We thank the reviewers for the detailed feedback! We will be sure to clarify the figures and the text as well as incorporate proof sketches as suggested. In the interest of space, we group and address the main concerns below:

Reviewer 1:

"[on] choosing the order of the model" + **Reviewer 2's** *"[what is a] systematic way for the practitioners to understand which order CGAs are suitable to their application context"* — The order k of the CGA model is dependent on the application at hand. It may be set to be the maximum k-way interaction that the practitioner expects to take place in the team. Our model is especially useful in settings like the NBA, in which team sizes are small relative to the overall number of players. This structural prior can be encoded in the order of the CGA. As an example, for the NBA, we expect at most 5-way interaction and so it is necessary to only consider compact, low-rank models.

Reviewer 2 (please see the above response to Reviewer 1 also):

"time and space complexities" — The time to compute the SV is the space complexity of the CGA model: the number of parameters. The complexity of learning a CGA depends on the training method employed to learn $\hat{v}$ (e.g. we use SGD).

"What happens if we increase the team size to a reasonable value such as 100 players? Does the proposed method scale " — We believe our method would scale well with the team size, and do even better if the data exhibits low-rank structure. Note that we considered doing this larger-scale experiment, but were limited by the speed of current MARL methods. Training $\binom{12}{3}$ teams already takes 4 days, so training $\binom{N}{100}$ teams for N much larger may well take on orders of months.

Reviewer 3:

"[How are] results proposed in this paper are related to existing work" + **Reviewer 4's** *"missing some important references...[thus] the novelty of this work may be limited* — Broadly, there has been ample, but more distantly related work that fall into the following three categories with which we characterize all the papers mentioned:

1) Fast methods to compute the SV (usually for certain subclasses of games): Fatima et al (only for voting games).

2) Sample complexity of approximating the SV: Maleki et al; Bachrach et al (only for simple games); Liben-Nowell et al (only for supermodular games).

3) Representation designed to allow for easy computation of the SV: Michalak et al (only for networks); Conitzer et al '04 (the '06 paper focuses on the core); Shoham et al (already cited).

To the best of our knowledge, and as also noted by Reviewer 1, all prior work in these three categories assume full knowledge of the CF. By contrast, our central premise is – **the CF is unknown and needs to be learnt from data**. We describe how our work differs from each of the three categories below and will be sure to append this to the paper:

1) our method holds for all cooperative games, to answer the question on *"assumptions you make regarding the game"*.

2) our bounds are on the sample complexity of learning the CF function with the CGA representation, thus drawing upon PAC/PMAC techniques. In contrast, work in (2) focus on the concentration of estimated SV values via standard concentration inequalities like Chebyshev/Hoeffding, and do not need to nor use learning theoretic methods. Thus, we respectfully disagree with that *"the techniques in these are similar to what is proposed in this paper"*.

3) we propose a representation that not only allows for fast SV computation, but also has provably good learning theoretic properties (crucial since we use it to learn the unknown CF from data). CGA is the first model to met not only the first desideratum (like work in (3)), but the second as well. We believe we are the **first to design a representation with learning from samples in mind**. Thus, we think there is ample and *"not limited"* novelty to our work.

"[how are]... teams in multiagent reinforcement learning... characterized" — We actually perform exactly this empirical evaluation in a large-scale MARL setting, where we exhaustively evaluate all counterfactual teams to characterize what a CF would look like. Please note also that we evaluate on two large scale settings and not one, as you stated.

Reviewer 4 (please see the first response to Reviewer 3 also):

"Generalisation to k-fold interactions seems not relevant in practice (according to their own experiments), but k=2 case was already known." — To clarify, while there has been work that experiments with order $k = 2$ models, no prior work has considered nor compared it with higher orders models. So it is not "already known" if second order models are indeed the best or offer the best tradeoff. Our experiments are the first to do precisely this comparison. Also, we found that the third order CGA doe perform better, but not to the extent that is worth the space and runtime tradeoff of $O(n^3)$ vs $O(n^2)$. And so, we think our comprehensive comparison and its conclusion is a novel contribution.

"missing ... ranking/skill estimation/matchmaking [models like] e.g., TrueSkill " — Note that all these models fall under the umbrella of the CGA family; it is a complete representation (Fact 1). For instance, TrueSkill is a first order CGA, which we include in our experiments, since the team performance is set to be the weighted sum of its players' skills.