We thank all the reviewers for their feedback. We are glad to see that the reviewers recognize the relevance of this work (R1, R2) and that they appreciate the clarity of our framework (R2, R4), the theoretical rigor of our approach (R4) and the extensiveness of our experiments (R2, R3). We hope this rebuttal addresses your main concerns, and we will incorporate all revisions into the updated version of this work.

Novelty. There are two main contributions in this paper: 1) A theoretical contribution that enhances our understanding of the trade-offs made by various methods and that shows why Shapley values are well justified for global feature importance. 2) An algorithmic contribution that shows how focusing on global importance enables us to improve computational performance by orders of magnitude compared to state-of-the-art local explanation methods.

Contribution 1: Theory. One of our main contributions is a new theoretical perspective on global feature importance that unites several existing methods (R1): we show that they all make different trade-offs regarding the choice of how to summarize interactions among each feature's contribution to the total predictive power. Revealing this connection allows explicit reasoning about what were previously implicit trade-offs people made when choosing a particular method. To address R1's concerns we will attempt to separate this contribution from the broader conceptual review.

Contribution 2: Computational performance. While our theory contribution strongly motivates Shapley values for global feature importance, Shapley values are known to be expensive to compute. Our second main contribution is an efficient sampling-based estimation algorithm that specifically targets global feature importance. While local Shapley based sampling methods exist, they are orders of magnitude slower than SAGE when applied to global feature importance. We apologize if this was obscured in Figure 1, where we showed the convergence of SAGE for a whole dataset against SHAP for a single instance. In reality, to compute SAGE with SHAP you would need to run it over hundreds or thousands of instances and then average the results (R1). In the updated Figure 1 we will show this more direct comparison (where SHAP is much slower).

Uncertainty and convergence. R1 and R2 asked about uncertainty estimates—Theorem 2 suggests how to calculate confidence intervals, and returning these is now default behavior. R1 asked how to determine the number of samples required—Supplement F describes how to determine convergence automatically using the width of the confidence intervals, which is now the default mode. Thank you for raising these valuable points.

Practical value and metrics. It is widely recognized that measuring feature importance is useful for understanding which features contribute the most information, for generating scientific hypotheses (see our BRCA experiment), and for performing feature selection and feature engineering (see R2). Our experiments demonstrate each of these use cases (R4). Since the first two are harder to evaluate for real datasets (see R2), we focused on unambiguous quantitative metrics. Our *cumulative importance correlation* metric (Figure 1 middle left) is itself a valuable contribution since this field lacks reliable quantitative metrics, and it is the most important evaluation for this problem. By this measure SAGE clearly performs the best overall (see supplementary Table 3). After updating our experiments per R4's request (see below), SAGE wins on 7/8 datasets (R1). Feature ablation beats SAGE on the Wine dataset but performs far worse on the remaining ones.

Models and datasets. Both R2 and R3 appreciated the extensiveness of our experiments, while R4 mentioned some concerns. Our goal was to demonstrate the efficiency of our method and the advantages of applying a game-theoretic approach across a variety of model types and datasets. Using state-of-the-art models, particularly those with heavy feature engineering, is largely orthogonal to our purpose. In response to R4 we should also point out that the Bike Sharing dataset was a Kaggle competition and that we used a GBM for our Credit dataset.

Following R4's recommendation, we re-ran experiments with the Bike Sharing, Bank and Credit datasets using XGBoost, and most methods' importance scores achieved higher correlation with the predictive power. SAGE's advantage over other methods either stayed the same or improved—on the Bank dataset SAGE now outperforms feature ablation by a large margin (average correlation of 0.984 vs. 0.921). Since SAGE outperforms other methods on DNNs, GBMs, SVMs, and random forests individually, we expect SAGE would also outperform other methods on the complex ensembled and stacked model compilations that typically win Kaggle competitions.

Clarifications and questions. The game-theoretic solution is the Shapley value [24] (R3). SAGE is model-agnostic because it works with any model class, unlike some methods listed in Table 1 (R3). The Bayes classifier minimizes the population risk when the loss function is cross entropy; this is widely known, but see our Supplement C.1 for a proof (R3). R2 raised several good questions. Working with simulated datasets is a good suggestion; we considered this but decided it was worth focusing on data with realistic feature interactions. Using a quantile is an interesting suggestion that may have certain advantages, but it would remove the connection with mutual information and change properties 1 and 5 from Section 3.1. We agree that applying SAGE as a regularizer would be a valuable application, but it would be difficult to use out-of-the-box with gradient-based optimization. We also thank R2 for the helpful writing suggestions.