Evolving Difficulty-Targeted Bouldering Routes

Daniel Tyebkhan

May 5, 2023

Abstract

The challenge of utilizing artificial intelligence to generate indoor rock climbing routes with a specific grade is an interesting and unsolved problem due to its complexity and subjectivity. We use MAP-Elites, an evolutionary, quality-diversity algorithm, in conjunction with GradeNet [8] to produce a set of disjoint MoonBoard climbing routes that sufficiently challenge a climber without exceeding their physical and technical limitations. We evaluate these routes through visual a assessment survey by climbers as well as an in-person study in which climbers attempt to climb the generated routes. While our algorithm generally performs well in producing complete or near-complete archives of diverse climbs at every difficulty level as assessed by GradeNet, they fall short when it comes to in person trials. Additionally, the data from user surveys, while supporting the claims of Duh and Chang [8] about GradeNet's superiority to human grading ability, is inconclusive in determining the success of our algorithm. These results leave open the path for future work to leverage the relative success of quality-diversity while accounting for the shortcomings of route quality and difficulty present in our system's design.

Contents

List of Figures

1 Introduction

Over the last two decades, the sport of rock climbing has seen a massive rise in popularity. It offers climbers of all skills a variety of mentally and physically challenging problems to engage with and solve. Although the sport originated in the outdoors, indoor climbing gyms have become a popular, safer, more consistent environment in which climbers of all abilities can train. One of the most important aspects of indoor climbing is *route setting*, the process of deciding where to place objects, called *holds*, on the wall for climbers to grab while climbing. The people who make these selections, called *setters*, have the difficult task of deciding how to create a variety routes that are appealing to climbers of all levels in order to keep gym members engaged. Additionally, climbers seeking improve their abilities want to climb routes that challenge them at or slightly beyond their current skill level.

Route setting is typically delegated to an experienced climber. After a route is set, its difficulty, or *grade*, is determined subjectively based on how difficult the setter and other experienced climbers believe the route to be. To keep things interesting, routes of similar difficulties should be diverse. In other words, each route at a given grade should offer climbers a different experience in terms of techniques and movements required. The ability to construct a variety of interesting routes that challenge a broad range of climbers is the mark of a good route setter. It is difficult for the average climber to create their own routes.

Due to the construction of modern-day climbing gyms, there are a near infinite number of ways in which a setter can arrange holds to form a route. This large domain combined with the desire for a variety of solutions makes this problem a prime candidate for the application of a *quality-diversity algorithm* (QDA), a type of evolutionary algorithm that seeks to find multiple viable solutions to a given problem. We address the question: *Can a quality-diversity algorithm be applied to route setting in order to create a variety of viable problems at a designated difficulty?* In this project, we apply the MAP-Elites QDA [15] to the route setting problem to generate a variety of routes at a given difficulty. We then test these routes with real life climbers and route setters to assess their viability. For ease of testing and transferring to real setting, we make some simplifications by reducing the domain from all setting configurations in a gym to a standardized training board, specifically the MoonBoard. That said, the nature of MAP-Elites should allow this approach to scale to larger route canvases depending on what resources a setter has to work with. Our setup builds on work in automatic route grading by Duh and Chang [8] and a selection of research applying QDAs to video game level design [20].

In Section 2 and Section 3 we discuss the necessary components, definitions, and details of rock climbing and QDAs respectively. In Section 4 we discuss the details of our experiments: behavioral descriptors and fitness function for MAP-Elites, and our methods for the online survey and in person testing of routes.

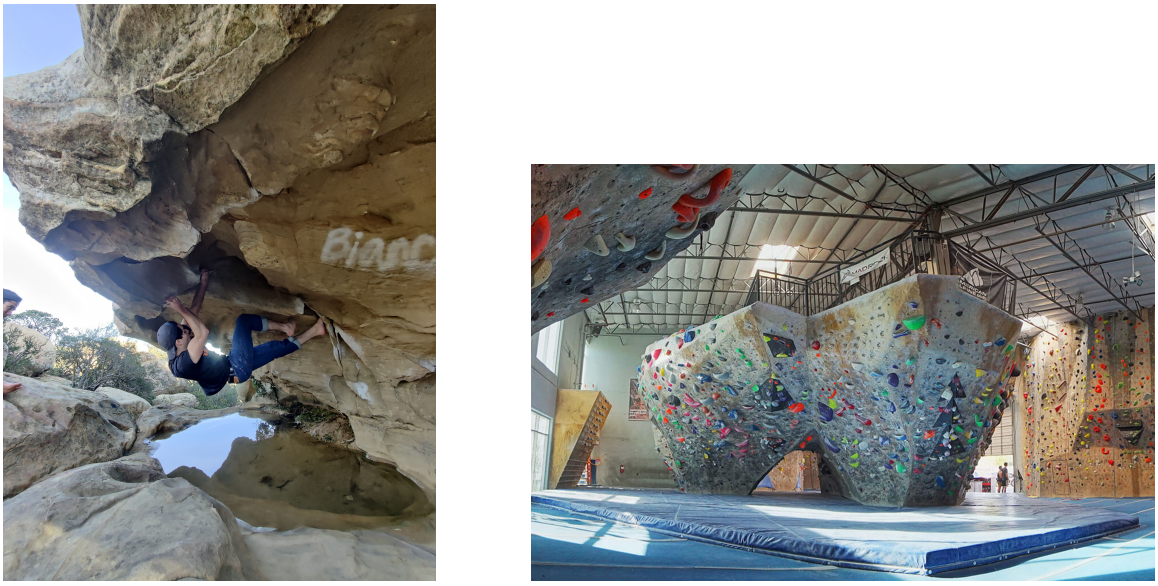


Figure 1: The researcher on an outdoor boulder (left) and an indoor bouldering wall [3] (right).

Then, in Section 5, we discuss the results of our algorithm from a QDA perspective, as well as the feasibility of routes as determined by the online survey and in person testing.

Our setup shows promise from a QDA perspective; generally, it finds routes with a variety of behaviors at each difficulty. However, our results from the online survey are inconclusive about the system's success, and in-person trials of routes indicate that they are generally significantly more difficult than intended.

2 Climbing

Here we give background information regarding various aspects of climbing.

2.1 Bouldering

In this project, we focus on the climbing style known as *bouldering*. Bouldering is done both indoors and outdoors on relatively short walls, usually not higher than 15-20 feet tall (see Figure 1). It mainly differentiates itself from other styles such as toproping and lead climbing by the lack of a rope. Due to the relatively short height, protection is provided by a crash pad placed at the bottom of the wall to soften the impact of falls. The lack of rope reduces the barrier of entry, making bouldering a low cost entry point to climbing. We select bouldering for our domain since these characteristics decrease its complexity for route generation, grading, and physical testing.

2.2 Grading

The difficulty of a climbing problem is called a *grade*. There are a variety of grading systems for each style of climbing (bouldering, traditional, etc) that indicate to a potential climber whether a given climb is below, within, or above their ability. Grades are an object of much controversy in the climbing community, mostly as a result of their subjectivity [7]. Typically, the first climber to ascend a route assigns it an initial grade based solely on their opinion. Then, as other climbers repeat the climb, they either confirm the route's current rating, or state that they believe it should be adjusted. Through this process, a consensus on the route's grade is eventually reached and canonized. The standard bouldering grading system in the U.S. is the *Vermin Scale* or *V Scale* which consists of grades beginning with the letter V followed by an integer between 0 and 17 (for example V5). A higher number indicates a more difficult climb. For example, a V7 climb is more difficult than a V6 climb, which is more difficult than a V5 climb.

2.3 Standardized Training Boards

A standardized training board is a small bouldering wall with pre-specified hold types and locations. Typically, each hold has an associated LED light positioned below it which, when turned on, indicates that it is a part of the route being used by the climber. A climber may use the hold if and only if its corresponding LED is on. These LEDs are controlled by apps which allow climbers and setters to enter routes then display them on the board. Examples of standardized training boards include the Moonboard [25] and the Tension Board [21]. For our experiments, we will focus on the MoonBoard (Figure 2) 2016 layout because that is the board we have access to for physical testing. Due to the nature of MoonBoard holds, grades on the MoonBoard are restricted to V4-V14.

2.4 The Route Setting Problem

Informally, the route setting problem involves finding a set of hand holds with designated start and ending holds of a prespecified difficulty. Here we define the problem rigorously:

Let η be a set of holds and π be a set of positions for those holds. Also, let $C = \eta \times \pi$.

Definition 1. *Call $p \in C$ a **position pair**.*

Position pairs are determined by setters when they place a given hold in a specific position on the wall.

Definition 2. *A **route** is defined as a 3-tuple (R, S, F) where*

- *$R \subseteq C$ with $|\{(h, p) \in R : p = p'\}| \leq 1$ for all $p' \in \pi$ are the position pairs,*

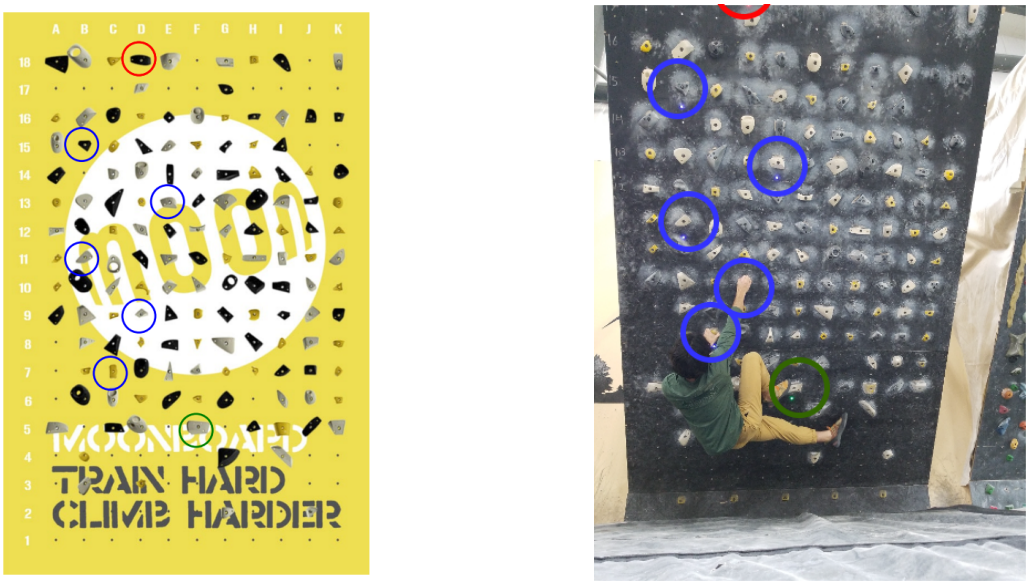


Figure 2: A MoonBoard route as viewed in the MoonBoard mobile application (left) and the researcher on the route (right). The start hold is circled in green and the end hold is circled in red. The climber may only use the circled holds while moving from the start hold to the end.

- $S \subseteq R$ *with* $1 \leq |S| \leq 2$ *are the starting position pairs, and*
- $F \in R \setminus S$ *is the terminal position pair.*

Our goal is ultimately to create these routes algorithmically based on a given grade.

Definition 3. *Let* $\mathcal{R} = \{R \subseteq C : R \text{ is a route}\}$. *A* ***difficulty function*** *is a function* $\delta : \mathcal{R} \to \mathbb{Q}$. *For a route* $R \in \mathcal{R}$ *we call* $\delta(R)$ *the* ***difficulty*** *of* R *and, for routes* R *and* R', *if* $\delta(R) > \delta(R')$, *we say* R *is more difficult than* R' *or, equivalently,* R' *is less difficult than* R.

Now that we have a defined route and a method of grading it we can properly describe our goal.

Definition 4. *The* ***route setting problem*** *is: Given* η, π, δ *as described and* $d \in \mathbb{Q}$, *construct a set of routes* S *such that* $\delta(R) = d$ *for all* $R \in S$.

In this paper, we consider a simplified version of the route setting problem for standardized training boards since they have fixed positions for each hold. Rather than taking η, π as inputs, we take a predefined set of position pairs C'. Thus we define:

Definition 5. *The* ***fixed-hold setting problem*** *is: Given* C', δ *as described and* $d \in \mathbb{Q}$, *construct a set of routes* S *such that* $\delta(R) = d$ *for all* $R \in S$.

In this research we attempt to solve the fixed-hold setting problem. Note that because position pairs are fixed tuples in the fixed-hold setting problem, for the remainder of the paper, we will refer to position pairs as *holds* to be consistent with climbing terminology.

2.5 Related Work

There are several attempts in the literature to assign grades based on a variety of computational techniques [5, 12]. Due to the aforementioned subjectivity of the process, many of these were unsuccessful. One approach that produced satisfactory results was Duh and Chang's GradeNet [8], a neural network to classify bouldering problems on a MoonBoard. The researchers trained the network on approximately 25,000 problems from the MoonBoard community database and the result had 85% accuracy when allowing ± 1 in the evaluated climb's grade (where grades range from V4 to V13). As far as we know, this is the most successful automatic route grading attempt for MoonBoards. In general, most human climbers will not agree completely on the grade of a route so, we consider this accuracy acceptable for determining routes of a specific difficulty. As a result, we select this as our grading mechanism (see Section 4.1).

GradeNet functions by first converting a climb to an anticipated sequence of moves where each move is encoded into a 22-dimensional vector with each element of the vector accounting for details on the move including the hold difficulty and placement. This set of vectors is then fed into a recurrent neural network that outputs a grade for the route as an integer between 4 and 13 inclusive. This integer output is the integer component of the V-grade described in Section 2.2. The range restriction of 4 to 13 is because the MoonBoard only allows grades between V4 and V14. GradeNet excludes V14 from the training because there are few routes and many of them were found to be low quality.

It is worth mentioning that are several prior projects dealing with the applications of artificial intelligence to grading climbing routes based on less subjective methods. One such method is the Whole History Rating system [18] which treats routes and climbers as adversaries and ranks them in a manner similar to Elo rating system in Chess [9]. When climbers successfully complete a climb, their rating increases and the climb's decreases by an amount based on the difference between their ratings. Scarff [18] had success with estimation of difficulty based on this system. However, it requires a large number of datapoints on what climbers climbed the route previously and what other routes those climbers ascended. This data requirement makes the approach impractical for our research on the generation of new routes for potentially unknown climbers.

There have been few previous attempts at solving the route setting problem procedurally. Those that do exist have significant drawbacks when it comes to scalability and objectivity. Stapel [19] proposed a heuristic-based greedy algorithm for the generation of routes on a MoonBoard. In general, their routes were completable, but often found relatively unsatisfactory by the climbers who tried them. Further, their approach relied heavily on human-designed heuristics such as having climbers rate how hard they thought specific moves were, rather than using a deterministic algorithm. Additionally, to find multiple routes, they

relied on a tree-based search approach. They acknowledge this approach creates a domain that is computationally slow to parse and infeasible when removed from the very limited domain of the MoonBoard. In contrast, by their nature, evolutionary algorithms excel at exploring large spaces making our approach more practically generalizable.

3 Evolutionary Algorithms

Evolutionary algorithms seek to simulate the Darwinian process of natural selection (c.f., e.g. [24]). They apply an iterative brute-force approach to discover a method to achieve a goal. In general, an evolutionary algorithm evolves a fixed number of generations. In each generation, they create a population of a given size bred from the previous generation, or selected randomly in the case of the first generation. They then take each individual in the population and apply a *fitness function* to it. After testing the fitness of all individuals, some combination of the most fit individuals are selected for the breeding of the next generation. After a sufficient number of iterations, there should be an individual with a high fitness score.

3.1 Quality-Diversity Algorithms

Quality-Diversity Algorithms (QDAs) are a variant of evolutionary algorithms that seek to discover a variety of high-fitness solutions to a given problem [15]. While traditional evolutionary algorithms are inspired by natural selection, quality-diversity algorithms derive motivation from the large number of ways in which animals accomplish tasks. For example, there are many different ways that sea creatures swim: Fish wiggle their bodies back and forth, while squids propel themselves by taking in and pushing out water. The diversity of solutions produced by QDAs is achieved by characterizing the performance of a specific set of parameters, called *elites* in the context of MAP-Elites (ME), based not only on their fitness, but also a variety of *behavioral descriptors*. In other words, rather than acting as a real-valued function mapping elites to fitness as a typical evolutionary algorithm does, ME maps elites to a multidimensional *behavior space* of which fitness is one dimension. The non-fitness dimensions of the behavior space, determined by the experiment designer based on what makes solutions distinct, are partitioned into equally-sized regions called *niches*. Rather than selecting for solely for the single most fit elite, ME selects for the most fit elite that falls into each of these niches. As a result, a desirable solution can be found for each niche in the behavior space. Since these niches are defined to be behaviorally disjoint, there should then be a diverse set of solutions to the problem. The set of niches is called an *archive*.

3.2 Related Work

QDAs are an active area of research and have been utilized successfully for a large number of applications including locomotion of tensegrity robots [6], developing gameplay strategy [17], and chemical design [23] amongst many others [13, 1, 10].

One analogue for the route setting problem is the task of difficulty-based level generation in video games. This field, generally known as procedural content generation (PCG) is vast and encompasses many techniques, but most relevant to us are attempts to use MAP-Elites or similar QDAs to generate unique levels with a specified degree of challenge. Here we explore and draw inspiration from a small number of these attempts.

Gravina et al. [11] originally introduced the idea of quality-diversity based PCG. They tested MAP-Elites on a variety of level generation tasks including the creation of diverse multi-dimensional objects and obstacles, the production of bullet-hell levels in which a large number of projectiles are fired at a player controlled character, as well as the generation of levels for *Super Mario Bros*, and decks for *Hearthstone* amongst several other tasks. They were ultimately able to successfully generate reasonable, playable content for a variety of different games.

Building off of Gravina et al. [11], several other authors progressed the use of QDAs, in particular MAP-Elites and variants of MAP-Elites, for PCG. Fontaine et al. [10] introduced a MAP-Elites variant to create decks for the videogame *Hearthstone* to support a wide variety of play styles. This paper is especially important as we base our mutation strategy off of their deck mutation strategy. Khalifa et al. [13] utilized MAP-Elites to create scripts for the bullet hell game *Talakat* that fired bullets at the player in diverse patterns. Charity et al. [4] utilized MAP-Elites in the development of a collaborative level design tool for the game *Baba is Y'all* in which humans in the game's community work with the QDA to design new levels for other players.

The details of many of these implementations follow a similar pattern so here we note some relevant takeaways reflected across the literature:

Remark 1. *For platformer and puzzle based games, typically the system is to use MAP-Elites to optimize the level in a game simulator running some variant of the A*-pathfinding algorithm. We started out along this path using an A* based climbing simulator [16] instead of GradeNet, but switched due to a variety of complications that arose with the simulator.*

Remark 2. *Another common theme is that a large part of the challenge of QDA based PCG is the selection of sufficient behavioral descriptors (see Section 4.4). We found it necessary to tweak our behavioral descriptors and partitions throughout our ongoing research. For example, we began this process thinking we would use average span*

rather than max span, but found that max span better represented the desired property of our climbs.

4 Methods

In this section, we discuss our methods including the design and hyperparameters for our algorithm, the design of the online grading survey, and the design of our in-person testing of routes.

4.1 Fitness Function and Grading

We grade routes using GradeNet. As the goal of our experiment is to create a route of a specific difficulty, we want our generated routes to converge to this difficulty. Thus, we minimize the distance between the target grade and the grade of the generated route by maximizing the following function. Given a target grade g and a route R, which GradeNet determines to have integer grade $grade(R)$, the fitness function is defined as

$$fitness(R, g) = \frac{1}{|g - grade(R)| + (0.01)(numMidHolds(R))}.$$

The $|g - grade(R)|$ term is the primary aspect that ensures we are minimizing the difference between the desired grade and the grade assigned to the route by GradeNet. The $(0.01)(numMidHolds(R))$ term was introduced after an observation that generated routes, especially at the lower grades, tended to have many extraneous holds that detracted from the climb. For example, there may a be a high-quality V4 route that contains an extra hold no climber would use and that makes no difference to difficulty. This hold's presence makes the route more confusing and less like a human-designed route. When a route contains many of these extraneous holds, the quality reduction effect is compounded. Thus, this term of the fitness function serves as a method to improve route quality. Without it, many grades would see archive saturation (no possible more fit solutions) quite early based on our results. Instead, a fitness function that continues to improve allows us to take advantage of additional computational power to produce better outputs by running for more iterations.

We use the reciprocal of the quantity we are minimizing so that we can treat it as a fitness and maximize on it. Another option was to use the additive inverse of our fitness function's denominator. Unlike some evolutionary algorithms, there is no difference in how these two options are treated evolutionarily since our selection and mutation strategy relies solely on the relative ranking of elites and not on the magnitude of fitness. This decision does, however, have some effect on the scales of our results and analysis.

Algorithm 1 mutateArchive

Input: An archive A to mutate and a population size n.
Output: A new population of n elites.

```
elites = emptyList()
for i ∈ [1..n] do
    elite = getRandomElite(A)
    k = pickK()
    indices = randomSampling([1..13], k)
    newElite = elite
    for index ∈ indices do
        newElite = swapHold(newElite, index)
    elites[i] = newElite
return elites
```

4.2 Parameter Space

A valid MoonBoard route has between 3 and 14 holds with a start hold or pair of holds in the first six rows middle holds in any of the rows except the top row, and a terminal hold in the top row. We assign each hold and pair of start holds an integer id. Our routes are then encoded as a 13 element vector. The first element represents a start hold or pair of start holds, so it accounts for up to two of the 14 possible holds. The second element represents the terminal hold. The remaining 11 elements represent middle holds and can have a either a hold value or a null value indicating there is no hold in that position. At each generation, MAP-Elites will produce a new set of elites consisting of these parameters based on the previous generation (or randomly in the case of the first generation).

4.3 Mutation Strategy

We use a k swaps mutation with geometric probability for k to construct each successive generation of elites. This technique is derived from the methods in Fontaine et al. [10]. For each individual in the new population, we select random route in the archive, then select a number k between 1 and 13 where there is a 50% chance of selecting 1, then a reduction of 50% probability for each successive possible k value until k reaches 13. Thus, the probability of selecting k is

$$P(k) = \begin{cases} \frac{1}{2^k} & \text{if } k < 13 \\ \frac{1}{2^{k-1}} & \text{if } k = 13 \end{cases}.$$

After randomly selecting k distinct indices, we swap them with a uniformly distributed random alternative from the possible hold/hold pair IDs and null to form the new route. This process is laid out in Algorithm 1.

Figure 3: The types of holds on the MoonBoard [2].

4.4 Behavioral Descriptors

We use two behavioral descriptors to classify the type of route we generated: variety of holds and maximum span.

1. **Variety of Holds**: There are several standard types of climbing holds that indicate the type of grip a climber uses to grab the holds. The MoonBoard hold configuration we consider (2016 with A, B, and Original School hold sets) has 3 types: pinches, finger pockets, and crimps (see Figure 3). We divide each of these categories into two classes: small and large. This gives us six total possible hold types. Our behavioral descriptor will then be the number of different hold types included in the climb. We omit the possibility of 1 hold type since by experimental observation, the algorithm was rarely able to find this type of climb at any grade. Thus our descriptor ranges from 2 to 6 and is partitioned into 4 sections.

   This descriptor is interesting because it separates climbs focused on specific grip types from those with a wider range of required grip techniques. The ability to construct routes with a variety of hold types is generally considered to be a strength of training boards.

2. **Maximum Span**: Another key characteristic of climbs is how spread out the climber's body is. Climbs where the climber must keep their limbs closer together require a different degree of balance and maintenance of distance from the wall than those which require the climber to stretch their body out as far as possible. To capture this aspect, we use the largest span between hands given by the moves generated by GradeNet during the grading process. This is divided into 1 foot intervals start with 0 and going up to 4.

With these partitions our archives will be 4x3 grids for a total of 12 niches.

4.5 Experiment Design

In our experiment, we apply MAP-Elites (utilizing the pyribs [22] library as a framework) with the k-swaps mutation described in Algorithm 1. Using the described setup, we run MAP-Elites with a budget

Algorithm 2 MAP-Elites Evolutionary Loop (Based on [15])

Input: A integer target grade g between 4 and 13, a population size n, and a number of generations G.
Output: An archive A where $grade(R) \approx g \; \forall R \in A$.

$A = emptyArchive()$
$P = makeRandomPopulation(n)$
for $generation \in [1..G]$ **do**
  for $R \in P$ **do**
    $diversity = getHoldDiversity(R)$
    $maxSpan = getMaxSpan(R)$
    $grade = runGradeNet(R)$
    $R.fitness = fitness(R, g)$
    $archive[diversity][maxSpan] = routeWithMaxFitness(R, A[diversity][maxSpan])$
  $population = mutateArchive(A, n)$
return A

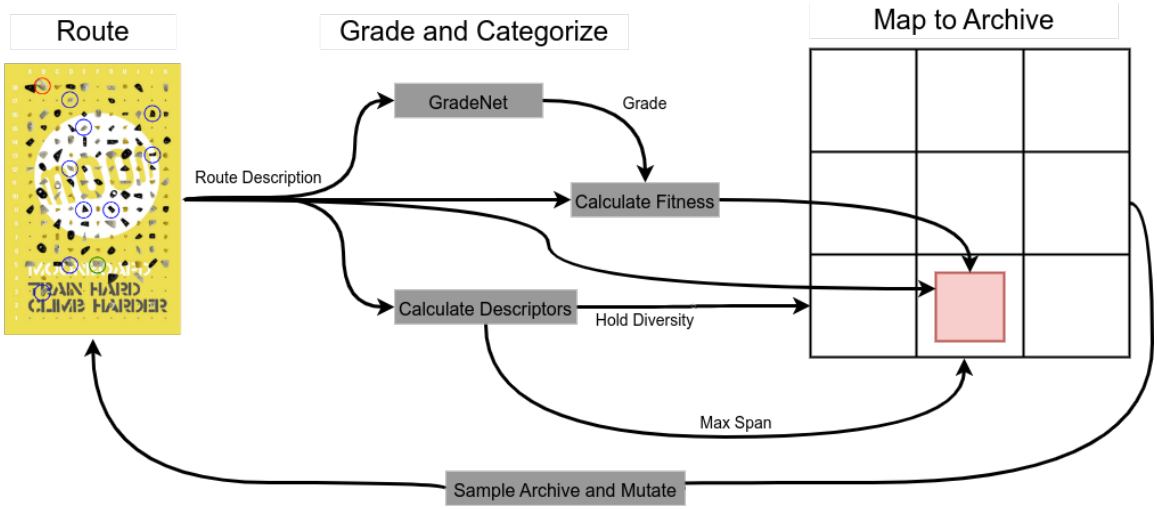


Figure 4: Evolution Cycle for Routes with MAP-Elites.

of 2000 iterations and 25 elites per generation. For the first generation, each route's vector is completely randomized. Then, we calculate each route's fitness and behavioral descriptors and place it into the archive accordingly. After that, we randomly select a route from the archive and mutate it. Selection and mutation is repeated 25 times to form the new population. Finally, the process repeats with assessment, selection, and mutation for 2000 generations. This experiment loop is also described in Algorithm 2 and Figure 4. At the end of the loop, we have an archive as described in Section 4.4 full of a diverse set of routes. We run this experiment 30 times for each grade to get a variety of solutions.

4.6 Human Testing of Routes

To evaluate the success of our algorithm in generating climbable routes of the correct difficulty, we ran two studies: an in-person study in which participants climbed generated routes, and an online survey in which

participants were asked to provide estimated grades based on images of the routes. We describe those processes here.

4.6.1 In-Person Testing of Routes

To check the validity of our generated routes in real life, we assess a climber's maximum grade by having them try a variety of well-established routes from the MoonBoard database known as *benchmark* routes. Each of these benchmark routes has a grade agreed upon by thousands of climbers in the MoonBoard community, and they are considered to be the standard for grading new routes. Climbers are given three tries to climb a benchmark from each grade, beginning with V4 and then increasing until they fail. Their maximum grade is the grade of the highest benchmark route that they successfully climb. After each climb, we ask the climber to provide their assessment of the climb's V-grade and a rating of quality from 1 to 5 with 1 being the lowest and 5 being the highest. After assessing this maximum grade, three generated routes of grade less than or equal to the maximum grade are randomly selected. For each of these three routes, the climber repeats the process of making up to three attempts on the climb, then reporting an estimated grade and quality score. Participants are not informed which routes are benchmarks and which are generated until after the study is complete, although it is possible that they may recognize some benchmark problems since these routes are publicly available on the MoonBoard app.

4.6.2 Route Assessment Survey

In an effort to reach a wider participant base, we also conduct an online survey using Qualtrics and CloudResearch in which users graded a set of routes. The survey contains one benchmark route and two generated routes of each grade from V4 to V13. We order the routes randomly for each participant. Then, for each route, participants provide an estimated grade (see Figure 5). Like the in-person testing, participants are not told which routes are benchmarks, although it is possible that they recognize some of them from climbing on the MoonBoard previously. We also collect the maximum benchmark grade which they previously climbed for further analysis. To screen participants and ensure they are familiar with climbing on a MoonBoard, we ask them to answer the following questions:

1. Do you have experience climbing on a MoonBoard with the 2016 hold set?
    - Possible Answers: "Yes", "No", or "Unsure". "No" or "Unsure" invalidate the result.
2. What angle (in degrees) was the MoonBoard you climbed on?
    - Possible Answers: Any text entry with any answer other than 40 invalidating the result.

Figure 5: An example question from the online survey. The parenthetical text next to the V-grade is the equivalent Font grade, a grading scale which might be more familiar to some participants.

3. Given an image of a route with one start hold, enter the coordinate of the start hold.
    - Possible Answers: Any text entry with any incorrect response invalidating the result.
4. Given an image of a route, enter the coordinate of the end hold.
    - Possible Answers: Any text entry with any incorrect response invalidating the result.

5 Results

Here we discuss the success of the algorithm from the perspective of theoretical QDA metrics as well as the results of our human trials.

5.1 QDA Results

In general, the experiment was successful by QDA metrics and we generated many routes of high fitness and the desired grade for every grade. We ran experiments 30 times and the results are contained in Figures 6 and 7. We examine two important metrics:

1. Quality-Diversity Score (QD-Score) is defined for an archive A and target grade g as

$$\sum_{elite \in A} fitness(elite, g).$$

   This represents the overall fitness of the archive. The logarithmic shape of these curves indicates that fitness rapidly grows early then eventually levels off.

2. Grade Difference is defined for an archive A and a target grade g as

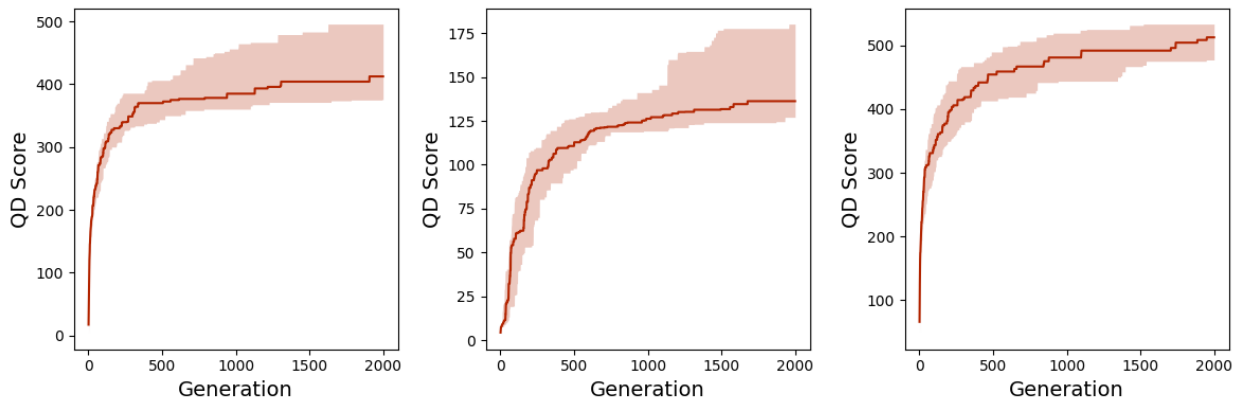
$$\sum_{elite \in A} (|g - grade(elite)|) + 9(numEmptyNiches(A)).$$

   This represents how close the archive is to having a climb of the desired grade in every niche. Empty niches are given the maximum possible distance from the target grade which is 9. We can see from the graphs that this reaches 0 for at least some archives at each grade, and all archives for most grades. In general this signifies that there is a climb of the correct grade in every niche in the archive most of the time. Also, these graphs show us that the difference often reaches 0 after only a couple hundred generations and from that point onward, the number of holds is being optimized.

These measurements show us that the algorithm is successful at finding routes of the correct grade, according to GradeNet, and all possible behaviors as defined by our descriptors.

It is interesting to note from the graphs however, that some grades appear more difficult for the algorithm to illuminate. In particular, V5 and V12 had much lower median fitness and took longer to get close to 0 grade difference, not fully reaching this mark for all 30 archives.

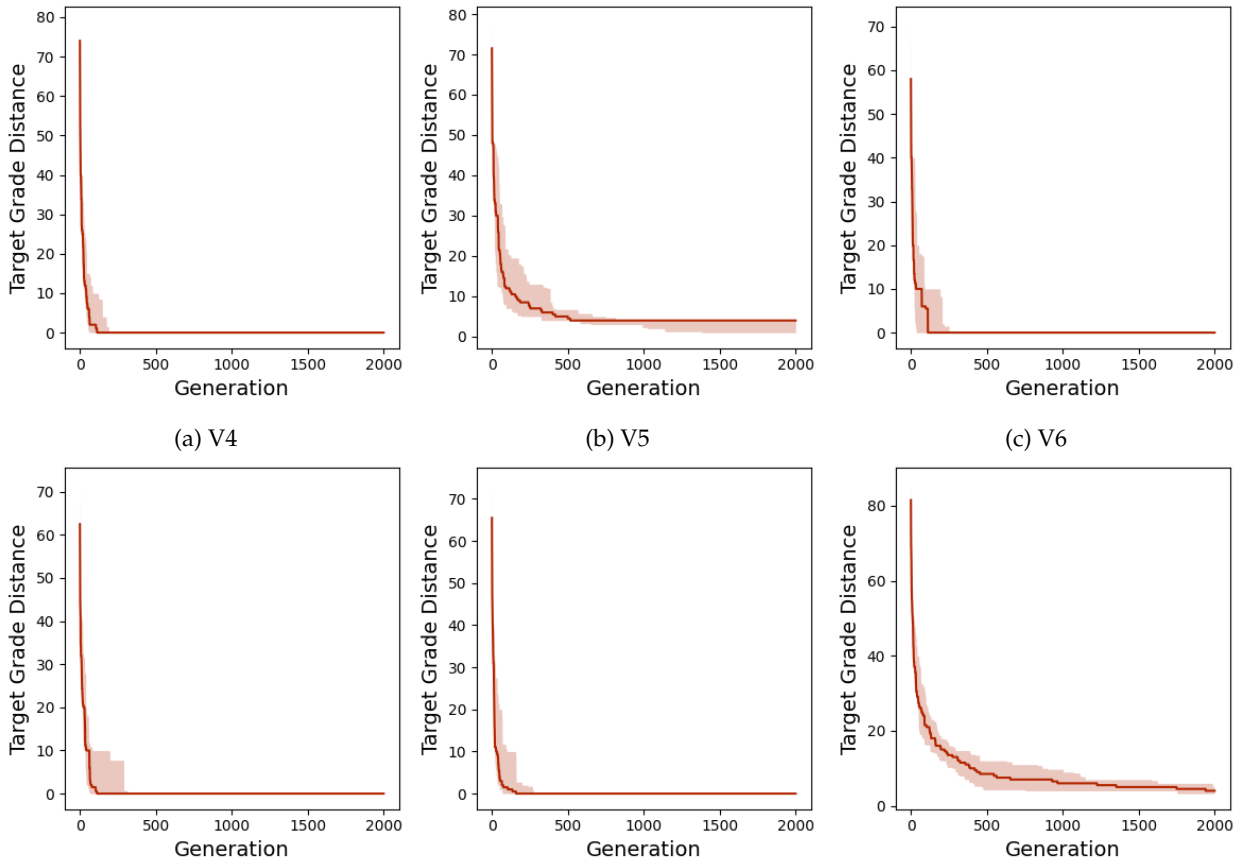
An explanation for this phenomenon may be available in the distribution that GradeNet assigns to random routes. We generated 10,000 random, valid MoonBoard routes and graded each of them using GradeNet. The results are visible in Figure 8. This graph clearly correlates with what we see in Figure 6. Looking at the distribution of randomly graded routes, we can see that very few routes of grades V5 and V12 are found. In fact, only 1 out of the 10,000 was graded V12 and only 17 were graded V5. This could be due either to a lower density of V5 and V12 routes in the space of all routes, or, to a bias in GradeNet. Either way, having far fewer routes of the correct grade via random generation means that it takes much longer to find a high fitness route in any niche when starting with a population of random routes. This directly translates into a longer evolutionary timeline and overall lower QD-score over time. Also, the difficulty in finding routes of this grade means we are less likely to get a grade difference of 0.



(a) V4 (b) V5 (c) V6

(d) V8 (e) V10 (f) V12

Figure 6: Quality-Diversity Score by generation for 30 repetitions of 2000 generations at different grades. The line is the median score of the thirty archives at each generation and the shaded region ranges the 25th percentile to the 75th percentile archive.



(a) V4 (b) V5 (c) V6

(d) V8 (e) V10 (f) V12

Figure 7: Grade Difference by generation for 30 repetitions of 2000 generations at different grades. The line is the median score of the five archives at each generation and the shaded region ranges the 25th percentile to the 75th percentile archive.

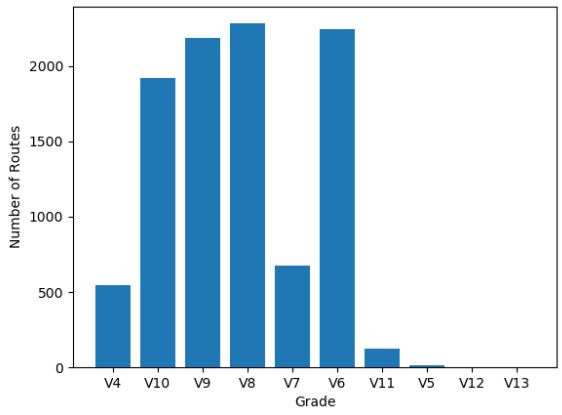


Figure 8: The grades of 10,000 randomly generated valid routes as assigned by GradeNet.

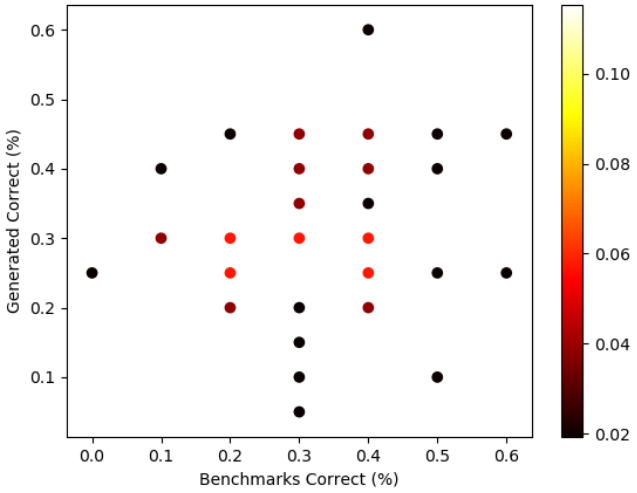


Figure 9: Data from online climber survey for all respondents. Color indicates frequency as a fraction of total responses.

5.2 Survey Results

We hypothesized that in our survey, there would be a positive correlation between the ability of participants to grade benchmark climbs and their likelihood to assign the algorithm's target grade to a route. Like GradeNet, we allow participants a ± 1 difference in grade prediction to be correct. Unfortunately, the results came back with no correlation between these quantities. We received a total of 832 responses to the survey. Of those, 52 are valid as defined in Section 4.6.2. From these results, we see no significant correlation between the prediction percentages. These percentages are displayed in Figure 9.

We also examine what happens looking only at climbers who climb higher grades to see if they are more successful at grading. Figure 10 shows the correlation (r) and statistical significance (p) as calculated by the Pearson-r test as we increase the minimum grade of participants. As can be seen from the correlation coefficients, these results remained inconclusive. Expanding the definition of correct to allow participants leeway of ± 3 V-grades, we do see a statistically significant positive correlation with $r = 0.32$ and $p = 0.02$. However, this would be expected with even random selection since it allows a range of 7 grades out of 10 total grades to be correct for any 1 target grade so it is not a particularly compelling result.

As our results are not significant, it is hard to draw any concrete conclusions that our system generates routes of the desired grade. Our results do however add support to the claim made by Duh and Chang [8] that GradeNet is more accurate than human visual assessment of routes since GradeNet grades at 85% accurate against human-graded MoonBoard database problems while our subject participants graded only 32% of benchmark routes correctly when allowing for ± 1 grade difference.

Minimum Grade Climbed	Sample Size	r	p
V4	52	0.14	0.31
V5	48	0.18	0.23
V6	41	0.14	0.37
V7	38	0.12	0.47
V8	25	0.15	0.48
V9	16	0.12	0.65
V10	10	-0.02	0.96
V11	5	-0.19	0.76
V12	3	0.50	0.67
V13	2	-1.00	1.00

Figure 10: Pearson r test values for online survey by minimum grade climbed by participant. The correlation coefficient is r and the statistical significance is p.

Climber	Calibration	R1 Int.	R1 Est.	R1 Qual.	R2 Int.	R2 Est.	R2 Qual.	R3 Int.	R3 Est.	R3 Qual.
1	V4	V4	V5	3	V4	V8	2	V4	V5	3
2	V5	V4	V5	2	V5	V7	3	V4	V8	3
3	V8	V4	V7	2	V7	V8	3	V8	V10	5
4	V5	V4	V6	3	V4	V5	2	V5	V6	5

Figure 11: Results of in person testing of routes. Each row shows a climber's calibration grade, then the intended (Int.) and estimated (Est.) grade and the quality (Qual. ranging from 1-5) for each route generated route (R1-R3) the climber attempted.

5.3 In-Person Testing Results

Four climbers of varying ability completed our in person tests. They completed maximum benchmarks ranging from V4 to V8. Their results are shown in Figure 11. Unfortunately, even the strongest climbers were unable to climb the generated routes at the lowest grade. In general, their comments reflect that some of the routes felt possible with more than three attempts, but they universally agree that generated climbs are harder than their intended grade, estimating them to be an average of 1.2 V-grades higher. Although it is hard to estimate a grade without climbing a route, this 1.2 average is relatively close to the ± 1 allowance for correct grading given to GradeNet. None of them estimate the grade of any route to be the grade intended by the algorithm. At the most extreme is a climber who estimates a generated route intended to be V4 as V8. Climbers also agree that while the holds selected by the algorithm generally felt right for the assigned grade, the positions in which climbers must put themselves in to transition between these holds is often awkward and extremely challenging making the routes more difficult than they initially seem.

Qualitatively, it is possible that with more experienced participants, or more available attempts, there would have been a greater success rate – it is hard to draw conclusions from a sample size of 4 that is not well distributed across skill levels. Some of the generated routes were given ratings as high as 5, the maximum rating, and one climber comments that some of the generated routes felt like real, human-set, MoonBoard routes that are out of their skill level.

5.4 Interpretation

While our algorithm appears successful from a QDA perspective, our human trials suggest that this was not the case. We believe this disconnect is due to the accuracy of GradeNet and our application of it. As mentioned previously, GradeNet is only accurate 85% of the time when allowing ± 1 in grading. This means that the climbs in the archive may not have been accurate to the grade despite GradeNet's verdict. Further, GradeNet's training data consists of human generated routes. When humans create routes, they implicitly take into account the route's "flow". That is, they pay attention to how ergonomic the movements are and consider the physical limitations of how climbers can position their bodies. As a result, GradeNet may not have learned to account for the flow of a climb. Thus, the unconventional approach of random hold selection by the algorithm may create routes that are not consistent with the neural network's training data, further reducing its ability to accurately assess grades.

6 Future Work

There are several possible future avenues for continued research. Most important would be the improvement of the algorithm to transfer to real life application. This could be accomplished in two ways:

1. The improvement of GradeNet's accuracy which is a current research project of the author of [19] who has promising early results in improving the network's capabilities. By improving the accuracy and capabilities of GradeNet, our fitness function would become more accurate.

2. The improvement of the fitness function to prioritize high quality routes. This is a difficult concept to quantify given how subjective and challenging it is to describe movements on the climbing wall. One possibility would be to train a classifier similar to GradeNet to give star ratings based on the user ratings of climbs in the MoonBoard database and incorporate its output into our notion of fitness. This would however, likely on be as successful and accurate as GradeNet, and does not overcome the issue of routes potentially not reflecting the training data. Another possibility would be to look into existing sports and/or videogame literature to look for metrics which might carry over into quantifying the quality of routes from a physical and mental perspective.

Assuming the successful improvement of the algorithm through the methods mentioned above, natural next steps would be to apply the algorithm to alternate MoonBoard hold setups as well as other standardized training boards such as the Kilter Board [14] and Tension Board [21].

7 Conclusion

In this paper, we discuss our experimental process and attempt at an automated solution to the Fixed-Hold Setting Problem using the MAP-Elites quality-diversity algorithm. We motivate the problem by discussing the current challenges in creating tailored routes. We discuss the basis that our methods have in both rock climbing and quality-diversity research, and how our project combines the two. After that, we introduce our experimental methods by which we link MAP-Elites and GradeNet to illuminate possible routes of a specific difficulty on a MoonBoard. Finally we discuss the results of our experiments, why we think they were less successful than desired, and lay out possibilities for future research in this vein.

8 Acknowledgement

This project received funding for the online survey from the Department of Computer Science at Union College.

I thank the fantastic professors of the Union College Computer Science department for their guidance and teachings over the last four years. In particular, thanks to professors Chris Fernandes and Nick Webb for their advisement in the early stages and formation of this project.

Special thanks to my advisor, Professor Matt Anderson, for his support, advice, and aid throughout the course of this research. His though provoking questions and comments on all aspects of the project, alongside his advice on overcoming each of the roadblocks that arose, were instrumental in its completion.

Finally, thanks to my family – Mom, Dad, Sarah, and Joshua – for their love and support throughout my college journey.

References

[1] Alberto Alvarez et al. "Empowering Quality Diversity in Dungeon Design with Interactive Constrained MAP-Elites". In: *2019 IEEE Conference on Games (CoG)*. 2019, pp. 1–8. DOI: 10.1109/CIG.2019.8848022.

[2] Behance. *Climbing+holds projects: Photos, videos, logos, illustrations and branding on Behance*. URL: https://www.behance.net/search/projects/?search=Climbing%5C%2BHolds.

[3] *California's most affordable indoor rock climbing gyms*. Jan. 2023. URL: https://climbhangar18.com/.

[4] Megan Charity, Ahmed Khalifa, and Julian Togelius. "Baba is Y'all: Collaborative Mixed-Initiative Level Design". In: *2020 IEEE Conference on Games (CoG)*. 2020, pp. 542–549. DOI: 10.1109/CoG47356.2020.9231807.

[5] Alejandro Dobles, Juan Carlos Sarmiento, and Peter Satterthwaite. "Machine learning methods for climbing route classification". In: *Web link: http://cs229. stanford. edu/proj2017/finalreports/5232206. pdf* (2017).

[6] Kyle Doney et al. "Behavioral repertoires for soft tensegrity robots". In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2020, pp. 2265–2271.

[7] Nick Draper. "14 Climbing grades". In: *The Science of Climbing and Mountaineering* (2016), p. 227.

[8] Yi-Shiou Duh and Ray Chang. "Recurrent Neural Network for MoonBoard Climbing Route Classification and Generation". In: *arXiv preprint arXiv:2102.01788* (2021). DOI: 10.48550/ARXIV.2102.01788. URL: https://arxiv.org/abs/2102.01788.

[9] A.E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., 1978. ISBN: 9780668047210. URL: https://books.google.com/books?id=8pMnAQAAMAAJ.

[10] Matthew C. Fontaine et al. "Mapping Hearthstone Deck Spaces through MAP-Elites with Sliding Boundaries". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. GECCO '19. Prague, Czech Republic: Association for Computing Machinery, 2019, pp. 161–169. ISBN: 9781450361118. DOI: 10.1145/3321707.3321794. URL: https://doi.org/10.1145/3321707.3321794.

[11] Daniele Gravina et al. "Procedural Content Generation through Quality Diversity". In: *2019 IEEE Conference on Games (CoG)*. 2019, pp. 1–8. DOI: 10.1109/CIG.2019.8848053.

[12] Lindsay Kempen. "A fair grade: assessing difficulty of climbing routes through machine learning". In: *Formal methods and tools, University of Twente* (2018).

[13] Ahmed Khalifa et al. "Talakat: Bullet Hell Generation through Constrained Map-Elites". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. GECCO '18. Kyoto, Japan: Association for Computing Machinery, 2018, pp. 1047–1054. ISBN: 9781450356183. DOI: 10.1145/3205455.3205470. URL: https://doi.org/10.1145/3205455.3205470.

[14] *Kilter Board*. URL: https://settercloset.com/pages/the-kilter-board.

[15] Jean-Baptiste Mouret and Jeff Clune. "Illuminating search spaces by mapping elites". In: *arXiv preprint arXiv:1504.04909* (2015).

[16] Kourosh Naderi, Joose Rajamäki, and Perttu Hämäläinen. "Discovering and Synthesizing Humanoid Climbing Movements". In: *ACM Trans. Graph.* 36.4 (July 2017). ISSN: 0730-0301. DOI: 10.1145/3072959.3073707. URL: https://doi.org/10.1145/3072959.3073707.

[17] Diego Perez-Liebana et al. "Generating Diverse and Competitive Play-Styles for Strategy Games". In: *arXiv preprint arXiv:2104.08641* (2021).

[18] Dean Scarff. *Estimation of Climbing Route Difficulty using Whole-History Rating*. 2020. DOI: 10.48550/ARXIV.2001.05388. URL: https://arxiv.org/abs/2001.05388.

[19] F.T.A. Stapel. *A Heuristic Approach to Indoor Rock Climbing Route Generation*. Jan. 2020. URL: http://essay.utwente.nl/80579/.

[20] Kirby Steckel and Jacob Schrum. "Illuminating the Space of Beatable Lode Runner Levels Produced By Various Generative Adversarial Networks". In: *arXiv preprint arXiv:2101.07868* (2021).

[21] *Tension Climbing*. URL: https://www.tensionclimbing.com/.

[22] Bryon Tjanaka and Matthew C. Fontaine. *pyribs: A bare-bones Python library for quality diversity optimization*. https://github.com/icaros-usc/pyribs. 2021.

[23] Jonas Verhellen and Jeriek Van den Abeele. "Illuminating Elite Patches of Chemical Space". In: (July 2020). DOI: 10.26434/chemrxiv.12608228.v1. URL: https://chemrxiv.org/articles/preprint/Illuminating_Elite_Patches_of_Chemical_Space/12608228.

[24] AR Wallace. "On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. III Tendency Var". In: *Depart Indefinitely Orig. Type J Proc Linn Soc Lond* (1858).

[25] *Welcome to training on the MoonBoard, climb on the same problems as other climbers from around the world.* URL: https://www.moonboard.com/.