A shape-constrained regression and wild bootstrap framework for reproducible drug synergy testing

Amir Asiaee[1*], James P. Long[2], Samhita Pal[1], Heather H. Pua[3], Kevin R. Coombes[4]

[1*]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, 37232, USA.
[2]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA.
[3]Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, 37232, USA.
[4]Department of Population Health Sciences, Georgia Cancer Center, Augusta University, Augusta, GA, 30912, USA.

*Corresponding author(s). E-mail(s): amir.asiaeetaheri@vumc.org;
Contributing authors: samhita.pal@vumc.org;

Abstract

High-throughput drug combination screens motivate computational methods to identify synergistic pairs, yet synergy is typically quantified by heuristic scores (Bliss, HSA, Loewe, ZIP) that provide no statistical inference and can be unstable or undefined when parametric dose–response fits fail. We present a nonparametric, assumption-light framework that defines interaction as the deviation from a monotone-additive null within a shared monotone model class. We fit a monotone surface by two-dimensional isotonic regression and a monotone-additive surface, compute an interaction surface, and summarize global interaction by a stable "interaction energy" statistic. A degrees-of-freedom-corrected wild bootstrap yields calibrated p-values for testing interaction in each dose–response matrix, enabling principled hit calling and multiple-testing control. On DrugCombDB, our method yields higher replicate concordance of interaction surfaces (median correlation 0.91 across 1,839 replicate pairs) than Bliss, HSA, Loewe, or ZIP (0.53–0.74), while avoiding the 20.9% Loewe and 3.6% ZIP failure rates. Because the fitted surface is generative, the method also predicts missing wells (median holdout RMSE 0.040 in viability units). By turning synergy scoring into statistically grounded outcomes (effect sizes with uncertainty), the framework provides more reliable targets for downstream machine learning models of combination response.

Keywords: Drug combinations, synergy, shape-constrained regression, isotonic regression, wild bootstrap, hypothesis testing, reproducibility

1 Introduction

Drug combinations are central to combination therapy and are routinely screened in vitro across dose matrices. The dominant computational practice is to assign each matrix a synergy *score* under one of several null models—Bliss independence [1], highest single agent (HSA), Loewe additivity [2], or ZIP [3]—and then rank "hits". However, the field remains fragmented over which null model is appropriate [4], and most synergy scores provide neither uncertainty quantification nor a principled global decision rule for whether interaction is present. These limitations are increasingly consequential: modern machine learning (ML) approaches often treat synergy scores as training labels [5, 6], but if labels are inconsistent across models or unstable to noise and missingness, predictive models are incentivized to chase idiosyncrasies of the chosen scoring rule rather than reproducible biology. In parallel, parametric dose–response modeling (e.g., Hill-type curves and their multi-parameter extensions) is widely used to smooth marginal responses, yet it can be sensitive to preprocessing and model misfit [7, 8] and can fail to converge on real screening data [3].

We introduce a statistical framework that (i) defines interaction in the same geometric model class used for estimation; (ii) avoids parametric curve fitting by using shape constraints motivated by basic pharmacology (monotonicity of effect with dose); and (iii) provides calibrated p-values for testing interaction per matrix via a wild bootstrap. The key idea is to compare a flexible monotone surface fit (2D isotonic regression) to a monotone-additive null fit on the same transformed scale. Their difference is an interaction surface that captures non-additivity while respecting the monotonicity constraint (increasing dose should not decrease effect). A global interaction statistic summarizes the matrix and, through bootstrap inference, supports reproducible discovery with error-rate control.

2 Results

2.1 Synergy labels are fragmented across null models

To quantify disagreement among widely used baselines, we analyzed DrugCombDB [9] blocks with finite Bliss, HSA, Loewe and ZIP scores (391,652 matrices) as provided by SynergyFinder conventions [3]. Baseline scores show strong dependence on the chosen null: Bliss and ZIP correlate highly (Pearson $r = 0.92$), while Loewe correlates weakly with ZIP ($r = 0.28$) and Bliss ($r = 0.33$) (Fig. 1). Disagreement is not limited to scaling: for 21–34% of matrices, only one of a baseline pair reports positive synergy (Fig. 1). Even "top hit" sets overlap modestly: among the top 5% most synergistic calls, the Loewe–ZIP Jaccard overlap is 0.36 (Fig. 1). These discrepancies imply that "synergy" labels used in downstream analyses—including ML training [5, 6]—can vary substantially with the scoring model.

2.2 A shape-constrained definition of interaction

We model a drug combination experiment as an $I \times J$ grid of responses $Y_{ij} \in [0,1]$ (viability) measured at increasing doses of two drugs. We transform responses to an unconstrained scale, $Z_{ij} = \mathrm{logit}(Y_{ij})$, and compute inverse-variance weights from within-cell replicates (Online Methods). We then fit two nested model classes: (i) a flexible monotone surface $\hat{\theta}^{\mathrm{iso}} \in \mathcal{M}$ via 2D isotonic regression; and (ii) a monotone-additive surface $\hat{\theta}^{\mathrm{add}} \in \mathcal{A} \subset \mathcal{M}$, $\theta_{ij} = \alpha + u_i + v_j$, under the same monotonicity direction. The interaction surface is the model-consistent deviation

$$\delta_{ij} = \hat{\theta}^{\mathrm{iso}}_{ij} - \hat{\theta}^{\mathrm{add}}_{ij}, \tag{1}$$

interpretable as synergy (for viability) when $\delta_{ij} < 0$ (lower viability than additive expectation). We summarize global interaction by the weighted interaction energy $S^2 = \sum_{ij} w_{ij}\delta^2_{ij}$, which measures total squared deviation from additivity across the grid—analogous to an F-statistic testing whether any regression coefficients are non-zero (Online Methods).

2.3 Calibrated hypothesis testing by df-corrected wild bootstrap

Baseline scores provide no statistical inference, which prevents principled error control in large screens. We therefore use a Rademacher wild bootstrap [10] under the monotone-additive null to obtain a p-value for each matrix. Because the null is fit to the observed data, naive residual resampling can understate noise; we apply a degrees-of-freedom correction to rescale residuals (Online Methods).

We assessed calibration on DrugCombDB via a pseudo-null experiment: we fit the null model to real matrices, generate synthetic responses by sign-flipping df-scaled residuals, and rerun the full bootstrap procedure. Under this constructed null, p-values are close to uniform (Fig. 4): across $n = 300$ pseudo-null matrices, the observed Type I error rates are 3.3% at $\alpha = 0.05$ and 5.7% at $\alpha = 0.10$. We additionally assessed power in simulation across interaction strengths, observing increasing power with stronger departures from additivity (Fig. 3).

2.4 Higher reproducibility and zero failures on replicate experiments

Reproducible synergy estimates are essential for translating screens into follow-up experiments. We benchmarked replicate concordance on 1,839 replicate pairs drawn from 1,209 DrugCombDB experiments with repeated measurements of the same drug pair in the same cell line (Online Methods). Our interaction surface on the transformed scale (δ_Z) is substantially more reproducible than baseline synergy surfaces: median

Table 1 Capabilities of the proposed framework versus common baselines. Baselines provide pointwise scores without uncertainty; some require parametric curve fitting, which can fail on real data.

Capability	Isotonic	Bliss	HSA	Loewe	ZIP
Statistical test	✓	–	–	–	–
Global summary	✓	–	–	–	–
Predicts missing wells	✓	–	–	–	–
Stable under perturbation*	✓	–	–	–	–
No fit failures	✓	✓	✓	–	–
Assumption	Monotone	Independence	Max effect	Dose equivalence	Independence + curve fit

*Perturbation stability evaluated in the accompanying repository experiments (Supplementary Methods).

replicate correlation is 0.91 for δ_Z, compared to 0.53 (Bliss), 0.61 (HSA), 0.74 (Loewe) and 0.71 (ZIP) (Fig. 5). In addition, parametric baselines can be undefined: Loewe and ZIP fail (non-finite output) in 20.9% and 3.6% of replicate experiments, respectively, while the proposed shape-constrained fits succeed in all cases (Fig. 5).

2.5 A generative surface model enables prediction of missing wells

Drug combination matrices are often incomplete due to plate layout constraints, assay failures, or adaptive designs. Because our method fits an explicit monotone surface, it naturally predicts unobserved dose pairs by evaluating $\hat{\theta}^{\text{iso}}$ and transforming back to viability. In a holdout benchmark on 200 DrugCombDB matrices, we withheld 20% of interior wells and predicted their viabilities from the remaining data. Predictions are accurate (median RMSE 0.040 in viability units), and the induced synergy summaries are similarly stable (median RMSE 0.035 for S_{proposed}) (Fig. 6).

3 Discussion

We presented a statistical framework for drug combination interaction that replaces fragile parametric curve fitting with shape-constrained regression and supplies calibrated p-values by a df-corrected wild bootstrap. The method addresses three recurrent problems in synergy analysis. First, baseline null models disagree substantially on the same data (Fig. 1), implying that "synergy" labels are not a stable ground truth. Second, baselines that rely on parametric marginal models (Loewe, ZIP) can fail on real matrices, whereas isotonic regression always yields a feasible monotone fit (Fig. 5). Third, the absence of uncertainty quantification in common scoring approaches prevents error-rate control and principled aggregation across experiments.

These issues are increasingly important in the era of ML-guided combination discovery. Many ML pipelines either predict dose–response surfaces and derive summary outcomes (e.g., IC_{50} or AUC) or directly predict synergy scores [5, 6]. Point estimates such as IC_{50} can be sensitive to curve fit and dose range [7], and our analysis shows that widely used synergy scores can be inconsistent and unstable. A natural implication is that predictive models trained on a single baseline score may learn to reproduce a particular scoring convention rather than robust interaction effects. By defining interaction within a minimal, biologically motivated model class and returning both effect sizes and calibrated p-values, our framework provides more defensible outcomes for screening, meta-analysis, and ML supervision.

Direction of interaction and multiplicity.

The global interaction energy S^2 is inherently two-sided: it tests whether the dose-response surface deviates from monotone additivity anywhere, without committing to a single sign across the grid. After rejecting the global null, users may summarize direction using directional energies (Supplementary Methods) and, if desired, apply sequential directional testing; because directional conclusions are conditioned on a significant global test, family-wise error can be controlled analogously to post-hoc direction tests following an F-test.

Monotonicity as a minimal assumption.

Monotonicity can fail in cases such as hormesis or biphasic responses [4]. Rather than assuming a parametric form, we treat monotonicity as a regularizing constraint—a weak assumption that stabilizes estimation on sparse, noisy grids without imposing a specific functional form. Dataset-scale diagnostics on DrugCombDB indicate that most apparent monotonicity violations are small in magnitude and consistent with noise (Supplementary Fig. 1), supporting monotone regression as a pragmatic default. When strong non-monotonicity is suspected, diagnostics can flag affected matrices for alternative modeling.

Outlook.

The framework readily extends to alternative response transforms, weighted designs, and different global statistics (Online Methods). More broadly, replacing heuristic scores with statistically grounded estimators can help reconcile the fragmented landscape of synergy models [4] by enabling calibrated discovery and reproducible benchmarking.

4 Online Methods

4.1 Data sources and preprocessing

We evaluated the method primarily on DrugCombDB [9], a curated database of drug combination experiments with dose matrices across many drugs and cell lines. We used viability responses scaled to $[0, 1]$ and standardized doses to a common unit when available. DrugCombDB provides baseline synergy scores for Bliss, HSA, Loewe and ZIP via SynergyFinder conventions [3]; we used these for baseline disagreement analyses.

4.2 Transform and weights

Let $Y_{ij} \in [0, 1]$ denote viability at dose pair (i, j). We transform responses to $Z_{ij} = \text{logit}(Y_{ij})$, clamping Y_{ij} to $[\epsilon, 1 - \epsilon]$ with $\epsilon = 0.001$ to avoid infinities. When replicate measurements are available, we compute inverse-variance weights

$$w_{ij} = \frac{m_{ij}}{\max(s_{ij}^2, \tau)}, \tag{2}$$

where m_{ij} is the number of replicates, s_{ij}^2 is the sample variance, and τ is a small variance floor.

4.3 Model classes and estimation

Let $\mathcal{M}$ denote the set of monotone surfaces on an $I \times J$ grid (non-increasing in each coordinate for viability). The isotonic estimator is the weighted least-squares projection

$$\hat{\theta}^{\text{iso}} = \arg\min_{\theta \in \mathcal{M}} \sum_{i,j} w_{ij}(Z_{ij} - \theta_{ij})^2. \tag{3}$$

Let $\mathcal{A} \subset \mathcal{M}$ denote the monotone-additive class $\theta_{ij} = \alpha + u_i + v_j$, with u and v constrained to be monotone in the same direction and with identifiability constraints. We compute $\hat{\theta}^{\text{add}}$ as the weighted least-squares projection onto $\mathcal{A}$. Both problems are solved as convex quadratic programs with linear inequality constraints using OSQP [11].

4.4 Interaction surface and summaries

We define interaction as $\delta = \hat{\theta}^{\text{iso}} - \hat{\theta}^{\text{add}}$. We report effect sizes on the viability scale as

$$S_{\text{proposed}} = \text{logit}^{-1}(\hat{\theta}^{\text{add}}) - \text{logit}^{-1}(\hat{\theta}^{\text{iso}}). \tag{4}$$

Global interaction is summarized by $S^2 = \sum_{ij} w_{ij}\delta_{ij}^2$ and its normalized variants (e.g., $S^2/\sum w$).

4.5 Wild bootstrap inference with df correction

To test the null hypothesis of monotone additivity, we use a Rademacher wild bootstrap [10]. We fit the null model, compute residuals $r_{ij} = Z_{ij} - \hat{\theta}_{ij}^{\text{add}}$, apply a degrees-of-freedom correction $\tilde{r}_{ij} = r_{ij}\sqrt{n_{\text{eff}}/(n_{\text{eff}} - df_{\text{null}})}$, generate bootstrap data $Z_{ij}^{\star} = \hat{\theta}_{ij}^{\text{add}} + \xi_{ij}\tilde{r}_{ij}$ with $\xi_{ij} \in \{-1, 1\}$, refit both models on $Z^{\star}$, and recompute the chosen statistic. The p-value is

$$p = \frac{1 + \#\{T_b^{\star} \geq T_{\text{obs}}\}}{B + 1}. \tag{5}$$

We estimate df_{null} from the number of pooled levels in the fitted monotone main effects (Online Note: `experiments/00-bootstrap-df-scaling.md`).

4.6 Benchmarks

Baseline disagreement: We computed pairwise correlations, sign disagreement rates, and top-hit overlaps across baseline scores on DrugCombDB blocks with finite Bliss, HSA, Loewe and ZIP scores (391,652 matrices). **Replicate concordance:** We sampled repeated experiments for the same drug pair and cell line, computed synergy/interaction surfaces for each replicate experiment, and evaluated replicate correlations and failure rates. **Missingness prediction:** For each of 200 matrices, we held out 20% of interior wells, fit the proposed model on the remaining wells, and evaluated RMSE on held-out viabilities and derived synergy summaries. **Pseudo-null calibration:** We generated pseudo-null data by sign-flipping df-corrected residuals from the fitted null and recomputed bootstrap p-values. **Simulation:** We simulated monotone-additive and interacting surfaces on 8×8 grids with Gaussian noise on the transformed scale and evaluated calibration and power across interaction strengths.

4.7 Code availability

All analyses and figure generation are implemented in an R pipeline in the accompanying repository, including scripts to reproduce all tables and figures in this manuscript. Code will be made available upon publication.

Acknowledgements

[To be added]

Author contributions

[To be added]

Competing interests

The authors declare no competing interests.

References

[1] Bliss, C. I. The toxicity of poisons applied jointly. *Annals of Applied Biology* **26**, 585–615 (1939).

[2] Loewe, S. & Muischnek, H. Über kombinationswirkungen. *Archiv für Experimentelle Pathologie und Pharmakologie* **114**, 313–326 (1926).

[3] Ianevski, A., He, L., Aittokallio, T. & Tang, J. Synergyfinder: a web application for analyzing drug combination dose–response matrix data. *Bioinformatics* **33**, 2413–2415 (2017).

[4] Meyer, C. T., Wooten, D. J., Lopez, C. F. & Quaranta, V. Charting the fragmented landscape of drug synergy. *Trends in Pharmacological Sciences* **41**, 266–280 (2020).

[5] Preuer, K. *et al.* Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* **34**, 1538–1546 (2018).

[6] Kuenzi, B. M. *et al.* Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684.e6 (2020).

[7] Huang, S. & Pang, L. Comparing statistical methods for quantifying drug sensitivity based on in vitro dose–response assays. *ASSAY and Drug Development Technologies* **10**, 88–96 (2012).

[8] Wooten, D. J., Meyer, C. T., Lubbock, A. L. R., Quaranta, V. & Lopez, C. F. Musyc is a consensus framework that unifies multi-drug synergy metrics for combinatorial drug discovery. *Nature Communications* **12**, 4601 (2021).

[9] Liu, H. *et al.* Drugcombdb: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Research* **48**, D871–D881 (2020).

[10] Wu, C. F. J. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* **14**, 1261–1295 (1986).

[11] Stellato, B., Banjac, G., Goulart, P., Bemporad, A. & Boyd, S. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation* **12**, 637–672 (2020).

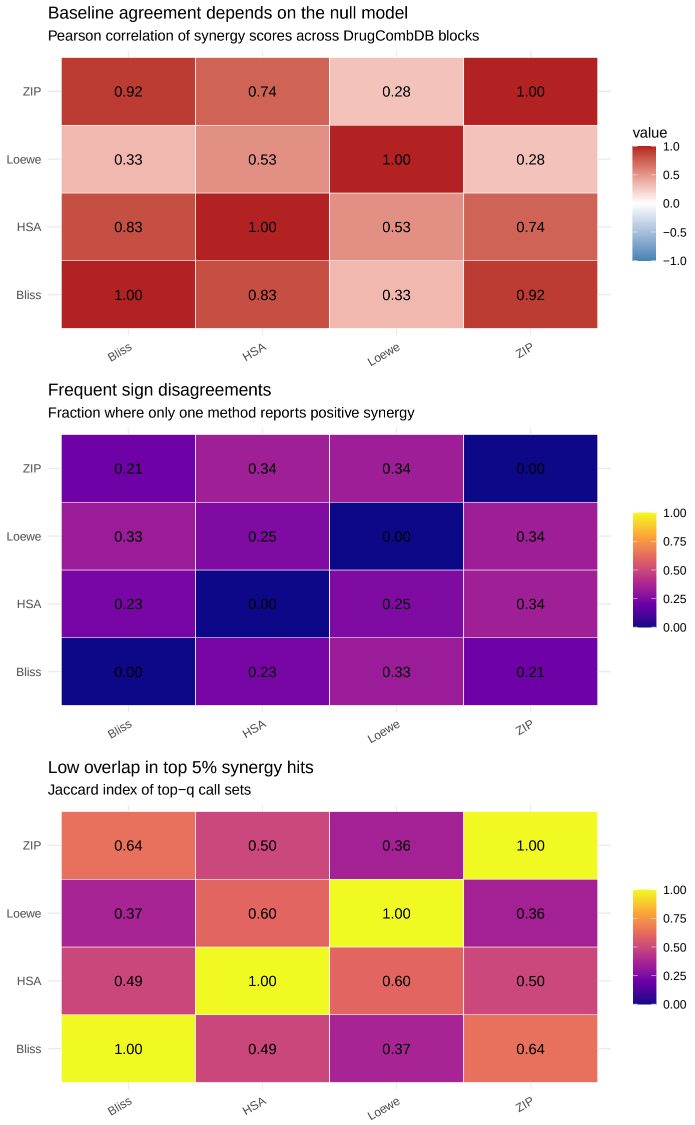

Fig. 1 Baseline synergy scores depend strongly on the null model. Pairwise Pearson correlations (top), fraction of blocks where only one method reports positive synergy (middle), and overlap of top 5% synergy calls (bottom; Jaccard index) across DrugCombDB blocks with finite scores for Bliss, HSA, Loewe and ZIP ($n = 391{,}652$). High correlation of Bliss and ZIP reflects shared independence assumptions, whereas Loewe often disagrees with both, consistent with its dose-equivalence construction and reliance on parametric marginal fits in common implementations.

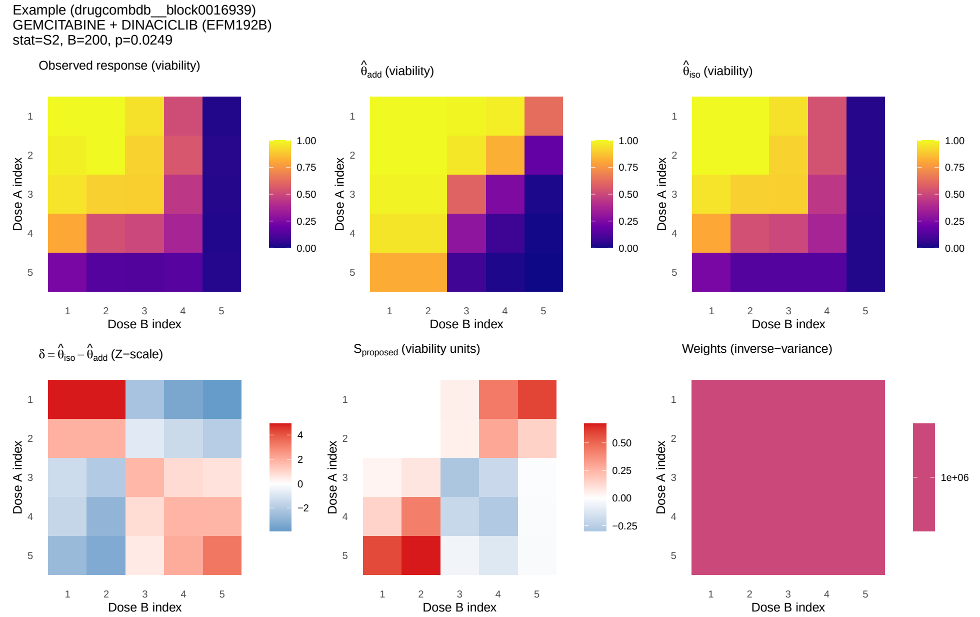

Fig. 2 Method overview on a DrugCombDB example. From observed viability on a dose grid (top left), we fit a monotone-additive null surface $\hat{\theta}^{\text{add}}$ and a monotone isotonic surface $\hat{\theta}^{\text{iso}}$ on the transformed scale. Their difference $\delta = \hat{\theta}^{\text{iso}} - \hat{\theta}^{\text{add}}$ defines interaction in the same constrained geometry. We report effect sizes on the viability scale as $S_{\text{proposed}} = \text{logit}^{-1}(\hat{\theta}^{\text{add}}) - \text{logit}^{-1}(\hat{\theta}^{\text{iso}})$, and compute a wild-bootstrap p-value for a global interaction statistic (example: $p = 0.0249$, $B = 200$). Inverse-variance weights (right) incorporate replicate variability.

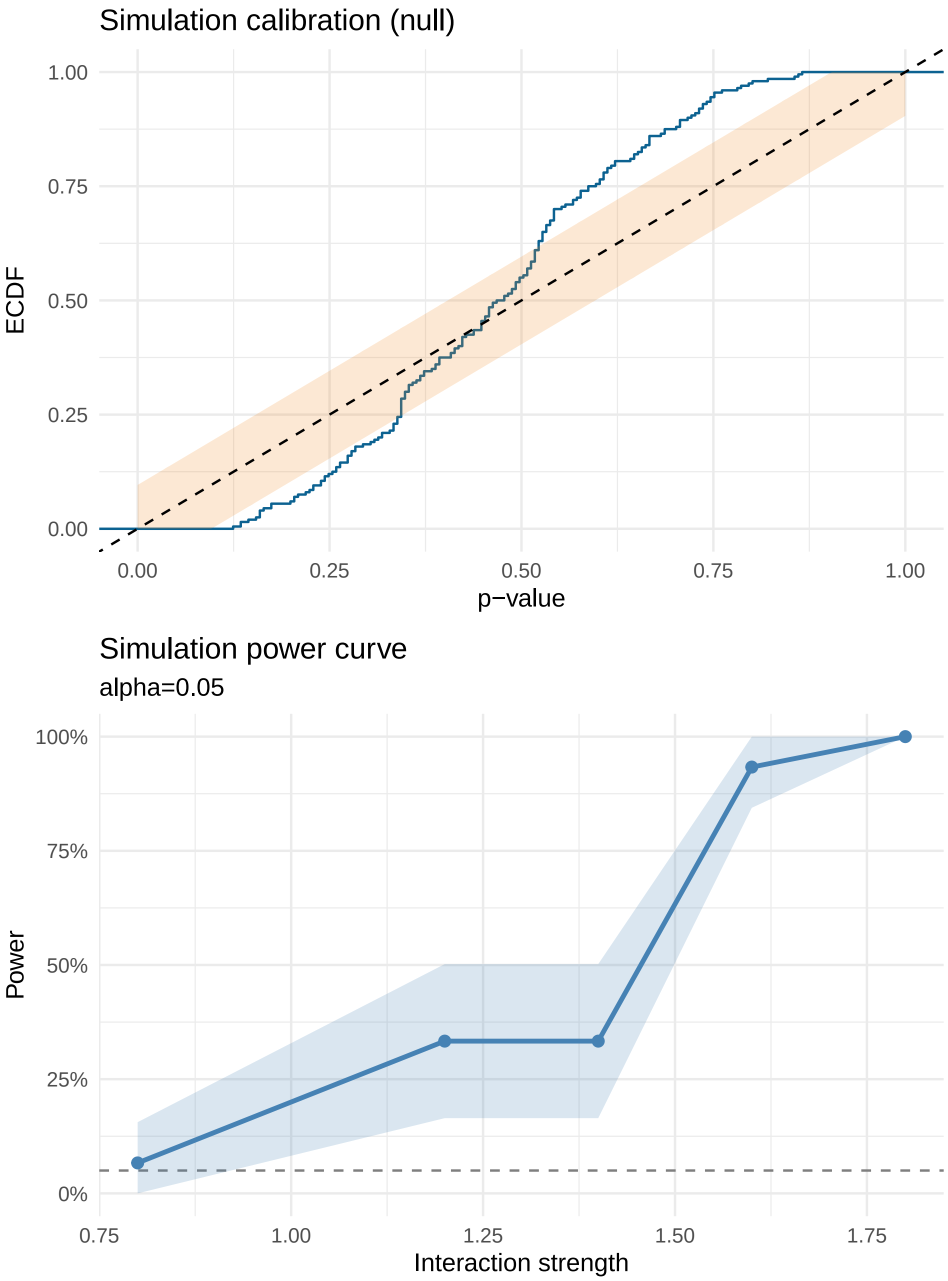

Fig. 3 Simulation calibration and power. Top: empirical CDF of p-values under a simulated additive null (8×8 grids; $\sigma = 0.1$ on the transformed scale; $n = 200$ simulations; $B = 200$ bootstraps) with a Dvoretzky–Kiefer–Wolfowitz band (95%). Bottom: power to detect interaction ($\alpha = 0.05$) as interaction strength increases ($n = 30$ simulations per strength). In this design, the global test is conservative under the null while achieving high power at larger interaction strengths.

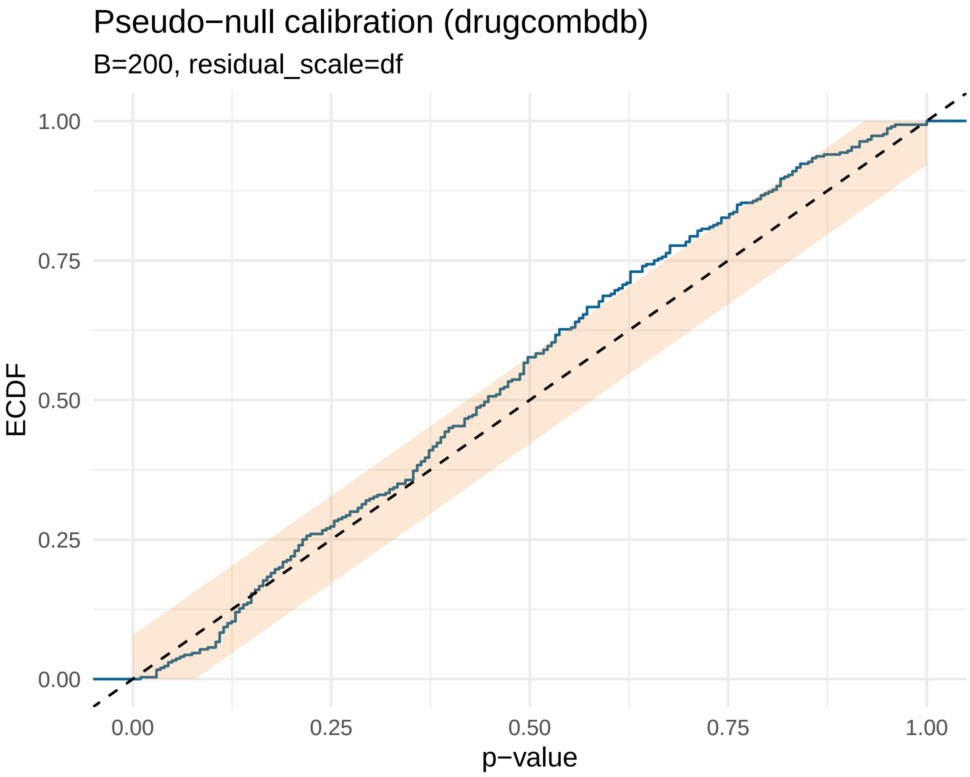

Fig. 4 Pseudo-null calibration on DrugCombDB. P-values from the df-corrected wild bootstrap are approximately uniform when data are generated from the fitted monotone-additive null by sign-flipping df-scaled residuals ($n = 300$ pseudo-null matrices; $B = 200$). The dashed diagonal is the Uniform$(0, 1)$ reference; the shaded region is a 95% Dvoretzky–Kiefer–Wolfowitz band.

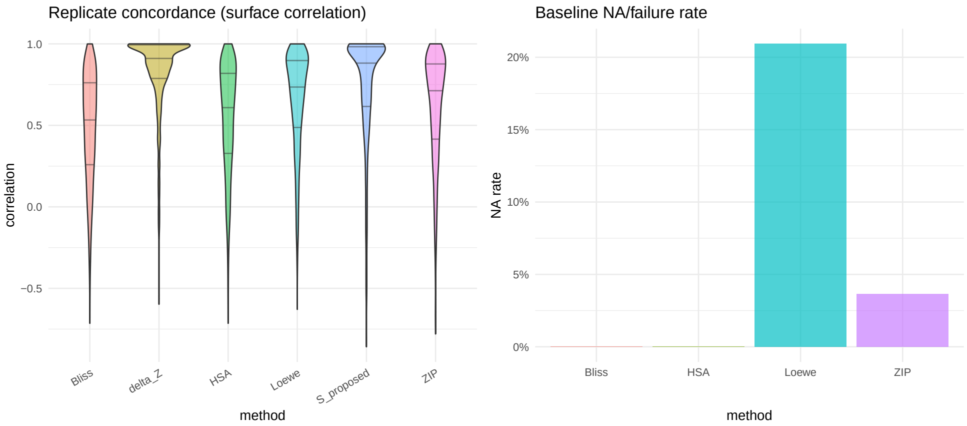

Fig. 5 Replicate reproducibility and baseline failure rates. Left: distributions of replicate correlations between synergy/interaction surfaces across 1,839 replicate pairs. Right: fraction of experiments where baseline methods return NA/non-finite values (Loewe: 20.9%; ZIP: 3.6%; n = 1,209 experiments), reflecting reliance on parametric curve fitting. The proposed interaction surface and summaries are defined for all experiments.

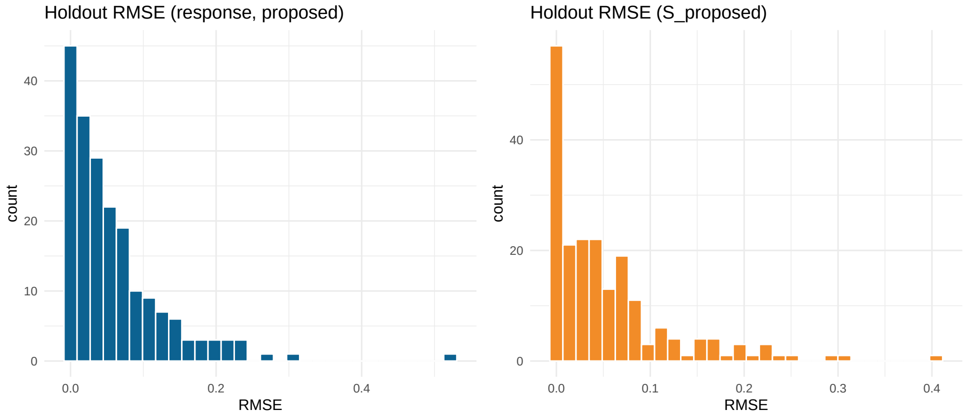

Fig. 6 Prediction under missingness. Holdout RMSE distributions for predicting viability responses (left) and the proposed synergy summary S_{proposed} (right) in a 20% interior-well holdout benchmark ($n = 200$ matrices). Baseline synergy scores are not generative and do not provide a comparable prediction task.