Alix VENTURES

Market Deep Dive Report

AI Enabled Drug Discovery

May 2020

Contents

1. Summary

Artificial intelligence is poised to exert a transformational impact on the global drug discovery pipeline. AI driven technologies combined with the rapid expansion of available datasets are already helping to screen drugs more quickly and with lower attrition in clinical trials. In the future, they will be used to 'imagine' new compounds, further offloading cost and time spend from a process that typically takes over 5 years and billions of dollars. The market environment for such technologies is favorable, as large biopharma companies increasingly look to outsource R&D and frequently partner with or acquire emerging companies and their proprietary technologies. Internal efforts at industry leading biopharmas are increasingly data centric, and have taken a portfolio approach, making partnerships with multiple AI companies a preferred strategy.

AI drug discovery is still a nascent field, the greatest challenge will be overcoming technical milestones. Though funding is accelerating rapidly (31.9% CAGR through 2027), the industry is in short supply of research talent due to the need for a multidisciplinary skill set and multiple adjacent competing industries in the AI space. Specifically in de novo AI drug design, serious limitations remain, such as a lack of chemical diversity, dangerous reactive groups, and the generation of compounds that are not synthetically feasible. Another challenge is the guarded industry culture around data sharing. Without access to large volumes of data, training and validation of AI models will remain lackluster, and early stage startups will have difficulty competing with existing companies who have many more years worth of data.

Still, this report finds a favorable opportunity for strong teams to make an impact on the drug discovery industry. The key will be finding the multidisciplinary talent and a management team with industry ties in order to facilitate partnerships. Startups should start with a limited scope where they can build a data advantage before attempting to compete with industry leaders. Ultimately, success in this space depends on rapid technological innovation and the ability to demonstrate scalable improvements in time/cost savings or diversity in drug pipelines.

2. Market Overview

The Artificial Intelligence in Drug Discovery Market was valued at USD 253.8 million in 2019 and is expected to reach USD 2.1 billion by the year 2027, at a CAGR of 31.9%. This rapid growth will be driven by cross-industry collaborations and partnerships, the increasing need to control drug discovery & development costs, the rising adoption of cloud-based applications & services, and the impending patent expiry of blockbuster drugs. The U.S. leads AI drug discovery both in investment (63.5%) and number of startups at (60%), followed next by the U.K.. China has recently become one of the fastest growing biopharma markets and has the most AI drug

discovery R&D research centers worldwide due to strong government commitment to the development of AI solutions.

Large cap biopharma companies are currently fighting slowing growth due to lack of innovation and drying up of drug discovery pipelines. The expected return on investment from drug development has declined steadily from 10.1 per cent in 2010 to 1.9 per cent in 2018[1]. As a result, industry leaders have increasingly resorted to acquisitions and partnerships to support R&D efforts, and built up considerable cash reserves to fuel such a strategy. Another key trend in the industry has been the growth of biologics due to lower attrition rates in clinical trials, strong safety profile, and defensibility from competing biosimilars. Small molecule therapeutics more easily lose market share to generics, as they are easier to manufacture and distribute. As blockbuster medicines come off patent, industry leaders have scrambled to diversify their drug pipelines and come up with new therapeutics.

2.1 Major Stakeholders

The major stakeholders to consider in this market include acquisition capable biopharma companies, startups and researchers with AI technology, and large tech companies. Current market dynamics in biopharma have encouraged partnerships or acquisitions of small and mid-size companies with promising platform technologies. AI drug discovery companies have been ideal partners as industry leaders look to bulk up data science capabilities and address increasingly inefficient drug discovery programs. For the smaller companies offering such technologies, partnerships with larger companies are often essential to long term survival (unless the company has a wetware component to deliver proprietary data at scale), due to access to larger and more comprehensive datasets, as well as the regulatory, marketing, R&D, and trial support skills that large companies have optimized for. Tech giants such as Google and Tencent have also started to enter the space using their AI expertise, but also as investors. Google DeepMind's AI technology has developed a deep learning algorithm that predicts the 3D structure of proteins from primary sequences more accurately than earlier techniques, and has engaged in partnership with Sanofi using these emerging technologies. Google and Tencent have both invested in and provided expertise to startups in the space, such as Atomwise and XtalPi.

On the regulatory side, we are still waiting from the FDA to provide guidelines for AI drug discovery programs. The purpose of such guidelines would be to support generalizability and to prevent potential bias from training datasets. Finally, medicines created by AI will also impact patients, not only by providing new cures for neglected diseases, but also by impacting drug prices via the cost savings of AI drug discovery. By speeding up the drug discovery process, patients will also get access to treatments faster, with less patients enrolled in clinical trials for ineffective compounds.

[1] Deloitte Research

2.2 Pitchbook Market Statistics

- Quick stats (All time)
    - No. Companies: 208
    - No. Deals: 735
    - No. Investors: 958
    - Largest deal: $400 M (Relay Therapeutics)
- Deal count (TTM): 103
- Funding per quarter
    - Last 3 years: $5.75 billion
    - Annual Total: $2.4 billion
- Most active VCs by deal count: JLabs, Y Combinator, GV, Felicis, Khosla, Alexandria Venture Investments

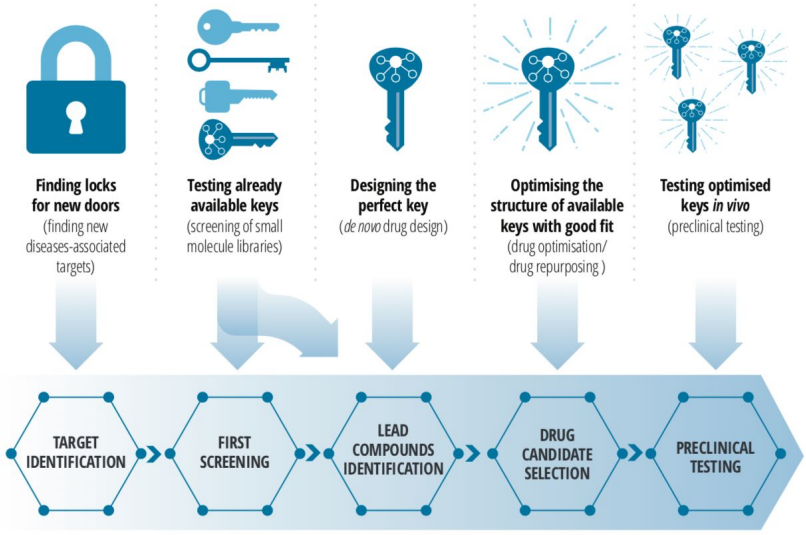
Stage	**Average Round Size**	**Average Post Valuation**
Angel	$0.6 M	$3.49 M
Seed	$3.21 M	$10.8 M
A	$13.41 M	$42.79 M
B	$36.25 M	$152.25 M
C	$55.25 M	$260 M
D	$44.25 M	$327.9 M

3. Technology Overview

3.1 The Drug Discovery Process

Drug discovery is the first part of the pharmaceutical value chain and involves identifying new candidate therapeutics for the treatment of disease. This process differs depending on the type of therapeutic (small molecule, biologics, etc). Small molecules still occupy roughly 90% of therapeutics on the market and current computational capabilities including those that power AI technologies have only been useful for such small molecules so the remainder of this section will describe the AI powered search for small molecules.

Once disease pathology and biology are understood, the discovery process typically starts with high throughput screening of existing compounds to identify leads. Downstream optimization of leads in order to improve certain characteristics such as binding affinity, toxicity, and other pharmacokinetic properties will occur thereafter. A summary is provided below:



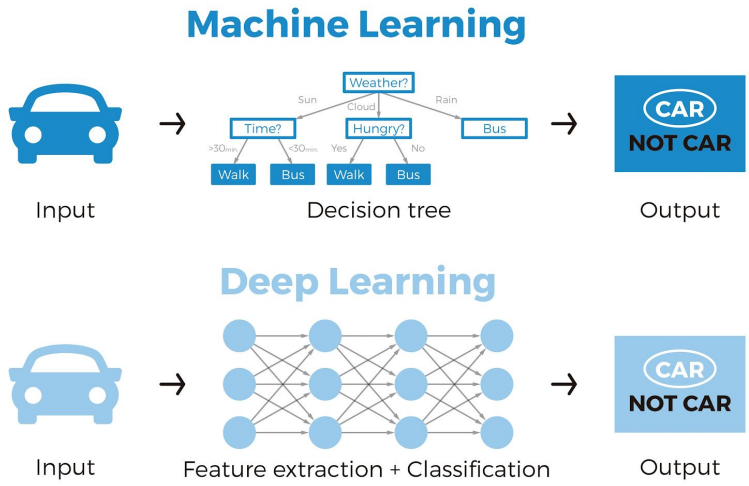
Summary of the five major stages of drug discovery.

This current approach is slow, has high attrition rates, and requires a robust physical infrastructure of testing equipment and compound libraries in order to show results. Out of the 10,000 molecules that are typically screened for a single indication, only 10 will make it to clinical trials. From there, the success rate for FDA approval is still only 10% once in trials. AI plugs into the drug discovery process at all five of the above stages in order to address speed, attrition, and infrastructure problems.

3.2 Defining AI in Drug Discovery

AI in this context, can mean several things. In general, they are a set of statistical learning algorithms that are capable of supervised or unsupervised 'learning' of a dataset to provide insights. Common subsets of AI used for drug discovery research include machine learning (ML), deep learning (DL), and natural language processing (NLP). Common tasks for ML architectures include classification and regression, where based upon past data, the algorithm can learn to make predictions (supervised learning). For example, an ML model can be trained to map a given set of characteristics about a molecule into a decision about whether it will perform well in a screening assay based on previous data. ML models can also be used for unsupervised learning, where just by being given a dataset, the algorithm can be used to cluster relevant traits or find commonalities between data points. This can be useful when given datasets representing healthy and diseased individuals, where unsupervised ML algorithms can identify disease specific traits.

In the context of drug discovery, these unsupervised learning algorithms are adept at analyzing multidimensional datasets and mapping associations between gene or protein networks for example. When datasets reach higher and higher degrees of complexity and dimensionality, DL models utilizing neural networks provide improved accuracy and speed for unsupervised learning tasks. DL algorithms are especially relevant for structural chemistry problems, given the high degree of complexity in molecule orientations. Finally, NLP algorithms are utilized to derive insights from text. The specific applications in drug discovery typically involve mining data from scientific abstracts to develop knowledge graphs and automating the generation of hypotheses for downstream use.



Simplification of ML vs DL

3.3 Applications of AI

In the target identification phase, AI is employed to analyze large datasets to search for novel exploratory hypotheses. At a high level, the idea is to offload some of the creativity of scientific inquiry to an AI so that we don't miss out ideas just because a scientist didn't think of it. NLP can be used to comb through scientific literature to generate new hypotheses for drug targets. For example, in 2017, IBM Watson identified novel RNA binding proteins for ALS. AI has also been used to understand biological systems through analysis of large multi-omics datasets. Machines are adept at clustering and elucidating biological pathways from multidimensional datasets that humans aren't able to see. Taking tissue samples from diseased and non-diseased patients and extracting genomic, proteomic, metabolomic, (etc) data, machines are able to identify disease associated traits and potential new drug targets. Roughly 50% of late stage clinical trials fail due to poor efficacy. ML techniques can be employed to generate higher quality drug targets to address this problem.

In the screening phase, deep learning techniques can better predict binding profiles and important other properties for successful clinical use including absorption, distribution, metabolism, excretion, and toxicity (termed ADMET properties). Vectorized representations of molecules in drug libraries can be mapped to 3D representations, which combined with a structural representation of a target, machines can be trained to discriminate how tightly the molecule will bind. Furthermore, chemical datasets describing results of various ADMET assays – ranging from physicochemical properties to animal-based PK properties, can be used to train ML models to predict these properties accurately for future unseen molecules. With this technology, thousands of molecules can undergo virtual screening, saving time and costs, and improving accuracy over legacy computational screening programs.

Novel platforms such as applying computer vision to in vitro drug screening have also shown promise (e.g. Recursion Pharmaceuticals). These platforms take pictures of cellular disease models over the course of a drug interaction to better predict pharmacological activity. Another ongoing area of development concerns new and improved structure featurization techniques for producing vectorized molecule representations. The outputs of AI models are dependent on features that can be incorporated into training vectors, and the accuracy of results can be improved by expanding the number of features used to describe molecules.

De novo drug design is perhaps one of the most promising areas that AI interfaces with drug discovery. AI systems will 'imagine' billions of possible atom configurations and via multiparameter optimization algorithms, select the configurations with the most pharmacological promise. This is significant because much of the current screening process involves repurposing of old drugs or known chemical scaffolds and modifying them for a new target. In theory, for these AI applications only the structure of the binding target along with discriminators for ADMET properties are necessary. This means that no knowledge of existing binders is required, removing bias and opening the possibility of truly novel drug structures.

Lead optimisation comes after a structurally promising compound has been identified. Repurposed drugs or newly designed drugs may demonstrate efficacy in an in-vitro or even in-vivo model, but may have adverse effects on other biologically important targets (so-called off-target effects), or they can be further optimized. Each structure might have millions of small modifications, such as a double bond instead of a single bond. AI algorithms can effectively screen these modifications in order to optimize formulations using previously described screening techniques.

Finally, AI algorithms can improve the accuracy of preclinical testing, shifting the risk earlier in the pipeline to avoid costs associated with clinical trials. Animal models in preclinical testing often are poor predictors of clinical performance. ML algorithms may be able to provide insights into cross-species differences in order to better inform decisions on escalating a drug to clinical trials.

3.4 Technology Summary

In summary, current generation AI algorithms have integrated into drug discovery platforms by:

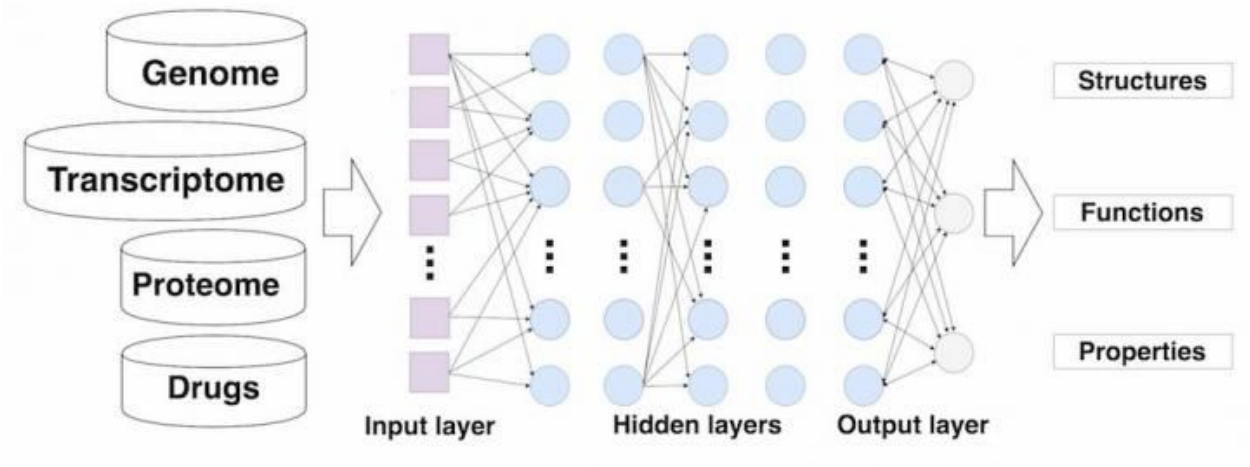
1. Target discovery
2. Predicting the drug-target interactions
3. Generating novel molecules
4. Predicting ADMET properties for translational researchers

Overall, the value of AI powered drug discovery can be broken down threefold:

1. Reducing timelines and costs by offloading physical screening to *in silico*
2. Increasing the accuracy of efficacy and safety predictions to frontload risk and reduce spending on more expensive downstream portions of the discovery pipeline
3. Diversifying drug pipelines with novel drug targets and structures for improved performance

The data utilized to generate these insights come from:

1. -omics data to characterize disease specific traits and identify biological targets
2. Small molecule libraries with structural information to train models
3. Previous screening data for both successful and unsuccessful compounds to train models
4. Abstracts and publications for NLP-powered learnings



-Omics Driven AI Approach

Summary of Major Techniques and Applications	
	Description
De novo Design	The *de novo* design of new therapeutics with novel structures is of great interest as companies look to diversify their drug pipelines. The current model relies heavily on screening old compounds and repurposing/modifying those with strong ADMET properties. AI models that can creatively design structures that fit desired properties can reduce reliance on time consuming and expensive screening processes.
Polypharmacology	The scalability of AI systems allows molecules to be screened on different cost functions easily. This means that it is relatively easy to screen molecules that can bind to multiple targets. Compounds that can bind to multiple targets are therapeutically promising, as many diseases are modulated by more than one target protein/receptor. Polypharmacology approaches have also been useful to optimize off target effects and ADMET properties.
Phenotypic Drug Discovery (-Omics Analysis)	Phenotypic drug design strategies do not rely on knowledge of the identity of a specific drug target or a hypothesis about its role in disease. With high dimensional -omics data, cells can be phenotyped so that assays are biology based and more informative of cellular response. This is contrast to traditional drug screening, where the purpose of a screen is to measure how tightly the drug binds to a purified target. AI algorithms plug in nicely to phenotypic drug discovery because they are able to analyze this highly dimensional data and provide insights.
Single Target Druggability	The simplest way that AI models have interfaced with drug discovery has been predicting single target druggability. Deep learning algorithms trained from structural binding data can more accurately predict binding to a target than previous generation methods.
Target ID & Hypothesis Generation	Natural language processing (NLP) methods have been employed to allow AI systems to read and mine through scientific abstracts and papers to extract insights. By creating a knowledge graph of all biological insights, drug targets and mechanisms in theory can be elucidated.
ADMET Prediction	Accurately predicting ADMET properties early in the discovery pipeline offers significant value to large biopharmas. Detecting compounds that will likely fail in more expensive phases of the drug development process saves companies from needing to test as many compounds. A variety of AI methods can be used to predict ADMET properties based on the structure of the compound. Essentially all companies have some form of ADMET prediction, companies will need to prove that their algorithms are more effective than previous generations.

3.5 Major Players

All major large cap biopharma companies have committed to the development of AI solutions as a part of their drug discovery process and have disclosed deals and partnerships with academic centers of startups providing such technology. While no company has produced yet a molecule or target that has reached clinical significance, a few startups have stood out as technological leaders:

Atomwise

Atomwise is a series A stage company that has raised $51.3 million from investors including Tencent, Baidu, Khosla, Y Combinator, and Monsanto Growth Ventures. Their CEO is Abraham Heifets. Their proprietary deep learning powered screening technology termed AtomNet screens more than 100 million compounds per day, improving the speed at which hits are generated and the quality of drug hits. Atomwise has assisted in the invention of new potential medicines for 27 disease targets via partnerships with over 60 academic institutions and companies.

Exscientia

Exscientia is a C stage company based in the U.K. that has raised $103 million from major life sciences companies including Novo Holdings, Bristol Myers Squibb, Evotec, GT Healthcare, and Celgene. Their platform relies on high dimensional phenotypic data, and as a result, they are highly active in terms of partnerships to obtain necessary clinical data. Partnerships include Roche, Celgene, Bayer, Sanofi, GSK, Sunovion, and lastly Sumitomo Danippon Pharma, who Exscientia designed the first AI conceived molecule with to reach Phase I clinical trials. The claim is that Exscientia's process cuts the time spent in discovery from 4.5 years to as little as one year, reduces discovery costs by 80 percent and results in one-fifth the number of synthesized compounds as is normally needed to produce a single winning drug.

BenevolentAI

BenevolentAI has raised $292 million from private investors since their founding in 2013. They use text mining to analyse the available patents and other genetics and biological information to create highly informative knowledge graphs. These graphs have been used to infer novel biology and drug targets. Benevolent has also developed benchmarking models for evaluating de novo molecular design, and has partnered with major pharmaceutical companies including Novartis, Janssen, and AstraZeneca.

Recursion Pharmaceuticals

Recursion is a C stage startup led by CEO Chris Gibson that has raised $226 million from investors including CRV, Obvious Ventures, TwoSigma, Felicis Ventures, and Menlo Ventures.

Their platform applies computer vision technology to analyze the impact of drug compounds on cellular disease models to better predict in-vivo and clinical activity. Recursion's technology is disease agnostic and currently has three Phase I compounds and one Phase II compound.

4. Historical Context, Key Trends, & Future Development

4.1 History of Drug Discovery

Small molecule therapeutics make up roughly 90% of drugs available on today's market. However, since 2014, biologics have accounted for 93% of spending growth in discovery platforms. Biologics have advantages over small molecule therapeutics such as higher specificity and better safety profiles, lower attrition in clinical trials, and less competition due to difficulty developing biosimilars. However, small molecules have simpler drug discovery and development processes and superior delivery and distribution characteristics. There is an industry wide shift towards biologics, however small molecules still dominate current markets.

In modern computer-aided small molecular drug discovery and development, ML methods, especially traditional learning methods, were widely used for building predictive models such as quantitative structure-activity relationship (QSAR) models and quantitative structure-property relationship (QSPR) models. These ML models provided a rough estimate of binding affinity and in some cases would predict ADMET properties by considering general drug design principles like Lipinski's Rule of Five. However, they were limited primarily by compute power, restricting their predictive power to low dimensional data inputs. In practice, they were used more to rule out obviously wrong compounds, than to precisely pick the few that might be successful.

4.2 Present Status

In recent years, new DL techniques have been adopted in drug discovery and development, opening a new door to computational decision making in pharmaceutical science. The success of DL techniques benefits from the rapid development of the DL algorithms, the progress in high-performance computing techniques, as well as the explosion of chemical information in chemical databases. This has allowed analysis of high dimensional data, allowing much greater precision than possible with earlier approaches.

Currently, AI has minimal usage in industry. Biopharma has historically been a tightly competitive industry with strong value placed upon IP assets. Only recently have there been efforts to share data in an attempt to encourage collaboration and progress in data intensive learning fields like AI (e.g. Alliance for Artificial Intelligence in Healthcare). The current technological status of AI is not yet mature enough to make a serious impact on drug discovery

systems. The technical proof of concepts have been established and initial partnerships with large biopharmas has started, but the general industry consensus is that it may take another 2-3 years for a successful compound to arise and progress through clinical trials.

Although AI has been able to design promising compounds that show results in vitro, few have made it out of animal testing and gone on to clinical trials. We still have no confirmation that these new molecules can be *better* than previous generations, especially because the models used to create them were trained on previous generation molecules. Technical challenges include making sure compounds are drug-like and synthetically feasible. AI systems often create molecules with dangerous reactive groups, or molecules that are very expensive or time consuming to synthesize, so a team of medicinal chemists is necessary to help guide the training of the models to discourage such features.

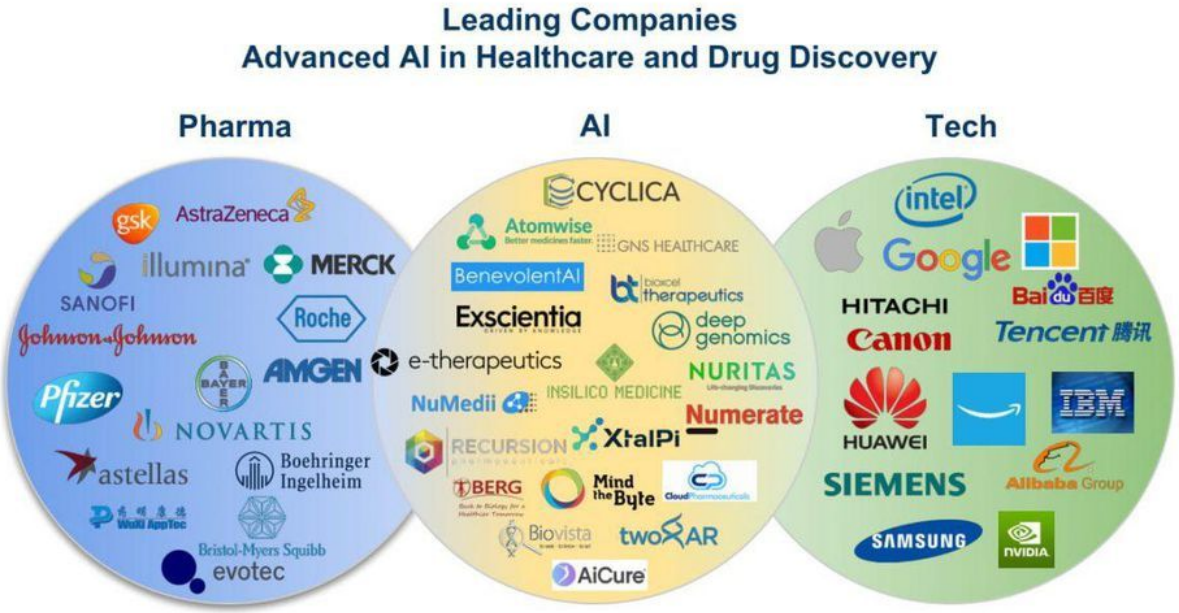
4.3 Future Predictions

In the near term, the performance of AI systems will increasingly be validated with influxes of more and more patient and wet lab validation data. Within a few years, we will see the results of the first AI designed molecules in human clinical trials, which if successful, would greatly catalyze growth in the industry. Successful *de novo* drug design from AI systems is still a question mark at this point, and will continue to remain primarily a research interest in the near future. On the target ID side, the rapid growth in wet lab data collection and availability has already fueled its use to identify several drug targets across diseases. Again, these have not yet been clinically validated, but the use of such technology is more widespread across both industry leaders and startups. Finally, on the screening side, AI systems have already demonstrated value as a method of reducing costs and time spend. The next few years will focus on refining screening algorithms and further integration in large pharma pipelines.

Thinking more long term, eventually AI systems will be sophisticated enough and will have enough training data to play a meaningful role at the very least in the drug screening process. AI systems will still work together with medicinal chemists, but much tighter integration and automation will inevitably occur. The timing from screening to preclinical testing will be reduced to a few months rather than five/six years, and new potential drug candidates will be identified at increasingly lower costs. New drugs capable of treating very precise pathologies will become the norm, facilitating the transition towards precision medicine. With increasing compute power, AI systems may be able to become more complex, and have impact past small molecules (e.g. vaccines, RNA therapies, proteins, and other biologics). These new approaches will be catalyzed by more efficient computing platforms such as quantum computing. In the end, multiple AI systems will be used in any one drug discovery platform. The best AI is not necessarily a single AI that can autonomously design a new drug, but one or many different AIs, that enable better understanding and the design of new inputs, throughout the drug discovery process from target selection, hit identification, lead optimization to preclinical studies and clinical trials.

5. Opportunities

There are over 200 drug discovery AI companies operating globally[2]. As this is a strongly hyped field, sifting through the noise and identifying those with legitimate 0-to-1 technological improvements and experienced teams will be the biggest challenge. This is especially difficult due to the fast paced and industry forward nature of the field. Verifying technological progress via scientific literature is difficult because it lags years behind cutting edge industry technology. This facilitates a competitive and secretive culture of the industry which disincentivizes companies from publishing their own innovation, further exacerbating the problem.



5.1 Startups to Watch

The following pre-series B startups have demonstrated significant promise as up and coming purveyors of novel technology:

TwoXar: TwoXAR is a series A stage AI pharmaceutical company that has raised $14.3 million from a16z and Softbank. The founding team is strong and has already entered partnerships with 6 large biopharma companies.

[2] BenchSci, List of AI drug discovery companies segmented by company type.

Anagenex: Anagenex has raised $3.5 million in a seed round from Khosla, Obvious, Menlo, Lux. The startup is applying machine learning towards DNA encoded libraries, an emerging data generation platform that has drastically enlarged the universe of testable drug compounds.

InVivo AI: Utilizing few shot learning algorithms to predict outcomes when data is sparse. Seed stage, has raised $1.8 million to date. Spun out of Mila - Quebec's AI Institute.

OneThree Biotech: OneThree Biotech integrates diverse data types to create algorithms to predict ADMET and binding. The company raised a seed round in January 2020 from Primary Venture Partners and Meridian Street Capital, bringing their total to $2.8 million raised to date.

Rahko AI - Rahko AI is a seed stage company based in London and founded in 2018 that has raised $1.45 million in venture funding. Rahko is a quantum machine learning company that announced a 3 year research collaboration with Merck in May 2020 and is an early partner of Amazon Quantum Solutions Lab.

LabGenius

LabGenius is a series A stage company that has raised $13.7 million and is backed by Lux Capital, Obvious Ventures, Felicis Ventures, among other top VC firms. They are the first biopharmaceutical company developing next generation protein therapeutics using a machine learning-driven evolution engine (EVA™). They use advanced deep-learning neural networks to explore protein fitness landscapes and improve multiple drug properties simultaneously. Protein therapeutics are a very promising and diverse therapeutic space, and if their platform shows promise, there would be a wide variety of applications.

Genesis Therapeutics

Genesis is a seed stage startup ($4M) funded notably by Andreesen Horowitz and Felicis Ventures. Genesis is developing tools to improve screening by better predicting pharmacological properties of drug compounds. Their AI platform, PotentialNet, predicts 20+ different ADMET properties and was invented by their CEO Evan Feinberg, while doing his Ph.D. in Vijay Pande's lab at Stanford. ADMET screening is a major challenge in the drug discovery pipeline and can vastly reduce downstream costs if done correctly. The team at Genesis has a strong background and with continued technological validation, will address a considerable market need.

Cyclica

Cyclica is a series A stage company that has raised $12.5 million to screen proteomes and utilize AI to predict off target effects for promising drug compounds. Cyclica is the first company to focus its efforts on polypharmacology - or multitargeted drug compounds. These approaches can provide information on unwanted drug targets, or lead prioritization for other diseases. This is promising because research has increasingly demonstrated that drug targets are one of many factors that contribute to pharmacological activity, and understanding system wide effects of a molecule has the potential to vastly improve our pre-clinical knowledge of efficacy. Cyclica has partnerships with Merck, Bayer, Eurofarma, and WuXi AppTec. Their CEO is Naheed Kurji.

Terray Therapeutics
Terray is a drug discovery company whose investors include Digitalis, KdT, and Two Sigma. Terray's proprietary screening platform generates more quantitative data across a larger chemical search space than any existing technology, providing the foundation for the company's computationally driven approach to their internal preclinical pipeline and partnered programs.

5.2 Emerging Technologies

Two technologies have arisen with the potential to vastly improve the drug discovery landscape: **implementation of quantum chemistry methods**, and **generative adversarial networks for *de novo* design**.

Quantum computing is a research stage technology that relies on qubits, trapped ions, and photons for computation, which allows further granularity past 'bits' that traditional computers use. These systems in theory have the potential to handle highly complex and specific modeling problems with exceptional efficiency and accuracy. A model of computation with built-in quantum mechanics will be best suited to simulate quantum systems, such as those used in quantum chemistry. The precision of quantum chemistry models is a vast improvement over the 'stick and ball' atom models that are currently used to model molecules. In fact, some of the efficacy of DL systems in modeling structural binding can be attributed to them 'learning' quantum chemistry reactions. Quantum computing will provide much more accurate drug-target binding models.

Quantum computers are also adept at multiparameter optimization, the key task that is needed when screening molecules. This means that molecules can be screened faster, and larger molecules with many more atoms can also be screened. This will be especially useful for screening and collecting data on non-small molecule therapeutics such as antibodies and other protein therapeutics. Machine learning can benefit from the use of quantum computing by getting more accurate and precise training and validation sets. Additionally, quantum computers will expand the feature set used to make predictions, by providing AI models quantum states that are otherwise impossible to describe with traditional computing approaches.

Generative adversarial networks (GANs) are another emerging technology poised to make an impact on *de novo* drug design. Initially described in 2014, GANs have been used widely in the software industry for applications such as DeepFakes, voice translation, and even art and music. GANs involve two deep learning networks, a generator which 'imagines' possible outputs, and a discriminator which decides whether the generated output satisfies certain criteria. In the context of drug discovery, GANs are promising because it offloads creative discovery work to machines without needing to screen billions of compounds.

With this technology, not only can speed be greatly improved, the molecular diversity of outputs is also expanded. Our current compound libraries are limited by imagination and synthetic feasibility and are 10^{50} times smaller than the universe of compounds with potential drug-like activity. GANs have the potential to bypass traditional human heuristics of what makes a good compound and devise compounds that humans have never thought of before. The landmark paper describing the use of GANs for drug discovery was published in September of 2019 by Zhavoronkov and colleagues at Insilico Medicine. A pharmacologically novel compound was discovered by their GAN based model in just 46 days. While this technology is still early and requires further validation, the time and cost savings of this approach are tremendous.

5.3 Emerging Applications

As the field experiences more and more technological progress, areas for application will expand to **two key areas: AI for protein therapeutics, and AI for rare/neglected disease states**.

AI for biologics and non-small molecule drug discovery is the next frontier for systematic drug discovery programs to tackle. Antibodies and other protein therapeutics have become a promising therapeutic strategy with a diverse range of indications. However, they are orders of magnitude larger than small molecule therapeutics, which makes them difficult to model and study systematically. Instead, biologics are mainly hypothesis driven, meaning that new therapeutics do not come from screening, but rather scientific insights.

As the technology of quantum computing accelerates, it is possible that a systematized and automated method of discovering biologics can be created, powered by an AI system. The vast complexity of proteins will require a computing platform capable of harnessing multidimensional data, and deep learning does just that. Currently, companies like LabGenius and BigHat Bio are tackling this problem, using deep learning to predict protein structure from an amino acid sequence. If this is possible, screening across libraries of protein sequences for specific targets can occur, similarly to how screening is done currently for vectorized representations of small molecules.

The other emerging application will be applying such technologies to rare or neglected disease states. This is made possible by changing the economics of drug discovery, because the development costs of AI derived molecules are much lower. Technologically, this is also powered by algorithms solving a ‘small data problem’. Novel targets for rare diseases or precision care rarely have enough clinical or screening data for any type of AI model to develop the predictive capabilities to screen molecules. To solve this problem, companies such as BenevolentAI and Exscientia have developed algorithms to mine through scientific literature for insights. In this

way, industry scientists can loosen their dependence on academics to fully describe mechanisms before approaching a drug design problem.

5.4 Industry Challenges

Artificial intelligence is a notoriously overhyped field, and as a result, industry specific challenges are often glanced over. This section identifies four major hurdles that startups face as they progress through the industry.

1. Compared to the amount of training data available for other AI applications, chemoinformatics **databases for DL modeling are far behind**. The major publicly available database, ChEMBL, has just under 2 million compounds, but activity data for building specific models are still limited. This is because there are also millions of targets, and when designing molecules for specific targets, datasets are often small and sparse. Algorithmically, a challenge has been how to learn quickly from very little data. Startups will need to acquire access to new and substantial datasets in order to develop technology that is relevant and not quickly outpaced.

2. Specific to deep learning algorithms, there has been a black box problem, where useful features or interpretations of those **features are difficult to extract from the models**. This is a problem native to deep learning, as once the training inputs go in, the only interpretable results come from the outputs, there is not a great way to determine how the AI is arriving at a certain result. This is a problem for drug discovery because scientists would like to know how and why a molecule was designed a certain way in order to better understand the underlying biology, or which specific or off target binding sites the drug may modulate. Regulatory and safety issues may come up if scientists are not able to describe how and why a drug compound is working.

3. For startups especially, **talent is in short supply**. To succeed in this field, a multidisciplinary skill set including backgrounds in medicinal chemistry/biophysics and specialization in machine learning is required. Computer-generated compounds can be riddled with components that are difficult to make, or contain dangerous reactive groups, and a chemistry team is needed to mitigate this. Computational chemists are already in short supply, and for those who also specialize in ML, the talent pipeline is even tighter as biopharma must compete with the tech industry for talent. Large biopharmas are ramping up hiring of data scientists as well, making a magnetic founding team essential for the success of an early stage startup.

4. Academically, this is a burgeoning field, and still **very early in terms of research output**. The field also suffers from overhype and extreme commercial interest, which in some senses has blunted academic progress. Industry leaders will poach scientists from academic labs, and as a

result, top publications in this field have come from private companies which often do not have the same interests in data sharing and collaboration. The overhype problem also has the potential to hinder technical progress when out of reach expectations are unmet.

Also important to recognize are non-challenges for the industry. These include regulatory, fundraising, and defensibility of IP. The regulatory environment for AI powered discovery is currently lax, as the FDA has not provided any guidance on emerging technologies. The expectation is that the FDA will eventually provide guidance, but the purpose would be to support generalizability and to prevent potential bias from training datasets. Thus from a regulatory point of view, most companies have not needed to significantly alter operations. As noted previously, AI drug discovery is one of the fastest growing fields in terms of fundraising (31.9% CAGR). Investors come in multiple forms including large biopharma companies and tech investors, in addition to the expected biotech focused investors. The funding and refinancing environment has thus far been favorable. Finally, the defensibility of IP and approach are addressed through innovation speed. Publications in this industry dominated field lag heavily behind the cutting edge. By the time a publication comes out, the publisher may very well be 3 years ahead of the research by some estimates. Thus, IP is 'defended' because leading companies are able to out innovate themselves in a short amount of time.

6. Conclusions

Overall, AI powered drug discovery is a rapidly evolving field that in theory provides significant value to the biopharmaceutical industry in the form of decreased costs and discovery timelines, as well as pipeline diversification, facilitating the development of novel drug compounds. The challenge with the space will continue to be the pace of technological innovation. Unlike other disease domains in biology, R&D in AI drug discovery is pushed forward primarily by industry innovation, meaning that technology has become the primary competitive differentiator, instead of perhaps GTM or business strategy. Large biopharma companies are investing heavily in both deepening and diversifying their data science capabilities and frequently collaborate with startups with emerging technologies, which is a mutually beneficial relationship, as large amounts of data are necessary for accurate models.

A winning team in this space will have a magnetic founding team that can attract talent and make partnerships. Their competitive advantage will be rapid execution of research goals applied towards indications or datasets where not at a data disadvantage. Below are the key takeaway points to understand while making investment decisions in this space.

6.1 Vertical Strengths

- Transformational technology with a very clear value proposition if done correctly
- Favorable acquisition and collaboration environment with large biopharmas
- Widespread availability of funding for refinancing or expansion
- Immature field with many areas for innovation and ample space to avoid competitive overlap

6.2 Vertical Weaknesses

- The field is very data driven, and without access to new high quality datasets, companies take on significant technical risk
- AI powered drug discovery is still a very young field. Molecules generated from AI driven insights still haven't demonstrated clinical efficacy.
- Talent is in short supply and will be expensive to acquire due to competition from large biopharmas and tech companies
- Advances in the field are held tight to the chest of innovating companies. Teams must have strong internal R&D capabilities and cannot rely on academia to support them
- Overhype can hurt a team by providing unrealistic expectations for growth. Teams should be confident in their technology but not promise too much

6.3 Opportunity Cost of Capital

AI drug discovery is a new and rapidly growing field powered by decades of foundational research in machine learning and enabled by the rapid scaleup of computing resources and availability of data. The biopharma industry and current market dynamics have facilitated a unique market structure that encourages collaboration and a hot acquisition market. This derisks a company's go to market strategy (allowing product focus), and provides plentiful opportunities for exits. Finally, AI drug discovery is by nature a platform approach, allowing the continued generation of assets from a single technology. Due to the diversity of biopharma and the countless range of indications and druggable targets, the size of the market will continue to rapidly expand in the coming years, making competition less of an immediate concern if companies specialize.

6.4 Investment Thesis Areas

This report finds AI drug discovery as a rich vertical with significant opportunity for startups to thrive. Specifically, companies that are able to address the following needs will have ample ground to build category defining leadership.

1. **Biologic design**: The biopharma industry has dramatically upscaled biologic development in the past decade and by all indications will continue this trend due to the defensibility and better targeting and safety profiles of biologic medicine. Applications of

AI towards protein and antibody therapeutics have already started development, but the market here is large and has not produced an industry leader.

2. **De novo design**: Creative molecule design is the next frontier to be crossed by machine technologies. Prediction of efficacy and safety will continue to improve with increasing data availability and for this reason will remain dominated by industry leaders such as Atomwise. Creative design technologies are the ultimate end goals, and teams with such technology should be looked at very carefully.

3. **New Biological Paradigms**: The single target action model for drug discovery is overly simplistic and limited as an approach for addressing disease pathology. Network driven models that are described by multiple proteins, complexes, and interactions are more accurate depictions of disease, but have not been explored due to various complexities including the ability to analyze data and understand how multiple compounds may best interact to address the disease. AI has begun to address this problem, and companies that can demonstrate the capability for AI to accurately deconstruct disease biology in a multifaceted network driven manner and to validate these insights in living models will be of high value.

Overall, this report concludes that investment into the AI drug discovery space is encouraged given that three key considerations are met:

1. The founding team has the relevant experience and capability to recruit scientists, interface with industry leaders, and rapidly produce research results.
2. The team has a plan to continuously access new and high quality data.
3. There is a competitive advantage already understood and patented by the founding team that thematically powers future technology development

8. References & Further Reading

1. Scientific American Descriptive Overview
2. Emerging Tech and Applications
3. Research review from March 2019
4. Deep Learning Drug Discovery Review March 2018
5. Nature Review (Exscientia)
6. Editorial Overview
7. Deloitte Whitepaper