Collaborative Problem Solving in an Open-Ended Scientific Discovery Game

AARON BAUER and ZORAN POPOVIĆ, University of Washington

Countless human pursuits depend upon collaborative problem solving, especially in complex, open-ended domains. As part of the growing technological support for such collaboration, an opportunity exists to design systems that actively guide and facilitate collaborative problem solving toward the most productive outcomes. A better understanding of the dynamics of open-ended collaboration on complex problems is needed to realize this opportunity. Motivated by this need for better understanding, we investigate the collaborative problem solving ecosystem of the scientific-discovery game *Foldit*. Our investigation is guided by two primary questions: how do the social aspects of *Foldit* impact an individual's behavior? and what factors have significant impact on group success? We find that collaboration and competition are associated with increased participation and that collaboration increases individual performance. We also find that measures of group skill, individual skill, and participation correlate with better group performance.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; • **Applied computing →** *Computer games*; Life and medical sciences;

Additional Key Words and Phrases: problem solving; online collaboration; scientific discovery games

ACM Reference Format:
Aaron Bauer and Zoran Popović. 2017. Collaborative Problem Solving in an Open-Ended Scientific Discovery Game. Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 22 (November 2017), 21 pages. https://doi.org/10.1145/3134657

1 INTRODUCTION

Collaborative problem solving is an integral part of many important tasks, from scientific discovery [34] to management [11]. The increasing prevalence of technologically-mediated collaboration provides an unprecedented opportunity to expand the scale, speed, and sophistication of collaboration on difficult problems. Progress in this direction could take many forms. Expanded scale and sophistication could arise from optimizing the design of collaborative mechanisms, such as how solvers are engaged in the problem or how problems are posed, and collaborative structures, such as how work is divided and shared. Creating layers of machine intelligence to schedule group work and dynamically adapt environmental parameters such as team size could achieve increased speed and solution quality, as has been applied to task routing in Wikipedia [5].

This opportunity is especially promising in creative, open-ended domains where good solutions are not known. It is not clear if existing models of collaboration (e.g., [1, 14]) apply to these complex

This work was supported by the National Science Foundation under Grant No.: DRL-1639576 and Grant No.: DGE-1546510 and by the Oak Foundation under Grant No.: 16-644. This work was also supported by the National Institutes of Health grant 1UH2CA203780, RosettaCommons, and Amazon. This material is based upon work supported by the National Science Foundation under Grant No. 1629879.

2573-0142/2017/11-ART22 $15.00
https://doi.org/10.1145/3134657

spaces at scale. Hence, a deep understanding of the entire collaborative problem-solving ecosystem in open-ended domains is crucial if new systems are to meet their potential. This understanding must contend with both individuals and groups, and account for instances of both success and failure.

We contribute to the development of this understanding by investigating open-ended collaborative problem solving in the scientific-discovery game *Foldit*. By modeling the functions of proteins, the workhorses of living cells, *Foldit* challenges its users to resolve the shape of proteins as a 3D puzzle. In addition to solving puzzles individually, users can join together in groups, sharing their own work and building off solutions from other group members. These puzzles are completely open and often under-specified, sharing many of the properties Jonassen attributes to *design problems* [15]. *Foldit* puzzles provide a *vague goal with few constraints* (i.e., find a good configuration of the protein), *answers that are neither right or wrong, only better or worse*, and *limited feedback* (i.e., real-time feedback is limited to a single numerical score corresponding to the protein's current energy state, and solvers must frequently progress through many low-scoring configurations to reach a good solution).

Developing users from novices to experts capable of overcoming these difficulties and solving protein structures currently unsolved by scientists is central to *Foldit*'s scientific-discovery community. Solutions generated by *Foldit* users have led to three results published in the journal *Nature* [4, 9, 17]. The ill-structured nature of the problems it poses and its objective of state-of-the-art biochemistry results make *Foldit* a highly suitable setting in which to study collaboration on real-world problems.

The structure of the collaboration itself also incorporates meaningful aspects of real-world settings. Collaboration in *Foldit* is flexible and endogenous. Users control how and when they collaborate, where they focus their efforts, and even whether they participate in collaboration at all. Groups are user-driven and their organization is ad-hoc rather than imposed top-down. This collaboration also occurs at multiple scales and across multiple channels. Users collaborate on individual puzzles, but may also remain an active member of a single group for years. *Foldit* facilitates direct sharing of solutions, but users also collaborate via text chat, screenshots, forums, and shared macro scripts called *recipes*.

In this work, our analysis is guided by two primary questions: (1) how do the social aspects of Foldit impact an individual's behavior? and (2) what factors have significant impact on group success? The first question motivates our investigation of the effects of early collaboration and early competitive success on a user's continued participation in *Foldit* and the impact joining a group has on individual performance. We find that early collaboration and competitive success are each associated with increased participation, and that joining a group leads to increased individual performance. In service of the second question, we explore how features of group activity at difference scales correlate with group performance. We find that features measuring group collaborative skill and individual group member skill correlate strongly with group performance, that participation has moderate correlation with performance, and that group and individual experience only correlate weakly with group performance.

2 RELATED WORK

Citizen science environments like *Foldit* have long been a fertile subject for study [13]. As these efforts have moved online, their contributions have spanned many fields including biology [29], environmental studies [18], and astronomy [22]. In their typology of citizen science, Wiggins and Crowston identify *virtual* projects as an important emerging type with high capacity for motivating continued participation in scientific research [32]. This, along with the scale at which virtual citizen science efforts can operate, has made them useful settings for researchers interested in understanding the dynamics of these collaborative spaces. For example, Rotman et al. studied the motivations that led both scientists and volunteers to engage in collaborative projects together [25]. They highlight the importance of feedback mechanisms in sustaining participation, such as the timing of motivational probes and the availability of information about how the data generated by the volunteers will be used. Though *Foldit* is the source of data for our analyses, our work does not focus on the dynamics of citizen science per se. Instead, the properties of the *Foldit* environment make highly suitable for studying a complex collaborative problem-solving ecosystem, and we are motivated by questions applicable to many real-world collaborative problem-solving settings.

As befits a task of such ubiquitous importance, collaborative problem solving has been long studied in a variety of ways. It has received particular attention for its role in learning, and numerous studies have shown how it can benefit learning [31]. Researchers have also conducted fine-grained, in-person, small-scale studies to explore the specific mechanisms at work [24]. More recently, the learning benefits of collaboration have been shown to extend to computer-mediated settings as well [26]. Our finding that group membership improves individual performance is consistent with this existing body of work. We show the benefits of collaboration extend to solving open-ended problems in an environment with user-driven ad hoc collaboration. Our result raises important questions about the mechanisms behind the observed benefits.

In addition to the educational lens, collaborative problem solving has been studied in terms of its efficacy for generating solutions. Early work found that group decision-making could exacerbate individual biases and that group discussion was plagued by inefficiency [16]. Other work, such as that by Stasser and Titus [28], highlights the way carefully structure group interactions can alleviate this friction. As prior results have demonstrated the ability of *Foldit* solvers to resolve previously unanswered scientific problems, we focus on understanding what contributes to success on *Foldit* puzzles, rather than on evaluating the efficacy of *Foldit* itself. Some work, such as Settles and Dow's study of online songwriting collaboration [27], has investigated the factors contributing to successful ad-hoc collaboration in an online setting, though not in the scientific-discovery domain.

Similar to our work in its subject of study, Tuasczik et al. investigated collaboration on real-world open-ended problems in *MathOverflow*, an online environment for solving novel mathematical problems [30]. The authors identified a set of collaborative acts users displayed and used regression models to assess their impact on solution quality. Cranshaw and Kittur use a combination of data analysis and visualization similar to our approach in this work to study the principles behind the success of the *Polymath Project* [7]. Our analysis focuses on the factors affecting the performance of groups and individuals rather than the identification and impact of specific actions or the principles that have led to the success of *Foldit.* Finally, collaborative problem solving has been studied in the context of evaluating systems designed to facilitate novel mechanisms of collaboration. For example, CoSolve [10] allows users to both pose and solve problems, visually representing the solving process as state-space search trees, The Climate CoLab [12] combines model-based planning, online debates, and electronic voting to enable collaborative development of plans to address climate change, and CrowdForge [21] presents a general framework for crowdsourcing complex tasks such as article writing.

A related, but distinct line of research deals with quantifying team performance and identifying the major factors involved. Much of this work has focused on developing frameworks for successful collaboration (e.g., [1, 14]). These frameworks deal with in-person and relatively small-scale settings, making analysis like ours a necessary step in developing frameworks that extend to the new collaborative environments made possible by recent technological advances. Other work leverages the concept of *collective intelligence* to explain team performance in domains such as competitive online games [19], as well as a variety of other tasks [33]. While measuring the collective intelligence of *Foldit* teams could certainly yield interesting predictive results, doing so would require soliciting the participation of *Foldit* users in additional data-gathering activities, and is beyond the scope of this work.

3 FOLDIT

Foldit is a scientific-discovery game that tasks human *solvers* with manipulating a 3D representation of a protein into the minimum energy configuration possible. Proteins are posed to solvers as *puzzles*, and solutions are evaluated in real-time according to their energy configuration. *Foldit* crowdsources protein structures by having solvers compete to produce the highest-scoring solutions.

In addition to generating solutions individually, *Foldit* solvers can join together in groups to tackle puzzles collectively. While solvers can communicate via a variety of typical online tools such as text chat and image sharing, *Foldit* provides an explicit mechanism for collaboration called *evolving*. Whenever a member of a group generates an individual solution (a *soloist* solution in *Foldit* parlance), they can choose to share it with the other members of their group. Those other members can then import this soloist solution directly into their client, and attempt to modify and improve it. If a teammate successfully improves on the soloist effort, the new, improved solution is recorded as an *evolver* solution. An evolver solution can in turn be evolved just like soloist solutions. A group's official solution for a puzzle is the highest scoring solution, soloist or evolver, produced by any member of that group. Prior work studying group behavior in *Foldit* has focused on the sharing of automated macro scripts called *recipes* [3], though there has been limited discussion of group solving dynamics [4].

Foldit puzzles fall into two main types: (1) *prediction* puzzles, the more common type of puzzle, in which the amino acids that make up the protein are known, but the way the protein folds up in 3D space is not know, and (2) *design* puzzles where solvers can modify which amino acids compose the protein in order to create a protein that fulfills a specific scientific purpose such as targeting key molecules in diseases. In this work we limit our analysis to prediction puzzles to avoid differences in puzzle type as a confounding factor.

4 BACKGROUND

The properties that make the *Foldit* ecosystem an attractive environment for the investigation of open-ended collaborative problem solving also present significant challenges. First, the variety of channels over which collaboration can occur means that only a portion of the collaborative activity is directly observable in the data *Foldit* makes available. Specifically, we can observe the solutions shared via the *Foldit* client itself and through this understand who is contributing to a group effort and whether their contribution consists of a de novo effort or directly builds on work by another group member. The forums hosted by the *Foldit* project and the recipes shared among groups are also observable, though we do not analyze these channels in this work. We cannot observe the sharing of ideas or other collaboration happening over other channels such as text chat. Due to this limited picture, fine-grained analysis of collaborative acts and problem-solving strategies requires a detailed look at low-level problem-solving behavior in order to infer the larger patterns at work. Hence, this work looks very broadly, focusing on the aggregate trends and correlations, with the goal of guiding future, more targeted analysis toward the most salient phenomena in need of deeper explanation.

A second challenge is the presence of many confounding variables. As *Foldit* is an active problem-solving community, randomized controlled trials and other experimentally controlled scenarios are absent from the data on its solvers' behavior. Hence, in assessing the impact and importance of various factors, establishing simple lines of cause and effect is frequently infeasible. We tackle this challenge in several ways. Throughout our analysis, we employ a combination of data analysis and visualization to illustrate the relevant trends and pair this with a broad discussion of the potential factors at work. This enables us to work toward an understanding of collaborative problem-solving in *Foldit* within the limitations imposed by the data. In the case of quantifying the effect of group membership on individual performance, the problem of confounding variables is particularly acute and we take the additional measure of carefully constructing a simulated controlled experiment. We identify a suitable *synthetic treatment* population (i.e., solvers who joined a group some time after they began playing *Foldit*), and pair each member of that population with the most similar member of the *synthetic control* population (i.e., solvers who never joined a group) in order to minimize the possibility that differences between the populations beyond group-joining itself are responsible for effects we observe.

A third challenge is one fundamental to the nature of *Foldit* puzzles. The lack of a metric for solution quality comparable across puzzles makes it difficult to quantify solver and group success over time. The only absolute metric of solution quality is the score computed from the protein's current configuration, but since this metric is contingent on various structural properties of the protein, there is no reliable baseline to use to compare solutions dealing different proteins. Interpretation of the best scores, as well as the differences in score among a set of solutions cannot easily generalize across multiple puzzles. Hence, a *relative* rather than absolute metric is the natural choice for tracking success in *Foldit* over time.

An obvious relative metric, a simple ranking of solutions by score, fails to capture any notion of puzzle difficulty or the degree to which top-scoring solutions exceed the competition. That is, if a puzzle is relatively easy, and there are dozens of similarly-scoring solutions, a ranking-based metric will treat this situation the same way it would treat a puzzle where one or two solutions dramatically outscore everything else. Since we are interested in quantifying success over time, we can afford to use a metric that dispenses with fine-grained distinctions within a single puzzle in favor of a metric that better accounts for difficulty and margin of success. Motivated by these considerations, in this analysis we use *the ratio of a solution's score to the score of the best solution* as our measure of performance. In other words, the performance of a solver or group on a given puzzle is measured as the ratio of their score to the best soloist or group solution, respectively. This metric accounts for the failings of ranking: if all the top solutions have similar scores, they will all have very similar performance; if one solution significantly exceeds the rest, our metric for performance will reflect this. We do not claim that this metric is the optimal choice for every analysis of performance in *Foldit*, but it is well-suited to the questions we seek to address in this work.

In working to overcome these challenges, we contribute a metric accounting for puzzle difficulty and margin of success as well as enabling post-hoc analysis that accounts for several important factors that were not controlled in the original data. We use these approaches to investigate individual and group behavior in the complex, open-ended domain of *Foldit* puzzles. We gain insight into the effects of collaboration and competition on individual behavior and the factors involved in group success, and discuss the consequences of these findings for the future design and augmentation of collaborative problem-solving systems.

5 EFFECTS OF EARLY COLLABORATION AND COMPETITIVE SUCCESS ON PARTICIPATION

In volunteer-based problem-solving communities such as *Foldit*'s, keeping solvers engaged is crucial to the success and longevity of the project. To contribute solutions to complex, open-ended problems, solvers must develop significant expertise, and thus must participate long enough to accomplish this. Furthermore, important social mechanisms such as collaboration and competition, both of which feature prominently in *Foldit*, require a critical mass of users to function effectively. Hence, understanding key drivers of long-term participation is important for the design of collaborative problem-solving systems. To this end, we assess how interaction with both collaboration and competition affects participation in *Foldit*.

5.1 Method

To address this question, we analyzed the data from all *Foldit* solvers across the 681 prediction puzzles released since early 2011 (puzzles released before then were not categorized). This dataset consists of 26,048 solvers who contributed 179,723 solutions. Since a solver's initial experiences in *Foldit* must necessarily play a significant role in their choice of whether to continue participating, we group the population in our dataset according to the presence of two experiences in each solver's first five puzzles.

Specifically, we consider a solver to have experienced *early success* if in any of their first five puzzles, their soloist solution ranked in the top 25 soloist solutions. We select the top 25 solutions as a threshold because that is how many solvers are shown on the first page of ranked soloist solutions for each puzzle on the *Foldit* website, and ranking is the primary mechanism for social recognition in *Foldit*. In other words, we consider a solver to have experienced early success if they can see themselves on the first page of ranked soloist solutions for any of their first five puzzles. We consider a solver to have experienced *early collaboration* if in any of their first five puzzles, they participated as a member of a group. *Foldit* has 3–7 puzzles available for solvers to contribute to at a time (individual puzzles expire and get replaced on a timeline of 1–2 weeks), so a solver's first five puzzles reasonably approximates the content available at the time they begin participating. These criteria give us the four non-overlapping classes listed in Table 1.

In terms of quantifying participation, the most relevant measure is also the most straightforward: the number of puzzles a solver contributed to. We compare the number of puzzles contributed to by solvers in each of the four classes.

5.2 Results

As Figure 1 shows, both early collaboration and early success are associated with increased participation. Each curve visualizes the rate at which solvers in a given class stop contributing to *Foldit*. For example, almost 40% of those who had both early success and early collaboration contributed to at least 100 puzzles, while only 29% of those with only early success did so. We use a Mann-Whitney U test for non-normally-distributed data to test significance and rank-biserial correlation coefficient (r) to measure effect size of the differences between classes (we use multiple Mann-Whitney U tests instead of a Kruskal-Wallis test in order to understand the magnitude of the effects). Specifically, we compare each class to the next best class in terms of participation (i.e., the class with early success and early collaboration is compared to the class with early success only; the class with early success only is compared to the class with early collaboration only, and so on). Summary statistics for each class and the results of our statistical comparisons are given in Table 1.

	solvers	mean puzzles	median puzzles	U	effect size r
Early success, collaboration	78	149.5	53	3294.5	0.188^{*}
Early success only	104	94.8	21	30932	0.649^{***}
Early collaboration only	1697	20.4	2	13621219	0.336^{***}
No early success, collaboration	24169	5.1	1	—	—

Table 1. The summary statistics for each class and results of statistical comparisons between classes. The U and r given for each class are the test statistic and effect size, respectively, of a Mann-Whitney U test comparing that class with the class on the next row (hence why the last row omits these). The p values for the Mann-Whitney test are indicated by: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

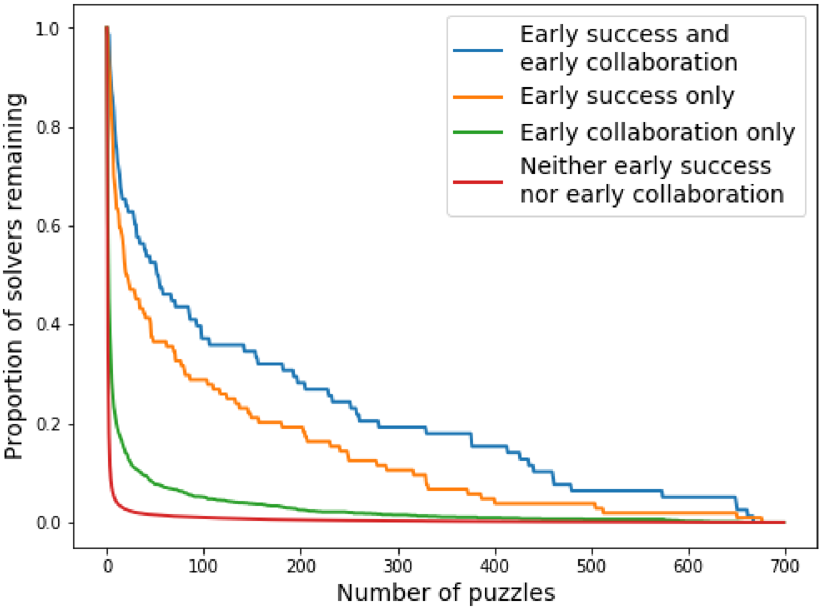

Fig. 1. The effects of early collaboration and success on continued participation. For a given number of puzzles on the x-axis, each curve indicates the proportion of solvers in that class who contributed to at least that many puzzles. This figure shows an association between increased participation and early success and early collaboration, with early success having the greater association. It is notable that early collaboration is associated with increased participation for solvers with and without early success.

5.3 Discussion

While early success was associated with greater participation than early collaboration, it is notable that early collaboration is associated with significant increases in participation for solvers independent of whether they experience early success. These results are consistent with existing research on participation in online communities, in particular the findings that recognition of user contributions (e.g., early success) and emphasis of social context (e.g., early collaboration) can increase participation [23].

It is also worth noting that though solvers with early success contribute to many more puzzles on average, they still stop contributing at an appreciable rate. In some sense, *Foldit* is failing to capture the talent of *all* its promising new solvers. It may be possible that well-designed and well-timed feedback could convert more of these solvers to long-term contributors. From this perspective, early success and collaboration could serve as indicators for identifying users with greater potential to become long-term contributors, and help guide attention from facilitators or the system itself toward integrating them into the community.

Though we identify significant differences in participation between groups that experience early success and collaboration and those that do not, we cannot establish clearly delineated cause and effect. It is possible that those who early on chose to join a group or put in the work to get a high score are already disposed to greater participation. Regardless of the dynamics at work, collaboration and competition can clearly provide experiences or opportunities with potentially long-term effects on the participation of new members. Designers of collaborative problem-solving systems interested in increasing participation should explore ways of both increasing the prevalence of these experiences and integrating them more tightly into the fabric of their systems.

6 EFFECT OF GROUP MEMBERSHIP ON INDIVIDUAL PERFORMANCE

The literature on learning and collaboration makes clear that collaboration can have significant learning benefits in both traditional and computer-mediated settings [26, 31]. It is not clear, however, if this effect extends to open-ended problem-solving domains such as *Foldit*. In *Foldit*, as in many other settings, developing solver skill is essential, and quantifying the role of collaboration in this process is necessary to construct a comprehensive understanding of the problem-solving ecosystem. Furthermore, characterizing the impact of group membership on individual performance in detail could serve as a guide to future analysis of the specific mechanisms at work.

6.1 Method

Given the number of confounding factors surrounding group membership in *Foldit*, isolating its effect on individual performance requires a carefully designed comparison. Ideally, we would have two randomly assigned, otherwise identical subsets of *Foldit* solvers where the members of one subset joined a group and members of the other subset did not. Since our data comes from an active scientific-discovery game rather than a controlled lab setting, such an ideal scenario does not exist. Hence, we construct two subsets of solvers from the available data in such a way as to control confounding factors.

In particular, we construct a *synthetic control* sample that never joins a group and a *synthetic treatment* sample that does. To measure the effect of group membership on individual performance, we compare how the performance of members of each sample develops before and after those in the synthetic treatment sample joined a group. We construct our two samples as follows. First, we identify the subset of solvers for the synthetic treatment sample who will support a robust comparison of their performance before and after they first join a group. Specifically, we select solvers who began not part of any group, who contributed to more than 30 puzzles and at least 10% of their total puzzles before they joined a group, and who contributed to at least as many puzzles after first joining a group as they did before joining any group. These criteria ensure a sufficient demonstration of each solver's performance before and after they joined a group. There are 92 solvers that meet our criteria.

In order to ensure the validity of comparisons between the synthetic treatment and control samples, we control for two potentially confounding variables in constructing the synthetic control sample. In particular, we construct the synthetic control sample by pairing each solver in the synthetic treatment sample with a solver that never joined a group minimizing the differences between each pair of solvers along two dimensions. First, we minimize the difference in median performance before the synthetic treatment sample solver joined a group, measuring performance as the ratio of the solver's solution score to the score of the best soloist solution. Minimizing performance differences before treatment occurs helps ensure that any differences that emerge after the treatment can be attributed to the treatment itself. Second, we minimize the difference in the total number of puzzles each solver contributed to. Here we use overall participation as a proxy to control for differences in overall engagement. As Figure 2 and Figure 3 show, this process results in very similar distributions of pre-treatment performance and overall participation for our synthetic treatment and control samples. Solvers in our synthetic treatment sample are diverse in terms of when they joined a group, as shown in Figure 4

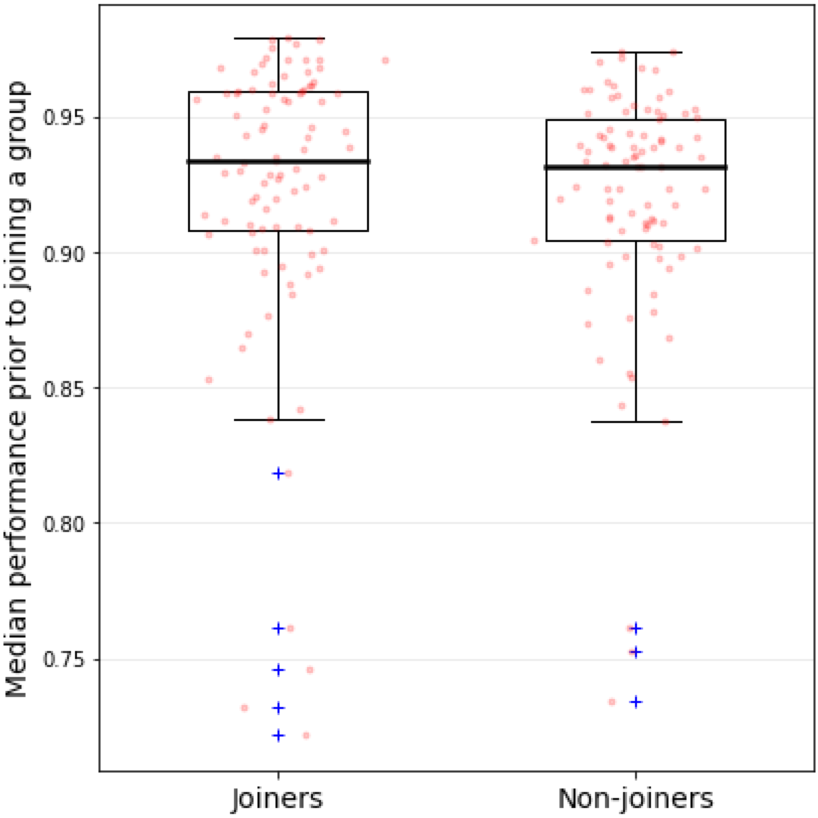

Fig. 2. The distribution of median performance for our synthetic treatment (i.e., group-joining) and control (i.e., non-group-joining) samples before treatment solvers joined a group. Red circles indicate individual solvers.

6.2 Results

To measure the effect of group membership on individual performance, we compare both the number of solvers who experienced an improvement in performance and the magnitude of the improvement in each of our two samples after treatment occurs. To compute the magnitude of improvement, for each pair of solvers, we measure their median performance over the same number of puzzles following the treatment as the number they contributed to prior to the treatment. For example, if a solver in the synthetic treatment sample contributed to 40 puzzles before joining a group, we measure their median performance and the median performance of the synthetic control sample solver they were paired with over their first 40 puzzles and then over their next 40 puzzles (i.e., their 41st through 80th puzzles). Then, we compute the difference between each solvers' median performance after the point of treatment and their median performance before treatment.

Comparing the number of solvers who experienced an improvement in performance in each sample (i.e., the solver had higher median performance after the point of treatment than before), more solvers improved in the synthetic treatment condition (58) than in the synthetic control condition (40). Pearson's χ^2 test indicates this can be attributed to a significant difference between

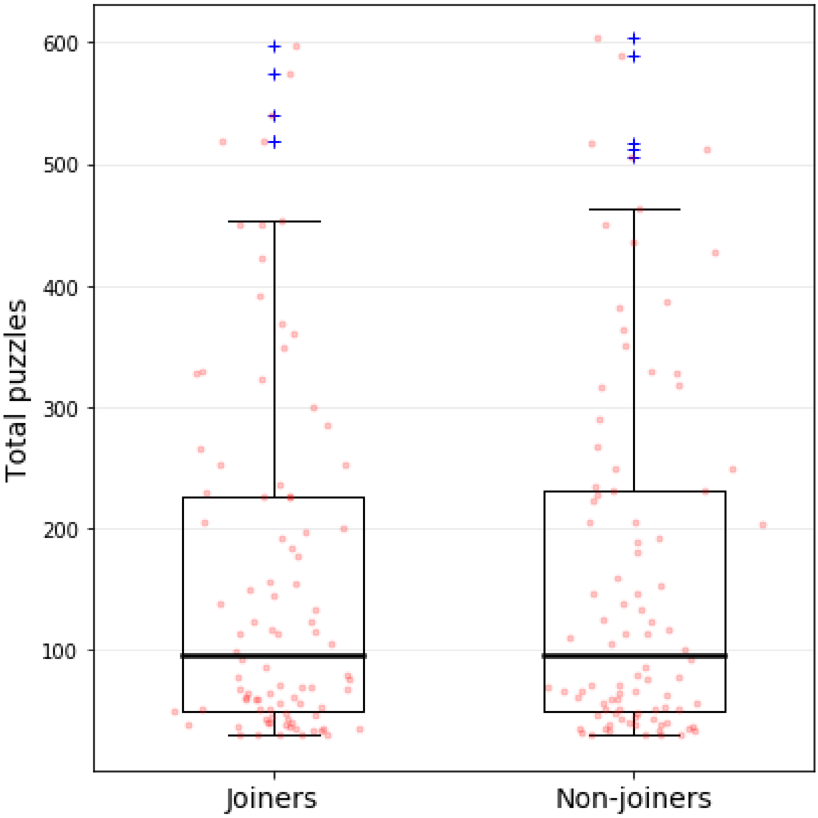

Fig. 3. The distribution of total puzzles contributed to for our synthetic treatment (i.e., group-joining) and control (i.e., non-group-joining) samples. Red circles indicate individual solvers.

the two conditions ($\chi^2(1, N = 184) = 6.31, p = 0.012$). Using a Mann-Whitney U test for non-normally-distributed data to test significance and rank-biserial correlation coefficient (r) to measure effect size, we find that among solvers who experienced improvement, those in the synthetic treatment sample improve their performance more than those in the control sample, but that this difference is not statistically significant. The 58 synthetic treatment solvers who improved have a mean increase in performance of 0.036 (median of 0.025) compared to a mean increase in performance of 0.023 (median of 0.018) for the 40 synthetic control solvers who improved ($U = 906$, $p = 0.067$, $r = 0.219$).

6.3 Discussion

The finding that collaboration improves individual performance in *Foldit* is not surprising, but serves as useful confirmation that this effect extends to ad-hoc collaboration on complex, open-ended problems in an entirely virtual environment. In particular, improvement was more widespread among group members than among non-group members. This finding motivates questions for deeper analysis into what about group membership in *Foldit* is contributing to the observed benefits. Any number of dynamics could be at work, including mentorship, exposure to new techniques and ideas, access to recipes shared with group members, or increased effort due to social motivation.

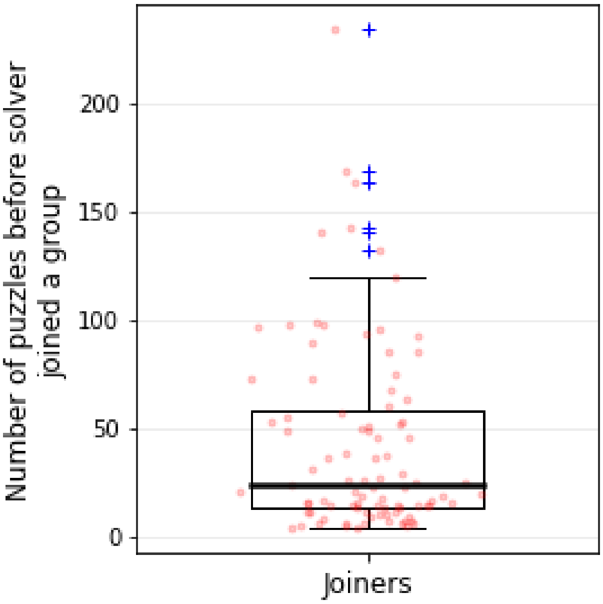

Fig. 4. The distribution of how many puzzles solvers in our synthetic treatment sample contributed to before joining a group. Red circles indicate individual solvers.

Investigating these dynamics is important for developing a comprehensive understanding of the collaborative problem-solving ecosystem, and could provide useful insights for designers of collaborative problem-solving systems.

An interesting aspect of these results is that many solvers in both samples had decreased performance after the point of treatment. This is not surprising as our metric for performance is relative rather than absolute, meaning if the solver community improves as a whole over time, overall performance should appear roughly the same. Nevertheless, this lack of improvement raises the question about what intervention apart from group membership might give these solvers the kind of boost experienced by the majority of those that joined a group.

Though we have taken steps to minimize the effect of confounding variables on our results, it is not possible to guarantee their absence. Regardless of any precautions, we are still necessarily comparing solvers who chose to join a group to those who did not, making our results vulnerable to selection bias. Due to our method of constructing our samples, however, any confounding factors that do play a role are not evident in solver performance pre-treatment or in overall participation. Furthermore, we measure improvement over the same number of puzzles for both conditions, so we control for increase in engagement associated with joining a group in terms of its effect on participation. Hence, we have confidence that group membership plays a significant role in the observed effect.

7 EXPLORATION OF GROUP PERFORMANCE

As in so much of real-world problem solving, group performance is a key driver of high quality solutions in *Foldit*. Developing a better understanding of what contributes to group success and failure could play a vital role in improving outcomes. This understanding is a necessary foundation for designing systems that structure group work in ways that enable more effective collaboration, or implementing layers of machine intelligence to schedule individual efforts in the most effective combination.

In this section we explore the relationship between group performance and a variety of factors. While the relationships we explore are correlational rather than explanatory, they provide a lay of the land, and, more importantly, generate questions that can guide future investigations. A deep understanding of group performance in *Foldit* will require fine-grained analysis of collaborative acts and the specific collaborative problem-solving behavior involved in high-quality solutions. While this fine-grained approach is beyond the scope of this work, our exploration offers a necessary foundation for future, more narrow dives into the dynamics behind what we observe.

7.1 Method

To conduct this exploration, we use data on group contributions on all 681 prediction puzzles released since early 2011. Of the 451 groups that participated in at least one of these puzzles, we restrict our analysis to the 66 groups that participated in at least five puzzles with at least two group members participating in each puzzle. These 66 groups contributed 13,471 solutions (the members of these groups contributed far more solutions than this, but *Foldit* considers a group's solution to be the best soloist or evolver solution contributed by one of its members).

Given our interest in collaborative problem-solving performance, it is natural to exclude those groups whose data does not demonstrate meaningful opportunity for collaboration. We ignore both ephemeral, short-lived groups (fewer than five puzzles) and groups clearly not collaborative in nature (fewer than five puzzles with multiple group members). The choice of five puzzles as a threshold is motivated by the release schedule of *Foldit* puzzles. Between three and seven puzzles are available at any given time, and a puzzle typically expires after 1–2 weeks, and is replaced with a new puzzle. Thus, a threshold of five puzzles nicely approximates a minimum engagement with the state of the game at a given point in time.

With this dataset, we explore group performance at both macro- and meso-scale. Specifically, we explore how features of a group's overall tenure correlate with overall group performance and how features of a group's effort on an individual puzzle correlate with performance on that puzzle. In both cases we measure performance as the ratio of a group's solution score to the score of the best group solution. As part of exploring overall performance, we also examine how these features differ between *high-performing* groups and other groups in our dataset. Given that a nuanced identification of high-performing groups would rely on the kind of deep understanding of group performance motivating our current exploration, we use a very simple threshold: we designate a group as *high-performing* if it contributed the single best solution (i.e., was ranked first) on at least one puzzle. Under this threshold, 15 of the 66 groups are considered high-performing. We view this threshold as providing a good upper bound on the number of high-performing groups — it is very unlikely that any group making a substantive collaborative contribution through *Foldit* has *never* contributed a puzzle's highest-scoring solution. In the context of an initial exploration, we view a broad, inclusive threshold as preferable to a narrow, overly-restrictive one.

Macro-scale features. We explore how the following features of a group's entire solving history relate to overall group performance. We apply these to the 66 groups in our dataset.

Group experience: the total number of puzzles the group has contributed to. Groups may improve their collaboration over time, which in turn may increase their performance.

Collaborative skill: the proportion of puzzles for which the group's solution was contributed by an evolver. Collaboration in *Foldit* consists of an ad-hoc back-and-forth as group members view, borrow from, and improve upon (i.e., evolve) each others' solutions, so no single feature will completely capture this complex process. As this feature measures the frequency with which a group's solution definitively represents an effort beyond what any individual member contributed alone, we hypothesize it approximates part of a group's overall ability to collaborate productively.

Group participation: the median number of group members participating on each puzzle over the group's entire history. Without sufficient participation, groups may be unable to benefit from collaboration.

Meso-scale features. We explore how the following features of a group's effort on an individual puzzle relate to group performance on that puzzle. We apply these to the 13,471 group contributions in our dataset.

Individual experience: the number of puzzles contributed to by the most experienced participating group member (only counting puzzles prior to the one in question). Since a group's solution is the best among those produced by all its members, we similarly use the most experienced participating group member as a measure of the experience the group brought to bear on a particular puzzle.

Individual skill: the median soloist performance of the best-performing participating group member (across all puzzles that group member has contributed to up until this point). As with the previous feature, the nature of group solutions in *Foldit* motivates our considering only the best-performing participating group member.

Soloist participation: the number of group members contributing as soloists. More soloists may provide a group with greater diversity of ideas and increased ability to pursue multiple approaches to a puzzle.

Evolver participation: the number of group members contributing as evolvers. More evolvers may enable a group to better or more quickly refine its solutions.

7.2 Results

We first describe our exploration of our three macro-scale features. For each feature, we plot the value of that feature for each group versus that group's median performance across all puzzles it contributed to, using color to distinguish top groups from other groups. Group experience is shown in Figure 5, collaborative skill is shown in Figure 6, and group participation is shown in Figure 7. In addition, we compute Spearman's rank correlation coefficient (Spearman's ρ) for each feature to measure its correlation with group performance, doing so separately for top groups and other groups. Finally, we perform for each feature a Mann-Whitney U test to determine if the difference between top groups and other groups is statistically significant, and compute the rank-biserial correlation coefficient r to measure the effect size for this test. We use these statistical measures as they are non-parametric and thus appropriate for non-normally-distributed data. These values are given in Table 2.

	Top groups		Other groups			
	median	Spearman's ρ	median	Spearman's ρ	U	effect size r
Group experience	574	0.517*	71	0.212	69	0.820***
Collaborative skill	0.288	0.743***	0.000	0.586***	14	0.963***
Group participation	6	0.706**	1	0.140	81.5	0.787***

Table 2. The statistical results of our exploration of macro-scale features are given here. For each feature, we list the median value and Spearman's rank correlation coefficient (ρ) for both top groups and other groups. We also list the test statistic U and rank-biserial correlation measure of effect size r for a Mann-Whitney U test of the difference between top groups and other groups. The p values for the correlation and Mann-Whitney test are indicated by: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

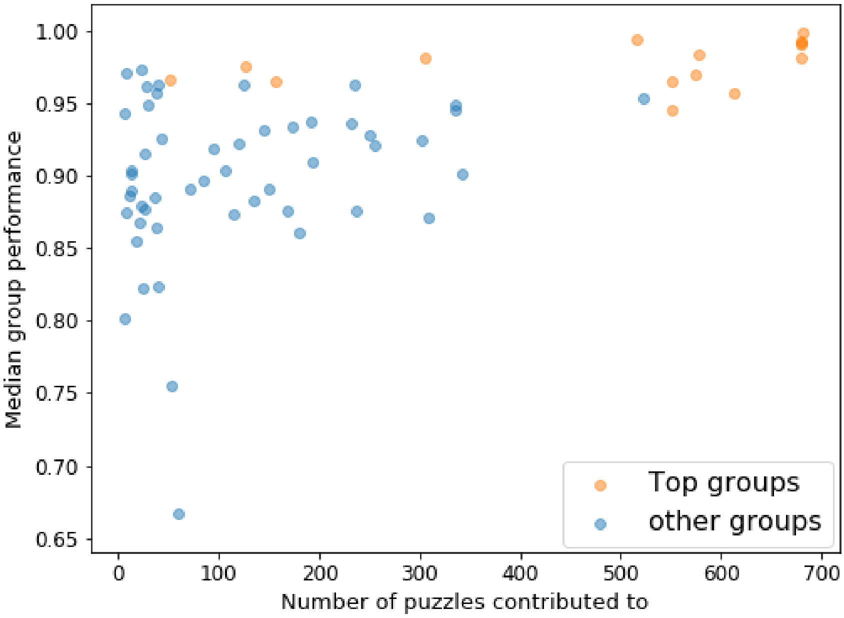

Fig. 5. The group experience feature vs median group performance.

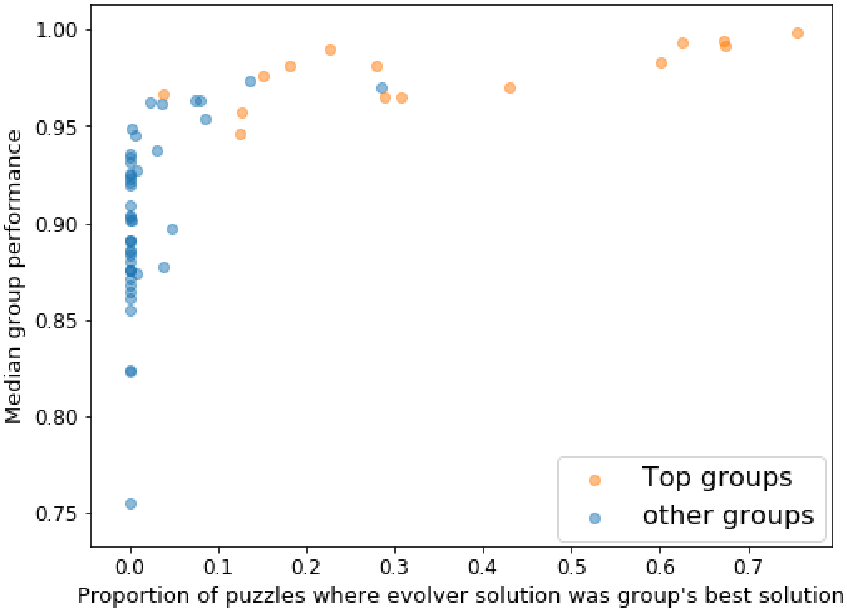

Fig. 6. The collaborative skill feature vs median group performance.

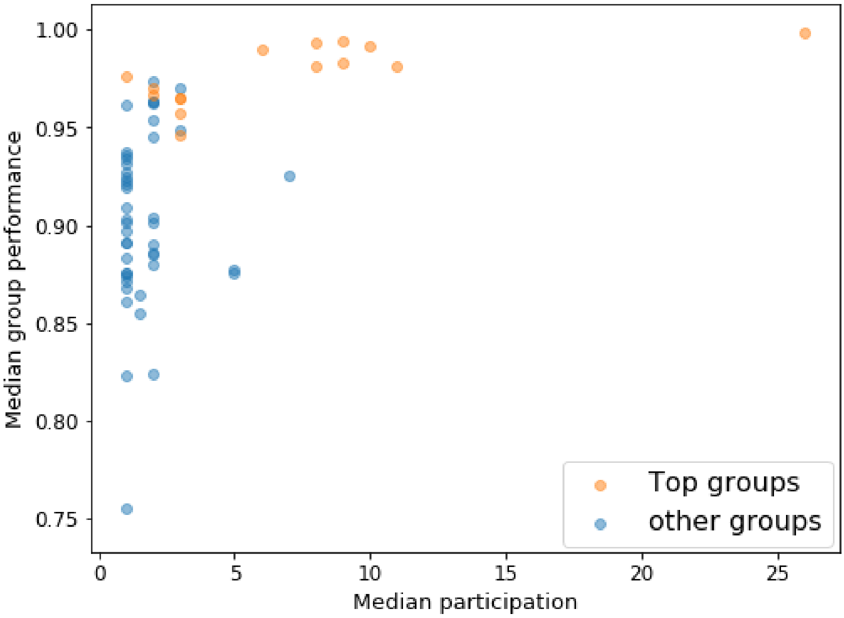

Fig. 7. The group participation feature vs median group performance.

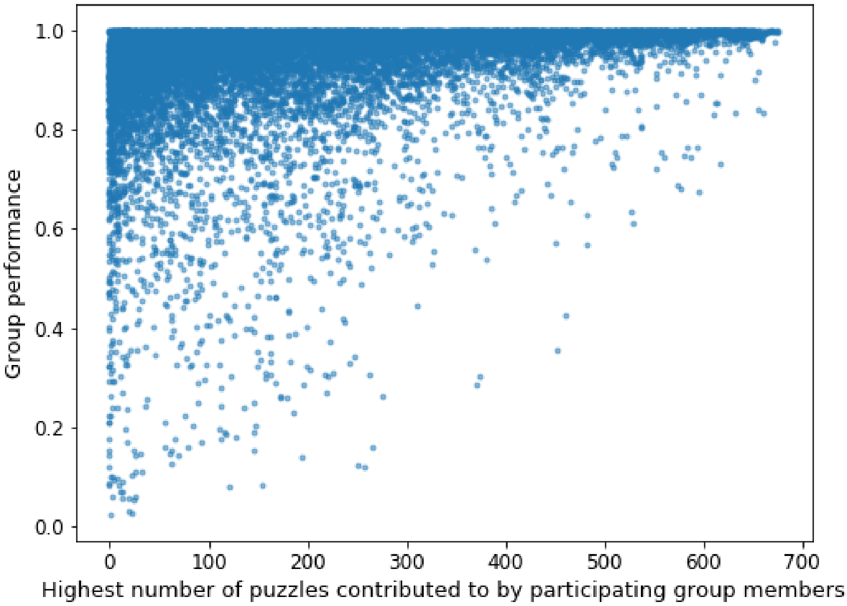

Fig. 8. The individual experience feature vs group performance.

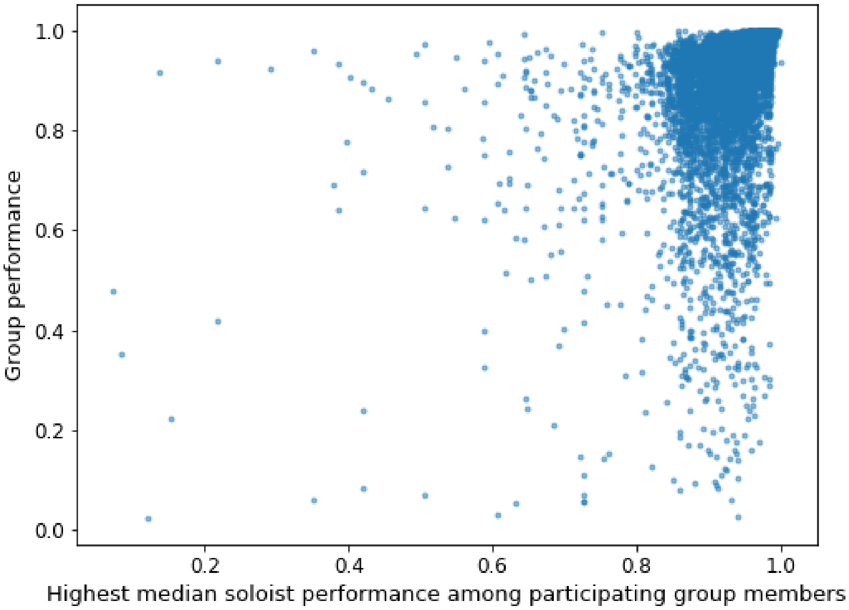

Fig. 9. The individual skill feature vs group performance.

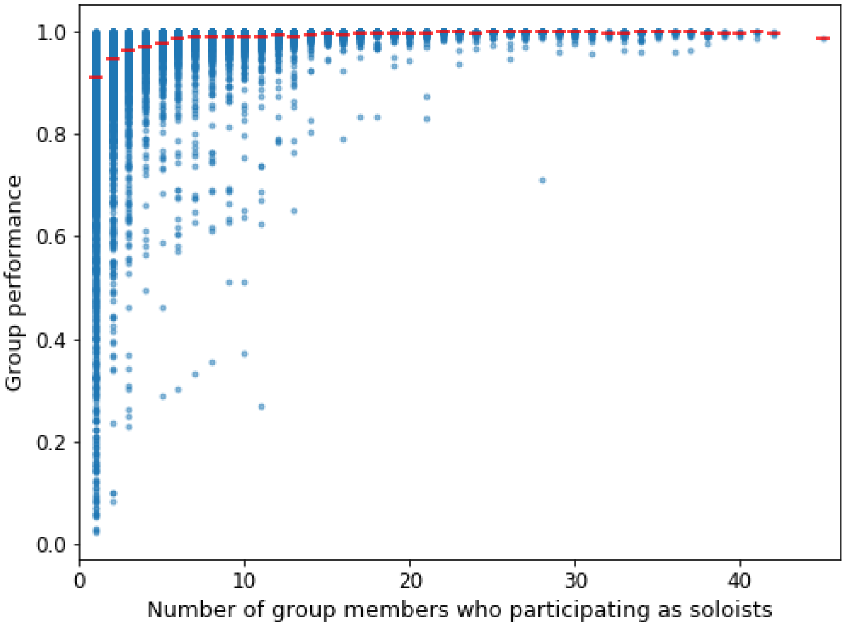

Fig. 10. The soloist participation feature vs group performance. Red lines indicate the median performance for each level of participation.

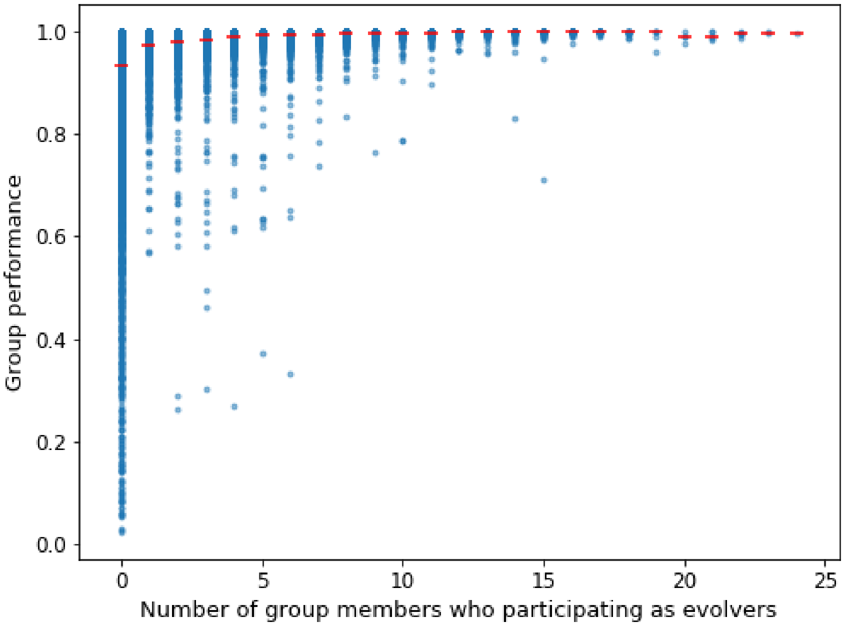

Fig. 11. The evolver participation feature vs group performance. Red lines indicate the median performance for each level of participation.

We find a significant difference between top groups and other groups for all three features. We also find significant correlation for top groups between those same three features and group performance, although at differing levels of significance. The collaborative skill feature has the largest effect size and the strongest correlations, including, unique among our three features, a significant correlation with group performance among non-top groups.

As with our macro-scale features, for each of our meso-scale features we plot the value of that feature for each group contribution (i.e., a separate data point for each puzzle each group participated in) versus the performance of that group on that puzzle. Individual experience is shown in Figure 8, individual skill is shown in Figure 9, soloist participation is shown in Figure 10, and evolver participation is shown in Figure 11. In addition, we compute Spearman's ρ for each feature to measure its correlation with group performance. These correlations are given in Table 3.

	Spearman's ρ
Individual experience	0.405^{***}
Individual skill	0.725^{***}
Soloist participation	0.687^{***}
Evolver participation	0.648^{***}

Table 3. The statistical results of our exploration of meso-scale features are given here. For each feature, we list the Spearman's rank correlation coefficient (ρ). All correlations are significant for $p < 0.001$.

We find significant correlation for all four meso-scale features, though the correlation is much weaker for individual experience than the other features. As with the macro-scale features, the measure of skill has the strongest correlation with performance. We note that participation for both soloists and evolvers appears to have diminishing returns. Using the median performance at each level of participation as a guide (shown as red hashes in Figures 10 and 11), we observe the benefits of increased participation diminish at more than six soloists and more than five evolvers.

7.3 Discussion

In our exploration of group performance in *Foldit*, we analyze the relationship of seven different features with group performance (three at the overall group level and four at the puzzle level). We also examine the differences in the former three features between top groups and other groups. We find that features measuring collaborative skill and individual skill correlate strongly with group performance, whereas features measuring group and individual experience only correlate weakly. Features measuring participation correlate moderately with group performance, though at the macro scale this correlation is only present for top groups. Finally, top groups have more experience, higher collaborative skill, and higher participation compared to other groups.

The relationship we find between collaborative skill and group performance is consistent with other work on solution quality in collaborative environments. Studies of collaboration in settings including *MathOverflow* [30] and *League of Legends* [19] find that collaborative acts improve group performance. In the context of *Foldit*, this finding highlights a promising focus for future, fine-grained analysis: investigating the collaborative organization and strategies used by groups when their best solution is an evolver solution. Understanding these events at the level of solver actions, including how solutions are shared and refined within a group, will be critical for characterizing why they are associated with long-term group success.

The relatively weak correlation between experience and group performance is surprising, especially since experience is often the strongest predictor of performance in a game [19]. One possible explanation is that the changing *Foldit* environment in terms of new puzzles and tools being added lessen the typical benefits of experience. Additional labeling of the puzzles in the *Foldit* dataset to better account for these differences would be one way to enable an investigation of this potential explanation. Another, more troubling explanation is that *Foldit* solvers are not learning from experience the way we would expect. Given the known importance of immediate feedback and reflection for learning [6], the lack of these practices as integral parts of *Foldit* may be contributing to reduced solver learning. Fortunately, existing work on improving crowdsourced solutions offer models of how reflection (in the form of self-assessment [8]) and real-time feedback (in the form of high-level expert guidance [2]) can be incorporated into online problem-solving environments. Our exploration suggests there may be an opportunity for these techniques to improve solution quality in the context of scientific-discovery games.

Participation's correlation with group performance in *Foldit* is exciting because participation in online communities is well-studied in the literature [23]. We observe that median participation is quite low for many groups, and the puzzle-level correlation of participation with group performance suggests increased participation could improve solution quality for those groups. A variety of mechanisms to increase participation have been identified, including sending personalized introductory messages emphasizing social interaction and explaining to members the value of their contributions [23]. The diminishing returns to increased participation we observe are consistent with analysis of collaboration among editors of Wikipedia articles [20]. Kittur and Kraut found that appropriate coordination techniques were necessary in order to benefit from adding more editors. A deeper look into the collaboration of *Foldit*'s top groups is necessary to reveal the coordination techniques they employ to capitalize on the benefits of higher participation.

Our initial exploration of group performance in *Foldit* is not without limitations. The thresholds we use to determine inclusion in the dataset and to differentiate top groups from other groups are clearly imperfect. Despite the requirement that every group in our dataset participate in at least five puzzles with at least two members, many of the groups appear to mostly participate with a single member (see the large number of groups with a median participation of one in Figure 7 and zero best solutions from evolvers in Figure 6). As for top groups, there are a handful we label as top groups that, at least by the features we measure, look a lot more like non-top groups. One result of our exploration is highlighting this diversity of group composition and behavior in *Foldit*.

Our exploration is limited by its focus on correlations rather than providing more explanatory or predictive insight. While building a regression model would be a natural approach to incorporate our features into a more predictive context, we fear such a model could easily be misleading due to complex and poorly understood dependencies between the features we study. Further investigation is needed before any such model can be constructed with confidence.

8 CONCLUSION

Developing a deep understanding of both individual and group success in an open-ended collaborative problem-solving environment is necessary for building the next generation of these environments that will improve outcomes through optimal design and machine intelligence. We progress toward this understanding by investigating the open-ended collaborative problem-solving ecosystem of the scientific-discovery game *Foldit*. Our investigation was motivated by two primary questions: (1) how do the social aspects of Foldit impact an individual's behavior? and (2) what factors have significant impact on group success? In order to carry out this investigation, we overcame significant challenges posed by the structure and complexity of *Foldit* data. In particular, we devised a metric of performance that accounts for puzzle difficulty and margin of success and conducted a simulated controlled experiment by pairing similar group-joining and non-group-joining to form synthetic treatment and control samples.

We analyzed the relationship between early collaboration and success and long-term participation, investigated the effect of group membership on individual performance, and explored how features of group activity in *Foldit* relate to group performance. We found that both early collaboration and early success were associated with increased participation. While those with early success participated more, those with early collaboration participated more than those without regardless of whether they had early success. We also found that group membership increased individual performance, though not universally. Finally, we found measures of group collaborative skill and individual group member skill correlate strongly with group performance, while group and individual experience only correlate weakly with performance. Participation had moderate correlation with group performance, with evidence of diminishing returns on additional participation.

We are far from exhausting the opportunity *Foldit* provides to study collaborative problem solving at scale on real-world problems. We limited our analysis here to *Foldit*'s prediction puzzles to avoid adding puzzle type as another variable, but an analysis of collaboration on design puzzles would be an exciting extension of our work. Not only do design puzzles offer a different, more creative objective, but they also incorporate a collaborative channel not present for prediction puzzles: expert feedback. For a subset of design puzzles, biochemists select a small number of solutions to try and synthesize in the lab. Experts also institute constraints on future deign puzzles to guide the solutions toward more promising designs. Other interesting variants of *Foldit* puzzles are also omitted from this work. *Hand-folding* puzzles take place over two rounds where the first round does not allow collaboration or recipes and solvers can import first-round solutions into the second round. These puzzles could provide a semi-controlled setting in which to study the effects of collaboration. *All-hands* puzzles treat the entire population of *Foldit* solvers as one big group (i.e., a shared soloist solution is shared with *everyone*). This alternative structure for collaboration could provide an illuminating contrast to the typical one where solvers are siloed into groups. Finally, *Foldit* does record some data on low-level solver activity that does not feature in our analysis. This data could be an invaluable resource for digging deeper into the findings presented here.

REFERENCES

[1] Brigid Barron. 2000. Achieving coordination in collaborative problem-solving groups. *The journal of the learning sciences* 9, 4 (2000), 403–436.

[2] Joel Chan, Steven Dang, and Steven P Dow. 2016. Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1223–1235.

[3] Seth Cooper, Firas Khatib, Ilya Makedon, Hao Lu, Janos Barbero, David Baker, James Fogarty, Zoran Popović, et al. 2011. Analysis of social gameplay macros in the Foldit cookbook. In *Proceedings of the 6th International Conference on Foundations of Digital Games*. ACM, 9–14.

[4] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

[5] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 32–41.

[6] National Research Council et al. 2000. *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press.

[7] Justin Cranshaw and Aniket Kittur. 2011. The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1865–1874.

[8] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1013–1022.

[9] Christopher B Eiben, Justin B Siegel, Jacob B Bale, Seth Cooper, Firas Khatib, Betty W Shen, Barry L Stoddard, Zoran Popovic, and David Baker. 2012. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature biotechnology* 30, 2 (2012), 190–192.

[10] Sandra B Fan, Tyler Robison, and Steven L Tanimoto. 2012. CoSolve: A system for engaging users in computer-supported collaborative problem solving. In *2012 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 205–212.

[11] J Richard Hackman. 2002. *Leading teams: Setting the stage for great performances*. Harvard Business Press.

[12] Joshua Introne, Robert Laubacher, Gary Olson, and Thomas Malone. 2011. The Climate CoLab: Large scale model-based collaborative planning. In *Collaboration Technologies and Systems (CTS), 2011 International Conference on*. IEEE, 40–47.

[13] Alan Irwin. 1995. *Citizen science: A study of people, expertise and sustainable development*. Psychology Press.

[14] Nicholas R Jennings. 1995. Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. *Artificial intelligence* 75, 2 (1995), 195–240.

[15] David H. Jonassen. 2000. Toward a design theory of problem solving. *Educational Technology Research and Development* 48, 4 (2000), 63–85.

[16] Norbert L Kerr and R Scott Tindale. 2004. Group performance and decision making. *Annu. Rev. Psychol.* 55 (2004), 623–655.

[17] Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, et al. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology* 18, 10 (2011), 1175–1177.

[18] Sunyoung Kim, Christine Robson, Thomas Zimmerman, Jeffrey Pierce, and Eben M Haber. 2011. Creek watch: pairing usefulness and usability for successful citizen science. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2125–2134.

[19] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 2017. What Makes a Strong Team?: Using Collective Intelligence to Predict Team Performance in League of Legends. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2316–2329.

[20] Aniket Kittur and Robert E Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 37–46.

[21] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.

[22] Raddick M Jordan, Bracey Georgia, and Gay Pamela. 2010. Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review* (2010).

[23] Sanna Malinen. 2015. Understanding user participation in online communities: A systematic literature review of empirical studies. *Computers in human behavior* 46 (2015), 228–238.

[24] Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*. Springer, 69–97.

[25] Dana Rotman, Jenny Preece, Jen Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. 2012. Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 217–226.

[26] Nikol Rummel and Hans Spada. 2005. Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The Journal of the Learning Sciences* 14, 2 (2005), 201–241.

[27] Burr Settles and Steven Dow. 2013. Let's get together: the formation and success of online creative collaborations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009–2018.

[28] Garold Stasser and William Titus. 2003. Hidden profiles: A brief history. *Psychological Inquiry* 14, 3-4 (2003), 304–313.

[29] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142, 10 (2009), 2282–2292.

[30] Yla R Tausczik, Aniket Kittur, and Robert E Kraut. 2014. Collaborative problem solving: A study of mathoverflow. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 355–367.

[31] Jonathan Tudge and Barbara Rogoff. 1999. Peer influences on cognitive development: Piagetian and Vygotskian perspectives. *Lev Vygotsky: critical assessments* 3 (1999), 32–56.

[32] Andrea Wiggins and Kevin Crowston. 2011. From conservation to crowdsourcing: A typology of citizen science. In *System Sciences (HICSS), 2011 44th Hawaii international conference on*. IEEE, 1–10.

[33] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

[34] Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. 2007. The increasing dominance of teams in production of knowledge. *Science* 316, 5827 (2007), 1036–1039.

Received June 2017; revised July 2017; accepted November 2017