ENUMERATION, CONFORMATION SAMPLING AND POPULATION OF LIBRARIES OF PEPTIDE MACROCYCLES FOR THE SEARCH OF CHEMOTHERAPEUTIC CARDIO-PROTECTION AGENTS

A thesis submitted in the fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

of

RHODES UNIVERSITY

By

LESTER SIGAUKE

February 2019

Abstract

Peptides are uniquely endowed with features that allow them to perturb previously difficult to drug biomolecular targets. Peptide macrocycles in particular have seen a flurry of recent interest due to their enhanced bioavailability, tunability and specificity. Although these properties make them attractive hit-candidates in early stage drug discovery, knowing which peptides to pursue is non-trivial due to the magnitude of the peptide sequence space. Computational screening approaches show promise in their ability to address the size of this search space but suffer from their inability to accurately interrogate the conformational landscape of peptide macrocycles. We developed an *in-silico* compound enumerator that was tasked with populating a conformationally laden peptide virtual library. This library was then used in the search for cardio-protective agents (that may be administered, reducing tissue damage during reperfusion after ischemia (heart attacks)). Our enumerator successfully generated a library of 15.2 billion compounds, requiring the use of compression algorithms, conformational sampling protocols and management of aggregated compute resources in the context of a local cluster.

In the absence of experimental biophysical data, we performed biased sampling during alchemical molecular dynamics simulations in order to observe cyclophilin-D perturbation by cyclosporine A and its mitochondrial targeted analogue. Reliable intermediate state averaging through a WHAM analysis of the biased dynamic pulling simulations confirmed that the cardio-protective activity of cyclosporine A was due to its mitochondrial targeting. Parallel-tempered solution molecular dynamics in combination with efficient clustering isolated the essential dynamics of a cyclic peptide scaffold. The rapid enumeration of skeletons from these essential dynamics gave rise to a conformation laden virtual library of all the 15.2 Billion

unique cyclic peptides (given the limits on peptide sequence imposed). Analysis of this library showed the exact extent of physicochemical properties covered, relative to the bare scaffold precursor. Molecular docking of a subset of the virtual library against cyclophilin-D showed significant improvements in affinity to the target (relative to cyclosporine A). The conformation laden virtual library, accessed by our methodology, provided derivatives that were able to make many interactions per peptide with the cyclophilin-D target. Machine learning methods showed promise in the training of Support Vector Machines for synthetic feasibility prediction for this library. The synergy between enumeration and conformational sampling greatly improves the performance of this library during virtual screening, even when only a subset is used.

Copyright

I hereby certify that, if appropriate, I have obtained and attached hereto a written permission statement from the owner(s) of any third_party copyrighted material in my thesis, for example visual images.

I undertake to be bound by Rule G 71 of the University's General Rules for Degrees, Diplomas and Certificates:

"If, at the date of the presentation, the thesis has not been published in a manner satisfactory to the Senate, the University shall have the right to make copies of the thesis from time to time, for deposit in other universities or research libraries, make additional copies of it, in whole or part from time to time and distribute the content in whatever format it deems fit, for the purpose of research. The University may for any reason, either at the request of the candidate or on its own initiative, waive its rights".

Signature of Student: ……

Date: ……

Table of Contents

Acknowledgements

For the success of this submission special mentions goes to the following:

Prof. Kevin Lobb (Supervisor),

Thank you so much for being a mentor and a shepherd. Your generosity, availability, persistence and thirst for the TRUTH has made a significant impact on my development as a Scientist. My appetite for learning and humility have been nurtured by your input and for that I am truly grateful. I began my journey as a novice researcher but I leave your research group as an ambitious reformer. I hope to collaborate with your group in the near future.

Members of the Computational Mechanistic Chemistry and Drug Discovery Research Group,

Arthur, Tenda, Faez, Kofoworola, Nonkosi, Kamogelo and Luxolo. I am indebted to your constructive feedback and participation in the discussions that we had during formal group meetings and over a beer at the Rat (Arthur) and coffee (Tenda – before work got hectic).

Members of the Medicinal Chemistry Research Group,

I am indebted to you for assisting me to get to grips with the medicinal chemistry concepts and objectives. I appreciate the criticisms and skepticism. Iron sharpens iron.

Dr Khanye (co-supervisor),

Thank you for your guidance and support. I appreciate the support that you gave my ideas.

Chapter 1: Prelude

Through the lens of George Lucas, “the galaxy far, far away” has dazzled generations of space opera enthusiasts for over 4 decades (Johnson & Brooker 2005). Characters such as Anakin Skywalker in “The Phantom Menace”, Han Solo in “A New Hope” and the Darth Vader fan boy, Kylo Ren, in “The Force Awakens” have all traversed the Star Wars universe in search of the pinnacle of their fulfillment (Schaefer 2015), harnessing or overcoming “the Force”. The “Force” in this universe is an energy field, created by all living things, that surrounds the inhabitants of the estimated 69 million colonies present in the Star Wars universe (Schaefer 2015). It is felt in the billions of stars that extend across a galaxy with a diameter of some 120,000 light years, enthralling us with the tangibility of more (Kapell & Lawrence 2006).

Although the Star Wars galaxy is a fictitious universe (unfortunately), the idea of “a galaxy far, far away” that can be traversed, explored, colonized and defended resonates with imaginators aplenty. By using the phrase an “Intergalactic Computer Network”, Licklider in 1963 was merely describing his latest invention “an electronic commons open to all” which we call the internet today (Leiner et al. 2003; Licklider 1963). He scarcely imagined that his work would contribute to the development of protocols and standards that would extend to an actual “Interplanetary Internet” in the true sense of his description (Bosnor 2000). Work by the Internet Society and affiliated special interest groups, has contributed to the advancement of space communication technology evolving the internet into a truly “Intergalactic” and multiplanetary internet of the internet (Townes et al. 2004; Kraft 2015). Visionaries like Mike Snell and Scott Burleigh, steering the InterPlanetary Networking Special Interest Group aspire to “move forward to an internet that is Interplanetary in scope and function” (Snell et al. 2018).

South Africa's most famous export, Elon Musk, is quoted as saying, "I think fundamentally the future is vastly more exciting and interesting if we're a spacefaring civilization and a multiplanet species than if we're or not" (Clifford 2018). I agree with him. Musk's enthusiasm for the ambitious pursuit to explore the "galaxy far, far away" motivated the genesis of SpaceX. Their greatest achievements to date have been the successful development of the orbital-sized Falcon 9 and the super-heavy Falcon Heavy re-usable lift launchers (Vozoff & Couluris 2008). The successes of the Falcon 9 rockets have since earned SpaceX multi-million dollar contracts with NASA (Mann 2018). The reusability of the Falcon 9 rockets allow NASA to pay less for more spacecraft and satellite launches. Consistent with his aspirations for a multiplanetary human presence, these launchers will propel a Big Falcon Spaceship capable of landing cargo and a team of crewmembers from the Mars100 astronaut selection process on Mars in the near future (Joseph 2010; Karcz et al. 2012).

The excitement surrounding quantum information processing and the advent of exascale computing into our horizon has brought the colonization of a different kind of universe within reach (Wendin 2017; Ananthraj et al. 2018). With these and the ubiquity of aggregated computing, the Chemical Space universe has stoked the imagination of chemists and material scientists (Osolodkin et al. 2015). Our vision is not merely to harness the force, to communicate through this intergalactic network, or to establish multiplanetary settlements. Our ambition is to mine and map this universe probing it for compounds and materials that have unique properties and behaviors (Díaz-Eufracio et al. 2018). These "intragalactic" units will offer solutions to the problems that humanity experiences today and the toys and gadgets that we will enjoy tomorrow.

This thesis will introduce the cost and efficiency challenges of Drug Discovery as well as the high attrition rates of candidates during development. We argue for the role that

computational approaches and peptides play in reducing the burdens experienced during Research and Development. Traversing through peptide space using computational tools helps to identify molecular probes located in “galaxies” or constellations of molecular and constitutional complexity “far, far away” from human intuition (Lau & Dunn 2017). At the center of our voyage through this universe is the enumerator that we developed with the sole purpose of propelling scaffolds and their chemical space into the grasp of rational and systematic interrogation for applications in molecule discovery. It is our hope that our work will be of value to the imaginators that traverse, explore, colonize and defend the chemical space universe.

“May the Force be with you.”

Chapter 2: Introduction and Literature Review

2.1 Context of Drug Discovery

2.1.1 Survey of the Pharmaceutical Industry

The time required for the development of a new molecular entity, to allow it to progress to approval as a drug in therapeutic interventions is between 10 and 15 years (Cheshire 2011; FDA 2018). According to surveys in 2017, investment in Research and Development (R&D) by members of PhARMA accounted for over $75 billion USD (PhRMA 2018). When this value is examined in the context of an economically constrained environment, capital expenditure within healthcare driven by Pharma R&D was the 2nd most expensive globally across all industries behind computing and electronics sectors (Jaruzelski et al. 2015). The appeal of Pharma generally as an investment destination has gradually lost its luster owing to the high costs in development as well as the latency between realizations of the value of investments (Danzon et al. 2005; Carter et al. 2016). The bulk of investment within Pharma in the 90's was sourced from Venture Capitalists, who intended to offset their investments in R&D by the future acquisition of small Biotech companies after their Initial Public Offering (IPO) to larger firms. This had as aim to send market signals from the capture of newly public listings (Reuer et al. 2012; Nicholson et al. 2005). The global recession of 2008 did not only affect the investment practices of the housing market but had an impact on the investment strategies of financial institutions who now rather seek to secure long-term returns on their portfolios by avoiding undue risk (Paul et al. 2010; Alessandri & Pattit 2014). Behavioral theory motivations showed that large pharmaceutical companies (tending to be risk averse) had reduced their propensity for R&D spending and had opted for low-return stock ownership or later post-IPO acquisition (Alessandri & Pattit 2014). The vacuum created by the later activity of large financial institutions and Pharma in the R&D sector gave rise to the overvaluation of

post-IPO products from reputable venture capitalists who benefit from the perceived informational advantage gained from early acquisition of low-quality products and the reputation of previously successful acquisitions (McNamee & Ledley 2015; Lindström & Kekkonen 2018). This overvaluation resulted in an increased competitiveness of Pharma which contributed towards driving R&D investments in neglected therapeutic areas with novel mechanisms of action that had a higher than normal failure rate (Pammolli et al. 2011).

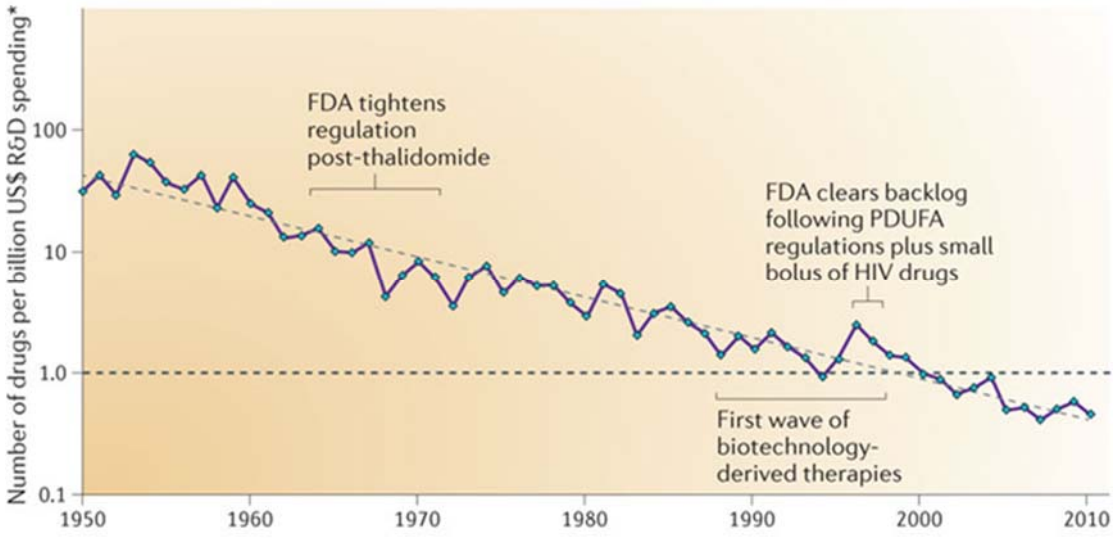


Figure 2.1: The number of drugs approved by the FDA (US) for every $1 Billion (USD, inflation adjusted) spent in R&D*. (Source: Nature Reviews Drug Discovery* (Scannell et al. 2012)*).*

When the inflation-adjusted trend in R&D is considered by looking at the number of drugs that are approved for every $1 Billion (USD) spent (Fig. 2.1) it is unanimously agreed that the overall trend for Pharma efficiency is negative (Scannell et al. 2012). This negative outlook is matched by the fact that despite all the investment in Pharma and its R&D spent, as of 2012 only 20 % of drugs report a positive return on investment compared to R&D costs (Khanna 2012). Of the approximately 4300 companies active in the Pharmaceutical sector over the last 6 decades only 261 of those companies have been able to boast of successfully achieving at least one drug approval (Munos 2009). A list of some of the reasons that contribute to the

high failure rate of the industry include the high cost of research, the shift towards rare disease cases, intensified regulation, globalization of research and the price-control of drug products (Khanna 2012). Concerning political implications, innovation within Policy Reform has the potential to ensure the protection of intellectual property across nations resulting in companies with marketed therapies receiving sufficient time to re-coup their investment. Decision making during the development stage may also include evidence of manufacturing implications, to ensure that a therapy intervention will be cost efficient in the market. In terms of R&D throughput, large companies (such as AstraZeneca) that are able to overcome the low productivity associated with the R&D phases have a significant chance of surviving within the pharma sector (Cook et al. 2014). Strategies to reduce deficiencies within R&D range from the optimization of the pre-clinical discovery stage, to focus on the development stage of progressing new molecular entities into marketed therapies (Orloff et al. 2009; Hay et al. 2014).

2.1.2 Drug Discovery Pipeline

The process followed to bring a drug to the market is a costly and time intensive process that requires an initial discovery and development stage (including multiple iterative sub-stages) which is followed by the rigorous two phases of assessment and then approval (Scannell et al. 2012). Strict checks are triggered at each of these three stages and only molecular entities that are deemed to have potential to succeed at future stages are graduated in order to maintain sustainable resource management practices (Rishton 2003). Each of the three progressive stages have different priorities that are emphasized and included for the classification of molecular entities as being viable for continued interest (Archer 2004). Not only are the molecular entities assessed in order to decrease the probability of failure in

subsequent phases, but the assessment from previous stages gives valuable insight into future improvement of the total efficiency through all stages (Meanwell 2016).

Research within the early discovery and development phase prioritizes “gaining insights into disease processes” in order to gain understanding of what the mechanisms of disease progression or treatment failure are (Archer 2004; Dowty et al. 2014). During this phase innovation aims to stretch the boundaries of what is known in order to identify new approaches that can be taken in order to either halt, prevent or reverse the effects of the condition (Brahmachari 2012). There are diverse approaches towards disease intervention which range from controlling the vector in the case of communicable diseases; perturbing the causative agent in the body in the case when the disease is caused by an infection; or probing the patient’s systems in order to make them resistant to manipulation by the disease causing agent (Csermely et al. 2013). In other interventions, the focus can shift to the reduction in the exposure of the disease-causing agent within the environment in the case of certain toxins. A robust understanding of the disease’s processes and it’s mechanisms of progression is important in order to inform researchers active within the discovery and development stage (Perrelli et al. 2011). In the case of chemotherapeutic interventions that rely on halting disease progression within the patient it is of paramount importance that key “druggable” targets are identified either within the disease causing agent in the case of an infection or within the patient (Rask-Andersen et al. 2014). The process of identifying a molecular entity that perturbs a particular druggable target is called Drug Design and relies on a strategic plan that aims to identify potent molecules, “Hits”, that are later optimized into “Leads” that are further optimized to potent and non-toxic developmental candidates (Hughes et al. 2011). Hits are those molecules that have been shown to possess molecular features that enable them to interact with the biomolecular target or hypothesized network with an affinity and

specificity that translates into an unambiguous physiological response in a reasonable time (Tarcsay & Keseru 2015; Csermely et al. 2013). The essential features of a Hit, critical for its hypothesized therapeutic response, are critically examined by the design, synthesis and testing of analogues of the Hit, allowing for the generation of models that relate structures and features to a particular response or activity (Mignani et al. 2016; Lam et al. 1991; Ghasemi et al. 2018). This Structure Activity Relationship enables the chaperoning of a Hit molecule and its analogues through a heuristic approach to physicochemical property optimization which results in compound clusters derived from the Hit precursor (He et al. 2016; Shultz 2013). Analogues derived during this Hit expansion stage, allow the discovery team to increase the probabilities of successfully identifying a Lead candidate from the Hit precursor (Rishton 2003; Dias & Ciulli 2014). Traits that determine the lead-likeness of Hit analogues are: high affinity towards the target, selectivity, efficacy, drug-likeness, low serum albumin affinity, low interference with P450 enzymes and P-glycoproteins, low cytotoxicity, metabolic stability, cell membrane permeability, water solubility, chemical stability, synthesizability, and patentability (Hann & Oprea 2004; Schneider 2013; Lagorce et al. 2011).

Any discussion around innovations that affect the efficiency of R&D within Pharma would be incomplete without an examination of the pre-clinical drug development stage, which accounts for an average of $164.7 Million USD and 67 **%** of failures (Morgan et al. 2011; DiMasi & Grabowski 2007). The goal of the pre-clinical development stages are to determine the pharmacokinetics and metabolism of lead candidates in order to ensure human safety and reduced toxicity (Law et al. 2014; Testa et al. 2000). In order to ensure that the Lead candidates identified during the discovery stages progress through this pre-clinical stage they undergo Lead-optimizations (LO) (Hughes et al. 2011; He et al. 2016). While the Hit identification stage focusses on drug-target analyses through biophysical examination and *in*

vitro assays that determine dose response curves and efficacy, the LO stage emphasizes *in vivo* pharmacokinetics (Johnson et al. 2009; Dowty et al. 2014). Modifications of the Lead candidates in preparation for pre-clinical development at this stage include chemical modifications of the lead candidates in order to optimize dosage and administration strategies (Dube et al. 2012; Ma et al. 2017). Each stage of the optimization cycle involves a delicate balance between enhancing the molecular properties of the Hit and not compromising on the features that contribute to its molecular potency and specificity (Wang et al. 2017).

Prior to Clinical trials of the Lead series little is known about the human safety, toxicity, pharmacokinetics and metabolism of the matured Leads (Hughes et al. 2011). The goal of the last phase of the pre-clinical stage in the Drug Discovery pipeline, is to identify clinical candidates that demonstrate reasonable safety within humans (Smith 2011; Hann & Oprea 2004). The clinical candidate undergoes an Investigational New Drug (IND) application process during this stage before a decision on its further testing to determine effectiveness in human patients during Clinical trial phases (Hay et al. 2014; Hughes et al. 2011). The IND application reviews the results of the Pre-Clinical stage that relies on testing in appropriate animal models to inform *in vivo* pharmacodynamics, pharmacokinetics, ADME and toxicology profiles of the Leads (Smith 2011). Success rates at this stage may be as low as 1 in 5,000 depending on the target class and the proposed therapy and protocols recommended (Nicolaou 2014; Bonabeau et al. 2008). Once the IND application is approved, the strategic plan culminates in the Clinical candidates becoming drug that is allowed to undergo testing in humans. At this stage the drugs are sent through a progressive series of Clinical trials that begin with testing for safety and dosage on a few healthy humans during Phase 1 studies and unhealthy humans in Phase 2 studies (Hughes et al. 2011; Chen et al. 2012). These tests are typically no more

than 2 years long with unacceptable toxicities within humans the main reason that drugs are eliminated from further testing at this stage (Medina-Franco et al. 2013; Leeson 2012). Testing for safety and efficacy in thousands of patients occurs over Phase 3 and 4 studies which may last up to 7 years depending on the medical condition (Carter et al. 2016; Orloff et al. 2009). Drugs in this phase of the IND process are eliminated on the grounds of inconclusive intended efficacy in human patients monitored throughout the clinical trials on patients that had the targeted disease (Kola & Landis 2004; Waring et al. 2015). Owing to the high attrition rates and cost associated with development, companies may opt to explore drugs that reach this stage for their potential as Leads for other therapies (Kinnings et al. 2011; Wang et al. 2013). Depending on the reasons for failure, modifications to the clinical candidates may be proposed in order to enhance their bioavailability and pharmacodynamic profile (Dube et al. 2012). The results from drugs that pass clinical trials are compiled into a report that is submitted by the company for marketing approval in the countries that they intend to distribute the drug or therapy (FDA 2018; Teague 2011; Hay et al. 2014). The approval bodies have the responsibility of granting marketing approval, with recommendations from a panel that consists of medical officers who review clinical study information and statisticians who review the protocols by evaluating the trial data and designs (Orloff et al. 2009; FDA 2018). Experts from pharmacology, pharmacokinetics and chemistry on the panel review data associated with the drug's properties and behavior which impact its administration strategy and quality control (Hughes et al. 2011; Mignani et al. 2016).

Although the process of successfully developing a drug and approving it for market has huge benefits both for patients who experience an improved quality of life, and pharmaceutical companies who experience the substantial dividends from sales, this process is riddled with casualties (Sams-Dodd 2013; Waring et al. 2015). There are extreme costs incurred in the

process and despite the increased productivity in drug discovery research the costs and productivity continue to rise at disproportionate rates (Grabowski et al. 2002; Khanna 2012). Innovations within the design of delivery systems and formulations have been sought as strategies to decrease the attrition rate of potent molecule candidates. Life-cycle design strategies that allow for earlier pharmacokinetic, pharmacodynamics and pharmacogenomics evaluations have been shown to reduce the rate of attrition during clinical trials (Cheshire 2011; Schneider 2013; Meanwell 2016). A bottleneck within the chemotherapeutic drug discovery process has been the relationship between the searchable space for chemical moieties and the targets for which they probe (Shan et al. 2011; Rask-Andersen et al. 2011; Y. Y. Li et al. 2011). Innovations within the realms of molecular modelling that enable the examination of molecular drug targets through integration of proteomics and bioinformatics tools have the power to cost effectively enhance the efficiency of the early Hit discovery stage (Kortagere & Ekins 2010; Sliwoski et al. 2014; Gilad et al. 2015). These innovations allow for the discovery of tunable target specific drugs to be designed contributing to enhanced compatibility with therapeutically relevant drug targets (Blaney 2012; T. Cheng et al. 2012; Khanna 2012; Abel, Mondal, et al. 2017). When insights derived from these innovations within the life-sciences are coupled with significant advances in computing the searchable chemical space landscape available to medicinal chemists can be substantially expanded (Medina-Franco et al. 2014; Dandapani & Marcaurelle 2010). This has led to the ideas surrounding what is thought to be a therapeutically relevant drug target to be re-considered (Shan et al. 2011; Doak et al. 2015).

2.1.3 Chemical Space

New drugs are typically discovered through testing of many compounds in order to identify probable hits that will then be graduated into the leads that can progress to clinical candidates

within the Drug discovery pipeline (Hughes et al. 2011). The search for new drugs with novel means of perturbation and scaffolds that interact with the drug target are identified from the chemical space universe (Leeson 2012; Díaz-Eufracio et al. 2018; Von Lilienfeld 2013). Depending on the resources available, medicinal chemists have the option to bias or restrict their search space due to the magnitude of the chemical space universe from which this search for Hits is performed (Rishton 2008; Lachance et al. 2012; McHugh, Rogers, Solomon, et al. 2016). A short discussion follows in order to emphasise the scale of the chemical space problem, this difficult problem may be described as "searching the chemical space universe for Hits with optimized properties for a specific clinical case".

The chemical space available to compounds is described as the potential of chemical realization that is present within the rules of material arrangement (Dobson 2004). By analogy, if one had an infinite amount of ink, paper, space and time, it would be possible for all the combinations of the 26 letters of the alphabet to be stored in a library of words of single letter, two letter, three letter up to 6 or 7 letter words. This large library exists, albeit conceptually, as the hypothetical "Library of Babel". A similar conceptual library of chemical compounds exists although no-one has claimed its ownership (Creighton 2015). Although in practice we do not discuss a "Chemical Molecules Library", we do describe this vastness of potential as the "chemical space universe". Von Lilienfeld (2013) defined the chemical compound space conceptually as, "some continuous observable space that is populated by all experimentally and theoretically possible chemicals with natural nuclear charges and real interatomic distances for which chemical interactions occur" (Von Lilienfeld 2013). Of relevance to those working in the field of drug discovery is the portion of the chemical space universe containing bioactive molecules (López-Vallejo et al. 2012). An estimate of the magnitude of this bioactive molecular chemical space (that also obeys the rules of chemical

bond formation and stability, and is restricted to a reasonable molecular size) is estimated by Lowe to be within the order of magnitude of 10^{60} compounds. This is an incomprehendable number of compounds (Lowe 2015).

In order for researchers to benefit from the sheer number of available compounds, the search for molecular leads (with new properties for applications in drug discovery) requires the researcher not to be overwhelmed but rather inspired by the potential of chemical space universe. This potential inspired the population of the GDB17 database by the Louis Reymond group (Ruddigkeit et al. 2012). Their database of 166 Billion compounds is derived from 17 membered combinations of heteroatoms, which was constructed through the use of graph theoretical methods to ensure a complete search of the available chemical space. The vastness of the GDB17 reduces the potential for missing active agents.

A broader discussion, however, is necessary to understand what chemical space is, its purpose and how it can be searched (Polishchuk et al. 2013). A more conservative definition of chemical space which relates pre-existing and accessible compounds views chemical space as a trait obtained from the combination of descriptors obtained from numerical abstractions or functional labels (Osolodkin et al. 2015). Instead of viewing the chemical space universe as something out there, it can rather be viewed as something that is already present and can be expanded into from what it currently is (Dobson 2004). A different analogy is relevant in the description of this view. For instance, one can gaze out the window and observe the landscape that exists (and upon which the discipline of geography depends). In a similar manner, rather than recreating chemical space in an exhaustive manner, another paradigm for chemical space is the recreating or re-imagining of compounds that are already available. In other words, “to probe the chemical space universe for Hits with optimized properties for a specific clinical case”.

2.2 Databases

The starting points for the identification of candidate hits are typically high-throughput screening assays of compound collections or the screening of libraries using assay platforms that are compatible with the targets of interest (Pereira & Williams 2007). These optimized assays take advantage of advances in miniaturization and automation, enabling rapid liquid handling and reliable activity detection to achieve fast turnaround times (Peakman et al. 2003; Archer 2004). Innovations in data analysis allow for the reliable examination of statistical profiles of compounds, enabling the identification of candidate hits for further secondary screening and development from the compound collections (Shterev et al. 2018; Malo et al. 2006).

Given the extent of possible chemical space (as before estimated at 10^{60} possible compounds)(Dobson 2004; Lowe 2015), exploring even a small fraction of this space using these methods appears to be a futile effort (Lowe 2015). In order to perform rapid exploration of chemical space, chemical vendors store large compound databases that are populated and screened to identify novel scaffolds with insightful chemistries. An example of this approach was when the 20 000 DIVERSet ChemBridge Corporation library of molecules optimized for diverse pharmacophore coverage and druglikeness, was tested successfully for candidate cardio-protection compounds that were able to reverse the inhibition of sarcoplasmic reticulum Ca-ATPase (SERCA) (Cornea et al. 2013). Databases exist that contain both reported compounds and compounds that may physically be obtained. Examples include the Chemical Abstract Database, the ZINC database and the PubChem database, who populate their databases from diverse sources such as from literature, from natural product collections or from reports of combinatorial approaches (Liu et al. 2017; Freeland et al. 1979; Irwin & Shoichet 2005; Edwards & Ericsson 1998). Coupling combinatorial approaches with

natural product screening libraries has allowed the emergence of physical chemical libraries composed of complex compounds that possess well-balanced physicochemical properties efficiently placed within the medicinally relevant chemical space of natural products (Leach & Hann 2011; López-Vallejo et al. 2012; Medina-Franco et al. 2014). By focusing on quality and accessibility of compound stocks, ChemBridge Corporation has developed a suite of screening libraries biased towards either diversity, targets specificity, macrocyclic scaffolds, fragments or the repurposing of Hits (Gilad et al. 2015).

There is not necessarily an upsurge in productivity of research groups in the identification of first in-class lead candidates, despite the astronomical surge in HTS capabilities over the last decade, and the number of citations in peer-reviewed scientific publications (Peakman et al. 2003; Sams-Dodd 2013; Meanwell 2016). The reasons for this pharmaceutical innovation gap, reflected by low production of in-class lead candidates, can be attributed to the way in which screening strategies have been made use of in discovery pipelines. The discovery paradigm over the last few decades has been dominated by target-based approaches that rely on programmes that ignore the complexity and variability of the entire organism but rather focus on knowledge of the target, the mode-of-action and the screening cascade (Sams-Dodd 2013). Within this target based discovery paradigm, optimizing the screening cascade has been seen as the silver bullet for achieving rapid success. It has also been thought that the drive for screening pipelines optimized for speed (achieved by reducing cycle times), has contributed to the reduced R&D productivity due to the development speed paradox (Lendrem & Lendrem 2014). By focusing on high-throughput and performing screens in parallel, the time to collate knowledge from failed tasks and screens is increased and the value of incremental sequential decision-making is lost. The rapid screens have also neglected tests of compounds under clinical conditions resulting in the lack of correlation between target

specific screening data and clinical development (Sams-Dodd 2013). The limitations of the promise of HTS and its lack of success, highlights a greater need for collaboration between process-focused innovations and disease-focused innovations. This collaboration may well be what is needed to improve target-based approaches reliant on HTS investment (Sams-Dodd 2013; Tajabadi et al. 2013).

Optimizing screening cascades towards speed (through automation) has an impact on the nature of the compounds and thus the chemistries exploited by hit candidates that emerge from HTS results (Lendrem & Lendrem 2014; Suryanarayana Birudukota et al. 2016). By biasing screening libraries towards compounds accessed via parallel synthesis, any hits that emerge from HTS tend to have high degrees of unsaturation and low degrees of complexity (Feher & Schmidt 2003). Although compounds accessed by this diversity oriented parallel synthesis are synthetically accessible, their pseudo-planar geometries severely limit their and their derivatives diversity in terms of chemical space (Lovering et al. 2009). High degrees of unsaturation contribute towards the rigidity of molecular scaffolds and thus their ability to match the rough topology of their targets of interest is reduced (Stockdale & Williams 2015). The decreased molecular complexity of small molecules has also been show to impact their physicochemical properties (Feher & Schmidt 2003; Meanwell 2016). Although the increased aromaticity of unsaturated species increases pi-pi interactions that may interact with relevant molecular targets, these conversely contribute to self-association reflected by increased melting points and poorer solubility profiles (Lovering et al. 2009). Compounds that are higher in complexity, saturation and number of chiral centres have been shown to possess higher probabilities of success in clinical stages of the development cycle and lower promiscuity (Lovering et al. 2009; Lovering 2013; Stockdale & Williams 2015). By transforming phenyl scaffolds into saturated dienediols using recombinant deoxygenation bacteria for [4 + 2]

cycloaddition reactions it was possible to test the “escape from flatland” hypothesis by activating chemical inert scaffolds by increasing their dimensionality (Suryanarayana Birudukota et al. 2016).

2.2.1 Natural product databases

This “escape from flatland” hypothesis is confirmed by the dominance within the approved small molecule drug space by natural product-derived drugs (Butler & Buss 2006; Harvey et al. 2015; Katz & Baltz 2016). Natural product-based screening libraries possess rich diversity and complexity as defined by stereochemistry, the degree of saturation and the presence of sp^3 centres (Harvey et al. 2015; Lovering et al. 2009). These features contribute to their chemical richness enabling them to be the “single most productive source of leads for the development of drugs” (Harvey 2008). This biologically relevant complexity and scaffold novelty exists because natural products have co-evolved with their biological receptors (Clemons et al. 2010; Thomas & Johannes 2011; Harvey et al. 2015). This has been in a manner that has allowed for the selection of the biosynthetic pathways that lead to the synthesis of specific natural products which possess the dimensionality, chirality and bioavailability required to complement the biological target (Nicolaou 2014; Clemons et al. 2010). Despite being privileged with such traits the pharmaceutical industry has decreased dependence on Natural-product derived screening libraries due to the increased turnaround times of pipelines that rely on natural-product screens (Harvey et al. 2015). Difficulties with product supply, structure elucidation of hits, accessibility of derivatives and the concerns revolving around the security of intellectual property have also contributed to the dominance of pseudo-planar compounds with simpler synthetic routes and chemistries within HTS (Rishton 2008; Katz & Baltz 2016; McChesney et al. 2007; Harvey et al. 2015).

2.2.2 Peptide databases

Although significant progress has been made towards the construction of highly active and potent secondary metabolite gene clusters that lead to natural products, the costs associated with the scale up of these processes and their development is still to be overcome (Katz & Baltz 2016). Despite these limitations, compounds derived from scaffolds that possess natural product-like features have allowed a new way of thinking about the "escape from flatland" conundrum within a target driven HTS paradigm (Lachance et al. 2012). Despite possessing low bioavailability profiles due to their larger size compared to small molecule drugs, peptide-derived therapeutics possess high specificity for their biological target and high stability within the biological environment (Uhlig et al. 2014). These properties correlate well with their safety and tolerability, reflected by their higher likelihood of success in clinical trials. This success is in comparison to therapies based on new biological entities (such as antibodies) and new chemical entities (small molecule drugs) (Lau & Dunn 2017). Naturally occurring peptides suffer from low membrane permeability and short half-life as a result of detection by host cell processes that also contribute to their rapid hydrolysis and predictable metabolism (Lau & Dunn 2017; Wu & Hancock 1999; Chatterjee et al. 2008).

Enthusiasm surrounding the therapeutic applications of naturally occurring peptides has grown due to advancements in proteomics and medicinal chemistry. These advancements reduce the suboptimal chemical and physical properties of naturally occurring peptides (Ma et al. 2017; Ahlbach et al. 2015). Proteomic advancements allow for the rational design of peptide therapeutics by translating knowledge obtained from the 3D structure of the natural peptide or protein effector into an effective ligand (Singh et al. 2015). The translation strategy involves the construction of focused libraries of peptides obtained through identification of the essential amino acids that perturb the receptor (identification may be through alanine

scanning, for example) (Chatterjee et al. 2008). Although recombinant techniques are routinely used to produce proteinaceous molecules, the total synthesis of targeted peptides with therapeutic potential has benefitted greatly from advances made in solid-phase peptide synthesis, in particular the use of Fmoc chemistry (Merrifield 1963; Fields & Noble 1990; Behrendt et al. 2016). For proteins too large to be accessible *via* solid-phase synthesis, native chemical ligation is a technique that allows for successful total synthesis. Challenging protein targets are therefore also accessible within focused libraries (Kulkarni et al. 2018a). Screening of the focused library allows for the identification of chemically important amino acids through the construction of structure activity relationships. The non-essential amino acids are marked as sites for modification, for the optimization of the peptides both for ease of synthesis and for bioavailability (Fosgerau & Hoffmann 2015). Medicinal chemistry strategies such as peptide acylation, and N-methylation, structure modifications such as the introduction of stabilizing α-helices and cyclization as well as the application of solubilizing agents such as cyclodextrin are routinely used to convert highly active peptides into drugs (Chatterjee et al. 2008; Craik et al. 2013; Lau & Dunn 2017). Growth in the peptide therapeutic landscape has been driven by advancements in rational design strategy which had contributed by 2017 to 60 FDA approved drugs, and over 150 peptide drugs active in clinical development (Lau & Dunn 2017). The design strategy employed to access therapeutic peptides is contingent on the advancements of peptide synthesis protocols.

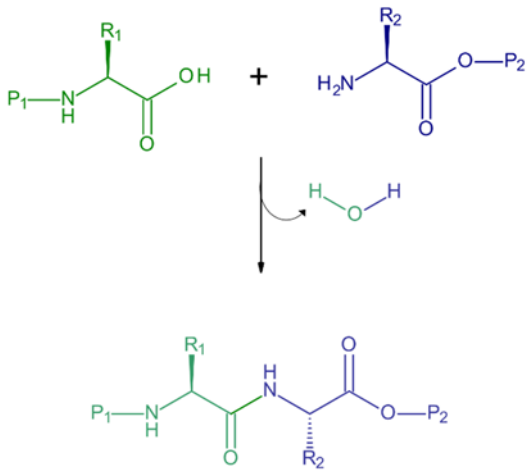


Figure 2.2: Formation of a peptide bond between two amino acids (R1, R2) (P = protecting group).

The synthesis of a peptide requires the chemically controlled formation of peptide bond through the condensation of a reactive carboxyl group on one amino acid with the amino group on a second amino acid (Fig. 2.2). The methods used for the targeted synthesis of specific therapeutic peptides can be divided into two main categories based on the dominant techniques used to access them: Enzymatic synthesis and chemical synthesis.

2.2.2.1 Enzymatic synthesis strategies

Advances in recombinant DNA technology and artificial synthetic DNA translation, used to facilitate molecular cloning and the expression of protein targets of therapeutic interest, has contributed towards the design and synthesis of therapeutic peptides and proteins (Gräslund et al. 2008; Satyawali et al. 2017). Protein synthetic approaches that aim to produce biologically active peptides may do so by making use of fluoroalkyl substituted amino acids, organofullerene-based frameworks or *para*-substitution of phenylalanine for example. Further, the application of peptide bond modification strategies such as the introduction of sulfonamides have contributed towards their improved bioavailability and metabolic stability

(Thust & Koksch 2003; Choi et al. 2003; Obreza & Gobec 2004; Minois et al. 2017; Han et al. 2013). Recombinant techniques that rely on ribosomal or non-ribosomal protein expression systems suffer from limitations associated with the limited ability to incorporate non-natural amino acids into the growing protein sequence (Josephson et al. 2005; Smolskaya et al. 2013). The purification, the identification of expression systems, the inability to accommodate non-natural amino acids and the low expression profiles of larger peptides is an issue. This limits the application of recombinant techniques to these systems (Kumar & Bhalla 2005; Tripathi 2016).

Peptide synthesis may make use of proteolytic enzymes to catalyze the formation of the peptide bond by thermodynamic manipulation of the reversible proteolysis (Fig. 2.3) (Bordusa 2002; Luethi & Luisi 1984). Parameters such as temperature, pH, substrate concentrations and the enzyme concentration affect the kinetic control of the proteolytic reaction while the incorporation of co-solvents are strategies that adjust the equilibrium state of the reaction (Bordusa 2002). Proteolysis enzymes such as papain, proteinase-K and trypsin are used for chemical synthesis of short peptide sequences owing to their high stereoselectivity, regio-selectivity and broad pH range stability (Ageitos et al. 2013; Viswanathan et al. 2010; Li et al. 2008).

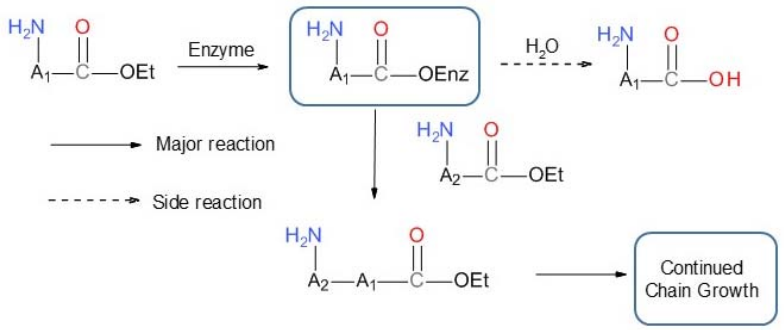


Figure 2.3: Peptide chain propagation catalysed by proteolytic enzymes.

Manipulations such as the incorporation of benzyl ester groups on alanine and glycine substrates have been shown to contribute towards the improved degree of polymerization and papain substrate specificity. Side-chain protection using *tert*-butylcarbonyl groups on lysine substrates alters kinetic behaviour and binding constants (Ageitos et al. 2016; Qin et al. 2014). Medium engineering (either homogeneous or heterogeneous), is used to manipulate the solutions driving the reactions towards peptide synthesis (Bordusa 2002; Guzmán et al. 2007). In homogenous systems, water and miscible organic solvent mixtures speedup reactions by: improving the solubility of non-polar substrates; shifting the thermodynamic equilibrium towards the products; enhancing the enantioselectivity; reducing microbial contamination; and stabilizing enzyme three-dimensional structure critical to its function (Torres & Castro 2004; de Gómez-Puyou & Gómez-Puyou 1998). Although inclusion of miscible organic solvents in exceptional circumstances is able to enhance enzymatic peptide synthesis, organic solvents are surfactants and tend to destabilize the enzymes in terms of catalytic throughput and thermostability (Quezada et al. 2017). Heterogeneous mediums, through the use of immiscible liquid phases, allow for enhanced availability of low solubility substrates, while maintaining an improved enzyme stability (Bordusa 2002). Macro-

heterogeneous systems make use of biphasic mixtures that allow for the catalysis reaction to occur within the protein-stable aqueous layer while unwanted hydrolysis is reduced by partitioning the peptide product to the organic layer (Barberis et al. 2006). An alternative heterogeneous strategy is the micro-heterogeneity approach, which stabilizes the enzyme activity during its immersion in a hydrophobic solvent. Micro-heterogeneous systems arise when an aqueous layer surrounding the enzyme is maintained by the encapsulation of the enzyme in either a denatured enzyme, the incorporation of water mimics such as ethylene glycol or the application of lyoprotecting agents such as crown ether derivatives (Tielmann et al. 2014; van Unen et al. 2002). The encapsulation of a proteolytic enzyme into a micro-heterogeneous environment has paved the way for approaches that make use of immobilization techniques; these improve the reusability and stability of the enzymes (Kumar & Bhalla 2005). Strategies of enzyme immobilization include making use of covalent fixing onto a solid matrix support or the use of a polymeric gel for entrapment (Kise & Hayakawa 1991; Bacheva et al. 2003).

The inability of enzymatic synthetic strategies to access larger peptides is a problem. Recombinant techniques that access larger peptides can be coupled with enzymatic approaches. This allows for the design and synthesis of novel peptides inaccessible to recombinant expression systems (Wallace 1995; Bordusa 2002). Chemoselective strategies such as the native chemical ligation, bis(2-sulfanylethyl)amido (SEA) ligation and the α-ketoacid-hydroxylamine (KAHA) amide forming ligation approaches have been used to extend the applicability of peptide synthesis, allowing for cyclization and condensation of unprotected fragments (Kulkarni et al. 2018b; Hou et al. 2011; Rohrbacher et al. 2015).

2.2.2.2 Chemical synthesis of peptides

Although the enzymatic approaches to the synthesis of peptides have experienced significant advancements, their incorporation in design strategies for therapeutic peptides is discouraged owing to the limited number of sequences and amino acids that can be incorporated into the growing peptide sequence (Kumar & Bhalla 2005; Satyawali et al. 2017). Enzyme-free methods that make use of matrix supports for the successive addition and systematic elongation of anchored peptide chains have experienced renewed attention owing to the maturity of coupling and protecting chemistries (Johnson et al. 2001; Guzmán et al. 2007). The first step of stepwise solid phase peptide synthesis (SPPS) is the anchoring of a protected amino acid onto the resin through covalent attachment (Fig. 2.4 A) (Merrifield 1963). The acid labile terminal protecting group is then removed in preparation for condensation with the next protected and activated incoming amino acid (Mende & Seitz 2011). The activated condensation and deprotection is repeated in a stepwise manner with thorough washing of the reaction vessel in order to avoid "translation" errors (Kulkarni et al. 2018a). The protecting group, orientation of synthesis (N to C, or C to N terminal), the coupling agents and the cleavage method impact the choice of the linker group on the resin, and the resin solid support for successful SPPS (Behrendt et al. 2016). In the case where an amide terminal peptide is desired, stabilization of the peptide is accomplished through the use of an amide forming resin such as the Rink amide resin or the Pal amide resin. Neutralization of C-terminal peptides is achieved through the use of acylating capping agents after cleavage from non-amide forming resins such as trityl chloride resins or the Wang resin (Tegge et al. 2007).

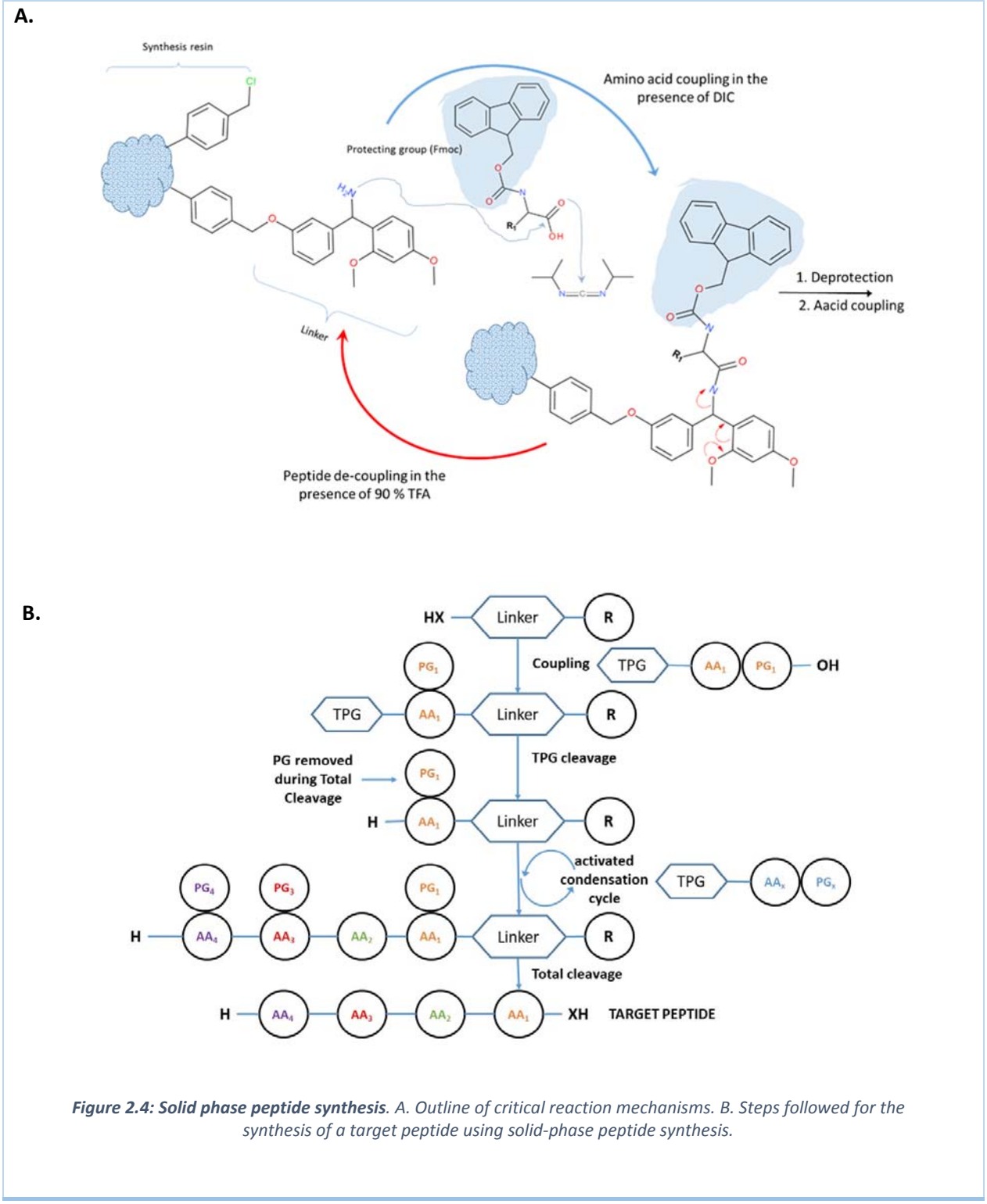


***Figure 2.4: Solid phase peptide synthesis**. A. Outline of critical reaction mechanisms. B. Steps followed for the synthesis of a target peptide using solid-phase peptide synthesis.*

In order to achieve successful elongation, base labile, temporary protecting groups (TPG) are used in SPPS to prevent self-condensation (Fig. 2.4 B). Acid labile, permanent protecting groups (PG) prevent side-chain groups from being incorporated into the growing peptide

backbone (Mäde et al. 2014). Side chain amino and guanidine groups are protected using Boc and Pbf protecting groups while acid, hydroxyl or sulfhydryl side chain groups are protected by Trt and OtBu acid labile groups (Fields & Noble 1990). During synthesis the amino group of the incoming acid is protected temporarily by a base labile Fmoc group that is deprotected in preparation for the next coupling reaction (Mende & Seitz 2011). Use of Fmoc protection was shown to require milder conditions of deprotection, protecting the peptide from acid damage when strong acids are used for deprotection (Mäde et al. 2014). Amino acid activation in Fmoc SPPS is achieved through the use of coupling agents such as anhydrides, active esters, acylazoles, acid halides, phosphonium salts or uronium salts (Albericio 2004; El-Faham & Albericio 2011). The coupling agents are used in activation because they form an intermediate that has an electron-withdrawing group introduced onto the carboxyl terminal of the incoming amino acid. This significantly enhances its electrophilicity and eventual attack by the amino nucleophile of the bound amino acid (El-Faham & Albericio 2011). The different types of coupling agents are also chosen in order to minimize the likelihood of side reactions that may give rise to racemization and loss of stereochemistry.

2.3 Probing the chemical space universe using peptide databases

Databases that correlate sequence and structure of peptides with their therapeutic potential have contributed significantly to the use of peptides and peptide-derived drugs in virtual screening protocols (Waghu et al. 2014; Qureshi et al. 2013; Kumar et al. 2015; Kapoor et al. 2012; Singh et al. 2015). The design and population of databases with peptides that possess therapeutic potential is through the use of either template-based methods, biophysical approaches or through the results of virtual screening (Fjell et al. 2012). Template-based methods rely on prior knowledge of peptide sequences in order to build potent derivatives. Strategies may make use of systematic mutation studies, the use of linguistic approaches for

pattern recognition or varying the length of sequence repeats. Strategies may also include the derivation of a generic template, based on residue properties or on the functionalizing of unmodified potent peptides to enhance their pharmacokinetic and pharmacodynamic profiles (Loose et al. 2006; Deslouches et al. 2005; Frecer et al. 2004; Jiang et al. 2011). Guided by observations that the shape and plasticity of proteins influence their classification (enabling better correlation with function), the design of peptides can be accomplished from biophysical based methods that examine their structures (Li & Koehl 2014; Fjell et al. 2012). Peptide design strategies that make use of biophysical methods depend on reliable structure determination of derivatives through the use of combinations of X-ray crystallography, nuclear magnetic resonance spectroscopy, circular dichroism, fluorescence spectroscopy, quartz crystal microbalance and molecular dynamics simulations (Rozek et al. 2003; Wu & Hancock 1999; McHugh, Rogers, Solomon, et al. 2016; Vasile et al. 2016; Li & Koehl 2014; Bürck et al. 2016; Tsai et al. 2009). Peptide design methods based on virtual screening strategies exploit computational approaches to expand and interrogate the search space occupied by peptides (Fjell et al. 2012). Virtual screening strategies typically leverage iterative molecular design theory to reduce experimental expenditure. They depend on the derivation of a fitness score that correlates some measure of biological activity (from peptide sequence) and structure data (Schneider et al. 2009; Tsai et al. 2011). Some examples of algorithms used in virtual screening design are evolutionary algorithms, random sequence generation, genetic algorithms, ant colony optimization and backbone template exploration algorithms (Schneider et al. 1994; Klepeis et al. 2004; Yagi et al. 2007; Hiss et al. 2007). Fitness scores that can be coupled to these algorithms are diverse and include fold stability evaluation, alignment with an experimentally optimized matrix or a Hidden Markov model. Fitness scores

may also involve maximizing an artificial neural network or the optimization of an interaction energy potential (Fjell et al. 2012).

The rapid searching of peptide databases improves the efficiency of rational therapeutic peptide design and reduces the time-gap between successive iterations during the molecular design process (Katz & Baltz 2016; Fosgerau & Hoffmann 2015; Ma et al. 2017). Rational design strategies of peptides may also include strategies that couple synthetic feasibility with virtual library generation. This will generate focused libraries that better identify lead candidates for clinical development (Lau & Dunn 2017). The following sections outline advancements made in virtual screening approaches and the different ways in which they are deployed in molecular design. These advancements allow for the extraction of design parameters for the construction of focused libraries for the rational design of potent and specific peptide therapeutics.

Although the goal of molecular modelling is to mimic molecular behavior, in the context of drug design, virtual screening aims to assist in a decision making process (Levitt & Warshel 1975; Warshel 1976; Fjell et al. 2012). Within the design stages of a discovery strategy, virtual screening allows for the rapid formation and testing of hypotheses' that lead to a greater understanding of the optimization space (Loughney et al. 2011; Kuhn et al. 2016). Virtual screening methods used in molecular design can be divided into structure based methods or ligand-based methods (T. Cheng et al. 2012; Anighoro & Rgen Bajorath 2016; Ripphausen et al. 2011). A pre-requisite in structure based methods is the availability of accurate and reliable structural data, such as the three-dimensional coordinates of targets. Ligand based methods rely on biological activity, such as affinity or inhibition by known ligands to derive a fitness assessment of compounds within a database (Raman et al. 2017; Shim et al. 2011; Bickerton et al. 2012).

2.3.1 Structure based methods

In drug-discovery, the fitness score for competitive inhibition is an interpretation of the association constant of the protein (P) and ligand (L) at equilibrium in solution:

$$P + L \leftrightarrow PL \ \therefore \ K_a = \frac{[PL]_{eq}}{[P]_{eq}\,[L]_{eq}} \quad (1)$$

Structure based methods derive an interpretation of K_a by computing the P-L free-energy of binding, $\Delta {G_b}^o$, using thermodynamic properties of the molecular-target interaction (Cross et al. 2009). If $C^o$is the constant that defines concentration of interacting partners, T, is the constant temperature and k_B is the Boltzmann factor, then:

$$\Delta {G_b}^o = -k_B T ln(K_a C^o) = k_B T ln\left(\frac{K_b}{C^o}\right) \quad (2)$$

If a rare protein – ligand binding event is described using the canonical ensemble's probability of transition from a low-energy microstate to a high-energy microstate, the free-energy change of binding, $\Delta {G_b}^o$, can be expressed in terms of probabilities, p, instead of equilibrium constants:

$$\Delta {G_b}^\circ = -k_B T \ln\left(\frac{p_{bound}(\boldsymbol{x})}{p_{unbound}(\boldsymbol{x})}\right) + k_B T \ln\left(\frac{C^{box}}{C^o}\right) \quad (3)$$

In structure based methods, ΔG_b, the free-energy change, is computed under non-standard conditions. In order to compute $\Delta {G_b}^\circ$ from ΔG_b a thermodynamic cycle that considers the volume change can be used (Figure 2.5).

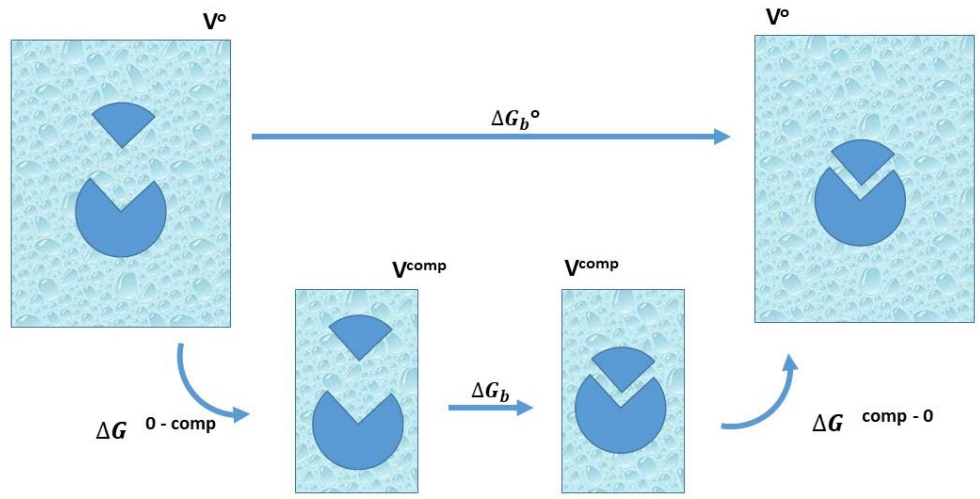


Figure 2.5: Thermodynamic cycle reflecting the transformations involved in determining the free-energy change of binding for a ligand to a protein under standard conditions, ΔG_b°, from the free-energy change of binding calculated under computational conditions, ΔG_b non-standard conditions. *The volume changes reflect the changes in the concentration states.*

This thermodynamic cycle allows us to determine a standard free-energy change from the simulated free-energy change (Fig. 2.5). The proportion of volume change during experimental and computational simulations (V^o and V^{comp}) can be related to the concentration C^o and C^{box}. The ΔG_b^0 according to the thermodynamic cycle is a sum of the three transitions:

$$\Delta G_b^0 = \Delta G^{0-comp} + \Delta G_b + \Delta G^{comp-0} \qquad (4)$$

The change in free energy from ΔG^{0-comp} is the free-energy change associated with a change in V from the experimental conditions to the computational conditions (both for the protein and for the ligand). If we assume that the internal energy change during this transition is negligible (ΔE = 0) the free energy change in the systems is because of a decrease in entropy due to a decrease in concentration:

$$\Delta G_b^{0-comp} = (\Delta E - T\Delta S)_P + (\Delta E - T\Delta S)_L = (-T\Delta S)_P + (-T\Delta S)_L = -2k_B T \ln \frac{V^{comp}}{V^0} \quad (5)$$

In the reverse scenario, the free-energy change is associated with an increase in entropy from V^{comp} to V^o. In this transformation, only the entropy change associated with the protein is considered due to the formation of a metastable complex PL (where ($\Delta S_{PL} = \Delta S_P$) as $\Delta S_L = 0$):

$$\Delta G_b^{comp-0} = (\Delta E - T\Delta S)_{PL} = -k_B T \ln \frac{V^0}{V^{comp}} = k_B T \ln \frac{V^{comp}}{V^o} \quad (6)$$

The thermodynamic cycle allows us to relate ΔG_b^0 for a standard process to ΔG_b access by computational approaches (if the concentrations of systems are known):

$$\Delta G_b^0 = \Delta G_b - k_B T \ln \frac{V^{comp}}{V^0} = \Delta G_b + k_B T \ln \frac{C^{comp}}{C^0} \; ; C^i = \frac{N^i}{V^i} \quad (7)$$

The ratio of C^{comp}/C^0 may be treated as being constant and unchanged in different PL systems investigated (using the same computational techniques). In this case, the $k_B T \ln \frac{C^{comp}}{C^0}$ term from Eq. 3 and Eq. 7, is insignificant for comparisons of ΔG_b during screening campaigns that make use of computationally derived energies. This is a gross oversimplification and only applies to systems that are compared under identical "simulation" conditions (General 2010). Naively one may consider that conversions from the computed ΔG_b to a standard free-energy can be made through inclusion of the C^o parameters. This correction alone, however, is insufficient due to the difficulty of reproducing low and high-energy micro-state population proportions that would be present in the canonical ensemble (Hermans & Wang 1997; Jorgensen & Thomas 2008). In binding affinity calculations this arises from a pervasive underestimation of the strength of non-bonded interactions in the unbound systems and limitations placed on the conformational searching (Wereszczynski & McCammon 2012; Gibbs 1902). Molecular docking approaches towards the calculation of

free-energy change of binding a query ligand to a target, are the workhorse method in structure based virtual screening techniques (Yagi et al. 2007). Other methods such as molecular dynamic simulations, alchemical free-energy calculations and combinations of both are able to sample states better than docking but are rarely used in early virtual screening stages due to their low-throughput (Genheden & Ryde 2015; Mobley & Klimovich 2012; Jiang & Roux 2010).

2.3.1.1 Molecular Docking

Molecular docking methods of virtual screening rely on the computation of a docking score of a small molecule and a macromolecule target. The docking score, within the context of a docking simulation, relies both on recovery of a ligands binding mode and estimation of the binding affinity of the ligand within the target "pocket" (Fang et al. 2016). ΔG_b behaves as a state function, and docking computes binding a the sum of the differences in energy, E, as a result of 1)conformational change that occurs in the protein P 2) conformational change that happens in the ligand L and 3) entropy lost (ΔS_{conf}) as a result of this binding. This is shown in Eq. 8:

$$\Delta G_b = (E_{bound}^{L-L} - E_{unbound}^{L-L}) + (E_{bound}^{P-P} - E_{unbound}^{P-P}) + \left(E_{bound}^{P-L} - E_{unbound}^{P-L} + \Delta S_{conf}\right) \quad (8)$$

Molecular docking relies on the precision of a scoring function that evaluates the intramolecular energetics of conformational change during association and the intermolecular energetics of complex formation (Cheng et al. 2009; Warren et al. 2006; S. Y. Huang et al. 2010). Docking experiments may include a re-scoring stage to enhance the sampling of ligand conformations within the receptor which may improve the visibility of the lowest energy conformation assumed to be the native binding mode (Campbell et al. 2014; Greenidge et al. 2014).

2.3.1.1.1 Scoring functions

Within a virtual screening context, a reliable scoring function is expected to identify the query compound that has the best affinity from a collection of compounds by ranking them according to the strength of their interaction with the receptor. A reliable scoring function is also expected to identify the pose that matches the native binding pose from a large collection of conformations and positions of that ligand query (Neudert & Klebe 2011). The pose matching ability of a scoring function is termed its "docking power" while its ability to predict the strength of interaction is its "ranking power" or its "scoring power" (Cheng et al. 2009). Scoring functions used in docking algorithms are based on the evaluation of either a force field derived function, an empirically derived function, a knowledge-based semi-empirical function or a descriptor-based function (Warren et al. 2006; Ding et al. 2015; Bryce 2011; Morris et al. 1998; Gohlke et al. 2000; Liu & Wang 2015).

By leveraging on advances made in the field of "molecular mechanics" domain, the free-energy of binding can be estimated from the solution to a Hamiltonian that describes the energetics of the molecular interaction. The solution to this Hamiltonian is derived from a scoring function that has parameters that describe energy components that are obtained by molecular mechanics force-fields or quantum mechanics derived terms (Bryce 2011). In molecular mechanics force-field approaches the energy components are a collection of gas phase precise atom-type specific bonded and non-bonded terms used to construct a master equation which is parameterized to include descriptions of physical factors such as polarization, solvation and entropy terms (Bohm & Stahl 2002; S. Y. Huang et al. 2010). Examples of this implementation were within the early versions of DOCK which made use of the Amber molecular mechanics force fields and the GOLD implementation which made use of atom types from the Tripos force field (Morris et al. 1998; Jones et al. 1997). The

contribution to the final free-energy of binding, ΔG_{bind}, is made possible through the force-field unbonded terms which involve a summation of the energy parameters across all the atom pairs in the bound and unbound state,

$$\mathrm{E} = \sum_i \sum_j \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i \, q_j}{\varepsilon(r_{ij}) r_{ij}} \right) \qquad (9)$$

where the receptor atom is i, ligand atom is j, distances between receptor and ligand atoms are r, the atomic charges are q and van der Waals parameters are A and B. The Lennard-Jones potential is typically used here to scale the VDW energy component. The Coulombic potential can be modified to include the distance dependant electrostatic constant, ε, of a pair of interacting atoms in the force field as in the AutoDock implementation (Morris et al. 1998; S. Y. Huang et al. 2010; Honig et al. 1993). Other modifications of the force-field derived master equations may include solvent parameterization through the incorporation of an implicit solvent term such as the Poisson-Boltzmann term (Ji & Zhang 2008). In order to ensure that parameterized force field derived functions do not overfit the solutions to the Hamiltonian, the force field parameters are calibrated by weighting factors derived from experimental data giving rise to the semi-empirical classification of some force-field derived scoring functions (Huey et al. 2007). This incorporation of weighting functions in the force field introduces the problem of transferability when the function is applied to systems outside the range of the experimental data used to "correct" the function. Force-field based functions that are derived from quantum chemical modelling, although computationally demanding have been deployed successfully as alternatives (Bryce 2011). By computing ligand terms within the receptor at the QM level and protein terms at the MM level with polarizability introduced via an induced charge method, QM/MM based scoring functions like QMScore, have been implemented in molecular docking approaches to enhance the precision scoring

(Illingworth et al. 2008; Bikadi & Hazai 2009; Brahmkshatriya et al. 2013). Alternatively the whole system can be treated at the QM level to further enhance the ranking power and the docking power of the scoring function at the expense of speed during the docking simulation (Bryce 2011). Performing fewer computations and making use of reliable linear scaling techniques have allowed for *ab initio* QM level Hamiltonians to be computed for an entire fully polarizable ligand-receptor (Grimme 2004; Fong et al. 2009; Morgado et al. 2007).

While force-field based methods derive a scoring function using parameters from molecular mechanics or quantum mechanics in their master function, empirical based methods use linear free-energy relationships from experimental systems (H. J. Huang et al. 2010). Empirical based scoring functions derive the parameters for the master equation from the linear regression of carefully selected experimental systems that have known binding energies (Morris et al. 1998; Eldridge et al. 1997; Wang et al. 2002; Friesner et al. 2004). To estimate the binding energy, empirical scoring functions use terms that describe the van Der Waals interactions, contributions to hydrogen bonding, estimations of electrostatic and the solvation terms of the protein - ligand complex (S. Y. Huang et al. 2010).

$$\Delta G_b = \sum \mathcal{W}_i * \Delta G_i, \qquad (10)$$

The relative flexibility of the systems and their sizes are considered by estimating the number of torsions and rotatable bonds in the query systems, N_{Tor}. To calculate the binding energy it is assumed that the free-energy change in binding that occurs when a ligand is docked to a protein is because of the deviations from the optimal positions of each of the terms across all the interacting atom pairs in the ensemble. Each of the terms is given a range upon which the contribution is either optimal, minimal or negligible. When the measured deviation is optimal, a higher scoring factor is assigned to the term while a lower value for the factor is assigned

when the deviation is minimal. In the case where the measured deviation is negligible the contribution from that interaction for the two atom pairs is ignored, implying that there is no contribution to the free – energy of binding by an interaction of those two atom pairs (Wang et al. 2002). Enhancements in the docking power of empirical based scoring functions include interventions such as increasing the number of protein-ligand complexes used for training and increasing the diversity of in the training sets (Morris et al. 2009; Wang et al. 2002). Glide 2.5 has its ChemScore scoring function improved by having emphases on the different types of hydrogen bonding contacts and metal-ligand interactions, by including protein-ligand Coulomb and vdW interaction energies, and by employing an explicit-water solvation approach to minimize the false positives (Friesner et al. 2004). By improving the quality of training protein-ligand complexes, a simpler version of the Vina scoring function was obtained. Quality improvement was as a result of reducing steric clash and hydrophobic interaction terms, adopting atom type specific atomic radii parameters and replacing Lennard-Jones type potentials with Gaussian attraction terms and quadratic repulsion terms instead (Quiroga & Villarreal 2016; Liu & Wang 2015).

Assumptions derived from inverse Boltzmann statistical mechanics approaches for fluids inspire knowledge-based potential functions. Knowledge-based scoring functions, although inspired by semi-empirical force-field methods, utilize the ensemble of structures to determine pair-wise mean-force potentials instead of affinities (Huey et al. 2007; Gohlke et al. 2000; Muegge 2006; Neudert & Klebe 2011). Reference structures from databases are used to derive a distribution function of the pairwise distances between different atom types typical of native structures (Muegge 2006). By assuming that the reference structures (ρ^*) provide a global minima, knowledge-based potentials penalize pairwise distances that deviate

from this minima contributing to its classification as an unfavourable conformation (S. Y. Huang et al. 2010).

$$\omega(r) = -k_B T \ln[\mathrm{g}(r)]\,, \mathrm{g}(r) = \rho(r)/\rho^*(r) \qquad (11)$$

The total score for the conformation is heavily influenced by the distribution function from the pairwise distances, and, as a result the precision of a knowledge-based scoring function is influenced by the resolution of the reference structure (Muegge & Martin 1999). Strategies exist to improve the resolution of the reference structure. Such strategies include treatment of protein-solvated and protein-bound solvent-mediated effects, and only dealing with specific receptor-ligand interactions. These then influence the calculation of the pairwise-distances distribution function for the atoms of interest (Gohlke et al. 2000). Reducing the dominance of solvent-mediated interactions is achieved through the use of a distance limit that emphasizes receptor-ligand interactions in the pair-wise distance-distribution function (Muegge & Martin 1999). Solvent-mediated effects that affect the Gibbs free energy within the distance limit are accounted for in the scoring function by a single-potential Solvent Accessible Surface Area (SASA) term (S. Y. Huang et al. 2010). An approximate cube algorithm is used to calculate this SASA term specific for solvent atoms, and for both the protein and the ligand in both their bound and unbound states (Böhm 1994). Changes to this term are captured into the scoring function allowing an implicit consideration of solvent effects in the molecular docking simulation (Gohlke et al. 2000). Another popular strategy to modify knowledge-based scoring functions is the inclusion of torsion angles in computing the total score of a conformation. In this implementation a torsion scoring term classifies a particular torsion system and applies the appropriate torsion function that favors near native torsion angles and penalizes unlikely torsion angles (Neudert & Klebe 2011).

Although the ranking power of a scoring function may be due to its being a linear combination of energy terms, this assumption is being challenged by approaches that consider the cooperativity of non-covalent interactions (Kinnings et al. 2011). Scoring functions derived from force-field based approaches, empirical based methods or knowledge-based methods assume that individual terms scale linearly to the total score of the interaction when in reality non-covalent interactions interact with each other in a non-linear fashion (Ding et al. 2015). Non-linear regression methods have seen increasing interest in the field of drug-discovery due to improved access to computational resources and advancements in the deployment of machine learning algorithms (Deng et al. 2004). Some of the machine learning classifiers used in molecular docking are naïve Bayesian methods, k-nearest neighbors, decision trees, support vector machines as well as deep-learning and convolutional neural networks (Liu & Wang 2015; Yan et al. 2017). These classifiers use features extracted from reliable training data to generate a model that will be deployed by the scoring function (Yan et al. 2017; Ragoza et al. 2017). The features extracted from the training data sets, and used to derive the docking scoring function can be from one of two general descriptor classes (L. Li et al. 2011). The features can belong to the simple element-element descriptor class (ideal for applications that emphasize "ranking" power) or the interaction-driven descriptor class (that emphasize "docking" power in their prediction) (Liu & Wang 2015; L. Li et al. 2011). Studies have shown that although non-parametric machine-learning methods are able to derive flexible scoring functions, they may fail to correlate the binding affinity with the binding mode predicted (Gabel et al. 2014). The incorporation of negative training data and careful choice between distance-driven and interaction-driven descriptors has been shown to improve the performance of machine-learning based methods in predicting the binding modes of novel ligands with a protein target (L. Li et al. 2011; Ding et al. 2015). Another strategy to improve

the precision of machine-learning scoring methods is to derive target-specific scoring functions that map protein pocket residues (Yan et al. 2017; Ballante et al. 2014; Matter et al. 2005).

2.3.1.1.2 Searching method

Molecular docking represents an optimization problem that requires the inspection of spatial parameters such as positional, orientational and conformational possibilities in order to conclude a search (Fu et al. 2015). It is often easier to predict the binding mode of a ligand in the native crystal protein (that has already accommodated that ligand) than it is to predict the binding mode of a ligand in a different protein. This is because the receptor cavity residues contort themselves to accommodate the guest ligand when the crystal structure is solved and this is difficult to replicate in rigid docking experiments (L. Li et al. 2011). Peptide ligands in particular offer difficulties due to the absence of double bonds and rigid rings increasing their flexibility and thus are more sensitive to searching and sampling biases (Rentzsch & Renard 2015). Docking algorithms attempt to survey the conformational landscape available to a ligand, within a receptor that is treated as either being rigid or flexible during the computation (Cross et al. 2009). The methods employed by docking algorithms are vast, in our discussion we will focus on the search methods that are most relevant for virtual screening of peptide libraries: grid-based searching algorithms used in AutoDock and Glide as well as the structure refinement algorithms used by HADDOCK and FlexPepDock for protein-protein docking (Morris et al. 1998; Friesner et al. 2004; Trellet et al. 2013). Programs such as SODOCK, which use docking algorithms incorporating of quantum-behaved particle swarm optimization (QPSO) solutions for flexible molecular docking problems, will be discussed briefly (Fu et al. 2015).

AutoDock's search method exploits the Lamarckian Genetic algorithm (LGA), a hybrid search algorithm that makes use of simulated annealing in performing local searches while incorporating a Genetic algorithm for a rapid global search (Morris et al. 1998; Morris et al. 2009). The LGA employed by AutoDock uses 3 "chromosomes" to direct its search. The first "chromosome" contains the set of "genes" associated with the translation of the ligand; the second "chromosome" contains the set of "genes" that influence the ligand orientation and the third "chromosome" codes for the ligand conformation if torsions exist. The translation "genes" are the *x*, *y*, and *z* co-ordinates within the grid of the search space. The four orientation "genes" are the quaternions that describe the rotation to be applied to all atoms of the query molecule in Euclidean space without introducing the gimbal lock. The torsional "genes" are assigned values that define the angles of each of the AutoTors torsions present in the query molecule identified. Conformation searching in AutoDock follows the generation of a population of individuals that are spawned at random and are evaluated for their fitness according to an efficient docking score function, that relies on a grid-based look-up table. Mutations of the torsion "genes" as well as crossover of the orientation and translation "genes" are allowed in order to expand the search space of individuals. In order to speed-up the searching, subsequent global searches are permitted by allowing "fit" individuals to transfer their traits to the next generation of individuals who will undergo the "local" mutation and crossover search. The process is repeated until the maximum number of evaluations is achieved and individuals are reported as being near native conformations based on how well they satisfy the fitness function (Morris et al. 1998).

The LGA method employed by AutoDock relies on the performance of AutoTors to define the rotatable bonds of the ligand during the preparation stage (Morris et al. 1998). In order to overcome AutoTors' inability to detect rotatable bonds present in constrained macrocyclic

systems it has been suggested that converting cyclic systems to their acyclic counterparts through the introduction of dummy atoms and the inclusion of a customized two-point attractor Lennard-Jones potential could manage macrocycle flexibility during the docking process (Forli & Botta 2007). Although the customized scoring function is able to bias ring closures, the introduction of flexibility into ring of a ligand reduces the precision, enriching false-positive poses due to expansion in the degrees of freedom of the search space (Zhong et al. 2017). The degree of exhaustiveness of the LGA search is specified by the user, who makes adjustments in the proportion of the positional and conformational landscape that is explored. This is done in a non-exhaustive, pseudo-stochastic manner (Morris et al. 2009). By treating optimization of flexible docking search methods as a "continuous optimization problem" probabilistic algorithms, such as the quantum-behaved particle swarm optimization, have been shown to improve the performance of molecular docking search methods (Korb et al. 2006; Chen et al. 2007; Fu et al. 2015).

In a manner similar to that of AutoDock, the Glide search algorithm makes use of a grid-based lookup method to evaluate the score of a particular ligand pose using scoring function energy fields. Glide however avoids a stochastic approach in its survey of the docking space and opts for a robust hierarchical searching filter. The filter begins with an exhaustive conformational search and minimization step of the ligand before these ligand poses undergo a phase space search to locate their optimum positions and orientations in the target. The first vacuum minimization step accounts for macrocyclic flexibility prior to phase space searching. A second minimization step of the ligand conformations within the receptor (pocket of the protein) prepares low energy conformations of the ligand for a torsion angle phase space optimization. This optimization uses a Monte Carlo procedure that refines the search and identifies probable binding modes. The energy evaluations may make use of the Emodel scoring

function which is a combination of an OPLS-AA force field based function with contracted van der Waals radii, the GlideScore empirical scoring function and a ligand strain energy term (Friesner et al. 2004).

Grid-based methods for predicting the bonding pose of a guest small molecule in the vicinity of a target have largely been successful and methods to assess their reliability are well advanced (Trellet et al. 2013; Rentzsch & Renard 2015). Molecular docking approaches for small molecules are better suited to re-docking ligands, or cognate docking into a crystal structure because this rigid protein will be preconfigured to the bound conformation (L. Li et al. 2011). Proteins and peptides undergo significant conformational changes when they transition from their unbound conformation to their bound conformation owing to their large flexibility (Raveh et al. 2010). Thus, computational approaches to predict the bound conformation from an unbound conformation requires exhaustive simultaneous sampling of side-chain flexibility as well as back-bone dynamics in the receptor and the incoming target (Gray et al. 2003; London et al. 2011). Scoring functions of small molecular docking approaches typically underestimate the impact of solvent effects in the binding process (Moreira et al. 2015; Lensink & Wodak 2013). This is justified by noting that most small molecule binding sites are located deep within buried pockets (Petsalaki & Russell 2008). However, the typical binding sites of peptides and proteins are surface sites that are exposed to the solvent and implicit solvent models therefore underestimate the impact of specific solvent mediated interactions (Trellet et al. 2013; Friesner et al. 2004). Owing to the fact that scoring functions utilize functions that have been parameterized using experimental binding data obtained from small molecule ligand databases the docking score may not be transferable to peptide or protein ligands (Iain H Moal et al. 2013). The prediction of binding affinities of peptide-protein interactions suffers from significant setbacks related to the

absence of experimental techniques that give reproducible estimations of binding affinities leading to low correlation between datasets and modelling experiments that accessed native conformations (Iain H. Moal et al. 2013; Gromiha et al. 2017; Moreira et al. 2015; Lensink & Wodak 2013).

Protein-protein docking programs such as the HADDOCK (high ambiguity driven docking) and FlexPepDock were originally devised to meet the need of refinement of solution structures of protein-protein interactions (Trellet et al. 2013). HADDOCK, one of the popular protein-protein docking programs, uses an ambiguous interaction restraint (AIR) score derived from user supplied experimental data that identifies residues that are actively or passively complicit in the interactions responsible for stabilizing the complex (Dominguez et al. 2003; Moreira et al. 2009). The three stage docking funnel for HADDOCK uses the AIR score in its interaction energy during the first rigid docking step of aligning the ligand and receptor partners from random orientations. This rapid alignment is followed by sequential simulated annealing steps that optimizes the local constraints and conformational rearrangements of the side chains and the backbone torsions. The final stage of the docking funnel is a simulation dependant optimization of the structure in the presence of explicit solvent. In order to enhance the stochasticity of the simulation, parallel tempering is used and the final complexes are arranged in clusters that are ranked according to their electrostatic, van der Waals and the AIR energy scores (de Vries et al. 2007; Dominguez et al. 2003). FlexPepDock was developed as a molecular docking tool to refine protein-protein complexes and its Monte Carlo with energy minimization approach has seen successful extensions in the prediction of the structure and binding energy of protein-protein complexes (Gray et al. 2003). To achieve reliable docking the FlexPepDock relies on the Rosetta docking protocol that begins with the creation of starting positions of the interacting partners. Low-resolution coarse-grained

alignments are performed by rigid-body Monte Carlo searching that uses residue-scale interaction potentials. The coarse-grained optimized models are sent through a high-resolution minimization stage. The high-resolution minimization protocol begins with residue packing and optimization of the coarse-grained model and concludes by performing a high-resolution rigid-body alignment and a sequential conformational search of interacting residues. Modifications of this algorithm include the incorporation of *ab initio* functionality that is able to dock folded peptides (from query peptide sequences) through the use of extensive backbone flexibility through the use of Rosetta backbone sampling tools (Raveh et al. 2011). Although FlexPepDock incorporates a statistically weighted full-atom scoring function that includes repulsive terms during its high resolution refinement stage, it is better suited for the refinement of binding modes of peptide-protein complexes (Gray et al. 2003; Raveh et al. 2011; London et al. 2011).

High-throughput virtual screening relies on the precision and efficiency of ranking protocols within molecular docking approaches.

2.3.1.2 Molecular Dynamics

While molecular docking approaches show promise in high-throughput structure based drug discovery due to their ability to predict the binding affinity of query molecules with target receptors, MD approaches are used for high-resolution interrogation of drug-target interactions. Through the incorporation of explicit solvent models, thermodynamic control, and receptor and ligand flexibility during system evolution, MD simulations are able to model the ligand-receptor conformational and energetic landscapes relevant for computationally driven drug discovery (De Vivo et al. 2016).

2.3.1.2.1 Theoretical underpinnings of MD

The central premise of MD simulations is Newton's second law of thermodynamics. Under this restriction, if the total enthalpy (H) of a microcanonical ensemble is kept constant, the time-dependent behaviour of *N* constituents of the system can be reproduced from their unconstrained Cartesian coordinate space ($\boldsymbol{x}$) and momentum ($\boldsymbol{p}$), by computing the potential (V) and kinetic (K) energies,

$$H(\boldsymbol{x},\boldsymbol{p}) = V(\boldsymbol{x}) + K(\boldsymbol{p}),\ \boldsymbol{x} = \{x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3\ \ldots . x_{\mathrm{N}}, y_N, z_N\} \quad (12)$$

A force field based function is used to compute $V(\boldsymbol{x})$ stored in the configuration of the system while a change in $K(\boldsymbol{p})$ can be observed if parts of the system are allowed to perturb over-time. In the case of isolated microscopic systems, accurate simulations depend on kinematic treatment of the spatial integral of the force, F, acting on an atom, *i*. To do this, classical mechanics asserts that F_i is constant over the small time step, δt, allowing an integrator calibrated with an appropriate relaxation time to update the configurational space, $\boldsymbol{x}$ (t + δt), of the system using average acceleration as an input to the Verlet method (Verlet 1967; Lippert et al. 2007). Prior to the update, algorithms such as SHAKE constrain covalent bond lengths and angles transforming the unconstrained configuration to a constrained configuration through the application of a restraining force on the particles computed from the previous step (Berendsen et al. 1995).

$$F_i(t) = m_i a_i(t) = -\frac{\partial V(\boldsymbol{x}(t))}{\partial x_i(t)} \quad (13)$$

The configuration of the system $\boldsymbol{x}$, as well as important thermodynamic observables such as pressure, temperature and potential energy at t + δt are recorded in a trajectory for later analysis. F_i (t + δt) is used for the next iteration until sufficient perturbations of the system

have been recorded. Most simulations of biomolecular systems make use of MD engines that compute energies of the systems using molecular mechanics based force-fields in explicit solvent models (Durrant & McCammon 2011). Although precision of the energies computed could improve with the use of QM based methods, the faster MM methods allow the simulations to sample the conformational landscape thoroughly within reasonable time (Wang et al. 2004; Genheden & Ryde 2015; Boulanger & Harvey 2018). Although for simplicity an all-atom Cartesian model is used to describe the configuration of the entire system, $\boldsymbol{x}$, there has been a growing interest in the use of computationally inexpensive bond/angle/torsion (BAT) coordinate models in internal coordinate molecular dynamics (ICMD) (Mazur & Abagyan 1989; Chen et al. 2005). By using BAT models in constrained ICMD simulations, performance improvements are achieved through the reduction in degrees of freedom sampled and the use of longer integration time steps allowing the dynamics of components that undergo large concerted conformational changes to be captured within MD simulation time scales (Wagner et al. 2013; Amadei et al. 1993).

Simulations of biomolecular systems make use of kinetic or configurational based isothermal-isobaric algorithms to approximate solutions to the Nosé-Hoover equations during the updating of the motion of the system by the integrator (Braga & Travis 2006; Beckedahl et al. 2016). Configurational based algorithms in temperature and pressure control have been shown to speed-up calculations of heat transfer allowing simulations to better mimic macroscopic behaviour (Travis & Braga 2006; Pronk et al. 2013). The use of periodic boundary conditions (PBC) and complementary distance dependent treatments of electrostatic and VDW non-bonded terms in the computations ensures that the simulations are robust and have reduced computational cost (Berendsen et al. 1995).

In order to extract statistically virtuous and ergodic ensembles, algorithms that speedup computations and enhance the sampling of classical MD simulations have been developed (Kasahara et al. 2018). These methods allow for detailed interrogations of molecule-receptor interactions such as binding-free energies and dissociation constants of weakly interacting solutes to be performed from the extraction of the MD-derived thermodynamic observables (De Vivo et al. 2016; Abrams et al. 2013). The central premise of these methods lies in overcoming the limitation placed by Boltzmann sampling in MD (Wereszczynski & McCammon 2012). During a simulation, trajectories are extracted along a path of the configurations computed in a non-deterministic manner (Lei & Duan 2007). In classical MD, these trajectories sample along a single trajectory and thus have a lower probability of visiting the high energy, less populated, configurational space of the system's canonical ensemble (Amadei et al. 1996). The canonical ensemble of microstates exists in a multidimensional space and only an infinitely long, ergodic simulation would ever sample this space (Gibbs 1902). However, classical MD simulations are non-ergodic and non-infinite, therefore, it is vital that thermodynamic descriptions associated with a transition of interest are obtained from methods that address sampling limitations of the configurational space (Abrams et al. 2013; Kasahara et al. 2018). Depending on the strategy they use to overcome the Boltzmann distribution, enhanced methods can either employ physical or non-physical sampling pathways.

2.3.1.2.2 Physical molecular dynamics methods

Enhanced sampling methods that are defined as "physical" recover thermodynamic observables from trajectories. But these trajectories sample the configurational space of an ensemble that is guided along a reaction coordinate or propagated towards a path in an impulsive way (Wereszczynski & McCammon 2012; De Vivo et al. 2016). For reaction-

coordinate biased physical sampling such as umbrella sampling (US), steered MD and metadynamics approaches, an “a posteriori” defined collective-variable (CV) is used to biase the simulation towards a reaction path of interest (Kumar et al. 1992; Torrie & Valleau 1974; Laio & Parrinello 2002). In order to recover observables that define the free energy change from binding, ΔG_b, umbrella sampling simulations generate a CV that follows the binding and unbinding reaction of a protein – ligand complex (Doudou et al. 2009). To sample this reaction, a force constant and a distant restraint on the ligand atoms is introduced into the H of the ensemble. This divides the path into snap-shot simulations of the protein - ligand complex at distances along the reaction coordinate (Doudou et al. 2009; Kumar et al. 1992). Free energies from each of the simulations undergo post-processing using a WHAM analysis that allows probability distributions from multiple simulations to be used in the estimation of a potential of mean force (PMF) along the “reaction” path (Kumar et al. 1992). In a drug-discovery context, the PMF can be used to extract thermodynamic descriptions such as ΔG_b as well as the kinetics associated with the ligand residence-time in the receptor (Doudou et al. 2009; Kumar et al. 1992; De Vivo et al. 2016; Pan et al. 2013). In instances where broader exploration of the free-energy expressed as a surface is desired, metadynamics can be applied (Bernardi et al. 2015). Whereas in umbrella sampling the ensembles are constrained along a collective variable by constant restraints, metadynamics simulations incorporate adaptive biasing restraints from an average of the instantaneous force of the ensemble during an unconstrained simulation (Laio & Parrinello 2002). This allows the simulation to follow a more ergodic collective variable with the force constraint depending on the derivative of the free energy, inserting a memory component to the sampling of phase space during the simulation (Bernardi et al. 2015; Darve & Pohorille 2001). Although this memory-laden approach to sampling allows an exhaustive search of the phase space, it’s practicality is limited to

interrogating free-energies of systems with smaller degrees of freedom such as those in conformational sampling and protein folding studies (Abrams et al. 2013; De Vivo et al. 2016; Lobb 2015).

Simulated annealing (SA) methods are examples of physical MD methods that do not follow the generation of a collective variable in their sampling (Bernardi et al. 2015). SA methods are optimization algorithms, analogous to the genetic algorithms introduced during the Molecular Docking discussion (Tsallis & Stariolo 1996). These optimizations rely on the generation of conformations by random sampling of a trajectory (followed by minimizations) to identify a range of low energy conformations (Agostini et al. 2006). The three main parameters manipulated during annealing (cooling) are the parameter that influences the distribution of torsions displaced (qv), a parameter that influences the probability of accepting (qa) a conformation and the parameter that influences the temperature at which acceptance is occurring (qt) (Agostini et al. 2006).

The most popular non-CV based physical methods that allow broad searching of the phase space are the replica-exchange MD methods (Bernardi et al. 2015). These methods enhance sampling through the cloning of a configuration into a set of replicas that have variable temperatures (parallel tempering, REMD) or scaled Hamiltonians (solute tempering, REST) of the system (Patriksson & Van Der Spoel 2008; Ostermeir & Zacharias 2014). During the simultaneous propagation of these replicas, probability driven “swapping” or “mutation” of configurations between a pair of replicas may occur on the satisfaction of a Monte Carlo criterion that considers the reference states (temperatures) and instantaneous free energies of the replica pairs (Patriksson & Van Der Spoel 2008). This mutation allows the configurations to overcome energy barriers allowing them to access energy minima within the phase space of the original configuration (Sugita & Okamoto 1999). When suitable mixing events have

been allowed across a diverse set of replicas, REMD strategies are able to enhance the sampling of the canonical ensemble facilitating protein folding studies dependent on extraction of sparsely populated portions of phase space (Patriksson & Van Der Spoel 2008; Ostermeir & Zacharias 2014). In the context of drug discovery, REMD approaches can be used to enhance the sampling of non-physical alchemical transformation MD strategies such as free-energy of perturbation simulations (De Vivo et al. 2016).

2.3.1.2.3 Non-physical molecular dynamics methods

Non-physical approaches rely on the robust treatment of free-energy change as a "state function" where the calculation of its change can be deconvoluted through the incorporation of a thermodynamic cycle involving an alchemical transformation (General 2010). Despite the advancements in the computational determination of thermodynamic properties such as binding energies, energy changes of different molecules are often meaningless and misleading in isolation and require consistent rescoring (Jorgensen & Thomas 2008; Eken et al. 2018). Artefacts introduced by reference force-fields and scoring functions within the computation of energy terms trivializes the impact of packing effects and resonance effects on systems for the sake of simplicity (Genheden & Ryde 2015; Boulanger & Harvey 2018; Bryce 2011). Enhancements of binding-energy computations that make use of combinations of molecular dynamics at the MM level with high resolution solvent models and energy computations at the QM and MM level still show limited correlation with their experimental values (Eken et al. 2018).

As is done during virtual screening, it is common for computed energy values to be expressed as relative state descriptions which cancels out artefacts (Yan et al. 2017; Lionta et al. 2014). Extensions of this allow the implications of the thermodynamic cycle of Fig. 2.5 to be extended to include non-physical transformations in the computation of the difference of free-energy

binding for different ligands. It is insisted arbitrarily in Fig. 2.5 that the thermodynamic cycle be of a physically interpretable transition of a system from standard conditions, $\Delta G_b°$, to non-standard conditions, ΔG_b. This does not have to be the case. The inclusion of alchemical transformations allows us to depart from this imperative relating the free-energy in terms of transformations of molecules from one state or identity to another identity (Jorgensen & Thomas 2008). Although hypothetical, this is consistent with assertions made by the early architects of the "Perturbation Theory" and its adoption in structure-based drug discovery with statistical mechanics (Zwanzig 1954; Wereszczynski & McCammon 2012; Gibbs 1902). Figure 2.6, highlights a thermodynamic cycle that allows the computation of energy change of binding by including energy changes of alchemical transformations, obtained from a series of free-energy perturbations (FEP).

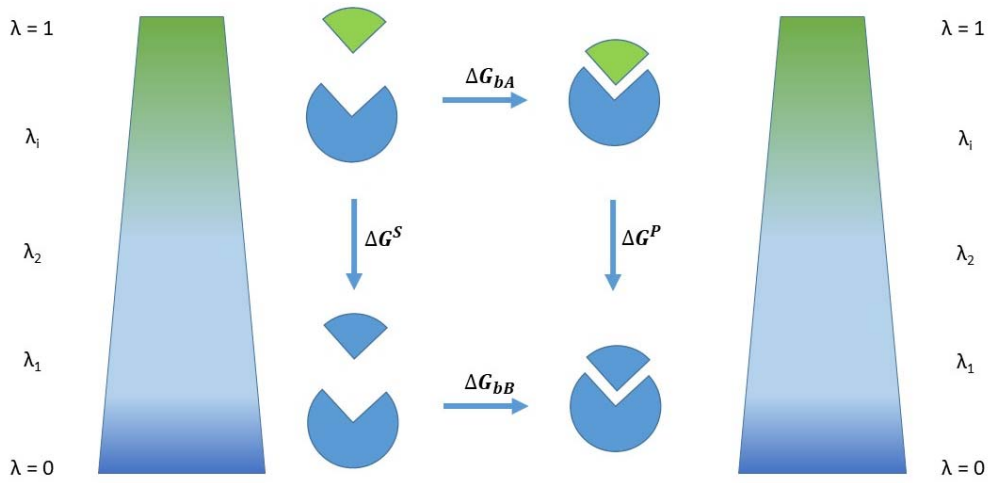


Figure 2.6: Thermodynamic cycle of the free-energy of transfer of compound A, ΔG_bA, and B, ΔG_bB, from the unbound to the bound state. *The vertical sides of the cycle reflect the discretized states, λi, of the alchemical transformations computed using FEP simulations with left in solvent, ΔG^S, and the right side , ΔG^P,*

This thermodynamic cycle reflects the free energy pathways that represent the transformation of compound A into compound B within the media of the protein binding pocket (Jorgensen & Thomas 2008).

$$\Delta\Delta G_{AB} = \Delta G^{P} - \Delta G^{S} = \Delta G_{bB} - \Delta G_{bA} \; ; \Delta G_{bB} = \Delta\Delta G_{AB} + \Delta G_{bA} \qquad (14)$$

The difference in the free energy change of binding in the mutation of compound A to compound B, $\Delta\Delta G_{AB}$, is obtained from the computation of ΔG_{bA}, and ΔG_{bB}. The $\Delta\Delta G_{AB}$ can be seen as a perturbing potential applied on A to achieve its alchemical transformation to B (Jorgensen 2008). Computational approaches are used to estimate the perturbing potential through a series of simulations that introduce a mixing parameter, λ_i, to scale the geometrical and force-field parameters, H, that enable the transformation of $A_{(\lambda=1)}$ to $B_{(\lambda=0)}$ (Hermans & Wang 1997; Procacci 2017).

$$\Delta\Delta G_{AB} = \sum_{i=0}^{i}(\lambda_i)\Delta G_{bB} + (1 - \lambda_i)\Delta G_{bA} \qquad (15)$$

Careful choice of the discretized mixing parameter widths allowed the demonstration of FEP in the lead optimization of catechol diethers with potency against reverse transcriptase HIV targets (Bollini et al. 2011). In this case, sampling of the protein-ligand complexes was achieved using a Monte Carlo approach while the free-energies were computed at the OPLS-AA force field for the protein, OPLS-/CM1A ligand level and explicit solvent. Replica-exchange MD simulations coupled with FEP were demonstrated to enhance sampling in fully automated computational solution, FEP+, for the lead-optimization of a broad range of target classes (Wang et al. 2015). The workflow developed accounted for transformations through the use of a perturbation map that is able to guide the substitutions of up to 10 heavy atoms present within the phase space of a query Hit molecule. The free-energy of binding of a molecule that had substantial enhancement of affinity was accessed by traversing pathways that connect

pairs of molecules within a closed cycle. This FEP enabled drug discovery pipeline was shown to contribute to a speed up advantage of up to 1000 x that of experimental approaches (Abel, Mondal, et al. 2017). It allowed the interrogation of over 3000 compounds that had optimized potency, solubility, stability and enhanced membrane permeability to probe JAK family kinases with greater specificity than Tofactinib, approved for rheumatoid arthritis therapies (Abel, Wang, et al. 2017; Dowty et al. 2014).

2.3.2 Ligand based methods

Whereas structure based methods rely on structural data of targets to derive a fitness score for assessment, ligand based virtual screening (LBVS) methods use fitness scores based on molecular similarity (Raman et al. 2017; S. Y. Huang et al. 2010). Molecular similarity is an ambiguous term that is often misrepresented when the worlds of cheminformatics and medicinal chemistry collide (Cereto-Massagué et al. 2015; Shim et al. 2011). The types of similarities that exist are a matter of perspective where two compounds that may appear to have similar physicochemical properties, for example, may possess vastly different activities as is characterized by the activity cliff phenomenon (Hu & Bajorath 2012; Stumpfe et al. 2014). Cheminformatics considers similarity as chemical (if reaction information is considered), molecular (if structural features such as functional groups and substructures are shared), biological (if activity data is considered), local (if the relationships between certain functional groups or substructures of molecules are considered) and global (if it is concerned with a whole compound view) (Maggiora et al. 2014). Thus searching for active compounds based on similarity may be misleading and similarity searches are better placed to describe relationships and correlate activities rather than predict activities of compounds that possess low similarity to active compounds (Shanmugasundaram & Rigby 2009; Maggiora 2006).

Ligand-based virtual screening (LBVS) methods rely on the availability of biological activity data of known ligands with targets or disease states of interest in order to derive a relationship that can be used during virtual screening campaigns (Shim et al. 2011; Bickerton et al. 2012; Nicholls et al. 2010). Ligand based methods use advances in data mining and cheminformatics approaches to interrogate the structural and chemical similarities and differences (Ruddigkeit et al. 2012; Melville et al. 2009; Yap et al. 2007). These comparisons extend virtual screening campaigns by incorporating activity data such as efficacy, selectivity, pharmacokinetics, activity and network polypharmacology in order to assist in prediction or interpretation of the activities of query molecules (Hopkins 2008; Kumar et al. 2015; Clemons et al. 2010; He et al. 2016). The main methods employed in determining ligand-based correlations in molecule sets are the Quantitative Structure Activity Relationship approach (QSAR) and the Pharmacophore dependant strategy (Yap et al. 2007; Van Drie 2007; Willett 2006).

2.3.2.1 Quantitative Structure Activity Relationships

The basis of QSAR methods for the correlation of activities in molecular sets is that a mathematical model can be built to correlate a biological response parameter from the linear free-energy relationships of molecules and their substituents (Singer & Purcell 1967). The Lipinski Rule of 5 is analogous of a QSAR where a property space restricts the region where orally bioavailable drugs are likely to exist (Lipinski et al. 2012). Low-dimensional linear-free energy relationships of different substituents can create a parametric structure activity response (SAR) for compounds related to a scaffold, while linear and nonlinear regression models derived from physicochemical and structural properties of training sets improve the resolution of predictive models and allow for more reliable inferences to be made from diverse compound sets (Yap et al. 2007). The observed molecular behavior and thus QSAR, of

a test molecule is present as a consequence of the cumulative effective of the descriptor properties that are influenced by the presence of functional groups and substituents (Hansch 1969). Apart from the Lipinski descriptors other descriptors obtained from experimental or computational approaches include the molecule volume, surface area, electronegativity, polarizability, ring types, shape, symmetry, planarity as well as functional groups present (Lovering 2013; Tiejun Cheng et al. 2007; Hou & Xu 2003). To highlight the magnitude and diversity of molecular descriptors available, 42 molecular structure descriptors were used to map the GDB-17 virtual library into a phase space (Ruddigkeit et al. 2013).

The statistical models used to derive a QSAR model for a molecular set are varied and typically include combinations of multiple linear regression, partial least squares and machine learning approaches. Multiple linear regression (MLR) methods are used in instances when a linear relationship between a set of *k* descriptors (χ) and a biological response exists for a training set of a large number of compounds (at least five times k). Training of the MLR QSAR model allows for a selection of constants (β) that minimize the sum of the squares of residuals from the difference between the predicted activity (y) and the experimentally determined activity (Abel, Mondal, et al. 2017; F. Cheng et al. 2012).

$$y = \beta 0 + \beta 1\chi 1 + \beta 2\chi 2 + \cdots \beta k\chi k \qquad (16)$$

The partial least squares (PLS) approach to model building aims to identify a linear combination of variables (χ) related to the molecular descriptors that maximizes the covariance between χ and the predicted activity (y) (Zhang et al. 2005). PLS methods are used when MLR methods are limited (in cases where datasets where the number of compounds < 5*k* where *k* is the number of descriptors) (Khajehsharifi et al. 2017).

Validated models built from machine learning methods have been used to correlate molecular properties with responses relevant to early stage drug discovery owing to their adaptability and versatility (Ghasemi et al. 2018). The k-Nearest Neighbour machine learning approach was used to estimate the endocrine disrupting activity of a diverse set of compounds from the average of members in a training set with close similarity as measured by the Euclidean distance (Asikainen et al. 2004; Sahigara et al. 2013). In another case, topological descriptors were used to train support vector machines (SVM) that use linear regression of a multidimensional feature space obtained from the input vectors of the training set (Oloff et al. 2005). These validated SVMs were used to mine a diverse collection of compound databases and resulted in the identification of compounds with atypical pharmaceutical profiles as dopaminergic antagonists. Supervised and unsupervised neural network (NN) methods have also shown promise in their relevance when deployed in virtual screening campaigns (Ghasemi et al. 2018). Deep learning NNs (DL NN) were shown to outperform commercial target prediction methods on the CHEMBL dataset when 43,340 features were used to train a multi-layer network (Unterthiner et al. 2015). Despite their complexity and computational cost, DL NN are attractive as machine learning methods for QSAR prediction because they are able to overcome redundancy, overfitting and the problems of the large number of molecular descriptors available that limits the reliable use of ordinary machine learning methods (Ghasemi et al. 2018). The use of multiple hidden layers of neural network training enhances the robustness of the trained model while coupling the deep learning neural net with neural networks that are solely responsible for feature selection. This ensures that the resulting QSAR models are able to outperform commercial target prediction machine learning methods (Wen et al. 2017; Zhang et al. 2017).

Although mathematical models can be built to predict and interpret relevant biological responses reliably, the inclusion of spatial descriptors has been shown to consistently contribute to the precision of QSAR models (Shim et al. 2011; Polanski 2009). Multi-dimensional QSAR models have been explored by incorporating interaction fields into 3D-QSAR grids and ligand flexibility as 4D-QSAR (Klebe et al. 1994; Clark 2009; Klein & Hopfinger 1998). Knowledge of receptor structures and solvent effects have been included in some models giving rise to 5D and 6D QSAR models (Polanski 2009).

2.3.2.2 Pharmacophore searching

Where multidimensional QSAR incorporates spatial descriptors to construct a reliable mathematical model, pharmacophore based virtual screening methods construct and represent models as ensembles of the common essential steric, electronic and function determining features that optimize supramolecular interaction or the pharmacological profile (Langer et al. 2006; Wolber et al. 2008). The understanding of pharmacophore modelling in drug design benefited from early ideas of ligand-receptor modulation highlighting that an effective drug needed to be of high specificity and complementarity with the receptor (Greene et al. 1994). Early ideas of pharmacophores were represented by the notion of privileged structures that were compounds that possessed substructures and motifs that were frequently associated with activity (Martin et al. 1973). Ensembles of motifs present in a collection of molecules can be considered as being pharmacophores if they:

I. Highlight functional groups that have common spatial dependent features that enable the design of new compounds with similar features and activities (Wang et al. 2017; Rush et al. 2005).
II. Can discriminate between enantiomers (Rognan et al. 1992).
III. Can discriminate between agonists and antagonists (Yin et al. 2016).

IV. Can account for inactive analogues of active privilege structures (de Lera & Ganesan 2016; Morphy et al. 2004).

Whereas early drug discovery identified pharmacophores from manual inspection, modern pharmacophores are constructed using computer algorithms from spatial features extracted from databases of flexible compounds (Leach et al. 2010; Martin et al. 1973). Computational tools used to derive pharmacophore models make use of either manual construction methods, automated perception approaches from multiple active structures, or receptor-based deduction strategies from crystallographic structures (Seidel et al. 2018). Software derived pharmacophores are typically a collection of high scoring models (hypotheses), that are generated from a search protocol that aligns features with structures during the elucidation process. The scoring function used in pharmacophore elucidation considers factors such as feature matching, selectivity, overlap and strain. The PHASE algorithm makes use of the scoring function in Eq. 17 to rank pharmacophores (Dixon et al. 2006),

$$score = F + w_v V\ - w_e E\ + w_m^{M-1}\ + w_s S \qquad (17)$$

A quality measure, F, is used to relate the angular and distance alignment of the features in a constructed pharmacophore hypothesis with the features on a reference molecule. In ligand-based virtual screening, the reference molecule may be an active compound or an arbitrarily chosen rigid molecule. Volume alignment, V, is a measure of average volume overlap of the non-reference structures used to derive the hypothesis, with the volume of the reference molecule. The strain energy, E, is an estimate of the energy difference of the lowest energy conformer that was used to construct the hypothesis and the energy of the reference conformer. The number of compounds in the database that matched the particular hypothesis, M, is a reflection of how well the compounds in the database are represented by that particular hypothesis while the selectivity term, S, reflects the fraction of molecules in a

decoy database that matches the hypothesis. A larger S term penalizes the hypothesis as it decreases the specificity of compounds that match the hypothesis (Dixon et al. 2006; Leach et al. 2010). Whereas in algorithms such as PHASE or comparative molecular field analysis (CoMFA) user defined weightings are used, w, to alter the emphases placed on individual terms, Pareto ranking may be used in order to reduce user-defined biases on the choice of pharmacophore hypothesis (Cramer et al. 1988; Cottrell et al. 2004).

Pharmacophore models built using peptide molecules benefit from pharmacophore algorithms that account for flexibility during their hypothesis elucidation. Algorithms use least-squares fitting to align features from pre-generated conformations or they may generate the thermally accessible conformers on-the-fly during the alignment process (Wolber et al. 2008). Hypothesis elucidation is thus dependent on the exhaustiveness of the conformational sampling, similarly to how structure based virtual screening methods also depend on good conformational sampling (Dixon et al. 2006). Where structural data is present for the target, methods such as the multicopy simultaneous approach, the GRID molecular interaction fields and the hydration-site-restricted methods generate pharmacophore models that incorporate receptor based features in their hypotheses (Joseph-McCarthy & Alvarez 2003; Hu & Lill 2012; Cross et al. 2012). The site identification by ligand competitive saturation (SILCS) method makes use of MD simulations to construct a pharmacophore hypothesis that considers protein flexibility and desolvation (Yu et al. 2015). Either hypotheses built from conformational content of active ligand databases or the target structures rapidly screen through compound databases in a manner analogous to structure based virtual screening (Shin et al. 2016). Owing to the abstractive nature of the pharmacophore models, the results from virtual screening using these hypotheses are used to generate focused libraries prior to high-resolution, low-throughput screening (Dixon et al. 2006; Leach et al. 2010).

By leveraging on algorithms that allow for rapid compound enumeration and intelligent interrogation (structure or ligand based), virtual screening approaches are able to populate enriched, activity-focused libraries of therapeutically relevant peptides (Schneider 2013; Schneider & Fechner 2005). Coupling structure based approaches with ligand based approaches allows virtual screening pipelines to have greater efficiency (Anighoro et al. 2016; Reutlinger et al. 2014). Evaluation of how well virtual libraries (derived from different approaches) sample the universal chemical space requires sophisticated mapping approaches. These approaches have to identify distributions of chemical features such as shape, polarity or chemical functionality (Wang et al. 2014; Gütlein et al. 2014). Owing to the degree of variation, these distributions are mapped using clustering algorithms incorporating multivariate principle component analysis (PCA) which expresses the variance of the combination of computed feature properties within a multi-dimensional space (Papa et al. 2009; Stein et al. 2006). The positional information possessed in these representations of chemical space allow relationships that enable rapid compound testing of bioactivity to be built (Ruddigkeit et al. 2013; Rush et al. 2005).

2.4 Study Case

Despite the decline in productivity of R&D within the pharmaceutical sector, surveys of clinical trials show a growth in the focus of biotechnology-derived agents such as antibodies, recombinant proteins and synthetic peptides (Rask-Andersen et al. 2014). By targeting drug-targets classified as being within the undruggable genome, these agents are providing novel ways of treating Alzheimer's disease, appetite regulation, diabetes, inflammatory diseases and tumours. (Hopkins et al. 2014). Targets are deemed as druggable if they possess protein domains that are amenable to perturbation by drug-like molecules (Russ & Lampel 2005; Orth et al. 2004). The growth in interest of beyond rule of 5 drugs (bRO5) is attributed to the

observations that the shape and size of binding sites impacts on the druggability of the target. Molecular entities must balance conformational flexibility with rigidity, and also balance ligand affinity with efficiency (Doak et al. 2015; Díaz-Eufracio et al. 2018). The tunability of peptides and advances made in their synthesis place them as starting points to be agents against difficult to drug binding sites (Zorzi et al. 2017). Our study focuses on the development of peptide based macrocycles for applications in cardiovascular diseases.

2.4.1 Cardiovascular Diseases

Growing worldwide mortality as a result of cardiovascular disease (CVD) has prompted the re-evaluation of cardiovascular risk assessment strategies in order to reduce the incidents of CVD (Leening et al. 2016; Recio et al. 2017). In South Africa the hidden menace of CVD accounts for over 195 deaths per day (Maredza et al. 2017). Treatment of CVD involves the prevention of vascular complications through the use of angiotensin converting enzyme inhibitors, angiotensin II reversing blockers, anticoagulants, cholesterol-lower statins, beta-blockers and anti-inflammatory NSAIDs and glucorcoticoids (Cushman et al. 1977; Sharp et al. 2015; Kassler-Taub et al. 1998; Hetzel & Sucker 2005; Desai & Sabatine 2015; Kernis et al. 2004; Recio et al. 2017). Vascular complications arise from hyperglycaemia, hypercholesterolemia and inflammation which induce the production and release of pro-inflammatory cytokines, resulting in atherothrombosis and atherosclerosis (Fuster et al. 2005; Libby et al. 2011). The inflammatory response cascades may in turn decrease the diameter of the coronary vessel causing an acute myocardial infarction (Virmani et al. 2000; Tabas 2010; Merched et al. 2008). Management of CVD is intended to reduce the likelihood of atherothrombosis and atherosclerosis due to cholesterol and lipid accumulation, the leading cause of myocardial infarction and strokes (Law et al. 2003; Ulven et al. 2016; Shepherd et al. 2008; Pignone et al. 2000).

The symptoms of the onset of myocardial infarction include persistent discomfort experienced on the shoulder, arm, back, neck and jaw as well as nausea, shortness of breath and fatigue (Steg et al. 2012). If untreated and undetected the myocardial infarction may result in heart failure or cardiac arrest (Thygesen et al. 2018). During a suspected myocardial infarction, an electrocardiogram (ECG) is used to differentiate between the two mechanisms that account for the onset of an myocardial infarction in order to determine the treatment intervention that will be employed (El-Menyar et al. 2011; Steg et al. 2012). In the case where ST segmentation elevation myocardial infaction (STEMI) is observed from the ECG trace, treatment is through the use of percutaneous coronary intervention (PCI) that includes coronary angioplasty is recommended in order to restore arterial blood flow inducing repurfusion of the heart tissue (Steg et al. 2012). If the ECG trace exhibits ST segmentation depression (non–ST elevation), treatment of the myocardial infarction is managed by the use of anticoagulants and bloodthinners such as heparin (Hetzel & Sucker 2005). Complications during STEMI treatment results in patients developing chronic heart failure. This requires careful monitoring due to the increased risk score that predicts a death by myocardial infarction of 19% after 5 years (Fox et al. 2010). A correlation between the histological size of myocardial necrosis and plaque formed as a result of the repurfusion during STEMI and the severity of the myocardial injury was shown to exist (MAROKO et al. 1971).

Conditioning interventions during and after STEMI treatment through the use of cardio-protective strategies have been shown to reduce the effects of repurfusion injury improving the clinical outcomes of patients experiencing myocardial infarction (Zhao et al. 2003; Vinten-Johansen & Shi 2011; Hausenloy 2013). Early endogeneous preconditioning strategies include the use of short cycles of ischemia and repurfusion. This has an effect on mitochondrial activity, energy production, pH sensitivity, charge (mediated by Ca+ concentration) and the

presence of reactive oxygen species in the myocardial tissue (Hausenloy 2013). Early extensions of the preconditioning protocols included remote-preconditioning, where ischemia and repurfusion cycles were conducted in organs distantly related to the cardiocyte cells (Przyklenk et al. 1993). The mechanisms of this successful intervention are not fully understood although observations suggested the existence of a cardioprotective factor (Gho et al. 1996).

Mitochondria have been targeted by post-conditioning protocols that regulate their control of homeostasis directly or indirectly through their interactions with heat shock protein 90 (HSP90) (Camara et al. 2011; Zhong et al. 2014). In these approaches, pharmacological interventions or reperfusion cycles are performed at the onset of reperfusion with similar outcomes of reduced size of infarction (Hausenloy 2013). Mounting evidence in support of mitochondrial activity in conditioning protocols led to the development of cardioprotective agents that target mitochondrial activity. This has led to the identification of connexin and sarcoplasmic reticulum as therapeutic targets for the regulation of pro-apoptotic pathways (Santillo et al. 2016). Targeting these effectors and pathways has an impact on the infarction size and extent of repurfusion injury contributing to a reduction in the incidents of chronic CVD improving patient quality of care (Steg et al. 2012). Agents such as nitrates, nitroprussides, calcium channel blockers, beta blockers, adenosine, nicorandil, atrial natriuretic peptides, angiotensin-converting enzyme inhibitors, statins, volatile anaesthetics as well as analogues of cyclosporine have been explored. These agents have ability to regulate pro-apoptic processes and coronary flow, impacting the growth of the myocardial infarction (Nusca et al. 2010; Al-Mallah et al. 2006; Kitakaze et al. 2007; Piot et al. 2008; Marzilli et al. 2000; Russo et al. 2004; Peter et al. 1978; Taniyama et al. 1997; Amit et al. 2006; Ambrosio et al. 2010). This study focusses on cyclosporine agents in this regard.

2.4.2 Cyclosporine-derived cardio protective agents

Numerous debates surround the clinical potential of cyclosporine-derived interventions in alleviating the proliferation of infarctions that arise during reperfusion treatment of STEMI (Heusler & Pletscher 2001; Dube et al. 2012; Ottani et al. 2016). Despite evidence from literature, an American College of Cardiology study showed that there was no clinical benefit of cyclosporine (CsA) over the placebo in the spread of the infarction during percutaneous coronary intervention (PCI) reperfusion procedures (Ottani et al. 2016). Although this evidence may appear to disprove the clinical benefits of CsA they actually confirm observations from an earlier study that showed that CsA has a competitive affinity for cytosolic cyclophilin D (cyl-D)(Dube et al. 2012). However, also in this study mitochondrial-targeted CsA (mtCsA) was shown to have increased cytoprotective capacity in rat cardiomycetes. The rational for this is supported by evidence of CsA's role in the prevention of the formation of the non-selective mitochondrial permeability transition pore (MPTP) necessary for post-ischemic homeostasis (Nazareth et al. 1991; Ahlbach et al. 2015). Formation of the MPTP is known to destabilise the proton electrochemical gradient vital for mitochondrial function and energy production, which triggers a cascade of processes which result in irreversible cell death (Kwong & Molkentin 2015; Nazareth et al. 1991). The contrast in observations by Ottani et al. (2016) and Dube et al. (2012), warrant further investigations as to whether the mitochondrial targeting of CsA would be more effective in therapy during PCI.

The wealth of experimental observations offered from the Dube et al. (2012) study offers us a unique opportunity to perform a computational search for cyclosporine analogues that have improved potency and selectivity for mitochondrial cyl-D. Macrocycles such as CsA provide a unique challenge in computational studies that aim to accelerate the drug discovery process.

This challenge arises from the difficulty in the conformational sampling required to identify conformations relevant for interrogations of their potency and selectivity (Kessler et al. 1996; Anighoro et al. 2016). As highlighted above, macrocyclic systems have gained attention in recent years as a source of under-explored molecular entities due to their high affinity for previously undruggable targets (Allen et al. 2016; Marsault & Peterson 2011; Sperandio et al. 2010). The sites on these targets are typically solvent exposed with flat surfaces and thus compounds with large contact points are favourable (Rask-Andersen et al. 2011; Dandapani & Marcaurelle 2010). Macrocyclic scaffolds are ideal because the cyclized nature restricts the torsional degree of freedom contributing to a lower entropy and thus increased binding Gibbs free energy (Rask-Andersen et al. 2014; Hewitt et al. 2015; Bockus et al. 2013).

2.5 Aims and Objectives

The aim of this study was to enhance the precision of library-based virtual screening approaches, leveraging on the biocompatibility, tunability and diversity of peptide derived scaffolds. The incorporation of computational approaches in the drug discovery pipeline has the potential of speeding up and reducing the cost of early discovery stages by prudently expanding and searching through the chemical and conformational space probed by hit candidates. The biggest limitation of virtual screening is difficulty identifying the bioactive conformations of constrained macrocycles. Our thesis is that improving the ability of these approaches to identify peptide macrocycle bioactive conformations will optimize the precision and efficiency of early drug discovery stages. If this is the case, then the identification of highly potent biocompatible clinical candidates will come earlier during development.

In order to realise these aims, our objectives were:

1. To develop a compound enumerator that could be used to populate conformation-laden virtual libraries of cyclic peptides within a confined sequence space.
2. To incorporate high-resolution and exhaustive sampling protocols that would identify the conformations that account for the essential dynamics of a peptide macrocycle. These conformations improve the precision of conformation search steps critical to virtual screening.
3. To coalesce acceleration and compression routines in order to populate a conformation laden virtual library that exhaustively probes the chemical space of a confined cyclic peptide sequence space within a reasonable time.
4. To demonstrate the impact of our protocol on the search for cardio-protection agents with improved selectivity and potency for cyclophilin-D using virtual screening techniques.

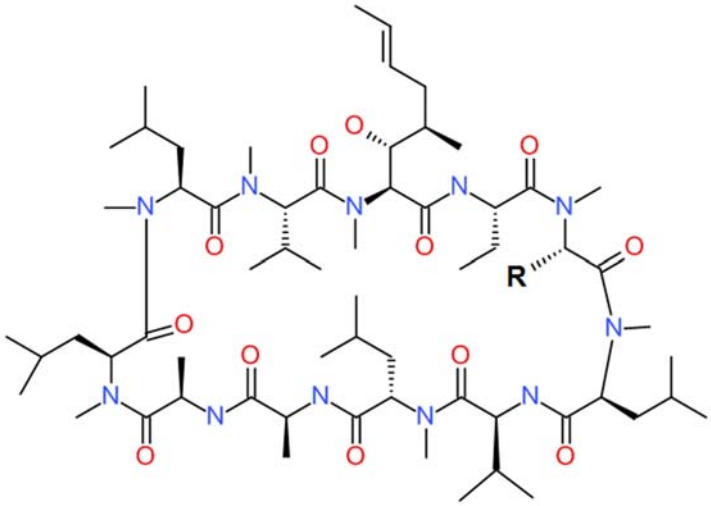
Chapter 3: Illuminating the Cyclophilin-D perturbation by cyclosporine and its mitochondrial targeted analogues

3.1 Introduction

Organelle specific targeting of cyclophilin-D (cyl-D) has been shown to increase the cytoprotective capacity of cyclosporine A (CsA) macrocycles in rat cardiomycetes cells (Dube et al. 2012). CsA inhibition of cyl-D prevents the formation of the non-selective mitochondrial permeability transition pore (MPTP). Formation of the MPTP is known to destabilize the proton electrochemical gradient that is vital for mitochondrial function (Kwong & Molkentin 2015). Disrupting mitochondrial energy production triggers a cascade of processes which result in irreversible cell death (Nazareth et al. 1991). Preventing the formation of the MPTP by targeting cyl-D mitochondrial provides the rationale for the development of mitochondrial targeted CsA (mtCsA) (Nazareth et al. 1991; Ahlbach et al. 2015). Enhancing the concentration of CsA present in the mitochondria has the logical benefit of reducing off target interactions and increasing drug present for interaction with cyl-D.

Although convincing biochemical data regarding the effects of CsA and mtCsA on cyl-D exists, the absence of structural data leaves many questions unanswered (Dube et al. 2012). These questions relate to the influence of the mitochondrial targeting groups on the drugging effect and the relation of these groups to the cytoprotection observed. An adjacent but related question relates to the possible improvements to the organelle specific targeting of cyclophilin-D: *Is it possible to design macrocyclic analogues that have stronger interactions with the cyclophilin-D target than CsA?* Answering these questions informs the design of more potent hit compounds for application as cardio-protective agents. The hypothesis that we aim to test in this Chapter is: “Do mtCsA and CsA have **similar** mechanisms of binding to cyl-D?” We aim to disprove this hypothesis through the application of molecular dynamics

simulations on reasonable models of the cyclophilin-D target in complex with the nascent CsA ligand (cyl-D-CsA) and the pharmacodynamically optimized analogue mtCsA (cyl-D-mtCsA).



Cyclosporine A

CsA **R** = **H**

mtCsA **R** = Mitochondrial targetting group

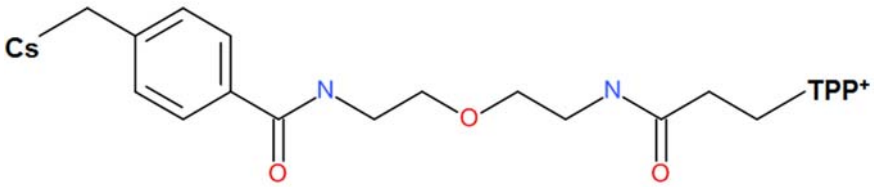


Figure 3.1: Cyclosporine analogues. *CsA and the mitochondrial targeted derivative (mtCsA)*

Simulation studies of biomolecular systems have gained attention through the ubiquity of open source molecular dynamics (MD) simulation software packages and the increased accessibility of high-performance computing architectures to end-users (Wu et al. 2015; Apol

et al. 2010). Typically, MD simulations are informed by crystal structure data, and provide the time evolution of the system's components, as a goal to the understanding of structure-function relationships (Teoh et al. 2014; Ribeiro & Ortiz 2015). Analysis of structural data over time gives evidence of molecular processes and imbues the practitioner with an armoury of evidence for why these processes occur (Kuhn et al. 2016). In this study GROMACS (Berendsen et al. 1995) was used to simulate the time evolution of the mtCsA analogue of CsA in complex with cyl-D (Fig. 3.1). The cyl-D-CsA crystal structure(2Z6W) was used as a reliable starting point for the simulations of CsA and its analogue under physiologically relevant conditions (Kajitani et al. 2008). High-throughput calculations of binding energies of protein-ligand conformations at each frame of the time-based trajectory, solving the MMPBSA equations for each, allowed for the interrogation of time resolved protein-ligand interactions (Miller et al. 2012; Kumari et al. 2014). The use of alchemical simulations with pathways defined by dynamic pulling simulations and the application of reliable intermediate state averaging through the WHAM analysis was used to enhance the precision of the protein-ligand interactions (Kumar et al. 1992; Procacci 2017).

3.2 Methods

3.2.1 Preparation of the CsA ligand topologies

CsA ligand coordinates were obtained from the X-ray crystal structure of the cyl-D-CsA complex (PDB: 2Z6W)(Kajitani et al. 2008). The small molecule was protonated at pH 7.4 using OpenBabel 2.3.2, prior to the generation of ligand topologies using the antechamber python parser interface (acpype v. 2012-09-13). Within acpype, the *bcc* MOPAC module calculated atom specific charges of 196 CsA ligand atoms. Minimization of the ligand system occurred after 500 steps using a steepest descent algorithm. The acpype topology generation procedure included 500 steps of restrained molecular dynamics using the Generalised Amber Force field (GAFF)(Wang et al. 2004).

3.2.2 Preparation of the mtCsA ligand model and topologies

Due to the absence of a mitochondrial targeted CsA crystal structure, the CsA crystal structure was mutated in order to derive the starting point for a model of the mtCsA-cyl-D complex. Discovery Studio (version 2017) was used to build the mitochondrial targeting triphenyl phosphine functional group prior to geometry optimisation of this functional group, first at the AM1 semi-empirical level, and finally at the HF/6-31G(d) level within Gaussian 09. The resulting vacuum optimised group was fused to CsA by substituting the appropriate hydrogen within the N-methyl group of CsA sarcosine with our optimised TPP fragment (at the appropriate bond length and with the appropriate angle and dihedral). This sarcosine modified CsA provided the mitochondrial targeted CsA (mtCsA) working model. In order to obtain ligand topologies of mtCsA using acpype, use of the *gas* modules allowed computation of atom specific charges (*bcc*/mopac was unable to provide these, given the ligand size of 269 atoms). Energy minimisation was satisfied after 72,000 steps before 500 ps of restrained dynamics were allowed as part of the acpype topology generation procedure.

3.2.3 Preparation of the Protein-Ligand complex

The GROMACS tool pdb2gmx generated the topologies of the protonated protein using the amber03 force field parameters. The protein and ligand structure files were merged into a single structure file of the complex before being placing at the centre of a cubic box with dimensions of 6.63 nm (cyl-D-CsA), 7.44 nm (cyl-D-mtCsA). The systems were solvated with 8,999 spce water molecules for cyl-D-CsA and 12,922 molecules for cyl-D-mtCsA. Ionic neutralisation was achieved by replacing 5 solvent molecules with 5 Cl^- ions in the cyl-D-CsA complex while 6 solvent molecules were replaced by 6 Cl^- ions in the cyl-D-mtCsA. Rapid (20 ps) steepest descent minimization was performed on the neutralized systems prior to temperature and pressure coupling.

3.2.4 Molecular Dynamic simulations

The modified Berendsen isotherm, the v-rescale thermostat, coupled the temperature of these systems to a 300 K heat bath during a protein-restrained simulation of 10 ps. The Parrinello-Rahman pressure-coupling algorithm was used to couple the pressure of these systems to 1 bar during a 10 ps simulation. In order to capture the stability of the complexes, simulations of 140 ns were performed on each complex. During the course of the simulations, force calculations consumed 60 % of the computation time with 16.7 % of the time spent on PME mesh computation. The remaining computation time was spent on neighbour searching and constraint calculations. Analyses of the trajectory outputs were performed in order to determine atomic level differences between the interactions of the original CsA ligand within its cyclophilin-D target and the modified mtCsA with the same target.

3.2.5 MMPBSA Analysis

Guided by analysis of the protein-ligand complexes, snapshots from a set of uncorrelated trajectories from a stable region between 90 - 120 ns of simulation were obtained for binding

energy analysis. The vacuum potential energy change and the sum of polar and non-polar solvation energies accounting for the desolvation energy terms were computed using the *g_mmpbsa* scripts in a single-trajectory manner. The polar solvation parameters for the positive and negative ionic concentrations were set at 0.150 M, the solute dielectric constant at 2 and the solvent constant at 80 while the vacuum dielectric constant was set to 1 with a solvent probe radius of 1.4 Å. A linear Poisson - Boltzmann equation was solved using APBS at 300 K. The solvent accessible volume (SAV) non-polar solvation model was calculated and used for computing binding free energies maintained during the region of interrogation (1.29 Å probe radius, 0.234 kJ mol^{-1} $Å^{-3}$ pressure coefficient, 0 kJ mol^{-1} offset and 20 quadrature grid points).

Interaction energies computed by our MMPBSA analysis did not include entropy contributions to the binding energy. The introduction of the Rosamine-linked-TPP^{+} functional group on mtCsA resulted in significant differences in the contributions of ligand-solvent interactions for the two ligands. Due to this disparity, decomposition of the binding energy on a per-residue basis enabled a reliable critique of ligand binding during the simulation. Statistical comparisons enabled magnification of significant changes in the ligand-receptor interactions highlighting whether a similar binding mechanism between the two ligands was plausible.

3.2.6 Dynamic pulling simulations

A reliable estimate of the energy cost of binding obtained from dynamic pulling simulations with the weighted histogram analysis method (WHAM) overcome the limitations in the estimation of the protein-ligand interaction. Instead of relying on the free-energy difference between ligand bound and unbound states computed from MMPBSA, coupling pulling dynamics with exhaustive averaging has been shown to be suited for constitutionally different ligands (Truong & Li 2018; Kumar et al. 1992; Lemkul & Bevan 2010).

We performed a dynamic pulling experiment on starting structures obtained after 100 ns of simulations. These starting structures were present within the stabilised region of the simulation. The biased pulling simulations and umbrella sampling experiments were implemented within the GROMACS package. By coupling umbrella sampling of the biased trajectory with a WHAM analysis, averaging of the binding event can be achieved (Kumar et al. 1992; Lemkul & Bevan 2010). This method relies on the precision of the starting complex and the setup of the pre-defined pulling coordinate, to provide an ensemble of "representative structures" extracted along the trajectory. Performing unbiased and unrestrained simulations of these structures simulates and umbrella sampling of the configurational space along the trajectory of the pulling coordinate. Recent studies have highlighted the reliability of these computations on the estimation of relative difference in the binding free-energy $\Delta\Delta G_{bind}$ of different ligand systems (De Vivo et al. 2016).

For our study a pulling force of 1000 $kJ \cdot mol^{-1} \cdot nm^{-2}$ was exerted on the two ligands with a pulling rate of 10 $nm \cdot ns^{-1}$ maintained along a direction-periodic single trajectory co-ordinate space. The Nose-Hoover isotherm algorithm was used to maintain a system temperature of 310 K. Structure restrained pulling simulations of 4 ns were sufficient to allow complete dissociation of ligand from target active site at 4 nm. Coordinate positions were computed every 2 fs while velocities and coordinates were recorded every 1 ps. Coordinates were extracted from the full pulling trajectories with a time gap of 50 ps. The distances between the ligand centres of mass (COM) and the protein COM for each extracted coordinate file was computed and stored. Representative structures with a distance gap of 0.15 nm were set aside as initial structures for the potential mean force calculation (PMF) from umbrella sampling and the WHAM analysis of the ensemble of simulations. To prepare for simulations, the 19 (CsA) and 18 (mtCsA) representative structures were coupled to 1 bar isobars and 310

K isotherms individually before performing 100 ps of minimization during a restrained simulation. Umbrella sampling of the coordinate space consisted of unrestrained production simulations of 500 ps for each representative system. By performing a WHAM analysis across the coordinate space, extraction of the weighted average energy of perturbation for each ligand allowed for the reliable interpretation of the ligand affinity to be obtained.

3.3 Results and Discussion

3.3.1 Molecular Dynamic simulations

3.3.1.1 Structural deviations and residual fluctuations

Root mean square deviation (RMSD) is one of the most informative indicators in establishing whether the given protein-ligand complex is stable and close to the likely experimental structure during simulation. We compared the total structural deviation of the CsA and mtCsA ligands individually through RMSD. Due to the increased flexibility of the TPP functional group, the mtCsA had larger structural deviations than the CsA ligand as expected (Fig. 3.2 A). A less obvious difference was observed between the deviations of the cyl-D bound to the different ligands. The mtCsA bound cyl-D had a slightly lower degree of deviation when compared to the CsA bound target (Fig 3.2 B). More evidence is required in order to determine whether a plausible case for an enhanced stabilization effect of mtCsA over the CsA ligand exists.

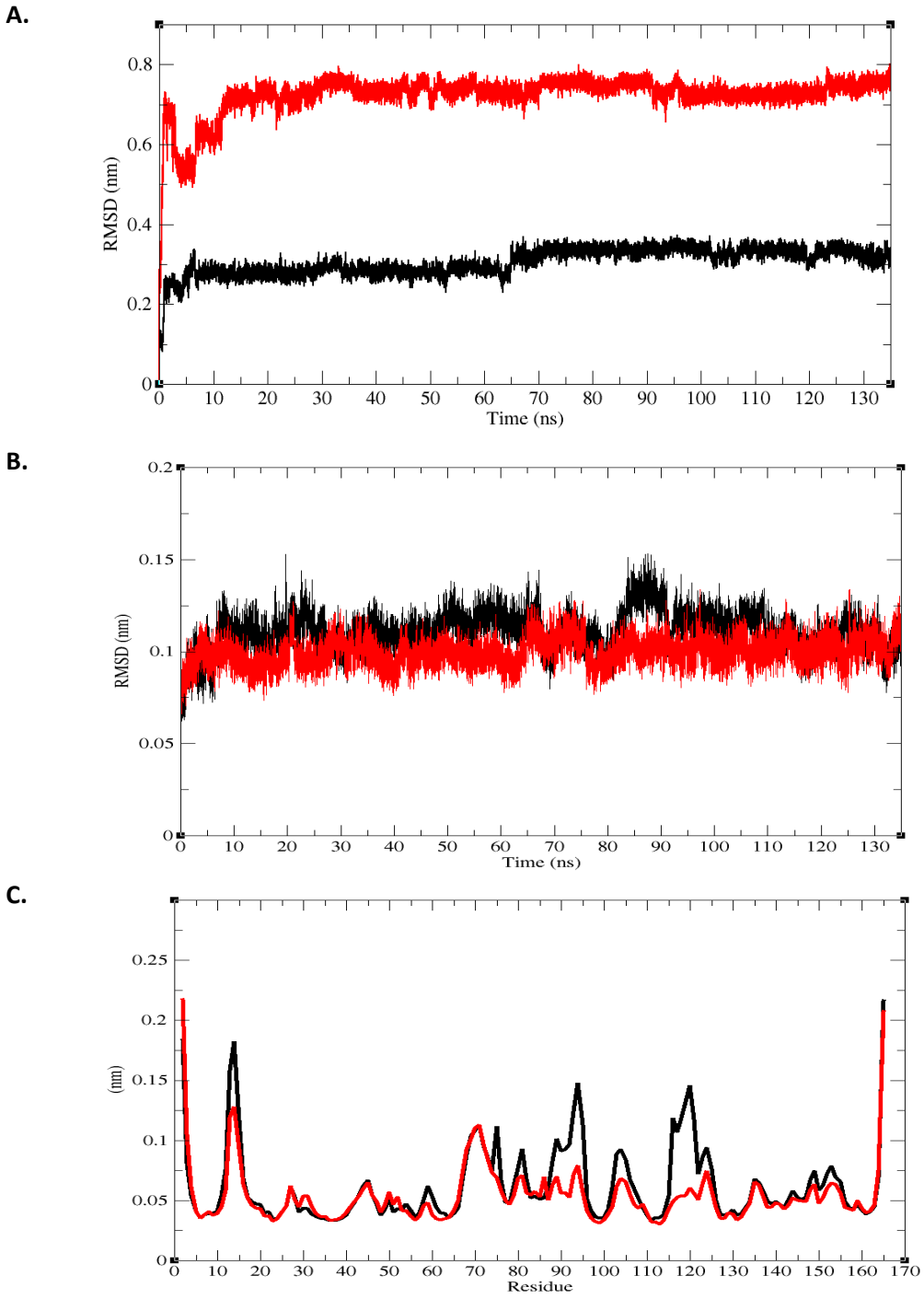


Figure 3.2: Plots of root mean square deviations (RMSD)*. RMSD plotted as a function of time obtained for (A)* ***Ligand*** *of CsA (black) and mtCsA (red). (B)* ***Protein*** *of cyl-D-CsA (black) and cyl-D-mtCsA (red), respectively. (C) Plots of root mean square fluctuations (RMSF) as a function of residue position of cyl-D protein in complex with CsA (black) and mtCsA (red).*

Vibrations around the equilibrium structure are not random and maybe influenced by local structure flexibility. To illustrate the average fluctuation of target residues during the simulation, the root mean square fluctuation (RMSF) of Cα atoms from their initial structure was plotted as a function of residue number (Fig. 3.2 C). The RMSF pattern of the residues within the targets bound to mtCsA and CsA showed differences in the fluctuations of specific residues.

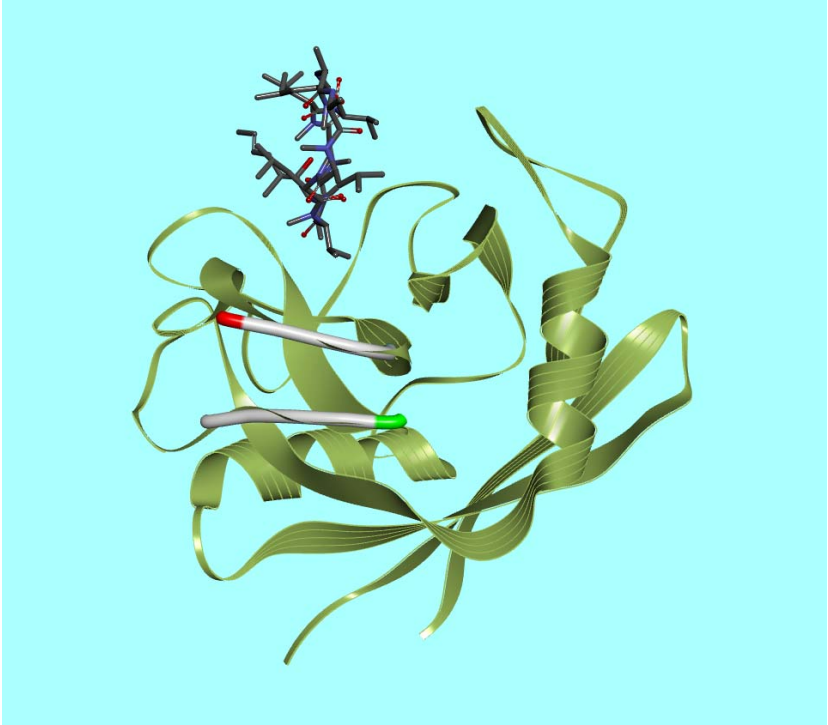


Figure 3.3: CsA (rod) ligand in complex with cyl-D (ribbon). *The location of the 90 – 95 residue loop with respect to the 115 – 121 residue loops are shown (pipe).*

The residues showing significant deviations in their fluctuations according to RMSF are: Lys 15, Phe 60, Gly 75, Arg 82, **Leu 90, Lys 91, His 92, Val 93, Gly 94, Pro 95,** Pro 105, **Cys 115, Thr 116, Ile 117, Lys 118, Thre 119, Asp 120, Trp 121.** The stretch of residues between position 90 – 95 and position 115 – 121 were in close proximity structurally (Fig. 3.3). It is possible that structural destabilisation was introduced between these residues in the presence of CsA but

absent in mtCsA. The MMPBSA study interrogated whether these differences were translated into differences in the contributions of active site residues and their decomposed energy pattern (Fig. 3.6).

3.3.1.2 Structure Compactness

The radius of gyration (R*g*) is a parameter linked to the structural compactness of a protein during MD simulation (Lobanov et al. 2008). The Rg for Cyl-D when bound to CsA increases, while it was found to be lower at some parts during the simulation when bound to mtCsA (Fig. 3.4 A). A stretch between 20 and 90 ns appears to reveal a difference in the structural compactness of the complexes. These differences are less significant after 90 ns. The solvent accessible surface area (SASA) is defined as the surface area of a protein which interacts with solvent molecules (Heffernan et al. 2015). An increase in the SASA reflects increased exposure of hydrophobic regions to solvent due to possible unfolding. The SASA of the cyl-D bound to the CsA and the cyl-D bound to the mtCsA do not appear to show any distinctions throughout the trajectories (Fig. 3.4 B).

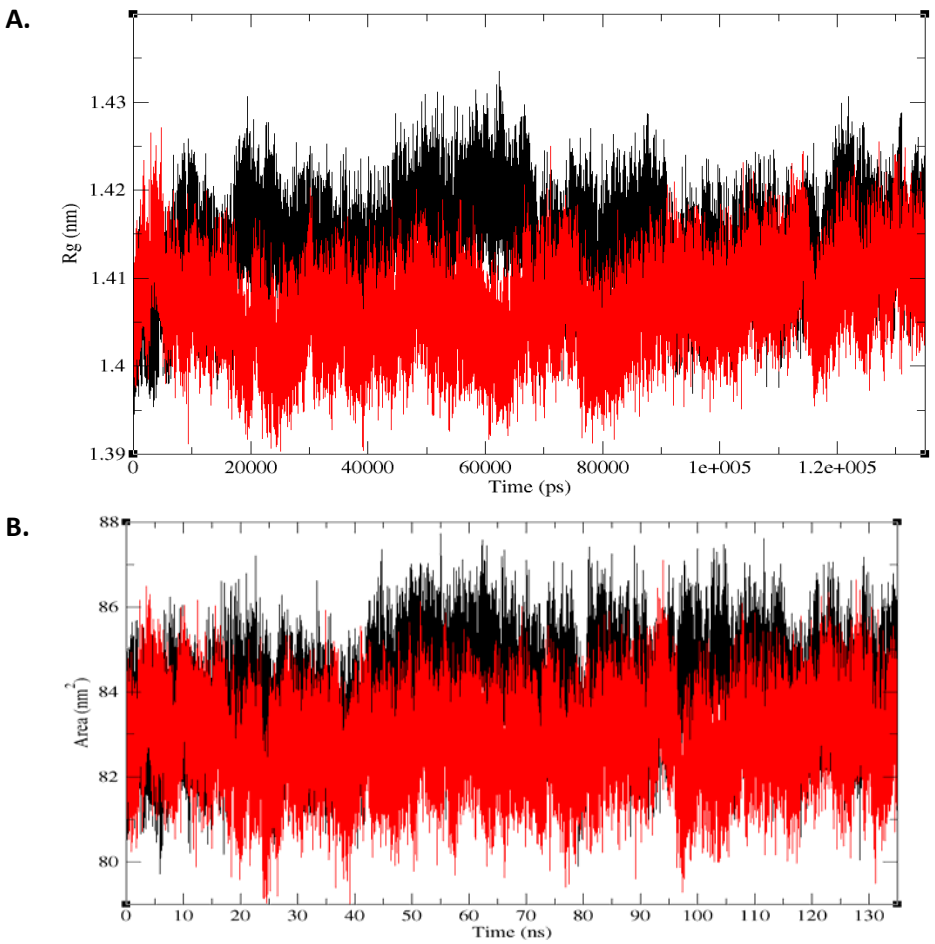


Figure 3.4: (A) Radius of gyration (Rg) and (B) Solvent Accessible Surface Area (SASA) as a function of time obtained for cyl-D-CsA (black) and cyl-D-mtCsA (red), respectively.

3.3.1.3 Hydrogen Bonding

The average number of hydrogen bonds formed between amino acid residues of the cyl-D with CsA and mtCsA were analysed during MD simulations. The predicted mean hydrogen bonds formed between cyl-D and CsA appeared to be more persistent than the bonds between cyl-D and mtCsA during the simulations. The hydrogen bond plots (Fig. 3.5) predict that the CsA binding to the active pocket of cyl-D would have more hydrogen bonds that would contribute to a stronger interaction. It was in our interests to interrogate the contribution of bonding interactions on the overall affinity of the ligands with the targets using the MMPBSA binding energy decomposition calculation.

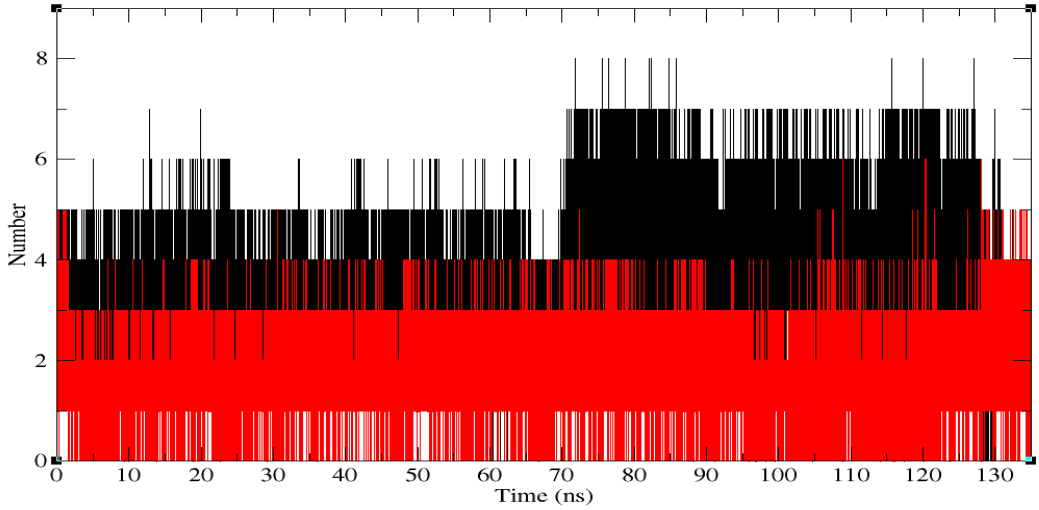


Figure 3.5: The number of hydrogen bonds formed by cyl-D with CsA (black) and mtCsA (red), respectively during MD simulations.

3.3.2 MMPBSA Analysis

Ligand binding energies during the simulations were performed using the MMPBSA approach in a high-throughput manner (Kumari et al. 2014). The *g_mmpbsa* scripts were developed to perform these MMPBSA based binding energy calculations within the context of the analysis of GROMACS trajectories. These scripts rely on the calculation of three energy terms in order to compute the strength of interaction of the ligand to the target in the present of solvent. The first term, ΔG_{MM}, is a force-field derived vacuum molecular mechanics term and is computed by addressing bonded and non-bonded terms of a complex in the bound and dissociated states. The solvation term addresses solvation energy changes that occur during complex association and is composed of both a polar electrostatic ΔG_{pol}, and a non-polar, non-electrostatic ΔG_{apol} potential (Miller et al. 2012). In order to avoid over-fitting, un-correlated snapshots between 90 and 120 ns timeframes were extracted for the binding energy calculations. The SAV parameters were used to compute the non-polar solvation energy of each snap shot and the mean binding energies were recorded (Table 3.1).

Table 3.1: MMPBSA Energy terms extracted from a collection of un-correlated snapshots extracted between 90 – 120 ns of simulations.

Energy Term	CsA (kJ mol^{-1})	mtCsA (kJ mol^{-1})
van der Waal	-0.004 (0.006)	-0.001 (0.001)
Electrostatic	0.722 (0.466)	-0.230 (0.134)
Polar Solvation	7.350 (21.126)	-9.013 (24.958)
Apolar Solvation (SAV)	-15.935 (95.010)	-29.348 (127.367)
Total Binding Energy	**-7.868 (96.807)**	**-38.592 (129.614)**

The mean binding energy contributions between the CsA (-7.868 kJ mol^{-1}) and mtCsA (-38.592 kJ mol^{-1}) ligands computed through the MMPBSA approach show a 390.5 % enrichment of the CsA binding energy by introduction of the mitochondrial targeting TPP moiety. This observation does not corroborate with the experimental data which shows a modest 0.5 % enrichment of affinity in mtCsA (Dube et al. 2012). Calculating the binding free energy change using the MMPBSA approach has limitations. It is heavily influenced by the unreliability of the parameters that describe solvent displacement, and the entropy contributions to free energy change is not accounted for (Kumari et al. 2014). These discrepancies insert irregularities in the computed binding free energy that mislead inferences based on this absolute energy difference.

The dominant energy terms that contribute to the inconsistent enrichment observed during the simulation are the descriptions of the solvation terms (polar 200 % and apolar 84 %). By observing that mtCsA had a larger surface area exposed to solvent than the CsA, it is probable that a misappropriation of solvent disruption could contribute to the disproportionately large enrichment due to binding affinity. Conclusions based on the absolute binding energy

computed from solving the MMPBSA equations in this case are unreliable due to the omission of solvent disruption accounted for by entropic contributions. Instead, to mitigate against solvent disruption terms and the inaccurate entropy contribution, the target response to the presence of ligand was interrogated. This response is monitored through a per residue decomposition of the binding energy contribution and performing statistical tests on the observed pattern to identify any significance in the pattern (Fig. 3.6).

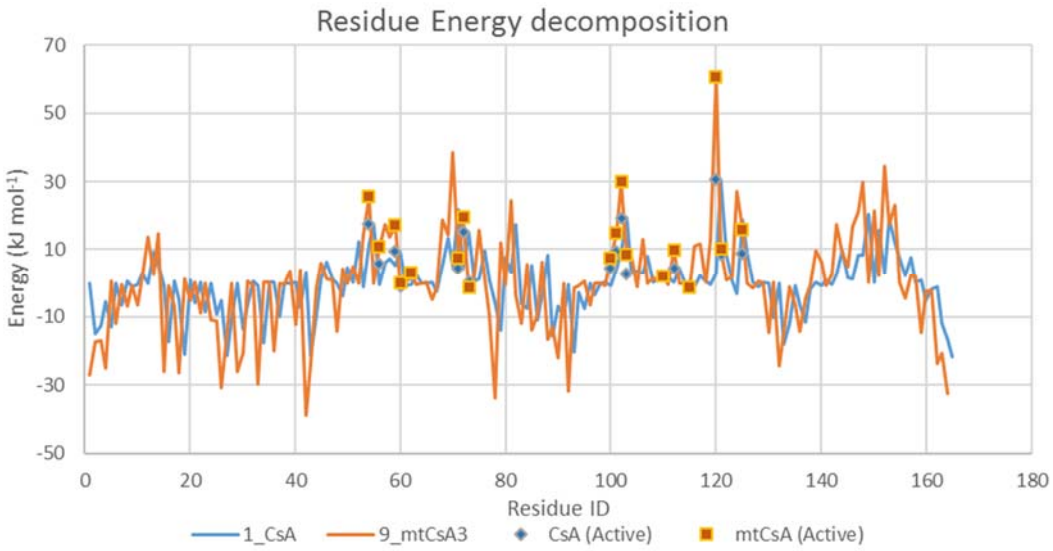


Figure 3.6: Residue energy decomposition of the MMPBSA computed binding energy for mtCsA and CsA ligands in complex with cyl-D.

Despite the differences in the MMPBSA binding energy, the energy decomposition pattern of cyl-D in complex with CsA is complementary to that of the cyl-D in complex with mtCsA. When a paired two-sample for means *t*-test was performed on the residue energy decomposition, we observed a Pearson Correlation of 94.7 % consistent with the trends observed from our residue energy decomposition plots. This strong correlation implies that the way in which the residues in the target in the different cases (drugging with CsA against drugging with mtCsA) vary with each other is not significantly different. The t-Critical two-tail statistic of 1.97 is

greater than the t-stat value of 0.014 (alpha 0.5) showing that there is no significant difference in the residue response to drugging in the two conditions. When contributions from active site residues are emphasised, a Pearson Correlation of 97.5 % shows that there are stronger correlations between host residues participating in the interaction in the different guests. These observations further re-enforce our claim that cyl-D perturbation due to CsA as a guest is comparable to that of mtCsA, contrary to the MMPBSA absolute binding energy inspection with misappropriated solvent terms.

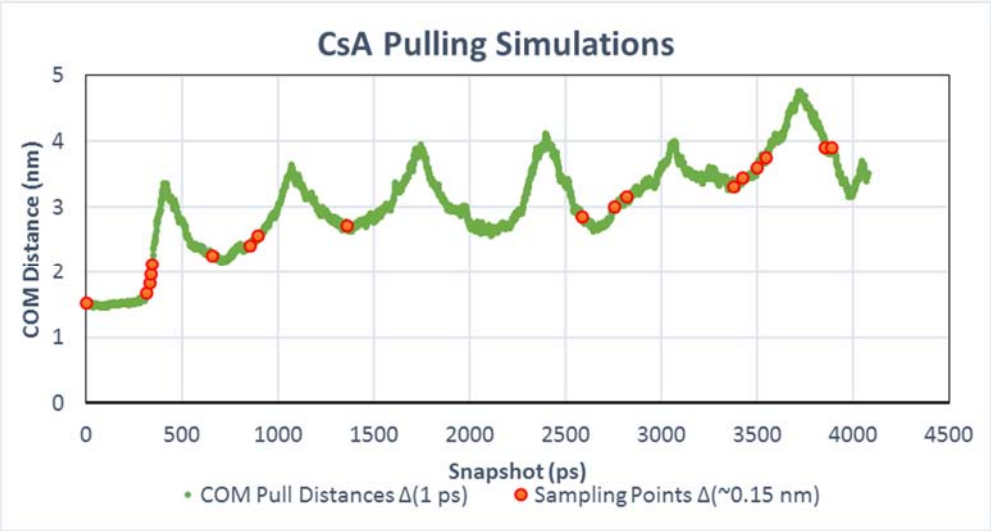
3.3.3 Dynamic pulling simulations

More precise binding estimations can be acquired by the application of biased simulation techniques such as dynamic pulling simulations. Umbrella sampling of the pulling trajectory implemented within GROMACS affords a reliable interpretation of the binding event. This is achieved by using the Weighted Histogram Analysis Method (WHAM) to efficiently average the ensemble of potential energies of perturbations sampled along the biased trajectory (Kumar et al. 1992; Lemkul & Bevan 2010). This method relies on the precision of the starting complex and the setup of the pre-defined pulling coordinate to provide an ensemble of "representative structures" extracted along the trajectory. Recent studies have highlighted the reliability of these computations on the estimation of relative binding free-energy $\Delta\Delta G_{bind}$ (De Vivo et al. 2016). This technique interrogated the differences in the binding energies between the mtCsA and CsA ligand complexes with cyl-D. A 10 $nm \cdot ns^{-1}$ pulling rate on the ligands allowed a sufficient COM difference between ligand and receptor after 4 ns of simulation. This allowed interaction thermodynamics to be estimated for the CsA and mtCsA ligands in complex with cyl-D.

From the pulling simulation snapshot analysis it can be seen that the larger mtCsA ligand had a steeper gradient of dissociation compared to that of the smaller CsA when a force of 1000

kJ mol^{-1}·nm^{-2} was exerted on the ligand (Fig. 3.7). This difference in response to pulling can be accounted for by the strength of the solvent attraction implicitly represented in our experiment, which marginally favours mtCsA dissociation over CsA dissociation.

A.



B.

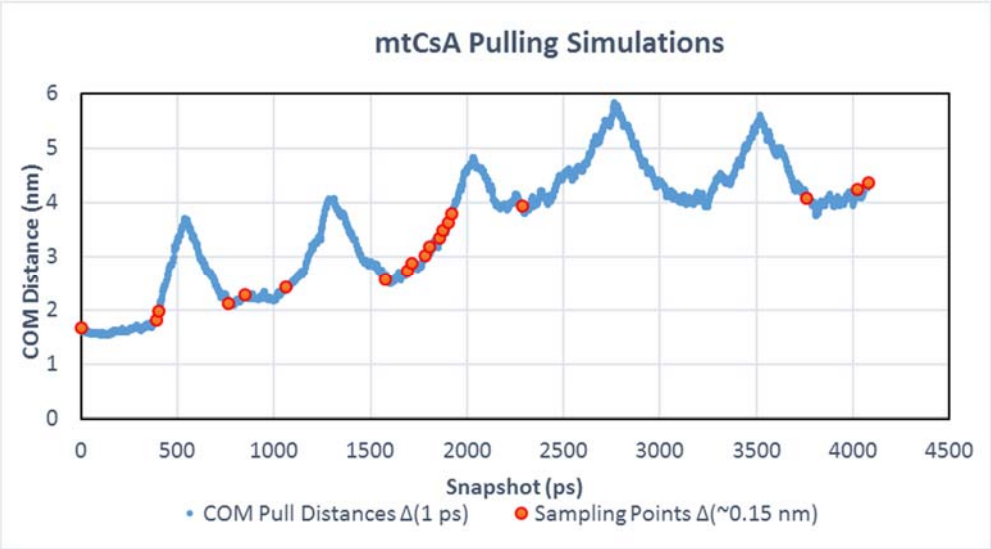
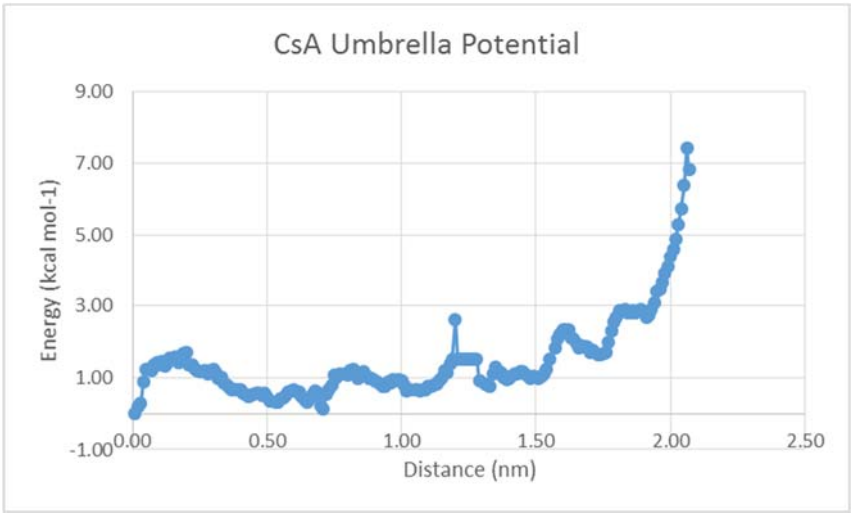


Figure 3.7: Pulling simulation snapshot analysis. *Windows for umbrella sampling were extracted from snapshots that were separated by an average of 0.15 nm across the Pull trajectory.*

A selection of snapshots that maintained an average COM distance separation of 0.15 nm from each other were extracted for the umbrella sampling experiment to derive a free-energy description of the dissociation pathway (Fig. 3.8).

A.



B.

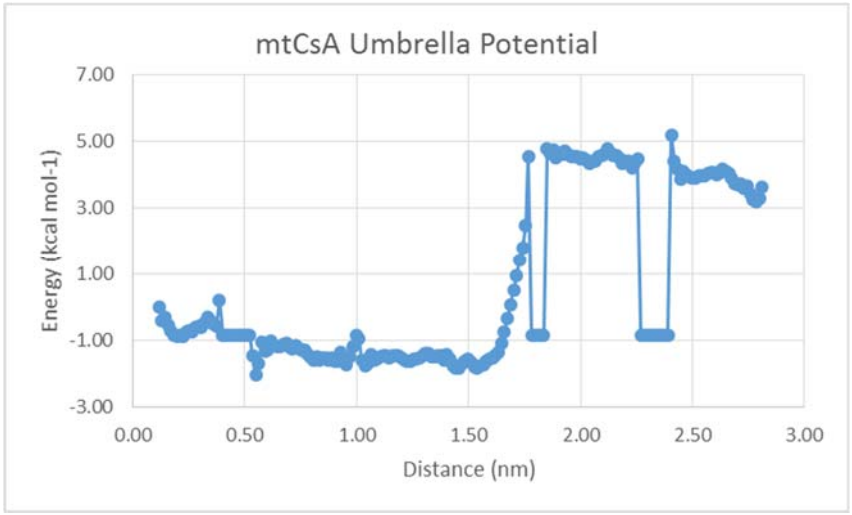


Figure 3.8: Map of the umbrella potentials from the WHAM analysis of unrestrained simulations of PMF calculations. *(A) cyl-D-CsA complex umbrella potential. (B) cyl-D-mtCsA complex umbrella potential.*

Due to the computation resources available, a maximum of 20 parallel simulations could be tethered together during the WHAM analysis. A COM pulling distance of 0.15 nm was chosen

in order to allow for a maximum of 20 simulations to be analysed simultaneously (Fig. 3.7). As part of the umbrella sampling experiment, 500 ps of unrestrained simulations of the representative ensembles sampled the coordinate space. Unrestrained simulations of these representative ensembles along the coordinate space traced the pulling event. Aggregating potential energies of ensembles along this pulling coordinate using molecular dynamics, allows us to estimate the binding cost from a WHAM analysis of these perturbations. Maps of the umbrella potentials obtained after the WHAM analysis plotted against the COM distance monitored during the simulation are shown (Fig 3.8). The difference between the extremes of the potential energy extracted from the WHAM analysis gives a more precise estimate of ΔG accounting for solvent contributions by implicitly modelling solvent behaviour. Umbrella sampling of the free energy of ensembles obtained from a pulling simulation using a WHAM analysis gave an energy cost of binding, ΔG, for CsA of -7.45 kcal·mol^{-1} and mtCsA -7.22 kcal·mol^{-1}. The binding energy estimated by this protocol corroborates with the assay data obtained by Dube et al. (2012). The starting coordinates of our pulling simulations were biased towards a single conformation. We were satisfied with using structures that lay within the stabilised region of interrogation extracted after 100 ns of unrestrained normal simulation. Although the maps of our potentials display artefacts in the path, we were only interested in the extreme values. The results we obtained gave binding energies within the range of the experimental observations.

3.4 Conclusions

The evidence from this study shows that the binding effect can be interrogated in the absence of crystal structure data of the mtCsA-cyl-D complex. There was no evidence from our study that suggested an impact of the TPP group on the mtCsA binding effect. An influence by TPP would motivate for a binding mechanism different to that of CsA. We have not been

successful in disproving the hypothesis that mtCsA and CsA have **similar** mechanisms of binding to the cyl-D target. The evidence gathered from our study showed that the introduction of the TPP mitochondrial targeting group did not have an impact on the binding effect of the mtCsA. It is thus plausible to assume that the experimental inhibition data obtained from binding assays was due to the direct interaction of the CsA pharmacophore group of both CsA analogues ligands on the cyclophilin-D target.

Analyses of unrestrained MD simulations predicted that the target residues could have different fluctuation patterns. A significant difference in the RMSD patterns of the ligands was attributed to the dynamics of the solvent exposed TPP functional group. Analyses of hydrogen bonding networks showed that there was a possibility that the CsA ligand maintained more Hbonds than the mtCsA ligand. Decomposition of the free energy of binding (ΔG_{bind}) on a per-residue basis allowed for a nuanced interpretation of the pattern of the receptors response to the ligand. Interrogating the binding interaction using target response expressed as a decomposition of the MMPBSA energy terms enabled a refined evaluation of the target response to ligand inclusion. Statistical tests of the energy decomposition response were unable to disprove the hypothesis that the introduction of the TPP group had no impact on the mechanism of binding of the mtCsA ligand. This approach circumvented irregularities arising from solvent displacement and entropy in describing the ΔG_{bind} obtained from the MMPBSA approach (Kumari et al. 2014). The propagation of biased simulations implicitly accounted for entropy changes and solvation effects in the estimation of binding energy (Lemkul & Bevan 2010). Assessment of the binding event through averaging of the perturbation space using a WHAM analysis of an ensemble of simulations for the two ligands with cyl-D allowed us to correlate our binding affinity with experimental data. We were able

to overcome limitations associated with entropy and solvent effects to obtain a reliable description of the strength of interactions of our ligands with their receptor.

Armed with this evidence we re-enforce that, contrary to the Ottani et al. (2016) findings, cyl-D perturbation is a plausible cardio-protection approach worthy of our pursuit. Our rationale hinges on the reliability of the Dube et al (2012) CsA mitochondrial targeting to minimize off-target effects. Our study highlighted that the improved dose response from mtCsA was due to mitochondrial accumulation and not due its enhanced affinity for cyl-D. It is our position that the search for novel cardio-protective agents with better selectivity and sensitivity for cyl-D than CsA derived agents must be pursued. We have defended the Dube et al. (2012) mitochondrial targeting by the TPP group as a method of optimisation while showing that opportunities for improvement are located in the design of peptide macrocycles inspired by CsA with stronger affinity and specificity to cyl-D. In this study, analyses of MD simulations, extraction of interaction energies from MMPBSA decomposition and calculations of the energy of binding from biased simulation provide evidence that the cyl-D-CsA and the cyl-D-mtCsA interactions occur through a similar mechanism. This was despite the presence of solvent exposed TPP functional group in mtCsA. These observations justify the library-based search for novel cardio-protection agents with reduced side effects and improved affinity for cyclophilin-D.

Chapter 4: Development of In-house Virtual Library Enumerator

4.1 Introduction

The rational design of novel peptide inhibitors benefits from knowledge gained through understanding of the structural complementarity of the proposed inhibitor and target receptor, and the potency and synthesizability of this novel peptide (Singh et al. 2015). If amino acid monomers are "words", then the large size of the "linguistically" accessible peptide search space makes it difficult to explore the plethora of structural relationships manually. A thought experiment that considers 11-membered peptides constructed from common, natural amino acid monomers spawns a search space of 25.6 Billion possible peptides. Computational modelling strategies that enumerate and screen virtual peptide libraries filter and therefore reduce numbers of peptides reaching test phases that depend on the *in vitro* screening of peptide based focused libraries (Loughney et al. 2011; Kuhn et al. 2016; Chatterjee et al. 2008). Focused libraries of therapeutically relevant peptides optimized for activity, synthesizability and potency also benefit from compound enumeration that considers structural relationships during virtual screening (Schneider 2013; Schneider & Fechner 2005).

Enumeration algorithms such as Schrödinger's MS Combi, the CDK-Taverna enumerator, ChemAxon's Plexus Suite Reactor and the cyclic peptide enumerator, CycloPS are recently-developed interventions (Truszkowski et al. 2011; Jeedimalla et al. 2015; Halls et al. 2013; Duffy et al. 2011). MS Combi boasts R-group enumeration, elemental enumeration and polymer enumeration within its suite (Halls et al. 2013). CDK-Taverna enables variable R-group functionality, atom variability and variable ring size in its enumeration functionality (Truszkowski et al. 2011). On the other hand, the Plexus suite provides options for scaffold and reaction based enumeration as well as extensive characterization of the virtual libraries

constructed (Jeedimalla et al. 2015). Relevant to peptide work, CycloPS combines the enumeration of its peptide libraries with a rules based triage to predict synthetic accessibility (Duffy et al. 2011). CycloPS has several options including the head-tail bond, disulfide bond, side-chain-to-side-chain bond, side-chain to N-terminal bond and side-chain to C-terminal bond cyclization options.

A major limitation of these enumeration protocols is their restrictiveness as many of the algorithms are protected by proprietary regulations. This limits their versatility and imposes restrictions on adopters of these protocols. Of interest in this study is that downfield virtual screening pipelines are limited by libraries that may be broad in terms of coverage of chemical space, however for which a single conformer is generated per chemical structure, and so the library explores limited conformational space. This places the entire burden of conformational search on the screening pipeline. The task of exploring chemical and conformational space is huge, and for virtual libraries to realize their full potential of scale in terms of their "linguistic accessibility" together with rich conformational content, enumerators require the recruitment of compression algorithms and acceleration schemes to achieve these ends. Flexibility is required if the tools are to be deployed in heterogeneous environments with storage, memory space and speed restrictions.

This chapter outlines our in-house developed enumerator that possesses this flexibility and utilizes compression algorithms and acceleration schemes to fulfill the ambitions for scale-up. The objective of the task was to build an enumerator capable of populating conformation-laden virtual libraries of cyclic peptides (within a confined sequence space).

4.2 Strategy Development

The aim of the work in this chapter was to develop a tool that will probe both the chemical and conformational space available to cyclic peptide systems through the enumeration of a conformation-laden virtual library. The collation of chemical and conformational content was attempted using two strategies. The first strategy was a tool that explored conformations **during** chain-growth (Prototype 1). The second strategy was a hybrid of conformational scaffold extraction followed by chemical enumeration, similar to the Plexus Suite and the CycloPS enumeration approaches (Prototype 2). This section outlines the construction of both enumeration Prototypes. The logic, implementation and pseudo-code of Prototype 1 and Prototype 2 are outlined, prior to description and analysis of the resultant conformation-laden libraries which each produces. A discussion of each strategy provides justification for the selection of the hybrid enumeration strategy (Prototype 2) over Prototype 1. This section concludes with a demonstration of the reliability, robustness and flexibility of the Prototype 2 implementation as regards the generation of a database of diverse scaffolds.

4.3 Prototype 1:

For Prototype 1, the strategy made use of enumeration and conformation searching simultaneously during the evolution of a peptide construction as part of the chain growth.

4.3.1 Logic

The rationale employed in the development of Prototype 1 was that, one could build a tool that could exhaustively search the conformational landscape of the backbone torsions of each amino acid while chain elongation was chaperoned. During this approach, each incoming amino acid had the three backbone omega (ω), psi (ϕ) and phi (ψ) torsions to rotate through (Figure 4.1). Elongation from the next incoming amino acid would occur at the terminal (carboxyl) end of each backbone rotamer. This process will repeat until termination. Termination is initiated in the event that the distance between the nitrogen atom in the

amino group of the first amino acid and the carbon atom in the carboxyl group of the last amino acid was within a specified tolerance that permitted cyclization. The torsions to be "searched" through were read from a list of sets of torsions defining the exact torsion values for a specified position within the peptide. For the incoming amino acid, the first torsion defined either a *cis* or *trans*-amide configuration (the energetically disfavoured *cis* arrangement was allowed). In the case where the incoming amino acid is the first amino acid of a linear peptide this position would only impact the terminal COOH conformation, while in a cyclized peptide this would have greater impact in terms of defining the *cis*/*trans* conformation of an amide bond. The second two positions each had 6 possible torsions defined as 60° rotations around their axis of rotation. Figure 4.1 illustrates the torsions available to influence the conformational search.

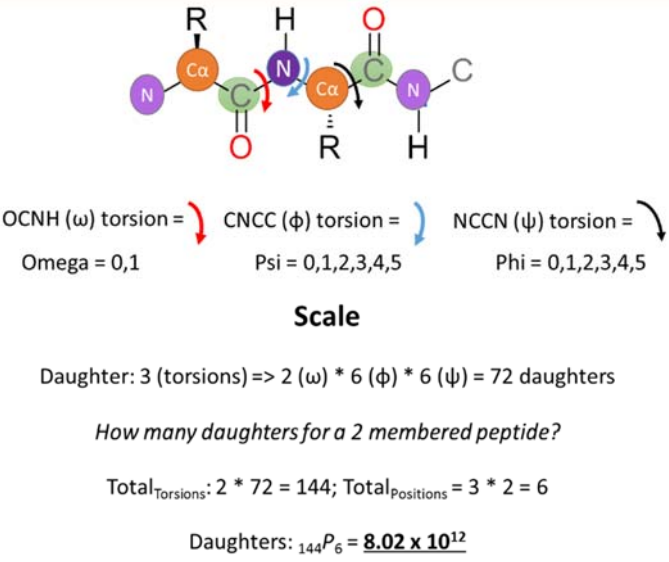


Figure 4.1: Torsions available to the peptide for conformational search. There are 2 torsions for omega; 6 torsions available for psi and phi dihedral torsions.

The number of daughter backbone conformers that graduate to the next level of elongation increases exponentially through each addition. In the case of a cyclizing peptide, the first amino acid has 72 daughter conformations whereas addition of the second gives rise to 8.02

x 10^{12} peptide conformers. The conformational number of the third amino acid reaches 373248, but this will be increased immensely by chemical space (where the amino acids are explored in addition to the conformations). Optimization of the final conformations may remove the number of conformations produced. In addition, if a 6 membered cyclic peptide is desired, for example, many of the conformations produced will not be appropriate for a final cyclic peptide, and there is risk in waste in computation before this limited set of conformers is generated, due to the exhaustiveness of the search.

4.3.2 Strategy Implementation

The functionality for the Prototype 1 strategy was implemented in python. The implementation had 5 Classes (*vector_Class, atom_Class, peptide_Class, labels,* and *recursivelabels*) that had functions that recursively build up the peptide while generating conformations. There were no special dependencies for this code other than the Python Standard Library which was used to construct all classes and methods employed by Prototype 1 (code for Prototype 1 is present in Supplementary Information).

4.3.2.1 vector_Class

Functions that perform vector calculations for geometric transformations were defined within a simple *vector_Class* class. The functions stored in *vector_Class* include: *setxyz* (sets the vector description for a *vector_Class* object, given 'xyz' co-ordinates defining a vector); *scale* (sets the bond distances to match bond type); *translatexyz* (changes the co-ordinates by a vector); *normalize* (sets the unit vector of the co-ordinates matrix to 1 by dividing by the square root of the magnitude of the vector); *dotproduct* (computation of the dot product between two sets of vectors); *crossproduct* (computation of the cross product of two vectors); *dihedral* (calculation of the dihedral bond between 4 atom pairs using their 'xyz' coordinates); *rotatez* (matrix computations to rotate *x, y* and *z* co-ordinates about the z axis through a

specific angle); *rotatev* (matrix computations to rotate *x, y* and *z* co-ordinates about a defined axis of rotation v through a specific angle).

4.3.2.2 atom_Class

Functions that act on parsed data from the PDB file referring to atom specific properties of an amino acid are stored in the *atom_Class* type objects. The functions available to the *atom_Class* include: *setname* (captures and stores the name of an atom)*; setelement* (captures the name of an element and stores it in a private variable)*; set* (combines setting both the atom name variable and also its 'xyz' co-ordinates)*; inbackbone* (marks the elements of the amino acid that would form the backbone of the completed peptide)*; printthis* (a function that reports atom specific properties such as the name and coordinates of the atom)*; detectNH* (captures the atoms of the amino acid that define the NH bond of the incoming amide group that will form the substitution point in the incoming amino acid)*; detectOH* (captures the atoms of the amino acid that define the OH bond of the terminal carboxyl group of the peptide. These will form the substitution point onto the backbone vector of the peptide for the incoming amino acid)*; printpdbline* (display the parameters of the atoms in PDB format)*; savepdbline* (write the parameters of the atoms in PDB format into the active file).

4.3.2.3 amino_class

Functions that perform amino acid manipulations completing the geometric orientation of whole amino acids are stored in the *atom_class* class. The functions stored in the *amino_class* include: *maindihedral* (specifies the atom positions of the OCCN dihedral of the amino and its angle)*; getCvector* (defines the hydroxyl CO bond and sets its bond distance to the NC bond distance of the peptide bond – see *peptide_Class addamino* function)*; getCOOvector* (defines the carbonyl CO bond and sets its bond distance to the CO carbonyl bond distance of the peptide)*; getNH1vector* (defines the amide NH bond and sets its bond distance)*; getNvector* (defines the NH bond of the peptide bond forming NH group that will be removed upon

formation of peptide bond – see *peptide_Class addamino* function)*; getNCAvector* (extracts the NCα bond information and defines the NCA vector from these two atoms)*; getCACvector* (extracts the CαC bond and defines a vector based on these two atoms)*; getCposition* (specify the position of the carboxyl group carbon of the amino acid)*; getNposition* (specify the position of the backbone Nitrogen atom of the amino acid)*; getCAposition* (specify the position of the Cα carbon of the amino acid)*; printthis* (displays the fields of the incoming amino acid)*; printpdbline* (display the parameters of the peptide atoms in PDB format)*; savepdbline* (write the parameters of the peptide atoms in PDB format into the active file)*; translate* (perform spatial translations on atom co-ordinates)*; rotate* (apply a rotation on the atoms about a specified angle)*; rotatev* (apply a rotation on all the atoms of the amino acid, about a specified angle along a vector)*; rotatev1* (apply a rotation on all the atoms of the amino acid except for NH_2, about a specified angle along a vector)*; rotatev2* (apply a rotation on only COOH atoms of the amino acid about a specified angle along a vector)*; setbackbone* (store the backbone atoms of the amino acid)*; setNH* (store the backbone NH atoms of the amino acid – preparation for substitution)*; setOH* (store the backbone OH atoms of the amino acid – preparation for substitution)*; openname* (read incoming file)*; addatom* (include atoms in atom list)*; readin* (extract atom parameters from incoming amino acid file)*; coords* (extract atom coordinates from incoming amino acid file).

4.3.2.4 peptide_Class

Functions that act on the growing peptide are stored in the *peptide_Class* class. The functions stored in the *peptide_Class* include: *distance_between_ends* (monitor the distance between the first N atom of peptide and the last C atom of the growing peptide)*; printalldihedrals* (measure the "main dihedral", the CCNC dihedral and the CNCC dihedral)*; getaminoacid* (return the *amino_class* object with specified number)*; addamino* (fix incoming amino acid at

terminal end of peptide by substituting the "*Nvector*" of the incoming aacid with the COH "*Cvector"* of the terminal amino acid of the growing peptide – see *amino_class getNvector* and *getCvector* functions)*; gettorsionvector* (define the bond along terminal aacid in order to introduce the specified dihedral angle)*; rotatepeptide* (apply the rotation along the bond of the terminal amino acid in order to introduce the specified dihedral angle)*; printpdbline* (display the parameters of the peptides in PDB format)*; savedpdbline* (write the parameters of the peptides in PDB format into the active file)*; printthis* (display the parameters of the peptides).

4.3.2.5 labels

Functions acting within an iterative loop to assign the torsions of the backbone of the peptide are stored in the *labels* class. The functions stored in the *labels* class include: *nameloop* (assign the torsion values of the CNCC dihedral (2), the NCACC dihedral (6) and the CACC dihedral (6) of the mutating terminal amino acid of the growing peptide)*; aminoloop* (apply mutations on the terminal amino acid according to the assigned torsion labels)*; printthis* (display the torsion labels for the terminal amino acid of peptide).

4.3.3 Program Description

The first step of the routine is to read the amino acid sequence, capture the PDB co-ordinate data for each amino acid type and to store these coordinates within appropriate *amino_Class* objects. Peptide labels are prepared for the peptide sequence using an iterator. Geometrical translation and rotation of the incoming amino acid atoms is performed in order to align this amino acid with the terminal C atom of the predecessor. This is performed before rotating about the 3 torsion vectors in order to adjust the conformation of the incoming amino acid. A peptide conformer is prepared for each label by reading through the list of peptide labels and applying the appropriate rotations to each part of the incoming amino acid. The distance between the terminal atoms are measured after each rotation. If the distance measured is

within a bond distance threshold of 1.5 Å, it is assumed that the conformer is a possible cyclic conformation. The process is repeated until all the torsions within the peptide are explored. The Pseudo-code for Prototype 1 is illustrated below (Fig. 4.2).

```
read aminoacid_sequence()
prepare_peptide_torsion_labels() #comprehensive list of torsions (e.g. tripepetide = 100_024_111)
foreach (peptide_label):
        foreach(amino_acid_in_sequence):
                translate_and_rotate(incoming_aminoacid)
generate_appropriate_conformation(incoming_aminoacid)
distance = distance_measure(terminal_atoms)
                if distance < 1.5 A:
                        save_the_peptide_conformation(name = peptide_label)
exit()
```

Figure 4.2: Pseudo-code for the python parser that implements the Prototype 1 simultaneous enumeration with conformation evolution during chain growth strategy.

To illustrate the *peptide_labels* conformer generation routine, Fig 4.3 outlines a 120° rotation about the *CACvector* torsion on the 3rd amino acid of a peptide made up of 5 glycine amino acids.

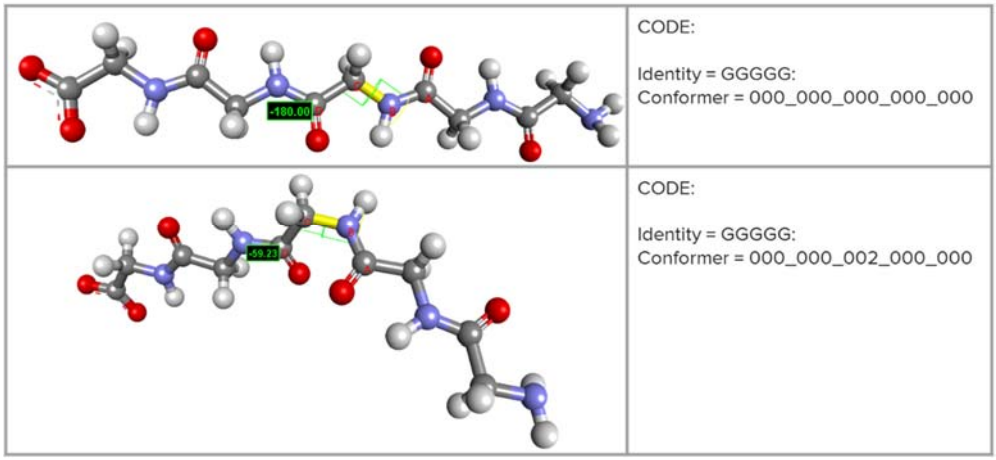


Figure 4.3: A conformer that is generated from a 120° rotation about the CACvector on the 3rd amino acid of a 5 membered glycine peptide.

4.3.4 Deployment of Prototype 1 on tripeptide virtual libraries

The Prototype 1 algorithm was deployed within a linux environment with amino acids obtained from a standard crystal structure database. The code was tested on the preparation of a tripeptide made solely of three Leucine amino acids.

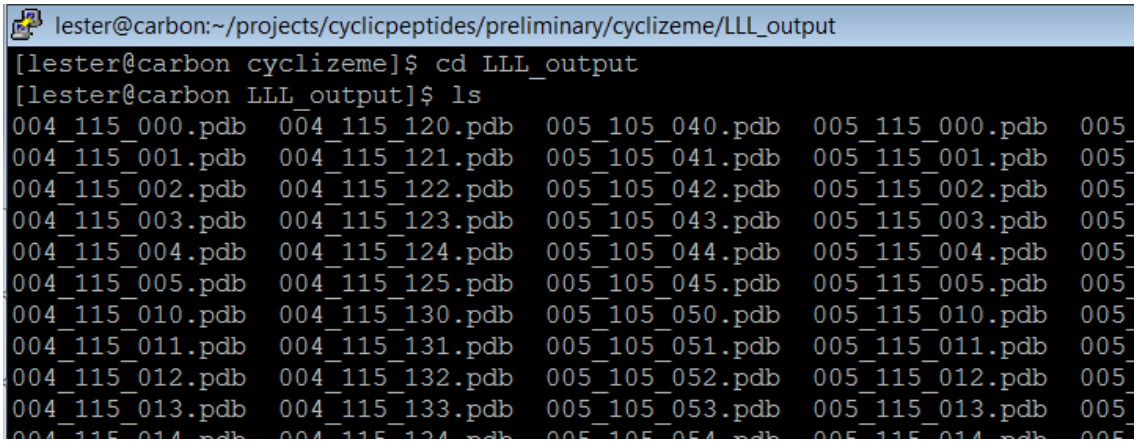


Figure 4.4: Stored conformers of the Leucine tripeptide that satisfied the cyclization metric.

There were 432 peptide conformers that achieved successful cyclization (stored in "~/cyclizeme/LLL_output", Fig. 4.4). These conformers were accessed after the preparation of 34,908,159 conformers (distance monitored in "~/cyclizeme/stats/peptide_distances.csv", Fig. 4.5) before termination was initiated after 5 days.

A

```
lester@carbon:~/projects/cyclicpeptides/preliminary/cyclizeme/stats
Filename, Distance
1.pdb, 10.781
2.pdb, 10.781
3.pdb, 10.781
4.pdb, 10.781
```

B

```
[lester@carbon stats]$ less peptide_distances.csv | grep -c "pdb"
34908159
[lester@carbon stats]$ less peptide_distances.csv | tail -20
34908140.pdb, 8.994
34908141.pdb, 11.976
34908142.pdb, 11.976
34908143.pdb, 11.976
34908144.pdb, 11.976
34908145.pdb, 10.781
34908146.pdb, 10.781
34908147.pdb, 10.781
34908148.pdb, 13.979
34908149.pdb, 14.618
34908150.pdb, 10.781
34908151.pdb, 10.781
34908152.pdb, 8.994
34908153.pdb, 11.976
34908154.pdb, 11.976
34908155.pdb, 11.976
34908156.pdb, 11.976
34908157.pdb, 10.781
34908158.pdb, 10.781
34908159.pdb[lester@carbon stats]$ less peptide_distances.csv | grep -c "pdb"
34908159
[lester@carbon stats]$
```

***Figure 4.5: Report of the conformers generated and measured**. A. Column labels for report. B. Evidence of conformations generated.*

Examples of some of the cyclized conformers are illustrated in Fig. 4.6. After 1 hour 30 minutes conformer “004_115_004.pdb” was recorded while conformer “005_115_143.pdb” was recorded after 1hr 45 minutes.

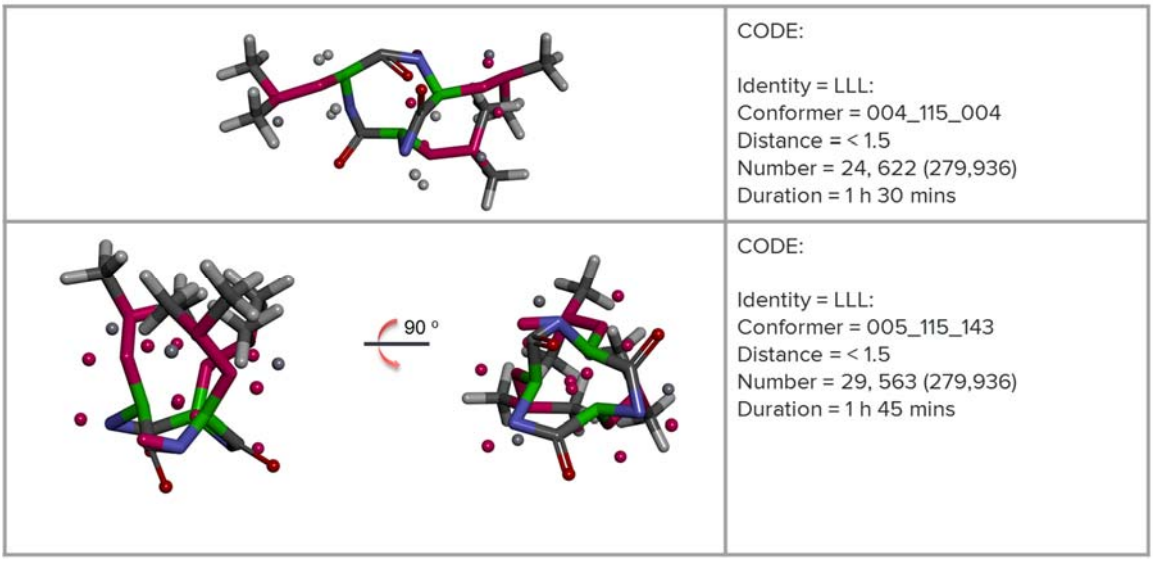


Figure 4.6: Examples of cyclized conformers prepared by Prototype 1 algorithm rendered using Discovery Studio Visualizer graphics.

Problems with side chain conformation were illustrated with the use of leucine for all residues in this tripeptide. In some cases the cyclized system was generated with acceptable side chain conformation; however this was not predictable being dependent on both the initial leucine conformation and the combination of 9 torsion angles in this case. Where the aim is to generate a diverse library that is conformationally laden (the focus being on ring conformation) computation directed at preventing steric clashes in the side chains at each step of this systematic approach will severely impact the efficiency with which cyclized conformations are identified.

Optimization of this code for speed through the use of parallelization could be explored as an option to speed-up the implementation of our algorithm. The limitation associated with the orientation of side chains would also need to be addressed. Peptides typically have alternating orientations of the side-chains in order to reduce steric clashes. Restricting our conformational search to a distance metric neglected the impact of steric clashes in restricting the likelihood of cyclization.

4.3.5 Conclusion

Although our flexible Prototype 1 algorithm could be used to populate a virtual library of tripeptides, the reliance on blind iterations to search the conformational landscape of conformers was a limitation in terms of computation and time. Speed-up of our implementation in order to reliably search the conformational and chemical space of peptides was not warranted given the successes of Prototype 2. Although this conformational search routine using distance metrics resulted in a large number of conformations, this number of conformations impacted their evaluation in terms of the search for bioactive conformations, which was judged to be too expensive computationally to follow. The strategy employed in the development of Prototype 1 was not adequate for the goal of populating conformation-laden virtual libraries of cyclic peptides within a defined sequence space, particularly in the context of Prototype 2.

4.4 Prototype 2:

In order to achieve the goals set out, an alternative strategy we pursued decoupled enumeration from conformational search. Thus Prototype 2, followed the strategy of identification of conformational diversity in a bare peptide backbone, and functionalization of this backbone ("decoration") rendered each conformation to be a polypeptide with the appropriate residues. Thus generation of libraries spanning both chemical and ring conformational space could be realized.

4.4.1 Logic

The limitations imposed by the simultaneous enumeration and conformational generation evident early in the development of Prototype 1, motivated the development of the hybrid Prototype 2 enumeration approach. Prototype 2 enumeration approach separated the two problems of searching the peptide space and that of searching the conformational space. This approach used a tool developed in the early stages of this project, which was capable of

modifying or adding functionality to 3d molecular scaffolds. Since the capability of the tool was modification of chemical functionality, using it to modify a conformational library of a poly-glycine backbone to generate many ring conformations of specific polypeptides was a feasible approach. Although the tool was used for preparation of polypeptides, its original purpose was lead modification, and it had been given the name DerivatizeME. Its full functionality will be discussed, and then specifically its use within the scheme of Prototype 2 for developing conformationally laden libraries of cyclic peptides. One strength of DerivatizeME is its ability to modify compounds in existing virtual libraries (such as SANCDB, the South African Natural Compound Database) that are not promising in terms of physico-chemical properties and generate a derivative library that contains many promising leads. As such the discussion about DerivatizeME will digress to explore this, before coming back to address its incorporation into Prototype 2 and how it may be used to generate conformationally laden cyclic peptide libraries.

4.4.2 DerivatizeME, a tool for directed functionalization of chemical scaffolds

To illustrate the fuller functionality of DerivatizeMe, Fig 4.7 shows the "mutation" or "decoration" of a bicyclic scaffold with 4 substituents. In this mode DerivatizeME recursively functionalizes a scaffold in order to generate all the possible derivatives. To do this a function, called *"fix"* is called recursively in order to prompt "mutations" of the scaffold for future levels, until all possibilities are generated.

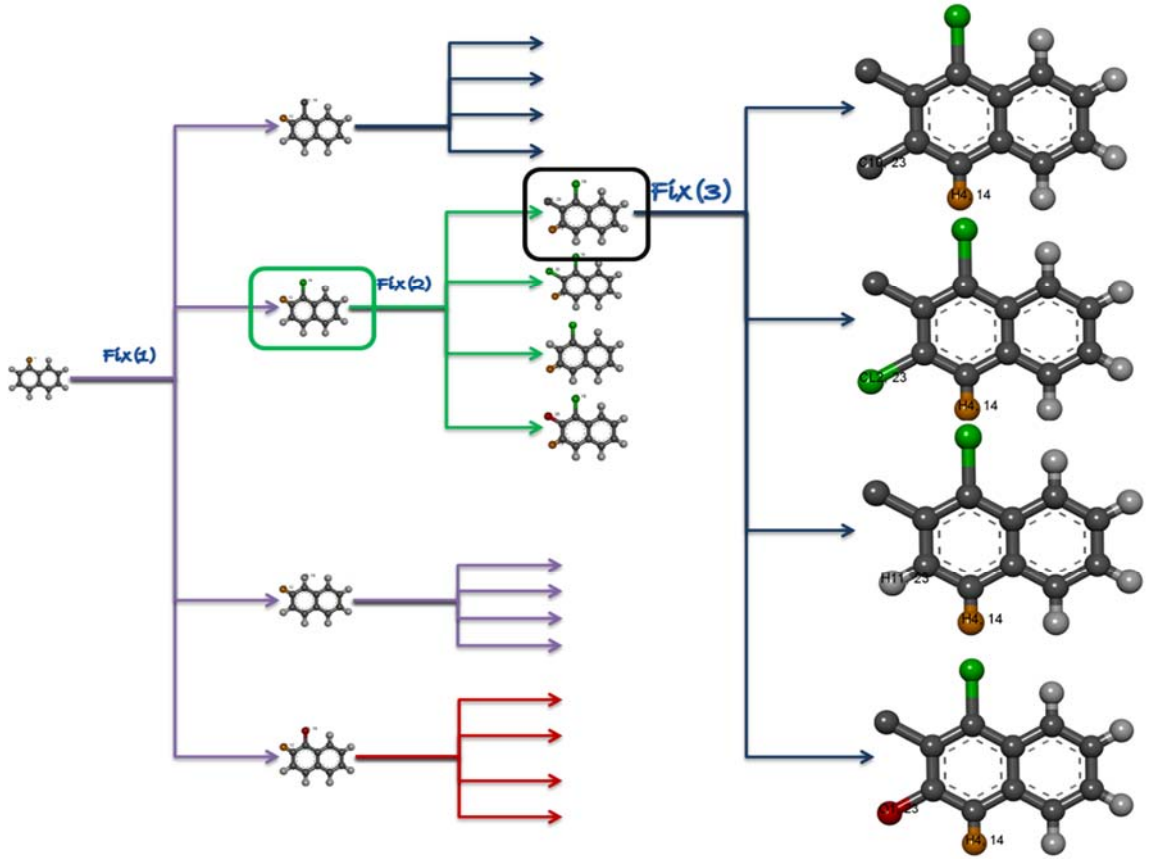


Figure 4.7: The decoration of a predefined scaffold using the Prototype 2 enumeration strategy. *Light grey atoms are substitutable hydrogens on the scaffold. The brown labelled atoms are the active substitution points. The green, dark grey, light grey and red labelled atoms are substituents. The arrow colours match the path followed by the derivative to the next stage of substitution.*

4.4.3 Strategy Implementation

The DerivatizeME was written in C++ and compiled in a linux environment using the gnu C++ compiler. A compromise in flexibility was made by making use of the OpenBabel API in order to access molecular manipulation methods that allowed us to implement our strategy. In order to successfully implement DerivatizeME, two header files were prepared to store functional routines (*"result.h"*) and geometric routines ("*multiple.h"*).

4.4.3.1 Functional routines

The *"result.h"* header file was used to store functional routines arranged in the *counter, errorfilemaker, logfilemaker* and *result* class objects. The *counter* object was used to communicate with all other objects, incrementing a counter every time it was accessed, such that other objects would be assured of writing structures to unique files. The *errorfilemaker* and *logfilemaker* classes stored functions that were used to open, close and write to the error and log files respectively. The core functional methods for DerivatizeME were stored in the *result* class. These methods included methods for **object communication and initialization**, (connecttocounter, connecttoerrorfile, connecttologfile, setacc, getacc, printacc, resetall, count), **structure manipulation**, (deleteallbondsinmolecule, deletedeletablehydrogens); **substitution routines** (makeallhydrogenssubstitutable, intelligentlymakehydrogensubstitutable, ignorehydrogen, makehydrogenssystematic, makerandomhydrogenssubstitutable, set_substupto_and_maxh, substitutablehydrogen, add), **io routines** (save, print), and **molecular descriptor routines** (setHBD, setHBA1, setHBA2, addHBD, addHBA1, addHBA2, addapproxmass, setpredlogp).

There were four substitution routines explored within the DerivatizeME development. The *makerandomhydrogenssubstitutable* routine made use of a parameter that could be tuned to specify what proportion of the available hydrogens could randomly be made substitutable. The *makehydrogenssystematic* routine prepared derivatives that had only been substituted

at one point i.e. single substitution derivatives. The *intelligentlymakehydrogensubstitutable* routine read a user-supplied list of integers that matched the hydrogen positions where substitution could be permitted. The *makeallhydrogenssubstitutable* routine made all hydrogen atoms of the scaffold substitution points (exhaustive substitution).

4.4.3.2 Geometric routines

The *"multiple.h"* header file was used to store functions and variables that enable critical geometric routines facilitating molecule transformation. The functions were arranged in the *Translate, Rotate, BondRotate, AtomDistance, badcontacts* and *split* routines made available by the research group. In order to accomplish the geometric operations described in *"multiple.h"* access to the OpenBabel API was through inclusion of appropriate headers such as *"obconversion.h"* and *"mol.h"*, and standard C++ headers such as *"mathlib.h"* and *"iostream.h"* were also referenced.

4.4.4 Program Description

The *functional* and *geometric* routines were coupled into the DerivatizeME enumeration engine. The four substitution routines were developed in order to bias substitution towards either exhaustive substitution or focussed substitution. Fig. 4.8 outlines the strategy for exhaustive (termed here as multifarious) substitution in order to demonstrate the enumeration routine and its impact on the derivatives generated.

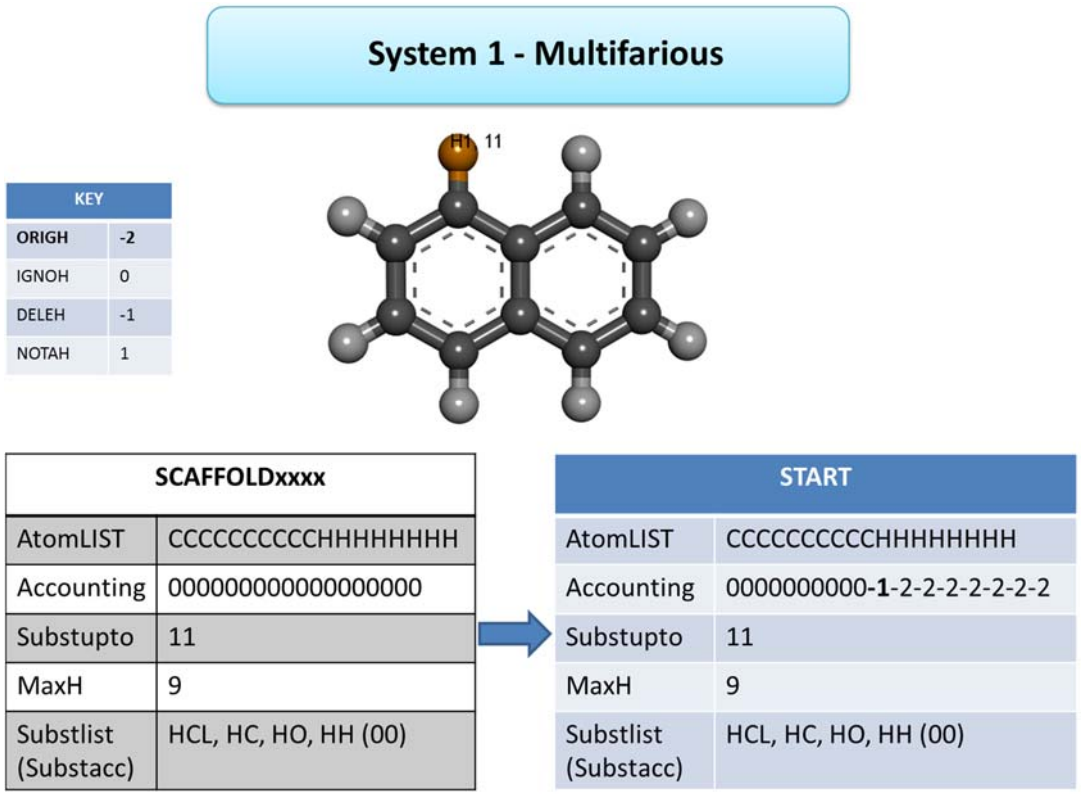


Figure 4.8: The description box of a scaffold with 8 hydrogens that can be substituted using the multifarious "makeallhydrogenssubstitutable" strategy*. There were four substituents used to construct derivatives of the scaffold.*

In this approach, each derivative is assigned parameters that keep track of the **AtomLIST, Accounting, Substupto, MaxH** and **Substlist** descriptions of the molecule**.** The **AtomLIST** description stores the atoms present in the molecule in order. The **Accounting** description records the current state of an atom within the molecule. In the event that an atom is available for substitution its value is set to be non-zero. If that atom is to be substituted in the current "fix" cycle its value becomes -1, but -2 if it is to be substituted at a later stage (arbitrarily chosen values). In the case of the exhaustive "multifarious" substitution strategy accessed by the *makeallhydrogenssubstitutable* setting, all hydrogen atoms in the scaffold are set within **Accounting** to be non-zero (Fig. 4.8). The **Substupto** description identifies the number of the atom that is to be checked for substitution (its value is incremented only after substitution has occurred to start the next *'fix'* cycle). The **MaxH** description allows us to track

how many hydrogens can be substituted at most within the scaffold. When the number of substitutions that have occurred is equal to **MaxH** – 1, the recursive *"fix"* cycle is terminated. At this stage no further derivatives are possible, no substituent atoms can be added to the **AtomLIST** and no hydrogens can be marked for deletion. The **Substlist** description stores the different substituents that will be used for substitution. In the given example in Figure 4.8 there are four substituents, HCl (for chloro substitution, the H provides the vector for substituent addition), HC (for methyl group substitution, final structures are not saved with hydrogen atoms to improve storage, CH_4 also works), HO (for hydroxy addition) and HH (to maintain all possible unsubstituted sites in addition to all possible substituted). In this exhastive approach, since 4 substituents are defined the number of generated compounds will increase in the sequence 1, 4, 16, 64 … 4^N, where N is the number of hydrogens substituted. The hydrogen atoms of the substituents are not included in **Accounting** to avoid confusion with the *"fix"* routines that act on native hydrogens present in the original molecule.

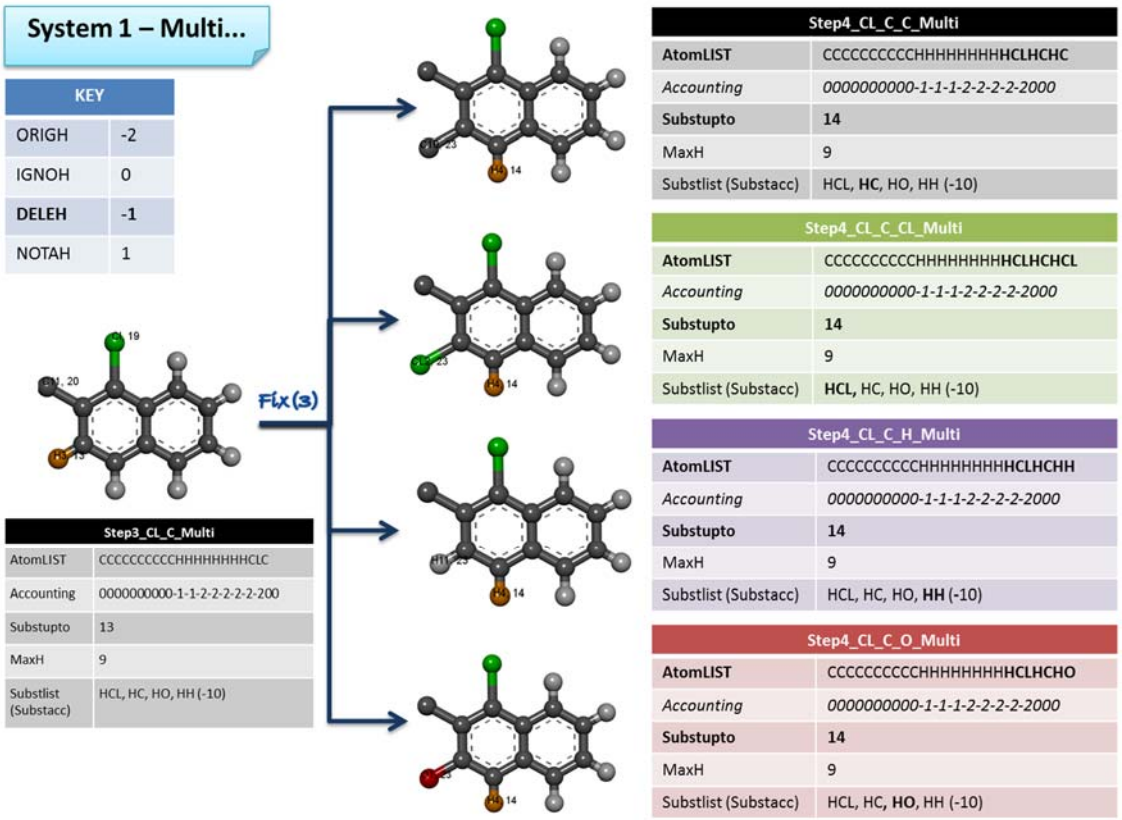


Figure 4.9: The progression of the adjustments that are made to the description box of the scaffold along its enumeration path. *The changes made during the third "fix" cycle of our scaffold are outlined with the description box for the 4th substitution shown.*

If the first two substitutions of the example in Figure 4.8 lead to Cl and C (methyl) substituents, the resultant disubstitution will then undergo a 3rd fix cycle to produce 4 derivatives by substitution at position 4 (Figure 4.9) of the 64 total compounds at this level. For the purpose of emphasis the **AtomLIST** substituent atoms are in bold and the hydrogen atoms are included. At every *"fix"* stage the hydrogen atoms present in the derivative are deleted, and the particular derivative is saved in *.xyz* format. In terms of *result* object creation, at each level of the "*fix*" cycle the *result* object is copied several times, the derivative changes made, the derivative saves a unique xyz file, and these objects recursively enter the next cycle of *"fix"*. Termination of the recursion is clear in terms of **Substupto** and **MaxH**. The different DerivatizeME enumeration strategies differ from this described approach in the manner in

which they define the hydrogens that are substitutable. The hydrogen description consistent with the enumeration strategy invoked is introduced onto the system as the first *result* object is spawned from the scaffold input.

4.4.5 Digression – The evaluation of DerivatizeME in the context of existing virtual libraries.

As already described, the DerivatizeME system was compiled in a standard gnu environment with OpenBabel dependencies. The input scaffolds and substituents were read in with the *'.xyz'* file format. Substituents were described (such as chloro, hydroxy etc.) together with a hydrogen atom; the hydrogen atom, although not included in the substitution, identified the orientation of the substituent and enabled correct 3D orientation of the substituent in creating derivatives. In order to check the capabilities of this enumeration approach three substituents were investigated, in the context of scaffolds obtained from the South African Natural Compound Database (SANCDB). The substituents chosen included the mono-atomic chloro substituent and the multi-atomic methyl and carboxylic acid substituents, which were used to decorate the SANCDB scaffolds. The collection of compounds in SANCDB were prepared by bioinformaticians that used a combination of automation and manual curation methods for the collation of a wide range of compounds derived from South African organisms in a single database. The objective of this derivatization of SANCDB scaffolds was two-fold: To demonstrate both the applicability of the different enumeration DerivatizeME strategies, and also to identify limitations in these decoration strategies with regards to structural restrictions of scaffolds and substituents.

4.4.5.1 Deployment of the DerivatizeME algorithm on SANCDB Scaffolds

The four DerivatizeME substitution routines we developed were tested on two SANCDB natural products. One natural product that had 5 substitutable hydrogens was used to test the exhaustive routine while a second natural product with 21 hydrogens was used to demonstrate the Random, List and Systematic routines (Figure 4.10). The performance of the

different routines were tested for their ability to run to completion, terminate independently and match the predicted number of derivatives.

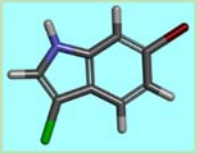
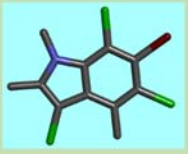
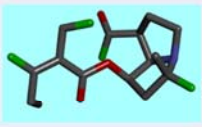
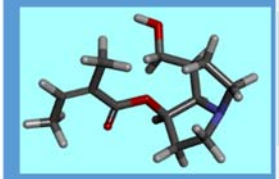
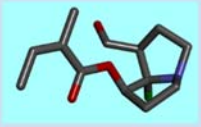
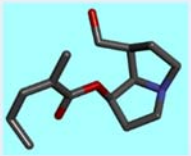
	Method	Predicted	# of Derivatives	Example
	All (5 H's)	$3^5 = 243$	243	
	Random	$3^8 = 6{,}561$ *See Fig 4.11	6,561	
	List (single position)	$3^1 = 3$	3	
21 Hydrogens	Systematic	3 * 21 = 63	63	

Figure 4.10: The four DerivatizeME routines. *All (makeallhydrogenssubstitutable), Random (makerandomhydrogenssubstitutable), List (intelligentlymakehydrogensubstitutable) and Systematic (makehydrogenssystematic) routines tested on SANC 488 that had 72 substitutable hydrogens. Examples of derivatives obtained from the Cl (green) substituent are shown as examples.*

The methods tested ran to completion uninterrupted confirming the robustness of the DerivatizeME program. The 'All' method was tested on a scaffold with only 5 hydrogens in order to reduce the size of the computation. The exhaustive 'All' method produced the expected 243 derivatives completing its process within a reasonable time. The 'Random' method produced 6,561 derivatives. These are achieved if the 'Random' sampling made 8 of the 21 hydrogens substitutable. The ability to make an intelligent substitution was tested by the 'List' method. This method generated derivatives at a single specified position and gave rise to the 3 expected derivatives. The 'Systematic' method produced 63 derivatives that had each been substituted at a single point consistent with its conservative approach to substitution.

The methodology followed in the 'Random' substitution strategy is outlined (Fig. 4.11). An iterator loops through all the H atoms of the query and uses a proportion value to determine the extent of substitutable hydrogens that are selected at random. The proportion value can be adjusted to adjust the extent of substitution. So, for example, if a maximum of 15 hydrogen atoms in the scaffold should undergo substitution, the code in Figure 4.11 will effect this random substitution as Rand(15).

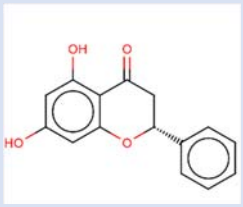
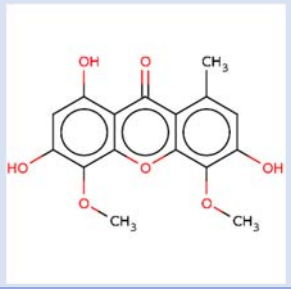
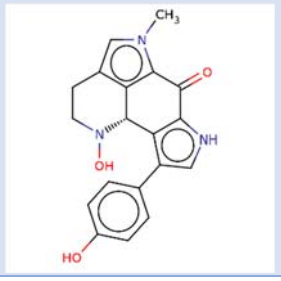
```
Loop through Hydrogens in atom to get H_total
Set proportion value to 15000/H_total
Loop through atoms in molecule:
        extract random no. such that it satisfies rand () % 1000
        if atom is Hyrdogen and random < proportion:
                set hydrogen as substitutable
Apply fix() on all substitutable Hydrogens
```

Figure 4.11: Pseudo-code for the DerivatizeME Random routine.

The Random (multifarious) and Systematic decoration schemes were tested on three representative compounds from SANCDB and assessed by *in silico* ADME-Tox screening using FAF-Drugs (Lagorce et al. 2015). The chosen compound scaffolds were scored as being in the three categories: Accepted (no structural alerts and ideal physico-chemical properties), Intermediate (low-risk structural alerts below a threshold) and Rejected (high-risk structural alerts and low-risk motifs that exceed a threshold) according to strict lead-like criteria. The Rejected candidate was SANC 105 identified as Pinocembrin isolated from the *Combretum apiculatum* (Aderogba et al. 2012). The Intermediate candidate was SANC 474 known as Drimiopsin D isolated from the *Drimiopsis maculata* (Mulholland et al. 2004). Finally, SANC 139, Tsitsikammamine B N-oxime isolated from the *Tsitsikamma favus* was chosen as the

Accepted candidate (Antunes et al. 2005). Table 4.1 lists these compounds and their calculated properties.

Table 4.1: Representative compounds from SANCDB that reflect Rejected, Intermediate and Accepted compound statuses.

Lead-like soft property	Rejected - SANC 105	Intermediate - SANC 474	Accepted - SANC 139
			
MW (150 - 400 Da)	256.25	318.28	335.36
logP (-3 to 4)	2.80	2.39	2.54
HBA (=< 7)	4	7	6
HBD (=< 4)	2	3	3
tPSA (=< 160)	66.76	109.36	84.65
Rotatable Bonds (=<9)	1	2	1
Rigid Bonds (=< 30)	18	17	25
Rings (=< 4)	2	1	2
Max Size System Ring (=< 18)	10	14	15
Carbons (3 - 35)	15	16	19
Hetero Atoms (1 - 15)	4	7	6
H/C Ratio (0.1 - 1.1)	0.27	0.44	0.32
Charges (=< 3)	0	0	0
Total Charge (-2 to 2)	0	0	0
Stereo centres (=< 2)	1	0	1
Hydrogens	12	14	16

Three Random strategies the Rand (15), Rand (8), Rand (3) strategies (substituting a maximum of 15, 8, and 3 positions on the natural product skeleton) that adjusted the degree of substitution were tested along with the Systematic approach. To illustrate the impact of DerivatizeME enumeration by each scheme the pharmaceutical profiles of the derivatives were calculated when the chloro, methyl and carboxylic acid substituent groups were used as

substituents. The number of compounds produced by each method within each of these categories is shown in Table 4.2.

Table 4.2: Derivatives generated by the Random and Systematic DerivatizeME methodologies.

	Rejected - SANC 105 (12 H's)	Intermediate - SANC 474 (14 H's)	Accepted - SANC 139 (16 H's)
Rand (15)	81	2187	19684
Rand (8)	27	729	6561
Rand (3)	3	3	27
Systematic	36 (12)	42 (14)	48 (16)

There is a difference in the number of substituents produced by each of the Random strategies with Rand(15) producing the most derivatives, the Rand(8) method producing the second highest and the Rand(3) producing the least. Rand(15) will render more hydrogens substitutable than Rand(8) or Rand(3) at each level, and this is evident in the big difference in numbers of derivatives formed. A difference of 12 hydrogens between SANC 105 and SANC 139 translated into 250 times the number of derivatives in the former over the later (for Rand(15)).

A survey of the physico-chemical properties of derivatives obtained from these three compounds (SANC 105 Accepted, SANC 474 Intermediate and SANC 139 Rejected) illustrates the impact of DerivatizeME enumeration strategies on these three different scaffolds.

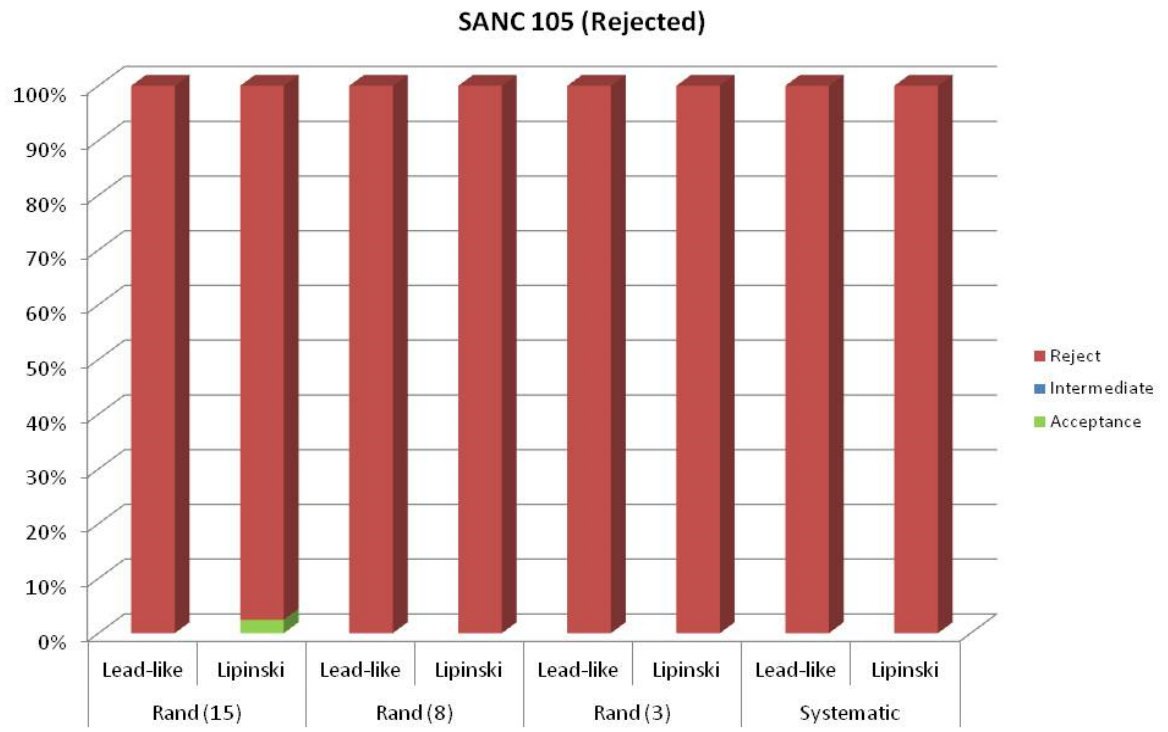


Figure 4.12: The diversity of derivatives of the Rejected SANC 105 scaffold generated by the different DerivatizeME methodologies

A small proportion of the derivatives of SANC 105 graduated to Accepted status based Lipinski criteria for the Rand(15) DerivatizeME methodology (Fig. 4.12). The application of Rand(8), Rand (3) and Systematic DerivatizeME approaches on this ADME-Tox Rejected compound did not appear to produce derivatives within either Intermediate or Accepted status, either in terms of Lipinski and Lead-likeness. This can be explained from the understanding that compounds are flagged as being Rejected if their physico-chemical properties exceed the filter threshold specified. Rejected status according to the FAF-Drugs criteria is either due to the detection of high-risk functional groups or due to compounds exceeding a threshold limit for low-risk motifs. DerivatizeME is not designed to remove problematic functional groups and is only designed to modify the physico-chemical properties of a query compound in a cumulative manner. Therefore, it was expected that few (if any) SANC 105 derivatives would

graduate from its Rejected status when the strict Lead-like parameters were implemented. The multifarious Rand (15) strategy was however able to modify the physico-chemical properties of 2.47 % of derivatives such that they were classed as Accepted when the more generous Lipinski RO5 filters were considered.

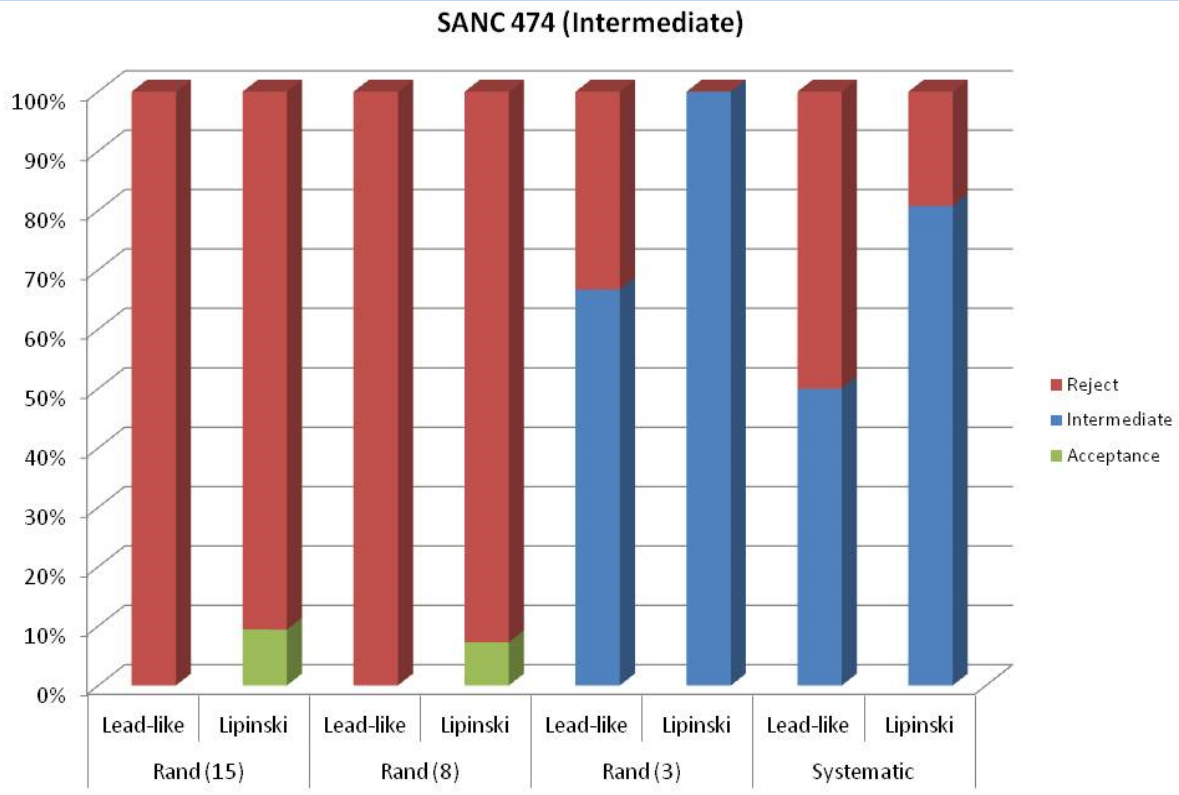


Figure 4.13: The diversity of Intermediate SANC 474 derivatives generated by the different DerivatizeME methodologies.

When the Intermediate compound SANC 474 was derivatized in the same way using DerivatizeME, positive results were observed with the different DerivatizeME strategies. The proportion of SANC 474 derivatives that graduated to Accepted status based on Lead-likeness or Lipinski criteria for the different DerivatizeME methodologies are shown (Fig. 4.13). 66% of derivatives generated using Rand(3) classified as Intermediate while the Systematic approach only saw 50% of its derivatives at Intermediate status (when the strict Lead-like filters were

considered). These observations are in contrast to the lack of generation of compounds classified as Lead-like when the Rand (15) and Rand (8) strategies were implemented using the specified substituents. The Rand (15) and Rand (8) strategies introduced more substituents into the scaffold increasing the likelihood of derivatives overcoming threshold limitations. Even when the more generous Lipinski factors filter was used, the Rand (15) and Rand (8) strategies produced few compounds with Accepted status. Surprisingly, neither the Rand (3) nor the Systematic DerivatizeME strategies produced compounds identifying as Accepted when the Lead-like or Lipinski filters were used. Despite the appearance of more derivatives that had Intermediate status according to the Lead-like parameters, there were derivatives of SANC 474 that had physico-chemical features that regressed them to the Rejected status. Despite having less substitution points, the Systematic strategy (50 %) introduced a larger proportion of derivatives that had unfavourable features when compared to the Rand (3) strategy (17 %).

The proportion of derivatives of an Accepted compound (SANC 139) classified by Lead-likeness or Lipinski criteria from the different methodologies are shown (Fig. 4.14). In terms of the Lead-like filtering parameters, the Systematic strategy (63.89 %) had a superior performance over the Rand (3) strategy (17.39 %) in producing compounds of Accepted status. In this instance, minor changes in the query compound were less likely to destabilise the positive traits already present in SANC 139, while a large number of substitutions contributed to the destabilisation of the compound when the Lead-like features were used for filtering. No Intermediate or Accepted derivatives in terms of Lead-like parameters were observed when the Rand (15) or Rand (8) strategies were employed for substitution. Instead, none of these derivatives maintained the positive traits of the query. In the case where the Lipinski parameters were used for filtering the Rand (3) strategy (100 %) had a remarkable

turnover rate for producing derivatives falling into the Accepted category. The Systematic strategy (71.30 %) was unable to match Rand (3) (100 %) in preventing attrition from derivatives when the Lipinski features were used for filtering.

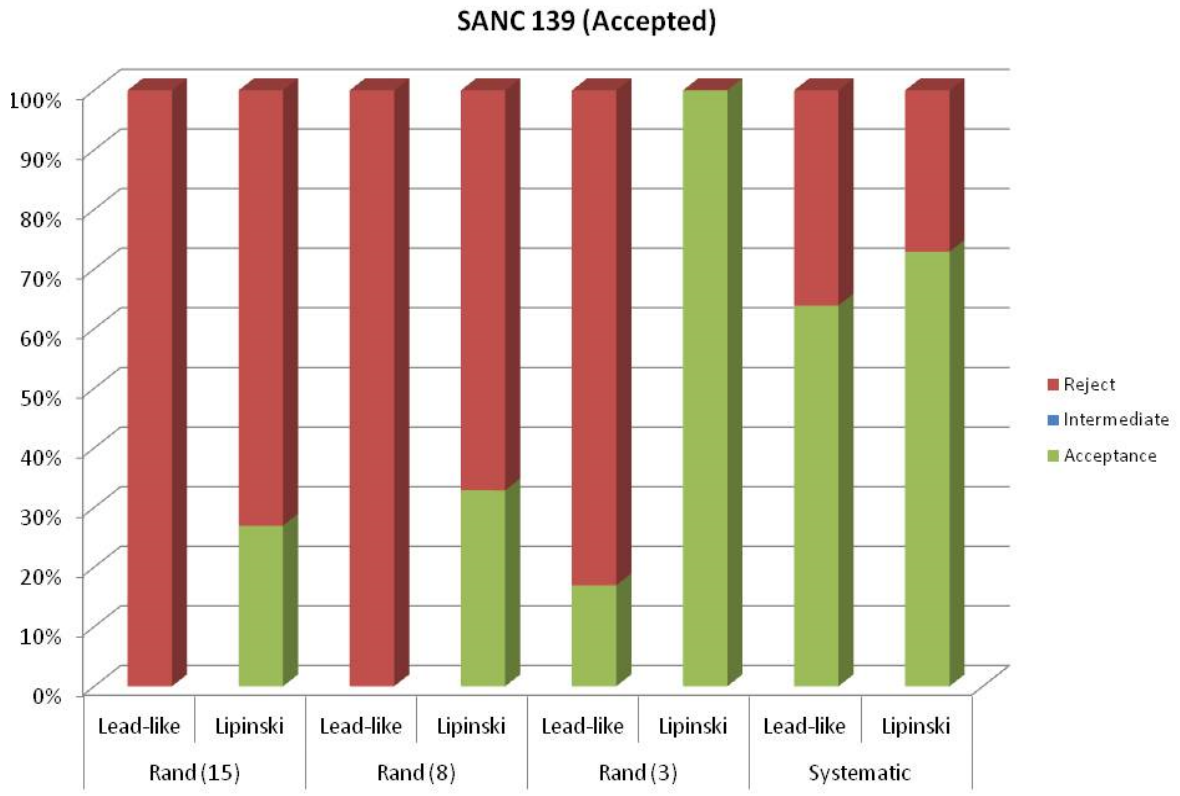


Figure 4.14: The diversity of Accepted SANC 139 derivatives generated by the different DerivatizeME methodologies.

In conclusion, the DerivatizeME methods are able to explore the chemical space of a compound making derivatives that have improved lead-likeness properties with different binding affinities available. The physico-chemical survey of derivatives shows that compounds that have scaffolds that possess problematic functional groups experience little to no improvement in their status (compared to the parent compound) with either exhaustive strategies or conservative decoration approaches. In order to "discharge" derivatives from their parent's unfavourable classification manipulation that unbalances the cumulative properties may be required as a future extension of the current DerivatizeME

implementation. The data from this study shows the reliability of the substitution protocols employed in DerivatizeME. From a survey of physico-chemical properties we were able to demonstrate that DerivatizeME enumeration is able to alter the chemical balance of a query compound through the introduction of different functional groups. The nature of the functional group, the number of substituents and the status of the query are all factors that affect the classification of derivatives produced by these enumeration experiments.

4.4.6 Deployment of DerivatizeME algorithm on peptide scaffolds (Prototype 2)

Consistent with the aims of developing a tool that will probe the chemical space available to cyclic peptide systems the "*makeallhydrogenssubstitutable"* routine was applied to peptide systems. The other DerivatizeME routines are available for further development although their use in the population of peptide virtual libraries was less facile. The main rationale for limiting our investment in these approaches was that we could not see an application where a peptide library populated by a multifarious (random) approach or systematic enumeration was preferred over a virtual library where an exhaustive routine was used. The list approach could have been used, but this required identification of exact atom numbers in the parent, and was not as convenient as the final approach followed. This section explored both single conformer peptide systems of six membered linear peptides and single conformer ten membered cyclic peptide systems. These experiments extracted results related to the timing of enumeration protocols, the appearance and geometry of the derivatives. The limitations of DerivatizeME on cyclic peptides associated with timing, virtuosity and exhaustiveness informed the use of DerivatizeME in generating cyclic peptide virtual libraries that possess conformational content (in Chapter 5).

4.4.6.1 Linear peptides

For the construction of virtual libraries for linear peptides, a six membered peptide of glycine residues was used to construct a backbone scaffold. The CycloPS enumerator was used to

generate the linear conformer before Discovery Studio visualizer was used to prepare the scaffold for decoration using DerivatizeME. In order to prepare for decoration all H atoms were deleted and single H's were added to the Cα backbone atoms (Fig. 4.15). The H atoms attached to the Cα were the sites of substitution for decoration. Appropriate chirality was maintained in the orientation of these H atoms. These backbones containing a single Cα-hydrogen per residue were termed "nude" scaffolds, for want of a better term.

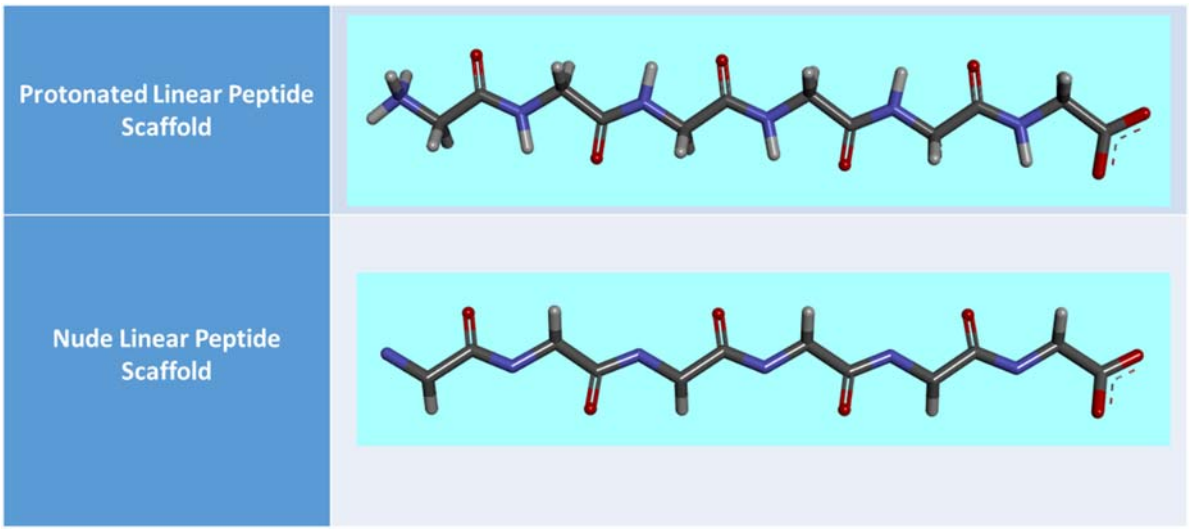


Figure 4.15: Six membered linear peptide of glycine residues. *The "nude" scaffolds were stripped of hydrogen atoms and hydrogen atoms remaining marked the substitution points.*

In the case of peptide decoration, the substituents were those that generated the naturally available L-amino acids. The substituents required to form the amino acids were prepared by extracting the side-chain, stripping all H atoms from this side chain, and replacing Cα with a single H atom. This, again, within the functioning of DerivatizeME ensured that substituents were added with the correct 3D orientation. These side chains, after preparation in this manner were stored in the *".xyz"* format (Figure 4.16).

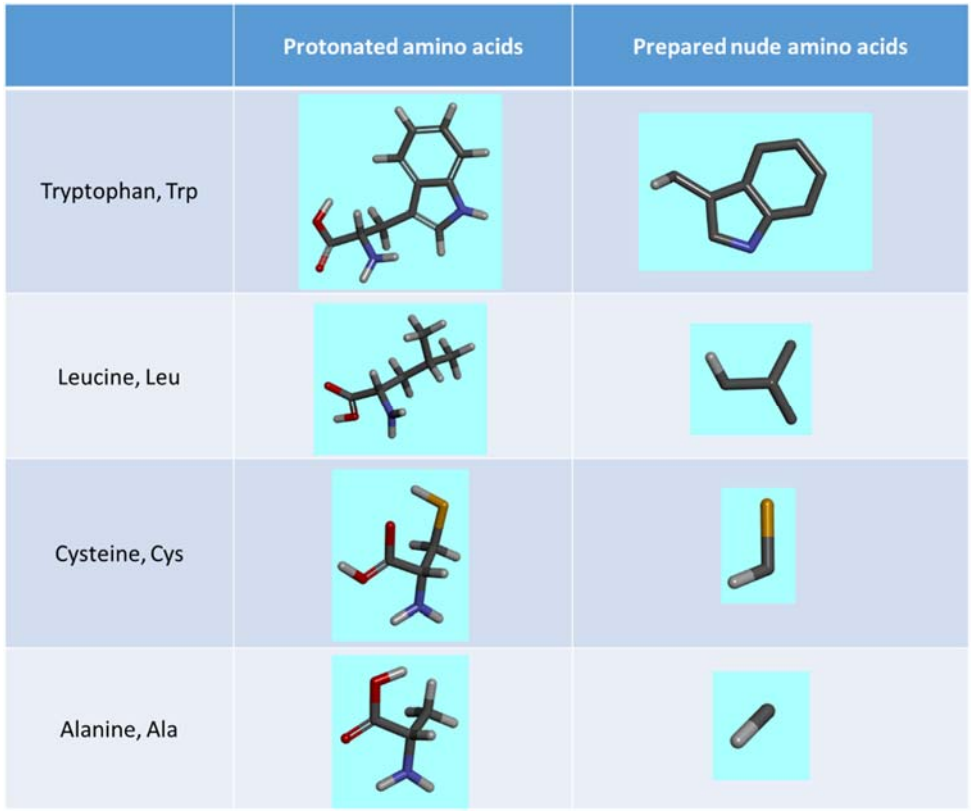


Figure 4.16: Examples of amino acids used for substitution. *The side chains of the amino acids were stripped of hydrogen atoms and a hydrogen atom remaining marked the substitution point.*

The approach we used to enable exhaustive decoration and efficient collation of the derivatives relied on the preparation of a subset of 6 amino acid substituents chosen from the 19 available amino acids used to construct the virtual library. An automated sequence from within a Python script allowed the exhaustive DerivatizeME enumeration protocol to be called, using these 6 substituents for derivatization. All the derivatives obtained from a single exhaustive run of DerivatizeME were stored in a single *".sdf"* file labelled by a description of the 6 membered substituent collection used. For the exhaustive decoration of a 6-membered linear peptide with 6 possible substitutions, a virtual library of 1,265,870,592 (1.2 Billion) individual peptides was populated (Fig 4.17). This number is consistent with what we expected statistically as there are 27,132 ($_{19}C_6$) different ways of choosing 6 substituents from

a collection of 19 amino acids while there are 46,656 (6^6) different ways in which derivatives can be made from 6 substituents.

```
glu_phe_val_cys_asn_ser.sdf  trp_arg_gly_asn_met_ser.sdf  val_tyr_his_asn_leu_ser.sdf
glu_phe_val_cys_his_asn.sdf  trp_arg_gly_cys_ala_asn.sdf  val_tyr_his_asn_met_ser.sdf
glu_phe_val_cys_his_ile.sdf  trp_arg_gly_cys_ala_his.sdf  val_tyr_his_ile_asn_leu.sdf
glu_phe_val_cys_his_leu.sdf  trp_arg_gly_cys_ala_ile.sdf  val_tyr_his_ile_asn_met.sdf
glu_phe_val_cys_his_met.sdf  trp_arg_gly_cys_ala_leu.sdf  val_tyr_his_ile_asn_ser.sdf
glu_phe_val_cys_his_ser.sdf  trp_arg_gly_cys_ala_met.sdf  val_tyr_his_ile_leu_met.sdf
glu_phe_val_cys_ile_asn.sdf  trp_arg_gly_cys_ala_ser.sdf  val_tyr_his_ile_leu_ser.sdf
glu_phe_val_cys_ile_leu.sdf  trp_arg_gly_cys_asn_leu.sdf  val_tyr_his_ile_met_ser.sdf
glu_phe_val_cys_ile_met.sdf  trp_arg_gly_cys_asn_met.sdf  val_tyr_his_leu_met_ser.sdf
glu_phe_val_cys_ile_ser.sdf  trp_arg_gly_cys_asn_ser.sdf  val_tyr_ile_asn_leu_met.sdf
glu_phe_val_cys_leu_met.sdf  trp_arg_gly_cys_his_asn.sdf  val_tyr_ile_asn_leu_ser.sdf
glu_phe_val_cys_leu_ser.sdf  trp_arg_gly_cys_his_ile.sdf  val_tyr_ile_asn_met_ser.sdf
glu_phe_val_cys_met_ser.sdf  trp_arg_gly_cys_his_leu.sdf  val_tyr_ile_leu_met_ser.sdf
[lester@g07s4027-3 outputdir]$ ls | grep -c ".sdf"
27132
[lester@g07s4027-3 outputdir]$ less trp_arg_gly_cys_his_leu.sdf
[lester@g07s4027-3 outputdir]$ less trp_arg_gly_cys_his_leu.sdf | grep -c "0000"
46656
[lester@g07s4027-3 outputdir]$
```

Figure 4.17: Screenshot of the linear peptide virtual library.

Remarkably, when this protocol was deployed within a server environment all derivatives were prepared within an acceptable time period (21 hours 19 minutes) (Fig. 4.18). The reader is referred to Chapter 6, “Towards the Population of a conformation-laden peptide library using DerivatizeME” for a discussion on the application of High-Performance Computing that facilitated the acceleration of our enumeration protocol.

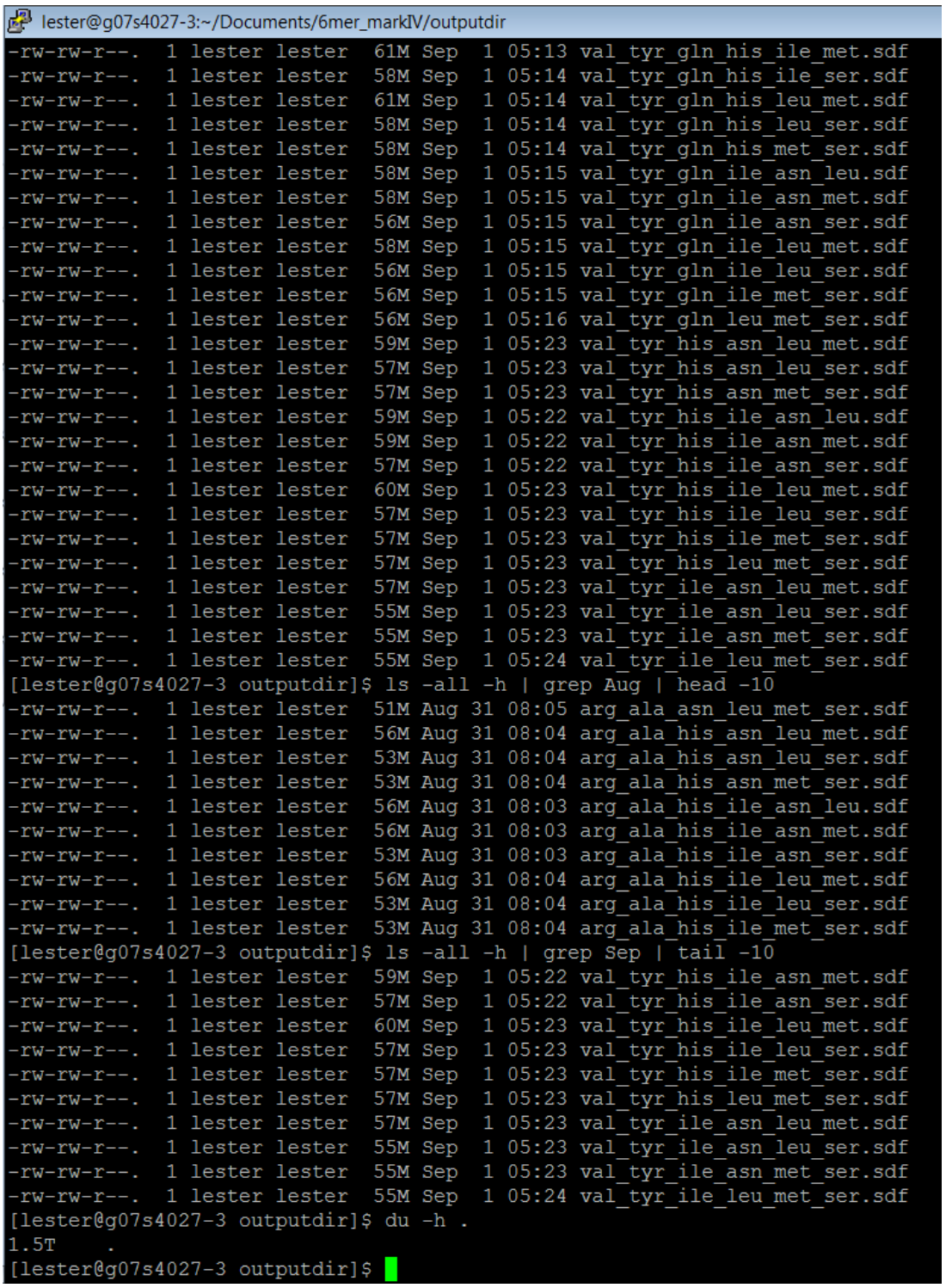


Figure 4.18: Screenshot of the time stamps and storage memory allocated to the linear peptide virtual library.

Owing to the difficulty to graphically illustrate all of the members of the linear peptide virtual library, an individual “nude” or bare peptide is illustrated with its protonated counterpart (Fig. 4.19).

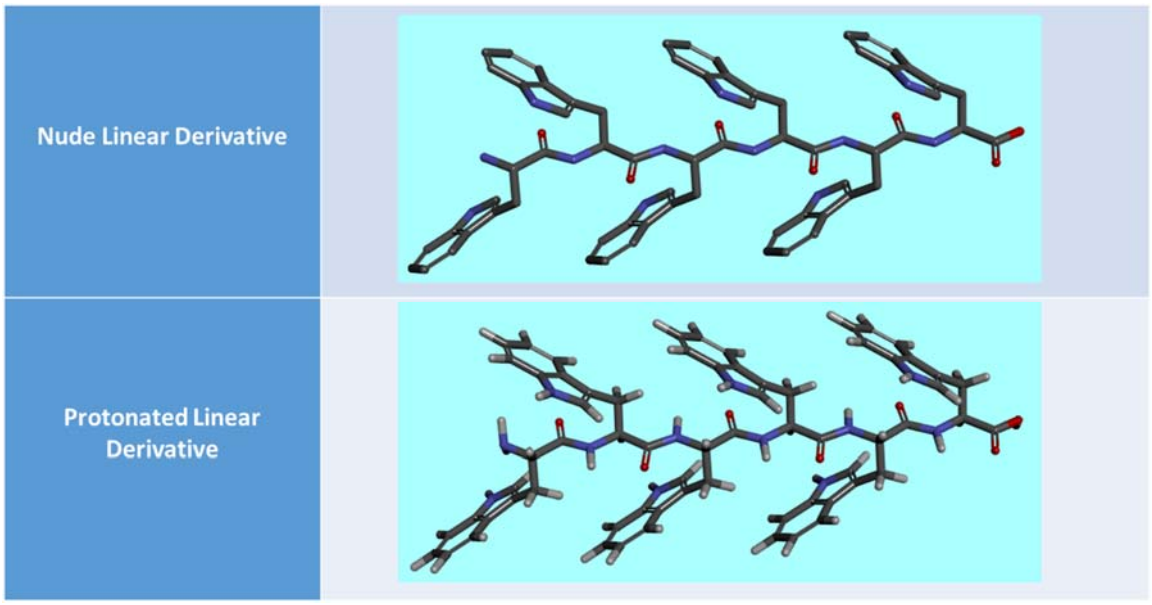


Figure 4.19: A single linear peptide derivative of the DerivatizeME enumeration protocol after exhaustive decoration (“nude”) and after protonation. *Example shows 6 tryptophan substituents immediately after enumeration (“nude”) and after protonation.*

The enumeration protocol produced peptides without protons, and we made use of the Discovery Studio Visualizer protonation protocol for the protonation and visualization of this peptide. As can be seen, the substitution points maintained bond distances and angles that satisfied rendering threshold limits for the expected hybridization (sp^3), and chirality was maintained. This was demonstrated by the number of H’s included on the Cα and Cβ atoms and the chirality of the Cα atoms. Although the Discovery Studio rendering thresholds are not a Universal benchmark we were satisfied with the geometry of derivatives produced by this enumeration protocol in terms of the bond lengths and angles at the points of introduction. The through-space transformations necessary to complete the introduction of the multi-atom substituents were shown to reliably position the entire substituents at their desirable point

on the query. The transformations were perfectly executed and the expected hybridization was maintained across the length of the peptide. We were satisfied with the performance of the DerivatizeME enumeration protocol on a linear peptide using amino acid substituents.

4.4.6.2 Cyclic peptides

For the population of a cyclic peptide virtual library, an 11 membered cyclic peptide of L-alanine residues was used to construct the backbone scaffold. Whereas the linear peptide had available 6 substitution points the cyclic peptides had 7 substitution points. The CycloPS enumerator was used to generate the cyclic conformer before Discovery Studio Visualizer was used to prepare the scaffold for decoration using the exhaustive DerivatizeME protocol. Biasing of the peptide scaffold towards cyclization was done by introducing two charged L-aspartic acid residues. It is known that during cyclic peptide synthesis, sequences that possess at least 1 charged residue between a stretch of five amino acids are less likely to aggregate during elongation (Duffy et al. 2011). In order to prepare for decoration all H atoms were deleted to prepare the "nude" cyclic peptide (Fig. 4.20). The 7 sites of substitution for the cyclic peptides were prepared by replacing each of the 7 alanine methyl group atoms with a single H atom. These 7 H atoms attached to the Cα directed all subsequent derivatives ensuring that the stereochemistry of each of the amino acids was preserved and correct. This was essential in assisting to reduce the complexity of the chemical space probed by the cyclic peptide library (compared to where both D- and L- amino acids could be used at each point). The substituents used for decoration were identical to those used for construction of the linear peptide library while the scaffold had an extra substitution point as a means of validating our enumeration protocol.

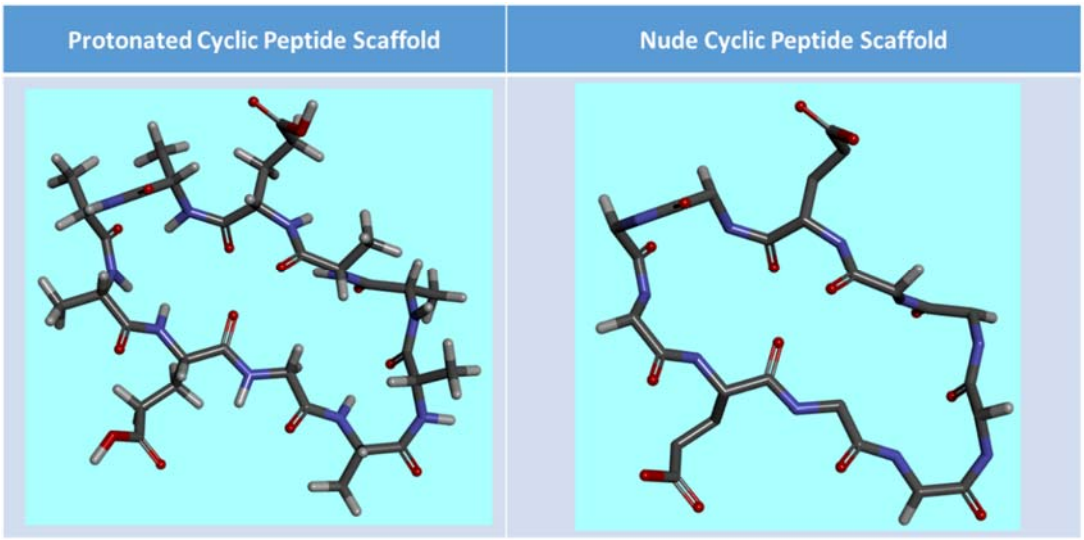


Figure 4.20: Ten membered cyclic peptide of alanine residues with charged aspartic acid groups. *The nude scaffolds were stripped of hydrogen atoms and hydrogen atoms remaining marked the substitution points.*

The protocol followed to allow for the population of the cyclic peptide library was identical to that followed for the linear peptide library. The virtual library of cyclic peptides accessed using DerivatizeME enumeration gave 7,595,553,552 (6 * 1,265,870,592) (7.5 Billion) individual peptides. This served as a further evidence for the reliability of the enumeration protocol. There were 27,132 different ways of choosing 6 substituents (as was the case for the linear peptides), but there were 279,936 (6^7) different ways in which 7 positions can be occupied by 6 substituents. The flexibility of our enumeration protocol allowed for the inclusion of compression protocols allowing compression of the target *".sdf"* to be performed immediately after it was written to the disk. The *"sdf"* file was removed from the disk immediately after the compressed *".zip"* file was written to the hard disk, in order to ensure sufficient storage space. Instead of requiring 9 Terabytes (1.5 T * 6) of storage for the cyclic peptide virtual library only 139 Gigabytes were required (Fig. 4.21).

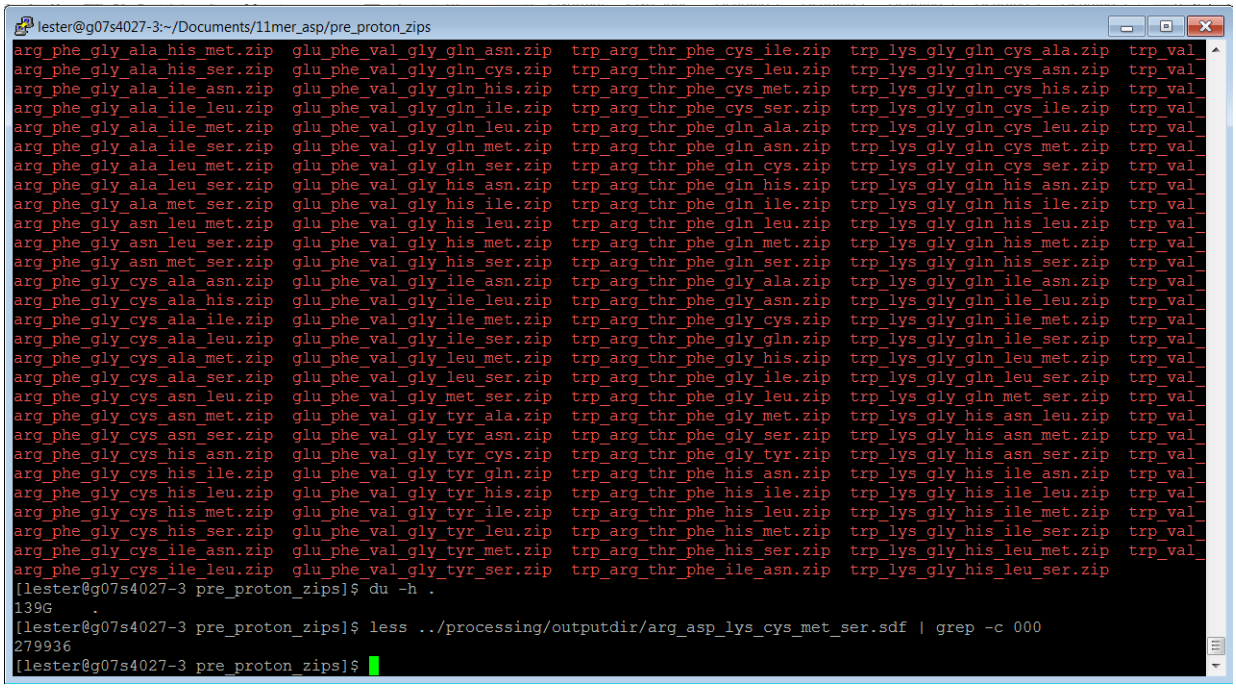


Figure 4.21: Screenshot of the storage memory allocated to the compressed cyclic peptide virtual library. *The total number of derivatives in a single collection is shown.*

A “nude” cyclic peptide is illustrated with its protonated counterpart illustrating the reliability of the enumeration protocol on cyclic peptide scaffolds (Fig. 4.22).

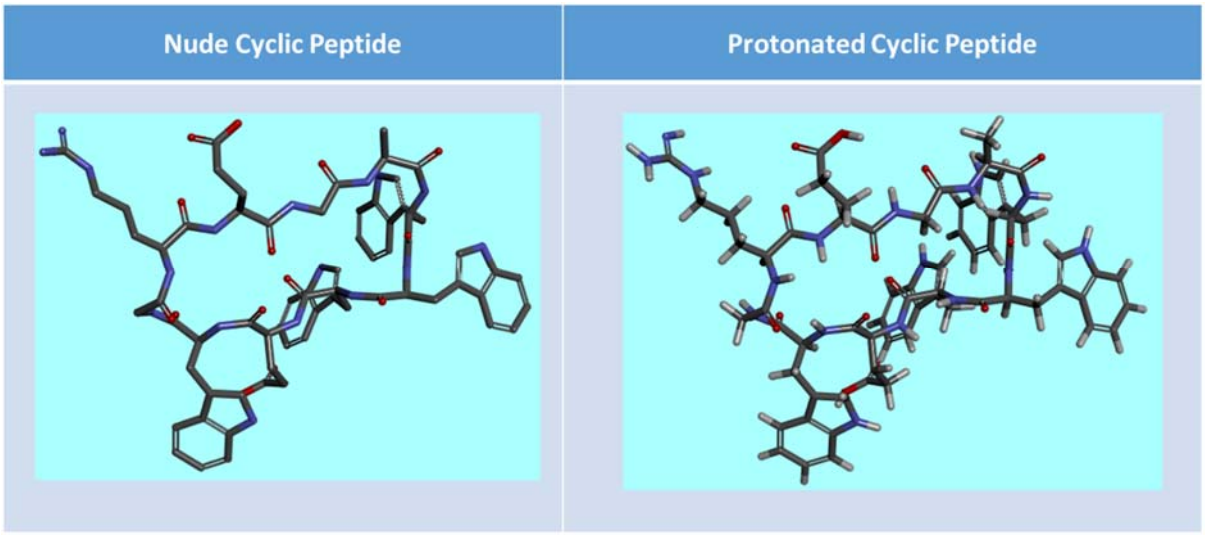


Figure 4.22: A single derivative of the DerivatizeME enumeration protocol after exhaustive decoration (nude) and after protonation*. Non-virtuous bonds highlighted (red).*

Local optimization of the close proximity sidechains may have the effect of optimizing the conformations produced by DerivatizeME (Fig. 4.22). The absence of energy minimization steps after substitution increases the computational efficiency but allows for probable steric clashes. In later versions of our DerivatizeME implementation it is envisaged that inclusion of energy minimization or conformational searching may enhance the precision of the derivatives produces by our algorithm. Later deployments of the exhaustive DerivatizeME enumeration for cyclic peptides will benefit from the use of a selection of scaffolds that survey the ring conformational space of the peptide scaffolds. A discussion on the choice of conformational search approaches and the scaffolds that approximate bioavailable conformations for cyclic peptides is conducted in Chapter 5. This combination in terms of ring conformation skeleton and enumeration of possible cyclic polypeptides forms the basis of Prototype 2, for creation of virtual libraries of cyclic polypeptides containing both chemical and conformational diversity.

4.5 Conclusion

The task of developing an in-house virtual library enumerator for the rational design of virtual libraries of peptide therapeutics has been presented. Two prototypes were explored in our development cycles. Prototype 1 was developed with the ambition of generating peptide conformations during chain elongation. Owing to the magnitude of the search problem and the inability to limit peptide conformations to bioactive conformations the strategy was not pursued to completion for drug discovery applications. Rather the more efficient protocol, Prototype 2, based on DerivatizeME, was developed with the conservative assumption that the problem of conformational search and enumeration could be decoupled. DerivatizeME was developed with a focus on efficient enumeration of a peptide scaffold. The DerivatizeME enumerator included the *makeallhydrogenssubstitutable*,

intelligentlymakehydrogensubstitutable, *makehydrogenssystematic* and the *makerandomhydrogenssubstitutable* routines which altered the choices of substitution during enumeration. For our purposes the exhaustive *makeallhydrogenssubstitutable* strategy was focused on for the population of virtual libraries, in which the scaffolds provided only had the appropriate hydrogen atoms present. The careful choice of multiple peptide scaffolds that survey the ring conformational space could be used to make a conformation-laden virtual library accessible. DerivatizeME worked well in a broader setting on a server, efficiently enumerating in concert with compression routines in order to manage storage space.

There are avenues for future development of the DerivatizeME-based enumeration strategy. For example simple substituent-based conformational searching may be implemented. In the context of cyclic peptides, with ready access to peptide topologies and API's such as OpenMM, full molecular mechanics optimizations will be possible, although this will carry a heavy computational cost. This Protocol 2 relies on other methods to generate ring conformations, and DerivatizeME is meant to focus on substitution patterns, and future development will rely more on coupling DerivatizeME, rather than introducing ring-conformation search routines. There are minor limitations relating to the peptide scaffold used by DerivatizeME, particularly where it comes to proline residues, which would have to be incorporated into the scaffold prior to decoration.

Despite the limitations discussed we were satisfied with DerivatizeME as an enumerator that could be used to populate conformation-laden virtual libraries of cyclic peptides within a confined sequence space. Further discussions in Chapter 5 resolve the conflicts associated with accessing bioactive ring conformations for cyclic peptides in the realization of our goals.

Chapter 5: Enhanced sampling of cyclic peptide scaffolds using Replica-Exchange Molecular Dynamics simulations

5.1 Introduction

Macrocyclic scaffolds are a source of under-explored bioactive molecular entities due to their high affinity for previously un-druggable targets and sites such as those involved in protein interactions (Allen et al. 2016; Marsault & Peterson 2011; Sperandio et al. 2010). The solvent-exposed, flat surfaces of difficult to drug targets require inhibitors with many potential interaction sites in order to maintain strong contacts with the active sites (Rask-Andersen et al. 2011; Dandapani & Marcaurelle 2010). Macrocyclic scaffolds are ideal for this while their cyclized nature restricts the large number of torsion degrees of freedom present (such as are present in linear molecules) with large numbers of contact points. Cyclization lowers their entropy, enhancing the Gibbs free energy associated with their improved binding and ligand efficiency (Rask-Andersen et al. 2014; Hewitt et al. 2015; Bockus et al. 2013).

However, cyclic peptide macrocycles provide a unique challenge to strategies that aim to accelerate the drug discovery process using computational approaches. This challenge arises from the difficulty in extracting and prioritising the bioactive conformations that are relevant to investigations of their potency and selectivity (Kessler et al. 1996; Anighoro et al. 2016). The screening methods used in the prediction of binding modes of ligands with their therapeutic targets employ conformational search procedures (in the presence of the target) for the reliable prediction of ligand binding poses (De Vivo et al. 2016; Campbell et al. 2014; Allen et al. 2016). Computational conformation search strategies are effective in exploring the conformational space of ligands that possess less than 20 torsion bonds (Coutsias et al. 2016). When the number of torsions exceed this limit, it rapidly becomes too computationally expensive to adequately explore the vast conformational landscape and successfully identify

bioactive conformations (Coutsias et al. 2016; Paissoni et al. 2015). Macrocyclic systems offer an added complication due to kinetic trapping that prevents conformation switching through pathways that involve high-energy concerted motions (Räder et al. 2018; McHugh, Rogers, Yu, et al. 2016).

Some of the options available for computational approaches to extend the exhaustiveness of conformational searching are the use of stochastic Monte Carlo based strategies (Diller & Merz 2002); the use of methods that bias conformational search along low frequency vibrations by Low Mode approaches (Labute 2010; Kolossvary & Keseru 2001); incorporating strategies such as Mixed Mode dynamics that use a combination of stochastic and Low Mode evaluation (Anighoro et al. 2016; Chen & Foloppe 2013); the application of energy function or substructure topology guided optimizations (Coutsias et al. 2016; Lei et al. 2004); or the use of molecular dynamics dependent approaches (McHugh, Rogers, Solomon, et al. 2016; Tsujishita & Hirono 1997; Abrams & Bussi 2014).

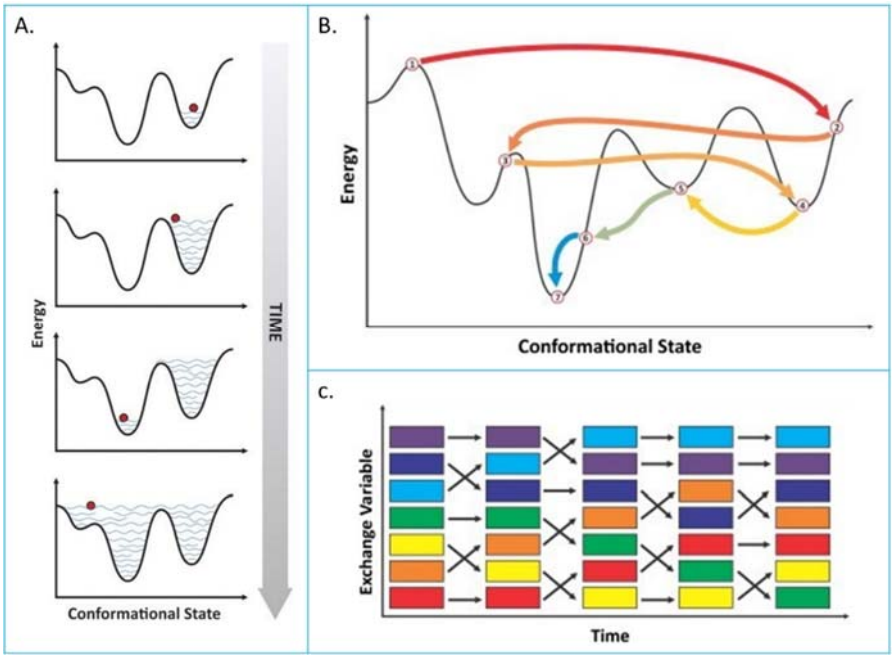


Figure 5.1: Enhanced sampling molecular dynamics approaches. *A. Metadynamics enhanced conformational search. B. Simulated annealing search. C. Replica exchange MD. Image source (Bernardi et al. 2015).*

Despite their computational expense, molecular dynamics (MD) based approaches with enhancements to mitigate against kinetic trapping are favourable towards conformation searches of macrocyclic systems (De Vivo et al. 2016; Oakley et al. 2013). Simulated annealing MD relies on performing simulations at a high temperature before rapidly decreasing the temperature to trap conformations (Fig. 5.1 B), while metadynamics prevents oversampling of a local energy minimum by introducing memory information during the simulation (Fig. 5.1 A) (Oakley et al. 2013; Paissoni et al. 2015). In Replica-Exchange Molecular Dynamics (REMD), replicas of simulations under different state conditions have over-lapping potential energies and are allowed to exchange coordinate information during the course of the REMD simulation (Fig. 5.1 C) (Abrams & Bussi 2014; Wakefield et al. 2015; Lei & Duan 2007). These exchanges enhance ergodicity within the simulation allowing conformations to overcome kinetic trapping that creates a barrier on the potential energy surface preventing the exploration of conformational space.

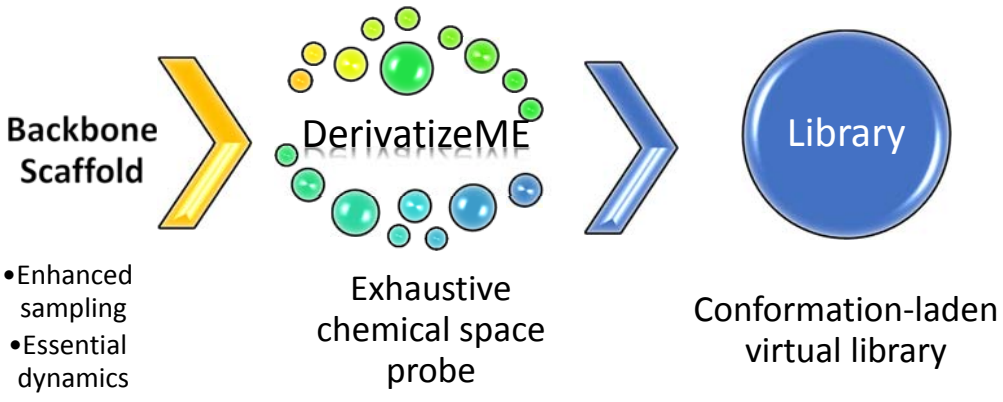


Figure 5.2: The strategy employed in the population of the peptide virtual library.

The acceptance of an exchange between two replicas in REMD is monitored by the satisfaction of a Metropolis criterion whose probability is dependent on the potential energies of the adjacent replicas at different temperatures (Okabe et al. 2001). The frequency and number of exchanges between replicas across a temperature series is impacted by the choice of temperatures, while the number of replicas used is influenced by the computation time and resources available (Wakefield et al. 2015). Numerous methods exist that guide the choice of temperatures used to achieve sufficient REMD mixing and sampling in simulations of coarse-grained models or atomistic polypeptide models (Rathore et al. 2005). In particular, the Patriksson Temperature Generator uses predefined constants from systems in equilibrium to predict the series of temperatures (Patriksson & Van Der Spoel 2008) (Eq. 18).

$$\langle P(T_1 \leftrightarrow T_2)\rangle = \int_{-\infty}^{\infty} P(T_1 \leftrightarrow T_2)\rho_{U_1-U_2}(u)\,du$$

$$= \frac{1}{2}\left[1 + erf\left(-\frac{\mu_{12}}{\sigma_{12}\sqrt{2}}\right)\right] + \frac{1}{2}e^{\left(C\mu_{12}+\frac{C^2\sigma_{12}^2}{2}\right)}\left[1 + erf\left(\frac{\mu_{12}+C\sigma_{12}^2}{\sigma_{12}\sqrt{2}}\right)\right] \quad (18)$$

The temperatures, T, that maintain a desired probability of exchange can be predicted from the number of water molecules, protein atoms, constraints and virtual sites of the polypeptide. These terms predict the average energy, μ_{12}, and the standard deviation, σ_{12}, used to describe the energy distributions of U_1 and U_2 (Eq. 18).

The aim of our work is to probe both the chemical and conformational space of cyclic peptide systems. The goal of the work presented in this Chapter was to identify the essential dynamics of a peptide macrocycle using replica exchange molecular dynamics (REMD). This enhanced sampling protocol facilitated high-resolution conformation extraction of the peptide to guide

the selection of scaffolds for the population of conformation-laden cyclic peptide virtual libraries using our DerivatizeME enumeration engine.

5.2 Methods

5.2.1 Analogues used in this study

There were two types of analogues used in this study. The first set of analogues investigated the impact of reducing the side-chain contribution to conformation by substituting some side-chain groups with methyl groups (mutating the residues to alanine) from the original cyclosporine-A (CsA) model derived from a crystal structure complex (PDB ID: 2Z6W) (Fig. 5.3). The CsA conformer (with a rotated Leucine 8) was obtained from the crystal structure and their co-ordinates were preserved. Mutations were introduced by substituting MLeu 2, MVal 4, Aba 6, and Val 9 with Ala groups (Fig 5.3).

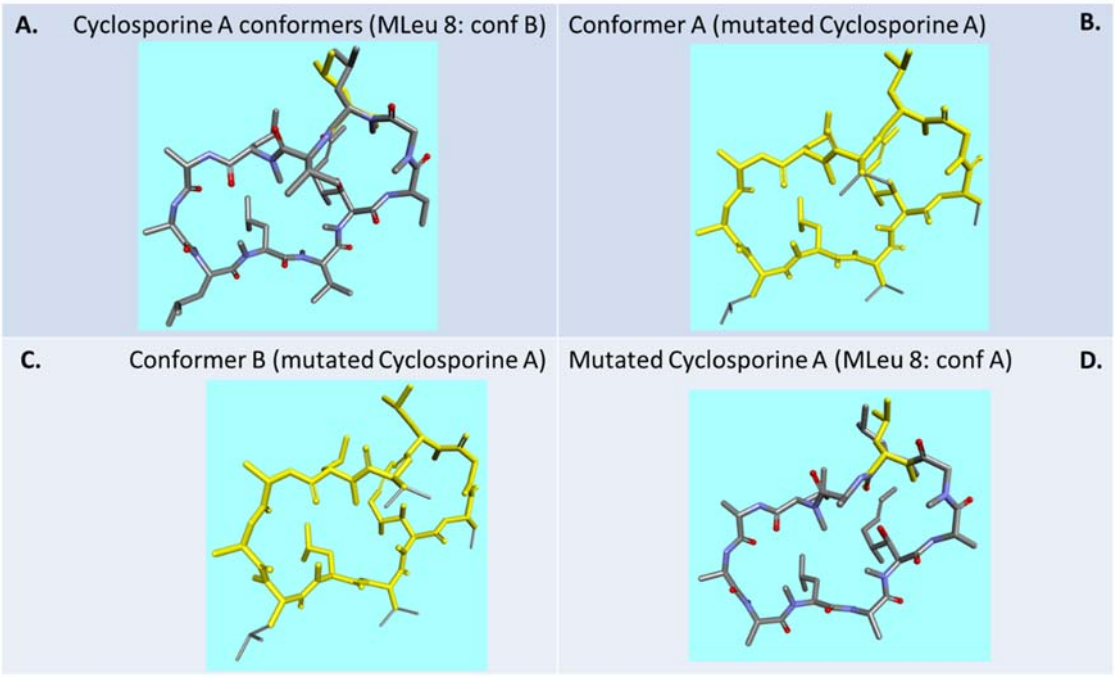


Figure 5.3: Investigative Cyclosporine-A analogues used in the study (brackets show highlighted analogue). *A. Overlap of cyclosporine A conformers. The rotated MLeu for conformer B is highlighted. B. Overlap of conformer A analogues. The mutated cyclosporine A is highlighted. C. Overlap of conformer B analogues. The mutated cyclosporine A is highlighted. D. Overlap of the mutated cyclosporine A analogues. The rotated MLeu from the conformer A is highlighted.*

The second set of analogues was of an example cyclic peptide. The CycloPS enumerator was used to generate an 11 membered (11mer) cyclic peptide of L-alanine residues with two charged L-aspartic

acid residues included in order to bias the peptide towards solid-phase synthesis. The same sequence was used to generate the scaffolds for the population of the virtual library (Fig. 5.4).

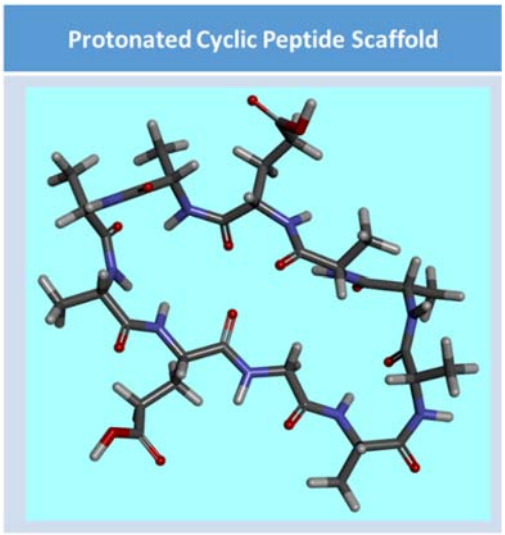


*Figure 5.4: **Eleven membered cyclic peptide of alanine residues with charged aspartic acid groups.***

5.2.2 Conformational Search

5.2.2.1 Molecular Dynamics simulations

The small peptides were protonated at pH 7.4 using OpenBabel 2.3.2, prior to the generation of ligand topologies using the antechamber python parser interface (acpype v. 2012-09-13). Within acpype, the bcc MOPAC module was used to calculate atom specific charges of the ligand atoms. Each cyclic peptide was minimized using a steepest descent algorithm for 500 steps, prior to being subjected to another 500 steps of restrained molecular dynamics using the Generalized Amber Force field (GAFF)(Wang et al. 2004).

The simulation systems of the macrocycle models were solvated by SPCE waters and neutralized by chlorine ions before 250 ns worth of classical simulations under a V-rescale modified Berendsen thermostat set at 300 K were performed. Isotropic pressure coupling was achieved through the Parrinello-Rahman coupling set at 1.0 bar. The amber03.ff parameters

for ions, solvent and ligand atoms were used solve the equations of motion under a leap-frog integrator with 2 fs time-step intervals. Removal of bond oscillations by treating all-bonds as constraints during the production run ensured that deductions of motions and configurational space from simulations that used longer time steps could be used reliably.

5.2.2.2 Replica Exchange Molecular Dynamics

The REMD approach was used in order to enhance the sampling of the macrocyclic systems. The equations of motion during the REMD dynamics were solved using forces calculated also using the amber 03 force field. The REMD Temperature Generator was used to define the temperature series for the REMD simulations that could maintain a probable exchange rate of 0.4. The REMD experiments performed were the exchange interval test, the substituent test and conformation extraction test. Table 5.1 shows the parameters used for the REMD Temperature Generator.

Table 5.1: Parameters used for the REMD Temperature Generator

Parameters	$N_{protein}$	N_{water}	N_H	***Constraints***	T_{limits}	P_{des}
Values	172	1,668	94	Fully Flexible	272 – 410	0.4

5.2.2.2.1 Exchange interval test

In order to determine the appropriate replica exchange interval that was efficient for the compute resources available three exchange intervals were tested (0.2 ps (replex 100), 1 ps (replex 500) and every 2 ps (replex 1000). The mutated cyclosporine conformer A (metA_ala) equilibrated macrocycle model (obtained after normal MD) was cloned into 24 replicas for each exchange experiment. Statistical tests on the average exchange rates maintained across the temperature series during the simulation were performed. The conformer diversity extracted from the lowest temperature trajectories for each exchange interval was assessed.

5.2.2.2.2 REMD of cyclosporine analogues

To test the robustness of the temperature series and the impact of different alanine mutations on cyclosporine, the 4 cyclosporine models were subjected to 40 ns of REMD simulations across 24 replicas with temperatures ranging from 296.1 K (replica 0) to 399.5 K (replica 23). For each of the 4 investigative systems, a stabilized macrocycle model obtained during normal MD was cloned into the 24 replicas.

5.2.2.2.3 REMD of cyclic peptide scaffold

The protonated 11mer cyclic peptide obtained from a normal MD run was cloned into 24 replicas and subjected to 40 ns of REMD simulation enhanced sampling.

5.2.2.3 MacroModel Conformer search

In order to test the extent of conformational sampling of the REMD approach the protonated 11mer macrocycle was subjected to conformational searches using different MacroModel algorithms optimized for macrocycle conformation searching (Watts et al. 2014). Four conformational search algorithms were used, *viz.* a) the MacroModel Monte Carlo multiple minimum search (torsional sampling) routine, b) low frequency molecular dynamics (the Low Mode sampling) approach, c) a hybrid of the torsional sampling and Low Mode algorithm, and d) a hybrid Largescale sampling with Low Mode technique. For each routine default parameters were used for water solvation during the conformation search and the best 10 conformers were recorded after 1000 search steps (minimization algorithm = PRCG; convergence threshold = 0.05 (1.0 for large scale); converge on gradient; energy window for saving structures = 21.0 kJ/mol; cut-off for different atoms = 0.5 Å). The OPLS 3 force field was used during these simulations.

5.2.3 Analysis of Molecular Dynamics

5.2.3.1 Trajectory analysis

For all MD simulations trajectories were extracted and analyzed for quality, compactness, RMSD convergence, distance fluctuations and hydrogen bond interactions. The multiple replica trajectories were aligned prior to concatenation in order to eliminate translational and rotational errors during simulation.

5.2.3.2 Diversity analysis

PCA analysis was performed on the trajectories in order to determine the extent of the sampling approach reflected as the diversity in structures. For multiple trajectory simulations the low energy replica (replica 0) trajectory was used for conformation analysis.

5.2.3.3 Extraction of essential dynamics

The essential dynamics of the simulations were obtained using clustering of the conformations from the RMSD matrix of the low temperature trajectory. The deterministic non-iterative Jarvis-Patrick clustering according to similarity routine was used to match structures into the same clusters if they were within an RMSD neighbor list. In order to determine the neighbors in common to cluster parameter when 10 members are examined at a time the generous (2), medium (3) and strict (4) members in common criteria were tested. Structures extracted using these routines were analyzed by computing the solvent accessibility, the radius of gyration and the total number of pharmacophore fingerprints (Hydrogen Bond acceptor (heavy atom), Hydrogen Bond donor (heavy atom), Negative ionizable, Positive ionizable and Hydrophobic groups) across the ensemble set using Discovery Studio.

5.2.4 Scaffold preparation

The Jarvis-Patrick clustering routine was used to interrogate the conformers that represent the Essential Dynamics of the systems being simulated. The representative conformers that

were the most populated were isolated and used as scaffolds whose conformations account for the essential dynamics from the enhanced sampling protocol. The representative conformers were prepared manually for enumeration as scaffolds for the population of a conformation laden virtual library of cyclic peptide.

5.3 Results and Discussion

5.3.1 Standard Molecular Dynamics Simulation

The temperature, pressure and density maintained during the simulations was analyzed in order to confirm the quality of the simulations. These thermodynamic properties reveal that the cyclosporine systems oscillated around mean values for temperature, pressure and density during the duration of the simulations as expected (Fig. 5.5).

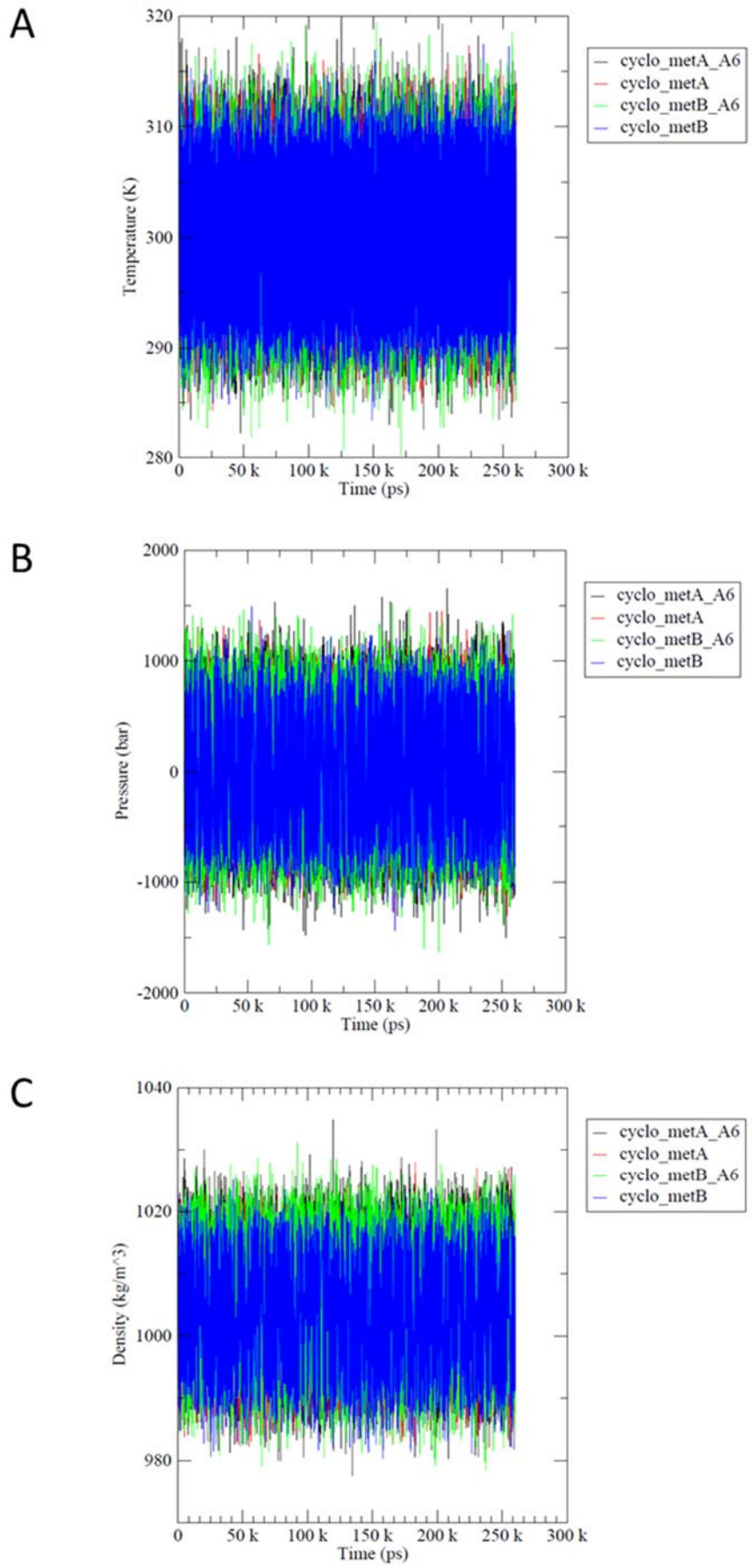


Figure 5.5: Thermodynamic properties of the cyclosporine analogues maintained during 250 ns of constant volume and pressure molecular dynamics simulations. *metA_A6 (black), metA (red), metB_A6 (green) and metB (blue). A. Temperature plots, B. Pressure plots, C. Density plots.*

The compactness of the peptides during the simulation was assessed by monitoring the radius of gyration (Rg)(Fig. 5.6). The Rg shows oscillations of the magnitude of compactness maintained during the simulations. The range in Rg is an indicator of the range of conformations generated from open to closed conformations. Further, although visual examination of the structure during dynamics affirms that no terminal residues have been introduced and that the cyclic peptide has not been linearized, the Rg would also show significant less-reversible reduction in Rg should linearization of the peptide have occurred.

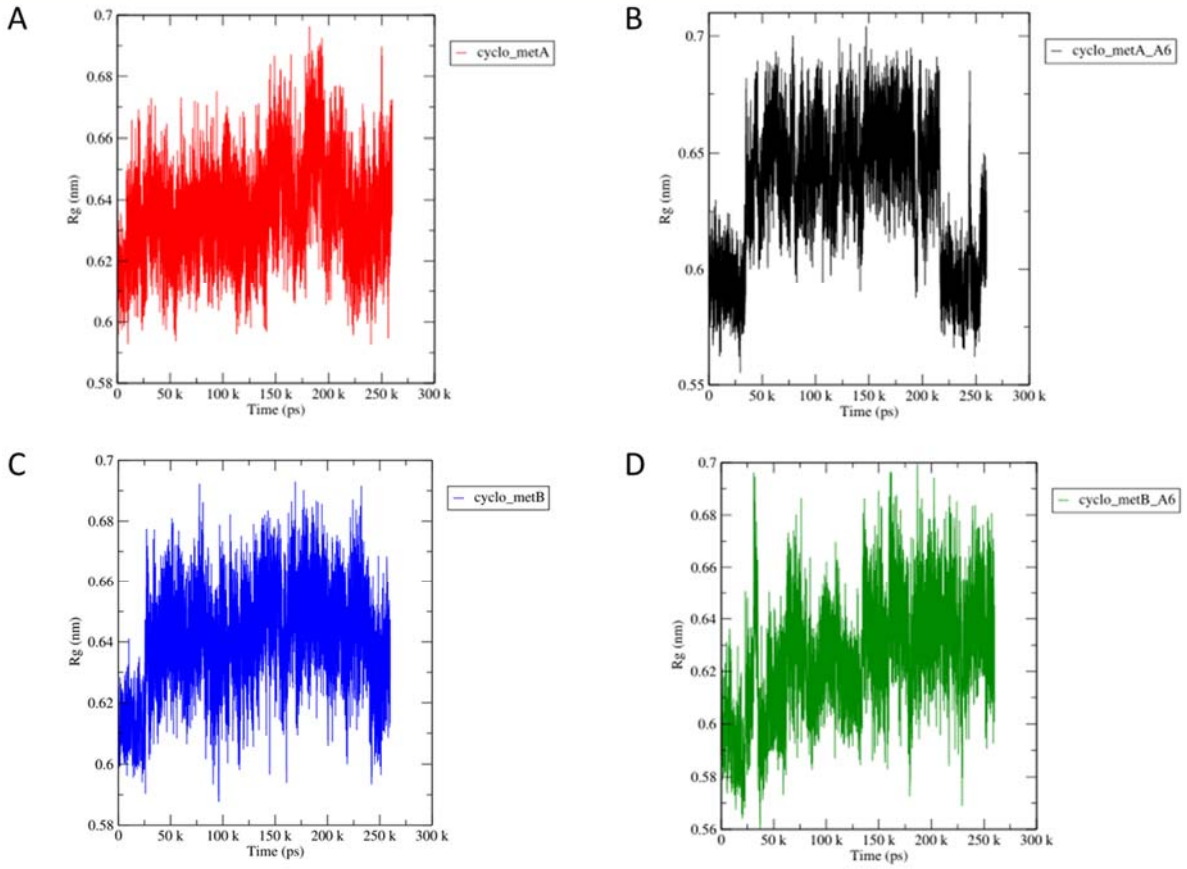


Figure 5.6: Peptide Radius of Gyration plots for cyclosporine analogues. A. metA (red), B. metA_A6 (black), C. metB (blue) and D. metB_A6 (green).

The structural changes that were occurring during the simulations were interrogated through an analysis of the change in the RMSD of the heavy atoms of the peptide from their starting structure (Fig. 5.7).

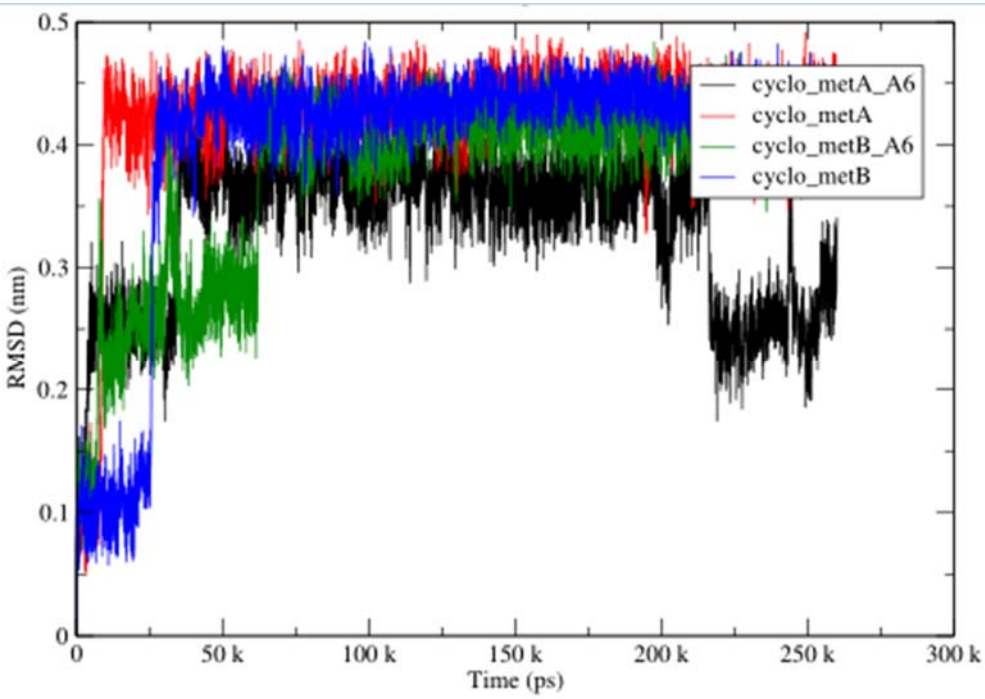


Figure 5.7: RMSD of the cyclosporine analogues compared to the original starting structure. *metA_A6 (black), metA (red), metB_A6 (green) and metB (blue).*

From the RMSD plots it can be seen that the peptide structures rapidly adopt different conformations from their starting structures evident from an increase in the RMSD for all systems at the start of simulation. The mutated cyclosporine systems (metA_A6 and metB_A6) display lower RMSD deviation than the cyclosporine conformers (metA and metB). RMSD is a cumulative term and the more atoms there are present within a system translates into an increase in RMSD. The mutations also reduced the number of rotating atoms from 196 to 172. This decrease in the number of atoms fluctuating translates into a 12.5 % difference in total contribution to RMSD. The magnitude of the change in RMSD of the mutants is much lower (~ 0.1 nm) than the original conformers (~ 0.3 nm). This observation is consistent with

the observation that the side-chains present in the conformers were biased towards the protein binding pocket of the original crystal structure. The reduced number of side-chains in the mutant conformers translates into a smaller contribution to the RMSD change.

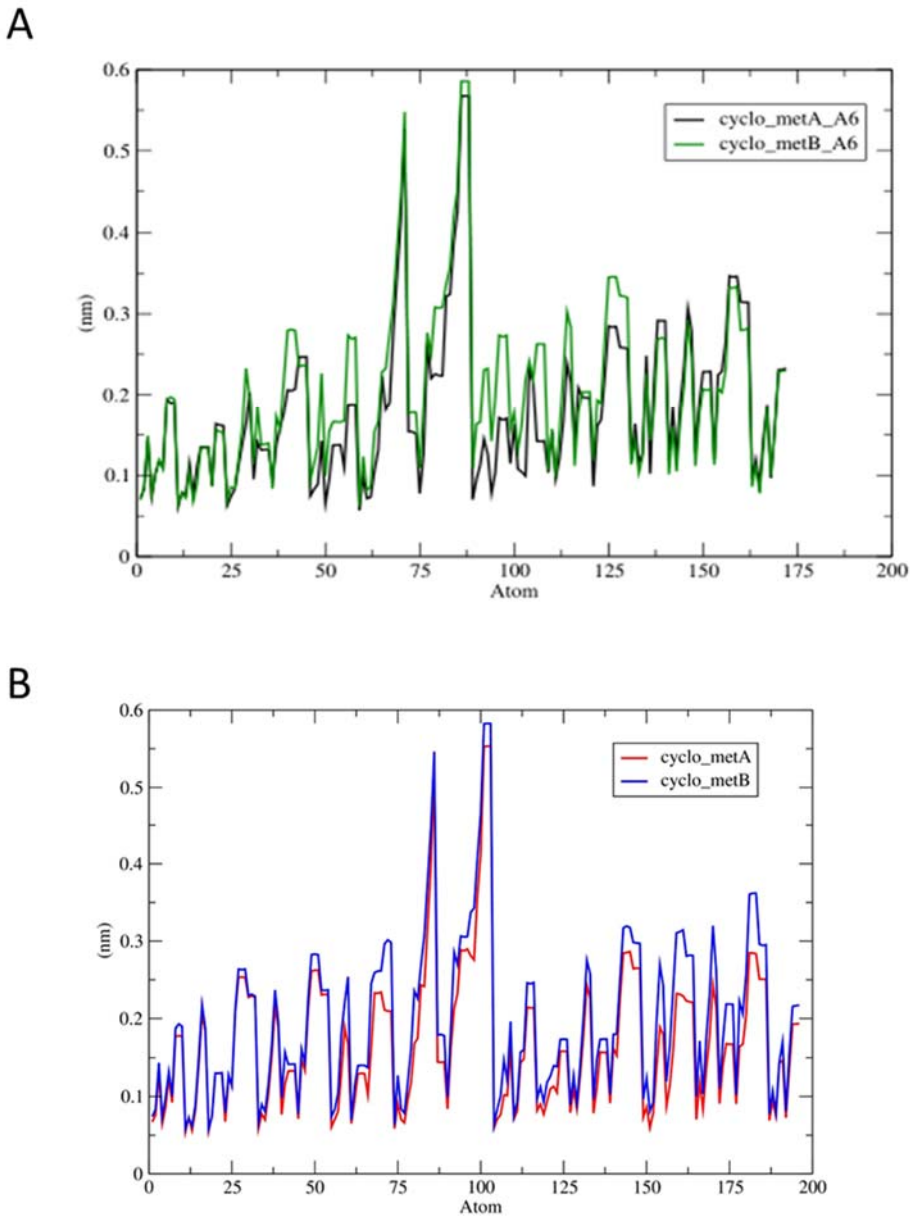


Figure 5.8: Root mean squared fluctuations (RMSF) of cyclosporine analogues obtained after constant volume and pressure MD simulation (250 ns). *A. mutant analogues metA_A6 (black) and metB_A6 (green). B. original conformers metA (red) and metB (blue).*

The trends observed from the RMSF plots show that the B conformers (with rotated leucine) have slightly higher fluctuations than the A conformers in both the mutants and the original conformers during the simulations (Fig. 5.8). Although there are significant fluctuations in the sidechain atoms, the backbone atoms have fluctuations lower than 0.1 nm for both sets of analogues. This highlights the rigidity of backbone atoms and shows that the majority of the structural deviation observed in this system under the simulation conditions were as a result of side-chain fluctuations. The observation that no atoms have an RMSF of 0 nm shows that the systems are fluid even though the backbone dynamics are small.

5.3.1.1 Dynamics of the 11mer cyclic peptide

The trends in thermodynamic and structural properties (temperature, pressure, density, Rg, RMSD, RMSF) during simulation of the 11mer cyclic peptide are illustrated in Fig. 5.9. The structural compactness (expected of a cyclic system) of the 11mer peptide is maintained throughout the simulation with oscillations highlighting the conformational sampling during the simulation (Fig 5.9 D). The compactness and structural deviation of the 11mer peptide is smaller than that of the cyclosporine analogues (metA, metB, metA_A6, metB_A6). The increased compactness and reduced flexibility of the 11mer peptide is reflected in the peak RMS fluctuation of less than 0.3 nm during the simulation where as a fluctuation of greater than 0.5 nm was observed in all the cyclosporine analogue atoms during the simulation. From these observations we can infer that the 11mer peptide has less conformational mobility during these simulations than the cyclosporine analogues.

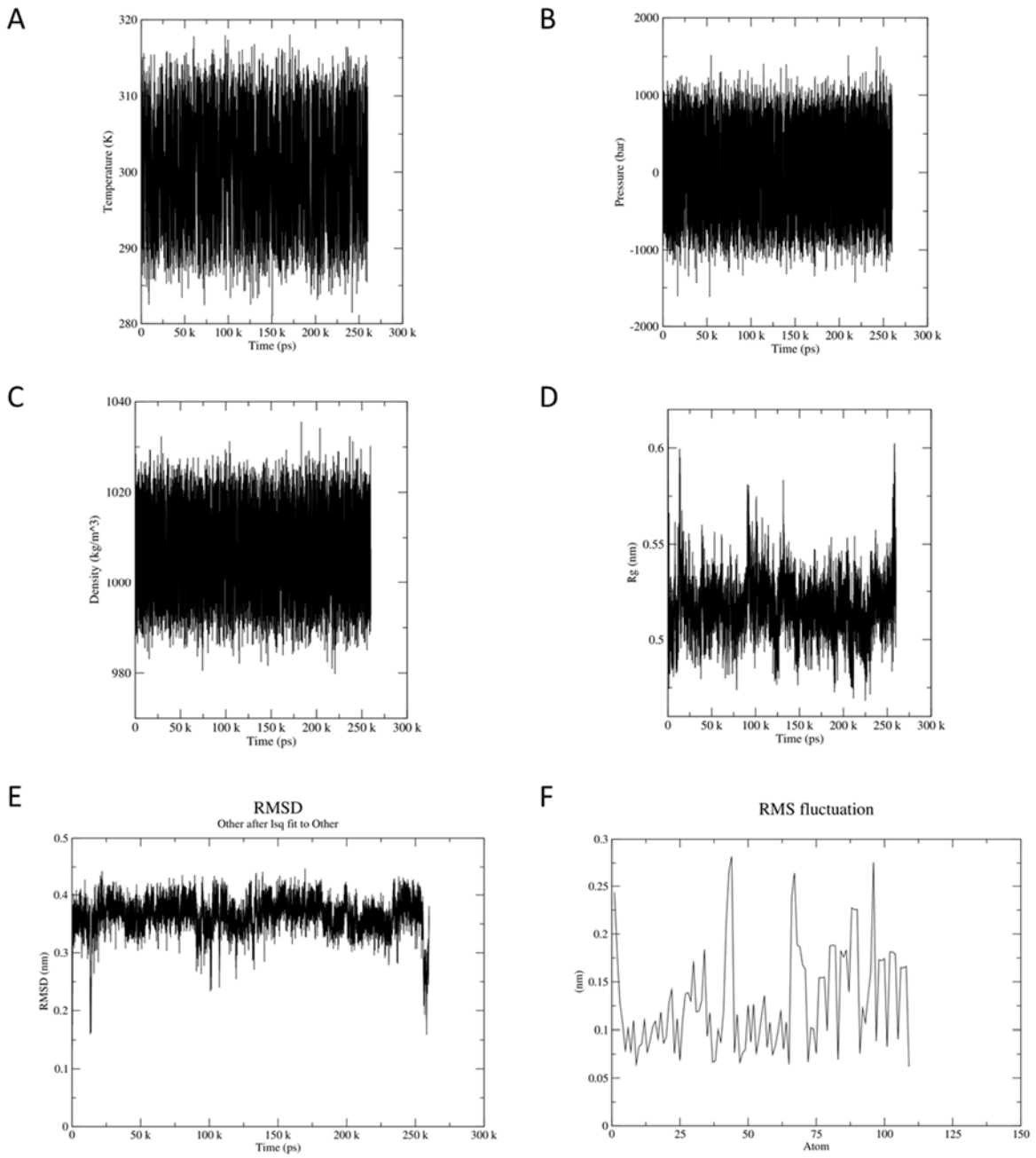


Figure 5.9: Thermodynamic and structural properties during constant volume and pressure simulation of the 11mer cyclic peptide. *A. Temperature plot. B. Pressure plot. C. Density plot. D. Radius of Gyration plot. E. Root Mean Square Deviation. F. Root Mean Square Fluctuation plot.*

5.3.2 Replica Exchange Molecular Dynamics

The same force-field topologies (as were used in normal MD) were used in the REMD calculations of the metA, meB, metA_A6 and metB_A6 cyclosporine analogues. The view was to explore the extent of conformational sampling from REMD.

5.3.2.1 Exchange interval test

Exchange interval tests (for REMD) were performed on the metA_A6 cyclosporine mutant analogue. The three exchange intervals tested were (0.2 ps (replex 100), 1 ps (replex 500) and every 2 ps (replex 1000) over the temperature range extracted from the temperature generator. A high rate of exchange will increasing mixing but reduce conformational transitions. A lower rate of exchange may increase conformational transitions but reduce the mixing and thus conformations accessed by transitions at higher temperatures might not be made visible at lower temperatures. The interval test also enable us to test the reliability of the temperature series that we selected. An optimum temperature series will allow potential energies between adjacent replicas to overlap across the simulation.

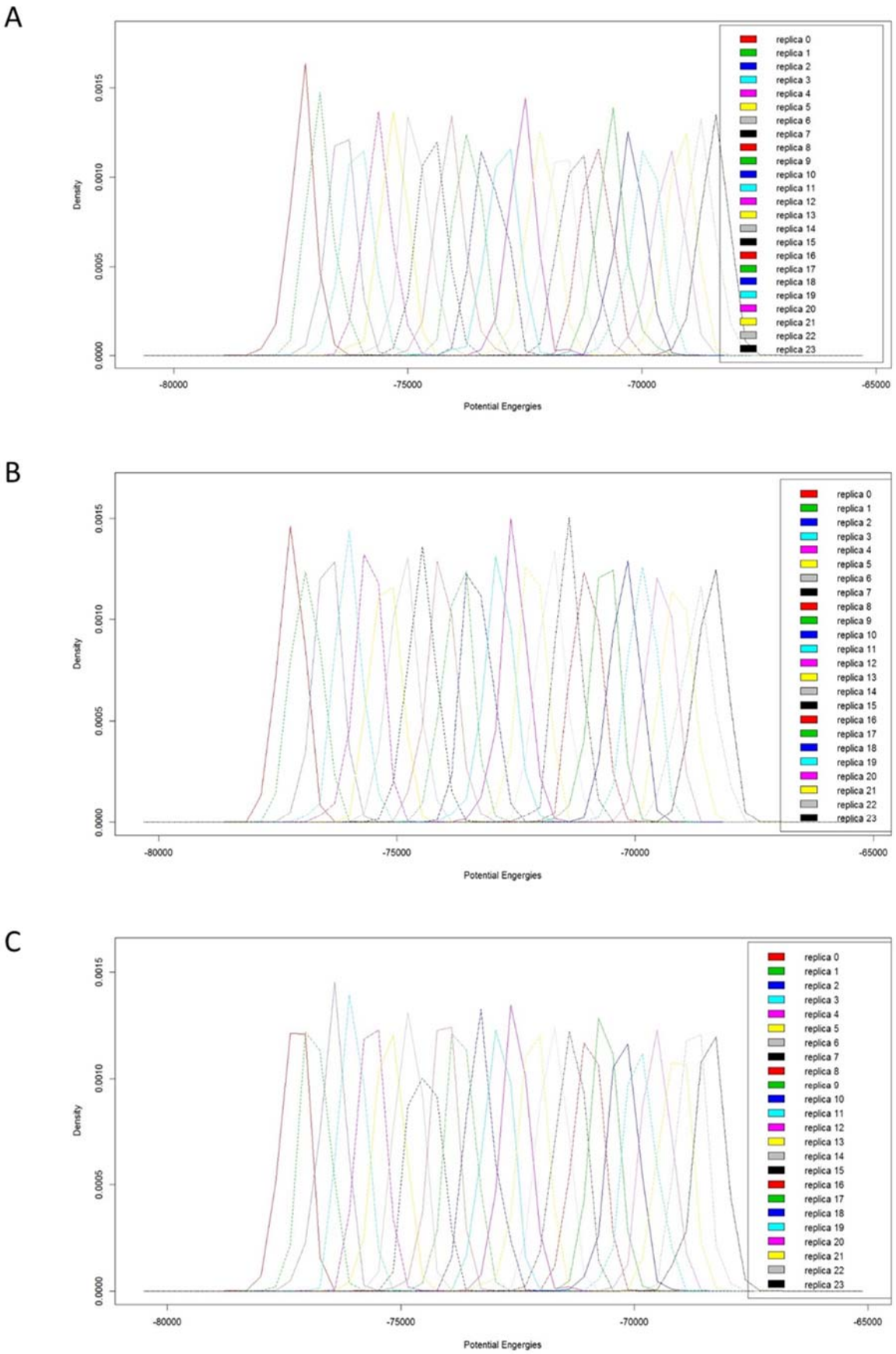


Figure 5.10: Potential energy distribution plots of REMD simulations of the metA_A6 analogue. *A. Replex 100, B. Replex 500, Replex 1000.*

An even distribution of potential energies was maintained throughout all the exchange intervals during the simulations (Fig. 5.10). There was satisfactory overlap between the potential energies of adjacent replicas in all three cases. The probability of exchange is influenced by the potential energy of adjacent replicas while the average exchange is influenced by the satisfaction of a Metropolis criterion for the acceptance of an exchange either forward or reverse, as described earlier. Data obtained from the Transition Matrix of the simulation is shown (Fig. 5.11).

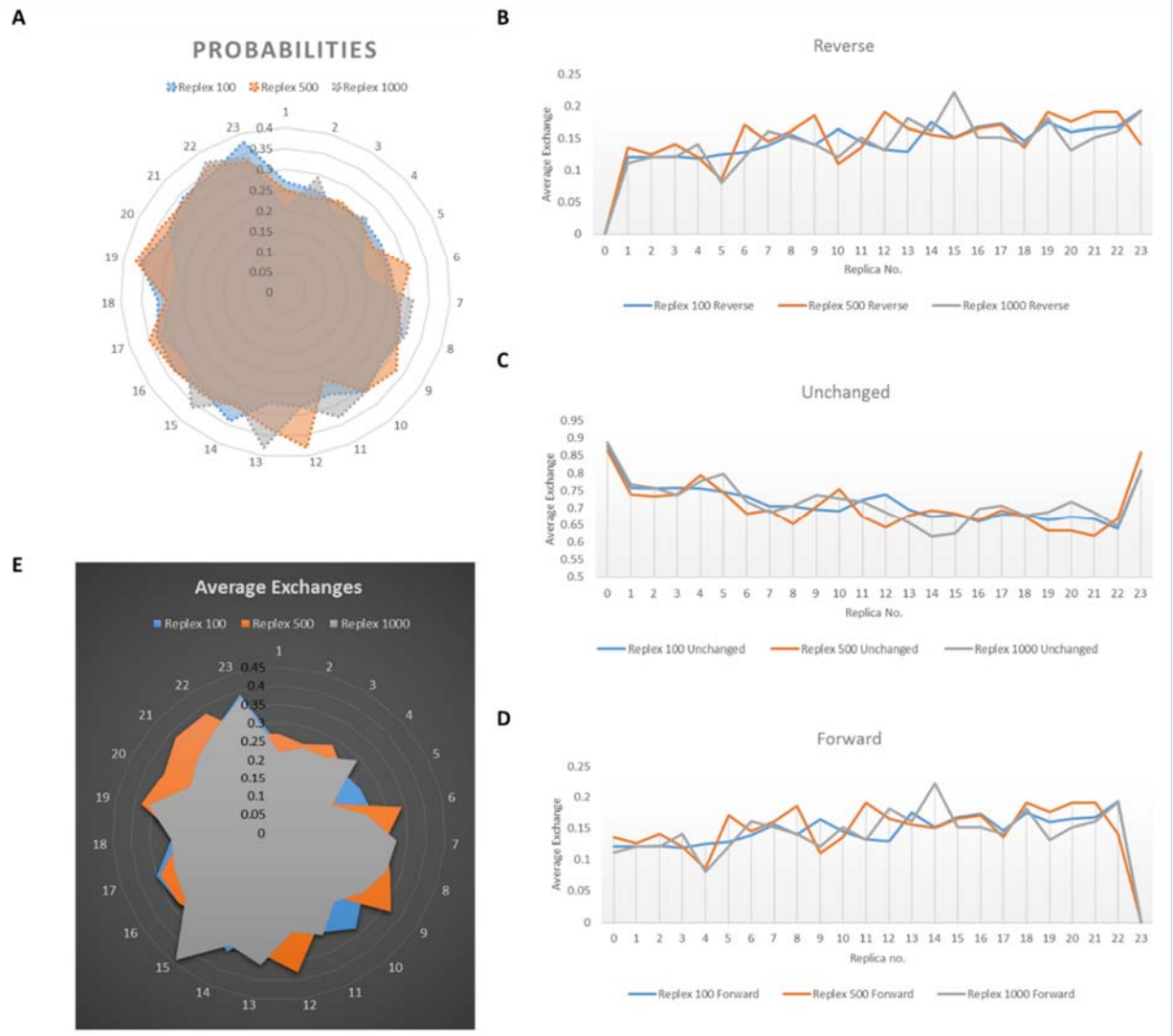


Figure 5.11: REMD exchange events between the different exchange intervals of analogue metA_A6. *A. Probability of exchange. B. Average reverse exchange acceptance. C. The average exchange rejection. D. The average forward exchange acceptance. E. The average forward and reverse exchange. Replex 100 (blue), Replex 500 (orange), Replex 1000 (grey).*

The metA_A6 cyclosporine analogue undergoes exchanges that are controlled by the satisfaction of the Metropolis criterion during the course of the simulation. The forward and reverse exchange events exhibited a similar trend across the different replicas. The Replex 100 represented an acceleration of the exchange process while Replex 1000 served as a delay of the exchange process. A trend in the unchanged replica highlights that there is an increase in the likelihood of exchange as temperature is increased (Fig 5.11 C). The Replex 500 exchange interval appears to have maintained a higher exchange rate of systems and this is more obvious at the higher temperature replicas. The implications of higher exchange rates at higher temperatures would mean that systems do not sample conformations accessible at these higher temperatures as efficiently as they sample conformations at lower temperatures. A positive implication for our study is that conformations that are accessed at higher temperatures have a higher probability of transitioning to the lower temperature replicas during this simulation. As such, for the remainder of the study the Replex 500 exchange interval was used for enhanced sampling.

5.3.3 Diversity analysis through PCA

In order to explore and visualize the diversity of the conformations extracted from both the Normal and the REMD simulations an analysis of the Principle Components of the variances of the structural deviations during the simulations was undertaken. A structure matrix is composed of the Cartesian coordinates of the atoms present in the peptide for each time stamp. After structure alignment a covariance matrix is compiled that records the difference of each position from an average position over the entire trajectory. Linear algebra is used to determine the Eigenvalues and Eigenvectors that span the Eigenspace of the covariance matrix. The Eigenvectors and Eigenvalues of the matrix can be extracted and plotted for each

time frame in order to allow for easier exploration of the diversity of the ensemble of structures generated by the simulation.

The Eigenvectors that describe the structures were plotted at each time step during the simulation (Fig 5.12). Each system was represented by the Eigenvector that accounts for the majority of the variance (Eigenvector 1). It must be noted that because a normalization constraint is applied during the linear algebra routines these projections are independent of time (they do not describe what happened at a particular time).

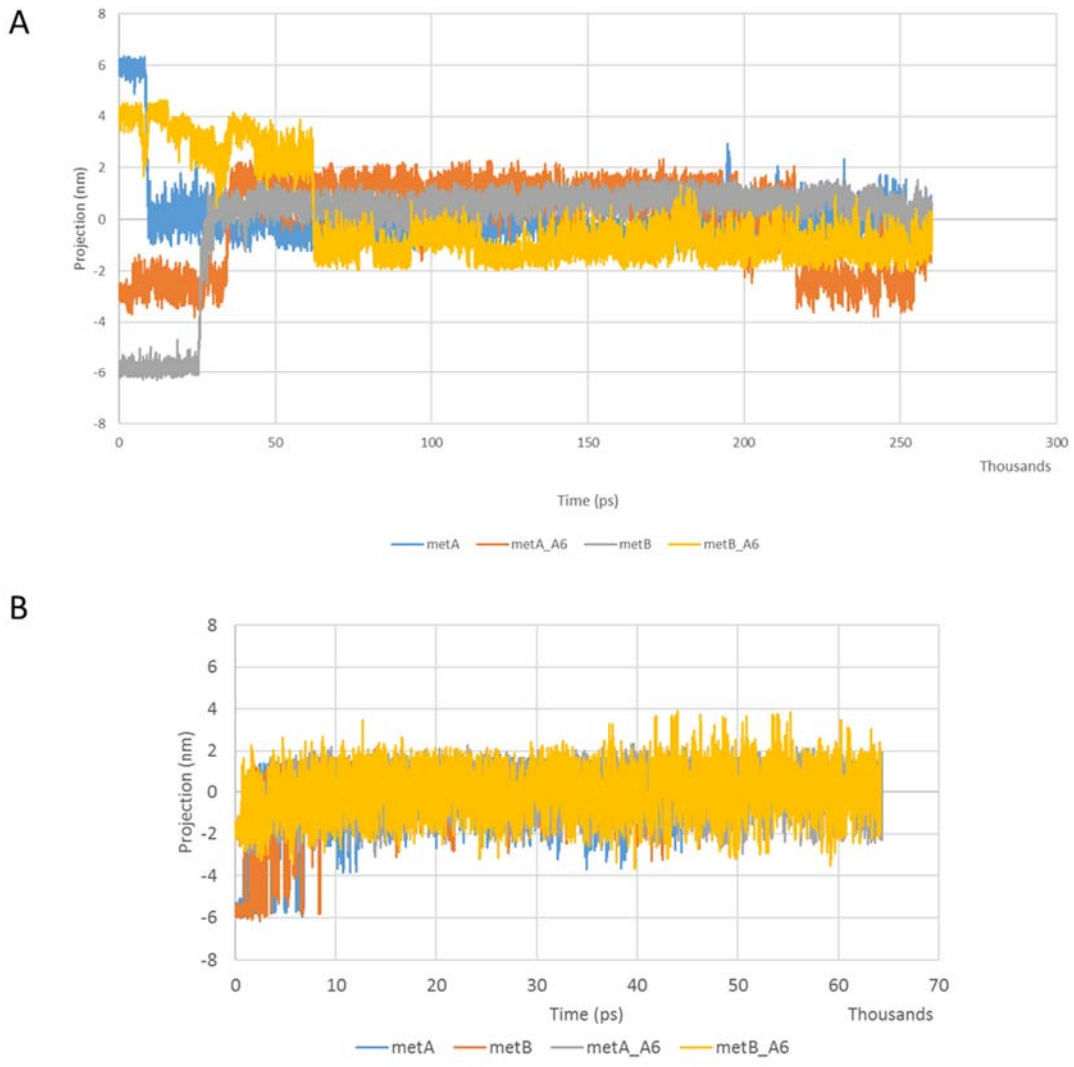


Figure 5.12: Plot of the Eigen vector 1 projection against time for each of the simulation systems. *A. Projection extracted from normal MD simulations of cyclosporine analogues. B. Projections extracted from lowest temperature replica trajectories of REMD simulations of cyclosporine analogues.*

As seen with the RMSD structural deviations during the normal MD simulations, prior to 50 ns, there was significant structural rearrangements of the analogues (Fig 5.12 A).

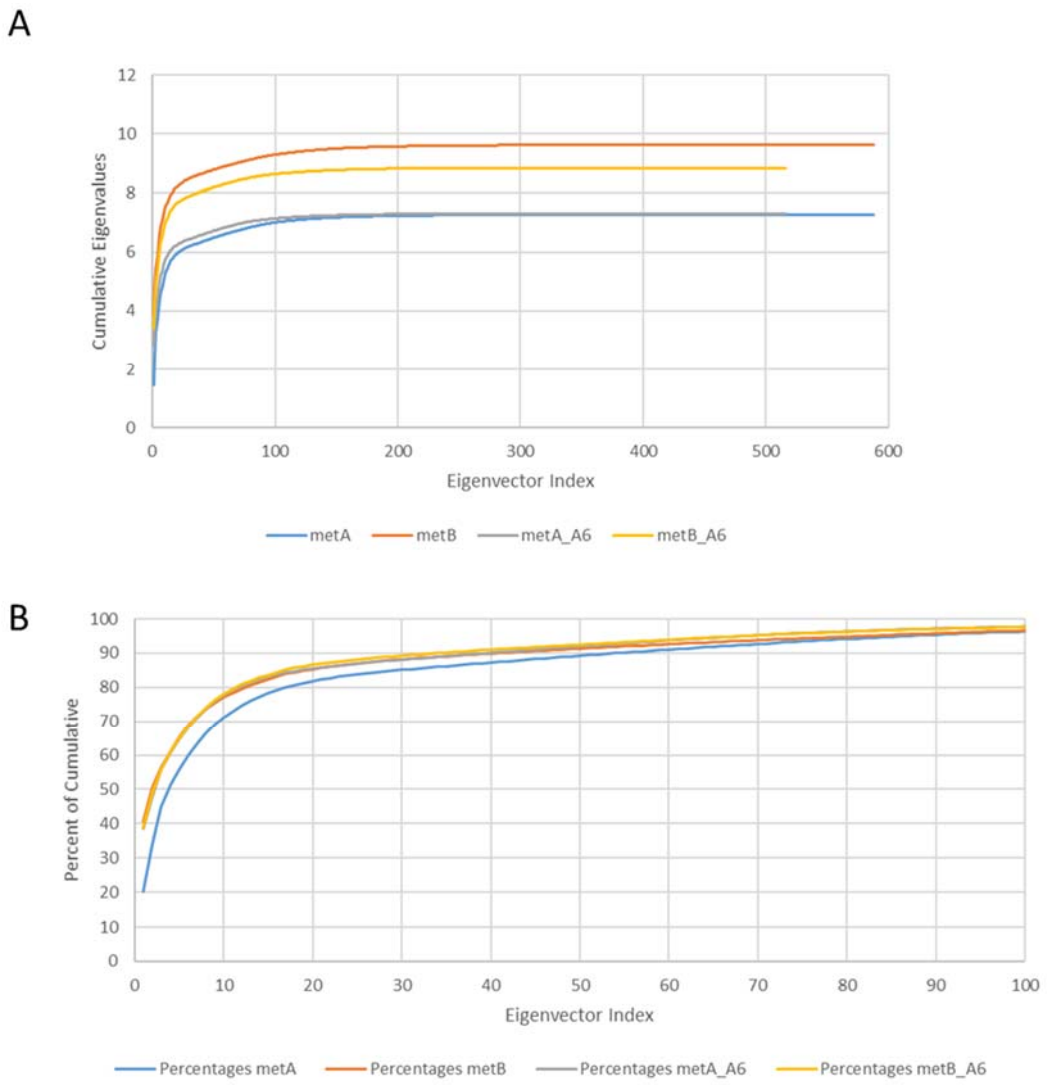


Figure 5.13: Eigen space of the normal MD simulations of the cyclosporine analogues. *A. Plot of the complete Eigen space of the matrix of the structure ensemble. B. Plot of the contribution of the top 100 Eigen vectors and their contribution to the Eigen space. MetA (blue), met B (orange), metA_A6 (grey), metB_A6 (yellow).*

In order to determine the diversity of the conformations present in the structure ensembles the dimensionality of the Eigenspace was analyzed. PCA analysis can reduce the complexity of data to make it easier to visualize. The data from 130 025 x 4 structures is presented in Fig 5.13.

Although the cumulated Eigenvalues for the metB analogues were higher than in metA (Fig. 5.13 A), the metA analogues required more Eigen vectors in order to account for the majority of the variance present in the structure matrix (Fig 5.13 B). From the normal MD simulation it was seen that there was greater structural diversity in the metA analogues than the metB analogues.

A plot of the relative inertia of the structures matrixes extracted from normal MD compared to the matrices of the REMD structure ensembles allowed for a comparison of the dimensionality of the enhanced sampling approach (Fig. 5.14).

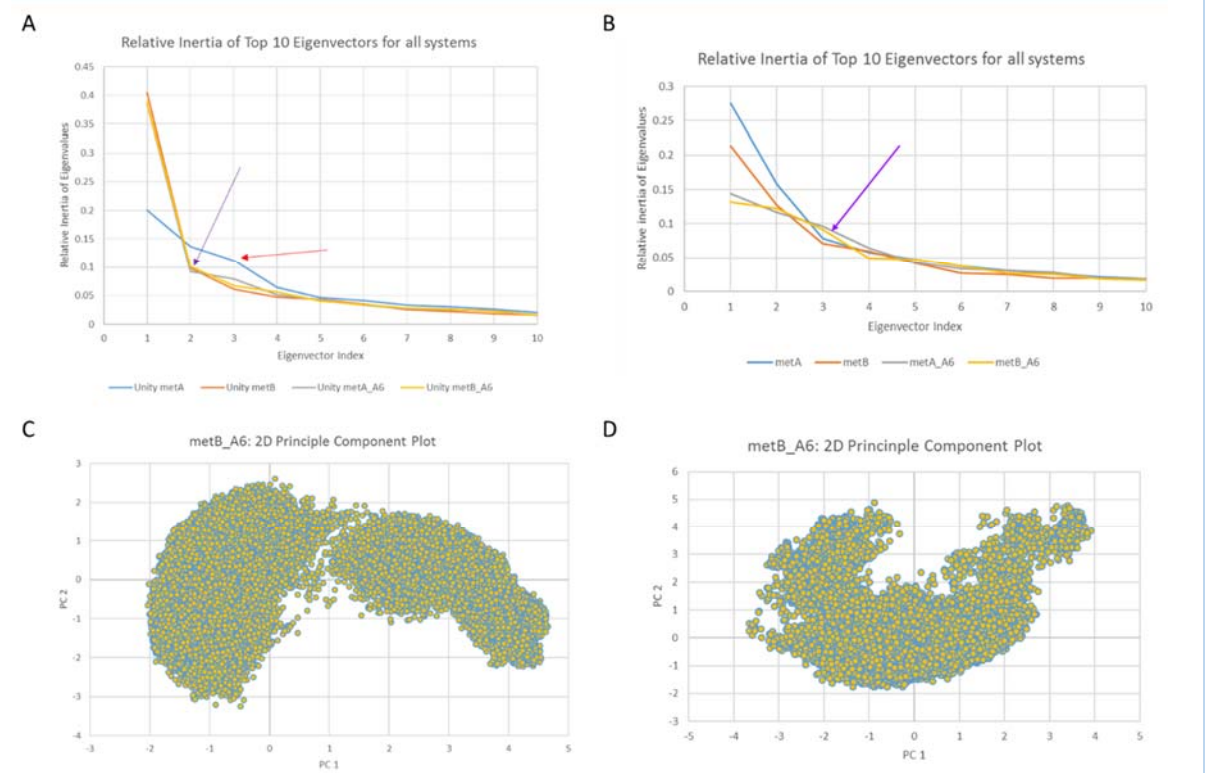


Figure 5.14: The dimensionality of the enhanced sampling REMD approach. *A. Scree plot of the relative inertia in the top 10 Eigen vectors in the structure ensemble of normal MD simulation for all cyclosporine analogues. B. Scree plot of the relative inertia in the top 10 Eigen vectors in the structure ensemble of replica 0 from the REMD simulation for all cyclosporine analogues. MetA (blue), met B (orange), metA_A6 (grey), metB_A6 (yellow) C. Two dimensional plot of the structures extracted from the metB_A6 analogue normal MD simulation. D. Two dimensional plot of the structures extracted from replica 0 of the metB_A6 analogue REMD simulation. (Eigen vector 1 (PC 1) and Eigen vector 2 (PC 2)).*

Enhanced sampling from REMD increased the dimensionality of the cyclosporine analogues by at least 1 Eigen vector in order to account for the majority of the variance (Fig. 5.14 B). It was observed that two dimensions were not sufficient to arrange the structures present in the replica 0 trajectory of analogue metB_A6 into distinct clusters. The structural diversity present in this ensemble required an additional dimension in order to resolve.

5.3.3.1.1 Extracting the Essential Dynamics through clustering

Experimental techniques such as deconvoluted Nuclear Magnetic Resonance are able to isolate conformations that will likely be present in solution by fitting approaches (Koivisto et al. 2010; Kolmer et al. 2015). In these kind of strategies back-calculation of NMR observables is used to match the experimental NMR pattern. The weights of individual conformers are then adjusted to match the NMR spectra and structures with low weights are less likely to contribute to a good fit to the observed NMR spectra and are thus excluded from the conformation ensemble. This conformation ensemble represents the essential dynamics of the query molecule and allows the practitioner the opportunity to isolate conformers that are present in solution from the conformational space.

In the absence of NMR observables for our peptide analogues, Essential Dynamics from enhanced sampling simulations were obtained by employing a clustering technique.

Clustering of the structures from the RMSD matrix of the low temperature trajectory using the nearest neighbor Jarvis-Patrick routine isolated representative structures (based on their RMSD neighbors). A structure was clustered with another structure if, within a collection of the 10 nearest structures, they had a certain number of similar neighbors. In order to determine this criteria we tested using generous (2), medium (3) and strict (4) criteria on only the metB cyclosporine analogue. The chemical space (defined by pharmacophores) probed

by the different clusters was also used to make a decision on which clustering technique was optimum for these cyclic peptides.

Table 5.2: Conformations extracted

	Generous (2)	Medium (3)	Strict (4)
Conformations	546	258	151
Pharmacophore Fingerprints	33,868	32,883	32,023

The strict method restricted the number of conformations to 151 conformers while the most generous criteria provided for 546 conformers (Table 5.2).

A

Anova: Single Factor — Solvent Accessibility Surface Area

SUMMARY

Groups	Count	Sum	Average	Variance
4_similarity	151	216502.5	1433.791	3121.555
3_similarity	258	370293.5	1435.246	3197.67
2_similarity	546	783073.5	1434.201	2934.064

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	262.0191	2	131.0096	**0.04317**	0.957751	**3.005179**
Within Groups	2889099	952	3034.768			
Total	2889361	954				

B

Anova: Single Factor — Radius Of Gyration

SUMMARY

Groups	Count	Sum	Average	Variance
4_similarity	151	1017.224	6.736581	0.044487
3_similarity	258	1739.802	6.743417	0.043482
2_similarity	546	3683.001	6.745422	0.042589

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.009247	2	0.004624	**0.107206**	0.898351	**3.005179**
Within Groups	41.05914	952	0.043129			
Total	41.06838	954				

C

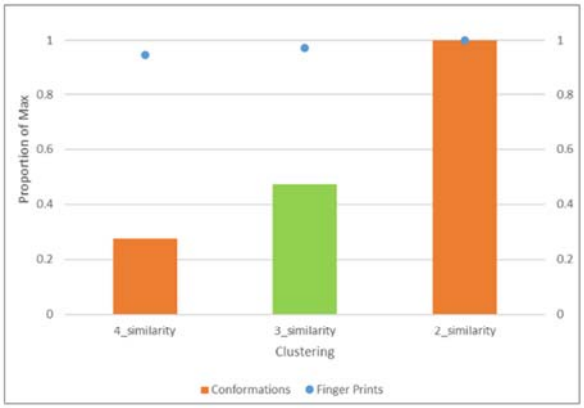


Figure 5.15: Statistical tests of the clustering techniques*. A. Analysis of Variance of the Solvent Accessibility. B. Analysis of variance of Radius of Gyration. C. Relationship between the number of conformations and the chemical space probed according to pharmacophore fingerprints.*

The pharmacophore fingerprints, SASA and Radius of gyration associated with each ensemble were computed using Discovery Studio. An Anova single factor test showed the absence of a statistically significant difference in the variances for the distributions of the SASA and Radius

of gyration in the conformation ensembles (Fig 5.15 A and B). It was thus concluded that adjusting the clustering parameters provided no improvement in the sampling of chemical and conformational space according to the SASA and Rg of the structure ensembles respectively. When the contributions of the parameters were measured against the pharmacophore fingerprints, a justification for the medium clustering technique was observed (Fig. 5.15 C). The strict clustering (4_similarity) technique recovered 94% of the pharmacophore diversity of the generous clustering (2_similarity) but only 27 % of the conformations. The medium clustering (3_similarity) recovered 47 % of the number of conformations and 97% of the pharmacophore diversity present (compared to using the generous approach). We were thus content with using the Jarvis-Clustering protocol that clustered structures that had at least three structures that were in common when a group of 10 neighboring structures was assessed. Careful choice of the clustering technique ensures that our chosen conformations have a balance between discrimination between structures and acceptance of similarity between structures.

5.3.3.2 REMD of cyclic-peptide scaffold

The REMD protocol demonstrated for the cyclosporine analogues (metA, metB, metA_A6 and metB_A6) was also applied for the 11mer cyclic peptide (that is used in later chapters to populate the conformation laden virtual library). A 24 replica temperature series (between 296.1 to 399.5 K) maintained an average exchange of 0.3 - 0.4 with an exchange interval test of 1 ps (Replex = 500) during the 65 ns of simulations. A plot of the distribution of the potential energies during the course of the simulation for each replica is outlined (Fig. 5.16). The distributions of potential energies maintained a bell-curve despite the mixing events that occurred between adjacent replicas. The degree of overlap of systems and the even distribution within each replica was acceptable for this simulation.

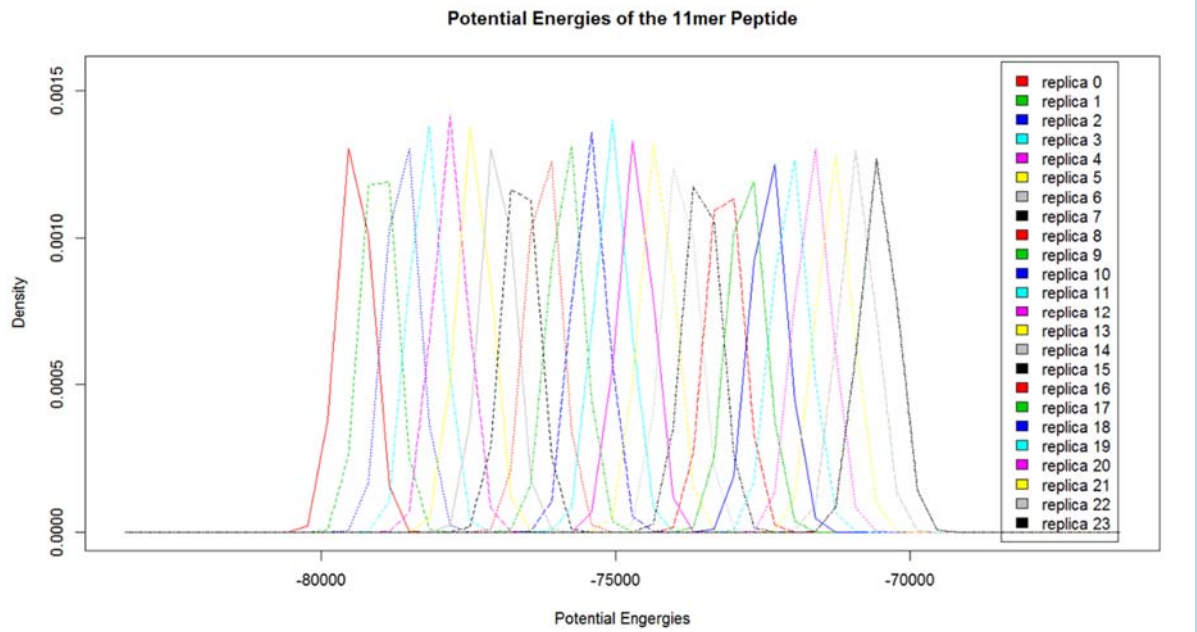


Figure 5.16: Potential energy distribution plot of the 24 replicas for the 11mer peptide REMD

5.3.4 Conformations from Normal and RE Molecular Dynamics for 11mer peptide

A PCA study of the conformations extracted from a normal MD simulation was used to highlight the enhanced sampling REMD protocol. The Eigenvectors and Eigenvalues were extracted from the covariance matrix (Fig 5.17).

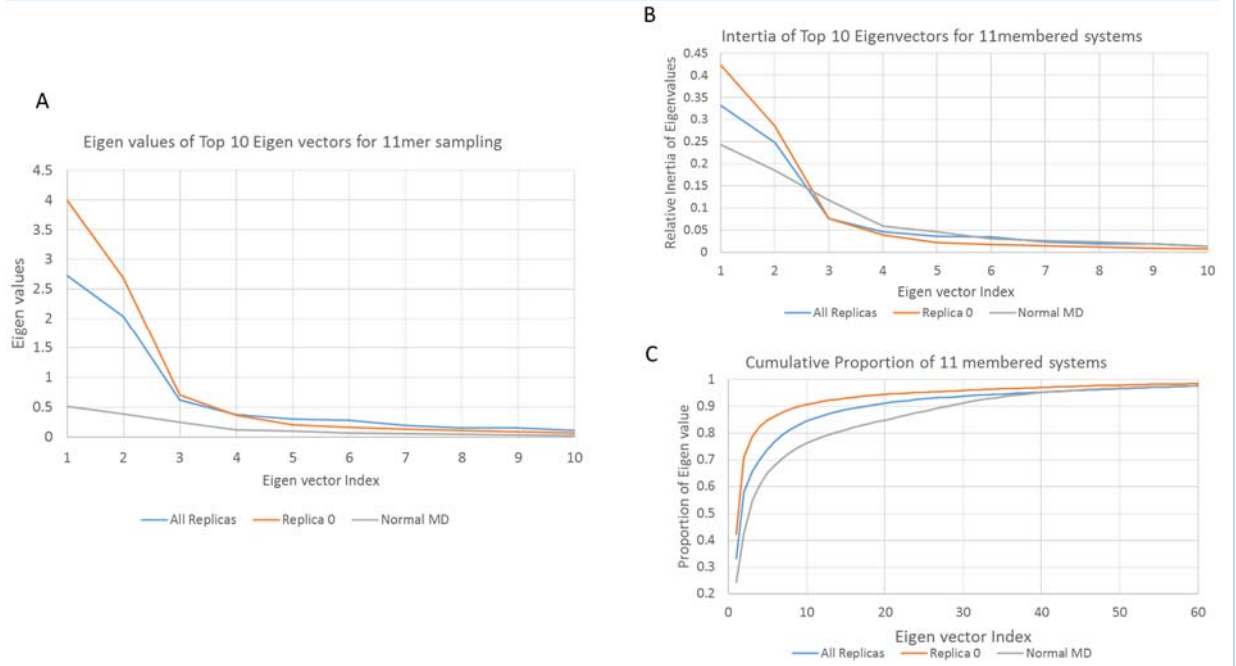


Figure 5.17: Interrogation of the Eigenvectors describing the sampling approach. *A. Eigen values over top 10 Eigenvectors. B. Relative Inertia of Eigenvalues. C. Proportion of total Eigenvalues. Structures obtained from All 24 Replicas (blue), structures obtained from Replica 0 (Orange) and structures obtained from normal*

Each of the data sets required 366 Eigenvectors to account for the diversity in their structures. The Eigenvalue of PC 1 for replica 0 was 3.99 while the normal MD had a PC 1 Eigenvalue of 0.5. This disparity highlights the magnitude of the variance and the degree of diversity within the set of structures present in the data sets. The dimensionality of the Scree plots on the inertia of Eigenvalues (Fig. 5.17 B) show that the "Replica-all" ensemble requires more Eigenvectors to account for the diversity and variance of the structures present within it. The proportion of the Eigenspace (total Eigenvalue) accounted for by the first 10 Eigen vectors in the structures obtained from the low temperature replica 0, was higher than in the normal MD simulation and the structures from all the Replicas (Fig. 5.17 C).

The magnitude of the diversity observed in the Replica 0 ensemble is larger than that observed in the normal MD ensemble because the majority of this diversity of the variance arises from backbone conformation transitions that contribute large degrees of variance. The diversity in the normal MD is influenced by side-chain fluctuations that have smaller contributions to variance. The smaller magnitudes present in the Eigenspace for normal MD highlights that these structures have smaller variations than those in the Replica ensembles (Replica 0 and Replica-all) (Fig. 5.17 A). Although we expect greater diversity and thus a higher Eigenvalue magnitude for PC1 in the Replica-all ensemble, the magnitude is offset by a smaller degree of similarity between the structures. Whereas in the normal MD fewer conformers are populated, the Replica-all ensemble has diverse conformations that are favored and populated evenly in the different replicas. This increased diversity results in a greater degree of dimensionality required to account for the majority of the variance (Fig. 5.17 C). The Replica 0 structures are all present in the Replica-all ensemble but their contribution to the diversity is decreased due to the inclusion of a normalizing step as part of the linear algebra routine that extracts the parameters of the Eigenspace.

The conformations present in the sampling ensembles were extracted using the clustering protocol. The Jarvis-Patrick clustering protocol examined 10 neighbors and chose structures that had at least 3 neighbors in common as belonging to the same cluster (Fig 5.18).

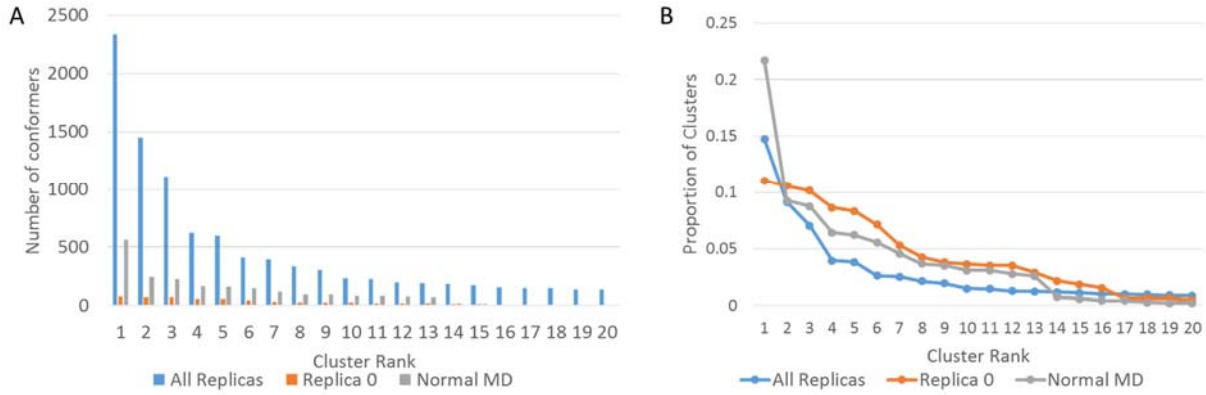


Figure 5.18: Clusters of structures extracted from 11 member peptide-sampling approaches. *A. The number of structures in each of the top 20 populated clusters. B. Proportion of the total number of conformers in the structure ensemble.*

The Replica-all ensemble had 15840 structures, the Replica 0 ensemble 660 structures while the normal MD ensemble had 2601 structures. From the cluster analysis, it was observed that despite the smaller number of conformers the Replica 0 structure ensemble was populated with a diverse set of conformers (Fig 5.18 A). The Replica 0 ensemble had the highest proportion of structures present in the top 5 most populated clusters for each of the sampling routines (Fig 5.18 B). A steep decline was observed between the most populated cluster, cluster rank 1, and the next most populated cluster, rank 2, for the normal MD ensemble. This steep decline shows that most of the structures present in the ensemble were similar to the top cluster. The shallowest decline is observed in the Replica 0 ensemble that benefits from a highly diverse population of structures due to the enhanced sampling protocol. The Replica-all ensemble had 2 486 clusters, the Replica 0 ensemble had 76 clusters while the normal MD ensemble had 355 clusters. The structures that were in the center of each of the clusters

were recorded as representing the “Essential dynamics” of the simulations. The normal MD structure ensemble shows a collection of conformers that are very similar. (Fig 5.19).

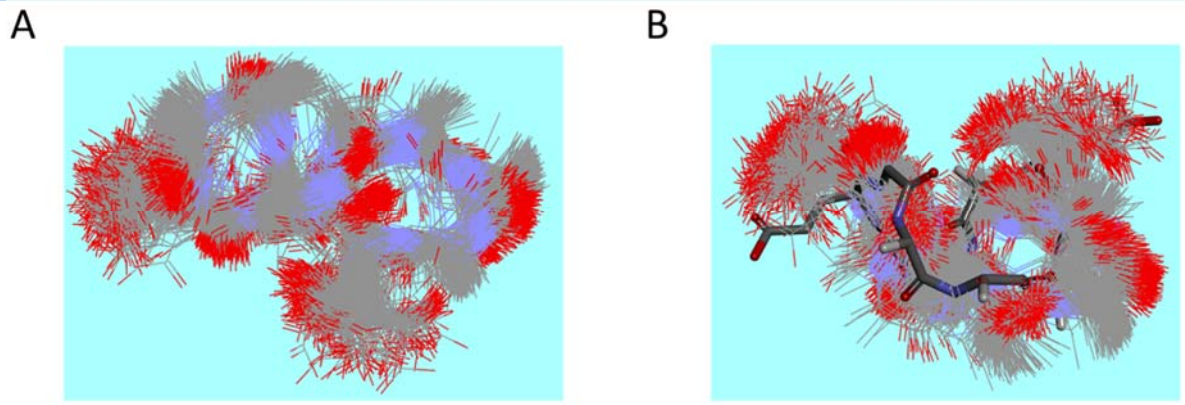


Figure 5.19: Essential Dynamics of the normal MD simulation. *A. Central structures of the 355 clusters. B. The central structures of the 355 clusters overlaid with the most dominant/populated cluster. Oxygen (red), Nitrogen (purple), Carbon (grey). Only hydrogens of the dominant cluster are shown (light grey).*

The decreased degree of diversity of the normal MD structure ensemble is seen through the porosity of the image of overlayed structures. The majority of the structures overlap over the same region of space showing this lack of conformational diversity in the simulation. Despite the lower number of conformations in its ensemble, the Replica 0 ensemble shows greater structural diversity in its essential dynamics (Fig 5.20).

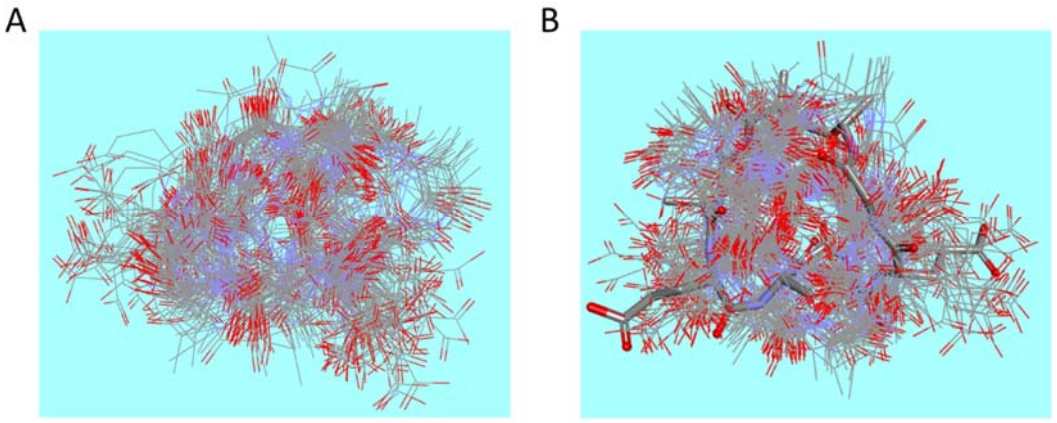


Figure 5.20: Essential Dynamics of the Replica 0 simulation. *A. Central structures of the 76 clusters. B. The central structures overlayed with the most dominant/populated cluster. Oxygen (red), Nitrogen (purple), Carbon (grey).*

In order to visualize the Essential dynamics of the Replica-all structure ensemble the clusters that had at least 4 structures were recorded after the Jarvis-Patrick clustering. This contraction gave 203 clusters and structures instead of the 2,486 of the entire structure ensemble (Fig 5.21).

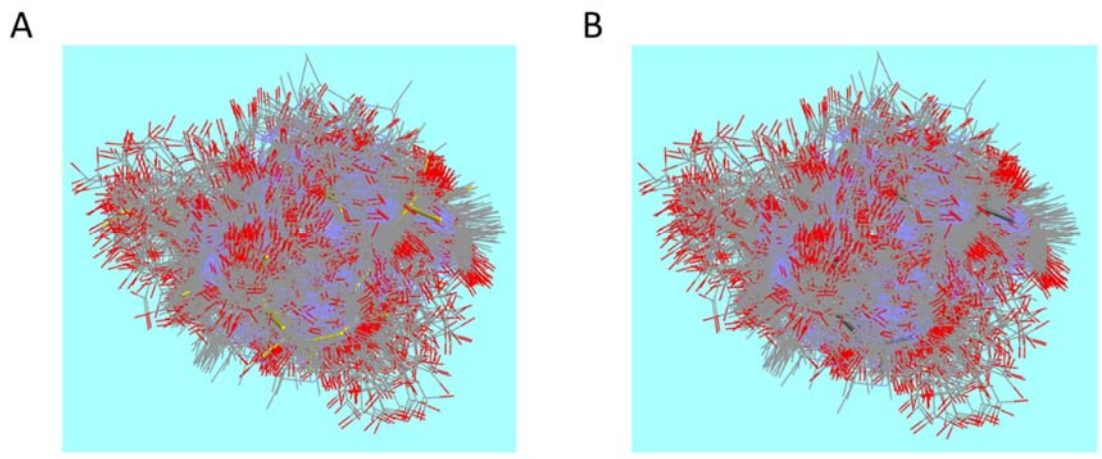


Figure 5.21: The top 203 clusters of the Replica-all structure ensemble. *A. Central structures of the 203 clusters overlayed with the most dominant/populated cluster (yellow). B. The central structures overlayed with the most dominant/populated cluster. Oxygen (red), Nitrogen (purple), Carbon (grey).*

Despite the cluster contraction, to resolve the Essential dynamics, the Replica-all simulation had a diverse set of conformers (Fig. 5.21). Although the contracted set has less conformers (203 structures), the Replica-all ensemble had less porosity than the normal MD simulation ensemble (366 structures) (Fig 5.19). The REMD enhanced sampling protocol is likely to access conformations that were not present in the normal MD simulation, and this is confirmed through these results.

5.3.5 Preparation for scaffold decoration for the 11mer peptide

The conformations of the 11mer peptide that are likely to be accessible to the Cyclophilin-D binding site are conformations that are probable in solution. The Replica 0 structures that persisted during the 296.1 K simulation are taken to represent the ensemble of likely solution accessible conformations. The Eigenvalues of the center structures from the 12 most

populated clusters were superimposed over the Eigenspace of the Replica 0 ensemble (Fig. 5.22). These structures can be seen to span the entirety of the Eigenspace of the Replica 0 simulations. These structures were recorded as the structures that will serve as the feedstock for the population of a conformation laden virtual library using the DerivatizeME enumeration protocol, in Chapter 6.

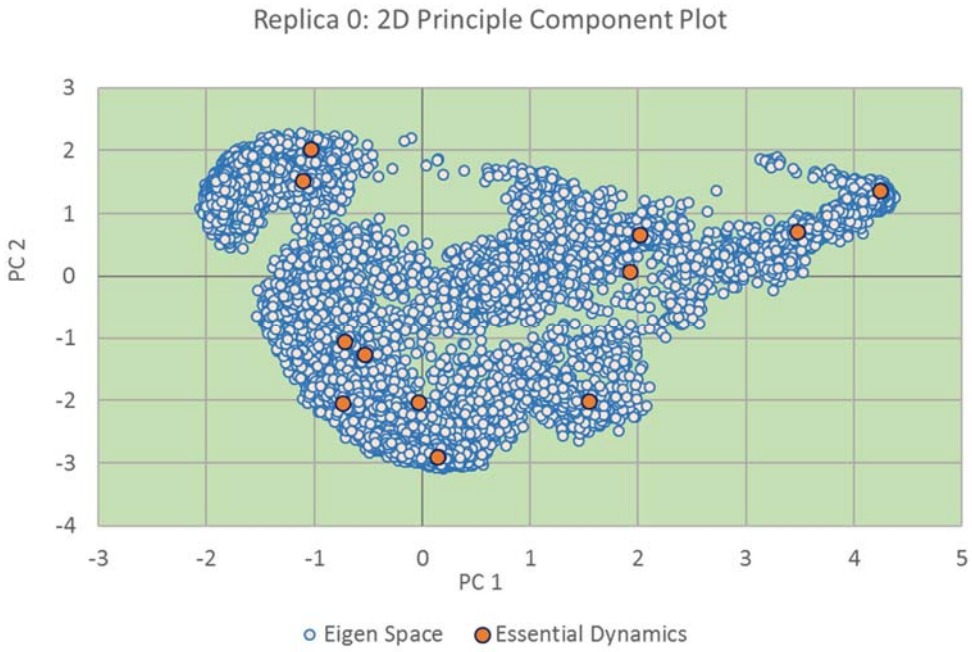


Figure 5.22: 2D Principle Component Plot of the Replica 0 Eigen space. *Eigen space structures (blue). Essential dynamics structures (orange).*

By inferring Boltzmann statistics, we would assume that the most populated conformations of the 11mer cyclic peptide in solution would be the lowest energy conformations. To test this scenario the difference in the potential energy of the systems was calculated at the DFT level in vacuum (M06-2X/6-31G ++ (gas phase)) (Fig. 5.23).

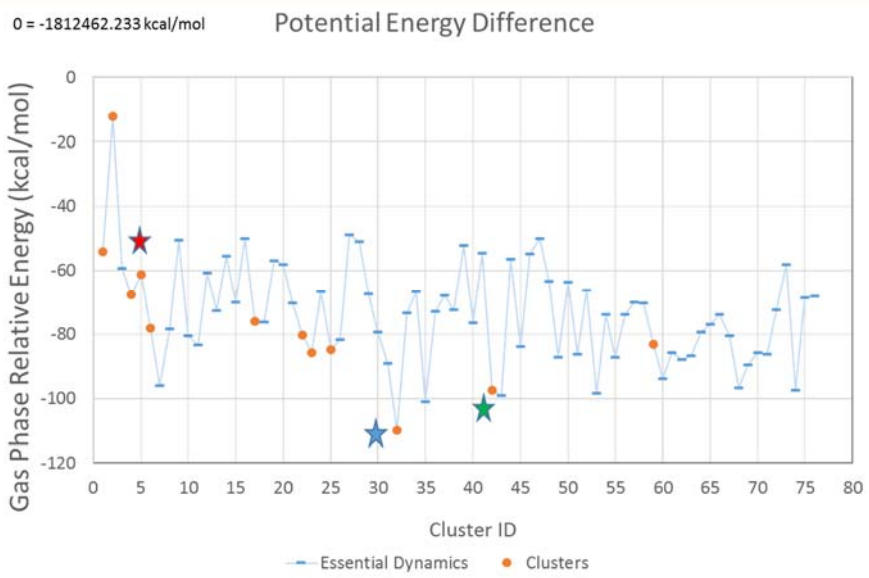


Figure 5.23: The Relative Gas Phase potential energy difference of the conformations. *The essential dynamics conformers (blue). The Scaffold conformers (orange).*

The structure with the lowest gas phase energy (cluster 32, blue star) present in the essential dynamics of the Replica 0 simulation was the first scaffold now chosen for the population of the virtual library using the cluster ranking approach (Fig. 5.23). This cluster 32 was the second most populated cluster (70 individuals) while cluster 4 (red star) had the highest number of structures present within its cluster (73) and cluster 42 was the third (green star) (67 individuals). The potential energy plot of structures extracted from the essential dynamics of the Replica 0 simulation show that there are some central structures reserved as scaffolds that have a relatively high energy (cluster 1, cluster 2 and cluster 4). These high-energy structures were made accessible through the application of the enhanced sampling protocol that populated conformations that would otherwise be unfavoured apart from the enhanced sampling protocol that we adopted. In solution at temperatures well above absolute zero, we expect higher energy conformations to become more populated.

5.3.6 MacroModel Conformer search

The best 10 structures obtained from the MacroModel Monte Carlo multiple minimum search (torsional sampling) routine, low frequency molecular dynamics (the Low Mode sampling) approach, a hybrid of the torsional sampling and Low Mode algorithm and a hybrid Largescale sampling with Low Mode technique were recorded (Fig. 5.24).

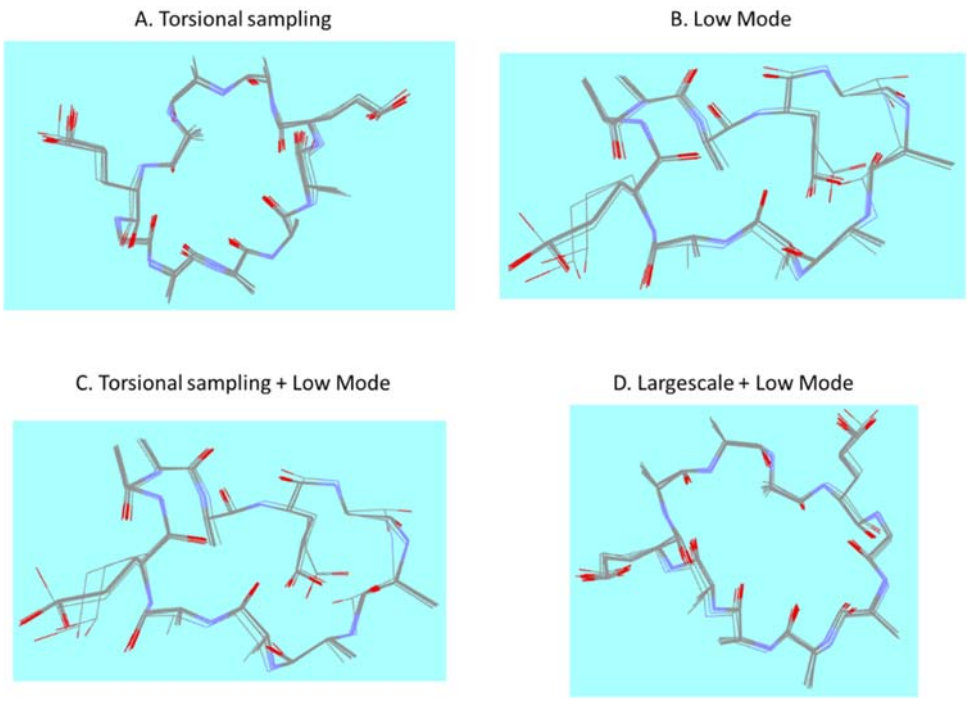


Figure 5.24: Top 10 conformers obtained from alternative macrocycle sampling protocols. *A. Torsional Sampling alone, B. Low Mode sampling alone, C. Torsional sampling combined with Low Mode sampling, D. Largescale sampling combined with Low Mode sampling.*

The best 10 conformers obtained by the MacroModel conformation search routines were not as diverse as the best conformations obtained from the REMD enhanced sampling routine. The best 10 conformers obtained by the MacroModel conformation search routines overlap significantly whereas those obtained from REMD are significantly disperse. The MacroModel search routines are optimized to extract the lowest energy conformers (global minimum) whereas the REMD conformations are not biased towards the lowest energy and are allowed to access seemingly unfavourable conformational space. Populating a virtual library using

scaffolds that have diverse conformations enables virtual screening algorithms to access large regions of conformational space during their affinity searching.

5.4 Conclusion

When experimental Nuclear Magnetic Resonance data is available, deconvolution of the NMR observables through least squares fitting of back-calculated NMR signals from conformation ensembles to the experimental data is able to isolate plausible solution available conformer states (Koivisto et al. 2010; Blundell et al. 2013). In this study, through the incorporation of conformational ensembles derived from replica exchange molecular dynamics, a nearest neighbour clustering strategy was used to extract plausible solution conformers for each of 5 peptides in the absence of experimental observables. Replica exchange molecular dynamics simulations of cyclic peptides was performed across a temperature series that maintained a satisfactory exchange frequency on similar sized systems. The diverse conformation milieu was assessed using a multivariate PCA analysis of the covariance of the RMSD of the structures traversed along the simulation trajectory. The clustering approach employed to identify the essential dynamics of the low temperature Replica 0 simulation used sampling of the Eigenspace of the structure ensemble. The scaffolds identified using this procedure were used (Chapter 6) to populate a virtual library (after side-chain modification) with backbone conformers providing for a virtual library that probes both chemical space and conformational space of the cyclic peptides extensively. In the absence of the protein target conformational content during the selection of these scaffolds, we are unable to make assertions that lead towards any deductions of the bioactive conformations, but docking validation and screening will be able to provide this information. The advantage of a diverse conformational landscape is that it enhances the ability of the near-bioactive conformation being accessed during the virtual screening stages.

Chapter 6: The population of a conformation-laden virtual library of cyclic peptides within a confined sequence space

6.1 Introduction

The large and diverse "linguistically" accessible search space available to peptides requires innovative approaches in order to inform rational molecular design strategies; particularly those that rely on the *in vitro* screening of peptide based focused libraries. In this chapter, we outline the deployment of an enumeration strategy for the purposes of populating a virtual library of macrocyclic peptides that includes both chemical and conformational information. Virtual screening of focused libraries in the search for novel peptide derived cardio-protective agents will benefit from the precision afforded by the conformational diversity of such a virtual library. The in-house enumerator, DerivatizeME, is used separately to a conformational search protocol during the population of this virtual library. The feedstock for the DerivatizeME enumeration were peptide scaffold conformations that captured the essential dynamics of an 11 membered cyclic peptide. The target enumerated cyclic peptides were chosen judiciously with head-to-tail cyclized solid-phase peptide synthesis in mind. Low temperature Essential Dynamics extracted from solute tempered REMD simulations biased the conformational diversity towards bioactive conformational space.

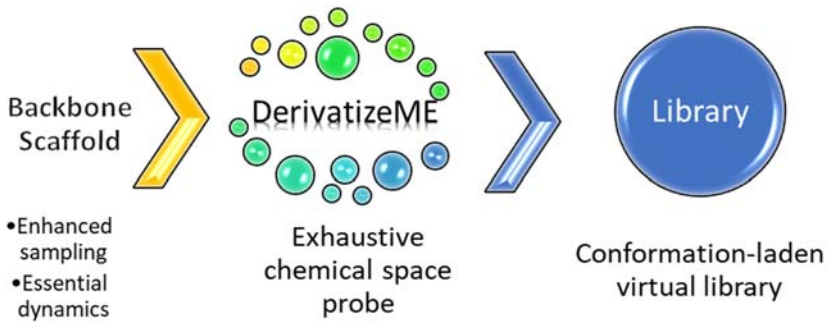


Figure 6.1: The strategy employed in the population of the peptide virtual library

6.2 Objectives

The objectives of this section were to populate a conformation laden virtual library within a confined sequence space. The essential dynamics of the 11 membered cyclic peptide represented by the 12 REMD accessed scaffolds were used as scaffolds for the population of the virtual library (Fig. 6.2). Each scaffold was limited to six substitution points in order to limit the scope of the library.

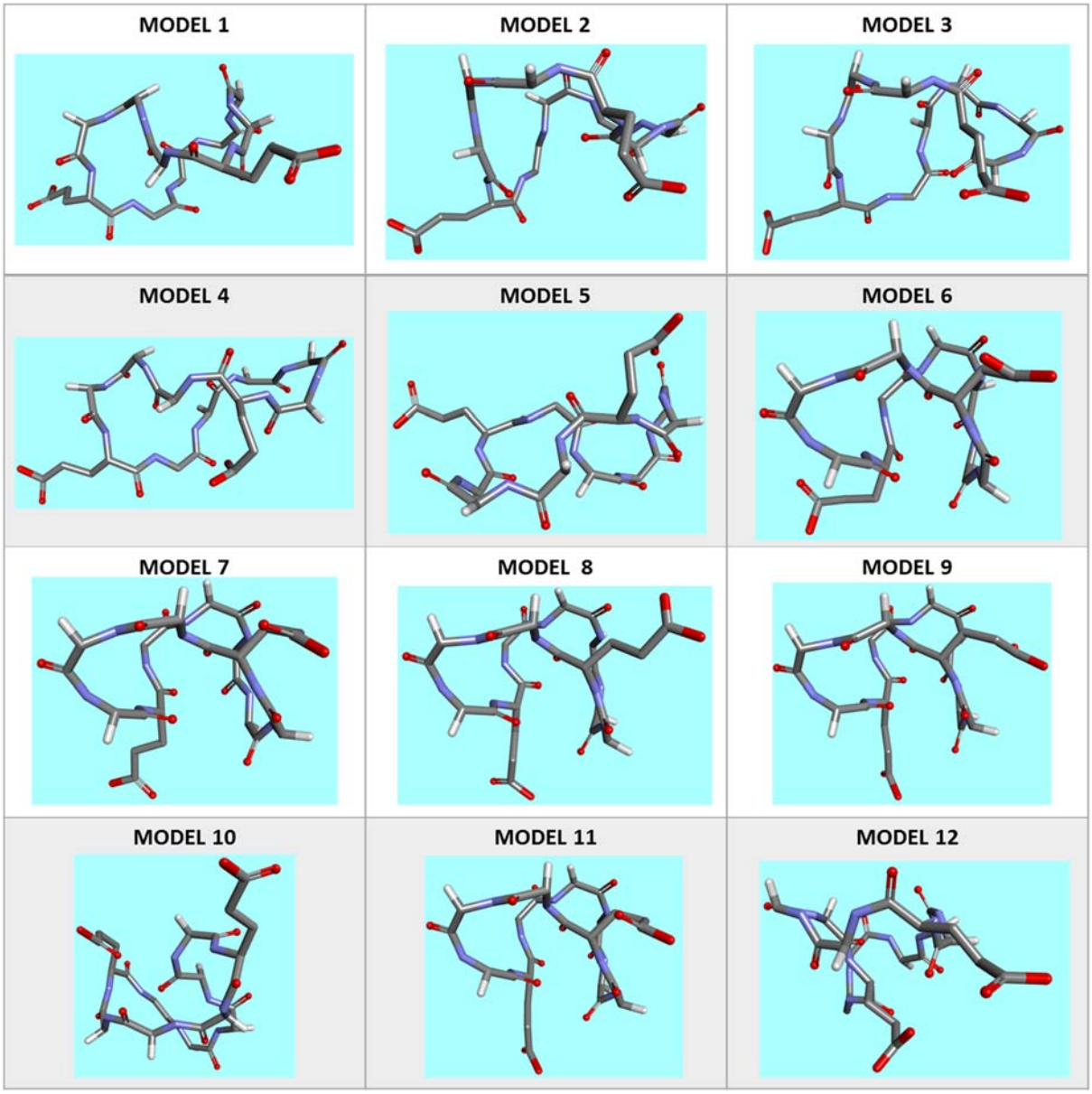


Figure 6.2: Cyclic peptide scaffolds from the Essential Dynamics of REMD enhanced sampling.

6.3 Methodology

6.3.1 Cluster architecture

In order ensure that the virtual library could be populated within a reasonable time we made use of a small computer cluster that aggregated compute nodes for the completion of the task. The local cluster, illustrated in Fig 6.3, made use of the Portable Batch System (PBS) to schedule tasks as jobs.

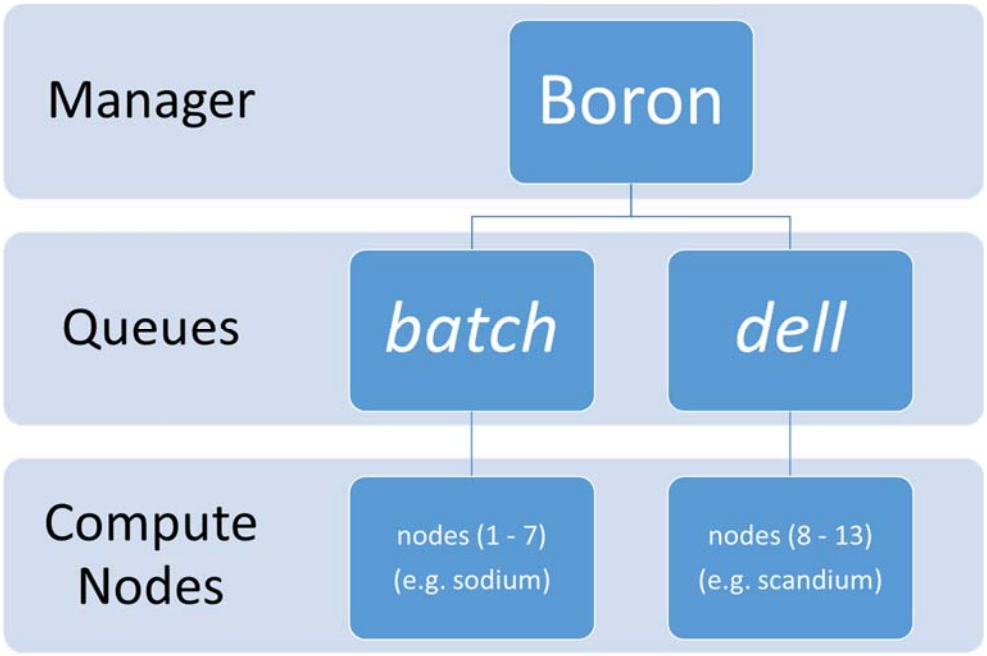


Figure 6.3: The cluster system architecture, together with host names (boron, sodium etc.), used for the population of the conformation laden virtual library.

Each compute node recruited 8 dynamically mapped CPUs, to which tasks were distributed according to the workload assigned to each node. Within the execution of our population-generating pipeline, tasks were allocated to either of the queues ('batch' or 'dell') with the knowledge of different efficiencies of computation on each queue. The distribution of jobs ensured that all nodes were active throughout the time that the population routine was in execution. The stages of enumeration were monitored individually, allowing for the

management of the enumeration and also ensuring that interrupted tasks could be easily retraced and repeated if needed.

6.3.2 Population routine

A single non-trivial python script (*populate_cl_vl.py*) automated the control of three processes: task division, preparation of customized job scripts and the submission of those jobs to the queue (Fig. 6.4). The script contained four functions *prepare_labels*, *derivatizeME,* *namewriter* and *readlist*:

prepare_labels: reads a list of labels from a file (e.g. 27000) and writes label_set files each containing a subset of the full list. An example of one of these labels is "trp_arg_gly_cys_his_leu". The label_set subsets are for distribution to nodes on the cluster.

derivatizeME: prepares job scripts for individual tasks. Each **task** is a decoration of a scaffold using the derivatizeME enumerator. The enumerator's inputs are read from the label_set files using the *readlist* function and this provides direct information about the amino acid substituents; together with the scaffold, there is a complete description of the enumeration task, which is written using the *namewriter* function. The *namewriter* function arranges the inputs for the enumeration task in an explicit array format for simplicity.

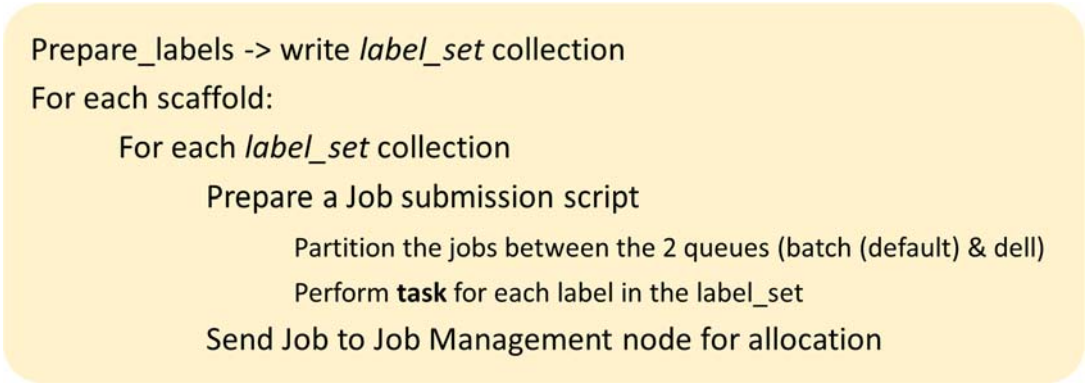


Figure 6.4: Pseudo code of the population script.

The **task** subroutine (Fig. 6.4) summons the enumerator for each label_set and compresses the file containing bare derivatives output from the enumerator. The **task** also controls the recruitment of a protonation script. This protonation script reads structures and follows a 3-step OpenBabel conversion cascade in order to protonate the derivatives. The proportion of derivatives from the derivative collection that are protonated can be adjusted from within the protonation script (to allow subsampling of the library). Due to the diverse protonation routines available, we separated the protonation steps from the enumeration protocol. In order to facilitate curation, bare derivatives (excluding hydrogen atoms) from the enumerator were stored in multistructure '.xyz' format. The protonation script takes this complete multistructure '.xyz' file as an input into its protonation routine (Fig. 6.5).

```
Read the derivative collection
Sample through the derivative collection
        Prepare single structure files of sampled derivatives
Loop through all the sampled structures
        convert '.xyz' to '.pdb'
        protonate '.pdb' structure, convert to '.sdf' and store within an open multistructure '.sdf'
Compress protonated multistructure 'sdf' file
Delete all files prepared in the sequence
```

Figure 6.5: Pseudo code of the protonation script recruited during the population script.

The sampling loop extracts a single derivative in '.xyz' format from the multistructure derivative collection and writes it to a single '.xyz' file in preparation for the three-step Open Babel conversion sequence. For sampling purposes, the sampling loop used a random number generator to extract an average of 10 derivatives for every 10,000 searches. The first stage of

the OpenBabel sequence is the extraction of a particular structure from the enumerated collection as an '.xyz' structure, followed by its conversion to '.pdb' format. The final step of the cascade is the simultaneous protonation and conversion of the '.pdb' structure file to a protonated '.sdf' structure. Protonation was performed at the default pH of 7.4. The protonated derivative was then included in a multi-structure ".sdf" file for the sampled derivative set. All intermediate files were deleted after the compression of the multi-structure ".sdf" file.

6.3.3 Property Space

The protonated derivatives were used as a sample set in order to interrogate the impact of decoration on the chemical space of the decoration scaffold. The chemical property space of each conformation collection (an enumeration on a particular scaffold conformation) was compared to the property space of the undecorated scaffold (11mer_Asp), cyclosporine A (CsA), the approved compound dataset from Drugbank (Drugbank) and a representative sample of compounds extracted from the ZINC database (Zinc). The datasets used for this study are summarized (Table 6.1).

Table 6. 1: Datasets of compounds used to visualize the chemical space.

Dataset	11mer Asp	CsA	Drugbank	Models (1 - 12)	Zinc
Compounds	1	1	1,742	1,520,001	3,083

The chemical property space probed by the data sets was identified by the physico-chemical properties (number of hydrogen bond acceptors, number of hydrogen bond donors, molecular weight, solubility as a partition coefficient and the polar surface area) of each of the compounds in the dataset. The physico-chemical properties were computed from the RDKit descriptor calculator made available by the KNIME analytics and reporting platform. Dataset boxplot statistics were computed using Conditional Box Plot KNIME nodes and plots

were prepared from the resultant statistics for analysis. The KNIME workflow used to read structure files, compute compound descriptors and plot the boxplots is illustrated (Fig. 6.6). In this figure, the different properties are treated with different Conditional Box Plot nodes, and then separate ".csv" files are written.

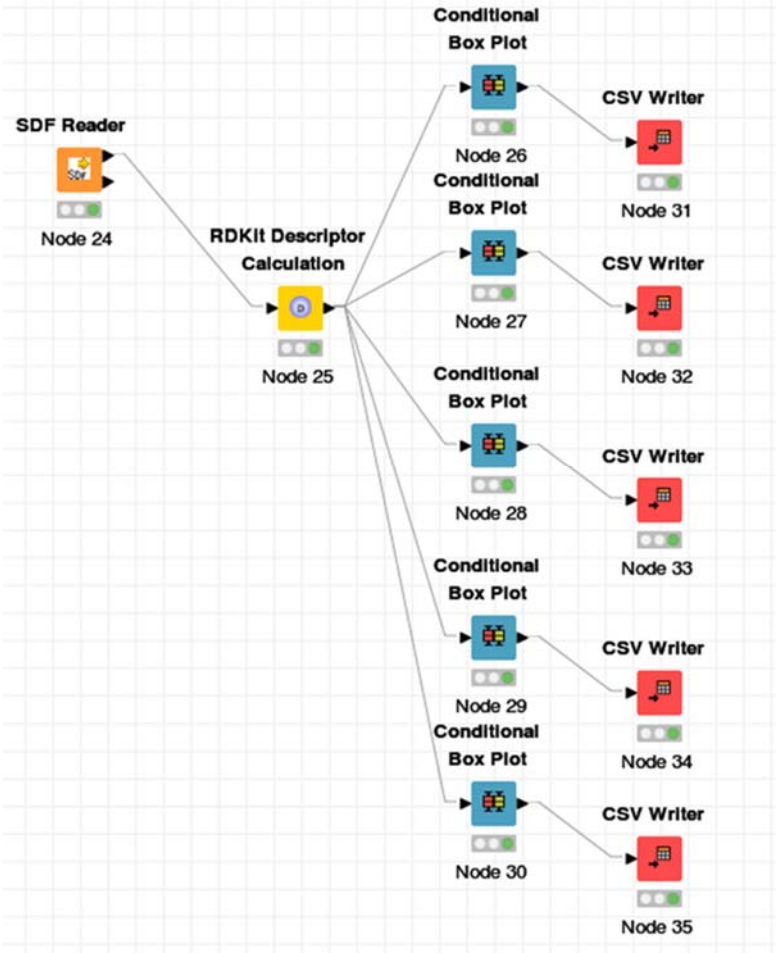


Figure 6.6: KNIME workflow to prepare box-plot statistics of datasets used in this study.

6.3.4 Synthetic feasibility determination

The protonated derivatives were used to determine the proportion of compounds within the virtual library that were likely to be synthesizable. Machine Learning approaches based on chemical and structural descriptors were used to define parameters that could be used for the prediction of peptide synthesizability of the protonated derivatives dataset. In the

absence of reliable experimental data associated with synthetic feasibility of this class of cyclic peptides, the CycloPS synthetic feasibility filter was used to prepare a set of training data for the use of machine learning model. Such a trained model is attractive due to its ability to deal efficiently with evaluation of large quantities of data. Scripts were used to prepare 120,000 sequences of 11-membered peptides within the sequence space of our virtual library. We used the CycloPS triage functionality to identify sequences that satisfied its synthetic feasibility restrictions (i.e. 5 or more amino acids; no more than 2 consecutive Proline residues; no forbidden amino acid pairs; Log P > 0; and 1 charged residue every 5 residues) (Duffy et al. 2011). Python scripts were used to match the CycloPS data to structures, including a labelling with respect to the CycloPS synthetic feasibility filter (*rule_applyer.py*).

KNIME workflows were used to perform the machine learning-based synthetic feasibility experiments. The data preparation, model training, prediction steps and hyper parameter tuning steps are outlined (Fig. 6.7). Low energy structures for training and prediction were produced by the LigPrep tool made available by Schrödinger nodes. Within LigPrep the OPLS 2005 force field was used and ionization was determined at a pH of 7 (ionization states within a range of +/- 2 were included). Features for the peptide structures were obtained by extracting chemical descriptors from the optimized and refined structures. Relevant chemical, structural and pharmaceutical descriptors of peptides were obtained with the use of a QikProp node (Jorgensen & Duffy 2000; Jorgensen & Duffy 2002). Non-zero and non-degenerate quantitative features that enhanced the relevance of the models produced were manually selected (Column filter) (Fig. 6.7A, uppermost workflow). These features were used to train a support vector machine (SVM) for synthetic feasibility prediction (Fig. 6.7A, lower workflow).

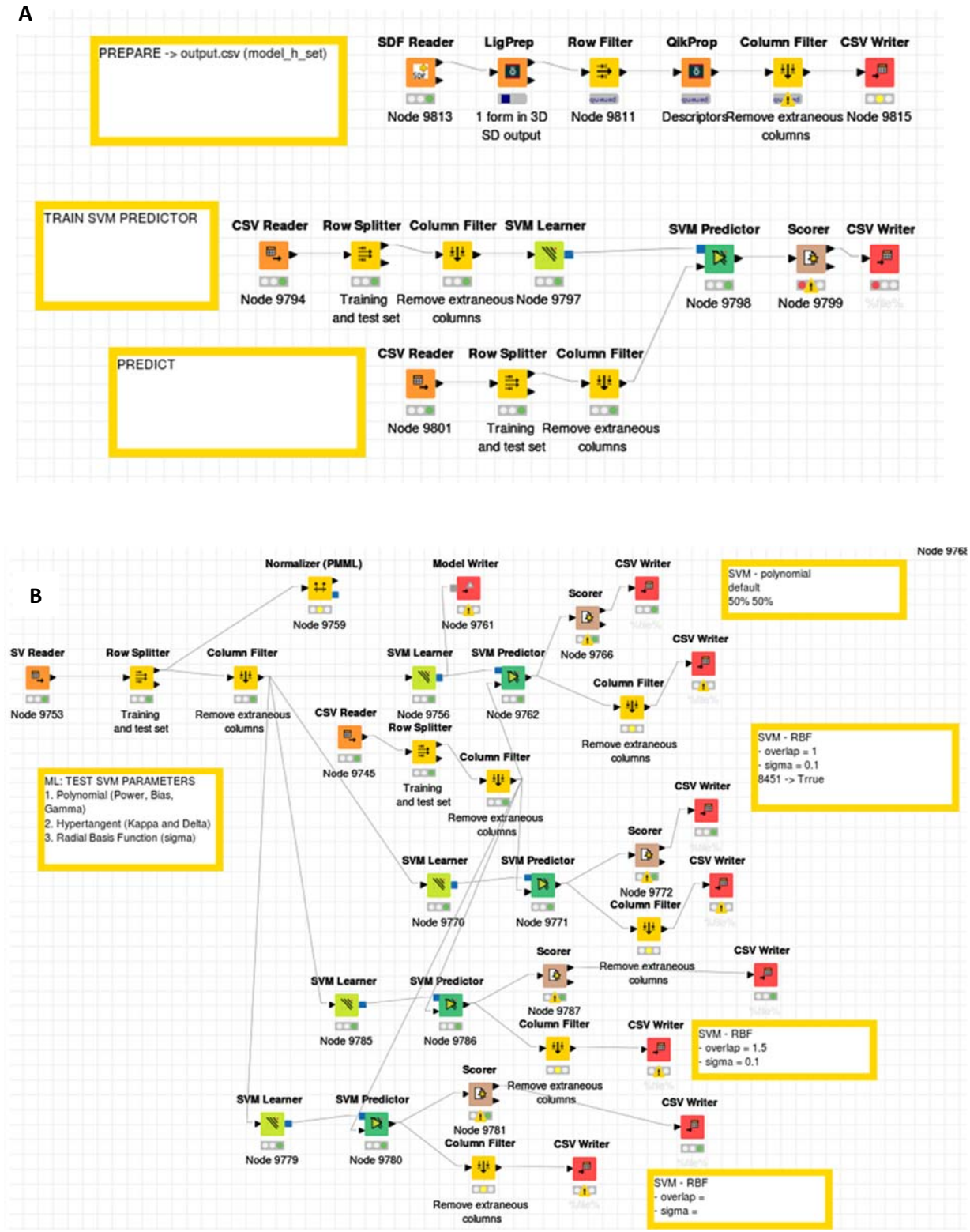


Figure 6.7: KNIME workflow to perform the Machine Learning steps to determine synthetic feasibility. *A. Scheme followed to prepare data, train support vector machine and predict outcome. B. Scheme followed to test SVM parameters.*

A series of SVM parameters were used to train 20,000 of the labelled sequences with the features embedded. Parameters that modified the HyperTangent (HT), Polynomial (POL) and Radial basis function (RBF) kernels of the SVM learner were used for hyper parameter optimization (Table 6.2).

Table 6.2: Support Vector Machine kernels and parameters tested during hyper parameter optimization.

Kernels	Parameters	Values Tested
Hyper Tangent (HT)	Overlap Penalty	0.5; 1.0; 1.5
	Kappa	0.1
	Delta	0.5
Polynomial (POL)	Overlap Penalty	0.5; 1.0; 1.5
	Power	0.1; 0.5; 1.0
	Bias	0.1; 0.5; 1.0; 5.0
	Gamma	0.1; 0.5; 1.0; 5.0
Radial Basis Function (RBF)	Overlap Penalty	0.5; 1.0; 1.5
	Sigma	0.1; 0.5; 1.0

The set of SV kernels and parameters that best reproduced the CycloPS classification of the remaining sequences (100,000 of the 120,000 sequences) were further used to predict the synthetic feasibility of the 1.5 million protonated derivatives.

6.3.5 Virtual Screening Test Case

A subset of the conformationally laden library was tested through molecular docking experiments. Screening, using molecular docking, was of a subset of 500 derivatives of

11mer_Asp (obtained by decoration of the 11mer_Asp CycloPS scaffold) and also a subset of our virtual library (500 x 12 obtained from derivatizing of the 12 conformations of 11mer_Asp). The CycloPS scaffold represents a single conformer test case that would be made available by existing enumerators while the 12 Model set represents biologically relevant conformers made available by our enhanced sampling approach. For this test, the screening problem was restricted to a set of 500 derivatives from each conformer set of the "trp_arg_gly_cys_his_leu" collection. A single dehydrated cyclophilin-D protein target was used to prepare a grid file for the glide docking experiments using the following parameters: the grid center was on the original crystal structure ligand; Hbond donors were allowed for aromatic hydrogens; the grid file prepared, was to allow for peptide macrocycle docking; and the OPLS3 force field atom type parameters were used for all calculations.

The cyclophilin-D target was prepared using Maestro Schrödinger tools with the following parameters: structure mistakes were automatically fixed; pretreat metals, if present; assign bond orders automatically; add hydrogens; convert selenomethionines if present; generate all possible ionization states of non-protein groups (protonate at pH of 7.0 with a pH range of +/- 2.0 using Epik); optimize protonation states at pH 7.0 using PROPKA; and perform restrained minimization of all atoms using the OPLS 2005 force field with an RMSD cutoff of 0.3 Å.

The derivatives were prepared using LigPrep from their protonated DerivatizeME '.sdf's with the following parameters (although not all apply to these systems): retain specified chirality's; desalt if present; generate possible tautomers; generate all possible ionization states of non-protein groups (protonate at pH of 7.0 with a pH range of +/- 2.0 using Ionizer); and apply atom types from the OPLS 2005 force field.

A High-throughput virtual screening Glide experiment was performed on each of the prepared ligands with the following parameters: docking (flexible ligand sampling; nitrogen inversions allowed during sampling; sampling of ring conformations permitted; and penalties for non-planar amide conformations included with Epik state penalties in docking score), ligand treatment (scale van der Waals radii by 0.80 with a partial charge cut off of 0.15; and docking restrictions of 300 atoms and 50 rotatable bonds), and output options (post-dock minimization of 5 poses per ligand; reject poses with Coulomb-vdW energy greater than 0.0 kcal/mol; conformationally distinct poses are identified from an RMSD of less than 0.5 Å and a maximum atomic displacement of 1.3 Å; and strain correction thresholds ignored).

Cyclosporine A was also docked on the cyclophilin-D protein under similar conditions and the results were analyzed. The dataset of compounds used for all virtual screenings is described (Table 6.3).

Table 6.3: Datasets of compounds used for the virtual screening test case.

Ligand Class	Description	Number of Derivatives
CsA	Cyclosporine A (approved drug)	1
Cyclops	Derivatives accessed by decoration of a single conformer CycloPS peptide scaffold.	500
Model_1, Model_2, Model_3, Model_4, Model_5, Model_6, Model_7, Model_8, Model_9, Model_10, Model_11, Model_12	Derivatives accessed by decoration of conformers that represent the essential dynamics of the single conformer CycloPS peptide scaffold.	6000 (500 x 12)

6.4 Results and Discussion

For the virtual library, there are 27,132 different ways of choosing 6 substituents from a collection of 19 amino acids. Each collection of 6 substituents could be arranged to produce 46,656 individual derivatives giving rise to a set of 1,265,870,592 (1.2 Billion) members with unique chemical properties and anticipated biological activity.

6.4.1 Database curation

The 12 scaffolds gave rise to a virtual library that contains 15,190,447,104 derivatives (15.2 Billion). The successful complete creation of this library was confirmed through surveys of the resultant database using BASH commands.

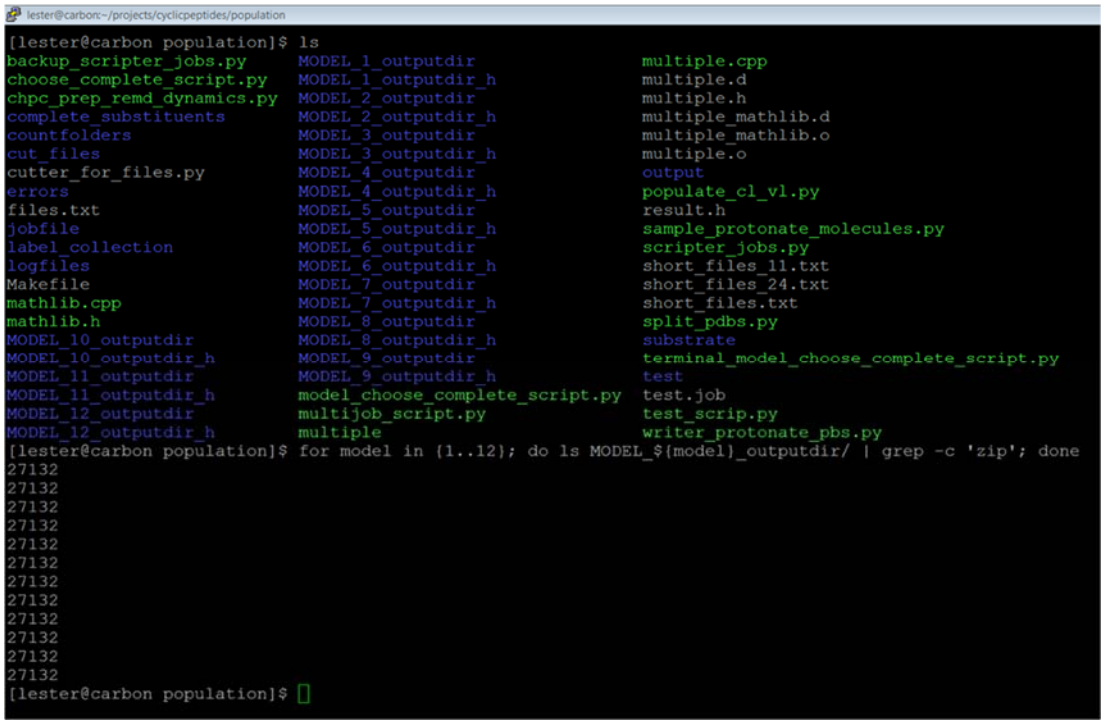


Figure 6.8: Evidence for the completed 6 membered combinations for each Conformation.

The unprotonated derivatives were stored in the MODEL_XX_outputdir folders, where "XX" denotes the conformation number from 1 to 12. Looping through each conformation folder

and counting the number of compressed collections gave evidence for the unprotonated derivatives (Fig. 6.8).

Each collection of the compressed conformations stored in “MODEL_XX_outputdir/” utilized 36 G of storage space (Fig. 6.9).

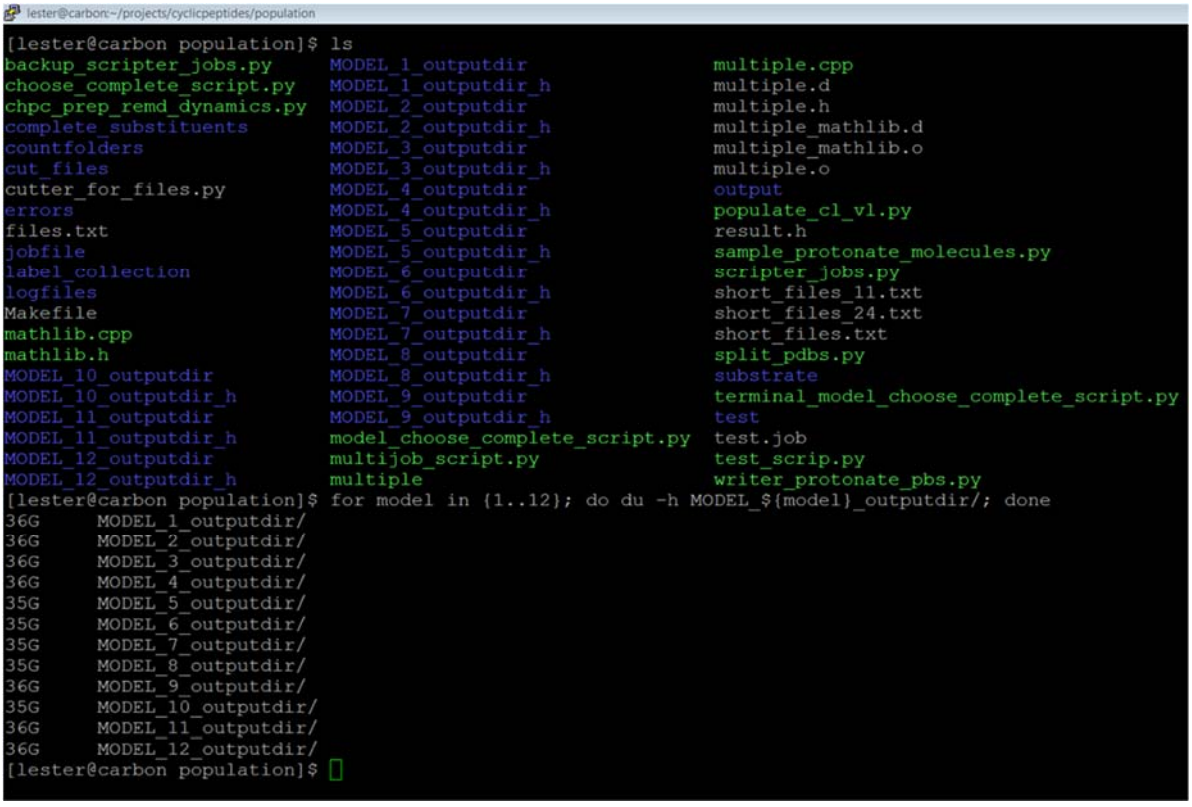


Figure 6.9: The storage space consumed by each Conformation collection.

For illustrative purposes, only one compound collection of each conformation set was inflated due to storage constraints. Inflation of the “trp_arg_gly_cys_his_leu” collection gave evidence to support the number of derivatives produced and stored within each compound collection (Fig. 6.10).

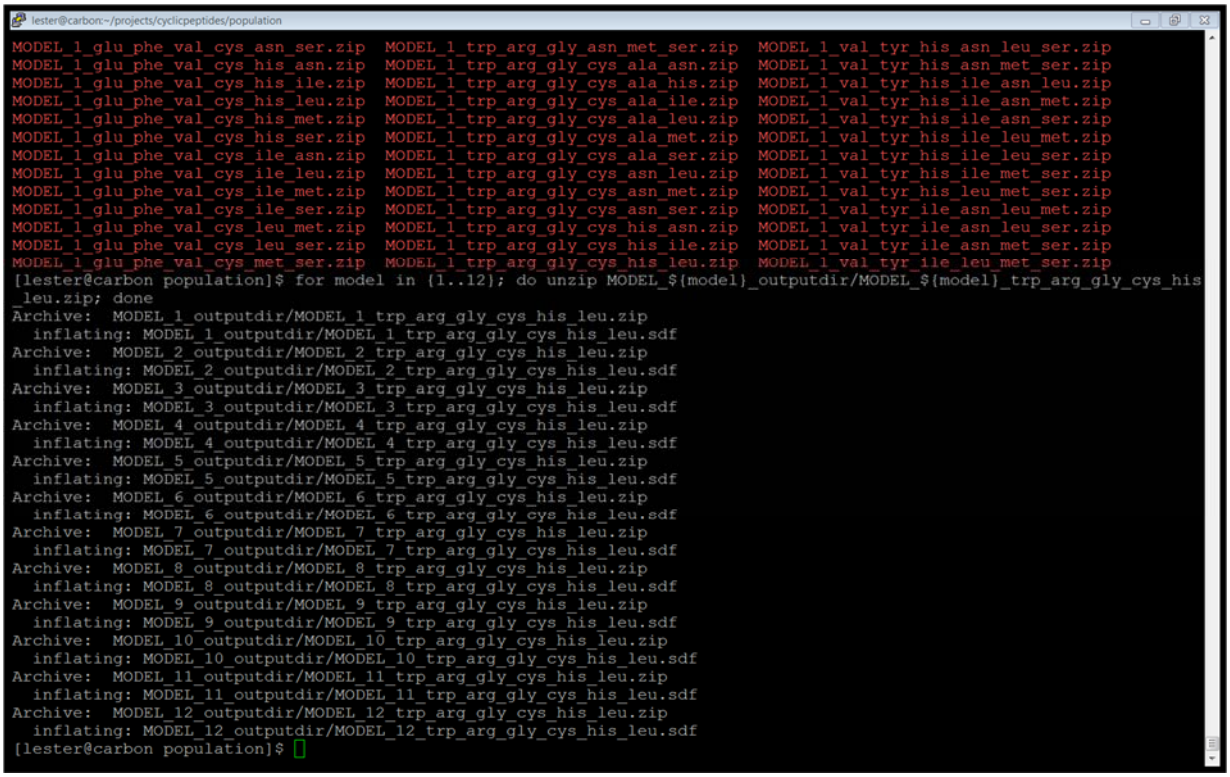


Figure 6.10: Inflation of the "trp_arg_gly_cys_his_leu" collection for each of the 12 Conformations.

The format of the inflated derivate database files for a single collection is outlined (Fig. 6.11). Although the files are labelled ".sdf", the format was changed to the ".xyz" structure format during the project to simply reduce the storage requirements. The "00000" string was used to count the derivatives present in the inflated collections by fishing for the labels of derivatives stored in each collection. Each of the inflated collections had the expected number of derivatives, 46,656.

From the evidence, we were confident that the conformationally laden cyclic peptide virtual library had been populated as intended. The 12 conformations that approximated the cyclic peptide essential dynamics were decorated exhaustively using our protocol. In summary, each conformation had 27,132 compound collections, which gave rise to 46,656 derivatives stored in compressed '.sdf' files. The virtual library of 15,190,447,104 (12 * 46,656 * 27,132)

curated derivatives consumed 427 GB of storage after compression. If the quoted compression efficiency of 2 % was used, inflation of the compressed library would require 21.35 TB of hard disk space.

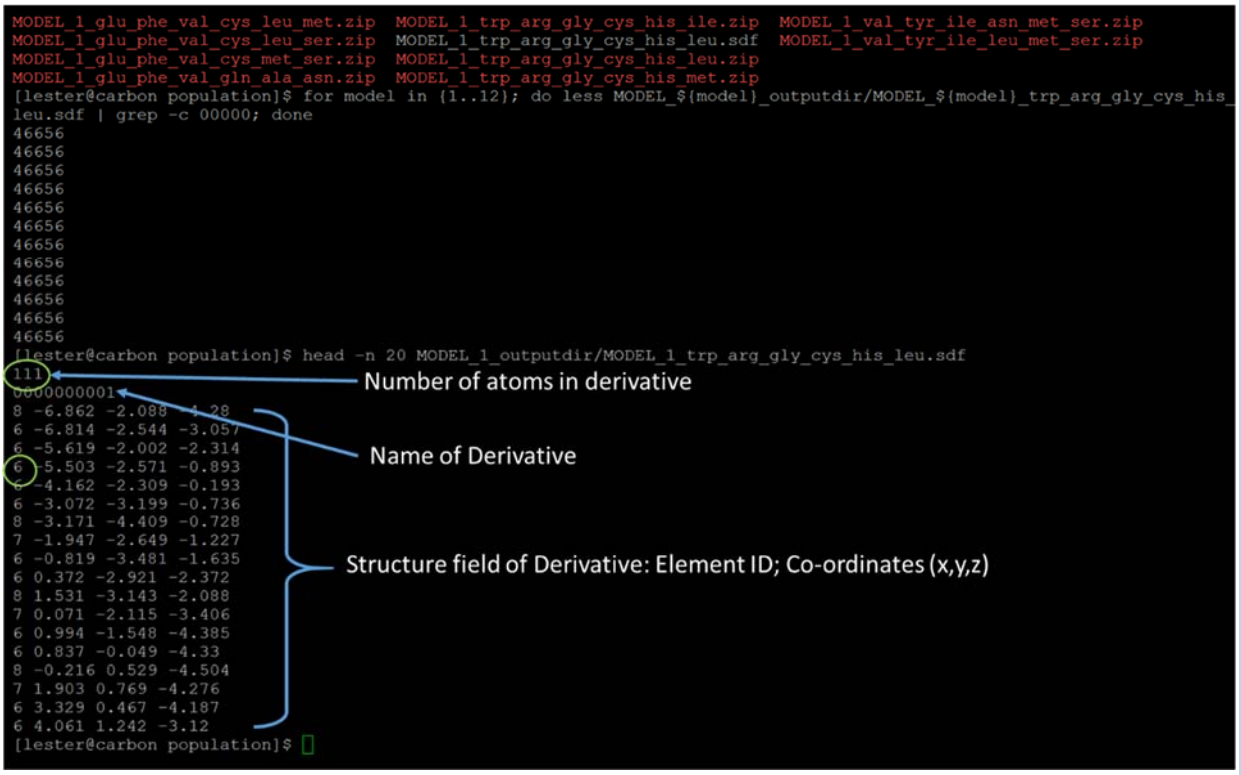


Figure 6.11: The evidence for the number of derivatives in the "trp_arg_gly_cys_his_leu" collection for each Conformation. *The format of the inflated '.sdf' format derivative collection is annotated.*

6.4.2 Physico-chemical Space

The physico-chemical properties of the subset of protonated derivatives were calculated in order to visualize the chemical space probed by the peptide space for each conformation. Calculation of these properties facilitated the visualization of the diversity of our enumerated compounds in the library in terms of chemical space.

6.4.2.1. *Curation of protonated sample*

The protonated sample stored as MODEL_XX_outputdir_h, where XX is the conformation number, was used for this study (Fig. 6.12). The compressed compound collections, appended as ".zip" files were counted (Fig. 6.12 A). The routine to access the inflated '.sdf' files (actually multimodel ".xyz" files) for the protonated "trp_arg_gly_cys_his_leu" compound collection is displayed (Fig. 6.12 B).

A.

```
[lester@carbon population]$ for model in {1..12}; do ls MODEL_${model}_outputdir_h/ | grep -c 'zip'; done
26865
26874
26864
26842
26877
26851
26878
26869
26853
26897
26865
26876
[lester@carbon population]$
```

B.

```
[lester@carbon population]$ for model in {1..12}; do unzip MODEL_${model}_outputdir_h/MODEL_${model}_trp_arg_gly_cys_his_leu_h.zip; done
Archive:  MODEL_1_outputdir_h/MODEL_1_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_1_outputdir_h/MODEL_1_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_2_outputdir_h/MODEL_2_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_2_outputdir_h/MODEL_2_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_3_outputdir_h/MODEL_3_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_3_outputdir_h/MODEL_3_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_4_outputdir_h/MODEL_4_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_4_outputdir_h/MODEL_4_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_5_outputdir_h/MODEL_5_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_5_outputdir_h/MODEL_5_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_6_outputdir_h/MODEL_6_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_6_outputdir_h/MODEL_6_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_7_outputdir_h/MODEL_7_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_7_outputdir_h/MODEL_7_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_8_outputdir_h/MODEL_8_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_8_outputdir_h/MODEL_8_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_9_outputdir_h/MODEL_9_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_9_outputdir_h/MODEL_9_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_10_outputdir_h/MODEL_10_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_10_outputdir_h/MODEL_10_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_11_outputdir_h/MODEL_11_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_11_outputdir_h/MODEL_11_trp_arg_gly_cys_his_leu_h.sdf
Archive:  MODEL_12_outputdir_h/MODEL_12_trp_arg_gly_cys_his_leu_h.zip
  inflating: MODEL_12_outputdir_h/MODEL_12_trp_arg_gly_cys_his_leu_h.sdf
```

Figure 6.12: Evidence for the protonated samples of the virtual library. *A. Count of compound collections of the compressed protonated samples. B. The inflation routine for the "trp_arg_gly_cys_his_leu" protonated samples.*

The random selection routine for the purposes of sampling meant that the total number of compound collections for each conformation did not exactly equal 27,132 (Fig. 6.12 A). Variations from 27,132 occur due to the random nature of the sampling routine (Fig. 6.12 A). For each of these 27,132 collections, a random selection of 10 derivatives (of the ~46,000 compounds in each collection was selected). This very small selection results in a protonated

subset of the library that contains ~1.4 million derivatives. The confirmation that a single protonated compound collection contains about 10 derivatives is seen for "trp_arg_gly_cys_his_leu" in Fig. 6.13.

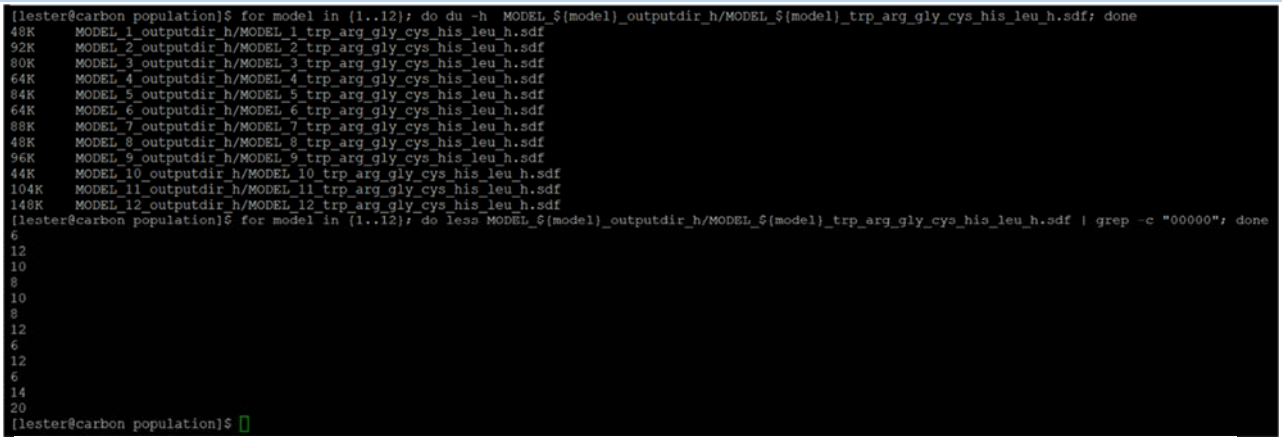


Figure 6.13: The size of the sample from the protonated "trp_arg_gly_cys_his_leu" compound collection.

The routine used for sampling gives different numbers of protonated derivatives reported for each of the "trp_arg_gly_cys_his_leu" compound collections due to the random nature of sampling (Fig. 6.13). The number of protonated derivatives for each conformation backbone (model) is summarized (Table 6.4).

Table 6.4: Survey of the number of derivatives from each conformation

Model #	1	2	3	4	5	6	7	8	9	10	11	12
Derivatives	126,224	126,004	126,995	126,890	127,127	126,569	126,356	126,049	126,786	126,550	127,430	127,021

6.4.2.2. Mapping of property space

The number of hydrogen bond acceptors, number of hydrogen bond donors and the molecular weights of the individual compounds for each of these 12 datasets were plotted (Fig. 6.13). Strongly electronegative atoms such as nitrogen and oxygen are included in the HBA atoms count (Veber et al. 2002). Heteroatoms (non-carbon) contribute to the hydrogen bond donor count if they are bound to hydrogen. The RDKit descriptor node uses the elemental weights of each of the atoms present in the compound in order to estimate the molecular weight of the compounds. Fig. 6.14 summarizes this data.

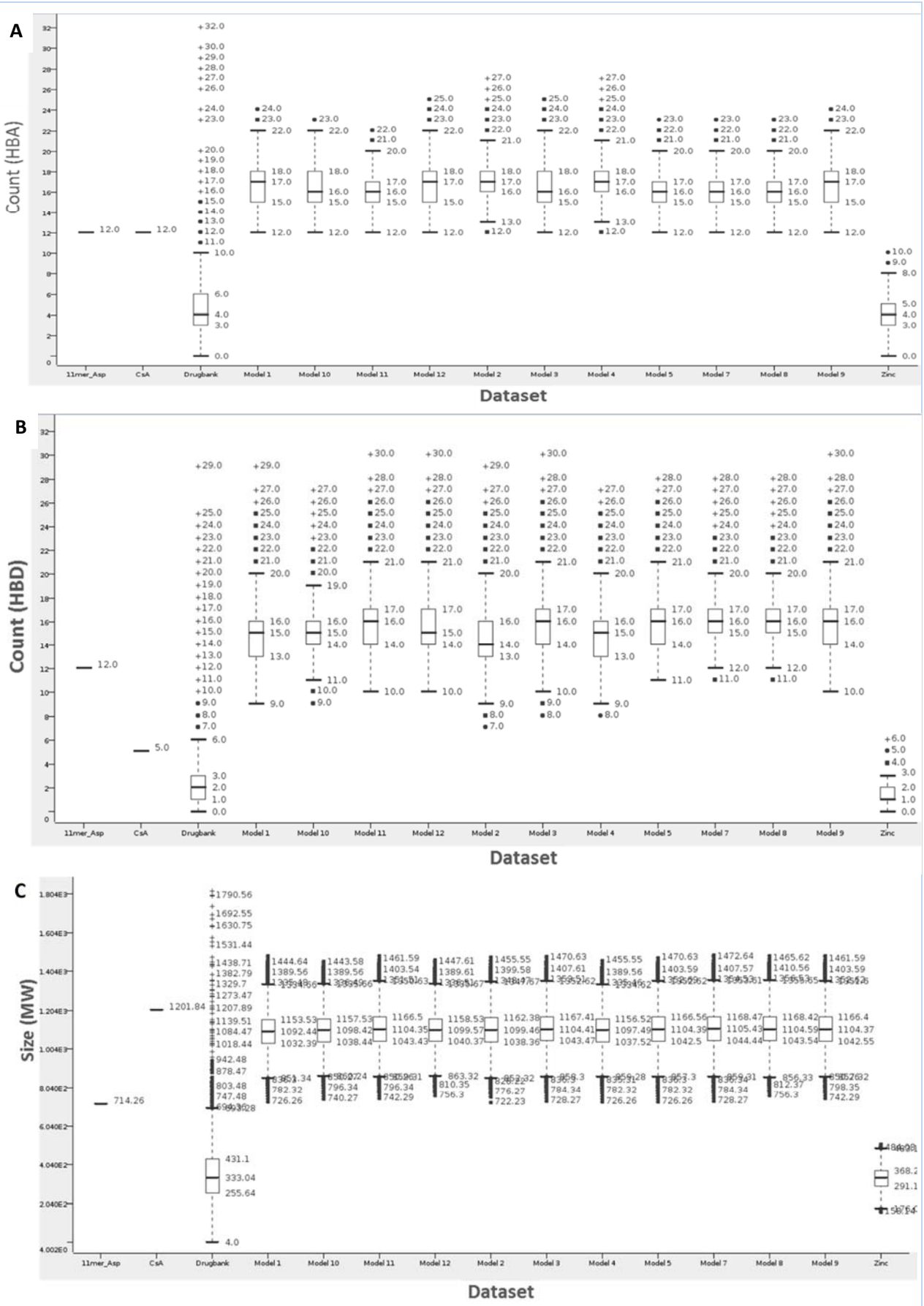


Figure 6.14: Box-plots of the chemical space of 11mer_Asp, CsA, Drugbank, protonated derivatives and Zinc datasets. *Chemical space as: A. count of HBA; B. count of HBD; C. MW space*

The decoration routine increases the HBA counts, the HBD counts and the MW of the derivatives as compared to the 11mer scaffold (11mer_Asp) (Fig. 6.14). Compounds within the cyclic peptide (Model 1-12) datasets have median HBA counts of between 16 and 17 compared to 12 recorded for 11mer_Asp and CsA (Fig. 6.14 A). The Drugbank drug-like compounds and the ZINC small molecule datasets have median HBA counts of four. Outlier compounds present in Drugbank record HBA counts as high as 190 (not shown) although 75% of its compounds have counts of 6 or lower similar to that of the ZINC small molecule dataset. The cyclic peptides do not exceed an HBA of 27 with their range spanning counts of 12 and 22.

The HBD count of 5 for the CsA drug is much lower than that of the 11mer scaffold with 12 (Fig. 6.14 B). Unlike the HBA counts, a significant number of derivatives with HBD counts lower than the 11mer scaffold are observed with some recording counts as low as 7. The median HBD counts for the derivatives sets lie between 14 and 15 with the majority of compounds recording counts lower than 21. The Drugbank and ZINC datasets have compounds with HBD counts as low as 0 and had median counts of 2 and 1 respectively. Although the Drugbank dataset had an extreme count of 113 (not shown), only a few extreme compounds matched the HBD space of these cyclic peptides. Although the CsA molecule had a larger HBA count than the majority of Drugbank compounds, its HBD count falls within the range of the Drugbank set. Lead-optimization stages of active hits aim to lower the HBA and HBD counts to increase compound bioavailability. For peptides this could be achieved through methylation of nitrogen atoms (Räder et al. 2018). Ultimately, N-methylation is intended for the conformationally laden library.

As expected the median MW for the cyclic peptides showed a significant increase to 1,000 g/mol compared to the 714 g/mol of the scaffold owing to the addition protocol (Fig. 6.14 C).

None of the derivatives had MW's lower or equal to the 11mer scaffold as expected. The CsA drug has a MW of 1,201 g/mol that falls within the range of the derivatives populated by our protocol. The majority of compounds present within the Drugbank dataset are smaller than our peptide derivatives although outliers exist that overlap with the peptide space. None of the compounds present in the ZINC small molecule set overlap with the peptide space.

Oral bioavailability characteristics of compounds that influence the formation of hydrogen bonding networks, and ion-dipole interactions which favour solubility can be defined by measuring the octanol/water partition coefficient (Log P) (Lemke et al. 2013). The adsorption of drug candidates from the gastrointestinal tract is also dependent on their transcellular passage which favours nonpolar lipophilic compounds obtained from their topological polar surface area (TPSA) (Smith et al. 1996; Manallack et al. 2014). For our study the RDKit descriptor used SLogP to estimate the Log P while the total polar surface area (PSA) contributions of the polar fragments that make up the candidate structures were used to estimate the TPSA (Ertl et al. 2000). The solubility and polarizability of compounds in our datasets are summarized (Fig. 6.15).

The solubility characteristics of 25 % of the cyclic peptide derivatives overlap with 50 % of compounds from the Drugbank databases (Fig. 6.15 A). The solubility score of 3.27 for the CsA drug lies within the middle 50 % of coumpounds present in the approved Drugbank and soluble small molecule ZINC datasets. The 11mer scaffold with a score of -8.37 falls outside the range of the Drugbank dataset. Over 75 % of the peptide derivatives have a solubility score better than that of the 11mer scaffold with the lower quartile as low -12.17. The median solubility for the cyclic peptide derivatives is between a score of -4.69 and -5.21. From these observations, it is plausible to assume that the majority of compounds from the peptide

derivatives would be insoluble and would require optimisation in order to enhance their bioavailability.

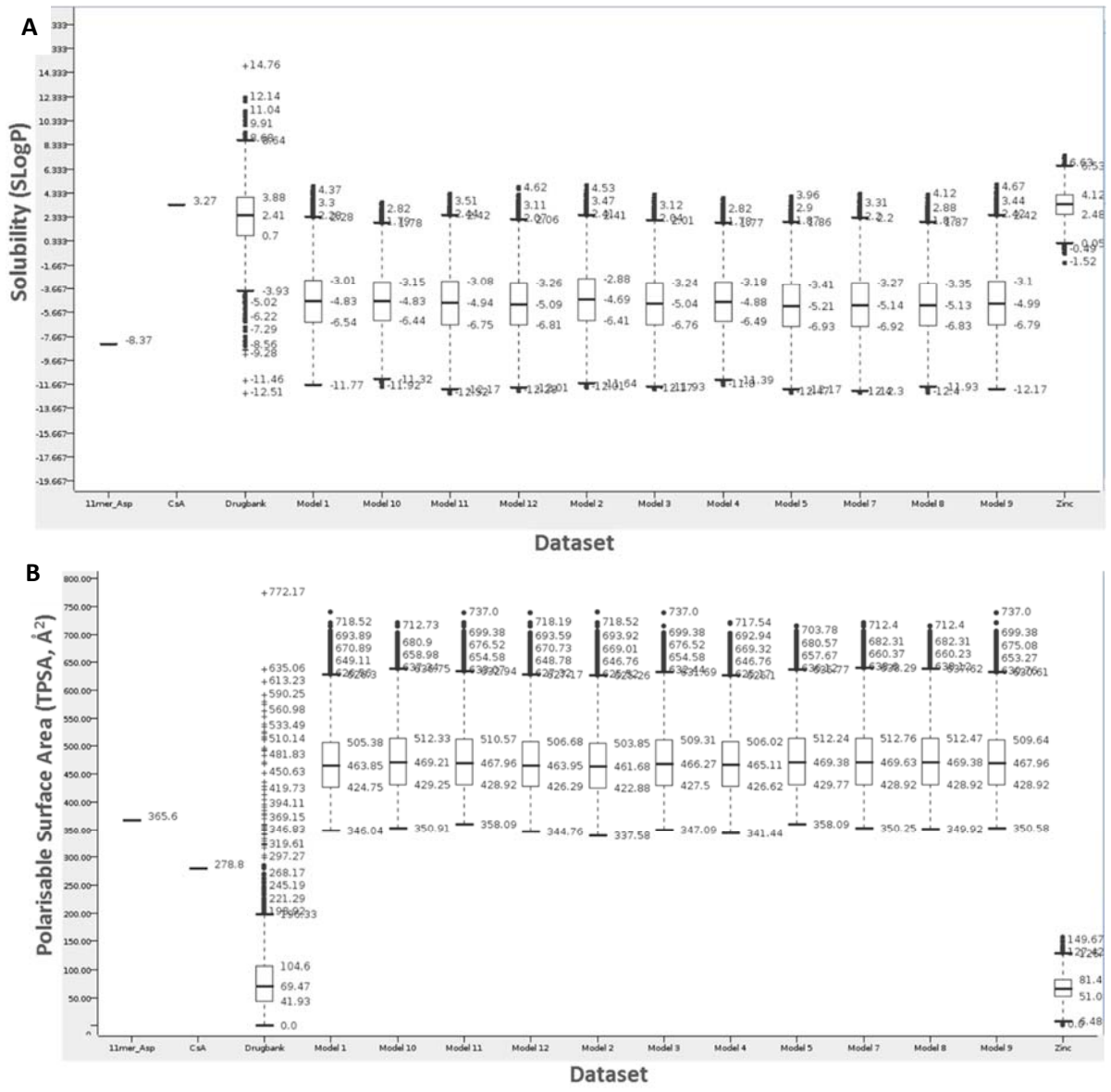


Figure 6.15: Box-plots of the solubility and polarizability space of 11mer_Asp, CsA, Drugbank, protonated derivatives and Zinc datasets. *Chemical space as: A. Solubility and B. Polarizable Surface Area.*

The polarizability plots (Fig. 6.15B) justify the trends observed from the solubility data. The peptide derivatives are likely to display lower or more unfavorable solubility due to their high polarizable surface area when compared to the drug like Drugbank and small molecule ZINC datasets (Fig. 6.15 B). While the CsA drug's TPSA of 278.8 $Å^2$ lay outside the Drugbank and

ZINC dataset limits, the CsA TPSA was significantly lower than that of all the peptides despite having a similar size w.r.t. MW (Fig. 6.14 C). The CsA drug is optimized for bioavailability through its low TPSA that enhances its solubility profile. This mitigates against its large size, which should reduce its bioavailability. All the peptide derivatives including the 11mer scaffold had TPSA's outside the limits of the Drugbank datasets. The median TPSA of the peptides lay between 463.85 $Å^2$ and 469.63 $Å^2$ highlighting a significant increase in the polarizable surface area compared to that of the 11mer scaffold; this increase is due to decoration, but is justified due to the inclusion of charged amino acids during decoration of the scaffold.

This discussion on the impact of decoration on the physico-chemical properties of the conformers provides further evidence for the reliability of the decoration protocol we implemented. Cumulative properties such as the HBD and HBA counts, MW and TPSA showed the property space expansion of the derivatives when compared to the 11mer scaffold.

6.4.3 Synthetic feasibility

The 117,129 sequences were passed through the CycloPS filter in order to determine which sequences satisfied the CycloPS synthesisability criteria. This sample set produced 49,704 synthesisable sequences and 67,425 unsynthesisable sequences. The manually selected chemical and structural descriptors computed by QikProp were used as features in the generation of Support Vectors, from the SVM learner, as a means to predict synthetic feasibility (Fig. 6.7A). The accuracy and precision in matching the CycloPS scores of the test dataset allowed for hyper parameter optimization (Table 6.5). The code used for classifications are: CycloPS non-Synthesisable (0), CycloPS Synthesisable (1), Predicted non-Synthesisable ('0') and Predicted Synthesisable ('1').

Table 6.5: Hyper parameter optimization of the HyperTangent, Radial Basis Function and Polynomial Support Vector Machine Kernels. *OP = Overlap Penalty; κ = Kappa; δ = Delta; σ = Sigma; P = Power; B = Bias; γ = Gamma.*

		True Negative	False Positive	False Negative	True Positive	
	OP_κ_δ	0:'0'	0:'1'	1:'0'	1:'1'	Total
HT	05_01_05	Maximum # of iterations reached				0
	10_01_05					
	15_01_05					
	OP_σ	0:'0'	0:'1'	1:'0'	1:'1'	Total
RBF	05_01	67333	0	49625	0	116958
	05_05	67333	0	49625	0	116958
	05_1	Maximum # of iterations reached				0
	10_01	67333	0	41174	8451	116958
	10_05	67333	0	41174	8451	116958
	10_1	67333	0	41174	8451	116958
	15_01	67333	0	41174	8451	116958
	15_05	Maximum # of iterations reached				0
	15_10	67333	0	41174	8451	116958
	OP_P	0:'0'	0:'1'	1:'0'	1:'1'	Total
POL	05_01	43113	24220	27159	22466	116958
	05_05	43107	24226	27167	22458	116958
	05_1	41670	25663	26119	23506	116958
	10_01	44084	23249	22995	26630	116958
	10_05	39940	27393	25826	23799	116958
	10_1	Maximum # of iterations reached				0
	15_01	42716	24617	26862	22763	116958
	15_05	41438	25895	26286	23339	116958
	15_10	37252	30081	26222	23403	116958
	B_γ	0:'0'	0:'1'	1:'0'	1:'1'	Total
POL 10_01	ALL	67333	0	49625	0	116958

The tests showed that the best performing parameters were the Radial Basis Function parameters. All of the HyperTangent parameter sets tested were unable to optimize within the limit of the number of iterations. Optimization of the Polynomial (OP = 1.0; Power = 0.1) kernel through varying the Bias and Gamma did not yield a satisfactory optimization. Table 6.3 showed that the testing datasets had a total of 116,958 sequences (171 members less than the original test set). The missing sequences were sequences who had missing features

during the QikProp calculation. The approach taken to test the sequence set for Table 6.3 included training data. This allowed us to eliminate kernels and their parameters that were unable to classify sequences as being either synthesisable or not.

When the RBF SVs were interrogated further they were unable to classify any of the sequences that were not used in training during the testing stages (Table 6.6). Although the RBF SVs could perceive sequences that were synthesisable (Table 6.5), classifications of any dataset not used in training resulted in no prediction of synthesisability (Table 6.6).

Table 6.6: RBF parameters tested on the dataset that included training data (Blue) and parameters tested when training data was excluded (Red).

	OP_σ	0:'0'	0:'1'	1:'0'	1:'1'	Total
RBF	10_01	67333	0	41174	8451	116958
	10_05	67333	0	41174	8451	116958
	10_1	67333	0	41174	8451	116958
	10_01	55814	0	41174	0	96988
	10_05	55814	0	41174	0	96988
	10_1	55814	0	41174	0	96988

At present, this SVM is unable to deal with data from the derivative library. Future work will concentrate on reducing the number of features used to create the SVM. This may aid both in the reduction of false negative results, but also enable the model to handle the new data (from our derivative library) in an adequate manner. Enhancement of this approach may include the application of Deep Learning Neural networks that will give rise to SVs that are likely to resolve the traits we are interested in (synthesisable or not). It is estimated through these experiments that this approach will improve the screening time by a factor of 10 over CycloPS.

6.4.4 Virtual Screening Test

Virtual screening using a molecular docking approach was used to test the impact of introducing conformation diversity into the virtual library. The set of derivatives were submitted to a docking experiment using Glide and the docking score results of the ligands were extracted (Fig. 6.16).

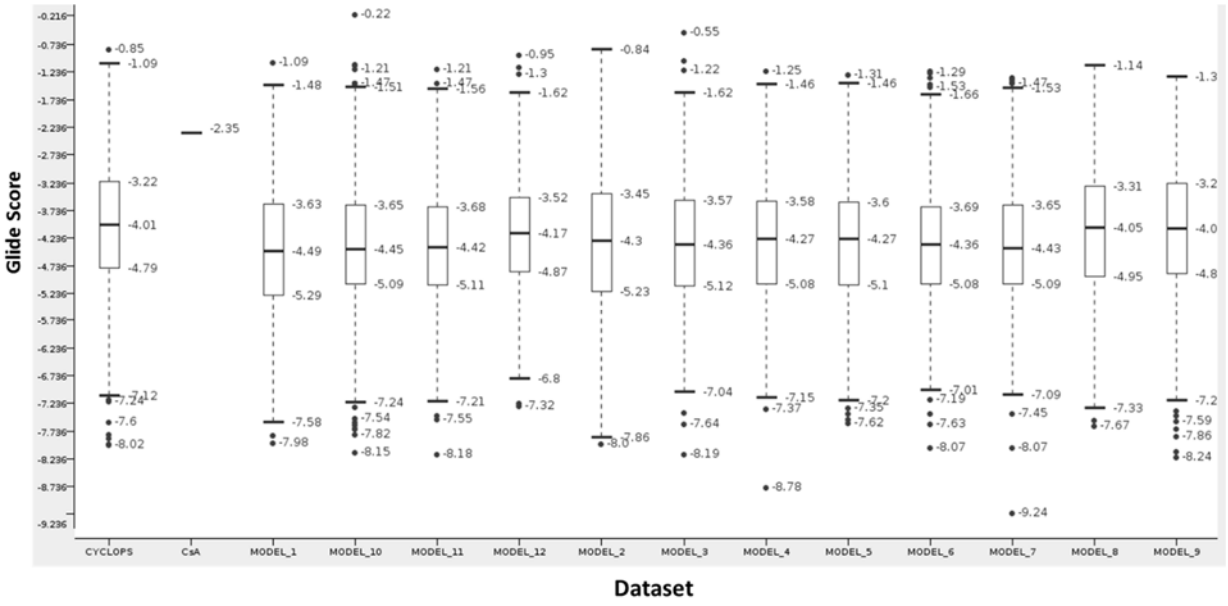


Figure 6.16: Boxplots of the Glide dock scores for the ligands from the Datasets docked against the cyclophilin-D target.

The Glide score reported for each of the docked ligands characterizes the strength of affinity of each of the ligands docked. For each of the datasets the majority of the derivatives generated from our decoration reported affinities that were better than those of the CsA approved drug. Although the range of docking scores for the derivatives from the diverse conformation datasets were similar to the range observed for the single conformation dataset (labelled CycloPS), differences can be seen in the individual ligands and also in the outliers. A derivative produced from the Model 7 conformer had the lowest score of -9.24. This is an

improvement over the single conformation screening with the best binding having a Glide score of -8.02 (label CYCLOPS). A survey of the docking results is shown (Fig. 6.17).

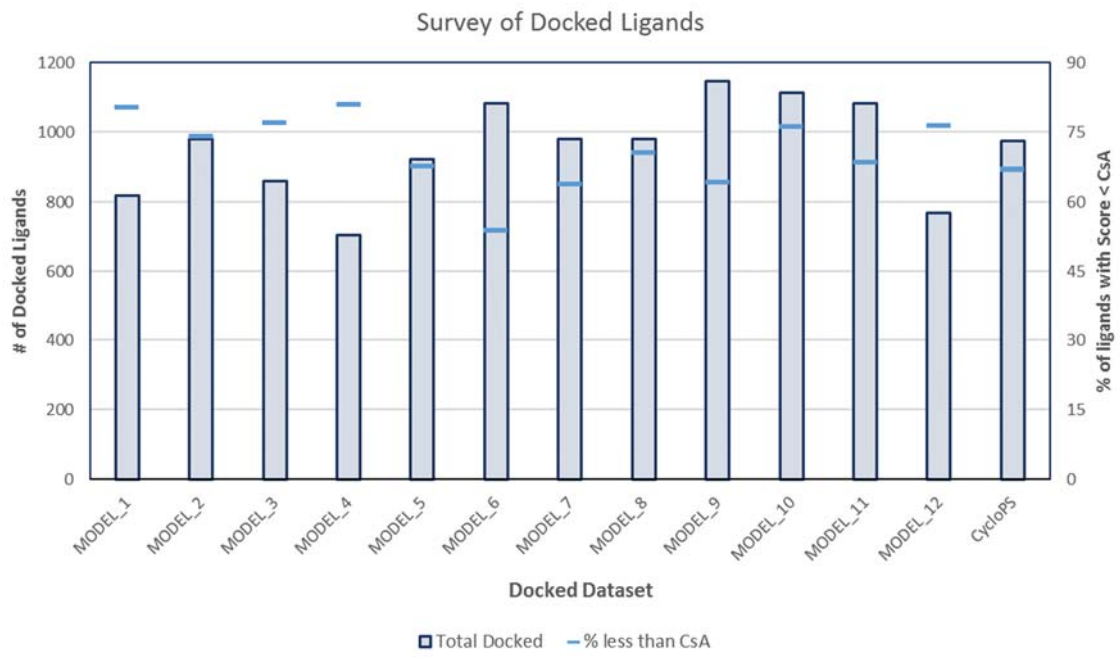


Figure 6.17: Survey of docked ligands and their performance against the CsA approved drug.

Each of the datasets required the docking of over 500 ligands. This is because LigPrep provides some variation on the input structures – this variation may be due to different tautomers or ionization states, or this could also be an additional ring conformation per input structure. Of interest was a count of ligands that had docking scores better than the CsA benchmark. Model 4 and Model 1 ligands performed best in this respect, with 81% and 80% of each of these sets respectively having a docking score better than CsA. This corresponds to a large number of structures even though LigPrep increases the number of structures docked by a small amount, *viz* from 500 to 705 and from 500 to 818 respectively. The Model 4 dataset boasted the second most potent ligand (Fig. 6.16). For the CycloPS single conformer dataset, only 66% scored

better than CsA in docking. This is comparable to three other conformation sets: the Model 9 (64 %); the Model 7 (63 %); and the Model 6 (53 %) datasets.

This high-resolution virtual screening includes more of the conformational landscape of these scaffolds whilst also considering derivatives with different chemistries. Some systems that would not have been identified through screening (Fig. 6.18) are identified as good inhibitors. The identities of the 10 ligands with the best affinities across the entire sets of data are considered (Fig. 6.18).

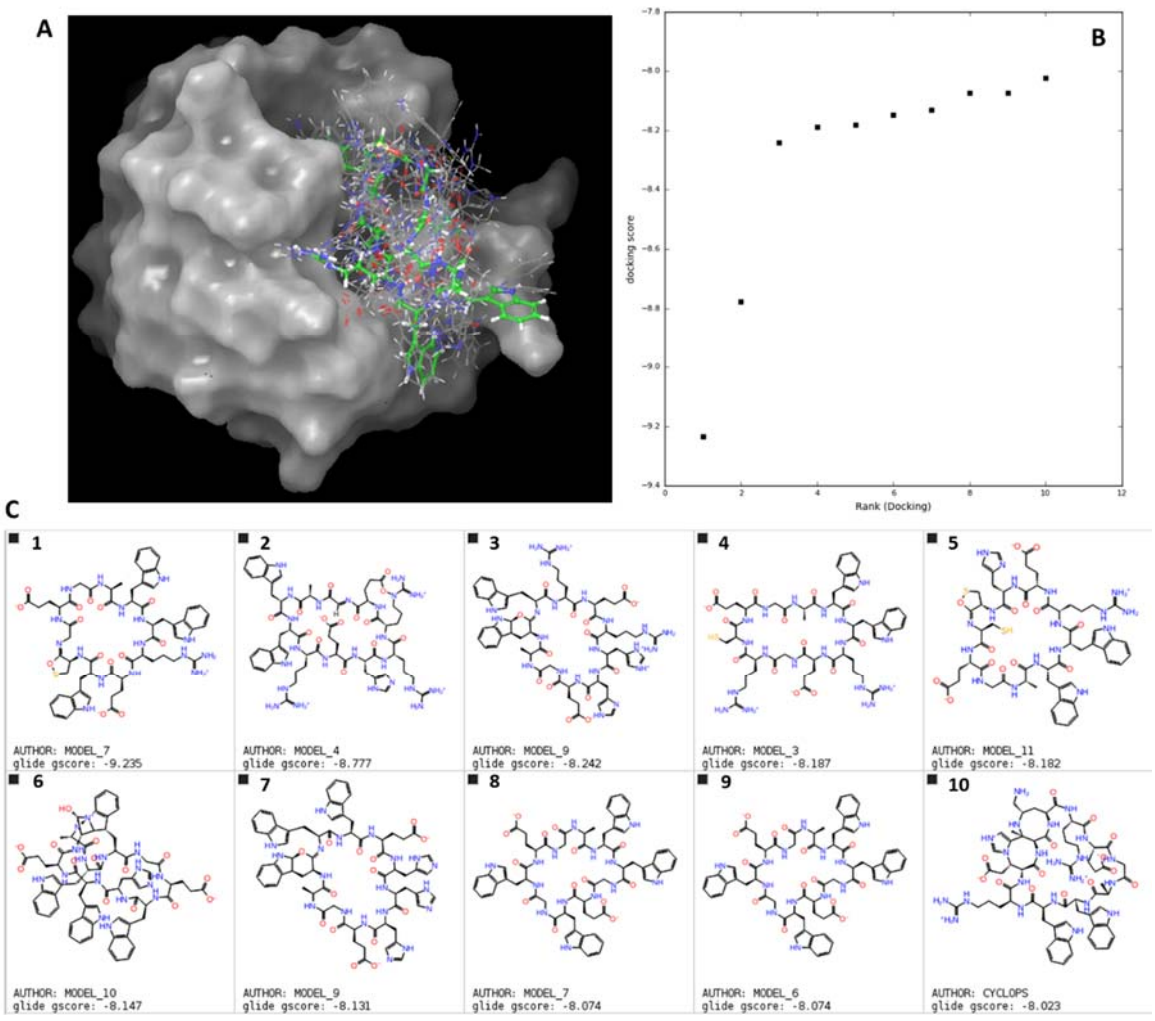


Figure 6.18: Assessment of the top 10 docked ligands. *A. docked positions of the top 10 ligands within the cyclophilin-D pocket. B. Glide docking scores of the top 10 ligands. C. Structures of the top 10 ligands (**"AUTHOR"** refers to the dataset source).*

Visualization of the docking poses showed that all 10 ligands exploit the same pocket (Fig. 6.18 A) as directed by the docking grid parameters. The nature of the chemistries used by the ligands to interact with the receptor surfaces leads to the differences in the docking scores observed. Conformations from the Model 7 and Model 9 datasets favoured 2 ligands each in the top 10 collection while none of the ligands from Model 1, Model 2, Model 5, Model 8 and Model 12 datasets were present in this set (Fig. 6.18 C). Ligand interaction diagrams highlight how the different ligands exploit different chemistries in their interactions (Fig. 6.19). From this it is clear that including ring conformations has led to some ring conformation sets binding well and some less well. On top of this, it is also clear that side chains may compensate for ring conformations that do not fit well, sometimes to the extent that the binding is better. In any case where ring conformation has not been addressed (CycloPS single conformation), conformational searching during docking does not arrive with the excellent complementarity observed when ring conformations are included (Models 1-12).

The top performing ligand from the Model 7 dataset (Fig. 6.19 A) recruited 12 interactions with binding pocket residues while the ligand from the CycloPS single conformation dataset which was the 10^{th} best performing ligand was only able to recruit 6 interactions in its conformation (Fig. 6.19 B). The conformationally laden library allowed for the identification of docking poses in which stabilizing π, π interactions were observed – something that was not observed where ring conformations were not included in the library. The best ligand from the single ring conformation library (Fig. 6.19 B) was stabilized exclusively by H-bond interactions with residues His54 (2), Asn71, Thr73, Asn102, and Trp121. The conformation from the high affinity Model 7 ligand also exploited a salt bridge (Arg55), H-bonds (Arg55, Asn102, Thr73, His126, and Trp121), pi-pi stacking (Trp121), and pi-cation (Lys148) types of interactions with the target receptor.

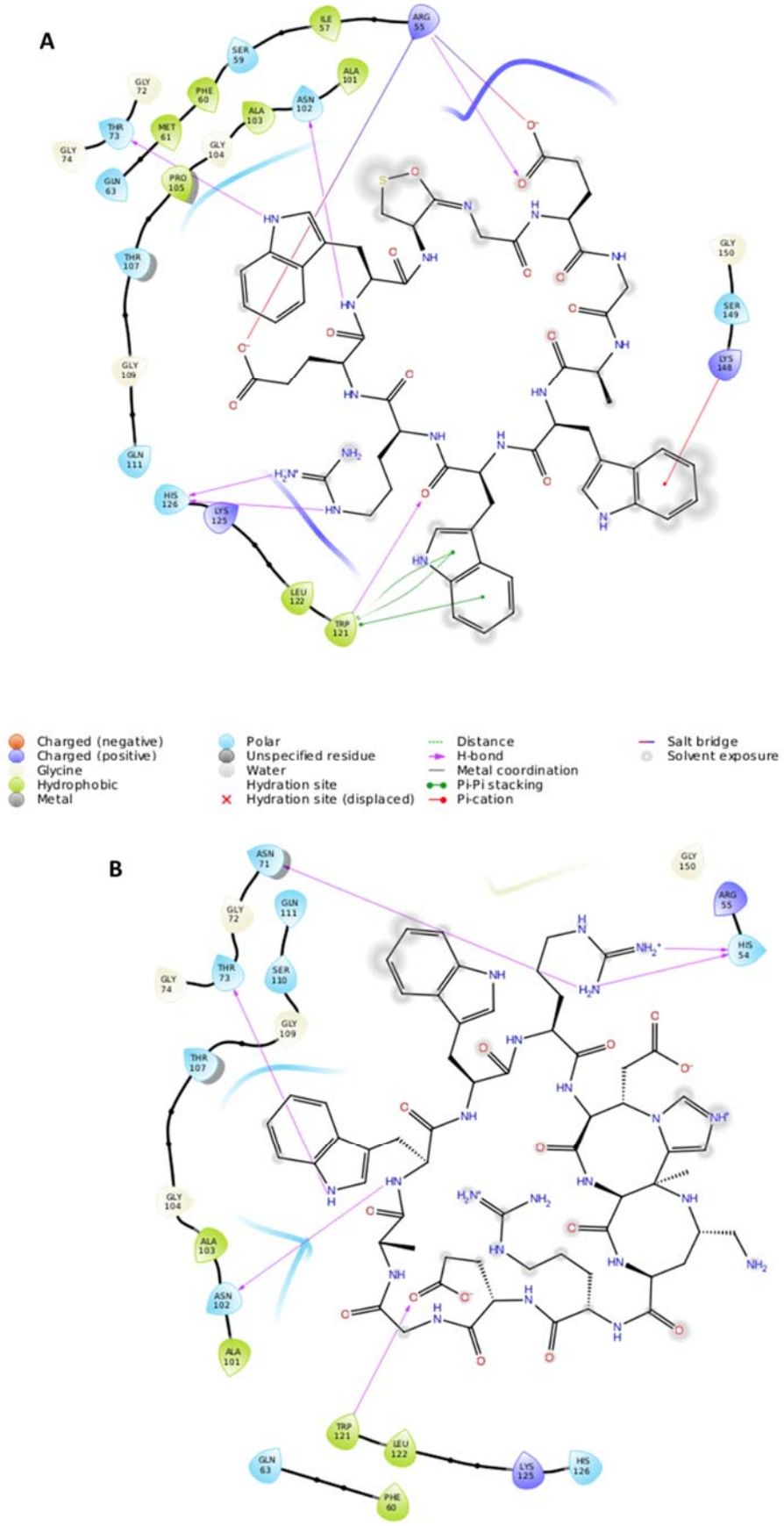


Figure 6.19: Ligand interaction diagrams of the docked poses. *A. The best binding ligand (Model 7 dataset). B. The 10th best binding ligand (Cyclops dataset).*

High-resolution screening, meaning that screening is of a conformationally laden dataset (particularly of ring conformations), within a high-throughput virtual screening paradigm allows for a more detailed search with better results. Although high precision virtual screening that makes use of REMD enhanced sampling is available (FEP+), its application is restricted to the lead optimization of a handful of derivatives (between 10-20). Our implementation makes use of REMD to define the conformations, but extends the level of search to be in a large sequence space of peptides (billions of structures). It identifies good binding candidates (even on a subset of the library) that would not have been identified through a straightforward docking protocol. This has a direct impact on the efficiency and productivity of the early stage Hit identification stages of the drug discovery pipeline.

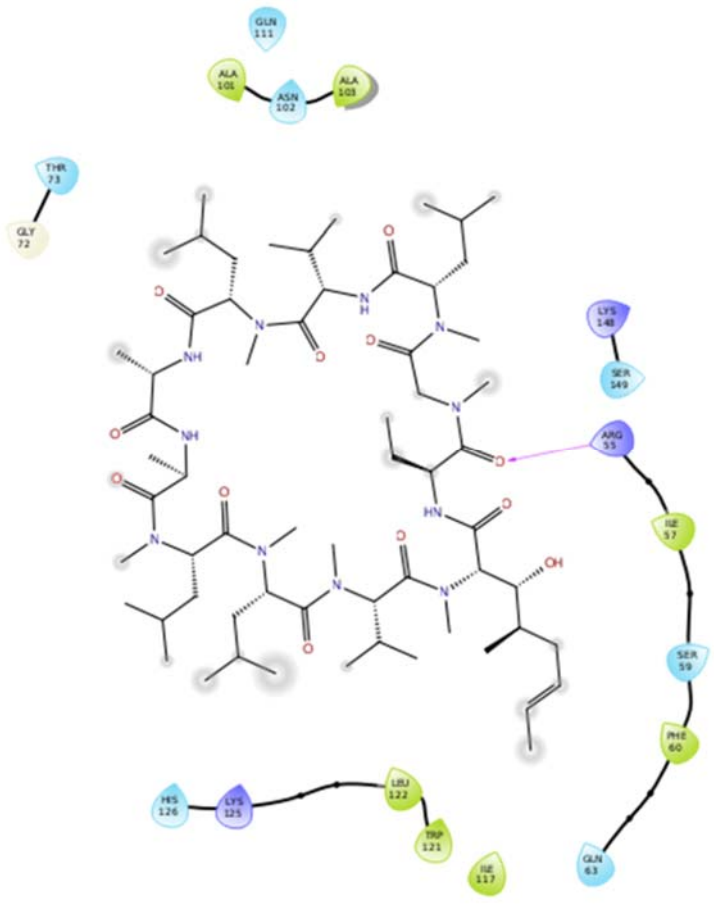


Figure 6.20: Ligand interaction diagram of the docked pose of cyclosporine-A

The lower docking score observed by CsA under our test conditions can be justified by interrogating the interaction map (Fig. 6.20). This ligand exploits a single H-bond interaction with Arg55. Solvent exposed surfaces of proteins are dominated by polar residues that are more likely to behave as H-bond acceptors (Doak et al. 2015), and the derivatives provide more of these interactions.

6.5 Conclusion

The use of aggregated computing allowed the population of a conformationally laden virtual library of cyclic peptides from a limited set of sequences to be completed within a reasonable time. The flexibility of DerivatizeME allowed us to recruit compression algorithms that allowed the library of 15.2 billion derivatives to be stored in 427 GB of storage. When a sample of derivatives were mapped onto plots of the Hbond acceptors, Hbond donors, size, solubility and polarizability, it could be seen that the virtual library expanded the chemical space of the query scaffold. The expansion of property space of the query peptide lead us to believe that it was possible to find derivatives that had improved potency and selectivity characteristics than the approved CsA drug for application as a cardio protection agent.

Enhancing the resolution of virtual screening by including conformation content in the virtual library was shown to contribute to high-throughput virtual screening of a ligand set. Including REMD sampling in the conformational search of a billion compounds was previously considered out of reach. Our approach of decoupling conformational search from enumeration allowed HTVS to efficiently leverage on this enhancement without compromising on precision. This approach could be used to exploit peptide macrocycles for the pursuit of molecular entities that perturb difficult to drug, beyond rule of 5 drug targets such as cyclophilin-D.

Chapter 7: Conclusion

A survey of the pharmaceutical industry highlights two trends related to the cost of research and Pharma efficiency. Despite compounded advancements in biotechnology and manufacturing technology, the number of drugs approved for every $1 Billion (USD) spent in research and development has been declining precipitously (DiMasi & Grabowski 2007; Scannell et al. 2012). Interventions to mitigate against these high costs have biased molecule screens toward small molecules that possess properties amenable to miniaturized and automated high-throughput screening (Zhang et al. 2011). The result has been an extremely high failure rate during the pre-clinical development stage (Sams-Dodd 2013). The success rate of pharmaceutical companies chaperoning a molecular entity to approval is as low as 6 in every 100 over the last 6 decades (Nicolaou 2014; Bonabeau et al. 2008).

Despite this negative outlook, a new trend is emerging. By drifting away from prior assumptions of the "druggability" of biological targets and the "druglikeness" of molecular probes, new targets and molecular screens are becoming valuable (Doak et al. 2014). This "beyond rule of 5" drug space is buoyed by evidence of the existence of complementarity between previously neglected drug targets and the molecules classes that perturb their function (Doak et al. 2015). These previously difficult to drug targets are amenable to macrocyclic entities that have large surface areas, increased flexibility and privileged bioavailability properties (Rask-Andersen et al. 2011; Lau & Dunn 2017). Akin to these ideas of druggability have been ideas that improve chemical search by exploring and expanding the universe of chemical diversity (Xie et al. 2015; Lowe 2015). The tuneability and diversity of peptide derived macrocycles lend themselves generously to the exploration of beyond rule of 5 molecular probes that possess privileged bioavailability (Uhlig et al. 2014). These properties

contribute to the observation that peptide-derived therapeutics have higher success rates in pre-clinical development over traditional small molecule therapeutics (Lau & Dunn 2017).

The exploration of peptide based therapeutics benefits significantly from the incorporation of molecular modelling techniques owing to the magnitude of the “linguistically” accessible chemical space probed by basic scaffolds of up to 13 members. Despite progress made in the use of molecular modelling for the rapid searching of small molecule compound libraries precision is lost in the interrogation of larger macrocycle libraries (McHugh, Rogers, Solomon, et al. 2016). Enhancements to the precision and resolution of high-throughput virtual screening approaches will contribute significantly to the design of peptide based focused libraries for lead identification. Optimizing virtual screening protocols to search a larger portion of peptide chemical space reliably will reduce the cost of investment in early stage R&D whilst improving efficiency and reducing the failure rate.

Our ambition in this study was to contribute towards the identification of peptide macrocycles that are engineered towards the perturbation of a validated target for cardio protection, cyclophilin-D. Our interest was the population of a virtual library that explores both chemical space in terms of the sequence space and of conformational diversity. This library would facilitate the identification of peptide Hit candidates from high-resolution virtual screening using existing high-throughput searching tools.

In order to overcome the limitations of existing enumerators we built DerivatizeME and embedded the ability to perform exhaustive, intelligent, systematic or random enumeration of a query bare scaffold from a limited set of substituents. DerivatizeME allowed us to decouple the exhaustive enumeration of sequence space from the problem of reliable conformational searching. Our approach was to decorate a collection of conformations that

best describe the essential dynamics of a peptide backbone. In this implementation, the peptide's essential dynamics were isolated by nearest neighbour clustering of conformations accessed by enhanced sampling using solute tempered Replica Exchange Molecular Dynamics simulations. Analysis of the low temperature trajectories reduced the likelihood of extracting conformations that only existed at high temperatures. One limitation of our enhanced sampling protocol was its reliance on nearest neighbour clustering for isolating the essential dynamics.

By aggregating computing power and recruiting compression algorithms our enumeration strategy populated a virtual library of 15.2 billion derivatives that explore the limited sequence space of an 11 membered peptide scaffold. A sample of this virtual library showed that the property space of the scaffold was expanded as expected due to this exhaustive enumeration. The conformational content of the library enhanced the virtual screening protocols demonstrating the relevance of our approach. Although rule-based filters have been used to filter peptides for synthesizability by CycloPS, we attempted to improve on this approach through machine learning derived regression. Our approach used CycloPS to provide data on synthetic feasibility, and this data was used to train a support vector machine (SVM). Future work may make use CycloPS data in a similar way but with deep learning and neural networks to improve reliability.

In conclusion, our study has demonstrated the importance of conformation in virtual libraries. The focus is on cardioprotective agents, for the treatement of perfusion injury. To this end a subset of our generated library of 15.2 billion cyclic peptides (with conformational information) was screened against cyclophilin-D. The result of this was the identification of cyclic peptides with a much higher affinity *in silico* for cyclophilin-D than the best known binder, cyclosporine A.

References

Abel, R., Mondal, S., et al., 2017. Accelerating drug discovery through tight integration of expert molecular design and predictive scoring. *Current Opinion in Structural Biology*, 43, pp.38–44. Available at: https://www.sciencedirect.com/science/article/pii/S0959440X16301701?via%3Dihub#sec0015 [Accessed September 25, 2018].

Abel, R., Wang, L., et al., 2017. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Accounts of Chemical Research*, 50(7), pp.1625–1632. Available at: http://pubs.acs.org/doi/10.1021/acs.accounts.7b00083 [Accessed September 25, 2018].

Abrams, C. et al., 2013. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy*, 16(1), pp.163–199. Available at: http://www.mdpi.com/1099-4300/16/1/163 [Accessed September 19, 2018].

Abrams, C. & Bussi, G., 2014. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1), pp.163–199.

Aderogba, M.A. et al., 2012. Isolation of antioxidant constituents from Combretum apiculatum subsp. apiculatum. *South African Journal of Botany*, 79, pp.125–131. Available at: http://dx.doi.org/10.1016/j.sajb.2011.10.004.

Ageitos, J.M. et al., 2013. Proteinase K-Catalyzed Synthesis of Linear and Star Oligo(l -phenylalanine) Conjugates. *Biomacromolecules*, 14(10), pp.3635–3642. Available at: http://pubs.acs.org/doi/10.1021/bm4009974 [Accessed October 10, 2018].

Ageitos, J.M. et al., 2016. The Benzyl Ester Group of Amino Acid Monomers Enhances Substrate Affinity and Broadens the Substrate Specificity of the Enzyme Catalyst in Chemoenzymatic Copolymerization. *Biomacromolecules*, 17(1), pp.314–323. Available at: http://pubs.acs.org/doi/10.1021/acs.biomac.5b01430 [Accessed October 10, 2018].

Agostini, F.P. et al., 2006. Generalized simulated annealing applied to protein folding studies. *Journal of Computational Chemistry*, 27(11), pp.1142–1155. Available at: http://doi.wiley.com/10.1002/jcc.20428 [Accessed September 21, 2018].

Ahlbach, C.L. et al., 2015. Beyond cyclosporine A: conformation-dependent passive membrane permeabilities of cyclic peptide natural products. *Future Medicinal Chemistry*, pp.1–10. Available at: http://www.future-science.com/doi/abs/10.4155/fmc.15.78 [Accessed October 22, 2015].

Al-Mallah, M.H. et al., 2006. Angiotensin-Converting Enzyme Inhibitors in Coronary Artery Disease and Preserved Left Ventricular Systolic Function. *Journal of the American College of Cardiology*, 47(8), pp.1576–1583. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0735109706002944 [Accessed October 22, 2018].

Albericio, F., 2004. Developments in peptide and amide synthesis. *Current Opinion in*

Chemical Biology.

Alessandri, T.M. & Pattit, J.M., 2014. Drivers of R&D investment: The interaction of behavioral theory and managerial incentives. *Journal of Business Research*.

Allen, S.E., Dokholyan, N. V. & Bowers, A.A., 2016. Dynamic Docking of Conformationally Constrained Macrocycles: Methods and Applications. *ACS Chemical Biology*, 11(1), pp.10–24. Available at: http://pubs.acs.org/doi/abs/10.1021/acschembio.5b00663 [Accessed March 15, 2017].

Amadei, A. et al., 1996. An efficient method for sampling the Essential subspace of proteins. *Journal of Biomolecular Structure and Dynamics*, 13(4).

Amadei, A., Linssen, A.B.M. & Berendsen, H.J.C., 1993. Essential dynamics of proteins. *Proteins: Structure, Function and Genetics*, 17(4), pp.412–425.

Ambrosio, G. et al., 2010. Chronic nitrate therapy is associated with different presentation and evolution of acute coronary syndromes: insights from 52 693 patients in the Global Registry of Acute Coronary Events. *European Heart Journal*, 31(4), pp.430–438. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19903682 [Accessed October 22, 2018].

Amit, G. et al., 2006. Intracoronary nitroprusside for the prevention of the no-reflow phenomenon after primary percutaneous coronary intervention in acute myocardial infarction. A randomized, double-blind, placebo-controlled clinical trial. *American Heart Journal*, 152(5), p.887.e9-887.e14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17070151 [Accessed October 22, 2018].

Ananthraj, V. et al., 2018. Towards Exascale Computing for High Energy Physics: The ATLAS Experience at ORNL. In *2018 IEEE 14th International Conference on e-Science (e-Science)*. IEEE, pp. 341–342. Available at: https://ieeexplore.ieee.org/document/8588705/ [Accessed February 7, 2019].

Anighoro, A., de la Vega de León, A. & Bajorath, J., 2016. Predicting bioactive conformations and binding modes of macrocycles. *Journal of Computer-Aided Molecular Design*, 30(10), pp.841–849. Available at: http://link.springer.com/10.1007/s10822-016-9973-5 [Accessed March 15, 2017].

Anighoro, A. & Rgen Bajorath, J., 2016. Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes. *Journal of Chemical Information and Modelling*, 56, pp.580–587.

Antunes, E.M. et al., 2005. Pyrroloiminoquinone and related metabolites from marine sponges. *Natural Product Reports*, 22(1), p.62. Available at: http://xlink.rsc.org/?DOI=b407299p [Accessed August 23, 2016].

Apol, E. et al., 2010. GROMACS Getting Started. *www.gromacs.org*.

Archer, J.R., 2004. History, evolution, and trends in compound management for high throughput screening. *Assay and drug development technologies*.

Asikainen, A.H., Ruuskanen, J. & Tuppurainen, K.A., 2004. Consensus kNN QSAR: a versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands.

Environmental science & technology, 38(24), pp.6724–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15669333 [Accessed October 1, 2018].

Bacheva, A. V. et al., 2003. Native and Modified Subtilisin 72 as a Catalyst for Peptide Synthesis in Media with a Low Water Content. *Russian Journal of Bioorganic Chemistry*, 29(5), pp.502–508. Available at: http://link.springer.com/10.1023/A:1026013912147 [Accessed October 11, 2018].

Ballante, F. et al., 2014. Hsp90 Inhibitors, Part 1: Definition of 3-D QSAutogrid/R Models as a Tool for Virtual Screening. *Journal of Chemical Information and Modeling*, 54(3), pp.956–969. Available at: http://pubs.acs.org/doi/10.1021/ci400759t [Accessed September 11, 2018].

Barberis, S. et al., 2006. Study of phytoproteases stability in aqueous-organic biphasic systems using linear free energy relationships. *Journal of Molecular Catalysis B: Enzymatic*, 38(2), pp.95–103. Available at: https://www.sciencedirect.com/science/article/pii/S1381117705002018?via%3Dihub [Accessed October 10, 2018].

Beckedahl, D. et al., 2016. On the configurational temperature Nose-Hoover thermostat. *Physica A: Statistical Mechanics and its Applications*.

Behrendt, R., White, P. & Offer, J., 2016. Advances in Fmoc solid-phase peptide synthesis. *Journal of Peptide Science*, 22(1), pp.4–27.

Berendsen, H.J.C., van der Spoel, D. & van Drunen, R., 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1), pp.43–56.

Bernardi, R.C., Melo, M.C.R. & Schulten, K., 2015. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et biophysica acta*, 1850(5), pp.872–877. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0304416514003559 [Accessed September 19, 2018].

Bickerton, G.R. et al., 2012. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2), pp.90–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22270643 [Accessed August 17, 2016].

Bikadi, Z. & Hazai, E., 2009. Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. *Journal of cheminformatics*, 1, p.15. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20150996 [Accessed September 6, 2018].

Blaney, J., 2012. A very short history of structure-based design: how did we get here and where do we need to go? *Journal of computer-aided molecular design*, 26(1), pp.13–4. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84857453737&partnerID=tZOtx3y1 [Accessed July 7, 2015].

Blundell, C.D., Packer, M.J. & Almond, A., 2013. Quantification of free ligand conformational preferences by NMR and their relationship to the bioactive conformation. *Bioorganic & Medicinal Chemistry*, 21(17), pp.4976–4987. Available at: http://www.sciencedirect.com/science/article/pii/S0968089613005968 [Accessed April

24, 2017].

Bockus, A.T., McEwen, C.M. & Lokey, R.S., 2013. Form and Function in Cyclic Peptide Natural Products: A Pharmacokinetic Perspective. *Current Topics in Medicinal Chemistry*, 13(7), pp.821–836. Available at: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1568-0266&volume=13&issue=7&spage=821 [Accessed September 14, 2016].

Bohm, H. & Stahl, M., 2002. The Use of Scoring Functions in Drug Discovery Applications. In B. Lipkowitz & D. B. Boyd, eds. *Reviews in Computational Chemistry*. Wiley & Sons.

Böhm, H.J., 1994. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of computer-aided molecular design*, 8(3), pp.243–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7964925 [Accessed September 10, 2018].

Bollini, M. et al., 2011. Computationally-Guided Optimization of a Docking Hit to Yield Catechol Diethers as Potent Anti-HIV Agents. *Journal of Medicinal Chemistry*, 54(24), pp.8582–8591. Available at: http://pubs.acs.org/doi/abs/10.1021/jm201134m [Accessed September 24, 2018].

Bonabeau, E., Bodick, N. & Armstrong, R.W., 2008. A More Rational Approach to New-Product Development. *Harvard Business Review*. Available at: www.hbr.org [Accessed June 9, 2016].

Bordusa, F., 2002. Proteases in Organic Synthesis†. *Chemical Reviews*, 102(12), pp.4817–4868. Available at: https://pubs.acs.org/doi/full/10.1021/cr010164d#_i3 [Accessed October 10, 2018].

Bosnor, K., 2000. How Interplanetary Internet Will Work. *HowStuffWorks*. Available at: https://computer.howstuffworks.com/interplanetary-internet.htm [Accessed February 7, 2019].

Boulanger, E. & Harvey, J.N., 2018. QM/MM methods for free energies and photochemistry. *Current Opinion in Structural Biology*.

Braga, C. & Travis, K.P., 2006. Configurational constant pressure molecular dynamics. *The Journal of Chemical Physics*, 124(10), p.104102. Available at: http://aip.scitation.org/doi/10.1063/1.2172601 [Accessed September 19, 2018].

Brahmachari, G., 2012. Natural Products in Drug Discovery: Impacts and Opportunities—An Assessment. *Bioactive Natural Products: Opportunities and Challenges in Medicinal Chemistry*.

Brahmkshatriya, P.S. et al., 2013. Quantum mechanical scoring: structural and energetic insights into cyclin-dependent kinase 2 inhibition by pyrazolo[1,5-a]pyrimidines. *Current computer-aided drug design*.

Bryce, R.A., 2011. Physics-based scoring of protein-ligand interactions: Explicit polarizability, quantum mechanics and free energies. *Future Medicinal Chemistry*.

Bürck, J. et al., 2016. Oriented Circular Dichroism: A Method to Characterize Membrane-Active Peptides in Oriented Lipid Bilayers. *Accounts of Chemical Research*.

Butler, M.S. & Buss, A.D., 2006. Natural products - The future scaffolds for novel antibiotics? *Biochemical Pharmacology*.

Camara, A. et al., 2011. *Acta Anaesthesiologica Croatica*, Unspecified. Available at: https://hrcak.srce.hr/70920 [Accessed October 22, 2018].

Campbell, A.J., Lamb, M.L. & Joseph-McCarthy, D., 2014. Ensemble-based docking using biased molecular dynamics. *Journal of chemical information and modeling*, 54(7), pp.2127–38. Available at: http://0-pubs.acs.org.wam.seals.ac.za/doi/full/10.1021/ci400729j [Accessed May 17, 2016].

Carter, P.H. et al., 2016. Investigating investment in biopharmaceutical R&D. *Nature Reviews Drug Discovery*.

Cereto-Massagué, A. et al., 2015. Molecular fingerprint similarity search in virtual screening. *Methods*, 71, pp.58–63. Available at: https://www.sciencedirect.com/science/article/pii/S1046202314002631 [Accessed September 29, 2018].

Chatterjee, J. et al., 2008. N-methylation of peptides: A new perspective in medicinal chemistry. *Accounts of Chemical Research*.

Chen, H.-M. et al., 2007. SODOCK: Swarm optimization for highly flexible protein–ligand docking. *Journal of Computational Chemistry*, 28(2), pp.612–623. Available at: http://doi.wiley.com/10.1002/jcc.20542 [Accessed September 17, 2018].

Chen, H. et al., 2012. A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. *Med. Chem. Commun.*, 3(3), pp.312–321. Available at: http://xlink.rsc.org/?DOI=C2MD00238H [Accessed August 17, 2016].

Chen, I.-J. & Foloppe, N., 2013. Tackling the conformational sampling of larger flexible compounds and macrocycles in pharmacology and drug discovery. *Bioorganic & Medicinal Chemistry*, 21(24), pp.7898–7920. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0968089613008596 [Accessed March 16, 2017].

Chen, J., Im, W. & Brooks, C.L., 2005. Application of torsion angle molecular dynamics for efficient sampling of protein conformations. *Journal of Computational Chemistry*, 26(15), pp.1565–1578. Available at: http://doi.wiley.com/10.1002/jcc.20293 [Accessed September 18, 2018].

Cheng, F. et al., 2012. Prediction of chemical-protein interactions: Multitarget-QSAR versus computational chemogenomic methods. *Molecular BioSystems*.

Cheng, T. et al., 2009. Comparative assessment of scoring functions on a diverse test set. *Journal of Chemical Information and Modeling*.

Cheng, T. et al., 2012. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *The AAPS Journal*.

Cheshire, D.R., 2011. How well do medicinal chemists learn from experience? *Drug discovery today*, 16(17–18), pp.817–21. Available at: http://www.sciencedirect.com/science/article/pii/S1359644611001838 [Accessed May

8, 2015].

Choi, H., Murray, T.F. & Aldrich, J. V., 2003. Synthesis and evaluation of derivatives of leucine enkephalin as potential affinity labels for δ opioid receptors. *Biopolymers*, 71(5), pp.552–557. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14635095 [Accessed October 12, 2018].

Clark, R.D., 2009. Prospective ligand- and target-based 3D QSAR: state of the art 2008. *Current topics in medicinal chemistry*, 9(9), pp.791–810. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19754395 [Accessed October 1, 2018].

Clemons, P.A. et al., 2010. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 107(44), pp.18787–92. Available at: http://0-www.pnas.org.wam.seals.ac.za/content/107/44/18787.full [Accessed September 22, 2015].

Clifford, C., 2018. Elon Musk defends plans to build a community on Mars after downbeat NASA report. *CNBC: Make It*, Entreprene. Available at: http://www.nature.com/articles/s41550-018-0529-6 [Accessed February 7, 2019].

Cook, D. et al., 2014. Lessons learned from the fate of AstraZeneca's drug pipeline: A five-dimensional framework. *Nature Reviews Drug Discovery*.

Cornea, R.L. et al., 2013. High-throughput FRET assay yields allosteric SERCA activators. *Journal of Biomolecular Screening*.

Cottrell, S.J. et al., 2004. Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. *Journal of computer-aided molecular design*, 18(11), pp.665–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15865060 [Accessed October 2, 2018].

Coutsias, E.A. et al., 2016. Exhaustive Conformational Sampling of Complex Fused Ring Macrocycles Using Inverse Kinematics. *Journal of chemical theory and computation*, 12, pp.4674–4687. Available at: http://pubs.acs.org/doi/pdf/10.1021/acs.jctc.6b00250 [Accessed March 29, 2017].

Craik, D.J. et al., 2013. The future of peptide-based drugs. *Chemical biology & drug design*, 81(1), pp.136–47. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23253135 [Accessed July 11, 2014].

Cramer, R.D., Patterson, D.E. & Bunce, J.D., 1988. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, 110(18), pp.5959–5967. Available at: http://pubs.acs.org/doi/abs/10.1021/ja00226a005 [Accessed October 2, 2018].

Creighton, J., 2015. Meet the Library of Babel: Every Possible Combination of Letters That has Been (or could be) Written. *Futurism*. Available at: https://futurism.com/meet-the-digital-library-of-babel-a-complete-combination-of-every-possible-combination-of-letters-ever/.

Cross, J.B. et al., 2009. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *Journal of Chemical Information and Modeling*.

Cross, S. et al., 2012. GRID-Based Three-Dimensional Pharmacophores I: FLAPpharm, a Novel Approach for Pharmacophore Elucidation. *Journal of Chemical Information and Modeling*, 52(10), pp.2587–2598. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22970894 [Accessed October 2, 2018].

Csermely, P. et al., 2013. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics*, 138(3), pp.333–408. Available at: http://www.sciencedirect.com/science/article/pii/S0163725813000284 [Accessed July 10, 2014].

Cushman, D.W. et al., 1977. Design of potent competitive inhibitors of angiotensin-converting enzyme. Carboxyalkanoyl and mercaptoalkanoyl amino acids. *Biochemistry*, 16(25), pp.5484–5491. Available at: http://pubs.acs.org/doi/abs/10.1021/bi00644a014 [Accessed October 22, 2018].

Dandapani, S. & Marcaurelle, L. a, 2010. Grand challenge commentary: Accessing new chemical space for "undruggable" targets. *Nature chemical biology*, 6(12), pp.861–863. Available at: http://www.nature.com/doifinder/10.1038/nchembio.479 [Accessed September 18, 2015].

Danzon, P.M., Nicholson, S. & Pereira, N.S., 2005. Productivity in pharmaceutical-biotechnology R&D: The role of experience and alliances. *Journal of Health Economics*.

Darve, E. & Pohorille, A., 2001. Calculating free energies using average force. *The Journal of Chemical Physics*, 115(20), pp.9169–9183. Available at: http://aip.scitation.org/doi/10.1063/1.1410978 [Accessed September 21, 2018].

Deng, W., Breneman, C. & Embrechts, M.J., 2004. Predicting Protein−Ligand Binding Affinities Using Novel Geometrical Descriptors and Machine-Learning Methods. *J. Chem. Inf. Comput. Sci.*, 44, pp.699–703. Available at: https://pubs.acs.org/doi/abs/10.1021/ci034246+?source=chemport [Accessed September 10, 2018].

Desai, N.R. & Sabatine, M.S., 2015. PCSK9 inhibition in patients with hypercholesterolemia. *Trends in Cardiovascular Medicine*, 25(7), pp.567–574. Available at: https://www.sciencedirect.com/science/article/pii/S1050173815000304?via%3Dihub [Accessed October 22, 2018].

Deslouches, B. et al., 2005. De novo generation of cationic antimicrobial peptides: Influence of length and tryptophan substitution on antimicrobial activity. *Antimicrobial Agents and Chemotherapy*.

Dias, D.M. & Ciulli, A., 2014. NMR approaches in structure-based lead discovery: recent developments and new frontiers for targeting multi-protein complexes. *Progress in biophysics and molecular biology*, 116(2–3), pp.101–12. Available at: http://www.sciencedirect.com/science/article/pii/S007961071400087X [Accessed May 18, 2015].

Díaz-Eufracio, B.I. et al., 2018. Exploring the chemical space of peptides for drug discovery: a focus on linear and cyclic penta-peptides. *Molecular Diversity*, 22(2), pp.259–267. Available at: http://link.springer.com/10.1007/s11030-018-9812-9 [Accessed October

23, 2018].

Diller, D.J. & Merz, K.M., 2002. Can we separate active from inactive conformations? *Journal of Computer-Aided Molecular Design*, 16, pp.105–112. Available at: http://download.springer.com/static/pdf/897/art%253A10.1023%252FA%253A1016320106741.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Farticle%2F10.1023%2FA%3A1016320106741&token2=exp=1493057557~acl=%2Fstatic%2Fpdf%2F897%2Fart%25253A10.1023%25252FA%25253A1016 [Accessed April 24, 2017].

DiMasi, J.A. & Grabowski, H.G., 2007. The cost of biopharmaceutical R&D: is biotech different? *Managerial and Decision Economics*, 28(4–5), pp.469–479. Available at: http://doi.wiley.com/10.1002/mde.1360 [Accessed October 24, 2018].

Ding, Y. et al., 2015. GeauxDock: A novel approach for mixed-resolution ligand docking using a descriptor-based force field. *Journal of Computational Chemistry*.

Dixon, S.L. et al., 2006. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *Journal of Computer-Aided Molecular Design*, 20(10–11), pp.647–671. Available at: http://link.springer.com/10.1007/s10822-006-9087-6 [Accessed October 2, 2018].

Doak, B.C. et al., 2015. How Beyond Rule of 5 Drugs and Clinical Candidates Bind to Their Targets. *Journal of medicinal chemistry*. Available at: http://0-pubs.acs.org.wam.seals.ac.za/doi/abs/10.1021/acs.jmedchem.5b01286 [Accessed March 21, 2016].

Doak, B.C. et al., 2014. Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chemistry & Biology*, 21(9), pp.1115–1142. Available at: https://www.sciencedirect.com/science/article/pii/S1074552114002890?via%3Dihub [Accessed October 23, 2018].

Dobson, C.M., 2004. Chemical space and biology. *Nature*, 432(7019), pp.824–8. Available at: http://dx.doi.org/10.1038/nature03192 [Accessed March 25, 2015].

Dominguez, C., Boelens, R. & Bonvin, A.M.J.J., 2003. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society*, 125(7), pp.1731–1737. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12580598 [Accessed September 14, 2018].

Doudou, S., Burton, N.A. & Henchman, R.H., 2009. Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force. *Journal of Chemical Theory and Computation*, 5(4), pp.909–918. Available at: http://pubs.acs.org/doi/abs/10.1021/ct8002354 [Accessed September 20, 2018].

Dowty, M.E. et al., 2014. Preclinical to clinical translation of tofacitinib, a Janus kinase inhibitor, in rheumatoid arthritis. *The Journal of pharmacology and experimental therapeutics*, 348(1), pp.165–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24218541 [Accessed September 25, 2018].

Van Drie, J.H., 2007. Monty Kier and the Origin of the Pharmacophore Concept. *Internet Electronic Journal of Molecular Design*, 6(9), pp.271–279. Available at: http://www.biochempress.comhttp//www.biochempress.com.http://www.biochempress.com [Accessed October 1, 2018].

Dube, H. et al., 2012. A mitochondrial-targeted cyclosporin A with high binding affinity for cyclophilin D yields improved cytoprotection of cardiomyocytes. *Biochem. J*, 441, pp.901–907. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3260541/pdf/bj4410901.pdf [Accessed March 9, 2017].

Duffy, F.J. et al., 2011. CycloPs: Generating Virtual Libraries of Cyclized and Constrained Peptides Including Nonnatural Amino Acids. *Journal of Chemical Information and Modeling*, 51(4), pp.829–836. Available at: http://pubs.acs.org/doi/abs/10.1021/ci100431r [Accessed August 7, 2017].

Durrant, J.D. & McCammon, J.A., 2011. Molecular dynamics simulations and drug discovery. *BMC biology*, 9(1), p.71. Available at: http://www.biomedcentral.com/1741-7007/9/71 [Accessed February 16, 2015].

Edwards, P.A. & Ericsson, J., 1998. PubChem Open Chemistry Database. *Current opinion in lipidology*.

Eken, Y. et al., 2018. SAMPL6 host–guest challenge: binding free energies via a multistep approach. *Journal of Computer-Aided Molecular Design*, pp.1–19. Available at: http://link.springer.com/10.1007/s10822-018-0159-1 [Accessed September 22, 2018].

El-Faham, A. & Albericio, F., 2011. Peptide Coupling Reagents, More than a Letter Soup. *Chemical Reviews*, 111(11), pp.6557–6602. Available at: http://pubs.acs.org/doi/abs/10.1021/cr100048w [Accessed October 13, 2018].

El-Menyar, A. et al., 2011. Atypical presentation of acute coronary syndrome: A significant independent predictor of in-hospital mortality. *Journal of Cardiology*, 57(2), pp.165–171. Available at: https://www.sciencedirect.com/science/article/pii/S0914508710002431 [Accessed October 23, 2018].

Eldridge, M.D. et al., 1997. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design*.

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, pp.3714–3717.

Fang, Y. et al., 2016. GeauxDock: Accelerating Structure-Based Virtual Screening with Heterogeneous Computing A. G. de Brevern, ed. *PLOS ONE*, 11(7). Available at: http://dx.plos.org/10.1371/journal.pone.0158898 [Accessed March 3, 2017].

FDA, 2018. The Drug Development Process. *US Department of Health and Human Services*. Available at: https://www.fda.gov/ForPatients/Approvals/Drugs/ucm405382.htm [Accessed July 8, 2018].

Feher, M. & Schmidt, J.M., 2003. Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*.

Fields, G.B. & Noble, R.L., 1990. Solid phase peptide synthesis utilizing 9-

fluorenylmethoxycarbonyl amino acids. *International Journal of Peptide and Protein Research*.

Fjell, C.D. et al., 2012. Designing antimicrobial peptides: Form follows function. *Nature Reviews Drug Discovery*.

Fong, P. et al., 2009. Assessment of QM/MM Scoring Functions for Molecular Docking to HIV-1 Protease. *Journal of Chemical Information and Modeling*, 49(4), pp.913–924. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19309119 [Accessed September 6, 2018].

Forli, S. & Botta, M., 2007. Lennard-Jones Potential and Dummy Atom Settings to Overcome the AUTODOCK Limitation in Treating Flexible Ring Systems. *Journal of Chemical Informatics*, 47(4), pp.1481–1492. Available at: http://pubs.acs.org/doi/full/10.1021/ci700036j [Accessed March 15, 2017].

Fosgerau, K. & Hoffmann, T., 2015. Peptide therapeutics: Current status and future directions. *Drug Discovery Today*.

Fox, K.A.A. et al., 2010. Underestimated and under-recognized: the late consequences of acute coronary syndrome (GRACE UK-Belgian Study). *European Heart Journal*, 31(22), pp.2755–2764. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20805110 [Accessed October 22, 2018].

Frecer, V., Ho, B. & Ding, J., 2004. De novo design of potent antimicrobial peptides. *Antimicrobial agents and chemotherapy*.

Freeland, R.G. et al., 1979. The Chemical Abstracts Service Chemical Registry System. II. Augmented connectivity molecular formula. *Journal of Chemical Information and Computer Sciences*.

Friesner, R.A. et al., 2004. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*.

Fu, Y. et al., 2015. A New Approach for Flexible Molecular Docking Based on Swarm Intelligence. *Mathematical Problems in Engineering*, 2015, pp.1–10. Available at: http://www.hindawi.com/journals/mpe/2015/540186/ [Accessed September 11, 2018].

Fuster, V. et al., 2005. Atherothrombosis and High-Risk Plaque. *Journal of the American College of Cardiology*, 46(6), pp.937–954. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0735109705013963 [Accessed October 22, 2018].

Gabel, J., Desaphy, J. & Rognan, D., 2014. Beware of machine learning-based scoring functions-on the danger of developing black boxes. *Journal of Chemical Information and Modeling*.

General, I.J., 2010. A Note on the Standard State's Binding Free Energy. *Journal of Chemical Theory and Computation*, 6(8), pp.2520–2524. Available at: http://pubs.acs.org/doi/abs/10.1021/ct100255z [Accessed September 22, 2018].

Genheden, S. & Ryde, U., 2015. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*.

Ghasemi, F. et al., 2018. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. *Drug Discovery Today*, 23(10), pp.1784–1790. Available at: https://www.sciencedirect.com/science/article/pii/S1359644617304762#bib0030 [Accessed October 1, 2018].

Gho, B.C. et al., 1996. Myocardial protection by brief ischemia in noncardiac tissue. *Circulation*, 94(9), pp.2193–200. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8901671 [Accessed October 22, 2018].

Gibbs, J.W., 1902. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*, Cambridge, USA: C. Scribner's sons. Available at: https://archive.org/details/elementaryprinc00gibbgoog [Accessed September 19, 2018].

Gilad, Y., Nadassy, K. & Senderowitz, H., 2015. A reliable computational workflow for the selection of optimal screening libraries. *Journal of Cheminformatics*.

Gohlke, H., Hendlich, M. & Klebe, G., 2000. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology*.

de Gómez-Puyou, M.T. & Gómez-Puyou, A., 1998. Enzymes in Low Water Systems. *Critical Reviews in Biochemistry and Molecular Biology*, 33(1), pp.53–89. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9543628 [Accessed October 10, 2018].

Grabowski, H., Vernon, J. & DiMasi, J.A., 2002. Returns on Research and Development for 1990s New Drug Introductions. *PharmacoEconomics*.

Gräslund, S. et al., 2008. Protein production and purification. *Nature Methods*, 5(2), pp.135–146. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18235434 [Accessed October 12, 2018].

Gray, J.J. et al., 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*.

Greene, J. et al., 1994. Chemical Function Queries for 3D Database Search. *Journal of Chemical Information and Modeling*, 34(6), pp.1297–1308. Available at: http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00022a012 [Accessed October 2, 2018].

Greenidge, P.A. et al., 2014. Improving docking results via reranking of ensembles of ligand poses in multiple X-ray protein conformations with MM-GBSA. *Journal of Chemical Information and Modeling*.

Grimme, S., 2004. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry*, 25(12), pp.1463–1473. Available at: http://doi.wiley.com/10.1002/jcc.20078 [Accessed September 6, 2018].

Gromiha, M.M., Yugandhar, K. & Jemimah, S., 2017. Protein–protein interactions: scoring schemes and binding affinity. *Current Opinion in Structural Biology*.

Gütlein, M., Karwath, A. & Kramer, S., 2014. CheS-Mapper 2.0 for visual validation of (Q)SAR models. *Journal of Cheminformatics*, 6(1), p.41. Available at: http://www.jcheminf.com/content/6/1/41 [Accessed May 27, 2017].

Guzmán, F., Barberis, S. & Illanes, A., 2007. Peptide synthesis: Chemical or enzymatic. *Electronic Journal of Biotechnology*.

Halls, M. et al., 2013. Virtual screening of electron acceptor materials for organic photovoltaic applications. *New Journal of Physics*, 15(10), p.105029. Available at: http://stacks.iop.org/1367-2630/15/i=10/a=105029?key=crossref.8e0d9411a08d34dd037ca7c1936e7f30 [Accessed October 30, 2018].

Han, J. et al., 2013. Design, Synthesis, and Biological Activity of Novel Dicoumarol Glucagon-like Peptide 1 Conjugates. *Journal of Medicinal Chemistry*, 56(24), pp.9955–9968. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24308627 [Accessed October 12, 2018].

Hann, M.M. & Oprea, T.I., 2004. Pursuing the leadlikeness concept in pharmaceutical research. *Current Opinion in Chemical Biology*, 8, pp.255–263.

Hansch, C., 1969. Quantitative approach to biochemical structure-activity relationships. *Accounts of Chemical Research*, 2(8), pp.232–239. Available at: https://pubs.acs.org/sharingguidelines [Accessed September 29, 2018].

Harvey, A.L., 2008. Natural products in drug discovery. *Drug Discovery Today*.

Harvey, A.L., Edrada-ebel, R. & Quinn, R.J., 2015. The re-emergence of natural products for drug discovery in the genomics era. *Nature Publishing Group*, 14(2), pp.111–129. Available at: http://dx.doi.org/10.1038/nrd4510.

Hausenloy, D., 2013. Cardioprotection Techniques: Preconditioning, Postconditioning and Remote Con-ditioning (Basic Science). *Current Pharmaceutical Design*, 19(25), pp.4544–4563. Available at: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1381-6128&volume=19&issue=25&spage=4544 [Accessed October 22, 2018].

Hay, M. et al., 2014. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1), pp.40–51. Available at: http://www.nature.com/doifinder/10.1038/nbt.2786 [Accessed June 9, 2016].

He, S. et al., 2016. Discovery, Optimization, and Characterization of Novel Chlorcyclizine Derivatives for the Treatment of Hepatitis C Virus Infection. *Journal of medicinal chemistry*. Available at: http://0-pubs.acs.org.wam.seals.ac.za/doi/abs/10.1021/acs.jmedchem.5b00752 [Accessed February 3, 2016].

Heffernan, R. et al., 2015. Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*, 5(1), p.11476. Available at: http://www.nature.com/articles/srep11476 [Accessed November 16, 2018].

Hermans, J. & Wang, L., 1997. Inclusion of Loss of Translational and Rotational Freedom in Theoretical Estimates of Free Energies of Binding. Application to a Complex of Benzene and Mutant T4 Lysozyme. Available at: https://0-pubs.acs.org.wam.seals.ac.za/doi/abs/10.1021/ja963568+ [Accessed September 22, 2018].

Hetzel, G.R. & Sucker, C., 2005. The heparins: all a nephrologist should know. *Nephrology Dialysis Transplantation*, 20(10), pp.2036–2042. Available at: http://academic.oup.com/ndt/article/20/10/2036/1934691/The-heparins-all-a-nephrologist-should-know [Accessed October 22, 2018].

Heusler, K. & Pletscher, A., 2001. The controversial early history of cyclosporin.

Hewitt, W.M. et al., 2015. Cell-Permeable Cyclic Peptides from Synthetic Libraries Inspired by Natural Products. *Journal of the American Chemical Society*, 137(2), pp.715–721. Available at: http://pubs.acs.org/doi/abs/10.1021/ja508766b [Accessed November 8, 2016].

Hiss, J.A. et al., 2007. Design of MHC I stabilizing peptides by agent-based exploration of sequence space. *Protein Engineering, Design and Selection*.

Honig, B., Sharp, K. & Yang, A.S., 1993. Macroscopic models of aqueous solutions: Biological and chemical applications. *Journal of Physical Chemistry*.

Hopkins, A.L., 2008. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11), pp.682–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18936753 [Accessed July 11, 2014].

Hopkins, A.L. et al., 2014. The role of ligand efficiency metrics in drug discovery. *Nature Reviews Drug Discovery*, 13(2), pp.105–121. Available at: http://www.nature.com/doifinder/10.1038/nrd4163 [Accessed June 10, 2016].

Hou, T.J. & Xu, X.J., 2003. ADME Evaluation in Drug Discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors. *Journal of Chemical Information and Computer Sciences*, 43, pp.2137–2152.

Hou, W. et al., 2011. Peptidyl *N* , *N* -Bis(2-mercaptoethyl)-amides as Thioester Precursors for Native Chemical Ligation †. *Organic Letters*, 13(3), pp.386–389. Available at: http://pubs.acs.org/doi/abs/10.1021/ol102735k [Accessed October 13, 2018].

Hu, B. & Lill, M.A., 2012. Protein Pharmacophore Selection Using Hydration-Site Analysis. *Journal of Chemical Information and Modeling*, 52(4), pp.1046–1060. Available at: http://pubs.acs.org/doi/10.1021/ci200620h [Accessed October 2, 2018].

Hu, Y. & Bajorath, J., 2012. Exploration of 3D Activity Cliffs on the Basis of Compound Binding Modes and Comparison of 2D and 3D Cliffs. *Journal of Chemical Information and Modeling*, 52(3), pp.670–677. Available at: http://pubs.acs.org/doi/10.1021/ci300033e [Accessed October 2, 2018].

Huang, H.J. et al., 2010. Current developments of computer-aided drug design. *Journal of the Taiwan Institute of Chemical Engineers*.

Huang, S.Y., Grinter, S.Z. & Zou, X., 2010. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Physical Chemistry Chemical Physics*.

Huey, R. et al., 2007. A semiempirical free energy force field with charge-based desolvation. *Journal of computational chemistry*, 28(6), pp.1145–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17274016 [Accessed January 18, 2015].

Hughes, J. et al., 2011. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6), pp.1239–1249. Available at: http://doi.wiley.com/10.1111/j.1476-5381.2010.01127.x [Accessed October 24, 2018].

Illingworth, C.J.R. et al., 2008. Assessing the Role of Polarization in Docking. *The Journal of Physical Chemistry A*, 112(47), pp.12157–12163. Available at: http://pubs.acs.org/doi/abs/10.1021/jp710169m [Accessed September 6, 2018].

Irwin, J.J. & Shoichet, B.K., 2005. ZINC - A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*.

Jaruzelski, B., Schwartz, K. & Staack, V., 2015. The 2015 Global Innovation 1000: Innovation's new world order (Study report). *PwC Strategy&*. Available at: https://www.strategyand.pwc.com/reports/2015-global-innovation-1000-media-report.

Jeedimalla, N. et al., 2015. Multicomponent assembly of 4-aza-podophyllotoxins: A fast entry to highly selective and potent anti-leukemic agents. *European Journal of Medicinal Chemistry*, 106, pp.167–179. Available at: http://dx.doi.org/10.1016/j.ejmech.2015.10.009 [Accessed October 30, 2018].

Ji, C.G. & Zhang, J.Z.H., 2008. Protein Polarization Is Critical to Stabilizing AF-2 and Helix-2' Domains in Ligand Binding to PPAR-γ. *Journal of the American Chemical Society*, 130(50), pp.17129–17133. Available at: http://pubs.acs.org/doi/abs/10.1021/ja807374x [Accessed September 6, 2018].

Jiang, W. & Roux, B., 2010. Free energy perturbation Hamiltonian replica-exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations. *Journal of Chemical Theory and Computation*.

Jiang, Z. et al., 2011. Rational Design of α-Helical Antimicrobial Peptides to Target Gram-negative Pathogens, Acinetobacter baumannii and Pseudomonas aeruginosa: Utilization of Charge, "Specificity Determinants," Total Hydrophobicity, Hydrophobe Type and Location as Design Parameters to Improve the Therapeutic Ratio. *Chemical Biology and Drug Design*.

Johnson, A.A. et al., 2001. Toxicity of antiviral nucleoside analogs and the human mitochondrial DNA polymerase. *The Journal of biological chemistry*, 276(44), pp.40847–40857. Available at: http://www.jbc.org/content/276/44/40847.long [Accessed July 16, 2014].

Johnson, D. & Brooker, W., 2005. Star Wars Fans, DVD, and Cultural Ownership: An Interview with Will Brooker. *The Velvet Light Trap*, 56(1), pp.36–44. Available at: http://muse.jhu.edu/content/crossref/journals/the_velvet_light_trap/v056/56.1johnson.html [Accessed February 7, 2019].

Johnson, T.W., Dress, K.R. & Edwards, M., 2009. Using the Golden Triangle to optimize clearance and oral absorption. *Bioorganic & Medicinal Chemistry Letters*, 19(19), pp.5560–5564. Available at: https://www.lundbeck.com/upload/drughunters/2014/pdf/Golden_triangle_BMCL_2009.pdf [Accessed June 10, 2016].

Jones, G. et al., 1997. Development and validation of a genetic algorithm for flexible

docking. *Journal of Molecular Biology*.

Jorgensen, W.L. & Duffy, E.M., 2000. Prediction of drug solubility from Monte Carlo simulations. *Bioorganic and Medicinal Chemistry Letters*.

Jorgensen, W.L. & Duffy, E.M., 2002. Prediction of drug solubility from structure. *Advanced Drug Delivery Reviews*.

Jorgensen, W.L. & Thomas, L.L., 2008. Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *Journal of Chemical Theory and Computation*, 4(6), pp.869–876. Available at: http://pubs.acs.org/doi/abs/10.1021/ct800011m [Accessed September 22, 2018].

Joseph-McCarthy, D. & Alvarez, J.C., 2003. Automated generation of MCSS-derived pharmacophoric DOCK site points for searching multiconformation databases. *Proteins: Structure, Function, and Genetics*, 51(2), pp.189–202. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12660988 [Accessed October 2, 2018].

Joseph, R., 2010. Marketing Mars: Financing the Human Mission to Mars and the Colonization of the Red Planet. *Journal of Cosmology*, 12, pp.4068–4080.

Josephson, K., Hartman, M.C.T. & Szostak, J.W., 2005. Ribosomal Synthesis of Unnatural Peptides. *Journal of the American Chemical Society*, 127(33), pp.11727–11735. Available at: http://pubs.acs.org/doi/abs/10.1021/ja0515809 [Accessed October 12, 2018].

Kajitani, K. et al., 2008. Crystal structure of human cyclophilin D in complex with its inhibitor, cyclosporin A at 0.96-Å resolution. *Proteins: Structure, Function, and Bioinformatics*, 70(4), pp.1635–1639. Available at: http://onlinelibrary.wiley.com/doi/10.1002/prot.21855/full [Accessed November 20, 2017].

Kapell, M.W. & Lawrence, J.S., 2006. *Finding the Force of the Star Wars Franchise*, New York: Die Deutsche Bibliothek.

Kapoor, P. et al., 2012. Tumorhope: A database of tumor homing peptides. *PLoS ONE*.

Karcz, J.S. et al., 2012. Red Dragon: Low Cost Access to the surface of Mars using commercial capabilities. In *Concepts and Approaches for Mars Exploration*. Houston: Lunar and Planetary Institution. Available at: https://ntrs.nasa.gov/search.jsp?R=20120013431 [Accessed February 7, 2019].

Kasahara, K., Sakuraba, S. & Fukuda, I., 2018. Enhanced Sampling of Molecular Dynamics Simulations of a Polyalanine Octapeptide: Effects of the Periodic Boundary Conditions on Peptide Conformation. *The Journal of Physical Chemistry B*, 122(9), pp.2495–2503. Available at: http://pubs.acs.org/doi/10.1021/acs.jpcb.7b10830 [Accessed September 19, 2018].

Kassler-Taub, K. et al., 1998. Comparative Efficacy of Two Angiotensin II Receptor Antagonists, Irbesartan and Losartan, in Mild-to-Moderate Hypertension. *American Journal of Hypertension*, 11(4), pp.445–453. Available at: https://academic.oup.com/ajh/article-lookup/doi/10.1016/S0895-7061(97)00491-3 [Accessed October 22, 2018].

Katz, L. & Baltz, R.H., 2016. Natural product discovery: past, present, and future. *Journal of Industrial Microbiology {&} Biotechnology*.

Kernis, S.J. et al., 2004. Does beta-blocker therapy improve clinical outcomes of acute myocardial infarction after successful primary angioplasty? *Journal of the American College of Cardiology*, 43(10), pp.1773–1779. Available at: https://www.sciencedirect.com/science/article/pii/S0735109704004668?via%3Dihub [Accessed October 22, 2018].

Kessler, H. et al., 1996. Conformation of cyclic peptides. Principle concepts and the design of selectivity and superactivity in bioactive sequences by "spatial screening." *Pure & Appl. Chem*, 68(6), pp.1201–1205.

Khajehsharifi, H. et al., 2017. The comparison of partial least squares and principal component regression in simultaneous spectrophotometric determination of ascorbic acid, dopamine and uric acid in real samples. *Arabian Journal of Chemistry*, 10, pp.S3451–S3458. Available at: https://www.sciencedirect.com/science/article/pii/S1878535214000343 [Accessed October 1, 2018].

Khanna, I., 2012. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discovery Today*, 17(19–20), pp.1088–1102. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1359644612001833 [Accessed June 9, 2016].

Kinnings, S.L. et al., 2011. A Machine Learning-Based Method To Improve Docking Scoring Functions and Its Application to Drug Repurposing. *Journal of Chemical Information and Modeling*, 51(2), pp.408–419. Available at: http://pubs.acs.org/doi/abs/10.1021/ci100369f [Accessed September 10, 2018].

Kise, H. & Hayakawa, A., 1991. Immobilization of proteases to porous chitosan beads and their catalysis for ester and peptide synthesis in organic solvents. *Enzyme and Microbial Technology*, 13(7), pp.584–588. Available at: https://www.sciencedirect.com/science/article/pii/014102299190094Q [Accessed October 11, 2018].

Kitakaze, M. et al., 2007. Human atrial natriuretic peptide and nicorandil as adjuncts to reperfusion treatment for acute myocardial infarction (J-WIND): two randomised trials. *The Lancet*, 370(9597), pp.1483–1493. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17964349 [Accessed October 22, 2018].

Klebe, G., Abraham, U. & Mietzner, T., 1994. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *Journal of Medicinal Chemistry*, 37(24), pp.4130–4146. Available at: http://pubs.acs.org/doi/abs/10.1021/jm00050a010 [Accessed October 1, 2018].

Klein, C.D. & Hopfinger, A.J., 1998. Pharmacological activity and membrane interactions of antiarrhythmics: 4D-QSAR/QSPR analysis. *Pharmaceutical research*, 15(2), pp.303–11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9523319 [Accessed October 1, 2018].

Klepeis, J.L. et al., 2004. Design of Peptide Analogues with Improved Activity Using a Novel

de Novo Protein Design Approach. *Industrial & Engineering Chemistry Research*.

Koivisto, J.J., Kumpulainen, E.T.T. & Koskinen, A.M.P., 2010. Conformational ensembles of flexible b-turn mimetics in DMSO-d 6. *Organic & biomolecular chemistry*, 8, pp.2103–2116. Available at: http://pubs.rsc.org/en/content/articlepdf/2010/ob/b921794k [Accessed April 24, 2017].

Kola, I. & Landis, J., 2004. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8), pp.711–716. Available at: http://www.nature.com/doifinder/10.1038/nrd1470 [Accessed June 9, 2016].

Kolmer, A. et al., 2015. Conformational analysis of small organic molecules using NOE and RDC data: A discussion of strychnine and α-methylene-γ-butyrolactone. *Journal of Magnetic Resonance*, 261, pp.101–109. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1090780715002402 [Accessed March 8, 2017].

Kolossvary, I.. & Keseru, G.M., 2001. Hessian-free low-mode conformational search for large-scale protein loop optimization: application to c-jun N-terminal kinase JNK3. *Journal of Computational Chemistry*, 22(1), pp.21–30. Available at: http://doi.wiley.com/10.1002/1096-987X%2820010115%2922%3A1%3C21%3A%3AAID-JCC3%3E3.0.CO%3B2-I [Accessed April 24, 2017].

Korb, O., Stützle, T. & Exner, T.E., 2006. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. In Springer, Berlin, Heidelberg, pp. 247–258. Available at: http://link.springer.com/10.1007/11839088_22 [Accessed September 17, 2018].

Kortagere, S. & Ekins, S., 2010. Troubleshooting computational methods in drug discovery. *Journal of Pharmacological and Toxicological Methods*, 61(2), pp.67–75. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1056871910000213.

Kraft, R., 2015. NASA, ESA Use Experimental Interplanetary Internet to Test Robot From International Space Station. *NASA Media Release*. Available at: https://www.nasa.gov/home/hqnews/2012/nov/HQ_12-391_DTN.html [Accessed February 7, 2019].

Kuhn, B. et al., 2016. A Real-World Perspective on Molecular Design. *Journal of medicinal chemistry*, 59, pp.4087–4102. Available at: http://0-pubs.acs.org.wam.seals.ac.za/doi/abs/10.1021/acs.jmedchem.5b01875 [Accessed February 21, 2016].

Kulkarni, S.S. et al., 2018a. Rapid and efficient protein synthesis through expansion of the native chemical ligation concept. *Nature Reviews Chemistry*.

Kulkarni, S.S. et al., 2018b. Rapid and efficient protein synthesis through expansion of the native chemical ligation concept. *Nature Reviews Chemistry*, 2(4), p.0122. Available at: http://www.nature.com/articles/s41570-018-0122 [Accessed October 13, 2018].

Kumar, D. & Bhalla, T.C., 2005. Microbial proteases in peptide synthesis: approaches and applications. *Applied Microbiology and Biotechnology*, 68(6), pp.726–736. Available at: http://link.springer.com/10.1007/s00253-005-0094-7 [Accessed October 11, 2018].

Kumar, S. et al., 1992. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *Journal of Computational Chemistry*, 13(8), pp.1011–1021. Available at: http://quantum.ch.ntu.edu.tw/ycclab/wp-content/uploads/2015/02/WHAM_1992.pdf.

Kumar, V., Krishna, S. & Siddiqi, M.I., 2015. Virtual screening strategies: Recent advances in the identification and design of anti-cancer agents. *Methods*, 71, pp.64–70. Available at: https://www.sciencedirect.com/science/article/pii/S1046202314002680 [Accessed September 29, 2018].

Kumari, R., Kumar, R. & Lynn, A., 2014. g_mmpbsa-A GROMACS Tool for High-Throughput MM-PBSA Calculations. *Journal of chemical information and modelling*, 54, pp.1951–1962. Available at: http://pubs.acs.org/doi/pdf/10.1021/ci500020m [Accessed August 14, 2017].

Kwong, J.Q. & Molkentin, J.D., 2015. Physiological and Pathological Roles of the Mitochondrial Permeability Transition Pore in the Heart. *Cell Metabolism*, 21(2), pp.206–214. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25651175 [Accessed April 24, 2017].

Labute, P., 2010. LowModeMD - Implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops. *Journal of Chemical Information and Modeling*, 50(5), pp.792–800. Available at: http://0-pubs.acs.org.wam.seals.ac.za/doi/pdf/10.1021/ci900508k [Accessed March 16, 2017].

Lachance, H. et al., 2012. Charting, navigating, and populating natural product chemical space for drug discovery. *Journal of Medicinal Chemistry*.

Lagorce, D. et al., 2015. FAF-Drugs3: a web server for compound property calculation and chemical library design. *Nucleic Acids Research*, 43(W1). Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv353 [Accessed September 8, 2016].

Lagorce, D. et al., 2011. The FAF-Drugs server: A multi-step engine to prepare electronic chemical compound collections.

Laio, A. & Parrinello, M., 2002. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20), pp.12562–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12271136 [Accessed September 20, 2018].

Lam, K.S. et al., 1991. A new type of synthetic peptide library for identifying ligand-binding activity. *Nature*.

Langer, T., Hoffmann, R.D. & Wiley InterScience (Online service), 2006. *Pharmacophores and pharmacophore searches*, Wiley-VCH. Available at: https://books.google.co.za/books?hl=en&lr=&id=LUpFcStB80cC&oi=fnd&pg=PA3&dq=%22Wermuth%22++Pharmacophores+and+pharmacophore+searches&ots=3Qujy1egmL&sig=xDvL-mj8NzAoPKfYwKIgifFnoK0&redir_esc=y#v=onepage&q=%22Wermuth%22 Pharmacophores and pharmacophore searches&f=false [Accessed September 29, 2018].

Lau, J.L. & Dunn, M.K., 2017. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic and Medicinal Chemistry*.

Law, M.R., Wald, N.J. & Rudnicka, A.R., 2003. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *BMJ (Clinical research ed.)*, 326(7404), p.1423. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12829554 [Accessed October 22, 2018].

Law, V. et al., 2014. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(Database issue), pp.D1091-7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965102&tool=pmcentrez&rendertype=abstract [Accessed April 1, 2015].

Leach, A.R. et al., 2010. Three-Dimensional Pharmacophore Methods in Drug Discovery. *Journal of Medicinal Chemistry*, 53(2), pp.539–558. Available at: http://pubs.acs.org/doi/abs/10.1021/jm900817u [Accessed September 29, 2018].

Leach, A.R. & Hann, M.M., 2011. Molecular complexity and fragment-based drug discovery: ten years on. *Current Opinion in Chemical Biology*, 15(4), pp.489–496.

Leening, M.J.G., Cook, N.R. & Ridker, P.M., 2016. Should we reconsider the role of age in treatment allocation for primary prevention of cardiovascular disease? *European Heart Journal*, 38(20), p.ehw287. Available at: http://eurheartj.oxfordjournals.org/lookup/doi/10.1093/eurheartj/ehw287 [Accessed October 22, 2018].

Leeson, P., 2012. Drug discovery: Chemical beauty contest. *Nature*, 481(7382), pp.455–6. Available at: http://www.nature.com/nature/journal/v481/n7382/full/481455a.html#ref2 [Accessed February 24, 2015].

Lei, H. & Duan, Y., 2007. Improved sampling methods for molecular simulation. *Current Opinion in Structural Biology*, 17(2), pp.187–191.

Lei, M. et al., 2004. Sampling protein conformations and pathways. *Journal of Computational Chemistry*, 25(9), pp.1133–1148. Available at: http://doi.wiley.com/10.1002/jcc.20041 [Accessed April 24, 2017].

Leiner, B. et al., 2003. "Origins of the Internet" in A Brief History of the Internet. *The Internet Society*. Available at: https://www.internetsociety.org/internet/history-internet/brief-history-internet/#Origins.

Lemke, T. et al., 2013. *Foye's Principles of Medicinal Chemistry* 7th ed., Philadelphia: Lippincott Williams and Wilkins.

Lemkul, J.A. & Bevan, D.R., 2010. Assessing the Stability of Alzheimer's Amyloid Protofibrils Using Molecular Dynamics. *Journal of Physical Chemistry B*, 114, pp.1652–1660. Available at: http://pubs.acs.org/doi/pdf/10.1021/jp9110794 [Accessed September 11, 2017].

Lendrem, D.W. & Lendrem, B.C., 2014. The development speed paradox: Can increasing development speed reduce R&D productivity? *Drug Discovery Today*.

Lensink, M.F. & Wodak, S.J., 2013. Docking, scoring, and affinity prediction in CAPRI. *Proteins: Structure, Function, and Bioinformatics*, 81(12), pp.2082–2095. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24115211 [Accessed September 15, 2018].

de Lera, A.R. & Ganesan, A., 2016. Epigenetic polypharmacology: from combination therapy to multitargeted drugs. *Clinical Epigenetics*.

Levitt, M. & Warshel, A., 1975. Computer simulation of protein folding. *Nature*.

Li, G. et al., 2008. Protease-Catalyzed Co-Oligomerizations of l -Leucine Ethyl Ester with l -Glutamic Acid Diethyl Ester: Sequence and Chain Length Distributions. *Macromolecules*, 41(19), pp.7003–7012. Available at: http://pubs.acs.org/doi/abs/10.1021/ma800946d [Accessed October 10, 2018].

Li, J. & Koehl, P., 2014. 3D representations of amino acids—applications to protein sequence comparison and classification. *Computational and Structural Biotechnology Journal*.

Li, L., Wang, B. & Meroueh, S.O., 2011. Support Vector Regression Scoring of Receptor–Ligand Complexes for Rank-Ordering and Virtual Screening of Chemical Libraries. *Journal of Chemical Information and Modeling*, 51(9), pp.2132–2138. Available at: http://pubs.acs.org/doi/abs/10.1021/ci200078f [Accessed September 10, 2018].

Li, Y.Y., An, J. & Jones, S.J.M., 2011. A Computational Approach to Finding Novel Targets for Existing Drugs P. E. Bourne, ed. *PLoS Computational Biology*, 7(9), p.e1002139. Available at: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002139 [Accessed July 8, 2015].

Libby, P., Ridker, P.M. & Hansson, G.K., 2011. Progress and challenges in translating the biology of atherosclerosis. *Nature*, 473(7347), pp.317–325. Available at: http://www.nature.com/articles/nature10146 [Accessed October 22, 2018].

Licklider, J.C., 1963. *"Topics for Discussion at the Forthcoming Meeting, Memorandum For: Members and Affiliates of the Intergalactic Computer Network,"* Washington, D. C. Available at: http://www.kurzweilai.net/memorandum-for-members-and-affiliates-of-the-intergalactic-computer-network.

Von Lilienfeld, O.A., 2013. First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties. *International Journal of Quantum Chemistry*.

Lindström, A. & Kekkonen, A., 2018. Do Mergers and Acquisitions Create Value for Acquirers? Short- and Long-Term Event Study on the Pharmaceutical Industry of Europe. Available at: https://lup.lub.lu.se/student-papers/search/publication/8945651 [Accessed October 20, 2018].

Lionta, E. et al., 2014. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current Topics in Medicinal Chemistry*.

Lipinski, C. a. et al., 2012. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 64, pp.4–17. Available at: http://dx.doi.org/10.1016/j.addr.2012.09.019.

Lippert, R.A. et al., 2007. A common, avoidable source of error in molecular dynamics integrators. *The Journal of Chemical Physics*, 126(4), p.046101. Available at: http://aip.scitation.org/doi/10.1063/1.2431176 [Accessed September 17, 2018].

Liu, J. & Wang, R., 2015. Classification of current scoring functions. *Journal of Chemical*

Information and Modeling.

Liu, R., Li, X. & Lam, K.S., 2017. Combinatorial chemistry in drug discovery. *Current Opinion in Chemical Biology*.

Lobanov, M.Y., Bogatyreva, N.S. & Galzitskaya, O. V., 2008. Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, 42(4), pp.623–628. Available at: http://link.springer.com/10.1134/S0026893308040195 [Accessed November 16, 2018].

Lobb, K.A., 2015. Isomerization of the 2-Norbornyl Carbocation. *European Journal of Organic Chemistry*.

London, N. et al., 2011. Rosetta FlexPepDock web server—high resolution modeling of peptide–protein interactions. *Nucleic Acids Research*, 39(suppl_2), pp.W249–W253. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr431 [Accessed September 14, 2018].

Loose, C. et al., 2006. A linguistic model for the rational design of antimicrobial peptides. *Nature*.

López-Vallejo, F. et al., 2012. Expanding the medicinally relevant chemical space with compound libraries. *Drug Discovery Today*, 17(July), pp.718–726.

Loughney, D., Claus, B.L. & Johnson, S.R., 2011. To measure is to know: An approach to CADD performance metrics. *Drug Discovery Today*.

Lovering, F., 2013. Escape from Flatland 2: complexity and promiscuity. *MedChemComm*, (4), pp.515–519.

Lovering, F., Bikker, J. & Humblet, C., 2009. Escape from flatland: increasing saturation as an approach to improving clinical success. *Journal of medicinal chemistry*, 52(21), pp.6752–6. Available at: http://0-pubs.acs.org.wam.seals.ac.za/doi/abs/10.1021/jm901241e [Accessed December 5, 2014].

Lowe, D., 2015. Chemical space is big. Really big. *MedChemComm*, 6, p.12. Available at: http://dx.doi.org/10.1039/C4MD90045F.

Luethi, P. & Luisi, P.L., 1984. Enzymic synthesis of hydrocarbon-soluble peptides with reverse micelles. *Journal of the American Chemical Society*, 106(23), pp.7285–7286. Available at: http://pubs.acs.org/doi/abs/10.1021/ja00335a092 [Accessed October 10, 2018].

Ma, L. et al., 2017. Peptide-Drug Conjugate: A Novel Drug Design Approach. *Current Medicinal Chemistry*.

Mäde, V., Els-Heindl, S. & Beck-Sickinger, A.G., 2014. Automated solid-phase peptide synthesis to obtain therapeutic peptides. *Beilstein Journal of Organic Chemistry*, 10(1), pp.1197–1212. Available at: http://www.beilstein-journals.org/bjoc/content/10/1/118 [Accessed September 14, 2016].

Maggiora, G. et al., 2014. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry*, 57(8), pp.3186–3204. Available at: http://pubs.acs.org/doi/10.1021/jm401411z [Accessed September 29, 2018].

Maggiora, G.M., 2006. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *Journal of Chemical Information and Modeling*, 46(4), pp.1535–1535. Available at: https://pubs.acs.org/doi/abs/10.1021/ci060117s [Accessed October 2, 2018].

Malo, N. et al., 2006. Statistical practice in high-throughput screening data analysis. *Nature Biotechnology*.

Manallack, D.T. et al., 2014. The Significance of Acid / Base Properties in Drug Discovery. *Chem. Soc. Rev.*, 42(2), pp.485–496.

Mann, A., 2018. Heavy-lift rocket poised to boost space science. *Science (New York, N.Y.)*, 359(6374), pp.376–377. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29371446 [Accessed February 7, 2019].

Maredza, M., Hofman, K.J. & Tollman, S.M., 2017. A hidden menace: Cardiovascular disease in South Africa and the costs of an inadequate policy response. *SA Heart*, 7(4), pp.48–57. Available at: http://www.journals.ac.za/index.php/SAHJ/article/view/1924 [Accessed October 22, 2018].

MAROKO, P.R. et al., 1971. Factors Influencing Infarct Size Following Experimental Coronary Artery Occlusions. *Circulation*, 43(1), pp.67–82. Available at: https://www.ahajournals.org/doi/10.1161/01.CIR.43.1.67 [Accessed October 22, 2018].

Marsault, E. & Peterson, M.L., 2011. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *Journal of Medicinal Chemistry*, 54(7), pp.1961–2004. Available at: http://pubs.acs.org/doi/abs/10.1021/jm1012374 [Accessed December 29, 2016].

Martin, Y. et al., 1973. Potential Anti-Parkinson Drugs Designed by Receptor Mapping. *Journal of Medicinal Chemistry*, 16(2). Available at: https://pubs.acs.org/sharingguidelines [Accessed October 1, 2018].

Marzilli, M. et al., 2000. Beneficial Effects of Intracoronary Adenosine as an Adjunct to Primary Angioplasty in Acute Myocardial Infarction. *Circulation*, 101, pp.2154–2159. Available at: http://www.circulationaha.org [Accessed October 22, 2018].

Matter, H. et al., 2005. Structural Requirements for Factor Xa Inhibition by 3-Oxybenzamides with Neutral P1 Substituents: Combining X-ray Crystallography, 3D-QSAR, and Tailored Scoring Functions. *Journal of Medicinal Chemistry*, 48, pp.3290–3312. Available at: https://pubs.acs.org/doi/10.1021/jm049187l [Accessed September 11, 2018].

Mazur, A.K. & Abagyan, R.A., 1989. New Methodology for Computer-Aided Modelling of Biomolecular Structure and Dynamics 1. Non-Cyclic Structures. *Journal of Biomolecular Structure and Dynamics*, 6(4), pp.815–832. Available at: http://www.tandfonline.com/doi/abs/10.1080/07391102.1989.10507739 [Accessed September 18, 2018].

McChesney, J.D., Venkataraman, S.K. & Henri, J.T., 2007. Plant natural products: Back to the future or into extinction? *Phytochemistry*.

McHugh, S.M., Rogers, J.R., Solomon, S.A., et al., 2016. Computational methods to design cyclic peptides. *Current Opinion in Chemical Biology*, 34, pp.95–102. Available at:

http://linkinghub.elsevier.com/retrieve/pii/S1367593116301016 [Accessed March 15, 2017].

McHugh, S.M., Rogers, J.R., Yu, H., et al., 2016. Insights into How Cyclic Peptides Switch Conformations. *Journal of Chemical Theory and Computation*, 12(5), pp.2480–2488. Available at: http://pubs.acs.org/doi/abs/10.1021/acs.jctc.6b00193 [Accessed November 5, 2016].

McNamee, L. & Ledley, F., 2015. What does the current biotech stock market value? *Nature Biotechnology*, 33(8), pp.813–814. Available at: http://www.nature.com/articles/nbt.3303 [Accessed October 20, 2018].

Meanwell, N.A., 2016. Improving Drug Design: An Update on Recent Applications of Efficiency Metrics, Strategies for Replacing Problematic Elements, and Compounds in Nontraditional Drug Space. *Chemical Research in Toxicology*, 29(4), pp.564–616. Available at: http://pubs.acs.org/doi/abs/10.1021/acs.chemrestox.6b00043 [Accessed June 9, 2016].

Medina-Franco, J.L. et al., 2013. Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discovery Today*, 18(May), pp.495–501.

Medina-Franco, J.L., Martinez-Mayorga, K. & Meuricea, N., 2014. Balancing novelty with confined chemical space in modern drug discovery. *Expert Opinion on Drug Discovery*, 9(2), pp.151–165.

Melville, J., Burke, E. & Hirst, J., 2009. Machine Learning in Virtual Screening. *Combinatorial Chemistry & High Throughput Screening*, 12(4), pp.332–343. Available at: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1386-2073&volume=12&issue=4&spage=332 [Accessed September 29, 2018].

Mende, F. & Seitz, O., 2011. 9-Fluorenylmethoxycarbonyl-based solid-phase synthesis of peptide α-thioesters. *Angewandte Chemie (International ed. in English)*, 50(6), pp.1232–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21290490 [Accessed November 9, 2015].

Merched, A.J. et al., 2008. Atherosclerosis: evidence for impairment of resolution of vascular inflammation governed by specific lipid mediators. *The FASEB Journal*, 22(10), pp.3595–3606. Available at: http://www.fasebj.org/doi/10.1096/fj.08-112201 [Accessed October 22, 2018].

Merrifield, R.B., 1963. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *Journal of the American Chemical Society*.

Mignani, S. et al., 2016. Why and how have drug discovery strategies in pharma changed? What are the new mindsets? *Drug Discovery Today*.

Miller, B.R. et al., 2012. MMPBSA.py : An Efficient Program for End-State Free Energy Calculations.

Minois, P. et al., 2017. [60]Fullerene l -Amino Acids and Peptides: Synthesis under Phase-Transfer Catalysis Using a Phosphine–Borane Linker. Electrochemical Behavior. *The Journal of Organic Chemistry*, 82(21), pp.11358–11369. Available at: http://pubs.acs.org/doi/abs/10.1021/acs.joc.7b01737 [Accessed October 12, 2018].

Moal, I.H. et al., 2013. Scoring functions for protein-protein interactions. *Current Opinion in Structural Biology*.

Moal, I.H. et al., 2013. The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinformatics*, 14(1), p.286. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-286 [Accessed September 15, 2018].

Mobley, D.L. & Klimovich, P. V., 2012. Perspective: Alchemical free energy calculations for drug discovery. *Journal of Chemical Physics*.

Moreira, I.S. et al., 2015. A new scoring function for protein–protein docking that identifies native structures with unprecedented accuracy. *Physical Chemistry Chemical Physics*, 17(4), pp.2378–2387. Available at: http://xlink.rsc.org/?DOI=C4CP04688A [Accessed September 15, 2018].

Moreira, I.S., Fernandes, P.A. & Ramos, M.J., 2009. Protein-protein docking dealing with the unknown. *Journal of Computational Chemistry*, 31(2), p.NA-NA. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19462412 [Accessed September 14, 2018].

Morgado, C. et al., 2007. Can the DFT-D method describe the full range of noncovalent interactions found in large biomolecules? *Phys. Chem. Chem. Phys.*, 9(4), pp.448–451. Available at: http://xlink.rsc.org/?DOI=B615263E [Accessed September 6, 2018].

Morgan, S. et al., 2011. The cost of drug development: A systematic review. *Health Policy*, 100(1), pp.4–17. Available at: https://www.sciencedirect.com/science/article/pii/S0168851010003659 [Accessed October 24, 2018].

Morphy, R., Kay, C. & Rankovic, Z., 2004. From magic bullets to designed multiple ligands. *Drug Discovery Today*.

Morris, G.M. et al., 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16), pp.2785–91. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2760638&tool=pmcentrez&rendertype=abstract [Accessed October 17, 2014].

Morris, G.M. et al., 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*.

Muegge, I., 2006. PMF scoring revisited. *Journal of Medicinal Chemistry*.

Muegge, I. & Martin, Y.C., 1999. A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *Journal of Medicinal Chemistry*, 42, pp.791–804. Available at: https://pubs.acs.org/doi/10.1021/jm980536j [Accessed September 10, 2018].

Mulholland, D.A. et al., 2004. Xanthones from Drimiopsis maculata. *Journal of natural products*, 67, pp.1726–1728.

Munos, B., 2009. Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery*.

Nazareth, W., Yafei, N. & Crompton, M., 1991. Inhibition of anoxia-induced injury in heart myocytes by cyclosporin A. *Journal of molecular and cellular cardiology*, 23(12), pp.1351–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1811053 [Accessed April 24, 2017].

Neudert, G. & Klebe, G., 2011. DSX: A knowledge-based scoring function for the assessment of protein-ligand complexes. *Journal of Chemical Information and Modeling*.

Nicholls, A. et al., 2010. Molecular Shape and Medicinal Chemistry: A Perspective. *Journal of Medicinal Chemistry*, 53(10), pp.3862–3886.

Nicholson, S., Danzon, P.M. & Mccullough, J., 2005. Biotech-Pharmaceutical Alliances as aSignal of Asset and Firm Quality. *Journal of Business*, 78(4), pp.1433–1464. Available at: http://repository.upenn.edu/hcmg_papershttp://dx.doi.org/10.1086/430865http://repository.upenn.edu/hcmg_papers/112 [Accessed October 20, 2018].

Nicolaou, K.C., 2014. Advancing the Drug Discovery and Development Process. *Angewandte Chemie (International ed. in English)*, p.n/a-n/a. Available at: http://doi.wiley.com/10.1002/anie.201404761%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/25045053.

Nusca, A. et al., 2010. Statin loading for acute coronary syndromes. *Current Opinion in Cardiology*, 25(4), pp.373–378. Available at: https://insights.ovid.com/crossref?an=00001573-201007000-00014 [Accessed October 22, 2018].

Oakley, M.T. et al., 2013. Computational and Experimental Investigations into the Conformations of Cyclic Tetra-α/β-peptides. *The Journal of Physical Chemistry B*, 117(27), pp.8122–8134. Available at: http://pubs.acs.org/doi/abs/10.1021/jp4043039 [Accessed November 8, 2016].

Obreza, A. & Gobec, S., 2004. Recent Advances in Design, Synthesis and Biological Activity of Aminoalkylsulfonates and Sulfonamidopeptides. *Current Medicinal Chemistry*, 11(24), pp.3263–3278. Available at: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=0929-8673&volume=11&issue=24&spage=3263 [Accessed October 12, 2018].

Okabe, T. et al., 2001. Replica-exchange Monte Carlo method for the isobaric-isothermal ensemble. *Chemical Physics Letters*, 335, pp.435–439. Available at: www.elsevier.nl/locate/cplett [Accessed November 23, 2018].

Oloff, S., Mailman, R.B. & Tropsha, A., 2005. Application of Validated QSAR Models of D1 Dopaminergic Antagonists for Database Mining. *Journal of Medicinal Chemistry*, 48, pp.7322–7335. Available at: https://cdn-pubs.acs.org/doi/full/10.1021/jm049116m?src=recsys& [Accessed October 1, 2018].

Orloff, J. et al., 2009. The future of drug development: Advancing clinical trial design. *Nature Reviews Drug Discovery*.

Orth, A.P. et al., 2004. The promise of genomics to identify novel therapeutic targets. *Expert Opinion on Therapeutic Targets*, 8(6), pp.587–596. Available at: http://www.tandfonline.com/doi/full/10.1517/14728222.8.6.587 [Accessed October

23, 2018].

Osolodkin, D.I. et al., 2015. Progress in visual representations of chemical space. *Expert Opinion on Drug Discovery*.

Ostermeir, K. & Zacharias, M., 2014. Hamiltonian replica-exchange simulations with adaptive biasing of peptide backbone and side chain dihedral angles. *Journal of Computational Chemistry*, 35(2), pp.150–158. Available at: http://doi.wiley.com/10.1002/jcc.23476 [Accessed September 21, 2018].

Ottani, F. et al., 2016. Cyclosporine A in Reperfused Myocardial Infarction: The Multicenter, Controlled, Open-Label CYCLE Trial. *Journal of the American College of Cardiology*, 67(4), pp.365–374. Available at: http://www.sciencedirect.com/science/article/pii/S0735109715074343 [Accessed April 24, 2017].

Paissoni, C. et al., 2015. Metadynamics Simulations Rationalise the Conformational Effects Induced by *N* -Methylation of RGD Cyclic Hexapeptides. *Chemistry - A European Journal*, 21(40), pp.14165–14170. Available at: http://doi.wiley.com/10.1002/chem.201501196 [Accessed November 8, 2016].

Pammolli, F., Magazzini, L. & Riccaboni, M., 2011. The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery*.

Pan, A.C. et al., 2013. Molecular determinants of drug-receptor binding kinetics. *Drug discovery today*, 18(13–14), pp.667–73. Available at: http://www.sciencedirect.com/science/article/pii/S1359644613000627 [Accessed May 17, 2016].

Papa, E. et al., 2009. Development, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers. *QSAR & Combinatorial Science*, 28(8), pp.790–796. Available at: http://doi.wiley.com/10.1002/qsar.200860183 [Accessed May 27, 2017].

Patriksson, A. & Van Der Spoel, D., 2008. A temperature predictor for parallel tempering simulations.

Paul, S.M. et al., 2010. How to improve RD productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9, pp.203–214. Available at: http://www.nature.com/doifinder/10.1038/nrd3078 [Accessed June 9, 2016].

Peakman, T. et al., 2003. Delivering the power of discovery in large pharmaceutical organizations. *Drug Discovery Today*.

Pereira, D.A. & Williams, J.A., 2007. Origin and evolution of high throughput screening. *British Journal of Pharmacology*.

Perrelli, M.-G., Pagliaro, P. & Penna, C., 2011. Ischemia/reperfusion injury and cardioprotective mechanisms: Role of mitochondria and reactive oxygen species. *World Journal of Cardiology*, 3(6), p.186. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3139040/ [Accessed August 7, 2017].

Peter, T.C. et al., 1978. Reduction of enzyme levels by propranolol after acute myocardial infarction. *Circulation*, 57(6), pp.1091–1095. Available at:

https://www.semanticscholar.org/paper/Reduction-of-enzyme-levels-by-propranolol-after-Peter-Norris/f048648bc59a669367f8b0a9a56693449a8e15e9 [Accessed October 22, 2018].

Petsalaki, E. & Russell, R.B., 2008. Peptide-mediated interactions in biological systems: new discoveries and applications. *Current Opinion in Biotechnology*, 19(4), pp.344–350. Available at: https://www.sciencedirect.com/science/article/pii/S0958166908000724?via%3Dihub [Accessed September 14, 2018].

PhRMA, 2018. Industry Profile. *Pharmaceutical Research and Manufacturers of America*. Available at: http://phrma.org/industryprofile/ [Accessed July 6, 2018].

Pignone, M., Phillips, C. & Mulrow, C., 2000. Use of lipid lowering drugs for primary prevention of coronary heart disease: meta-analysis of randomised trials. *BMJ (Clinical research ed.)*, 321(7267), pp.983–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11039962 [Accessed October 22, 2018].

Piot, C. et al., 2008. Effect of Cyclosporine on Reperfusion Injury in Acute Myocardial Infarction. *New England Journal of Medicine*, 359(5), pp.473–481. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMoa071142 [Accessed October 22, 2018].

Polanski, J., 2009. Receptor dependent multidimensional QSAR for modeling drug--receptor interactions. *Current medicinal chemistry*, 16(25), pp.3243–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19548875 [Accessed October 1, 2018].

Polishchuk, P.G., Madzhidov, T.I. & Varnek, a., 2013. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design*, 27(8), pp.675–679.

Procacci, P., 2017. Alchemical determination of drug-receptor binding free energy: Where we stand and where we could move to. *Journal of Molecular Graphics and Modelling*, 71, pp.233–241. Available at: https://www.sciencedirect.com/science/article/pii/S1093326316304302 [Accessed September 23, 2018].

Pronk, S. et al., 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics (Oxford, England)*, 29(7), pp.845–54.

Przyklenk, K. et al., 1993. Regional Ischemic “Preconditioning” Protects Remote Virgin Myocardium From Subsequent Sustained Coronary Occlusion. *Circulation*, 87(3), pp.893–899. Available at: http://ahajournals.org [Accessed October 22, 2018].

Qin, X. et al., 2014. Influence of N_{ε} -Protecting Groups on the Protease-Catalyzed Oligomerization of l -Lysine Methyl Ester. *ACS Catalysis*, 4(6), pp.1783–1792. Available at: http://pubs.acs.org/doi/10.1021/cs500268d [Accessed October 10, 2018].

Quezada, A.G. et al., 2017. Interplay between Protein Thermal Flexibility and Kinetic Stability. *Structure (London, England : 1993)*, 25(1), pp.167–179. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28052236 [Accessed October 10, 2018].

Quiroga, R. & Villarreal, M.A., 2016. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLoS ONE*.

Qureshi, A., Thakur, N. & Kumar, M., 2013. HIPdb: A Database of Experimentally Validated HIV Inhibiting Peptides. *PLoS ONE*.

Räder, A.F.B. et al., 2018. Improving oral bioavailability of cyclic peptides by N-methylation. *Bioorganic & Medicinal Chemistry*, 26(10), pp.2766–2773. Available at: https://www.sciencedirect.com/science/article/pii/S0968089617312166 [Accessed October 15, 2018].

Ragoza, M. et al., 2017. Protein–Ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 57(4), pp.942–957. Available at: http://pubs.acs.org/doi/10.1021/acs.jcim.6b00740 [Accessed September 10, 2018].

Raman, E.P. et al., 2017. Estimation of relative free energies of binding using pre-computed ensembles based on the single-step free energy perturbation and the site-identification by Ligand competitive saturation approaches. *Journal of Computational Chemistry*.

Rask-Andersen, M., Almén, M.S. & Schiöth, H.B., 2011. Trends in the exploitation of novel drug targets. *Nature reviews. Drug discovery*, 10(8), pp.579–90. Available at: http://dx.doi.org/10.1038/nrd3478 [Accessed August 31, 2015].

Rask-Andersen, M., Masuram, S. & Schiöth, H.B., 2014. The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual review of pharmacology and toxicology*, 54, pp.9–26. Available at: http://0-www.annualreviews.org.wam.seals.ac.za/doi/abs/10.1146/annurev-pharmtox-011613-135943 [Accessed March 15, 2016].

Rathore, N., Chopra, M. & De Pablo, J.J., 2005. Optimal allocation of replicas in parallel tempering simulations. *The Journal of Chemical Physics*, 122. Available at: http://jcp.aip.org/jcp/copyright.jsp [Accessed November 23, 2018].

Raveh, B. et al., 2011. Rosetta FlexPepDock ab-initio: Simultaneous Folding, Docking and Refinement of Peptides onto Their Receptors V. N. Uversky, ed. *PLoS ONE*, 6(4), p.e18934. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21572516 [Accessed September 14, 2018].

Raveh, B., London, N. & Schueler-Furman, O., 2010. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Structure, Function, and Bioinformatics*, 78(9), p.NA-NA. Available at: http://doi.wiley.com/10.1002/prot.22716 [Accessed September 14, 2018].

Recio, C. et al., 2017. The Potential Therapeutic Application of Peptides and Peptidomimetics in Cardiovascular Disease. *Frontiers in Pharmacology*, 7, p.526. Available at: http://journal.frontiersin.org/article/10.3389/fphar.2016.00526/full [Accessed October 22, 2018].

Rentzsch, R. & Renard, B.Y., 2015. Docking small peptides remains a great challenge: An assessment using AutoDock Vina. *Briefings in Bioinformatics*.

Reuer, J.J., Tong, T.W. & Wu, C.-W., 2012. A SIGNALING THEORY OF ACQUISITION PREMIUMS: EVIDENCE FROM IPO TARGETS. *Source: The Academy of Management Journal*, 55(3), pp.667–683. Available at: https://www.jstor.org/stable/pdf/23317495.pdf?refreqid=excelsior%3Aa5d6d903c41d8aaf05dc5a1ca9276d93 [Accessed October 20, 2018].

Reutlinger, M. et al., 2014. Multi-objective molecular de novo design by adaptive fragment prioritization. *Angewandte Chemie - International Edition*.

Ribeiro, A.A.S.T. & Ortiz, V., 2015. MDN: A Web Portal for Network Analysis of Molecular Dynamics Simulations. *Biophysical Journal*, 109(6), pp.1110–1116.

Ripphausen, P., Nisius, B. & Bajorath, J., 2011. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today*.

Rishton, G.M., 2008. Natural Products as a Robust Source of New Drugs and Drug Leads: Past Successes and Present Day Issues. *American Journal of Cardiology*.

Rishton, G.M., 2003. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today*, 8(2), pp.86–96.

Rognan, D. et al., 1992. Structure and molecular modeling of GABAA receptor antagonists. *Journal of medicinal chemistry*.

Rohrbacher, F. et al., 2015. Spontaneous head-to-tail cyclization of unprotected linear peptides with the KAHA ligation. *Chem. Sci.*, 6(8), pp.4889–4896. Available at: http://xlink.rsc.org/?DOI=C5SC01774B [Accessed September 14, 2016].

Rozek, A. et al., 2003. Structure-Based Design of an Indolicidin Peptide Analogue with Increased Protease Stability. *Biochemistry*.

Ruddigkeit, L. et al., 2012. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52, pp.2864–2875.

Ruddigkeit, L., Blum, L.C. & Reymond, J.L., 2013. Visualization and virtual screening of the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 53(1), pp.56–65.

Rush, T.S. et al., 2005. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of Medicinal Chemistry*, 48(5), pp.1489–1495.

Russ, A.P. & Lampel, S., 2005. The druggable genome: an update. *Drug discovery today*, 10(23–24), pp.1607–10. Available at: http://www.sciencedirect.com/science/article/pii/S1359644605036664 [Accessed May 31, 2016].

Russo, I. et al., 2004. The activity of constitutive nitric oxide synthase is increased by the pathway cAMP/cAMP-activated protein kinase in human platelets. New insights into the antiaggregating effects of cAMP-elevating agents. *Thrombosis Research*, 114(4), pp.265–273. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15381390 [Accessed October 22, 2018].

Sahigara, F. et al., 2013. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *Journal of Cheminformatics*, 5(1), p.27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23721648 [Accessed October 1, 2018].

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, Available at: http://linkinghub.elsevier.com/retrieve/pii/S1359644612003674 [Accessed June 9, 2016].

Santillo, E. et al., 2016. Cardioprotection by Conditioning Mimetic Drugs. *Anti-inflammatory & anti-allergy agents in medicinal chemistry*, 15(1), pp.15–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27439628 [Accessed October 22, 2018].

Satyawali, Y., Vanbroekhoven, K. & Dejonghe, W., 2017. Process intensification: The future for enzymatic processes? *Biochemical Engineering Journal*, 121, pp.196–223. Available at: https://www.sciencedirect.com/science/article/pii/S1369703X17300323 [Accessed October 12, 2018].

Scannell, J.W. et al., 2012. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3), pp.191–200. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22378269 [Accessed October 23, 2018].

Schaefer, S., 2015. Star Wars 101: A guide for the non-geek prepping for “The Force Awakens.” *Boston Herald (MA)*. Available at: http://0-search.ebscohost.com.wam.seals.ac.za/login.aspx?direct=true&db=nfh&AN=2W63730109517&site=eds-live.

Schneider, G., 2013. *De novo Molecular Design*, John Wiley & Sons. Available at: https://books.google.com/books?id=1QFRAQAAQBAJ&pgis=1 [Accessed September 16, 2015].

Schneider, G. et al., 2009. Voyages to the (un)known: adaptive design of bioactive compounds. *Trends in Biotechnology*.

Schneider, G. & Fechner, U., 2005. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*.

Schneider, G., Schuchhardt, J. & Wrede, P., 1994. Artificial neural networks and simulated molecular evolution are potential tools for sequence-oriented protein design. *Bioinformatics*.

Seidel, T., Wolber, G. & Murgueitio, M.S., 2018. Pharmacophore Perception and Applications. In *Applied Chemoinformatics*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, pp. 259–282. Available at: http://doi.wiley.com/10.1002/9783527806539.ch6f [Accessed October 1, 2018].

Shan, Y. et al., 2011. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24), pp.9181–3. Available at: http://dx.doi.org/10.1021/ja202726y [Accessed May 17, 2016].

Shanmugasundaram, K. & Rigby, A., 2009. Exploring Novel Target Space: A Need to Partner High Throughput Docking and Ligand-Based Similarity Searches? *Combinatorial Chemistry & High Throughput Screening*, 12(10), pp.984–999. Available at: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1386-2073&volume=12&issue=10&spage=984 [Accessed September 29, 2018].

Sharp, S. et al., 2015. Pharmacodynamic effects of C-domain-specific ACE inhibitors on the renin-angiotensin system in myocardial infarcted rats. *Journal of the Renin-Angiotensin-*

Aldosterone System, 16(4), pp.1149–1158. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25757657 [Accessed October 22, 2018].

Shepherd, J. et al., 2008. Intensive Lipid Lowering With Atorvastatin in Patients With Coronary Heart Disease and Chronic Kidney Disease. *Journal of the American College of Cardiology*, 51(15), pp.1448–1454. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0735109708003549 [Accessed October 22, 2018].

Shim, J., Mackerell, A.D. & Jr., 2011. Computational ligand-based rational design: Role of conformational sampling and force fields in model development. *MedChemComm*, 2(5), pp.356–370. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21716805 [Accessed March 29, 2017].

Shin, W.-H. et al., 2016. PL-PatchSurfer2: Improved Local Surface Matching-Based Virtual Screening Method That Is Tolerant to Target and Ligand Structure Variation. *Journal of Chemical Information and Modeling*, 56(9), pp.1676–1691. Available at: http://pubs.acs.org/doi/10.1021/acs.jcim.6b00163 [Accessed September 29, 2018].

Shterev, I.D. et al., 2018. Bayesian Multi-Plate High-Throughput Screening of Compounds. *Scientific Reports*.

Shultz, M.D., 2013. *Setting expectations in molecular optimizations: Strengths and limitations of commonly used composite parameters*, Available at: http://www.sciencedirect.com/science/article/pii/S0960894X1300958X [Accessed May 28, 2017].

Singer, J.A. & Purcell, W.P., 1967. Relationships Among Current Quantitative Structure-Activity Models. *Journal of Medicinal Chemistry*.

Singh, S. et al., 2015. SATPdb: A database of structurally annotated therapeutic peptides. *Nucleic Acids Research*.

Sliwoski, G. et al., 2014. Computational methods in drug discovery. *Pharmacological reviews*, 66(1), pp.334–95. Available at: http://pharmrev.aspetjournals.org/content/66/1/334.full [Accessed January 9, 2015].

Smith, D.A., Jones, B.C. & Walker, D.K., 1996. Design of Drugs Involving the Concepts and Theories of Drug Metabolism and Pharmacokinetics. *Medicinal Chemistry Reviews*, 16(3), pp.243–266.

Smith, G.F., 2011. Designing Drugs to Avoid Toxicity. In *Progress in Medicinal Chemistry*. pp. 1–47.

Smolskaya, S., Zhang, Z.J. & Alfonta, L., 2013. Enhanced Yield of Recombinant Proteins with Site-Specifically Incorporated Unnatural Amino Acids Using a Cell-Free Expression System J. D. Hoheisel, ed. *PLoS ONE*, 8(7), p.e68363. Available at: http://dx.plos.org/10.1371/journal.pone.0068363 [Accessed October 12, 2018].

Snell, M. et al., 2018. InterPlanetary Networking Special Interest Group (IPNSIG). Available at: http://ipnsig.org/ [Accessed February 7, 2019].

Sperandio, O. et al., 2010. Rationalizing the chemical space of protein–protein interaction inhibitors. *Drug Discovery Today*, 15(5), pp.220–229.

Steg, P.G. et al., 2012. ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. *European Heart Journal*, 33(20), pp.2569–2619. Available at: https://academic.oup.com/eurheartj/article/33/20/2569/447818 [Accessed October 22, 2018].

Stein, S.A.M. et al., 2006. Principal Components Analysis: A Review of its Application on Molecular Dynamics Data. *Annual Reports in Computational Chemistry*, 2(06), pp.233–248. Available at: http://www.worldscientific.com/worldscibooks/10.1142/p784#t=aboutBook.

Stockdale, T.P. & Williams, C.M., 2015. Pharmaceuticals that contain polycyclic hydrocarbon scaffolds. *Chemical Society reviews*, 44(21), pp.7737–7763. Available at: http://xlink.rsc.org/?DOI=C4CS00477A [Accessed August 17, 2016].

Stumpfe, D. et al., 2014. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *Journal of Medicinal Chemistry*, 57(1), pp.18–28. Available at: http://pubs.acs.org/doi/10.1021/jm401120g [Accessed October 2, 2018].

Sugita, Y. & Okamoto, Y., 1999. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1–2), pp.141–151. Available at: https://www.sciencedirect.com/science/article/pii/S0009261499011239?via%3Dihub [Accessed September 21, 2018].

Suryanarayana Birudukota, N. V., Franke, R. & Hofer, B., 2016. An approach to “escape from flatland”: Chemo-enzymatic synthesis and biological profiling of a library of bridged bicyclic compounds. *Organic and Biomolecular Chemistry*.

Tabas, I., 2010. Macrophage death and defective inflammation resolution in atherosclerosis. *Nature Reviews Immunology*, 10(1), pp.36–46. Available at: http://www.nature.com/articles/nri2675 [Accessed October 22, 2018].

Tajabadi, F.M., Campitelli, M.R. & Quinn, R.J., 2013. Scaffold Flatness: Reversing the Trend. *Springer Science Reviews*, 1(1–2), pp.141–151. Available at: http://link.springer.com/10.1007/s40362-013-0014-7 [Accessed May 24, 2016].

Taniyama, Y. et al., 1997. Beneficial Effect of Intracoronary Verapamil on Microvascular and Myocardial Salvage in Patients With Acute Myocardial Infarction. *Journal of the American College of Cardiology*, 30(5), pp.1193–1199. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0735109797002775 [Accessed October 22, 2018].

Tarcsay, Á. & Keseru, G.M., 2015. Is there a link between selectivity and binding thermodynamics profiles? *Drug Discovery Today*, 20(1), pp.86–94.

Teague, S.J., 2011. Learning lessons from drugs that have recently entered the market. *Drug discovery today*, 16(9–10), pp.398–411. Available at: http://www.sciencedirect.com/science/article/pii/S1359644611000754 [Accessed May 23, 2016].

Tegge, W., Bautsch, W. & Frank, R., 2007. Synthesis of cyclic peptides and peptide libraries on a new disulfide linker. *Journal of peptide science : an official publication of the European Peptide Society*, 13(10), pp.693–9. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/17668890 [Accessed November 9, 2015].

Teoh, T.C., Salmah, I. & Tang, J.M., 2014. Molecular Dynamics and Docking of Biphenyl: A Potential Attachment Inhibitor for HIV-1 gp120 Glycoprotein. *Tropical Journal of Pharmaceutical Research*, 13(March), pp.339–346.

Testa, B. et al., 2000. The influence of lipophilicity on the pharmacokinetic behavior of drugs: Concepts and examples. *Perspectives in Drug Discovery and Design*, 19, pp.179–211.

Thomas, G.L. & Johannes, C.W., 2011. Natural product-like synthetic libraries. *Current opinion in chemical biology*, 15(4), pp.516–22. Available at: http://www.sciencedirect.com/science/article/pii/S1367593111000913 [Accessed September 18, 2015].

Thust & Koksch, 2003. Protease-Catalyzed Peptide Synthesis for the Site-Specific Incorporation of α-Fluoroalkyl Amino Acids into Peptides†. *Journal of Organic Chemistry*, 68(6), pp.2290–2296. Available at: https://pubs.acs.org/doi/10.1021/jo020613p [Accessed October 12, 2018].

Thygesen, K. et al., 2018. Fourth universal definition of myocardial infarction (2018). *European Heart Journal*. Available at: https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehy462/5079081 [Accessed October 23, 2018].

Tiejun Cheng et al., 2007. Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. Available at: http://pubs.acs.org/doi/abs/10.1021/ci700257y [Accessed May 27, 2017].

Tielmann, P. et al., 2014. Increasing the activity and enantioselectivity of lipases by sol–gel immobilization: further advancements of practical interest. *Nanoscale*, 6(12), pp.6220–6228. Available at: http://xlink.rsc.org/?DOI=C3NR06317H [Accessed October 10, 2018].

Torres, S. & Castro, G.R., 2004. Non-Aqueous Biocatalysis in Homogeneous Solvent Systems. *Food Technol. Biotechnol.*, 42(4), pp.271–277. Available at: https://pdfs.semanticscholar.org/e8b5/ce43ddeaf909d681c4124d3af3551f84a2bb.pdf [Accessed October 10, 2018].

Torrie, G.M. & Valleau, J.P., 1974. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chemical Physics Letters*, 28(4), pp.578–581. Available at: https://www.sciencedirect.com/science/article/pii/0009261474801090?via%3Dihub [Accessed September 20, 2018].

Townes, S. et al., 2004. *The Mars Laser Communication Demonstration*,

Travis, K.P. & Braga, C., 2006. Configurational temperature and pressure molecular dynamics: review of current methodology and applications to the shear flow of a simple fluid. *Molecular Physics*, 104(22–24), pp.3735–3749. Available at: http://www.tandfonline.com/doi/abs/10.1080/00268970601014880 [Accessed September 19, 2018].

Trellet, M., Melquiond, A.S.J. & Bonvin, A.M.J.J., 2013. A Unified Conformational Selection

and Induced Fit Approach to Protein-Peptide Docking O. Keskin, ed. *PLoS ONE*, 8(3), p.e58769. Available at: http://dx.plos.org/10.1371/journal.pone.0058769 [Accessed September 11, 2018].

Tripathi, N.K., 2016. Production and Purification of Recombinant Proteins from *Escherichia coli*. *ChemBioEng Reviews*, 3(3), pp.116–133. Available at: http://doi.wiley.com/10.1002/cben.201600002 [Accessed October 12, 2018].

Truong, D.T. & Li, M.S., 2018. Probing the Binding Affinity by Jarzynski's Nonequilibrium Binding Free Energy and Rupture Time. *The Journal of Physical Chemistry B*, 122(17), pp.4693–4699. Available at: http://pubs.acs.org/doi/10.1021/acs.jpcb.8b02137 [Accessed March 28, 2019].

Truszkowski, A. et al., 2011. New developments on the cheminformatics open workflow environment CDK-Taverna. *Journal of Cheminformatics*, 3(1), p.54. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22166170 [Accessed October 30, 2018].

Tsai, C.-W. et al., 2009. Coupling Molecular Dynamics Simulations with Experiments for the Rational Design of Indolicidin-Analogous Antimicrobial Peptides. *Journal of Molecular Biology*.

Tsai, T.Y., Chang, K.W. & Chen, C.Y.C., 2011. IScreen: World's first cloud-computing web server for virtual screening and de novo drug design based on TCM database@Taiwan. *Journal of Computer-Aided Molecular Design*.

Tsallis, C. & Stariolo, D.A., 1996. Generalized simulated annealing. *Physica A: Statistical Mechanics and its Applications*, 233(1–2), pp.395–406. Available at: https://www.sciencedirect.com/science/article/pii/S0378437196002713?via%3Dihub [Accessed September 21, 2018].

Tsujishita, H. & Hirono, S., 1997. CAMDAS: An automated conformational analysis system using molecular dynamics. *Journal of Computer-Aided Molecular Design*, 11, pp.305–315.

Uhlig, T. et al., 2014. The emergence of peptides in the pharmaceutical business: From exploration to exploitation. *EuPA Open Proteomics*.

Ulven, S.M. et al., 2016. Exchanging a few commercial, regularly consumed food items with improved fat quality reduces total cholesterol and LDL-cholesterol: a double-blind, randomised controlled trial. Available at: https://doi.org/10.1017/S0007114516003445 [Accessed October 22, 2018].

van Unen, D.-J., Engbersen, J.F.J. & Reinhoudt, D.N., 2002. Why do crown ethers activate enzymes in organic solvents? *Biotechnology and Bioengineering*, 77(3), pp.248–255. Available at: http://doi.wiley.com/10.1002/bit.10032 [Accessed October 10, 2018].

Unterthiner, T. et al., 2015. *Deep Learning as an Opportunity in Virtual Screening*, Available at: http://www.datascienceassn.org/sites/default/files/Deep Learning as an Opportunity in Virtual Screening.pdf [Accessed October 1, 2018].

Vasile, F. et al., 2016. Thermodynamically–Weighted Conformational Ensemble of Cyclic RGD Peptidomimetics from NOE Data. *The Journal of Physical Chemistry B*, 120(29), pp.7098–7107. Available at: http://pubs.acs.org/doi/abs/10.1021/acs.jpcb.6b04941

[Accessed November 8, 2016].

Veber, D.F. et al., 2002. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, 45, pp.2615–2623.

Verlet, L., 1967. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159(1), pp.98–103. Available at: https://link.aps.org/doi/10.1103/PhysRev.159.98 [Accessed September 17, 2018].

Vinten-Johansen, J. & Shi, W., 2011. Perconditioning and Postconditioning: Current Knowledge, Knowledge Gaps, Barriers to Adoption, and Future Directions. *Journal of Cardiovascular Pharmacology and Therapeutics*, 16(3–4), pp.260–266. Available at: http://cpt.sagepub.com [Accessed October 22, 2018].

Virmani, R. et al., 2000. *Lessons From Sudden Coronary Death A Comprehensive Morphological Classification Scheme for Atherosclerotic Lesions*, Available at: http://www.atvbaha.org [Accessed October 22, 2018].

Viswanathan, K. et al., 2010. Protease-Catalyzed Oligomerization of Hydrophobic Amino Acid Ethyl Esters in Homogeneous Reaction Media Using l -Phenylalanine as a Model System. *Biomacromolecules*, 11(8), pp.2152–2160. Available at: http://pubs.acs.org/doi/abs/10.1021/bm100516x [Accessed October 10, 2018].

De Vivo, M. et al., 2016. The Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of medicinal chemistry*. Available at: http://0-pubs.acs.org.wam.seals.ac.za/doi/abs/10.1021/acs.jmedchem.5b01684 [Accessed January 31, 2016].

Vozoff, M. & Couluris, J., 2008. SpaceX Products-Advancing the Use of Space. In *AIAA SPACE 2008 Conference & Exposition*. Reston, Virigina: American Institute of Aeronautics and Astronautics. Available at: http://arc.aiaa.org/doi/10.2514/6.2008-7836 [Accessed February 7, 2019].

de Vries, S.J. et al., 2007. HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.726–733. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17803234 [Accessed September 14, 2018].

Waghu, F.H. et al., 2014. CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Research*.

Wagner, J.R. et al., 2013. Advanced techniques for constrained internal coordinate molecular dynamics. *Journal of Computational Chemistry*.

Wakefield, A.E., Wuest, W.M. & Voelz, V.A., 2015. Molecular Simulation of Conformational Pre-Organization in Cyclic RGD Peptides. *Journal of Chemical Information and Modeling*, 55(4), pp.806–813. Available at: http://pubs.acs.org/doi/abs/10.1021/ci500768u [Accessed November 8, 2016].

Wallace, C.J., 1995. Peptide ligation and semisynthesis. *Current Opinion in Biotechnology*, 6(4), pp.403–410. Available at: https://www.sciencedirect.com/science/article/pii/0958166995800697 [Accessed

October 13, 2018].

Wang, J. et al., 2004. Development and testing of a general Amber force field. *Journal of Computational Chemistry*.

Wang, L. et al., 2015. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society*, 137(7), pp.2695–2703. Available at: http://pubs.acs.org/doi/10.1021/ja512751q [Accessed September 24, 2018].

Wang, L. et al., 2017. Accurate Modeling of Scaffold Hopping Transformations in Drug Discovery. *Journal of Chemical Theory and Computation*, 13(1), pp.42–54. Available at: http://pubs.acs.org/doi/10.1021/acs.jctc.6b00991 [Accessed September 26, 2018].

Wang, L. et al., 2013. TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *The AAPS journal*, 15(2), pp.395–406. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3675739&tool=pmcentrez&rendertype=abstract [Accessed May 18, 2016].

Wang, R., Lai, L. & Wang, S., 2002. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design*.

Wang, X. et al., 2014. IDrug: A web-accessible and interactive drug discovery and design platform. *Journal of Cheminformatics*, 6(1), pp.1–8.

Waring, M.J. et al., 2015. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, 14(7), pp.475–486. Available at: http://www.nature.com/doifinder/10.1038/nrd4609 [Accessed June 9, 2016].

Warren, G.L. et al., 2006. A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry*.

Warshel, A., 1976. Theoretical Studies of Enzymic Reactions. *Chemical Physics*.

Watts, K.S. et al., 2014. Macrocycle Conformational Sampling with MacroModel. *Journal of Chemical Information and Modeling*, 54(10), pp.2680–2696. Available at: http://pubs.acs.org/doi/10.1021/ci5001696 [Accessed December 4, 2018].

Wen, M. et al., 2017. Deep-Learning-Based Drug–Target Interaction Prediction. *Journal of Proteome Research*, 16(4), pp.1401–1409. Available at: http://pubs.acs.org/doi/10.1021/acs.jproteome.6b00618 [Accessed October 1, 2018].

Wendin, G., 2017. Quantum Information Processing with Superconducting Circuits: a Review. *Reports on Progress in Physics*, 80(10), p.106001. Available at: https://arxiv.org/pdf/1610.02208.pdf [Accessed February 7, 2019].

Wereszczynski, J. & McCammon, J.A., 2012. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. *Quarterly reviews of biophysics*, 45(1), pp.1–25. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3291752&tool=pmcentrez&rendertype=abstract [Accessed February 16, 2015].

Willett, P., 2006. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, 11(23–24), pp.1046–1053. Available at: https://www.sciencedirect.com/science/article/pii/S1359644606004193?via%3Dihub [Accessed September 29, 2018].

Wolber, G. et al., 2008. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today*, 13(1–2), pp.23–29. Available at: https://www.sciencedirect.com/science/article/pii/S1359644607003996?via%3Dihub [Accessed October 2, 2018].

Wu, M. & Hancock, R.E.W., 1999. Improved derivatives of bactenecin, a cyclic dodecameric antimicrobial cationic peptide. *Antimicrobial Agents and Chemotherapy*.

Wu, X. et al., 2015. Molecular dynamics simulation and free energy calculation studies of kinase inhibitors binding to active and inactive conformations of VEGFR-2. *Journal of Molecular Graphics and Modelling*, 56, pp.103–112. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1093326314002095.

Xie, T. et al., 2015. Review of natural product databases. *Cell Proliferation*, 48(4), pp.398–404.

Yagi, Y. et al., 2007. In silico panning for a non-competitive peptide inhibitor. *BMC Bioinformatics*.

Yan, Y. et al., 2017. Protein-Ligand Empirical Interaction Components for Virtual Screening. *Journal of Chemical Information and Modeling*.

Yap, C. et al., 2007. Regression Methods for Developing QSAR and QSPR Models to Predict Compounds of Specific Pharmacodynamic, Pharmacokinetic and Toxicological Properties. *Mini-Reviews in Medicinal Chemistry*, 7(11), pp.1097–1107. Available at: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1389-5575&volume=7&issue=11&spage=1097 [Accessed September 29, 2018].

Yin, J. et al., 2016. Structure and ligand-binding mechanism of the human OX1 and OX2 orexin receptors. *Nature Structural & Molecular Biology*.

Yu, W. et al., 2015. Pharmacophore Modeling Using Site-Identification by Ligand Competitive Saturation (SILCS) with Multiple Probe Molecules. Available at: https://pubs.acs.org/sharingguidelines [Accessed October 2, 2018].

Zhang, L. et al., 2017. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, 22(11), pp.1680–1685. Available at: https://www.sciencedirect.com/science/article/pii/S1359644616304366?via%3Dihub [Accessed October 1, 2018].

Zhang, M.H., Xu, Q.S. & Massart, D.L., 2005. Boosting Methodology for Regression Problems. *Analytical Chemistry*, 77(5), pp.1423–1431. Available at: http://www-stat.stanford.edu/ [Accessed September 29, 2018].

Zhang, S. et al., 2011. A robust high-throughput sandwich cell-based drug screening platform. *Biomaterials*.

Zhao, Z.-Q. et al., 2003. Inhibition of myocardial injury by ischemic postconditioning during

reperfusion: comparison with ischemic preconditioning. *American journal of physiology. Heart and circulatory physiology*, 285(2), pp.H579-88. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12860564 [Accessed October 22, 2018].

Zhong, B. et al., 2017. Rational design of cyclic peptides to disrupt TGF-B/SMAD7 signaling in heterotopic ossification. *Journal of Molecular Graphics and Modelling*, 72, pp.25–31.

Zhong, G.-Q. et al., 2014. Novel functional role of heat shock protein 90 in protein kinase C-mediated ischemic postconditioning. *Journal of Surgical Research*, 189(2), pp.198–206. Available at: https://www.sciencedirect.com/science/article/pii/S002248041400078X?via%3Dihub#bib4 [Accessed October 22, 2018].

Zorzi, A., Deyle, K. & Heinis, C., 2017. Cyclic peptide therapeutics: past, present and future. *Current Opinion in Chemical Biology*.

Zwanzig, R.W., 1954. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics*, 22(8), pp.1420–1426. Available at: http://aip.scitation.org/doi/10.1063/1.1740409 [Accessed September 24, 2018].