ELSEVIER

Teaser Identifying the right synergy metric for large-scale oncology combination screens is far from trivial: what to look out for, based on real-world data analyses.

Applying synergy metrics to combination screening data: agreements, disagreements and pitfalls

Anna H.C. Vlot[1,2], Natália Aniceto[1], Michael P. Menden[3,4,5], Gudrun Ulrich-Merzenich[6] and Andreas Bender[1]

[1] Department of Chemistry, Centre for Molecular Science Informatics, University of Cambridge, Cambridge, CB2 1EW, UK
[2] Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), 10115, Berlin, Germany
[3] Institute of Computational Biology, Helmholtz Zentrum München — German Research Centre for Environmental Health, 85764, Munich, Germany
[4] Department of Biology, Ludwig-Maximilians University Munich, Martinsried, 82152, Germany
[5] German Centre for Diabetes Research (DZD e.V.), Neuherberg, 85764, Germany
[6] University Hospital Bonn (UKB), Medical Clinic III, AG Synergy Research, Bonn, 53127, Germany

Synergistic drug combinations are commonly sought to overcome monotherapy resistance in cancer treatment. To identify such combinations, high-throughput cancer cell line combination screens are performed; and synergy is quantified using competing models based on fundamentally different assumptions. Here, we compare the behaviour of four synergy models, namely Loewe additivity, Bliss independence, highest single agent and zero interaction potency, using the Merck oncology combination screen. We evaluate agreements and disagreements between the models and investigate putative artefacts of each model's assumptions. Despite at least moderate concordance between scores (Pearson's r >0.32, Spearman's ρ > 0.34), multiple instances of strong disagreement were observed. Those disagreements are driven by, among others, large differences in tested concentrations, maximum response values and median effective concentrations.

Introduction

Several high-throughput combination screens against cancer cell lines have been published recently [1–3]; however, their analysis, in particular with respect to synergistic and antagonistic effects, is not obvious. A synergistic drug combination (see Glossary) elicits an effect that is higher than the expected additive monotherapy potency, whereas an antagonistic drug combination elicits an effect that is lower than the expected additive monotherapy potency. If the elicited effect of a drug combination is equal to the expected additive monotherapy potency, the combination is referred to as 'being additive'.

Corresponding author: Bender, A. (ab454@cam.ac.uk)

Anna Vlot is currently a PhD student in the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS), where she is working on the development of methods for pattern identification in single-cell sequencing data. She was part of the Bender group as a visiting student in the first half of 2018, where she worked on the evaluation of synergy metrics and predictive combination modelling. Previously, her research focused on mechanistic modelling of drug distribution and drug¨target affinity prediction. She received her Masters in bio-pharmaceutical sciences, specialising in systems pharmacology, from Leiden University in 2018.

Dr NatÃlia Aniceto is a Postdoctoral Research Associate in the Bender group, within the Centre for Molecular Informatics at the University of Cambridge. Her recent research focuses on finding new ways of leveraging transcriptomics data to understand and model synergy in cancer. Other research topics in her recent work include improving high-throughput screening with the use of prediction confidence. Natalia received her PhD in cheminformatics from the University of Kent in 2018.

Dr Andreas Bender is a Reader for Molecular Informatics with the Centre for Molecular Informatics at the Department of Chemistry of the University of Cambridge, leading a group of 20 postdocs, PhD and graduate students and academic visitors. In his work, Andreas is involved with the integration and analysis of chemical and biological data, aimed at understanding phenotypic compound action (such as cellular readouts and also organism-level effects) on a mechanistic level, ranging from compound efficacy to toxicity. He received his PhD from the University of Cambridge as a Cambridge Gates Scholar in 2005 and worked in the Lead Discovery Informatics group at Novartis in Cambridge, MA, as well as at Leiden University in The Netherlands, before his current post.

https://doi.org/10.1016/j.drudis.2019.09.002

GLOSSARY

Antagonism When two compounds together elicit an effect that is lower than would be expected based on the effect elicited by each drug alone (and according to the reference model chosen) based on the effect elicited by each drug alone
Conditional probability The probability of observing one event, given that another event has occurred
Confidence interval The interval in which the true value of a parameter is likely to be found with a certain confidence
Half-maximal effect concentrationL The concentration of a drug that induces the half-maximal response
Reference model A model used to define the null hypothesis
Standard deviation Measure of the variation or dispersion of a set of data values
Synergy When two compounds together elicit an effect that is higher than would be expected based on the effect elicited by each drug alone (and according to the reference model chosen)

Impact of experimental design on synergy estimations

The experimental design of high-throughput experiments can have a strong impact on the observed responses. For example, position biases and edge effects can occur with the use of 96-, 384- or 1536-well microtiter plates, into which cancer cell lines are commonly seeded [4]. To quantify cell viability in such plate-based approaches, three experimental conditions are explored in different wells, namely treatment, control and blank (Fig. 1a). At day zero, cell lines are seeded into control and treatment wells, whereas blank wells remain empty. Most commonly, a waiting time of 48 h is employed to ensure linear growth of cell lines after a lag phase [5]. The linear growth of a cell line can also be influenced by seeding density, the optimality of which is dependent on the microtiter plate size and the cell type. After n days, the cells will be fixated, stained with either Syto^{TM}60 (DNA staining) or CellTiter-Glo® (ATP staining) and dead cells and dye are washed off. Finally, the staining intensity of each condition is measured and quantified. To correct for the intensity measured as a result of residual luminescence from dye that was not fully washed from the wells, the intensity of the blank well is subtracted from the intensity of the treatment and the control wells. Subsequently, the cell viability is defined as the ratio of the corrected intensity of the treatment well to the corrected intensity of the control well (Fig. 1b).

The simplest experimental setup to estimate synergy is to test three treatment conditions at single concentrations: (i) monotherapy of drug A; (ii) monotherapy of drug B; and (iii) the combination of drug A and B (Fig. 1c). However, for this, the minimal inhibitory concentration (MIC) of both monotherapies must be known to avoid over- or under-treatment, which would result in either unspecific cytotoxicity or no response, respectively. In real-world applications, the MIC and the concentration required to reduce cell viability by half (IC_{50}) might be known for the standard of care but will most certainly be unknown for experimental drugs. In clinical trials, patients are generally treated with an established drug in combination with a novel compound. This situation can be mimicked in an anchored screening approach (Fig. 1d). In this approach, the established drug is the anchored drug that is fixed at a single concentration, whereas the experimental drug is screened at n titrations. The experimental drug is also screened as monotherapy, so that synergy can be claimed based on the difference of IC_{50} values in all three conditions. The most comprehensive and widely used but also the most laborious and expensive approach *in vitro* is to screen complete drug combination matrices (Fig. 1e) [1–3]. Here, the first row and first column of such a matrix contain the monotherapies used to calculate the theoretical reference surface. This theoretical surface is then subtracted from the experimentally observed surface, and the volume difference between the experimental and theoretical surface corresponds to the synergy score. Depending on whether this volume difference is positive or negative, the score would be classified as either synergistic or antagonistic, respectively. Variability in estimated synergy scores might occur as a result of smoothing of the experimental surface. To summarise, there are many experimental factors that can influence observed synergy scores, even beyond the synergy metrics we will discuss in the main part of this article.

Reference models for compound combination response

Defining a theoretical reference model of the expected combination response is far from trivial and, therefore, several reference models of the expected combination effect have been developed and proposed, most of which originate from toxicity research. Of those metrics, the most well-known ones will be discussed: Loewe additivity [6–8], Bliss independence [9] and highest single agent (HSA) [10]. Additionally, we considered the most recently proposed approach, zero interaction potency (ZIP), which is a hybrid model based on Loewe additivity and Bliss independence [11]. The formulation of the Loewe additivity score in this research was obtained through a mathematical formulation as presented by Chou and colleagues [11–13]. In practice, these reference models are used to obtain a synergy score that characterises the deviation of the 'observed' response from the 'expected' response. Because each of these different synergy scores is defined differently, each possesses situation-dependent characteristics, advantages and shortcomings.

The theoretical limitations of the methods mentioned above have been discussed by several authors [14–16]. For example, it has been reported that the accuracy of the Loewe additivity method across the whole tested concentration range can only be guaranteed when applied to compound combinations that show a similar pharmacodynamic profile in terms of having the same maximum response and a constant potency ratio [3,10,11]. The influence of these assumptions was recently comprehensively formulated in the Loewe Additivity Consistency Condition published by Lederer *et al.* [17]. As for Bliss independence, it has been argued that achieving a certain effect through independent processes (Box 1) is biologically unlikely because crosstalk between biological pathways is abundant with respect to the same biological endpoint [14]. However, independence in the model refers to statistical independence and does not require pharmacological independence. Regarding HSA, the most commonly voiced concern is that, according to this model, combinations are commonly classified as synergistic, even when a compound is tested in combination with itself, which could also be a challenge for Bliss independence [15]. Additionally, the scores obtained from HSA tend to be overly optimistic because it has the lowest threshold for assigning synergy. As for ZIP, despite taking

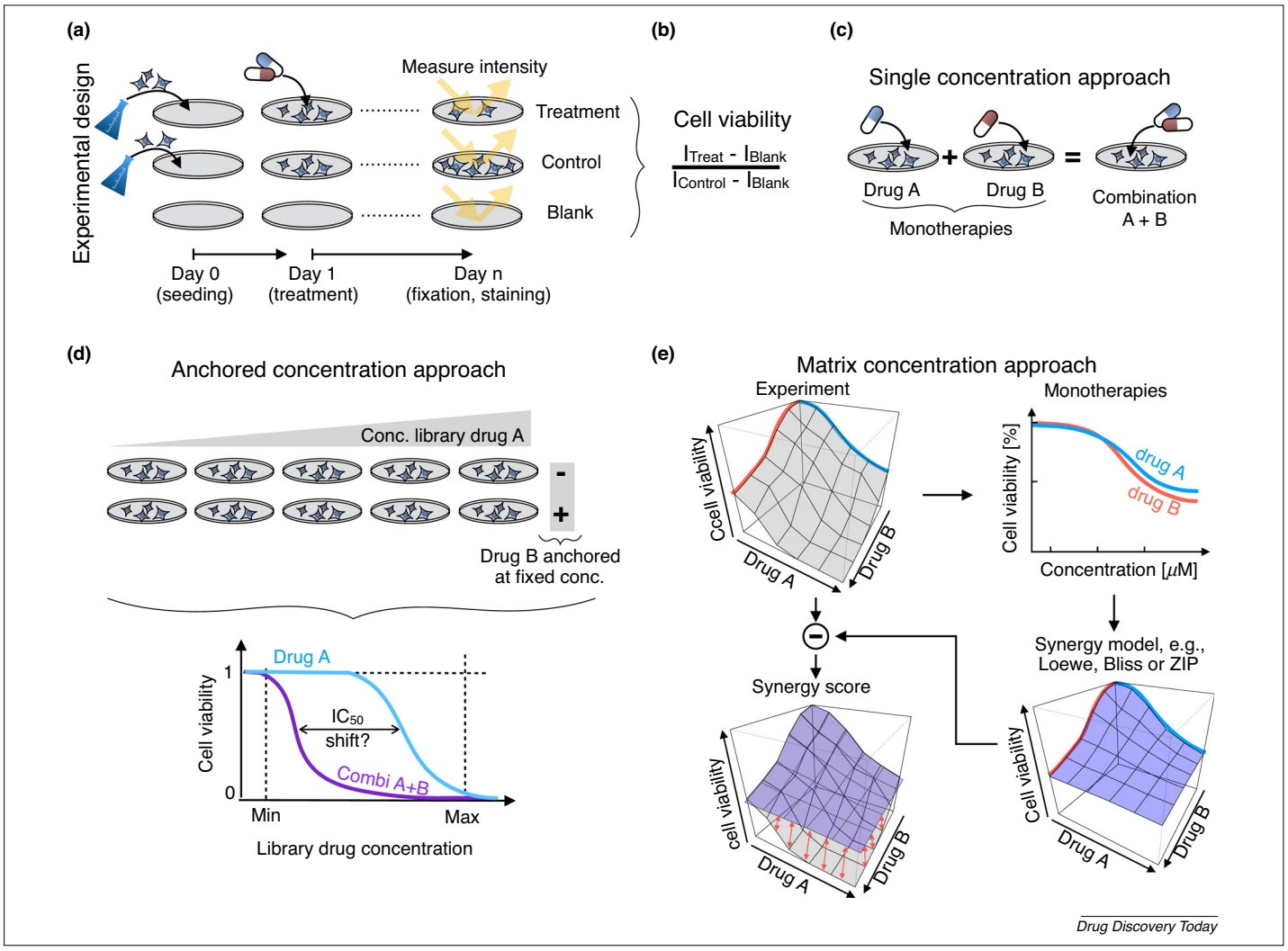

FIGURE 1

Experimental screening design and synergy modelling. **(a)** Shows the simplest experimental plate design to estimate **(b)** the cell viability and **(c)** drug synergy; complemented with **(d)** the anchored concentration approach, where the anchored drug is given at fixed concentration and any library drug would be explored as a mono- and combinatorial-therapy. **(e)** A matrix approach exploring drug combination concentrations in a grid. First column and first row contain both monotherapies, which enable calculation of the theoretical additivity model. Synergy is estimated by calculating the volume between the experimental and theoretical additivity model.

much inspiration from the Loewe additivity and Bliss independence model, this hybrid model has been reported not to be affected by assumptions considering the pharmacodynamic profile of compounds in a combination [11]. However, a downside to the ZIP model is that it depends on the accurate fitting of the dose–response curve to a log–logistic curve to obtain an accurate estimate of the relative half-maximal effect concentration (EC_{50}) and the slope parameter. This could be challenging, especially when the dose–response data quality is low or the dose–response curve results from a complex process and does fit the theoretical dose–response models. Other emerging reference models, like SANE [18], BRAID [19], Schindler's Hill partial differential equation [20], the effective dose model [21] and MuSyC [22], try to address the described limitations but will not be considered in detail in this review.

Previous studies comparing the behaviour of synergy metrics

Additionally, some comparisons between different combination effect metrics have been reported. The most commonly reported comparisons are made between the Loewe additivity and the Bliss independence model. For example, an extensive mathematical comparison of Loewe additivity and Bliss independence has been described by Goldoni and Johansson, albeit in regard to toxicity modelling [23]. Furthermore, Drescher and Boedeker [24] reported the differences between the concentration addition (CA) method (i.e., Loewe additivity) and the independence action (IA) method (i.e., Bliss independence) in terms of an analytics comparison for concentrations tending to 0 and concentrations tending to infinity. This work extended upon previous work by Chen and Christensen [25] who described that the expected effect values for CA are larger than values for IA when considering steeper dose–response curves, whereas values for CA were smaller than values for IA for flatter dose–response curves. Additionally, Beader *et al*. stated that the expected response for Bliss independence is generally higher than that for Loewe additivity for nonlinear dose–response curves [26]. For linear dose–response curves, both scores should give the same expected response. Currently, it is

BOX 1

Metrics*Loewe additivity*

The Loewe additivity metric [6–8] is based on the premise that no drug can interact with itself and that both compounds in a combination are interchangeable. Mathematically, this metric relies on the dose equivalence principle and the sham experiment principle [15]. The dose equivalence principle states that, for a given effect of drug A, an equivalent dose for drug B exists. As a result, it is assumed that the response curves have the same minimum and maximum response because, if this is not the case, an equivalent dose between A or drug B does not exist [34]. According to the sham experiment principle, this equivalent dose of drug B produces the expected additive effect of the drug combination of drug A and drug B. From this it follows that Loewe additivity is observed when the conditions in Eq. 1 are met. In this equation, a is the dose of drug A, b is the dose of drug B and A and B are the equivalent doses that obtain the combined response for drug A and B, respectively.

$$\frac{a}{A}+\frac{b}{B}=1 \tag{1}$$

The derivatisation to obtain a mathematical equation for the expected effect was reported by Yadav *et al.* [11] and based on the mathematical formulation of Loewe additivity as first presented by Chou *et al.* [12,13] (see supplementary material online).

Bliss independence

The Bliss independence model [9] was developed on the premise that the effect of two compounds is statistically independent. Biologically, this can translate to two drugs achieving the same effect (e.g., inhibition of cancer cell growth) independently (i.e., via two independent pathways). That is to say, the model assumes that the relative effect of a drug is not dependent on the presence of another drug. In reality, however, the effect is merely required to be statistically independent. The expected effect E_{AB} is given by Eq. 2.

$$E_{AB} = E_A + E_B - E_A E_B \tag{2}$$

Highest single agent

The highest single agent (HSA) model, also known as Gaddum's pharmacological independence, deems a combination to be synergistic when an excess over the highest single-agent effect is observed [10]. This approach is simple and intuitive because it relies on the observation of an improvement in efficacy upon administration of a second drug. Formally, the expected effect E_{AB} is given by Eq. 3.

$$E_{AB} = \max\{E_A, E_B\} \tag{3}$$

Zero interaction potency

Zero interaction potency (ZIP) is the most recently introduced metric. This model is based on the notion that the expected combined effect is observed when the dose–response curve remains unaltered upon addition of the second drug [11]. As such, this metric assesses changes in potency and the shape of the dose–response curve for a compound combination compared with those of its individual components. For this metric, the expected effect E_{AB} is given by Eq. 4, where [A] and [B] are the concentrations of drug A and drug B, respectively, $EC_{50,A}$ and $EC_{50,B}$ are the concentrations at which the half-maximal response of drug A and drug B, respectively, are observed, and λ_A and λ_B are the slope parameters for the dose–response curves of drug A and drug B.

$$E_{AB} = \frac{\left(\frac{[A]}{EC_{50,A}}\right)^{\lambda_A}}{1+\left(\frac{[A]}{EC_{50,A}}\right)^{\lambda_A}} + \frac{\left(\frac{[B]}{EC_{50,B}}\right)^{\lambda_B}}{1+\left(\frac{[B]}{EC_{50,B}}\right)^{\lambda_B}} - \frac{\left(\frac{[A]}{EC_{50,A}}\right)^{\lambda_A}}{1+\left(\frac{[A]}{EC_{50,A}}\right)^{\lambda_A}} \frac{\left(\frac{[B]}{EC_{50,B}}\right)^{\lambda_B}}{1+\left(\frac{[B]}{EC_{50,B}}\right)^{\lambda_B}} \tag{4}$$

commonly accepted that the assessment of combination effects will give different results between Loewe additivity and Bliss independence methods.

Despite efforts to better understand the behaviour of different metrics for the assessment of synergy, the selection of a metric for the assessment of large combination screens appears largely arbitrary. In the Merck combination screen [2], for example, synergy and antagonism were defined according to Bliss independence and HSA. In a subsequent study, however, preference was given to the use of synergy scores according to the Loewe additivity method in the development of a predictive model using this data [27]. The Loewe additivity score was also employed in the DREAM challenge on synergy prediction on a recent AstraZeneca combination screen [3]. In the analysis of the NCI ALMANAC screen [1], a modified Bliss score: the ComboScore, was used. This score was retained as the score to be predicted in recently developed synergy prediction models using the NCI ALMANAC screen [28]. In the ComboScore, a growth ceiling of 100% was applied to remove the influence of so-called reversals (i.e., hormesis), where enhancement of growth is observed upon treatment with a single drug. Overall, all three traditional synergy scores are commonly used in the assessment of combination screens, but the arguments underlying the decisions are not always clearly communicated.

A comparison of the ZIP score and the other synergy metrics described above was performed in the original ZIP publication [11]. Here, the comparability of ZIP to Loewe additivity, Bliss independence and HSA-based methods was explored for a small dataset ($n = 459$) by Mathews Griner *et al.* [29]. It was found that

scoring according to ZIP was more similar to scoring according to Bliss independence (rank correlation of 0.77) and HSA (rank correlation of 0.85) based models than Loewe additivity models (rank correlation of 0.5). Additionally, in a recent publication these four synergy metrics plus the ComboScore, used in the original analyses of the NCI ALMANAC screen, were compared when applied to 3647 drug pairs from the NCI ALMANAC screen [30]. In this study, the highest agreement was observed between HSA and Loewe additivity (Pearson's r of 0.66, Spearman's ρ of 0.69), and the lowest agreement was observed between Bliss independence and ComboScore (Pearson's r of 0.21) and the majority of assigned synergistic and antagonistic labels were found to be unique to a certain metric. This illustrates that the choice of a synergy metric influences the conclusions drawn from a combination screen substantially.

Comparative approach taken in the current work: four common synergy metrics, analysing large-scale combination screening data

In this review, we present a systematic comparison of the four previously mentioned synergy metrics: Loewe additivity, Bliss independence, HSA and ZIP, when applied to a large-scale combination screen published by Merck & Co. [2]. The methods used can be found in Box 2. Herein, we evaluate their commonalities and differences in behaviour between metrics, during which we focus specifically on the influence of the observed single-agent and combination dose–response relationship. These differences in behaviour are of significant practical relevance for tasks like modelling combination screening data in a pharmaceutical setting [31].

Interbatch comparability within and between metrics

To assess the robustness of different synergy metrics to experimental variability, the difference between the dose–response data and their corresponding synergy scores originating from different batches ($n = 315$) was analysed (Box 2). The correlation between the cell viability (%) for compound combination and cell line pairs between different batches is high (r >0.9), with mean absolute differences of $-0.77 \pm 11.18\%$ viability between batch 1 and 3, and $-4.19 \pm 6.93\%$ viability between batch 2 and 3 (Table S4, see supplementary material online). Those differences are small considering that the minimum response ranges from 1.71×10^{-27} to 1.69% viability and the maximum response values range from 100.32 to 160.33% viability across all three batches, covering a range of minimally 98.31% and maximally 160% viability within each batch (Table S5, see supplementary material online).

The agreement between the synergy scores for the data origination from different batches is considerably smaller, with *r* ranging from 0.16 to 0.87 and ρ ranging from 0.52 to 0.85 between any two batches (Table S6, see supplementary material online).

The discrepancy between the calculated synergy scores is especially large for certain instances of S_{LOEWE} between batch 2 and 3, as can be seen from the mean absolute differences ($\overline{\Delta}$) and standard deviations (*SD,* Table S6, see supplementary material online). For the other metrics, the *SD* values are lower and close to those observed for the viability values (Table S5, Table S6, see supplementary material online), indicating that those other metrics are more robust to changes in the dose–response relationship, despite the *SD* values being high. The highest robustness, characterised by high values of *r* and ρ, is found for the S_{HSA} ($r_{1,3} = 0.78$, $\rho_{1,3} = 0.63$, $r_{2,3} = 0.95$, $\rho_{2,3} = 0.85$) and S_{BLISS} values ($r_{1,3} = 0.73$, $\rho_{1,3} = 0.57$,

BOX 2

Methods*Data*

The data used in this review are from a large-scale combination oncology screen published by Merck & Co. [2]. The dataset contains dose–response values for 22 737 combination–cell line pairs. In this study, 583 pairwise combinations of 38 different compounds (Table S2, see supplementary material online) were used, where 22 of those 38 were tested exhaustively and 16 were only tested with compounds from the exhaustive set. The compound combinations were tested on 39 cancer cell lines, originating from seven different tissues (Table S3, see supplementary material online). Both the combination–response values for the 22 737 combinations for 4 × 4 concentration combinations, and the single agent dose–response values for all 38 compounds in each cell line, were supplied. The concentrations between the single-agent and combination experiments do not align, but a 5 × 5 response surface per cell was obtained through inter- and extra-polation of the fitted dose–response curves. The response is the normalised fractional cell viability.

Synergy scores (S)

The synergy scores (*S*) according to Loewe additivity, Bliss independence, HSA and ZIP were calculated for the 5 × 5 combination matrix from the single-agent dose–response values using SynergyFinder™ (version 1.6.1) [33]. No additional baseline correction was applied, and E_{min} and E_{max} were not fixed. The values of *S* are defined such that $S = 0$ represents additivity, $S > 0$ indicates synergy and $S < 0$ indicates antagonism. Values of 0 were forcibly assigned to all instances in the 5 × 5 synergy score matrix where the concentration of one of the combinations is zero. The calculation of the synergy score according to each metric was successful for 19 119 combination–cell line pairs. Of those 19 119 combination–cell line pairs, 315 were tested in multiple batches. To encapsulate the whole 5 × 5 synergy surface for each compound combination in one score, we propose the use of a synergy weighted-sum-score (S_{WSS}).

Classification

The *S* and S_{WSS} values according to each metric were classified as synergistic when $S_{(WSS)} < 0$, $0 \notin CI_{99.9\%}$, as antagonistic when $S_{(WSS)} > 0$, $0 \notin CI_{99.9\%}$, and as additive otherwise, as adapted from previous work [37]. The *CI* was determined based on four technical replicates, and only for those compound combinations for which four technical replicates were available ($n = 18\ 225$).

Metric of comparison

Batch-to-batch experimental variation and variation between metrics was assessed based on the mean, median, standard deviations (*SDs*), Pearson correlation coefficients (r) and Spearman rank correlation coefficients (ρ) [38] for real values. For classified data, Cohen's κ [39] and conditional probabilities of encountering antagonistic classifications for one and antagonistic classifications for another metric were analysed. A more detailed overview of the methods and the mathematical and theoretical background of the synergy metrics can be found in the supplementary material online.

$r_{2,3} = 0.93$, $\rho_{2,3} = 0.83$). The robustness of S_{ZIP} ($r_{1,3} = 0.64$, $\rho_{1,3} = 0.52$, $r_{2,3} = 0.72$, $\rho_{2,3} = 0.75$) and S_{LOEWE} is notably lower ($r_{1,3} = 0.68$, $\rho_{1,3} = 0.58$, $r_{2,3} = 0.19$, $\rho_{2,3} = 0.69$). From this, it follows that HSA and Bliss independence are more robust to small changes in the dose–response relationship. We hypothesise that this could be the result of the dependency on accurate log–logistic curve fitting to get the maximum response values and the median effective concentration, needed to determine S_{LOEWE} and S_{ZIP}. Especially in situations where the response values are irregular or when the dose–response curve has not reached the maximum efficacy value, this might prove difficult if not impossible. This is also illustrated by the inability to obtain values for S_{LOEWE} and S_{ZIP} for ~16% of the combinations in the original data (see supplementary material online).

Comparability between synergy metrics

Next, we consider the complete set of synergy scores obtained for the data (Box 2). The first differences between the metrics are observed when considering the data summaries of the synergy scores according to each metric. As expected from their inherently different nature, the value distributions of S (synergy) and S_{WSS} (synergy weighted-sum-score) differ per metric (Table S7, Table S8, see supplementary material online). Considering that a value of zero is the theoretical threshold between antagonism and synergy, Loewe additivity ($mean_{LOEWE} = -2.1$, $SD_{LOEWE} = 15.4$) is the most stringent metric given its scores tend towards negative values, followed, in order, by Bliss ($mean_{BLISS} = 0$, $SD_{BLISS} = 9.4$), ZIP ($mean_{ZIP} = 1.5$, $SD_{ZIP} = 9.7$) and HSA ($mean_{HSA} = 2.4$, $SD_{HSA} = 10.1$, Table S7, see supplementary material online). From the corresponding SD values, we conclude that the dispersion of assigned scores is found to be the highest for Loewe additivity. Interestingly, when considering the maximum and minimum values for the weighted-sum-score, $S_{WSS,ZIP}$ gives the highest values ($S_{WSS,ZIP,max} = 669.4$, $S_{WSS,ZIP,min} = -336.2$, Table S8, see supplementary material online) and exceeds $S_{WSS,HSA}$ ($S_{WSS,HSA,max} = 521.2$, $S_{WSS,HSA,min} = -867.7$, Table S8, see supplementary material online). Negative values, corresponding to antagonism, are found most commonly for $S_{WSS,LOEWE}$ at 89% of their full range of values, compared with 62%, 59% and 33% for $S_{WSS,BLISS}$, $S_{WSS,HSA}$ and $S_{WSS,ZIP}$, respectively. In line with these findings, the largest number of combinations classified (Box 2) as antagonistic are indeed found for S_{LOEWE} (Table 1). Notably, the second-largest fraction of combinations classified as antagonistic is found for S_{ZIP}, despite the small range of negative values overall. However, as expected considering the S_{ZIP} values, the highest proportion of combinations classified as synergistic are also found for S_{ZIP}. When considering the S_{WSS} values, those observations shift. Here, the second-most-prevalent antagonistic classifications are found for $S_{WSS,ZIP}$ and the most prevalent synergistic classifications are found for $S_{WSS,HSA}$. Overall, the values for HSA and ZIP tend towards higher values than Loewe additivity and Bliss independence. These results provide a first benchmark of the differences in synergy scores that are assigned to the same data using different metrics.

Next, we describe the comparability in terms of correlations and classification agreement between metrics (Box 2). The moderate-to-low values for r, ρ and Cohen's κ suggest that the metrics behave differently with respect to (i) their raw values, (ii) their ranked values and (iii) their resulting antagonistic, additive and synergistic class membership (Fig. 2; Table S9, see supplementary material online). The highest agreement is observed between S_{BLISS} and S_{HSA} ($r = 0.86$, $\rho = 0.76$, $\kappa = 0.68$), whereas the lowest agreement is observed between S_{LOEWE} and S_{ZIP} ($r = 0.32$, $\rho = 0.34$, $\kappa = 0.22$). The high correlations found between S_{BLISS} and S_{HSA} are in contrast with low correlation reported by Gilvary *et al.* (r = 0.34) [30]. Those contrasting observations could be the result of the differences between the compound combinations included in the differences datasets, or result from differences in data preparation related to quality control of combinations to include in the analysis. Notably, the ranked correlation and class agreement of S_{ZIP} versus all other metrics is low ($\rho = 0.34$–0.53, $\kappa = 0.22$–0.28). The low agreement found between S_{LOEWE} and S_{ZIP} is in line with results from the comparison of the ZIP metric to other metrics in its initial publication [11] as well as low correlations between these two metrics reported by Gilvary *et al.* [30]. Similar trends were observed for the S_{wss} values (Table S10, see supplementary material online). Overall, our results are in line with the low r and ρ values and frequently conflicting synergistic and antagonistic assignments recently reported in a preprint by Gilvary *et al.* [30]. However, results differ with respect to quantitative high and low agreement between pairs of individual metrics. Of note, high Pearson's r values are not essential to demonstrate comparability, given the inherently different nature of the metrics and their differing distributions. Spearman's ρ and Cohen's κ are therefore more relevant measures because a low Spearman's ρ shows that the order in which the combinations are scored differs between two synergy metrics, and a low Cohen's κ shows that the classification of a compound combination differs between metrics.

Antagonistic–synergistic classification (dis)agreement between metrics

To further define the extent of disagreement in membership for compound–cell line combinations, we explored how commonly compound–cell line combinations are classified as synergistic by one and antagonistic by another metric. This is equivalent to

TABLE 1

Percentage of compound combination-cell line pairs classified into a certain class per metric.

Metric	Full dose–response surface			S_{WSS} (synergy weighted sum score)		
	Antagonistic	Additive	Synergistic	Antagonistic	Additive	Synergistic
Loewe	6	91.4	2.6	13.7	80.4	5.9
Bliss	3.2	94.5	2.2	5.5	89.8	4.7
HSA	2.2	93	4.8	1.9	84.5	13.6
ZIP	4.8	88.3	6.9	4.6	87.7	7.7

The antagonistic classification is assigned to $S < 0$, where $0 \notin CI$; the synergistic classification is assigned to $S > 0$, where $0 \notin CI$, and an additive classification is assigned in all other instances. *CI* is the 99.9% confidence interval.

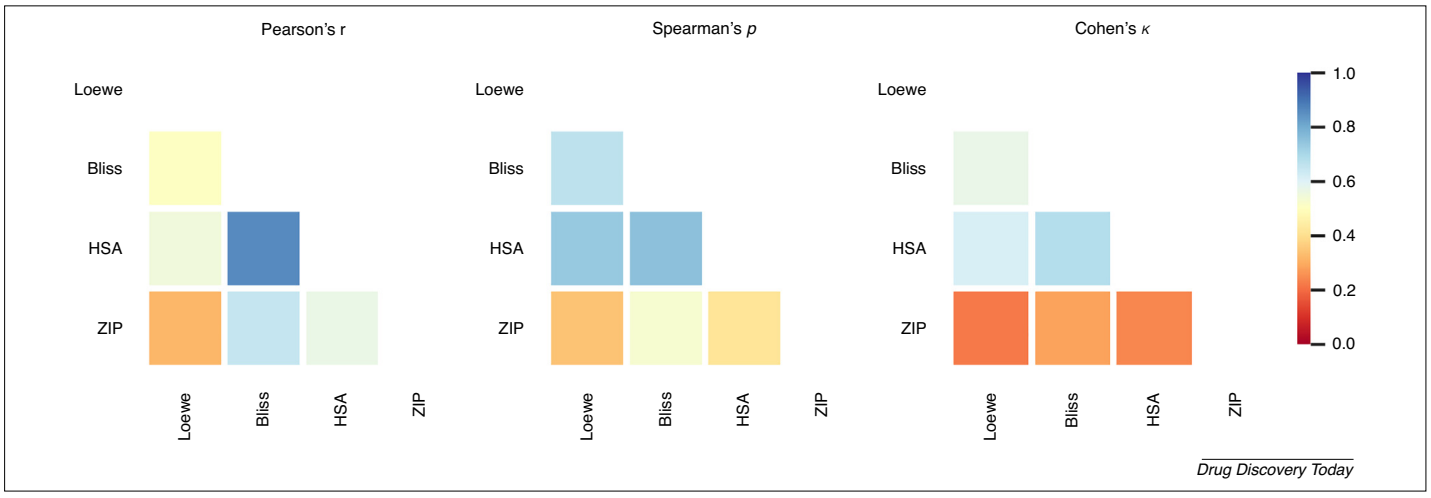

FIGURE 2

Pearson's correlation *r* values, Spearman's ranked correlation ρ values, and Cohen's κ values between *S* calculated according to different reference models, namely Loewe additivity, Bliss independence, HSA and ZIP. The moderately low values of Pearson's *r*, Spearman's ρ and Cohen's κ suggest that the metrics behave differently with respect to (i) their raw values, (ii) their ranked values and (iii) their resulting antagonistic, additive and synergistic class membership. The values underlying these heatmaps are provided in Table S9 (see supplementary material online).

exploring how commonly different metrics confidently characterise the same compounds at opposite ends of the synergy scoring range.

First, we consider the classifications according to the S_{WSS} values. Here, clashes in synergistic and antagonistic classifications are observed for merely 1.7% (317/18 225) of the total number of combination–cell line pairs. Their classifications and corresponding $CI_{99.9\%}$ values are provided (see supplementary material online). Out of those combinations, 22.7% (72/317) have clashing classifications in more than one cell line, whereas a clash between more than two metrics is observed in 37.2% (118/317). Disagreements in antagonistic–synergistic label assignments between multiple metrics in multiple cell lines are observed for 7.6% (24/317) of the total number of clashing instances.

The largest portion of clashes is observed between classifications according to the value of $S_{WSS,LOEWE}$ and $S_{WSS,ZIP}$ (30.0%, Fig. 3; Table S11, see supplementary material online), whereas the smallest

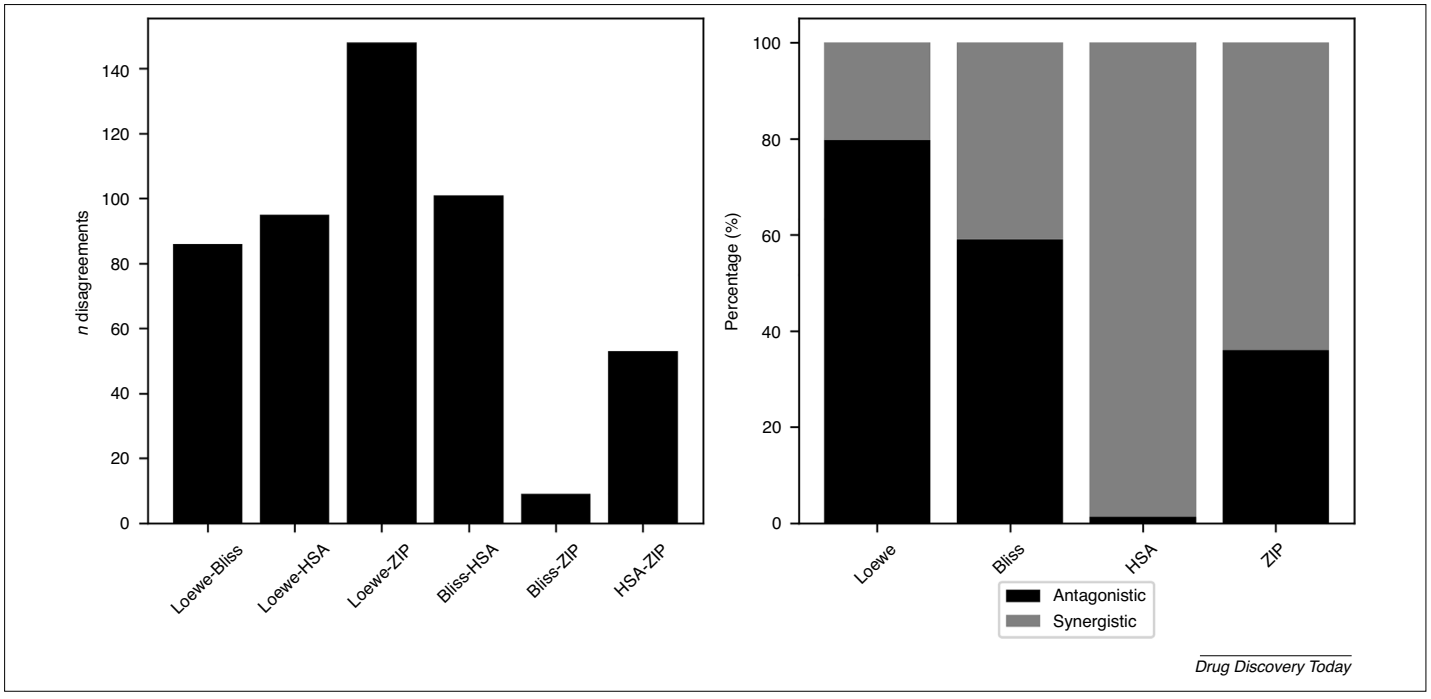

FIGURE 3

Number of disagreements in antagonistic and synergistic classifications per metric pair (left), and the percentage of antagonistic and synergistic classification in those disagreement instances per metric (right). The frequency of observing clashes is dependent on the metrics considered (left). Clashes in which Loewe additivity scores are antagonistic are most common, whereas HSA classifications are most commonly synergistic (right). The values underlying these graphs are presented in Table S11 (left) and Table S12 (right) (see supplementary material online).

fraction of clashes is observed between classifications according to $S_{WSS,BLISS}$and $S_{WSS,ZIP}$ (1.9%, Fig. 3; Table S11, see supplementary material online). In line with the largest proportion of antagonistic classifications overall, the classification according to Loewe is most commonly antagonistic in these clashes (Fig. 3; Table S12, see supplementary material online). As also expected from the previously discussed data summary, synergistic classifications are most commonly found for classifications according to $S_{WSS,HSA}$.

Considering the classifications according to the S values, most observations are similar, except the relatively lower proportion of the disagreements between classifications according to S_{BLISS} and S_{HSA} (Table S13, see supplementary material online). Another notable result is that a large proportion of synergistic classifications are observed for instances where S_{ZIP} provides a clashing classification (Table S14, see supplementary material online). This could be a result of the limited range of negative values: $-50 \leq S < 0$, for S_{ZIP} (Table S7, see supplementary material online), increasing the chance of observing *CI*s overlapping 0, resulting in additive classifications through the classification method used in this analysis. Next, we illustrate how probable it is to obtain a contrasting classification according to another metric given a certain classification by one metric. For this, we provide a table of conditional probabilities of observing a certain combination of contrasting classifications according to S and S_{WSS} values (Table 2). Notably, the conditional probabilities found for S values (0–38%) are larger than those observed for S_{WSS} values (0–10%).

The highest conditional probability of observing an antagonistic–synergistic classification disagreement for S values is found for synergistic classifications according to S_{ZIP} given that the score according to another metric is antagonistic (Table 2). The conditional probabilities of finding an antagonistic classification according to S_{ZIP} given that the classification according to the S_{LOEWE}, S_{BLISS} or S_{HSA} is synergistic, are 22, 27.7 and 38%, respectively. Those findings are in line with findings of high classification disagreement according to S_{ZIP} as illustrated by low Cohen's κ values (Fig. 2; Table S9, Table S10, see supplementary material online). Those high probabilities of contrasting synergistic and antagonistic classifications suggest that this is where a high proportion of the disagreements found for classifications according to S_{ZIP} arises from. This suggests that compounds that are classified as synergistic according to S_{ZIP} might not exceed the concentration-specific effect of the most potent drug in the combination. Otherwise, they would be classified as synergistic or additive according to the S_{HSA}. These results illustrate that the choice of metric for the assessment of synergy can influence the interpretation of results from large combination screens. When considering classifications according to the S_{WSS} scores, the conditional probabilities of observing a metric-specific disagreement in antagonistic–synergistic classifications are relatively low (0–10%, Table 2). This suggests that such discrepancies are seldom observed across the full concentration–response surface for a given combination in a given cell line.

The highest conditional probabilities for disagreements in classification are found for instances where classification according to $S_{WSS,HSA}$ is synergistic, given antagonistic classifications according to $S_{WSS,LOEWE}$ (P = 3.8%), $S_{WSS,BLISS}$ (P = 6.2%) or $S_{WSS,ZIP}$ (P = 10.0%, Table 2). This is in line with general expectations of higher values for S_{HSA} than S_{LOEWE} and S_{BLISS} in particular; and with the description of S_{WSS} values previously described in this review

TABLE 2

The conditional probabilities of finding an antagonistic classification for one, and a synergistic classification for another metric for the same compound combination-cell line pair.

A		B		*P(B|A)* (%)*	*P(B|A)* (%)*
				Full dose-response surface	S_{WSS} (Synergy weighted sum score)
Antagonistic	Bliss	Synergistic	Loewe	0.55	2.61
	HSA			0	0
	ZIP			0.57	1.55
Synergistic	Bliss	Antagonistic	Loewe	9.05	6.97
	HSA			6.02	3.84
	ZIP			19.07	9.62
Antagonistic	Loewe	Synergistic	Bliss	3.36	2.41
	HSA			0	0.29
	ZIP			0.12	0.12
Synergistic	Loewe	Antagonistic	Bliss	0.68	2.4
	HSA			1.93	4.04
	ZIP			13.06	0.57
Antagonistic	Loewe	Synergistic	HSA	4.84	3.81
	Bliss			2.86	10.02
	ZIP			2.12	6.19
Synergistic	Loewe	Antagonistic	HSA	0	0
	Bliss			0	0.12
	ZIP			12.06	0.07
Antagonistic	Loewe	Synergistic	ZIP	22.01	5.42
	Bliss			27.71	0.8
	HSA			37.98	0.29
Synergistic	Loewe	Antagonistic	ZIP	1.05	1.2
	Bliss			0.26	0.12
	HSA			2.12	2.1

* *P(A|B)* gives the conditional probability, in percentages, of observing classification A using a certain metric, given a certain classification B using another metric.

(Table S7, see supplementary material online). Similar conditional probabilities are observed for antagonistic classifications based on $S_{WSS,LOEWE}$ given synergistic classifications according to $S_{WSS,BLISS}$ ($P = 3.8\%$), $S_{WSS,HSA}$ ($P = 7.0\%$), and $S_{WSS,ZIP}$ ($P = 9.6\%$). This does not correspond to the notion that the expected response for Bliss independence is generally higher than the expected response for Loewe additivity, as suggested by Baeder *et al*. [26]. When the expected response is higher, the synergy score should generally be lower. Therefore, given the highest expected response for Bliss independence, one would expect those antagonistic classifications to be more common for $S_{WSS,BLISS}$ than for $S_{WSS,LOEWE}$.

For several reasons, these results cannot be directly compared to results on classification agreement by Gilvary *et al*. [30]. First, the classification approach differs between the two studies. Where we used a confidence-interval-based approach, Gilvary *et al*. used a percentile-based approach in which the top 5% of scores were classified as synergistic and the bottom 66.67% were classified as antagonistic. Furthermore, Gilvary *et al*. reported the proportional overlap of synergistic classifications and antagonistic classifications between scores, whereas we report the disagreement of synergistic and antagonistic classifications. Quantitative results that can be extracted from their study are a proportional overlap of 0.40 between Loewe additivity and HSA and 0.35 between HSA and ZIP despite these metrics being relatively highly correlated in their analysis of the NCI ALMANAC screen. Thus, both pieces of work illustrate that the agreement of synergistic and antagonistic classifications between metrics is limited.

Underlying reasons for disagreements between synergy classifications

To identify the conditions under which these antagonistic–synergistic disagreements occur, we investigate the dose–response profiles for several combinations for which such disagreements occur. Here, we use the agreements observed for classifications according to the S_{WSS} values so that the combinations can be considered as a whole. To characterise how divergent the S_{WSS} values are for these combinations, the relative ranking (*rr*) of the S_{WSS} values will be reported. The relative ranking was performed so that all ranks lie within the interval 0 to 1 and were obtained by dividing the rank by the total number of combinations.

Surprising synergistic classifications according to $S_{WSS,ZIP}$

A number of unexpected synergistic compound combinations according to $S_{WSS,ZIP}$ were observed in this study. An example of this is the synergistic classification of MK-5108 and oxaliplatin in NCIH-520, for which a synergistic classification according to $S_{WSS,ZIP}$ and an antagonistic classification according to $S_{WSS,BLISS}$ were observed at a relative ranking difference of 0.91. Nearly all the values for S_{ZIP} indicate synergism, whereas nearly all the S_{BLISS} values indicate antagonism. The values for S_{LOEWE} and S_{HSA} are most commonly low-positive or high-negative, corresponding to additivity (Table S15, see supplementary material online). From the combination–response surface (Table S16, see supplementary material online) no indication of synergism is observed because the combination effect is not much higher than the single agent responses (Figure S2, see supplementary material online). Therefore, it seems that, with a value of 286, an unexpectedly high $S_{WSS,ZIP}$ value is obtained. Similarly, for mitomycin and geldanamycin in the Caov-3 cell line, all classifications except the one according to $S_{WSS,ZIP}$ are antagonistic. Upon inspection of the scores obtained across the full dose–response matrix, it is observed that all three other metrics give negative values of *S* across all concentration pairs, yet the value of S_{ZIP} is commonly close or equal to zero (Table S17, see supplementary material online). Combined with the single-agent dose–response curves (Figure S3, see supplementary material online) and the combination dose–response surface (Table S18, see supplementary material online) it seems again that $S_{WSS,ZIP}$ is unreasonably high. This is especially true at the first concentration point where the observed combination effect does not exceed the effect of mitomycin individually. Generally, however, the S_{ZIP} values for this combination do correspond less to convincing synergism. Yet, together with previously mentioned results of high conditional probabilities of finding a synergistic classification given antagonistic classifications according to other metrics, this suggests that, in certain cases, values of $S_{WSS,ZIP}$ could be misleadingly high.

Loewe additivity commonly produces unexpected antagonistic classifications

A striking observation is that, in every classification disagreement involving Loewe additivity, the classification according to $S_{WSS,LOEWE}$ was always antagonistic. For example, for the combination of lapatinib and bortezomib in MSTO, where the classification according to $S_{WSS,BLISS}$ ($rr_{BLISS} = 0.99$), $S_{WSS,HSA}$ ($rr_{HSA} = 0.99$) and $S_{WSS,ZIP}$ ($rr_{ZIP} = 0.99$) is synergistic, yet the classification according to $S_{WSS,LOEWE}$ ($rr_{LOEWE} = 0.06$) is antagonistic. Evidently, the classification according to the $S_{WSS,LOEWE}$ is the anomaly here. Similarly, for the combination of cyclophosphamide and MK-8669 in the OCUB-M cell line, the classification according to $S_{WSS,LOEWE}$ is the only one to stem from a negative value for $S_{WSS,LOEWE}$ ($S_{WSS,LOEWE} = -444$, $rr_{LOEWE} = 0.03$). Although here the only antagonistic classification was found for $S_{WSS,BLISS}$ ($S_{WSS,BLISS} = 212$, $rr_{BLISS} = 0.99$), the ranking for $S_{WSS,HSA}$ ($rr_{HSA} = 0.93$) and $S_{WSS,ZIP}$ ($rr_{ZIP} = 0.95$) are also very high, confirming that $S_{WSS,LOEWE}$ is the anomaly here. The lack of synergistic classification despite high relative ranking is that the experimentally based 99.9% confidence interval for the combinations includes values ≤ 0. One last example is the combination of dexamethasone and dinaciclib in LNCaP, where, again, the classification according to $S_{WSS,LOEWE}$ ($rr_{LOEWE} = 0.002$) is the only one to be antagonistic, whereas the classification according to the $S_{WSS,ZIP}$ is positive ($rr_{ZIP} = 0.94$) alongside positive $S_{WSS,BLISS}$ and $S_{WSS,HSA}$ values, although not classified as synergistic owing to reasons stated above. In all these examples, the classification according to $S_{WSS,LOEWE}$ is clearly the anomaly, suggesting that the assumptions of the Loewe additivity model might be violated in these instances.

Disagreements can result from a non-constant potency ratio between the drugs in a combination

We furthermore identified cases where disagreements in synergistic classifications can result from a non-constant potency ratio between the two drugs in a combination. The most compelling evidence from these examples comes from the combination of lapatinib and bortezomib in MSTO. From the *S* values, it can be seen that, for the identified concentration ranges of lapatinib, from 1.1 to 5 μM, and bortezomib, from 0.002 to 0.04 μM, the scores for S_{LOEWE} are almost identical to the scores according to

(most of) the other metrics (Table S19, see supplementary material online). From the dose–response curves we can visually infer that this is exactly the concentration range where the slopes of the dose–response curves appear identical resulting in a constant potency ratio (Figure S4, see supplementary material online). Although it has been asserted that a constant potency ratio is not necessary for the assessment of Loewe additivity using the parameterised method implemented in SynergyFinder™ [32,33], accuracy cannot be guaranteed if this condition is not ensured as this example suggests [14,17].

Negative single-agent responses lead to spurious classifications

Another notable observation is that a negative response (i.e., an increase in cell viability) was observed for cyclophosphamide in the OCUB-M cell line (Figure S5, see supplementary material online) and no response was observed for dexamethasone in the LNCaP cell line (Figure S6, see supplementary material online). Similarly, for outliers detected upon plotting the S and S_{WSS} values against one another (Figure S7, Figure S8, see supplementary material online), the two most prominent outliers were observed to be due to $S_{WSS,LOEWE}$ values too low to be theoretically reasonable. These outliers correspond to the combination of SN-38 and ABT-888 ($S_{WSS,LOEWE} = -6000$) and SN-38 and erlotinib ($S_{WSS,LOEWE} = -4000$) in the SK-MEL-30 cell line (Figure S9, see supplementary material online). One initial observation is the large difference in tested concentrations between SN-38 and the compounds it was tested in combination with (Table S20, see supplementary material online). This is not only true for those two extreme outliers but also for other combination for which very low values of $S_{WSS,LOEWE}$ were found. Additionally, in line with previously stated observations, it is observed that for ABT-888 no response is observed for any of the concentrations (Fig. 4b; Table S21, see supplementary material online) and for erlotinib a response is only observed at the highest tested concentration (Fig. 4b; Table S21, see supplementary material online). In these cases, the assessment of synergy is uninformative by definition, because synergy can only be observed when the compounds in the combination individually illicit an effect. From the results described above, it seems that the Loewe additivity model effectively removes these pairs from consideration because a negative $S_{WSS,LOEWE}$ score is assigned to the combinations considered. However, from the combination dose–response surfaces (Fig. 4c; cyclophosphamide and MK-8669 on OCUB-M, Table S22, SN-38 and ABT-888 on SK-MEL-30, Table S23, SN-38 and erlotinib on SK-MEL-30, Table S24, see supplementary material online) one would not expect strong negative scores across most of the surface as were observed here (Table S25, Table S26, see supplementary material online), because the combined effect is mostly at least as high or higher than that for each of the drugs independently. As for known limitations of the Loewe additivity model, a non-constant potency ratio is observed for the compounds in these combinations, which might explain the unreasonably low S_{LOEWE} and $S_{WSS,LOEWE}$ values [14,17]. Additionally, our results suggest that incompatibility with the Loewe additivity model (as implemented in SynergyFinder™) could arise when the differences in concentrations at which the individual drugs are tested is large. From our results, it is suggested that violations of this requirement result in unreliable and strongly negative values for S_{LOEWE} and $S_{WSS,LOEWE}$.

Notably, opposing observations are made for synergy scores according to the Bliss independence and the HSA model. For example, for PD325801 and dasatinib in the NCIH-2122 cell line, we observe a negative single-agent response for dasatinib in the NCIH-2122 cell line (Figure S10, see supplementary material online). This combination is classified as synergistic according to the $S_{WSS,BLISS}$ but as antagonistic according to the $S_{WSS,LOEWE}$ and the $S_{WSS,HSA}$. Given that the $S_{WSS,HSA}$ is generally expected to be higher than the other metrics, this suggests that the score for $S_{WSS,BLISS}$ here is inexplicably high. As described before, the low negative value for $S_{WSS,LOEWE}$ could be a result of non-compatibility of the Loewe additivity model because the maximum response of each compound in the combination is different, leading to the Loewe additivity being ill-defined [34]. The positive $S_{WSS,BLISS}$ score for a combination where one of the single-agent responses is positive is in line with findings described previously in this review for cyclophosphamide and MK-8869 in OCUB-M. In the case of Bliss independence, incompatibility scenarios expectedly arise from the fact that the accuracy of Bliss independence requires the single-agent responses to be positive and monotonically increasing [14].

Overall, the observations described above suggest that, for compound combinations in which one of the drugs has a negative effect (as opposed to the expected direction), the $S_{WSS,LOEWE}$ values correspond to antagonism whereas values of $S_{WSS,BLISS}$, and in some instances $S_{WSS,ZIP}$ values, correspond to synergistic classifications. Essentially, the assessment of synergy in such combinations does not lead to meaningful values because synergy does not exist when one of the drugs does not elicit a positive response. Instead, an increase in efficacy for such combinations corresponds to nonsynergistic potentiation or enhancement [12]. However, theoretically, these combinations might still possess a promising therapeutic potential if dose-reduction can be achieved by administration of the combination. To select those compound combinations that are most likely to provide a considerable potentiating effect, we propose that those combinations for which S_{BLISS}, S_{HSA} and S_{ZIP} indicate synergy should be selected as promising combinations from large combination screens.

Theoretical limitations cannot predict all metric disagreements

Theoretical limitations cannot predict all metric disagreements and the largest difference in relative rank was found to be observed for the combination of L-778123 and MK-2206 in the LoVo cell line between $S_{WSS,BLISS}$ and $S_{WSS,HSA}$ ($\Delta_{rr} = 0.74$, $S_{WSS,BLISS} = -62.5$, $S_{WSS,HSA} = 142.7$, Table S28, see supplementary material online); and between $S_{WSS,HSA}$ and $S_{WSS,ZIP}$ ($\Delta_{rr} = 0.81$, $S_{WSS,HSA} = 142.7$, $S_{WSS,ZIP} = -79.1$, Table S28, see supplementary material online). The disagreement for this combination is observed between all metrics, with assignments of synergism according to $S_{WSS,LOEWE}$ and $S_{WSS,HSA}$ and assignments to the antagonistic class based on $S_{WSS,BLISS}$ and $S_{WSS,ZIP}$. Here, all values of S_{LOEWE} and S_{HSA} are positive, and all but one of the S_{BLISS} and S_{ZIP} values are negative (Table S27, see supplementary material online). Thus, the disagreement of classification is not merely a result of differences in the magnitude of concentration-specific scores.

The dose–response curves for this combination reveal a difference in slope between L-77123 and MK-2206 (Figure S11, see supplementary material online), a situation in which the Loewe additivity model is known to be unreliable [17]. The same can,

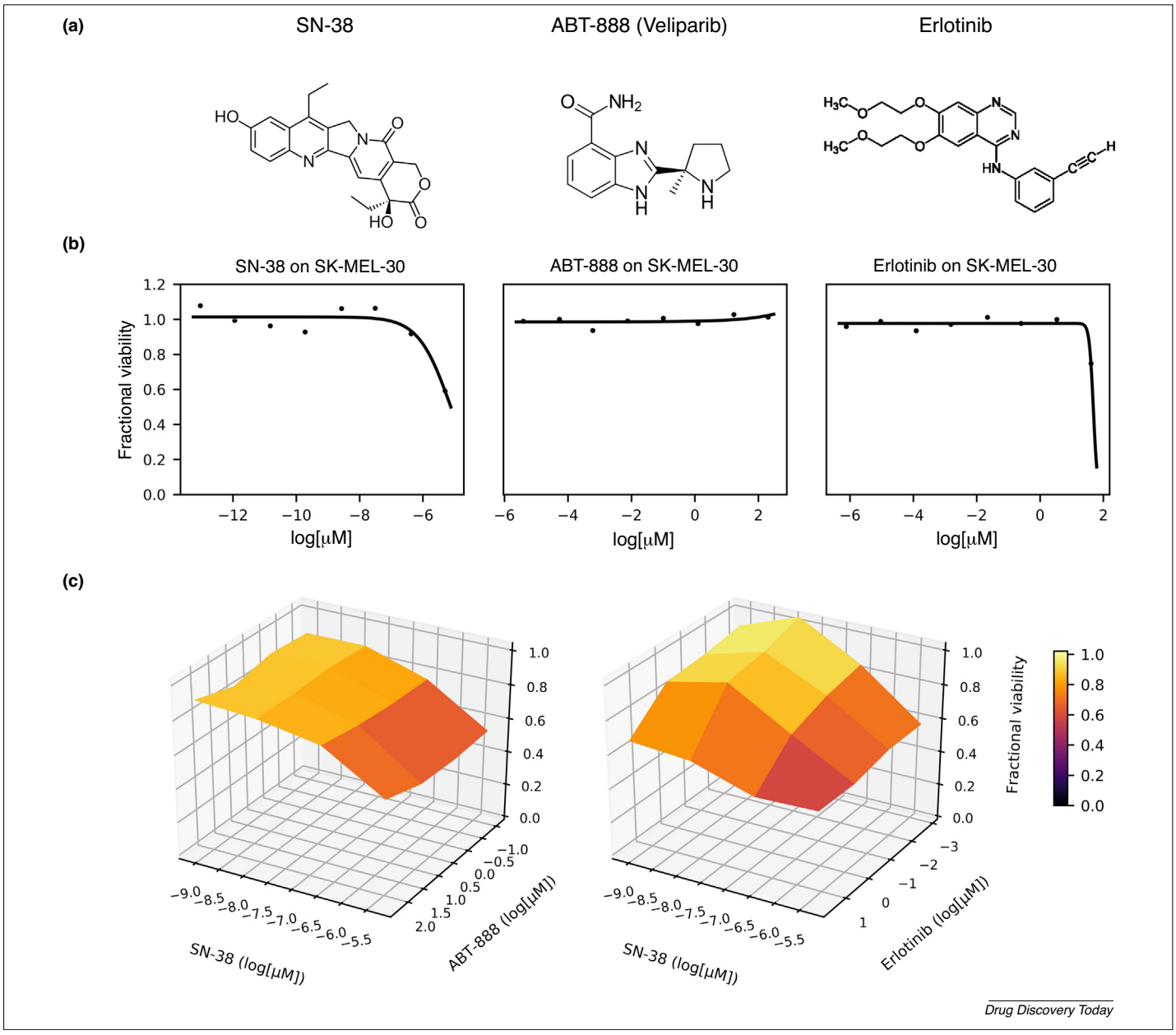

FIGURE 4

The structure, dose–response curves and combination response surface of SN-38, ABT-888 and erlotinib. **(a)** Compounds that are part of the combinations that give very low values for $S_{WSS,LOEWE}$. **(b)** Single-agent, fitted dose–response curves for SN-38, ABT-888 and erlotinib on the SK-MEL-30 cell line. SN-38 and erlotinib were fitted to a log–logistic curve, whereas ABT-888 was fitted to a linear curve. **(c)** The dose–response surface of the combination of SN-38 and ABT-888 in the SK-MEL-30 cell line for which a very low $S_{WSS,LOEWE}$ value of −6000 was found (left), and for the combination of SN-38 and erlotinib in the SK-MEL-30 cell line for which a very low $S_{WSS,LOEWE}$ value of −4000 was found. These very low $S_{WSS,LOEWE}$ values were visually identified as outliers in the correlation plots shown in Figure S9 (see supplementary material online).

however, not be said for the HSA model, from which the same classification was obtained. This example reveals that, even when considering the individual dose–response profile of a combination, unexpected disagreements between metrics can still occur and synergy scores alone do not provide a conclusive answer as to whether a compound combination is synergistic or not.

Perspectives

In this review, we show that metrics for the assessment of synergy should not be applied to real-world experimental large-scale oncology data in a purely mechanistic manner. To prevent erroneous results, the dose–response relationships for all tested compound combinations should be inspected before selection of a suitable synergy reference model for the data at hand. Furthermore, to fully comply with the definition for synergy, compound combinations for which no or a negative response is observed should be removed. In this regard, setting a growth ceiling of 100%, as done for the ComboScore [1] used in the NCI ALMANAC, is not sufficient because it does not eliminate the compound combination from consideration. Additionally, when applying the Loewe additivity model, compound combinations with very different dose–response behaviours cause problems. This includes combinations

where the difference in the slope of dose–response curves is large, or when different maximum response values are observed for each drug in the combination. Importantly, this is not a straightforward task to automate and doing it manually is highly time-consuming and unfeasible when working with large-scale data. Therefore, metrics for which the influence of the shape of the respective dose–response curves is minimal might be easier to implement for the initial analysis of large-scale combination screens. Additionally, choosing the most appropriate model for individual experiments might introduce biases when analysing synergy scores from several experiments, for example for the identification of biomarkers of synergy across multiple cell lines treated with the same drug combination.

Beyond curve shape, data curation should be performed with particular attention being paid to quality control, because the confidence with which a metric can be applied is fundamentally dependent on the experimental noise. One efficient way to deal with this is the use of a quality control metric like mQC, a data-driven metric that assigns a 'good', 'medium' or 'bad' rating with an associated confidence score, to the quality of the single dose–response matrix [35]. This empirical score is produced by an AdaBoost Decision Tree model trained using crowdsourced labels assigned by experts and it can be used in a similar manner as the Z′ value commonly used for biochemical screens [35]. Additionally, when the tested concentrations for the single agent experiments and the combination experiments do not align, as in the Merck dataset used in this analysis, the selection of an appropriate dose–response curve that accurately represents the data is essential. However, given that in many cases dose–response curves cannot easily be fitted, this step is often difficult to perform. Most commonly, dose–response curves are fitted to a Hill equation, for example a 4-parameter log–logistic model. However, in some cases or in certain parts of the dose–response curve, multiphasic [36], exponential or linear curves might be more suitable. Accurate fitting of the correct dose–response curve is of particular importance because, when the curve fit is suboptimal, the interpolated and extrapolated values will be inaccurate, as will be any subsequently calculated synergy value. Besides the selection of the right type of dose–response fit, it is essential that enough concentrations at the right intervals and absolute concentration are used to ensure the capture of the full dose–response curve. This is particularly important when using a metric for which the minimum effect, maximum effect and EC_{50} values need to be accurately estimated, such as Loewe additivity and ZIP. Ensuring data quality and the accurate representation of the dose–response profile will decrease the chance of observing false-positive and false-negative label assignments, thus improving the accuracy of the whole analysis.

From the metrics discussed here, HSA is the only metric where the score is not dependent on the shape of the dose–response relationship and makes no assumptions about the origin of the pharmacodynamic effect. Although it should be noted that HSA is not recommended as a sensitive synergy metric owing to the high false-positive rate, this makes the HSA score easy to interpret. This could be beneficial in the assessment of large combination screens where inspection of each individual is not feasible, the underlying mechanisms are not the same for each compound–cell line combination, and drug combinations that show potentiation (augmentation, enhancement) should not be excluded. For Loewe additivity and ZIP, accurate fitting to a log–logistic function is required for their accurate assessment. Besides, for Loewe additivity, a constant potency ratio and equal efficacy are required. As for the assumption regarding the origin of pharmacodynamic effects, in the Bliss independence reference model, drugs are assumed to act through independent pathways; and in the Loewe additivity reference model drugs are expected to act via the same conceptual mechanism.

Indeed, a large proportion of false positives will generally be encountered when using HSA as a reference model in the assessment of large combination screens. To illustrate, HSA scores are generally higher (i.e., higher leniency) resulting in 39% of all scores having values larger than 0 in our analysis, compared with 29% for Loewe additivity, 31% for Bliss independence and 37% for ZIP. This need not be a problem, especially in cases where only the relative scoring of a compound combination is required to prioritise candidates for subsequent investigation. Besides, HSA does not give an indication whether the combined effect is higher than would be expected based on the individual effects of both drugs in the combination. As a result, HSA scores cannot be used to differentiate between potentiation and synergy. Additionally, as also seen for Bliss independence [14], when dose–response curves are steep, a combination that might get a score that indicates synergy, administration of a slightly higher dose might be more beneficial in terms of the desired effect than a combination with another drug. However, this is dependent on the complete desired effect versus side-effect profile, and the intended therapeutic effect. Another concern is that, in cases of combined negative and positive responses for the single agents in a combination, care should be taken to identify the desired direction of the observed effect. For example, if a positive response is desired, the highest negative score should be selected from single-agent responses where both responses are negative, whereas the lowest negative value should be selected in case a negative response is desired.

Additionally, more-sophisticated metrics like Loewe additivity and Bliss independence could have better predictive power for therapeutic potential for specific compound combinations. To illustrate, it has been asserted that Loewe additivity or the CI-index is more suitable for mechanistic and clinical research if the dose–effect curves are well characterised [15]. Additionally, when considering an oncology screen most effects will be unspecific cytotoxic responses through similar mechanisms or nonallosteric targeted therapies, in which case the use of Loewe would seem justified. However, the recent publication by Gilvary *et al.* [30] found that Bliss independence corresponds best to combination successes in the clinic within the NCI ALMANAC screen which contains cytotoxic and target-specific antineoplastic compounds. In conclusion, there is no silver bullet when selecting the appropriate synergy metric for the analysis of large combination screens owing to the complex combination of advantages and disadvantages concerning different metrics. However, with more data becoming available, biological interpretation and synergy biomarkers could guide the choice of synergy metrics.

Concluding remarks

Upon investigation of the literature and the analysis of a large-scale combination screen we conclude that the diversity of dose–response profiles of individual drugs render the application of

one-size-fits-all, automated approaches to analyse large-scale compound synergies problematic. Therefore, great care should be taken when selecting a synergy metric for the assessment of such data. We hope that this work will hereby serve as a reference for important considerations in the process.

Acknowledgements

We thank our colleague Dr F. Richards for his constructive feedback and helpful comments on this work. This work was supported by the Leiden International Study Fund, the Netherlands (grant number L17077) to A.H.C.; and the Jo Kolk study fund, the Netherlands to A.H.C.; and the Erasmus + Traineeship fund to A.H.C.; and the KNMP stipend, the Netherlands to A.H.C.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.drudis.2019.09.002.

References

1 Holbeck, S.L. *et al.* (2017) The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res.* 77, 3564–3576
2 O'Neil, J. *et al.* (2016) An unbiased oncology compound screen to identify novel combination strategies. *Mol. Cancer Ther.* 15, 1155–1162
3 Menden, M.P. *et al.* (2019) Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* 10, 2674
4 Lundholt, B.K. *et al.* (2003) A simple technique for reducing edge effect in cell-based assays. *J. Biomol. Screen.* 8, 566–570
5 Iloki Assanga, S.B. *et al.* (2013) Cell growth curves for different cell lines and their relationship with biological activities. *Int. J. Biotechnol. Mol. Biol. Res.* 4, 60–70
6 Loewe, S. *et al.* (1926) Über Kombinationswirkungen. *Naunyn. Schmiedebergs. Arch. Exp. Pathol. Pharmakol* 114, 313–326
7 Loewe, S. (1928) Die quantitativen probleme der pharmakologie. *Ergebnisse. Der Physiol.* 27, 47–187
8 Loewe, S. (1953) The problem of synergism and antagonism of combined drugs. *Arznei-Forschung* 2, 285–290
9 Bliss, C. (1939) The toxicity of poisons applied jointly. *Annu. Appl. Biol.* 26, 585–615
10 Gaddum, J.H. (1940) *Pharmacology*. pp. 378–383, Oxford University Press, London
11 Yadav, B. *et al.* (2015) Searching for drug synergy in complex dose?response landscapes using an interaction potency model. *Comput. Struct. Biotechnol. J.* 13, 504–513
12 Chou, T.C. (2006) Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacol. Rev.* 58, 621–681
13 Chou, T.C. and Talalay, P. (1983) Analysis of combined drug effects: a new look at a very old problem. *Trends Pharmacol. Sci.* 4, 450–454
14 Greco, W.R. (1995) The search for synergy: a critical review from a response surface perspective. *Pharmacol. Rev.* 47, 332–382
15 Foucquier, J. and Guedj, M. (2015) Analysis of drug combinations: current methodological landscape. *Pharmacol. Res. Perspect.* 3, e00149
16 Roell, K.R. *et al.* (2017) An introduction to terminology and methodology of chemical synergy-perspectives from across disciplines. *Front. Pharmacol.* 8, 1–11
17 Lederer, S. *et al.* (2018) Additive dose response models: explicit formulation and the Loewe additivity consistency condition. *Front. Pharmacol.* 9, 1–11
18 Di Veroli, G.Y. *et al.* (2016) Combenefit: an interactive platform for the analysis and visualization of drug combinations. *Bioinformatics* 32, 2866–2868
19 Twarog, N.R. *et al.* (2016) BRAID: a unifying paradigm for the analysis of combined drug action. *Sci. Rep.* 6, 25523
20 Schindler, M. (2017) Theory of synergistic effects: Hill-type response surfaces as 'null-interaction' models for mixtures. *Theor. Biol. Med. Model.* 14, 15
21 Zimmer, A. *et al.* (2016) Prediction of multidimensional drug dose responses based on measurements of drug pairs. *Proc. Natl. Acad. Sci. U. S. A.* 113, 10442–10447
22 Meyer, C. *et al.* (2019) Quantifying drug combination synergy along potency and efficacy axes. *Cell Syst.* 8 97-108.e16
23 Goldoni, M. and Johansson, C. (2007) A mathematical approach to study combined effects of toxicants *in vitro*: evaluation of the Bliss independence criterion and the Loewe additivity model. *Toxicol. Vitr.* 21, 759–769
24 Drescher, K. and Boedeker, W. (1995) Assessment of the combined effects of substances: the relationship between concentration addition and independent action. *Biometrics* 51, 716–730
25 Chen, C.-Y. and Christensen, E.R. (1985) A unified theory for microbial growth under multiple nutrient limitation. *Water Res* 19, 791–798
26 Baeder, D.Y. *et al.* (2016) Antimicrobial combinations: Bliss independence and Loewe additivity derived from mechanistic multi-hit models. *Philos. Trans. R. Soc. B* 371, 20150294
27 Preuer, K. *et al.* (2018) DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* 34, 1538–1546
28 Sidorov, P. *et al.* (2019) Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *Front. Chem* 7, 509
29 Mathews Griner, L.A. *et al.* (2014) High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell-like diffuse large B-cell lymphoma cells. *PNAS* 111, 2349–2354
30 Gilvary, C.M. *et al.* (2019) Multi-task learning predicts drug combination synergy in cells and in the clinic. *BioRxiv* . http://dx.doi.org/10.1101/576017
31 Bulusu, K.C. *et al.* (2016) Modelling of compound combination effects and applications to efficacy and toxicity: state-of-the-art, challenges and perspectives. *Drug Discov. Today* 21, 225–238
32 Chou, T.C. (2010) Drug combination studies and their synergy quantification using the Chou-Talalay method. *Cancer Res.* 70, 440–446
33 Ianevski, A. *et al.* (2017) SynergyFinder: a web application for analyzing drug combination dose-response matrix data. *Bioinformatics* 33, 2413–2415
34 Jonker, D.M. *et al.* (2004) Towards a mechanism-based analysis of pharmacodynamic drug-drug interactions *in vivo*. *Pharmacol. Ther.* 106, 1–18
35 Chen, L. *et al.* (2016) mQC: a heuristic quality-control metric for high-throughput drug combination screening. *Sci. Rep.* 6, 37741
36 Di Veroli, G.Y. *et al.* (2015) An automated fitting procedure and software for dose-response curves with multiphasic features. *Sci. Rep.* 5, 14701
37 Martin-Betancor, K. *et al.* (2015) Defining an additivity framework for mixture research in inducible whole-cell biosensors. *Sci. Rep.* 5, 17200
38 Spearman, C. (1904) The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101
39 Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46