AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-scale Examination of Bio-pharma Firms

Bowen Lou
School of Business
University of Connecticut
bowen.lou@uconn.edu

Lynn Wu
The Wharton School
University of Pennsylvania
wulynn@wharton.upenn.edu

Abstract

Advances in artificial intelligence (AI) could potentially reduce the complexities and costs in drug discovery. Using a resource-based view, we conceptualize an AI innovation capability that gauges a firm's ability to develop, manage and utilize AI resources for innovation. Using patents and job postings to measure AI innovation capability, we find that it can affect a firm's discovery of new drug-target pairs for preclinical studies. The effect is particularly pronounced for developing new drugs whose mechanism of impact on a disease is known and for drugs at the medium level of chemical novelty. However, AI is less helpful in developing drugs when there is no existing therapy. AI is also less helpful for drugs that are either entirely novel or those that are incremental "follow-on" drugs. Examining AI skills, a key component of AI innovation capability, we find that the main effect of AI innovation capability comes from employees possessing the combination of AI skills and domain expertise in drug discovery as opposed to employees possessing AI skills only. Having the combination is key because developing and improving AI tools is an iterative process requiring synthesizing inputs from both AI and domain experts. Taken together, our study sheds light on both the advantages and the limitations of using AI in drug discovery and how to effectively manage AI resources for drug development.

Keywords: Artificial intelligence, drug discovery, IT Innovation, biotech & pharmaceutical industries, AI capability

Introduction

The drug discovery process is extremely complex (Dougherty and Dunne 2012). It requires navigating a combinatorial space of more than 10^{60} molecules to find a suitable drug candidate (Agrawal et al. 2019; Mak and Pichika 2019). This vast search space is simply too large for human beings to process effectively using technologies without an artificial intelligence (AI) component. As a result, drug candidates discovered using conventional methods are often clustered in small areas of the innovation space and many of them provide only slight improvements over existing drugs (Trafton 2020). Accordingly, it is hard to find new drugs, especially those that differ substantially from existing drugs and that provide big improvements (Krieger et al. 2018; Rotman 2019; Scannell et al. 2012).

This problem also presents opportunities to accelerate compound discovery and the overall drug development process if the search through the combinatoric space could be expedited. Advances in AI, especially with the recent advances in digitization and machine learning, could help address the intractable search problem of discovering new drug candidates (Agrawal et al. 2019). Because AI excels in automating predictions and identifying hidden patterns in data, it facilitates recombination in innovation (Wu et al. 2020), accelerating the discovery of novel chemical compounds under certain conditions. For instance, AI can screen compounds 100 times faster than humans can using conventional methods (Smith 2020), and AI algorithms can constantly improve their prediction accuracy through feedback data. Using AI to aid drug development, medical and machine learning specialists succeeded in finding a novel antibiotic, Halicin and other drug candidates, out of more than 100 million molecules, in a fraction of the time that traditional methods require (Stokes et al. 2020). Not only can Halicin kill many species of antibiotic-resistant bacteria in animal studies, it is also structurally distinct from prior antibiotics (Marchant 2020). This discovery is groundbreaking because antibiotic-resistant

"superbugs" are a major public health issue that traditional methods have largely failed to address. In addition to antibiotics, AI has also accelerated the search for coronavirus vaccines. A collaborative effort between AI and medical experts created an AI system that can identify optimal mRNA sequences in just 16 minutes (Liang 2020; Zhang et al. 2020). In addition to speed, mRNA sequences found through AI have a more robust secondary structure that can produce a stable and efficacious mRNA vaccine against COVID-19.[1]

As AI technologies are increasingly adopted in drug discovery and anecdotal evidence about AI's effect on drug discovery continues to grow, it is important to examine the effect of AI on drug discovery systematically. As with all technologies, AI has both advantages and disadvantages when applied to improve busines and innovation processes (Wu et al. 2020). Firms lacking an understanding of the benefits and limitations of AI could misallocate valuable AI resources to the types of projects where AI provides minimal benefits. Failed outcomes from these investments could then create disincentives to invest in AI even in areas where AI can clearly offer benefits. This can in turn hurt firms' long-term competitiveness (Aral and Weill 2007).

In this study, we examine what AI can and cannot do for drug innovation, how firms can develop and manage AI, and how these AI practices differ from practices using earlier generations of information technologies (IT). Drawing from a resource-based and IT capability framework, we conceptualize an AI innovation capability (AIIC) – the firm's ability to develop, manage and utilize AI resources for scientific discovery and research and development (R&D). We measure AIIC using patents and job postings. They collectively capture three types of AI resources that are key for creating AIIC in firms: 1) tangible AI assets such as data and

[1] https://syncedreview.com/2020/05/12/new-baidu-algorithms-boost-mrna-vaccine-development-for-sars-cov-2/

infrastructure; 2) AI skills that are critical to create, implement and deploy AI tools for scientific discovery; 3) AI-enabled intangibles such as firm practices and knowledge assets that complement the use of AI. We then apply the AIIC across a wide range of bio-pharma firms to examine its effect on drug innovation. In assessing a firm's AI skills, a key part of the AIIC, we quantify the total AI skills in a firm and also distinguish employees that individually possess a combination of AI skills and domain knowledge from those with just AI skills. Possessing this combination is key to AI-based innovation because effective use of AI for drug discovery is an iterative process that requires a continuous synthesis of knowledge from both AI and medical experts. We then estimate how the AIIC can impact the quantity and chemical novelty of drugs that are developed under preclinical studies.

Our findings show that firms with higher AIIC can better support the compound discovery for preclinical studies than can other firms. This is consistent with AI's abilities to accelerate search and discovery in a previously computational infeasible space. We also find that the effective management of AI for drug development involves not just hiring employees with AI skills but hiring those with a combination of AI and medical knowledge. We show that there are both advantages and limitations of using AI for developing drugs. As of this writing, the current state of AI is not mature enough to produce drugs with a full spectrum of novelty for preclinical studies. Instead, our findings suggest that AI works best for drugs aimed at an intermediate level of novelty that is neither too novel nor too incremental and for drugs whose mechanism for attacking the disease is known.

Theory and Hypotheses

Our investigation covers four areas. First, we develop the concept of AIIC from the resource-based view of the firm. We then examine AI and compound discovery for preclinical

studies. Next, we examine how AI can aid drug discovery by investigating cases when the mechanisms of impact for treating the disease are known. Lastly, we probe into the effect of AI on the novelty of drugs.

AI Innovation Capability and the Resource-based View of the Firm

Firms have developed IT capabilities to gain competitive advantage (Bharadwaj 2000; Santhanam and Hartono 2003). Bharadwaj (2000) defines IT capability as "the ability to mobilize and deploy IT-based resources in combination or co-present with other resources and capabilities." The combination of several IT-related sources such as IT infrastructure, practices, and employee skills helps firms create IT capability (Bharadwaj 2000; Ravichandran et al. 2017; Tambe and Hitt 2012). Such combinations are often valuable, rare, difficult to imitate, and non-substitutable, allowing firms to differentiate themselves, improve R&D productivity, and outperform competitors (Bardhan et al. 2013; Joshi et al. 2010; Kleis et al. 2012).

Drawing from the resource-based and IT capability view, we argue that firms can also benefit from creating AIIC. Specifically, we define a firm's AIIC as the ability to develop, use, and manage AI resources in combination or co-present with other resources and capabilities to effectively conduct scientific discovery and R&D. Grant (1991) suggests that firms can leverage three types of resources to create competitive advantage: tangible resources, personnel-based resources, and other intangible resources. We similarly examine AIIC using the three types. First, tangible AI resources may include data, infrastructure and algorithms that are customized for specific scientific discovery (Varian 2018). They form the foundation upon which firms can develop AIIC.

Second, having AI skills, as embodied in the employees, is critical to create, implement and deploy AI tools for scientific discovery (Babina et al. 2020). Creating an AI tool for drug

discovery requires continuous syntheses of inputs from both AI and medical scientists during both the development and the operational stages of the tool. Accordingly, AI skills should not be constrained by hiring employees with AI expertise only. They also need to have some working pharmacology knowledge so they can speak the same language as the medical scientists to effectively communicate and work with them. Having a multi-disciplinary perspective helps AI and medical scientists to select and solve a problem that AI is suited to address. However, this does not necessarily mean that firms should hire employees who are experts in both AI and drug discovery, a combination which may be difficult to find. Rather, employees can be expert in one domain and possess some working knowledge of the other. For example, an AI expert does not need to have an advanced degree in pharmacology or medicine but does need some domain knowledge, so she can communicate and collaborate effectively with domain experts in drug discovery. Thus, it is critical to distinguish the need for personnel who have a combination of AI and domain expertise from those with either only AI skills or only domain expertise. For instance, the discovery of Halicin was the result of a successful multi-disciplinary collaboration between medical and computer scientists, many of whom have a main expertise in one of the two areas (either machine learning or medical sciences) and some working knowledge in the other. This collaboration has created a machine learning platform that can broaden the search space for biodiversity and led to the discovery of a new type of antibiotics (Stokes et al. 2020). The platform is also constantly improving as operational data are fed back to improve algorithmic performance, creating a virtuous cycle that can dramatically accelerate the drug innovation process. Having domain expertise is key to distinguishing spurious correlations from causations in the output of AI, so the right feedback and training data can be selected to train AI algorithms.

The combination of AI and domain expertise contrasts to the use of prior generations of IT that often centralize IT employees because any expertise they might possess in other areas is generally not put to use beyond the tool development stage. Additional domain skills, such as understanding business processes, are useful primarily at the design stage of an IT tool but not in the operational stage (Hammer 1990). AI, on the other hand, requires a constant and iterative approach, which relies heavily on using operational data to improve algorithmic performance, an approach that traditional IT implementations do not use. Thus, even after an AI tool is designed and deployed, AI and domain experts still need to put their cross-disciplinary skills to use, so that they can continuously work together to improve algorithmic performance and to keep up with AI's rapid technical advances. Accordingly, having the combination of AI and domain expertise within individual employees is critical to creating AIIC because it is an indicator of how well a firm can utilize both AI and domain expertise to innovate.

Lastly, AI-enabled intangibles, including firm practices and knowledge assets that can empower and foster the use of AI, are key drivers for improving firm performance and competitiveness (Barney 1991; Teece 1998). Firms that recognize what AI can offer to their businesses are more likely to invest in AI technologies, invent new AI methods, and develop practices that complement AI. Overall, we expect firms that have acquired tangible and intangible AI assets as well as employees with AI skills, especially those with a combination of AI and domain knowledge, can have more capabilities to innovate using AI.

AI and Drug Discovery

Developing drugs is perhaps one of the most expensive and riskiest processes in the world, costing over $2 billion for a typical drug (DiMasi et al. 2016). About 90% of potential drugs fail to attain FDA approval (Smietana et al. 2016). The early stage of developing drugs

primarily comprises the discovery and preclinical trials that often involve animal testing where drug candidates are proposed to address certain biological targets that cause a disease. Once a drug candidate-target pair is found and verified during the preclinical trial phase, the drug enters the later clinical trials stage. If the drug succeeds in these trials, the last stage involves the FDA deciding whether to grant final approval.

We focus on AI's effect on the early stage of drug discovery because the early stage is particularly crucial to the entire innovation process in that discovering more drug candidates initially will likely lead to more clinical trials and approvals. Furthermore, because an average drug can take more than 10 years to develop and because many advances in machine learning are too recent, it may be too early to capture AI's effect on later stages. Thus, we mainly focus on the effect of AI on the early stage of drug innovation. But the role of AI in the late-stage innovations might one day be examined in a similar way.

The early-stage drug discovery process has been inherently slow because it involves searching for chemical compounds in a large complex space spanning multiple scientific disciplines. The breadth of the search area ranges from genetics to protein synthesis, from biological and chemical synthetic processes to drug mechanisms (Dougherty and Dunne 2012; Vamathevan et al. 2019). The drug discovery process requires an understanding of the human biological system consisting of 25,000 genes and millions of proteins, all of which can create complex interactions with each other (Pisano 2006). The difficulty in managing this complexity is a key reason for the high failure rate in developing drugs (Dougherty and Dunne 2012).

Using IT to find new chemical compounds and to advance drug discovery started decades ago. IT is an underpinning in the fields of cheminformatics and bioinformatics studies. Yet past generations of IT have had only limited success in discovering new drugs (Brown 1998; Drews

2000). However, AI, with recent advances in digitization and machine learning, can fundamentally accelerate the discovery of new drug candidates for two reasons. First, the digitization of scientific knowledge has enlarged the digital search space and AI is especially suited for taking advantage of the enlarged data domain to identify new drug candidates (Jayaraj and Gittelman 2018). Instead of being hampered by their complexity, more data and more fine-grained data can improve the accuracy of AI algorithms. Second, human experts using traditional methods tend to find drugs within a narrow spectrum of novelty (Trafton 2020). AI, on the other hand, can overcome the limitations of human searches to explore a much larger innovation space that can greatly facilitate recombination in innovation (Wu et al. 2020). By automatically collecting, analyzing, and detecting complex patterns in the existing data, AI algorithms can search through the combinatoric space to identify new compounds with desired pharmacological effects. Compared to other technologies, AI can do this without having explicit instructions. Through many examples of input-output pairs, supervised learning (an area of machine learning) can accurately uncover linkages and make better and faster predictions than humans can in many areas (He et al. 2015; Hu et al. 2018). This can be particularly helpful in situations where the bottleneck of scientific discovery lies in the inability to navigate large complex data to make predictions, the type of problem faced in compound discovery (Hughes et al. 2011). By providing a ranked list of potentially promising drug candidates for human researchers to investigate, AI can accelerate the early stage of drug development.

Hypothesis 1: AI innovation capability has a positive effect on drug-target identification at the early stage of the drug development process.

An important part of AIIC is the ability of AI and medical scientists to collaborate. Their collaboration was key to finding Halicin and other novel chemical compounds for preclinical

studies. A successful collaboration requires that AI scientists have domain knowledge in drug discovery and medical scientists to have some AI skills. Medical scientists do not need to be AI experts and vice versa. AI scientists with some domain knowledge can effectively work with medical scientists in choosing the appropriate training and feedback data for AI to use. Having some AI skills can help medical scientists understand the output of AI algorithms, and make the appropriate judgements and actions. They are more likely to know when to trust algorithmic outputs, detect false positives, and understand how to communicate their concerns to further improve AI tools. Ultimately, the success of developing, managing and using AI for scientific discovery requires that AI scientists and domain experts collaborate (Hitt et al. 2018); having employees who individually possess both AI and domain knowledge could facilitate the collaboration.

Hypothesis 2: Firms with employees who individually possess both AI and domain knowledge are more likely to discover drug candidates than firms with employees who possess only AI knowledge but lack domain knowledge.

AI and Discovering Drugs with Known Mechanisms of Impact

Having a large amount of training data is necessary for machine learning to make useful predictions. Diseases with known treatments are more likely to fulfill the data requirement because existing treatments and mechanisms are usually documented in the literature. Furthermore, when the mechanism of impact or existing treatments are known, experts can more easily distinguish likely drug candidates from spurious ones. The disambiguation process is especially important because AI is an effective prediction machine for finding correlations, but it cannot directly provide causal inference. When the mechanism of impact for treating a disease is known, it is easier for scientists to check and verify whether the drug candidates have the desired

mechanisms to attack the disease ("Mechanism matters" 2010), and complement AI by choosing the best drug candidates for clinical trials. Although data analytics can automate the disambiguation process to some extent if the mechanisms are digitized, the most valuable knowledge still lies in forms of human intuition growing out of experiences that are difficult to digitize (Wu et al. 2020). It is thus important to have domain experts ensure that the drug candidates possess the desired properties. However, when the mechanism of impact is not known, human researchers are less able to distinguish true drug candidates from false ones without conducting large and expensive clinical trials. Having too many false positives recommended by AI could impede scientific progress if these false positives are selected for clinical trials. Thus, we expect AI to have a stronger effect for finding drug candidates with known mechanisms of impact than for finding drugs whose mechanisms of impact are unknown.

Hypothesis 3: AI innovation capability helps in discovering drug candidates with known mechanisms of impact better than it helps in discovering those whose mechanisms are unknown.

AI and Drug Novelty

While AI can accelerate the discovery of drug candidates for preclinical studies, it is unclear whether the drug candidates it discovers represent incremental or large improvements over existing drugs. Research has shown that discovering drug candidates aimed at novel therapies that represent big leaps is much harder than discovering therapies offering incremental improvements ("follow-on" drugs) (Rajkumar 2020). However, on average, once they gain FDA approval, the return on these novel drugs is substantially higher than that on "follow-on" drugs (Krieger et al. 2018). Although it is difficult to assess a drug's therapeutic impact at the early stage, the chemical novelty of a drug is often a proxy, providing an ex-ante measure of drug effectiveness (Krieger et al. 2018).

As discussed above, AI can help in identifying new molecules with desired pharmacological effects when there are abundant data available for AI to search through for finding hidden patterns, especially when an existing treatment or drug mechanism is known. For example, in 46 days, Insilico Medicine uncovered and experimentally tested six new compounds that can inhibit discoidin domain receptor 1 (DDR1), a tyrosine kinase target implicated in fibrosis and other diseases, because those well-known DDR1 and common kinase inhibitors are already well-documented (Zhavoronkov et al. 2019). Similarly, the discovery of Halicin is possible with AI because the mechanisms of how antibiotics work are known. However, novel drugs that differ radically from existing treatments have almost no precedents, and machine learning is ill-suited to support discovering them (Wu et al. 2020). Inferences based on limited data may depend heavily on tacit knowledge that is inherently costly to collect and transfer, and therefore can be difficult to digitize for AI consumption (Nonaka and Von Krogh 2009; Von Hippel 1994). Developing sufficiently novel drugs also requires deeper understanding of a narrow domain with tacit knowledge to which AI can add limited benefit at best. For example, the discovery of artemisinin for treating malaria was fundamentally driven by limited data and human ingenuity. The only reference to the drug treatment appeared in one sentence in an ancient book written in the 3rd century that was not directly related to malaria. Dr. Youyou Tu, the inventor of artemisinin, combined her clinical experience with the ancient text to create a paradigm shift in antimalarial drug development (Tu 2011). Current AI technology is not capable of effectively understanding the meanings in the ancient texts needed to make the necessary link to treating the disease (Marcus and Davis 2019). Even if it could, a single data point, such as the case in this instance, would hardly be sufficient for machine learning to make useful inferences.

Turning to the other end of the innovation spectrum, we expect AI to provide similarly limited value in facilitating the development of incremental drugs because many firms already have capabilities for discovering such drugs (Krieger et al. 2018) that don't require advanced technology to succeed. Scientists are at least as effective at finding incremental drugs as AI is because the search space for incremental drugs is relatively small; it usually entails searching the space around existing drug therapies. AI, on the other hand, can provide the most power when searching through a large and disconnected space. Furthermore, deploying AI can be expensive given the substantial upfront investments and strategic planning required for the necessary digital transformation (Bughin et al. 2017; WIPO 2019). There are also recurring costs of employing AI specialists to curate data and train AI algorithms. Using AI to discover incremental drugs would thus provide marginal benefits at best. Thus, we expect AI is most effective at developing drugs that are of intermediate novelty—those that are neither too radically different nor too incremental. Intermediate-novelty drug candidates stand to benefit more than their counterparts from broad searches and finding patterns in diverse data that AI can support.

Hypothesis 4: The effect of AI innovation capability on drug development is more positive for drugs that have an intermediate degree of novelty than for drugs that have an either lower or higher degree of novelty.

Data and Measurement

First, we discuss how we examine a firm's drug portfolio. Next, we discuss how we measure drug novelty and drug mechanisms. Lastly, we show how we operationalize the firm's AIIC.

Drug Portfolio

We focus on the global biotechnology and pharmaceutical industry, and collect drug development data from two leading sources: the Informa Pharmaprojects database and the investigational drug database from Clarivate Analytics (Hess and Rothaermel 2011; Kapoor and Klueter 2015; Krieger et al. 2018). The drug dataset spans 25 years (1995-2019), has a comprehensive coverage of drug candidates, and has information about their detailed development stages. We focus primarily on the compound discovery and preclinical research stage where drug candidates are proposed to address certain biological targets that cause a disease. In addition, our data include the originators and licensees and all other firms involved in the development process. We account for the transfer of drug patents and their rights using data from the Securities Data Company (SDC) to ensure that drugs in our drug databases are correctly matched to the firms responsible for their original development (Eklund 2018). Thus, we can observe a firm's drug portfolio, pipelines and the originators for each drug.

Existing Mechanisms of Impact

We create a binary variable measuring whether a mechanism of pharmacological impact for treating a disease condition is known. Our drug development database ties drugs to the disease conditions they treat over time. For detecting available mechanisms of impact, we label a drug as having no known mechanisms when it has "Unidentified pharmacological activity" or "Not applicable" indications under drug mechanisms. We also use the free text in the DrugBank database that contains comprehensive information about how each drug works (Wishart et al. 2018; Wishart et al. 2008) to ascertain the drug mechanism.[2]

Drug Novelty

[2] We can search for keywords "unclear," or "unknown" to infer whether the drug mechanism is known. For example, Modafinil (https://www.drugbank.ca/drugs/DB00745) shows "The exact mechanism of action is unclear, although in vitro studies have shown it to inhibit the reuptake of dopamine ..." We can thus infer the drug mechanism for Modafinil to be unknown.

We focus on small-molecule drugs because they constitute the majority of drugs in development (Krieger et al. 2018). In total, we measure the chemical novelty of 13,699 drugs based on their known chemical structures.[3] Using methods suggested in recent research literature on chemical informatics (Backman et al. 2011; Cao et al. 2008), we measure novelty by assessing the deviation of the chemical structure of a drug candidate from the structures of all prior drugs. The detail of this calculation is shown in Appendix 1. Each drug is assigned a novelty score between 0 and 1. A higher novelty score indicates a more novel drug.

Patent Stock

We use global patents from the worldwide patent statistical database PATSTAT[4] that offers bibliographical data for over 100 million patents from 90 global patent-issuing authorities. Each patent record contains a detailed patent application, citations, a title, an abstract, and legal persons (e.g. firms or any organizations) filing the patent application. It identifies whether the patent owners are business enterprises, education institutions, governmental agencies or individuals (Du Plessis et al. 2009). It also develops a comprehensive approach to standardize the original names of patentees (Magerman et al. 2006). We match the names of the firms from our drug database to patent assignees in PATSTAT.[5] We also adjust the assignee names to represent

[3] Our databases provide detailed historical development records of over 60,000 drugs. But the chemical structure information for most compounds that never progress beyond the very early discovery stage is not available. The same is true for large-molecule drugs (known as biologics). Our empirical analyses mainly focus on the small-molecule drugs with known chemical structures.

[4] We first harvest the PATSTAT data in the version 2017b to cover patent records till 2017. Since Google launched its public datasets of worldwide patents on BigQuery in 2017 (https://cloud.google.com/blog/products/gcp/google-patents-public-datasets-connecting-public-paid-and-private-patent-data), we further augment our data to cover the time period from 2017 to 2019 by retrieving records of global patents through Google Patents Public Datasets. Similar approaches are employed to match them to other datasets used in our analysis.

[5] We primarily use the PATSTAT standardized name (PSN_NAME) in the company sector (PSN_SECTOR is referred to as COMPANY) to determine the assignees of patents for the firms in our sample for this matching process. We also test our matching by using other harmonized names available, such as DOCDB standardized name, and the OECD HAN name as recorded in PATSTAT. As the accurate sector assignment is provided for the PATSTAT standardized name (PSN_SECTOR for PSN_NAME), we choose to use PSN_NAME for our matching procedure.

the original company that filed the patent after accounting for their merger and acquisition (M&A) history using Thomson Reuters SDC and the Zephyr databases from Bureau Van Dijk. Based on these matched firms, we then retrieve their patent application documents from PATSTAT and extract filing years, titles, abstracts, and citations for these patents. Following the convention in the R&D literature (Griliches et al. 1986; Hall et al. 2001), we use the patent filing year (as opposed to the publication year) because it more closely approximates the date when the firm produced and used the innovation. Thus, we can measure a firm's general investment in patent inventions using the accumulated stock of patent applications by the firm with an annual depreciation rate of 15% (Hall 1990; Hall et al. 2005).[6]

AIIC

Based on the resource-based and IT capability view, we operationalize AIIC by using texts in patents and job postings to capture a wide range of AI-related concepts, definitions, applications and fundamentals. Patents are primarily used to capture tangible and intangible assets; job postings are primarily used to capture employee resources. Thus, the use of patents and job postings can collectively capture tangible, intangible and employee resources, all three types of resources related to AIIC. We also distinguish machine learning from other types of AI investments (Cockburn et al. 2018; WIPO 2019). Below we explain how we find AI-related patents and skills.

AI-related Patents

We use patents to capture the ability of a firm to either innovate AI technology itself directly or use AI to innovate generally (Webb 2019). In the pharmaceutical and biotechnology

[6] We apply a standard perpetual inventory equation with declining balance depreciation to measure patent stock (Hall 1990): $P_t = (1-\delta)P_{t-1} + R_t$, where P_t is the end-of-period patent stock and R_t is the contemporaneous patent inventions during the year t. We use the conventional 15% per year for the depreciation rate δ.

industry, patents have been ubiquitously used to gauge the ability of a firm to innovate. A single patent in the drug space can exert unique influence not only on a product, but also on the technology used for drug discovery itself (Markman et al. 2004). New drugs are typically patented (Abrams and Sampat 2017; Hemphill and Sampat 2011), because patents are the most effective way to prevent imitations and substitutions (Cohen et al. 2000; Levin et al. 1987). Patents can discourage reverse engineering which is much easier than discovering a new drug, thereby preventing theft of intellectual properties (Gilchrist 2016). Accordingly, patents can serve as a key indicator of pharmaceutical innovation.

Patents can thus represent both the tangible and intangible resources, the first and the third characteristics respectively for creating AIIC. Patents are tangible innovation outcomes because they are highly portable and transferable (Markman et al. 2004). Patents can also approximate for intangibles and knowledge assets that are critical for future innovation and productivity (Tambe et al. 2019) because they are the culmination of complementary processes, skills, and firm practices. Firms with more AI-related intangible assets are more innovative and more productive than their counterparts (Brynjolfsson et al. 2018).

To find patents related to AI, we use both AI patent classes and the free text in each patent. AI patent classes can directly measure whether a patent contributes to the core AI technologies. The United States Patent and Trademark Office (USPTO) designated a specific patent class for AI technologies: Class 706 for "Data Processing – Artificial Intelligence." This class has a large set of subclasses including "neural networks" and "machine learning." To find patents that employ AI to solve problems in other domains without directly contributing to the core advances in AI technology itself, we look for AI-related words from a glossary of validated words and phrases in patent titles and abstracts. This glossary includes words related to the three

interrelated technological subfields within AI: robotics, symbolic systems, and learning (Cockburn et al. 2018). We also follow a widely accepted computing classification system from the Association of Computing Machinery Computing Classification System that accounts for the dynamic change of AI technologies (WIPO 2019). Because this method has been used for over 50 years to organize the classification of concepts and trends of technologies, it can significantly mitigate the subjective classification of AI.[7] Furthermore, we also include phrases related to a variety of AI technologies from outside vendors because off-the-shelf AI technologies (e.g., PyTorch and TensorFlow) can also be used for scientific discoveries (Raymond et al. 2019). Lastly, we test several variants of these keywords in our dictionary; they do not qualitatively change the classification. A list of AI-related keywords used to identify AI-related patents is in Appendix 2.

In total, we find 7,433 AI-related patents developed by the biotechnology and pharmaceutical firms from our dataset spanning 1995 to 2019. Similar to the way we measure general patent stock, we track the development of AI over time and use the accumulated AI-related patent stock with an annual depreciation rate of 15% (Hall 1990; Hall et al. 2005). While our analysis primarily focuses on the years from 2010 to 2019, patents from prior years are used to construct a patent stock for each firm. The stock-based measure also aligns with the spirit of the standard innovation production function that models new knowledge as a function of existing knowledge stock combined with resources devoted to produce the new knowledge (Jones 2005; Romer 1990).

[7] There are three major hierarchies to develop AI-related phrases for classification: (i) the "AI" hierarchy, comprised of AI functional applications such as natural language processing, computer vision, knowledge representation and reasoning, simulation of human cognitive tasks, and AI techniques used to realize those functions; (ii) a "machine learning" hierarchy that unveils numerous learning-based AI techniques; and (iii) a "life and medical sciences" hierarchy under the "applied computing" category that covers activities pertinent to intelligent computing for producing medicines.

We plot the growth in AI-related patents in Figure 1 and distinguish the three types of AI technologies: expert systems, machine learning, and other AI applications. Overall, we see a tremendous growth in patents related to AI, with the biggest growth in machine learning.

AI-related Job postings

In addition to the tangible and intangible resources represented by patents, we also collect job postings that can be used to create constructs related to personnel skills, the second type of the resources in creating AIIC. Job postings related to AI can be used to gauge the latent demand for AI skills in firms (Alekseeva et al. 2019), which is critical to using AI tools to innovate. AI skills can also capture intangible human capital that the firms need to foster innovation in the long run, as well as approximate for innovation outputs that are not patentable. As such, AI skills can also measure certain tangible and intangible assets that are not captured in patents. Together, AI-related patents and skills are thus complementary representations of a firm's overall AIIC.

We measure the total AI skills in firms from job postings, similar to Babina et al. (2020).[8] Our job posting data come from a leading analytics company collecting job posts from over 40,000 online job boards and company websites from 2010 to 2019. We examine both the skill requirements and job titles listed in each of the postings. To measure AI skills in job postings, we search for a similar set of AI-related words that are used in classifying AI-related patents. These job postings have a time stamp and the name of the hiring firms, allowing us to create a metric about AI skills for each firm in each year. We also use the job title classification from O*NET to identify AI-related positions, similar to how IT and analytics labor are distinguished from other employees in earlier work (Tambe and Hitt 2012; Wu et al. 2019). If any of the skills listed under the job title is related to AI, we treat the posting with that job title as requiring AI skills. We

[8] Babina et al. (2020) also document consistent patterns across various measures of AI using job postings.

aggregate these individual-level skills for each firm-year observation, assuming that firm- and occupation-specific factors with respect to the likelihood of posting a job are uncorrelated.[9] AI skills at a firm are thus estimated by the number of AI-related job postings. We then measure AIIC in a firm using the standardized sum of the standardized values of the variables about AI-related patents (AI patents) and skills (AI skills) as shown in equation 1. The standardization procedure (norm) first subtracts the mean from the variable and then divides the resulting difference by the standard deviation.

$$\text{AIIC} = \text{norm}(\text{norm}(\text{AI patents}) + \text{norm}(\text{AI skills})) \quad (1)$$

We also separately measure the number of job postings that require both AI skills and domain knowledge in medicine because each job posting has detailed skill descriptions for the job. In particular, the job postings indicate specific skillsets related to domain expertise for drug innovation, so we also search for a set of domain knowledge-related words such as pharmaceutical industry knowledge, drug development, molecular biology, medical and clinical research for further classification. Thus, we can segregate AI skills into (1) those requiring a hybrid of AI skills and domain knowledge, and (2) those requiring only AI skills without domain knowledge. We estimate their effects on drug innovation separately.

IT Innovation Capability

Similar to how we measure AIIC, we use IT patents, and job postings that require IT skills to measure IT innovation capability (ITIC), which is the innovation capability that firms develop from their information technology investments. IT patents are identified using Category 2 in patent classification that includes computer hardware and software, communications, computer

[9] To the extent such matchings vary systematically across firms, the problem can be alleviated by firm-fixed effects.

peripherals, and information storage (Hall et al. 2001).[10] We identify IT skills using the skill requirements in the job postings as well as the job titles. For example, IT skills listed in a job posting can include software development as well as hardware support. IT-related job titles can include software engineer or systems analyst. If the job posting also contains keywords such as computer, website, software, and telecommunication, we identify it as requiring IT skills. Similar to our construction of AI skills, we aggregate the talents with IT skills in each firm. The ITIC of each firm can thus be calculated as the standardized sum of the standardized values of patents and skills measures related to IT.

Control Variables

We primarily rely on Crunchbase, PitchBook and Bureau van Dijk Orbis databases to incorporate firm characteristics into our empirical models. These three databases provide rich information about both public and private firms in the biotechnology and pharmaceutical industry. Because entrepreneurial exits have been shown to affect organizational innovation outcomes (Bernstein 2015), we control for a firm's financial ownership status over years (a dummy variable indicating whether it is publicly held). We also control for firm age, number of employees, and R&D spending. The founding year of each firm is collected and verified using BioCentury, Moody's, Renaissance Capital and Thomson Financial Securities. Workforce headcount and R&D expenses come from Compustat Global, BioCentury, and Bureau van Dijk Orbis.

[10] We also use multiple alternative methods from Forman et al. (2016), such as incorporating electronics-related patents about electrical and semiconductor devices identified from Category 4 in Hall et al. (2001) and searching IT-related phrases on the titles and abstracts of patents. These approaches yield directionally consistent results in our estimation on the effect of IT innovation capability.

Table 1 shows the summary statistics and the correlations of all the variables for those firms with AIIC. We observe that a firm's AI-related patents and skills, the two data sources that we use to measure AIIC, are not highly correlated with each other.

Empirical Strategy and Identification

The data used for our primary analysis cover the years between 2010 and 2019, which capture the period of rapid advances in AI technologies (Fleming 2018). In total, we have 2,043 global bio-pharma firms, of which 644 firms have AIIC. We primarily focus on the 644 firms in our analysis; results using the full sample are shown in Appendix 3.

We use firm-level analyses to examine AI's effect on the quantity of drug candidates discovered for preclinical studies using equation 2. Due to the risky nature of drug discovery, the number of drug candidates is highly skewed, with many firms having no drug candidates at a particular stage in a typical year. This number is further skewed if we only use drugs with sufficient chemical novelty. Thus, we take the logarithm of one plus the raw number of drug candidates in our main analysis. We also include firm-fixed effects γ_i to control for any unobserved time-invariant differences in firm characteristics, and year-fixed effects y_t to account for temporal shocks. We control for a firm's financial ownership structure to account for different innovation priorities between public and private firms. We also control for firm size (total employees), firm age, patent stock, and R&D expenditure. Our main coefficient of interest, β_1, captures the marginal effect of the AIIC on drugs developed at the preclinical trials stage. We explore AI's effect on drug novelty measured by the number of drugs in three categories on the spectrum of chemical novelty that correspond to incremental, intermediate and highly novel drugs, and estimate equation 2 for each novelty range.

$$\ln(Number\ of\ Drugs)_{it} = \beta_0 + \beta_1\, AIIC_{it} + Controls_{it} + y_t + \gamma_i + \epsilon_{it}. \quad (2)$$

We use instrumental variables to address a potential upward bias of AI if high performing firms with slack resources choose to invest in AI. The instrumental variables are derived from a yearly patent-citation network for each firm in our sample that has AIIC. In the network, each node is a firm, and each link is the aggregate patent citations. For example, if patents in firm A cited patents from firm B 5 times in the current year, a directed link between A and B would have a weight of 5. In this example, the relationship is not reciprocal: B's patents don't cite A's patents. Thus, B is A's neighbor because A has drawn knowledge from B, but A is not B's neighbor since B did not cite A, and thus there is no observed information flow from A to B. We use the total number of neighboring firms with AIIC to instrument for a focal firm's AIIC. We also use two variations of these instrumental variables: (1) the average number of AI patents in the neighboring firms; (2) the average ratio of a firm's AI patents to total patents for these neighbors. Similar to the network-based approaches to construct instrumental variables in Wu et al. (2017), we focus on knowledge flows across firms that serve as a proxy for the cost of accessing AI-related knowledge and reflect the ease of accessing external AIIC from neighboring firms (see Figure 2 for one example). We exclude direct competitors, defined as network neighbors that are in the same industry as the focal firm. Thus, the network neighbors used to create the instruments are firms not in the bio-pharma sectors. The citation neighbors also vary substantially in their industries and geographical locations and are thus less likely to be affected by common industry or region-specific shocks and competitive pressure. The associated F-statistic in the first stage is 15.4, passing the threshold for the weak instrument test.

Findings and Discussion

AI and Drug Discovery

We first explore the relationship between a firm's AIIC and the number of new drug candidates a firm discovered for preclinical studies in a year (Table 2). After applying firm and year fixed effects and controlling for a firm's cumulated patent stock, financial ownership status, age, total number of employees and R&D expenses, we find that AIIC is positively associated with the number of new drugs developed at the preclinical trial stage (Column 1). This effect is separate from IT's effect (Column 2). Specifically, a one-standard-deviation increase in a firm's AIIC is associated with a 3% increase in new drug candidates. Given that it is extremely difficult to identify suitable chemical compounds with pharmacological effects for the preclinical trial stage, a 3% increase can be substantial, especially if it leads to a major discovery such as Halicin. We also find more than 40% of the effect of AI comes from machine learning (Column 3), suggesting that the key advance in AI facilitating drug development comes from the ability to navigate a large search space as enabled by machine learning. The 2SLS estimations on AI's effect show consistent results (Table 5). All estimations are conducted with standard errors clustered at the firm level. Overall, they support Hypothesis 1.

We also estimate the effect of AI patents and AI skills separately (Table 3).[11] Because the patent data are also available prior to 2010, we compare AI patents' effect before and after 2010. We find that, likely because of advances, recent AI technologies exhibit a greater effect on drug discovery than AI did in earlier years (Column 1 and Column 2).

We also categorize the overall employees with AI skills into (1) those possessing both AI skills and domain knowledge in drug development, and (2) those possessing AI skills only. We find that the effect of AI skills comes primarily from those with a combination of AI skills and domain knowledge (Column 5). The effect from employees with only AI skills is small and only

[11] Job positions take time to fill, so we use lagged 1-year and 2-year of AI skills, and the results are similar.

statistically significantly at the p<0.1 level. This suggests that acquiring AI skills alone is not sufficient; having employees who can straddle between AI and pharmacology is key to facilitating the drug development process. These results support Hypothesis 2.

AI, Mechanisms of Impact, and Drug Discovery

If the mechanism of impact for existing drug treatments is known, AI should be more effective at developing new drugs to treat similar conditions. In our firm-fixed effects analysis, we find that AIIC is positively associated with the number of drug candidates at the preclinical trials stage when the mechanisms of impact are known (Column 2 in Table 4), and the size of the effect is also greater than that for drugs whose mechanisms are unknown (Column 3 in Table 4). The 2SLS estimations show similar results (Columns 2 and 3 in Table 5), suggesting that AI is more capable of exploiting existing drug treatments to find new drug candidates that treat a similar condition than it is in finding treatments where no prior therapeutic drugs have identified mechanisms of impact. These findings support Hypothesis 3.

AI and Drug Novelty

We also examine how a firm's AIIC can affect the discovery of chemically novel drugs. Table 4 shows that the effect of AIIC is small and statistically insignificant for both incremental drugs (chemical novelty between 0 and 0.3) and drugs at the more novel end of the spectrum (chemical novelty between 0.7 and 1). However, within the middle range of chemical novelty, from 0.3 to 0.7, the estimate of the effect of AIIC is positive and much greater than the estimates for very incremental and highly novel drugs (Columns 1-3, the differences are significant at p<0.01). On average a one-standard-deviation increase in a firm's AIIC is associated with a 2.8% increase in the number of drugs in this intermediate novelty range. The 2SLS estimations yield directionally similar results (Columns 4-6 in Table 5). These results suggest that a firm's AIIC

can help primarily in discovering medium-novel drugs rather than in discovering either completely novel ones or incremental derivatives of prior drugs. Separately estimating AI patents and AI skills yields similar results (Appendix 5). Overall, these results support Hypothesis 4.

We also find that relative to the AIIC, a firm's ITIC has no significant effect on the process of drug discovery for preclinical studies (Table 2).[12] While IT is still important to support drug discovery, they do not provide a competitive edge in drug development, possibly because most firms have already invested in IT, and earlier best IT practices may have already diffused throughout the industry. Our results show that AIIC is primarily responsible for improving the early stage of drug innovation process before clinical trials occur. Results using the full sample of firms are also consistent (Appendix 3). Although there are many firms with no AIIC in the full sample, we continue to find that AI can affect the early stage of the drug innovation process. The effect is particularly salient for those drug candidates when the mechanisms of impact are known and for those drugs at the intermediate level of novelty. In the full sample analysis, we also address selection biases of firms choosing to invest more in AI, using the Propensity Score Matching and Coarsened Exact Matching (Blackwell et al. 2009; Ho et al. 2007; Rosenbaum and Rubin 1983). These results are consistent with our main results. We also used different functional forms such as Poisson regressions (Appendix 4). They yield directionally similar results.

Viewing these results together could also help address certain selection biases caused by the fact that innovative firms are also more likely to invest in AI. It is difficult to envision a scenario where a firm chooses to use AI only to find drugs of intermediate novelty and only

[12] Removing AI from general IT investment can cause downward bias of the IT estimates. Our alternative robustness tests use AIIC as an instrument for ITIC to isolate the part of IT that is associated with AI. We find a positive effect from the AI component of IT, but no effect from the remaining parts of IT.

when prior therapies already exist. Presumably innovative firms are likely to try and develop drugs across all degrees of novelty. The more plausible explanation is that AI can be particularly helpful in this scenario precisely because the innovation capabilities it can provide in aiding drug discovery are best suited to making discoveries within this set of conditions.

Limitations and Future Research

To the best of our knowledge, our study is the first to systematically measure AIIC and examine its effect on compound discovery in drug development. However, there are a few limitations that future research should address. First, we primarily examine the effect of AIIC on compound discovery for preclinical trials, the early stage in the drug innovation process. Future research should examine its effect on later stages, especially the clinical trial phases. The use of AI in these later stages may differ from that in the early stage as problems faced in clinical trials tend to differ from those in compound discovery. Currently, it may not be feasible to detect the effect of AI's applications in the early stage on a drug's final approval from the FDA; machine learning is just beginning to have an effect on compound discovery and it could take more than a decade for a drug to get through clinical trials. However, the same method used in this study to examine AIIC's impact on the early stage can be used to study the later stages as well. Also, in our study, we primarily focus on small-molecule drugs whose chemical structures are readily available; it would be important to extend the study to include large-molecule drugs which have become increasingly important (Krieger et al. 2018).

Second, it is important to improve the measurement of the actual use of AI. A task-level analysis is key to getting at how AI use affects the management of specific business processes. Large-scale firm-level studies may not be well-suited to capture the specific uses of AI on a particular task which could differ across firms or between divisions within firms (Burton-Jones

and Straub Jr 2006). A task-based examination coupled with measuring specific uses of AI can help advance the understanding of how different complementary practices are needed to leverage AI for specific tasks.

Third, we note that the job posting data are not the same as the employment data (Tambe and Hitt 2012) and our study suffers from the limitation that the data do not fully represent the human capital in a firm. Posted job openings may not be filled, and therefore may not be representative of the actual human capital in a firm. While we can make assumptions about the fill rate from job postings to approximate the actual human capital at a firm, detailed employment data are preferred because they provide a direct measurement of human capital. However, it is encouraging that a recent study by Babina et al. (2020) finds that AI skills measured using job postings are highly correlated with those using resume data that have detailed employment history, suggesting that job posting data are a suitable tool to approximate AI skills in firms. However, better measurements of AI skills and management practices are needed to advance our understanding of how to effectively manage AI.

Lastly, our paper focuses on the effect of AIIC on drug innovation. Our findings may extend beyond drug discovery to have broader effects on general scientific discovery and R&D outcomes. Future work should consider the broader implications of AIIC on all innovations.

Implications and Concluding Remarks

In this study, we use a resource-based view of firms to develop a bio-pharma firm's AIIC and examine its effect on the early stage of drug development. Our study makes two important contributions. First, we provide a theoretical extension on IT capabilities to understand what capabilities AI can provide for innovation. To the best of our knowledge, our study is the first to systematically examine the link between AIIC and drug development. We show both the

advantages and disadvantages of AIIC in compound discovery for developing new drugs. Specifically, we find that AIIC can help develop new drugs at the intermediate level of novelty and new drugs whose mechanism of impact for treating a condition is known.

Second, we create a multi-dimensional yardstick for measuring AIIC. While IT capability can be operationalized in many distinct ways depending on the context (Chan and Levallet 2013), we are the first to use patents and job postings on a large scale to create a metric that can gauge a firm's ability to innovate by developing, using and managing AI resources. We also count the number of employees that individually possess AI skills and domain knowledge in drug development as an important component of AIIC. This multi-dimensional measure contrasts with efforts in a burgeoning area of research on AI productivity that primarily tabulates either number of employees or dollar investments to measure AI investment or use (Brynjolfsson et al. 2018; Dixon et al. 2021).

Our findings suggest several important managerial implications. First, it is crucial for firms to recognize the nature of the collaboration between AI and medical experts as ongoing rather than one-off. Having individuals with multi-disciplinary knowledge is key to facilitating the ongoing collaboration required for AI-assisted drug innovation. Developing and using new AI tools for drug discovery is an iterative process that requires inputs from both AI and medical experts. Although the IT alignment literature also shows that the effective use of IT requires the combination of IT- and business-related knowledge (Bassellier and Benbasat 2004; Kearns and Sabherwal 2006), it is limited to the upfront development of the tool (Hammer 1990). By contrast, the effective use of AI requires that the collaboration of AI and domain experts continues beyond the tool development stage because wringing the best performance out of AI algorithms requires continuously training with new operational data. Thus, the management of

AIIC requires individuals with a combination of AI and domain expertise who can persistently develop and use AI tools operationally. This does not necessarily mean that firms should hire employees who are experts both in AI technologies and in drug discovery. Instead, it requires hiring employees who have at least the working knowledge of both.

A second managerial implication is that AI resources should be managed and used in areas where tasks are heavily dependent on automatic data processing and reasoning, and involve the navigation of a large search space. However, special attention should be paid to separating spurious correlations from causation; experience, intuition, and human expertise are all necessary for the effective use of AI, and they complement AI investments to facilitate innovation.

Third, while AI can expand and navigate the search space for innovation, it is not applicable to creating innovation across the entire spectrum of novelty. AI is effective for discovering new drug candidates with known mechanisms of impact because it is easier to distinguish between spurious correlations and causations in drug candidates. Similarly, we find that AI provides the biggest benefit for discovering drug candidates that possess intermediate novelty because AI is best at navigating a large search space to combine existing technologies in a new way; as such these innovations tend to be of intermediate novelty. In keeping with this point, AI is of limited use in developing either very novel or very incremental drug candidates, possibly because the problems AI has at the extremes of the novelty spectrum are different both from the moderate level of novelty and from each other. At the novel end, AI has a capability problem: it is incapable of discovering truly new treatments for which little or no data are available to detect patterns. By contrast, the problem at the incremental end is less due to capability but to price: it is not worth the effort and cost of deploying AI to find drugs that constitute marginal improvements over existing drugs when scientists are already proficient at

identifying incremental drugs. As AI tools become cheaper to adopt and use, it may become cost effective to use AI for finding incremental drugs in the future. For certain treatments that may require drugs that represent a radical break from the past, the current state of AI is ill-suited for their discovery. Thus, firms should avoid indiscriminately applying AI to accelerate drug discovery without regard to the degree of novelty of the drug they seek.

This study shows the nuances of managing and applying AI for discovery of new drugs for further development. Contrary to popular believes that firms can invest in AI skills simply by hiring AI engineers, we show at least in terms of drug discovery, employees who can straddle between AI and pharmacology are required. Similarly, AI cannot create a competitive advantage in all arenas. Rather, AI can be very useful for discovering drugs whose mechanisms of impact are known. AI is also useful for discovering intermediately novel drugs; AI does not provide sufficient capabilities and is too expensive to discover incremental drugs, and AI is of virtually no use for radically novel drugs. Taken together, these findings represent an argument for the thoughtful development, management and application of AI.

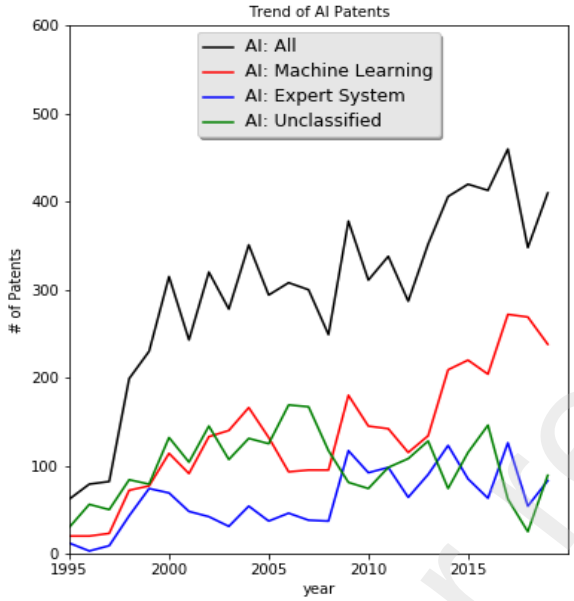


Figure 1. Rising trend of AI adoption as measured by AI-related patents in bio-pharma firms. Keywords related to machine learning include neural network, and support vector machine; keywords related to expert system include rule-based inference, and symbolic reasoning.

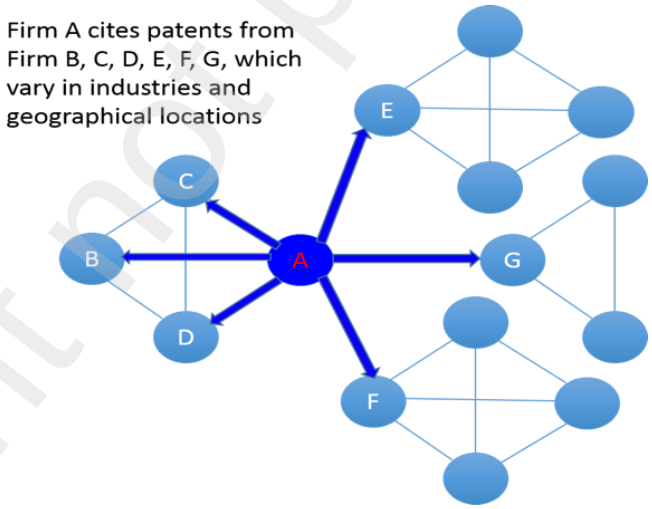


Figure 2. An illustration of our citation-network based instruments. Each node represents a firm in the network, and each edge represents an interfirm citation flow. A directed edge exists between firm A and firm B if firm A cites a patent from firm B.

Table 1. Summary Statistics and Correlation Table

Variables	Mean	Std dev.	1	2	3	4	5	6	7	8	9	10	11	12	13
1. ln(Number of Drugs)	0.065	0.30	1.00												
2. ln(Number of Drugs, with Known Mechanisms)	0.062	0.29	0.88	1.00											
3. ln(Number of Drugs, without Known Mechanisms)	0.0066	0.071	0.19	-0.27	1.00										
4. ln(Number of Drugs, Novelty: [0-0.3])	0.020	0.14	0.35	0.35	0.00	1.00									
5. ln(Number of Drugs, Novelty: [0.3-0.7])	0.050	0.27	0.60	0.49	0.18	-0.41	1.00								
6. ln(Number of Drugs, Novelty: [0.7-1])	0.0041	0.057	0.08	0.03	0.14	-0.13	-0.10	1.00							
7. ln(AI Stock)	0.78	0.96	0.51	0.46	0.12	0.06	0.45	0.10	1.00						
8. ln(AI Skills)	0.35	1.23	0.50	0.45	0.12	-0.02	0.48	0.02	0.44	1.00					
9. ln(Patent Stock)	3.60	2.15	0.44	0.42	0.04	-0.02	0.44	0.04	0.61	0.47	1.00				
10. Public Status	0.29	0.45	0.19	0.12	0.14	-0.09	0.26	0.06	0.27	0.21	0.35	1.00			
11. ln(Firm Age)	3.12	0.82	0.37	0.33	0.06	-0.04	0.38	0.01	0.42	0.33	0.34	0.19	1.00		
12. ln(Number of Employees)	6.80	2.94	0.34	0.27	0.11	0.00	0.33	0.04	0.47	0.35	0.35	0.10	0.32	1.00	
13. ln(R&D)	21.46	2.70	0.16	0.16	-0.01	-0.08	0.15	0.11	0.25	0.13	0.16	-0.11	0.15	0.32	1.00

Note:
(1). The summary statistics are reported based on the sample of firms with AIIC (the number of observations is 4,122).
(2). We add one to actual values of the variables to avoid the possibilities of taking a natural logarithm of zero.
(3). The stock-based measure of AI patents is referred to as AI stock (AI Stock).

Table 2. AI on Drugs: AIIC and Number of Drugs

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs)	(3) ln(Number of Drugs)
AIIC	0.0301***	0.0332***	
	(0.00437)	(0.00514)	
ITIC		-0.00766	
		(0.00674)	
AIIC, Machine Learning			0.0128***
			(0.00342)
ln(Patent Stock)	-0.00356	-0.00262	0.00410
	(0.00386)	(0.00394)	(0.00368)
Public Status	-0.0317**	-0.0315**	-0.0276**
	(0.0128)	(0.0128)	(0.0129)
ln(Firm Age)	0.0388***	0.0375**	0.0365**
	(0.0146)	(0.0147)	(0.0147)
ln(Number of Employees)	-0.00178	-0.00177	-0.00141
	(0.00111)	(0.00111)	(0.00112)
ln(R&D)	0.00380***	0.00383***	0.00375***
	(0.00140)	(0.00140)	(0.00141)
Observations	4,122	4,122	4,122
R-squared	0.836	0.836	0.834
Year FE	YES	YES	YES
Firm FE	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure.
(2). Column 1 shows the results for the number of drugs discovered for preclinical trials. We take ITIC into account in Column 2. In Column 3, we examine how machine learning could be linked to drug discovery. The machine learning-based AIIC is constructed from patents and skills in job postings that are related to machine learning only.
(3). *** p<0.01, ** p<0.05, * p<0.1

Table 3. AI on Drugs: AI Patents, AI Skills and Number of Drugs

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs)	(3) ln(Number of Drugs)	(4) ln(Number of Drugs)	(5) ln(Number of Drugs)
ln(AI Stock)	0.0204**	0.0302***		0.0327***	0.0330***
	(0.00936)	(0.00654)		(0.00653)	(0.00654)
ln(AI Skills)			0.0250***	0.0269***	
			(0.00513)	(0.00513)	
ln(AI Skills, Hybrid)					0.0269***
					(0.00756)
ln(AI Skills, Pure)					0.0127*
					(0.00760)
ln(Patent Stock)	0.0222***	-0.000576	0.00930***	-0.00219	-0.00239
	(0.00661)	(0.00404)	(0.00333)	(0.00404)	(0.00404)
Public Status	-0.0418**	-0.0282**	-0.0307**	-0.0322**	-0.0329**
	(0.0204)	(0.0129)	(0.0129)	(0.0128)	(0.0128)
ln(Firm Age)	0.0434***	0.0355**	0.0354**	0.0389***	0.0393***
	(0.0143)	(0.0147)	(0.0147)	(0.0146)	(0.0146)
ln(Number of Employees)	0.00842***	-0.00167	-0.00119	-0.00171	-0.00171
	(0.00258)	(0.00112)	(0.00112)	(0.00112)	(0.00112)
ln(R&D)	-0.00351	0.00382***	0.00345**	0.00374***	0.00369***
	(0.00285)	(0.00141)	(0.00140)	(0.00140)	(0.00140)
Observations	4,170	4,122	4,122	4,122	4,122
R-squared	0.857	0.835	0.835	0.836	0.836
Year FE	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure.
(2). We separately estimate the effect of each main component of our AIIC—AI patents and AI skills—on the discovery of drugs for preclinical studies. The stock-based measure of AI patents is referred to as AI stock (AI Stock). As our patent data are available prior to 2010, in Column 1, we estimate the effect of AI stock prior to 2010 and Column 2 estimates the effect after 2010.
(3). We examine the effect of AI skills in Column 3, and the effect of AI stock and AI skills together in Column 4. In Column 5, we segregate AI skills into those requiring a hybrid of AI skills and domain knowledge (AI Skills, Hybrid) and those requiring only AI skills without domain knowledge (AI Skills, Pure), and we estimate their effects as well.
(4). *** p<0.01, ** p<0.05, * p<0.1

Table 4. AI on Drugs:
AIIC, Number of Drugs with Known Mechanisms, and Novelty of Drugs

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs with Known Mechanisms)	(3) ln(Number of Drugs without Known Mechanisms)	(4) ln(Number of Drugs, Novelty: 0-0.3)	(5) ln(Number of Drugs, Novelty: 0.3-0.7)	(6) ln(Number of Drugs, Novelty: 0.7-1)
AIIC	0.0301***	0.0280***	0.00305	0.00452	0.0281***	0.00162
	(0.00437)	(0.00426)	(0.00195)	(0.00299)	(0.00405)	(0.00149)
ln(Patent Stock)	-0.00356	-0.000710	-0.00277	0.00161	-0.00326	-0.000574
	(0.00386)	(0.00376)	(0.00172)	(0.00264)	(0.00357)	(0.00131)
Public Status	-0.0317**	-0.0254**	-0.00612	0.00934	-0.0383***	-0.00402
	(0.0128)	(0.0125)	(0.00571)	(0.00879)	(0.0119)	(0.00436)
ln(Firm Age)	0.0388***	0.0359**	0.00352	-0.00231	0.0481***	0.000707
	(0.0146)	(0.0143)	(0.00651)	(0.0100)	(0.0136)	(0.00497)
ln(Number of Employees)	-0.00178	-0.00199*	3.61e-05	-0.000664	-0.00211**	0.000398
	(0.00111)	(0.00109)	(0.000497)	(0.000764)	(0.00103)	(0.000379)
ln(R&D)	0.00380***	0.00323**	0.000861	0.00131	0.00287**	0.000320
	(0.00140)	(0.00137)	(0.000624)	(0.000960)	(0.00130)	(0.000476)
Observations	4,122	4,122	4,122	4,122	4,122	4,122
R-squared	0.836	0.834	0.422	0.643	0.828	0.475
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure.
(2). Column 1 shows the results for the number of drugs discovered for preclinical trials. We examine the number of drugs with or without known mechanisms in Column 2 and Column 3.
(3). Column 4-6 shows the estimates for the drugs with chemical novelty scores in a specific range. Three ranges are created: incremental drugs with novelty score between 0 and 0.3, medium-novel drugs with novelty score between 0.3 and 0.7, and highly novel drugs with novelty score between 0.7 and 1.
(4). *** p<0.01, ** p<0.05, * p<0.1

Table 5. AI on Drugs, 2SLS

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs with Known Mechanisms)	(3) ln(Number of Drugs without Known Mechanisms)	(4) ln(Number of Drugs, Novelty: 0-0.3)	(5) ln(Number of Drugs, Novelty: 0.3-0.7)	(6) ln(Number of Drugs, Novelty: 0.7-1)
AIIC	0.0779**	0.0753**	0.00749	0.0189	0.0916**	-0.0122
	(0.0375)	(0.0362)	(0.00833)	(0.0155)	(0.0414)	(0.00967)
ln(Patent Stock)	-0.0251	-0.0220	-0.00477	-0.00489	-0.0319*	0.00566
	(0.0165)	(0.0154)	(0.00546)	(0.00672)	(0.0181)	(0.00413)
Public Status	-0.0389*	-0.0326	-0.00679*	0.00715	-0.0480***	-0.00192
	(0.0206)	(0.0208)	(0.00406)	(0.00906)	(0.0186)	(0.00485)
ln(Firm Age)	0.0489**	0.0458**	0.00445	0.000726	0.0615***	-0.00221
	(0.0195)	(0.0191)	(0.00402)	(0.00903)	(0.0217)	(0.00639)
ln(Number of Employees)	-0.00273*	-0.00293*	-5.20e-05	-0.000951	-0.00337**	0.000673
	(0.00160)	(0.00156)	(0.000339)	(0.000848)	(0.00146)	(0.000516)
ln(R&D)	0.00421***	0.00364***	0.000899**	0.00143*	0.00342***	0.000200
	(0.00123)	(0.00117)	(0.000358)	(0.000760)	(0.00124)	(0.000313)
Observations	4,122	4,122	4,122	4,122	4,122	4,122
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure.
(2). We use the total number of neighboring firms with AIIC to instrument for a focal firm's AIIC. We also use two variations of these instrumental variables: (1) the average number of AI patents in the neighboring firms; (2) the average ratio of a firm's AI patents to total patents for these neighbors. The associated first-stage F-statistic (15.4) is above the threshold of passing the weak instrument test.
(3). The effect of AIIC on drugs with or without known mechanisms of impact is reported in Column 2-3. Column 4-6 shows the estimates on the effects of AIIC on incremental drugs, medium-novel drugs and highly novel drugs.
(4). *** p<0.01, ** p<0.05, * p<0.1

Appendix

1. Chemical similarities and novelty of drugs

We measure the chemical similarity of two molecules by drawing on a central concept in chemistry, the "Similarity Property Principle". It states that structurally similar molecules should also have similar physicochemical properties and biological activities (Johnson and Maggiora 1990). We measure drug similarities by finding their maximum common substructure (MCS), a characteristic which can be used to differentiate novel chemical compounds that could potentially offer novel treatments from incremental compounds that are derivatives of existing drugs. We calculate the similarities of our focal drugs to all prior drugs and take the maximum pair-wise score to be the similarity score of the focal drug.[13]

Specifically, we calculate the pair-wise similarity score between any two drugs, X and Y, using the "Tanimoto coefficient", which is the ratio of the atoms in MCS that appears in both X and Y and all atoms that appear in both (Cao et al. 2008; Krieger et al. 2018; Nikolova and Jaworska 2003):[14]

$$Similarity_{X,Y} = \frac{N_{X\&Y}}{N_X + N_Y - N_{X\&Y}}, \qquad (1)$$

where N_X and N_Y are the total number of atoms in chemical structures of drug X and drug Y respectively, and $N_{X\&Y}$ is the total number of atoms in MCS that appears in both drugs X and Y.[15] Thus, a similarity score of zero means that the two drugs have no common components. A

[13] As robustness checks, we also restrict the previous drugs to be those within a certain time range, so that our novelty score does not automatically decrease for irrelevant structural reasons as the base of comparison becomes larger over time. Our results using a 5-year range are similar.

[14] Conventionally, any non-hydrogen atoms are included for computation. Our drug data provide the simplified molecular-input line-entry system (SMILES) code, which is a chemical notation language mainly designed for digital processing (Weininger 1988). We convert the SMILES code of each drug to its graph representation and use the graph to compute pair-wise similarity scores.

[15] MCS is one of the most accurate ways to calculate similarity and additionally provides a more flexible and efficient way of identifying important local structures (Cao et al. 2008). Although many algorithms can compute MCS in general graphs (Conte et al. 2004), they can't be applied to the study of chemical structures that tend to be represented as small and sparse graphs. Thus, we use a novel backtracking graph-matching algorithm to pinpoint the

similarity score of 1 indicates that they have the same set of atoms and bonding, although it does not imply that the two molecules are identical because MCS does not take into account the orientation in space of each molecule. Although different orientations could give them different chemical properties, they are still more similar on average to each other than to other compounds. Because similarity is highly correlated with chemical novelty, it is widely used to screen for groups of related drugs, and to digitally quantify certain chemical properties without human or animal testing (Wawer et al. 2014).

For example, we show the similarity of two drugs, Imatinib mesylate[16] and Bafetinib,[17] with their MCS highlighted in colors (see Figure A1). Imatinib mesylate is a first-generation tyrosine kinase inhibitor for treating chronic myelogenous leukemia (CML). Bafetinib is developed as a more powerful treatment and an alternative for patients who have become resistant to Imatinib mesylate. In terms of the size of their chemical structures, they both contain 42 atoms in total, with 35 of them appearing in the MCS. Therefore, their pair-wise similarity score is calculated as

$$\frac{35}{42+42-35} = 0.714.$$

This suggests 71.4% of their chemical substructure is common. After all pair-wise drug similarity scores are calculated, we can derive the maximum similarity score to all previously developed drug candidates and subtract it from 1 to calculate the novelty score. To identify a set of previously developed drugs, we use the time that development of the drug first began, and

MCS in our chemical graph representations. The core idea of this algorithm is to identify and enumerate all possible combinations of a node (for atom)/edge (for bond) mapping for a pair of chemical graphs, and then arrange these mappings into a tree-like representation with the leaf being the largest set of node/edge correspondences. The generated common substructure from this approach is then the largest overlap between the graphs of the chemical structures for our drugs. Our MCS approach eliminates any mismatches and is thus rigidly identified; it provides a lower bound for the pair-wise similarity score (Wang et al. 2013).

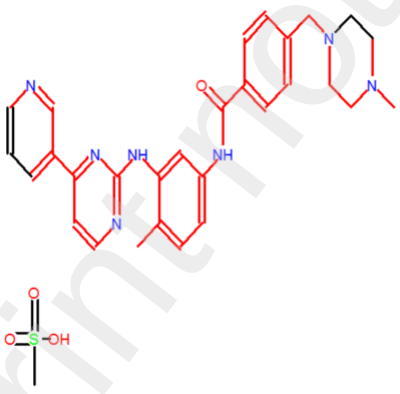
[16] PubChem profile of Imatinib mesylate: https://pubchem.ncbi.nlm.nih.gov/compound/Imatinib_mesylate

[17] PubChem profile of Bafetinib: https://pubchem.ncbi.nlm.nih.gov/compound/859212-16-1

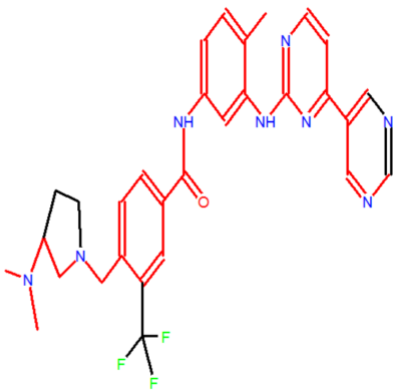
thus the novelty measure is based on the ex-ante chemical structure at the earliest development stage.

$$Drug\ Novelty_i = 1 - \max_{j \in P_i} Similarity_{i,j}, \quad (2)$$

where P_i is all drug candidates that have reached at least the Phase I stage of clinical trials prior to the initial development of the focal drug i (Krieger et al. 2018). Therefore, a novel drug candidate should have a higher novelty score and is likely to possess a molecular structure that is distinct from previous drug candidates.[18] Figure A2 plots the distribution of the drug novelty scores. Consistent with findings in Krieger et al. (2018), the novelty score can capture a substantial number of variations in drug novelty and it has been extensively tested to show its effects on drug risks, revenues, and impact. Overall, while more novel drugs are less likely to be approved by the FDA, those that are approved are more likely to be clinically effective, generate more valuable patents, and have a higher impact on the firm's market cap than are more incremental drugs (Krieger et al 2018).



Drug name: Imatinib Mesylate
Molecular formula: $C_{30}H_{35}N_7O_4S$



Drug name: Bafetinib
Molecular formula: $C_{30}H_{31}F_3N_8O$

[18] Krieger et al. (2018) discuss several limitations to the Tanimoto similarity metric, but it is still widely used for measuring drug novelty. Often the chemical properties of the most similar compound are used to estimate a newly discovered compound. Despite the broad coverage of drug information in our drug databases, we may still miss drugs at the earliest stage of development that are not recorded in the database. We address this issue by using a rolling 5-year window to compare a drug to prior drug candidates that have reached at least the Phase I stage of clinical trials.

Figure A1. Visualization of the chemical structures of two drugs: Imatinib Mesylate (left part of the plot) and Bafetinib (right part of the plot) as well as their maximum common substructure highlighted in red. For simplicity and clarity of visualization, the skeletal structural representation of chemical compounds is shown featuring the unlabeled attachment of hydrogen atoms to carbon atoms represented by the vertices of line segment for bonding together. Those atoms other than hydrogen and carbon are explicitly labeled in vertices (e.g., N, S, F). The octet rule in chemistry is satisfied to determine number of hydrogens attached to carbon atoms and number of line segments bonding the atoms. For the identification of maximum common substructure in a pair of chemical compounds, we perform an exact matching without allowing any atom or bond mismatches to be reflected in this visualization.

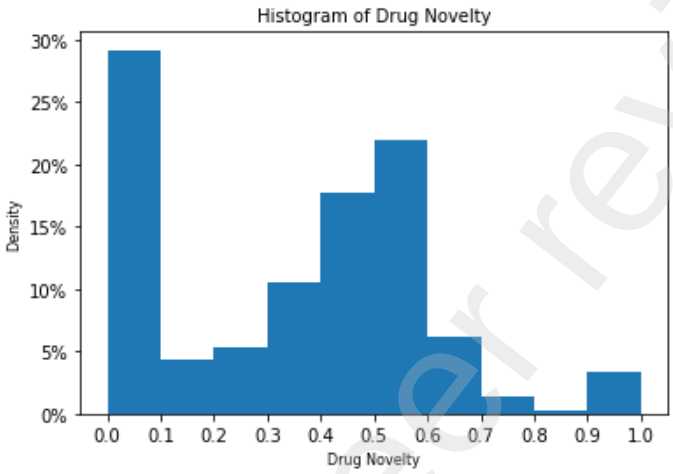


Figure A2. Histogram of novelty scores of drugs with SMILES information developed after 1995 from our drug databases.

2. A dictionary of exemplar keywords to identify AI-related patents and skills (normalized into lower case)

3d imaging	data mining	genetic algorithm	machine learning	neural network	semantic analysis
adaboost	decision tree	graphical model	maximum entropy	pattern recognition	stochastic gradient descent
anomaly detection	deep learning	hidden markov	maximum a posteriori	predictive analysis	supervised learning
artificial intelligence	defuzzification	hyperspectral imaging	maximum likelihood	predictive model	support vector machine
cloud computing	dimensionality reduction	inference engine	mechatronic	pytorch	symbolic reasoning
cluster analysis	expert system	logic program	motif discovery	random forest	tensorflow
computer vision	feature selection	logic system	motion capture	reinforcement learning	unsupervised learning
conditional random field	fuzzy logic	machine intelligence	natural language	robot	xgboost

3. Full sample analysis: estimations on the effect of AIIC on compound discovery for developing new drugs for all firms including those without AIIC

To generate the full sample of all the bio-pharma firms obtained from our drug dataset, we first aggregate patent records at the firm-year level and then match the firm-year pairs to the drug dataset. We assign zero to the number of drugs to a patent filing firm in a particular year if it developed no drugs that year. For example, if a bio-pharma firm A filed a patent in year 2016, but did not develop any drug in 2016, the number of drugs of firm A in the year 2016 is zero. However, in our data, we cannot reliably distinguish missing data from cases where a firm developed no drug in that year. This may artificially inflate the correlations among drug-related variables because a firm producing zero drugs in a given year would have zero level on any of the drug properties. Thus, when we calculate the correlations among drug-related variables, we restrict the observations to when the number of drugs developed is greater than zero. When possible, we also use a dummy to control for missing values in the data. The firm fixed effect and year fixed effect in our regression specification can also alleviate the concern about missing drug development records across the years.

Table A1. AI on Drugs:
AIIC (AI Patents and AI Skills) and Number of Drugs

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs)	(3) ln(Number of Drugs)	(4) ln(Number of Drugs)	(5) ln(Number of Drugs)
AIIC	0.0220***	0.0261***			
	(0.00368)	(0.00494)			
ITIC		-0.00678			
		(0.00545)			
AIIC, Machine Learning			0.00941***		
			(0.00324)		
ln(AI Stock)				0.0313***	0.0313***
				(0.00692)	(0.00692)
ln(AI Skills)				0.0146***	
				(0.00344)	
ln(AI Skills, Hybrid)					0.0120**
					(0.00497)
ln(AI Skills, Pure)					0.00932*
					(0.00523)
ln(Patent Stock)	-0.000146	0.000458	0.00321	-0.000841	-0.000826
	(0.00291)	(0.00295)	(0.00287)	(0.00303)	(0.00304)
Public Status	0.00329	0.00343	0.00505	0.00355	0.00363
	(0.00727)	(0.00727)	(0.00727)	(0.00728)	(0.00728)
ln(Firm Age)	0.0329***	0.0327***	0.0322***	0.0334***	0.0334***
	(0.00989)	(0.00989)	(0.00992)	(0.00991)	(0.00992)
ln(Number of Employees)	-0.000348	-0.000329	-0.000167	-0.000364	-0.000363
	(0.000814)	(0.000814)	(0.000815)	(0.000815)	(0.000815)
ln(R&D)	0.00196**	0.00198**	0.00198**	0.00198**	0.00197**
	(0.000808)	(0.000808)	(0.000809)	(0.000809)	(0.000809)
Observations	14,183	14,183	14,183	14,183	14,183
R-squared	0.641	0.641	0.640	0.641	0.641
Year FE	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure.
(2). Column 1 shows the results for the number of drugs discovered for preclinical studies in our fixed effect estimations. We take ITIC into account in Column 2. In Column 3, we examine how machine learning could be linked to drug discovery. The machine learning type of AIIC is constructed based on machine learning-related patents and skills in job postings.
(3). In Column 4-5, we estimate the effect of each main component of our AIIC—AI patents and AI skills—on the drugs discovered for preclinical studies. The stock-based measure of AI patents is referred to as AI stock (AI Stock). In Column 5, we segregate AI skills into those requiring a hybrid of AI skills and domain knowledge (AI Skills, Hybrid) and those requiring only AI skills without domain knowledge (AI Skills, Pure), and we estimate their effects as well.
(4). *** p<0.01, ** p<0.05, * p<0.1

Table A2. AI on Drugs:
AIIC, Number of Drugs with Known Mechanisms, and Novelty of Drugs

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs with Known Mechanisms)	(3) ln(Number of Drugs without Known Mechanisms)	(4) ln(Number of Drugs, Novelty: 0-0.3)	(5) ln(Number of Drugs, Novelty: 0.3-0.7)	(6) ln(Number of Drugs, Novelty: 0.7-1)
AIIC	0.0220***	0.0207***	0.00165	0.00398	0.0205***	0.000694
	(0.00368)	(0.00353)	(0.00133)	(0.00262)	(0.00283)	(0.000901)
ln(Patent Stock)	-0.000146	0.000502	-0.000508	0.00342*	-0.00249	-0.000859
	(0.00291)	(0.00279)	(0.00105)	(0.00208)	(0.00224)	(0.000713)
Public Status	0.00329	-0.00386	0.00743***	0.00279	-0.00382	0.00137
	(0.00727)	(0.00697)	(0.00263)	(0.00519)	(0.00559)	(0.00178)
ln(Firm Age)	0.0329***	0.0270***	0.00767**	0.00540	0.0289***	0.000174
	(0.00989)	(0.00948)	(0.00358)	(0.00706)	(0.00760)	(0.00242)
ln(Number of Employees)	-0.000348	-0.000850	0.000439	7.89e-05	-0.00116*	0.000362*
	(0.000814)	(0.000781)	(0.000295)	(0.000581)	(0.000626)	(0.000199)
ln(R&D)	0.00196**	0.00182**	0.000318	0.000398	0.00157**	0.000135
	(0.000808)	(0.000775)	(0.000293)	(0.000577)	(0.000621)	(0.000198)
Observations	14,183	14,183	14,183	14,183	14,183	14,183
R-squared	0.641	0.640	0.438	0.513	0.678	0.439
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure.
(2). Column 1 shows the results for the number of drugs discovered for preclinical studies in our fixed effect estimations. We examine the number of drugs with or without known mechanisms in Column 2 and Column 3.
(3). Column 4-6 shows the fixed effect estimations for the number of drugs with their chemical novelty scores in a specific range. Three ranges are created: incremental drugs with novelty score between 0 and 0.3, medium-novel drugs with novelty score between 0.3 and 0.7, and highly novel drugs with novelty score between 0.7 and 1.
(4). *** p<0.01, ** p<0.05, * p<0.1

There could be selection biases towards firms choosing to invest more in AI. To address this concern, we use the Propensity Score Matching (PSM) and Coarsened Exact Matching (CEM) method to match AI and non-AI firms based on the firm characteristics including financial ownership status, firm age, total number of employees and R&D expenses (Blackwell et al. 2009; Ho et al. 2007; Rosenbaum and Rubin 1983). Therefore, firms that exhibit similar characteristics without AIIC could be added into our AI firm sample for estimations.

Table A3. AI on Drugs:
AIIC, Number of Drugs with Known Mechanisms, and Novelty of Drugs (PSM)

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs with Known Mechanisms)	(3) ln(Number of Drugs without Known Mechanisms)	(4) ln(Number of Drugs, Novelty: 0-0.3)	(5) ln(Number of Drugs, Novelty: 0.3-0.7)	(6) ln(Number of Drugs, Novelty: 0.7-1)
AIIC	0.0313***	0.0297***	0.00227	0.00474	0.0293***	0.00118
	(0.00462)	(0.00451)	(0.00175)	(0.00304)	(0.00411)	(0.00141)
ln(Patent Stock)	-0.00473	-0.00257	-0.00204	0.00224	-0.00471	-0.00101
	(0.00400)	(0.00390)	(0.00151)	(0.00263)	(0.00355)	(0.00122)
Public Status	-0.0390***	-0.0337***	-0.00623	-0.00200	-0.0397***	0.00157
	(0.0124)	(0.0121)	(0.00471)	(0.00818)	(0.0110)	(0.00380)
ln(Firm Age)	0.0484***	0.0453***	0.00376	0.00630	0.0501***	0.000564
	(0.0156)	(0.0152)	(0.00592)	(0.0103)	(0.0139)	(0.00478)
ln(Number of Employees)	-0.00131	-0.00153	0.000110	-0.000458	-0.00189*	0.000519
	(0.00119)	(0.00116)	(0.000451)	(0.000782)	(0.00106)	(0.000364)
ln(R&D)	0.00467***	0.00427***	0.000662	0.00150*	0.00330***	0.000283
	(0.00139)	(0.00135)	(0.000525)	(0.000912)	(0.00123)	(0.000424)
Observations	5,775	5,775	5,775	5,775	5,775	5,775
R-squared	0.794	0.792	0.513	0.676	0.786	0.467
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure. Column 1-6 shows the results estimated by utilizing the propensity score matching approach on the sample that includes both firms with AIIC and firms without.
(2). Column 1 shows the results for the number of drugs discovered for preclinical studies. We examine the number of drugs with or without known mechanisms in Column 2 and Column 3.
(3). Column 4-6 shows the fixed effect estimations for the number of drugs with their chemical novelty scores in a specific range. Three ranges are created: incremental drugs with novelty score between 0 and 0.3, medium-novel drugs with novelty score between 0.3 and 0.7, and highly novel drugs with novelty score between 0.7 and 1.
(4). *** p<0.01, ** p<0.05, * p<0.1

Table A4. AI on Drugs:
AIIC, Number of Drugs with Known Mechanisms, and Novelty of Drugs (CEM)

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs with Known Mechanisms)	(3) ln(Number of Drugs without Known Mechanisms)	(4) ln(Number of Drugs, Novelty: 0-0.3)	(5) ln(Number of Drugs, Novelty: 0.3-0.7)	(6) ln(Number of Drugs, Novelty: 0.7-1)
AIIC	0.0310***	0.0284***	0.00290	0.00323	0.0277***	0.00149
	(0.00578)	(0.00559)	(0.00189)	(0.00379)	(0.00459)	(0.00144)
ln(Patent Stock)	-0.000966	0.00112	-0.00238	-0.000412	-0.000997	1.15e-05
	(0.00550)	(0.00533)	(0.00180)	(0.00361)	(0.00437)	(0.00137)
Public Status	0.00365	-0.00560	0.00853*	0.00557	-0.00405	0.00303
	(0.0134)	(0.0130)	(0.00438)	(0.00879)	(0.0106)	(0.00333)
ln(Firm Age)	0.00478	0.00258	0.00291	0.00209	0.00598	-0.00396
	(0.0177)	(0.0171)	(0.00578)	(0.0116)	(0.0140)	(0.00440)
ln(Number of Employees)	-0.000212	-0.000939	0.000622	-9.20e-05	-0.000939	0.000458
	(0.00140)	(0.00136)	(0.000459)	(0.000920)	(0.00111)	(0.000349)
ln(R&D)	0.00194	0.00188	0.000374	-0.000776	0.00269**	0.000361
	(0.00139)	(0.00134)	(0.000454)	(0.000909)	(0.00110)	(0.000345)
Observations	5,938	5,938	5,938	5,938	5,938	5,938
R-squared	0.674	0.671	0.500	0.604	0.689	0.562
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure. Column 1-6 shows the results from the coarsened exact matching method on the sample that includes both firms with AIIC and firms without.
(2). Column 1 shows the results for the number of drugs discovered for preclinical studies. We examine the number of drugs with or without known mechanisms in Column 2 and Column 3.
(3). Column 4-6 shows the fixed effect estimations for the number of drugs with their chemical novelty scores in a specific range. Three ranges are created: incremental drugs with novelty score between 0 and 0.3, medium-novel drugs with novelty score between 0.3 and 0.7, and highly novel drugs with novelty score between 0.7 and 1.
(4). *** p<0.01, ** p<0.05, * p<0.1

4. Estimations on the effect of AIIC on number of drugs using Poisson regressions

Table A5. AI on Drugs, Poisson Regressions

DV	(1) Number of Drugs	(2) Number of Drugs with Known Mechanisms	(3) Number of Drugs without Known Mechanisms	(4) Number of Drugs, Novelty: 0-0.3	(5) Number of Drugs, Novelty: 0.3-0.7	(6) Number of Drugs, Novelty: 0.7-1
AIIC	0.509***	0.484***	1.118	0.311	0.633***	0.738
	(0.150)	(0.153)	(0.721)	(0.509)	(0.161)	(1.726)
ln(Patent Stock)	-0.216	-0.112	-0.992*	-1.036	-0.124	-2.393
	(0.162)	(0.180)	(0.540)	(0.716)	(0.170)	(1.513)
Public Status	-0.0995	-0.0230	-13.95	1.048*	-2.112***	1.150
	(0.362)	(0.374)	(1,350)	(0.564)	(0.781)	(1.466)
ln(Firm Age)	1.472***	1.415***	-1.136	0.406	2.734***	-0.535
	(0.503)	(0.522)	(1.271)	(0.998)	(0.673)	(1.997)
ln(Number of Employees)	0.0202	0.0116	0.370	-0.146	0.0997	0.0279
	(0.0553)	(0.0566)	(0.283)	(0.113)	(0.0997)	(0.221)
ln(R&D)	0.0328	0.0132	0.392	-0.201	0.0228	0.655
	(0.0660)	(0.0670)	(0.458)	(0.159)	(0.0721)	(0.901)

Note:
(1). We perform Poisson regressions that model the raw number of drugs discovered for preclinical studies on the same sample (N=4,122) as shown in the main body of the paper. Each drug used in the analysis has a known chemical structure.
(2). Column 1 shows the results for the number of drugs discovered for preclinical studies. We examine the number of drugs with or without known mechanisms in Column 2 and Column 3. Column 4-6 looks into three ranges of chemical novelty scores based on the chemical structures of drugs.
(3). *** p<0.01, ** p<0.05, * p<0.1

5. Estimations on the effect of AI patents and AI skills on number of drugs with known mechanisms and novelty of drugs (for firms with AIIC)

We show the effect of AI patents and AI skills separately. Table A6 estimates the effect of AI patents. Tables A7 and Table A8 estimate the effect of hybrid and pure AI skills respectively. Table A9 shows the estimates about the effect of AI patents and two measures of AI skills together.

Table A6. AI on Drugs:
AI Patents, Number of Drugs with Known Mechanisms, and Novelty of Drugs

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs with Known Mechanisms)	(3) ln(Number of Drugs without Known Mechanisms)	(4) ln(Number of Drugs, Novelty: 0-0.3)	(5) ln(Number of Drugs, Novelty: 0.3-0.7)	(6) ln(Number of Drugs, Novelty: 0.7-1)
ln(AI Stock)	0.0302***	0.0268***	0.00478*	0.00716	0.0258***	0.00191
	(0.00654)	(0.00638)	(0.00290)	(0.00447)	(0.00606)	(0.00222)
ln(Patent Stock)	-0.000576	0.00249	-0.00307*	0.00114	0.000362	-0.000513
	(0.00404)	(0.00394)	(0.00179)	(0.00276)	(0.00375)	(0.00137)
Public Status	-0.0282**	-0.0221*	-0.00583	0.00976	-0.0350***	-0.00385
	(0.0129)	(0.0125)	(0.00570)	(0.00878)	(0.0119)	(0.00436)
ln(Firm Age)	0.0355**	0.0327**	0.00336	-0.00255	0.0448***	0.000557
	(0.0147)	(0.0143)	(0.00651)	(0.0100)	(0.0136)	(0.00497)
ln(Number of Employees)	-0.00167	-0.00186*	1.98e-05	-0.000690	-0.00197*	0.000400
	(0.00112)	(0.00109)	(0.000497)	(0.000765)	(0.00104)	(0.000380)
ln(R&D)	0.00382***	0.00323**	0.000879	0.00133	0.00287**	0.000323
	(0.00141)	(0.00137)	(0.000624)	(0.000960)	(0.00130)	(0.000476)
Observations	4,122	4,122	4,122	4,122	4,122	4,122
R-squared	0.835	0.833	0.422	0.643	0.826	0.475
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure. The stock-based measure of AI patents is referred to as AI stock (AI Stock).
(2). Column 1 shows the results for the number of drugs discovered for preclinical studies in our fixed effect estimations. We examine the number of drugs with or without known mechanisms in Column 2 and Column 3.
(3). Column 4-6 shows the fixed effect estimations for the number of drugs with their chemical novelty scores in a specific range. Three ranges are created: incremental drugs with novelty score between 0 and 0.3, medium-novel drugs with novelty score between 0.3 and 0.7, and highly novel drugs with novelty score between 0.7 and 1.
(4). *** p<0.01, ** p<0.05, * p<0.1

Table A7. AI on Drugs:
Hybrid AI Skills, Number of Drugs with Known Mechanisms, and Novelty of Drugs

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs with Known Mechanisms)	(3) ln(Number of Drugs without Known Mechanisms)	(4) ln(Number of Drugs, Novelty: 0-0.3)	(5) ln(Number of Drugs, Novelty: 0.3-0.7)	(6) ln(Number of Drugs, Novelty: 0.7-1)
ln(AI Skills, Hybrid)	0.0307***	0.0304***	0.000398	-0.000145	0.0327***	0.00203
	(0.00650)	(0.00633)	(0.00288)	(0.00444)	(0.00602)	(0.00220)
ln(Patent Stock)	0.00928***	0.0112***	-0.00141	0.00365	0.00860***	0.000107
	(0.00333)	(0.00325)	(0.00148)	(0.00228)	(0.00309)	(0.00113)
Public Status	-0.0316**	-0.0256**	-0.00571	0.0100	-0.0389***	-0.00408
	(0.0129)	(0.0126)	(0.00572)	(0.00880)	(0.0119)	(0.00437)
ln(Firm Age)	0.0347**	0.0322**	0.00291	-0.00328	0.0446***	0.000517
	(0.0147)	(0.0143)	(0.00651)	(0.0100)	(0.0136)	(0.00497)
ln(Number of Employees)	-0.00123	-0.00148	9.60e-05	-0.000574	-0.00160	0.000428
	(0.00112)	(0.00109)	(0.000495)	(0.000762)	(0.00103)	(0.000378)
ln(R&D)	0.00337**	0.00282**	0.000832	0.00127	0.00245*	0.000294
	(0.00140)	(0.00137)	(0.000624)	(0.000960)	(0.00130)	(0.000476)
Observations	4,122	4,122	4,122	4,122	4,122	4,122
R-squared	0.835	0.833	0.422	0.643	0.827	0.475
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure.
(2). Column 1 shows the results for the number of drugs discovered for preclinical studies in our fixed effect estimations. We examine the number of drugs with or without known mechanisms in Column 2 and Column 3.
(3). Column 4-6 shows the fixed effect estimations for the number of drugs with their chemical novelty scores in a specific range. Three ranges are created: incremental drugs with novelty score between 0 and 0.3, medium-novel drugs with novelty score between 0.3 and 0.7, and highly novel drugs with novelty score between 0.7 and 1.
(4). *** p<0.01, ** p<0.05, * p<0.1

Table A8. AI on Drugs:
Pure AI Skills, Number of Drugs with Known Mechanisms, and Novelty of Drugs

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs with Known Mechanisms)	(3) ln(Number of Drugs without Known Mechanisms)	(4) ln(Number of Drugs, Novelty: 0-0.3)	(5) ln(Number of Drugs, Novelty: 0.3-0.7)	(6) ln(Number of Drugs, Novelty: 0.7-1)
ln(AI Skills, Pure)	0.0253***	0.0243***	0.00163	0.00236	0.0259***	0.00117
	(0.00656)	(0.00639)	(0.00291)	(0.00447)	(0.00607)	(0.00222)
ln(Patent Stock)	0.00959***	0.0115***	-0.00142	0.00361	0.00896***	0.000136
	(0.00333)	(0.00325)	(0.00148)	(0.00228)	(0.00309)	(0.00113)
Public Status	-0.0282**	-0.0222*	-0.00573	0.00992	-0.0352***	-0.00383
	(0.0129)	(0.0126)	(0.00571)	(0.00878)	(0.0119)	(0.00436)
ln(Firm Age)	0.0355**	0.0329**	0.00307	-0.00298	0.0453***	0.000505
	(0.0147)	(0.0143)	(0.00651)	(0.0100)	(0.0136)	(0.00497)
ln(Number of Employees)	-0.00110	-0.00135	0.000102	-0.000566	-0.00147	0.000435
	(0.00112)	(0.00109)	(0.000495)	(0.000762)	(0.00103)	(0.000378)
ln(R&D)	0.00353**	0.00298**	0.000834	0.00127	0.00262**	0.000305
	(0.00141)	(0.00137)	(0.000623)	(0.000959)	(0.00130)	(0.000476)
Observations	4,122	4,122	4,122	4,122	4,122	4,122
R-squared	0.834	0.833	0.422	0.643	0.826	0.475
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure.
(2). Column 1 shows the results for the number of drugs discovered for preclinical studies in our fixed effect estimations. We examine the number of drugs with or without known mechanisms in Column 2 and Column 3.
(3). Column 4-6 shows the fixed effect estimations for the number of drugs with their chemical novelty scores in a specific range. Three ranges are created: incremental drugs with novelty score between 0 and 0.3, medium-novel drugs with novelty score between 0.3 and 0.7, and highly novel drugs with novelty score between 0.7 and 1.
(4). *** p<0.01, ** p<0.05, * p<0.1

Table A9. AI on Drugs:
AI Patents, Hybrid AI Skills, Pure AI Skills, Number of Drugs with Known Mechanisms, and Novelty of Drugs

DV	(1) ln(Number of Drugs)	(2) ln(Number of Drugs with Known Mechanisms)	(3) ln(Number of Drugs without Known Mechanisms)	(4) ln(Number of Drugs, Novelty: 0-0.3)	(5) ln(Number of Drugs, Novelty: 0.3-0.7)	(6) ln(Number of Drugs, Novelty: 0.7-1)
ln(AI Stock)	0.0330***	0.0295***	0.00484*	0.00719	0.0286***	0.00209
	(0.00654)	(0.00638)	(0.00291)	(0.00448)	(0.00606)	(0.00222)
ln(AI Skills, Hybrid)	0.0269***	0.0270***	-0.000184	-0.00123	0.0289***	0.00212
	(0.00756)	(0.00738)	(0.00337)	(0.00519)	(0.00701)	(0.00257)
ln(AI Skills, Pure)	0.0127*	0.0115	0.00192	0.00329	0.0121*	0.000152
	(0.00760)	(0.00742)	(0.00339)	(0.00522)	(0.00705)	(0.00259)
ln(Patent Stock)	-0.00239	0.000704	-0.00312*	0.00110	-0.00154	-0.000631
	(0.00404)	(0.00394)	(0.00180)	(0.00277)	(0.00374)	(0.00137)
Public Status	-0.0329**	-0.0267**	-0.00589	0.00979	-0.0399***	-0.00418
	(0.0128)	(0.0125)	(0.00572)	(0.00881)	(0.0119)	(0.00437)
ln(Firm Age)	0.0393***	0.0363**	0.00358	-0.00224	0.0486***	0.000752
	(0.0146)	(0.0143)	(0.00652)	(0.0100)	(0.0136)	(0.00498)
ln(Number of Employees)	-0.00171	-0.00190*	2.57e-05	-0.000677	-0.00201*	0.000395
	(0.00112)	(0.00109)	(0.000498)	(0.000766)	(0.00103)	(0.000380)
ln(R&D)	0.00369***	0.00311**	0.000880	0.00134	0.00273**	0.000313
	(0.00140)	(0.00137)	(0.000624)	(0.000961)	(0.00130)	(0.000477)
Observations	4,122	4,122	4,122	4,122	4,122	4,122
R-squared	0.836	0.834	0.423	0.643	0.828	0.475
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES	YES

Note:
(1). Each drug used in the analysis has a known chemical structure. The stock-based measure of AI patents is referred to as AI stock (AI Stock).
(2). Column 1 shows the results for the number of drugs discovered for preclinical studies in our fixed effect estimations. We examine the number of drugs with or without known mechanisms in Column 2 and Column 3.
(3). Column 4-6 shows the fixed effect estimations for the number of drugs with their chemical novelty scores in a specific range. Three ranges are created: incremental drugs with novelty score between 0 and 0.3, medium-novel drugs with novelty score between 0.3 and 0.7, and highly novel drugs with novelty score between 0.7 and 1.
(4). *** p<0.01, ** p<0.05, * p<0.1

References

"Mechanism matters". 2010. *Nature Medicine* (16:4), pp. 347-347.

Abrams, D., and Sampat, B. 2017. "What's the Value of Patent Citations? Evidence from Pharmaceuticals."

Agrawal, A., McHale, J., and Oettl, A. 2019. "Artificial Intelligence, Scientific Discovery, and Commercial Innovation," Working Paper.

Alekseeva, L., Azar, J., Gine, M., Samila, S., and Taska, B. 2019. "The Demand for Ai Skills in the Labor Market," *Available at SSRN:* *https://ssrn.com/abstract=3470610*).

Aral, S., and Weill, P. 2007. "It Assets, Organizational Capabilities, and Firm Performance: How Resource Allocations and Organizational Differences Explain Performance Variation," *Organization science* (18:5), pp. 763-780.
Babina, T., Fedyk, A., He, A. X., and Hodson, J. 2020. "Artificial Intelligence, Firm Growth, and Industry Concentration," *Firm Growth, and Industry Concentration (July 14, 2020)*).
Backman, T. W., Cao, Y., and Girke, T. 2011. "Chemmine Tools: An Online Service for Analyzing and Clustering Small Molecules," *Nucleic acids research* (39:suppl_2), pp. W486-W491.
Bardhan, I., Krishnan, V., and Lin, S. 2013. "Research Note—Business Value of Information Technology: Testing the Interaction Effect of It and R&D on Tobin's Q," *Information Systems Research* (24:4), pp. 1147-1161.
Barney, J. 1991. "Firm Resources and Sustained Competitive Advantage," *Journal of management* (17:1), pp. 99-120.
Bassellier, G., and Benbasat, I. 2004. "Business Competence of Information Technology Professionals: Conceptual Development and Influence on It-Business Partnerships," *MIS quarterly*), pp. 673-694.
Bernstein, S. 2015. "Does Going Public Affect Innovation?," *The Journal of Finance* (70:4), pp. 1365-1403.
Bharadwaj, A. S. 2000. "A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Investigation," *MIS quarterly*), pp. 169-196.
Blackwell, M., Iacus, S., King, G., and Porro, G. 2009. "Cem: Coarsened Exact Matching in Stata," *The Stata Journal* (9:4), pp. 524-546.
Brown, F. K. 1998. "Chemoinformatics: What Is It and How Does It Impact Drug Discovery," *Annual reports in medicinal chemistry* (33), pp. 375-384.
Brynjolfsson, E., Rock, D., and Syverson, C. 2018. "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," 0898-2937, National Bureau of Economic Research.
Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., Henke, N., and Trench, M. 2017. "Artificial Intelligence–the Next Digital Frontier," *McKinsey Glob Institute*).
Burton-Jones, A., and Straub Jr, D. W. 2006. "Reconceptualizing System Usage: An Approach and Empirical Test," *Information systems research* (17:3), pp. 228-246.
Cao, Y., Jiang, T., and Girke, T. 2008. "A Maximum Common Substructure-Based Algorithm for Searching and Predicting Drug-Like Compounds," *Bioinformatics* (24:13), pp. i366-i374.
Chan, Y., and Levallet, N. 2013. "It Capabilities–Quo Vadis?,").
Cockburn, I. M., Henderson, R., and Stern, S. 2018. "The Impact of Artificial Intelligence on Innovation," National Bureau of Economic Research.
Cohen, W. M., Nelson, R. R., and Walsh, J. P. 2000. "Protecting Their Intellectual Assets: Appropriability Conditions and Why Us Manufacturing Firms Patent (or Not)," National Bureau of Economic Research.
Conte, D., Foggia, P., Sansone, C., and Vento, M. 2004. "Thirty Years of Graph Matching in Pattern Recognition," *International journal of pattern recognition and artificial intelligence* (18:03), pp. 265-298.
DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. 2016. "Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs," *Journal of health economics* (47), pp. 20-33.
Dixon, J., Hong, B., and Wu, L. 2021. "The Robot Revolution: Managerial and Employment Consequences for Firms," *Management Science*:forthcoming).
Dougherty, D., and Dunne, D. D. 2012. "Digital Science and Knowledge Boundaries in Complex Innovation," *Organization Science* (23:5), pp. 1467-1484.
Drews, J. 2000. "Drug Discovery: A Historical Perspective," *science* (287:5460), pp. 1960-1964.
Du Plessis, M., Van Looy, B., Song, X., and Magerman, T. 2009. "Data Production Methods for Harmonized Patent Indicators: Assignee Sector Allocation," *Luxembourg: EUROSTAT Working Paper and Studies*).

Eklund, J. 2018. "The Knowledge-Incentive Trade-Off: Understanding the Organization Design & Innovation Relationship," *Academy of Management Proceedings*: Academy of Management Briarcliff Manor, NY 10510, p. 16747.
Fleming, N. 2018. "How Artificial Intelligence Is Changing Drug Discovery," *Nature* (557:7707), p. S55.
Forman, C., Goldfarb, A., and Greenstein, S. 2016. "Agglomeration of Invention in the Bay Area: Not Just Ict," *American Economic Review* (106:5), pp. 146-151.
Gilchrist, D. S. 2016. "Patents as a Spur to Subsequent Innovation? Evidence from Pharmaceuticals," *American Economic Journal: Applied Economics* (8:4), pp. 189-221.
Grant, R. M. 1991. "The Resource-Based Theory of Competitive Advantage: Implications for Strategy Formulation," *California management review* (33:3), pp. 114-135.
Griliches, Z., Pakes, A., and Hall, B. H. 1986. "The Value of Patents as Indicators of Inventive Activity." National Bureau of Economic Research Cambridge, Mass., USA.
Hall, B. H. 1990. "The Manufacturing Sector Master File: 1959-1987," National Bureau of Economic Research.
Hall, B. H., Jaffe, A., and Trajtenberg, M. 2005. "Market Value and Patent Citations," *RAND Journal of economics*), pp. 16-38.
Hall, B. H., Jaffe, A. B., and Trajtenberg, M. 2001. "The Nber Patent Citation Data File: Lessons, Insights and Methodological Tools," National Bureau of Economic Research.
Hammer, M. 1990. "Reengineering Work: Don't Automate, Obliterate," *Harvard Business Review* (68:4), pp. 104-112.
He, K., Zhang, X., Ren, S., and Sun, J. 2015. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification," *Proceedings of the IEEE international conference on computer vision*, pp. 1026-1034.
Hemphill, C. S., and Sampat, B. N. 2011. "When Do Generics Challenge Drug Patents?," *Journal of Empirical Legal Studies* (8:4), pp. 613-649.
Hess, A. M., and Rothaermel, F. T. 2011. "When Are Assets Complementary? Star Scientists, Strategic Alliances, and Innovation in the Pharmaceutical Industry," *Strategic Management Journal* (32:8), pp. 895-909.
Hitt, L., Wu, L., Campbell, K., Jeafarqomi, K., Ashtiani, H., and Levesque, L. 2018. "Corporate Data Literacy: Scoring Firms and Firm Performance," IHS Markit.
Ho, D. E., Imai, K., King, G., and Stuart, E. A. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political analysis* (15:3), pp. 199-236.
Hu, J., Shen, L., and Sun, G. 2018. "Squeeze-and-Excitation Networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141.
Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. 2011. "Principles of Early Drug Discovery," *British journal of pharmacology* (162:6), pp. 1239-1249.
Jayaraj, S., and Gittelman, M. 2018. "Scientific Maps and Innovation: Impact of the Human Genome on Drug Discovery." Doctoral dissertation, Rutgers University.
Johnson, M. A., and Maggiora, G. M. 1990. *Concepts and Applications of Molecular Similarity*. Wiley.
Jones, C. I. 2005. "Growth and Ideas," in *Handbook of Economic Growth*. Elsevier, pp. 1063-1111.
Joshi, K. D., Chi, L., Datta, A., and Han, S. 2010. "Changing the Competitive Landscape: Continuous Innovation through It-Enabled Knowledge Capabilities," *Information Systems Research* (21:3), pp. 472-495.
Kapoor, R., and Klueter, T. 2015. "Decoding the Adaptability–Rigidity Puzzle: Evidence from Pharmaceutical Incumbents' Pursuit of Gene Therapy and Monoclonal Antibodies," *academy of management journal* (58:4), pp. 1180-1207.
Kearns, G. S., and Sabherwal, R. 2006. "Strategic Alignment between Business and Information Technology: A Knowledge-Based View of Behaviors, Outcome, and Consequences," *Journal of management information systems* (23:3), pp. 129-162.

Kleis, L., Chwelos, P., Ramirez, R. V., and Cockburn, I. 2012. "Information Technology and Intangible Output: The Impact of It Investment on Innovation Productivity," *Information Systems Research* (23:1), pp. 42-59.
Krieger, J. L., Li, D., and Papanikolaou, D. 2018. "Developing Novel Drugs," National Bureau of Economic Research.
Levin, R. C., Klevorick, A. K., Nelson, R. R., Winter, S. G., Gilbert, R., and Griliches, Z. 1987. "Appropriating the Returns from Industrial Research and Development," *Brookings papers on economic activity* (1987:3), pp. 783-831.
Liang, V. 2020. "Baidu's Ai-Related Patented Technologies: Doing Battle with Covid-19," in: *Wipo Magazine*. Geneva, Switzerland: Wipo.
Magerman, T., Van Looy, B., and Song, X. 2006. "Data Production Methods for Harmonized Patent Statistics: Patentee Name Harmonization,").
Mak, K.-K., and Pichika, M. R. 2019. "Artificial Intelligence in Drug Development: Present Status and Future Prospects," *Drug discovery today* (24:3), pp. 773-780.
Marchant, J. 2020. "Powerful Antibiotics Discovered Using Ai," *Nature*).
Marcus, G., and Davis, E. 2019. *Rebooting Ai: Building Artificial Intelligence We Can Trust*. Pantheon.
Markman, G. D., Espina, M. I., and Phan, P. H. 2004. "Patents as Surrogates for Inimitable and Non-Substitutable Resources," *Journal of management* (30:4), pp. 529-544.
Nikolova, N., and Jaworska, J. 2003. "Approaches to Measure Chemical Similarity–a Review," *QSAR & Combinatorial Science* (22:9‐10), pp. 1006-1026.
Nonaka, I., and Von Krogh, G. 2009. "Perspective—Tacit Knowledge and Knowledge Conversion: Controversy and Advancement in Organizational Knowledge Creation Theory," *Organization science* (20:3), pp. 635-652.
Pisano, G. P. 2006. *Science Business: The Promise, the Reality, and the Future of Biotech*. Harvard Business Press.
Rajkumar, S. V. 2020. "The High Cost of Prescription Drugs: Causes and Solutions." Nature Publishing Group.
Ravichandran, T., Han, S., and Mithas, S. 2017. "Mitigating Diminishing Returns to R&D: The Role of Information Technology in Innovation," *Information Systems Research* (28:4), pp. 812-827.
Raymond, P., Yoav, S., Erik, B., Jack, C., John, E., Barbara, G., Terah, L., James, M., Saurabh, M., and Carlos, N. J. 2019. "The Ai Index 2019 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University.
Romer, P. M. 1990. "Endogenous Technological Change," *Journal of political Economy* (98:5, Part 2), pp. S71-S102.
Rosenbaum, P. R., and Rubin, D. B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* (70:1), pp. 41-55.
Rotman, D. 2019. "Ai Is Reinventing the Way We Invent," *MIT Technology Review*).
Santhanam, R., and Hartono, E. 2003. "Issues in Linking Information Technology Capability to Firm Performance," *MIS quarterly*), pp. 125-153.
Scannell, J. W., Blanckley, A., Boldon, H., and Warrington, B. 2012. "Diagnosing the Decline in Pharmaceutical R&D Efficiency," *Nature reviews Drug discovery* (11:3), p. 191.
Smietana, K., Siatkowski, M., and Møller, M. 2016. "Trends in Clinical Success Rates," *Nature Reviews Drug Discovery* (15:6), pp. 379-380.
Smith, W. 2020. "How Atomwise Uses Artificial Intelligence for Drug Discovery."
Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., and Bloom-Ackerman, Z. 2020. "A Deep Learning Approach to Antibiotic Discovery," *Cell* (180:4), pp. 688-702. e613.
Tambe, P., and Hitt, L. M. 2012. "The Productivity of Information Technology Investments: New Evidence from It Labor Data," *Information Systems Research* (23:3-part-1), pp. 599-617.
Tambe, P., Hitt, L. M., Rock, D., and Brynjolfsson, E. 2019. "It, Ai and the Growth of Intangible Capital," *Available at SSRN 3416289*).

Teece, D. J. 1998. "Capturing Value from Knowledge Assets: The New Economy, Markets for Know-How, and Intangible Assets," *California management review* (40:3), pp. 55-79.
Trafton, A. 2020. "Artificial Intelligence Yields New Antibiotic," in: *MIT News*.
Tu, Y. 2011. "The Discovery of Artemisinin (Qinghaosu) and Gifts from Chinese Medicine," *Nature medicine* (17:10), p. 1217.
Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., and Spitzer, M. 2019. "Applications of Machine Learning in Drug Discovery and Development," *Nature Reviews Drug Discovery*), p. 1.
Varian, H. 2018. "Artificial Intelligence, Economics, and Industrial Organization," 0898-2937, National Bureau of Economic Research.
Von Hippel, E. 1994. ""“Sticky Information” and the Locus of Problem Solving: Implications for Innovation," *Management science* (40:4), pp. 429-439.
Wang, Y., Backman, T. W., Horan, K., and Girke, T. 2013. "Fmcsr: Mismatch Tolerant Maximum Common Substructure Searching in R," *Bioinformatics* (29:21), pp. 2792-2794.
Wawer, M. J., Li, K., Gustafsdottir, S. M., Ljosa, V., Bodycombe, N. E., Marton, M. A., Sokolnicki, K. L., Bray, M.-A., Kemp, M. M., and Winchester, E. 2014. "Toward Performance-Diverse Small-Molecule Libraries for Cell-Based Phenotypic Screening Using Multiplexed High-Dimensional Profiling," *Proceedings of the National Academy of Sciences* (111:30), pp. 10911-10916.
Webb, M. 2019. "The Impact of Artificial Intelligence on the Labor Market," *Available at SSRN 3482150*).
Weininger, D. 1988. "Smiles, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules," *Journal of chemical information and computer sciences* (28:1), pp. 31-36.
WIPO. 2019. "Wipo Technology Trends 2019: Artificial Intelligence."
Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., and Sayeeda, Z. 2018. "Drugbank 5.0: A Major Update to the Drugbank Database for 2018," *Nucleic acids research* (46:D1), pp. D1074-D1082.
Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. 2008. "Drugbank: A Knowledgebase for Drugs, Drug Actions and Drug Targets," *Nucleic acids research* (36:suppl_1), pp. D901-D906.
Wu, L., Hitt, L., and Lou, B. 2020. "Data Analytics, Innovation, and Firm Productivity," *Management Science* (66:5), pp. 2017-2039.
Wu, L., Jin, F., and Hitt, L. M. 2017. "Are All Spillovers Created Equal? A Network Perspective on Information Technology Labor Movements," *Management Science* (64:7), pp. 3168-3186.
Wu, L., Lou, B., and Hitt, L. M. 2019. "Data Analytics Supports Decentralized Innovation," *Management Science* (65:10).
Zhang, H., Zhang, L., Li, Z., Liu, K., Liu, B., Mathews, D. H., and Huang, L. 2020. "Lineardesign: Efficient Algorithms for Optimized Mrna Sequence Design," *arXiv preprint arXiv:2004.10177*).
Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., and Asadulaev, A. 2019. "Deep Learning Enables Rapid Identification of Potent Ddr1 Kinase Inhibitors," *Nature biotechnology* (37:9), pp. 1038-1040.

Acknowledgments

We would like to thank the senior editors, the associate editor, and three anonymous reviewers for constructive comments and suggestions. We also appreciate the generous financial support from Mack Institute for Innovation Management of The Wharton School, University of Pennsylvania.

About the Authors

Bowen Lou is an assistant professor of Operations and Information Management at the School of Business, University of Connecticut. He received his Ph.D. from The Wharton School, University of Pennsylvania. He conducts research on economics of artificial intelligence and innovation. His work has appeared in *Management Science*.

Lynn Wu is an associate professor of operation, information and decisions at the Wharton School of Business, University of Pennsylvania. She researches and teaches how emerging information technologies, such as artificial intelligence and data analytics, affect innovation, business strategy, and productivity. She has won several best paper awards from top journals and flagship conferences in information systems and received two early career awards in information systems.