Plant extracts and natural products - Predictive structural and biodiversity-based analyses of uses, bioactivity, and 'research and development' potential

Vafa Amirkia

Thesis submitted in accordance with the requirements of the UCL School of Pharmacy for the degree of Doctor of Philosophy

November 2016

UCL SCHOOL OF PHARMACY

29-39 Brunswick Square

London WC1N 1AX

Declaration

This thesis describes research conducted in the School of Pharmacy, University College London from February 2013 to February 2016 under the supervision of Professor Michael Heinrich and Dr Jose M. Prieto. I certify that the research described is original and that I have written all the text herein and have clearly indicated by suitable citation any part of this dissertation that has already appeared in publication.

1 November 2016

Vafa Amirkia

Abstract

The process of drug discovery and development over the last 30 years has been increasingly shaped by formulaic approaches and natural products – integral to the drug discovery process and widely recognized as the most successful class of drug leads – have significantly been deprioritized by a struggling worldwide pharmaceutical industry. Alkaloids - historically the most important superclass of medically important secondary metabolites - have been used worldwide as a source of remedies to treat a wide variety of illnesses yet, there exists a wide discrepancy between their historical and modern significances.

To understand these trends from an insider's perspective, 52 senior-stakeholders in industry and academia were engaged to provide insights on a series of qualitative and quantitative aspects related to developments in the process of drug discovery from natural products. Stakeholders highlighted the dissonance between the perceived high potential of natural products as drug leads and overall industry and company level strategies. Many industry contacts were highly critical to prevalent company and industry-wide drug discovery strategies indicating a high level of dissatisfaction *within* the industry. One promising strategy which respondents highlighted was virtual screening which, to a large extent has not been explored in natural products research strategies.

Furthermore, the physicochemical features of 27,783 alkaloids from the Dictionary of Natural Products were cross-referenced to pharmacologically significant and other metrics from various databases including the European Bioinformatics Institute's ChEMBL and Global Biodiversity Information Facility's GBIF biodiversity data. The combined dataset revealed that a compound's likelihood of medicinal use can be linked to its host species' abundance and was input into target-independent machine learning algorithms to predict likelihood of pharmaceutical use. The neural network model demonstrated an accuracy of >57% for all pharmaceutical alkaloids and 98% of all alkaloids.

This study is the first to incorporate the biodiversity of host organisms in a machine learning scheme characterizing druglikeness and thus demonstrates the link between host species' abundance and druglikeness. These findings yield new insights into cost-effective, real-world indicators of drug development potential across the diverse field of natural products.

Acknowledgements

This doctoral thesis would not have been possible without the support of many people.

Firstly, I wish to express my gratitude to my primary supervisor Professor Michael Heinrich, who was generously, abundantly, and consistently encouraging and offered irreplaceable guidance throughout the course of this research. There is no doubt that this research could not be completed without his constant and wise encouragement (often in the form gentle nudges to stretch my intellectual capacities each of the many times we encountered awkward pauses during our discussions). I also wish to express my gratitude to my secondary supervisor Dr Jose M. Prieto whose scientific proficiency and high standard of excellence has propelled this research forward in an invaluable way; he is a rising star in the field of pharmacognosy.

Furthermore, I wish to thank my classmates and colleagues at UCL which have provided a highly enriching experience during my studies and upheld the UCL value of 'collegiality and community-building' throughout the years. Outside the UCL community, I must acknowledge the encouragement from a wide range of acquaintances over the years; many of which are scattered across the globe working to enrich the collective life of society.

Moreover, I am deeply appreciative of the loyal support from a smaller circle of true friends who have stood shoulder-to-shoulder with me throughout various developments and exemplified excellence in various aspects of life.

Lastly, I would like to thank my parents whom been unconditionally supportive and infinitely trusting throughout the successive stages of my life. The profound friendship and deep spiritual bond we have nurtured, has, and will continue to guide my life. This research could not be possible if my father did not enable me 'to receive training at school and to be instructed in such arts and sciences as are deemed useful and necessary' and my mother had not always strived create 'conditions as would be most conducive to both [my] material and spiritual welfare and advancement'.

Publications Resulting from the Thesis

Journal articles:

1. Amirkia, V., & Heinrich, M. (2014). Alkaloids as drug leads– A predictive structural and biodiversity-based analysis. *Phytochemistry Letters*, 10, xlviii-liii.

2. Amirkia, V., & Heinrich, M. (2015). Natural products and drug discovery: a survey of stakeholders in industry and academia. *Frontiers in Pharmacology*, 6.

3. Amirkia, V., & Heinrich, M. (2016). Machine learning 'drug-likeness' in pure alkaloids (Unpublished Manuscript)

Conference proceedings:

1. Amirkia, V., & Heinrich, M. (2014) Alakaloids as drug leads- A predictive structural and biodiversity-based analysis. Poster presentation. 13th Meeting of Consortium for Globalization of Chinese Medicine (CGCM). 27-29 August 2014, Beijing, China.

2. Amirkia, V., & Heinrich, M. (2014) Alakaloids as drug leads- A predictive structural and biodiversity-based analysis. Poster presentation. 62nd International Congress and Annual Meeting of the Society for Medicinal Plant and Natural Product Research (GA). 31 August–4 September 2014, Guimaraes, Portugal.

3. Amirkia, V., & Heinrich, M. (2015) Natural product development: current industry and academia insights. Oral presentation. APS PharmSci 2015 – The Science of Medicines. 7–9 September 2015, Nottingham, UK.

Table of Contents

Table of Figures

Table of Tables

List of Abbreviations

AA – Amino Acid

ADME – Absorption, Distribution, Metabolism, and Excretion

ANN – Artificial Neural Network

API – Active Pharmaceutical Ingredient

BBB – Blood–brain Barrier

CLogP – Calculated Log P

CNS – Central Nervous System

DNP – Dictionary of Natural Products

EBI – European Bioinformatics Institute

EMBL – European Molecular Biology Laboratory

GBIF – Global Biodiversity Information Facility

GI – Gastrointestinal

HBA – Hydrogen Bond Acceptor

HBD – Hydrogen Bond Donor

HTS – High Throughput Screening

IP – Intellectual Property

LogD – Distribution Coefficient

LogP – Partition Coefficient

M&A – Mergers and Acquisitions

MWT – Molecular Weight

NP – Natural Products

OECD - Organization for Economic Co-operation and Development

pK_a – Acid Disassociation Constant

PSA – Polar Surface Area

QSAR - Quantitative Structure Activity Relationship

RAE – Relative Absolute Error

R&D – Research and Development

Ro3 – Rule of Three

Ro5 – Lipinski Rule of Five

ROI – Return on Investment

SAR – Structure-activity Relationships

SD – Standard Deviation

SVM – Support Vector Machine

TCM – Traditional Chinese Medicine

Tox – Toxicity

VS – Virtual Screening

WoK - Web of Knowledge

1. General introduction

1.1. Defining the field of natural products

1.1.1. Humans, nature, and natural products

Human relationship with nature has been long and complicated. From the beginning of the historical record, geography, time, and culture, have not stopped humans from continuously probing and exploring the wide range of natural phenomena. The earliest records of human activity often reveal impressive and ingenious 'responses' to these natural phenomena and to this day, human life is inseparable from the rich tapestry nature offers human existence. Yet, concurrent to humans becoming more adept at tapping into the resources inherent to their surroundings, they have also, in many cases, come closer to recognizing their own limitations in fully understanding scientific phenomena as well as understanding the consequences of human activity on nature.

Overwhelming evidence from the historical record suggest that essentially all aspects of human life have greatly benefitted through an ever-evolving, ever-penetrating understanding of the natural world. It very well may be this demonstrable, derived benefit across essentially all facets of humanity's collective life that has fuelled an insatiable thirst for more and what some characterize as the 'pillaging' of the world's resources with no regard for consequence. Nevertheless, many would agree that humans relationship with natural resources over the years has not only reaffirmed their value, but have also led to a greater collective consciousness on how to more *effectively* tap into these vast arrays of resources to further human civilization.

It could be argued that every knowledge system in existence today is somehow connected to or has sought inspiration from products of the natural world; collectively and in the broadest sense referred to as 'natural products'. Contemporary knowledge systems related to industry, commerce and healthcare have been intimately and remain connected with the natural world. In the cases of healthcare and medically-related scientific advances, nature has served as an indispensible platform in facilitating advances in the field. Consistently relied on over thousands of years, healing systems such as Ayurveda and Traditional Chinese Medicine (TCM) are a testimony to human's salutary dependency on nature. These ancient healing systems have not only provided an inestimable increase in the standard of life for large populations throughout the centuries, but also provided impetus to contemporary research efforts to focus on the vast reservoir of natural products.

In the modern day, some researchers studying natural products through a Darwinian lens are of the conviction that natural products carry an intrinsic superiority to man-made alternatives in the modern world of pharmacology. These researchers assert that these products must have added to a plant or organism's fitness in order to remain as metabolic product after thousands of years of selection. Firn and Jones capture this school of thought:

> The simplest evolutionary model accounting for natural product diversity thus demands that each natural product retained in a population must have a value to the producers. Organisms might retain for a short time some 'redundant' natural products in their chemistry (products whose production once enhanced fitness but which no longer do so), but natural selection would be expected to continuously prune such dead wood from the thicket. Redundant molecules could, of course, take on a new role as precursors of new generations of compounds that do enhance fitness (Firn and Jones, 2003).

And thus, if the development of natural products fits an evolutionary model, how specifically can their evolution be linked to an organism's fitness? In a subsequent publication, Firn and Jones (2009) identify five potential outcomes of one or more metabolic mutations on a natural product in this evolutionary process:

1. Possess properties that are new and enhance the functioning of the cell and hence the organism

2. Possess properties that are new and adversely affect the cell and hence the organism

3. Possess properties that are new but have no impact on the functioning of the cell or the organism other than the imposed metabolic cost of production

4. Possess properties that can substitute for an existing, necessary property with no impact on the functioning of the cell or the organism other than the imposed metabolic cost of production, but with the accrual of potential functional redundancy

5. Possess properties that can substitute for an existing, necessary property with a negative impact on the functioning of the cell hence the organism (*via*, for example, diversion of substrates)

Ji *et al.* (2009) extend this commentary by linking evolved, inherent ‘advantage’ (referred to as enhanced ‘fitness’ by Firn and Jones) of a natural product within a host species, with potential medicinal or biological advantages in human applications:

> As these compounds proved to be advantageous, they became a trait on which natural selection could act, and were retained and improved throughout the course of evolution. Given the similarities between aspects of human physiology and that of other animals, it is not surprising that such molecules can also exert biological effects in humans. For example, many chemicals that plants evolved to defend themselves against herbivores are now used as laxatives, emetics, cardiotonics or muscle relaxants in humans. In addition, humans have taken advantage of some of the discovered properties of natural compounds: those that are able to interact with or suppress the growth of bacteria, for example, are now used as antimicrobial drugs in medicine.

Lastly, it must be noted that increased human involvement with and proficiency in tapping into the reservoir of nature’s resources has, in many cases blurred the line between what is truly ‘natural’ product and what has been synthetically modified by chemical and genetic processes or by other selective human interference. This point will be explored in subsequent sections of this chapter. At this point, recognizing that natural products are seen as a source of latent potential for the advancement of various human endeavours and defining them as naturally occurring compounds is sufficient.

1.1.2. Classifying natural products

Classification schemes in any scientific field are complex and the classification of natural products is certainly no exception. In the case of biological taxonomy - referred to as 'the world's oldest' profession - organisms are grouped together taxonomically by identifying shared characteristics such as physical/genetic traits (often referred to as phylogeny) or by evolutionary relationships (Knapp, 2010). These classifications schemes have been and continue to remain hotly debated. Typically, these schemes receive impetus and are adjusted concurrent to technological advances such as genetic screening or advances in bioinformatics.

Various schemes have been put forth with respect to the classification ofnatural products and this thesis will not fully explore the details of all such schemes. Yet, one of the earliest and most influential with respect to natural products produced by plants, referred to as primary and secondary metabolites, was put forth by a German physiological chemist named Albrecht Kössel. Kössel, who as Finn and Jones (2009) characterized it, 'unknowingly initiated a schism when he proposed that plants had two distinct types of metabolism, 'primary' and 'secondary''. Primary metabolites at the time were characterized by their commonality among all organisms and their role in the basic cellular processes. Before Kössel's proposition in 1891, these primary metabolites were studied intensely by a wide number of chemists and were considered to be of primary importance for the survivability of an organism, such as cell division and growth, respiration, storage, and reproduction. Kössel proposed that in addition to these basic compounds, there exist another group of metabolic products called secondary metabolites. These natural compounds were expounded upon in a significant way about 30 years after Kössel's proposition through the work of Czapek in 1921 'who dedicated an entire volume of his 'plant biochemistry' series to what he named '*Endproduckt*'. Bourgaud continues that 'according to him [Czapek], these products could well derive from nitrogen metabolism by what he called 'secondary modifications' such as deamination. Compared to the main, 'primary' molecules found in plants, these secondary metabolites were soon defined by their low abundance, often less than 1% of the total carbon, or a storage usually occurring in dedicated cells or organs" (Czapek, 1921).

From its beginnings in 1891, this simple classification scheme of primary and secondary metabolites has inevitably evolved through a series of challenges and refutations. Not all such developments will be stated in this thesis but it is important to note that modifications to these schemes have also been punctuated by a few voices highlighting the artificiality of all such labels, and suggestions to do away with these artificial labels altogether. In their textbook *Natural Products from Plants*, Kaufman *et al.* (1999) summarize the inaccuracies of the primary/secondary metabolite labels and why they have decided to exclude them within their text:

> Regarding terminology pertaining to plant metabolites, we shall refrain from using the terms “primary and secondary metabolites”. These labels have caused a lot of confusion in the literature and their continued use certainly cannot be defended on chemical grounds. So-called “primary metabolites” have referred to those compounds that produce energy, such as adenosine triphosphate. So called “secondary metabolites” have referred to those compounds synthesized by plants that do not produce energy. These ideas, in our view, are obsolete and not useful. Why? Because many of the compounds/metabolites considered to be “secondary” are really essential for carbon fixation and reduction through photosynthesis, glycolysis, fermentation, and the tricarboxylic acid cycle. Also, many of the metabolites that have been classified as “secondary” are really essential to the survival of the plant at particular times in its developmental life cycle. So, we shall abandon the older terminology of “primary and secondary metabolites” and simply use the terms metabolite and product for any of the compounds that plants synthesize because, as far as we can ascertain, all have some survival value to the plant in both time and space (Kaufman et al., 1999).

Although the debate surrounding terminology continues to this day, the purpose of the aforementioned details is to convey an understanding that these terms are to a large extent fluid and dynamic. Schemes can vary so widely that there are no hard and fast rules to defining what a natural product is (particularly with an every-increasing involvement by humans into the natural world). For the purposes of this thesis the following scheme (Fig. 1) will be utilized which to a large extent is widely accepted as a sufficiently representative classification scheme of natural products across all living organisms. Differing from Kaufman *et al.*, this scheme divides natural products into primary and secondary metabolites and specifically identifies polyketides, fatty acids, terpenoids, steroids, phenylpropanoids, alkaloids, specialized amino acids (AA) and peptides, and specialized carbohydrates as secondary metabolites. It must be noted that what chemical and physical properties constitute compounds belonging to the aforementioned classes of compounds is also a point of debate. In the case of the alkaloids, a more specific definition will be explored in subsequent chapters of this thesis.

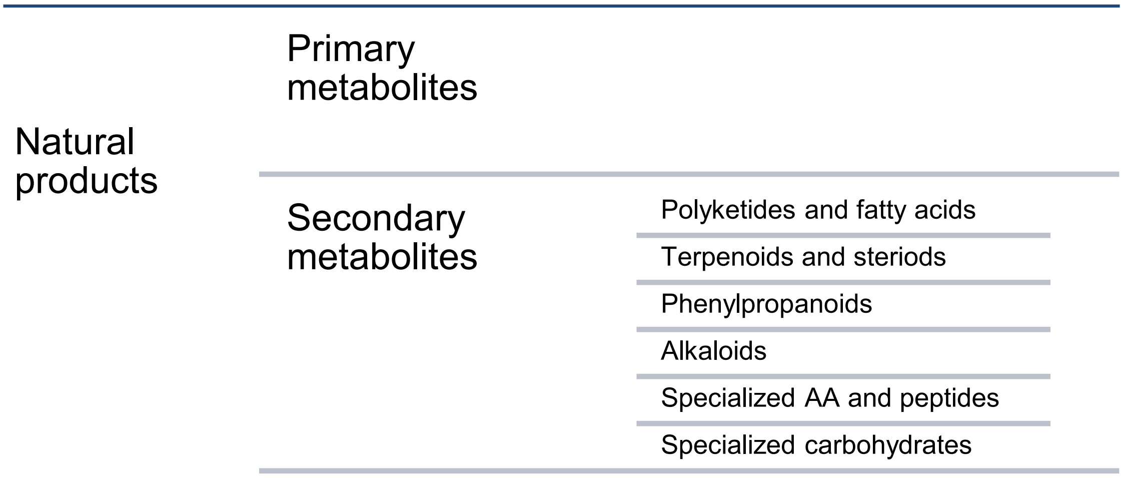


Figure 1: Commonly agreed upon classes and sub classes of natural products (Ji *et al.*, 2009)

1.2. Significance in modern pharmaceutics

One would imagine that following millennia of use as traditional remedies and centuries of increasingly effective isolation and purification, natural products would be an integral part of modern pharmaceutical science. Indeed, they are. Natural products do play a critical role in the discovery and development of modern medicines and are widely regarded as the most successful class of compounds as drug leads. Their sheer diversity alone is unparalleled by any other class of compounds.

Beginning in 1997, Newman and Cragg began a series of highly acclaimed reviews looking at ‘Natural Products as Sources of New Drugs’ in which they traced what percentage of newly approved molecular entities (NMEs) were split across the following major categories of compound classes (with abbreviations):

- “B” Biological; usually a large (>45 residues) peptide or protein either isolated from an organism/cell line or produced by biotechnological means in a surrogate host.

- “N” Natural product.

- “NB” Natural product “Botanical” (in general these have been recently approved).

- “ND” Derived from a natural product and is usually a semisynthetic modification.

- “S” Totally synthetic drug, often found by random screening/modification of an existing agent.

- “S*” Made by total synthesis, but the pharmacophore is/ was from a natural product.

- “V” Vaccine.

- “/NM” Mimic of natural product

Cragg and Newman’s five highly cited reviews in 1997, 2003, 2007, 2012, and 2014 have essentially become the ‘gold-standard’ in highlighting the role natural products research in the drug discovery world. Their reviews consistently show two major trends. The first (Fig. 2) is that there has been no upward surge or downward fall in the rate at which NMEs are approved by regulatory authorities, implying that either regulators are more strictly regulating the approval of new drugs, the industry is just not growing to be more innovative, or a combination of the two. A more detailed analysis of these macro trends are covered in subsequent sections of this thesis as well as in the published observations of a large number of industry stakeholders. The second trend (Fig. 3) relates specifically to natural products and shows that natural products, depending on the year, have occupied anywhere between 12% and 47% of all new small-molecule approvals. From 1981 to 2014, there have only been two years when this percentage has decreased below 20% (1997 and 2013) and on average this percentage is greater than 30% which is the highest percentage of any single class of compounds.

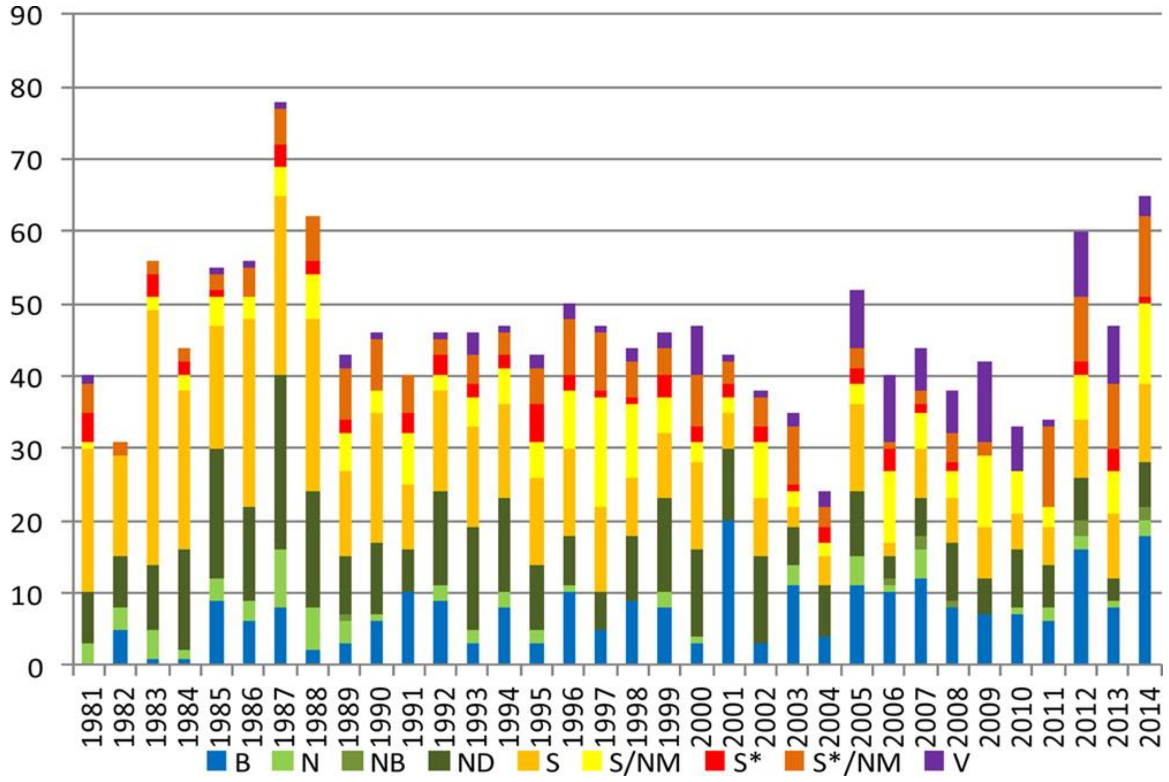


Figure 2: All small-molecule approved drugs by source/year (Newman and Cragg, 2014).

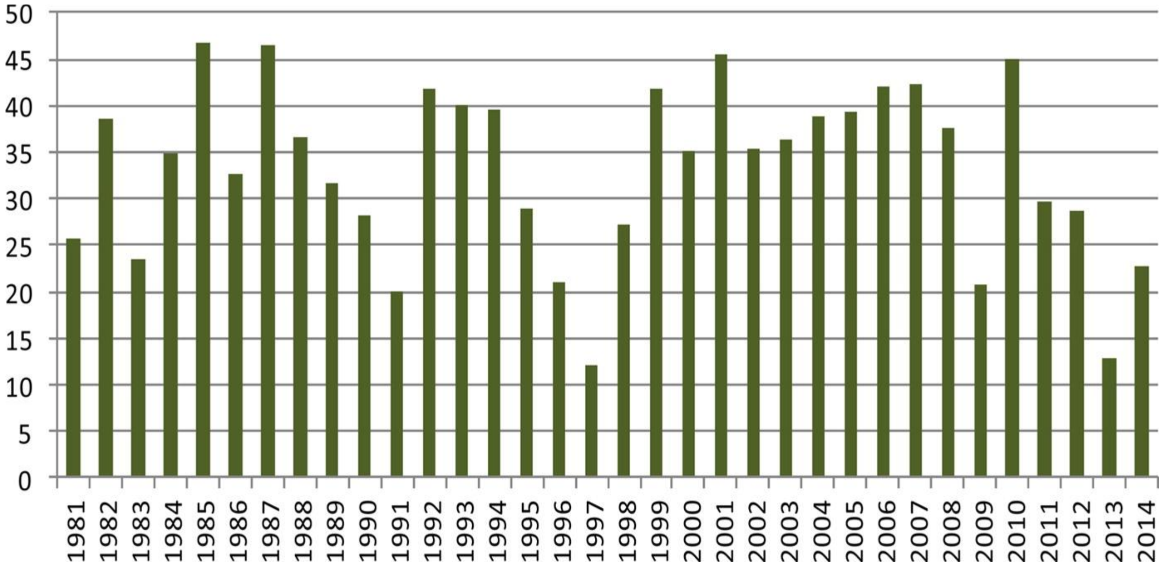


Figure 3: Percent of approved small-molecule natural product, natural botanical, and natural products derived by year, 1981−2014 (Newman and Cragg, 2014).

Cragg and Newman's analysis and their accompanying commentary have been echoed by many others throughout the years. A more detailed analysis of other views on the current and future potential of natural products in pharmaceutics is presented in subsequent sections of this thesis.

In their 2012 review, Cragg and Newman summarize their findings covering several of their previous publications and link these findings to the vast potential they judge, is still latent within the natural products arena:

> In this review, as we stated in 2003 and 2007, we have yet again demonstrated that natural products play a dominant role in the discovery of leads for the development of drugs for the treatment of human diseases. As we mentioned in earlier articles, some of our colleagues argued (though not in press, only in personal conversations at various forums) that the introduction of categories such as “S/NM” and “S*/NM” is an overstatement of the role played by natural products in the drug discovery process. On the contrary, we would still argue that these further serve to illustrate the inspiration provided by Nature to receptive organic chemists in devising ingenious syntheses of structural mimics to compete with Mother Nature’s longstanding substrates.
>
> Even if we discount these categories, the continuing and overwhelming contribution of natural products to the expansion of the chemotherapeutic armamentarium is clearly evident…and as we stated in our earlier papers, much of Nature’s “treasure trove of small molecules” remains to be explored, particularly from the marine and microbial environments.

1.3. Evolving attitudes towards their use as drugs

A wider selection of natural product development and drug discovery-related opinion, review, and primary literature published over the last two decades shows a range of varied, often contrasting viewpoints on the potential of natural products as drug leads/candidates (Table 1). The majority of published literature hails the potential of natural products as sources of structurally novel, highly diverse compounds and cites examples of how natural products comprise a high proportion of successfully marketed new drugs over the last 20 years. The voice of optimism is loud and clear and has generally overshadowed a number of critical voices which have pointed out major challenges in natural product drug development such as extraction and supply issues (McChesney, 2007). Most of these publications focus on plants-derived natural products but some also touch on compounds isolated from marine, fungal, or bacterial hosts. Many have added additional perspectives to this exploration into the potential of natural products development through focusing on academia-industry partnership initiatives, inter-disciplinary approaches such as virtual screening methods and genomics efforts; one example being Shen's paper in 2003 which outlined three main advantages of virtual screening of natural products. Shen argued that virtual screening provides higher hit rates as compared with typical HTS assays thus saving time/cost and considers it to be a more effective strategy in investigating the 90% of 'natural diversity' which so far has not been explored (defined as species which have yet to be studied systematically in research settings), and lastly more effective in increased prediction of ADME/Tox and other drug like properties which may show promise in diminishing missed/failed hits (Shen, 2003; Bohlin, 2010).

Title	Author & Year of publication	General outlook/tone
Recent Natural Products Based Drug Development: A Pharmaceutical Industry Perspective	Shu, 1998	Optimistic
Natural Product Drug Discovery in the Next Millennium	Cragg and Newman, 2001	Optimistic
Natural Products in the Process of Finding New Drug Candidates	Vuorela *et al.*, 2004	Optimistic
The Role of Natural Product Chemistry in Drug Discovery	Butler, 2004	Neutral
The Renaissance of Natural Products as Drug Candidates	Paterson and Anderson, 2005	Optimistic
Drug Discovery from Medicinal Plants	Balunas and Kinghorn, 2005	Optimistic

The Evolving Role of Natural Products in Drug Discovery	Koehn and Carter, 2005	Optimistic
Drug Discovery from Natural Products	Gullo *et al.*, 2006	Optimistic
Drug Discovery from Natural Sources	Chin *et al.*, 2006	Optimistic
Plant natural products: Back to the future or into extinction?	McChesney *et al.*, 2007	Pessimistic
Challenges and Opportunities in Drug Discovery from Plants	Jachak and Saklani, 2007	Optimistic
A Review of High Throughput Technology for the Screening of Natural Products	Mishra *et al.*, 2007	Neutral
New Aspects of Natural Products in Drug Discovery	Lam, 2007	Neutral
The Value of Natural Products to Future Pharmaceutical Discovery	Baker *et al.*, 2007	Neutral/ Pessimistic
Molecular understanding and modern application of traditional medicines: triumphs and trials	Corson and Crews, 2007	Neutral/ Optimistic
Natural Products in Drug Discovery	Harvey, 2008	Optimistic
Natural Products as a Robust Source of New Drugs and Drug Leads: Past Successes and Present Day Issues	Rishton, 2008	Neutral
Drug Discovery and Natural Products: End of an Era or an Endless Frontier?	Li and Vederas, 2009	Neutral
Modern Natural Products Drug Discovery and Its Relevance to Biodiversity Conservation	Kingston, 2010	Optimistic

The impact of the United Nations Convention on Biological Diversity on natural products research	Cragg *et al.*, 2012	Neutral
The Pharmaceutical Industry and Natural Products: Historical Status and New Trends	David *et al.*, 2014	Neutral
The Re-emergence of Natural Products for Drug Discovery in the Genomics Era	Harvey *et al.*, 2015	Optimistic

Table 1: Selection of representative publications on the outlook of natural products as drug leads in modern drug discovery programs and their overall levels of optimism (based on this author’s assessment)

Overviews of merits in natural product development are characterized by Harvey’s assertion in 1999 that ‘the major advantage of natural products for random screening is the structural diversity provided by natural products, which is greater than provided by most available combinatorial approaches based on heterocyclic compounds’. An increasing number of statements such as these indicate that the debate of whether or not natural products may serve as drug leads has evolved into a debate of how best tap into the potential latent in such a diverse and rich class of compounds. These observations also dovetail with Knight’s summary in 2003 of the advantages and challenges of natural product development. It is important to note that although Knight’s, and many others’, research primarily focused on seeking solutions to rising microorganism resistances towards traditional antibiotics through diversifying microbe genomes in order to diversify natural products their analysis is very much related to primary as well as sub-classes of secondary metabolites. While the generalized advantages listed below (Table 2) have been elucidated upon by many, less attention has been given to understanding multi-disciplinary yet specific challenges (Table 3) associated with natural product development.

With these observations in mind, one of the key aims of this thesis is to investigate to what extent challenges such as ‘characterization and isolation of the active compounds from natural product extracts are extremely labour intensive and time consuming' or 'the lack of systematic exploitation of ecosystems for the discovery of novel microbial compounds had resulted in random sampling and has missed the true potential of many regions' hold true.

Advantages in natural product development programs
Natural products offer unmatched chemical diversity with structural complexity and biological potency (Verdine, 1996).
Natural products have been selected by nature for specific biological interactions. They have evolved to bind to proteins and have drug-like properties (Nisbet and Moore, 1997).
Natural product resources are largely unexplored and novel discovery strategies will lead to novel bioactive compounds. Natural product extracts are complementary to synthetic and combinatorial libraries. About 40% of the natural product diversity is not represented in synthetic compounds libraries (Henkel *et al.*, 1999).
Research on natural products has led to the discovery of novel mechanisms of action, for example, the discovery of the role of guggulsterone (Urizal *et al.*, 2002)
Natural products are powerful biochemical tools, serving as "pathfinders" for molecular biology and chemistry and in the investigation of cellular functions (Hung *et al.*, 1996).
Natural products can guide the design of synthetic compounds (Breinbauer *et al.*, 2002).

Table 2: Perceived inherent advantages in natural product development (Knight *et al.*, 2003)

Challenges in natural product development programs
The lack of systematic exploitation of ecosystems for the discovery of novel microbial compounds had resulted in random sampling and has missed the true potential of many regions (Czárán *et al.*, 2002).

The characterization and isolation of the active compounds from natural product extracts are extremely labor intensive and time consuming (Monaghan *et al.*, 1995)
The production of adequate quantities of the active compound needed for drug profiling may require extensive media optimization and scale-up (Strobel, 2002).

Table 3

Perceived inherent challenges in natural product development (Knight *et al.*, 2003)

One limitation of many, if not all, of the aforementioned studies evaluating natural products is that in essence they are opinion papers not based on empirical data from relevant stakeholders. The authors normally had not engaged with any or a substantial number of stakeholders from *within* the pharmaceutical industry; most importantly those who *currently* work in the industry. Understandably so, not only is it challenging to track down a meaningful number of industry decision makers with experience in natural product drug development, but perhaps the larger challenge is eliciting their views (which can be critical of their superiors) pertaining to their company's strategy and/or industry trends. This internal lens, through angles such as commercial operations, strategic planning, research and development, and senior management is essential in gaining a clearer understanding of the role of natural product discovery and development as it contributes to drug development in general, as well as the gaps, and potential advances in academic-industry partnerships to advance drug discovery efforts.

1.4. Objectives

Thus, in light of the historical and modern importance of natural products in pharmaceutics, optimism among academics that natural products can continue to contribute significantly to the advancement of drug discovery and the declining productivity of the drug discovery process within industry, the following steps will be taken to assess how natural products can continue to drive drug discovery. Firstly, assumptions about advantages and challenges must be validated in light of stakeholder experience within the pharmaceutical industry. Subsequently, these findings allow for a fuller understanding of what factors are serve as 'bottlenecks' in the development of natural products and lastly, help understand how such factors can be aptly characterized and practically alleviated. This approach is explored in the subsequent chapters of this thesis.

2. Strategies, challenges and perceptions in modern drug discovery

2.1. Prevalent trends – a review of the literature

2.1.1. Macro trends: Industry

2.1.1.1. Rising costs and lessening productivity

The origins of the modern pharmaceutical industry have been debated over the years with some arguing that that the modern pharmaceutical industry traces its earliest roots back to 'apothecaries and pharmacies' which can be dated to the Middle Ages (Walsh, 2010) and focusing on earlier times associated with crude traditional remedies. Whatever the 'origin story', it is clearly evident that the modern pharmaceutical industry, as it is recognized today, was heavily shaped by events in the 18th and 19th centuries. Summarizing the narrative surrounding these early breakthroughs in the context of widely recognized modern pharmaceutical companies, Walsh writes:

> Whilst the scientific revolution of the 17th century had spread ideas of rationalism and experimentation, and the industrial revolution had transformed the production of goods in the late 18th century, the marrying of the two concepts for the benefit of human health was a comparatively late development.
>
> Merck in Germany was possibly the earliest company to move in this direction. Originating as a pharmacy founded in Darmstadt in 1668, it was in 1827 that Heinrich Emanuel Merck began the transition towards an industrial and scientific concern, by manufacturing and selling alkaloids. Similarly, whilst GlaxoSmithKline's origins can be traced back as far as 1715, it was only in the middle of the 19th century that Beecham became involved in the industrial production of medicine, producing patented medicine from 1842, and the world's first factory for producing only medicines in 1859
>
> Meanwhile, in the USA, Pfizer was founded in 1849, by two German immigrants, initially as a fine chemicals business. They expanded rapidly during the American civil war as demand for painkillers and antiseptics rocketed. Whilst Pfizer was providing the medicines needed for the Union war effort, a young cavalry commander named Colonel Eli Lilly was serving in their army. A trained pharmaceutical chemist, Lilly was an archetype of the dynamic and multi-talented 19th century American industrialist, who after his military career, and trying his hand at farming, set up a pharmaceutical business in 1876. He was a pioneer of new methods in the industry, being one of the first to focus on R&D as well as manufacturing. Another military man in the drugs business was Edward Robinson Squibb, who as a naval doctor during the Mexican-American war of 1846–1848 threw the drugs he was supplied with overboard due to their low quality. He

> set up a laboratory in 1858, like Pfizer supplying Union armies in the civil war, and laying the basis for today’s BMS.
>
> Switzerland also rapidly developed a home-grown pharmaceutical industry in the second half of the 19th century. Previously a center of the trade in textiles and dyes, Swiss manufacturers gradually began to realise their dyestuffs had antiseptic and other properties and began to market them as pharmaceuticals, in contrast to the origin in pharmacies of other enterprises. Switzerland’s total lack of patent laws led to it being accused of being a “pirate state” in the German Reichstag. Sandoz, CIBA-Geigy, Roche and the Basel hub of the pharmaceutical industry all have their roots in this boom.
>
> It wasn’t just Swiss companies had their roots in the dye trade. Bayer was founded in 1863 as a dye maker in Wuppertal, the hometown of Karl Marx’s collaborator Friedrich Engels. It later moved into medicines, commercialising aspirin around the turn of the 20th century, one of the most successful pharmaceuticals ever at that point.

Walsh continues his narrative by describing how ‘national rivalries and conflicts’ as well as the first and second world wars accelerated the need for medicines (particularly insulin and penicillin) and thrust the industry into a globalized setting. Furthermore, the emergence of social healthcare systems and other healthcare infrastructure helped structure industry growth and soon the industry welcomed ‘blockbuster’ drugs with record sales. After tracing the further growth of the industry through the Second World War, Walsh continues by highlighting how pharmaceutics have grown from their aforementioned humble beginnings to one of the largest, most powerful industries in the world.

Yet, this growth has been met in recent years by tremendous challenges with respect to highly prohibitive research spending, rising regulatory costs, and stagnant, if not decreasing ‘productivity’ at the level of the marketability of new molecular entities. Pharmaceutical research and development spending clearly rose exponentially from the 1980s to the mid-2000s and that rise was subsequently followed by a consistent plateau in spending in the last decade (Fig. 4). Additionally, when this trend in spending is correlated to newly approved drugs, no major trend can be observed which in essence indicates that the industry’s higher budgets - as compared with previous decades - do not have any noticeable impact on the ‘productivity’ in discovering new drugs. According to Tufts Center for the Study of Drug Development, the cost of developing a new drug, both in the pre-clinical and clinical stages is skyrocketing (Fig. 5) and is now estimated to be close to 2.5 billion USD. It is important to note that throughout the 1970's and 80's preclinical costs outweighed clinical costs but this trend was reversed in the 90's and 00's. This trend may indicate that the regulatory environment tightening requirements towards safety or efficacy requirements.

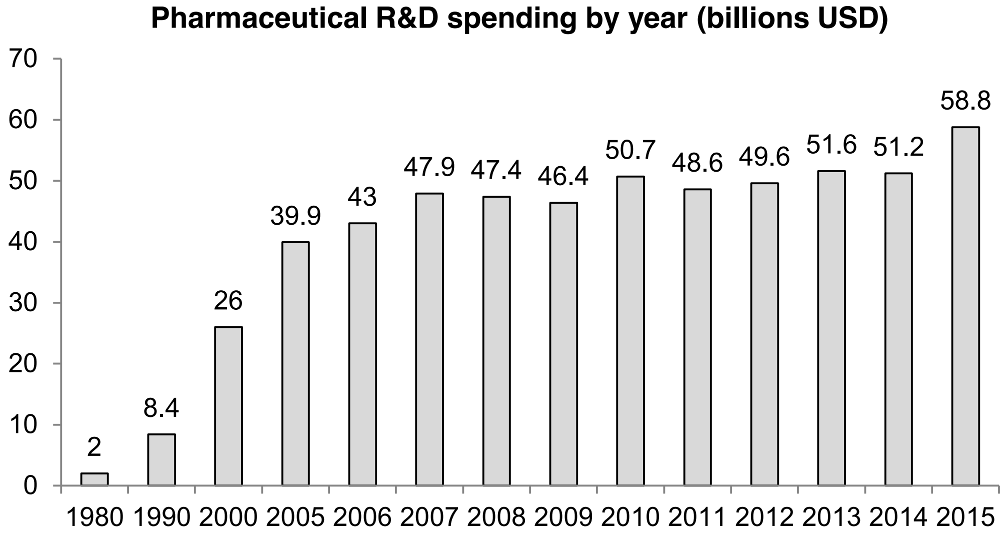


Figure 4: R&D spending representative of 37 global pharmaceutical companies (PhRMA)

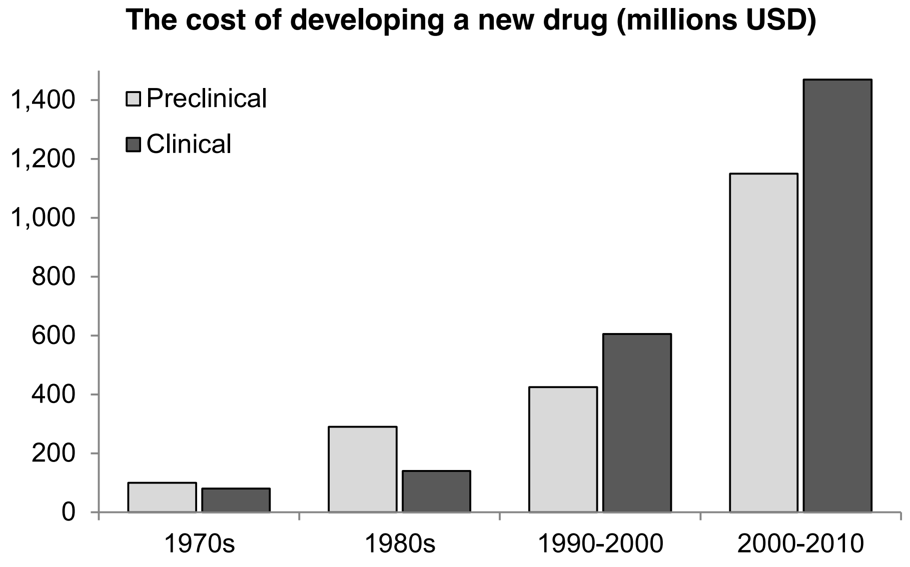


Figure 5: Tufts Center for the Study of Drug Development cost estimates for developing a new drug. Estimates in 2016 range from 2.6-2.8 billion USD for the full discovery and development of a new molecular entity (NME).

A more specific question that merits one's attention is where specifically in the drug discovery and development pipeline are costs greatest and therefore most prohibitive? In 2010, Paul *et al.* developed a model to estimate R&D costs at differing stages along the drug discovery and development process. Their model was based on a set of industry-appropriate R&D assumptions (industry benchmarks and data from Eli Lilly and Company) and sought to define 'performance' of the R&D process at each stage of development. In total, they estimated that to launch a new molecular entity (NME) in 2010 costs more than 1.7 million USD with the largest portion of capitalized cost/launch (an estimated 23%) coming from 'lead optimization' (Paul *et al.*, 2010). 23% represents the largest percentage of total cost among the eight stages (target-to-hit, hit-to-lead, lead optimization, preclinical, phase I, phase II, phase III, and submission to launch) Paul identifies. With such a sharp increase in costs, does it mean natural products research should be downscaled? Cragg and Newman, two highly respected figures in the world of natural product development believe this heavy investment into leads can be mitigated by "…expanding, not decreasing, the exploration of Nature as a source of novel active agents which may serve as the leads and scaffolds for elaboration into desperately needed efficacious drugs for a multitude of disease indications" (Cragg and Newman, 2012).

2.1.1.2. Outsourcing research and development risk through mergers and acquisitions

Another separate yet related trend of the pharmaceutical industry emerging in recent decades has been that of accelerating mergers and acquisitions (M&A). One could assume that if cost is a risk, as drug discovery and development costs rise, cost-associated risk rises as well. In fact, this trend between rising cost and risk, in the context of M&A, has become a key research area by researchers such as Danzon *et al.* (2004) and Higgins and Rodriguez (2006) which focus on M&A effectiveness and the role of M&A in driving R&D productivity respectively.

The effectiveness of these increasingly common M&As are regularly challenged and generally doubted by industry observers. Patent expirations, gaps in a company's product pipeline, and financial struggles are all reasons common reasons for M&A activity. Interestingly, reasons such as these have collectively been labelled as a 'desperation index' by a number of industry observers. One does not have to look far to see the never-ending challenge to sustain and grow in a tightly regulated, fiercely competitive operating environment and it is precisely for this reason that 'effectiveness' of such activity is scrutinized so closely. For example, do larger corporations with more resources at their disposal typically drive innovation or generate value in ways more beneficial to some or all aspects of the drug discovery process? To what extent does M&A activity adequately respond to the issues it seeks to address?

Before addressing the 'creating value' proposition often put forth to justify pharmaceutical industry M&A activity, Higgens and Rodriguez (2006) summarize a few theories surrounding value creation in general:

> A significant quantity of research has been dedicated to understanding for whom and how value is created through acquisitions. Many theories have emerged, for example, the monopoly theory of mergers (Mueller, 1985; Eckbo, 1992; and Ravenscraft and Scherer, 1987); the synergies approach (Bradley *et al.*, 1989); economies of scale (Ravenscraft and Scherer, 1989; Houston *et al.*, 2001); to gain market power (Anand and Singh, 1997; Baker and Bresnehan, 1985; Barton and Sherman, 1984); redeployment of assets (Capron, 1999); and, diversification (Berger and Ofek, 1995).
>
> The conclusion one draws from the bulk of the research focusing on whether value is "created" or "destroyed" is that the return to acquiring firm shareholders, on average, is essentially zero. The majority of the value flows to the target firm shareholders (Jensen and Ruback, 1983; Brickley and Netter, 1988; Bruner, 2002). Relatively few studies have been able to demonstrate meaningful value gains on behalf of acquiring firms in non-tender offer acquisitions. Relatively few studies have been able to demonstrate meaningful value gains on behalf of acquiring firms in non-tender offer acquisitions. Andrade et al (2001) suggests that the underlying strategic motivation for a particular transaction may provide a fruitful avenue for identifying how value is created through acquisitions for acquirer shareholders.

Furthermore, Higgens and Rodriguez focus their analysis on one of the primary motivations of M&A in the pharmaceutical industry; namely that of M&A as a method for outsourcing R&D. Within their sample of 60 public firms that formed at least one strategic alliance between 1994 and 2001, they identify examples of

M&A as a facilitator in fueling R&D activity in cases such as Gilead Sciences' acquisition of Triangle Pharmaceuticals and Merck's acquisition of Aton Pharmaceuticals Inc.. They conclude that 'on average…companies experiencing a deterioration of their research pipeline and product sales were more likely to engage in an acquisition. Moreover, these firms were either able to stabilize or to reverse the pipeline declines that they were experiencing' which affirms that M&A is an effective strategy in specifically boost R&D activity. It is important to note that Higgens and Rodriguez's analysis does not take into the account other aspects of companies such as the financial performance of the companies involved; nevertheless, other researchers have focused their research on such gaps.

In 2007, the Merck Company Foundation funded three prominent economists at the University of Pennsylvania, Yale University, and Cornell University (Danzon *et al.*) to perform a thorough analysis of this wider angle of overall firm performance. Their unprecedented research sampled '383 firms in the pharma–biotech industry and 165 'transforming mergers', defined as transactions that are sufficiently large that post-merger integration will require reorganization and potentially have an observable impact on accounting measures of performance'. The specificity of their analysis: 'distinguishes between small biotech firms and large pharmaceutical firms, because small firms, which account for almost half the firms in our sample, face very different production and cost functions'. In essence, their study established several competing hypotheses to 'explain firm-specific merger activity and to generate a measure of each firm's propensity to… [merge]' and then '[measure] the effects of mergers on a range of performance measures' their one key finding was 'although merger in the pharma–biotech industry is a *response* to being in trouble for both large and small firms, there is no evidence that it is a *solution'* for the issues it seeks to address.

In 2007, another group of researchers from the University of Sheffield and Bahcesehir University also assessed the overall effectiveness of the M&A activity in the pharmaceutical industry (Demirbag *et al.*, 2007). The difference between this group's analysis by the aforementioned analysis from Danzon *et al.* was that it not only includes Danzon *et al.*'s financial performance analysis (in the form of profit analyses), but also augments it with research productivity and return on investment (ROI) metrics. Table 4 summarizes their key findings and compares them with previously published research. Their findings with respect to research productivity, ROI, and profit margins largely support previously published (both pharmaceutical industry and other industry) data and show that commonly perceived value generation in M&A such as increased research productivity, increased ROI, and increased profit margin are not as obvious as many perceive.

Findings from Demirbag *et al.*	Existing literature
Research Productivity	
M&As exhibited poorer results than their pre-M&A firms and non-M&A rivals	Supports the work of Hitt *et al.* (1991) that M&As have negative effects on firm innovation and lack
No value creation from M&A	Supports James' (2002) argument about synergy creation. Supports Bergren's (2003) proposition that integration and harmonization issues will affect R&D productivity. Supports Sudarsanam (2003) that decline in innovation output is due to provision of improper technology inputs and poor integration management. Supports Gaughan (2001) that M&A might not be the ideal alternative to access the necessary resources. Contradicts the findings the finding of Higgins and Rodriguez (2004) that the pursuit of M&A pipeline helps stabilize or reverse pipeline decline.
ROI	
M&As exhibited poorer results that their pre-M&A firms	Supports the finding of Heracleous and Murray (2001) that ROI of M&As appears to be lower than their pre-M&A firms.

M&As fared better than their non-M&A rivals	Supports Dickerson *et al.* (1997), James (2002), King *et al.* (2004) that M&As do not have a net beneficial effect on company profitability.
M&As were not better than their pre-M&A firms but better than their non-M&A rivals	Supports Healy *et al.* (1992) that M&As experience improvements in asset productivity as compared with non-M&A rivals. Contradicts the findings of Gaughan (2001), Capron (1999), Chatterjee (1986) that synergy in terms of economies of scale and scope.
Profit Margin	
M&As performed better than their pre-M&A firms and almost on par with non-M&A rivals	Supports the work of Singh and Montgomery (1987) that value creation occurs after M&A activity. Contradicts the findings of Danzon *et al.* (2004), Heracleous and Murray (2001), and Dickerson *et al.* (1997) that M&As experienced slower operating profit growth, no change in enterprise value or turnover

Table 4: Summary of Demirbag's findings on the effect of M&A on research productivity, ROI and profit margin as compared with existing literature. shows that commonly perceived value generation in M&A such as increased research productivity, increased ROI, and increased profit margin are not as obvious as many perceive (Demirbag *et al.*, 2007)

Thus, examining Danzon, Higgens and, Dermirbag's findings in a more cohesive manner suggests that there is general dissatisfaction with M&A activity creating substantial and enduring value or innovation within the pharmaceutical industry. Rather, M&A is perceived as a measure taken to decrease risk and as Demirbag's research shows actually can *decrease* research productivity (defined as number of NMEs developed in relation to total R&D expenditure). This reported decrease in research productivity both pre- and post-M&A activity is not only of concern at an industry-level but negatively impacts specific research areas such as natural products research. As reported by Bruno *et al.* in 2014, these programs are already seeing sharp decreases in funding, and likely as a symptom of the aforementioned trends, have virtually been abandoned by all major global pharmaceutical companies (this point will be covered later in this chapter). This point is ironic considering natural products are historically the most consistent source of pre-cursors and leads for modern drugs. This point between the dissonance between current prevalent strategies and perceptions of usefulness of these strategies by stakeholders will be explored in sections to come.

2.1.2. Micro trends: Industry and academia

2.1.2.1. Druglikeness: emerging in the last two decades

Advances made in the field of human drug discovery over the last century have been unprecedented. It is difficult to doubt the significance of advances in approaches to the development of drugs and medicines as a whole. High-throughput screening and other technological advances have only worked to add momentum to the field's development. Yet, what is interesting is that many observers agree that this spur in momentum, specifically in terms of launching new drugs to market, has in a large sense plateaued out. Often citing reasons related to costs and regulations associated with of discovery and development of drugs, many analysts have a bleak outlook when it comes to the future of drug discovery. A direct result of this 'drying up' of the drug pipeline was a shift towards high-throughput screening and in essence, a scramble to screen as many products through molecular, ligand-based or other popular bioassays (such as a cytotoxicity MTT assay). The torrent of data and published papers that followed allowed researchers to view drug development through an unprecedented lens. Now that thousands of compounds were being screened weekly (and eventually daily) one could theoretically see an increased correlation between screened compounds and drugs successfully making it to market. With all of this new data how was 'market potential' characterized?

Much of the debate in drug development in the last 15 years has revolved around these trends, particularly with respect to modelling 'druglike' properties (commonly referred to as 'druglikeness'). As one explores the literature, it is very clear that what exactly druglikeness entails really depends on the intended application of the compound. Properties appropriate for successful metabolism of an orally administered drug differ greatly from for example, transdermal injections. This is precisely why popular rules, such as Lipinski's Rule of Five (Ro5), have been fiercely debated (Kenny and Montanari, 2013). Before exploring some of these rules, it is important to also understand that there is a significant counter voice to these rules and the presentation of them does not imply that all those interested in discovering new drugs abide by them. Despite the wide popularity of these rules, there are many who cite exceptions and trends which clearly go against it (i.e. the number of new chemical entities reaching the market has remained constant or continued on a downward trend) (Brown and Superti-Furga, 2003). For example, Abad-Zapatero's following viewpoint illustrates the doubt and scepticism by many with regards a magic formula with regards to drug discovery:

> Rules, commandments or absolute certainties are dangerous. Even if they can be found to be approximately true at the beginning, they are eventually superseded by a deeper understanding of the problem. They might provide an initial guidance of our decisions or actions but not a solid compass for long and tortuous journeys, because if we follow them strictly they might obscure our way to discovery. Navigating through reality is much more suitable and requires a delicate balance between rules and *insight* (Abad-Zapatero, 2007).

Nevertheless, in understanding what role natural products have played and will continue play in their development as drugs it is helpful to analyze current prominent trends with relation to these rules (and permutations of such rules) in modern drug development (Table 5).

Rule/Publication	Year Proposed	Summary	Application/Scope	# of Citations[1]
Lipinski's Rule of Five (Lipinski *et al.*, 1997)	1997	$\leqslant$5 HBD $\leqslant$10 HBA $\leqslant$500 MWT $\leqslant$5 CLogP	Orally active drugs in humans	7,462
A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. A Qualitative and Quantitative Characterization of Known Drug Databases (Ghose *et al.*, 1998)	1998	40-130 molar refractivity 20-70 total atoms 160-480 MWT -0.4-5.6 CLogP	Central nervous system, cardiovascular, cancer, inflammation, and infection disease drugs	526
Molecular Properties That Influence the Oral Bioavailability of Drug Candidates (Veber *et al.,* 2002)	2002	<10 rotatable bonds <140 Å^2 PSA (or <12 HBA +HBD)	Oral bioavailability in rats	1,291
Computational approaches to the prediction of the blood–brain distribution (Norinder and Haeberlein, 2002)	2002	$\leqslant$5 N+O atoms >0 log P - (N+O atoms)	CNS-active substances (penetration through the BBB)	187

[1] # of citations as searched on http://scholar.google.com in March 2014

A 'Rule of Three' for Fragment-based Lead Discovery (Congreve, 2003)	2003	$\leqslant 3$ rotatable bonds $\leqslant 3$ HBD $\leqslant 3$ HBA $\leqslant 300$ MWT $\leqslant 3$ CLogP	Fragments leading to drug candidates	435
A Comparison of Physiochemical Property Profiles of Development and Marketed Oral Drugs (Wenlock, 2003)	2003	$\leqslant 4$ HBD $\leqslant 7$ HBA $\leqslant 473$ MWT $\leqslant 5.5$ CLogP $\leqslant 4.3$ CLog $D_{7.4}$	Marketed orally active drugs in humans	342
In silico ADMET Traffic Lights as a Tool for the Prioritization of HTS hits (Lobell *et al.*, 2006)	2006	$\leqslant 7$ rotatable bonds <140 Å^2 PSA >50 mg L^{-1} solubility $\leqslant 400$ MWT $\leqslant 3$ CLogP	Small-molecule oral drugs	33
The Rule of Five Revisited: Applying Log D in Place of Log P in Drug-Likeness Filters (Bhal *et al.*, 2007)	2007	<5 HBD <10 HBA <500 MWT <5 CLogD	Orally active drugs in humans	57

Table 5: Overview of rules of thumb for druglikeness/lead-likeness

As already alluded to, another aspect of such an analysis of modern rules of thumb is their general non-applicability to natural products. The discrepancy between the latent potential of natural products as drug leads and the proliferation of druglikeness rules which are for the most not applicable to natural products is increasingly apparent. As Zhang and Wilkinson bluntly state:

> A major class of drug molecules that are excluded from the original analysis which generated the 'rule-of-five' are natural products. This is a serious defect as it is known that a large percentage of marketed drugs are natural products and their semisynthetic derivatives (Zhang and Wilkinson, 2007).

Parallel to this process of searching for physiochemical factors indicative of druglikeness in synthetic and semi-synthetic compounds was a consistent conviction, by a number of researchers, of the potential of natural products as drug leads. In foreseeing limitations in traditional approaches to developing synthetic and semi-synthetic compounds, some earlier than others, turned their attention to natural products development hoping that this vast, diverse source of compounds could meet the needs of an ever-growing global pharmaceutical industry. One of the earlier studies in 1998 showed that of the 520 new drugs approved between 1983 and 1994, 39% were natural products or derived from natural products and 60–80% of antibacterials and anti-cancer drugs were derived from natural products (Cragg *et al.*, 1997). Another study from 2000 estimated that less than 10% of the world's biodiversity has been tested for biological activity, thus outlining the vast potential that natural products hold (Harvey, 2000). Artemisinin, triptolide, celastrol, capsaicin, and curcumin were singled out by Corson and Crews (2007) to highlight the vast potential of traditional medicines in driving the modern drug discovery process through 'traditional knowledge'. Fewer studies have limited their focus primarily to the alkaloids but nevertheless the developmental potential of alkaloids has been a part of the debate on natural product development as a whole. In recent years, the outlook of natural product development remained generally optimistic among the academic community. As one scholar noted in 2008, 'natural products have been the single most productive source of leads for the development of drugs' and that to accelerate their development 'approaches are being developed to improve the ease with which natural products can be used in drug discovery campaigns, and data mining and virtual screening techniques are also being applied to databases of natural products' (Harvey, 2008).

Yet, interestingly such calls for an increased emphasis on tapping into the latent biodiversity by Harvey and others have largely remained unanswered. In the last five years, some such as Backlund, have begun probing into the data mining approach through chemical space or cross-database referencing exploration (Rosen *et al.*, 2009; Wolf and Siems, 2007). This being the case, data mining and database studies specifically focusing on secondary metabolites remain virtually absent from this debate.

When one widens this specific exploration into data mining approaches of alkaloids into that of an exploration of natural products in general, one can then begin to see that data mining/database approaches are abundant. Like Lipinski and the aforementioned others who proposed rules of thumb for specific pharmaceutical applications (i.e. oral drug candidates) there have been, albeit to a significantly lesser extent, some explorations into identifying trends and patterns in ‘drugs’ from natural sources. Unlike Lipinski and the aforementioned others, these natural product druglike propositions are non-specific. Lipinski’s initial dataset was comprised of Pfizer and Merck orally administered drug candidates. Trends in natural product development are quite non-specific. Reasons for a lack of specific ‘rules of thumb’ could include the extreme diversity of compounds in the natural product world, the challenges associated with sourcing them and thus a lack of comparable data across classes/sub-classes, or even a more fundamental challenge such as the one outlined by Jürg Gertsch in his paper “How scientific is the science in ethnopharmacology? Historical perspectives and epistemological problems”. In essence, Gertsch asserts that it is highly unlikely to be able to reproduce bioassay results in natural products research. He argues that there are simply too many variables particular to specific ethnopharmacological and natural product research settings which do not lend themselves to reproducibility. Therefore, how does one then equate or apply these findings to propel progress into the development of other natural products at large? Also, how do the potential of natural products fit into the current drug development paradigm which is heavily focused on ‘druglikeness’?

However premature these efforts, some have ventured to highlight trends with the intention of streamlining this drug discovery process. For example Zhang and Wilkinson outline the following trend related to drug discovery productivity in natural products:

> The high productivity of natural product-based drug discovery may be related to the fact that their chemical structures have been biologically prevalidated by evolutionary selection which defines structural prerequisites for binding to proteins. For example, an analysis of over 154,000 natural products showed that the majority of them have molecular volumes ranging from 100 to 500 Å^3 (Koch *et al.*, 2005). In the meantime, the volumes found in over 18,000 binding cavities of protein targets range from 300 to 800 Å^3. Therefore, the average volumes of natural products correlate with the average dimensions of protein cavities (note that protein ligands often do not fill the entire volume of a given protein cavity (Zhang and Wilkinson, 2007).

They continue with a more specific physiochemical analysis of properties leading to druglikeness and also include observations derived from Grabowski and Schneider's computer-based analysis of chemotype/molecular properties.

> For example, when analysed by either a size-independent 'chemistry space filter' or 'support-vector-machine' approach, natural products exhibit better scores of 'druglikeness' than synthetic compounds. Further, natural products contain on average twice as many oxygen atoms and three times fewer nitrogen atoms than synthetic drug molecules. They also contain a slightly higher number of hydrogen-bond donors than do synthetic drugs. Natural products contain approximately four times more chiral centers and far fewer aromatic rings, a fact which may engender upon natural products better selectivity when binding to stereo-defined sites (Grabowski and Schneider, 2007).

It is important to note that observations such as these are limited in that they are modelling relatively small datasets filled with compounds from *all* natural product classes. A small or large scale analysis, of how these or other patterns may apply to the development of secondary metabolites such as alkaloids, or a subclass of alkaloids has never been proposed.

2.1.2.2. Risky simplistic paradigms

Unsurprisingly, simplistic rules (such as those cited previously) are closely related to simplistic drug discovery paradigms. Few who have disagreed with the whole notion of strict rules and concepts aiming to hone in on druglikeness or lead-like compounds have proposed other viable, specific alternatives. For example, one general call to action by Abad-Zapatero (2007) is: 'We have to look deeper into the underlying principles that govern the interactions between drugs and targets'. This vague suggestion does not contain any information related to methods and/or techniques and is not particularly helpful in advancing the debate of the usefulness of druglikeness rules. He continues: 'I would argue that we have to look for more subtle variables and concepts than simple 'quinarian' or 'ternarian' rules in order to guide our efforts to discover better drugs more effectively…. We need to combine and reduce the multitude of variables spreading through the myriad of columns of our spreadsheets' (Abad-Zapatero, 2007).

This issue of simplistic, single-target paradigms which reduce the drug discovery process to a mere game of ligand 'hit or miss' is widely recognized as a problem and many have commented on its prevalence. In 2005, Samms-Dodd stated "there has been a tendency to focus narrowly on the target and to underestimate the complexity of the physiological role of the target in the intact organism. As a consequence the validity of the target was not questioned sufficiently, and this meant that programs have continued beyond the point where they could and should have been terminated – and this reduces the productivity of the industry" (Samms-Dodd, 2005). Overington *et al.* in 2006 and Medina-Franco in 2013 are a few of the many voices which share a similar conviction to those views espoused through Samms-Dodd's critique (Overington *et al.*, 2006; Medina-Franco, 2013).

The suggestion of looking for more subtle variables in natural products research and consolidating, without oversimplifying, large datasets is a primary aim of this thesis. Some have proposed other paradigms in drug discovery with more comprehensively look at additional determinate variables. One example is Hopkin's proposed 'network pharmacology' model proposed in 2008 which integrates 'network biology and polypharmacology' in order to 'validate target combinations and optimize multiple structure-activity relationships while maintaining druglike properties’ for drug leads (Hopkins, 2008). To build on such propositions and incorporate the aforementioned reservoir of potential in natural products research, a novel-screening paradigm involving a specific, robust biodiversity based screening paradigm for natural products is presented in a subsequent chapter of this thesis.

2.1.3. Rise and fall of interest/funding/etc. of natural products

Historically, natural products (natural product) development has been a field of immense interest to medical, commercial and scientific communities worldwide. As isolation and purification techniques advanced, natural products increasingly became prime candidates for drug leads and drug discovery efforts (Cragg and Newman, 1997). Their diversity and abundance characterized them as a virtually limitless source of novel lead compounds. Yet in the last decade, the majority of multinational pharmaceutical companies have reduced natural product Research and Development (R&D) expenditures (David, 2014).

Arrested prospecting programs	Continued prospecting programs
Abbott	Dabur
Astellas	Eisai
Bayer	Novartis
Boehringer Ingelheim	Otsuka
Bristol-Myers Squibb	Pierre Fabre
Daiichi Sankyo	Piramal
Eli Lilly	
GlaxoSmithKline	
Johnson and Johnson	
Kyowa Hakko	

Merck Sharp and Dohme	
Novo Nordisk	
Pfizer	
Roche	
Sanofi	
Servier	
Takeda	

Table 6: Big/medium pharma companies which ceased bioprospecting between 2000 and 2013 or were still bioprospecting in 2014 (David, 2014).

What are elements behind this trend and where are these reallocated efforts ending up? What are the common drivers and barriers in natural product development? How can efforts to understand such drivers and barriers (Amirkia and Heinrich, 2014) enhance the ability to further leverage the potential of natural products? If natural products have historically been such an important source of new drugs, what insights can one gain into the heavily academia-driven natural product drug development process as compared with the widely recognized slowdown of industry efforts? This research seeks to gain insight into these questions by directly soliciting the views of an unprecedentedly large panel of pharmaceutical industry experts who *currently serve* in senior positions in academic and or commercial organizations. Yet before such results are presented, it is important to further understand previous efforts in the context of natural development drug discovery within the pharmaceutical industry.

2.1.3.1. Bioprospecting and Shaman pharmaceuticals

Within the last few decades, there is no doubt that natural products research aimed to drive discovery and development of novel pharmaceutical products has, at times, sharply trended upwards. This increased effort has, without surprise, also been met with a considerable amount of critical analysis. One example of this critical critique of natural product development can be found in *Valuing Diversity of*

Medicinal Plants by Principe. It is important to remember that this analysis was put forth in 1991, a time when natural products research was beginning to gain momentum and not yet integrated at a large scale into the strategies of multinational pharmaceutical companies. Principe outlines three 'traditional factors that have influenced pharmaceutical research in the Western countries away from the direction of medicinal plants'. These being:

> 1. The screening-type of research required to identify a plant of interest and isolate the pharmacologically-active ingredient(s) is tedious, difficult, very costly, and often unrewarding, both for researchers and the company;
>
> 2. Once the ingredient(s) have been isolated, if it (they) cannot be wholly synthesized, the company then has to be concerned about securing its supply of raw plant material. This has become a considerable concern given that most plants of interest are usually indigenous only to developing countries, and dealing with these countries has become increasingly more complicated (and expensive); and
>
> 3. In some cases, especially where the active ingredient cannot be identified or isolated (an increasingly rare occurrence), the patent laws of most OECD counties, which do not permit the patenting of natural products, can present a barrier to marketing the new product. The pharmaceutical firm must either try to create a proprietary drug or modify the active mixture so that it can be patented, otherwise it would not be able to recover the cost of bringing a new product to market, through all of the laboratory and clinical testing and government regulation, which now costs about $100 million in the United States (Anon., 1985).

Principe reinforces this outlook on natural product development potential by citing the following comment by an executive in the pharmaceutical industry: in drug development from natural products, only between 1 in 10,000 to 1 in 40,000 compounds screened is likely to yield a marketable product (Principe, 1991). Homing in on a specific class of natural products, if this ratio can be agreed upon and it is assumed that there are about 27,000 alkaloids of which are randomly being selected from, one should expect an output between 0.6 and 2.7 alkaloids. This range is much lower than what one sees in reality. Though it must be understood that such an estimate was provided in the 90's when HTS and database modelling approaches were not as effective as they are now. Principe continues by stating: In the drug development business, failure is the rule rather than the exception. In contrast, a 2002 study of the ‘new drug discovery paradigm’ estimates that 1 in 1 million compounds go from HTS hit to marketed drug (Oprea, 2002). Merck chief executive P. R. Vagelos (1991) noted, ‘the odds against success, whether statistical or financial, are daunting. Most research projects fail’, as do many, if not most, start-up firms’.

Most of the critical analysis surrounding natural products development can be directly related to the following four motifs and concepts: ‘bioprospecting’, ‘indigenous rights (often referred to as ‘access rights’), ‘ethnopharmacology’, and ‘sustainability’ among others (Cox, 2008; Gertsch; 2009; Brito and Nunes, 1997).

Interestingly, all these correlate closely to the rise and fall of Shaman Pharmaceuticals, once regarded as the poster boy for revolutionizing the capitalization of ethnomedicine. Shaman spent over $90 million dollars in commercialization of natural products identified through bioprospecting programs and never was able to put forth convincing enough evidence to the FDA in commercializing its medicines (Clapp and Crook, 2002). The example of Shaman Pharmaceuticals has been studied in detail and is not included at a detailed level in this thesis. Rather this thesis aims to generate those new models which Clapp and Crook call for:

> This analysis of Shaman’s history suggests that risks inherent in drug development, company risk associated with Shaman’s drug development strategy, and technological change in the industry all contributed to its failure. The authors examine the opportunities and constraints encountered in bioprospecting and ethnobotanical searches and argue that natural products will remain important to drug development. Technological change, however, means that new models and new institutional structures are required for drug development based on natural products (Clapp and Crook, 2002).

One of the most fundamental challenges in natural product research is that of finding or ‘prospecting’ material of interest. Various strategies have been argued to be superior. Artuso encapsulates prospecting strategies used by pharmaceutical companies interested in natural product development into the following four methods: random, biomedical, ethnobotanical, and ecological (Artuso, 1997). His findings are summarized in Table 7:

Prospecting strategy	**Pros**	**Cons**
Random Search	-Effective when few leads exist -Can potentially uncover previously unknown and structurally unique families of compounds	-Low success rate per extract tested -Slow and time consuming (although HTS is changing this)
Taxonomic and Biomedical Information Search	-Relatively rapid and less expensive -Focused and targeted search	-Need clues and characteristics of a potentially effective compound -Often yields compounds similar to those already known
Ethnobiological Search	-Compounds have known therapeutic or toxicological effects	-Specific therapeutic objectives difficult to correlate to active compounds
Ecologically Informed Search	-Reduction in number of species and tests in the screening process	-Collection of material -Cost and time of gathering ecological information

Table 7: Summary of prospecting strategies in natural products and alkaloid research (Artuso, 1997)

The aim of this thesis is not to test the validity or effectiveness of each prospecting strategy in the larger context of natural product research, but rather gain a deeper insight into which of these prospecting strategies can potentially be most relevant to the development of classes of natural products such as alkaloids, which have been regarded as a highly productive class of natural products.

After all, the use of one or more of the aforementioned strategies depends on the specific application or goal of the screen at hand. The results of such an analysis may support the notion that a specific alkaloid, regardless of how it was 'bioprospected', has more or less an equal chance of being commercialized in the pharmaceutical arena. That is, that there is no significant correlation between prospecting strategy and development status of any given alkaloid; particularly with those which have made it to the marketed drug phase. By focusing on the alkaloids as a sub-class of natural products, rather than the several hundred thousand known natural products, it is hoped that one could better understand this challenging question.

Another concept which has proven a challenge in natural product research is that of indigenous rights. To what extend this hinders or enriches natural product research is highly debated. Some debate that intellectual property or ethical considerations only serve hinder the development process while others argue that ethnobotanical strategies serve to benefit indigenous populations. Clapp and Crook aptly summarize various aspects of this challenge by stating:

> Drawing on the ethnobotanical knowledge of indigenous peoples adds another layer of complications. Although there is no legal requirement under the Convention on Biological Diversity to gain the consent of indigenous people, the Convention strongly encourages it. Likewise, the 1989 Convention Concerning Indigenous and Tribal Peoples in Independent Countries and the UN Draft Declaration on the Rights of Indigenous People both support the principle of gaining the prior informed consent of indigenous people for the use of their resources and traditional knowledge (ILO, 1989; WIPG, 1993). Gaining the consent of indigenous people involves identifying appropriate indigenous communities with which to work, gaining their approval to share knowledge and resources, and negotiating appropriate contracts and compensation packages. This can be a very time-consuming undertaking. In the International Cooperative Biodiversity Group bioprospecting program operating in Peru, it took well over 2 years of negotiating to reach a final agreement with the indigenous partners (Crook, 2001). Perhaps the most problematic aspect of the Philippines' access legislation, at least from the point of view of companies seeking to gain

> access, is the requirement to gain the prior informed consent of local and indigenous communities to work in their territories (Reid *et al.*, 1995).

Thus, one must ask to what extent has this challenge above been or is relevant to the development of natural products? Are intellectual property issues relevant to the development of a few isolated natural products (ex. P57 in *Hoodia*) or do such considerations affect a significant portion of compounds of interest? Insights into these questions may very determine the extent and focus of future screens of natural products for drug discovery purposes. In the context of alkaloids, judging from the 51 alkaloids (listed in Appendix 3) which have made it into the pharmaceutical arena, nearly all are non-patented or patents have expired (which is expected since these products have been on the market for decades).

Most such alkaloids have no ties with indigenous populations and have become commodities which are freely manufactured and traded across worldwide markets. How does one quantify such indigenous rights considerations on a larger scale across widely differing contexts? One simple categorization scheme is to categorize each alkaloid into the following IP (intellectual property) statuses: 1. No registered IP belonging to a specific party, 2. Registered/claimed IP associated an indigenous or another group in a single country, 3. Registered/claimed IP associated groups in more than one country. Data for such a query would then come from one of the numerous patent search databases such as the World Intellectual Property Organization's PATENTSCOPE or Google Patents. Such a scheme, although simple, when applied to non-essential/essential groups of alkaloids may shed insights into whether or not alkaloids adhere to the trends described above in natural product development. This is a research area of interest which is not covered by the methods used in this research.

Another major challenge in drug discovery from natural products relates to the relationship between traditional medicines and marketed pharmaceutical products. The crux of this challenge lies in the following question: To what extent can naturally occurring plant preparations that have been used by ethnic groups be effectively developed into marketed pharmaceutical products, and how does the exponential rise of reported pharmacological activity of alkaloids contribute/counteract the development process of natural products?

A brief historical overview of what is debated ‘science’ by some may prove helpful in an initial exploration of these questions. In the 80s and 90s, some such as Farnsworth and others drew attention to the vast potential they believed ethnomedicine held for the future of drug development. The database he championed, NAPRALERT, is one tool to assist in identifying plants for inclusion into drug discovery programs. He explained that in North America there are two major reasons for studying ethnopharmacology: First, to utilize the information as a guide to drug development under the assumption that if a plant has been used by indigenous cultures over a long period of time, there should be a valid drug potential in the plant and second to validate scientifically their effects and side effects to a point where they can recommended for use in developing countries where they would be culturally acceptable and allow such countries to conserve hard currency and reduce health care costs (Farnsworth, 1993). This viewpoint was supported by many, including Tulp who lists plant species with traditional uses at the top of his list of unconventional natural sources which show the highest potential (Tulp and Bohlin, 2004).

It is important to note that both of these assertions are primarily based on data published in natural products related journals. Some such as Gertsch challenge this literature based approach and believe that ‘in ethnopharmacology and pharmacognosy literature we find thousands of claims of bioactive natural products and extracts with potential therapeutic applications. But hopes and promises are rarely tested on a rational basis’, and further state that ‘drug discovery lacks mathematical precision and ethnopharmacology is certainly not an exact science and interpretations are often ambiguous’ (Gertsch, 2009). Thus any serious data-driven analysis of natural products would acknowledge that such biases exist and try to mitigate them through a statistically significant volume of carefully collected data.

When one looks specifically at the alkaloids which have made it to the market as pharmaceuticals (Appendix 3) one sees that few have been strong ethnopharmacological leads, that is, many have emerged as a result of screening random or targeted physiochemical screening programs. Lastly, this debate is not simply divided into supporters and non-supporters of ethnomedical lead approaches, there are others like Cox who weigh strengths and limitations of such methods (Cox, 1994).

The concept of sustainability, as viewed in the context of extraction and cultivation of plant material, is another obvious challenge in natural product development (Cordell and Colvard, 2005). Similar to the other challenges mentioned earlier in this thesis, the specific task at hand is understanding, on a large scale, to what extent availability or extraction methods of a plant correlate with its ability to be developed and marketed as a drug. Again, similar to aspects mentioned earlier, the crux of such an issue lies in its ability to be quantified. What tools exist to quantify availability or biomass of plant species of every plant which is known to produce alkaloids? A large scale data-based approach is described in a subsequent chapter of this thesis which proposes a novel method which allows for a better understanding of this relationship.

In conclusion, Newman and Cragg state the direction they would like to see the natural products field develop in:

> To us, a multidisciplinary approach to drug discovery, involving the generation of truly novel molecular diversity from natural product sources, combined with total and combinatorial synthetic methodologies, and including the manipulation of biosynthetic pathways, will continue to provide the best solution to the current productivity crisis facing the scientific community engaged in drug discovery and development (Newman and Cragg, 2012).

This ‘multidisciplinary approach’ is proposed in subsequent sections of this thesis.

2.2. A study into perceptions of natural products: industry and academia

2.2.1. Results from insider perspectives

To validate the aforementioned assumptions regarding the advantages and challenges of natural product drug discovery, one needs to elicit feedback from stakeholders within industry. Furthermore, such insider perspectives will indicate current prevalent strategies within drug discovery processes as well as identify gaps which can begin to be addressed through various disciplines.

In order to begin gleaning such insights, a panel of industry and academic contacts (some of which are personal contacts) were personally invited through email and phone correspondence to participate and submit insights to a natural products development survey which was hosted online (Google Forms – http://forms.google.com). The participants were informed that results from the survey are to be used non-commercially, anonymously, and for the purposes of doctoral research. A screenshot of the survey can be found in Appendix 1.

A snowballing strategy was used to increase the number of contacts (Table 8). Industry contacts represented many of the major multinational pharmaceutical companies such as Merck, Novartis, GSK, Pfizer, AZ and Bayer among others. Seniority of each respondent varied with respect to his or her organization. Titles of respondents included: Chief Scientific Officer (CSO), President, Vice President (VP), Group Leader, Senior Analytical Chemist, and Senior Principal Scientist among others.

Academic contacts originated from eight different countries including, Brazil, Oman, New Zealand, UK, and USA. The majority of academic respondents were full-time academics, five of which also hold senior roles in pharmaceutical-company related organizations (consultancy, clinical research and/or pharmaceutical entities). It must be mentioned that the panel is clearly limited in its geographical coverage of smaller pharmaceutical markets such as Asia, Japan, and Latin America; markets which represented approximately 11%, 9% and 5% of 2014 total worldwide pharmaceutical sales respectively (IMS Health, 2014). Nevertheless, barring the extreme of labelling the panel as strictly representative of ‘the industry’, it is felt that the panel of contacts is generally representative of trends of interest within the industry (Amirkia and Heinrich, 2015).

There were four primary goals which were considered in designing the questionnaire:

1. To understand drivers and barriers in natural product drug discovery efforts

2. To understand what respondents identify as '*current preferred strategies'* for discovering new drugs in industry today

3. As HTS stands as a prevalent tool in in drug discovery today, the goal is to elicit *perceptions of the efficacy of natural products* as compared with other classes of compounds in screens

4. To understand the respondent's general outlook on future drug discovery as a whole. This approach would allow the authors to better understand the perceived effectiveness of past, present, and future natural product drug discovery efforts and more importantly compare any potential similarities and differences in insights between academic and industry respondents.

Age	Highest Educational Degree	Years of pharma industry exp.	Country	Size of current institution	Title	Natural product expertise comes from:
65	PhD / MD / PharmD / Doctorate	35	USA	11-100	CSO	Industry (Research)
58	PhD / MD / PharmD / Doctorate	25	USA	1-10	President	Industry (Research)
67	PhD / MD / PharmD / Doctorate	40	UK	1-10	Director	Industry (Management)
57	PhD / MD / PharmD / Doctorate	25	UK	999-10,000	Professor	Academia
67	PhD / MD / PharmD / Doctorate	40	UK	1-10	Director	Industry (Management)
52	PhD / MD / PharmD / Doctorate	26	UK	1-10	Project Manager	Industry (Management)
54	PhD / MD / PharmD / Doctorate	16	UK	10,000+	Doctor	Academia
45	PhD / MD / PharmD / Doctorate	9	USA	10,000+	Principal Scientist	Academia
72	PhD / MD / PharmD / Doctorate	40	USA	10,000+	Retired CSO	Industry (Research)
75	PhD / MD / PharmD / Doctorate	37	USA	10,000+	Retired	Industry (Research)
70	PhD / MD / PharmD / Doctorate	42	USA	1-10	Ph.D.	Industry (Research)
59	PhD / MD / PharmD / Doctorate	35	USA	10,000+	Research Fellow	Industry (Research)
49	PhD / MD / PharmD / Doctorate	22	Belgium	10,000+	Scientific Affairs	Industry (Management)
42	PhD / MD / PharmD / Doctorate	13	Switzerland	10,000+	Senior investigator	Industry (Research)
76	PhD / MD / PharmD / Doctorate	0	USA	999-10,000	Professor	Academia
59	PhD / MD / PharmD / Doctorate	30	Italy	11-100	Professor	Academia
62	Bachelors	15	USA	1-10	VP Analytical Chemistry	Academia
56	PhD / MD / PharmD / Doctorate	30	USA	101-999	Professor and Director	Academia
75	PhD / MD / PharmD / Doctorate	25	USA	10,000+	Retired Chief	Industry and Government
39	PhD / MD / PharmD / Doctorate	0	UK	101-999	Reader	Academia
58	PhD / MD / PharmD / Doctorate	30	Germany	10,000+	CVP	Industry (Research)
53	PhD / MD / PharmD / Doctorate	21	USA	11-100	Senior Scientist	Industry (Research)
51	PhD / MD / PharmD / Doctorate	22	USA	10,000+	Executive Director	Academia
47	PhD / MD / PharmD / Doctorate	0	USA	11-100	Sr. Research Scientist	Academia
54	PhD / MD / PharmD / Doctorate	28	USA	10,000+	Professor and Director	Industry (Research)
34	PhD / MD / PharmD / Doctorate	10	USA	10,000+	Managing Director	Industry (Research)
46	PhD / MD / PharmD / Doctorate	15	Germany	10,000+	Dr.	Industry (Research)
40	PhD / MD / PharmD / Doctorate	12	UK	11-100	Asso. Dir. of Discovery	Industry (Research)
57	PhD / MD / PharmD / Doctorate	27	USA	10,000+	Director	Industry (Management)
55	PhD / MD / PharmD / Doctorate	25	France	999-10,000	Dir. of Bot. and R&D Sourcing	Industry (Management)
29	PhD / MD / PharmD / Doctorate	1	USA	999-10,000	Research Scholar	Academia
42	PhD / MD / PharmD / Doctorate	13	Switzerland	10,000+	Director	Industry (Research)
42	PhD / MD / PharmD / Doctorate	13	Switzerland	10,000+	Senior Investigator	Industry (Regulatory)
35	PhD / MD / PharmD / Doctorate	5	USA	999-10,000	Research Scientist	Academia
45	PhD / MD / PharmD / Doctorate	20	India	101-999	Senior Manager	Industry (Research)
39	PhD / MD / PharmD / Doctorate	0	Oman	11-100	Associate professor	Academia
50	PhD / MD / PharmD / Doctorate	10	USA	101-999	Professor	Academia
38	PhD / MD / PharmD / Doctorate	0	New Zealand	10,000+	Senior Lecturer	Academia
60	PhD / MD / PharmD / Doctorate	10	USA	10,000+	CSO	Academia
52	PhD / MD / PharmD / Doctorate	20	Spain	10,000+	Director	Industry (Research)
52	PhD / MD / PharmD / Doctorate	25	UK	10,000+	Group Leader	Industry (Research)
60	PhD / MD / PharmD / Doctorate	29	USA	10,000+	Senior Principal Scientist	Industry (Research)
40	PhD / MD / PharmD / Doctorate	10	USA	10,000+	Chief Scientist	Industry (Research)
62	PhD / MD / PharmD / Doctorate	40	UK	1-10	CEO	Development
40	Masters	18	UK	999-10,000	Chief	Industry (Management)
60	PhD / MD / PharmD / Doctorate	37	Germany	999-10,000	VP	Industry (Management)
53.8	***Avg. age***	***20.6***	***Avg. years of pharma exp.***			

Table 8: Profile of selected respondents (46 of the total 52 respondents) sorted in order of received respons

The survey consisted of a series of six quantitative and qualitative close-ended questions followed by ten profile and background related questions. Close-ended questions with several choices had an 'other' box for the respondent to fill in his/her response. Multiple choice selections were displayed in randomized order for each survey so as to control for position bias in responses.

Close-ended questions were comprised of the following:

1. In your opinion, what are the top 2 current preferred strategies for drug discovery?
2. Based on your experience or on your assessment, approximately how many agents based on natural products and alkaloids researched in commercial R&D facilities make it to market as pharmaceutical products?
3. From your experience, what have been the major drivers to natural product development in industry?
4. From your experience, what have been the major barriers to natural product development in industry?
5. Drug Discovery is a history of triumphs and failures. Compared to last decades how successful is the industry today in discovering new medicines?
6. What is your outlook on the future viability (rate at which pharmaceuticals are developed and launched to market) of natural products, serving either as final pharmaceutical products or as leads to the development of the final pharmaceutical products?

The six close-ended questions each had an open field for participants to provide additional thoughts. Total completions of the survey ended at 52 responses after 14 weeks spanning from January and May 2015.

One major, consistent theme across respondents was the dissonance between the perceived potential of natural product development among individuals in industry and overall industry and/or company level strategies. Large scale, structural modification processes (i.e. HTS) have become Big Pharma's 'go-to-strategy’ for honing in on successful leads. HTS typically avoids the need to continuously source and verify new natural product material, which matches the highest citied barrier from industry contacts in the survey (i.e. a secure supply).

Additionally, large HTS screening programs are argued by many (Macarron, 2011) to be more cost-effective in the long run which is also in line with the third largest barrier cited by the industry contacts (i.e. cost/funding/budget). General results are summarized below (Table 9):

	Industry Respondents (n=33)	**Academia** Respondents (n=19)
Average Age	53	48
Average years of experience in the pharmaceutical industry	25	10
Male/Female	82% Male, 18% Female	89% Male, 11% Female
Drug Discovery is a history of triumphs and failures. Compared to last decades how successful is the industry today in discovering new medicines?	**Avg. of all responses (1=Full of Triumph, 7= Full of Failure)**	
	3.9 (SD:1.22)	4.3 (SD:1.37)
In your opinion, what are the top current preferred strategies for drug discovery?	1. High throughput screening (HTS) – **31%** 2. Physiochemical – modifications to existing leads – **25%** 3. Virtual/Computational prospecting/modelling **- 19%**	1. High throughput screening (HTS) – **34%** 2. Physiochemical – modifications to existing leads – **24%** 3. Virtual/Computational prospecting/modelling **-18%**

Top **Drivers** to natural product development in industry[2]	1. Structural Novelty and Bioactivity – **47%** 2. Efficacy and/or chemical viability (solubility, stability, toxicity, etc.) – **18%** 3. Supply - **11%**	1. Structural Novelty and Bioactivity – **42%** 2. Efficacy and/or chemical viability (solubility, stability, toxicity, etc.) – **24%** 3. Cost/Funding/Budget – **15%**
Top **Barriers** to natural product development in industry[2]	1. Supply – **26%** 2. Structural Complexity – **20%** 3. Cost/Funding/Budget - **19%**	1. Cost/Funding/Budget – **25%** 2. Structural Complexity – **23%** 3. Supply – **25%**
What is your outlook on the future viability (rate at which pharmaceuticals are developed and launched to market) of natural products, serving either as final pharmaceutical products or as leads to the development of the final pharmaceutical products? [3]	Optimistic - **52%** Unsure/'Hard to say' – **21%** Pessimistic – **27%**	Optimistic – **63%** Unsure/'Hard to say' - **26%** Pessimistic – **11%**

Table 9: Results of the survey including selected close ended and profile questions

Two other major themes emerged from participants writing insights in the open space provided after each close-ended question and are illustrated with a selection of verbatim statements pertinent to each theme:

[2] Respondents could select or list more than one response. All responses were added together and a 'response rate' was calculated by taking the percentage of a particular response as a total of all responses.

[3] Respondents selected one response. All responses were added together and a 'response rate' was calculated by taking the percentage of a particular response as a total of all responses.

Ineffectiveness of current HTS drug discovery programs (industry efforts which boast large libraries and cutting edge screening technologies have gained momentum which in turn has overshadowed smaller, more unique and fruitful discovery efforts):

- *"The industry focus on numbers (quantity vs. quality) has counted against natural products discovery - and the belief that supply of material on a suitable scale might be difficult (which may be a misconception)."*

- *"Industry is driven by numbers and processes; HTS, you could include fragment screening in this too – or even billions of compounds on encoded libraries (as we have at [X] company). I am part of a group who strongly advocate the huge impact of proper attention to physical properties and efficiency as existing leads are optimized (sadly mostly derived from the numbers generated above). HTS yet is the adopted strategy; in my opinion is probably isn't the most preferred!"*

- *"These approaches [High throughput screening (HTS), Combinatorial Chemistry] are favored by many pharmaceutical companies, even though they have not been notably successful."*

- *"HTS depends on large libraries, most of which have been so thoroughly studied that their utility going forward must be considered modest."*

- *"My understanding is that physicochemical modifications of existing leads represents the vast majority of drug discovery, and there are few places which are supporting anything beyond HTS or medicinal chemistry cycles."*

- *"Based on our internal track record, the outcome of HTS or VS is heavily dependent on the quality (control) of the actives and their ligand efficiency and the access to orthogonal assays to confirm the activity. These methods also complement each other and can be supported by additional methods, e.g. fragment-based. They are also generally easier to strip to the 'core' and obtain initial SAR. With natural products, you need to be lucky with the minor metabolites yielding some useful SAR.*

  Nevertheless, our experience at X University…screening endogenous X species was successfully generating leads that were NOT pursued as chemists perceived the SAR work to have low feasibility."

- *"In industry, modification of existing structures whether already in-house identified compounds or to bypass other structures with patent protection is much more common. This allows for the creation of "me too" therapeutic agents. Bioprospecting is much more common in academia, but natural product identification seems to be decreasing on the whole. Whether this is purely due to* *funding issues or a broader shift in the field is not certain. Similarly, virtual/computation approaches are used in refining structure in industry but are essentially never used to de novo identify a drug.*

Many academic labs have used such strategies as well, but with few successes. HTS is still fairly common place in industry and is gaining greater traction in academic settings with more and more universities creating screening facilities. Serendipity is certainly an important part of drug discovery, especially in areas such as neurology, but no one would bet on winning the lottery to fund their lab."

- *"HTS has been an abject failure in terms of discovery, due in most cases to not thinking about transfer across membrane issues when trying to go from a hit to an active in cells/animals.*

If one uses phenotypic screening (a dirty term amongst screeners in Pharma!), then if you see a valid effect, you will be well ahead of any HTS assay in vitro."

- *"In Pharma, HTS is the buzzword. I know of screens where over 1 million synthetic compounds have produced nothing, many times. Natural products in phenotypic screens are between 10,000 and 100,000 depending upon what is known about potential mechanisms etc."*

The second key theme that emerged centers around the lack of support/interest in organization for natural product drug development efforts (industry strategy over the last few decades has taken its form against a natural product-centric strategy and is unlikely to change)

- *"Don't fit company strategy."*

- *"Executive management fiat. Senior and executive scientific management at most Big Pharma wrote off natural products in the late '80s and early '90s with the advent of HTS, believing that HTS would have all of the answers."*

- *"Hostility; No support"*

- *"Lack of will to study them."*

- *"Natural product discovery tends to require a group to champion the approach. In my experience med chemists don't switch between synthetic chemistry and natural product chemistry. The latter requires an infrastructure and senior champions who believe in the potential of the approach. The novelty of the structures that result often go beyond anything that a med chemist might consider synthesizing as such this can take you to places you wouldn't have got to by any other route."*

- *“Screening of synthetic chemicals in massive libraries is cheap and most often results in hits that can be optimized as leads effective against sign targets. This process discounts any deep understanding of the biological processes involved in a disease state, other than the role played by an individual target biomolecule (kinases, etc.). And, the chemistry involved in elaborating these often simple structures is easy and high throughput - so from the chemists standpoint - why knock yourself out with natural product modifications which are often more difficult? Regrettably in industry little credit is given for the extra effort and overall productivity will appear low.”*

- *“The major driver to natural product development in industry is to eliminate it, which is what most of the large pharma companies have in fact done.”*

- *“From my perspective, today, natural products make only sense as starting materials for further optimization. I am convinced that we will see less and less original natural products that make it to the market in human pharma (animal health may be a different story). Also TCM et al may be a different story.”*

- *“natural products are currently not the "flavor of the month or decade" but now days, chemists are looking for structural leads that may well have activity, due to the failure of combi-chem as a discovery tool.”*

It is interesting to note that such an open question in fact only elucidated two key reasons why natural products are poorly represented in such drug discovery processes. Open questions are often used to elicit a wider set of views (Heinrich *et al.*, 2009) and here a clear focus on two concerns emerged.

2.2.2. Perceived ‘hit rates’ by compound class

Gaining insights into perceptions of the drivers and barriers of natural product drug discovery is a helpful yet limited step in providing insights into the drug development process. This data does not convincingly indicate the ‘effectiveness’ of natural products in drug discovery *as compared to* other commonly researched classes of compounds. Thus, respondents were asked to provide an approximate ratio of ‘success rates’ for several classes of compounds in the following way:

> *Based on your experience or on your assessment, approximately how many synthetic [or Biologics, Natural Products, Alkaloids] agents researched in commercial R&D facilities make it to market as pharmaceutical products?*

Interestingly, all perceived 'hit rates' for industry respondents are higher than those academia respondents reported (Table 10). Industry respondent 'hit rates' were higher at a rate ranging between 1.8-3.1-times. This may reaffirm the other finding that in general, senior stakeholders in industry typically do support natural product centric discovery strategies, and hence the more frequent 'hits'. Conversely, this indicates that screening programs related to academic efforts, particularly with respect to HTS, are not perceived as being as useful as widespread industrial efforts. Does this mean that industry is more 'productive' than academia in screening for natural products? Not necessarily, as this question does not attempt to equalize all screening methods but rather gain a general indication of respondent's *perceptions* towards screening efforts in the most general sense possible. Additionally, it is also surprising to note that there is a larger gap between 'hit rates' reported between natural products and synthetics for industry versus academia respondents; 8-times vs. 5-times, respectively. This also indicates a reaffirmation of the previous observation that many working in industry - regardless of their role and their level of dissatisfaction with the strategic direction of their organization, still perceive strong relative potential in natural product drug development as compared with currently prevalent synthetic-centric strategies.

Furthermore, the goal in asking this question is twofold; to gain a general indicator of perceived 'success/hit rates' of natural products against other compound classes as well as compare the perception of 'success rates' against previous claims published over the years by industry observers (Shen, 2003). Yet, there are two limitations to this question. The first is that each respondent may define 'researched' in a completely differing way. To one respondent a compound is not 'researched' until it perhaps enters a HTS program, while to another, a compound merely existing in a company compound library may count as being 'researched'. The second is the definition of the compound class (ex. Where does a natural product which has been structurally modified fit?). It goes without saying that there are numerous variables in any screen (compound library itself, target/ligands, parameters for defining a successful ‘hit’, purpose of screen, etc…) which make a particular screen entirely unique and incomparable to another.

Compound class	**Industry** respondents (n=33) (logarithmic value[4])	**Academia** respondents (n=19) (logarithmic value)
All natural products[5]	4.03 (1 in 10,723)	4.53 (1 in 33,598)
Alkaloids	4.27 (1 in 18,738)	4.53 (1 in 33,598)
Synthetic compounds	4.97 (1 in 93,260)	5.26 (1 in 183,298)
Biologics	3.67 (1 in 4,642)	4.11 (1 in 12,743)
Overall average	***4.24 (1 in 17,179)***	***4.61 (1 in 40,504)***

Table 10: Respondent’s estimates of how many agents researched commercial R&D facilities make it to market as pharmaceutical products (commonly referred to as ‘hit rate’)

35 of the 52 *respondents* (35 of all 105 *individual answers* to this open ended question) listed HTS as a ‘top preferred current strategy’; it is natural that this be a focal point of analysis.

Many publications have cited barriers to natural product drug development. In 2004 Jean-Yves Ortholand, who at the time worked at Merck in France, listed six major drawbacks in programs screening natural products: expense, time, novelty, tractability, scale-up and intellectual property (Ortholand, 2004). In looking to the industry feedback, each of Ortholand’s ‘drawbacks’ are corroborated to some extent with a particular focus on supply and cost/funding. It is noteworthy that these two highly cited barriers do not directly involve the actual screen itself but rather affect the feasibility of pre/post-screen efforts. The most frequently cited barriers seem to be those which prevent a screen from happening in the first

[4] Respondents selected from a range of six responses beginning at ‘1 in 100’ and ending at ‘1 in 1,000,000+’. Averages were calculated by assigning a value between 2 and 7 to each response and extrapolating through a logarithmic calculation (ex. 10^2=100, 10^6=1,000,000)

[5] All natural products includes alkaloids

place (i.e. cost or company strategy) or from moving from early stage screening to pre-clinical development (i.e. supply, scale-up). Therefore, besides proposing the obvious that costs should be reduced and/or funding increased for natural product drug discovery efforts, are there potential cost-sensitive resolutions to the supply/scale-up barrier?

2.3. Insights derived from the survey:

2.3.1. Can cost-effectiveness and natural products research coexist?

Supply as an unmet need has been mentioned not only by industry outsiders, but confirmed by industry and academic expert interviews presented earlier in this thesis. Yet, one question which deserves additional analysis is to what extent is the industry missing the target of cost effective natural product development. Is this a lofty goal, only to be attained is in the distant future? How do current efforts size up against sustainable natural products research and how actually 'sustainable' are current efforts?

McChesney *et al.* provide a number of highly pertinent insights into these questions through a presentation of two sets of in-depth analyses. The first titled 'the utilization of the world's plants', evaluates how much of the world's biodiversity are humans currently tapping into in the fields of food and medicine, and the second being a scenario analysis of biomass required for use of natural products in the treatment of acute and chronic conditions.

McChesney *et al.* begin with the assumption that 'it is generally estimated that there are approximately 300,000 species of higher plants'. This assumption is based on previous research by systemic botanists (Lawrence, 1951), and it is important to note that there is variability between estimates ranging from 250,000 to 500,000 species (Lawrence, 1951). Of these 300,000 species, only 1% has been utilized *in* food and 5% of that 1% has been utilized *for* food. Furthermore they state: the vast majority of caloric intake derives from about 20 species of plants. These plants represent the basis upon which the world's population is fed, representing a very narrow foundation supporting the world's human population. This first analysis continues with a comparison of food sources to sources of medicines. They state that 10,000 of the world's plants have documented medicinal use; more than three times more than what constitute main sources of human caloric intake but furthermore, *only* 150-200 species have produced agents used in western medicines. This is a highly significant indicator that shows the colossal potential yet uncovered by recent drug discovery efforts. So how cost-effective is it to tap into this reservoir?

McChesney continues with a second set of analyses looking at how much biomass of source plant material is required to carry forward initial assessments, verifications, clinical work, and eventual treatment of a relatively small patient population suffering from an *acute* condition. These estimates are summarized as follows (order of presentation is reversed from McChesney's original publication for ease of readability):

Step/Assumption	Quantity of material
Starting material – dry plant biomass	5 kg
Isolation yield from biomass	0.001%
Active substance	50 mg

Table 11 – Dry plant biomass required to perform initial efficacy testing of a natural product

Step/Assumption	Quantity of material
Starting material – dry plant biomass	100 kg
Isolation yield from biomass	0.001%
Active substance	1 g

Table 12 – Dry plant biomass required to perform initial efficacy testing of a natural product and secondary biological assays (toxicology and *in vivo* evaluations)

Step/Assumption	Quantity of material
Starting material – dry plant biomass	200,000 kg
Isolation yield from biomass	0.001%
Active substance	2 kg

Table 13 – Dry plant biomass required to perform initial efficacy testing of a natural product, secondary biological assays (toxicology and *in vivo* evaluations), and clinical environment testing (clinical trials)

Step/Assumption	Quantity of material
Starting material – dry plant biomass	2,000,000 kg (2,220 tons)
Isolation yield from biomass	0.001%
Active substance/year	20 kg
Patient population/year (acute condition)	10,000 patients
Grams/needed for course of therapy	2 g

Table 14 – Dry plant biomass required to treat an acute condition for 10,000 patients/year

This analysis continues with an additional assumption; a ten times larger patient population of 100,000 patients suffering from a *chronic* condition for which a patient would consume 50 mg/day for the course of the year. A ten times larger patient population would increase the requirement of starting material by two orders of magnitude to 200,000,000 kg of dry plant biomass! The authors argue that a required biomass of this magnitude is entirely feasible *if* one compares these volumes to the production of large scale commodities such as wheat, corn, or soybeans which hold the status of heavily developed mega-crops. The authors also state that the isolation yield assumption of 0.001% of dry plant biomass is the 'worst-case' and that this percentage may be up to a thousand times higher in certain species.

Compounding these three barriers - low penetration of the world's biodiversity, low rates of extraction, and low hit rates - reveals a large gap between current efforts in natural products development and what will actually sustain the field of pharmaceutics moving forward. Yet, the argument put forth in this thesis is that it would be incorrect to dismiss the potential of natural products development based on the current deficiencies in approaches used. As shown earlier, HTS is recognized by industry insiders as the preferred strategy but is widely believed to be inefficient in the face of supply issues. It is for this reason that some who have recognized these deficiencies have sought out efforts to devise novel methods of natural product synthesis which break away from the 'harvesting raw biomass' model.

2.3.2. Efforts to augment supply of natural products

A look at the historical record shows that before being isolated and identified as specific sources of active compounds, natural products were often used in relatively crude ways. As natural products were increasingly isolated, researched both in *in vitro* and *in vivo* settings and their effects gradually pinpointed, certain natural products became indispensible to the modern pharmaceutical environment. Alkaloids have been no exception and have many a time been at the forefront of new isolations, discoveries, and applications over the past two centuries and it is precisely such alkaloids that issues of supply (cited by industry insiders and observers) relate to the most. An illustrative example of this dynamic is the widely used pharmaceutical alkaloid morphine.

With use documented as early as the Byzantine Empire and isolated in 1804 (Hodgson, 2001), morphine represents one of the most studied and familiar alkaloids in the world (Hodgson, 2001). The role, usefulness and merits of morphine are not a focus of this thesis and thus will not be examined in depth, yet, because of a wide use over a long-period of recorded history, there is an abundance of data related to its supply and procurement. Both the natural supply and, more recently, synthetic sources of morphine deserve closer consideration if one is to understand the effectiveness of efforts to augment the supply of natural products, and more specifically that of alkaloids. Morphine can serve as an effective case study of such efforts.

The first total synthesis of morphine was published by Gates in 1952. Although this was an incredible breakthrough, this initial effort reported a very small overall yield of 0.06%

(Gates and Tschudi, 1952). No less than 25 additional unique total syntheses of morphine and/or its derivatives were published following Gates' initial publication (Table 15).

Principle author	Year	Target	Steps	Overall yield (as reported)
Gates	1952	Morphine	31	0.06%
Ginsburg	1954	*rac*-Dihydrothebainone	21	8.9%
Grewe	1967	*rac*-Dihydrothebainone	9	0.81%
Rice	1980	Dihydrocodeinone	14	29.7%
Evans	1982	*rac*-O-Me-thebainone A	12	16.7%
White	1983	Codeine	8	1.8%
Rapoport	1983	*rac*-Codeine	26	1.2%
Fuchs	1987	*rac*-Codeine	23	1.3%
Tius	1992	*rac*-Thebainone-A	24	1.1%
Parker	1992	*rac*-Dihydrocodeinone	11	11.1%
Overman	1993	Dihydrocodeinone	14	1.9%
Mulzer	1996	Dihydrocodeinone	15	9.1%
Parsons	1996	Morphine	5	1.8%
White	1997	*ent*-Morphine	28	3%
Mulzer	1997	Dihydrocodeinone	18	5.7%
Ogasawara	2001	Dihydrocodeineone ethylene ketal	21	1.5%
Taber	2002	Morphine	27	0.51%
Trost	2002	Codeine	15	6.8%
Fukuyama	2006	*rac*-Morphine	25	6.7%
Hudlicky	2007	*ent*-Codeine	15	0.23%
Iorga/Guillou	2008	*rac*-Codeine	17	0.64%
Chida	2008	*rac*-Dihydroisocodeine	24	3.8%
Hudlicky	2009	Codeine	18	0.19%
Magnus	2009	*rac*-Codeine	13	20.1%
Stork	2009	*rac*-Codeine	22	2%
Fukuyama	2010	Morphine	18	4.8%

Table 15: All total syntheses of morphine and morphine derivatives between 1952 and 2010 (Rinner and Hudlicky, 2011).

One key aim of this thesis is to understand the strengths and limitations of such approaches. In this respect, the analysis which follows will focus on two aspects related to the synthesis of morphine, namely time/complexity constraints and low yield.

148 years passed between the isolation of morphine from opium by Serteurner in 1804 its synthesis by Gates. Following isolation, it took nearly a century to elucidate the full structure of the compound. Today, technological innovations have ensured that isolation and structural elucidation of alkaloids are no longer major hurdles, but, simply looking at the large volume of discovered alkaloids and the time it would take identify chemical syntheses for each of the 51 pharmaceutical alkaloids (let alone the pool of 27,000+ alkaloids altogether) indicates that this is no small task. Granted, many alkaloids are structurally related, differ merely by one or two atoms, and can serve precursors to one another during these synthesis; for example, codeine and morphine. Yet, if one assumes that on average seven alkaloids are linked to one total synthesis process, that still leaves well over 3,000 chemical syntheses processes which have to be elucidated for all alkaloids to be covered. This number would likely be reduced to several hundred if only the 'essential' alkaloids used in industry, medicine and other fields were to be focused. Additionally this task is even appears more daunting when one considers the complexity of each synthesis. The list of syntheses in Rinner's review average 18 steps for each process; these are not simple or short syntheses, even though morphine is not considered a highly 'complex' alkaloid. Morphine has a molecular weight of 285 g/mol which is smaller and therefore assumed less complex than the average molecular weight of across all alkaloids in the Dictionary of Natural Products (DNP) which is 485 g/mol. This indicates that to develop additional chemical syntheses for the average alkaloid would be a considerably tougher task.

Another, and possibly even more straightforward, analysis into the effectiveness of these efforts is one which scrutinizes overall yield. An average of all overall yields for morphine and morphine derivatives in Rinner's review is a mere 5.44%. This average decreases to 1.79% when only the morphine syntheses are considered. Such a low-yield inevitably is cost prohibitive and as Rinner points out, is out computed by low wages for workers:

> To date there is no practical source of morphine, either by chemical synthesis or through fermentation, that would compete with the cost of isolation. Of course, part of the reason that natural morphine is so inexpensive is the low-wage investment in harvesting it, mostly in Afghanistan, Turkey, and India. Were the workers there paid "western" wages, the price could never be as low as it is today (~$400–700/kg).

It is important to also note that standards for reporting yields in syntheses such as these are not fully agreed on by chemists and thus each synthesis is not fully comparable (Wernerova and Hudlicky, 2010). Nevertheless, these reported yields are a good indicator which can benchmark overall effectiveness of such efforts and with an estimated global consumption of 474 tons in 2012; chemical synthesis of morphine at a yield in the range of 5% is simply not feasible (INCB 2013).

Rinner continues by summarizing efforts to augment supply of morphine through chemical and bio-synthesis by stating:

> Eight total syntheses of morphine or congeners have been reported in the last 5 years, attesting to no shortage of new ideas or strategies. The interest in this fascinating molecule will no doubt continue, yet a truly practical synthesis of the title alkaloid [morphine] still remains a distant dream. In order even to approach the current price per kilogram, a synthesis would have to be five to six steps long starting with commodity chemicals. A potential for a practical synthesis may exist in the realm of fermentation provided the biosynthetic pathway could be coded into a single plasmid and used to over-express the required enzymes in a robust bacterial carrier. A proof of principle has been attained through the work of Kutchan with the cloning and expression of codeinone reductase in *E. coli.*
>
> Another possibility for practical synthesis could come from the combination of fermentation for attaining specific steps with semi synthesis to complete the preparation. Currently, we are fully dependent on natural sources of morphine and all medicinally useful derivatives are made by semi synthesis. Perhaps more important goals for the future generations of chemists would be to focus on the de novo total synthesis of the derivatives themselves rather than morphine or codeine. Perhaps we will see some effort devoted to this most worthwhile task in the near future (Rinner and Hudlicky, 2011).

Efforts to boost secondary metabolite production through genetic modification have generally fallen short in putting forth viable solutions. Facchini *et al.* propose gene sequencing and combinational chemistry in their model based on genetically modified yeast producing secondary metabolites, such as alkaloids, but little is proposed beyond a theoretical approach and no experimental data is presented (Facchini *et al.*, 2012). Nakagawa *et al.* propose a novel scheme to leverage an *E. coli.* fermentation system which begins with simple carbon sources such as L-tyrosine and ends at (*S*)-reticuline, a branch point intermediate in the biosynthesis of many types of benzylisoquinoline alkaloids (Nakagawa *et al.*, 2011).

In their conclusion, Nakagawa comment on the viability of such an approach and challenges involved:

> Although the overall yield of (*S*)-reticuline in this system was low, we estimate that the production cost of (*S*)-reticuline using this method would be much lower than that of previous microbial systems in terms of substrate costs. In addition to reduced production costs, another advantage of this system is the simple and effective purification procedure, which results in (*S*)-reticuline with little contamination by other undesired BIAs. (*S*)-Reticuline was purified from the culture medium using solid-phase extraction and high-performance liquid chromatography. This procedure recovered more than 90% of purified (*S*)-reticuline in two steps. A simple purification procedure resulting in high yields makes this production system economically viable (Nakagawa *et al.*, 2011).

Thus, although this proposes a novel approach which builds on previous low-yielding set-ups through innovations during purification, overall yields remain prohibitively low. There is no doubt that such efforts are improving in effectiveness with time, but regardless, the practice of sourcing raw material from nature and isolating an alkaloid continues to showcase itself as the most economically viable sourcing strategy.

3. Alkaloids as a historical and modern source of medicines

3.1. History of alkaloids

3.1.1. Discovery and isolation

The subsequent chapters of this thesis focus on the alkaloids, a class of highly diverse and widely distributed natural products. This methodology culminates in alkaloid specific insights that are believed to be applicable to natural product drug discovery at-large.

The archaeological and historical record shows that peoples across Asia, Europe, and Africa used alkaloid-containing plants as early as 2000 BCE (Aniszewski, 2007). Specific applications included empirical medicines for animals and humans, sources of poison for hunting expeditions or executions (Wink, 1998), and a variety of other dietary uses which Cordell elucidates:

> 'we are aware that the use of alkaloid-containing beverages as stimulants (e.g. tea and coffee) is very old. Cultivation of tea, for example, probably dates back to the 12th century B.C.E. in Sichuan Province in China, where cultivation continues today. The legend of coffee, 'Qahwah', begins in Ethiopia, or may be derived from indigenous groups in Central Africa who used the stimulant properties of coffee beans during long treks (Quimme, 1976). Other alkaloid-based stimulants (khat, betel nut, etc.) may date back much further in time (Cordell *et al.*, 2001).

Following this rich historical record, alkaloids reached a turning point in the early 19th century with breakthroughs in their isolation, and structural elucidation purified compounds. In the early years of the 19th century, Friedrich Sertürner isolated what is now known as morphine. This catalysed a cascade of highly important isolations and discoveries by several European scientists including the isolation of xanthine (1817), strychnine (1818), atropine (1819), quinine (1820), and caffeine (1820) (Heinrich *et al.*, 2012). This burst of single compound isolation has been characterized by many, including Sneader, as 'the greatest advance in the process of drug discovery' (Sneader, 2005).

It would be a mischaracterization to merely associate the isolation and discovery of alkaloids with events belonging to the early 19th century. As overviewed in chapter 2, the vast majority of *discovered* alkaloids have not been studied in depth and alkaloid containing genera are abundantly under neglected. Thus, when one begins looking at efforts to isolate and discover new alkaloids following that initial burst of discovery, one can clearly see a steady stream of discoveries up through the early 21st century.

In 2008, Clement *et al.* discovered three completely new alkaloids from the tropical ascidian *Lissoclinum* cf. *badium* - isolissoclinotoxin B, diplamine B, and lissoclinidine B – which, in their tests, stabilized the tumor suppressor p53 (Clement *et al.*, 2008). In 2013, Wang *et al.* discovered and elucidated structures of nine new alkaloids isolated from the club moss *Lycopodium japonicum* Thunb. They found that although these nine alkaloids were completely novel, they shared structural elements related to lycopodine (Wang *et al.*, 2013). This discovery was followed most recently in 2016 by Zhang *et al.*'s discovery of two new monoterpenoid indole alkaloids, named 14,15-dihydro-14β,15β-epoxy-10-hydroxyscandine and 15α-hydroxy-meloscandonine (Zhang *et al.*, 2015). Both were isolated from isolated from the aerial parts of *Melodinus hemsleyanus* Diels and were tested for PTP1B inhibitory activity in the hopes of applications related to type-2 diabetes therapies. These recent discoveries attest to the fact that the natural products and alkaloids 'well' is certainly not dry.

3.1.2. Definition and distribution

As mentioned earlier, there is considerable variation in definitions of alkaloids as a subclass of natural products. For the purposes of this thesis, alkaloids are defined as: "a large group of nitrogen-containing secondary metabolites of plant, microbial or animal origin. These include the majority of nitrogen-containing natural products with the exception of the simple amino acids, proteins and nitrogen-containing substances of polyketide origin such as the aminoglycoside antibiotics" (Buckingham, 2000). Such a definition is relatively broad and thus, one has to reference specific natural product datasets to specify concrete numbers as to how many alkaloids actually have been 'discovered'.

Much of the uncertainty of how many alkaloids actually exist stems from various issues including: poor chemical identification or structure elucidation, lack of dereplication, chemical ambiguities, and the varying definitions of what exactly constitutes an alkaloid (Rates, 2001). As with natural products as a whole, many have proposed differing classificatory schemes for alkaloids. One popular scheme divides the whole class of compounds into three categories:

- True alkaloids (compounds which derive from amino acid and a heterocyclic ring with nitrogen),
- Protoalkaloids (compounds, in which the N atom derived from an amino acid is not a part of the heterocycle), and
- Pseudoalkaloids (compounds, the basic carbon skeletons of which are not derived from amino acids).

One of the largest compilations of discovered and recorded alkaloids yields 27,783 compounds. This collection is housed in the (DNP) and it is important to note that the DNP's categorization scheme differs from many others in that it categorizes based on biogenetic origin rather than purely on the basis of structural features. This could explain why the raw number of alkaloids in the DNP is significantly more than other datasets. A full breakdown of the dictionary's specific categorization scheme can be found in Appendix 2.

The scope of this thesis encompasses all such variations in definitions by taking the widest categorization of alkaloids as a class of compounds; the 27,783 found in the DNP (as of April 2014).

Alkaloids are not ubiquitously distributed in plants but rather represent a class of richly diverse compounds which are found far and wide. For example, Cordell *et al.* (2001) performed an in-depth taxonomical distribution analysis of all alkaloids contained in NAPRALERT (which at the time of Cordell's study contained 26,900 alkaloids in total). This analysis revealed that 67 of 83 of higher plant orders contain alkaloids and that 'alkaloids are distributed in 7,231 species of higher plants in 1,730 genera (approx. 14.2%) within 186 plant families'.

3.1.3. Use as non-medicines

Alkaloids may best be known throughout history for their medicinal properties and more modernly as excellent drug leads. Their medicinal applications will explored in subsequent sections of this thesis. Yet, what is often overlooked is the use of alkaloids in various other non-medicinal applications. As a highly diverse and widespread class of compounds, alkaloids are increasingly and used for, as one scholar encapsulates it, ‘murder, magic and medicine’ (Mann, 1992).

For example, the highly potent toxicity of some alkaloids have lent them to be used as very useful poisons in ancient times and more recently as pesticides when applied in lower concentrations to agricultural settings. Benzophenanthridine alkaloids, which belong to the larger group of isoquinoline alkaloids, are a class of alkaloids which have shown promise with regards to controlling fungal diseases of garden and ornamental crops (Howell *et al.*, 1973). Sanguinarine and chelerythrine are two commonly used ingredients in a number of pesticides which demonstrate strong antifungal activity (Newman *et al.*, 1999). Some have also cited related insect-repellent activity of various alkaloids from the diterpenoid and norditerpenoid alkaloid families (Isman, 2006; Ulubelen *et al.*, 2001).

Alkaloids have also commonly been used as tranquilizers (in animal husbandry and more recently clinical settings) and one of the most notable is the curare alkaloid *d*-tubocurarine. Initial descriptions of this strange substance by explorers were surrounded in a shroud of mystery:

> The exact origin of this "flying death" is still veiled in mystery. Its actual preparation is surrounded by all the esoteric magic and superstition of these strange people descended from the Aztecs, and very little more is known about it now than was discovered by Charles Waterton, a traveller of Lancashire origin, in his journey to the wilds of Demerara in 1812-a journey undertaken with the special object of investigating the origin and preparation of the Wourali poison (Gray *et al.*, 1946)

And its isolation summarized by Mahfouz:

> The need for isolating a purified active principle from crude curare started at an early date in South America, when Boussingault and Roulin (1828) succeeded in obtaining a bitter principle which they differentiated from strychnine, isolated eight years previously. Although the problem was somewhat clarified by the work of Preyer (1865) and Boehm (1886, 1897), it was not until 1935 that the active alkaloidal salt, dextro-tubocurarine chloride, was isolated in a pure crystalline state by King from a sample of native tube-curare. The same alkaloid was obtained in a good yield by Wintersteiner and Dutcher (1943) from a single plant species, *chondrodendron tomentosum*, which is probably its chief botanical source (Mahfouz, 1949).

And its uses and mechanism of action, expounded on by Sobell:

> [it] has been used for centuries by South American Indians to prepare poison arrows for hunting wild animals for food. Death results from respiratory paralysis and subsequent asphyxiation. Its major action is the interruption of transmission of a nerve impulse at the neuromuscular junction. This is thought to reflect complex formation between the drug and cholinergic receptors located at the postjunctional membrane, competitively blocking the transmitter action of acetylcholine (Sobell *et al.*, 1972)

With such a wide array of diverse and abundant alkaloids, many of which have minimally been researched in formal settings, it is certain that additional applications will emerge over time.

The process of drug discovery as it stands today differs greatly from the ones prominent throughout most of the 20th century decades. Highly popular, yet debated empirical rules aiming to enhance the selectivity of drug candidates have for many years been in the spotlight. As mentioned previously, popular terms such as ‘lead-like’ and ‘druglike’ have gained prominence though the work of Lipinski and Congreve (Lipinski, 2000; Rees *et al.*, 2004). As one explores the literature, it is quite clear that what exactly druglikeness entails really depends on the intended application of the compound. Properties appropriate for successful metabolism of an orally administered drug differ greatly from, for example, transdermal injections. The applicability and application of such rules to other research areas is an active debate in drug research and development.

One conspicuously lacking class of compounds in this debate about druglikeness and associated rules has been natural products, which, however, are well known to be of major importance as drugs (Cragg and Newman, 2005; Newman and Cragg, 2007). It could be argued that the sheer diversity of natural products does not allow for adherence to such rules, yet nevertheless the importance of natural products (and specifically alkaloids) in modern drug discovery cannot be overestimated as their use has been linked closely the history of human use of such resources (Heinrich, 2013).

3.2. Alkaloids in *modern* medicine

One would assume that with a 4,000+ year history of use, often acting as remedies for a variety of illnesses, alkaloids and alkaloid containing taxa would play an important and visible role in modern drug development (Bruhn, 1973). Or in the words of Cordell focusing on local and traditional uses: 'For thousands of years, indigenous groups around the world discovered, through self-experimentation with locally available plant extracts, that they could provide materials for hunting prey, culinary enhancement, amelioration from disease, relief of pain, and healing…in this [last] 200-year period, many alkaloids became critical components of the global pharmaceutical armamentarium, and tremendous healing has resulted from their clinical application' (Cordell *et al.*, 2001). The search using the *Dictionary of Alkaloids* (Buckingham 2010) and other sources identified a total of 51 pure, naturally occurring alkaloids used currently or within the last 50 years for regulated pharmaceutical uses (Appendix 3). This means that to date less than 0.002% (51/27,783) of alkaloids or alkaloid-based drugs are marketed for such uses internationally. It is not surprising that such a diverse set of natural products and their derivatives yield drugs which are used in a variety of applications ranging from cough-suppressants to antimalarial agents. However, in the last 25 years only galanthamine and Taxol™ were newly introduced into biomedicine, and the former in essence through an extension of the therapeutic claims (i.e. from poliomyelitis to Alzheimer's disease, Heinrich and Teoh, 2004). There are only less than 200 others which are commonly used in industrial processes and the manufacturing of commercial goods (for example: N,N'-dioctadecanoylethanediamine is an antifoaming agent used in the polymer industry and methylamine hydrochloride is used in the tanning industry).

Cordell *et al.*(2001) performed highly insightful analyses which highlight the vast potential of alkaloids in drug discovery in the context of two major points; the number of alkaloids scrutinized under ‘biological evaluations’ and ‘some poorly evaluated alkaloid-containing families by genus’.

The first point highlights the fact that although alkaloids represent about 15.6% of all natural products, they constitute 50% of natural products significant to modern pharmaceutics. NAPRALERT also contains *in vitro* and *in vivo* biological test data and when this data is extracted for the alkaloids, 76.3% of all alkaloids have never been tested in a biological assay. Only 0.79% of all alkaloids have been examined biologically in 20+ assays. This is highly significant indicator which further points to the vast untapped potential alkaloids hold.

Number of ‘biological tests’	Number of alkaloids	%
0	16,132	76.38%
1	2,291	10.85%
2-5	1,995	9.45%
6-10	366	1.73%
11-15	119	0.56%
16-20	50	0.24%
20+	167	0.79%
Total[6]	***21,120***	

Table 16: The biological evaluation of alkaloids from higher plants in NAPRALERT (Cordell *et al.,* 2001)

Cordell’s second analysis again highlights the high level of potential associated with the structurally diverse alkaloids. While 21,120 alkaloids can be found in higher plants, there are a very high number of plant genera, all rich in alkaloids, which have yet to be studied. Important families, which contain many genera not yet studied well for alkaloids are listed in Table 17. From this list, it is seen that there are several large families such as the Orchidaceae, Asteraceae, and

[6] Not all alkaloids originate from plants therefore this table only covers about 81% of all alkaloids.

Poaceae which contain at least 500 genera each, which contain dozens if not hundreds of alkaloids, and which have not reached the 10% threshold of being 'studied'. This noteworthy finding has the potential to further highlight discovery efforts to explore 'high potential families and genera.

Family	Genera studied/total genera	% Studied	Number of alkaloids isolated
Orchidaceae	18/1,000	1.8%	53
Theaceae	1/40	2.5%	54
Malpighiaceae	2/60	3.3%	22
Arecaceae	7/200	3.5%	21
Crassulaceae	2/25	4.0%	47
Flacourtiaceae	4/85	4.7%	21
Poaceae	36/500	7.2%	256
Asclepiadaceae	19/250	7.6%	180
Bignoniaceae	8/110	8.0%	53
Malvaceae	6/75	8.0%	48
Proteaceae	6/75	8.0%	61
Verbenaceae	8/100	8.0%	24
Asteraceae	92/1,100	8.4%	705
Acanthaceae	22/250	8.8%	92
Gentianaceae	7/75	9.3%	25
Araliaceae	7/70	10.0%	55
Campanulaceae	7/70	10.0%	53
Icacinaceae	5/50	10.0%	26
Loranthaceae	7/70	10.0%	20
Nyctaginaceae	3/30	10.0%	22
		Total	***1,838***

Table 17: Some poorly evaluated alkaloid-containing families by genus in NAPRALERT (Cordell *et al.,* 2001)

It is important to note that Cordell's dataset is dated to over a decade ago.

Nevertheless, when one considers the decreased attention of industrial players in the natural products arena over the last ten years and surveys published literature, it is apparent that there has not been an unprecedented breakthrough in alkaloid or natural product drug discovery since this data has been published, and thus such data is deemed to be relevant.

With these points in mind, this chapter seeks extend the aforementioned findings by:

1. Performing a comprehensive analysis to identify what alkaloids have made it through the drug discovery process as modern medicines.

2. Investigating physiochemical indicators between these medicinally important ('pharmaceutical') alkaloids as compared to ('non-pharmaceutical') alkaloids

3. Understanding how such indicators relate to prevalent rules of thumb in drug discovery and how alkaloid-specific indicators can be leveraged to enhance natural product drug development at large.

3.3. Methods and datasets used

3.3.1. Dictionary of Natural Products

A maximum of 33 data types, both qualitative and qualitative, were exported for each of the 27,783 alkaloids (Table 18). Modifications were made to the format of some data to ensure consistency.

Alkaloid Name	Hazard and Toxicity	Solubility
Accurate Mass	Melting Point	Source/Synthesis
Biological Source	Molecular Formula	Structure Drawing
Biological Use/Importance	Molecular Weight	Supplier
Boiling Point	Optical Rotation	Synonym

CAS No.	Partition Coefficient	Type of Compound
CRC Registry No.	Percent Composition	Type of Organism
Density	Physical Description	Use/Importance
Development Status	Refractive Index	UV Maxima
Dissociation Constant	Rotation Conditions	
General Statement	RTECS	

Table 18: Data types extracted from the DNP

Additionally, of these 33 data types, some nonessential columns of data such as CAS No., CRC Registration No., and Structure Drawing were deemed relevant to the aims of this thesis and thus not analysed. The most critical and complete data types included: Name, Synonym, Accurate Mass, Biological Source, Melting Point, and Optical Rotation. These six variables served as the foundation to which all of the other data in subsequent analyses was built on (i.e. >95% of all alkaloids contained this data).

3.3.2. Chemical and physical data inputs - EMBL-EBI

The next dataset linked to the DNP was extracted from ChEMBL. ChEMBL is an open-data database containing binding, functional and ADMET information for a large number of druglike bioactive compounds. In 2012, the database contained over 1 million compounds with data that is manually abstracted from the primary published literature on a regular basis, then further curated and standardized to maximize their quality and utility across a wide range of chemical biology and drug discovery research problems (Gaulton *et al.*, 2012). ChEMBL can be accessed freely at https://www.ebi.ac.uk/chembl/. The database was established by the European Molecular Biology Laboratory - European Bioinformatics Institute that describes itself as a non-profit, intergovernmental organization funded by EMBL member states (http://www.ebi.ac.uk/about).

There are four types of search queries in ChEMBL; compounds, targets, assays, and documents. Only queries for compounds were included in this analysis. When querying a compound there are a maximum of 20 quantitative/qualitative data types (Table 19).

ACD BpKa	ACD LogP	ACD LogD	#Rotatable Bonds[7]	Heavy Atoms	Num Alerts	QED Weighted
HBA[8]	HBD[9]	#Ro5 Violations[10]	Aromatic Rings	Passes Rules of Three	Med Chem Friendly	ACD ApKa
Mol reg no.	Compound	Synonyms	Max Phase	Parent Molecular Weight	ALogP	PSA

Table 19: Data Types Extracted from ChEMBL (Ghose *et al.,* 1998; Ertl et al., 2000)

Through its web interface, ChEMBL was manually queried for each of the compounds listed in the initial DNP extract (synonyms from each of the two datasets also included). Due to the wide variance between keywords and formats between the two datasets, automating this process would not yield many 'hits'. Therefore this initial 'bridging' of datasets was performed manually in the form of each query and subsequent data import being performed manually. This initial effort yielded 2,015 'hits' (approximately 7%) and it is estimated that there are <500 potential remaining compounds that exist in both datasets. Similar to the

[7] Number of rotatable bonds in the molecule. Rotatable bonds are defined as single bonds between heavy atoms. Doesn't include ring bonds, those connected to a heavy atom that is attached to only hydrogen atoms or amide bonds (Goujon *et al.*, 2010).

[8] Where a hydrogen bond acceptor is defined as oxygen, nitrogen, sulphur, or phosphorus with one or more lone pairs. The following are excluded: atoms with positive formal charges, amide nitrogens, pyrrole-type nitrogens, aromatic oxygen and aromatic sulphur atoms.

[9] Where a hydrogen bond donor is defined as oxygen, nitrogen, sulphur, or phosphorus with one or more attached hydrogen atoms.

[10] Number of properties defined in Lipinski's Rule of 5 (Ro5) that the compound fails.

DNP, not all data types, for example Mol reg no., Max Phase, and Med Chem Friendly were deemed relevant and analysed.

3.3.3. Dataset combination

Data extracts from the two aforementioned datasets were combined into one Microsoft Excel spread sheet. Table 20 shows a 'complete' set of data for a particular alkaloid:

Data source	Variable	Value
DNP	Name	2-Amino-1-phenyl-1-propanol; (1R,2R)-form
	Synonym(s)	D-threo-form, Nor-ψ-ephedrine, Norpseudoephedrine, Norisoephedrine, Cathine
	Molecular formula	$C_9H_{13}NO$
	Accurate mass	151.099714
	Biological source	Found in "Ma Huang" and *Catha edulis* (Celastraceae) (Khat), used as a stimulant in Arab countries,
	CAS no.	37577-07-4
	CRC registry No.	BCM24
	Melting point	77° C
	Molecular weight	151.208
	Optical rotation	33.14
	Percent composition	C=071.49 H=008.67 N=009.26 O=010.58
	Physical description	Plates (MeOH)
	Rotation conditions	EtOH
	Type of compound	ZQ1400 VX2010
	Type of organism	ZQ1400
ChEMBL	Max phase	0
	Parent mol weight	151.21

	ALogP	0.8
	PSA	46.25
	HBA	2
	HBD	2
	# Ro5 violations	0
	# Rotatable bonds	2
	Passes Ro3	Y
	Med chem friendly	Y
	ACD ApKa	12.07
	ACD BpKa	8.47
	ACD LogP	0.36
	ACD LogD	-1.52
	Aromatic rings	1
	Heavy atoms	11
	Num alerts	0
	QED weighted	0.66

Table 20: Example of a ‘complete entry’ in the combined spread sheet

A disproportionately small number of alkaloids have been developed into marketed pharmaceutical products. To begin to understand how best to prioritize the 27,000+ alkaloids for bioassay screening, HTS or other methods, one has to begin to systematically look at what has made it through the drug pipeline. Those alkaloids which have reached this stage have been listed earlier in this thesis. At the most basic level, an initial analysis (Table 21 and Fig. 6) of 13 basic physiochemical properties of two sets of alkaloids (those used in marketed pharmaceutical products and those which are not) shows averages of variables ranging from -56 to +34% ((Pharma Avg./Total Avg.) – 1). The variable which exhibits the largest difference between the two sets is the distribution coefficient (log D)[11] followed by hydrogen bond donors (HBD), the partition coefficient (log

[11] The distribution coefficient is the ratio of the sum of the concentrations of all forms of the compound (ionized plus un-ionized) in each of the two phases

P)[12], and polar surface area (PSA) respectively. The log D, HBD, log P, and PSA of marketed pharmaceutical products is on average, ranging 31-55% lower than that of other alkaloids. These observations do not completely deviate from those general rules of thumb outlined in the literature review section of this thesis but rather indicate that adjustments to purely computational screening methods must be made to enhance alkaloid based drug discovery.

	Pharmaceuticals/Drugs in ChEMBL (n=47)		**Other Alkaloids in ChEMBL (n=1,968)**		**% Difference in Avg. (Pharma Avg./Total Avg.) – 1)**
Variable	**Range 90% of Values Fall Within (SD)**	**Average**	**Range 90% of Values Fall Within (SD)**	**Average**	
MWT	162.23 - 809.41 (164.78)	375.88	219.23 - 840.70 (178.70)	446.26	-15.7%
ALogP	-0.02 – 4.89 (1.46)	1.46	-0.92 - 7.2 (2.42)	2.97	-14.3%
PSA	32.78 – 153.45 (42.24)	65.91	28.23 – 243.51 (61.11)	96.29	-31.5%
HBA	2 – 12 (2.93)	5.17	2 - 13 (3.40)	5.76	-10.2%
HBD	0 – 3 (1.02)	1.19	0 – 6 (1.90)	2.27	-47.3%
#Rotatable Bonds	0 – 10 (3.33)	4	0 – 16 (5.10)	4.87	-17.9%
ApKa	8.60 – 13.93 (2.44)	11.44	3.41 – 13.71 (3.54)	10.12	13.0%
BpKa	5.90 - 9.98 (1.54)	7.91	1.01 - 10.56 (3.31)	6.41	23.3%
ACDLogP	-0.65 – 5.75 (2.03)	1.84	-1.21 – 7.72 (2.85)	3.11	-40.6%

[12] The partition coefficient is a ratio of concentrations of un-ionized compound between the two solutions.

LogD	-1.09 – 5.2 (2.06)	0.98	-2.52 - 7.33 (2.89)	2.22	-55.7%
Aromatic Rings	1 – 4 (0.99)	1.89	0 – 4 (1.32)	1.61	17.4%
Heavy Atoms	13 – 46 (11.92)	27.60	16 – 59 (12.54)	31.62	-12.6%
QED Weighted	0.25- 0.88 (0.22)	0.66	0.11 – 0.85 (0.23)	0.49	35.2%
#Ro5 Violations	0 – 2 (0.65)	0.28	0 – 2 (0.90)	0.64	-56.0%

Table 21: Comparison of average of physiochemical properties between pharmacologically significant and insignificant alkaloids

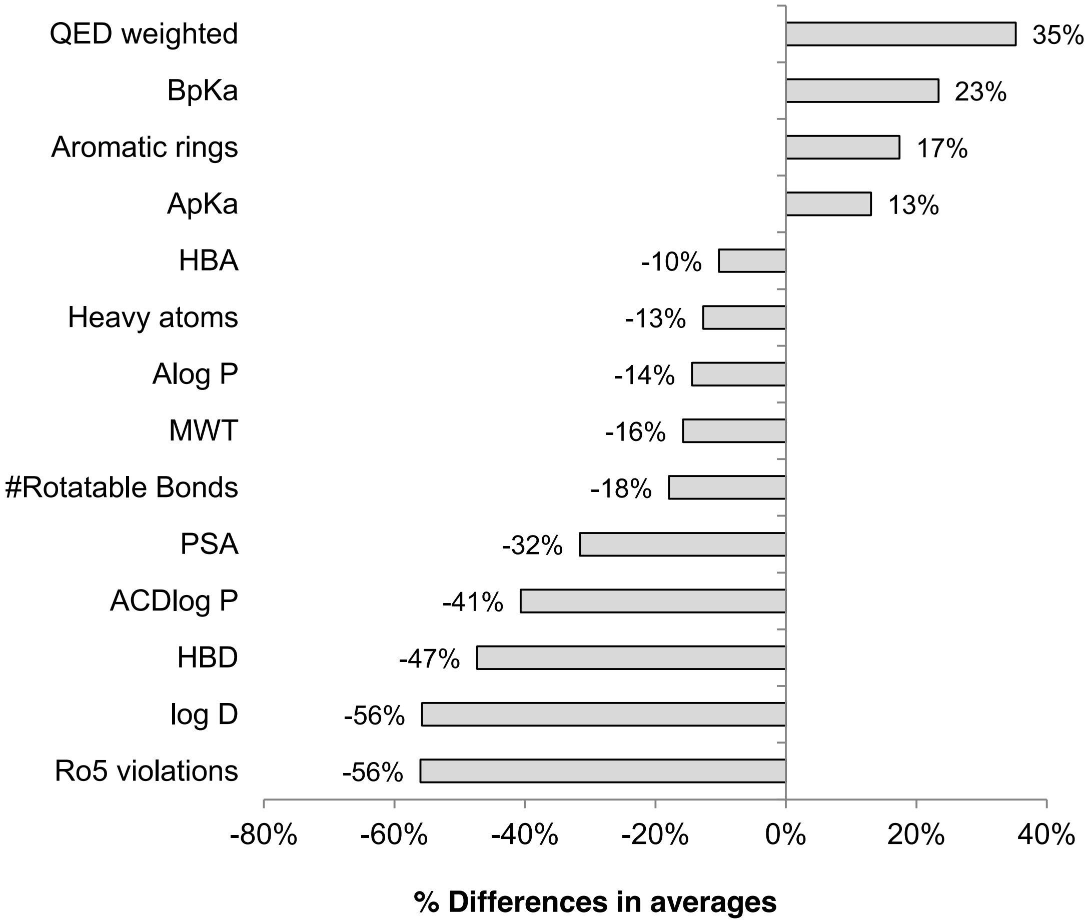


Figure 6: % Difference in averages of physiochemical properties between pharmaceutically significant and insignificant alkaloids

Average log D values for pharmaceutical alkaloids are less than half as compared to other non-pharmaceutical alkaloids and average log P values for pharmaceutical alkaloids are less than 40% as compared to other non-pharmaceutical alkaloids. This suggests that ionization, acidity (log D is decreased as a function of increased pH), and ultimately solubility are potentially the most weighty factors in alkaloid development. These observations are somewhat confirmed by the aforementioned Ro5/Ro3 in that they state that log P values should be <5.0 and <5.6 respectively. The greatest concentration (Fig. 7) of medicinal alkaloid hits lies between a log P value of -1 and 4.

Another variable which shows considerable variation between the two sets of data is PSA. Average PSA values for pharmaceutical alkaloids are less than 35% as compared to other non-pharmaceutical alkaloids. PSA is regarded as a key determinate in intestinal absorption, BBB penetration and several other drug characteristics (Kubinyi and Folkers, 2008). Ertl and Stenberg both independently verified the correlation between PSA and intestinal absorption. Their work suggested that PSA values of <60 Å^2 lead to significantly higher absorption and values >140 Å^2 indicate less than 10% absorption (Stenberg *et al.*, 1999; Ertl *et al.*, 2000). This general rule follows quite closely what one observes in the pharmaceutical alkaloids set (Fig. 8).

Lastly, the average HBD for pharmaceutical alkaloids is 47% less as compared to other non-pharmaceutical alkaloids. The effect of increased hydrogen bond donors on decreasing permeability across lipid bilayers and thus general solubility has been demonstrated by many (Abraham *et al.*, 1994; and Paterson et al., 1994). Thus, it is no surprise that *only* one of the pharmaceutical alkaloids has more than 3 hydrogen bond donors (Fig. 9). The vast majority have less than two HBD with the category average being close to one. These results reaffirm Lipinski's key findings as proposed in his Rule of Five work.

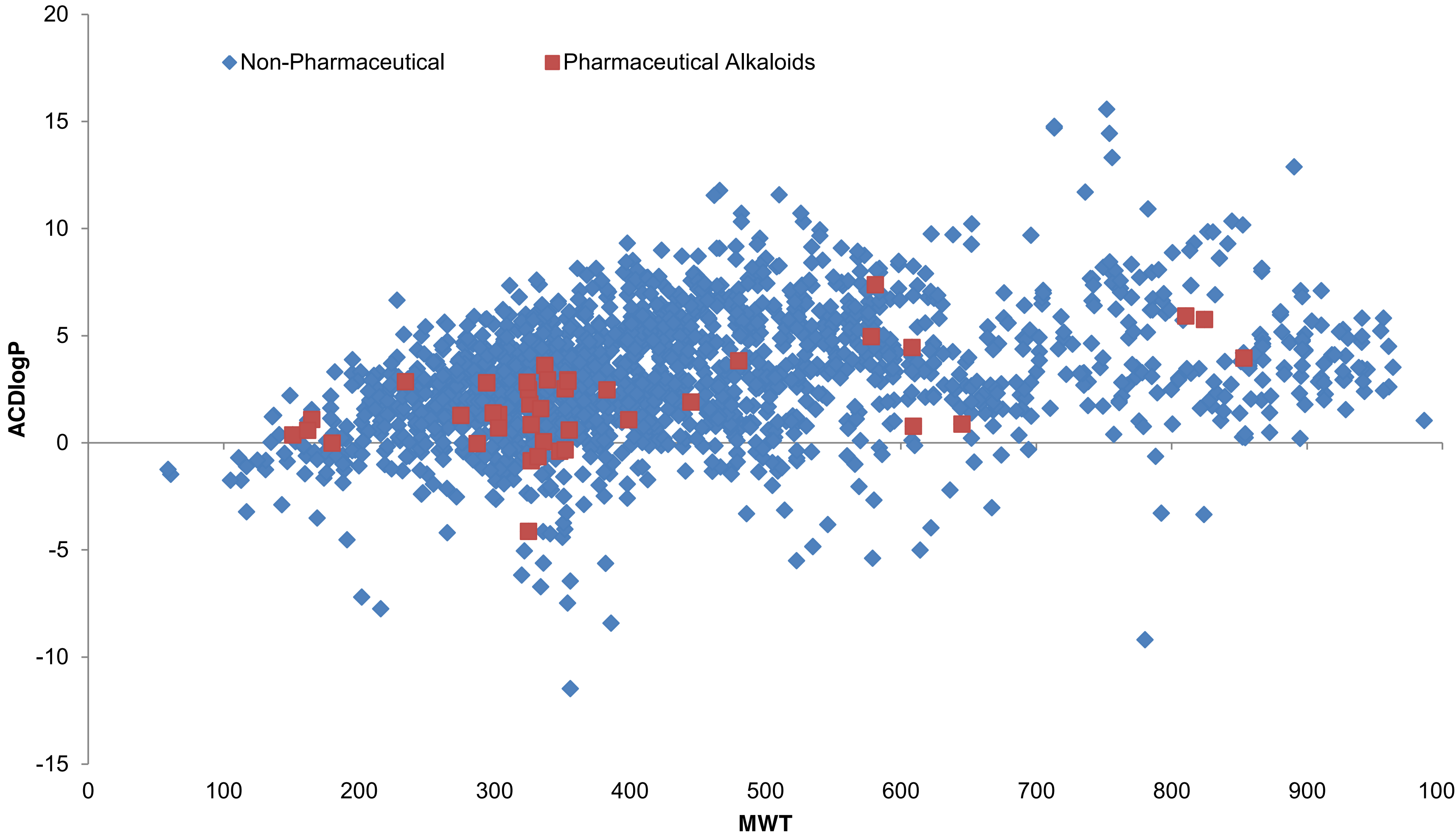


Figure 7: Molecular weight (g/mol) and partition coefficient distribution (ACDlog P) for pharmaceutical/non-pharmaceutical alkaloids (n=2,015)

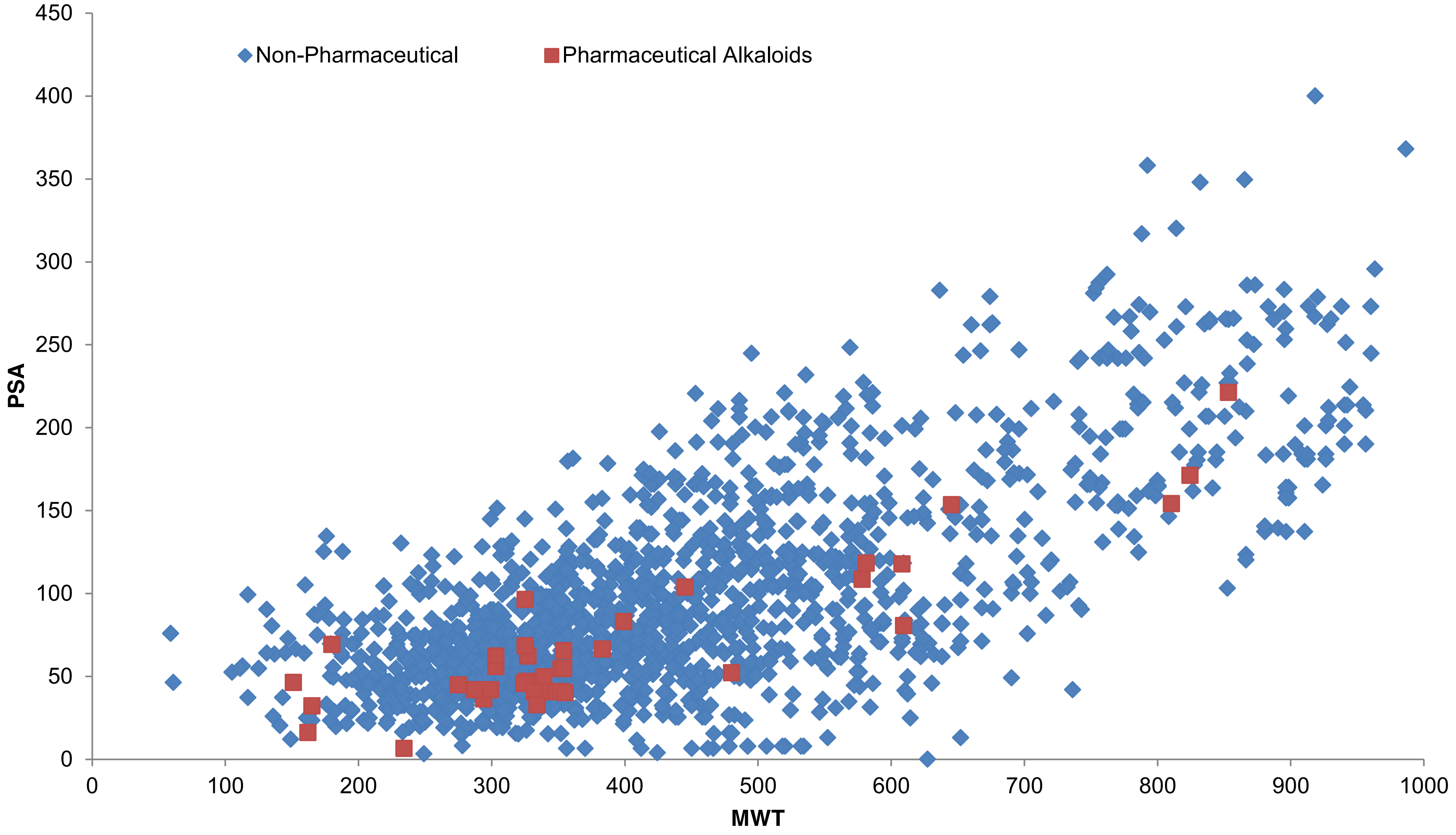


Figure 8: Molecular weight (g/mol) and PSA (Å^2) for pharmaceutical/non-pharmaceutical alkaloids (n=2,015)

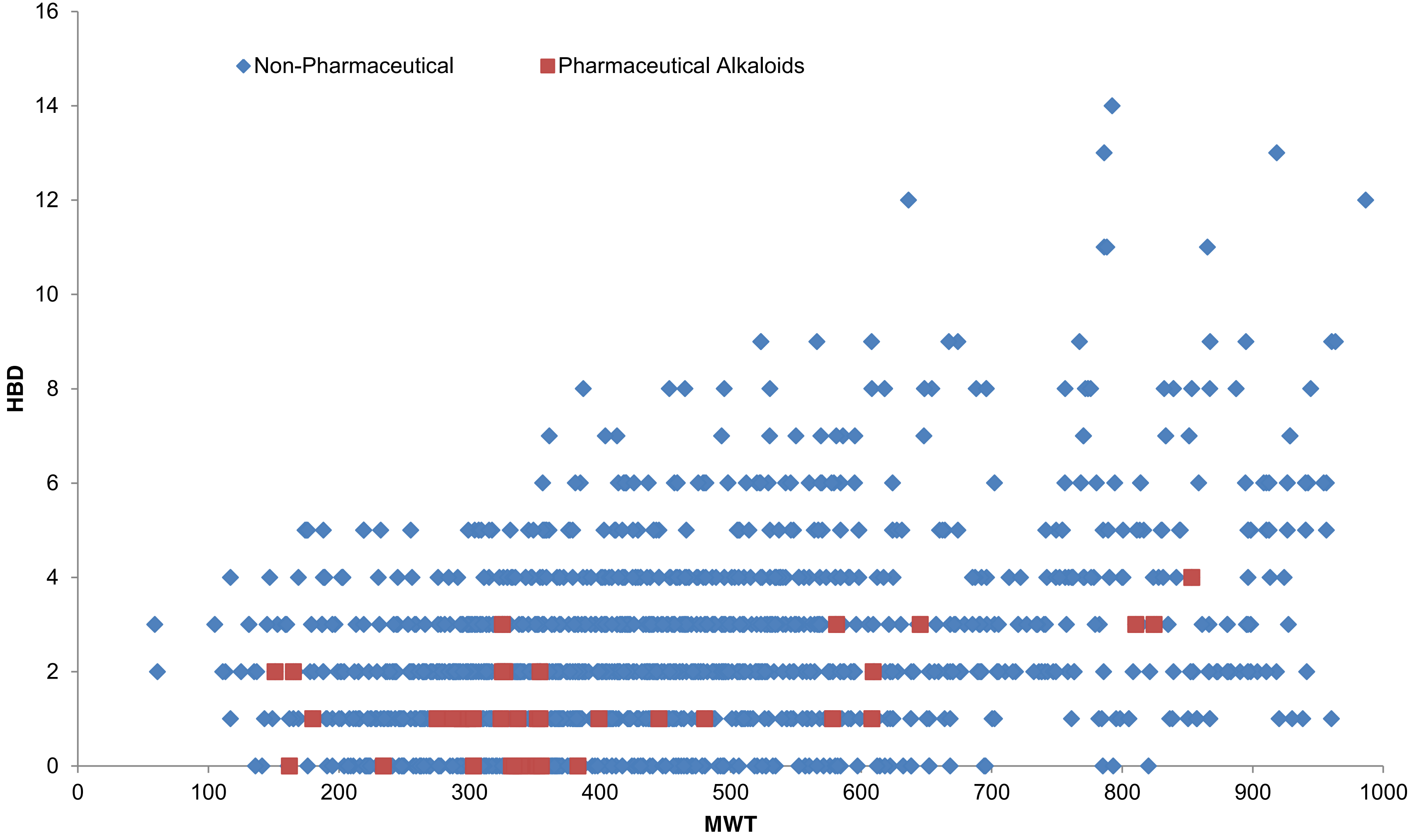


Figure 9: Molecular weight (g/mol) and HBD for pharmaceutical/non-pharmaceutical alkaloids (n=2,015)

In addition to these three chemical descriptors, other chemical descriptors such as MWT, heavy atoms[13], and rotatable bonds are on average less, although not as significantly as log P, log D, PSA and HBD, for the pharmaceutical alkaloids dataset. This confirms and extends many of the observations made by those working in the field of drug discovery. These results are logical and fit into the prevalent debate in modern pharmacology i.e. it is expected that among all alkaloids, those which are developed into drugs are not large molecules with poor aqueous solubility. This leads to another important related question which arises is how these averages fit into other 'non-alkaloid' drugs. How do these results relate to other studies of this nature?

There have been a few studies which have looked at larger drug indices such as an analysis of the World Drug Index (Lutz and Kenakin, 1999). Although at the time of their study in 1999 they cited the World Drug Index as having over 43,000 compounds, their analysis (Fig. 10) of calculated properties shows much fewer compounds. Nevertheless their analysis shows that average log P values hover around 3 and HBD values around 1-2 which is similar to the medicinal alkaloid dataset used in this thesis.

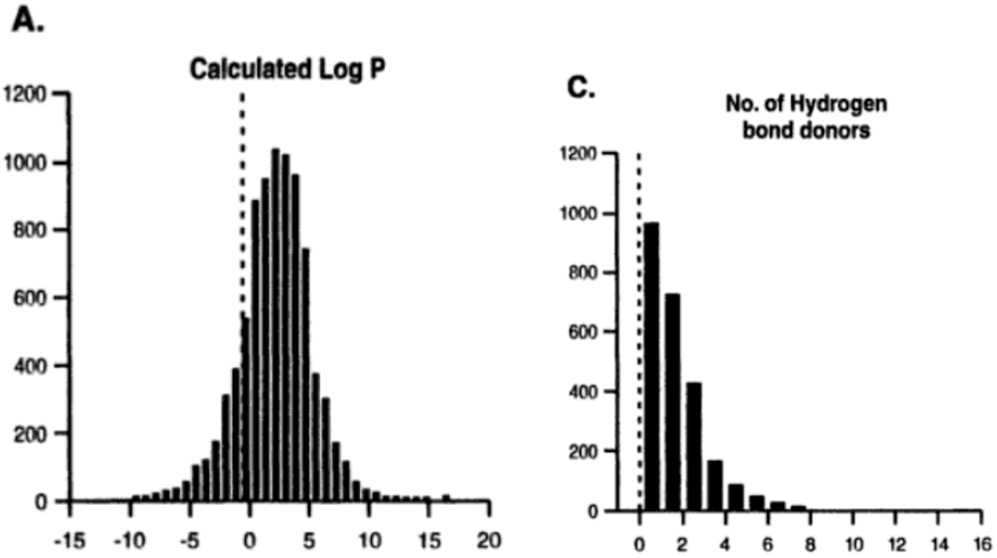


Figure 10: Lutz and Kenakin's histograms of HBD (x-axis), Log P (x-axis), MWT (y-axis) for the World Drug Index (Lutz and Kenakin, 1999).

[13] The number of non-hydrogen atoms in the molecule is the number of 'heavy' atoms (Goujon *et al.*, 2010).

3.4. Alkaloid specific rules of thumb

As mentioned earlier rules such as the Ro3 and Ro5 were not designed with pure natural products in mind.

It is important to note that the key area of inquiry of this thesis is not to merely propose an additional, albeit more specific, rule of thumb to more accurately characterize druglike natural products. The main object of inquiry is to what extent chemical and physical properties play a role in the *overall* development of a natural product, and more specifically, of an alkaloid. Some researchers have set to devise increasingly accurate virtual filters based on such rules and the proposed scheme may contribute to further such efforts, yet this is not a key aim of this thesis.

It can be seen that the pharmaceutical alkaloids have 56% less Ro5 violations when compared with alkaloids at large, thus suggesting that such rules of thumb are relatively effective indicators in alkaloid development processes. In looking strictly at MWT values in the DNP for 27,783 alkaloids its seen that 27% pass the Ro3 while 77% pass the Ro5 (Fig. 11).

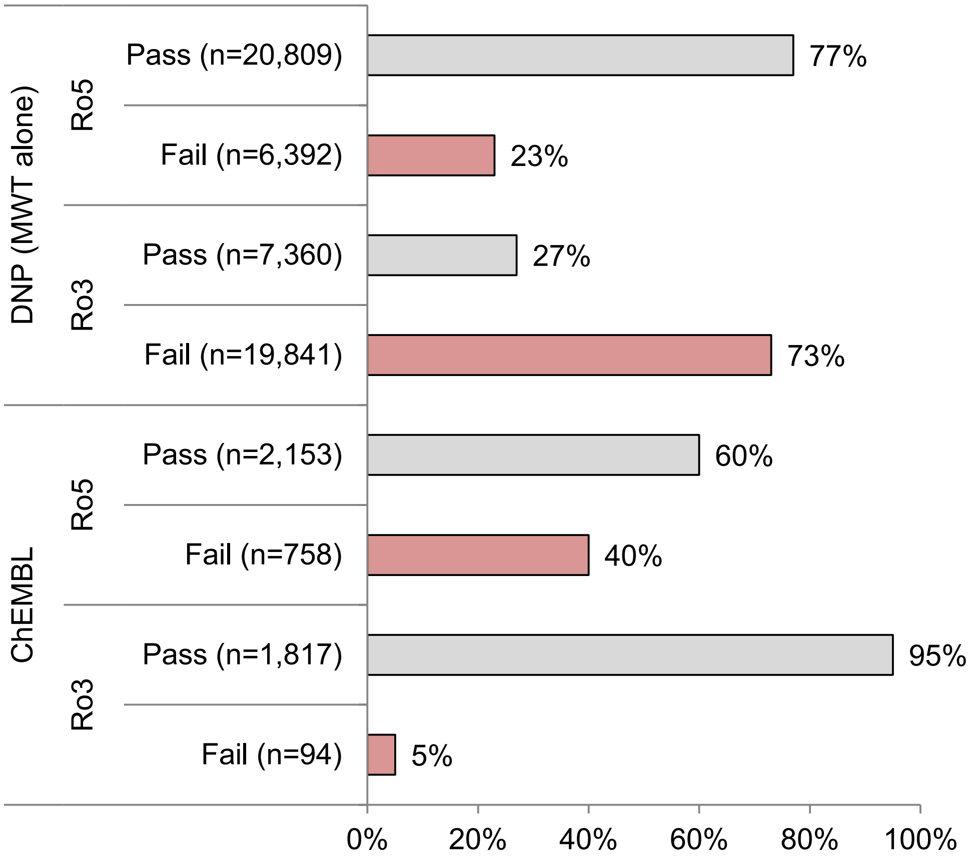


Figure 11: Total alkaloids that pass/fail the Ro3/Ro5 on *MWT alone* (DNP) and pass/fail the Ro3/Ro5 (ChEMBL)

Thus, in working towards the objective in this thesis of investigating the applicability and enhancing the effective of such approaches, a few modifications to such rules of thumb, based on the current dataset, are proposed.

For example, the following proposed scheme correctly filters over 90% of the pharmaceutical alkaloids:

1. MWT/PSA ≥ 3
2. HBD ≤ 4
3. BpKa 6 – 10
4. LogP -1 – 7
5. Ratio MW / Heavy Atoms 13 – 13.2

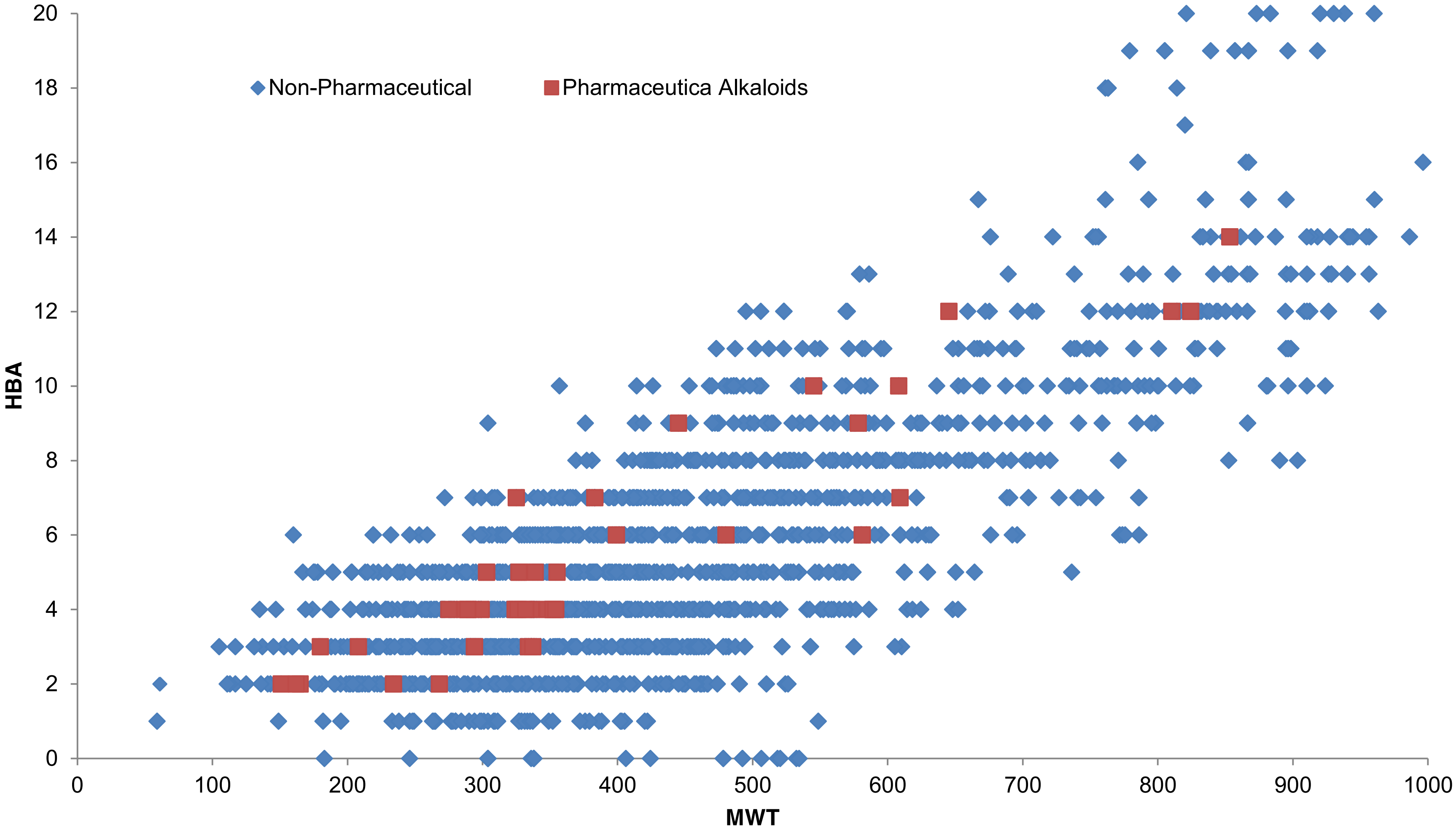


Figure 12: Molecular weight (g/mol) and HBA for pharmaceutical/non-pharmaceutical alkaloids (n=2,015)

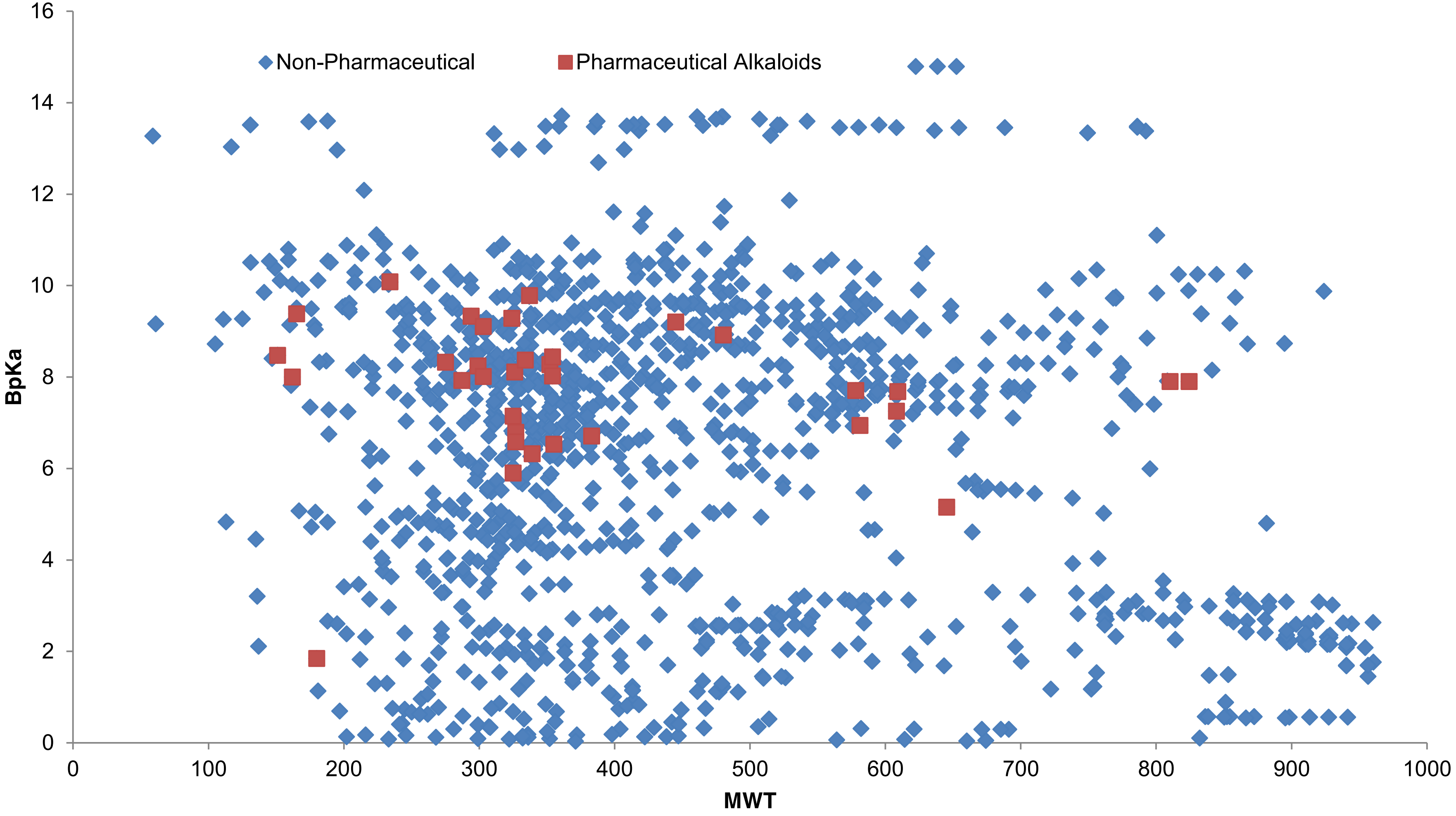


Figure 13: Molecular weight (g/mol) and BpKa for pharmaceutical/non-pharmaceutical alkaloids (n=2,015)

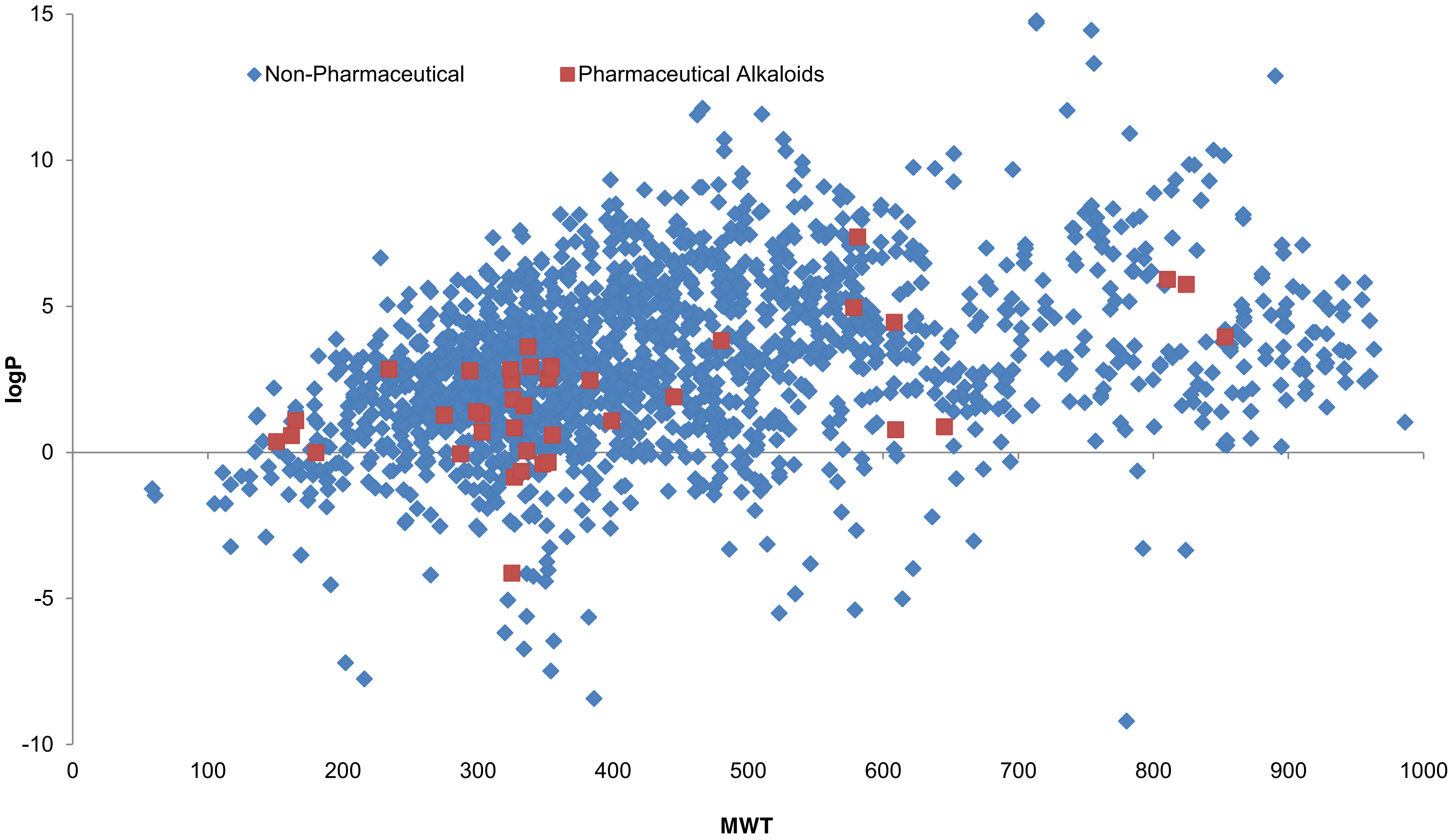


Figure 14: Molecular weight (g/mol) and logP for pharmaceutical/non-pharmaceutical alkaloids (n=2,015)

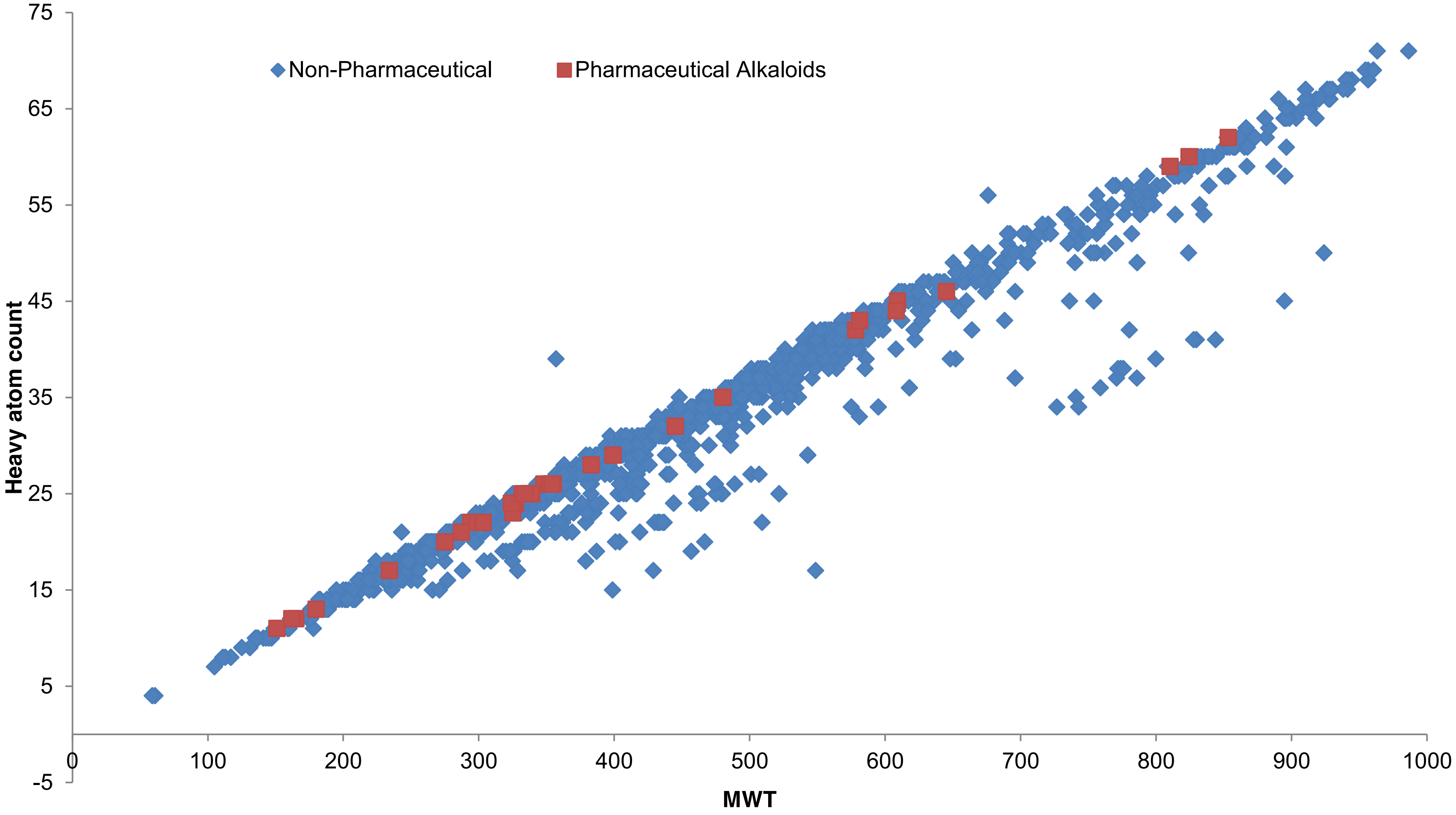


Figure 15: Ratio of molecular weight (g/mol) and heavy atom count for pharmaceutical/non-pharmaceutical alkaloids (n=2,015)

When such rules are used to filter the ChEMBL dataset the total alkaloids go from 2,015 down to 672 (Fig. 16). Rule 3 (pKa 6-10) is the most selective in that it filters out 25% of the total alkaloids. The 672 alkaloids represent exactly one third of the total dataset. If this number is 100% accurate and assuming that there are no supply, commercial, and/or identification issues that leaves over 600 alkaloid candidates which have the chemical profile to serve in some commercial pharmaceutical capacity. Extrapolating this liberal estimate to the larger DNP dataset (which includes a total of 27,000+ alkaloids) suggests, that there may be upwards of 6,000-7,000 alkaloids which carry this 'development potential'.

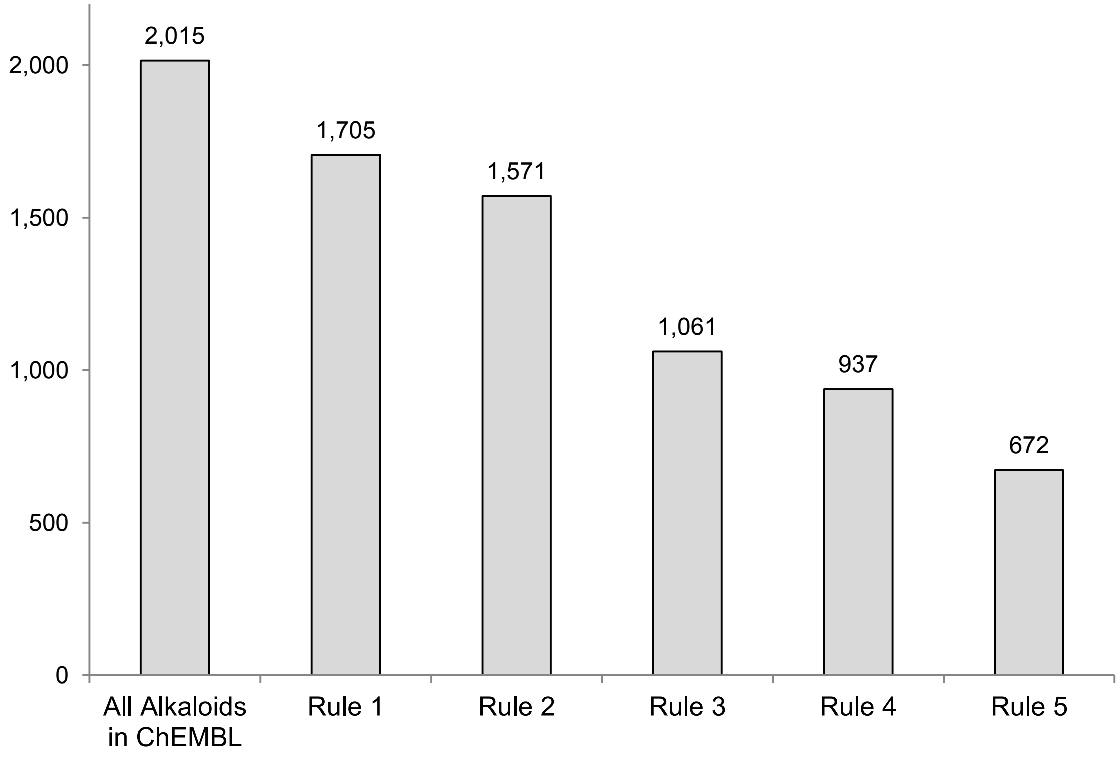


Figure 16: Number of total alkaloids (ChEMBL) and number filtered through each proposed rule

With this in mind, it is important to reiterate that the purpose of this thesis is not merely to put forth one or more rules which extend those rules previously outlined by others. Rather, the central challenge at the center of this thesis is to understand what other metrics are at play in alkaloid development. How are other such metrics weighted? Such an understanding may very well give credence to the possibility that non-chemical characteristics trump chemical properties when considering the *totality* of natural product drug development process.

3.5. Alkaloids which have failed to enter *modern* medicine

An equally, if not more important, question relates to understanding compounds discovered, investigated and eventually rejected for end-use. By being able to examine a large sample of such compounds one would theoretically be able to gain a more comprehensive picture of common stumbling blocks in the development process. For example, to what extent do solubility, toxicity, and/or stability issues influence the stages to which a potential alkaloid-based drug reaches? Is solubility or compound availability more selective in such endeavours? Insights into these questions could highlight and prioritize specific bioassays or bioprospecting strategies which could more effectively screen compounds of interest, which in turn would lead to more compounds entering the R&D pipeline.

These are much more challenging questions for two reasons; such data is extremely fragmented and is typically proprietary in nature. Manufacturing, consumer goods, and pharmaceutical companies, which actively research and develop such compounds, have no incentive to share such data publically and in fact sharing this data may potentially be damaging. Nevertheless one can begin to piece together data leading to insights into this question by looking at other sources such as the *Dictionary of Alkaloids* (Roberts *et al.,* 2010), books such as *Modern Alkaloids* (Berlinck *et al.*, 2007), and other published literature. The identification and tracking of those compounds is more straightforward due to tighter regulations, for example by the FDA, which clearly stipulates why certain compounds are banned or limited in their use. Many of these can be accessed on the United States' FDA's website (http://www.FDA.gov) as well as other open-access regulatory websites. Results for a select number of pharmaceutical products are summarized in Table 22.

Alkaloid name	Stage reached	Reason for failure
Acronycine	Phase I	Moderate potency and poor solubility in aqueous solutions, dose-limiting GI toxicity after oral administration
Camptothecin	Phase II	Caused unpredictable and severe side effects
Curacin A	Preclinical	Solubility issues
Dolastatin 10	Phase I	Discontinued due to hypertension side-effects
Ellipticine	Preclinical	Cardiovascular toxicity and hemolysis
Trabectedin (Ecteinascidin 743)	Phase II/III	Licensed to Ortho Biotech (J&J)

Table 22: Examples of alkaloids which have failed in the drug development pipeline (Fattorusso and Taglialatela-Scafati, 2008; Newman and Cragg, 2004)

This brief examination into 'failed' alkaloids indicates that toxicity is crucial in the drug discovery process (similar to other types of compounds used in drug discovery and development) (Cook *et al.*, 2014; Roberts *et al.*, 2014). Unlike the relatively simple physical and chemical properties identified and analysed earlier in this chapter, toxicity presents a more significant challenge in that its characterization is much more ambiguous. There are accepted consensuses on conventions relating to the measurement of solubility, acidity, or weight of an alkaloid. But these consensuses and standards are much harder to specify with toxicity data. Assays, and the data they produce, vary considerably depending on targets which can be organisms, organs, tissues or cells. This has proved to be an enduring and costly challenge in drug discovery. Some, such as Vedani *et al.* (2012) have modeled toxicity computationally in the context of natural products. These efforts are novel and hold promise, yet remain in preliminary stages and require more effort to determine their impact to modern drug discovery processes.

4. Machine learning and drug discovery processes

4.1. Machine learning efforts in drug discovery efforts – an overview

4.1.1. Virtual screening paradigms in drug discovery

Screening efforts in the drug discovery world are wide and varied in their objectives. A scan of published literature reveals that typically, screens can be traced back to a few overarching general screening paradigms. Previous research by Guiguemde *et al.* (2012) and Bleicher *et al.* (2003) has categorized these as virtual, target based, and phenotypic (Fig. 17). Natural product screens for the most part fall into the virtual filtering and profiling categories. These computational exercises seldom incorporate pure natural products or targets such as specific ligands. In this context, this thesis proposes a *predictive* screening model that incorporates additional non-target based metadata in the form of biodiversity data and hereafter referred to as 'biodiversity based screening'.

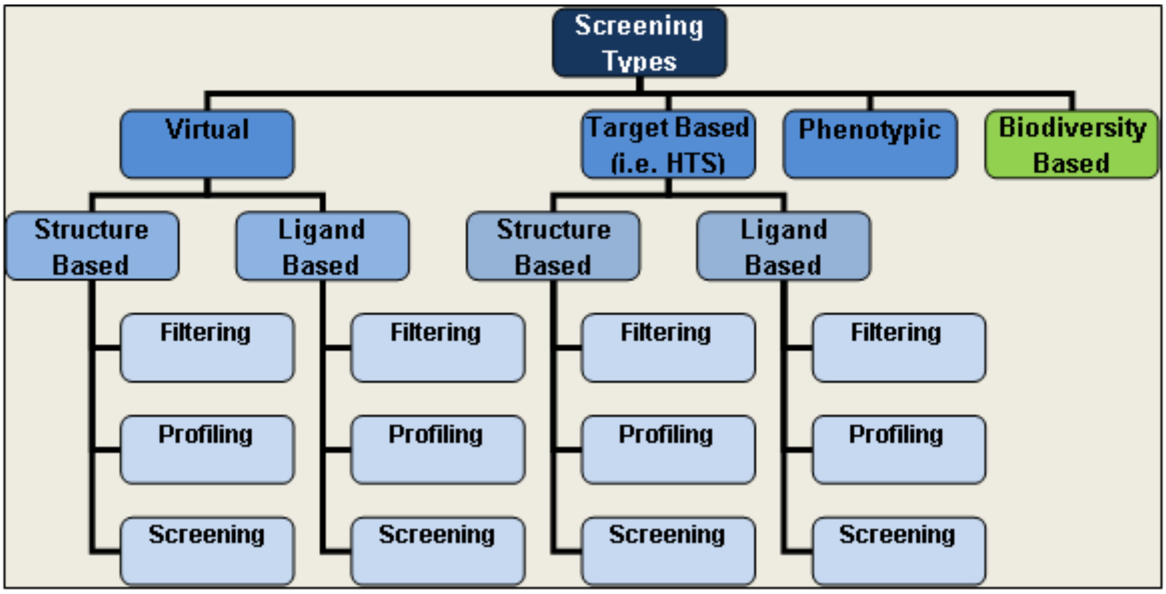


Figure 17: Overview of commonly used screening strategies in modern drug discovery programs with the addition of this thesis' proposed 'biodiversity based screening' (Guiguemde *et al.*, 2012; Bleicher *et al.*, 2003).

Compounds subject to screening programs in modern drug discovery efforts overwhelmingly undergo synthetic modifications (size, lipophilicity, etc.). Most often, these compounds are subsequently filtered by well-known rules such as the Ro5 or Ro3 to quickly screen for fragments, leads, or compounds of 'interest' (Lipinski, 2004). This thesis argues that there are two major limitations to the usefulness of this popular method. The first being that these rules have proven to be accurate in a number of settings, most notably in the published internal data of drug candidates from a few pharmaceutical companies (i.e. Pfizer), but are widely recognized as not being good fits for natural products (Owens, 2003). In fact, very few screens focus solely on natural product sub-classes such as alkaloids or even natural products as a whole (possibly because natural products are an immensely large, diverse, and unwieldy 'superclass' of compounds). The second limitation relates to the debate surrounding the *real-world* applicability of single ligand based screening in the context of drug discovery (Sams-Dodd, 2005; Morphy *et al.*, 2004; Sams-Dodd, 2013). Many have commented on this approach in industry (i.e. solutions to complex medical conditions can be discovered through the isolation of single key biological targets) and believe it to be a major contributor to the stagnant level of innovation within the pharmaceutical landscape. The industry has obviously invested heavily in single target/ligand based screening as its primary discovery strategy. Are there real-world indicators or descriptors, which, when integrated into screening efforts, can enhance efforts in approaching the vast reservoir of yet-to-be researched natural products?

4.1.2. Examples of machine learning efforts in drug discovery

A scan of literature in the areas of machine learning and drug discovery over the years shows a wide range of approaches and methods. Due to their robustness and effectiveness across various applications, two models have received considerable attention: artificial neural network (ANN) and support vector machine (SVM) based models. A typical ANN is represented in figure 7 and shows the process of inputs eventually feeding an output through a hidden network while a typical SVM is shown in figure 8. Artificial neural networks can be represented by supervised or unsupervised learning algorithms while support vector machines only represent supervised learning schemes.

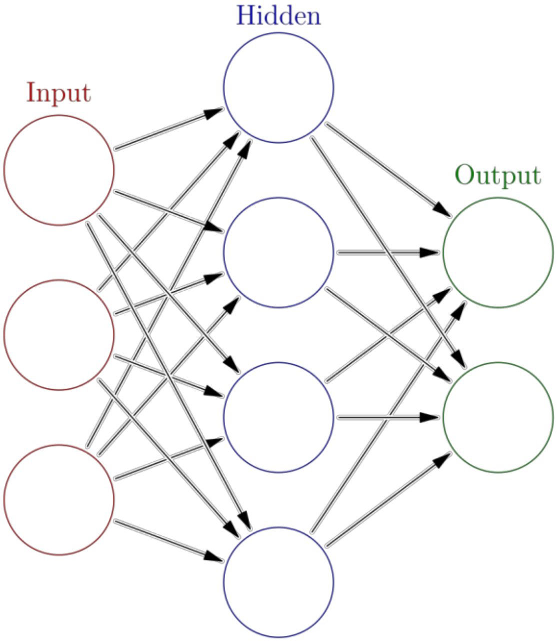


Figure 18: Schematic of a simple artificial neural network showing three inputs feeding into a hidden layer and producing two outputs

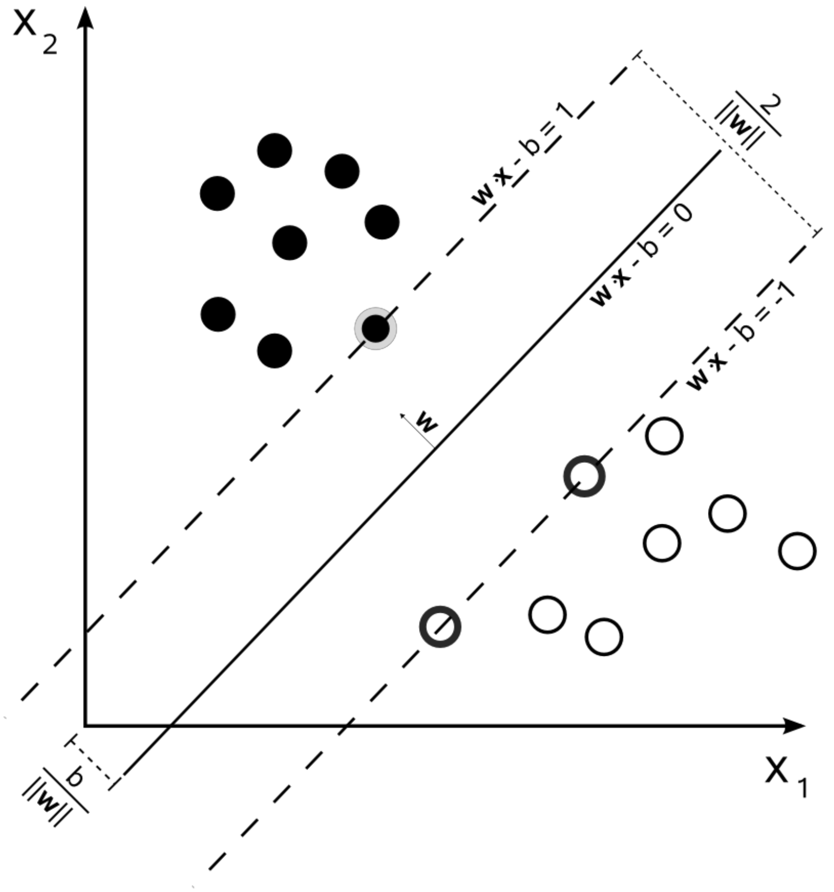


Figure 19: Maximum-margin hyper plane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

The value of these models is that they seek to enhance discovery efforts, typically in earlier stages of the drug discovery process. Objectives of these models can range from 'specific', such as testing a ligand or receptor against a library of 100,000 semi-synthetic fragments and leads, to 'wide', such as filtering structural features of a library of drug/druglike molecules against a set of simple rules. As previously mentioned, there are a subset of publications which extend these exercises to the level of measuring the ability or potential of a compound in acting as a drug. 'Druglikeness' has become the *de facto* term used to encapsulate this concept. One aim of this thesis is to model *real-life applicability* in the natural product drug discovery process rather than simply characterize 'hits' through the computation of physical and chemical data.

In terms of overall accuracy of 'hits', the majority of predictive druglikeness models published over the last 15 years fall between 60-80% (Table 23). This level of accuracy is impressive when one considers how prohibitively expensive drug discovery has become.

Model type	**Metric**	**Overall accuracy/'Hit rate' of model**	**Reference**
ANN	Druglike/non-druglike	61-84%	Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? (Walters and Murcko, 1998)
ANN	Drug/nondrug	72-95%	Prediction of 'drug-likeness' (Walters and Murcko, 2002)
Decision trees, SVM,	SAR models for *Salmonella* mutagenicity	63-79%	Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds (Helma *et al.*, 2004)
ANN, SVM	Drug/nondrug	72-82%	Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification (Byvatov *et al.,* 2003)
ANN, Ro5, SVM	Drug/nondrug	68-75%	Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions (Zernov *et al.*, 2003)

SVM	Active compounds (HIV-1 protease inhibitors, dopamine receptor antagonists, etc.)	2-95%	A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor (Han *et al.*, 2008)

Table 23: Selection of representative drug discovery machine learning efforts published throughout the years

4.2. Incorporating biodiversity as a data input

The quest to model more accurately biological, chemical, docking activity and QSAR through virtual screening efforts has growth with considerable strength over the last two decades. Such approaches have developed in parallel with a large number of both freely available open-source and commercial datasets containing various sets of metadata. A closer look at a selection of freely available datasets (Table 24) reveals no shortage of chemical, biological, physical descriptors covering hundreds of thousands of compounds. This data typically feeds into the screening, profiling and filtering schemes already elucidated on earlier in this chapter. It is important to note that the data contained in these databases is usually a mixture of experimental or empirical and computational or predicted data.

Database name	Description
1-Click Docking	Docking to 9,871 targets or user targets: https://mcule.com/apps/1-click-docking
ACD/Percepta	Prediction of ADME/T and physico-chemical properties: http://www.acdlabs.com/products/percept
ADMET Predictor™	Prediction of ADME/T and physico-chemical properties: http://www.simulations-plus.com
Chembench	Chemoinformatics research support by integrating robust model builders, generators of descriptors, property and activity predictors, virtual libraries of available chemicals with predicted biological and drug-like properties, and special tools for chemical library design: http://chembench.mml.unc.edu
Chemistry Development Kit (CDK)	The CDK is a Java library for structural chemo- and bioinformatics applications. It includes the generation of 260 types of descriptors: http://cdk.sourceforge.net
DIGEP-Pred	Prediction of drug-induced changes in the gene expression profile based on the structural formulae of drug-like compounds: http://www.way2drug.com/GE
Discovery Studio	QSAR modelling and pharmacophore generation, for data analysis and structure optimisation: http:// accelrys.com/products/discovery-studio
GUSAR	QSAR modelling, antitarget interactions and LD50 value prediction based on atom-centric Quantitative Neighbourhoods of Atoms (QNA) and Multilevel Neighbourhoods of Atoms (MNA) descriptors: http:// www.way2drug.com/GUSAR
GUSAR (web-service)	Prediction of acute rodent toxicity (LD50 values), interaction with antitargets and ecotoxicity endpoints: http://www.way2drug.com/GUSAR

INVDOCK	Automatically searches a protein and nucleic acid 3D structure database (this database currently covers 9,000 protein and nucleic acid entries) to identify the protein, RNA or DNA molecule that the small molecule can bind to: http://bidd.nus.edu.sg/group/softwares/invdock.htm
KNIME	Graphical workbench for the entire analysis process, including plug-ins for descriptor generation, creation of QSAR models, and work with SD files: http://www.knime.org
Molecular Operating Environment (MOE)	Calculates over 600 molecular descriptors including topological indices, structural keys, E-state indices, physical properties, topological polar surface area (TPSA) and the Chemical Computing Group's (CCG's) van der Waals surface area (VSA) descriptors. MOE includes tools for the creation of QSAR/QSPR models using probabilistic methods and decision trees, PCR and PLS methods: http://www.chemcomp.com/ software-chem.htm
Molinspiration	Cheminformatics software with tools supporting molecule manipulation and processing, including SMILES and SDfile conversion, normalisation of molecules, generation of tautomers, molecule fragmentation, and calculation of various molecular properties needed in QSAR, molecular modelling and drug design: http://www.molinspiration.com
OpenTox	Interoperable, standards-based framework for the support of predictive toxicology including APIs and services for compounds, datasets, features, algorithms, models, ontologies, tasks, validation, and reporting which may be combined into multiple applications satisfying a variety of different user needs: http://www.opentox.org
OSIRIS	Guides the performance of risk assessment and integrated testing strategies on skin sensitisation, repeated dose toxicity, mutagenicity, carcinogenicity, bioconcentration factors, and aquatic toxicity: http://osiris.simpple.com/OSIRIS-ITS/itstool.do

PASS Online	Prediction of several thousand types of biological activity, including pharmacological effects, mechanisms of action, toxic and adverse effects, interaction with metabolic enzymes and transporters, influence on gene expression based on the structural formula of the chemical: http:// www.way2drug.com/PASSOnline
PreADMET	Calculates more than 2000 2D and 3D descriptors, with prediction of ADME/T and drug-likeness properties: http://preadmet.bmdrc.org
PredictFX™	QSAR modelling and simulation suite that provides prediction of off-target pharmacology, associated side effect profile and affinity profiles on 4,790 targets for drug lead compounds: http:// www.certara.com/products/molmod/predictfx
Prediction of Activity Spectra for Substances (PASS)	PASS is software for the creation of SAR models based on MNA descriptors and modified Bayesian algorithm. It predicts several thousand types of biological activity, including pharmacological effects, mechanisms of action, toxic and adverse effects, interaction with metabolic enzymes and transporters, and influence on gene expression: http://www.way2drug.com
QSARpro®	QSAR modelling including calculation of over 1000 molecular descriptors of various classes: http:// www.vlifesciences.com/products/QSARPro/Product_QSARpro.php
RS-WebPredictor	Prediction of cytochrome P450-mediated sites of metabolism on drug-like molecules: http:// reccr.chem.rpi.edu/Software/RS-WebPredictor
Scigress Explorer, SCIGRESS	Molecular and QSAR modelling including generation of physico-chemical descriptors for small organic molecules, inorganics, polymers, materials systems and whole proteins: http://www.fqs.pl/ chemistry_materials_life_science/products/scigress_explorer

Selnergy™	Combination of docking software to predict interaction energies of a ligand with a protein, database of 7000 protein structures with annotated biological properties and Greenpharma Core Database: http:// www.greenpharma.com/services/selnergy-tm
Small-molecule drug discovery suite	2D/3D QSAR with a large selection of fingerprint options, shape-based screening, with or without atom properties, ligand-based pharmacophore modelling, docking, and R-group analysis: http:// www.schrodinger.com/productsuite/1
SMARTCyp	Prediction of the sites in molecules that are most liable to cytochrome P450-mediated metabolism: http://www.farma.ku.dk/smartcyp
StarDrop™	QSAR modelling, data analysis and structures optimisation, R-group analysis and ADME/T prediction: http://www.optibrium.com
SYBYL®-X Suite6	QSAR modelling, pharmacophore hypothesis generation, molecular alignment, conformational searching, ADME prediction, docking and virtual screening: http://www.tripos.com
Target Fishing Dock (TarFisDock)6	Identification of drug targets from the Potential Drug Target Database with a docking approach: http:// www.dddc.ac.cn/tarfisdock
Toxicity Estimation Software Tool (T.E.S.T.)	Estimation of toxicity values and physical properties of organic chemicals based on the molecular structure of the organic chemical entered by the user: http://www.epa.gov/nrmrl/std/qsar/qsar.html

Table 24: Commercial and freely available software for prediction of biological activity, docking, generation of descriptors and QSAR modelling. Adapted from (Lagunin *et al.*, 2014)

One limitation of these datasets is that they are primarily oriented towards strong coverage of synthetic, semi-synthetic, and other non-naturally occurring natural product derivatives; coverage of pure natural products is highly incomplete. Nevertheless, this source of data is important in that it can and will continue to serve as a valuable resource in narrowing the search for druglike compounds.

With regards to data sources specifically focusing on natural products, there are a significant number – albeit fewer – databases which compile biological (including ethnobiological) and phytochemical information used in virtual screening efforts (Table 25). These databases typically contain lists of plant species relevant to a particular geography (i.e. plant species found in the Amazon) therapeutic area (i.e. medicinal plant species used to treat diabetes). At the time of Lagunin's publication in 2014, the GBIF dataset is was the largest in terms of covering the most plant species at over 1.4 million species.

Source	Description and URL	# of species
A Guide to Medicinal and Aromatic Plants	Information about medicinal, spice and aromatic plants: http:// www.hort.purdue.edu/newcrop/med-aro/ toc.html	510
AGRIS	International information system for agricultural sciences and technology. Bibliographic data: http://agris.fao.org/ agris-search/index.do	Not defined
Ayurvedic Medicinal Plants of Sri Lanka	Medicinal plants used in all of the traditional medicine systems in Sri Lanka and Ayurveda: http://www. ayurvedicmedicinalplantssrilanka.org/	1,635
Botanical Dermatology Database (BoDD)	Description of plants used in the treatment of dermatological diseases, medicinal use and adverse effects: http:// www.botanical-dermatology-database.info/	300
Botanical.com	The electronic version of "A Modern Herbal" by Maud Grieve, published in 1931: http://www.botanical.com/ botanical/mgmh/comindx.html	800
Chemical Abstracts Service (CAS)	The collection and organization of all publicly disclosed chemical substance information including plant components: http://www.cas.org	Not defined
Chinese Herbal Medicine Dictionary	Includes also examples of recipes and dosages of plants: http:// alternativehealing.org/	~900

	chinese_herbs_dictionary.htm	
ClinicalTrials.gov	Database of publicly and privately supported clinical studies of human participants including studies of plant extracts: http://clinicaltrials.gov/	Not defined
Customary Medicinal Knowledgebase (CMKb)	Medicinal plants used by Australian Aborigines: http://biolinfo.org/cmkb	456
Cardiovascular Disease Herbal Database (CVDHD)	Provides docking results between phytocomponents and 2398 target proteins, cardiovascular-related diseases, pathways and clinical biomarkers: http:// pkuxxj.pku.edu.cn/CVDHD/index.php	3,518
Database on Ethno- Medicinal Plants	Medicinal plants and their active components that can be used for the development of new drugs: http:// www.assamphytocure.org/scien.php	80
Dictionary of Natural Products (DNP)	Major commercial source of chemical information on natural products: http:// dnp.chemnetbase.com	Not defined
Dr Duke's Phytochemical and Ethnobotanical Databases	Provides search tools for plant selection and information on ethnobotanical use, phytochemicals and activities: http:// www.ars-grin.gov/duke	1000
ethnoBotany DataBase (eBDB)	International ethnobotany patabase that provides multilingual data on plants from Ecuador, Peru, Kenya and Hawai'i: http://ebdb.org	Not defined
EcoPort	Wiki-like database including ethnobotanical data: http://ecoport.org/ep	88,291
Ethnobotany of the Peruvian Amazon	Medicinal and useful plants in the Amazonian region of Peru ́: http:// www.biopark.org/Plants-Amazon.html	16
EXTRACT database	An expert-based knowledge system on medicinal plants: http://www.plant-medicine.com/index.asp	24
FDA Poisonous Plant Database	References in the literature describing studies of the toxic effects of plants: http://www.accessdata.fda.gov/scripts/	Not defined

	plantox/index.cfm	
FRLHT Indian Medicinal Plants Database	Covers natural resources used in the Indian system of medicine, geo-distribution data, propagation and trade information: http://envis.frlht.org/	6,198
Global Biodiversity Information Facility (GBIF)	GBIF database also includes data on medicinal plants: http://www.gbif.org/	1,454,695
Glob*in*Med	Data on medicinal herbs and plants from different countries including dosage and interactions with drugs and herbs: http:// www.globinmed.com	Not defined
HerbalThink-TCM	Interactive software to learn aspects of Traditional Chinese Medicine: http:// www.rmhiherbal.org/herbalthink/ index.html	430
Herbalist	Description of the principles of the therapeutic use of medicinal plants and data on medicinal plants: http:// www.hoptechno.com/herbmm.htm	161
HerbMed	Categorised, evidence-based resource for herbal information, with hyperlinks to clinical and scientific publications: http://herbmed.org/	242
MedlinePlus: Herbs and Supplements	Dietary supplements and herbal remedies, their effectiveness, dosage, and drug interactions: http:// www.nlm.nih.gov/medlineplus/druginfo/ herb_All.html	80
Herbs & Ayurveda	Ayurveda plants: http:// herbsandayurveda.wordpress.com	20
Indian–Russian Traditional Indian Medicine Database	Plants used in Traditional Indian Medicine, including pharmacological activities of plants and their phytoconstituents (experimental and predicted by PASS software): http:// ayurveda.pharmaexpert.ru/	50
InterBioScreen (IBS) natural products library	Information on natural compounds and their derivatives, with samples available for biological activity screening: http:// www.ibscreen.com/	Not defined
KNApSAcK Core	Metabolites related to plants, medicinal/	1,432

DB	edible plants that are related to geographic zones: http://kanaya.naist.jp/ KNApSAcK_Family/	
MAROWINA FACTS®	Natural remedies, dietary supplements, medicinal plants and herbs of Surinam: http://www.tropilab.com/medsupp.html	43
Myanmar Medicinal Plant Database (MMPD)	MMPD: http://www.tuninst.net/MMPD/ MMPD-indx.htm	100
Medicinal Plants of Bangladesh Database (MPBD)	MPBD: http://www.mpbd.info/	900
NAPRALERT®	Database of natural products, extracts of organisms, case reports, non-clinical and clinical studies: http://napralert.org/	Not defined
Native American Ethnobotany Database	Plants used as drugs, foods, dyes, and more, by native peoples of North America with links to plants database: http:// herb.umd.umich.edu/	4,029
Natural Standard	Systematic reviews of foods, herbs and supplements including drug interactions, dosages and clinical trials: http://www.naturalstandard.com	Not defined
National Center for Complementary and Alternative Medicine (NCCAM), Herbs at a Glance	A series of brief fact sheets that provides basic information about specific herbs or botanicals: http://nccam.nih.gov/health/ herbsataglance.htm	48
PLANTS Database	Standardised information about the vascular plants, mosses, liverworts, hornworts, and lichens of the US: http:// plants.usda.gov/java/	1,049
Plants For A Future (PFAF)	A resource and information center for edible and otherwise useful plants: http://www.pfaf.org/user/default.aspx	7,000
Prelude Medicinal Plants Database[1]	The use of plants in different traditional veterinary and human medicines in Africa: http://www.africamuseum.be/ collections/external/prelude	2,357

PROSEA	Plant Resources of South-East Asia: http://proseanet.org/prosea/eprosea.php	6,697
PROTA	Plant Resources of Tropical Africa: http:// www.prota.org	7,400
International Organization for Plant Information, Provisional Global Plant Checklist	Taxonomic records from 6 major floristic datasets and 7 specialised plant family datasets: http://bgbm3.bgbm.fu-berlin.de/IOPI/GPC/query.asp	201,397
PubChem Substance Database	Samples from a variety of sources including medicinal plants, and links to biological screening results: http:// www.ncbi.nlm.nih.gov/pcsubstance	Not defined
Raintree	Phytochemical information, taxonomic, ethnobotanical and clinical data for plants of the Amazon Rainforest: http:// www.rain-tree.com/	251
Richters Catalog	Description of plants and their parts, which are sold: http://www.richters.com/ Web_store/web_store.cgi	1,062
RxList Supplements	Descriptions of herbs, and dietary supplements, their mode of action and interactions with drugs: http:// www.rxlist.com/supplements/article.htm	Not defined
SuperNatural II Database[1]	A Database of purchasable natural products: http:// bioinformatics.charite.de/main/content/ databases_and_applications.php	Not defined
TCMID	Traditional Chinese Medicine Information Database: http:// tcm.cz3.nus.edu.sg/group/tcm-id/ tcmid.asp	1,098
The Plant List	The accepted Latin names with links to all synonyms by which that species has been known in other databases: http:// www.theplantlist.org/	1,244,871
TIPdb	Database of anti-cancer, anti-platelet, and anti-tuberculosis phytochemicals from indigenous plants in Taiwan: http:// cwtung.kmu.edu.tw/tipdb/	Not defined
TradiMed	Commercial database of plants with symptom(s), efficacy, target organ(s),	502

	property, safety measures: http:// www.tradimed.com/	
TRAMEDIII	South African Traditional Medicines Database: http://www.mrc.ac.za/Tramed3	Not defined
TRAMIL	Traditional Medicines in the Islands (Carribean): http://www.tramil.net/	365
Tropicos®	The nomenclatural, bibliographic, and specimen data collected for the past 25 years: http://www.tropicos.org/ Home.aspx	1,200,000

Table 25: Commercial available data sources used in screens linked to biological activity and QSAR modelling in medicinal plants (Lagunin *et al.*, 2014). Langunin's list excludes commonly used universal databases such as Web of Knowledge (WoK), Medline, and Index Medicus, GBIF represents one of the most comprehensive plant species databases with over 1.4 million species.

In light of the heavy emphasis on filtering, profiling, and screening for druglikeness through the use of chemical and physical properties, critical variables linked to 'success' in the natural product drug discovery processes such as supply have historically been excluded. In this thesis, this gap is addressed through the incorporation of one of the most comprehensively available biodiversity datasets in order to screen for druglikeness.

4.3. Datasets used and methods

4.3.1. GBIF

Principe and others have cited supply constraints as a key obstacle in the development of natural products. For example, Harvey states that natural products are unattractive to many drug discovery companies because of perceived difficulties relating to the complexities of natural product chemistry and to the access and supply of natural products and thus the technical difficulties relating to isolation and structural elucidation of bioactive natural products are being solved by contributions from many different natural product researchers. The prime challenge of quantification of this issue was detailed in the literature review of this thesis. Comprehensive tools in quantifying abundance and/or distribution of plant species on a large scale are essentially non-existent. One effort which shows much promise was put forth by the Global Biodiversity Information Facility (GBIF), which is self-described as operating ‘through a network of nodes, coordinating the biodiversity information facilities of participant countries and organizations, collaborating with each other and the Secretariat to share skills, experiences and technical capacity’.[14] Biodiversity data in GBIF is served through four ‘portals’; occurrences (records that document evidence of a named organism in nature), datasets (smaller datasets endorsed and subsequently published by GBIF through partnering institutions, for example: a dataset from a project by the Taiwan Endemic Species Research Institute enabling Facebook users to upload images of moths, along with dates, locations and species identification), species, and countries/territories. The database can be accessed freely at http://www.gbif.org/.

One approach in gaining insights into the relationship between the prevalence of a host plant/organism and the ‘development status’ of the alkaloids it produces, is to plot out occurrences in a dataset, such as GBIF, against drug/non-drug status of such compounds and observe the presence or lack of any significant trends. The accuracy of this method is, of course, founded upon two points: the assumption that GBIF occurrences are sufficiently representative and correlate to the actual biodiversity of a host/plant species worldwide, and the accuracy of listed host species the natural product (i.e. alkaloid) originates from. Table 25 shows that GBIF is one of the most comprehensive datasets currently available. As of March 2014, GBIF contained 424,254,844 occurrences of organisms in nature including 117,909,945 (27.8%) records from the kingdom Plantae. 1,360,782 species of plants are covered in the database. Occurrences include collected and documented specimens, citations, and records in nature. For example: the DNP reports that the alkaloid monocrotaline was recorded in five species of plants: *Crotalaria retusa* L., *C. spectabilis* Roth, *C.aegyptiaca* Benth., *C. burhia* Benth, and *Lindelofia spectabilis* (Fabaceae, Boraginaceae). Occurrences in GBIF for these five plant species total to 3,222 (2,575, 440, 144, 27, and 36 respectively).

[14] http://www.gbif.org/whatisgbif

A preliminary calculation of this nature has been made for 14% of all the alkaloids listed in the DNP (4,061/27,783) and an initial analysis of the results of this approach are shared in the following section of this thesis.

4.4. Biodiversity as a key criterion of druglikeness

The more challenging question regarding the quantification of alkaloid biodiversity has already been outlined in previous sections. The utility and limitations of the GBIF database have also been outlined in previous sections. It is important to note that data was extracted for 4,062 of the total alkaloid set (dataset is 14.6% complete) and preliminary results are shown in figure 21. Stripes of horizontal data points refer to families of alkaloids all derived from the same natural sources. It can be seen that 80% of all pharmaceutical alkaloids have more than 100 occurrences and less than two alkaloids have less than two occurrences.

When averaging the two data sets, the average of the pharmaceutical alkaloids set is 15,192 (SD=26,164) occurrences while the non-pharmaceutical set averages at 4,925 occurrences (SD=17,880). The standard deviation of the non-pharmaceutical set is significantly higher when calculated as a percent of the category average. This is logical considering the wide variation of abundances of alkaloid producing plants around the globe. These observations support those who argue that supply issues overshadow over research and development of natural products.

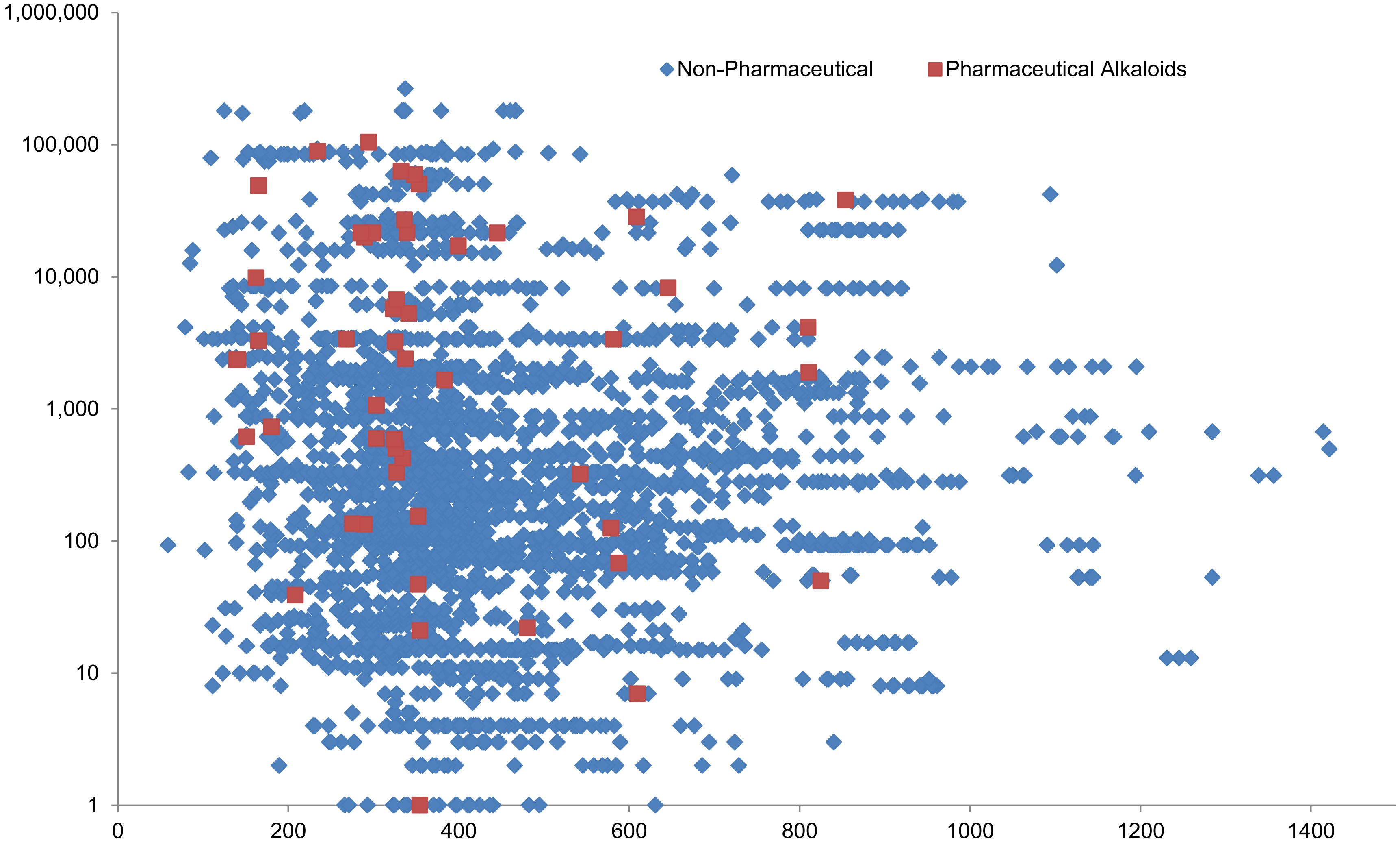


Figure 20: Molecular weight (g/mol) and GBIF occurrences

4.5. Predictive modelling with physical, chemical and biodiversity data

4.5.1. Dataset and methods

Over the years fewer drug discovery screening programs have focused on strictly screening pure natural products. Most libraries are boosted with large additions of semi-synthetic derivatives or even fully synthetic leads which often are related to natural product originators in some way. Thus, knowing that commonly used drug discovery rules of thumb were not designed for, and do not perform strongly with natural products, the following question arises: To what extent can a simple combination of structural data and biodiversity species abundance data predict the likeliness of a natural product to be identified as a drug?

Previous research as part of this thesis (Amirkia and Heinrich, 2014) supports that there is a highly significant discrepancy not only related to physico-chemical properties, but also relating to a species' abundance as exemplified for one class (i.e. the alkaloids) of drug and non-drug natural products. As a next step it is essential to understand how can such insights apply to the continued discovery of pharmaceutically relevant natural products in the form of druglikeness? This research argues that druglikeness in alkaloids can not only be modelled using commonly used modelling schemes, but also predicted accurately by leveraging GBIF host species' abundance data.

A series of machine learning tools was used to develop predictive models. WEKA has been used in previous SAR machine learning studies to model compound ligand activity. All predictive models were executed in WEKA version 3.6.12 (Hall *et al.*, 2009). WEKA allows for a wide range of data inputs as well as predictive models yet, the following algorithms were deemed most useful for this study due to their use across drug discovery efforts as well as high accuracy. In this thesis, the following four algorithms were used in the predictive model (Hall *et al.*, 2009):

Algorithm/Model	WEKA Definition
RandomTree	Class for constructing a tree that considers K randomly chosen attributes at each node. Performs no pruning
RandomForest	Class for constructing a forest of random trees
BayesNet	Bayes Network learning using various search algorithms and quality measures. Base class for a Bayes Network classifier
NaiveBayes	Class for a Naive Bayes classifier using estimator classes
J48 (decision tree)	Class for generating a pruned or unpruned C4.5 decision tree
MultilayerPerceptron (ANN)	A Classifier that uses back propagation to classify instances

Each algorithm was trained with the following four WEKA training settings (Hall *et al.*, 2009):

1. Use training set: The classifier is evaluated on how well it predicts the class of the instances it was trained on

2. Percentage split 30%: randomly splits a dataset according to the given percentage (30%) into a train and a test file

3. Percentage split 50%: randomly splits a dataset according to the given percentage (50%) into a train and a test file

4. Cross validation (10 folds): splits dataset into 10 pieces and performs stratified cross-validation with each of the 10 folds

Alkaloids, as classified in the DNP, and thus input in this thesis' dataset, number 27,783. Following this initial import, extended physical property data was imported from the European Bioinformatics Institute's ChEMBL database and cross referenced against the DNP import. Only 2,015 (7.5%) of the 27,783 alkaloids had a full-set of complete entries for both the DNP and ChEMBL databases. Following these imports, GBIF host species' abundance data (www.gbif.org/occurrence) was manually queried and added as an additional data point for each of the 2,015 alkaloids, thus the final dataset for each alkaloid included the following 17 metrics: Accurate Mass, Max Phase, ALogP, PSA, HBA, HBD, #Ro5 Violations, #Rotatable Bonds, ACD ApKa, ACD BpKa, ACD LogP, ACD LogD, Aromatic

Rings, Heavy Atoms, Num Alerts, QED Weighted, and GBIF occurrences. For the sake of comparison, another dataset was also constructed which contained all 27,783 alkaloids with zeros for missing data. Modelling of this larger dataset proved significantly less accurate, and was discarded for use in the models.

To best ensure that the accuracy of each model could be measured in a meaningful manner, two types of error calculations were chosen, both which are presented in the results below. The first indicator of accuracy is the raw percentage correct pharmaceutical or non-pharmaceutical predictions of the model. Only 2% of the alkaloids in the dataset are labelled as pharmaceutical alkaloids so it is important to look at both true positives and negatives. Looking at true negatives (i.e. non-pharmaceutical alkaloids) alone can falsely portray accuracy (which is in fact the practice among many machine learning druglikeness related publications). This is a quicker approach, analogous to the 'hit rate' of the virtual screen. The second indicator is a more sophisticated statistical analysis of the data which essentially calculates the summation in error for predicted/actual values as a proportion of the summation of mean/actual values. This indicator is referred to as relative absolute error (RAE) and indicates accuracy not only for the individual predicted outcomes but also the algorithm as a whole. RAE values >100% indicate that the model is performing worse than just predicting the mean of the dataset.

$$RAE = \frac{\sum_{i=1}^{N} |\hat{\theta}_i - \theta_i|}{\sum_{i=1}^{N} |\bar{\theta} - \theta_i|}$$

4.5.2. Discussion of algorithm outputs

Of the six algorithms RandomTree yielded the highest percentage of correctly predicted pharmaceutical alkaloids (100%) and lowest overall RAE (0%) (Fig. 21). RandomForest also produced a model with nearly 100% correct pharmaceutical alkaloid predictions with less than 25% RAE. The next accurate model is the

MultilayerPerceptron (ANN), which can predict pharmaceutical alkaloids up to an accuracy of 57%. Dissimilar to the ‘yes/no’ decision-making tree models, the nodes and weights this model uses allow for increased flexibility and applicability to other potential relevant datasets.

The level of accuracy is significantly increased when accuracy of pharmaceutical alkaloids is substituted by *overall* accuracy of the model (Fig. 22). Due the disproportionate number non-pharmaceutical alkaloids in the dataset (98%), by including them in the accuracy calculation of each model, accuracy of all models jumps to >96%. This indicates that the models do much better (almost 2-times better) at predicting non-pharmaceutical alkaloids than they do pharmaceutical alkaloids. It is argued that with a larger more complete set of data the accuracy of all the models would reach >99%. As the dataset stands now, the high level of accuracy achieved comes from a model which merely represents inputs from only 7.5% of all alkaloids.

A closer look at the decision tree models, RandomTree and RandomForest, reveals that their construction is more of a ‘one-time use’ model which in essence, ‘over-fits’ the provided dataset. Publications over the years are rife with over fitted, ‘one-time use’ models which are intolerant to other data sources. Applicability to other datasets is typically limited and previously published decision trees successfully have acted as a ‘proof-of-concepts’, but have not significantly been applied in real-world situations as documented in the published literature (Ehrman *et al.*, 2007; Burbridge *et al.*, 2001; Weston *et al.*, 2002). This research argues that the use of biodiversity and host species’ abundance data in natural product screens can help move purely computational machine learning efforts towards real life applicability. Figure 24 shows a simplified schematic of the random tree model for the alkaloid dataset. In order to maximize the number of correct predictions, the model generates a tree with dozens of exceptions which cannot be used to accurately predict druglikeness in other datasets.

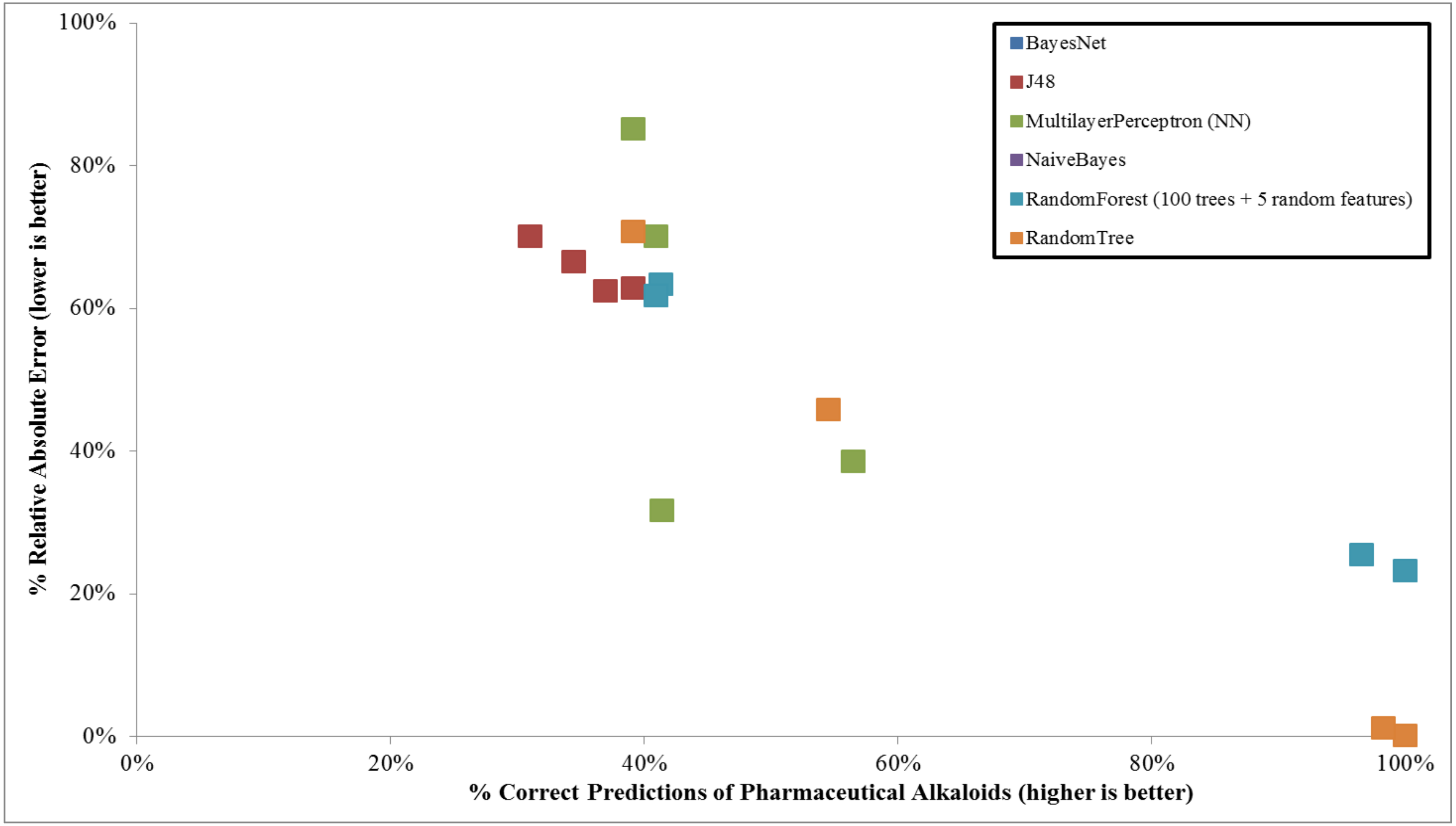


Figure 21: Summary of correct predictions of *pharmaceutical alkaloids* vs. relative absolute error (RAE) for each algorithm and training method (each dot of the same colour represents one training scheme within one algorithm)

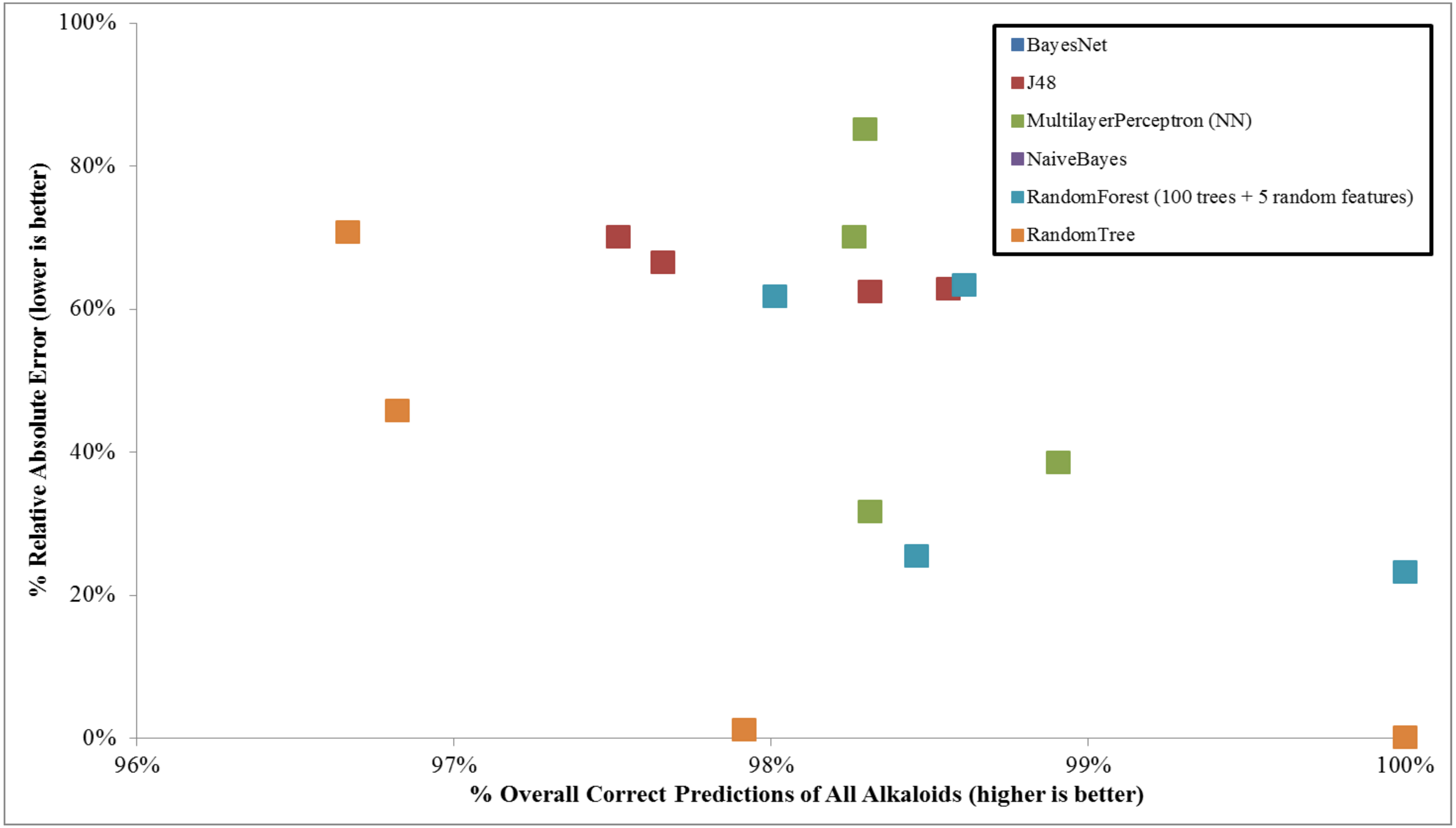


Figure 22: Summary of correct *overall* predictions of all alkaloids vs. RAE for each algorithm and training method (each dot of the same colour represents one training scheme within one algorithm)

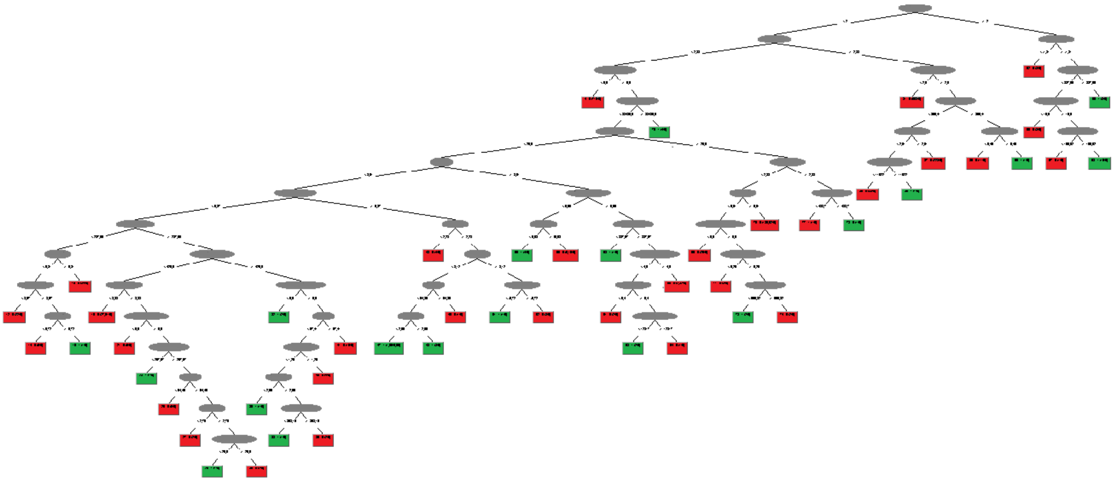


Figure 23: Simplified schematic of the random tree model with pharmaceutical alkaloids represented in green and non-pharmaceutical alkaloids in red.

The ANN output in this thesis is in line with previous target or ligand machine learning studies but is the first *non-target*, species abundance-based modelling strategy which demonstrates the potential of non-traditional alternative data sources in virtual screening and drug discovery processes (Warmuth *et al.*, 2003). Although this effort solely focuses on the alkaloid class of secondary metabolites, it is believed to be applicable to other classes of natural products such as terpenoids. Undergoing similar analyses with other natural product classes or sub-classes may very well demonstrate that other classes of natural products as inputs can predict even higher degrees of accuracy in predicting drugs or druglike compounds. The determining factor in accurately modelling natural products is the high level of completeness and accuracy of the ‘drugs’ (or druglike compounds) which the models are initially trained with (Table 25):

	Pharmaceutical Alkaloids	**Non-Pharmaceutical Alkaloids**
Correctly Predicted (n=1,992)	1. Ajmalicine 2. Berberine 3. Chelerythrine 4. Chondocurine 5. Cinchonidine 6. Cinchonine 7. Codeine 8. Colchicine 9. Deserpidine 10. Emetine 11. Ephedrine 12. Ergotamine 13. Galanthamine 14. Harringtonine 15. Narceine 16. Nicotine 17. Pilocarpine 18. Quinine 19. Reserpine 20. Sanguinarine 21. Scopolamine 22. Sparteine 23. Taxol 24. Theophylline 25. Tropine tropate 26. Vincristin	1. Acivicin 2. Agelongine 3. Arginine 4. Betaine 5. Calcimycin 6. Canavanine 7. Chromophenazine B 8. Cordifoline 9. Cordifoline 10. Discorhabdin K 11. Domoic acid 12. Dysidine 13. Dysiherbaine 14. Dysinosin C 15. Fasciospongine A 16. Flazine 17. Lonijaposide C 18. Melodinine C 19. Montipyridine 20. Platencin 21. Platensimycin A5 22. Pulchellamine F 23. Pyranonigrin B **+1,943 others**
Incorrectly Predicted (n=22)	1. Aconitine 2. Ajmaline 3. Argemonine 4. Boldine 5. Cathine 6. Cocaine 7. Hydrastine 8. Lobeline 9. Lysergic acid 10. Monocrotaline 11. Palmatine 12. Papaverine 13. Physostigmine 14. Quinidine 15. Strychnine 16. Tetrahydropalmatine 17. Tropine tropate 18. Vinblastine 19. Vincamine 20. Yohimbine	1. Lupanine 2. Quinicine

Table 25: Correct and incorrect predictions of pharmaceutical and non-pharmaceutical alkaloids using the artificial neural network model

Although the results show an impressive level of *overall* accuracy, it is important to look closely at the 20 incorrectly predicted pharmaceutical alkaloids with hopes of better understanding the why the proportion of incorrectly pharmaceutical alkaloids (false negatives) is significantly higher than incorrectly predicted non-pharmaceutical alkaloids (also false negatives). One meaningful way to understand which metrics the neural network is weighting most heavily (and thus using to judge between a drug/non-drug), is to average values of inputs across the two groups of false negatives and to see if there are any significant differences. When this calculation is made, it can be seen (Fig. 24) that correctly predicted pharmaceutical alkaloids on average have 8.9 times more GBIF occurrences than incorrectly predicted pharmaceuticals alkaloids. This not only reaffirms findings in previous research that species abundance is highly correlated to druglikeness of alkaloids, but additionally indicates that species abundance (that as a data source within this model) is a statistically significant factor in machine learning efforts in predicting of pharmaceutical alkaloids.

A closer look at the pharmaceutical alkaloids reveals that there are a few pairs or closely related pharmaceuticals which have been split by the model into pharmaceutical and non-pharmaceutical predictions. In the case of codeine and papaverine, both are derived from *Papaver somniferum* L. and therefore are reported as being equally abundant, Yet, codeine is a slightly smaller molecule with a lower partition coefficient while papaverine has a five times larger distribution coefficient. Another pair is vincristine and vinblastine first isolated from *Catharanthus roseus* (L.) G.Don, which are even more alike as compared with papaverine and codeine. This pair presents an exception to the model because the DNP lists vinblastine as originating from three species while there are only two species for the correctly predicted vincristine. This result indicates that the nature of this modelling work is limited in its specificity of predictions with regards to closely linked compounds. This exception can also possibly be traced back to the quality of the DNP species lists for each alkaloid. Not all host species are listed and/or specified. Another possible explanation for this result is that the model is trained to output a binary yes/no result. It is highly likely that outputs for two similar compounds are just on either side of the >1 (druglike) and <1 (non-druglike) threshold. Placing predictions on a spectrum of druglike and non-drug-like would help to alleviate this limitation. Lastly, it must be noted that the pharmaceutical set of alkaloids is miniscule with a total of 46 compounds. Given a full set of physical and chemical properties, there are at least 11 additional pharmaceutical alkaloids which could better train the model; morphine is one notable omission. Additional omissions are: pseudoephedrine, hemsleyadine, granatonine, belladonnine, dregamine, eschscholtzine, lauroscholtzine, pelletierine, protopine, protoverine, and synephrine.

Examining the incorrectly predicted non-pharmaceutical alkaloids suggests that there are commonalities among them, the first being that both lupanine and quinicine are highly toxic and poisonous alkaloids. This result seems to be in line with what is expected from a dataset that does not contain any metrics directly representing or correlating to compound toxicity. Secondly, host species (as listed in the DNP) for both compounds are diverse. Lupanine is listed as originating from *Lupinus albus* L., *Lupinus termis* Forssk. [a taxon also recorded as a synonym of *L. albus* (www.theplantlist.org)], *Podalyria buxifolia* Willd, and *Virgilia capensis* (L.) Lam. [also considered to be a subspecies of *Virgilia oroboides* (P.J.Bergius) T.M.Salter ssp. *Oroboides* (www.theplantlist.org)] while quinicine originates from ‘many *Cinchona* species’. Since in the model species abundance of host species is a key criterion this wider abundance in combination with strong pharmacological (albeit toxic) effects is involved in the incorrect prediction of these two being pharmaceutical alkaloid.

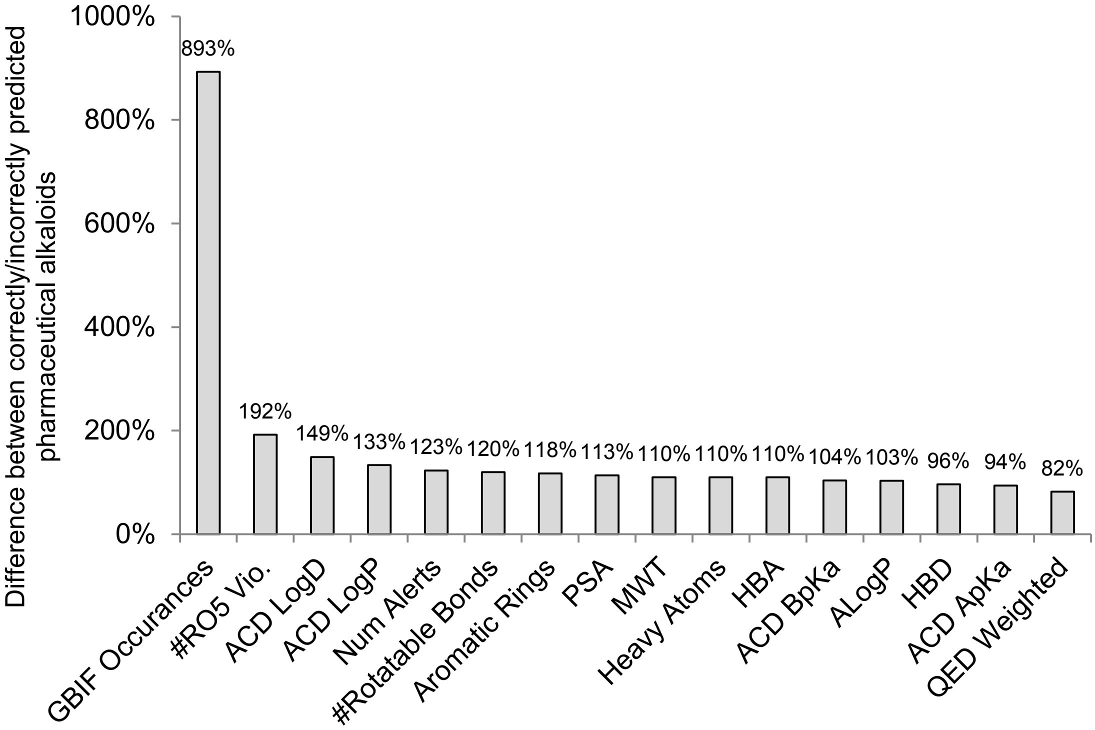


Figure 24: The difference in average of each input between correctly and incorrectly predicted pharmaceutical alkaloids

GBIF species abundance data has continuously and plans to continually be updated with additional novel species as well as occurrence data for pre-existing species. These additions will only increase the quality and robustness of the data, which in turn can help drive more complex modelling efforts. One possible source of bias is that the GBIF dataset is generated from species which are most commonly researched and published. A bias of this kind could possibly self-fulfil one’s hypothesis that species abundance is highly correlated to druglikeness. Looking more closely at the generation of the GBIF dataset shows that as of January 2016, 983 institutions have submitted 12,760 databases containing species occurrence data. In 2007, Yesson studied the comprehensiveness of the GBIF dataset in the context of Legume species abundance and found that 84% of occurrences passed their own internal ‘geographical validation’ and 3.6% of listed Legume species could not be validated according to any of their listed criteria (Yesson *et al.*, 2007). At the time of the Yesson study, only 199 institutions were providing occurrence data to GBIF, now there are close to 1000. Certainly, this sharp rise in data volume (169 million plant occurrences alone) paired with the frequency of newly generated data is sufficiently ‘random’ enough to avoid any significant bias.

It is widely accepted that there is a great challenge in the drug discovery world today in both in terms of approach (ex. HTS and single target focus screening paradigm, incremental structural modification paradigm) and climate (ex. astronomical costs, regulatory pressures). Alkaloids remain vital to drug discovery efforts and best estimates today cite that humans have only systematically explored about 10% of all natural product host species. How can the research community potentially refine its approach to maximize this historically proven wealth of potential? How can current cutting-edge approaches be paired with bygone experiences in the commercialization of natural products (ex. Shaman Pharmaceuticals) be utilized to advance the collective ability to advance drug discovery efforts worldwide (King and Carlson, 1995; Clapp and Cook, 2002)?

Machine learning efforts have without a doubt enhanced the collective ability to successfully discover myriads of new and highly diverse fragments and leads and can reduce the gap between technological advances and innovation in drug discovery. Fragments and leads are disproportionally representative of compounds derived from natural products yet researchers rarely use pure natural product compounds in machine learning efforts. This research argues that natural products (i.e. alkaloids) can be screened in a highly impactful way if real-world indicators, such as species abundance, are incorporated in predictive models. Additionally, it is demonstrated that novel machine learning efforts focusing specifically on alkaloids have the potential to accurately predict a high number of marketed drugs and even a higher number of correctly non-pharmaceutical natural products.

5. Looking ahead to the future of natural products

5.1. Application to modern drug discovery and screening paradigms

The perceived failure of current drug discovery paradigms to adequately tap into the vast reservoir latent in natural products calls for more efficient and impactful approaches to be devised. Both previous published research as well as insights from industry stakeholders reaffirm this latent potential. Supply is undoubtedly a key challenge in natural products research and thus, any bioprospecting, machine learning, or other strategy cannot overlook it.

The thesis examined the problems of supply in the context of host species' abundance data of pharmaceutical alkaloids and it was shown that source species of pharmaceutical alkaloids are on average 4.3 times more 'abundant' (GBIF) species abundance dataset) than a randomly picked non-pharmaceutical alkaloid (Amirkia and Heinrich, 2014). Alkaloid containing species yielding medicines are thus much more widely distributed than species which yield alkaloids not used pharmaceutically. This suggests that such a dataset is sufficiently significant for modelling supply constraints which are so often cited in natural product related literature.

This unique and 'real-world' data point can be further leveraged by the myriad machine learning schemes prevalent in the modern drug paradigm. Machine learning efforts have been integral to drug discovery programs yet their contribution to drug discovery strategies is limited. The preceding insights demonstrate how target-independent machine learning efforts can more effectively capture the potential of natural products and help prioritize efforts. The demonstrated predictive model shows overall accuracy comparable to previously published target-based machine learning efforts of synthetic and semi-synthetic centric studies. The model incorporates species abundance data and can consistently predict >50% of all pharmaceutical alkaloids and 98% of all alkaloids, which are regarded as a highly important class of natural products in historical as well as modern drug discovery.

5.2. Future work

Methods described in the preceding chapters provide novel insight into how traditionally used data inputs and empirical rules seeking to model druglikeness can be augmented by non-traditional data sources, such as host species biodiversity data, to accelerate innovation and efficiency in modern drug discovery.

However it must be realized that the preceding insights, although novel, are based on a few simple metrics which have potential to be deepened. With respect to the GBIF database and as mentioned previously, its data volume and accuracy is making tremendous strides forward. Yet, the extracted dataset and accompanying modelling work was merely based on one metric within the GBIF database; GBIF occurrences. While this 'proof of concept' is certainly exciting, these insights could certainly be extended and applied to more specific sets of circumstances.

The following questions for example would extend these findings into further specific geographic, cultural, and industrial applications.

1. Are host species widely spread across one region or densely found in smaller areas?

2. How many countries does the species naturally grow in?

3. How are occurrences of the species in the dataset distributed across time? Were instances discovered and recorded decades ago or have records been relatively consistent?

4. Are there potential indicators between species abundance and species biomass?

5. Are such occurrences linked to the weediness and as such the ease of access of the source species (Stepp and Moerman, 2001; Stepp, 2004)?

A systematic assessment of a species' abundance can play a constructive pre-screening or filtration role in natural product drug discovery programs which addresses one of the key concerns of the stakeholders who contributed to this thesis. Costs for such analyses are minimal compared to R&D budgets common to pharmaceutical companies today. Additionally such analyses need not necessarily be exclusively seen as applicable to screening programs for candidates or leads but may prove to be of value in other natural product related endeavours. Companies which are heavily invested or interested in TCM, Ayurveda, and other traditional medicine centric portfolios may use this approach to optimize procurement or investment processes. Compounds which originate from host species which are becoming increasingly abundant may hold more promise long-term sustainability in production and marketability.

References

Abad-Zapatero, C. (2007). A Sorcerer's apprentice and the rule of five: from rule-of-thumb to commandment and beyond. *Drug Discovery Today*, 12(23), 995-997.

Abraham, M. H., Chadha, H. S., Whiting, G. S., & Mitchell, R. C. (1994). Hydrogen bonding. 32. An analysis of water‐octanol and water‐alkane partitioning and the Δlog p parameter of seiler. *Journal of Pharmaceutical Sciences*, 83(8), 1085-1100.

Ajay, A., Walters, W. P., & Murcko, M. A. (1998). Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *Journal of Medicinal Chemistry*, 41(18), 3314-3324.

Amirkia, V., & Heinrich, M. (2014). Alkaloids as drug leads–A predictive structural and biodiversity-based analysis. *Phytochemistry Letters*, 10, xlviii-liii.

Amirkia, V., & Heinrich, M. (2015). Natural products and drug discovery: a survey of stakeholders in industry and academia. *Frontiers in Pharmacology*, 6, 237

Aniszewski, T. (2007). Biological significance of alkaloids. *Alkaloids–Secrets of Life: Alkaloid Chemistry, Biological Significance and Ecological Role*, 141-180.

Artuso, A. (1997). Drugs of natural origin: economic and policy aspects of discovery, development, and marketing. New York: Pharmaceutical Products Press.

Baker, D. D., Chu, M., Oza, U., & Rajgarhia, V. (2007). The value of natural products to future pharmaceutical discovery. *Natural Product Reports*, 24(6), 1225-1244.

Balunas, M. J., & Kinghorn, A. D. (2005). Drug discovery from medicinal plants. *Life Sciences*, 78(5), 431-441.

Berlinck, R. G. S., Kossuga, M. H., Fattorusso, E., & Taglialatela-Scafati, O. (2007). Modern Alkaloids. *Modern Alkaloids.*

Bhal, S. K., Kassam, K., Peirson, I. G., & Pearl, G. M. (2007). The Rule of Five revisited: applying log D in place of log P in drug-likeness filters. *Molecular Pharmaceutics*, 4(4), 556-560.

Bleicher, K. H., Böhm, H. J., Müller, K., & Alanine, A. I. (2003). Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery*, 2(5), 369-378.

Bohlin, L., Göransson, U., Alsmark, C., Wedén, C., & Backlund, A. (2010). Natural products in modern life science. *Phytochemistry Reviews*, 9(2), 279-301.

Breinbauer, R., Vetter, I. R., & Waldmann, H. (2002). From protein domains to drug candidates—natural products as guiding principles in the design and synthesis of compound libraries. Angewandte Chemie International Edition, 41(16), 2878-2890.

Brito, A. R., & Nunes, D. S. (1997). Ethnopharmacology and the sustainable development of new plant-derived drugs. *Ciência e Cultura* (Säo Paulo), 49(5/6), 402-50.

Brown, D., & Superti-Furga, G. (2003). Rediscovering the sweet spot in drug discovery. *Drug Discovery Today*, 8(23), 1067-1077.

Bruckingham, J. (2000). Dictionary of natural products on CD-ROM. New York: Champman and Hall.

Burbidge, R., Trotter, M., Buxton, B., & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & Chemistry*, 26(1), 5-14.

Butler, M. S. (2004). The role of natural product chemistry in drug discovery. *Journal of Natural Products*, 67(12), 2141-2153.

Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences*, 43(6), 1882-1889.

Chin, Y. W., Balunas, M. J., Chai, H. B., & Kinghorn, A. D. (2006). Drug discovery from natural sources. *The AAPS journal*, 8(2), E239-E253.

Clapp, R. A., & Crook, C. (2002). Drowning in the magic well: Shaman Pharmaceuticals and the elusive value of traditional knowledge. *The Journal of Environment & Development*, 11(1), 79-102.

Clement, J. A., Kitagaki, J., Yang, Y., Saucedo, C. J., O'Keefe, B. R., Weissman, & McMahon, J. B. (2008). Discovery of new pyridoacridine alkaloids from *Lissoclinum* cf. *badium* that inhibit the ubiquitin ligase activity of Hdm2 and stabilize p53. *Bioorganic & Medicinal Chemistry*, 16(23), 10022-10028.

Congreve, M., Carr, R., Murray, C., & Jhoti, H. (2003). A 'rule of three' for fragment-based lead discovery? Drug Discovery Today, 8(19), 876-877.

Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., & Pangalos, M. N. (2014). Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*, 13(6), 419-431.

Cordell, G. A., & Colvard, M. D. (2005). Some thoughts on the future of ethnopharmacology. J*ournal of Ethnopharmacology*, 100(1), 5-14.

Cordell, G. A., Quinn ‐ Beattie, M. L., & Farnsworth, N. R. (2001). The potential of alkaloids in drug discovery. *Phytotherapy Research*, 15(3), 183-205.

Corson, T. W., & Crews, C. M. (2007). Molecular understanding and modern application of traditional medicines: triumphs and trials. *Cell*, 130(5), 769-774.

Cox, P. A. (1990). Ethnopharmacology and the search for new drugs. In *Ciba Foundation Symposium 154-Bioactive Compounds from plants* (pp. 40-55). John Wiley & Sons, Ltd..

Cox, P. A., & Balick, M. J. (1994). The ethnobotanical approach to drug discovery. *Scientific American* (June), 60-65.

Cragg, G. M., & Newman, D. J. (2001). Natural product drug discovery in the next millennium. *Pharmaceutical Biology*, 39(sup1), 8-17.

Cragg, G. M., Newman, D. J., & Rosenthal, J. (2012). The impact of the United Nations Convention on Biological Diversity on natural products research. *Natural Product Reports*, 29(12), 1407-1423.

Cragg, G. M., Newman, D. J., & Snader, K. M. (1997). Natural products in drug discovery and development. *Journal of Natural Products*, 60(1), 52-60.

Crook, C. (2001). Biodiversity prospecting agreements, evaluating their economic and conservation benefits in Costa Rica and Peru. Unpublished dissertation.

Czárán, T. L., Hoekstra, R. F., & Pagie, L. (2002). Chemical warfare between microbes promotes biodiversity. *Proceedings of the National Academy of Sciences*, 99(2), 786-790.

Czapek, F., (1921). Spezielle Biochemie, Biochemie der Pflanzen, vol. 3, G. Fischer Jena, p. 369.

Danzon, P. M., Epstein, A., & Nicholson, S. (2004). Mergers and acquisitions in the pharmaceutical and biotech industries (No. w10536). National Bureau of Economic Research.

David, B., Wolfender, J. L., & Dias, D. A. (2014). The pharmaceutical industry and natural products: historical status and new trends. *Phytochemistry Reviews*, 14(2), 299-315.

Demirbag, M., Ng, C. K., & Tatoglu, E. (2007). Performance of mergers and acquisitions in the pharmaceutical industry: a comparative perspective. *Multinational Business Review*, 15(2), 41-62.

Ehrman, T. M., Barlow, D. J., & Hylands, P. J. (2007). Virtual screening of Chinese herbs with random forest. *Journal of Chemical Information and Modeling*, 47(2), 264-278.

Ertl, P., Rohde, B., & Selzer, P. (2000). Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20), 3714-3717.

Facchini, P. J., Bohlmann, J., Covello, P. S., De Luca, V., Mahadevan, R., Page & Martin, V. J. (2012). Synthetic biosystems for the production of high-value plant metabolites. *Trends in Biotechnology*, 30(3), 127-131

Farnsworth, N. R. (1993). Ethnopharmacology and future drug development: the North American experience. *Journal of Ethnopharmacology*, 38(2), 137-143.

Fattorusso, E., & Taglialatela-Scafati, O. (Eds.). (2008). *Modern alkaloids: structure, isolation, synthesis, and biology*. John Wiley & Sons.

Firn, R. D., & Jones, C. G. (2003). Natural products–a simple model to explain chemical diversity. *Natural Product Reports*, 20(4), 382-391.

Firn, R. D., & Jones, C. G. (2009). A Darwinian view of metabolism: molecular properties determine fitness. *Journal of Experimental Botany*, 60(3), 719-726.

Gates, M., & Tschudi, G. (1956). The synthesis of morphine. *Journal of the American Chemical Society*, 78(7), 1380-1393.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey & Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1), D1100-D1107.

Gertsch, J. (2009). How scientific is the science in ethnopharmacology? Historical perspectives and epistemological problems. *Journal of Ethnopharmacology*, 122(2), 177-183.

Ghose, A. K., Viswanadhan, V. N., & Wendoloski, J. J. (1998). Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *The Journal of Physical Chemistry*, 102(21), 3762-3772.

Ghose, A. K., Viswanadhan, V. N., & Wendoloski, J. J. (1999). A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *Journal of Combinatorial Chemistry*, 1(1), 55-68.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., & Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic Acids Research*, *38*(suppl 2), W695-W699.

Grabowski, K., & Schneider, G. (2007). Properties and architecture of drugs and natural products revisited. *Current Chemical Biology*, 1(1), 115-127.

Gray, T. C., & Halton, J. (1946). A milestone in anaesthesia?:(d-tubocurarine chloride). *Proceedings of the Royal Society of Medicine*, 39(7), 400.

Guiguemde, W. A., Shelat, A. A., Garcia-Bustos, J. F., Diagana, T. T., Gamo, F. J., & Guy, R. K. (2012). Global phenotypic screening for antimalarials. *Chemistry & Biology*, 19(1), 116-129.

Gullo, V. P., McAlpine, J., Lam, K. S., Baker, D., & Petersen, F. (2006). Drug discovery from natural products. *Journal of Industrial Microbiology and Biotechnology*, 33(7), 523-531.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18.

Han, L. Y., Ma, X. H., Lin, H. H., Jia, J., Zhu, F., Xue, Y. & Chen, Y. Z. (2008). A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *Journal of Molecular Graphics and Modelling*, 26(8), 1276-1286.

Hanson, J. R. (2003). *Natural products: the secondary metabolites* (Vol. 17). Royal Society of Chemistry.

Harvey, A. (2000). Strategies for discovering drugs from previously unexplored natural products. *Drug Discovery Today*, 5(7), 294-300.

Harvey, A. L. (1999). Medicines from nature: are natural products still relevant to drug discovery? *Trends in Pharmacological Sciences*, 20(5), 196-198.

Harvey, A. L. (2008). Natural products in drug discovery. *Drug Discovery Today*, 13(19), 894-901.

Harvey, A. L., Edrada-Ebel, R., & Quinn, R. J. (2015). The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery*, 14(2), 111-129.

Heinrich, M. (2013). Ethnopharmacology and drug discovery. *Comprehensive Natural Products II: Chemistry and Biology, Development & Modification of Bioactivity*, *3*, 351-381.

Heinrich, M., & Gibbons, S. (2001). Ethnopharmacology in drug discovery: an analysis of its role and potential contribution. *Journal of Pharmacy and Pharmacology*, 53(4), 425-432.

Heinrich, M., Edwards, S., Moerman, D. E., & Leonti, M. (2009). Ethnopharmacological field studies: a critical assessment of their conceptual basis and methods. *Journal of Ethnopharmacology*, 124(1), 1-17.

Heinrich, M., & Teoh, H. L. (2004). Galanthamine from snowdrop—the development of a modern drug against Alzheimer's disease from local Caucasian knowledge. *Journal of Ethnopharmacology*, 92(2), 147-162.

Helma, C., Cramer, T., Kramer, S., & De Raedt, L. (2004). Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *Journal of Chemical Information and Computer Sciences*, 44(4), 1402-1411.

Henkel, T., Brunne, R. M., Müller, H., & Reichel, F. (1999). Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angewandte Chemie International Edition*, 38(5), 643-647.

Higgins, M. J., & Rodriguez, D. (2006). The outsourcing of R&D through acquisitions in the pharmaceutical industry. *Journal of Financial Economics*, 80(2), 351-383.

Hodgson, B. (2001). *In the arms of morpheus: The tragic history of laudanum, morphine, and patent medicines*. Firefly Books Limited.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11), 682-690.

Howell, C. R., Bell, A. A., & Stipanovic, R. D. (1973). Virulence to cotton and tolerance to sanguinarine among Verticillium species. *Canadian Journal of Microbiology*, 19(11), 1367-1371.

Hung, D. T., Jamison, T. F., & Schreiber, S. L. (1996). Understanding and controlling the cell cycle with natural products. *Chemistry & Biology*, 3(8), 623-639.

IMS Health (2014). IMS Health Report - Total Unaudited and Audited Global Pharmaceutical Market By Region

International Labour Organization (ILO) (1989). Indigenous and Tribal Peoples Convention, [Online]. C169. http://www.refworld.org/docid/3ddb6d514.html [Accessed 13 May 2015].

International Narcotics Control Board (2013). Comments on the reported statistics on narcotic drugs [Online]. https://www.incb.org/documents/Narcotic-Drugs/Technical-Publications/2013/Part_2_Comments_E.pdf. [Accessed 20 May 2013].

Isman, M. B. (2006). Botanical insecticides, deterrents, and repellents in modern agriculture and an increasingly regulated world. *Annu. Rev. Entomol.*, *51*, 45-66.

Jachak, S. M., & Saklani, A. (2007). Challenges and opportunities in drug discovery from plants. *Current Science-Bangalore*, 92(9), 1251.

Ji, H. F., Li, X. J., & Zhang, H. Y. (2009). Natural products and drug discovery. *EMBO Reports*, 10(3), 194-200.

Kaufman, P. B., Cseke, L. J., Warber, S., Duke, J. A., & Brielmann, H. L. (1999). *Natural Products from Plants* (No. 581.192 N3). Boca Raton FL: CRC press.

Kenny, P. W., & Montanari, C. A. (2013). Inflation of correlation in the pursuit of drug-likeness. *Journal of Computer-aided Molecular Design*, 1-13.

Khanna, I. (2012). Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discovery Today*, 17(19), 1088-1102.

King, S. R., & Carlson, T. J. (1995). Biocultural diversity, biomedicine and ethnobotany: the experience of Shaman Pharmaceuticals. *Interciencia Caracas*, 20, 134-134.

Kingston, D. G. (2010). Modern natural products drug discovery and its relevance to biodiversity conservation. *Journal of Natural Products*, 74(3), 496-511.

Knapp, S. (2010). What's in a name? A history of taxonomy. Natural History Museum. http://www.nhm.ac.uk/nature-online/science-of-naturalhistory/taxonomy-systematics/history-taxonomy/index.html [Accessed 31 June 2015].

Knight, V., Sanglier, J. J., DiTullio, D., Braccili, S., Bonner, P., Waters, J., & Zhang, L. (2003). Diversifying microbial natural products for drug discovery. *Applied Microbiology and Biotechnology*, 62(5-6), 446-458.

Koch, M. A., Schuffenhauer, A., Scheck, M., Wetzel, S., Casaulta, M., Odermatt, & Waldmann, H. (2005). Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proceedings of the National Academy of Sciences of the United States of America*, 102(48), 17272-17277.

Koehn, F. E., & Carter, G. T. (2005). The evolving role of natural products in drug discovery. *Nature Reviews Drug Discovery*, 4(3), 206-220.

Kossel A (1891). "Ueber die chemische Zusammensetzung der Zelle" [The chemical composition of the cell]. Archiv für Physiologie (in German): 181–186.

Kubinyi, H., & Folkers, G. (2008). *Molecular Drug Properties*. R. Mannhold (Ed.). John Wiley & Sons. 111-126

Lagunin, A. A., Goel, R. K., Gawande, D. Y., Pahwa, P., Gloriozova, T. A., Dmitriev, & Druzhilovsky, D. S. (2014). Chemo-and bioinformatics resources for in silico drug discovery from medicinal plants beyond their traditional use: a critical review. *Natural Product Reports*, 31(11), 1585-1611.

Lam, K. S. (2007). New aspects of natural products in drug discovery. *Trends in Microbiology*, 15(6), 279-289.

Lawrence, G.H.M., (1951). *The Taxonomy of Vascular Plants*. The Macmillan Company, New York

Li, J. W. H., & Vederas, J. C. (2009). Drug discovery and natural products: end of an era or an endless frontier? *Science*, 325(5937), 161-165.

Lipinski, C. A. (2004). Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4), 337-341.

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1), 3-25.

Lobell, M., Hendrix, M., Hinzen, B., Keldenich, J., Meier, H., Schmeck, C., & Hillisch, A. (2006). In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem*, 1(11), 1229-1236.

Lutz, M., & Kenakin, T. (1999). *Quantitative molecular pharmacology and informatics in drug discovery*. John Wiley & Sons.

Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes & Sittampalam, G. S. (2011). Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3), 188-195.

Mahfouz, M. (1949). The fate of tubocurarine in the body. *British Journal of Pharmacology and Chemotherapy*, 4(3), 295-303.

McChesney, J. D., Venkataraman, S. K., & Henri, J. T. (2007). Plant natural products: back to the future or into extinction? *Phytochemistry*, 68(14), 2015-2022.

Medina-Franco, J. L., Giulianotti, M. A., Welmaker, G. S., & Houghten, R. A. (2013). Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discovery Today*, 18(9), 495-501.

Mishra, K. P., Ganju, L., Sairam, M., Banerjee, P. K., & Sawhney, R. C. (2008). A review of high throughput technology for the screening of natural products. *Biomedicine & Pharmacotherapy*, 62(2), 94-98.

Monaghan, R. L., Polishook, J. D., Pecore, V. J., Bills, G. F., Nallin-Omstead, M., & Streicher, S. L. (1995). Discovery of novel secondary metabolites from fungi-is it really a random walk through a random forest? *Canadian Journal of Botany*, 73(S1), 925-931.

Morphy, R., Kay, C., & Rankovic, Z. (2004). From magic bullets to designed multiple ligands. *Drug Discovery Today*, 9(15), 641-651.

Nakagawa, A., Minami, H., Kim, J. S., Koyanagi, T., Katayama, T., Sato, F., & Kumagai, H. (2011). A bacterial platform for fermentative production of plant alkaloids. *Nature Communications*, 2, 326.

Newman, D. J., & Cragg, G. M. (2004). Marine natural products and related compounds in clinical and advanced preclinical trials. *Journal of Natural Products*, 67(8), 1216-1238.

Newman, D. J., & Cragg, G. M. (2007). Natural Products as Sources of New Drugs over the Last 25 Years. *Journal of Natural Products*, 70(3), 461-477.

Newman, D. J., & Cragg, G. M. (2012). Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of Natural Products*, 75(3), 311-335.

Newman, D. J., Cragg, G. M., & Snader, K. M. (2000). The influence of natural products upon drug discovery. *Natural Product Reports*, 17(3), 215-234.

Newman, D. J., Cragg, G. M., & Snader, K. M. (2003). Natural products as sources of new drugs over the period 1981-2002. *Journal of Natural Products*, 66(7), 1022-1037.

Newman, S. E., Roll, M. J., & Harkrader, R. J. (1999). A naturally occurring compound for controlling powdery mildew of greenhouse roses. *HortScience*, 34(4), 686-689.

Niedergassel, B., & Leker, J. (2009). Open innovation: chances and challenges for the pharmaceutical industry. *Future medicinal chemistry*, 1(7), 1197-1200.

Nisbet, L. J., & Moore, M. (1997). Will natural products remain an important source of drug research for the future? *Current Opinion in Biotechnology*, 8(6), 708-712.

Norinder, U., & Haeberlein, M. (2002). Computational approaches to the prediction of the blood–brain distribution. *Advanced Drug Delivery Reviews*, 54(3), 291-313.

Oprea, T. I. (2002). Current trends in lead discovery: Are we looking for the appropriate properties? *Journal of Computer-aided Molecular Design*, 16(5-6), 325-334.

Ortholand, J. Y., & Ganesan, A. (2004). Natural products and combinatorial chemistry: back to the future. *Current Opinion in Chemical Biology*, 8(3), 271-280.

Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery*, 5(12), 993-996.

Owens, J. (2003). Chris Lipinski discusses life and chemistry after the Rule of Five. *Drug Discovery Today*, 8(1), 12-16.

Paterson, D. A., Conradi, R. A., Hilgers, A. R., Vidmar, T. J., & Burton, P. S. (1994). A Non ‐ aqueous Partitioning System for Predicting the Oral Absorption Potential of Peptides. *Quantitative structure ‐ Activity Relationships*, 13(1), 4-10.

Paterson, I., & Anderson, E. A. (2005). The renaissance of natural products as drug candidates. *Science,* 310(5747), 451.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3), 203-214.

Pharmaceutical Research and Manufacturers of America (2015). Biopharmaceutical research industry profile. Washington, DC: PhRMA.

Principe, P. P. (1991). Valuing the biodiversity of medicinal plants. *Conservation of Medicinal Plants*, 79-124.

Rates, S. M. K. (2001). Plants as source of drugs. *Toxicon*, 39(5), 603-613.

Reid, W. V., Barber, C. V., & La Vina, A. (1995). Translating genetic resources rights into sustainable development: gene cooperatives, the biotrade and lessons from the Philippines. *Plant Genetic Resources Newsletter (IPGRI/FAO)*.

Rinner, U., & Hudlicky, T. (2011). Synthesis of morphine alkaloids and derivatives. In *Alkaloid Synthesis* (pp. 33-66). Springer Berlin Heidelberg.

Rishton, G. M. (2008). Natural products as a robust source of new drugs and drug leads: past successes and present day issues. *The American Journal of Cardiology*, 101(10), S43-S49.

Roberts, R. A., Kavanagh, S. L., Mellor, H. R., Pollard, C. E., Robinson, S., & Platz, S. J. (2014). Reducing attrition in drug development: smart loading preclinical safety assessment. *Drug Discovery Today*, 19(3), 341-347.

Roberts, A. D., Baggaley, K. H., & Buckingham, J. (2010). *Dictionary of Alkaloids*.

Rosén, J., Gottfries, J., Muresan, S., Backlund, A., & Oprea, T. I. (2009). Novel chemical space exploration *via* natural products. *Journal of Medicinal Chemistry*, 52(7), 1953-1962.

Sams-Dodd, F. (2005). Target-based drug discovery: is something wrong? *Drug Discovery Today*, 10(2), 139-147.

Sams-Dodd, F. (2013). Is poor research the cause of the declining productivity of the pharmaceutical industry? An industry in need of a paradigm shift. *Drug Discovery Today*, 18(5), 211-217.

Shen, J., Xu, X., Cheng, F., Liu, H., Luo, X., Shen, J. & Jiang, H. (2003). Virtual screening on natural products for discovering active compounds and target information. *Current Medicinal Chemistry*, 10(21), 2327-2342.

Shu, Y. Z. (1998). Recent natural products based drug development: a pharmaceutical industry perspective. *Journal of Natural Products*, 61(8), 1053-1071.

Sobell, H. M., Sakore, T. D., Tavale, S. S., Canepa, F. G., Pauling, P., & Petcher, T. J. (1972). Stereochemistry of a curare alkaloid: O, O′ , N-trimethyl-d-tubocurarine. *Proceedings of the National Academy of Sciences*, 69(8), 2212-2215.

Stenberg, P., Luthman, K., Ellens, H., Lee, C. P., Smith, P. L., Lago, & Artursson, P. (1999). Prediction of the intestinal absorption of endothelin receptor antagonists using three theoretical methods of increasing complexity. *Pharmaceutical Research*, 16(10), 1520-1526.

Stepp, J. R., & Moerman, D. E. (2001). The importance of weeds in ethnopharmacology. *Journal of Ethnopharmacology*, 75(1), 19-23.

Stepp, J. R. (2004). The role of weeds as sources of pharmaceuticals. *Journal of Ethnopharmacolog*y, 92(2), 163-166.

Strobel, G. A. (2002). Rainforest endophytes and bioactive products. *Critical Reviews in Biotechnology*, 22(4), 315-333.

Tralau-Stewart, C. J., Wyatt, C. A., Kleyn, D. E., & Ayad, A. (2009). Drug discovery: new models for industry–academic partnerships. *Drug Discovery Today*, 14(1), 95-101.

Tulp, M., & Bohlin, L. (2004). Unconventional natural sources for future drug discovery. *Drug Discovery Today*, 9(10), 450-458.

Urizar, N. L., Liverman, A. B., D'Nette, T. D., Silva, F. V., Ordentlich, P., Yan, Y.& Moore, D. D. (2002). A natural product that lowers cholesterol as an antagonist ligand for FXR. *Science,* 296(5573), 1703-1706.

Ulubelen, A., Mericli, A. H., Meriçli, F., Kilinçer, N., Ferizli, A. G., Emekci, M., & Pelletier, S. W. (2001). Insect repellent activity of diterpenoid alkaloids. *Phytotherapy Research*, 15(2), 170-171.

Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45(12), 2615-2623.

Vedani, A., Dobler, M., & Smieško, M. (2012). VirtualToxLab—a platform for estimating the toxic potential of drugs, chemicals and natural products.*Toxicology and Applied Pharmacology*, 261(2), 142-153.

Verdine, G. L. (1996). The combinatorial chemistry of nature. *Nature,* 384(6604), 11-13.

Vuorela, P., Leinonen, M., Saikku, P., Tammela, P., Rauha, J. P., Wennberg, T., & Vuorela, H. (2004). Natural products in the process of finding new drug candidates. *Current Medicinal Chemistry*, 11(11), 1375-1389.

Walsh, R. (2010). A history of: The pharmaceutical industry. Pharmaphorum [Online]. http://pharmaphorum.com/views-and-analysis/a_history_of_the_pharmaceutical_industry/. [Accessed 13 January 2016].

Walters, W. P., & Murcko, M. A. (2002). Prediction of ‘drug-likeness’. *Advanced Drug Delivery Reviews*, 54(3), 255-271.

Wang, X. J., Li, L., Si, Y. K., Yu, S. S., Ma, S. G., Bao, X. & Li, Y. (2013). Nine new lycopodine-type alkaloids from Lycopodium japonicum Thunb. *Tetrahedron*, 69(30), 6234-6240.

Warmuth, M. K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., & Lemmen, C. (2003). Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences*, 43(2), 667-673.

Wenlock, M. C., Austin, R. P., Barton, P., Davis, A. M., & Leeson, P. D. (2003). A comparison of physiochemical property profiles of development and marketed oral drugs. *Journal of Medicinal Chemistry*, 46(7), 1250-1256.

Wernerova, M., & Hudlicky, T. (2010). On the practical limits of determining isolated product yields and ratios of stereoisomers: reflections, analysis, and redemption. *Synlett*, 2010 (18), 2701-2707.

Weston, J., Perez-Cruz, F., Bousquet, O., Chapelle, O., Elisseeff, A., & Schölkopf, B. (2002). Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics*, 1(1), 1-8.

Wolf, D., & Siems, K. (2007). Burning the Hay to Find the Needle Data Mining Strategies in Natural Product Dereplication. *CHIMIA International Journal for Chemistry*, 61(6), 339-345.

Working Group on Indigenous Populations (WGIP) (1993). UN draft declaration on the rights of indigenous peoples [Agreed to at 11th session of WGIP].

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M. & Culham, A. (2007). How global is the global biodiversity information facility. *PLOS ONE*, 2(11), e1124.

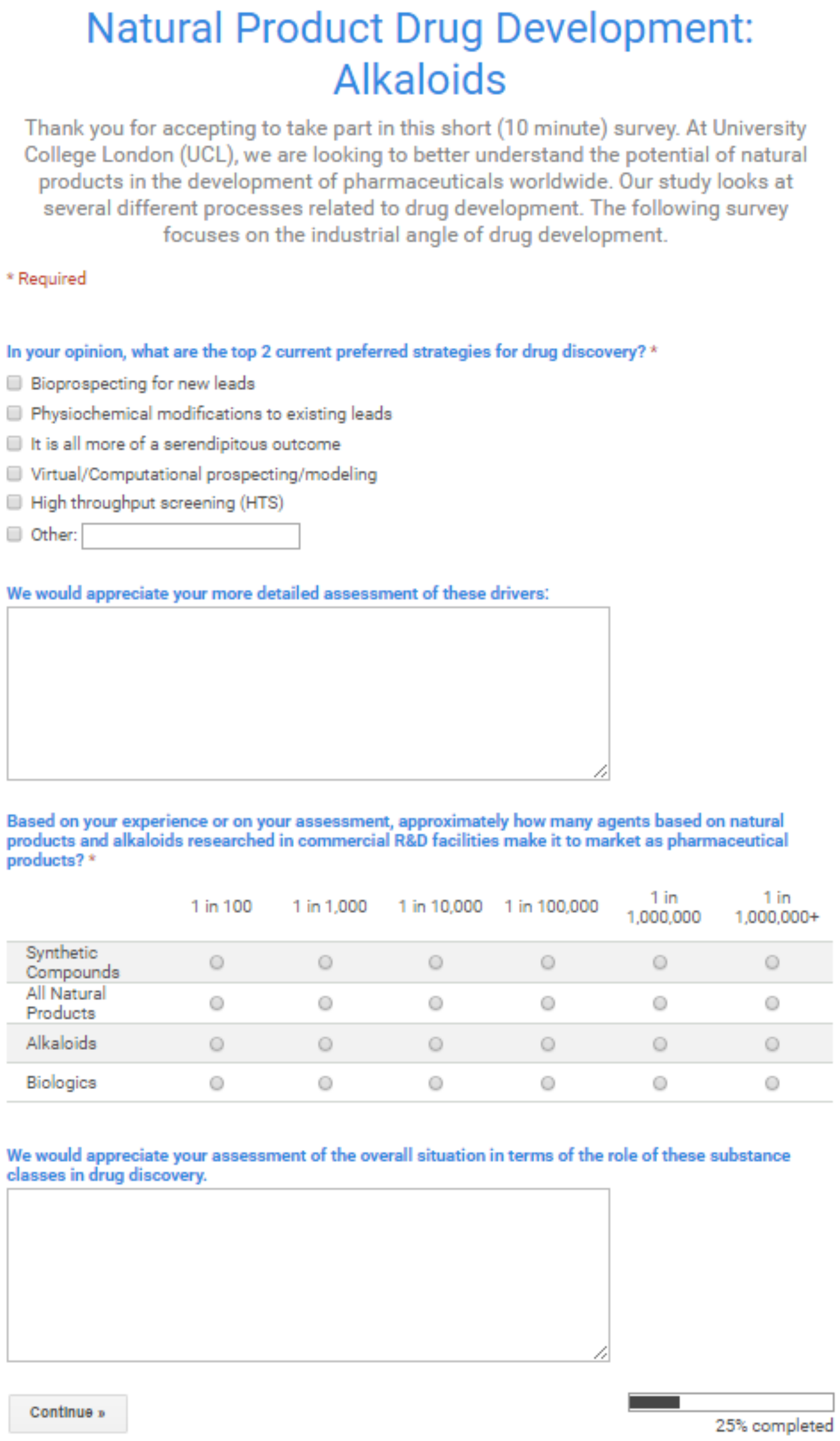
Zernov, V. V., Balakin, K. V., Ivaschenko, A. A., Savchuk, N. P., & Pletnev, I. V. (2003). Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *Journal of Chemical Information and Computer Sciences*, 43(6), 2048-2056.

Zhang, M. Q., & Wilkinson, B. (2007). Drug discovery beyond the ‘rule-of-five’. *Current Opinion in Biotechnology*, 18(6), 478-488.

Zhang, P. Z., Zhang, Y. M., Gu, J., & Zhang, G. L. (2016). Two new alkaloids from Melodinus hemsleyanus Diels. *Natural Product Research*, 30(2), 162-167.

Appendix

Appendix 1: Screenshot of online natural product drug development survey for stakeholders in industry and academia. Survey hosted by Google Forms.



Natural Product Drug Development: Alkaloids

Thank you for accepting to take part in this short (10 minute) survey. At University College London (UCL), we are looking to better understand the potential of natural products in the development of pharmaceuticals worldwide. Our study looks at several different processes related to drug development. The following survey focuses on the industrial angle of drug development.

* Required

In your opinion, what are the top 2 current preferred strategies for drug discovery? *

□ Bioprospecting for new leads

□ Physiochemical modifications to existing leads

□ It is all more of a serendipitous outcome

□ Virtual/Computational prospecting/modeling

□ High throughput screening (HTS)

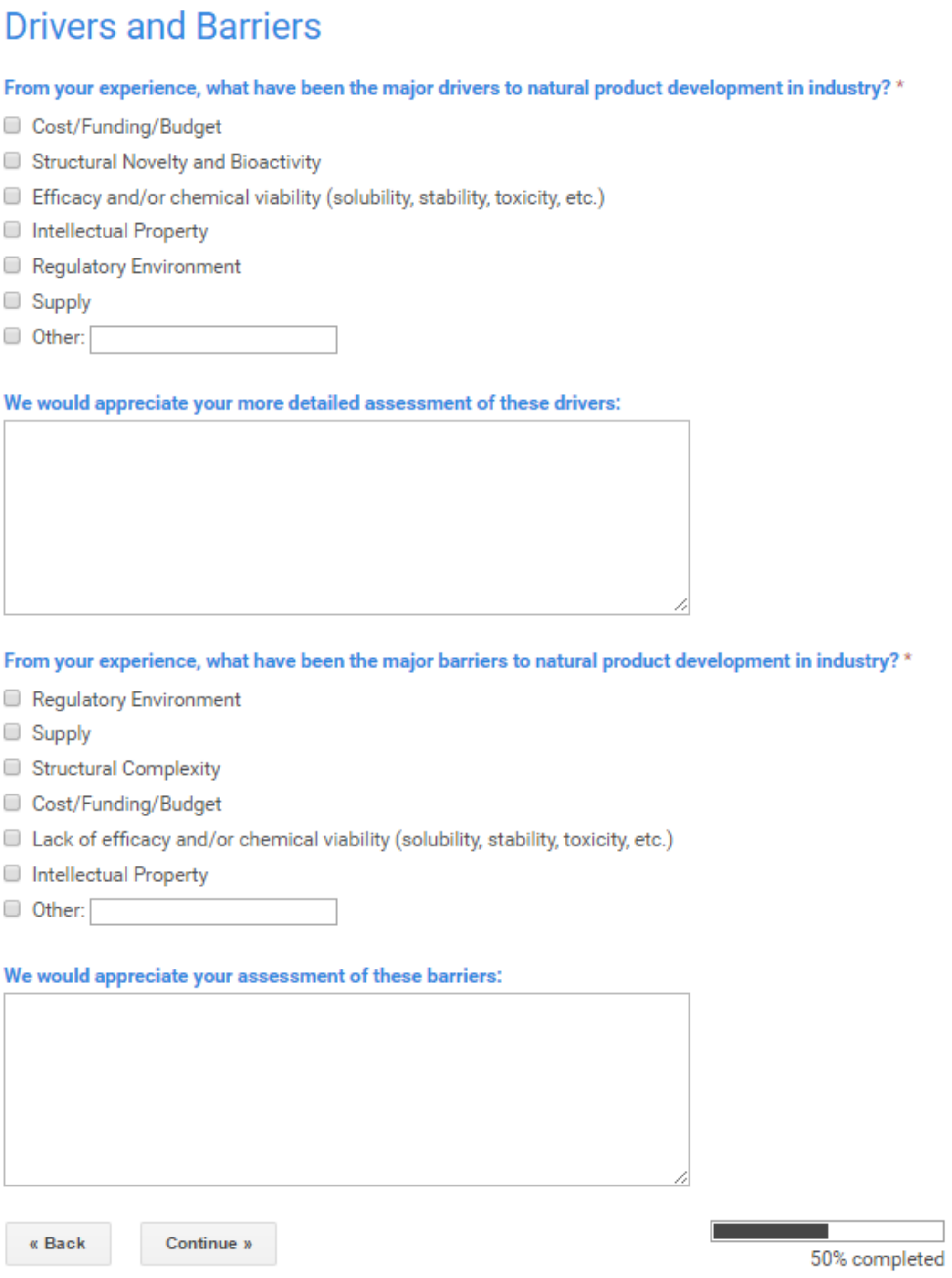
□ Other: ______

We would appreciate your more detailed assessment of these drivers:

Based on your experience or on your assessment, approximately how many agents based on natural products and alkaloids researched in commercial R&D facilities make it to market as pharmaceutical products? *

	1 in 100	1 in 1,000	1 in 10,000	1 in 100,000	1 in 1,000,000	1 in 1,000,000+
Synthetic Compounds	○	○	○	○	○	○
All Natural Products	○	○	○	○	○	○
Alkaloids	○	○	○	○	○	○
Biologics	○	○	○	○	○	○

We would appreciate your assessment of the overall situation in terms of the role of these substance classes in drug discovery.



Drivers and Barriers

From your experience, what have been the major drivers to natural product development in industry? *

□ Cost/Funding/Budget

□ Structural Novelty and Bioactivity

□ Efficacy and/or chemical viability (solubility, stability, toxicity, etc.)

□ Intellectual Property

□ Regulatory Environment

□ Supply

□ Other: __________

We would appreciate your more detailed assessment of these drivers:

From your experience, what have been the major barriers to natural product development in industry? *

□ Regulatory Environment

□ Supply

□ Structural Complexity

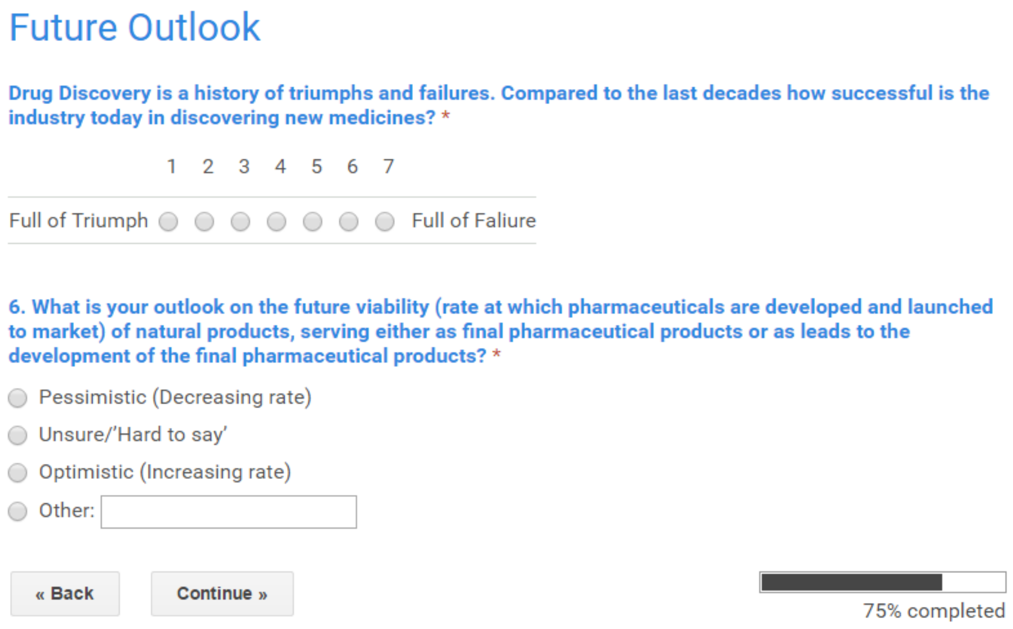
□ Cost/Funding/Budget

□ Lack of efficacy and/or chemical viability (solubility, stability, toxicity, etc.)

□ Intellectual Property

□ Other: __________

We would appreciate your assessment of these barriers:



Future Outlook

Drug Discovery is a history of triumphs and failures. Compared to the last decades how successful is the industry today in discovering new medicines? *

	1	2	3	4	5	6	7	
Full of Triumph	○	○	○	○	○	○	○	Full of Faliure

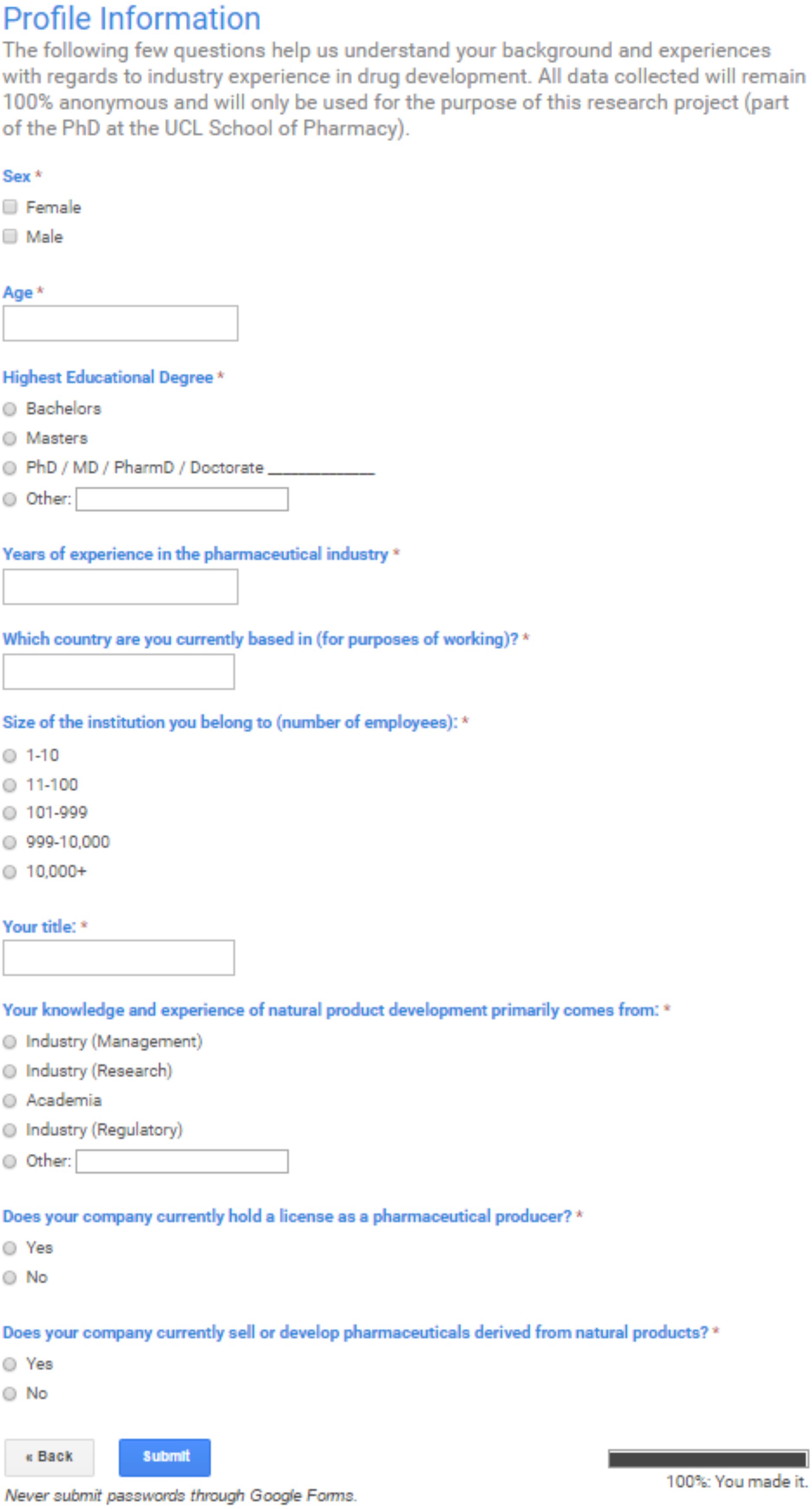
6. What is your outlook on the future viability (rate at which pharmaceuticals are developed and launched to market) of natural products, serving either as final pharmaceutical products or as leads to the development of the final pharmaceutical products? *

○ Pessimistic (Decreasing rate)

○ Unsure/'Hard to say'

○ Optimistic (Increasing rate)

○ Other: □



Profile Information

The following few questions help us understand your background and experiences with regards to industry experience in drug development. All data collected will remain 100% anonymous and will only be used for the purpose of this research project (part of the PhD at the UCL School of Pharmacy).

Sex *

□ Female

□ Male

Age *

Highest Educational Degree *

○ Bachelors

○ Masters

○ PhD / MD / PharmD / Doctorate ____________

○ Other:

Years of experience in the pharmaceutical industry *

Which country are you currently based in (for purposes of working)? *

Size of the institution you belong to (number of employees): *

○ 1-10

○ 11-100

○ 101-999

○ 999-10,000

○ 10,000+

Your title: *

Your knowledge and experience of natural product development primarily comes from: *

○ Industry (Management)

○ Industry (Research)

○ Academia

○ Industry (Regulatory)

○ Other:

Does your company currently hold a license as a pharmaceutical producer? *

○ Yes

○ No

Does your company currently sell or develop pharmaceuticals derived from natural products? *

○ Yes

○ No

100%: You made it.

Never submit passwords through Google Forms.

Appendix 2: DNP categorizations of alkaloids (Buckingham, 2000)

Major Class	Subclass
Alkaloids derived from ornithine	1. Simple ornithine alkaloids 2. Chromone alkaloids 3. Tropane alkaloids 4. Pyrrolizidine alkaloids 5. Miscellaneous ornithine-derived alkaloids
Alkaloids derived from lysine	6. Simple piperidine alkaloids 7. Lobelia alkaloids 8. More complex lysine-derived alkaloids 9. Lycopodium alkaloids 10. Lythraceae alkaloids
Alkaloids of polyketide origin	11. Naphthalene-isoquinoline alkaloids 12. Cytochalasan alkaloids
Alkaloids derived from anthranilic acid	13. Simple anthranilic acid derivatives 14. Simple quinoline alkaloids 15. Quinazoline alkaloids 16. Acridone alkaloids 17. Acridone-coumarin alkaloid dimers 18. 1,4-Benzoxazin-3-one alkaloids 19. Benzodiazepine alkaloids 20. Cryptolepine-type alkaloids
Alkaloids derived wholly or in part from phenylalanine or tyrosine	21. Simple tyramine alkaloids 22. Cinnamic acid amides 23. Securinega alkaloids 24. Betalain alkaloids

Isoquinoline alkaloids	25. Simple isoquinoline alkaloids 26. Benzylisoquinoline alkaloids 27. Pseudobenzylisoquinoline alkaloids 28. Bisbenzylisoquinoline alkaloids 29. Secobisbenzylisoquinoline alkaloids 30. Cularine group alkaloids 31. Secocularine alkaloids 32. Cancentrine-type alkaloids 33. Quettamine alkaloids 34. Dibenzopyrrocoline alkaloids 35. Pavine and isopavine alkaloids 36. Proaporphine alkaloids 37. Aporphine alkaloids 38. Morphine alkaloids 39. Dibenzazecine and Hasubanan alkaloids 40. Protoberberine alkaloids 41. Narceine and phthalideisoquinoline alkaloids 42. Protopine alkaloids 43. Rhoeadine alkaloids 44. Spirobenzylisoquinoline alkaloids 45. Benzo[c]phenanthridine alkaloids 46. Phenethylisoquinoline alkaloids 47. Homoaporphine alkaloids 48. Homoerythrina alkaloids 49. Colchicine-like alkaloids 50. Dibenzocycloheptylamine alkaloids 51. Erythrina and cephalotaxus alkaloids 52. Amaryllidaceae alkaloids 53. Mesembrenoid alkaloids 54. Emetine group alkaloids 55. Phenanthroindolizidine and phenanthroquinolizidine alkaloids

Alkaloids derived from tryptophan	56. Simple tryptamine alkaloids 57. Physostigmine-like alkaloids 58. Carbazole alkaloids 59. Miscellaneous tryptophan derivatives 60. β-Carboline alkaloids 61. Aristotelia alkaloids 62. Borreria alkaloids 63. Ergot alkaloids
Monoterpenoid indole alkaloids	64. Monoterpenoid-derived indole alkaloid glycosides 65. Camptothecin-like alkaloids 66. Indoloquinolizidine alkaloids 67. Corynanthe alkaloids 68. Corynanthe tryptamine alkaloids 69. Ajmalicine-like alkaloids 70. Oxindole alkaloids 71. Gelsemium alkaloids 72. Yohimbinoid alkaloids 73. Akuammiline alkaloids 74. Sarpagine alkaloids 75. Ajmaline alkaloids 76. Pleiocarpamine alkaloids 77. Cinchona alkaloids 78. Strychnos alkaloids 79. Condylocarpan alkaloids 80. Secodine alkaloids 81. Aspidosperma alkaloids 82. Kopsane alkaloids 83. Quebrachamine and pandoline alkaloids 84. Iboga alkaloids 85. Pyridocarbazole alkaloids 86. Uleine-dasycarpidan alkaloids 87. Eburna alkaloids 88. Bisindole alkaloids

Terpenoid alkaloids	89. Monoterpenoid alkaloids 90. Dendrobium alkaloids 91. Nuphar alkaloids 92. Macrocyclic sesquiterpene alkaloids 93. Erythrophleum alkaloids 94. C_{19} and C_{20} Diterpenoid alkaloids and 4-nor analogues 95. Miscellaneous diterpenoid alkaloids 96. Olivoretin group 97. Daphniphylline alkaloids
Steroidal alkaloids	98. Steroidal alkaloids (pregnane type) 99. Steroidal alkaloids (conanine type) 100. Steroidal alkaloids (spirosolane and solanidine type) 101. Steroidal alkaloids (buxus type) 102. Steroidal alkaloids (salamandra type) 103. Miscellaneous steroidal alkaloids
Imidazole alkaloids	
Oxazole alkaloids	
Thiazole alkaloids	
Pyrazine and quinoxaline alkaloids	
Pyrrole alkaloids	
Putrescine alkaloids	
Spermine and spermidine alkaloids	
Peptide alkaloids	
Purines	
Pteridines and analogues	

Appendix 3: Record of alkaloids used in marketed drugs and clinical environments (excluding additive used in extending the storage life of drawn blood) (The Royal Society of Chemistry, 1971; Cordell, 1981; Schmeller and Wink, 1998; Buckingham, 2010). Derivatives not included.

Alkaloid	Synonyms	Applications	Example product
Aconitine		Rheumatism, neuralgia, sciatica	Aconitysat™, Bronpax™, Pectovox™, Vocadys™
Adenine		Antiviral agent, pharmaceutical aid used to extend storage life of whole blood	Adenosine, Ansyr®
Ajmaline	Ajimaline, Gilurytmal, Merabitol, Raugalline, Rauwolfine, Rytmalin, Tachmalin	Antiarrhythmic agent	Aritmina™, Gilurytmal™, Rauwopur™, Ritmos™
Atropine	Tropine tropate	Antispasmodic, anti-parkinson, cycloplegic drug	Abdominol™, Espasmo™, Protecor™, Tonaton™
Berberine	Berbericine, Umbellatine,	Eye irritations, AIDS, hepatitis	Kollyr™, Murine™, Sedacollyre™
Boldine		Cholelithiasis, vomiting, constipation	Boldoflorine™, Boldosal™, Oxyboldine™, Sambil™
Caffeine		Neonatal apnea, atopic dermatitis	Agevis™, Anlagen™, Thomapyrine™, Vomex A™
Canescine	Harmonyl, Raunormine, Recanescine, Reserpidine	Antihypertensive agent	Deserpidine
Cathine	Norpseudoephedrine, Norisoephedrine	Anorectic drug	Amorphan™, Eetless™, Recatol™

Cinchonidine	Cinchonan-9-ol	Increases reflexes, epileptiform convulsions	Quinimax™, Paluject™
Cocaine		Local anesthetic	Used in highly regulated clinical environments
Codeine	Methylmorphine, Codicept, Kodein, Tussipan	Antitussive, analgesic	Antituss™, Codicaps™, Tussipax™
Colchicine		Amyloidosis treatment, acute gout	ColBenemid™, Colgout™, Verban™
Diethanolamine	2,2'-Dihydroxydiethylamine, Diolamine	Base used in pharmaceuticals etc.	*Menbutone Diethanolamine*
Emetine	Ipecine, Methylcephaleine	Intestinal amoebiasis, expectorant drug	Cophylac™, Ipecac™, Rectopyrine™
Ephedrine		Nasal decongestant, bronchodilator	Amidoyna™, Bronchicum™, Peripherin™, Solamin™
Ergometrine	Ergonovine, Ergotrate, Ergobasine, Ergotocin, Ergostetrine	Postpartum/postabortal hemorrhage	Ergometron™, Ergotrate Maleat™, Syntometrine™
Ergotamine		Migrane treatment	Ergostat™, Lingraine™, Migral™, Virdex™
Eserine	Physostigmine	Ophthalmology, antidote/poisoning	Anticholium™, Antilirium™, Piloserine™
Galanthamine	Galantamine, Jilkon, Karantonin, Lycoremine	Muscle relaxant, Alzheimer’s	Nivalina™

Hydrastine		Gastrointestinal disorders	Gine Sedans™, Kollyr™
Hyoscine	Scopolamine	Motion sickness,	Buscopan™, Hyospasmol™, Lotanal™, Transcop™
Hyoscyamine	Daturine, Duboisine	Antispasmodic, anti-parkinson, cycloplegic drug	Bellatard™, Cystospaz™, Donnatab™, Urised™
Lobeline		Anti-smoking, asthma, cough	Citotal™, Lobatox™, Refrane™, Stopsmoke™
Morphine		Pain relief, diarrhea	Diastat™, Duromorph™, Oramprph™, Spasmofen™
N,N - Diallylbisnortoxinerine	Alcuronium chloride	Short acting muscular relaxant	Alloferin
Narceine		Cough suppressant	Peneraj™
Nicotine		Anti-smoking	Nicabate™, Nicoderm™, Nicorette™, Stubit™
Noscapine	Narcotine	Cough suppressant	Bequitusin™, Degoran™, Tossamine™, Tussisedal™
Papaverine	Papaveroline tetramethyl ether	Vasodilator, gastrointestinal disorders	Acticarbine™, Opdensit™, Pameion™, Vasocalm™
Pelletierine		Tenia infestations	Pelletierine tannate USP

Pilocarpine	Ocucarpine, Pilocarpol, Syncarpine	Miotic in treatment of glaucoma, leprosy	Frikton™, Piladren™, Salegen, Thiloadren™, Vistacarpin™
Quinidine	Conquinine, Conchinine, Pitayine	Ventricular and supraventricular arrhythmias, malaria, cramping	Cardioquin™, Duraquin™, Quindex™, Rhythomochin 1™
Quinine	6'-Methoxycinchonan-9-ol, Chinin	Malaria, babesiosis, myotonic disorders	Adaquin™, Biquinate™, Quinoctal™, Zynedo-B™
Raubasine	Ajmalicine	Vascular disorders	Circolene™, Cristanyl™, Duxil™, Sarpan™
Rescinnamine	Reserpinine, Anaprel, Apoterin, Cinnaloid, Rescaloid, Moderil, Scinnamina	Hypertension	Detensitral™, Diuraupur™, Rauwopur™
Reserpine		Hypertension, psychoses	Abicol™, Briserin™, Sandril™, Terbolan™
Rotundine	Argemonine , Bisnorargemonine,	Analgesic, sedative, hypnotic agent	Rotundin-BVP, Transda
Sanguinarine		Antiplaque agent	Toothpastes and mouthwashes
Sparteine		Uterine contractions, cardiac arrhythmias	Anxoral™, Diffucord™, Normotin™, Tachynerg™

Strychnine	Strychnidin-10-one	Eye disorders	Dysurgal™, Pasuma™ Retinovix™, Senirakt™
Synephrine		Vasoconstrictor, conjunctival decongestant, weight loss	Oxedrine, Sympatol
Taxol	Paclitaxel, Taxol A, Anzatax, Yewtaxan	Mamma and ovary carcinoma	Taxol™
Theobromine		Asthma, diuretic agent	Atrofed™, Circovegetalin™, Dynamol™, Urodonal™
Theophylline	Austyn, Elan, Elixophyllin, Euphyllin, Nuelin	Asthma, bronchospasms	Adenovasin™, Aerobin™, Euphyllin™, Theochron™
Turbocuranine	Tubarine	Muscle relaxant	Jexin™, Tubarine™
Vinblastine		Hodgkin’s disease, testicular cancer, blood disorders	Periblastine™, Velban™, Velbe™, Velsar™
Vincamine		Vasodilator	Aethroma™, Angiopac™, Pervin™, Vincimax™
Vincristine		Burkitt’s lymphoma	Norcristine™, Oncovin™, Vincrisul™
Vindesine		Chemotherapy	DAVA, Eldesine, Eldisine

Yohimbine	Aphrodine, Corymbin, Corynine, Yohimex, Hydroergotocin, Quebrachine, Yohimvetol	Aphrodisiac, urinary incontinence	Aphrodyne™, Pasuma™, Prowess™, Yohimex™

Appendix 4: CD-ROM containing the following:

- Executable file of WEKA modelling software
- Instruction guide of WEKA modelling software
- Full Excel dataset containing all alkaloids and their physical, chemical, and biodiversity dataset
- Excel dataset of the pharmaceutical/non-pharmaceutical data input for ‘full’ entries as used in each predictive modelling algorithm