ANTIMICROBIAL DRUG REPURPOSING THROUGH MOLECULAR MODELLING

Acquisition, Analysis and Prediction

NHAT PHUONG DO

Doctor of Philosophy

ASTON UNIVERSITY
December 2021

Abstract

ANTIMICROBIAL DRUG REPURPOSING THROUGH MOLECULAR MODELLING

Acquisition, Analysis and Prediction

Nhat Phuong Do

Doctor of Philosophy

Year of submission: 2021

Antimicrobial resistance has sparked unprecedented medical crises around the world, not only increasing the mortality rate but also impacting nosocomial resources. Methicillin-resistant *Staphylococcus aureus* (MRSA) has consistently evaded the available range of antibiotics and is a typical case study for new generation drugs. Drug development has been conventionally suffering from exceedingly high costs and overdrawn timelines. Drug Repurposing can be a solution to alleviate those burdens. Put simply, DR is a mechanism to identify new usages of existing drugs, typically targeted to treat diseases different to the ones that these were initially intended for.

This inherently interdisciplinary research targets to identify the best MRSA drug candidates analysing protein (BIG) data, in the process developing a combination of techniques from stochastic mathematics, statistics and data analytics that can generically identify drug targets from the databank. Structure-based virtual screening was used to repurpose an extensive range of marketed drugs and Phase I/II/III trials. Molecular docking methods were used for virtual screening against MRSA targets based on sequence alignment to match gene sequences against proteins in the Protein Data Bank (PDB). Ligands from the Database of Useful Decoys - Enhanced were docked against MRSA-oriented target proteins using 10 open-source docking programmes for benchmark. The novel consensus scoring methods prove superior to other reported consensus scores in terms of discrimination between decoys and active ligands concerning MRSA drug target identification. The consensus scoring predictions are then applied to docking data between MRSA targets and compounds from the Repurposing Hub to identify a list of potential drug candidates for anti-MRSA treatment.

MRSA is currently an apocalypse across the world with limited prevention and medications. This study provided more potential candidates to help fight against MRSA. The consensus scoring developed in this study can be generically implemented to unlock other antimicrobial drug candidates.

Key words: drug repurposing, Methicillin-resistant *Staphylococcus aureus*, virtual screening, molecular docking, consensus score.

Declaration

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions. The work was done under the guidance of Dr Amit K Chattopadhyay, Department of Engineering and Applied Science, Aston University and Dr Darren Flower, Department of Health and Life Science, Aston University.

Nhat Phuong Do

In my capacity as supervisor of the candidate's thesis, I certify that the above statements are true to the best of my knowledge.

Dr Amit K Chattopadhyay

Date: 31 December 2021

Acknowledgements

To begin, I would like to thank my supervisor, Dr Amit K Chattopadhyay, for his unwavering support of my PhD study and associated research, as well as his patience, inspiration, and vast expertise. His advice was invaluable during the research and writing of this thesis. For my PhD studies, I could not have asked for a greater advisor and mentor.

Apart from my supervisor, I would like to thank my co-supervisor Dr Darren Flower for his insightful remarks and instruction, as well as the difficult question that prompted me to broaden my research to include several perspectives.

I am grateful to Alex Brulo from Aston University's IT Team who was a key help with my computational setup, especially in my initiation days at Aston. I would also like to express my gratitude to Dr Michael Stich and Dr Tomas Johansson for allowing me to attend their Mathematics classes, a subject that until then was more a tool for me, given my original background in pharmacy.

I am genuinely grateful for the financial support provided by Project 911, Ministry of Education and Training, throughout my research. This research would not have been possible without this support.

I am also grateful to have benefitted from Aston University's emergency Covid fund support that was extremely useful to me at a time when I was badly placed in terms of my limited resource.

Last but not least, I would like to express my gratitude to my family: my wonderful wife for wholeheartedly supporting me and taking care of our kid so that I could focus on my studies; my little daughter for bringing me motivations when I needed it the most; my parents and my sister for spiritually supporting me during my studies and life in general.

Contents

List of Figures

List of Tables

List of Publications

1. "Towards Effective Consensus Scoring in Structure-Based Virtual Screening" by Nhat Phuong Do, D R Flower, S Chattopadhyay, and A K Chattopadhyay. Under review at the journal Interdisciplinary Sciences: Computational Life Sciences.

Chapter 1

Reader's Summary of the Thesis Work

ANTIMICROBIAL DRUG REPURPOSING THROUGH MOLECULAR MODELLING

Acquisition, Analysis and Prediction

Supervisors:
Dr Amit K Chattopadhyay (PI) and Dr Darren Flower (CoI)
Aston University, Birmingham UK

Student: Nhat Phuong Do

1 Introduction

Antibiotics are widely regarded as the antibacterial panacea, magic bullets that can treat all forms of bacterial infections thereby saving lives. However, indiscriminate overuse as also natural bacterial immunity against such therapeutics has led to antimicrobial resistance that has sparked unprecedented medical crisis around the world, not only increasing the mortality rate but also impacting nosocomial resources, including other long-termed illnesses. Methicillin-resistant *Staphylococcus aureus* (MRSA) is a poignant case in hand. This bacterium has consistently evaded the available range of antibiotics and is a typical case study for new generation drugs.

Drug development and subsequent manufacturing have been conventionally (wet) laboratory-based, that, apart from the obvious issue of exhausting the chemical space available for targeting new drugs, suffers from three key issues – cost, overdrawn timelines and side effects. These limitations have driven attempts to develop new drugs by repurposing the existing ones, a technology now popularly referred to as Drug Repurposing (DR hereafter). DR is an area of translational biology that identifies new or different therapeutically useful indications for marketed drugs by targeting alternative diseases. Put simply, DR is a mechanism to identify new usages of existing drugs, typically targeted to treat diseases different to the ones that these were initially intended for.

Most drugs have significant off-target activity, thus potential new therapeutic uses should be identifiable for molecules known to be free of toxicity or side effects. Molecules that have passed safety evaluation in Phase I trials but proved ineffective for efficacy reasons in Phase II or Phase III against some other disease can also be repurposed. Successful examples of drug repositioning abound: thalidomide in severe erythema nodosum leprosum; antidepressant Zyban, used successfully for smoking cessation; Parkinson's disease drug apomorphine, now treats erectile dysfunction; even Viagra began as a heart medicine. Repurposing has huge untapped potential for identifying novel, safe, tested, patent-protected medicines.

DR is not new though. It has been traditionally implemented using molecular docking (or simply 'docking') that uses computational algorithms to map the detailed conformation of a molecule with reference to another. A typical example could be docking a protein against a ligand to find possible drug targets. This is a useful method only when implemented across a very wide range of drug targets through multiple (100+) docking programmes but otherwise has low accuracy because of shortcomings in current scoring functions. To alleviate these issues, as also to accommodate attempts to reduce the false positive and false negative rate, combining information from multiple docking programmes has been suggested. This method is called "consensus scoring". In this work, primary docking scores from a small finite number of randomly chosen docking programmes have been statistically combined to avail a much wider protein:ligand mapping space than is accorded by individual molecular docking. For comparison, other consensus scores are also duplicated using the same data. The proposed novel consensus scoring methods prove superior to other reported consensus scores in terms of discrimination between decoys and active ligands concerning MRSA drug target identification.

2 Methodology

Unlike conventional *in silico* virtual screening, the proposed computational DR methodology is based on a biaxial structure: Consensus Scoring. It was independently implemented, results compared, risk validated against standard results from conventional docking platforms and then a set of (probabilistically) highly accurate MRSA drug targets identified.

Preparations for docking

Starting from the Database of Useful Decoys - Enhanced (DUD-E), latter version with additional targets as used in conventional (computational) DR, the following steps were sequentially followed:

- The targets from DUD-E were chosen based on the structural similarity to MRSA targets. After that, the decoys and active ligands from DUD-E set were docked to the corresponding MRSA target. To provide a general overview on evaluation, both receiver operating characteristics and enrichment factor were chosen as metrics to evaluate the performance of docking programmes as well as consensus scores, whereas receiver operating characteristics represents the degree of discrimination between decoys and active ligands with enrichment factor representing the retrieving of true active ligands among top-scored ligands.

- Essential genes i.e. genes encoded for proteins that play a vital part in the survival of organisms were identified and analysed. Database of Essential Genes provides a library of essential genes for *Staphylococcus aureus*. Sequence alignment was used to compare these essential genes against protein structures in Protein Data Bank (PDB). For those hits with a high matching score, the proteins were selected based on the resolution of the structures and the availability of the co-crystallised ligands. For those with moderate matching scores, homology modelling is used to predict the structures of the proteins.

- Repurposing Hub is a library containing candidates for repurposing tasks. After filtering with Lipinski's rule, remaining ligands were retained for docking against MRSA proteins. These ligands were converted to three-dimensional structures using the programme OpenBabel, followed by energy minimisation. Depending on each docking programme, the ligand chemical format can be converted to suit the requirement.

- The docking of ligands against MRSA targets across 10 open-sourced docking programmes: DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, rDock, Smina, Autodock Vina and VinaXB. These programmes were employed for both benchmarking and data acquisition.

CS Scoring Method With the success in the benchmark for MRSA targets, the ligands from Repurposing Hub were docked against MRSA hits from essential gene sequence alignment. Once the docking was done, the docking scores were input in the same fashion as in the consensus score.

A combination of traditional statistical descriptors such as Minimum, Maximum, Mean, Median, Euclidean Distance, Cubic Mean and Deprecated Sum Rank as well as newly developed score Exponential Consensus Rank were used to compare with the proposed novel consensus scores:

$$S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^n$$

$$S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[{S_{i,j}}^n]$$

$$S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{S_i})^n$$

$$S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[(S_i - \bar{S_i})^n]$$

$$S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{SD_i})^n$$

$$S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[(S_i - \bar{SD_i})^n]$$

Here S_c is the combined score, S_i is the docking score of ligands for programmes i = 1, 2, ... 10, $x_{i,j}$ are coefficients of the docking programmes i (DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, rDock, Smina, Autodock Vina and VinaXB) that are the weight factors of those docking outcomes in the combinatorics in the j_{th} iteration, $\bar{S_i}$ is the mean of the set from the programme i, $\bar{SD_i}$ is the standard deviation of the set from the programme i, n represents the combinatorial order real values only (n = 1 implies linear combination). The six equations were iterated over a total of $\binom{29}{9}$ ensembles involving 10 docking programmes, each weighing between 0 and 1, incremented in steps of 0.05 each. S_i represents the arithmetic means of the docking scores of all ligands for the same target for each docking programme used. The rank of active ligands before and after combining were then compared to evaluate the improvement produced by the consensus algorithm.

The metrics used for comparison between individual docking programmes and consensus scores include: median rank, receiver operating characteristic and enrichment factor.

3 Results and Discussion

Single docking programme produced average results (AUCROC from 0.495 to 0.623). This mean the single docking programmes did not provide good discrimination between actives and decoys. Traditional consensus scores did not significantly improved docking performance (AUCROC increased to 0.704 while enrichment factor decreased from 5.5 to 2.1).

The first key outcome of the consensus module is that linearised docking combinatorial scores provide better active ligand ranking than higher-order consensus formulae as some previous proponents of the CS method attempted before. One more finding was that odd-ordered CS combinations (formulae 1a-d) consistently outperform their even ordered counterparts. This means linearised consensus scores were better at discriminating between actives and decoys. The findings also indicate that linear combinations using absolute values in the statistical norm showed the area ratios of the histograms of median ranks obtained from novel consensus models better than using true value (0.648 compared to 0.532). This results in better consistency in the consensus model.

Another benchmark result of the proposed approach is to establish an improvement threshold of the CS scoring methodology, in other words, quantify how many individual docking methods needed. While consensus scoring predictions did initially improve with

the added number of docking inputs, this improvement did not improve when the number of docking programmes continued added. Using 5-6 docking programs will improve the docking power but does not increase running time.

The novel consensus scores using standard deviation produce significant outcomes. The statistics from ROC and enrichment factor for the proposed CS model are consistently higher than the highest values from single docking programmes. This means using combined scores from multiple docking programs can recognise actives from an ensemble of with higher ratio. When using MRSA dataset, the enrichment factor increased from 5.5 to 19.1, which means in the subset, the probability of finding an active was 19 times higher than in the entire ensemble.

4 Conclusions

MRSA is a malignant pathogen that requires expanding research for more cures. There are various strategies to overcome the problem, but drug repurposing is an encouraging approach. Thanks to the availability of a safety profile, drug repurposing can help us to cut down the cost and time optimise resources. Consensus scoring algorithms were investigated using MRSA dataset and ten docking programmes (DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, rDock, Smina, Autodock Vina and VinaXB) leading to the following key conclusions:

- The novel consensus score consistently gives better predictions for active compounds in terms of AUCROC (0.833 to 0.873 compared to 0.623) than conventional (single docking based) *in silico* virtual screening.
- The algorithmic modelling based on Consensus Scores has identified a list of potential MRSA drug candidates (30 candidates for each target) that are now candidates for wet laboratory investigation.
- The 'accuracy plots' establish the strength of the CS method: a) only a handful of docking programmes (5-6 programmes) are required; b) the choice of these docking platforms can be completely random; c) the extant results are more accurate than individual docking.
- The consensus model can be exploited in other virtual screening.

Chapter 2

Introduction

1 Thesis Overview

Antibiotics were widely regarded as the antibacterial panacea, magic bullets that could treat all forms of bacterial infections thereby saving lives. However, indiscriminate overuse as also natural bacterial immunity against such therapeutics has led to antimicrobial resistance that has sparked unprecedented medical crisis around the world, not only increasing the mortality rate but also impacting nosocomial resources, including other long-termed illnesses. Methicillin-resistant *Staphylococcus aureus* (MRSA) was a poignant case in hand. This bacterium has consistently evaded the available range of antibiotics and was a typical case study for new generation drugs.

Drug development and subsequent manufacturing have been conventionally (wet) laboratory-based, that, apart from the obvious issue of exhausting the chemical space available for targeting new drugs, suffered from three key issues – cost, overdrawn timelines and side effects. These limitations have driven attempts to develop new drugs by repurposing the existing ones, a technology now popularly referred to as Drug Repurposing (DR hereafter). DR is an area of translational biology that identifies new or different therapeutically useful indications for marketed drugs by targeting alternative diseases. Put simply, DR is a mechanism to identify new usages of existing drugs, typically targeted to treat diseases different to the ones that these were initially intended for.

Most drugs have significant off-target activity, thus potential new therapeutic uses should be identifiable for molecules known to be free of toxicity or side effects. Molecules that have passed safety evaluation in Phase I trials but proved ineffective for efficacy reasons in Phase II or Phase III against some other disease can also be repurposed. Successful examples of drug repositioning abounded thalidomide in severe erythema nodosum leprosum; antidepressant Zyban, used successfully for smoking cessation; Parkinson's disease drug apomorphine, now used to treat erectile dysfunction; similarly, Viagra began as a heart medicine. Repurposing has huge untapped potential for identifying novel, safe, tested, patent-protected medicines.

DR is not new though. It has been traditionally implemented using molecular docking (or simply 'docking') that uses computational algorithms to map the detailed conformation of a molecule with reference to another. A typical example could be docking a protein against a ligand to find possible drug targets. This is a useful method only when implemented across a very wide range of drug targets through multiple (100+) docking programmes but otherwise has low accuracy because of shortcomings in current scoring functions. To alleviate these issues, and also to accommodate attempts to reduce the false positive and false negative rate, combining information from multiple docking programmes has been suggested. This approach was called "consensus scoring". In this work, primary docking scores from a handful of randomly chosen docking programmes have been statistically combined to avail a much wider protein:ligand mapping space than accorded by individual molecular docking platforms. For comparison, other consensus scores were also duplicated using the same data. The proposed novel consensus scoring methods proved superior to other reported consensus scores in terms of discrimination between decoys and

active ligands concerning MRSA drug target identification.

The rest of this chapter describes the discovery of antibiotics and the rise of antibiotics resistance with a focus on a molecular view which is the principle for this study. The spread of antibiotic resistance has led to a severe loss of lives and expense. Meanwhile, the treatment and prevention seem not commensurate when the resistance rate is fast and the productivity of drug discovery and development has been flattened. The main reason is that drug discovery and development is a costly and lengthy process. Drug Repurposing is a promising strategy when cutting down the cost and time when utilising the pharmacological data available.

Chapter 2 summarises the existing DR approaches and the reason why this study chooses virtual screening for DR. Then it focuses on the main backbone of structure-based virtual screening, the molecular docking methods. The fundamentals of docking programmes are summarised and also the limitations of the existing docking programmes. This lead to the attempts to incorporate data from multiple docking programmes to improve docking competence in this study.

Chapter 3 carries the main works of this study. It describes the methods used to build the *S. aureus* protein structures from essential genes. It also describes how to obtain the collection of compounds to screen against MRSA targets using docking programmes. Before the main virtual screening is carried out, the docking programmes are benchmarked using a library of decoys and actives. Also, a number of consensus scores are proposed and compared with traditional consensus scores to prove their superiority. Finally, the best version of the novel consensus score is applied to the main dataset of docking to identify compounds with a high probability of being active against MRSA targets.

Chapter 4 features the main findings and the data obtained. It presents the protein structure from sequence alignment as well as homology modelling. It also confirms the performance of chosen docking programmes after benchmarking. Most importantly, the consensus scores proposed has been proved to improve the docking capability. While many consensus scores use sophisticated methods but the results dependent on the nature of target proteins, this study seeks to combine the docking score in a simple way but with remarkable improvement. Finally, potential compounds for DR are chosen by applying the consensus score to the docking dataset.

Chapter 5 listed the candidates after applying chosen model to the docking data between compounds from Repurposing Hub and MRSA proteins. The lists contained 30 ligands for each MRSA target, as a results of 0.5% cut-off from a total of 5902 ligands.

Finally, conclusions are drawn and how the findings in this study can be applied in Chapter 6. It also discusses the limitations of this study and the suggestions for future research.

2 Antibiotic Resistance Apocalypse

2.1 Role of Antibiotics

Antibiotics belong to a class of drugs that can act on microorganisms. These are drugs with antimicrobial activities used in the treatment and prevention of infections caused by bacteria. To grow and divide, bacteria need to parasitically dominate in the human or animal body. They consume essential substances available inside the body and excrete toxic metabolites, causing disorders and diseases to humans, sometimes death. The effects of antibiotics are demonstrated by the ability to cease or inhibit the cell growth of bacteria but they do not affect viruses. Their mechanism includes inhibition of the wall synthesis, inhibited nucleic synthesis and competition with the essential substances for the growth of bacteria. Before the discovery of the first antibiotics in the 20th century, treatment for infection was no more than traditional medicine (Lindblad, 2008). Therefore, it led to a search for a cure to decrease death by wound infections from natural products in the 19th century.

Many attempts have been made to help fight against the deadly infections caused by bacteria. Arsphenamine was synthesised by Alfred Bertheim in 1907 (Williams, 2009), as

an antiseptic agent. The first antibiotic, penicillin, was discovered by Alexander Fleming in 1928 (Ligon, 2004) and employed pure penicillin for the first time to treat streptococcal meningitis in 1942 (Fleming, 1943). Thanks to the discovery, Alexander Fleming, Howard Florey and Ernst Boris Chain shared the Nobel Prize in Physiology or Medicine in 1945. After penicillin, many antibiotics were discovered and synthesised. The first sulphonamide Prontosil was developed in 1932 by Bayer Laboratories, (Aminov, 2010) or streptomycin, a first-line anti-tuberculosis drug, was developed in 1943 by Selman Waksman, for which he received Nobel Prized in Medicine in 1952 (Woodruff, 2014). The late 19th century witnessed the outbreak in discovery with hundreds of synthesised antimicrobial drugs. Antibiotics have brought a new age in the fight against deadly infections.

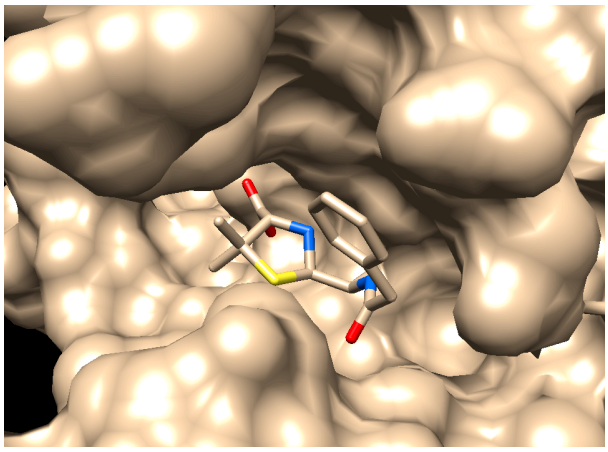
2.2 Molecular Mechanism of Antibiotics

Antibiotics inhibit the growth or cause the death of bacterial cells by affecting vital biological processes. To exert such a bactericidal effect, the antibiotic molecules need to bind to a target within the bacterial cell. For instance, the cell wall of Gram-positive bacteria has the typical characteristics of gram-positive bacterial cell walls. The cell wall is a firm external structure that encloses the cellular membranes and prevents cell blast due to osmotic pressure. It appears like a reasonably thick (approximately 20 to 40nm) homogeneous layer under the microscope (Kim et al., 2015). The cell wall has been known to be composed of polysaccharides and peptides the peptidoglycan for a long time (Salton et al., 2002). The polysaccharide backbone contains N-acetylglucosamine and N-acetylmuramic acid, with a five-aminoacid peptide, called stem peptide, attached to acetylmuramic acid (Sidow et al., 1990). The chains are cross-linked by a group of five glycines, bonded to the lysine (location 3) on one stem peptide and the alanine (location 4) on another, forming a mesh-like structure around the cell (Labischinski, 1992). The peptidoglycan contains pentaglycine as a distinctive characteristic in Staphylococci, providing both toughness and flexibility to endure strong intracellular and external pressure.

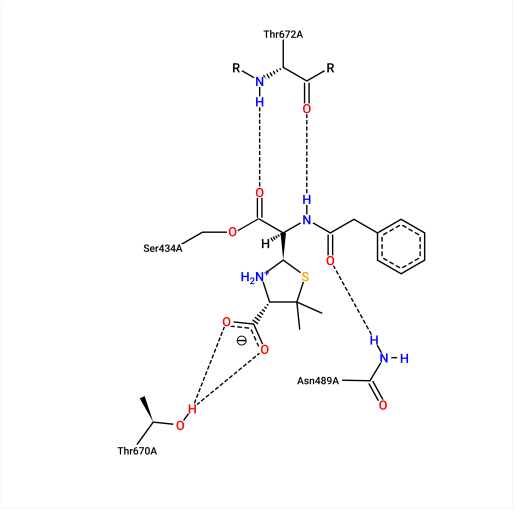
Transpeptidases, which are termed penicillin-binding proteins (PBPs) (also call DD-transpeptidase), catalyse the process of cell wall cross-linking (Labischinski, 1992). To the current knowledge, although there are eight types of staphylococcal PBPs (Templin and Höltje, 2013), only four types of PBP are widely documented. There is evidence that PBP-1 is vital for staphylococci survivability in the presence of β-lactams (Beise et al., 1988; Reynolds et al., 1988). PBP structure contains one domain for transpeptidation (cross-linking). The β-lactam ring of penicillin blocks the transpeptidation region of PBPs, inhibiting the cross-linking process by resembling the terminal alanine link of the stem peptide. The activity of the enzyme is inhibited when bounded to β-lactam and therefore can no longer catalyse the synthesis of the cross-links. As a result, the cell wall becomes weak lacking cross-linking of the peptidoglycan, leading to some intracellular contents leaking out and the cell ceasing to grow (Giesbrecht et al., 1998). Figure 2.1 provides a visual view of how a penicillin molecule binds to a penicillin-binding protein.

2.3 Antibiotic Resistance

As a result of an evolutionary process that enables living objects to self-sustain, bacteria have developed a resistance mechanism to survive antimicrobial agents. Bacteria have obtained resistance via many ways: changes in the permeability of cell membrane, secretion of enzymes to destroy the structure of drugs, creation of a system to pump out the drug molecules, changes in biosynthesis, changes in the protein structures which are receptors for antibiotics. Antimicrobial resistance decreases the success rate of treatment, increase the cost, the hospital duration. Sometimes inpatients suffer hospital infections more than the initial reasons of admission. The first case of antimicrobial resistance against penicillin was reported just 4 years after its mass production (Spink and Ferris, 1947). Methicillin resistance *Staphylococcus aureus* (MRSA) was first filed in Britain in 1961 (Barber, 1961; Jevons, 1961) and has become one of the most common reasons for hospital infections. Other antibiotics including penicillins, cephalosporins, fluoroquinolones, have been also



(a)



(b)

Figure 2.1: The depiction of penicillin G (open form) in the active domain of the penicillin-binding protein and the intermolecular interaction.The code of protein from Protein Data Bank: 3UDI. 2.1a) penicillin G molecule (open form) is surrounded by the protein surface, generated by Chimera (Pettersen et al., 2004). 2.1b) The intermolecular interactions between penicillin molecule and protein 3UDI were captured using PoseView (Stierand et al., 2006).

reported to be resisted by increasing numbers of bacterial strains.

The mechanisms underlying decreased permeability in bacteria differ between Gram-positive and Gram-negative bacteria. Penicillin resistance in Gram-positive bacteria is caused by alterations in the cell wall. In *S. aureus*, for example, the creation of an extra PBP, known as PBP2a, with a decreased affinity for penicillin and β-lactam antibiotics is what causes the resistance (Karaman et al., 2020). Mutations in the structure and quantity of porins cause resistance in Gram-negative bacteria (Breijyeh et al., 2020). The number of porins is lowered in bacteria such as *Pseudomonas aeruginosa*; nevertheless, altered porins such as non-specific porins that cannot transport penicillin are found in bacteria such as *Escherichia coli*, *Klebsiella pneumonia* and *Enterobacter* species (Pagès et al., 2008).

2.4 Worldwide Antibiotic Resistance

At the present resistance scheme, 23,000 patients were reported dead due to antibiotic resistance in the US (CDC, 2013) and 25,000 deaths in Europe were recorded for the same cause (European Centre for Disease Prevention and Control, 2015). According to the annual report of the World Health Organization, seven bacteria have resisted common antibiotics at the concerned level: *Escherichia coli*, *Neisseria gonorrhoeae* and *Klebsiella pneumoniae* have developed resistance to 3rd generation cephalosporins; *Staphylococcus aureus* has been resistant to β-lactams; *Staphylococcus pneumonia* and *Shigella* have been resistant to penicillin and non-typhoidal *Salmonella* resistant to fluoroquinolones (WHO, 2014). Report of Antimicrobial Resistance Surveillance in Europe also shows the same theme of these bacteria. *K. pneumonia* have gained 7.4% in resistance to carbapenem in 2014. *P. aeruginosa* was reported with an average resistance of more than 10% to fluoroquinolones and aminoglycosides. *P. aeruginosa* also shows resistance to carbapenem at a rate ranging from less than 10% to more than 50%. This range shows that different countries have various resistance levels due to their medical control (European Centre for Disease Prevention and Control, 2015). According to the WHO report, the numbers of infections and deaths related to antibiotic resistance were 2,036,100 and 22,618 in the US in 2013, 87,751 and 38,481 in Thailand in 2012, 386,100 and 25,100 in Europe in 2007 (World Health Organization, 2017). Antibiotic resistance has led the community to a point where medications are less effective against the emerging strains.

3 Methicillin Resistant *Staphylococcus aureus* (MRSA)

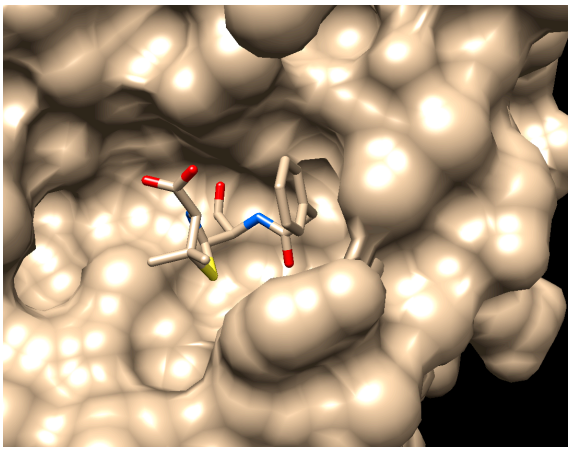
3.1 Biology of MRSA

Staphylococcus aureus is a Gram-positive, round-shaped bacterium that is commonly present in the upper respiratory tract and on the skin. *Staphylococcus aureus*, a Gram-positive coccus belonging to the *Micrococcaceae* family, has cells that appear individually or in pairs, tetrads and unique irregular "grape-like" clusters if dividing cells do not split. The name "staphylococcus" comes from the Greek language "grapes" and the colour of *Staphylococcus* colonies is described by the Latin word "aureus", which means "gold". *S. aureus* colonises usually exposed skin areas and the upper respiratory tract, especially the nasal airways, in humans. Healthy people are usually unaware that they have the staphylococcal carriage, but they can get mild skin diseases like blisters and ulcers. *S. aureus*, on the other hand, is an aggressive bacterium that can produce more severe infections under certain circumstances. *S. aureus* intrusion is typically seen in burns and post-operative infections, where the toxin from the bacteria can produce toxic shock, which causes fever, nausea and, in certain circumstances, fatality. Pneumonia, mammary gland infection, skin infections, bone infections endocarditis and bacteraemia are all infections caused by *S. aureu. S. aureus* can potentially induce food poisoning as a result of the development of enterotoxins. If left untreated, *S. aureus* can lead to pneumonia and bloodstream infections, both of which can be fatal. *S. aureus* can access the underlying tissues or the circulation when the epidermal and mucosal barriers are compromised, such as by chronic skin diseases, wounds or surgical intervention. *S. aureus* infection is especially dangerous for people who have intrusive medical equipment (such as peripheral and central catheterisation) or have weakened immune systems (Lowy, 1998).

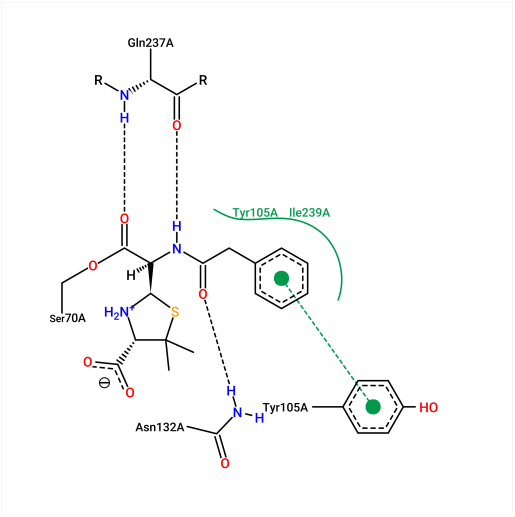
Before the production of penicillin in the early 1940s, the fatality rate of those infected with *S. aureus* was around 80% (Skinner and Keefer, 1941). Penicillin helped to fight against *S. aureus* after the production of penicillin in 1940 (Tan and Tatsumura, 2015). However, the first case of penicillin-resistant *S. aureus* isolate was reported in a hospital in 1942, not too long after penicillin was approved for medical use (Spink and Ferris, 1947). Penicillin-resistant *S. aureus* strains were widely discovered in the population later on. Benzylpenicillin (penicillin G), a lactam antibiotic, was used to treat infections caused by *S. aureus* before the 1950s, but by the late 1950s, the resistance of *S. aureus* variants to benzylpenicillin were already creating an alarming situation.

Methicillin-resistant *Staphylococcus aureus* (MRSA), a type of *Staphylococcus aureus* that is unsusceptible to antibiotics in the β-lactam class. MRSA was first documented in Britain in 1961 (Jevons, 1961), shortly after the use of methicillin became commonplace. β-Lactamase, an enzyme that inactivated the lactam antibiotics, was produced by resistant variants. The goal was to develop penicillin analogues that were resistant to β-lactamase hydrolysis. The synthesis of methicillin, which had the phenyl ring of benzylpenicillin attached with the two methoxy groups, was accomplished in 1959. Two methoxy groups created spatial obstacles around the amide link, which reduced the attraction of the amide link for staphylococcal β-lactamases. Figure 2.2 provides a visual view of how a penicillin molecule binds to a β-lactamase. Unfortunately, methicillin-resistant *S. aureus* (MRSA) strains were recorded not too long after methicillin clinical application. Resistance was owing to the expression of an extra penicillin-binding protein (PBP2a) obtained from another organism that was resistant to the action of antibiotic rather than β-lactamase formation (Chambers, 1997). Methicillin was once commonly used, but due to its toxicity, it is no longer licensed for human use and has been mainly replaced by more robust β-lactam antibiotics such as oxacillin, flucloxacillin and dicloxacillin. Despite this, the term methicillin-resistant *Staphylococcus aureus* is still being in use. MRSA was accountable for clinical outbreaks in many regions around the world in the dozen years following its observation (Chambers and DeLeo, 2009).

Staphylococcal resistance to penicillin is caused by the synthesis of penicillinase (a member of the β-lactamase family), an enzyme that breaks down the β-lactam ring of the penicillin molecule, making the antibiotic impotent. Methicillin, nafcillin, oxacillin,



(a)



(b)

Figure 2.2: The depiction of penicillin G (open form) in the active domain of the β-lactamase and the intermolecular interactions.The code of protein from Protein Data Bank: 1GHP. 2.2a) penicillin G molecule (open form) is surrounded by the β-lactamase surface, generated by Chimera (Pettersen et al., 2004). 2.2b) The intermolecular interactions between penicillin molecule and β-lactamase 1GHP were captured using PoseView (Stierand et al., 2006).

cloxacillin, dicloxacillin and flucloxacillin are penicillinase-resistant β-lactam antibiotics that may withstand degradation by penicillinase. To undertake cross-linking processes, the so-called PBP 2a, which is responsible for staphylococcal methicillin resistance, appears to require appropriate pentaglycine interpeptide bridges (Kopp et al., 1996). When cultivated in the vicinity of β-lactam antibiotics, mutant strains with shorter interpeptide bridges (mono-, di- and triglycine peptides) revealed dramatically reduced resistance and cross-linking (Ubukata et al., 1989).

The methicillin resistance leads to the remaining choice of vancomycin in the treatment of *S. aureus* infections at that point. However, in the 1990s, *S. aureus* strains with reduced susceptibility to vancomycin were also reported (Daum et al., 1992; Hiramatsu et al., 1997; Paterson, 1999). Due to the cross-resistance, now MRSA is also resistant to amoxicillin, oxacillin and other common antibiotics in the cephalosporin group. If the rapid development of antibiotic resistance remains, the possibility of MRSA evolving resistant to all antibiotics continues growing, making MRSA a more serious epidemiological threat.

3.2 MRSA Threat

With the acquired antibiotic resistance, MRSA has become a huge threat to the community. The WHO considered MRSA as a "major cause of morbidity and mortality worldwide" while the CDC considered MRSA at a serious level and high priority for the Public Health Agency of Canada (World Health Organization, 2017).

In the United States, an estimated 2.5 million people were found infected with MRSA in 2005 (Graham et al., 2006). In the United States health system, 75% of bacterial infection cases were due to invasive MRSA infection (Liu et al., 2011). Normal patients paid US$ 29,455 while patients with Methicillin-susceptible *Staphylococcus aureus* infections paid US$ 52,791 and patients with MRSA surgical site infections paid US$ 92,363 in 2003 (Engemann et al., 2003). According to the US *Centers for Disease Control and Prevention* (CDC), there were 80,461 invasive MRSA infections and 11,285 MRSA-related deaths each year (CDC, 2013). In the United Kingdom, the proportion of isolated samples resistant to MRSA increased from 2% in 1990 and 1991 to a record of 43% in 2002, with a minor dip in 2004 (Johnson et al., 2005). The number of deaths in the United Kingdom due to MRSA was estimated to be around 3,000 per year in early 2005 (Johnson et al., 2005). In Europe, MRSA prevalence varied greatly with percentages ranging from 0.9 per cent to 56.0 per cent. The EU/EEA population-weighted mean MRSA percentage decreased dramatically was 17.4% in 2014 (European Centre for Disease Prevention and Control, 2015).

In the 2013 annual report of the US Centre for Disease Control and Prevention, *Clostridium difficile* was held responsible for the death of 14,000 patients out of 23,000 deaths by antibiotic resistance. Bacterial strains with substantial resistance that cause serious diseases were also listed. Several bacteria were also named in ECDC and WHO reports such as *Enterococcus*, *P. aeruginosa*, non-typhoidal *Salmonella*, *S. pneumonia* (CDC, 2013). MRSA was also listed as one of the most dangerous antibiotic-resistant infections for humans, according to the US Centers for Disease Control and Prevention (CDC) (CDC, 2013). Besides, the report indicated multi-drug-resistant tuberculosis and MRSA became less urgent owing to intensive control and prevention. However, these pathogens remained a threat to the community. In other report, antibiotics infections cost from $10,500 (2004 US dollar) to $111,000 (2006 US dollar) for each patient with transplant. This cost accumulated up to $17 billion for hospital infections (Klevens et al., 2007). These numbers were calculated for developed countries with well-administered health programmes. In developing countries where this problem is not sufficiently addressed, the antibiotic resistance picture is likely to be more gloomy and unpredictable. 10 million deaths were predicted every year around the world with ca US$100 trillion expected to be invested in health prevention regimes by the year 2050 (O'Neill and Grande-Bretagne, 2014).

In particular, Methicillin-resistant *Staphylococcus aureus* (MRSA) is one of the most current concerns. According to the CDC, MRSA occupied 8% cases of hospital infections in 2006 and 2007 (Hidron et al., 2008). Infections by MRSA have increased during the last ten years and in 2011 there were more than 11,000 deaths related to MRSA (CDC, 2013). In Britain, deaths by MRSA infections were calculated to be 3,000 in 2005, which turned out to be a major issue to be debated in the general election in the same year (Koteyko et al., 2008).

Methicillin-resistant *Staphylococcus aureus* is expected to exist side-by-side with humans for the time being. Although the number of hospital cases related to MRSA decreased from 2012 (401,000 cases) to 2017, it remains high (323,700 cases). Patients with injection exposed 16 times higher infected with MRSA (World Health Organization, 2017). Despite numerous attempts to batter the problem, the biomedical scientific community is advised to pursue a variety of MRSA-related research, including finding new medications. MRSA was classified into "High priority" group by WHO (World Health Organization, 2017), along with other multi-drug-resistant micro-organisms, which demands increased endeavour in the discovery and development of new antibiotics and novel prophylactic strategies.

3.3 Anti-MRSA Antibiotic Research

MRSA infection requires immediate treatment and any postponement might be lethal. Antibiotics can be administered by intravenous, oral or a combination of both routes, depending on the conditions of the patient. Current antibiotics that are effective against MRSA include clindamycin, daptomycin, linezolid, quinupristin-dalfopristin, rifampin, telavancin, tetracyclines, trimethoprim/sulfamethoxazole, vancomycin (Liu et al., 2011). Nonetheless, some antibiotics have been reported to be unsusceptible to MRSA. For instance, some new MRSA strained were reported resistant to vancomycin and teicoplanin (Sieradzki and Tomasz, 1997; Schito, 2006).

Due to the rapidly increasing resistance against antibiotics, strict criteria are set out to restrict the resistance. Antibiotics must be strictly prescribed to confirmed infections, as well as narrow-spectrum antibiotics, are prescribed with priority. Besides, sufficient doses must be given from the beginning rather than increased doses. Thanks to intensive control, MRSA has not developed into an epidemic. However, prevention is not enough to keep down the rate of resistance and demand for new antibiotics remains high, especially to treat MRSA infections.

Looking back at the antibiotic discovery timeline from the 1980s to 2010s, the general trend is a decline in the number of newly approved drugs. During the period 1980-1984, 17 new antibiotics were brought to the market and that quantity is 12 for the 1985-1989 interval. However, the number of new antibiotics from 2005 to 2007 and from 2008 to 2011 remained 2, which mean approximate one antibiotic was approved every year (Bassetti

et al., 2013).

Substantial efforts are in line to identify and develop new antibiotics. A range of approaches, involving both conventional laboratory-based experiments as also computerised tools (Artificial Intelligence) is in line. A conventional approach is to screen a large database of compounds with antibacterial potential. After that, intensive pre-clinical and clinical trials will continue to filter the actual drug with relative efficacy. This conventional approach takes years to bring a new drug to the market, some drugs even take up to 17 years (Ashburn and Thor, 2004). Another approach is extraction and derivatives synthesis from natural products (Wright, 2014; Genilloud, 2014; Takano et al., 2012; Pitscheider et al., 2012). Nonetheless, this approach requires knowledge of natural products with antimicrobial in addition to intensive laboratory works to bring a new drug to the clinic. A combination of these approaches can shorten the timescale and cost of new antibiotics (Pereira et al., 2015; Macherla et al., 2013; Wong et al., 2012). However, the limit of natural products does not allow much search space for this combination. There are several minor approaches such as biotechnology with antimicrobial peptide (Gomes et al., 2014) or quorum sensing (Naik and Mahajan, 2013).

Since MRSA is one of the substantial causes of nosocomial infection, it requires huge attention in finding new remedies. Although the Infectious Diseases Society of America and the British Society for Antimicrobial Chemotherapy has provided a strategy for a prescription for MRSA, it is essential to discover new medicines. The list of drugs on the pipeline has been reviewed and concluded not encouraging (Kurosu et al., 2013; Liu et al., 2011; Nathwani et al., 2008). The conventional approach using combinatorial chemistry and the biological assay was supplanted by modified techniques (Thomas et al., 2008; Fletcher et al., 2007; Nicolaou et al., 2001; Wilkening et al., 1999; Ratcliffe et al., 1999). Treatment using phages were proposed and still needs further research (Kurosu et al., 2013). New macromolecules were also discovered to be potential for MRSA treatment (Lau et al., 2015a; Tomoda, 2016). Even the genome was investigated for antibacterial activity (Chu et al., 2016). However, these approaches need further investigation and it will take some time to bring new results.

4 Drug Repurposing (DR) or Drug Repositioning

4.1 Traditional Drug Discovery

Traditional drug discovery and development mainly relies on the outcome from high throughput screening, which is an automated procedure to evaluate a large library of substances for a certain biological target. With the advances in computational power and chemoinformatics, virtual screening became an integral of the process. Nonetheless, the drug discovery and development processes still consume a massive amount of time and cost while the efficiency is low. In general, the drug discovery and development processes can take up to 13.5 years (Paul et al., 2010) and more than 2.0 billion dollars (Paul et al., 2010; Adams and Brantner, Apr) for a drug to go from scratch to the counter.

In a series of articles about research and development costs spanning over a decade, DiMasi and his colleagues have estimated the cost of drug discovery and development. In 1991, he took a survey of 93 randomly chosen drugs from 12 pharmaceutical companies. In this survey, the cost of unsuccessful drug candidates was also linked with the cost of approved drugs. The expense to bring the drug to the market was estimated at $231 million (1987 US dollars) (DiMasi et al., 1991). The clinical trial cost for each drug was approximately $93 million (DiMasi et al., 1995). In an updated study, this number was increased to $802 million (DiMasi et al., 2003). In 2016, a new study report the amount of $2558 million (2013 US dollars) to discover and develop a new drug and the overall success rate was 11.83% (DiMasi et al., 2016). Other investigators also pointed out similar figures. Grabowski estimated the costs for each new drug in the late 1990s had increased more than six times compared to the drug in the 1970s ($802 million versus $138 million) (Grabowski, 2011). In an independent study, Adam and Brantner estimated the expenses range from $500 million to $2000 million, depending on the category and company (Adams

and Brantner, Apr). In another more detailed study, the average expense each year on drugs in human clinical trials was estimated approximately $27M, with $17M in Phase I, $34m in Phase II and $27m in Phase III of the human clinical trials (Adams and Brantner, 2010). Wouters also pointed to the cost of a new drug as one ranging from $314 million to $2.8 billion (Wouters et al., 2020). Another study also indicated a new drug took on average approximately 13.5 years to reach the market and the cost of a new drug ranged between approximately $0.9 billion and $2.7 billion (Paul et al., 2010; Mathieu, 2008).

Despite the massive amount of time and cost, the productivity of drug discovery and development tends to decline. In 2007, 19 new drugs were approved by the FDA, which is the lowest number since 1983 and the number slightly increased to 21 in 2008 and 24 in 2009 (Paul et al., 2010; Mathieu, 2008). In 2010, DiMasi estimated the approval probabilities for each clinical stage in the development process. For all compounds, the average clinical success ratio was 19% and only 16% for self-emerged compounds (DiMasi et al., 2010). Meanwhile, the rate for a drug to be successfully launched was 11.8% for small molecules (DiMasi et al., 2010, 2016). This means out of every nine drugs investigated, just one drug successfully reached approval and other drugs had to stop along the road. In another study by Ashburn, the success rate was estimated to be less than 10% (Ashburn and Thor, 2004).

Therefore, the traditional drug discovery and development process is a lengthy and pricey expedition, also involving substantial invasive treatments or therapeutic regimes even at development stages. On the contrary, Drug Repurposing (DR) could save up to millions of dollars (Mullard, 2012, 2014) for each new drug since the existing drugs have been fully investigated pharmacology and toxicity profile. Even drugs that dropped out at late stages have a certain safe profile after qualifying through phase I clinical trials. In addition, DR can substantially reduce the timeline of drug development from ca two decades to less than 5 years. In fact, aided by the new line of Machine and Deep Learning approaches, the timeline could be further reduced. And this is not restricted to therapeutic discovery alone, the same approach could well be implemented in vaccine discovery. A typical case in hand is the range of Covid-19 vaccines that were made available in less than a year's time which was only possible through computer modelling (Machine/Deep Learning) approaches. Last but not least, DR can help patients suffering from rare diseases, which are not generally the priority of pharmaceutical companies. As a consequence, Drug Repurposing has now assumed a frontline status in drug discovery and development strategies.

4.2 Drug Repurposing (also called Drug Repositioning)

Apart from the huge cost and time involved, conventional drug discovery has a very high attrition rate as well. According to the National Institute of Health (NIH) estimates, ca 80-90% drug candidates get rejected at the Phase I/II/III trial stages, even before market testing (Waring et al., 2015). Although rejected after Phase I/II/III clinical trials, these potential drug candidates have by then already obtained their pharmacokinetic and safety profiles. Both conventional, *i.e.* laboratory-based and computerised molecular docking based DR benefit from such abundant pharmaceutical data in lining up prospective drug candidates from single or assayed molecules.

Serendipitous discovery of off-target indications of many drugs lead to the appeal to explore the currently approved repository and also the disregarded compounds. There is no official definition for drug repurposing. In the academic publication, drug repurposing can be referred to as "drug repositioning", "drug redirecting", "drug reprofiling" or "therapeutic switching". Some authors have attempted to introduce various definitions, varying from simple to detailed statements. For example, Dudley et al. define drug repurposing as "finding a new use for an existing drug" (Dudley et al., 2011a). Endeavouring to give a full definition, Ashburn and Thor described drug repurposing as "the process of finding new uses outside the scope of the original medical indication for existing drugs" (Ashburn and Thor, 2004). The terms "drug repurposing" and "drug repositioning" have been most interchangeably used in academic articles (Ashburn and Thor, 2004).

Although the simple term describes "finding new uses for existing drugs", drug repositioning is not only limited to approved drugs but also include active substances dropped out of the clinical trials and drugs withdrawn from the market due to severe adverse reactions. This means the substances that have not entered the clinical phases or discarded should be excluded.

4.3 Polypharmacology/ Drug Promiscuity

Drug repurposing or finding a new indication for an existing drug is possible and widely applied, thanks to the underlying principle of polypharmacology or drug promiscuity. Traditional remedies derived from plant and animal sources have been used from time immemorial, based on individualised therapeutic evidence provided by persons suffering from a variety of ailments (Schmidt et al., 2008). Such nature-based pharmaceutical products extracted from active compounds have been there for centuries in most older civilisations but the process of lining these up for industrial level medicinal production started just over a century ago (Ban, 2006). The advances in analytical chemistry and pharmacology led to the relationship between active compounds, proteins and diseases. Over time, these proteins have been acknowledged as "pharmaceutical targets" for active compounds and the quest to discover new drugs moved from time-consuming and fortuitous phenotype-based research to more oriented and governable (Brown, 2007).

Drug development efforts have traditionally aimed to create candidates that are both extremely selective and effective for a certain biological target. Based on the "one drug-single target" assumption, the main goal of drug discovery for years was to identify highly selective (and of course, highly potent) against the biological targets of interest. In addition, owing to the limit of resources at those times, the investigation to unravel indications for small-molecule compounds led to incomplete drug's profile (Mestres et al., 2008). Recently, numerous public and private actions to acquire and archive drug-target interaction statistics in bibliographical assets have made major contributions to changing this skewed notion of pharmacological selectivity. Selective medicines are now widely acknowledged to be the minority rather than the standard, with a large quantity of biologically active compounds interacting with multiple proteins. Consequently, the phenomenon of one drug exposing non-selective interaction with multiple proteins is being reviewed in a more holistic view. It has been acknowledged that the majority of medicines work by regulating several targets and pathways. The philosophy of drug design has been shifted from "one drug - one target" to "one drug - multiple targets", termed as polypharmacology (Hopkins, 2008; Apsel et al., 2008; Hopkins, 2009; Briansó et al., 2011; Simon et al., 2012; Paolini et al., 2006).

The term "polypharmacology" was first used by Kenny et al. in 1997 (Kenny et al., 1997) to describe the non-selectivity of indoramin (poly + pharmakologos: the multiple knowledge of medications). Polypharmacology is also referred to as functional promiscuity, cross-reactivity, poly-reactivity, poly-specificity or multi-specificity. At present, the definition of polypharmacology is recognised and broadly accepted. Another term that is also used interchangeably is "drug promiscuity". The difference between "polypharmacology" and "drug repurposing" is that polypharmacology is a concept that describes a drug that can bind to multiple targets while drug repurposing is an application of polypharmacology to find new usage for existing drugs. To exert a biological effect, a drug molecule needs to bind to a specific location called an active site within the target or protein. Only drugs that structurally match a cavity can form intermolecular interactions. However, observation showed that one drug could bind to other targets at other sites rather than at the expected site. This is also the cause of the side-effects of drugs besides the main indication. The most common adversity is the side-effects, due to the interaction of drugs with unexpected targets. On the other hand, it provides an open probability to unveil more indications for existing drugs (drug repurposing) and enhanced efficacy (Mencher and Wang, 2005). Therefore, polypharmacology can be both, a blessing and a curse.

There have been numerous attempts to understand the rationale of polypharmacology. Most studies exploit the relationship between the extent of polypharmacology and the

properties of drugs and targets. There are inconsistent conclusions about the correlation between drug promiscuity and molecular weight. Several found a weak relationship (Azzaoui et al., 2007) while others concluded inverse correlation (Hopkins et al., 2006; Morphy and Rankovic, 2007; Mestres et al., 2009; Gleeson et al., 2011). Nonetheless, some works conclude no clear relationship (Leeson and Springthorpe, 2007; Peters et al., 2009). Regardless of whatever relationship between molecular weight, those aforementioned studies found a weak correlation between drug promiscuity and drug hydrophobicity. One possible explanation is that dug molecules accumulate at the phospholipid layer due to the non-selective nature of hydrophobic intermolecular interaction, leading to interactions with signalling molecules (Leach and Hann, 2011). However, drug promiscuity appears to have a weak correlation with drug flexibility and binding site similarity (Haupt et al., 2013).

Attempts were made to search for promiscuous drugs, particularly for complex illnesses (Hopkins, 2008). This can be undertaken through a variety of ways targeting to discover new therapeutic targets and chemical pathways (Xie et al., 2012), varying from genome-wide association research (Sanseau et al., 2012), gene expression information (Dudley et al., 2011b) and networks (Pujol et al., 2010; Kalinina et al., 2011; Daminelli et al., 2012) to structural methods (Kinnings et al., 2010; Haupt and Schroeder, 2011). Alignment methods (Esther et al., 2008; Xie et al., 2009; Konc et al., 2012) and alignment-free techniques, such as employing fingerprints (Schalon et al., 2008; Liu et al., 2011), are two types of structural binding pocket comparison strategies. The latter has the benefit of successfully discovering even distant similarities but does not present an aligned structure. Another application to find off-targets of the existing compounds is called target fishing (Patel et al., 2015).

Polypharmacology is now heralded as the new dawn of drug discovery (Yildirim et al., 2007; Hopkins, 2008; Durrant et al., 2010; Oprea and Mestres, 2012; Boran and Iyengar, 2010a,b; Xie et al., 2012). A single drug interacting with several targets of a single disease pathway, or a single drug interacting with multiple targets related to multiple disease pathways, are examples of polypharmacology. A number of drugs are notable for their multi-targeting capabilities, even though their development were serendipitous. Aspirin, for example, is commonly used as a painkiller to reduce mild pains or as an antipyretic to relieve fever (Knox et al., 2011); it also has anti-inflammatory effects and is used for rheumatoid arthritis (Simpson et al., 1966), pericarditis (Berman et al., 1981) and Kawasaki diseases (Durongpisitkul et al., 1995). It has also been used to prevent transient ischemic attacks (Grundmann et al., 2003), strokes, heart attacks (Amory and Amory, 2007), miscarriage (Daya, 2003) and recently cancer (Baron, 2012).

In the past, a drug bound to several targets was unfavourable to the pharmaceutical industry due to side effects. Nonetheless, not many vastly selective sole-target drugs were discovered. The high rates of attrition in the late stages of the drug discovery process and the capacity to bind to different targets of drugs suggested an opportunity to reveal new indications for known drugs and unsuccessful candidates. Therefore drug repurposing based on the principle of polypharmacology can potentially provide opportunities for new indications from existing drugs.

4.4 Advantages of Drug Repurposing

The advantages of drug repurposing are based on the availability of existing drug or chemical safety and pharmacokinetic profiles. Most repurposed drugs, for example, have undergone *in vitro* and *in vivo* testing, lead optimisation, toxicological profiling, large-scale manufacturing, formulation development and even early clinical trials. Even withdrawn drugs still can be re-used if adverse effects can be avoided. Therefore, drug repurposing provides the ability to lower the cost, time and raise the success rate of developing a new medicine compared to conventional drug development strategies. On average, it takes 1-2 years for investigators to uncover new therapeutic targets and 8 years to develop a repurposed drug (Sertkaya et al., 2014), compared to 13.5 years in the traditional drug discovery process (Paul et al., 2010). With reduced time, drug repurposing can be appropriate in case of an epidemic such as Ebola (Kouznetsova et al., 2014; Veljkovic et al., 2015) or Zika (Xu et al., 2016; Shiryaev et al., 2017) Moreover, when compared to conventional

approaches, the cost of research and development for drug repurposing is reduced. For many nations, drug repurposing alleviates the cost barriers. A drug repurposing approach costs more than $1.6 billion to produce a new drug, compared to $12 billion for a standard approach (Deotarse et al., 2015). As a result, drug repurposing allows many companies to develop drugs with lower expenditures. Furthermore, drug repurposing is a low-risk method. Ashburn described drug repurposing offers a better risk-to-reward bargain while the traditional approach gains approximate 10% of success rate (Ashburn and Thor, 2004; Hay et al., 2014).

Another significant benefit of drug repurposing is the potential to find cures for patients suffering from rare diseases (or orphan diseases). A rare (or orphan) disease, which is defined by the US Orphan Drug Act of 1983 (US Food and Drug Administration, 1983), is a disease that affects only a small proportion of a population. Under the Orphan Drug Act, a disease is qualified as an orphan drug (or rare disease) if its prevalence is less than 200,000 in the US. In Europe, a disease is considered rare when 5 people in 10,000 are affected. Because this percentage changes by year and area, it is difficult to give a global figure (Sardana et al., 2011). Although the number of persons with each specific rare disease is limited, the number of people suffering from at least one rare disease is very considerable due to the large variety of these diseases. More than 7,000 rare diseases have been discovered, affecting more than 300 million individuals around the world (Bloom, 2016). For instance, Hutchinson-Gilford progeria syndrome (HGPS), a genetic abnormality that causes premature ageing, affects 1 in every 8 million babies born (Prakash et al., 2018), whereas Huntington's disease, a well-known rare ailment, impacts 3 to 7 per 100000 people in Europe (Rawlins et al., 2016). In another publication, 25-30 million patients were reported suffering these conditions due to the unavailability of treatments (Chakraborti et al., 2019). Due to their relatively "small" prevalence, rare diseases have drawn little attraction from the pharmaceutical companies due to the lack of economic motivation. As a consequence, drug discovery for rare diseases remains a challenge for academia. Drug repurposing will be a promising way of compensating for the lack of cures for these diseases.

Thanks to the availability of data regarding the safety and pharmacological profile, drug repurposing can help to find new drugs involving far fewer resources than traditional drug discovery. It is hence not a surprise that repurposed drugs accounted for 30% of the 51 approved drugs in 2009 (Graul et al., Feb). Drug repurposing has gained more and more attraction from the scientific community and industrial sector to find more indications.

4.5 Examples of Successful Drug Repurposing

Drug repurposing has proved its effectiveness via typical examples. Although drug repurposing gradually received increasing attention with advances in computational methods, in the beginning, drug repurposing was treated as an accidental coincidence in the clinical trial stages. Researchers discovered the additional biological effects, which were attributed to side-effects or otherwise ready for pharmacological use.

A well-known example of drug repurposing is the marketing of sildenafil by its treatment for erectile dysfunctions (trade-name as Viagra). Sildenafil was primarily discovered by Pfizer with the initial aim of was the treatment of hypertension. Sildenafil inhibits phosphodiesterase-5, an enzyme to degrade the level of cGMP inside the cells, leading to the dilation of the vessel. Thus, sildenafil was primarily investigated for the aim of hypertension and angina treatment. However, phosphodiesterase-5 also plays a major role in hydrolysing cGMP in corpus cavernosum and inhibition of cGMP results in penile erection enhancement (Terrett et al., 1996; Park et al., 1998). This so-called side-effect has been reported by a number of subjects and a study on this new exploration was carried out. The company set up a large-scale trial with more than 3,700 participants who are men with erectile dysfunction and with a promising outcome, sildenafil became a blockbuster (Morales et al., 1998; Montorsi, 1998; Wagner, 1998).

Another example is the case of thalidomide which is revived for the treatment of cancer. Thalidomide was firstly developed in the 1950s as an anticonvulsant drug. Early trials showed inefficiency but sedative effect. It was launched in Germany in 1957 under

the brand name of Contergan and in the United Kingdom in the next year as Distaval (Smithells and Newman, 1992) with the initial indication for morning sickness in pregnant women (Ashburn and Thor, 2004). It was sold over-the-counter in Germany and on prescription in the UK but was broadly used. Later, combination drugs that combine thalidomide and others were advertised for a wide range of indications: Asmaval to treat asthma, Tensival to treat hypertension, Valgraine to treat migraine and so forth. The introduction of these drugs put great dependence on the safety of thalidomide. Unusual cases of babies born with limb defects were first observed in Germany in 1960. In 1961 the number of reported cases increased and were linked to thalidomide administration of pregnant mothers. At the same time, the same inference was made in the UK, Sweden, Switzerland, Belgium, the Netherlands, Canada and Brazil. Approximate 10,000 babies around the world were born with limb malformations associated with the administration of thalidomide (Lenz, 1988). The mortality rate among 'thalidomide babies' was almost 40%, mostly due to major internal abnormalities (Lenz, 1988). As a result, the ratio of internal abnormalities is substantially less among survivors than it was in the general population at birth. It was withdrawn from the European market in 1961. It was revisited in 1965 since it posed potential to treat erythema nodosum leprosum lesions and, in 1998, the FDA granted approval for this indication (Teo et al., 2002). Recently, when people understood its mechanism to cause deformity in babies, it was repurposed for the treatment of cancer, particularly multiple myeloma (D'Amato et al., 1994). Recently, thalidomide shows potential to be repurposed for treatment of SARS-CoV-2 (Sundaresan et al., 2021).

These examples recommend that repurposing is possible for every drug, even when they are withdrawn.

4.6 Drug Repurposing for MRSA

Given the phenomenal development of antimicrobial resistance coupled with lack of funding and timing for new laboratory-based MRSA treating drugs, numerous attempts have been made to find new compounds with anti-MRSA competence. Lau et al. screened 1162 approved drugs by the Food and Drug Administration for activities toward MRSA to treat MRSA skin infections (Lau et al., 2015b). They have used an experiment-based method called 10 μM single-point assay on MRSA strain USA100 and found 6 candidates with Floxuridine as the most promising compound for repurposing for the treatment of MRSA skin infections. Niu et al. screened a compound collection comprising 1524 compounds from the Johns Hopkins Clinical Compound Library against MRSA strain USA300 using a growth inhibition assay. They pinpointed 9 candidates, 5 of which (chloroxine, thiostrepton, clofazimine, spiramycin and carbomycin) are antibiotics used to treat other infections and 4 of which (pyrvinium pamoate, quinaldine blue, dithiazanine iodide and closantel) are drugs used for other indications (Niu et al., 2017), which also need further study. Recently, Gilbert-Girard et al. screened 774 compounds from Screen-Well FDA Approved Drug Library Version 2 against *Staphylococcus aureus* strain ATCC 25923 and found 45 candidates for further *in vitro* tests (Gilbert-Girard et al., 2020). Once again, such activities again MRSA need further investigation. Prieto et al. screened 1.280 off-patent FDA-approved drugs in the Prestwick Chemical library using high throughput β-galactosidase-based screening for inhibition of GraXRS, a two-component system that determines bacterial resilience against host innate immune barriers and found VER as a promising candidate for sensitizing *S. aureus* (Prieto et al., 2020). Sedlmayer et al. used high throughput screening to screen 5,283 compounds against methicillin-resistant *S. aureus* (MRSA) ATCC 43300 for the *in vitro* inhibition of AI-2, which triggers biofilm formation. At a dose much lower than that for anti-cancer treatment, they found 5-FU effectively reduced MRSA adhesion and inhibited biofilm formation *in vitro*. In a mouse model, 5-FU was found to restore antibiotic susceptibility of MRSA infection (Sedlmayer et al., 2021).

Numerous endeavours have been made in the search for a new drug to treat MRSA infections. Either the experiment-based or *in silico* methods have been used in those studies. These studies are limited to specific target proteins (Prieto et al., 2020) or using

virtual screening with a specific class of the drug such as ethnomedical drugs (Dou et al., 2016), salicylanilide anthelmintic drugs (Rajamuthiah et al., 2015), ebselen (Thangamani et al., 2015), antihistamines (Perlmutter et al., 2014). Some researches utilized a larger library but aimed for the ordinary *S. aureus* strains (Lau et al., 2015b; Yeo et al., 2018).

Given the increasing alert of MRSA strain during the past decade and the lack of newly developed drugs to treat MRSA, it is essential to look for a new approach with an effective outcome. Drug repurposing with the aim for drugs with MRSA treatment indication is a promising strategy with a higher prospective success rate while lowering the time and cost.

No study utilising all possible repurposable compounds has been carried out. Furthermore, there has been no research to exploit the entire range of MRSA targets. The present thesis addresses these two open questions by combining multiple input sources (e.g. Phase I/II/III MRSA candidates) in a systematic manner and processing them for best protein:ligand overlap (matching).

An excess of 5900 compounds as possible ligand matches, all from the Repurposing Hub, were scrutinised and then virtually screened against more than 140 MRSA target proteins to find the best MRSA drug candidates with anti-MRSA activity. In addition, multiple docking programmes have been used for screening of drugs that have been approved or dropped out in late-stage clinical trials against the whole set of receptors from MRSA strain. Furthermore, to overcome the limitation of the current docking methods, a novel mechanism of "consensus scoring" (CS), that is essentially a statistical combination of all available docking outputs from multiple docking programmes to arrive at a holistic 3-dimensional representation of the molecules, that were established as being superior to any individual docking architecture, faster in processing and more accurate in prediction. This CS model boosts up the enrichment ability for potential candidates with antimicrobial activities, followed by an enumeration of the top ligands.

Chapter 3

Literature Review

1 Drug Repurposing Approaches and Methodologies

The aim of Drug Repurposing (DR) is to find a new indication for an existing drugs, a late-stage trial substance or a withdrawn drug. Drug repurposing can use the same methods as traditional drug research and development, but on a shorter timeline. In general, there are two basic drug repurposing principles. First, because of the interdependence of various ailments, medications developed for one disease may also be effective for other diseases. Second, because medications are inherently confusing, they can be linked to a variety of pathways and targets (Ashburn and Thor, 2004). As a result, DR research can be divided into two categories based on where the discoveries come from: drug-based strategies, in which the discovery is based on drug knowledge and disease-based strategies, in which the discovery is based on disease information. For forecasting treatment potentials and innovative indications for current treatments, drug-based techniques rely on data connected to drugs, such as chemical, molecular, biological, pharmacological and genetic information.

Drug-based strategies are employed when there is a large amount of medication-related information or a strong need to learn about the contribution of pharmacological properties to drug repurposing. A greater amount of research in this category shares the proposition that if two medications have comparable profiles and modes of action and one drug is utilised to cure a certain disease, the other is a strong candidate for treating the same disease. The genome strategy and the chemical structure and molecular information strategy are the two sub-strategies that fall within this category. Data on diseases, such as phenotypic features, side effects and indications, is used as the foundation for disease-based techniques to forecast therapeutic potentials and novel indications for current treatments. When there is a lack of drug-related data or a motivation to learn how pharmacological properties can help with drug repurposing aimed for a specific illness, disease-based strategies are applied (Ashburn and Thor, 2004). If two diseases have comparable profiles and indications and one drug is utilised to cure one disease, then that drug can be considered a strong contender for curing the other disease. The phenome strategy is the primary approach that speaks for this category.

Drug Repurposing is carried based on two main approaches: experiment-based and computer-based. These approaches also resemble approaches in drug discovery and development, with the growth of computer-based approach. The experiment-based strategy (or activity-based drug repurposing) refers to the use of actual experiments to screen chemicals or drugs for additional therapeutic usage. The experiments can be carried out *in vitro* and/or *in vivo*, using cell assays, animal models or clinical trials (Oprea and Overington, Aug; Lionta et al., 2014), without the need for structural data of the target proteins. Recently, there is a new application called "airway-on-a-chip" that can be used for drug repurposing (Si et al., 2021). The advantages of experiment-based approaches include a diverse range of screening assays for targets and cells, simple to evaluate screening results and the reduced proportion of false positive hits (Shim and Liu, 2014). Nonetheless, the disadvantages are similar to traditional drug discovery: cost and time consuming; the prerequisite of physicochemical properties and structural information. On the other

hand, *in silico* repurposing carries out computational screening of library of chemicals or drugs to identify possible active compounds. Similar to computer-aided drug design, this latter approach has become fruitful since an enormous volume of structural information on macromolecules has been compiled in recent years in conjunction with the growth of bioinformatics and computational science. In this approach, the intermolecular interaction between the compounds and the target of interest is exploited (Talevi, 2018). This approach can reduce time and expense, but it requires knowledge of the structural data of the target proteins. It also necessitates disease/phenotype data or gene expression profiles of medicines. Since the growth of bioinformatics and computer power, a large volume of data about the structure of proteins has been acquired over recent years, this latter approach has proven fruitful.

The drug repurposing techniques can be classified into three groups: drug-based, (ii) target-based and (iii) disease-based, based on the quantity and quality of data provided. In the drug-based methodology, the structural features of drug molecules, bioactivities, adverse reactions and toxicity are investigated to find new biological effects. This approach is based on traditional pharmacology and drug discovery concepts, in which experiments are typically undertaken to establish the biological response of drugs without information of the physiological targets. One well-known successful example in this method is sildenafil (Koch et al., 2014). Target-based methodology consists of *in silico* or virtual screening a library of drugs or compounds against a target protein to identify potential ligands with possible interactions. Since most pharmacological targets directly reflect disease pathways/mechanisms, this methodology has a higher success rate than the drug-based methodology (Napolitano et al., 2013). When additional information about the disease model is available, the use of disease-based technique in drug repurposing becomes relevant. In this case, drug repurposing can be directed by the illness and/or treatment information that is provided by disease-related target proteins, genetic data, metabolic pathways/profile as well toxicity, therapeutical targets, disease pathways, pathological conditions, adverse and side effects regarding the disease and treatment. As a result, it necessitates the creation of specialised disease networks, including knowledge of genetic expression, main targets and disease-induced receptor related to the diseases (Chong et al., 2006).

In comparison to standard techniques, incorporating target information into the drug repurposing process increases the likelihood of discovering effective medications. Target-based approaches, such as docking, allow researchers to screen practically all drugs or chemicals with defined structure information in a matter of days. This is the reason why many pharmaceutical companies rely on these techniques to discover new indications (Jin and Wong, 2014). In this work, thanks to the availability of MRSA protein structural information, it is possible to use molecular docking for virtual screening of drugs and compounds against these target proteins to identify the potential candidates with anti-MRSA activity.

Databases for Drug Repurposing

Apart from the available information, drug repurposing mirrors the advantages and disadvantages of drug discovery and development. Similar to drug discovery and development, drug repurposing integrating computational methods is less resource-consuming than experiment-based methods. Therefore, with the availability of structural information, virtual screening is favoured over other methods. Thanks to the advances in computer science and chemoinformatics and the observation in the correlation between structure and bioactivity, high speed and cloud computing allow intensive computer calculations. Along with the enhancement of computational power, the augmentation in chemoinformatics also allows expanding the chemical space. Millions of compounds yet to be synthesised can be virtually explored. A number of multi-purpose chemical libraries are freely available: ZINC (Sterling and Irwin, 2015), Pubchem (Kim et al., 2020), ChEMBL (Mendez et al., 2019) and commercially supplied: Boehringer Ingelheim's BI-Claim (Lessel et al., 2009), Eli Lilly's Proximal Collection (Nicolaou et al., 2016), Pfizer global virtual library (Hu et al., 2012), and Merck's Accessible inventory (Lyu et al., 2019).

For repurposing objective, the main focus is approved drugs, usually by the United States Food and Drug Association. However, substances at clinical trials also make good

candidates, as they have passed the tests for safety and proved to exert certain biological reactions. Withdrawn drugs is also a good source for repurposing. Although they have been withdrawn due to severe adverse reactions, a new indication is still possible with appropriate administration. There are many libraries that contain the drugs for repurposing. One main source containing the FDA-approved drugs is the DrugBank (Knox et al., 2011). Investigational and withdrawn drugs are also available but limited. Compounds at clinical trial stages are also available at www.clinicaltrials.gov with untreated information (Tse et al., 2009; Zarin et al., 2011). A number of libraries have been created with specific disease-orientation: GCPR-targeted (Sriram and Insel, 2018). Recently, Corsello et al. have constructed a repository with more than 10,000 compounds, including approved drugs and substances that have reached the clinical stages, namely Repurposing Hub (Corsello et al., 2017). Repurposing Hub has been exploited and resulted in a promising candidate, halicin, with potential broad-spectrum antibacterial activities (Stokes et al., 2020).

Protein structure databases are also expanded, providing necessary information for *in silico* experiments. Some repositories are open available: Protein Data Bank (Berman et al., 2000), ModBase (Pieper et al., 2011), UniProt (The UniProt Consortium, 2019). PDB is a leading protein and macromolecule database with more than 170,000 experimental structures contributed by more than 40,000 data depositors around the world (Burley et al., 2021).

With such an abundance of structural information from databases, drug repurposing using *in silico* methods become much easier than before. Thus, drug repurposing using structure-based virtual screening for MRSA treatments is possible providing the availability of large libraries containing approved and trial drugs, as well as structural information of MRSA targets.

2 Virtual Screening

The term "virtual screening" (VS) was first mentioned in a publication in 1997 (Horvath, 1997). This breakthrough precipitated a multitude of publications in the subject. A simple search for "virtual screening" in Google Scholar produces 1.97 million hits which obviously indicates to the importance and applicability of the technique in factual terms. The purpose of virtual screening is to narrow down the portion of lead-like hits against the chosen target. Virtual screening is a more straightforward and rational drug discovery strategy than conventional experimental high-throughput screening and it has the benefits such as low expense and efficient screening (Moitessier et al., 2008; Bailey and Brown, 2001). As no experiments are carried out, neither chemicals are needed nor equipment to be operated (such as in high-throughput screening). Consequently, virtual screening is cost-effective. Meanwhile, virtual screening can save time due to the cut of experiments or compound synthesis. Furthermore, virtual screening is labour efficient owing to the exclusion of laborious work.

Virtual screening applications works based on the information about the target and ligands. Depending on the extent of available information, virtual screening can be divided into two main approaches: ligand-based virtual screening and structure-based virtual screening. In the first approach, the information of the target is not fully provided and that of ligands is available, usually known as active ligands. Ligand-based virtual screening is based on the similarity concept that is structurally similar molecules tend to have similar chemical and biological properties. The chemical library can be scanned for compounds with similar properties. Therefore, the heart of ligand-based virtual screening methods is the measurement of similarity, which ranges from two-dimensional descriptors, such as fingerprints, to three-dimensional descriptors like pharmacophores.

As for structure-based virtual screening, information about the structure of the targets and compounds within the library that is available are scanned to estimate the probability of binding with high affinity. The core is commonly used in this approach but is not limited to molecular docking. Molecular docking is a mathematical technique that relies upon two components: searching algorithms to search for possible conformations of the compounds

and scoring functions to evaluate the binding interaction of each conformation and rank the top scores among them. Molecular docking has been the most used strategy since the early 1980s to present. Structure-based drug design is an essential part of drug discovery and development (Warren et al., 2012; Merz Jr et al., 2010; Pei et al., 2014). Many commercial medicines are products from structure-based drug design method (Sliwoski et al., 2014).

In silico investigations, molecular docking is an essential element that is extensively exploited in current drug design and discovery. Contemporary molecular docking techniques have developed to the point where they are regarded as a valuable tool in rational drug design, thanks to substantial developments in terms of innovative computational algorithms and powerful computational resources. In this method, the magnitude of intermolecular interaction is attempted to be correctly estimated based on the structural and experimental data available. It is worth noting that protein docking (or protein-protein docking), which is the docking of two macro-molecules, is not discussed in this work.

3 Molecular Docking

3.1 Overview of Molecular Docking

The therapeutic effect of the drug molecules in particular as well as the biological effect of the small molecules in general is due to a mechanism known as molecular recognition, which is a very basic occurrence. Modulation of biological signals and cellular reactions are regulated by a range of such recognition events. These processes occur at the molecular level and shape the principle of ligand-receptor interactions. Non-covalent interactions such as intermolecular van der Waals, hydrogen bonds and electrostatic interactions are used to establish connections in physiology and pharmacology (Brooijmans and Kuntz, 2003). The study of chemical characteristics that are accountable for specific biological responses, as well as the anticipation of molecule alterations that boost potency, are not trivial tasks. A better understanding of ligand-receptor recognition can lead to a breakthrough in structure-based drug design. Molecular docking is one of the useful methods that can give an insight into such ligand-receptor interactions.

Molecular docking is terminology for the prediction of position, orientation and conformation (usually termed as docking pose or pose) of a small molecule in reference to a biomacromolecule. Molecular docking is used in opposition to protein docking (or protein-protein docking), docking of two biomacromolecules, which is not discussed in this study. Docking studies were pioneered during the 1970s and 1980s (Levinthal et al., 1975; Salemme, 1976; Kuntz et al., 1982). Since the beginning, binding between ligand and protein was supposed to be between two rigid molecules and many studies followed the theory “lock-and-key” of Koshland in 1958 (Koshland, 1958). This theory is based on the model of “lock” and “key”, whereas protein plays a role as lock and ligand as key. A right “key” to a “lock” will exert the appropriate biological reactions. Nonetheless, the “lock-and-key” model was not sufficient to explain all the experimental results.

The next theory that was more advanced than the “lock-and-key” was the “induced-fit” model (Wei et al., 2004). This theory is based on an observation: the protein changed its conformation to adopt the complex with ligand with the lowest energy. During the course of the docking process, the ligand and the protein change their shape to obtain a general "best-fit" state, just like “hand-and-glove”. Another theory is confirmation selection and population shift which is based on the energy landscape (Knegtel et al., 1997). The protein structure adopts various conformations from which corresponding energies can be demonstrated as a map of energies. These conformations are interchangeable and some adopt local lowest energies and one lowest of all. A ligand can bind to one of the energy canyons and other conformations will shift to this state, which leads to the lowest energy complex and with the highest frequency.

At present, docking remains a burgeoning field of study (Kitchen et al., 2004). A quick PubMed search for articles including the keywords "docking" and "ligand" was undertaken to gain a better understanding of the extent to which docking investigations have penetrated the research community. The number of articles regarding docking has steadily

increased from 1986 to present, with an approximate 2,141 in 2020 and 2,239 by the end of 2021. This is a rough estimate and the actual figures require further investigation, but it can bring a general view of how common docking studies are.

3.2 Molecular Components of Molecular Docking

The molecules of receptor and ligand are the two key interacting components in any molecular docking. The three-dimensional structure of the receptor that is used in the calculation for a molecular docking simulation is either an experimental (solved three-dimensional structures of proteins stored in the database of proteins) or an anticipated structure using different prediction techniques such as "comparative modelling" (homology and threading). The binding pocket can then be determined using the structural information of the protein. The binding pocket can be defined with certainty if the target is co-crystallised with the ligand; or else, the binding pocket must be deduced in one of the other ways, such as using methods to predict protein function from structural analogues or from the investigation of physicochemical and geometrical features of the protein geometry. The binding pocket is not necessarily the biggest, but usually the cavity with the highest physiochemical criteria. The ligand is the other essential element of molecular docking. For ligand processing, two approaches have been used: whole-molecule approaches and fragment-based methods. In whole-molecule approaches, the pool of conformations of the ligand is explored and each conformation is docked into the binding pocket of the protein. In fragment-based approaches, a library of fragments is prepared from the ligand structure and the docking is completed using either fragments docked separately before being reconnected or using fragments as "anchor" gradually enlarge the ligand during the docking phase (Rarey et al., 1996). Depending on the extent of ligand and protein flexibility, different levels of approximation will be applied for scoring functions.

3.3 Docking Methodology

In a molecular docking simulation, the flexibility of ligand and protein is determined as six degrees of translational (along x-, y- and z-axes) and rotational freedom. Furthermore, depending on the torsion angles of each rotatable bond, conformational degrees of freedom is also accountable for each ligand and receptor (Gani, 2007; Leach et al., 2006). Because finding the entire conformational space is a challenging task that takes a lot of time and resources, an approximation for lowering the dimension of the search space is necessary. The degree of molecular flexibility is usually used to categorise molecular docking techniques. There are three extents of approximation in docking approaches from the perspective of flexibility: rigid docking, semi-flexible docking and flexible docking.

The key parameters in molecular docking are accuracy and speed. The oldest and most basic molecular docking methods had the basic form rigid-body approximation. The structure of the ligand and protein is not changed in the docking process in rigid docking and both are treated as rigid objects. The protein is considered to be solid in the second strategy, semi-flexible docking. In molecular docking, rigid-body assignment of two components (i.e. ligand and protein) is speedier than when flexibility is added in the docking process (Morris and Lim-Wilby, 2008). This is because the search space is relatively small. Although the speed of a docking simulation is advantageous, the impact of various ligand and protein conformations in bound and unbound states cannot be overlooked. Major conformational changes in the protein structure, such as reorganisation of side chains and movement of loops and domains, can happen when the ligand binds to the target ("induced-fit" theory) (Wei et al., 2004). As a result, because a real biological structure has multiple degrees of freedom, the flexibility of both ligand and receptor should be considered for an ideal scenario. It is critical to developing programmes that are able of addressing these concerns for this goal. Nevertheless, when the protein is also considered flexible, docking can take extremely long time, even weeks. As a result, the most typical technique is keeping the protein stiff whereas the ligand undergoes a conformational change during docking, which is likewise a trade-off between accuracy and computational speed. This methodology

has been used by almost all docking programmemes and is broadly utilised in numerous studies (Rarey et al., 1996; Morris et al., 1998).

Nonetheless, flexible ligand and rigid protein have not fully answered the question in the task of docking. Protein motility is correlated with ligand binding activity (Teague, 2003). Therefore, in flexible docking, both the ligand and the protein conformational degrees of freedom, as well as translational and rotational degrees of freedom, are taken into account (Hung and Chen, 2014; Leach et al., 2006; Sousa et al., 2013). However, introducing protein flexibility into docking poses a huge challenge in computational cost. To overcome this problem, the protein is not treated fully flexible, but other approximation methods are applied, such as flexibility of side chains (Morris et al., 2009) or ensemble of rigid conformations (Knegtel et al., 1997).

3.4 Binding Site Prediction

In addition to three-dimensional protein data, identifying overlapping regions between proteins and ligands is essential for virtual screening. A binding site is a region within proteins where ligands can form intermolecular iterations. The interaction of small molecules to a protein at various binding sites may trigger a biological reaction. If this biological reaction is relevant to a particular disease or disorder, this interaction can be considered as a target for the treatment of the disease or disorder. If the molecule binds to the protein at the binding site and exerts a more powerful biological reaction, it can act as an agonist. On the contrary, if the interaction ceases or weakens an activity, it can be developed as an inhibitor. Traditionally, binding sites can be determined by co-crystallisation of a complex of protein and ligand. However, such co-bound ligands are not always available. In case only the protein structure is present, the binding site can be detected by comparison to other structures or sequences with an existing binding site or by prediction using computational tools.

A wide range of algorithms have been developed to predict binding pockets of ligands to protein. These algorithms are classified into two categories: residue-based, surface-based and interaction-based (Ehrt et al., 2018). Some binding site prediction tools available: IsoCleft Finder (Kurbatova et al., 2013), IsoMIF (Chartier and Najmanovich, 2015), SiteHopper (Batista et al., 2014), SMAP (Xie et al., 2009). These prediction tools usually return a list of putative sockets, ranked in terms of physicochemical or geometrical properties. Nonetheless, confirmation in conjunction with visualisation is recommended.

3.5 Mechanics of Docking

The objective of molecular docking is to use computer aided techniques to anticipate the three-dimensional structure of the ligand-receptor complex. Docking is accomplished in two steps: firstly, sampling ligand poses in the binding pocket of the protein receptor and then ranking these poses using a scoring function. When addressing various degrees of freedom, such as translational and rotational freedom, the computational cost of docking for arranging ligand and receptor near each other is enormously large. As a systematic method, one could construct and investigate all conceivable binding modes for the receptors of interest using their three translational and three rotational freedoms, although this approach would be impractical given the computing capability of recently developed computer resources. For example, assuming increments of 10 degrees in angle and 6^{14} conformations, it will take a processor capable of processing 10000 confirmations per second around 2×10^3 years to finish the simulations (Taylor et al., 2002). As a result, it is essential to create efficient search algorithms and scoring functions to set a balance between computing costs and the ligand conformation space.

Search Algorithms

Search algorithms are also preferred to as "matching algorithms" or "sampling algorithms". Prior to the prediction of the binding state between a ligand and a protein, docking programmes have to obtain an ensemble of ligand conformations (and protein conformations

in case of ensemble docking). These conformations can be generated by translational, rotational and vibrational changes in the ligand structure. In theory, the search space is made up of all potential protein and ligand positions, orientations and conformations. However, with existing computer resources, exhaustively investigating the search space is inconceivable. Several algorithms have attempted to reproduce the binding mode obtained via experiments (can be referred to as search algorithm, sampling algorithm or placement algorithm). Shape-complementary methods use geometric descriptions to map a ligand into an active location of a protein. Another approach is molecular dynamics, which simulates the molecular level interaction between all molecules to explore the entire conformational space, essentially using Newtonian kinetics to quantify collision between any two molecules. Other widely used genetic algorithms conceptualises the idea of the GA stems from Darwin's theory of evolution. The whole ligand pose is considered as "chromosome" and the ligand fragments as "gene". Mutation makes random changes to the genes, resulting in a new ligand conformation. Monte Carlo algorithms generating random conformations by rotation or translation of the bonds is another option. These are largely restricted to static configurations by identifying the lowest energy manifolds (Meng et al., 2011). Recent advances include generalisation of the original Monte Carlo algorithm to incorporate time evolution of interacting variables, a method popularly referred to as *kinetic Monte Carlo* (Chattopadhyay and Marenduzzo, 2007).

Scoring Functions

In the docking process, a search algorithm may generate a large number of conformations to be docked into predicted binding sites of protein, depending on the number of rotatable bonds the ligand possesses. The follow-up target is to identify the correct binding pose within that ensemble at an acceptable accuracy within reasonable time. Scoring functions are used in computer-aided drug discovery such as virtual screening, lead optimisation and structure-based drug design. In molecular docking, scoring functions are methods to predict the affinity between two molecules, in particular, protein and ligand. Scoring functions aim to discriminate the true binding poses from false-positive prediction and to rank the binding strength of non-covalent interactions among the set of results using approximate algorithms.

Force field is a function to calculate the potential energy of an ensemble of atoms in molecular mechanics and molecular dynamics simulations. Force field functions consist of terms to calculate various types of interaction within the system. Force-field-based scoring functions predict the binding affinity by calculating all the intermolecular interactions (van der Waals and electrostatic) between ligands and proteins (due to the nature of interactions between protein and ligand). A force-field-based scoring function generally takes the following form:

$$E = \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{torsions} K_\varphi[1+cos(n\varphi-\delta)] + \sum_{vdW} \varepsilon_{ij}\left[\left(\frac{R_{min,_{ij}}}{R_{ij}}\right)^{12} - 2\left(\frac{R_{min,_{ij}}}{R_{ij}}\right)^{6}\right] + \sum_{electrostatics} \frac{q_i q_j}{4\pi D r_{ij}} \tag{3.1}$$

where b is the interatomic distance, k_b is the stiffness of the bond, b_0 is the equilibrium length of the bond, θ is the angle formed by the two bond vectors, the values of θ_0 and k_0 are the stiffness and equilibrium geometry of the angle, respectively. The torsional potential in the equilibrium is characterized by cosine function, where ϕ is the torsional angle, δ is the phase, and n is the dihedral potential. The last term $\varepsilon_i j$ is a parameter based on the two interacting atoms i and j, q_i and q_j are the effective charge on i and j atoms, and $R_{min,_{ij}}$ is the distance at which the energy of Lennard–Jones equation is at minimum (Huang et al., 2006a). Since there are versions of scores using a component such as Poisson–Boltzmann or Generalized Born solvation model not based on the force field.

Some docking softwares use force-field-based scoring functions such as DOCK (Meng et al., 1992; Allen et al., 2015), GOLD (Jones et al., 1997; Verdonk et al., 2003), AutoDock

(Morris et al., 2009) and other methods for refinement such as linear interaction energy and free-energy perturbation. Due to the great number of the pairwise atom interactions that could be generated within the complex, the force-field-based scoring function is computationally expensive and time-consuming. Therefore, there is a cut-off in distance to increase the speed but also lead to a decrease in accuracy. Changes in conformational entropy and changes in solvational entropy are the two main factors that affect binding entropy. Since the binding process results in the loss of conformational degrees of freedom for both the ligand and the protein, the conformational entropy change is often negative. On the other hand, because of the partial or complete desolvation of the binding cavity during binding, the solvation entropy is usually always in favor (Huang et al., 2006a; Singh and Warshel, 2010).

The second type of scoring function is empirical scoring functions, which approximate the binding interactions of a complex based on a set of weighted energy terms. Empirical scoring functions use the value from a training set of ligand-protein complexes. The empirical scoring functions take the following form:

$$\Delta G = \sum_{i} W_i \Delta G_i \tag{3.2}$$

In Equation 2.2, ΔG_i may stand for H-bond, vdW energy, electrostatics, desolvation, entropy and hydrophobicity. Fitting the known binding affinity values of a collection of protein-ligand complexes with accessible three-dimensional structural data yields the relevant coefficients W_i. To fit the experimentally measured affinity values with the calculated binding score, the coefficients are provided using linear regression, such as multiple linear regression, or non-linear regression, such as support vector machine methods. The empirical scoring functions are quicker in generating the binding score than the FF scoring functions due to the simplicity of energy terms, but they may have a restricted applicability area due to the amount and variety of protein-ligand complexes in the training set. Docking programmes using empirical scoring functions include: FlexX (Rarey et al., 1996), Glide (Friesner et al., 2004), ICM (Abagyan et al., 1994), Surflex (Jain, 2003).

Another type of scoring function is knowledge-based scoring functions, which take into account the frequencies of occurrence and distance of various types of atom pair interactions. They are based on the observation that the favourability of interaction is directly proportional to the occurrence frequency of that interaction and therefore favourably contributes to the binding affinity. The distributions of frequency of interactions from a training set are used to compute the potential of mean force. Potential of mean force is the potential which provides the average force across all coordinates. It is also the free energy profile along a preferred coordinate. Potential of mean force can be used to reflect the energetics of a range of biological systems, such as interactions between molecules, conformational changes within a molecule and protein folding and unfolding (Mitchell et al., 1999b). The knowledge-based scores sum up pair-wise statistical potentials between protein and ligand and are generally expressed as:

$$Score = \sum_{i}^{ligand} \sum_{protein}^{j} \omega_{ij}\left(r\right) \tag{3.3}$$

The inverse Boltzmann formula is used to derive pair-wise potentials (X') straight from the occurrence frequency of atom pairs in a dataset.

$$\omega_{ij}(r) = -k_B T \ln[g_{ij}(r)] = -k_B T \ln\left[\frac{\rho_{ij}(r)}{\rho^*_{ij}}\right] \tag{3.4}$$

where k_B is the Boltzmann constant, T is the absolute temperature of the system, $\rho(r)$ is the number density of the protein-ligand atom pair at radius r and ρ^* is the pair density in a reference state where the intermolecular interactions are either assumed zero or they consist of non-specific interactions that are common to all sorts of the atom. Free energies are estimated using radial distribution function, g(r), pertaining to the fraction

$\frac{\rho(r)}{\rho^*}$. The most common method for obtaining the requisite pair-wise potentials is to use a big collection of protein/ligand complex structures as the training set, often known as the "knowledge base". Atoms of protein and ligand atoms are degenerated and divided into several categories. The above formula is then used to generate distance-dependent potentials for each probable atom pair from the occurrence frequency of this atom pair found in the training set. Knowledge-based scoring functions are far more effective than FF or physics-based techniques due to their pair-wise features, which frequently require intensive solvent processing. Knowledge-based potentials, unlike empirical scoring functions, are produced through statistical analysis of structural data without the use of experimental binding affinity measures. Scoring functions belonging to this category include: PMF (Muegge, 1999), DrugScore (Gohlke et al., 2000), SMoG (DeWitte and Shakhnovich, 1996), BLEEP (Mitchell et al., 1999b,a), GOLD/ASP (Mooij and Verdonk, 2005).

Machine learning scoring functions, a new type scoring functions has been proposed recently. For modelling of diverse physicochemical and biological features, the quantitative structure-activity relationship technique has traditionally been widely used. The descriptors in quantitative structure-activity relationship are obtained from the representation of the compounds and contain chemical or topological information that can be used to investigate protein-ligand interaction mode. In the advancement of docking scoring functions, descriptors such as traditional ligand-based attributes (number of rotatable single bonds, number of H-bond donor/acceptor, molecular weight), structural connectivity fingerprints of protein-ligand complex, geometrical features (surface or shape or volume characteristics) and specific interactions attributes (hydrogen bonds, electrostatic interactions or aromatic stacking) can be used. Furthermore, statistical models for generating binding scores can be developed using linear regression methods or machine learning approaches such as random forest, Bayesian classifiers, neural networks and support vector machines. Machine learning scoring functions may resemble empirical scoring functions in appearance, but they normally have a much higher number of descriptors and are not always in a linear functional form but mostly rely on the machine-learning technique used. Some scoring functions in this category such as LigScore (Krammer et al., 2005), RF-Score (Ballester and Mitchell, 2010), NN-Score (Durrant and McCammon, 2010) SFCScore (Zilian and Sotriffer, 2013), ID-Score (Li et al., 2013).

The criteria for an ideal scoring function include accuracy and promptness. However, available scoring functions had to sacrifice either accuracy or promptness. This obstacle has not been overcome and an adequately efficient function is still not yet achieved. According to a review of Liu and Wang, scoring functions can be reconsidered into 4 concepts: physics-based methods, empirical scoring functions, knowledge-based potentials and descriptor-based scoring functions (Liu and Wang, 2015) since some versions can include components such as Poisson–Boltzmann or Generalized Born solvation model which are not covered in conventional terms. Besides these types of scoring functions, consensus scores that integrate more than one scoring function have been widely reported to improve the performance of virtual screening.

3.6 Docking Performance Validation

Given the immense types and options available as docking programmemes, it is imperative that consideration be given as to their quality and a general base level performance. In other words, given a select set of proteins and ligands, certain docking programmemes are expected to perform better (meaning more accurate structure classification and faster convergence) than others. The basic question is how accurate the search algorithms generate the native pose regarding the known complex and how accurate the scoring functions rank the correct pose among other poses. If a ligand predicted pose in the active site was closer than a certain threshold compared to the X-ray structure, docking is declared successful. Usually, the threshold is set to 1.5 to 2Å(Dixon, 1997; Bissantz et al., 2000). For instance, FlexX was tested on a sample of 19 protein-ligand complexes before being tested on a broader sample of 200 complexes (Rarey et al., 1996). Glide's accuracy was tested

by redocking ligands from 282 PDB complexes (Friesner et al., 2004), whereas GOLD's docking accuracy was tested on 100 and 305 PDB complexes (Jones et al., 1997). DOCK has been validated on various proteins over the years (Shoichet et al., 1993; Bodian et al., 1993; Debnath et al., 1999), while LigandFit has been evaluated with 19 protein-ligand complexes (Venkatachalam et al., 2003). The same test set of 30 protein-ligand complexes with experimental binding affinities was used to calibrate AutoDock (Osterberg et al., 2002) and AutoDock Vina (Trott and Olson, 2010). The following criteria were proposed to evaluate various aspects of docking programmes.

"Docking power" refers to the capability of a docking programme or a scoring algorithm to distinguish the true ligand binding pose among computationally putative decoys. In an ideal condition, the top-ranked binding pose should be designated as the native binding pose. For each protein-ligand complex, a decoy set of ligand binding poses was built. Then, for each set, each scoring function is used to rank all ligand binding poses. The distance between the best-scored binding pose and the true binding pose is calculated by computing the RMSD value using the Hungarian function (Allen and Rizzo, 2014). This complex was designated as a successful example for the provided scoring algorithm if the distance between the docked pose and the known native ligand is less than a predefined threshold (e.g., usually no more than 2.0Å(Plewczynski et al., 2011)). The success rate over the whole test set is estimated as the docking power of a given scoring algorithm or docking programme (Cheng et al., 2009; Li et al., 2014; Liu et al., 2017; Su et al., 2019)

"Ranking power" refers to the capability of a docking programme or a scoring algorithm to accurately rank the different know ligands bound to the same protein by their binding affinities given true binding poses of these ligands are provided. To measure ranking power, each cluster of complexes in the test set contains three complexes generated by the same protein with substantially varying binding affinities. The success rate of accurately ranking the three complexes in each cluster throughout the whole test set is used to determine a scoring function's ranking power. One point was assigned to a docking programme or scoring algorithm that correctly score the three members of a complex cluster as the best $>$ the medium $>$ the lease order of binding affinities. Once this analysis was finished for the whole test set, an overall success rate was calculated.

Ranking power does not require scrutiny in the correlation between docking scores and experimental binding affinities like a linear correlation in scoring power, as long as the rank orders of binding ligands are correctly retrieved. The essence of ranking power can recompense the shortcoming of scoring functions that is the accuracy of scoring functions are still far from perfect. However, in virtual screening using molecular docking, the priority is to enrich the possible actives, which is adequate with ranking power.

"Scoring power" refers to the capability of a docking programme or a scoring algorithm to compute the binding scores in a linear fashion in accordance with experimental binding affinities, given the known protein-ligand complexes. On the contrary to ranking power, scoring power highlights a scoring algorithm's ability to execute across a variety of protein-ligand complexes. It assesses a scoring function's overall capability to estimate binding affinity, which is perhaps the most challenging task in virtual screening works. The binding scores of a number of complexes in the test set were computed using each scoring algorithm. The Pearson correlation coefficient between each scoring algorithm's predicted binding values and the observed binding data was used to quantitatively assess its scoring power on this test set.

"Screening power" refers to the capability of a docking programme or a scoring algorithm to identify the true binding ligands among a cluster of arbitrary compounds to the same target/protein. In reported works, screening power was assessed in a cross-docking design. The test set consists of a number of clusters of complexes, whereas each cluster contains three or five ligands bound to the same protein. Therefore, ligands bound to other proteins are hypothesised as non-binders to that protein. The scoring algorithm's screening power was measured by how adequately it sorted the native ligands at the top.

3.7 Database for Benchmark/Computational Validation

Virtual screening hits are predictions that need to be certified both *in silico* and *in vitro* or *in vivo* to establish their accuracy because virtual screening workflows comprise a course of computational procedures. Computational validation is frequently carried out by running virtual screenings on a collection of active compounds and a set of inactive or decoy compounds altogether.

Actives (or active compounds) are substances that have been confirmed to exert a high extent of activity against a certain target. The specific threshold at which a compound is determined active is arbitrary, but compounds with IC50, K_i, or EC50 values in the range from nM to mM (or nano-molar to the micro-molar range) are frequently termed actives (Gimeno et al., 2019). The virtual screening methods are more rigorous when the threshold is set higher. It is acknowledged that a drawback of virtual screening is that the virtual screening methods may not represent the actual mechanism of the compound (for example, the ligand may attach to an allosteric location of the protein rather than the catalytic pocket) (Scior et al., 2012). Because their binding modes are different, the VS should not be able to recognise these chemicals as actives, as this would weaken the virtual screening outcome.

Inactives (or inactive compounds) are substances with low (or no) affinity toward a certain target. Similar to actives, an activity cutoff below which substances are considered inactive should be chosen (Gimeno et al., 2019). Usually, there should be an interval between actives and inactive threshold to avoid indetermination (Mysinger et al., 2012). Information about actives is usually obtained from a common compound database such as PubChem (Kim et al., 2016) and ChEMBL (Gaulton et al., 2017).

Decoys are substances that share similar properties with actives but haven't been confirmed for affinity toward the target of interest. Since they are likely to be inactive, they are putative decoys (Gimeno et al., 2019). Decoys are typically acquired by scanning for substances with physical characteristics that are comparable to active compounds but are chemically distinct. Due to the limited number of reports and the resulting paucity of information on inactive compounds, decoys are commonly utilised instead of inactives in benchmark processes (Kirchmair et al., 2008). Analogously to actives, the decoys are putative inactive substances despite their action profile with the target of interest have not been reported. Therefore, a modest percentage of them might be positively active, which leads to reduced performance. This is also known as another intrinsic drawback of virtual screening methods (Scior et al., 2012). Information about decoys can be accessed from databases like DUD (Huang et al., 2006b), DUD-E (Mysinger et al., 2012) or through generating tools like DecoyFinder (Cereto-Massagué et al., 2012), which allows users to find sets of decoys that match the physiochemical properties of given actives.

For self-evaluation, each developing team used a different training set of proteins and ligands for docking simulation. Although the selection was made with care, the composition of these training libraries is skewed towards some families of proteins. To provide a means for comparison and evaluation of the performance of the docking programmes, a number of decoy and active libraries have been developed for structure-based virtual screening. The composition of decoys and actives was made to cover most of the range of protein families. Some of these libraries are used in studies to benchmark docking programmes. By introducing a new method garnered from spatial statistics, Maximum Unbiased Validation (MUV) was developed to offer unbiased samples in respect to both false enrichment and similar bias, containing 18 targets with a number of 30 active ligands and 15,000 decoys for every target (Rohrer and Baumann, 2009). Demanding Evaluation Kits for Objective *in Silico* Screening (DEKOIS) introduced in 2011 (Vogel et al., 2011) and upgraded in 2013 to a newer version, DEKOIS 2.0 (Bauer et al., 2013), which contained 81 sets of actives and decoys for 11 target classes, was designed to avoid the biases into the decoy sets, i.e. analogue bias and artificial enrichment. The Directory of Useful Decoys (DUD), a database with 2,950 documented ligands for 40 various targets and a ratio of one active to 36 decoys, was created as a benchmarking collection for docking studies with the goal of reducing bias (Huang et al., 2006b). Nonetheless, several investigations indicated that particular

structures were over-represented in the active portion, that charge was not taken into account when ligand sampling was done and that actual ligands might be detected in decoy samples (Good and Oprea, Apr; Hawkins et al., Apr; Mysinger and Shoichet, 2010). To overcome the bias in DUD, Database of Useful Decoys - Enhanced (DUD-E) was introduced in 2012, spanning from a range of 102 targets and 22886 ligands and decoys (Mysinger et al., 2012). The Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt BDB), which contains 9,905 active ligands interacting with 27 nuclear receptors, was established to address the shortage of information and pharmaceutical profile in current nuclear receptor (Lagarde et al., 2014). Despite attempts to build libraries of decoys and actives with care, existing databases still exert biases, due to the limitation of ligand selection, the difference between actives and decoys and the putative decoy determination (Réau et al., 2018).

3.8 Post-docking Evaluation

The goal of virtual screening is to choose a subset of the input library, usually the best-scored ligands. An ideal docking programme would be able to score or rank the true active ligands over the inactives or decoys. Nonetheless, such capability of existing docking programmes is still far from perfect.

Average Rank

The average rank of the active represents the central rankings of all compounds, making it a more useful and less arbitrary metric of enrichment (Fernandes et al., 2004; Kairys et al., 2006). Note that a random sample of evenly distributed actives and decoys should produce an average rank of 50%. The meaning of average rank is that half of the actives would be found before the threshold set at that value. Average rank can take a value of the mean or the median rank of active ligands. However, due to its simplicity and dependence on the number of actives and decoys, average rank is not as common as other sophisticated metrics.

Receiver Operating Characteristic

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems.Receiver Operating Characteristic began in psychology and radiology and is currently used in a variety of disciplines like healthcare, acoustics, meteorology and criminology to analyse the reliability of a particular sensor system and, as a result, to make effective choices based on the given measures. Indeed, ROC curve analysis aids in answering two critical questions: a) In comparison to another system, how successful is the proposed model at recognising known active ligands and rejecting decoys? b) Where should the cutoff be established to discriminate between ligands that should be further investigated and those that should be rejected? (Triballeau et al., 2005)

		True active	
		True	False
Virtual creening	Positive	True Positive (TP)	False Positive (FP)
	Negative	True Negative (TN)	False Negative (FN)

Table 3.1: Confusion matrix of actives and decoys in virtual screening

A confusion matrix is utilised as a tool to better understand the ROC curve, as it enables to quickly calculate the sensitivity and the specificity based on a comparison of given datasets (active/decoy) and subsets from virtual screening (chosen/discarded). Confusion matrix was used in the area of drug development by Manallack et al. (Manallack et al., 2002). The elements in the confusion matrix are: a) the number of true ligands found within the chosen threshold is the true positive (TP), b) the number of inactive or decoys found within the chosen threshold is the false positive (FP), c) the number of true inactives or decoys found within the discarded fraction is the true negative (TN) and d) the number

of true ligands found within the discarded fraction is the false negative (FN) (See Table 3.1). Hence, (TP+FN) is the total number of active ligands and (FP+TN) is the total number of decoys.

Sensitivity (Se) in the area of drug development is defined as the percentage of truly active ligands chosen from a virtual screening process: the number of true positive (TP) ligands divided by the total number of true positives and false negatives (FN).

$$Se = \frac{N_{\text{selected actives}}}{N_{\text{total actives}}} = \frac{TP}{TP + FN} \tag{3.5}$$

This percentage can range from 0 (when all active ligands are unavailable) to 1 (when all active ligands are present). As a result, sensitivity provides information about active ligands that would otherwise be neglected: false negatives. The lower this value, the higher the sensitivity and the better the test for choosing active ligands.

$$Sp = \frac{N_{\text{discarded inactives}}}{N_{\text{total inactives}}} = \frac{TN}{TN + FP} \tag{3.6}$$

Specificity can range from 0 (all inactives are picked) to 1 (all decoys are rejected), providing information on decoys that have been misclassified: false positives. The lower this value, the higher the specificity and the better the test for removing inactive ligands.

Sensitivity and specificity will develop in opposite directions when different thresholds are chosen from the lowest to the highest screening score provided by the virtual screening, ranging between zero and 1. When the cutoff is set to the lowest score, all ligands are chosen, regardless of whether they are actives or decoys, resulting in (Se = 1, Sp = 0). When the cutoff is set higher than the highest score, all ligands are excluded, resulting in (Se = 0, Sp = 1). As a result, optimising both sensitivity and specificity at the same time is unachievable and a tradeoff should be made. The ROC curve enables the user to make such a selection by offering a detailed overview of a screening capacity to distinguish across all selection cutoffs (Zweig and Campbell, 1993).

The ROC curve plots sensitivity against specificity together, Se as a function of $(1 - Sp)$. In other words, at all possible cutoffs, the percentage of actives is plotted against the observed percentage of decoys. A diagonal going from the origin to the upper right corner represents a random classification of the ligands, whereas a virtual screening able to recognise the true actives will have a ROC plot above the diagonal. In the case of ideal distribution where all the actives are retrieved before the decoys, the curve goes up vertically to the upper-left corner ($Se = Sp = 1$) where actives are fully distinguished from the decoys and afterwards joins the horizontal line to the upper-right corner. As a consequence, the stronger the virtual screening result looks like the ROC curve bending towards the upper left corner.

The ROC curve is not as smooth in fact as it appears in the conceptual depiction, but rather jagged and bumpy. Since the sensitivity and specificity can only take discrete values, the confusion matrices would be loaded with integers, resulting in the jagged feature of the ROC curve. Indeed, as the cutoff increases by one, the increase of a true positive result in a vertical line, but the addition of a false positive result in a horizontal displacement. When the ligand library contains more actives and decoys, the curves become less serrated.

The area under the curve (AUROCC) is a useful method of analysing the overall performance of the tests compared to the relative locations of ROC plots. The virtual screening is deemed to be weak if the AUROCC is close to 0.5 (random); the highest feasible AUROCC is 1, which represents an ideal case. The higher the AUROCC, the better the virtual screening method is in distinguishing between actives and decoys. An AUROCC of 0.9 indicates that a randomly chosen active has a probability to have a higher score than a randomly chosen decoy 9 times out of 10. However, this interpretation does not imply that a positive active is found with a chance of 0.9 or that 90 per cent of the chosen ligands are actives (Truchon and Bayly, 2007). According to a suggestion by Swets, values of AUROCC between 0.50 and 0.70 indicate a rather low accuracy while AUROCC values between 0.70 and 0.90 indicate accuracy that is useful depending on the purpose and

higher values indicate a high level of accuracy (Swets, 1988). However, it is known that there is no standard rule to balance the errors (Neyman and Pearson, 1992). The balance between the number of false positive and false negative is left to the subjective decision of the investigator (Hubbard and Bayarri, 2003), depending on how large the subset to be further tested and the available resources. That means choosing a wide threshold can retrieve more active ligands but also more false positive chosen ligands. In fact, choosing a point in the left-upper corner where the curve is skyrocketing would give a significant advantage (Triballeau et al., 2005).

Enrichment Factor

Enrichment Factor (EF) simply describes how many times the actives are found in the best-scored fraction than in the entire dataset. Depending on the chosen threshold, the value of the Enrichment Factor may vary. For instance, the EF threshold can be chosen at 1%, 2% or 5%. Therefore, Enrichment Factor at these thresholds can be annotated EF1%, EF2% or EF5%, respectively. The formula of Enrichment Factor is:

$$EF = \frac{\frac{Active_{subset}}{Total_{subset}}}{\frac{Actives}{Total}} \tag{3.7}$$

or

$$EF = \frac{\frac{TP}{TP+FP}}{\frac{TP+FN}{TP+FP+TN+FN}} \tag{3.8}$$

whereas $Active_{subset}$ and $Total_{subset}$ are the number of actives and the number of compounds at the chosen threshold, *Actives* and *Total* are the number of actives in the entire dataset and the total number of compound in the dataset.

One disadvantage of the Enrichment Factor is that the number is dependent on the value of actives and decoys. For instance, the EF1% of a library of 100 actives and 1,000 decoys would be different from EF1% of a library of 100 actives and 10,000 decoys with the same distribution of actives. Thus, it describes the absolute ratio of actives found with the tested compound library, not for prospective screening. Another disadvantage is that it equally ranks the actives within the threshold, making it difficult to recognise that all actives are ranked at the top of the subset and or just before the threshold (Truchon and Bayly, 2007).

Other Metrics for Early Recognition

It is hard to recognise the distribution of actives within the chosen set with ROC and EF as they treat the actives evenly. More interest to identify the "early recognition" of actives has been growing. Robust Initial Enhancement (RIE) is a metric that uses an exponential weight that decreases as rank increases. The rationale for RIR is that it is less sensitive to big changes than the EF metric when there is only a limited amount of actives. RIE succeeds where ROC fails to recognise if the actives are distributed at the beginning, in the middle or at the end of a set of sorted ligands. RIE takes into account the exponential function of the negative value of ligand rank. Therefore, if the actives are more prone to the beginning of the ranked list compared to the case where actives are more prone to the end, while ROC gives equal values of area under the curves. The meaning of the RIE metric is similar to EF in that it indicates how many times the exponential average of the screening distribution is better than that of random distribution (Truchon and Bayly, 2007). Nonetheless, RIE also suffers a similar disadvantage like EF metric, that minimum value and maximum value are both dependent on the total number of ligands and the number of actives (Truchon and Bayly, 2007). RIE is also reported to be linearly related to ROC (Truchon and Bayly, 2007).

Boltzmann-enhanced discrimination of receiver operating characteristic metric (BEDROC) is another metric proposed to address the "early recognition" issue by utilising a continuously declining exponential weight as a function of ligand rank (Truchon and Bayly, 2007). Fundamentally, BEDROC can be regarded as a weighted modification of the AUROCC value, with the beginning of the ROC curve receiving more weight. As a result, BEDROC reflects early enrichment rather than the overall performance. BEDROC is biased to report higher values with smaller decoy sets (Jain and Nicholls, 2008). Furthermore, BEDROC and RIE are reported to have a linear relationship (Zhao et al., 2009).

In addition, there are other newly developed metrics. Clark et al. presented a new statistic, pROC, which based on the negative logarithmic function of false positive rates, rather than relying on the rankings of actives (Clark and Webster-Clark, Apr). Others metrics that can be counted are Predictiveness Curve (Empereur-mot et al., 2015) or different transformations of EF and ROC .

Although a number of metrics have been developed to address the issue “early recognition” in virtual screening, most of them suffer from the sophistication and may require some while for wider recognition from the community. In the meantime, the most well-documented metrics that are widely employed are Receiver Operating Characteristic and Enrichment Factor. Therefore, in this study, these two metrics are use in addition to average rank of active ligands.

3.9 Pitfalls of Docking

The prediction of molecular docking is primarily the estimation of binding likelihood between small molecules and macromolecules based on the knowledge of intermolecular interactions. Although it has become a more common and essential part of drug design, there are still questions about the application of molecular docking.

The first factor is the structures of the proteins. Protein structures from repositories such as Protein Data Bank are usually obtained from experimental methods, such as X-ray diffraction, nuclear magnetic resonance or electron microscopy. Depending on the method and the condition of the experiment, the protein can be captured at a single conformation or an ensemble of conformations. Therefore, such protein structures downloaded from Protein Data Bank do not necessarily represent the true state of the proteins but rather just a snapshot. As a consequence, docking using such protein structures does not represent the true nature of the binding between the ligands and the proteins. The rigid receptor is one of the most difficult obstacles to overcome in the realm of docking. Depending on the substrate it binds to, a protein can assume a variety of distinct conformations. As a result, docking with a rigid receptor corresponds to a single protein conformation, resulting in false negatives in many cases when the ligand was later discovered to be active. This occurs because a protein might be in a continual state of motion between distinct structural states with similar energies, which is typically overlooked in docking.

Another factor is the reliability of docking programmes. Most available docking programmes have to trade off the speed over the accuracy. Depending on the methods and scoring functions used in docking, the accuracy may vary across different docking programmes. However, such expected accuracy is still far from perfection. In fact, many existing programmes successfully predict the binding mode between the ligand and the target with various accuracy: DOCK6 73.3% (Allen et al., 2015), Autodock Vina at 80% (Trott and Olson, 2010), Gemdock at 79% (Yang and Chen, 2004; Hsu et al., 2011), ADFR at 74% (Ravindranath et al., 2015), Ledock at 75% (Zhang and Zhao, 2016), PLANTS 72% (Korb et al., 2012), PSOVina 63% (Ng et al., 2015), QuickVina2 63% (Alhossary et al., 2015), Smina more than 90% (Koes et al., 2013) and VinaXB 46% (Koebel et al., 2016), in term of binding pose prediction. However, there is still a poor correlation between docking scores and binding free energy (Wang and Zhu, 2016). In addition, to achieve sufficient speed, the docking programmes have to exclude many elements from the environment such as aqueous solvent, ions or pH.

Other factors that can be taken into account are the lack of environment such as water molecules and ions, pH condition, the isomerisation of the ligand, the prediction of the

binding sites. All these aspects affect the final docking performance. Therefore, every step taken with care can reduce the false positive and false negative rates.

3.10 Consensus Scores Improve Docking Performance

As discussed above, four types of scoring functions were developed to tackle the issue of imperfect accuracy. Although the new scoring functions claimed to improve the accuracy, it is still insufficient to use a single scoring function for virtual screening. To overcome this problem, the idea of combining multiple docking programmes and scoring functions has been implemented. Over previous decades, consensus scores have gained popularity, due to their superior performance over individual docking scores (Wang and Wang, 2001; Clark et al., 2002; Feher, 2006). Consensus scores are now becoming the norm (Perez-Nueno et al., 2009; Park et al., 2014), reflecting their success in responding to recent epidemic outbreaks, such as Ebola (Onawole et al., 2018), Zika (Bowen et al., 2019) and SARS-Cov-2 (Amendola et al., 2021). The success of consensus scores is ascribed to the fact that repeated observations are statistically expected to lead to the true value (O'Boyle et al., 2009). A major advantage of consensus scores is the ability to reduce false positives and false negatives in virtual screening, thereby hugely optimising the time and resource of testing. Consensus scores have been employed in both structure-based and ligand-based virtual screening (Oda et al., 2006; Schultes et al., 2015).

Initially conceptualised by Charifson (Charifson et al., 1999), consensus scoring uses one scoring function to rank the poses and another scoring function to re-score the best-docked pose. Another approach is to combine the output from multiple docking programmes and scoring functions for a unique consensus score. Most consensus score protocols use established statistical concepts (Ginn et al., 2000) (summation, minimum, maximum and median of scores or ranks). These values are directly input for the so-called "rank-by-number" and "rank-by-rank" because of their complete compatibility. The prerequisite requirement for statistical consensus scores is that the initial scores must be homologous. For instance, the docking scores were uniformly generated (Wang and Wang, 2001) or rescored with the same docking engine (Stahl and Rarey, 2001; Bissantz et al., 2000). Another way to directly combine the docking scores is to used output from docking programmes based on the same core (for instance, Autodock Vina and Smina) (Masters et al., 2020).

One more method is to combine the output from multiple programmes or scores using data fusion. However, most docking programmes apply various scoring functions, resulting in diverse ranges of docking scores. For instance, docking scores from Autodock Vina usually range from -15 to 0, while docking scores from DOCK vary from -100 to 0. In some cases in DOCK, some outliers obtained extremely high positive values. Therefore, it is essential to bring such different data to a unified scale. In case the docking scores from different docking programmes have different units and ranges, normalisations are applied to bring these values to a unified scale. Many authors have used different normalisation methods for such purposes across the literature. These normalisations include simple normalisations: rank transform (Clark et al., 2002; Feher, 2006), minimum-maximum scaling (Oda et al., 2006; Carta et al., 2007) and z-score scaling (Vigers and Rizzi, 2004; Jacobsson and Karlén, 2006) prior to combination. Although the normalisation leads to scale uniformity, it may sometimes shift the data to another distribution that may lead to partial information loss.

Recently, machine learning models were applied to utilise the docking output with enhanced results (Brylinski, 2013; Fang et al., 2015; Pereira et al., 2016; Ericksen et al., 2017). These machine learning-based consensus scores are sophisticated models and tend to favour specific datasets.

As summarised in this chapter, with the advances in biochemistry and chemoinformatics, MRSA target data and libraries of compounds that can be used for repurposing practice are adequately available or easily obtained. With such an abundance of information, structure-based virtual screening is an appropriate approach. Molecular docking is a

substantial method to exploit the interactions between the compounds and MRSA targets to explore new targets for those compounds. Although molecular docking has been widely recognised by the scientific community and has been integrated into the drug discovery and development process in past decades, the performance of existing docking programmes is still far from perfect. Consensus scores using multiple docking programmes is an alternative way to improve the ability to identify the active compounds but still make use of docking information. Many attempts have been made to address the early enrichment of active compounds in screened libraries via various metrics but Receiver Operating Characteristic and Enrichment Factor are still favoured due to their long establishment.

The next chapter will discuss how the study was carried out. The target proteins were built from *S. aureus* essential genes using sequence alignment and homology modelling. The ligands including approved drugs and compounds from clinical trials that are ready to use for repurposing and were obtained from a tailored library named Repurposing Hub. Ten docking programmes were used to predict the binding likelihood between the ligands and targets. A database of decoys was put in use to benchmark the ability of docking programmes to recognise the active ligands amongst others. Traditional consensus scores were computed to compare with the single docking programmes. After docking, a proposed consensus score was applied to improve the performance of docking methods. Finally, this consensus score was applied to obtain the subset of the potential candidates for repurposing.

Chapter 4

Methodology

1 MRSA Protein Acquisition

1.1 *S. aureus* Essential Genes

The starting point for the task of drug repurposing for MRSA treatment is the Database of Essential Genes (Zhang et al., 2004; Luo et al., 2014), which is a repository containing the essential genes of organisms. These genes are vital to the existence of micro-organisms since they are encoded for structures and functions that play important roles in the growth and reproduction (Itaya, 1995). The collections of essential genes deposited to Database of Essential Genes were determined by experimental methods, whereas the first version MRSA essential gene collection, namely N315, was obtained by antisense RNA method (Ji et al., 2001). The collection was expanded using Transposon-Mediated Differential Hybridisation in 2009, named NCTC 8325 (Chaudhuri et al., 2009). In this work, the revised version of *S. aureus* essential genes are used in the sequence alignment to search for the corresponding protein structures from PDB.

1.2 Gene Sequence Alignment

A list of 351 *S. aureus* gene sequences from the Database of Essential Genes was used to scan throughout Protein Data Bank to find the encoded proteins using the NCBI Basic Local Alignment Search Tool (BLAST) programme (Altschul et al., 1990). BLAST is an application that can identify sequence similarities between known sequences and sequences within a database. The ability to find sequence analogue make it possible to identify prospective proteins from a gene sequence. The Basic Local Alignment Search Tool (BLAST) identifies areas where sequences are locally similar. The program computes the statistical significance of matches between nucleotide or protein sequences and sequence databases. BLAST operates in three steps. Firstly, it cleaves the query sequence into small sequences of typically 3-4 amino acids for proteins or 10-12 nucleotides for DNA sequences. Secondly, these short sequences are used to search for perfect matches across all the entries in the database. Thirdly, when a match is found it then tries to extend the alignment to determine whether this match is part of a longer matching sequence. For each new pair of letters, it evaluates whether it is a good match. If it is a good match then the score is increased and if it is a bad match the score is reduced. The score table for each pair of amino acids or nucleotides is precomputed and incorporated into the BLAST algorithm. When testing on a database of actual sequences, BLAST was effective at rapidly identifying alignments with high scores (Altschul et al., 1990). Here the lasted stand-alone version, BLAST+ (Camacho et al., 2009), was used for sequence alignment of *S. aureus* genes against protein structures from Protein Data Bank. The command line for the alignment was:

blastp -query input -db pdb -remote -out output -entrez_query “Staphylococcus aureus”

whereas *blastp* is the built-in module to search and compare with protein structures, *-query* is the option for input sequence, *input* is the query file containing the gene sequence, *-db* is the option for the target database and *pdb* stands for Protein Data Bank, *-remote* is the option to search for an online database, *-out* is the option for the outcome and *output* is the file containing the results; *-entrez_ query* is the option for a specific organism and *Staphylococcus aureus* is the bacterium of interest.

After the sequence alignment, there can be several possible outcomes: i) The outcome protein has one protein with high identity and full coverage. That protein is chosen as the matching protein of MRSA genes. ii) There is more than one matching protein. Those structures are inspected and selected based on the resolution and availability of the co-crystallised ligand, with favour to the availability of the ligand. iii) The sequence has no matches in PDB. The corresponding gene is kept unattended. iv) The sequence has one or more proteins with moderate identity and coverage. Homology modelling will be applied to identify the structure of the protein.

1.3 Homology Modelling

When there are no structures from PDB that adequately match the *S. aureus* gene sequence, homology modelling was used to build their structures. Homology modelling is a method for the construction of an unknown protein structure from its sequence and an existing structure of homologous proteins. In this study, SWISS-MODEL was used for this particular purpose (Waterhouse et al., 2018). SWISS-MODEL is a web-based server for homology modelling of protein structures. The four major processes in creating a homology model are i) finding protein template(s), ii) aligning the query sequence and template structure(s), iii) model construction, and iv) model quality assessment. SwissModel will automatically choose templates based on the most closely aligned protein sequence that has a three-dimensional structure available for it. When the template search is finished, the output page includes a main table showing the list of available templates ranked according to the expected quality of the resulting models. There are multiple templates which cover the complete sequence and share a considerable sequence identity with our target sequence. Depending on the reference of the user, the template with best match to the query can be choose for modelling. Models can be displayed interactively using the 3D viewer. By default, models are coloured by model quality estimates assigned by QMEAN to highlight regions of the model which are well- or poorly modelled. The target/template alignment is used as input for ProMod-II to create an all-atom model for the target sequence once templates are chosen for model creation, either via the automated or manual selection option Guex and Peitsch (1997). If loop modeling using ProMod-II does not produce satisfying results, MODELLER is used to construct an alternate model (Šali and Blundell, 1993). *S. aureus* gene sequences with low identity and coverage from the previous stage were inputted to the SWISS-MODEL server to search the templates. The template with the best coverage, identity and Global Model Quality Estimate (GMQE) score was then selected for the modelling.

2 Benchmark using Ranking Order as Evaluating Metric

2.1 Ligand and Protein Selection

Since molecular docking methods use known molecular interactions to predict the binding affinities between ligands and proteins, the ability to recognise active compounds is highly dependent on the protein structures utilised and the extent of similarity between the screened ligands and native ligand from the protein-ligand complex (Broccatelli and Brown, 2014; Pinzi et al., 2018; Jain, 2009; Verdonk et al., 2008). To favour the findings toward MRSA treatment, this benchmark intentionally chose the targets that feature the MRSA structural information. The structures of MRSA targets retrieved by sequence alignment and DUD-E targets were compared using the Dali server (Holm and Rosenström, 2010). Those targets that share similar structures were extracted and clustered.

Then the DUD-E decoys and actives were cross-docked against MRSA targets that share similar structures with their targets from DUD-E.

For each ligand set in DUD-E, after filtering with Lipinski's rule, 999 decoys and one active were randomly chosen. Only one active chosen as the median rank, a simple metric, was used to evaluate the docking performance. After docking, for each target, the active was ranked amongst the decoys and the median of all ranks of the ligands was calculated.

2.2 Ligand Preparation

Before the docking against protein targets, ligands needed to be properly prepared. The preparation processed mostly involves three-dimensional (3D) structure generation, protonation and energy minimisation. This was done using OpenBabel (O'Boyle et al., 2011), a computational tool mainly used for chemoinformatics and interconversion between chemical file formats. OpenBabel is a popular and open chemical toolkit to for the inter-conversion of computational chemistry file formats as well as the processing of physiochemical properties of the molecules. Other chemical toolboxes includes RDKit (Landrum, 2010) and CDK (Steinbeck et al., 2003; Willighagen et al., 2017), that also offer quick access to molecular information. One advantage of OpenBabel is being written in C++ and the source code and bindings are available to allow coding using Bash or Python. Up to date, OpenBabel was cited in the reference of more than 3600 articles (Web of Science). OpenBabel has been validated with the error rate in chemistry format conversion and canonicalization algorithm decreased to less than 0.01% and 0.001%, respectively (O'Boyle et al., 2011). Subsequently, OpenBabel was chosen for the processing of chemical properties in this study. Conversion from string formats like SMILES to 3D formats like SDF is made possible by coordinate creation in 3D. The 3D structure generator creates linear elements from the ground up using geometrical rules based on atom-atom hybridization. Ring systems employ single-conformer ring templates. From largest to smallest, the templates are iterated through in the template matching algorithm in search of matches. The process continues if a match is found, but it won't match any previously templated ring atoms unless there is a single overlap (the two ring systems of a spiro group) or an overlap involving precisely two nearby atoms (two fused ring systems). The stereochemistry (cis/trans and tetrahedral) is adjusted to match the input structure after an initial structure has been produced (O'Boyle et al., 2011).

Decoys and actives were already available in MOL2 and SDF formats with hydrogen atoms added. Therefore, only chemical format conversion was needed. Depending on the requirement of each docking programme, an appropriate format was obtained using the OpenBabel programme. ADFR, PSOVina, QuickVina2, Smina, Autodock Vina and VinaXB require the input ligands in PDBQT format. DOCK, Ledock and PLANTS use MOL2 as the default format of the ligands. SD format is required by rDock while Gemdock prefers MOL extension.

2.3 Protein Preparation

The preparation of MRSA protein targets mainly consisted of residue correction, protonation, binding site prediction and grid generation. First, the protein structures were inspected for any wrong or collided residues. Next, the protonation was accomplished using a built-in *DockPrep* module in Chimera (Pettersen et al., 2004). The prepped structure was saved in MOL2 or PDB format. Protein structure in MOL2 format was required for DOCK and rDock and PDB was for Gemdock and Ledock. PLANTS also needed MOL2 but the preparation was conducted using its own companion *SPORES* (ten Brink and Exner, 2009, 2010). Ledock also used its preparation tool *lepro* (Zhang and Zhao, 2016) to automatically process the protein and generate an input file for docking. Conversion to PDBQT for ADFR, PSOVina, QuickVina2, Smina, Autodock Vina and VinaXB was carried out in AutodockTool4 (Morris et al., 2009). For ADFR, it used its own Autosite (Ravindranath and Sanner, 2016) module to predict the binding site.

For other programmes, binding site prediction mainly relied on the FTSite server (Ngan

et al., 2012). The output included three predicted clusters with an order from high to low measure. These predicted clusters were visually compared with the co-crystallised complex. Usually, the predicted cluster (or clusters) that coincided with the ligand from the complex (if present) was assigned as a binding site with confidence. In case more than one cluster coincided with the ligand, all of them were selected.

Autodock and other derivatives required an input configuration file containing the information about the receptor, the ligands and the binding site instead of putting all details in the command line. The binding site parameters were generated using AutodockTools4 (Morris et al., 2009). The chosen clusters from FTSite output were used to define a box with a minimum site that contains all the residues in the clusters. The grid box format is with one centre (x, y, x coordinates) and sizes. Meanwhile, Ledock also accepted the same box but is defined with coordinates of 8 corners. PLANTS required a sphere with the same centre and the radius is calculated as half of the main diagonal. This was to ensure the least dissimilarity in the binding pocket between each docking programme.

2.4 Docking of Ligands and Proteins

Ten docking programmes were chosen in view of their ease of use and prominence: ADFR (Ravindranath et al., 2015), UCSF DOCK (Allen et al., 2015), Gemdock (Yang and Chen, 2004; Hsu et al., 2011), Ledock (Zhang and Zhao, 2016), PLANTS (Korb et al., 2012), PSOVina (Ng et al., 2015), QuickVina2 (Alhossary et al., 2015), Smina (Koes et al., 2013), Autodock Vina (Trott and Olson, 2010) and VinaXB (Koebel et al., 2016). All protein structures chosen above were downloaded from the Protein Data Bank (PDB) (Berman et al., 2000, 2002). Prior to docking, protein structures were stripped off small molecules, ion and water molecules, followed by protonation. Decoys and ligands were prepared in a three-dimensional structure with an appropriate format.

Binding site prediction was carried out using FTSite server (Ngan et al., 2012) for ADRF, DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, Smina, Autodock Vina and VinaXB while rDock used its own package (Ruiz-Carmona et al., 2014). Finally, 999 decoys and one active ligand were docked against each chosen MRSA target. Each docking programme generated various conformation of ligands within the binding pocket and used its underlying scoring function to estimate the likelihood of binding for each ligand conformation. Only the best-scored pose was retained for each decoy and ligand. All protein structures used here were downloaded from the Protein Data Bank (PDB) (Berman et al., 2000, 2003). All decoys and actives were docked against 29 targets using 10 docking programmes. These programmes have been benchmarked in other works but inconsistent due to various dataset. Therefore in this study, benchmark of these docking programmes was carried but oriented to MRSA targets. The parameters were set in line with those used in published works to prevent the abundance of docked poses and excess amount of running time.

2.4.1 ADFR

ADFR (Ravindranath et al., 2015) used its package Autosite (Ravindranath and Sanner, 2016) to generate a TRG file containing the geometry information. The docking procedure also required target and ligand in PDBQT format. PDBQT ligands were docked against the PDBQT receptor. In ADFR, the ADFR score uses an energy function which is a weighted sum of terms representing van der Waals, hydrogen bond, electrostatic, and desolvation contributions, computed between pairs of atoms.

$$\begin{aligned} E = w_{vdW} \sum_{i,j} \left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}} \right) + w_{hbond} \sum_{i,j} E(t) \left(\frac{C_{i,j}}{r_{i,j}^{12}} - \frac{D_{i,j}}{r_{i,j}^{10}} \right) \\ + w_{elec} \sum_{i,j} \left(\frac{q_i q_j}{\epsilon r_{i,j}.r_{i,j}} \right) + w_{sol} \sum_{i,j} \left(S_i V_j + S_j V_i \right) e^{\frac{-r_{i,j}^2}{2\sigma^2}} \end{aligned} \tag{4.1}$$

ADFR uses this energy formula to estimate the affinity between atoms from three groups: Ligand (L), Rigid Receptor (RR) and Flexible Receptor (FR). The final score is the summation of these interaction terms:

$$S_{ADFR} = E_{L-L} + E_{L-RR} + E_{L-FR} + E_{FR-FR} + E_{FR-RR} \tag{4.2}$$

Only the first two terms, ligand intramolecular and ligand-rigid receptor intermolecular interactions—are taken into account when a rigid receptor is involved. When receptor atoms are appointed flexible, the additional terms (ligand-flexible receptor inter-molecular, flexible-flexible receptor inter-molecular, and flexible-rigid receptor inter-molecular interactions) are automatically incorporated into the scoring functions. Each term in the scoring function can have a weight assigned to it.

2.4.2 DOCK

UCSF DOCK (Allen et al., 2015) uses standard protein preparation starting from two structures, one of which encapsulates a protein appended with hydrogens and charges, using the Dock Prep module in software Chimera (Pettersen et al., 2004) and saved in MOL2 format for docking performance. The other structure represented a hydrogen-stripped protein prepared for the generation of a molecular surface using module DMS in Chimera. The molecular surface of the protein was then generated by rolling a ball of the size of a water molecule over the Van der Waals surface of the protein. Next, collections of overlapping spheres at surface invaginations were produced using SPHGEN and only the largest sphere associated with each surface atom is kept. The sphere collection was then clustered using a linkage algorithm. All spheres within 10Åof each atom within the co-crystallised ligand with the protein were retained for grid generation. Finally, the module GRID was used to prepare Van der Waal and electrostatic energy grids, which were used to speed up docking calculations. MOL2 ligands were docked against receptors using rigid docking. The primary energy scoring component of DOCK is a type of force field scoring, consisting of van der Waals and electrostatic components similar to the terms in ADFR:

$$E = \sum_{i=1}^{lig} \sum_{j=1}^{rec} \left(\frac{A_{i,j}}{r_{i,j}^{a}} - \frac{B_{i,j}}{r_{i,j}^{b}} + 332 \frac{q_i q_j}{D_{ij}} \right) \tag{4.3}$$

where each term is a double sum over ligand atoms i and receptor atoms j. In latest version, DOCK was added with new scores, including Hawkins score, Poisson–Boltzmann with solvent-accessible surface area solvation score and Amber score (Goodford, 1985; Meng et al., 1992; Lang et al., 2009; Allen et al., 2015).

2.4.3 Gemdock

Gemdock (Yang and Chen, 2004; Hsu et al., 2011) used target protein structures downloaded from the PDB removing all water molecules and irrelevant atoms. The position and size of the binding site were determined by taking into account all protein atoms with a distance less than 8Åfrom each atom of ligand. The ligand was then removed. Docking of ligands and protein was carried out with module mod_ga which defines the core of Gemdock. Gemdock initialised the orientation and conformations of ligands to generate an initial population size of 200. For each ligand screened, Gemdock stopped when the generation number reaches 70. The score and pose for each ligand were then saved. Gemdock used an empirical scoring function given as:

$$E_{tot} = E_{inter} + E_{intra} + E_{penal} \tag{4.4}$$

where E_{inter} and E_{intra} are the intermolecular and intramolecular energy, respectively, and E_{penal} is a large penalty number if the ligand is outside of the search range. E_{penal} is set to the value of 10,000.

The intermolecular energy is defined as:

$$E_{inter} = \sum_{i=1}^{lig} \sum_{j=1}^{pro} \left[F\left(r_{ij}^{B_{ij}}\right) + 332.0 \frac{q_i q_j}{4 r_{ij}} \right] \tag{4.5}$$

where r_{ij} is the distance between the atoms i and j, q_i and q_j are the formal charges, and 332.0 is a converting factor from the electrostatic energy into kilocalories per mole. The lig and pro are the numbers of the heavy atoms in the ligand and receptor, respectively. $F\left(r_{ij}^{B_{ij}}\right)$ is a simple atomic pairwise potential function. $\left(r_{ij}^{B_{ij}}\right)$ is the distance between the atoms i and j with the interaction B_{ij} made by the pairwise heavy atoms between ligands and proteins where B_{ij} is either a hydrogen or a steric bond.

The intramolecular energy of a ligand is:

$$E_{intra} = \sum_{i=1}^{lig} \sum_{j=i+2}^{lig} F\left(r_{ij}^{B_{ij}}\right) + \sum_{k=1}^{dihed} A[1 - cos(m\theta_k - \theta_0)] \tag{4.6}$$

where $F\left(r_{ij}^{B_{ij}}\right)$ is atomic pairwise potential function but the value is 1000 to reject unrealistic conformations when $\left(r_{ij}^{B_{ij}}\right)$ is less than 2.0 Åand dihed is the number of rotatable bonds (Yang and Chen, 2004; Hsu et al., 2011).

2.4.4 Ledock

Ledock (Zhang and Zhao, 2016) used protonated conformations with hydrogens stripped from proteins using lepro. Ledock requires a configuration file including a binding cavity box. Binding pockets were detected using the FTSite server (Ngan et al., 2012). The binding cavity box was defined by a lower and upper coordinate in the x-axis, y-axis and z-axis. Ligands in MOL2 format were docked into protein with default parameters and docked poses were returned in DOK format.

$$\Delta G_{bind} = \alpha \sum_{i \in lig} (E_i^{vdW} + E_i^{hb}) \Theta(E_{co} - E_i^{vdW} - E_i^{hb}) + \beta(r) \sum_{i \in lig} \sum_{j \in pro} \frac{q_i q_j}{r_{ij}} + \gamma E_{lig}^{strain} \tag{4.7}$$

The first term is the summation of van der Waal interaction E_{vdw} and hydrogen bonding energy E_{hb}, where *theta* is the Heaviside step function and E_{co} is the limit energy to enable soft docking. The second term is the electrostatic interaction energy with a distance function

$$\beta(r)$$

accounting for both electrostatic screening and desolvation effects, where q is the partial atomic charge and r is the distance between pairwise atoms. The third term is the ligand conformational strain upon binding, and is made up by the intramolecular clash and/or torsion strain. Coefficients α, β and γ were empirically identified (Zhang and Zhao, 2016).

2.4.5 PLANTS

To confirm compatibility with PLANTS (Korb et al., 2012), protein targets and the ligands were protonated with SPORES (ten Brink and Exner, 2009, 2010), mode = complete. The binding site sphere was defined using the same coordinates from FTSite (Ngan et al., 2012), with a little modification. PLANTS required a sphere defined by a centre and radius. The centre of the binding site sphere was the same as the centre of the grid box from FTSite and the radius was calculated as half of the internal diagonal. The virtual screening of PLANTS was done with mode = screen. Two empirical scoring functions are offered in PLANTS: the CHEMPLP scoring function and a modified piecewise linear potential PLP version. The PLP scoring function, f_{PLP}, used in PLANTS is modelled after those described in

Gehlhaar et al. (1995) and Verkhivker (2004) using just distance-based potentials. It has the following structure:

$$f_{PLP} = f_{plp} + f_{tors-lig} + f_{clash-lig} + 0.3f_{score-prot} - 20.0 \tag{4.8}$$

In the first component, called f_{plp}, steric interactions between the protein and the ligand are primarily modeled. Metal ions in the protein binding site are taken into account, as well as the occlusion of polar atoms by nonpolar ones via distance-based potentials. The parameter r represents the Euclidean separation between a ligand and a protein atom, and the resultant potential value is PLP(r). Depending on the type of protein and ligand atom, parameters A to F define the form of the potential. A simple clash term ($f_{clash-lig}$), which prevents the atoms of the ligand from getting too close together, plus a torsional potential make up the intramolecular ligand scoring function ($f_{tors-lig}$) (Korb et al., 2007). The scoring function CHEMPLP, abbreviated as $f_{CHEMPLP}$, has the functional form:

$$f_{CHEMPLP} = f_{plp} + f_{chem-hb} + f_{tors-lig} + f_{clash-lig} + 0.3f_{score-prot} - 20.0 \tag{4.9}$$

The PLP scoring function mentioned above is used in the first part (f_{plp}) of the intermolecular score, despite utilizing different parameter settings. In the second part ($f_{chem-hb}$), the hydrogen bonding and metal-acceptor interactions between the protein and the ligand are taken into account. The protein and intramolecular ligand terms are the same as those that were mentioned in the PLP case (Korb et al., 2007). Finally, a penalty term is introduced to both PLP and CHEMPLP scoring functions if the ligand falls outside the predetermined binding site of the protein (Korb et al., 2007).

2.4.6 Autodock Vina

For Autodock Vina (Trott and Olson, 2010) and other derivatives (PSOVina (Ng et al., 2015), QuickVina2 (Alhossary et al., 2015), Smina (Koes et al., 2013), VinaXB (Koebel et al., 2016)) both proteins and ligands were prepared in PDBQT format. The docking was carried out with parameters from the configuration file. The maximum iteration of running with the option *exhaustiveness* was set to 20. The output files included a PQDBQT file which contained the same number of docked poses as in the option *exhaustiveness* and a log file which contained all of the binding affinities and RMSD scores. The first pose was regarded as the best-docked pose in the Autodock Vina output log file and had the RMSD value of 0Å. The RMSDs of the rest of the poses were calculated from this pose. The sum of distance-dependent atom pair interactions is used to predicte the binding energy.

$$E = \sum e_{pair}(d) \tag{4.10}$$

Here d is the surface distance calculated of the pairwise atoms. Every pair of atom interacts through a steric interaction. Depending on the type of the atoms, additional hydrophobic and non-directional H-bonding interactions could be added:

$$e_{pair} = \begin{cases} w_1 * Gauss_1(d) + \\ w_2 * Gauss_2(d) + \\ w_3 * Repulsion(d) + \\ w_4 * Hydrophobic(d) + \\ w_5 * HBond(d) \end{cases} \tag{4.11}$$

whereas w_1, w_2, w_3, w_4, w_5 are predefined weights for each term (Trott and Olson, 2010).

2.4.7 PSOVina

To address the conformational search problem in docking, PSOVina merged the particle swarm optimization (PSO) algorithm with the effective Broyden-Fletcher-Goldfarb-Shannon (BFGS) local search approach used in AutoDock Vina. The position, orientation,

and torsional angle of each rotatable bond collectively produce a solution vector, which is a potential ligand conformation, in the flexible ligand approach. The current search challenge is to identify the solution vector that produces the lowest Vina scoring function score (Ng et al., 2015).

2.4.8 QuickVina2

Autodock Vina uses the BFGS method and the Markov chain of the modified Monte Carlo algorithm with restart, respectively, to explore the molecular docking search space. The element of optimization that requires the most time is local search. The first-order-necessary-condition heuristics in QVina2 limit the use of local search to docked conformation candidates that are considered important. This is made possible by maintaining a circular database of 10N last-assessed docked conformations, where N is the total number of design variables. As many as 2N nearest (in terms of Euclidean distance) neighbors for each newly randomized candidate of docked conformation are obtained from the database, and then a significance test is run to assess whether a local search from docked conformation is required. (Alhossary et al., 2015)

2.4.9 Smina

In addition to the Gaussian, repulsion, hydrogen bonding, and hydrophobic terms that make up the Autodock Vina scoring function, an electrostatic term, an AutoDock4 desolvation term (Morris et al., 2009), a non-hydrophobic contact term, and a Lennard-Jones 4–8 van der Waals term are added to the scoring function. Only heavy atom interactions between the ligand and protein atoms are considered in docking (Koes et al., 2013).

2.4.10 VinaXB

VinaXB uses an halogen bond scoring function based on Autodock Vina scoring function. An empirical scoring function for halogen bonding is presented along with its implementation in AutoDock Vina. The halogen bonding term is defined based on the overlap of van der Waals radii of interacting atoms. Due to the anisotropic charge on halogen, an angle term accounts for the varying positive charge on the atom. The XBSF scoring function (E) is defined using these three terms: weight, angle factor, and distance factor as follows:

$$E = W\theta D \tag{4.12}$$

where W = weight, θ = angle factor, D = distance factor. (Koebel et al., 2016)

2.5 Traditional Consensus Scores

To compare with individual docking programmes and other consensus scores, the most common methods of normalisation were applied to bring docking scores to their united representations before combination. The three most common normalisation procedures were employed:

i) Rank normalisation - Ranks represent docking scores for each target assigned against ascending ranks. This implies that ligands with more negative scores rank higher on this scale. Each docking score in one ligand set was replaced by its position (rank) in the ordered array counted from the smallest value (most negative).

ii) Minimum-maximum normalisation (henceforth referred to as min-max normalisation), also known as min-max scaling or [0-1] scaling, is a simple method of transforming the entire range of values to the range of [0, 1]. The normalised docking scores were computed by:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{4.13}$$

where x' is the normalised docking score, x is a primary docking score, min(x) and max(x) is minimum and maximum docking score from the same ligand set for each target, respectively.

iii) z-score normalisation (or standardisation) is a method to transform data to a distribution with a mean of zero and a standard deviation of 1. The meaning of a z-score gives an idea of how far from the centre of the data. The z-score of each docking score was calculated by:

$$x' = \frac{x - \mu}{\sigma} \tag{4.14}$$

where x' is the normalised docking score, x is a primary docking score, μ is the mean of the docking score set for each target and σ is the standard deviation of the docking score set.

A drawback of these normalisation methods is that they shift the relative distribution of scores towards each other. This may cause a loss of information. For example, in a set of ligands [A, B, C] with docking scores of [-8 -4, -6], rank normalisation will return a list of ranks [1, 3, 2]. This normalisation method gives a rough idea of the relative position of a ligand of interest in the entire list. However, the above list will get the same list of ranks with a set of ligands [a, b, b] with docking scores of [-10, -4, -6]. Hence, the difference in the absolute value is lost after rank normalisation.

Traditional consensus score refers to those scores using statistical concepts such as minimum, mean, maximum to combine the docking scores from multiple programmes or scoring functions. These consensus scores have been used in numerous virtual screening studies. In this study, 8 consensus scores were used to compare the joint performance of docking programmes compared to the individual programmes. The scores Mean (MEAN), Median (MED), Minimum (MIN), Maximum (MAX) (Ericksen et al., 2017), Deprecated Sum Rank (DSR) (Willett, 2013) and Euclidean Distance (EUC) (Feher, 2006) were most common amongst publications. Cubic Mean score (CBM) was added in line with Euclidean Distance score. A newly developed score Exponential Consensus Rank (ECR) (Palacio-Rodríguez et al., 2019) was also exploited. These consensus score lines across ten sets of normalised docking scores (S_i) were calculated as follows:

$$MEAN = \text{mean}\{S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}\} \tag{4.15a}$$

$$MED = \text{median}\{S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}\} \tag{4.15b}$$

$$MIN = \text{minimum}\{S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}\} \tag{4.15c}$$

$$MAX = \text{maximum}\{S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}\} \tag{4.15d}$$

$$EUC = \left[\sum_{i=1}^{10} {S_i}^2\right]^{1/2} \tag{4.15e}$$

$$CBM = \left[\sum_{i=1}^{10} {S_i}^3\right]^{1/3} \tag{4.15f}$$

$$ECR = \sum_{i=1}^{10} \exp(S_i) \tag{4.15g}$$

$$DSR = \frac{\sum_{i=1}^{10} S_i}{\text{maximum}\{S_i\}} \tag{4.15h}$$

Here the traditional consensus scores were calculated based on normalised docking scores across 10 docking programmes for each ligand:target pair. The MEAN, MEDIAN, MIN and MAX scores take the mean, median, minimum and maximum values of such normalised docking scores, respectively. The Euclidean Distance and Cubic Mean scores take the root mean square and cubic mean of the scores accordingly. The Exponential Consensus Rank takes the rank of the docking scores or in another way, rank-normalised

scores, assuming that each docking score is scored with the best pose of each ligand. Similarly, the Deprecated Sum Rank calculates the sum value of the rank of the docking scores, after discarding the worst rank.

Finally, the active ligands were ranked amongst each ligand set and the median rank was calculated for each consensus score.

2.6 Novel Consensus Scores

Molecular docking is a procedure to generate different conformation of poses of ligands for predicting the intermolecular interactions based on varying sets of physicochemical properties, e.g. hydrogen bonding, hydrophobicity, hydrophilicity and a multitude of others. The consensus scoring approach takes into account these interactions to design an overall score that depicts the ensemble representation of the 3D molecule rather than its pose specific description. To avoid information loss while using normalisation methods, in this work, the novel consensus algorithms statistically combined raw information from all docking platforms and then outlined four independent optimised functional ensemble representations of the real molecule in the real solvent:

$$S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^n \tag{4.16a}$$

$$S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[{S_i}^n] \tag{4.16b}$$

$$S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{S_i})^n \tag{4.16c}$$

$$S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[(S_i - \bar{S_i})^n] \tag{4.16d}$$

Here S_c is the combined score, S_i is the docking score of ligands for programmes i = 1, 2,... 10, $x_{i,j}$ are coefficients of the docking programmes i (ADFR, DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, Smina, Autodock Vina and VinaXB) that are the weight factors of those docking outcomes in the combinatorics, in the j^{th} iteration, $\bar{S}$ is the mean of the score set from the programme i, n represents the combinatorial order real values only (n = 1 implies linear combination). Equations 4.16a-4.16d were iterated over a total of approximate $\binom{29}{9}$ ensembles involving 10 docking programmes, each weighing between 0 and 1, incremented in steps of 0.05 each. The rank of active ligand before and after combining was then compared to evaluate the improvement produced by the novel consensus algorithm. The pseudo-code for these models provided in Appendix 6.

In this benchmark, primary docking scores from diverse docking platforms were directly combined representing the entire ensemble. For comparison purposes, various normalisation methods were also used to bring the diverse docking output to a unified scale for traditional consensus scores. Although consensus scores were widely used in virtual screening, it is not clear how many programmes should be inputted to achieve the most efficient consensus outcome. One computational experiment was carried out by O'Boyle by generating putative scores and suggested at least 4 programmes should be used for consensus scores (O'Boyle et al., 2009). In this work, the effect of the number of docking programmes over the novel consensus models was also exploited.

3 Benchmark using ROC and EF as Evaluating Metrics

After running docking with ADFR, the running time for docking was reported particularly prolonged compared to other docking programmes. For that reason, ADFR was substituted with rDock in this benchmark.

3.1 Ligand and Protein Selection

In order to benchmark docking programmes for MRSA targets, the same subset of MRSA proteins and ligands from the Database of Useful Decoys - Enhanced (DUD-E) were selected like in the previous benchmark (Section 2.1. However, in this section, the metrics Receiver Operating Characteristics (ROC) and Enrichment Factor (EF) were used instead of median rank. The combinations of possible targets and the corresponding set of decoys and actives resembled the combinations in the previous benchmark. A set of 1000 decoys and 40 active ligands were randomly chosen for each target.

3.2 Ligand and Protein Preparation

The ligands and targets were prepared similarly as in the previous benchmark (Section 2.2 and 2.3. The ligands were undertaken conversion to the three-dimensional structure along with protonation and energy minimisation using OpenBabel. Meanwhile, the target structures were prepped with protonation, residue corrections and appropriate chemical format conversions.

3.3 Docking of Ligands and Proteins

Ten docking programmes were chosen in view of their ease of use and prominence: UCSF DOCK (Allen et al., 2015), Gemdock (Yang and Chen, 2004; Hsu et al., 2011), Ledock (Zhang and Zhao, 2016), PLANTS (Korb et al., 2012), PSOVina (Ng et al., 2015), QuickVina2 (Alhossary et al., 2015), rDock (Ruiz-Carmona et al., 2014), Smina (Koes et al., 2013), Autodock Vina (Trott and Olson, 2010) and VinaXB (Koebel et al., 2016). As mentioned above, ADFR was substituted with rDock, due to the lengthened running time.

All MRSA protein structures chosen were downloaded from the Protein Data Bank (PDB) (Berman et al., 2000, 2002). After the preparation of ligands and proteins, the ligands were docked into protein at the binding site. Docking of 9 other docking programmes was carried out similar to docking in the previous benchmark (Section 2.4).

rDock

For rDock (Ruiz-Carmona et al., 2014), the search space was automatically created with the following conditions, using the crystal structure ligand coordinates as a reference: small sphere = 1.0, max cavities = 1; radius = 6.0; small sphere = 1.0; max cavities = 1; radius = 6.0; small sphere = 1.0; small sphere = 1.0 To allow some motion for target H-bond donors and acceptors, rDock was run with receptor flex = 3.0. rDock returned SCORE.TOTAL and SCORE.INTER for each pose. Although these scores are highly associated, SCORE.INTER performed somewhat better on average (Ericksen et al., 2017), hence it was utilised for all evaluations. In this project, the number of the maximum run was set to 100 which is the recommended setting for exhaustive docking. Docking files consisted of an SD file containing docked poses. The intermolecular (S_{inter}), ligand intramolecular (S_{intra}), site intramolecular (S_{site}), and external constraint terms ($S_{restraint}$) are weighted sums that make up the rDock master scoring function (S_{total}). The major term of importance is sinter, which stands for the protein-ligand interaction score (or RNA-ligand interaction score). The ligand conformation's relative energy is represented by S_{intra}. Similar to S_{site}, this term denotes the relative energy of the active site's flexible regions. $S_{restraint}$ is a set of non-physical restraint functions that can be used to the docking calculation to influence it in a number of beneficial ways (Ruiz-Carmona et al., 2014).

$$S_{total} = S_{inter} + S_{intra} + S_{site} + S_{restraint} \tag{4.17}$$

3.4 Traditional Consensus Scores

In addition to rank normalisation, min-max normalisation and z-score normalisation, quantile normalisation was added to provide more diversity. Quantile normalisation is a statistical technique for making two datasets statistically equal. To quantile normalise two or

more datasets, first, they are sorted, then the average (often, the arithmetic mean) of the datasets is calculated. Next, the greatest value in all is turned into the mean of the highest values, the second-highest value is turned into the mean of the second highest values, and so forth. Quantile normalisation was preferred over the more popular z-score and min-max normalisation, because modified score distributions reach a common shape, ensuring equal weights among programme scores, hence docking score outliers were likewise unaffected (Ericksen et al., 2017). This makes up to 4 normalisation schemes to bring docking scores to a unified distribution.

Eight traditional consensus scores were used (Mean (MEAN), Median (MED), Minimum (MIN), Maximum (MAX), Euclidean Distance (EUC), Cubic Mean (CBM) (Feher, 2006), Exponential Consensus Rank (ECR) (Palacio-Rodríguez et al., 2019) and Deprecated Sum Rank (DSR) (Willett, 2013)) across ten sets of normalised docking scores, similar to previous benchmark (Section 2.5).

3.5 Novel Consensus Scores

In addition to Equations 4.16a-4.16d in the previous benchmark, in this section, a new descriptor is additionally explored to examine the ability to discriminate between active ligands and decoys. Standard deviation is used in two models, along with previous models without descriptor and with mean.

$$S_c = \sum_{i=1}^{10} x_{i,j} \left(S_i - S\bar{D}_i\right)^n \tag{4.18a}$$

$$S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[(S_i - S\bar{D}_i)^n] \tag{4.18b}$$

Here S_c is the combined score, S_i is the docking score of ligands for programmes i = 1, 2,... 10, $x_{i,j}$ are coefficients of the docking programmes i (DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, rDock, Smina, Autodock Vina and VinaXB) that are the weight factors of those docking outcomes in the combinatorics, in the j^{th} iteration, $S\bar{D}_i$ is the standard deviation of the score set from the programme i, n represents the combinatorial order real values only (n = 1 implies linear combination). Equations 4.18a-4.18b were iterated over a total of approximate $\binom{29}{9}$ ensembles involving 10 docking programmes, each weighing between 0 and 1, incremented in steps of 0.05 each. AUROCC and EF05 of each ligand set (containing decoys and actives) before and after combining were then compared to evaluate the improvement produced by the proposed consensus algorithm. The pseudo-code for these models provided in Appendix 6.

3.6 Consensus Score Evaluation

For post-docking analysis, the docking scores are sorted from the best to the worst in terms of favor to binding affinity. In most of the cases, ligands with best docked scores are usually selected up for further investigation. It is expected that the active ligands or the true drugs are found within the list of top scores. If the docking performance is good enough, a large proportion of active ligands would be found with a small fraction of total docked ligands.

In this benchmark, two typical metrics were used to highlight the improvement. The first one is Receiving Operating Characteristics (ROC), a well-established metric to measure the discrimination between two populations. A Receiver Operating Characteristic curve is a graphical representation of the analytical performance of a binary classifier methodology at various classification cutoffs whereas the area under the ROC curve (AUROCC) represents the extent or measure of discrimination. Although AUROCC is a global measure of overall performance and it is independent of the number of actives and inactives, it does not emphasise early recognition, which is a concern in virtual screening practice. Therefore, Enrichment Factor (EF), a measure to estimate how good one subset shifted toward one extremum of the entire dataset is additionally employed. EF is measured by

the ratio of positively predicted actives in the chosen percentage of best-ranked ligands, divided by the ratio of active equally spread among datasets.

There is no standard for how large is the subset of best scored ligands. It is subjected to the experience of the performer and the trade-off between the number of ligand retrieved and the cost to test the subset. When the threshold is bigger (for example 10%), more active ligands are retrieved but the time and cost would increase. If the threshold is too small (for example 0.1%) very few active ligands are found within the top scored ligands, hence insufficient. Usually a threshold is set for a subset depending on how the researchers are willing to sacrifice the cost over the ratio of active ligands retrieved. It is advised to choose a threshold of 0.5%, 1%, 2% or 5% of the entire ensemble (Jain and Nicholls, 2008). In this study, the cut-off was was chosen at 0.5% without sacrifice the number of active ligands. The AUCROC and EF at the threshold of 0.5% (abbreviated as EF05), were calculated, resulting in approximate 30 chosen ligands for target.

AUROCC and EF05 were computed for each ligand set across all targets. The mean values of AUROCC and EF05 were calculated to represent each docking programme and consensus scores, including traditional and novel consensus scores. Here the mean values were calculated instead of median values since for EF05, in some consensus scores, a large number of ligand sets return an EF05 value of zero, resulting in a median of zero too.

4 Docking of Repurposable Ligands to MRSA Targets

4.1 Ligand Selection

After benchmarking with MRSA targets and a compound library containing decoys and actives from DUD-E, the exact procedure was applied to the prospective set of ligands and the full range of MRSA proteins.

The ligands were selected based on the library from Repurposing Hub, a library specifically developed for drug repurposing. Repurposing Hub is a repository containing approved drugs and clinical trial drugs that can be exploited for repurposing intention. Initially launched in 2017, Repurposing Hub contained 5691 compounds and this number has increased to 6798 compounds (16 September 2021, https://clue.io/repurposing) (Corsello et al., 2017). Recently, Stokes et al. discovered a compound, halicin, which is a potential broad-spectrum antibiotic from Repurposing Hub (Stokes et al., 2020).

In order to cut down unnecessary time by discarding the compounds unlikely to be drugs, drug-like properties are usually applied to increase the successful rate. There are a number of rules attempted to generalise the drug likeness, such as Lipinski's rule (Lipinski et al., 1997), Viber rule (Veber et al., 2002), Waring rule (Waring, 2009) or Golden Triangle rule (Johnson et al., 2009). These rules generally cover physiochemical properties which are essential for a drug such as absorption, permeability or distribution. Lipinski's rule is the most commonly used rule for an oral drug. It is a rule in which an oral drug has to meet at least 3 out of 4 following criteria: the number of hydrogen bond donors is no more than 5 (the total number of nitrogen–hydrogen and oxygen-hydrogen bonds), the number of hydrogen bond acceptors is no more than 10 (all nitrogen or oxygen atoms), the molecular mass is less than 500 daltons and the octanol-water partition coefficient (log P) is no more than 5. Additionally, a threshold of 10 rotatable bonds was set for the compounds. After filtering, 5092 compounds remained.

Next, the same procedure is applied to generate a three-dimensional structure from the SMILE format. First, the SMILE string was used by the OBBuilder to create a 3D structure using rules and fragment templates. Then, 250 steps of a steepest descent geometry optimisation with the MMFF94 forcefield were carried out. Next, 200 iterations of a Weighted Rotor conformational search (optimising each conformer with 25 steps of the steepest descent) were performed. Finally, 250 steps of conjugate gradient geometry optimisation were implemented. Depending on the input requirement, the formats of these ligands were converted to suit each docking programme.

4.2 Protein Selection

MRSA is a malignant pathogen and is named in WHO's priority list for medication. In order to reduce the cost and time of the drug discovery process for MRSA, in this work, the drug repurposing approach with the help of structure-based virtual screening is exploited. The MRSA targets are obtained based on the list of MRSA essential genes. Essential genes are those that are essential for cellular survival. These genes make up the bare minimum of a live cell's gene set. As a result, the functions encoded by this gene set are critical and could even be called a basis of life (Itaya, 1995). Critical gene products of microbial cells are attractive novel targets for antibacterial medications because essential cellular functions are the targets for most antibiotics. Therefore, in this work, essential genes are the source for finding MRSA antibiotics.

The Database of Essential Genes is a repository containing an indispensable set of genes for a wide range of microorganisms. It was initially launched in 2004 and frequently updated, it contains 53,885 essential genes and 786 essential non-coding sequences from 85 species, including bacteria, archaea and eukaryotes (Luo et al., 2021). Specifically, for MRSA, the first version included 302 essential genes in 2001 (Ji et al., 2001; Forsyth et al., 2002) and was updated to 351 genes in 2009 (Chaudhuri et al., 2009). Sequence alignment was carried out for MRSA essential genes against the Protein Data Bank using BLAST+ (Camacho et al., 2009) to find the matching proteins. Those hits with high identity and coverage were directly processed with docking. The sequences with no hits were discarded and those with moderate identity and coverage were input for SWISS-MODEL (Waterhouse et al., 2018) for the prediction of the protein structures.

4.3 Docking of Repurposable Ligands Against MRSA Hits

The docking of ligands from Repurposing Hub against MRSA targets was carried out in the same fashion as in the benchmark of MRSA targets. Prior to docking, overlapped residues (if available) were also corrected and the protein structures were stripped off the water, ion and trivial molecules, followed by protonation and the binding site prediction (See Section 2.3). The ligands were processed through three-dimensional structure generation, protonation and energy minimisation (See Section 2.2). Ten docking programmes were chosen in view of their ease of use and prominence: UCSF DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, rDock, Smina, Autodock Vina and VinaXB.

Finally, the docking scores were ready to be processed with an appropriate consensus score to choose potential ligands for repurposing. The best setting of the novel consensus score was applied to raw docking scores to obtain one single combined score. A cutoff is chosen at 0.5% of the best-ranked compounds to subset a list of potential activities against MRSA targets.

Chapter 5

Results and Discussions

1 Ligand and Protein Selection

1.1 Sequence Alignment of MRSA Essential Genes

The library of MRSA proteins was built with the starting point of the essential genes from *S. aureus*, obtained from the Database of Essential Genes using BLAST+, a standalone version of BLAST. For each sequence, BLAST+ returned either a list of "hits" (proteins) with significant similarity to the query sequence or an empty list with "No hits". The matching protein was a singular protein chain rather than a full protein. For example, a BLAST+ execution gives the result of chains A and B of the same protein, if they share a similar residue sequence. If the protein contained chains with different compositions, BLAST+ returned the only matching protein chain or chains. In the BLAST+ output, the matching proteins were listed in descending order of measure "Score" which calculates the number of pairwise matchings between the query sequence and the protein sequence and "expect value" (E-value) that measures how many matches would have been returned at a given score by chance (Camacho et al., 2009). Another measure also calculated was the length of the matching protein chain. Those hits with the same length and "Score" were clustered into the same entry. This means these protein chains shared a similar composition and can be referred to interchangeably. The full version of BLAST+ results is available in Appendix A.2.

For docking purposes, the protein chains with the approximately same length and with highest "Score" and E-value were inspected and chosen based on the availability of an adequate co-crystallised ligand and the resolution of the structure. To help with a better preparation for subsequent docking, the hits with a co-crystallised ligand was more favourable to the hit with a better resolution. For instance, protein A which has a ligand and a resolution of 2.5 Å is more favourable than protein B with a resolution of 1.6Å but without ligand. Finally, 78 MRSA genes with PDB code and chain were retained as target proteins for corresponding gene sequences. These protein structures were then prepared for molecular docking (See Table 5.1).

Table 5.1: MRSA target hits

Protein target hits from sequence alignment of MRSA essential genes against PDB. The column from left to right: ID of gene sequence from the Database of Essential Genes; length of the input sequence; length of the matching protein in PDB, Score represents the number of matching pairs; Identity is the extent to which two (nucleotide or amino acid) sequences have the same residues at the same positions in an alignment, expressed as a percentage; E-value, PDB code of matching protein. The letter after the PDB code represents the chain of the protein.

DEG ID	Query Length	Protein length	Score	Identity	E-value	PDB code
DEG10170006	428	449	890	100	0	6R1N.A
DEG10170023	205	205	417	100	4E-151	4HLC.A
DEG10170029	190	198	390	100	8E-141	4YLY.A
DEG10170032	267	291	532	95	0	6CLV.A
DEG10170033	121	121	246	99	3E-86	2NM3.A
DEG10170034	158	161	322	99	6E-115	5ETR.A

Continuation of Table 5.1

DEG ID	Query Length	Protein length	Score	Identity	E-value	PDB code
DEG10170048	693	693	1434	100	0	2XEX.A
DEG10170051	328	331	640	98	0	4E4R.A
DEG10170053	306	308	621	99	0	2X7I.A
DEG10170054	327	331	682	100	0	2HK2.A
DEG10170057	124	127	246	100	7E-86	2FRH.A
DEG10170062	132	132	268	100	2E-94	2B7L.A
DEG10170067	307	326	625	100	0	1HSK.A
DEG10170073	311	312	610	98	0	4GCM.A
DEG10170075	195	203	400	100	2E-144	5VZ2.A
DEG10170077	336	338	689	100	0	3LVF.O
DEG10170078	396	403	783	100	0	4DG5.A
DEG10170079	253	254	518	99	0	3M9Y.A
DEG10170080	505	513	1037	100	0	4QAX.A
DEG10170081	434	442	880	100	0	5BOE.A
DEG10170094	124	127	256	100	9E-90	4M20.A
DEG10170096	443	446	885	98	0	3FF1.B
DEG10170098	313	313	632	99	0	1ZOW.A
DEG10170099	414	437	847	100	0	2GQD.A
DEG10170104	256	282	514	99	0	4D44.A
DEG10170105	493	501	1014	99	0	4C12.A
DEG10170108	397	397	800	99	0	5ZH8.A
DEG10170109	88	88	175	100	8E-59	1KA5.A
DEG10170112	183	194	369	99	1E-132	1LM4.A
DEG10170116	160	160	330	100	5E-118	4NAT.A
DEG10170122	104	106	207	99	4E-71	3DIE.A
DEG10170123	266	286	545	99	0	2JFQ.A
DEG10170130	470	484	932	99	0	3WQT.A
DEG10170131	390	396	776	100	0	4DXD.A
DEG10170133	917	917	1896	99	0	1QU2.A
DEG10170134	207	210	423	100	6E-155	4QRH.A
DEG10170143	308	316	621	99	0	3IM9.A
DEG10170144	244	252	495	100	2E-180	3SJ7.A
DEG10170145	77	101	149	100	2E-48	4DXE.H
DEG10170150	245	269	506	99	0	3KY7.A
DEG10170152	294	301	596	100	0	6G15.A
DEG10170159	256	256	526	99	0	4H8E.A
DEG10170161	567	567	1161	99	0	5ZNJ.A
DEG10170180	61	63	125	100	5E-40	2X4K.A
DEG10170181	420	426	863	100	0	1LRZ.A
DEG10170184	159	157	321	100	1E-114	6PBO.X
DEG10170185	318	321	665	99	0	4DQ1.A
DEG10170190	323	330	657	99	0	6NDL.A
DEG10170193	90	98	180	100	1E-60	4QJU.A
DEG10170195	219	219	442	100	2E-160	2H92.A
DEG10170202	451	451	929	100	0	2VPQ.A
DEG10170204	185	185	379	100	3E-138	6RK3.A
DEG10170217	189	189	392	100	1E-141	2H29.A
DEG10170228	420	420	866	99	0	1QE0.A
DEG10170237	106	106	215	100	3E-74	4PEO.A
DEG10170248	645	645	1337	99	0	1NYR.A
DEG10170252	585	606	1170	99	0	3T0T.A
DEG10170253	307	330	622	100	0	5XZ7.A
DEG10170254	314	327	640	100	0	5KDR.A
DEG10170255	285	285	593	100	0	5KDR.B
DEG10170260	420	420	866	100	0	1JIL.A
DEG10170270	252	252	517	100	0	1QXY.A
DEG10170271	243	243	494	100	0	5N9M.A
DEG10170272	437	437	909	99	0	6H5E.B
DEG10170273	315	337	613	97	0	2QV7.A
DEG10170274	475	483	978	100	0	3IP4.B
DEG10170275	485	485	982	99	0	3IP4.A
DEG10170276	100	100	202	100	3E-69	3IP4.C
DEG10170281	309	317	624	99	0	4RPA.A
DEG10170290	119	143	246	100	9E-86	4DXE.A
DEG10170292	356	360	735	100	0	2I80.A

Continuation of Table 5.1

DEG ID	Query Length	Protein length	Score	Identity	E-value	PDB code
DEG10170299	286	292	581	100	0	4TO8.A
DEG10170301	267	273	546	100	0	5JIC.A
DEG10170303	451	455	917	100	0	6GYZ.A
DEG10170312	72	72	149	100	5E-49	2N8N.A
DEG10170343	130	130	266	100	2E-93	5X1X.A
DEG10170346	388	388	799	99	0	1XPK.A
DEG10170350	117	119	229	98	2E-81	6D1R.A

Certain hits from BLAST+ results were chains of ribosomal proteins, mainly the component chains from the 30S and 50S ribosomes. Since the docking of small ligands to multiple ribosomal chains has been a challenge for many docking programmes, in this study, these hits were retained for future research. The list of hit ribosomal proteins is available in Appendix 3. The gene sequences without any hit in the PDB were discarded. Those with moderate identity and coverage were inputted for homology modelling.

1.2 Homology Modelling of MRSA Essential Genes

MRSA genes with moderate values of coverage and identity were inputted for SWISS-MODEL (Waterhouse et al., 2018). SWISS used both BLAST (Altschul et al., 1990) and HHblits (Remmert et al., 2012) to search for the templates with the highest similarity to the query sequences. Although some authors suggested that with good coverage, an identity of more than 30% can be sufficient for homology modelling (Xiang, 2006), the rate of false-negative was reported to significantly increase in the "twilight zone" (identity from 20% to 35%) (Rost, 1999). To avoid the decreasing accuracy at such an edge, the threshold of identity was set at 40%. In addition, since the coverage of MRSA genes varied, the cutoff was set at 90%. Another measure, the Global Model Quality Estimate (GMQE), which combines features from the target-template alignment and the template structure (Biasini et al., 2014), was set at 75%. The list of 72 templates that met these cutoffs is listed in Table 5.2. There was one noticeable result from DEG10170188 where the identity was 100% and the coverage was 92%. However, the corresponding results from BLAST+ in sequence alignment were 98% and 92%, respectively. Although the identity and the coverage were quite high, the GMQE from DEG10170188 was just 0.79. Therefore, it was considered not a hit but inputted for modelling.

Table 5.2: Results from template searching in SWISS-MODEL
Results from template searching in SWISS-MODEL. The column from left to right: code name of gene sequence from the Database of Essential Genes; identity; coverage of the matching protein in PDB, GMQE of the sequence alignment method; QMEAN of the sequence alignment method; PDB code of matching protein used as the template. The letter after the PDB code represents the chain of the protein.

DEG ID	Identity	Coverage	GMQE	QMEAN	Template
DEG10170002	55.2	0.99	0.8	-1.25	4TR6.A
DEG10170006	52.25	0.99	0.79	-1.74	2DQ3.A
DEG10170012	62.71	0.99	0.77	-1.57	4JIS.A
DEG10170015	56.92	0.99	0.8	0.41	4DD5.A
DEG10170020	58.82	0.99	0.78	-0.48	4E1L.C
DEG10170022	53.85	0.99	0.76	1.93	1YBX.A
DEG10170026	49.11	1	0.77	-0.49	1G97.A
DEG10170027	77.53	0.98	0.8	-0.34	1DKU.E
DEG10170035	72.56	0.99	0.81	-0.88	3A74.A
DEG10170039	49.15	0.98	0.75	-1.84	5EUL.C
DEG10170040	63.04	0.99	0.75	-2.91	4V9H.T
DEG10170041	52.19	0.99	0.79	0.42	3QOY.A
DEG10170043	59.17	0.98	0.81	0.15	1DD4.A
DEG10170045	56.67	0.96	0.75	-2.41	5TW1.E
DEG10170049	73.79	1	0.9	0.22	2C78.A
DEG10170052	51.39	0.99	0.78	-1.93	2P5I.A

Continuation of Table 5.2

DEG ID	Identity	Coverage	GMQE	QMEAN	Template
DEG10170070	61.39	0.99	0.77	-2.33	1TF2.A
DEG10170071	48.94	0.98	0.77	-1.15	1GQE.A
DEG10170072	86.27	0.99	0.86	0.23	1KO7.A
DEG10170074	79.46	1	0.9	0.41	2PPV.A
DEG10170084	63.01	0.97	0.78	-0.53	2D2E.A
DEG10170086	60.84	0.98	0.84	0.01	5J8Q.A
DEG10170092	62.34	0.99	0.79	-0.39	4BPF.A
DEG10170101	62.5	1	0.81	-1.22	1I6K.A
DEG10170102	79.39	1	0.81	0.25	1Z3E.A
DEG10170103	57.2	0.98	0.82	0.18	4DY6.A
DEG10170113	75.21	0.98	0.84	-0.12	1W85.A
DEG10170121	77.63	1	0.87	-0.5	2RHS.B
DEG10170128	46.85	0.99	0.79	-0.92	3LK7.A
DEG10170139	44.29	0.98	0.78	-0.72	5UMF.A
DEG10170142	53.61	0.97	0.79	0	1U7N.A
DEG10170153	48.31	0.99	0.79	-0.81	4XX0.B
DEG10170156	45.91	0.96	0.76	-1.53	3AVX.A
DEG10170158	48.07	0.98	0.81	-0.1	1IS1.A
DEG10170161	56.21	0.99	0.76	-1.37	2J3M.A
DEG10170165	50	0.91	0.75	-0.87	3V7Q.A
DEG10170169	44.34	1	0.79	-0.35	3ZQ4.B
DEG10170172	76.24	0.99	0.87	-0.9	4LNI.H
DEG10170174	49.84	0.98	0.79	-1.99	5ND6.A
DEG10170188	100	0.92	0.79	-1.09	2OLV.A
DEG10170194	73.39	1	0.77	-1.07	4DCS.A
DEG10170197	74.32	1	0.92	0.62	1WTF.A
DEG10170200	43.36	0.98	0.77	-1.77	1H9A.A
DEG10170201	66.88	0.99	0.81	-0.82	2ZYA.B
DEG10170204	54.35	0.99	0.79	0.13	1YBY.A
DEG10170206	40.63	1	0.75	-1.52	3GNL.A
DEG10170210	46.21	0.97	0.77	-1.44	3IEV.A
DEG10170214	59.15	0.93	0.79	0.87	4B9Q.A
DEG10170221	59.12	0.96	0.79	-2.09	1VHX.A
DEG10170222	54.12	0.99	0.81	-2.11	5US5.A
DEG10170224	67.6	0.97	0.79	0.04	2HMA.A
DEG10170225	87	1	0.95	1.13	2HMA.A
DEG10170226	43.88	0.99	0.76	-1.69	1P3W.A
DEG10170227	52.01	0.97	0.78	-1.52	1L0W.A
DEG10170233	58.59	0.98	0.79	-0.41	6BLB.A
DEG10170242	59.07	0.98	0.8	-0.52	1SUL.A
DEG10170253	52.15	0.99	0.8	-1.41	1ZXX.A
DEG10170266	64.47	0.96	0.81	-0.34	5T8S.A
DEG10170279	60.38	0.97	0.78	-0.25	3HMQ.A
DEG10170280	59.14	1	0.79	-0.91	2F7F.A
DEG10170284	58.85	1	0.8	0.01	4V4O.M
DEG10170285	54.26	1	0.78	-1.26	4V4O.O
DEG10170287	42.17	0.97	0.76	-1.91	3ZET.B
DEG10170294	45.07	0.97	0.78	-1.51	3AZ9.A
DEG10170295	63.16	0.99	0.82	-1.07	3R38.A
DEG10170297	59.04	0.99	0.75	-1.64	1ZBT.A
DEG10170300	54.41	0.99	0.79	-0.84	4ZDK.A
DEG10170303	68.16	0.99	0.82	-0.99	3PDK.A
DEG10170304	58.85	1	0.8	0.01	4V4O.M
DEG10170313	73.95	1	0.88	0.25	5G40.A
DEG10170339	45.98	0.98	0.77	-1.38	1LK7.A
DEG10170348	53.8	0.97	0.76	-1.85	2ZXI.A
DEG10170349	42.34	0.97	0.75	-2.19	1XZQ.A

SWISS-MODEL used the templates listed in Table 5.2 to create the structures of corresponding MRSA genes. The remaining templates with insufficient coverage, identity or GMQE values were discarded. The same reason for those sequences without appropriate templates was due to the lack of experimentally determined structures deposited in the protein database.

Certain templates from SWISS-MODEL results were chains of ribosomal proteins, sim-

ilar to hits from BLAST+ results. These hits were also retained for future research. The list of hit ribosomal proteins was available in Appendix 3.

2 Results and Discussions of Benchmark using Median Rank as Evaluation Metric

There were 102 targets from DUD-E and 78 MRSA proteins structurally cross-compared using the Dali server (Holm and Rosenström, 2010). The results show that there were 6 clusters in which 3 clusters contained more than one DUD-E or MRSA protein sharing a similar structure. To fully estimate all possibilities, all possible matching targets were paired and docking was run of DUD-E ligands against corresponding MRSA protein. For instance, DHI1 and INHA (DUD-D proteins) shared structural similarities with 3OSU and 4D44 (MRSA proteins). Therefore, decoys and actives of DHI1 and INHA targets were interchangeably docked against 3OSU and 4D44 proteins. As a result, there were 29 sets of [protein:ligand group pairs] obtained (Table 5.3).

Table 5.3: List of structurally similar targets of DUD-E targets and MRSA targets. For each column, the targets from DUD-E and MRSA targets shared a similar protein structure, which means two DUD-E targets or two MRSA targets in the same column also had similar structures. Those targets of the same column were cross-paired for docking of DUD-E ligands again respective MRSA targets.

DUD-E targets	DEF	DYR	ADA, ALDR GLCM, PYRD	DHI1, INHA	HXK4	TYSY
MRSA target	1LM4	2W9H	3M9Y, 3T05 4HB7, 4TO8, 5BOE	3OSU, 4D44	3WQT, 5JIC	4DQ1

The set of decoys and actives were obtained from the DUD-E repository. For each target, 999 decoys and one active ligand were randomly chosen. The total amount of 1000 ligands was then docked against each corresponding target using ten docking programmes (ADFR, DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, Autodock Vina and VinaXB), producing 10 matrices of 1000 x 29. For consensus scores, docking results of each ligand:target pair were combined using Eqns. 4.16a-4.16d. While analysing a new set of combined scores, for each target, all combined scores were picked in descending order, starting with the best binding energy score and then progressing towards the worst (that is negative infinity to positive infinity). The medians of these re-positioned values were then used to calculate the histogram leading to the probability distribution function.

2.1 Statistical Ranking of Docking Scores (DUD-E Database)

In this study, the median ranking order was used for evaluation. First, active ligands for 29 targets were ranked among 1000 ligand (docked) arrays. The anticipated median rank of a subset of randomly chosen actives out of 1000 ligands is 500. The ability of consensus scores to improve the ranks of the actives when the ranks of the active scores varied from the top to the bottom was tested. The average rank of a set of actives for each target would result in an average value (Kairys et al., 2006; Truchon and Bayly, 2007). Therefore, only one active was randomly chosen for each target. For each docking programme, the docking score of activities for each target was ranked among 1000 ligands, resulting in a set of 29 ranks of active ligands. The median ranks obtained from 10 docking programmes verified that median ranks of active ligands (e.g. 250 from ADFR) were better than a median from a random selection. For simplicity, each set of ranks (for 10 programmes) was represented by a single median rank, as detailed in Table 5.4.

Smina returned the best median rank of 150, followed by PLANTS with the median rank of 163 and 185 in QuickVina2 while Autodock Vina and Gemdock showed comparative median ranks of 191 and 192. Surprisingly, the highly popular DOCK turned up the worst program (median rank of 423). This can be explained that DOCK was not particularly sensitive to MRSA-related targets. On the other hand, Autodock Vina and its derivatives showed promising results. Based on this evaluation, Smina was the single best performing docking station for the DUD-E set of ligands. In another word, if only a single docking

programme	Median rank
ADFR	337
DOCK	423
Gemdock	192
Ledock	387
PLANTS	163
PSOVina	375
QuickVina2	185
Smina	150
Autodock Vina	191
VinaXB	224

Table 5.4: The median rank of actives across 29 targets using 10 docking programmes. The active ligand for each target was ranked together with other 999 decoys. The median rank takes the median value of 29 ranks of the actives across 29 targets for every single programme.

programme is designated for the task of virtual screening against MRSA-related targets, Smina is a suitable choice. The median rank can be expressed as the recovery rate or the threshold for the retrieval of 50% of the actives. Recovery rate is defined as the ratio of the actives that can be recovered when screening a certain fraction of the whole dataset in ranked order. Therefore, the recovery rate of ADFR, DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, Smina, Autodock Vina and VinaXB were 33.7%, 42.3%, 19.2%, 38.7%, 16.3%, 37.5%, 18.5%, 15%, 19.2% and 22.4%, respectively. This means, should Smina be chosen for virtual screening, 15% of the best-ranked ligands after docking would contain half of the active ligands. The first plot of Figure 5.1a showed the individual performance of docking programmes across 29 sets of target:ligands.

In an attempt to improve the ranking of actives over the full dataset, the docking scores from 10 docking programmes were combined in the so-called consensus scores. Here ten traditional consensus scores were computed: MEAN, MEDIAN, MIN, MAX, EUC, CBM, ECR and DSR. One obstacle was that the docking scores from different docking programmes span across various scales. For that reason, the three most common methods including rank normalisation, min-max normalisation and z-score normalisation were used to bring the docking scores to a unified scope. 8 consensus scores were calculated based on normalised docking scores. That made up to 24 possible combined scores. Here the traditional consensus scores that are commonly used in the literature were used to compare the ability to improve the ranking of the actives.

	MAX	MIN	MEAN	MEDIAN	EUC	CBM	ECR	DSR
min-max	228	246.5	184	202.5	206	201	217	224
rank	191	195	271	205.5	176	174	207.5	183
z-score	203	209	256	231	1000	220	192	205

Table 5.5: Median rank of traditional consensus scores over normalisation methods. The median ranks were obtained in the same manner as the median rank from each individual docking programme. Each median rank represented the combination of 10 docking programmes after normalised with respective methods.

As demonstrated in Figure 5.1, these conventional consensus scores showed no noticeable improvement over the individual programmes across three different normalisation methods (Figures 5.1b, 5.1c and 5.1d. In fact, the best median the Cubic Mean score could reach was 174 while all other consensus scores declined, compared to 150 in Smina. This was probably due to the lack of sensitivity of MRSA data to these consensus scores.

2.2 Novel Consensus Scores

For each docking programme, the median ranks of active ligands across 29 targets were obtained and histograms plotted for visual presentation. To establish the improved performance of consensus scores (CS) over individual docking, the scores from the individual best performer Smina were compared against the CS score line. This was estimated from

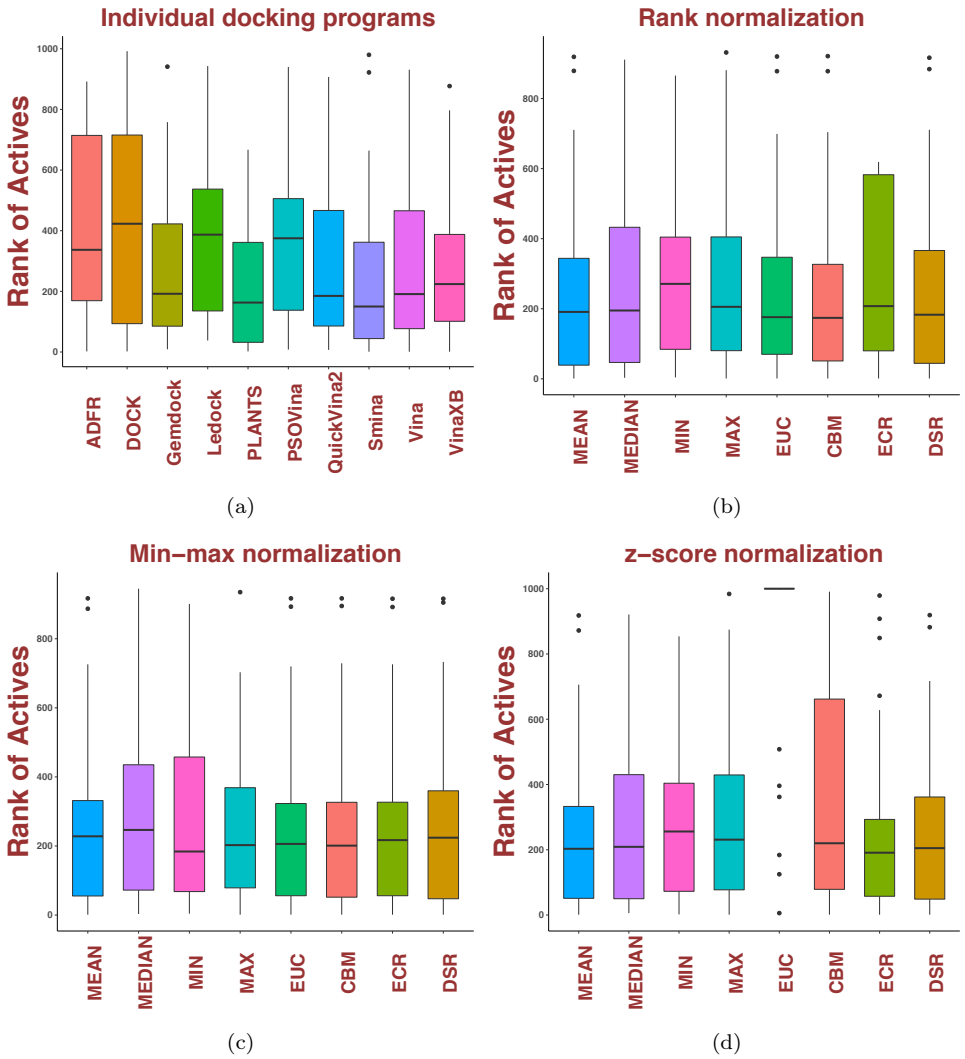
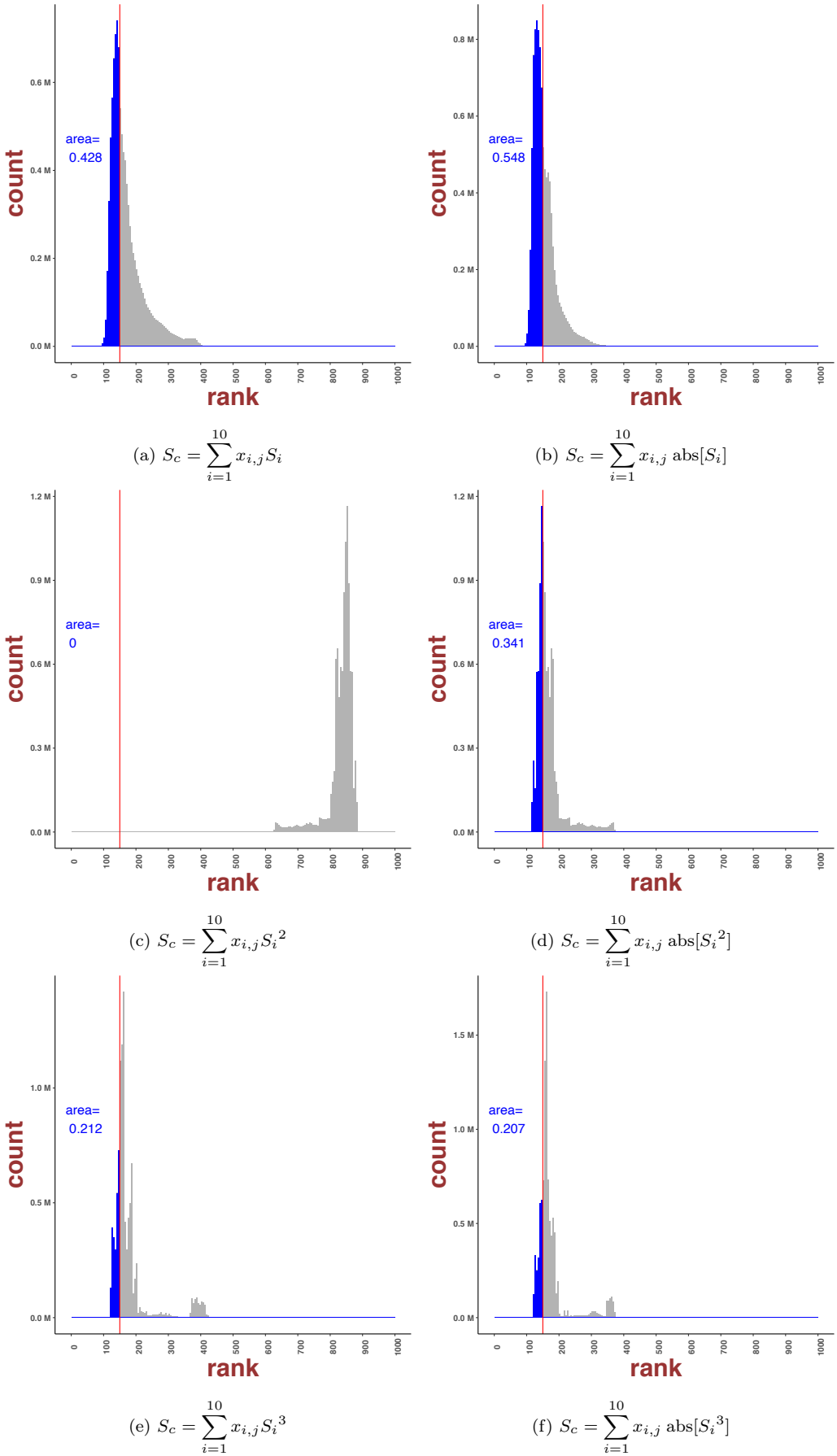


Figure 5.1: Box plots demonstrate the ranks of actives from programmes and consensus scores. Each box plot illustrates the ranks of active ligands across 29 targets using each docking programme or consensus score. The lines parallel to the x-axis in each box represents the median ranks or the quantitative measure for the performance of each docking programme or consensus score. a) Ranks of the actives from individual docking programmes. b) Ranks of the actives from various consensus scores after rank normalisation. c) Ranks of the actives from various consensus scores after min-max normalisation. d) Ranks of the actives from various consensus scores after z-score normalisation.

the area patches to the left (since binding energy is negative) of the best performing individual docking platform (Smina, identified by the solid line close to the maxima of the histograms). The greater the patch area, the better the CS score line (compared to Smina).

As clearly demonstrated in Figures 5.2 and 5.6, the linear consensus model consistently turned the best performer, with CS docking score progressively declining with increasing values of n, where n is the exponent in the statistical norm. It was noted that three out of four linear combinations ($n = 1$) demonstrated higher ranks compared to the individual best performer Smina (82, 83 and 82 for model 4.16a, 4.16b and 4.16c, respectively). Another suggestive trend was the dominance of the odd n values against their even counterpart. This was expected as the docking scores were energy affinity measures, hence negative, that could be compensated by the absolute (consensus) values (as in models in Eqn. 4.16b and Eqn. 4.16d). Model 4.16d was the worst scorer, while linear combinations of models 4.16a, 4.16b and 4.16c showed similar behaviour with approximate best ranks and comparable histograms (non-normalised probability density functions).



(a) $S_c = \sum_{i=1}^{10} x_{i,j} S_i$

(b) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[S_i]$

(c) $S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^2$

(d) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[{S_i}^2]$

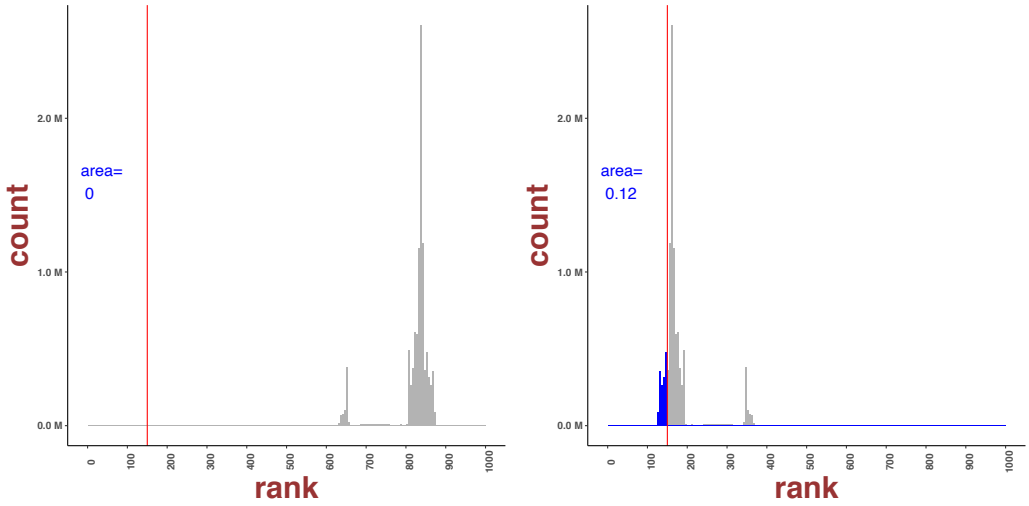
(e) $S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^3$

(f) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[{S_i}^3]$

Figure 5.2: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with the vertical red line) of the total histogram area, evaluated for order ranging from 1 to 3 as in Eqns. 4.16a (left)-4.16b (right).

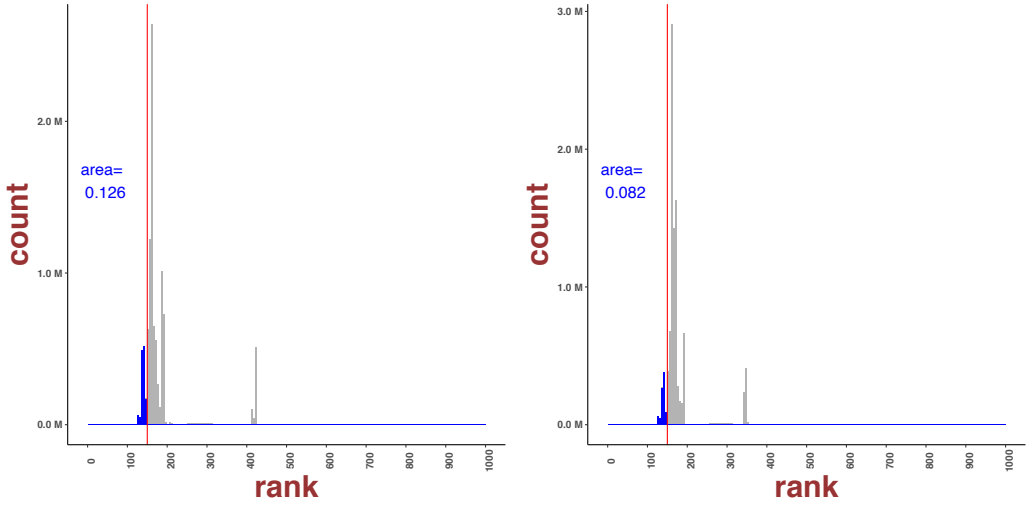
As explained, the solid red lines in these histograms represent the individually best-performing docking standard while the blue patches to the left of these red lines (docking scores represent attractive energy measures which are negative, hence to the left) represent

the fractional betterment in docking scores due to the CS methodology. The grey patches to the right or left of the red lines indicate "no shows", implying that the CS method did not improve the individual best (docking) scores in those regions. The histograms are non-scaled representations of the Probability Density Functions (PDFs), or in other words, Figures 5.2, 5.3, 5.4, etc pictorially demonstrated the improvement in docking standards by the usage of the CS method as opposed to individual best scorers.



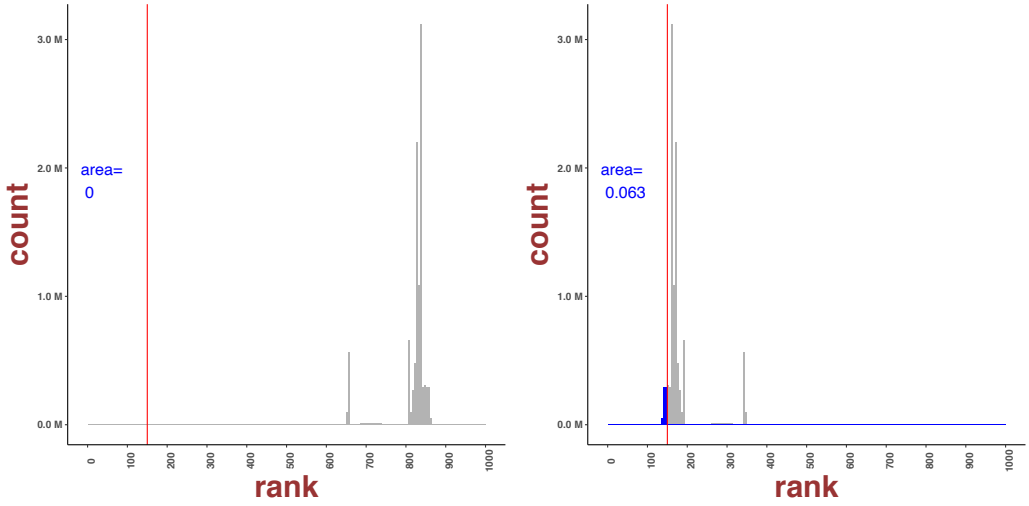
(a) $S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^4$

(b) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[{S_i}^4]$



(c) $S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^5$

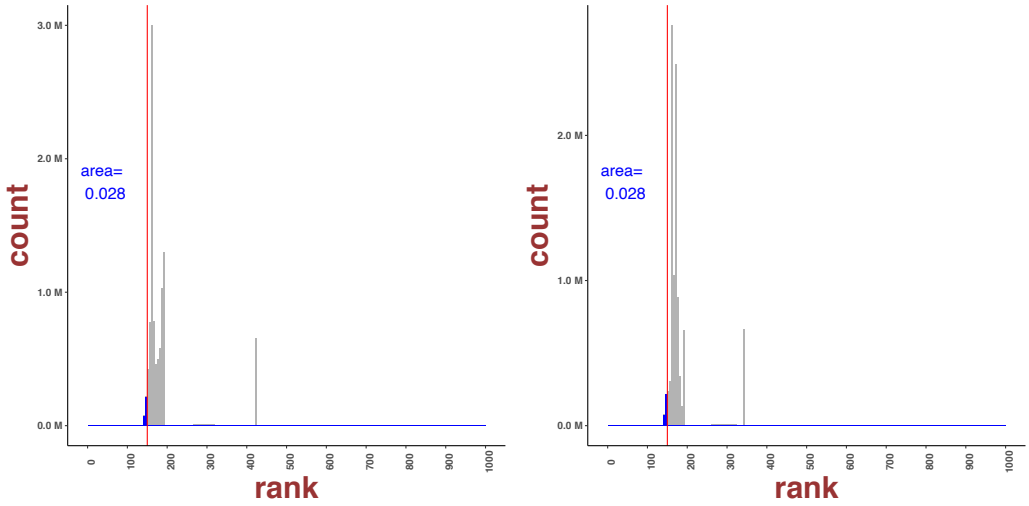
(d) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[{S_i}^5]$



(e) $S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^6$

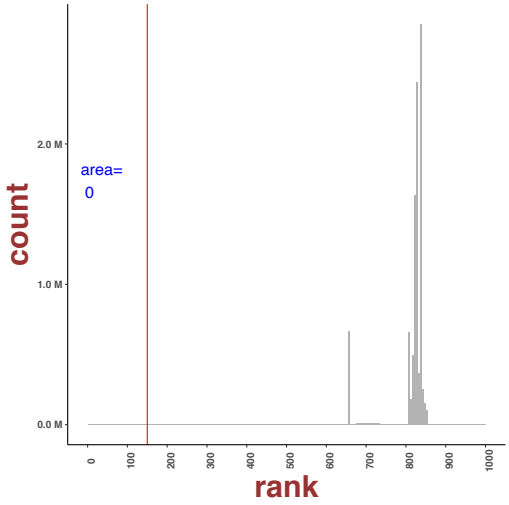
(f) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[{S_i}^6]$

Figure 5.3: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with the vertical red line) of the total histogram area, evaluated for order ranging from 4 to 6 as in Eqns. 4.16a (left)-4.16b (right).

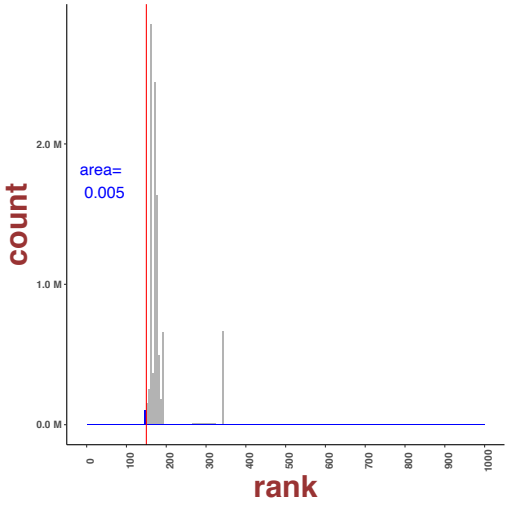


(a) $S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^7$

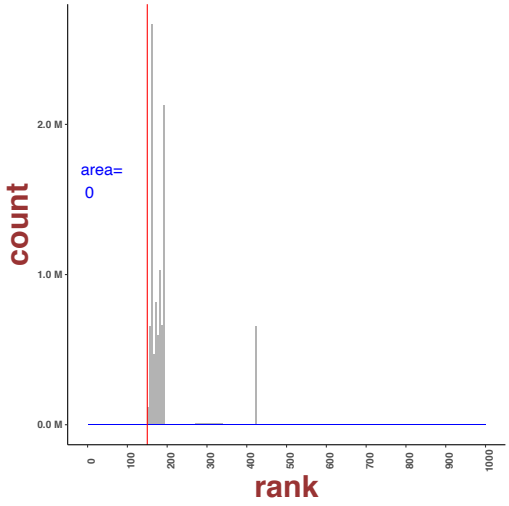
(b) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[{S_i}^7]$



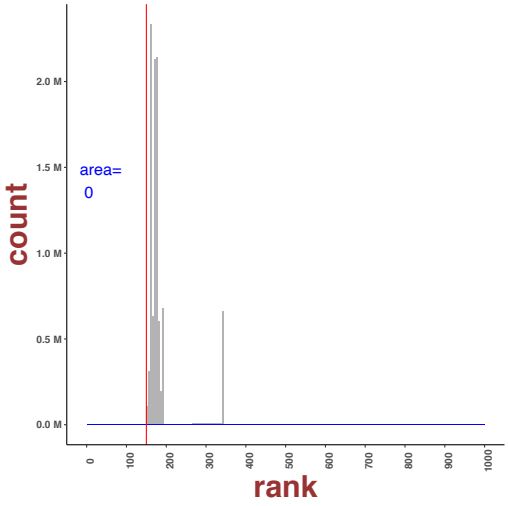
(c) $S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^8$



(d) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[{S_i}^8]$

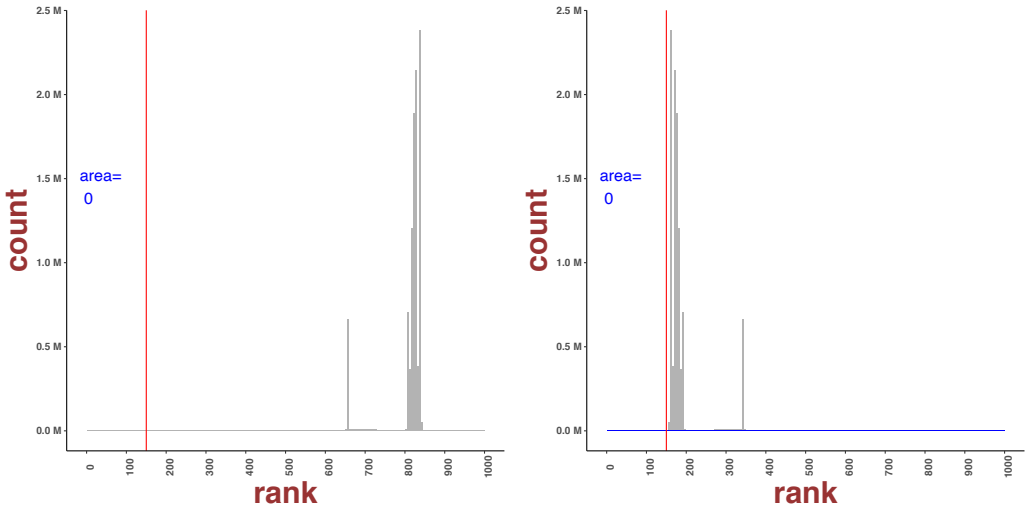


(e) $S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^9$



(f) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[{S_i}^9]$

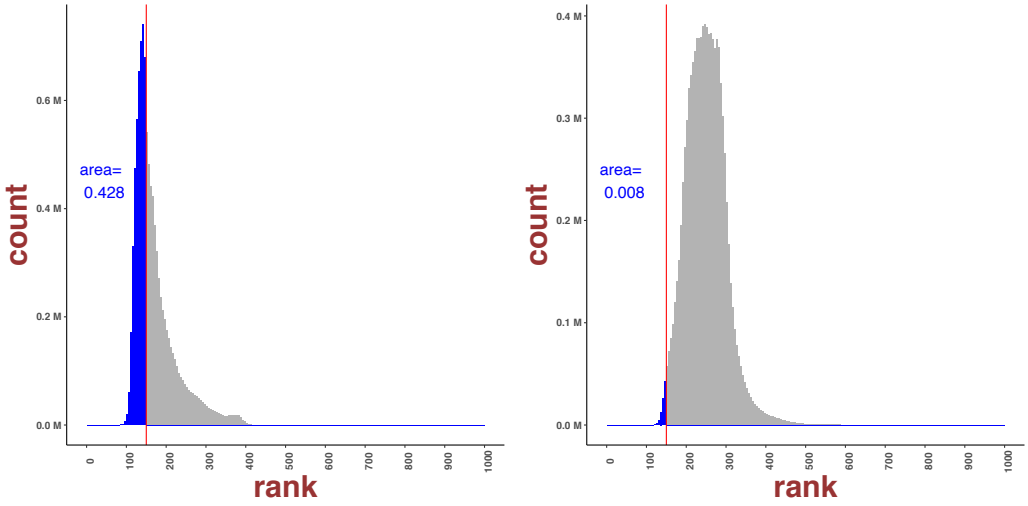
Figure 5.4: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with the vertical red line) of the total histogram area, evaluated for order ranging from 7 to 9 as in Eqns. 4.16a (left)-4.16b (right).



(a) $S_c = \sum_{i=1}^{10} x_{i,j} {S_i}^{10}$ (b) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[{S_i}^{10}]$

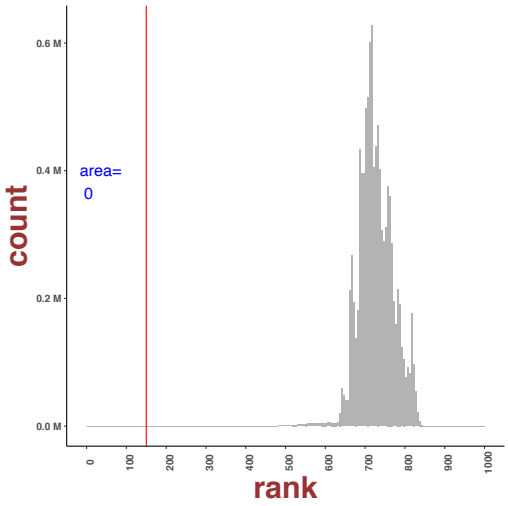
Figure 5.5: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with the vertical red line) of the total histogram area, evaluated for order 10 as in Eqns. 4.16a (left)-4.16b (right).

The lack of any blue patch in Figure 5.5 indicates that the individual docking standard is the best and the same as the CS scoreline.

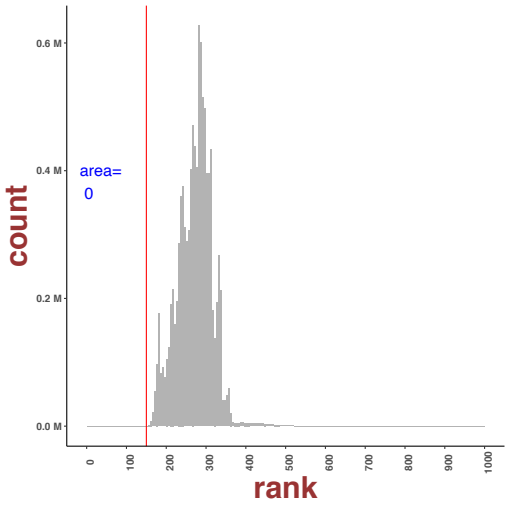


(a) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - \bar{S}_i)$

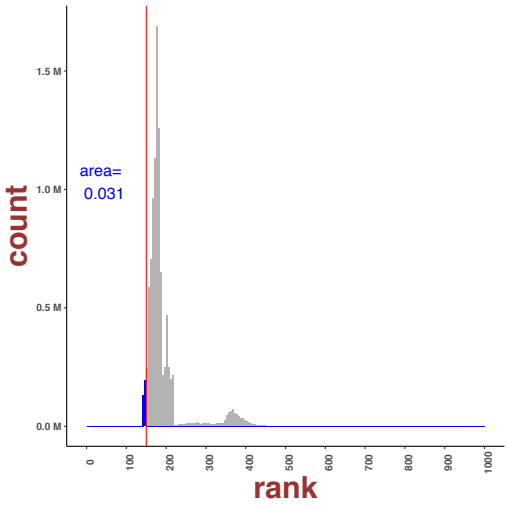
(b) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[S_i - \bar{S}_i]$



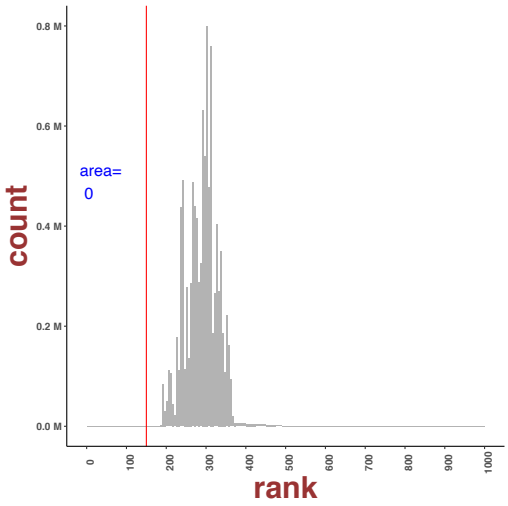
(c) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - \bar{S}_i)^2$



(d) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_i - \bar{S}_i)^2]$

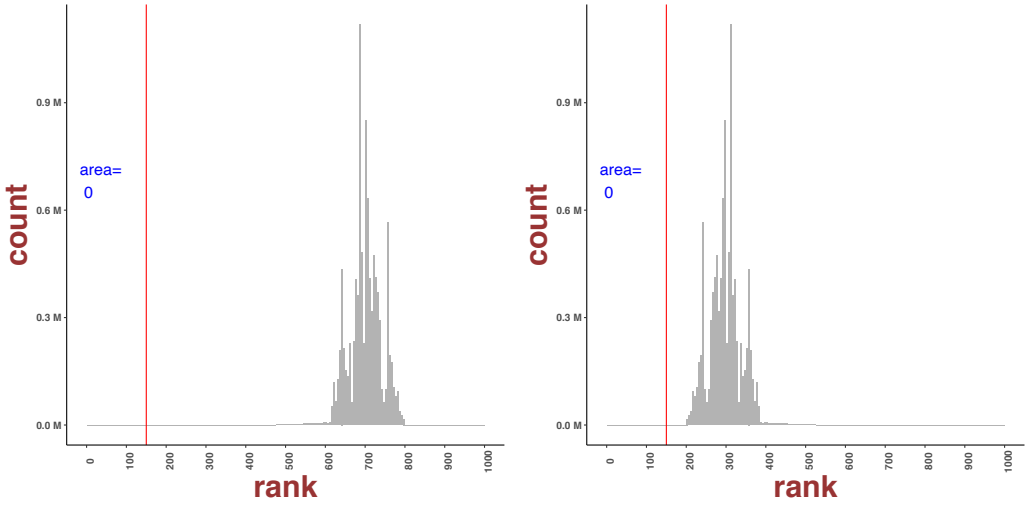


(e) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - \bar{S}_i)^3$



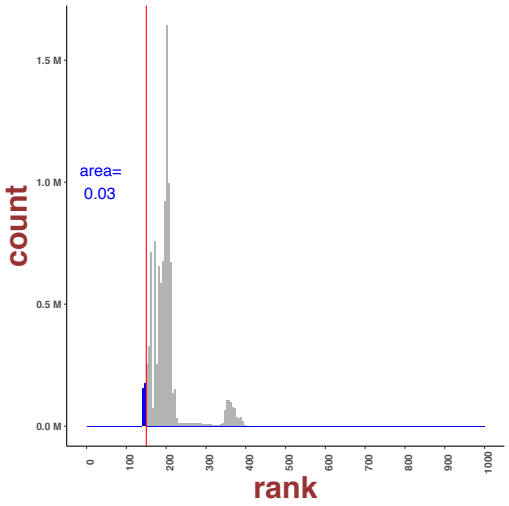
(f) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_i - \bar{S}_i)^3]$

Figure 5.6: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with the vertical line) of the total histogram area, evaluated for order ranging from 1 to 3 as in Eqns. 4.16c (left)-4.16d (right).

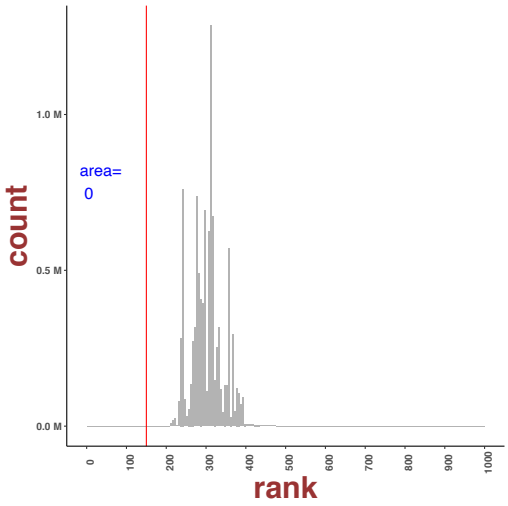


(a) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ (S_i - \bar{S}_i)^4$

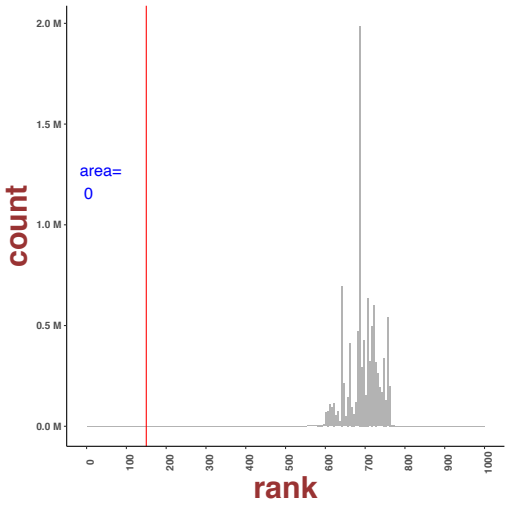
(b) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[(S_i - \bar{S}_i)^4]$



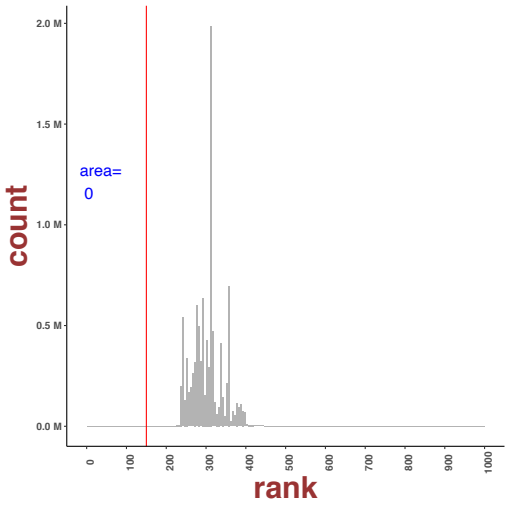
(c) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ (S_i - \bar{S}_i)^5$



(d) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[(S_i - \bar{S}_i)^5]$

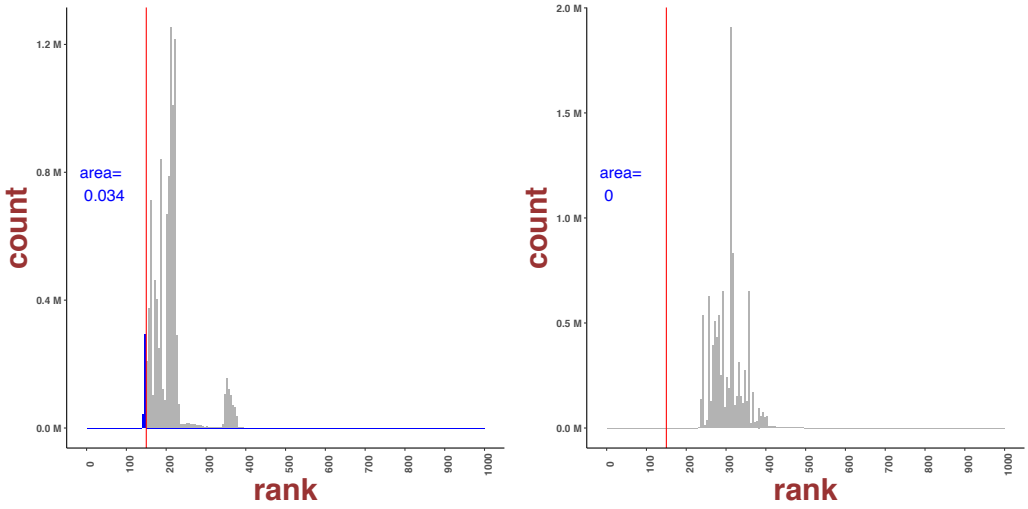


(e) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ (S_i - \bar{S}_i)^6$



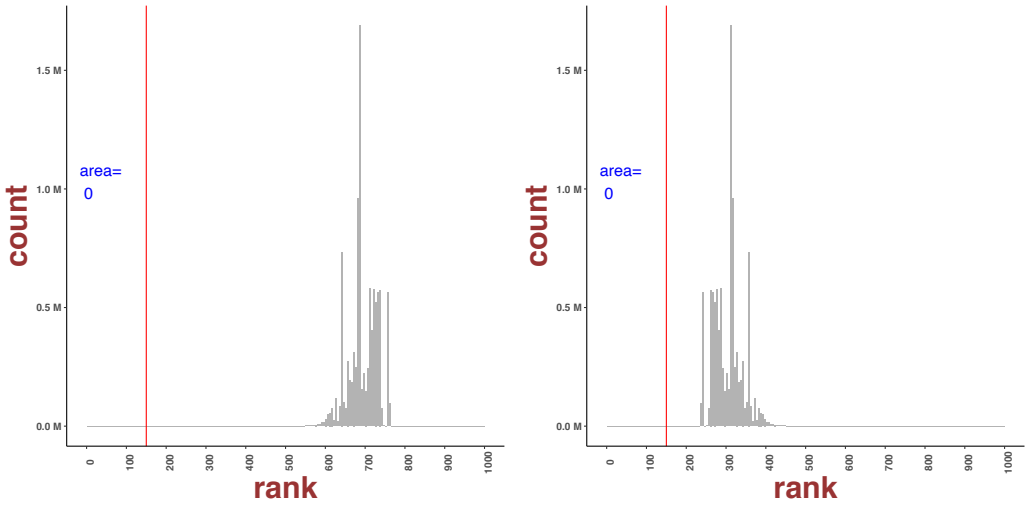
(f) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[(S_i - \bar{S}_i)^6]$

Figure 5.7: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with the vertical line) of the total histogram area, evaluated for order ranging from 4 to 6 as in Eqns. 4.16c (left)-4.16d (right).



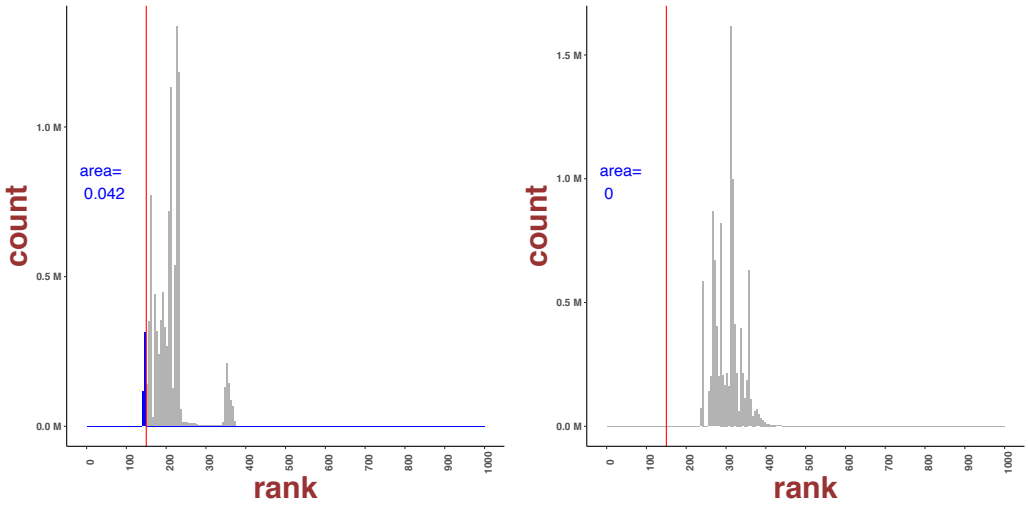
(a) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ (S_i - \bar{S}_i)^7$

(b) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[(S_i - \bar{S}_i)^7]$



(c) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ (S_i - \bar{S}_i)^8$

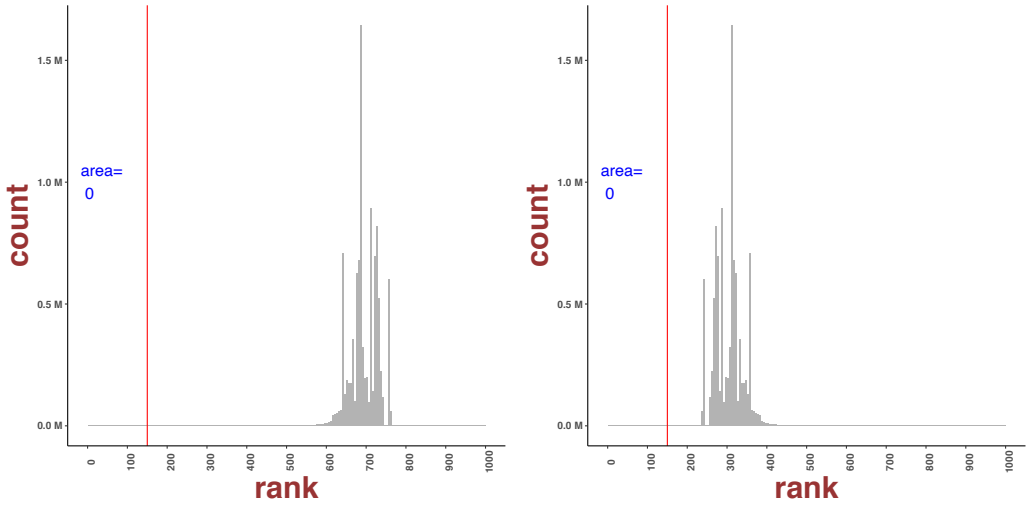
(d) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[(S_i - \bar{S}_i)^8]$



(e) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ (S_i - \bar{S}_i)^9$

(f) $S_c = \sum_{i=1}^{10}\sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[(S_i - \bar{S}_i)^9]$

Figure 5.8: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with the vertical line) of the total histogram area, evaluated for order ranging from 7 to 9 as in Eqns. 4.16c (left)-4.16d (right).



(a) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - \bar{S}_i)^{10}$ (b) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \text{abs}[(S_i - \bar{S}_i)^{10}]$

Figure 5.9: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with the vertical line) of the total histogram area, evaluated for order 10 as in Eqns. 4.16c (left)-4.16d (right).

As evident from Figures 5.6, linear regression over the set of 10 docking scores involving the ligand:protein sets returned a better docking score than equations with higher order. When the order value increased, the median ranks tended to converge a center, for instance, around median rank of 112 in case of equation 1.2. and 179 in case of equation 1.4. This effect was more obvious in cases of equations 1.1. and 1.3 due to the fluctuation. However, when n value increased, the best median rank deteriorated (the left end of the blue patch), ignoring the even values of n, which is not in favor of virtual screening. The same trend was observed for the area ratio.

	Eqn. 1.1		Eqn. 1.2		Eqn. 1.3		Eqn. 1.4	
Power	Best rank	Area ratio	Best rank	Area ratio	Best rank	Area ratio	Best rank	Area ratio
1	82	0.532	83	0.648	82	0.532	119	0.02
2	558	0	109	0.541	395	0	152	0
3	109	0.45	109	0.413	112	0.107	177	0
4	579	0	109	0.289	399	0	174	0
5	110	0.295	110	0.18	118	0.085	177	0
6	572	0	111	0.117	399	0	17	0
7	111	0.137	111	0.078	116	0.086	177	0
8	556	0	112	0.047	399	0	182	0
9	112	0.07	112	0.038	119	0.087	179	0
10	543	0	112	0.005	399	0	179	0

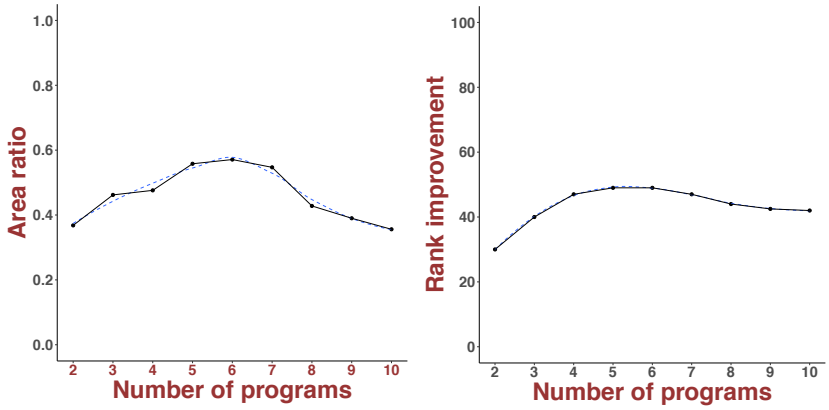
Table 5.6: Table of best ranks and area ratios from the histograms. The rank improvement is the difference between the best rank that each novel consensus score can reach amongst 10015005 combinations, represented by the far left end of the blue shaded portion of the histogram, and the median rank by the best programme, milestoned by the red vertical line. The area ration the area of histogram of median rank that is counted better than the supposedly best docking programmes. The Best rank is the highest rank that 10015005 combinations achieved.

2.3 Consensus Model Accuracy Convergence

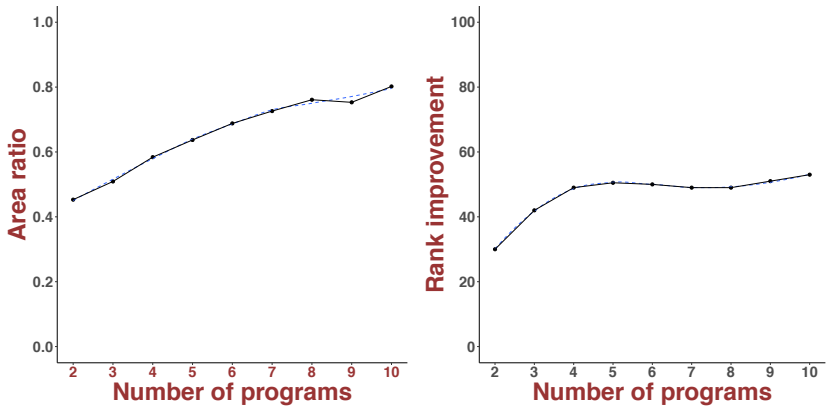
For enumerating the strength of linear combination in each model, the correlation between the number of docking programmes and the consensus performance was estimated. Two types of measures were calculated: area ratio and rank improvement, a relative comparison of which are encapsulated in Table 5.6. An additional measure, rank improvement, was calculated to assess the advancement of consensus scores. Rank improvement is defined as the difference between the best rank each model can achieve and the rank from the best individual programme (Smina). The model in Eqn. 4.16a defines an explicit correlation between the number of docking programmes and consensus outcome. The area ratio considerably increased from 2 to 7 programmes and then saturated after around 8 docking combinations (Figure 5.10b). Similarly, rank improvement drastically increased from 2 to

4 programmes and flattened after 5 programmes (Figure 5.10f). Comparison between these two measures suggested that numerous docking programmes do not necessarily contribute to the overall performance. Models 4.16a and 4.16c showed similar saturation patterns both for area ratio and rank improvement. The consensus effect tends to increase from combinations of 2 programmes and maximise after 5 or 6 programmes (Figures 5.10a, 5.10c, 5.10e, 5.10g). Model 4.16d showed poor improvement in both area ratio and rank, area ratio mostly remaining zero (Figure 5.10d) while rank showed negative changes around n = 8 programmes (Figure 5.10h), indicating no improvement.

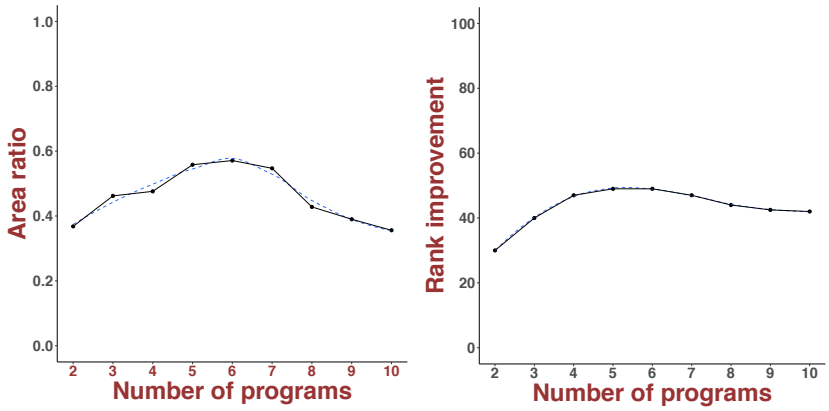
A possible reason for the lack of convergence in Figures 5.10b and 5.10f was the lack of fluctuations due to the consideration of absolute values, causing gradual increments ("accumulation" effect) with increasing number of docking programmes unlike in models 4.16a and 4.16a for which the consensus accuracy converges faster by 4 or 5 programmes.



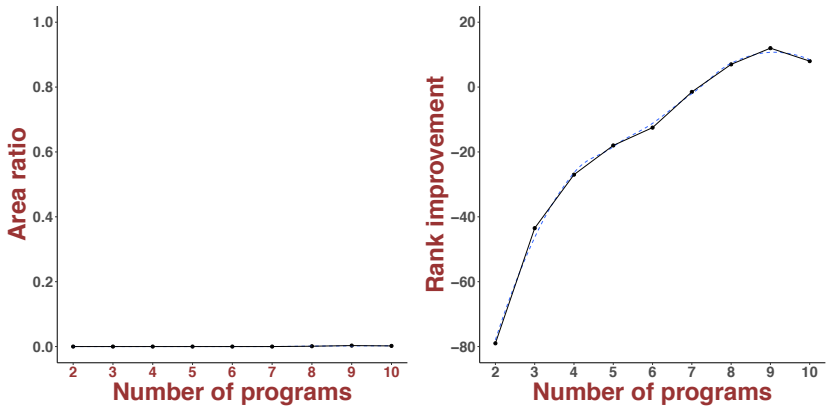
(a) (b)



(c) (d)



(e) (f)



(g) (h)

Figure 5.10: Performance versus the number of docking programmes. The figures on the left represent the area ratio versus the number of programmes and the figures on the right represent the best rank versus the number of programmes. From top to bottom: area ratio and rank improvement of model 4.16a, 4.16b, 4.16c, 4.16d

The ideal way to enumerate the CS coefficients would be through probabilistic modelling of the data for each docking programme, ideally using machine learning or deep learning, an approach that was taken in the later subsection. Here, small incremental changes to the relative weights were used and compared each against the other, retaining only the top scoring ones. The quality of this prediction has been favourably compared against the machine-learning outcome, as shown in the next section.

2.4 Conclusion

Consensus scoring algorithms using MRSA datasets and ten docking programmes (ADFR, DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, Smina, Autodock Vina and VinaXB) were investigated. The performance benchmark was the median rank of active ligands. The individual docking programmes with conventional consensus scores (minimum, maximum, mean, median, reciprocal rank and Euclidean distance) were also compared in this section, including the newly reported Exponential Consensus Rank score.

Before consensus scoring, the distribution of docking scores was altered with three normalisation methods (rank, min-max scaling, and z-scores) to offer a direct combination with commonly used statistical consensus scores. Comparisons show that insensitivity of the MRSA dataset to conventional consensus scores and no improved rank compared to 150 from Smina. Nonetheless, the novel consensus scores consistently perform better than individual docking programmes on the MRSA benchmark dataset. In this work, the raw docking scores from ten docking programmes (ADFR, DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, Smina, Autodock Vina and VinaXB) were directly combined. Due to an exhaustive search of combinations, there was no obligation for data normalisation. Results showed that the novel model gave better rankings of active ligands across benchmark datasets.

A key outcome of the novel consensus module is the preponderance of linear combination of docking scores towards improved active ligand ranking over higher-order consensus formulas. Given that such complex systems are known to be inherently higher-order, such a linear mapping is interesting and potentially more efficient than higher-order scores. As of the higher-order scores, as in Eqns. 4.16a-4.16d, odd ordered combinations show consistently better combinatorics than their even ordered counterparts. These findings also indicate that linear combinations using absolute values (model 4.16b) converge towards a better functional relationship between the number of docking programmes and consensus performance. While consensus prediction accuracy does increase with an increasing number of docking associations, as shown in Figure 5.10, that number is not a monotonically diverging quantity, rather it saturates beyond a certain finite number of docking programs, typically 5-7 for the sets of ligands and MRSA proteins. This is a remarkable feature of the consensus approach. It should allow for the systematic substitution of weaker docking programmes with programmes exhibiting a higher scoring accuracy, as they arise over time since consensus scoring will always outperform even the best performing individual docking programme.

3 Results and Discussions of Benchmark using ROC and EF as Evaluation Metrics

In the previous benchmark using median rank as an evaluation metric, the running in ADFR was particularly time-consuming. Therefore in this section, ADFR was substituted with rDock. The docking evaluation metrics were also changed to Receiver Operating Characteristic (ROC) and enrichment factor (EF). Therefore, more actives were needed for each target. Similar to the previous benchmark, 29 sets of target:ligands were obtained from the DUD-E repository. For each target, 1000 decoys and 40 active ligands were randomly chosen. The total amount of 1040 ligands were then docked against each corresponding target using ten docking programmes (DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, rDock, Smina, Autodock Vina, and VinaXB), producing 10 matrices of 1040

x 29. For consensus scores, docking results of each target:ligand set were combined using Eqns. 4.16a-4.16d and 4.18a-4.18b. AUROCC and EF05 were used to quantitatively evaluate the ability to discriminate actives from decoys. There is a minor difference to the previous benchmark as the mean of AUROCC and EF05 were calculated instead of the median as in some cases, the majority of EF05 values are zero, resulting in zero value of the median.

3.1 Statistical Ranking of Docking Scores (DUD-E Database)

In this study, ROC and EF were used as metrics for the evaluation of docking programmes and consensus scores. For the quantitative purpose, the area under the ROC curves (AUROCC) and Enrichment Factor at 0.5% (EF05) were calculated. First, active ligands for 29 targets were randomly chosen and then ranked across 1040 ligands (including actives and decoys). A uniform distribution of actives and decoys from 1040 ligands would lead to an AUROCC of 0.5. The AUROCC obtained from 10 docking programmes verified that most programmes are able to distinguish between decoys and actives better than a random pick up. (as detailed in Table 5.7). In addition to ROC, EF was used to evaluate the ability of the docking programmes to recognise the actives among the top-ranked ligands. One limitation of EF is that it greatly depends on the number of decoys and actives. Therefore, ROC and EF are exploited simultaneously to assess the performance of docking programmes and consensus scores. The AUROCC curves of 29 target:ligand sets from 10 docking programmes are shown in Figure 5.11 and 5.12.

Smina showed a number of ROC curves skewed towards the right upper corner, which represents the better discrimination between two groups of ligands. Meanwhile, ROC curves in rDock were evenly distributed above and below the diagonal (Figure 5.12). This indicated that actives and decoys were not separated for all targets.

From the mean AUROCC values from ten programmes, Smina showed the best discrimination between decoys and actives, with an AUROCC value of 0.623. Other docking programmes showed approximate AUROCC (DOCK 0.580, Gemdock 0.610, PLANTS 0.597, QuickVina2 0.596, Autodock Vina 0.602 and VinaXB 0.604) (Table 5.7). rDock showed the worst results, really close to 0.5, indicating that the result of rDock is as good as random selection. For confirmation, the AUROCC from ADFR was 0.542, which was a little higher than rDock AUROCC but less than the other nine programmes. Therefore, the substitution of ADFR with rDock was not a detriment. Ledock and PSOVina shared the second-worst value of AUROCC (0.559 and 0.560). Surprisingly, the popular DOCK turned up with an average AUROCC (0.580). In general, Autodock Vina and its derivatives showed promising results. Based on this evaluation, Smina was the single best performing docking station for the DUD-E set of ligands in terms of discrimination between decoys and actives.

	DOCK	Gemdock	Ledock	PLANTS	PSOVina	QuickVina2	rDock	Smina	Autodock Vina	VinaXB
AUROCC	0.580	0.610	0.559	0.597	0.560	0.596	0.492	0.623	0.602	0.604
EF05	2.5	3.0	1.3	5.2	4.3	5.4	0.0	5.5	5.1	5.7

Table 5.7: Area under ROC curve and Enrichment Factor at 0.5% of individual docking programmes. Each value was obtained by taking the mean of AUROCC or EF05 values from 29 targets and across 10 docking programmes.

On the other hand, EF05 showed that PLANTS, QuickVina2, Smina, Vina and VinaXB had a better ratio of active ligands in top-ranked ligands (EF05 approximate 5). VinaXB had the best value of EF05 (5.7) and followed by Smina (EF05 of 5.5). rDock continued to show that it is deficient in retrieving actives in 0.5% of the top-ranked ligands (EF05 0.0). Ledock also remained the second-worst programme (EF05 1.3).

When considering both AUROCC and EF in a combination fashion, PLANTS, QuickVina2, Smina, Autodock Vina and VinaXB had better AUROCC and EF05 while Ledock and PSOVina showed worse values in both. On the other hand, DOCK and Gemdock had good AUROCC (0.580 and 0.610) but modest EF05 (2.5 and 3.0). This indicates when using DOCK and Gemdock, actives can be retrieved but within a larger propor-

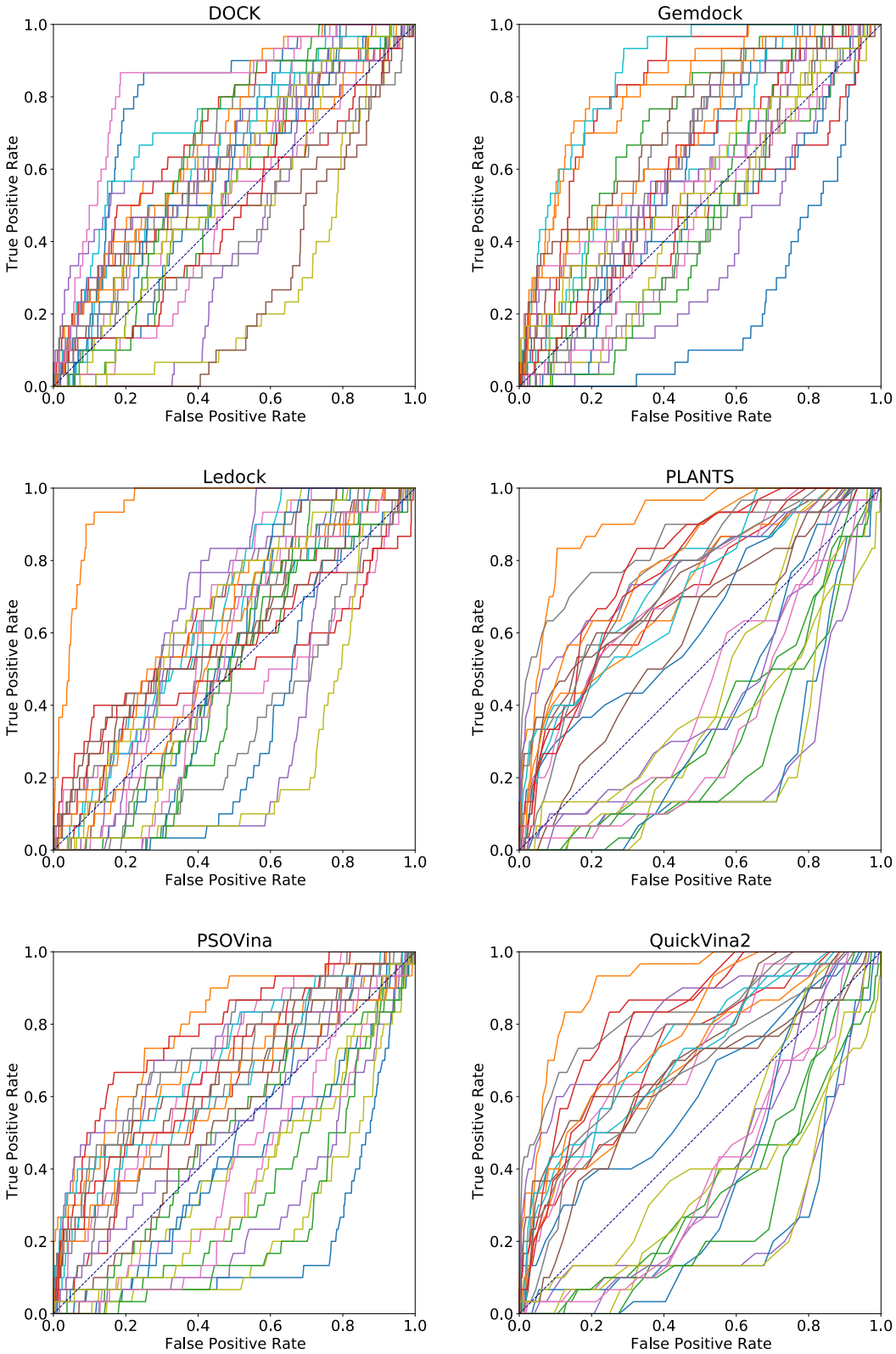


Figure 5.11: ROC curves across 29 targets across from DOCK, Gemdock, Ledock, PLANTS, PSOVina and QuickVina2. Each target is represented by one colour across 6 ROC figures. The diagonal represents a result from evenly distributed actives. The ROC curves skewed to left upper corner indicated a better discrimination between actives and decoys.

tion of top-ranked ligands. Meanwhile, rDock suffered poor results in both AUROCC and EF05. Subsequently, Smina and VinaXB showed consistent superior to other programmes in both AUROCC and EF05. These findings were in line with results from the previous benchmark, where Smina appeared to be the programme with the best performance for the chosen dataset. Similarly, the performance of DOCK was estimated amongst the worst.

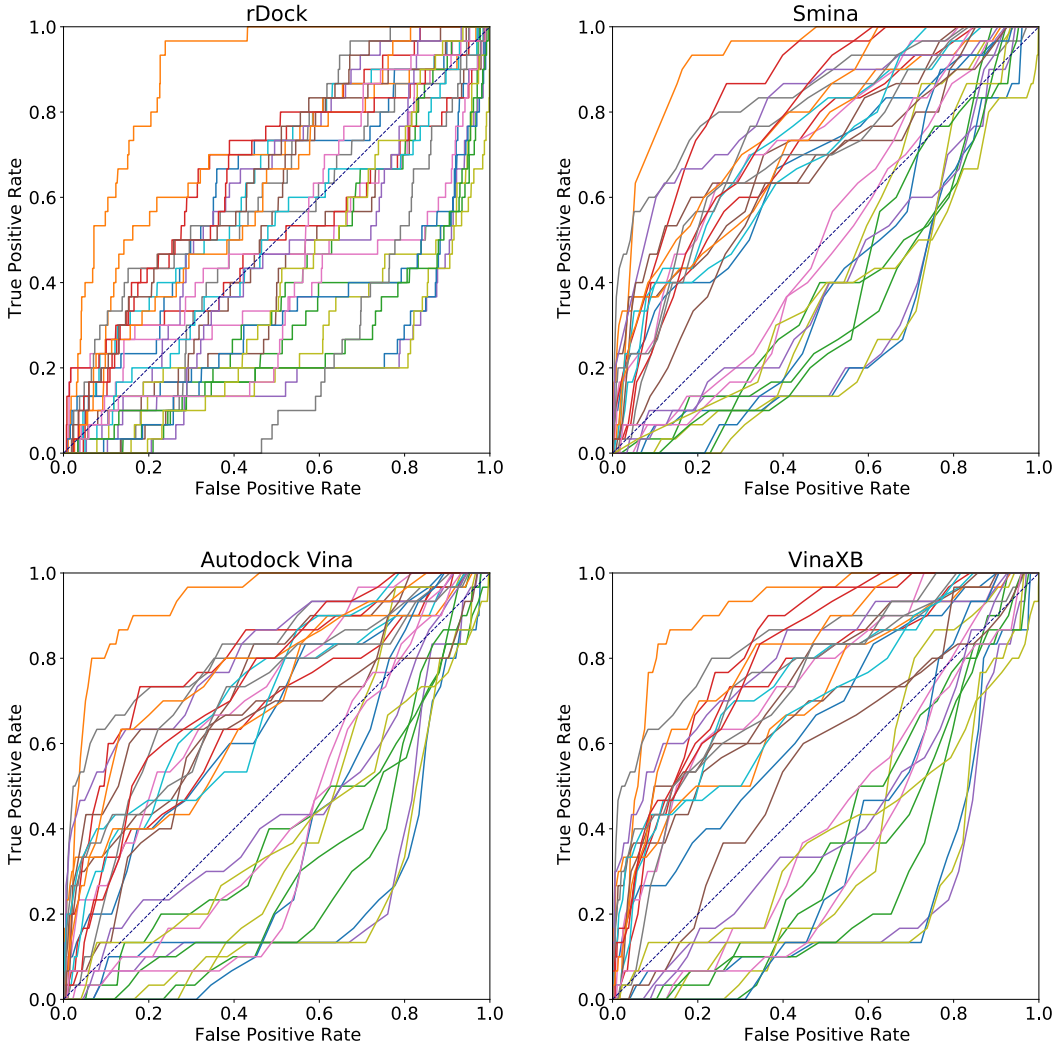
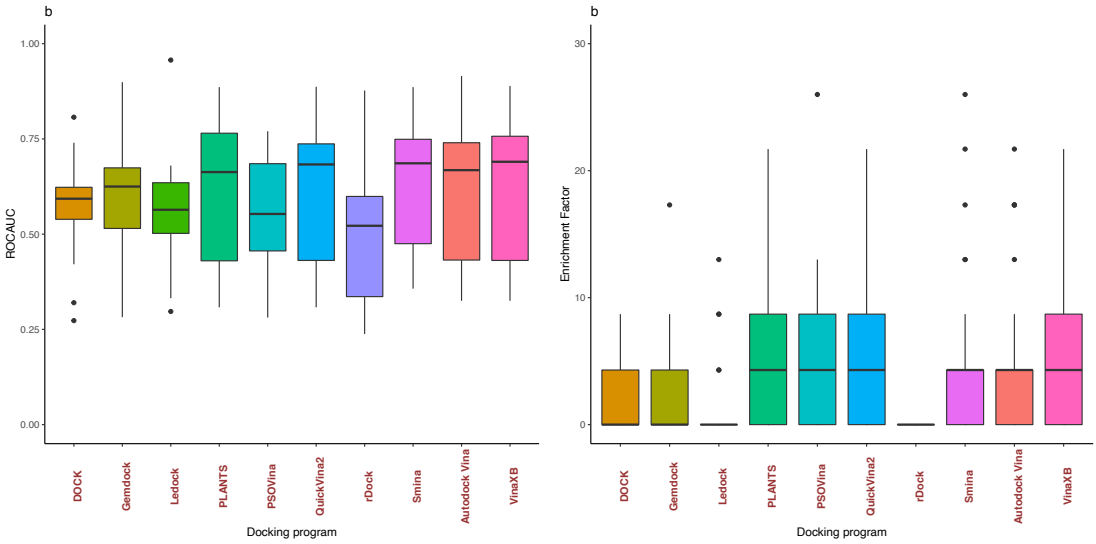


Figure 5.12: ROC curves across 29 targets across from rDock, Smina, Autodock Vina and VinaXB. Each target is represented by one colour across 4 ROC figures. The diagonal represents a result from evenly distributed actives. The ROC curves skewed to left upper corner indicated a better discrimination between actives and decoys.



(a) (b)

Figure 5.13: Box plots demonstrate the AUROCC and EF05 of actives from 10 docking programmes. Each box plot illustrates the AUROCC and EF05 across 29 targets using each docking programme. The median line in the middle of each box plot represents the average performance of each docking programme using ROC or EF.

This means the results using ROC and EF to evaluate relatively agree with the results using median rank in the previous benchmark. Therefore, Smina and VinaXB were chosen for the virtual screening task given only one single programme is needed. For comparison, AUROCC of 0.623 and EF05 of 5.7 were set as milestones to compare with results from traditional and novel consensus scores.

Similar to the previous benchmark, 8 traditional consensus scores (MAX, MIN, MEAN, MEDIAN, EUC, CBM, ECR and DSR) were used to test the joint performance. Three normalisation methods were exploited to bring the docking scores from different scales to unified distribution. Another method, quantile normalisation, was added for the same purpose. Finally, the mean of AUROCC and EF05 values for 29 targets were obtained to represent each consensus score.

	MAX	MIN	MEAN	MEDIAN	EUC	CBM	ECR	DSR
min-max	0.562	0.623	0.614	0.612	0.613	0.611	0.614	0.612
rank	0.576	0.601	0.608	0.609	0.603	0.599	0.704	0.610
z-score	0.576	0.597	0.611	0.611	0.424	0.605	0.603	0.611
quantile	0.576	0.601	0.612	0.610	0.397	0.612	0.591	0.610

Table 5.8: Mean values of AUROCC represented docking programmes after different normalisation schemes. After normalised and combined, AUROCC was calculated for each target. The mean of 29 AUROCC values was obtained to evaluate how separated the actives and the decoys were for each target.

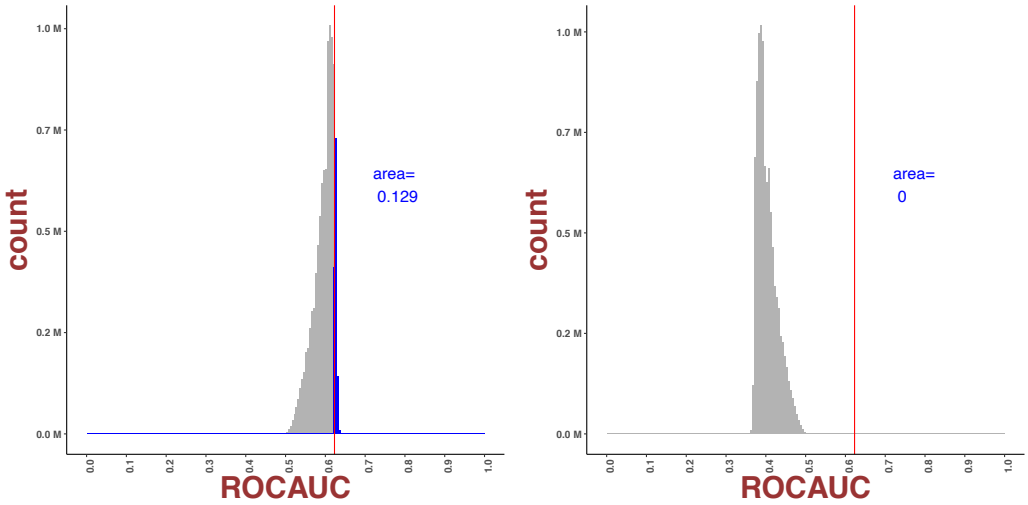
	MAX	MIN	MEAN	MEDIAN	EUC	CBM	ECR	DSR
min-max	2.4	3.0	5.5	5.4	4.6	3.4	5.2	5.2
rank	2.1	3.0	3.3	4.8	2.5	2.5	2.1	6.3
z-score	1.6	0.9	5.1	4.8	1.5	1.3	4.2	6.1
quantile	2.1	3.0	4.9	5.4	0.3	3.1	5.2	6.3

Table 5.9: Mean values of Enrichment Factor at 0.5% represented docking programmes after different normalisation schemes. After normalised and combined, EF at 0.5% was calculated for each target. The mean value of 29 EF05 was obtained to evaluate how much the ratio of actives in 0.5% top-ranked ligands was higher than that ratio in the entire set of ligands.

As demonstrated in Table 5.8 and 5.9, the conventional consensus scores in general showed no remarkable improvement compared to individual docking programmes. One exception for AUROCC values was 0.704 from the ECR score in the ranking normalisation scheme but the corresponding EF05 was reduced to 2.1. On the other hand, EF05 for DSR score in ranking, z-score and quantile normalisations slightly increased (6.3, 6.1 and 6.3) but the corresponding AUROCC dropped. This phenomenon is probably due to the insensitivity of MRSA data to these consensus scores. Despite numerous successful reports when applied consensus scores, in this case, the traditional consensus scores failed to improve the discrimination between decoys and actives for MRSA targets. Therefore, new novel consensus scores were expected to enhance the outcome for the task of virtual screening for MRSA proteins.

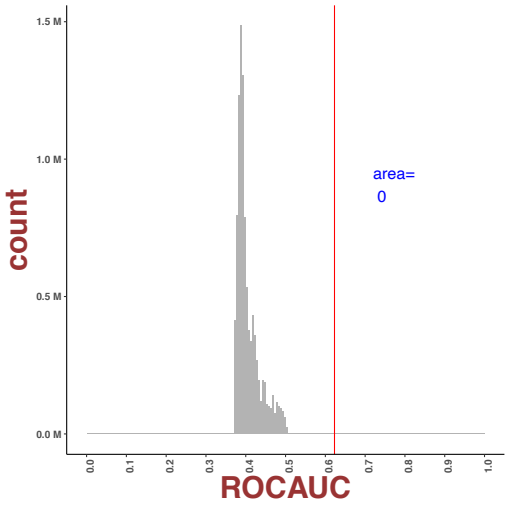
3.2 Novel Consensus Scores

For each docking programme, the AUROCC and EF05 of active ligands across 29 targets have been obtained and histogram plotted for visual presentation. To establish the improved performance of consensus scores (CS) over individual docking, the scores from the individual best performers were compared against the CS score line. This was estimated from the area patches to the right (better AUROCC and EF05) of the best performing individual docking platform (0.623 for AUROCC and 5.7 for EF05, identified by the solid vertical line in the histograms). The greater the patch area, the better the CS score line (compared to maximum AUROCC and EF05).

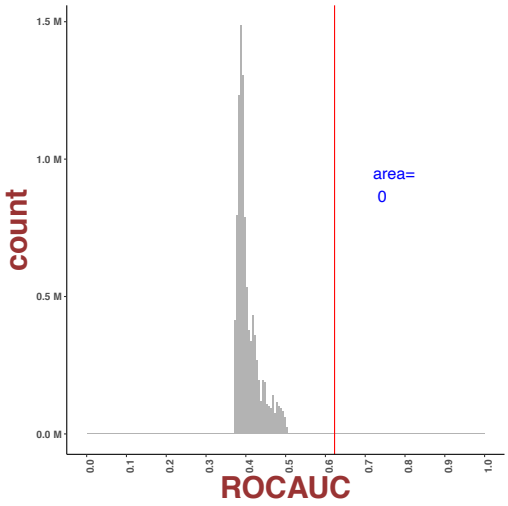


(a) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)$

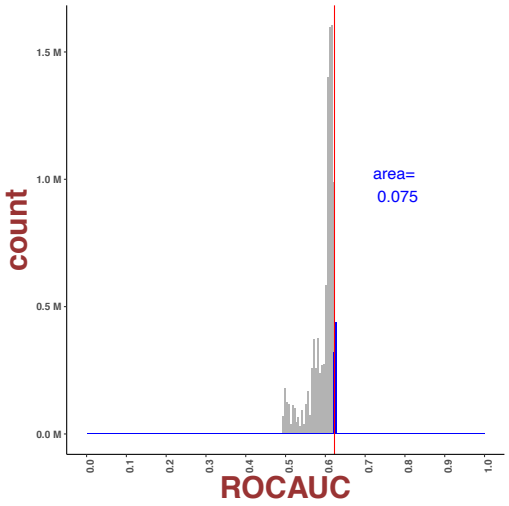
(b) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[S_i]$



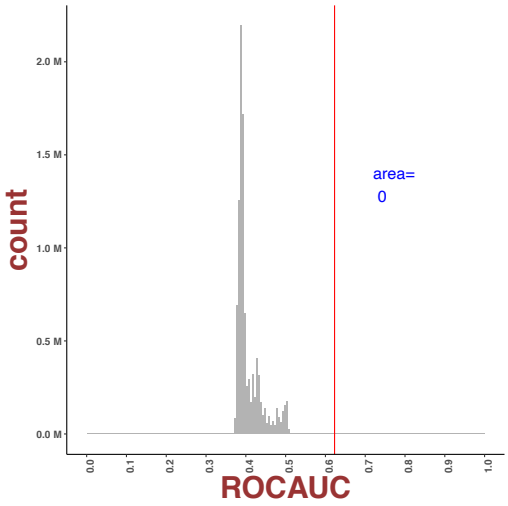
(c) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)^2$



(d) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[{S_i}^2]$

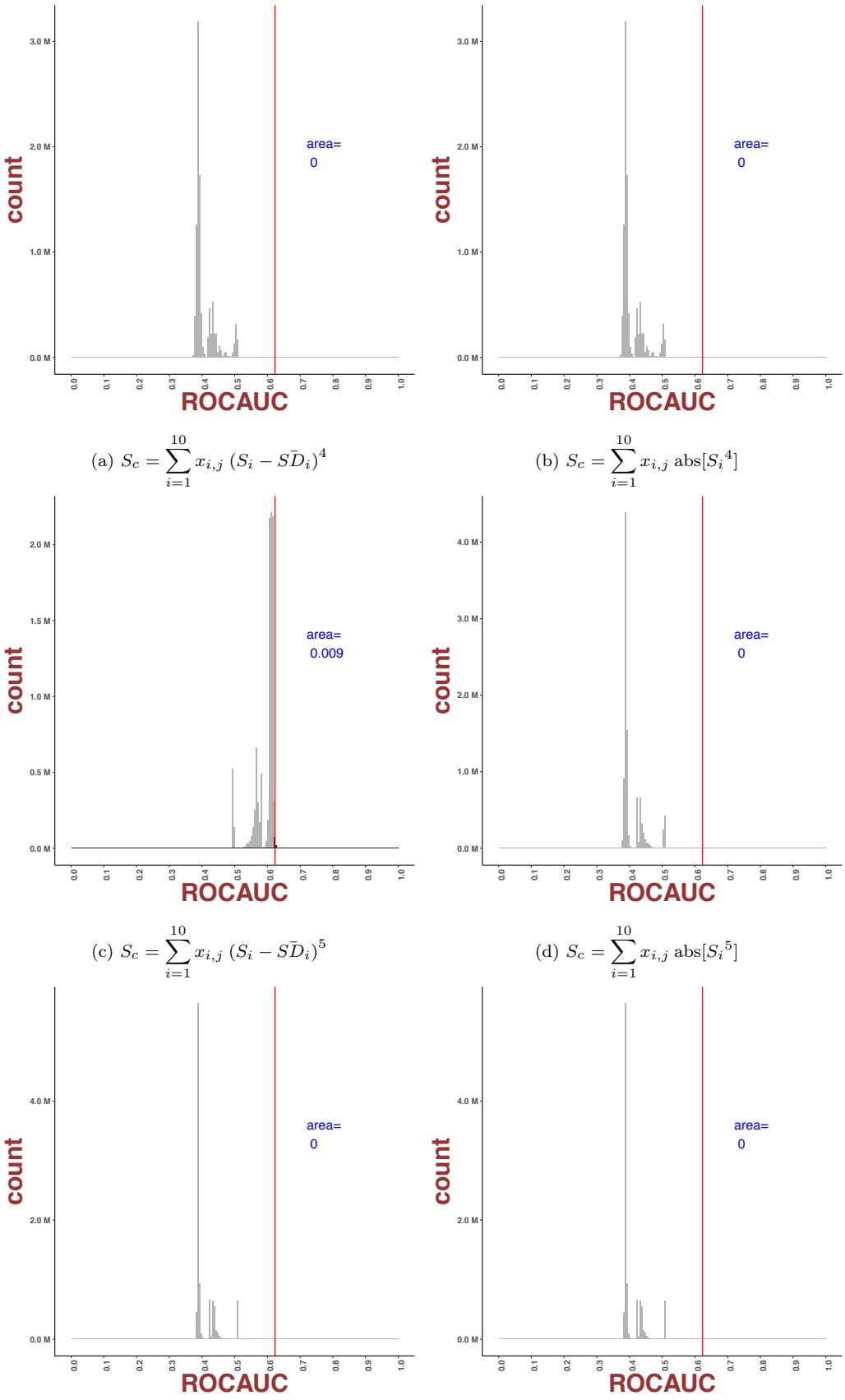


(e) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)^3$



(f) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[{S_i}^3]$

Figure 5.14: Histogram of mean AUROCC from models 4.16a and 4.16b with power from 1 to 3. Histograms from the model 4.16a were located on the left and histograms from the model 4.16b were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.



(a) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)^4$

(b) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[{S_i}^4]$

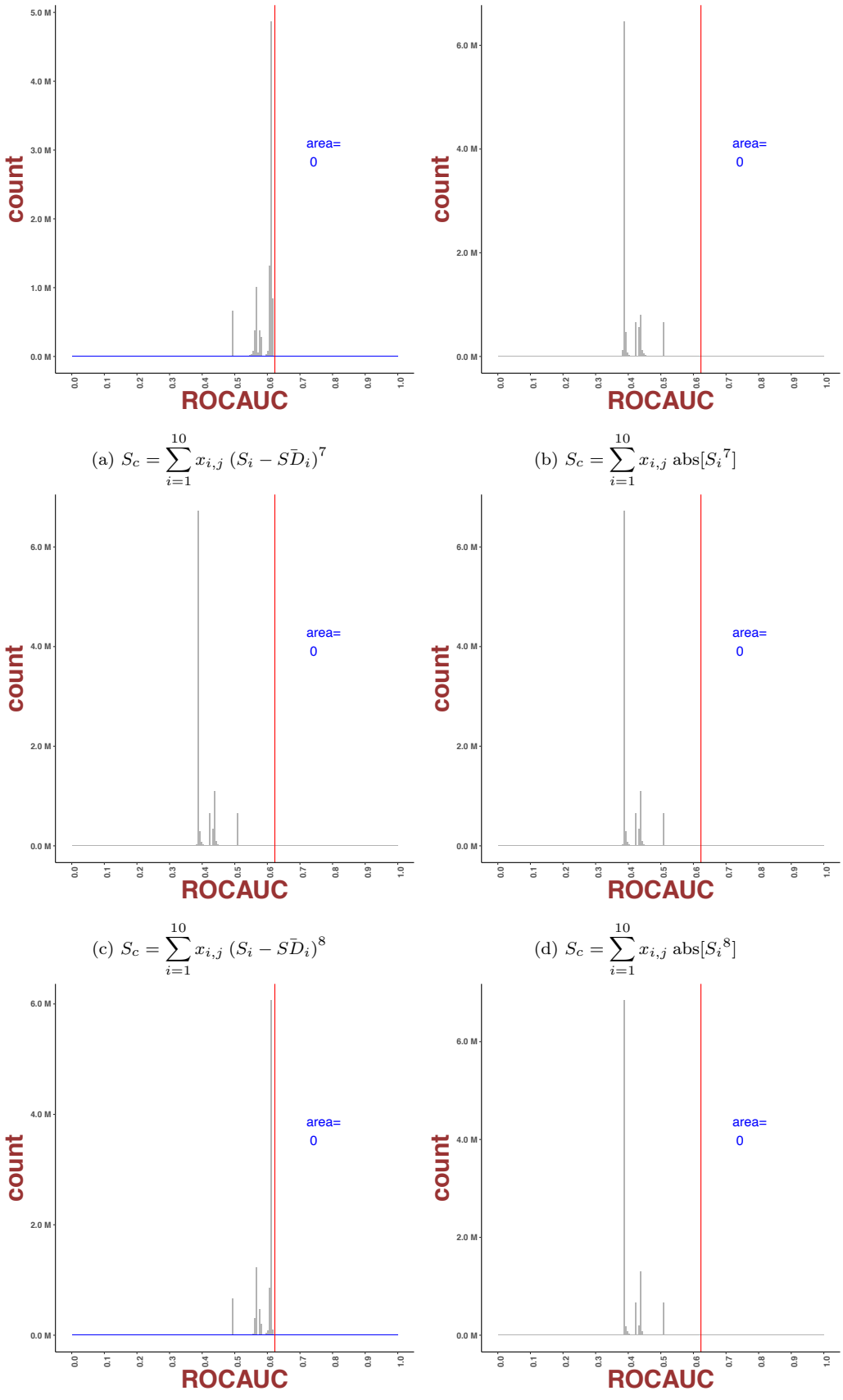
(c) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)^5$

(d) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[{S_i}^5]$

(e) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)^6$

(f) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[{S_i}^6]$

Figure 5.15: Histogram of mean AUROCC from models 4.16a and 4.16b with power from 4 to 6. Histograms from the model 4.16a were located on the left and histograms from the model 4.16b were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.



(a) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)^7$

(b) $S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[{S_i}^7]$

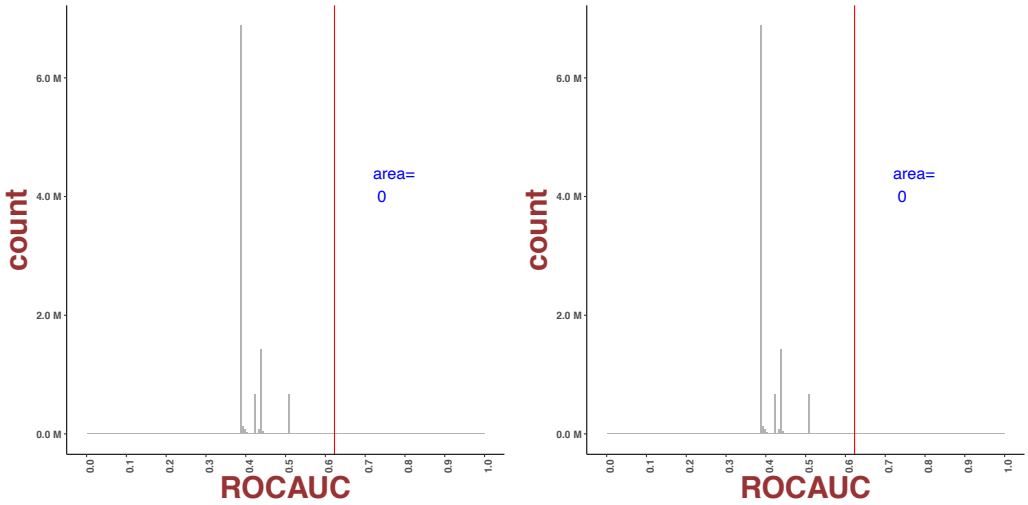
(c) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)^8$

(d) $S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[{S_i}^8]$

(e) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)^9$

(f) $S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[{S_i}^9]$

Figure 5.16: Histogram of mean AUROCC from models 4.16a and 4.16b with power from 7 to 9. Histograms from the model 4.16a were located on the left and histograms from the model 4.16b were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.



(a) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - S\bar{D}_i)^{10}$ (b) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[{S_i}^{10}]$

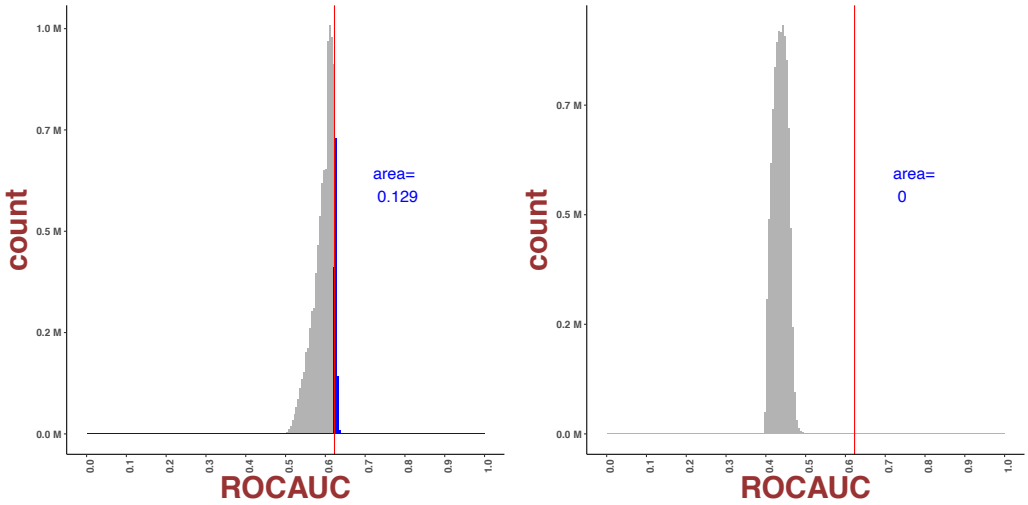
Figure 5.17: Histogram of mean AUROCC from models 4.16a and 4.16b with power 10. Histograms from the model 4.16a were located on the left and histograms from the model 4.16b were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.

Similar to the previous benchmark, the same trend was observed in histograms for AUROCC. The histogram fluctuated between two endpoints and between even and odd orders. The fluctuation was not observed using absolute values since all variables took positive values. This was explained by the fact that the docking scores mostly take negative values, so absolute values will take positive values and therefore causing the reverse in the ranks of actives and decoys. Therefore models using absolute values provided better consistency when the power varied. From this point forward, the even powers in models 4.16a, 4.16c and 4.18a will be ignored for a smoother comparison.

Another propensity was histograms tended to have a wider spread when the power values were small and tended to converge to a centre in the higher orders. This resulted in the fact that linear combinations reached higher maximum AUROCC values.

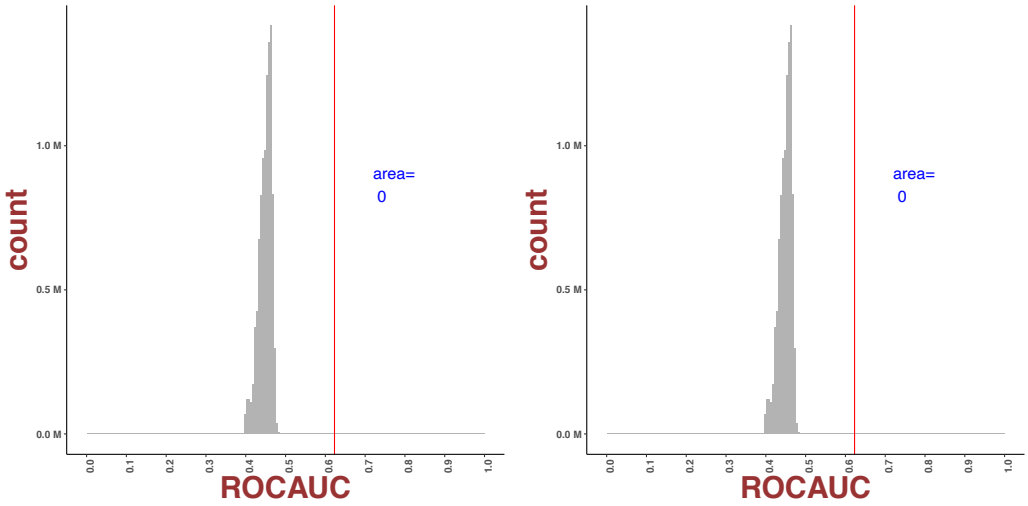
Unlike when using median rank to evaluate virtual screening performance, here the linear combination (n=1) from Eqns. 4.16a and 4.16c were not remarkable. The best AUROCC from both models were 0.702. There was no single improvement in models 4.16b and 4.16d. This was partly because the actives in the previous benchmark were randomly chosen rather than a set of actives in this benchmark.

Last but most importantly, the models 4.18a and 4.18b using standard deviation showed surprisingly encouraging results. As clearly demonstrated in Figure 5.14, 5.18 and 5.22, the histograms for models 4.18a and 4.18b were entirely located to the right of the red lines, given odd values were ignored for model 4.18a. Furthermore, AUROCC values spanned across from more than 0.623 to a maximum value of 0.873 in the linear model of Eqn. 4.18b. This value indicated a significant improvement, compared to the not-so-impressive AUROCC values from single docking programmes, suggesting that model 4.18b in linear mode is a good consensus score.



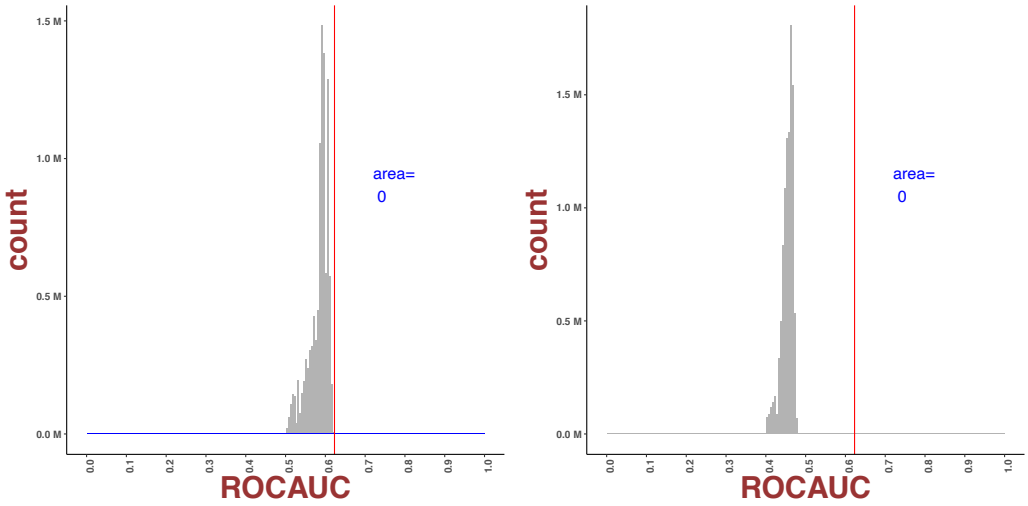
(a) $S_c = \sum_{i=1}^{10} x_{i,j}\ (S_i - \bar{S}_i)$

(b) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[S_i - \bar{S}_i]$



(c) $S_c = \sum_{i=1}^{10} x_{i,j}\ (S_i - \bar{S}_i)^2$

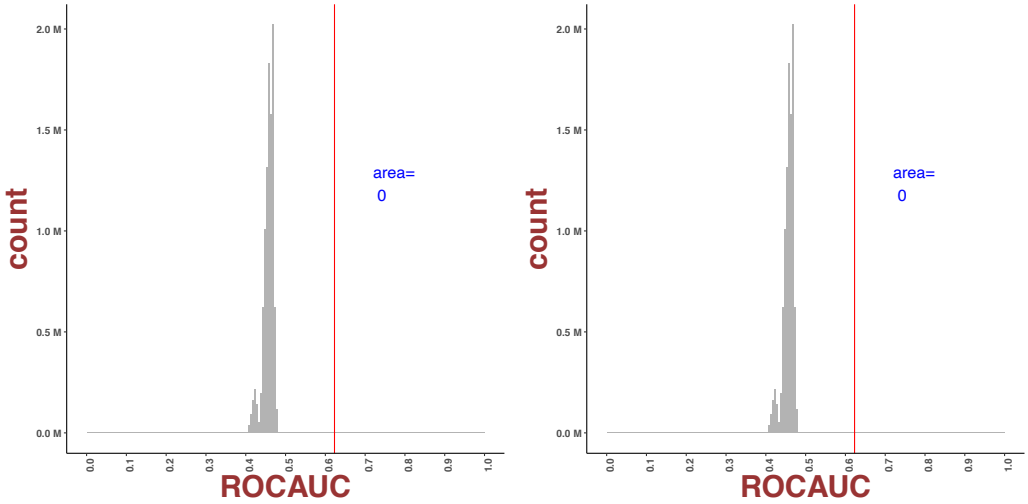
(d) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[(S_i - \bar{S}_i)^2]$



(e) $S_c = \sum_{i=1}^{10} x_{i,j}\ (S_i - \bar{S}_i)^3$

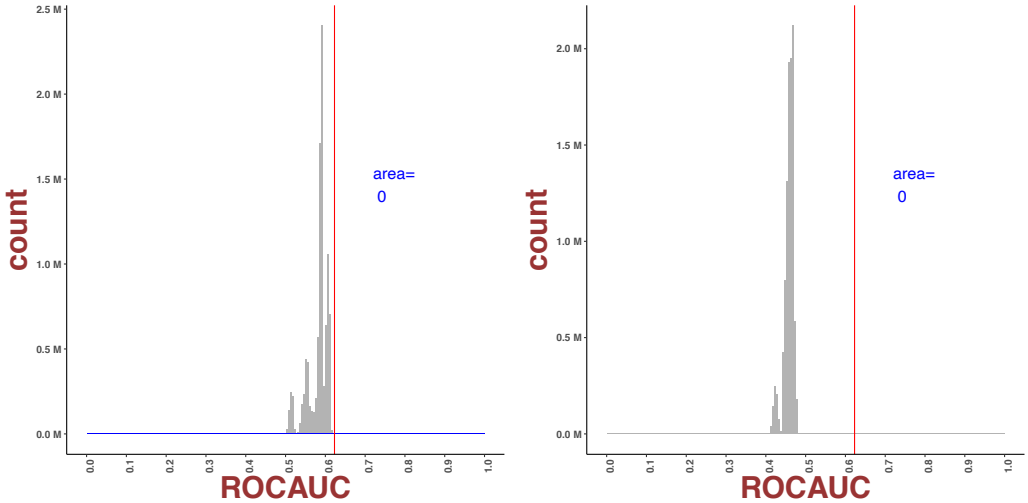
(f) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[(S_i - \bar{S}_i)^3]$

Figure 5.18: Histogram of mean AUROCC from models 4.16c and 4.16d with power from 1 to 3. Histograms from the model 4.16c were located on the left and histograms from the model 4.16d were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.



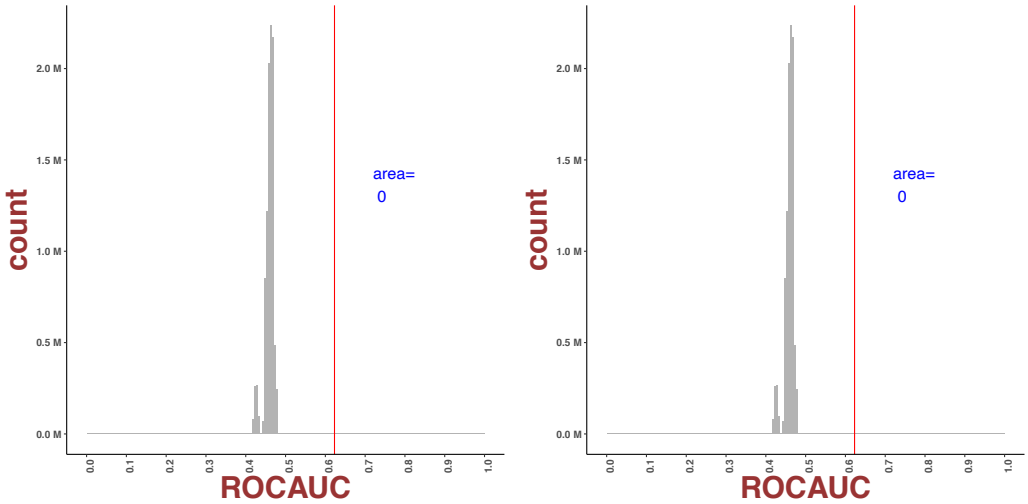
(a) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{S}_i)^4$

(b) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[(S_i - \bar{S}_i)^4]$



(c) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{S}_i)^5$

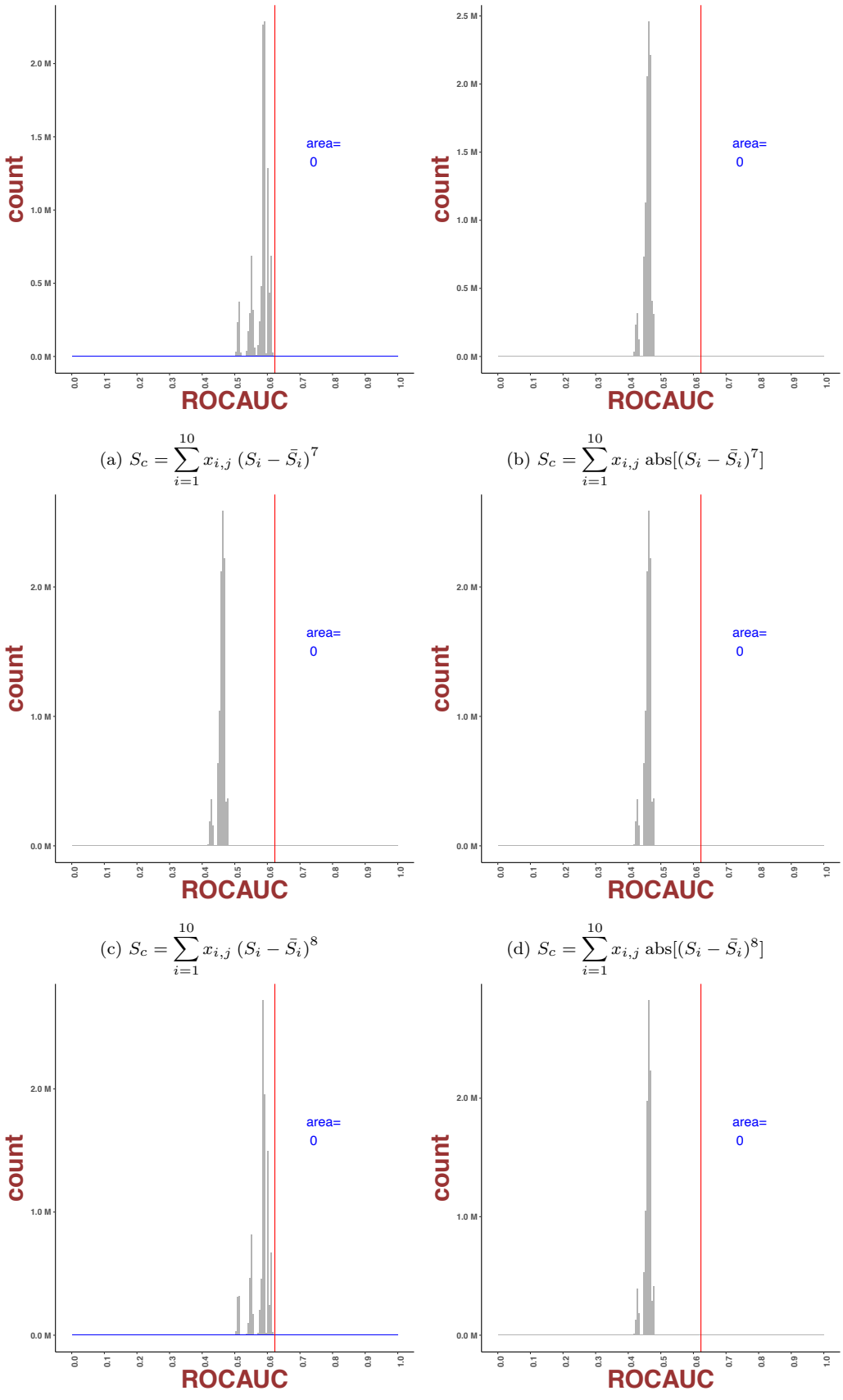
(d) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[(S_i - \bar{S}_i)^5]$



(e) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{S}_i)^6$

(f) $S_c = \sum_{i=1}^{10} x_{i,j} \, \text{abs}[(S_i - \bar{S}_i)^6]$

Figure 5.19: Histogram of mean AUROCC from models 4.16c and 4.16d with power from 4 to 6. Histograms from the model 4.16c were located on the left and histograms from the model 4.16d were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.



(a) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{S}_i)^7$

(b) $S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[(S_i - \bar{S}_i)^7]$

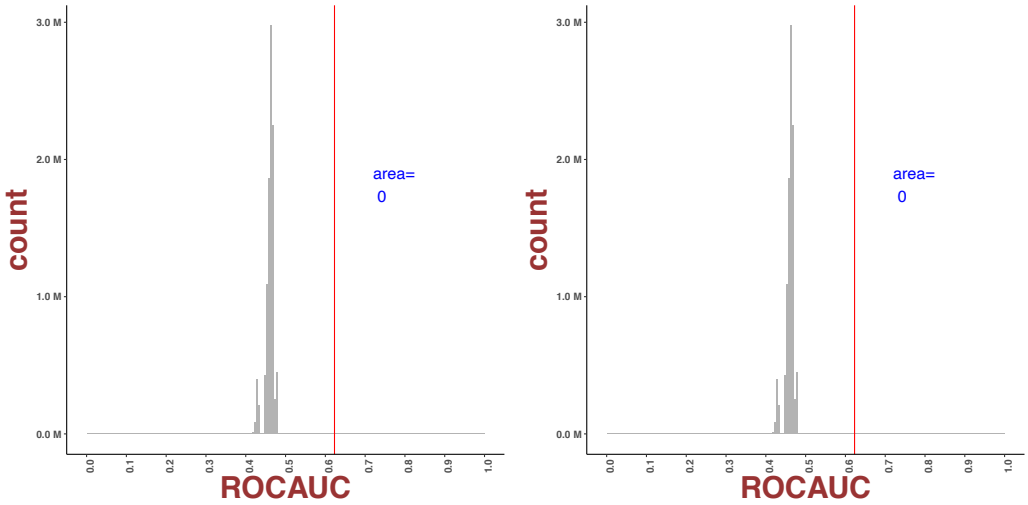
(c) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{S}_i)^8$

(d) $S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[(S_i - \bar{S}_i)^8]$

(e) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{S}_i)^9$

(f) $S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[(S_i - \bar{S}_i)^9]$

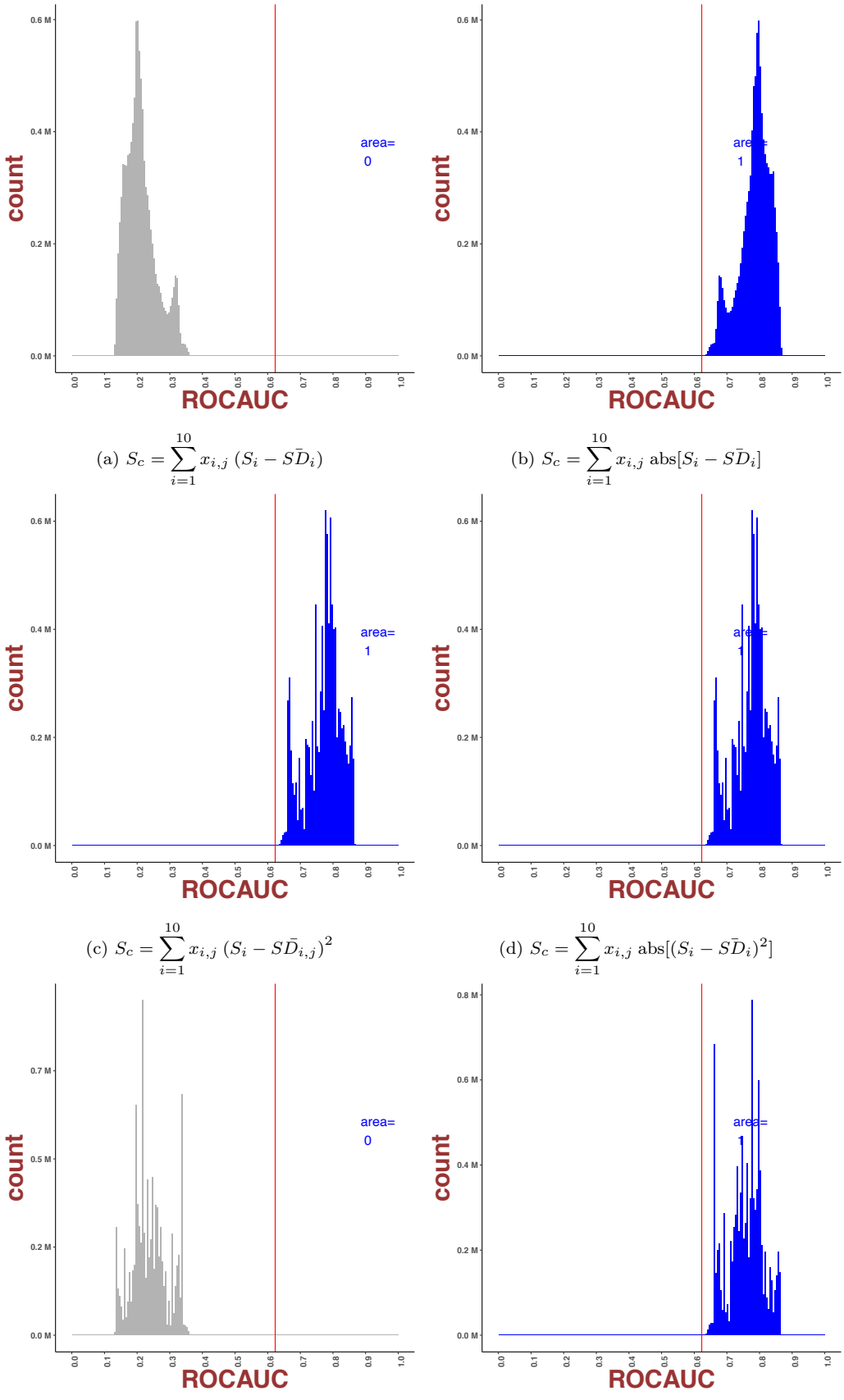
Figure 5.20: Histogram of mean AUROCC from models 4.16c and 4.16d with power from 7 to 9. Histograms from the model 4.16c were located on the left and histograms from the model 4.16d were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.



(a) $S_c = \sum_{i=1}^{10} x_{i,j} \, (S_i - \bar{S_i})^{10}$ (b) $S_c = \sum_{i=1}^{10} x_{i,j} \, \mathrm{abs}[(S_i - \bar{S_i})^{10}]$

Figure 5.21: Histogram of mean AUROCC from models 4.16c and 4.16d with power 10. Histograms from the model 4.16c were located on the left and histograms from the model 4.16d were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.

The same phenomenon was observed in histograms for EF05 (See Appendix 4). Again, for the model 4.18b, a major combination showed better early enrichment in the top-ranked ligands. The best EF05 was also obtained in model 4.18b using absolute values. The best value of EF05 was 19.1, which is also much higher than 5.7 in VinaXB. This value of EF05 means for this benchmark dataset, after using the consensus model 4.18b, the top 0.5% ranked ligands could contain 19.1 times higher than the ratio of actives in the entire dataset.



(a) $S_c = \sum_{i=1}^{10} x_{i,j}\ (S_i - S\bar{D}_i)$

(b) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[S_i - S\bar{D}_i]$

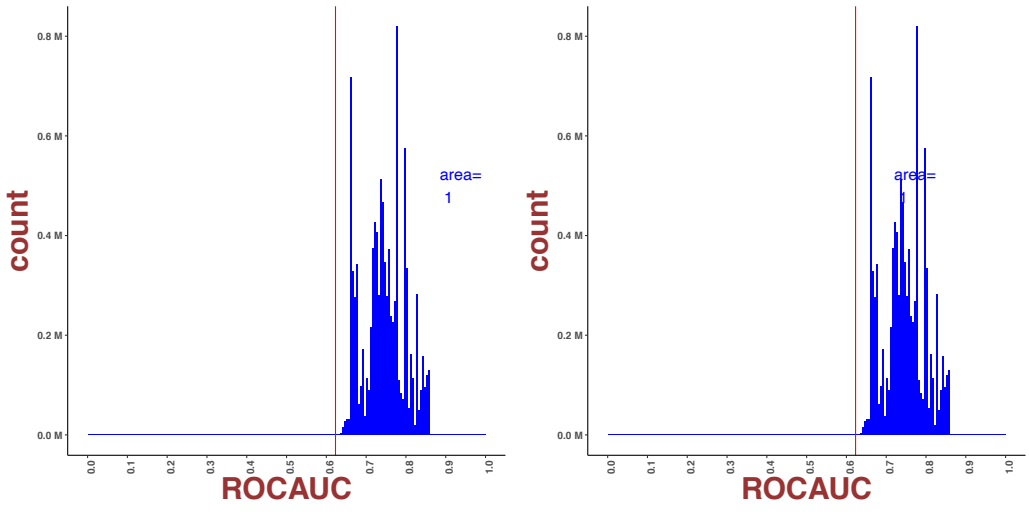
(c) $S_c = \sum_{i=1}^{10} x_{i,j}\ (S_i - S\bar{D}_{i,j})^2$

(d) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[(S_i - S\bar{D}_i)^2]$

(e) $S_c = \sum_{i=1}^{10} x_{i,j}\ (S_i - S\bar{D}_{i,j})^3$

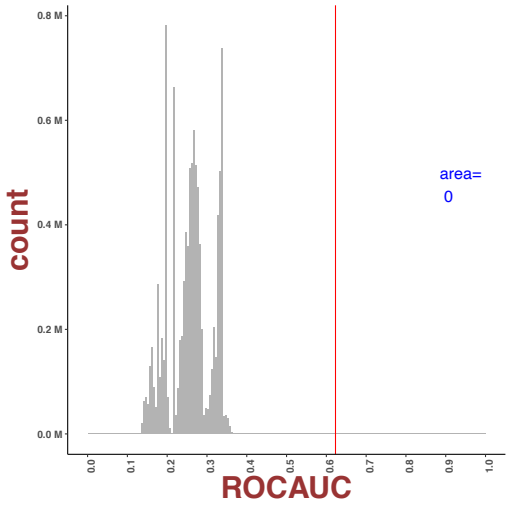
(f) $S_c = \sum_{i=1}^{10} x_{i,j}\ \mathrm{abs}[(S_i - S\bar{D}_i)^3]$

Figure 5.22: Histogram of mean AUROCC from models 4.18a and 4.18b with power from 1 to 3. Histograms from the model 4.16c were located on the left and histograms from the model 4.16d were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.

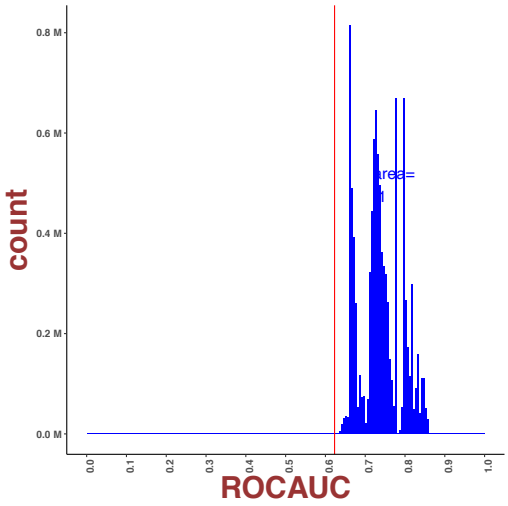


(a) $S_c = \sum_{i=1}^{10} x_{i,j}\,(S_i - S\bar{D}_{i,j})^4$

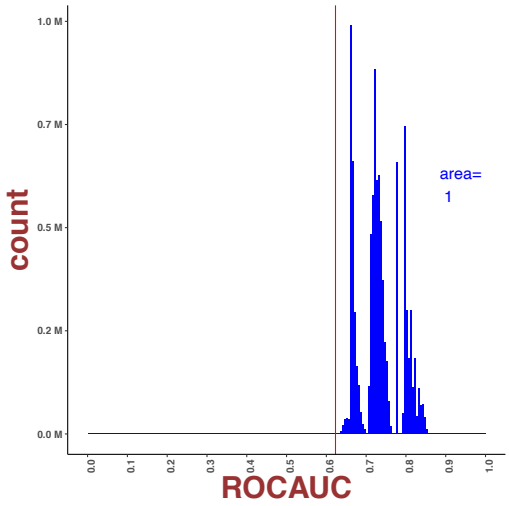
(b) $S_c = \sum_{i=1}^{10} x_{i,j}\,\mathrm{abs}[(S_i - S\bar{D}_i)^4]$



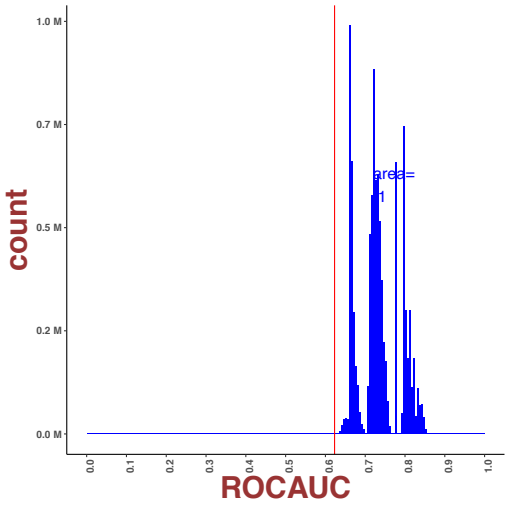
(c) $S_c = \sum_{i=1}^{10} x_{i,j}\,(S_i - S\bar{D}_{i,j})^5$



(d) $S_c = \sum_{i=1}^{10} x_{i,j}\,\mathrm{abs}[(S_i - S\bar{D}_i)^5]$

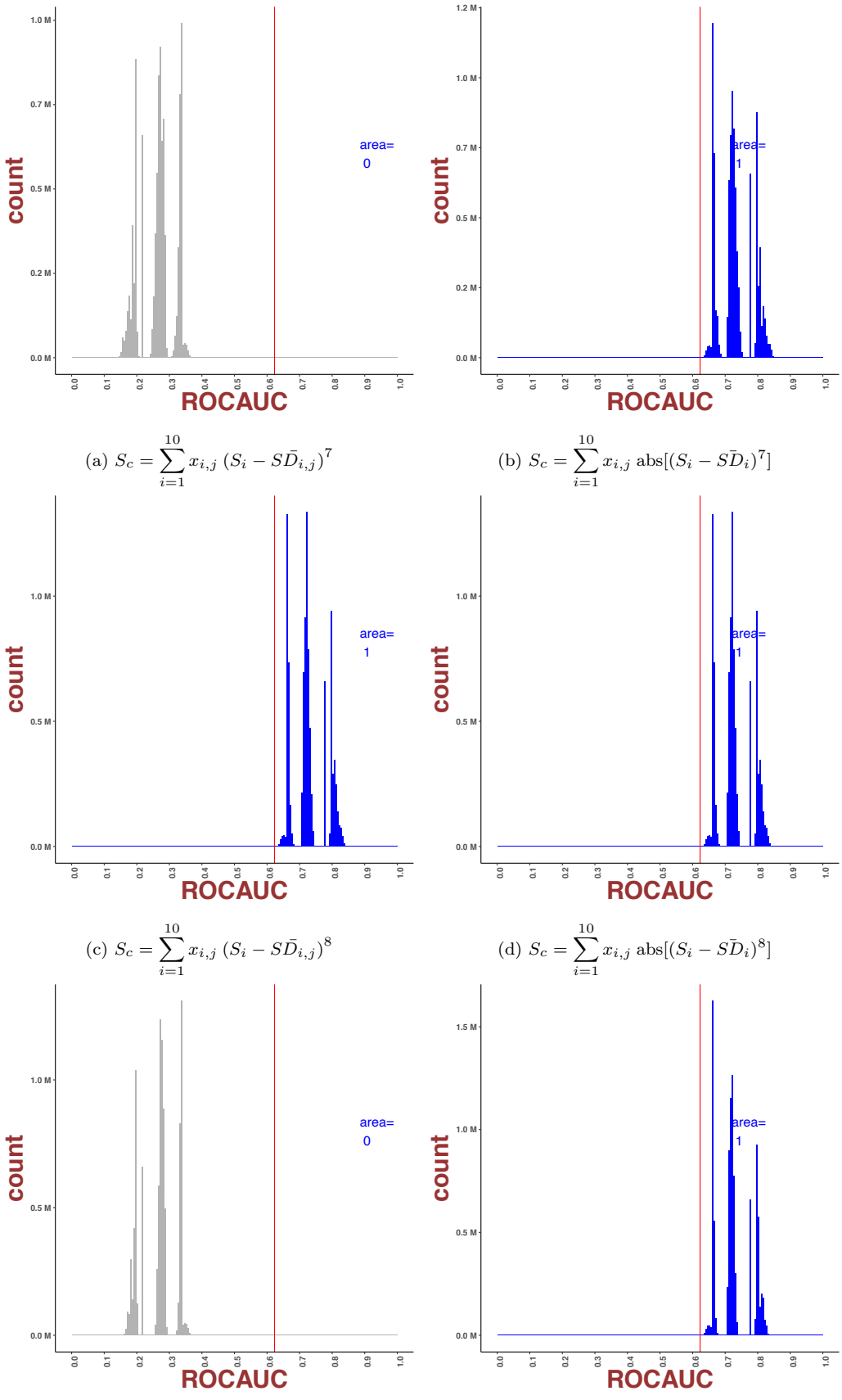


(e) $S_c = \sum_{i=1}^{10} x_{i,j}\,(S_i - S\bar{D}_{i,j})^6$



(f) $S_c = \sum_{i=1}^{10} x_{i,j}\,\mathrm{abs}[(S_i - S\bar{D}_i)^6]$

Figure 5.23: Histogram of mean AUROCC from models 4.18a and 4.18b with power from 4 to 6. Histograms from the model 4.16c were located on the left and histograms from the model 4.16d were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.



(a) $S_c = \sum_{i=1}^{10} x_{i,j}\,(S_i - S\bar{D}_{i,j})^7$

(b) $S_c = \sum_{i=1}^{10} x_{i,j}\,\mathrm{abs}[(S_i - S\bar{D}_i)^7]$

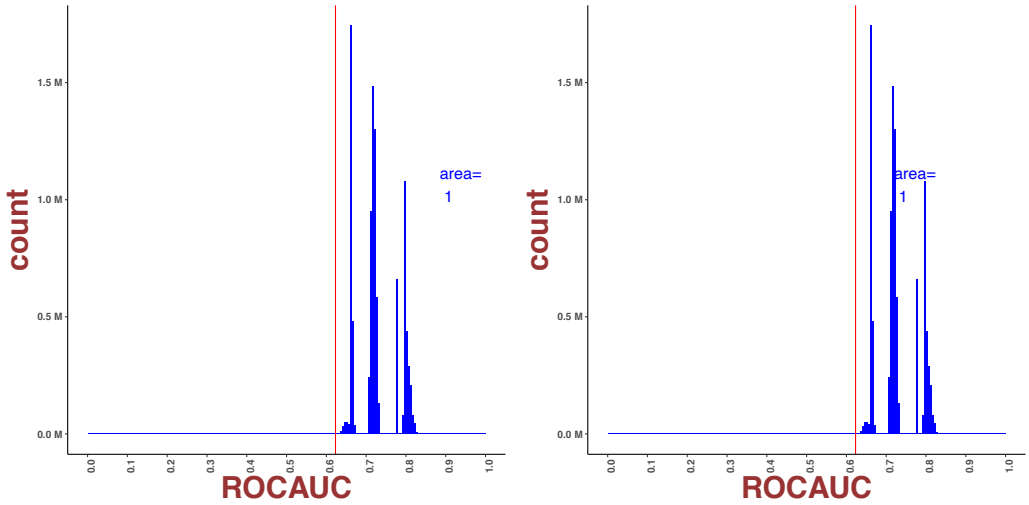
(c) $S_c = \sum_{i=1}^{10} x_{i,j}\,(S_i - S\bar{D}_{i,j})^8$

(d) $S_c = \sum_{i=1}^{10} x_{i,j}\,\mathrm{abs}[(S_i - S\bar{D}_i)^8]$

(e) $S_c = \sum_{i=1}^{10} x_{i,j}\,(S_i - S\bar{D}_{i,j})^9$

(f) $S_c = \sum_{i=1}^{10} x_{i,j}\,\mathrm{abs}[(S_i - S\bar{D}_i)^9]$

Figure 5.24: Histogram of mean AUROCC from models 4.18a and 4.18b with power from 7 to 9. Histograms from the model 4.16c were located on the left and histograms from the model 4.16d were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.



(a) $S_c = \sum_{i=1}^{10} x_{i,j}\,(S_i - S\bar{D}_{i,j})^{10}$ (b) $S_c = \sum_{i=1}^{10} x_{i,j}\,\mathrm{abs}[(S_i - S\bar{D}_i)^{10}]$

Figure 5.25: Histogram of mean AUROCC from models 4.18a and 4.18b with power 10. Histograms from the model 4.16c were located on the left and histograms from the model 4.16d were located on the right. The number of combinations is 10015005. The red vertical line represents the best AUROCC amongst 10 docking programmes, which is 0.623. The portion of the histogram to the right of the vertical line was shaded blue, representing the combinations with better AUROCC. The rest of the histogram was shaded grey.

These findings of the model 4.18b suggested that it consistently showed significant improvement in both AUROCC and EF05. Consequently, the model 4.18b with the linear model was chosen to be the consensus score for virtual screening of repurposable ligands against MRSA targets for potential candidates with anti-MRSA activity. Note, since the ROC scores are positive definite compared to negative docking (attractive energy based) scores, the blue patches sat to the right of the best individual docking scoreline (red line).

As evident from Figures 5.14, 5.18 and 5.22, linear regression over the set of 10 docking scores involving the ligand:protein sets returned better docking scored than higher-order model.

Power	AUROCC	EF05	
1	0.873	19.1	
2	0.87	18.1	
3	0.864	16.9	
4	0.86	16.6	
5	0.859	16.3	
6	0.856	15.7	
7	0.852	15.4	
8	0.845	14.6	
9	0.839	14.5	
10	0.833	14.5	

Table 5.10: Table of maximum AUROCC and EF05 that model 4.18b achieved from power 1 to 10. The maximum AUROCC and EF05 decreased when the power increased.

EF05 also witnessed similar patterns and also agreed model 4.18b is the best model amongst all the models used. Histograms for metric EF05 available in Appendix 4.

When combining both AUROCC and EF05, model 4.18b appeared to be the best model with the ability to improve such the discrimination between actives and decoys and boosted the early enrichment of actives at the top of the ranked list of ligands. Given that power 1 of model 4.18b has the potential to enhance the AUROCC and EF05 better than higher powers, the linear form of model 4.18b was chosen as the best model to apply for virtual screening of prospective MRSA dataset.

In the next sessions, a subset of combinations that produced the best AUROCC and

EF05 was extracted to process for a distinct consensus score.

3.3 Conclusion

In this benchmark, consensus scoring algorithms using MRSA datasets and ten docking programmes (DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, rDock, Smina, Autodock Vina and VinaXB) were investigated. The performance benchmark metrics were the ROC and EF. The performance of single docking programmes was found to be relatively in line with the previous benchmark with Smina showing the best AUROCC while DOCk showed poor results in both benchmarks. Other programmes showed moderate performance.

The individual docking programmes were also compared with conventional consensus scores (minimum, maximum, mean, median, Reciprocal Rank, Euclidean distance, Cubic Mean and Deprecated Sum Rank) and recently reported Exponential Consensus Rank score. Before consensus scoring, the distribution of docking scores was altered with 4 normalisation methods (rank, min-max scaling, z-scores and quantile) to offer a direct comparison with commonly used statistical consensus scores. Comparisons show that the MRSA-related dataset is not sensitive to traditional consensus scores, showing no improved AUROCC and EF05 compared to the best single programme.

The novel consensus scores consistently perform better than individual docking programmes on the MRSA benchmark dataset. In this work, raw docking scores from ten docking programmes (DOCK, Gemdock, Ledock, PLANTS, PSOVina, QuickVina2, rDock, Smina, Autodock Vina and VinaXB) were directly combined. Due to an exhaustive search of combinations, there was no obligation for data normalisation. Results showed that the novel model gave better AUROCC and EF05 of active ligands across benchmark datasets.

One outcome of the novel consensus module was the preponderance of linear combination of docking scores towards improved active ligand ranking over higher-order consensus formulas. As of the higher-order scores, as in Eqns. 4.16a-4.16d and 4.18a-4.18b, odd ordered combinations show consistently better combinatorics than their even ordered counterparts.

One key finding in this benchmark was that consensus models using standard deviation (models 4.18a and 4.18b) produced significantly improved AUROCC and EF05 values, with model 4.18b more consistent than model 4.18a. With the maximum AUROCC and EF05 values in linear combinations higher than those in a higher power, the linear form of model 4.18b was marked as the best-optimised model to be applied in the virtual screening for MRSA-targeted ligands.

Chapter 6

Enriched Subset of Potential Candidates for Anti-MRSA Repurposing

1 Enriched subset of potential candidates for anti-MRSA repurposing

After confirmation with the previous benchmark section, the Eqn. 4.18b was chosen as the consensus model. A total of 10015005 coefficient ensembles were clustered using hierarchical clustering. The cluster with best AUROCC and EF05 was chosen, followed by execution for the square root means, resulting in the consensus equation:

$$\begin{aligned} S_c = & abs(0.093 * DOCK) + abs(0.434 * Ledock) + abs(0.038 * PLANTS) \\ & + abs(0.244 * PSOVina) + abs(0.027 * QuickVina2) + abs(0.154 * rDock) \\ & + abs(0.036 * Smina) + abs(0.031 * Vina) + abs(0.022 * VinaXB) \end{aligned} \quad (6.1)$$

From the equation 6.1, the programmes that contribute the most to Eqn. 6.1 included Ledock, PSOVina and rDock in descending order. Others programmes show minor beneficence, while there was no contribution from Gemdock to the final consensus results. To confirm the plausibility of Eqn. 6.1, the variables were substituted with MRSA benchmark docking scores. The results showed a range of promising AUROCC and EF05 across 29 targets, with the mean AUROCC of 0.865 and mean EF05 of 17.6. This result

The model 6.1 was applied to the docking data of repurposable ligands against MRSA targets. 0.5% of top scored from 5902 ligands were chosen for each protein, resulting in a subset of 30 ligands in Table 6.1 and 6.2.

Table 6.1: List of proteins and 30 potential ligands.
The protein name and chain in the first column represents the MRSA target hit from the BLAST alignments. The ligands in the second column were 30 top-ranked compounds of each target after applying the model 6.1 to the MRSA docking dataset. These ligands were listed in descending order in terms of predicted binding affinity to their corresponding protein target. These compounds were obtained from Repurposing Hub (Corsello et al., 2017)

Protein	**Ligands**
1HSK_A	candicidin, echinomycin, actinomycin-d, dactinomycin, KB-SRC-4, amphotericin-b, epacadostat, salvianolic-acid-A, adenosine-triphosphate, uridine-5-triphosphate, INS316, guadecitabine, EPZ-5676, baricitinib, S-3304, glipizide, SR-3306, rifamycin, XL147, DOTMP, gliquidone, TCN201, MK-7246, pyrintegrin, citicoline, SDZ-220-581, GSK3326595, SirReal-2, PF-04217903, ramatroban

Continuation of Table 6.1	
Protein	**Ligands**
1JIL_A	echinomycin, dactinomycin, candicidin, everolimus, actinomycin-d, epicatechin-gallate-(-), epigallocatechin-gallate-(-), guadecitabine, tetrahydrofolic-acid, citicoline, lometrexol, uridine-5-triphosphate, banoxantrone, 1,5-dicaffeoylquinic-acid, reynoutrin, theaflavin, riboflavin-5-phosphate-sodium, cefpirome, rutin, salvianolic-acid-A, cefonicid, adenosine-triphosphate, EB-47, JNJ-64619178, PF-477736, diosmin, DOTMP, neohesperidin, YM-201636, myricitrin
1KA5_A	echinomycin, candicidin, INS316, uridine-5-triphosphate, mocetinostat, triciribine-phosphate, IB-MECA, CF102, folic-acid, fludarabine-phosphate, pazopanib, VU591, STAT3-inhibitor-VI, adenosine-triphosphate, CGP-71683, darolutamide, glipizide, JI-101, pevonedistat, tetrahydrofolic-acid, chlorogenic-acid, BVD-523, H2L-5765834, PF-573228, pyrintegrin, adenosine-phosphate, edaglitazone, diosmin, URB597, KS-176
1LM4_A	echinomycin, CDBA, riboflavin-5-phosphate-sodium, uridine-5-triphosphate, riboflavin, INS316, rutin, adenosine-triphosphate, famotidine, guadecitabine, isoquercitrin, kuromanin, raltitrexed, citicoline, tucatinib, bisindolylmaleimide-IX, DOTMP, FCE-22250, lometrexol, folic-acid, MIW-815, C188-9, HER2-Inhibitor-1, KB-SRC-4, methotrexate, etoposide-phosphate, PIK-294, GNTI, tetrahydrofolic-acid, azosemide
1LRZ_A	dactinomycin, candicidin, TPPS4, ammonium-glycyrrhizinate, guadecitabine, diosmin, alpha-glucosyl-hesperidin, sennoside-A, icariin, procyanidin-B-2, riboflavin-5-phosphate-sodium, apramycin, hesperidin, CGP-71683, sennoside-protonated, digitoxin, etoposide-phosphate, PSB-603, fostamatinib, PF-573228, WAY-600, KB-SRC-4, hypericin, adavivint, BMS-935177, CDBA, riboflavin, EPZ-5676, adenosine-triphosphate, elinogrel
1NYR_A	echinomycin, dactinomycin, actinomycin-d, evans-blue, PF-05212384, candicidin, CDBA, sirolimus, TPPS4, amphotericin-b, uridine-5-triphosphate, INS316, guadecitabine, citicoline, adenosine-triphosphate, rutin, safflower-yellow, thiamine-pyrophosphate, procyanidin-B-2, DOTMP, hyperin, inarigivir, PDD-00017273, cromoglicic-acid, ribostamycin-sulfate, neomycin, astilbin, riboflavin-5-phosphate-sodium, kuromanin, deforolimus
1QE0_A	echinomycin, TG-100801, Mps1-IN-5, candicidin, ivermectin, CDBA, TPPS4, rifapentine, MIW-815, HER2-Inhibitor-1, BMS-817378, sennoside-protonated, rutin, INS316, guadecitabine, 4-galactosyllactose, hydroxysafflor-yellow-A, uridine-5-triphosphate, TAK-243, SU11274, adenosine-triphosphate, baicalin, myricitrin, isepamicin, R-428, tucatinib, KB-SRC-4, epirubicin, dihydroergotamine, ceftobiprole
1QU2_A	echinomycin, KB-SRC-4, evans-blue, VER-155008, adenosine-triphosphate, JNJ-64619178, sennoside-protonated, alpha-glucosyl-hesperidin, uridine-5-triphosphate, entasobulin, hesperidin, R-428, integrin-antagonist-1, HER2-Inhibitor-1, MK-8033, rutin, CK-101, sulfatinib, baicalin, salvianolic-acid-A, diosmin, adenosine-phosphate, AMG900, ZM-241385, tetrahydrofolic-acid, inarigivir, pexidartinib, cot-inhibitor-1, nafamostat, AZD8835

Continuation of Table 6.1	
Protein	**Ligands**
1QXY_A	echinomycin, candicidin, gamithromycin, rifaximin, AZD5991, uridine-5-triphosphate, citicoline, adenosine-triphosphate, INS316, brivudine, metafolin, E7449, famotidine, trifluridine, TAK-243, idoxuridine, rutin, triciribine-phosphate, tetrahydrofolic-acid, minodronic-acid, valganciclovir, diminazene-aceturate, Ro-9187, broxuridine, alanosine, monosodium-alpha-luminol, resmetirom, polyinosine, lometrexol, tucatinib
1XPK_A	echinomycin, actinomycin-d, candicidin, adenosine-triphosphate, dactinomycin, evans-blue, INS316, phthalylsulfathiazole, triciribine-phosphate, CHIR-98014, uridine-5-triphosphate, pazopanib, polyinosine, PF-05089771, tedizolid-phosphate, benzthiazide, mibampator, PF-915275, CGP-71683, EB-47, danirixin, L-368899, ICA-121431, BEBT-908, TCN201, cefonicid, TPPS4, GSK1292263, HER2-Inhibitor-1, 4SC-202
1ZOW_A	dactinomycin, actinomycin-d, echinomycin, candicidin, INS316, linsitinib, PF-573228, LY2784544, RGFP966, MK-5108, XL147, PSB-06126, eletriptan, integrin-antagonist-1, GS-143, PF-06873600, metafolin, VER-155008, hPGDS-IN-1, sulfatinib, ceftiofur, PF-03814735, PF-8380, C646, ONO-8130, isoquercitrin, adenosine-triphosphate, uridine-5-triphosphate, lifirafenib, adoprazine
2B7L_A	echinomycin, candicidin, dactinomycin, rutin, NMS-E973, DOTMP, KB-SRC-4, INS316, pirarubicin, mangafodipir, ML228, peposertib, ticagrelor, hyperin, GSK2126458, JW-74, hydroxysafflor-yellow-A, COH29, LX1031, uridine-5-triphosphate, cilengitide, epigallocatechin-gallate-(-), avagacestat, BMS-214662, cilofexor, etoposide-phosphate, AZD3264, astilbin, SDZ-220-040, OTX015
2FRH_A	candicidin, dactinomycin, actinomycin-d, SR-3306, adenosine-triphosphate, amuvatinib, INS316, HER2-Inhibitor-1, KB-SRC-4, cefmenoxime, cefonicid, mocetinostat, triciribine-phosphate, uridine-5-triphosphate, AMG900, theaflavin, diosmin, GSK2126458, cefamandole, guadecitabine, grapiprant, TC-G-1008, hypericin, lurasidone, pazopanib, 4EGI-1, LY393558, XL147, CGP-71683, G-749
2GQD_A	echinomycin, dactinomycin, actinomycin-d, eprinomectin, rose-bengal, uridine-5-triphosphate, raltitrexed, adenosine-triphosphate, folic-acid, 4EGI-1, GS-9973, nafamostat, sapropterin, tiotidine, CHIR-98014, TWS-119, AMG900, pradefovir, zaprinast, indacaterol, fosfructose, carmoterol, R112, cromoglicic-acid, aminopterin, methotrexate, ebrotidine, telatinib, genistein, MGCD-265
2H29_A	echinomycin, actinomycin-d, candicidin, deforolimus, everolimus, dactinomycin, sirolimus, rifabutin, nystatin, amphotericin-b, rifaximin, adenosine-triphosphate, tedizolid-phosphate, iloperidone, uridine-5-triphosphate, PF-05089771, cefonicid, epigallocatechin-gallate-(-), aminopterin, methotrexate, CHIR-98014, PF-04937319, tedizolid, folic-acid, dacinostat, pazopanib, phlorizin, hematoporphyrin, MK-8033, pemetrexed

Continuation of Table 6.1	
Protein	**Ligands**
2H92_A	candicidin, echinomycin, actinomycin-d, INS316, uridine-5-triphosphate, adenosine-triphosphate, salazodine, triciribine-phosphate, PF-04217903, Ro-61-8048, fosfructose, salidroside, sitaxentan, VU591, telatinib, STAT3-inhibitor-VI, TCN201, fludarabine-phosphate, NT157, adenosine-phosphate, 4EGI-1, AMG-208, SNS-314, darolutamide, sinefungin, bicalutamide, SR-27897, basmisanil, UNC2327, GR-113808
2HK2_A	echinomycin, actinomycin-d, candicidin, everolimus, dactinomycin, deforolimus, zotarolimus, temoporfin, adenosine-triphosphate, epigallocatechin-gallate-(-), epicatechin-gallate-(-), uridine-5-triphosphate, amygdalin, folic-acid, tetrahydrofolic-acid, naringin-dihydrochalcone, KB-SRC-4, T-5224, cefonicid, ZCL-278, INS316, AMG900, ICA-121431, sodium-picosulfate, PF-05089771, telatinib, PF-477736, reynoutrin, EB-47, rutin
2I80_A	candicidin, dactinomycin, echinomycin, actinomycin-d, uridine-5-triphosphate, INS316, SCH-58261, ebrotidine, epacadostat, L-694247, labetalol, resminostat, VU591, tiotidine, rosmarinic-acid, MK-5046, AMG-517, pentamidine, adenosine-triphosphate, delphinidin, TC1, TCS-2210, adenosine-phosphate, leteprinim, NT157, arotinolol, AT13148, famotidine, acedapsone, R112
2JFQ_A	candicidin, digitoxin, echinomycin, actinomycin-d, dactinomycin, olsalazine, famotidine, pirinixic-acid, benserazide, risedronate, imidurea, triciribine-phosphate, fosfructose, zoledronic-acid, nifuroxazide, HSR6071, GGsTop, olomoucine, dynasore, amiloride, tiludronate, adaprev, PIK-93, epacadostat, adefovir, luteolin, SR-27897, VU591, rufloxacin, sulfasalazine
2N8N_A	dactinomycin, actinomycin-d, candicidin, astilbin, DOTMP, hydroxysafflor-yellow-A, INS316, adenosine-triphosphate, epicatechin-gallate-(-), GSK2239633A, demeclocycline, prednisolone-sodium-phosphate, myricitrin, ticagrelor, PSB-06126, kuromanin, hypericin, rutin, uridine-5-triphosphate, hydrocortisone-phosphate, pazopanib, epigallocatechin-gallate-(-), 4-galactosyllactose, XL147, LB42708, licogliflozin, sulfatinib, PIK-294, GSK2334470, phlorizin
2NM3_A	echinomycin, actinomycin-d, CHIR-98014, kuromanin, adenosine-triphosphate, INS316, uridine-5-triphosphate, CGP-71683, esculin, guadecitabine, DOTMP, baicalin, XL228, EMD-1214063, hypericin, reynoutrin, losartan, CaMKII-IN-1, ceftriaxone, raltitrexed, ganetespib, LY3295668, cot-inhibitor-2, epirubicin, cefpirome, casanthranol-variant, azilsartan, MCC950, ZD-7155, rutin
2QV7_A	echinomycin, uridine-5-triphosphate, INS316, adenosine-triphosphate, guadecitabine, cefonicid, baicalin, amygdalin, metafolin, citicoline, folic-acid, BMS-599626, theaflavin, GNTI, rutin, EB-47, epacadostat, procyanidin-B-2, TAK-243, hydroxysafflor-yellow-A, cefazolin, epigallocatechin-gallate-(-), PK-44, taprenepag, canagliflozin, cefotetan, XL228, PSB-603, hesperidin, raltitrexed

Continuation of Table 6.1	
Protein	**Ligands**
2VPQ_A	everolimus, dactinomycin, sirolimus, candicidin, echinomycin, actinomycin-d, zotarolimus, rifabutin, nystatin, rifapentine, AZD5991, cefonicid, evans-blue, adenosine-triphosphate, guadecitabine, hesperidin, EB-47, uridine-5-triphosphate, diosmin, ceftriaxone, m-THP, rutin, tedizolid-phosphate, naringin, INS316, T-5224, AMI-1, inarigivir, riboflavin-5-phosphate-sodium, JNJ-64619178
2X4K_A	candicidin, echinomycin, CDBA, TPPS4, INS316, uridine-5-triphosphate, ML228, 4EGI-1, AST-1306, ML193, TC-G-1008, N-(2-chlorophenyl)-2-((2E)-2-[1-(2-pyridinyl)ethylidene]hydrazinocarbothioyl)hydrazinecarbothioamide, G-749, thiamine-pyrophosphate, XL147, trabodenoson, CGP-78608, verdinexor, lometrexol, JNJ-64619178, CGP-71683, epacadostat, gepotidacin, MK-0773, PRX-08066, JTE-013, CGP-60474, folic-acid, EPZ015666, C188-9
2X7I_A	dactinomycin, echinomycin, candicidin, actinomycin-d, eprinomectin, adenosine-triphosphate, fludarabine-phosphate, uridine-5-triphosphate, INS316, pradefovir, TG-100713, aminopterin, SAM-315, adenosine-phosphate, azosemide, thiamine-pyrophosphate, fosfructose, TG100-115, phlorizin, CL316243, ebrotidine, imidurea, dextrorotation-nimorazole-phosphate-ester, sulfisomidin, VU591, sotagliflozin, NT157, PF-05089771, sulfasalazine, adrafinil
2XEX_A	actinomycin-d, dactinomycin, candicidin, everolimus, sirolimus, deforolimus, rifapentine, echinomycin, amphotericin-b, doramectin, zotarolimus, AMI-1, uridine-5-triphosphate, metafolin, aminopterin, nafamostat, methotrexate, folic-acid, sulfasalazine, pafuramidine, tetrahydrofolic-acid, TW-37, COH29, NT157, famotidine, furamidine, lometrexol, EED226, JNJ-26481585, tedizolid-phosphate
3DIE_A	echinomycin, dactinomycin, deforolimus, sirolimus, rifampin, KB-SRC-4, AZD5991, HER2-Inhibitor-1, GSK2239633A, AMG900, BMS-599626, rutin, AGI-6780, Mps1-IN-1, ticagrelor, TG-100801, fostamatinib, GLPG0187, imatinib, TAK-632, GW-627368, YM-201636, pyronaridine, diosmin, Mps-BAY-2a, rifapentine, OICR-9429, PIK-294, flumatinib, WIKI4
3FF1_B	candicidin, dactinomycin, echinomycin, uridine-5-triphosphate, INS316, hyperin, isoquercitrin, BMS-817378, guadecitabine, myricitrin, reynoutrin, riboflavin-5-phosphate-sodium, citicoline, kuromanin, rutin, baicalin, cromoglicic-acid, famotidine, linagliptin, resmetirom, PDD-00017273, CEP-33779, diosmin, procyanidin-B-2, DOTMP, adenosine-triphosphate, azilsartan-medoxomil, SDZ-220-040, folic-acid, AC-55541
3IM9_A	echinomycin, actinomycin-d, dactinomycin, candicidin, uridine-5-triphosphate, adenosine-triphosphate, epicatechin-gallate-(-), DOTMP, riboflavin-5-phosphate-sodium, INS316, salvianolic-acid-A, PF-573228, sodium-picosulfate, SDZ-220-581, bisindolylmaleimide-IX, LY311727, astilbin, isoquercitrin, avanafil, AMZ30, SDZ-220-040, rutin, T-025, taprenepag, kuromanin, JPH203, triciribine-phosphate, epigallocatechin-gallate-(-), KD025, polyinosine

Continuation of Table 6.1	
Protein	**Ligands**
3IP4_A	candicidin, echinomycin, dactinomycin, actinomycin-d, doramectin, everolimus, ivermectin, deforolimus, rifabutin, nystatin, sennoside-protonated, paromomycin, TD139, guadecitabine, ceftriaxone, CGP-71683, alpha-glucosyl-hesperidin, EB-47, GSK2239633A, ZCL-278, CaMKII-IN-1, bafetinib, ouabain, VER-155008, HER2-Inhibitor-1, LP-533401, diosmin, PF-05089771, imatinib, AMG900
3IP4_B	dactinomycin, TG-100801, candicidin, hypericin, epigallocatechin-gallate-(-), uridine-5-triphosphate, epicatechin-gallate-(-), ciaftalan-zinc, neohesperidin, myricitrin, rutin, paromomycin, LY2090314, KB-SRC-4, kuromanin, neomycin, IPI549, neohesperidin-dihydrochalcone, polyinosine, casanthranol-variant, theaflavin, naringin, isoquercitrin, citicoline, bisantrene, hyperin, idarubicin, 4-galactosyllactose, VX-11e, larotrectinib
3IP4_C	actinomycin-d, dactinomycin, candicidin, INS316, uridine-5-triphosphate, E7449, adenosine-triphosphate, 8-bromo-cGMP, TAK-243, polyinosine, resmetirom, rutin, GS-6201, CVT-10216, diclazuril, folic-acid, altanserin, AMG-337, citicoline, olaparib, AZD2461, epigallocatechin-gallate-(-), INC-280, guadecitabine, elinogrel, dasabuvir, pelanserin, tetrahydrofolic-acid, CFI-402257, famotidine
3KY7_A	candicidin, dactinomycin, actinomycin-d, theaflavin, uridine-5-triphosphate, guadecitabine, INS316, kuromanin, hyperin, riboflavin-5-phosphate-sodium, salvianolic-acid-A, adenosine-triphosphate, nafamostat, methotrexate, myricitrin, CGP-78608, GNTI, isoquercitrin, famotidine, EB-47, TCS-2210, rutin, epicatechin-gallate-(-), citicoline, lometrexol, PDD-00017273, ebrotidine, GSK2126458, 3-MPPI, preladenant
3LVF_O	actinomycin-d, evans-blue, sennoside-protonated, madecassoside, DOTMP, safflower-yellow, adenosine-triphosphate, procyanidin-B-2, GNF-5837, simeprevir, ceftriaxone, ST-2825, sennoside-A, uridine-5-triphosphate, EB-47, inarigivir, alpha-glucosyl-hesperidin, TD139, KB-SRC-4, dioscin, ginsenoside-RE3, AMI-1, ammonium-glycyrrhizinate, CGP-71683, HER2-Inhibitor-1, bisoctrizole, EPZ005687, aclarubicin, TC-S-7004, CaMKII-IN-1
3M9Y_A	actinomycin-d, candicidin, INS316, uridine-5-triphosphate, adenosine-triphosphate, epacadostat, methotrexate, azosemide, epigallocatechin-gallate-(-), epicatechin-gallate-(-), indisulam, procyanidin-B-2, triciribine-phosphate, ACT-132577, myricitrin, cefmenoxime, SDZ-220-040, fludarabine-phosphate, fosphenytoin, fosfructose, LXS196, DOTMP, pevonedistat, famotidine, pyrantel, hyperin, SB-772077B, briciclib, dextrorotation-nimorazole-phosphate-ester, PF-06273340
3SJ7_A	echinomycin, evans-blue, cefsulodin, adenosine-triphosphate, citicoline, baicalin, cefonicid, acarbose, ceftriaxone, EB-47, bekanamycin, CGP-71683, diosmin, alpha-glucosyl-hesperidin, AMI-1, digitoxin, BMS-817378, cefazolin, cefotetan, GLPG0187, elinogrel, GDC-0834, fimasartan, apramycin, EPZ-5676, astilbin, cadazolid, AMG900, CX-5461, epacadostat

Continuation of Table 6.1	
Protein	**Ligands**
3T0T_A	candicidin, dactinomycin, echinomycin, PF-05212384, uridine-5-triphosphate, INS316, epigallocatechin-gallate-(-), rutin, adenosine-triphosphate, citicoline, riboflavin-5-phosphate-sodium, tetrahydrofolic-acid, guadecitabine, kasugamycin, hyperin, DOTMP, diosmin, kuromanin, neohesperidin, ribostamycin, bekanamycin, kanamycin, myricitrin, resmetirom, amygdalin, procyanidin-B-2, ribostamycin-sulfate, acarbose, JNJ-7706621, isoquercitrin
3WQT_A	echinomycin, actinomycin-d, dactinomycin, deforolimus, TPPS4, lurbinectedin, trabectedin, nystatin, uridine-5-triphosphate, rifaximin, hesperidin, INS316, adenosine-triphosphate, eprinomectin, lometrexol, diosmin, evans-blue, A922500, baicalin, cromoglicic-acid, pazopanib, 1,5-dicaffeoylquinic-acid, folic-acid, cefmenoxime, telotristat, cefonicid, aminopterin, LP-533401, pyrintegrin, cefamandole
4C12_A	actinomycin-d, echinomycin, candicidin, deforolimus, eprinomectin, sirolimus, emamectin, dactinomycin, tubocurarine, ciaftalan-zinc, rose-bengal-lactone, INS316, uridine-5-triphosphate, aminopterin, folic-acid, lometrexol, citicoline, adenosine-triphosphate, ebrotidine, tetrahydrofolic-acid, metafolin, GSK3326595, pemetrexed, salazodine, pranlukast, salvianolic-acid-A, cefazolin, TC-S-7004, CA-4948, CGP-71683
4D44_A	actinomycin-d, echinomycin, candicidin, rifabutin, dactinomycin, cefonicid, adenosine-triphosphate, cilofexor, rutin, cot-inhibitor-2, KB-SRC-4, sitaxentan, INS316, SirReal-2, lifitegrast, T-1095, AMG900, R-428, epigallocatechin-gallate-(-), 4EGI-1, guadecitabine, cefmenoxime, triciribine-phosphate, IOWH032, folic-acid, CGP-71683, taprenepag, ceftriaxone, ML277, LY311727
4DG5_A	candicidin, echinomycin, actinomycin-d, dactinomycin, INS316, uridine-5-triphosphate, adenosine-triphosphate, cefonicid, theaflavin, MIW-815, MK-8033, 4EGI-1, cefazolin, naringin, canagliflozin, isepamicin, VU591, fostemsavir, salazodine, ceftriaxone, CGP-71683, ebrotidine, rolitetracycline, cefamandole, fosfructose, pevonedistat, NTNCB, CPI-444, beta-amyloid-synthesis-inhibitor, PF-05089771
4DQ1_A	actinomycin-d, echinomycin, dactinomycin, candicidin, CDBA, amphotericin-b, oligomycin-A, FCE-22250, acarbose, uridine-5-triphosphate, adenosine-triphosphate, diosmin, INS316, sennoside-protonated, AMG900, fostemsavir, cot-inhibitor-1, ceftriaxone, DOTMP, cefozopran, guadecitabine, ticagrelor, KB-SRC-4, fostamatinib, tetrahydrofolic-acid, T-5224, bisoctrizole, cefonicid, paritaprevir, rutin
4DXD_A	dactinomycin, candicidin, eprinomectin, echinomycin, uridine-5-triphosphate, INS316, adenosine-triphosphate, cefonicid, thiamine-pyrophosphate, triciribine-phosphate, EPZ015666, lometrexol, CDBA, NT157, cefmenoxime, folic-acid, cefazolin, briciclib, hPGDS-IN-1, YM-201636, ceftiofur, PF-04217903, cefamandole-nafate, fosfructose, raltitrexed, JNJ-64619178, ceftobiprole, chlorogenic-acid, polydatin, sodium-picosulfate

Continuation of Table 6.1	
Protein	**Ligands**
4DXE_A	echinomycin, PF-05212384, dactinomycin, TG-100801, beta-carotene, nystatin, uridine-5-triphosphate, INS316, adenosine-triphosphate, neomycin, SDZ-220-040, SDZ-220-581, fosfructose, LY2979165, minodronic-acid, DOTMP, epacadostat, XL147, tenofovir, citicoline, SGX523, thiamine-pyrophosphate, riboflavin-5-phosphate-sodium, CS-917, hygromycin-B, ceftiofur, L-690330, CGP-71683, rosmarinic-acid, triciribine-phosphate
4DXE_H	echinomycin, candicidin, dactinomycin, CDBA, sirolimus, astaxanthin, madecassoside, KPT-9274, ivermectin, vinflunine, quizartinib, rifapentine, oligomycin-A, doramectin, amphotericin-b, TFC-007, ammonium-glycyrrhizinate, avatrombopag, MK-4074, selamectin, sennoside-A, GR-127935, nystatin, lurbinectedin, SB-216641, evans-blue, E7449, safflower-yellow, GNF-7, deslanoside
4E4R_A	echinomycin, kuromanin, trabectedin, uridine-5-triphosphate, API-1, casanthranol-variant, INS316, isoquercitrin, rhein, bendroflumethiazide, vipadenant, aztreonam, NBQX, BI-78D3, hypericin, delphinidin, adenosine-triphosphate, 4EGI-1, bisindolylmaleimide-IX, amiloride, hyperin, 5-amino-3-D-ribofuranosylthiazolo[4,5-d]pyrimidin-2,7(3H,6H)-dione, cidofovir, pelanserin, penciclovir, epacadostat, E7449, sardomozide, riboflavin-5-phosphate-sodium, IDO5L
4GCM_A	candicidin, dactinomycin, zotarolimus, echinomycin, amphotericin-b, everolimus, rifabutin, rifapentine, FCE-22250, erythromycin, sirolimus, deforolimus, pimecrolimus, rifaximin, oligomycin-A, ascomycin, AMG900, hesperidin, INS316, guadecitabine, adenosine-triphosphate, HER2-Inhibitor-1, YM-244769, uridine-5-triphosphate, folic-acid, diosmin, glipizide, tacrolimus, 4SC-202, PGL5001
4H8E_A	candicidin, dactinomycin, actinomycin-d, TPPS4, everolimus, doramectin, thiamine-pyrophosphate, adenosine-triphosphate, INS316, GSK2239633A, tedizolid-phosphate, MBX-2982, YM-244769, CGP-71683, uridine-5-triphosphate, XL228, hPGDS-IN-1, TCN201, glipizide, R306465, sodium-picosulfate, fludarabine-phosphate, adenosine-phosphate, masitinib, IOWH032, sitaxentan, CaMKII-IN-1, fosfructose, BAY-87-2243, imatinib
4HLC_A	echinomycin, MIW-815, salvianolic-acid-A, inarigivir, HER2-Inhibitor-1, naringin, uridine-5-triphosphate, INS316, procyanidin-B-2, adenosine-triphosphate, fosaprepitant-dimeglumine, rutin, KB-SRC-4, S49076, WIKI4, acalabrutinib, diosmin, etrasimod, neomycin, cromoglicic-acid, fostamatinib, lometrexol, enzastaurin, IWP-L6, dasabuvir, hypericin, TW-37, GSK256066, nafamostat, grapiprant
4M20_A	uridine-5-triphosphate, epigallocatechin-gallate-(-), procyanidin-B-2, guadecitabine, INS316, cefonicid, naringin, SCH-58261, mangafodipir, bisantrene, sulfatinib, TAK-243, rutin, fostamatinib, DOTMP, etoposide-phosphate, BEBT-908, clindamycin-phosphate, idarubicin, cefazolin, glipizide, ipragliflozin-L-proline, EB-47, hesperidin, pyrantel, SDZ-220-040, epicatechin-gallate-(-), JNJ-64619178, thiamine-pyrophosphate, myricitrin

Continuation of Table 6.1	
Protein	**Ligands**
4NAT_A	echinomycin, TD139, neohesperidin, riboflavin-5-phosphate-sodium, INS316, fostemsavir, VER-155008, XL147, epigallocatechin-gallate-(-), CGP-71683, theaflavin, GSK2126458, rutin, uridine-5-triphosphate, diosmin, naringin, cefonicid, hydroxysafflor-yellow-A, fostamatinib, C188-9, T-5224, cefmenoxime, KD025, MIW-815, ribostamycin, 4-galactosyllactose, BMS-754807, WIKI4, adenosine-triphosphate, IACS-10759
4PEO_A	dactinomycin, echinomycin, actinomycin-d, hypericin, ciaftalan-zinc, INS316, TPPS4, CDBA, E7449, PSB-06126, L-798106, indacaterol, INC-280, uridine-5-triphosphate, riboflavin-5-phosphate-sodium, RWJ-21757, SR-2640, 7-hydroxystaurosporine, avitinib, BAY-61-3606, PD-407824, canagliflozin, N-benzylnaltrindole, R112, LY2784544, CKD-712, candesartan, ML323, PF-02545920, AM-1241
4QAX_A	dactinomycin, actinomycin-d, echinomycin, eprinomectin, rutin, uridine-5-triphosphate, INS316, naringin, sennoside-protonated, procyanidin-B-2, astragaloside-a, proscillaridin-A, salvianolic-acid-A, DOTMP, hesperidin, IPI549, baicalin, neohesperidin, hypericin, presatovir, paromomycin, epicatechin-gallate-(-), ceftriaxone, casanthranol-variant, ipragliflozin-L-proline, riboflavin-5-phosphate-sodium, citicoline, adenosine-triphosphate, licogliflozin, CaMKII-IN-1
4QJU_A	actinomycin-d, dactinomycin, hypericin, paritaprevir, DBPR-211, R-428, icariin, KB-SRC-4, uridine-5-triphosphate, SGI-1027, SR-3306, GSK461364, lifirafenib, nilotinib, PRN1008, adavivint, LX1031, BMS-587101, amphotericin-b, flumatinib, BMS-833923, deltarasin, CFI-402257, GLPG0187, GSK2239633A, tropifexor, GDC-0834, beclabuvir, TW-37, enzastaurin
4QRH_A	actinomycin-d, echinomycin, candicidin, dactinomycin, theaflavin, adenosine-triphosphate, INS316, uridine-5-triphosphate, cefmenoxime, resmetirom, silymarin, sinefungin, citicoline, EB-47, cefonicid, candesartan, thiamine-pyrophosphate, guadecitabine, Ro-5126766, inarigivir, salvianolic-acid-A, GSK2334470, baricitinib, diosmin, hydroxysafflor-yellow-A, fludarabine-phosphate, naringin-dihydrochalcone, adenosine-phosphate, cefamandole, epigallocatechin-gallate-(-)
4RPA_A	echinomycin, dactinomycin, actinomycin-d, uridine-5-triphosphate, riboflavin-5-phosphate-sodium, INS316, adenosine-triphosphate, polyinosine, thiamine-pyrophosphate, E7449, citicoline, kuromanin, baicalin, myricitrin, SDZ-220-040, isoquercitrin, UBP-310, nafamostat, TAK-733, phlorizin, INC-280, ZK-200775, famotidine, UBP-302, furamidine, leteprinim, reynoutrin, irosustat, altanserin, CT7001

Continuation of Table 6.1	
Protein	**Ligands**
4TO8_A	echinomycin, candicidin, dactinomycin, eprinomectin, amphotericin-b, INS316, uridine-5-triphosphate, baicalin, hesperidin, riboflavin-5-phosphate-sodium, minodronic-acid, rutin, hyperin, TAK-243, polyinosine, hygromycin-B, triciribine-phosphate, kuromanin, procyanidin-B-2, canagliflozin, amrubicin, E7449, naringin-dihydrochalcone, famotidine, adenosine-triphosphate, 4-galactosyllactose, CHIR-98014, isoquercitrin, raltitrexed, 2-hydroxy-4-((E)-3-(4-hydroxyphenyl)acryloyl)-2-((2R,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)-6-((2S,3R,4R,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)cyclohexane-1,3,5-trione
4YLY_A	candicidin, digitoxin, INS316, uridine-5-triphosphate, citicoline, guadecitabine, CGP-78608, adenosine-triphosphate, rutin, riboflavin-5-phosphate-sodium, reynoutrin, E7449, ciaftalan-zinc, kuromanin, elinogrel, apigenin, famotidine, baicalin, myricitrin, omeprazole, myricetin, ceftriaxone, 8-bromo-cGMP, kaempferol, hyperin, epacadostat, metafolin, PDD-00017273, isoquercitrin, polyinosine
5BOE_A	candicidin, dactinomycin, echinomycin, madecassoside, digitoxin, doramectin, nystatin, everolimus, sirolimus, eprinomectin, actinomycin-d, ivermectin, emamectin, rifapentine, guanosine, cidofovir, Ro-9187, R-1479, minodronic-acid, HA-1004, fosfructose, famotidine, guanidinoethyldisulfide-bicarbonate, tipiracil, 5-amino-3-D-ribofuranosylthiazolo[4,5-d]pyrimidin-2,7(3H,6H)-dione, citicoline, ellagic-acid, inosine, forodesine, tioguanine
5ETR_A	dactinomycin, actinomycin-d, adenosine-triphosphate, paritaprevir, naringin-dihydrochalcone, rutin, hydroxysafflor-yellow-A, procyanidin-B-2, mangafodipir, TD139, fostemsavir, baicalin, XL228, PF-05089771, triciribine-phosphate, inarigivir, acarbose, CUDC-907, uridine-5-triphosphate, epacadostat, INS316, isepamicin, PK-44, AMG900, sennoside-protonated, etoposide-phosphate, pilaralisib, m-THP, paromomycin, riboflavin-5-phosphate-sodium
5JIC_A	actinomycin-d, candicidin, zotarolimus, everolimus, dactinomycin, sirolimus, doramectin, TPPS4, echinomycin, INS316, uridine-5-triphosphate, adenosine-triphosphate, cefonicid, indisulam, radotinib, XL147, cefazolin, ebrotidine, aminopterin, fananserin, ticagrelor, neohesperidin-dihydrochalcone, SirReal-2, epacadostat, cefmenoxime, EB-47, IACS-10759, IOWH032, BMS-817378, SC-9
5KDR_A	candicidin, actinomycin-d, dactinomycin, echinomycin, TPPS4, PF-05212384, CDBA, zotarolimus, triciribine-phosphate, INS316, everolimus, sirolimus, AZD1480, baicalin, neohesperidin, uridine-5-triphosphate, PIK-75, adenosine-triphosphate, astilbin, APY-29, fosfructose, naringin, XL147, pazopanib, EB-47, myricitrin, avanafil, epigallocatechin-gallate-(-), kuromanin, polyinosine
5KDR_B	dactinomycin, actinomycin-d, echinomycin, candicidin, TPPS4, CDBA, everolimus, zotarolimus, deforolimus, PF-05212384, safflower-yellow, eprinomectin, sirolimus, ivermectin, astaxanthin, MK-4074, doramectin, emamectin, FCE-22250, rifapentine, abamectin, nystatin, TFC-007, m-THP, madecassoside, grazoprevir, evans-blue, deslanoside, rifaximin, GR-127935

Continuation of Table 6.1	
Protein	**Ligands**
5N9M_A	echinomycin, dactinomycin, candicidin, INS316, uridine-5-triphosphate, triciribine-phosphate, E7449, indisulam, chlorthalidone, ellagic-acid, adenosine-triphosphate, dorzolamide, astilbin, epacadostat, sitaxentan, WAY-316606, tenofovir, sulphadimethoxine, TAME, PK-44, trichlormethiazide, lomeguatrib, ipragliflozin-L-proline, tiludronate, idelalisib, methyclothiazide, risedronate, licogliflozin, 4EGI-1, 8-bromo-cAMP
5VZ2_A	candicidin, dactinomycin, CDBA, EW-7197, BLZ945, CT7001, AGI-6780, EMD-66684, XL228, ENMD-2076, PF-06409577, SR-3306, zidovudine, alovudine, XL147, AMI-1, ETC-159, quercetin, CGP-78608, E7449, R-428, baicalin, delphinidin, KHS-101, beta-amyloid-synthesis-inhibitor, DBeQ, risdiplam, INC-280, RX-3117, rutin
5X1X_A	candicidin, echinomycin, dactinomycin, tenalisib, SDZ-220-040, TAK-243, hypericin, 5-amino-3-D-ribofuranosylthiazolo[4,5-d]pyrimidin-2,7(3H,6H)-dione, astilbin, PDE10-IN-1, guanosine, riboflavin-5-phosphate-sodium, casanthranol-variant, CEP-33779, baicalin, reynoutrin, kuromanin, TG100-115, acalisib, candesartan, rutin, XL147, peficitinib, BAY-61-3606, riboflavin, uridine-5-triphosphate, penciclovir, A-839977, AR-C155858, BMS-986142
5XZ7_A	echinomycin, actinomycin-d, ivermectin, adenosine-triphosphate, uridine-5-triphosphate, INS316, triciribine-phosphate, pyrintegrin, baicalin, riboflavin-5-phosphate-sodium, EB-47, guadecitabine, inarigivir, TAK-243, naringin, thiamine-pyrophosphate, zidovudine, fostamatinib, E7449, JNJ-64619178, pazopanib, GSK2239633A, famotidine, pemetrexed, sulfatinib, INC-280, amygdalin, MK-8033, PF-562271, paromomycin
5ZH8_A	echinomycin, adenosine-triphosphate, uridine-5-triphosphate, triciribine-phosphate, LY311727, hesperidin, hypericin, epacadostat, cefmenoxime, INS316, doripenem, fludarabine-phosphate, FG-4592, SDZ-220-040, adenosine-phosphate, betamethasone-phosphate, polyinosine, ceftiofur, ceftobiprole, MK-8245, rutin, salvianolic-acid-A, cot-inhibitor-1, GDC-0980, SDZ-220-581, AZ-7371, hydrocortisone-phosphate, naringin, thiamine-pyrophosphate, etoposide-phosphate
5ZNJ_A	echinomycin, dactinomycin, candicidin, actinomycin-d, eprinomectin, evans-blue, INS316, guadecitabine, isoquercitrin, uridine-5-triphosphate, inarigivir, adenosine-triphosphate, rutin, procyanidin-B-2, kuromanin, daidzin, myricitrin, tucatinib, acarbose, epigallocatechin-gallate-(-), citicoline, TWS-119, neohesperidin, salvianolic-acid-A, naringin-dihydrochalcone, BMS-986142, pyrintegrin, naringin, reynoutrin, PSB-603
6CLV_A	actinomycin-d, dactinomycin, adenosine-triphosphate, naringin, alpha-glucosyl-hesperidin, HER2-Inhibitor-1, methotrexate, CHIR-98014, naringin-dihydrochalcone, salazodine, tetrahydrofolic-acid, diosmin, PF-06273340, hesperidin, PF-562271, AMG900, KD025, BEBT-908, EB-47, glipizide, inarigivir, MGCD-265, aminopterin, tucatinib, folic-acid, triciribine-phosphate, epacadostat, adavivint, TC-G-1008, ebrotidine

Continuation of Table 6.1	
Protein	**Ligands**
6D1R_A	actinomycin-d, dactinomycin, echinomycin, PSB-06126, etoposide-phosphate, fenoverine, sulfatinib, enzastaurin, CaMKII-IN-1, ertugliflozin, ICA-121431, C188-9, LB42708, SRT1720, SRT2104, LY2784544, YM-022, GSK2239633A, GSK2126458, zosuquidar, PF-573228, BMS-779788, pilaralisib, KB-SRC-4, methotrexate, CGP-71683, pazopanib, tivantinib, G-749, purmorphamine
6G15_A	adenosine-triphosphate, uridine-5-triphosphate, INS316, moxalactam, famotidine, cefonicid, sodium-picosulfate, rutin, CGP-71683, T-1095, HER2-Inhibitor-1, LY311727, L-694247, L-798106, XL147, paliperidone, PF-573228, fludarabine-phosphate, triciribine-phosphate, AMZ30, adenosine-phosphate, ceftiofur, thiamine-pyrophosphate, baricitinib, riboflavin-5-phosphate-sodium, TPPS4, KD025, NT157, folic-acid, cot-inhibitor-1
6GYZ_A	candicidin, dactinomycin, actinomycin-d, echinomycin, uridine-5-triphosphate, guadecitabine, adenosine-triphosphate, riboflavin-5-phosphate-sodium, mangafodipir, ceftriaxone, INS316, cefonicid, triciribine-phosphate, inarigivir, linagliptin, epicatechin-gallate-(-), AMI-1, E7449, resmetirom, cefmenoxime, moxalactam, pyrintegrin, cefamandole, AMG900, CGP-71683, PDD-00017273, silibinin, integrin-antagonist-1, ouabain, acarbose
6H5E_B	dactinomycin, echinomycin, uridine-5-triphosphate, actinomycin-d, INS316, hPGDS-IN-1, guadecitabine, pemetrexed, CGP-71683, adenosine-triphosphate, famotidine, citicoline, sulfatinib, baricitinib, lometrexol, PF-04217903, sitaxentan, telatinib, GS-143, TCN201, baicalin, Ro-5126766, aztreonam, AMG-337, nilotinib, XL228, ebrotidine, candesartan, nemiralisib, PGL5001
6NDL_A	dactinomycin, candicidin, adenosine-triphosphate, INS316, S-3304, neohesperidin, SR-3306, aminopterin, dioscin, guadecitabine, rutin, metafolin, uridine-5-triphosphate, BMS-986020, masitinib, salazodine, LY2874455, sulfasalazine, folic-acid, WZ-3146, VER-155008, avitinib, 4SC-202, citicoline, evocalcet, MBX-2982, salvianolic-acid-A, radezolid, YM-244769, lometrexol
6PBO_X	guadecitabine, AMI-1, adenosine-triphosphate, KB-SRC-4, diosmin, CFI-400945, TD139, SR-3306, TG-100801, INS316, CUDC-907, YM-201636, IOWH032, AMG900, R-428, Mps1-IN-1, PRN1008, neohesperidin-dihydrochalcone, EB-47, ML193, ZCL-278, BMS-599626, LX1031, S-3304, 4SC-202, TC-S-7004, eltrombopag, HER2-Inhibitor-1, IACS-10759, adavivint
6R1N_A	dactinomycin, candicidin, actinomycin-d, INS316, uridine-5-triphosphate, guadecitabine, citicoline, HER2-Inhibitor-1, rutin, adenosine-triphosphate, ceftriaxone, CHIR-98014, naringin-dihydrochalcone, lometrexol, BMS-817378, naringin, neomycin, neohesperidin, tetrahydrofolic-acid, pyrintegrin, folic-acid, riboflavin-5-phosphate-sodium, cefsulodin, dibekacin, cilofexor, reynoutrin, hydroxysafflor-yellow-A, resmetirom, daidzin, fostemsavir

Continuation of Table 6.1	
Protein	**Ligands**
6RK3_A	candicidin, procyanidin-B-2, rutin, guadecitabine, CaMKII-IN-1, INS316, KB-SRC-4, AMG900, metafolin, myricitrin, neohesperidin, fostamatinib, CGP-71683, TG-101209, 2-hydroxy-4-((E)-3-(4-hydroxyphenyl)acryloyl)-2-((2R,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)-6-((2S,3R,4R,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)cyclohexane-1,3,5-trione, linsitinib, TAK-243, PHA-665752, pazopanib, theaflavin, VER-155008, cefpirome, reynoutrin, kuromanin, AMI-1, folic-acid, adavivint, dihydroergotamine, MM-206, tetrahydrofolic-acid

Table 6.2: List of proteins and 30 potential ligands. target hit from the BLAST alignments. The ligands in the second column were 30 top-ranked compounds of each target after applying the model 6.1 to the MRSA docking dataset. These ligands were listed in descending order in terms of predicted binding affinity to their corresponding protein target. These compounds were obtained from Repurposing Hub (Corsello et al., 2017)

Protein	**Ligands**
DEG10170002	vericiguat, SNS-314, APY-29, AZD1480, GPi-688, VLX600, tasisulam, phthalylsulfathiazole, IB-MECA, L-phenylisopropyladenosine, nelociguat, CH-5183284, CQS, imidurea, merbarone, fludarabine-phosphate, TCS-359, XL228, beta-amyloid-synthesis-inhibitor, exo-IWR-1, JTE-013, R112, TC-G-1008, yoda-1, phthalylsulfacetamide, PIK-75, BAM7, BML-284, torasemide, dynasore
DEG10170006	sennoside-protonated, adenosine-triphosphate, guadecitabine, sennoside-A, PF-562271, tigecycline, MIW-815, PF-431396, myricitrin, PF-573228, CHIR-98014, fostamatinib, metafolin, alpha-glucosyl-hesperidin, fostemsavir, eravacycline, TG-101209, BMS-599626, procyanidin-B-2, acarbose, adavivint, inarigivir, KB-SRC-4, naringin-dihydrochalcone, rutin, tetrahydrofolic-acid, tradipitant, paritaprevir, cefonicid, cilengitide
DEG10170012	danirixin, alendronate, pamidronate, minodronic-acid, zoledronic-acid, azaguanine-8, neridronic-acid, NK-252, sulfasalazine, imidurea, famotidine, etidronic-acid, risedronate, guanidinoethyldisulfide-bicarbonate, E7449, valaciclovir, PIK-93, pevonedistat, AZ-10417808, tiludronate, adenosine-triphosphate, ibandronate, zanamivir, XL228, AZD1480, HA-1004, nifursol, TH-302, cefazolin, amiloride
DEG10170015	uridine-5-triphosphate, INS316, cefmenoxime, G007-LK, lometrexol, PF-573228, cefoselis, metafolin, fostamatinib, imidurea, famotidine, NS-3623, ceftriaxone, adenosine-triphosphate, fosaprepitant-dimeglumine, ceftiofur, pemetrexed, AZD1480, cefozopran, AMG900, tozasertib, bisantrene, PRT062607, benzthiazide, BMS-754807, KD025, CCT196969, R406, DSR-6434, AZ-10417808
DEG10170020	XL147, epacadostat, APY-29, triciribine-phosphate, CEP-32496, BW-348U87, amuvatinib, TC-G-1008, cefazolin, denotivir, uridine-5-triphosphate, A-804598, CP-471474, indisulam, NVP-BHG712, RAF265, guadecitabine, IMREG-1, PF-03758309, VE-821, zibotentan, PF-573228, SCH-58261, CH-5183284, sal003, MLN2480, crizotinib-(S), regadenoson, R406, tiotidine

Continuation of Table 6.2	
Protein	**Ligands**
DEG10170022	CDBA, madecassoside, actinomycin-d, echinomycin, dactinomycin, safflower-yellow, sirolimus, nystatin, deforolimus, amphotericin-b, candicidin, dioscin, ascomycin, everolimus, lurbinectedin, oligomycin-A, sennoside-protonated, sennoside-A, vinblastine, eprinomectin, omeprazole-magnesium, emamectin, rifaximin, doramectin, FCE-22250, m-THP, vinorelbine, astragaloside-a, ammonium-glycyrrhizinate, theaflavin
DEG10170026	INS316, XL147, PF-562271, AZD3264, cefmenoxime, uridine-5-triphosphate, defactinib, PF-431396, KD025, PF-573228, salvianolic-acid-A, adenosine-triphosphate, ceftriaxone, pimodivir, GSK2126458, BAY-1251152, atuveciclib, cefazolin, cefotetan, CZC-54252, sulfatinib, epicatechin-gallate-(-), PTC-209, BMS-986158, epigallocatechin-gallate-(-), imidurea, enasidenib, ribostamycin-sulfate, myricitrin, NS-11021
DEG10170027	epacadostat, CGP-78608, minodronic-acid, chlorthalidone, DNQX, rimeporide, zoledronic-acid, guanidinoethyldisulfide-bicarbonate, 5-amino-3-D-ribofuranosylthiazolo[4,5-d]pyrimidin-2,7(3H,6H)-dione, risedronate, HA-1004, hydroflumethiazide, acetazolamide, alendronate, NG-nitro-arginine, famotidine, pamidronate, zanamivir, neridronic-acid, azathioprine, eniporide, E7820, guanosine, sparfosate, azosemide, CGP-57380, IDO5L, sulfaguanidine, NK-252, taurolidine
DEG10170035	telatinib, ripretinib, KD025, azosemide, HTH-01-015, INS316, tucatinib, APY-29, SD-208, ceftriaxone, cefazolin, AZ960, NS-11021, riboflavin-5-phosphate-sodium, ravoxertinib, XL147, LY3214996, GSK3179106, CFI-402257, CW-008, TC-G-1008, CC-930, fostamatinib, pimodivir, SNS-314, CF102, famotidine, nafamostat, pazopanib, PF-573228
DEG10170039	dactinomycin, actinomycin-d, candicidin, GTP-14564, deforolimus, pyrazolanthrone, echinomycin, 6-aminochrysene, zotarolimus, CDBA, everolimus, sasapyrine, sirolimus, semaxanib, sanguinarium-chloride, E7449, STF-083010, nystatin, vadadustat, dithranol, 3-bromo-7-nitroindazole, amlexanox, CGS-15943, necrostatin-2, PIT, PP242, tanshinone-IIA, (R)-(-)-apomorphine, CKD-712, hydroxytacrine-maleate-(R,S)
DEG10170041	MLN2480, pilaralisib, NS-11021, adenosine-triphosphate, ceftriaxone, CHIR-98014, PRT062607, uridine-5-triphosphate, cefmenoxime, CW-008, PF-05089771, INS316, KG-5, imidurea, losartan, epacadostat, guadecitabine, ZCL-278, CGP-71683, R547, ZD-7155, gliquidone, opicapone, ceftiofur, SPP301, phthalylsulfathiazole, SNS-314, candesartan, tigecycline, triciribine-phosphate
DEG10170043	CHIR-98014, INS316, ZCL-278, XL147, N-(2-chlorophenyl)-2-((2E)-2-[1-(2-pyridinyl)ethylidene]hydrazinocarbothioyl)hydrazinecarbothioamide, CHIR-99021, uridine-5-triphosphate, cefmenoxime, epacadostat, pimodivir, SNS-314, LY2784544, NS-11021, rebastinib, AGI-6780, APY-29, bisantrene, Ro-4987655, BAY-61-3606, CCT196969, JTE-013, NS-3623, R547, adenosine-triphosphate, ebrotidine, selonsertib, TC-G-1008, triciribine-phosphate, azosemide, CGP-71683

Continuation of Table 6.2	
Protein	**Ligands**
DEG10170045	AZD4635, LXR-623, abafungin, LY2157299, ciaftalan-zinc, lesinurad, TG100-115, AH-7614, epacadostat, PIK-293, FR-180204, isoxicam, sulfachlorpyridazine, CCG-63802, KD025, BTT-3033, LIMKi-3, ML281, CQS, GW-438014A, SB-2343, tivantinib, CFM-2, hydroxyfasudil, SKF-86002, tenalisib, XL147, 8-bromo-cGMP, SB-415286, benzamil
DEG10170049	KPT-9274, flumatinib, GNF-5837, KB-SRC-4, pilaralisib, AMI-1, fostamatinib, LY2801653, tucatinib, BT-11, BTT-3033, R-428, NVP-BHG712, SNX-5422, GDC-0834, DBPR-211, Mps-BAY-2a, VX-11e, elinogrel, TC-S-7004, ITI214, masitinib, HER2-Inhibitor-1, TD139, adavivint, PF-573228, radotinib, entrectinib, IACS-10759, KRCA-0008
DEG10170052	MRK-409, MLN0128, L-838417, PP-121, AZD1480, EW-7197, APY-29, BIBX-1382, FR-180204, tiotidine, talniflumate, PSB-06126, SB-525334, HA-1004, TG100-115, vorasidenib, epacadostat, voxtalisib, CHR-6494, niflumic-acid, IDO5L, lomeguatrib, cariporide, famotidine, iclaprim, CS-917, AZD6738, TA-01, zaprinast, SD-208
DEG10170070	epacadostat, famotidine, benzamil, azathioprine, sulfaquinoxaline, AG-490, minodronic-acid, EED226, tiotidine, PRT062607, TC-S-7009, sangivamycin, trabodenoson, cefdinir, TAME, althiazide, benzthiazide, phthalylsulfathiazole, adenosine-phosphate, sulfamethizole, CQS, succinylsulfathiazole, sulfadoxine, INS316, sulfameter, trichlormethiazide, Ro-61-8048, HA-1004, cyclopenthiazide, cefixime
DEG10170071	micronomicin, azosemide, PRT062607, dibekacin, uridine-5-triphosphate, epacadostat, ceftriaxone, R-428, ceftiofur, cefazolin, kanamycin, cefozopran, PF-05089771, cefpirome, PF-562271, mangafodipir, adenosine-triphosphate, tivantinib, bekanamycin, PSB-603, CPI-444, doripenem, GSK9027, ARRY-334543, sotrastaurin, cefuroxime, HER2-Inhibitor-1, sisomicin, triciribine-phosphate, darglitazone
DEG10170072	eravacycline, adenosine-triphosphate, INS316, vemurafenib, guadecitabine, bisindolylmaleimide-IX, edaglitazone, fostamatinib, cefazolin, VX-11e, L-368899, avitinib, MIW-815, bekanamycin, neomycin, tigecycline, acarbose, AMG900, PF-573228, cefotetan, epacadostat, neohesperidin, R-428, TP-0903, CEP-32496, lifitegrast, NMS-1286937, cefonicid, paromomycin, azilsartan-medoxomil
DEG10170074	LY393558, guadecitabine, ciaftalan-zinc, HER2-Inhibitor-1, PLX8394, edoxaban, cefonicid, EMD-66684, PDD-00017273, alpha-glucosyl-hesperidin, KB-SRC-4, uridine-5-triphosphate, DOTMP, MIW-815, XL228, fostemsavir, EB-47, rutin, CaMKII-IN-1, PF-573228, zoliflodacin, JW-74, AMG900, CHIR-98014, NS-11021, PF-562271, apramycin, GNF-7, BMS-986142, INS316
DEG10170084	cefotetan, fostemsavir, adenosine-triphosphate, folic-acid, cefmenoxime, INS316, PF-05089771, PF-03814735, cefmetazole, uridine-5-triphosphate, epacadostat, PF-573228, tetrahydrofolic-acid, CHIR-98014, ceftiofur, FN-1501, GNTI, NS-11021, ebrotidine, BEBT-908, dabrafenib, famotidine, cefamandole, guanidinoethyldisulfide-bicarbonate, cefozopran, trabodenoson, minodronic-acid, TC-G-1008, AMG900, benzamil

Continuation of Table 6.2	
Protein	**Ligands**
DEG10170086	candicidin, echinomycin, INS316, uridine-5-triphosphate, epacadostat, TC-S-7009, famotidine, pazopanib, LY2603618, HSR6071, tiotidine, adenosine-triphosphate, chlorthalidone, CHR-6494, triciribine-phosphate, acetazolamide, taurolidine, UBP-302, regadenoson, ebrotidine, TC-G-1008, adenosine-phosphate, benserazide, PF-573228, UBP-310, fludarabine-phosphate, PIK-93, azosemide, NS-5806, CS-917
DEG10170092	pamidronate, zoledronic-acid, risedronate, clodronic-acid, alendronate, guanidinoethyldisulfide-bicarbonate, minodronic-acid, isoxicam, etidronic-acid, neridronic-acid, epacadostat, kinetin, zaltidine, 1-phenylbiguanide, lidamidine, tiludronate, sardomozide, azaguanine-8, hydrochlorothiazide, LTA, amiloride, 6-benzylaminopurine, kifunensine, ditolylguanidine, m-chlorophenylbiguanide, phenformin, sulfamonomethoxine, butalbital, lamotrigine, sulfaguanidine
DEG10170101	adenosine-triphosphate, GSK256066, BEBT-908, guadecitabine, CaMKII-IN-1, IPI549, ceftriaxone, ceftiofur, XL228, nafamostat, EB-47, imidurea, icariin, elinogrel, epacadostat, pilaralisib, vemurafenib, VU0364439, EW-7197, ZM-241385, bisantrene, CUDC-907, PF-573228, XL147, folic-acid, cefazolin, aztreonam, ticagrelor, C188-9, cefpirome
DEG10170102	PF-562271, bisantrene, ceftriaxone, JNJ-7706621, GNTI, PRT062607, ciaftalan-zinc, GSK1838705A, PF-431396, hypericin, lometrexol, CZC-54252, tucatinib, trilaciclib, adenosine-triphosphate, NMS-1286937, tetrahydrofolic-acid, PF-573228, uridine-5-triphosphate, cefazolin, bisindolylmaleimide-IX, BMS-754807, radotinib, CW-008, INS316, 2-hydroxy-4-((E)-3-(4-hydroxyphenyl)acryloyl)-2-((2R,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)-6-((2S,3R,4R,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)cyclohexane-1,3,5-trione, A-839977, KD025, XL228, EB-47
DEG10170103	cefmenoxime, I-BRD9, epacadostat, BAY-1251152, azosemide, TH-302, AMG900, JNJ-7706621, vatalanib, BMS-754807, telatinib, tivantinib, vericiguat, EED226, FR-180204, NS-5806, SirReal-2, XL147, cefazolin, torasemide, tropifexor, 4EGI-1, ARRY-334543, Ro-3306, MM-206, Ro-5126766, XL228, defactinib, NT157, talniflumate
DEG10170113	tucatinib, purmorphamine, ARRY-334543, dabrafenib, AST-1306, cefmenoxime, TC-G-1008, JNJ-64619178, APY-29, enasidenib, INC-280, HER2-Inhibitor-1, 4EGI-1, avitinib, ceftiofur, guadecitabine, PF-573228, AMG900, BMS-779788, cefazolin, COH29, presatovir, T-025, JPH203, seletalisib, TAK-243, EW-7197, talmapimod, CID-2745687, salvianolic-acid-A
DEG10170121	HER2-Inhibitor-1, INS316, baricitinib, adenosine-triphosphate, ceftiofur, ertapenem, MIW-815, cefotetan, DOTMP, enasidenib, AMZ30, bisindolylmaleimide-IX, SNS-314, BMS-214662, CHIR-98014, epicatechin-gallate-(-), cefazolin, guadecitabine, imidurea, BMS-587101, tucatinib, AC-55541, PF-431396, uridine-5-triphosphate, cefmenoxime, dabrafenib, T-5224, EB-47, acarbose, GNTI

Continuation of Table 6.2	
Protein	**Ligands**
DEG10170128	MIW-815, paromomycin, cefotetan, fostamatinib, DOTMP, inarigivir, guadecitabine, ceftobiprole, neomycin, INS316, BMS-587101, tetrahydrofolic-acid, cefmenoxime, isepamicin, AMG900, radotinib, myricitrin, epacadostat, zoliflodacin, 8-bromo-cGMP, ceftiofur, pilaralisib, cromoglicic-acid, dabrafenib, PRT062607, adenosine-triphosphate, azilsartan-medoxomil, cefonicid, nilotinib, SRT1720
DEG10170139	iproniazid, FN-1501, PF-05089771, HER2-Inhibitor-1, CK-101, PF-573228, nafamostat, DOTMP, CB-5083, tiotidine, TCN201, PRT062607, purvalanol-B, CEP-33779, tucatinib, abafungin, epacadostat, guadecitabine, LY3009120, BMS-754807, INC-280, IPI549, pilaralisib, AZD3264, LY3295668, mocetinostat, uridine-5-triphosphate, APY-29, micronomicin, adenosine-triphosphate
DEG10170142	NVP-BHG712, nilotinib, GLPG0187, PF-573228, cefonicid, PRT062607, PF-562271, gliquidone, PF-05089771, tucatinib, aminopterin, cot-inhibitor-2, R406, guadecitabine, PSB-06126, bisindolylmaleimide-IX, TG-101209, NS-5806, GSK256066, SRT1720, HTH-01-015, oglemilast, WIKI4, AMG900, azosemide, tradipitant, BAY-61-3606, CUDC-907, SC-51089, AV-608
DEG10170153	cefmenoxime, nifursol, cefazolin, PLX8394, bisantrene, cefoselis, DOTMP, cilengitide, defactinib, epacadostat, tradipitant, uridine-5-triphosphate, adenosine-triphosphate, fostamatinib, fostemsavir, famotidine, GSK256066, cefonicid, dibekacin, metafolin, folic-acid, ceftriaxone, hypericin, pimodivir, rutin, ZCL-278, AZD3264, PF-06273340, cefmetazole, betamethasone-phosphate
DEG10170156	PF-573228, aminopterin, folic-acid, methotrexate, neomycin, ML786, ammonium-glycyrrhizinate, ceftriaxone, fostamatinib, PF-562271, imidurea, rebastinib, sulfatinib, adenosine-triphosphate, guadecitabine, metafolin, AMG900, nilotinib, GNF-5837, cefmenoxime, EB-47, GLPG0187, GNTI, hesperidin, KD025, PF-431396, ceftobiprole, ebrotidine, PF-05089771, diosmin
DEG10170158	adenosine-triphosphate, aminopterin, folic-acid, guadecitabine, ML193, tucatinib, pradefovir, BX-912, INS316, uridine-5-triphosphate, epacadostat, lometrexol, tiotidine, EB-47, cefozopran, VX-11e, tetrahydrofolic-acid, CUDC-907, TAK-243, BEBT-908, ceftobiprole, methotrexate, cilofexor, ceftriaxone, pevonedistat, HER2-Inhibitor-1, MK-8033, AMG-517, BMS-626529, cefmenoxime
DEG10170161	PF-573228, BEBT-908, XL147, PF-562271, cefmenoxime, cefoselis, BMS-754807, folic-acid, cefozopran, ceftriaxone, cefixime, SKLB-1028, BT-11, adenosine-triphosphate, TAK-632, aminopterin, EB-47, SRT3190, bisantrene, eltrombopag, avitinib, CUDC-907, guadecitabine, HER2-Inhibitor-1, defactinib, ceftiofur, inarigivir, avapritinib, telotristat, cefepime
DEG10170165	epacadostat, INS316, cefazolin, cefotetan, tradipitant, nifursol, cefmenoxime, Ro-4987655, azosemide, TAK-632, SB-415286, cefmetazole, DOTMP, PF-04217903, 4EGI-1, BMS-582949, A-839977, PF-573228, guadecitabine, bisindolylmaleimide-IX, beclabuvir, PLX8394, ceftobiprole, XL147, BMY-45778, PRT062607, ertapenem, ML193, VX-702, benzthiazide

Continuation of Table 6.2	
Protein	**Ligands**
DEG10170169	everolimus, deforolimus, tiotidine, HA-1004, rifapentine, abamectin, deslanoside, ivermectin, TAME, rifabutin, actinomycin-d, LMK-235, ammonium-glycyrrhizinate, amphotericin-b, CDBA, sirolimus, zotarolimus, sardomozide, L-arginine, licostinel, didox, vinorelbine, oligomycin-A, vinblastine, guanidinoethyldisulfide-bicarbonate, zeatin, glecaprevir, zaltidine, doramectin, APY-29
DEG10170172	cefonicid, uridine-5-triphosphate, tradipitant, PK-44, ceftriaxone, epacadostat, telatinib, AMG900, ceftiofur, XL228, cefotetan, TD139, AC-55541, methacycline, PF-05089771, PF-573228, JW-74, azosemide, MRK-560, GSK2239633A, vapendavir, XL147, cefazolin, BI-78D3, CID-16020046, PF-04217903, apafant, cefmenoxime, benzthiazide, ceftobiprole
DEG10170174	rutin, acarbose, isoquercitrin, adenosine-triphosphate, uridine-5-triphosphate, neomycin, bisindolylmaleimide-IX, hyperin, EB-47, zoliflodacin, metafolin, AZD3264, pilaralisib, tetrahydrofolic-acid, FN-1501, kasugamycin, NS-11021, sotrastaurin, HER2-Inhibitor-1, reynoutrin, CHIR-98014, go-6983, myricitrin, ebrotidine, methotrexate, famotidine, TCS-21311, aminopterin, APY-29, cilengitide
DEG10170188	G-749, ACTB-1003, CHIR-99021, Mps-BAY-2a, R406, elinogrel, PLX8394, AMG900, cot-inhibitor-2, flumatinib, SGI-1027, PF-562271, ML193, rebastinib, AMI-1, LY2801653, BMS-833923, epacadostat, INS316, VX-11e, acalabrutinib, VE-822, KPT-9274, NVP-TAE684, WZ-4002, KG-5, HER2-Inhibitor-1, bromosporine, pranlukast, adavivint
DEG10170194	cot-inhibitor-1, SDZ-220-040, epacadostat, EW-7197, ebrotidine, AMG900, SDZ-220-581, TAK-243, aztreonam, protirelin, adenosine-triphosphate, NVP-BHG712, bendroflumethiazide, BMS-754807, tucatinib, benzamil, bisantrene, NS-11021, VX-11e, selinexor, nilotinib, CCT196969, R-428, Mps-BAY-2a, JTE-013, GSK1838705A, SNS-314, tiotidine, AR-12, NS-3623
DEG10170197	radotinib, adenosine-triphosphate, elinogrel, AMI-1, JNJ-64619178, mocetinostat, tetrahydrofolic-acid, PF-562271, SR-3306, NS-11021, ebrotidine, nafamostat, pazopanib, HER2-Inhibitor-1, taltirelin, CZC-54252, lifirafenib, pemetrexed, PF-477736, ARRY-334543, KG-5, uridine-5-triphosphate, PLX8394, INS316, avatrombopag, L-phenylisopropyladenosine, BAY-1251152, CFI-402257, R-428, LY2857785
DEG10170200	HER2-Inhibitor-1, tetrahydrofolic-acid, PF-573228, paromomycin, ceftobiprole, DOTMP, guadecitabine, PRT062607, epacadostat, neomycin, AMG900, ebrotidine, uridine-5-triphosphate, cefonicid, AZD3264, fostamatinib, ML786, hygromycin-B, inarigivir, kanamycin, metafolin, rutin, naringin-dihydrochalcone, adavivint, INS316, PF-562271, cefotetan, folic-acid, cilengitide, tobramycin
DEG10170201	ciaftalan-zinc, INCB-057643, SCH-58261, SDZ-220-040, naltriben, sotrastaurin, MM-206, SDZ-220-581, epacadostat, NS-11021, uridine-5-triphosphate, bumetanide, KAF-156, LY2090314, XL228, indisulam, TAK-659, GF109203X, minodronic-acid, perfluorodecalin, vatinoxan, nafamostat, piretanide, pyrantel, SB-415286, tenalisib, famotidine, INS316, azosemide, IKK-2-inhibitor-V

Continuation of Table 6.2	
Protein	**Ligands**
DEG10170204	NS-11021, JTE-013, EW-7197, pimodivir, imidurea, CT-7758, kuromanin, XL228, INS316, N-(2-chlorophenyl)-2-((2E)-2-[1-(2-pyridinyl)ethylidene]hydrazinocarbothioyl)hydrazinecarbothioamide, BX-912, PF-477736, uridine-5-triphosphate, CGP-71683, SRT3190, CGP-53353, epacadostat, PLX4720, danirixin, enasidenib, C188-9, glipizide, adenosine-triphosphate, GSK1838705A, CT7001, AGI-6780, E7046, PF-03758309, PF-573228, ZCL-278
DEG10170206	DOTMP, PLX8394, bisindolylmaleimide-IX, sotrastaurin, metafolin, fostamatinib, ceftobiprole, HER2-Inhibitor-1, imidurea, NVP-AUY922, TG-100801, guadecitabine, INS316, SU11274, bisantrene, PF-573228, hypericin, icariin, TD139, alpha-glucosyl-hesperidin, raltitrexed, CEP-37440, adavivint, casanthranol-variant, MK-8245, NS-5806, ceftiofur, XL228, CHIR-98014, radotinib
DEG10170210	adenosine-triphosphate, epacadostat, TC-S-7004, aminopterin, uridine-5-triphosphate, EB-47, AZD2858, INS316, SB-334867, indisulam, fludarabine-phosphate, R547, imidurea, U-104, AGI-6780, AMG-925, HSR6071, I-BRD9, JNJ-64619178, VU591, TAK-243, inarigivir, IDO5L, methotrexate, puromycin, edoxaban, tedizolid-phosphate, PF-573228, COH29, cyclic-AMP
DEG10170214	folic-acid, adenosine-triphosphate, epacadostat, defactinib, cefetamet, ebrotidine, eniporide, CHIR-98014, DOTMP, famotidine, cefonicid, MK-8245, fostamatinib, rimeporide, TAK-243, cefoselis, PF-562271, ZCL-278, INS316, cefmenoxime, HA-1004, PF-573228, trabodenoson, HER2-Inhibitor-1, ceftizoxim, FR-180204, uridine-5-triphosphate, brivanib-alaninate, SNS-314, protirelin
DEG10170221	HER2-Inhibitor-1, cefotetan, neomycin, INS316, avitinib, apramycin, ceftriaxone, GSK256066, CHIR-98014, uridine-5-triphosphate, guadecitabine, XL228, adenosine-triphosphate, cefonicid, isepamicin, defactinib, DOTMP, ebrotidine, imidurea, paromomycin, bekanamycin, cilengitide, epacadostat, PF-573228, cefmenoxime, PF-562271, rutin, PRT062070, ceftobiprole, GNTI
DEG10170222	AMG900, KB-SRC-4, nilotinib, sulfatinib, NVP-BHG712, olmutinib, BMS-833923, MIW-815, CEP-37440, conivaptan, R-428, imatinib, CGP-71683, entrectinib, radotinib, GSK2126458, pranlukast, tucatinib, AMG-PERK-44, CUDC-907, LY2090314, SR-3306, zosuquidar, AZD3264, LY2801653, TC-S-7003, ZCL-278, lifirafenib, mocetinostat, ONO-4059
DEG10170224	HER2-Inhibitor-1, cefotetan, VU591, BMS-599626, MK-8033, adavivint, nilotinib, hPGDS-IN-1, PF-573228, flumatinib, GLPG0187, BMS-833923, radotinib, MGCD-265, cefmetazole, Ro-5126766, cefmenoxime, CFI-402257, MBX-2982, tivozanib, imidurea, KB-SRC-4, aminopterin, LY2801653, INS316, BX-912, PLX8394, I-BRD9, KD025, CCT196969
DEG10170225	elinogrel, diosmin, icariin, tetrahydrofolic-acid, ZCL-278, folic-acid, alpha-glucosyl-hesperidin, cefmenoxime, CVT-10216, KB-SRC-4, T-5224, BMS-599626, HER2-Inhibitor-1, hyperin, lometrexol, salvianolic-acid-A, LX1031, tedizolid-phosphate, Mps1-IN-1, CaMKII-IN-1, JW-74, MLN2480, MK-3207, FN-1501, TG-100801, kuromanin, GS-9973, LY2801653, R547, ceftiofur

Continuation of Table 6.2	
Protein	**Ligands**
DEG10170226	indisulam, epacadostat, oglufanide, cefmenoxime, PF-06273340, nifursol, IMREG-1, HSR6071, PF-562271, FN-1501, XL147, adenosine-phosphate, telatinib, benzamil, PD-318088, acetazolamide, INS316, AMZ30, SB-415286, bendroflumethiazide, chlorthalidone, cyclopenthiazide, CCMI, GSK356278, BMS-707035, DNQX, CV-1808, molidustat, sitaxentan, azosemide
DEG10170227	epacadostat, JTE-013, amiloride, AZ-10417808, ditolylguanidine, SB-747651A, IDO5L, guanidinoethyldisulfide-bicarbonate, XL147, CZC-54252, SB-772077B, GSK2256098, HSR6071, penciclovir, CHIR-98014, ETC-159, SCH-900776, famotidine, VS-4718, CGP-52411, chlorthalidone, pirenoxine, TAK-659, CGP-57380, chlorothiazide, chlorproguanil, E7449, hydrochlorothiazide, polyinosine, entecavir
DEG10170233	adenosine-triphosphate, azilsartan-medoxomil, cefmenoxime, lometrexol, ceftobiprole, cefonicid, metafolin, fostamatinib, guadecitabine, epacadostat, GSK2239633A, INS316, defactinib, cefozopran, PF-573228, ebrotidine, paromomycin, compound-w, JW-74, tucatinib, GDC-0941, PRT062070, tetrahydrofolic-acid, ceritinib, lifitegrast, BAY-1251152, ceftriaxone, famotidine, CHIR-98014, PF-562271
DEG10170242	GSK1838705A, avitinib, BMS-817378, ribostamycin, nafamostat, protirelin, FN-1501, adenosine-triphosphate, AGI-6780, FR-180204, paromomycin, alpha-glucosyl-hesperidin, AMG900, PLX8394, CHIR-99021, tobramycin, folic-acid, ceftazidime, LXS196, cefonicid, SRT3190, diosmin, INS316, kanamycin, CP-316819, metafolin, cefixime, ceftriaxone, epacadostat, SRT1720
DEG10170253	elinogrel, ACTB-1003, TAK-632, gliquidone, cefmenoxime, adenosine-triphosphate, SB-683698, guadecitabine, tegobuvir, PRT062607, CA-4948, fluralaner, ML193, tivozanib, folic-acid, glipizide, AMI-1, CGP-71683, AMG900, regorafenib, cefazolin, TG-101209, S1P1-agonist-III, sulfatinib, bisantrene, CW-008, DCC-2618, TAK-593, BTT-3033, SNS-314
DEG10170266	sotrastaurin, bisindolylmaleimide-IX, uridine-5-triphosphate, adavivint, myricitrin, adenosine-triphosphate, indisulam, LY2090314, epacadostat, NS-11021, AZ-10417808, INS316, GF109203X, nemiralisib, azosemide, famotidine, cefotetan, PRT062607, cefonicid, guadecitabine, I-BRD9, EED226, PF-4981517, BAY-61-3606, isoquercitrin, KB-SRC-4, SNS-314, AMG-548, enzastaurin, imidurea
DEG10170279	naringin-dihydrochalcone, alpha-glucosyl-hesperidin, rutin, AR-12, etoposide-phosphate, HER2-Inhibitor-1, sennoside-A, sennoside-protonated, neomycin, metafolin, azilsartan-medoxomil, madecassoside, AMG900, ceftobiprole, MIW-815, EW-7197, tucatinib, GNTI, pazopanib, cefonicid, ebrotidine, adenosine-triphosphate, cefmenoxime, evans-blue, TD139, Ro-5126766, BAY-87-2243, MGCD-265, R406, GNF-5837
DEG10170280	famotidine, LY2979165, amiloride, guanaben-acetate, guanidinoethyldisulfide-bicarbonate, zaltidine, NS-3623, HA-1004, cariporide, sardomozide, acitazanolast, FR-180204, ICI-162846, lidamidine, taminadenant, SU3327, sangivamycin, LY215490, guanfacine, azathioprine, tenofovir, sulfisomidin, tenoxicam, 5-amino-3-D-ribofuranosylthiazolo[4,5-d]pyrimidin-2,7(3H,6H)-dione, cimetidine, iobenguane, tiotidine, chlorproguanil, nolatrexed, chlorothiazide

Continuation of Table 6.2	
Protein	**Ligands**
DEG10170284	DOTMP, rifampin, defactinib, INS316, BMS-599626, madecassoside, adenosine-triphosphate, JD-5037, PRT062607, naringin-dihydrochalcone, BEBT-908, cefmenoxime, ceftriaxone, uridine-5-triphosphate, LY2090314, nilotinib, cefotetan, MIW-815, ceftobiprole, guadecitabine, zoliflodacin, CUDC-907, sennoside-protonated, fostamatinib, isepamicin, moxalactam, rutin, adavivint, bisindolylmaleimide-IX, cilengitide
DEG10170285	zaltidine, radotinib, BW-348U87, minodronic-acid, TPCA-1, GSK2256294A, VLX600, sparfosate, thiophanate, siguazodan, AZ-10417808, benserazide, HA-1004, adaprev, guanidinoethyldisulfide-bicarbonate, acitazanolast, SB-200646, MLN2480, BMS-345541, m-chlorophenylbiguanide, nilotinib, TS-011, ellagic-acid, folic-acid, iobenguane, nelociguat, tifenazoxide, chlorproguanil, amiloride, APY-29
DEG10170287	evans-blue, PLX8394, tucatinib, ZCL-278, HER2-Inhibitor-1, IPI549, INS316, AMG900, cefazolin, guadecitabine, adenosine-triphosphate, sennoside-protonated, ebrotidine, CK-101, GS-9973, ceftriaxone, entrectinib, CFI-402257, sennoside-A, cefmenoxime, MK-8033, cefalonium, MGCD-265, radotinib, bisindolylmaleimide-IX, GLPG0187, pazopanib, adavivint, SU11274, PF-573228
DEG10170294	INS316, PIK-93, PF-05089771, uridine-5-triphosphate, DSM265, NS-11021, adenosine-triphosphate, epacadostat, N-(2-chlorophenyl)-2-((2E)-2-[1-(2-pyridinyl)ethylidene]hydrazinocarbothioyl)hydrazinecarbothioamide, triciribine-phosphate, succinylsulfathiazole, SDZ-220-040, cefazolin, SDZ-220-581, thiamine-pyrophosphate, AMG319, adenosine-phosphate, selinexor, guanidinoethyldisulfide-bicarbonate, AT-9283, IMREG-1, sulfasalazine, phthalylsulfathiazole, BW-348U87, tiotidine, WAY-213613, ARRY-334543, bisantrene, ebrotidine, BLZ945
DEG10170295	MIW-815, etoposide-phosphate, DOTMP, nilotinib, cilengitide, TD139, rutin, adavivint, KB-SRC-4, tucatinib, GNF-5837, HER2-Inhibitor-1, nemorubicin, DBPR-211, hesperidin, sennoside-protonated, m-THP, ceftriaxone, KPT-9274, SRT2104, mangafodipir, GNTI, AMG900, fostemsavir, NVP-BHG712, BMS-626529, BQ-123, alpha-glucosyl-hesperidin, EPZ005687, IPI549
DEG10170297	PF-562271, DOTMP, rutin, hyperin, adenosine-triphosphate, uridine-5-triphosphate, kuromanin, myricitrin, guadecitabine, PLX8394, epacadostat, bekanamycin, bisindolylmaleimide-IX, polyinosine, ribostamycin-sulfate, PF-573228, PRT062607, ebrotidine, dabigatran, defactinib, LY2979165, paromomycin, INS316, cefotetan, FR-180204, PF-431396, famotidine, JNJ-64619178, HER2-Inhibitor-1, cefozopran
DEG10170300	ceftriaxone, ceftiofur, tucatinib, cefozopran, azilsartan-medoxomil, PF-573228, ceftobiprole, adavivint, ebrotidine, guadecitabine, NS-5806, cefoselis, cefmenoxime, diosmin, CEP-32496, alpha-glucosyl-hesperidin, PF-05089771, epacadostat, PF-562271, NS-11021, uridine-5-triphosphate, APY-29, HER2-Inhibitor-1, VX-11e, cefpirome, MK-2461, cefonicid, bisantrene, bisindolylmaleimide-IX, PF-431396

Continuation of Table 6.2	
Protein	**Ligands**
DEG10170303	ceftriaxone, adenosine-triphosphate, cefoselis, radezolid, CZC-54252, BT-11, XL228, epacadostat, avapritinib, fostamatinib, cefazolin, pilaralisib, tozasertib, GSK1838705A, DOTMP, BMS-986142, PF-05089771, INS316, neohesperidin, TC-S-7004, acarbose, tetrahydrofolic-acid, cefonicid, ML193, KPT-9274, metafolin, Ro-5126766, ceftobiprole, CUDC-907, TD139
DEG10170304	azosemide, adenosine-triphosphate, trabodenoson, NS-11021, PF-05089771, LY2090314, INS316, uridine-5-triphosphate, epacadostat, polyinosine, ceftriaxone, FN-1501, SNS-314, LB42708, ebrotidine, benzthiazide, bisindolylmaleimide-IX, nafamostat, famotidine, ceftiofur, fludarabine-phosphate, lometrexol, cefixime, theaflavin, rimegepant, LP-533401, TG100-115, larotrectinib, Ro-61-8048, cefotetan
DEG10170313	INS316, trabodenoson, uridine-5-triphosphate, epacadostat, aminopterin, salazodine, AMZ30, cefotaxime, PF-04937319, cefmenoxime, citicoline, NS-11021, moxalactam, L-694247, PF-05089771, adenosine-triphosphate, SRT3190, TC-G-1008, PF-04217903, guadecitabine, famotidine, cefazolin, BI-78D3, MI-14, nafamostat, vericiguat, GPBAR-A, radotinib, VER-155008, cefonicid
DEG10170339	diclofenamide, epacadostat, adenosine-triphosphate, famotidine, ACT-132577, cefmenoxime, indisulam, KD025, PF-04217903, minodronic-acid, cefazolin, clorsulon, PF-573228, hyaluronic-acid, SB-747651A, CHIR-98014, AM-1241, ebrotidine, thiamine-pyrophosphate, dorzolamide, AZD8797, PF-06273340, brinzolamide, AZ960, molidustat, amiloride, zoledronic-acid, benzamil, hydrochlorothiazide, PU-H71
DEG10170348	R-428, Mps1-IN-1, R406, ebrotidine, folic-acid, nifursol, casanthranol-variant, TGR-1202, ARRY-334543, adenosine-triphosphate, DOTMP, PLX8394, cefmenoxime, KB-SRC-4, NS-3623, tucatinib, cilofexor, TAK-243, CEP-32496, GSK256066, azosemide, bendroflumethiazide, rebastinib, TC-S-7006, epacadostat, EED226, VER-155008, verdinexor, AMG900, GSK3326595
DEG10170349	tucatinib, pyrintegrin, DOTMP, cot-inhibitor-2, ciaftalan-zinc, FN-1501, sulfatinib, TC-S-7006, uridine-5-triphosphate, PIK-294, lorlatinib, presatovir, ARRY-334543, NS-11021, XL147, HER2-Inhibitor-1, cefmenoxime, fostemsavir, m-THP, adenosine-triphosphate, cot-inhibitor-1, SNS-314, MK-3207, fostamatinib, epacadostat, ABBV-744, abafungin, cefazolin, evobrutinib, PF-573228

For further investigation, molecular dynamic simulations can be exploited to ascertain the stability of those ligand:protein pairs ($30 \times 78 = 2340$ pairs), which might be intensive work and *in vitro* tests are also required. Another straightforward option is to start with biological assays. This approach is more applicable as a number of ligands in the enrichment subset recur for various targets. This table made up to a list of 626 ligands with uridine-triacetate appears in enrichment subsets for 78 MRSA proteins and 714 ligands for 72 modelled MRSA proteins. Table 6.3 listed ligands appeared in the enriched subsets with the frequencies in descending order. The frequency (column "Freq") here indicated how many times these compounds appeared in the enriched subsets. With a high frequency in the subsets of compounds with high predicted affinities toward MRSA proteins, these ligands stand a high chance of being active against one of the MRSA targets.

name	**freq**	**name**	**freq**
uridine-5-triphosphate	64	epacadostat	42
adenosine-triphosphate	61	adenosine-triphosphate	41
echinomycin	58	INS316	36
INS316	58	PF-573228	32
dactinomycin	56	uridine-5-triphosphate	31
candicidin	52	cefmenoxime	30
actinomycin-d	46	guadecitabine	26
rutin	36	HER2-Inhibitor-1	23
guadecitabine	29	ceftriaxone	21
citicoline	23	cefazolin	21
riboflavin-5-phosphate-sodium	23	famotidine	20
CGP-71683	20	AMG900	20
triciribine-phosphate	20	ebrotidine	19
folic-acid	20	NS-11021	17
diosmin	19	PF-562271	17
cefonicid	19	tucatinib	17
KB-SRC-4	18	cefonicid	17
baicalin	17	DOTMP	16
kuromanin	17	fostamatinib	15
epigallocatechin-gallate-(-)	17	cefotetan	14
famotidine	16	ceftiofur	14
EB-47	16	azosemide	14
DOTMP	16	PRT062607	14
epacadostat	15	imidurea	14
AMG900	15	bisindolylmaleimide-IX	13
procyanidin-B-2	14	CHIR-98014	13
myricitrin	14	ceftobiprole	13
HER2-Inhibitor-1	14	XL147	13
tetrahydrofolic-acid	14	folic-acid	13
lometrexol	13	PF-05089771	12
naringin	13	XL228	12
E7449	12	APY-29	11
thiamine-pyrophosphate	12	bisantrene	11
XL147	12	adavivint	11
CDBA	12	tetrahydrofolic-acid	11
everolimus	12	PLX8394	11
TPPS4	12	SNS-314	11
salvianolic-acid-A	12	metafolin	11
hypericin	12	radotinib	10
ceftriaxone	12	tiotidine	10
isoquercitrin	12	rutin	10
sirolimus	12	EB-47	9
fosfructose	11	nilotinib	9
inarigivir	11	guanidinoethyldisulfide-bicarbonate	9
hesperidin	11	MIW-815	9
reynoutrin	10	defactinib	9

Table 6.3: List of Repurposing Hub ligands with the highest frequency among top-ranked ligands in the enrichment subsets. The first column lists the compounds with high frequency in the enriched subsets for MRSA hits and the third column lists the compounds with high frequency in the enriched subsets for modelled MRSA proteins. The columns "freq" illustrate how many enriched subsets that contained the compound in the previous column.

Chapter 7

Conclusion

This chapter will wrap up this study by summarising the major findings in connection to the research objectives and questions, as well as discussing their worth and contribution. It will also go through the limitations of the study and make recommendations for further research.

Since the 1920s, thanks to the continuous discovery and development of a series of antibiotics and millions of lives have been saved. However, with the evolutionary progress of nature, bacteria have developed resistant mechanisms against antibiotics. The serious bacterial infections that seemed to be pushed back have become more serious with the resistant bacterial strains. Antibiotic resistance causes prolonged hospital admission, increased hospital charges and increased death rate. Many strains have been reported to be a serious threat and listed as a priority in finding additional medications. MRSA has been posing a burgeoning threat to the community by lifting the fatality rate and the expense of treatment. MRSA is classified as high priority threat by a number of health organisations.

On the other hand, the treatments for MRSA are considered to be inadequate. The key reason is the antimicrobial resistance of MRSA to current (antibiotic) medications. For instance, shortly after the introduction of penicillin into clinical treatments, *S. aureus* developed resistance against penicillin. It was also resistant against methicillin, from which the term Methicillin-resistant *Staphylococcus aureus* came. Recently, MRSA was also reported insusceptible to vancomycin. If the resistance trend remains, MRSA will become resistant to more antibiotics and cause more damage to the community. Another factor is the burden of developing a new drug. A drug can take up to 13 years and more than 2.0 billion dollars to develop from scratch until it reaches approval for clinical use. Furthermore, the success rate for a new drug is low, resulting in flattened productivity in recent years. Thus, it is vital to seek out new medicines for the treatment and prevention of MRSA infections.

Drug Repurposing (DR) is an approach to find new indications for existing drugs by exploiting the available pharmacokinetic information and safety profiles of not only existing and withdrawn drugs but also the compounds at clinical trial stages. DR can help to cut down the expenses off of millions of US dollars and 2 to 10 years while increasing the success rate in finding a new therapeutic indication, depending on the profile of the testing compounds. Another advantage is that it is promising for patients suffering rare diseases and useful in case of a pandemic.

Due to the need for further MRSA medications, the aim of this study is to investigate the existing compounds which can be repurposed for possible anti-MRSA activities. This study has successfully explored the space of repurposable compounds using computational docking tools in a consensus fashion. The main finding of this study is the list of potential candidates for anti-MRSA repurposing. This list consists of 621 ligands that have high predicted affinities toward MRSA targets. The final list is sorted in descending order of the frequency of the ligands showing high predicted affinities toward MRSA targets. The compounds in the list, with further biological assessments, can fill in the demand for new medications to treat MRSA infections.

Another finding is the novel consensus score that was able to discriminate between

the actives and other compounds in the chosen dataset. This consensus score helped to narrow down the list of potential candidates for MRSA repurposing. This study has explored various consensus models using mean and standard deviation as descriptors, with power increases from 1 to 10. The models took in the raw docking scores with exhaustively weighted combinations. Two benchmarks indicated models with absolute values showed consistent performance over models using real values, as the even power caused shifting in the ranking of actives. Another finding was the models showed better performance with linear settings compared to non-linear forms. Thus the linear form of the model using standard deviation and absolute values was marked as the best model for virtual screening for anti-MRSA candidates.

A key finding is that the number of docking programmes needed for consensus scores is small, only 5 or 6 programmes. If accuracy measures from CS models were monotonically increasing with the number of docking programmes used, the method would have been suggestive at best and not too useful at worst. Fortunately, this is not the case. The consensus effects increased significantly up to a certain number of docking programmes beyond which more docking scores hit the plateau region of the accuracy plots (Figure 5.10). Adding more programmes simply sacrificed computational time without gaining much improvement.

This study has applied a more general approach compared to other studies which focused on a few MRSA targets and smaller libraries of compounds. Here in this study, all available MRSA targets were investigated. Starting with a crucial point of essential genes, sequence alignment was employed to identify matching proteins, which encode the important biomolecules for the survival and growth of MRSA. Furthermore, to make use of the genes with insufficient matches, homology modelling was applied to search for homology structures. This results in the total of 150 MRSA targets to be explored. In terms of ligands, 5902 compounds for virtual screening were obtained from the Repurposing Hub, which contained all approved drugs, withdrawn drugs and compounds at clinical trials, all of which are ready for repurposing. Therefore the docking between the chosen ligands and proteins provided a broad investigation.

With available structural information from targets and compounds, virtual screening using molecular docking was employed as the main method in this study. A new consensus score that is more efficient in recognising the actives amongst other compounds was developed. The consensus score in this study employs the raw docking data with ease and high efficiency while other studies achieved less consistent improvement using traditional consensus scores or required sophisticated systems like machine learning. The novel consensus score also confirmed consensus strategies employing multiple docking programmes improved the overall performance with the appropriate setting. While individual programmes and traditional consensus scores produced AUROCC values of approximate 0.6, several novel consensus scores were able to generate AUROCC values of 0.833-0.873. This was a remarkable advancement, providing the ideal AUROCC was 1.0 when all the active ligands were perfectly scored at the top. .

The list of ligands provided by this study can be further investigated for anti-MRSA activity with high probability. For the next stage, biological experiments or molecular dynamics can be used to confirm the binding affinity between ligands and their targets. With experimental and clinical confirmation, a number of the compounds may reveal the anti-MRSA activities and further in clinical use. In addition, the consensus model developed in this work can be used in other virtual screening studies as well as other fields that employ binary classification such as diagnostic imaging.

The limitations of this study lie in the quantity and quality of current protein structural data. One limitation is the unavailability of a large number of proteins encoded by the MRSA essential genes. Although the Protein Data Bank is one of the biggest databases of structural information of biological macromolecules, the number of structures available is still far from the true number of true possible structures in all species. Therefore, it may need time for biologists and biochemists to expand the volume of information. In the meantime, this leads to the fact that a number of MRSA gene sequences had no matching proteins or templates, hence were discarded in this study. One more limitation lies in

the nature of protein structures obtained from the Protein Data Bank. The structural data is just a snapshot of the protein, reflecting only one of the numerous states of the protein dynamics. Therefore, docking does not fully assess the binding between proteins and ligands. Besides, with the current technologies, the quality of the protein structure is still not the best, represented by the low resolution of the protein structure, resulting in possible wrong residues in the protein structures as well as the binding sites. This could lead to the approximation in the residue identity, hence the accuracy of the following methods. In addition, the unavailability of a co-crystallised ligand in some structures led to uncertainty in binding site prediction.

Another source of limitation is the accuracy of docking programmes. Docking programmes exploit the existing intermolecular interaction information together with physicochemical properties to predict the affinity likelihood of prospective protein-ligand complexes. Although many scoring functions have been developed to improve the power of docking programmes, an sufficient performance to recognise the activity amongst a library of the compounds is still desired. As mentioned above, the structural data of proteins represents one of the infinite states of the proteins. Many docking programmes have integrated the flexibility of the protein in docking but this will significantly increase the computational cost.

In this study, one limitation is the limited availability of computer resources, with limited time for each High-Performance Computer system during the research and the incompatibility of the systems.

For future research, it is recommended to start with the proposed list of potential ligands from this study for anti-MRSA activity. Molecular dynamics or biological assays is adequately suitable for this task. One more suggestion is to fulfil this study by exploring the remaining space of MRSA targets and repurposable compounds. One direction is to re-screen the gene sequences against PDB for missing targets. The PDB is a repository for experimental structures, hence its capacity is currently constrained. However, it is possible to find a new hit with a better resolution or a template with better coverage later on as the database is growing along with time. Another direction is to adopt the remaining chemical space outside Lipinski's rule. In this study, compounds that satisfied Lipinski's rule and with less than 10 rotatable bonds were selected, leaving the relatively large molecules. This option might take a longer computational time. Another direction is to add the flexibility of the proteins by using flexible docking. However, this approach might increase the computational time as well. One more suggestion is to exploit the models built in this study with more possibilities of combinations, for example using scores from different normalisation schemes rather than the raw docking scores.

In conclusion, this study performed a workflow to repurpose drugs and compounds for anti-MRSA activity, with the results serving as a list of candidates. It also built a consensus model that is highly capable of recognising active ligands from a library of compounds. With further research, the list of candidates can be verified using biological testing. The consensus model can be improved and applied to other virtual screening studies.

Appendix A

Appendix

1 Essential Genes of *Staphylococcus aureus*

Table A.1: Essential genes of Staphylococcus aureus
The column "DEG ID" contains gene codes from the Database of Essential Genes, the column "Gene Name" represents gene named regarding *S. aureus* and the column "Function" lists the function of each gene.

DEG ID	**Gene Name**	**Function**
DEG10170001	dnaA	chromosomal replication initiation protein
DEG10170002	SAOUHSC_00002	DNA polymerase III subunit beta
DEG10170003	SAOUHSC_00003	hypothetical protein
DEG10170004	SAOUHSC_00005	DNA gyrase
DEG10170005	SAOUHSC_00006	DNA gyrase
DEG10170006	SAOUHSC_00009	seryl-tRNA synthetase
DEG10170007	SAOUHSC_00015	hypothetical protein
DEG10170008	SAOUHSC_00018	replicative DNA helicase
DEG10170009	SAOUHSC_00020	two-component response regulator
DEG10170010	SAOUHSC_00021	sensory box histidine kinase VicK
DEG10170011	SAOUHSC_00223	teichoic acid biosynthesis protein F
DEG10170012	ispD	2-C-methyl-D-erythritol 4-phosphate cytidylyl-transferase
DEG10170013	SAOUHSC_00226	hypothetical protein
DEG10170014	SAOUHSC_00227	hypothetical protein
DEG10170015	SAOUHSC_00336	acetyl-CoA acetyltransferase
DEG10170016	SAOUHSC_00345	hypothetical protein
DEG10170017	rpsF	30S ribosomal protein S6
DEG10170018	SAOUHSC_00349	bacteriophage L54a
DEG10170019	rpsR	30S ribosomal protein S18
DEG10170020	guaA	GMP synthase
DEG10170021	SAOUHSC_00442	DNA polymerase III
DEG10170022	SAOUHSC_00444	hypothetical protein
DEG10170023	tmk	thymidylate kinase
DEG10170024	SAOUHSC_00454	DNA polymerase III
DEG10170025	SAOUHSC_00461	methionyl-tRNA synthetase
DEG10170026	glmU	bifunctional N-acetylglucosamine-1-phosphate uridyltransferase/glucosamine-1-phosphate acetyltransferase
DEG10170027	SAOUHSC_00472	ribose-phosphate pyrophosphokinase
DEG10170028	SAOUHSC_00474	50S ribosomal protein L25/general stress protein Ctc
DEG10170029	SAOUHSC_00475	peptidyl-tRNA hydrolase
DEG10170030	SAOUHSC_00482	hypothetical protein
DEG10170031	SAOUHSC_00484	hypothetical protein

Continuation of Table A.1		
DEG ID	**Gene Name**	**Function**
DEG10170032	SAOUHSC_00489	dihydropteroate synthase
DEG10170033	SAOUHSC_00490	dihydroneopterin aldolase
DEG10170034	SAOUHSC_00491	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase
DEG10170035	lysS	lysyl-tRNA synthetase
DEG10170036	gltX	glutamyl-tRNA synthetase
DEG10170037	SAOUHSC_00510	serine acetyltransferase
DEG10170038	cysS	cysteinyl-tRNA synthetase
DEG10170039	secE	preprotein translocase subunit SecE
DEG10170040	rplK	50S ribosomal protein L11
DEG10170041	rplA	50S ribosomal protein L1
DEG10170042	rplJ	50S ribosomal protein L10
DEG10170043	rplL	50S ribosomal protein L7/L12
DEG10170044	rpoB	DNA-directed RNA polymerase subunit beta
DEG10170045	SAOUHSC_00525	DNA-directed RNA polymerase subunit beta
DEG10170046	rpsL	30S ribosomal protein S12
DEG10170047	SAOUHSC_00528	30S ribosomal protein S7
DEG10170048	SAOUHSC_00529	elongation factor G
DEG10170049	SAOUHSC_00530	elongation factor Tu
DEG10170050	SAOUHSC_00549	putative GTP cyclohydrolase
DEG10170051	eutD	phosphotransacetylase
DEG10170052	SAOUHSC_00575	hypothetical protein
DEG10170053	SAOUHSC_00577	mevalonate kinase
DEG10170054	SAOUHSC_00578	mevalonate diphosphate decarboxylase
DEG10170055	SAOUHSC_00579	phosphomevalonate kinase
DEG10170056	argS	arginyl-tRNA synthetase
DEG10170057	SAOUHSC_00620	accessory regulator A
DEG10170058	SAOUHSC_00640	teichoic acid biosynthesis protein
DEG10170059	SAOUHSC_00641	teichoic acids export protein ATP-binding subunit
DEG10170060	SAOUHSC_00642	teichoic acid biosynthesis protein
DEG10170061	SAOUHSC_00643	tagB protein
DEG10170062	SAOUHSC_00645	glycerol-3-phosphate cytidylyltransferase
DEG10170063	SAOUHSC_00728	hypothetical protein
DEG10170064	nrdI	ribonucleotide reductase stimulatory protein
DEG10170065	SAOUHSC_00742	ribonucleotide-diphosphate reductase subunit alpha
DEG10170066	nrdF	ribonucleotide-diphosphate reductase subunit beta
DEG10170067	murB	UDP-N-acetylenolpyruvoylglucosamine reductase
DEG10170068	SAOUHSC_00760	hypothetical protein
DEG10170069	SAOUHSC_00762	hypothetical protein
DEG10170070	secA	preprotein translocase subunit SecA
DEG10170071	SAOUHSC_00771	peptide chain release factor 2
DEG10170072	SAOUHSC_00781	HPr kinase/phosphorylase
DEG10170073	SAOUHSC_00785	thioredoxin reductase
DEG10170074	SAOUHSC_00788	hypothetical protein
DEG10170075	clpP	ATP-dependent Clp protease proteolytic subunit
DEG10170076	SAOUHSC_00793	hypothetical protein

Continuation of Table A.1		
DEG ID	**Gene Name**	**Function**
DEG10170077	SAOUHSC_00795	glyceraldehyde-3-phosphate dehydrogenase
DEG10170078	pgk	phosphoglycerate kinase
DEG10170079	tpiA	triosephosphate isomerase
DEG10170080	SAOUHSC_00798	phosphoglyceromutase
DEG10170081	eno	phosphopyruvate hydratase
DEG10170082	SAOUHSC_00803	ribonuclease R
DEG10170083	smpB	SsrA-binding protein
DEG10170084	SAOUHSC_00847	ABC transporter
DEG10170085	SAOUHSC_00848	hypothetical protein
DEG10170086	SAOUHSC_00849	aminotransferase
DEG10170087	SAOUHSC_00850	hypothetical protein
DEG10170088	SAOUHSC_00851	hypothetical protein
DEG10170089	SAOUHSC_00868	hypothetical protein
DEG10170090	SAOUHSC_00869	D-alanine–poly(phosphoribitol) ligase subunit 1
DEG10170091	SAOUHSC_00870	dltB protein
DEG10170092	SAOUHSC_00871	D-alanine–poly(phosphoribitol) ligase subunit 2
DEG10170093	SAOUHSC_00872	extramembranal protein
DEG10170094	SAOUHSC_00881	hypothetical protein
DEG10170095	SAOUHSC_00892	hypothetical protein
DEG10170096	pgi	glucose-6-phosphate isomerase
DEG10170097	SAOUHSC_00903	Signal peptidase IB
DEG10170098	SAOUHSC_00920	3-oxoacyl-(acyl carrier protein) synthase III
DEG10170099	SAOUHSC_00921	3-oxoacyl- synthase
DEG10170100	SAOUHSC_00922	hypothetical protein
DEG10170101	SAOUHSC_00933	tryptophanyl-tRNA synthetase
DEG10170102	spxA	transcriptional regulator Spx
DEG10170103	ppnK	inorganic polyphosphate/ATP-NAD kinase
DEG10170104	SAOUHSC_00947	enoyl-(acyl carrier protein) reductase
DEG10170105	SAOUHSC_00954	UDP-N-acetylmuramoylalanyl-D-glutamate–L-lysine ligase
DEG10170106	SAOUHSC_00957	hypothetical protein
DEG10170107	SAOUHSC_00980	1
DEG10170108	SAOUHSC_00998	fmt protein
DEG10170109	SAOUHSC_01028	phosphocarrier protein HPr
DEG10170110	SAOUHSC_01035	hypothetical protein
DEG10170111	SAOUHSC_01036	hypothetical protein
DEG10170112	def	peptide deformylase
DEG10170113	SAOUHSC_01040	pyruvate dehydrogenase complex
DEG10170114	SAOUHSC_01050	hypothetical protein
DEG10170115	SAOUHSC_01063	hypothetical protein
DEG10170116	coaD	phosphopantetheine adenylyltransferase
DEG10170117	SAOUHSC_01077	hypothetical protein
DEG10170118	SAOUHSC_A01041	hypothetical protein
DEG10170119	SAOUHSC_01078	ribosomal protein L32
DEG10170120	pheS	phenylalanyl-tRNA synthetase subunit alpha
DEG10170121	pheT	phenylalanyl-tRNA synthetase subunit beta
DEG10170122	SAOUHSC_01100	thioredoxin
DEG10170123	SAOUHSC_01106	glutamate racemase
DEG10170124	SAOUHSC_01119	hypothetical protein
DEG10170125	SAOUHSC_01144	cell division protein
DEG10170126	SAOUHSC_01145	penicillin-binding protein 1

Continuation of Table A.1		
DEG ID	**Gene Name**	**Function**
DEG10170127	mraY	phospho-N-acetylmuramoyl-pentapeptide-transferase
DEG10170128	murD	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase
DEG10170129	SAOUHSC_01148	cell division protein
DEG10170130	SAOUHSC_01149	cell division protein
DEG10170131	SAOUHSC_01150	cell division protein FtsZ
DEG10170132	SAOUHSC_01154	hypothetical protein
DEG10170133	ileS	isoleucyl-tRNA synthetase
DEG10170134	gmk	guanylate kinase
DEG10170135	SAOUHSC_01178	phosphopantothenoylcysteine decarboxylase/phosphopantothenate–cysteine ligase
DEG10170136	SAOUHSC_01179	primosomal protein N
DEG10170137	SAOUHSC_01183	methionyl-tRNA formyltransferase
DEG10170138	SAOUHSC_01188	hypothetical protein
DEG10170139	SAOUHSC_01189	ribulose-phosphate 3-epimerase
DEG10170140	SAOUHSC_01190	hypothetical protein
DEG10170141	rpmB	50S ribosomal protein L28
DEG10170142	SAOUHSC_01197	putative glycerol-3-phosphate acyltransferase PlsX
DEG10170143	SAOUHSC_01198	malonyl CoA-acyl carrier protein transacylase
DEG10170144	SAOUHSC_01199	3-oxoacyl-(acyl-carrier-protein) reductase
DEG10170145	acpP	acyl carrier protein
DEG10170146	SAOUHSC_01205	signal recognition particle-docking protein FtsY
DEG10170147	SAOUHSC_01207	signal recognition particle protein
DEG10170148	rpsP	30S ribosomal protein S16
DEG10170149	rimM	16S rRNA-processing protein RimM
DEG10170150	trmD	tRNA (guanine-N(1)-)-methyltransferase
DEG10170151	rplS	50S ribosomal protein L19
DEG10170152	rbgA	ribosomal biogenesis GTPase
DEG10170153	sucC	succinyl-CoA synthetase subunit beta
DEG10170154	SAOUHSC_01222	DNA topoisomerase I
DEG10170155	rpsB	30S ribosomal protein S2
DEG10170156	tsf	elongation factor Ts
DEG10170157	pyrH	uridylate kinase
DEG10170158	frr	ribosome recycling factor
DEG10170159	SAOUHSC_01237	undecaprenyl pyrophosphate synthase
DEG10170160	SAOUHSC_01238	phosphatidate cytidylyltransferase
DEG10170161	SAOUHSC_01240	prolyl-tRNA synthetase
DEG10170162	polC	DNA polymerase III PolC
DEG10170163	nusA	transcription elongation factor NusA
DEG10170164	SAOUHSC_01244	hypothetical protein
DEG10170165	SAOUHSC_01245	hypothetical protein
DEG10170166	infB	translation initiation factor IF-2
DEG10170167	SAOUHSC_01249	riboflavin biosynthesis protein RibF
DEG10170168	rpsO	30S ribosomal protein S15
DEG10170169	SAOUHSC_01252	hypothetical protein
DEG10170170	SAOUHSC_01260	CDP-diacylglycerol–glycerol-3-phosphate 3-phosphatidyltransferase
DEG10170171	SAOUHSC_01285	glutamine synthetase repressor
DEG10170172	SAOUHSC_01287	glutamine synthetase

Continuation of Table A.1		
DEG ID	**Gene Name**	**Function**
DEG10170173	SAOUHSC_01333	LexA repressor
DEG10170174	SAOUHSC_01337	transketolase
DEG10170175	SAOUHSC_01350	hypothetical protein
DEG10170176	SAOUHSC_01351	DNA topoisomerase IV subunit B
DEG10170177	SAOUHSC_01352	DNA topoisomerase IV subunit A
DEG10170178	SAOUHSC_01359	hypothetical protein
DEG10170179	SAOUHSC_01361	transcriptional regulator
DEG10170180	SAOUHSC_01362	4-oxalocrotonate tautomerase
DEG10170181	SAOUHSC_01373	methicillin resistance factor
DEG10170182	SAOUHSC_01374	methicillin resistance factor
DEG10170183	murG	undecaprenyldiphospho-muramoylpentapeptide beta-N- acetylglucosaminyltransferase
DEG10170184	SAOUHSC_01434	dihydrofolate reductase
DEG10170185	thyA	thymidylate synthase
DEG10170186	SAOUHSC_01462	hypothetical protein
DEG10170187	recU	Holliday junction-specific endonuclease
DEG10170188	SAOUHSC_01467	penicillin-binding protein 2
DEG10170189	SAOUHSC_01470	hypothetical protein
DEG10170190	SAOUHSC_01473	BirA bifunctional protein
DEG10170191	SAOUHSC_01474	tRNA CCA-pyrophosphorylase
DEG10170192	SAOUHSC_01477	hypothetical protein
DEG10170193	SAOUHSC_01490	DNA-binding protein HU
DEG10170194	engA	GTP-binding protein EngA
DEG10170195	cmk	cytidylate kinase
DEG10170196	SAOUHSC_01501	elastin binding protein
DEG10170197	SAOUHSC_01504	ferredoxin
DEG10170198	SAOUHSC_01592	transcriptional regulator
DEG10170199	SAOUHSC_01598	AtsA/ElaC family protein
DEG10170200	SAOUHSC_01599	glucose-6-phosphate 1-dehydrogenase
DEG10170201	SAOUHSC_01605	6-phosphogluconate dehydrogenase
DEG10170202	SAOUHSC_01623	acetyl-CoA carboxylase biotin carboxylase subunit
DEG10170203	SAOUHSC_01624	acetyl-CoA carboxylase
DEG10170204	SAOUHSC_01625	elongation factor P
DEG10170205	SAOUHSC_01627	hypothetical protein
DEG10170206	SAOUHSC_01661	hypothetical protein
DEG10170207	SAOUHSC_01662	RNA polymerase sigma factor RpoD
DEG10170208	SAOUHSC_01663	DNA primase
DEG10170209	SAOUHSC_01666	glycyl-tRNA synthetase
DEG10170210	era	GTP-binding protein Era
DEG10170211	SAOUHSC_01672	hypothetical protein
DEG10170212	rpsU	30S ribosomal protein S21
DEG10170213	SAOUHSC_01682	chaperone protein DnaJ
DEG10170214	dnaK	molecular chaperone DnaK
DEG10170215	SAOUHSC_01684	heat shock protein GrpE
DEG10170216	holA	DNA polymerase III subunit delta
DEG10170217	SAOUHSC_01697	nicotinate (nicotinamide) nucleotide adenylyltransferase
DEG10170218	SAOUHSC_01700	GTP-binding protein YqeH
DEG10170219	SAOUHSC_01701	hypothetical protein
DEG10170220	greA	transcription elongation factor GreA
DEG10170221	SAOUHSC_01720	Holliday junction resolvase-like protein

Continuation of Table A.1		
DEG ID	**Gene Name**	**Function**
DEG10170222	SAOUHSC_01721	hypothetical protein
DEG10170223	alaS	alanyl-tRNA synthetase
DEG10170224	SAOUHSC_01725	tRNA methyl transferase
DEG10170225	SAOUHSC_01726	(5-methylaminomethyl-2-thiouridylate)-methyltransferase
DEG10170226	SAOUHSC_01727	hypothetical protein
DEG10170227	aspS	aspartyl-tRNA synthetase
DEG10170228	hisS	histidyl-tRNA synthetase
DEG10170229	SAOUHSC_01739	hypothetical protein
DEG10170230	SAOUHSC_01741	D-tyrosyl-tRNA(Tyr) deacylase
DEG10170231	SAOUHSC_01742	GTP pyrophosphokinase
DEG10170232	SAOUHSC_01746	bifunctional preprotein translocase subunit SecD/SecF
DEG10170233	ruvB	Holliday junction DNA helicase RuvB
DEG10170234	ruvA	Holliday junction DNA helicase RuvA
DEG10170235	obgE	GTPase ObgE
DEG10170236	rpmA	50S ribosomal protein L27
DEG10170237	SAOUHSC_01756	hypothetical protein
DEG10170238	rplU	50S ribosomal protein L21
DEG10170239	SAOUHSC_01766	folylpolyglutamate synthase/dihydrofolate synthase
DEG10170240	valS	valyl-tRNA synthetase
DEG10170241	SAOUHSC_01770	hypothetical protein
DEG10170242	engB	ribosome biogenesis GTP-binding protein YsxC
DEG10170243	SAOUHSC_01782	hypothetical protein
DEG10170244	rplT	50S ribosomal protein L20
DEG10170245	rpmI	50S ribosomal protein L35
DEG10170246	infC	translation initiation factor IF-3
DEG10170247	SAOUHSC_01787	hypothetical protein
DEG10170248	thrS	threonyl-tRNA synthetase
DEG10170249	SAOUHSC_01791	primosomal protein DnaI
DEG10170250	SAOUHSC_01792	hypothetical protein
DEG10170251	coaE	dephospho-CoA kinase
DEG10170252	SAOUHSC_01806	pyruvate kinase
DEG10170253	SAOUHSC_01807	6-phosphofructokinase
DEG10170254	SAOUHSC_01808	acetyl-CoA carboxylase carboxyltransferase subunit alpha
DEG10170255	SAOUHSC_01809	acetyl-CoA carboxylase subunit beta
DEG10170256	SAOUHSC_01811	DNA polymerase III alpha subunit superfamily protein
DEG10170257	SAOUHSC_01827	septation ring formation regulator EzrA
DEG10170258	rpsD	30S ribosomal protein S4
DEG10170259	SAOUHSC_01837	1-acyl-sn-glycerol-3-phosphate acyltransferase domain-containing protein
DEG10170260	SAOUHSC_01839	tyrosyl-tRNA synthetase
DEG10170261	murC	UDP-N-acetylmuramate–L-alanine ligase
DEG10170262	SAOUHSC_01866	hypothetical protein
DEG10170263	SAOUHSC_01871	polysaccharide biosynthesis protein
DEG10170264	leuS	leucyl-tRNA synthetase
DEG10170265	SAOUHSC_01908	hypothetical protein
DEG10170266	SAOUHSC_01909	S-adenosylmethionine synthetase
DEG10170267	SAOUHSC_01928	transposase family protein

Continuation of Table A.1		
DEG ID	**Gene Name**	**Function**
DEG10170268	SAOUHSC_01930	hypothetical protein
DEG10170269	SAOUHSC_01979	hypothetical protein
DEG10170270	SAOUHSC_02102	methionine aminopeptidase
DEG10170271	SAOUHSC_02106	hypothetical protein
DEG10170272	SAOUHSC_02107	UDP-N-acetylmuramyl tripeptide synthetase
DEG10170273	SAOUHSC_02114	putative lipid kinase
DEG10170274	gatB	aspartyl/glutamyl-tRNA amidotransferase subunit B
DEG10170275	gatA	aspartyl/glutamyl-tRNA amidotransferase subunit A
DEG10170276	gatC	aspartyl/glutamyl-tRNA amidotransferase subunit C
DEG10170277	SAOUHSC_02122	DNA ligase
DEG10170278	SAOUHSC_02123	ATP-dependent DNA helicase PcrA
DEG10170279	nadE	NAD synthetase
DEG10170280	SAOUHSC_02133	nicotinate phosphoribosyltransferase
DEG10170281	SAOUHSC_02140	putative manganese-dependent inorganic pyrophosphatase
DEG10170282	SAOUHSC_02151	hypothetical protein
DEG10170283	SAOUHSC_02152	ABC transporter
DEG10170284	groEL	chaperonin GroEL
DEG10170285	groES	co-chaperonin GroES
DEG10170286	SAOUHSC_02260	delta-hemolysin
DEG10170287	SAOUHSC_02277	putative DNA-binding/iron metalloprotein/AP endonuclease
DEG10170288	SAOUHSC_02279	hypothetical protein
DEG10170289	SAOUHSC_02280	hypothetical protein
DEG10170290	acpS	4'-phosphopantetheinyl transferase
DEG10170291	SAOUHSC_02317	UDP-N-acetylmuramoylalanyl-D-glutamyl-2
DEG10170292	ddl	D-alanyl-alanine synthetase A
DEG10170293	SAOUHSC_02327	hypothetical protein
DEG10170294	fabZ	(3R)-hydroxymyristoyl-ACP dehydratase
DEG10170295	SAOUHSC_02337	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
DEG10170296	SAOUHSC_02357	hypothetical protein
DEG10170297	prfA	peptide chain release factor 1
DEG10170298	rpmE2	50S ribosomal protein L31 type B
DEG10170299	SAOUHSC_02366	fructose-bisphosphate aldolase
DEG10170300	pyrG	CTP synthetase
DEG10170301	SAOUHSC_02371	pantothenate kinase
DEG10170302	SAOUHSC_02399	glucosamine–fructose-6-phosphate aminotransferase
DEG10170303	glmM	phosphoglucosamine mutase
DEG10170304	SAOUHSC_02407	hypothetical protein
DEG10170305	rpsI	30S ribosomal protein S9
DEG10170306	rplM	50S ribosomal protein L13
DEG10170307	rplQ	50S ribosomal protein L17
DEG10170308	SAOUHSC_02485	DNA-directed RNA polymerase subunit alpha
DEG10170309	SAOUHSC_02486	30S ribosomal protein S11
DEG10170310	rpsM	30S ribosomal protein S13
DEG10170311	rpmJ	50S ribosomal protein L36
DEG10170312	infA	translation initiation factor IF-1

Continuation of Table A.1		
DEG ID	**Gene Name**	**Function**
DEG10170313	adk	adenylate kinase
DEG10170314	secY	preprotein translocase subunit SecY
DEG10170315	rplO	50S ribosomal protein L15
DEG10170316	rpmD	50S ribosomal protein L30
DEG10170317	rpsE	30S ribosomal protein S5
DEG10170318	rplR	50S ribosomal protein L18
DEG10170319	rplF	50S ribosomal protein L6
DEG10170320	rpsH	30S ribosomal protein S8
DEG10170321	rpsN	30S ribosomal protein S14
DEG10170322	rplE	50S ribosomal protein L5
DEG10170323	rplX	50S ribosomal protein L24
DEG10170324	rplN	50S ribosomal protein L14
DEG10170325	rpsQ	30S ribosomal protein S17
DEG10170326	SAOUHSC_02504	50S ribosomal protein L29
DEG10170327	rplP	50S ribosomal protein L16
DEG10170328	rpsC	30S ribosomal protein S3
DEG10170329	rplV	50S ribosomal protein L22
DEG10170330	rpsS	30S ribosomal protein S19
DEG10170331	rplB	50S ribosomal protein L2
DEG10170332	rplW	50S ribosomal protein L23
DEG10170333	rplD	50S ribosomal protein L4
DEG10170334	rplC	50S ribosomal protein L3
DEG10170335	SAOUHSC_02527	FmhB protein
DEG10170336	SAOUHSC_02571	secretory antigen precursor
DEG10170337	SAOUHSC_02572	hypothetical protein
DEG10170338	SAOUHSC_02575	hypothetical protein
DEG10170339	SAOUHSC_02612	ribose-5-phosphate isomerase A
DEG10170340	SAOUHSC_02623	isopentenyl pyrophosphate isomerase
DEG10170341	SAOUHSC_02720	hypothetical protein
DEG10170342	SAOUHSC_02757	hypothetical protein
DEG10170343	SAOUHSC_02791	pyrophosphohydrolase
DEG10170344	SAOUHSC_02805	hypothetical protein
DEG10170345	SAOUHSC_02859	hydroxymethylglutaryl-CoA reductase
DEG10170346	SAOUHSC_02860	HMG-CoA synthase
DEG10170347	SAOUHSC_03049	hypothetical protein
DEG10170348	SAOUHSC_03052	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA
DEG10170349	trmE	tRNA modification GTPase TrmE
DEG10170350	rnpA	ribonuclease P
DEG10170351	rpmH	50S ribosomal protein L34

2 Results of Sequence Alignment of MRSA Targets

Table A.2: MRSA hits in PDB from BLAST sequence alignment
The table displays the hits from sequence alignment of *S. aureus* against the Protein Data Bank. The hits were chosen with Identity from sequence alignment is equal or greater than 95%. The letter after the PDB entry represents the protein chain.

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
DEG10170006	428	449	890	100%	0	6R1N_A

Continuation of Table A.2

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
DEG101700023	205	205	417	100%	4E-151	2CCJ_A, 2CCJ_B, 2CCK_A, 2CCK_B, 4GFD_A, 4GFD_B, 4GSY_A, 4GSY_B, 4HDC_A, 4HDC_B, 4HEJ_A, 4HEJ_B, 4HLC_A, 4HLC_B, 4HLD_A, 4HLD_B, 4QG7_A, 4QG7_B, 4QGA_A, 4QGA_B, 4QGF_A, 4QGF_B, 4QGG_A, 4QGG_B, 4QGH_A, 4QGH_B, 4XWA_A, 4XWA_B
		229	418	100%	7E-151	4DWJ_A, 4DWJ_B, 4DWJ_C, 4DWJ_D, 4DWJ_E, 4DWJ_F, 4DWJ_G, 4DWJ_H, 4EAQ_A, 4EAQ_B, 4F4I_A, 4F4I_B
		225	418	100%	7E-151	2CCG_A, 2CCG_B
		208	407	99%	8E-147	4MQB_A, 4MQB_B
DEG10170029	190	198	390	100%	8E-141	4YLY_A, 4YLY_B
DEG10170032	267	270	545	99%	0	4HB7_A, 4HB7_B
		270	534	96%	0	1AD1_A, 1AD1_B, 1AD4_A, 1AD4_B
		291	532	95%	0	6CLU_A, 6CLU_B, 6CLU_C, 6CLU_D, 6CLV_A, 6CLV_B, 6CLV_C, 6CLV_D
DEG10170033	121	121	246	99%	3.00E-86	1RRI_A, 1RRW_A, 1RRY_A, 1RS2_A, 1RS4_A, 1RSD_A, 1RSI_A, 1U68_A, 1DHN_A, 2DHN_A, 2NM2_A, 2NM2_B, 2NM2_C, 2NM2_D, 2NM3_A
DEG10170034	158	161	322	99%	6E-115	3QBC_A, 3QBC_B, 4AD6_A, 4AD6_B, 4CRJ_A, 4CWB_A, 4CYU_B, 4CYU_C, 4CYU_D, 5ETR_B, 5ETR_A, 5ETS_B, 5ETS_A, 5ETT_B, 5ETT_A
		161	329	99%	3E-114	4CYU_A

Continuation of Table A.2

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
		161	318	99%	2E-113	5ETQ_A, 5ETQ_B, 5ETV_A
DEG10170048	693	693	1434	100%	0	2XEX_A, 2XEX_B
		693	1432	99%	0	3ZZ0_A, 3ZZ0_B
		693	1432	99%	0	3ZZT_A, 3ZZT_B
		693	1430	99%	0	3ZZU_A, 3ZZU_B
DEG10170051	328	332	640	98%	0	4E4R_A
DEG10170053	306	308	621	99%	0	2X7I_A
DEG10170054	327	331	682	100%	0	2HK2_A, 2HK2_B, 2HK3_A, 2HK3_B
DEG10170057	124	127	246	100%	7E-86	2FRH_A, 2FRH_B
		124	244	99%	4.E-87	2FNP_A, 2FNP_B
		123	243	100%	6.E-85	1FZP_D, 1FZP_B
DEG10170062	132	132	268	100%	2E-94	2B7L_A, 2B7L_B, 2B7L_C, 2B7L_D
DEG10170067	307	326	625	100%	0	1HSK_A, 1HSK_B
DEG10170073	311	312	610	98%	0	4GCM_A, 4GCM_B
DEG10170075	195	195	400	100%	0	1E-144, 5DL1_A, 5DL1_B, 5DL1_C, 5DL1_D, 5DL1_E, 5DL1_F, 5DL1_G, 5DL1_H, 5DL1_I, 5DL1_J, 5DL1_K, 5DL1_L, 5DL1_M, 5DL1_N
		203	400	100%	1.E-144	3QWD_A, 3QWD_B, 3QWD_C, 3QWD_D, 3QWD_E, 3QWD_F, 3QWD_G, 3QWD_H, 3QWD_I, 3QWD_J, 3QWD_K, 3QWD_L, 3QWD_M, 3QWD_N, 3V5E_A, 3V5E_B, 3V5E_C, 3V5E_D, 3V5E_E, 3V5E_F, 3V5E_G, 3V5E_H, 3V5E_I, 3V5E_J, 3V5E_K, 3V5E_L, 3V5E_M, 3V5E_N
		197	400	100%	2.E-144	3ST9_A, 3ST9_B, 3ST9_C, 3ST9_D, 3ST9_E, 3ST9_F, 3ST9_G, 3STA_V, 3STA_A, 3STA_B, 3STA_C, 3STA_E, 3STA_F, 3STA_G, 3STA_I, 3STA_K, 3STA_L, 3STA_M, 3STA_N, 3STA_S, 3STA_T

Continuation of Table A.2

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
		203	400	100%	2.E-144	4EMM_V, 4EMM_A, 4EMM_B, 4EMM_C, 4EMM_D, 4EMM_E, 4EMM_F, 4EMM_G, 4EMM_H, 4EMM_I, 4EMM_J, 4EMM_K, 4EMM_L, 4EMM_M, 5VZ2_A, 5VZ2_B, 5VZ2_C, 5VZ2_D, 5VZ2_E, 5VZ2_F, 5VZ2_G, 5VZ2_I, 5VZ2_K, 5VZ2_L, 5VZ2_M, 5VZ2_N, 5VZ2_S, 5VZ2_T, 5W18_A, 5W18_B, 5W18_C, 5W18_D, 5W18_E, 5W18_F, 5W18_G, 5W18_I, 5W18_K, 5W18_L, 5W18_M, 5W18_N, 5W18_S, 5W18_T
		203	399	99%	4.E-144	3V5I_A, 3V5I_B, 3V5I_C, 3V5I_D, 3V5I_E, 3V5I_F, 3V5I_G, 3V5I_H, 3V5I_I, 3V5I_J, 3V5I_K, 3V5I_L, 3V5I_M, 3V5I_N, 3V5I_O, 3V5I_P, 3V5I_Q, 3V5I_R, 3V5I_S, 3V5I_T, 3V5I_U, 3V5I_V, 3V5I_W, 3V5I_X, 3V5I_Y, 3V5I_Z, 3V5I_AA, 3V5I_BB
		195	397	99%	3.E-143	5C90_A, 5C90_B, 5C90_C, 5C90_D, 5C90_E, 5C90_F, 5C90_G, 5C90_H, 5C90_I, 5C90_J, 5C90_K, 5C90_L, 5C90_M, 5C90_N
		200	397	99%	3.E-143	4EMP_V, 4EMP_A, 4EMP_B, 4EMP_C, 4EMP_E, 4EMP_F, 4EMP_G, 4EMP_I, 4EMP_K, 4EMP_L, 4EMP_M, 4EMP_N, 4EMP_S, 4EMP_T

Continuation of Table A.2

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
		195	399	99%	8.E-146	DKF_A, DKF_B, DKF_C, DKF_D, DKF_E, DKF_F, DKF_G, DKF_H, DKF_I, DKF_J, DKF_K, DKF_L, DKF_M, DKF_N,
		195	398	99%	1.E-145	4MXI_A, 4MXI_B, 4MXI_C, 4MXI_D, 4MXI_E, 4MXI_F, 4MXI_G,
DEG10170077	336	338	689	100%	0	3LVF_P, 3LVF_R, 3LVF_O, 3LVF_Q
		336	689	100%	0	3K73_Q, 3K73_O, 3K73_P, 3K73_R, 3L6O_Q, 3L6O_P, 3L6O_R, 3L6O_O, 3LC2_O, 3LC2_Q, 3LC2_R, 3LC2_P
		339	689	100%	0	3LC7_O, 3LC7_R, 3LC7_Q, 3LC7_P
		336	687	99%	0	3LC1_P, 3LC1_R, 3LC1_O, 3LC1_Q
		336	686	99%	0	3HQ4_R, 3HQ4_O, 3HQ4_P, 3HQ4_Q, 3KV3_O, 3KV3_Q, 3KV3_R, 3KV3_P
		336	685	99%	0	5T73_A, 5T73_B, 5T73_C, 5T73_D
		336	684	99%	0	3K9Q_Q, 3K9Q_O, 3K9Q_P, 3K9Q_R, 3L4S_Q, 3L4S_P, 3L4S_O, 3L4S_R
		334	689	100%	0	3VAZ_A, 3VAZ_B, 3VAZ_O, 3VAZ_P, 3VAZ_Q, 3VAZ_R,
DEG10170078	396	403	783	100%	0	4DG5_A
DEG10170079	253	254	518	99%	0	3M9Y_A, 3M9Y_B
		261	517	99%	0	3UWU_A, 3UWU_B, 3UWV_A, 3UWV_B, 3UWW_A, 3UWW_B, 3UWY_A, 3UWY_B, 3UWZ_A, 3UWZ_B
DEG10170080	505	513	1037	100%	0	4MY4_A, 4NWJ_A, 4NWX_A, 4QAX_A
DEG10170081	434	442	880	100%	0	5BOE_A, 5BOE_B, 5BOF_A, 5BOF_B
DEG10170094	124	127	256	100%	9.E-90	4M20_A, 4M20_B, 4M20_C, 4M20_D, 5EP5_A, 5EP5_B, 5EP5_C, 5EP5_D

Continuation of Table A.2						
Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
		127	253	99%	9.E-89	4YBV_A, 4YBV_B, 4YBV_C, 4YBV_D
DEG10170096	443	446	885	98%	0	3FF1_A, 3FF1_B
DEG10170098	313	313	632	99%	0	1ZOW_A, 1ZOW_B, 1ZOW_C, 1ZOW_D, 3IL7_A, 3IL7_B
DEG10170099	414	437	847	100%	0	2GQD_A, 2GQD_B
DEG10170104	256	260	517	99%	0	3GNS_A, 3GR6_A, 3GR6_D, 3GR6_G, 3GR6_J
		256	516	99%	0	3GNT_A, 3GNT_B
		277	516	99%	0	4ALL_A, 4ALL_B, 4ALL_C, 4ALL_D
		279	516	99%	0	4NZ9_A, 4NZ9_B
		256	515	99%	0	4FS3_A

Continuation of Table A.2

Essential gene	Query length	Protein length	Score	Identity	E-value	PDB
		282	514	99%	0	4ALI_A, 4ALI_B, 4ALI_C, 4ALI_D, 4ALI_E, 4ALI_F, 4ALI_G, 4ALI_H, 4ALJ_A, 4ALJ_B, 4ALJ_C, 4ALJ_D, 4ALJ_E, 4ALJ_F, 4ALJ_G, 4ALJ_H, 4ALK_A, 4ALK_B, 4ALK_C, 4ALK_D, 4ALK_E, 4ALK_F, 4ALK_G, 4ALK_H, 4ALM_A, 4ALM_B, 4ALM_C, 4ALM_D, 4ALN_A, 4ALN_B, 4ALN_C, 4ALN_D, 4ALN_E, 4ALN_F, 4ALN_G, 4ALN_H, 4ALN_I, 4ALN_J, 4ALN_K, 4ALN_L, 4BNF_A, 4BNF_B, 4BNF_C, 4BNF_D, 4BNF_E, 4BNF_F, 4BNF_G, 4BNF_H, 4BNG_A, 4BNG_B, 4BNG_C, 4BNG_D, 4BNG_E, 4BNG_F, 4BNG_G, 4BNG_H, 4BNH_A, 4BNH_B, 4BNH_C, 4BNH_D, 4BNH_E, 4BNH_F, 4BNH_G, 4BNH_H, 4BNI_A, 4BNI_B, 4BNI_C, 4BNI_D, 4BNI_E, 4BNI_F, 4BNI_G, 4BNI_H, 4BNJ_A, 4BNJ_B, 4BNJ_C, 4BNJ_D, 4BNJ_E, 4BNJ_F, 4BNJ_G, 4BNJ_H, 4BNK_A, 4BNK_B, 4BNK_C, 4BNK_D, 4BNK_E, 4BNK_F, 4BNK_G, 4BNK_H, 4BNL_A, 4BNL_B, 4BNL_C, 4BNL_D, 4BNL_E, 4BNL_F, 4BNL_G, 4BNL_H, 4BNM_A, 4BNM_B, 4BNM_C, 4BNM_D, 4BNM_E, 4BNM_F, 4BNM_G, 4BNM_H, 4BNN_A, 4BNN_B, 4BNN_C, 4BNN_D, 4BNN_E, 4BNN_F, 4BNN_G, 4BNN_H,

Continuation of Table A.2

Essential gene	Query length	Protein length	Score	Identity	E-value	PDB
		261	514	99%	0	6TBB_A, 6TBC_A, 6TBC_B, 6TBC_C, 6TBC_D, 6TBC_E, 6TBC_F, 6TBC_G, 6TBC_H,
DEG10170105	493	501	1014	99%	0	4C13_A, 4C12_A
DEG10170108	397	397	800	99%	0	5ZH8_A, NEW, 5ZH8_B
DEG10170109	88	88	175	100%	8E-59	1KA5_A
DEG10170112	183	183	375	99%	8E-135	1LQW_A, 1LQW_B, 2AI9_A, 2AI9_B
		191	371	99%	3.E-133	1Q1Y_A, 3U7K_A, 3U7L_A, 3U7M_A, 3U7N_A
		194	369	99%	1.E-132	1LM4_A, 1LM4_B
		184	359	99%	9.E-129	1LMH_A
		184	375	99%	1.E-136	6JFQ_A
		183	370	99%	4.E-135	6JFO_A
DEG10170116	160	160	330	100%	5E-118	4NAH_A, 4NAH_B, 4NAH_C, 4NAH_D, 4NAH_E, 4NAH_F, 4NAT_A, 4NAT_B, 4NAT_C, 4NAU_A, 4NAU_B, 4NAU_C
		168	330	100%	1.E-119	3F3M_A
DEG10170122	104	107	211	100%	2E-72	2O7K_A
		107	208	99%	2.E-71	2O85_A
		106	207	99%	4.E-71	3DIE_A, 3DIE_B
		106	206	99%	1.E-70	2O87_A
		107	205	98%	3.E-70	2O89_A
DEG10170123	266	286	545	99%	0	2JFQ_A, 2JFQ_B
DEG10170130	470	484	932	99%	0	3WQT_A, 3WQT_B, 3WQT_C, 3WQT_D, 3WQU_A, 3WQU_B, 3WQU_C, 3WQU_D
DEG10170131	390	392	778	100%	0	3VO8_A, 3VO8_B, 3WGN_A, 3WGN_B
		396	776	100%	0	4DXD_A
		390	776	99%	0	3WGL_A, 3WGL_B, 3WGM_A, 3WGM_B
		390	763	99%	0	3WGK_A, 3WGK_B
DEG10170133	917	917	1896	99%	0	1QU2_A, 1QU3_A, 1FFY_A

Continuation of Table A.2

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
DEG10170134	207	207	422	100%	1E-152	2J41_A, 2J41_B, 2J41_C, 2J41_D
		210	423	100%	6.E-155	4QRH_A, 4QRH_B, 4QRH_C, 4QRH_D
DEG10170143	308	316	621	99%	1E-152	3IM9_A
DEG10170144	244	252	495	100%	2E-180	3SJ7_A, 3SJ7_B
		246	47	97%	6.E-171	3OSU_A, 3OSU_B
DEG10170145	77	101	149	100%	2E-48	4DXE_H, 4DXE_L, 4DXE_K, 4DXE_G, 4DXE_J, 4DXE_I
DEG10170152	294	294	598	100%	0	6G14_A, 6G14_B
		301	596	100%	0	6G0Z_A, 6G0Z_B, 6G12_A, 6G12_B, 6G15_A, 6G15_B
DEG10170159	256	256	526	99%	0	4H8E_A, 4U82_A, 3WYI_A
DEG10170161	567	567	1161	99%	0	5ZNJ_A, 5ZNK_A
DEG10170180	61	63	125	100%	5E-40	2X4K_A, 2X4K_B
DEG10170181	420	426	863	100%	0	1LRZ_A
DEG10170184	159	159	325	100%	4.E-116	2W9G_A, 2W9H_A
		163	325	100%	4.E-116	4FGG_A, 4FGH_A, 6PR7_A, 6PRA_A
		167	325	100%	7.E-116	3SQY_X, 3SRQ_X, 3SRR_X, 3SRS_X, 3SRU_X, 3SRW_X, 4LAE_X, 4LAG_X, 4LAH_X, 4LEK_X
		160	323	100%	2.E-115	4XE6_X
		158	323	100%	3.E-115	3FYV_X, 3FYW_X, 3FRD_X, 3FRE_X, 3FRF_X
		160	323	99%	3.E-115	4Q67_A
		158	322	99%	5.E-115	3FY8_X, 3FY9_X, 3FRA_X, 3FRB_X
		166	322	100%	7.E-115	3SR5_X
		160	322	100%	1.E-114	4Q6A_A
		157	321	100%	1.E-114	3F0B_X, 3FQ0_A, 3FQC_A, 3FQC_B, 3FQZ_A, 3F0Q_X, 3F0S_X, 4TU5_X, 4XEC_X, 5HF2_X, 5JG0_X, 6ND2_X, 6P9Z_X, 6PBO_X

Continuation of Table A.2

Essential gene	Query length	Protein length	Score	Identity	E-value	PDB
		160	321	100%	2.E-114	5HF0_X
		161	321	100%	2.E-114	3M08_A
		157	320	99%	3.E-114	3F0U_X, 3FQF_A, 3FQO_A, 3FQV_A, 3F0V_X, 3F0X_X
		161	320	99%	4.E-114	3M09_A
		160	320	99%	4.E-114	5ISP_X
		167	320	100%	5.E-114	3SGY_A, 3SGY_B, 3SH2_A, 3SH2_B
		160	319	99%	2.E-113	5IST_X
		168	319	99%	2.E-113	3LG4_A, 3LG4_B
		157	318	99%	2.E-113	3I8A_X
		160	318	99%	4.E-113	5ISQ_X
		182	326	100%	5.E-118	6E4E_A
		162	323	100%	5.E-118	6PR9_A, 6PRB_A, 6PRD_A
DEG10170185	318	321	665	99%	0	4DQ1_A, 4DQ1_B
DEG10170190	323	323	658	99%	0	3RIR_A, 3RKW_A, 3RKX_A, 3RKY_A
		329	657	99%	0	3V7C_A, 3V7S_A, 6AQQ_A, 6ORU_A
		329	657	99%	0	3V7R_A, 3V8J_A, 3V8K_A, 3V8L_A
		328	655	99%	0	4HA8_A
		328	655	99%	0	4DQ2_A
		329	657	99%	0	6APW_A
		330	657	99%	0	6NDL_A
DEG10170193	90	98	180	100%	1.E-60	4QJN_A, 4QJN_B, 4QJN_C, 4QJN_D, 4QJU_A, 4QJU_B
DEG10170195	219	219	442	100%	2.E-160	2H92_A, 2H92_B, 2H92_C
DEG10170202	451	451	929	100%	0	VPQ_A, 2VPQ_B
DEG10170204	185	191	379	100%	2.E-138	6RJI_A
		185	379	100%	3.E-138	6RK3_A
DEG10170217	189	189	392	100%	1E-141	2H29_A, 2H29_B
		189	388	99%	7.E-140	2H2A_A, 2H2A_B
DEG10170228	420	420	866	99%	0	1QE0_A, 1QE0_B

Continuation of Table A.2

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
DEG10170236	94	94	186	100%	3.E-63	4WCE_T, 4WF9_T, 4WFA_T, 4WFB_T, 5HL7_T, 5HKV_T, 5NRG_T, 6SJ6_Z
DEG10170237	106	106	215	100%	3E-74	4PEO_A, 4PEO_B
DEG10170238	102	102	202	100%	3.E-69	4WCE_O, 4WF9_O, 4WFA_O, 4WFB_O, 5HL7_O, 5HKV_O, 5NRG_O, 6DDD_D, 6DDG_D, 6HMA_P, 6SJ6_U
		105	203	100%	2.E-71	6WQN_D, 6WQQ_D, 6WRS_D, 6WRU_D
DEG10170244	118	118	238	100%	4.E-83	4WCE_N, 4WF9_N, 4WFA_N, 4WFB_N, 5HL7_N, 5HKV_N, 5NRG_N, 6DDD_C, 6DDG_C, 6SJ6_T, 6WQN_C, 6QQQ_C, WRS_C, 6WRU_C
		116	235	100%	1.E-83	6HMA_O
DEG10170245	66	66	127	100%	9.E-41	4WCE_3, 4WF9_3, 4WFA_3, 4WFB_3, 5HL7_3, 5HKV_3, 5NRG_3, 6SJ6_7
		65	126	100%	4.E-42	6DDD_Q, 6DDG_Q, 6WQN_Q, 6WQQ_Q, 6WRS_Q, 6WRU_Q
		54	124	100%	3.E-41	6HMA_3
DEG10170248	645	645	1337	99%	0	1NYQ_A, 1NYQ_B, 1NYR_A, 1NYR_B
DEG10170252	585	606	1170	99%	0	3T05_A, 3T05_B, 3T05_C, 3T05_D, 3T07_A, 3T07_B, 3T07_C, 3T07_D, 3T0T_A, 3T0T_B, 3T0T_C, 3T0T_D
DEG10170253	307	322	623	100%	0	5XOE_A
		330	622	100%	0	5XZ6_A, 5XZ7_A, 5XZ8_A, 5XZ9_A, 5XZA_A
DEG10170254	314	327	640	100%	0	2F9I_A, 2F9I_C, 5KDR_A
DEG10170255	285	285	593	100%	0	2F9I_B, 2F9I_D, 5KDR_B
DEG10170260	420	420	866	100%	0	1JII_A, 1JIJ_A, 1JIK_A, 1JIL_A
DEG10170270	252	252	517	100%	0	1QXW_A, 1QXY_A, 1QXZ_A
DEG10170271	243	243	494	100%	0	5N9M_A, 5N9M_B

Continuation of Table A.2

Essential gene	Query length	Protein length	Score	Identity	E-value	PDB
		251	494	100%	0	6GS2_A, 6GS2_C, 6H5E_A, 6H5E_C
DEG10170272	437	437	909	100%	0	6GS2_B, 6GS2_D, 6H5E_B, 6H5E_D
DEG10170273	437	437	909	100%	0	6GS2_B, 6GS2_D, 6H5E_B, 6H5E_D
DEG10170274	475	483	978	100%	0	2DF4_B, 2DQN_B, 2F2A_B, 2G5H_B, 2G5I_B, 3IP4_B
DEG10170275	485	485	982	99%	0	2DF4_A, 2DQN_A, 2G5H_A, 2G5I_A, 3IP4_A, 2F2A_A
DEG10170276	100	100	202	100%	3.E-69	2DF4_C, 2DQN_C, 2F2A_C, 2G5H_C, 2G5I_C, 3IP4_C
DEG10170281	309	317	624	99%	0	4RPA_A, 4RPA_B
DEG10170290	119	122	247	100%	3.E-86	5CXD_A, 5CXD_B, 5CXD_C
		143	246	100%	9.E-86	4DXE_A, 4DXE_B, 4DXE_C, 4DXE_D, 4DXE_E, 4DXE_F, 4JM7_A, 4JM7_B, 4JM7_C
DEG10170292	356	360	735	100%	0	2I80_A, 2I80_B
		364	735	100%	0	2I87_A, 2I87_B, 2I8C_A, 2I8C_B
		364	730	99%	0	3N8D_A, 3N8D_B
DEG10170299	286	292	581	100%	0	4TO8_A, 4TO8_B
DEG10170301	267	287	546	100%	0	2EWS_A, 2EWS_B
		285	546	100%	0	4NB4_A, 4NB4_B, 4NB4_C, 4NB4_D, 4NB4_E, 4NB4_F, 4NB4_G, 4NB4_H
		273	546	100%	0	4M7X_A, 4M7Y_A, 4M7Y_B, 5ELZ_A, 5JIC_A
		267	546	100%	0	6EBV_A, 6EBV_A, 6EBV_C, 6EBV_D
		266	545	100%	0	6AWG_A, 6AWG_B, 6AWG_C, 6AWG_D, 6AWH_A, 6AWH_B, 6AWH_C, 6AWH_D, 6AWI_A, 6AWI_B, 6AWI_C, 6AWI_D, 6AWJ_A, 6AWJ_B, 6AWJ_C, 6AWJ_D
		265	543	100%	0	6AVP_A, 6AVP_B, 6AVP_C, 6AVP_D
DEG10170303	451	455	917	100%	0	6GYZ_A, 6GYZ_B

Continuation of Table A.2

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
DEG10170306	145	145	301	100%	5.E-107	4WCE_G, 4WF9_G, 4WFA_G, 4WFB_G, 5HL7_G, 5HKV_G, 5NRG_G, 6DDD_V, 6DDG_V, 6HMA_H, 6SJ6_M, 6WQN_V, 6WQQ_V, 6WRS_V, 6WRU_V
DEG10170307	122	122	239	100%	4.E-83	S 4WCE_K, 4WF9_K, 4WFA_K, 4WFB_K, 5HL7_K, 5HKV_K, 5NRG_K, 6DDD_Z, 6DDG_VZ, 6SJ6_Q, 6WQN_Z, 6WQQ_Z, 6WRS_Z, 6WRU_Z
		120	234	100%	4.E-83	6HMA_L
DEG10170311	37	37	72.8	100%	1.E-19	4WCE_4, 4WFA_4, 4WFB_4, 5HL7_4, 6DDD_R, 6DDG_R, 6HMA_4, 6WQN_R, 6WQQ_R, 6WRS_R, 6WRU_R
DEG10170315	146	146	289	100%	3.E-102	4WCE_I, 4WF9_I, 4WFA_I, 4WFB_I, 5HKV_I, 6DDD_X, 6DDG_X, 6HMA_J, 6SJ6_O, 6WQN_X, 6WQQ_X, 6WRS_X, 6WRU_X
		140	278	100%	9.E-98	5HL7_I, 5NRG_I
DEG10170316	59	59	117	100%	9.E-37	4WCE_W, 4WF9_W, 4WFA_W, 4WFB_W, 5HL7_W, 5HKV_W, 5NRG_W, 6DDD_M, 6DDG_M, 6SJ6_2
		58	114	100%	1.E-37	6HMA_X, 6WQN_M, 6WQQ_M, 6WRS_M, 6WRU_M
DEG10170318	119	119	238	100%	6.E-83	4WCE_L, 4WF9_L, 4WFA_L, 4WFB_L, 5HL7_L, 5HKV_L, 5NRG_L, 6DDD_a, 6DDG_a, 6HMA_M, 6SJ6_R, 6WQN_a, 6WQQ_a, 6WRS_a, 6WRU_a
DEG10170319	178	178	358	100%	1.E-128	4WCE_E, 4WF9_E, 4WFA_E, 4WFB_E, 5HKV_E, 5NRG_E

Continuation of Table A.2

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
		178	355	99%	2.E-127	5HL7_E
		175	352	100%	5.E-128	6HMA_G
		175	351	99%	1.E-127	6WRU_U
DEG10170322	179	179	360	100%	3.E-129	4WCE_D, 4WF9_D, 4WFA_D, 4WFB_D, 5HL7_D, 5HKV_D, 5NRG_D
		158	318	100%	5.E-115	6HMA_F
DEG10170323	105	105	206	100%	7.E-71	4WCE_R, 4WF9_R, 4WFA_R, 4WFB_R, 5HL7_R, 5HKV_R, 5NRG_R, 6DDD_R, 6DDG_R, 6SJ6_X
		105	205	99%	5.E-72	6WQN_G, 6WQQ_G, 6WRS_G, 6WRU_G
		103	202	100%	6.E-71	6HMA_S
DEG10170324	122	122	241	100%	8.E-84	4WCE_H, 4WF9_H, 4WFA_H, 4WFB_H, 5HL7_H, 5HKV_H, 5NRG_H, 6DDD_W, 6DDG_W, 6HMA_I, 6SJ6_N, 6WQN_W, 6WQQ_W, 6WRS_W, 6WRU_W
		142	243	100%	2.E-86	6SJ5_C, 6SJ5_D
DEG10170326	69	69	132	100%	3.E-42	4WCE_V, 4WF9_V, 4WFA_V, 4WFB_V, 5HL7_V, 5HKV_V, 5NRG_V, 6DDD_K, 6DDG_K, 6SJ6_1
		72	127	99%	2.E-42	6WQN_K, 6WQQ_K, 6WRS_K, 6WRU_K
		67	127	100%	3.E-42	6HMA_W
DEG10170327	144	144	290	100%	2.E-102	4WCE_J, 4WF9_J, 4WFA_J, 4WFB_J, 5HL7_J, 5HKV_J, 5NRG_J, 6DDD_Y, 6DDG_Y, , 6SJ6_P, 6WQN_Y, 6WQQ_Y, 6WRS_Y, 6WRU_Y
		137	275	100%	7.E-99	6HMA_K

Continuation of Table A.2

Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
DEG10170329	117	117	233	100%	1.E-80	4WCE_P, 4WF9_P, 4WFA_P, 4WFB_P, 5HL7_P, 5HKV_P, 5NRG_P, 6SJ6_V
		116	231	100%	5.E-82	6DDD_E, 6DDG_E
		117	230	99%	1.E-81	6WQN_E, 6WQQ_E, 6WRS_E, 6WRU_E
DEG10170331	277	277	556	100%	0	4WCE_A, 4WF9_A, 4WFA_A, 4WFB_A, 5HL7_A, 5HKV_A, 5NRG_A, 6SJ6_D, 6WQN_B, 6WQQ_B, 6WRS_B, 6WRU_B
		276	555	100%	0	6DDD_B, 6DDG_B
		274	550	100%	0	6HMA_C
DEG10170332	91	91	184	100%	2.E-62	4WCE_Q, 4WF9_Q, 4WFA_Q, 4WFB_Q, 5HL7_Q, 5HKV_Q, 5NRG_Q, 6DDD_F, 6DDG_F, 6SJ6_W
		91	181	99%	4.E-63	6WQN_F, 6WQQ_F, 6WRS_F, 6WRU_F
		89	180	100%	1.E-62	6HMA_R
DEG10170333	207	207	421	100%	3.E-152	4WCE_C, 4WF9_C, 4WFA_C, 4WFB_C, 5HL7_C, 5HKV_C, 5NRG_C, 6SJ6_F
		207	419	99%	2.E-153	6DDD_S, 6DDG_S, 6WQN_S, 6WQQ_S, 6WRS_S, 6WRU_S
		207	418	100%	4.E-153	6HMA_E
DEG10170334	220	220	441	100%	9.E-160	4WCE_B, 4WF9_B, 4WFA_B, 4WFB_B, 5HL7_B, 5HKV_B, 5NRG_B, 6SJ6_E
		217	435	100%	2.E-159	6DDD_L, 6WQN_L, 6WQQ_L, 6WRS_L, 6WRU_L
		215	531	100%	8.E-158	6HMA_D
DEG10170343	130	130	266	100%	2E-93	5X1X_A
DEG10170346	388	388	806	100%	0	1XPK_C
		388	801	99%	0	1XPK_B
		388	799	99%	0	1XPK_A, 1XPK_D
		388	798	99%	0	1TVZ_A, 1TXT_A, 1TXT_B, 1TXT_C, 1TXT_D

Continuation of Table A.2						
Essential gene	**Query length**	**Protein length**	**Score**	**Identity**	**E-value**	**PDB**
		390	796	99%	0	1XPL_A, 1XPL_B, 1XPL_C, 1XPL_D, 1XPM_A, 1XPM_B, 1XPM_C, 1XPM_D
DEG10170350	117	117	234	100%	3E-81	1D6T_A
		119	229	98%	0	2.E-81, 6D1R_A
DEG10170350	45	45	87.4	100%	3.E-25	4WCE_2, 4WF9_2, 4WFA_2, 4WFB_2, 5HL7_2, 5HKV_2, 5NRG_2, 6DDG_P, 6SJ6_6
		50	87.4	100%	2.E-27	6DDD_P, 6WQN_P, 6WQQ_P, 6WRS_P, 6WRU_P
		43	82.8	100%	2.E-25	6HMA_2

3 Essentian genes hit ribosomal proteins

Table A.3: List of essential genes that hit ribosomal proteins.
List of ribosomal protein chains matching essential genes using both BLAST+ and SWISS-MODEL template searching. The numbers and figures after the protein entry represent chains of ribosomal component.

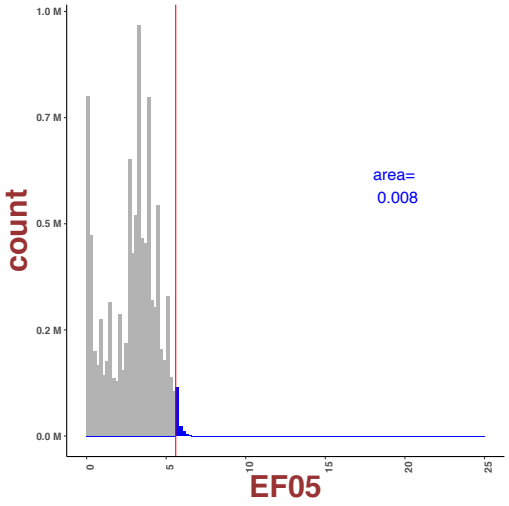
DEG ID	**Ribosome**	**Ligands**
DEG10170017	30S ribosomal protein S6	5ND8_F, 5ND9_F, 5NGM_F, 5LI0_F, 5NG8_A.11, 5TCU_E, 5T7V_L
DEG10170019	30S ribosomal protein S18	5ND8_R, 5ND9_R, 5NGM_R, 5LI0_R
DEG10170028	50S ribosomal protein L25	4WCE_S, 4WF9_S, 4WFA_S, 4WFB_S, 5HL7_S, 5HKV_S, 5NRG_S, 6SJ6_Y, 6WQN_H, 6WQQ_H, 6WRS_H, 6WRU_H
DEG10170046	30S ribosomal protein S12	5ND8_L, 5ND9_L, 5NGM_L, 5TCU_K, 5T7V_D, 5NG8_A.11, 5LI0_L
DEG10170141	50S ribosomal protein L28	4WCE_U, 4WFA_U, 4WFB_U, 5HL7_U, 5HKV_U, 6DDJ_J, 6DDG_J, 6SJ6_O, 6WQN_J
DEG10170148	30S ribosomal protein S16	5ND8_P, 5ND9_P, 5NGM_P, 5LI0_P, 5NG8_A.15, 5TCU_O, 5T7V_F
DEG10170155	30S ribosomal protein S2	5ND8_B, 5ND9_B, 5NGM_B, 5LI0_B
DEG10170168	30S ribosomal protein S15	5ND8_O, 5ND9_O, 5NGM_O, 5LI0_O, 5NG8_A.63, 5TCU_N, 5T7V_E
DEG10170212	30S ribosomal protein S21	5ND8_U, 5ND9_U, 5NGM_U, 5NG8_A.20
DEG10170236	50S ribosomal protein L27	4WCE_T, 4WF9_T, 4WFA_T, 4WFB_T, 5HL7_T, 5HKV_T, 5NRG_T, 6SJ6_Z, 6DDD_I, 6DDG_I, 6WQN_I, 6WQQ_I, 6WRS_I, 6WRU_I, 6HMA_U
DEG10170238	50S ribosomal protein L21	4WCE_O, 4WF9_O, 4WFA_O, 4WFB_O, 5HL7_O, 5HKV_O, 5NRG_O, 6DDD_D, 6DDG_D, 6HMA_P, 6SJ6_U, 6WQN_D, 6WQQ_D, 6WRS_D, 6WRU_D

Continuation of Table A.3		
DEG ID	**Ribosome**	**Ligands**
DEG10170244	50S ribosomal protein L20	4WCE_N, 4WF9_N, 4WFA_N, 4WFB_N, 5HL7_N, 5HKV_N, 5NRG_N, 6DDD_C, 6DDG_C, 6SJ6_T, 6WQN_C, 6QQQ_C, WRS_C, 6WRU_C, 6HMA_O
DEG10170245	50S ribosomal protein L35	4WCE_3, 4WF9_3, 4WFA_3, 4WFB_3, 5HL7_3, 5HKV_3, 5NRG_3, 6SJ6_7, 6DDD_Q, 6DDG_Q, 6WQN_Q, 6WQQ_Q, 6WRS_Q, 6WRU_Q, 6HMA_3
DEG10170258	30S ribosomal protein S4	5ND8_D, 5ND9_D, 5NGM_D, 5LI0_D, 5NG8_A.3, 5TCU_C, 5T7V_J
DEG10170298	50S ribosomal protein L31	5ND8_K, 5ND9_K, 5NGM_K, 5LI0_I, 5NG8_A.44, 5TCU_P
DEG10170305	30S ribosomal protein S9	5ND8_L, 5ND9_L, 5NGM_L, 5LI0_L, 5NG8_A.57, 5TCU_H
DEG10170306	50S ribosomal protein L13	4WCE_G, 4WF9_G, 4WFA_G, 4WFB_G, 5HL7_G, 5HKV_G, 5NRG_G, 6DDD_V, 6DDG_V, 6HMA_H, 6SJ6_M, 6WQN_V, 6WQQ_V, 6WRS_V, 6WRU_V
DEG10170307	50S ribosomal protein L17	4WCE_K, 4WF9_K, 4WFA_K, 4WFB_K, 5HL7_K, 5HKV_K, 5NRG_K, 6DDD_Z, 6DDG_VZ, 6SJ6_Q, 6WQN_Z, 6WQQ_Z, 6WRS_Z, 6WRU_Z, 6HMA_L
DEG10170309	30S ribosomal protein S11	5ND8_K, 5ND9_K, 5NGM_K, 5LI0_K, 5NG8_A.59, 5TCU_J, 5T7V_C
DEG10170310	30S ribosomal protein S13	5ND8_M, 5ND9_M, 5LI0_M, 5NG8_A.12, 5TCU_L
DEG10170311	50S ribosomal protein L36	4WCE_4, 4WFA_4, 4WFB_4, 5HL7_4, 6DDD_R, 6DDG_R, 6HMA_4, 6WQN_R, 6WQQ_R, 6WRS_R, 6WRU_R
DEG10170315	50S ribosomal protein L15	4WCE_I, 4WF9_I, 4WFA_I, 4WFB_I, 5HKV_I, 6DDD_X, 6DDG_X, 6HMA_J, 6SJ6_O, 6WQN_X, 6WQQ_X, 6WRS_X, 6WRU_X, 5HL7_I, 5NRG_I
DEG10170316	50S ribosomal protein L30	4WCE_W, 4WF9_W, 4WFA_W, 4WFB_W, 5HL7_W, 5HKV_W, 5NRG_W, 6DDD_M, 6DDG_M, 6SJ6_2, 6HMA_X, 6WQN_M, 6WQQ_M, 6WRS_M, 6WRU_M
DEG10170317	30S ribosomal protein S5	5ND8_E, 5ND9_E, 5NGM_E, 5LI0_E, 5NG8_A.53, 5TCU_D, 5T7V_K
DEG10170318	50S ribosomal protein L18	4WCE_L, 4WF9_L, 4WFA_L, 4WFB_L, 5HL7_L, 5HKV_L, 5NRG_L, 6DDD_a, 6DDG_a, 6HMA_M, 6SJ6_R, 6WQN_a, 6WQQ_a, 6WRS_a, 6WRU_a
DEG10170319	50S ribosomal protein L6	4WCE_E, 4WF9_E, 4WFA_E, 4WFB_E, 5HKV_E, 5NRG_E, 5HL7_E, 6HMA_G, 6WRU_U
DEG10170320	30S ribosomal protein S8	5ND8_H, 5ND9_H, 5NGM_H, 5LI0_H, 5NG8_A.56, 5TCU_G
DEG10170321	30S ribosomal protein S14	5ND8_N, 5ND9_N, 5NGM_N, 5LI0_N, 5NG8_A.62, 5TCU_M
DEG10170322	50S ribosomal protein L5	4WCE_D, 4WF9_D, 4WFA_D, 4WFB_D, 5HL7_D, 5HKV_D, 5NRG_D, 6HMA_F

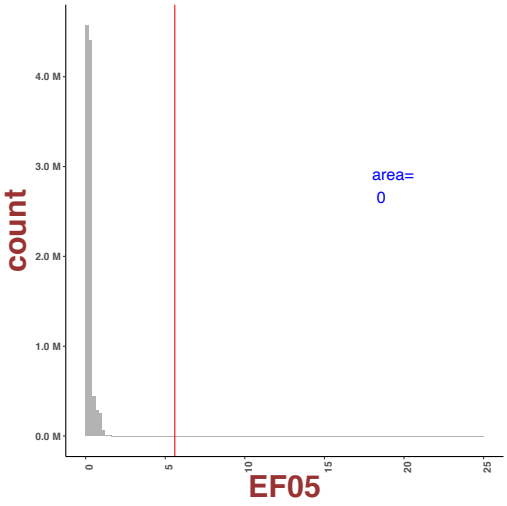
Continuation of Table A.3

DEG ID	Ribosome	Ligands
DEG10170323	50S ribosomal protein L24	4WCE_R, 4WF9_R, 4WFA_R, 4WFB_R, 5HL7_R, 5HKV_R, 5NRG_R, 6DDD_R, 6DDG_R, 6SJ6_X, 6WQN_G, 6WQQ_G, 6WRS_G, 6WRU_G, 6HMA_S
DEG10170324	50S ribosomal protein L14	4WCE_H, 4WF9_H, 4WFA_H, 4WFB_H, 5HL7_H, 5HKV_H, 5NRG_H, 6DDD_W, 6DDG_W, 6HMA_I, 6SJ6_N, 6WQN_W, 6WQQ_W, 6WRS_W, 6WRU_W, 6SJ5_C, 6SJ5_D
DEG10170325	30S ribosomal protein S17	5ND8_Q, 5ND9_Q, 5NGM_Q, 5LI0_Q
DEG10170326	50S ribosomal protein L29	4WCE_V, 4WF9_V, 4WFA_V, 4WFB_V, 5HL7_V, 5HKV_V, 5NRG_V, 6DDD_K, 6DDG_K, 6SJ6_1, 6WQN_K, 6WQQ_K, 6WRS_K, 6WRU_K, 6HMA_W
DEG10170327	50S ribosomal protein L16	4WCE_J, 4WF9_J, 4WFA_J, 4WFB_J, 5HL7_J, 5HKV_J, 5NRG_J, 6DDD_Y, 6DDG_Y, 6HMA_K, 6SJ6_P, 6WQN_Y, 6WQQ_Y, 6WRS_Y, 6WRU_Y
DEG10170328	30S ribosomal protein S3	5ND8_C, 5ND9_C, 5NGM_C, 5LI0_C, 5NG8_A.51, 5TCU_B
DEG10170329	50S ribosomal protein L22	4WCE_P, 4WF9_P, 4WFA_P, 4WFB_P, 5HL7_P, 5HKV_P, 5NRG_P, 6SJ6_V, 6DDD_E, 6DDG_E, 6WQN_E, 6WQQ_E, 6WRS_E, 6WRU_E, 6HMA_Q
DEG10170330	30S ribosomal protein S19	5ND8_S, 5ND9_S, 5NGM_S, 5LI0_S, 5NG8_A.18, 5TCU_R
DEG10170331	50S ribosomal protein L2	4WCE_A, 4WF9_A, 4WFA_A, 4WFB_A, 5HL7_A, 5HKV_A, 5NRG_A, 6SJ6_D, 6WQN_B, 6WQQ_B, 6WRS_B, 6WRU_B, 6DDD_B, 6DDG_B, 6HMA_C
DEG10170332	50S ribosomal protein L23	4WCE_Q, 4WF9_Q, 4WFA_Q, 4WFB_Q, 5HL7_Q, 5HKV_Q, 5NRG_Q, 6DDD_F, 6DDG_F, 6SJ6_W, 6WQN_F, 6WQQ_F, 6WRS_F, 6WRU_F, 6HMA_R
DEG10170333	50S ribosomal protein L4	4WCE_C, 4WF9_C, 4WFA_C, 4WFB_C, 5HL7_C, 5HKV_C, 5NRG_C, 6SJ6_F, 6DDD_S, 6DDG_S, 6WQN_S, 6WQQ_S, 6WRS_S, 6WRU_S, 6HMA_E
DEG10170334	50S ribosomal protein L3	4WCE_B, 4WF9_B, 4WFA_B, 4WFB_B, 5HL7_B, 5HKV_B, 5NRG_B, 6SJ6_E, 6DDD_L, 6WQN_L, 6WQQ_L, 6WRS_L, 6WRU_L, 6HMA_D
DEG10170351	50S ribosomal protein L34	4WCE_2, 4WF9_2, 4WFA_2, 4WFB_2, 5HL7_2, 5HKV_2, 5NRG_2, 6DDG_P, 6SJ6_6, 6DDD_P, 6WQN_P, 6WQQ_P, 6WRS_P, 6WRU_P, 6HMA_2

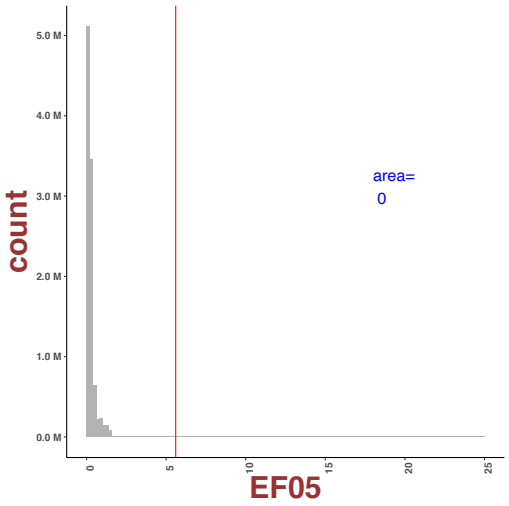
4 Histogram of Consensus Models using EF05 as Evaluation Metric



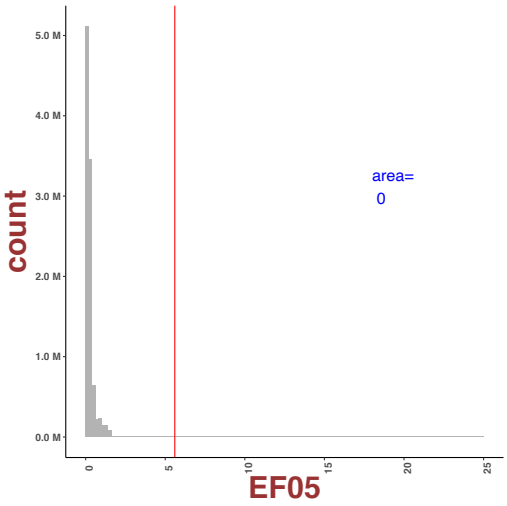
(a) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} S_{i,j}$



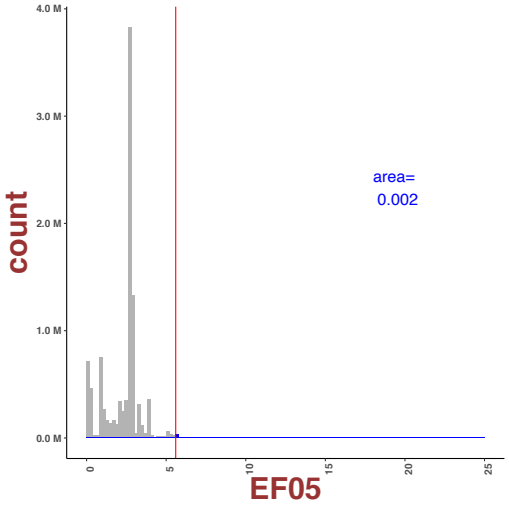
(b) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[S_{i,j}]$



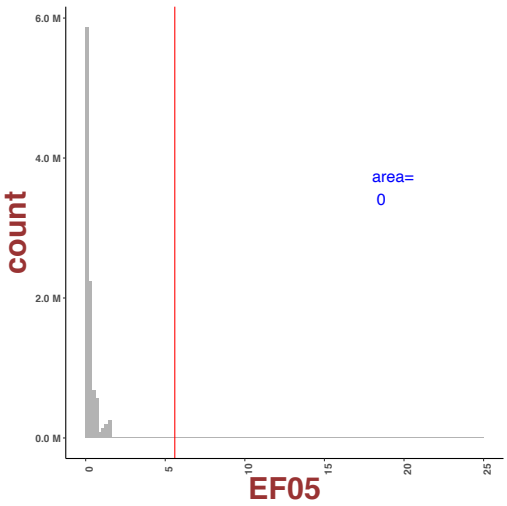
(c) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} {S_{i,j}}^2$



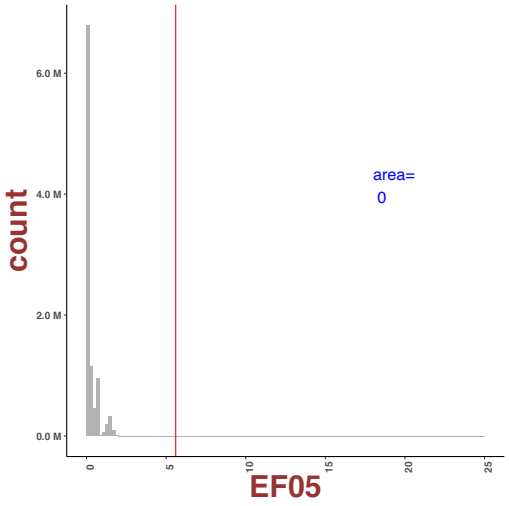
(d) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[{S_{i,j}}^2]$



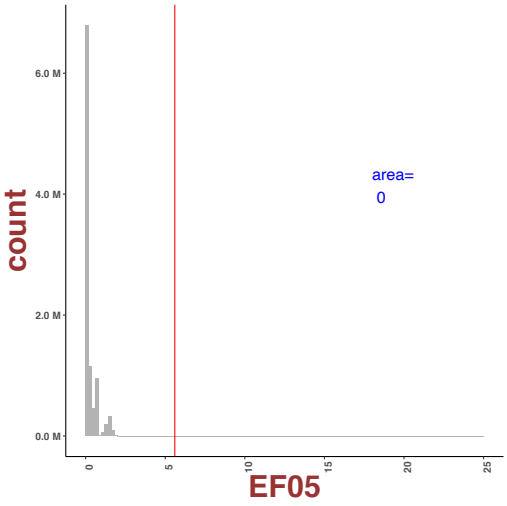
(e) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} {S_{i,j}}^3$



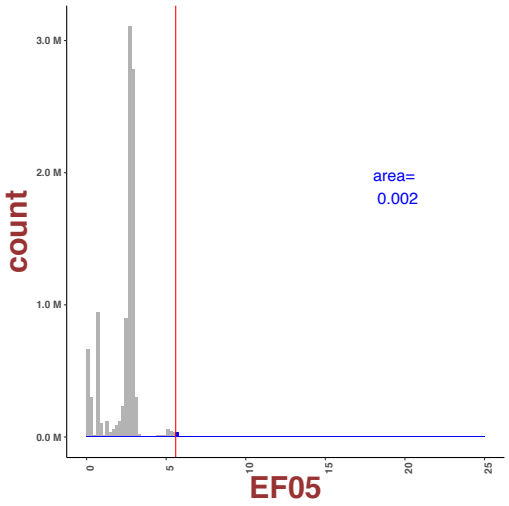
(f) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[{S_{i,j}}^3]$



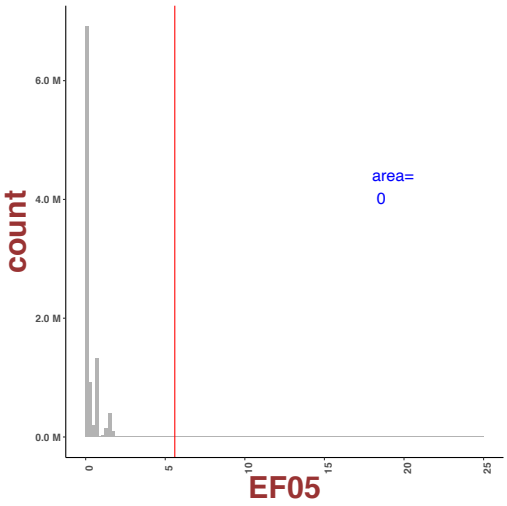
(g) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} {S_{i,j}}^4$



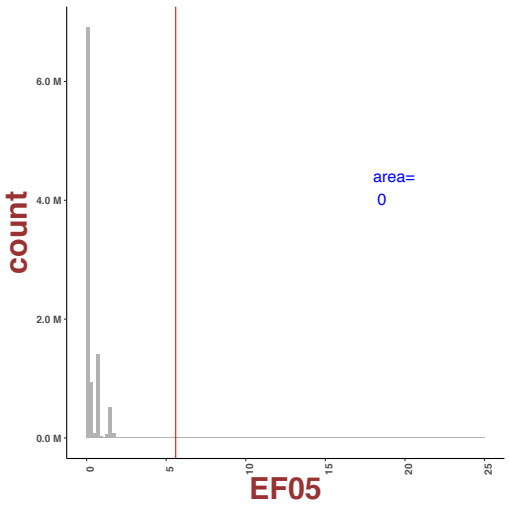
(h) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[{S_{i,j}}^4]$



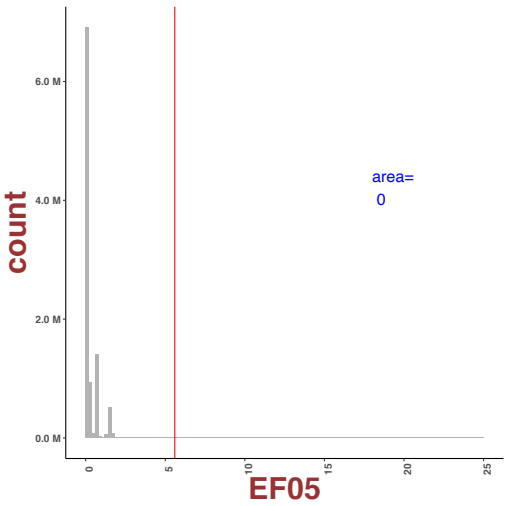
(i) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} {S_{i,j}}^5$



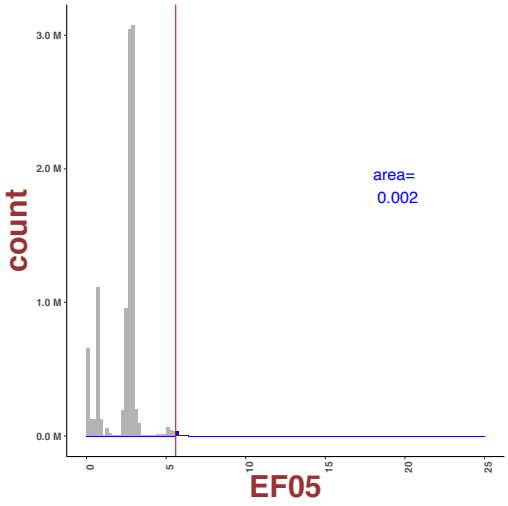
(j) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[{S_{i,j}}^5]$



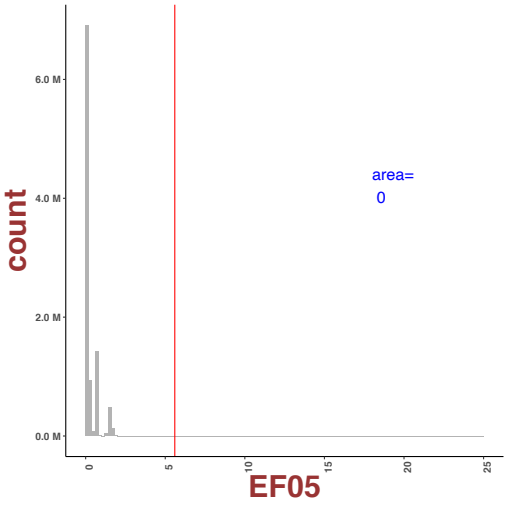
(k) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} {S_{i,j}}^6$



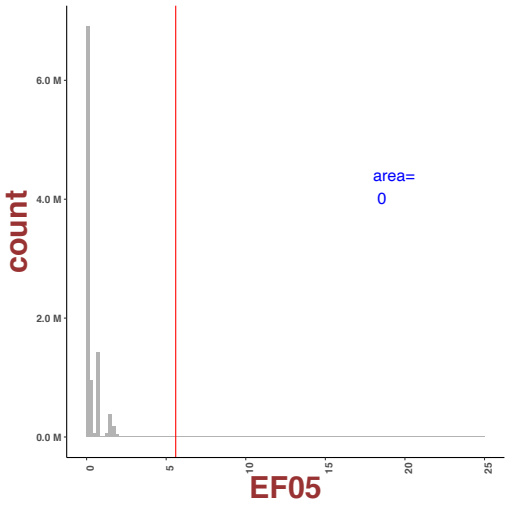
(l) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[{S_{i,j}}^6]$



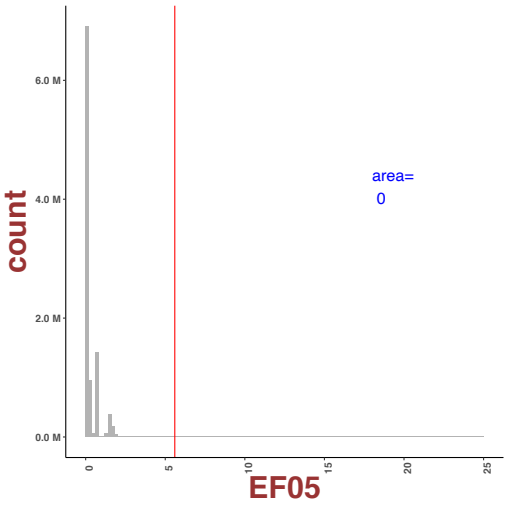
(m) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} {S_{i,j}}^7$



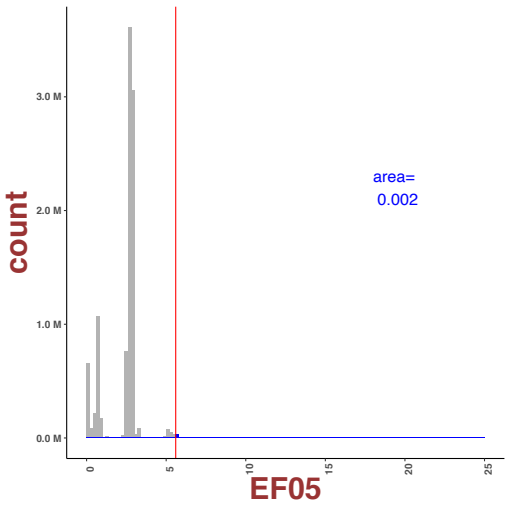
(n) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \text{ abs}[{S_{i,j}}^7]$



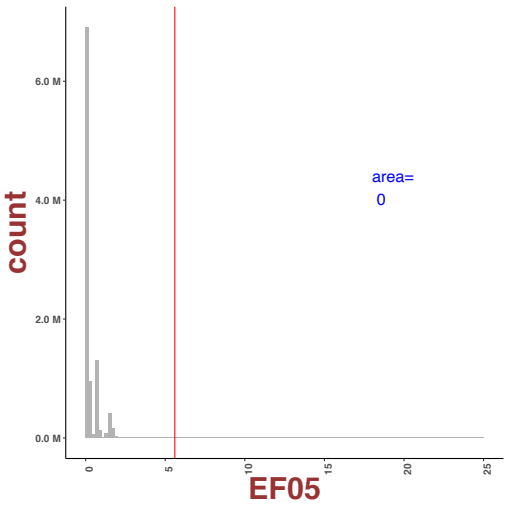
(o) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} {S_{i,j}}^8$



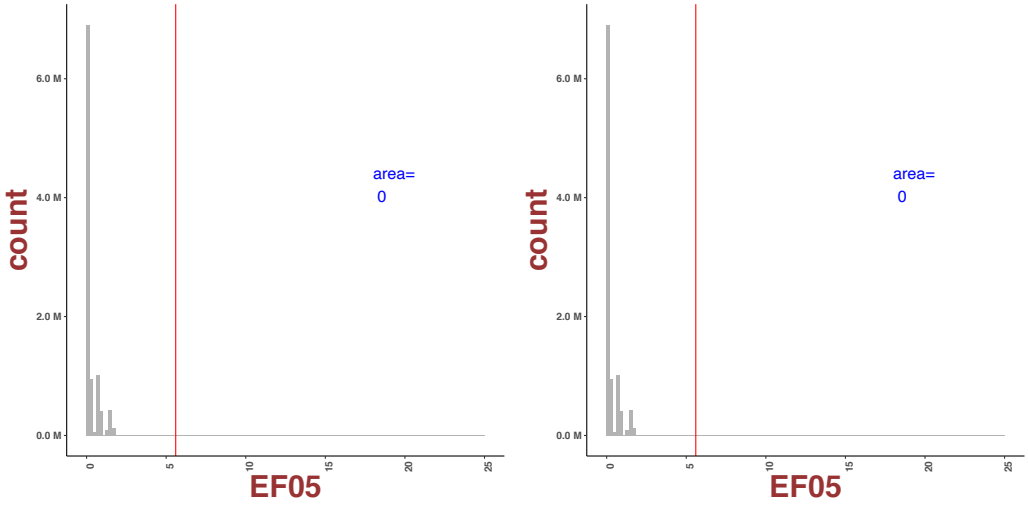
(p) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \text{ abs}[{S_{i,j}}^8]$



(q) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} {S_{i,j}}^9$



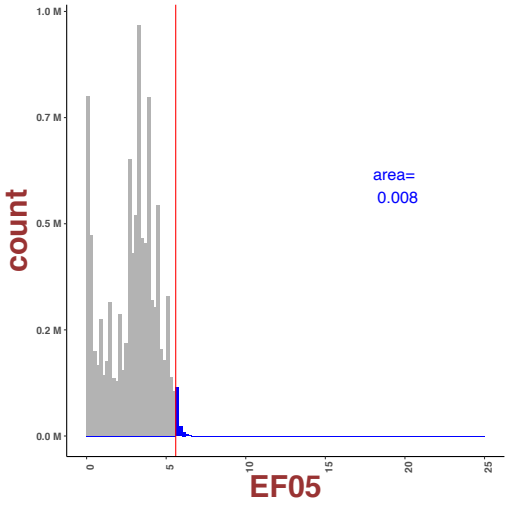
(r) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \text{ abs}[{S_{i,j}}^9]$



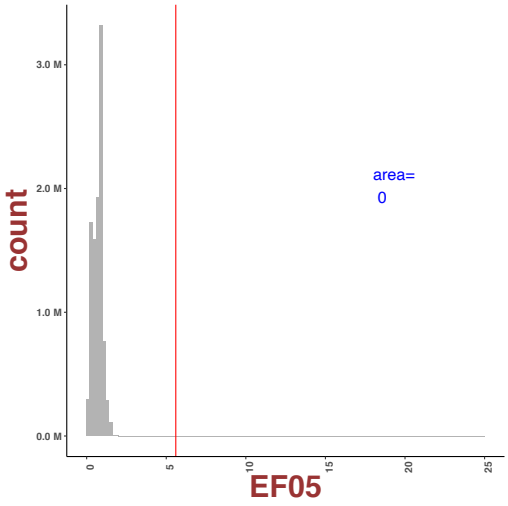
(s) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} {S_{i,j}}^1 0$

(t) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j}\ \mathrm{abs}[{S_{i,j}}^1 0]$

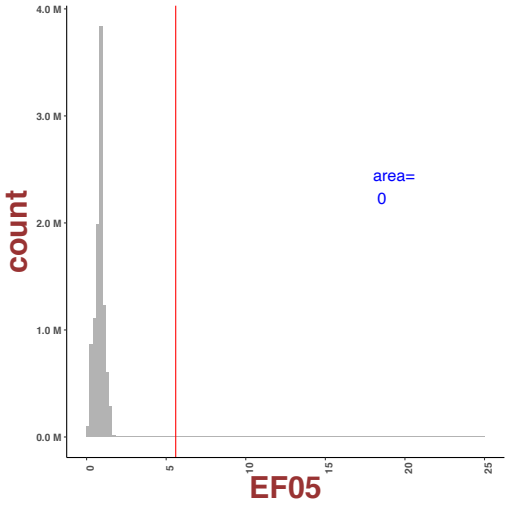
Figure A.1: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with a straight line) of the total histogram area, evaluated for order ranging from 1 to 10 as in Eqns (4.16a (left)-4.16b (right)).



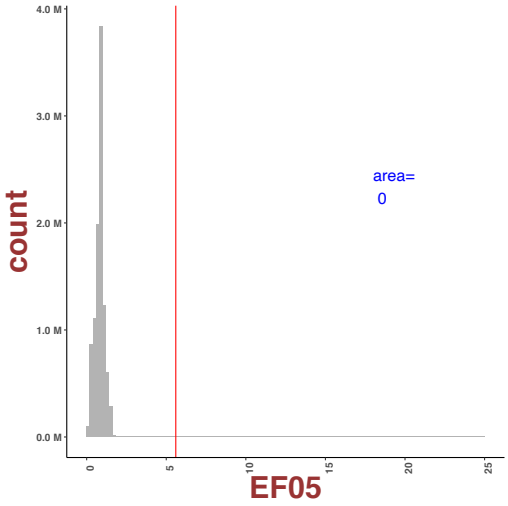
(a) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_{i,j} - \bar{S_i})$



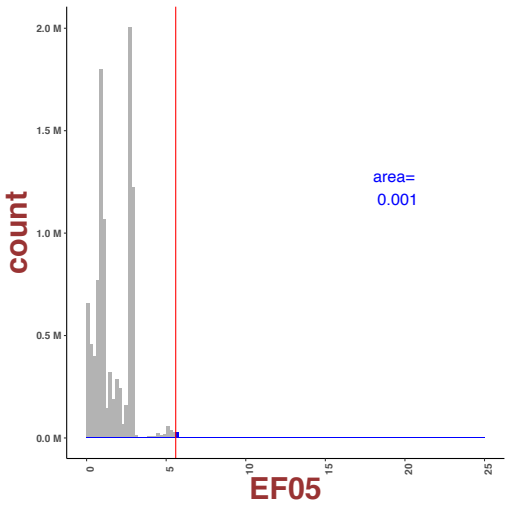
(b) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \text{abs}[S_{i,j} - \bar{S_i}]$



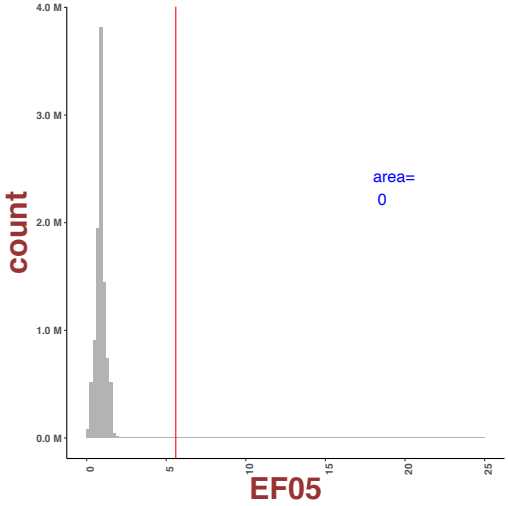
(c) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - \bar{S_{i,j}})^2$



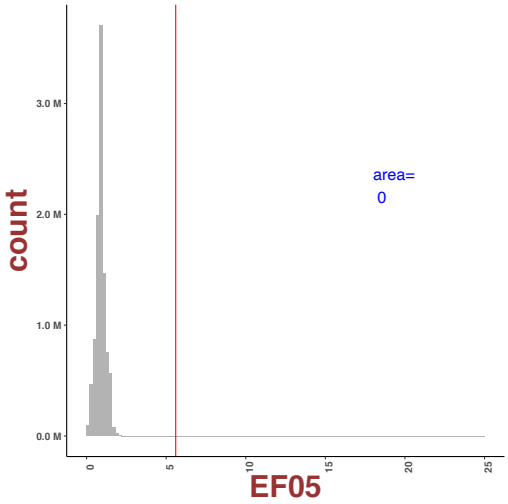
(d) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \text{abs}[(S_{i,j} - \bar{S_i})^2]$



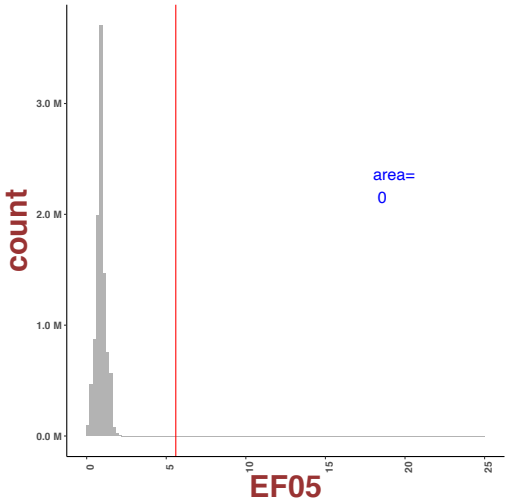
(e) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - \bar{S_{i,j}})^3$



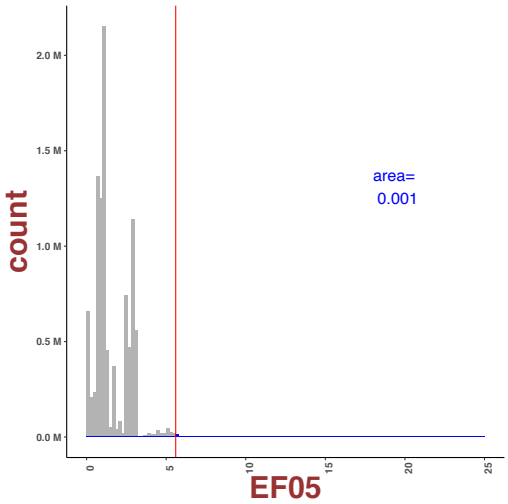
(f) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \text{abs}[(S_{i,j} - \bar{S_i})^3]$



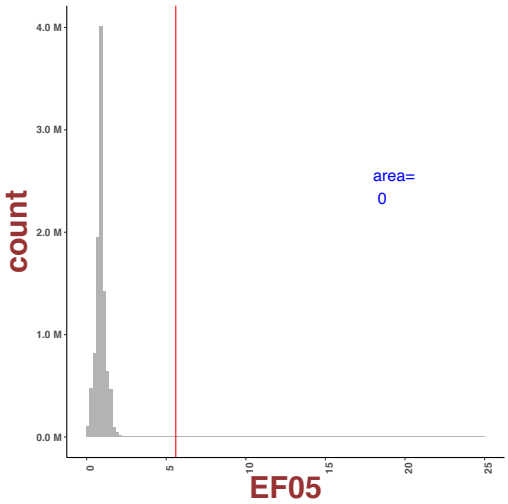
(g) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S_{i,j}^{-})^4$



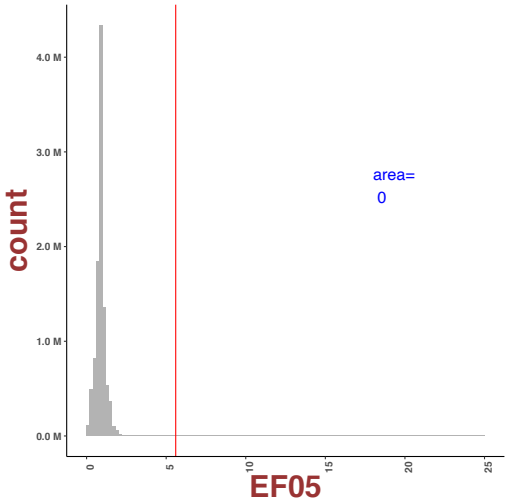
(h) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - \bar{S_i})^4]$



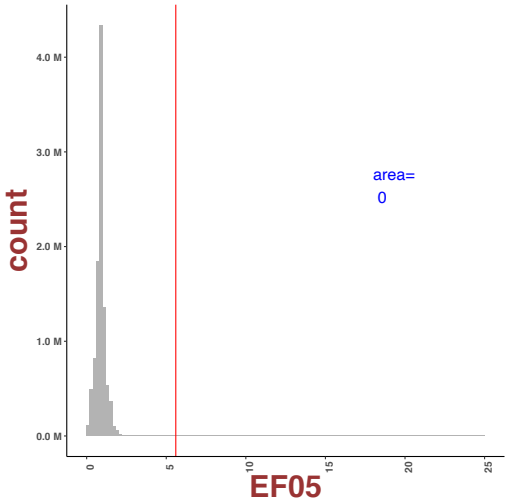
(i) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S_{i,j}^{-})^5$



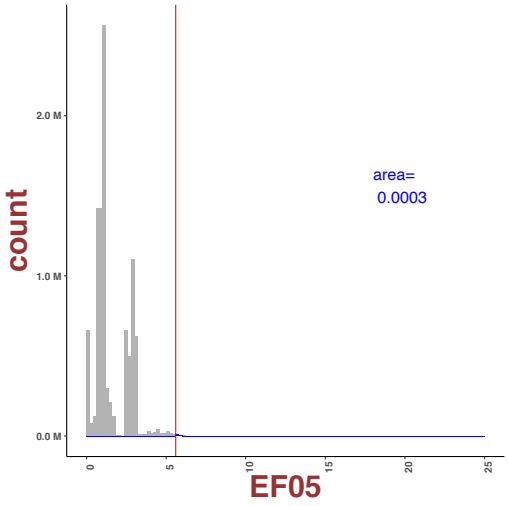
(j) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - \bar{S_i})^5]$



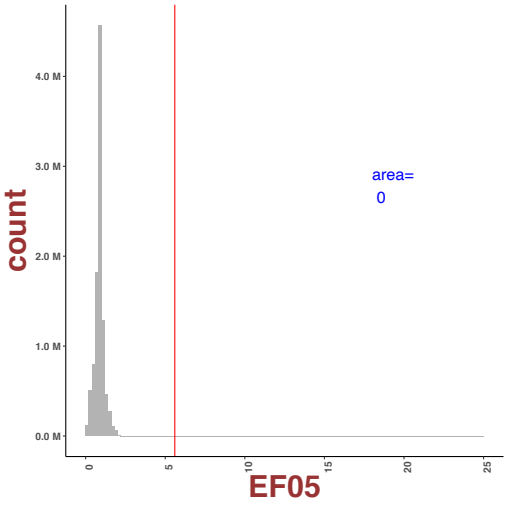
(k) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S_{i,j}^{-})^6$



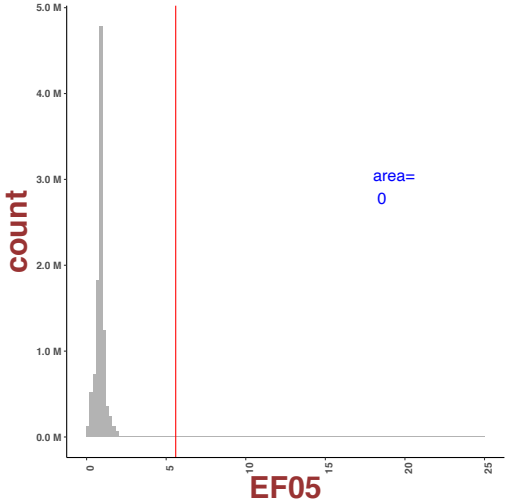
(l) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - \bar{S_i})^6]$



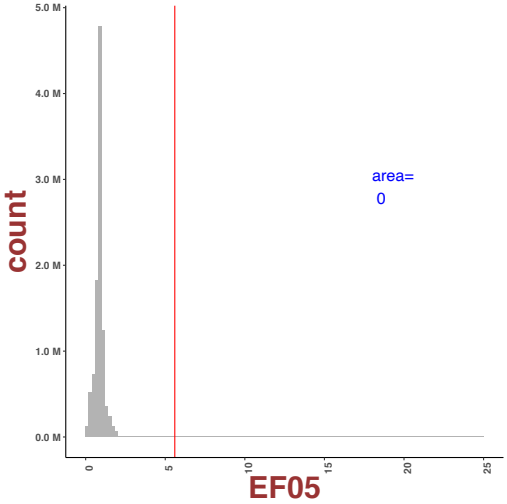
(m) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S_{i,j}^{-})^7$



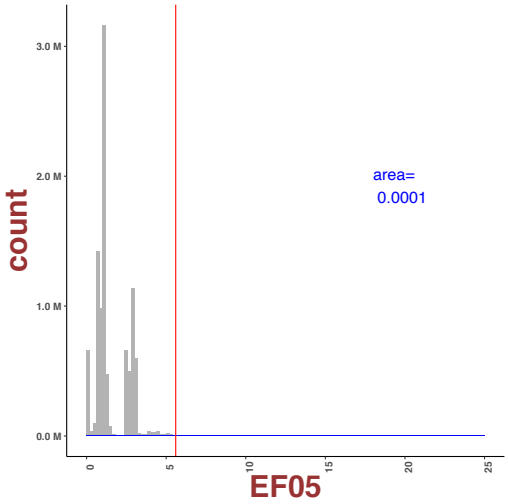
(n) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - \bar{S_i})^7]$



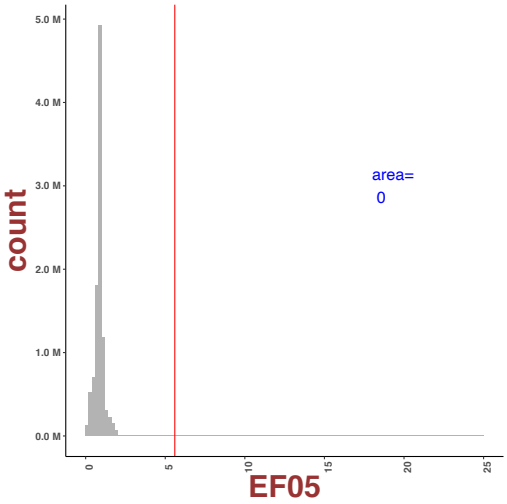
(o) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S_{i,j}^{-})^8$



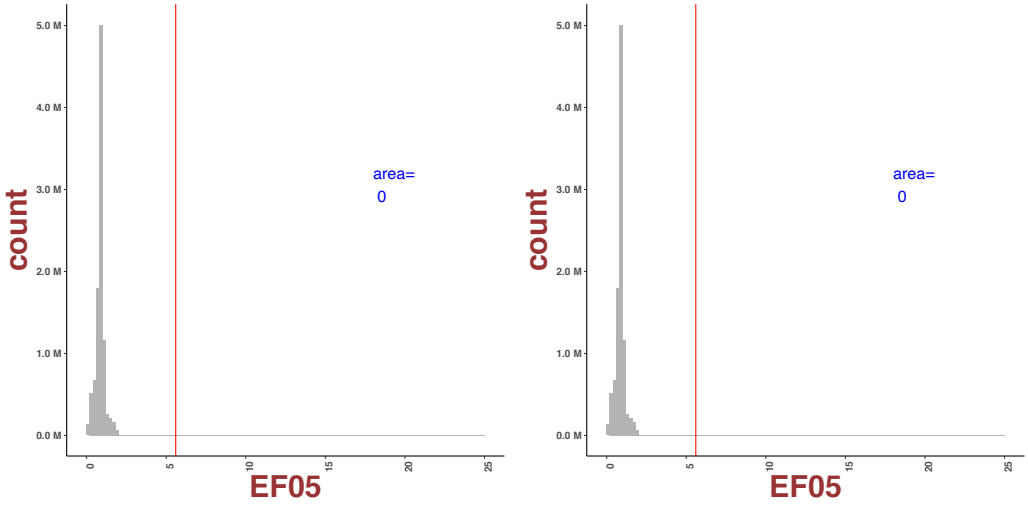
(p) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - \bar{S_i})^8]$



(q) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S_{i,j}^{-})^9$



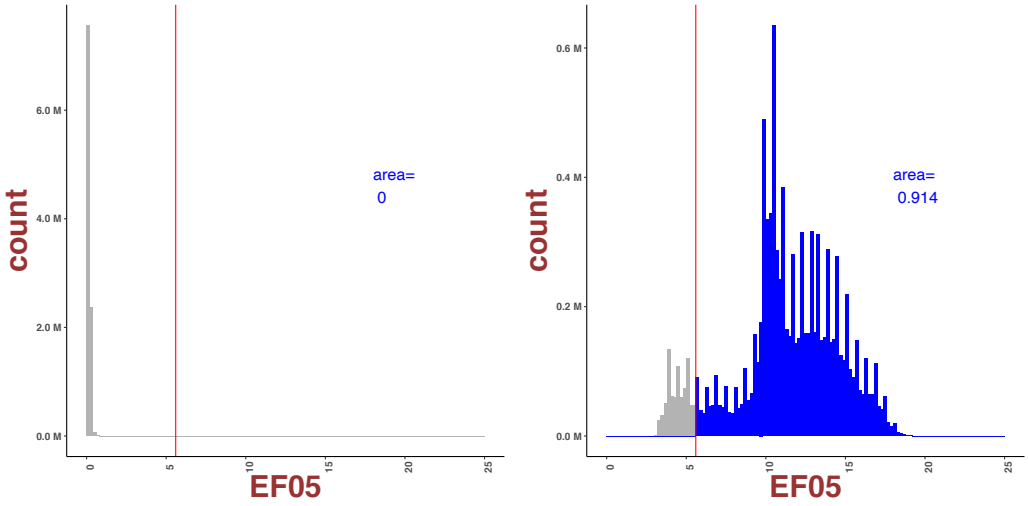
(r) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - \bar{S_i})^9]$



(s) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S_{i,j}^{-})^1 0$

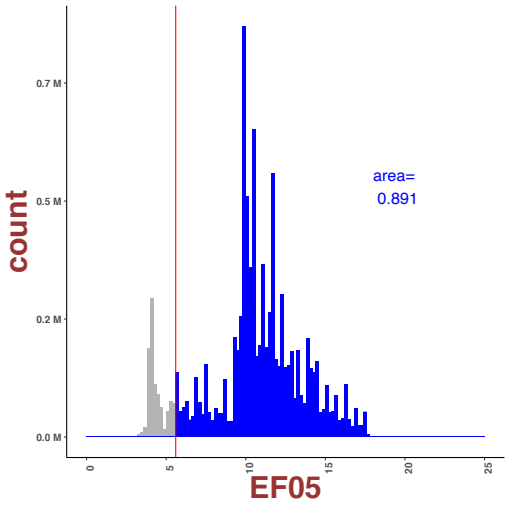
(t) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - \bar{S}_i)^1 0]$

Figure A.2: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with a straight line) of the total histogram area, evaluated for order ranging from 1 to 10 as in Eqns (4.16c (left)-4.16d (right)).

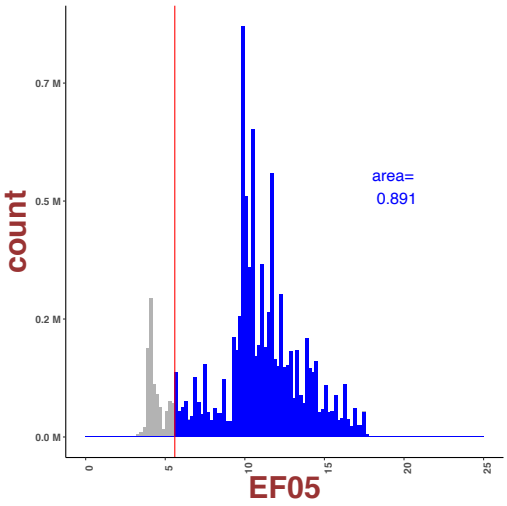


(a) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_{i,j} - S\bar{D}_i)$

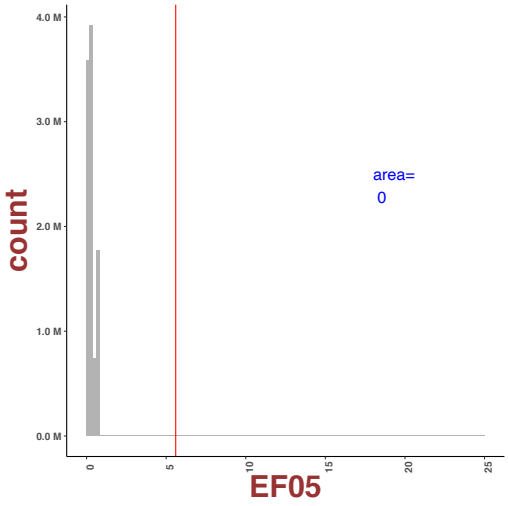
(b) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[S_{i,j} - S\bar{D}_i]$



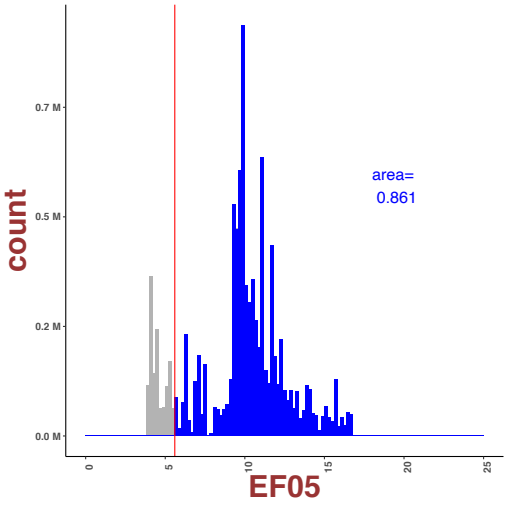
(c) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S\bar{D}_{i,j})^2$



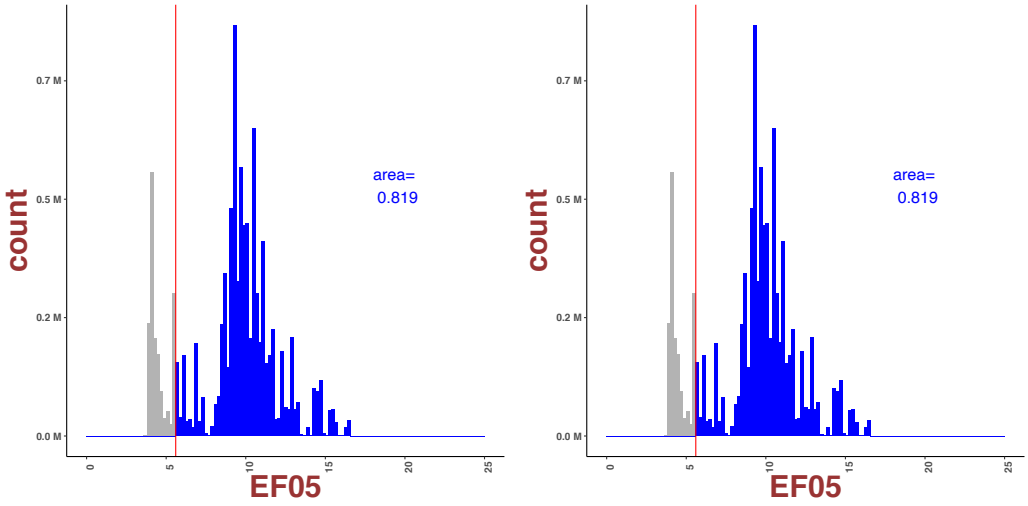
(d) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - S\bar{D}_i)^2]$



(e) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S\bar{D}_{i,j})^3$

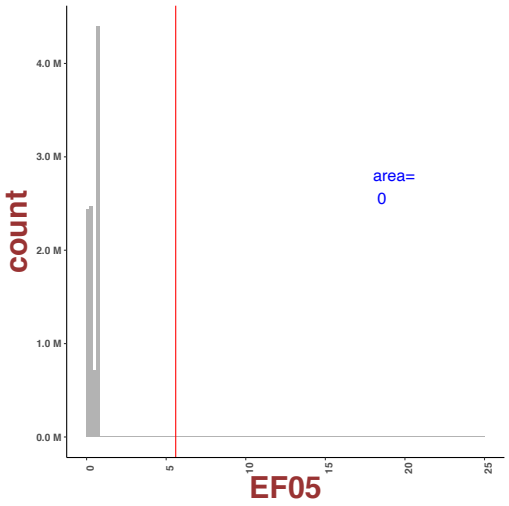


(f) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - S\bar{D}_i)^3]$

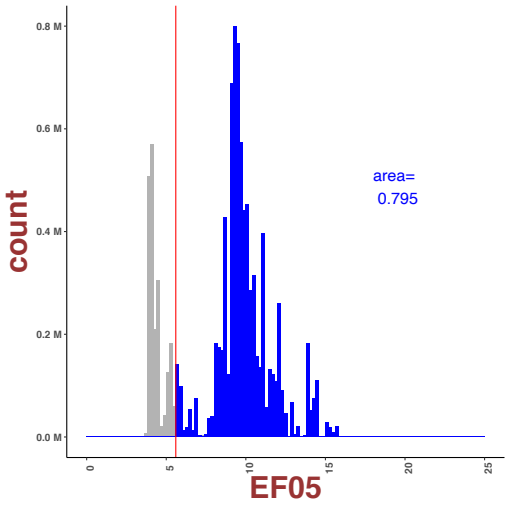


(g) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S\bar{D}_{i,j})^4$

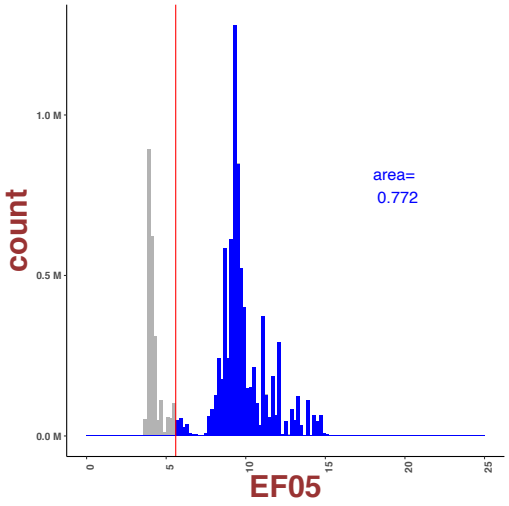
(h) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - S\bar{D}_i)^4]$



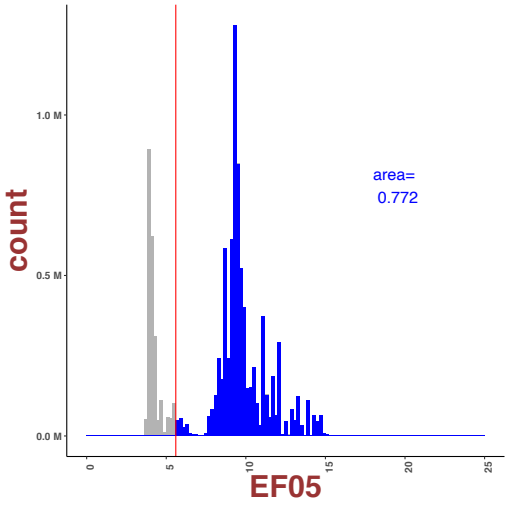
(i) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S\bar{D}_{i,j})^5$



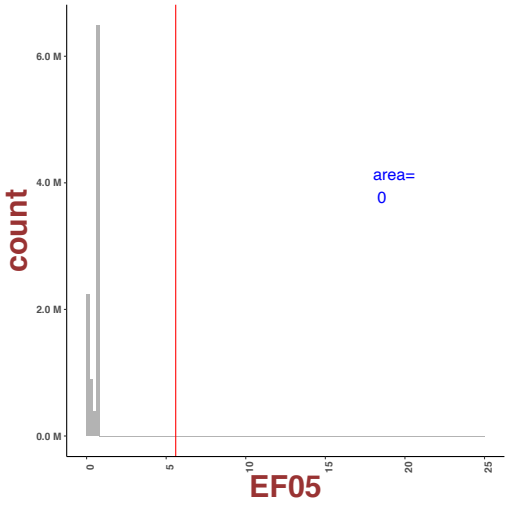
(j) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - S\bar{D}_i)^5]$



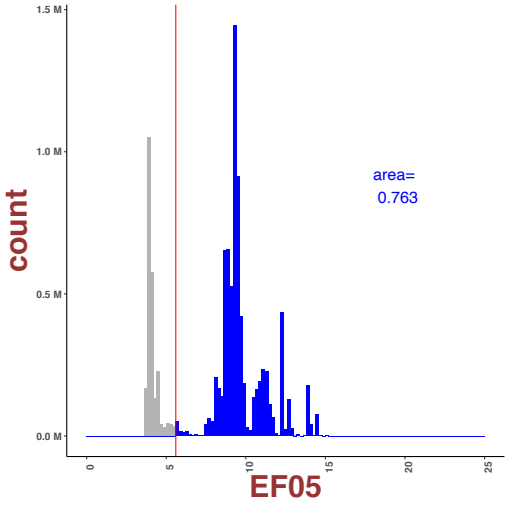
(k) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S\bar{D}_{i,j})^6$



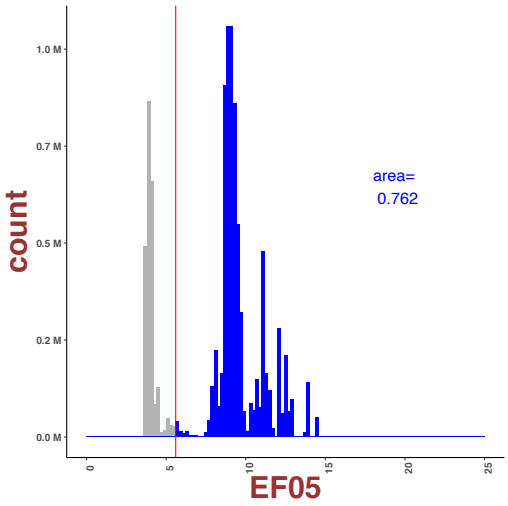
(l) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - S\bar{D}_i)^6]$



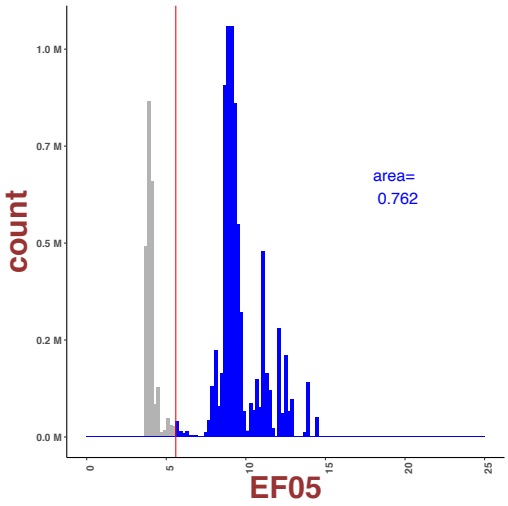
(m) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S\bar{D}_{i,j})^7$



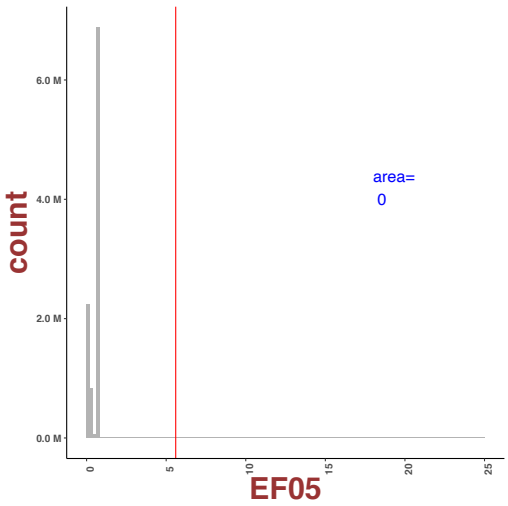
(n) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - S\bar{D}_i)^7]$



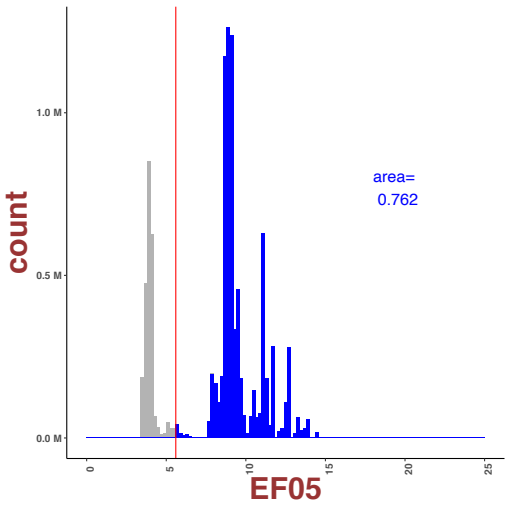
(o) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S\bar{D}_{i,j})^8$



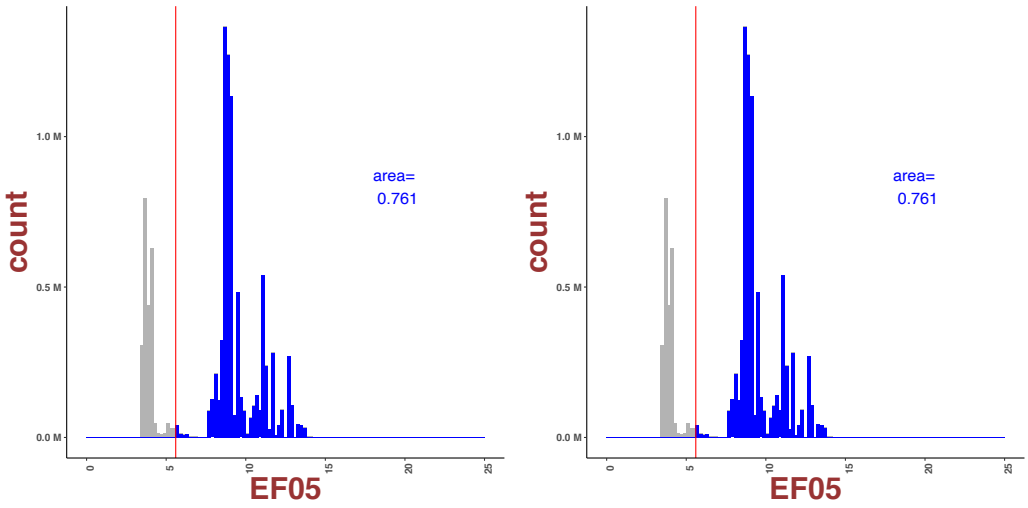
(p) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - S\bar{D}_i)^8]$



(q) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S\bar{D}_{i,j})^9$



(r) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - S\bar{D}_i)^9]$



(s) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, (S_i - S\bar{D}_{i,j})^1 0$

(t) $S_c = \sum_{i=1}^{10} \sum_{j=0}^{20} x_{i,j} \, \mathrm{abs}[(S_{i,j} - S\bar{D}_i)^1 0]$

Figure A.3: Consensus scores, defined as area fraction (to the left of the best performing individual docking score marked with a straight line) of the total histogram area, evaluated for order ranging from 1 to 10 as in Eqns (4.18a (left)-4.18b (right)).

5 Potential candidates for anti-MRSA repurposing

Table A.4: Potential candidates for repurposing
List of Repurposing Hub ligands with the highest frequency among top-ranked ligands in the enrichment subsets. The first and third column lists the compounds with high frequency in the enriched subsets for MRSA proteins. The columns "Freq" illustrate how many enriched subsets that contained the compound in the previous column.

Compound	**Freq**	**Compound**	**Freq**
uridine-5-triphosphate	64	monosodium-alpha-luminol	1
adenosine-triphosphate	61	UNC2327	1
echinomycin	58	Mps-BAY-2a	1
INS316	58	trabodenoson	1
dactinomycin	56	tropifexor	1
candicidin	52	MK-7246	1
actinomycin-d	46	temoporfin	1
rutin	36	tenalisib	1
guadecitabine	29	MK-5108	1
citicoline	23	CGP-60474	1
riboflavin-5-phosphate-sodium	23	trichlormethiazide	1
CGP-71683	20	TG-101209	1
triciribine-phosphate	20	TG-100713	1
folic-acid	20	trifluridine	1
diosmin	19	UBP-310	1
cefonicid	19	cefozopran	1
KB-SRC-4	18	MM-206	1
baicalin	17	mibampator	1
kuromanin	17	FG-4592	1
epigallocatechin-gallate-(-)	17	TC1	1
famotidine	16	GNF-5837	1
EB-47	16	grazoprevir	1
DOTMP	16	quercetin	1
epacadostat	15	GS-6201	1
AMG900	15	quizartinib	1
procyanidin-B-2	14	R-1479	1
myricitrin	14	radezolid	1
HER2-Inhibitor-1	14	radotinib	1
tetrahydrofolic-acid	14	ramatroban	1
lometrexol	13	guanidinoethyldisulfide-bicarbonate	1
naringin	13	resminostat	1
E7449	12	RGFP966	1
thiamine-pyrophosphate	12	HA-1004	1
XL147	12	rifampin	1
CDBA	12	APY-29	1
everolimus	12	risdiplam	1
TPPS4	12	AR-C155858	1
salvianolic-acid-A	12	HSR6071	1
hypericin	12	GR-113808	1
ceftriaxone	12	AMG-208	1
isoquercitrin	12	ascomycin	1
sirolimus	12	purmorphamine	1
fosfructose	11	fimasartan	1
inarigivir	11	A-839977	1

Continuation of Table A.4			
Compound	**Freq**	**Compound**	**Freq**
hesperidin	11	abamectin	1
reynoutrin	10	AC-55541	1
neohesperidin	10	acalabrutinib	1
deforolimus	10	polydatin	1
eprinomectin	10	fosaprepitant-dimeglumine	1
evans-blue	10	pranlukast	1
polyinosine	10	adaprev	1
cefmenoxime	10	presatovir	1
hyperin	10	adrafinil	1
epicatechin-gallate-(-)	10	alanosine	1
pazopanib	10	gepotidacin	1
amphotericin-b	9	GGsTop	1
TAK-243	9	ginsenoside-RE3	1
ebrotidine	9	alovudine	1
SDZ-220-040	9	gliquidone	1
theaflavin	9	arotinolol	1
aminopterin	9	AST-1306	1
adenosine-phosphate	9	carmoterol	1
nystatin	8	JPH203	1
AMI-1	8	silymarin	1
astilbin	8	bicalutamide	1
rifapentine	8	BLZ945	1
hydroxysafflor-yellow-A	8	sotagliflozin	1
methotrexate	8	BMS-587101	1
metafolin	8	L-368899	1
sennoside-protonated	8	BMS-754807	1
fostamatinib	8	BMS-935177	1
4EGI-1	8	labetalol	1
fludarabine-phosphate	8	SRT1720	1
PF-05089771	7	larotrectinib	1
GSK2239633A	7	BVD-523	1
sulfatinib	7	tacrolimus	1
CaMKII-IN-1	7	CA-4948	1
JNJ-64619178	7	lomeguatrib	1
CHIR-98014	7	TAK-632	1
zotarolimus	7	LY2090314	1
etoposide-phosphate	7	JW-74	1
resmetirom	7	JNJ-7706621	1
cefazolin	7	rose-bengal-lactone	1
pyrintegrin	7	SGI-1027	1
doramectin	7	rufloxacin	1
naringin-dihydrochalcone	7	avagacestat	1
alpha-glucosyl-hesperidin	6	idelalisib	1
VU591	6	RX-3117	1
raltitrexed	6	IDO5L	1
PF-573228	6	R306465	1
glipizide	6	AZD3264	1
XL228	6	salidroside	1
nafamostat	6	azilsartan	1
ivermectin	6	A922500	1
R-428	6	irosustat	1
VER-155008	6	SB-772077B	1

Continuation of Table A.4			
Compound	**Freq**	**Compound**	**Freq**
SR-3306	6	SC-9	1
acarbose	6	bendroflumethiazide	1
NT157	6	IWP-L6	1
INC-280	5	JI-101	1
4-galactosyllactose	5	benzthiazide	1
BMS-817378	5	Mps1-IN-5	1
PDD-00017273	5	fosphenytoin	1
ticagrelor	5	myricetin	1
sodium-picosulfate	5	CK-101	1
paromomycin	5	lifitegrast	1
cromoglicic-acid	5	SU11274	1
cefamandole	5	S49076	1
ceftiofur	5	T-025	1
sitaxentan	5	TAK-733	1
neomycin	5	losartan	1
tedizolid-phosphate	5	TAME	1
rifabutin	5	lurasidone	1
TG-100801	5	luteolin	1
rifaximin	5	LXS196	1
TD139	5	LY2874455	1
adavivint	5	LY2979165	1
fostemsavir	5	LY3295668	1
salazodine	5	LY393558	1
PF-05212384	5	telotristat	1
MIW-815	5	cefamandole-nafate	1
tucatinib	5	MCC950	1
GSK2126458	5	CFI-400945	1
casanthranol-variant	5	CF102	1
digitoxin	5	tioguanine	1
TCN201	5	tipiracil	1
IOWH032	4	tivantinib	1
SDZ-220-581	4	chlorthalidone	1
hPGDS-IN-1	4	sulphadimethoxine	1
baricitinib	4	sulfisomidin	1
FCE-22250	4	broxuridine	1
candesartan	4	KS-176	1
LY311727	4	beta-carotene	1
4SC-202	4	silibinin	1
elinogrel	4	simeprevir	1
telatinib	4	kaempferol	1
PSB-06126	4	betamethasone-phosphate	1
madecassoside	4	kanamycin	1
C188-9	4	kasugamycin	1
T-5224	4	BI-78D3	1
phlorizin	4	KHS-101	1
ciaftalan-zinc	4	KPT-9274	1
CGP-78608	4	SNS-314	1
YM-201636	4	astragaloside-a	1
pemetrexed	4	adefovir	1
minodronic-acid	4	BMS-214662	1
PF-04217903	4	L-690330	1
KD025	4	BMS-779788	1

Continuation of Table A.4			
Compound	**Freq**	**Compound**	**Freq**
mangafodipir	4	BMS-833923	1
safflower-yellow	4	BMS-986020	1
amygdalin	4	SR-2640	1
sulfasalazine	4	SRT2104	1
MK-8033	4	ST-2825	1
canagliflozin	4	brivudine	1
cot-inhibitor-1	4	cilengitide	1
SirReal-2	3	CKD-712	1
enzastaurin	3	N-(2-chlorophenyl)-2-({(2E)-2-[1-(2-pyridinyl)ethylidene]hydrazino}carbothioyl)hydrazinecarboth	1
paritaprevir	3	MK-0773	1
isepamicin	3	NMS-E973	1
GNTI	3	selamectin	1
BEBT-908	3	dibekacin	1
sennoside-A	3	ZD-7155	1
ipragliflozin-L-proline	3	ZK-200775	1
TC-S-7004	3	NTNCB	1
PK-44	3	zoledronic-acid	1
LX1031	3	API-1	1
PIK-294	3	olsalazine	1
taprenepag	3	dynasore	1
G-749	3	OTX015	1
ammonium-glycyrrhizinate	3	edaglitazone	1
pevonedistat	3	EED226	1
integrin-antagonist-1	3	EMD-66684	1
bisindolylmaleimide-IX	3	8-bromo-cAMP	1
R112	3	entasobulin	1
S-3304	3	EPZ005687	1
imatinib	3	peposertib	1
AZD5991	3	PF-03814735	1
indisulam	3	PF-04937319	1
EPZ-5676	3	PF-915275	1
"5-amino-3-D-ribofuranosylthiazolo[4,5-d]pyrimidin-2,7(3H,6H)-dione"	3	phthalylsulfathiazole	1
BMS-599626	3	PIK-93	1
cilofexor	3	deltarasin	1
m-THP	3	nifuroxazide	1
ICA-121431	3	WAY-600	1
delphinidin	3	CVT-10216	1
riboflavin	3	clindamycin-phosphate	1
PSB-603	3	MK-5046	1
ceftobiprole	3	MK-8245	1
cefpirome	3	cadazolid	1
YM-244769	3	ML323	1
GLPG0187	3	tubocurarine	1
oligomycin-A	3	UBP-302	1
LY2784544	3	CPI-444	1
TC-G-1008	3	URB597	1
WIKI4	3	CS-917	1

Continuation of Table A.4			
Compound	**Freq**	**Compound**	**Freq**
IACS-10759	3	N-benzylnaltrindole	1
azosemide	3	WAY-316606	1
ZCL-278	3	CX-5461	1
licogliflozin	3	BAY-87-2243	1
TW-37	3	verdinexor	1
neohesperidin-dihydrochalcone	3	C646	1
emamectin	3	vipadenant	1
Mps1-IN-1	2	dacinostat	1
flumatinib	2	nemiralisib	1
GSK3326595	2	danirixin	1
pilaralisib	2	VX-11e	1
zidovudine	2	DBPR-211	1
STAT3-inhibitor-VI	2	JTE-013	1
PF-562271	2	SGX523	1
GS-143	2	JNJ-26481585	1
GSK2334470	2	benserazide	1
trabectedin	2	7-hydroxystaurosporine	1
amiloride	2	ENMD-2076	1
ribostamycin	2	PD-407824	1
tiotidine	2	PDE10-IN-1	1
MK-4074	2	peficitinib	1
MGCD-265	2	pentamidine	1
tenofovir	2	ertugliflozin	1
SR-27897	2	erythromycin	1
chlorogenic-acid	2	esculin	1
cefotetan	2	pexidartinib	1
BMS-986142	2	PF-02545920	1
penciclovir	2	ETC-159	1
ribostamycin-sulfate	2	rifamycin	1
tiludronate	2	PF-06409577	1
TCS-2210	2	PF-06873600	1
L-798106	2	PF-8380	1
apramycin	2	PHA-665752	1
briciclib	2	etrasimod	1
lifirafenib	2	evocalcet	1
leteprinim	2	EW-7197	1
GDC-0834	2	fananserin	1
moxalactam	2	PIK-75	1
dioscin	2	pimecrolimus	1
TWS-119	2	EMD-1214063	1
COH29	2	paliperidone	1
furamidine	2	eltrombopag	1
pradefovir	2	zosuquidar	1
ML228	2	valganciclovir	1
ML193	2	vinflunine	1
PF-477736	2	NBQX	1
lurbinectedin	2	DBeQ	1
PF-06273340	2	WZ-3146	1
guanosine	2	demeclocycline	1
LP-533401	2	YM-022	1
PGL5001	2	diclazuril	1

Continuation of Table A.4			
Compound	**Freq**	**Compound**	**Freq**
AMZ30	2	zaprinast	1
altanserin	2	ZM-241385	1
pyrantel	2	diminazene-aceturate	1
risedronate	2	pafuramidine	1
CUDC-907	2	rose-bengal	1
GR-127935	2	OICR-9429	1
CT7001	2	olaparib	1
linsitinib	2	olomoucine	1
grapiprant	2	doripenem	1
AMG-337	2	omeprazole	1
L-694247	2	dorzolamide	1
nilotinib	2	ONO-8130	1
daidzin	2	3-MPPI	1
masitinib	2	eletriptan	1
CFI-402257	2	pirarubicin	1
beta-amyloid-synthesis-inhibitor	2	pirinixic-acid	1
aztreonam	2	preladenant	1
EPZ015666	2	AZ-7371	1
mocetinostat	2	apigenin	1
hydrocortisone-phosphate	2	SB-216641	1
Ro-5126766	2	Ro-61-8048	1
"1,5-dicaffeoylquinic-acid"	2	rolitetracycline	1
dasabuvir	2	H2L-5765834	1
pelanserin	2	IB-MECA	1
Ro-9187	2	AT13148	1
"2-hydroxy-4-((E)-3-(4-hydroxyphenyl)acryloyl)-2-((2R,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)-6-((2S,3R,4R,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)cyclohexane-1,3,5-trione"	2	avatrombopag	1
SCH-58261	2	RWJ-21757	1
dextrorotation-nimorazole-phosphate-ester	2	idoxuridine	1
ouabain	2	iloperidone	1
cot-inhibitor-2	2	rhein	1
AGI-6780	2	AZD1480	1
LB42708	2	AZD2461	1
T-1095	2	sapropterin	1
bekanamycin	2	sardomozide	1
dihydroergotamine	2	azilsartan-medoxomil	1
linagliptin	2	inosine	1
PRN1008	2	bafetinib	1
BAY-61-3606	2	banoxantrone	1
indacaterol	2	basmisanil	1
deslanoside	2	beclabuvir	1

Continuation of Table A.4			
Compound	**Freq**	**Compound**	**Freq**
imidurea	2	hematoporphyrin	1
8-bromo-cGMP	2	GW-627368	1
cefsulodin	2	acalisib	1
TG100-115	2	proscillaridin-A	1
hygromycin-B	2	acedapsone	1
epirubicin	2	forodesine	1
astaxanthin	2	aclarubicin	1
bisoctrizole	2	fenoverine	1
bisantrene	2	ACT-132577	1
rosmarinic-acid	2	prednisolone-sodium-phosphate	1
TFC-007	2	adoprazine	1
icariin	2	gamithromycin	1
avanafil	2	ganetespib	1
CEP-33779	2	GDC-0980	1
avitinib	2	genistein	1
idarubicin	2	amuvatinib	1
darolutamide	2	PRX-08066	1
IPI549	2	GNF-7	1
MBX-2982	2	AM-1241	1
ellagic-acid	2	pyronaridine	1
sinefungin	2	GS-9973	1
cidofovir	2	AMG-517	1
methyclothiazide	1	GSK1292263	1
SAM-315	1	GSK256066	1
CL316243	1	GSK461364	1
ML277	1	amrubicin	1
tedizolid	1	AZD8835	1

Table A.5: Potential candidates for repurposing
List of Repurposing Hub ligands with the highest frequency among top-ranked ligands in the enrichment subsets. The first and third column lists the compounds with high frequency in the enriched subsets for modelled MRSA proteins. The columns "Freq" illustrate how many enriched subsets that contained the compound in the previous column.

Compound	**Freq**	**Compound**	**Freq**
epacadostat	42	althiazide	1
adenosine-triphosphate	41	AMG319	1
INS316	36	cefamandole	1
PF-573228	32	beta-amyloid-synthesis-inhibitor	1
uridine-5-triphosphate	31	sanguinarium-chloride	1
cefmenoxime	30	GPi-688	1
guadecitabine	26	AMG-PERK-44	1
HER2-Inhibitor-1	23	KAF-156	1
ceftriaxone	21	SPP301	1
cefazolin	21	AMG-925	1
famotidine	20	amlexanox	1
AMG900	20	cefepime	1
ebrotidine	19	SB-200646	1
NS-11021	17	purmorphamine	1
PF-562271	17	AZD8797	1
tucatinib	17	ibandronate	1

Continuation of Table A.5			
Compound	**Freq**	**Compound**	**Freq**
cefonicid	17	SB-334867	1
DOTMP	16	adaprev	1
fostamatinib	15	pirenoxine	1
cefotetan	14	cimetidine	1
ceftiofur	14	abamectin	1
azosemide	14	CID-2745687	1
PRT062607	14	CID-16020046	1
imidurea	14	PIT	1
bisindolylmaleimide-IX	13	CFM-2	1
CHIR-98014	13	SCH-900776	1
ceftobiprole	13	SC-51089	1
XL147	13	acalabrutinib	1
folic-acid	13	ivermectin	1
PF-05089771	12	PLX4720	1
XL228	12	fosaprepitant-dimeglumine	1
APY-29	11	SB-772077B	1
bisantrene	11	PSB-603	1
adavivint	11	ticagrelor	1
tetrahydrofolic-acid	11	merbarone	1
PLX8394	11	thiophanate	1
SNS-314	11	AG-490	1
metafolin	11	didox	1
radotinib	10	CGP-53353	1
tiotidine	10	GDC-0941	1
rutin	10	baricitinib	1
EB-47	9	iproniazid	1
nilotinib	9	SB-525334	1
guanidinoethyldisulfide-bicarbonate	9	PTC-209	1
MIW-815	9	CB-5083	1
defactinib	9	SGI-1027	1
KB-SRC-4	9	TCS-359	1
alpha-glucosyl-hesperidin	8	radezolid	1
KD025	8	sal003	1
TC-G-1008	8	sulfadoxine	1
cefozopran	8	ribostamycin	1
HA-1004	8	rifampin	1
aminopterin	8	licostinel	1
FN-1501	8	rifaximin	1
minodronic-acid	8	rimegepant	1
nafamostat	8	sulfaquinoxaline	1
ZCL-278	8	selonsertib	1
neomycin	7	sulfamonomethoxine	1
cilengitide	7	hPGDS-IN-1	1
R-428	7	TCN201	1
GNTI	7	iclaprim	1
FR-180204	7	Ro-3306	1
paromomycin	7	sulfamethizole	1
PF-431396	7	sulfameter	1
pilaralisib	7	L-368899	1
CUDC-907	7	TA-01	1
ARRY-334543	7	atuveciclib	1

Continuation of Table A.5			
Compound	**Freq**	**Compound**	**Freq**
benzamil	6	hydroxyfasudil	1
TAK-243	6	ascomycin	1
myricitrin	6	sulfachlorpyridazine	1
AZD3264	6	lesinurad	1
triciribine-phosphate	6	ABBV-744	1
R406	6	brivanib-alaninate	1
elinogrel	6	BMS-817378	1
TD139	6	AST-1306	1
BMS-754807	6	astragaloside-a	1
lometrexol	6	brinzolamide	1
amiloride	6	IACS-10759	1
cefoselis	6	BQ-123	1
fostemsavir	6	larotrectinib	1
NVP-BHG712	6	apafant	1
sennoside-protonated	6	CA-4948	1
indisulam	6	GSK2256294A	1
VX-11e	6	GSK356278	1
EW-7197	6	LY2857785	1
LY2801653	5	sitaxentan	1
LY2090314	5	lamotrigine	1
ML193	5	kifunensine	1
azilsartan-medoxomil	5	LY2603618	1
BEBT-908	5	SKLB-1028	1
acarbose	5	LY2157299	1
JTE-013	5	AZD4635	1
CGP-71683	5	LX1031	1
benzthiazide	5	RAF265	1
sulfatinib	5	salazodine	1
AZ-10417808	5	GSK3179106	1
sotrastaurin	5	LXR-623	1
inarigivir	5	tasisulam	1
GSK256066	5	ravoxertinib	1
pazopanib	5	lomeguatrib	1
avitinib	5	GTP-14564	1
pimodivir	5	lurbinectedin	1
NS-5806	5	tanshinone-IIA	1
tradipitant	5	tenofovir	1
ciaftalan-zinc	5	safflower-yellow	1
NS-3623	5	LP-533401	1
CZC-54252	5	BLZ945	1
nifursol	5	guanaben-acetate	1
trabodenoson	5	BML-284	1
GSK1838705A	5	IKK-2-inhibitor-V	1
HSR6071	4	guanfacine	1
SRT3190	4	guanosine	1
dabrafenib	4	GW-438014A	1
PF-04217903	4	reynoutrin	1
cefixime	4	MI-14	1
AZD1480	4	A-804598	1
naringin-dihydrochalcone	4	citicoline	1
JNJ-64619178	4	siguazodan	1
EED226	4	talmapimod	1

Continuation of Table A.5			
Compound	**Freq**	**Compound**	**Freq**
cefmetazole	4	candesartan	1
AGI-6780	4	SKF-86002	1
telatinib	4	taminadenant	1
BAY-1251152	4	LTA	1
R547	4	LXS196	1
GNF-5837	4	LY215490	1
IDO5L	4	LY2784544	1
CW-008	4	TC-S-7003	1
zoledronic-acid	4	LY3009120	1
zaltidine	4	TCS-21311	1
fludarabine-phosphate	4	LY3295668	1
GLPG0187	4	CCG-63802	1
sennoside-A	4	CCMI	1
MLN2480	4	LY393558	1
I-BRD9	4	tegobuvir	1
phthalylsulfathiazole	4	telotristat	1
JW-74	4	cefalonium	1
bekanamycin	4	cefdinir	1
adenosine-phosphate	4	cefetamet	1
zoliflodacin	4	masitinib	1
BMS-599626	4	cefotaxime	1
methotrexate	4	MBX-2982	1
CEP-32496	4	ceftazidime	1
KPT-9274	4	cefuroxime	1
enasidenib	4	CEP-33779	1
CFI-402257	4	ceritinib	1
SB-415286	4	losartan	1
CCT196969	4	lorlatinib	1
diosmin	4	TAK-593	1
chlorthalidone	4	L-694247	1
IPI549	4	SirReal-2	1
rebastinib	4	sisomicin	1
Ro-5126766	4	kasugamycin	1
AMI-1	4	kinetin	1
SRT1720	3	BIBX-1382	1
icariin	3	KRCA-0008	1
kanamycin	3	SNX-5422	1
alendronate	3	L-arginine	1
PF-06273340	3	BMS-214662	1
TC-S-7004	3	BMS-345541	1
ammonium-glycyrrhizinate	3	BMS-582949	1
flumatinib	3	BMS-707035	1
BX-912	3	BMS-779788	1
cot-inhibitor-2	3	LMK-235	1
risedronate	3	L-838417	1
hypericin	3	BMS-986158	1
BMS-833923	3	BMY-45778	1
sirolimus	3	SRT2104	1
salvianolic-acid-A	3	LB42708	1
kuromanin	3	STF-083010	1
CQS	3	sulfisomidin	1
Mps-BAY-2a	3	bumetanide	1

Continuation of Table A.5			
Compound	**Freq**	**Compound**	**Freq**
tigecycline	3	LIMKi-3	1
BT-11	3	SU3327	1
4EGI-1	3	S1P1-agonist-III	1
protirelin	3	T-025	1
N-(2-chlorophenyl)-2-({(2E)-2-[1-(2-pyridinyl)ethylidene]hydrazino}carbothioyl)hydrazinecarbothioamide	3	TGR-1202	1
azathioprine	3	CF102	1
KG-5	3	CGP-52411	1
BW-348U87	3	EMD-66684	1
hyperin	3	NVP-TAE684	1
mocetinostat	3	1-phenylbiguanide	1
TG-101209	3	dithranol	1
chlorproguanil	3	oglufanide	1
TG100-115	3	"2-hydroxy-4-((E)-3-(4-hydroxyphenyl)acryloyl)-2-((2R,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)-6-((2S,3R,4R,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)cyclohexane-1,3,5-trione"	1
bendroflumethiazide	3	omeprazole-magnesium	1
TAK-632	3	ONO-4059	1
sardomozide	3	DSR-6434	1
CHIR-99021	3	dynasore	1
SDZ-220-040	3	edaglitazone	1
CaMKII-IN-1	3	ellagic-acid	1
vericiguat	3	AM-1241	1
tivantinib	3	epigallocatechin-gallate-(-)	1
SDZ-220-581	3	zibotentan	1
entrectinib	3	PD-318088	1
E7449	3	PDD-00017273	1
candicidin	3	eprinomectin	1
MGCD-265	3	EPZ005687	1
IMREG-1	3	penciclovir	1
pamidronate	3	perfluorodecalin	1
echinomycin	3	PF-03814735	1
BTT-3033	3	PF-04937319	1
AMZ30	3	PF-4981517	1
everolimus	3	evobrutinib	1
neridronic-acid	3	FCE-22250	1
PIK-93	3	PIK-294	1
actinomycin-d	3	NT157	1
acetazolamide	3	zeatin	1
hydrochlorothiazide	3	CGP-78608	1
polyinosine	3	CT7001	1
m-THP	3	CGS-15943	1
CDBA	3	tifenazoxide	1

Continuation of Table A.5			
Compound	**Freq**	**Compound**	**Freq**
madecassoside	3	methacycline	1
abafungin	3	TP-0903	1
gliquidone	3	TPCA-1	1
MK-8033	3	CKD-712	1
isepamicin	3	clodronic-acid	1
cefpirome	3	clorsulon	1
deforolimus	3	CP-471474	1
BAY-61-3606	3	CPI-444	1
vinblastine	2	MRK-560	1
MK-3207	2	crizotinib-(S)	1
molidustat	2	CVT-10216	1
cilofexor	2	ZD-7155	1
MM-206	2	vapendavir	1
INC-280	2	vatalanib	1
JNJ-7706621	2	VE-822	1
sangivamycin	2	verdinexor	1
doramectin	2	vorasidenib	1
GF109203X	2	nemiralisib	1
Ro-61-8048	2	VX-702	1
epicatechin-gallate-(-)	2	NG-nitro-arginine	1
lifirafenib	2	WAY-213613	1
T-5224	2	WZ-4002	1
neohesperidin	2	denotivir	1
VU591	2	deslanoside	1
NMS-1286937	2	betamethasone-phosphate	1
PF-03758309	2	JD-5037	1
eravacycline	2	trichlormethiazide	1
m-chlorophenylbiguanide	2	seletalisib	1
AR-12	2	WIKI4	1
eniporide	2	niflumic-acid	1
8-bromo-cGMP	2	nolatrexed	1
edoxaban	2	yoda-1	1
nystatin	2	diclofenamide	1
TC-S-7006	2	zaprinast	1
oligomycin-A	2	ZM-241385	1
casanthranol-variant	2	NVP-AUY922	1
nelociguat	2	zosuquidar	1
C188-9	2	dioscin	1
pranlukast	2	oglemilast	1
sulfaguanidine	2	olmutinib	1
rimeporide	2	doripenem	1
hesperidin	2	dorzolamide	1
apramycin	2	opicapone	1
tivozanib	2	3-bromo-7-nitroindazole	1
ribostamycin-sulfate	2	DSM265	1
amphotericin-b	2	6-aminochrysene	1
BMS-587101	2	emamectin	1
GSK2126458	2	paritaprevir	1
TH-302	2	BAY-87-2243	1
BMS-626529	2	entecavir	1
GS-9973	2	U-104	1
BMS-986142	2	enzastaurin	1

Continuation of Table A.5			
Compound	**Freq**	**Compound**	**Freq**
glipizide	2	semaxanib	1
tenalisib	2	ETC-159	1
CEP-37440	2	phenformin	1
ertapenem	2	DCC-2618	1
BI-78D3	2	darglitazone	1
SD-208	2	VU0364439	1
tiludronate	2	CT-7758	1
CS-917	2	ML281	1
GSK2239633A	2	trilaciclib	1
Mps1-IN-1	2	MK-2461	1
moxalactam	2	MLN0128	1
cot-inhibitor-1	2	tropifexor	1
ML786	2	TS-011	1
COH29	2	compound-w	1
MK-8245	2	conivaptan	1
CK-101	2	CP-316819	1
tozasertib	2	UBP-310	1
CHR-6494	2	MRK-409	1
tobramycin	2	cromoglicic-acid	1
CH-5183284	2	vadadustat	1
TAME	2	nemorubicin	1
regadenoson	2	valaciclovir	1
CGP-57380	2	CV-1808	1
thiamine-pyrophosphate	2	bromosporine	1
aztreonam	2	UBP-302	1
TG-100801	2	cyclic-AMP	1
mangafodipir	2	vatinoxan	1
tedizolid-phosphate	2	VE-821	1
HTH-01-015	2	naltriben	1
Ro-4987655	2	dabigatran	1
LY2979165	2	necrostatin-2	1
TC-S-7009	2	voxtalisib	1
cariporide	2	VS-4718	1
cyclopenthiazide	2	exo-IWR-1	1
vemurafenib	2	E7046	1
VER-155008	2	E7820	1
vinorelbine	2	ICI-162846	1
A-839977	2	amuvatinib	1
evans-blue	2	G007-LK	1
etoposide-phosphate	2	riboflavin-5-phosphate-sodium	1
etidronic-acid	2	rifabutin	1
PF-477736	2	rifapentine	1
pevonedistat	2	ripretinib	1
AC-55541	2	hyaluronic-acid	1
pemetrexed	2	hydroflumethiazide	1
acitazanolast	2	"hydroxytacrine-maleate-(R,S)"	1
ACTB-1003	2	hygromycin-B	1

Continuation of Table A.5			
Compound	**Freq**	**Compound**	**Freq**
"5-amino-3-D-ribofuranosylthiazolo[4,5-d]pyrimidin-2,7(3H,6H)-dione"	2	AT-9283	1
presatovir	2	IB-MECA	1
DNQX	2	AV-608	1
chlorothiazide	2	GSK9027	1
ditolylguanidine	2	eltrombopag	1
PRT062070	2	avatrombopag	1
PSB-06126	2	R112	1
zotarolimus	2	imatinib	1
zanamivir	2	AZD2858	1
dibekacin	2	AZD6738	1
NK-252	2	sasapyrine	1
DBPR-211	2	butalbital	1
danirixin	2	BAM7	1
dactinomycin	2	SB-683698	1
VLX600	2	beclabuvir	1
taurolidine	2	ITI214	1
theaflavin	2	regorafenib	1
micronomicin	2	GSK3326595	1
avapritinib	2	phthalylsulfacetamide	1
lifitegrast	2	GDC-0834	1
SU11274	2	PIK-293	1
azaguanine-8	2	PIK-75	1
torasemide	2	piretanide	1
sulfasalazine	2	PK-44	1
succinylsulfathiazole	2	fluralaner	1
SB-747651A	2	PP-121	1
L-phenylisopropyladenosine	2	PP242	1
TAK-659	2	pradefovir	1
selinexor	2	ACT-132577	1
sparfosate	2	G-749	1
benserazide	2	procyanidin-B-2	1
isoxicam	2	AH-7614	1
SR-3306	2	glecaprevir	1
AZ960	2	raltitrexed	1
iobenguane	2	taltirelin	1
talniflumate	2	PU-H71	1
isoquercitrin	2	puromycin	1
lidamidine	2	purvalanol-B	1
SCH-58261	2	pyrantel	1
AMG-548	1	GNF-7	1
GPBAR-A	1	go-6983	1
INCB-057643	1	pyrazolanthrone	1
tenoxicam	1	pyrintegrin	1
SB-2343	1	AMG-517	1
LY3214996	1	GSK2256098	1
ceftizoxim	1	CC-930	1
JPH203	1	(R)-(-)-apomorphine	1
6-benzylaminopurine	1		

6 Pseudocode of Consensus Scores

For 10 coefficients x_i, with the increments of 0.05 and the sum of 10 coefficient is 1.0, there were $\binom{29}{9}$ combinations (10015005). Therefore, the code run over 10015005 sets of ten coefficients.

for j=1 $\rightarrow$ 10015005
    $S_c = \sum_{i=10} x_{i,j} S_i$
    Calculate $rank_j$, $AUCROC_j$, EF_j
end

Draw histograms of rank, AUCROC, EF.
Calculate blue patches better than best individual program.

References

Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, 15(5):488–506.

Adams, C. P. and Brantner, V. V. (2006 Mar-Apr). Estimating the cost of new drug development: Is it really 802 million dollars? *Health Affairs (Project Hope)*, 25(2):420–428.

Adams, C. P. and Brantner, V. V. (2010). Spending on new drug development. *Health Economics*, 19(2):130–141.

Alhossary, A., Handoko, S. D., Mu, Y., and Kwoh, C.-K. (2015). Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics*, 31(13):2214–2216.

Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., Case, D. A., Kuntz, I. D., and Rizzo, R. C. (2015). DOCK 6: Impact of New Features and Current Docking Performance. *Journal of computational chemistry*, 36(15):1132–1156.

Allen, W. J. and Rizzo, R. C. (2014). Implementation of the Hungarian Algorithm to Account for Ligand Symmetry and Similarity in Structure-Based Design. *Journal of Chemical Information and Modeling*, 54(2):518–529.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Amendola, G., Ettari, R., Previti, S., Di Chio, C., Messere, A., Di Maro, S., Hammerschmidt, S. J., Zimmer, C., Zimmermann, R. A., Schirmeister, T., Zappalà, M., and Cosconati, S. (2021). Lead Discovery of SARS-CoV-2 Main Protease Inhibitors through Covalent Docking-Based Virtual Screening. *Journal of Chemical Information and Modeling*, 61(4):2062–2073.

Aminov, R. I. (2010). A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. *Frontiers in Microbiology*, 1.

Amory, J. K. and Amory, D. W. (2007). Dosing frequency of aspirin and prevention of heart attacks and strokes. *The American Journal of Medicine*, 120(4):e5; author reply e7.

Apsel, B., Blair, J. A., Gonzalez, B. Z., Nazif, T. M., Feldman, M. E., Aizenstein, B., Hoffman, R., Williams, R. L., Shokat, K. M., and Knight, Z. A. (2008). Targeted polypharmacology: Discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nature chemical biology*, 4(11):691.

Ashburn, T. T. and Thor, K. B. (2004). Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683.

Azzaoui, K., Hamon, J., Faller, B., Whitebread, S., Jacoby, E., Bender, A., Jenkins, J. L., and Urban, L. (2007). Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem*, 2(6):874–880.

Bailey, D. and Brown, D. (2001). High-throughput chemistry and structure-based design: Survival of the smartest. *Drug Discovery Today*, 6(2):57–59.

Ballester, P. J. and Mitchell, J. B. O. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175.

Ban, T. A. (2006). The role of serendipity in drug discovery. *Dialogues in Clinical Neuroscience*, 8(3):335–344.

Barber, M. (1961). Methicillin-resistant staphylococci. *Journal of Clinical Pathology*, 14(4):385–393.

Baron, J. A. (2012). Aspirin and cancer: Trials and observational studies. *Journal of the National Cancer Institute*, 104(16):1199–1200.

Bassetti, M., Merelli, M., Temperoni, C., and Astilean, A. (2013). New antibiotics for bad bugs: Where are we? *Annals of Clinical Microbiology and Antimicrobials*, 12(1):22.

Batista, J., Hawkins, P. C., Tolbert, R., and Geballe, M. T. (2014). SiteHopper - a unique tool for binding site comparison. *Journal of Cheminformatics*, 6(1):P57.

Bauer, M. R., Ibrahim, T. M., Vogel, S. M., and Boeckler, F. M. (2013). Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public Library of Challenging Docking Benchmark Sets. *Journal of Chemical Information and Modeling*, 53(6):1447–1462.

Beise, F., Labischinski, H., and Giesbrecht, P. (1988). Selective inhibition of penicillin-binding proteins and its effects on growth and architecture of Staphylococcus aureus. *FEMS Microbiology Letters*, 55(2):195–201.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, 10(12):980.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallographica. Section D, Biological Crystallography*, 58(Pt 6 No 1):899–907.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Berman, J., Haffajee, C. I., and Alpert, J. S. (1981). Therapy of symptomatic pericarditis after myocardial infarction: Retrospective and prospective studies of aspirin, indomethacin, prednisone, and spontaneous resolution. *American Heart Journal*, 101(6):750–753.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L., and Schwede, T. (2014). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(Web Server issue):W252.

Bissantz, C., Folkers, G., and Rognan, D. (2000). Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *Journal of Medicinal Chemistry*, 43(25):4759–4767.

Bloom, B. E. (2016). Recent successes and future predictions on drug repurposing for rare diseases. *Expert Opinion on Orphan Drugs*, 4(1):1–4.

Bodian, D. L., Yamasaki, R. B., Buswell, R. L., Stearns, J. F., White, J. M., and Kuntz, I. D. (1993). Inhibition of the fusion-inducing conformational change of influenza hemagglutinin by benzoquinones and hydroquinones. *Biochemistry*, 32(12):2967–2978.

Boran, A. D. and Iyengar, R. (2010a). Systems approaches to polypharmacology and drug discovery. *Current opinion in drug discovery & development*, 13(3):297.

Boran, A. D. W. and Iyengar, R. (2010b). Systems Pharmacology. *The Mount Sinai journal of medicine, New York*, 77(4):333.

Bowen, L. R., Li, D. J., Nola, D. T., Anderson, M. O., Heying, M., Groves, A. T., and Eagon, S. (2019). Identification of potential Zika virus NS2B-NS3 protease inhibitors via docking, molecular dynamics and consensus scoring-based virtual screening. *Journal of Molecular Modeling*, 25(7):194.

Breijyeh, Z., Jubeh, B., and Karaman, R. (2020). Resistance of Gram-Negative Bacteria to Current Antibacterial Agents and Approaches to Resolve It. *Molecules (Basel, Switzerland)*, 25(6):1340.

Briansó, F., Carrascosa, M. C., Oprea, T. I., and Mestres, J. (2011). Cross-Pharmacology Analysis of G Protein-Coupled Receptors. *Current topics in medicinal chemistry*, 11(15):1956.

Broccatelli, F. and Brown, N. (2014). Best of Both Worlds: On the Complementarity of Ligand-Based and Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling*, 54(6):1634.

Brooijmans, N. and Kuntz, I. D. (2003). Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure*, 32:335–373.

Brown, D. (2007). Unfinished business: Target-based drug discovery. *Drug Discovery Today*, 12(23):1007–1012.

Brylinski, M. (2013). Nonlinear Scoring Functions for Similarity-Based Ligand Docking and Binding Affinity Prediction. *Journal of Chemical Information and Modeling*, 53(11):3097–3112.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D. S., Ghosh, S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Lawson, C. L., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Persikova, I., Randle, C., Rose, A., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Tao, Y.-P., Voigt, M., Westbrook, J. D., Young, J. Y., Zardecki, C., and Zhuravleva, M. (2021). RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49(D1):D437–D451.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10:421.

Carta, G., Knox, A. J. S., and Lloyd, D. G. (2007). Unbiasing Scoring Functions: A New Normalization and Rescoring Strategy. *Journal of Chemical Information and Modeling*, 47(4):1564–1571.

CDC (2013). Antibiotic resistance threats in the United States, 2013. Technical report, Centers for Disease Control and Prevention.

Cereto-Massagué, A., Guasch, L., Valls, C., Mulero, M., Pujadas, G., and Garcia-Vallvé, S. (2012). DecoyFinder: An easy-to-use python GUI application for building target-specific decoy sets. *Bioinformatics (Oxford, England)*, 28(12):1661–1662.

Chakraborti, S., Ramakrishnan, G., and Srinivasan, N. (2019). Repurposing Drugs Based on Evolutionary Relationships Between Targets of Approved Drugs and Proteins of Interest. In Vanhaelen, Q., editor, *Computational Methods for Drug Repurposing*, Methods in Molecular Biology, pages 45–59. Springer New York, New York, NY.

Chambers, H. F. (1997). Methicillin resistance in staphylococci: Molecular and biochemical basis and clinical implications. *Clinical Microbiology Reviews*, 10(4):781.

Chambers, H. F. and DeLeo, F. R. (2009). Waves of Resistance: Staphylococcus aureus in the Antibiotic Era. *Nature reviews. Microbiology*, 7(9):629–641.

Charifson, P. S., Corkery, J. J., Murcko, M. A., and Walters, W. P. (1999). Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *Journal of Medicinal Chemistry*, 42(25):5100–5109.

Chartier, M. and Najmanovich, R. (2015). Detection of Binding Site Molecular Interaction Field Similarities. *Journal of Chemical Information and Modeling*, 55(8):1600–1615.

Chattopadhyay, A. K. and Marenduzzo, D. (2007). Dynamics of an anchored polymer molecule under an oscillating force. *Physical Review Letters*, 98(8):088101.

Chaudhuri, R. R., Allen, A. G., Owen, P. J., Shalom, G., Stone, K., Harrison, M., Burgis, T. A., Lockyer, M., Garcia-Lara, J., Foster, S. J., Pleasance, S. J., Peters, S. E., Maskell, D. J., and Charles, I. G. (2009). Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC genomics*, 10:291.

Cheng, T., Li, X., Li, Y., Liu, Z., and Wang, R. (2009). Comparative Assessment of Scoring Functions on a Diverse Test Set. *Journal of Chemical Information and Modeling*, 49(4):1079–1093.

Chong, C. R., Chen, X., Shi, L., Liu, J. O., and Sullivan, D. J. (2006). A clinical drug library screen identifies astemizole as an antimalarial agent. *Nature Chemical Biology*, 2(8):415–416.

Chu, J., Vila-Farres, X., Inoyama, D., Ternei, M., Cohen, L. J., Gordon, E. A., Reddy, B. V. B., Charlop-Powers, Z., Zebroski, H. A., Gallardo-Macias, R., Jaskowski, M., Satish, S., Park, S., Perlin, D. S., Freundlich, J. S., and Brady, S. F. (2016). Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nature chemical biology*, 12(12):1004–1006.

Clark, R. D., Strizhev, A., Leonard, J. M., Blake, J. F., and Matthew, J. B. (2002). Consensus scoring for ligand/protein interactions. *Journal of Molecular Graphics and Modelling*, 20(4):281–295.

Clark, R. D. and Webster-Clark, D. J. (2008 Mar-Apr). Managing bias in ROC curves. *Journal of Computer-Aided Molecular Design*, 22(3-4):141–146.

Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., Johnston, S. E., Vrcic, A., Wong, B., Khan, M., Asiedu, J., Narayan, R., Mader, C. C., Subramanian, A., and Golub, T. R. (2017). The Drug Repurposing Hub: A next-generation drug library and information resource. *Nature Medicine*, 23(4):405–408.

D'Amato, R. J., Loughnan, M. S., Flynn, E., and Folkman, J. (1994). Thalidomide is an inhibitor of angiogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 91(9):4082–4085.

Daminelli, S., Haupt, V. J., Reimann, M., and Schroeder, M. (2012). Drug repositioning through incomplete bi-cliques in an integrated drug-target-disease network. *Integrative Biology: Quantitative Biosciences from Nano to Macro*, 4(7):778–788.

Daum, R. S., Gupta, S., Sabbagh, R., and Milewski, W. M. (1992). Characterization of Staphylococcus aureus Isolates with Decreased Susceptibility to Vancomycin and Teicoplanin: Isolation and Purification of a Constitutively Produced Protein Associated with Decreased Susceptibility. *The Journal of Infectious Diseases*, 166(5):1066–1072.

Daya, S. (2003). Recurrent spontaneous early pregnancy loss and low dose aspirin. *Minerva Ginecologica*, 55(5):441–449.

Debnath, A. K., Radigan, L., and Jiang, S. (1999). Structure-based identification of small molecule antiviral compounds targeted to the gp41 core structure of the human immunodeficiency virus type 1. *Journal of Medicinal Chemistry*, 42(17):3203–3209.

Deotarse, P. P., Jain, A. S., Baile, M. B., Kolhe, N. S., and Kulkarni, A. A. (2015). Drug repositioning: A review. *Int. J. Pharma. Res Rev*, 4:51–58.

DeWitte, R. S. and Shakhnovich, E. I. (1996). SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *Journal of the American Chemical Society*, 118(47):11733–11744.

DiMasi, J. A., Feldman, L., Seckler, A., and Wilson, A. (2010). Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs. *Clinical Pharmacology & Therapeutics*, 87(3):272–277.

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47:20–33.

DiMasi, J. A., Grabowski, H. G., and Vernon, J. (1995). R&D Costs, Innovative Output and Firm Size in the Pharmaceutical Industry. *International Journal of the Economics of Business*, 2(2):201–219.

DiMasi, J. A., Hansen, R. W., and Grabowski, H. G. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185.

DiMasi, J. A., Hansen, R. W., Grabowski, H. G., and Lasagna, L. (1991). Cost of innovation in the pharmaceutical industry. *Journal of Health Economics*, 10(2):107–142.

Dixon, J. S. (1997). Evaluation of the CASP2 docking section. *Proteins*, Suppl 1:198–204.

Dou, J.-L., Jiang, Y.-W., Xie, J.-Q., and Zhang, X.-G. (2016). New Is Old, and Old Is New: Recent Advances in Antibiotic-Based, Antibiotic-Free and Ethnomedical Treatments against Methicillin-Resistant Staphylococcus aureus Wound Infections. *International Journal of Molecular Sciences*, 17(5).

Dudley, J. T., Deshpande, T., and Butte, A. J. (2011a). Exploiting drug–disease relationships for computational drug repositioning. *Briefings in Bioinformatics*, 12(4):303–311.

Dudley, J. T., Sirota, M., Shenoy, M., Pai, R., Roedder, S., Chiang, A. P., Morgan, A. A., Sarwal, M., Pasricha, P. J., and Butte, A. J. (2011b). Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine*, 3(96):96ra76.

Durongpisitkul, K., Gururaj, V. J., Park, J. M., and Martin, C. F. (1995). The prevention of coronary artery aneurysm in Kawasaki disease: A meta-analysis on the efficacy of aspirin and immunoglobulin treatment. *Pediatrics*, 96(6):1057–1061.

Durrant, J. D., Amaro, R. E., Xie, L., Urbaniak, M. D., Ferguson, M. A. J., Haapalainen, A., Chen, Z., Guilmi, A. M. D., Wunder, F., Bourne, P. E., and McCammon, J. A. (2010). A Multidimensional Strategy to Detect Polypharmacological Targets in the Absence of Structural and Sequence Homology. *PLoS Computational Biology*, 6(1).

Durrant, J. D. and McCammon, J. A. (2010). NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling*, 50(10):1865–1871.

Ehrt, C., Brinkjost, T., and Koch, O. (2018). A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLoS Computational Biology*, 14(11).

Empereur-mot, C., Guillemain, H., Latouche, A., Zagury, J.-F., Viallon, V., and Montes, M. (2015). Predictiveness curves in virtual screening. *Journal of Cheminformatics*, 7.

Engemann, J. J., Carmeli, Y., Cosgrove, S. E., Fowler, V. G., Bronstein, M. Z., Trivette, S. L., Briggs, J. P., Sexton, D. J., and Kaye, K. S. (2003). Adverse Clinical and Economic Outcomes Attributable to Methicillin Resistance among Patients with Staphylococcus aureus Surgical Site Infection. *Clinical Infectious Diseases*, 36(5):592–598.

Ericksen, S. S., Wu, H., Zhang, H., Michael, L. A., Newton, M. A., Hoffmann, F. M., and Wildman, S. A. (2017). Machine Learning Consensus Scoring Improves Performance Across Targets in Structure-Based Virtual Screening. *Journal of chemical information and modeling*, 57(7):1579–1590.

Esther, K., Claire, S., and Didier, R. (2008). How to Measure the Similarity Between Protein Ligand-Binding Sites? *Current Computer-Aided Drug Design*, 4(3):209–220.

European Centre for Disease Prevention and Control (2015). Antimicrobial resistance surveillance in Europe 2014. Technical report, European Centre for Disease Prevention and Control.

Fang, J., Yang, R., Gao, L., Yang, S., Pang, X., Li, C., He, Y., Liu, A.-L., and Du, G.-H. (2015). Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery. *Molecular Diversity*, 19(1):149–162.

Feher, M. (2006). Consensus scoring for protein–ligand interactions. *Drug Discovery Today*, 11(9):421–428.

Fernandes, M. X., Kairys, V., and Gilson, M. K. (2004). Comparing Ligand Interactions with Multiple Receptors via Serial Docking. *Journal of Chemical Information and Computer Sciences*, 44(6):1961–1970.

Fleming, A. (1943). STREPTOCOCCAL MENINGITIS TREATED WITH PENICILLIN.: MEASUREMENT OF BACTERIOSTATIC POWER OF BLOOD AND CEREBROSPINAL FLUID. *The Lancet*, 242(6267):434–438.

Fletcher, J. T., Finlay, J. A., Callow, M. E., Callow, J. A., and Ghadiri, M. R. (2007). A Combinatorial Approach to the Discovery of Biocidal Six-Residue Cyclic D, L-α-Peptides Against Bacteria MRSA and E. coli. and the Biofouling Algae Ulva linza and Navicula perminuta. *Chemistry (Weinheim an der Bergstrasse, Germany)*, 13(14):4008–4013.

Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., Wall, D., Wang, L., Brown-Driver, V., Froelich, J. M., C, K. G., King, P., McCarthy, M., Malone, C., Misiner, B., Robbins, D., Tan, Z., Zhu Zy, Z.-y., Carr, G., Mosca, D. A., Zamudio, C., Foulkes, J. G., and Zyskind, J. W. (2002). A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. *Molecular Microbiology*, 43(6):1387–1400.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749.

Gani, O. A. B. S. M. (2007). Signposts of docking and scoring in drug design. *Chemical Biology & Drug Design*, 70(4):360–365.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, 45(Database issue):D945.

Genilloud, O. (2014). The re-emerging role of microbial natural products in antibiotic discovery. *Antonie van Leeuwenhoek*, 106(1):173–188.

Giesbrecht, P., Kersten, T., Maidhof, H., and Wecke, J. (1998). Staphylococcal Cell Wall: Morphogenesis and Fatal Variations in the Presence of Penicillin. *Microbiology and Molecular Biology Reviews*, 62(4):1371–1414.

Gilbert-Girard, S., Savijoki, K., Yli-Kauhaluoma, J., and Fallarero, A. (2020). Screening of FDA-Approved Drugs Using a 384-Well Plate-Based Biofilm Platform: The Case of Fingolimod. *Microorganisms*, 8(11).

Gimeno, A., Ojeda-Montes, M. J., Tomás-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Mulero, M., Pujadas, G., and Garcia-Vallvé, S. (2019). The Light and Dark Sides of Virtual Screening: What Is There to Know? *International Journal of Molecular Sciences*, 20(6).

Ginn, C. M., Willett, P., and Bradshaw, J. (2000). Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design*, 20(1):1–16.

Gleeson, M. P., Hersey, A., Montanari, D., and Overington, J. (2011). Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nature reviews. Drug discovery*, 10(3):197–208.

Gohlke, H., Hendlich, M., and Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions11Edited by R. Huber. *Journal of Molecular Biology*, 295(2):337–356.

Gomes, E. S., Schuch, V., and de Macedo Lemos, E. G. (2014). Biotechnology of polyketides: New breath of life for the novel antibiotic genetic pathways discovery through metagenomics. *Brazilian Journal of Microbiology*, 44(4):1007–1034.

Good, A. C. and Oprea, T. I. (2008 Mar-Apr). Optimization of CAMD techniques 3. Virtual screening enrichment studies: A help or hindrance in tool selection? *Journal of Computer-Aided Molecular Design*, 22(3-4):169–178.

Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*, 28(7):849–857.

Grabowski, H. (2011). The Evolution of the Pharmaceutical Industry Over the Past 50 Years: A Personal Reflection. *International Journal of the Economics of Business*, 18(2):161–176.

Graham, P. L., Lin, S. X., and Larson, E. L. (2006). A U.S. population-based survey of Staphylococcus aureus colonization. *Annals of Internal Medicine*, 144(5):318–325.

Graul, A. I., Sorbera, L., Pina, P., Tell, M., Cruces, E., Rosa, E., Stringer, M., Castañer, R., and Revel, L. (2010 Jan-Feb). The Year's New Drugs & Biologics - 2009. *Drug News & Perspectives*, 23(1):7–36.

Grundmann, K., Jaschonek, K., Kleine, B., Dichgans, J., and Topka, H. (2003). Aspirin non-responder status in patients with recurrent cerebral ischemic attacks. *Journal of Neurology*, 250(1):63–66.

Guex, N. and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *ELECTROPHORESIS*, 18(15):2714–2723.

Haupt, V. J., Daminelli, S., and Schroeder, M. (2013). Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS ONE*, 8(6).

Haupt, V. J. and Schroeder, M. (2011). Old friends in new guise: Repositioning of known drugs with structural bioinformatics. *Briefings in Bioinformatics*, 12(4):312–326.

Hawkins, P. C. D., Warren, G. L., Skillman, A. G., and Nicholls, A. (2008 Mar-Apr). How to do an evaluation: Pitfalls and traps. *Journal of Computer-Aided Molecular Design*, 22(3-4):179–190.

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1):40–51.

Hidron, A. I., Edwards, J. R., Patel, J., Horan, T. C., Sievert, D. M., Pollock, D. A., Fridkin, S. K., and Facilities, N. H. S. N. T. a. P. N. H. S. N. (2008). Antimicrobial-Resistant Pathogens Associated With Healthcare-Associated Infections: Annual Summary of Data Reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2006–2007. *Infection Control & Hospital Epidemiology*, 29(11):996–1011.

Hiramatsu, K., Hanaki, H., Ino, T., Yabuta, K., Oguri, T., and Tenover, F. C. (1997). Methicillin-resistant Staphylococcus aureus clinical strain with reduced vancomycin susceptibility. *Journal of Antimicrobial Chemotherapy*, 40(1):135–136.

Holm, L. and Rosenström, P. (2010). Dali server: Conservation mapping in 3D. *Nucleic Acids Research*, 38(Web Server issue):W545–W549.

Hopkins, A. L. (2008). Network pharmacology: The next paradigm in drug discovery. *Nature Chemical Biology*, 4(11):682–690.

Hopkins, A. L. (2009). Drug discovery: Predicting promiscuity. *Nature*, 462(7270):167–168.

Hopkins, A. L., Mason, J. S., and Overington, J. P. (2006). Can we rationally design promiscuous drugs? *Current Opinion in Structural Biology*, 16(1):127–136.

Horvath, D. (1997). A virtual screening approach applied to the search for trypanothione reductase inhibitors. *Journal of Medicinal Chemistry*, 40(15):2412–2423.

Hsu, K.-C., Chen, Y.-F., Lin, S.-R., and Yang, J.-M. (2011). iGEMDOCK: A graphical environment of enhancing GEMDOCK using pharmacological interactions and post-screening analysis. *BMC Bioinformatics*, 12(Suppl 1):S33.

Hu, Q., Peng, Z., Sutton, S. C., Na, J., Kostrowicki, J., Yang, B., Thacher, T., Kong, X., Mattaparti, S., Zhou, J. Z., Gonzalez, J., Ramirez-Weinhouse, M., and Kuki, A. (2012). Pfizer Global Virtual Library (PGVL): A Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information. *ACS Combinatorial Science*, 14(11):579–589.

Huang, N., Kalyanaraman, C., Bernacki, K., and Jacobson, M. P. (2006a). Molecular mechanics methods for predicting protein–ligand binding. *Physical Chemistry Chemical Physics*, 8(44):5166–5177.

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006b). Benchmarking Sets for Molecular Docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801.

Hubbard, R. and Bayarri, M. J. (2003). Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing. *The American Statistician*, 57(3):171–178.

Hung, C.-L. and Chen, C.-C. (2014). Computational approaches for drug discovery. *Drug Development Research*, 75(6):412–418.

Itaya, M. (1995). An estimation of minimal genome size required for life. *FEBS letters*, 362(3):257–260.

Jacobsson, M. and Karlén, A. (2006). Ligand Bias of Scoring Functions in Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling*, 46(3):1334–1343.

Jain, A. N. (2003). Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *Journal of Medicinal Chemistry*, 46(4):499–511.

Jain, A. N. (2009). Effects of Protein Conformation in Docking: Improved Pose Prediction through Protein Pocket Adaptation. *Journal of computer-aided molecular design*, 23(6):355.

Jain, A. N. and Nicholls, A. (2008). Recommendations for evaluation of computational methods. *Journal of Computer-Aided Molecular Design*, 22(3):133–139.

Jevons, M. P. (1961). "Celbenin" - resistant Staphylococci. *British Medical Journal*, 1(5219):124–125.

Ji, Y., Zhang, B., Van, S. F., Horn, n., Warren, P., Woodnutt, G., Burnham, M. K., and Rosenberg, M. (2001). Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science (New York, N.Y.)*, 293(5538):2266–2269.

Jin, G. and Wong, S. T. (2014). Toward better drug repositioning: Prioritizing and integrating existing methods into efficient pipelines. *Drug discovery today*, 19(5):637–644.

Johnson, A. P., Pearson, A., and Duckworth, G. (2005). Surveillance and epidemiology of MRSA bacteraemia in the UK. *The Journal of Antimicrobial Chemotherapy*, 56(3):455–462.

Johnson, T. W., Dress, K. R., and Edwards, M. (2009). Using the Golden Triangle to optimize clearance and oral absorption. *Bioorganic & Medicinal Chemistry Letters*, 19(19):5560–5564.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen. *Journal of Molecular Biology*, 267(3):727–748.

Kairys, V., Fernandes, M. X., and Gilson, M. K. (2006). Screening Drug-Like Compounds by Docking to Homology Models: A Systematic Study. *Journal of Chemical Information and Modeling*, 46(1):365–379.

Kalinina, O. V., Wichmann, O., Apic, G., and Russell, R. B. (2011). Combinations of Protein-Chemical Complex Structures Reveal New Targets for Established Drugs. *PLoS Computational Biology*, 7(5).

Karaman, R., Jubeh, B., and Breijyeh, Z. (2020). Resistance of Gram-Positive Bacteria to Current Antibacterial Agents and Overcoming Approaches. *Molecules (Basel, Switzerland)*, 25(12).

Kenny, B., Ballard, S., Blagg, J., and Fox, D. (1997). Pharmacological Options in the Treatment of Benign Prostatic Hyperplasia. *Journal of Medicinal Chemistry*, 40(9):1293–1315.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. (2020). PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., and Bryant, S. H. (2016). PubChem Substance and Compound databases. *Nucleic Acids Research*, 44(Database issue):D1202.

Kim, S. J., Chang, J., and Singh, M. (2015). Peptidoglycan Architecture of Gram-positive Bacteria by Solid-State NMR. *Biochimica et biophysica acta*, 1848(0):350.

Kinnings, S. L., Xie, L., Fung, K. H., Jackson, R. M., Xie, L., and Bourne, P. E. (2010). The Mycobacterium tuberculosis Drugome and Its Polypharmacological Implications. *PLoS Computational Biology*, 6(11).

Kirchmair, J., Markt, P., Distinto, S., Wolber, G., and Langer, T. (2008). Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? *Journal of Computer-Aided Molecular Design*, 22(3):213–228.

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949.

Klevens, R. M., Edwards, J. R., Richards, C. L., Horan, T. C., Gaynes, R. P., Pollock, D. A., and Cardo, D. M. (2007). Estimating Health Care-Associated Infections and Deaths in U.S. Hospitals, 2002. *Public Health Reports*, 122(2):160–166.

Knegtel, R. M. A., Kuntz, I. D., and Oshiro, C. M. (1997). Molecular docking to ensembles of protein structures. *Journal of Molecular Biology*, 266(2):424–440.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. (2011). DrugBank 3.0: A comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, 39(Database issue):D1035.

Koch, U., Hamacher, M., and Nussbaumer, P. (2014). Cheminformatics at the interface of medicinal chemistry and proteomics. *Biochimica Et Biophysica Acta*, 1844(1 Pt A):156–161.

Koebel, M. R., Schmadeke, G., Posner, R. G., and Sirimulla, S. (2016). AutoDock VinaXB: Implementation of XBSF, new empirical halogen bond scoring function, into AutoDock Vina. *Journal of Cheminformatics*, 8(1):27.

Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013). Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *Journal of chemical information and modeling*, 53(8):1893–1904.

Konc, J., Česnik, T., Konc, J. T., Penca, M., and Janežič, D. (2012). ProBiS-Database: Precalculated Binding Site Similarities and Local Pairwise Alignments of PDB Structures. *Journal of Chemical Information and Modeling*, 52(2):604.

Kopp, U., Roos, M., Wecke, J., and Labischinski, H. (1996). Staphylococcal peptidoglycan interpeptide bridge biosynthesis: A novel antistaphylococcal target? *Microbial drug resistance (Larchmont, N.Y.)*, 2(1):29–41.

Korb, O., Olsson, T. S. G., Bowden, S. J., Hall, R. J., Verdonk, M. L., Liebeschuetz, J. W., and Cole, J. C. (2012). Potential and Limitations of Ensemble Docking. *Journal of Chemical Information and Modeling*, 52(5):1262–1274.

Korb, O., Stützle, T., and Exner, T. E. (2007). An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intelligence*, 1(2):115–134.

Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis*. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98–104.

Koteyko, N., Nerlich, B., Crawford, P., and Wright, N. (2008). 'Not rocket science' or 'No silver bullet'? Media and Government Discourses about MRSA and Cleanliness. *Applied Linguistics*, 29(2):223–243.

Kouznetsova, J., Sun, W., Martínez-Romero, C., Tawa, G., Shinn, P., Chen, C. Z., Schimmer, A., Sanderson, P., McKew, J. C., Zheng, W., and García-Sastre, A. (2014). Identification of 53 compounds that block Ebola virus-like particle entry via a repurposing screen of approved drugs. *Emerging Microbes & Infections*, 3(1):1–7.

Krammer, A., Kirchhoff, P. D., Jiang, X., Venkatachalam, C. M., and Waldman, M. (2005). LigScore: A novel scoring function for predicting binding affinities. *Journal of Molecular Graphics & Modelling*, 23(5):395–407.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2):269–288.

Kurbatova, N., Chartier, M., Zylber, M. I., and Najmanovich, R. (2013). IsoCleft Finder – a web-based tool for the detection and analysis of protein binding-site geometric and chemical similarities. *F1000Research*, 2.

Kurosu, M., Siricilla, S., and Mitachi, K. (2013). Advances in MRSA drug discovery: Where are we and where do we need to be? *Expert Opinion on Drug Discovery*, 8(9):1095–1116.

Labischinski, H. (1992). Consequences of the interaction of beta-lactam antibiotics with penicillin binding proteins from sensitive and resistant Staphylococcus aureus strains. *Medical microbiology and immunology*, 181(5):241–265.

Lagarde, N., Ben Nasr, N., Jérémie, A., Guillemain, H., Laville, V., Labib, T., Zagury, J.-F., and Montes, M. (2014). NRLiSt BDB, the manually curated nuclear receptors ligands and structures benchmarking database. *Journal of Medicinal Chemistry*, 57(7):3117–3125.

Landrum, G. (2010). RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org.

Lang, P. T., Brozell, S. R., Mukherjee, S., Pettersen, E. F., Meng, E. C., Thomas, V., Rizzo, R. C., Case, D. A., James, T. L., and Kuntz, I. D. (2009). DOCK 6: Combining techniques to model RNA–small molecule complexes. *RNA*, 15(6):1219–1230.

Lau, Q. Y., Ng, F. M., Cheong, J. W. D., Yap, Y. Y. A., Tan, Y. Y. F., Jureen, R., Hill, J., and Chia, C. S. B. (2015a). Discovery of an ultra-short linear antibacterial tetrapeptide with anti-MRSA activity from a structure–activity relationship study. *European Journal of Medicinal Chemistry*, 105:138–144.

Lau, Q. Y., Tan, Y. Y. F., Goh, V. C. Y., Lee, D. J. Q., Ng, F. M., Ong, E. H. Q., Hill, J., and Chia, C. S. B. (2015b). An FDA-Drug Library Screen for Compounds with Bioactivities against Meticillin-Resistant Staphylococcus aureus (MRSA). *Antibiotics*, 4(4):424–434.

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

Leach, A. R., Shoichet, B. K., and Peishoff, C. E. (2006). Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps. *Journal of Medicinal Chemistry*, 49(20):5851–5855.

Leeson, P. D. and Springthorpe, B. (2007). The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews. Drug Discovery*, 6(11):881–890.

Lenz, W. (1988). A short history of thalidomide embryopathy. *Teratology*, 38(3):203–215.

Lessel, U., Wellenzohn, B., Lilienthal, M., and Claussen, H. (2009). Searching Fragment Spaces with Feature Trees. *Journal of Chemical Information and Modeling*, 49(2):270–279.

Levinthal, C., Wodak, S. J., Kahn, P., and Dadivanian, A. K. (1975). Hemoglobin interaction in sickle cell fibers. I: Theoretical approaches to the molecular contacts. *Proceedings of the National Academy of Sciences of the United States of America*, 72(4):1330–1334.

Li, G.-B., Yang, L.-L., Wang, W.-J., Li, L.-L., and Yang, S.-Y. (2013). ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein–Ligand Interactions. *Journal of Chemical Information and Modeling*, 53(3):592–600.

Li, Y., Liu, Z., Li, J., Han, L., Liu, J., Zhao, Z., and Wang, R. (2014). Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *Journal of Chemical Information and Modeling*, 54(6):1700–1716.

Ligon, B. L. (2004). Penicillin: Its discovery and early development. *Seminars in Pediatric Infectious Diseases*, 15(1):52–57.

Lindblad, W. J. (2008). Review Paper: Considerations for Determining if a Natural Product Is an Effective Wound-Healing Agent. *The International Journal of Lower Extremity Wounds*, 7(2):75–81.

Lionta, E., Spyrou, G., Vassilatis, D. K., and Cournia, Z. (2014). Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Current Topics in Medicinal Chemistry*, 14(16):1923–1938.

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

Liu, C., Bayer, A., Cosgrove, S. E., Daum, R. S., Fridkin, S. K., Gorwitz, R. J., Kaplan, S. L., Karchmer, A. W., Levine, D. P., Murray, B. E., J Rybak, M., Talan, D. A., Chambers, H. F., and Infectious Diseases Society of America (2011). Clinical practice guidelines by the infectious diseases society of america for the treatment of methicillin-resistant Staphylococcus aureus infections in adults and children. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 52(3):e18–55.

Liu, J. and Wang, R. (2015). Classification of Current Scoring Functions. *Journal of Chemical Information and Modeling*, 55(3):475–482.

Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., and Wang, R. (2017). Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. *Accounts of Chemical Research*, 50(2):302–309.

Lowy, F. D. (1998). Staphylococcus aureus Infections. *New England Journal of Medicine*, 339(8):520–532.

Luo, H., Lin, Y., Gao, F., Zhang, C.-T., and Zhang, R. (2014). DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Research*, 42(D1):D574–D580.

Luo, H., Lin, Y., Liu, T., Lai, F.-L., Zhang, C.-T., Gao, F., and Zhang, R. (2021). DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. *Nucleic Acids Research*, 49(D1):D677.

Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K., Tolmachev, A. A., Shoichet, B. K., Roth, B. L., and Irwin, J. J. (2019). Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229.

Macherla, V., Hensler, M., Houson, H., Adhikari, P., Thienphrapa, W., Beverage, J., Nizet, V., and Esquenazi, E. (2013). Optimized Marine Natural Products Discovery and Screening: Searching for Novel "Superbug" Antibiotics. In *Planta Medica*, volume 79, page PK12.

Manallack, D. T., Pitt, W. R., Gancia, E., Montana, J. G., Livingstone, D. J., Ford, M. G., and Whitley, D. C. (2002). Selecting Screening Candidates for Kinase and G Protein-Coupled Receptor Targets Using Neural Networks. *Journal of Chemical Information and Computer Sciences*, 42(5):1256–1262.

Masters, L., Eagon, S., and Heying, M. (2020). Evaluation of consensus scoring methods for AutoDock Vina, smina and idock. *Journal of Molecular Graphics and Modelling*, 96:107532.

Mathieu, M. P. (2008). *Parexel's Bio/Pharmaceutical R&D Statistical Sourcebook 2008/2009*. Parexel Publishing, United States.

Mencher, S. K. and Wang, L. G. (2005). Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clinical Pharmacology*, 5:3.

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., Hersey, A., and Leach, A. R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(Database issue):D930–D940.

Meng, E. C., Shoichet, B. K., and Kuntz, I. D. (1992). Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry*, 13(4):505–524.

Meng, X.-Y., Zhang, H.-X., Mezei, M., and Cui, M. (2011). Molecular Docking: A powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157.

Merz Jr, K. M., Ringe, D., and Reynolds, C. H. (2010). *Drug Design: Structure-and Ligand-Based Approaches*. Cambridge University Press.

Mestres, J., Gregori-Puigjané, E., Valverde, S., and Solé, R. V. (2008). Data completeness—the Achilles heel of drug-target networks. *Nature Biotechnology*, 26(9):983–984.

Mestres, J., Gregori-Puigjané, E., Valverde, S., and Solé, R. V. (2009). The topology of drug-target interaction networks: Implicit dependence on drug properties and target families. *Molecular bioSystems*, 5(9):1051–1057.

Mitchell, J. B. O., Laskowski, R. A., Alex, A., Forster, M. J., and Thornton, J. M. (1999a). BLEEP—potential of mean force describing protein–ligand interactions: II. Calculation of binding energies and comparison with experimental data. *Journal of Computational Chemistry*, 20(11):1177–1185.

Mitchell, J. B. O., Laskowski, R. A., Alex, A., and Thornton, J. M. (1999b). BLEEP—potential of mean force describing protein–ligand interactions: I. Generating potential. *Journal of Computational Chemistry*, 20(11):1165–1176.

Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., and Corbeil, C. R. (2008). Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *British Journal of Pharmacology*, 153(Suppl 1):S7.

Montorsi, F. (1998). Clinical safety of oral sildenafil citrate (VIAGRATM) in the treatment of erectile dysfunction—by Morales et al. *International Journal of Impotence Research*, 10(2):73–74.

Mooij, W. T. M. and Verdonk, M. L. (2005). General and targeted statistical potentials for protein–ligand interactions. *Proteins: Structure, Function, and Bioinformatics*, 61(2):272–287.

Morales, A., Gingell, C., Collins, M., Wicker, P. A., and Osterloh, I. H. (1998). Clinical safety of oral sildenafil citrate (VIAGRA TM) in the treatment of erectile dysfunction. *International Journal of Impotence Research*, 10(2):69–73.

Morphy, R. and Rankovic, Z. (2007). Fragments, network biology and designing multiple ligands. *Drug Discovery Today*, 12(3-4):156–160.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of computational chemistry*, 30(16):2785–2791.

Morris, G. M. and Lim-Wilby, M. (2008). Molecular docking. *Methods in Molecular Biology (Clifton, N.J.)*, 443:365–382.

Muegge, I. (1999). The Effect of Small Changes in Protein Structure on Predicted Binding Modes of Known Inhibitors of Influenza Virus Neuraminidase: PMF- Scoring in DOCK4. *Medicinal Chemistry Research*, 9(7):490–500.

Mullard, A. (2012). Drug repurposing programmes get lift off. *Nature Reviews Drug Discovery*, 11(7):505–506.

Mullard, A. (2014). Bank tests drug development waters. *Nature Reviews Drug Discovery*, 13(9):643–644.

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594.

Mysinger, M. M. and Shoichet, B. K. (2010). Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *Journal of Chemical Information and Modeling*, 50(9):1561–1573.

Naik, V. and Mahajan, G. (2013). Quorum sensing: A non-conventional target for antibiotic discovery. *Natural Product Communications*, 8(10):1455–1458.

Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., and Greco, D. (2013). Drug repositioning: A machine-learning approach through data integration. *Journal of Cheminformatics*, 5(1):30.

Nathwani, D., Morgan, M., Masterton, R. G., Dryden, M., Cookson, B. D., French, G., Lewis, D., and British Society for Antimicrobial Chemotherapy Working Party on Community-onset MRSA Infections (2008). Guidelines for UK practice for the diagnosis and management of methicillin-resistant Staphylococcus aureus (MRSA) infections presenting in the community. *The Journal of Antimicrobial Chemotherapy*, 61(5):976–994.

Neyman, J. and Pearson, E. S. (1992). On the Problem of the Most Efficient Tests of Statistical Hypotheses. In Kotz, S. and Johnson, N. L., editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer Series in Statistics, pages 73–108. Springer, New York, NY.

Ng, M. C. K., Fong, S., and Siu, S. W. I. (2015). PSOVina: The hybrid particle swarm optimization algorithm for protein-ligand docking. *Journal of Bioinformatics and Computational Biology*, 13(3):1541007.

Ngan, C.-H., Hall, D. R., Zerbe, B., Grove, L. E., Kozakov, D., and Vajda, S. (2012). FTSite: High accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics*, 28(2):286–287.

Nicolaou, C. A., Watson, I. A., Hu, H., and Wang, J. (2016). The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *Journal of Chemical Information and Modeling*, 56(7):1253–1266.

Nicolaou, K. C., Hughes, R., Pfefferkorn, J. A., Barluenga, S., and Roecker, A. J. (2001). Combinatorial Synthesis through Disulfide Exchange: Discovery of Potent Psammaplin A Type Antibacterial Agents Active against Methicillin-Resistant Staphylococcus aureus (MRSA). *Chemistry – A European Journal*, 7(19):4280–4295.

Niu, H., Yee, R., Cui, P., Tian, L., Zhang, S., Shi, W., Sullivan, D., Zhu, B., Zhang, W., and Zhang, Y. (2017). Identification of Agents Active against Methicillin-Resistant Staphylococcus aureus USA300 from a Clinical Compound Library. *Pathogens*, 6(3).

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33.

O'Boyle, N. M., Liebeschuetz, J. W., and Cole, J. C. (2009). Testing Assumptions and Hypotheses for Rescoring Success in Protein-Ligand Docking. *Journal of Chemical Information and Modeling*, 49(8):1871–1878.

Oda, A., Tsuchida, K., Takakura, T., Yamaotsu, N., and Hirono, S. (2006). Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling*, 46(1):380–391.

Onawole, A. T., Kolapo, T. U., Sulaiman, K. O., and Adegoke, R. O. (2018). Structure based virtual screening of the Ebola virus trimeric glycoprotein using consensus scoring. *Computational Biology and Chemistry*, 72:170–180.

O'Neill, J. and Grande-Bretagne (2014). *Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations*. Review on Antimicrobial Resistance.

Oprea, T. I. and Mestres, J. (2012). Drug Repurposing: Far Beyond New Targets for Old Drugs. *The AAPS Journal*, 14(4):759–763.

Oprea, T. I. and Overington, J. P. (2015 Jul-Aug). Computational and Practical Aspects of Drug Repositioning. *Assay and Drug Development Technologies*, 13(6):299–306.

Osterberg, F., Morris, G. M., Sanner, M. F., Olson, A. J., and Goodsell, D. S. (2002). Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins*, 46(1):34–40.

Pagès, J.-M., James, C. E., and Winterhalter, M. (2008). The porin and the permeating antibiotic: A selective diffusion barrier in Gram-negative bacteria. *Nature Reviews Microbiology*, 6(12):893–903.

Palacio-Rodríguez, K., Lans, I., Cavasotto, C. N., and Cossio, P. (2019). Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Scientific Reports*, 9.

Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S., and Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nature Biotechnology*, 24(7):805–815.

Park, H., Eom, J.-W., and Kim, Y.-H. (2014). Consensus Scoring Approach To Identify the Inhibitors of AMP-Activated Protein Kinase $A2$ with Virtual Screening. *Journal of Chemical Information and Modeling*, 54(7):2139–2146.

Park, K., Moreland, R. B., Goldstein, I., Atala, A., and Traish, A. (1998). Sildenafil Inhibits Phosphodiesterase Type 5 in Human Clitoral Corpus Cavernosum Smooth Muscle. *Biochemical and Biophysical Research Communications*, 249(3):612–617.

Patel, H., Lucas, X., Bendik, I., Günther, S., and Merfort, I. (2015). Target Fishing by Cross-Docking to Explain Polypharmacological Effects. *ChemMedChem*, 10(7):1209–1217.

Paterson, D. L. (1999). Reduced susceptibility of Staphylococcus aureus to vancomycin–a review of current knowledge. *Communicable diseases intelligence*, 23(3):69–73.

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

Pei, J., Yin, N., Ma, X., and Lai, L. (2014). Systems Biology Brings New Dimensions for Structure-Based Drug Design. *Journal of the American Chemical Society*, 136(33):11556–11565.

Pereira, F., Latino, D. A. R. S., and Gaudêncio, S. P. (2015). QSAR-Assisted Virtual Screening of Lead-Like Molecules from Marine and Microbial Natural Sources for Anti-tumor and Antibiotic Drug Discovery. *Molecules*, 20(3):4848–4873.

Pereira, J. C., Caffarena, E. R., and dos Santos, C. N. (2016). Boosting Docking-Based Virtual Screening with Deep Learning. *Journal of Chemical Information and Modeling*, 56(12):2495–2506.

Perez-Nueno, V. I., Pettersson, S., Ritchie, D. W., Borrell, J. I., and Teixidó, J. (2009). Discovery of Novel HIV Entry Inhibitors for the CXCR4 Receptor by Prospective Virtual Screening. *Journal of Chemical Information and Modeling*, 49(4):810–823.

Perlmutter, J. I., Forbes, L. T., Krysan, D. J., Ebsworth-Mojica, K., Colquhoun, J. M., Wang, J. L., Dunman, P. M., and Flaherty, D. P. (2014). Repurposing the Antihistamine Terfenadine for Antimicrobial Activity against Staphylococcus aureus. *Journal of Medicinal Chemistry*, 57(20):8540–8562.

Peters, J.-U., Schnider, P., Mattei, P., and Kansy, M. (2009). Pharmacological promiscuity: Dependence on compound properties and target specificity in a set of recent Roche compounds. *ChemMedChem*, 4(4):680–686.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612.

Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E. C., Pettersen, E. F., Huang, C. C., Datta, R. S., Sampathkumar, P., Madhusudhan, M. S., Sjölander, K., Ferrin, T. E., Burley, S. K., and Sali, A. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research*, 39(Database issue):D465–474.

Pinzi, L., Caporuscio, F., and Rastelli, G. (2018). Selection of protein conformations for structure-based polypharmacology studies. *Drug Discovery Today*, 23(11):1889–1896.

Pitscheider, M., Mäusbacher, N., and Sieber, S. A. (2012). Antibiotic activity and target discovery of three-membered natural product-derived heterocycles in pathogenic bacteria. *Chemical Science*, 3(6):2035–2041.

Plewczynski, D., Łaźniewski, M., Augustyniak, R., and Ginalski, K. (2011). Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of Computational Chemistry*, 32(4):742–755.

Prakash, A., Gordon, L. B., Kleinman, M. E., Gurary, E. B., Massaro, J., D'Agostino, Sr, R., Kieran, M. W., Gerhard-Herman, M., and Smoot, L. (2018). Cardiac Abnormalities in Patients With Hutchinson-Gilford Progeria Syndrome. *JAMA Cardiology*, 3(4):326–334.

Prieto, J. M., Rapún-Araiz, B., Gil, C., Penadés, J. R., Lasa, I., and Latasa, C. (2020). Inhibiting the two-component system GraXRS with verteporfin to combat Staphylococcus aureus infections. *Scientific Reports*, 10.

Pujol, A., Mosca, R., Farrés, J., and Aloy, P. (2010). Unveiling the role of network and systems biology in drug discovery. *Trends in Pharmacological Sciences*, 31(3):115–123.

Rajamuthiah, R., Fuchs, B. B., Conery, A. L., Kim, W., Jayamani, E., Kwon, B., Ausubel, F. M., and Mylonakis, E. (2015). Repurposing Salicylanilide Anthelmintic Drugs to Combat Drug Resistant Staphylococcus aureus. *PLoS ONE*, 10(4).

Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 261(3):470–489.

Ratcliffe, R. W., Wilkening, R. R., Wildonger, K. J., Waddell, S. T., Santorelli, G. M., Parker, D. L., Morgan, J. D., Blizzard, T. A., Hammond, M. L., Heck, J. V., Huber, J., Kohler, J., Dorso, K. L., St. Rose, E., Sundelof, J. G., May, W. J., and Hammond, G. G. (1999). Synthesis and properties of 2-(naphthosultamyl)methyl-carbapenems with potent anti-MRSA activity: Discovery of L-786,392. *Bioorganic & Medicinal Chemistry Letters*, 9(5):679–684.

Ravindranath, P. A., Forli, S., Goodsell, D. S., Olson, A. J., and Sanner, M. F. (2015). AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *PLoS Computational Biology*, 11(12).

Ravindranath, P. A. and Sanner, M. F. (2016). AutoSite: An automated approach for pseudo-ligands prediction—from ligand-binding sites identification to predicting key ligand atoms. *Bioinformatics*, 32(20):3142–3149.

Rawlins, M. D., Wexler, N. S., Wexler, A. R., Tabrizi, S. J., Douglas, I., Evans, S. J. W., and Smeeth, L. (2016). The Prevalence of Huntington's Disease. *Neuroepidemiology*, 46(2):144–153.

Réau, M., Langenfeld, F., Zagury, J.-F., Lagarde, N., and Montes, M. (2018). Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Frontiers in Pharmacology*, 9.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2):173–175.

Reynolds, P. E., Actor, P., Daneo-Moore, L., Higgings, M. L., Salton, M. R. J., and Shockman, G. D. (1988). Antibiotic inhibition of bacterial cell surface assembly and function. *American Society for Microbiology, Washington, DC*.

Rohrer, S. G. and Baumann, K. (2009). Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *Journal of Chemical Information and Modeling*, 49(2):169–184.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94.

Ruiz-Carmona, S., Alvarez-Garcia, D., Foloppe, N., Garmendia-Doval, A. B., Juhos, S., Schmidtke, P., Barril, X., Hubbard, R. E., and Morley, S. D. (2014). rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Computational Biology*, 10(4).

Salemme, F. R. (1976). An hypothetical structure for an intermolecular electron transfer complex of cytochromes c and b5. *Journal of Molecular Biology*, 102(3):563–568.

Šali, A. and Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234(3):779–815.

Salton, M. R. J., Seltmann, G., and Holst, O. (2002). *The Bacterial Cell Wall*. Springer Science & Business Media.

Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., and Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, 30(4):317–320.

Sardana, D., Zhu, C., Zhang, M., Gudivada, R. C., Yang, L., and Jegga, A. G. (2011). Drug repositioning for orphan diseases. *Briefings in Bioinformatics*, 12(4):346–356.

Schalon, C., Surgand, J.-S., Kellenberger, E., and Rognan, D. (2008). A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins*, 71(4):1755–1778.

Schito, G. C. (2006). The importance of the development of antibiotic resistance in Staphylococcus aureus. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, 12 Suppl 1:3–8.

Schmidt, B., Ribnicky, D. M., Poulev, A., Logendra, S., Cefalu, W. T., and Raskin, I. (2008). A natural history of botanical therapeutics. *Metabolism - Clinical and Experimental*, 57:S3–S9.

Schultes, S., Kooistra, A. J., Vischer, H. F., Nijmeijer, S., Haaksma, E. E. J., Leurs, R., de Esch, I. J. P., and de Graaf, C. (2015). Combinatorial Consensus Scoring for Ligand-Based Virtual Fragment Screening: A Comparative Case Study for Serotonin 5-HT3A, Histamine H1, and Histamine H4 Receptors. *Journal of Chemical Information and Modeling*, 55(5):1030–1044.

Scior, T., Bender, A., Tresadern, G., Medina-Franco, J. L., Martínez-Mayorga, K., Langer, T., Cuanalo-Contreras, K., and Agrafiotis, D. K. (2012). Recognizing Pitfalls in Virtual Screening: A Critical Review. *Journal of Chemical Information and Modeling*, 52(4):867–881.

Sedlmayer, F., Woischnig, A.-K., Unterreiner, V., Fuchs, F., Baeschlin, D., Khanna, N., and Fussenegger, M. (2021). 5-Fluorouracil blocks quorum-sensing of biofilm-embedded methicillin-resistant Staphylococcus aureus in mice. *Nucleic Acids Research*, 49(13):e73.

Sertkaya, A., Birkenbach, A., Berlind, A., and Eyraud, J. (2014). Examination of clinical trial costs and barriers for drug development: Report to the Assistant Secretary of Planning and Evaluation (ASPE). *Washington, DC: Department of Health and Human Services*.

Shim, J. S. and Liu, J. O. (2014). Recent Advances in Drug Repositioning for the Discovery of New Anticancer Drugs. *International Journal of Biological Sciences*, 10(7):654–663.

Shiryaev, S. A., Mesci, P., Pinto, A., Fernandes, I., Sheets, N., Shresta, S., Farhy, C., Huang, C.-T., Strongin, A. Y., Muotri, A. R., and Terskikh, A. V. (2017). Repurposing of the anti-malaria drug chloroquine for Zika Virus treatment and prophylaxis. *Scientific Reports*, 7(1):15771.

Shoichet, B. K., Stroud, R. M., Santi, D. V., Kuntz, I. D., and Perry, K. M. (1993). Structure-based discovery of inhibitors of thymidylate synthase. *Science (New York, N.Y.)*, 259(5100):1445–1450.

Si, L., Bai, H., Rodas, M., Cao, W., Oh, C. Y., Jiang, A., Moller, R., Hoagland, D., Oishi, K., Horiuchi, S., Uhl, S., Blanco-Melo, D., Albrecht, R. A., Liu, W.-C., Jordan, T., Nilsson-Payant, B. E., Golynker, I., Frere, J., Logue, J., Haupt, R., McGrath, M., Weston, S., Zhang, T., Plebani, R., Soong, M., Nurani, A., Kim, S. M., Zhu, D. Y., Benam, K. H., Goyal, G., Gilpin, S. E., Prantil-Baun, R., Gygi, S. P., Powers, R. K., Carlson, K. E., Frieman, M., tenOever, B. R., and Ingber, D. E. (2021). A human-airway-on-a-chip for the rapid identification of candidate antiviral therapeutics and prophylactics. *Nature Biomedical Engineering*, pages 1–15.

Sidow, T., Johannsen, L., and Labischinski, H. (1990). Penicillin-induced changes in the cell wall composition of Staphylococcus aureus before the onset of bacteriolysis. *Archives of microbiology*, 154(1):73–81.

Sieradzki, K. and Tomasz, A. (1997). Inhibition of cell wall turnover and autolysis by vancomycin in a highly vancomycin-resistant mutant of Staphylococcus aureus. *Journal of Bacteriology*, 179(8):2557–2566.

Simon, Z., Peragovics, A., Vigh-Smeller, M., Csukly, G., Tombor, L., Yang, Z., Zahoránszky-Kohalmi, G., Végner, L., Jelinek, B., Hári, P., Hetényi, C., Bitter, I., Czobor, P., and Málnási-Csizmadia, A. (2012). Drug effect prediction by polypharmacology-based interaction profiling. *Journal of Chemical Information and Modeling*, 52(1):134–145.

Simpson, M. R., Simpson, N. R., and Masheter, H. C. (1966). New drugs. 8. Flufenamic acid in rheumatoid arthritis. Comparison with aspirin and the results of extended treatment. *Annals of Physical Medicine*, 8(6):208–213.

Singh, N. and Warshel, A. (2010). A comprehensive examination of the contributions to the binding entropy of protein-ligand complexes. *Proteins*, 78(7):1724–1735.

Skinner, D. and Keefer, C. S. (1941). Significance of bacteremia caused by Staphylococcus aureus: A study of one hundred and twenty-two cases and a review of the literature concerned with experimental infection in animals. *Archives of Internal Medicine*, 68(5):851–875.

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W. (2014). Computational Methods in Drug Discovery. *Pharmacological Reviews*, 66(1):334–395.

Smithells, R. W. and Newman, C. G. (1992). Recognition of thalidomide defects. *Journal of Medical Genetics*, 29(10):716–723.

Sousa, S. F., Ribeiro, A. J. M., Coimbra, J. T. S., Neves, R. P. P., Martins, S. A., Moorthy, N. S. H. N., Fernandes, P. A., and Ramos, M. J. (2013). Protein-ligand docking in the new millennium–a retrospective of 10 years in the field. *Current Medicinal Chemistry*, 20(18):2296–2314.

Spink, W. W. and Ferris, V. (1947). PENICILLIN-RESISTANT STAPHYLOCOCCI: MECHANISMS INVOLVED IN THE DEVELOPMENT OF RESISTANCE. *Journal of Clinical Investigation*, 26(3):379–393.

Sriram, K. and Insel, P. A. (2018). G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? *Molecular Pharmacology*, 93(4):251–258.

Stahl, M. and Rarey, M. (2001). Detailed Analysis of Scoring Functions for Virtual Screening. *Journal of Medicinal Chemistry*, 44(7):1035–1042.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E. (2003). The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2):493–500.

Sterling, T. and Irwin, J. J. (2015). ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337.

Stierand, K., Maaß, P. C., and Rarey, M. (2006). Molecular complexes at a glance: Automated generation of two-dimensional complex diagrams. *Bioinformatics*, 22(14):1710–1716.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackerman, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., and Collins, J. J. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell*, 180(4):688–702.e13.

Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2019). Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling*, 59(2):895–913.

Sundaresan, L., Giri, S., Singh, H., and Chatterjee, S. (2021). Repurposing of thalidomide and its derivatives for the treatment of SARS-coV-2 infections: Hints on molecular action. *British Journal of Clinical Pharmacology*, page 10.1111/bcp.14792.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science (New York, N.Y.)*, 240(4857):1285–1293.

Takano, E., Bovenberg, R. A. L., and Breitling, R. (2012). A turning point for natural product discovery – ESF-EMBO research conference: Synthetic biology of antibiotic production. *Molecular Microbiology*, 83(5):884–893.

Talevi, A. (2018). Drug repositioning: Current approaches and their implications in the precision medicine era. *Expert Review of Precision Medicine and Drug Development*, 3(1):49–61.

Tan, S. Y. and Tatsumura, Y. (2015). Alexander Fleming (1881–1955): Discoverer of penicillin. *Singapore Medical Journal*, 56(7):366–367.

Taylor, R. D., Jewsbury, P. J., and Essex, J. W. (2002). A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design*, 16(3):151–166.

Teague, S. J. (2003). Implications of protein flexibility for drug discovery. *Nature Reviews. Drug Discovery*, 2(7):527–541.

Templin, M. F. and Höltje, J.-V. (2013). Chapter 764 - Murein dd-Endopeptidase/PBP-7. In Rawlings, N. D. and Salvesen, G., editors, *Handbook of Proteolytic Enzymes (Third Edition)*, pages 3452–3454. Academic Press.

ten Brink, T. and Exner, T. E. (2009). Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein-Ligand Docking Results. *Journal of Chemical Information and Modeling*, 49(6):1535–1546.

ten Brink, T. and Exner, T. E. (2010). pKa based protonation states and microspecies for protein–ligand docking. *Journal of Computer-Aided Molecular Design*, 24(11):935–942.

Teo, S. K., Resztak, K. E., Scheffler, M. A., Kook, K. A., Zeldis, J. B., Stirling, D. I., and Thomas, S. D. (2002). Thalidomide in the treatment of leprosy. *Microbes and Infection*, 4(11):1193–1202.

Terrett, N. K., Bell, A. S., Brown, D., and Ellis, P. (1996). Sildenafil (VIAGRATM), a potent and selective inhibitor of type 5 cGMP phosphodiesterase with utility for the treatment of male erectile dysfunction. *Bioorganic & Medicinal Chemistry Letters*, 6(15):1819–1824.

Thangamani, S., Younis, W., and Seleem, M. N. (2015). Repurposing ebselen for treatment of multidrug-resistant staphylococcal infections. *Scientific Reports*, 5(1):11596.

The UniProt Consortium (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515.

Thomas, G. L., Spandl, R. J., Glansdorp, F. G., Welch, M., Bender, A., Cockfield, J., Lindsay, J. A., Bryant, C., Brown, D. F. J., Loiseleur, O., Rudyk, H., Ladlow, M., and Spring, D. R. (2008). Anti-MRSA Agent Discovery Using Diversity-Oriented Synthesis. *Angewandte Chemie International Edition*, 47(15):2808–2812.

Tomoda, H. (2016). New Approaches to Drug Discovery for Combating MRSA. *Chemical & Pharmaceutical Bulletin*, 64(2):104–111.

Triballeau, N., Acher, F., Brabet, I., Pin, J.-P., and Bertrand, H.-O. (2005). Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *Journal of Medicinal Chemistry*, 48(7):2534–2547.

Trott, O. and Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of computational chemistry*, 31(2):455–461.

Truchon, J.-F. and Bayly, C. I. (2007). Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *Journal of Chemical Information and Modeling*, 47(2):488–508.

Tse, T., Williams, R. J., and Zarin, D. A. (2009). Reporting "Basic Results" in ClinicalTrials.gov. *CHEST*, 136(1):295–303.

Ubukata, K., Nonoguchi, R., Matsuhashi, M., and Konno, M. (1989). Expression and inducibility in Staphylococcus aureus of the mecA gene, which encodes a methicillin-resistant S. aureus-specific penicillin-binding protein. *Journal of Bacteriology*, 171(5):2882–2885.

US Food and Drug Administration (1983). Orphan Drug Act of 1983.

Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002). Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *Journal of Medicinal Chemistry*, 45(12):2615–2623.

Veljkovic, V., Loiseau, P. M., Figadere, B., Glisic, S., Veljkovic, N., Perovic, V. R., Cavanaugh, D. P., and Branch, D. R. (2015). Virtual screen for repurposing approved and experimental drugs for candidate inhibitors of EBOLA virus infection. *F1000Research*, 4.

Venkatachalam, C. M., Jiang, X., Oldfield, T., and Waldman, M. (2003). LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of Molecular Graphics and Modelling*, 21(4):289–307.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623.

Verdonk, M. L., Mortenson, P. N., Hall, R. J., Hartshorn, M. J., and Murray, C. W. (2008). Protein-Ligand Docking against Non-Native Protein Conformers. *Journal of Chemical Information and Modeling*, 48(11):2214–2225.

Vigers, G. P. A. and Rizzi, J. P. (2004). Multiple Active Site Corrections for Docking and Virtual Screening. *Journal of Medicinal Chemistry*, 47(1):80–89.

Vogel, S. M., Bauer, M. R., and Boeckler, F. M. (2011). DEKOIS: Demanding evaluation kits for objective in silico screening–a versatile tool for benchmarking docking programs and scoring functions. *Journal of Chemical Information and Modeling*, 51(10):2650–2665.

Wagner, G. (1998). Clinical safety of oral sildenafil citrate (VIAGRATM) in the treatment of erectile dysfunction—by Morales et al. *International Journal of Impotence Research*, 10(2):74–74.

Wang, G. and Zhu, W. (2016). Molecular docking for drug discovery and development: A widely used approach but far from perfect. *Future Medicinal Chemistry*, 8(14):1707–1710.

Wang, R. and Wang, S. (2001). How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *Journal of Chemical Information and Computer Sciences*, 41(5):1422–1426.

Waring, M. J. (2009). Defining optimum lipophilicity and molecular weight ranges for drug candidates—Molecular weight dependent lower logD limits based on permeability. *Bioorganic & Medicinal Chemistry Letters*, 19(10):2844–2851.

Waring, M. J., Arrowsmith, J., Leach, A. R., Leeson, P. D., Mandrell, S., Owen, R. M., Pairaudeau, G., Pennie, W. D., Pickett, S. D., Wang, J., Wallace, O., and Weir, A. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, 14(7):475–486.

Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A., and Warren, S. D. (2012). Essential considerations for using protein–ligand structures in drug discovery. *Drug Discovery Today*, 17(23):1270–1281.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., and Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303.

Wei, B. Q., Weaver, L. H., Ferrari, A. M., Matthews, B. W., and Shoichet, B. K. (2004). Testing a Flexible-receptor Docking Algorithm in a Model Binding Site. *Journal of Molecular Biology*, 337(5):1161–1182.

WHO (2014). ANTIMICROBIAL RESISTANCE: Global report on surveillance. Technical report, World Health Organization.

Wilkening, R. R., Ratcliffe, R. W., Wildonger, K. J., Cama, L. D., Dykstra, K. D., DiNinno, F. P., Blizzard, T. A., Hammond, M. L., Heck, J. V., Dorso, K. L., Rose, E. S., Kohler, J., and Hammond, G. G. (1999). Synthesis and activity of 2-(sulfonamido)methyl-carbapenems: Discovery of a novel, anti-MRSA 1,8-naphthosultam pharmacophore. *Bioorganic & Medicinal Chemistry Letters*, 9(5):673–678.

Willett, P. (2013). Combination of Similarity Rankings Using Data Fusion. *Journal of Chemical Information and Modeling*, 53(1):1–10.

Williams, K. (2009). The introduction of 'chemotherapy' using arsphenamine – the first magic bullet. *Journal of the Royal Society of Medicine*, 102(8):343–348.

Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C. T., Guha, R., and Steinbeck, C. (2017). The Chemistry Development Kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics*, 9(1):33.

Wong, W. R., Oliver, A. G., and Linington, R. G. (2012). Development of Antibiotic Activity Profile Screening for the Classification and Discovery of Natural Product Antibiotics. *Chemistry & Biology*, 19(11):1483–1495.

Woodruff, H. B. (2014). Selman A. Waksman, Winner of the 1952 Nobel Prize for Physiology or Medicine. *Applied and Environmental Microbiology*, 80(1):2–8.

World Health Organization (2017). Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis. Technical Report 9789240026438 (electronic version), World Health Organization, Geneva.

Wouters, O. J., McKee, M., and Luyten, J. (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844–853.

Wright, G. D. (2014). Something old, something new: Revisiting natural products in antibiotic drug discovery. *Canadian Journal of Microbiology*, 60(3):147–154.

Xiang, Z. (2006). Advances in Homology Protein Structure Modeling. *Current protein & peptide science*, 7(3):217.

Xie, L., Xie, L., and Bourne, P. E. (2009). A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, 25(12):i305.

Xie, L., Xie, L., Kinnings, S. L., and Bourne, P. E. (2012). Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annual Review of Pharmacology and Toxicology*, 52:361–379.

Xu, M., Lee, E. M., Wen, Z., Cheng, Y., Huang, W.-K., Qian, X., Tcw, J., Kouznetsova, J., Ogden, S. C., Hammack, C., Jacob, F., Nguyen, H. N., Itkin, M., Hanna, C., Shinn, P., Allen, C., Michael, S. G., Simeonov, A., Huang, W., Christian, K. M., Goate, A., Brennand, K. J., Huang, R., Xia, M., Ming, G.-l., Zheng, W., Song, H., and Tang, H. (2016). Identification of small-molecule inhibitors of Zika virus infection and induced neural cell death via a drug repurposing screen. *Nature Medicine*, 22(10):1101–1107.

Yang, J.-M. and Chen, C.-C. (2004). GEMDOCK: A generic evolutionary method for molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 55(2):288–304.

Yeo, W.-S., Arya, R., Kim, K. K., Jeong, H., Cho, K. H., and Bae, T. (2018). The FDA-approved anti-cancer drugs, streptozotocin and floxuridine, reduce the virulence of Staphylococcus aureus. *Scientific Reports*, 8.

Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L., and Vidal, M. (2007). Drug-target network. *Nature Biotechnology*, 25(10):1119–1126.

Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., and Ide, N. C. (2011). The ClinicalTrials.gov Results Database — Update and Key Issues. *New England Journal of Medicine*, 364(9):852–860.

Zhang, N. and Zhao, H. (2016). Enriching screening libraries with bioactive fragment space. *Bioorganic & Medicinal Chemistry Letters*, 26(15):3594–3597.

Zhang, R., Ou, H.-Y., and Zhang, C.-T. (2004). DEG: A database of essential genes. *Nucleic Acids Research*, 32(Database issue):D271–D272.

Zhao, W., Hevener, K. E., White, S. W., Lee, R. E., and Boyett, J. M. (2009). A statistical framework to evaluate virtual screening. *BMC Bioinformatics*, 10(1):225.

Zilian, D. and Sotriffer, C. A. (2013). SFCscoreRF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *Journal of Chemical Information and Modeling*, 53(8):1923–1933.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577.