Playing the Data: Video Games as a Tool to Annotate and Train Models on Large Datasets

Parham Ghasemloo Gheidari [1] **Kai-Hsiang Chang** [1] **Renata Mutalova** [1] **Alexander Butyaev** [1] **Attila Szantner** [1] **Roman Sarrazin-Gendron** [2] **Jérôme Waldispuhl** [1]

Abstract

Citizen science platforms can generate vast quantities of labeled data by engaging non-expert human contributors in solving tasks relevant to AI model development. In this work, we present insights from two deployed citizen science projects—Borderlands Science and Project Discovery—that have engaged millions of participants in annotating complex biological data. We discuss how human feedback collected via these platforms can be used to train or fine-tune AI models, with implications for learning from noisy demonstrations, preference aggregation, and biological discovery inspired by innate human intuition. We demonstrate how data from citizen science can be systematically used to train and evaluate machine learning models for biological sequence alignment and clustering, and propose a framework for aggregating and leveraging noisy human strategies at scale.

1. Introduction

AI models increasingly rely on large volumes of labeled data to learn accurate representations of complex phenomena. In domains such as biology, generating such labels is particularly challenging (Sapoval et al., 2022). High-quality annotations often require expert knowledge—e.g., in interpreting microscopy images, aligning genetic sequences, or classifying cell types (Sullivan et al., 2018; Sarrazin-Gendron et al., 2024). This makes large-scale data generation and annotation for model training prohibitively expensive and logistically demanding. As a result, biological datasets often suffer from limited sample size, sparse coverage, or underrepresentation of edge cases, all hindering downstream AI models' development and generalization.

To address this bottleneck, researchers have turned to human-assisted methods for scalable data manipulation (Mosqueira-Rey et al., 2022). Amazon Mechanical Turk (AMT) has become a standard tool in this space, offering a flexible workforce for annotation tasks (Crowston, 2012). However, the quality and engagement of AMT workers can vary widely. Since participants are compensated per task and often prioritize speed, complex or cognitively demanding problems can suffer from low attention, high label noise, and limited motivation for accuracy (Binder, 2022).

An alternative approach is to use game-based systems to collect data in ways that are intrinsically motivating. While originally designed for labeling tasks in domains like natural language processing and computer vision, "Games with a Purpose" (GWAPs) have since been adapted to support scientific data collection and model training. This approach has enabled the integration of complex scientific tasks by embedding data annotation in the form of a puzzle or strategy game. Notable examples include Foldit (Cooper et al., 2010), which crowdsourced solutions for protein folding, and Phylo (Kwak et al., 2013), which gamified DNA sequence alignment through a tile-matching game. While these platforms can attract highly motivated users and sustain engagement more effectively than traditional microtask systems, their reach is often limited to communities with pre-existing interests in science or puzzle-solving. Moreover, the overhead involved in developing and maintaining a standalone game significantly restricts scalability.

A more recent and promising direction comes from the Massively Multiplayer Online Science (MMOS) model, which embeds citizen science tasks directly into existing commercial video games. By leveraging the infrastructure, popularity, and production value of AAA games, this model can engage millions of users in scientific data curation without requiring them to seek out or download a separate application (Sarrazin-Gendron et al., 2024). These platforms offer not only scale but also deep engagement, as users participate during extended gameplay sessions and often return repeatedly. Additionally, the immersive nature of the hosting game

[1]Department of Computer Science, McGill University [2]Département d'informatique, Université du Québec à Montréal. Correspondence to: Jérôme Waldispuhl <jerome.waldispuhl@mcgill.ca>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

can provide context and motivation for the scientific task, improving both data quality and player retention.

In this paper, we examine how human feedback collected through large-scale, game-embedded citizen science platforms can be used to train AI models for biological problems. We focus on two case studies: Borderlands Science (BLS), which reimagines microbial DNA alignment as a tile-based puzzle embedded in Borderlands 3, and Project Discovery (PD), which integrates real-world scientific challenges—such as protein localization and cell classification—into the gameplay of EVE Online. Together, these projects demonstrate a new paradigm for collecting large-scale, non-expert feedback with relevance to machine learning, particularly in domains where data is costly or difficult to annotate using traditional means. However, learning from such data introduces formal challenges, including label noise modeling, annotator bias correction, and the design of robust aggregation strategies that can generalize across heterogeneous human input.

2. Background and Related Work

Various paradigms have emerged to incorporate human participation into large-scale data generation for computational systems. These include microtask crowdsourcing platforms such as Amazon Mechanical Turk (AMT) (Mosqueira-Rey et al., 2022), game-based approaches commonly known as Games With a Purpose (GWAP), and more recently, Massively Multiplayer Online Science (MMOS) initiatives that embed scientific tasks directly into commercial video games. Each of these approaches differs in scale, motivation structures, and the types of tasks they are best suited for, but all share a common goal: leveraging distributed human input to solve problems that are difficult to automate.

Learning from distributed human input—particularly in the form of large-scale, non-expert annotations—raises several technical challenges. In the case of reinforcement learning from human feedback (RLHF), reward modeling often relies on sparse or noisy preference signals, making it difficult to ensure stability and alignment without careful calibration or regularization (Christiano et al., 2017). Imitation learning similarly depends on capturing consistent behavior across demonstrators, yet citizen science settings frequently involve diverse strategies and variable expertise. To aggregate such input effectively, techniques like expectation-maximization-based label modeling (Dawid & Skene, 1979), confidence-weighted voting, and Bayesian crowd models (Raykar et al., 2010) have been explored to estimate both ground truth and annotator reliability. These methods remain critical when applying citizen-generated data to train robust, generalizable machine learning models.

A growing number of citizen science projects demonstrate how non-expert human contributions can support both scientific discovery and AI development. In Foldit (Cooper et al., 2010), for example, players manipulate 3D protein structures to discover energetically favorable conformations—an intuition-driven task where automated methods often fell short at the time. Similarly, Eyewire (Kim et al., 2014) is a game-based citizen science project that engages players in tracing neuron segmentation in 3D space, producing data used in neuroscience research and machine learning. In contrast, Galaxy Zoo (Lintott et al., 2008) is a non-gamified citizen science platform that has mobilized hundreds of thousands of volunteers to classify galaxy morphologies from telescope images, resulting in a benchmark dataset widely used in astronomy and computer vision. Building on the success of Galaxy Zoo, the broader Zooniverse platform (Simpson et al., 2014) hosts a wide range of non-game citizen science projects across disciplines such as astronomy, ecology, and medicine. Volunteers contribute to structured annotation tasks like identifying wildlife in camera trap images, transcribing historical documents, and classifying medical scans or biological samples.

Other projects have extended the scope of citizen science into behavioral and cognitive domains. One notable example is Sea Hero Quest, a mobile video game developed to assess human spatial navigation ability at scale (Coutrot et al., 2019). Designed in collaboration with neuroscientists, the game presents players with navigational tasks that require recalling and executing routes through virtual mazes—tasks analogous to those used in clinical cognitive tests. Since its launch, the game has been played by over 4.3 million users across 195 countries, generating the largest dataset ever collected on human spatial navigation performance (Coutrot et al., 2018; Spiers et al., 2019). The data enabled researchers to identify how age, culture, and geography affect navigation ability and revealed population-level benchmarks that can inform early diagnosis of neurodegenerative diseases such as Alzheimer's. In addition, the game has provided a foundation for modeling individual differences in spatial cognition and has been used to train AI systems on realistic navigation behavior. These findings highlight the versatility of citizen science in not only generating annotations but also capturing rich behavioral patterns, decision heuristics, and cognitive variability—factors increasingly relevant for designing human-aligned AI systems.

From a machine learning perspective, the data generated by citizen science and crowdsourcing platforms offers several distinct advantages. As outlined by Good et al. (Good & Su, 2013), these platforms contribute across a spectrum of task types critical to bioinformatics, including information collection, annotation, transcription, evaluation, and problem solving. However, how these tasks are embedded into interactive systems—particularly games—affects both the data they generate and the machine learning ap-

proaches they support. For example, projects like Foldit (Cooper et al., 2010) represent competitive games focused on individual discovery and structure optimization, where human intuition is used directly. Platforms such as EyeWire (Kim et al., 2014) and Sea Hero Quest (Coutrot et al., 2019) collect rich behavioral trajectories that can inform cognitive modeling and imitation learning. Others, like Galaxy Zoo (Lintott et al., 2008), are optimized for large-scale image classification tasks and rely on consensus aggregation for supervised learning. Meanwhile, systems like Project Discovery and Borderlands Science illustrate two design strategies for embedding scientific tasks into existing games: one by integrating real-world labeling tasks directly, and the other by rethinking game mechanics around scientific data, enabling imitation learning and alignment optimization at scale.

When such data is collected from diverse contributors, it captures not only human expertise but also behavioral variability—factors that are valuable for training machine learning models, learning ranking functions, or simulating decision heuristics. We now introduce a refined framework (Fig. 1), inspired by Good et al., specifically tailored to game-based citizen science. This alternative structure is grounded in practical examples and emphasizes how the nature of the task and the type of data collected influence the choice of downstream AI methods.

In the next section, we describe two large-scale MMOS projects—Borderlands Science and Project Discovery—that represent a new phase in this trajectory, embedding scientific annotation tasks in the flow of commercial gameplay and generating AI-relevant biological data at unprecedented scale.

3. Citizen Science Systems

3.1. Borderlands Science

Borderlands 3 is a first-person shooter role-playing game developed by Gearbox Software and released in 2019. Known for its fast-paced action, humor, and stylized visuals, the game has sold over 15 million copies worldwide as of early 2022 (Strickland, 2022). In April 2020, the game introduced a citizen science mini-game called *Borderlands Science*, developed in collaboration with Massively Multiplayer Online Science (MMOS), McGill University, and The Microsetta Initiative. This initiative, described in a 2020 correspondence in *Nature Biotechnology* (Waldispühl et al., 2020), integrates scientific research tasks into the gaming experience. Players can access the mini-game via an arcade machine aboard their spaceship, the "Sanctuary III," allowing them to engage with scientific puzzles seamlessly during gameplay. This design ensures minimal disruption to the core gaming experience while enabling optional, recurring participation in scientific research tasks.

The gameplay of *Borderlands Science* involves solving tile-matching puzzles that correspond to multiple sequence alignments (MSAs). An MSA is a way of arranging DNA, RNA, or protein sequences from different organisms so that regions of similarity—with DNA and RNA representing the genetic instructions and messages within cells—reflecting structural or functional relationships, are aligned in columns. These alignments are crucial in bioinformatics because they enable the comparison of homologous sequences across organisms, which is foundational for tasks such as phylogenetic tree construction, gene annotation, and evolutionary analysis. In the minigame, the columns are rotated 90 degrees to introduce the concept of gravity for tile manipulation, aiming to make the game more intuitive and more fun. Therefore, each row corresponds to a column of an MSA. Each puzzle represents short sections of microbial DNA, where colored bricks stand in for nucleobases (A, T, G, C). Players are instructed to align these bricks with static "guides" displayed on the left edge of the grid, correcting misalignments by inserting special yellow "gap" bricks to improve matching. Gameplay rewards precision and coverage: full rows of aligned bricks grant bonus points, and players must reach a target "par" score to progress to the next puzzle. Beyond visual alignment, the players' input helps refine ambiguous regions in the original multiple sequence alignments (MSAs), contributing to the biological task of improving 16S rRNA sequence alignments — a highly conserved yet containing highly variable genetic marker used to identify and compare bacteria across species, and therefore essential in gut microbiome research. Accurate MSAs are essential for constructing phylogenetic trees, which represent the evolutionary relationships among microbial taxa (Thompson et al., 1999). These phylogenetic trees enable researchers to infer the structure of microbial communities, understand the relationships between species, and trace their evolutionary history. In the context of the gut microbiome, well-resolved phylogenies are crucial for understanding microbial diversity, identifying disease-associated taxa, and linking microbial profiles to host lifestyle, diet, or health outcomes (Johnson et al., 2019). By helping correct alignment errors in conserved but variable regions like the V4 domain of 16S rRNA, player contributions ultimately lead to more accurate phylogenetic inference and more reliable biological conclusions.

Under the hood, the BLS system employs a full pipeline that begins with a global alignment of over 1 million sequences. From these, localized windows with uncertain alignments are extracted and converted into puzzles. These regions are often challenging for automated algorithms because of their low sequence conservation, the presence of insertions or deletions, or alignment ambiguity due to repetitive motifs or short conserved regions. Tools like MAFFT and MUS-

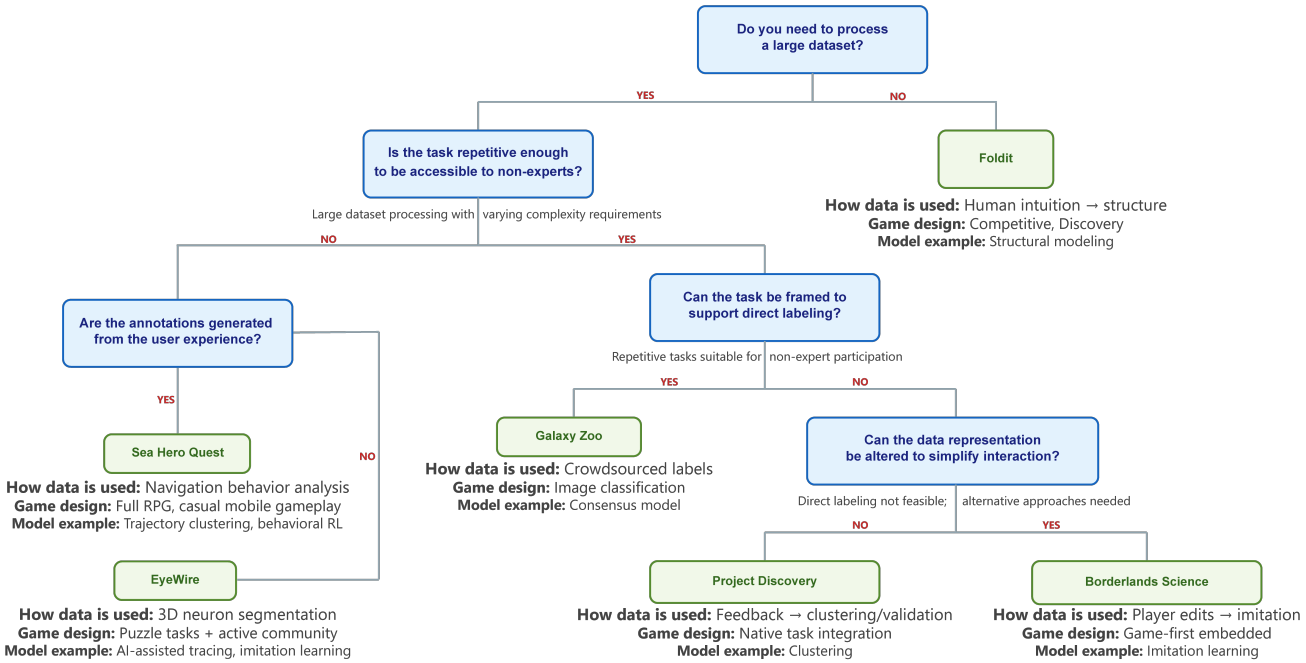

Figure 1. Typology of game-based citizen science tasks with machine learning applications, inspired by Good and Su (2013) (Good & Su, 2013). The figure refines the original categories—information collection, transcription, annotation, evaluation, and problem solving—by mapping them to interactive game designs, cognitive demands, and their alignment with downstream AI workflows.

CLE tend to either over-align (forcing artificial similarity) or under-align (leaving informative homology unrecognized) in such ambiguous cases, especially in highly diverse microbiome datasets (Edgar, 2004; Katoh & Standley, 2016). This makes the refinement of these regions a difficult, underdetermined optimization problem where multiple alignments may score similarly under a given algorithm's objective function, yet differ significantly in biological validity.

Each puzzle is then played by 45 participants, and the collected solutions are filtered, aggregated, and compared against reference alignments. Human pattern recognition allows players to spot compensatory gap placements and local homology that algorithms may miss. Importantly, humans tend to apply only the most meaningful edits, helping to avoid over-alignment—a common pitfall in algorithmic methods where too many gaps are inserted to force artificial similarity. The final set of player-informed alignments is then reintegrated into the broader MSA and subjected to downstream analysis.

3.2. Project Discovery

EVE Online is a space-based PvE and PvP MMO developed by CCP Games, first released in 2003. The game features a vast universe with over 7,800 star systems, along with a player-driven economy and community system. On average, just over 125 thousand players are online daily in 2025 (Population, 2025). Project Discovery is a citizen science game integrated into EVE Online in 2020 and features multiple phases. In phases one and two, over 322,000 participations and more than 33 million classification patterns were generated for the image segmentation task of fluorescently labeled human cells under a microscope. These data were embedded in the game, resulting in an enhanced learner capable of characterizing subcellular protein distribution, achieving an F1 score of 0.72 (Sullivan et al., 2018).

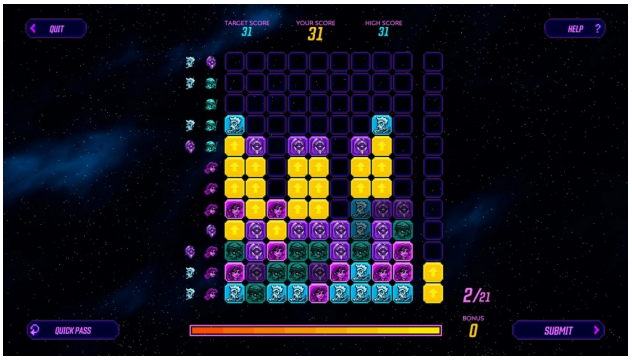

Figure 2. Borderlands Science game environment

In phases three and four of Project Discovery, flow cytometry data from the blood of COVID-19 patients is encoded

into two-dimensional scatter plots. This visualization approach was first introduced by Butyaev et al. (Butyaev et al., 2022) in the Colony B project for the clustering of bacterial cells. In Project Discovery, each data point represents a signal corresponding to a single detected cell, with the X and Y coordinates reflecting its light absorbance and fluorescence characteristics, respectively. Players are tasked with drawing enclosed shapes around clusters in these scatter plots, effectively classifying cell populations based on their distribution. Each dataset is assigned a difficulty level, and players are gradually presented with more complex data as their performance improves. Upon completing puzzles, players are rewarded with in-game currency, with the amount based on their accuracy and clustering performance. These player annotations provide valuable insight into complex clusters that have distributions that are challenging for traditional clustering algorithms. In particular, they help researchers explore the relationships between COVID-19 infection and the immune response.

At launch, over 2,000 unique puzzles were made available, each solved by multiple participants. High-quality solutions from players were then selected to train the imitation algorithm using a convolution and U-Net architecture, with a full description in the results section. By leveraging the collective intelligence of players, Project Discovery contributes to automated cell clustering techniques and may assist in identifying biologically relevant subclusters that conventional computational methods might overlook. For instance, the proximity thresholds between clusters are often smaller than those in traditional models, which could allow for increased sensitivity to subtle, high-resolution differences.

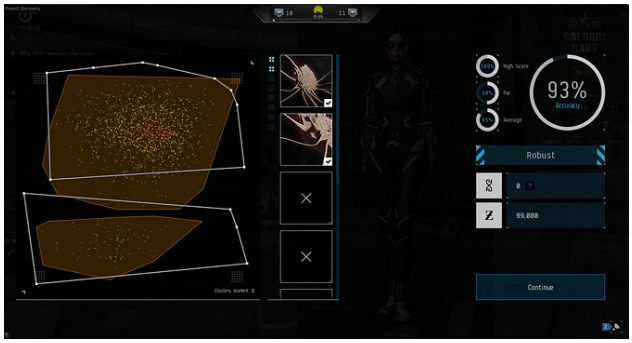

Figure 3. Project Discovery game environment

3.3. Game as a Platform for Data Annotation

In both Borderland Science and Project Discovery, the consensus of human strategies not only yields solutions comparable in quality to those derived from models trained solely on expert-labeled data, but also introduces valuable diversity in problem-solving approaches. This diversity reflects the wide range of human heuristics and intuitions, which is particularly beneficial in tasks that are ambiguous or lack a single definitive answer.

Across both platforms, we find that human strategies are not only effective but also generalizable—especially in underdetermined problems with multiple plausible optima. In such cases, the aggregation of non-expert input often reveals solution patterns that may be overlooked by purely algorithmic or deterministic methods. These findings support the potential of game-based annotation systems as scalable, engaging alternatives to conventional data labeling pipelines, capable of capturing nuanced or context-sensitive insights from a broad participant base.

4. Results and Analysis

In this section, we consolidate our discussion of the data pipeline, modeling, and outcomes from Borderlands Science and Project Discovery. Rather than isolating preprocessing and model integration, we consider how human feedback collected via each platform was structured, processed, and ultimately contributed to downstream machine learning models and scientific findings. We organize this section into two parts, each corresponding to one of the citizen science systems introduced in Section 3.

4.1. Borderlands Science

The first major study based on Borderlands Science was published in the CHI 2023 proceedings (Mutalova et al., 2023a) and focused on understanding whether human players contribute meaningfully to solving complex sequence alignment problems. Analyzing over 1.1 million player solutions spanning 25,000 puzzles, the researchers examined player behavior across different puzzle difficulty levels. When compared to baseline algorithms like greedy heuristics, progressive profile strategies (PPS), and random placements, human players consistently achieved higher alignment scores. For instance, humans placed gaps more efficiently than simple algorithms (Mutalova et al., 2023a), often focusing on regions with clear visual misalignments and avoiding unnecessary edits. This contrasts with heuristic or exhaustive search methods like Needleman-Wunsch, which can struggle in complex puzzles by over-inserting or misplacing gaps in ambiguous regions (Mutalova et al., 2023b). As difficulty increased, human solutions became significantly more optimal than algorithmic counterparts, particularly in their ability to balance game score with gap usage.

Beyond performance, the study investigated the nature of human strategies. It found that players tend to use specific gap placement patterns that differ from those of algorithmic solvers, placing more one-gap columns than gap-free ones,

contrary to most heuristics. These behavioral patterns were then modeled using imitation learning. Both transformer and fully connected neural network (FCN) architectures were trained on Pareto-optimal player solutions. These behavior cloning models achieved high cosine similarity (up to 0.79) and low Levenshtein distances compared to human consensus alignments. Notably, even when not exactly imitating human moves, the models achieved comparable alignment scores, suggesting that they captured underlying strategic structures present in human gameplay. This points to the broader conclusion that human decision patterns in puzzle-based alignments can not only outperform naive algorithms but also be successfully generalized and modeled by machine learning systems.

In a follow-up study presented at Collective Intelligence 2023, researchers introduced a player-guided learning system—Generative Adversarial Imitation Learning (GAIL)—designed to enhance alignment performance using the BLS dataset. GAIL was trained on player-generated alignment solutions, and its performance was compared against a standard Deep Q-Network (DQN) agent. The contrast was substantial: while DQN hovered just above the minimum required score on average, the GAIL model consistently produced solutions near or above average human performance, especially in higher difficulty puzzles. These improvements were attributed to the fact that GAIL learned not only how to solve puzzles efficiently, but also how to balance alignment score with the number of gaps inserted—a core tradeoff in the game. The GAIL model extended the DQN framework by introducing Generative Adversarial Networks (GANs), where the generator (the DQN agent) learns from both its own experiences and adversarial feedback provided by a discriminator. The discriminator, trained on expert player solutions, distinguishes between human-generated and agent-generated solutions, guiding the agent to adopt more human-like strategies.

Quantitatively, BLS-aligned sequences yielded phylogenetic trees—tree structures that represent the evolutionary relationships among microbial species—that were closer to a curated expert-derived reference tree than those generated by all other methods tested. For example, BLS achieved the lowest combined phylogenetic distance score (48.4), compared to 67.1 for a post-processed alignment baseline and 87.2 for a raw alignment generated without refinement. This indicates that the human-aligned sequences were more consistent with expert-curated microbial phylogenies than purely algorithmic approaches.

Crucially, the improved alignments also had tangible effects on real-world microbiome analysis. Using diversity metrics computed across 74 host-related variables (e.g., diet, lifestyle, health status), the authors found that the BLS-derived phylogenies enabled stronger separation between microbial communities. In variables such as teeth brushing frequency, prior *Clostridium difficile* infection, diabetes, and antibiotic history, the BLS trees exhibited larger effect sizes than the baseline method derived from the expert reference. In fact, BLS led to significantly stronger associations (higher significance category) for 13 variables, while the baseline did so for 10—mostly in less biologically relevant categories. This demonstrates that the BLS-enhanced phylogenies preserve subtle but meaningful signals in microbiome data that are relevant to human health.

Finally, the structural integrity of the BLS alignments was validated by comparing them to known secondary structures of the 16S rRNA gene—patterns of nucleotide base pairing (such as stems and loops) that contribute to the molecule's functional three-dimensional shape. The BLS alignment preserved a higher proportion of nucleotide bases that could be mapped to this structural model and showed strong agreement in regions of known biological function, such as binding sites—locations where helper proteins, like S8 and S15, attach to the RNA to help it fold correctly and carry out its role in building other proteins. These findings support the conclusion that player-generated alignment adjustments—even in small puzzle segments—capture biologically valid structural constraints that are often missed by traditional algorithms.

4.2. Project Discovery

Phases three and four of Project Discovery focus on exploring performance, benefits, and applications of imitation models on human strategies versus machine learning based models. After analyzing over one million player-submitted solutions across more than 2,000 samples, the data reveals that human consensus performs comparably to baseline algorithms—such as Agglomerative, Gaussian Mixture, BIRCH, and KMeans—on simpler puzzles. However, as puzzle complexity increases, human performance begins to exceed that of the algorithms. A clear example of this is when two clusters are positioned closely together: baseline methods often merge them into a single group, whereas human players are more adept at distinguishing them as separate entities. In addition to identifying the correct number of clusters, players also reach consensus on spatial arrangements that closely match the true underlying structure of the data.

To assess how well algorithmic clustering mimics human behavior, the study utilized several metrics, including intersection over union (IoU), which measures the overlapping spatial area between algorithmic and player outputs; the Dice coefficient, which balances precision and recall; and pixel-wise accuracy, which evaluates the proportion of correctly classified pixels (Zou et al., 2004; Cho, 2024). When comparing baseline models to player solutions, our preliminary results found that no existing model perfectly

replicates human decision-making. Agglomerative clustering achieved the highest IoU and pixel-wise accuracy but displayed a significantly lower Dice coefficient, indicating that it approximates human spatial patterns while exhibiting greater variability in the shape and consistency of clusters. By contrast, Gaussian Mixture, BIRCH, and KMeans models consistently scored lower across all three metrics, suggesting they are less effective at capturing the subtle patterns that human intuition can detect.

Table: Mean performance metrics for clustering models. Agglomerative clustering shows the best alignment with human annotations.

Model Type	IoU	Dice Coefficient	Accuracy
Agglomerative	0.1236	0.0039	0.4157
Birch	0.0721	0.0473	0.2316
GaussianMixture	0.0687	0.0460	0.2232
KMeans	0.0692	0.0463	0.2235

In addition to baseline models, an imitation learning approach was implemented to evaluate how effectively machine learning models can replicate human behavior. The imitation learning model outperformed all baseline algorithms, achieving an Intersection over Union (IoU) score of 56.36%, demonstrating strong alignment with human-generated results. This indicates that the model effectively captures key elements of human decision-making, particularly in identifying cluster boundaries and spatial arrangements. These results highlight the potential of imitation learning to replicate intuitive human strategies, offering promising applications for complex data clustering tasks in the future.

Finally, using the imitation learning algorithm, insights into human decision-making are also explored, as shown in Figure 4. Using the human-inspired model trained with player solutions, reliable identification of separate clusters can be detected when inter-cluster distances are sufficiently large. As shown in Figure 4, we artificially created three clusters with varying densities and dispersions to test our model's ability to detect each cluster accurately. The classification results indicate the existence of a critical threshold distance that governs the transition between detecting clusters as separate entities and merging them. In general, data with high spread and low density tends to result in merging, whereas data that is more compact and high-density tends to be separated more reliably.

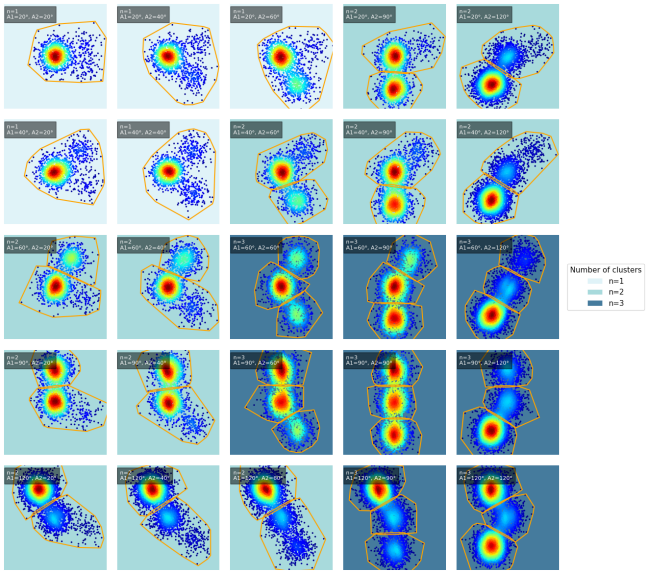

Figure 4. Human-inspired imitation model making predictions on artificially generated puzzles with three overlapping clusters with varying densities and spatial arrangements. Predictions revealed human biases toward recognizing denser and compact regions as distinct clusters, merging sparse or heavily overlapping areas into fewer clusters.

5. Discussion

Training robust and accurate AI models capable of understanding underlying data representations relies heavily on the quantity and quality of labeled data. However, in fields such as biology where data acquisition is expensive and expert annotation is limited, producing high-quality training datasets remains a significant challenge (Sapoval et al., 2022). Moreover, while AI models excel at pattern recognition and large-scale data analysis, there are many scenarios particularly in creative domains and complex decision-making—where mimicking human strategies offers clear advantages. For instance, generative models, although capable of producing novel images, often "hallucinate" or exhibit biases toward objects that appear more frequently in the training data (Aithal et al., 2024). On the other hand, decisions that may seem simple, such as whether to turn left or right while driving or choosing which car to buy, often involve complex contextual reasoning and causal inference (Ratcliff et al., 2016). These connections are difficult to capture with purely learned hidden representations. Humans, by contrast, are naturally adept at navigating ambiguity, applying contextual understanding, and recognizing patterns in noisy or incomplete data—particularly when those patterns are not explicitly defined by algorithms. Tasks such as image segmentation, cell clustering, and sequence alignment still benefit significantly from expert intuition and visual reasoning, which remain challenging to replicate with AI alone.

Incorporating human-like strategies enables the development of hybrid systems in which human insight enhances

and guides machine learning processes, resulting in models that are more interpretable, adaptable, and generalizable. This approach is especially valuable in early-stage scientific discovery or hypothesis generation, where rigid algorithmic models may overlook subtle but meaningful features (Koivisto & Grassini, 2023; Nomura et al., 2024). As titled by Jaeger et al (Jaeger, 2025) artificial intelligence is Algorithmic Mimicry, by integrating human reasoning—either directly through citizen science or indirectly through cognitive modeling—these systems can address key limitations of data-only approaches. In this context, the role of the human is not to be replaced, but rather extended and embedded within AI workflows to support the development of more robust, context-aware technologies.

Our experiments and results from Borderland Science and Project Discovery demonstrate that embedding data annotation tasks within popular games allows researchers to collect high-quality and diverse annotations at a scale comparable to established platforms like Amazon Mechanical Turk. When designed effectively, citizen science platforms can mobilize a broad and engaged contributor base, generating valuable training data for AI systems.

One of the central challenges of this approach lies in translating complex scientific data into intuitive and engaging game experiences without oversimplifying the underlying science. This requires careful attention to both scientific integrity and user experience, particularly when managing sensitive data and ensuring confidentiality (Waldispühl et al., 2020). Designing annotation tasks that are scientifically meaningful yet accessible to non-experts is critical for ensuring broad participation and meaningful contributions. To sustain long-term player engagement, it is essential to align scientific objectives with compelling game mechanics, reward systems, and narrative integration. Gamification elements—such as progression, achievements, social recognition, and storyline immersion—not only keep players motivated but also foster a sense of purpose and connection to real-world impact. Additionally, regular updates, feedback on scientific contributions, and community-building initiatives help maintain player interest over time (Sarrazin-Gendron et al., 2024).

Importantly, embedding citizen science into games introduces ethical considerations that must be addressed proactively. Transparency regarding how player data and contributions are used is essential to building trust. Informed consent, robust data privacy practices, and clear opt-out mechanisms must be implemented and communicated. Fair recognition of contributors—whether through in-game rewards, public acknowledgment, or co-authorship—further ensures that players are valued ethically and respectfully. Designing for inclusivity and accessibility is also key to enabling equitable participation across diverse demographics.

These systems must tap into both intrinsic motivators—such as curiosity, learning, and a sense of purpose—and extrinsic motivators, including competition, recognition, and in-game rewards. While the full scope of scientific problems amenable to gamification is still being explored, our previous projects, such as Project Discovery and Borderlands Science, have demonstrated promising results in both cluster-based problems and NP-hard tasks like sequence alignment. The structure and design of the citizen science platform play a crucial role in shaping the quality and utility of the collected data. Moreover, citizen-contributed annotations introduce natural variability and diverse interpretations often absent from expert-curated datasets. This variability is a valuable feature when training AI models that must generalize and perform reliably in noisy, real-world environments.

6. Future Directions

Both Borderlands Science and Project Discovery have demonstrated the potential of converting scientific annotation tasks into large-scale game projects, effectively crowdsourcing valuable data for machine learning tasks. By engaging players in interactive gameplay while contributing to complex scientific problems, these platforms offer a novel approach to addressing challenging tasks in fields like bioinformatics and sequence alignment. These projects, by transforming traditionally tedious and specialized tasks into engaging, game-like environments, not only increase the scale of data collection but also enhance the diversity and richness of human input. Players are incentivized through in-game rewards, thus contributing to the rapid annotation of vast datasets that would otherwise be difficult or time-consuming to process manually.

However, while these approaches provide valuable insights into the potential of game-based citizen science for data collection, the focus has largely been on biological datasets, particularly in sequence alignment and clustering problems. These types of datasets often require specialized knowledge and expertise, which makes them well-suited for human involvement in scientific games. Players can, through their interaction, solve alignment tasks with a degree of accuracy and consistency that traditional computational methods may struggle to achieve, especially when considering complex or ambiguous sequences. Despite these successes, the current reliance on biological datasets presents several limitations. To fully understand the broader implications and benefits of this method, it is important to generalize the approach to datasets from non-biological domains and across various types of data. For instance, applying this method to fields such as climate science, chemistry, or social sciences could reveal different patterns of performance and offer insights into how the game-based data collection method scales across various disciplines. By expanding beyond the scope of biological data, we can better assess the adaptabil-

ity and robustness of these platforms for a range of scientific questions. Potential Drawbacks

While the potential for large-scale citizen science projects is clear, there are several drawbacks and challenges that need to be addressed for broader adoption. One key limitation is the accuracy and quality control of player-generated data. Despite the engagement of many players, the quality of their contributions may vary significantly. While some players may perform tasks with high precision, others may make errors or take shortcuts. Ensuring that the data collected is reliable and accurate, without the need for extensive post-processing, can be difficult. This issue becomes more significant as the complexity of the scientific problem increases, requiring more careful validation and error-checking systems.

Additionally, there is the issue of data privacy. While most data collected in these platforms may be anonymized, it's important to ensure that any personal or identifiable information remains protected. Players might be unknowingly providing data that could be misused, and adequate safeguards need to be in place to protect the privacy of participants, especially in regions with strict data protection laws.

7. Informed Consent

Before participation, individuals are informed about what the data is about, how it can be useful and the fact that they will help correct mistakes made by computers. The official video can be view here. Both biological data and user information are anonymized to ensure privacy, security, and confidentiality. The intended use of the data is regularly updated on the official website and reflected in all related publications, ensuring transparency throughout the project. This approach ensures participants are consistently aware of how their data is being used, maintaining ethical standards of consent and privacy.

8. Data Availability

All data obtained from Borderland Science Project are publicly available at https://games.cs.mcgill.ca/bls/. The data are also publicly available on mirror sites at https://gitlab.com/borderlands-science/BLS1 and https://doi.org/10.6084/m9.figshare.24962349. These data include all the puzzles produced by the project, all the solutions submitted by the players and related data. The detailed description of the data and instruction on how to download them are also available. The players are identified by unique alphanumeric strings as it was not possible to obtain their informed consent to share their personal information, as these data were collected through a video game.

Currently, data from EVE online are not publicly available, but the detailed information is available at https://games.cs.mcgill.ca/project/project-discovery/ and https://www.eveonline.com/discovery.

9. Model Training Details

During training of imitation algorithm on the Borderland Science Project, the DQN agent (the generator) updates its policy through experience replay and Q-value updates, using a learning rate of 0.001 and a discount factor of 0.99. The discriminator's weights are updated using the GAIL loss, which utilizes Binary Cross-Entropy loss, and is optimized with Adam at a learning rate of 0.0001. The adversarial feedback from the discriminator helps refine the agent's behavior, encouraging it to generate solutions that closely match human decision-making. GAIL was trained on large volumes of player-generated alignment data, allowing it to learn from implicit human preferences and problem-solving heuristics.

For Project Discovery, a convolutional neural network (CNN) was first trained to predict cluster counts from input images using mean squared error loss. The predicted counts were then rescaled and spatially expanded to form a fourth channel, which was concatenated with the original 3-channel image data. Using this 4-channel input, a modified U-Net architecture was trained for the segmentation task, optimized with binary cross-entropy loss and evaluated based on pixel-wise accuracy.

Code availability

All the code used to generate data, process data and compute results presented in this paper is freely available as a GitLab repository linked on the project website: https://games.cs.mcgill.ca/bls/. The code is also publicly available on mirror sites at https://gitlab.com/borderlands-science/BLS1 for Borderland Science Project. All code used to generate results for Project Discovery is available at https://github.com/renata-nerenata/ccp.

Acknowledgments

We would like to acknowledge all participants from Borderland Science and EVE Online for contributing to this project. We would also like to acknowledge the Massively Multiplayer Online Science (MMOS) team, based in Gryon, Switzerland, specifically Kornél Erhart and Attila Szantner, for their collaboration and input.

Our gratitude extends to Gearbox Studio Québec, Québec, QC, Canada, particularly David Bélanger, Michael Bouffard, Mathieu Falaise, Vincent Fiset, Steven Hebert, Jonathan Huot, Jonathan Moreau-Genest, Ludger Saintélien, Amélie

Brouillette, Gabriel Richard, and Sébastien Caisse, for their technical assistance and feedback. We also thank Gearbox Entertainment Company, Frisco, TX, USA, and its team members, including Joshua Davidson, Dan Hewitt, Seung Kim, David Najjab, Steve Prince, and Randy Pitchford, for their support in the development and integration of the system.

Special thanks to the Department of Pediatrics, Department of Computer Science, Department of Bioengineering, and the Center for Microbiome Innovation, all at the University of California, San Diego, La Jolla, CA, USA, with particular recognition to Daniel McDonald and Rob Knight for their insights into microbiome-related data and their expertise.

Impact Statement

This work establishes a novel, scalable paradigm for training and evaluating machine learning models by embedding scientific tasks directly into mainstream video games which is a significant departure from traditional citizen science platforms and educational tools. By leveraging the infrastructure and reach of commercial gaming, we demonstrate that high-quality human feedback can be gathered across multiple data enabling applications in sequence alignment, clustering, and policy imitation.

Crucially, our framework introduces a confidentiality-preserving approach that encodes both input data and player-generated solutions. This addresses long-standing ethical concerns around privacy and data protection in crowdsourced AI pipelines, offering a balanced methodology that supports scale without compromising trust. Moreover, this collaboration between academic researchers and a commercial game studio illustrates that partnerships across sectors can be both feasible and highly productive, enabling access to larger, more diverse populations and edge case detection. Our results show that models trained through this integrated feedback loop are not only accurate but better aligned with human reasoning in tasks marked by ambiguity and limited information. This work broadens the potential of citizen science in AI development, transforming gameplay into a powerful, ethical engine for machine learning progress.

At a societal level, the approach promotes inclusive participation in research, lowering barriers to entry while surfacing new ethical questions around data governance and representation. As citizen science becomes more deeply embedded in AI workflows, our work provides a foundational model for how such systems can scale responsibly.

References

Aithal, S. K., Maini, P., Lipton, Z. C., and Kolter, J. Z. Understanding hallucinations in diffusion models through mode interpolation. arXiv preprint arXiv:2406.09358, 2024.

Binder, C. C. Time-of-day and day-of-week variations in amazon mechanical turk survey responses. Journal of Macroeconomics, 71:103378, 2022. doi: 10.1016/j.jmacro.2021.103378.

Butyaev, A., Drogaris, C., Tremblay-Savard, O., and Waldispühl, J. Human-supervised clustering of multidimensional data using crowdsourcing. Royal Society Open Science, 9:211189, 2022. doi: 10.1098/rsos.211189.

Cho, Y.-J. Weighted intersection over union (wiou) for evaluating image segmentation. Pattern Recognition Letters, 2024:9 pages, 11 figures, 2024. doi: 10.1016/j.patrec.2024.07.011.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30, 2017.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., and Players, F. Predicting protein structures with a multiplayer online game. Nature, 466(7307):756–760, 2010. doi: 10.1038/nature09304.

Coutrot, A., Silva, R., Manley, E., de Cothi, W., Sami, S., Bohbot, V. D., Wiener, J. M., Hölscher, C., Dalton, R. C., Hornberger, M., and Spiers, H. J. Global determinants of navigation ability. Current Biology, 28(17):2861–2866.e4, 2018. doi: 10.1016/j.cub.2018.06.009.

Coutrot, A., Schmidt, S., Coutrot, L., Pittman, J., Hong, L., Wiener, J. M., Hölscher, C., Dalton, R. C., Hornberger, M., and Spiers, H. J. Virtual navigation tested on a mobile app is predictive of real-world wayfinding navigation performance. PloS One, 14(3):e0213272, 2019. doi: 10.1371/journal.pone.0213272.

Crowston, K. Amazon mechanical turk: A research tool for organizations and information systems scholars. In Shaping the Future of ICT Research. Methods and Approaches, volume 389 of IFIP Advances in Information and Communication Technology, pp. 210–221. Springer, 2012. doi: 10.1007/978-3-642-30925-7_19.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28(1):20–28, 1979. doi: 10.2307/2346806.

Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5):1792–1797, 2004. doi: 10.1093/nar/gkh340.

Good, B. M. and Su, A. I. Crowdsourcing for bioinformatics. Bioinformatics, 29(16):1925–1933, 2013. doi: 10.1093/bioinformatics/btt333.

Jaeger, J. Artificial intelligence is algorithmic mimicry: Why artificial "agents" are not (and won't be) proper agents. Journal Section, 2025.

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., and Weinstock, G. M. Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. Nature Communications, 10(1):5029, 2019. doi: 10.1038/s41467-019-13036-1.

Katoh, K. and Standley, D. M. A simple method to control over-alignment in the mafft multiple sequence alignment program. Bioinformatics, 32(13):1933–1942, 2016. doi: 10.1093/bioinformatics/btw108.

Kim, J. R., Greene, M. J., Zlateski, A., Lee, K., Richardson, M., Turaga, S. C., Purushothaman, G., Gray Roncal, W., Vogelstein, J. T., and Seung, H. S. Space-time wiring specificity supports direction selectivity in the retina. Nature, 509(7500):331–336, 2014. doi: 10.1038/nature13240.

Koivisto, M. and Grassini, S. Best humans still outperform artificial intelligence in a creative divergent thinking task. Scientific Reports, 13:13601, 2023. doi: 10.1038/s41598-023-40858-3.

Kwak, D., Kam, A., Becerra, D., Zhou, Q., Hops, A., Zarour, E., Kam, A., Sarmenta, L., Blanchette, M., and Waldispühl, J. Open-phylo: a customizable crowd-computing platform for multiple sequence alignment. Genome Biology, 14(10):R116, 2013. doi: 10.1186/gb-2013-14-10-r116.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. Monthly Notices of the Royal Astronomical Society, 389(3):1179–1189, 2008. doi: 10.1111/j.1365-2966.2008.13689.x.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review, 56(4):3005–3054, 2022. doi: 10.1007/s10462-022-10246-w.

Mutalova, R., Sarrazin-Gendron, R., Cai, E., Richard, G., Ghasemloo Gheidari, P., Caisse, S., Knight, R., Blanchette, M., Szantner, A., and Waldispühl, J. Playing the system: Can puzzle players teach us how to solve hard problems? In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–15, 2023a.

Mutalova, R., Sarrazin-Gendron, R., Ghasemloo Gheidari, P., Cai, E., Richard, G., Caisse, S., Knight, R., Blanchette, M., Szantner, A., and Waldispühl, J. Player-guided ai outperforms standard ai in sequence alignment puzzles. In Proceedings of The ACM Collective Intelligence Conference, CI '23, pp. 53–62, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701139. doi: 10.1145/3582269.3615597.

Nomura, M., Ito, T., and Ding, S. Towards collaborative brain-storming among humans and ai agents: An implementation of the ibis-based brainstorming support system with multiple ai agents. In Proceedings of the ACM Collective Intelligence Conference, CI '24, pp. 1–9, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705540. doi: 10.1145/3643562.3672609.

Population, M. Eve online - mmo population. https://mmo-population.com/game/eve-online, 2025. Accessed: 2025-05-20.

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. Diffusion decision model: Current issues and history. Trends in Cognitive Sciences, 20(4):260–281, 2016. doi: 10.1016/j.tics.2016.01.007.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. Journal of Machine Learning Research, 11:1297–1322, 2010.

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C. J., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. A. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C. R., Yan, Z., Yao, V., and Treangen, T. J. Current progress and open challenges for applying deep learning across the biosciences. Nature Communications, 13(1728), 2022. doi: 10.1038/s41467-022-29246-9.

Sarrazin-Gendron, R., Ghasemloo Gheidari, P., Butyaev, A., Keding, T., Cai, E., Zheng, J., Mutalova, R., Mounthanyvong, J., Zhu, Y., Nazarova, E., Drogaris, C., Erhart, K., Team, B. S. D., players, B. S., Brouillette, A., Richard, G., Pitchford, R., Caisse, S., Blanchette, M., McDonald, D., and Waldispühl, J. Improving microbial phylogeny with citizen science within a mass-market video game. Nature Biotechnology, 43(1):76–84, 2024. doi: 10.1038/s41587-024-02175-6.

Simpson, R., Page, K. R., and De Roure, D. Zooniverse: observing the world's largest citizen science platform. Proceedings of the 23rd international conference

on World wide web, pp. 1049–1054, 2014. doi: 10.1145/2567948.2579215.

Spiers, H. J., Coutrot, A., Manley, E., de Cothi, W., Hornberger, M., and Bohbot, V. D. Virtual navigation tested on a mobile app is predictive of real-world wayfinding navigation performance. PLOS ONE, 16(3):e0246405, 2019. doi: 10.1371/journal.pone.0246405.

Strickland, D. Borderlands 3 sales hit 15 million, franchise sales at 74 million. TweakTown, 2022. URL https://www.tweaktown.com/news/84667/borderlands-3-sales-hit-15-million-franchise-at-74/index.html.

Sullivan, D. P., Winsnes, C. F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., Nordgren, A., Smith, K., Revaz, B., Finnbogason, B., Szantner, A., and Lundberg, E. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. Nature Biotechnology, 36(9):820–828, 2018. doi: 10.1038/nbt.4225.

Thompson, J. D., Plewniak, F., and Poch, O. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Research, 27(13):2682–2690, 1999. doi: 10.1093/nar/27.13.2682.

Waldispühl, J., Szantner, A., Knight, R., Caisse, S., and Pitchford, R. Leveling up citizen science. Nature Biotechnology, 38:1129–1130, 2020. doi: 10.1038/s41587-020-0589-0.

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells, W. M., r., Jolesz, F. A., and Kikinis, R. Statistical validation of image segmentation quality based on a spatial overlap index. Academic Radiology, 11(2):178–189, 2004. doi: 10.1016/s1076-6332(03)00671-8.