

SCIENTIFIC MAPS AND INNOVATION:  
IMPACT OF THE HUMAN GENOME MAP ON DRUG DISCOVERY

by

SEBASTIAN JAYARAJ

A Dissertation submitted to the  
Graduate School-Newark  
Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Dr. Michelle Gittelman

And approved by

Dr. Michelle Gittelman: \_\_\_\_\_

Dr. Sara Parker-Lue: \_\_\_\_\_

Dr. Jerry Kim: \_\_\_\_\_

Dr. Bhaven Sampat: \_\_\_\_\_

Newark, New Jersey

May, 2018

© 2018

Sebastian Jayaraj

ALL RIGHTS RESERVED

ABSTRACT OF THE DISSERTATION

SCIENTIFIC MAPS AND INNOVATION: IMPACT OF THE HUMAN GENOME ON

DRUG DISCOVERY

By SEBASTIAN JAYARAJ

Dissertation Director: Dr. Michelle Gittelman

Scientific mapping projects like the ongoing US BRAIN initiative and Human Cell Atlas are data and resource intensive endeavors that curate immense amounts of information. While both public and private funds support such mapping of the scientific knowledge landscape, the impact of such projects on innovation is not well understood. Innovation can be conceptualized as a spatial process, where firms and inventors search either locally or in distant, lesser explored, territories. In my dissertation, I asked whether scientific maps influence technological search, and if so how.

To address this empirically, I chose the Human Genome Project (HGP) as my context and asked how it affected drug discovery. HGP, the largest publicly-funded biology project, released the complete human genome map in 2000, enabling scientists to identify and focus on disease-causing genes as targets for drug discovery. Using a novel dataset of chemistry drug patents, I tracked firm search processes pre- and post-HGP. To measure how the HGP map impacts firm exploration, I also developed a novel method based on chemical similarities to capture search trajectories over time. My conclusions on how the HGP map influenced inventive activity, innovation strategy and outcomes are

detailed in three essays and build on existing theories in the innovation literature. Briefly, I found that the HGP map increased the rate of novel compound production, exploration and impacted firm innovation strategies. This was influenced by prior firm knowledge, market competition and product market specialization.

This study informs on mechanisms by which basic science driven scientific maps influence industry innovation. My findings bear implications on public policy, R&D management and firm strategy.

## DEDICATION

For Aakanksha,  
my Saraswati

## ACKNOWLEDGEMENTS

This has been a long journey made possible through the efforts of many people – past and present. I would like to express my sincere thanks and gratitude to my advisor Dr. Michelle Gittelman for mentoring and guiding me through this intellectual marathon. I really appreciate her contributions of time, stimulating discussions and invaluable feedback on my research. She has been so positive and supportive throughout the PhD process. Having the right ensemble, can raise the experience to another level. For this, I would like to thank my stellar dissertation committee members: Dr. Sara Parker Lue, Dr. Bhaven Sampat and Dr. Jerry Kim. Each of them has been incredibly generous with their time, insights and support.

I cannot thank my wife, Aakanksha Singhvi, enough for her enduring support and encouragement. From being my sounding board for ideas, to constantly reminding me of the larger picture and keeping me on track, she has been invaluable to this journey.

My sincere thanks to Batia Wiesenfeld and Lori Rosenkopf for getting me started in management research. The wonderful faculty, students and administration at Rutgers University have played an important part in shaping my research interests and PhD experience. I thank - Dr. Fariborz Damanpour, Dr. Deborah Dougherty, Dr. Petra Christmann, Dr. John Cantwell, Dr. Marya Doerfel and Goncalo Filipe. My undergraduate students for teaching me a valuable life skill – entertaining millennials through music and lecture slides!

This research was shaped by industry experts who shared their experiences and insights on the biopharmaceutical industry. I thank Ragu Bharadwaj, George Johnston, Jefferson Tilley, Mahmud Hassan, Derek Lowe, Daniel V. Santi, Ralph Reid, Steven Dong, Jeff Blaney, Craig Funt and Sumati Murli for their time and discussions. To the late Dave Weininger, thank you for inventing SMILES – which has been indispensable to this research, showing me the Rio Grande and for being such an inspiring human being.

My dissertation has benefitted immensely from the valuable insights and feedback the innovation community at the CCC Doctoral conference and the Academy of Management conference; especially, Gino Cattani, Chris Tucci, Rajshree Agarwal, Rahul Kapoor and Keld Laursen. I am grateful to the Wharton Mack Institute Emerging Scholar Workshop and thank Daniel Levinthal, Ron Adner, Connie Helfat and Sidney Winter for introducing me to evolutionary perspectives on strategic management.

I am grateful to the funding sources that made my dissertation research possible. I was supported by the National Science Foundation's Science of Science Policy and Innovation grant award # 1664380, the Rutgers Advanced Institute for the Study of Entrepreneurship and Development research grant and the Dean's summer research fellowships.

Where we think also matters. I am grateful to the gorgeous reading rooms and books made available by the New York Public Library (especially, the Map Division and Rose

room for invigorating this research on scientific maps and knowledge exploration), The Rockefeller University Library and the Seattle Public Library.

I would like to express my gratitude to my family for encouraging me throughout this process. My dad, Victor Jayaraj, for teaching me life's most valuable lessons: I wish you could have seen me complete this journey. Prof. Ashok Singhvi for being an intellectual parent and gracious mentor.

My gurus, Indrajit and the late Subroto Roy-Chowdhury, for teaching me the Sitar – a source of peace and tranquility the last few years. To all my friends who have supported me through this process, and I have not been able to mention here, thank you.

## TABLE OF CONTENTS

Acknowledgements	v
List of Tables	x
List of Figures	xi
Summary	1
1. Scientific Maps And Innovation: Impact Of The Human Genome On Drug Discovery	8
1.1 Introduction	9
1.2 Theory	11
1.3 Research Setting	17
1.4 Data	21
1.5 Research Design	36
1.6 Results	43
1.7 Discussion and Conclusions	60
1.8 References	64
1.9 Appendix	71
2. Scientific Maps and Innovation Strategy: Impact of the Human Genome on the Adoption of Targeted Strategies	82
2.1 Introduction	83
2.2 Theory	86
2.3 Research Setting	93

2.4 Data	98
2.5 Research Design	113
2.6 Results	117
2.7 Discussion and Conclusions	129
2.8 References	134
2.9 Appendix	139
3. Scientific Maps and Exploration: Tracking Technological Search Trajectories in Drug Discovery	143
3.1 Introduction	144
3.2 Background on Technological Search	146
3.3 Data	150
3.4 Method	153
3.5 Results	168
3.6 Discussion and Conclusions	177
3.7 References	180
3.8 Appendix	185

## LIST OF TABLES

Table 1.1	Competition and Crowding around Top Genes	46
Table 1.2	Descriptive Statistics	50
Table 1.3	Differences-in-Differences Estimate (Drugs vs Materials Science)	51
Table 1.4	Differences-in-Differences Estimate (Target vs Non-Target)	58
Table 2.1	Top Patenting Firms	104
Table 2.2	Correlation Table	119
Table 2.3	Logistic Regression Estimates	121
Table 2.4	Ordinary Least Squares Estimates	123
Table 3.1	Pairwise Tanimoto Coefficients	157
Table 3.2	Sample Markush Patents	164
Table 3.3	Application of Clustering Method	165

## LIST OF FIGURES

Figure 1.1	Example Markush Structure	23
Figure 1.2	Original Markush Patent	26
Figure 1.3	Patent Applications by Patent Office	31
Figure 1.4	Distribution of Drug Patents	32
Figure 1.5	Sample Patent Abstract	35
Figure 1.6	Comparison of Novel Compound Production	41
Figure 1.7	Experimental Setup	43
Figure 1.8	Industry Patenting	44
Figure 1.9	Firm Patenting	44
Figure 1.10	Firm Level Increase in Patenting	45
Figure 1.11	Changes in Firm Strategies	45
Figure 1.12	Novel Compounds by Search Strategy	49
Figure 1.13	Novel Compound Production (Control vs Treatment)	50
Figure 1.14	Novel Compound Production (Target vs Non-Target)	57
Figure 2.1	Differences in Search Process	97
Figure 2.2	Distribution of WIPO Patents	101
Figure 2.3	Construction of Longitudinal Panel	102
Figure 2.4	Sample Patent Abstract	108
Figure 2.5	Increase in Target-based Adoption	118
Figure 3.1	Chemical Similarity	156
Figure 3.2	Structural Comparisons of Compounds	156
Figure 3.3	Sample Markush Structure	161
Figure 3.4	Tanimoto based Clustering	162

Figure 3.5	Jaccard Distances	176
Figure 3.6	Firm Exploration using Chemical Distances (3 Firms)	169
Figure 3.7	Firm Exploration using Chemical Distances (578 Firms)	169
Figure 3.8	Firm Exploration using Chemical Distances (578 Firms)	170
Figure 3.9	Firm Exploration for Target-Based	170
Figure 3.10	Firm Exploration for Non-Target Based	171
Figure 3.11	Firm Exploration using Euclidean Distances	174
Figure 3.12	Firm Exploration using Euclidean Distances (3 Firms)	174
Figure 3.13	Industry Exploration using Euclidean Distances	175
Figure 3.14	Industry Exploration using Chemical Distances	176

## **SUMMARY: IMPACT OF THE HUMAN GENOME ON DRUG DISCOVERY**

Scientific mapping projects like the ongoing US BRAIN initiative and Human Cell Atlas are data and resource intensive endeavors that curate immense amounts of information. In the past, early human dissections and detailed anatomical maps laid the groundwork for modern surgical training and practice. While both public and private funds support such mapping of the scientific knowledge landscape, the economic value of such projects to markets and innovation is not yet well established. Theorists indicate that empirical studies on how scientific knowledge influences technological search can lead to new insights on the process and mechanisms driving innovation. I address this gap by building on existing theories of science as a map, technological change and evolutionary perspectives (Nelson & Winter, 1982; Utterback, 1994; Fleming & Sorenson, 2001). In my dissertation, I asked whether scientific maps influence technological search, and if so how. To address this empirically, I chose the Human Genome Project (HGP) as my context and examined how it affected drug discovery.

The Human Genome Project (HGP) is the largest publicly-funded biology project. In 2000, HGP released a precise and detailed map of the human genome, which allowed scientists to predict disease-related gene targets and enabled targeted drug design. These changes in the nature of drug discovery provide a unique opportunity to study the process of innovation. Prior studies on the HGP have examined the role of institutional arrangements and patenting strategies related to innovation (Huang & Murray, 2010; Huang & Murray, 2013). Recent research on the human genome examines the role of

intellectual property rights on follow-on innovation by comparing the effect of patent protected and publicly available gene sequences (Williams, 2013; Sampat & Williams, 2017). Other work on the role of mapping on innovation indicate the positive effect of publicly available geographical maps on gold-mining efforts (Nagaraj, 2015). But the mechanisms by which a scientific map influences innovation remain yet to be established. Building on this literature, I examine how the HGP map impacts the process of innovation and outcomes in the drug industry.

The process of innovation is conceptualized as a spatial process (technological search), where firms and inventors search either locally or in unknown, new areas (March, 1991). This search process progresses by combining new and existing technological knowledge technologies (Nelson & Winter, 1982; Henderson & Clark, 1990). Innovation theorists suggest that scientific knowledge enables predictive capabilities and reduces uncertainty in the search for new products. (Fleming & Sorenson, 2001). Predictive learning routines based on abstracted representations of the search space are theorized as leading to successful outcomes in complex technological landscapes as they provide an unbiased representation (Arora and Gambardella, 1998; Gavetti and Levinthal, 2000). Puranam & Swamy (2010) argue that mental representations like maps, even if incomplete, can be useful in situations involving coordinated problem-solving. Therefore, it is theorized that mapping of scientific knowledge can help firms predict innovation outcomes without full experimentation.

My research setting is drug discovery in the biopharmaceutical industry – a sector driven by high costs and high failure rates. In 2015, the US biopharma industry spent \$59 billion on R&D; it costs \$2.6 billion and 10-15 years to develop a single drug, of which

less than 12% succeeded (PhRMA, 2016). Drug discovery is a complex problem as drug-target interactions are unpredictable and can lead to off-target effects (Scannell, et al, 2012; Gittelman, 2016). In drug discovery, scientists work towards identifying drug targets and biological mechanisms responsible for the disease. Given a drug target, medicinal chemists then try to design compounds that can bind to the drug target. To do this, they comb through a large theoretical landscape of solutions called chemical space – containing more than  $10^{60}$  possible combinations. Without a map to navigate this high-dimensional search space, medicinal chemists rely on available disease knowledge and prior experience to make calculated guesses on the types of chemicals that could work.

The human genome map helped in navigating his chemical search space by providing an accurate list of about 10,000 potential disease targets. This made it possible to model disease targets and design compounds that could fit precisely with the target (Drews, 2000; Gittelman, 2016). Thus, the availability of an accurate human genome allowed for predictive search and a high level of specificity in the search for novel compounds.

I utilize a novel dataset of chemistry-based drug patents called Markush patents to contrast the process of technological search before and after the HGP map. Patents are central to intellectual property protection and appropriation in biopharmaceutical industry and widely used in economic analyses (Scott & Sampat, 2012). As of 2016, more than 90% of marketed drugs are small molecules making this class of drugs economically important and relevant for this study. Very early in the drug discovery stage, a special type of patent known as Markush patent allows applicants to apply claims for broad

regions of chemical space. These Markush patents predate all other drug-related patents and mark the starting point of small-molecule discovery projects.

While not all Markush patent applications end up becoming granted patents or lead to a drug, they represent the universe of early stage small-molecule R&D activity – not just the discovery projects that become successful. This makes Markush patents especially useful for analyzing the effects of technological change on search trajectories of firms. For my empirical study, I have collected about 39,000 drug-related Markush patent applications filed between 1990-2004, and extracted millions of novel compounds embedded in them to analyze the regions captured in chemical space. A detailed discussion of data sampling and Markush patents is provided in the data section of essay one of the dissertation.

The dissertation is composed of three essays that study the impact of the human genome map on drug innovation. In essay one, I explore the overall effect of the human genome on drug discovery at the industry level. Two mechanisms are explored by which scientific maps can influence the search process: i) predictive power of the map's impact on novel compound production and firm search strategies, and ii) open accessibility of the map's impact on competition. Using a differences-in-differences estimation I test the impact of the human genome map on novel compound production.

In essay two, I examine the map's impact on the adoption of targeted drug strategies by incumbent firms. The map increased the number of disease gene targets and made accessible new regions of the genome, opening up new market opportunities. Since the detailed map was beneficial mainly to target-based strategies, firms experienced in this approach had an advantage compared to other firms. To unpack the organizational

factors driving adoption of target-based strategies, I examine the role of prior firm knowledge, specialized capabilities and competition.

In essay three, I examine how the map impacted firm exploration in chemical space. I build on prior methodological advances in the innovation literature and computational chemistry to introduce a novel technique to measure technological distance between patents based on chemical structures. Using computational chemistry search techniques, I measure chemical structural similarities to calculate technological distance between patents (Johnson & Maggiora, 1992). This technique is applied on a sample of small molecule drug patents to examine exploration trajectories pre and post the HGP map and to assess the role of search strategies in exploration.

To summarize, in my dissertation research I found that the human genome map increased the rate of invention, novel compound production and altered technological search strategies in firms. Further, I found that prior firm knowledge, market competition and product market specialization are influential in adopting targeted strategies. Overall, targeted strategies lead to broad exploration compared to non-target based strategies. These results reveal an interesting finding: scientific maps not only increase the rate and intensity of technological search but also impact exploration. This study informs on mechanisms by which basic science driven scientific maps influence industry innovation. These findings bear implications for public policy, R&D management and firm strategy.

## References

- Arora, Ashish, and Alfonso Gambardella. "The changing technology of technological change: general and abstract knowledge and the division of innovative labour." *Research policy* 23.5 (1994): 523-532.
- Cohen, Wesley M., and Daniel A. Levinthal. "Absorptive capacity: A new perspective on learning and innovation." *Administrative science quarterly* (1990): 128-152.
- Drews, Jürgen. "Drug discovery: a historical perspective." *Science* 287.5460 (2000): 1960-1964.
- Gittelman, Michelle. "The revolution re-visited: Clinical and genetics research paradigms and the productivity paradox in drug discovery." *Research Policy* (2016).
- Hemphill, C. Scott, and Bhaven N. Sampat. "When do generics challenge drug patents?." *Journal of Empirical Legal Studies* 8.4 (2011): 613-649.
- Huang, Kenneth G., and Fiona E. Murray. "Does patent strategy shape the long-run supply of public knowledge? Evidence from human genetics." *Academy of Management Journal* 52.6 (2009): 1193-1221.
- Johnson, M. A., and G. M. Maggiora. "Concepts and Applications of Molecular Similarity, John Wiley & Sons." (1992).
- March, James G. "Exploration and exploitation in organizational learning." *Organization science* 2.1 (1991): 71-87.
- Nagaraj, A. (2015, November 6). The Private Impact of Public Maps— Landsat Satellite Imagery and Gold Exploration. Job Market Paper, MIT Sloan School of Management, Cambridge, MA. Available: [http://web.mit.edu/nagaraj/files/nagaraj\\_jmp\\_nov6.pdf](http://web.mit.edu/nagaraj/files/nagaraj_jmp_nov6.pdf)

Nelson, Richard R., and G. Sidney. "Winter. 1982." An evolutionary theory of economic change (1982): 929-964.

Puranam, Phanish, and Murali Swamy. "Expeditions without Maps: Why Faulty Initial Representations May Be Useful in Joint Discovery Problems." Available at SSRN 1153142 (2010).

Sampat, Bhaven, and Heidi L. Williams. How do patents affect follow-on innovation? Evidence from the human genome. No. w21666. National Bureau of Economic Research, 2015.

Scannell, Jack W., et al. "Diagnosing the decline in pharmaceutical R&D efficiency." Nature reviews

Utterback, James. "Mastering the dynamics of innovation: how companies can seize opportunities in the face of technological change." (1994).

Williams, Heidi L. "Intellectual property rights and innovation: Evidence from the human genome." Journal of Political Economy 121.1 (2013): 1-27.

# **1. SCIENTIFIC MAPS AND INNOVATION: IMPACT OF THE HUMAN GENOME ON DRUG DISCOVERY**

## **Abstract**

How scientific maps influence technological search is not well understood. In this study, I analyze the impact of the human genome - a precise scientific map, on drug discovery. In the drug industry, deep understanding of the disease is essential for designing new therapeutics. Two mechanisms are explored by which scientific maps influence technological search: 1) predictive power and precision of the map can alter existing search strategies and rate of invention 2) public nature of the map can influence competition and intensity of search. Using a unique data set of chemistry drug patents, analyses show a rapid increase in patenting and novel compounds correlated with the availability of the human genome. Difference-in-differences estimation indicate that the map increases novel compound production across the industry by 64%, controlling for other technological changes like combinatorial chemistry. Overall, scientific maps increase the rate of invention and novel compound production suggesting that scientific mapping projects can enhance industry innovation.

Keywords: scientific maps, technological search, drug innovation, human genome

## **Introduction**

Scientific maps like cartographic maps provide a high-level representation of the knowledge landscape aiding researchers in charting new discoveries and exploration. For example, early human dissections and detailed drawings (anatomical maps) laid the groundwork for modern surgical training and practice. Public scientific mapping projects like the Human Genome Project (HGP), the on-going Human Cell Atlas, European Human Brain Project and US BRAIN initiative organize immense amounts of scientific knowledge offering researchers a birds-eye view of the terrain and new paths to explore. While considerable resources are spent on public mapping projects, the economic value of maps to markets and innovation is not well understood. Theorists suggest that empirical studies on how scientific knowledge influences technological search can lead to new insights on the process and mechanisms driving innovation (Fleming & Sorenson, 2001). In this study, I examine how the introduction of a scientific map influences the process of innovation across an industry.

The role of scientific maps on technological innovation or mapping as an economic activity is not yet well established. Recently, scholars have examined the role of the Human Genome in the context of follow on innovations and intellectual property rights (Sampat & Williams, 2015; Williams, 2013). Others have examined how differences in the accessibility and quality of geographic maps impact gold mining activity (Nagaraj, 2015). But very few empirical studies are available on the impact of scientific maps on technological search and innovation. Science mapping provides organized and detailed representations of scientific knowledge, that facilitate rapid

knowledge sharing and advancement. For example, the Allen Brain Atlas details taxonomies, images and cellular descriptions of human and mouse brains facilitating research in academia and industry. Some scholars indicate that new scientific knowledge can direct innovated-related search towards more useful knowledge combinations in complex landscapes and predict their outcomes (Nelson, 1982; Fleming & Sorenson, 2004). Other scholars have argued that in complex fields, such predictive search may be of limited value; instead, experiential, feedback-based learning is associated with creative insights, and therefore, more successful technological innovation (Gittelman, 2016; Nightingale, 1998; Vincenti, 1990). Therefore, the exact mechanisms by which science and mapping influence processes of innovation are not well understood.

In 2000, the Human Genome Project (HGP), the world's largest publicly-funded biology project, made available the first draft of the human genome. This was essentially a scientific map that allowed for identification of human genes and their exact DNA sequences enabling a high level of specificity in the search for potential drug candidates. At around the same time, a novel technology called combinatorial chemistry emerged, that allowed chemists to create millions of related compounds at relatively low cost in a single step. Together, the technological and scientific breakthroughs transformed the search process by making possible efficient search of much greater scale, along with a higher precision in the target of search. These changes in the nature of discovery provide a unique opportunity to study the process of technological innovation under different informational and technological contexts like a scientific map.

Prior research on the HGP has examined how the map influenced institutional arrangements and intellectual property rights related to innovation, but the role of the human genome as a scientific map remains unexplored (Huang & Murray, 2010; Williams, 2013). I explore two mechanisms by which scientific maps influence technological search: 1) predictive power and precision of a scientific map increases knowledge recombination leading to more search and diverse solutions 2) public nature of the map increases competition and stronger efforts to protect intellectual property.

I have assembled an original and unique database of 39,000 small molecule drug patent applications over 15-years. Small-molecules drugs are economically important as they account for 90% of all marketed drugs as of 2016. These early-stage chemistry drug patent applications known as Markush patents allow for granular and precise tracking of firm level technological search pre and post the human genome map. Results presented in this exploratory study shed new light on important but largely unexplored questions on the scientific maps and their impact on innovative activity. Understanding the role of scientific maps in innovation is important for management theory, public policy and firm strategy.

## **Theory**

### **Scientific Maps and Technological Search**

In evolutionary perspectives, innovations have two dimensions: technical novelty and market selection, and firm innovative activities are aimed at developing novelty of economic value (Nelson & Winter, 1982). The generation of technical novelty – driven

by new knowledge and technology - leads to diversity and a better chance of market selection. In this search for technical novelty, firms can both explore new technological space and/or exploit prior knowledge (March, 1991). The balance depends partly on the relative costs of exploration and exploitation and the ability to apply prior expertise towards future projects.

Innovation scholars theorize invention as a search process over technological landscapes, involving recombination of new and existing component technologies (Henderson & Clark, 1990; Fleming & Sorenson, 2001). In the absence of broader knowledge about the search space or predictive theory, experiential, feedback-based learning guides the search process. (Cohen & Levinthal, 1990). Some scholars indicate that scientific knowledge can act as a map that improves firms' ability to explore the technological landscape, by reducing uncertainty in technological search and lower the cost of exploration. And that science can help direct towards more useful new combinations when the search landscape is complex – like drug discovery, which requires inter-disciplinary knowledge (Fleming & Sorenson, 2004). Predictive learning routines based on abstracted representations of the search space are theorized as leading to successful outcomes in complex landscapes as they provide an accurate, unbiased representation of the search space (Arora and Gambardella, 1998; Gavetti and Levinthal, 2000). Thus, scientific knowledge is theorized as allowing firms to predict outcomes without full experimentation, enabling them to work with a smaller set of possible combinations.

In the context of small molecule drug discovery, searching for an optimal compound in chemical space is a time-consuming and complex task (Scannell, et al, 2012). This is partly because chemical space is a high-dimensional search space with  $10^{60}$  possible combinations of novel compounds. The process of drug discovery can be conceptualized as a Kauffman fitness landscape<sup>1</sup>, where two main components (biological gene targets and chemical agents;  $N=2$ ) interact with high-interdependence ( $K$ ). In this landscape multiple compounds can interact with varying efficacy with one target or more targets, thus making the interactions multiplex, interdependent and overlapping. Interdependence or coupling between components occurs where changes made to one component requires changes to another for the system to work properly (Ulrich, 1995; Fleming & Sorenson, 2001). In the case of drug discovery, this implies that disease target and designed compounds are tightly inter-related and change in knowledge about one of them will impact the other, thereby impacting coordinated outcomes.

With an estimated 20,000 human gene targets and  $10^{60}$  possible novel compounds, for a drug to work effectively there needs to be tight coupling between the disease target and therapeutic compound. Mismatches in coupling can lead to suboptimal outcomes, such as off-target effects that result in clinical trial failures. Therefore, biological targets and chemistry-based drugs are tightly-coupled components with high degree of interdependence. The large scale of possible interactions makes it an increasingly difficult task to find useful combinations (i.e. successful drugs) (Kauffman et al, 2000;

---

<sup>1</sup> Drawing on evolutionary biology, Stuart Kauffman's fitness landscape is a mathematical model to describe the evolution of populations. It is visualized as having peaks and valleys across which populations evolve and follow different paths.

Rivkin, 2000). In fact, phenotypic screening and serendipitous discoveries have led to a majority of modern medicine's successful drugs (Sampat, 2012). Without a scientific map of either component available, novel recombination would be difficult. But with access to a scientific map of at least one component (i.e. human genome), recombination and predictive capabilities become feasible. New knowledge occurs when new information is integrated and/or recombined with existing knowledge of the problem giving rise to breakthrough ideas and innovations (Schilling & Green, 2011).

Puranam & Swamy (2010) argue that mental representations like maps, even if incomplete, can be useful in situations involving coordinated problem-solving. In drug discovery, when local search breaks down due to reuse and re-optimization of existing compounds (i.e. phenotypic search paradigm), science acts as map in transforming the process of drug discovery into more directed identification of useful components. Access to new technological tools and scientific maps can be expected to reduce firms' dependence on local, feedback-based search, and learning-by-doing, in their R&D activities. The new gains in predictive power and increased efficiency could result in modified or new search strategies. Thus, a scientific map like the human genome can direct search towards more useful combinations and potentially alter modes of search in drug discovery.

*Hypothesis 1a: Scientific maps will increase overall inventive activity*

*Hypothesis 1b: Scientific maps will alter technological search strategies*

## **Scientific Maps and Competition**

Drug patenting can be compared to the mineral claim system instituted by the American West in the middle-to-late 1800's (Kitch, 1977). Exclusive prospecting rights whether gold, oil or valuable minerals were given to the finder who was first to file claims to a piece of land. Similarly, in fields like chemicals and pharmaceuticals, firms aggressively try to "patent block" rivals from patenting related inventions or developing substitutes. These defensive patents fence off competitors and are also used to force rivals into negotiations (Cohen, Nelson & Walsh, 2000; Ziedonis, 2004; McGrath & Nerkar, 2003). In the drug industry, firms patent aggressively to block rivals from entering their product markets and disease areas.

The emergence of combinatorial chemistry and the human genome within the same decade provided a synergistic effect in the search for drugs. The map was a source of new genetic information that could be exploited to identify new drug targets and disease mechanisms (Triggle, 2006). The map supplied much useful information about "where to search" for new targets compared to hypotheses generated by researchers. Easy accessibility to off-the-shelf tools enabled assembly of combinatorial chemistry platforms in-house. Instead of just shooting at a broad region of a target with a single-loading rifle, firms now could locate and take aim at precise molecular targets (with the genome map), and fire at them with an automatic weapon (combinatorial chemistry tools).

The chemical space associated with new drug targets is prime real-estate and can lead to aggressive patenting and patent blocking activity – where competing firms try to

capture as much territory as possible. This has important financial and downstream implications – firms can enter new disease and product markets, out license valuable chemical space and potentially block competitors from entering markets. These conditions can lead to patent races and aggressive protection of firm intellectual properties in an attempt to obtain exclusive and broad prospecting rights. Williams' (2010) analysis of differences in intellectual property rights between the Human Genome Project (public domain) and Celera's gene sequences (privately held) indicate a 20-30% drop in scientific research and product development. Even small firms will try to capture as much chemical territory as possible:

“Smaller companies are quicker to patent to establish room to operate. You would think that a (large) firm misses something, let's do something similar with better properties. You make a series of compounds, write down 18 permutations that you could exploit that their patent didn't cover very well, then find one that's interesting, better pharmacokinetics, more potent, etc. To the point where you have a patentably distinct advantage over theirs (large firms).” – Interview with Industry Expert, 2016

With specific gene targets as starting points for discovery, firms and inventors could narrow their focus and concentrate their efforts on making only target-specific solutions. One could argue, that firms will now make a smaller set of accurate keys to fit the lock, resulting in fewer but more specific compounds. But the public nature of the map – that is, competitors also get to work on same targets, combined with availability of patents in the public domain, could instead drive inventors and firms to develop more compounds to block off competitors. Patent protection of novel compounds gives firms the ability to exclude competitors from copying, entering a product market or downstream negotiation rights. But publishing the compounds and their structures in patents has a flipside – competing firms can examine, learn and copy the designs.

Medicinal chemists call such unpatented chemical opportunities “holes” and try to capture closely related chemical space, as their closeness is also translatable into biological activity (Southall & Ajay, 2005). For example, Pfizer’s Viagra and Eli Lilly’s Cialis are closely related chemical structures. Hence, to fence out regions of chemical space and block competitors from entering, search efforts around gene targets will intensify after the availability of the map. This competitive pressure and race to patent chemical space will increase rate and intensity of search activity.

Therefore, the human genome map can be expected to increase industry competition, resulting in an increase in diversity of solutions produced.

*Hypothesis 2: Scientific maps will increase intensity of search*

### **Research Setting**

The research setting for this study is drug discovery in the biopharmaceutical industry – a sector driven by high research and development (R&D) investment, long development cycles and a high rate of failure. In 2015, the US biopharma industry spent \$59 billion on R&D; it costs \$2.6 billion and 10-15 years to develop a single drug, of which less than 12% succeeded (PhRMA, 2016). The human body is a complex system where drug-target interactions are unpredictable and lead to off-target effects and undesirable drug-drug interactions making drug discovery a challenging and costly task (Drews, 2000; Scannell, et al, 2012).

*Process of Drug Discovery:* In the early days of drug discovery, natural products, crude extracts or purified compounds were screened for biological activity. When an active compound or natural product was found, it was tested or modified into a therapeutic agent. For example, aspirin is a natural product derived from trees and flowers. This approach of separating bioactive compounds, or synthesizing and modifying them without a clear understanding of the drug mechanisms involved, is known as *phenotypic drug discovery* (Scannell, et al, 2012; Lipinski & Hopkins 2004).

In contrast to this trial-and-error process, *target-based discovery* starts with a therapeutic area of interest (e.g. coronary artery disease) along with a competitive analysis of existing drugs and patent landscapes. A drug target is any protein (e.g. enzymes, receptors) or nucleic acid (DNA, RNA) involved in a disease related biological pathway to which a drug can bind and alter its state. For example, Pfizer's Lipitor, the best-selling drug of all time, is a lipid-lowering compound which works by inhibiting HMG-CoA reductase, an enzyme responsible for cholesterol production in the liver. Starting with a drug target, medicinal chemists design and synthesize compounds that can bind to the target. For this, medicinal chemists search in *chemical space* – which is extremely large containing  $10^{60}$  possible combinations of unique drug-like molecules. Through design and testing, a candidate compound is developed that is as specific as possible to the drug target and disease (Drews, 2000).

*Human Genome Project and Drug Discovery:* On June 26, 2000, the Human Genome Project (HGP), an international public-private consortium, sequenced and publicly

released the first complete draft of the human genome. It was the world's largest collaborative biology project costing \$3 billion and an important development in human biology that was aimed, in part, at revolutionizing the way in which new drugs were discovered (Gittelman, 2016). The human genome is a digital map of 3 billion DNA base pairs revealing the location and identity of genes which encoded various proteins in cells. An important milestone of the HGP was the revision of possible human genes - the number of available disease targets increased from a few hundred to at least 10,000 targets (Drews, 2000; Tripp & Grueber, 2011). These were important advances for drug discovery, as the new map allowed for predictive search and a high level of specificity in the search for lead candidates. The human genome map was a major scientific advance that drew a lot of public attention.

“Without a doubt, this is the most important, most wondrous map ever produced by humankind.” - US President Bill Clinton, 2000

For drug discovery, the human genome provided a precise map of genes and gene variants. Scientists could predict 3-dimensional protein structures using gene sequences – a field known as structural genomics. This substituted for the prior method in which labs would spend many years trying to solve the 3-dimensional crystal structures of proteins. Prior to HGP, there were less than 2000 human protein structures available (RCSB PDB, 2016). But having the ability to predict protein structures using gene sequences allowed scientists to rapidly identify disease mechanisms. The map was a disruptive event that fundamentally changed the way pharmaceutical firms approached the problem of drug discovery.

*Combinatorial Chemistry and High-throughput Capabilities:* Automated technology complemented the discovery of new compounds impacting the cost and speed of search. A novel technique called combinatorial chemistry emerged in the mid 1990's, that allowed chemists to create millions of related compounds in a single step. This technology significantly reduced the time taken to synthesize novel compounds which at the time was a rate limiting issue in drug discovery. In 1996, a chemist made 4 compounds/month at an average cost of \$7500/compound. Using combinatorial chemistry, a chemist could make 3,300 compounds at an average cost of \$12/compound (See cost comparison in Appendix). This led to rapid adoption of combinatorial chemistry in the drug industry.

“In the late ‘90s people started to believe that if you made larger “lead-like”<sup>2</sup> libraries of compounds, your odds of success would improve. A lot of technologies evolved to enable machine-based synthesis in parallel with computer-based tracking of compounds for screening. The excitement around the technologies inspired creation of a new ACS Journal of Combinatorial Chemistry”. - Interview with Industry Expert, 2016

Computational tools allowed modeling of the 3-dimensional structures of gene targets with millions of chemical structures. This allowed efficient and rapid sampling of the very large chemistry landscape. Parallel advances in industrial robotics facilitated screening of millions of compounds for pharmacological properties – a process known as high-throughput screening. Together the combined availability of powerful computational tools, combinatorial techniques and the human genome map gave firms the

---

<sup>2</sup> A lead compound is a chemical compound that has pharmacological and biological activity that could be therapeutically useful, but needs to be optimized to fit better with the disease target

ability to search rapidly in chemical space. These changes were expected to, in turn, vastly improve the number of new drugs entering into trials and the rate at which new compounds could be discovered.

## **Methods**

### **Data**

In this study, I am interested testing the impact of the human genome map on drug discovery projects. So ideally, this would include project level data that captured the early stage drug discovery efforts of all firms, which could then be analyzed for influences of the human genome map. Such a set up would allow me to categorize projects and test firm and industry level effects of the scientific map. Given the difficulty and secrecy involved in obtaining firm's internal R&D investments, I looked for external indicators of early stage R&D activity that could be both comprehensive across the industry and also accurate. In the innovation and strategy field, patents are well established as a measure of innovation and have been used to study R&D activity and knowledge in firms for a long time (Jaffe, 1986; Jaffe, et al; 1993).

For my empirical research, I utilize a novel dataset of chemistry-based drug patents called Markush patents<sup>3</sup> to contrast the process of technological search before and after the map. These are early stage drug discovery patents and are well suited to test the effect of the human genome map. This section provides an introduction to Markush patents and describes the study's data collection process in further detail.

---

<sup>3</sup> These chemistry patents contain a special place-holder structure called Markush structure (first appeared in a patent in 1924) named after their inventor Eugene Markush, to capture a broad set of compounds.

*Drug Patents:* Patents are central to intellectual property protection and appropriation in biopharmaceutical industry and widely used in economic analyses (Scott & Sampat, 2012). In the United States, there are four main classes of patents: composition of matter, method, machine and article of manufacture. While there are several types of patents available, the following are relevant for the drug industry: composition of matter or product patent, process patent, formulation patent and method of use patent. Process patents claim the specific compound or processes involved in making the drug. Formulation patents contain the pharmaceutical dosage form of the drug. Method of use patents capture the use of the drug for a particular disease, preventing others from doing the same.

Drug patents can be grouped based on the type of drug product – namely biologics (protein, antibodies, DNA based therapies) and small molecules (chemistry-based drugs). The drug discovery process is very different for each of these drug therapies, involves very specialized knowledge and the products mode of intervention is also distinct. For my analysis, I focus only small molecule patents as they comprise more than 90% of marketed drugs and are economically important. Focusing on only one type of drug patent (chemistry-based) implies uniformity in selection and comparison of drug projects across firms.

### **Markush Patents in Drug Discovery**

New chemical substances are specified as composition of matter patents. This type of product patent is considered superior to the other patents in terms of broad claims and

gives the patent assignee full rights to make, sell or license this property. The new chemical entities can be claimed by chemical name or by chemical structure, or claimed within a broad Markush structure. Apart from the drug industry, Markush patents are also used in other areas which involve chemicals like materials science, industrial chemicals and agrobiotechnology. The common aspect in all Markush patents is the presence of a special place-holder structure called Markush structure named after their inventor Eugene Markush, to capture a broad set of compounds and was used for the first time in a patent in 1924. Figure 1 below shows a Markush structure that claims a broad range of actual compounds that can be synthesized.

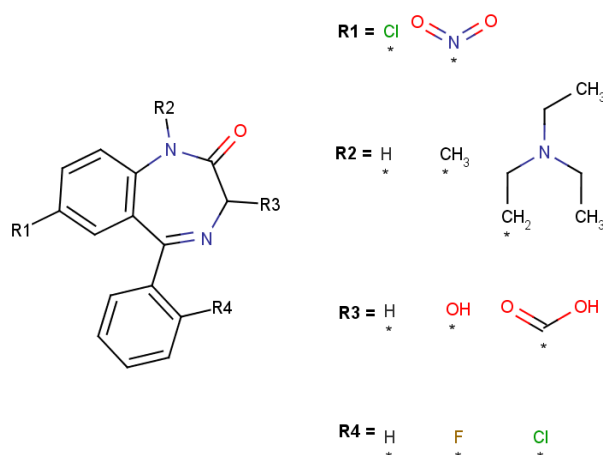


Figure 1: Example Markush structure with R groups representing various chemical groups. Each such Markush structure can represent hundreds of actual compounds (image source: ChemAxon Inc.)

The chemical structure on the left is the backbone structure on which various R-groups are situated on. This structural backbone is known as the chemical scaffold. In the

example provided above, R1-R4 are the placeholders that can accept different chemical groups for each position. In this case R1 has 2 options, R2-R4 have 3 options, thus yielding  $2 \times 3 \times 3 \times 3 = 54$  novel combinations. Instead of listing out these 54 combinations, the Markush structure efficiently captures this combinatorial set of novel compounds. A Markush structure is used in patent applications to define the scope of a chemical series (Langdon, et al, 2011). Markush patents can try to capture a broad region sometimes going into thousands of combinations, but if these claims are construed as too broad they can be rejected by the patent examiner.

Markush patents are filed very early in the discovery stage where a firm makes general claims for a number of molecules without revealing the identity of the exact compound or compounds it is pursuing. In the drug industry, Markush patents are very specific to small molecule or chemistry drugs (compared to biologics patents) and mark the starting point where firms begin to claim intellectual property rights by planting “flags” in chemical space (Southall & Ajay, 2005). The compounds of interest are hidden among hundreds of other similar looking compounds with the same basic scaffold structure (Figure 1). Other drug patents like process, method of use and formulation patents sequentially follow the Markush patent in terms of timing and normally reference this starting patent. In fact, some Markush patents are also found in the Food & Drug Administration’s Orange Book database of approved drug products.

Are Markush patents optional? In the drug industry, Markush patents are de facto starting points that all firms tend to file to protect their intellectual property. It does not

make sense for a firm to file only one chemical compound (and structure) if competitors can find ways to replicate it by making slight alterations to the protected structure.

“I’m not sure if I’ve ever seen a (modern) non-Markush drug patent, come to think of it. The only time you’d do that is if you’re claiming a completely new (unexpected, not taught toward, etc.) use for a known compound, and then you’re not claiming the chemical matter per se. But I’ve never seen a chemical matter claim that didn’t have variable groups in it, somewhere.” – Interview with Industry expert

An interesting feature of Markush patents is that they encapsulate the actual compounds made, and those that the firm or inventor plans to protect for future use – sometimes, running into the hundreds or thousands of novel compounds. Thus, these compounds within Markush patents represent the explorative effort undertaken by the firm or inventor in chemical space and the chemical compounds they wish to protect. For example, the drug project and compound that became the blockbuster drug Viagra was initially aimed at reducing hypertension and chest pain due to heart disease (Figure 2). A side effect of the drug allowed it to be repurposed to a new market. While not all Markush patent applications end up becoming granted patents or lead to a drug, they represent the small-molecule related R&D activity and explorative effort of the firm – not just the compounds that end up becoming successful. In 2002, Pfizer used a Markush patent US6469012 B1 to sue ICOS and Eli Lilly for marketing Cialis and infringing upon their chemical space (a similar looking chemical structure).


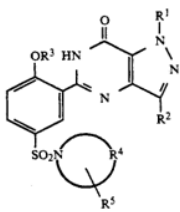
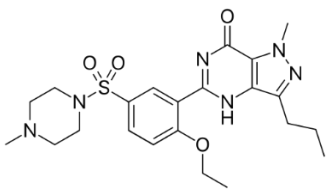
 US005250534A		
<b>United States Patent</b> [19]		[11] <b>Patent Number:</b> <b>5,2</b>
<b>Bell et al.</b>		[45] <b>Date of Patent:</b> <b>Oct.</b>
[54] <b>PYRAZOLOPYRIMIDINONE ANTIANGINAL AGENTS</b>		Attorney, Agent, or Firm—Peter C. Richards C. Benson; James T. Jones
[75] Inventors: <b>Andrew S. Bell; David Brown; Nicholas K. Terrett</b> , all of Groton, Conn.		[57] <b>ABSTRACT</b> Compounds of the formula:
[73] Assignee: <b>Pfizer Inc.</b> , New York, N.Y. [21] Appl. No.: <b>882,988</b> [22] Filed: <b>May 14, 1992</b>		
<b>Related U.S. Application Data</b> [63] Continuation of Ser. No. 717,227, Jun. 18, 1991, abandoned.		
<b>Foreign Application Priority Data</b> Jun. 20, 1990 [GB] United Kingdom ..... 9013750		
[51] Int. Cl. <sup>5</sup> ..... <b>A61K 31/505; C07D 487/04</b> [52] U.S. Cl. .... <b>514/258; 544/262</b> [58] Field of Search ..... <b>544/262; 514/258</b>		
<b>References Cited</b> <b>U.S. PATENT DOCUMENTS</b> 4,052,390 10/1977 Broughton et al. .... 544/118 <b>FOREIGN PATENT DOCUMENTS</b> 1095688 8/1988 Australia . 3309689 10/1989 Australia . 0201188 12/1986 European Pat. Off. . 0347146 12/1989 European Pat. Off. . 0349239 1/1990 European Pat. Off. . 0351058 1/1990 European Pat. Off. . 0352960 1/1990 European Pat. Off. . 0371731 6/1990 European Pat. Off. . <b>OTHER PUBLICATIONS</b> Hamilton, et al., J. Med. Chem., 30, 91-96 (1987). Primary Examiner—Nicholas S. Rizzo Assistant Examiner—Y. N. Gupta		<p>wherein R<sup>1</sup> is H, C<sub>1</sub>-C<sub>3</sub> alkyl, C<sub>3</sub>-C<sub>5</sub> cycloalkyl, C<sub>1</sub>-C<sub>3</sub> perfluoroalkyl; R<sup>2</sup> is H, C<sub>1</sub>-C<sub>6</sub> alkyl substituted by OH, C<sub>1</sub>-C<sub>3</sub> alkoxy or C<sub>3</sub>-C<sub>6</sub> perfluoroalkyl; R<sup>3</sup> is C<sub>1</sub>-C<sub>6</sub> alkyl, C<sub>3</sub>-C<sub>6</sub> cycloalkyl, C<sub>3</sub>-C<sub>7</sub> cycloalkyl, C<sub>1</sub>-C<sub>6</sub> alkyl or (C<sub>3</sub>-C<sub>6</sub> cycloalkyl)C<sub>1</sub>-C<sub>6</sub> alkyl; R<sup>4</sup> together with the nitrogen atom to which it is attached completes a pyrrolidinyl, piperidino, morpholino, 4-N-(R<sup>6</sup>)-piperazinyl group; R<sup>5</sup> is H, C<sub>1</sub>-C<sub>3</sub> alkyl, C<sub>1</sub>-C<sub>3</sub> alkoxy, NR<sup>7</sup>R<sup>8</sup>, or CONR<sup>7</sup>R<sup>8</sup>; R<sup>6</sup> is alkyl, (C<sub>1</sub>-C<sub>3</sub> alkoxy) C<sub>2</sub>-C<sub>6</sub> alkyl, hydroxyalkyl, (R<sup>7</sup>R<sup>8</sup>)C<sub>2</sub>-C<sub>6</sub> alkyl, (R<sup>7</sup>R<sup>8</sup>NCOC<sub>1</sub>-C<sub>6</sub> alkyl, CONR<sup>7</sup>R<sup>8</sup>, CSNR<sup>7</sup>R<sup>8</sup> or C(NH)NR<sup>7</sup>R<sup>8</sup>; R<sup>7</sup>, R<sup>8</sup> each independently H, C<sub>1</sub>-C<sub>4</sub> alkyl, (C<sub>1</sub>-C<sub>3</sub>) C<sub>2</sub>-C<sub>4</sub> alkyl or hydroxy C<sub>2</sub>-C<sub>4</sub> alkyl; and pharmaceutically acceptable salts thereof, are selective cGMP inhibitors useful in the treatment of cardiovascular disorders such as angina, hypertension, heart failure, and atherosclerosis.</p>
		
		<b>Chemical structure for Sildenafil (Viagra)</b>
		<b>8 Claims, No Drawings</b>

Figure 2: Original Markush patent assigned to Pfizer that covers the blockbuster drug Viagra (left) and actual chemical structure (right). The Markush structure is shown in the abstract with the R-groups.

Even though, composition of matter or formulation patents that cover the actual drug compound (e.g. Sildenafil) can be used to represent a firm's drug efforts, they do not completely represent the origin or original range of compounds claimed. Since, they are filed after the original Markush patent, a granted composition of matter or formulation patent does not capture the correct time of the project's origin. That is, the explorative effort in chemical space or project timing is missed when sampling non-Markush patents.

Importantly, these follow on patents can bias analyses due to endogeneity by representing only the projects that progressed far enough or were successful enough to enter clinical trials and get approved for market use.

Thus, in drug discovery Markush patents are a) specific to only small molecule or chemistry-based drug projects b) mark the origin of the drug discovery project c) capture a region of chemical space the applicant is interested in filing claims for, and d) a Markush patent can lead to multiple follow on patents (republished Markush patents, composition of matter and process patents). These unique properties of Markush patents and applications, and their starting position in terms of timing make them especially useful for analyzing the evolution of technological search in firms. I exploit these unique features of Markush patents to measure longitudinal R&D activity and track search trajectories over time.

### **Data Collection of Markush Patents**

*Data source:* The source for Markush patents is the Chemical Abstract Society's (CAS) Scifinder product. Scifinder is the world's largest repository of chemical structures, published articles and patents and provides access to MARPAT – a comprehensive database of Markush patents that cover all 9 major patent offices and 63 patent authorities worldwide. More than 1 million Markush structures and about 481,000 Markush patents and applications from 1988-present are available for searching. In a recent comparative analysis of patented compound databases, CAS's Scifinder was found to be more

comprehensive and accurate compared to the Derwent World Patents Index and Reaxys (Ede, et al, 2016).

Since MARPAT contains Markush patents that also belong to other chemistry-based industries, I had to sort these patents into drug and non-drug related categories. For this, I consulted database experts at CAS, and mapped out how each Markush patent is curated, categorized and organized. Scifinder database has 80 section codes under the broad categories of Biochemistry, Organic, Macromolecular, Applied and Physical, Inorganic & Analytical. Each new Markush patent is manually analyzed by a trained chemist and assigned to a category using a specific section code. I consulted both CAS and industry experts in medicinal chemistry to help identify and shortlist section codes that were at risk of being drug patents. This exercise resulted in a set of 28 section codes that are drug-related. Full list in Appendix Table A.

Using the Scifinder web-based search interface, I restricted search to only 28 specific section codes (e.g. Pharmacology, Heterocyclic compounds, Pharmaceuticals, etc.) and manually downloaded the Markush records. This strategy captured drug patents and applications while eliminating non-drug chemistry areas like Ceramics, Dyes, Fuels and Materials Science.

For the study, I am interested in capturing effects of the human genome on drug discovery and needed a sample set that included pre-and post-release of the map (between 1998-2000). Though the Human Genome Project was started in 1990, gene sequences were not released publicly until 1998. In June 2000, the first complete draft of the human genome was made available. Hence, I restricted the search for patent

applications filed between 1990 and 2004 to cover a window of time broad enough to measure pre-post effects of the map. This resulted in a dataset of 88,160 Markush patent records, comprised of 29,309 granted patents and 58,851 patent applications.

### Sample selection

The dataset is comprised of WIPO patent applications (38,672), US patents and patent applications (12,656) and EP patents and patent applications (10,055) accounting for 70% of all Markush records collected between 1990-2004. Patents are associated with a distinct code appended to their identifiers, known as patent kind codes which identifies their status such as A, A1, A2, B1, B2. For example, the code A1 indicates patent application publication, while A2 indicates patent application republication. 'A' indicates a US grant date (replaced by B1 and B2 since 2001). Thus, a Markush patent application (first to be filed), can be refiled or result in a granted patent.

For example, the Markush patent US5137884A (A indicating granted status) assigned to Merck & Co has a grant date of 1992-08-11. But the original Markush application is actually US06552570 filed in 1983-11-16, almost a decade earlier. In addition, the granted patent could differ in content (due to republication, modifications) compared to the original Markush patent filed. Also, only a small proportion of all Markush applications become granted patents – most are rejected, discontinued or abandoned projects. Thus, if Merck had invested in 10 drug discovery projects leading up to a Markush application stage in 1983 and only 3 got granted status, we would be

missing the remaining 7 projects. Thus, sampling only granted patents would not be reliable.

Markush patent applications comprise 68% of the downloaded dataset (patents + applications). A breakup of this based on priority application country is shown in Figure 3. Patents filed in the World Intellectual Property Organization (WIPO) patent office form the majority of patents (74%), followed by the European Patent Office (17%) and United States Patent & Trademark Office (8%). Under a revised law in all US patent applications filed on or after November 29, 2000 were published within 18 months. Prior to this they were filed in the patent office but not available or disclosed to others. Corroborating with changes in the patent offices, in my sample I observed that US applications are under-represented as they were not published prior to 2000. Thus, including US patent applications would represent an unbalanced dataset and skew representation post-2000.

European patents comprise about 16% of the patent applications, but are not well represented post-1995. This could possibly be due to applicants switching to WIPO applications for broad coverage instead of filing EP patents. The WIPO's Patent Cooperation Treaty (PCT) assists applicants in seeking international patent protection for their inventions. That is, by filing one international patent application in the WIPO office, applicants can simultaneously seek protection for an invention in a large number of countries. In 2018, this included 152 PCT countries. This wide international coverage also implies that WO applications by applicants are also more likely to be important and

broader as they are intended for multiple markets, and may therefore also be used to keep out smaller firms or competitors in other countries. Thus, the WIPO patent applications in the sample represent a strong and reliable measure of firm R&D activity in small-molecule drug discovery.

For cross-sectional analyses, including both EP and US patent applications with the WIPO patent applications would be appropriate depending on the selected time period. But for the longitudinal analysis in this study covering 15 years, I risk under-representing the US and EP Markush patents for the entire time period. For these reasons, I have selected only the WIPO or WO Markush patent applications for empirical analysis in the time period 1990-2003. There are no granted WIPO Markush patents in the dataset. After eliminating redundant records based on unique priority application number, a total of 38, 673 WIPO Markush patent applications were obtained. A breakdown of the assignees and priority application countries are shown in Figure 4 below.

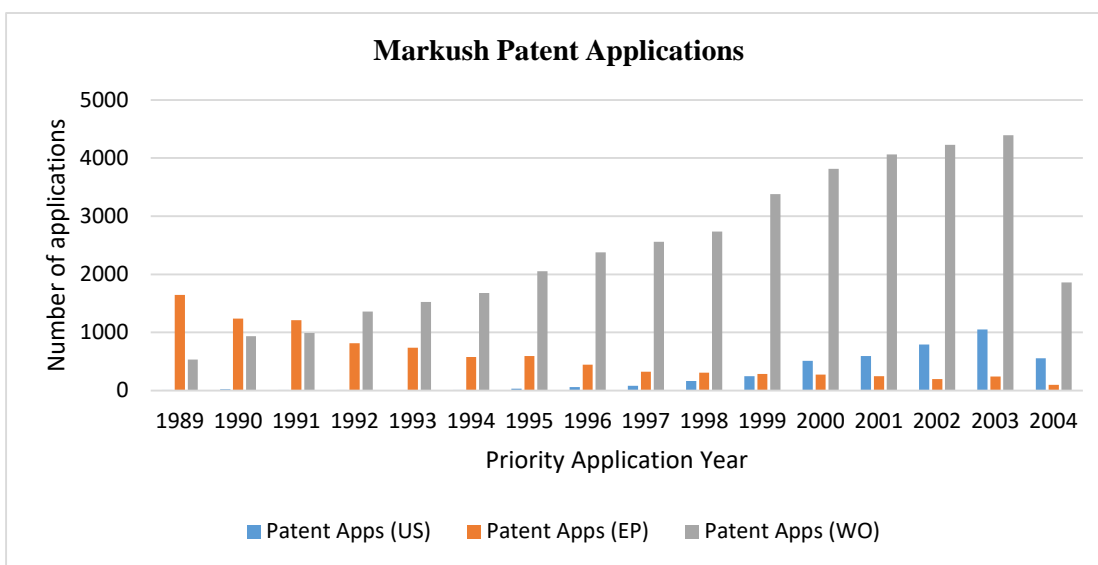


Figure 3: Applications separated by the top three patent offices – WIPO (WO), EP and US

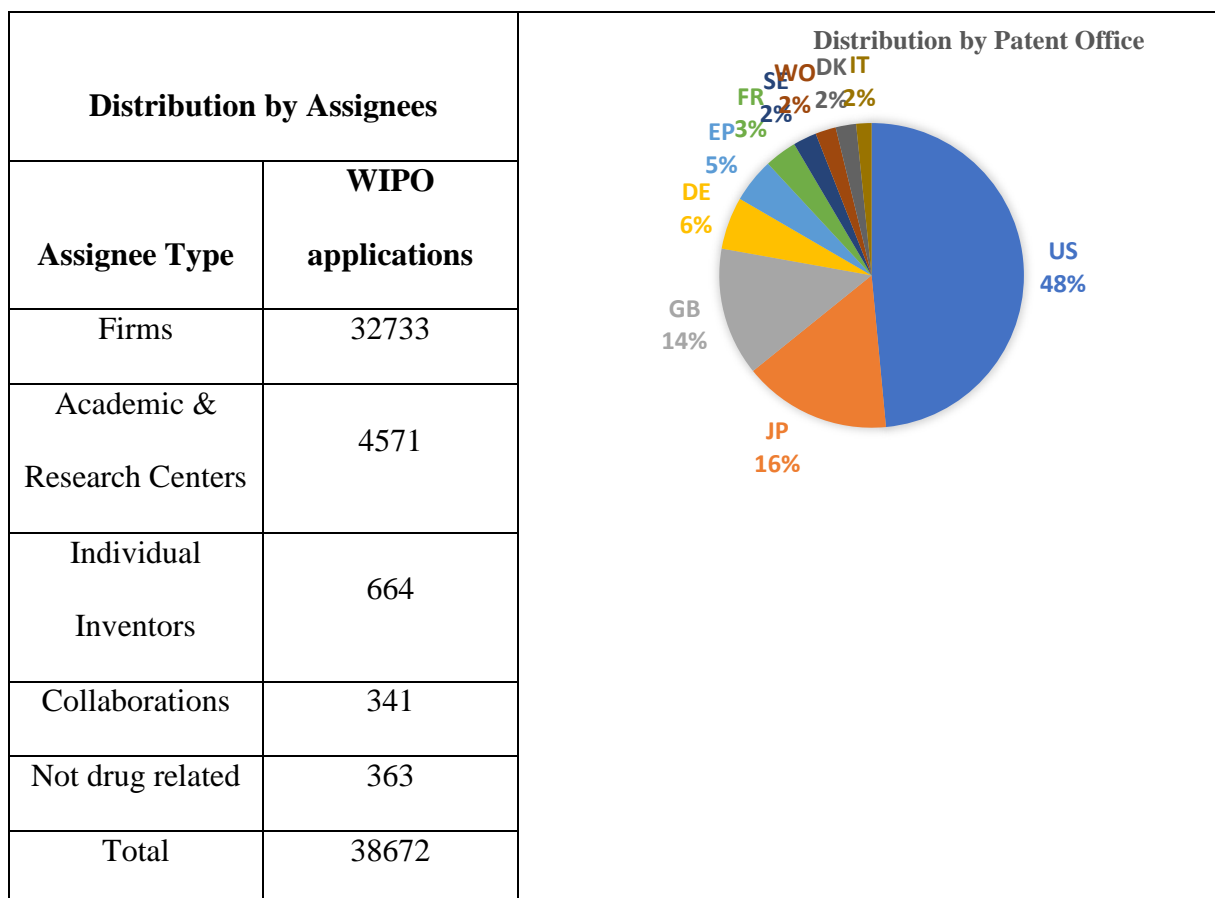


Figure 4: Distribution of drug patents in sample after selecting only WIPO patent applications

A sample CAS record and fields in a Markush patent are provided in the Appendix. Using a standard text parser written in Python, each of the patent fields was extracted and stored in a MySQL relational database for querying.

*Exemplified Compounds:* Exemplified compounds, also known as prophetic substances, are the novel compounds that are listed in the Markush patent records obtained from Scifinder. The Scifinder database extracts the novel compounds registered in the Markush patent application and assigns a unique CAS identification number to each compound (e.g. the CAS Registry Number for Viagra is 139755-83-2). From the WIPO patent applications records that were downloaded, 8,217,103 exemplified compounds and unique identifiers were extracted. Using these CAS numbers, I create a count of novel compounds reported for each patent.

*Patent Assignees:* Assignees were extracted from patent records and stored separately. Assignees have a name and location associated with them. Each assignee is categorized into a separate category: firms, universities, medical centers and collaborations. This classification process yielded 4486 firms, 1500 universities and medical centers. For the analysis, only firm assigned Markush patent applications were selected and does not include any collaborations between firms or between academia-industry.

### Identification of Drug Discovery Strategies

I interviewed discovery scientists to understand what drug discovery strategies firms and how to identify them using patent records. Target-based patents in their description would normally name a disease target gene or gene symbol in addition to mentioning specific keywords that indicated the search strategy used to make the compounds. Figure 3 below shows the abstract of a Markush patent where a molecular target (COX-2,

cyclooxygenase-2), specific keywords related to target-based approach (inhibiting activity) and disease indications (inflammation) can be obtained by reading the abstract.

To identify the presence of targeted strategies, I developed a text-based classification method based on patent text. Initial versions of this method included the text of the title, abstract and description section of the patent. But this yielded false positives, as inventors and examiners would reference other patents, titles or references that could contain a gene name, but not be used in the discovery approach. Hence, a non-targeted approach could be mis-classified as a target-based strategy. To overcome this problem, I compared and contrasted the merits of using abstract versus description before settling on just using the patent title and abstract that captured the main aspects of the patent and its strategy. In addition, CAS provides two human curated fields known as Index Terms and Supplementary Terms – these are short keyword summaries of the patent provided by internal knowledge experts. The text from these fields is then compared to gene names and gene symbols derived from two sources: the National Center for Biotechnology Information (NCBI) database and the HUGO Gene Nomenclature Committee's (HGNC) database. The HGNC is a committee of the Human Genome Organization that sets standards for naming genes and assigning gene symbols. For example, the drug target for Viagra is the enzyme phosphodiesterase 5A, written using the symbol PDE5A or PDE5. A patent could use any of these versions of the gene name or symbol, hence, the algorithm should be able to account for this variation.

To counter this problem, the gene database from the National Center for Biotechnology Information (NCBI) was also used as it provides a list of gene synonyms and old names that were previously used. Thus, together a comprehensive collection of 61,561 gene symbols and identifiers were created. Using this combined data, patent abstracts, titles and curated sections from Scifinder were scanned using custom software developed for this purpose.

United States Patent		4,693,008
Ando, et al.		August 5, 2003
Sulfamoylheteroaryl pyrazole compounds as anti-inflammatory/analgesic agents		
Abstract		
This invention relates to a compound of the formula: #STR1## or a pharmaceutically acceptable salt thereof, wherein A and R, esp. 1 are each an optionally substituted 5 to 6-membered heteroaryl, wherein the heteroaryl is optionally fused to a carbocyclic ring or 5 to 6-heteroaryl; R, esp. 2 is NH, sub 2; R, esp. 3 and R, esp. 4 are each hydrogen, halo, (C, sub 1 -C, sub 4)alkyl optionally substituted with halo and the like; and X, sub 1 to X, sub 4 are each hydrogen, halo, hydroxy; (C, sub 1 -C, sub 4)alkyl optionally substituted with halo and the like. These compounds have <a href="#">FIG. 5</a> inhibiting activity and thus useful for treating or preventing inflammation or other <a href="#">FIG. 6</a> related diseases.		
Inventors:	Ando; Kenzo (Aichi-Ken, JP), Kawamura; Kiyoshi (Aichi, JP)	
Assignee:	Pfizer Inc. (New York, NY)	
Family ID:	22613376	
Appl. No.:	09/729,661	
Filed:	November 28, 2000	

Figure 5: Sample patent abstract from United States Patent & Trademark Office database is analyzed using custom algorithms to identify target-based keywords, gene target and disease indications and classified as using a target-based drug discovery strategy.

A custom algorithm was implemented in the Python language using the open source Natural Language Toolkit library to mine the text of all the Markush patents in the sample. In addition to the gene identifiers, a set of keywords gathered from interviews indicating a target-based approach were also screened. These words include versions of relevant keywords like 'gene', 'genes', 'genomic', 'genomics', 'receptor', 'receptors', 'inhibit', 'inhibitor', 'inhibitors', 'target', 'targeted'. Wild-card matching (e.g. inhibit\*) was implemented to account for these variants.

The text mining and classification using this approach led to identification of gene names and symbols appearing patent titles and abstracts. False positives can occur when

the algorithm tags a patent as target-based when the actual gene symbol could be used in a different context. For example, the valid gene symbols ‘CAT’, ‘MICE’ and ‘PIGS’ can be misconstrued as gene names when used in the context of compounds being tested for toxicity on these animals. Or the gene ‘BOC’ has a different connotation when used in a chemistry patent. In chemistry, BOC groups refer to *tert*-butyloxycarbonyl, a protecting group in organic synthesis. To ensure reliability, I tag false positives using a special list of such terms (see Appendix) and manually inspect patents to ensure that the algorithm driven classification is accurate and false positives eliminated. This sorting created two broad categories of patents based on their search strategy: 60% non-target based and 40% target-based.

## **Research Design**

### ***Dependent variable***

*Number of exemplified (novel) compounds:* Each novel compound listed in a Markush patent is an original and tangible chemical entity with a unique chemical structure. These chemical compounds are structurally based on the Markush structure, and required to support the invention and substantiate patent claims. Patents that do not have broad claims and considered weak have a stronger likelihood of being challenged or broken by competitors (Hemphill & Sampat, 2011). Hence, a set of exemplified compounds in a patent indicate the firm’s search activity in chemical space and intent to protect a region of chemical space. Novel compounds arise as a result of knowledge recombination, search strategy and explorative intensity.

The number of compounds encoded in Markush patents is then a factor of search effort, dedicated resources and strategic intent of the firm and inventors (Drews, 2000; DiMasi et al, 2010). And, more compounds in a patent imply the intensity of research efforts, knowledge flows and specialized skills in the firm (Jaffe & Trajtenberg, 1999; Alacacer & Gittelman, 2006; Henderson & Cockburn, 1996). Thus, number of patented compounds provide a quantitative representation of the intensity of search effort and suited as a measure for analysis of innovation novelty.

### **Difference-in-Differences Estimation**

Difference-in-Differences (DID) estimations are used to study the role of interventions like economic policies and events in natural experiments. This statistical technique analyzes the differential effect of the intervention or treatment on a treatment group and compares it to a control group that is not exposed to the treatment (Angrist & Krueger, 2001; Imbens & Woolbridge, 2007). A common experimental setup is observing two groups over two-time periods, where the first group is exposed to treatment in the second-time period, while the control group is not exposed to the treatment the entire period. To eliminate biases in second time period comparisons due to trends or inherent differences, the average gains in the control group is subtracted from the average gain in the treatment group (Card & Krueger, 1994).

In this study, the dependent variable (exemplified compounds) is affected by two well technological factors – the availability of combinatorial chemistry and the human genome. While combinatorial chemistry preceded the arrival of the map, it is difficult to

pinpoint its exact arrival or separate out only those patents that used it for making compounds (Asgari et al, 2016; Persidis, 1998). In contrast the public release of human genome DNA sequences is well documented. In 1998, only about 8% of the genome was publicly available, but within two years 90% of the map was made public in 2000 (Collins, et al, 1998). The first complete draft of the human genome was completed and announced on June 26, 2000. Thus, the partial map of the human genome was available 1998 onwards, while the complete draft appeared in 2000. These two-time points can be used as intervention periods in the estimation.

To estimate only the effect of the human genome on small molecule drug discovery using a difference-in-differences approach, a control set of non-drug patents was required that was also impacted by combinatorial chemistry but not by the human genome. I surveyed other chemistry industries that were also impacted by combinatorial chemistry and found two fields: materials science and agrochemicals.

The agrochemical industry is involved in making plant-related products like fertilizers, herbicides and insecticides. The products they produce can also include small-molecules and utilize similar search processes. While agrochemicals had a number of similarities to the drug industry, it did not satisfy an important criterion – not being impacted by scientific maps. Private and public plant genome mapping efforts were already in place during the time period of interest and could influence the direction of search in the agrochemical industry (Waterhouse & Helliwell, 2003).

In contrast, the materials science industry was only influenced by combinatorial chemistry technologies but not plant or human genomes. After detailed comparisons and analysis of patent data, I chose Materials Science patents as a control group for the difference-in-differences estimation.

*Combinatorial chemistry in Materials Science:* Since the early 1990's combinatorial chemistry methods have been used in drug discovery (Ellman, et al, 1997). During this time, a parallel adoption of the technique took place in materials science. Materials science is a broad interdisciplinary area drawing from physics, chemistry, engineering, biology, medicine and nanotechnology, and is used in the development of new materials for electronics, energy systems, aerospace, nanotechnology, industrial chemicals, polymers and catalysts. Adoption of combinatorial chemistry in non-drug industries quickly followed:

"Now this idea, known as "combinatorial chemistry" ... is spreading outside medicine. The electronics industry, for example, thrives on new materials with exotic properties, such as emitting light of a certain colour when pumped with electricity (electroluminescence), or conducting electricity without resistance (superconductivity). But these materials are usually compounds of several elements arranged in complicated crystalline structures. ... researchers have just published the results of their attempts to make this technique work." - Combinatorial Chemistry Material Gains, The Economist, 1998.

"Pioneered by the pharmaceutical industry and adapted for the purposes of materials science and engineering, the combinatorial approach represents a watershed in the process of accelerated discovery, development and optimization of materials. To survey large compositional landscapes rapidly, thousands of compositionally varying samples may be synthesized, processed and screened in a single experiment." - Koinuma & Takeuchi, 2004

Adoption of combinatorial chemistry in materials science research led to new types of compounds and associated materials for electroluminescence, photoluminescence and semi-conductors (Koinuma & Takeuchi, 2004; Economist, 1998). New techniques were developed for manufacturing advanced inorganic materials like optical, glass, dielectric and magnetic materials, catalytic powders, polymers and biofunctional materials (Resetar & Eiseman, 2001; Takeuchi et al, 2005). Based on sampling and suitability for difference-in-differences analysis, materials science broad chemistry patents were selected as a control group. Using similar techniques that I had used to separate out only drug patents, I collected materials science patents using relevant CAS category codes. From this I selected patents filed only in the United States. See Figure 6 and Appendix showing comparative growth and adoption of combinatorial chemistry in materials science and drug discovery research patenting and publications.

To ensure that the differences-in-differences estimation is robust and the comparison of drug patents is made to a relevant control group, I collected and tested patents from the field of agribiotech and materials science. Agribiotech was also influenced by combinatorial chemistry, but it was unsuitable as control since the agribiotech sector was also being influenced by the availability of plant genomes. Comparatively, materials science as an industry was going through similar changes as the drug industry and adopting combinatorial chemistry in its product development cycles (Appendix Figures C-E). Compound growth trends for both drug industry and material sciences is shown in Figure 7. There are remarkable similarities in compound growth rates over a 15-year period, with drug patents having much higher compound production.

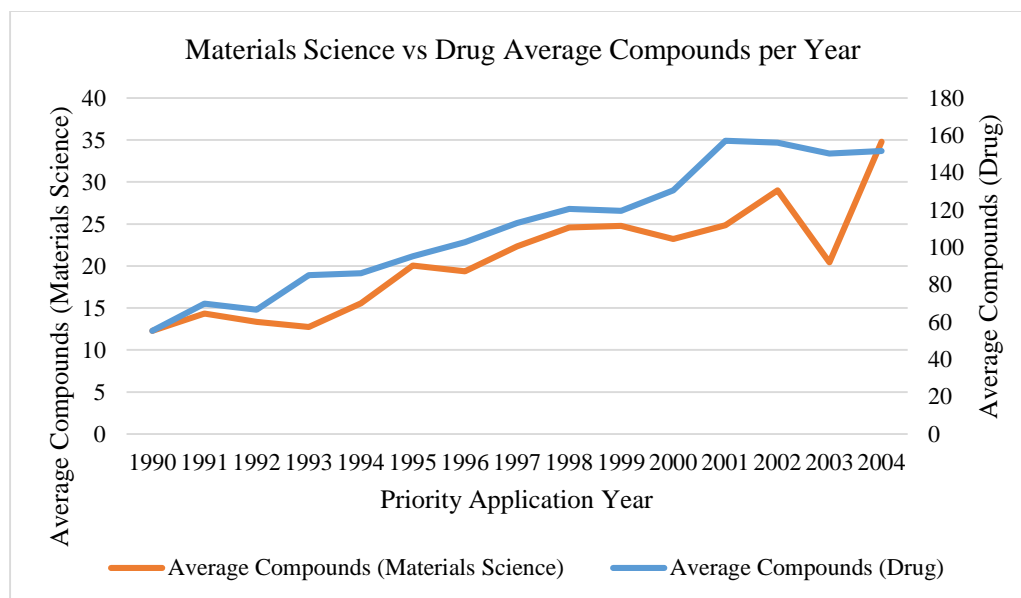


Figure 6: Comparison of novel compound production derived from Markush patent applications for treatment (drug) and control group (materials science).

*Data Matching for Difference-in-Differences Estimation:* For the difference-in-differences estimation, a focused subset of the drug patents is used. Countries can differ in access to new technologies (like combinatorial chemistry, high-speed computers), scientific resources and manpower, and organizational cultures. To control for these exogenous factors that could bias the analysis, only WIPO patent applications filed in the United States are selected. The United States accounts for nearly half of all the applications filed world-wide, and is suitable for this selection. Similarly, the control group is also comprised of only materials science patents filed in the United States. I restricted both control and treatment samples to the same country to ensure reduce unexplained effects due to inter-country variation in firm R&D. Combinatorial chemistry was a novel technique with specialized supporting infrastructure like robotics and high-throughput screening. The United States is also the largest market for chemicals and

firms interested in this market would typically file patents here (Ahuja & Lampert, 2001). Restricting the sample to the United States maintains consistency, reliability and comparability in factors like accessibility to combinatorial chemistry technologies and resources needed for technological search.

Treatment group: 14,677 US small-molecule drug firm patent applications (1990-2005)

Control group: 4,313 US materials-science firm patents and applications (1990-2005)

Intervention Period: 1998. See genome sequence release dates in Appendix Figure A.

### **Model**

The generic difference-in-differences (DID) model is specified as:

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2.dB + u$$

where  $y$  is the outcome variable (compounds),  $d2$  is a dummy variable for the second time period (post-Human Genome Map), the dummy variable  $dB$  captures differences between treatment and control groups before the second time period (treatment). The coefficient of interest is  $\delta_1 d2.dB$ . The difference-in-differences estimate is,

$$\tilde{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1})$$

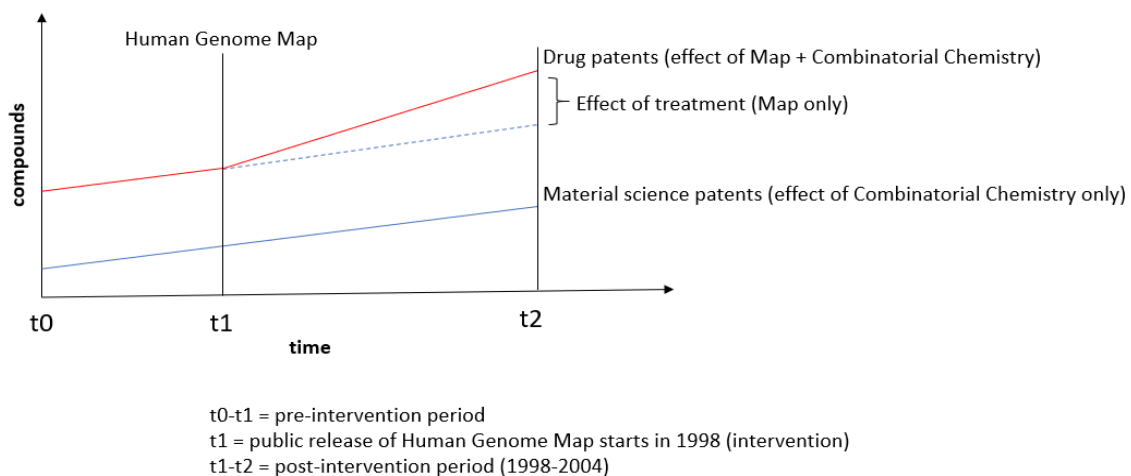


Figure 7: Experimental setup to test the effect of the human genome and combinatorial chemistry on novel compound production. In the post-intervention time period, the separate effect of the map is calculated by subtracting out the effect of only combinatorial chemistry.

Therefore, Estimation of HGP Map Effect = (Drug compounds – Material Science compounds)<sub>post-Map</sub> – (Drug compounds – Material Science compounds)<sub>pre-Map</sub>

## Results

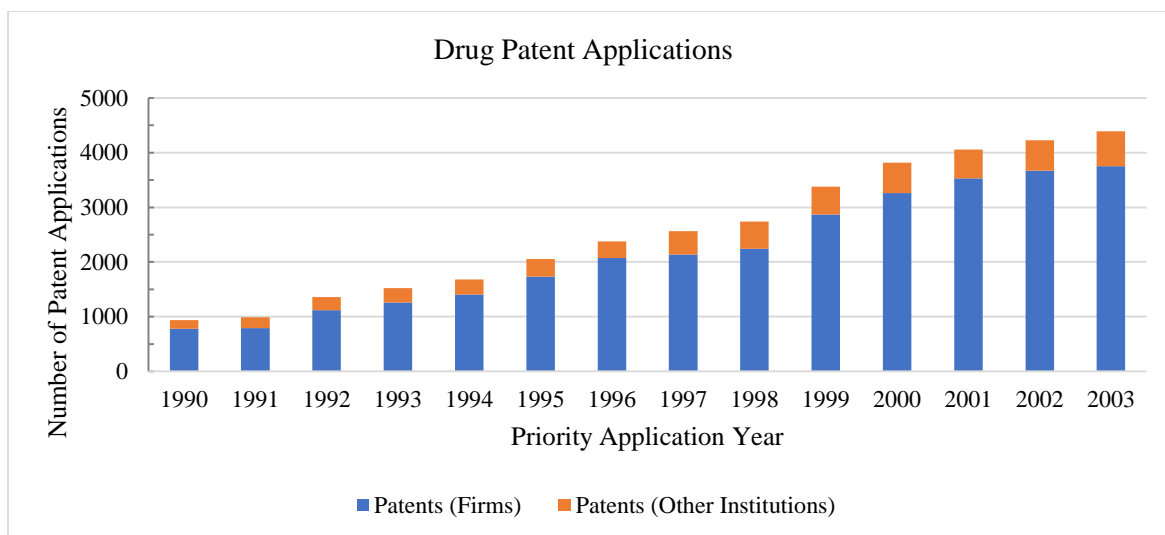


Figure 8: Industry patenting increases 60% between 1998-2003

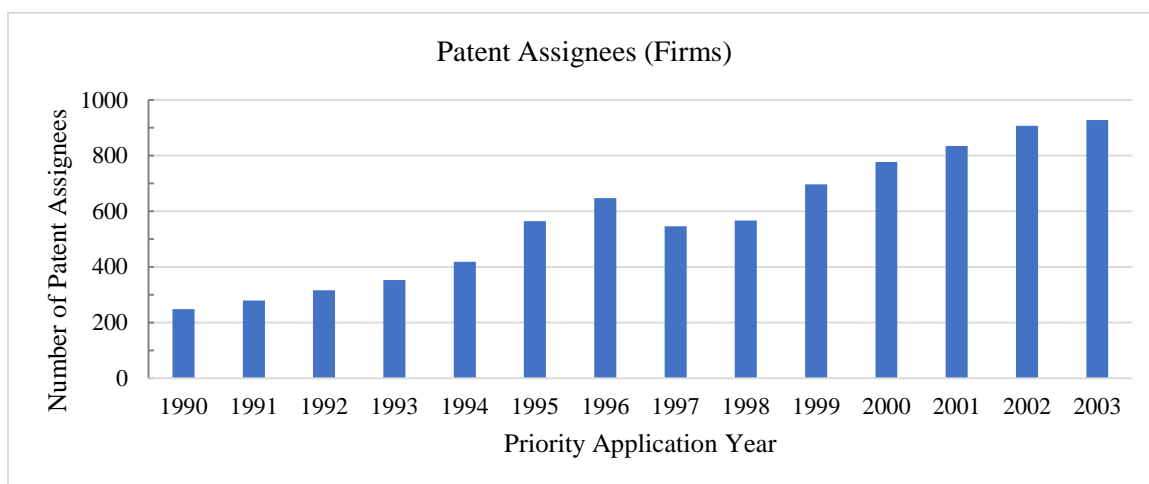


Figure 9: Number of firms patenting per year increases 70% between 1998-2003

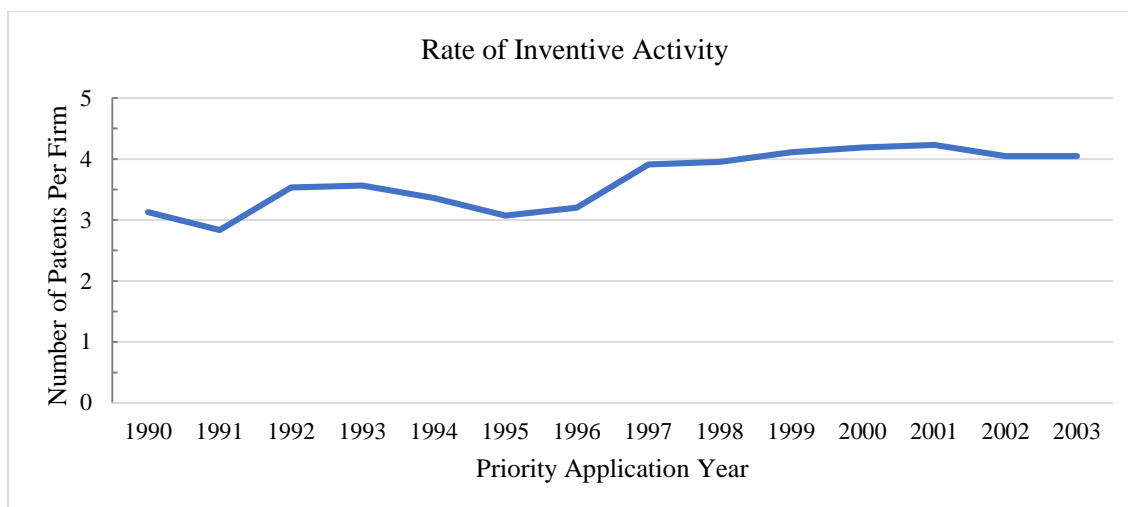


Figure 10: Average number of patent applications filed per firm increases 33% after 1998.

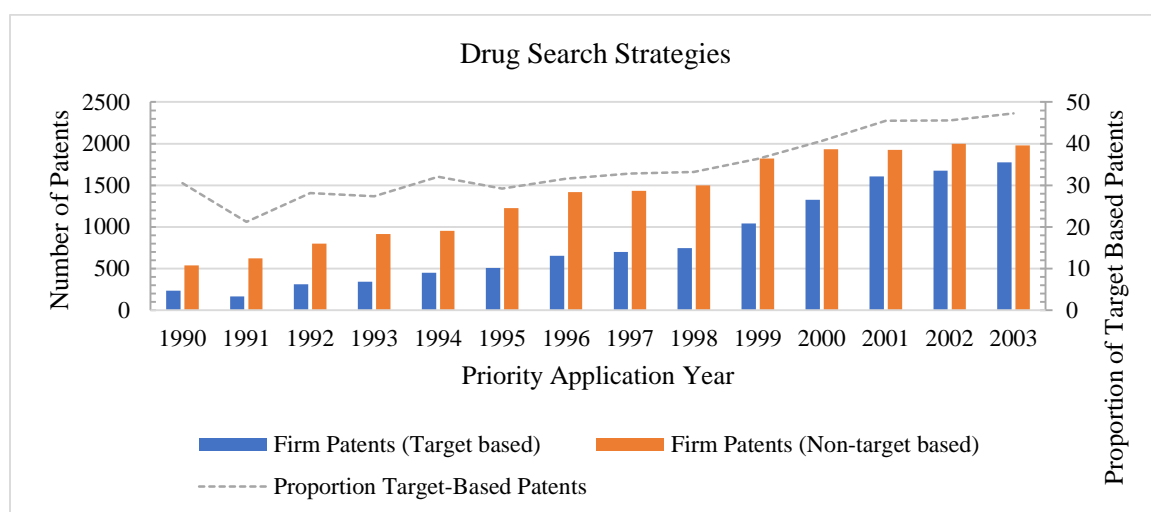


Figure 11: Targeted strategies increase from ~30% of total projects in the early 1990's to ~50% in 2003.

Changes in firm patenting activity are shown in Figures 8-11. Figure 8 shows a 60% increase in firm patenting between 1998-2003, the period when more than 90% of the human genome sequences was released (see Appendix A). In Figure 9, number of firms between 1998-2003 go up by 70% indicating an increase in firms engaged in drug

discovery. To capture the rate of inventive activity among drug firms, I calculated an average score (total patents per year/total unique firms per year) and for each year. Compared to gross yearly patent counts that do not account for firm-level heterogeneity in patenting, this measure provides an average estimate of overall industry inventive activity across time. The trend line is shown in Figure 10. I observe a 33% increase over the 14-year time period with firm inventive activity being the highest post-1998.

Changes in drug discovery strategies are shown in Figure 11. Firm patents are clustered based on their search strategies (target or non-target based) and plotted along with the target-to-non-target proportion over time. An industry-wide shift towards target-based strategies is shown (almost 90% by 2003) with a nearly 50% increase in target-based strategies between 1998-2003. In sum, post human genome I observe more patenting activity, more number of firms engaged in drug discovery, drastic change in drug search strategies and overall inventive activity increasing across the industry (Figures 8-11), supporting Hypothesis 1a and 1b.

Gene Name	Gene Symbol or keywords	Therapeutic Area	Patent Applications citing gene (pre-map)	Patents Applications citing gene (post-map)	Firms (pre-map)	Firms (post-map)
tumor necrosis factor receptor	tnf	Oncology	174	291	67	96

cyclooxygenase	cox	Inflammation & Pain	123	281	61	104
mitogen-activated protein (MAP) kinases	p38	Autoimmune	57	140	26	42
peroxisome proliferator-activated receptors	ppar	Metabolic diseases (like Diabetes)	22	128	11	54
vascular endothelial growth factor	vegf	Oncology, Ophthalmology	7	104	2	40
mitogen-activated	map	Autoimmune	11	74	8	36

protein (MAP) kinases						
Dipeptidyl peptidase-4	dpp- 4/dpp-IV	Oncol ogy, Viral	3	66	1	31

Table 1: Competition and crowding around top gene targets: List of small-molecule patents and drug firms citing top gene targets before-after the human genome map.

Genentech's drug Avastin is a VEGF inhibitor that was approved in 2004 and had sales of \$7 billion in 2015. Tnf is the second most studied gene in the human genome (Dolgin, 2017).

Table 1 shows a list of top gene targets that experienced rapid increase in patenting and in number of firms working on them. In the post-map period, I see more firms and patents around the same gene targets indicating intense competition for disease markets in the industry. These point to an increase in patenting around similar disease targets. Thus, with more firms, more patent applications and crowding around genes in the drug industry the search intensity increases, supporting Hypothesis 2.

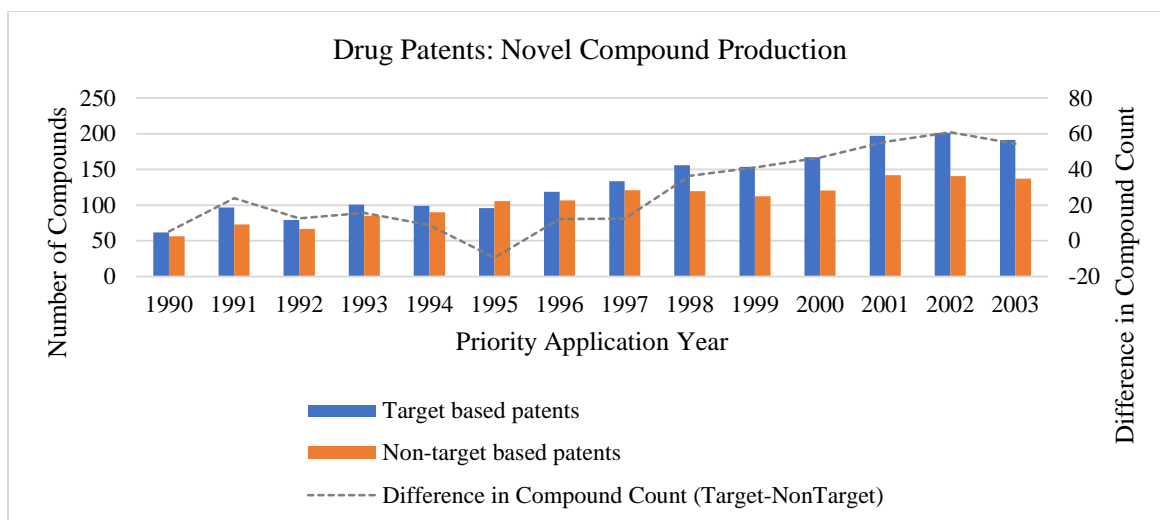


Figure 12: Novel compounds for target and non-target patents, where the gap increases after 1998.

Figure 12 captures an interesting trend – novel compounds are consistently higher for target-based search strategies, and the gap between target non-target strategies widens after the release of gene sequences. Search strategies have differential compound yield and the difference in novel compounds patented is amplified with the availability of a scientific map. There appears to be a specific advantage or characteristic that target based strategies have over non-target based approaches in producing novel compounds.

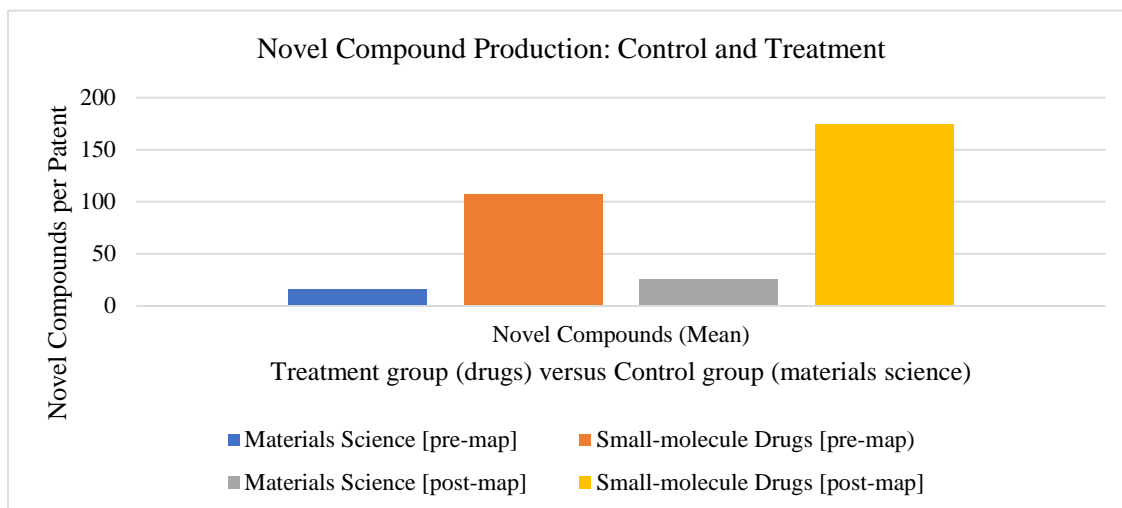


Figure 13: Novel compounds increase 63% for drugs and 56% for materials science from 1998-2004 (post-map). Mean value of compounds for drug patents pre-map is 107 and post-map is 174.

Drug Patent Applications (Treatment)		Materials Science Patents (Control)
Patent Priority Application Country	United States	United States
Time Period	1990-2005	1990-2005
Observations	14,677	2,811
Compound Output	Mean: 136.21	Mean: 18.42
	Min: 0	Min: 1
	Max: 8916	Max: 757
	Std Dev: 324.70	Std Dev: 26.28

Table 2: Descriptive statistics for treatment and control groups

HGP Map Treatment Period	1998-2004		2000-2004	
	(1)	(2)	(3)	(4)
VARIABLES <sup>a</sup>	Model 1:	Model 2:	Model 3:	Model 4:
	OLS Fixed	OLS	OLS Fixed	OLS
	Effects		Effects	
DtrXDpost	68.80*** (10.10)	58.12*** (9.864)		
Dpost	8.294** (3.496)	9.090*** (1.872)		
Dtr	31.61 (33.98)	91.00*** (10.48)	43.92 (33.51)	101.0*** (11.09)
DtrXDpost2			63.27*** (11.46)	57.28*** (12.34)
Dpost2			8.095** (4.079)	7.438*** (2.472)
Constant	56.73** (26.00)	16.03*** (0.873)	59.13** (25.84)	17.90*** (0.928)
Observations	20,340	20,340	20,340	20,340
R-squared	0.009	0.040	0.008	0.041
Number of orgnamecode	3,054		3,054	

Robust standard errors in parentheses. Standard errors clustered at firm level.

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

<sup>a</sup>Legend:

Dtr = Dummy for treatment (Drug patents)

Dpost = Dummy for Post-intervention period (1 if year between 1998-2004)

DtrXDpost = Interaction term for Treatment and Post-intervention period 1998-2004  
(Diff-in-Diff estimator)

Dpost2 = Dummy for Post-intervention period (1 if year between 2000-2004)

DtrXDpost2 = Interaction term for Treatment and Post-intervention period 2000-2004  
(Diff-in-Diff estimator)

Table 3: Difference-in-differences estimate for the effect of the human genome map on novel compound production. Treatment group is US drug and control is US materials science patents and applications.

Table 2 and 3 show descriptive statistics and results for the Difference-in-Differences estimation.

In Model 1, the difference-in-differences estimator (DtrXDpost) is calculated using Ordinary Least Squares (OLS) regression with firm fixed effects and the standard errors clustered for robustness at firm level for the intervention time period 1998-2004. The difference-in-difference estimate for the effect of only the human genome map is 68.80. To check whether these effects are consistent even after the arrival of the complete human genome map in 2000, Model 2 measures the effect for the 2000-2004 time period. Results are similar and statistically significant.

This implies that controlling for the independent effect of combinatorial chemistry (using materials science patents), the genome map accounted for 69 more compounds for each patent. This result is statistically significant and consistent with Model 2 and 4 (run without fixed effects). This implies that the availability of human genome map increased compound output by 69 compounds per patent in the 1998-2004 intervention period, and by 63 compounds per patent with the arrival of the complete human genome (2000-2004). These results support Hypothesis 2.

The DID estimator in all four regression models are positive and statistically significant at  $p < 0.01$ . The difference-in-differences estimation (Model 1) indicates a 64% increase in novel compound production across the industry attributed to the effect of the human genome.

### **Robustness Checks**

Our analysis of small-molecule patents shows a transition from non-target to target-based approaches in the mid-to-late 1990's. This observation in my data (Figure 11) is supported by a number of reports in drug discovery and management research that also point to a shift in search strategies in the 1990's marking the entry of the genomics era (Scannell, et al, 2012; Drews, 2000; Triggle, 2006; Zucker & Darby, 2002; Lipinski & Hopkins, 2004; Gittelman, 2016). The change in internal search processes has been attributed to the availability of high-throughput technologies and genomics-driven search paradigms.

“The 1990s saw a major shift in small-molecule drug discovery strategies, from iterative low-throughput in vivo screening and medicinal chemistry optimization to target-based high-throughput screening (HTS) of large compound libraries. ... the former is slow and expensive in terms of the number of compounds that can be tested, whereas the latter is fast and cheap.” – Scannell, et al, 2016

Given that the uncertainty in drug discovery and pharma industry extremely competitive, it is expected that drug firms will try to capture as much chemical space as possible since it is not clear which of these compounds will do well in preclinical and clinical trials. This could explain the increase in novel compound production with the arrival of the human genome map.

I have tried to rule out alternate explanations for the increase in compound production in drug patents like combinatorial chemistry (see Discussion), R&D expenditures and outsourcing of R&D projects. One possible rationale for increased compound production could be higher R&D investments, which would mean more resources for each project. This could be translated into more manpower, funding or technologies to make new compounds. Appendix Figure F shows R&D spending by drug and chemical companies from 1990-2000. The trends do not show a major increase in R&D spending after mid-1990's or after 1998 that could account for the increase in compound counts I observe in the period after the human genome was made public.

Difference-in-difference estimates are run with clustered standard errors and firm fixed effects. Clustered standard errors provide a way to get unbiased standard errors of OLS coefficients and account for heteroscedasticity. Standard errors are clustered at the firm level and firm fixed effects are used in the OLS regression to account for omitted

variable bias. In this panel, longitudinal observations are captured for the same firms and compound means are being measured for each observation, hence, the firm fixed effects model is suitable for this analysis.

To rule out any specific year's influence on the overall difference-in-differences trends, robustness tests were carried out by dropping out high-influence years like 2003. In figure 6, the gap between average compounds for materials science and drugs is maximum due to the sharp drop in materials science compounds. To ensure that this increased does not bias the difference-in-differences (DID) estimator I ran the DID estimation for the entire time period except observations for the year 2003. The DID estimator for this robustness checking model (excluding year 2003) was 65 compounds, consistent with the main model results and these were significant at  $p < 0.001$ . This robustness check model is included in the Appendix Table D.

*Value and Importance of Markush Patents in Sample:* Firms could be patenting large portions of chemical space using the Markush patents to fence out competitors or claim important space. And it could be argued that firms could be blocking out regions of space without really meaning to pursue those compounds into marketable drugs. To test whether the patent applications in my sample was not just placeholder patents in an inter-firm fencing strategy, I had to check whether these patents were important to the firm and industry. To do this, I screened all the patents in the data sample against the Organization for Economic Cooperation and Development (OECD) Triadic patent database. This database captures triadic patent families – defined as a set of patents registered in the

three major patent offices: United States Patent and Trademark Office (USPTO), the European Patent Office (EPO) and the Japan Patent Office (JPO).

Filing patents in all three offices indicates high-value and intent to take the drug to market in economically important markets. By screening the World Intellectual Property Office (WIPO) patent applications against this OECD Triadic database, I could identify those patents that were part of patent triads filed in USPTO, EPO and JPO. I found that about 38% (37.78) of the patent applications in the sample are part of the OECD Triadic database families and are economically important to the firm. This implies that a significant portion of the patents in this study are economically important and indicate a focused approach by the firms to strategically capture chemical space and fence out smaller firms and competitors.

*Effect of Drug Discovery Strategy on Novel Compound Production:* There is also an underlying factor that could contribute to the increase in compound production – the search strategies involved. Figure 14 shows differences in compound output based on search strategies.

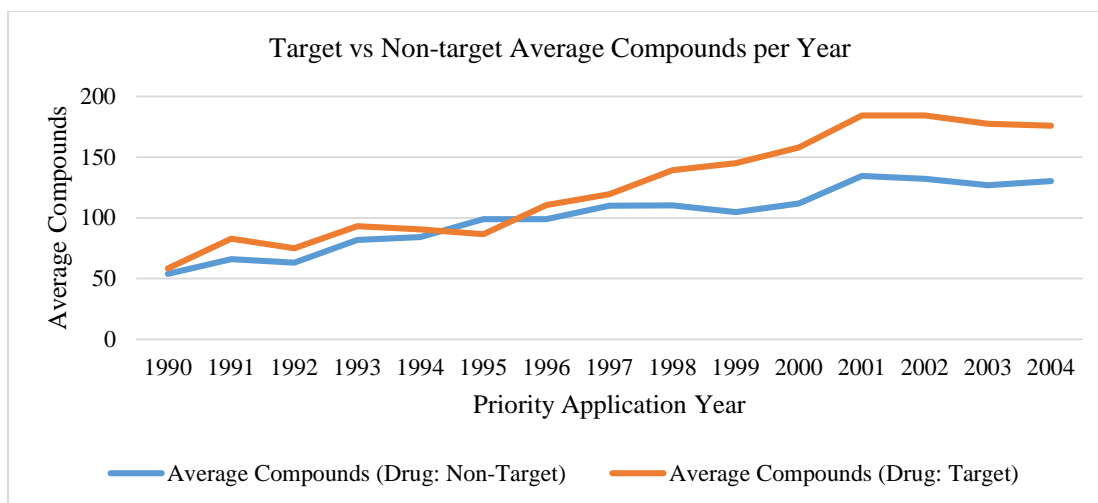


Figure 14: Comparison of novel compound production for the two drug discovery strategies

Markush patent applications that indicate a target-based strategy have generally higher compound output compared to the non-targeted approach. This gap in novel compound production increases considerably post-1997. To estimate, the effect of just target-based strategies on novel compound production, I used the same differences-in-differences estimation approach for the same time period, but only on the Markush drug patent applications filed in the United States. Similar to the main model, I used two intervention time periods to test the impact of the map on drug discovery strategies. The target-based patents are the treatment group as their approach benefits most with the availability of the human genome. The non-target-based patents act as a control. Both groups are equally impacted by combinatorial chemistry and the standard errors of the difference-in-differences estimation is clustered at the firm-level. This controls for within firm differences in use of technologies needed for drug discovery. Results are shown in Table 4.

HGP Map Treatment Period	1998-2004		2000-2004	
	(1)	(2)	(3)	(4)
VARIABLES <sup>b</sup>	Model 1:	Model 2:	Model 3:	Model 4:
	OLS Fixed	OLS	OLS Fixed	OLS
	Effects		Effects	
TgtXDpost	41.23** (18.10)	60.54*** (15.18)		
Dpost	13.92 (25.13)	-21.99 (18.90)		
is_target	15.78 (12.72)	5.526 (9.622)	17.92** (8.390)	14.02* (7.801)
TgtXDpost2			49.26*** (15.27)	60.66*** (14.92)
Dpost2			-6.354 (17.87)	-27.58 (21.31)
Constant	80.18*** (18.57)	99.69*** (14.07)	92.44*** (11.14)	100.1*** (15.58)
Observations	15,602	15,602	15,602	15,602
R-squared	0.012	0.015	0.012	0.015
Number of orgnamecode	2,247		2,247	

Robust standard errors in parentheses. Standard errors clustered at firm level.

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

<sup>bs</sup>Legend:

Is\_target = Dummy for treatment (Target-based drug patents)

Dpost = Dummy for Post-intervention period (1 if year between 1998-2004)

TgtXDpost = Interaction term for Treatment and Post-intervention period 1998-2004  
(Diff-in-Diff estimator)

Dpost2 = Dummy for Post-intervention period (1 if year between 2000-2004)

TgtXDpost2 = Interaction term for Treatment and Post-intervention period 2000-2004  
(Diff-in-Diff estimator)

Table 4: Difference-in-differences estimate for the effect of the human genome map on novel compound production based on drug discovery strategies. Treatment group is target-based patent applications and control is non-target based patent applications.

Difference-in-differences estimation shows that the human genome map increases novel compound production in target-based strategies by 34% in the 1998-2004 time period, and by 38% in the time period when the complete map is available (2000-2004). Thus, the map impacts targeted strategies more and this is evidenced by a higher production of novel compounds in those Markush patents.

## Discussion and Conclusions

Using a novel dataset of chemistry drug patent applications, I find that the human genome map is correlated with an increased rate in drug patenting, novel compounds and

a change in discovery strategies. These findings point to the role of scientific maps as catalysts in knowledge recombination, focusing search processes and enabling firms to explore more of the technological landscape.

The economic and social cost of being unable to develop adequate drugs for unmet medical needs is high. This is further compounded by not having enough research funding for basic science projects, which is required to understand the science behind diseases and their mechanisms. Despite considerable pushback, the Human Genome Project – a basic science project was launched with public-funding in 1990 with much skepticism about its impact on biology and disease. By the time the project was completed, new firms were launched to exploit the genome, large amounts of investor funds were pumped into the biotechnology industry and academic scientist-entrepreneur led firms became more common (Zucker & Darby, 1996). While most industry and academic articles report an overall decline in new drug development in the post-genome era, there has been no systematic empirical analysis of how the processes of drug discovery changed with the advent of the human genome map (Scannell, 2012; Gittelman, 2016). In this study, I open the black box of drug innovation using a novel and focused data set of early stage drug patents that have not been used previously in management research.

The human genome provided a precise map of gene targets that could be used to predict outcomes and focus the search for new drugs – complementing the search process. Combinatorial chemistry represented a substitute technology that could replace the skills and tacit knowledge of experienced medicinal chemists which was a prized

commodity of larger pharmaceutical firms. I proposed that the map facilitated focused search, while the new tools made compound production more efficient - resulting in more inventive activity and adoption of targeted strategies. I also predicted that the open accessibility of the map combined with low cost screening technologies, would lead to patent races prompting firms to capture valuable chemical space and intensify intellectual property protection. My results support these hypotheses (1a, 1b and 2).

These results also support prior studies on the role of mapping and science as a map. By showing a significant effect on the production of compounds in the time period 1998-2000 when only a portion of the map was available, this results empirically support Puranam & Swamy's (2010) recent theoretical models on the role of incomplete maps on exploration and innovation. The significant effect of the map on novel compounds in the time periods (2000-2004) when the complete map was available, and the increase in production seen in targeted strategies support prior work on the idea that science acts as map in accelerating the rate and intensity of technological search (Fleming & Sorenson, 2004).

The human genome was a technological disruption that impacted the biopharmaceutical industry. Availability of a precise scientific map and genetic targets allowed firms to sharpen or enhance their search efforts. In this regard, the map acts as a form of complementary asset that allowed firms to map out their search strategies in advance and create search efficiencies in the drug discovery process (Teece, 1986). Having a target to focus their search compared to working through trial-and-error was itself a paradigm shift

in firms' modes of drug discovery. But the map also presents a dilemma that firms have to respond to. The public nature of the map meant that competitors and new entrants also had access to this search-enhancing tool and could potentially lead them to valuable new targets and compounds. The lowering of entry barriers to new scientific knowledge provided by the map could potentially create a race like condition to capture chemical space around new disease targets.

In parallel, the adoption and wide spread availability of combinatorial chemistry created an opportunity to make novel and more compounds at low cost. Combinatorial chemistry technology in effect was a substitute to the resource intensive manual process of making novel compounds using medicinal chemists. The ability to assemble in-house combinatorial chemistry capabilities using off-shelf robotic components and vendor libraries had a major impact on drug discovery groups across the industry. In the late 1990's firms were replacing synthetic and medicinal chemists with combinatorial chemistry tools. Industry experts we interviewed had accounts of chemistry teams being replaced with automated combinatorial chemistry workflows during this time period. The important change during this period was the lowering of the cost per molecule achieved due to combinatorial chemistry.

Thus, a dual force was at play here: publicly available search enhancing-capability provided by the map and lowered cost of compound production due to combinatorial chemistry. This coupled with the lowered barriers to entry created race-like conditions leading to an increase in novel compound production per patent as shown in our results.

The separate effect of the map (60+ compounds) per patent is supported by the difference-in-difference estimations controlling for just the effect of combinatorial chemistry.

From a policy perspective, this result is interesting as the map shows a stronger and separate effect on new compound production compared to just the productivity gains achieved by combinatorial chemistry. This is due to the map opening up new territories in chemical space that was previously unexplored. Thus, public investment in basic science projects like the Human Genome led to more exploration of chemical space, increased novel compound output per patent and created increased market competition – in essence, creating conditions ripe for technological innovation.

With a number of scientific mapping projects in progress, understanding their economic impact and how they influence innovation processes is important for public policy, R&D management and firm strategy. By linking a basic science project with industry innovation, this empirical analysis sheds new light on important but largely unexplored questions in the literature on technological search. A scientific map by providing accurate coordinates of the solution landscape makes the search process less undirected and random, but also increases crowding. These results reveal an interesting finding: scientific maps not only increase the rate of invention, but also impact the nature of technological search and underlying strategies, suggesting that scientific mapping projects can enhance the quality of industry innovation.

## References

Ahuja, Gautam, and Curba Morris Lampert. "Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions." *Strategic management journal* 22.6?7 (2001): 521-543.

Alcacer, Juan, and Michelle Gittelman. "Patent citations as a measure of knowledge flows: The influence of examiner citations." *The Review of Economics and Statistics* 88.4 (2006): 774-779.

Angrist, Joshua, and Alan B. Krueger. *Instrumental variables and the search for identification: From supply and demand to natural experiments*. No. w8456. National Bureau of Economic Research, 2001.

Arora, Ashish, and Alfonso Gambardella. "The changing technology of technological change: general and abstract knowledge and the division of innovative labour." *Research policy* 23.5 (1994): 523-532.

Asgari, Navid, Kulwant Singh, and Will Mitchell. "Alliance portfolio reconfiguration following a technological discontinuity." *Strategic Management Journal* 38.5 (2017): 1062-1081.

Azoulay, Pierre, et al. *Public R&D investments and private-sector patenting: evidence from NIH funding rules*. No. w20889. National Bureau of Economic Research, 2015.

Card, David, and Alan B. Krueger. "Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: reply." *The American Economic Review* 90.5 (2000): 1397-1420.

Cohen, Wesley M., and Daniel A. Levinthal. "Absorptive capacity: A new perspective on learning and innovation." *Administrative science quarterly* (1990): 128-152.

Collins, Francis S., et al. "New goals for the US human genome project: 1998-2003."

Science 282.5389 (1998): 682-689.

Combinatorial Chemistry Material Gains, The Economist, March 12, 1998

Darby, Michael R., and Lynne G. Zucker. Going public when you can in biotechnology.

No. w8954. National Bureau of Economic Research, 2002.

Derek Lowe, How do you find a New Compound to Patent?, Science Translational

Medicine, Dec 10, 2015

DiMasi, Joseph A., et al. "Trends in risks associated with new drug development: success

rates for investigational drugs." Clinical Pharmacology & Therapeutics 87.3 (2010): 272-

277.

Dolgin, 2017, "The most popular genes in the human genome", Nature News Feature,

November, 2017

Drews, Jürgen. "Drug discovery: a historical perspective." Science 287.5460 (2000):

1960-1964.

Ellman, Jonathan, Barry Stoddard, and Jim Wells. "Combinatorial thinking in chemistry

and biology." Proceedings of the National Academy of Sciences 94.7 (1997): 2779-2782.

Fleming, Lee, and Olav Sorenson. "Science as a map in technological search." Strategic

Management Journal 25.8?9 (2004): 909-928.

Fleming, Lee, and Olav Sorenson. "Technology as a complex adaptive system: evidence

from patent data." Research policy 30.7 (2001): 1019-1039.

Fleming, Lee. "Recombinant uncertainty in technological search." Management science

47.1 (2001): 117-132.

Gittelman, Michelle. "The revolution re-visited: Clinical and genetics research paradigms and the productivity paradox in drug discovery." *Research Policy* (2016).

Gunther McGrath, Rita, and Atul Nerkar. "Real options reasoning and a new look at the R&D investment strategies of pharmaceutical firms." *Strategic Management Journal* 25.1 (2004): 1-21.

Hemphill, C. Scott, and Bhaven N. Sampat. "Evergreening, patent challenges, and effective market life in pharmaceuticals." *Journal of health economics* 31.2 (2012): 327-339.

Hemphill, C. Scott, and Bhaven N. Sampat. "When do generics challenge drug patents?." *Journal of Empirical Legal Studies* 8.4 (2011): 613-649.

Huang, Kenneth G., and Fiona E. Murray. "Entrepreneurial experiments in science policy: Analyzing the Human Genome Project." *Research Policy* 39.5 (2010): 567-582.

Imbens, Guido, and Jeffrey Wooldridge. "what's new in econometrics" lecture 5. NBER Summer Institute, 2007.

Jaffe, Adam B. "Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value." (1986).

Jaffe, Adam B., and Manuel Trajtenberg. "International knowledge flows: evidence from patent citations." *Economics of Innovation and New Technology* 8.1-2 (1999): 105-136.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. "Geographic localization of knowledge spillovers as evidenced by patent citations." *the Quarterly journal of Economics* 108.3 (1993): 577-598.

Johnson, M. A., and G. M. Maggiora. "Concepts and Applications of Molecular Similarity, John Wiley & Sons." (1992).

Kauffman, Stuart, José Lobo, and William G. Macready. "Optimal search on a technology landscape." *Journal of Economic Behavior & Organization* 43.2 (2000): 141-166.

Kitch, Edmund W. "The nature and function of the patent system." *The Journal of Law and Economics* 20.2 (1977): 265-290.

Koinuma, Hideomi, and Ichiro Takeuchi. "Combinatorial solid-state chemistry of inorganic materials." *Nature materials* 3.7 (2004): 429.

Langdon, Sarah R., Nathan Brown, and Julian Blagg. "Scaffold diversity of exemplified medicinal chemistry space." *Journal of chemical information and modeling* 51.9 (2011): 2174-2185.

Laursen, Keld, and Ammon Salter. "Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms." *Strategic management journal* 27.2 (2006): 131-150.

Lipinski, Christopher, and Andrew Hopkins. "Navigating chemical space for biology and medicine." *Nature* 432.7019 (2004): 855-861.

March, James G. "Exploration and exploitation in organizational learning." *Organization science* 2.1 (1991): 71-87.

Nagaraj, A. (2015, November 6). The Private Impact of Public Maps— Landsat Satellite Imagery and Gold Exploration. Job Market Paper, MIT Sloan School of Management, Cambridge, MA. Available: [http://web.mit.edu/nagaraj/files/nagaraj\\_jmp\\_nov6.pdf](http://web.mit.edu/nagaraj/files/nagaraj_jmp_nov6.pdf)  
National Academies of Sciences, Engineering, and Medicine. 2016. Advancing Concepts and Models for Measuring Innovation: Proceedings of a Workshop. Washington, DC: The National Academies Press. doi: 10.17226/23640.

Nelson, Richard R., and G. Sidney. "Winter. 1982." An evolutionary theory of economic change (2005): 929-964.

Nightingale, Paul. "A cognitive model of innovation." Research policy 27.7 (1998): 689-709.

Ouellette, Lisa Larrimore. "How many patents does it take to make a drug? Follow-on pharmaceutical patents and university licensing." Michigan Telecommunications and Technology Law Review 17 (2010): 299.

Persidis, Aris. "Combinatorial chemistry." Nature biotechnology 16.7 (1998): 691-693.

Pharmaceutical Research & Manufacturers of America (PhRMA) Industry, 2016.

<http://www.phrma.org/report/industry-profile-2016>

Puranam, Phanish, and Murali Swamy. "Expeditions without Maps: Why Faulty Initial Representations May Be Useful in Joint Discovery Problems." Available at SSRN 1153142 (2010).

Resetar, Susan, and Elisa Eiseman. Anticipating technological change: combinatorial chemistry and the environment. No. RAND/MR-1394.0-EPA. RAND CORP SANTA MONICA CA, 2001.

Rivkin, Jan W. "Imitation of complex strategies." Management science 46.6 (2000): 824-844.

Rosenberg, Nathan. "Why do firms do basic research (with their own money)?." Research policy 19.2 (1990): 165-174.

Sampat, Bhaven N. "Mission-oriented biomedical research at the NIH." Research Policy 41.10 (2012): 1729-1741.

- Sampat, Bhaven, and Heidi L. Williams. How do patents affect follow-on innovation? Evidence from the human genome. No. w21666. National Bureau of Economic Research, 2015.
- Scannell, Jack W., et al. "Diagnosing the decline in pharmaceutical R&D efficiency." *Nature reviews Drug discovery* 11.3 (2012): 191-200.
- Tripp, Simon, and Martin Grueber. "Economic impact of the human genome project." Battelle Memorial Institute (2011).
- Southall, Noel T., and Ajay. "Kinase patent space visualization using chemical replacements." *Journal of medicinal chemistry* 49.6 (2006): 2103-2109.
- Takeuchi, Ichiro, Jochen Lauterbach, and Michael J. Fasolka. "Combinatorial materials synthesis." *Materials today* 8.10 (2005): 18-26.
- Teece, David J. "Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy." *Research policy* 15.6 (1986): 285-305.
- Triggle, David J. "Drug discovery and delivery in the 21st century." *Medical Principles and Practice* 16.1 (2006): 1-14.
- Tripsas, Mary. "Unraveling the process of creative destruction: Complementary assets and incumbent survival in the typesetter industry." *Strategic Management Journal* (1997): 119-142.
- Ulrich, Karl. "The role of product architecture in the manufacturing firm." *Research policy* 24.3 (1995): 419-440.
- Vincenti, Waler G. "What Engineers Know and How They Know it Analytical Studies from Aeronautical History." (1990).
- Waterhouse, Peter M., and Christopher A. Helliwell. "Exploring plant genomes by RNA-induced gene silencing." *Nature Reviews Genetics* 4.1 (2003): 29.

Williams, Heidi L. "Intellectual property rights and innovation: Evidence from the human genome." *Journal of Political Economy* 121.1 (2013): 1-27.

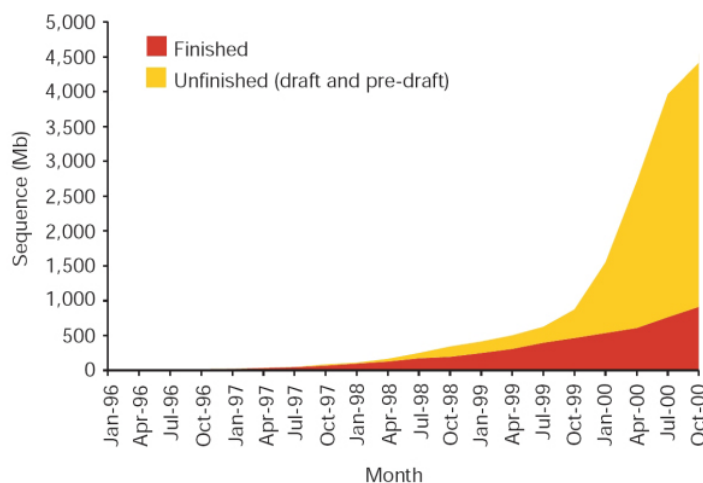
Wood, Denis. *Rethinking the power of maps*. Guilford Press, 2010.

Ziedonis, Rosemarie Ham. "Don't fence me in: Fragmented markets for technology and the patent acquisition strategies of firms." *Management science* 50.6 (2004): 804-820

Zucker, Lynne G., and Michael R. Darby. "Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry." *Proceedings of the National Academy of Sciences* 93.23 (1996): 12709-12716.

## APPENDIX

**Figure A:** Release of the human genome sequences in Megabases (Y axis) – the standard measure of genome segments. While 6% of genome was available in 1998, the complete draft released in 2000 covered more than 90% (Collins, et al. Science, 1998). Difference between the draft-finished genome versions is defined by coverage, number of gaps and error rate. Image: Lander, et al, Nature, 2001



**Figure B:** Cost comparison of traditional versus combinatorial chemistry (source: Persidis, 1998)

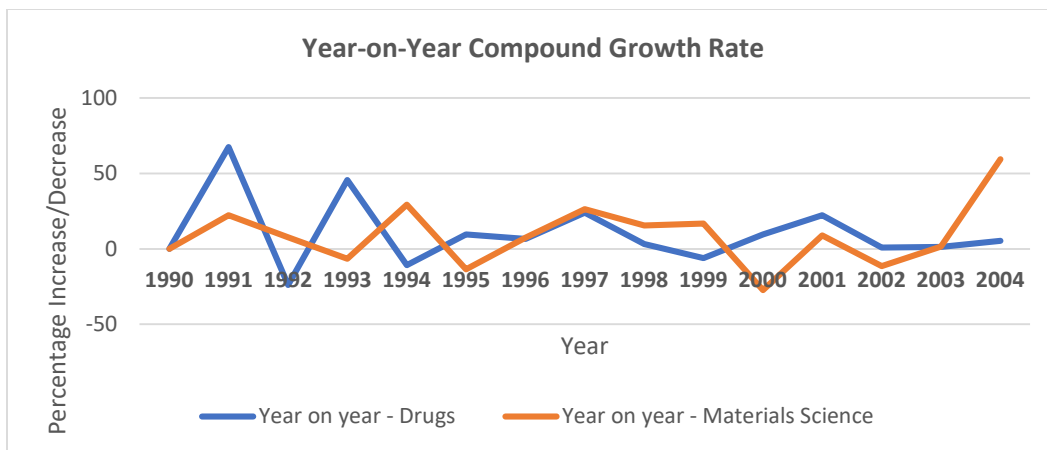
Table 1. The power of combinatorial chemistry		
	Traditional chemistry	Combinatorial chemistry
Compounds per one chemist month	4	3,300
Total cost	\$30,000	\$40,000
Cost per compound	\$7,500	\$12
Source: Booz, Allen & Hamilton 1996 Survey		

Richard A. Houghten, president of Torrey Pines Institute for Molecular Studies, San Diego, and one of the pioneers of the field says: " It's really quite spectacular how quickly it's become almost an accepted routine. Whereas 10 years ago a good medicinal chemist might make 50 to 100 compounds a year, that same chemist is probably expected to make in the thousands or tens of thousands today."

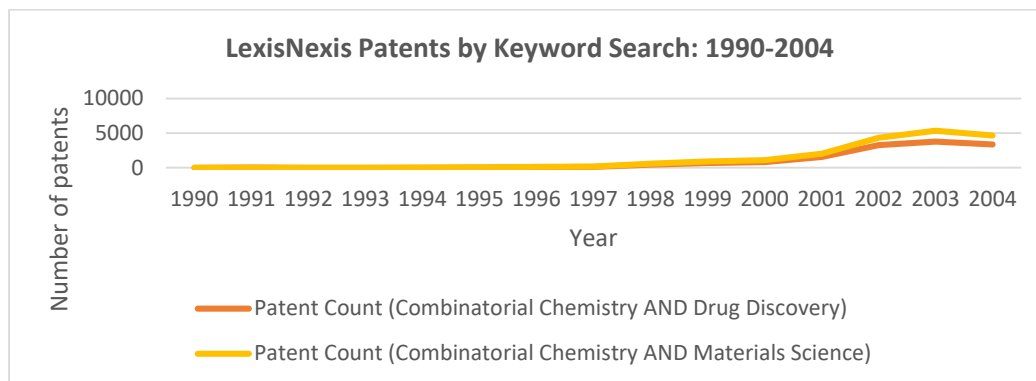
Peter L. Myers, chief scientific officer of CombiChem, San Diego, agrees. "The pharmaceutical industry, in particular, has embraced it totally. Every company now has some aspect of this in-house." – Chemical & Engineering News, April 6, 1998

**Figure C:** Year-on-year compound growth rate for material science and drug patents

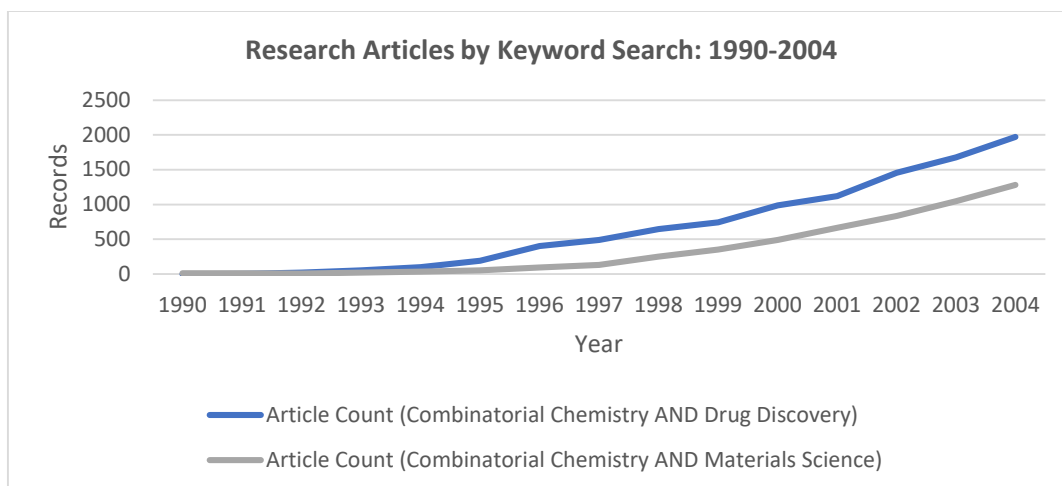
follow similar trends indicating related adoption of combinatorial chemistry



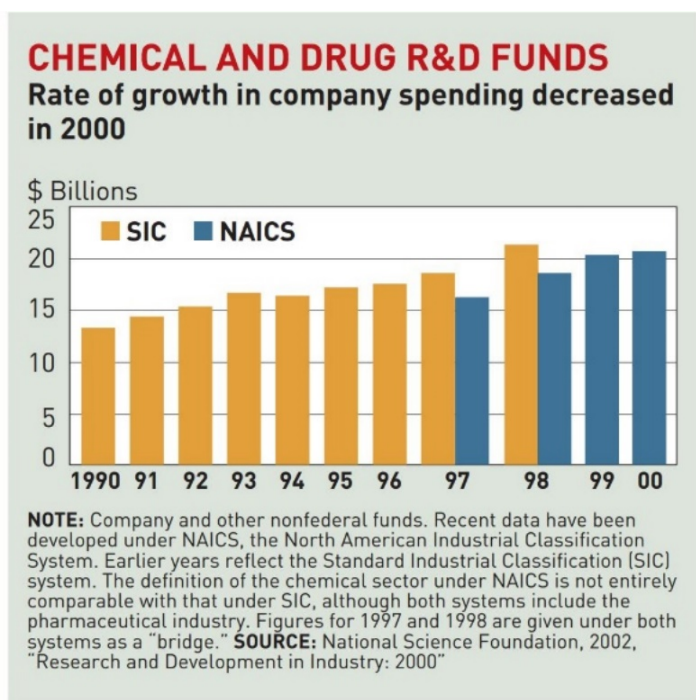
**Figure D:** Comparison of patents found in LexisNexis for material science and drug discovery



**Figure E:** Comparison of scientific publications for material science and drug discovery (source: Scopus)



**Figure F:** R&D investment from 1990-2000 (American Chemical Society report, Chemical & Engineering News, 2008)



**Table A:** Curated list of Drug-related categories used to search Scifinder

CA Section Title	

Heterocyclic Compounds (More Than One Hetero Atom)	
Heterocyclic Compounds (One Hetero Atom)	
Benzene, Its Derivatives, and Condensed Benzenoid Compounds	
Pharmacology	
Amino Acids, Peptides, and Proteins	
Pharmaceuticals	
Aliphatic Compounds	
Carbohydrates	
Organometallic and Organometalloidal Compounds	
Biomolecules and Their Synthetic Analogs	
Industrial Organic Chemicals, Leather, Fats, and Waxes	
Fermentation and Bioindustrial Chemistry	
Biochemical Methods	
Steroids	
Alicyclic Compounds	
Terpenes and Terpenoids	
Alkaloids	
Enzymes	
General Organic Chemistry	
Biochemical Genetics	
Radiation Biochemistry	
Mammalian Hormones	
General Biochemistry	

Toxicology	
Microbial, Algal, and Fungal Biochemistry	
Mammalian Biochemistry	
Microbial Biochemistry	
Immunochemistry	

**Table B:** Sample Markush record obtained from the CAS Scifinder database for this study.

<p>START_RECORD</p> <p>FIELD Copyright:Copyright (C) 2017 American Chemical Society (ACS). All Rights Reserved.</p> <p>FIELD Database:CAPLUS</p> <p>FIELD Title:Preparation of 1-[(indolylazacycloalkyl)alkyl]-2,1,3-benzothiadiazole 2,2-dioxides exhibiting 5-HT<sub>2A</sub> receptor activity.</p> <p>FIELD Accession Number:AN 2000:592717</p> <p>FIELD Abstract: The title compds. [I; m = 0-2; n = 1-2; p = 1-2; q = 1-6; r = 0-3; R1 = halo, alkyl, CN, etc.; R2, R3 = H, alkyl; R4 = H, alkyl, (un)substituted Ph, phenylalkyl; R5 = alkyl, alkoxy, CO<sub>2</sub>H, etc.] and their salts which are active at 5-HT<sub>2A</sub> receptor, were prepd. and formulated. E.g., a synthesis of II.HCl which showed K<sub>i</sub> of &lt; 15 nM against 5-HT<sub>2A</sub> receptor binding, was given. [on SciFinder(R)]</p> <p>FIELD Author:</p> <p>FIELD Chemical Abstracts Number(CAN):CAN 133:177178</p> <p>FIELD Section Code:28-10</p>
---

FIELD Section Title:Heterocyclic Compounds (More Than One Hetero Atom)

FIELD CA Section Cross-references:1, 63

FIELD Corporate Source:

FIELD URL:

FIELD Document Type:Patent

FIELD CODEN:PIXXD2

FIELD Internat.Standard Doc. Number:

FIELD Journal Title:PCT Int. Appl.

FIELD Full Journal Title:

FIELD Language:written in English

FIELD Volume:

FIELD Issue:

FIELD Page:31 pp.

FIELD Publication Year:2000

FIELD Publication Date:20000824

FIELD Index Terms:5-HT receptors Role: BSU (Biological study, unclassified), MSC (Miscellaneous), BIOL (Biological study) (5-HT2A; prepn. of 1-[(indolylazacycloalkyl)alkyl]-2,1,3-benzothiadiazole 2,2-dioxides exhibiting 5-HT2A receptor activity)

FIELD Index Terms(2):

FIELD CAS Registry Numbers:288606-13-3P Role: BAC (Biological activity or effector, except adverse), BSU (Biological study, unclassified), SPN (Synthetic preparation), THU (Therapeutic use), BIOL (Biological study), PREP (Preparation),

USES (Uses) (prepn. of 1-[(indolylazacycloalkyl)alkyl]-2,1,3-benzothiadiazole 2,2-dioxides exhibiting 5-HT<sub>2A</sub> receptor activity); 4760-34-3; 120192-70-3; 180161-14-2; 288606-14-4; 709046-15-1 Role: RCT (Reactant), RACT (Reactant or reagent) (prepn. of 1-[(indolylazacycloalkyl)alkyl]-2,1,3-benzothiadiazole 2,2-dioxides exhibiting 5-HT<sub>2A</sub> receptor activity); 399-51-9P (6-Fluoroindole); 443987-59-5P Role: RCT (Reactant), SPN (Synthetic preparation), PREP (Preparation), RACT (Reactant or reagent) (prepn. of 1-[(indolylazacycloalkyl)alkyl]-2,1,3-benzothiadiazole 2,2-dioxides exhibiting 5-HT<sub>2A</sub> receptor activity); 749837-42-1P Role: SPN (Synthetic preparation), PREP (Preparation) (prepn. of 1-[(indolylazacycloalkyl)alkyl]-2,1,3-benzothiadiazole 2,2-dioxides exhibiting 5-HT<sub>2A</sub> receptor activity)

FIELD Supplementary Terms: indolylazacycloalkylalkylbenzothiadiazole dioxide prepn formulation serotonin receptor selective ligand; benzothiadiazole indolylazacycloalkylalkyl dioxide prepn formulation serotonin receptor selective ligand

FIELD PCT Designated States: Designated States W: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, AM, AZ, BY, KG, KZ, MD, RU, TJ, TM.

FIELD PCT Reg. Des. States:Designated States RW: AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, BF, BJ, CF, CG, CI, CM, GA, ML, MR, NE, SN, TD, TG.

FIELD Reg.Pat.Tr.Des.States:

FIELD Main IPC:C07D417-14.

FIELD IPC:

FIELD Secondary IPC:A61K031-41; A61P025-00.

FIELD Additional IPC:

FIELD Index IPC:

FIELD Inventor Name:Fairhurst, John.

FIELD National Patent Classification:

FIELD Patent Application Country:Application: WO

FIELD Patent Application Date:20000211.

FIELD Patent Application Number:2000-GB469

FIELD Patent Assignee:(Eli Lilly and Company Limited, UK).

FIELD Patent Country:WO

FIELD Patent Kind Code:A1

FIELD Patent Number:2000049017

FIELD Priority Application Country:GB

FIELD Priority Application Number:1999-3784

FIELD Priority Application Date:19990218

FIELD Citations:Boehringer Ingelheim Ltd; EP 0058975 A 1982|Eli Lilly And Company Limited; EP 0897921 A 1999|Eli Lilly And Company Ltd; EP 0854146 A

1998|Janssen Pharmaceutica N V; EP 0013612 A 1980|Janssen Pharmaceutica N V; EP  
 0184258 A 1986|Rhone-Poulenc, S; EP 0433149 A 1991|Roussel-Uclaf; FR 2621588  
 A 1989|Takeda Chemical Industries Ltd; WO 9914203 A 1999

FIELD DOI:

END\_RECORD

**Table C:** Gene Symbols that cause False Positives during Classification

ALK

BOC

CR2

C3

C2

C6

C7

CS

NA

AR

MICE

NME2

COMP

WAS

PREP

SPN

PIGS
MAX
MSC
T
REST
IV
CAT
LARGE
SET
PROC
NM
IMPACT
ICOS
HOAC

Table D: Differences-in-Differences Robustness Test (without year 2003). The effect of the map (DtrXDpost) is consistent and significant.

---

VARIABLES	Model without Year 2003
<hr/>	
DtrXDpost	65.06***
	(9.672)
Dpost	8.933**
	(3.823)
Dtr	33.11
	(36.85)
Constant	56.61**
	(28.41)
Observations	17,399
Number of orgnamecode	2,707
R-squared	0.009
<hr/>	
Robust standard errors in parentheses	

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

## **2. SCIENTIFIC MAPS AND INNOVATION STRATEGY: IMPACT OF THE HUMAN GENOME ON THE ADOPTION OF TARGETED STRATEGIES**

### **Abstract**

When new technologies emerge, incumbent firms are faced with decisions regarding the adoption of the new advances into their innovation processes. What drives firm decision-making in this tension between existing capabilities and new market opportunities? Using the context of drug discovery, I analyze how drug firms changed innovation strategies in response to a technological event. The human genome map was a major breakthrough for both the academic community and drug industry. The map increased the number of disease gene targets and made accessible new regions of the genome, opening up new market opportunities. Since the detailed map was beneficial mainly to target-based strategies, firms experienced in this approach had an advantage compared to other firms. To unpack the organizational factors driving adoption, I examine the role of prior knowledge, specialized capabilities and competition. Using a novel dataset of early stage drug projects, I find that the human genome map moderates the impact of these firm-related factors. This study provides novel insights on firm preadaptation, knowledge capabilities and the moderating effect of a scientific map in refocusing innovation strategies.

Keywords: technological change, firm adaptation, strategic refocusing, innovation strategy

## **Introduction**

Firms maintain a fine balance between what they know and what products they make. When technological breakthroughs emerge, firms have to carefully strategize how they design new products and enter new markets. These strategic decisions are driven by prior knowledge of the firm, product portfolios and adaptive organizational capabilities (Zott & Amit, 2007; Helfat & Raubitschek, 2000; Teece, et al, 1997). The process of technological change and its broad impact on markets has been explored by management scholars for a long time but the mechanisms and conditions affecting innovation are not well understood (Utterback, 1994). Innovation scholars suggest that regular experimentation and engagement in learning allow firms to develop absorptive capacities over time, which in turn allows them to adapt during periods of technological change (Nelson & Winter, 1982; Cohen & Levinthal, 1990). Also, by chance some firms due to prior experimentation and explorative activities may be preadapted to exploit emergent technological change (Cattani, 2005; 2006). Over time, firm level differences in capabilities, resources and knowledge due to varying levels of intent and investment form the basis for firm heterogeneity and differential performance (Wernerfelt, 1984). Therefore, what firms know and the capabilities they have evolved play a critical role in understanding how firms adapt during periods of technological change.

Innovation scholars argue that firms develop resilience to technological change as a result of deeply entrenched capabilities that accumulate and coevolve with products and markets (Helfat & Raubitschek, 2000; Zollo & Winter, 2002). Other scholars indicate

that technological change can be Schumpeterian in nature destroying markets and incumbent advantages, instead favoring the competencies of new entrants unencumbered by firm history (Tripsas, 1997; Tripsas & Gavetti, 2000; Christensen, 1993). Hence, a debate exists between the role of path-dependent capabilities of incumbent firms and opportunity seeking nimbleness of entrants in managing technological change. While prior studies have shown the effect of new to market technologies like digital cameras, typesetting and disk drives on incumbent strategies, few studies examine the effect of novel or breakthrough scientific discoveries on innovation. In this study, I address this gap in empirical research by examining how a disruptive technological change influences the role of organizational capabilities and market conditions in determining innovation strategy.

In the early 1990's, the predominant drug discovery strategy was non-target based or phenotypic drug discovery. Biopharmaceutical firms engaged in relatively low levels of target-based strategies as it was resource intensive and required deep background knowledge of gene targets, computational drug design and high-throughput screening capabilities. In contrast, the phenotypic approach based on iterative trial-and-error learning was long established in firms (Gittelman, 2016). At the time, few disease related genes were well studied, hence the range of disease markets that firms could exploit were also limited. Adding to this, private firms and consortiums were patenting new disease related genes causing high-walled barriers to entry. Under these conditions, firms engaged in target-based research were either large firms that were diversified, resource-

rich and had found a niche space, or smaller specialized biotech firms with unique intellectual properties (Zucker, et al, 1994).

In the late 1990's the largest publicly funded science project, the Human Genome Project was working towards completely mapping the 3 billion base pairs of human DNA.

Researchers in academia and industry could use this map to understand basic mechanisms in biology and potentially develop new life-saving drugs. Following an open-sharing agreement known as the Bermuda Principle, HGP scientists started releasing gene sequences into the public domain as soon as they became available. Between 1998-2000, more than 99% of the human genome was released. The public nature of a detailed, scientific map was a major breakthrough for researchers (Lander, et al, 2001).

A few characteristics of the human genome make it interesting in the context of drug discovery: a) the map was made freely available on the internet for download and intellectual property rights on individual genes were restricted (Cook-Deegan & Heaney, 2010) b) scientists had developed sophisticated software, high-throughput robotics and screening systems to speed up the search process. The advance in supporting technological capabilities like genomics and high-throughput screening lowered the cost of adopting the targeted strategies. This, combined with the public availability of the human genome map potentially enhanced the appeal of target-based strategies to drug firms.

Recent research on the human genome examines the role of intellectual property rights on follow-on innovation by comparing the effect of patent protected and publicly available gene sequences (Williams, 2013; Sampat & Williams, 2017). Other research on the role of mapping on innovation indicates the positive effect of publicly available geographical maps on gold-mining efforts by entrants (Nagaraj, 2015). Building on this literature, I find that the role of scientific maps' impact on the process of innovation and an understanding of underlying mechanisms remains unknown. I address this gap in the literature by examining incumbent firms' search strategies in response to a technological change, like the human genome map, shining a lens on the role of prior firm capabilities and market conditions during this phase of transition. Empirical analyses show that certain firm specific factors change their effect during this technological change, suggesting a moderating effect of a scientific map on factors driving innovation strategy.

## **Theory**

The nature of technological change can be competence enhancing or destroying, in turn altering the trajectories of technological search and innovation strategies. Scholars have shown that factors internal and external to the firm also contribute to shaping these trajectories. An important factor driving innovation in firms is organizational learning (Levitt & March, 1998; Argote & Miron-Spektor, 2011). Another is selection, which is strongly influenced by downstream market forces and product areas that firms derive their revenues from (Kapoor & Klueter, 2015). Competition can also influence R&D strategy and product lines (March, 1991; Cockburn, et al, 2000). Thus, a host of firm and

market related factors work in tandem to affect innovation strategies and outcomes. I examine the literature on these factors in detail below.

*Organizational learning and search:* In evolutionary perspectives, firm survival is an outcome of fit between the environment and the firm. This fit is dependent on routines and firm search for new solutions. These routines stabilize over time and are influenced by learning and experience which in turn drive search processes. The stochastic nature of market selection leads to differential search outcomes leading to variation in firm growth and survival. Thus, through joint action of search and selection, firms evolve over time (Nelson & Winter, 1982; Cyert & March, 1992). A key aspect of the search and selection process in knowledge-driven firms is organizational learning. Cohen & Levinthal (1990) indicate that firms need to invest in R&D projects and continual learning in order to make effective use of new knowledge that is external to the firm.

Underlying systems of learning can over time lead to the co-evolution of products, capabilities and knowledge leading to valuable learning routines that will condition firms' adaptation to any exogenous shift like new scientific knowledge (Helfat & Raubitschek, 2000). The type of search strategy is also deeply entrenched in the organizational core capabilities and routines as a result of path-dependent processes and technological paradigms (Dosi, 1982; Nelson & Winter, 1982). Some scholars have argued that in complex fields, predictive search may be of limited value; instead, experiential, feedback-based learning is likely to be associated with creative insights, and therefore, more successful technological innovation (Gittelman, 2016; Nightingale, 1998;

Vincenti, 1990). Hence, the types of learning and path-dependent nature of knowledge and evolved capabilities can have influence how firms respond to new scientific knowledge and utilize them.

The degree to which scientific maps and new analytical tools increase exploratory search is likely to be unevenly distributed across firms. While science is a public good, it is not “free” in the sense that firms may exploit knowledge off the shelf; prior investments in R&D will influence firms’ ability to identify, evaluate, and assimilate new scientific knowledge (Rosenberg, 1990). Cattani (2005) indicates that technological preadaptation (i.e. skills, capabilities, knowledge) are beneficial during periods of technological change leading to more valuable inventions. Preadaptation with regard to knowledge and specialized capabilities is similar to the concept of absorptive capacity wherein it allows firms to capture the gains of new technological advances. In the drug discovery context, preadaptation could be prior experience or knowledge in genome technologies, combinatorial chemistry tools, or experience in certain diseases that are more amenable to target-based approaches. Therefore, firms engaged in explorative research and learning that is aligned with target-based search could be already preadapted to exploit the opportunities provided by the human genome.

Internal research projects, collaborations with academia and scientific publishing can help bridge resource gaps that incumbent firms may face and increase absorptive capacities (Cockburn & Henderson, 1998; Gittelman & Kogut, 2003). Another aspect of acquiring knowledge that can be relevant during periods of technological change is

through in-licensing or acquisition of knowledge assets. These complementary assets can immediately provide specific domain knowledge (or resources) to the focal firm and/or can be repurposed for new opportunities during later periods of technological change (Helfat & Lieberman, 2002; Veugelers, 1997). While not actively planned for, the prior exposure and experience are repurposed to take advantage of new opportunities. A new innovation strategy may require intricate coordination among various organizational units and supporting specialists, and having prior knowledge of the search strategy or possessing specialized capabilities could be an advantage in adopting it. Thus,

*H1a: Prior R&D knowledge and related experience will positively impact adoption of target-based strategies*

*H1b: Specialized organizational capabilities will positively impact adoption of target-based strategies*

*Product market focus:* Strategy scholars examining the fit between product market strategy and business models, show that product strategies that focus on specialized market segments (differentiation) have better firm performance (Zott & Amit, 2007). Other scholars have indicated that the commercialization environment in product markets is tightly linked to entrepreneurial ideas and firm strategy (Gans & Scott, 2003). Broader studies examining product diversification at the country level indicate a negative relationship with innovation (Hitt, et al, 1994; 1997). Hence, the products and markets in which firms operate are closely tied to research and development (R&D) strategy. That

is, whether firms are specialized in their product offering (e.g. Novo Nordisk focuses on diabetes drugs) or diversified (e.g. Pfizer makes drugs for cardiology, cancer, pain) influences the selection on new research projects. Such product market focus can guide the direction and composition of upstream R&D strategy (Kapoor & Klueter, 2015). The pressure to serve existing customers, available infrastructure and organizational capabilities can influence how business managers decide upon new R&D projects. Therefore, depending on the nature of technological change and firm preparedness, selection of R&D projects can vary based on firm characteristics like prior market focus and internal capabilities.

Availability of the human genome map did not automatically provide therapeutic targets for all disease areas. Areas like cancer which has been supported by strong academic research (i.e. knowledge of disease mechanisms, gene targets and pathways) benefitted enormously from the map and new gene targets, while nascent areas like neurology and neuroscience were not readily amenable to target-based research. If benefits of target-based projects (i.e. relevance of gene targets to existing markets) are perceived to not be aligned with existing market strategies or supported by internal capabilities, firms would bypass or reduce exposure such projects. Diversified firms that address multiple disease markets would be more open to targeted approaches if their activities fall within the scope of a viable gene target. In contrast, specialized firms that focused on a narrow range of diseases would be skeptical of adopting targeted strategies if there is no fit between their activities and available gene targets. Therefore, firms' product market focus will be a strong determinant of adoption of targeted strategies, especially when

knowledge of the human genome is incomplete (before the map). Hence, firms with diverse disease market interests will be stronger candidates for target-based discovery, while specialized firms will be less prone to adopting target-driven strategies prior to having a complete genome map.

*H2: Disease market diversification will positively impact adoption of target-based strategies prior to the human genome*

*Competition:* Bringing a drug to market takes about a decade, involves high R&D costs and the chances of success are very low. Firms take a real-options view of R&D investing and try to offset their bets on projects by taking on risky projects or entering new disease areas (McGrath, 1997; McGrath & Nerkar, 2004). The drug industry also faces intense competition in lucrative disease markets with multiple players offering alternative product options. When competitors adopt new innovation strategies or new technologies, focal firms are under pressure to play catch up or increase investments in innovation (Utterback & Suarez, 1993). Thus, when close competitors adopt target-based drug discovery, it can put pressure on focal firms to also adopt similar strategies (Henderson & Cockburn, 1996). The easy accessibility of the human genome combined with complementary technologies like high-throughput screening potentially enhances the viability of target-based approach, especially for firms that did not have deep knowledge or prior experience. This could in essence bring in more entrants and firms to switch search strategies thereby increasing overall competition in the market place. Thus, the

public nature of the map creates race-like conditions to capture new territory and disease markets by enticing firms to be the first to file patents around new disease targets. Hence,

*H3: Competitive pressure from other firms engaged in targeted drug discovery will positively impact adoption of target-based strategies*

*Moderating Effect of a Scientific Map:* The strategy and innovation literature has explored how disruptive technological changes can diminish the advantages of incumbents leading to gales of creative destruction. Scholars have shown the effects of complementary assets and competence enhancing innovations during technological change (Anderson & Tushman, 1990; 1991). The Human Genome Project's release of gene sequences in the public domain made knowledge of the human genome accessible to all altering the intellectual property landscape (Cook-Deegan & Heaney, 2010; Gittelman, 2016). The map was not as much a technological discontinuity or competence destroying as it was a complementary innovation that could be competence enhancing (Tripsas & Gavetti, 2000; Rothaermel, 2010). The human genome provided basic scientific knowledge that allowed scientists to better understand the molecular basis of biological processes (Scannell, 2012; Drews, 2000). Using the map as guide, researchers could probe and test hypotheses related to the cause of disease and explore underlying disease mechanisms. This knowledge was useful to biologists, medicinal chemists and pharmacologists – enhancing coordinated search efforts in drug discovery. But the map had special relevance for target-based drug discovery: an approach that was based on molecular targets and focused search processes was poised to gain from an accurate

parts-list of 10,000 druggable genes. For such directed target-based discovery strategies, the map complemented the search strategies by providing a clearer picture of the target landscape.

The availability of the map also influenced other factors that were linked to the adoption of target-based strategies. By opening up the landscape of gene targets, the human genome map made the target-based approach more accessible and valuable to disease specialized firms. Secondly, firms that had deep knowledge and specialized capabilities related to target-based paradigms like genomics would have an advantage compared to others at exploiting the map. Therefore, firms with related absorptive capacities will have a stronger effect on target-based adoption. Competitive pressures directly created by the open access nature of the human genome, will also have a positive effect on target-based adoption. Thus,

*H4: Public availability of the human genome map will increase adoption of target-based strategies by positively moderating the effect of firm's product market focus, knowledge capabilities and external competition.*

## **Research Setting**

Our research setting is drug discovery in the biopharmaceutical industry – a sector driven by high research and development (R&D) investment, long development cycles and a high rate of failure. In 2015, the US biopharma industry spent \$59 billion on R&D; it

costs \$2.6 billion and 10-15 years to develop a single drug, of which less than 12% succeeded (PhRMA, 2016). The human body is a complex system where drug-target interactions are unpredictable and lead to off-target effects and undesirable drug-drug interactions making drug discovery a challenging and costly task (Drews, 2000; Scannell, et al, 2012; Gittelman, 2016).

### Technological Search Landscape

In the early days of drug discovery, natural products, crude extracts or purified compounds were screened for biological activity. In most cases, the scientists did not know what the mechanism of action was or the drug target. Once, an active compound or natural product was found, it was tested and/or modified into a therapeutic agent. For example, Aspirin - one of the oldest marketed drugs since the late 1800's, had its origins a few thousand years ago, as a natural product derived from trees and flowers as a therapeutic agent. This method of separating bioactive compounds, or synthesizing compounds *de novo* and modifying them without a clear understanding of the drug target or mechanisms involved, and testing them for efficacy in animals is known as classical pharmacology, forward pharmacology or *phenotypic drug discovery* (Scannell, et al, 2012; Lipinski & Hopkins 2004).

In modern drug discovery, the search for a new drug starts with a therapeutic area of interest (e.g. coronary artery disease) along with a competitive analysis of existing drugs and patent landscapes. Scientists mine scientific literature to identify drug targets, genes and biological mechanisms responsible for the disease. A drug target is any protein (e.g.

enzymes, receptors) or nucleic acid (DNA, RNA) involved in a disease related biological pathway to which a drug can bind and alter its state. For example, Pfizer's Lipitor, the best-selling drug of all time, is a lipid-lowering compound which works by inhibiting HMG-CoA reductase, an enzyme responsible for cholesterol production in the liver. With information on the drug target, medicinal chemists design and synthesize compounds that can bind to the target. To do this, drug discovery scientists comb through a large theoretical landscape of solutions called chemical space.

*Chemical space* is an extremely large landscape containing  $10^{60}$  possible combinations of unique drug-like molecules. Through iterative cycles of design and testing, a candidate compound is developed that is as specific as possible to the drug target and disease. The top lead candidates are then selected for advancement into *clinical trials*. This approach which starts with knowledge of the disease mechanisms and then translates into a viable drug candidate is known as reverse pharmacology, rational or *target-based drug discovery* (Drews, 2000).

### Technological Change

The Human Genome Project (HGP), was the world's largest collaborative biology project and an important development in human biology that was aimed, in part, at revolutionizing the way in which new drugs were discovered (Gittelman, 2016). The human genome is a digital map of 3 billion DNA base pairs revealing the location and identity of genes which encoded various proteins in cells. An important milestone established by the HGP was revising the number of human genes to only 21,000. This list

of human gene sequences increased the number of available disease targets to at least 10,000 (Drews, 2000; Tripp & Grueber, 2011). The map allowed for predictive search and a high level of specificity in the search for lead candidates, chemical starting points in the search for novel drugs.

Scientists could use this map to search for specific genes and gene variants. A powerful use of the genome was the ability to predict 3-dimensional protein structures using just the gene sequences – a field known as structural genomics. This substituted for the prior method in which firms and academic labs would spend many years trying to solve the 3-dimensional crystal structures of proteins. Prior to HGP, there were less than 2000 human protein structures available (RCSB PDB, 2016). Having a protein structure (even a predicted model structure) allowed scientists to identify active sites responsible for biological action and modify compounds to block this activity. The human genome represents a disruptive event that empowered how pharmaceutical firms approached drug discovery by providing new knowledge and information of the problem space.

### Innovation Strategies

In the non-target or phenotypic search paradigm, medicinal chemists were central actors in the process of drug discovery. They were central to the process and responsible for conceiving, designing and optimizing drug candidates. Apart from the ability to synthesize organic molecules, medicinal chemists were required to have a good understanding of biology, pharmacology and drug toxicity. Experiential learning acquired by working with particular class of compounds (e.g. statins, antipsychotics) made some

medicinal chemists develop specialized skills and tacit knowledge of in optimizing compounds. This helped them pick the most promising drug candidates – a key differentiating factor in successful drug discovery and a source of firm competitive advantage.

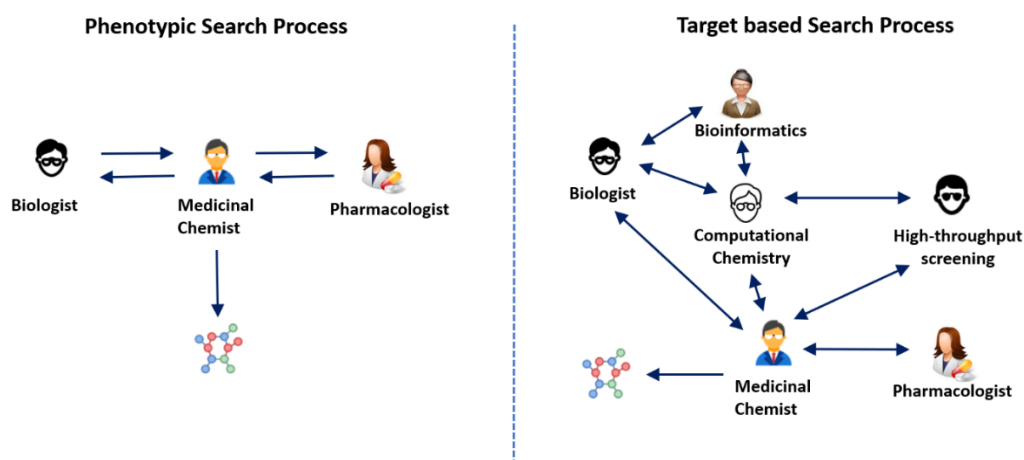


Figure 1: Differences in workflow of the two drug discovery strategies are shown here. Organizational processes and underlying capabilities are distinctively different in the two search strategies.

In the target-based approach, the organization of drug discovery involved more actors and interactions, facilitated by new specialists and firm capabilities like genomics, bioinformatics and high-throughput screening. This difference became more pronounced with the entry of the new tools like combinatorial chemistry, computational sciences and the human genome. With increasing knowledge specialization and a greater reliance on technological tools, new scientific team members emerged – computational chemists, bioinformatics specialists, and robotics personnel, along with high-throughput screening facilities, new software tools and techniques. The medicinal chemist's increased

interactions and coordination with team members indicates changing organizational processes and related knowledge capabilities (Figure 1). In sum, the availability of high-throughput methods accompanied with the precision of the genome map made viable the targeted approach guided by the genomics – the adoption of which altered innovation processes, workflows and organizational capabilities.

## **Methods**

### **Data**

This study introduces a novel and unique data set known as broad chemistry or Markush patents. Data collection and sample selection criteria are described below.

Patents are central to intellectual property protection and appropriation in biopharmaceutical industry and widely used in economic analyses (Scott & Sampat, 2012). As of 2016, more than 90% of marketed drugs are small molecules making chemistry-based patents relevant for this analysis. I utilize a novel dataset comprised of a special type of chemistry drug patent called Markush patents to test my method and reveal insights on firm level exploration. An inventor or firm uses a Markush patent to make general claims for use of a molecule without revealing its exact structure – this compound is hidden among hundreds of other similar looking compounds.

*Markush or Broad chemistry patents:* Markush patents are used across industries which work with chemicals like drugs, materials science, industrial chemicals and agrobiotechnology. These patents contain a special structure called Markush structure,

named after Eugene Markush, to capture a broad set of compounds. This type of patents is filed very early in search process where a firm makes general claims for use of a molecule without revealing the exact compound they are pursuing. The compound that could make it to market is hidden among hundreds of other similar looking compounds. In the drug discovery stage, Markush patents are very specific to small molecule drugs (compared to biologics or process patents) and mark the starting point where firms begin to claim intellectual property rights in chemical space (Southall & Ajay, 2005). Follow on drug patents and granted patents are based on this original Markush patent.

Another interesting feature of Markush patents is that they encapsulate the actual compounds made, and those that the firm or inventor plans to protect for future use – sometimes, running into the hundreds or thousands of novel compounds. Thus, these compounds within Markush patents represent the explorative effort undertaken by the firm or inventor in chemical space. While granted patents or single compound patents that cover a drug (like Viagra) are important for intellectual property protection, they do not indicate the original explorative effort undertaken by the firm. Invariably, all such granted patents will reference the starting Markush patent or application. While not all Markush patent applications end up becoming granted patents or lead to a drug, they represent capture all small-molecule related R&D activity and explorative research – not just the ones that become successful.

I exploit these unique features of Markush patents and the “flags” firms plant in chemical space to measure their exploration over time. This non-bias in the nature of Markush patent applications make them especially useful for analyzing overall technological

search trajectories of firms and estimating the broad effects of technological change. These unique properties of Markush patents make them well suited to study firms' technological search trajectories over time. For my empirical research, I have collected 39,000 drug related Markush patent applications filed in the WIPO patent office from 1990-2004, and extracted millions of novel compounds embedded in them.

*Data source and collection:* The source for Markush patents is the Chemical Abstract Society's (CAS) Scifinder product. Scifinder is the world's largest repository of chemical structures, published articles and patents and provides access to MARPAT – a comprehensive database of Markush patents that cover all 9 major patent offices and 63 patent authorities worldwide. More than 1 million Markush structures and about 481,000 Markush patents and applications from 1988-present are available for searching. In a recent comparative analysis of patented compound databases, CAS's Scifinder was found to be more comprehensive and accurate compared to the Derwent World Patents Index and Reaxys (Ede, et al, 2016).

A full discussion of data sampling and Markush patents is provided in essay one of the dissertation

In this study I focus on the impact of the human genome on the adoption of targeted-strategies by incumbent firms, and hence eliminate patents assigned to academic, medical and research institutes. After eliminating non-firm assignees and collaborations, I have 32,733 Markush patent applications assigned to firms. These Markush patent applications

cover multiple priority application countries and their share of the patents is shown in Figure 2.

The top patenters in the sample were United States, Japanese and European biopharmaceutical companies. With the human genome being publicly accessible these firms and their modes of drug discovery were equally at risk of being impacted by the human genome map. Therefore, the sample for this study includes international WIPO Markush patent applications filed assigned only to firms.

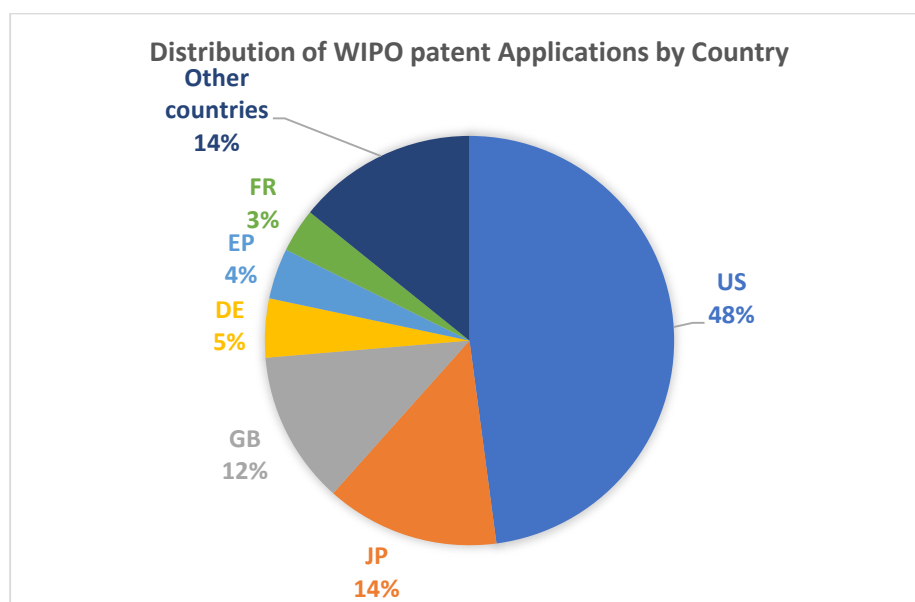


Figure 2: Distribution of WIPO patent applications by patent application country.

*Firm selection and Sample selection:* In this study, I examine how the search strategies of incumbent firms changed after a technological event – i.e. the human genome map. To analyze this empirically, I would need to measure differences in drug discovery strategies at the firm-level pre and post the arrival of the map.

To be able to test changes in search strategy, I only require firms that experienced the technological change and are supported by sufficient number of projects to make inferences on their search strategies. For example, a firm that emerged after 2000 would already be exposed to the treatment (map) and will not provide information on how their strategies changed in response to the map. Similarly, a firm that patented only in the pre map period or did not consistently engage in R&D activity will not be able to inform on strategic changes in drug discovery across the entire time period. Figure 3, describes the process of constructing a longitudinal panel with the firm level observations needed for this analysis.

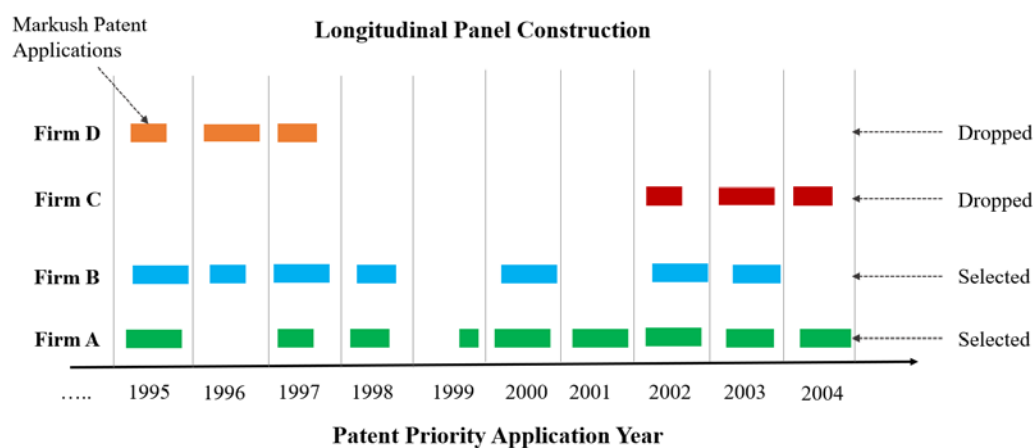


Figure 3: View of the selection criteria for constructing the longitudinal panel needed for testing change in drug discovery strategies. Firms with observations that are not present both pre-and post map period are dropped. Each observation is a Markush patent application filed by a firm in a given year.

To ensure that the longitudinal panel I constructed is consistent and reliable, I restricted selection of firms to the following criteria: i) firms had to be present in both pre (1990-1997) and post (1998-2004) human genome time periods ii) firms needed to show that they patented in most years of this time period. This requirement indicates continuous R&D activity and shows firm exposure to the technological change iii) firms on average needed to have a minimum of 2 patents a year resulting in a cut-off of at least 30 patents in the sample (i.e.  $2 \times 15$  years). Having these cutoff criteria ensured that I did not bias the sample with firms that emerged only after the human genome or had firms that only saw the first period but not the impact of the human genome. Also, having a minimal count cutoff of 30 patents eliminated firms that were not actively engaged in R&D activities throughout this period.

Applying these cutoffs, resulted in 131 firms having 30 or more patents. The constructed panel contains 21,315 patent applications from 131 firms that were filed between 1990-2004. This sample represented more than 65% of all WIPO firm patent applications initially collected for analysis.

The patent application counts and proportion of targeted-strategies for the top 30 firms are shown in Table 1 below.

<b>Firm</b>	<b>Patent Application Count</b>	<b>Target-Based Strategy</b>	<b>% Target- based</b>
Smithkline Beecham	1325	601	45.36
Merck & Co	1139	474	41.62
Pfizer	832	339	40.75
AstraZeneca	788	408	51.78
Eli Lilly	682	315	46.19
Pharmacia	559	200	35.78
Glaxo	550	303	55.09
Novartis	527	233	44.21
Schering	512	240	46.88
Bristol-Myers Squibb	484	225	46.49
Bayer	458	122	26.64
Warner- Lambert	458	161	35.15
Wyeth-AHP	443	187	42.21
F.Hoffmann- La Roche	397	208	52.39
Takeda Chemical Industries	397	166	41.81

Novo-Nordisk	389	150	38.56
Fujisawa Pharmaceutica l Co	349	92	26.36
Abbott Laboratories	347	114	32.85
Boehringer Ingelheim	346	131	37.86
Janssen Pharmaceutica	304	103	33.88
Aventis Pharma	288	73	25.35
DuPont	270	80	29.63
Rhone- Poulenc Rorer	270	90	33.33
BASF	243	58	23.87
Merck Sharp & Dohme	243	176	72.43
The Procter & Gamble	234	69	29.49
G.D. Searle	213	48	22.54
Zeneca	205	54	26.34

Merck Patent GmbH	204	70	34.31
Shionogi & Co	203	66	32.51
Total	13,659	5556	

Table 1: Top 30 firms account for 64% of all Markush patent applications filed between 1990-2004 (total: 21,315)

### *Identification of Drug Discovery Strategies*

I interviewed discovery scientists to understand what drug discovery strategies firms and how to identify them using patent records. Target-based patents in their description would normally name a disease target gene or gene symbol in addition to mentioning specific keywords that indicated the search strategy used to make the compounds. Figure 3 below shows the abstract of a Markush patent where a molecular target (COX-2, cyclooxygenase-2), specific keywords related to target-based approach (inhibiting activity) and disease indications (inflammation) can be obtained by reading the abstract.

To identify the presence of targeted strategies, I developed a text-based classification method based on patent text. Initial versions of this method included the text of the title, abstract and description section of the patent. But this yielded false positives, as inventors and examiners would reference other patents, titles or references that could contain a gene name, but not be used in the discovery approach. Hence, a non-targeted approach

could be mis-classified as a target-based strategy. To overcome this problem, I compared and contrasted the merits of using abstract versus description before settling on just using the patent title and abstract that captured the main aspects of the patent and its strategy. In addition, CAS provides two human curated fields known as Index Terms and Supplementary Terms – these are short keyword summaries of the patent provided by internal knowledge experts. The text from these fields is then compared to gene names and gene symbols derived from two sources: the National Center for Biotechnology Information (NCBI) database and the HUGO Gene Nomenclature Committee's (HGNC) database. The HGNC is a committee of the Human Genome Organization that sets standards for naming genes and assigning gene symbols. For example, the drug target for Viagra is the enzyme phosphodiesterase 5A, written using the symbol PDE5A or PDE5. A patent could use any of these versions of the gene name or symbol, hence, the algorithm should be able to account for this variation.

To counter this problem, the gene database from the National Center for Biotechnology Information (NCBI) was also used as it provides a list of gene synonyms and old names that were previously used. Thus, together a comprehensive collection of 61,561 gene symbols and identifiers were created. Using this combined data, patent abstracts, titles and curated sections from Scifinder were scanned using custom software developed for this purpose.

United States Patent		6,603,008
Ando, et al.		August 5, 2003
Sulfamoylbenzoyl pyrazole compounds as anti-inflammatory/analgesic agents		
Abstract		
This invention relates to a compound of the formula: $\text{R}^1\text{STR}_1\text{R}^2$ or a pharmaceutically acceptable salt thereof, wherein A and R, sup. 1 are each an optionally substituted 5 to 6-membered heteroaryl, wherein the heteroaryl is optionally fused to a carbocyclic ring or 5 to 6-heteroaryl, R, sup. 2 is NH, sub. 2; R, sup. 3 and R, sup. 4 are each hydrogen, halo, (C, sub. 1 - C, sub. 0)alkyl optionally substituted with halo and the like; and X, sup. 1 to X, sup. 4 are each hydrogen, halo, hydroxy, (C, sub. 1 - C, sub. 0)alkyl optionally substituted with halo and the like. These compounds have <a href="#">FIG. 1</a> inhibiting activity and thus useful for treating or preventing inflammation or other <a href="#">FIG. 2</a> related diseases.		
Inventors:	Ando, Kazuo (Aichi-Ken, JP), Kawamura, Kiyoshi (Aichi, JP)	
Assignee:	Pfizer Inc. (New York, NY)	
Family ID:	22613376	
Appl. No.:	09/723,661	
Filed:	November 28, 2000	

Figure 4: Sample patent abstract from United States Patent & Trademark Office database is analyzed using custom algorithms to identify target-based keywords, gene target and disease indications and classified as using a target-based drug discovery strategy.

A custom algorithm was implemented in the Python language using the open source Natural Language Toolkit library to mine the text of all the Markush patents in the sample. In addition to the gene identifiers, a set of keywords gathered from interviews indicating a target-based approach were also screened. These words include versions of relevant keywords like 'gene', 'genes', 'genomic', 'genomics', 'receptor', 'receptors', 'inhibit', 'inhibitor', 'inhibitors', 'target', 'targeted'. Wild-card matching (e.g. inhibit\*) was implemented to account for these variants.

The text mining and classification using this approach led to identification of gene names and symbols appearing patent titles and abstracts. False positives can occur when the algorithm tags a patent as target-based when the actual gene symbol could be used in a different context. For example, the valid gene symbols 'CAT', 'MICE' and 'PIGS' can be misconstrued as gene names when used in the context of compounds being tested for toxicity on these animals. Or the gene 'BOC' has a different connotation when used in a chemistry patent. In chemistry, BOC groups refer to *tert*-butoxycarbonyl, a protecting group in organic synthesis. To ensure reliability, I tag false positives using a special list

of such terms (see Appendix) and manually inspect patents to ensure that the algorithm driven classification is accurate and false positives eliminated. This sorting created two broad categories of patents based on their search strategy: 60% non-target based and 40% target-based.

### *Disease Market Focus*

Firms can be focused on developing drugs for multiple disease areas (diversified) or a few areas (specialized). A Herfindahl-Hirschman Index (HHI) is an established measure of market concentration, used for identifying market monopolies and competition. It is calculated by taking the market share of each firm, squaring it and summing the end result. A resulting measure is a proportion between 0 and 1; and score under 0.15 indicates an unconcentrated industry and greater than 0.25 high concentration. This method and measure is applied to calculating disease market concentration for each firm in the sample. It is calculated as

$$HHI = \sum_{i=1}^n d_i^2$$

where  $d$  is a disease area's share of the total diseases explored in clinical trials in the prior 3 years.

A firm's engagement in a clinical trial indicates intent to enter that disease market.

Successful clinical trial results in market entry or product launch for a specific disease area.

Using Clarivate Analytics Cortellis database (formerly Thomson Reuters) of clinical trials outcomes, I collected disease areas that firms had run clinical trials from mid-1980s to 2004. Using Cortellis's disease hierarchy I categorized 881 diseases into 27 broad categories (e.g. Ocular Disease, Respiratory). See Appendix Table A for full list of broad categories. For each firm and specific year, the prior 3 years market share of each disease is calculated (derived from clinical trials participated in) and a HHI score generated. Thus, for any given firm-year, a HHI score provides an accurate picture of the firm's prior 3 years disease focus. I chose a prior 3-year window as it adequately captures firm activities in various disease areas as compared to calculating HHI in the current year. For instance, diversified firms may be conducting clinical trials in non-overlapping disease areas across years and it will be hard to capture this diversity with single year measures. A similar time window is also used to capture firm experiences and knowledge capabilities in biotechnology by other scholars (Kapoor & Klueter, 2015).

*Genomics Capabilities:* Ability to engage in gene related research that requires high levels of expertise is a strong indicator of firm capabilities in genomics. When firm scientists publish gene related research in journals they are required to submit the gene sequences that they cite in their research to a central, public sequence repository like GenBank. GenBank is the largest genome database for genome sequences hosted and supported by the National Center for Biotechnology Information (NCBI). GenBank curates the sequence, provides an accession number to the gene and makes it accessible to the community. Zucker & Darby (2001) use GenBank sequence counts assigned to Japanese inventors to measure the effect of star scientists. I use a similar approach to

capture firm-level genomics capabilities by identifying GenBank sequences assigned to firms. For each firm in my sample, I retrieved all genomic sequences (for any organism) submitted to GenBank from mid-1980s to 2004.

Genomic sequence counts using firm name was carried out using a BioPython script that queried the publicly available NCBI GenBank web database for each year in the time period (1990-2004). Using a sliding window of 3 years, the program counted the number of gene sequences submitted by a firm in the prior 3 years starting from 1990. Thus, if Merck submitted DNA sequences between 1987-1989, the total sequence count was assigned to Merck-1990. The GenBank sequence count provides an unbiased estimate of the firm's engagement in genomics research and capabilities accumulated in the prior 3 years. This data collection was carried out for all 131 firms in the sample.

#### *Related R&D Knowledge*

Prior scholars in the innovation field have used firm publications as a measure of internal knowledge stocks (Cockburn & Henderson, 1998; Gittelman & Kogut, 2003). There are both private and publicly available literature databases like Scopus, Thomson Web of Science and Google Scholar. The NCBI's PubMed is the largest open access citation repository of life science related research literature containing more than 27 million records obtained from MEDLINE, books, journals and other scientific sources. I comparing these various databases for coverage specifically related to life sciences and drug discovery publishing and found the PubMed database to be most comprehensive.

An interesting feature of the PubMed database is the categorization of each article using special scientific terms called Medical Subject Headings (MeSH). MeSH are the National Library of Medicine's controlled vocabulary thesaurus used for indexing articles in PubMed. Recent work by Li, et al (2015) uses MeSH terms to indicate whether patents are disease targeted. These features of PubMed make it appropriate for collecting research & development related scientific publications for the firms in the sample.

To identify publications that are related specifically to the target-based drug discovery, I interviewed drug discovery scientists and asked them to curate a list of target-based MeSH terms. A full list of the MeSH terms is included in the Appendix. Using these target-specific MeSH terms and keywords selected from MEDLINE's target-based vocabulary and the firm names from the sample, I created a Python program to query PubMed for research articles. This resulted in articles that used target-based technologies like combinatorial chemistry, genomics, computer-based drug design, protein modeling and computational chemistry.

Using the PubMed database and creating yearly cutoff windows for each firm, I collected field-specific research articles affiliated with the firm in the prior 3 years. Other scholars have also used similar time windows to capture recently acquired firm knowledge and capabilities (Kapoor & Klueter, 2015; Henderson & Cockburn, 1996). The surveyed articles also included those published by the firm in collaboration with academic centers or other firms as it was evidence that the firm scientists were engaged in cutting edge

research. The target-related publication counts are linked to each firm in the sample by year.

### *Firm Information*

Assignees were extracted from patent records and stored separately. Assignees have a name and location associated with them. Each assignee is categorized into a separate category: firms, universities, medical centers and collaborations. This classification process yielded 3300 firms, 1500 universities and medical centers. For the analysis, only 131 firms and the Markush patent applications assigned to them were selected.

## **Research Design**

### Unit of Analysis

The unit of analysis is a drug patent application filed each year by a firm (total: 21,315 patents). Hence, each observation is an individual patent application which employs a specific drug discovery strategy and is associated with originating firm characteristics like knowledge, capabilities and experience.

e.g. Patent number [US6603008] filed in [1999] having a [Target-based] strategy belongs to [Pfizer] which had [310 publications] and [713 genbank sequences]

### Dependent Variable

*Is Target Based (adoption model):* The dependent variable *Is\_Target* is a binary variable – 1 if the patent employs a target-based strategy, 0 if it is a non-target based strategy.

Since, I interested in the adoption of target-based search strategies and the various factors that influence it, a simple categorical dependent variable is specified for logistic regression purposes.

### Independent Variables

*Disease Market Specialization:* Herfindahl-Hirschman index of firm's disease market focus in prior 3 years (specialized vs diversified) calculated as the sum of the squared disease shares for each firm.

*Genomics Capabilities:* This is a count of all DNA sequence records submitted to the public GenBank database by the firm the in prior 3 years. This measure of genomics capabilities indicates if firms had any genomics experience prior to when the small molecule patents were being filed.

*Related R&D Knowledge:* This is a count of all scientific publications related to combinatorial chemistry and genomics published by the firm in the prior 3 years. I use the 3-year cutoff as this captures a time period when the knowledge is relatively recent and can influence the selection of current projects. This measure captures combinatorial chemistry or genome specific learning that the firm has engaged in the prior 3 years.

*Human Genome Map:* This is a categorical variable (0/1) indicating public availability of human genome sequences released by the Human Genome Project. The public access to gene sequences started in 1998, hence, 1998-2004 is the post-map period and 1990-1997 is the pre-map period.

*Target-based Experience:* This is the number of target based patent applications filed by the firm in prior 3 years. I apply the same logic for time window selection as R&D knowledge above, given that target-based discovery strategies are knowledge and capabilities driven.

*Competition:* This is the yearly increase in the number of target-based patents filed by competitors in the same year. Instead of defining market-specific competition for each firm, all target-based patents filed minus the focal firm's patents are counted and the difference with the previous year is calculated (i.e. competitor's target patents in year<sub>x</sub> – competitor's target patents in year<sub>x-1</sub>). This year-on-year increase in competitor's target patents captures the yearly difference in level of competition on the focal firm. That is, it represents how actively other competing firms are engaged in target-based drug discovery. Thus, the competition variable is the increase in number of target-based patents compared to the previous year. Though we cannot completely separate out the effect of the map on competitor's target-based patents, the year-on-year increase measure allows us to reliably detect changes in competition with regard to the availability of the map.

*Firm size:* The Mergent Online database of firms was to categorize firm sizes. Large firms were coded based on a threshold of having at least 10,000 employees. This resulted in 83 large firms and 48 small firms (total 131).

### Logistic Regression Model

Logistic regression is a non-parametric model suited to predict the probability of a binary response on one or more independent variables. The logit estimate is a function of the predictors allowing us to capturing effects for a per unit increase in predictor. It allows one to say whether the presence of a risk factor (e.g. specialized capabilities) increases the odd of a given outcome (i.e. adoption of target-based) by a specific factor, all other factors held constant. The predictors in this model include firm level characteristics (size, prior experience in target-based, specialized capabilities related to target-based, knowledge), market factors (competition, product market focus) and role of technological event (introduction of human genome map). Prior innovation studies have used logistic regression to estimate R&D strategy models (Cassiman & Veugelers, 2006). The logit model is suited to test the effect of these predictors on the adoption of target-based strategy.

The logit model is specified as:  $y = \beta_0 + \beta_1 X + e$

where,  $e$  is an error distributed by a standard logistic distribution and  $X$  is the predictor.

Interaction Effects: To test for moderating effects, an interaction between the independent variable and moderating variable is specified. A moderating variable (M) changes the direction or magnitude of the relationship between two variables. The test of the moderating effect of M is a comparison of the following two logit specifications:

$$y = \beta_0 + \beta_1 X + e$$

and

$$y = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 XM + e$$

where the  $\beta_3 XM$  is the product of the predictor and moderating variable.

For model fitting, logistic regression uses a maximum likelihood approach to find the smallest possible deviance between the observed and predicted values.

## Results

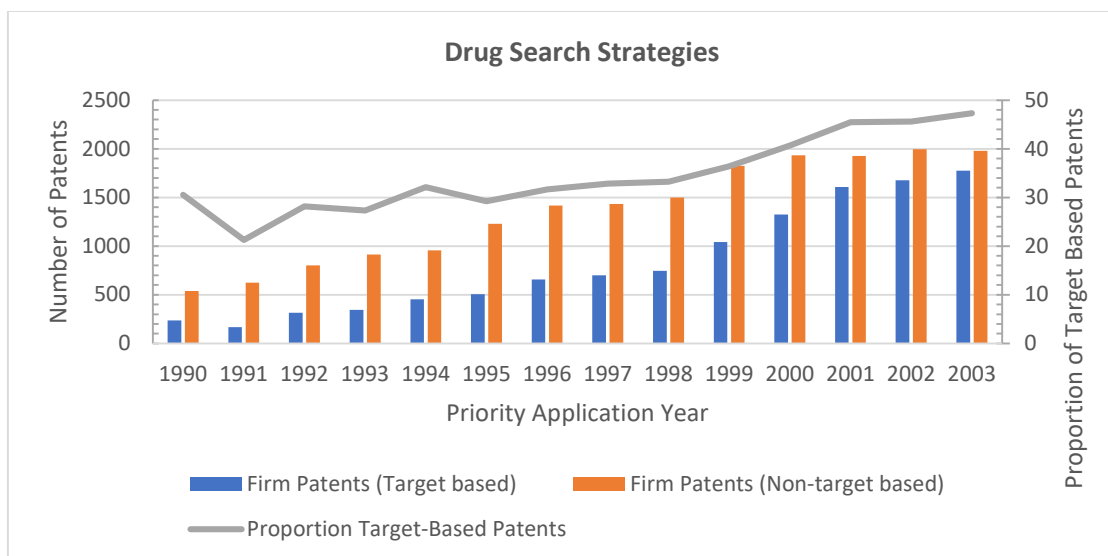


Figure 5: Increase in the adoption of target-based drug discovery strategies observed in patents between 1990-2003

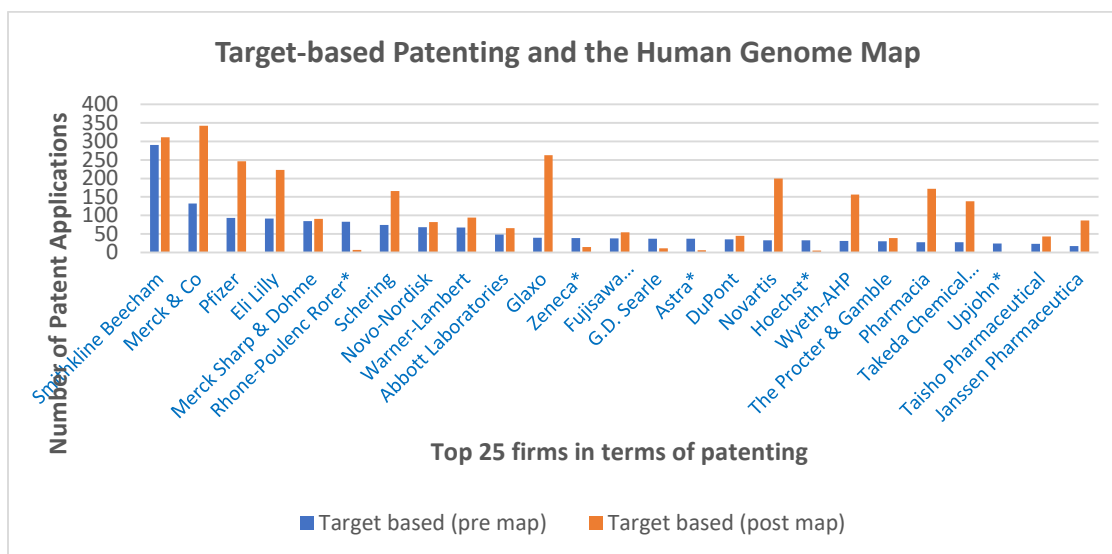


Figure 6: Changes in target-based drug discovery strategy pre-post human genome map.

Asterisks indicate firms that were acquired or merged during this time period.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
is_target (1)	1													
Compound (2)	0.0699*	1												
Firm size (3)	-0.0214*	-0.0083	1											
Disease focus (4)	0.0014	0.0072	-0.1276*	1										
Genbank (5)	0.0753*	0.0866*	0.1526*	0.0411*	1									
Publications (6)	0.0690*	0.0597*	0.2719*	0.1310*	0.5403*	1								

Priortarg etexp (7)	0.1 29 8*	0.0 44 3*	0.2 41 7*	0.0 89 9*	0.4 81 6*	0.6 25 9*	1							
Competit ion (8)	- 0.0 31 5*	- 0.0 09 8	- 0.0 11 2	- 0.0 00 7	- 0.0 16 6*	- 0.0 55 4*	- 0.1 94 8*	1						
Map (9)	0.1 28 6*	0.0 86 3*	- 0.1 06 8*	- 0.1 38 3*	0.2 84 4*	0.0 47 2*	0.2 39 4*	0.0 196 *	1					
disease_f ocus_ma p (10)	0.0 81 9*	0.0 59 6*	- 0.2 46 3*	0.4 98 0*	0.2 30 8*	0.0 99 3*	0.2 15 6*	0.4 25 133	0.4 8*	1				
genbank_ map (11)	0.0 81 8*	0.0 84 3*	0.1 39 0*	0.0 30 2*	0.9 86 7*	0.5 23 4*	0.4 86 7*	- 0.0 150 *	0.3 21 8*	0.2 46 9*	1			
publicati on_map (12)	0.1 01 1*	0.0 79 0*	0.1 81 7*	0.0 09 8	0.6 58 9*	0.8 11 4*	0.6 38 3*	- 0.0 481 *	0.3 57 1*	0.2 39 1*	0.6 73 0*	1		

priortarg	0.1	0.0	0.1	0.0	0.5	0.5	0.9	-	0.1	0.3	0.2	0.5	0.6	
etexp_ma	35	55	89	25	22	27	36	861	80	76	36	95		
p (13)	1*	6*	1*	7*	1*	4*	3*	*	2*	9*	9*	2*	1	
competiti	-		-	-		-	-						-	
on_map	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.9	0.1	0.1	0.0	0.0	0.1	
(14)	11	05	04	35	41	44	42	819	74	02	49	24	06	
	6	1	2	8*	4*	5*	4*	*	5*	5*	7*	8*	4*	1

Table 2: Correlation table of outcome variable and predictors. Asterisk indicates  $p < 0.05$

Logit Model				
	(1)	(2)	(3)	(4)
VARIABLES	Model 1	Model 2	Model 3	Model 4
diseasefocus	-0.143 (0.156)	0.373 (0.271)	0.0769 (0.162)	-0.143 (0.156)
genbank	-0.00108** (0.000431)	6.79e-07 (0.000162)	-5.60e-05 (0.000147)	-0.00108** (0.000431)
publications	-3.32e-05 (0.00105)	4.21e-05 (0.000497)	3.59e-05 (0.000463)	-3.32e-05 (0.00105)
priortargetexp	0.00654***	0.00482***	0.00496***	0.00654***

	(0.00250)	(0.00112)	(0.00112)	(0.00250)
competition	0.00116**	-9.33e-05	-7.68e-05	0.00116**
	(0.000504)	(7.99e-05)	(7.52e-05)	(0.000504)
map			0.465***	0.447***
			(0.0669)	(0.136)
disease_focus_map				0.516*
				(0.281)
genbank_map				0.00108**
				(0.000430)
publications_map				7.54e-05
				(0.00107)
priortargetexp_map				-0.00172
				(0.00267)
competition_map				-0.00125**
				(0.000518)
Constant	-0.975***	-0.529***	-0.928***	-0.975***
	(0.107)	(0.0933)	(0.0818)	(0.107)
Observations	7,504	13,620	21,124	21,124

---

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 3: Logistic regression estimates (baseline probabilities) for the adoption of target-

based strategy with clustered standard errors

OLS Model with Firm Fixed Effects				
	(5)	(6)	(7)	(8)
VARIABLES	Model 5	Model 6	Model 7	Model 8
diseasefocus	-0.0495*	-0.0226	-0.0456**	-0.0367
	(0.0276)	(0.0700)	(0.0229)	(0.0234)
genbank	-3.19e-06	0.000109***	8.95e-05***	3.43e-05
	(5.70e-05)	(3.07e-05)	(2.93e-05)	(7.42e-05)
publications	0.000350	8.23e-05	5.99e-05	0.000186
	(0.000218)	(0.000216)	(9.45e-05)	(0.000179)
priortargetexp	0.000136	0.000369	0.000609*	0.000710
	(0.000586)	(0.000356)	(0.000324)	(0.000642)
competition	0.000276***	-4.02e-05**	-3.00e-05	0.000272***
	(9.79e-05)	(1.96e-05)	(1.89e-05)	(9.76e-05)
map			0.0906***	0.142***
			(0.0147)	(0.0251)
disease_focus_map				-0.0481
				(0.0627)
genbank_map				7.03e-05
				(7.84e-05)
publications_map				-0.000200

				(0.000260)
priortargetexp_map				-0.000107
				(0.000671)
competition_map				-0.000307***
				(0.000102)
Constant	0.270***	0.405***	0.310***	0.270***
	(0.0122)	(0.0280)	(0.0179)	(0.0169)
Observations	7,504	13,620	21,124	21,124
R-squared	0.003	0.005	0.019	0.020
Number of firmid	122	127	131	131

---

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 4: Ordinary Least Squares with fixed effects estimates for adoption of target-based strategy with clustered standard errors

Figures 5 and 6 are descriptive charts indicating the increase in target-based strategies in the pre and post human genome time periods. Figure 6 shows this change in strategy at the firm level for the top patenting firms in the sample. Table 2 provides a correlation measure of the dependent and independent variables.

### **Adoption of Target-Based Search Strategy**

Logistic regression is a non-parametric model suited to predict the probability of a drug

patent adopting a target-based or non-target based approach for drug discovery. The predictors in this model include firm level characteristics (size, prior experience in target-based, specialized capabilities related to target-based, knowledge), market factors (competition, product market focus) and role of technological event (introduction of human genome map). Logistic regression analyses were performed using the STATA 14.2 statistical package and results are presented in Table 3 and 4.

*Related R&D Experience (publications and prior target experience):* This measure captures the number of scientific articles related to the target-based approach published by the firm in prior 3 years. Related scientific knowledge has a weak but positive effect on the adoption of target-based strategies (Models 2 and 3). My second measure of related R&D experience was the number of target-based patents filed by the firm in prior 3 years. This firm-specific predictor is positive and significant ( $p < 0.001$ ) across models 1-4 and when controlled for firm fixed effects in model 7. This effect is also present when interacted with the availability of the human genome (1998-2004). Together, these results indicate that prior R&D experience in terms of scientific publishing and patenting that is target-oriented is a strong predictor of adoption of target-based strategy. This result supports Hypothesis 1a.

*Specialized Genomics Capabilities (genbank):* The number of gene sequences submitted in prior 3 years to the GenBank database is used as a proxy for the firm's specialized capabilities in genomics – the ability to manipulate and analyze genes or parts of the genome. In the models tested (1 and 4) specialized genomics capabilities had a negative

and significant effect ( $p < 0.001$ ) on adoption of target-based strategy for drug discovery. This is in contrast to what I predicted and does not support Hypothesis 1b.

*Product Market Focus (disease\_focus):* In the period before the human genome map was available publicly (pre-map), disease specialization is a negative predictor of the target-based drug strategy (model 1 and 5). But in the period after the genome became accessible (post-map), disease specialization is a positive predictor of adoption of the target-based approach (model 2) but is not significant. Thus, product market focus has a contextual effect on adoption of target-based strategy – disease diversification helped prior to the map but disease specialization appears to be a predictor in the time period after the human genome. This result provides partial support to Hypothesis 2.

*Competition:* Competitive pressure from other drug firms in the industry to engage in target-based strategy was coded as the year-on-year increase in target-based drug patents filed by other firms in prior 3 years. Though this does not capture direct product-market competition (i.e. for specific firm products or markets), it is an indirect measure of the general trend in the drug industry. In an industry as competitive as the biopharmaceutical industry where intellectual property rights play a strong role in strategic positioning, competitor actions and industry trends are a considerable force in setting firm R&D strategy. In models 1-4, competitive pressure has a positive and significant effect on the adoption of target-based strategies. This effect is also positive and consistent when estimated with firm fixed effects (models 5-8). These results support Hypothesis 3.

*Moderating Effects of the Human Genome:* The time period during which the majority of the map (>90%) became publicly available was coded with a dummy variable and interacted with the main predictors. The availability of the map has a strong and significant impact on the adoption of target-based strategies as shown in model 3 and 4. When estimated with firm fixed effects (model 7 and 8) I observe a similar positive effect on adoption of the target-based approach. The moderating effect of the map on the other predictors is tested using interactions shown in model 4. Disease specialized firms have a stronger and significant effect on adoption of target strategies with the availability of the map. Compared to models 1-3, firms with genomics experience have a strong and positive effect on the adoption of target-based strategy with the availability of the human genome in model 4. Similarly, I see scientific publishing having a weak but positive effect on adoption, and that prior target-based experience have a negative effect but these effects are not statistically significant. A moderator changes the direction or magnitude of the relationship between two variables and I see the availability of the human genome following this behavior. These results support the moderating effect of the human genome map on factors which influence the adoption of target-based strategy. Thus, supporting Hypothesis 4.

*Robustness checks:* To check for multicollinearity among the predictor variables in the OLS regressions, a variance inflation factor (VIF) measure was calculated. The variance inflation factor is the ratio of the variance in a regression model with multiple predictors, divided by the variance of the model with one predictor only. This calculation quantifies the severity of multicollinearity and provides an index measure. The VIF was calculated

in STATA and mean value of 1.98 was obtained for the predictors. VIF less than 10 indicates low collinearity between the predictors in the model. Hence, multicollinearity effects were ruled out in this regression model.

A Wald Chi squared test was used to test the statistical significance of each coefficient in the model (in STATA 14.2) including interaction effects (moderation). That is, whether they can be removed from the model without affecting it in any meaningful way (null hypothesis). The Wald test uses a Z-statistic to yield a chi-squared distribution and appropriate for large sample data (Frazier, et al, 2004). Two Wald tests were conducted, one with only the primary predictors ( $\chi^2=90.75$ ,  $p=0.0$ , degrees of freedom=6) and another with the interaction effects included ( $\chi^2=761.09$ ,  $p=0.0$ , degrees of freedom=11). Both Wald tests resulted in chi-squared parameters greater than zero, indicating that the effect of these predictors are significant and necessary for this regression model.

## **Discussion and Conclusion**

This study examines how technological search strategies are impacted when disruptive new technologies enter the market. I focused on understanding the organizational and external conditions that influenced how incumbent drug firms selected their drug discovery strategies with the arrival of the human genome. Logistic regression models show that prior R&D experience, specialized genomics capabilities, product market focus and competition have a significant effect on the adoption of target-based strategies.

In my analysis, related R&D knowledge has a positive impact on the adoption of target-based strategies. This empirical result supports other studies in the literature that point to the role of absorptive capacities and explorative research in adapting to new technological change and exploiting them (Cohen & Levinthal, 1990; Cockburn & Henderson, 1998; Gittelman & Kogut, 2003). Firms with diverse portfolios can capture internal/external knowledge spillovers through cross-pollination of ideas and movement of personnel. Prior studies show that alignment of complementary projects can have a positive internal spillover effect (Henderson & Cockburn, 1996; Cassiman & Veugelers, 2006). Firm investments in basic projects and learning, builds up knowledge stocks that can be applied to new to industry technologies providing firms with strategic and competitive advantages (Cattani, 2005). Hence, firms that were preadapted in target-related technologies like genomics and high-throughput screening were in a favorable position to adopt target-based strategies.

The direct role of specialized genomic capabilities (Hypothesis 1a) shows a negative effect on adoption, but these specialized capabilities show a positive effect on adoption when moderated by the availability of the human genome map. This result is very interesting as it adds a contextual aspect to the role of specialized capabilities in firm strategy. Firm capabilities or core competencies provide platforms for product exploration and represent investments in future opportunities. For example, Merck, a global leader in small-molecule drug discovery, led a collaboration with Washington University in 1995 to sequence 300,000 gene sequences. These investments are path

dependent and irreversible due to complex interdependencies between organizational and technological elements (Kogut & Kulatilaka, 2001).

Adner & Levinthal (2004) explain that this path dependence that affects technological search arises due to endogenous rather than exogenous factors. Recent research by Teodoridis et al (2017) argues that in fast-evolving knowledge domains like theoretical mathematics, specialists have an advantage in identifying new creative opportunities. Building on these studies, my results show that while firms engaged in genomics capabilities prior to the availability of the map, it did not have an effect on target-based projects. But with the availability of the human genome, specialists in genomics capabilities could exploit the map better than others – resulting in more target-based discovery projects. Thus, firm investments in developing specialized capabilities and exploratory research may not have immediate payoffs but provide strategic advantages when competence enhancing technological changes emerge.

Diversified disease market focus had a positive effect in the time period prior to the map. This finding is consistent what I predicted and with the strategy literature on diversification strategy (Rumelt, 1982; Hitt, et al, 1994). But this effect changes in the time period after the map, showing that disease market specialization in firms is a positive indicator of adoption. This reflects a larger underlying change in R&D strategy at the industry level, where firms are becoming more diversified in general (see Table 2). Specialized firms that were reluctant to engage in target-based search previously (before the map) are adopting it at higher levels after the map. This result is interesting from a

strategic perspective – open access technologies and publicly available knowledge appears to be a catalyst in promoting a switch in innovation strategy.

Competition from other firms that are adopting the target-based strategy has a positive effect on focal firm's adoption of similar strategy. This result supports prior studies that report the influence of competitive pressure on firm strategy (Hoskisson, et al, 2000; Zahra, 1996). The availability of the genome is in essence providing all firms with the same map of the search landscape. Privately held advantages related to the map, through private sequencing efforts or collaborations, are eliminated with the public release of sequence information (Williams, 2013; Sampat & Williams, 2015). This puts pressure on both incumbents and entrants to dig deeper and wider on the human genome before competitors lock up valuable regions. This effect is positive in the time period prior to the map.

In this background of organizational and market factors, the human genome plays a significant role in moderating the influence of firm specific capabilities. The precise, detailed nature of the map by providing the location and identity of genetic sequences is directly beneficial to firms engaged in the target-based approach (Drews, 2000). In the statistical models, we find that the introduction of the human genome map positively moderates the effect of competition, specialized genomics capabilities and the firm's disease market focus on adoption of targeted strategies. This has interesting strategic implications as firms that had invested in genomics capabilities were better positioned to exploit the rich new information provided by the map. For example, firms like Merck and

GlaxoSmithKline reorganized their internal R&D groups and invested heavily in genomics technologies, gene sequencing projects and collaborations with academia to prepare for the flood of genomic data that would arrive with the completion of the Human Genome Project. Thus, the map was seen as a new tool, technology and knowledge source to complement their existing knowledge capabilities and reinforce their target-based strategies.

The results from the statistical models support these trends. While diversified firms were at a higher risk of adopting target-based prior to the map, we see firms that were disease market focused increased their adoption of targeted strategies after the map. Disease specialized firms' prior knowledge and capabilities in specific product markets serves as an advantage in exploiting the new map. Instead of substituting their strategic capabilities, the human genome map acts as a complement and strengthens their adoption of target-based strategies. For example, the discovery of new gene analogs for existing disease targets that firms were already working on or had expertise in, opened up new markets and better clinical targeting of existing drugs and indications. Thus, the human genome map enhances the adoption of targeted strategies by positively moderating the effect of prior specialized genomics capabilities, product market focus and competition. Hence, the map acts as a complement in the adoption of targeted strategies.

This study addresses a gap in empirical research: understanding how technological change impacts innovation strategy. My analyses show a nuanced understanding of the conditions and factors that influence strategic refocusing related to R&D projects.

Contrary to creative destruction, the human genome enhances adoption of targeted strategies by incumbent firms and increases exploration by specialized firms. For practitioners, firm capabilities have to continually evolve during disruptive technological change to avoid losing their innovative edge. Engaging in explorative research projects and continuous organizational learning buffer the effects of disruptive change and preadapt firms to take advantage of emerging new technologies.

## References

- Adner, Ron, and Daniel Levinthal. "Doing versus seeing: Acts of exploitation and perceptions of exploration." *Strategic Entrepreneurship Journal* 2.1 (2008): 43-52.
- Anderson, Philip, and Michael L. Tushman. "Managing through cycles of technological change." *Research-Technology Management* 34.3 (1991): 26-31.
- Anderson, Philip, and Michael L. Tushman. "Technological discontinuities and dominant designs: A cyclical model of technological change." *Administrative science quarterly* (1990): 604-633.
- Cassiman, Bruno, and Reinhilde Veugelers. "In search of complementarity in innovation strategy: Internal R&D and external knowledge acquisition." *Management science* 52.1 (2006): 68-82.
- Cattani, Gino. "Preadaptation, firm heterogeneity, and technological performance: a study on the evolution of fiber optics, 1970–1995." *Organization Science* 16.6 (2005): 563-580.
- Christensen, Clayton M. "The rigid disk drive industry: A history of commercial and technological turbulence." *Business history review* 67.4 (1993): 531-588.

Cockburn, Iain M., and Rebecca M. Henderson. "Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery." *The Journal of Industrial Economics* 46.2 (1998): 157-182.

Cook-Deegan, Robert, and Christopher Heaney. "Patents in genomics and human genetics." *Annual review of genomics and human genetics* 11 (2010): 383-425.

Dosi, Giovanni. "Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change." *Research policy* 11.3 (1982): 147-162.

Drews, Jürgen. "Drug discovery: a historical perspective." *Science* 287.5460 (2000): 1960-1964.

Frazier, Patricia A., Andrew P. Tix, and Kenneth E. Barron. "Testing moderator and mediator effects in counseling psychology research." *Journal of counseling psychology* 51.1 (2004): 115.

Gans, Joshua S., and Scott Stern. "The product market and the market for "ideas": commercialization strategies for technology entrepreneurs." *Research policy* 32.2 (2003): 333-350.

Gittelman, Michelle, and Bruce Kogut. "Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns." *Management Science* 49.4 (2003): 366-382.

Gittelman, Michelle. "The revolution re-visited: Clinical and genetics research paradigms and the productivity paradox in drug discovery." *Research Policy* (2016).

Gunther McGrath, Rita, and Atul Nerkar. "Real options reasoning and a new look at the R&D investment strategies of pharmaceutical firms." *Strategic Management Journal* 25.1 (2004): 1-21.

Henderson, Rebecca, and Iain Cockburn. "Scale, scope, and spillovers: the determinants of research productivity in drug discovery." *The Rand journal of economics* (1996): 32-59.

Hitt, Michael A., Robert E. Hoskisson, and Hicheon Kim. "International diversification: Effects on innovation and firm performance in product-diversified firms." *Academy of Management journal* 40.4 (1997): 767-798.

Hitt, Michael A., Robert E. Hoskisson, and R. Duane Ireland. "A mid-range theory of the interactive effects of international and product diversification on innovation and performance." *Journal of management* 20.2 (1994): 297-326.

Hoskisson, Robert E., et al. "Strategy in emerging economies." *Academy of management journal* 43.3 (2000): 249-267.

Kapoor, Rahul, and Thomas Klueter. "Decoding the adaptability–rigidity puzzle: Evidence from pharmaceutical incumbents' pursuit of gene therapy and monoclonal antibodies." *academy of management journal* 58.4 (2015): 1180-1207.

Lander, Eric S., et al. "Initial sequencing and analysis of the human genome." *Nature* 409.6822 (2001): 860-921.

Li, Danielle, Pierre Azoulay, and Bhaven N. Sampat. "The applied value of public investments in biomedical research." *Science* 356.6333 (2017): 78-81.

March, James G. "Exploration and exploitation in organizational learning." *Organization science* 2.1 (1991): 71-87.

McGrath, Rita Gunther. "A real options logic for initiating technology positioning investments." *Academy of management review* 22.4 (1997): 974-996.

Nagaraj, A. (2015, November 6). The Private Impact of Public Maps— Landsat Satellite Imagery and Gold Exploration. Job Market Paper, MIT Sloan School of Management, Cambridge, MA. Available: [http://web.mit.edu/nagaraj/files/nagaraj\\_jmp\\_nov6.pdf](http://web.mit.edu/nagaraj/files/nagaraj_jmp_nov6.pdf)

Nelson, Richard R., and G. Sidney. "Winter. 1982." *An evolutionary theory of economic change* (1982): 929-964.

Nelson, Richard R., and G. Sidney. "Winter. 1982." *An evolutionary theory of economic change* (2005): 929-964.

Nightingale, Paul. "A cognitive model of innovation." *Research policy* 27.7 (1998): 689-709.

Rumelt, Richard P. "Diversification strategy and profitability." *Strategic management journal* 3.4 (1982): 359-369.

Sampat, Bhaven, and Heidi L. Williams. How do patents affect follow-on innovation? Evidence from the human genome. No. w21666. National Bureau of Economic Research, 2015.

Scannell, Jack W., et al. "Diagnosing the decline in pharmaceutical R&D efficiency." *Nature reviews Drug discovery* 11.3 (2012): 191-200.

Scannell, Jack W., et al. "Diagnosing the decline in pharmaceutical R&D efficiency." *Nature reviews Drug discovery* 11.3 (2012): 191-200. Tripp, Simon, and Martin Grueber. "Economic impact of the human genome project." Battelle Memorial Institute (2011).

Teece, David J., Gary Pisano, and Amy Shuen. "Dynamic capabilities and strategic management." *Strategic management journal* (1997): 509-533.

Teodoridis, Florenta, Keyvan Vakili, and Michaël A. Bikard. "Can Specialization Foster Creativity? Mathematics and the Collapse of the Soviet Union." *Academy of Management Proceedings*. Vol. 2017. No. 1. Academy of Management, 2017.

Utterback, James M., and Fernando F. Suárez. "Innovation, competition, and industry structure." *Research policy* 22.1 (1993): 1-21.

Utterback, James. "Mastering the dynamics of innovation: how companies can seize opportunities in the face of technological change." (1994).

Vincenti, Waler G. "What Engineers Know and How They Know It Analytical Studies From Aeronautical History." (1990).

Wernerfelt, Birger. "A resource-based view of the firm." *Strategic management journal* 5.2 (1984): 171-180.

Williams, Heidi L. "Intellectual property rights and innovation: Evidence from the human genome." *Journal of Political Economy* 121.1 (2013): 1-27.

Zahra, Shaker A. "Technology strategy and financial performance: Examining the moderating role of the firm's competitive environment." *Journal of Business venturing* 11.3 (1996): 189-219.

Zott, Christoph, and Raphael Amit. "The fit between product market strategy and business model: implications for firm performance." *Strategic management journal* 29.1 (2008): 1-26.

Zucker, Lynne G., and Michael R. Darby. "Capturing technological opportunity via Japan's star scientists: Evidence from Japanese firms' biotech patents and products." *The journal of Technology transfer* 26.1-2 (2001): 37-58.

Zucker, Lynne G., Michael R. Darby, and Marilyn B. Brewer. Intellectual capital and the birth of US biotechnology enterprises. No. w4653. National Bureau of Economic Research, 1994.

## APPENDIX

**Table A:** List of main diseases categorized based on the Cortellis database clinical disease hierarchy. 881 different diseases were grouped into 27 broad categories which were used to calculate the Herfindahl-Hirschman Index for each firm.

Main Disease Category	Number of Specific Indications
Infectious Disease	179
Cancer	89
Neuro	81
Uncategorized	55
GI	55
Cardio	53
Hematological	47
Endocrine Disease	34

Dermatological	33
Metabolic	30
Genitourinary	30
Respiratory	25
Ocular Disease	25
Gynecology Obstetrics	25
Immune	23
Musculoskeletal Disease	22
Inflammatory Disease	14
Psychiatric Disorder	13
Toxicity Intoxication	11
Genetic Disorder	11
Andrology	9
Injury	6
Otorhino	4
Nutritional	3
Mouth Disease	2
Fatigue	1
Growth Disorder	1

**Table B:** List of MeSH keywords used to select target-based publications assigned to firms

Base Sequence

Genes

Catalytic Domain

Crystallography, X-Ray

Chemical Synthesis

DNA, Complementary/genetics

Models, Molecular

Molecular Conformation

Molecular Structure

Algorithms\*

Combinatorial Chemistry Techniques

Computer Graphics

Computer Simulation

Drug Design\*

Models, Genetic

Models, Molecular

Molecular Structure

Mutation

Models, Chemical

Pharmaceutical Preparations/chemical synthesis\*

Biopolymers/chemistry

Database Management Systems

Models, Molecular\*

Solvents

Structure-Activity Relationship

CDC2-CDC28 Kinases\*

Combinatorial Chemistry Techniques

Crystallography, X-Ray

Drug Design

Enzyme Inhibitors/chemical synthesis

Enzyme Inhibitors/chemistry\*

Fluorenes/chemical synthesis

Fluorenes/chemistry\*

Isoindoles

Magnetic Resonance Spectroscopy

Models, Molecular

Protein Binding

Protein-Serine-Threonine Kinases/chemistry

Proto-Oncogene Proteins\*

Pyridines/chemical synthesis

Pyridines/chemistry\*

Structure-Activity Relationship

Urea/analogs & derivatives\*

Urea/chemical synthesis

Urea/chemistry\*

Biotechnology\*

Centrifugation

Computer Simulation\*

Computer-Aided Design

Models, Biological\*

Software\*

Software Design

### **3. SCIENTIFIC MAPS AND EXPLORATION: TRACKING TECHNOLOGICAL SEARCH TRAJECTORIES IN DRUG DISCOVERY**

#### **Abstract**

Innovation scholars theorize that explorative search can lead to novel knowledge recombination and valuable outcomes, but this type of search is both costly and risky. How scientific maps influence firm exploration is not well established. In this study, I explore how the human genome map impacted firms' exploration for novel compounds. I build on prior methodological advances in the innovation literature and computational chemistry to introduce a novel technique to measure technological distance between patents based on chemical structures. This technique is applied on a sample of small molecule drug patents to examine exploration trajectories over a 14-year time period. Analyses show that overall firm search trajectories become narrower over time, and that targeted strategies have broader exploration compared to non-targeted strategies. By capturing firm level technological search trajectories over time, this study provides insights on the evolution of firm exploration and the effect of a scientific map on search trajectories.

Keywords: technological search, technological distances, chemical similarities, drug discovery

## Introduction

The process of innovation is conceptualized as a spatial process (technological search), where firms and inventors search either locally or in unknown, new areas (Cyert & March, 1963; March, 1991). Innovation theorists suggest that broad exploration can lead to new knowledge and increase technical novelty (Nelson & Winter, 1982). Innovation theorists suggest that scientific knowledge can aid in the navigation of such complex landscapes, enable predictive capabilities and reduce uncertainty in the search for new products (Fleming & Sorenson, 2001). In the context of small molecule drug discovery, searching for an optimal compound in chemical space is time-consuming and complex (Scannell, et al, 2012). This is mainly because chemical space is a high-dimensional search space with  $10^{60}$  possible combinations of novel compounds. To empirically test the influence of a map on technological search trajectories, I build on existing theories in the innovation literature and introduce a novel technological distance measure to capture firm exploration over time.

In 2000, the Human Genome Project (HGP), the world's largest publicly-funded biology project, made available the first draft of the human genome. The human genome is a digital map of 3 billion DNA base pairs revealing the location and identity of genes which encode various proteins in cells. The human genome map provides a list of about 10,000 potential disease targets (Drews, 2000). Given a drug target, medicinal chemists can design and synthesize small molecules or compounds that can bind to it. To do this, they comb through a large theoretical landscape of solutions called chemical space -

containing  $10^{60}$  possible combinations. So how does a scientific map like the human genome map influence how scientists search for new compounds?

In a high-dimensional search space like chemical space, which has a large number of possible solutions identifying the correct or even a set of optimal solutions is a hard task. Without a map to navigate this massive landscape, medicinal chemists would rely on disease knowledge and prior experience to make calculated guesses on the types of chemicals that could work (Scannell, 2012). In contrast, having the human genome map allowed scientists to construct detailed models of disease targets using computational tools. This drastically reduced the search space (chemical space) of possible options, making it possible for the medicinal chemists to sample a reduced search space and design compounds that could fit precisely with the target (Lipinski & Hopkins, 2004; Gittelman, 2016). Therefore, the availability of a scientific map like the human genome had important consequences for the nature of drug discovery and its outcomes.

I use this context of small molecule drug discovery, to test the effect of the human genome map on firm exploration for new drugs. My dataset is comprised of a special type of drug patent called Markush or broad chemistry patents. In drug discovery, Markush patents are very specific to small molecule drugs and mark the starting point where firms begin to claim intellectual property rights in chemical space. An inventor or firm uses a Markush patent to make general claims for use of a molecule without revealing its exact structure – this compound is hidden among hundreds of other similar looking compounds. Thus, these compounds within Markush patents represent the explorative

effort undertaken by the firm or inventor. I exploit these unique features of Markush patents and the “flags” firms plant in chemical space to measure their exploration over time. For this, I introduce a novel technique from computational chemistry to capture chemical structural similarities between patents (Johnson & Maggiora, 1992). I test this method on a sample of Markush patents to identify firm level search trajectories (exploration) across time periods and based on search strategies. Results presented in this study are exploratory and inform on the nature of technological search and exploration when impacted by scientific maps.

### **Background on Technological Search and Distances**

Innovation scholars theorize invention as a search process over technological landscapes combining new and existing knowledge and technologies (Henderson & Clark, 1990). This search is tied to the generation of technical novelty and driven by new knowledge and technology, which can lead to a better chance of market selection (Nelson & Winter, 1982). In this search for technical novelty, firms can both explore new technological space and/or exploit prior knowledge (March, 1991). Explorative search can lead to novel knowledge recombination and valuable outcomes, but this type of search is both costly and risky. The balance depends partly on the relative costs of exploration and exploitation and the ability to apply prior expertise as a useful input for future projects. Some innovation scholars argue that firms tend to search locally, enter markets related to their capabilities, and use prior expertise to select future projects (Sorenson & Fleming, 2004; Stuart & Podolny, 1996; Helfat & Raubitschek, 2000). Thus, exploration in knowledge

landscapes is conditioned by what firms already know and the relative costs of undertaking the search.

The process of drug discovery can be conceptualized as a Kauffman fitness landscape, where two main components (biological gene targets and chemical agents;  $N=2$ ) interact with high-interdependence ( $K$ ). In this landscape multiple compounds can interact with varying efficacy with one target or more targets, thus making the interactions multiplex, interdependent and overlapping. Interdependence or coupling between components occurs where changes made to one component requires changes to another for the system to work properly (Ulrich, 1995; Fleming & Sorenson, 2001). In the case of drug discovery, this implies that disease target and designed compounds are tightly inter-related, and that deep knowledge and prior experience can influence how technological exploration occurs.

Empiricists utilize a range of tools and methods to probe how firms navigate this technological landscape. Technological positions are identified, distances between positions measured and high dimensional search spaces are computationally modeled. In the innovation literature exploration in technological space is captured through various measures like SIC codes, networks, patent classes, patent citations and recently, topic modeling. Stuart & Podolny (1996) use network ties to track the evolution of firms' technological positions. Ahuja & Katila (2002) follow patent citations and patterns to identify search scope and depth in technological search. Rosenkopf & Nerkar (2001) have examined exploration across organizational and technological boundaries using the

3-digit technical classes in patents. Yoon & Kim (2012) utilize the text description of patents to extract technological positions between patents. Tzabbar (2009) uses vector distances calculated from patent technology classes to identify technological repositioning in firms. Recently, Aharonson & Schilling (2016) measure technological distance between patents using vector distance computed from the patent subclasses. Kaplan & Vakili (2015) use topic modeling of patent text to examine technological distances and breakthrough innovation.

Thus, a range of measures have been developed and established to capture exploration in technological space. And most of these methods develop distance measures based on citations or patent technology classes. To classify patents, patent offices have categories of patent classes or subclasses that examiners assign new patents to. Hence, a semiconductor chemistry patent will be assigned differently than a drug chemistry patent. Using patent classes to differentiate between these two patents or measure distance using vectors works well in this case. Most of the patent innovation literature builds on these methodical advances to empirically capture firm and industry exploration over time.

But if we would like to measure exploration between firms in the same industry, operating in similar product markets, this approach of using patent classes becomes less reliable. For example, if we compare technological distance or exploration for two drug companies making small molecule colon cancer drug their patents classes will overlap for the most part, indicating close distance between the two patents. This makes sense as both firms operate in similar areas. But what if the two drugs operated on completely

different drug targets or employed vastly different chemical structures that are not related? Drug industry experts would categorize these patents as being distant, while their patent classes would make them appear close. This also has implications for studying longitudinal aspects of exploration.

Drug firms that only specialize in small-molecule drugs or work in a specific disease area will show no exploration if we rely only on patent classes or subclasses. Even if these firms are using new methods, disease targets or exploring completely new chemical scaffolds to design their drugs. We risk losing this granularity and nuance by aggregating patents at the class and subclass level, even though patent class-based differences are an excellent and robust method to capture distance at the industry level. For example, how does one test if Firm A's compounds in 2004 look similar to prior years or if they made completely new structures. This question is similar to examining Intel's processors from different time periods: how different are the Xeon processors from the old Pentium processors. They are in the same product category - processors, but the underlying architecture and design is very different - implying broad exploration in their innovative processes. To address issues like this, I introduce a granular technique drawing from computational chemistry. I exploit the richness of chemistry patents and the chemical structures contained within them to determine distance between patents and firm level exploration. By examining differences in chemical structures (between patents and firms) I can reconstruct their search paths in chemical space over time. A full description of this method is discussed below.

## Data

Patents are central to intellectual property protection and appropriation in biopharmaceutical industry and widely used in economic analyses (Scott & Sampat, 2012). As of 2016, more than 90% of marketed drugs are small molecules making chemistry-based patents relevant for this analysis. I utilize a novel dataset comprised of a special type of chemistry drug patent called Markush patents to test my method and reveal insights on firm level exploration.

An inventor or firm uses a Markush patent to make general claims for use of a molecule without revealing its exact structure – this compound is hidden among hundreds of other similar looking compounds.

*Markush or Broad chemistry patents:* Markush patents are used across industries which work with chemicals like drugs, materials science, industrial chemicals and agrobiotechnology. These patents contain a special structure called Markush structure, named after Eugene Markush, to capture a broad set of compounds. This type of patents is filed very early in search process where a firm makes general claims for use of a molecule without revealing the exact compound they are pursuing. The compound that could make it to market is hidden among hundreds of other similar looking compounds. In the drug discovery stage, Markush patents are very specific to small molecule drugs (compared to biologics or process patents) and mark the starting point where firms begin to claim intellectual property rights in chemical space (Southall & Ajay, 2005). Follow on drug patents and granted patents are based on this original Markush patent.

Another interesting feature of Markush patents is that they encapsulate the actual compounds made, and those that the firm or inventor plans to protect for future use – sometimes, running into the hundreds or thousands of novel compounds. Thus, these compounds within Markush patents represent the explorative effort undertaken by the firm or inventor in chemical space. While granted patents or single compound patents that cover a drug (like Viagra) are important for intellectual property protection, they do not indicate the original explorative effort undertaken by the firm. Invariably, all such granted patents will reference the starting Markush patent or application. While not all Markush patent applications end up becoming granted patents or lead to a drug, they represent capture all small-molecule related R&D activity and explorative research – not just the ones that become successful.

I exploit these unique features of Markush patents and the “flags” firms plant in chemical space to measure their exploration over time. This non-bias in the nature of Markush patent applications make them especially useful for analyzing overall technological search trajectories of firms and estimating the broad effects of technological change. These unique properties of Markush patents make them well suited to study firms’ technological search trajectories over time. For my empirical research, I have collected 39,000 drug related Markush patent applications filed in the WIPO patent office from 1990-2004, and extracted millions of novel compounds embedded in them.

*Data source and collection:* The source for Markush patents is the Chemical Abstract Society's (CAS) Scifinder product. Scifinder is the world's largest repository of chemical structures, published articles and patents and provides access to MARPAT – a comprehensive database of Markush patents that cover all 9 major patent offices and 63 patent authorities worldwide. More than 1 million Markush structures and about 481,000 Markush patents and applications from 1988-present are available for searching. In a recent comparative analysis of patented compound databases, CAS's Scifinder was found to be more comprehensive and accurate compared to the Derwent World Patents Index and Reaxys (Ede, et al, 2016).

A full discussion of data sampling and Markush patents is provided in essay one of the dissertation. To measure technological distance, a sample of 21,334 drug-related Markush patents and patent applications filed between 1990-2004 are used. This sample includes 570 firms.

*Data source for Chemical Structures:* SureChEMBL is an open access database that contains compounds extracted from the full text, images and attachments of patent documents. The SureChEMBL database contains 17 million compounds extracted from 14 million patent documents belonging to the US, European and WIPO patent office (Papadatos, et al, 2015). The extraction of chemical content is performed in an automated way by mining patent text, images and associated structure files for compound information. Methods like name-to-structure and image-to-structure conversion are applied to capture patented compounds. The extracted compounds are available in an open data format known as SMILES that allows distance measures (Weininger, 1988).

The entire compound database of SMILES strings was downloaded and linked to the Markush patents using unique patent identifiers assigned by the patent offices (e.g. WO-1234567-A1 or US-5678901). Using this concordance, Markush patent compounds for firms in the dataset were extracted and organized by year.

*ChemmineR tools software:* This is an open source cheminformatics toolkit created by an academic group (Cao, et al, 2008). The software provides tools for data conversion, compound mining, structural similarity searching and clustering of small molecules. It is written in R and C++ and is among the fastest implementations for chemical structure comparisons. This software was used to cluster the patented compounds for the time period 1990-2004. These are computationally intensive processes and can take a few days to a week to run on a standard single-processor laptop for a single firm.

## **Method**

### **Technological Distance Measure based on Chemical Structural Similarity**

I introduce a novel technological distance measure using the compounds filed by the firms and inventors in the Markush patents. To understand how and what is measured, we need to understand the nature of the solution space – the region from where the solution to a problem exists. In the case of small molecule drug discovery where finding an optimal compound for a disease target is the problem, the solution space is the entire universe of compounds, called chemical space.

Chemical Similarity: Chemical space is immense – more than  $10^{60}$  possible combinations for just known organic compounds, and all known drugs inhabit this space and occupy specific coordinates defined by chemical structure. The drug Fentanyl occupies a very different location than Aspirin. An important concept in chemistry is that structurally similar looking compounds have similar properties (Johnson & Maggiora, 1992). This similarity between chemical structures is captured as the inverse of the distance in chemical space (further apart indicates less similarity, just as Aspirin's properties are quite different from Fentanyl's). One such standard distance measure is the Tanimoto similarity, a form of Jaccard coefficient<sup>4</sup>, widely used in chemical informatics and drug discovery for chemical structure similarity measurements (Bajusz, et al, 2015). Recently, Krieger et al (2017) have used the Tanimoto similarity measure to capture novelty of new drugs by comparing to previously approved drugs.

Tanimoto similarity coefficient: The Tanimoto similarity is central to this method and captures distance between compounds.

“The Tanimoto coefficient is defined as  $c/(a+b+c)$ , which is the proportion of the features shared among two compounds divided by their union. The variable  $c$  is the number of features (or on-bits in binary fingerprint) common in both compounds, while  $a$  and  $b$  are the number of features that are unique in one or the other compound, respectively. The

---

<sup>1</sup> The Jaccard index or similarity coefficient is a statistic to measure the overlap that two sample sets share in their attributes.

Tanimoto coefficient has a range from 0 to 1 with higher values indicating greater similarity than lower ones.” – ChemMine Tools tutorial

The Tanimoto coefficient between two points,  $a$  and  $b$ , with  $k$  dimensions is calculated as:

$$\frac{\sum_{j=1}^k a_j \times b_j}{(\sum_{j=1}^k a_j^2 + \sum_{j=1}^k b_j^2 - \sum_{j=1}^k a_j \times b_j)}$$

In the Tanimoto approach, compound structures are converted into 2-D fingerprints, a string of unique bits, and compared for similarity (Lipinski, 2000; Lipinski & Hopkins, 2004; Johnson & Maggiora, 1992). A well established 2-D format in the cheminformatics area is known as the simplified molecular-input line-entry system (SMILES). The SMILES notation facilitates accurate capture of compounds and efficient search methods (Weininger, 1988). See Figure 1 for calculation of chemical fingerprints and Tanimoto similarity.

### Tanimoto coefficient

$$S_{AB} = \frac{c}{(a+b-c)}$$

a: number of "on" in compound A

b: number of "on" in compound B

c: number of "on" in both compound A and B

### Pairwise Tanimoto similarity

	A	B	Z
A	1	0.57	0.50
B	0.57	1	0.86
Z	0.50	0.86	1

### Partial list of fingerprints for 3 compounds

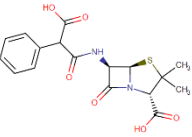
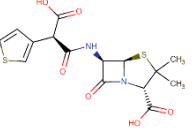
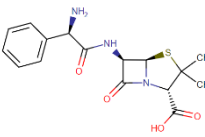
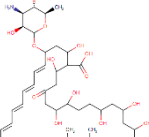
	Bit 1	Bit 2	Bit 3	Bit 4	...	...	Bit 1024
A	0	1	0	1	0	1	1
B	1	1	1	1	1	0	1
Z	0	0	0	1	0	1	0

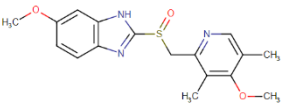
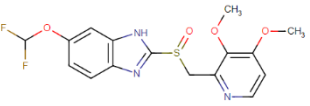
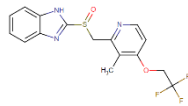
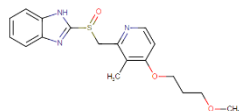
**Figure 1:** Chemical similarity used as a measure of distance between compounds (figure source: Jeliaskowa & Jaworska, Proctor & Gamble, 2005)

### Examples of Chemical Similarities using Tanimoto Coefficient

To understand how the Tanimoto coefficient measure can be applied to identifying chemical similarities, I

Present below two well-known classes of chemical drugs – antibiotics and proton-pump inhibitors and their differences. Proton-pump inhibitors are gastrointestinal drugs used to reduce gastric acid production.

Antibiotics			
 <p>Carbenicillin</p>	 <p>Ticarcillin</p>	 <p>Ampicillin</p>	 <p>Nystatin</p>

<b>Proton-Pump Inhibitors (PPI)</b>			
 Omeprazole	 Pantoprazole	 Lansoprazole	 Rabeprazole

**Figure 2:** Structural comparison of compounds for two classes of drugs – Antibiotics and Proton-Pump Inhibitors.

The antibiotic compounds Carbenicillin, Ticarcillin and Ampicillin look closer to each other in structure compared to Nystatin. Based on simple visual similarities, we can expect their Tanimoto coefficients to reflect their structural similarity (See top row Figure 2). For the PPI compounds, the Omeprazole and Pantoprazole look similar compared to the other two compounds (See bottom row Figure 2). The calculated pairwise Tanimoto coefficients for these 8 compounds are shown below. For example, Carbenicillin is compared to Ticarcillin, Ampicillin and so on and its coefficients recorded.

#### Pairwise Chemical Similarity Scores using Tanimoto Atom Pair Similarity

Compound Name	Carbenicillin	Ticarcillin	Ampicillin	Nystatin
Carbenicillin	1	0.79	0.72	0.06

Ticarcillin	0.79	1	0.58	0.04
Ampicillin	0.72	0.58	1	0.05
Nystatin	0.06	0.04	0.05	1

### Proton Pump Inhibitors

Compound Name	Omeprazole	Pantoprazole	Lansoprazole	Rabeprazole
Omeprazole	1	0.51	0.38	0.41
Pantoprazole	0.51	1	0.41	0.38
Lansoprazole	0.38	0.41	1	0.53
Rabeprazole	0.41	0.38	0.53	1

### Comparison of

#### Both

Compound Name	Carbenicillin	Ticarcillin	Omeprazole	Pantoprazole
Carbenicillin	1	0.79	0.11	0.11
Ticarcillin	0.79	1	0.09	0.09
Omeprazole	0.11	0.09	1	0.51
Pantoprazole	0.11	0.09	0.51	1

**Table 1:** Pairwise Tanimoto coefficients shown for each compound. Scores above 0.60 are highlighted. In the third table, two representatives of each group are selected and

compared to show differences in similarity scores. Tanimoto coefficients were calculated using the software ChemmineR Tools.

The coefficient values in Table 1 correlate with the structural differences that can be observed in Figure 2. Carbenicillin is closest in structure to Ticarcillin and Ampicillin; the Tanimoto coefficients are 0.79 and 0.72 (closer to 1 means more similar). Similarity cut-offs above 0.60 are a recommended threshold to identify similar looking compounds. Thus, we see 3 antibiotics (Carbenicillin, Ticarcillin and Ampicillin) and 2 PPI (Omeprazole and Pantoprazole) drugs that appear closer in distance in chemical space. When sample compounds of each group are compared (Group 3), we see the Tanimoto coefficients reduce further, thereby implying more distance between the compound structures for these two different classes of drugs.

Clustering of chemical compounds: For pairwise similarities between compounds, applying the Tanimoto coefficient is a reliable method in chemical informatics. To calculate distances for large numbers of compounds (as in a patent), these compounds are compared using a clustering technique. Using the pairwise Tanimoto distances, these 8 compounds can be clustered to identify similar clusters of compounds. For this, a clustering approach called binning clustering from ChemmineR Tools is used.

“Binning clustering assigns compounds to similarity groups based on a user-definable similarity cutoff. For instance, if a Tanimoto coefficient of 0.6 is chosen then compounds

will be joined into groups that share a similarity of this value or greater using a single linkage rule for cluster joining.” – ChemmineR tutorial

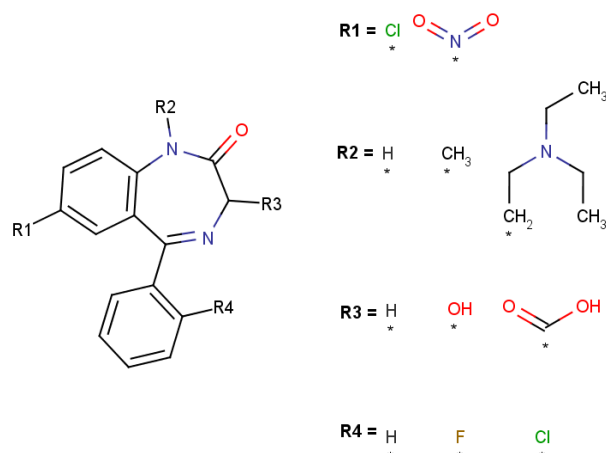
Results of clustering the 8 compounds are shown in Appendix Figure B. The three antibiotics Carbenicillin, Ticarcillin and Ampicillin fall in the same cluster, while all the other compounds appear in separate, independent clusters. This technique can be used to identify clusters of structurally similar compounds. The ChemmineR software is implemented for high-throughput clustering and is very efficient at calculating pairwise Tanimoto coefficients.

#### *Applying Tanimoto Coefficients to Measure Chemical Distance Between Markush*

##### *Patents:*

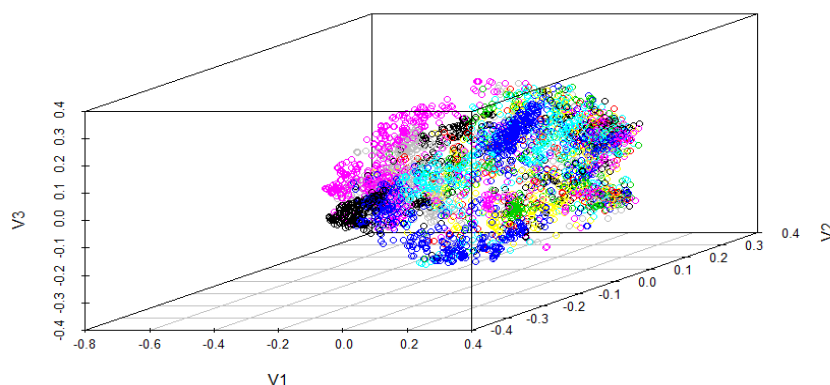
Broad chemistry or Markush patents contain unique compounds encoded in the description and methods section based on a central design known as a chemical scaffold. This generic chemical scaffold is also called a Markush structure first appeared in 1923 and is named after its inventor, Eugene Markush.

Compounds listed in a broad chemistry patent have to be unique (exact similarity to any prior known compound cannot be patented), so the complete set of listed patent compounds constitutes a unique and diverse set of molecules (see Figure 3).



**Figure 3:** Example Markush structure with R groups representing various chemical groups. Each such Markush structure can represent hundreds of actual compounds (image source: ChemAxon Inc.)

Using the SureChEMBL database of patents and patented compounds, novel compounds contained within each Markush patent in my sample was collected and stored into separate firm specific files that could be processed. The ChemMine Tools software is applied on this data to calculate Tanimoto similarity scores for the exemplified compounds in the Markush patents. Compounds for a set of Markush patents are clustered based on the similarity cut-off (0.6). Compounds that fall in the same cluster are structurally similar and closer in chemical space. These clusters of exemplified compounds of different Markush patents are then examined for their source (firm names), year and patent identifiers. Patents that contain structurally similar compounds will show up occupying similar regions in chemical space. Figure 4 is a clustering of patented compounds for one company across several years.



**Figure 4:** Tanimoto similarity coefficient-based clustering of Bristol-Myer Squibb's Markush compounds between 1989 and 1997. Colored groups represent similar clusters, for example pink dots represent a neighborhood of chemical space occupied by structurally similar compounds. Image created using ChemmineR and R Studio.

Using these Tanimoto clustering measures, we can now look at measuring the distance between patents at the firm level. For this, I apply set theory logic of intersections and overlaps between sets of objects.

*Distance Measure Between Patents Using Tanimoto Coefficients:* The compound clusters for a pair of patents represents a set of objects that can be used estimate distance between the two patents. If two patents have more compound clusters in common, then building from Tanimoto coefficient logic, we can infer that these two patents (and their compounds) are closer in chemical space. For two patents that represent different regions of chemical space, we do not expect to see high overlap in clusters.

To calculate this overlap, a measure called Jaccard index is used. A Jaccard Index is the intersection of two sets divided by the union.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

This measure introduced by Paul Jaccard, is a statistic used to compare similarity and diversity of sample sets. Application of this measure is found in object recognition in computer vision and machine learning algorithms. The Jaccard Distance which measures dissimilarity between sample sets (in this case compounds) is complementary to the Jaccard Index (JI), obtained by subtracting the JI from 1.

Thus, Jaccard Distance = 1 – Jaccard Index

For a given pair of Markush patents, the Jaccard Distance (JD) represents a distance in chemical space (0 to 1) calculated using the set objects (in this case, the compound clusters). A score of 0 indicates perfect similarity while a distance of 1 implies complete dissimilarity. Two patents with a score of 1 imply no compound structural features in common, implying a novel Markush scaffold. This would represent a new region in chemical space and thus, exploration by the firm.

*Example of Chemical Distance Between Two Markush Patents using Jaccard Index:*

Two patents belonging to the same firm, Pfizer, were selected for this analysis. The patents are two years apart and related to the same class of drugs – Phosphodiesterase (PDE) Inhibitors. This class of drugs contains drugs like Viagra, a Pfizer product and

PDE inhibitor. The number of exemplified compounds in each patent is also provided. Given the similarity in scope and class of drug target, these two patents have a higher probability of having similar looking compounds. The third patent is a completely different class of drugs known as Opiates (like Fentanyl). This is added as a control to test this measure of distance.

<b>Patent Application Number</b>	<b>Title</b>	<b>Priority Application Year</b>	<b>Patent Assignee</b>	<b>Com- pounds</b>	<b>Type</b>
WO- 200102711 3-A2	Preparation of 5-(3-pyridyl)- substituted pyrazolo[4,3- d]pyrimidinones as phosphodiesterase inhibitors.	1999	Pfizer Limited, UK	18 4	<i>PD E Inhi- bitor</i>
WO- 200207477 4-A1	Preparation of pyrazolopyrimidines as cyclic guanosine 3',5'-monophosphate phosphodiesterase inhibitors for sexual dysfunction.	2001	Pfizer Limited, UK	77	<i>PD E Inhi- bitor</i>

WO- 200303562 2-A1	3-Azabicyclo[3.1.0]hexane derivatives as opioid receptor antagonists.	2001	Pfizer Product s Inc., USA.	11 9	<i>Opi ate</i>
--------------------------	---	------	--------------------------------------	---------	--------------------

**Table 2:** Three sample Markush Patents belonging to Pfizer are used to measure distance

Using ChemMine Tools, each of the compounds in the two PDE inhibitor patents are compared structurally to calculate Tanimoto similarity scores. These scores are then used to cluster the compounds. Each cluster is then analyzed to generate a Jaccard Index for a patent pair. Shown below is a table used to derive the Jaccard Index.

	Cluster Ids																	
<b>Patent Applicatio n Number</b>	2 1 5 7	6 9 4 4	4 0 2 8		2 0 5 4	4 0 3 4	4 0 0 9	4 0 2 7	4 0 1 2	4 0 3 9	4 0 1 8	2 0 6 5	4 0 2 2	4 0 3 6	2 0 6 7	2 0 1 9	2 0 0 0	<b>Overlap</b>
WO- 20010271 13-A2	4	1	2	1	7	1	1	1	1	1	1	2	1	1	4	1	1	31

WO- 20020747 74-A1																			
	4	11	2	3	2	1	1	1	1	1	1	4	1	1	1	1	1		37

**Table 3:** Application of the clustering method to two patents

Jaccard Index = Intersection of Sets / Union of Sets

Jaccard Index (WO-2001027113-A2, WO-2002074774-A1) = Total Compounds in

Intersection/ Total Compounds in Both Patents

Jaccard Index (WO-2001027113-A2, WO-2002074774-A1) = 68/261 => 0.26

Therefore, Jaccard Distance (WO-2001027113-A2, WO-2002074774-A1) = 1 – 0.26 =

**0.74 or 74%** dissimilar.

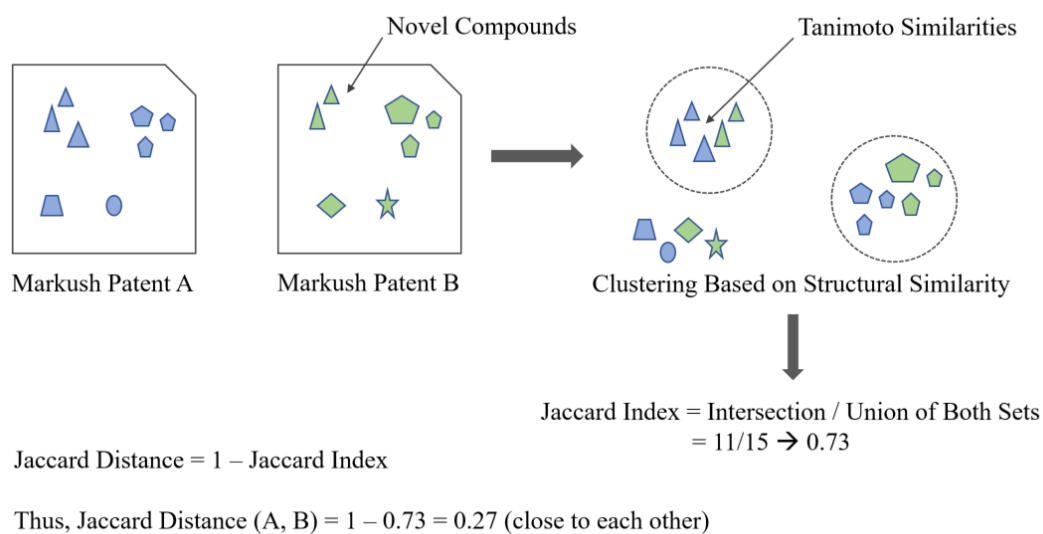
A similar table was created to measure the distance between a PDE inhibitor patent (WO-2001027113-A2) and the opioid receptor patent (WO-2003035622-A1). Only one cluster was found in common between these two patents containing 16 compounds (out of a total of 184 + 119 compounds). Thus,

Jaccard Index = 16/303 => 0.05

Jaccard Distance (WO-2001027113-A2, WO-2003035622-A1) = 1 – 0.05 = **0.95 or 95%** dissimilar.

Thus, this method shows that the PDE inhibitors are closer in chemical space compared to their distance to the Opiate patent, controlled for the same firm. Figure 5 provides an

overview of the chemical distance measurement for any two patents containing compounds.



**Figure 5:** Overview of technological distance measure using chemical structural similarities for two patents containing compounds

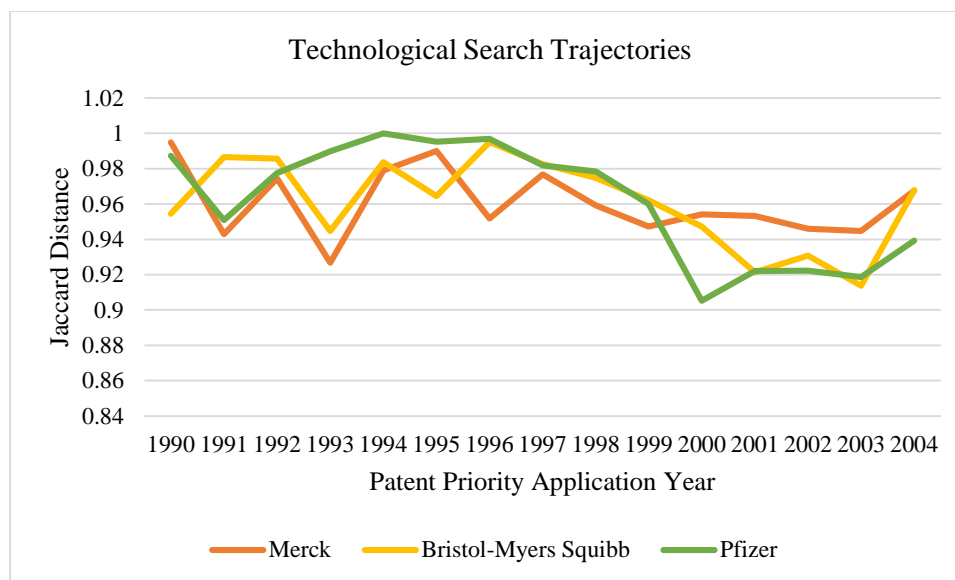
*Firm-level chemical distances:* For each firm in the sample, Markush patents are collected and organized by year. The first year in the sample is marked as the reference year. The compounds that were created in this year are marked in chemical space as “reference markers” – base to which subsequent patents are measured against. Every novel compound in every subsequent patent is measured against these reference marker compounds and their distance in chemical space recorded using Tanimoto coefficients. Building up from the pairwise compound similarities and cluster similarities, the average chemical distance is calculated for each year.

Two versions of firm-level distances have been implemented: a) distance to prior year and b) cumulative distance. In the distance to prior year, the reference markers are updated to the prior year's compounds in chemical space. In the cumulative distance, all prior years' compounds are marked for comparison. These two approaches provide a finer-grained approach to measuring exploration in relation to previous year's search behavior and cumulative search behavior over a period of time.

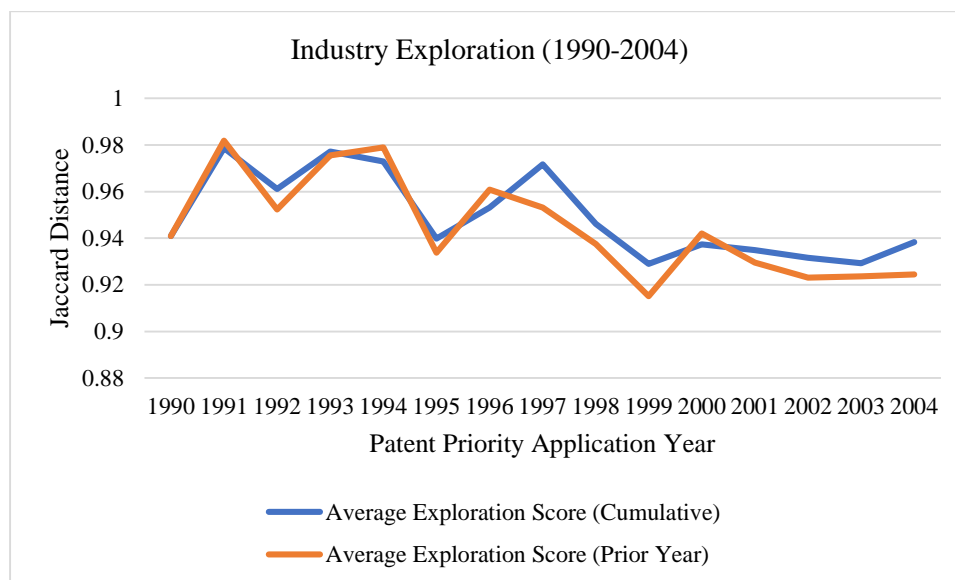
## **Results**

### **Application of the Chemical Distance Method to Small Molecule Markush Patents**

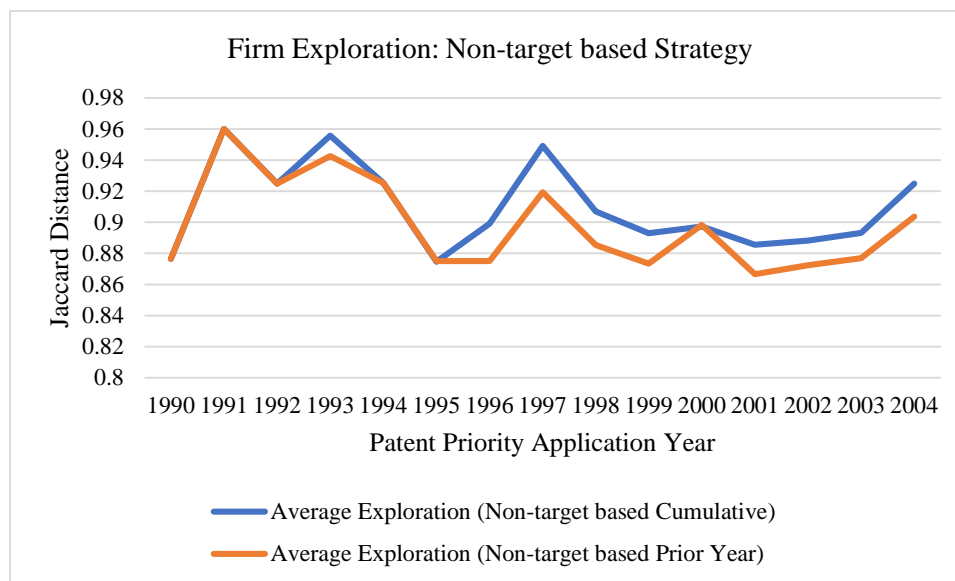
I apply the technological distance measure to a sample of Markush drug patent applications and granted patents using the chemistry-based method introduced here. The exploration scores are aggregated by year, drug discovery strategy and trends for cumulative and prior year are compared (see Figures 6-10).



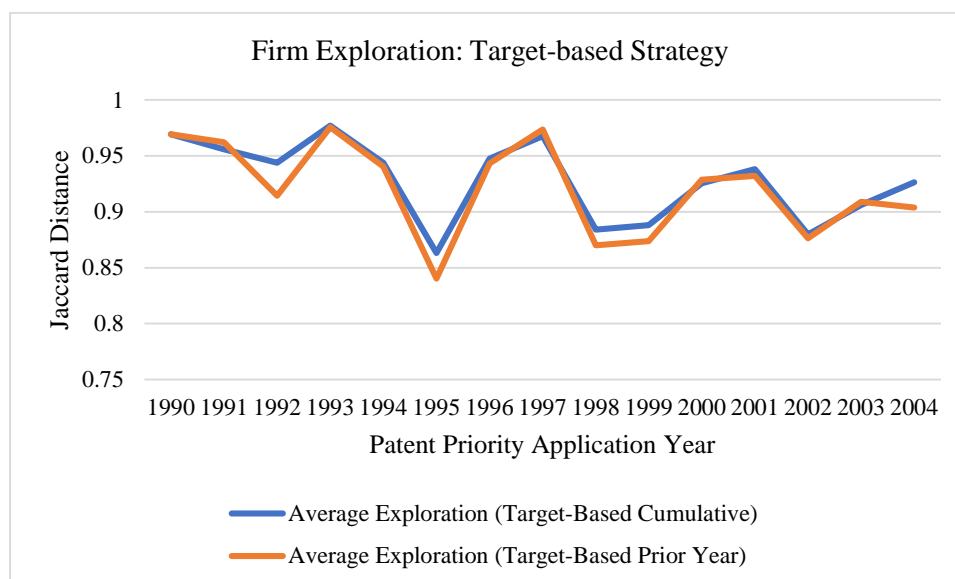
**Figure 6:** Firm exploration calculated using chemical structural similarities for 3 firms between 1990-2004. Average cumulative exploration scores for each patent are grouped by year and compared.



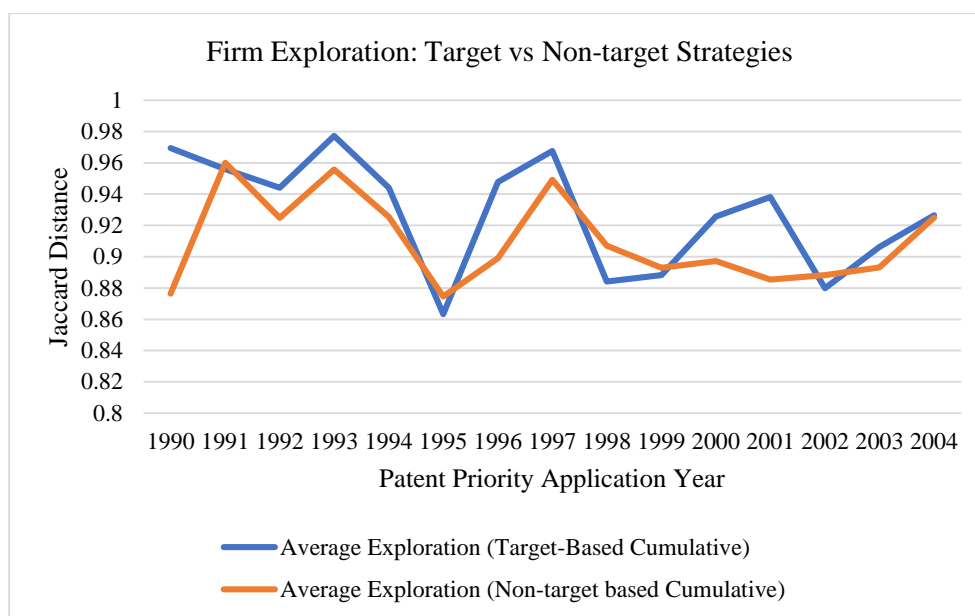
**Figure 7:** Firm exploration calculated using chemical structural similarities within patents. Markush patents (granted and applications) for 578 firms are represented here from 1990-2004.



**Figure 8:** Firm exploration for target-based drug discovery is shown here.



**Figure 9:** Firm exploration for non-target based drug discovery is shown here.



**Figure 10:** Targeted strategies show more exploration compared to non-target based drug discovery

Figure 6 shows variation in explorative paths for three firms. A score closer to 1 indicates maximal dissimilarity – indicating exploration. The similarity here is calculated based on chemical scaffolds, which are the primary backbone structures of compounds. The Jaccard distance for any given year (e.g. 0.96) is read as: average dissimilarity to prior year is 0.96 or about 96% of current portfolio of compounds are not similar to all the prior year's compounds. Hence, indicating a higher level of exploration.

The average of all such explorative paths is shown in Figure 7. As one would expect, the prior year comparison trends are slightly higher compared to cumulative. Matching only against prior year compounds is a subset of the history of all compounds made by the firm. Figures 8-9 compare the explorative paths of the firms based on their search

strategy. Figure 10 compares average firm exploration based on search strategies (cumulative comparison). Overall, targeted strategies indicate more exploration than non-targeted strategies.

### **Comparison to Euclidean Distance Measure**

In the innovation literature, technological distances have been previously measured using different spatial distance measures like Euclidean, Cosine and Jaccard distances. These distances are derived from spatial coordinates specified by patent classes and subclasses. The United States Patent Office provides a well-structured patent class system that is used to capture distances. For example, a patent with class 514 and subclass 18.7 indicates Drug and Body Treating Compositions (class 514) for an Anti-Inflammatory (sub-class 18.7). Using such detailed classes, vector distances can be calculated for any two patents or groups of patents. A well-used method for measuring technological distances in the innovation literature is the Euclidean distance based on patent classes (Jaffe, 1989; Aharonson & Schilling, 2016; Tzabbar, 2009; McNamee, 2013). To compare the chemical distance method to existing approaches, I focused on technological distances that use patent class information derived from the US patent classification system.

An algorithm was implemented in the Python programming language using the Scientific Python (SciPy) library to capture the inter-patent distances using USPTO classes and subclasses. To measure technological distance between two sets of patents, centroid distances were used. Patents were grouped by year for a firm or group of firms depending

on the analysis. The centroid for a group of patents (or n-dimensional space object) is the arithmetic mean position of all the points in the shape. The Euclidean distance between any two patent-years is calculated using the centroids between those years. This centroid-based measure captured the technological distance groups of patents for a firm or collection of firms grouped by years.

The general formula for Euclidean distance is,

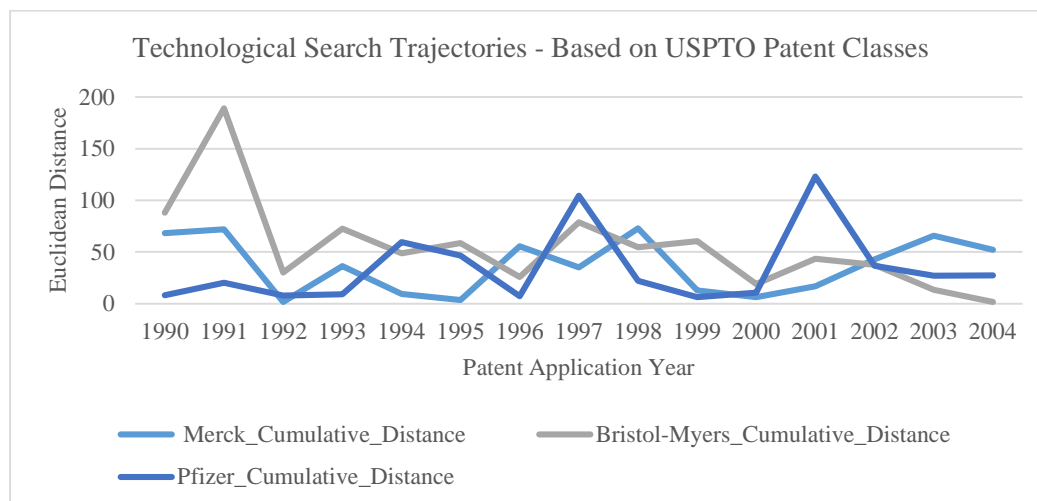
$$d = |\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}.$$

where, x and y are any two vectors.

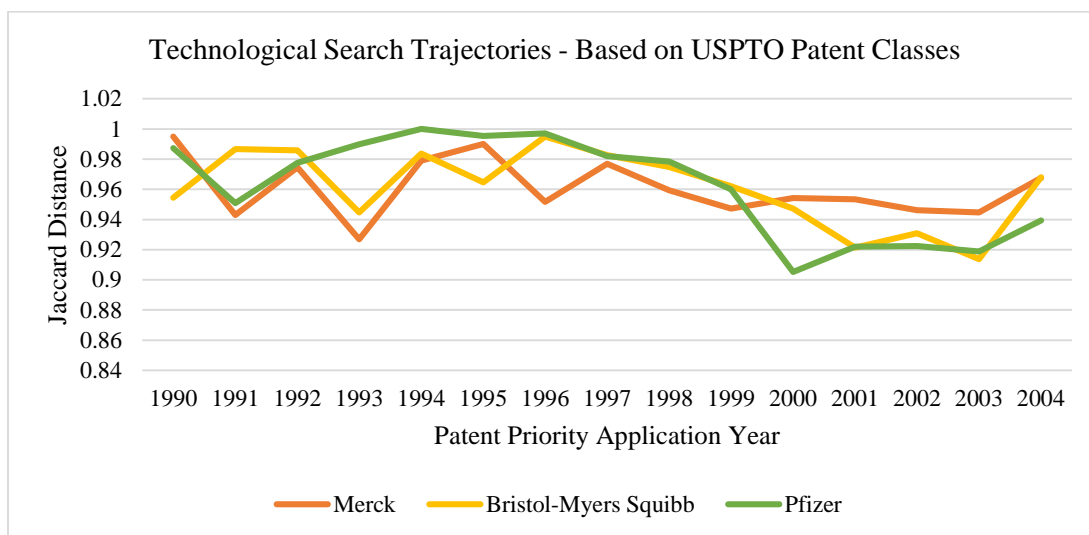
**Cumulative versus Prior-Year Distances:** Similar to the chemical distance measure, two versions of distance measures were implemented to capture firm level exploration. A cumulative distance measure captures distance compared to all the previous years' patent classes. That is patents in 1999 are compared to all prior year patents (1990-1998). A prior-year distance measure compares the classes to only the prior year's classes. That is 1999 is compared to 1998 only. This dual measure provides a granular account of the firm's year-by-year exploration.

#### Firm-Level Technological Distances

*Data sample:* For this firm level comparison, I selected about 1000 US patent applications that had patent class information for three firms – Merck (422), BMS (312), Pfizer (290). Results are show in Figure 11 and 12 below.



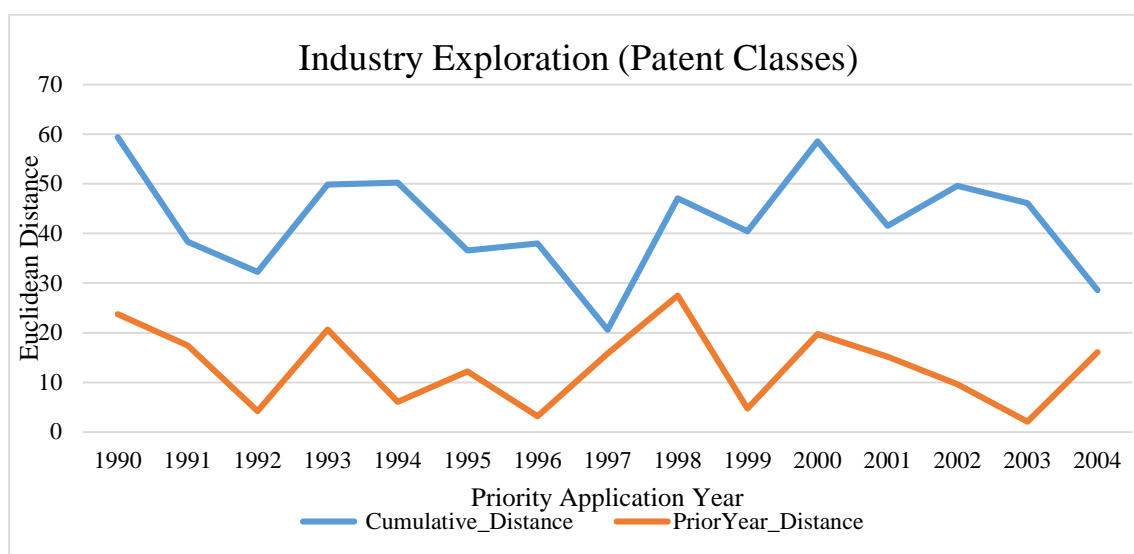
**Figure 11:** Firm exploration calculated using USPTO classes (Euclidean distance) for 3 firms between 1990-2004. Patents are grouped by year, their centroids calculated and distance compared.



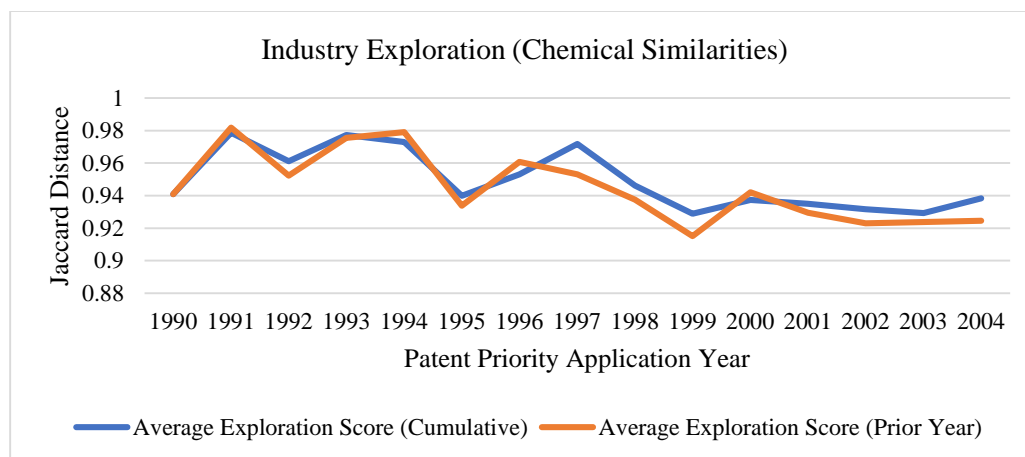
**Figure 12:** Firm exploration calculated using chemical structural similarities for 3 firms between 1990-2004. Average cumulative exploration scores for each patent are grouped by year and compared.

### Industry-level Technological Distances

*Data sample:* For the industry level comparison, I selected about US patent applications for 3000 firms filed between 1990-2004. Results are show in Figure 13 and 14 below.



**Figure 13:** Industry level exploration calculated using USPTO classes (Euclidean distance) between patents. Markush patents (granted and applications) for 3000 firms are represented here from 1990-2004.



**Figure 14:** Firm exploration calculated using chemical structural similarities within patents. Markush patents (granted and applications) for 570 firms are represented here from 1990-2004.

#### Discussion of Chemical and Euclidean distances

Euclidean distance (based on patent classes) indicates an increase in exploration 1999-2004, while chemical structure based distances show a decrease in exploration. Both measures are looking at very different levels of observation: patent classes are high level based on meta-data (macro classification based on US patent office), while chemistry-similarity based measure is micro-level and granular, examining differences in the type of compounds firms make and their trajectory in chemical space. Interpreting these results should take into consideration these macro-micro levels of analysis. Patent-based distances are restricted by the level of granularity the USPTO or patent body provides. Also, these measures are susceptible to changes in patent classes or subclasses (i.e. addition of new subclasses can increase or decrease distance and may need to be controlled for longitudinal studies). In the chemical-similarity based measure, Jaccard

distances based on Tanimoto chemical similarities are used to capture centroid distances between groups of patents. This measure is built upon actual compounds made and registered in patents and is an accurate measure of the exploration in chemical space by the firm.

The industry exploration for patent based approach does not show any clear or consistent trends over time. The exploration trends for chemical based distances show a clear decrease in exploration over time, showing that firms are tending to make the same kind of compounds over time. Thus, showing that technological exploration is decreasing with time.

## **Discussion and Conclusions**

To empirically assess the impact of the human genome map on exploration a novel distance measure was developed and tested on a sample of Markush drug patents and applications comprising 570 firms from 1990-2004. Results indicate an overall decrease in exploration, while targeted strategies showing more exploration than non-target based approaches (Figures 6-10). These results support underlying events in the context of drug discovery and are discussed in detail below.

Overall, we see a general narrowing of firm exploration between 1990-2004. Two interesting trends are consistent across these figures – a dip in exploration in 1995 and a period of increased exploration subsequent to that. In this particular time period, two major events occurred related to drug discovery and novel compound search.

Combinatorial chemistry, a novel technology to make large number of compounds at low cost and high diversity entered the industry (Persidis, 1998). From the late 1990's leading up to June 2000, the Human Genome Map was released in the public domain. These two developments are critical to understanding the exploration trends we observe in Figures 6-10.

In general, exploration could be declining over time as firms tend to make compounds similar to what they know or what their inventors are experts in. In our interviews with industry experts, we found that some medicinal chemists have preferred chemical scaffolds in which they are specialists and tend to exploit their expertise in manipulating those structures across projects. It is also possible that firms could be working on similar looking chemical structures driven by focused disease markets and targets areas that they already operate in (e.g. antibiotics versus pain drugs). These constraints could be narrowing the search process over time.

In the mid-1990's, there was broad adoption of combinatorial chemistry techniques across the drug industry. Firms built their own libraries of compounds and also had access to vendor libraries that contained millions of new compounds. Since firms were no longer limited by their medicinal chemists' skills, with combinatorial chemistry they could to experiment with new compound structures and screen them in a high-throughput manner. This accessibility to new diverse compounds and novel chemical scaffolds due to combinatorial chemistry could explain the sudden increase in chemical exploration across the industry post-1995.

We observe targeted strategies having higher levels of exploration in most years compared to non-targeted. The underlying differences in search strategies could also explain the contrasting exploration trends between targeted and non-targeted strategies. Non-target based or phenotypic search is based on iterative testing and modification of a small set of chemical structures, until the desired effect is found. It is possible, that chemical space exploration in the non-targeted approach is encumbered by two limiting factors: medicinal chemist's bias in selection of chemical structures based on preferences and focused iteration/selection of compounds based on privately accumulated internal libraries. In contrast, the target-based approach is defined by having a focused disease target (a lock) for which combinatorial libraries can be screened in a high-through way for available hits (keys). The diverse nature of externally sourced libraries could be the reason for the increase in chemical space exploration compared to non-target based exploration that we observe in Figure 10. Here, the disease or gene target combined with combinatorial libraries is driving exploration, not the tacit knowledge and prior experience of medicinal chemists and firms.

We also observe an increase in exploration for target-based strategies post-2000 coinciding with the release of the human genome (Figures 9-10). A possible explanation for this bump in exploration could be the availability of about 10,000 targets – many new genes and exact DNA sequences that could be used to model the target protein structures. The availability of new gene targets combined with combinatorial chemistry libraries (new, diverse chemical scaffolds) could be increasing overall exploration in chemical

space. Interestingly, we observe exploration also increasing in non-targeted strategies towards the early 2000's suggesting that firms could be integrating combinatorial chemistry and high-throughput methods across both strategies (Kotz, 2012).

While the technological distance measure introduced here can provide granular tracking of exploration paths compared to patent classes, more testing is needed using different patent datasets and across different time periods to improve robustness. In summary, this study introduces a novel technique using chemical similarities to measure technological distance and firm exploration using patents. By capturing technological search trajectories over time this study provides insights on the effect of the human genome map on chemical space exploration.

## References

- Aharonson, Barak S., and Melissa A. Schilling. "Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution." *Research Policy* 45.1 (2016): 81-96.
- Arora, Ashish, and Alfonso Gambardella. "The changing technology of technological change: general and abstract knowledge and the division of innovative labour." *Research policy* 23.5 (1994): 523-532.
- Backman, Tyler WH, Yiqun Cao, and Thomas Girke. "ChemMine tools: an online service for analyzing and clustering small molecules." *Nucleic acids research* 39.suppl\_2 (2011): W486-W491.

Bajusz, Dávid, Anita Rácz, and Károly Héberger. "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?." *Journal of cheminformatics* 7.1 (2015): 20.

ChemMine Tools Tutorial: <http://chemmine.ucr.edu/help/>

ChemmineR: Cheminformatics Toolkit for R. (2018)

<http://www.bioconductor.org/packages/release/bioc/vignettes/ChemmineR/inst/doc/ChemmineR.html>

Cohen, Wesley M., and Daniel A. Levinthal. "Absorptive capacity: A new perspective on learning and innovation." *Administrative science quarterly* (1990): 128-152

Cyert, Richard M., and James G. March. "A behavioral theory of the firm." Englewood Cliffs, NJ 2 (1963): 169-187.

*Drug discovery* 11.3 (2012): 191-200. Tripp, Simon, and Martin Grueber. "Economic impact of the human

Fleming, Lee, and Olav Sorenson. "Science as a map in technological search." *Strategic Management Journal* 25.8?9 (2004): 909-928.

Fleming, Lee, and Olav Sorenson. "Technology as a complex adaptive system: evidence from patent data." *Research policy* 30.7 (2001): 1019-1039.

Fleming, Lee. "Recombinant uncertainty in technological search." *Management science* 47.1 (2001): 117-132.

genome project." Battelle Memorial Institute (2011).

Gittelman, Michelle. "The revolution re-visited: Clinical and genetics research paradigms and the productivity paradox in drug discovery." *Research Policy* (2016)

- Helfat, Constance E., and Ruth S. Raubitschek. "Product sequencing: co-evolution of knowledge, capabilities and products." *Strategic management journal* (2000): 961-979.
- Hemphill, C. Scott, and Bhaven N. Sampat. "When do generics challenge drug patents?." *Journal of Empirical Legal Studies* 8.4 (2011): 613-649.
- Henderson, Rebecca M., and Kim B. Clark. "Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms." *Administrative science quarterly* (1990): 9-30.
- Johnson, M. A., and G. M. Maggiora. "Concepts and Applications of Molecular Similarity, John Wiley & Sons." (1992).
- Journal of Economic Behavior & Organization* 43.2 (2000): 141-166.
- Kaplan, Sarah, and Keyvan Vakili. "The double-edged sword of recombination in breakthrough innovation." *Strategic Management Journal* 36.10 (2015): 1435-1457.
- Katila, Riitta, and Gautam Ahuja. "Something old, something new: A longitudinal study of search behavior and new product introduction." *Academy of management journal* 45.6 (2002): 1183-1194.
- Kauffman, Stuart, José Lobo, and William G. Macready. "Optimal search on a technology landscape."
- Kotz, Joanne. "Phenotypic screening, take two." *SciBX: Science-Business eXchange* 5.15 (2012).
- Cyert, Richard M., and James G. March. "A behavioral theory of the firm." Englewood Cliffs, NJ 2 (1963): 169-187.
- Lipinski, Christopher A. "Drug-like properties and the causes of poor solubility and poor permeability." *Journal of pharmacological and toxicological methods* 44.1 (2000): 235-249.

Lipinski, Christopher, and Andrew Hopkins. "Navigating chemical space for biology and medicine." *Nature* 432.7019 (2004): 855-861.

Lipinski, Christopher, and Andrew Hopkins. "Navigating chemical space for biology and medicine." *Nature* 432.7019 (2004): 855-861.

March, James G. "Exploration and exploitation in organizational learning." *Organization science* 2.1 (1991): 71-87.

Nagaraj, A. (2015, November 6). The Private Impact of Public Maps— Landsat Satellite Imagery and Gold Exploration. Job Market Paper, MIT Sloan School of Management, Cambridge, MA. Available: [http://web.mit.edu/nagaraj/files/nagaraj\\_jmp\\_nov6.pdf](http://web.mit.edu/nagaraj/files/nagaraj_jmp_nov6.pdf)

Nelson, Richard R., and G. Sidney. "Winter. 1982. An evolutionary theory of economic change."

Papadatos, George, et al. "SureChEMBL: a large-scale, chemically annotated patent document database." *Nucleic acids research* 44.D1 (2015): D1220-D1228.

Persidis, Aris. "Combinatorial chemistry." *Nature biotechnology* 16.7 (1998): 691-693

Puranam, Phanish, and Murali Swamy. "Expeditions without Maps: Why Faulty Initial Representations May Be Useful in Join Discovery Problems." Available at SSRN 1153142 (2010)

Rosenkopf, Lori, and Atul Nerkar. "Beyond local search: boundary?spanning, exploration, and impact in the optical disk industry." *Strategic Management Journal* 22.4 (2001): 287-306.

Scannell, Jack W., et al. "Diagnosing the decline in pharmaceutical R&D efficiency." *Nature reviews*

- Southall, Noel T., and Ajay. "Kinase patent space visualization using chemical replacements." *Journal of medicinal chemistry* 49.6 (2006): 2103-2109.
- Stuart, Toby E., and Joel M. Podolny. "Local search and the evolution of technological capabilities." *Strategic Management Journal* 17.S1 (1996): 21-38.
- Tzabbar, Daniel. "When does scientist recruitment affect technological repositioning?." *Academy of Management Journal* 52.5 (2009): 873-896.
- Ulrich, Karl. "The role of product architecture in the manufacturing firm." *Research policy* 24.3 (1995):419-440.
- Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." *Journal of chemical information and computer sciences* 28.1 (1988): 31-36.
- Yoon, Janghyeok, and Kwangsoo Kim. "TrendPerceptor: A property–function based technology intelligence system for identifying technology trends from patents." *Expert Systems with Applications* 39.3 (2012): 2927-2938.

## Appendix

Figure A: Structural comparison of 8 compounds

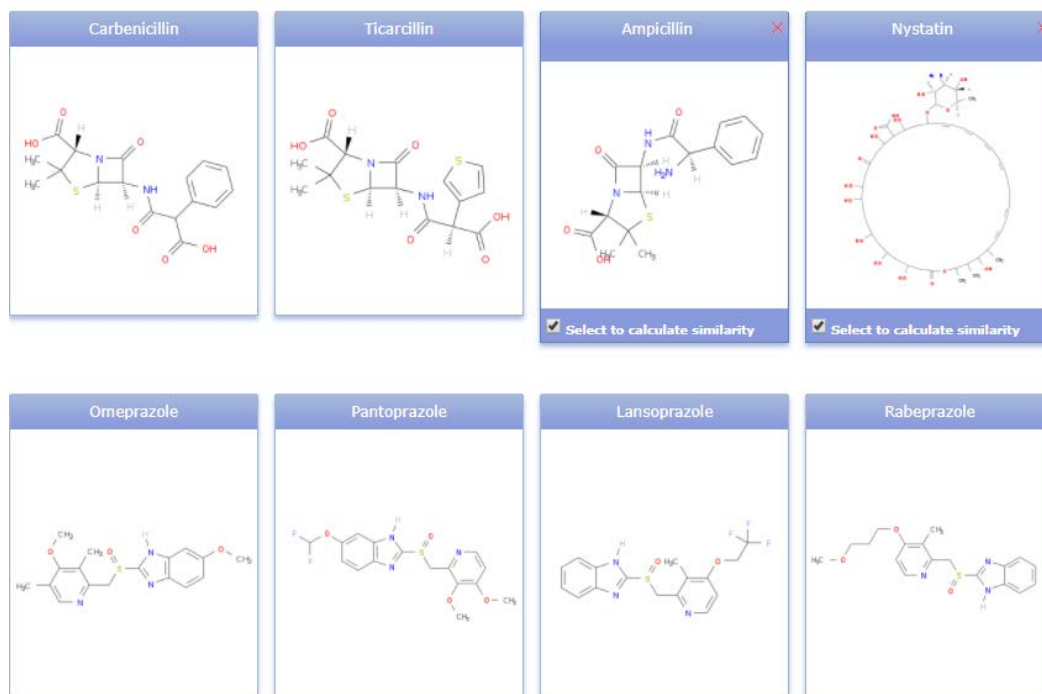


Figure B: Binning clustering algorithm results from ChemmineR

WORKBENCH

- [My Compounds](#)
- [Add Compounds](#)

TOOLS

- [Past Jobs](#)
- [Upload Numeric Data](#)
- [Cluster](#)
- [Physicochemical Properties](#)
- [Similarity Workbench](#)

SEARCH

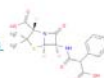
- [PubChem Similarity Search](#)

## Binning Clustering Results

Job Start Time	July 28, 2017, 1:59 p.m.
Options	Similarity Cutoff: 0.6
Results	<a href="#">Download CSV »</a>

[Hide Example Structures](#)

Bin 1 Size: 3



Members: [Carbenicillin](#) [Ticarcillin](#) [Ampicillin](#)

Bin 4 Size: 1



Bin 5 Size: 1

