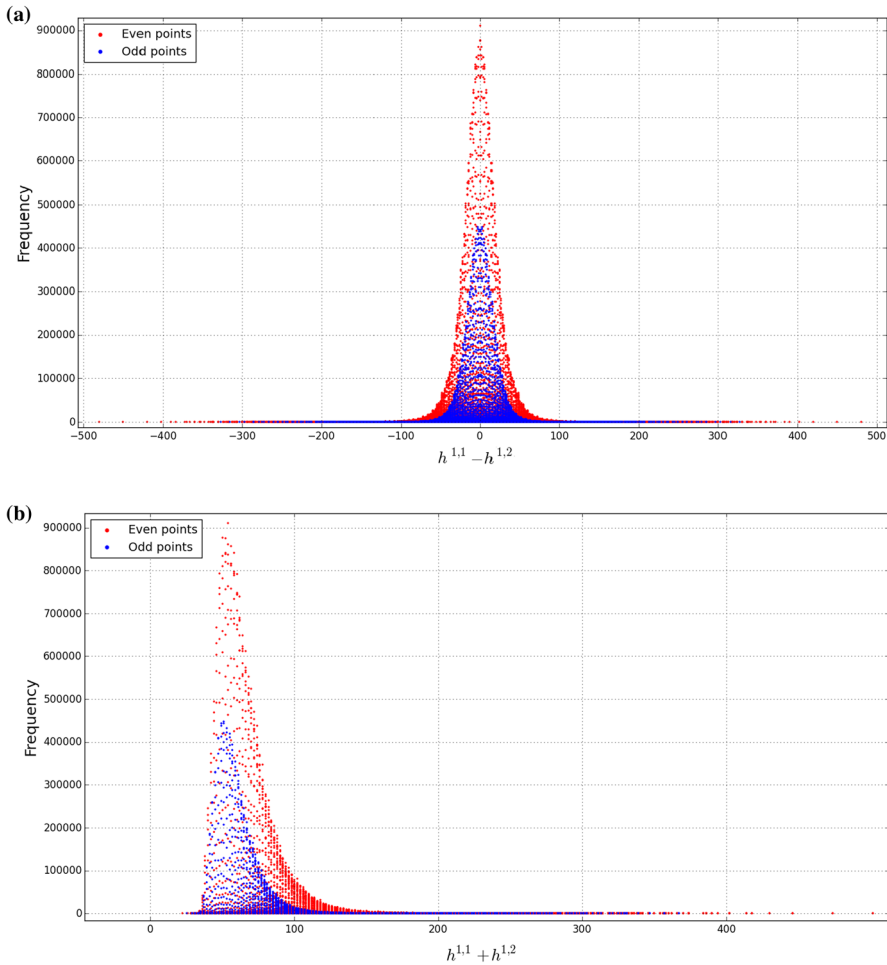
To facilitate the analysis, we plot  $(h^{1,1} - h^{1,2}, f)$  and  $(h^{1,1} + h^{1,2}, f)$  as shown in (a) and (b) of Fig. 2, respectively. Recall that the Euler number  $\chi = 2(h^{1,1} - h^{1,2})$ . We will use the difference  $h^{1,1} - h^{1,2}$  rather than the Euler number. In the simplest heterotic constructions,  $|h^{1,1} - h^{1,2}|$  corresponds to the index of the Dirac operator and gives the number of generations of particles in the low-energy spectrum [1].

By inspection, these plots already exhibit two patterns. Firstly, in both the  $h^{1,1} - h^{1,2}$  and  $h^{1,1} + h^{1,2}$  plots, there appears to be an inner distribution contained within the outer distribution. We find that these inner and outer distributions are related to the parity of  $h^{1,1} \pm h^{1,2}$ . Figure 3 elucidates this point by having the odd and even values in different colors.

Though this parity structure may be a result of the Kreuzer–Skarke algorithm, its consistent appearance means we need to treat the distributions of even and odd distinctly for now.

The second evident structure which can be seen by inspection, is that the outer edge of the distribution of  $h^{1,1} - h^{1,2}$  (Fig. 3a) appears to follow a normal like curve, whereas the edge of  $h^{1,1} + h^{1,2}$  (Fig. 3b) follows a Planck like curve. It is through the analysis of these distributions that we deduce their characteristic behavior and underlying structure. In the main body of this paper, we outline the results and analysis of only the even distributions for  $h^{1,1} - h^{1,2}$  and  $h^{1,1} + h^{1,2}$ , except where it is important to present both. It turns out that any structure and patterns which are found in the even distributions for  $h^{1,1} - h^{1,2}$  and  $h^{1,1} + h^{1,2}$  are found identically in the odd distribution (see “Appendix” for various plots).

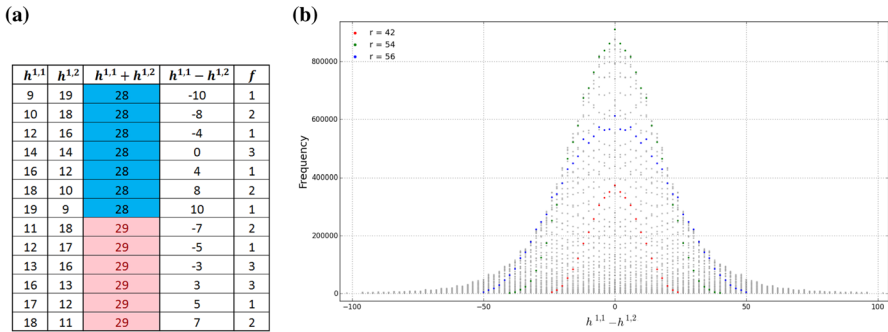
*2.1. Analysis of  $h^{1,1} - h^{1,2}$ .* Before we can present the results, it is important to explain some notation. When working with the distribution of  $h^{1,1} - h^{1,2}$ , we find that it is composed of many curves, whose individual structure is the same as the “edge” or boundary of the distribution mentioned earlier. As a consequence of this, we refer to



**Fig. 3.** **a** The  $h^{1,1} - h^{1,2}$  distribution for threefolds, highlighting the two sub-distributions, where *red* and *blue* data points correspond to even and odd values of  $h^{1,1} - h^{1,2}$ , respectively; **b** the same, but for  $h^{1,1} + h^{1,2}$  (color figure online)

$h^{1,1} - h^{1,2}$  as being composed of a “family of curves.” Each curve is then classified by its *r*-value, where  $r = h^{1,1} + h^{1,2}$ . It is important to be clear that in this analysis, although  $h^{1,1} - h^{1,2}$  is just half the Euler number, we are not summing over all the possible values of  $h^{1,1} + h^{1,2}$ . We are keeping these values distinct: hence, the *r*-curves we obtain. Later on in Sect. 2.3 we sum over all possible values of  $h^{1,1} + h^{1,2}$  to get two plots representing the full Euler number distribution.

Consider the example in Fig. 4a. By ordering the data in terms of  $h^{1,1} + h^{1,2}$ , one can classify data sets within  $h^{1,1} - h^{1,2}$  by an *r*-value. Holding *r* fixed, we can plot the frequency *f* versus the difference  $h^{1,1} - h^{1,2}$ . We call each value of *r* a curve, which we can overlay on the same plot. In this example, we tabulate data for curves identified by  $r = 28$  and  $r = 29$ . As a further illustration, we show explicitly the curves of the even distribution within  $h^{1,1} - h^{1,2}$  for  $r = 42, 54, 66$  in Fig. 4b. By mirror symmetry, the curve is symmetric about the vertical axis, where  $h^{1,1} - h^{1,2} = 0$ .



**Fig. 4.** **a** Example of repeated values of the sum  $h^{1,1} + h^{1,2}$  being 28 and 29; **b** three highlighted curves ( $r = 42, 54, 66$ ) within the even  $h^{1,1} - h^{1,2}$  distribution. The transparent grey data dots are all the data plots for the distribution. Refer to Fig. 23 for the corresponding odd plot

We can now perform a regression analysis for each individual curve, in the quest of obtaining a function describing the distribution. In the analysis, we indeed find an approximate function predicting the fine structure of the data. We operate with one caveat: we ignore data points which have a frequency lower than 2000. At large  $r$ , the data, whose frequency is below 2000, begins to deviate from our model. The reason for such deviations, comes down to the fact that our model, though remarkably accurate, is still an approximation. We suspect that with further modifications, such deviations can be accounted for and that consequently, it may be possible to find an exact function to map the frequency distribution of  $h^{1,1} - h^{1,2}$ . Such statements also apply to the distribution of  $h^{1,1} + h^{1,2}$ .

*2.1.1. A pseudo-Voigt fit* Due to the normally-distributed, peak-like nature of these curves, we performed a regression analysis using the following models: Gaussian; Cauchy (Lorentzian); Pearson7; Breit–Wigner; Voigt; and pseudo-Voigt. In the “Appendix A.1.2”, we perform a side by side comparison. It turns out that both the Voigt model (25e) as well as the pseudo-Voigt model (25f) give excellent fits.

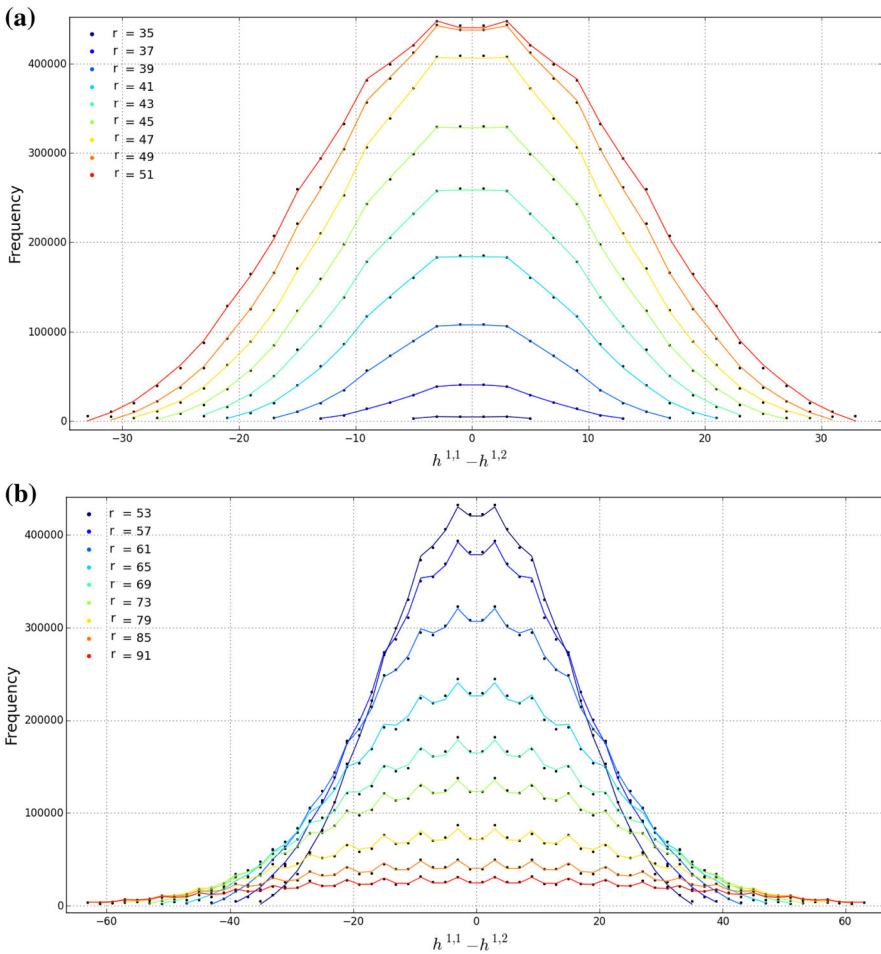
We focus on the **pseudo-Voigt model** as it gives the best fits. This is a linear combination of a Gaussian and Lorentzian (Cauchy) distribution:

$$f(x, A, \mu, \sigma, \alpha) = (1 - \alpha) \frac{A}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} + \alpha \frac{A}{\pi} \left[ \frac{\sigma^2}{(x - \mu)^2 + \sigma^2} \right], \tag{2.2}$$

with amplitude ( $A$ ), center ( $\mu$ ), Gaussian width ( $\sigma$ ), and fractional parameter alpha ( $\alpha$ ). However, we can modify the above distribution slightly so that the amplitude  $A$  of the distribution has an oscillating component

$$A(x, A_0, a, b) = A_0 + a \cos(2\pi b \cdot x), \tag{2.3}$$

where  $A_0$  is the original amplitude of a particular curve described by the pseudo-Voigt distribution,  $a$  is the amplitude of oscillations, and  $b$  represents the period. By doing a regression analysis one curve at a time using this modified pseudo-Voigt model, we are almost able to replicate not just the basic structure of each curve, but even the individual behavior of each data point in the entire distribution. (See “Appendix A.1.3” for a comparative plot of the all the regression curves using the standard, unmodified, pseudo-Voigt model.)



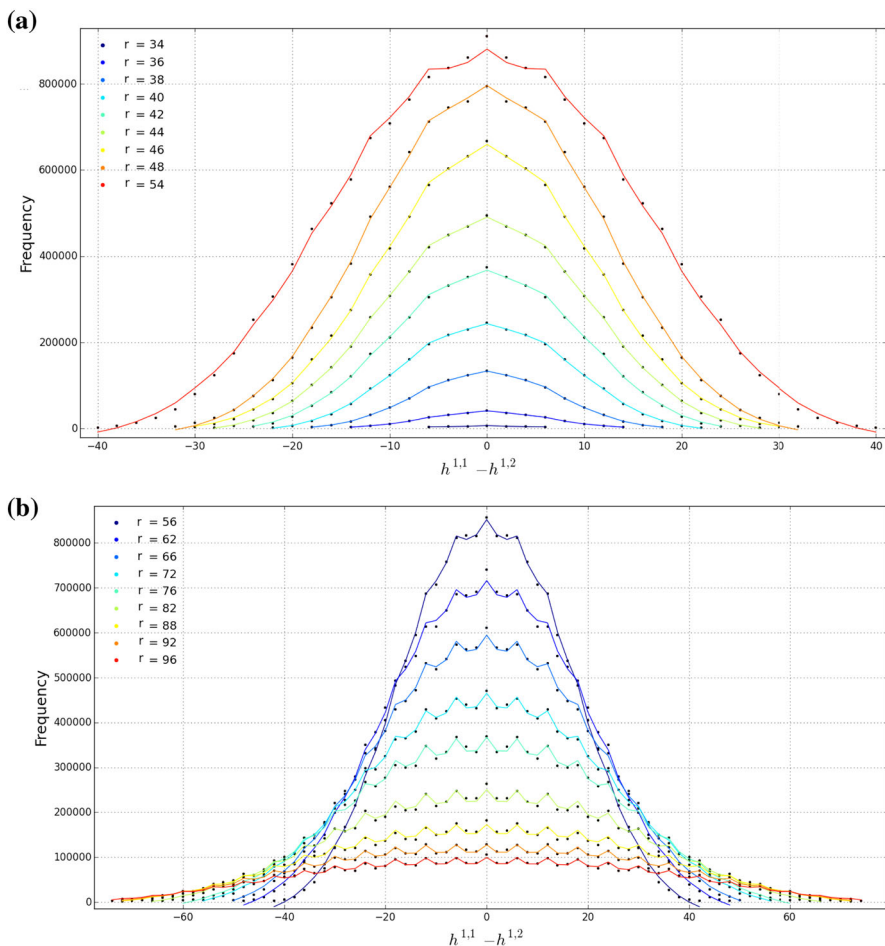
**Fig. 5.** Plots of frequency against  $h^{1,1} - h^{1,2}$  for various odd values of  $r$ . Each line represent a modified pseudo-Voigt profile based on the regression analysis for each curve. See Fig. 28a for a plot of all even curves. **a** Regression lines for all odd  $r$  valued curves, with  $r \in [35, 51]$ . **b** regression lines for few select odd  $r$  values, with  $r > 51$

We plot the frequency against  $h^{1,1} - h^{1,2}$  for various values of  $r$  (odd and even). Figures 5 and 6 are striking in their accuracy.

As these figures illustrate, each curve follows a pseudo-Voigt profile, however the individual data points seem to “jump” up and down, as if oscillating. It is this behavior of the data points which can be accounted for by the modified pseudo-Voigt model. To do the regression analysis, we used Python *lmfit* with a custom model which is just the modified pseudo-Voigt model. The parameters that were fitted are  $(A_0, a, b, \sigma, \alpha)$ . Due to mirror symmetry,  $\mu = 0$ . In “Appendix A.1.4”, one can find a table with the value of every parameter for every curve as well as their reduced  $\chi^2$  values.

A few comments explicate the regression lines and the behavior of the distributions.

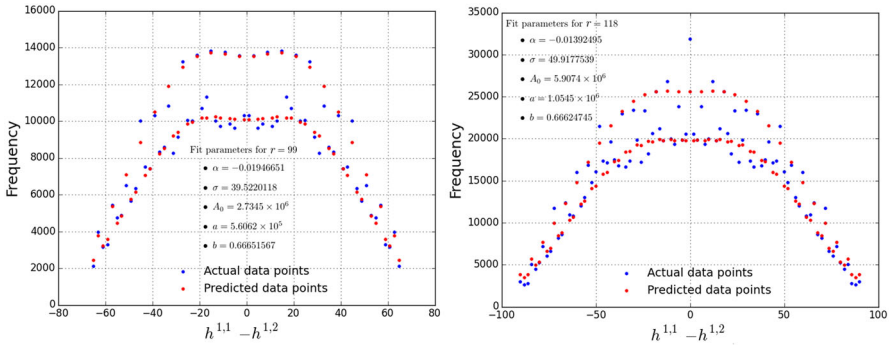
1. When we refer to the model as being an “excellent fit,” it is principally a statement made by inspection of the curves and the data. If one inspects the reduced  $\chi^2$  values



**Fig. 6.** Plots of frequency against  $h^{1,1} - h^{1,2}$  for various even values of  $r$ . Each line represent a modified pseudo-Voigt profile based on the regression analysis for each curve. See Fig. 28b for a plot of all odd curves. **a** Regression lines for few select even  $r$  values, with  $r \leq 54$ . **b** Regression lines for few select even  $r$  values, with  $r > 54$

(Fig. 29), the numbers are large, which statistically does not refer to a good fit. This is misleading however. Firstly, we need to consider that the number of parameters used in the model is five. This allows for a larger  $\chi_R^2$  value. Secondly, the distribution is based on a discrete set of data. When doing a regression analysis using the modified pseudo-Voigt model, one obtains an equation which describes a continuous curve. Lastly, the frequency values span over several orders of magnitude. The tiniest deviation from a parametric model—in this case, the modified pseudo-Voigt profile—will be detected in cases where there is such a huge sample size. Typically the predicted model gives data points which are in the range of 0.02–3% accuracy from the actual data point. The tail behavior of the model is less accurate however, here the predicted values can be off from between 60 and 80%. For cases with a very poor fit, the last data point (large value of  $h^{1,1} - h^{1,2}$ ) can have an error of up to 300%—this is another example of the model being less accurate at lower frequency. When one is dealing with such



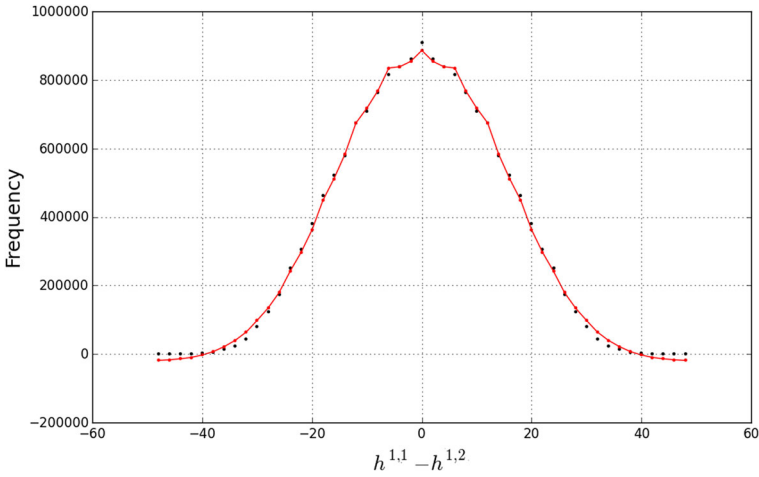


**Fig. 7.** These two plots serve two purposes. The first is to show how the modeled data should really look by using data points (*red points*) instead of the (perhaps misleading) lines (refer to Comment 1 below). The second purpose is to illustrate that as  $r$  becomes large (*left plot* has  $r = 99$ , *right plot* has  $r = 118$ ), the actual data points deviate more and more from the modeled data, implying that there is a missing function in the modified pseudo-Voigt model which would allow one to describe the data at much lower frequencies (color figure online)

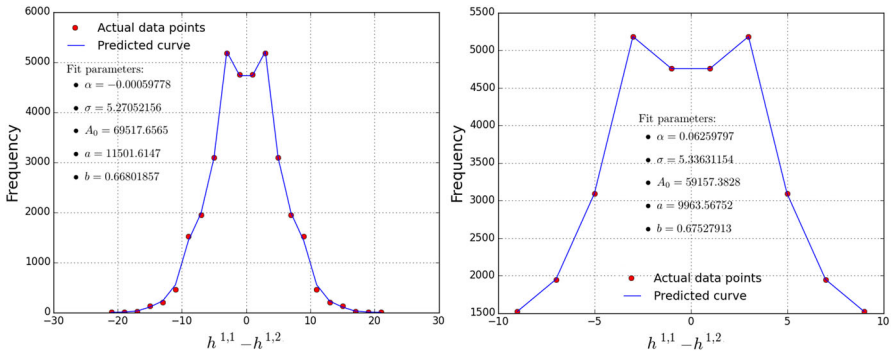
sample sizes, even a 1% error can give a difference of up to a couple of thousand. This difference summed over all the data points for a particular curve result in a large  $\chi_R^2$  value. Due to the discussion in Sect. 2.4 we from now on ignore the  $\chi_R^2$  as a test for model validation. Instead we opt for probability plots—which can also be seen in Sect. 2.4.

2. One obtains a continuous model to describe the discrete data, in reality, we should not be plotting fitted curves, but rather fitted data points—as can be seen in Fig. 7. It is just illustratively more clear to display the curves. One could in principal work out what the discrete approximation is to our continuous model.
3. Although the modified pseudo-Voigt distribution does a good job to model the behavior of the data, one still needs to address the problems experienced with our model at low frequency. A problem which is hidden, by virtue of our cut-off frequency, is that the tail of our models predicts negative values, Fig. 8. There is a possibility that by having different variances  $\sigma_g, \sigma_c$  for the mixing of the two distributions (Gaussian, Cauchy), one could adjust the tail behavior. Introducing more and more parameters however does not always resolve the problem, as it is possible to over-fit the data. Yes, the model may be more accurate, but one loses physical significance. In a situation like ours, where one does not have any physical backing for choice in models, this line between fitting and over fitting is not so clear.
4. The odd distribution’s behavior is more regular. In comparison to the even distribution, as one increases in  $r$  value, the behavior of the individual data points remain somewhat constant relative to the fitted curve. The even distribution becomes more and more irregular as one increases the  $r$  value. This suggests that there is an added parameter which seems as if it should be function of  $r$ . By regular and irregular we are referring to how well the data point is described by the model.
5. Both distributions become very irregular as the value of  $r$  becomes large ( $r > 100$  and  $r > 120$  for odd and even distributions respectively—see Fig. 7). A large  $r$  value refers to curves which have a relatively low frequency. Again this suggests that the pseudo-Voigt model needs to some how have some function of  $r$  which “distorts” the behavior of the curves as  $r$  increases (by the looks of how the real data deviates from the modeled one, it seems that the missing functions is also oscillating in nature).





**Fig. 8.** By considering the entire frequency range, the model is not able to adequately describe the tail behavior. The model goes into the negative frequency range instead of tapering off to 0



**Fig. 9.** *Left plot* shows the modeled line according to the modified pseudo-Voigt distribution with no cutoff frequency. We obtain a good fit to the data. The *right plot* has a cutoff frequency of 460, which is equivalent to a percentage cut off of 9.68% (calculated relative to the peak frequency for that  $r$ -curve). This curve is exact

There exist, however, certain cases where the model is exact. In other words predicted values are the same as the actual values. This happens when one adjusts the frequency cutoff for each  $r$  curve individually. That is to say, we only examine data points with at least  $f_0$  reflexive polytopes with a given value of  $r$  and  $h^{1,1} - h^{1,2}$ . If there are fewer than  $f_0$  cases, the data is ignored.

This trend persists for all values of  $r$ , however what becomes apparent is that it's not the percentage cutoff frequency that determines whether or not one gets an exact fit, but rather, the number of data points that remains after the percentage cut of has been effected. Figure 30 gives a table of how many data points remain after an appropriate cut off percentage has been chosen to achieve a perfect fit. From this table we see that for even curves, one almost always requires 7 data points to achieve a perfect fit; for the odd curves, the number of data points is 10. The reason for this constant number throughout all the curves is that the centers of all the distributions for the various curves are all similar. As soon as one includes a larger number of data points we cannot achieve exact

fits, and the model becomes approximate. At very low  $r$  values the number of data points remaining after cutoff are not too different to the total number of points. As  $r$  increase, the total number of points increase—the fact that we can achieve exact fits becomes less meaningful. The other models—even when including an oscillatory component were unable to give exact fits.

The model is thus much more accurate at low  $r$  values, and as  $r$  increases the actual data deviates more and more from the fit. This reinforces the statements from the comments that the pseudo-Voigt model can be modified further with some function  $g(r, x)$  such that it will greatly improve the accuracy of the fit, and perhaps even become exact.

After the above analysis, we return to our goal of finding a single function describing the distributions. It is clear from the above that the function has to be a function of at least two variable,  $f = f(x, r)$ . We thus continue the analysis by plotting all the parameters versus  $r$ , in search for any relationships. We find that three parameters  $\sigma, b$  and  $\alpha$  can be expressed in terms of  $r$ , the other parameters, while they show trends, do not give a precise relationship with  $r$ . For the even distribution of  $h^{1,1} - h^{1,2}$ , the  $r$  values range from 36 to 110, whereas for the odd distribution (see Fig. 24a, b) the  $r$  values range from 37 to 99. By looking at Fig. 10a, it turns out that:

$$\alpha(r) = c_\alpha, \quad b(r) = c_b, \quad \sigma(r) = c_{\sigma_1}r + c_{\sigma_2}. \tag{2.4}$$

Our model of  $h^{1,1} - h^{1,2}$  now looks as follows:

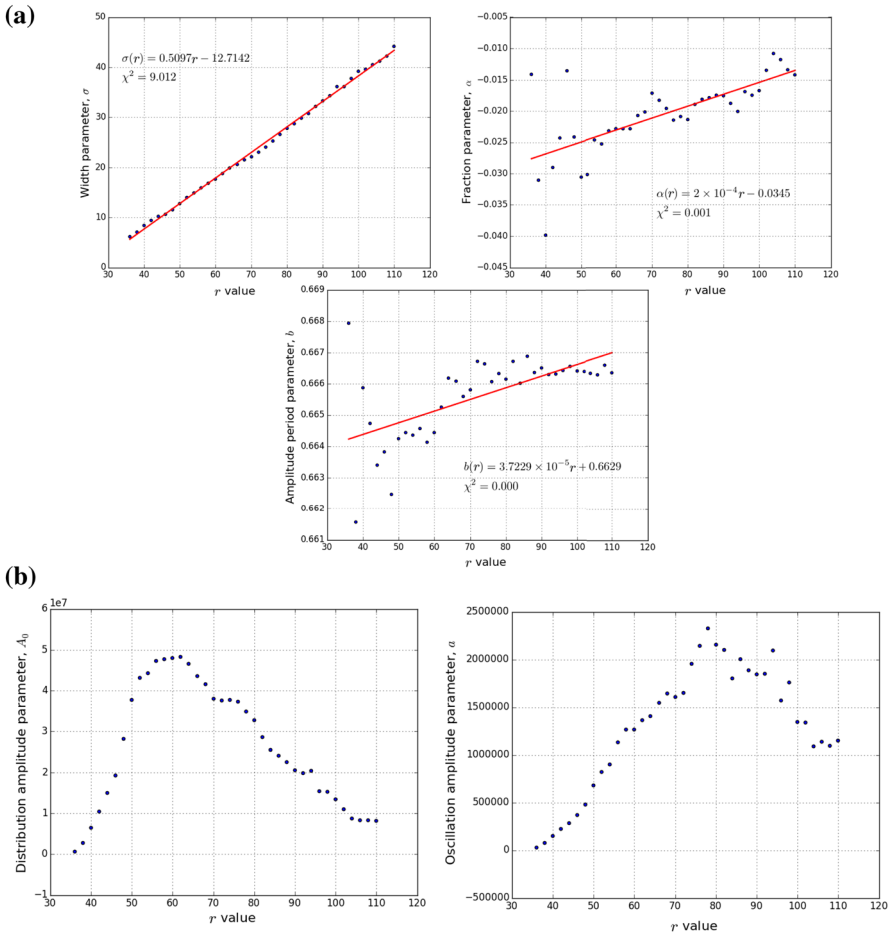
$$f(x, r, A_0, a) = (1 - c_\alpha) \frac{A_0(r) + a(r) \cos(2\pi c_b \cdot x)}{\sqrt{2\pi(c_{\sigma_1}r + c_{\sigma_2})}} e^{\frac{-(x)^2}{2(c_{\sigma_1}r + c_{\sigma_2})^2}} + c_\alpha \frac{A_0(r) + a(r) \cos(2\pi c_b \cdot x)}{\pi} \left[ \frac{(c_{\sigma_1}r + c_{\sigma_2})^2}{x^2 + (c_{\sigma_1}r + c_{\sigma_2})^2} \right], \tag{2.5}$$

where  $A_0(r)$  and  $a(r)$  are two unknown functions yet to be determined (see Fig. 10b for relationship plots). For replicating the plots as precisely as possible, one would need to keep the parameters, as they are, up to their 17 decimal values, without excluding terms as we have done. If one wants to reproduce the data from the model, one has to use the exact expressions. Making an approximation from an already approximate model leads to large errors.

The first plot in Fig. 10a in particular evinces a sinusoidal fluctuation about the mean. This again indicates the possibility of refining the plots by adding an extra function.

2.2. *Analysis of  $h^{1,1} + h^{1,2}$ .* We begin by classifying the curves within the  $h^{1,1} + h^{1,2}$  distribution (Fig. 2) in an analogous way to how it was explained before. This time, we order the data by  $h^{1,1} - h^{1,2}$  such that a single curve within  $h^{1,1} + h^{1,2}$  can be identified by its  $q$ -value, where  $q = h^{1,1} - h^{1,2}$ . Due to mirror symmetry, the curve for  $q = -a$  is the same curve as  $q = a$ , thus within our two-dimensional plots will only have  $q > 0$ . In continuation to the analysis on  $h^{1,1} - h^{1,2}$ , we use a cutoff frequency of 2000 and only present results from the even distribution within  $h^{1,1} + h^{1,2}$ , unless stated otherwise. As an example, illustrating the classification of curves within  $h^{1,1} + h^{1,2}$ , consider the curves  $q = 0, 18, 30$  in Fig. 11.

2.2.1. *A Planckian fit* Each curve within the  $h^{1,1} + h^{1,2}$  distribution behaves the same. Just like in the  $h^{1,1} - h^{1,2}$  distribution, we do a regression analysis for each curve within

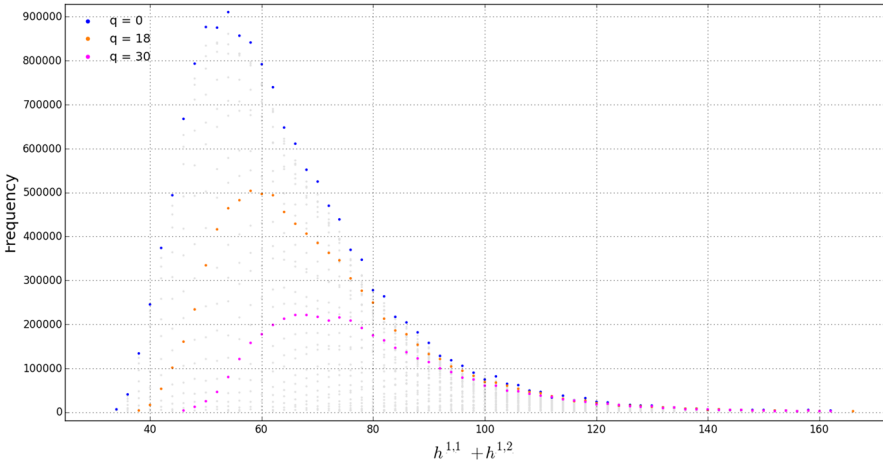


**Fig. 10.** For the even distribution of  $h^{1,1} - h^{1,2}$ . **a** The width parameter  $\sigma$  has a linear relationship with  $r$  such that  $\sigma(r) = 0.5097r - 12.7142$ . The amplitude period parameter,  $b$ , also has a linear relationship, however, since  $r$  is at most order 3 in magnitude, we can regard it as a constant such that  $b(r) = 0.6629 \sim 2/3$ . The same goes for the fraction parameter,  $\alpha$ ; we can regard it as a constant such that  $\alpha(r) = -0.0345$ . For odd parameter fit statistics see Fig. 24a; **b** plots of  $A_0$  versus  $r$  (left) and  $a$  versus  $r$  (right). Both exhibit a similar pattern, however it is difficult to discern any nice relationships. For odd parameter plots see Fig. 24b

the distribution independently, in the quest to describe the entire  $h^{1,1} + h^{1,2}$  with a single function. The model we chose to describe  $h^{1,1} + h^{1,2}$  is the simplest possible Planckian model

$$f(x, A, n, b) = \frac{A}{x^n} \frac{1}{e^{b/(x-22)} - 1} \tag{2.6}$$

The parameter names in the fit results are the amplitude  $A$ , the power  $n$ , and some real constant  $b$ . The shift in  $x$ -axis is so that the distribution begins at 0 as the smallest  $h^{1,1} + h^{1,2}$  above the cutoff is 22. The choice of a Planckian model in the above form is greatly motivated by the blackbody distribution  $f(T, \lambda)$ . The  $q$  curves within  $h^{1,1} + h^{1,2}$  appear to behave in a manner analogous to the curves of constant  $T$  within the blackbody



**Fig. 11.** Three curves ( $q = 0, 18, 30$ ) within the even  $h^{1,1} + h^{1,2}$  distribution. The transparent *grey data dots* are all the data plots for the distribution. Refer to Fig. 31 to see the same example for the classification of odd curves within the odd distribution

distribution. This is an initial trial. Later, we will discover additional structure in the distribution by trying to mimic the blackbody distribution exactly. It turns out that the general behavior of the distribution is modeled very well, cf. Fig. 12a.

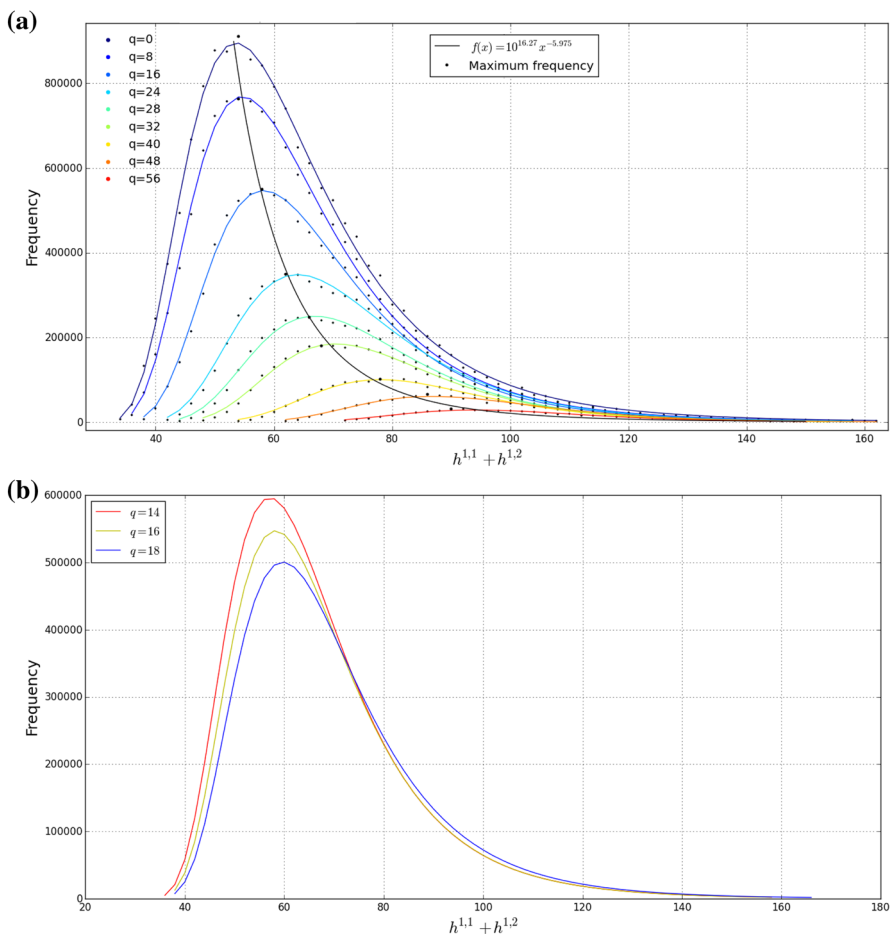
Consider the maximum of each of the curves. As indicated in Fig. 12a, we can fit the maxima to a curve as indicated using the data plotted for the given values of  $q$ . From the above analysis, the  $h^{1,1} + h^{1,2}$  distribution behaves analogously to a blackbody spectrum—except for one small subtlety. It is in this subtlety that the added structure within  $h^{1,1} + h^{1,2}$  is observed.

Just as was seen in Fig. 2,  $h^{1,1} + h^{1,2}$  appears to split up into two smaller distributions based on the parity of  $h^{1,1} + h^{1,2}$ . One can then further break up both the even and odd distributions into three further sets. The manner we observed this added fine structure is again motivated by a blackbody spectrum. In a true blackbody distribution, the curves of constant  $T$  never overlap. However, if you consider the lines of best fit only, when looking at our distribution one sees an overlap of certain curves. For example, observe the following plot of curves which clearly cross in Fig. 12b.

It turns out that this overlapping occurs consistently to the point where one can classify the curves (defined by their  $q$  value) into residue classes  $q_n$  distinguished by  $n \bmod 6$ . On the left hand side of the  $h^{1,1} + h^{1,2}$  axis, the curves are ordered with red (residue class  $q_2$ ) above yellow (residue class  $q_4$ ) above blue (residue class  $q_0$ ), whereas on the right hand side of the axis, the order is reversed. Similar behavior is observed in the odd distribution of  $h^{1,1} + h^{1,2}$  with the curves in the residue classes  $q_1, q_3$ , and  $q_5$  (see Fig. 32b).

The clusters of curves constitute an entire set of mod 6 residue classes. These classes now define a set of curves which belong to very “nice” distributions that behave exactly like a blackbody distribution.<sup>1</sup> Compare, for example, a plot of the all the curves for even distribution of  $h^{1,1} + h^{1,2}$ , separated into their residue classes, Fig. 13

<sup>1</sup> Of course  $h^{1,1} + h^{1,2}$  is not continuous. It is discrete. However, the structure of the best fit curve to the data points appears very similar to that of a continuous blackbody distribution.

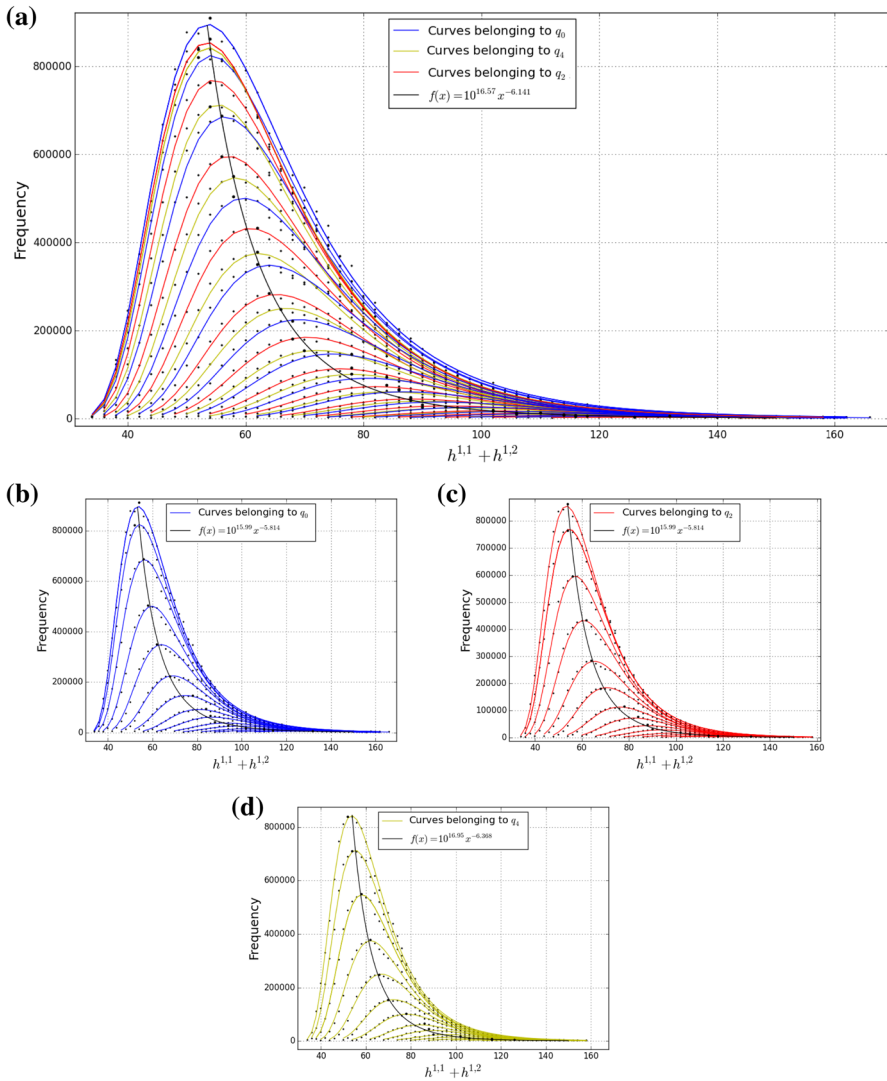


**Fig. 12.** In the attempt to describe the data analogously to a blackbody distribution (a), we discover some subtle structure (b). **a** Lines of best fit from a regression analysis for a few select curves. The *black data* points represent the maximum frequency for that particular  $q$ -curve. The *Black line* is a line of best fit to describe the points of maximum frequency—this is analogous to a blackbody spectrum. See Fig. 32a for the curves within the odd distribution. **b** The curves segregate into three classes determined by the value of the even integer modulo 6. A similar pattern occurs in the odd distribution; see Fig. 32b

As a first approximation we have successfully modeled the general trend of the data. There is, however, a fine structure to the individual data points that we would like to model. Introducing an oscillating term in the amplitude, as seen in the analysis of  $h^{1,1} - h^{1,2}$ , unfortunately did not seem to improve the fits.

Again, it appears that the least number of variables our functions can have is two,  $f = f(x, q)$ . This function will be slightly different in the values of coefficients, depending on which residue class one is modeling.

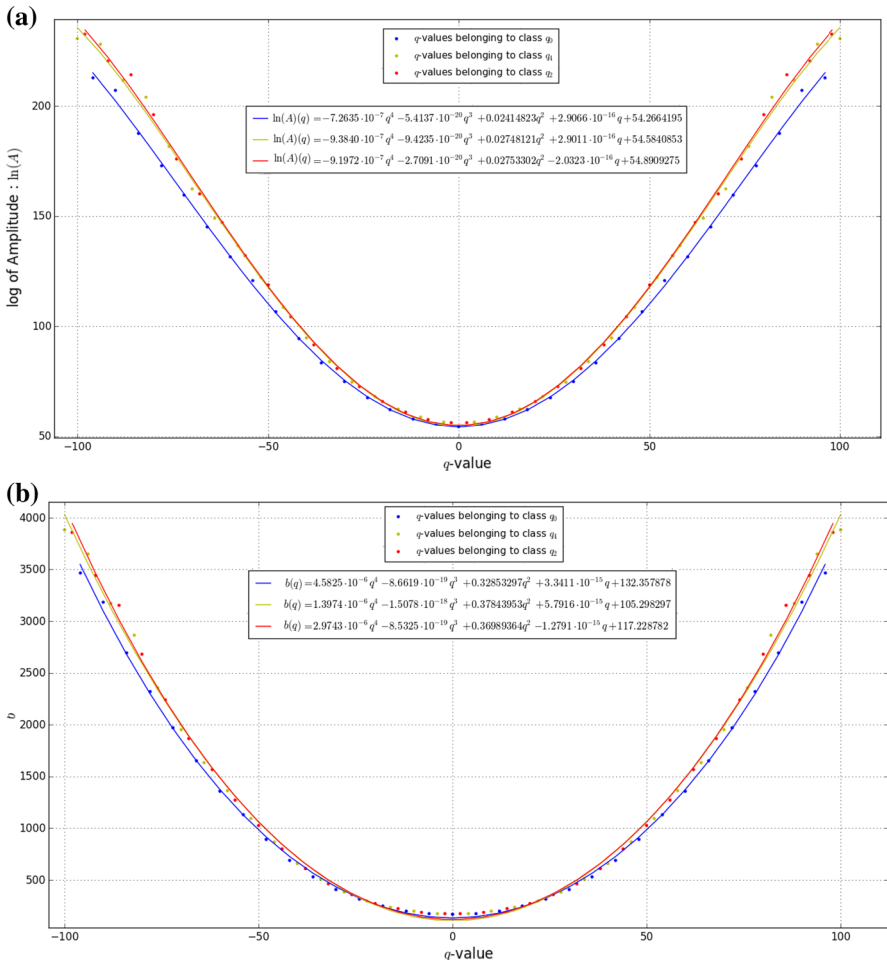
Just as for  $h^{1,1} - h^{1,2}$ , we wish to express the parameters for the  $h^{1,1} + h^{1,2}$  model (2.6) in terms of  $q$ . We therefore write  $A = A(q)$ ,  $b = b(q)$ ,  $n = n(q)$  and seek to find expressions for the coefficients.



**Fig. 13.** We illustrate the added structure for even  $h^{1,1} + h^{1,2}$  data, by displaying how the regression curves can be divided into residue classes. For the list of odd curves, refer to Fig. 33. **a** All the curves color coded according to what residue class their curves  $q_n$  belongs to. **b** Family of curves all belonging to  $q_0$ . **c** Family of curves all belonging to  $q_2$ . **d** Family of curves all belonging to  $q_1$  (color figure online)

While the  $x$ -axis of  $h^{1,1} + h^{1,2}$  has only positive  $q$  values—due to the fact the data points will overlap—when plotting them against the parameter values, we also have to consider the negative values of  $q$ . We present the various relationships (see Fig. 34 for the plots for the odd distribution of  $h^{1,1} + h^{1,2}$  analogous to Fig. 14).

Each distribution has an equation with different parameter values. However, the fact that we can express all the parameters in terms of  $q$  means we are able to get a generalized formula to describe the entire  $h^{1,1} + h^{1,2}$  distribution—as long as the frequency is above 2000. For succinctness we use the following notation for the coefficients



**Fig. 14.** The parameter plots are *color* coded according to what residue class their  $q$  value belong to. **a** Plotting the  $q$ -value parameter versus the  $\log(A)$  parameter. **b** Plotting the  $q$ -value parameter versus the  $b$  parameter. **c** Plotting the  $q$ -value parameter versus the power  $n$  parameter (color figure online)

$$A_{k,i}, n_{k,i}, b_{k,i}, \tag{2.7}$$

where the subscript  $k = 0, 1, 2, 3, 4, 5$  refers to residue class  $q_k$ , and  $i = 0, 1, 2, 3, 4$  refers to the coefficient of the  $i^{th}$  power of  $q$ . Thus, we have:

$$A_k(q) = \exp\left(\sum_{i=0}^4 A_{k,i}q^i\right), \quad n_k(q) = \sum_{i=0}^4 n_{k,i}q^i, \quad b_k(q) = \sum_{i=0}^4 b_{k,i}q^i, \tag{2.8}$$

where the matrix of coefficient values for  $A_{k,i}, n_{k,i}$  and  $b_{k,i}$  can be found in ‘‘Appendix A.2.2’’.<sup>2</sup> Our function (2.6) now is able to approximately describe the entire  $h^{1,1} + h^{1,2}$  distribution:

<sup>2</sup> Perhaps it is important to state explicitly—due to potential confusion—that the coefficients  $A_{k,i}$  refers to the natural logarithm of the amplitude values while  $A_k$  is the actual amplitude seen in the model.

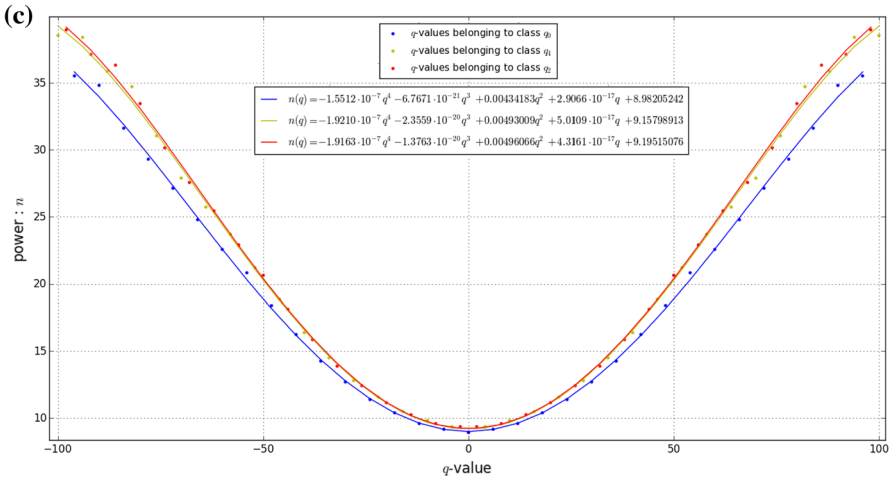


Fig. 14. continued

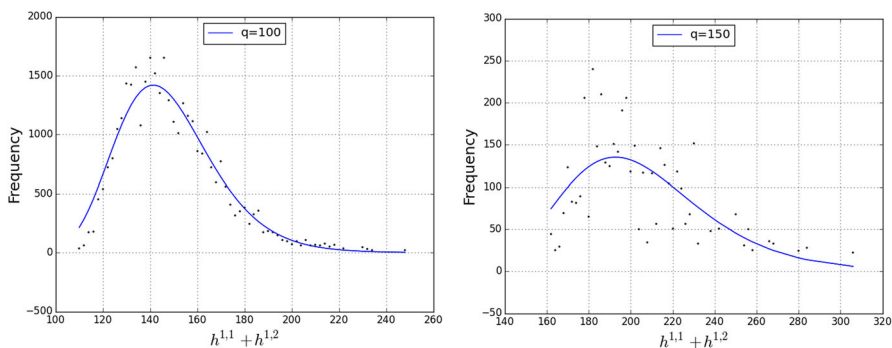
$$f_k(x, q) = \frac{e^{\sum_{i=0}^4 A_{k,i} q^i}}{x \sum_{i=0}^4 n_{k,i} q^i} \frac{1}{\left( e^{\frac{\sum_{i=0}^4 b_{k,i} q^i}{(x-22)}} - 1 \right)}, \tag{2.9}$$

Of course there are certain constraints on the values of  $q$ . For a given  $k$ ,  $q$  has to be an integer which falls within the residue class  $q_k$ . For even values of  $k$ ,  $x = 2m$ , and for odd  $k$ ,  $x = 2m + 1$ . We have  $m > 12$ .

A few comments about the analysis on the  $h^{1,1} + h^{1,2}$  distribution are in order.

1. The Planckian model used in (2.6) could be modified in some manner such that there is some oscillating behavior in the amplitude. Any kind of oscillatory term we introduce, only has a mild effect on the model’s behavior. As the  $q$  values exceed 100, the model is not able to describe the data very well.
2. Assuming one adds an oscillatory component to the model, the module used in python to do the regression analysis called *lmfit* is sensitive to the initial conditions set by the user. Since the model is a custom model, it is difficult to find the correct initial conditions such that the best fit line oscillates close to every point (as with  $h^{1,1} - h^{1,2}$ ).
3. It is possible that the model used does not have the features required to describe the oscillatory “up and down” behavior of the data points. The Planckian model was chosen in that the  $h^{1,1} + h^{1,2}$  distribution resembled a blackbody distribution.
4. In choosing a polynomial model for Fig. 14a–c, we picked the lowest order polynomial that gave the best fit. Choosing the order to be four for all the plots appeared to be convenient. However, it is apparent that the parameter relationship plot in Fig. 14b would be better described by a polynomial of order 6. One could use an order 6 polynomial for all the other relationships plots too, but doing so might not have any physical significance. One can achieve an arbitrarily good fit the larger the order of the polynomial used, but that does not necessarily mean the chosen model is the correct model.





**Fig. 15.** *Left figure is the fitted model (blue line) for a  $q$  value of 100 and right has a  $q$  value of 150. As the  $q$ -value increases, the scattering of the data points within  $h^{1,1} + h^{1,2}$  increases to the point where the model works no longer. For an example of how the model begins to break down at large  $q$ , see Fig. 35 (color figure online)*

**2.3. The distribution of the Euler number.** The Euler number for Calabi–Yau threefolds is

$$\chi = 2(h^{1,1} - h^{1,2}). \tag{2.10}$$

As mentioned previously, we are summing over all the various  $r$ -curves to obtain the full-Euler number distribution. A plot of  $\chi$  versus frequency yields the pseudo-Voigt distribution. In particular, we can model the behavior of the distribution almost perfectly using the modified pseudo-Voigt curve (2.11) and (2.12), which is repeated here for convenience:

$$f(x, A, \sigma, \alpha) = (1 - \alpha) \frac{A}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} + \alpha \frac{A}{\pi} \left[ \frac{\sigma^2}{x^2 + \sigma^2} \right], \tag{2.11}$$

where

$$A(x, A_0, a, b) = A_0 + a \cos(2\pi b \cdot x). \tag{2.12}$$

The results of the regression analysis for the Euler number distribution is presented in Fig. 16a.

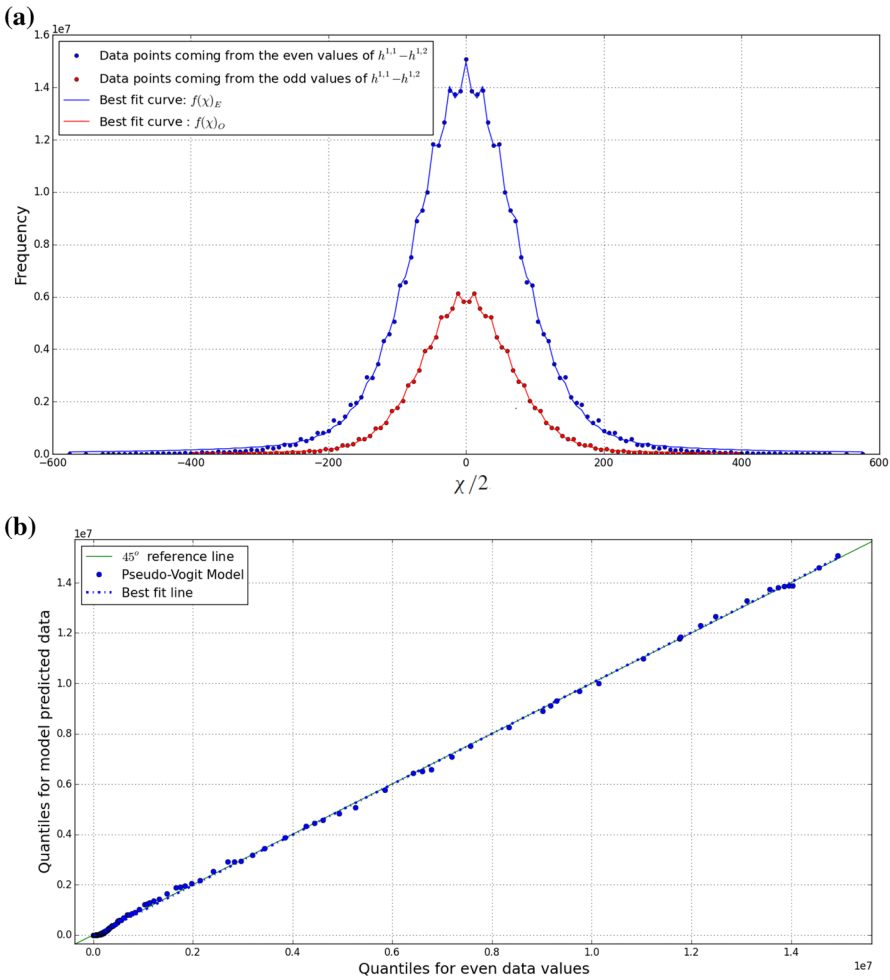
The fitted parameter values for  $f(\chi)_E$  corresponding to even values of  $h^{1,1} - h^{1,2}$  are:

$$(A_0, \sigma, \alpha, b, a) = (1.9032 \times 10^9, 75.8305889, 0.00718459, 0.58347826, 8.7427 \times 10^7). \tag{2.13}$$

Likewise, the fitted parameter values for  $f(\chi)_O$  corresponding to odd values of  $h^{1,1} - h^{1,2}$  are:

$$(A_0, \sigma, \alpha, b, a) = (7.6043 \times 10^8, 64.9735680, 0.00549425, 0.83357720, 3.6881 \times 10^7). \tag{2.14}$$

Although  $\chi$  is only even, the two curves originate from the fact that if you take  $\chi/2$  you get even and odd values. The two curves arise from the parity of  $\chi/2$  and are presented in Fig. 16a.



**Fig. 16.** Various plots illustrating the actual fit of the modified pseudo-Voigt model. We can tell we have a good fit by looking at the probability plots for the quantiles of the standard pseudo-Voigt distribution versus quantiles for the actual data. The  $R^2$  values in (b) and (c) are given relative to the line  $y = x$ . **a** The distribution of Euler numbers fitted to a modified pseudo-Voigt curve. The *blue* curve  $f(\chi)_E$  represents even values of  $\chi/2$ . The *red* curve  $f(\chi)_O$  represents odd values. **b** Probability plot for the even values of  $\chi/2$ . The model fits the data with  $R^2 = 0.99944$ . **c** Probability plot for the odd values of  $\chi/2$ . The model fits the data with  $R^2 = 0.99965$  (color figure online)

**2.4. Goodness-of-fit.** A goodness-of-fit test is implemented as a means of testing how well a given model describes some given data. Typically the model validation process consists of only quoting a single statistically generated number like the  $R^2$ ,  $\chi^2$  or  $p$  values. Based on the size of this number, one then makes inferences on how well the chosen model fits the observation. One needs to be careful however of misusing such indicators as an absolute measure for assessing goodness-of-fit.

For a structural equation model (SEM)—in our case, the modified pseudo-Voigt and Planckian models—this assessment is not so straight forward as it would be for a simple regression analysis. To quantify the predictive power of an SEM, a single statistical

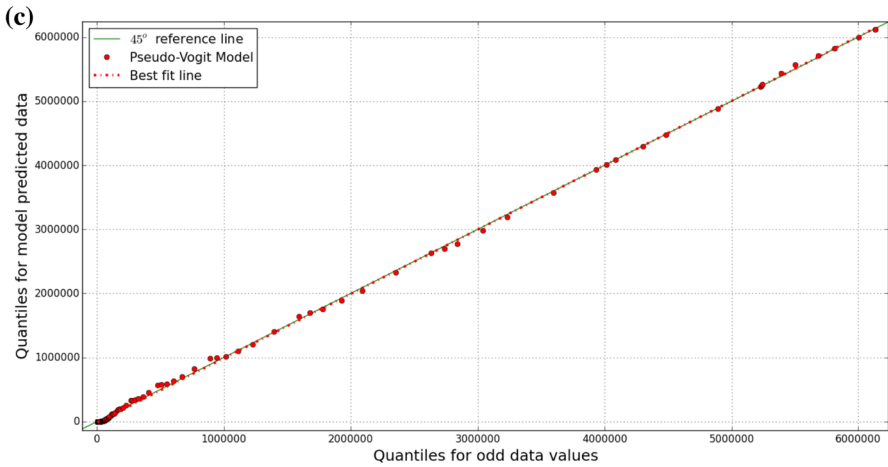


Fig. 16. continued

test does not suffice - in fact, there is no single test. According to [41], the best one can do is assess three different aspects of what it means to have a good fit, these are: overall fit, comparative fits to a test model and model parsimony.<sup>3</sup> The only real test available is the chi-squared ( $\chi^2$ ) test, when it comes to overall fit, this  $\chi^2$  statistic is the most popular test. The  $\chi^2$  test compares observed and predicted correlation matrices with each other, and so, statistical significance is evaluated based on the value of  $\chi^2$ . A large  $\chi^2$  value signifies a considerable difference between the correlation matrices. A low value indicates there is little statistical difference between matrices. Since the  $\chi^2$  test is between actual and predicted matrices only, when looking for overall fit, one searches for non-significant differences between the correlation matrices. Often, rather than presenting the  $\chi^2$  or  $\chi^2_R$  (the chi-squared value relative to the degrees of freedom for the model) value, a  $p$  value is given instead. The  $p$  value, in a way, informs us whether one should reject a null hypothesis or not. A small  $p$ -value suggests that the differences in observed versus predicted are too large to be consistent with the null-hypothesised model i.e. assuming the null-hypothesised model, the probability of observing what we did is relatively small, suggesting either an absolutely fluke experimental outcome or an incorrect model null-hypothesis. The  $p$ -values can be determined by a  $p$ -value calculator by inputting the  $\chi^2_R$  value. There is no standard way of choosing a significance level for the  $p$ -value, but typically  $p < 0.05$  is considered statistically significant.

In general, statistical non-significance given by appropriate values of the  $\chi^2$  fit statistics is adequate. However, one must be careful of drawing similar conclusions for structural equation modeling. The fit statistic makes a statement of the correlation matrices only, not about whether or not the correct model is identified. This is largely due to the sensitivity to sample size of the  $\chi^2$  test. In our analysis, the sample size (number of reflexive polytopes) is enormous—almost one billion! For large samples ( $> 200$ ) the  $\chi^2$  test will give significant differences for any model used. This sensitivity to a sample size, together with an *effect size* and *alpha value*, is related to what one calls the power of a test - the probability of not incorrectly accepting a null hypothesis that is actually false.

<sup>3</sup> Parsimony refers to the ability of a model to give a certain degree of fit whilst having the least required number of predictor variables.

Without worrying too much about what an effect size and alpha value is; for any alpha value, the greater the sample size, the greater the power of the statistical test. However, increasing the sample size beyond a certain amount, can result in the test having “too much” power.<sup>4</sup> Perceived effects in very large sample sizes, will always become significant.<sup>5</sup> Observe how in Figs. 29 and 36 the  $\chi^2_R$  values for all the different curves is extremely large, naively indicating that we have a horrible fit—which would be an incorrect conclusion.

It is clear from the above discussion that we cannot use the  $\chi^2$  or  $p$  values in validating our choice in model. What is not so clear, is the additional subtlety in using purely statistical means to assess goodness-of-fit for our data. This subtlety lies at the heart of almost all statistical tests—the construction of a null hypothesis. The term frequency, as used in the statistical sense, refers to the number of outcomes for a certain event. The measurement of this outcome will often have certain known or unknown factors affecting it. These tests check for the probability that the errors found are too significant to be solely due to random variations in the data. For example, assume that statistical tests give non-significant results. If the residuals are small enough to be considered random errors in the measurement of the frequency, we could say that the model is appropriate. If however, the residuals are too large or present additional structure, we could say the model is good, but not quite the correct one as the residual errors are not “random enough”. In our case, there is no notion of measured frequency and error in measurement of frequencies. Our frequencies are generated as a result of a combinatoric calculation. Statistical tests assume that the input is from measurement and observations (obeying some null-hypothesis), thus they are inherently constructed with this notion in mind. By inputting our data, the tests are trying to calculate something from a data set which does not obey the very assumption they use in their calculations. We are not exactly clear how much this affects statistical outcomes, but it is important to keep in mind.

How do we validate then, that our chosen models are a good fit, or that our model is the best one at describing the data? We implement graphical methods. The first graphical method is obviously through pure inspection—this is not quite statistically quantifiable. There is a statistically based graphical method to assess goodness-of-fit called probability plots, Q-Q plots or P-P<sup>6</sup> plots. These plots were initially constructed to test the “normality” of a data set when the sample size is too large to depend on the  $\chi^2$  and  $p$  values. In principle, a standard probability plot tells you the likelihood that the a sample’s distribution of data obeys a normal distribution—hence checking for normality. The answer to the question is not given by a statistical value, but rather by a graphical representation—from which one can extract statistical numbers. If the plotted data on this probability plot is a straight line, then we can determine that the sample set is normally distributed.

We can extend this concept further: we can take two different samples, and take a probability plot to determine if two data sets come from populations with a common distribution. Such a probability plot is referred to as a Q-Q (quantile–quantile) plot. Extending this concept one more time—as for our use—we will take the quantiles of our theoretical distribution (the modified pseudo-Voigt and Planckian profiles) as our

<sup>4</sup> Power is the probability that you do detect deviations from your null-hypothesised model, when the null-hypothesised model is, in fact, incorrect.

<sup>5</sup> Conversely is also true, for extremely small sample sizes, any effect which should be significant, becomes insignificant.

<sup>6</sup> A P-P plot is the plot of the cumulative distribution frequency of the one data set against the CDF of the other. P-P plots are not as useful as Q-Q plots, thus are seldom used.

“first sample” and plot them against the quantiles of our data as our “second sample”; this will give us our probability plot. In all the probability plots, it is the quantiles of the respective data sets which are plotted against each other.

Quantiles are basically just a generalization of quartiles. For example, the  $k$ th percentile of a set of values divides them, such that the number of values which lie below is  $k\%$ , and the number of values which lie above is  $(100 - k)\%$ . The 25th percentile is the lower quartile or the  $\frac{1}{4}$  quantile. Quantiles are the same as percentiles, but indexed by sample fractions rather than by sample percentages. Suppose that  $p \in [0, 1]$ , the aim is to find the value that is the fraction  $p$  of the way through the ordered data set. As an example, if  $p = \frac{1}{2} = 0.5$ , we want to know what is the value that sits at  $p = 0.5$  of the way through i.e. half way. The value that sits there (this value may have to be interpolated) will be called the quantile for the fraction  $p = 0.5$ . There are many different algorithms for generating the quantiles for a given data set, we use python to generate the quantiles in a manner similar to that discussed above. For an ordered data set,  $x_1 \leq x_2 \leq x_1 \dots \leq x_{n-1} \leq x_n$ , the most common way of calculating quantiles is to first compute the empirical distribution function:

$$F(x) = \frac{1}{n} \sum_{i=1}^n 1(x_i \leq x), \quad x \in \mathcal{R}, \tag{2.15}$$

and then define the quantile function to be the inverse of  $F(x)$ :

$$F^{-1}(p) = \min\{x \in \mathcal{R} : F(x) \geq p, \quad p \in (0, 1)\}. \tag{2.16}$$

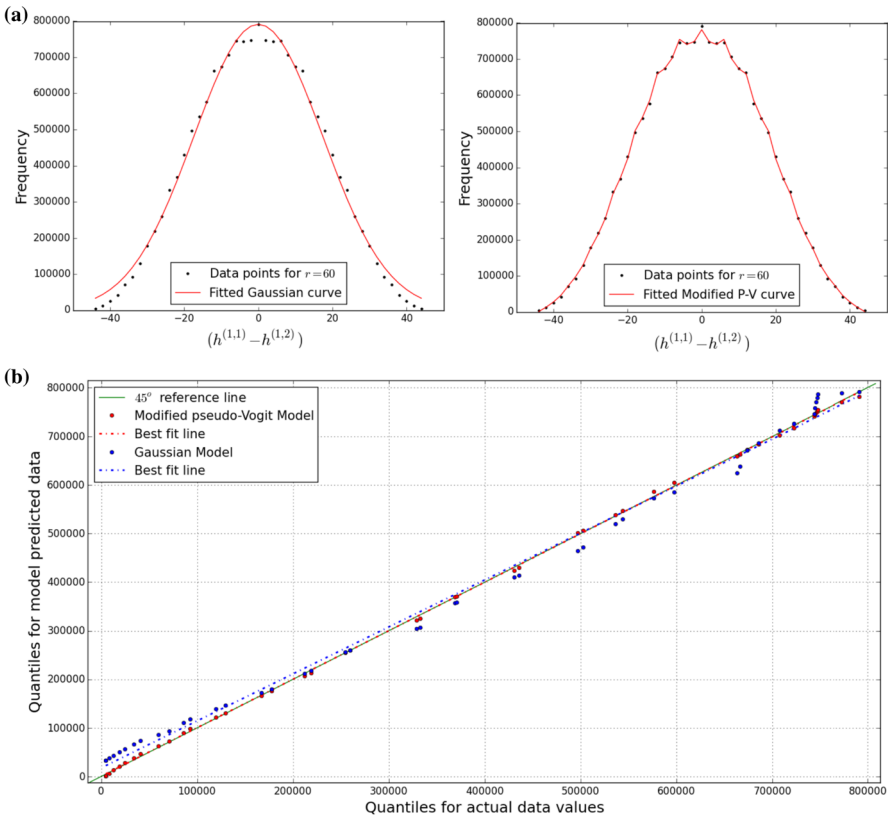
By generating the quantiles of some theoretical model and comparing them to the quantiles of a given data set of equal length, one can determine if the data set belongs to the same distribution as the data set belonging to the theoretical model—i.e., does the data fit the model. If the quantiles are roughly equal the plots will all be more or less on a straight line.

In probability plots:

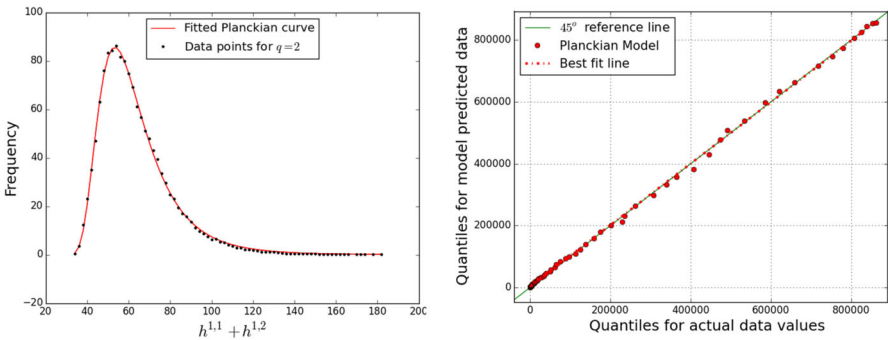
1. The length of data set needs to be equal. For unequal lengths, one must perform an interpolation of data.
2. If two identical data sets were compared to one another, the points would lie exactly on a 45 degree line. Thus, for two different data sets, the deviation from this reference line determines the likelihood that the sets belong to similar distributions. To quantify this likelihood, one can calculate the  $R^2$ -value of the data, relative to the  $y = x$  reference line.
3. Q–Q plots are not only limited to determining similarity in data sets. By analyzing the deviations which occur, one can determine how the scale and location of the data is shifted - the data would follow some line  $y = mx + c$ , where  $m, c$  would be the estimates of these shifts in scale and location. Also, from the distribution of points above or below the reference line, one can infer aspects of the tails and skewness in the data.

Consider the following curves for the  $h^{1,1} - h^{1,2}$  distribution with  $r = 60$  in Fig. 17a, b.

For the  $h^{1,1} + h^{1,2}$  distribution we just plot the data of  $q = 2$  together with the corresponding probability plot in Fig. 18.



**Fig. 17.** Using probability plots, we are able to statistically see which model provides the better fit. We employ such graphical methods as standard goodness-of-fit tests such as the  $\chi^2$  fail to give meaningful results. **a** Best fit curve for  $r = 60$  based on the *left* Gaussian model, *right* modified pseudo-Voigt model. **b** Probability plot for Fig. 17a. The *x*-axis represents the quantiles for the actual data, the *y*-axis represents the theoretically predicted quantiles—dependent on the model chosen (*red* modified pseudo-Voigt model ( $R^2 = 0.99974$ ); *blue* Gaussian model ( $R^2 = 0.99334$ )). The  $R^2$  values are not relative to the best fit lines, but are relative to the  $45^\circ$  reference line  $y = x$ . The closer the  $R^2$  value is to 1, the more similar the predicted quantiles are to the actual ones, thus, the better the model describes the data (color figure online)



**Fig. 18.** *Left* best fit curve of  $h^{1,1} - h^{1,2}$  distribution for curve  $q = 2$  based on the Planckian model. *Right* probability plots of our fitted theoretical Planck model versus the  $q = 2, h^{1,1} - h^{1,2}$  distribution

In its current form, the probability plots do not allow us to calculate  $p$ -values of the various models. This due to the same issue encountered previously. If one however standardizes the data according to the  $Z$ -standardization:

$$Z = \frac{X - \mu}{\sigma}, \tag{2.17}$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation, it is possible to calculate the  $p$ -values since the magnitude of each sample gets rescaled. The probability plot of all the models is displayed in the “Appendix”, with the relative  $p$ -values for each model—Fig. 25g, h. What we see is that the modified pseudo-Voigt is statistically the model which provides the best fit.

*2.5. Implications for physics.* Calabi–Yau threefold compactifications of string theory have been the traditional approach to obtaining interesting phenomenological models. The plethora of geometries and configurations, ranging from heterotic strings on Calabi–Yau threefolds endowed with stable bundles, to D-brane probes on local Calabi–Yau varieties, to F-theory compactification on elliptic fibrations, has over the years justified the landscape and inspired various statistical analyses of the space of vacua.

Of particular interest has been the investigation of further structures in the Kreuzer–Skarke database, including identification of “the tip” where Hodge numbers are small [21,35,46], the top bounding curves where Hodge numbers are large [43], identifying elliptically fibered threefolds [28,29,42,44], finding further fibrations such as K3-fibers [33,45], or a step-by-step construction of all possible smooth Calabi–Yau hypersurfaces from the reflexive polytope data [19], etc. Now, it should be emphasized that each of the some 473 million reflexive polytopes admits, as an ambient toric variety, many<sup>7</sup> so-called maximal projective crepant partial (MPCP) desingularization, each of which gives rise to a different Calabi–Yau threefold. Therefore, the actually number of Calabi–Yau threefolds from the Kreuzer–Skarke database is many orders of magnitude larger than  $10^{10}$ . While manifolds coming from the same reflexive polytope have different geometrical data such as triple intersection numbers, which in the standard embedding in heterotic compactification correspond to Yukawa couplings, they do share the same Hodge numbers because these, by virtue of (2.1), depend only on the combinatorics of the polytope. We need to wait for significant theoretical and/or computational advances to have the full data of the Hodge pairs in view of the Calabi–Yau manifolds themselves, which might give new statistics. It would be perhaps even more interesting if the statistics remain largely the same, thereby hinting at some universality in the distribution of such topological data.

In the context of the recent works on F-theory, it is an important fact the vast majority of the Kreuzer–Skarke threefolds are elliptic fibrations over some complex surface, and in fact birational to [42,44,45] a Weierstrass model. For example, some  $10^6$  alone [42] come from elliptic fibrations over  $\mathbb{P}^2$ . Therefore the Kreuzer–Skarke dataset is directly relevant to F-theory. In the more classical context of heterotic strings, the Hodge numbers dictate the number of (anti-)generations in the standard embedding. In our above plots, the Euler number  $\pm 6$  indicate the three generation models. The generic paucity of  $\chi = \pm 6$  manifolds led to the industry of non-standard embedding where extra vector bundle and Wilson line information is needed. The advantage of F-theory models is that

---

<sup>7</sup> The actual numbers are not yet known, but even up to  $h^{1,1} = 7$ , we already see from tens to thousands and with the number increasing potentially exponentially as we go up in Hodge number [19].



the compactification data comes only from the Calabi–Yau manifold. In particular, the intersection theory of the cycles and fiber-degeneration structure determine the gauge group, anomaly cancellation, matter content, and Yukawa couplings. Much of this can be extracted from the polytope data.

F-theory compactifications on threefolds, resulting in six dimensional gauge theories have been considered from the point of view of systematically classifying the base complex surfaces [44] and the statistics have been performed therein. Non-toric bases were considered and a number of Calabi–Yau threefolds beyond the Kreuzer–Skarke data were found. It is remarkable that the overall distribution of Hodge numbers remains largely unchanged. Indeed, in unpublished work of Kreuzer–Skarke, where they extended the hypersurface in toric fourfolds to double hypersurfaces in fivefolds, obtaining some  $10^{10}$  more manifolds and the shape of Fig. 1 persists. All these point to the Kreuzer–Skarke data being a robust representative in the space of Calabi–Yau threefolds. Our distribution subsequently seems a representative sample, and we speculate that analyses of string vacua, in any context, should be thus weighted. For example, in study of the “typical” number of generations in four dimensional heterotic compactification, or of charged matter in six dimensional F-theory compactification, one should superpose our pseudo-Voigt profile.

### 3. Calabi–Yau Twofolds: K3 Surfaces

As noted in the Introduction, there are 4319 data points, corresponding to hypersurfaces as Calabi–Yau twofolds, i.e., K3 surfaces, in reflexive three dimensional polytopes. Being algebraic K3 surfaces, there is only one relevant topological invariant, the Hodge number,  $h^{1,1} = 19$ . However, there is a further refined algebraic quantity for the K3 surface  $X$ , the rank of the Neron–Severi lattice  $H^2(X; \mathbb{Z}) \cap H^{1,1}(X)$ , which is the **Picard Number**  $\rho(X)$  and which enumerates the number of divisors on the surface up to algebraic equivalence. The Picard numbers of the 4319 K3 surfaces were computed in [12]. We present the distribution thereof in Fig. 19a.

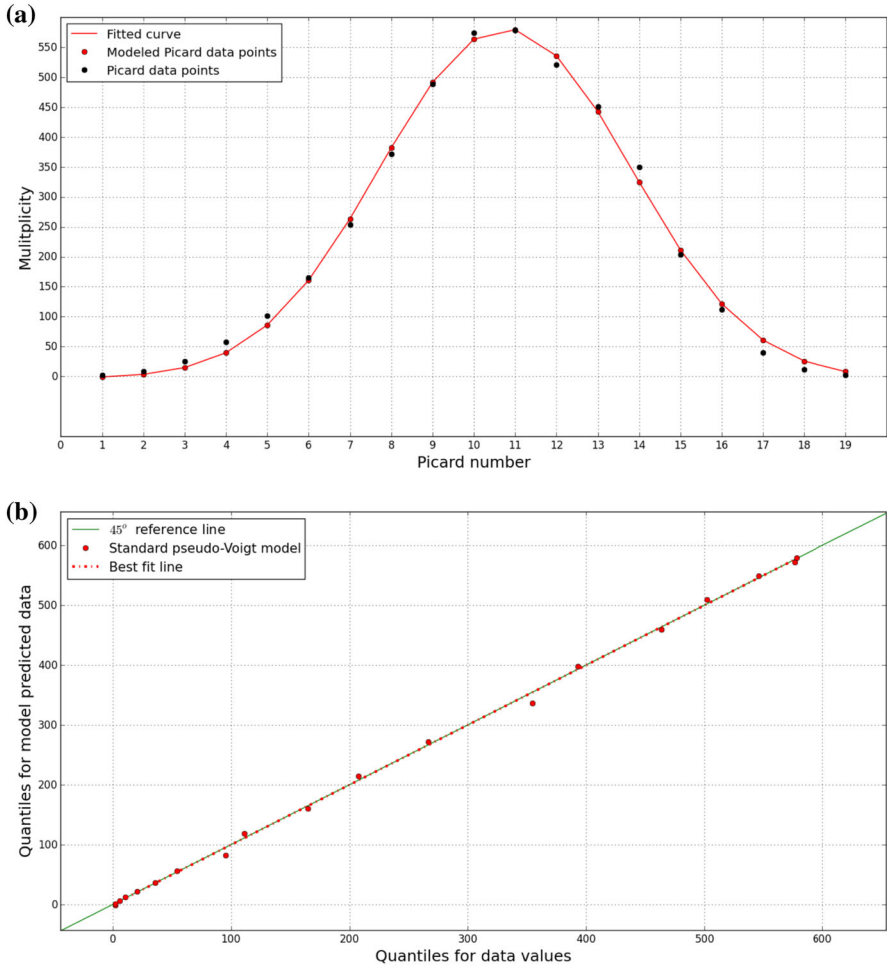
We only used the standard pseudo-Voigt profile as the modified one did not change the fit significantly. Here are the fit statistics for best fit curve:  $(A, \mu, \sigma, \alpha) = (4517.45, 10.76, 2.97, -0.031)$ , as shown in Fig. 19.

What is interesting about Fig. 19a is that the “oscillations” of the actual data points above and below the modeled curve is very apparent, yet modifying the pseudo-Voigt profile is unable to give any significant improvement. This leads to two potential conclusions: (a) the pseudo-Voigt profile is not the best profile to use in combination with an oscillatory component; (b) the manner in which the oscillations occur is not so straightforward as introducing a simple cosine function. An interesting exercise would be to superimpose a cosine function along the distribution, by rotating it as one traverses the profile. As long as the wavelength, amplitude and angle of rotation are all small enough, the continuously rotated cosine function should remain a function everywhere along the profile.

### 4. Calabi–Yau Fourfolds

The analysis of the four fold data is performed in the same spirit as the threefold data. We aim to look for patterns in the frequency plots. Due to complex conjugation and Poincaré duality, the only topological invariants of fourfolds that vary are  $h^{1,1}$ ,  $h^{1,2}$ ,  $h^{1,3}$ , and  $h^{2,2}$ . Three of these are independent [15]:





**Fig. 19.** Using probability plots, we are able to statistically see which model provides the better fit. We employ such graphical methods as standard goodness-of-fit tests, such as the  $\chi^2$  test, fail to give meaningful results. **a** For K3 surfaces, the multiplicity is plotted against Picard number with a pseudo-Voigt fit. **b** Probability plot for the multiplicity quantiles versus the fitted standard pseudo-Voigt quantiles. The  $R^2$  value is 0.99908

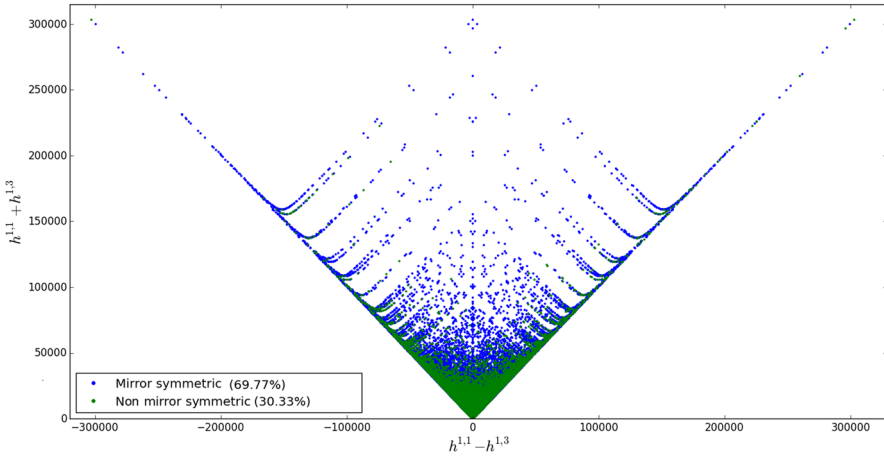
$$h^{2,2} = 44 + 4h^{1,1} - 2h^{1,2} + 4h^{1,3}. \tag{4.1}$$

We compiled a database for the frequency of the triplets  $(h^{1,1}, h^{1,2}, h^{1,3})$  to then obtain the following data structure

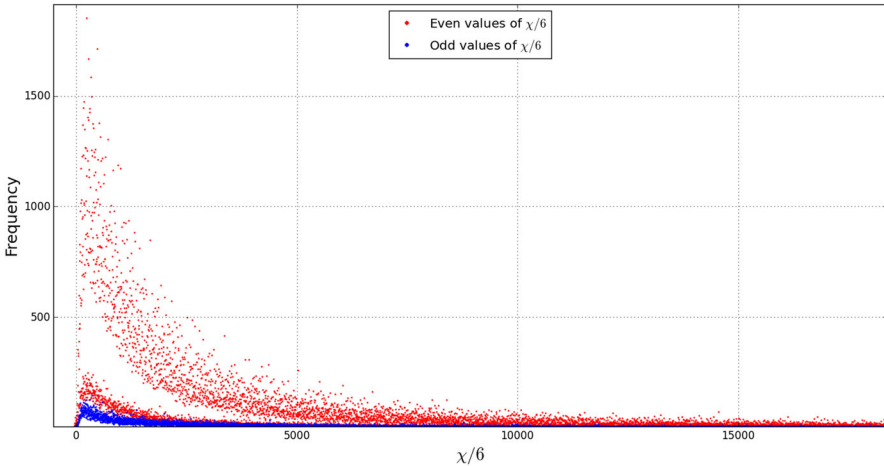
$$(h^{1,1}, h^{1,2}, h^{1,3}, f).$$

Since one expects mirror symmetry within the invariants  $(h^{1,1} \pm h^{1,3})$  [40], a plot of  $h^{1,1} - h^{1,3}$  against  $h^{1,1} + h^{1,3}$  (Fig. 20) should be symmetric about the line  $h^{1,1} - h^{1,3} = 0$ .

Doing a quick analysis of the data yields the following observations: only partial mirror symmetry is found. For 69.77% of data points, the point  $(h^{1,1} - h^{1,3}, h^{1,1} + h^{1,3})$  is accompanied by the point  $(-h^{1,1} + h^{1,3}, h^{1,1} + h^{1,3})$ . Taking frequency into account,



**Fig. 20.** The blue points correspond to manifolds with a mirror symmetric counterpart in the data set (color figure online)



**Fig. 21.** Frequency of Calabi–Yau fourfolds with a given Euler number

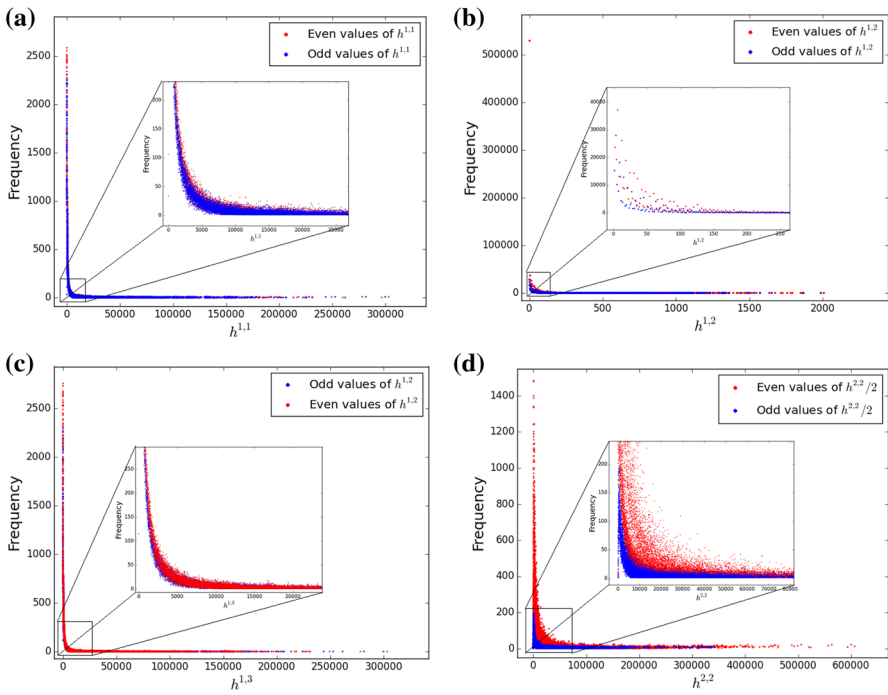
the percentage drops to 27.35%—see Fig. 37 in the “Appendix”. This is most likely due to an incomplete data base.

For now, we have performed a primary analysis on the Euler distribution only. The Euler number for fourfolds is [15]:

$$\chi = 6(8 + h^{1,1} - h^{1,2} + h^{1,3}). \tag{4.2}$$

Interestingly enough, the distinction between even and odd distributions persist in the fourfold data base. For illustrative purposes, we show the distribution of  $\chi/6$  against frequency.

It is not immediately clear what is the reason for the gap, presumably it could be a cluster of data points which is missing from the data base. Until one obtains the complete fourfold data base of Hodge numbers, one can’t say much else. We also preset plots of the individual Hodge numbers  $h^{i,j}$  versus frequency.



**Fig. 22.** The frequency for all the hodge  $h^{i,j}$  numbers. Red points and blue are odd and even points respectively for the various Hodge numbers. The data points are very dense close to the origin making it difficult to properly illustrate the mixing of odd and even Hodge numbers. Only  $h^{2,2}$  c has a clear separation between of an even values. **a**  $h^{1,1}$  versus frequency. **b**  $h^{1,2}$  versus frequency. **c**  $h^{1,3}$  versus frequency. **d**  $h^{2,2}$  versus frequency (color figure online)

### 5. Conclusions and Outlook

By examining the distribution of Hodge numbers of Calabi–Yau manifolds of complex dimension two, three and four, realized as hypersurfaces in toric varieties of one higher dimension as constructed by Kreuzer and Skarke based on the results of Batyrev and Borisov, we have found many hithertofore undiscovered patterns. We summarize our key points as follows.

- For threefolds, there are 30108 distinct pairs of Hodge numbers  $(h^{1,1}, h^{1,2})$  from 473800776 reflexive polytopes, the frequency of both the half-Euler number  $h^{1,1} - h^{1,2}$  and the sum  $h^{1,1} + h^{1,2}$  are distributed according to whether the value is odd or even;
  - The half-Euler number  $h^{1,1} - h^{1,2}$  follows a modified pseudo-Voigt distribution

$$f(x) = (1 - \alpha) \frac{A'}{\sigma \sqrt{2\pi}} e^{-\frac{(x)^2}{2\sigma^2}} + \alpha \frac{A'}{\pi} \left[ \frac{\sigma^2}{x^2 + \sigma^2} \right],$$

where the modification is made in the amplitude  $A$  of the distribution, such that

$$A' = A_0 + b \cos(2\pi \cdot b).$$

There is fine periodic substructure in terms of curves indexed by an integer  $r$ . Our model is accurate for low  $r$ -values ( $r \in [36, 110]$  and  $r \in [37, 99]$ ); using

probability plots as test for goodness of fit, this modified pseudo-Voigt model is indeed the best one out of several standard candidates (cf. Fig. 29 for all the  $R^2$  and  $p$  values).

Among  $A, \sigma, \alpha, b, a$ , the parameters  $\sigma, b, \alpha$  have a strong linear relationship with  $r$ :

Even $r$	Odd $r$
$\sigma(r) = 0.5097r - 12.7142$	$0.51379r - 13.2494$
$\alpha(r) = 2 \times 10^{-4}r - 0.0345$	$2.25 \times 10^{-4}r - 0.0388,$
$b(r) = 3.7299 \times 10^{-5}r + 0.6629$	$7.9101 \times 10^{-5}r + 0.65956$

For a small subset of curves with a low  $r$ -value and an appropriate cut-off frequency, it is extraordinary that the model *exactly fits the data*. That is, it appears that the number of data points for each curve required, such that the model will result in a perfect fit is: 7 for even  $r$ -valued curves and 10 for the odd valued  $r$ -curves, see Fig. 30.

- The quantity  $h^{1,1} + h^{1,2}$  follows a Planckian distribution

$$f(x) = \frac{A}{x^n} \frac{1}{e^{b/(x-22)} - 1}$$

There is a substructure of curves, indexed by an integer  $q$ , each Planckian and with some periodic behavior. The curves  $q_n$  appear clustered into groups of residue classes distinguished by  $n \bmod 6$ , and the parameters  $\log(A), n, b$  all have extremely strong relationships with the  $q$  value.

By substituting this relationship into the model, we have a function  $f_k(x, q)$  that approximately describes the entire  $h^{1,1} + h^{1,2}$  distribution up to a  $q$  value of 69, 100:

$$f_k(x, q) = \frac{e^{\sum_{i=0}^4 A_{k,i} q^i}}{x^{\sum_{i=0}^4 n_{k,i} q^i}} \frac{1}{\left( e^{\frac{\sum_{i=0}^4 b_{k,i} q^i}{(x-22)}} - 1 \right)}, \tag{5.1}$$

with  $k = 0, 1, \dots, 5$  and the coefficients given in A.8, A.9, A.10.

- The Euler number  $\chi = 2(h^{1,1} - h^{1,2})$  follows the modified pseudo-Voigt distribution composed with a sinusoidal  $A + A_0 + a \cos(2\pi b \cdot x)$  which is almost an exact fit, with the coefficients given by  $(A_0, \sigma, \alpha, b, a) = (1.9032 \times 10^9, 75.8305889, 0.00718459, 0.58347826, 8.7427 \times 10^7)$ , at  $R^2 = 0.99944$  for even  $\chi$  and  $(1.9032 \times 10^9, 75.8305889, 0.00718459, 0.58347826, 8.7427 \times 10^7)$  at  $R^2 = 0.99965$  for odd  $\chi$ ,

The modified pseudo-Voigt distribution is remarkably accurate in predicting the overall and fine sub-structure of the Euler number distribution.

- For K3 surfaces, we have looked at the distribution of the multiplicity with Picard number. We find that this distribution follows a standard pseudo-Voigt profile. Adding in the sinusoidal modification does not significantly increase the overall fit. The parameters are given by  $(a, \mu, \sigma, \alpha) = (4517.45, 10.76, 2.97, -0.031)$  with  $R^2 = 0.99908$ .
- For Calabi–Yau fourfolds, there is no exact mirror symmetry, due to incompleteness of available data. Nevertheless, by breaking up the data into three groups, we have
  - Mirror symmetric partners with the same frequency: 27.35%

- Mirror symmetric partners without the same frequency: 42.22%
- Non mirror symmetric partners: 30.33%

By plotting the various  $h^{i,j}$  versus frequency we see there is no distinction between even and odd data values for  $h^{i,j}$ , expect for  $h^{2,2}/2$ . This distinction is carried out further in the Euler number distribution where odd points are clustered on a band with much lower frequencies. The even values of  $\chi/6$  appear to be distributed along to separate bands.

It is remarkable how well the pseudo-Voigt distribution, modified with a sinusoidal component, fits the distribution of topological numbers of toric Calabi–Yau manifolds, often giving an exact fit. Of course, what we are studying at heart is the number of integer points inside (cf. (2.1)) reflexive polytopes. This is a highly non-trivial counting problem whose answer will ultimately give full analytic results for our distributions and we suspect that the answer should be some generalized pseudo-Voigt function.

Now, in addition of Calabi–Yau manifolds, stable vector bundles over various such manifolds in a variety of construction beyond Kreuzer–Skarke have also been studied algorithmically over the years in the context of heterotic compactification (cf. e.g., [23–26]). One can see a somewhat pseudo-Voigt profile in these as well, even though there is no underlying polytope and the counting problem is dictated by certain Diophantine system. It would be interesting to see why this shape is universal in such classifications.

*Acknowledgements.* We are grateful to Cyril Matti for collaboration during the early stages of this project. We thank Mark Dowdeswell for his input with regards to the goodness-of-fits for the various plots. YHH is indebted to the Science and Technology Facilities Council, UK, for grant ST/J00037X/1, the Chinese Ministry of Education, for a Chang-Jiang Chair Professorship at NanKai University, and the city of Tian-Jin for a Qian-Ren Award. YHH is also perpetually indebted to Merton College, Oxford for continuing to provide a quiet corner of Paradise for musing and contemplations. VJ and LP are supported by the South African Research Chairs Initiative of the Department of Science and Technology and the National Research Foundation.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## A. Appendix

Here we include all additional plots to supplement the main body. This includes the relevant plots for the odd distributions—since in the main text we only presented the plots for even distributions—as well as the regression analysis statistics and parameter values for both distributions.

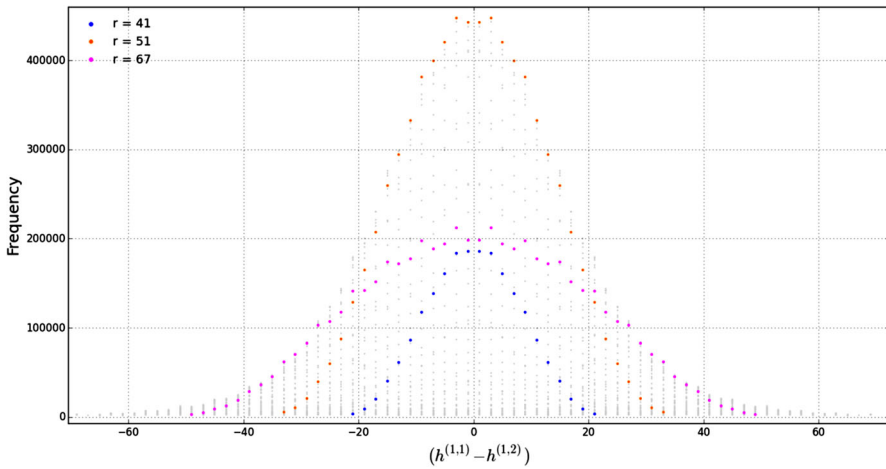
*A.1. Supplementary plots for the  $h^{1,1} - h^{1,2}$  distribution.* All even plot counterparts will be referenced in the figures. The plots appear in the same order as in the main body, with descriptions only if necessary.

*A.1.1. Plots for the odd distribution as counterparts to the even ones*

*A.1.2. Comparative plots* Here we present a comparison of various models we used, by plotting them side by side with the relevant fit-statistics. We choose a single even curve,  $r = 54$ , and odd curve,  $r = 51$ , to illustrate the difference between models.

### Gaussian Model

$$f(x, A, \mu, \sigma) = \frac{A}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{A.1}$$



**Fig. 23.** Three highlighted curves ( $r = 41, 51, 67$ ) within the odd  $h^{1,1} - h^{1,2}$  distribution. The transparent grey data dots is the rest of the distribution. Refer to Fig. 4 for the even plot

**Lorentzian Model**

$$f(x, A, \mu, \sigma) = \frac{A}{\pi} \left[ \frac{\sigma}{(x - \mu)^2 + \sigma^2} \right] \tag{A.2}$$

**Pearson7 Model**

$$f(x, A, \mu, \sigma, m) = \frac{A}{\sigma \beta(m - \frac{1}{2}, \frac{1}{2})} \left[ 1 + \frac{(x - \mu)^2}{\sigma^2} \right]^{-m}, \tag{A.3}$$

where  $\beta$  is the Beta function.

**Breit–Wigner Model**

This model is based on the Breit-Wigner function.

$$f(x, A, \mu, \sigma, t) = \frac{A(t\sigma/2 + x - \mu)^2}{(\sigma/2)^2 + (x - \mu)^2} \tag{A.4}$$

**Voigt Model**

$$f(x, A, \mu, \sigma, \gamma) = \frac{a \text{Re}[z]}{\sigma \sqrt{2\pi}} \tag{A.5}$$

where

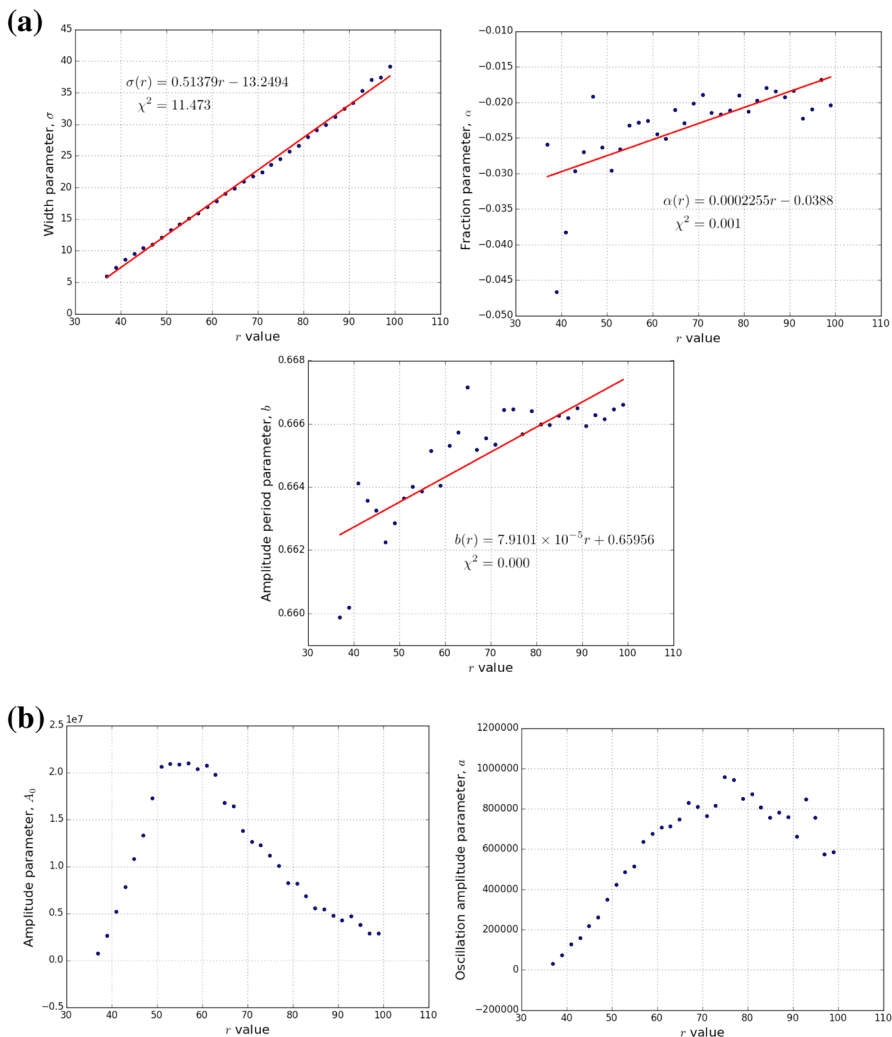
$$z = \frac{x - \mu + i\gamma}{\sigma \sqrt{2}}, \quad w(z) = e^{-z^2} \text{erfc}(-iz) \tag{A.6}$$

The Voigt model is a convolution of the Gaussian and Lorentzian models.

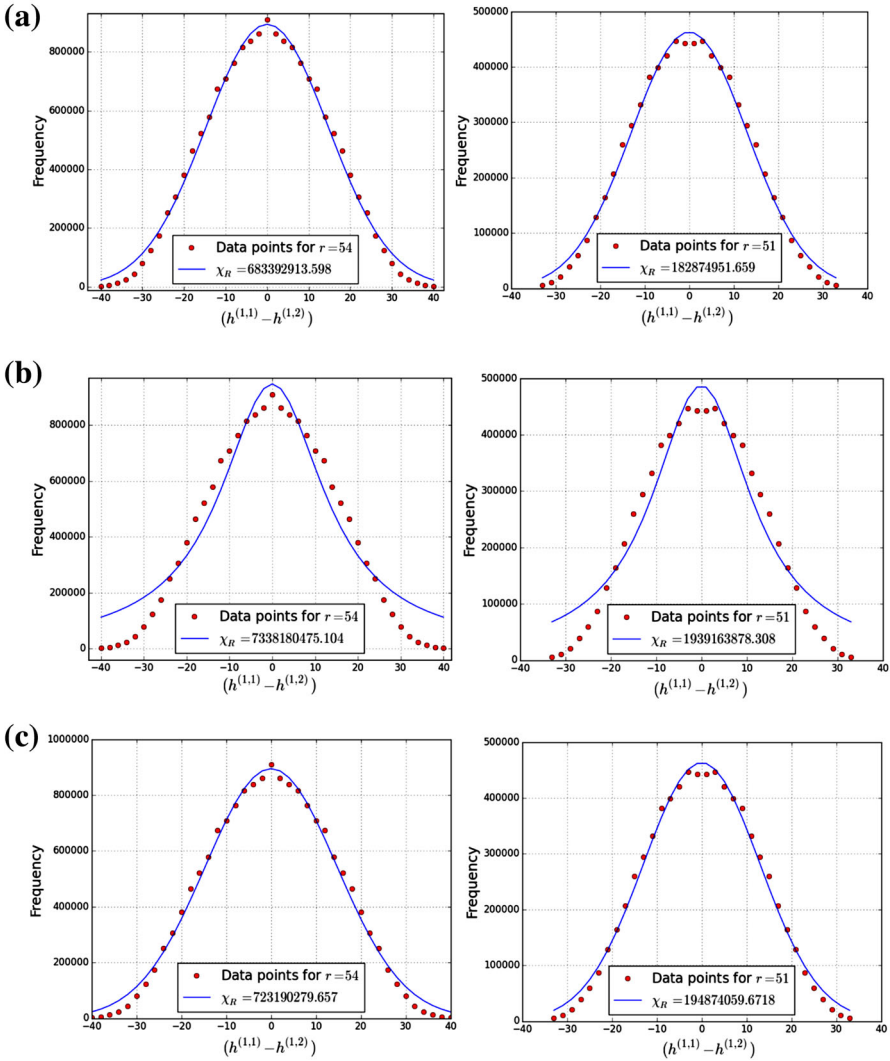
**Pseudo-Voigt Model**

$$f(x, A, \mu, \sigma, \alpha) = (1 - \alpha) \frac{A}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} + \alpha \frac{A}{\pi} \left[ \frac{\sigma^2}{(x - \mu)^2 + \sigma^2} \right] \tag{A.7}$$

We present the standardized and shifted probability plots for the above comparisons:



**Fig. 24.** The plots of the various parameters  $A$ ,  $\sigma$ ,  $\alpha$ ,  $b$ ,  $a$  versus  $r$  for odd values of  $r$ . **a** The width parameter  $\sigma$  has a linear relationship with  $r$  such that  $\sigma(r) = 0.51379r - 13.2494$ . The amplitude period parameter,  $b$ , also has a linear relationship, however, since  $r$  is at most order 3 in magnitude, we can regard it approximately as a constant such that  $b(r) = 0.65956 \sim 2/3$ . The same goes for the fraction parameter,  $\alpha$ , we can regard it as a constant such that  $\alpha(r) = -0.0388$ . For even parameter fit statistics see Fig. 10. **b** Plots of  $A_0$  versus  $r$  (left) and  $a$  versus  $r$  (right). Both exhibit a similar pattern, however it is difficult to find any nice relationships. For even parameter plots see Fig. 10



**Fig. 25.** For all models, the *left* hand graph is for  $r = 54$  and the right is for  $r = 51$ . The probability plot presents all the models together. All the above mentioned modeled are included to compare their resemblance with the actual data. The larger the  $p$  value the better the line  $y = x$  fits the data, implying the better the model is at describing the data. **a** Gaussian model. **b** Lorentzian (Cauchy) model. **c** Pearson7 model. **d** Breit–Wigner model. **e** Voigt model. **f** Pseudo-Voigt model. **g** The probability plot for  $r = 51$ . **h** The probability plot for  $r = 54$

**A.1.3. A first approximation to the data** The overall behavior of the data across each curve is modeled extremely well using the pseudo-Voigt model. Here we present a few plots illustrating a first approximation to the data. A second approximation can be made by introducing an oscillating amplitude as described in Sect. 2.1

**A.1.4. Table of parameter values and statistics** Here we present the parameter values as well as the reduced  $\chi$  value,  $\chi_R$ , in a tabular format for all even  $r$  curves— $r \in [34, 120]$ —and for all odd  $r$  curves— $r \in [35, 99]$ .



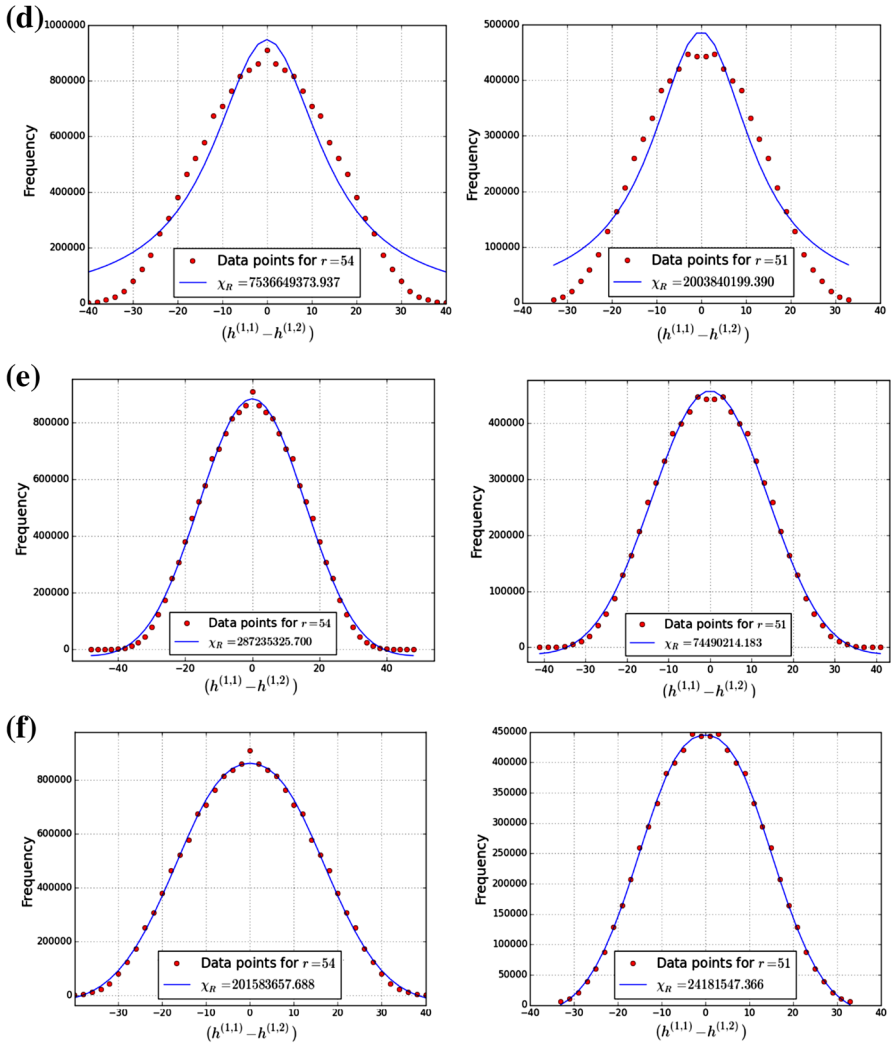


Fig. 25. continued

A.2. Supplementary plots for the  $h^{1,1} + h^{1,2}$  distribution.

A.2.1. Plots for the odd distribution as counterparts to the even ones All even plot counterparts will be referenced in the figures. The plots appear in the same order as in the main body, with descriptions only if necessary.

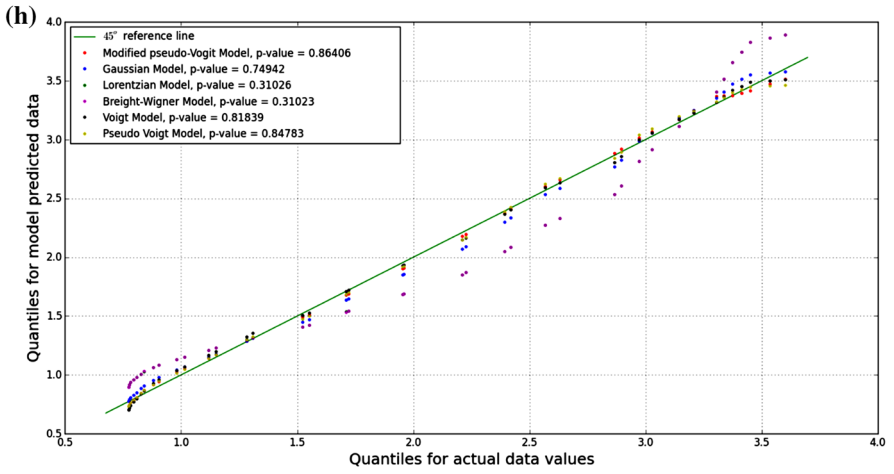
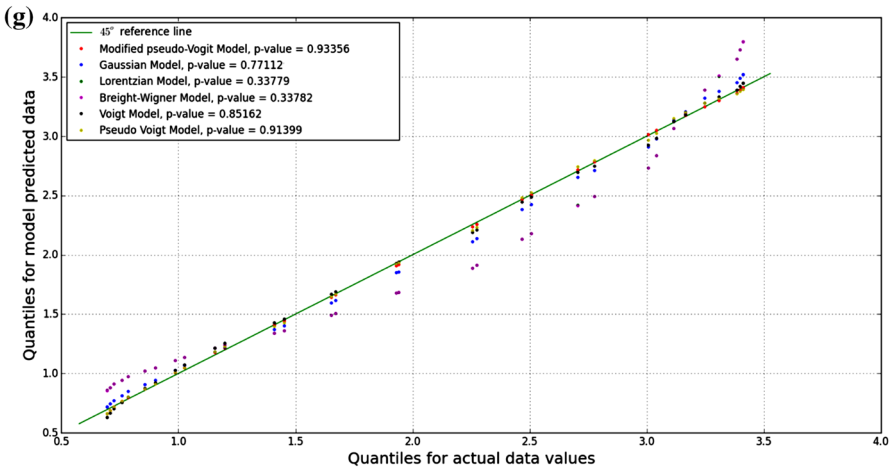
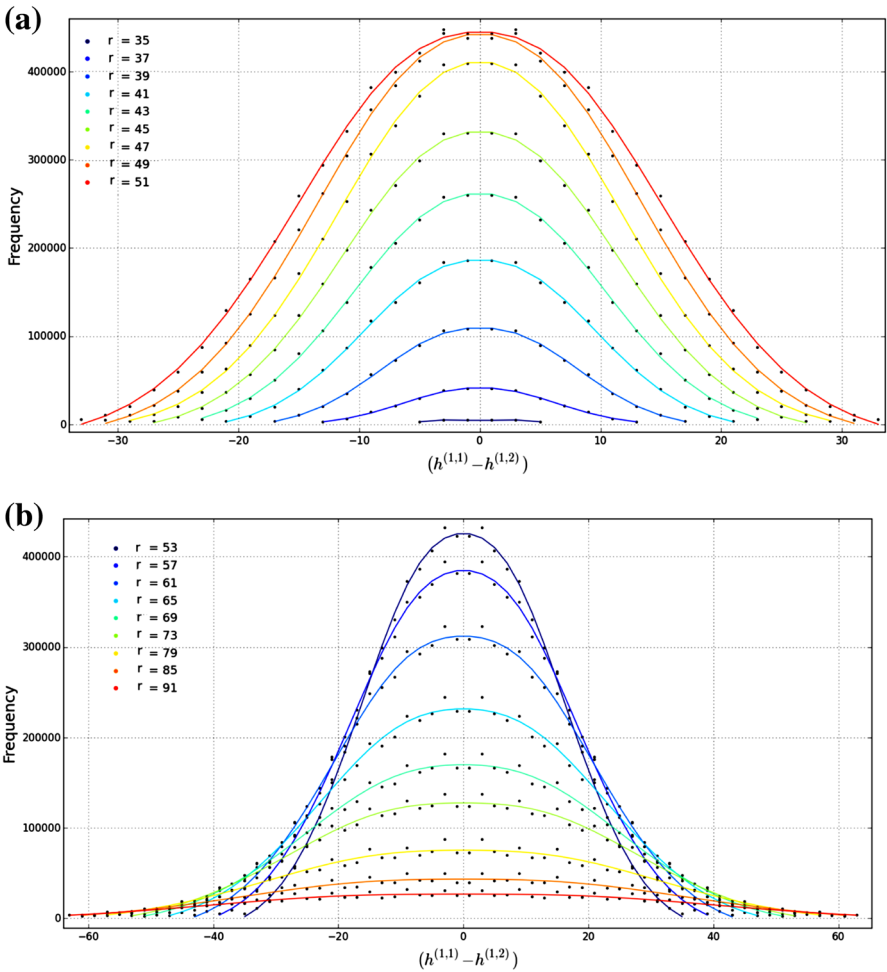


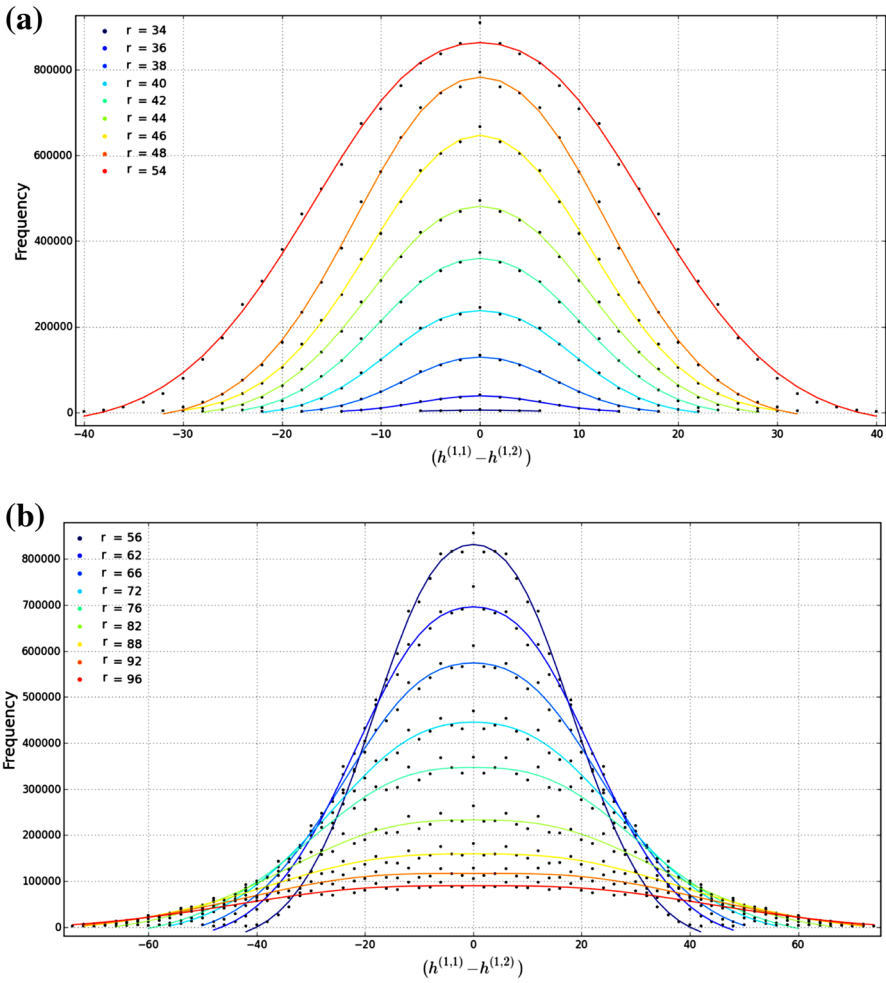
Fig. 25. continued

A.2.2. Table of parameter values, coefficient values and statistics **Coefficient values for the description of the entire  $h^{1,1} + h^{1,2}$  distribution**



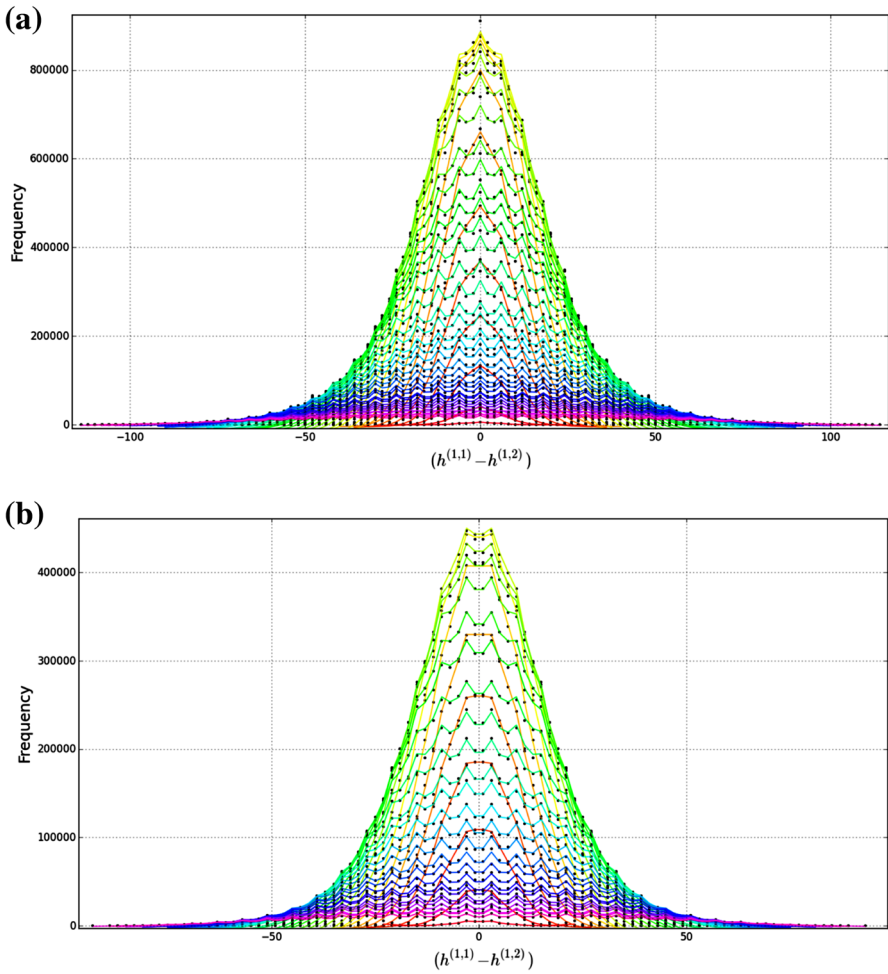
**Fig. 26.** Best fit curve based on the pseudo-Voigt model for the same sets of curves as seen in Fig. 5. **a** Regression lines for few select even  $r$  values, with  $r \in [35, 51]$ . **b** Regression lines for few select even  $r$  values, with  $r > 51$

$$A_{k,i} = \begin{pmatrix} 54.2664195 & 2.9066 \times 10^{-16} & 0.02414823 & -5.4137 \times 10^{-20} & -7.2635 \times 10^{-7} \\ 65.0676835 & -2.0296 \times 10^{-16} & 0.03354614 & 3.7552 \times 10^{-19} & -3.1443 \times 10^{-7} \\ 54.8909275 & -2.0323 \times 10^{-16} & 0.02753302 & -2.7091 \times 10^{-20} & -9.1972 \times 10^{-7} \\ 62.6423777 & 1.2736 \times 10^{-16} & 0.03020535 & -1.1234 \times 10^{-19} & -8.6929 \times 10^{-7} \\ 54.5840853 & 2.9011 \times 10^{-16} & 0.02748121 & -9.4235 \times 10^{-20} & -9.3840 \times 10^{-7} \\ 64.2001359 & -1.3980 \times 10^{-16} & 0.03700128 & 8.3795 \times 10^{-20} & -1.3712 \times 10^{-7} \end{pmatrix} \tag{A.8}$$



**Fig. 27.** Best fit curve based on the pseudo-Voigt model for the same sets of curves as seen in Fig. 6. **a** Regression lines for few select even  $r$  values, with  $r \leq 54$ . **b** Regression lines for few select even  $r$  values, with  $r > 54$

$$b_{k,i} = \begin{pmatrix} 132.357878 & 3.3411 \times 10^{-15} & 0.32753297 & -8.6619 \times 10^{-19} & 4.5825 \times 10^{-6} \\ 184.853063 & -5.7999 \times 10^{-17} & 0.31981034 & 1.0014 \times 10^{-18} & 3.9052 \times 10^{-5} \\ 117.228782 & -1.2791 \times 10^{-15} & 0.36989364 & -8.5325 \times 10^{-20} & 2.9743 \times 10^{-6} \\ 173.033950 & -1.1829 \times 10^{-15} & 0.31584408 & 8.9872 \times 10^{-19} & 2.5454 \times 10^{-5} \\ 105.298297 & 5.7916 \times 10^{-15} & 0.37843953 & -1.5078 \times 10^{-18} & 1.3974 \times 10^{-6} \\ 171.521189 & 1.5811 \times 10^{-15} & 0.36410293 & -2.5726 \times 10^{-19} & 2.5139 \times 10^{-5} \end{pmatrix} \quad (\text{A.9})$$



**Fig. 28.** This is what the entire distribution looks like using our modified pseudo-Voigt model. See Fig. 29 for the fitted coefficients as well as the fits for every curve given by the probability plots. **a** Every fitted even curve from  $r = 34$  until  $r = 120$ . **b** Every fitted even odd from  $r = 35$  until  $r = 99$

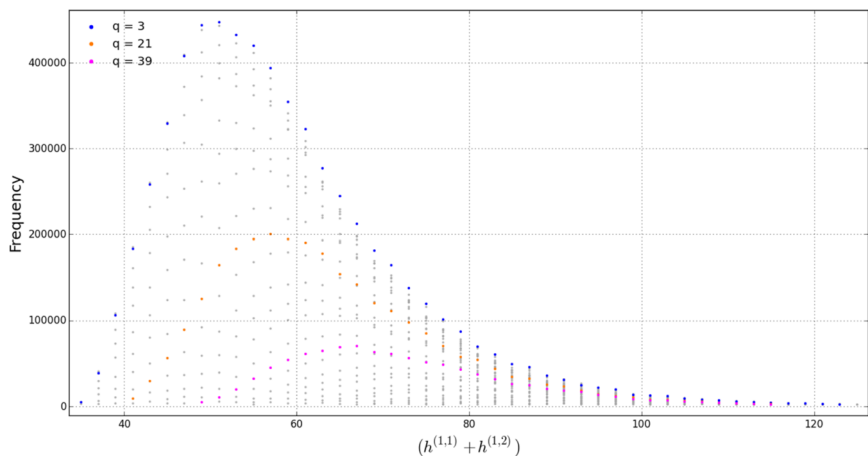
$$n_{k,i} = \begin{pmatrix} 8.98205242 & 2.9066 \times 10^{-17} & 0.00434183 & -6.7671 \times 10^{-21} & -1.5512 \times 10^{-7} \\ 11.6018246 & 5.1148 \times 10^{-17} & 0.00644305 & 0 & -1.7241 \times 10^{-7} \\ 9.19515076 & 4.3161 \times 10^{-17} & 0.00496066 & -1.3763 \times 10^{-20} & -1.9163 \times 10^{-7} \\ 11.0620173 & -1.1446 \times 10^{-18} & 0.00570064 & 2.8085 \times 10^{-20} & -2.4813 \times 10^{-7} \\ 9.15798913 & 5.0109 \times 10^{-17} & 0.00493009 & -2.3559 \times 10^{-20} & -1.9210 \times 10^{-7} \\ 11.4578629 & -6.0813 \times 10^{-18} & 0.00705818 & 9.2055 \times 10^{-21} & -3.5862 \times 10^{-7} \end{pmatrix} \tag{A.10}$$

*A.3. Supplementary plots for the fourfold data.* When looking for mirror symmetry in the fourfold data, we only observed partial mirror symmetry. Below is the full break down of the data set.

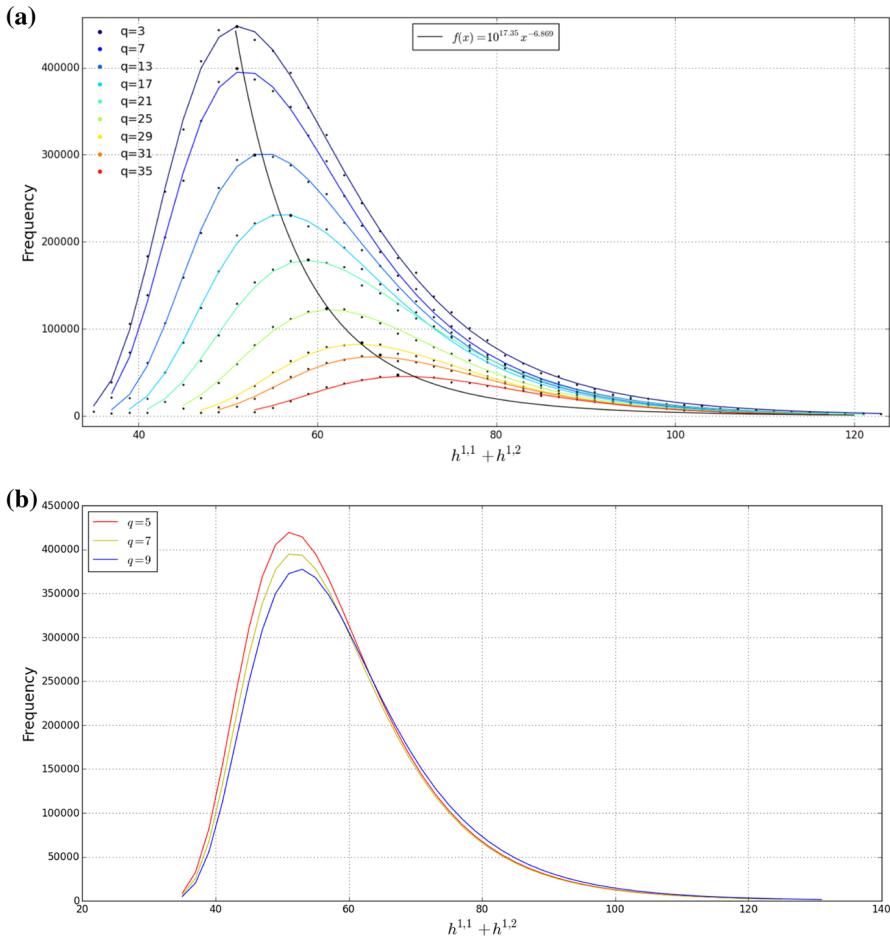
$r$	$A_0$	$\sigma$	$\alpha$	$b$	$a$	$\chi^2_k$	$R^2$	$p$	$r$	$A_0$	$\sigma$	$\alpha$	$b$	$a$	$\chi^2_k$	$R^2$	$p$
34	74808.082	5.61029	0.003764698	0.671324793	11882.85554	1913.32108	1	1	35	69517.6993	5.27052174	-0.00059769	0.66801823	11501.6207	2615.83922	0.98471088	0.87358981
36	62112.5048	6.14542	-0.009876003	0.667458363	28438.58633	40004.88553	0.99902292	0.91717668	37	66682.118	5.8927626	-0.01836059	0.66878222	27241.556	53480.6329	0.99993762	0.97888572
38	2545950.511	7.04214	-0.021320726	0.661106908	70029.4258	77592.3274	0.99989369	0.96783088	39	2418862.36	7.26453884	-0.03399152	0.66624643	67134.7518	157852.39	0.9998798	0.97276103
40	5997498.444	8.38473	-0.027896601	0.664313042	15021.8977	167626.1551	0.99981973	0.95476703	41	4946884.13	8.55589054	-0.02144231	0.66478941	118674.657	141478.1	0.99965101	0.9841959
42	10051953.39	9.93476	-0.023538526	0.664313085	214365.7566	1124838.16	0.99960558	0.94463388	43	74335.112	9.48971721	-0.02378655	0.66566727	149650.971	440063.37	0.99983906	0.95857146
44	1438706.27	10.1952	-0.019800045	0.663363561	27921.3165	1015624.68	0.999910944	0.97052621	45	1041088.73	10.352866	-0.02296192	0.66294576	209654.283	662453.8	0.99987931	0.94607889
46	1893027.6	10.6388	-0.01402811	0.664897086	363905.1143	1263490.69	0.999827533	0.938642702	47	13000374.5	10.354384	-0.0176622	0.6624124	254360.624	570395.19	0.99985986	0.96200221
48	2691644.43	11.52	-0.01907394	0.665758045	456461.6269	4801834.01	0.99960734	0.948013265	49	16832005.1	11.990223	-0.02272541	0.66219193	334570.92	1417879.3	0.99988501	0.9711674
50	35415476.33	12.7568	-0.02505604	0.663930639	648055.0793	15949031.33	0.999163785	0.888518617	51	19624474.4	13.181995	-0.02585479	0.66328760	420525.527	2427117.9	0.99977948	0.93356734
52	4053164.09	13.9486	-0.025150833	0.663741398	770677.2752	17803931.93	0.999329887	0.887202079	53	2004655.14	14.1510006	-0.02328645	0.66408311	465066.332	24857709	0.9997502	0.93861201
54	4208478.16	14.9145	-0.02088962	0.664242039	851781.3562	17783077.53	0.998744804	0.86462454	55	20316683.4	14.9770179	-0.02144823	0.66400324	479764.277	1827092.5	0.99954727	0.9163107
56	45318925.17	15.9308	-0.02245431	0.664639342	1081188.801	91024311.13	0.99961346	0.902838544	57	2064175.15	15.8792577	-0.0219817	0.66478694	418779.2	1999973.66	0.9997336	0.941276
58	4577655.84	16.87025	-0.020439012	0.66390829	116222.825	8951308.6	0.99950544	0.915806564	59	19628194.9	16.839466	-0.02028947	0.6646865	650103.969	1193409.5	0.99976798	0.93882446
60	4583436.12	17.6159	-0.019455309	0.664461299	1195317.91	6732478.91	0.99975046	0.937493846	61	19631203.5	17.88731	-0.02146377	0.6650147	671348.429	1263047.1	0.99979416	0.9418219
62	4589423.65	18.7829	-0.02008061	0.664693685	1299727.161	9592028.64	0.99925833	0.885179311	63	1881116.5	18.997709	-0.0226892	0.66582004	79837.657	16307189.7	0.99981185	0.8886674
64	4462920.2	19.8429	-0.020615871	0.665093096	1347466.7	78028169.68	0.99988021	0.973313468	65	16183205.1	19.772633	-0.01933318	0.66722014	714485.873	1330073.2	0.99989661	0.9565573
66	41519682.02	20.5755	-0.018305682	0.666138254	1466283.568	54033929.56	0.999339136	0.882676184	67	15604477.3	20.8972797	-0.02026454	0.66500765	798587.968	8036678.17	0.99938287	0.8313905
68	3971267.25	21.4871	-0.0176357	0.66544129	16021010.61	409337945.3	0.999205972	0.892005972	69	13104503.1	21.6694017	-0.0184441	0.66559333	76710.979	6136235	0.99891504	0.88422783
70	3680787.68	22.0999	-0.015684425	0.665847362	1557320.642	33793439.97	0.999174607	0.878072158	71	12181331.1	22.345088	-0.01735848	0.66535239	73723.294	7981281.47	0.99934345	0.88305521
72	36162771.81	23.0026	-0.016476545	0.666763067	1581895.238	21524913.66	0.999863961	0.917179602	73	11688172.3	23.5912389	-0.02059164	0.66581685	77831.62	69955.7	0.999573	0.9013663
74	36144785.21	24.0403	-0.01782108	0.666499201	1872368.376	2864010.59	0.99932336	0.87539384	75	10374775.4	24.4411893	-0.018943	0.66648826	888710.412	2162581.62	0.99959387	0.93108108
76	3449648.35	25.2339	-0.01860589	0.666121761	1800563.649	44636083.29	0.998671434	0.824721608	77	9517481.18	25.6570194	-0.01923599	0.66566431	893712.308	282101.59	0.99845812	0.84029783
78	32892615.97	26.1559	-0.018823059	0.666381088	2189453.136	33663175.17	0.998301995	0.82834669	79	7885048.98	26.6103987	-0.0174641	0.6650459	817061.611	2805786.84	0.99843597	0.87271938
80	30667275.71	27.8144	-0.019368889	0.666925848	2025953.144	2721895.144	0.999747881	0.898977327	81	45734480.8	27.9448661	-0.0201979	0.6650062	898533.404	1066668.68	0.99846468	0.83248626
82	27351655.4	28.6931	-0.017490104	0.666752586	2011512.935	2628425.4	0.99726603	0.770596127	83	6530766.47	29.1128881	-0.0184932	0.66597173	771581.758	154872.99	0.99621037	0.6969065
84	24566921.31	29.8261	-0.016927048	0.666202732	1732875.478	2309745.94	0.99559355	0.706834711	85	5276286.62	29.9134628	-0.01662316	0.6662315	721999.658	101236.36	0.99647031	0.66124585
86	2299616.56	30.8169	-0.01644321	0.666095338	18911979.09	14479232.26	0.997492504	0.74438007	87	5180484.66	31.1367808	-0.0174286	0.66615796	74982.031	76958.124	0.99744609	0.72808704
88	2153840.71	32.153	-0.01614982	0.666181087	1804104.196	18088956.61	0.995744097	0.67804628	89	4543978.97	32.4824891	-0.01819232	0.66650101	24406.102	47057.134	0.9955444	0.6467876
90	19886629.73	33.3369	-0.01668135	0.666516895	1783587.312	11527988.91	0.998471494	0.69352255	91	4114525.48	33.32904	-0.0178282	0.66659154	64509.413	43027.338	0.99581489	0.6019004
92	1848699.11	34.3926	-0.017242388	0.66626368	1741577.927	537588.556	0.99848737	0.769625304	93	4317572.14	35.4965215	-0.0206715	0.66636079	78310.682	40816.122	0.99672509	0.7503965
94	1889782.16	36.3033	-0.018185807	0.666336104	1925176.842	835284.233	0.987709885	0.561674719	95	3525255.74	37.183239	-0.01908989	0.66644729	708049.931	53115.258	0.9881393	0.507914
96	1488989.87	38.1253	-0.01814316	0.66646019	1920878.376	4987610.018	0.99861999	0.71928676	97	2781781.76	37.9448661	-0.01605578	0.66642265	55843.656	63366.622	0.9818748	0.3367676
98	14740741.13	37.7735	-0.016672525	0.666516294	16933174.327	4104612.131	0.996716296	0.711028368	99	2721520.91	39.2526357	-0.0196972	0.66656673	558530.974	196164.967	0.99611455	0.6092328
100	1373455.92	39.1882	-0.016641518	0.666363484	1338642.314	3437131.53	0.994423193	0.5603671									
102	1115067.97	39.4973	-0.013892005	0.666445058	1599292.476	468926.072	0.999951211	0.499880801									
104	933946.392	40.153	-0.0121718	0.666411434	1159602.308	37172.62	0.99154808	0.46670647									
106	868491.797	41.0922	-0.0126555	0.666256758	1185946.732	264510.089	0.993791084	0.560245553									
108	838082.944	42.139	-0.013639004	0.666558606	1104233.414	1209490.828	0.996572623	0.62430704									
110	8195376.057	44.0026	-0.014513112	0.666306562	1162489.945	922920.416	0.996020933	0.58859257									
112	7991588.586	44.9151	-0.01519218	0.666095555	1170845.208	578070.616	0.99570384	0.646776002									
114	7502725.304	47.0497	-0.015129142	0.666362124	1206184.755	1116881.881	0.993672141	0.627571007									
116	6781922.138	48.3831	-0.015377227	0.714161497	-72768.74336	4239217.226	0.991182119	0.440649433									
118	6003445.42	49.8567	-0.014286543	0.666240388	107317.828	1377485.637	0.974412536	0.334976999									
120	5081179.345	50.9995	-0.013978989	0.6662345881	907002.9689	520853.6576	0.985749662	0.266875287									

Fig. 29. Left list of best fit coefficients for all even curves  $r \in [34, 120]$ . Right list of best fit coefficients for all odd curves  $r \in [35, 99]$ . In both figures, the last two columns represent the  $R^2$  and  $p$  values for the probability plot for each curve. The  $p$ -values were obtained by first performing a Z-Standardization on the data

Even				Odd					
r-value	Max F	% Cut off	Number of data points	r-value	Max F	% Cut off	Number of data points		
			Total	At cut off			Total	At cut off	
28	3	0	7	7	29	3	6	6	
30	99	13.13	11	9	31	22	9.09	12	8
32	768	9.6	23	9	33	553	4.88	20	10
34	6258	15.1	25	9	35	5180	19.3	22	10
36	40739	24.35	27	9	37	40607	16.25	24	10
38	133355	35.99	31	9	39	108236	32.34	28	10
40	244716	50.26	35	9	41	185481	46.9	30	10
42	373126	69.68	33	7	43	259859	53.49	34	10
44	494185	76.89	37	7	45	330009	59.99	36	10
46	666992	73.76	41	7	47	408779	61.89	38	10
48	793852	80.74	43	7	49	443162	69.95	40	10
50	877191	82.42	43	7	51	447109	74.45	42	10
52	875275	86.6	45	7	53	432081	76.37	46	10
54	910113	84.6	49	7	55	419456	77.24	46	10
56	816288	92.86	49	7	57	393842	86.33	48	10
58	793170	92.54	51	7	59	354495	81.52	5	

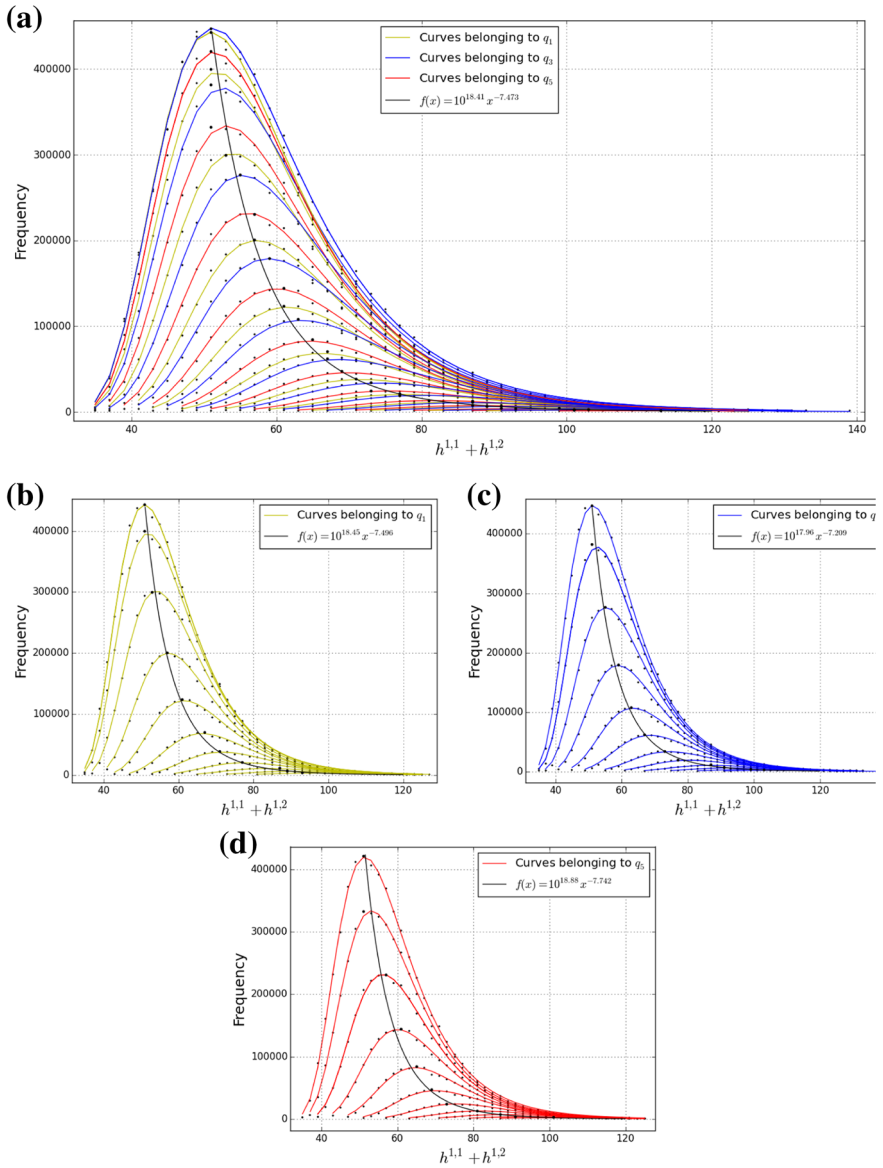


**Fig. 31.** Three highlighted curves ( $q = 3, 19, 31$ ) within the odd  $h^{1,1} + h^{1,2}$  distribution. The transparent grey data dots are all the data plots for the distribution. Refer to Fig. 11 for the even plot

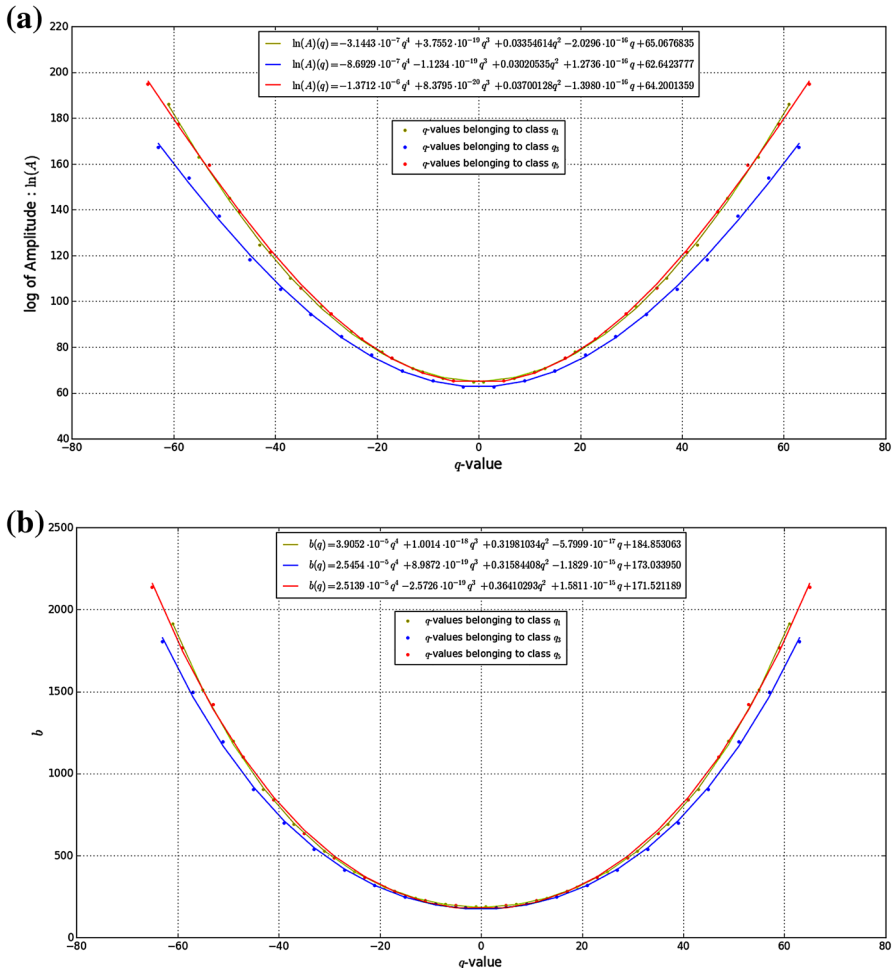


**Fig. 32.** In the attempt to describe the data analogously to a blackbody distribution (a), we discover some subtle structure (b). These are the odd counterparts to Fig. 12. **a** Lines of best fit from a regression analysis for a few select curves. The *black data* points represent the maximum frequency for that particular  $q$  – curve. The *black line* is a line of best fit to describe the points of maximum frequency—this is analogous to a blackbody spectrum. See Fig. 12a for the curves within the even distribution. **b** The curves segregate into three classes determined by the value of the even integer modulo 6. A similar pattern occurs in the even distribution; see Fig. 12b





**Fig. 33.** We illustrate the added structure for odd  $h^{1,1} + h^{1,2}$  data, by displaying how the regression curves can be divided into residue classes. For the list of even curves, refer to Fig. 13. **a** All the curves color coded according to what residue class their curves  $q_n$  belongs to. **b** Family of curves all belonging to  $q_1$ . **c** Family of curves all belonging to  $q_3$ . **d** Family of curves all belonging to  $q_5$



**Fig. 34.** The parameter plots are *color* coded according to what residue class their  $q$  value belong to. For the relationships in the even distribution, see Fig. 14. **a** Plotting the  $q$ -value parameter versus the  $\log(A)$  parameter. **b** Plotting the  $q$ -value parameter versus the  $b$  parameter. **c** Plotting the  $q$ -value parameter versus the power  $n$  parameter (color figure online)

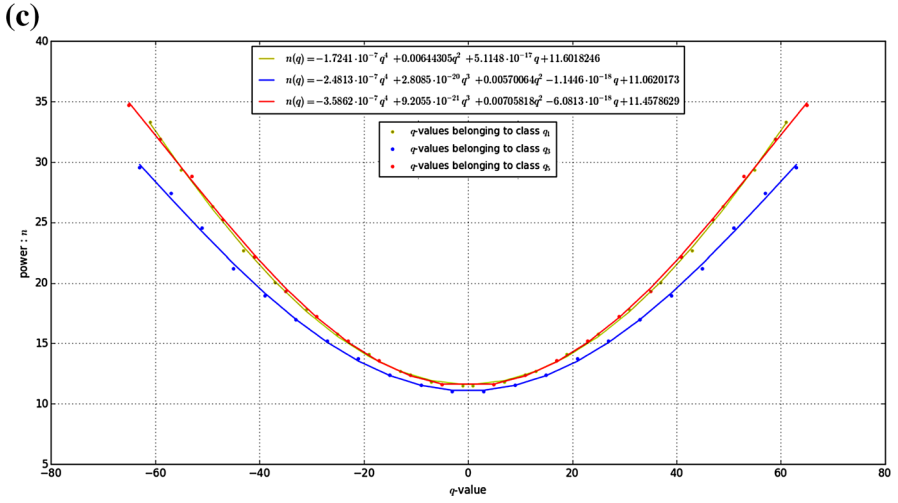


Fig. 34. continued

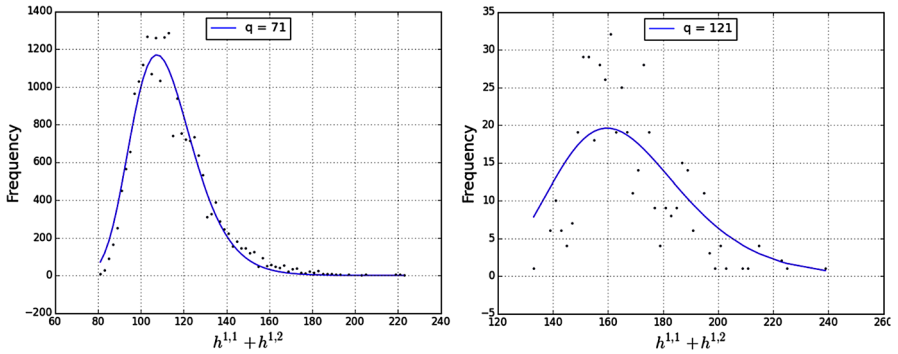


Fig. 35. Left figure is the fitted model (blue line) for a  $q$  value of 71 and right has a  $q$  value of 121. As the  $q$ -value increases, the scattering of the data points within  $h^{1,1} + h^{1,2}$  increases to the point where the model works no longer. For an example of how the model begins to break down at large  $q$ , see Fig. 15



## References

1. Candelas, P., Horowitz, G.T., Strominger, A., Witten, E.: Vacuum configurations for superstrings. Nucl. Phys. B **258**, 46 (1985)
2. Candelas, P., Dale, A.M., Lutken, C.A., Schimmrigk, R.: Complete intersection Calabi–Yau manifolds. Nucl. Phys. B **298**, 493 (1988)
3. Candelas, P., Lutken, C.A., Schimmrigk, R.: Complete intersection Calabi–Yau manifolds. 2. Three generation manifolds. Nucl. Phys. B **306**, 113 (1988)
4. Gagnon, M., Ho-Kim, Q.: An exhaustive list of complete intersection Calabi–Yau manifolds. Mod. Phys. Lett. A **9**, 2235 (1994)
5. Hitchin, N.: Generalized Calabi–Yau manifolds. Quart. J. Math. **54**, 281. [arXiv:math.DG/0209099](https://arxiv.org/abs/math/0209099)
6. Douglas, M.R.: The statistics of string/M theory vacua. JHEP **0305**, 046 (2003). [arXiv:hep-th/0303194](https://arxiv.org/abs/hep-th/0303194)
7. Candelas, P., Lynker, M., Schimmrigk, R.: Calabi–Yau manifolds in weighted  $P(4)$ . Nucl. Phys. B **341**, 383 (1990)
8. Batyrev, V.: Dual Polyhedra and Mirror Symmetry for Calabi–Yau Hypersurfaces in Toric Varieties. [arXiv:alg-geom/9310003](https://arxiv.org/abs/alg-geom/9310003)
9. Batyrev, Victor V., Borisov, Lev A.: On Calabi–Yau complete intersections in toric varieties. In: Andreatta, M., Peternell, T. (eds.) Higher Dimensional Complex Varieties, Proceedings of the International Conference, pp. 39–65. Waller de Gruyter, Trento, Italy, Berlin (1996). [arXiv:alg-geom/9412017](https://arxiv.org/abs/alg-geom/9412017)
10. Kreuzer, M., Skarke, H.: On the classification of reflexive polyhedra. Commun. Math. Phys. **185**, 495 (1997). [arXiv:hep-th/9512204](https://arxiv.org/abs/hep-th/9512204)
11. Avram, A.C., Kreuzer, M., Mandelberg, M., Skarke, H.: The web of Calabi–Yau hypersurfaces in toric varieties. Nucl. Phys. B **505**, 625 (1997). [arXiv:hep-th/9703003](https://arxiv.org/abs/hep-th/9703003)
12. Kreuzer, M., Skarke, H.: Classification of reflexive polyhedra in three-dimensions. Adv. Theor. Math. Phys. **2**, 847 (1998). [arXiv:hep-th/9805190](https://arxiv.org/abs/hep-th/9805190)
13. Kreuzer, M., Skarke, H.: Reflexive polyhedra, weights and toric Calabi–Yau fibrations. Rev. Math. Phys. **14**, 343 (2002). [arXiv:math/0001106](https://arxiv.org/abs/math/0001106) [math-ag]
14. Kreuzer, M., Skarke, H.: Complete classification of reflexive polyhedra in four-dimensions. Adv. Theor. Math. Phys. **4**, 1209 (2002). [arXiv:hep-th/0002240](https://arxiv.org/abs/hep-th/0002240)
15. Kreuzer, Maximilian, Skarke, Harald: Calabi–Yau 4-folds and toric fibrations. J. Geom. Phys. **26**, 272–290 (1998). [arXiv:hep-th/9701175v1](https://arxiv.org/abs/hep-th/9701175v1)
16. Gray, J., Haupt, A., Lukas, A.: Calabi–Yau fourfolds in products of projective space. Proc. Symp. Pure Math. **88**, 281 (2014)
17. Gray, J., Haupt, A., Lukas, A.: All complete intersection Calabi–Yau four-folds. JHEP **1307**, 070 (2013). [arXiv:1303.1832](https://arxiv.org/abs/1303.1832) [hep-th]
18. Anderson, L.B., Apruzzi, F., Gao, X., Gray, J., Lee, S.J.: A new construction of Calabi–Yau manifolds: generalized CICYs. Nucl. Phys. B **906**, 441–496 (2016). [arXiv:1507.03235](https://arxiv.org/abs/1507.03235) [hep-th]
19. Altman, R., Gray, J., He, Y.H., Jejjala, V., Nelson, B.D.: A Calabi–Yau database: threefolds constructed from the Kreuzer–Skarke list. JHEP **1502**, 158 (2015). [arXiv:1411.1418](https://arxiv.org/abs/1411.1418) [hep-th]
20. Davies, R.: The expanding zoo of Calabi–Yau threefolds. Adv. High Energy Phys. **2011**, 901898 (2011). [arXiv:1103.3156](https://arxiv.org/abs/1103.3156) [hep-th]
21. Candelas, P., Davies, R.: New Calabi–Yau manifolds with small Hodge numbers. Fortsch. Phys. **58**, 383 (2010). [arXiv:0809.4681](https://arxiv.org/abs/0809.4681) [hep-th]
22. He, Y.H.: Calabi–Yau geometries: algorithms, databases, and physics. Int. J. Mod. Phys. A **28**, 1330032 (2013). [arXiv:1308.0186](https://arxiv.org/abs/1308.0186) [hep-th]
23. Anderson, L.B., He, Y.H., Lukas, A.: Heterotic compactification, an algorithmic approach. JHEP **0707**, 049 (2007). doi:10.1088/1126-6708/2007/07/049. [arXiv:hep-th/0702210](https://arxiv.org/abs/hep-th/0702210) [hep-th]
24. Gabella, M., He, Y.H., Lukas, A.: An abundance of heterotic vacua. JHEP **0812**, 027 (2008). doi:10.1088/1126-6708/2008/12/027. [arXiv:0808.2142](https://arxiv.org/abs/0808.2142) [hep-th]
25. Gao, P., He, Y.H., Yau, S.T.: Extremal Bundles on CalabiYau Threefolds. Commun. Math. Phys. **336**(3), 1167 (2015). doi:10.1007/s00220-014-2271-y. [arXiv:1403.1268](https://arxiv.org/abs/1403.1268) [hep-th]
26. Anderson, L.B., Gray, J., Lukas, A., Palti, E.: Heterotic line bundle standard models. JHEP **1206**, 113 (2012). doi:10.1007/JHEP06(2012)113. [arXiv:1202.1757](https://arxiv.org/abs/1202.1757) [hep-th]
27. Braun, V., He, Y.H., Ovrut, B.A., Pantev, T.: The exact MSSM spectrum from string theory. JHEP **0605**, 043 (2006). doi:10.1088/1126-6708/2006/05/043. [arXiv:hep-th/0512177](https://arxiv.org/abs/hep-th/0512177)
28. Taylor, W.: On the Hodge structure of elliptically fibered Calabi–Yau threefolds. JHEP **1208**, 032 (2012). [arXiv:1205.0952](https://arxiv.org/abs/1205.0952) [hep-th]
29. Taylor, W., Wang, Y.N.: A Monte Carlo exploration of threefold base geometries for 4d F-theory vacua. JHEP **01**, 137 (2016). [arXiv:1510.04978](https://arxiv.org/abs/1510.04978) [hep-th]
30. Gao, X., Shukla, P.: On classifying the divisor involutions in Calabi–Yau threefolds. JHEP **11**, 170 (2013). [arXiv:1307.1139](https://arxiv.org/abs/1307.1139) [hep-th]
31. Blumenhagen, R., Jurke, B., Rahn, T.: Computational tools for cohomology of toric varieties. Adv. High Energy Phys. **2011**, 152749 (2011). [arXiv:1104.1187](https://arxiv.org/abs/1104.1187) [hep-th]

32. Gray, J., He, Y.-H., Jejjala, V., Jurke, B., Nelson, B.D., Simon, J.: Calabi–Yau manifolds with large volume vacua. *Phys. Rev. D* **86**, 101901 (2012). [arXiv:1207.5801](#) [hep-th]
33. Candelas, P., Constantin, A., Skarke, H.: An abundance of K3 fibrations from polyhedra with interchangeable parts. *Commun. Math. Phys.* **324**(3), 937–959 (2013). [arXiv:1207.4792](#) [hep-th]
34. Braun, V.: On free quotients of complete intersection Calabi–Yau manifolds. *JHEP* **1104**, 005 (2011). [arXiv:1003.3235](#) [hep-th]
35. Candelas, P., de la Ossa, X., He, Y.H., Szendroi, B.: Triadophilia: a special corner in the landscape. *Adv. Theor. Math. Phys.* **12**, 429 (2008). [arXiv:0706.3134](#) [hep-th]
36. Kreuzer, M., Skarke, H.: PALP: a package for analyzing lattice polytopes with applications to toric geometry. *Comput. Phys. Commun.* **157**, 87 (2004). [arXiv:math/0204356](#) [math-sc]
37. Braun, A.P., Knapp, J., Scheidegger, E., Skarke, H., Walliser, N.O.: PALP—a User Manual. [arXiv:1205.4147](#) [math.AG]
38. The On-Line Encyclopedia of Integer Sequences. <http://oeis.org>, Number A090045
39. He, Y.H., Lee, S.J., Lukas, A.: Heterotic models from vector bundles on toric Calabi–Yau manifolds. *JHEP* **1005**, 071 (2010). [arXiv:0911.0865](#) [hep-th]
40. Lynker, M., Schimmrigk, R., Wisskirchen, A.: Landau–Ginzburg vacua of string, M theory and F theory at  $c = 12$ . *Nucl. Phys. B* **550**, 123 (1999). [arXiv:hep-th/9812195](#)
41. Stamatis, D.H.: *Six Sigma and Beyond: Statistics and Probability*, vol. 3, 1st edn. CRC Press (2002)
42. Braun, V.: Toric elliptic fibrations and F-theory compactifications. *JHEP* **1301**, 016 (2013). doi:[10.1007/JHEP01\(2013\)016](#). [arXiv:1110.4883](#) [hep-th]
43. Johnson, S.B., Taylor, W.: Calabi–Yau threefolds with large  $h^{2,1}$ . *JHEP* **1410**, 23 (2014). doi:[10.1007/JHEP10\(2014\)023](#). [arXiv:1406.0514](#) [hep-th]
44. Taylor, W., Wang, Y.N.: Non-toric Bases for Elliptic Calabi–Yau Threefolds and 6D F-Theory Vacua. [arXiv:1504.07689](#) [hep-th]
45. Anderson, L.B., Gao, X., Gray, J., Lee, S.J.: Multiple fibrations in Calabi–Yau geometry and string dualities. *JHEP* **1610**, 105 (2016). doi:[10.1007/JHEP10\(2016\)105](#). [arXiv:1608.07555](#) [hep-th]
46. Candelas, P., Constantin, A., Mishra, C.: Calabi–Yau Threefolds With Small Hodge Numbers. [arXiv:1602.06303](#) [hep-th]
47. Bianchi, M., Ferrara, S.: Enriques and octonionic magic supergravity models. *JHEP* **0802**, 054 (2008). doi:[10.1088/1126-6708/2008/02/054](#). [arXiv:0712.2976](#) [hep-th]

Communicated by Y. Yin