AI, Labor, and Economy Case Studies

Zymergen

PARTNERSHIP ON AI

Contents

04 Observations 23

05 Reflections and implications 34

06 Appendix 39

Preface

Background and objectives

This case study is a part of a compendium of ongoing research by the Partnership on AI (PAI) investigating the impact of artificial intelligence (AI) technologies in the workplace. The objective is to illustrate the tradeoffs and challenges associated with the introduction of AI technologies into business processes. Through this series of case studies, we intend to document the different types of AI techniques implemented, as well as discuss the real-world impacts of AI on labor, the economy, and society broadly. Researchers often struggle to understand the economic and social consequences of AI and its wide-ranging implications for society. For instance, contemporary economists grapple with the question of why ongoing AI and broader digitization efforts have not yet produced clearly measurable productivity gains for the global economy.[1] At the same time, one major question for the public and policymakers has been AI's impact on the workforce, both in the changing nature of work and net job loss or creation. Our hope is to help ground the conversations around productivity and workforce impact in examples of real-world AI implementation while highlighting nuances across sectors, geographies, and type of AI techniques used. This case study specifically looks at how a biotech startup in the San Francisco Bay Area applies machine learning and automation as an alternative to conventional R&D and scientific experimentation practices.

Methodology

Subject organizations were recruited from a pool of 100+ candidates that 'AI, Labor, and the Economy' Working Group members submitted to the case study project team. The final set of organizations prioritized for study reflects a combination of their willingness to participate in the project and the intention to profile organizations representing a variety of sizes, geographies and industries. The following case study was developed over the course of five months in the summer and fall of 2018 to help answer the above objectives.[2] The methodology included interviews with a set of management stakeholders at the subject organization Zymergen, including three senior executives, one data scientist, two business development representatives, and four scientists at different levels of seniority. These managers were directly or indirectly involved in Zymergen's AI processes, including the implementation, operation, and use of genomic libraries; its machine learning-enabled experiment design engine; and its 'automated wet lab' [3] processes. The interview subjects include founders of the company, senior executives, and managers from the research, data science, and business development teams. As a result, the case study primarily reflects a managerial perspective, rather than the views of personnel working with these technologies (though management often speaks to and shares data about workers' perspectives and experiences).

Representatives from non-profit and for-profit organizations affiliated with the Partnership on AI's Working Group on "AI, Labor, and the Economy" supported the case study development by conducting interviews, drafting write-ups, and supplementing the case with external research or expert consultations on the industry or macroeconomic dynamics.

[1] For more, see "Is the Solow Paradox Back?", McKinsey Quarterly, June 2018

[2] Zymergen has raised capital from McKinsey & Company, a co-author of the case studies.

[3] A wet lab is a scientific laboratory designed to handle chemicals and avoid contamination, often built with specific equipment and requirements to reduce human contact with the chemicals. At Zymergen, the equipment and tools used in the automated wet lab include but are not limited to: liquid handling systems, robotic colony pickers, barcoders, acoustic dispensers, automated plate readers, robotic rule-based scripts, and systems or software used to operate this equipment.

Definition of terms

While we acknowledge that there is no consensus on the definition of terms such as AI and automation, we would like to explain how these terms are used in the compendium:

Artificial intelligence/AI is a notoriously nebulous term. Following the Stanford 100 Year Study on Artificial Intelligence, we embrace a broad and evolving definition of AI. As Nils J. Nilsson has articulated, artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment. (Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements,* (Cambridge, UK: Cambridge University Press, 2010).

Our definition of **automation** is based on the classic human factors engineering definition put forward by Parasuraman, Sheridan, and Wickens in 2000: https://ieeexplore.ieee.org/document/844354, in which automation refers to the full or partial replacement of a function previously carried out by a human operator.[4] Following Parasuraman et al.'s definition, levels of automation also exist on a spectrum, ranging from simple automation requiring manual input to a high level of automation requiring little to no human intervention in the context of a defined activity.

Explainable AI or **Explainability** is an emerging area of interest in communities ranging from DARPA to criminal justice advocates. Broadly, the terms refer to a system that has not been "black-boxed," but rather produces outputs that are interpretable, legible, transparent, or otherwise explainable to some set of stakeholders.

In this compendium, a **model** refers to a simplified representation of formalized relations between economic, engineering, manufacturing, social, or other types of situations and natural phenomena, simulated with the help of a computer system.

Analytics and AI techniques used

Zymergen uses a range of analytics and AI techniques for its experiment recommendation engine and data normalization and data-cleaning processes in its automated wet lab. These techniques include linear regression, polynomial regression, Bayesian hierarchical modeling, and convolutional neural networks (CNNs).

[4] Our definition draws on the classic articulation of automation described by Parasuraman, Sheridan, and Wickens (2000): https://ieeexplore.ieee.org/document/844354

01 Introduction

Introduction

Can a combination of AI, automation, and genomics solve challenging problems in biology and material sciences in completely new ways? The four co-founders of Zymergen, a company that aims to solve problems "beyond the bounds of human intuition,"[5] attempt to answer this question with a new approach to conventional R&D. The company is seeking to leverage automation and AI to develop microbial strains faster, more cheaply, and more consistently. In so doing, the company is working to build a more comprehensive understanding of the microbial genome to develop new materials and products for unmet market needs.

Six years into the startup's history, the company has made significant progress, yet it is far from reaching its ambition. The company provides a glimpse into the unique labor, economic, and organizational challenges of an "AI-native" company [6] — one that was founded to use machine learning and artificial intelligence technologies as a key differentiator.

This case study examines the costs and benefits of Zymergen's approach. Does applying machine learning and automation to scientific experiment design and execution deliver higher business impact than conventional approaches? Additionally, what are the labor ramifications of such an approach for the startup, its customers, and the broader economy?

[5] Company website.

[6] An 'AI-native' refers to a company that was founded with a stated mission of leveraging artificial intelligence or machine learning as a key enabling technology. 'AI-natives' can build infrastructure from the ground-up without the need to shift from legacy systems (e.g., on-premise to cloud-based storage).

1.1. Background on Zymergen

Zymergen is a private molecular technology company based in California's San Francisco Bay Area that specializes in the development of new material products and the improvement of existing products for pharmaceutical, agricultural, materials, personal care, and electronics. Zymergen uses machine learning and automated wet labs to accelerate and improve experiment design and execution cycles. Through these approaches, Zymergen aims to improve the economics of its clients in sectors that rely on fermentation as a means of production, such as in agriculture, industrials, chemicals, and pharmaceuticals. Zymergen also leverages its insights and platform to develop its own products for the electronics (films, coatings, adhesives), marine, and personal care industries.

The company was co-founded in 2013 by Joshua Hoffman (CEO); Zach Serber (Chief Science Officer); Jed Dean (VP of Operations and Engineering); and Richard Hansen (member of founding team, former employee). Since then, Zymergen has raised $574M from investors such as SoftBank Vision Fund, Data Collective, ICONIQ Capital, and McKinsey & Company (a co-author of the case studies report), among others. [7] Zymergen has a team of 700 to 800 full-time equivalent (FTE) [8] employees and is still growing rapidly to meet increasing customer demand.

Zymergen utilizes the hypothesis-driven scientific method, and has adopted a test-everything mentality with an atheoretic (not based on theory), data-driven approach to innovation. In this sense, the company integrates a more data-driven method to experiment design, yet still relies on the expert knowledge of scientists.

Zymergen was built with an AI-first mindset, presenting a clean slate of opportunities and challenges. Because of this, Zymergen hired employees who were already invested in its mission of leveraging AI and automation to advance biology. Its infrastructure was designed for AI and automated wet labs. For that reason, the need to shift from "the old way of doing things" does not apply to Zymergen, whereas other organizations adopting AI may struggle with such a transformation. As a Zymergen business development employee put it, "the utility of machine learning was a hypothesis on which the company was founded."

[7] During the time of writing the case study in fall 2018, the company had raised $174M. On December 13, 2018, the company announced a $400M Series C round from multiple investors. See coverage of the announcement on Bloomberg and the Wall Street Journal.

[8] FTE refers to full-time equivalent employees at the company. The term is a business acronym and not intended to be a reductionist or derogatory term, but is a conventional unit of measure to compare workloads across different business contexts.

1.2. Zymergen's approach to biotech and molecular technology

Zymergen distinguishes itself with three proprietary assets and approaches (See Exhibit 1):

Genetic library

Zymergen is building a genomic library currently composed of two terabases[9] of physical and digital DNA data. The dataset is the foundation for a genomic "search" platform that is enabled by machine learning.

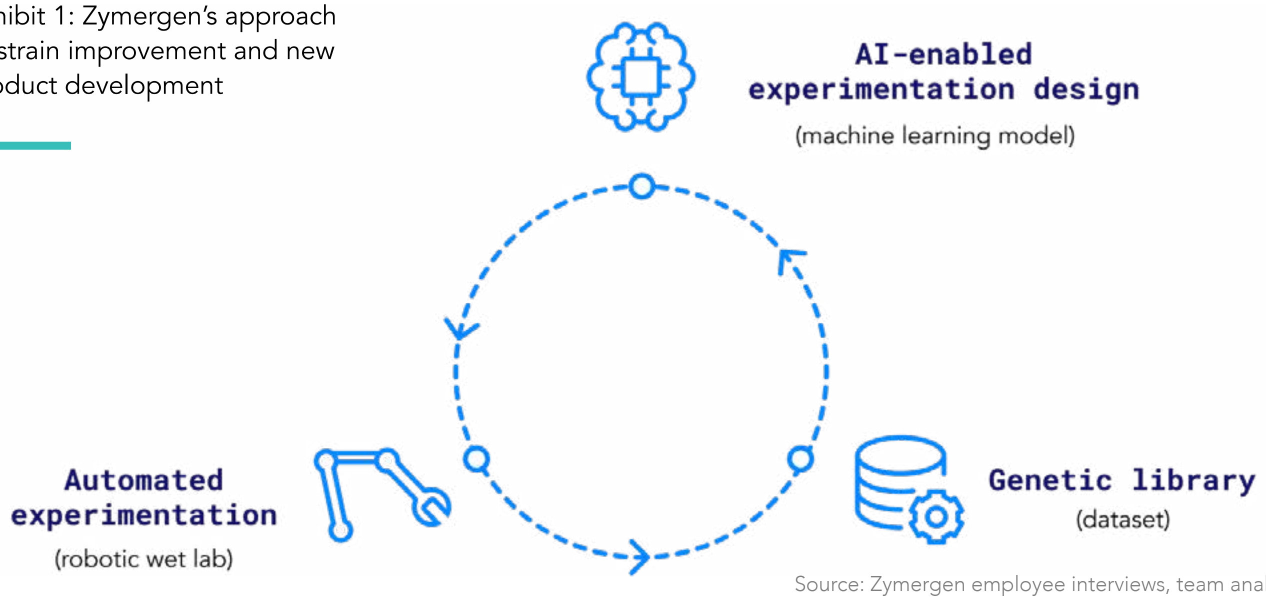
AI-enabled experimentation design

Zymergen leverages data science and machine learning techniques to improve experiment design and the iteration process.

Automated experimentation

Zymergen has a highly automated wet lab that uses robotics and sensors to improve accuracy, reduce human errors, and increase speed compared with traditional human-operated wet labs.

Exhibit 1: Zymergen's approach to strain improvement and new product development



Source: Zymergen employee interviews, team analysis.

[9] A terabase refers to genetic sequence data equivalent to 1012 base pairs.

Zymergen has two main lines of business (See Exhibit 2):

Internal product development

Internal product development accounts for approximately 20 percent of projects. Zymergen leverages its genetic library and genomic search platform to develop new products internally. Zymergen aims to directly commercialize new products such as new materials (e.g., specialty polymers using monomers produced from genetically-engineered, single-cell microbes) with a range of potential applications (e.g., electric vehicles, consumer electronics).

External services

Client service programs account for 80 percent of Zymergen's projects. Zymergen leverages its genomic model for microbial strain[10] optimization across clients in agriculture, consumer electronics, and industrial chemicals, among others. Strain optimization involves experimenting with different variations of a microbial strain to achieve certain objectives in large-scale fermentation (e.g., a strain that entails reduced sugar consumption per unit output of product). Zymergen's external services program aims to improve economics, accelerate commercialization, or develop entirely new products for clients.

Exhibit 2: Zymergen has two core lines of business with differentiation through proprietary data, machine learning models, and automated wet labs

LINES OF BUSINESS	
PRODUCT DEVELOPMENT (INTERNAL)	SERVICES (EXTERNAL)
Zymergen leverages its **genetic library and genomic search platform** to develop new products internally Aim is to commercialize **new products (e.g., specialty polymers using monomers produced from genetically engineered, single- cell microbes)** with a range of potential applications (e.g., electric vehicles, consumer electronics)	Zymergen leverages its global genomic model for strain optimization across **clients in agriculture, pharma, and industrial chemicals, among others** Aim is to **improve economics, accelerate commercialization, or develop entirely new products** for the client

Source: Zymergen employee interviews, team analysis.

[10] A strain is a genetic variant or subtype of a microorganism (e.g., virus or bacterium or fungus). These strains can be of use in large-scale fermentation processes.

02 Industry context and Zymergen’s value proposition

2.1. Background on Zymergen's customers

Zymergen's customers operate in a variety of different sectors, yet they often have similar scientific and business challenges. The clients are typically Fortune 2000 companies in the agriculture, electronics, chemicals, animal nutrition, or food ingredients product sectors.

Zymergen's core customers often use fermentation — a global market estimated to be roughly $150 billion in 2016[11] — in the material development and manufacturing process. The fermentation production process is a highly competitive business, with high-volume products, very thin margins, and highly capital-intensive operations. These businesses often have R&D departments focused on identifying economic savings through improving microbial strains to optimize for yield or energy usage in the downstream production of certain outputs. Optimizing strains is a key route to improving the economics of downstream production, and therefore, customers of companies like Zymergen are willing to invest heavily in multi-year R&D efforts to improve a microbial strain for large-scale production. As one Zymergen representative explained, these companies are typically making large-scale products "where a small saving per unit can have a very large impact."

For example, a small improvement in microbe efficiency can lead to large impact on a company's bottom line: "Making these microbes more efficient leads to reduced sugar consumption per unit output of product," the business development representative said. "Dropping the sugar consumption by a little can have an outsized impact when the production scale is massive, and the margins are thin in a commodity product. If you can save $10-20/ton [of product output] on sugar costs and you are producing 1M tons [of product] per year, it becomes a very compelling business case."

Customers could realize economic benefits in two ways:

- **Input material productivity gains through yield improvements (operating expense savings):** Strain improvement programs could reduce production costs though realizing savings on inputs (e.g., sugar, feedstock). A company could achieve the same level of production output in the fermentation process with a lower consumption of inputs, leading to savings in the form of lower operating expenses (OpEx).
- **Capital productivity gains (capital expense savings):** The company could also increase the production rate of strains, enabling it to produce more volume in shorter periods of time with the same inputs. Increasing the production rate of strains is a less capital-intensive route to increasing production volumes compared with alternatives. By increasing the efficiency of the production assets (e.g., manufacturing plant, fermentation tanks), the company could realize capital expense (CapEx) benefits through forgone investments in new production facilities, e.g., fermentation tanks, while still adding production capacity.

The natural trade-offs in the strain improvement business (e.g., improving economics of existing strains vs. investing in new plants and machinery for additional capacity) lead to broader economic consequences and labor implications across the manufacturing value chain:

- **Upstream impact:** Investment in strain improvement programs could lead to improved economics for manufacturing (e.g., higher yield). This includes the R&D process (e.g., strain improvement program), which may use internal or third-party R&D services such as Zymergen.
- **Downstream impact:** Strain improvement programs could result in lower demand for inputs in manufacturing (e.g., raw materials, labor, energy) or greater capital efficiency (e.g., forgone investment in new manufacturing facilities, equipment, and associated supply chain labor).

[11] "Fermentation Products Market by Type - Global Opportunity Analysis and Industry Forecast," Allied Market Research, June 2017.

2.2. Conventional approaches to strain improvement and Zymergen's differentiation

Traditional strain improvement programs use companies' internal R&D teams to improve strains over a period of years: It can take eight to ten years to manufacture and commercialize a particular strain. These strain improvement programs have traditionally relied on a combination of two conventional approaches:

1

Hypothesis-driven experimentation

In this approach, an R&D team composed of Ph.D. scientists and research associates (RAs) conducts strain-specific research to identify ways to create a desired variant (phenotype)[12] of a strain, often informed by related academic research. The research and individual expertise informs the creation of a hypothesis, which is then tested by RAs who execute the experiments manually in a lab (e.g., using hand pipetting). While the hypothesis-driven experiment design primarily relies on data from previous experiments and related academic research, there are still elements of human intuition introduced by the scientists in honing in on hypotheses.

2

Mutagenesis

This is a process by which a microbe's DNA is altered through spontaneous mutations. While this can occur naturally, it is also used as a laboratory technique whereby DNA mutations are intentionally induced to produce certain proteins or changes in strains of an organism to achieve desired improvements (e.g., more product per unit sugar consumed in fermentation).[13] A research team studies what inducements lead to improvements in strain properties, and whether the changes have unintended consequences. For example, while a change could mean the process uses less sugar, it may also take longer to produce the desired product.

[12] The composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical or physiological properties, behavior, and products of behavior.

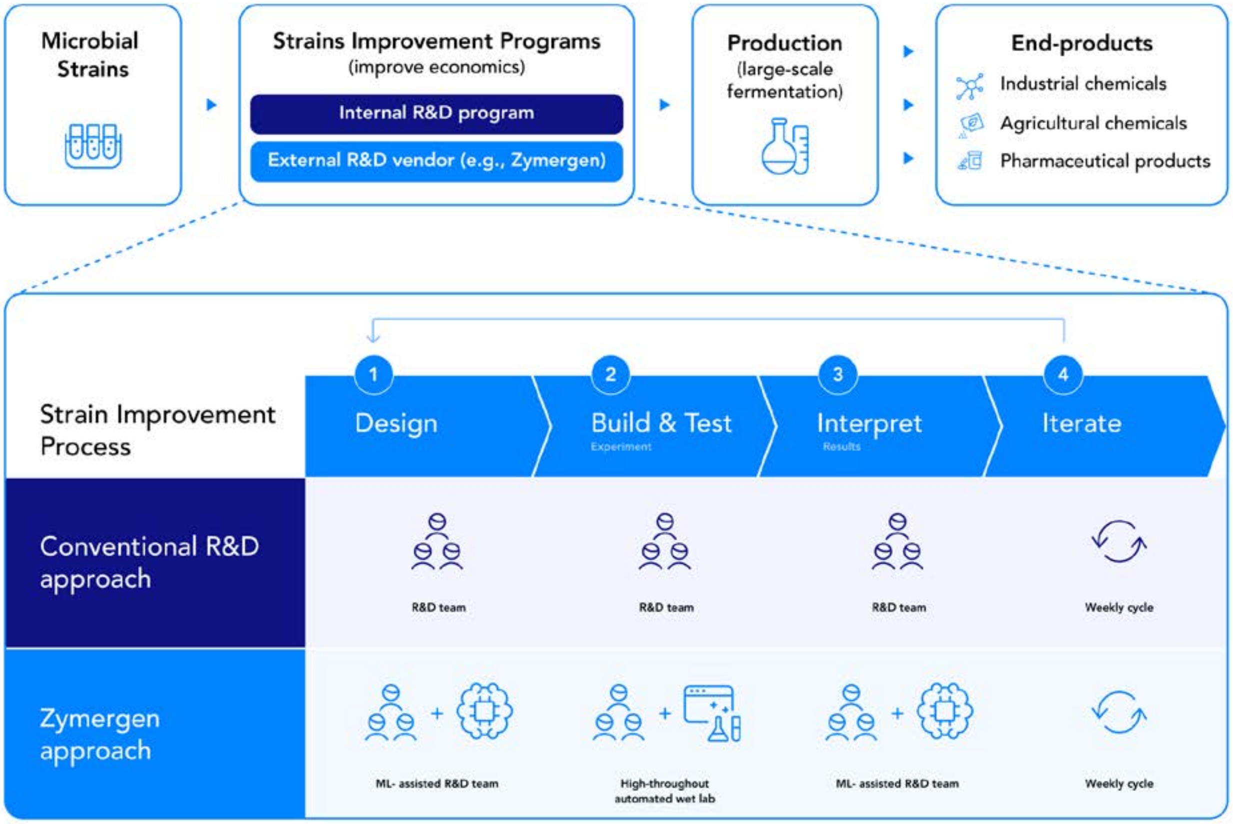
[13] Fanli, Zeng (2017). "Multiple-site fragment deletion, insertion and substitution mutagenesis by modified overlap extension PCR". *Biotechnology & Biotechnological Equipment.*

Zymergen's approach introduces a third means of strain improvement - often viewed by clients as an option to further improve the economics of an existing strain or accelerate time-to-market for projects early in development (See Exhibit 3):

3 Machine learning-enabled, atheoretic strain improvement

This approach uses data science to prioritize among experiment designs and to build a more comprehensive understanding of the genome. Nonetheless, Zymergen does not exclusively approach strain improvement through an atheoretical approach (also referred to as hypothesis-agnostic) but rather uses it in tandem with a hypothesis-driven approach (as described above), bringing in the deep expertise of the company's scientists. Zymergen finds that this combination of human-driven and machine-driven approaches leads to a targeted yet more systematic alternative to conventional methods.

Exhibit 3: Overview of Zymergen's place in the value chain



Source: Zymergen employee interviews, team analysis.

2.3. Zymergen's value proposition to its customers

Zymergen reports that most customers use its services because of its innovative approach to scientific experimentation (e.g., heavily leveraging automation and machine learning). Companies with mature products, such as citric acid, lactic acid or amino acid, usually consider Zymergen after identifying about 80 percent of potential strain improvements through conventional approaches internally. Thus, they are reaching diminishing returns — or no returns — with their internal strain improvement processes; they might then engage with Zymergen to capture the last increments of improvement.

The conventional hypothesis-driven methodology "works quite well, until a certain point. The system is so complex that you reach a limit or ceiling of what you can do using human intuition," a Zymergen business development representative said.

Clients may also engage Zymergen for new product development, rather than improving existing strains. Zymergen's approach can generate data from experiments (through automation and automatic data capture) and process data (through data science and analytics) at a faster pace and therefore can uncover patterns faster or more reliably than a human can. Nonetheless, humans are still an integral part of Zymergen's R&D process, especially in earlier stages of discovery where data coverage[14] is limited.

[14] Data coverage refers to the quantity of data available compared to what may be required for a sufficiently predictive model. In earlier stage projects, the quantity of data may be limited as fewer experiments have been run for the phenotype.

03 Opportunities and challenges for Zymergen as an ‘AI-native’ company

3.1. Challenges in molecular biology R&D and opportunities for Zymergen

Zymergen was founded to solve fundamental challenges in biology and material science by leveraging machine learning, sophisticated data infrastructure, and an engineering mindset. Because the company is AI-native, it makes conscious design choices around its data infrastructure and data sciences methodology, allowing it to tackle challenges in novel ways. Zymergen has opportunities to address several types of challenges in the field:

3.1.1. Searching a complex genomic space that is largely unknown

The genomic space is complex and largely not well understood.[15] A business development representative compares it to the game Go, which has simple rules but a huge number of possible moves: "The genomic space is vast. There are 3,000 genes in a simple organism. The number of total interactions across the genes and the number of permutations is close to the number of moves you have in Go" (See Exhibit 9 in appendix). Additionally, he said, the fermentation process is complex: "This is not just a single cell fermentation, but also how that cell is affected by the billions of cells in a fermentation tank. There are gradations of temperature, different atmospheric pressures at various levels of the tank — and this is happening over the course of four to five days."

Zymergen's approach is to catalog all the potential perturbations that could occur in the microbe, akin to how a search engine attempts to catalog the web. A scientist could then "search" the genome, but instead of finding the optimal restaurant for a given set of criteria (e.g., downtown San Francisco, 4-star rating, and Chinese cuisine), they could find the optimal microbial strain for a given set of criteria (e.g., X strain and interaction, more product per unit sugar consumed in fermentation). This is done by using data science and machine learning techniques to "recommend" strains. AI is used as a prioritization tool: "The design space is close to infinite and we are just scratching the surface, so the AI helps prioritize and focus our next set of experiment designs," said a business development representative. "It often brings in changes you never would have found or thought of otherwise. It is also faster and more resource efficient."

[15] The genomic space refers to the size of the problem space — the complete genome, in Zymergen's case, the microbial genome (vs. human genome). The size of the human genome, for example, is over 3 billion base pairs residing in 23 pairs of chromosomes within the nucleus of human cells (or $4^{3,200,000,000}$ in Exhibit 9). The microbe genome is considerably smaller than the human genome yet still large. See "The Human Genome Project Completion,"National Human Genome Research Institute (NHGRI); and, "Microbial genomes,"University of Leicester.

3.1.2. Improving the reproducibility and consistency of experimentation

Despite the emphasis in science on repeatability and reproducibility, scientists often struggle to reproduce the results of others' experiments. "Two scientists with similar experience levels may have five to ten percent variance in results for the same experiment, which seems minor, but when doing this at multiple iterations, the variation becomes significant," a business development representative said.

Zymergen's approach is to leverage robotics and advanced machinery to automate the wet lab. Using robotics for experimentation makes the results more reproducible through enhanced standardization. Automated workflows implement standardized processes, reducing unwanted variations between trials. Zymergen uses data normalization to restructure large datasets to improve quality and reduce redundancies. Robotic wet labs also have vastly higher throughput than human scientists running experiments. Zymergen reports that it can run 1,000 experiments per week, whereas a human-operated lab could conduct between 10 and 100 per week.

3.1.3. Capturing data in a systematic and comprehensive way

Infrastructure in conventional labs often has not been designed to capture data. Small variations across experiments (e.g., humidity, measurement of liquids) can have a dramatic influence on the reliability of the results and thus create challenges in interpreting "noisy" data." There is often systematic bias and noise. People often reach wrong conclusions because the methodology is inconsistent across experiments or there is bias in the approach," a co-founder said.

Because Zymergen's wet lab and data infrastructure have been designed for data capture, its research teams capture more data points than in traditional experiments — 3.5 million data points per week, a business development representative said — and the data is higher quality and more standardized than in conventional labs. The hope is that this will lead to better results with the company's machine learning models.

3.2. Technical opportunities and challenges

3.2.1. Integrating machine learning to improve experiment design

Zymergen integrates machine learning and analytic techniques in multiple areas. Two of them stand out as being most beneficial and are widely used within the company[16]:

1 Experiment recommendation engine

There are thousands of experiments a scientist could undertake for a given strain. Zymergen's recommendation engine — referred to internally as the Consolidation Recommender — helps scientists prioritize which experiments to execute first. Using techniques including polynomial regressions, convolutional neural networks (CNNs), and representational learning methods to rank and order experiments, the recommendation engine makes suggestions to scientists in a way similar to online product recommendation engines. The predictive power of the engine is driven by the long-term memory of the models: The engine uses all the data it has collected over time, not just the most recent, as humans tend to do.

2 Data normalization and cleansing

With thousands of experiments run each week, there is a high volume of data with many influencing variables on an experiment's results.[17] In other settings, scientists and research associates can spend hours cleaning data for interpretation, or the data can have inherent biases from small variance in experiment execution (e.g., position of the well plate, instrument(s) the samples run on, experiment batch). As a co-founder notes, "biology is intrinsically noisy. This simple application has been extraordinarily important for us." The data normalization application – the first application the data science team built – applies Bayesian hierarchical modelling[18] for normalization. Normalization is a key data transformation step to improve quality of the machine-learning recommendation engine, while also saving time and resources.

[16] Other examples include computer vision to identify viable cell colonies on agar plate (e.g., petri dishes) with desirable properties that can be picked and placed in microtiter wells.

[17] Data generated through experiments in the lab, using historical experiments to normalize, does not yet include sensor data from the labor for exogenous variables (e.g., humidity in the lab, lab technicians) but this is planned to be included in future iterations.

[18] While not considered machine learning, Bayesian hierarchical modelling is a form of statistical modelling to make scientific inferences on a specific object or populations based on multiple observations. In simple terms, the approach tries to answer "what is the probability that X is the true value given the current data?" In hierarchical modelling, this concept is applied to multiple previous observations of similar objects (e.g., experiments). For more, see this lecture from Angie Wolfgang (Penn State).

Zymergen's researchers have also learned some important lessons about using AI to design experiments:

- **Advanced AI and analytical techniques:** More advanced techniques do not always produce better results. Simple models are often "easier to explain [to scientists] and interpret and therefore are easier to implement," said a data scientist. On the other hand, "more complex models didn't improve accuracy enough to justify the costs in time, money, complexity, understanding and buy-in." More advanced techniques may become more useful as data availability increases; the volume of data required for a deep learning model is much larger compared to a linear regression. Data is also required across each strain at a certain level of depth to derive valuable insights, rather than in aggregate across the whole genome. Though Zymergen has a vast genomic database (two terabases), it does not (yet) have enough data coverage to achieve all that it wants to achieve.

- **Long-term memory:** Arguably, what is more valuable than the methodology used is the "memory" of datasets. Human scientists tend to exhibit a recency bias when designing and picking experiments. "AI has found opportunities for us in genome engineering based on experiments we ran a year or more prior – AI is the elephant that never forgets," said Zymergen's chief technology officer. "Humans were looking mostly at results from the previous three to six months. Even memory notwithstanding, humans mostly work in tools like spreadsheets, and most spreadsheets become unworkable after a thousand or so rows. So even with data-assisted tooling, humans tend to cull the data to more recent or data they perceive a priori as 'more valuable.' AI finds the diamonds in the rough."

- **Unexplainable models:** Given Zymergen's approach of receiving recommendations from an algorithm, scientists often experience occasions when they are unsure of the reason a particular experiment was recommended. Some describe this as the "black-box" phenomenon.[19] For scientists at Zymergen, explainability of models was initially a challenge. "Scientists are the hardest customers to work for because they want to know all the ins and outs of the model," a data scientist said. The culture has shifted over time as scientists have seen the recommendation engine deliver successful results and thus have become more comfortable with the culture of increased experimentation. For some stakeholders, such as P&L owners at Zymergen's clients, the impact on the bottom line (profitability) is more important than being able to fully understand the strain improvement process from a scientific perspective. "We are able to show them the improved economics, not through theory but through empirical data. This is what is critical," a co-founder said.

Strain improvement: A false dichotomy

To test the effectiveness of Zymergen's AI recommendation engine, Zymergen pitted its AI recommendation engine against one of its human scientists. Both the AI and human scientist suggested 100 strains for a major client program, which were built and tested in the wet lab. The AI recommendation engine's suggestions showed higher average improvement, and the AI engine also suggested the single strain with the highest absolute improvement. Yet in practice, the comparison between AI and humans is not so simple; humans remain vital to Zymergen's AI recommendation today.

[19] In general, a black box system references a system in which only the inputs and outputs are visible; what goes on "inside" the black box is unknown or not easily explained -- the causes for this opacity vary, and may be due to technical or proprietary characteristics of the algorithm. When the inner-workings of a system remain unknown, this can raise issues of trust, transparency, fairness, and accountability.

3.2.2. Leveraging automation to improve experiment throughput and productivity

Advances in robotics and computing have given rise to high-throughput screening, or wet lab automation, in which researchers can execute greater quantities of experiments and at a lower cost per experiment. High-throughput screening uses computer software, robotics such as acoustic dispensers and liquid handling systems, and automated plate readers. This type of automation is becoming a preferred method for labs in pharmaceuticals, biology, and chemistry. It is also used for scientific experimentation in drug discovery. Zymergen reports that this approach has led to significant productivity gains, with Zymergen having seen output roughly 10 times higher than conventional R&D labs.[20]

One limitation of high-throughput screening is that it requires significant capital investment. Many academic labs or smaller corporate R&D labs are unable to purchase the necessary equipment, so many may outsource these services to companies that do have high-throughput screening labs, such as Zymergen.

For Zymergen, this increase in productivity has meant changes in work assignments, responsible tasks, and skill sets for the people working in the lab, particularly the research associates. Research associates at Zymergen may have a master's or bachelor's degree in biology, chemistry, or even computer science. In many conventional labs, RAs are responsible for the "manual" work to conduct an experiment. In the high-throughput environment, RAs may be responsible for preparing specifications for experiments to send to the wet lab factory (instead of executing the experiments); ensuring quality control for experiments; post-processing of large datasets of experiment results; and even managing partially automated workflows. The RAs are not responsible for the manual execution of experiments at Zymergen, but instead offload this work to the automation systems and a "factory team" of lab technicians to run the experiments in the robotic wet lab.[21]

The changes in work can extend to the more educated workforce as well. By offloading more mundane tasks to automation and robotics, Zymergen's proposition is that scientists (Ph.D.s) should have more time and opportunity to work on higher-value work, leading to higher net value created." If you hire a Ph.D. to do miniprep,[22] it is not the best use of their time, and it is actually performed better by robotics," a co-founder said.

3.2.3. Building a genetic library dataset and infrastructure to learn across experiments

Zymergen has designed its wet lab to optimize for consistent and automatic data capture to build up a better understanding of the microbial genome. Today, its proprietary dataset includes two terabases of genomic data. Robotics machinery has been designed to capture all types of data that could influence experiments, such as humidity in the lab or the relative position of well plates. The company has also designed automated workflows so that data are captured and structured in a convenient manner for post-processing by scientists or machine learning algorithms. Zymergen reports that the data it has generated and recorded -and the infrastructure it has created to capture it in an organized way - represent one of Zymergen's chief advantages.

[20] Zymergen employees have reported that the robotic wet lab can conduct roughly 1,000 experiments per week (output measured as unique phenotypes in this case), while typical R&D labs without the automated wet lab machinery typically can conduct roughly 100 unique phenotypes, although output can vary by lab.

[21] The "factory team" refers to lab technicians who operate and monitor the automated wet lab. While the lab is heavily automated, some technicians are still needed to operate the lab machinery, execute orders from core research teams, and monitor the lab

[22] Miniprep is a term used to describe a form of plasmid preparation for DNA extraction. Miniprep is one method used to purify plasmid DNA from bacteria. Miniprep is for a small-scale DNA yield, ~50 to 100 µg. The process can be done manually on paper; however, machinery has been developed to automate this process for increased throughput and consistency (Bouchard, Roland; et al. (2010). *Laboratory Methods in Microbiology*. Universal Scientific. pp. 119–126.*)*

04 Observations

4.1. Labor implications for Zymergen and its customers

4.1.1. Overview

Zymergen's business model has implications for both its own labor force and that of its customers. Because Zymergen is relatively young, the effects of its business model on internal hiring are clearer than the effects on hiring at its customers', which have had a limited time to adjust to Zymergen as a market participant. Nonetheless, a look at both types of labor effects gives a sense of the direction hiring may be heading – and raises interesting questions about the future of the workforce more generally.

For Zymergen's own highly educated (a third have Ph.D.'s) workforce, there have been no reductions in total workforce due to automation and machine learning. Yet, the relationship with labor is distinctive for a company that was founded on the principle of using automation to reduce manual lab testing and using machine learning to minimize reliance on human "intuition" within the hypothesis-driven scientific method. The heavy use of automation and machine learning within the company has given its scientists the time to do "higher-value" or more creative work. Yet it is also apparent that Zymergen is able to conduct more experiments with fewer resources fully dedicated to a project, demonstrating labor efficiency compared to conventional R&D labs.

The use of automation and machine learning has not only shifted the company's approach to science and experimentation, it also has had an impact on the nature of work, the resources required, and where those resources are allocated within the organization.

4.1.2. Internal impact: increased labor productivity

Wet lab automation and machine learning assisted experiment design have allowed Zymergen to do more R&D with fewer labor resources — an effect that is increasing in magnitude.

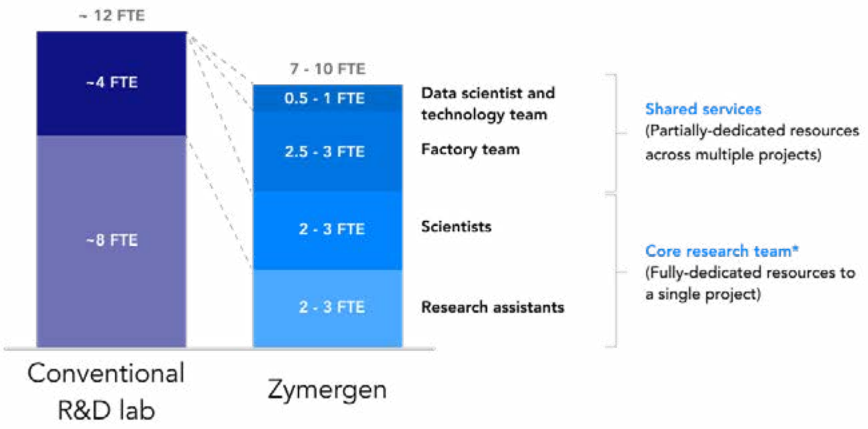
> *"In our six-year history, the role of automation and machine learning has grown, and the role of labor has grown leaner," a co-founder said. "Zymergen has been reducing the staff devoted to each project but rapidly increasing the number of projects we take on."*

Comparing Zymergen to another biotech company for new product development, Zymergen's core research teams [23] appear to have about 50 percent fewer researchers (scientists and research associates)[24] with significantly higher output (a tenfold increase in productivity compared with conventional labs). This is largely driven by productivity gains from automation in the wet lab. (See Exhibit 5: Comparison of staffing models.) Research associates at Zymergen focus on preparing experimental specifications, quality control, and analysis of the experimentation results, compared to manually conducting experiments as they would in more conventional R&D labs.

Unlike a conventional R&D lab, which may lack high-throughput screening, a Zymergen team is supported by centralized resources of data scientists, automation engineers, and "factory" teams (high-throughput screening lab technicians). Accounting for these shared services, who are partially-dedicated to research teams, a Zymergen team may have closer to a total of seven to ten full-time equivalents directly or indirectly contributing to a project. When accounting for these support resources, Zymergen may have closer to 25 percent fewer workers per program, rather than a 50 percent figure suggested above; however, these support staff may scale more efficiently as Zymergen's operations grow in experience and accumulated learnings.

Exhibit 4: Zymergen has a radically different staffing model vs. conventional R&D labs

Staffing level of a typical strain optimization project, full-time equivalents (FTE)



* Typically, Scientists are PhD-equivalents. Research Associates typically have a Master's or Bachelor's degree or equivalents.

Source: Zymergen employee interviews, team analysis.

[23] Core research teams only include scientists and research associates fully-dedicated to a single project. However, core research teams are supported by 'shared services' teams at Zymergen, composed of data scientists, automation engineers, and "factory" teams (lab technicians). This tally does not account for corporate functions or business development and sales teams.

[24] Compares only research teams (scientists and RA): ~12 FTE (conventional R&D lab) to ~6 FTE (Zymergen), or roughly 50 percent..

4.1.3. Internal impact: shift in project team composition

The composition of teams is different at Zymergen than those at biotech peers and other R&D labs. Research associates, who often have a bachelor's or master's degree, are typically responsible for supporting the scientists through analysis, mini-prep of experiments, and data cleansing. Scientists, who typically have a Ph.D. in a field such as biology, are responsible for designing experiments and interpreting the results.

Core research teams at Zymergen have a 50/50 split of scientists to research associates, compared with a 30/70 split across peer R&D teams. As stated above, the core research teams at Zymergen, e.g., scientists and research associates, are about half the size of other R&D teams, as well (See Exhibit 5: Comparison of staffing models.)

The lower ratio of research associates on a core team is due to the role of automation in the wet lab and machine learning at Zymergen, which has replaced many of the tasks that an RA could be responsible for in a typical R&D lab. At Zymergen, an RA would never manually pipet in the wet lab, but instead will "ship" the experiment to Zymergen's automated wet lab, or "factory," for execution. At Zymergen, machine learning tools may also be used for data cleansing and normalization, which have historically been typical tasks for an RA.

However, Zymergen's approach can also create additional work. As Zymergen can produce 10 times more output (in the form of unique strains) than a conventional R&D team may be able to, it has more data to interpret and experiments to analyze.

4.1.4. Internal impact: centralization of resources and shared services

While core strain improvement program teams at Zymergen may be smaller by about 50 percent, the core research teams are enabled by an increased number of shared services throughout the organization.[25] Compared to peers in biotech and R&D teams at clients, Zymergen likely has a higher ratio of labor dedicated to enabling functions or shared services. Considering partial resources dedicated to a project, Zymergen may have a larger team size, including shared services, for a typical project than a conventional R&D team. The full-time equivalent (FTE) team size, however, is lower compared to conventional R&D teams.

The most prominent examples of these shared services are data science teams that develop machine learning tools; data and software engineering teams that develop systems and software for information management and automatic data capture; and "factory" teams, which are responsible for operating and improving the automated wet lab and include automation engineers, process engineers, and technicians. Core program teams are internal customers of various services provided by the automated wet lab teams and the data science teams. These shared services are critical drivers of the increased productivity for client programs, such as the tenfold higher output of unique strains.

From a labor perspective, Zymergen's organization appears to have a 'top-heavy' research-intensive team of scientists and research associates with more diverse, specialized, and distributed teams of scientists, engineers, roboticists, and data scientists as support and shared services. For example, data engineering teams are critical to ensuring that data can be properly leveraged by machine learning models.

[25] As stated above, the 50 percent decrease accounts only for core research teams (scientists, research associates). Accounting for these shared service resources, Zymergen may have closer to 25 percent fewer workers per team, compared to a conventional lab of only research scientists.

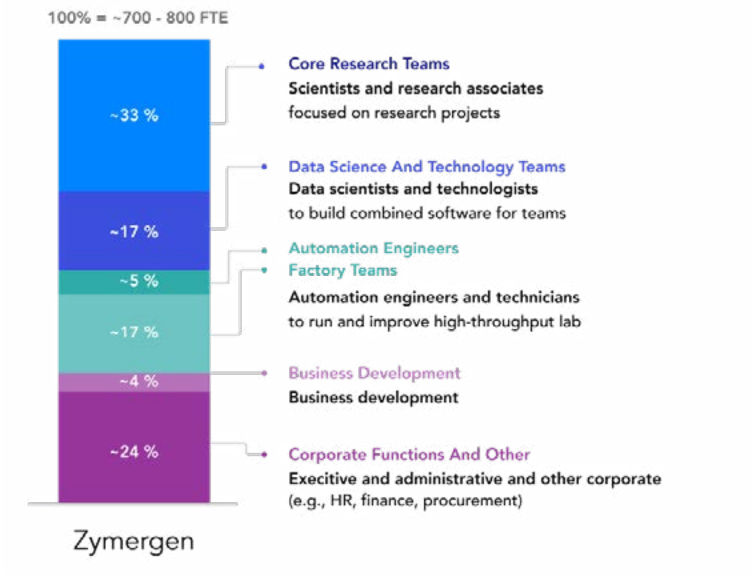
This difference in functional labor breakdown between Zymergen and its peer organizations is explained by two main factors: First, automation and machine learning tools require significant human resources to develop, support, and improve continuously. Second, Zymergen can achieve economies of scale with centralized automation and AI resources, involving both technical talent as well as the capital-intensive infrastructure investments (e.g., automated wet lab, cloud storage). Knowledge sharing across Zymergen programs is also a benefit of a shared services model.

Zymergen's organizational structure reflects these differences. In contrast to a traditional R&D lab which has a much larger core research team, out of about 700 to 800 total FTEs at Zymergen, only 33 percent are fully dedicated core research teams involved in client programs and internal product development. Another 17 percent are involved in shared services, partially dedicated to projects. Some 22 percent of FTEs make up factory teams (shared services, partially dedicated to project teams) that build and test strains in the lab. Corporate functions and business development account for approximately 28 percent of FTE (not considered shared services for this analysis). Business development is responsible for identifying potential clients, developing relationships, and interfacing closely with clients and Zymergen core teams throughout a program. Corporate functions include executive leadership, finance, HR, and other administrative roles (See Exhibit 6: Zymergen's organizational distribution).

The centralized resource model may not be the most appropriate model for all companies, such as Zymergen's customers. The upfront costs associated with automation in a high-throughput screening lab or in machine learning models and infrastructure might not make economic sense due to: (1) fewer programs to realize marginal benefits, (2) capital intensity and limited funding, and (3) lack of available technical talent and organizational hurdles.

Exhibit 5: Zymergen's organizational distribution

Approximate allocation of employees (FTE) across function as % of total FTE



Approximate FTE as of time of case study research in fall 2018

Source: Zymergen company data, Linkedin data, analysis.

4.1.5. Internal impact: allocation of labor across a project life cycle

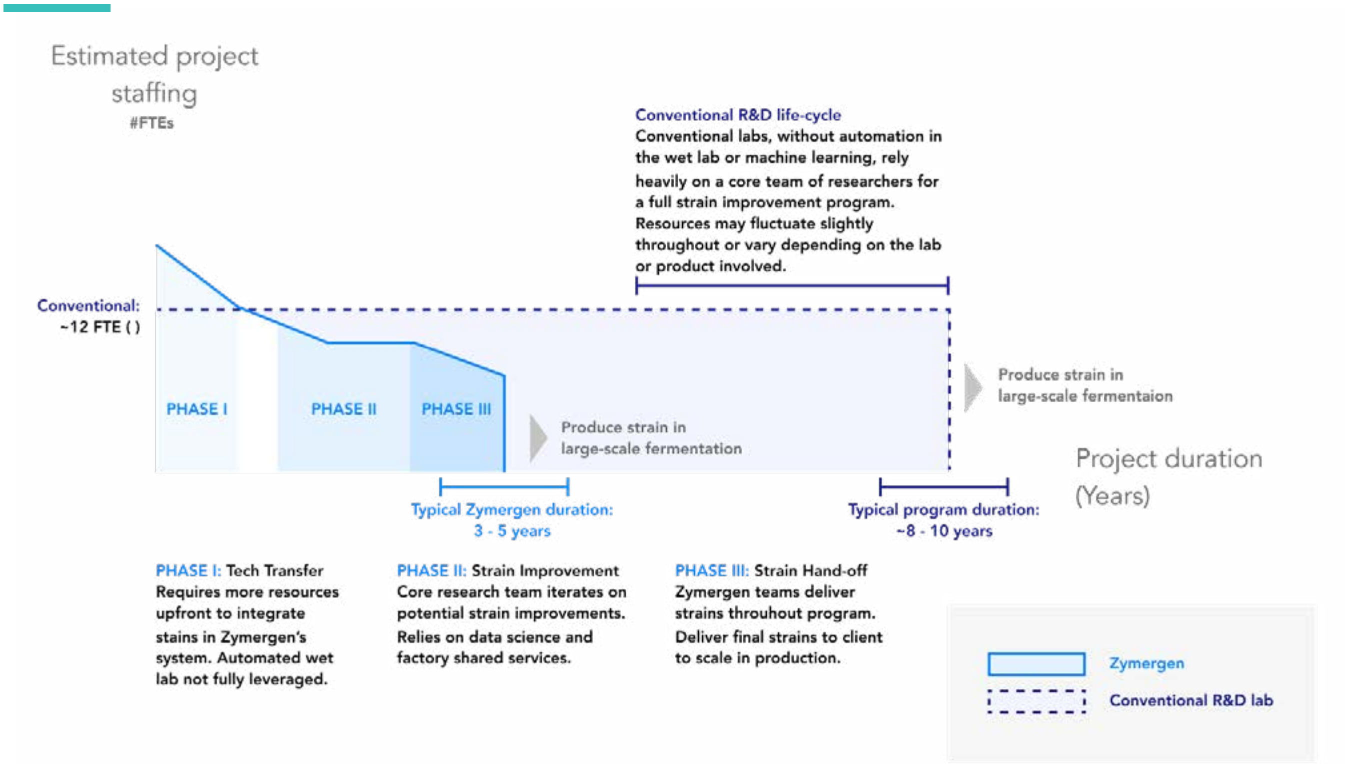
In a conventional strain improvement program, an R&D team of about 12 FTE would likely have been fully dedicated to the program for its full eight to ten years. Within Zymergen, program staffing is more dynamic, with a larger input of resources (some fully-dedicated, some partially-dedicated) required upfront staffing in the tech transfer phase (defined below), followed by fewer fully-dedicated resources to sustain the program during the strain improvement phase. The programs at Zymergen offer additional productivity gains through shorter projects.

During the first phase of a project (the tech transfer phase), Zymergen integrates the client's specifications into the Zymergen platform and ensures that the client's strain can be integrated into Zymergen's "factory" workflows (systems in the high-throughput screening/automated wet lab). During this phase, Zymergen will typically require more worker resources, such as specialists in Zymergen's workflows, testing teams, and the core research team of scientists.

During the strain improvement phase, the program is in steady-state and is iterating through highly automated experimentation cycles to identify optimizations to the strain for the client's desired product and program objectives (e.g., improve economics, accelerate commercialization).

Exhibit 6: Zymergen has significantly shorter project durations

Staffing level across a typical strain optimization project timeline



Conventional R&D life-cycle
Conventional labs, without automation in the wet lab or machine learning, rely heavily on a core team of researchers for a full strain improvement program. Resources may fluctuate slightly throughout or vary depending on the lab or product involved.

PHASE I: Tech Transfer
Requires more resources upfront to integrate stains in Zymergen's system. Automated wet lab not fully leveraged.

PHASE II: Strain Improvement
Core research team iterates on potential strain improvements. Relies on data science and factory shared services.

PHASE III: Strain Hand-off
Zymergen teams deliver strains throuhout program. Deliver final strains to client to scale in production.

Source: employee interviews, expert interviews, team analysis.

4.1.6. Internal impact: hiring for hybrid profiles

Zymergen is a hybrid company that takes an engineering approach to biology. When hiring, it looks for hybrid backgrounds: cross-functional candidates with backgrounds in both engineering and biology. “Most biologists are trained to generate designs [for strain experiments]. To thrive at Zymergen, you need to put that mindset in the back seat,” a co-founder said. “We look for people with backgrounds in both biology and engineering. We need both the math and science background.”

Finding employees with this profile has not always been easy. Because of its unique hiring needs, Zymergen requires an ecosystem of specialization within and beyond its organization — everything from automation engineers to biomaterial patent lawyers. Given the specificity of certain profiles sought, Zymergen can encounter geographic challenges in sourcing its labor, as well as limitations from broader academic or training institutions depending on whether and how well they encourage such cross-functional profiles.

4.1.7. External impact: labor impact on customers’ R&D teams

It is too early to tell how Zymergen’s strain improvement programs will affect the R&D teams at other companies, such as its customers. So far, it appears that there has been minimal impact. In the future however, there could be potential for some labor displacement of customer R&D teams or lower hiring rates of in-house R&D teams.

A key question: Is Zymergen’s services-based business model — and, by extension, AI and machine learning — competing with clients’ internal R&D teams, or complementing them? In some cases, clients have too many projects for the R&D staff and resources they have. Once Zymergen takes some of the work, the clients’ staff may get redeployed to other projects.

However, there can sometimes be tension between the R&D team of a client and Zymergen’s team in working on a finite-seeming set of projects.

“The answer is nuanced here,” a co-founder said. “There is a strong argument to make that our work is synergistic or complementary to existing R&D teams at our clients.” Typically, the clients have already optimized a strain to 80 to 90 percent of its potential yield. The final portion of potential optimization requires Zymergen’s systematic approach rather than a conventional hypothesis-driven approach.

4.1.8. External impact: potential impact on manufacturing labor force of customers

It is possible that the savings resulting from a Zymergen program could increase demand for fermentation products or biomaterials, resulting in increased demand for manufacturing labor by customers, or even new downstream job opportunities associated with sale of newly developed products.[26]

However, Zymergen’s programs could also potentially decrease labor need associated with manufacturing:

Labor at fermentation plants: If production volume can be increased due to a strain improvement program, the labor productivity of manufacturing operations will increase, i.e., in theory, less labor will be needed to manufacture the same volume of product. In practice, some manufacturers increase the volume of their production while keeping the labor force constant, rather than decrease their labor force.

[26] Savings achieved through use of Zymergen’s AI methods may result in partially-lower product costs, and thus greater demand for products than would otherwise be experienced.

Forgone construction or associated supply chain labor: Zymergen customers may also look for productivity gains as an alternative to making a capital-intensive investment in new facilities for additional capacity. This could have downstream impacts on labor in the form of forgone construction labor for a new plant or forgone labor associated with the equipment and construction supply chain, as well as the labor to staff such a plant. However, to date, no direct workforce reduction in the manufacturing labor of Zymergen's customers has been reported.

4.1.9. Potential long-term workforce impacts

If Zymergen achieves its founders' goals, the impact on highly skilled workers will be profound: These workers could face the same drop in available jobs that others have identified in traditional manufacturing jobs, borne from factory automation." Our vision is to have all of the manual parts of the work done by robotics and to have all the intellectual efforts, such as design and interpretation, to be done with machine learning and big data," a co-founder said. "There is a lot of anxiety among members of the staff that their job could be at risk in the long term."

The current picture, however, suggests more of an evolution than a dramatic change. Some trends observed at Zymergen and its clients could signal broader workforce changes in the long term, especially for biology R&D teams, biotech companies, and their customers, but with less profound impacts observed to-date.

A number of the micro labor trends observed at Zymergen could offer clues to the future of the broader workforce:[27]

1. Relative decline in the number of highly-educated scientists doing manual lab work.
2. Relative increase in the number of shared services for machine learning and automation; job gains for specialists required to enable automation and machine learning tools, especially data engineers, data scientists, and automation engineers.
3. Increased demand for specialization in multiple domains (e.g., biology, engineering, data science) or for hybrid profiles (e.g., automation engineers with biology background).
4. Faster project staffing cycles due to shorter total project durations.
5. Slightly fewer aggregate workers per project relative to its R&D clients.

[27] Felten, Edward W.; Raj, Manav; Seamans, Robert (2018). "A Method to Link Advances in Artificial Intelligence to Occupational Abilities." *AEA Papers and Proceedings 2018.*

4.2. Productivity and business results

Zymergen's techniques led to reported cost savings for its customers on both R&D and manufacturing, as well as higher business productivity internally. Zymergen's customers operate in a set of industries that are capital-intensive, highly competitive, and often with thin margins. As one Zymergen representative noted of a Fortune 100 client in a commodity sector, "there are not many success stories in this sector." Yet, the productivity gains from Zymergen have not come without cost; indeed, Zymergen has invested heavily to facilitate these benefits.

4.2.1. R&D productivity

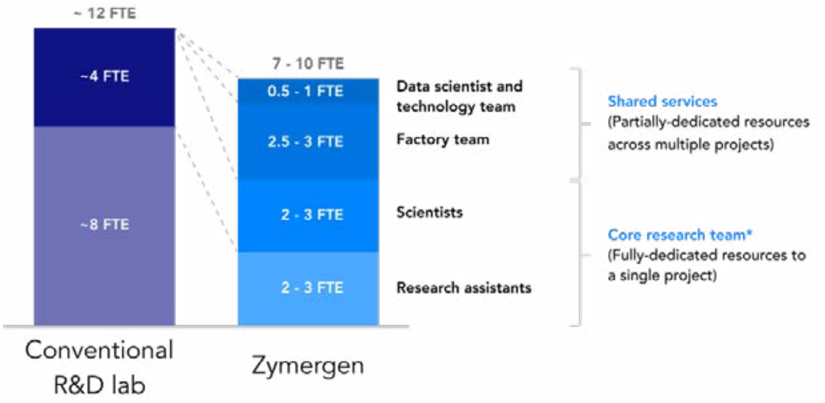
Zymergen reports that its most significant productivity gains, compared to conventional R&D labs, are realized through high-throughput screening in the automated wet lab. Zymergen aims to improve the economics of a customer's strains and deliver the improved strains in much shorter time frames than traditional R&D labs. These internal gains translate to benefits for Zymergen's customers as well, typically through improved economics in downstream manufacturing of a product (See section 4.1.2).

Zymergen's strain improvement programs typically require just three to five years to reach the same levels of improvement that would take eight to ten years at their clients' organizations. While the project duration depends on the type of strain involved, as well as the data availability from past experiments, the time savings can be significant. For one Fortune 100 client, Zymergen was able to "deliver the same level of yield improvements in about two years that would have typically taken the client six to seven years without Zymergen," said a business development representative.

Zymergen reports that its fast cycle times allow research teams to meet ambitious timelines due to automation in the lab and confidence in the data, leading to shorter project durations overall compared to conventional R&D labs. Experimentation throughput at Zymergen is significantly higher than at conventional R&D labs: from tens or hundreds of experiments and iterations to thousands of experiments per week in a single project, a tenfold increase in output. "Zymergen has focused on testing as many strains as rapidly as possible. We are constantly working with our data science team to think of ways to improve this and automate workflows," said a business development representative.

Exhibit 7: Zymergen has clear throughput advantage over conventional R&D labs

Unique genotypes produced per week (single project)



Source: Employee interviews,expert interviews, team analysis.

4.2.2. Manufacturing savings for Zymergen’s customers

Zymergen’s strain improvement programs typically target products with large-scale production, where a small improvement in yield can lead to significant economic savings. The focus is typically either on improving yield, thereby saving money on raw materials (e.g., reduced sugar consumption per unit output of product), or improving the speed of production (e.g., faster fermentation with the same level of inputs). Both of these illustrate productivity increases for Zymergen’s customers. From a limited sample set of Zymergen programs, Zymergen finds that it has been able to deliver a two- to- three time yield improvements compared to in-house R&D teams, which can translate to significant savings in the manufacturing costs.

For example, a Zymergen customer may produce citric acid, an organic acid produced in significant volumes and used as a natural preservative in food and beverages or to add an acidic taste. Production is consolidated among a few global producers, mostly based in China, and overcapacity in the industry has created price competition and pressured margins. If Zymergen can double or triple yields, this can mean significant benefits for the producer. From Zymergen’s perspective, though its services likely cost more than internal strain improvement programs, the return on investment appears to justify this cost.
Zymergen’s use of AI helps its clients save resources in multiple ways:

- **Reduced energy costs:** Many fermentation reactions require cooling. A beneficial attribute for a phenotype is ‘thermotolerance’: maintaining the same productivity at a relatively higher temperature. Improving thermotolerance would reduce the energy costs required to cool a fermentation tank.
- **Use of biomaterials:** If Zymergen delivers an improved strain that uses biomaterials instead of petrochemicals for the product’s inputs, it could eliminate contaminants that might have to be removed in downstream processing, which could reduce overall production costs.
- **Reduced manufacturing labor overhead:** If a company has a fixed amount of labor and increases production volume, the per-unit labor costs of manufacturing decrease.
- **Forgone capital expenditures:** Improving the economics of an existing strain can be an alternative to large capital expenditures for installing additional manufacturing capacity. For instance, if Zymergen is able to increase the production rate by two to three times for a high-volume product, the client could increase the volume of the product with the same level of inputs (e.g., raw materials, energy, labor) and with current plant capacity. Improving the productivity of the fermentation process might be an alternative to building an additional industrial fermentation plant, which could cost anywhere from $100 million to $200 million for a 1 million-liter capacity plant, resulting in CapEx savings.[28]

[28] Industry expert estimate (n=2)..

4.2.3. Investments required for AI and automation systems

Although Zymergen offers significant benefits to customers, AI and automation can be expensive. Significant investments have included:

- **Infrastructure investments:** Zymergen made significant investment in infrastructure to enable its business, including robotics and machinery in the wet lab, sensors for automatic data capture, and IT/data infrastructure to centralize and store data. “Automation is expensive. You can spend a few million dollars very, very quickly,” said an automation engineering executive at Zymergen. “The barrier to entry for automation is high.”

- **Investment in staff:** The automation team at Zymergen is composed of roughly 40 FTEs,[29] or close to 5 - 6 percent of the workforce. Additionally, the data science and software engineering team accounts for roughly 30 percent of the workforce. There will likely be increased economies of scale on labor in the future, as machine learning and automation improve. Theoretically, as Zymergen scales the number of programs, data science and automation engineers could scale at a slower rate due to reusability of machine learning models and wet lab infrastructure across multiple programs. This shared services model could thus allow for future labor efficiencies internally at Zymergen.

- **Unconventional cost structure:** Zymergen has a different cost base than conventional R&D teams or other biotech firms: lower estimated labor costs as a percent of total costs, with significantly higher consumables costs (reagents, DNA, raw materials for experiments) because of the increased volume of experiments performed (see next section); slightly higher IT infrastructure costs (cloud and storage); and high upfront investments in machinery and robotics.

- **High cost of data acquisition:** The automated wet lab makes it easy to “test everything,” and Zymergen uses that advantage. The ‘more shots on goal’ strategy, in theory, will lead to higher quality data that can be leveraged across programs and will improve the machine learning recommendation engines over time. Yet the company has major consumable costs associated with its data acquisition strategy: To run each experiment, the company needs to purchase an inventory of supplies, from pipettors to the microbial cultures, reagents, and DNA used in experiments, which can add up quickly with each marginal experiment. The company has a lower unit cost per build than most conventional R&D labs due to better negotiated pricing on consumables. In absolute figures, however, the consumable costs are much higher than those of most labs. In some ways, Zymergen is burdened by its chief advantage: The company is enabled by machine learning and automation to experiment on vastly more strains, yet these experiments also carry significant costs.

[29] FTE refers to full-time equivalent employees at the company. The term is a business acronym and is a conventional unit of measure to compare workloads across different business contexts. Because labor may be undertaken by part-time employees, it is useful to standardize work amounts across full-time equivalents, rather than total worker counts.

05 Reflections and implications

Zymergen's experience holds potential lessons for those interested in how AI and machine learning could change the workplace and the workforce. As an AI-native company, Zymergen is faced with a unique set of challenges and opportunities. As the company matures, the implications on the business, productivity, and labor become more salient, as well as the role that AI and ML have in influencing these changes. Although the ultimate effects on the broader workforce and labor market remain to be seen, reflecting on Zymergen's experience holds some indications for the future.

5.1. Reflections on change management for an AI-native company

An 'AI-native' company is faced with a different set of opportunities and challenges than an incumbent company seeking to deploy AI. Zymergen had no legacy systems or processes to contend with. Rather, it had a clean slate on which it could shape an IT infrastructure and web lab automation designed for its needs. Its early investment in building strong data assets could provide a competitive advantage in the long run. Zymergen also faced some challenges: It required more upfront investment in data and automation infrastructure, had to work harder to find employees with the hybrid profiles it needed, and needed to train employees. But the people it hired had already bought into the AI-first mission, so change management was less of an issue.

Still, AI presents technical and cultural hurdles for both AI-native companies and incumbents alike. Zymergen's management team noted several ways the move to AI affected how people approached their work:

- **Unexplainable models:** It was much easier to get buy-in from the client's R&D teams and Zymergen's in-house scientists if a model was easily understood. "When an AI algorithm proposes — using an unexplainable model like deep learning — that you target a specific market, or set prices a particular way, and this proposal differs from what human judgment would have led to, will your people believe it and act upon it?" said the CTO. "Even at Zymergen, scientists struggle to accept the counterintuitive proposals AI emits, even when on average, following the AI models has a statistically measurable benefits. What chances do companies with a decades-long process have in making this shift?" Nonetheless, for many decision-makers, such as P&L owners at Zymergen's clients, an impact on profitability (as a result of the model's recommendations) is more important than being able to fully understand the rationale behind the model itself. Yet even within this approach, understanding a model's processes can be important for debugging the model's performance and for avoiding possible complications and risks, particularly in high-stakes domains.
- **Incremental approach to trusting an AI system:** Starting with simpler models earlier in the company's lifecycle helped gain trust of the end-users — Zymergen's scientists. In the long-term, this approach can help employees learn to trust the models, which ultimately will be more important than having the most powerful models immediately. As the CTO noted, "don't expect miracles overnight," suggesting that the performance gains from integrating machine learning into the scientific process takes time and requires incremental trust-building with many stakeholders. Zymergen's team reported that the approach and AI-related projects need committed sponsorship and leadership for a long period of time, often two to four years, to see a project through completion. Interestingly, change management around the adoption of machine learning for scientists proved to be more challenging than for lab technicians around the adoption of robotic automation. Whereas lab technicians welcomed automation in the wet lab because the volume of work was rapidly increasing, scientists felt some friction in handing off work to an AI system, driven perhaps in part by pride. "This is something that our scientists have studied for years to get a Ph.D.," the CTO said. "To be told it can be reduced to a statistical analysis is challenging. We try to shift the conversation to new challenges where our scientists can better focus their attention to. The reality is that we have an endless amount of knowledge work to be done, and we are only handing off a fraction with an AI system."

5.2. Technical reflections on the integration of ML and automation

Working with AI systems has provoked a number of reflections within Zymergen about both the promise and the limitations of AI, machine learning, and automation:

- **Risks of sophistication:** More sophisticated techniques or AI methodologies often do not always produce better results. They are more difficult to explain and require more data. Their implementation may also be more challenging. Thus, simpler techniques, e.g., linear regressions, can be effective tools.
- **Increasing comfort:** To Zymergen's customers and many of Zymergen's scientists, results from strain improvement and the consequent impact to the bottom line (e.g., savings on raw materials) are more important than being able to explain why a strain was recommended by the machine learning algorithms. Explainable models in the context of material science R&D research may become less important as results are demonstrated and stakeholders become more comfortable with unexplainable recommendations.
- **Limits of machine learning:** Machine learning approaches to experiment design are both an alternative and are complementary to a hypothesis-driven approach. In conventional experiment design, domain experts such as biologists design experiments based on years of experience. The machine learning approach is more systematic, yet it may be susceptible to its own form of biases and trust issues. The machine learning approach also requires vast quantities of data to be effective.
- **Advantages of computational scale and data "memory":** For the machine learning models used by Zymergen, the value is located in the immense scale and timeframes of relevant data — the data "memory" — rather than an advanced methodology. The bias exhibited by humans who tend to focus on the most recent data can negatively impact the design of experiments.
- **Costs and benefits of high-throughput screening:** High-throughput screening is a key productivity advantage for Zymergen, yet it comes with significant data acquisition costs, in the form of costs to run an experiment. Zymergen's vastly higher throughput — 10 times higher than a typical lab — is a key advantage because it provides more data for machine learning models. Yet each experiment comes with costs, including consumables, energy, and increased CapEx for machinery as programs scale. The consumable costs for Zymergen are significantly higher as a portion of costs because of this approach. It is not clear that Zymergen's "experiment everything" mindset will continue to prove economically viable.
- **Adaptability**: Automation and machine learning can be less adaptable than humans, which can make continued changes to models, processes, or robotics equipment costly and complex. As the CTO notes, "Automation is a double-edged sword. What was automated, now runs in high-throughput. When you want to change it, this is now an engineering problem, not a re-training [of humans] problem." If a process needs to be redesigned, this can have real cost implications. Humans, on the other hand, tend to adapt to changes more readily. While they may not be as efficient in the long term, they can be more adaptable when a process is not fully defined or codified.

5.3. Implications of business model on labor and the workforce

Zymergen takes a fundamentally different approach from that of traditional peer companies both to its scientific projects and to its staffing. Using automation and machine learning propels Zymergen beyond what is possible with human-driven science. Zymergen's employees are more diverse and specialized than those at traditional R&D labs, and they are typically distributed across multiple projects. Zymergen does not claim that its employees are smarter than its clients' R&D teams — but that they are equipped with better tools (e.g., a recommendation engine and a high-throughput lab).

Zymergen's workforce model may give a glimpse of how AI could affect the nature of work, especially in R&D-intensive industries:

- **Services-based business model:** A services-based business model can provide economies of scale for automation and machine learning investments, as well as data acquisition costs. Machine learning models can 'learn' across programs, potentially lowering the marginal development and training costs as deployments scale. Zymergen can justify its investment in automation in the wet lab and development of machine learning models because these assets and human resources are shared across multiple programs. A client may not be able to justify the upfront capital investment for a limited number of strain improvement programs. As Zymergen increases its number of projects, it will likely achieve better economies of scale. Services-based models could also impact Zymergen's customers' workforces. If the outsourced, services-based model – especially for AI-related applications – proves to be more effective and efficient, it could displace employees in the customer organizations in the long-term.
- **Data acquisition:** Zymergen effectively passes its data acquisition costs through to the customer by using its baked-in fees to cover the costs of such experiments. In exchange for improved economics for its customers, Zymergen keeps the data generated through experiments and can use it to build Zymergen's broader genomic dataset and help train the machine learning recommendation engine for all programs. However, given the relatively low data coverage of the microbial genome overall, the data is not always valuable across various programs.
- **Shared services**: Zymergen uses smaller core research teams but more shared services to enable core research teams to be able to experiment and interpret at higher throughput. This larger support staff enables its fully-dedicated research teams: Zymergen has more data scientists and automation engineers (as shared services) but fewer scientists and research associates (fully-dedicated) on the core research team.
- **Barbell effect on Zymergen's workforce:** Changes in demand for different types of workforce at Zymergen – in part due to offloading certain tasks to robotics and machine learning systems – could shed light on potential future changes in the workforce. A co-founder speculated that, based on what he observes in Zymergen's workforce today, the so-called "barbell effect" could increase in the future. The barbell effect describes higher demand for Ph.D. scientists and lower-paid support staff, but lower demand for the research associates in the middle. This could mean higher job availability for those at the top and bottom ends of the scale, though with others perhaps displaced. However, it is unclear whether this dynamic will fully play out. For instance, demand for labor to run operations in the wet lab, such as lab technicians, could decrease as the automation becomes more advanced. Conversely, middle-skilled RA demand could increase if demand for Zymergen's services increases, leading to increased capacity and operating labor required for the wet lab.
- **Changing allocation of labor**: The introduction of dynamic AI processes may affect the steadiness at which labor and capital are deployed across project lifecycles. For instance, within Zymergen, projects experience higher upfront investment for implementation and then less labor to maintain the program thereafter. In contrast, a conventional project would have a roughly constant number of FTEs over the project's life-cycle.

5.4. Conclusion

At its six-year mark, Zymergen is still early in its life cycle, and the full effects of machine learning and automation on its own business, its customers' businesses, and the relevant markets it operates in have not been fully demonstrated. Still, the case offers valuable insight into how AI might impact workplaces of the future, especially for new companies that are "AI-native" by default.

The case of Zymergen raises important questions about the future of work. The company was built on the principle that certain tasks traditionally held by scientists could be automated: designing scientific experiments and the manual labor work to conduct them. Yet today, scientists are still integral to the process at Zymergen. By pairing AI with human scientists, Zymergen discovered new value that enhanced the traditional R&D process. For instance, the combination of its AI experiment recommendation engine with its high-throughput automated wet lab have yielded valuable experiments that may not otherwise have been tested. Yet, of note, introducing the AI engine was not a panacea in and of itself; it operated within a more complex network of enabling systems and automated processes.

Zymergen's use of automation and machine learning also holds implications for the required skills of the future, especially for highly-educated workers (e.g., Ph.Ds in biology or research associates). While workforce reductions have not been reported, differences compared to conventional R&D labs are apparent, with demand for some specialized skill sets increasing and demand for others declining. Manual labor previously used in scientific experimentation has largely been eliminated at Zymergen, yet lab technicians are still required to keep its automated wet lab up and running. The influence of these technologies on labor will likely continue as AI-related technologies mature. Ongoing trends from Zymergen indicate that the highly-educated workforce will be involved in some stages of product development for the foreseeable future. The growing need for hybrid profiles (e.g., data science and biology) at Zymergen suggest a trend in increased specialization in multiple domains necessary to support AI-related efforts. Within Zymergen's areas of focus specifically, concentrating on tasks where humans have a comparative advantage could lead to more human involvement in early product development (in which AI is not yet fully spun up) and post-development review, testing, and analysis (in which the AI system's outputs need to be evaluated and iterated upon).

The nuance of Zymergen's current and potential impact on labor and the economy is a testament that AI and automation's effects are often multi-faceted. Zymergen's business model is designed around the use of AI, machine learning, and automation to improve its own labor productivity, and the operating expenditure and capital productivity of its clients. Its investments in automation, machine learning, and AI have led to increased experimentation speeds, reduced labor costs, and also reduced project times. However, the net effect on other measures of productivity is more difficult to observe. While Zymergen has lower labor costs, it also has large upfront capital investment costs along with major consumable costs. Likewise, the labor effects might not be observable within Zymergen, but instead cascade outside to other parts of the business ecosystem.

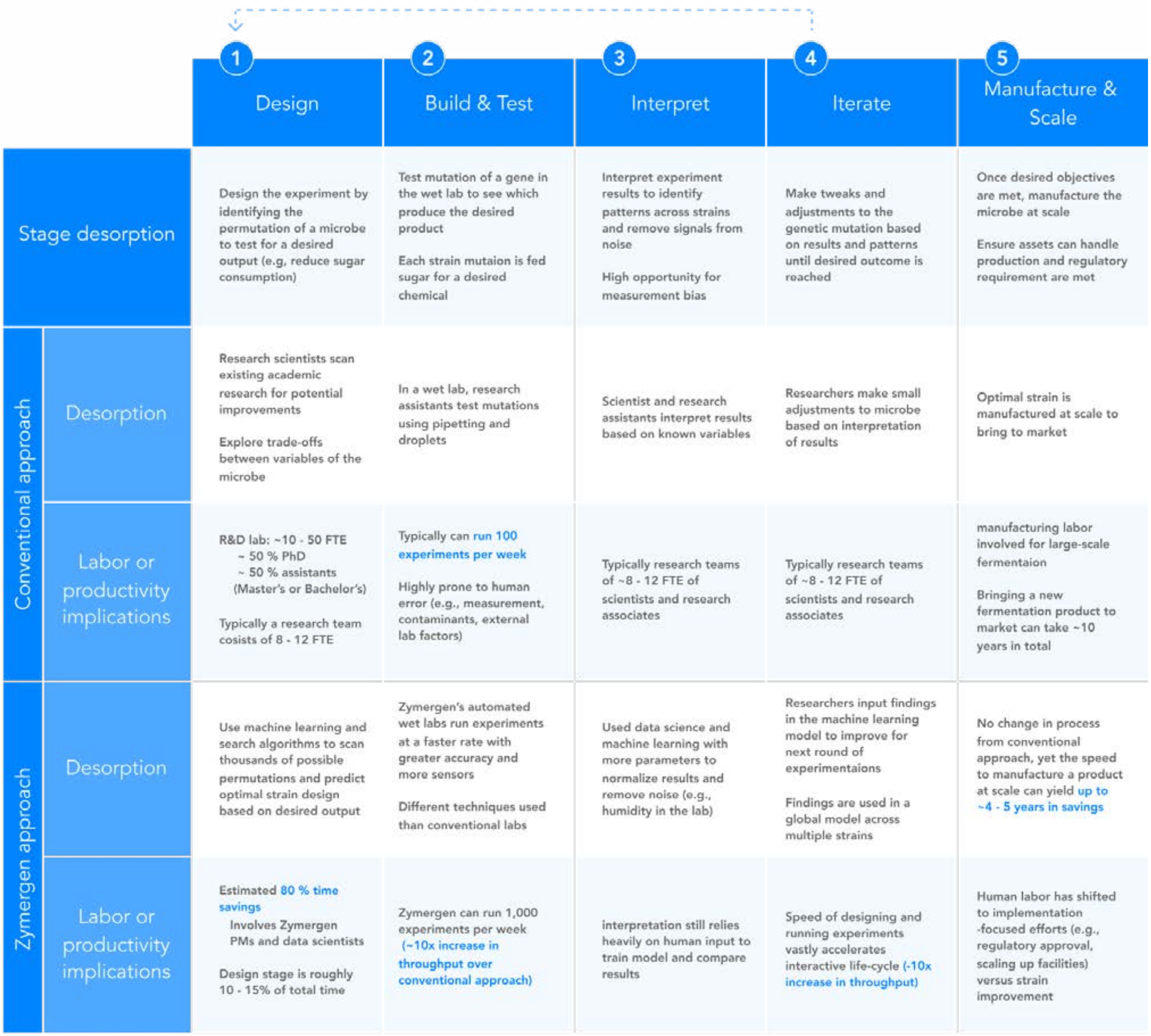
06 Appendix

6.1. Open questions for further research

The Zymergen case poses a number of questions: With a large portion of manual work conventionally done by research associates and scientists offloaded to automated systems, will the more cognitively-driven tasks also be replaced by AI systems or will the new technologies be complementary? Similarly, as automation enters into the most manual and repetitive human tasks (like pipetting), what will be the trend for less repetitive, more novel tasks? How will human agility for learning compare with the time needed to design and maintain automation systems for varied tasks? Finally, what will be the cascading impacts of AI be on external companies and industries? Given Zymergen's relatively small employee count, these may be the most profound implications of its reported productivity enhancements and merits further research. While many impacts and ramifications of AI-related technologies on labor and the economy remain unknown, the case of Zymergen calls attention to trends that suggest the need for further study.

Exhibit 8. Comparing Zymergen's process to the conventional strain development process

Exhibit 8: Zymergen has leveraged machine learning and automation to drastically accelerate the strain improvement process

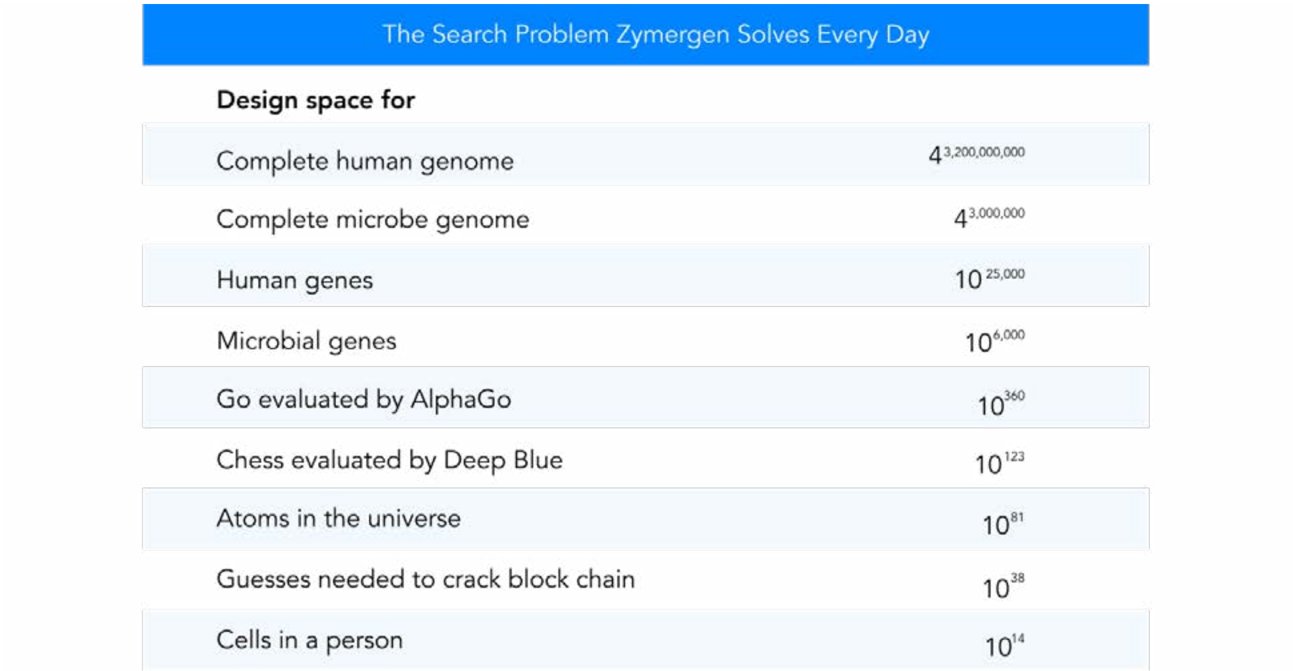


		1 Design	2 Build & Test	3 Interpret	4 Iterate	5 Manufacture & Scale
Stage desorption		Design the experiment by identifying the permutation of a microbe to test for a desired output (e.g, reduce sugar consumption)	Test mutation of a gene in the wet lab to see which produce the desired product Each strain mutaion is fed sugar for a desired chemical	Interpret experiment results to identify patterns across strains and remove signals from noise High opportunity for measurement bias	Make tweaks and adjustments to the genetic mutation based on results and patterns until desired outcome is reached	Once desired objectives are met, manufacture the microbe at scale Ensure assets can handle production and regulatory requirement are met
Conventional approach	Desorption	Research scientists scan existing academic research for potential improvements Explore trade-offs between variables of the microbe	In a wet lab, research assistants test mutations using pipetting and droplets	Scientist and research assistants interpret results based on known variables	Researchers make small adjustments to microbe based on interpretation of results	Optimal strain is manufactured at scale to bring to market
	Labor or productivity implications	R&D lab: ~10 - 50 FTE ~ 50 % PhD ~ 50 % assistants (Master's or Bachelor's) Typically a research team cosists of 8 - 12 FTE	Typically can run 100 experiments per week Highly prone to human error (e.g., measurement, contaminants, external lab factors)	Typically research teams of ~8 - 12 FTE of scientists and research associates	Typically research teams of ~8 - 12 FTE of scientists and research associates	manufacturing labor involved for large-scale fermentaion Bringing a new fermentation product to market can take ~10 years in total
Zymergen approach	Desorption	Use machine learning and search algorithms to scan thousands of possible permutations and predict optimal strain design based on desired output	Zymergen's automated wet labs run experiments at a faster rate with greater accuracy and more sensors Different techniques used than conventional labs	Used data science and machine learning with more parameters to normalize results and remove noise (e.g., humidity in the lab)	Researchers input findings in the machine learning model to improve for next round of experimentaions Findings are used in a global model across multiple strains	No change in process from conventional approach, yet the speed to manufacture a product at scale can yield up to ~4 - 5 years in savings
	Labor or productivity implications	Estimated 80 % time savings Involves Zymergen PMs and data scientists Design stage is roughly 10 - 15% of total time	Zymergen can run 1,000 experiments per week (~10x increase in throughput over conventional approach)	interpretation still relies heavily on human input to train model and compare results	Speed of designing and running experiments vastly accelerates interactive life-cycle (-10x increase in throughput)	Human labor has shifted to implementation -focused efforts (e.g., regulatory approval, scaling up facilities) versus strain improvement

Source: employee interviews, expert interviews, team analysis.

Exhibit 9. Size of the genomic ‘search space’

Exhibit 9: Size of the genomic ‘search space’



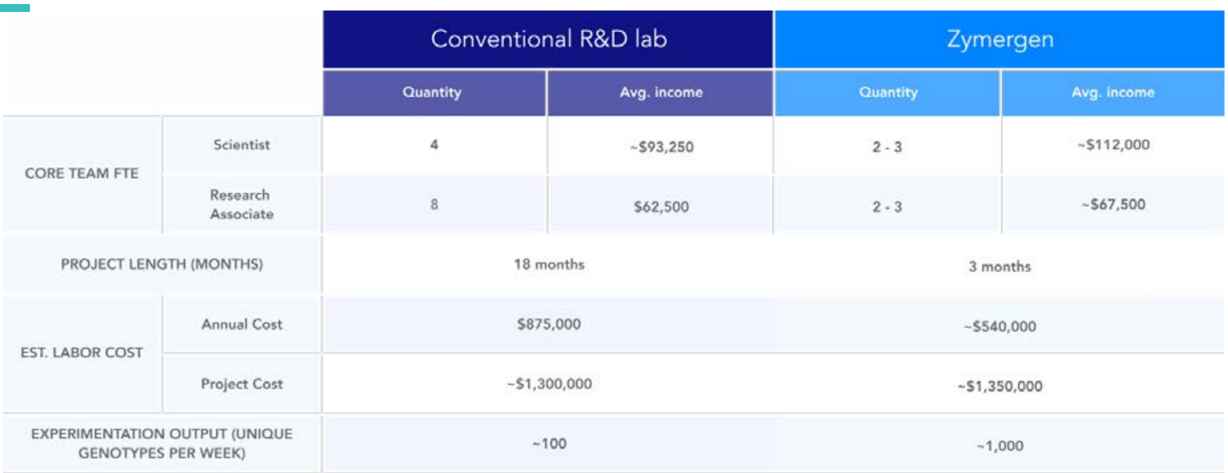
The Search Problem Zymergen Solves Every Day

Design space for	
Complete human genome	$4^{3,200,000,000}$
Complete microbe genome	$4^{3,000,000}$
Human genes	$10^{25,000}$
Microbial genes	$10^{6,000}$
Go evaluated by AlphaGo	10^{360}
Chess evaluated by Deep Blue	10^{123}
Atoms in the universe	10^{81}
Guesses needed to crack block chain	10^{38}
Cells in a person	10^{14}

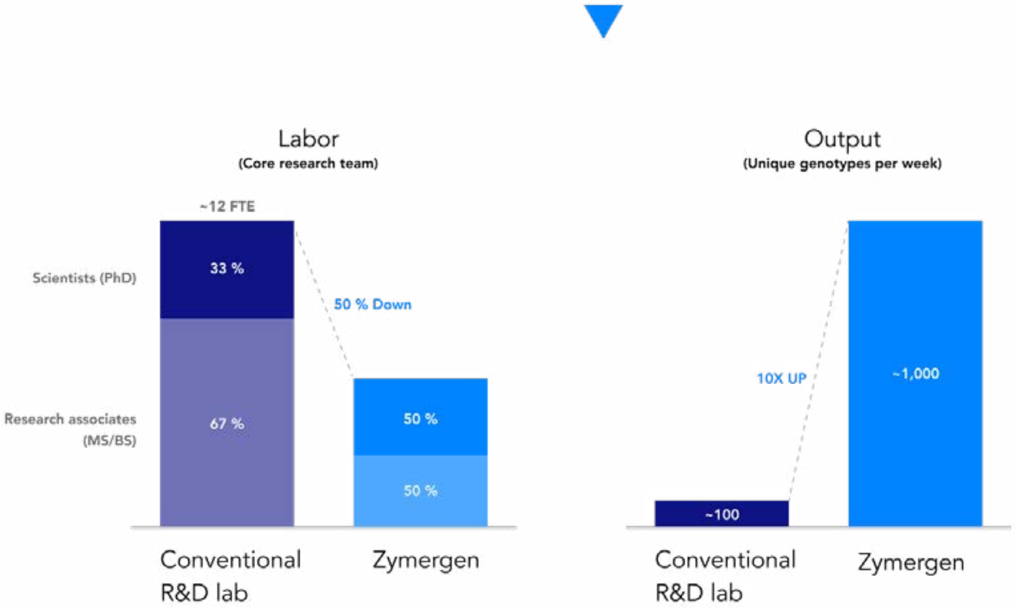
Source: Zymergen company documents, Medium blog post from Data Collective (Zymergen Investor).

Exhibit 10. Comparison of a team size and labor cost for a new product development project

Exhibit 10: Comparison of team size and labor costs for a new product development project



		Conventional R&D lab		Zymergen	
		Quantity	Avg. income	Quantity	Avg. income
CORE TEAM FTE	Scientist	4	~$93,250	2 - 3	~$112,000
	Research Associate	8	$62,500	2 - 3	~$67,500
PROJECT LENGTH (MONTHS)		18 months		3 months	
EST. LABOR COST	Annual Cost	$875,000		~$540,000	
	Project Cost	~$1,300,000		~$1,350,000	
EXPERIMENTATION OUTPUT (UNIQUE GENOTYPES PER WEEK)		~100		~1,000	



1 Does not leverage machine learning to the extent of other projects within Zymergen, but heavily leverages automated wet lab. The comparison is for a particular product and is not 1:1; the attributes of the type of product are comparable (e.g., similar complexity of organism)

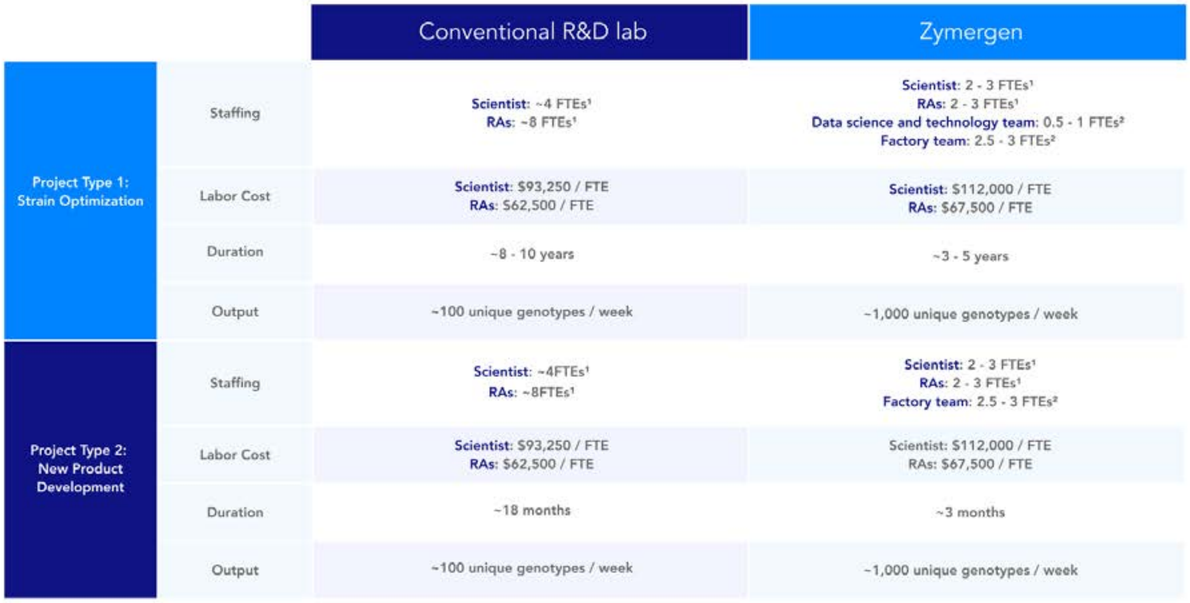
2 Includes core project team. Does not include support team or shared services (e.g., manufacturing, data science)

3 Typically, Scientists are PhD-equivalents. Research Associates typically have a Master's or Bachelor's degree or equivalent.

SOURCE: employee and expert interviews; incomes based off of Glassdoor crowd-sourced salaries in San Francisco, CA, team analysis.

Exhibit 11. Fact sheet - conventional R&D lab and Zymergen comparison

Exhibit 11: Fact sheet - conventional R&D lab and Zymergen comparison



		Conventional R&D lab	Zymergen
Project Type 1: Strain Optimization	Staffing	**Scientist**: ~4 FTEs[1] **RAs**: ~8 FTEs[1]	**Scientist**: 2 - 3 FTEs[1] **RAs**: 2 - 3 FTEs[1] **Data science and technology team**: 0.5 - 1 FTEs[2] **Factory team**: 2.5 - 3 FTEs[2]
	Labor Cost	**Scientist**: $93,250 / FTE **RAs**: $62,500 / FTE	**Scientist**: $112,000 / FTE **RAs**: $67,500 / FTE
	Duration	~8 - 10 years	~3 - 5 years
	Output	~100 unique genotypes / week	~1,000 unique genotypes / week
Project Type 2: New Product Development	Staffing	**Scientist**: ~4FTEs[1] **RAs**: ~8FTEs[1]	**Scientist**: 2 - 3 FTEs[1] **RAs**: 2 - 3 FTEs[1] **Factory team**: 2.5 - 3 FTEs[2]
	Labor Cost	**Scientist**: $93,250 / FTE **RAs**: $62,500 / FTE	Scientist: $112,000 / FTE RAs: $67,500 / FTE
	Duration	~18 months	~3 months
	Output	~100 unique genotypes / week	~1,000 unique genotypes / week

1 Fully dedicated team to a single project

2 Partially dedicated shared services across multiple projects

SOURCE: employee and expert interviews; incomes based off of Glassdoor crowd-sourced salaries in San Francisco, CA, team analysis.

PARTNERSHIP ON AI