Cranfield University

GlaxoSmithKline

JESSICA BROTHWOOD

DRUGGABLE AND BIOPHARMABLE GENOME ANNOTATION PIPELINE DEVELOPMENT

Cranfield Health

Applied Bioinformatics

MSc Thesis

Academic year: 2011-12

Supervisors:

Dr Michael Cauchi (Cranfield)

Dr Hannah Tipney (GSK)

September 2012

Cranfield University

GlaxoSmithKline Research & Development Ltd

Cranfield Health

Applied Bioinformatics

MSc Thesis

Academic Year 2011-12

Jessica Brothwood

Druggable and biopharmable genome annotation pipeline development

Supervisors:

Dr Michael Cauchi (Cranfield University)

Dr Hannah Tipney (GlaxoSmithKline)

September 2012

This thesis is submitted in partial fulfilment of the requirements for the degree of Master of Science.

Abstract

The identification of proteins which could be potential targets for new pharmaceutical products is invaluable for the continued improvement people's quality of life and expansion of available treatment options. In order to aid the discovery of new drug targets, predictions of every human gene likely to be exploitable by compounds and biotechnology were generated using open source tools and publicly available data. An automated pipeline was produced in order to minimise the effort required to reproduce, update and expand this work.

In total, using various different prediction techniques, over 15,000 genes were predicted to code potential targets. An optimistic estimate of the druggable genome at 5,097 genes was produced. These genes contain one or more of the same Pfam protein domains as a drug target (a protein displaying significant activity with a phase four drug from ChEMBL database). The preliminary techniques explored here estimate the biopharmable genome to encompass between 3,169 and 8,117 genes. However, as they failed to identify many of the known approved targets, it is likely this is not an accurate representation.

An easy to run, updated and expandable prediction pipeline, which annotates genes with a predicted druggable target class as well as a ChEMBL target class, if available, has been produced in Perl and implemented within GlaxoSmithKline.

Acknowledgements

This dissertation would not have been possible without advice, guidance and support of Dr Hannah Tipney, to whom I am extremely grateful. I would also like to thank Peter Woollard for his recommendations and help with Perl and Dr Michael Cauchi for his advice and support and for his assistance in producing the ROC curves.

I am obliged to Kieran Todd for his invaluable work on the biopharmable genome. I would also like to thank Dr David Michalovich, for information on drug target classes, and Stefan Senger, for his explanations of compound activities, as well as everyone else in the Computational Biology department at GlaxoSmithKline for their warm welcome and help.

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under *grant agreement* n° 115191, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution (www.imi.europa.eu).

Table of Contents

List of Figures

List of Tables

Abbreviations

AUC: Area under Curve

cGMP: Cyclic Guanosine Monophosphate

CTLA-4: Cytotoxic T-Lymphocyte Antigen 4

CYP: Cytochrome P450

dbGAP: database of Genotypes and Phenotypes

DO: Disease Ontology

ECM: Extracellular Matrix

EGFR: Epidermal Growth Factor Receptor

EMA: European Medicines Agency

FDA: U S Food and Drug Administration

GO: Gene Ontology

GPCR: G-Protein Coupled Receptor

GSK: GlaxoSmithKline

GWAS: Genome Wide association Study

IMI: Innovative Medicines Initiative

IND: Investigational New Drug

NBE: New Biological Entity

NME: New Molecular Entity

NTD: Novel Target Drug

OMIM: Online Mendelian Inheritance in Man

PDB: Protein Data Bank

PPP: Public-Private Partnerships

QED: Quantitative Estimate of Drug-likeness

R&D: Research and Development

ROC: Receiver Operating Characteristic

SGC: Structural Genomics Consortium

1 Introduction

1.1 Drug Discovery and Development

The discovery and development of new pharmaceutical products is essential to expand available treatment options and continue to improve people's quality of life. Areas of particular importance include those diseases with limited or absent treatment options and those where existing medication only works for selected patients. However, getting a new drug to market is not a simple task. *De novo* drug discovery, shown in Figure 1.1, takes approximately 10-17 years from the discovery of a new medicine to when it is available to treat patients. The clinical trials alone take between 5-6 years and involve thousands of volunteers. Ultimately, for every 5,000-10,000 compounds which enter the research and development (R&D) pipeline, only one is approved (PhRMA, 2007, Ashburn and Thor, 2004).

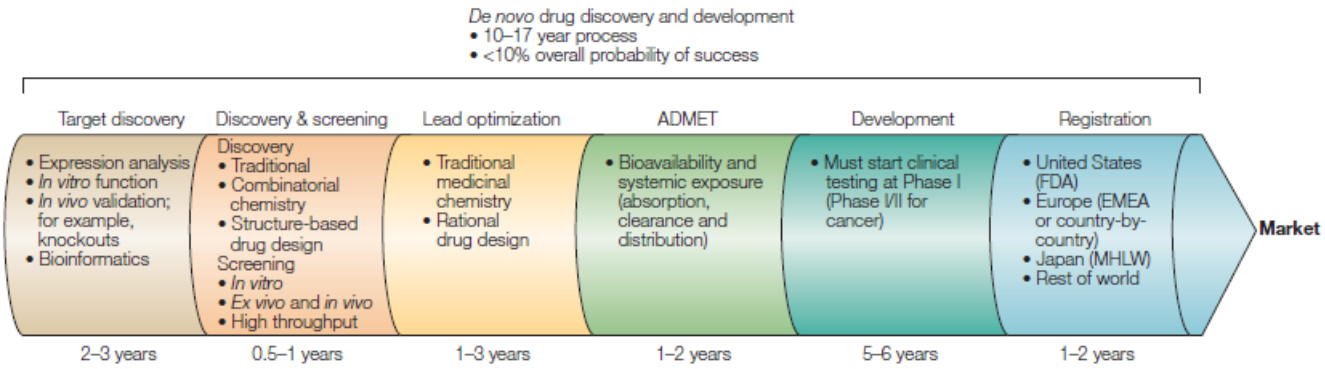

Figure 1.1: **The R&D process.**
The phases of drug development and estimates of the amount of time each stage of development takes. The probability of success is lower than 10% (from Ashburn and Thor, 2004, Figure 2) .

1.1.1 Target Identification and Validation

The pre-discovery process revolves around gaining an understanding of a chosen disease and attempting to uncover the underlying causes of the condition (PhRMA, 2007). At this stage, one of three methods is usually adopted. A mechanism based approach targeting compounds with a specific mode of action, a function based approach which attempts to normalise a disease specific abnormality or a physiology based approach, aiming to reduce disease specific symptoms (Sams-Dodd, 2006).

The mechanism based approach starts with the identification of a molecular target to attempt to modulate with a drug. In most cases this target will be a protein, since the majority of successful drug activity is achieved though binding to and modifying the

activity of proteins (Hopkins and Groom, 2002). It is critical to ensure that both the target selected is involved in a disease and that it is compound binding or accessible to biotechnology (i.e. druggable or biopharmable). In other words, the chosen target should have the potential to interact with a drug and this interaction should potentially affect a disease state (PhRMA, 2007). A drug target can be upstream of a disease causing process, for example a receptor linked to a cascade linked to a disease state. However, the further upstream the target the more chance of affecting other processes and causing unwanted side effects.

Advantageously, the target-centric approach to drug discovery can utilise high-throughput small molecular screening strategies (Swinney and Anthony, 2011) where the chosen target is assayed against large compound libraries, often containing hundreds of thousands of drug candidates, and any binding activity is detected. These compounds of interest can be identified in a number of ways. Traditionally the majority of substances were identified from nature, for example the antibiotic penicillin from *Penicillium* fungi (PhRMA, 2007), and between 1999 and 2008 18 new drugs resulted from modifications to natural substances (Swinney and Anthony, 2011). A *de novo* approach uses computer modelling to create molecules from scratch or modify known compounds of interest. Biological therapeutics are produced by genetically engineered organisms (PhRMA, 2007).

Phenotypic assays are employed in the function and physiology based approaches to drug discovery. These approaches do not require specific understanding of the molecular mode of action; instead, candidate drugs are selected based on functional activity observed in a cell or animal based assay. Since drug selection is based on observed activity an effect in humans can often be established more easily, however it is a challenge to optimise the properties of a drug without prior knowledge of the biological mechanism (Armstrong, 1999, Swinney and Anthony, 2011).

Both target based and phenotypic approaches have specific benefits. Target based assays can identify compounds of interest more quickly than phenotypic screening, which is considerably lower throughput. The identification of potential targets based on supporting molecular knowledge appears to be the fastest and most efficient approach. Between 1999 and 2008, 100 new drugs were discovered by target based methods, compared to 58 by phenotypic approaches. However, phenotypic screening shows more success when identifying previously unknown molecular targets and producing first-in-class drugs. In the same time period, 28 first-in-case drugs came from the phenotypic approach versus 17 drugs from the target based method (Swinney and Anthony, 2011). The use of a dual approach to drug discovery in industry is therefore justified in order to exploit the advantages of both methods, however this thesis will focus down to only the target-centric approach.

1.1.2 Lead Compound Selection and Optimisation

Lead compounds are those which were shown to bind the identified target or have an effect in phenotypic assays. Early safety tests are then preformed on each lead in an attempt to ensure the drug will be successfully absorbed into the bloodstream, distributed to the site of action, metabolised effectively, excreted from the body and is nontoxic. Any leads which demonstrate these properties are then optimised, with hundreds of different variants of the compound produced and retested, and the most promising one is selected as the drug candidate (PhRMA, 2007).

1.1.3 Preclinical Testing

The European Medicines Agency (EMA) and US Food and Drug Administration (FDA) require each candidate drug to undergo extremely thorough testing before it can be deemed safe enough to be tested on humans. Tests are carried out *in vitro*, in the lab, and *in vivo*, using cell cultures and model organisms (PhRMA, 2007) to ensure the identified target is non-essential and non-toxic.

Typically two species are used in animal studies – one rodent and one non-rodent (Bode et al., 2010). The primary model for genetic studies has traditionally been the mouse, which is ideal for studying single mutations but limited as a model for complex human disease. Human disease is generally polygenic, but genetic manipulations of mice generally test the effect of only a single major gene. Additionally, diseases which occur spontaneously in humans must be induced in mice. For the non-rodent model, dogs (Karlsson and Lindblad-Toh, 2008) and the minipig are considered to be favourable models due to their relative genetic and physiological similarities to man and applicability to many experimental studies. For example, the minipig is considered an advantageous model for general toxicology studies, particularly in regard to the cardiovascular system, due to its comparatively similar biology to humans (Bode et al., 2010).

1.1.4 Clinical Trials

If a candidate drug still appears promising after preclinical trials, its safety and efficacy must then be tested in humans. Phase I clinical trials test the drug in a small group of healthy volunteers to determine whether the compound is safe in humans. Phase II trials test the drug on a small group of patients with the disease or condition under study to examine any side effects or risks associated with the drug. If the drug continues to show promise, Phase III trials will be launched, testing the drug against placebos on a large group of patients to show its safety and efficacy (PhRMA, 2007).

After completion, if findings demonstrate the medicine is both safe and effective, the EMA or FDA will review the study and, upon approval, allow the drug to be manufactured for the treatment of patients (PhRMA, 2007). Selecting the right target with potentially druggable features should reduce attrition in these trials due to aspects such as drug target potency and selectivity.

The focus of this thesis is on the very beginning of this pipeline: the identification of molecular targets for new drugs.

1.2 Challenges Facing the Pharmaceutical Industry

With the average drug taking 13 years to develop, at an ever increasing cost, a main focus for the pharmaceutical industry is reducing the time and cost of getting successful drugs to market (Buchan et al., 2011).

Despite the pharmaceutical industry as a whole spending an estimated US$67.4 billion on R&D in 2010, compared to US$47.6 billion in 2004 (PhRMA, 2011), the rate of innovation has remained alarmingly stagnant (Figure 1.2). The estimated capitalized cost per approved new small molecule $1318 million, with an approved biological product costing nearly the same at $1241 million (DiMasi and Grabowski, 2007). A constant of around 30 new molecular entities (NMEs) are approved each year (Lindsay, 2003), with around 18 targeting human proteins, and, of these, only around four of these are novel target drugs (NTDs) which act on previously unexploited targets encoded by the human genome (Rask-Andersen et al., 2011).

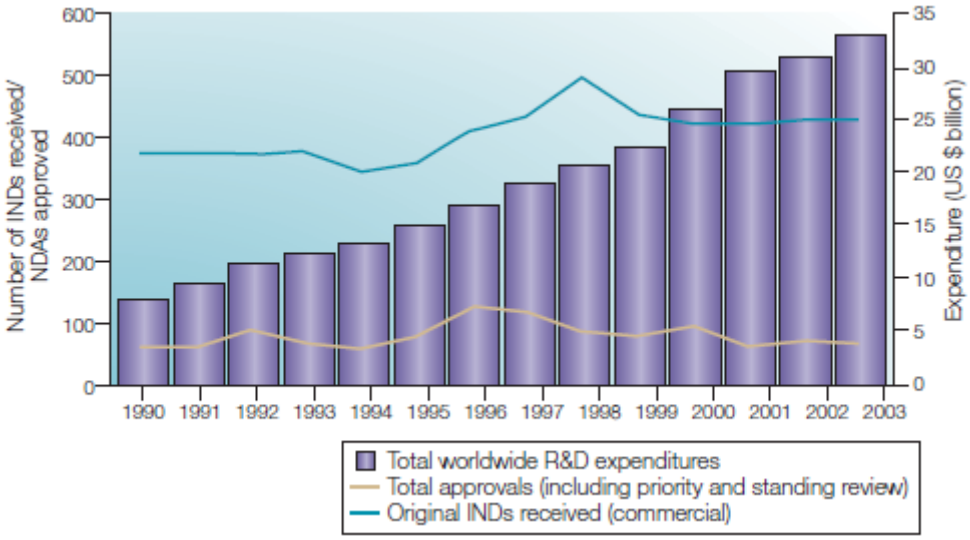

Figure 1.2: **Rise in average drug development costs from 1990 to 2003.**
Despite enormous increases in spending in novel technologies, R&D productivity has decreased since the mid-1990s. This is measured by the number of approvals or original Investigational New Drug (IND) applications received by the FDA per dollar spent (from Ashburn and Thor, 2004, Figure 1).

The rising cost of R&D can be attributed to several factors. Firstly, most of the "low-hanging fruit", or easily discovered drugs, have already been exploited leaving only those drugs which are harder to discover. Additionally, R&D efforts have been increasingly focused towards more complex diseases, for which drug discovery is much harder (Malik, 2009).

Trials have also become longer, involving more volunteers, making them more costly for each compound. In addition, there has been an increase in agents reaching clinical studies, with 2,700 agents investigated in 2008 versus 2000 agents in 2003, but no

notable increase in the number of drugs approved. Therefore clinical trials have higher failure rates than previously, and with 50% of all drugs which reach Phase III development failing, this is extremely costly. For example, in 2006 Pfizer stopped development of a cholesterol drug, *torcetrapib*, after investments of $1 billion because they found it actually increased the risk of cardiac problems (Malik, 2009).

The pharmaceutical industry is constantly pushing for new drugs, not only to provide new treatments for patients but also to replace the high numbers of blockbuster drugs which are losing patent protection. These drugs have sales in excess of US$1 billion annually and generate the majority of the income of pharmaceutical firms, so profits will decrease when their patents expire and they face generic competition. Consequently, new approaches to drug discovery need to be considered to respond to increasing pressure to deliver new products (Malik, 2009).

1.3 Drug Types

There are two kinds of drugs currently marketed: small molecule drugs and biotechnology products. Both are used for the treatment, prevention or cure of disease in humans. Small molecular weight drugs are chemically synthesised, with well defined structures, and far less complex than the large biological products, derived from living material of humans, animals or microorganisms. Monoclonal antibodies and fusion proteins are common types of therapeutic biotechnology (FDA, 2009).

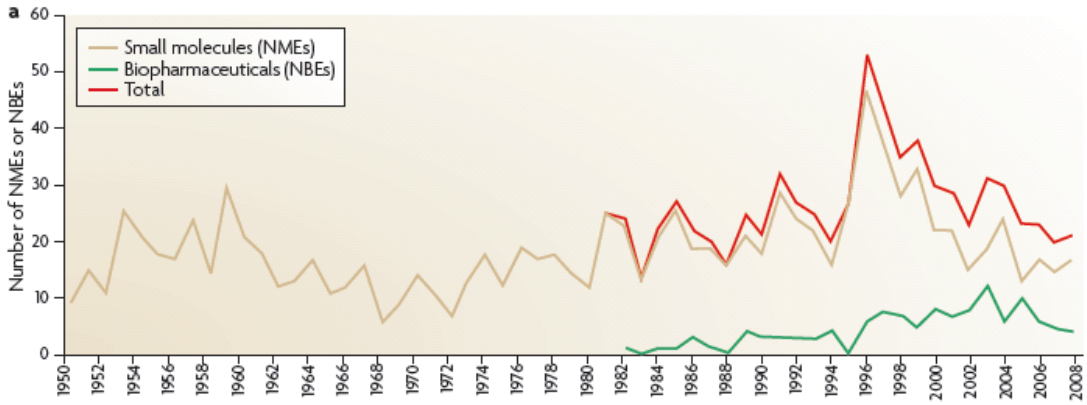

Figure 1.3: **FDA approvals of new chemical entities and new biological entities from 1950-2008.** Timeline displays the approval of 1,103 small molecules and 119 biologics. The surge observed around 1997 can be attributed to a clearing of the backlog of new drug applications, possibly due to accelerated reviewing of applications due to the passing of the Prescription Drug User Fee Act (from Figure 1a, Munos, 2009).

The majority of approved drugs are classed as small molecules, as can be seen in Figure 1.3. Novel target drugs (NTDs) act on previously unexploited targets from the human genome. Although over 60% of NTDs in the last two decades have been small molecule drugs, biologics are strongly emerging therapeutics, with monoclonal antibodies representing 20% of NTDs between 2001 and 2010 (Rask-Andersen et al., 2011).

Common routes of small molecular drugs administration include oral, sublingual (absorption through the blood vessels under the tongue), rectal, topical and parenteral (intravenous, intramuscular, subcutaneous). Although oral administration is the most common and convenient route, other routes offer benefits in specific circumstances, for example where a patient is unable to swallow, to reduce systemic side effects whilst managing localised disease, if an immediate or delayed onset of action is required or to allow the use of drugs which are poorly absorbed, inactive or ineffective if given orally (NursingTimes, 2007). The majority of biologics need to be delivered by injection, causing complications with compliance and effectiveness, although orally bioavailable formulations are in development (PRNewswire, 2011).

1.3.1 Estimates of Drug-likeness

Orally administered small molecule drugs, not derived from natural products, usually comply with the 'rule of five'. Lipinski *et al.*'s rules state that poor adsorption or permeation of a compound is more likely when there are more than 5 hydrogen bond donors and 10 hydrogen bond acceptors, the molecular mass is higher than 500 Da and the lipophilicity is high (with a Log P value greater than 5). Excessive numbers of hydrogen bond donor and acceptor groups will hinder permeability across the membrane bi-layer, as will excessive compound size. Excessive lipophilicity will result in absorption problems due to poor aqueous solubility, preventing the flux of drug across the intestinal membrane into the blood (Lipinski et al., 2001). These rules only apply to the passive diffusion of compounds through cell membranes and are not applicable to compounds which are actively transported by transporter proteins. Although originally only intended to predict the absorption of compounds, conformance with the rule of five can also be used to predict the overall drug-likeness of a compound (Leeson, 2012).

Although the rule of five is predictive of bioavailability, it has limitations. Up to 16% of oral drugs violate at least one of the criteria and several high profile drugs fail more than one (Bickerton et al., 2012), such as *montelukast*, an oral treatment for chronic asthma (Young, 2001). By filtering compounds using the rule of five, undesirable compounds could be considered drug-like by only just meeting the four criteria whereas better compounds could fail by missing just one of the cut offs (Bickerton et al., 2012).

A recently proposed method (Bickerton et al., 2012) has been developed to quantify drug-likeness. Called the quantitative estimate of drug-likeness (QED) it produces a value between 0 (unfavourable) and 1 (favourable) based on the desirability of the compounds properties, rather than the basic pass or fail provided by the rule of five. It takes into account the number of aromatic rings and rotatable bonds in a molecule, the polar surface area (measuring hydrophilicity) and the number of groups in the molecule

known to cause toxicity in addition to eight physical properties proposed to be important for oral drugs (Bickerton et al., 2012, Leeson, 2012).

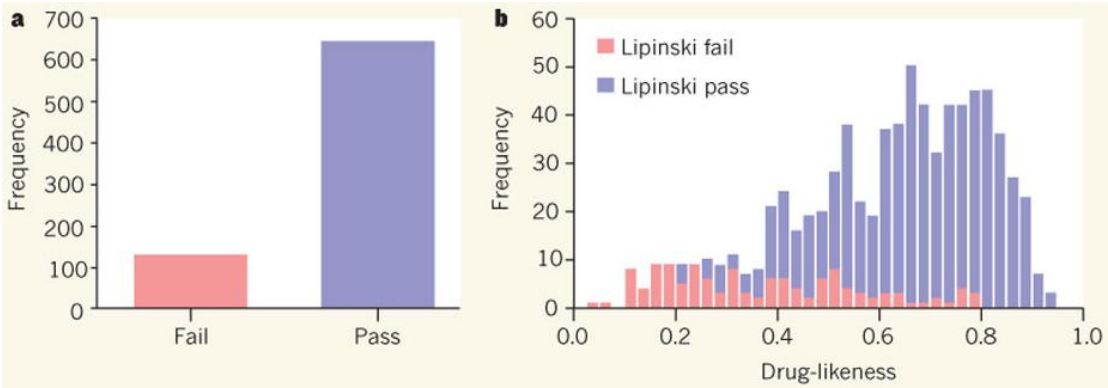

Figure 1.4: **Assessment of drug-likeness of FDA approved oral drugs using the rule of five and QED methods.**

a) The number of FDA approved oral drugs that fail or pass the rule of five, based on a set of 771 drugs. b) Chart showing the distribution of drug-likeness for these same drugs calculated using the QED method. Some very drug-like molecules, with a score over 0.6, fail the rule of five and some very un-drug-like molecules, with scores below 0.4, pass it (from Figure 1, Leeson, 2012).

A comparison of the two methods can be seen in Figure 1.4, generally showing consensus at the extremes of the scale, with most drugs with very high or very low QED scores passing and failing the rule of five respectively. However some notable differences are seen in the middle of the scale, showing that some drugs which failed the rule of five are considered to be very drug-like using the QED method.

The QED method is highly customisable, allowing users to set relative weightings for properties as desired. Importantly, it allows a threshold of drug-likeness to be set, rather than just a pass or fail, and can be applied to sets other than oral drugs, such as those administered intravenously, therefore providing possibly the best estimate of how drug-like a compound is to date (Leeson, 2012).

1.3.2 Small Molecular Drug Targets

Agonist small molecular drugs generally mimic the endogenous ligands of a protein, binding in the same pocket, whereas antagonist drugs block the action of agonists. Phosphodiesterase type 5 inhibitors, such as *sildenafil* (trade name *Viagra*) in Figure 1.5, are examples of chemical compounds with very similar structures to the endogenous molecule with which it competes, cyclic guanosine monophosphate (cGMP). Sildenafil acts as substrate analog, inhibiting the catalytic site of phosphodiesterase and therefore preventing the degradation of cGMP. Elevated cGMP causes smooth muscle tissues to relax, resulting in vasodilation and increased blood flow (Corbin and Francis, 2011).

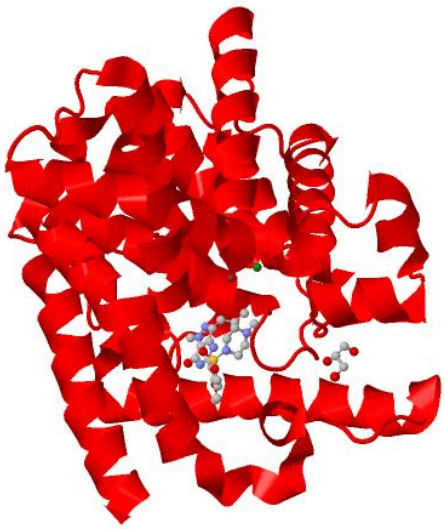

Figure 1.5: **The catalytic domain of human phosphodiesterase 5A complexed with sildenafil.** Enzyme shown in red, ligands coloured by element include sildenafil, glycerol and magnesium and zinc ions. Grey, carbon; blue, nitrogen; red, oxygen; yellow, sulphur; green, magnesium; and dark red, zinc (PDB ID: 1TBF, Zhang et al., 2004).

A ‘one drug, one target’ assumption must be made during drug development in order to optimise binding to a disease relevant protein, although this is rarely the case and a single drug will often have multiple molecular targets. Even so, a high level of specificity for the intended target is generally required, both for efficacy and in order to reduce the chance of any side effects (Johnson, 2009).

For example *imatinib*, used to treat chronic myeloid leukaemia and gastro-intestinal stromal tumours, targets inactive conformation of ABL tyrosine kinase and shows high specificity for its target, more so than drugs which target the active state, such as *dasatinib* (Johnson, 2009). Even so, imatinib shows specificity for tyrosine kinases other than ABL. The off target effects on KIT and PDGFR allow imatinib to be employed as a treatment for gastrointestinal stromal tumours caused by mutations of KIT or PDGFRα (Lee and Wang, 2009).

In another example, some drugs, such as antipsychotic drugs (for example *chlorpromazine*), actually benefit from their promiscuity, becoming more effective through modulation of a spectrum of receptors (Rask-Andersen et al., 2011, Bianchi, 2010). Targets such as dopamine and serotonin receptors produce the therapeutic effect on mood (Kusumi et al., 2000) but chlorpromazine's multitude of other targets result side effects such as mild antihistaminic activity, lactation and reduced gonadotropin levels, dry mouth and muscular disorders (Walsh and Schwartz-Bloom, 2004).

Similarly, the identification of a single druggable target can result in the production of many different drugs, since it is possible for multiple compounds to target the same single protein binding site but produce different therapeutic effects. For example, *aspirin* irreversibly inhibits cyclooxygenase enzymes, whereas *ibuprofen* and *naproxen* bind reversibly. All three molecules bind at the same substrate site, but aspirin's irreversible molecular mode of action translates into a long-lasting effect in platelets, which are unable to resynthesize new enzymes, causing additional functionality as an antiplatelet drug. Many different biochemical features can contribute to the specific functional response of a drug, including residence time; irreversible binding; transient binding; uncompetitive inhibition, in which binding only occurs to the enzyme-substrate complex; and non-competitive inhibition, where binding is equally successful whether the substrate is bound or not (Swinney and Anthony, 2011).

1.3.3 Biological Therapeutics

Biological therapeutics, also known as biologics, consist of a variety of different engineered proteins with medicinal applications. Monoclonal antibodies are useful in the treatment of disorders which cause the target to be expressed at higher levels, such as cancer and inflammatory diseases. One such example is the over expression of receptor tyrosine kinases in tumours (Berg et al., 2007).

Elevated levels of epidermal growth factor receptor (EGFR), a member of the erbB family of receptor tyrosine kinases, are observed in some human epithelial cancers. EGFR consists of an extracellular domain which binds ligands, a transmembrane domain and an intracellular tyrosine kinase domain. The presence of these receptors increases the likelihood that the cell will inappropriately grow and divide, since upon activation EGFR initiates signal-transduction cascades involved in cell proliferation and survival. Blocking activation of these receptors is particularly effective in preventing tumour growth since the receptor is capable of dimerization even in the absence of EGF (Kirkpatrick et al., 2004).

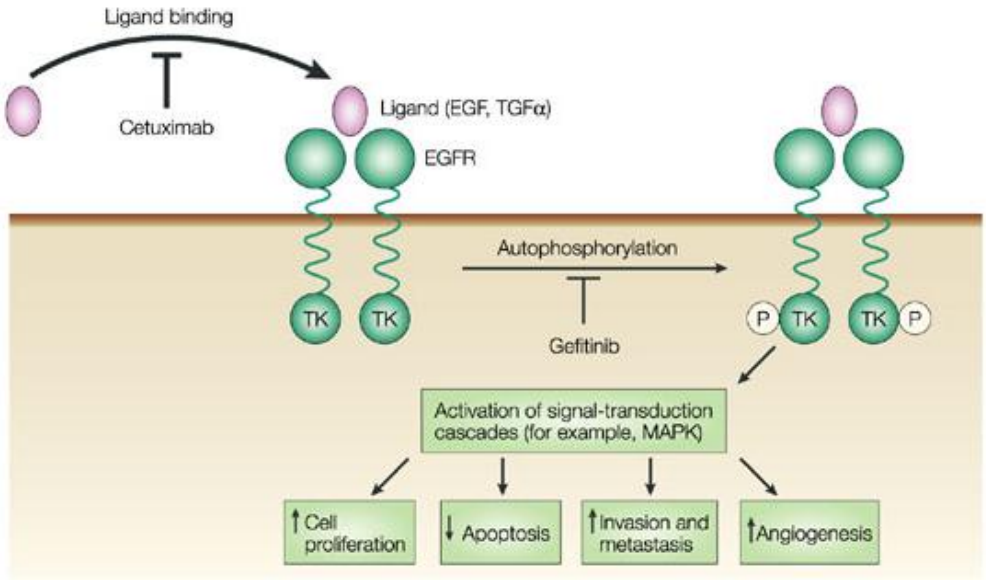

Figure 1.6: **The EGFR pathway and cetuximab's mode of action.**
This simplified illustration of the EGFR signal transduction pathway shows how cetuximab blocks the receptor from binding a ligand and prevents dimerization. *Gefitinib*, an agent that inhibit the tyrosine kinase activity of EGFR, is also shown. MAPK, mitogen-activated protein kinase; TGF-α, transforming growth factor-α; TK, tyrosine kinase domain (from Figure 1, Kirkpatrick et al., 2004).

Cetuximab is a recombinant, human/mouse chimeric monoclonal antibody used in the treatment of large bowel, head and neck cancers. As seen in Figure 1.6, cetuximab competes with EGF for the extracellular binding site of EGFR, inhibiting the receptor by blocking the change in conformation which exposes the dimerization arm and, therefore, prevents the EGFR-controlled pathway from being initiated (Berg et al., 2007, Kirkpatrick et al., 2004). Therapeutic antibodies are generally composed of mouse antibodies with the substitution of human Fc regions in order to reduce the host's anti-antibody immune response (Hwang and Foote, 2005).

1.4 The Druggable and Biopharmable Genome

1.4.1 Genomic Sequencing

In order to predict the druggable and biopharmable genes in an organism, a thorough understanding of its genome is required. The initial sequencing of the human genome (Venter et al., 2001, Consortium, 2001) and the complete (~99%) updates (Human Genome Sequencing, 2004), along with transcriptomics and gene prediction improvements, provide the essential basis for identifying gene and protein sequences with druggable features. In the pre-genomic era only phenotypic assay approaches to drug discovery could be employed, however, faster target centric approaches are now able to exploit these significant advances in our knowledge (Lander, 2004).

As well as providing a comprehensive reference for human genomic studies, the sequencing of the human genome has allowed the biology of model systems to be related more accurately to humans. Other key genomes sequenced included mouse, dog, rat and chimpanzee (Lander, 2011) allowing the best model organisms for toxicology and efficacy testing to be selected based on homology with man. For example, drug metabolism mediated by cytochrome P450 (CYP) enzymes can be modelled by animal drug-metabolising systems. As there are many subfamilies of CYP enzymes, multiple model organisms are needed to ensure the best fit for each family. For CYP1A-mediated pathways most commonly used experimental models were deemed appropriate, however other pathways required close homologues to be selected. The dog was identified as a good model for processes depending on the CYP2D, *Maccacus rhesus* to represent CYP2C and CYP3A seemed to be well modelled by pig or minipig homologue CYP3A29 (Zuber et al., 2002).

1.4.2 Determining Target Druggability

Although a therapeutically relevant target must be disease modifying (Sakharkar and Sakharkar, 2007), in theory the chemical tractability of a protein family can be determined by the presence of protein folds that enable interactions with drug-like compounds. Cellular location should also be considered since 60% of current drug targets are located at the cell surface, compared with only around 22% of all human proteins (Overington et al., 2006). Similarly, to be exploitable by biotechnology the target must be present in an accessible location, either extracellular or pericellular, since antibodies cannot enter cells.

Proteins without these structural features are unlikely to be modulated by pharmaceuticals, so, although a protein may control an interesting pathway, if it does not possess a druggable domain it cannot be easily targeted. Likewise, a protein may have a druggable structure, but modulating its function may not provide any therapeutic benefit. Therefore actual drug targets will be a subset of druggable proteins, as visualised in Figure 1.7, the validation of which will come with successful clinical use (Russ and Lampel, 2005). It is also worth noting that this view of druggability only represents the current state of our abilities: domains which are druggable now, not those that additionally might be in the future, and therefore will evolve over time (Lander, 2004).

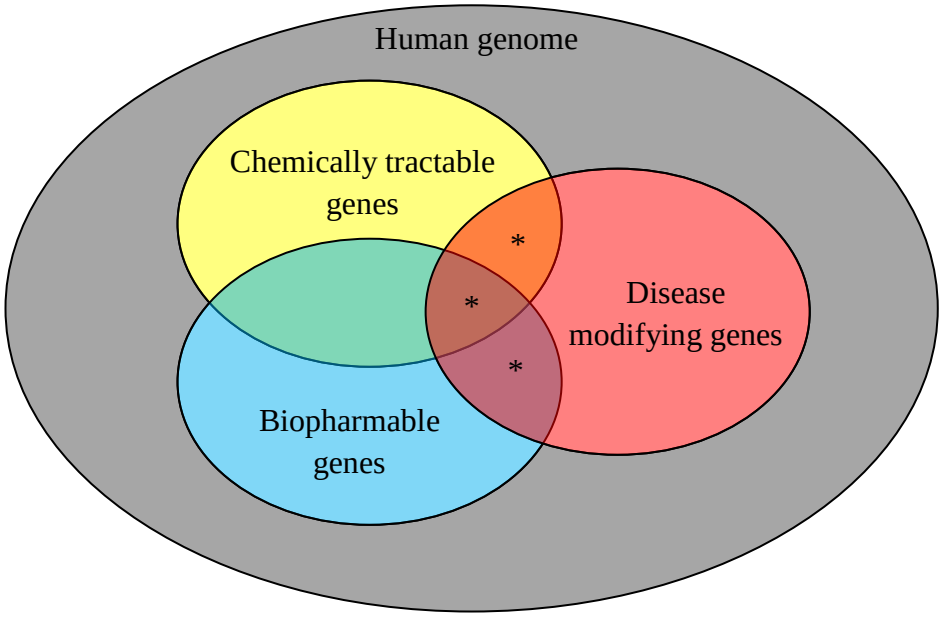

Figure 1.7: **Drug targets are the overlap between chemically tractable or biopharmable genes and disease modifying genes.**
Areas containing drug targets are represented with an asterisk (adapted from Figure 2, Hopkins and Groom, 2002)

One approach to determining druggability is to assess the presence and/or number of ligand binding domains on potential targets, giving an indication of the number of

points where small molecule drugs could potentially act. As protein binding sites usually exist due to functional necessity, most successful drugs gain activity through competing for binding sites with endogenous small molecules, making the potency with which the drug binds to its target critical (Hopkins and Groom, 2002).

Another approach is to identify the protein domain most likely to have made an existing target druggable and assume other proteins with this domain will share these characteristics. If the identified domain binds and interacts with drugs, it stands to reason proteins with these domains could be open to drug interference. Similarly, proteins in the same family of (or which share close homology with) known drug targets are very likely to have shared function and therefore may also be open to modulation by similar small molecules.

Although a few drugs exist which bind to either ribosomes or DNA, or have unknown modes of action, most bind to and inhibit proteins (Overington et al., 2006). The most commonly targeted protein classes are shown in Figure 1.8.

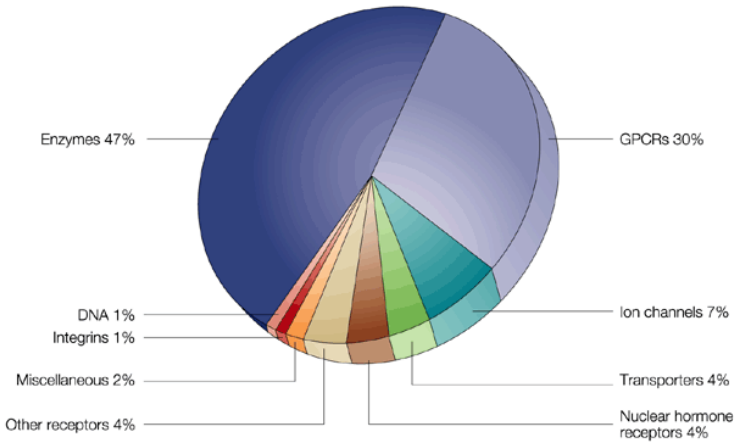

Figure 1.8: **Biochemical classes of marketed small molecule drug targets.**
The majority of drugs marketed target enzymes, G-protein-coupled receptors (GPCRs) or ion channels (from Figure 1, Hopkins and Groom, 2002).

Considering the properties of known drug targets may be beneficial in understanding what makes a protein druggable. GPCRs, a highly targeted family, are present in almost every organ system representing one of the most universal ways in nature to transmit signals into cells. They therefore present a wide range of opportunities as therapeutic targets for many conditions (Filmore, 2004). It should be also be noted that many GPCRs are unlikely to be useful as drug targets since they are not involved in disease processes, for example olfactory GPCRs.

GPCRs are located at the interface between a cell's internal and external environments, making them accessible for modulation. As they have a range of natural ligands, including amines, ions, nucleosides, lipids, peptides and proteins, the composition compound targeting each individual receptor can vary, allowing for greater specificity. A conformational change occurs upon binding a ligand at the active site, signalling the

coupled G protein inside the cell to release components which control various cellular mechanisms (Filmore, 2004).

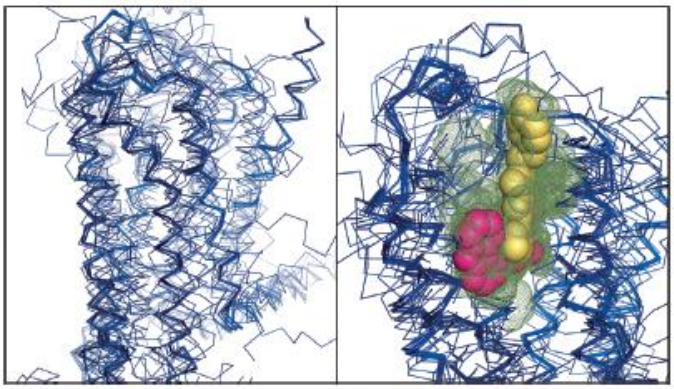

Figure 1.9: **Non-rhodopsin GPCR structure.**
Overlay of the structure of 20 non-rhodopsin GPCRs (left) and merged binding sites (right). Ribbon structures are shown, extracellular side up, with almost perfect helical overlay. Two reference inverse agonists, *carazolol* (2RH1, b2AR) and *ZM241385* (3EML, A2AR), are shown as magenta and yellow spheres in the binding sites (from Figure 1, Kolb and Klebe, 2011).

As seen in Figure 1.9, the structure of the GPCRs appear to be very similar, supporting the assumption that proteins in the same family are similar and, therefore, will have similar functionality. This means that any remaining, untargeted GPCRs could represent exploitable drug targets, although some receptors are unlikely to provide a therapeutic use, for example the olfactory receptors. Additionally, the structure of already identified drug target families could be used as a basis for identifying new druggable targets through homology searches.

1.4.3 Known Drug Targets

Over 21,000 marketed drug products are recognised (defined as the active drug ingredient in association with inactive ingredients in tablet, capsule, cream or liquid form) (Overington et al., 2006). However when duplicate active ingredients and other additional supplements were taken into account by Rask-Andersen *et al.* in 2011, only 1,542 unique drug compounds were identified (Rask-Andersen et al., 2011) from the DrugBank database (Knox et al., 2011). After removing 225 drugs with no known target, 1,236 protein targets were assigned to the remaining 1,317 drugs. Upon the further removal of drugs with no known human target and those with only a non-therapeutic target, for example drug metabolising cytochrome P450 enzymes, only 989 drugs acting on 435 human, therapeutic effect-mediating targets remained (Rask-Andersen et al., 2011). As of 2006, only 166 marketed drug products were biologicals, with just 15 human proteins targeted by monoclonal antibodies and only 9 marketed drug targets modulated by both small molecule and biological drugs (Overington et al., 2006).

With an estimated 22,000 protein coding genes (Pertea and Salzberg, 2010), and, if one includes post-translational modification and complex assembly, an even larger number of different protein isoforms (Hopkins and Groom, 2002) few proteins are currently exploited. Since 584 human proteins have been found to be associated with cardiovascular disease alone (Johnson et al., 2005) it is clear that the number of currently exploited targets is comparatively low, so identification of novel protein targets should aid drug development.

1.4.4 The Druggable Genome

Identifying all druggable genes in the human genome has been attempted before. Estimates of the number of currently known human drug targets range from ~200-400, depending which methods were used and how the targets were defined (Rask-Andersen et al., 2011). In 1996, Drews and Ryser were the first to present an overall estimate of the number of predicted drug targets, identifying 483 targets using drugs listed in the ninth edition of The Pharmacological Basis of Therapeutics and 5,000-10,000 potential targets on the assumption that the human genome contained 300,000 genes (Drews, 1996), shown in Figure 1.10.

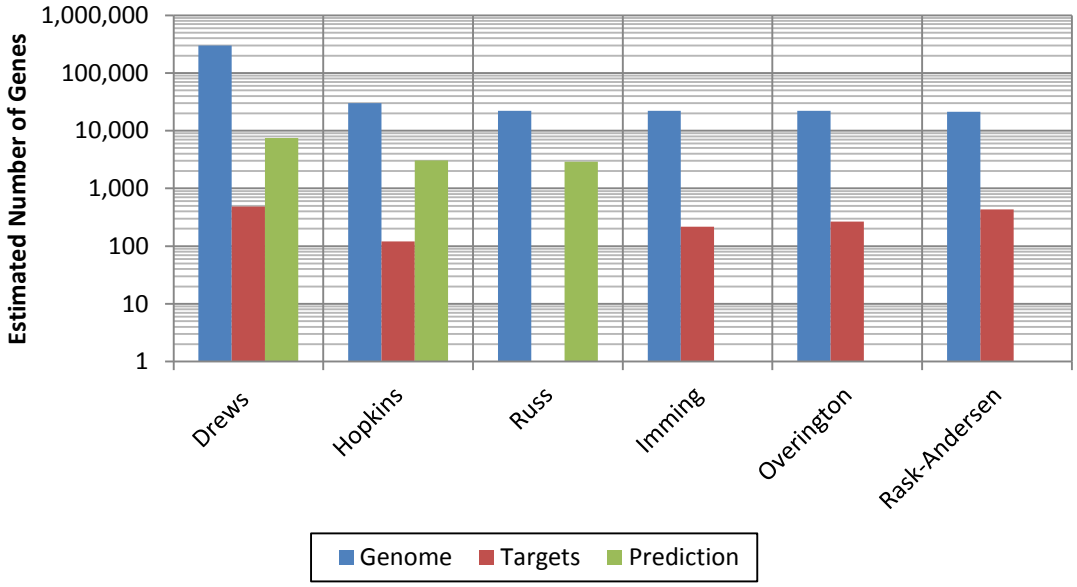

Figure 1.10: **Comparing previous known human drug targets and druggable target estimates in the context of the perceived genome space.**
Data from (Drews, 1996, Hopkins and Groom, 2002, Russ and Lampel, 2005, Imming et al., 2006, Overington et al., 2006, Rask-Andersen et al., 2011). Human genome size was assumed to be the same from the Russ and Imming publications as Overington and Rask-Andersen genome size the same as the current human genome build.

Six years later, in 2002, Hopkins and Groom revised this, identifying 120 targets of currently marketed small molecule drugs and 399 molecular targets capable of binding

rule of five compliant compounds with affinities below 10 µM, regardless of whether these drug-like ligands were actually approved. In contrast to the previous work, they based their estimate of the druggable genome on the predicted human proteome at 21,688 and the assumption that functional similarities are conserved within protein families. Therefore, if one member is able to bind a drug, other members will also likely be able to bind similar compounds. Using this reasoning they predicted 3,051 human genes code for a protein with a precedent for binding a drug-like molecule and, using the yeast genome as a model, the overlap of those related to disease (around 2–5% of the genome) suggests 600-1,500 human genes could be pharmologically exploitable by small-molecule drugs (Hopkins and Groom, 2002).

In 2006, Russ and Lampel released an update to Hopkins and Groom's work, and by using an updated version of the InterPro domains set produced an estimate of 3,533 druggable genes, 482 more than identified in 2002. By removing olfactory GPCRs, the estimate was reduced to 3,050 genes. However, to address the problem of over prediction, 182 equivalent Pfam domains were manually chosen from the 108 InterPro definitions, which resulted in a count of 2,917 genes after removing sensory receptors (Figure 1.10) (Russ and Lampel, 2005).

Although neither published a druggable estimate, in 2006 Imming *et al.* produced a figure of 218 targets of approved drugs (Imming et al., 2006) and Overington *et al.* supplied 266 human proteins as targets, through a systematic review of the US Food and Drug Administration (FDA) Orange Book and the Centre for Biologics Evaluation and Research (CBER) website (Overington et al., 2006). The same year the publicly available DrugBank database was launched (http://www.drugbank.ca/), drawing heavily from these earlier datasets, listing drug-target interactions, accession numbers and pharmacological agents (Wishart et al., 2008).

Rask-Andersen *et al.* analysed the complete data set of pharmacological agents from the DrugBank database as of May 2009, focusing on individual genes, without filtering with regard to gene family redundancies or for rule of five compliant molecules. This resulted in the identification of 435 known human drug targets, 989 unique marketed drugs and 2,242 known drug-target interactions. Through manual curation of the DrugBank database, removing drugs without known protein targets (for example, those which act on DNA, dietary supplements or those with unknown targets), those administered as prodrugs (as these are metabolised to the active form, which is included if relevant) and those acting on non-human targets (such as antibiotics, antiparasitics and antifungals) a list of 1,092 pharmacological agents acting on 1,044 human protein targets was produced. Non-therapeutic targets were then removed through validation of each drug, drug target and interaction using current medical literature and public databases, identifying therapeutically irrelevant targets and side effect mediating targets to produce the final, strict dataset of 435 known effect mediating drug targets (Rask-Andersen et al., 2011).

Each target's encoding gene was also assigned a class according to their molecular function. The largest group, comprising 44% (193) of known human drug targets, is the receptor category, most commonly G protein-coupled receptors (GPCRs), usually the target of anti-hypertensive and anti-allergic drugs. The second largest receptor target class, ligand-gated ion channels, are usually the target of sedatives, and the third, receptor tyrosine kinases, are targeted by anticancer drugs. The next largest group (29%) are enzymes, with the top three largest targets being hydrolases (EC 3), oxidoreductases (EC1) and transferases (EC 2). Targeted enzymes are most commonly soluble proteins (78%) rather than membrane associated. The third largest known target class, with 15% of the human targets, are transporter proteins, which includes voltage gated ion channels (Rask-Andersen et al., 2011).

Protein-protein interaction networks (also called interactome networks) have been a recent focus of network biology and drug-target interactions can also be studied in this context, providing a new avenue for target prediction. A drug-target network, viewed in Cytoscape, of the 989 drugs and 435 protein targets identified by Rask-Andersen *et al.* showed that almost half of the drugs on the market interact with similar targets, producing large networks which exploit only a limited part of the proteome. Therefore the smaller NTD networks are of particular interest as they often represent novel molecular mechanisms (Rask-Andersen et al., 2011).

1.4.5 The Secretome and Biopharmable Genome

Secreted proteins often circulate the body, gaining access to most organs and tissues, and many factors act as therapeutic agents. Secreted proteins play important roles in many different biological functions, such as signalling pathways, blood coagulation, structural scaffolding, enzymatic action and immune defence. Extracellular matrix (ECM) molecules also interact directly with cell surface receptors, therefore improved knowledge of the secretome could unveil novel therapeutics as well as drug targets (Huxley-Jones et al., 2008, Xu, 2007). Secreted proteins have high specificity for receptors, making them attractive candidates for therapeutics. However downstream components of receptor signalling cascades are often shared, so targeting processes with these agents is still expected to produce unwanted side effects. Similarly, although secreted proteins have the benefit of acting at low concentrations, they have limited oral availability and short half-lives, requiring frequent administration (Bonin-Debs et al., 2004).

Classically secreted proteins can be identified based on specific signatures at the amino terminus of the protein, known as signal peptides (Bonin-Debs et al., 2004). Signal peptides function like a postal address, identifying proteins destined for secretion or for specific organelle for further processing. When proteins reach their targeted locations these signal peptides are cleaved off and degraded (Choo et al., 2005). Not all proteins

display a signal peptide sequence in their primary structure, nor are they all released through the classical endoplasmic reticulum-Golgi pathway (Prudovsky et al., 2003) but signal peptides allow predictions of the specific destinations of the proteins that do.

Although the ECM can act as a barrier to effective drug action, its numerous components (Figure 1.11) also represent possible therapeutic targets. Many drugs in development target components of the ECM (Huxley-Jones et al., 2008), for example *firategrast* is an orally administered monoclonal antibody therapy currently in development as a treatment for multiple sclerosis. Acting against α4β-integrin, a glycoprotein of the ECM, Phase II trials have shown firategrast reduces trafficking of mononuclear white blood cells across the blood brain barrier (Miller et al., 2012).

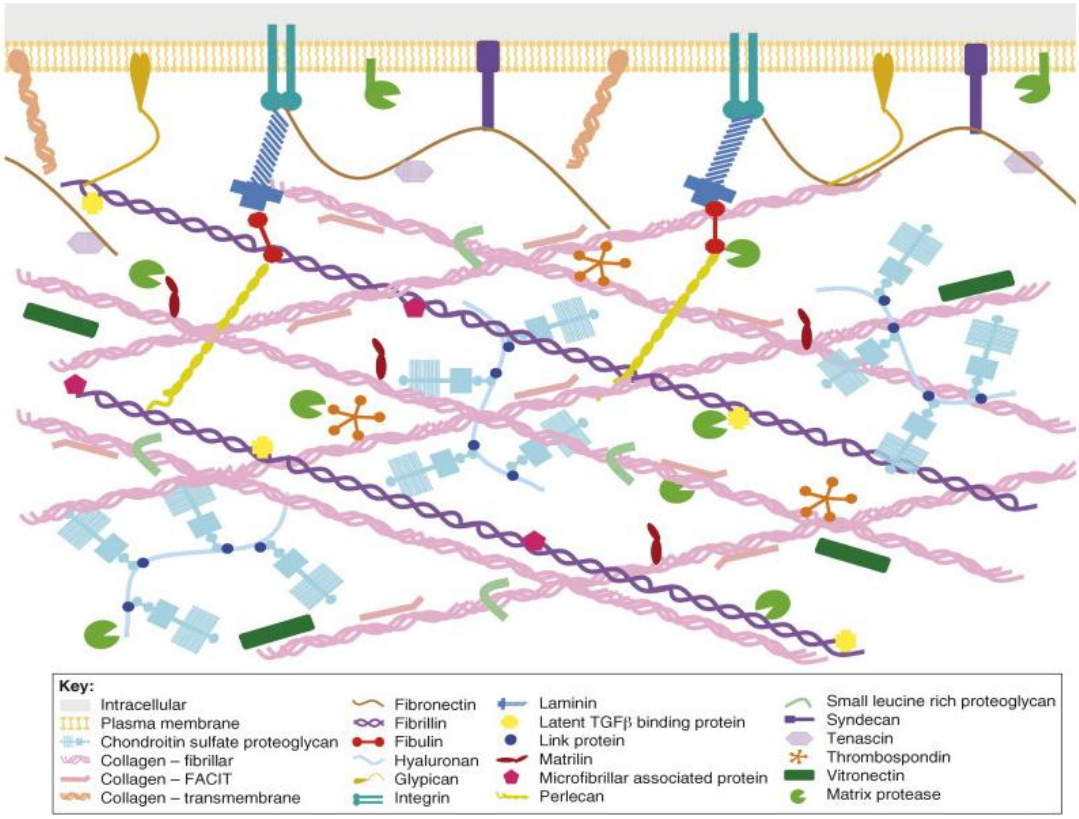

Figure 1.11: **Diagrammatic representation of the extracellular matrix.**
Depiction of all the major classes of extracellular matrix molecules (from Figure 1, Huxley-Jones et al., 2008).

The impact of existing secreted protein therapies has been huge, for example until the discovery of insulin for the treatment of diabetes mellitus in 1921 the diagnosis would have been incurable. It is estimated that 10% of human genes encode secreted proteins; however, deciphering the secretome and function of these proteins presents a major challenge due to the complexity of these protein classes. One approach is to screen the secretome of a model organism, such as zebrafish, *in vivo* and apply this to homologues in man (Xu, 2007).

1.5 Strategy to New Drug Discovery

In addition to new scientific approaches to drug discovery, new business models are increasingly important. One key approach is to increase collaboration between pharmaceutical companies and other key players in healthcare, such as academics and small enterprises. This would enhance competitiveness in pharmaceutical industry for the benefit of patients and scientists. Public-private partnerships (PPP) involving both for-profit companies and non-profit institutions aim to encourage these collaborative efforts, supporting drug discovery and development (Goldman, 2012). An example collaboration, the Medicine for Malaria Venture and Novartis, has successfully launched and distributed a malaria treatment, *Coartem Dispersible*, formulated especially for children (MWV, 2012). The largest PPP in life sciences R&D is the European Union initiative called the Innovative Medicines Initiative (IMI) which hopes to address bottlenecks in the drug development process and develop new tools for predicting drug safety and efficacy (Goldman, 2012).

Precompetitive research such as IMI does not reduce commercial advantage and, therefore, allows collaboration and pooling of resources between pharmaceutical companies. This facilitates the desire of many pharmaceutical companies such as GSK to promote public domain discovery systems involving open-access tools and promotes collaboration with partners outside of industry. IMI provides neutral platform for transparent exchanges between researchers to support the development of new medicines in a non-competitive environment for the benefit of both industry and society (Goldman, 2012). The project described in this thesis will contribute potentially druggable and biopharmable targets to OpenPHACTS, an IMI funded project.

OpenPHACTS (http://www.openphacts.org) is a three year project, ending in March 2014. The consortium contains 23 partners from a wide range of backgrounds, including academia, pharmaceutical (including GSK) and biotechnology companies. The project aims to reduce the barriers between drug discovery in industry and academia by providing an open access innovation platform, called Open Pharmacological Space. This will comprise of data, vocabularies and infrastructure needed to accelerate drug-oriented research, which is open to all users and available in the public domain.

Similarly, the use of open source methodologies can accelerate the discovery process, allowing the reuse of existing resources and preventing wasted time and effort. The accessibility of open source can ultimately lead to more people becoming involved and an inevitably faster rate of innovation. Besides speed, another advantage is the transparency of the process, making open source methods easily understood and, therefore, adapted to evolving needs. Since everything is available on the web, an open soruce project can be picked up in the future by those not originally involved, meaning it will not cease in the event of the graduation of a student, termination of funding or

departure of an investigator. Freely available software is also subject to extensive and continuous peer review through commenting systems (Woelfle et al., 2011).

Research conducted in this manner has the potential for significant impact on human health, with an example of successful open research seen in the improvement of an off patent drug, *praziquantel* (used in the treatment of the parasitic infection schistosomiasis), to produce a purer product at a lower cost (reviewed in Woelfle et al., 2011).

2 Aims and Objectives

This project has a number of aligned aims:

1. Druggable genome prediction
2. Biopharmable genome prediction
3. Pipeline production

Most importantly, it aims to produce estimates of all human genes likely to be druggable or biopharmable. These predictions can then be compared to known, currently approved small molecule drug and biotechnology targets. This will be achieved using:

- open source software
- publically available data

The work of Hopkins and Groom in 2002 will be replicated a decade on to produce the first prediction of the druggable genome. Fundamentally, this should:

- draw comparisons between findings
- produce a list of gene and protein identifiers for reference (not included in the original publication)

Druggable domain definitions will then be updated programmatically in an attempt to:

- minimise the required manual curation effort
- provide easily updatable predictions

To align produced predictions with existing drug target data, target classes will be predicted using regular expressions and predicted targets will be annotated with their ChEMBL target class (where available).

Predictions of the biopharmable genome will be produced, expanding on work carried out by Kieran Todd and Alan Lewis at GlaxoSmithKline to:

- draw comparisons between the druggable and biopharmable genome estimates
- evaluate the success of signal peptide and transmembrane region prediction tools

The pipeline should also be easily updated, therefore it aims to produce:

- an easy to use, fully automated prediction pipeline
- well documented scripts which can be modified and updated in the future

Since the human genome assembly is being constantly revised, first in class drug targets constantly being discovered and our knowledge of disease ever expanding others should be able to easily replicate this work in the future.

3 Materials and Methods

3.1 Programming Languages

All research was carried out using scripts written in Perl version 5.8.5 (perl.org), which can be found in the uploaded ZIP file under the 'scripts' directory. Most of the scripts make use of the core modules in order to be as portable as possible. However, in order to query a local ChEMBL database instance the DBI module (dbi.perl.org) is required and in order to parse the DrugBank XML database file the XML-TreeBuilder module (search.cpan.org/~jfearn/XML-TreeBuilder-4.1) is required.

All graphs were plotted using R version 2.14.0 (r-project.org). The scripts used can be found in the uploaded ZIP file under the 'analysis' directory. The ggplot2 package (cran.r-project.org/web/packages/ggplot2, Wickham, 2009) is required to produce bar plots, whilst ROC curves require the included rocit function (courtesy of Dr Michael Cauchi, Cranfield University), and Venn diagrams require the VennDiagram package (cran.r-project.org/web/packages/VennDiagram, Chen and Boutros, 2011).

3.2 Mutual Resources

The ChEMBL (www.ebi.ac.uk/chembl, Gaulton et al., 2012) and DrugBank (www.drugbank.ca, Knox et al., 2011) databases were used as resources for known targets of approved small molecule and biotechnology drugs.

BioMart's (www.ensembl.org/biomart/martview, Kinsella et al., 2011) web service was used to query Ensembl human genome assembly GRCh37.p6 via XML using various filters. The 'protein coding' filter was used to generate a universe of protein coding genes, removing pseudogenes, which contained 21,405 genes and 96,535 proteins. This collection of proteins was used in all predictions/comparisons/analysis unless otherwise stated and can be found under the 'outputs' directory of the attached ZIP file in the 'universe' folder.

3.3 Chemically Tractable Resources and Filters

Protein domains from InterPro (www.ebi.ac.uk/interpro, Hunter et al., 2012) and Pfam (pfam.sanger.ac.uk, Punta et al., 2012) were used to predict genes which have drug binding properties. InterPro is a consortium of 11 major signature databases, allowing predictions of structure, function or family membership to be made based only upon sequence. Pfam is a member of this consortium, but as it clearly defines domains based on multiple sequence alignments and profile hidden Markov models it should, therefore, provide more constrained estimates.

A synopsis of all of the filters used in BioMart queries to produce the chemically tractable predictions can be seen in Figure 3.1.

A list of druggable InterPro domains was provided in (Hopkins and Groom, 2002), and an up to date version of this list including replacements (Table 3.1) was used as a BioMart filter and is labelled as *(3)* in Figure 3.1.

The Protein Data Bank (PDB) (www.pdb.org, Berman et al., 2000) web service was used to determine which of the associated Pfam domains have an experimentally determined structure with a free ligand via an XML query with Perl managed output.

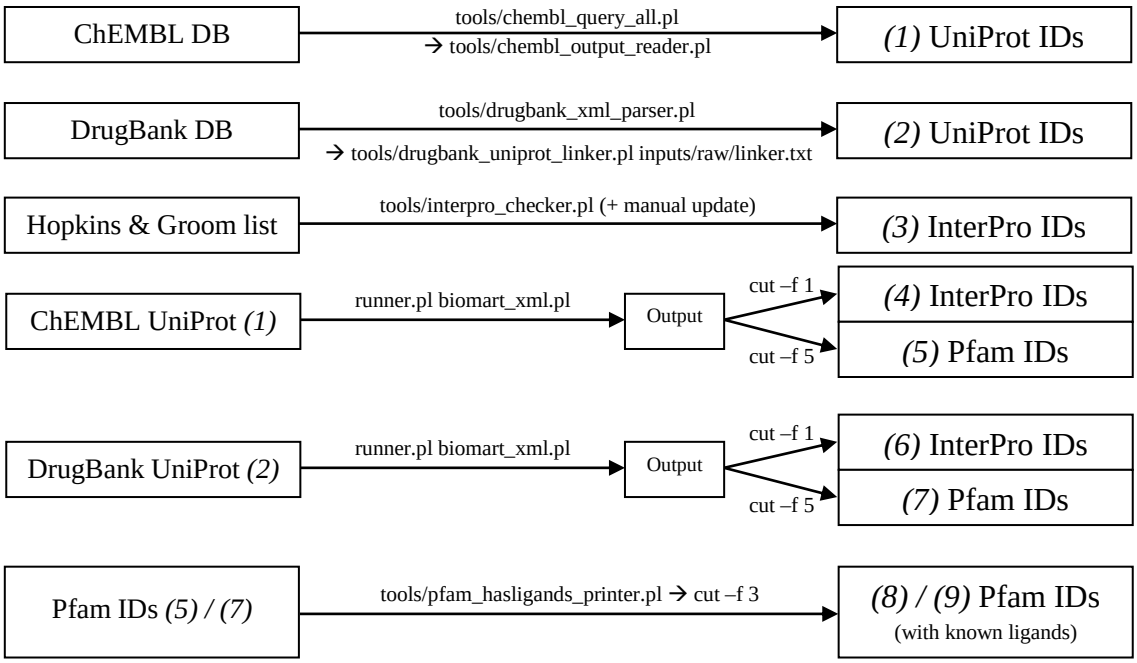

Figure 3.1: **Generation of the nine BioMart query filters for estimating the known (*1* & *2*) and potentially chemically tractable target lists.**

Bracketed numbers indicate an input which is used as a filter in a BioMart query, large arrows indicate the processing of the original input using scripts (included) or unix commands, smaller arrows indicate the output of the previous script is processed. ChEMBL is a local database, DrugBank is an XML file.

Table 3.1: **Modifications to the original Hopkins and Groom list of InterPro domains.**

Original name	Original InterPro Accession	Updated name(s)	Updated InterPro Accession(s)	Reason
Cation channels	IPR000636	Ion transport domain	IPR005821	Replaced
Neutral zinc metallopeptidases zinc-binding region	IPR000130	Peptidase, metallopeptidase	IPR006026	Replaced
Domain in various γ-carboxylases and other proteins	IPR001870	Vitamin K-dependent gamma-carboxylase	IPR007782	Manual update (false positive matches against B30.2/SPRY)
SAM-binding motif	IPR000051	MCP methyltransferase, CheR-type, SAM-binding domain, C-terminal	IPR022642	Manual update
Neurotransmitter-gated ion channel	IPR001175	Neurotransmitter-gated ion-channel	IPR006201	Replaced
		Neurotransmitter-gated ion-channel ligand-binding	IPR006202	Replaced
Thioredoxin	IPR000063	Thioredoxin domain	IPR013766	Replaced
Aldehyde dehydrogenase family	IPR002086	Betaine aldehyde dehydrogenase	IPR011264	Replaced
		Aldehyde dehydrogenase domain	IPR015590	Replaced
H+/K+- and Na+/K+-transporting ATPase	IPR000661	ATPase, P-type cation-transporter, C-terminal	IPR006068	Replaced
		ATPase, P-type cation-transporter, N-terminal	IPR004014	Replaced
		ATPase, P-type cation exchange, alpha subunit	IPR006069	Replaced
Amino-acid permease	IPR002027	Amino acid permease domain	IPR004841	Replaced
		Amino acid permease, conserved site	IPR004840	Replaced
Na+/H+ exchanger	IPR000676	Cation/H+ exchanger	IPR006153	Replaced
		Na+/H+ exchanger	IPR004709	Replaced
Aminotransferase class-III pyridoxal-phosphate	IPR000954	Adenosylmethionine--8-amino-7-oxononanoate aminotransferase BioA	IPR005815	Replaced
		Aminotransferase class-III	IPR005814	Replaced
DNA-directed DNA polymerase family B	IPR002064	DNA-directed DNA polymerase, family B	IPR006172	Replaced
		DNA-directed DNA polymerase, family B, multifunctional domain	IPR006134	Replaced
		DNA-directed DNA polymerase, family B, exonuclease domain	IPR006133	Replaced
Glyceraldehyde 3-phosphate dehydrogenase	IPR000173	Glyceraldehyde 3-phosphate dehydrogenase, active site	IPR020830	Manual update
		Glyceraldehyde 3-phosphate dehydrogenase, NAD(P) binding domain	IPR020828	Manual update
		Glyceraldehyde 3-phosphate dehydrogenase, catalytic domain	IPR020829	Manual update
Aromatic-ring hydroxylase	IPR000733	Aromatic-ring hydroxylase-like	IPR003042	Replaced
Poly(ADP-ribose) polymerase; catalytic region	IPR001290	Poly(ADP-ribose) polymerase, catalytic domain	IPR012317	Replaced
Aspartate and ornithine carbamoyltransferase family	IPR002029	Aspartate/ornithine carbamoyltransferase, Asp/Orn-binding domain	IPR006131	Replaced
		Aspartate/ornithine carbamoyltransferase, carbamoyl-P binding	IPR006132	Replaced
		Aspartate/ornithine carbamoyltransferase	IPR006130	Replaced
Prolyl 4-hydroxylase; α-subunit, C terminus	IPR003865	Prolyl 4-hydroxylase, alpha subunit	IPR006620	Manual update
		Prolyl 4-hydroxylase alpha-subunit, N-terminal	IPR013547	Manual update

Analysis involving the Pfam domains used in Russ and Lampel 2005 could not be completed since a list was not included with the publication and despite best efforts could not be obtained.

The Hopkins and Groom InterPro domains, seen as filter (3) in Figure 3.1 are used as the first method of predicting the druggable genome, which consists of all human genes/proteins which contain one or more of these domains.

Filters (1) and (2) from Figure 3.1 are used to generate each set of known approved drug targets (from ChEMBL and DrugBank). If a gene/protein from the Ensembl human genome is associated with a UniProt identifier on either list, information including its Ensembl identifiers, description and contained InterPro and Pfam domains is returned.

All associated InterPro domains are then extracted from each set of known targets to create the filters (4) and (6) for the ChEMBL InterPro and DrugBank InterPro methods. All genes/proteins in the human genome which contain one or more of these domains are predicted to be druggable.

Similarly, all associated Pfam domains are extracted from each set of known targets, creating the filters (5) and (7) for the ChEMBL Pfam and DrugBank Pfam methods. All human genes/proteins which contain one of more these domains is predicted to be druggable. To produce filters (8) and (9) these two lists of Pfam domains are queried against the PDB, returning only domains associated with a free ligand. Genes/proteins which contain one or more of these domains are predicted to be druggable by the ChEMBL/DrugBank PDB methods.

3.4 Chemically Tractable Genome Outputs

For each druggable genome prediction method, the input files shown in Figure 3.1 and contained within the ‘inputs’ directory of the attached ZIP file, were taken as the BioMart query filter and all associated genes and proteins from Ensembl human genome assembly GRCh37.p6 were returned in a tab delimited format with each new line containing: Interpro ID, Interpro Short Description, UniProt/SwissProt Accession, PDB ID, PFAM ID, EntrezGene ID, Ensembl Gene ID, Ensembl Transcript ID, Ensembl Protein ID, Associated Gene Name, Associated Transcript Name, Status (gene), Status (transcript), Ensembl Family Description and Description.

As shown in Figure 3.2, these outputs can be parsed to allow gene/proteins identified to be compared between methods via Venn diagram. Each prediction method can also be evaluated against the two public sets of known targets (ChEMBL and DrugBank) and the protein coding universe as defined by Ensembl GRCh37.p6 in order to assess the merits and caveats of each method.

In addition, a complete summary file can be produced from the output of all seven prediction methods which includes various gene and protein identifiers for each predicted druggable target, whether this protein is already listed as a known target in ChEMBL or DrugBank, the methods which predicted this protein as well as a predicted protein class using Perl regex captures and target classifications from ChEMBL if available.

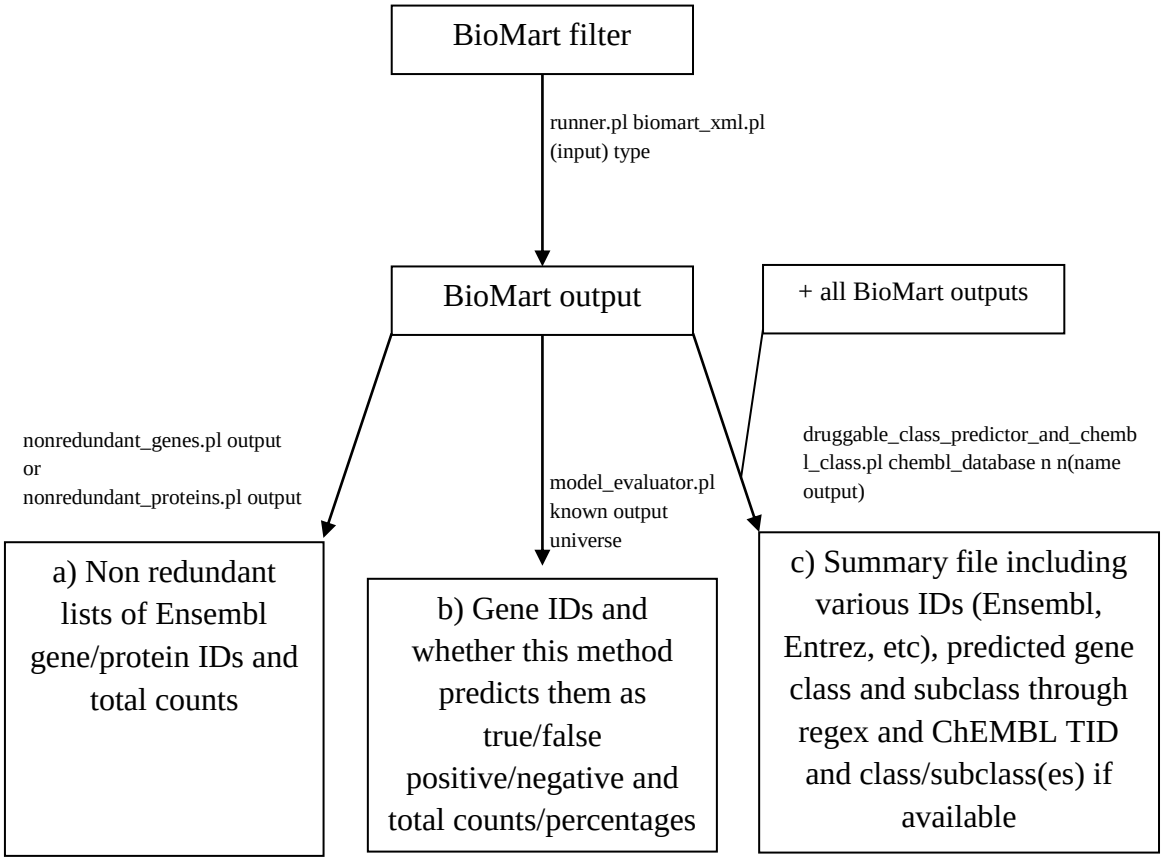

Figure 3.2: **Pipeline producing data for comparisons, method evaluations and a complete summary of all the chemically tractable outputs for each prediction method.**

Arrows indicate the processing of each file with the scripts annotated on the arrow. BioMart filters are the InterPro domains etc seen on the right in Figure 3.1 and these files are the input for each query, unless the type is universe, sigp or tmhmm, in which case no input is required. Available query types are uniprot, interpro, pfam, pdb, go, sigp, tmhmm or universe. This output file can then be parsed to produce (a) total counts of the number of genes and proteins identified, (b) evaluated against known targets (e.g. from ChEMBL/DrugBank) and the universe to identity true/false positive/negatives and the sensitivity/specificity and (c) entered alongside all other used prediction methods (n = 7) to predict each protein's target class and annotate it with other known information.

3.5 Biopharmable Resources and Filters

Gene Ontology (GO) (Consortium, 2004) provides a structured, controlled vocabulary describing biological processes, cellular components and molecular functions. In this case GO cellular component terms were used as an indication that a gene is associated with an accessible location, either in the extracellular space, extracellular region and plasma membrane.

SignalP (Petersen et al., 2011) predicts the presence of a secretory signal peptide, a protein sorting signal which, in eukaryotes, targets its passenger protein for translocation across the endoplasmic reticulum membrane. TMHMM (Krogh et al., 2001) predicts the presence of transmembrane helices. Both methods were trained using a hidden Markov model. The two resources were used here to produce predictions of all the secreted and transmembrane proteins, and therefore genes, in the human genome.

A synopsis of all of the filters used to produce biopharmable estimates can be seen in Figure 3.3.

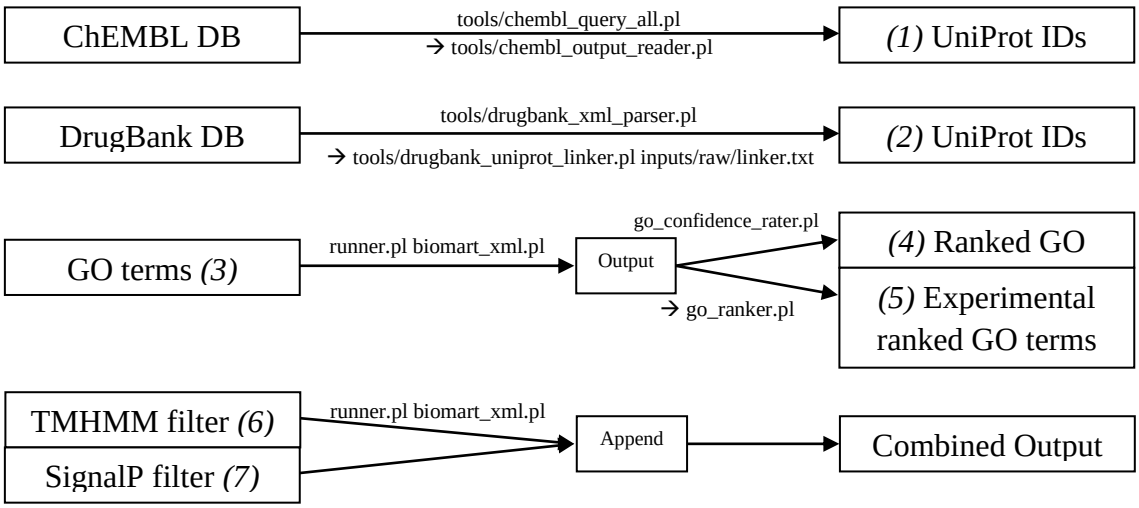

Figure 3.3: **Generation of the seven BioMart query filters for estimating the known (*1* & *2*) and potentially biopharmable targets.**
Bracketed numbers indicate an input which is used as a filter in a BioMart query, large arrows indicate the processing of the original input using scripts (included) or unix commands, smaller arrows indicate the output of the previous script is processed. ChEMBL is a local database, DrugBank is an XML file. The filters *6* and *7* are combined within the pipeline to produce a single BioMart output.

Filters (1) and (2) from Figure 3.3 are used to generate each set of known approved biotechnology targets (from ChEMBL and DrugBank). If a gene/protein from the Ensembl human genome is associated with a UniProt identifier on either list, information including its Ensembl identifiers and description.

All genes/proteins in the Ensembl human genome associated with a GO term indicating an extracellular or membrane bound location make up the biopharmable genome

prediction when (3) is used as a BioMart query filter. The output of this GO method is then filtered to include only genes/proteins which score above five on their experimental evidence code (allowing the possibility of medium or high) and the confidence this GO term is accessible to biotechnology (again, medium or high, but not both medium) – the GO above five method. This is further filtered to allow only genes/proteins with an experimental evidence code to support their associated GO term, the GO experimental rank above five method.

The SignalP and TMHMM method predicts all genes/proteins in the Ensembl human genome which have a predicted signal peptide or transmembrane region to be biopharmable.

3.6 Biopharmable Genome Outputs

For each biopharmable genome prediction method, the input files shown in Figure 3.3 and contained within the 'inputs' directory of the attached ZIP file, were taken as the BioMart query filter and all associated genes and proteins from Ensembl human genome assembly GRCh37.p6 were returned in a tab delimited format with each new line containing: Interpro ID, Interpro Short Description, UniProt/SwissProt Accession, PDB ID, PFAM ID, EntrezGene ID, Ensembl Gene ID, Ensembl Transcript ID, Ensembl Protein ID, Associated Gene Name, Associated Transcript Name, Status (gene), Status (transcript), Ensembl Family Description and Description. If GO terms were included in the query, GO terms are also returned.

As for the chemically tractable predictions, outputs can be compared using produced non redundant gene and protein lists and each method was evaluated against known biotechnology targets and the universe of protein coding genes.

A complete summary file is also produced, showing the Ensembl, Entrez and UniProt identifiers for each predicted gene, a predicted target class and subclass, a ChEMBL identifier, class and subclass(es) if applicable, whether it is known to be approved and, if so, which database this information came from (ChEMBL or DrugBank) and which method(s) predicted it. Figure 3.4 shows a synopsis of the output files produced.

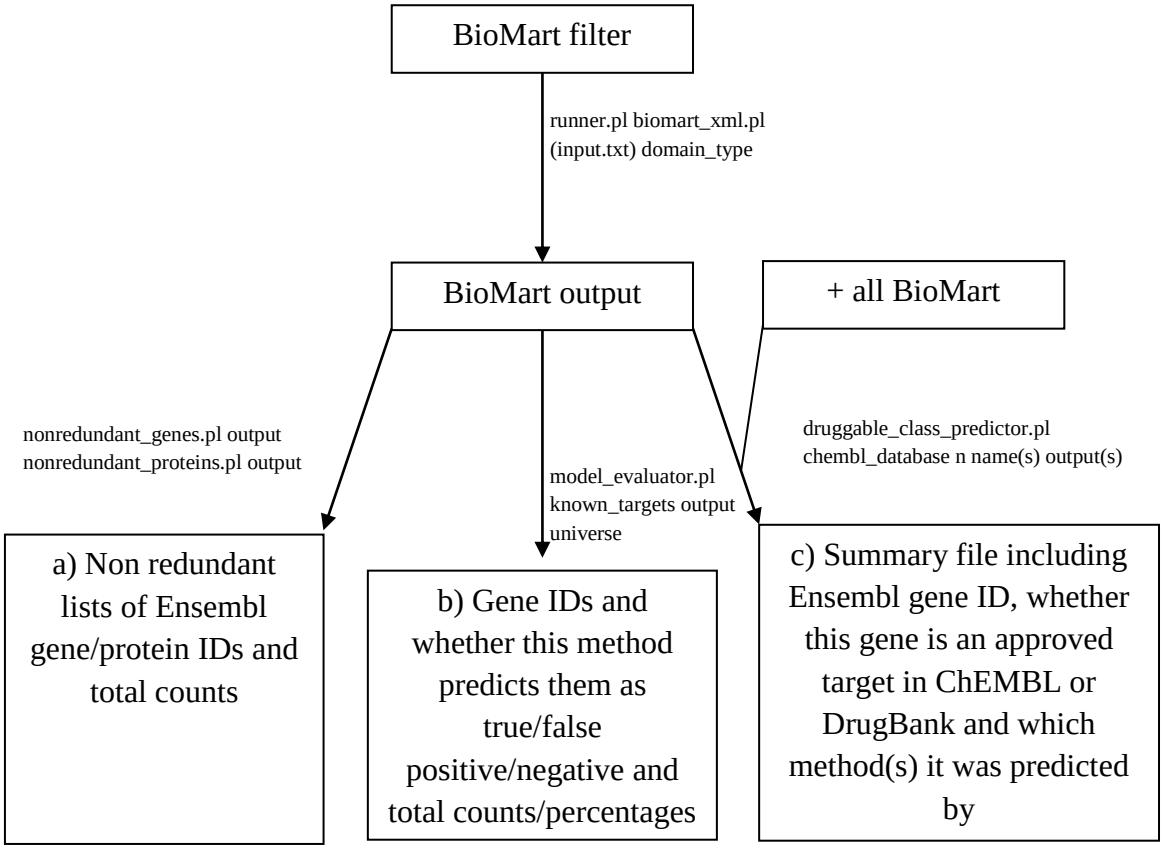

Figure 3.4: **Pipeline producing summary (a) data for comparisons, (b) method evaluations and (c) a complete summary of all the biopharmable outputs.**

Arrows indicate the processing of each file with the scripts annotated on the arrow. BioMart filters are the InterPro domains etc seen on the right in Figure 3.1 and these files are the input for each query, unless the type is universe, sigp or tmhmm, in which case no input is required. Available query types are uniprot, interpro, pfam, pdb, go, sigp, tmhmm or universe. This output file can then be parsed to produce (a) total counts of the number of genes and proteins identified, (b) evaluated against known targets (e.g. from ChEMBL/DrugBank) and the universe to identity true/false positive/negatives and the sensitivity/specificity and (c) entered alongside all other used prediction methods (n = 4) to predict each protein's target class and annotate it with other known information.

Additionally, GO terms were filtered according to their evidence code, as shown in Table 3.2 (see Appendix 8.5 for an explanation of each code), and whether the GO term is likely to be in an accessible location, an example of which is shown in Table 3.3. Terms which least ambiguously describe a location likely to be accessible by biotechnology are ranked more highly than more ambiguous terms or those which are less likely to be accessible.

Table 3.2: **Confidence ranking assigned to each GO evidence code.**

GO Evidence Code	Confidence
TAS, EXP, IDA, IPI, IMP, IGI, IEP	high (3)
ISS, ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD, RCA, IEA	medium (2)
NAS, IC, ND, NR	low (1)

Table 3.3: **Example of the confidence level assigned to each returned child GO term.**
The GO terms input as the BioMark filter are shown in bold.

Go Term	Go Term Name	Confidence Biopharmable
GO:0005615	**extracellular space**	**high (3)**
GO:0005578	proteinaceous extracellular matrix	high (3)
GO:0005576	**extracellular region**	**high (3)**
GO:0005886	**plasma membrane**	**medium (2)**
GO:0071438	invadopodium membrane	high (3)
GO:0046691	intracellular canaliculus	low (1)
GO:0009898	internal side of plasma membrane	no (N/A)

Ranking each child GO term is useful where an ambiguous parent term, such as plasma membrane which could refer to proteins exposed externally or internally to the cell, has less ambiguous child terms, for example as seen in Figure 3.5 where “integral to plasma membrane” indicates a protein would be inaccessible but its child term, “integrin complex”, indicates it is able to bind ligands from the extracellular matrix.

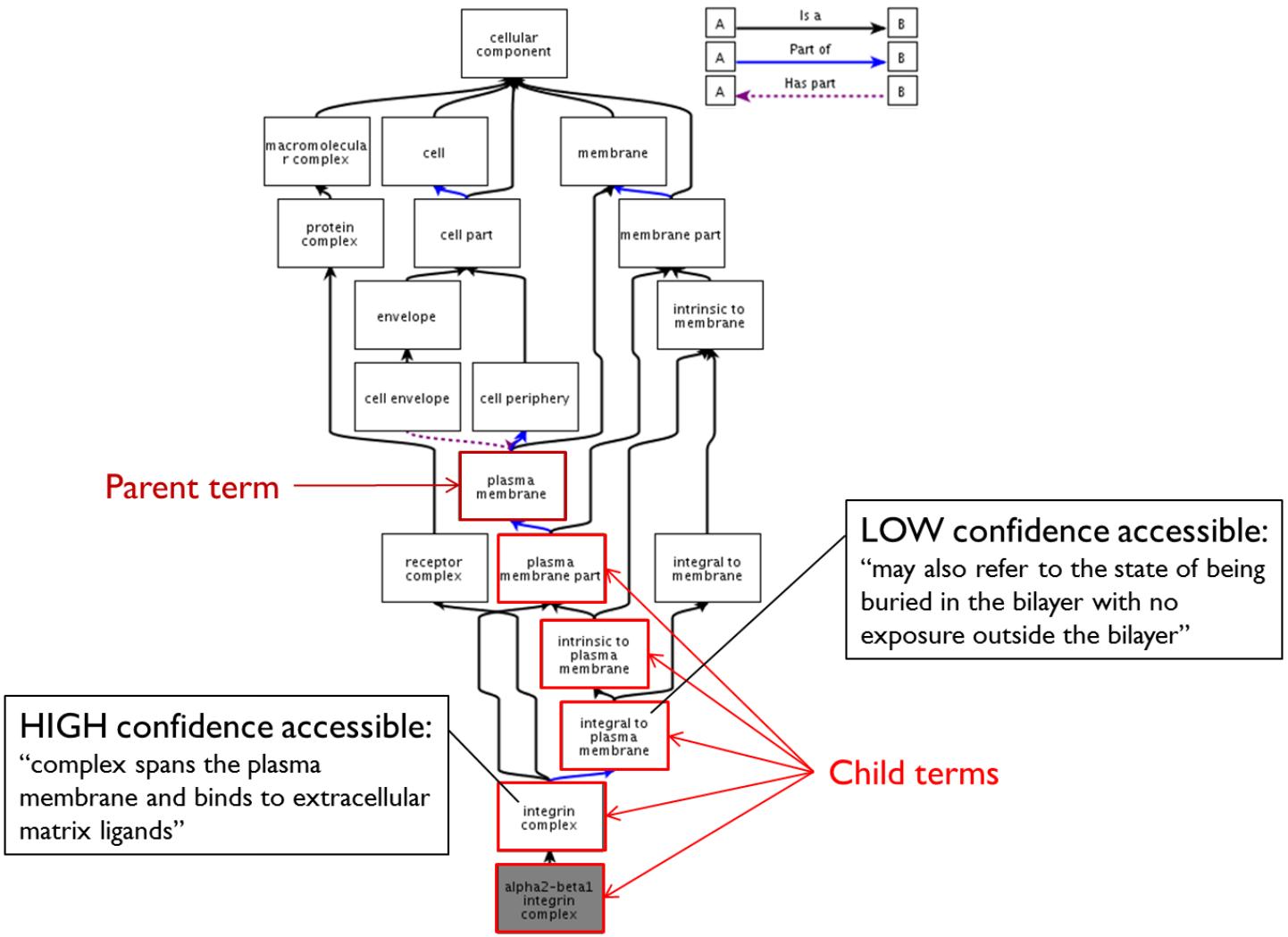

Figure 3.5: **Example of GO terms with the parent term "plasma membrane" with different accessibility confidence levels (image edited from Binns et al., 2009).**

3.7 Prediction Method Evaluation

For all chemically tractable or biopharmable target prediction methods, the protein coding universe of the human genome was taken to be 21,405 genes (coding 96,535 proteins) extracted from Ensembl human genome assembly GRCh37.p6. Genes present in the universe which were not predicted by a given method or a known to be a target will be considered to be true negatives.

When predicting chemically tractable targets, the ChEMBL 13 database (accessed on 07/08/12, ChEMBL 14 is now available) provided 763 genes, coding 1,122 proteins, as known targets of phase IV small molecule drugs and the Drugbank database (downloaded on 10/06/12) lists 1,870 genes, coding 2,649 proteins, as the targets of approved small molecule drugs.

To attempt to ensure only real targets of phase IV drugs were captured the ChEMBL targets all display significant activity with the phase IV drug. Multiple measures of activity were used in order not to exclude any targets on the basis, for example, that they are targeted by agonists but only inhibition was measured therefore capturing only the targets of antagonists. This was taken to be an activity concentration (AC50), inhibitory concentration (IC50), effective concentration (EC50) or effective dose (ED50) above 100nM, an inhibitory constant (Ki) above 0.1nM or a recorded activity or inhibition above 25%.

For biotechnology targets, the ChEMBL database returned 79 genes (coding 118 proteins) as the targets of phase IV protein type drugs and the Drugbank database listed 1,171 genes (coding 1,665 proteins) as the targets of approved biotechnology.

A target class was predicted using regex capture for each predicted protein, for example if the description contained the phrase “ase”, “p450” or “rhodanese” and not the phrase “bcl2”, “release”, “ase coupled” or “ase activated” then it is pulled into a loop which then attempts to find the subclass of the enzyme, for example if it contains the phrase “peptidase” or “protease” it is placed into the enzyme major class and protease subclass. Transporters are also captured within this loop, in addition to separately, if the description matches “transport”, “neurotransmitter” or “carrier” as some were being found to contain the phrase “ATPase”. The full script can be found in the attached ZIP file at ‘scripts/druggable_class_predictor_and_chembl_class.pl’.

4 Results

Using a combination of every prediction method, 12,577 genes (58.8% of all protein coding genes) were predicted to be chemically tractable and 9,660 genes (45.1% of all protein coding genes) were predicted to be biopharmable. Figure 4.1 shows that, if the predicted genes are taken to be true, there are 6,024 genes exclusively open to modulation by small molecule drugs and 3,107 genes which could be exploitable only through targeting with biotechnology.

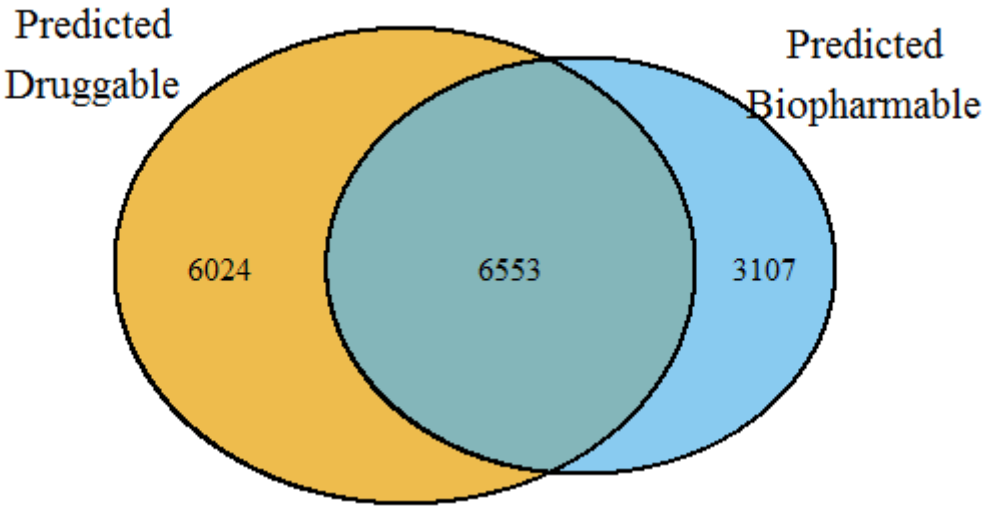

Figure 4.1: **Comparison of the number of genes identified using all prediction methods as potentially druggable and biopharmable.**

A more conservative estimate which requires each druggable gene to be predicted by four methods reduces the druggable estimate by 54.5% (6,856 genes) to a total of 5,721 predicted druggable genes. Requiring each biopharmable gene to be predicted by three methods reduces the estimate by 54.1% (5,229 genes) to a total of 4,431 predicted biopharmable genes. As shown by Figure 4.2, the mutually exploitable overlapping area decreases as well as the overall number of genes predicted.

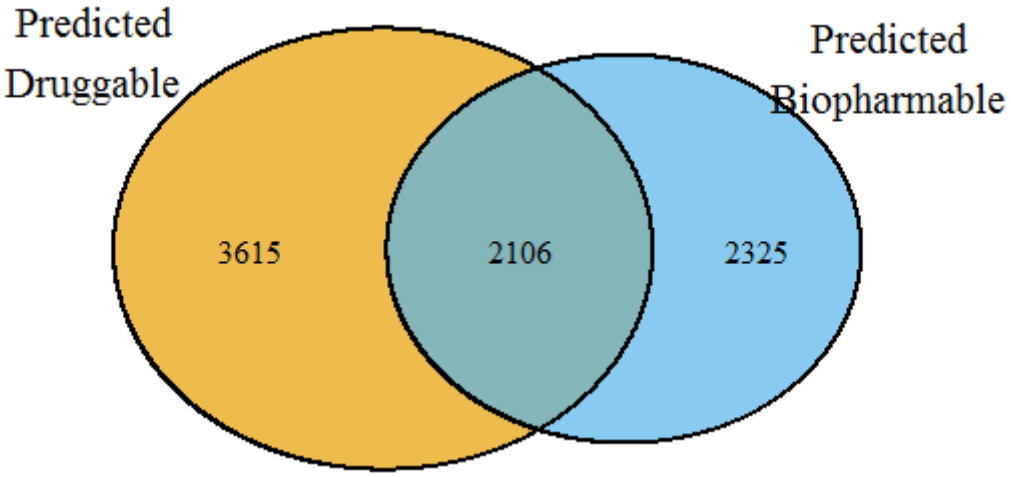

Figure 4.2: **Comparison of the number of genes identified using four or more druggable methods and three or more biopharmable methods as potential target genes.**

4.1 Druggable Genome

The ChEMBL database provides 763 genes (coding 1,122 proteins) as known targets of phase IV small molecule drugs and the Drugbank database lists 1,870 genes (coding 2,649 proteins) as the targets of approved small molecule drugs. As seen in Figure 4.3, there are fewer targets from ChEMBL than from DrugBank.

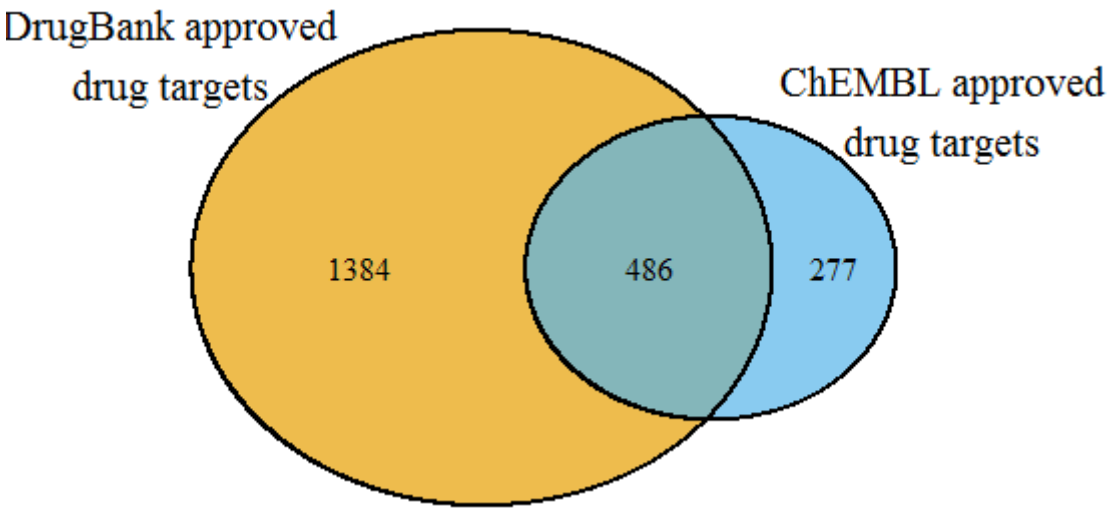

Figure 4.3: **Comparison of genes listed as the target of approved small molecule drugs in DrugBank and those showing significant activity with a phase IV small molecule drug in ChEMBL.**

The two sets also show little overlap, with only 486 indicated as targets in both databases. This represents 26% of data from DrugBank, which captures an additional 1,384 genes, and 64% of data from ChEMBL, which provides an additional 277 genes. Therefore each target prediction method will be evaluated on its ability to correctly predict each data set individually.

4.1.1 Prediction of ChEMBL database small molecule targets

The method which identifies genes with one of the same InterPro domains as an approved DrugBank target correctly predicted the most ChEMBL targets (679 genes, 89% of all known targets) out of all the methods. However, it also predicts a total of 12,225 genes to be druggable, over 57% of the total number of protein coding genes in the human genome and much more than any other method (Figure 4.4). The method which identifies genes with one of the same InterPro domains from a ChEMBL phase four drug target correctly predicted 87% of the available targets and offered a much more conservative estimate of 5,744 druggable genes (26.8% of the genome).

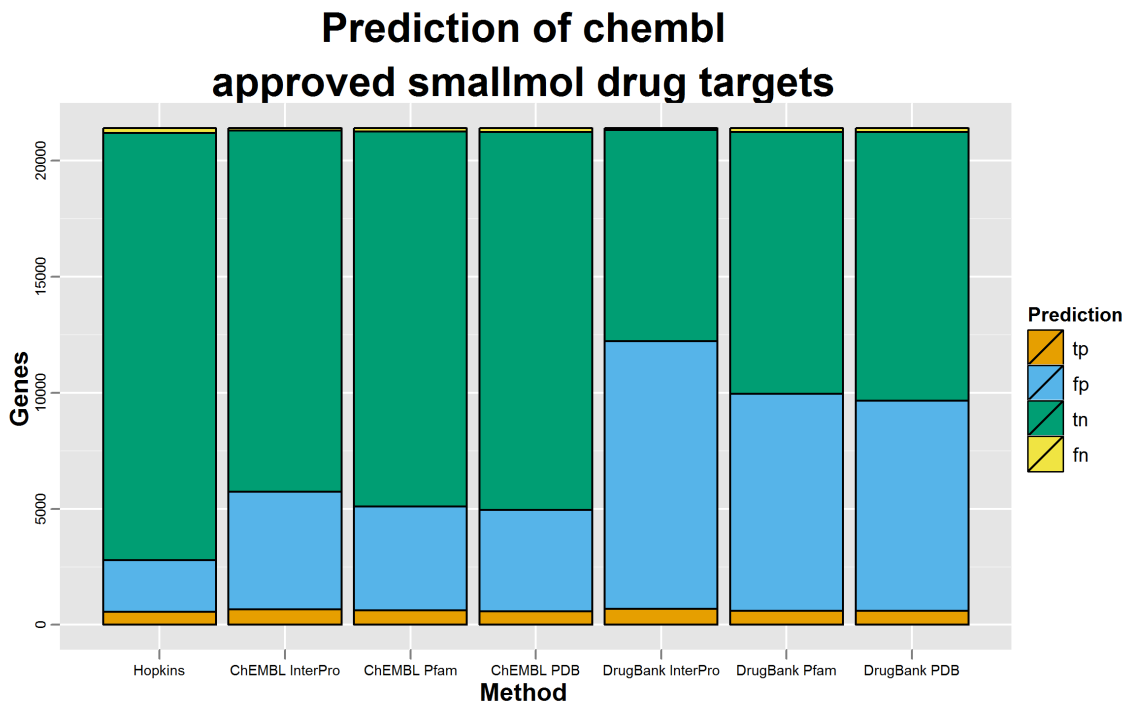

Figure 4.4: **Proportion of ChEMBL database phase IV small molecule target genes predicted correctly by each method.**

The orange section (tp, true positive) represents correctly predicted targets; blue (fp, false positive) indicates genes which are not approved targets in the ChEMBL database which have been predicted to be druggable by this method; green (tn, true negative) represents protein coding genes which were not predicted as druggable and are not known to be drug targets; and yellow (fn, false negative) represents the number of known targets which have been missed by each prediction method. The sum of the orange and blue areas indicates the proportion of the genome predicted as druggable.

Identifying genes coding a protein which contains one or more of the Hopkins and Groom identified druggable InterPro domains offers the most conservative estimate, correctly predicting the fewest ChEMBL targets (552 genes, around 72% of all approved targets) but also predicting the fewest genes overall. A total of 2,779 genes were predicted, around 13% of the protein coding human genome.

Of interest, there was little difference observed between identifying genes containing one or more of every Pfam domain found in an approved drug target (i.e. ChEMBL Pfam or DrugBank Pfam) and only those Pfam domains more likely to be drug binding as it has been found to be associated with a free ligand in PDB. Each method predicted a total of 5,097 and 4,960 druggable genes respectively using ChEMBL targets, both around 23% of the human genome. Of the 137 fewer genes predicted by the more conservative ChEMBL PDB method, 26 were approved ChEMBL drug targets.

The Pfam domains from DrugBank targets which are associated with a free ligand PDB correctly predicted slightly more targets than the same from ChEMBL targets, predicting five more genes correctly. This equates to a very small difference percentage wise, with 78% of possible targets correctly predicted by DrugBank PDB and 77% by ChEMBL PDB. But since the ChEMBL PDB method produced a much more conservative druggable genome estimate, with 23% of the human protein coding genes predicted against 45% using DrugBank PDB (but only missed approximately 1% more true targets), this method could be considered more reasonable.

4.1.2 Prediction of DrugBank database small molecule targets

Once again the DrugBank target InterPro method correctly predicted the most approved drug targets, correctly predicting 1,846 out of a total of 1,870 approved targets (nearly 99%). Again, however, it does predict over 57% of the human genome to be druggable (12,225 genes) as seen in Figure 4.5.

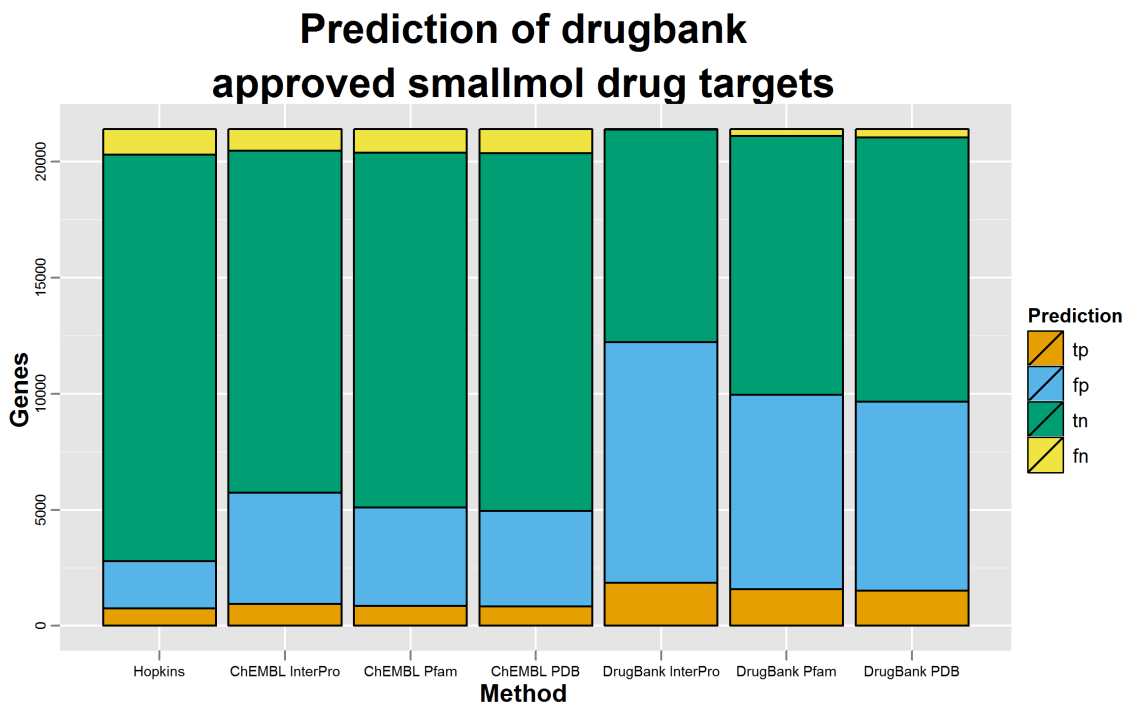

Figure 4.5: **Proportion of DrugBank database approved small molecule target genes predicted correctly by each method.**

The orange section (tp, true positive) represents correctly predicted targets; blue (fp, false positive) indicates genes which are not approved targets in the ChEMBL database which have been predicted to be druggable by this method; green (tn, true negative) represents protein coding genes which were not predicted as druggable and are not known to be drug targets; and yellow (fn, false negative) represents the number of known targets which have been missed by each prediction method. The sum of the orange and blue areas indicates the proportion of the genome predicted as druggable.

The Hopkins and Groom method offered the most conservative estimate, but missed 1,110 approved drug targets from DrugBank (Figure 4.5), compared to only 211 false negatives using ChEMBL data (Figure 4.4). Therefore, only 41% percent of DrugBank targets were predicted by this method compared to 72% of ChEMBL targets.

Similarly, little difference was observed between all Pfam domains from approved targets and only using those with structures with free ligands in PDB (292 fewer genes, 66 of which were true positives). Again, 23% of the genome was estimated to be druggable using ChEMBL PDB and 45% using DrugBank PDB, however DrugBank data was significantly better at correctly predicting DrugBank targets than ChEMBL

data (81% versus 45%) when compared to predicting ChEMBL targets (78% versus 77%).

Higher numbers of false negatives are observed overall in Figure 4.5 compared to Figure 4.4, particularly from more conservative methods such as the Hopkins and Groom InterPro domains or ChEMBL Pfam domains. This could be due to the larger overall number of known targets in DrugBank or the DrugBank database may list targets which display low levels of activity with approved drugs or more types of drug.

4.1.3 Druggable Predictions

When identifying genes containing InterPro domains or Pfam domains from the same set of known targets (i.e. ChEMBL or DrugBank), Pfam offers the more conservative estimate. Predicting known targets from ChEMBL using Pfam or InterPro domains from the ChEMBL targets, the same 4,587 genes were predicted as druggable using both methods, with extra 510 predicted using Pfam domains, as opposed to the extra 1,157 using InterPro domains (Figure 4.6).

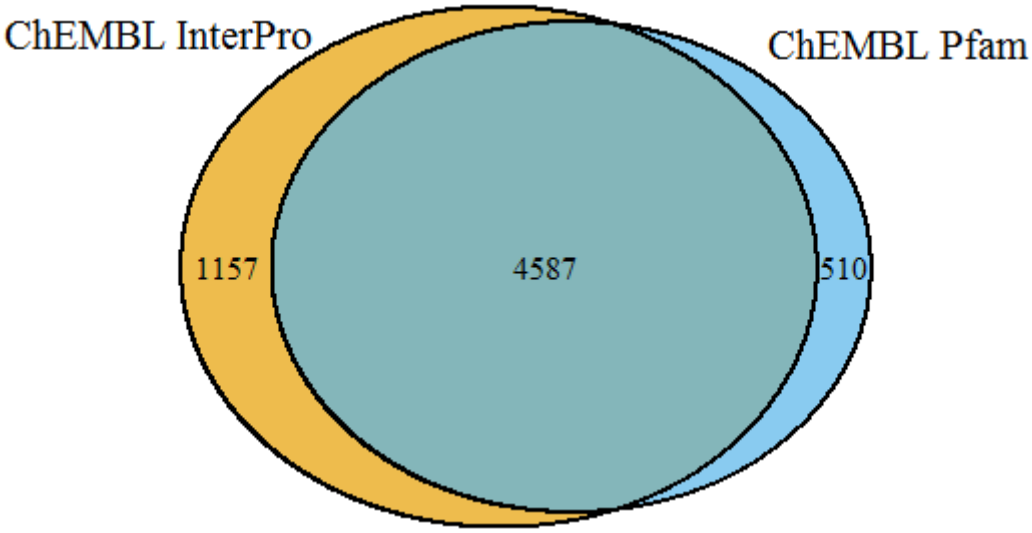

Figure 4.6: **Comparison of predicted druggable genes using InterPro or Pfam domains from proteins showing significant activity with a phase IV small molecule drug in ChEMBL.**

Querying PDB to return only Pfam domains associated with a free ligand excluded 77 Pfam domains from ChEMBL targets with either no known ligand or no structure in PDB, leaving 372 Pfam domains in total and resulting in a prediction 137 genes smaller. Similarly, the PDB query excluded 208 Pfam domains from DrugBank targets, leaving 978 Pfam domains in total and resulting in a prediction 292 genes smaller.

When two of the most conservative estimates were compared (Hopkins and Groom InterPro and ChEMBL Pfam) it was found that 2,100 genes were common to both, as shown in figure 4.7. The ChEMBL Pfam method identified 2,318 more genes overall, with 2,997 of these not containing an InterPro domain identified by Hopkins and Groom. Identified InterPro and Pfam domains can be found under the 'inputs' directory of the submitted ZIP file, in a directory called the database name, 'chembl' or 'drugbank'.

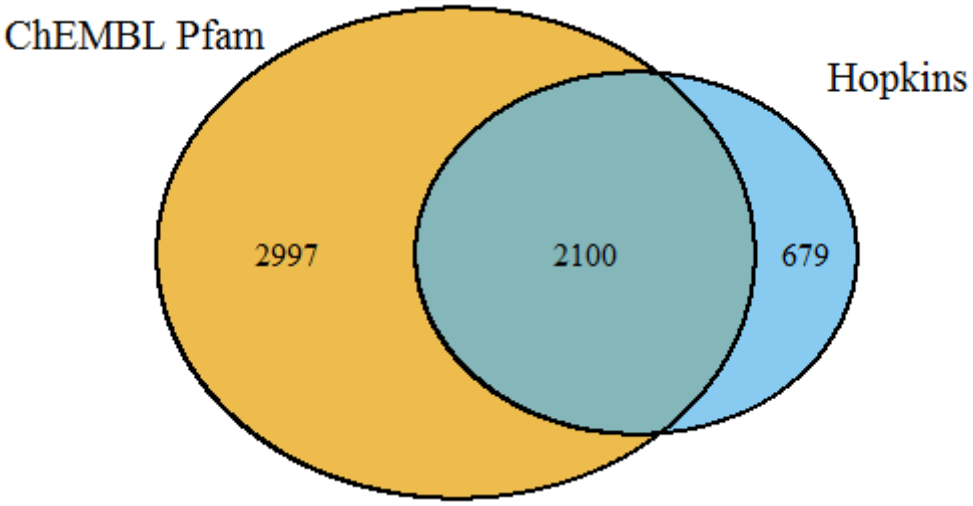

Figure 4.7: **Comparison of genes identified as potentially druggable using the Hopkins and Groom InterPro domains from 2002 and all Pfam domains from proteins showing significant activity with a phase IV small molecule drug in ChEMBL.**

Figure 4.8 compares the most conservative estimate, Hopkins and Groom InterPro domains, against the largest estimate, DrugBank InterPro domains. Almost 98% of genes predicted as druggable by the Hopkins and Groom InterPro were also predicted by DrugBank InterPro. However, 58 genes were missed by DrugBank InterPro and this method also predicts 9,504 genes extra genes, around 4.4 times as many as predicted by Hopkins and Groom InterPro.

There were 42 InterPro domains included in the Hopkins and Groom set not observed in DrugBank targets, 102 domains in common and 2,036 extra domains gained from DrugBank targets. Similarly, there were 69 InterPro domains from Hopkins and Groom which were not observed in ChEMBL targets, 75 domains in common and 857 extra InterPro domains obtained from the ChEMBL targets.

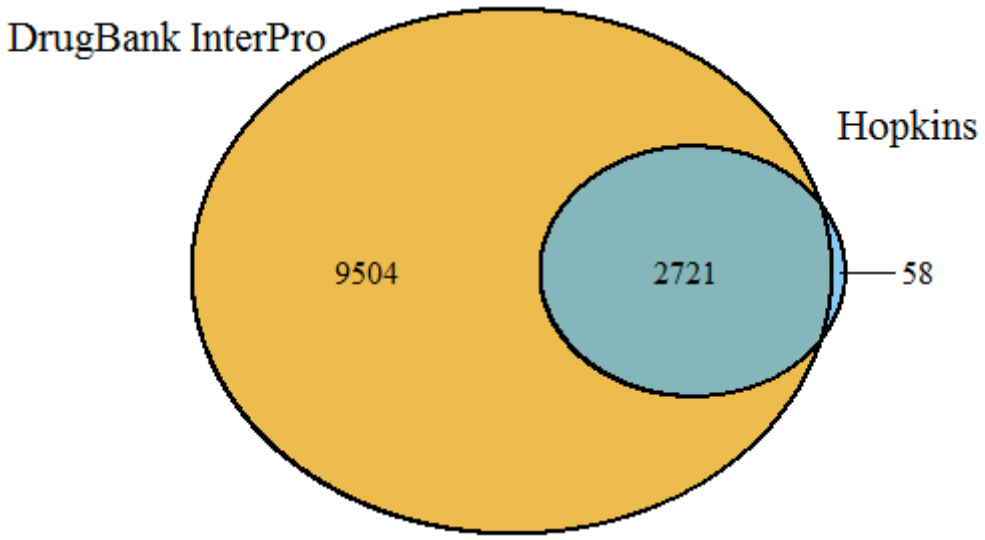

Figure 4.8: **Comparison of genes identified as potentially druggable using the Hopkins and Groom InterPro domains from 2002 and all InterPro domains of proteins listed as the target of an approved drug in DrugBank.**

4.1.4 Sensitivity and Specificity of methods predicting ChEMBL small molecule targets

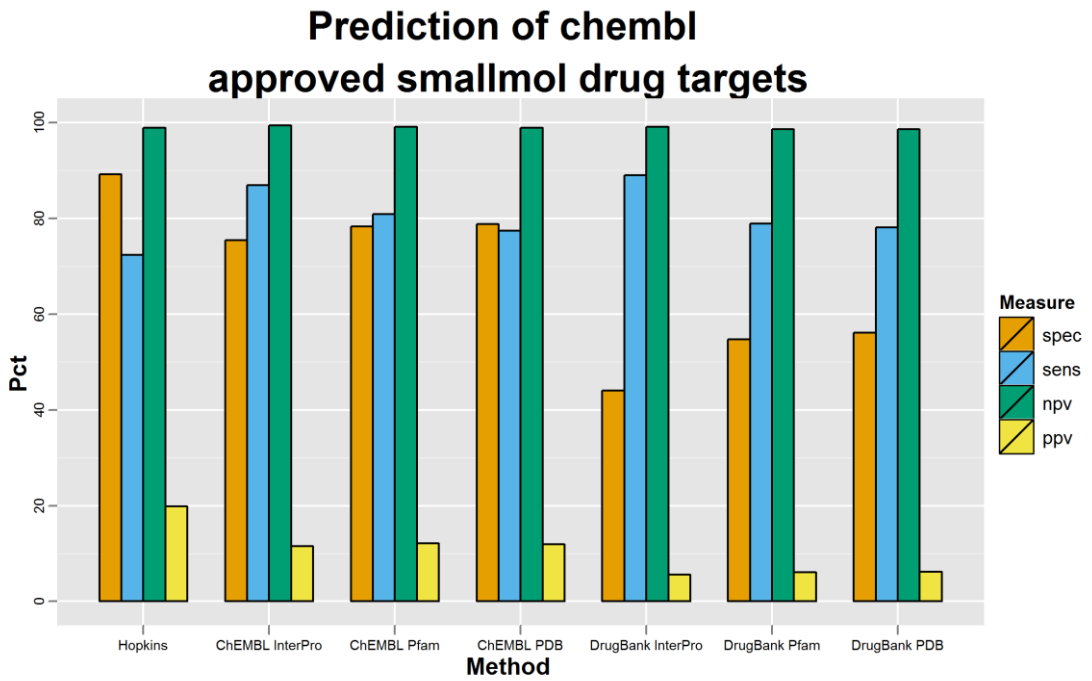

Figure 4.9: **Evaluation of each method when predicting ChEMBL phase IV drug targets.** Orange bars (spec) show specificity; blue (sens), sensitivity; green (npv), negative predictive value; and yellow (ppv), positive predictive value.

Specificity is the percentage of genes which are not known drug targets and were not predicted to be druggable. Sensitivity is the percentage of known drug targets which are correctly predicted to be druggable. The negative predictive value indicates the percentage of genes not known to be druggable which were not identified as druggable. The positive predictive value indicates the percentage of genes identified as druggable which are known drug targets.

Seen in Figure 4.9, the prediction method using Hopkins and Groom InterPro domains is the most specific, correctly identifying 89.2% of true negatives, with the next highest, using PDB confirmed ChEMBL target Pfam domains, at 78.8%. The ChEMBL-derived methods are significantly more specific than the DrugBank-derived methods.

The Hopkins and Groom method also provides the highest positive prediction value, showing it correctly predicts a druggable target 19.9% of the time with the next best percentage coming from the use of ChEMBL Pfam domains (12.1%). The DrugBank methods show poor positive predictive power, from 5.6-6.2%.

Using all InterPro domains from DrugBank targets results in the most sensitive method, correctly identifying 89% of ChEMBL targets, with the ChEMBL InterPro domains showing the second highest sensitivity, identifying 86.9% of targets correctly. Overall there is very little variation in sensitivity compared to specificity.

All methods had high negative predictive values of above 98%, however the ChEMBL InterPro method correctly predicts true negatives best at 99.4% of the time, with the DrugBank InterPro next best at 99.1%.

Therefore, when tested against ChEMBL data, the Hopkins and Groom, PDB confirmed ChEMBL Pfam and ChEMBL Pfam methods provide the most conservative predictions of the druggable genome. Using all InterPro domains from known targets provides the largest druggable genome estimate, with InterPro domains from DrugBank targets the most numerous.

4.1.5 Sensitivity and Specificity of methods predicting DrugBank small molecule targets

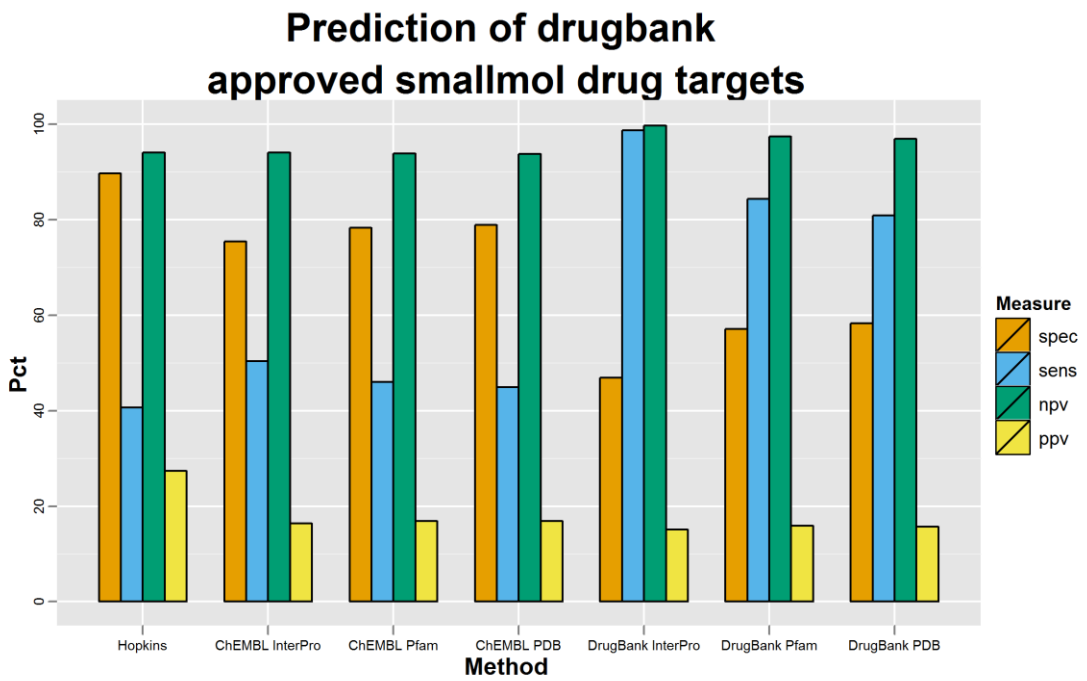

Figure 4.10: **Evaluation of each method when predicting DrugBank approved drug targets.** Orange bars (spec) show specificity; blue (sens), sensitivity; green (npv), negative predictive value; and yellow (ppv), positive predictive value.

Again, as seen in Figure 4.9, Figure 4.10 shows the most specific method for predicting druggable targets is using the Hopkins and Groom InterPro domains, correctly predicting 89.7% of negative genes, with PDB confirmed ChEMBL target Pfam domains next best at 78.9%.

Hopkins and Groom InterPro domains also produce the best positive predictive value, at 27.3%, showing just over a quarter of genes predicted were true positives, with the next best value (16.9%) seen when using PDB confirmed ChEMBL Pfam domains, with all

ChEMBL methods correctly predicting drug targets around 16% of the time and DrugBank methods at around 15%.

The same as when predicting the ChEMBL targets, the DrugBank InterPro method provides the best sensitivity when predicting DrugBank targets. It predicted almost all genes of approved targets (98.7%), with the DrugBank target Pfam domain method showing the second best at 84.3%.

Negative predictive values show much more variation between methods when predicting DrugBank targets than those from ChEMBL. The DrugBank InterPro method is best at correctly predicting negative genes, with 99.7% of genes predicted as negative correctly, and DrugBank Pfam the second best predicting 97.5%. However the Hopkins and Groom domains correctly predict true negatives 94% of the time, indicating this method misses the highest percentage of true DrugBank targets.

Therefore, the Hopkins and PDB confirmed ChEMBL Pfam methods provide the smallest estimates of the druggable genome when tested against DrugBank data. The DrugBank target associated InterPro and Pfam domains provide the most inclusive druggable genome estimates.

4.1.6 Receiver operating characteristic curves for methods predicting small molecule targets

In a receiver operating characteristic (ROC) curve, the area under curve (AUC) represents the method's discriminatory power, a value of 1 would represent a perfect test, with 100% specificity and 100% sensitivity, whereas a value of 0.5 would indicate that the test has no discriminatory power, performing no better than chance.

Figure 4.11 shows that, when predicting significant targets of phase IV drugs from the ChEMBL database, the ChEMBL InterPro method shows the greatest predictive power (0.811) closely followed by the Hopkins and Groom method (0.808). Figure 4.12 shows that, when predicting listed targets of approved drugs from the DrugBank database, the DrugBank target associated InterPro domains provided the best predictive power (0.728).

Each ROC curve also shows the methods' specificity and sensitivity in relation to each other. In Figure 4.11 the most sensitive methods can be seen to be ones using InterPro domains as they have the highest points, whereas the most specific methods are those shifted furthest to the left, for example the Hopkins and ChEMBL Pfam and PDB methods. Additionally, the two ChEMBL Pfam methods show similar specificity and sensitivity values in comparison to the other methods, which tend to be either much more specific or much more sensitive.

Similarly, in Figure 4.12 the highest peak is observed using the DrugBank InterPro method, showing it is most sensitive, and the peak furthest to the left is the Hopkins method, showing it is, again, most specific. However, since these peaks are of smaller magnitude and less centralised, either the DrugBank Pfam or DrugBank PDB method may provide the best all round prediction since it shows similar sensitivity and specificity and good predictive power, with an AUC of 0.707 and 0.696 respectively.

Overall, better predictive power was observed when predicting the ChEMBL targets (Figure 4.11) than DrugBank targets (Figure 4.12), with best AUCs of 0.811 and 0.728.

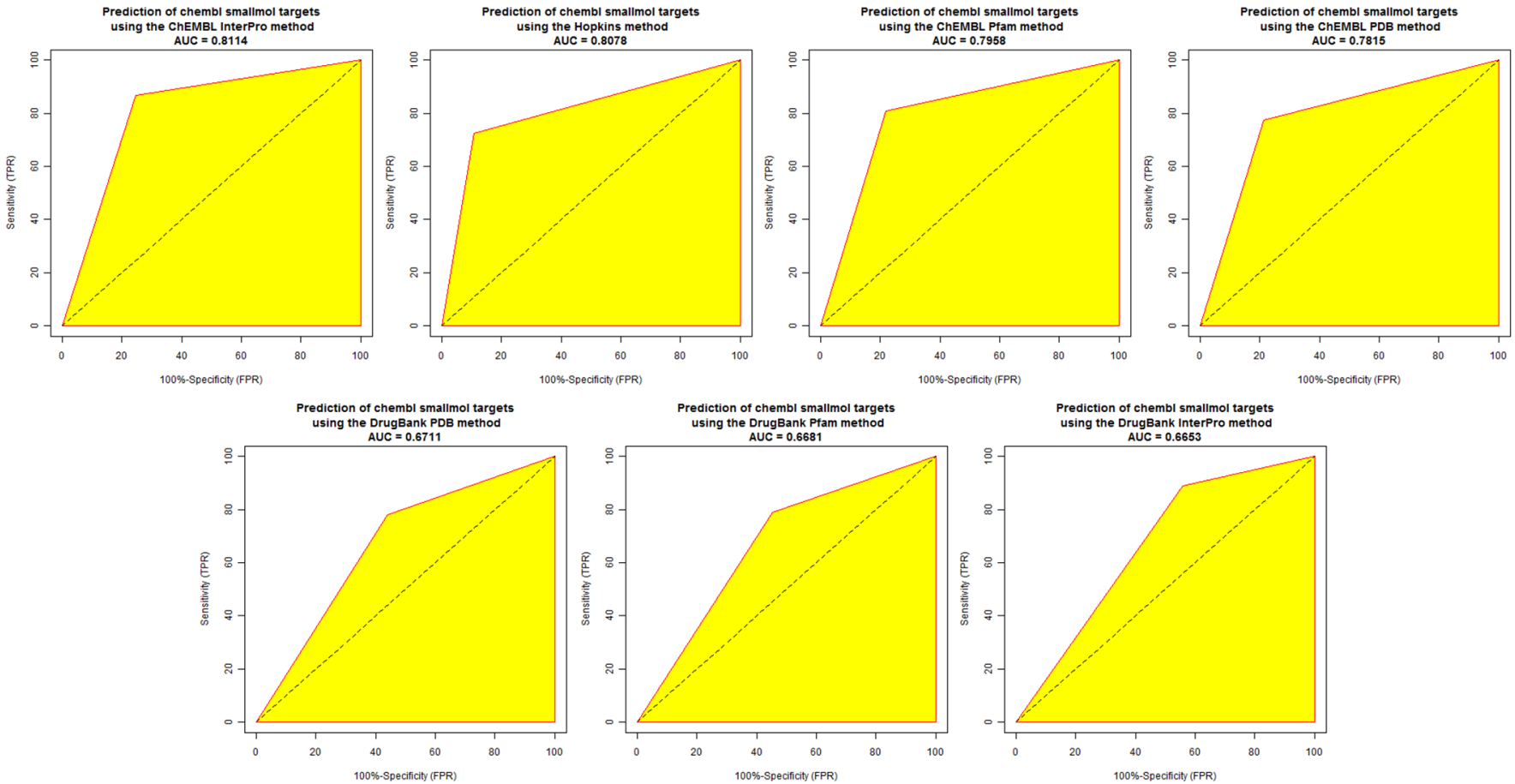

Figure 4.11: **ROC curves showing the predictive power of each method for the ChEMBL targets.**
AUC, area under curve; TPR, true positive rate; and FPR, false positive rate. The dotted line indicates an AUC of 0.5 which is the predictive power which could be observed by chance.

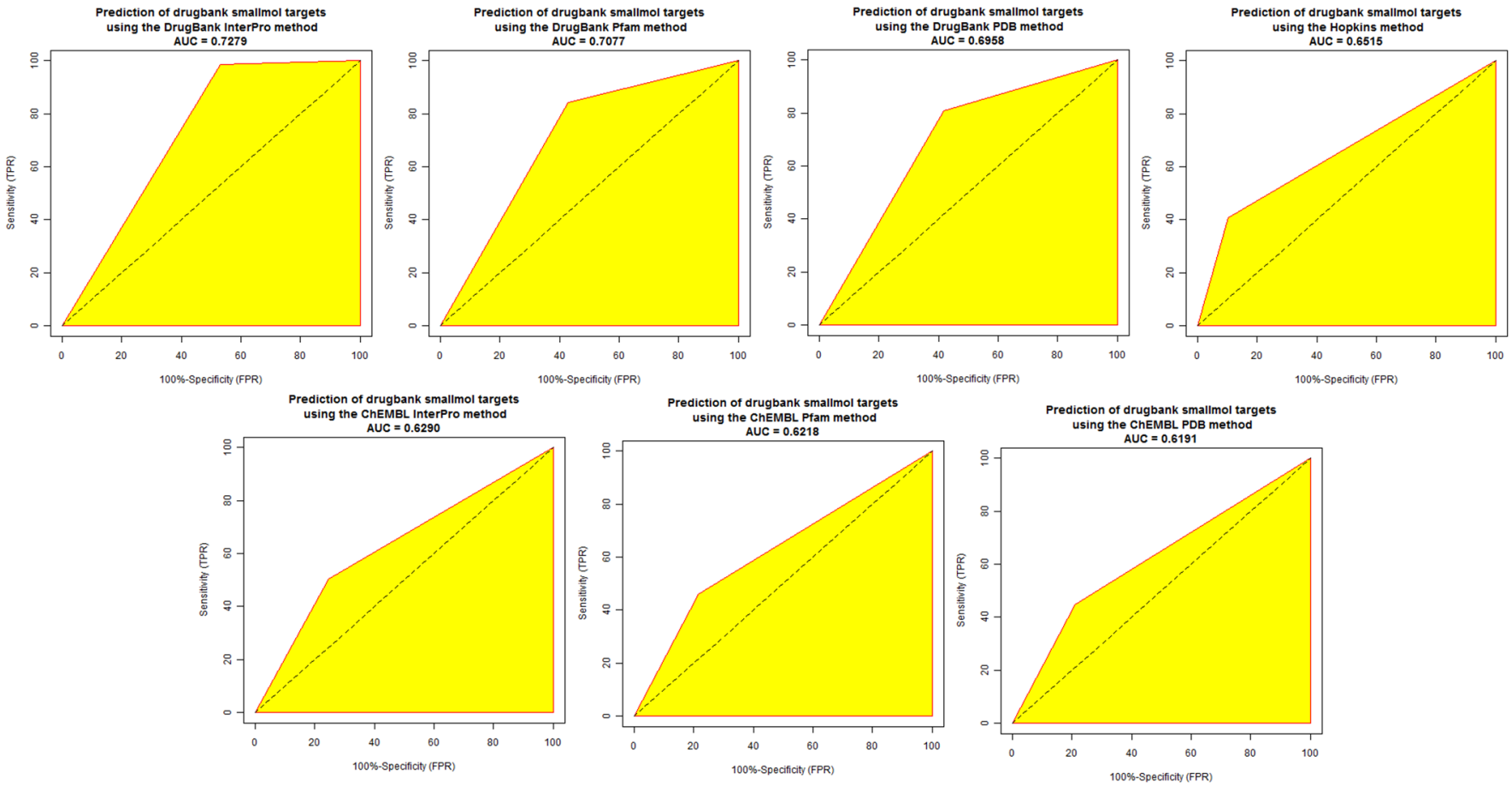

Figure 4.12: **ROC curves showing the predictive power of each method for the DrugBank targets.**
AUC, area under curve; TPR, true positive rate; and FPR, false positive rate. The dotted line indicates an AUC of 0.5 which is the predictive power which could be observed by chance.

4.1.7 Druggable Predictive Power

Both when predicting targets from ChEMBL (which show significant activity with a phase IV small molecule drug) or targets from DrugBank (of approved small molecule drugs), the most specific method used the Hopkins and Groom InterPro domains, which correctly predicted 89.2% and 89.7% of the protein coding genome respectively. The next most specific method was using PDB confirmed ChEMBL target associated Pfam domains, predicting around 79% correctly in each case.

When predicting both the ChEMBL and DrugBank known targets, the most sensitive method was using DrugBank target associated InterPro domains, correctly predicting 89% of ChEMBL targets and 99% of DrugBank targets. For ChEMBL targets, the next most sensitive method was using ChEMBL InterPro domains, at 87%, whereas for DrugBank targets it was using DrugBank Pfam domains, at 84%.

From plotting the ROC curves and calculating the AUC, the greatest predictive power for ChEMBL targets was seen when using ChEMBL InterPro domains, with an AUC of 0.811, classing it as a good prediction (from 0.8 to 0.9) according to Muller *et al.* (Muller et al., 2005). The Hopkins method scored 0.808, also classing itself as a good test, and the ChEMBL Pfam and ChEMBL PDB methods scored 0.796 and 0.782 respectively, classing them as fair tests (0.7-0.8).

The greatest predictive power for DrugBank targets was observed using DrugBank InterPro domains, at 0.728, and DrugBank Pfam domains, at 0.708, classing them as fair. The rest of the methods were poor predictors of DrugBank targets (AUC<0.7).

4.1.8 Hopkins and Groom comparison

By identifying protein coding genes with the InterPro domains from Hopkins and Groom 10 years later, the 2012 predicted druggable space is now 272 genes smaller with 2,779 genes identified compared to the previously identified 3,051. As shown in Figure 4.13, genes coding proteins with GPCR domains now represent the largest percentage of the druggable genome, and protein kinases represent 2% less. The gamma-carboxylase group is now represented by a single gene and, along with metallopeptidases, has been replaced in the top 10 by neurotransmitters and transporters.

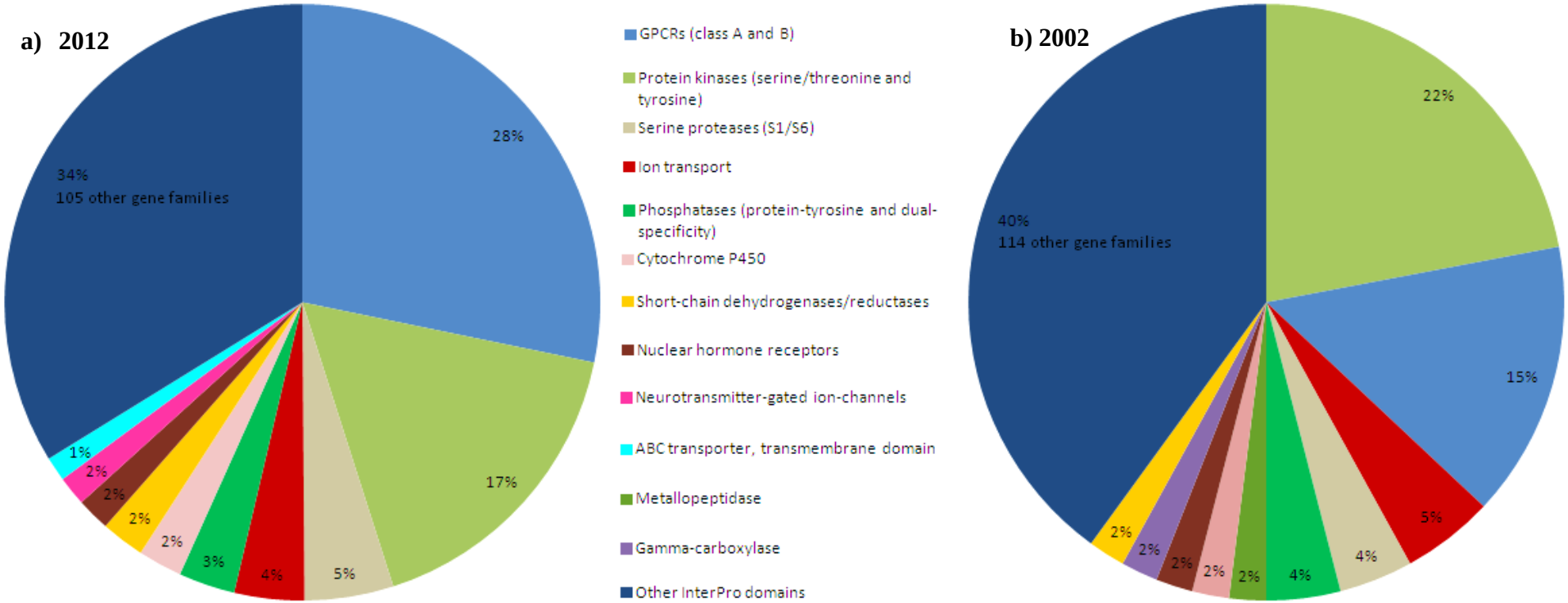

Figure 4.13: **The distribution of the Hopkins and Groom druggable InterPro domains contained within the a) 2,779 genes identified by this method in 2012 and b) 3,051 genes predicted in the orignal paper from 2002. Showing the top 10 most frequently present InterPro domain groups.**
InterPro domains described as rhodopsin-like and secretin-like GPCRs have been combined, as have serine-threonine/tyrosine- and serine-threonine- / dual-specificity protein kinase catalytic domains and the two neurotransmitter-gated ion-channel domains. Every other domain is represented individually.

4.1.8 Predicted Druggable Target Classes

From 1,499 genes which show activity with a phase IV drug, 1,290 genes (86%) had their ChEMBL major target class correctly predicted through regex capture, 169 (11.3%) were incorrectly classified and 40 (2.7%) remained unclassified. Figure 4.14 shows the predicted target classes of all genes predicted as druggable using all seven methods. This takes into account genes returned with more than one description which resulted in a different major classification (e.g. if a BioMart error occurred and an enzyme was returned without a description, the gene would be found in both the enzyme and unclassified category) but if a gene was only associated with one class it is only represented once.

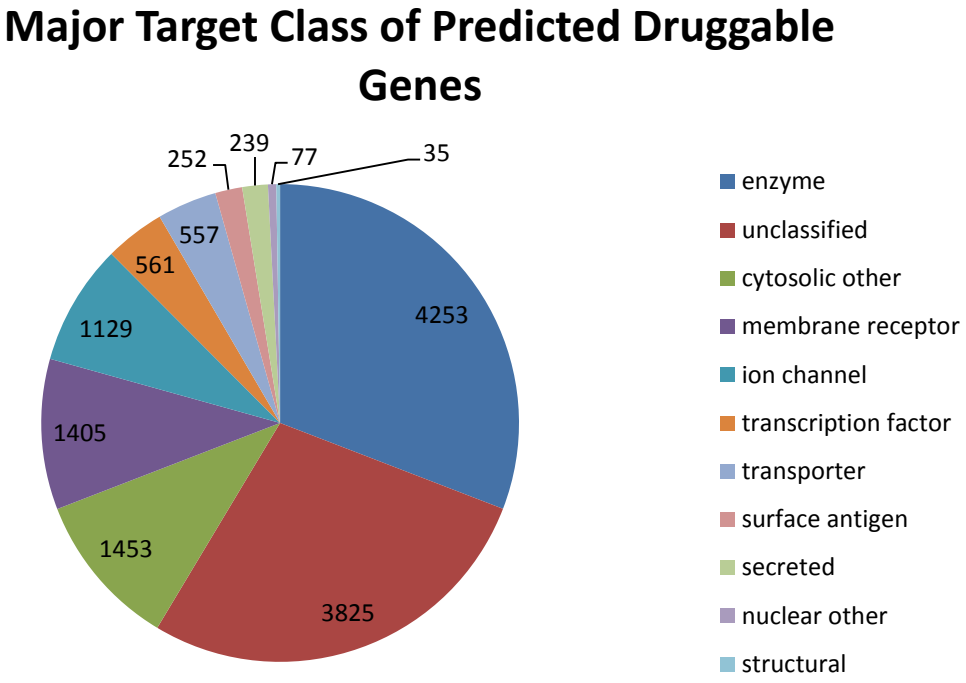

Figure 4.14: **The predicted target class of all genes predicted to be druggable using all prediction methods.**
The total of 13,786 takes into account genes with multiple descriptions resulting in different target classes, the total number of unique genes is 12,225.

The majority (30.9%) of predicted targets were enzymes and 27.8% of targets were not classified using regular expressions which correctly captured the majority of major target classes of known drug targets. Cytosolic proteins were the next largest class at 10.5% of genes, 10.2% of potential targets were receptors, and the remainder were classified as channels (8.2%), transcription factors (4.1%) and transporters (4%). Other target classes represented less than 2% each and less than 5% of the total space.

An attempt was also made to identify some of the major subclasses of each identified gene, the results of which can be seen in Table 4.1. This takes into account genes with differing descriptions causing them to fit into more than one subclass (e.g. if a gene

were to code a protein with a description captured by the kinase regex and also one captured as a protease). Although specific subclasses proved difficult to determine through programmatic capturing of keywords, the majority of enzymes identified were kinases (796 genes), the majority of receptors appear to be olfactory or taste GPCRs (492) and most of the identified ion channels matched as ligand gated (121).

Table 4.1: **The predicted target class and subclass of all genes predicted as druggable.**
The total of 14,281 takes into account genes with multiple descriptions resulting in different target classes or subclasses, the total number of unique genes is 12,225. Seven transmembrane (7TM) refers to members of the GPCR family, 7TM1 for class A GPCRs, 7TM2 for class B and 7TM3 for class C. Other GPCRs include classes such as frizzled and smoothened.

Predicted Target Class / Subclass	Number Of Genes
Enzyme	**4343**
N/A	2451
Kinase	796
Protease	543
Phosphatase	250
Reductase	164
P450	90
Phosphodiesterase	45
Aminoacyltransferase	4
Unclassified	**3825**
Membrane Receptor	**1802**
N/A	565
7TM1 Olfactory/Taste	492
7TM1	300
Other GPCR	218
Nuclear	107
7TM2	60
Integrin	34
7TM3	26
Cytosolic Other	**1453**
Ion Channel	**1137**
N/A	799
Ligand Gated	121
Ryanodine Receptor	114
Voltage Gated	103
Transcription Factor	**561**
Transporter	**557**
Surface Antigen	**252**
Secreted	**239**
Nuclear Other	**77**
Structural	**35**
Total	**14281**

4.2 Biopharmable Genome

The ChEMBL database contains only 79 genes (coding 118 proteins) as known targets of phase IV biotechnology and the Drugbank database lists 1,171 genes (coding 1,665 proteins) as the targets of approved biotechnology. Of the limited ChEMBL coverage of biotechnology targets, less than half of these (39 genes) were also found in the DrugBank database, as seen in Figure 4.15.

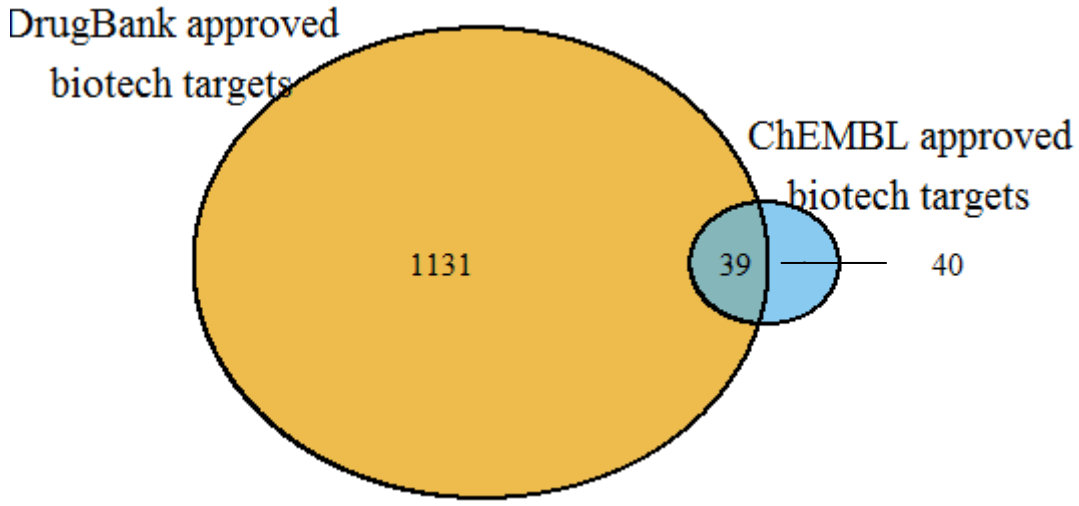

Figure 4.15: **Comparison of genes listed as the target of approved small molecule drugs in DrugBank and those showing significant activity with a phase IV small molecule drug in ChEMBL.**

4.2.1 Prediction of ChEMBL database biotechnology targets

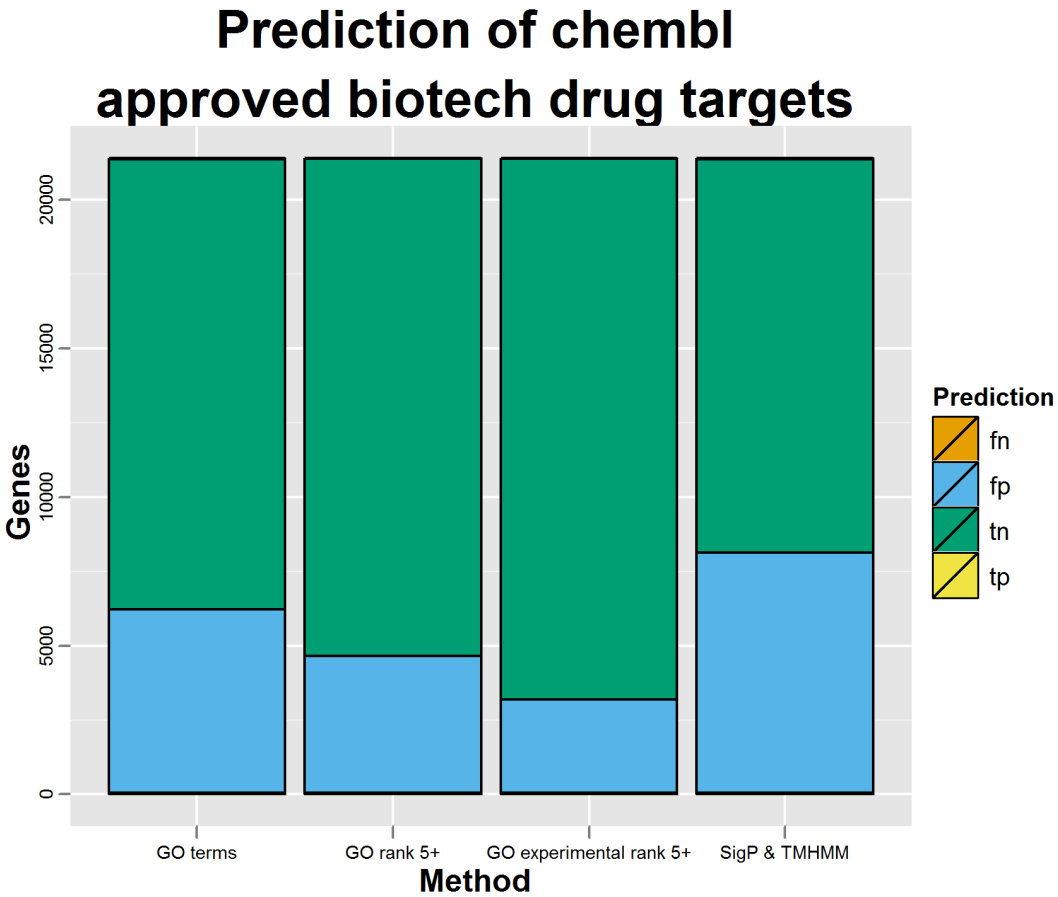

Figure 4.16: **Proportion of ChEMBL database phase IV biotechnology target genes predicted correctly by each method.**

The orange section (tp, true positive) represents correctly predicted targets; blue (fp, false positive) indicates genes which are not approved targets in the ChEMBL database which have been predicted to be biopharmable by this method; green (tn, true negative) represents protein coding genes which were not predicted as biopharmable and are not known to be drug targets; and yellow (fn, false negative) represents the number of known targets which have been missed by each prediction method. The sum of the orange and blue areas indicates the proportion of the genome predicted as biopharmable.

As can be seen in Figure 4.16, the poor coverage of biotechnology targets in ChEMBL means known targets represent less than 0.4% of the protein coding genome. Genes associated with a GO term indicating extracellular or plasma membrane bound location predicted 44% of the known targets (35 genes). The GO terms predicted a total of 6,209 genes and a combination of SignalP and TMHMM predicted 8,117 genes as being biopharmable. The most conservative estimate, the genes associated with GO terms for an accessible location which were then filtered to so only those with experimental evidence for the association and medium or high confidence the associated GO term indicates an accessible location, predicted 30 genes correctly and 3,169 biopharmable genes in total.

4.2.2 Prediction of DrugBank database biotechnology targets

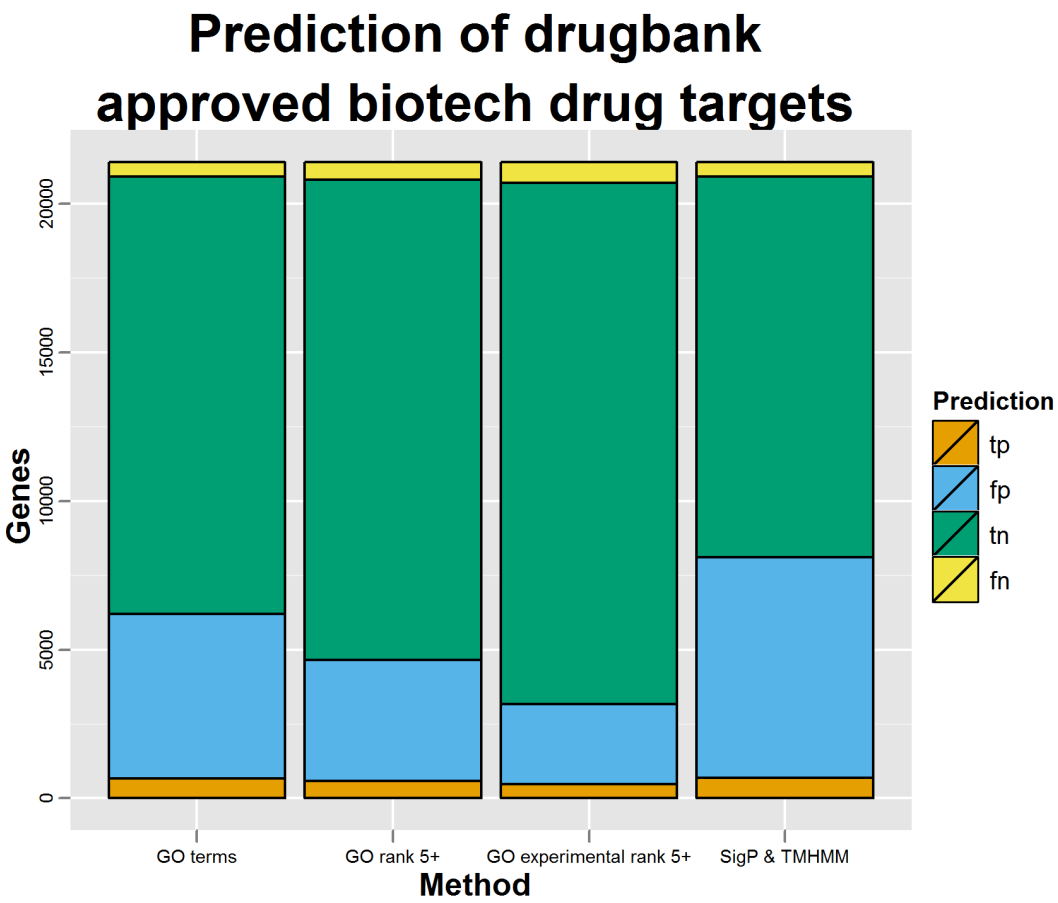

Figure 4.17: **Proportion of DrugBank database approved biotechnology target genes predicted correctly by each method.**
The orange section (tp, true positive) represents correctly predicted targets; blue (fp, false positive) indicates genes which are not approved targets in the DrugBank database which have been predicted to be biopharmable by this method; green (tn, true negative) represents protein coding genes which were not predicted as biopharmable and are not known to be drug targets; and yellow (fn, false negative) represents the number of known targets which have been missed by each prediction method. The sum of the orange and blue areas indicates the proportion of the genome predicted as biopharmable.

Shown in Figure 4.17, the greatest number of true positives (686 genes, ~58% of all approved targets) was identified through filtering genes using a combination of signal peptides (SigP) and transmembrane helices (TMHMM). The second highest, 675 genes, were identified using GO terms for an extracellular or plasma membrane location.

However, these methods also produced the greatest numbers of false positives, with SigP/TMHMM predicting 8,117 genes to be biopharmable (~39% of the human protein coding genome). The GO terms method predicted biopharmable 6,209 genes (29% of the universe).

Filtering the GO term predictions by rank results in fewer false positives, but also fewer true positives, with those ranked above five (meaning when medium equals two and

high equals three, they scored five or above on the sum of the level of confidence in the evidence associating the gene to the GO term and the confidence this GO term indicates an accessible location) predicting 91 fewer target genes than the original prediction. Similarly, the same prediction but with genes removed which did not have experimental evidence to support their location, 200 approved targets are lost, though the number predicted, or false positive, genes is reduced by 2,840, or just over half (~51%).

4.2.3 Biopharmable Predictions

Of the two prediction methods, genes coding proteins predicted to contain a signal peptide or transmembrane region predicted a larger biopharmable genome than those associated with a GO term for extracellular or plasma membrane location. There were 4,666 commonly predicted biopharmable genes, with 3,451 extra genes predicted only by the SigP and TMHMM method and 1,543 unique to the GO term method (Figure 4.18).

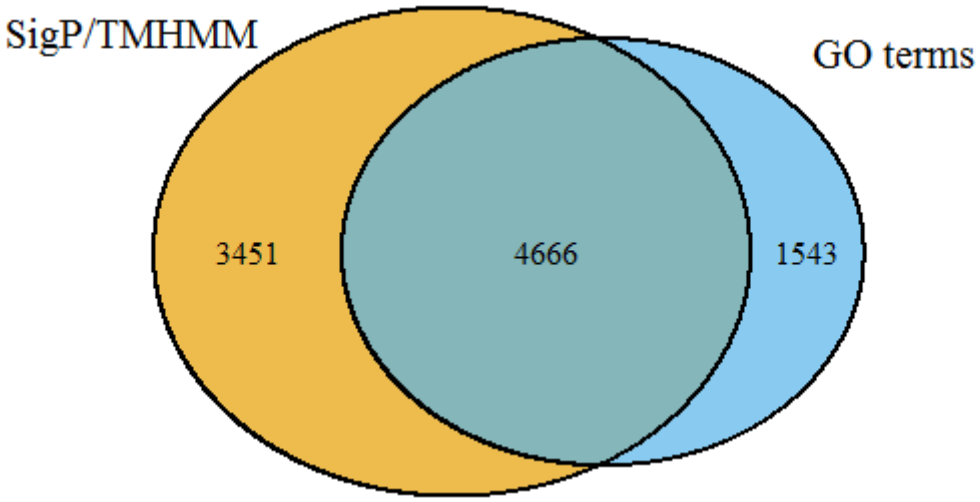

Figure 4.18: **Comparison of the number of genes identified using the two main biopharmable prediction methods.**

4.2.4 Sensitivity and Specificity of methods predicting ChEMBL biotechnology targets

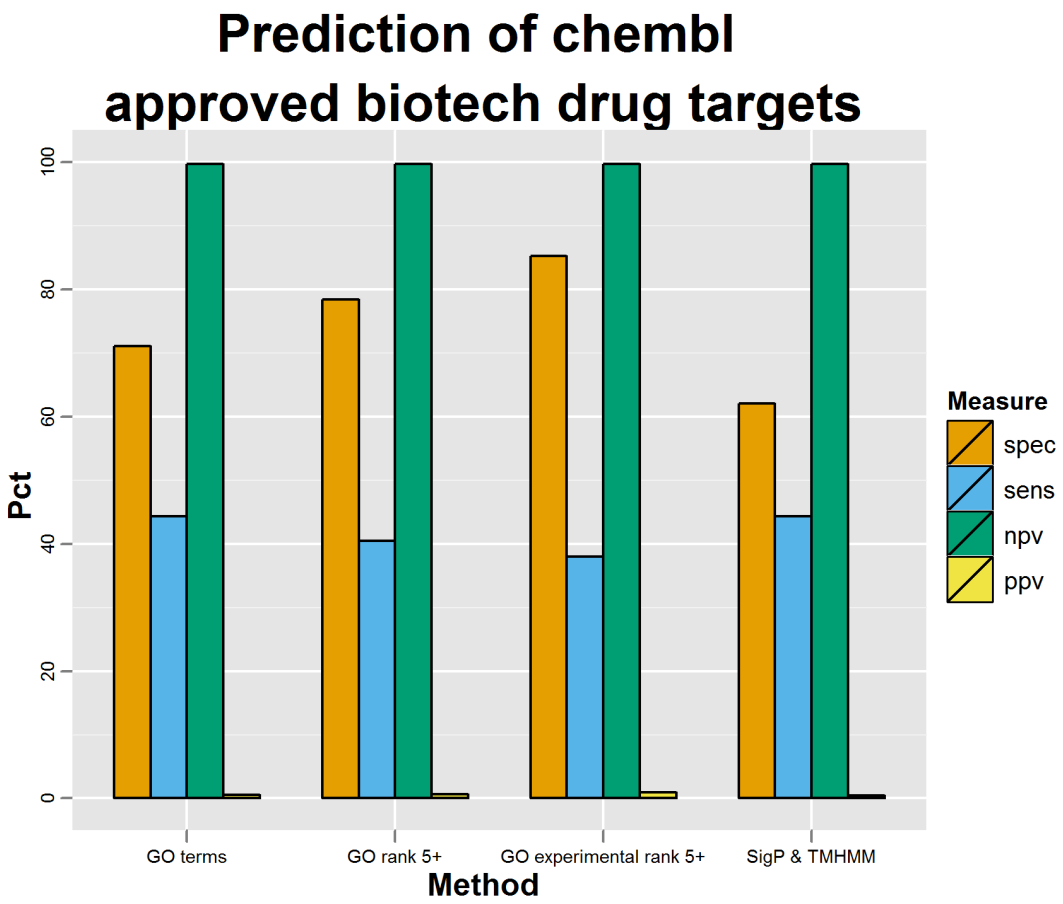

Figure 4.19: **Evaluation of each method when predicting ChEMBL phase IV biotechnology targets.**

Orange bars (spec) show specificity; blue (sens), sensitivity; green (npv), negative predictive value; and yellow (ppv), positive predictive value.

The highest specificity (85%) was achieved when genes associated with the extracellular or plasma membrane GO terms were filtered to leave only those with experimental evidence and a ranking of above five. The SignalP and TMHMM method had the lowest specificity at 62%, as shown in Figure 4.19.

Both the SigP and TMHMM and GO terms methods had the highest sensitivity at 44.3% each as they both identified 35 biopharmable genes correctly. The least sensitive method, the ranked experimental GO terms, only correctly predicted 40% of the biopharmable ChEMBL targets.

All methods had a negative predictive value of around 99.7%. The best positive predictive value was seen using the ranked experimental GO terms, however this was only 1%, indicating these methods correctly predicted a known target a very small percentage of the time due to the very small dataset.

4.2.5 Sensitivity and Specificity of methods predicting DrugBank biotechnology targets

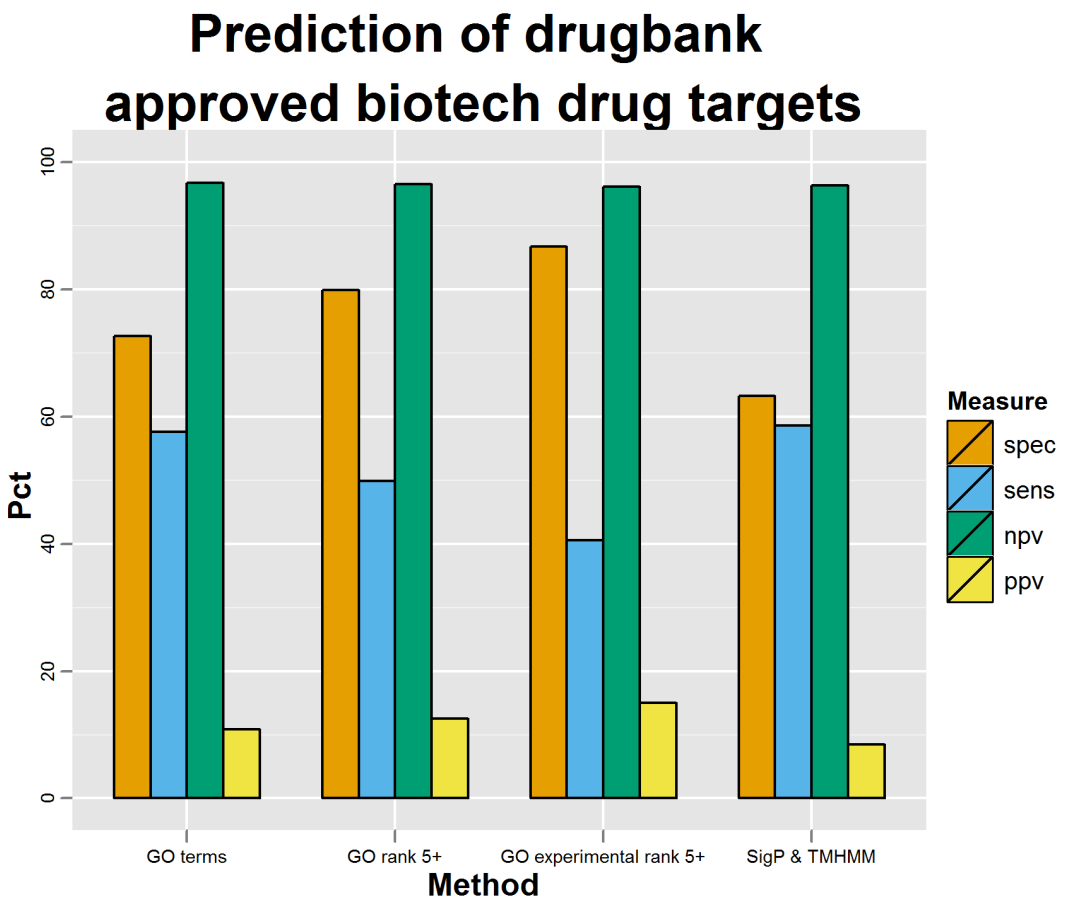

Figure 4.20: **Evaluation of each method when predicting DrugBank approved biotechnology targets.** Orange bars (spec) show specificity; blue (sens), sensitivity; green (ppv), positive predictive value; and yellow (npv), negative predictive value.

In Figure 4.20, as the specificity of a method increases the sensitivity decreases. SignalP and TMHMM is the most sensitive, correctly identifying 59% of available targets. Compared to the SigP/TMHMM method, genes identified via GO terms seem to show increase in specificity (73% versus 63%) without much of a decrease in sensitivity (58% versus 59%).

The most specific method was the GO terms filtered by experimental evidence code and either medium or high confidence that the term indicates an accessible location, showing 87% specificity. However, sensitivity for this method drops to 41%, showing it misses a large number of known biopharmable targets.

4.2.6 Receiver operating characteristic curves for predicting targets of biotechnology

The ChEMBL predictions seen in Figure 4.19 and Figure 4.21 are based on a very small dataset. As there were so few true positives it is hard to evaluate the true power of these methods to predict biopharmable targets. Nevertheless, the method with the best predictive power was the GO term query with experimental evidence which ranked five or higher for accessibility, with an AUC of 0.62.

When predicting the more numerous DrugBank targets (Figure 4.22), the method using all GO term results provided the most predictive power, with an AUC of 0.65. In contrast to predicting ChEMBL targets, the predictive power of the GO terms decreased as they were filtered by evidence code and confidence the associated term indicated accessibility.

For predicting ChEMBL biotech targets the SignalP and TMHMM method showed little to no predictive power, with an AUC of only 0.53. It performed slightly better when predicting DrugBank targets, with an AUC of 0.6.

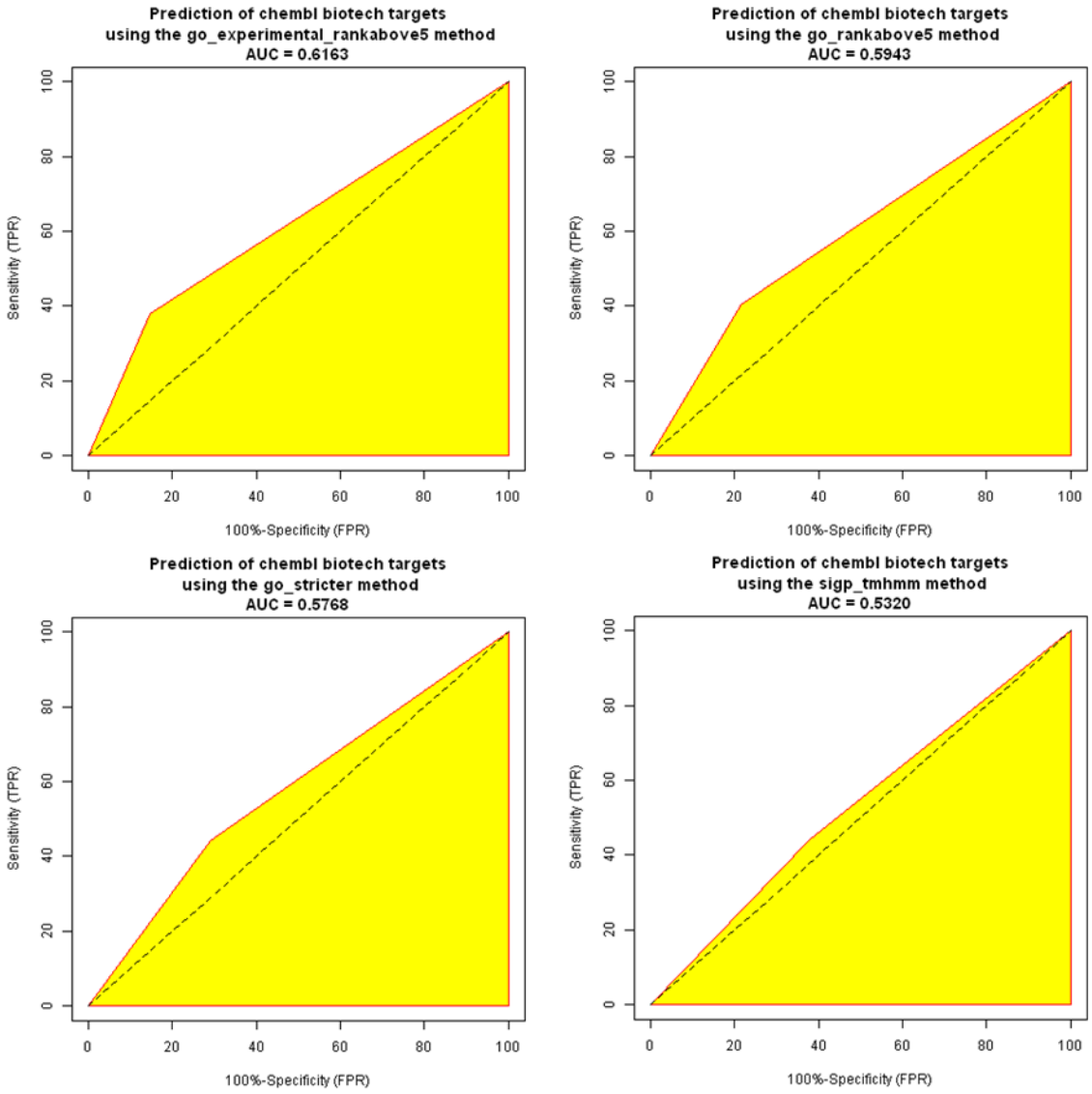

Figure 4.21: **ROC curves showing the predictive power of each method for the ChEMBL targets.** AUC, area under curve; TPR, true positive rate; and FPR, false positive rate. The dotted line indicates an AUC of 0.5 which is the predictive power which could be observed by chance.

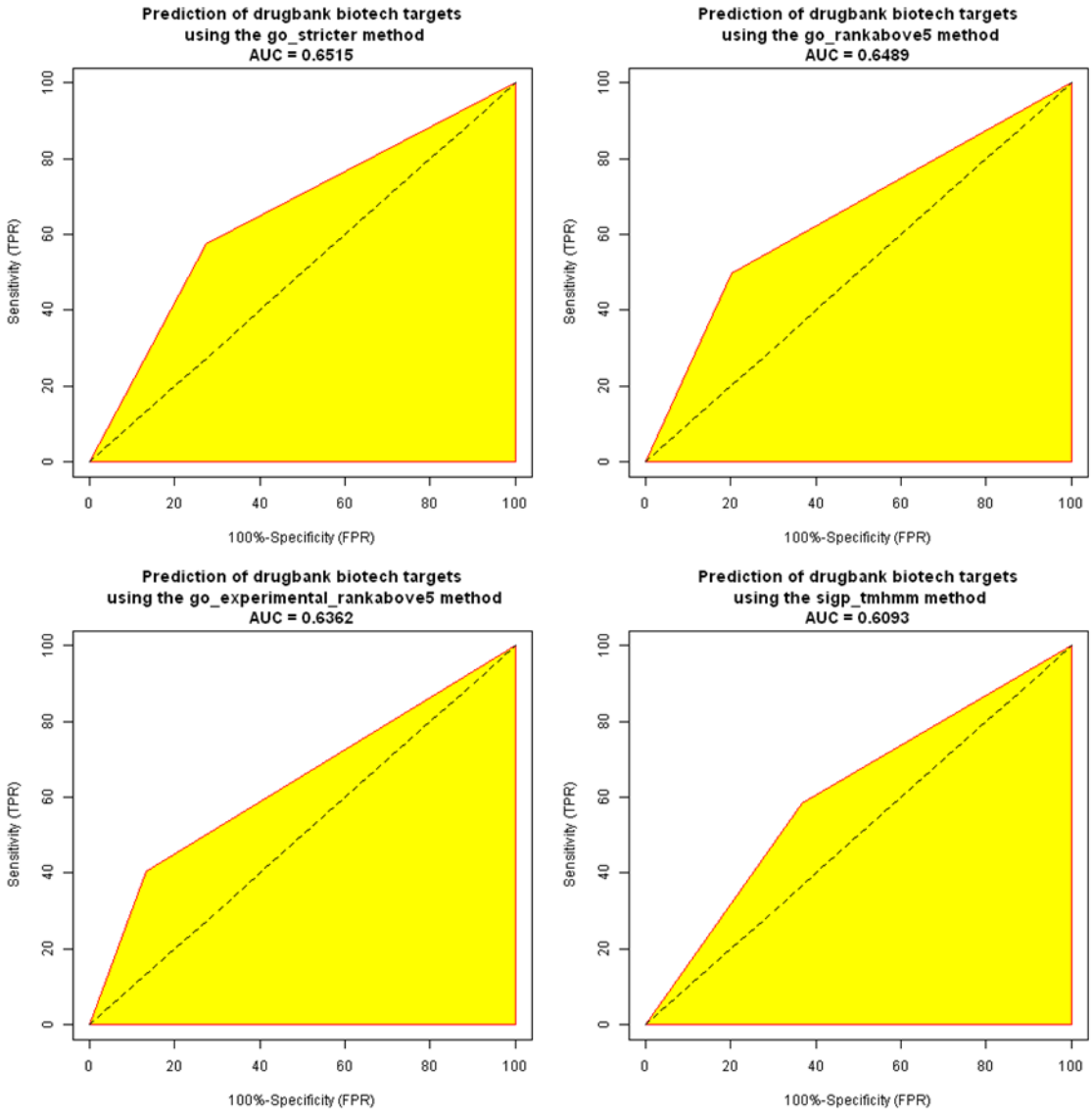

Figure 4.22: **ROC curves showing the predictive power of each method for the DrugBank targets.** AUC, area under curve; TPR, true positive rate; and FPR, false positive rate. The dotted line indicates an AUC of 0.5 which is the predictive power which could be observed by chance.

4.2.7 Biopharmable Predictive Power

The most specific method for predicting both ChEMBL phase IV biotechnology targets and DrugBank approved biotechnology targets, at 85% and 87% respectively, was the GO terms which were filtered for only those with an experimental evidence code and the child term had medium or high likelihood of indicating an accessible location.

Just using the parent GO terms without filtering and the SignalP and TMHMM methods gave the most sensitive predictions, both predicting 44.3% of known biopharmable ChEMBL targets and each method predicting 58% and 59% of known DrugBank targets respectively. However, the SignalP andTMHMM method had lower specificity

than the GO terms (73% versus 63%) when predicting known biopharmable DrugBank targets.

From the ROC curves, when predicting ChEMBL targets the GO terms filtered by experimental evidence and ranking five or above showed the best predictive power with a score of 0.616 indicating it is a poor prediction. Every other method scored under 0.6, making them almost unusable for predicting ChEMBL targets, with SigP/TMHMM scoring only 0.532, just above that observed through chance.

However, every method scored above 0.6 when predicting DrugBank targets, with GO terms seen as the best method at 0.652, which is still a poor predictor but more useful than those predicting ChEMBL biotechnology targets.

4.2.8 Predicted Biopharmable Target Classes

From an identified 807 known biopharmable genes which show activity with a phase IV drug, 701 genes (86.9%) had their ChEMBL major target class correctly predicted through regex capture, 90 (11.2%) were incorrectly classified and 16 (2%) remained unclassified. Figure 4.23 shows the predicted target classes of all genes predicted as biopharmable using all four methods, taking into account genes returned with more than one description resulting in a different classification, but genes with only one associated description (therefore, only one associated class) will be represented once.

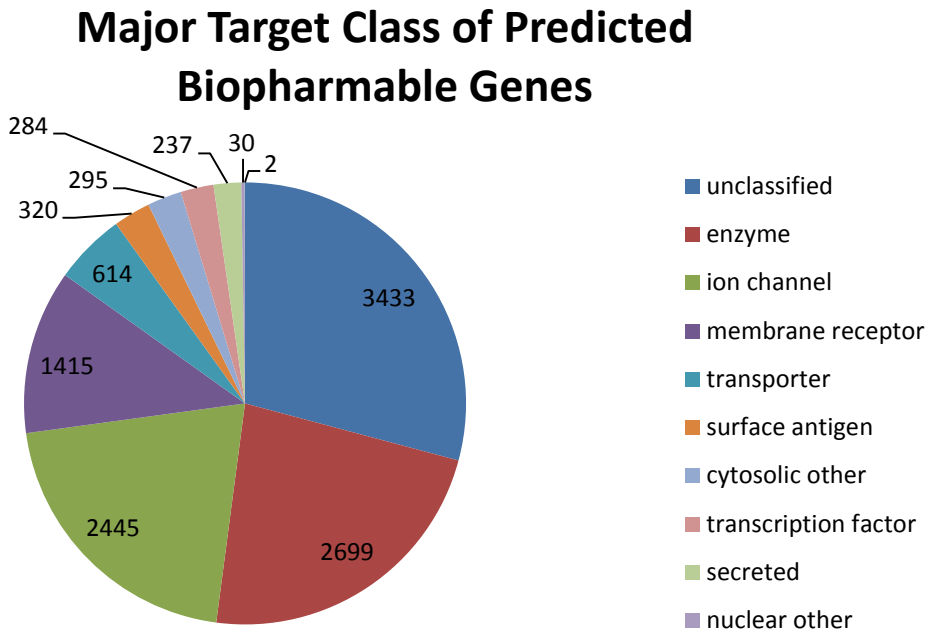

Figure 4.23: **The predicted target class of all genes predicted to be biopharmable using all prediction methods.**

The total of 11,774 takes into account genes with multiple descriptions resulting in different target classes, the total number of unique genes is 9,660.

The majority of predicted biopharmable genes were unclassified (29.2%), with enzymes the most abundant class at 22.9%. Channels the second best represented at 20.8% and receptors were found to make 12% of the predicted biopharmable genome. Transporters represent 5.2% and other classes, such as surface antigens (2.7%), made up less than 3% of the estimate each.

The identified major subclasses of each gene can be seen in Table 4.2. As with Figure 4.23, if a gene was returned with more than one description which resulted in it being assigned more than one subclass it will be represented more than once. Although specific subclasses proved difficult to determine through programmatic capturing of keywords, the majority of enzymes identified were proteases (447 genes), the majority of channels appear to be ligand gated (204) and most of the identified receptors were olfactory or taste GPCRs (492).

Table 4.2: **The predicted target class and subclass of all genes predicted as biopharmable.** The total of 12,267 takes into account genes with multiple descriptions resulting in different target classes or subclasses, the total number of unique genes is 9,660. Seven transmembrane (7TM) refers to members of the GPCR family, 7TM1 for class A GPCRs, 7TM2 for class B and 7TM3 for class C. Other GPCRs include classes such as frizzled and smoothened.

Predicted Target Class / Subclass	Number of Genes
Unclassified	**3433**
Enzyme	**2757**
N/A	1565
Protease	447
Kinase	373
Phosphatase	149
Reductase	99
P450	95
Phosphodiesterase	27
Aminoacyltransferase	2
Ion Channel	**2479**
N/A	2103
Ligand Gated	204
Voltage Gated	149
Ryanodine Receptor	23
Membrane Receptor	**1816**
N/A	602
7TM1 Olfactory/Taste	492
7TM1	301
Other GPCR	228
Nuclear	65
7TM2	64
Integrin	38
7TM3	26
Transporter	**614**
Surface Antigen	**320**
Cytosolic Other	**295**
Transcription Factor	**284**
Secreted	**237**
Nuclear Other	**30**
Structural	**2**
Grand Total	**12267**

4.3 Results Summary

In total, 12,577 genes were predicted by the seven different druggable genome prediction methods. If only the genes predicted by four or more of the methods are taken as the estimate, it stands at 5,721 genes. In the Hopkins and Groom replication, the major target class is now GPCRs instead of protein kinases and the estimate size has reduced overall from 3,051 genes in 2002 to 2,779 in 2012.

Druggable predictions were reasonably specific, with the Hopkins and Groom InterPro domains which correctly predicting 89.2% and 89.7% of the protein coding genome when testing against ChEMBL and DrugBank targets respectively. Similar specificity was observed using PDB confirmed ChEMBL target associated Pfam domains, predicting around 79% correctly in each case. It is also possible to be very sensitive, with the DrugBank target associated InterPro domains method correctly predicting 89% of ChEMBL targets and 99% of DrugBank targets.

However, since no method was both very sensitive and very specific, the appropriate druggable target prediction method to use should be based on whether a smaller, higher confidence prediction is required or a larger prediction which includes as many targets as possible is required. The use of DrugBank InterPro domains excludes the fewest potential targets, but predicts over half of the protein coding genome. Hopkins and Groom InterPro domains or ChEMBL PDB confirmed Pfam domains create the smallest, most conservative estimates, but also show the highest numbers of false negatives.

On the assumption that the extracted ChEMBL targets represent all real targets of approved drugs, the ChEMBL Pfam domains prediction may offer a compromise with 80.9% sensitivity and 78.3% specificity. As it had an area under curve of 0.796 this puts it on the boundary between a fair and good test and since the aim is to predict unexploited drug targets some “false positives” are required.

The four biopharmable genome prediction methods estimated the druggable genome at 9,660 genes or, when taking only genes predicted by three or more methods, 5,229. The largest target class were unclassified genes, probably due to the regex capturing typical compound target classes or possibly due to BioMart returning genes with no description. The second largest class was the enzymes.

The most sensitive (58.6%) method was when predicting the known DrugBank targets using the SignalP and TMHMM method. Best specificity (86.7%) was observed when predicting DrugBank known targets using GO terms supported by experimental evidence and a rank above five. The best predictive power, with an AUC of 0.652, came from using GO terms which indicate an extracellular or plasma membrane bound location to predict DrugBank known targets.

4.4 Prediction Pipeline

A collection of easily updated scripts can be found on the attached DVD, along with complete lists of all of the genes predicted by each method, a drug target class predicted through regular expressions, a ChEMBL target class (if known) and whether this gene is listed as an approved target in ChEMBL or DrugBank. A fully automated prediction process can be launched using the pipeline script. All of which were generated through the course of this project.

The druggable predictions produced using the Hopkins and ChEMBL methods have been used at GlaxoSmithKline to annotate the small molecule tractability of genes and create a list of tractable genes to form a focussed screening set which could be run through an animal model.

Extensive documentation for these scripts was also produced in the form of PerlDocs, Wiki pages and README files, which are housed on GSKs internal intranet system.

5 Discussion

Knowledge of which human proteins are likely to be druggable and/or biopharmable is important in the development of new therapeutics. These estimates highlight the genes and proteins which may be targeted by pharmaceutical and biotechnology companies exclusively and areas may be a future focus of both. Proteins of interest in disease states can be checked against these target lists to gage whether they are likely to be successful drug targets, potentially cutting down expenditure on assaying compounds against targets unlikely to bind them.

Additionally, known targets (and proteins similar to them) already have compounds they are known to bind which provide a starting point for new compound assays, meaning there may be less expenditure required to reach an optimum lead. Since they are less of a risk than novel targets, they are less likely to cause attrition at later stage. The use of open source and publicly available data should allow academics to assess whether proteins they have identified as potential targets are likely to be drug binding, aiding drug target discovery outside of industry.

5.1 Recommendations

The best estimate of the druggable genome depends on its intended use. If a conservative, high confidence dataset with as few false positives as possible is required, the Hopkins and Groom compiled druggable InterPro domains offer the smallest druggable estimate. The removal of known olfactory or taste receptors and targets known to be toxic would reduce the number of predicted proteins unlikely to be drug target in this prediction. Additionally, the unpublished druggable Pfam domains used by Russ and Lampel in 2005 may also be of interest for this purpose, but our inability to obtain this data prevented testing of their utility.

But the Hopkins and Russ domains have been manually curated, requiring a high level of validation and effort to update. The most conservative druggable genome estimate using an automatically updating method came from using the Pfam domains of ChEMBL associated with a ligand in PDB. Although 2,181 more genes were predicted than the Hopkins and Groom method there were 39 less known targets from ChEMBL missed and 80 from DrugBank.

If, however, a large, lower confidence dataset is required, excluding as few potential druggable targets as possible, the DrugBank InterPro method only misses 84 ChEMBL targets (127 less than the Hopkins and Groom method) and 24 DrugBank targets. It should be noted that it also predicts a very high proportion of the genome as druggable, only excluding 43% of the protein coding human genes from this list of potential targets.

A compromise may exist in using a combination of these prediction techniques. A high percentage of ChEMBL known targets (80.2%) are predicted by four or more of the prediction methods whilst maintaining a specificity of 75.2%. In both regards, however, the ChEMBL target Pfam domains outperform this consensus method with a sensitivity and specificity of 80.9% and 78.3% respectively. Since it also offers good best predictive power with an AUC of 0.796, the use of Pfam domains from ChEMBL targets may offer the best automated solution.

The best biopharmable prediction is hard to determine. No method managed to predict even half of the known ChEMBL targets and the most successful prediction of the DrugBank data captured 58.6% of the targets but also only 63% specificity. Using a consensus of targets predicted by three or more prediction techniques resulted in a slightly greater level of specificity (80.9%) than all other methods, expect the GO term method filtered by experimental evidence and confidence that it indicates an accessible location. However, the specificity drops, predicting less than half of known DrugBank targets (48.5%). Problems in predicting extracellular proteins may stem from difficulties in viewing them at an intact molecular level, since they are difficult to capture using crystallography or NMR (Vakonakis and Campbell, 2007). Signal peptides and transmembrane helices are hard to detect and distinguish, particularly for eukaryotes. Both areas are hydrophobic, with transmembrane helices having typically longer hydrophobic regions and no cleavage site, but the cleavage site pattern is insufficient to distinguish the two types of sequence (Petersen et al., 2011).

From these tests, determining the disease relevance of proteins should be the primary approach for the selection of targets for biotechnology since antibodies can be developed against many targets, whereas compound libraries already contain the drugs and require a target. Known receptors, ion channels, antigens and other cell surface protein classes, as well as components of the extracellular matrix, can be considered accessible to modulation, though transmembrane and secreted proteins appear hard to predict.

Figure 4.13 shows that using Hopkins and Groom druggable InterPro domains from 2002 now produces an estimate which is 272 genes smaller. Some of this is due to this work's removal of the supplied B30.2/SPRY domain, which does not appear to be druggable, and replacement with Vitamin K-dependent gamma-carboxylase, with which it had apparent false matches. A domain which once represented 2% of the druggable genome is now represented by only a single gene.

Also, work on different target classes since 2002 could have resulted in better annotation of their family and domains, further reducing the estimate size. For example, in 2003, Fredriksson *et al*. identified more than 800 human GPCR sequences and analysed 342 unique functional non olfactory human GPCR sequences. Five main families, glutamate, rhodopsin, adhesion, frizzled and secretin, were identified

(Fredriksson et al., 2003). Similarly, in 2008, the determined crystal structure of GPCR opsin was determined, supplying information into GPCR ligand binding and activation (Park et al., 2008).

Proteins which were considered to contain a given domain may have been reconsidered due to updates in the InterPro consortium members' alignment techniques, for example, higher family specific inclusion thesholds in Pfam. Additionally 17 of the original InterPro domains had been removed, reducing estimate size, or replaced, possibly with a narrower domain definition. The features of other domains may have also been reclassified, resulting in a more specific signature which would be present in fewer genes.

Improvements to the estimates produced using the Hopkins and Groom technique could come from manually updating the existing domains with the advanced knowledge available 10 years later. For example, any domains which bind newer drugs which were not included in the original list should be included. On the other hand, Pfam domains corresponding to these InterPro domains have been shown to produce more conservative estimates, therefore obtaining the Russ and Lampel list and making similar updates would likely produce an even more conservative druggable estimate than the smallest observed in this thesis. Estimates produced using InterPro domains from known targets could be improved by only including entries described as active sites or binding sites on the InterPro website, and it is possible this filtering process could be automated.

5.2 Comparisons to Previous Work

The number of approved human drug targets has varied between previous work on the druggable genome. In 2002, Hopkins and Groom estimated the number of targets at 120 proteins (Hopkins and Groom, 2002) meaning that the InterPro domains used in repeating this analysis have been extracted from 399 rule of five compliant compound binding proteins. In contrast, InterPro domains used in the ChEMBL InterPro and DrugBank InterPro predictions came from 1,122 and 2,649 proteins and 932 and 2,138 InterPro domains respectively.

Even the smaller ChEMBL target set contains over double the number of proteins as the original Hopkins and Groom set, explaining why so many more InterPro domains were identified (932 versus 144). Also the Hopkins and Groom set only contains druggable domains from targets, not all associated InterPro domains. Additionally, not all InterPro domains from the Hopkins and Groom set were from approved drug targets, explaining why there was not complete consensus observed between the two sets. The identified DrugBank targets contained 2,138 InterPro domains compared to the total of 6,816

identified in the human protein coding genome explaining how so much of the human genome was predicted as druggable. As InterPro predicts the occurrence of functional domains, repeats and important sites, not all domains represent an active site.

Estimates of the number of approved drug targets had doubled by 2006, with Imming *et al.* identifying 218 and Overington *et al.* finding 266 (Imming et al., 2006, Overington et al., 2006). Through manual curation of the DrugBank database in 2011, Rask-Andersen *et al*. identified 435 therapeutically relevant human protein targets. Ensembl genes and proteins associated with UniProt identifiers of targets in ChEMBL with significant activity with a phase four drug gave 1,122 target proteins (763 genes). This is more than expected when compared to previous work (which identified between 218 and 435 targets), which could indicate either more proteins assayed against phase four drugs show activity than expected or that the significance cut-off used may have been too low.

Similarly, DrugBank listed 2,649 proteins (1,870 genes) as the target of an approved drug. As nontherapeutic targets were not removed, this could explain the higher number of targets provided, as Rask-Andersen *et al.* manually removed over half of targets from DrugBank using current medical literature (Rask-Andersen et al., 2011). Therefore, the targets of nutritional supplements like *tetrahydrofolic acid* would be removed. Additional annotation in the DrugBank database to allow identification of therapeutic drugs would help to overcome this issue, or perhaps another database such as Therapeutic Target Database would be a better choice for this purpose.

A comparison of druggable genome estimates produced using various techniques is shown in Figure 5.1 in the context of the perceived genomic space. The ChEMBL Pfam estimate is shown to predict 23.8% of genes, over predicting compared to Russ and Lampel's methods by around 14%. Generally estimates have reduced over time to reflect the corresponding reduction in protein coding genes.

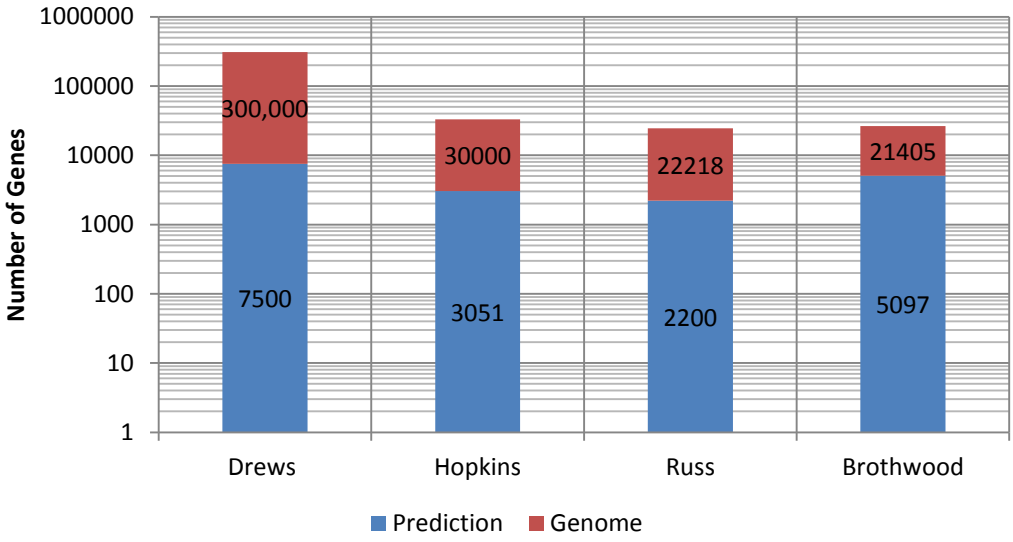

Figure 5.1: **Comparing the predicted druggable genome against the perceived human genome, using extracted ChEMBL Pfam domains as the Brothwood estimate.**

In line with the work by Russ and Lampel in 2005 using Pfam domains equivalent to the InterPro domains identified by Hopkins and Groom, enzymes are the largest target class, with kinases the largest class within this. The class "cytosolic other" which includes Bcl-2 and nuclear factors, has now become a major class not considered by the Russ paper. Rhodopsin-like GPCRs remain a major target class, though the target prediction here captured 492 sensory receptors compared to ~400 from Russ, though this may just be due to the larger prediction.

However, the known 118 proteins (79 genes) biotechnology targets from ChEMBL is much closer to the 76 identified by Overington *et al.* in 2006, suggesting this dataset has a different definition of a biologic than the 1,665 proteins (1,171 genes) returned from DrugBank. The ChEMBL query appears to only return targets of monoclonal antibodies, whereas the DrugBank query returns the targets of other biological therapeutics such as the recombinant human erythropoietin *epoetin alfa*. This treatment for anaemia increases haemoglobin by stimulating erythropoiesis (Littlewood et al., 2001). It also returns nontherapeutic biologics which have approved use for other purposes, such as *secretin* which is used to diagnose exocrine pancreatic function and targets the secretin receptor (Conwell et al., 2003).

5.3 Successfully Predicted Examples

All seven druggable genome prediction methods predicted muscarinic acetylcholine receptor M3 (UniProt identifier: P20309), a type 1 GPCR encoded by the human gene CHRM3, as chemically tractable. On the 23rd of July 2012 the FDA approved the rule of five compliant compound *aclidinum bromide* (*Tudorza Pressair*) as long term treatment of chronic obstructive pulmonary disease. Inhibition of this receptor decreases intracellular calcium levels and causes smooth muscle relaxation in the walls of the bronchioles, leading to bronchodilation (Kruse et al., 2012). Since the Hopkins and Groom InterPro domains were published in 2002, but have been used to predict a target approved in 2012, the use of protein domain druggability prediction techniques appears justified.

Similarly, all seven methods predicted another GPCR, β3-adrenergic receptor (UniProt identifier: P13945) as a possible drug target. A novel, first-in-class agonist, *mirabegron* (*Myrbetriq*) was approved for the treatment of overactive bladder on the 28th of June 2012. It increases bladder capacity by relaxing the detrusor smooth muscle (Tyagi and Tyagi, 2010) in a mechanism similar to aclidinum bromide, showing that all seven prediction methods have the potential to predict novel, previously unexploited drug targets, particularly of the GPCR class.

Another target of a first-in-class drug, *Ivacaftor* (*Kalydeco*), was predicted, once again, by all seven methods. It was approved as a treatment for of a rare form of cystic fibrosis on the 31st of January, 2012. The target, CFTR protein (UniProt identifier: P13569), is an ABC-class chloride-ion transporter (Yu et al., 2012) proving that all seven druggable prediction methods are capable of successfully predicting new drug targets of different classes, not just novel GPCRs. These three examples show the developed pipeline is able to predict some of the most recently approved drug targets.

Although the biopharmable genome predictions did miss a high percentage of known targets, some novel targets were successfully identified. Epidermal Growth Factor Receptor 2 (ERBB2, Uniprot identifier: P04626) is a cell membrane receptor over expressed in breast cancer, shown in Figure 5.2. It was identified by all four biopharmable prediction methods and is associated with the GO term indicating a basolateral plasma membrane location, supported by evidence from a direct assay, and scored an overall biopharmable rank of five out of six. ERBB2 is targeted by the monoclonal antibodies *Pertuzumab* (recently approved by the FDA on the 8th of June 2012) and *Trastuzumab* in the treatment of treatment of late stage metastatic breast cancer (Cho et al., 2003).

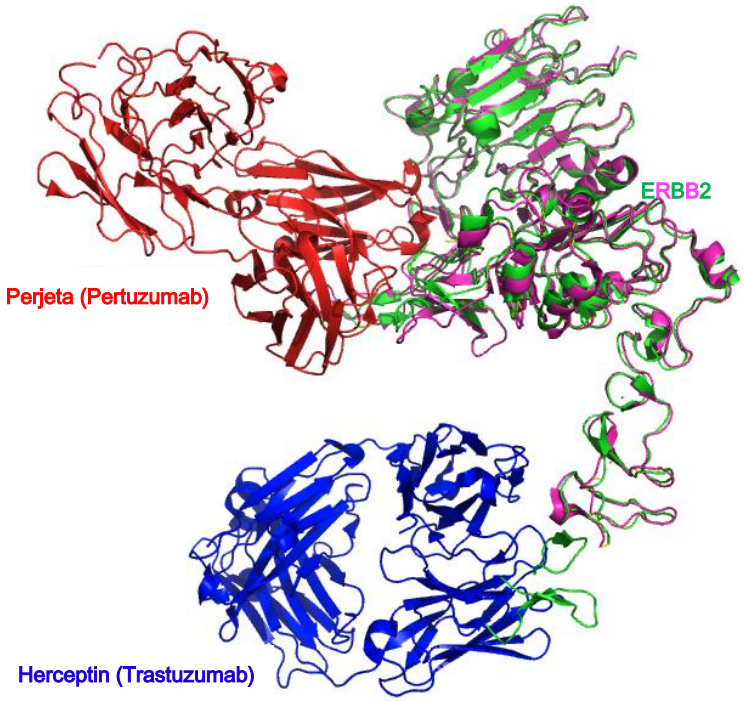

Figure 5.2: **Binding of approved monoclonal antibodies (Pertuzumab and Trastuzumab) to identified therapeutic breast cancer target Epidermal Growth Factor Receptor 2.**
Image from PDB (Berman et al., 2000, PDB:1n8z)

Another recently approved biologic, *Ipilimumab* (*Yervoy*), was approved for the treatment of malignant melanoma on the 25th of March 2011. Its target, cytotoxic T lymphocyte antigen 4 (CTLA-4, UniProt identifier: P16410), was predicted by all four biopharmable prediction methods. CTLA-4 naturally restricts the antitumour immune response so binding a monoclonal antibody to it helps to overcome CTLA-4–mediated T-cell suppression (Weber, 2007). These examples show the biopharmable prediction methods are capable of predicting some of the most recently approved biotechnology targets.

5.4 Caveats

One of the main caveats to the approach of capturing every InterPro or Pfam domain from each known druggable target is domains will be captured which do not bind the drug, resulting in false positives which do not contain the drug binding domain, illustrated in Figure 5.3.

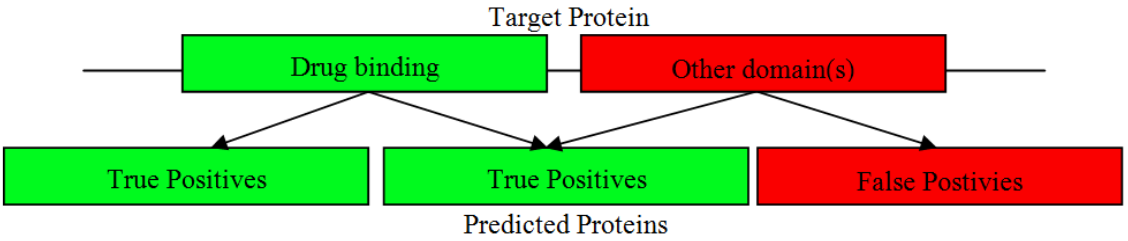

Figure 5.3: **An example target protein containing two domains, one which is drug binding (green) and one which does not bind a drug (red). Proteins predicted which contain only the red domain will be false negatives with no indication of drug binding function.**

The use of InterPro domains resulted in a far larger number of druggable target predictions than Pfam domains as InterPro has broader domain definitions and, therefore, coverage. Since each single InterPro entry unifies signatures from at least one of its member databases, for example it could bring together domains from Pfam, PRINTS, PROSITE, SMART, ProDom, PIRSF, SUPERFAMILY, PANTHER, CATH-Gene3D, TIGRFAMs and HAMAP. Since each source has its own methodology of signature production (Consortium et al., 2002, Hunter et al., 2012) each InterPro definition will be present in more proteins than each Pfam domain. Pfam families are identified from constructing a hidden Markov model using multiple sequence alignments and searching this against the UniProtKB sequence database, including only sequence regions which score above a family-specific cut off value (Punta et al., 2012). This highly curated, conserved domain approach differs from classifications by ProDom, which involves automated clustering of protein segments using PSI-BLAST (Servant et al., 2002). False negatives could have occurred either because an approved target did not have an associated InterPro or Pfam domain or this domain was not linked to or returned by the Ensembl BioMart query.

PDB was used in an attempt to reduce the number of domains which are not directly involved in drug binding domains used in the druggable predictions. All domains from within an approved drug target were tested to see whether there is any experimental evidence of binding, and therefore, the ability to bind the drug. However, if the structure has not been determined and listed in PDB the domain will be excluded. The caveats of this approach are detailed in Figure 5.4, though this does reduce the number of false positives by removing 77 Pfam domains derived from ChEMBL targets (137 less genes predicted) and 208 Pfam domains from DrugBank targets (292 less genes predicted).

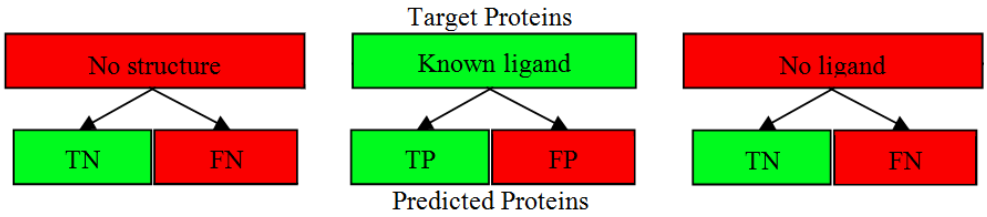

Figure 5.4: **By filtering to include only Pfam domains with a known ligand in PDB, domains with no determined structure and domains with no known ligand will be excluded.**
Druggable domains with no determined structure will result in false negatives, domains with a ligand which is not a drug will produce false positives (although it can be argued this are likely to be druggable) and if a structure has a ligand which has not been determined this will result in a false negative. TP, true positive; FP, false positive; TN, true negative; and FN, false negative.

Additionally, not all domains appear to be captured. The 42 InterPro domains provided in the Hopkins and Groom set which were not observed in DrugBank targets mostly consisted of domains which do not appear to targeted by an approved drug, such as the Serpin protein domain family (protease inhibitors). Some others do appear to be druggable, such as the catalytic domain of 3'5'-cyclic nucleotide phosphodiesterase.

The first group could be due to misclassification of the binding domain by Hopkins and Groom or a change in domain definition by InterPro, as observed when B30.2/SPRY domain was falsely matched against a provided γ-carboxylase domain (IPR001870). The second group likely indicates either that the target is not listed DrugBank, a failure to capture all target UniProt identifiers from the DrugBank database or of BioMart to return all InterPro domains within each result.

Additionally, there are 30 genes identified by the Hopkins and Groom method not identified by any other. These include plasma alpha-l-fucosidase from the InterPro domain for glycoside hydrolase family 29. Similarly, electroneutral sodium bicarbonate exchanger is returned from the sodium bicarbonate transporter domain. These domains show evidence of binding a rule of five compliant compound, but do not appear to be found in an ChEMBL phase four or DrugBank approved target, showing they may only bind an experimental compound.

5.5 Future Work

An advised improvement for the future would to programmatically work with protein identifiers, rather than their encoding gene identifiers, since these are often the actual drug targets. Instead of using Ensembl gene identifies, working with the Ensembl protein identifiers or using the UniProt BioMart and UniProt identifiers as the primary reference would make aligning data with ChEMBL and DrugBank data much easier. In addition, by using UniProt accessions problems experienced due to inconsistent linking between of Ensembl identifiers to UniProt identifiers would be alleviated. Since the actual target protein would be identified directly, this provides a clearly starting point for disease linking, for example using genome-wide association studies (GWAS) and assaying against potential drugs in pharmaceutical companies. Importantly, it would still be possible to draw comparisons against previous work since it is possible to obtain gene identifiers from the protein identifiers and it is a safer assumption that one protein is coded by one gene than that one gene codes one protein, which is not always the case.

Expansion to include rule of five compliant compound binding sites or to include data from other databases could also reveal more useful information. The BindingDB (Liu et al., 2007) is a highly curated database including binding affinity data for small molecules to proteins, focussing on quantitative data (such as Ki, Kd and IC50 measurements). It contains 5,583 protein targets compared to the 8,900 protein targets recorded in ChEMBL. Other databases include PubChem (Wang et al., 2012), containing data from ChEMBL and BindingDB amongst other sources, and Therapeutic Target Database (Zhu et al., 2012), housing 2,085 (364 successful) fully referenced drug targets, with additional relevant information such as protein function, sequence, 3D structure, therapeutic class and ligand binding properties (Nicola et al., 2012).

Using only data from the binding assays in ChEMBL rather than all assays, which include functional assays, could improve known target data quality. This is important since the number of known targets with significant activity appeared high compared to previous work. Additionally the activity level at which a target is considered significant could be adjusted higher, however this would exclude some targets of promiscuous drugs which may be therapeutically relevant but only display low levels of activity.

Additionally, the use of proprietary in house data from pharmaceutical companies would improve the quality of drug-target assay data and, therefore, the extrapolated predictions. OpenPHACTS could build upon the information provided by this thesis, incorporating information from other companies and sources to create a consensus of various predictions to create one, higher confidence list.

The use of homology would also be an interesting future step. Identifying proteins with significantly similar sequences to known targets may provide a better estimate, as would the use of druggable domains sequences from ChEMBL's DrugEBIlity database. Homology searches could be adjusted to create very conservative estimates, with only

targets very similar to the known examples or broader estimates allowing some less similar proteins to be predicted.

Identification of targets based on structural information would give a high confidence estimate, ensuring that not only is a binding domain present, it is correctly folded and in an accessible location on the protein. However, many drug targets are difficult to crystallise, for example those which are membrane bound rather than globular, and therefore do not have determined structures. But as the Structural Genomics Consortium (SGC) and other groups continue to determine proteins structures, the PDB could be provide an even more useful resource in the future. FTMAP, which identifies protein sites with high binding affinity, may also prove a useful tool for structural based prediction methods (Brenke 2009).

For biopharmable predictions, it is advisable to evaluate potential targets on their involvement in disease, for example through literature searches for strong or weak associations, using Disease Ontology (DO) terms or databases such as Online Mendelian Inheritance in Man (OMIM) or GWAS. The DO provides well-defined, standardised terms for genetic, environmental and infectious diseases so any gene associated to a DO term may be potentially of interest (Schriml et al., 2011). OMIM focuses on genetic disorders, containing over 12,000 genes and their association to a specific phenotype (OMIM, 2012) whereas GWAS focuses on the analysis of genetic differences between healthy individuals and those with specific illnesses, with information contained within the database of genotype and phenotype (dbGaP).

The use of other signal peptide predictors (such as SignalCF or TatP), secretome predictors (such as SecretomeP), transmembrane predictors (for example, HMMTOP, DAS or SOSUI) and subcellular location predictors (such as TargetP) should be explored.

Additional filtering of all produced datasets to remove targets known to be toxic, or those which are unlikely to be suitable targets, would also improve prediction quality. Targets known to cause a harmful effect when targeted can be obtained from databases such as the Toxin Toxin-Target Database (Lim et al., 2010) or SuperToxic (Schmidt et al., 2009) and, if any of these targets are predicted, they could be flagged as toxic or removed from the prediction entirely.

Replication of the existing prediction methods in model organisms or pathogens, through querying a different Ensembl genome, could provide potentially useful target information. The replication in model organisms would ensure any developed drug could be tested effectively and easily. Identified pathogen targets could also be tested through to ensure no close homologues existing in humans.

6 Conclusion

An automated pipeline was created to generate both a druggable and a biopharmable genome. The pipeline lends itself to be updated as new knowledge emerges, for example new druggable protein domains. This has already been used by GlaxoSmithKline and has created great interest in the IMI Openphacts project.

In conclusion, the biopharmable genome is very difficult to accurately predict. Characteristics which make a protein druggable are more clearly defined (there must be a pocket able to bind a compound like molecule) whereas those which allow a target to be modulated by biotechnology are more poorly characterised. The known characteristics, such as accessibility, are hard to predict as multiple secretion pathways exist and not every protein with a transmembrane region is exposed on the cell surface. None of the prediction models managed to capture a high percentage of the known targets even though the biotechnology targets provided by ChEMBL appear to be almost exclusively receptors. Although the target class predictions were not fully implemented for the biopharmable classes, receptors and ion channels appear to make up the majority of the results, implying the GO terms for plasma membrane and TMHMM methods are successfully predicting potential targets.

Estimates of the druggable genome have been more successful, capturing a much higher percentage of the known targets. As expected, the Hopkins and Groom InterPro domains produced an estimate in line with previous estimates, at of 2,779 genes and other predictions were at around 3,000 druggable genes. The use of every InterPro domain from known targets should be considered too broad, as over predicting to the point of including the majority of the genome is not a accurate estimate.

Therefore, to create a broad estimate which is still within reasonable boundaries, the Pfam domains with a confirmed ligand in PDB could be used. This estimate was greater than previously predicted estimates, at 4,960 genes, but missed far fewer known targets that the Hopkins and Groom InterPro domains. Since these methods cannot take into account the discovery of first in class targets, the actual size of the druggable genome is likely to be somewhere in between these two estimates at around 3,000-5,000 genes.

The fully automated pipeline developed as part of this thesis using open source and publicly accessible data is available for update and modification for any required future work.

7 Bibliography

ARMSTRONG, J. W. 1999. *A review of high-throughput screening approaches for drug discovery* [Online]. http://www.combichemistry.com/statdir/stat.php?id=pdf15: HTS Consulting Ltd. [Accessed 24/05/12 2012].

ASHBURN, T. T. & THOR, K. B. 2004. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov,* 3**,** 673-683.

BERG, J. M., TYMOCZKO, J. L. & STRYER, L. 2007. *Biochemistry,* New York, W.H. Freeman.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research,* 28**,** 235-242.

BIANCHI, M. T. 2010. Promiscuous modulation of ion channels by anti-psychotic and anti-dementia medications. *Medical Hypotheses,* 74**,** 297-300.

BICKERTON, G. R., PAOLINI, G. V., BESNARD, J., MURESAN, S. & HOPKINS, A. L. 2012. Quantifying the chemical beauty of drugs. *Nat Chem,* 4**,** 90-98.

BINNS, D., DIMMER, E., HUNTLEY, R., BARRELL, D., O'DONOVAN, C. & APWEILER, R. 2009. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics,* 25**,** 3045-3046.

BODE, G., CLAUSING, P., GERVAIS, F., LOEGSTED, J., LUFT, J., NOGUES, V. & SIMS, J. 2010. The utility of the minipig as an animal model in regulatory toxicology. *Journal of Pharmacological and Toxicological Methods,* 62**,** 196-220.

BONIN-DEBS, A. L., BOCHE, I., GILLE, H. & BRINKMANN, U. 2004. Development of secreted proteins as biotherapeutic agents. *Expert Opinion on Biological Therapy,* 4**,** 551-558.

BUCHAN, N. S., RAJPAL, D. K., WEBSTER, Y., ALATORRE, C., GUDIVADA, R. C., ZHENG, C., SANSEAU, P. & KOEHLER, J. 2011. The role of translational bioinformatics in drug discovery. *Drug Discovery Today,* 16**,** 426-434.

CHEN, H. & BOUTROS, P. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics,* 12**,** 35.

CHO, H.-S., MASON, K., RAMYAR, K. X., STANLEY, A. M., GABELLI, S. B., DENNEY, D. W. & LEAHY, D. J. 2003. Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. *Nature,* 421**,** 756-760.

CHOO, K., TAN, T. & RANGANATHAN, S. 2005. SPdb - a signal peptide database. *BMC Bioinformatics,* 6**,** 249.

CONSORTIUM, G. O. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research,* 32**,** D258-D261.

CONSORTIUM, I. H. G. S. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

CONSORTIUM, T. I., MULDER, N. J., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BINNS, D., BISWAS, M., BRADLEY, P., BORK, P., BUCHER, P., COPLEY, R., COURCELLE, E., DURBIN, R., FALQUET, L., FLEISCHMANN, W., GOUZY, J., GRIFFITH-JONES, S., HAFT, D., HERMJAKOB, H., HULO, N., KAHN, D., KANAPIN, A., KRESTYANINOVA, M., LOPEZ, R., LETUNIC, I., ORCHARD, S., PAGNI, M., PEYRUC, D., PONTING, C. P., SERVANT, F. & SIGRIST, C. J. A. 2002. InterPro: An integrated documentation resource for protein families, domains and functional sites. *Briefings in Bioinformatics,* 3**,** 225-235.

CONWELL, D. L., ZUCCARO JR, G., VARGO, J. J., TROLLI, P. A., VANLENTE, F., OBUCHOWSKI, N., DUMOT, J. A. & O'LAUGHLIN, C. 2003. An endoscopic pancreatic function test with synthetic porcine secretin for the evaluation of chronic abdominal pain and suspected chronic pancreatitis. *Gastrointestinal Endoscopy,* 57**,** 37-40.

CORBIN, J. D. & FRANCIS, S. H. 2011. Conformational conversion of PDE5 by incubation with sildenafil or metal ion is accompanied by stimulation of allosteric cGMP binding. *Cellular Signalling,* 23**,** 1578-1583.

DIMASI, J. A. & GRABOWSKI, H. G. 2007. The cost of biopharmaceutical R&D: is biotech different? *Managerial and Decision Economics,* 28**,** 469-479.

DREWS, J. 1996. Genomic sciences and the medicine of tomorrow. *Nature Biotechnology,* 14**,** 1516-&.

FDA. 2009. *Frequently Asked Questions About Therapeutic Biological Products* [Online]. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved

/ApprovalApplications/TherapeuticBiologicApplications/ucm113522.htm: U.S. Food and Drug Administration. [Accessed 15/05/12 2012].

FILMORE, D. 2004. It's a GPCR world. Modern drug discovery.

FREDRIKSSON, R., LAGERSTRÖM, M. C., LUNDIN, L. G. & SCHIÖTH, H. B. 2003. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Molecular Pharmacology,* 63**,** 1256-1272.

GAULTON, A., BELLIS, L. J., BENTO, A. P., CHAMBERS, J., DAVIES, M., HERSEY, A., LIGHT, Y., MCGLINCHEY, S., MICHALOVICH, D., AL-LAZIKANI, B. & OVERINGTON, J. P. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research,* 40**,** D1100-D1107.

GOLDMAN, M. 2012. The Innovative Medicines Initiative: A European Response to the Innovation Challenge. *Clin Pharmacol Ther,* 91**,** 418-425.

HOPKINS, A. L. & GROOM, C. R. 2002. The druggable genome. *Nature Reviews Drug Discovery,* 1**,** 727-730.

HUMAN GENOME SEQUENCING, C. 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-945.

HUNTER, S., JONES, P., MITCHELL, A., APWEILER, R., ATTWOOD, T. K., BATEMAN, A., BERNARD, T., BINNS, D., BORK, P., BURGE, S., DE CASTRO, E., COGGILL, P., CORBETT, M., DAS, U., DAUGHERTY, L., DUQUENNE, L., FINN, R. D., FRASER, M., GOUGH, J., HAFT, D., HULO, N., KAHN, D., KELLY, E., LETUNIC, I., LONSDALE, D., LOPEZ, R., MADERA, M., MASLEN, J., MCANULLA, C., MCDOWALL, J., MCMENAMIN, C., MI, H., MUTOWO-MUELLENET, P., MULDER, N., NATALE, D., ORENGO, C., PESSEAT, S., PUNTA, M., QUINN, A. F., RIVOIRE, C., SANGRADOR-VEGAS, A., SELENGUT, J. D., SIGRIST, C. J. A., SCHEREMETJEW, M., TATE, J., THIMMAJANARTHANAN, M., THOMAS, P. D., WU, C. H., YEATS, C. & YONG, S.-Y. 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research,* 40**,** D306-D312.

HUXLEY-JONES, J., FOORD, S. M. & BARNES, M. R. 2008. Drug discovery in the extracellular matrix. *Drug Discovery Today,* 13**,** 685-694.

HWANG, W. Y. K. & FOOTE, J. 2005. Immunogenicity of engineered antibodies. *Methods,* 36**,** 3-10.

IMMING, P., SINNING, C. & MEYER, A. 2006. Opinion - Drugs, their targets and the nature and number of drug targets. *Nature Reviews Drug Discovery,* 5**,** 821-834.

JOHNSON, L. N. 2009. Protein kinase inhibitors: contributions from structure to clinical compounds. *Quarterly Reviews of Biophysics,* 42**,** 1-40.

JOHNSON, R. J., WILLIAMS, J. M., SCHREIBER, B. M., ELFE, C. D., LENNON-HOPKINS, K. L., SKRZYPEK, M. S. & WHITE, R. D. 2005. Analysis of gene ontology features in microarray data using the Proteome BioKnowledge (R) Library. *In Silico Biology,* 5**,** 389-399.

KARLSSON, E. K. & LINDBLAD-TOH, K. 2008. Leader of the pack: gene mapping in dogs and other model organisms. *Nat Rev Genet,* 9**,** 713-725.

KINSELLA, R. J., KÄHÄRI, A., HAIDER, S., ZAMORA, J., PROCTOR, G., SPUDICH, G., ALMEIDA-KING, J., STAINES, D., DERWENT, P., KERHORNOU, A., KERSEY, P. & FLICEK, P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database,* 2011.

KIRKPATRICK, P., GRAHAM, J. & MUHSIN, M. 2004. Cetuximab. *Nat Rev Drug Discov,* 3**,** 549-550.

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. & WISHART, D. S. 2011. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research,* 39**,** D1035-D1041.

KOLB, P. & KLEBE, G. 2011. The Golden Age of GPCR Structural Biology: Any Impact on Drug Design? *Angewandte Chemie International Edition,* 50**,** 11573-11575.

KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E. L. L. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology,* 305**,** 567-580.

KRUSE, A. C., HU, J., PAN, A. C., ARLOW, D. H., ROSENBAUM, D. M., ROSEMOND, E., GREEN, H. F., LIU, T., CHAE, P. S., DROR, R. O., SHAW, D. E., WEIS, W. I., WESS, J. & KOBILKA, B. K. 2012. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature,* 482**,** 552-556.

KUSUMI, I., TAKAHASHI, Y., SUZUKI, K., KAMEDA, K. & KOYAMA, T. 2000. Differential effects of subchronic treatments with atypical antipsychotic drugs on dopamine D₂ and serotonin 5-HT_{2A} receptors in the rat brain. *Journal of Neural Transmission,* 107**,** 295-302.

LANDER, E. S. 2004. Eric S. Lander. *Nat Rev Drug Discov,* 3**,** 730-730.

LANDER, E. S. 2011. Initial impact of the sequencing of the human genome. *Nature,* 470**,** 187-197.

LEE, S. & WANG, J. 2009. Exploiting the promiscuity of imatinib. *Journal of Biology,* 8**,** 30.

LEESON, P. 2012. Drug discovery: Chemical beauty contest. *Nature,* 481**,** 455-456.

LIM, E., PON, A., DJOUMBOU, Y., KNOX, C., SHRIVASTAVA, S., GUO, A. C., NEVEU, V. & WISHART, D. S. 2010. T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Research,* 38**,** D781-D786.

LINDSAY, M. A. 2003. Target discovery. *Nat Rev Drug Discov.* England.

LIPINSKI, C. A., LOMBARDO, F., DOMINY, B. W. & FEENEY, P. J. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews,* 46**,** 3-26.

LITTLEWOOD, T. J., BAJETTA, E., NORTIER, J. W., VERCAMMEN, E. & RAPOPORT, B. 2001. Effects of epoetin alfa on hematologic parameters and quality of life in cancer patients receiving nonplatinum chemotherapy: results of a randomized, double-blind, placebo-controlled trial. *J Clin Oncol,* 19**,** 2865-74.

LIU, T., LIN, Y., WEN, X., JORISSEN, R. N. & GILSON, M. K. 2007. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research,* 35**,** D198-D201.

MALIK, N. N. 2009. Key issues in the pharmaceutical industry: consequences on R&D. *Expert Opinion on Drug Discovery,* 4**,** 15-19.

MILLER, D. H., WEBER, T., GROVE, R., WARDELL, C., HORRIGAN, J., GRAFF, O., ATKINSON, G., DUA, P., YOUSRY, T., MACMANUS, D. & MONTALBAN, X. 2012. Firategrast for relapsing remitting multiple sclerosis: a phase 2, randomised, double-blind, placebo-controlled trial. *The Lancet Neurology,* 11**,** 131-139.

MULLER, M. P., TOMLINSON, G., MARRIE, T. J., TANG, P., MCGEER, A., LOW, D. E., DETSKY, A. S. & GOLD, W. L. 2005. Can Routine Laboratory Tests Discriminate between Severe Acute Respiratory Syndrome and Other Causes of Community-Acquired Pneumonia? *Clinical Infectious Diseases,* 40**,** 1079-1086.

MUNOS, B. 2009. Lessons from 60 years of pharmaceutical innovation. *Nat Rev Drug Discov,* 8**,** 959-968.

MWV. 2012. *Medicines for Malaria Venture: Coartem® Dispersible* [Online]. http://www.mmv.org/achievements-challenges/achievements/coartem-d. [Accessed 25/05/12 2012].

NICOLA, G., LIU, T. & GILSON, M. K. 2012. Public Domain Databases for Medicinal Chemistry. *Journal of Medicinal Chemistry*.

NURSINGTIMES. 2007. *The administration of medicines* [Online]. http://www.nursingtimes.net/nursing-practice/clinical-specialisms/prescribing/the-administration-of-medicines/288560.article: NursingTimes. [Accessed 22/05/12 2012].

OMIM. 2012. *Online Mendelian Inheritance in Man* [Online]. http://omim.org/: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). [Accessed 19/08/12 2012].

OVERINGTON, J. P., AL-LAZIKANI, B. & HOPKINS, A. L. 2006. How many drug targets are there? *Nat Rev Drug Discov,* 5**,** 993-996.

PARK, J. H., SCHEERER, P., HOFMANN, K. P., CHOE, H. W. & ERNST, O. P. 2008. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature,* 454**,** 183-187.

PERTEA, M. & SALZBERG, S. 2010. Between a chicken and a grape: estimating the number of human genes. *Genome Biology,* 11**,** 206.

PETERSEN, T. N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Meth,* 8**,** 785-786.

PHRMA. 2007. *Drug Discovery and Development: Understanding the R&D Process* [Online]. http://www.phrma.org/sites/default/files/159/rd_brochure_022307.pdf: PhRMA. [Accessed 24/05/12 2012].

PHRMA. 2011. *Pharmaceutical Industry Profile 2011* [Online]. http://www.phrma.org/sites/default/files/159/phrma_profile_2011_final.pdf: Pharmaceutical Research and Manufacturers of America. [Accessed 16/05/12.

PRNEWSWIRE. 2011. *New Oral Biologics Delivery Company, Entrega, Announces Strategic Partnership with Pharma* [Online]. http://www.prnewswire.com/news-releases/new-oral-biologics-delivery-company-entrega-announces-strategic-partnership-with-pharma-113197424.html: PRNewswire. [Accessed 26/07/12 2012].

PRUDOVSKY, I., MANDINOVA, A., BAGALA, C., SOLDI, R., BELLUM, S., BATTELLI, C., GRAZIANI, I. & MACIAG, T. 2003. Chapter 327 - Nonclassical Pathways of Protein Export. *In:* EDWARD, A. D. (ed.) *Handbook of Cell Signaling.* Burlington: Academic Press.

PUNTA, M., COGGILL, P. C., EBERHARDT, R. Y., MISTRY, J., TATE, J., BOURSNELL, C., PANG, N., FORSLUND, K., CERIC, G., CLEMENTS, J., HEGER, A., HOLM, L., SONNHAMMER, E. L. L., EDDY, S. R., BATEMAN, A. & FINN, R. D. 2012. The Pfam protein families database. *Nucleic Acids Research,* 40**,** D290-D301.

RASK-ANDERSEN, M., ALMEN, M. S. & SCHIOTH, H. B. 2011. Trends in the exploitation of novel drug targets. *Nature Reviews Drug Discovery,* 10**,** 579-590.

RUSS, A. P. & LAMPEL, S. 2005. The druggable genome: an update. *Drug Discovery Today,* 10**,** 1607-1610.

SAKHARKAR, M. K. & SAKHARKAR, K. R. 2007. Targetability of human disease genes. *Current drug discovery technologies,* 4**,** 48-58.

SAMS-DODD, F. 2006. Drug discovery: selecting the optimal approach. *Drug Discovery Today,* 11**,** 465-472.

SCHMIDT, U., STRUCK, S., GRUENING, B., HOSSBACH, J., JAEGER, I. S., PAROL, R., LINDEQUIST, U., TEUSCHER, E. & PREISSNER, R. 2009. SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Research,* 37**,** D295-D299.

SCHRIML, L. M., ARZE, C., NADENDLA, S., CHANG, Y.-W. W., MAZAITIS, M., FELIX, V., FENG, G. & KIBBE, W. A. 2011. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*.

SERVANT, F., BRU, C., CARRÈRE, S., COURCELLE, E., GOUZY, J., PEYRUC, D. & KAHN, D. 2002. ProDom: Automated clustering of homologous domains. *Briefings in Bioinformatics,* 3**,** 246-251.

SWINNEY, D. C. & ANTHONY, J. 2011. How were new medicines discovered? *Nature Reviews Drug Discovery,* 10**,** 507-519.

TYAGI, P. & TYAGI, V. 2010. Mirabegron, a beta(3)-adrenoceptor agonist for the potential treatment of urinary frequency, urinary incontinence or urgency associated with overactive bladder. *IDrugs,* 13**,** 713-22.

VAKONAKIS, I. & CAMPBELL, I. D. 2007. Extracellular matrix: from atomic resolution to ultrastructure. *Curr Opin Cell Biol.* United States.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J., GABOR MIKLOS, G. L., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, J., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z., FRANCESCO, V. D., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W., GONG, F., GU, Z., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R.-R., KE, Z., KETCHUM, K. A., LAI, Z., LEI, Y., LI, Z., LI, J., LIANG, Y., LIN, X., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B., SUN, J., WANG, Z. Y., WANG, A., WANG, X., WANG, J., WEI, M.-H., WIDES, R., XIAO, C., YAN, C., et al. 2001. The Sequence of the Human Genome. *Science,* 291**,** 1304-1351.

WALSH, C. T. & SCHWARTZ-BLOOM, R. D. 2004. *Levine's Pharmacology: Drug Actions and Reactions,* New York, Taylor & Francis.

WANG, Y., XIAO, J., SUZEK, T. O., ZHANG, J., WANG, J., ZHOU, Z., HAN, L., KARAPETYAN, K., DRACHEVA, S., SHOEMAKER, B. A., BOLTON, E., GINDULYTE, A. & BRYANT, S. H. 2012. PubChem's BioAssay Database. *Nucleic Acids Research,* 40**,** D400-D412.

WEBER, J. 2007. Review: Anti–CTLA-4 Antibody Ipilimumab: Case Studies of Clinical Response and Immune-Related Adverse Events. *The Oncologist,* 12**,** 864-872.

WICKHAM, H. 2009. ggplot2: elegant graphics for data analysis. Springer New York.

WISHART, D. S., KNOX, C., GUO, A. C., CHENG, D., SHRIVASTAVA, S., TZUR, D., GAUTAM, B. & HASSANALI, M. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research,* 36**,** D901-D906.

WOELFLE, M., OLLIARO, P. & TODD, M. H. 2011. Open science is a research accelerator. *Nat Chem,* 3**,** 745-748.

XU, J. 2007. Medicinal proteomics: Searching from within. *Drug Discovery Today: Disease Mechanisms,* 4**,** 37-40.

YOUNG, R. N. 2001. 5 Discovery of Montelukast: a Once-a-Day Oral Antagonist of Leukotriene D4 for the Treatment of Chronic Asthma. *In:* OXFORD, F. D. K. A. A. W. (ed.) *Progress in Medicinal Chemistry*. Elsevier.

YU, H., BURTON, B., HUANG, C. J., WORLEY, J., CAO, D., JOHNSON, J. P., JR., URRUTIA, A., JOUBRAN, J., SEEPERSAUD, S., SUSSKY, K., HOFFMAN, B. J. & VAN GOOR, F. 2012. Ivacaftor potentiation of multiple CFTR channels with gating mutations. *J Cyst Fibros.* Netherlands: 2012 European Cystic Fibrosis Society. Published by Elsevier B.V.

ZHANG, K. Y. J., CARD, G. L., SUZUKI, Y., ARTIS, D. R., FONG, D., GILLETTE, S., HSIEH, D., NEIMAN, J., WEST, B. L., ZHANG, C., MILBURN, M. V., KIM, S.-H., SCHLESSINGER, J. & BOLLAG, G. 2004. A Glutamine Switch Mechanism for Nucleotide Selectivity by Phosphodiesterases. *Molecular Cell,* 15**,** 279-286.

ZHU, F., SHI, Z., QIN, C., TAO, L., LIU, X., XU, F., ZHANG, L., SONG, Y., LIU, X., ZHANG, J., HAN, B., ZHANG, P. & CHEN, Y. 2012. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Research,* 40**,** D1128-D1136.

ZUBER, R., ANZENBACHEROVÁ, E. & ANZENBACHER, P. 2002. Cytochromes P450 and experimental models of drug metabolism. *Journal of Cellular and Molecular Medicine,* 6**,** 189-198.

8 Appendix

8.1 Pipeline

Please see the uploaded ZIP file for all scripts needed to run this analysis, in the directory 'scripts' and subdirectory 'tools'. To edit the automated running of the pipeline please see 'scripts/pipeline.pl'. The data used within this thesis is included alongside a distribution ready folder containing only the required inputs and scripts except the required drugbank database saved as 'inputs/raw/drugbank_xml.xml' and 'inputs/raw/drugbank_target_links.csv' to be downloaded from http://www.drugbank.ca/downloads (described as 'All Drugs, including target, transporter, carrier, and enzyme information' and 'Links to external databases and external identifiers for drug targets' respectively). Additionally, ChEMBL database access information, such as username and password, must be completed in 'scripts/tools/chembl_query_all.pl' and 'scripts/tools/chembl_query_confident_smallmol.pl', however sample data from ChEMBL 13 is supplied.

8.2 Druggable genome list

Please see the attached ZIP file 'outputs/smallmol/all_smallmol_predictions.txt' for a complete tab delimited list of all genes/proteins predicted as druggable. Per new line the file includes: Ensembl gene ID, Uniprot ID, Entrez ID, HGNC symbol, Predicted target class, Predicted target subclass, Chembl TID, Chembl class, Chembl subclass(es), Approved in and Prediction method(s).

8.3 Biopharmable genome list

Please see the attached ZIP file 'outputs/biotech/all_biotech_predictions.txt' for a complete tab delimited list of all genes/proteins predicted as biopharmable. Per new line the file includes: Ensembl gene ID, Uniprot ID, Entrez ID, HGNC symbol, Predicted target class, Predicted target subclass, Chembl TID, Chembl class, Chembl subclass(es), Approved in and Prediction method(s).

8.4 Updated Hopkins and Groom InterPro domains list

InterPro Domains			
IPR002290	IPR004709	IPR001054	IPR000811
IPR000276	IPR001431	IPR001148	IPR001170
IPR001245	IPR001604	IPR001179	IPR001273
IPR005821	IPR000322	IPR001192	IPR001424
IPR001254	IPR000907	IPR002018	IPR006131
IPR000387	IPR001124	IPR002085	IPR006132
IPR006026	IPR001429	IPR000834	IPR006130
IPR000832	IPR000889	IPR001395	IPR006620
IPR001128	IPR001129	IPR000337	IPR013547
IPR000536	IPR001241	IPR000917	IPR004307
IPR007782	IPR002937	IPR000300	IPR000183
IPR002198	IPR000022	IPR000734	IPR000590
IPR001140	IPR000602	IPR001327	IPR000903
IPR022642	IPR005815	IPR001807	IPR000933
IPR006202	IPR005814	IPR002398	IPR001295
IPR000215	IPR001365	IPR006068	IPR001369
IPR006201	IPR006172	IPR004014	IPR001514
IPR013766	IPR006134	IPR006069	IPR001747
IPR002073	IPR006133	IPR001433	IPR001796
IPR002181	IPR002173	IPR001873	IPR002088
IPR000477	IPR000043	IPR004841	IPR000312
IPR001190	IPR020830	IPR004840	IPR000398
IPR001763	IPR020828	IPR000451	IPR000788
IPR001876	IPR020829	IPR000836	IPR001267
IPR003594	IPR000643	IPR001412	IPR001631
IPR000712	IPR003042	IPR001969	IPR001866
IPR000175	IPR001093	IPR003024	IPR001985
IPR000217	IPR012317	IPR001102	IPR002060
IPR000403	IPR001330	IPR001211	IPR002202
IPR000413	IPR002205	IPR001375	IPR002365
IPR000169	IPR002516	IPR002007	IPR002755
IPR001320	IPR002657	IPR002117	IPR002948
IPR015590	IPR000146	IPR000542	IPR000489
IPR011264	IPR000323	IPR000994	IPR000531
IPR002113	IPR000572	IPR002129	IPR000821
IPR006153	IPR002089	IPR002369	IPR001847

8.5 GO terms ranked by accessibility

GO Term	GO Term Name	Confidence Biopharmable
GO:0016020	membrane	low
GO:0016021	integral to membrane	low
GO:0005789	endoplasmic reticulum membrane	no
GO:0005743	mitochondrial inner membrane	no
GO:0005886	plasma membrane	medium
GO:0032587	ruffle membrane	medium
GO:0005615	extracellular space	high
GO:0005576	extracellular region	high
GO:0016324	apical plasma membrane	medium
GO:0031205	endoplasmic reticulum Sec complex	no
GO:0016323	basolateral plasma membrane	medium
GO:0030659	cytoplasmic vesicle membrane	low
GO:0031901	early endosome membrane	no
GO:0034707	chloride channel complex	medium
GO:0005741	mitochondrial outer membrane	no
GO:0005887	integral to plasma membrane	low
GO:0005923	tight junction	medium
GO:0005578	proteinaceous extracellular matrix	high
GO:0005643	nuclear pore	no
GO:0005925	focal adhesion	high
GO:0005765	lysosomal membrane	no
GO:0000139	Golgi membrane	no
GO:0030667	secretory granule membrane	high
GO:0030658	transport vesicle membrane	low
GO:0030672	synaptic vesicle membrane	low
GO:0031315	extrinsic to mitochondrial outer membrane	no
GO:0005747	mitochondrial respiratory chain complex I	no
GO:0031966	mitochondrial membrane	no
GO:0070469	respiratory chain	no
GO:0030126	COPI vesicle coat	no
GO:0045121	membrane raft	medium
GO:0034364	high-density lipoprotein particle	high
GO:0034366	spherical high-density lipoprotein particle	high
GO:0001772	immunological synapse	high
GO:0008305	integrin complex	high
GO:0009897	external side of plasma membrane	high
GO:0031088	platelet dense granule membrane	high
GO:0010008	endosome membrane	no
GO:0030670	phagocytic vesicle membrane	no
GO:0031092	platelet alpha granule membrane	high
GO:0042383	sarcolemma	medium

GO:0030121	AP-1 adaptor complex	low
GO:0030130	clathrin coat of trans-Golgi network vesicle	low
GO:0030666	endocytic vesicle membrane	no
GO:0032281	alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid selective glutamate receptor complex	high
GO:0005891	voltage-gated calcium channel complex	high
GO:0005905	coated pit	high
GO:0030117	membrane coat	medium
GO:0030131	clathrin adaptor complex	low
GO:0019898	extrinsic to membrane	low
GO:0030897	HOPS complex	no
GO:0031225	anchored to membrane	low
GO:0042622	photoreceptor outer segment membrane	medium
GO:0031528	microvillus membrane	high
GO:0031965	nuclear membrane	no
GO:0019815	B cell receptor complex	high
GO:0001518	voltage-gated sodium channel complex	high
GO:0005901	caveola	high
GO:0031227	intrinsic to endoplasmic reticulum membrane	no
GO:0043020	NADPH oxidase complex	low
GO:0071438	invadopodium membrane	high
GO:0005778	peroxisomal membrane	no
GO:0030173	integral to Golgi membrane	no
GO:0032580	Golgi cisterna membrane	no
GO:0031012	extracellular matrix	high
GO:0005746	mitochondrial respiratory chain	no
GO:0005750	mitochondrial respiratory chain complex III	no
GO:0031902	late endosome membrane	no
GO:0042101	T cell receptor complex	high
GO:0042612	MHC class I protein complex	high
GO:0005589	collagen type VI	high
GO:0045211	postsynaptic membrane	medium
GO:0055038	recycling endosome membrane	low
GO:0012506	vesicle membrane	low
GO:0032592	integral to mitochondrial membrane	no
GO:0005637	nuclear inner membrane	no
GO:0042589	zymogen granule membrane	high
GO:0016942	insulin-like growth factor binding protein complex	high
GO:0070821	tertiary granule membrane	high
GO:0009279	cell outer membrane	high
GO:0005890	sodium:potassium-exchanging ATPase complex	high
GO:0005581	collagen	high
GO:0042613	MHC class II protein complex	medium
GO:0005579	membrane attack complex	high
GO:0032311	angiogenin-PRI complex	high

GO:0034045	pre-autophagosomal structure membrane	no
GO:0031264	death-inducing signaling complex	high
GO:0031265	CD95 death-inducing signaling complex	high
GO:0008076	voltage-gated potassium channel complex	high
GO:0031080	Nup107-160 complex	no
GO:0046930	pore complex	low
GO:0031307	integral to mitochondrial outer membrane	no
GO:0043190	ATP-binding cassette (ABC) transporter complex	low
GO:0033177	proton-transporting two-sector ATPase complex, proton-transporting domain	low
GO:0005779	integral to peroxisomal membrane	no
GO:0000421	autophagic vacuole membrane	low
GO:0030055	cell-substrate junction	high
GO:0005604	basement membrane	high
GO:0070083	clathrin sculpted monoamine transport vesicle membrane	no
GO:0034362	low-density lipoprotein particle	high
GO:0009898	internal side of plasma membrane	no
GO:0016328	lateral plasma membrane	high
GO:0016342	catenin complex	no
GO:0043296	apical junction complex	high
GO:0032585	multivesicular body membrane	no
GO:0005614	interstitial matrix	high
GO:0005640	nuclear outer membrane	no
GO:0030122	AP-2 adaptor complex	medium
GO:0060171	stereocilium membrane	high
GO:0005915	zonula adherens	high
GO:0030176	integral to endoplasmic reticulum membrane	no
GO:0000221	vacuolar proton-transporting V-type ATPase, V1 domain	no
GO:0033162	melanosome membrane	no
GO:0005594	collagen type IX	high
GO:0030665	clathrin coated vesicle membrane	low
GO:0031201	SNARE complex	medium
GO:0005606	laminin-1 complex	high
GO:0005610	laminin-5 complex	high
GO:0005605	basal lamina	high
GO:0031258	lamellipodium membrane	high
GO:0031527	filopodium membrane	high
GO:0033017	sarcoplasmic reticulum membrane	no
GO:0005607	laminin-2 complex	high
GO:0005834	heterotrimeric G-protein complex	no
GO:0031234	extrinsic to internal side of plasma membrane	no
GO:0005592	collagen type XI	high
GO:0033116	endoplasmic reticulum-Golgi intermediate compartment membrane	no
GO:0042734	presynaptic membrane	medium
GO:0030118	clathrin coat	no

GO:0030056	hemidesmosome	high
GO:0031526	brush border membrane	medium
GO:0030663	COPI coated vesicle membrane	no
GO:0030315	T-tubule	low
GO:0060201	clathrin sculpted acetylcholine transport vesicle membrane	low
GO:0060203	clathrin sculpted glutamate transport vesicle membrane	low
GO:0061202	clathrin sculpted gamma-aminobutyric acid transport vesicle membrane	low
GO:0042584	chromaffin granule membrane	low
GO:0008282	ATP-sensitive potassium channel complex	high
GO:0034673	inhibin-betaglycan-ActRII complex	medium
GO:0016327	apicolateral plasma membrane	medium
GO:0031235	intrinsic to internal side of plasma membrane	no
GO:0030132	clathrin coat of coated pit	medium
GO:0070772	PAS complex	no
GO:0030119	AP-type membrane coat adaptor complex	no
GO:0017071	intracellular cyclic nucleotide activated cation channel complex	no
GO:0016469	proton-transporting two-sector ATPase complex	low
GO:0005774	vacuolar membrane	no
GO:0033180	proton-transporting V-type ATPase, V1 domain	low
GO:0030128	clathrin coat of endocytic vesicle	no
GO:0070062	extracellular vesicular exosome	high
GO:0012507	ER to Golgi transport vesicle membrane	no
GO:0034704	calcium channel complex	high
GO:0060170	cilium membrane	medium
GO:0031095	platelet dense tubular network membrane	no
GO:0005749	mitochondrial respiratory chain complex II	no
GO:0019867	outer membrane	no
GO:0005889	hydrogen:potassium-exchanging ATPase complex	high
GO:0097025	MPP7-DLG1-LIN7 complex	high
GO:0031253	cell projection membrane	medium
GO:0045092	interleukin-18 receptor complex	high
GO:0045323	interleukin-1 receptor complex	high
GO:0043205	fibril	high
GO:0005588	collagen type V	high
GO:0005892	acetylcholine-gated channel complex	high
GO:0070765	gamma-secretase complex	low
GO:0005587	collagen type IV	high
GO:0035631	CD40 receptor complex	high
GO:0005642	annulate lamellae	no
GO:0030867	rough endoplasmic reticulum membrane	no
GO:0009925	basal plasma membrane	medium
GO:0031314	extrinsic to mitochondrial inner membrane	no
GO:0005597	collagen type XVI	high
GO:0030669	clathrin-coated endocytic vesicle membrane	no

GO:0034359	mature chylomicron	high
GO:0034360	chylomicron remnant	high
GO:0034361	very-low-density lipoprotein particle	high
GO:0034363	intermediate-density lipoprotein particle	high
GO:0042627	chylomicron	high
GO:0031362	anchored to external side of plasma membrane	high
GO:0031313	extrinsic to endosome membrane	no
GO:0043083	synaptic cleft	high
GO:0005757	mitochondrial permeability transition pore complex	no
GO:0042765	GPI-anchor transamidase complex	no
GO:0031305	integral to mitochondrial inner membrane	no
GO:0031224	intrinsic to membrane	low
GO:0032588	trans-Golgi network membrane	no
GO:0001891	phagocytic cup	no
GO:0005639	integral to nuclear inner membrane	no
GO:0030125	clathrin vesicle coat	no
GO:0017059	serine C-palmitoyltransferase complex	no
GO:0035339	SPOTS complex	no
GO:0031233	intrinsic to external side of plasma membrane	high
GO:0043257	laminin-8 complex	high
GO:0043259	laminin-10 complex	high
GO:0043256	laminin complex	high
GO:0034676	alpha6-beta4 integrin complex	high
GO:0005927	muscle tendon junction	medium
GO:0014701	junctional sarcoplasmic reticulum membrane	no
GO:0031301	integral to organelle membrane	no
GO:0005899	insulin receptor complex	high
GO:0005924	cell-substrate adherens junction	high
GO:0042022	interleukin-12 receptor complex	high
GO:0072536	interleukin-23 receptor complex	high
GO:0005753	mitochondrial proton-transporting ATP synthase complex	no
GO:0000275	mitochondrial proton-transporting ATP synthase complex, catalytic core F(1)	no
GO:0045261	proton-transporting ATP synthase complex, catalytic core F(1)	low
GO:0044420	extracellular matrix part	high
GO:0042567	insulin-like growth factor ternary complex	high
GO:0005900	oncostatin-M receptor complex	high
GO:0005742	mitochondrial outer membrane translocase complex	no
GO:0001401	mitochondrial sorting and assembly machinery complex	no
GO:0030526	granulocyte macrophage colony-stimulating factor receptor complex	high
GO:0016471	vacuolar proton-transporting V-type ATPase complex	no
GO:0033178	proton-transporting two-sector ATPase complex, catalytic domain	low
GO:0030127	COPII vesicle coat	no
GO:0030868	smooth endoplasmic reticulum membrane	no

GO:0030660	Golgi-associated vesicle membrane	no
GO:0031090	organelle membrane	no
GO:0005608	laminin-3 complex	high
GO:0030285	integral to synaptic vesicle membrane	low
GO:0008250	oligosaccharyltransferase complex	no
GO:0019897	extrinsic to plasma membrane	medium
GO:0016012	sarcoglycan complex	high
GO:0044459	plasma membrane part	medium
GO:0046691	intracellular canaliculus	medium
GO:0032589	neuron projection membrane	medium
GO:0005577	fibrinogen complex	high
GO:0005744	mitochondrial inner membrane presequence translocase complex	no
GO:0070044	synaptobrevin 2-SNAP-25-syntaxin-1a complex	medium
GO:0017146	N-methyl-D-aspartate selective glutamate receptor complex	high
GO:0008328	ionotropic glutamate receptor complex	high
GO:0032983	kainate selective glutamate receptor complex	high
GO:0002080	acrosomal membrane	no
GO:0070032	synaptobrevin 2-SNAP-25-syntaxin-1a-complexin I complex	medium
GO:0030904	retromer complex	no
GO:0072534	perineuronal net	high
GO:0070022	transforming growth factor beta receptor complex	high
GO:0072563	endothelial microparticle	high
GO:0031232	extrinsic to external side of plasma membrane	high
GO:0033165	interphotoreceptor matrix	high
GO:0031240	external side of cell outer membrane	high
GO:0005584	collagen type I	high
GO:0016010	dystrophin-associated glycoprotein complex	high
GO:0071986	Ragulator complex	no
GO:0016600	flotillin complex	no
GO:0032584	growth cone membrane	medium
GO:0030673	axolemma	medium
GO:0044214	fully spanning plasma membrane	high
GO:0070743	interleukin-23 complex	high
GO:0005754	mitochondrial proton-transporting ATP synthase, catalytic core	no
GO:0034706	sodium channel complex	high
GO:0042105	alpha-beta T cell receptor complex	high
GO:0045171	intercellular bridge	no
GO:0005751	mitochondrial respiratory chain complex IV	no
GO:0005595	collagen type XII	high
GO:0046658	anchored to plasma membrane	medium
GO:0042825	TAP complex	no
GO:0071556	integral to lumenal side of endoplasmic reticulum membrane	no
GO:0042824	MHC class I peptide loading complex	no
GO:0017090	meprin A complex	no

GO:0034098	Cdc48p-Npl4p-Ufd1p AAA ATPase complex	no
GO:0043514	interleukin-12 complex	high
GO:0031226	intrinsic to plasma membrane	medium
GO:0033179	proton-transporting V-type ATPase, V0 domain	low
GO:0005590	collagen type VII	high
GO:0005784	Sec61 translocon complex	no
GO:0005787	signal peptidase complex	no
GO:0048179	activin receptor complex	high
GO:0071439	clathrin complex	no
GO:0032590	dendrite membrane	medium
GO:0000276	mitochondrial proton-transporting ATP synthase complex, coupling factor F(o)	no
GO:0031306	intrinsic to mitochondrial outer membrane	no
GO:0001527	microfibril	high
GO:0031302	intrinsic to endosome membrane	no
GO:0002079	inner acrosomal membrane	no
GO:0044421	extracellular region part	high
GO:0042175	nuclear outer membrane-endoplasmic reticulum membrane network	no
GO:0002081	outer acrosomal membrane	no
GO:0030061	mitochondrial crista	no
GO:0005922	connexon complex	high
GO:0034358	plasma lipoprotein particle	high
GO:0043509	activin A complex	high
GO:0043512	inhibin A complex	high
GO:0045263	proton-transporting ATP synthase complex, coupling factor F(o)	low
GO:0032937	SREBP-SCAP-Insig complex	no
GO:0097060	synaptic membrane	medium
GO:0032809	neuronal cell body membrane	medium
GO:0043260	laminin-11 complex	high
GO:0005898	interleukin-13 receptor complex	high
GO:0000815	ESCRT III complex	no
GO:0031260	pseudopodium membrane	medium
GO:0005896	interleukin-6 receptor complex	high
GO:0070110	ciliary neurotrophic factor receptor complex	high
GO:0000836	Hrd1p ubiquitin ligase complex	no
GO:0031228	intrinsic to Golgi membrane	no
GO:0046696	lipopolysaccharide receptor complex	high
GO:0035354	Toll-like receptor 1-Toll-like receptor 2 protein complex	high
GO:0035355	Toll-like receptor 2-Toll-like receptor 6 protein complex	high
GO:0005895	interleukin-5 receptor complex	high
GO:0005583	fibrillar collagen	high
GO:0005585	collagen type II	high
GO:0071953	elastic fiber	high
GO:0042406	extrinsic to endoplasmic reticulum membrane	no

GO:0031256	leading edge membrane	medium
GO:0031304	intrinsic to mitochondrial inner membrane	no
GO:0045273	respiratory chain complex II	no
GO:0045281	succinate dehydrogenase complex	no
GO:0045282	plasma membrane succinate dehydrogenase complex	medium
GO:0005785	signal recognition particle receptor complex	no
GO:0034679	alpha9-beta1 integrin complex	high
GO:0005591	collagen type VIII	high
GO:0070435	Shc-EGFR complex	high
GO:0042568	insulin-like growth factor binary complex	high
GO:0016013	syntrophin complex	medium
GO:0044425	membrane part	low
GO:0032998	Fc-epsilon receptor I complex	high
GO:0034667	alpha3-beta1 integrin complex	high
GO:0002095	caveolar macromolecular signaling complex	low
GO:0071133	alpha9-beta1 integrin-ADAM8 complex	high
GO:0072517	host cell viral assembly compartment	no
GO:0020017	flagellar membrane	low
GO:0071914	prominosome	high
GO:0032579	apical lamina of hyaline layer	high
GO:0071062	alphav-beta3 integrin-vitronectin complex	high
GO:0031259	uropod membrane	no
GO:0097197	tetraspanin-enriched microdomain	high
GO:0071065	alpha9-beta1 integrin-vascular cell adhesion molecule-1 complex	high
GO:0060342	photoreceptor inner segment membrane	no
GO:0045271	respiratory chain complex I	no
GO:0005586	collagen type III	high
GO:0034678	alpha8-beta1 integrin complex	high
GO:0035003	subapical complex	no
GO:0005602	complement component C1 complex	high
GO:0016011	dystroglycan complex	high
GO:0032591	dendritic spine membrane	medium
GO:0031309	integral to nuclear outer membrane	no
GO:0032002	interleukin-28 receptor complex	high
GO:0042720	mitochondrial inner membrane peptidase complex	no
GO:0005596	collagen type XIV	high
GO:0032473	external side of mitochondrial outer membrane	no
GO:0005600	collagen type XIII	high
GO:0012510	trans-Golgi network transport vesicle membrane	no
GO:0070702	inner mucus layer	medium
GO:0070703	outer mucus layer	high
GO:0019866	organelle inner membrane	no
GO:0005582	collagen type XV	high
GO:0034365	discoidal high-density lipoprotein particle	high

GO:0042025	host cell nucleus	no
GO:0020005	symbiont-containing vacuole membrane	no
GO:0020003	symbiont-containing vacuole	no
GO:0034666	alpha2-beta1 integrin complex	high

8.6 GO term evidence codes

Experimental Evidence Codes:

EXP: Inferred from Experiment

IDA: Inferred from Direct Assay

IPI: Inferred from Physical Interaction

IMP: Inferred from Mutant Phenotype

IGI: Inferred from Genetic Interaction

IEP: Inferred from Expression Pattern

Computational Analysis Evidence Codes:

ISS: Inferred from Sequence or Structural Similarity

ISO: Inferred from Sequence Orthology

ISA: Inferred from Sequence Alignment

ISM: Inferred from Sequence Model

IGC: Inferred from Genomic Context

IBA: Inferred from Biological aspect of Ancestor

IBD: Inferred from Biological aspect of Descendant

IKR: Inferred from Key Residues

IRD: Inferred from Rapid Divergence

RCA: inferred from Reviewed Computational Analysis

Author Statement Evidence Codes:

TAS: Traceable Author Statement

NAS: Non-traceable Author Statement

IC: Inferred by Curator

ND: No biological Data available

Automatically-assigned Evidence Codes:

IEA: Inferred from Electronic Annotation

Obsolete Evidence Codes:

NR: Not Recorded