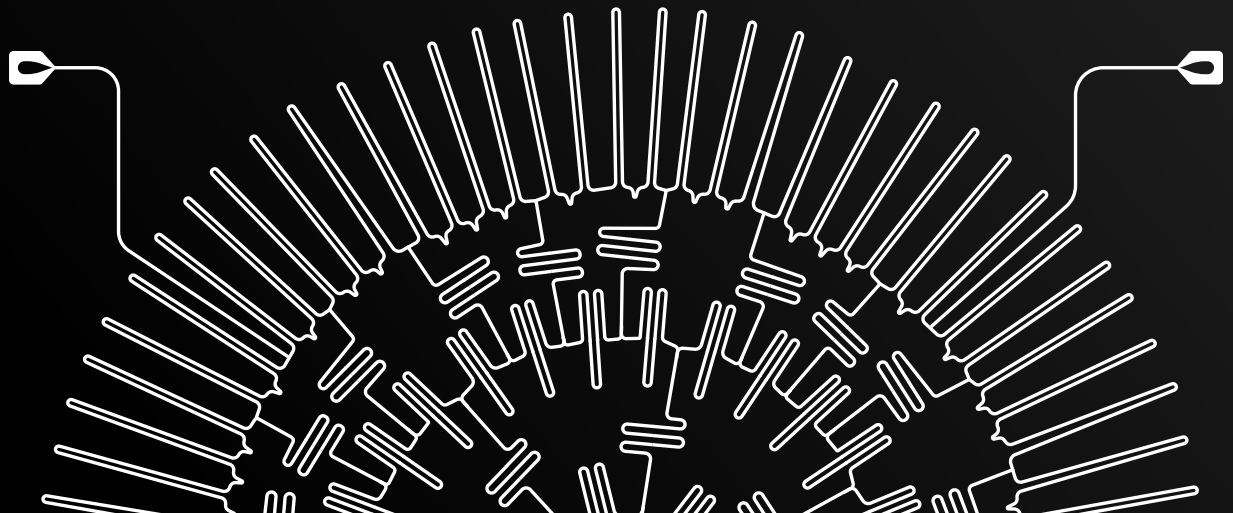
ZAPATA

The Near-Term Promise of **Quantum Generative AI**



Executive Summary:

Recent research suggests that generative AI could be the first domain where quantum computers will provide a practical advantage[1]. In this report, we will:

- Review the state of classical machine learning today and the emerging bottlenecks
- Discuss recent progress in quantum computing technology
- Examine recent research that demonstrates how quantum generative AI can enhance existing classical techniques
- Explore potential applications of the techniques developed by this research

Below is a summary of the report's key findings:

Classical Machine Learning (ML) is Reaching its Limits

Challenges related to data management, computational limits, and operational complexity are constraining further improvements in classical ML, demanding new breakthroughs for progress.

Quantum Computing is Advancing Rapidly

Although we are still in the era of noisy, intermediate-scale Quantum (NISQ) devices, recent advances in hardware, algorithms, error correction and mitigation promise to unlock value in the next 3-5 years.

Generative AI Leads the Race for Practical Quantum Advantage

Generative modeling is a subfield of ML wherein models generate novel data that resembles the data used to train the model, rather than simply classifying data. Quantum computers have the potential to outperform classical computers in these tasks due to their ability to encode and sample from complex probability distributions that are classically intractable.

1. In this paper, any references to quantum advantage will refer to practical quantum advantage, wherein a quantum algorithm or quantum heuristic can outperform state-of-the-art classical methods in a relevant real-world application. This contrasts with an algorithmic quantum advantage, which has been claimed by Google and others.

Quantum and Classical: Better Together

By combining quantum and classical approaches to generative modeling, we can outperform both approaches in isolation. Our research demonstrates this performance boost in three ways: in a hybrid quantum-classical image generation task, in the classically-enhanced training of a quantum generative model, and in improving upon the results of a classical solver for investment portfolio optimization using a quantum-inspired model.

Generative Modeling Enhances Classical Optimization Solvers

Our research demonstrates how generative modeling can boost classical optimization solvers. Any generator could be used, including classical generators, quantum-inspired models, as well as fully quantum models. As quantum hardware matures, more powerful quantum generative models could be plugged into this generator-enhanced optimization (GEO) framework to achieve a more significant advantage over classical solvers in isolation.

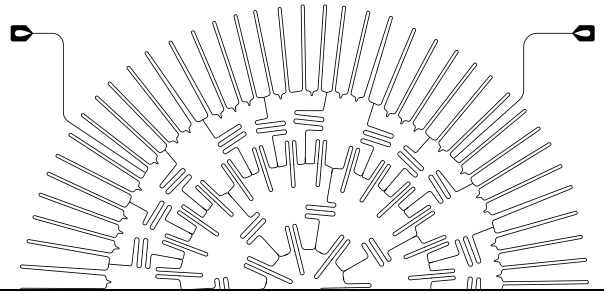
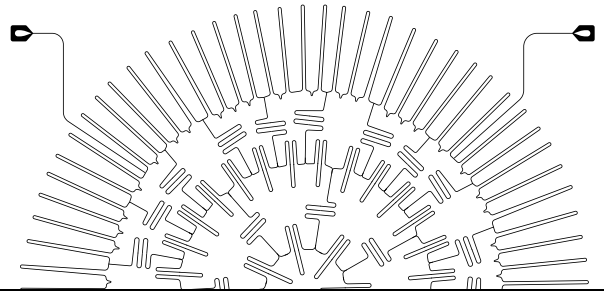


Table of Contents



Introduction

Industry spending on artificial intelligence and machine learning (AI/ML) continues to grow, with IDC predicting that global spending will more than double in the near term – from $85.3 billion in 2021 to $204 billion by 2025.

The leading spenders are the retail and banking sectors, with a focus on automated customer service solutions and recommendation engines for the former and automated threat intelligence and fraud analysis for the latter. However, AI/ML applications are gaining traction in other industries as well, including natural language processing (NLP) and medical use cases such as medical imaging, genetics, proteomics, and drug discovery.

Until recently, most AI/ML applications involved analyzing data, whether for classifying that data or for making predictions. However, the emerging field of generative AI goes beyond simply analyzing data — it generates new data.

Recent examples of generative models such as ChatGPT and DALL-E have generated impressive results in text and image generation, respectively. But generative modeling could do far more valuable work than generating images and text — as we will show in this paper, generative modeling could be used to generate novel solutions to complex business problems.

As adoption grows, companies face several significant challenges when it comes to unlocking the potential of both analytical and generative AI/ML. Those challenges fall into three broad categories:

Data challenges

- Data can be too big: the challenges of storing and managing huge amounts of data can make it hard to train AI/ML models.
- Data can also be too small: Some use-cases will have very few data points (e.g., natural disasters, financial crises, etc.), and the scarcity of data is a problem for training and running models. The challenge is to generate new synthetic data or create models that could be trained with less data.
- The intense operational demands of data preparation, particularly when it comes to labeling data for supervised learning approaches, can slow down ML applications.

Compute challenges

- Requisite computing power drastically intensifies as the number of parameters in models increases. These parameters now reach into the billions with an estimated cost of $1 per thousand parameters.
- The extreme compute costs of generative AI models such as GPT-3 and DALL-E have rendered them expensive and difficult to access.
- As Moore's Law runs out of steam and we reach the limits of circuit miniaturization, we lose the ability to manufacture chips capable of handling the increasing computational load of advanced AI.

Operational challenges

- Researchers and AI-driven companies must rely on an increasingly complex and distributed array of compute and data resources.
- Managing, fine tuning and debugging these distributed systems demands a growing range of expertise from an increasingly limited talent pool.
- Enterprise IT architectures are growing more complex and layered, making it increasingly difficult to integrate new AI applications in a way that adds value.

As AI/ML pushes against the limits of classical computing, quantum computing, even in its current "NISQ" stage (Noisy Intermediate-Scale Quantum), can unlock new frontiers in AI/ML, addressing some of the data and compute challenges outlined above. However, quantum computing also shares some of these challenges, particularly on the operational front.

In this report, we will describe how quantum computing can accelerate and improve certain generative AI approaches, in addition to enhancing classical solutions for combinatorial optimization problems, such as portfolio optimization. We will also introduce new tools for operationalizing and deploying quantum-enhanced ML applications in production at enterprise scale, overcoming some of the challenges outlined above. Based on our findings, we believe that any organization currently investing in AI/ML at the enterprise level will find quantum computing to be increasingly important when it comes to ongoing progress and, ultimately, transformative breakthroughs.

The State of AI Today

Surveying the current state of the field in their most recent Hype Cycle for AI report, Gartner notes a recent shift to a data-centric approach. As Gartner puts it, "the AI community has traditionally focused on improving outcomes from AI solutions themselves, but data-centric AI shifts the focus toward enhancing and enriching the data used to train the algorithms." This shift in focus acknowledges the growing data challenges outlined in this paper's introduction that are constraining further AI development.

In confronting the challenge of data scarcity, Gartner highlights the growing innovation in synthetic data: data that is artificially generated rather than empirically observed. In many cases, synthetic data used to train machine learning models is itself created by generative machine learning models. Gartner predicts a massive increase in synthetic data adoption due to its avoidance of personally identifiable information, reduced costs and time savings, and positive impact on ML performance.

Gartner also points to recent innovations in operationalizing AI as business applications, referring specifically to advances in ModelOps, AI cloud services, and decision intelligence. As Gartner describes it, decision intelligence refers to the practical discipline of improving decision making by "explicitly understanding and engineering how decisions are made and how outcomes are evaluated managed and improved via feedback."

Operationalizing AI in a way that measurably enhances decision making has been an ongoing challenge for AI, one that Gartner also highlighted in their 2021 Hype Cycle report. It is no longer enough for AI to be a proof-of-concept — it must deliver real business value. Given the growing complexity of enterprise data and compute architectures, the task of integrating practical AI applications has become increasingly difficult. According to Senior Principal Analyst Shubhangi Vashisth, "on average, it takes about eight months to get an AI-based model integrated within a business workflow and for it to deliver tangible value."

"On average, it takes about eight months to get an AI-based model integrated within a business workflow and for it to deliver tangible value."

Shubhangi Vashisth, Senior Principal Analyst at Gartner

The State of AI Report, compiled by investors Nathan Beinach and Ian Hogarth, comes to similar conclusions about the emerging challenges. In 2021, they noted: “with the increasing power and availability of ML models, gains from model improvements have become marginal. In this context, the ML community is growing increasingly aware of the importance of better data practices, and more generally better MLOps (DevOps for ML), to build reliable ML products.” In other words, improvements in ML performance are no longer coming from model improvements, but from new efficiencies in data management and MLOps practices.

The industry shift from building models to running models was already underway in 2020, when Beinach and Hogarth noted that “25% of the 20 fastest growing GitHub projects in Q2 2020 concern ML infrastructure, tooling and operations.”

Despite growing spending, Beinach and Hogarth also draw attention to looming limitations in the field. In some cases, as we mentioned at the outset, these limitations are associated with data. For some models, training data is relatively scarce — for example, a model predicting stock market crashes would have very limited sample size of training data. In these cases, synthetic data would be valuable for augmenting the training data. In other cases, the problem is too much data. Beinach and Hogarth note that “working with massive datasets is cumbersome and expensive,” and that while careful data selection can focus resources on the most valuable examples, “classical methods often become intractable at scale.”

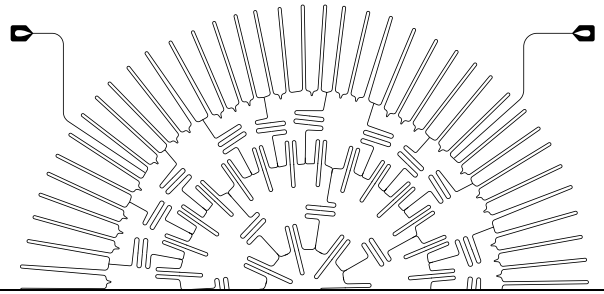
Greater limitations exist on the compute side, however. Outrageous computational, economic, and environmental costs are increasingly required to gain incrementally smaller improvements in model performance. Case in point, an MIT-led study showed that, given the current state of technology, it would cost over 100 quintillion dollars and 50 quadrillion tons of carbon emissions (or, a billion times the average annual global emissions) to decrease the ImageNet error rate from 9% to 1%.

The MIT study focuses on the “voracious appetite for computing power” required by deep learning models and states that advances here are “strongly reliant on increases in computing power.” This reliance, they say, “reveals that progress along current lines is rapidly becoming economically, technically, and environmentally unsustainable.”

Rather than expanding existing ML models along their unsustainable trajectory, novel ML techniques will be required to achieve further progress. The good news is that compute infrastructure is evolving to support AI initiatives, according to Gartner.

This evolution involves everything from GPU accelerators, application-specific integrated circuits and more exotic compute technologies such as neuromorphic hardware.

Quantum computing is another breakthrough technology that could have a real role to play in alleviating these problems. Indeed, major players such as IBM and Goldman Sachs are already incorporating quantum computing into their roadmaps to accelerate growth in compute power, efficiency and utility. In the last few years, novel approaches in the design and implementation of quantum computing devices have proliferated rapidly and could deliver new breakthroughs in machine learning.



What is Quantum Computing?

Classical computers encode information in “bits” that can have one of two values: 1 or 0. This means that any two bits can hold four potential values: 00, 01, 10, 11. While that might not sound like a lot, if you have a 32-bit processor in your computer, that processor can handle over 4.2 billion different combinations of values. Still, by their very nature, each bit can only hold one value at a time.

Quantum computing encodes information in “qubits.” While a qubit can also represent a 1 or 0, it can also be a blend of both states at once due to a quantum property called superposition. Qubits can also be entangled, meaning the state of one qubit can be influenced by that of another qubit. Thanks to superposition, entanglement and another property known as interference, there are certain problems a quantum computer can theoretically solve much faster than a classical computer.

Several engineering challenges, particularly the fragility of quantum states that makes qubits “noisy” or prone to error, mean that quantum devices need to devote considerable time and resources to error correction. As a result, these devices are not yet generally faster than their classical counterparts. Nevertheless, we are entering a phase in which today’s quantum devices can outperform their classical counterparts in specific circumstances.

“Quantum supremacy” refers to the point when a quantum device outperforms classical devices. Researchers at both Google (2019) and the University of Science and Technology of China in Hefei (2020) claim to have achieved quantum supremacy. In Google’s case, they used a 53-qubit device to sample one instance of a quantum circuit a million times in about 200 seconds — a task that, they asserted, would take one of the world’s fastest supercomputers 10,000 years to compute[2]. In the Hefei demonstration, the team used a photonic quantum computer to perform a calculation in 200 seconds that would take a world-class supercomputer 2.5 billion years to complete.

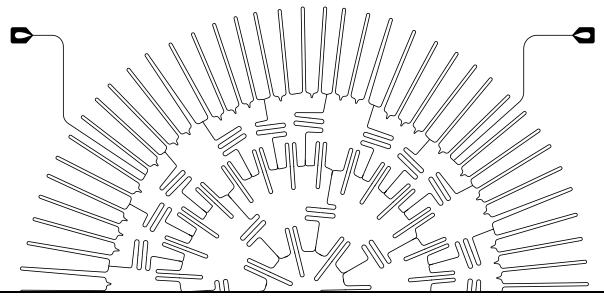
While these demonstrations of quantum supremacy are contentious in their practicality, the recent progress in quantum hardware development is undeniable. Earlier this year, Quantinuum’s System Model H1-1 set a record with a Quantum Volume of 8192, a benchmark introduced by IBM in 2019 to measure the overall capability and performance of quantum computers. IBM itself recently released its 433-qubit Osprey processor and will unveil Condor, the first quantum computer with over 1,000 qubits, by next year.

2. Detractors at IBM, who developed the Summit supercomputer, were quick to rebut that it would take closer to two days for Summit to complete the calculation. In 2022, researchers in China demonstrated that the same calculation could be done in a few hours on ordinary processors, and using the same approach, a supercomputer could do it in seconds.

There are also now more distinct quantum computing architectures than ever. In addition to the superconducting quantum computers developed by companies such as IBM and Rigetti, there are trapped ion computers from companies including IonQ and Quantinuum; quantum annealing devices from D-Wave; neutral or cold atom architectures from QuEra, ColdQuanta and others; photonic architectures from the likes of Xanadu and PsiQuantum; and silicon spin qubit architectures from companies such as Quantum Motion and Silicon Quantum Computing. In addition to fully quantum architectures, there are also quantum-inspired options available, such as NVIDIA's cuQuantum, which simulate quantum circuits on classical computers.

Meanwhile, Google recently marked a major milestone in quantum error correction. By grouping several physical qubits together, they created a logical qubit with less error than the physical qubits on the chip. This is an essential step to creating fully fault-tolerant quantum computers that can outperform classical computers.

The growing capability of quantum computing devices and algorithms positions them to meet the market's voracious appetite to move AI/ML applications past their emerging limits. Indeed, we at Zapata have recently demonstrated how quantum-inspired techniques can already boost the performance of classical generative ML models.



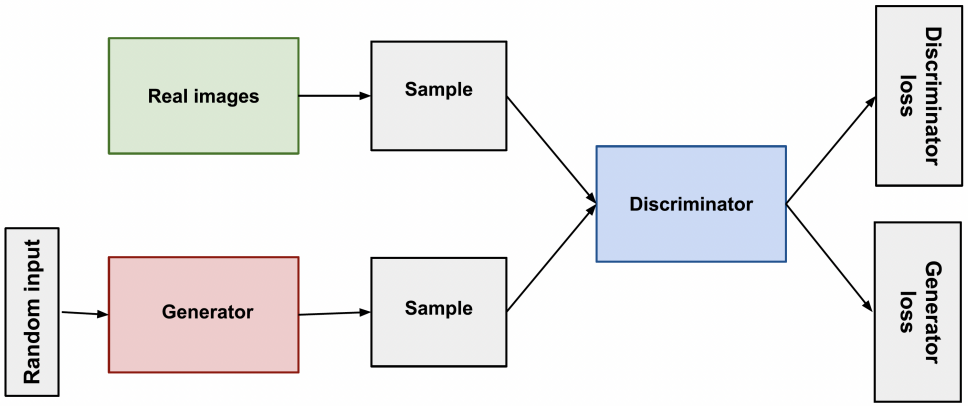
Generative AI: A Leader in the Race for Quantum Advantage

As previously mentioned, generative AI refers to the application of ML models to generate synthetic data that falls within the probability distribution of some existing data of interest. In other words, generative models generate new data that resembles the data they were trained on, without repeating it.

In this way, generative AI can create, for example, new faces based on a training set of faces, or new passages in the style of an author based on a training set of that author's work. Examples of generative models include OpenAI's GPT-2, GPT-3 and ChatGPT which generate text, OpenAI's DALL-E, which generates images from text, and StyleGAN 2, which produces faces of people who do not exist. Taken further, generative models could also generate novel solutions to difficult problems of business value. We will return to this application later.

A Closer Look: Generative Adversarial Networks

The "GAN" in StyleGAN stands for generative adversarial network. GANs were first introduced in 2014 and, in addition to generating images of faces, have since been used to generate images based on text, enhance the resolution of blurry images, and generate 3D objects. So, how do they work?



The basic structure of a GAN. Source: Google

At the most basic level, a GAN consists of two neural networks, one called the “generator” and the other called the “discriminator.” To develop a GAN that can produce novel images, the generator and discriminator are trained together on real images in an adversarial game. The discriminator’s job is to decide which images are real and which are imposters created by the generator. Meanwhile, the generator’s job is to create fake images good enough to trick the discriminator.

Using only feedback from the discriminator, the generator gradually learns to create images that the discriminator can no longer tell apart from real images. As soon as the discriminator’s ability to discriminate approaches guessing — that is, when it’s correct about 50% of the time — the generator has effectively stumped it and can no longer continue improving. At this point, it must be decided whether the data generated by the generator is good enough or if the GAN must be recalibrated.

As GPT-3, DALL-E, and StyleGAN make clear, classical computers are certainly competent at generative modeling. But that doesn’t mean they can’t be improved by quantum methods. In fact, generative modeling is one the most promising areas where quantum computing could add value in the near term.

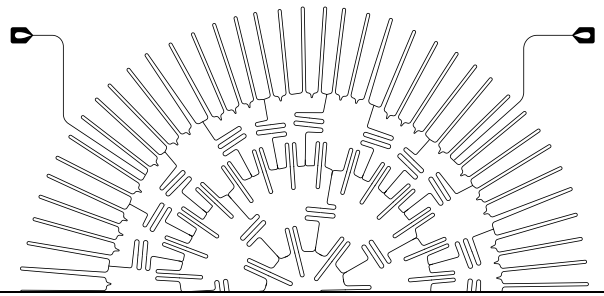
At a high level, the reason quantum computers have the potential to excel at generative modeling is because they are much better at encoding and sampling from complex probability distributions compared to classical computers. This is due to quantum computers' ability to express distant correlations through quantum entanglement, as well as the inherent probabilistic nature of measuring a quantum state. In fact, extensive theoretical research has shown that quantum neural networks can encode and sample from probability distributions that are not tractable classically.

This ability stems from the fact that a quantum state can itself be described as a series of probabilities. For example, if you measure the state of a single qubit, it could be 30% likely to collapse to one state and 70% likely to collapse to the other. With multiple qubits, each of the exponentially large number of possible measurement outcomes will have different relative probabilities. With a quantum computer, we generally have control over these percentages.

Quantum generative models also allow for an efficient sampling procedure from these probability distributions — that is, they allow you to generate data samples that exactly follow the encoded probability distribution without any bias or additional algorithmic steps. This is due to the "wavefunction collapse" phenomenon, wherein the quantum state collapses to one state or another when it is measured. While wavefunction collapse often represents a challenge for quantum computing practitioners, it is actually a valuable resource in the case of generative models. Even if a classical and a quantum model

encoded the same distribution, it is possible that the quantum model would be much more proficient at generating data from it.

Recently, our own research at Zapata has demonstrated clearly how quantum computing can improve upon the results of classical generative models.



Harnessing the Generative Power of Quantum

Generative models work by capturing essential underlying features of the training data and using those features to produce similar, but novel data. Our hypothesis was that quantum generative models could better capture these essential features because quantum devices can express data distributions that are out of reach for classical devices.

To test that assumption, we conducted several experiments resulting in published research papers, three of which we will discuss here. The first involved the first ever generation of high-resolution images using a quantum computer[3]. The second concerns a shortcut to quantum advantage in generative modeling, in which we used a quantum-inspired classical model to enhance the training of a quantum model. Finally, the third showed how quantum-inspired generative modeling could be applied in practice to boost the results of a classical solver for combinatorial optimization problems, using an investment portfolio optimization problem as a benchmark.

Generating Handwritten Digits

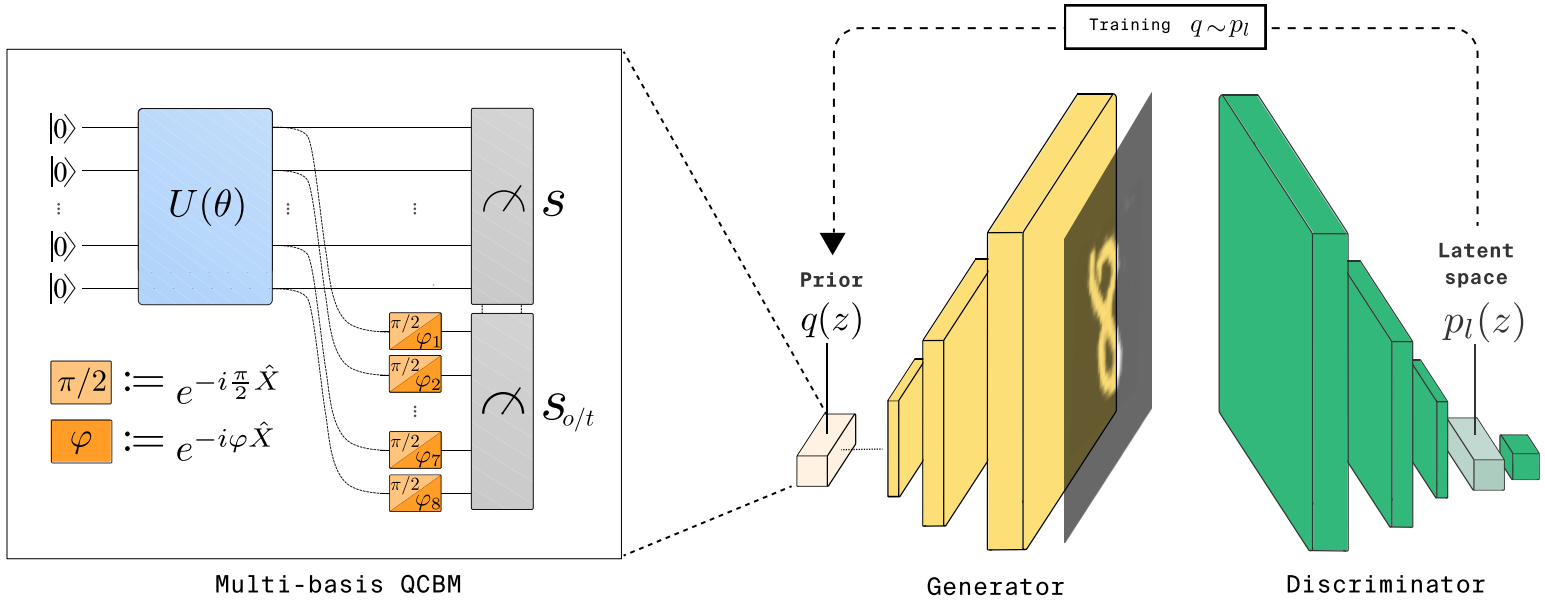
We described earlier how, in a GAN, the generator tries to "trick" the discriminator. The generator does so by taking the output of the discriminator – its identification of images as real or fake – and uses that to improve the images it generates. We also mentioned that sometimes a point is reached where the output of the discriminator becomes less and less helpful and the whole network needs to be recalibrated.

We hypothesized that, by modifying the GAN and inserting a quantum circuit, we could link the latent space of the discriminator (which provides a condensed, compressed description of the input data that approximately captures the data in fewer variables) with the prior distribution of the generator (the random component in the generator that drives it to create unique data) — thus improving the training of the model. First, we used a modified version of a GAN known as an associative adversarial network (AAN), which gives the generator indirect access to the latent space of the discriminator. From there, we modeled the prior distribution of the generator using an approach we invented known as a multi-bases quantum circuit Born machine (QCBM).

In this framework, the information from the latent space is encoded as a rich probability distribution by the QCBM, such that it can then provide the generator with samples that resemble the discriminator's understanding of the data. This made it easier for

3. Huang et al. previously demonstrated a quantum algorithm that could generate images of handwritten digits using a NISQ device, but the images generated had significantly lower resolution.

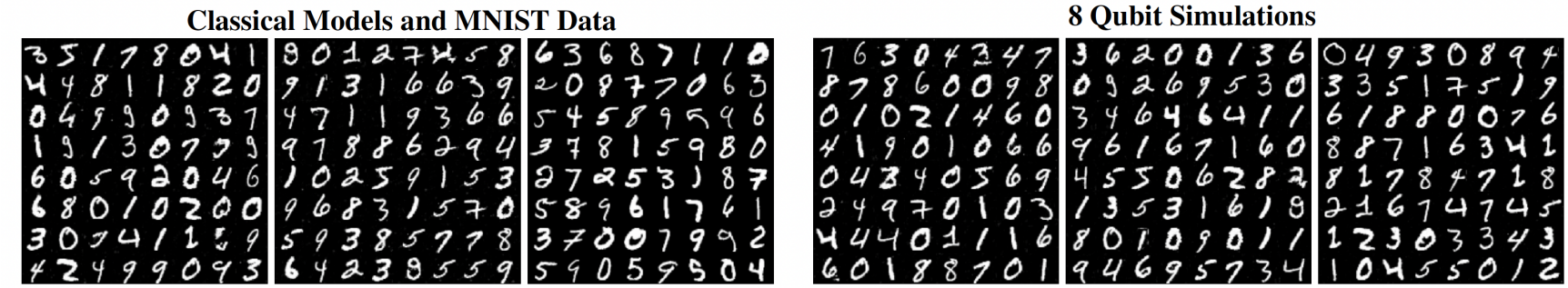
the generator to accomplish its challenging learning task and improved its ability to fool the discriminator. The quantum-enhanced machine learning model we implemented looked like this:



The multi-basis-enhanced Quantum Circuit Associative Adversarial Network (QC-AAN) framework.

We trained this hybrid (i.e., quantum-enhanced) algorithm on the MNIST dataset of handwritten digits, a canonical dataset for evaluating generative models. Using IonQ's 8-qubit ion-trap quantum device, we were able to generate high-quality images that were competitive with those generated by comparable classical GANs, as well as conventional AANs utilizing Restricted Boltzmann Machines (another type of classical generative model) trained on the same data.

In the figure below, the Inception Score (IS) evaluates the quality and diversity of images generated in GANs. A model with a high IS produces very diverse images of high-quality handwritten digits. As you can see, the IS scores of the quantum-enhanced AANs matched or exceeded those of the classical GANs.



To the best of our knowledge, this is the first practical implementation of a quantum-classical algorithm generating high-resolution images on a NISQ device.

To be fair, one might be able to achieve better classical ML performance by choosing more sophisticated variants of GANs. However, for this simple demonstration on a relatively early 8-qubit quantum device, we believe the comparison with classical GANs captures the unique capabilities of quantum devices and is appropriate. The point is that this paper demonstrates a framework for testing practical quantum advantage in generative modeling on today's hardware for real-world use cases — and provide competitive results.

This is just the beginning of what quantum could accomplish, and the work we have done is scalable to both larger and more sophisticated classical models and quantum devices. Considering companies like IBM are planning to unveil devices with over 1,000 qubits by 2023, our results illustrate how a QC-AAN framework could measurably improve the performance of classical GANs and advance the capabilities of generative modeling.

This work has since been published in Physical Review X.

Busting a Common QML Misconception

When people hear the term "quantum machine learning," it's not uncommon for people to assume this refers to classical data loaded into an ML algorithm running fully on a quantum device. It thus seems natural to write-off QML because such techniques will be limited by low qubit counts and high error rates for the foreseeable future. While this is true for some approaches, it's not true for all. In the near-term, quantum devices will add the most value when they are used for only a small piece of an otherwise fully classical ML workflow. This small-quantum-slice approach has been the case for all of Zapata's research demonstrating the potential for a quantum advantage in ML.

In these small-slice cases, quantum devices are only used for processes where there is a limit to classical computing power, while the rest of the ML workflow is fully classical. As an example, quantum computers are well suited for generating randomness within a complex probability distribution for data that would be intractable to represent classically (the dimensionality of that distribution is limited by qubit counts, however). If you load this distribution (imagine weighting some dice to bias their outcomes) and then take samples from it (rolling those dice), we observe improved performance for the (classical) generative machine learning model that this quantum sampler is feeding. It generates samples (for example: images of handwritten digits) that the same generator without the quantum sampler would be unable to find.

Shortcutting Quantum Advantage in Generative Modeling

Trainability refers to the ability to tune the parameters of a model to fit the data of interest reliably and accurately. To fully exploit the benefits of quantum computing for generative modeling, we need to be able to train these models well. If we can't train a model to generate data that aligns with the training data, then it has poor trainability, and is of little use.

Quantum circuits for generative modeling, as with other applications of parametrized quantum circuits, are known to have many training issues, including barren plateaus. To understand barren plateaus, imagine that training a model is like trying to climb Mount Everest without knowing where it's located. If you start randomly walking, you won't know which direction to walk to get closer to Everest. This is where the training gets stuck: modifying the parameters in any direction doesn't yield any significant changes in the output, so the model doesn't know in which direction to "walk towards" to improve its results.

But the closer your starting location is to Everest, the more likely you are to make it to the top. Similarly, we hypothesized that if you could start training your model from a good place using classical methods, you could improve the subsequent training of the quantum model. As it turned out, our research demonstrated that combining classical and quantum methods could improve the training of these quantum models, overcoming barren plateaus to outperform both classical and quantum methods in isolation.

In our research, we used a tensor network-based generative model, whose architecture is known as a matrix product state (MPS). This is a quantum-inspired model which runs on a classical computer. MPSs are much easier and cheaper to train than quantum circuits, but they can't model long-range correlations between bits without using an exponential amount of computing resources. In contrast, quantum circuits are much better at expressing distant correlations but are much harder to train, due to barren plateaus.

Our approach was to first train an MPS as much as we can, then map it into a quantum circuit, extend the quantum circuit with additional quantum gates that would not be feasible on classical computers, and continue training on a quantum computer. The MPS learns the short-range correlations, and once it reaches the limits of its computational resources, it passes a relatively well-trained model to the quantum circuit. The quantum circuit then continues the training with the long-range correlations.

To give an analogy, imagine the quantum circuit is the wings of an airplane, and the MPS is the wheels. Obviously, an airplane can't start flying immediately from rest. We use the wheels to gain speed until the plane can start using the wings. The wings can take the airplane to places that the wheels alone couldn't possibly carry it. But it's the combination of wheels and wings which helps the airplane get to its destination. In the same way, the synergy between classical and quantum is what will ultimately expand what's possible for both in generative modeling.

This work represents a paradigm shift in the race for practical quantum advantage. It is no longer a question of quantum vs. classical, but rather how the two can be used synergistically together to get better results, faster.

Additionally, such a framework gives rise to a well-defined notion of practical quantum advantage. If a classical model cannot be further optimized on classical hardware to solve a task of interest, and a quantum computer is able to improve on its performance, there is real practical value generated using a quantum computer that could not have been attained without it.

Generator-Enhanced Optimization

Generative modeling is not limited to generating images and text. As we will demonstrate here, it can also generate novel solutions to complex optimization problems that are common across industries.

For example, consider the process of creating optimal investment portfolios. Not only are there countless investment options, but predicting the future performance of any particular investment requires modeling dozens of scenarios involving countless variables. Improved approaches to this problem and similar optimization problems would have a transformative impact on the field of quantitative finance and other industries with complex optimization problems.

Today, quantum and quantum-inspired generative models can boost classical solutions to these complex combinatorial optimization problems. They can do this by improving on the ability of classical models to generalize from the training data and uncover new solutions to the problem.

In generative machine learning, generalization refers to the ability of a model to generate novel data resembling the training data. In other words, it asks if a model truly understands the essential features of the training data, rather than simply memorizing and repeating the training data. In the context of portfolio optimization, better generalization results in the generation of risk-optimized portfolios that were previously unconsidered.

Similar to the QC-AAN we used to generate high-resolution digits, our recent work combined a classical approach to solving this optimization with a quantum-inspired “booster.” In this case, we used MPS, the quantum-inspired model from the previous section, to improve the output of a classical solver for portfolio optimization. In doing so, we were able to expand the possibilities of potential portfolios. Using real data from the S&P 500 and other financial stock indexes, we showed how our quantum-inspired model could generate portfolios with the same level of return but with lower levels of risk than the portfolios generated by the classical solvers alone.

We called this strategy “Generator-Enhanced Optimization”, or GEO for short. To the best of our knowledge, this is the first demonstration of the generalization capabilities of quantum-inspired generative models in the context of a real-world application in an industrial-scale setting. For more details, see the research paper we published on this work

“GEO shows that generative AI can do more than generate text and images: it can enhance existing solutions to complex optimization problems, working with classical and quantum-inspired generative models today, or with powerful, fully quantum models in the future"
Alejandro Perdomo-Ortiz, Research Director, Quantum AI

We have since gone beyond this first demonstration, developing tailored metrics for generalization that showed superior performance for the MPS over a GAN in solving the portfolio optimization problem. With these metrics, we finally have a quantitative tool to evaluate practical quantum advantage in generative modeling.

More recently, we used these metrics to demonstrate the generalization capabilities of real quantum generative models, particularly the QCBM, the same generative model used in the QC-AAN framework to generate high-resolution images of digits. These works are the first promising steps towards a more thorough study of the generalization capabilities of different classes of generative models in the quest for practical quantum advantage.

Significantly, our research demonstrated that QCBMs can generate samples with higher quality than the samples in the training set. The ability to generate higher quality samples has major implications for solving combinatorial optimization problems, such as the portfolio optimization problem. A QCBM could be swapped in for the quantum-inspired model used in our GEO strategy, which already could improve upon results from classical solvers. In other words, as QCBMs become more powerful, they could be a valuable asset in the race for practical quantum advantage through our GEO strategy.

More recent work has since improved our ability to solve optimization problems with quantum-inspired generative modeling.

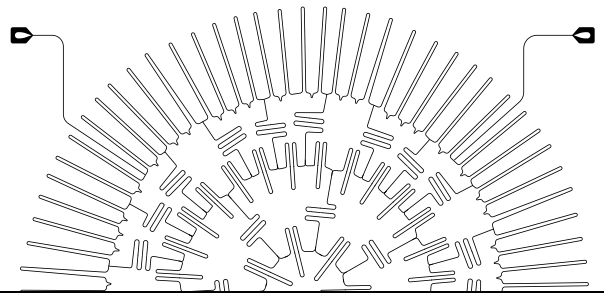
Many combinatorial optimization problems have a high number of constraints that a solution must satisfy to be considered valid, known as equality constraints, which narrows the space of valid solutions. But rather than searching only for solutions within that valid space, traditional optimization solvers often generate many possible solutions that may or may not be valid, resulting in expensive and inefficient searches.

Oftentimes, equality constraints dramatically reduce (sometimes even exponentially so) the likelihood that you generate valid samples if you don’t have a native way of handling them. Our recent work addresses this problem.

In our recent work, we developed an approach to encode equality constraints directly into a symmetric MPS, a variant of the quantum-inspired model we mentioned earlier. By utilizing sparse tensor structures that are imposed by the problem's symmetries, the MPS is constrained to only output valid samples.

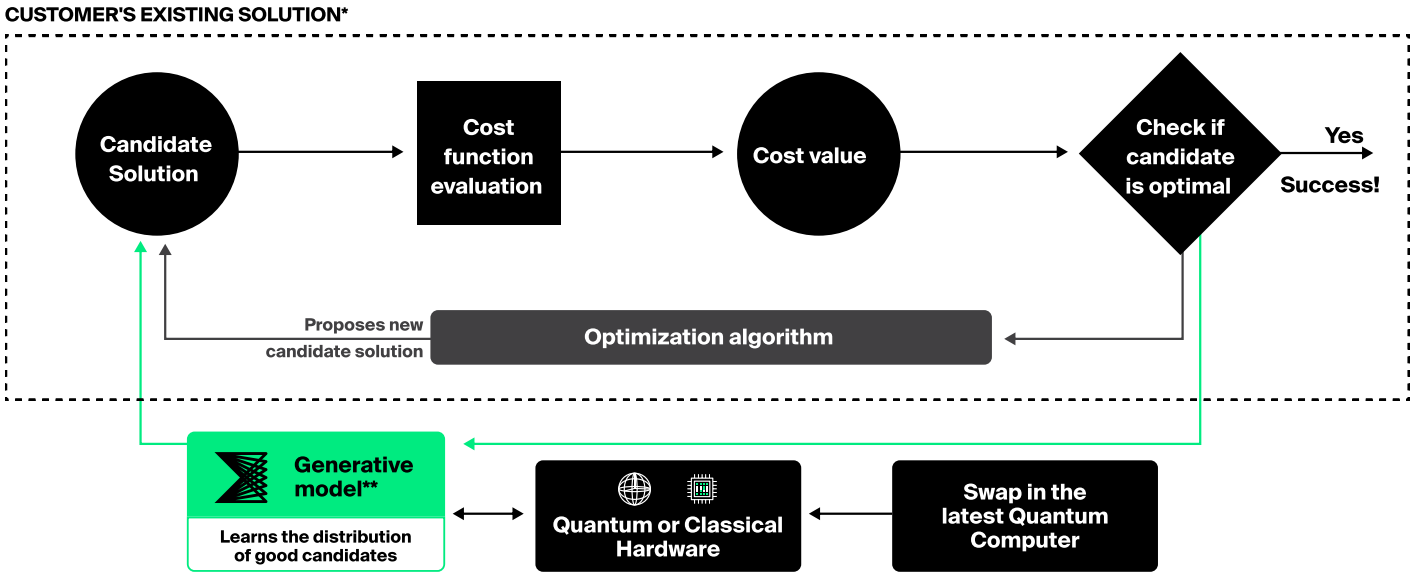
Directly encoding the hard constraints of the problem into your model can significantly improve its ability to adequately tackle highly constrained tasks by, for example, improving the model's generalization behavior, because it can only generalize into the valid space. Importantly, we found that the constrained MPS outperform their standard unconstrained counterparts in finding novel and higher quality solutions to combinatorial optimization problems.

Additionally, while most traditional methods suffer when the number of constraints is increased, since more constraints need to be validated, our new approach actually benefits from more constraints, because the more constraints we have, the sparser the generative model can be parametrized, which in turn leads to better computational performance with significantly reduced computational resources.



Applications of Quantum Generative AI

Zapata's Generator Enhanced Optimization (GEO) is designed to boost customers' existing optimization solutions



*GEO can augment customers' existing solutions, or Zapata can help build the classical optimizer in cases where it does not yet exist.
**Goal is to provide better candidate solutions with fewer cost function evaluations

Applications of GEO

Portfolio optimization is just one of many combinatorial optimization problems where GEO can be used to boost classical solvers. Existing classical solvers, such as Gurobi and CPLEX, can already deliver highly effective solutions for optimization problems across industries. GEO does not compete with these classical solvers — it enhances them.

The GEO approach can use any generator, quantum or classical. In the near-term, generators will likely be classical or quantum-inspired, for example using MPS as we demonstrated in the portfolio optimization research. But as quantum hardware matures, users that have already built their GEO application workflow will be able to plug in more powerful quantum generators to achieve a more significant advantage.

Potential use cases by industry:

Energy & Chemistry	Manufacturing	Consumer Goods	Health & Pharma	Finance	Logistics & Transportation	Telecom & IT
Unit Catalyst Optimization	Manufacturing process optimization	Inventory optimization (finished goods)	Molecular structure analysis	Capital allocation optimization	Delivery routing optimization	Communications traffic optimization
Smart grid optimization	Design process optimization	Merchandizing optimization	Active site modeling and optimization	ATM cash supply optimization	Workforce scheduling optimization	Workflow optimization
Chemical reaction process Optimization	Preventative maintenance	Advertising Optimization	Healthcare resource optimization	Asset pricing optimization	Warehouse optimization	Network optimization
Well drilling optimization	Inventory optimization (raw materials)		Clinical trial patient feature selection	Trading optimization	Rideshare route optimization	ETL optimization
Mine location Selection	Factory floor automation		Protein and ligand design optimization	Arbitrage optimization	Supply chain optimization	Spark optimization
Battery composition optimization			Ambulance dispatch optimization	Loan portfolio optimization		

Molecular Discovery

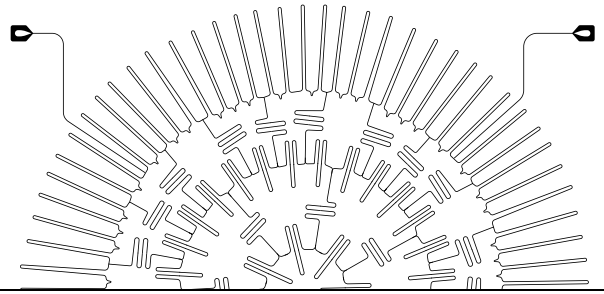
Developing a new drug can take 10-15 years and cost billions of dollars. Synthesizing and testing thousands of potential molecules in search of promising candidates for further study accounts for about a third of that overall expenditure. The discovery and development of new molecules in other fields of materials science involve similarly lengthy pipelines. To speed up the molecular discovery process, generative models could generate simulated molecular structures. These simulations could be optimized for desired chemical and physical properties, such as an affinity for binding with a target receptor on a particular pathogen. Classical, quantum-inspired, and hybrid-quantum models, such as GANs, QC-AANs or other quantum-enhanced generative models, could be used to explore the full range of possible molecular structures in chemical space very efficiently.

Our experiments with image generation have shown that a quantum-boosted AAN can search the latent space more effectively and thoroughly than the classical GAN alone to generate higher-quality and more diverse images. This makes quantum-boosted AANs great candidates to search the space of possible molecules and generate higher-quality candidate molecules as a result. This technique could be applied for chemical and materials discovery beyond the pharmaceutical industry as well.

Synthetic Data Generation

It can often be difficult to obtain comprehensive data for training discriminative or predictive ML models, for instance those used for quality control assessment, maintenance predictions, financial forecasts, or disaster modeling. Sometimes this is due to the rarity of certain scenarios and the resulting scarcity of data, in other cases it is due to the cost of obtaining data. In such cases, it would be valuable to generate synthetic data to augment the datasets used to train these models.

Research suggests that quantum-boosted generative modeling could generate synthetic data that expresses a wider range of probabilities than would be possible with a purely classical model. While the dimensionality of these synthetic data probability distributions would be limited by low qubit counts, quantum computers will be able to express more complex distributions as they become more powerful, and quantum-inspired or quantum-classical hybrid models can already express larger dimensionalities. This will ultimately enable these predictive and discriminative models to be trained more comprehensively, allowing them to identify and predict rarer scenarios.



Turning Research into Code: Operationalizing QML

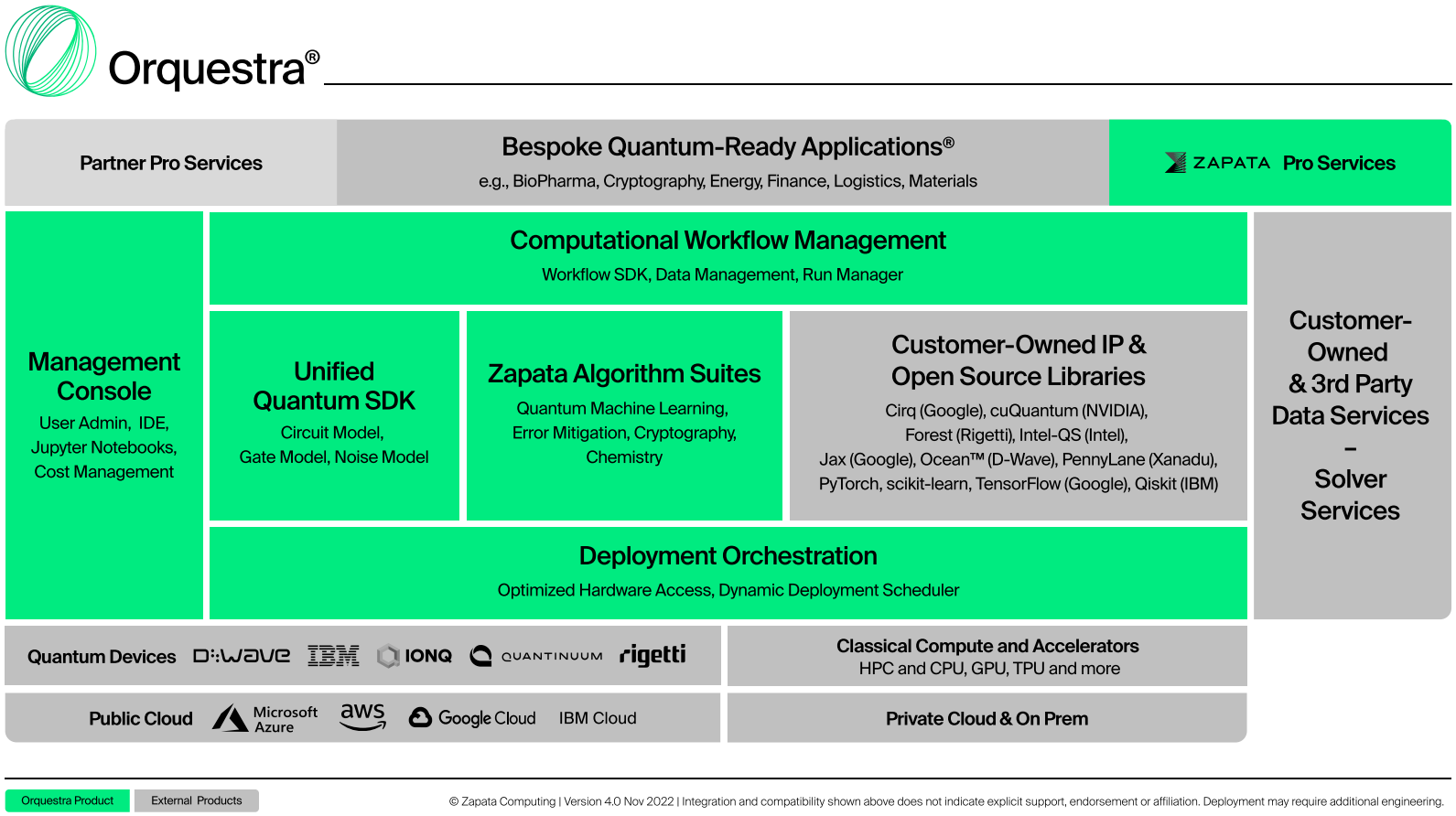
Generative modeling could provide the fastest path to practical quantum advantage, whether in solving optimization problems through GEO or other generative modeling applications. However, scaling challenges abound. Demonstrating practical quantum advantage in a research setting is one thing, deploying a quantum-enhanced application that can deliver value at enterprise scale is another.

As we noted early on in this paper, citing Gartner and Beinach and Hogarth's State of AI report, the complexity of operationalizing ML and managing data, models and compute is the greatest challenge to unlocking the full potential of ML. This remains true with classical ML and is even more true when quantum resources are added to the mix. Any quantum advantage can easily be negated by inefficient data and compute architectures.

At Zapata, we built our software platform Orquestra® to address the operational challenges of deploying enterprise ML applications in addition to quantum-specific challenges.

The Orquestra Platform Architecture

Orquestra supports the entire ML application cycle — from research to development to deployment — by providing a framework for modular computational workflows. Workflows abstract the complexities of operating ML applications by codifying each step into containerized tasks that can be run in parallel or serially. With modular workflows, any component of an application can easily be substituted: data, models, or compute.



This enables users to test different hardware backends, deploy the hardware best suited for their specific application, and swap in more powerful devices when they become available. Orquestra also addresses the data challenges associated with ML and is optimized for data velocity, accommodating data from data warehouses, public and private clouds, on-premises and edge sources in addition to requisite ETL and data cleaning processes.

The modular approach also applies to models and algorithms, whether from open-source libraries or Zapata's own proprietary algorithm libraries. The Quantum Machine Learning suite is the latest addition to the Zapata algorithm library. It takes all the previously mentioned research in QML and makes it accessible to customers as code for building quantum-ready applications.

The Orquestra QML Suite

Software at scale

- Single source of truth in code across different research projects
- Interoperability for multiple ML frameworks (Jax, Torch, Julialang) and quantum backends

Research at scale

- Composition of applications from a library of components
- Simple interfaces allow us to plug-and-play new experiments and applications
- Extension with new models, apps and features

Computing at scale

- Deploy applications and experiments for large-scale benchmarking to accelerate the R&D cycle
- Easily scale GPU and compute resources according to your needs

The Orquestra QML Suite includes a catalog of quantum and quantum-inspired generative models for building quantum machine learning applications. This includes purely quantum generative models, such as QCBMs, and quantum-inspired models, such as MPS, that were used in the previously mentioned research on high-resolution digits generation and GEO, respectively. These quantum and quantum-inspired models can be benchmarked against conventional neural-net based models in Orquestra.

Data Scientists and Machine Learning Engineers familiar with packages like PyTorch, Tensorflow or XGBoost will be able to get up-to-speed on QML Suite easily. We provide cookie cutter examples for each application that can be run with the click of a button. Each example can be configured and customized with simple extensions of our framework, for example to use a different objective function, dataset, or model architecture. Once these models are integrated in application workflows, the platform orchestrates their deployment across all the necessary quantum, classical, and hybrid compute resources.

We at Zapata believe quantum generative modeling will be one of the fastest paths to quantum advantage, and we built the QML Suite to help our customers achieve it by harnessing the latest breakthroughs in QML research. To get started on building quantum-ready applications for machine learning and optimization, get in touch today.

Request a briefing for near-term QML use cases in your industry

Contact info@zapatacomputing.com to get started

About Zapata Computing

Zapata Computing, Inc. is the enterprise quantum software company solving the world's most computationally complex problems with quantum and other forms of Big Compute™. Our mission is to deliver first-mover quantum advantage to the enterprise. Our platform, Orquestra®, enables users to build quantum-ready applications that leverage the most advanced classical, quantum-inspired and quantum technology to solve real-world problems. The Quantum AI team is the largest research group at Zapata, and our quantum generative AI/ML IP portfolio ranks second in the US as of 2021. Zapata has pioneered new methods in ML, optimization, and simulation to maximize value from near-term quantum devices, and work with ecosystem hardware and software providers such as Amazon, Nvidia, D-Wave, Google, Quantinuum, IBM, IonQ and Rigetti.

For more information, visit

www.zapatacomputing.com