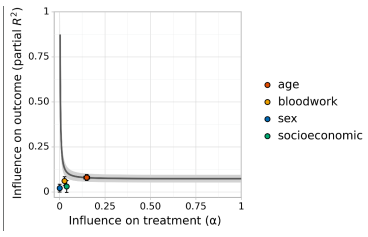
We thank the reviewers for the insightful comments and questions, and for their support. Using the camera ready 9th page, we have added a discussion section, and expanded the related work.

In the new discussion, we emphasize that Austen plots are most useful in situations where the conclusion from the plot would be 'obvious' to a domain expert. For instance, the LaLonde RCT plot shows that a confounder would have to be much stronger than the observed covariates to induce substantial bias. Similarly, the LaLonde observational plot shows that confounders similar to the observed covariates could induce substantial bias. Such conclusions would not be affected by mild perturbations of the dots or the line. By contrast, Austen plots should not be used to draw conclusions such as, "I think a latent confounder could only be 90% as strong as 'age', so there is a small non-zero effect". Such nuanced conclusions might depend on the particular sensitivity model we use, or statistical misestimation or incautious interpretation of the calibration dots—the latter two concerns raised by the reviewers. Drawing precise quantitative (rather than qualitative) conclusions about induced bias from Austen plots would require careful consideration of these issues, and expert statistical guidance. Hence, we recommend that the plots should be used mainly with domain experts to guide qualitative conclusions ("this job program likely works", "this study doesn't establish drug efficacy").

R1 *Cinelli and Hazlet* (CH) warn against 'informal' benchmarking procedures. Our bias calculation is based on a formal bound, so this critique doesn't directly apply. To illustrate the issue, CH provide an example where the change in ATE induced by leaving out an observed covariate X is smaller than the bias induced by omitted variable U, even though X and U come into the model in an identical manner. In contrast, the bias estimate in the Austen plot model is typically higher than the change in ATE induced by leaving out a variable—this is shown in our model conservativism experiments. In particular, Austen plots (correctly) anticipate that the bias from omitting U can be higher than the ATE difference induced by computing without X, even when U and X have identical confounding strength.

The CH example draws attention to a point that requires some care. The crux of it is 1. strength of influence of U is computed *conditional on the observed covariates*, and 2. the reduction in uncertainity in going from only X to both X, U may be greater than the reduction in uncertainity in going from nothing to X. Thus, it's possible in principle for the influence of U to be larger than the estimated (observed) influence of X, even if X and U are similar causal variables. In other words, the omission of U could cause us to underestimate the influence of X from the observed data. Accordingly, when the domain expert compares the strength of the unobserved confounder to a reference dot for X on the plot, they must also ask if knowing U could have substantially increased the predictive power of X. In cases where this seems plausible—e.g., the domain expert expects an important interaction between U and X—then naively eyeballing the dot vs line position may be unreliable and further careful thought is required. However, we note that examples of this kind are somewhat contrived. Indeed, we usually expect the opposite effect. If U and X are dependent, then some of the information in X will be redundant, and the measured R^2 and α will *overestimate* the true influence. That is, the CH effect tends to make our sensitivity analysis conservative. This is reflected by the fact that grouping similar covariates (to mitigate redundancy) led to higher computed influence in every example considered in the paper. Although the CH effect is an important conceptual point and the domain expert should consider it as part of due diligence, it doesn't seem to have much impact on the practical use of Austen plots. We also note that this conceptual subtlety is a generic feature of calibrating sensitivity analyses, not particular to our method. This is reflected, e.g., in Franks et al §5.2.1, which describes their calibration procedure based on looking at variance explained by observed Z conditioned on $X \setminus Z$—a procedure similar to ours, carrying the same nuance. We have clarified this point in the newly added discussion section. We thank the reviewer for bringing this to our attention.

R2 *Inference* We agree that inference is a key issue. We have added some additional detail estimation via the plug-in estimators. For handling uncertainty, we suggest bootstrapping several plot versions and trusting only conclusions supported by all plots. We tried this on the examples in the paper, and found no change in conclusions. We have added an appendix describing how to visually summarize the uncertaintity; see fig (cf. Fig 1 in paper). Note: α uncertainty is plotted, but too small to show up clearly. Formal uncertainity quantification is an interesting direction for future work; particularly important because bootstrapping requires model refitting, which can be computationally intensive for the ML models that motivate the paper.



We emphasize that the unobserved confounder bias we address exists even in the (very) large data regime where plug-in estimators work very well. Indeed, this is the setting where misidentification bias matters most relative to statistical error. Accordingly, we believe the paper, using the plug-in estimators, is a significant contribution even in the absence of efficiency guarantees. In the discussion, we caution that the plots may be misleading if there's large uncertainity in the estimaton of Q and g. As part of the discussion on estimation issues, we note efficient estimation as a good direction for future work.

Related work Thank you for the pointers to the additional related literature. As you suggest, we have used the extra page to substantially expand the related work section.