Towards Personalized Medicine: Computational Approaches to Support Drug Design and Clinical Decision Making

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

M. Sc. Charlotta Pauline Irmgard Schärfe

aus Walsrode

Tübingen
2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	03.12.2018
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Debora Marks

If it were not for the great variability among individuals, medicine might as well be a science, not an art.

Sir William Osler, 1892

Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

Towards Personalized Medicine: Computational Approaches to Support Drug Design and Clinical Decision Making

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe.
Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden.
Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Ort, Datum	Unterschrift

Abstract

The future looks bright for a clinical practice that tailors the therapy with the best efficacy and highest safety to a patient. Substantial amounts of funding have resulted in technological advances regarding patient-centered data acquisition — particularly genetic data. Yet, the challenge of translating this data into clinical practice remains open.

To support drug target characterization, we developed a global maximum entropy-based method that predicts protein-protein complexes including the three-dimensional structure of their interface from sequence data. To further speed up the drug development process, we present methods to reposition drugs with established safety profiles to new indications leveraging paths in cellular interaction networks. We validated both methods on known data, demonstrating their ability to recapitulate known protein complexes and drug-indication pairs, respectively.

After studying the extent and characteristics of genetic variation with a predicted impact on protein function across 60,607 individuals, we showed that most patients carry variants in drug-related genes. However, for the majority of variants, their impact on drug efficacy remains unknown. To inform personalized treatment decisions, it is thus crucial to first collate knowledge from open data sources about known variant effects and to then close the knowledge gaps for variants whose effect on drug binding is still not characterized. Here, we built an automated annotation pipeline for patient-specific variants whose value we illustrate for a set of patients with hepatocellular carcinoma. We further developed a molecular modeling protocol to predict changes in binding affinity in proteins with genetic variants which we evaluated for several clinically relevant protein kinases.

Overall, we expect that each presented method has the potential to advance personalized medicine by closing knowledge gaps about protein interactions and genetic variation in drug-related genes. To reach clinical applicability, challenges with data availability need to be overcome and prediction performance should be validated experimentally.

Zusammenfassung

Therapien mit der besten Wirksamkeit und höchsten Sicherheit werden in Zukunft auf den Patienten zugeschnitten werden. Hier haben erhebliche finanzielle Mittel zu technologischen Fortschritten bei der patientenzentrierten Datenerfassung geführt, aber diese Daten in die klinische Praxis zu übertragen, bleibt aktuell noch eine Herausforderung.

Um die Wirkstoffforschung in der Charakterisierung therapeutischer Zielproteine zu unterstützen, haben wir eine Maximum-Entropie-Methode entwickelt, die Protein-Interaktionen und ihre dreidimensionalen Struktur aus Sequenzdaten vorhersagt. Darüber hinaus, stellen wir Methoden zur Repositionierung von etablierten Arzneimitteln auf neue Indikationen vor, die Pfade in zellulären Interaktionsnetze nutzen. Diese Methoden haben wir anhand bekannter Daten validiert und ihre Fähigkeit demonstriert, bekannte Proteinkomplexe bzw. Wirkstoff-Indikations-Paare zu rekapitulieren.

Unsere Analyse genetischer Variation mit einem Einfluss auf die Proteinfunktion in 60.607 Individuen konnte zeigen, dass nahezu jeder Patient funktionsverändernde Varianten in Medikamenten-assoziierten Genen trägt. Der direkte Einfluss der meisten beobachteten Varianten auf die Medikamenten-Wirksamkeit ist jedoch noch unbekannt. Um dennoch personalisierte Behandlungsentscheidungen treffen zu können, präsentieren wir eine Annotationspipeline für genetische Varianten, deren Wert wir für Patienten mit hepatozellulärem Karzinom illustrieren konnten. Darüber hinaus haben wir ein molekulares Modellierungsprotokoll entwickelt, um die Veränderungen in der Bindungsaffinität von Proteinen mit genetischen Varianten voraussagen.

Insgesamt sind wir davon überzeugt, dass jede der vorgestellten Methoden das Potential hat, Wissenslücken über Proteininteraktionen und genetische Variationen in medikamentenbezogenen Genen zu schließen und somit das Feld der personalisierten Medizin voranzubringen. Um klinische Anwendbarkeit zu erreichen, gilt es in der Zukunft, verbleibende Herausforderungen bei der Datenverfügbarkeit zu bewältigen und unsere Vorhersagen experimentell zu validieren.

Acknowledgments

They say it takes a village to raise a child, and similarly, many colleagues and friends contributed to my academic coming of age. I want to express my gratitude to you:

To my advisors Oliver Kohlbacher and Debora Marks for taking me on as their PhD student. Their feedback, mentorship, and consistent pushing of my thinking challenged me in all the right ways. To Chris Sander for his mentorship and guidance. A special thanks to Debbie and Chris for their hospitality in Boston and the many inspiring discussions about computational biology and science in general.

To the additional members of my examining committee, Kay Nieselt, Nico Pfeifer and Daniel Huson.

To my lab mates in the Kohlbacher and Marks labs and my academic friends in Tübingen and Boston for stimulating conversations and a mind-opening research environment. A big *thank you* to Thomas Hopf, Benjamin Schubert, Fabian Aicheler and Philipp Thiel for proof-reading parts of this thesis and to Luis Luis Luis de la Garza for being such a wonderful office-companion. To Li Yang Smith for her hospitality and encouragement.

To all my collaborators, co-authors and students for their valuable scientific contributions and insightful discussions.

To the administrative and technical teams in Tübingen and at HMS, in particular Claudia Walter, for keeping things running smoothly and their constant help.

To the Fulbright Commission whose support made it possible to spend part of my PhD studies in the US for connecting me with a network of inspiring people from around the world.

To my friends and family, in particular my partner, Henry, for their unconditional support and well needed distraction when things did not work out as they should. A special thanks to my dad for inspiring discussions and proof-reading this thesis.

Thank You!

General Remarks

- In accordance with the standard scientific protocol, the personal pronoun ***we*** will be used to indicate the reader and the writer, or my scientific collaborators and myself.
- Unless stated otherwise, all figures of protein structures were generated using the freely available visualization software Pymol[378].

Contents

Chapter 1

Introduction

Motivation

The sequencing of the first full human genome in the Human Genome Project took more than ten years to complete and cost a little less than three billion US dollars. The sequencing cost not only dropped by a 100-fold over the course of the project [226], but has recently reached its all time low: a complete whole personal genome is now available for less than $1,000 including its analysis [118] and within a few days [125]. The massive technological progress that came with the hunt for ever cheaper availability of human genomes has also led to advances in other biological fields as it transformed gene sequencing and related methods, suitable for regular use in a research lab.

The broad availability of full genome sequences for a wide range of organisms has advanced our understanding of the evolutionary history within individual species and across all domains of life [76]. Through genetic material only, it is also possible to characterize long-extinct species, such as the woolly mammoth or the Neanderthal at a precision never seen before [129,285,310,343]. However, one of the main promises that came with the thorough understanding of the human genome — that of helping to cure diseases — has not been fulfilled yet [126].

Among the unexpected complexity of the human genome [126], scientific and technical challenges include data processing and storage as well as suitable data analysis and mining algorithms [405]. The genetic and epigenetic markup of a patient, the exact status of expressed proteins in the affected cells and other features of their disease may in the future be used to improve 1) therapy decisions and 2) drug action and metabolism in the patient, based on a better understanding of disease etiology. These goals can only be reached by integrating computational analysis through novel algorithms, pipelines and user interfaces into the process of disease research, drug discovery and medical decision making.

The increasingly obvious lack of a thorough understanding of many cellular processes has further complicated the translation of sequencing-based insights into clinical practice. It turns out that many genes previously linked to disease etiology may very well be unrelated[236], and even newly designed drugs can be rendered useless when the cell activates alternative pathways to evade drug action[158]. We still do not know the complete mechanism of action for many drugs, let alone the three-dimensional structure or dynamic behavior of many disease- and drug-associated proteins. In addition to the interpretation of personal genomes, it is thus essential to shed light on these unknowns — a process that can leverage computational methods.

Only in the recent past it has become possible to infer higher order knowledge about a protein's molecular structure, fitness and involvement in interactions using the data produced by the genomic revolution. One example for such a method is the inference of residue interactions ("contacts") in the three-dimensional structure of a protein based on the protein's evolutionary record. This allowed to predict the protein structure *de novo* for a wide set of proteins, including many pharmacologically relevant proteins[159,267].

In this thesis, we present several approaches to translating the wealth of genetic data into a thorough understanding of disease with the ultimate goal of improving patient care. Our contributions are distributed along the personalized health value chain and range from predicting residue-level protein-protein interaction interfaces to developing a pipeline that could support the medical community in interpreting a patient's genome.

Part I: Phenotype discovery based on the evolutionary record of proteins

Protein-protein interactions (PPI) are of pivotal importance for the pathogenesis of various diseases due to their involvement in many cellular processes[270,362], but often elude experimental characterization. Predicting which proteins interact and how genetic variants affect that state can thus serve as an entry point for the identification of new therapeutic targets[362].

In the first part of the thesis, we developed a method using the evolutionary record of proteins, determined by gene sequencing, to gain phenotypic insights of protein complexes. Building on co-evolution approaches for monomer structure prediction we developed a method capable of predicting protein interactions at residue-resolution as well as identifying specific interacting protein pairs in larger proteins complexes.

Part II: Genetic variation in drug-related genes

Affordable genotyping methods can be used to profile patients for genetic determinants of drug efficacy or toxicity (pharmacogenetics/pharmacogenomics, short PGx)[260] in population-stratified[282,311,468] or individualized settings[258,341]. Even though early PGx studies mainly

focused on genes of particular interest for pharmacodynamics (PD) or -kinetics (PK), larger genome-wide association studies (GWAS) have recently been able to identify variants associated with drug efficacy or toxicity in a less biased manner[70,296]. So far, these approaches have led to the identification of more than 10,000 associations between genetic variants and alteration in drug response (collated in PharmGKB[453]). In addition to that, drug labels for 202 medical compounds now include PGx data with some requiring genetic testing prior to prescription (including 36 drugs with PGx testing required by the U.S. Food and Drug Administration, short FDA)[340].

Recent advances in the creation of large reference sets of human variation[63,236,427] now allow combining knowledge of drug mechanism with the prevalence of non-functional alleles in drug-related genes. In the second part of the thesis, we use such data sets to take a complementary approach to classic PGx research by including existing biological information about drug action and disease etiology to study the prevalence of functional variation in drug-related genes. The results of our survey can be used to identify potential genetic variants for targeted PGx studies and the extension of PGx screening panels even though our knowledge of the pharmacological action of drugs and understanding of genetic interaction in a cellular or tissue-specific context remains incomplete[279].

Part III: Drug repurposing

The complete depletion of a drug target may result in non-rescuable resistance to that drug in a patient. Here, alternative therapy strategies need to be explored using either drugs that treat the disease through a different mechanism of action or, possibly, through drugs repurposed from other indications.

Inspired by success stories of compounds serendipitously repositioned for another indication, a plethora of systematic drug repurposing methods have been developed over the last decade using for example guilt-by-association (GBA) approaches of chemical and protein similarity[203,369] or knowledge about cellular protein interaction networks[249,272,435].

In the third part of the thesis, we have built on our insights of the previous two chapters and present a heterogeneous information network (HIN) that incorporates different objects and relations to connect drugs, diseases and genetic variants through genes. The goal of this project was to use this network to identify drugs whose target genes are in the genetic neighborhood of the genes involved in the disease ("neighborhood targets"). By defining meta-paths in the HIN, we developed two approaches for predicting novel indications for existing drugs — in an unsupervised and supervised manner.

Part IV: Clinical reporting and in silico treatment stratification

From a patient perspective, identifying variants affecting drug efficacy is only the first step towards better treatment decisions. The true values lies in utilizing the molecular profile of their personal disease to guide therapy decisions[51,213,380]. While newly identified variants need characterization, validation and ultimately representation in dosing and treatment guidelines, quickly providing a comprehensive overview over existing clinical knowledge to the treating physician can already improve the treatment decision process[202].

In the last part of the thesis, we developed a pipeline that annotates genetic variation found in a patient with existing knowledge to support physicians in their clinical decision making. To include previously unknown variants in drug-related genes, we further devised a molecular modeling workflow that predicts the effect of such variants on the pharmacological phenotype of a protein.

Chapter 2

Background

2.1 Genetic Variation in Humans

2.1.1 The Human Genome

The instructions needed to build and direct activities of an organism are mostly stored in its genome - a set of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA, only in the case of some viruses) molecules, often grouped into chromosomes[77]. The DNA molecule contains regions that code for functional units called protein-coding *genes* which are translated to *proteins*, intermixed with large regions that are not translated and fulfill other (regulatory) functions[332]. Proteins then perform specific actions in a cell of the organism, including transport of molecules, catalysis of chemical reactions, and housekeeping functionalities.

The human genome is with 3.2 billion base pairs (3.2 Gb) among the larger genomes currently known and is estimated to contain approximately 20,500 genes[332], which take up only about 1% of the total genome[304]. In its most recent versions the reference assemblies, GRCH37 and GRCh38, of the human genome covers now close to 90% of the full genome[180], with more than 300 gaps remaining between scaffolds in the most recent release, GRCh38.p12 (as of August 2018).

To understand how this low number of genes can result in the complex structure of the human organism that grows from 1 zygote to 37 trillion cells[26] of more than 200 different cell types[351] one needs to take a closer look at the genome and protein-coding gene structure:

Each gene in the genome consists of an open reading frame (ORF) with *exons*, elements that are transcribed and translated into a protein, and *introns*, elements that are transcribed, but then removed prior to translation. The untranslated regions (UTRs) at the 5' and 3' ends contain ribosome binding sites and the start/stop codons. For a gene to be expressed, a promoter region at it's 5' terminus binds transcription factors (proteins regulating the rate of transcription) and RNA polymerase which is responsible for the transcription of

the gene into messenger RNA, mRNA. Through *alternative splicing* of the same mRNA a multitude of protein isoforms with different functionality can be encoded in a single gene. Multiple promoter regions further increase the combinatorial space arising from different transcription alternatives[293]. Through these mechanisms, up to 95% of the genes in the human genome with more than one exon give rise to multiple, different protein products[321].

The human genome is organized into 23 chromosomes, which are compact structures consisting of DNA wrapped around histones that package the DNA tightly enough to fit into the cell's nucleus while also structuring it for later expression. Except from germ cells, human cells contain two sets of chromosomes, one from each parent. These cells are called diploid — compared to haploid germ cells that only contain one set of chromosomes and give rise to gametes used in sexual reproduction.

2.1.2 Evolution and Population Genetics

In the course of the past 3.8 billion years, life on earth has been able to survive in different living circumstances through the gradual introduction of random smaller and larger changes into the genome that have helped the offspring adapt to and survive in a particular habitat[289]. Our understanding of this evolutionary process driven by natural selection on variation in a population produced by genetic mutation and recombination[171] can be traced back to major contributions from Charles Darwin, Alfred Wallace and Gregor Mendel, synthesized by Roland Fisher in his book "The Genetical Theory of Natural Selection" in 1930[94,108].

The *theory of natural selection and adaptation* was formulated independently by Alfred Wallace and Charles Darwin after they made (at that time curious) observations about slight variations of related species in different environmental niches. To explain these observations they formulated a hypothesis that can be broken down in three pillars:

1. Phenotypic variation: traits among individuals in a population can vary in respect to physiology, morphology, and behaviors contains one allele for a gene

2. Differential fitness: different traits confer different rates of survival into adulthood and following from this into reproduction

3. Heritability of fitness: traits can be passed onto a child generation

They also concluded that different environments might affect the fitness of an organism resulting in selection of distinct traits in disparate *niches*[237].

While Darwin had no precise explanation for how exactly new species arise, in 1865/66 the Austrian monk Gregor Mendel was the first to devise the modern theory of genetics following a series of breeding experiments with purple and white flower pea plants. From

his observations, he formulated the idea of hereditary units, now known as genes, that may occur in different forms, now known as *alleles*, and result in a particular phenotypic trait of the organism. In this case the gene for pea flower color exists in two alleles, one for each color. An organism then obtains (inherits) an allele from each of its parents and thus may carry two identical (*homozygous*) or two different alleles (*heterozygous*) for a specific trait. This theory can be formulated as the **three laws of inheritance**:

1. Law of segregation: during gamete formation in sexual reproduction, each gamete only contains one allele for a gene
2. Law of independent assortment: the alleles of different genes segregate independently
3. Law of dominance: some alleles can *dominate* the phenotype as soon as they occur in the organism, while others are *recessive* and need to occur in the homozygous state to show their effect.

The analysis of how the genetic markup (*genotype*) changes the organism's observable properties (*phenotype*) are still active research areas.

Natural Selection and Fitness

One fundamental concept of the modern theory of evolution is *natural selection*, which describes the phenomenon that some individuals in a population are more likely to survive and produce offspring than others under certain conditions. This is due to the advantages their genotype provides in the current environment. While this appears simple at first sight, its nuances are not as trivial to understand and have been discussed elsewhere[171].

One particular phenotype often used in this context is the organism's *fitness* which corresponds to its survival potential and its success at reproduction.

Fitness, as other traits, is significantly determined by the organism's genotype. Genetic variation randomly arising in the population is thus one essential driving force of natural selection by introducing novel alleles into the population that may prove to be advantageous in the environment: Individuals with certain alleles may show traits that allow them to survive under certain harsh conditions or reproduce more than others, thus spreading their particular alleles to their offspring. These variants get fixed in the population, which, as a consequence, evolves to show these traits at a larger frequency[171].

Genetic Variation

Genetic variation in an organism is introduced through random mutations of the genome during gamete formation in the parents (*germline variation*) or in the tissue during the

organism's lifetime (*somatic variation*). Especially larger structural changes can be introduced through sexual reproduction by equal or unequal recombination of the genetic material of the parents including the *crossover* of parental chromosomes.

Genetic variation comes in multiple forms: from substitutions of single bases (*single nucleotide variant*, SNV), insertions or deletions of genomic segments (*indels*), up to the inversion, duplication or deletion of full genes (*copy number variants*, CNV)[103] (Figure 2.1). These events can occur at any place in the genome, such as protein-coding segments, transcription-modulating regulatory elements, splice sites, and regions with currently unknown direct connections to protein function[46]. The phenotype can then be affected through direct alterations of the protein transcript as well as modulation of gene expression. Larger structural changes mainly influence the transcription of a gene similar to variants outside coding regions that modulate regulatory elements. Duplication and deletion events result either in amplification or reductions of transcription levels of the affected genes[31].

Genetic changes can affect the function of the gene's protein product in multiple ways and thereby alter the offspring's phenotype: it can reduce a protein's ability to fulfill its cellular function (non-functional), inactivate the protein completely (loss-of-function, LoF), or confer a new molecular function (gain-of-function, GoF). If one of the two copies of a gene in the diploid genome is entirely lost, the second copy on the homologous chromosome has to fulfill the its function — in the case that this copy also harbors a variant, a disease phenotype can arise[103].

2.1.3 Genetic Variation in Disease

While each human genome is different, most variants observed in an individual are common in the population showing allele frequencies (AF) $> 5\%$ in the reference population with only a small percentage (1-4%) being seen in fewer than 0.5% of the total population[63]. While many of the variants may be of neutral effect under our current living conditions, some of them can be associated with beneficial or deleterious effects on health, causing or altering susceptibility to a disease[263,286].

While rare diseases with a strong inheritance pattern are usually characterized by targeted sequencing efforts involving one or more affected families, many common disorders are not amenable to this approach. Here, GWAS have become a popular tool to screen larger cohorts of patients in a genome location-agnostic manner for statistical association between the trait of interest and any locus. The idea behind this approach is, that many common diseases have a different underlying genetic architecture than rare Mendelian disorders. While rare diseases are often caused by single rare variants, common disorders are hypothesized to being influenced by multiple common variants in the population (*common*

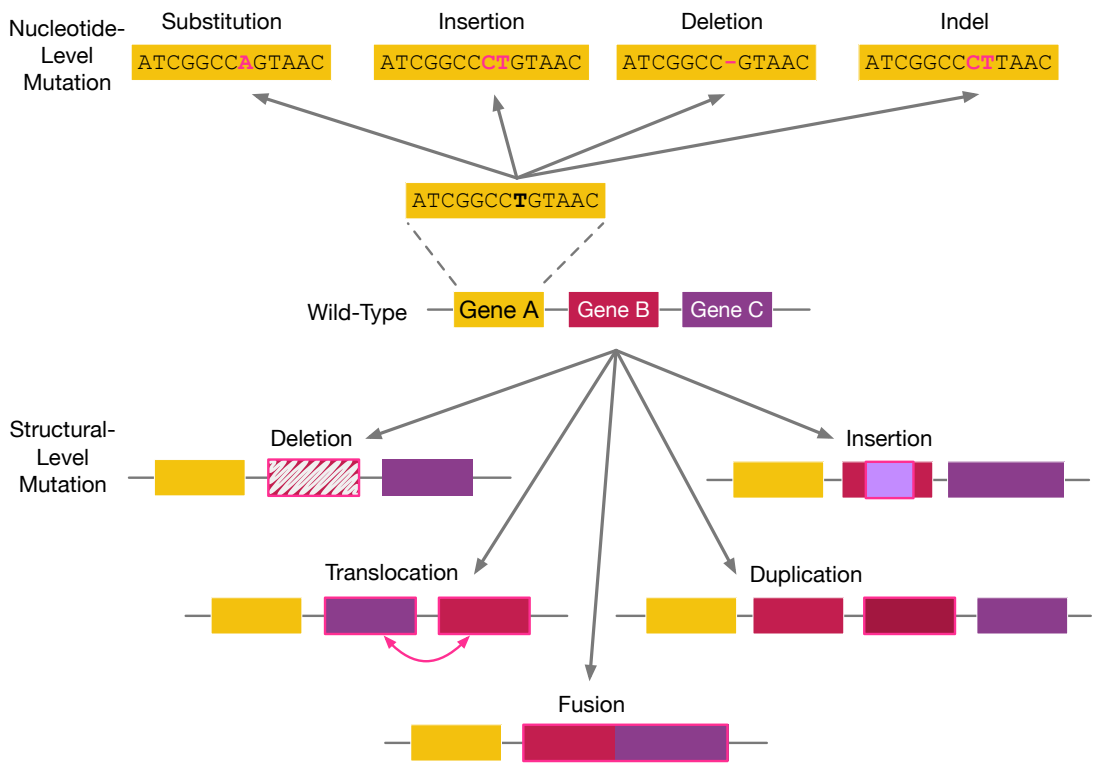


Figure 2.1: Genetic changes can be either on the nucleotide level through nucleotide substitution, insertion, or deletion (or a combination thereof) (top row) or as large scale structured variation of the genome (bottom row). In that case the structure of the genome may be modified through the deletion, duplication, fusion or inversion of larger chromosomal segments. Figure inspired by Cardoso *et al.*[46]

disease-common variant hypothesis)[43]. Such patterns can then be found by testing for statistical associations between a patient cohort and variant.

However, there are problems with genome-wide approaches[432], such as very small effect sizes, and failure to explain the full extent of heritability of the trait[122,482]. Enthusiasm was further damped by the high proportion of association signal being found not within but between gene coding regions, complicating their immediate interpretation[263].

2.2 Drug Discovery

Deregulation of cellular processes caused by genetic variation, infection with a pathogen or other factors can result in the dysfunction of parts or all of an organism, associated with a decline of the well-being (i.e., pathology or *disease*). The aim of pharmaceutical treatment of a diseased individual (*patient*) is to restore their well-being by either alleviating disease-associated symptoms such as pain or to correct the deregulation and stop the dysfunction altogether. Depending on the causal background (*etiology*) of the disease, this can be done by a variety of measures, including surgical procedures or administration of medicines.

The protocol for disease treatment has undergone significant changes over the past centuries from praying, bleeding a patient with leeches, and administering herbal tinctures to the using high-purity therapeutic molecules often targeted specifically at the molecular pathways that are involved in disease etiology. Nowadays, two main classes of drug molecules can be distinguished: 1) small organic or non-organic molecules (*small molecules*) and 2) larger biomolecules such as antibodies (*biologics* or *large molecules*). The discovery process with which therapeutics are found has evolved in parallel with the changes in medical practice, from the serendipitous discovery of pharmacological effects of herbal remedies elicited through one or more unknown compounds therein to a process of systematic search for such molecules.

2.2.1 Drug Pharmacology

Usually, therapeutics act on the organism in one of two ways: 1) *etiology-specific*, i.e., targeting the underlying cause of a disorder, or 2) *palliative*, i.e., alleviating symptoms, but not targeting disease-mechanism[246,470]. A drug's effect is further determined by the molecule's effect on the target organism (*pharmacodynamics*, PD) and the organism's effect on the drug molecule (*pharmacokinetics*, PK).

Pharmacodynamics

The PD profile of a drug molecule describes what the molecule does to the organism. It relates to the *mechanism of action* (MoA) of a compound and includes its molecular effects

on the body (including biochemical and physiological effects such as receptor binding and chemical interactions). Usually, the effect of a drug is dependent on its dose ranging from ineffective if given in too small quantities to toxic if given in too high a dose[258]. The relationship between a compound's dose and the body's response is determined during the drug's development and is required to anticipate later dosing regimens and off-target effects. This dose-response relationship can be described by several parameters, including *affinity*, *potency*, and *efficacy*. While the first describes the ability of a compound to bind to a cellular target, the second describes how much of a compound is needed to provoke a effect and the last the maximal response that can be achieved by a compound.

A therapeutic elicits its MoA through interaction with biomolecules in the cell, such as receptors, enzymes or ion channels. In the past, a multitude of different mechanisms have been discovered[412], including

- Stimulation or depression of a receptor (receptor agonism)
- Blocking of a receptor (receptor antagonism)
- Stabilization of receptor activation
- Inhibition of the activity of an enzyme
- Direct chemical reaction

The way in which this result can be obtained varies by drug, either through competitive interaction with a target, i.e., the drug competes with the natural substrate, or through non-competitive or allosteric effects[412].

Pharmacokinetics

The PK profile of a drug molecule describes what the organism does to a therapeutic molecule. This includes the molecule's passage through the body, from its liberation, **a**bsorption, and **d**istribution to its **m**etabolism and **e**xcretion (ADME). This scheme is often extended to include a drug's toxicological effects (ADMET) as these are often caused by a compound's metabolites.

Before a compound can reach its target in the cell, it is usually taken up by the bloodstream. From there it gets distributed through the different compartments of the body, if not hindered by natural barriers (e.g., the blood-brain barrier). Once in the cell, the drug's response is determined by its PD profile. Near-instantly after entering the body, the molecule also starts being broken down — primarily in the liver. Prior to excretion (for example through the kidneys), therapeutics and their degradation intermediates can interact with enzymes and thus affect the normal metabolism, cause adverse events, and alter the breakdown pathways of other therapeutics.

A variety of proteins is involved in handling therapeutic compounds in the body (sometimes referred to as *pharmacogenes*). These include transporters and carriers responsible for shuttling the therapeutic in and out of the cell and metabolic enzymes involved in its breakdown[218]. The most prominent class of enzymes, involved in the metabolism of up to 80% of all currently available drugs, is the cytrochome P450 family (CYP450)[114].

Pharmacokinetics further influences the drug's bioavailability, i.e., the proportion of the drug that reaches the site of action in the cell. Together with a drug's PD profile it also influences a drug's dose-response relationship and plays a crucial role in finding a therapeutic's optimal formulation, route of administration and dosing. Nowadays, modeling of the PD and PK characteristics of a molecule belongs to the standard procedure in the drug discovery process[347,426].

2.2.2 The Drug Discovery Pipeline

Even today, the discovery process of new therapeutics is a complicated endeavor that does not fit a single cookie-cutter recipe. Nevertheless, we can distinguish some general steps in modern drug discovery and development[412]: the discovery pipeline usually starts with research into the mechanism of the disorder to identify molecular key players that could be modified to alleviate the disorder (*target identification*).

A target can be a protein, gene, or RNA and has to fulfill a catalog of criteria to be considered suitable. In particular, it has to be *druggable*, meaning that it is accessible to putative drug molecules and elicits a measurable biological response upon binding[168]. Then follows the search for suitable molecules (small-molecules or biologics) (*hit and lead identification*) and their optimization to improve their pharmacological and safety profile (*lead optimization*). Afterwards, one or two final candidates are selected for *clinical development* in which the candidates are vigorously tested for efficacy and safety *in vivo*. If they succeed in the clinical trials, the developing instance (often a pharmaceutical company) can apply for marketing approval.

2.2.3 Hurdles in the Drug Development Process

Developing new drugs is a time and cost expensive process which has - despite all technological advances - experienced a decline in productivity over the previous years[284,370]. For example, the widely established estimate of development cost per new molecular entity by Paul *et al.*[326] of $1.8B has been recently updated to $2.8B[79]. Due to increasing constraints imposed by legislators on proving efficacy and safety of drugs, only few chemical compounds in the pipeline make it to the market[370] and many drug candidates fail in the late steps of development or post-marketing[65,214,284,441]: only one in 24 drug candidates entering clinical development, will pass all clinical trial phases[41].

There are multiple reasons why a drug may not work or cause unexpected adverse events in a patient, including 1) mechanistic problems, such as a wrong or insufficiently established target[41,284], 2) cross-reactivity of compounds with *off-targets* (proteins and tissues)[361] and 3) pharmacogenomic effects, meaning that the drug targets or pharmacokinetic genes contain genetic alterations that alter the drugs PD/PK profile[144] (see also Section 2.2.4). The failure of a single-target based therapeutic can in many cases be backtracked to the intrinsic robustness of the target system, as organisms evolved to maintain functionality and handle perturbations in their environment through chemicals, toxins and genetic mutations. Sometimes, a diseased system also remains robust against drug-induced perturbations through shifts in the cellular network flow which results in lack of drug efficacy. At the same time, a drug may interfere with other components of the cellular network, leading to side effects[210]. Especially when there is a high level of genetic diversity, as found in cancer cells and pathogens, chances are high that some cells evade the drug action and lead to resistant subclones[210].

Failures in the late stages of the development cycle are particularly expensive because they require the repetition of most earlier phases to find a new suitable candidate. One of the outspoken goals of the pharmaceutical companies is thus the reduction of late-stage attrition by addressing and solving all potential risks early in the process or kill a project if necessary ("fail early, fail cheap")[223,326]. Developing teams try to incorporate as much high-quality information about the disease, targets, compounds and target organism early on into the drug discovery pipeline[55,65].

The payoffs of recent changes in the development process and ongoing massive investment in the establishment of *in vitro* methods for toxicology screening[441], inclusion of big data[55] and systems-medicine approaches[162], the movement away from the single-target paradigm[277], and towards open innovation[284] remain to be seen in the future.

2.2.4 Influence of Genetic Variation on Drug Efficacy

Even after approval of a drug, patients respond differently to drugs[374] — some suffer from serious adverse events, and others do not experience any effect at all[52]. A multitude of factors can affect the response of a patient to a drug, including the complexity of the disease, drug-drug interactions, environmental factors, age, sex, and also the genetic mutations in drug-related genes in the patient[2,63]. Gaining a thorough understanding of the impact of genetic variations in pharmacogenes on a drug's efficacy has been of scientific interest in the past. PGx aspects now have found their way into the drug development[144] and approval process[481].

The two most important aspects in which genetic variation can affect a therapeutic are: i) modulation of the drug's efficacy and PD properties, if they occur in the therapeutic

target or associated pathways. The breast cancer drug herceptin, for example, is now prescribed to patients whose tumors overexpress its drug target *ERBB2* as this renders them more susceptible to said compound[358]. ii) Modulation of the metabolism of a drug through variants in ADMET genes, giving rise to poor, intermediate, extensive and ultra-rapid metabolizers[144]. Another breast cancer medication, tamoxifen, requires activation through cytochrome P450 enzymes and can be affected by genetic variants in these enzymes that result in their overly slow or quick metabolic activation that can be associated with over-/under-dosing[393].

PGx can be included early in the discovery pipeline[358], and also to guide treatment decisions after marketing approval[453]. Through large-scale sequencing efforts in the past years, there have been efforts from different groups — including us — to quantify and characterize the extent of variants in drug-related genes[44,114,218,302,459].

2.3 Molecular Modeling Methods

Molecular modeling algorithms are commonly used in the *in silico* drug discovery process for modeling, representing and manipulating structures and reactions of molecules (e.g., drug targets, target-ligand complexes, ...). They can be used to characterize a molecule by predicting properties dependent on its three-dimensional structure and to simulate the behavior of molecules.

2.3.1 Molecular Mechanics (MM)

Molecular mechanics refers to the use of classical mechanics (Newtonian mechanics) to describe the energetic properties of the molecule. Atoms, as the smallest individual units considered here, are described as point charges with masses and linked by covalent interactions (*bonds*) and noncovalent interactions. A potential energy function can then be used on this system of N particles (*atoms*) to calculate the potential energy $\mathcal{V}(\boldsymbol{r}^N)$ for the given coordinates $\boldsymbol{r}$ in space. This function describes the potential energy of the system as a sum of covalent (E_{cov}) and non-covalent ($E_{\text{non\_cov}}$, i.e., electrostatic and van der Waals forces) energy terms.

The parameterized form of a potential energy function is called a *force field* in which parameters for equilibrium bond, angle, and dihedral values, van der Waals multipliers and other constants are usually determined empirically. Based on equilibrium constants and the observed terms in the structural conformation at hand, the overall energy of the system's given conformation can be calculated.

Two commonly used force field families for the simulation of biomolecules are *Optimized Potentials for Liquid Simulations* (OPLS)[17,194,454] and *Assisted Model Building with Energy Refinement* (AMBER)[30,66,215,244,446,447]. The functional form of both force field families

is very similar and builds on previous work by Lifson and Warshel[241], but differs in the details of its parameterization. For example, while the authors of AMBER used quantum mechanical calculations to derive the electrostatic potential, the OPLS authors derived charges empirically. Different to the first AMBER force field, the OPLS force field also focused on modeling systems in the presence of an explicit solvent.

The general form of the second generation AMBER force field, was defined by Cornell *et al.*[66] as

$$\begin{aligned}\mathcal{V}(\boldsymbol{r}^N) = &\sum_{\text{bonds}} k_l(l_i - l_{i,eq})^2 + \\ &\sum_{\text{angles}} k_\theta(\theta_i - \theta_{i,eq})^2 + \\ &\sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\omega - \gamma)] + \\ &\sum_{i<j}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right)\end{aligned}$$

Here, the first term sums over the bonds in the molecule, interpreted as ideal spring forces and defined as harmonic potential centered around equilibrium bond-length values ($l_{i,eq}$). The bond-stretching constant k_l controls the stiffness of the bond spring. Parametrization of this term in the force field consists of determining values for k_l and $l_{i,eq}$. The second term similarly sums over the bond angles centered around equilibrium angles ($\theta_{i,eq}$).

The third term sums over torsion angles and represents how the energy changes as a bond rotates using a Fourier series over multiple terms. A bond twist can occur for different reasons, including bond order (e.g., double bonds), neighboring bonds or lone electrons. Here, V_n is the barrier to free rotation, n the periodicity of rotation, ω the observed torsion angle and γ the angle at which the potential energy achieves its minimum. Each of these first three terms thus represents the energy difference between an ideal geometry for bond lengths, bond angles and dihedral angels compared to the actual geometry[446].

The non-bonded energy between all atom pairs is covered by the fourth term of the force field, including van der Waals and electrostatic energies. The van der Waals term is computed as the Lennard-Jones potential in which attractive forces fall off with distance r between two atoms with r^{-6} while repulsive forces fall off with r^{-12}. The electrostatic term, on the other hand, is described using Coulomb's law of atom-centered point charges, falling off with r^{-1} for the atom point charges q_i and q_j.

The AMBER force field — like all others — makes several simplifications. Electrostatic and van der Waals interactions for example are only calculated between atoms separated by at least three bonds and between molecules[66].

2.3.2 Molecular Dynamics (MD)

To model the movements of a molecule (often together with water molecules and ions), MD simulations can be used. Here, Newton's laws of motion are employed to generate a *trajectory* by propagating the potential energy of the system over time. The position and velocity of each atom are updated by solving the differential equation embodied in Newton's second law of motion, $\boldsymbol{F} = m\boldsymbol{a}$, where $\boldsymbol{F}$ is the force acting on the atom calculated from its mass m and its acceleration $\boldsymbol{a}$. The velocity of an atom describes how fast it changes with time, $\boldsymbol{v} = d\boldsymbol{r}/dt$. Acceleration is the derivative of velocity, defined as $\boldsymbol{a} = d\boldsymbol{v}/dt$, the motion of the atom along the coordinates r can thus be written as

$$\frac{d^2\boldsymbol{r}}{dt^2} = \frac{\boldsymbol{F}}{m} = \boldsymbol{a} \tag{2.1}$$

$\boldsymbol{F}$ can be derived from the inter-atomic potentials as described in the potential energy function $F(\boldsymbol{r}) = -\boldsymbol{\nabla}\mathcal{V}(\boldsymbol{r})$

From the trajectory one can then sample the conformation distribution by taking discrete snapshots and calculate time averages of properties.

2.3.3 Molecular Docking

The structure of an inter-molecular complex of two molecules (e.g., a protein *receptor* and a small molecule *ligand*) can be predicted using *molecular docking*[29]. Most molecular docking algorithms consist of two steps: 1) generation of potential receptor-ligand conformations through **sampling** of translational and rotational conformations of the molecules relative to each other (often also including the conformational freedom within each molecule), and 2) **scoring** the solutions towards their likelihood of representing the "true" complex conformation. This second step typically uses a knowlegde-based or empirical scoring function. Particularly when evaluating a large set of ligands for their ability to bind to a protein, the scoring step further has to be able to rank the ligands relative to each other. Most scoring functions try to achieve this task by estimating the free energy of binding ΔG_{bind}.

While scoring methods have become very good at predicting the true complex conformation with near atomic resolutions, they still have problems at predicting accurate binding affinities[363].

2.3.4 Molecular Mechanics Combined with Continuum Models

To improve the accuracy of predicted binding free energies of a ligand-receptor complex, the electrostatic contribution of the solvation free energy can be added to the MM terms of the energy function. Common approaches in this context are MM combined with Poisson-Boltzmann (PB) or generalized Born (GB) and surface area (SA) approaches (MM-PBSA and MM-GBSA) which have been shown to be more accurate than purely empirical scoring functions, particularly when also averaging interaction energies over MD or Monte Carlo simulations[337]. In this method, the standard force field terms are combined with a polar solvation energy obtained from PB or GB continuum-solvation methods, a non-polar continuum solvation energy estimated as linear relation to the solvent-accessible surface area (SASA) and the entropy calculated at the MM level. For MM-PBSA, for example, the overall free energy of a system is then computed as $G = E_{MM} + G_{PBSA} - TS_{MM}$ with $E_{MM} = E_{\text{cov}} + E_{\text{electrostatic}} + E_{\text{vdW}}$ and $E_{\text{cov}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}}$. The solvation free energy $G_{PBSA} = G_{\text{polar}} + G_{\text{non\_polar}}$ can be calculated from the PB equation and an estimate for the non-polar free energy. The solute entropy TS_{MM} can be estimated from the MD trajectory by quasi-harmonic analysis.

The binding affinity is then calculated from the energies of the receptor-ligand (RL) complex as well as structures of ligand (L) and receptor (R) that are obtained by removing the respective partner[338]:

$$\Delta G_{\text{bind}} = \left\langle G_{RL} - G_R - G_L \right\rangle_{RL} \tag{2.2}$$

2.4 Protein Interactions

Proteins often interact with other biomolecules to fulfill their function[21]: cellular processes such as signal transduction, cell-to-cell communication, membrane transport as well as transcription and replication are orchestrated by a multitude of interacting proteins that form binary or higher-order *complexes*[270].

2.4.1 Identification of Protein-Protein Interactions

The process of identifying interacting partners in the cellular protein-protein interaction (PPI) network and their binding sites faces several inherent challenges: the multitude of genes, proteins with their isoforms, and modification states in a cell, low protein abundance and problems with the purification of many proteins therein[172]. PPIs are dynamic and the lifetime of many protein complexes is limited, i.e., transient, to the duration of a particular process which complicates their detection using conventional approaches[205].

Such interactions are usually not stable in experimental conditions, thus complicating co-purification of the interacting partners. Similarly, some interactions occur only in certain tissues and may be missed by experimental approaches using the wrong tissue or cell line[130].

Experimental Methods for PPI Determination

A multitude of experimental methods have been developed to characterize PPIs. Yeast two-hybrid (Y2H) has been shown capable of identifying binary interactions between proteins at reasonable accuracy[355]. In such studies yeast cells are transfected with a *bait* and a *prey* plasmid. The bait is the protein of interest fused with a DNA-binding domain of a transcription factor, while the prey confers a potential interacting protein fused to an activation domain of the transcription factor. If bait and prey interact, the transcription of reporter genes occurs, so that the presence of gene products resultant of the reporter gene expression serves as indicator of the interaction[456]. Other biochemical and mass spectrometry (MS) approaches in which the protein complexes are first purified through immunoprecipitation (IP-MS)[148] or affinity purification based on epitope-tagged baits (AP-MS)[172,403] have been used for proteome-wide experimental elucidation of the PPI network. However, none of these methods give the three-dimensional structure of the protein complex. To yield these information experimentally, more complex experiments, such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or high-resolution cryo-electron microscopy (cryo-EM) are required[309].

PPI Prediction

Due to the difficulty of experimental PPI elucidation, computational methods have been developed that can be used to characterize different aspects of protein interactions (recently reviewed by Keskin *et al.*[205]). These methods have previously demonstrated their utility for coarse-grained interaction prediction, while residue-level prediction was less successful[136]. Methods utilizing global criteria of genome structure make use of data in genomic databases and can be grouped into gene neighborhood-based, fusion-based, interaction homolog search-based, phylogenetic similarity-based, coexpression-based and network topology-based approaches[205].

Particularly in bacteria, interacting proteins tend to be close on the genome, which led to the development of methods that search for co-localized genes (e.g., on the same operon in bacteria[71,116]). Furthermore, some interacting proteins can be found in a fused form in some organisms. Some computational approaches search for such examples and use these to identify potentially interacting proteins in other organisms ("Rosetta Stone")[264,265]. Similarly, proteins that are known to interact in one species are likely to also have a conserved interaction in their homologs in other species giving rise to methods that mine

databases for such interaction homologs. Due to the assumption of interaction conservation, such methods work primarily for protein complexes that are obligate[205]. Similarities between the expression profile of genes[418] or their evolutionary record, e.g., based on the structure of their phylogenetic trees[196,328], have also successfully been employed to identify PPIs. These coarse-grained methods can be combined with additional approaches for prediction of binding surface patches at residue-level and docking[205]. Methods that use the full amino-acid sequence of proteins could also be used to predict protein interactions at a high resolution. Here, co-evolution based methods using *mutual information* (MI) between residues of two proteins have been proposed in the early 1990s[119]. However, due to limitations of the employed statistical method and the large number of sequences required for their successful use, they have only recently seen further development and application.

2.4.2 The Human Interactome

The set of all protein interactions in an organism is called *interactome*. For humans, the interactome is estimated to consist of up to 650,000 binary interactions (0.2% of the minimally $\sim 20,000 \times 20,000$ possible pairwise interactions if each gene only coded for one protein)[408] with a variety of interface types[5]. However, we still know only a small fraction of the interacting pairs and types[5,279].

While many anecdotal reports of PPIs can be found in the literature, so far only small sections have been charted systematically in larger projects. These include the probing of 13,000 genes using Y2H[355] and the MS-based probing of 7,668 genes[172]. These studies found that each protein node in the interaction network is on average connected to three other proteins[172,355]. Sub-networks exist, however, in which nodes are more connected to each other (*modules*) and a relatively small number of genes have many more connections to other genes than the rest (*hubs*)[18]. Still incomplete, such reference maps of the human interactome have already been used to study disease mechanisms[18,431], find novel drug targets[277] and analyze the effects of gene mutations[130,153].

To facilitate the use of these reference maps for computational analyses, database resources exist. The molecular interaction (MINT) database, for example, collates experimentally verified PPIs from literature[240] and covers 25,000 interactors across 650 species including 8,000 human proteins and the Database of Interacting Proteins (DIP)[364] comprises 29,000 proteins across 850 species including 9,100 interactions of 5,000 human proteins (as of July 2018). Consortia have further compiled meta-databases from such primary databases, integrating experimental and literature sources. Widely used examples include STRING[182,183], IMEx[315], and GeneMania[440]. Some resources further incorporate knowledge about functional PPI modules (i.e., pathways) including KEGG[198] and PathwayCommons[50].

2.4.3 The Role of the Human Interactome in Disease

The dysfunction of cellular processes due to the disruption of PPIs is often implicated in the emergence of disease. The understanding of protein interactions in healthy and diseased cells can thus give valuable, molecular-level information about the links between proteins, genetic variants and diseases[205].

One of the first studies that explored the human disease network was published 2007 by Goh *et al.*[121] where the authors conclude that complex diseases can be caused by different mutations in a common functional module that perturbs cellular function as "genes that contribute to a common disorder (i) show an increased tendency for their products to interact with each other through protein-protein interactions, (ii) have a tendency to be expressed together in specific tissues, (iii) tend to display high co-expression levels, (iv) exhibit synchronized expression as a group, and (v) tend to share Gene Ontology (GO) terms"[121].

Disease Modules in the Interactome

Genes connected to the same disease tend to cluster in the PPI network[121,279]. This clustering is coupled to an increase in biological similarity of the disease genes according to the GO terms for biological processes, molecular function, and cellular localization and could be used in the target identification step of the drug discovery process[279].

The small fraction of diseases that have overlapping disease neighborhoods, however, show a statistically higher GO annotation similarity, coexpression correlation, symptom similarity (from medical records) and relative risk of comorbidity (from disease history of 30 million individuals)[235,323]. This implies that the network-based distance between diseases correlates with pathobiological and clinical similarity and distances between disease modules have predictive power for identifying overlapping disease pairs that also share high levels of comorbidity and biological mechanism (shared pathways etc.)[279].

2.4.4 Network Connection Between Drug Targets and Disease Genes

Only a small proportion of validated disease genes also act as targets for approved drugs[470]. Retrospective studies have, however, shown that compounds that are effective against a protein whose involvement in the disease has been established, have a higher likelihood of gaining FDA approval[292]. The minimum shortest path in the PPI network for disease-drug pairs can be associated with their palliative or etiology-based use: compounds whose shortest path to the disease gene are shorter than expected at random are more likely to treat disease-causation rather than symptoms[470]. This was explained through one or more of the following factors: 1) direct druggability of disease genes (e.g., because they are aberrant membrane-bound receptors that can be targeted by substrate-mimicking molecules), 2)

a rather thorough understanding of etiology or 3) targeted design of therapeutics (e.g., inhibitors of hyperactive oncogenes in cancer)[470].

2.5 Drug Repurposing

Approximately 10-14% of the human proteome is estimated to be druggable and between 1,500 and 3,500 proteins are estimated to be directly involved in disease[163,314]. Given that roughly the same number of genes are already used as drug targets[314,318], novel strategies for finding new drugs targeting previously untreated diseases have to be devised. As stated by Nobel laureate Sir James Black, *"The most fruitful basis of the discovery of a new drug is to start with an old drug"*[348]. Recycling existing compounds for novel indications appears attractive because their pharmacological profile and safety in humans are already established[257]. These compounds can be the starting point for the conventional drug discovery pipeline with additional efforts to optimize their side activities using experimental[451,452] and computational[24] methods. While this approach may circumvent the need for the initial steps in the pharmaceutical discovery process (Figure 2.2), relatively few systematic repurposing programs exist so far in practice, and the area remains of high research interest[325].

2.5.1 Repurposing Using Classical Drug Discovery Methods

A straightforward approach for finding additional therapeutic indications for approved drugs — and late-stage candidates that may have failed for their primary indication — is to use them as libraries in chemical screens, either *in vitro* or *in silico*:

Compounds with a high similarity with respect to their chemical structure are often also similar in their properties (*similar property principle*), including activity[190]. Searching a molecule library for compounds similar to those showing initial activity against a target is thus a common strategy in drug discovery. By using only existing drug molecules as the search library, the same ligand-based computational chemistry strategies can be applied in repurposing[203,204,242,322].

Protein structure-based drug discovery approaches such as molecular docking can similarly be used with only known drugs as the virtual screening libraries[146,207,451,452,473]. This method was applied, for example, for the identification and experimental validation of synapsin-I as a target for staurosporine[73]. The major disadvantage of protein structure-based approaches is the dependence on structural data either of the protein-ligand complex or at least of the target. Furthermore, one has to acknowledge that even though molecular docking can adequately reproduce the binding pose of a ligand in the binding site, docking scores often do not correlate with measured binding affinities[336,342,399,400].

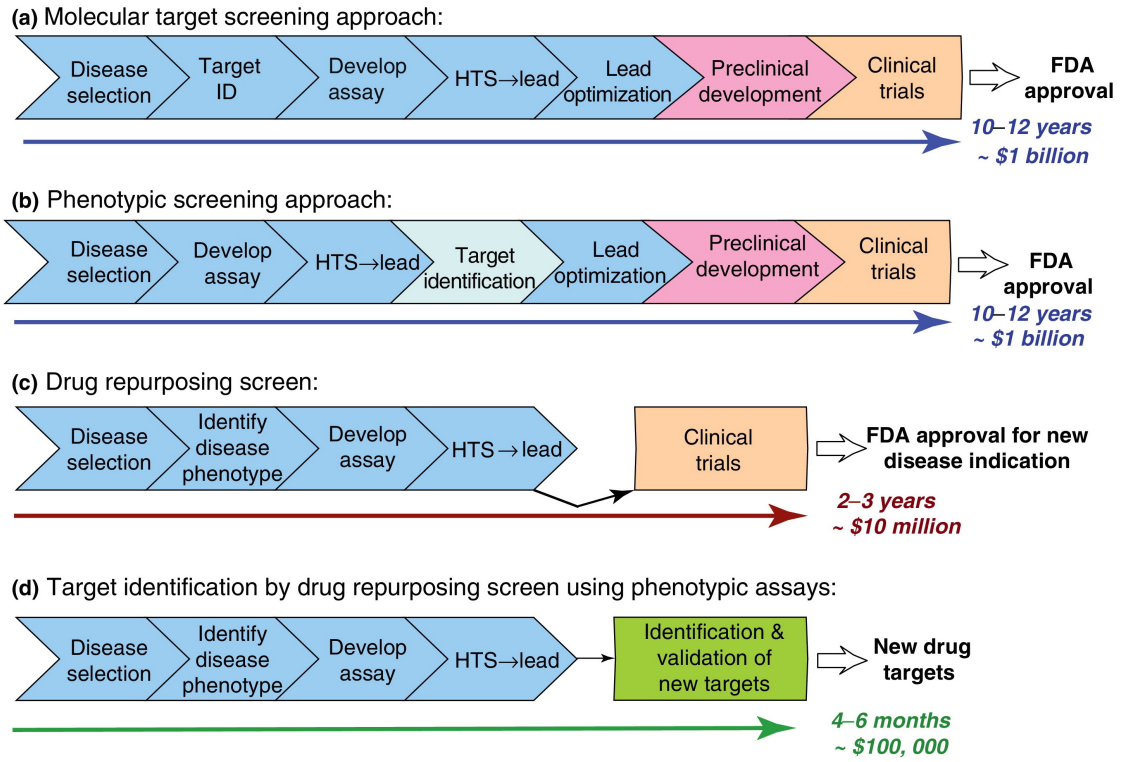


Figure 2.2: Drug discovery pipeline with repurposing. By using repurposing and phenotypic screening, the average development cycle of a new drug can be shortened from 10-12 years through conventional targeted or phenotypic approaches (a, b) to a few years. (c) Using established compounds in phenotypic repurposing screens, the clinical development phase for showing the drug's efficacy for the novel indication can be drastically shortened. (d) Repurposing screens can further be used to identify novel disease targets. Here, the activity of a compound with known molecular target on a disease phenotype can point towards novel disease targets. Reprinted from Zheng *et al.*[479], with permission from Elsevier.

Phenotypic screening is one of the main drivers for experimentally identifying new lead compounds[91,412]. Especially phenotypic animal models provide useful model systems for the disease and results are clinically translatable[367,479]. Using such an experimental approach with approved drugs minimizes the need for further validation after screening, saving development time and cost (Figure 2.2).

2.5.2 Genetic Variation- and Gene Expression-Based Repurposing

The involvement of a gene/protein in the disease etiology is crucial (but not sufficient) to be considered a potential drug target[163] and modern genomic technologies have been widely used for target identification[254]. If the genetic basis of the disease is known, this knowledge can also be used to find existing drugs that act on these proteins: for a genetic

disease caused by the LoF (or GoF) mutation of a gene, the treatment with an agonistic (or antagonistic) drug may be a potential strategy[39,349,365,439].

There are some pitfalls with the approach, requiring further development. The main point to consider in predictions is the directionality of a mutation effect (LoF vs. GoF) and drug action (agonist or antagonist) — both often not recorded in the literature[439]. Other problems arise from the quality of the included data: in GWAS many variants associated with a phenotype do not fall in a protein-coding region complicating their interpretation[63,243,349,365]. Even when using curated databases such as OMIM, the data can be polluted by non-causal gene-disease links[236], stemming from the era before whole genome sequencing (WGS) when the genetic characterization of a disease was expensive, and often only a small set of genes was explored[286].

Compared to full genetic profiles of patients, it is much easier and cheaper to obtain data on gene expression changes in diseases and on cellular effects of drug perturbation through microarrays and RNA sequencing. By comparing the gene expression profile (GEP) induced by drug treatment with that of the diseased tissue, one can search for matching pairs of cellular perturbation[169].

The most famous example of this approach is the *Connectivity Map* (CMap) — a compendium of drug perturbation GEPs in a set of cell lines[224,225] which has been used for the development of several repurposing approaches[175–177,184,206,273,394]. Several of the strongest predictions could later be validated including topiramate for inflammatory bowel disease[88], ursolic acid in muscle atrophy[221] and the HDAC inhibitor vorinostat for gastric cancer[61].

While disease GEP signatures are shown to be robust across tissues[87], this does not necessarily transfer to the cell lines used in the main public databases[81]. Additionally, while many FDA-approved drugs now have established GEPs, such profiles often exist only for a single drug dose. Predictive power despite those drawbacks could, however, be overcome by the creation of cell type-specific drug perturbation profiles[222] and the inclusion of background gene expression networks[417].

2.5.3 Network-Based Repurposing

The guilt-by-association (GBA) approaches is based on the hypothesis that if two diseases can be modulated by the same drugs, then their underlying mechanisms may be similar. Thus, if two diseases share a large number of drugs, those drugs that are currently only used in one of the two indications, may be repurposed to the other[59].

This approach has inspired the development of a multitude of different networks containing various node and edge types[166,238,249,271,272,344,435,466] where paths through the network are used to identify new drug-disease associations (see Figure 2.3 for an example). Some

of these methods are built on heterogeneous information networks[154], while others follow a semantics-based approach[67].

An alternative approach is leveraging the modularity of the protein-protein network (see also Chapter 2.4.3) in a pathway-agnostic manner to identify drug-disease pairs through drug-gene-disease co-modules. In this case, modules are identified in the gene network through partition methods on gene closeness data[137,477].

Methods for drug-target or disease-gene predictions based on network knowledge can also be used in repurposing studies, as they can be easily integrated with the missing disease-gene or drug-target links if the molecular target is established[56,78,99,300,463].

Network-based approaches have the advantage that they allow the prediction of drug MoA on top of the repositioning. Once created, the networks can also easily be extended for a disease or drug of interest with a known target gene. However, the construction of a reliable network is complicated by a plethora of data sources with partially overlapping entries and little quantitative information about the interaction strength. For the protein-protein interactome, only a small proportion is currently known[279], but noise is introduced by predicted connections or by those obtained in high-throughput setups. Thus each additional integrated data source may add false associations, resulting in different performances of the same method depending on the integrated database[166].

2.5.4 Hybrid and Machine Learning-Based Repurposing

While some of the network-based methods are *de facto* machine learning (ML)-based, using the network or structures derived from it as features, there are also other methods that combine multiple types of data and prior knowledge for supervised learning approaches. The problem is then usually formulated as a binary classification problem to predict associated drug-disease pairs. The two gold standard methods are PREDICT[127] and PreDR[438]. While PREDICT uses a logistic regression classifier that employs multiple similarity measures for drug-drug and disease-disease relationships as features[127], PreDR employs a kernel-based approach on the same dataset[438].

Two approaches treating the problem of drug-disease association prediction as supervised network inference have been shown to outperform PREDICT and PreDR[45,179,369]. Here, the features for drugs contain chemical fingerprints or a phenotypic profile while diseases are encoded using molecular features (disease genes, diagnostic biomarkers, pathways and environmental factors). Then either logistic regression[179,369], template matching[369] or a Random Forest (RF) are employed for classification[45].

Many of the proposed ML methods require prior knowledge encoded as feature vectors and result in a black-box model. Another limitation of these approaches is that they often require at least some information about the included feature types for a drug and the

diseases[45]. Furthermore, the data in the training sets are often unbalanced, and even though many groups try to reduce redundancy between training and test data, the included data sources are frequently not independent.

In their performance, the ML-based methods tend to outperform all other proposed repurposing approaches[45,179,369]. Approaches that infer edges in the drug-gene-disease network, on the other hand, result in predictions with higher interpretability, than regression-based predictions[369], but fail to reach comparable prediction performance.

2.6 Phenotype Inference using Evolutionary Sequence Records

Genome sequencing has become a popular method to study the effects elicited by specific genetic variants on the phenotype. Apart from identifying statistical associations between a variant and a particular phenotype in GWAS, computational approaches have been developed to utilize the data from such projects. Here, the availability of genetic information of a protein of interest across multiple species was shown to be predictive of several phenotypes including the 3D structure of proteins[33,268,419,475] and RNA[448], as well as the overall fitness of the organism[161].

2.6.1 Computational Methods to Predict Protein 3D Structure

Consideration of a protein's 3D structure has been shown to be helpful in the analysis of its functionality. As discussed in previous sections, knowledge of the 3D structure of a protein can then be used to study effects of genetic variation, design new drugs (Chapter 2.2.2), and study protein interactions (Chapter 2.4). Many proteins do not have a resolved structure yet[333], thus computational methods that either rely on a template for modeling the structure (*comparative modeling*) or predict the structure *de novo* have been developed.

Comparative modeling of 3D structures is based on the fact that protein structure is more conserved then amino acid sequence with proteins sharing the same fold despite only approximately 30% pairwise sequence identity[60]. If a solved homologous structure is available, spatial coordinates can be transferred from this template structure to the target protein using the aligned sequences of target and template[13].

If no template structure exists, a 3D model of the protein can be created *ab initio*. For smaller proteins (less than 100 amino acids) the simulation of the *in vivo* folding process based on MD simulations and a force field yield reasonable 3D models given the appropriate compute power[245,386]. For larger proteins, a starting structure is assembled from fragments with known 3D structure that have been identified through a coarse-grained search of the conformational space. This 3D model is then refined[33,36,475].

Several ML-based approaches have further aimed at learning structural features of sequences based on the solved structure space and utilize those for novel predictions[283,462].

To constrain the conformational search space explored by *de novo* prediction methods, orthogonal data from experiments, such as constraints identified by MS or NMR experiments, can be included in the prediction.

A breakthrough in *de novo* modeling was obtained using correlated mutations in a protein family to infer interacting residues from sequence data alone (Figure 2.4). Compensatory mutations in proteins have been observed early on in the field of sequence analysis and occur if a second residue is required to change after an initial point mutation to rescue the protein's structure or function[6,313]. From this, the idea of using correlated mutations to predict contacts in a protein arose in the mid-1990s[119,301,390,413,414]. However, while observing correlated mutations at sites known to be close in the structural context is relatively trivial, the inverse of inferring actual contacts from a multiple sequence alignment is a challenging problem.

2.6.2 Co-Evolution-Based Prediction of Protein Structure

The concept of correlated mutation can be modeled, statistically, as the dependency of amino acids between columns in a multiple sequence alignment (MSA). Initially, primarily pair-independent measures were used to infer co-varying residues, including MI between the distributions of amino acids occurring in pairs of columns i, j in the MSA (Figure 2.4). The counts of each letter in the alphabet of $\Sigma = \{1, ..., q\}$ with $q = 21$ (representing the 20 proteinogenic amino acids plus the gap character) can be converted to a frequency vector f_i for position i in the alignment. Observations of pairs of amino acids at positions i, j can be similarly represented as a frequency matrix $f_{i,j}$ with each cell in the matrix representing a particular combination of characters $\sigma_i \in \Sigma$.

For a multiple sequence alignment with B sequences of a protein $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_N)$ of length N, this can be formally described as

$$\begin{aligned} f_i(k) &= \frac{1}{B} \sum_{b=1}^{B} \delta(\sigma_i^b, k), \\ f_{i,j}(k,l) &= \frac{1}{B} \sum_{b=1}^{B} \delta(\sigma_i^b, k) \delta(\sigma_j^b, l) \end{aligned} \tag{2.3}$$

The MSA in this case is defined as an integer array $\{\boldsymbol{\sigma}^b\}_{b=1}^{B}$ with a column per position in the protein sequence and a row for each sequence in the alignment. The Kronecker symbol $\delta(a, b)$ takes value 1 if $a = b$ and 0 otherwise.

The MI of two sites in the alignment can be calculated from the frequencies of single sites and pairs of sites as

$$MI_{ij} = \sum_{k,l=1}^{21} f_{i,j}(k,l) \ln \frac{f_{i,j}(k,l)}{f_i(k) f_j(l)} \tag{2.4}$$

Those pairs (i, j) with high MI are then predicted as contacts in the 3D structure.

Many of the early papers in this field found, however, that only few of the identified pairs were indeed close in the three-dimensional structure of the protein. This can be attributed to confounding factors such as uneven representation of protein family members, inadequate numbers of sequences in the MSA, phylogenetic effects and indirect and transitive signals obfuscating direct co-evolving signals[42,229,269].

Global maximum-entropy models

Inspired by statistical physics approaches dealing with similar problems in coupled spin systems, global probability models have been recently employed to identify co-evolving residues while removing transitive effects. These models treat pairs of residues as dependent on each other while also including the single-position conservation of each residue[269].

Several such global methods have been proposed in the past years: either based on Bayesian network modeling[42] or entropy maximization under data constraints known as *Potts models* (or *pairwise graphical models*). The latter is often also referred to as *direct coupling analysis* (DCA)[96,269,291].

The Potts model is the least biased probabilistic model capable of reproducing the observed frequencies $f_i(k)$ and $f_{i,j}(k,l)$ as defined in equation 2.3, while explaining underlying patterns in the data through hidden constraints[95,181,229],

$$\begin{aligned} P(\sigma_i = k) &= \sum_{\boldsymbol{\sigma}, \sigma_i = k} P(\boldsymbol{\sigma}) = f_i(k), \\ P(\sigma_i = k, \sigma_j = l) &= \sum_{\boldsymbol{\sigma}, \sigma_i = k, \sigma_j = l} P(\boldsymbol{\sigma}) = f_{i,j}(k,l), \end{aligned} \tag{2.5}$$

Under the Potts model, the probability of sequence $\boldsymbol{\sigma}$ is defined as

$$\begin{aligned} P(\boldsymbol{\sigma}|\mathbf{h}, \mathbf{J}) &= \frac{1}{Z(\mathbf{h}, \mathbf{J})} \exp(H(\boldsymbol{\sigma})) \\ H(\boldsymbol{\sigma}) &= \underbrace{\sum_{i=1}^{N} \mathbf{h}_i(\sigma_i)}_{\text{conservation}} + \underbrace{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{J}_{ij}(\sigma_i, \sigma_j)}_{\text{co-conservation}} \end{aligned} \tag{2.6}$$

Where the partition function Z normalizes distribution and $H(\boldsymbol{\sigma})$ describes the statistical energy of the system — defined as the sum of all its single-site constraints $\mathbf{h}$ and pairwise coupling constraints $\mathbf{J}_{ij}$.

The approach of inferring the parameters $\{\mathbf{h}, \mathbf{J}\}$ based on the constraints defined in equation 2.5 from a set of independent equilibrium configurations $\{\boldsymbol{\sigma}^b\}_{b=1}^{B}$ of the model observed in reality is called the *inverse Potts problem*. It would usually be solved by maximum likelihood estimation, optimizing the likelihood function

$$\begin{aligned}\mathcal{L}(\mathbf{h}, \mathbf{J}) &= \prod_{\sigma \in B} P(\sigma | \mathbf{h}, \mathbf{J}) \\ &= \prod_{\sigma \in B} \frac{1}{Z(\mathbf{h}, \mathbf{J})} \exp\left(\sum_{i=1}^{N} \mathbf{h}_i(\sigma_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{J}_{ij}(\sigma_i, \sigma_j) \right)\end{aligned} \tag{2.7}$$

so that the probability of observing the sequence $\boldsymbol{\sigma}$ in the alignment B is maximized. This process requires the calculation of the partition function Z

$$\begin{aligned}Z(\mathbf{h}, \mathbf{J}) &= \sum_{\sigma_1=1}^{21} ... \sum_{\sigma_N}^{21} P(\sigma_1, ..., \sigma_N | \mathbf{h}, \mathbf{J}) \\ &= \sum_{\sigma_1=1}^{21} ... \sum_{\sigma_N}^{21} \exp\left(\sum_{i=1}^{N} \mathbf{h}_i(\sigma_i) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{J}_{ij}(\sigma_i, \sigma_j) \right)\end{aligned} \tag{2.8}$$

for any configuration $\{\mathbf{h}, \mathbf{J}\}$. This partition function sums over all system configurations, i.e., possible sequences, and ensures that $P(\boldsymbol{\sigma})$ is a valid probability distribution in which the probabilities of all sequences that could be generated sum up to one. This corresponds to the calculation of 21^N possibilities for a sequence of length N and 21 possible states at each position before summing them up. Solving systems of the size usually encountered in protein structure determination is thus computationally intractable.

Therefore mathematical approaches have been devised that can approximate the parameters in the inverse Potts problem at faster compute times. Some of these have been successfully implemented for protein contact prediction, including Monte Carlo optimization[229], naive mean-field inversion (NMFI)[268,291], and *pseudo-likelihood maximization* (PLM) which has shown to further improve accuracy of the inferred contacts[16,95,96].

The general concept of the DCA approach to obtain interaction scores from a multiple sequence alignment is illustrated in Figure 2.5.

Sequence re-weighting

One of the main assumptions used in maximum likelihood inference of Potts models is that the B sample configurations (i.e., sequences) are independent and identically distributed (i.i.d). Sequences in an MSA do not fulfill these assumptions due to sampling bias and phylogeny. The information content of each sample does therefore not match that of a truly independent sample. One way to mitigate this sample bias is through re-weighting of the B configurations used in parameter estimation to penalize samples with very high similarity. A weight w_b can be added to each sequence $\boldsymbol{\sigma}^b$ that depends on its similarity to the others.

In DCA the similarity measure is based on the fraction of identical amino acids in two sequences $\boldsymbol{\sigma}^a$ and $\boldsymbol{\sigma}^b$. Using a predefined similarity threshold x, one can compute the number of similar sequences m_b to sequence $\boldsymbol{\sigma}^b$. The weight w_b of $\boldsymbol{\sigma}^b$ is then defined as $w_b = \frac{1}{m_b}$.

This weighting procedure further allows to estimate the number of "effective" samples as $N_{\text{eff}} = \sum_{b=1}^{B} w_b$ which is a better estimator of the amount of available sample information.

Regularization

Since the number of parameters of the model outnumbers the amount of available sequences for most protein families for model inference, measures such as *regularization* help to avoid overfitting. Here, a penality term is added to the objective function before optimizing,

$$\arg\max_{\mathbf{h},\mathbf{J}}(log\mathcal{L}(\mathbf{h},\mathbf{J}) - \mathcal{R}(\mathbf{h},\mathbf{J})) \tag{2.9}$$

where the optimization now consists of a trade-off between maximizing the probability of the data (log of the pseudo-likelihood $\mathcal{L}(\mathbf{h},\mathbf{J})$) and minimizing the complexity of the model, as described by the regularisation term $\mathcal{R}(\mathbf{h},\mathbf{J})$.

Both l_1 [15,192] and l_2 regularization have been employed in protein structure prediction, with l_2 being more commonly used [95,197,268,419,448],

$$\mathcal{R}_{l_2}(\mathbf{h},\mathbf{J}) = \lambda_h \sum_{i=1}^{N} ||\mathbf{h}_i||_2^2 + \lambda_J \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} ||\mathbf{J}_{ij}||_2^2 \tag{2.10}$$

Here, the regularization parameters λ_h and λ_J determine the strength of regularization for each parameter type.

Co-evolution-based interaction scores

Scalar scores for the coupling strength of residue pairs in the protein can be calculated by summarizing the $\mathbf{J}_{ij}$ matrices. The cell values inferred for each pair-matrix $\mathbf{J}_{ij}$ correspond

to the preferences for observing a particular amino acid combination σ_k, σ_l with $k, l \in q$. If the cells in $\mathbf{J}_{ij}$ show very strong differences in values, this indicates that this pair of residues in the sequence is evolutionarily constrained.

In PLM-based DCA (plmDCA), the coupling scores are calculated as the Frobenius norm (FN) of the transformed $\mathbf{J}_{ij}$ matrices. By shifting the coupling matrix in the so-called *zero-sum gauge*, their rows and columns means are centered around zero, as

$$\mathbf{J}'_{ij}(k,l) = \mathbf{J}_{ij}(k,l) - \mathbf{J}_{ij}(\cdot,l) - \mathbf{J}_{ij}(k,\cdot) + \mathbf{J}_{ij}(\cdot,\cdot) \tag{2.11}$$

where $\cdot$ denotes the average across the matrix row/column entries. $\mathbf{J}'_{ij}$ is then summarized by calculating the FN

$$\mathcal{S}_{ij}^{\mathrm{FN}} = ||\mathbf{J}_{ij}||_2 = \sqrt{\sum_{k}^{q}\sum_{l}^{q} \mathbf{J}'_{ij}(k,l)^2} \tag{2.12}$$

which sums across all 21×21 combinations of amino acids k, l in positions i and j.

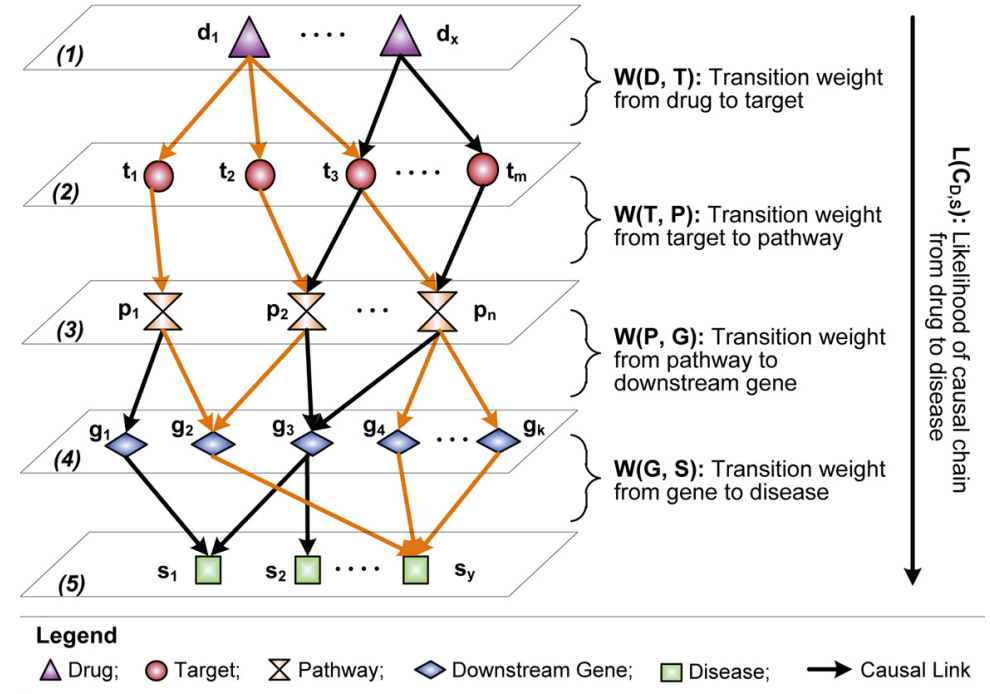
Due to confounding factors such as phylogeny and insufficient sampling, residue pairs with high FN scores still suffer from low accuracy despite sample re-weighting and regularization. This problem can partially be mitigated by an additional correction applied to the matrix of all FN scores for the protein called *average product correction* (APC)[89]. Based on the assumption that each individual site should only be sparsely coupled, APC approximates the coupling background and removes it from the score. The final coupling score (*evolutionary coupling (EC) score*) $\mathcal{S}_{ij}^{\mathrm{CN}}$ of a site pair (i, j) is thus calculated as

$$\mathcal{S}_{ij}^{\mathrm{CN}} = \mathcal{S}_{ij}^{\mathrm{FN}} - \frac{\mathcal{S}_{i\cdot}^{\mathrm{FN}}\mathcal{S}_{\cdot j}^{\mathrm{FN}}}{\mathcal{S}_{\cdot\cdot}^{\mathrm{FN}}} \tag{2.13}$$

where $\cdot$ denotes the row/column average so that $\mathcal{S}_{i\cdot}^{\mathrm{FN}}$, for example, represents the average FN score of all site pairs involving sequence position i.

An $N \times N$ EC matrix can then be constructed from the $\mathcal{S}_{ij}^{\mathrm{CN}}$ scores and the cells with the highest scores correspond to the site pairs that are strongly coupled in the evolutionary record.

A



B

Drug	Selected Causal Chains Supporting This Prediction ([D] Drug; [T] Target; [P] Pathway; [G] Downstream Gene; [S] Disease)
Anakinra (score=5.29)	[D] Anakinra → [T] IL1R → [P] Osteoclast differentiation → [G] NCF4 → [S] CD
	[D] Anakinra→ [T] IL1R→ [P] Amoebiasis→ [G] FN1→ [S] CD
Rifabutin (score=4.30)	[D] Rifabutin → [T] HSP90AA1 → [P] NOD-like receptor signaling pathway → [G] IL6, TNF, NLRP3, NOD2 → [S] CD
	[D] Rifabutin → [T] HSP90AA1→ [P] Glucocorticoid receptor regulatory network → [G] IL6 → [S] CD
Nedocromil (score=4.00)	[D] Nedocromil → [T] HSP90AA1 → [P] NOD-like receptor signaling pathway → [G] IL6, TNF, NLRP3, NOD2 → [S] CD
	[D] Nedocromil → [T] FPR1 → [P] Staphylococcus aureus infection → [G] FPR2, IL10 → [S] CD

Figure 2.3: Schematic setup of CauseNet. A) The causal inference chain passes through several layers of distinct node types: drug (d) → target gene (t) → pathway (p) → downstream gene (g) → disease (s). Transition weights w between the different nodes are learned from a set of "treatment enriched chains" of known disease treatments. B) examples of drug-disease predictions and the causal chain supporting said predictions. Reprinted from Li *et al.*[238], available under Creative Commons Attribution 2.0 license.

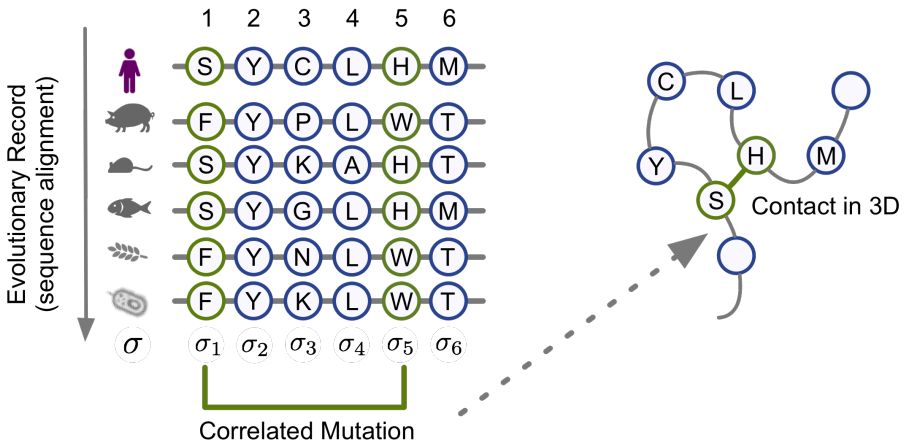


Figure 2.4: Compensatory mutations in proteins can be observed for residue pairs whose interaction is relevant for protein structure or function. Co-conservation can thus be used to infer strongly coupled residue pairs and use this to predict the structure of a protein *de novo*. Figure adapted from Marks *et al.*[267], available under Creative Commons Attribution 2.0 license.

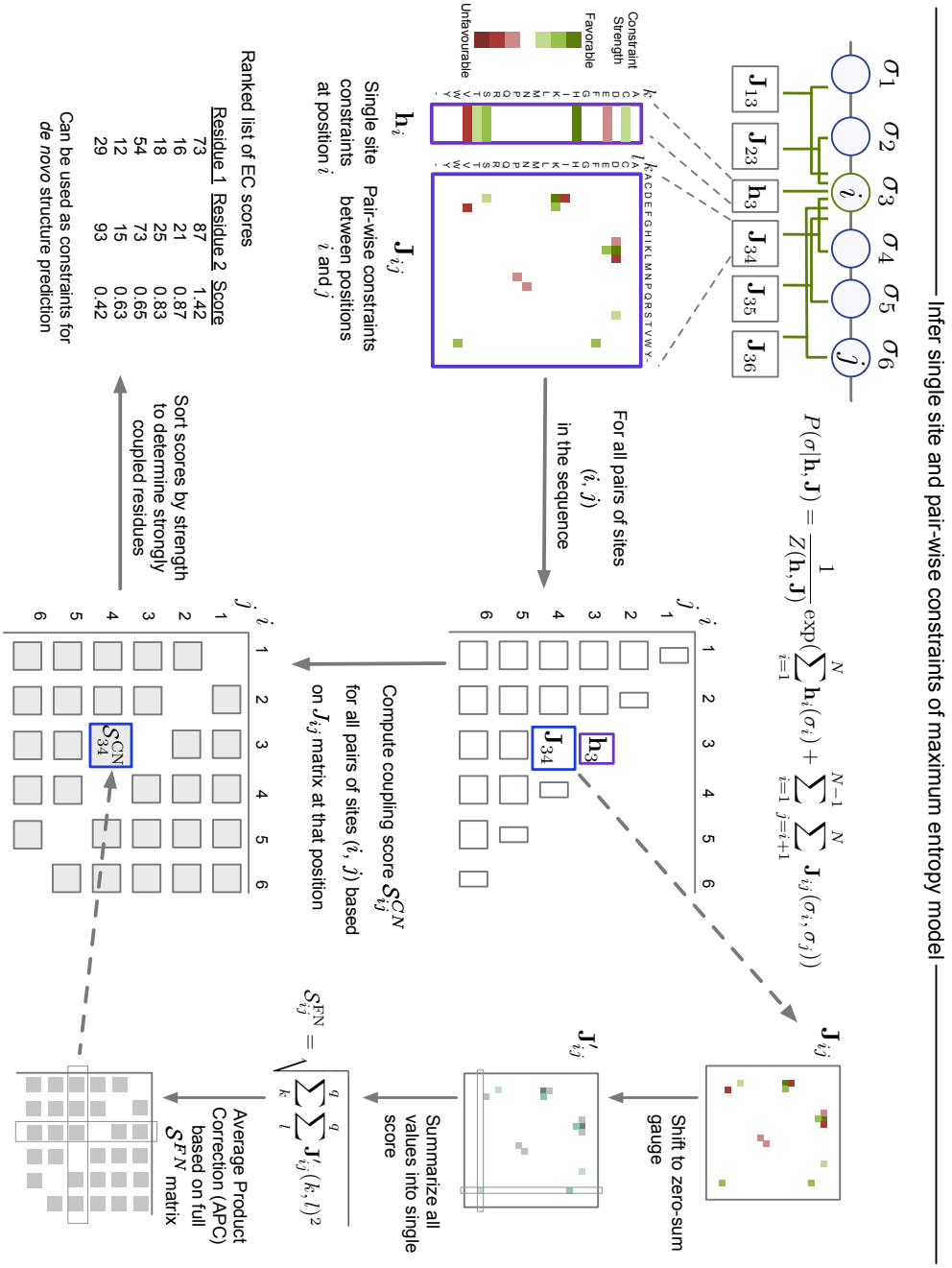


Figure 2.5: Predicting coupling scores from sequence alignments starting from sequence alignments. From the sequence alignment for a sequence σ, a global maximum entropy model is inferred, constrained by the observed conservation of amino acids at single sites and co-conservation at pair-sites for σ. Coupling scores for all pairs of residues l and k (EC scores) are obtained by summarizing the $\mathbf{J}_{ij}$ matrix for site-pairs using the Frobenius norm, followed by a correction for confounding bias (using average product correction, APC). Figure inspired by Hopf *et al.*[161].

Chapter 3

Predicting Protein Interactions using Co-evolution

Parts of this chapter have also been published in the following article:

Sequence co-evolution gives 3D contacts and structures of protein complexes.[160]

3.1 Introduction

The cell is a crowded place in which proteins constantly interact with each other[270]. For most proteins this interaction is essential to fulfill their cellular functions, and while some form stable long-lived complexes, short-lived transient interactions are not rare, especially in cellular signaling cascades[335,416]. The detection of protein interactions has seen unprecedented progress in recent years through the development and improvement of several high throughput approaches, e.g., AP-MS, a method combining affinity purification of epitope-tagged proteins with MS[172], but particularly transient interactions remain difficult to observe experimentally.

Especially the determination of the 3D structure of protein complexes remains challenging, despite recent advances in structural biology[12,58,165,478]. We thus still have little or no 3D information for 80% of the currently known protein interactions in bacteria, yeast and mammals. This amounts to a total of ∼30,000/∼6,000 incompletely characterized interactions in *Homo sapiens* (*H. sapiens*) and *Escherichia coli* (*E. coli*), respectively[294,346]. While 3D information is not required to analyze global properties of cellular systems, it is necessary for residue-level studies for drug development and the study of phenotype-altering mutations.

To address this knowledge gap between PPIs and their 3D structure, all types of available structural information of the complex subunits (including low resolution and

homology models) can be combined with force-field driven (RosettaDock), data-driven, and cross-linking approaches[54,82,200,216,217,219,354,373,409,430,443] (see Chapter 2.4 for a more detailed review of the methodology) to obtain 3D models of the complexes.

In addition to structural data, the use of sequence data has been explored in the past. Particularly, mutations in one protein partner could result in compensatory mutations in the other partner to rescue residue interactions across the PPI interface. MI methods to identify such correlated mutations in the sequence alignments of interacting proteins were proposed for the first time two decades ago[119,328–330]. These models could not discriminate between transitive (indirect) and direct contacts, resulting in reduced prediction accuracy[268]. Data about compensatory mutations were also incorporated into machine learning-based scoring functions for interface residues[9,10,101], but suffered from the same problem of identifying transitive interaction pairs. Recently, this problem was overcome for protein contact prediction within proteins using direct coupling analysis. Employing a global maximum entropy model, it was possible to 1) accurately infer residue contacts, and 2) predict protein 3D structures *de novo*[159,268,291].

We thus hypothesized that such a generalized global statistical model employed for monomer prediction would also be capable of inferring direct co-evolving residues in PPIs to: 1) identify the interaction interface at residue-level resolutions, 2) predict the protein partners in large protein complexes or even across a whole genome.

Goals of the Project

The goal of this project was to develop a method that infers residues interacting across a protein-protein interface based on global maximum entropy methods for direct coupling analysis. It was further the goal to show that the method is capable of identifying the physically interacting subunits in large protein complexes, and that inferred residue pairs can be used to produce 3D structures of protein complexes.

We evaluated the approach on a catalog of known PPIs in bacteria as well as large protein complexes, to show that it is possible to use the evolutionary record of proteins to infer interactions.

3.2 Materials and Methods

The proposed method, EVcomplex, identifies interacting residues between two proteins based on a statistical model similar to DCA for structure prediction of protein monomers. The approach consists of several steps (summarized in Figure 3.1):

Creation of a joint sequence alignment. First, the evolutionary record of both proteins is retrieved by creating multiple sequence alignments of the proteins. Putatively

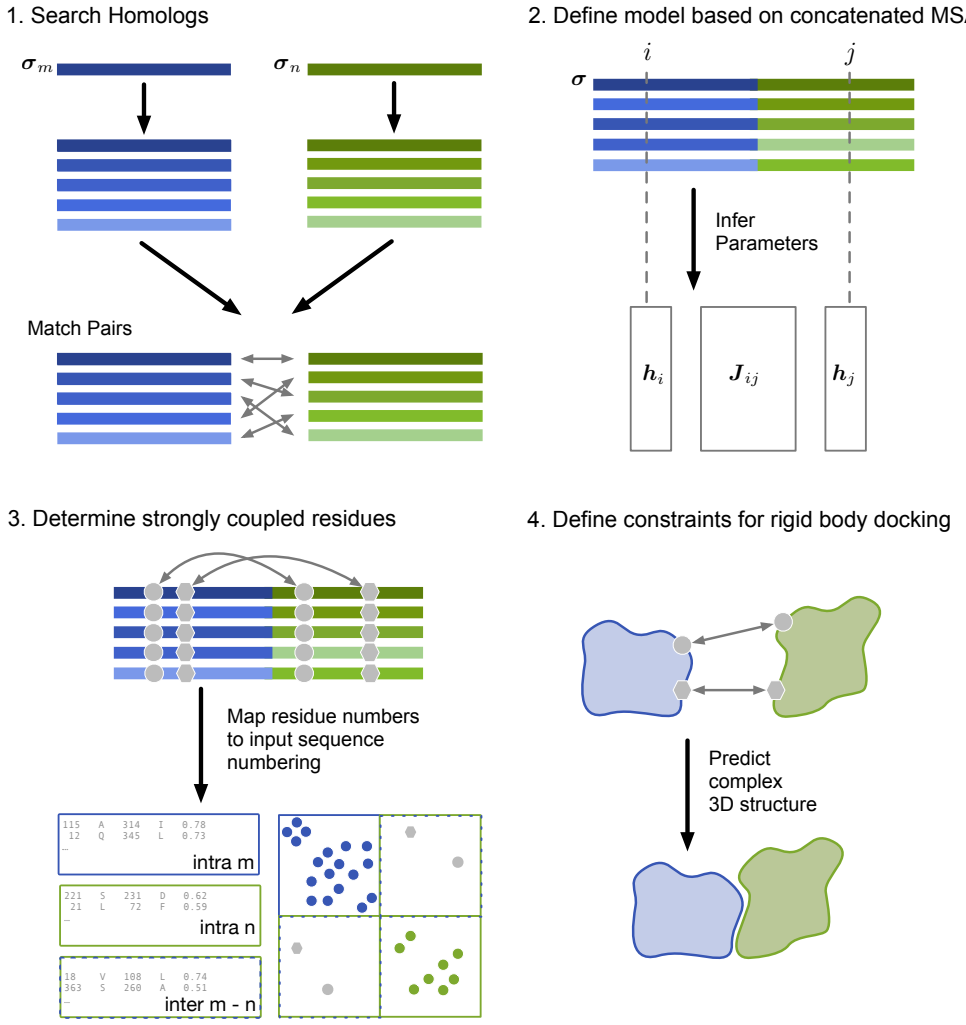


Figure 3.1: Contact prediction using evolutionary couplings. The evolutionary record of both proteins is retrieved from a large sequence database by creating multiple sequence alignments. Puatatively interacting pairs of sequences have to be identified and concatenated before inferring the co-evolution model using pseudo-likelihood maximisation. The model is summarized to infer particularly strong couplings. These can then be used as constraints for docking protocols.

interacting pairs of sequences are then concatenated so that an overall sequence $\boldsymbol{\sigma}$ is obtained from the two monomer sequences $\boldsymbol{\sigma}_m$ and $\boldsymbol{\sigma}_n$.

Inference of co-evolution model. We employ a Potts model as the global probability model from which interaction strengths between residues are inferred (the details of model and parameter inference using pseudo-likelihood estimation are explained in depth in Chapter 2.6)

Mapping and scoring. After model inference, a re-mapping step then translates the residue indices in the concatenated sequence back to the corresponding indices in $\boldsymbol{\sigma}_m$ and $\boldsymbol{\sigma}_n$. Given the global approach, not only co-evolving residues between the two proteins (inter-contacts) are inferred, but also within each monomer (intra-contacts). Raw coupling scores depend on the alignment length and depth and are further transformed to scale-free *EVcomplex scores* using the method described below to facilitate comparison of predictions between proteins. Based on a benchmark set, it was possible to determine a universal threshold for separating high-confidence scores from noise. Strong inter-protein couplings can then be used as constraints in docking tools such as HADDOCK to predict the 3D structure of the complexes.

3.2.1 Method Details

Monomer sequence search

The first step in the EVcomplex method is the creation of MSAs for the two interacting proteins. A key challenge here is to create sufficiently diverse alignments (evolutionary depth) to facilitate statistical inference while not including proteins that have diverged too far and thus have changed structure or are not involved in the same conserved interaction. Iterative hidden Markov model (HMM)-based homology search methods, in which position-specific residue and transition probabilities are determined from seed alignments, have been shown to be more successful in identifying even distant homologs of a protein, both in respect to accuracy and sensitivity, than other standard sequence search tools[93,353], including BLAST[7,8], the *de facto* standard in sequence search.

jackhmmer allows creating accurate MSAs by searching for homologs using iteratively extended profile HMMs[189]. The tool allows to modulate the evolutionary depth of the created alignment using domain specific inclusion thresholds based on 1) bit score and 2) E-value. The first is solely dependent on the profile HMM and the aligned sequence $\boldsymbol{\sigma}$ and is computed as the log-odds ratio of the probability of $\boldsymbol{\sigma}$ given the profile HMM and the probability of observing $\boldsymbol{\sigma}$ under a null model:

$$S = \log_2 \frac{P(\boldsymbol{\sigma}|\text{HMM})}{P(\boldsymbol{\sigma}|\text{null})}. \tag{3.1}$$

In the case of *jackhmmer* two null models are used, first a one-state HMM describing random sequences as i.i.d. with a pre-defined amino acid composition and secondly an *ad hoc* model to correct for biased composition regions in the target sequence. The bit score thus describes how well the particular sequence matches the HMM compared to the null model: if the score is positive, the sequence hit is better described by the HMM than the random distribution[92].

Based on the bit score, the E-value describes the statistical significance of obtaining a match for a sequence of a particular length and with the specific bit score by calculating the number of matches expected to obtain the particular bit score in a database of random, non-homologous sequences of the same size[105].

While previous studies of co-evolution in proteins usually used E-value thresholds to determine inclusion of a protein sequence in the alignment[120,159], we employ bit score-based membership determination in this project based on the assumption that using the same average sequence similarity per residue (bits/residue) for each protein will result in alignments with comparable evolutionary depth (which is typically not the case for the same E-value). For a protein with sequence length L, we define the inclusion threshold for the domain-specific bit score (–incdomT in *jackhmmer*) as $x \times L$ with x being the expected similarity in bits/residue. While the optimal coverage of the full domain length is reached at different thresholds depending on the protein, we found that $x = 0.5$ serves as a good starting point. For the high-throughput evaluation of our method, we refrained from a protein-specific threshold optimization, choosing a bit score of $0.5 \times L$ as the protein specific inclusion threshold. We want to highlight that varying the expected similarity to obtain wider and narrower alignments can result in improved prediction accuracy and should be evaluated when focusing on a particular protein complex.

Sequence matching

Once the homologs of each protein across species are identified and aligned, the specific interacting protein pairs need to be paired up before identifying co-evolved residues. This step is crucial in the overall process because if non-interacting proteins are combined, the inferred residue pairs will not correspond to co-evolution driven by interaction. However, the specificity and conservation of interactions across orthologous proteins is rarely known, i.e., for a species S we do not know which of the paralogs of protein A interacts with which paralog of protein B. We thus need to revert to approximate approaches to construct paired alignments with sufficient numbers of diverse sequences. In this project, we implement two such approaches, based on different assumptions (Figure 3.2): 1) the paralogs with the highest sequence similarity to the query interact, 2) the paralogs with the closest reciprocal distance on the genome to each other interact.

The most obvious strategy is based on choosing the most similar homologs of both proteins A and B to the query proteins in each species, assuming that the interaction is conserved across orthologs. Due to its low computational cost, our approach is based on the absolute number of identical residues shared between the query proteins as a measure for sequence similarity of the orthologs in each species. Then only the paralogs A* and B* that show the highest percentage of sequence identity to the query proteins A and B

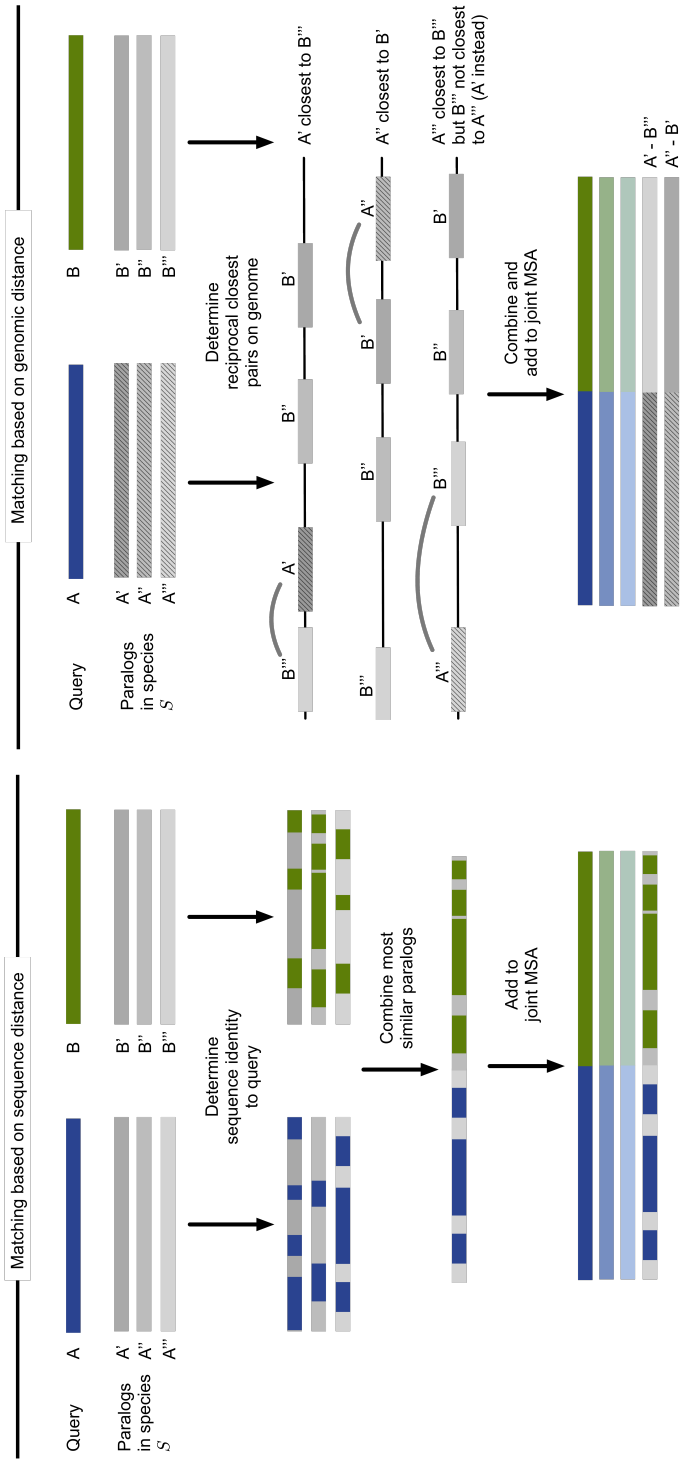


Figure 3.2: Approaches to sequence concatenation. We implemented two different methods for identifying possible pairs of sequences: (Left) among all paralogs in one species, the one with the highest sequence identity is selected as putative ortholog to the query sequence for both proteins A and B; (right) by choosing those proteins that are located with the smallest reciprocal distance on the genome (e.g., on the same operon in bacteria).

are concatenated into a single record of the matched MSA. All other paralogs of A and B are discarded as the optimal way of pairing these further cannot be determined without any other information. The advantage of this method is that it is not restricted to any particular kingdom and can thus theoretically be applied to any protein complex. Since only one pair of sequences is concatenated in each species, the resulting alignment used for parameter inference of the global model may contain too few sequences. This approach is thus not feasible if the protein interaction is only present in a small fraction of species (e.g., only in eukaryotes).

Based on the observation that interacting proteins in bacteria are often proximal on the genome (e.g., same operon), co-evolution studies of the histidine kinase-response regulator complex matched sequences based on operon structure[395,445]. We generalized this approach in order to match more than one paralog per species and thus improve the sample size of sequences used in the parameter inference step. The genomic location of the genes in the alignments are retrieved from their coding sequences (CDS) as matched through the ENA database[320]. For all possible combinations of paralogs to A and B in a species, the genomic distance is computed using their CDS coordinates. Pairs that were not present together on an ENA contig (or whole genome sequence) were excluded. Based on the pairwise distances, we then determine those pairs that are mutually closest to each other, i.e., A' is closest to B' and B' is closest to A' and not another A paralog A". If this criterion is fulfilled, the paralogs are matched and added to the concatenated MSA. To exclude noise introduced by pairs that only had one paralog each and thus were paired even though they do not reside close on the genome, the alignment can be filtered for such distant pairs (e.g., pairs with a genomic sequence distance greater than 1000 nucleotides). While this concatenation approach is capable of including multiple paralogs of a species, which is especially useful in protein complexes known to have multiple copies in each genome, it is unfortunately not applicable to eukaryotic proteins due to the different genomic architecture.

Recently, several alternative algorithms have been proposed, that may help matching multiple paralogs per species without simultaneous co-localization in the genome. These methods are based on the assumption that correctly matched sequences also maximize the inter-protein co-evolution signal and can thus be found through optimizing the co-evolution score while testing all possible combinations of paralogs per species[136]. Especially for proteins unique to only certain branches of the tree of life, such methods may further increase the applicability of co-evolution in protein interaction determination, but the computational complexity of this iterative optimization problem has so far hindered its wide applicability[136]. Given that our work was mainly aimed at proving the general applicability of co-evolution for protein complex determination, such more elaborate concatenation strategies have not been evaluated in our work, but could be included in the future.

After construction of the joint MSA, we employ several post-processing steps to ensure a certain quality of the data included in model inference. These are mainly the exclusion of positions with low sequence coverage (i.e., many gaps) and the clustering of sequences at a certain level of sequence identity (usually in the range of 70 to 90%) to reduce redundancy. Each sequence is then re-weighted in proportion the cluster size to reduce the individual influence a redundant sequence has in the model inference step.

Statistical model

We built our pipeline on the previously established EVfold approach[159,268,404] that is described in detail in Chapter 2.6.

This global maximum entropy model is constrained by the observed conservation of amino acids at single sites and co-conservation at pair-sites for a concatenated sequence $\boldsymbol{\sigma}$ and is defined as

$$P(\boldsymbol{\sigma}|\mathbf{h},\mathbf{J}) = \frac{1}{Z(\mathbf{h},\mathbf{J})}\exp(\sum_{i=1}^{N}\mathbf{h}_i(\sigma_i) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbf{J}_{ij}(\sigma_i,\sigma_j)) \tag{3.2}$$

in which $\mathbf{h}_i$ represents a vector containing parameters of the single site conservation and $\mathbf{J}_{ij}$ a 21×21 parameter matrix constrained by observed pairwise amino-acid frequencies. The $\mathbf{h}$-vectors and $\mathbf{J}$-matrices for all positions i, j in the sequence can then be inferred using optimization approaches. In the context of protein co-evolution, these have mainly been mean-field (MF) approximation[268,291] and pseudo-likelihood estimation (PLM)[15,95].

Due to its higher accuracy of predicted contacts[95], we chose a modified implementation of plmDCA[95] for parameter estimation in the EVcomplex approach. We then obtained EC scores for residues i and j by first shifting the $\mathbf{J}_{ij}$ matrix for this pairwise site to zero-sum gauge ($\mathbf{J}'_{ij}$), and then summarizing it using the Frobenius norm. The resulting score $\mathcal{S}_{ij}^{\mathrm{FN}}$ was corrected for confounding bias using APC to obtain the final EC score $\mathcal{S}_{ij}^{\mathrm{CN}}$.

$$\begin{aligned}\mathcal{S}_{ij}^{\mathrm{FN}} &= ||\mathbf{J}_{ij}||_2 = \sqrt{\sum_{k}^{q}\sum_{l}^{q}\mathbf{J}'_{ij}(k,l)^2} \\ \mathcal{S}_{ij}^{\mathrm{CN}} &= \mathcal{S}_{ij}^{\mathrm{FN}} - \frac{\mathcal{S}_{i\cdot}^{\mathrm{FN}}\mathcal{S}_{\cdot j}^{\mathrm{FN}}}{\mathcal{S}_{\cdot\cdot}^{\mathrm{FN}}}\end{aligned} \tag{3.3}$$

Scoring

Every pair of residues in each protein as well as each possible combination of residues between the two interaction partners has an EC score, but given that most residues ac-

tually do not interact within or between proteins, weakly coupled residue pairs should be excluded from further steps. In monomer structure prediction, a rank-based approach using an average of one to two constraints per residue combined with several biochemically motivated filtering-steps has been shown to be robust due to the large number of constraints[120,149,159,268]. Reasoning that we only need a small number of inter-protein couplings to constrain 3D complex prediction combined with the observation that the co-evolutionary signal between proteins is usually weaker than within a protein, a stricter inclusion approach is required for protein complex prediction.

It is further necessary to determine a score threshold at which two proteins will not be considered interacting. While the shape of the score distribution is similar between all proteins, the absolute value ranges are not. We thus developed the EVcomplex score $\mathcal{Q}_{ij}$ to estimate the significance of the individual couplings while facilitating the comparison of predictions between proteins.

This score is based on several observations regarding the general EC score distribution (see Figure 3.3):

- The zero-sum gauge transformation in the EC score calculation step results in a zero-mean distribution (yellow dashed line in Figure 3.3)
- Most couplings are weak, i.e. have a score close to zero (red dashed line in Figure 3.3)
- A one-sided tail of couplings with positive scores can be observed in the distribution (see Figure 3.4 A)
- Couplings in the tail are more likely to be close in 3D space than those with a score close to zero (blue line in Figure 3.3).

From these observations we reason that the large fraction of couplings in the middle part of the distribution correspond to a symmetric Gaussian-like "background" distribution containing non-interacting residue-pairs. Based on the assumption that the more a coupling in the tail exceeds the boundaries of the symmetric background distribution, the more likely it reflects true co-evolution of the residue pair, we quantify the raw EC reliability as the ratio by which the EC score $\mathcal{S}_{ij}^{\mathrm{CN}}$ exceeds the noise level, i.e.,

$$\mathcal{Q}_{ij}^{\mathrm{raw}} = \frac{\mathcal{S}_{ij}^{\mathrm{CN}}}{|\min_{i,j}(\mathcal{S}_{ij}^{\mathrm{CN}})|} \tag{3.4}$$

The accuracy of EC scores critically depends both on the number and diversity of sequences in the input alignment and the size of the statistical inference problem[192,267,291].

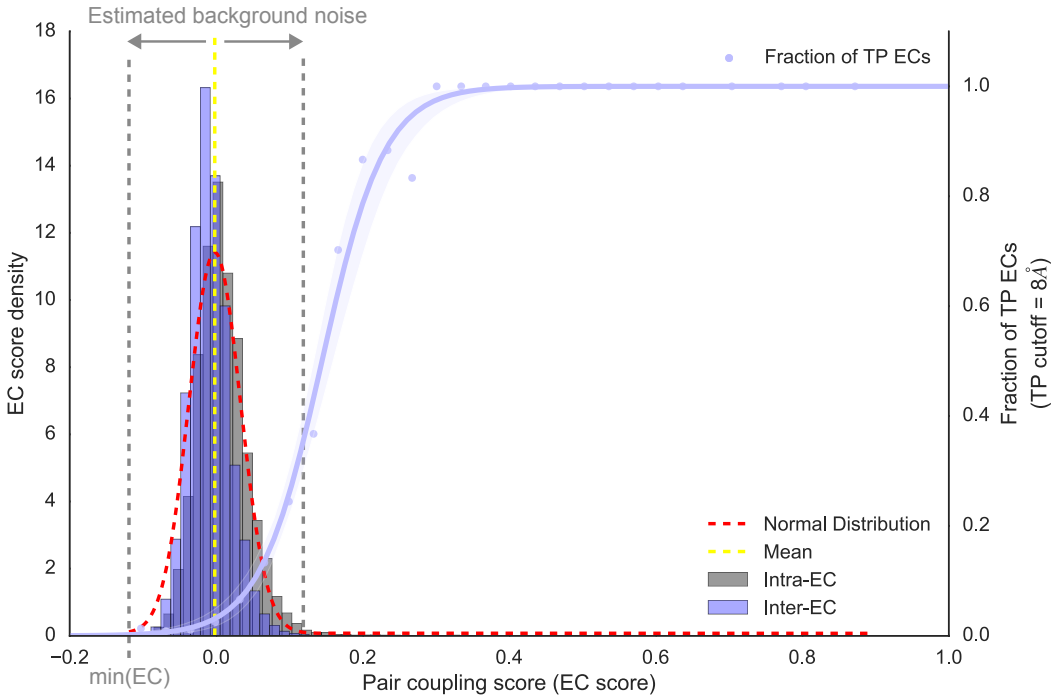


Figure 3.3: EC score distribution illustrated for inter- and intra-EC pairs of the methionine transporter complex, MetNI. Scores are zero-mean distributed (red dashed curve estimating the Gaussian distribution fitted to the data). Couplings with a score close to zero are less likely to be close in 3D space compared to those in the tail of the distribution (blue line representing the proportion of couplings with a distance less than 8 in the MetNI structure (PDB: 3tui)). This background noise is estimated by a symmetric range around 0 with the width being estimated by the minimum EC score.

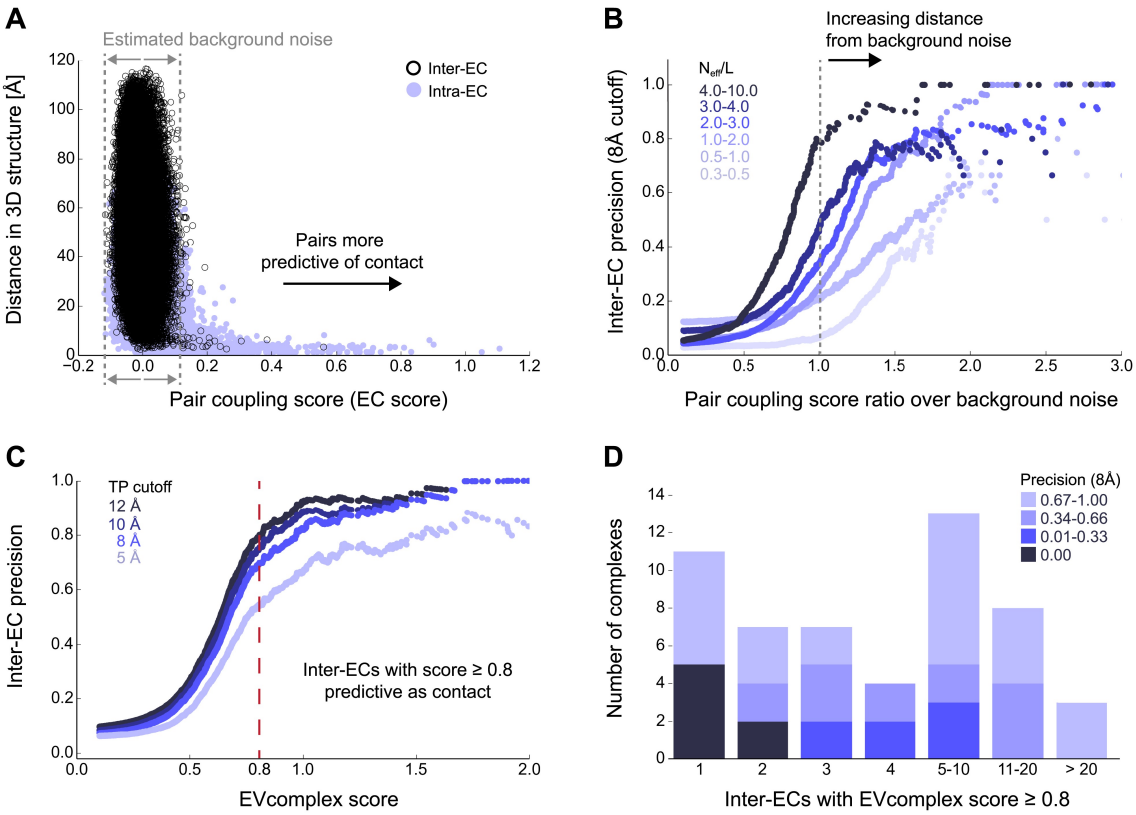


Figure 3.4: Evolutionary couplings capture interacting residues in protein complexes. (A) Inter- and intra-EC pairs with high coupling scores above the background level of the score distribution largely correspond to proximal pairs in the three dimensional structure of the complex. For the protein complexes in the evaluation set, this distribution is compared to the distance in the known 3D structure of the complex (shown here for the methionine transporter complex, MetNI) (B) An increasing distance from the background noise (ratio of EC score over background noise line) results in more accurate contacts. Additionally, the larger the alignment the more reliable the inferred coupling pairs are which then reduces the required distance from noise (different shades of blue). Residue pairs with a minimum atom distance of 8 Å between the residues are defined as true positive contacts, and precision = TP/(TP + FP). (Plot limited to range (0,3) excluding the outlier histidine kinase-response regulator complex (HK-RR) with extremely high number of sequences) (C) To compare different protein complexes and to estimate the average inter-EC precision for a given score threshold independent of sequence numbers, we normalized the raw couplings score for the number of sequences in the alignment (EVcomplex score). In this work, inter-ECs with an EVcomplex score ≥ 0.8 are used. Note: the shown plot is cut off at a score of 2 in order to zoom in on the phase change region excluding the high sequence coverage outlier HK-RR. (D) For complexes in the benchmark set, inter-EC pairs with EVcomplex score ≥ 0.8 give predictions of interacting residue pairs between the complex subunits to varying accuracy (8 Å TP distance cutoff). Reproduced from Hopf et al. (2014)[160].

We therefore incorporate a normalization factor to make the raw reliability score comparable across different protein pairs. This normalized EVcomplex score Q_{ij} is defined as

$$\mathcal{Q}_{ij} = \frac{\mathcal{Q}_{ij}^{\text{raw}}}{1 + \sqrt{\frac{N_{\text{eff}}}{L}}} \tag{3.5}$$

where N_{eff} is the effective number of sequences after the clustering step and L the length of the concatenated alignment.

As can be seen from the normalization factor, this score will be close to the raw EC score in the case of abundant training data while reducing the confidence in predictions in cases of limited data for the model size.

Comparison of predicted and known complex 3D structure

Once the raw and normalized coupling scores are computed, the residues with the strongest couplings within each protein can be used for structure prediction. To evaluate the prediction performance, couplings can be compared to their pair distance in the 3D structure. We calculated the precision (positive predictive value, PPV) of the couplings with a score above or equal to a range of thresholds,

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.6}$$

True positive (TP) couplings are here defined as those that are also close in the structure, in respect to minimum atom distance between two residues, while couplings with a high score, but large distance in 3D are considered false positives (FP). Following the general consensus in the field residues with a minimum atom distance of 8 Å or less are considered interacting between two complex subunits[317].

Using top ranking couplings (with a EVcouplings score above a certain threshold), monomer structures can be docked HADDOCK[82]. Residues with coupling scores above the EVcomplex score threshold are implemented as unambiguous distance constraints on C_α atoms in the docking protocol, with an effective distance of 5 Å (d_{eff}) and upper and lower bounds of 2 Å.

In the HADDOCK procedure a series of models is generated (500/100/100 in each of the following three steps: rigid-body energy minimization, semi-flexible refinement in torsion angle space, and finally refinement in explicit solvent/water). The models are scored by HADDOCK using a weighted sum of van der Waals and electrostatic energies (E_{vdw} and E_{elec}) together with an empirical desolvation term (E_{desolv}). The score would usually also include a distance restraint energy term in the last iteration, that is excluded in our evaluation protocol to allow comparison of scores between multiple runs.

After obtaining the 3D model of the complex, we evaluated the agreement of the models with experimentally determined structures. Following the CAPRI guideline for protein complex predictions[280], we focused on the binding interface for which we computed the root mean square deviation (RMSD) of all heavy atoms in the protein backbone at the interface,

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{3.7}$$

The RMSD measures the average distance between a defined set of N atoms in two superimposed structures. In protein structures comparison the Euclidean distance between the atom coordinates is commonly used distance measure δ. While monomer structure prediction is often evaluated based on backbone C_α atoms[192,267,291], the CAPRI guidelines[280] for the evaluation of the protein complex prediction focus on correctly modeling binding interface between the interaction partners. Here, the interface is defined as all residues with any heavy backbone atom (C_α, C, N, O) closer than 6 Å to any backbone atom in the interacting protein[280]. The interface RMSD (iRMSD) was computed using Python and ProFit v3.1 (available at `http://www.bioinf.org.uk/software/profit`).

Determination of interacting subunits in a macromolecular assembly

Co-evolution scores can also be used to predict *if* two proteins interact, through calculation of the EVcomplex scores between arbitrary proteins. Based on the scores obtained in the PPI evaluation data set, a reliability score threshold was determined to discriminate between pairs that show co-evolution signals with high confidence and those that lack such evidence: subunits with a maximal EVcomplex score greater or equal to 0.8 likely, those between 0.75 and 0.8 as maybe interact, while interactions between subunits that had a score below 0.75 could be rejected.

3.2.2 Data Sets

Validation and prediction data sets

The EVcomplex method was evaluated using a data set of known binary PPI in *E. coli* with known 3D structure, compiled from yeast two-hybrid experiments and literature-curated databases[346]. We further extended this data set with three additional pairs based on our analysis of other subunits in the same macromolecular complex: ButC/BtuF, MetI/MetQ, and ATP synthase subunits a and b. 3D structures of protein complexes in this set without existing structural information were then predicted using the same protocol.

As one of the two matching methods used to pair up sequences from the two protein monomer MSA is based on genome distance, we filtered the original data set presented in Rajagopala *et al.*[346] to remove all pairs that are separated by more than 20 genes in the *E. coli* genome (as determined by an ordered list of genes in the UniProt database[64]). We then removed pseudo-homomultimeric complexes and annotated interaction homologs by mapping the proteins in the set to protein families as defined in the Pfam database, and then filtering those in which both partners were annotated with the same Pfam domain as well as flagging pairs whose specific combination of Pfam domains was also observed in other PPIs in the set.

The data set was then split into a validation subset with known structures and a prediction set. To assign complexes to these two sets we combined the annotation of known 3D structures[346] with a search for homologous structures in the PDB. First, each of the two partners was used as a search query, then the results were intersected, keeping those where both partners were found in the same PDB entry. We assigned those complexes whose structures where only solved at resolutions above 5 Å to the prediction set (affecting eight entries). This approach resulted in a final set of 93 non-redundant bacterial complexes whose subunits are proximal in the genome for evaluation and a set of 32 complexes without structural data.

Since the complexity of the model increases quadratically with the length of the query sequence, in multi-domain proteins we focused the protocol on interacting domains (informed by the known 3D structures and assigned Pfam domains).

MSA generation

Multiple sequence alignments of the individual proteins were created by jackhmmer[189] with five iterations, a bit score domain inclusion threshold of 0.5 × L (where L is the length of the protein) searching the UniProt database[64]. Genomic locations for the coding sequences (CDS) were then retrieved for all members of the alignments from the ENA database[320]. We then paired the sequences in the alignments of the two PPI partners using the matching protocols described in Chapter 3.2.1. To compare the results to previous reports, we mainly evaluated the performance of the distance-based matching approach. We further excluded sequence pairs that had a genomic distance greater than 10 k nucleotides to avoid inclusion of false-positive pairs.

We then clustered the paired sequences at 80% sequence identity and re-weighted them according to their cluster membership. Those columns in the alignment that contained more than 80% gaps were excluded from the prediction.

3D structures of monomers

Three-dimensional structures of the complexes in our evaluation set were obtained from the Protein Data Bank[22]. Further, out of the 22 complexes in the evaluation set that had five or more inter-protein couplings, a diverse set of 15 was chosen for 3D complex prediction. Where available, we selected unbound structures of the monomers in the docking step to avoid overfitting and compared these to the 3D structure of the cognate complex. If no unbound structures existed, we randomized side chain placement using either Schrödinger Protein Preparation Wizard[368] or SCWRL[220] followed by a short optimization step in Chimera[339]. We excluded domains that were reported to be highly flexible, namely the two C-terminal helices of AtpE in the AtpE-AtpG interaction, the CA domain of histidie kinase in the two-component signaling complex, and subunits 1 and 2 of COX 2 in ubiquinol oxidase (here only in the docking evaluation).

Not all complexes in our prediction set had previously solved monomer structures. In these cases, we predicted the 3D structure of the monomers using one of two strategies: If a homologous structure existed, we created a comparative model of the structure in *E. coli* using SwissModel[27,34] (MetQ and IlvB). Alternatively, if no structure template for comparative modeling was available, we built the structures *de novo* following the EVfold protocols[159,268] using PLM[15,95] and sequence clustering at 90% sequence identity. This was particularly the case for the subunit a of ATP synthase and UmuC.

Evaluation of interacting subunits prediction in a macromolecular assembly

We tested the usefulness of our co-evolution approach as a method to discern interacting and non-interacting partners in larger protein complexes on the eight subunits of *E. coli* ATP synthase F_0 and F_1 complex. For all 28 possible pairwise combinations we computed EVcomplex scores and considered the highest inter-monomer score as a proxy for the likelihood of interaction. Predictions were then compared to observable interaction from the known partial complex structure (PDB: 3oaa, 1ds0, 2a7u) as well as literature reports of interacting subunits as determined by crosslinking, cryo-EM and other experiments[37,75,275,379].

3.2.3 EVcomplex Webserver

We implemented the EVcomplex pipeline as a command line tool in Python that is available through a webserver (`www.evcomplex.org`, source code available from `https://github.com/debbiemarkslab/evcomplex_server`). This webserver was built on top of the pipeline using Python Flask[356] as a low weight back-end following the common model-view-controller design pattern, separating classes responsible for computations from those delivering visualisations to the user.

The front-end was implemented in HTML5, CSS and JavaScript, including the extensions D3.js and jquery. Twitter bootstrap (`http://getbootstrap.com`) was used as the main front-end framework due to its responsive and mobile first approach. The familiar style of input elements also reduces entrance hurdle to new users. Interaction between back- and front-end is facilitated through the Jinja templating engine[357]. For dynamic content including long polling of currently active prediction runs in a constantly updating website, we further incorporated the JavaScript framework AngularJS (`https://angularjs.org/`).

Past and active jobs are stored in a MySQL database. Each step in the pipeline as described in Chapter 3.2.1, namely monomer homology search, alignment concatenation, couplings prediction, post-processing and re-scoring, as well as comparison to known 3D structures, if available, tracks its progress in the database and upon successful completion the user is provided a view of the results. If the job fails in the process, the user will also be informed of the event and the step at which the problem occurred.

We include and implement several visualizations of the data produced in the run to support the user in his analysis, such as glyph-based sequence logos of the concatenated alignment[261] and contact maps displaying strongly coupled residue pairs within and between the input proteins.

3.3 Results

3.3.1 Performance of the Algorithm on Known Protein Complex Structures

Of the 93 complexes in the our evaluation set, 76 had sufficient sequence data available to allow EVcomplex predictions (minimally 0.3 non-redundant sequences per residue). For this set we analyzed the relationship between the EVcomplex score and the precision of the pair predictions compared to known structures. We observe that 53 out of the 76 complexes had at least one inter-protein coupling with a score above 0.8 and that 69% of the predicted pairs with a score above this threshold are accurate within 8 Å to the cognate complex structure (Figure 3.4C). A small number of complexes showed more than 20 inter-protein couplings with scores above 0.8 with precision over 80% (namely, histidine kinase and response regulator, t-RNA synthetase, and vitamin B importer complex) (Figure 3.4D), while 23 out of the 76 had no high-scoring coupling even though it is well known that the subunits interact. Reasons for this are discussed in Chapter 3.4.

Contacts predicted with coupling scores below 0.8 are not necessarily false. Especially those that cluster at the same interface as the contacts predicted with higher scores, can be correct. In the ethanolamine ammonia-lyase complex, for example, three inter-contacts are predicted above a threshold of 0.8 while five additional contact pairs with scores slightly

below that threshold cluster with the contacts on the monomers, indicating their high likelihood of being also correct.

On the other hand, as shown in Figure 3.4, not all inter-protein contacts with a high coupling score are close in known structure models. This can be due to inaccurate assumptions of the method in some cases, including the assumption that interaction is conserved across paralogs and orthologs, as well as the assumption that co-evolution stems from structural contact. It can be further the result of highly flexible assemblies that can undergo large conformational changes of which only a subset has been captured by 3D structures (e.g., the vitamin B12 transporter complex BtuCDF in the process of transporting B12 across the membrane[173]).

Evolutionary couplings produce accurate three-dimensional complex structures through docking

Protein-protein docking usually relies on predicted or experimental constraints that approximate the binding mode to reduce the search space of possible complex conformations and produce meaningful results[428]. To evaluate the performance of ECs as constraints for protein docking, we selected a diverse set of 15 protein complexes (Figure 3.5) with at least five inter-protein contacts with an EVcomplex score above 0.8. We then generated 100 models per complex using ECs as distance constraints in HADDOCK[82] . Additionally, we also produced control models to assess the amount of information added to the docking protocol by ECs (500 models per complex, no constraints except center of mass).

For 13 out of the 15 complexes, the top-ranked model obtained using ECs as HADDOCK constraints had iRMDS under 6 Å when compared to their cognate structure, as did the best-performing model in all 15 cases (Table 3.1). Overall, 74.4% of the 1500 generated models using EC constraints are close to the experimental structures of the complexes ($<$ 6 Å backbone iRMSD), compared to less than 2% of the 7500 controls. Not surprisingly, complexes with a high number of true positive contacts showed the highest docking performance, with a weak correlation between number of ECs and iRMSD (Pearson's r=-0.28, p=0.3) and TP rate and iRMSD (Pearson's r=-0.48, p=0.068). For example, the 30S ribosomal proteins RS3 and RS14 showed a precision of 0.91 for 11 ECs and an iRMSD of 1.1 Å for the top-ranked model.

In the methionine-importing transmembrane transporter complex, MetI interacts with the ATP binding protein MetN. Overall, MetI-MetN had 14 inter-protein ECs with a precision of 0.86 and an average iRMSD of 1.4 Å over all 100 models (1.5 Å for the top ranked model, Figure 3.6A). The three top-ranking inter-protein ECs (K136-E108, A128-L105, and R124-E74, MetI-MetN respectively) constitute a network coupling the ATP-binding pocket in MetN to the open and closed conformations of the membrane transporter MetI[186]. Using

Table 3.1: EVcomplex predictions and docking results for 15 protein complexes

		EVcomplex contacts			**Docking quality (iRMSD)**	
Complex name	**Subunits**	**Seqs**	**ECs**	**TP rate**	**Top-ranked model**	**Best model**
ATP synthase ϵ and γ subunits	AtpE:AtpG	2.9	15	0.53	1.4	1.4
Vitamin B12 uptake system	BtuC:BtuD	9.8	21	0.88	1.1	0.9
Vitamin B12 uptake system	BtuC:BtuF	3.2	5	0.6	2.8	2.8
Carbamoyl-phosphate synthase	CarB:CarA	2.3	17	0.88	1.9	1.9
Ubiquinol oxidase	CyoB:CyoA	1	11	0.55	1.8	1.2
Dihydroxyacetone kinase	DhaL:DhaK	1.4	12	0.42	6.7	2.4
Outer membrane usher protein/ Chaperone protein	FimD:FimC	3.6	6	0.83	3.2	3
Aminomethyltransferase/ Glycine cleavage system H protein	GcsH:GcsT	2.9	5	0.2	5.4	5.4
Histidine kinase/ response regulator (*T. maritima*)	KdpD:CheY	95.4	78	0.72	2.1	2
Methionine transporter complex	MetN:MetI	1.9	14	0.86	1.5	1.2
Molybdopterin synthase	MoaD:MoaE	3.6	8	1	4.4	4.1
IIA-IIB complex of the N,N'-diacetylchitobiose (Chb) transporter	PtqA:PtqB	3.1	5	0.2	7.2	5.5
30 S Ribosomal proteins	RS10:RS14	1.2	6	1	5.3	2.5
30 S Ribosomal proteins	RS3:RS14	1.4	11	0.91	1.1	1.1
Succinatequinone oxido-reductase flavoprotein/iron-sulfur subunits	SdhB:SdhA	3	8	0.62	1.4	1.4

Seqs = Number of non-redundant sequences in concatenated alignment normalized by alignment length, ECs = inter-ECs with EVcomplex score ≥ 0.8, top-ranked model = iRMSD of model from known structure, for docked model with best HADDOCK score, best model = lowest iRMSD observed across all models.

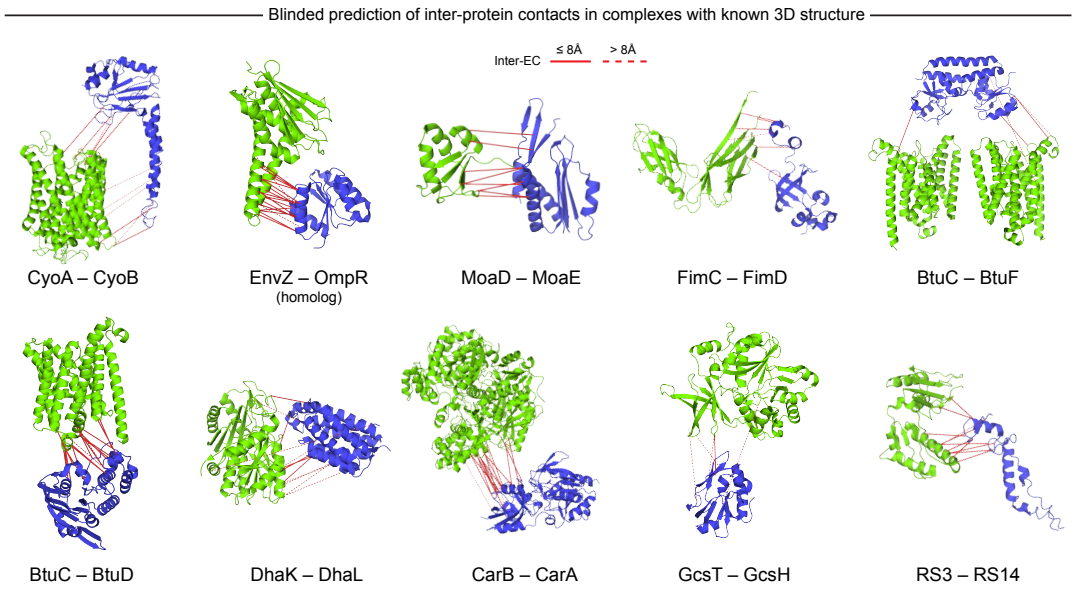


Figure 3.5: Blinded prediction of evolutionary couplings between complex subunits with known 3D structure. Inter-ECs with EVcomplex score ≥ 0.8 on a selection of benchmark complexes (monomer subunits in green and blue, inter-ECs in red, pairs closer than 8 Å by solid red lines, dashed otherwise). The predicted inter-ECs for these ten complexes were then used to create full 3D models of the complex using protein-protein docking. Reproduced from Hopf *et al.* (2014)[160]

the docking protocol, this interaction network could be reproduced accurately in the 3D structure (Figure 3.6A).

Nevertheless, the protocol also appears to be robust against a high number of false positive constraints as both Ubiquinol oxidase (6 out of 11 constraints correct) and the ATP synthase F_1 ϵ and γ subunits (8 out of 15 correct) also resulted in top-ranked complexes with iRMSDs of 1.8 and 1.4 Å, respectively. In ATP synthase F_1, residue D82 in the ϵ subunit and R222 in the γ subunit together with high scoring intra-γ couplings connect the ϵ subunit to the core of the ATP synthase and its catalytic function which was also reproduced in the docked models (Figure 3.6B).

Evolutionary coupling analysis thus not only identifies conserved residue networks, but also adds significant information to the docking protocol. This results in an drastically increased docking accuracy compared to center of mass docking (75% vs 2% correctly docked models with and without ECs).

Strength of evolutionary couplings can help predict interacting subunits

After showing that ECs can successfully identify evolutionary interaction constraints between known interacting subunits of a complex, we asked whether it is also possible to distinguish between interacting and non-interacting proteins based on the strength of their

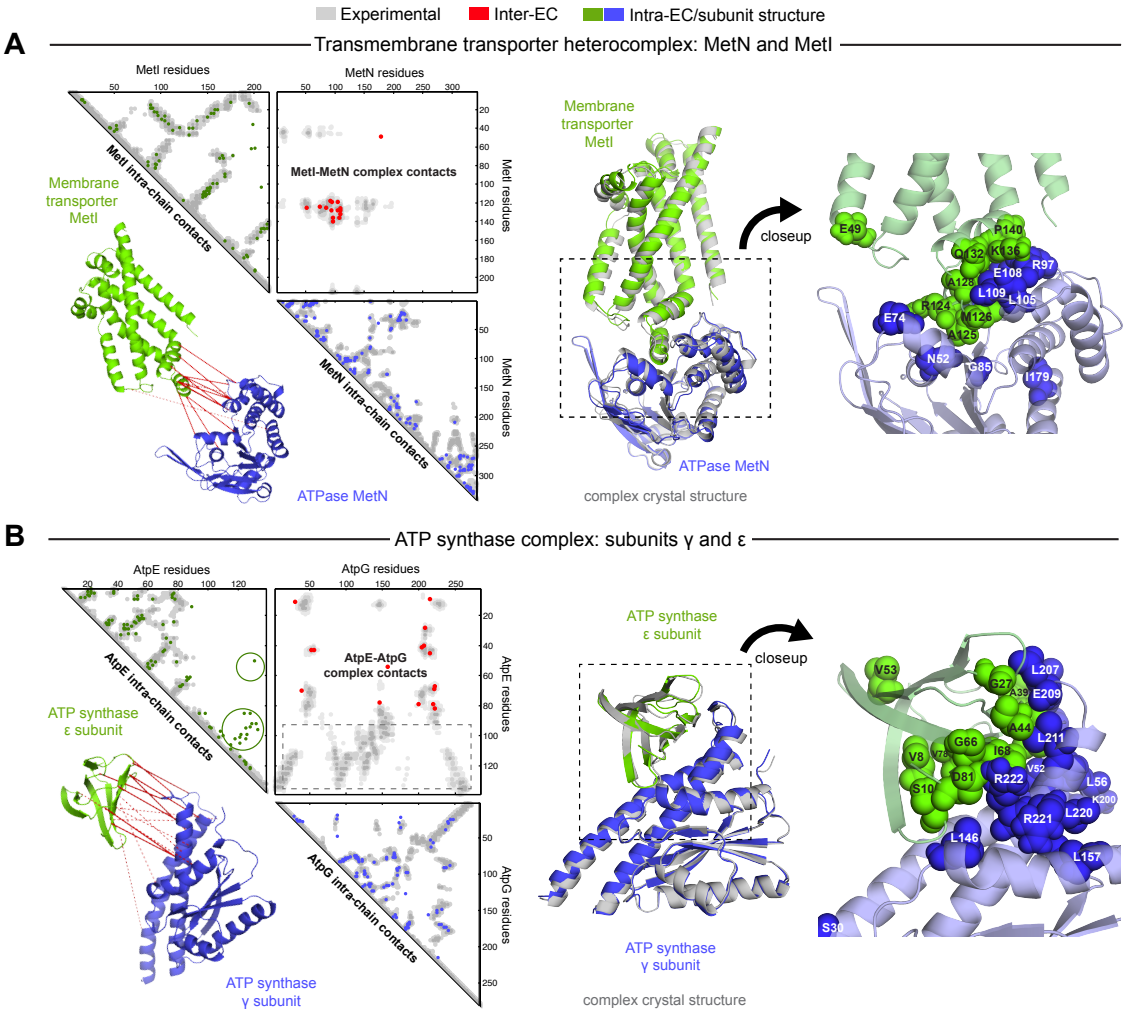


Figure 3.6: Evolutionary couplings give accurate 3D structures of complexes. EVcomplex predictions and comparison to crystal structure for (A) the methionine-importing transmembrane transporter heterocomplex MetNI from *E. coli* (PDB: 3tui) and (B) the gamma/epsilon subunit interaction of *E. coli* ATP synthase (PDB: 1fs0). Left panels: complex contact map comparing predicted inter-ECs with EVcomplex score ≥ 0.8 (red dots, upper right quadrant) and intra-ECs (up to the last chosen inter-EC rank; green and blue dots, top left and lower right triangles) to close pairs in the complex crystal (dark/mid/light gray points for minimum atom distance cutoffs of 5/8/12 Å for inter-subunit contacts and dark/mid gray for 5/8 Å within the subunits). Inter-ECs with an EVcomplex score ≥ 0.8 are also displayed on the spatially separated subunits of the complex (red lines on green and blue cartoons, couplings closer than 8 Å in solid red lines, dashed otherwise, lower left). Right panels: superimposition of the top ranked model from 3D docking (green/blue cartoon, left) onto the complex crystal structure (gray cartoon) and close-up of the interface region with highly coupled residues (green/blue spheres). Reproduced from Hopf *et al.* (2014)[160].

co-evolution signal. We chose ATP synthase as well characterized test case, even though some aspects of its 3D structure still remain unknown (see Section 3.3.2). This macromolecule consists of eight subunits, of which five are located in the cytoplasm ($\alpha, \beta, \gamma, \delta, \epsilon$) and three are anchored in the membrane (a, b, c). ATP synthase is ubiquitous in all kingdoms of life and is essential in ATP production. To see if the EVcomplex score can be used as a proxy for existing interaction, we calculated the EC and EVcomplex scores for all 28 possible subunit combinations. Using the previously determined threshold of 0.8, it was possible to correctly classify 24 out of 28 (86%) interactions. Furthermore, 80% of the interacting residue pairs covered in the known partial crystal structure of ATP synthase (PDB: 3oaa) were correct at 10 Å minimum atom distance.

The four interactions that are wrongly classified as not interacting at an EVcomplex threshold of 0.8 can partially be restored at a lower threshold of 0.75, at the expense of introducing two new false positives (correctly classified: $\beta - b$, $\epsilon - c$, now false positive: $\beta - c$, $\delta - a$). For some of the interactions in the complex, evidence diverges in the literature. For these cases it is interesting to note that some are supported by strong ECs (e.g., $\alpha - \epsilon$ and $\beta - \delta$) while others lack this strong signal ($\beta - \epsilon$ and $\gamma - c$, classified as false negative in the statistics presented above). The interaction between β and ϵ is a special case as can be seen in the crystal structure of 3oaa that involves a highly extended conformation of the last two helices of the ϵ subunit into the enzyme that is not present in other structures. The lack of a reasonably strong signal for this interaction could either result from the transience of the interaction or a lack of conservation of this interaction across homologs (see also Section 3.4).

Based on the high success rate of classification of interacting subunits we suggest that EVcomplex and the EVcomplex score can be used on a larger scale for genome-wide interaction predictions as well as identification interacting subunits in large complexes.

3.3.2 Novel Predictions of Protein Complexes

The data set of 3,449 *E. coli* PPI used to derive our evaluation and prediction sets[346] did not contain any high-resolution structural information for 229 complexes in which the interacting partners are close on the genome (subunits less than twenty genes apart). 82 of these complexes had sufficiently large and diverse concatenated sequence alignments ($N_{\text{eff}/L} \geq 0.3$) and no known structure of interacting homologs. For these complexes we predicted evolutionary couplings and after transforming the raw EC score into EVcomplex scores for inter-protein couplings, 32 of the complexes had at least one inter-protein contact with a score above 0.8.

Based on observations in the evaluation set that protein pairs with a high number of contacts with a high confidence score tend to also have a higher proportion of correct

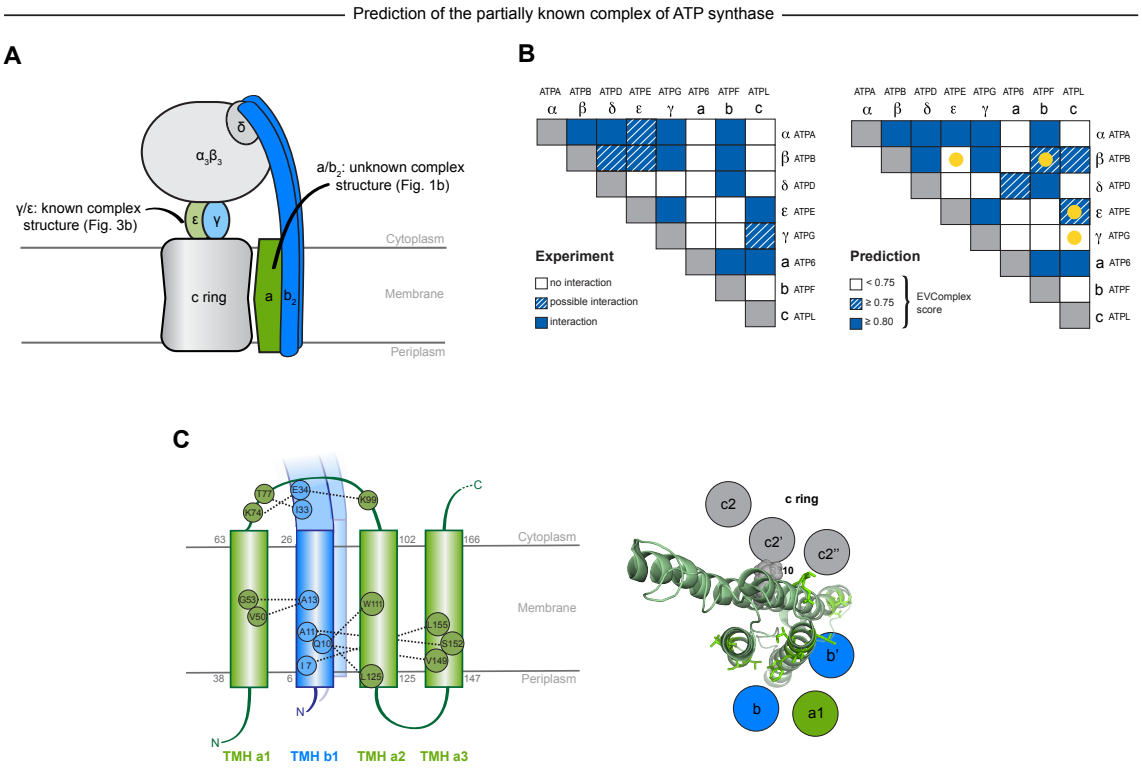


Figure 3.7: Predicted interactions between the a-, b-, and c-subunits of ATP synthase. (A) The a- and b- subunits of *E. coli* ATP synthase are known to interact, but the monomer structures of subunits a and b and the structure of their interaction in the complex are unknown. (B) EVcomplex prediction (right matrix) for ATP synthase subunit interactions compared to experimental evidence (left matrix), which is either strong (left, solid blue squares) or indicative (left, crosshatched squares). Interactions that have experimental evidence, but are not predicted at the 0.8 threshold are indicated as yellow dots. (C) Left panel: residue detail of predicted residue-residue interactions (dotted lines) between subunit a and b (residue numbers at the boundaries of transmembrane helices in gray). Right panel: proposed helix-helix interactions between ATP synthase subunits a (green), b (blue, homodimer), and the c ring (gray). Reproduced from Hopf *et al.* (2014) [160]

contacts in that set (Figure 3.4D), we evaluated some examples of such complexes with several contacts above the threshold of 0.8 in the light of prior biological information. These examples include subunits I and Q of the D-methionine transporter (MetI - MetQ), the UmuC and UmuD complex of the *E. coli* stress response system, subunits a and b of ATP synthase as well as small and large subunits of Acetolactate synthase (IlvB - IlvN) (Figure 3.8).

While the interaction between methionine permease MetI and the ATPase MetN of the D-methionine ABC transporter has been evaluated in our benchmark set (Figure 3.6), the interaction between MetI and the periplasmic methionine-binding protein MetQ [281] is the prediction with the highest number of strong couplings (24 contacts with score above 0.8).

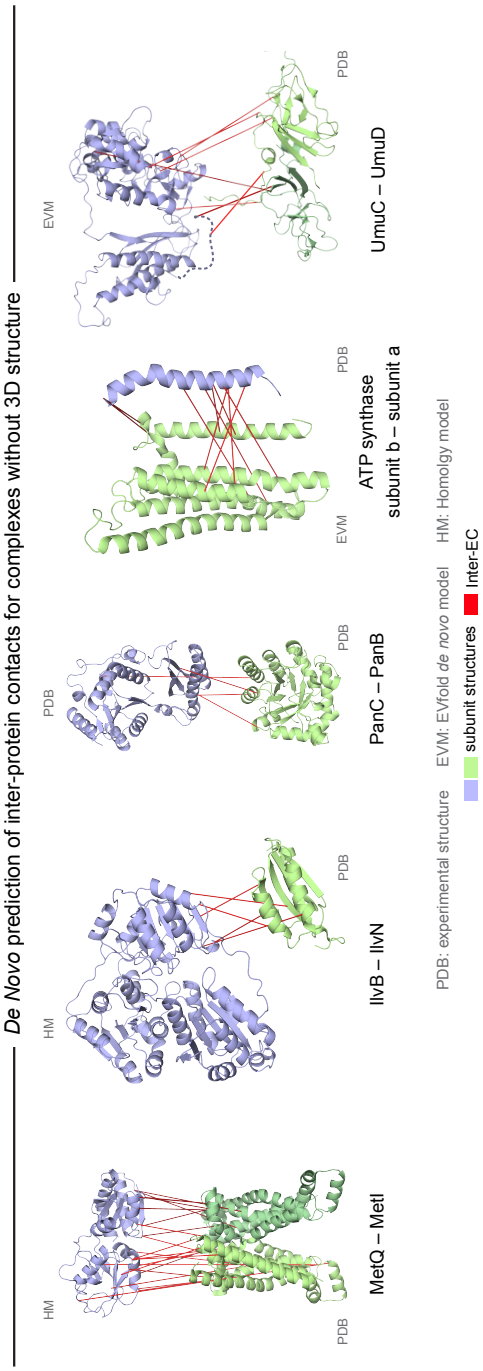


Figure 3.8: Evolutionary couplings in *E. coli* complexes of unknown 3D structure (subunits: blue/green cartoons; inter-ECs with EVcomplex score ≥ 0.8: red lines). For complex subunits which homomultimerize (light/dark green cartoon), inter-ECs are placed arbitrarily on either of the monomers to enable the identification of multiple interaction sites. Left to right: (1) the membrane subunit of methionine-importing transporter heterocomplex MetI (PDB: 3tui) with its periplasmic binding protein MetQ (Swissmodel: P28635); (2) the large and small subunits of acetolactate synthase IlvB (Swissmodel: P08142) and IlvN (PDB: 2lvw); (3) panthotenate synthase PanC (PDB: 1iho) and ketopantoate hydroxymethyltransferase PanB (PDB: 1m3v); (4) subunits a and b of ATP synthase (EVfold-membrane model for subunit a, PDB: 1b9u for subunit b), and (5) UmuC (EVfold-model) with one possible conformation of UmuD (PDB: 1i4v). Reproduced from Hopf *et al.* (2014)[160]

All of the top 15 contacts between MetI and MetQ are located at the periplasmic face of MetI and one interface of MetQ. The identified binding interfaces is thus consistent with the known periplasmic binding of the two subunits.

As part of the bacterial stress/SOS response, UmuD is cleaved to form UmuD'$_2$ which then interacts with the DNA polymerase UmuC to copy damaged DNA[25]. Six out of the seven ECs above the score threshold are co-located on the same face of the dimer. Furthermore, while the location of the residues from inter-protein ECs are located in two separate domains of UmuC, intra-UmuC ECs further support the spatial proximity of of these residues, supporting the accuracy of the interaction interface.

Similar clustering patterns can be observed for the other predictions, thus adding confidence to the prediction, showing that EC analysis can be capable of identifying interaction interfaces in previously uncharacterized protein complexes.

Post hoc evaluation of predictions using recently solved complex structures

While the manuscript[160] was in preparation, the previously unsolved structure of DinJ-YafQ, a toxin/antitoxin complex was published (PDB: 4mlo[239] and 4qlu[359]), allowing to further quantify the accuracy of our predictions. We compared our predicted inter-protein couplings to the 3D structure and found that 17 out of the 19 couplings with an EVcomplex score above 0.8 have a minimum atom distance below 8 Å in the structure. When employing our docking protocol (as described in section 3.2) to this structure, we further find that the predicted complex agrees well with the solved structure (overall C_α RMSD of the best model: 4.6 Å).

This post hoc analysis again demonstrates that our predicted inter-EC contacts agree well with experimental data and the method is capable of identifying *de novo* interaction patterns.

Prediction of residue details of subunit interactions in ATP synthase

While most of the 3D structure of ATP synthase is known[436], some parts of this complex remain elusive. Particularly, the interactions between the a, b and c subunits as well as the monomer structure of a have not yet be solved experimentally. After the successful evaluation of inter-protein EC predictions, including the correct identification of interacting subunits in the overall ATP synthase complex, we moved on to predict details on the interaction of the previously uncharacterized interfaces. There exists only a homology model for the a subunit of the complex, developed in 1999 (PDB: 1c17[350]) based on constraints obtained from mutation experiments. From these a model of transmembrane helix (TMH) 2 to 5 was inferred. This topology was later supported by crosslinking data for all possible combinations of helix pairs of TMH2 - TMH5[381]. To incorporate TMH1 into a

model of the full *a*-subunit, we used the EVfold protocol for transmembrane proteins[159]. The helix packing arrangement of the resulting model is supported by several crosslinking data sets[74,104,253] including the exact or near exact correspondence between the crosslinks obtained by Schwem and Fillingame (2006)[381] and the top-ranking intra-subunit ECs. There is, however, only a weak coupling signal for TMH1 to the helical bundle suggesting that this helix is not part of the bundle, again supported by studies that failed to detect crosslinks between the first transmembrane helix and the four-helix bundle[104].

We then predicted inter-protein ECs based on the previously established inter-protein co-evolution protocol and detect ten high confidence contacts. All of these are between the membrane helices (TMH1-3 and THM5) in *a* and the membrane-integral part of *b*. Again, these interactions can be supported by crosslink data as they either directly correspond to known cross-links or are in the direct neighborhood thereof[74,75,253]. Unfortunately, it was not possible to construct an explicit 3D model of the *a*-*b* complex due to the conflicting constraints between the individual helices in *a* to the *b* homo-dimer. It is thus necessary to resolve these conflicting information in further work that may result from an intertwined arrangement of the two subunits.

For the interaction between subunits *a* and *c*, we also analysed high confidence inter-protein ECs and found that the top ranked inter-protein coupling is close to the known interaction interface (aR201 - cD61)[80].

While it was not possible to create a full three-dimensional model of the ATP synthase *a* and *b* interaction at the time, our study still shows that it is feasible to add orthogonal information to previous data. With more work on disentangling contradicting couplings it may well be possible to reconstruct the full 3D structure of ATP synthase. This is supported by two recent cryo-EM structures of bovine and *L. pneumonia* ATP synthase[476,480] which are consistent with the predicted ECs.

3.4 Discussion

In this chapter, we were able to show that it is possible to reconstruct PPI interfaces from sequence information alone. The statistical approach presented here is based on a global maximum entropy model that successfully detects direct co-evolution between residues within and between two proteins. We were able to show that co-evolved residues between two subunits are indeed close in the complex structure and can successfully be used as constraints in protein docking to determine the correct three-dimensional structure of the complex. We applied this approach to a set of unknown complexes for *de novo* interface prediction. This work, together with an independent and parallel work by Ovchinnikov *et al.*[317] constitute the first generalized application of a global co-evolution model to PPIs.

While the presented work can be seen as a proof-of-concept for several aspects of co-evolution based predictions in the context of protein interactions, limitations remain.

Availability of sequence information

The dependence on large numbers of evolutionary related, but diverse, sequences is one of the main limitations of the method and depends on two factors: i) the availability of enough homologous monomer sequences in public databases and ii) the concatenation of those to form a joint MSA. With the dramatic increase of available genome sequences, the problem of finding sufficient numbers of homologs for each interacting subunit will likely subside in the future. We hypothesize that once there are more than 10,000 (i.e., more than ten times the concatenated length of proteins, on average) fully sequenced bacterial genomes of sufficient diversity, it will become feasible to test all possible interactions in a typical bacterial genome, and for those with strong evolutionary couplings infer the three-dimensional structure of the complex.

For protein complexes that are unique to a single domain of life, the scenario is worse due to the lower availability of sequence data. It is easy, though, to imagine targeted sequencing efforts in a particular set of species to further such approaches. For PPIs with a large number of paralogous interactions, but only present in a small number of species, the matching of monomer sequences poses the largest limitation of our approach.

Correct matching of protein monomer sequences

As mentioned before, the accuracy of the proposed method depends on correctly matching the found monomer sequences because each wrongly matched pair in the alignment adds misinformation to the statistical model. It is evident from the results presented here that sufficiently large sequence alignments result in more precise predictions of the interacting residues. In this work we implemented two approaches to match the homologous sequences of both subunits across multiple species: one based on the assumption that the two homologs in a species with the highest sequence identity to the query proteins also interact with each other; the other matching method is based on the observation that interacting proteins in bacteria often reside on the same operon in the genome. It has been shown to be successful in identifying multiple interaction homologs per species (e.g., multiple copies of the two-component signaling and toxin/anti-toxin systems), but is not applicable to interactions only found in eukaryotes due to the absence of operons in their genome architecture.

Solving the matching problem in general is not a trivial task because we usually do not know 1) how conserved an interaction is compared to the monomer sequence, and 2) which pairs of paralogs interact in a species. In the case of non-specific or promiscuous interactions between multiple pairs of paralogs this task is further exacerbated as there

exists no clear 1:1 matching. The effect of incorrect matching can be observed in the prediction result as a strong (and correct) intra-protein co-evolution signal, while no strong inter-ECs can be identified. This means that while co-evolution within proteins remains detectable since it is independent of the sequence order in the alignment, the wrong pairing of interacting sequences results in the loss of detectable couplings between the proteins.

There is still a demand for more elaborate sequence matching approaches that likely require additional prior knowledge to optimally pair multiple paralogs even in non-bacterial organisms. Several approaches can be envisioned that can either include further phylogenetic data, additional interaction predictions such as those described in Chapter 2.4.1, or an iterative optimization approach that combines sequences matching with simultaneous co-evolution detection. The latter has been tested recently by Gueudre *et al.*[136]. Here, the authors hypothesize that the alignment that maximizes the inter-protein co-evolution signal is most likely the correct one. While generally promising, this method currently has some shortcomings. Detection of interacting proteins may be biased by initial assumptions that the two protein families are indeed interacting as this served as the basis for maximizing the inter-family co-evolutionary signal. The method is thus not suitable to predict *if* two proteins interact in the first place. Secondly, the problem definition results in an exponential search space in respect to the number of species in the alignment and a super-exponential search space in the number of paralogs inside each species. The authors use a different definition of the statistical model that allows determining heuristic solutions to the problem in reasonable time, but further development is required to apply such approaches in the EVfold and plmDCA based framework suggested here.

In other cases, such as cross-species contacts involved in pathogen-host or antibody-antigen interactions, currently no general matching strategies exist, but the required paired sequences may be obtained by targeted sequencing of specific specimen in a cohort of infected individuals.

Discrimination between monomer- and homooligomeric signals

Another limitation of the proposed method is that it is not capable of detecting homo-oligomeric interactions. In the case that a protein interacts with other copies of itself, this poses an evolutionary constraint on the monomer sequence, which cannot be found through our approach. Discriminating between homo-oligomeric inter-contacts and within monomer intra-contacts has so far been mainly done based intra-contacts determined from experimental monomer structures[83,120,306,307]. While this approach can be useful if there exists structural data about the monomers, it cannot blindly discriminate between those contact classes in the absence of any structural information. Possible strategies in this context could include a folding strategy that simultaneously folds the monomers

and assembles them, at the same time in- and excluding constraints to the two contact classes based on observed clashes or energy changes of the resulting structures — similar to approaches used to disambiguate constraints for structure reconstruction from NMR and cross-linking data[102,308]. This approach may particularly be helpful in cases where the monomer structure is not stable on its own and an intertwined assembly, such as the one seen in many ion channels, is required to stabilize the complex.

Chapter 4

Genetic Variation in Drug Targets and Other Pharmacogenes

Parts of the content of this chapter have also been published in the following article:

Genetic variation in human drug-related genes[371]

4.1 Introduction

About three in five Americans aged 20 and above take prescription drugs every month[199] and many either encounter adverse events or reduced treatment effects[230,374]. In addition to environmental factors and heterogeneity of the disease, the genetic markup of the patient is expected to contribute substantially to the effectiveness of treatments through variants in drug metabolizing enzymes and drug targets[260] that affect the drug's PK and PD profile in the patient[341]. The identification of genetic determinants of drug efficacy and toxicity is thus of large interest for the development of effective medicines[246], to optimize dosing, and minimize side effects for the patient.

The advent of affordable genotyping of patients has resulted in the rise of PGx studies searching for such determinants[260] in population stratified[282,311,468] or individualized settings[258,341]. Population genetics studies investigating drug-related genes have so far mainly focused on a limited set ADME pharmacogenes[44,114,218,287,459] in smaller patient cohorts (Supplementary Table E.1). Geographic ancestry was only considered in few studies and was mainly derived from either two different American populations or five global superpopulations in the 1,000 Genomes Project[63]. Overall, differences in AF between populations could be observed in all studies that considered geographic ancestry[90,114,218,459], indicat-

ing the need to overcome annotation bias and extend study cohorts to better represent populations of Non-European descent.

Even though early PGx studies have mainly focused on genes of particular interest for PD or PK, larger GWAS have recently been able to identify PGx variants in a less biased manner[70,296]. Much of the current knowledge on genetic variation and drug response is summarized in PharmGKB[453], but the majority of drugs are missing PGx information. The extent to which genetic variation affects drug response can be illustrated by the class of anti-hypertensive drugs, now taken by a quarter of all adults in the US[199], where over 200 publications have reported PGx variants in close to 250 different genes (`https://www.pharmgkb.org/disease/PA444552`, accessed 09/03/18). Most of these variants only have a small effect[256], but the analysis of drug target-associated variants in addition to other PGx genes has shown promise to predict part of the treatment response[123]. Such connections between polymorphisms in disease genes/drug targets include two common functional variants in the β1-adrenergic receptor (*ADRB1*, Ser49Gly and Gly389Arg) that are linked to different responses to channel blockage by beta-blockers[187,248,401].

Translation of known PGx factors into clinical prescribing has only taken place for a few drugs, many of which are targeting cancer and cancer-type specific somatic mutations, such as the anti-HER2/neu antibody trastuzumab (trade name Herceptin) for breast cancer[167,396]. The dosage of few other medications is already based on genetic variants in the patient. In the case of the commonly used anticoagulant warfarin, taken by roughly 1.5% of the US population, the correct dosage depends on the haplotype of the metabolizing enzyme *CYP2C9* as well as variants in the drug targets *VKORC1* and *PROS*. Roughly 40% of the variation in dosing requirements of patients can be attributed to variants in these genes[191] leading to the development of genotypic dosing regimens[188] and a range of dosing algorithms[115,212,250] to avoid detrimental side effects such as major bleeding and death.

These examples show that PGx knowledge has the potential to be translated into treatment benefits for the patient. In this chapter we show, that most genetic variants seen in drug-related genes remain unstudied, however, and many more studies will be required to complete our knowledge.

Goals of the Project

The goal of this project was to create an inventory of the extent to which drugs and drug-related genes are affected by genetic variation. We focus on genetic variation in known drug-related genes, namely drug targets and ADME genes to 1) determine the amount of genetic variability due to common and rare polymorphisms in pharmacogenes, 2) investigate how much an individual in the average population is at risk of containing

functional genetic variants that may affect drug action, and 3) quantify how much this risk may change depending on geographic ancestry and drug.

Ultimately, the effect of genetic variation in pharmacogenes is just a proxy for the extent to which a drug's efficacy may be affected by altered PD or PK properties. Since the MoA for many drugs depends on the modulation of several cellular targets (Chapter 2.2), this required the development of a drug-centered summary score that combines the risk of all molecular drug targets.

We then mined a data set collating genetic variants found in 60,706 patients[236] for those drugs that showed the highest risk for being affected by functional variation in their drug targets either in the overall population or in particular subpopulations.

4.2 Materials and Methods

4.2.1 Data Sets and Data Preparation

To study the prevalence of genetic variation in drug targets and other pharmacogenes, we utilized the reported genetic variants and their respective AF in ExAC[236] for the total cohort and for different ethnic subgroups. Based on the assumption that especially variants resulting in the non-functional transcripts (including LoF) will also show an effect on drug efficacy, we focused on the set of variants in our data set that were predicted to elicit such an effect.

ExAC release 0.3 was downloaded as a variant call format (VCF) file from the project website and variants were filtered for quality, only retaining those passing the QC filter. The VCF file was annotated using Variant Effect Predictor (VEP, release 83) using the standard annotation features including SIFT and PolyPhen annotation of coding variants. We then selected only the canonical transcript for further consideration. A variant was classified to affect the function of a protein (short *non-functional*) if it fulfilled one of the following criteria:

- PolyPhen prediction `probably damaging` or `possibly damaging` and SIFT prediction `deleterious`,
- ExAC provided LoF prediction: `high confidence`

We further excluded variants whose locus was not observed at least once in each subpopulations (AN > 0, for all populations p with $p \in$ (AFR, AMR, SAS, EAS, FIN, NFE)) and in less than half of all alleles in the population.

4.2.2 Cumulative Allele Probability of a Gene

We then calculated the cumulative allele probability (CAP) for observing one of a set of non-functional variants in a gene on the population scale based on the reported AF of these variants in the ExAC data set. The AF corresponds to the probability of observing a variant allele in the population ($P(\text{allele A}) = AF(A)$). Using standard frequentist statistical approaches, the overall probability of observing one of multiple independent events is computed by multiplication of the probabilities of observing any of the events. The probability of not observing any variant allele in a set of alleles $A = (\text{variant a}, \text{variant b}, \text{variant c}, ...)$ in gene g for an individual i can thus be calculated as

$$P_{\text{ref}}(g) = \prod_{v \in A} (1 - AF(v))^2 \tag{4.1}$$

From this we computed the cumulative allele probability, that is the probability of at least one variant allele occurring in the gene on at least one chromosome as

$$\text{CAP}(g) = 1 - P_{\text{ref}}(g) \tag{4.2}$$

While these probabilities can be computed for arbitrary set of alleles in a gene g, we focused our analyses on non-functional variants, if not otherwise noted.

4.2.3 Probability of Observing a Variant in a Gene Set

For some analyses we grouped genes into larger sets associated with a particular property, e.g., all drug targets for drug d. To estimate the likelihood of an individual to carry at least one non-functional variant in this set of genes G, we further extend the statistic approach above to combine multiple genes, to a drug risk probability (DRP) for drug d with genetic targets G, defined as

$$\text{DRP}(d) = 1 - \prod_{g \in G} P_{\text{ref}}(g) \tag{4.3}$$

Comparison of Non-Functional Variants Across Populations

To investigate whether there is a difference in prevalence of non-functional variants in drug targets, we compare the probabilities for observing such variants and the resulting gene sets per drug between the human populations included in the ExAC dataset. Comparison is done using risk ratio (RR), and absolute risk difference (RD).

We calculate the *risk ratio* based on the probability of containing non-functional variants between the population exhibiting the minimal risk (group 1) and the maximal risk (group

2) for containing such a variant in a drug target. A RR of one means that the risk in both groups is identical, while a value smaller than one means that members of group 1 are RR-times less likely to have a non-functional variant in the gene or gene set that group 2.

$$RR = \frac{\text{DRP(group 1)}}{\text{DRP(group 2)}} \tag{4.4}$$

The disadvantage of ratio-based measures is that the absolute magnitude of the risk is not considered. To focus on genes with an overall high risk of containing functional variants we also compute the *absolute risk difference*,

$$RD = |\text{DRP(group 2)} - \text{DRP(group 1)}|. \tag{4.5}$$

4.3 Results

4.3.1 Genetic Variability of Drug-Related Genes Across 60K Individuals

Drug-related genes and non-functional variants

First, we investigated the extent of all non-synonymous genetic variants for 806 drug-related genes collated from DrugBank[232] and other sources[114,218] in the 60,706 human individual exomes from ExAC[236] (Figure 4.1): the AF distribution of non-synonymous variants in drug-related genes is almost identical to that of all genes in ExAC (n=17,758) with 97.5% being rare (AF $< 0.1\%$[236]) (Figures 4.1, 4.2). When filtered for variants with a negative effect on protein function, 98.7% of the 61,134 non-functional variants in drug-related genes are rare.

Nevertheless, 43% of the drug-related genes with predicted non-functional variants have at least one non-functional variant with AF $\geq 0.1\%$. The drug-related genes with the most frequent non-functional variants are membrane transporter genes related to drug efflux and uptake such as *ABCB5* (three LoF, six damaging), *SLC22A1* (nine damaging), and *SLC22A14* (eight damaging). In the clinically highly important polymorphic cytochrome P450 enzyme *CYP2D6* also eight damaging variants have been identified. Since the ExAC cohort contains an order of magnitude more individuals than previously available, it also allowed us to identify genes with many different non-functional variants even though each variant may be individually rare. The ADME genes with the most non-functional variants per residue reflect similar findings from smaller cohort studies and include the glutathione S-transferase sodium/bile transporter *SLC10A1* (0.36 variants/residue), *GSTA5* (0.31 variants/residue), and some cytochromes P450s such as *CYP1A1* (0.30 variants/residue) and *CYP2C19* (0.28 variants/residue)[218]. Furthermore, our analysis revealed drug target genes with comparable numbers of non-functional variants per residue including the dofetilide

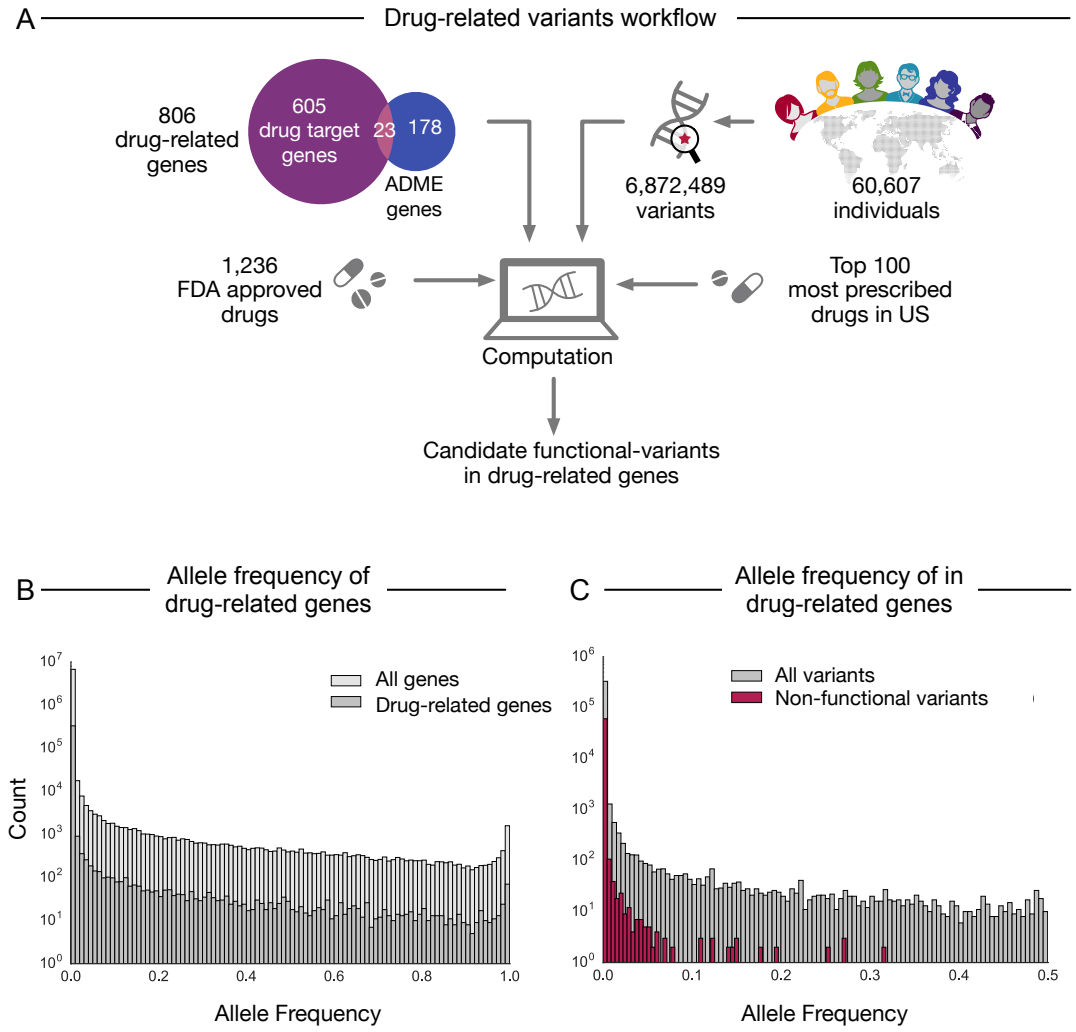


Figure 4.1: Analysis of genetic variation in drug-related genes. A) The analysis pipeline was based on exome data from ExAC[236], drug-gene relationships from DrugBank[232], and prescription information[86], followed by filtering steps and subsequent computational analysis to investigate drug-specific risks of pharmacogenetic alterations in patients. B) Comparison of the allele frequency distribution between non-synonymous variants of all human genes covered in ExAC (n=17,758) and non-synonymous variants in drug-related genes (n=806). C) Comparison of the allele frequency distribution between non-functional variants as predicted by LOFTEE[259], Polyphen-2[4] and SIFT[303] and all non-synonymous variants in the drug-related genes.

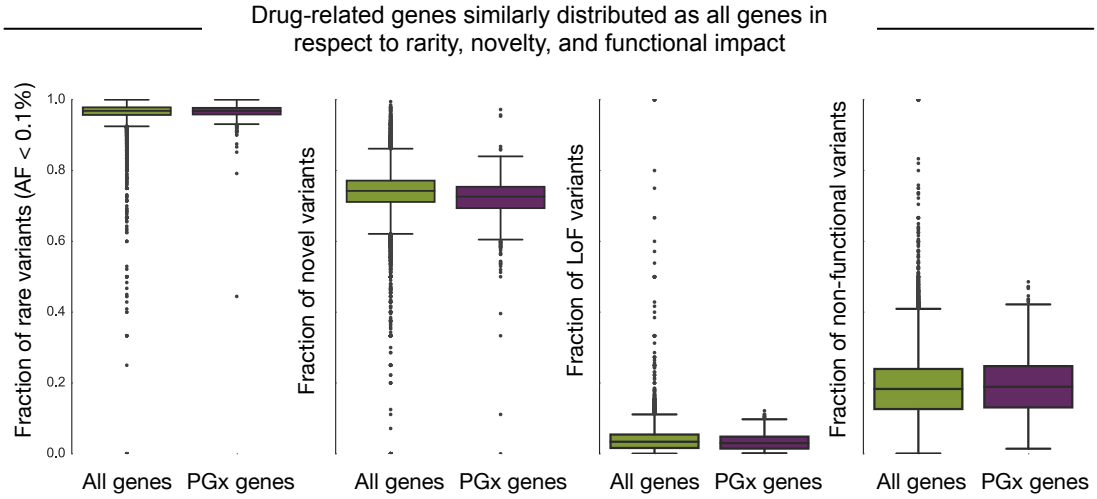


Figure 4.2: Distribution of variant properties by gene in the non-synonymous subset of the ExAC collection. From left to right: fraction of variants in each gene with allele frequencies (AF) below 0.1% for all genes compared to drug-related genes; fraction of variants in gene without corresponding entries in dbSNP, thus deemed novel; fraction of variants that result in the complete loss of the protein product (LoF) in the full data set; fraction of variants in gene that are predicted to be non-functional (LoF or damaging as predicted by SIFT and PolyPhen).

target *KCNJ12* (0.31 variants/residue) and the target for the rheumatoid arthritis drug niflumic acid, *PLA2GLB* (0.30 variants/residue).

While both metrics described above may be useful to evaluate the extent of genetic variation in the human population, they do not quantify the risk of an individual person in the population to carry non-functional variants in a particular gene. Amongst the genes with the highest probability of being affected by a non-functional variant, as computed by the CAP score, are both, ADME genes and drug targets. The ADME genes with the highest CAP scores include *NAT2* (81%, involved in metabolizing arylamine and hydrazine drugs), *CYP2D6* (59.6%, involved in the metabolism of 20% of most prescribed drugs in the US[474]) and the transporter gene *SLCO1B1* (26.0%, a high risk gene for simvastatin-related myopathy/rhabdomyolsis[295]). The drug target genes with comparable high CAPs scores include tyrosinase (*TYR*; 62.4%, targeted by the acne drug azelaic acid), the alpha-4 subunit of the GABAA receptor *GABRA4* (53%, targeted by benzodiazepines) and *F5* (20.1%, targeted by drotrecogin alpha which was withdrawn from the market due to an unacceptably high number of adverse drug reactions) (Figure 4.3). The major proportion of the CAP score for these highest "risk" genes derives from common genetic variants many of which have been observed previously. Nevertheless, for many genes a non-negligible proportion of the score is contributed by rare non-functional variants, which were identified

through the sufficiently large cohort size (lines in light purple and light blue in Figure 4.3, respectively). In addition, we estimate that more than 60% of the drug-related genes in our set are putative novel candidates for pharmacogenomic research, so far missing relevant information from clinical studies[453].

Cancer drug-target genes with many germline non-functional variants

Especially in cancer therapy, genetic variation in drug targets has been recognized to play a crucial role for treatment success[178,360]. While some cancer drugs do not act in the tumor tissue, the cancer drug's primary site of action usually is in the tumor, whose genome contains tumor-specific somatic variants as well as a subset of patient-specific germline variants[406]. Information on somatic variants from tumor samples is thus increasingly used to enable research on drug design and to implement stratified or personalized cancer therapy. However, the patient's germline genome is routinely masked in these tumor sequencing analysis protocols[178,360].We thus wanted to assess whether target genes of drugs used in cancer therapy contain germline variants in the population that may affect the drug action and may be missed by current analysis protocols. More than 15% of the drugs in this report (193 of the 1,236) are used in oncology (as defined by the WHO ATC code[458]) and between them have 163 gene targets. Several of these targets have high probabilities of having a non-functional variant in the germline genome. For some of these targets the germline risk directly corresponds to potential altered treatment effects. This is the case for the kinase *KDR* (also known as *VEGFR2*) (CAP=25%), which is targeted by sorafenib and sunitib to inhibit vascularization of the tumor site[3]. Other drug targets for cancer therapeutics with high CAP scores include *MAP4* (60%) and *TUBB1* (30%) that are targets of paclitaxel, *MAP1A* (42%) a target of estramustine, *CD3G* (39%) a target of muromonab and *PARP1* (37%) a target of olaparib (Figure 4.3). Overall, 40 cancer drug target genes, including 34 target genes with kinase domains, show CAP scores >1%. For these examples, functional germline variants are only relevant for treatment response if the tumor genome also carries them. While there is not a complete overlap between both germline and tumor genome due to loss of heterozygosity and other alterations in carcinogenesis[406], our analysis suggests that a large percentage of the population may contain non-functional variants in cancer therapeutic targets in the germline that may carry over to the cancer genome and could be easily overlooked by current sequencing analysis protocols.

Aggregating risk for non-functional variants in targets by drug highlights drug candidates for future pharmacogenomics research

About 70% of the FDA-approved drugs analyzed here do not have any pharmacogenomics data associated with them in public repositories[453]. However, our analysis shows that there

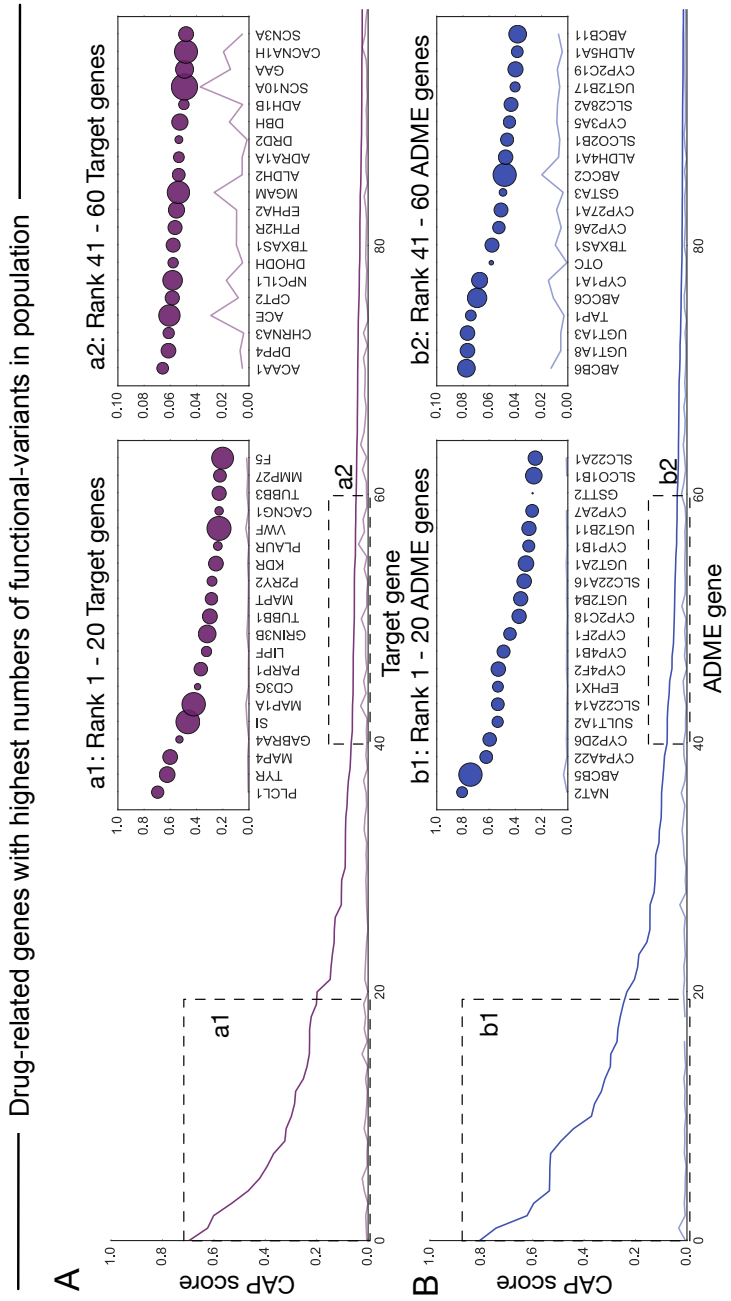


Figure 4.3: Drug-related genes with highest probability of having non-functional variants. A) Protein-centered cumulative allele probability (CAP) scores for the 100 drug targets with highest scores (purple) and the contribution of CAP scores as determined from rare variants alone (light purple). a1) Top 20 target genes with highest CAP score, a2) Examples of target genes with lower CAP scores, B) 100 ADME-genes with highest CAP scores (blue), and the corresponding CAP score determined from rare variants alone (light blue). b1) Top 20 ADME-genes with highest CAP scores, b2) Examples of ADME-genes with lower CAP scores. Bubble size corresponds to the number of non-functional variants observed for the respective gene.

are many non-functional variants in their target genes (Figure 4.4). To estimate how much each drug can be affected by non-functional variants in its target genes and to highlight possible candidates for future research, we computed the DRP for these durgs. For all FDA-approved drugs considered here (n=1,236), 43% have a DRP greater than 1%. The DRPs are weakly correlated to the number of targets (linear regression, $r^2 = 0.28$), leaving many drugs with few targets but higher than expected DRPs (determined by root mean square errors, short RMSE, of the model). For instance, one of the two human targets of azelaic acid, tyrosinase (*TYR*) is highly mutated in the population causing a DRP of 62.5% for this drug, which results in an RMSE of 0.34.

Drugs with the top DRP scores are paclitaxel and docetaxel (82%), quinacrine (70%), azelaic acid (63%), triazolam and other benzodiazepines (>50%). This means that any individual in the population has a probability of more than 50% to carry a non-functional variant that may affect the medication outcome of these drugs. Several of the drugs with high DRPs are considered "essential medicines" by the WHO[382]. In addition to paclitaxel and docetaxel, these include the opioid methadone (13.6%), the diuretic amiloride (11.7%), and the local anesthetic lidocaine (11.4%). For instance, the drug methadone targets the D- and M-type opioid receptors (*OPRD1*, *OPRM1*) and whilst some non-coding variants and a single coding variant (rs1799971) have previously been associated with required dose adjustments and treatment response, we observe another 132 non-functional variants in these target genes, which could therefore be candidates for further testing. Since variants with predicted damaging effects dominate especially the rather high DRPs, we filtered the variants for only those resulting in LoF. Restricting to these high confidence variants, the DRP decreases below 10% and the drugs with the highest DRP include the anti-cancer drug marimastat (8.3%), the anti-ulcer medication sulfacrate (8.2%), the anti-flu drug oseltamivir (6.0%) which targets human *CES1* for activation, and several liptins used for diabetes that inhibit *DPP4* (5.6%).

Commonly used drugs have high probability of a non-functional variant in their targets

We then focused our analysis on the top 100 most prescribed medications in the US (from 2013[86]) which results in a list of 77 unique drug compounds for further investigation. 42% of these drugs have a DRP score greater than 1% of containing a non-functional variant and the probability of an individual carrying a non-functional variant in any of the targets for these 77 top prescribed drugs is 81%. For some of these drugs it is already well established that there is some genetic component to drug response, even if the details are debated[28]. For instance, five of the top fifteen most prescribed drugs in the US are asthma drugs (budesonide, salbutamol, salmeterol, fluticasone, and tiotropium). Whilst each of the

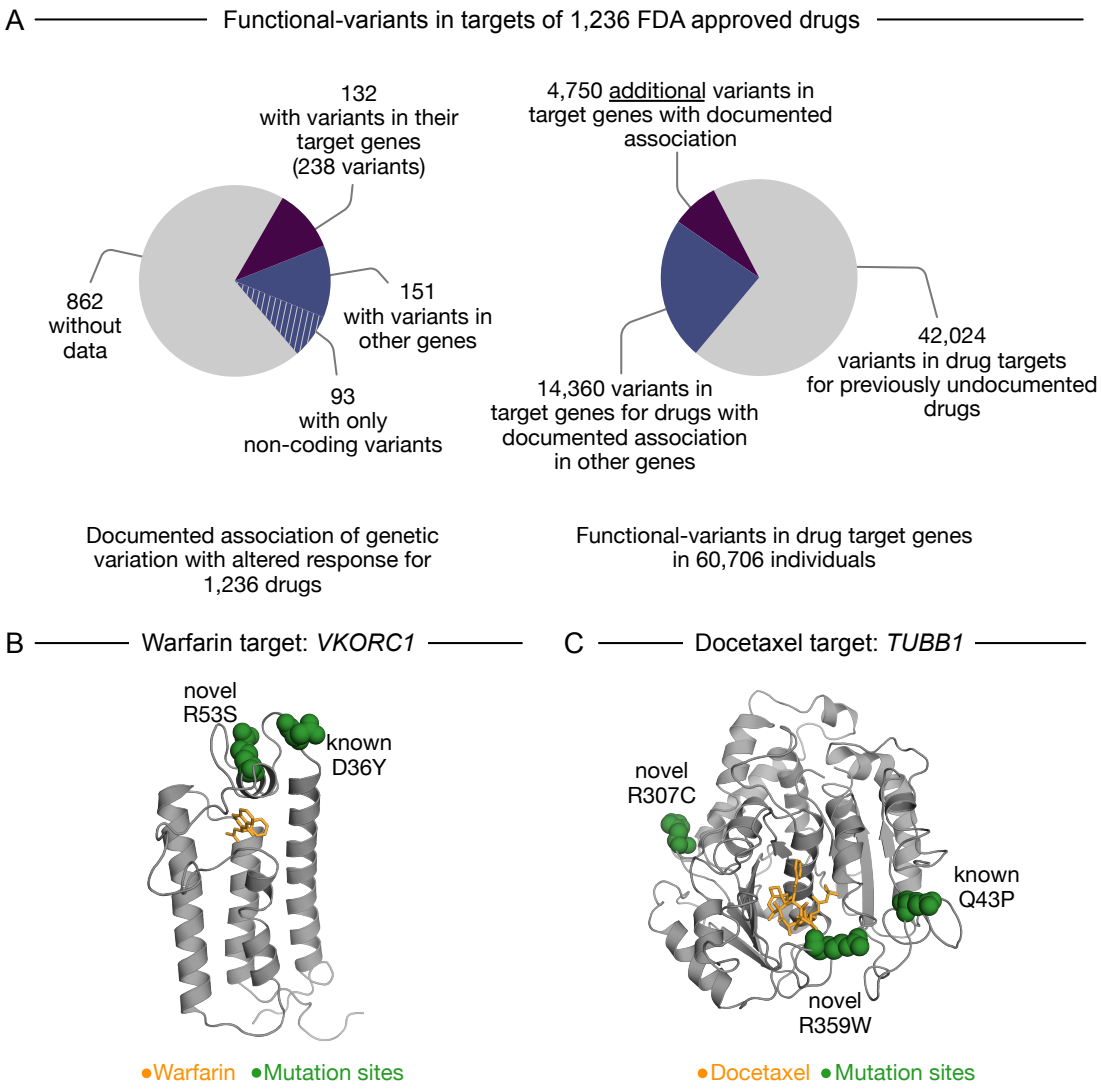


Figure 4.4: Knowledge gap between observed genetic variants in the population and documented pharmacogenomics data. A) Availability of documented pharmacogenetic associations for 1,236 FDA-approved drugs in public repositories such as the PharmGKB database[453] (left), is less abundant than non-functional variants observed in the population for the drug target genes (right). B) and C) Examples of known and novel genetic variants (green) in the target genes of warfarin and taxanes that could affect drug efficacy due to effects on the binding site (ligand highlighted in purple).

DPRs is not particularly high (ranging from 0.06% to 0.25%), their widespread prescription rate (> 100 million prescriptions in 2013) still results in thousands of individuals who may be affected by a non-functional variant. Similarly, statins (e.g., atorvastatin and rosuvastatin) are prescribed to nearly one in five adults in the US[199] and primarily target *HMGCR*. Due to genetic variation in this target gene statins have a DRP of 0.18%. This means that of the 40 million individuals who are prescribed a statin in the US, more than 80,000 individuals could be at risk of altered pharmacodynamics of statin treatment due to a non-functional variant in the target *HMGCR*. This finding is underlined by previous pharmacogenetic studies showing that *HMGCR* is the most important polymorphic gene for treatment success of statins[53].

Overall, the genetic-variability of drug targets of many of the top 100 prescribed drugs has not been systematically annotated so far (Figure 4.5), including the Alzheimer's drug memantine (DRP=7.2%), the pain-medication acetaminophen (DRP=4.7%) and the proton-pump inhibitor esomeprazole (DRP=3.1%) that all have high DRPs. While these drugs, to our knowledge, do not have known associations between non-functional variants in drug targets with drug action, clinical studies show that certain proportions of patients treated with them do not respond to treatment. The extent of this non-response is reflected by the number needed to treat, NNT[437]. For instance, for every one patient successfully treated for Alzheimer's diseases with memantine, between two and seven patients do not respond to treatment[251] (NNT=3 to 8). Similarly, the NNT for acetaminophen and its indication of pain is five[290] and for esomeprazole and reflux disease is 54[117].

4.3.2 Human Ancestry and Drug Targets

Drug-related genes show geographic difference in genetic variability

Individuals with different geographic ancestry carry genomic variants with different frequencies[150]. The six populations differentiated in ExAC are African, South Asian, East Asian, Finnish, Non-Finnish European, and the Latino[236]). About half of all non-functional variants in drug-related genes (M = 54%, SD = 15.2%) are unique to only one of the six populations and only 0.1% of non-functional variants occur with an AF $\geq$ 0.1% across all populations. Consequently, this results in drug-related genes that have a high risk of non-functional variants depending on geographic ancestry. For instance, using a cutoff of CAP > 1%, we found that 231 drug-related genes have functional variants in the cohort of European ancestry compared to 298 genes with functional variants for the cohort of African ancestry.

Nevertheless, 114 drug-related genes showed a CAP score above 1% in each population indicating that there are genes with a similar world-wide pharmacogenetic relevance. Not surprisingly, among those genes with the highest difference in CAP score between popula-

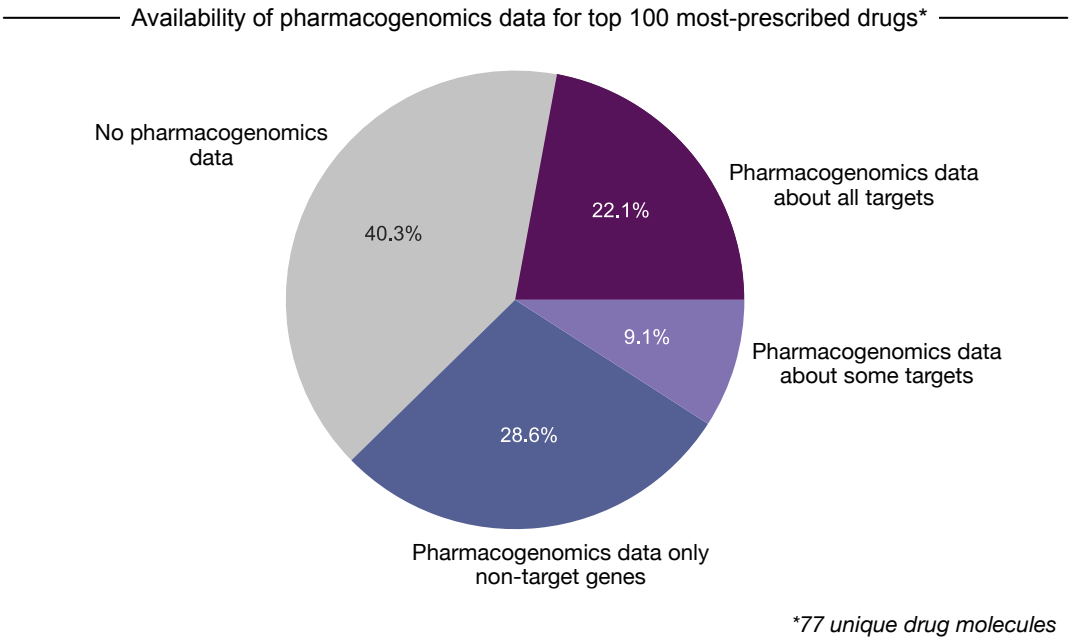


Figure 4.5: Fractions of the top 100 most prescribed drugs in the US that have established pharmacogenomics data documented in PharmGKB, either for genes documented to be the drug's pharmacological target in DrugBank (purple) or other genes, such as those related to drug ADME.

tions are many cytochrome P450s and phase II enzymes, as noted in previous studies of smaller population sizes[114]. Similarly, we observe drug target genes with markedly different CAP scores across populations. Among the target genes with the highest absolute CAP score difference are *VWF* (which is targeted by antihemophilic factor), *SIRT5* (targeted by suramin for treating sleeping sickness), and the gastric lipase *LIPF* (targeted by orlistat for obesity treatment). The latter has 65 non-functional variants and the most frequent variants differ especially between African and East Asian cohorts (CAP 8% vs 51%). Target genes with high subpopulation differences also include several targets for antineoplastic agents, such as the olaparib-target *PARP1*, for which the CAP score ranges from 10.2% in patients of African ancestry to 69.6% in Latino patients. While the efficacy of olaparib depends on the tumor genome and not the germline, the risk to carry germline-originated variants in the tumor should not be ignored. We also observed population differences in the nucleoside transporter *SLC28A1*. While the CAP score is 4% in Non-Finish Europeans, individuals with an East Asian ancestry have a risk of 60%. Interestingly, several variants in *SLC28A1* have been associated with different outcomes in non-small cell lung cancer and breast cancer[402,457] when treated with gemcitabine, suggesting that variant differences across the populations may be involved.

Analysis of the DRP score reveals a population-specific risk for several drugs

Of the 1,236 FDA approved drugs considered, 241 have more than 10% absolute difference in DRP scores between at least two sub-population cohorts and 24 of these have more than 30% DRP difference. Out of this subset of drugs, 11 belong to the 100 most prescribed drugs in the US and 28 are recommended worldwide by the WHO for their therapeutic use, including oxcarbazepine, amobarbital and dolasetron. 312 of the 1,236 drugs have a high risk (DRP>1%) in all six sub-populations (Figure 4.6A, and the DRP top 20 drugs stratified by population are illustrated in Figure 4.6B).

Well-known differences, such as response to disulfiram (treatment for chronic alcoholism), are recapitulated in the data (Figure 4.6). Specifically, the genetic variant E487K in the disulfiram target *ALDH2* (rs671) is seen in the ExAC East Asian population at similarly high frequencies as seen in previous genetic studies[98].

The different responses in the asthma-medication salbutamol and the blood-thinner warfarin have been attributed to variants in their respective drug targets, including R16G in *ADRB2* (rs1042713) for salbutamol[247] and 1639G>A (rs9923231) in *VKORC1* for warfarin[319]. Since the well-known response altering variants were not annotated by mutation prediction software as non-functional variants, we did not expect to see the drugs appear high in our ranked list of risk differences across the populations (see Discussion). Nevertheless, our analysis shows that salbutamol still has a high risk ratio between populations,

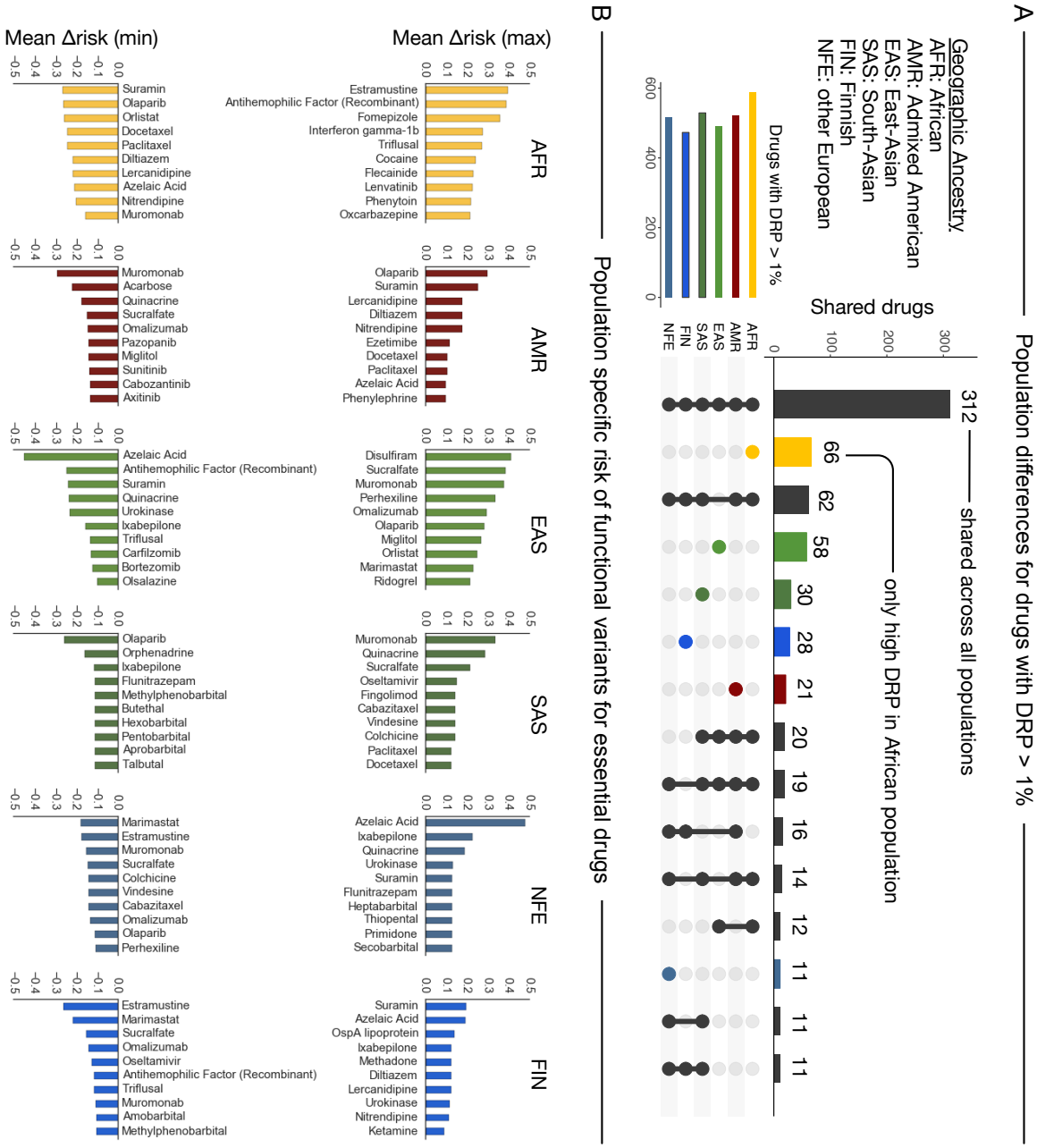


Figure 4.6: Variability of drug risk probabilities across populations. A) Number of drugs with shared (black) or private (colored) drug risk probabilities (DRP) for non-functional variants in their pharmacological target genes greater than 1%. DRP scores were calculated by aggregating the risk of functional variation across all documented pharmacological target genes of that drug. B) Drugs with highest (top) or lowest (bottom) mean DRP difference compared to all other populations indicating for which this population is at higher/lower risk of encountering non-functional variants in the target for a drug and thus higher/lower impact on drug effect.

caused by 29 variants with a dominant contribution from one variant separating the individuals of Finnish ancestry from African ancestry (rs201257377, N69S, AF_{FIN}=0.01). To our knowledge this variant has not been functionally characterized or previously associated with salbutamol response. Similarly, we observe 19 non-functional variants in the warfarin target *VKORC1* that are population-specific, including a non-functional variant observed most frequently in individuals of Non-Finnish European or Latino ancestry, (rs61742245, D36Y, AF_{NFE}=0.003, AF_{Latino}=0.001), that has been previously associated with predisposition for warfarin resistance[252]. However, 16 of the non-functional variants may be novel risk factors including a non-functional variant primarily observed in individuals of East Asian ancestry (R53S, ENST00000394975.2:c.157C>A, AF_{EAS}=0.001). Using a recent protein 3D model[69,387] of *VKORC1*, we mapped the R53S variant to the putative warfarin binding pocket (Figure 4.4). Furthermore, analysis of co-evolution in the protein using EVfold[267] shows that R53 is strongly coupled to other residues in the protein and changes in this site are predicted by EVmutation[161] to affect protein fitness due to epistatic variant effects. Together, this suggests that this mutation might be negatively associated to warfarin binding.

Trifusal, a drug for stroke re-occurrence, targets four genes (*PTGS1* (also known as Cox-1), *NOS2*, *NFKB1*, and *PDE10A*) that together have more non-functional variants in the African population than in any other population (DRP_{AFR}=37%, Figure 4.6B). This difference between populations is mainly due to a variant in *NOS2*, which occurs in the population of African ancestry with higher than average frequency (rs3730017, AF_{AFR}=19% vs AF_{global}=4%) and while not functionally characterized, has been associated with protection against cerebral malaria[423]. In *PTGS1*, three non-functional variants have allele frequencies above 0.1% in the cohort of African ancestry. The most frequent variant (rs5789, L237M, AF_{AFR}=0.5% vs AF_{global}=1.7%) lies on the dimer interface and has previously been associated with reduced metabolic activity of the enzyme[234]. A second variant is an indel, which is predicted to result in the total loss of protein function (AF_{AFR}=0.3% vs AF_{global}=0.02%). The effects of the third non-functional variant common in the African cohort (rs139956360, E259A, AF_{AFR}=0.2% vs AF_{global}=0.02%) on enzyme activity or drug binding is less clear from the three-dimensional structure of the protein and would require further exploration. Since triflusal is prescribed for prophylactic use in the same way as aspirin for stroke prevention, it is clearly worth further investigating the effects of these observed non-functional variants.

4.4 Discussion

In this chapter, we analyzed the extent of functional genetic variation in drug-related genes and its implication for 1,236 FDA-approved drugs in sequencing data of 60,706 exomes.

We show that not only the risk of carrying non-functional variants in ADME genes, but also in drug targets is high for any patient. For ADME genes this observation is in line with previous studies[90,218,459], but novel for drug-target genes. We observed non-functional variants in 98% of the drug-related genes and at least one high confidence LoF variant in 93% of the genes. The prevalence of non-functional variants in drug-related genes is thus higher than previously shown[459]. When considering drug target genes for the 100 most prescribed medications in the US the probability of carrying at least one non-functional variant is above 80% for each patient. Together with the high risk for clinically actionable variants in ADME genes (estimated at 98%[90]) these findings indicate that genetic variability may contribute significantly to observed differences in drug response between patients.

While individualized cancer therapies often focus on the somatic variants present only in tumor tissue, we can show that functional germline variants, which are routinely masked out in the analysis of somatic variants, are common in many cancer drug targets. By excluding germline variants that the tumor inherited from its progenitor cell from cancer genome analysis in the context of therapeutic decision-making may thus result in the oversight of important determinants for treatment response or resistance development. To what extent the tumor genome varies from the germline genome, is dependent on patient and cancer type. Loss of heterozygosity, where the germline allele is lost in the disease progression and copy number alterations can indeed result in drastic changes between genetic variants observed in the normal tissue of a patient and the cancer[255,406]. The presented results should thus be seen as a motivation to include all types of variants seen in the tumor tissue for clinical decision making. The high prevalence of variants in systemic cancer therapy targets, such as *KDR* for sorafenib, further indicates, that the germline variants of target genes in addition to ADME genes could be of interest for clinical decision making.

Geographic ancestry is a well-established confounding factor for drug response, but few drugs have been assessed in their efficacy across global populations. Even where clinical trials have been carried out in different populations, particularly non-European and non-Asian individuals remain understudied. By calculating risk probabilities for drugs and different populations, we showed that the frequency of non-functional variants in drug-related genes varies widely across populations. Even for drugs where population differences in response are observed, additional patient groups may be at high risk of altered PD due to genetic variants in drug targets. Especially for drugs commonly used around the world, such as those on the WHO Essential Medicines list, this could result in large numbers of patients with reduced drug efficacy in some, but not all, of the populations they are applied in.

Confidence in drugs-gene-associations

The analysis in this study relied on external data for drug variant annotation and drug-gene associations. Even though it was possible to estimate the burden of functional variation in drug-related genes and quantify to which extent individual drugs may be affected, there remain certain limitations. First of all, even manually curated drug-target associations and pharmacogenomics data are susceptible to spurious annotations. For example, some subunits of the GABA receptors including *GABRA4* are generally thought to give rise to receptors resistant to classic benzodiazepines such as diazepam[288], but have been annotated as targets for benzodiazepines. Comparison to a different, independently curated set of drug-target associations[366] further shows that annotation of drug-target pairs does not always agree.

Furthermore, to quantify the real risk for a drug, drug-specific ADME-gene relations should be incorporated into the DRP calculation. For example, optimal warfarin dosing is known to be dependent on variants in *CYP2C9* in addition to *VKORC1*[188] and variants in the ADME-gene *UGT1A1* are documented to contribute to different responses to the cancer drug irinotecan around the globe[262]. Unfortunately, comprehensive inclusion of ADME-genes in the DRP calculations is currently not possible because sufficient data for ADME-genes is lacking for most FDA approved drugs including the relative contribution of each enzyme. Our DRP estimates thus probably still underestimate the drug-specific risk of functional variation as well as population differences.

Variant effect prediction

The vast majority of variants in drug-related genes considered in this study has not been seen previously and thus lacks validated knowledge about their functional impact on drug efficacy. We therefore had to rely on predictions of their impact on protein function. The probabilities presented are based on the assumption that the functional classification is correct and represents enzyme activity or drug efficacy. The relative risk between genes is based on the assumption that there has not been a significant bias in assessment when genes already have known deleterious mutations. However, as all ML-based prediction tools, the software used in this study for variant effect prediction has several shortcomings: they are usually trained on biased sets of disease-causing variants only and have issues with circularity in training and evaluation data[134,161]. Except from LOFTEE, which is specifically developed to predict complete loss of protein function, the tools cannot distinguish between activating and deactivating effects of variants. When assessing variants not for their immediate pathological impact (for which the classifiers were originally developed), but overall effect on protein functionality, this distinction could be crucial as it may also affects downstream response to therapy.

The discrepancy between observed and predicted functional effects can be illustrated on the well-studied PGx variants in the anti-asthmatics target *ADRB2* (R16G/rs1042713, Q27E/rs1042714 and T164I/rs1800888) that all are classified as benign[247,316]. To alleviate this problem, one could include additional prediction algorithms, which comes at the risk of reduced specificity (in some cases more than half of all non-synonymous variants were classified as non-functional[218]) as all currently available methods have their individual drawbacks[140]. Reliable computational classification methods for variant effects on drug response remain scarce due to insufficient training data[140], but may arise in the future if efforts are increased to create such data, for example using novel high throughput methods such as deep mutational scans[110,278]. For the present study we chose a conservative approach to variant annotation that requires the complete loss of the protein product, which should have a marked impact on the drug, or the consensus prediction of two independent prediction tools at the expense of missing some known variants. It is thus likely that we underestimated the effect of the non-functional variants in our study.

Quality and availability of sequencing data

The use of whole exome sequencing data comes with the intrinsic limitation that only variants in protein coding regions can be detected, potentially missing pharmacologically relevant non-coding variants[141] or larger structural changes of the genome. Furthermore, even at low false-positive rates many called variants can be inaccurate[389] and several pharmacologically relevant gene families - namely CYPs, HLA and UGTs - are at high risk for variant calling errors due to the complex genetic structure of their loci[85,420]. While members of the cytochrome P450 family have indeed been found to be problematic in short-read sequencing[114], this does not apply for most other drug-related genes[218,459]. To reduce the false-positive variant calls in our survey, we included only variants of sufficient locus coverage and high quality.

Furthermore, the ExAC cohort, despite being very large in total, does not cover all populations at equal depth[236]. The power to detect very rare variants thus differs by an order of magnitude between the individual populations (from 0.01% AF for the Finnish and East Asian populations to 0.001% for Non-Finnish European). Due to legal restrictions in the underlying exome sequencing projects, sample-specific data including haplotype phase is also missing in ExAC. Epistatic effects of variants could thus not be investigated, even though they are known to exist. For example, while the single variant rs12248560 (*CYP2C18**17) results in increased CYP2C19 activity, the combination with another variant (rs28399504) is associated with LoF of the protein (*CYP2C19**4B)[218].

Future implications

Many major medical institutions have started implementing genotyping protocols for pre-emptive pharmacogenetic testing[1,84,352]. However, these usually focus on a small number of ADME-genes[90] and often only test a subset of established actionable variants using microarrays[384]. While these arrays facilitate fast and cheap screening, we show here that the vast majority of variants in drug-related genes seen in the human population is not covered. We further want to motivate that the number of genes with pharmacogenomic variants should systematically include genes implicated in drug mechanism even though only very few examples in such genes have yet been characterized well enough to be part of a dosing guideline. Furthermore, with allele frequencies below 0.1%, many non-functional variants in drug-related genes are so rare that they cannot be observed in clinical trial cohorts, but may contribute to adverse events or diffuse lack of efficacy post-marketing. In the future, this should be considered in all phases of clinical drug development and the effects of genetic variants in genes associated with PD and PK of the drug candidate should be systematically characterized.

In conclusion, large-scale sequencing efforts can be used to identify and quantify the extent of genetic variation in genes relevant for drug action and metabolism. Identification of such variants is only the first step towards better treatment decisions. Newly identified variants of pharmacogenomic importance require validation and ultimately updated dosing guidelines. The development of quality-controlled and patient-centered software solutions to combine available knowledge of pharmacologically actionable variants with a patient's genome as well as fast and accurate approaches (experimental and computational) to functionally classify novel variants will thus be of high importance for a future of personalized medicine.

Chapter 5

Drug Repurposing using the myDrug Network

5.1 Introduction

From a patient perspective, the identification of variants affecting drug efficacy and safety is only one step towards better care: their main goal is to find the right treatment for their disease. However, the loss of the molecular target for the standard of care drug due to a genetic variant can result in resistance to that drug in that patient. In this case, alternative therapeutic strategies have to be explored. Such alternatives could be derived from identifying drugs with different MoA, but also from drugs repositioned from their approved indication (*drug repurposing*).

Systematic drug repurposing has seen several success stories, particularly in cancer therapy, and relies on a multitude of computational approaches (see Chapter 2.5 for a detailed overview). While personalized medicine is moving into the clinic[90], repurposing methods include patient-specific attributes through patient-specific expression profiles[175–177,184,206,273,394]. Genetic variation, another data type associated with personalized treatment decision, has so for mainly been considered as a tool to identify new drug targets[349,365,439].

Goals of the Project

Because proteins in a cell do not act independently, we hypothesized that targeting proteins interacting with genes in the PPI neighborhood of a disease protein can affect the disease state. This project was aimed at developing drug repurposing methods that are based on exploring this neighborhood. Our approach utilizes biological networks to identify drugs in in the network-neighborhood of mutated genes and genes associated with a disease through prior knowledge (*neighborhood targets*).

The first goal of the project was thus to build a heterogeneous information network (called `myDrug HIN`) by integrating data about drug structure and mechanism with curated knowledge about diseases, cellular processes, and genetic variation. To describe effects of genetic variants on diseases and drugs, we explicitly modeled genetic variation.

We then aimed at developing repurposing methods, formulated as edge prediction problem in the `myDrug` network. Here we explored and compared a rule-based approach using neighborhood targets and a machine-learning approach where different meta paths between drugs and diseases in the network are used as input features for a Random Forest classifier.

Related Work

Due to the high popularity of drug repurposing, several methods with similar approaches have been published while this project was ongoing. One such approach also uses a heterogeneous information network to train a logarithmic regression classifier to predict novel drug-disease edges in a network[154]. This project has similar intentions as `myDrug`, but is not targeted towards personalized medicine and focused on gene expression data.

Other resources whose data structure is similar to the `myDrug HIN` are chemProt[211] and PHAROS[305] in which disease genes are linked to manually curated drug-target pairs. While chemProt implemented several GBA drug-target prediction algorithms[211], PHAROS was mainly mined for novel insights about global properties of the drug-target space[314,366] and could serve as an additional source for the `myDrug HIN`in the future.

5.2 Materials and Methods

5.2.1 Data Sets and Data Preparation

We built the `myDrug` data base from multiple data sources containing data about drugs, their target genes, PPIs, genetic variants associated with disease or altered drug response and protein domains (Figure 5.1, Supplementary Table E.2 and Supplementary Table E.3). Data integration was performed using KNIME[23] and Jupyter notebooks[334] (available at `https://github.com/kohlbacherlab/mydrug`). An automation script coordinates the initial data download and pre-processing. The user can configure this script to adjust for changing sources and user authentication for those resources that require it. After pre-processing (data cleaning and identifier mapping), the data was imported into a relational database. The imported data was then used to build an instance of the `myDrug HIN` as a graph object and further transferred into a graph database for easy access. Both, the graph object as well as the graph database instance of the `myDrug HIN` can be filtered and queried in R or Python for subsequent analysis.

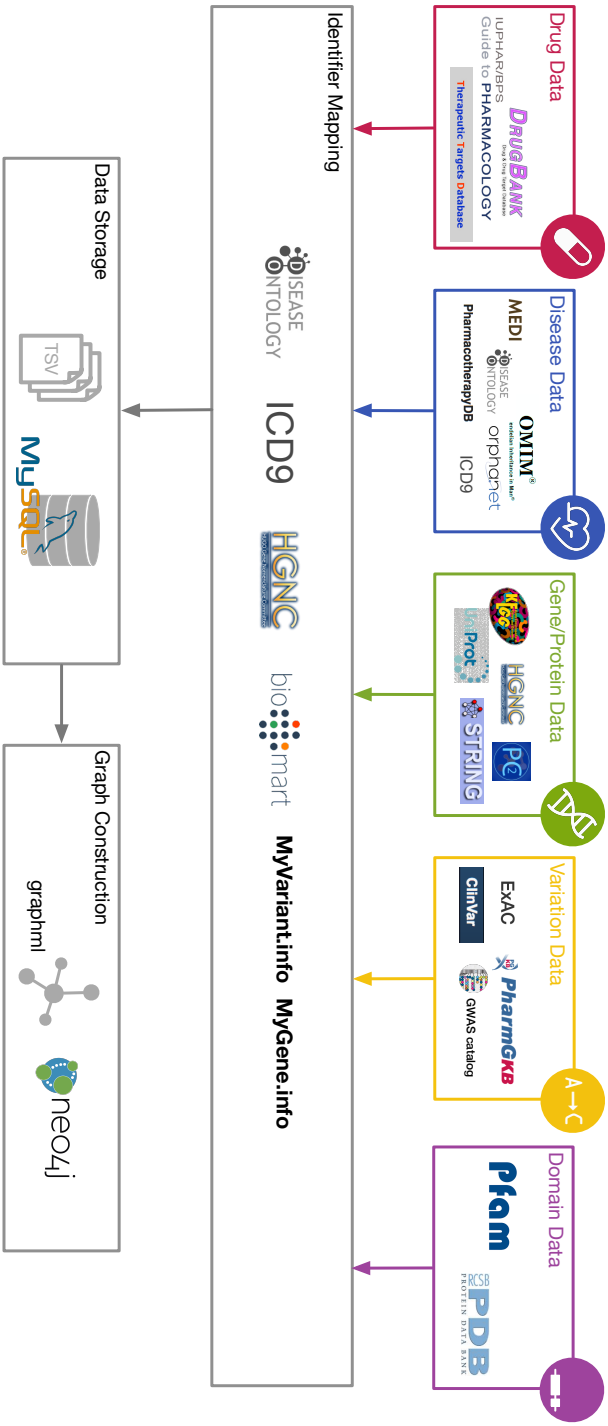


Figure 5.1: The `myDrug` construction workflow. Data about different node types including drugs, diseases and genes, were collated from multiple sources. Entity identifiers were then mapped to create a unique set of nodes and edges. The resulting data was stored in a relational database and used to build an instance of the `myDrug HIN` in a graph database.

Drug Data

Data about therapeutic compounds and biologics was downloaded from DrugBank (version 4.3)[232]. To allow later integration of other data sources, compounds were assigned internal identifiers (`mydrug IDs`). Unique SMILES were created for each small molecule compound using RDKit[227]. From the SMILES three different molecular fingerprints (Morgan, RDKit and MACCS) were computed to create pairwise similarity measures between drugs (see Drug Relations below for more information).

Disease Data

We included multiple disease dictionaries into the `myDrug` database due to their complementary advantages: i) Online Mendelian Inheritance in Men (OMIM)[139] for diseases with a strong level of genetic evidence, ii) International Classification of Disease (ICD) codes[48] for a broad coverage of diseases known in humans, and iii) Disease Ontology (DO)[375] as a common ontology for a range of disease vocabularies. Cross-mapping between identifiers was done using mapping files provided by DO, Orphanet[449], the Unified Medical Language System (UMLS)[32] and supplemental data from network medicine studies[121,235,323].

Gene/Protein Data

The HUGO Gene Nomenclature Commitee (HGNC), as the global authority for a standardised human gene nomeclature[128], provided the base set for gene and protein data. Protein-coding genes were annotated with information from the UniProt database[64] and protein domain-information from the high confidence subset in Pfam (Pfam A)[106].

Pathway names and the gene sets comprising these pathways were extracted from KEGG pathway xml files[198,312].

Variation Data

Allele frequencies and variation effect information of genetic variants in human genes was obtained from ExAC[236]. Furthermore, variants from dbSNP[388] with implications in disease or pharmacogenomics were included.

Drug Relations

Several types of relations involving drug compounds were integrated in the `myDrug` resources:

- **Molecular Similarity**: drug-drug edges for small molecule drugs with RDKit-fingerprint-based Tanimoto similarity > 0.85

- **Pharmacokinetic Interactions**: drug-gene interactions implicated in drug mechanism were obtained from the DrugBank[232] xml file and the IUPHAR "interactions" table[385]
- **Drug Indications**: drug-disease links from TTD[465], pharmacotherapyDB[154], MEDI[444]
- **Drug-Related Variants**: drug-variant links for pharmacogenomic associations from PharmGKB anntotation files[453]. Only associations labeled significant by PharmGKB were included in the `myDrug` HIN.

Disease Relations

Genetic information about disease etiology was added using OMIM, CTD and orphanet. Pharmacologically relevant associations between diseases and genes were further incorporated through TTD and CTD, but these partially stem from transitive closure of drug-target and drug-disease associations[465] and may thus contain noise. Specific disease-variant edges were extracted from ClinVar[228] and the GWAS Catalog[450] (with p-value $< 5 \times 10^{-8}$ and not classified as "benign").

Gene and Protein Relations

Sequence similarity-based protein-protein links were included from SIMAP2[11] for all pairs of proteins in the Uniprot database. In addition to that we included pairwise PPI information from KEGG (parsing the individual pathway xml files), STRING database[111], and Pathway Commons[50].

STRING contains protein associations from co-expression data, high and low throughput protein interaction experiments, literature mining and protein homology. A connection of two proteins/genes is associated with a confidence score, ranging from 1.0 (total confidence) to 0.0 (no confidence). We filtered the full data set to only contain high confidence associations (score ≥ 0.7) that showed up in at least two of the following channels: low throughput databases, high throughput databases, text-mining and co-expression.

Variants falling into coding regions or adjacent UTRs of a gene were linked to these genes using dbSNP identifiers and myvariant.info[461].

Domain Relations

Protein domains were linked to genes using the human proteome data set from the Pfam FTP server[106]. To link drugs to protein domains, we extracted ligand binding sites from protein 3D structures in the PDB[22] (minimal atom distance of protein to ligand of 4Å) and mapped them to Pfam domains using SIFTS[429].

Identifier Mapping

Drug IDs were mapped to `mydrug IDs` via DrugBank IDs, where available, and otherwise through name and synonym mapping. For MEDI, we used drug SMILES from chemspider[455] using the Python module chemspipy[411] and matched them using RDK-based canonical SMILES. Diseases were mapped using the ICD system (for TTD and MEDI) or DO terms (pharmacotherapyDB) using KNIME.

Variants were required to contain a dbSNP identifier (thus excluding structural variants) and diseases were mapped to the three supported vocabularies either using the Experimental Factor Ontology (EFO) for the GWAS catalog or using mapping files in the case of ClinVar.

Ensembl protein IDs were mapped to HGNC through the STRING alias file, myGene.info[460] and biomart[398] queries.

5.2.2 myDrug Database

The pre-processing step converted the data after identifier-mapping into structured data tables that were stored in a relational database implemented in MySQL. Here, we retained meta-information about the original data sources as columns in the database tables.

We used the Python package SQLAlchemy[19] to fill and query the MySQL database.

Update Protocol

The `myDrug` database can be updated through a complete re-population in which current releases from all source databases are downloaded, mapped, and stored as text files. These can then be imported into a new version of the database.

Each of the included databases has their own update schedule ranging from weekly (ClinVar, OMIM) to sporadic (DrugBank) to never (MEDI). To balance the overhead of extracting and unifying the data from the individual sources with the higher precision obtained by the most up-to-date version of the underlying databases, we suggest an update cycle that follows that of most included sources, which is at least once per year, but this can be adjusted as needed.

5.2.3 myDrug Network Construction

A heterogeneous information network G_{myDrug} was defined as the directed graph $G_{\text{myDrug}} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. The network schema $T_{\text{myDrug}} = (\mathcal{A}, \mathcal{R})$ consist of the vertex typing function $\tau = v \to \mathcal{A}$ that relates each vertex $v \in \mathcal{V}$ to a type in the vertex type set $\mathcal{A} = (D, M, P, V, N, R)$ and an edge typing function $\phi : \mathcal{E} \to \mathcal{R}$ with $\mathcal{R}$, the relation type set, containing all pairwise combinations of vertex types that have data in the database.

Several network instances were constructed from different subset of the data in the MySQL database with vertices being collated from the tables for drugs (D), diseases (M), genes/proteins (P), genetic variants (V), pathways (N) and protein domains (R). Edges were constructed from the relation tables and typed by their source and target vertex types. By default all edges were defined as bi-directional unless directionality information existed. We created a default instance of the `myDrug HIN`, initialised using a subset of the data included in the `myDrug` database. Specifically, only PPI present in KEGG were included for protein-protein relations and DO was chosen as disease vocabulary. To query the neo4j instance of the `myDrug HIN` we used the REST interface provided by the neo4j server and the py2neo Python package[397] .

Edge weights

Biologically meaningful information about edge weights only exists for edges based on similarity (gene-gene, drug-drug). The raw data of the gene interaction networks in STRING further contain confidence scores. However, due to the sparsity of this information, the `myDrug HIN` is unweighted for subsequent analyses, unless otherwise stated.

We further implemented a topology-based edge weight based on the arithmetic mean of the edge-type specific node degrees. For edge-type et between vertices u and v, we define the promiscuity weight w as,

$$w_{uv} = \frac{\deg^+_{et}(u) + \deg^-_{et}(v)}{2} \tag{5.1}$$

with $\deg^+_{et}(v)$ denoting the outdegree and $\deg^-_{et}(v)$ the indegree of v when only considering edges of edge type et.

To down-weigh paths leading through promiscuous nodes, i.e., hubs, in the network, we define the weight of a path $p = v_1 \xrightarrow{e_{et_1}} v_2 \rightarrow ... \xrightarrow{e_{et_{n-1}}} v_n$ in the graph as

$$w_{v_1 v_n} = \frac{1}{\sum_{e \in E} w_e} \tag{5.2}$$

with $E = (e_{et_1}, ..., e_{et_{n-1}})$ defined as the set of edges of pre-defined edge types on the path p.

For the analyses of path-weight distributions in the rule-based predictions, we further normalized the path weights for each predicted drug-disease pair to z-scores, i.e. subtracting the average path weight of all predictions and dividing by the standard deviation.

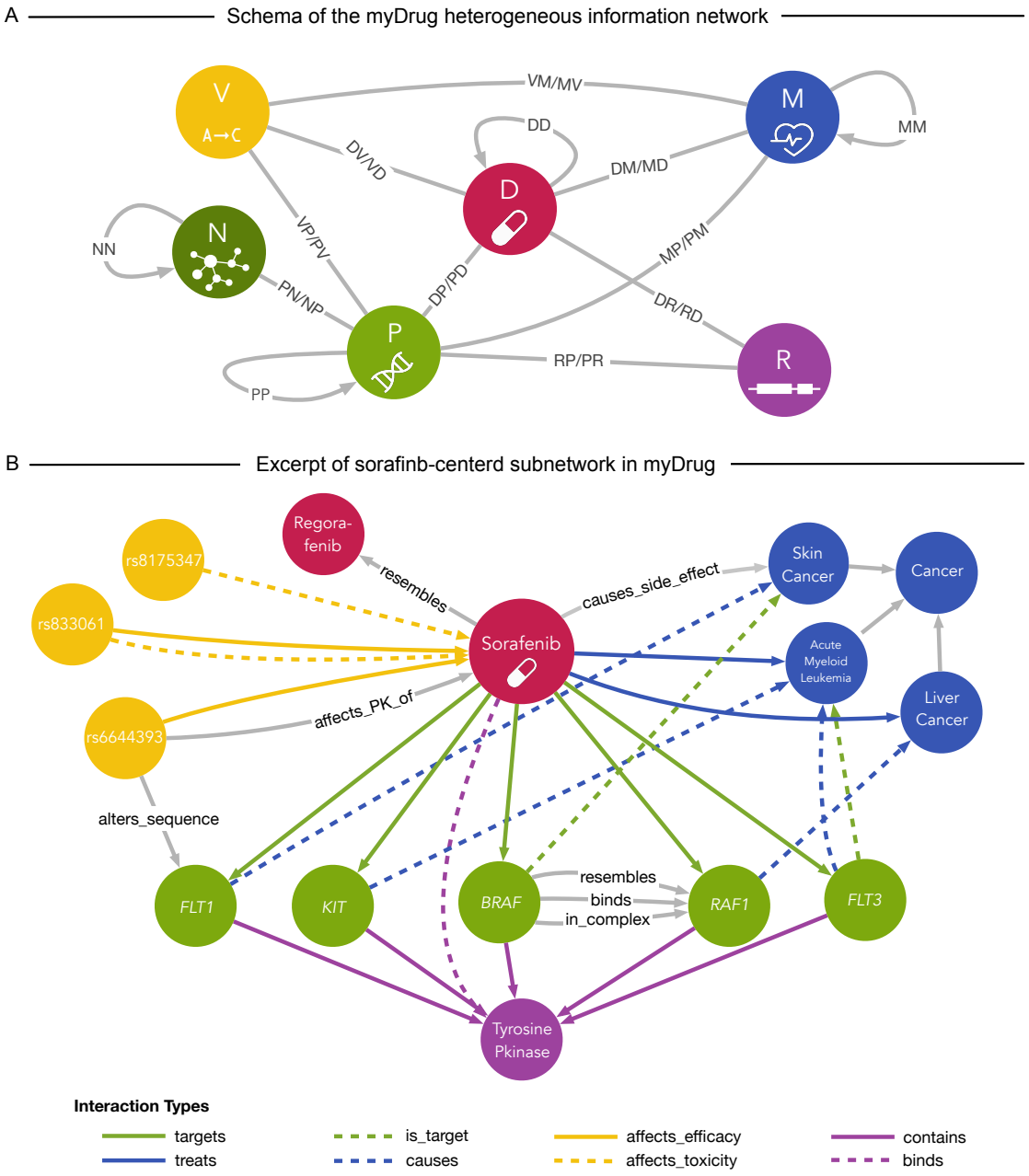


Figure 5.2: The network schema of the `myDrug` heterogeneous information network. A) Visualization of the overall network schema containing six different vertex types, corresponding to drugs (D), target genes/protein (P), genetic variants (V), protein domains (R), cellular pathways (N), and diseases (M), as well as edges between those nodes. B) Illustration of a populated network `myDrug HIN` instance using part of the sub-network centered around the cancer drug sorafenib.

5.2.4 Rule-Based Drug-Disease Edge Inference

We devised an unsupervised edge-prediction method that searches for drugs whose pharmacological targets are adjacent to genes involved in a disease based on the `myDrug HIN` using two meta paths in the network (Table 5.1).

Table 5.1: Meta paths and path instances used for rule-based neighborhood drug repurposing

Rule	Meta Path
Drug and Disease share target gene	D - P - M
Drug target adjacent to gene involved in disease in cellular gene network	D - P - P - M

Algorithm

We implemented the rule-based approach as an extension of the breadth-first search (BFS) algorithm in which the user can define a meta path n and a set of query nodes s from which the search on graph G is initiated (Algorithm 1). The edge sequence can be defined as the successive pairing of the node types along the meta path or by edge attributes if different types of edges exist.

To exclude results that could have been found with shorter meta path queries, these subpaths of the query are explored simultaneously in the BFS and nodes reached by them are stored in each step. For example, to find all possible diseases linked to a drug's neighborhood target, we would define the edge sequence as `[Compound-Gene,Gene-Gene,Gene-Disease]`. We would further explore the supporting edge sequences `[Compound-Disease]` to exclude the drug's main indication from the result set and `[Compound-Gene,Gene-Disease]` for predictions obtainable through a shorter path.

Implementation Details

We implemented the extended BFS algorithm and functions for data handling in the Python package `mydrug` (available at `https://github.com/kohlbacherlab/mydrug`). The class `myDrug` is implemented as an extension of a network-unaware base class `myDrugBase` and uses the python bindings of iGraph[68] to model graph objects.

All operations on the graph are implemented as class functions including the extended BFS for meta path extraction, setter and getter methods, and functions that retrieve vertex- and edge-properties from the `myDrug` database.

Algorithm 1: Extended breadth-first search used to find novel indications for drugs starting with a graph, a query node and a predefined meta path edge sequence.

```
Data: graph G, source node s, edge sequence es, supporting edge sequence ses
Result: Path: object containing information about the edge sequences and the meta path instance.
Function BFSmodified(G, s, es, ses)
    level = {s: ∅}; parent = {s: None}; i = 1;
    frontier = [s];
    medges={}; sedges={};
    while frontier ≠ ∅ do
        next = [];
        # Initialize output edges
        medges[(i-1, es[i-1])] = [];
        for supp ∈ ses[i-1] do
            sedges[(i-1, supp)] = [];
        end
        for u ∈ frontier do
            # Add edges along main meta path
            for e ∈ G.edges(source=u, edgetype=es[i-1]) do
                v = e.target;
                if v ∉ level then
                    level[v] = i;
                    parent[v] = u;
                    next.append(v);
                    medges[(i, es[i-1])] += e;
                end
            end
            # Add edges along supporting meta path
            for et ∈ ses[i-1] do
                for e ∈ G.edges(source=u, edgetype=et) do
                    v = e.target;
                    if v ∉ level then
                        level[v] = i;
                        parent[v] = u;
                        sedges[(i, es[i-1])] += e;
                    end
                end
            end
            frontier = next;
            i += 1;
            # End search if end of meta path is reached.
            if i > |es| then
                break;
            end
        end
    end
    return medges, sedeges
end
```

5.2.5 Evaluation of Repurposing Predictions

The rule-based repurposing approach was first evaluated by estimating its potential to retrieve known drug-disease associations. For each drug we computed how many original indications could be retrieved using the two rules. For this, we did not filter out known indications in the result set and then determined the fraction pr_d of known drug indications I_d in the predicted set P_d for a drug as

$$pr_d = \frac{|\{x : x \in P_d \wedge x \in I_d\}|}{|P_d|} \tag{5.3}$$

In analogy to the validation method employed by Chiang and Butte[59], we then quantified the enrichment of drug-disease pairs in the prediction set that are established as label or off-label use. For this we devised the confusion matrix, Table 5.2 and define the enrichment as $e = a * (c + d)/c * (a + b)$.

Table 5.2: Confusion matrix between drug-disease pairs in prediction set and already used in the clinical practice

		Used in clinical practice		
		Yes	No	Total
In prediction	Yes	a	b	$a + b$
	No	c	d	$c + d$
	Total	$a + c$	$b + d$	N

If $e > 1$, predicted drug-disease pairs are e-times more likely to be already applied in medical practice. We further calculated the significance of this enrichment using a hypergeometric test employing the functions implemented in SciPy[193].

External Validation Using Clinical Trials Data

To further validate our approach, we tested how many predicted drug-disease pairs are already under clinical investigation because this indicates that sufficient medical evidence data existed to support the possibility of a mechanistic association.

Clinical trial data was downloaded from the aact database (∼210,000 records, `clinicaltrials.org`, March 2016). We then mapped intervention and condition IDs (NCT_IDs) of the clinical trials database to medical subject headings (MeSH terms) using the MeSH thesaurus. Intervention MeSH terms were then converted to DrugBank IDs (`https://raw.githubusercontent.com/olegursu/drugtarget/master/identifiers.tsv`), followed by the internal mapping to myDrug IDs. Condition MesH terms were mapped to Disease Ontology IDs (DOIDs) using external identifiers provided in the Disease Ontology.

Enrichment of predicted drug-disease pairs under clinical evaluation was calculated analogously to that of approved drug-disease pairs, computing the confusion matrix of drug-disease pairs in the prediction set and in the clinical trials database.

5.2.6 Machine Learning-Based Drug - Disease Edge Inference

The mechanisms in which drugs act in a cell to alleviate a disease are diverse[121]. We thus extended the rule-based repurposing strategy to incorporate additional rules and learned their respective importance from known drug-disease pairs using a Random Forest classifier.

Meta Path features

Features to be used in the RF classifier were defined from a set of biologically plausible meta paths connecting a drug to a disease (Table 5.3) and extracted from the neo4j `myDrug` instance using Cypher queries for each `Compound-Disease` (D-M) pair in the network.

Two types of feature vectors were implemented and used as input to training the classifer in order to evaluate if network topology data could yield more accurate predictions: 1) a binary vector where the bit is set to 1 if at least one instance of that meta path exists and 0 otherwise, and 2) a vector recording the absolute count of observed instances per meta path and drug-disease pair.

Random Forest Classifiers for Drug-Disease Edge Prediction

We used *Random Forests*[38,157], as implemented in scikit-learn[331], to classify drug-disease pairs as new edges in the graph, due to their overall robustness against noisy data[38,145]. An RF classifier consists of an ensemble of decision trees in which each tree is built from a bootstrap sample of the full data set and a random subset of features at each split. This combination of bootstrap aggregation and random variable selection results in classifiers that are robust to noise in the data set, less prone to overfitting[145] and are not affected by correlated features[38].

The optimal classifier for a given set of features was determined using exhaustive grid search of included trees and maximum tree depth. The search space contained forest sizes of 100 to 10,000 trees and unpruned trees as well as maximum tree depths between five and nine levels. Within the grid search, 5-fold CV was employed to evaluate the models. We then performed the assessment of RF classifier performance for different feature subsets based on the output of the grid search, using ensembles consisting of 10,000 unpruned decision trees. The maximum number of features to consider in each split was set to $\sqrt{n}$, with n being the length of the feature vector[145].

The data set used for classifier training and evaluation was built from positive drug-disease tuples obtained through the meta path `Compound-treats-Disease`, and a negative

Table 5.3: Meta paths and path instances used for feature creation of neighborhood drug repurposing

Feature	Rule	Meta Path
F1	Drug resembles drug used to treat disease	D - D - M
F2	Drug target is a disease gene	D - P - M
F3	Drug targets a gene that is mutated in disease	D - P - V - M
F4	Drug is affected by variant that also affects disease	D - V - M
F5	Drug binds to domain that also binds a drug used to treat disease	D - R - D - M
F6	Drug binds to domain that also occurs in disease gene	D - R - P - M
F7	Drug resembles drug that binds gene that is also a disease gene	D - D - P - M
F8	Drug binds to same domain as other drug used to treat disease	D - R - D - M
F9	Drug binds to same domain as other drug that shares a target with disease	D - R - D - P - M
F10	Drug resembles a drug that binds to domain that also occurs in disease gene	D - D - R - P - M
F11	Drug targets gene in same pathway as disease gene	D - P - N - P - M
F12	Drug binds to domain of a gene that is in the same pathway as the disease gene	D - R - P - N - P - M
F13	Drug is affected by variant in a gene that is mutated in disease	D - V - P - V - M
F14	Drug is affected by variant in a gene that is in the same pathway as a gene mutated in disease	D - V - P - N - P - V - M
F15	Drug target gene is neighbor to gene involved in disease	D - P - P - M
F16	Drug target gene is close to disease gene	D - P - P - P - M
F17	Drug resembles drug that binds gene adjacent to disease gene	D - D - P - P - M
F18	Drug resembles drug that binds gene close to disease gene	D - D - P - P - P - M
F19	Drug binds to same domain as a drug that targets a gene adjacent to a disease gene	D - R - D - P - P - M
F20	Drug targets gene that is adjacent to a gene mutated in disease	D - P - P - V - M
F21	Drug is affected by variant in a gene that is adjacent to a disease gene	D - V - P - P - M
F22	Drug is affected by variant in a gene that is adjacent to a gene mutated in disease	D - V - P - P - V - M
F23	Drug resembles a drug that binds to domain that also occurs in a gene adjacent to a disease gene	D - D - R - P - P - M

D = drug, P = gene (protein), M = disease (morbidity), R = domain (region), V = variant, N = pathway (network).

set contained pairs that were linked in the network through non-causal edges (`Compound-[treats_not,treats_symptoms]-Disease`). The resulting data set is highly imbalanced with 6,047 positive drug-disease pairs and only 631 negative pairs. Assuming that the majority of all possible drug-disease pairs is in reality not connected, we enriched the negative set with 14,974 randomly chosen unconnected compound-disease pairs, resulting in a total set of 21,652 pairs. This set was then split in half, resulting in a training set and a validation set for performance evaluation and all data points in which none of the meta paths existed in the `myDrug HIN` (i.e., all features were zero) were excluded.

We evaluated the classifiers for drug-disease edge prediction using different combinations of the features defined in Table 5.3 as feature vectors:

1. gene-gene edges of different data sources treated separately
2. gene-gene edges of different data sources and interaction sub-types (e.g. `interacts_with`,`subsequent-catalysis`, ...) treated separately

Cross-Validation

All classifiers were evaluated using 10-fold cross-validation and several scores for classifier performance were calculated. These included the accuracy, precision, recall, F1 score, as well as the area under the receiver operating curve for non-binary classifiers. The average precision score describes the mean of precisions obtained every time a new positive sample is recalled over the threshold interval of (0,1) and approximates the area under the precision-recall curve.

5.3 Results

5.3.1 Overview over the myDrug Network

The full `myDrug` network contains 3,065,016 edges between 95,334 vertices (excluding genetic variants) including 8,221 drugs, 9,304 DO diseases, and 16,306 protein domains (Supplementary Tables E.2 and E.3). Furthermore, 45,746 variants that are directly linked to diseases or drugs are also included in the `myDrug HIN` used for subsequent analyses.

Disease- and drug-relations in the network

2,234 of the 8,221 drugs were FDA-approved and 1,845 drug compounds were linked to 554 distinct disease DO terms (6,678 edges). 2,233 diseases had genetic data associated and 375 diseases had both drug and gene data. On average, if a compound or disease had any information, a drug is connected to 3.6 diseases (median=2.0) and a disease is associated with 12.0 compounds (median=3.0). The disease with the highest number of associated drugs was acute myocardial infarction (total=258). Doxorubicin was the drug associated with the most diseases (total=32, 88% of these are cancer types).

Protein-relations in the network

The collection of KEGG pathways contained information about 6,998 genes of which 5,053 were connected in the network. This covered 36.0% of the 19,465 protein coding genes

included in our data set. About 60% (4,888) of the drug compounds have links to drug-related genes, including 2,395 drug target genes and 346 PD genes. Drugs with target information had on average 2.5 drug targets (max=144 for NADH, median=1.0) and each gene in the network was on average targeted by 0.3 drugs. If a gene serves as a target, it had an average of 5.1 (max=140 for *CDK2*, median=2.0) drugs associated with it.

Protein domains targeted by drugs

Based on co-crystallized drug-protein complexes, we also extracted associations between compounds and the protein domains they bind to. Using this approach, we extracted domain information for 58% of the compounds in the `myDrug` network. These bind on average to 3.6 different Pfam A domains (median=1.0). 214 compounds have been co-crystallized with ten or more domains and 13 compounds with more than 140 domains. Each Pfam domain is associated with an average of one drug. Again, this distribution is skewed, and those domains that have any drug association, bind 5.4 drugs on average (median=2.0). The domains linked to the most drugs are the kinase domains (451 ligands bind *protein kinase* domain, 11 of these FDA-approved, and 123 ligands bind to *protein tyrosine kinase* domain, 9 of which are FDA-approved), *trypsin* (332 ligands, 23 FDA approved) and the ligand-binding domain of *nuclear hormone receptor* (191, 25 FDA-approved).

5.3.2 Neighborhood-Targets

We predicted indications for 2,234 FDA-approved drugs using the meta paths F2 (`drug-gene-disease`, D-P-M) and F15 (`drug-gene-gene-disease`, D-P-P-M) where only KEGG data was used for gene-gene edges. Drug-disease prediction returned at least one new indication for 1,457 drugs using the D-P-M meta path and for 1,204 drugs using D-P-P-M meta path. 756 of the FDA-approved drugs did not contain target annotation and thus no paths could be constructed. For 22 drugs no instance of F2 and for 274 drugs no instance of the F15 existed despite annotated drug targets. Based on the general architecture of the algorithm, no pairs already covered in the shorter meta path D-P-M were allowed in the D-P-P-M set. For the 1,457 drugs with predictions using the F2 rule, an average of 38 diseases (median=26) were predicted and for the 1,204 drugs with DPPM-based predictions 70 indications (median=43) (Figure 5.3).

Based on the included data and rational drug development concepts, we expect a significant enrichment of known drug indications in the F2- and less so in the F15-based predictions. Indeed, 1,258 of the 55,358 DPM-inferred pairs (2.3%) are known drug indications represented in the `myDrug` network. This is a four-fold enrichment over the 2,559 indications in the non-predicted set of all 513,864 possible pairs of drugs and diseases that

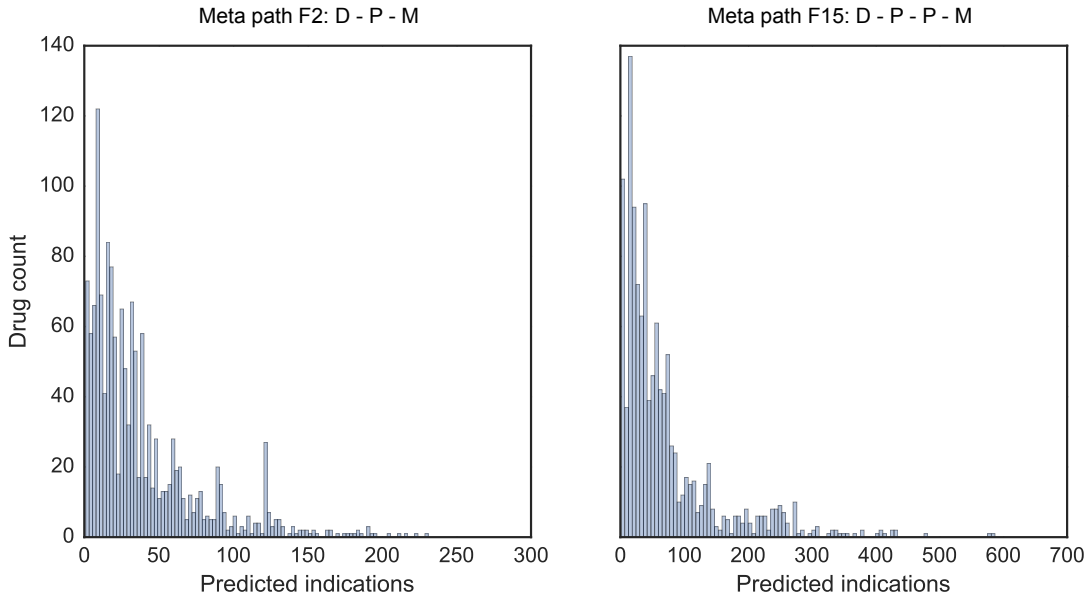


Figure 5.3: Distribution of disease predictions per drug based on a) `drug-gene-disease` meta path and b) `drug-gene-gene-disease` meta path using interactions between proteins based on KEGG pathways.

are potentially reachable in the network (29x enrichment over all 3,269,112 pairs predictable from 2,233 diseases with genetic information). Of the 84,815 F15-based predictions, only 313 are known indications (0.4%).

External evaluation

The comparison of drug-disease predictions to known treatment options indicates that most F15-based predictions are new. To test if these predictions have additional evidence in clinical studies, we evaluated the overlap between the prediction set and 15,485 drug-disease pairs in the FDA clinical trials database (covering only the drugs and diseases also present in the prediction set): Of the 55,358 F2-inferred pairs, 1,809 drug-disease combinations were covered by clinical trials (8x enrichment). Additionally, 1,560 of the 84,815 F15-inferred drug-disease edges were under clinical investigation (4.5x enrichment).

Impact of hubs on prediction accuracy

We further expect that hubs in the network add noise to the predictions. Therefore we implemented edge weights reflecting the node promiscuity as the average of the source vertex' outdegree and the target's indegree. Meta path instances were then assigned a promiscuity-score calculated as the reciprocal of the cumulative edge weights (see Methods). These scores where then normalized by subtracting the average promiscuity score of the set

and dividing by the standard deviation. Comparison of the promiscuity-score distribution in the full set of predictions to those deemed correct due to their membership in either the indication or the clinical trial set, shows that correct predictions deviate significantly from the overall distribution with non-zero average z-scores ($mean_{all}$=1.04e-13, $mean_{indication}$=-0.05, $mean_{clinical\ trials}$=-0.14, $p<0.05$ for both subsets using two-sided Kolmogorov-Smirnov tests).

5.3.3 Systematic Repurposing Using Random-Forest Classifiers

Next, we defined the drug repurposing problem as a supervised edge-prediction task (specifically, drug - disease edges). To include the full extent of information available in the `myDrug HIN`, we extended the rule-set to incorporate a variety of different meta paths (Table 5.3). These meta paths were used to construct feature vectors for drug-disease pairs which served as input to train several RF classifiers.

Binary feature vector compared to path-instance counts

Meta paths in the network could be used in two ways to construct feature vectors as input for the RF. While a binary feature vector only includes the information whether at least one such meta path exists for a drug-disease pair, absolute count features harbor implicit information about network topology and hubs.

We compared RF classifiers trained on binary feature vectors to those using absolute feature counts to assess which yield more accurate predictions, but for RFs of 10,000 trees hardly any difference in the performance metrics could be observed between binary and absolute count feature vectors (average precision = 0.85, Table 5.4).

Table 5.4: Classification performance for binary and absolute count features for RFs consisting of 10,000 unpruned classification trees.

Metric	Binary	Absolute Counts
F1	0.796	0.804
Accuracy	0.816	0.821
Precision	0.824	0.822
Recall	0.770	0.786
MCC	0.630	0.639
AUC	0.882	0.888
Average precision	0.851	0.854

RF performance based on different gene interaction networks

We then compared the performance of a classifier trained on features created from a HIN instance built from all data sources, including all three PPI network sources (KEGG, String and PathwayCommons), to classifiers utilizing only interactions from a single PPI source for gene-gene meta path features. Here, mean prediction accuracy ranges from 0.70 (KEGG) to 0.80 (PathwayCommons) and AUCs from 0.76 (KEGG) to 0.87 (PathwayCommons), which were all outperformed by the classifier built from all features (Table 5.5).

Table 5.5: Classification performance for absolute-count features across different gene network sources for RFs consisting of 10,000 unpruned classification trees

Metric	All	KEGG	PC	STRING
F1	0.805	0.619	0.789	0.755
Accuracy	0.822	0.699	0.806	0.76
Precision	0.823	0.757	0.802	0.722
Recall	0.787	0.524	0.777	0.790
MCC	0.641	0.402	0.609	0.523
AUC	0.888	0.761	0.868	0.824
Average precision	0.854	0.752	0.841	0.805

Furthermore, the feature vectors containing a full set of features resulted in overall increased prediction performance including higher AUCs and average precision than a classifier trained with a reduced set of features ignoring interaction sub-types (Figure 5.4). Here, the AUC increases from 0.87 for a classifier built on feature vectors using a single feature per network to 0.89 when treating PPI sources and interaction sub-types separately in the feature vector. Average precision changes from 0.84 to 0.85.

5.3.4 Relative Feature Importance

The previous subsection demonstrated that information about interaction types in the feature vector increase classifier performance and we wanted to know which of them contribute most to prediction accuracy. We obtained the relative ranking of feature importance for the RF classifier with highest prediction performance using the decrease in averaged weighted Gini impurity — the measure for impurity used for splitting the decision trees in the RF — of all trees in the forest. The features with highest impact are listed in Table 5.6 and correspond to meta paths along the genetic neighborhood of drug targets, but also moving out of the direct neighborhood by considering aspects such as drug similarity.

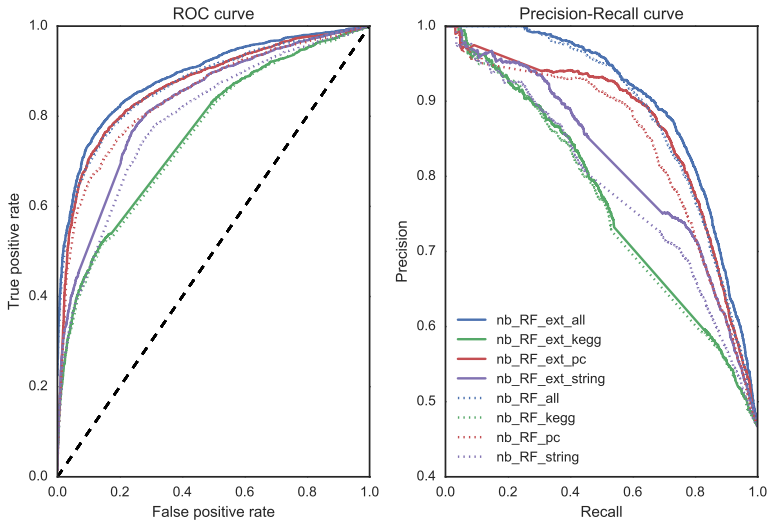


Figure 5.4: ROC curves for different feature sets. We used absolute count features and features built on meta paths that included gene-gene links, different data sources could either include data from any PPI source (blue), only KEGG (green), Pathway Commons (red) or only String (purple).

Table 5.6: Most important features in RF classifier

Feature meta path	PPI data source	Importance
D-binds-R-binds-D-targets-P-catalysis_precedes→P-associates-M	KEGG	0.134
D-binds-R-binds-D-targets-P-controls_expression_of→P-associates-M	KEGG	0.039
D-binds-R-binds-D-targets-P-in_complex_with-P-associates-M	KEGG	0.035
D-binds-R-binds-D-targets-P-interacts_with-P-associates-M	KEGG	0.026
D-binds-R-binds-D-targets-P-p3-P-associates-M	KEGG	0.024
D-resembles-D-binds-R-contains-P-catalysis_precedes→P-associates-M	KEGG	0.023
D-resembles-D-binds-R-contains-P-controls_expression_of→P-associates-M	KEGG	0.020
D-resembles-D-binds-R-contains-P-in_complex_with-P-associates-M	KEGG	0.017
D-resembles-D-binds-R-contains-P-interacts_with-P-associates-M	KEGG	0.014
D-resembles-D-binds-R-contains-P-p3-P-associates-M	KEGG	0.013
D-resembles-D-targets-P-catalysis_precedes→P-catalysis_precedes→P-associates-M	KEGG	0.012

5.4 Discussion

In this chapter, we presented a heterogeneous information network modeling the interaction between drugs, genes and diseases and evaluated two different methods to identify novel drug candidates for diseases in this network. We showed that both methods result in meaningful predictions and were able to enrich for clinical treatment candidates.

Prediction performance

We evaluated our repurposing methods using standard techniques such as cross-validation[97,127,238,369,438] and comparison of prediction results to external data[59,238,349]. We further tested the robustness of the RF classifier through 10-fold cross-validation. With AUCs between 0.8 and 0.88, the RF approach lies in the same performance range as many other published methods built on similar data types. The current gold standard methods for ML-based repurposing, PREDICT and PreDR, are reported to reach AUCs between 0.85 and 0.9[127,179,369,438] showing that our approach is on par with these methods despite excluding gene expression data of drug perturbations and disease states — information utilized in other approaches[154].

For the rule-based approach, prediction evaluation was non-trivial because the approach aimed to move away from conventional similarity-based methods. By calculating the enrichment of pairs also present in clinical trials, we could compare our approach to the GBA method proposed by Chiang and Butte[59]. Here, enrichment values of more than 4-fold for clinical trials outperform those obtained with GBA[59]. Our prediction performance is in line with previous studies, observing that approaches that infer edges in the drug-gene-disease network result in predictions with higher interpretability than regression-based predictions[369], but fail to reach comparable prediction performance.

Inclusion of heterogeneous data sources

When comparing different repurposing methods, performance usually depends on two factors: 1) the computational approach, 2) the underlying data used for model training and edge inference[127,154,179,272,369,438]. Method and data are usually closely linked, which makes independent evaluation hard.

Inclusion of more data sources seemed to generally increase prediction performance in past studies[127,154,179,369,438] even though increasing numbers of false-positive associations add noise. Different identifier systems used by the distinct data sources of the same association types further complicate their seamless integration. Particularly for disease codes the n-to-n mapping between different coding systems such as ICD and MeSH terms results in decreasing accuracy with each mapping step. Nevertheless, our evaluation supports this hypothesis as the inclusion of different PPI networks into the feature vector increased classifier performance.

A second problem arises from the automatic inclusion of heterogeneous biological data: some relations are trivial to a medical expert but are not directly linked in the ontology[154]. Especially the rule-based approach proposed in this chapter can be affected by this problem.

Important features in RF classifiers

The importance of each feature in an RF can be measured by the extent to which its exclusion affects performance[415]. Features ranked highly in our RF classifiers were those considering the genetic neighborhood of drug targets in addition to other aspects such as compound similarity. These meta paths are to some extent more complicated than those reported to be of high importance in a recent logistic-regression HIN-based method[154]. However, our ranking is likely biased through correlated features[407] in which all correlated variables receive small weights as no single one dominates prediction accuracy in a substantial subset of the RF decision trees[415].

Some of the meta paths used to construct the feature vector are related in the sense that they describe the same biological mechanism considering different data sources (e.g., KEGG, PC and STRING for PPIs). Furthermore, several relations included in the network are not independent from each other in the cell, such as the association of chemically similar drugs binding to the same protein. Thus, several clusters of correlated variables exist in our feature set (Supplementary Figure D.2), likely affecting the ranking presented in our analysis. However, given the random selection of features included the nodes in each decision tree, the confounding effects of correlated features do not affect classifier performance[38]. We thus chose an overall robust classifier with likely redundancy in the included features over missing potentially important information, but acknowledge that the listing of important features may be skewed.

Limitation of the network approach

The `myDrug HIN` is built on a multitude of different data sources, representing multiple vertex and edge types (see Methods). Due to the connectedness of the network, novel predictions for drugs or diseases do not require complete knowledge about the different association types. However, disease and drug vertices need to be connected to other vertices in the network to facilitate drug-disease edge prediction. It is thus not possible to make predictions between vertices separated by multiple components in the network.

Future development

In this chapter, we demonstrated how two approaches using meta paths in a HIN — one unsupervised and one unsupervised — can be useful to identify repurposing hypotheses. Prediction performance is comparable to that of other repurposing approaches even though

common data types such as gene expression data were not included in the network. The presented RF models appear robust against false-positive associations in the training and feature set, resulting in improved prediction performance the more data sources are included. Due to the reported benefits of gene expression data in other studies[127,154,179,272,369,438], the inclusion of gene expression data may be capable of improving predictor performance further.

By including knowledge about *in vivo* cell-type-specific gene expression landscapes, like the Human Cell Atlas[351] or GTEx[135] for filtering the HIN instance, the proposed methods could also be used for tissue-specific repurposing predictions (e.g., to adapt predictions for neuronal diseases to genes only expressed in neurons). Personalized predictions for treatment options could be made by adapting the `myDrug HIN` instance to reflect a patient's specific gene expression and genetic variation pattern.

We showed that low topology-based path weights in the rule-based approach enrich for true-positive predictions. A thorough evaluation of relative association importance in the network may thus help to prioritize predictions in both methods. Here, a logical extension of the presented approach would be to include edge and node weights in feature computation to downweigh highly connected or low confidence nodes. One possible weighing scheme worth evaluation are *degree-weighted paths*[154].

In addition to the repurposing methods presented in this chapter and through publicly available web server (`midrug.org`), the `myDrug` data repository that forms the basis for the HIN can also be used to gain general insights into disease and drug mechanism in the future. One such example will be shown in the next chapter.

Chapter 6

Personalized Pharmacogene Analysis

6.1 Introduction

A patient's molecular profile of their disease can be used for personalized therapy decisions[51,213,380]. Specifically in the oncology setting, the arduous task of extracting actionable variants from the long list of all mutations found in the affected tissue is delegated to a specialized group of geneticists, bioinformaticians, and oncologists to make treatment decisions on a patient by patient basis[213,380]. With the cost of genome sequencing dropping to a level that incentivises medical insurance providers to reimburse genome sequencing in cancer patients[201], large amounts of genomics data have become available from patients. With this influx of data, existing strategies for interpreting genomic results need to be streamlined and automated to enable medical experts to utilize the wealth of information. This means that tools need to be developed that allow physicians to receive an overview of the molecular driving forces of the disease as well as be advised of actionable therapeutic targets.

Many major medical institutions have started implementing genotyping protocols for preemptive pharmacogenetic testing, but these usually focus on the PD and PK of drugs[90]. However, as shown in Chapter 4, other drug-related genes are also relevant for a thorough understanding of the particular disease in the patient and to guide treatment decisions. In cancer, for example, driver gene mutations offer a selective growth advantage to the cells in which they occur, and the identification of such genes and their interaction networks present a paradigm for our understanding of cancer progression and therapeutic resistance[360,433]. For instance, in some instances of colorectal cancer treated with the *EGFR* antibody cetuximab *KRAS* mutations arise, requiring secondary therapy strategies to be implemented[392].

A number of publicly available databases have been created in the past years to annotate genetic variants in respect to their therapeutic actionability specifically for cancer[109,124,133].

These resources suffer from two major problems: 1) Manually querying them is not feasible for all sequenced patients and, therefore, automatic mechanisms of data retrieval from diverse resources are required. 2) Many of the genetic alterations observed in cancer have not yet been functionally characterized and it thus remains unknown to what extent they may influence the disease and treatment. For example, while certain variants including the commonly seen V600E polymorphism in the serine-threonine protein kinase *BRAF* have been thoroughly characterized in manually curated resources such as mycancergenome[469], PharmGKB[453] and CIViC[133], many less frequent variants lack data about their effect on targeted therapeutics (e.g., in mycancergenome.com only six of the 19 variants described in *BRAF* have annotated effects on targeted therapies, accessed 11/2017). Furthermore, experimental assays such as KinomeScan, commonly employed to test protein-specific ligand-affinities[72], only incorporate few clinically established variants[100], and thus currently do not provide an sufficient data source for individualized treatment decisions.

To identify variants that may be either amenable to drug treatment or confer treatment resistance in a patient-focused context, comprehensive probing of the drug target space and downstream drug effects is required. Several interactive and command line analysis tools for automated cancer genome annotation have been developed, including Drug-SNPing[464], Virtual Pharmacist[57], BALL-SNP[298], IntoGen[360], cBioPortal[49], the Broad Tumor Portal[233], and IMPACT[156], but usually solely rely on collating existing information from the literature.

Goals of the Project

The aim of this project was to create a computational infrastructure that can be used to gain drug-related insights on the effects of genetic variants. This includes mining the `myDrug HIN` presented in the previous chapter for pharmacologically relevant data that may be affected by genetic variants. To increase the resolution at which the classification of variants is performed, we developed an MM protocol to model mutations in drug targets and to quantify the extent to which the binding affinity may be altered.

We further introduce a clinical reporting pipeline, which annotates variants with pharmacogenomic and drug actionability information.

6.2 Materials and Methods

6.2.1 Structure-Based Modeling of Mutation Effects on Drug Binding

We developed a molecular modeling protocol for predicting the effects of a genetic variant on protein stability and activity similar to those commonly used in the field of computational protein design[209,297].

This protocol was implemented as a series of workflows that model the effect of one or more single nucleotide variants in a protein on the binding affinity of a ligand given the protein's 3D structure:

1. Introduction of variant into structure
2. Molecular dynamics simulation
3. Extraction of representative frames using PCA
4. Virtual screening of ligand library using molecular docking
5. Pose analysis

To predict changes in binding affinity of drug-like ligands due to variation in the protein, we extracted a representative set of frames from the MD trajectory spanning all conformations observed during the simulation using principal component analysis (PCA) followed by k-means and hierarchical clustering and prepared them for molecular docking. By docking ligands into the wild-type and variant structures, we then tried to estimate the change in binding affinity of a ligand between the native protein and its variants. We evaluated our approach based on its ability to reproduce known conformations and changes in binding affinities of 15 protein kinase variants.

Molecular Dynamics protocol

MD simulations of variant proteins were applied in protein design for testing the structural integrity of designs, mapping out molecular interactions and ensuring that the protein variant is at a potential energy minimum[209,297]. Extensive changes of the protein sequence, such as those introduced by multiple variants in the active site, can result in non-native conformations becoming thermodynamically more favorable than the native conformation. These changes would not be observed without the modeling of protein dynamics over a certain time[209,297].

After creating a model of the altered structure by replacing the variant residues in the wild-type input structure, we used MD simulation to simulate resulting structural changes: To obtain the overall conformation of the modified protein at its energy minimum we minimized the structure and then simulated the system for at least 20 ns using the OPLS2005 force field[17]. This protocol was implemented in KNIME[23] with the Schrödinger extension for structure preparation and Desmond[35,391] (version 38017) for system relaxation and MD simulation. We used the Schrödinger tool multisim (version 3.8.5.11) to coordinate the multiple stages of the Desmond simulation.

To keep the number of particles in the system small, it was solvated with TIP3P water[195] in an orthorhombic or a rhombic dodecahedron xy-hexagon box (depending on

the protein shape). Periodic boundary conditions were satisfied using an electrostatic cutoff of 9 Å and a box size buffer of 7 Å. The system was neutralized by adding Na^+ and Cl^- counter ions to charged groups. We further simulated a cytoplasmic salt concentration of 0.15 M by adding additional explicit Na^+/Cl^- ions.

Preceding the 20 ns production MD simulation, each system — protein wild type and variants — was relaxed with Desmond. Relaxation steps included two stages of minimization (first solute-restrained, then unrestrained), followed by four stages of short equilibration MD runs (72 ps) with gradually diminishing restraints. Here, the system was heated up from 10 K to 310 K in four steps (12 ps at NVT conditions with restrained non-hydrogen solute atoms restraint and 10K, then 1 step of 12 ps at NPT conditions and restrained non-hydrogen solute atoms and 10 K and 1 atm, followed by two 24 ps runs at NPT without restraints at 300 K and 1 atm). For the production simulation we simulated the system for 20 ns with fixed particle number, temperature, and pressure (NPT) in a Nose-Hoover chain thermostat/Martyna-Tobias-Klein barostat, at 310 K and 1 atm. Long-range electrostatic interactions were modeled using the particle-mesh Ewald (PME) method with a long-range cutoff of 10 Å. We recorded energies every 1.2 ps and trajectory frames of the structure in 4.8 ps intervals for in the production MD simulation.

Docking pre-processing protocol

Representative frames were extracted from the trajectories by clustering and exporting cluster representatives. If the variant was modeled in more than one PDB structure for a distinct kinase conformation, we clustered those trajectories together. First, the Desmond trajectories were converted to the CHARMM DCD trajectory format using CatDCD (version 4.0) provided by the Visual Molecular Dynamics (VMD) tool[170] to make them compatible with BALL[152] for downstream analyses. To concatenate the frames of multiple trajectories, we further reordered the atoms in the DCD files according to the order in one reference trajectory file. Highly flexible residues at the protein termini were removed using a sliding window approach determining terminal residues with a sudden change in root mean square fluctuation (RMSF)[107].

Dimensionality of the data was reduced by principal component analysis (PCA) prior to clustering. Here, we used a protocol developed by Fischer *et al.* using torsional space (i.e., using backbone dihedral angles) as input to the PCA because it was shown to result in a more diverse set of conformations extracted from the trajectory[107]. The dimensionality of the complete input matrix was reduced by projecting it into the space spanned by the first m principal components covering 95 % of the overall variance.

We extracted representative frames from the MD trajectory in a two-step process employing k-means clustering followed by complete-linkage hierarchical clustering in R.

During the first step we grouped the n frames in the trajectory into k clusters, following a common rule of thumb[266] of $k = \sqrt{\frac{n}{2}}$. To account for the randomly assigned initial cluster centroids in k-means, we repeated this process ten times and selected the run that best represented the full trajectory by comparing the convex hulls of the covered space spanned by the cluster representatives[107].

In a second step, cluster representatives were grouped using complete-linkage hierarchical clustering to obtain a reduced set of structures for docking. This step was performed using the RMSD on backbone Cartesian coordinates to account for positional similarity between conformers[107]. The binary tree obtained in this process was cut at several levels in the hierarchy to obtain 5, 10, 15, and 20 clusters, selecting the structure per cluster with the overall lowest RMSD to all other members as representatives.

Docking protocol

The extracted trajectory frames were prepared for docking by converting them to the Schrödinger-native maestro file format (.mae) and annotating them with variant- and clustering meta-data. Then the docking grid was created for each structure using the centroid of the co-crystallized ligand to define the inner grid box. The outer box limits were set to 30 Å to facilitate the docking of larger ligands in the library.

We compiled a screening library from the following three data sources:

- co-crystallized ligands of kinases in PDB
- ligands with measured binding affinities by Davies *et al.*[72]
- all FDA-approved small molecule drugs from DrugBank[232]

These molecules were desalted and tautomerized using the Schrödinger LigPrep tool (version 3.4), while retaining original ionization states and chirality specified in the input structures. The ligand library was docked into the binding pocket using the Glide XP docking protocol[113] implemented in the Schödinger Glide program[112,138] (version 59047). Here, the initial number of retained poses was set to 5000 (MAXKEEP=5000) and 1000 poses were kept for energy minimization (MAXREF=1000), following previous benchmark protocols[422]. Ligand poses were optimized post-docking (POSTDOCK=TRUE, POSTDOCK_NPOSE=10).

The docking step was followed by a rescoring step using MM energy calculation combined with the generalized Born and surface area continuum solvation (MM-GBSA) implemented in the Prime tool (version 3.0) to obtain more accurate estimates for relative binding affinities of the ligands[132].

Initial Glide XP scores as well as MM-GBSA estimates for the free binding energy (dG) were extracted from the Maestro poseviewer files and stored in a matrix with individual

ligand-variant/conformation combination per row and 20 columns (one for each cluster representative). All docking results for a conformation and variant were aggregated by calculating the median score. For those ligands where multiple isomers/tautomers where docked, only the conformer with the lowest score was included. The variants' effects on binding affinity were calculated as the difference between the predicted affinity in the wild-type simulations and those for the variant.

6.2.2 Evaluation of the Modeling Protocols

We tested and evaluated the described protocol for several protein kinases, a protein family of clinical interest in the field of oncology due to a large number of approved therapeutics against these targets (Chapter 5.3.1). Protein kinases transfer phosphate groups from adenosine triphosphate (ATP) to a residue in a protein. This step serves as an activation signal for many cellular functions including cell cycle and cell death[142]. Depending on the amino acid to which the phosphate group gets attached, one discriminates between tyrosine and serine-threonine kinases as well as several atypical kinases which — despite demonstrated kinase activity — do not share sequence similarity to the two other kinase families.

The modeled proteins are summarized in Table 6.1 and contain members of both large kinase subfamilies as well as structures representing the active and inactive conformation of the protein depending on the "Asp-Phe-Gly" (DFG) motif. This motif is located at the N-terminus of the activation loop and facilitates catalysis if the side chain of the aspartate residue faces into the active site while the phenylalanine occupies a hydrophobic pocket adjacent to the ATP-binding site ("DFG-in")[142].

Evaluation of MD protocol

We evaluated the MD protocol by comparing conformations obtained from the simulation to known 3D structures of kinases harboring the variant. Since certain variants are known to cause a shift from one kinase conformation to the other, we expected that if the simulation sufficiently covered conformational space, we would observe this shift also in the frames extracted from the trajectory. However, upon inital inspection of the results, it appeared that a simulation length of 20 ns may be too short to cover the full conformational space of this protein family. We thus extended the simulation duration for a single kinase, *BRAF*, to 100 ns and also evaluated the conformational space coverage in this longer simulation through PCA.

To judge the quality of individual simulations and resulting models, we devised a workflow that computes the standard set of MD quality control (QC) measures for each

Table 6.1: Protein kinases modeled using the proposed pipeline.

Protein	Structure	Catalaytic Activity	Conformation	Kinase Family	Uniprot Accession
EGFR	2GS2, 2GS6	active	DFG-in	tyrosine	P00533
	3GT8, 3W23	inactive	DFG-out	tyrosine	P00533
BRAF	3TV6, 3C4C	active	DFG-in/αC-in	serine-threonine	P15056
	4E26, 3PRF	intermediate	DFG-in/αC-out	serine-threonine	P15056
	3IDP, 4KSP	inactive	DFG-out/αC-out	serine-threonine	P15056
KIT	1PKG	active	DFG-in	tyrosine	P10721
	3G0E, 1T45	inactive	DFG-out	tyrosine	P10721
FLT3	1RJB	inactive	DFG-out	tyrosine	P36888
ROS1	3ZBF	inactive	DFG-out	tyrosine	P08922
ERBB4	2R4B, 3BCE	active	DFG-in	tyrosine	Q15303
	3BBW, 3BBT	inactive	DFG-out	tyrosine	Q15303
KDR	3VHE, 3VHK	inactive	DFG-out	tyrosine	P35968
MTOR	4JSN, 4JSP	inactive	DFG-out	atypical	P42345
TTN	1TKI, 4JNW	inactive	DFG-out	atypical	Q8WZ42
EPHA3	3FY2, 4GK3	active	DFG-in	tyrosine	P29320
	2QOQ,2QO2	inactive	DFG-out	tyrosine	P29320
EPHA5	2R2P	active	DFG-in	tyrosine	P54756
PI3K	2RD0	active	DFG-in	atypical	P42336

simulation (energy analysis, RMSD, RMSF, radius of gyration, number of hydrogen-bonds, solvent accessible surface area (SASA) and Ramachandran plots).

Evaluation of docking protocol

We evaluated the docking protocol by re-docking co-crystallized ligands to the cognate structures to test if it allowed reproducing the native binding conformation. The correctness of the docking pose was assessed by calculating the RMSD to the crystallized conformation. For data storage and retrieval we devised a mySQL database that extents the `myDrug` database introduced in the previous chapter with tables storing meta-data and modeling results.

To quantify the extent to which predicted changes in binding affinity caused by a mutation agree with experimental evidence, we calculated Spearman's rank correlation between predicted and measured differences in binding affinities (as estimated by the Glide XP score and *in vitro* measurements of the dissociation constant K_d, respectively) for a subset of 72 ligands and 15 variants in four kinases (*EGFR, BRAF, KIT, FLT3*) from Davis *et al.*[72].

To evaluate the changes in prediction performance after re-scoring by MM-GBSA, we also calculated ρ between predicted ΔG differences of wild type and mutant and experimental measurements. Due to the high temporal demand and license restrictions of this process, we focused on five *EGFR* mutations for the this step. Experimental measurements of $K_d > 10$ nM were set to null in the published data set and excluded from our analyses[72].

The statistical tests were performed using the SciPy stack[193] and Jupyter notebooks[334].

6.2.3 Annotation of Patient Genomes by Known Drug Effects

Overall server architecture

Due to the high computational demand of the complete protocol described above, the prediction of variant effects based on their structural impact cannot be performed for all possible mutations in all human genes at the moment. However, particularly in the clinical setting, physicians and consulting experts may be interested in an aggregated collection of mutation and gene-specific data. Such data are collacted in the `myDrug` database. To make it accessible to the outside world, we developed a Flask web server, which supports different types of queries and implements a RESTful interface to the database (code available at `https://github.com/kohlbacherlab/mydrug-server`) that allow the search of the genetic neighborhood for additional treatment candidates.

- **Search by gene**: query by gene symbol and retrieve list of neighboring genes, their associated drugs and diseases
- **Search by compound**: query by drug name and retrieve list of diseases sharing the drug target gene or lying in the genetic neighborhood of the drug target
- **Search by disease**: query by disease name, Disease Ontology ID or OMIM code and retrieve list of drugs modulating one of the known disease target genes or affecting neighboring genes
- **Search by mutations**: Annotate a set of variants, represented as a VCF file or list of HGVS variant names, for their involvement in drug action and diseases. For individual genes, it also allows the search for drugs targeting gene neighbors to the mutated gene in the `myDrug` network. We further implemented the annotation of specific observed variants with pharmacogenomic data from PharmGKB.

The overall structure of the service follows the model-view-controller design pattern. The web application is modularized by individual Flask blueprints for the different use cases, as depicted in Figure 6.3. The data model representing the data stored in the `myDrug` MySQL database is shared between the individual blueprint modules using SQLAlchemy [19], while forms (`forms.py`) and controllers (`views.py`) are placed inside the individual blueprint module directories. Jinja templates return data from the controller to build the final view and are defined in a separate static templates directory, whose structure follows the blueprint modularization pattern.

The RESTful API is implemented using the Flask Restplus module [147] which automatically builds a documentation view using Swagger [410] about API usage options. Data is returned as JSON object which allows the direct conversion in python dictionary objects for subsequent analyses by the end user.

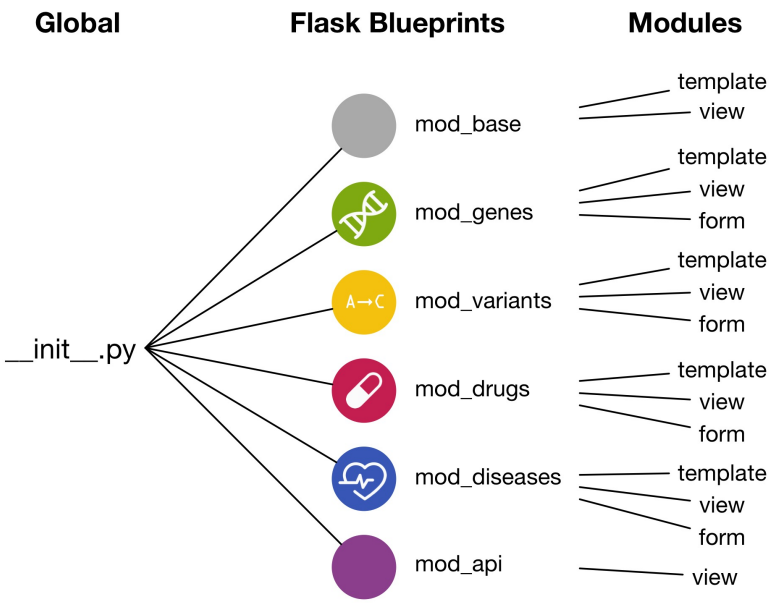


Figure 6.1: Internal structure of the myDrug webservice. Each blueprint serves the logic and view of a particular use case, including querying by drug, disease and gene.

Evaluation of approach for 198 patients with hepatocellular carcinoma (HCC)

The mutation view in the myDrug app can be used to create patient reports from VCF files of genetic variants found in a patient. One particular use case for this pipeline is the annotation of tumor-specific variants to explore targeted treatment options.

To evaluate the usefulness of our reporting approach, we tested the protocol on 197 freely available somatic exomes of liver cancer patients in the The Cancer Genome Atlas (TCGA) Liver Hepatocellular Carcinoma (LIHC) study. We obtained somatic mutations using the R/Bioconductor package TCGAbiolinks[62] for the TCGA-LIHC project as a single list with all variants observed in all samples. This list was split and converted into individual VCF files using Python. We used the `myDrug` RESTful API to annotate the individual patient exomes, running Ensembl VEP[276] (release 84) to annotate the variants based on the reference genome prior to `myDrug` annotation. Subsequent analyses were performed in Python using Jupyter notebooks[334], pandas[274], pyVCF[47] and the SciPy stack[193]. Variants predicted to result in the loss of protein function by the LOFTEE VEP plugin[259] or unanimously predicted to have a deleterious effect on the protein product by SIFT[303] and PolyPhen-2[4] were defined as non-functional.

6.3 Results

6.3.1 Prediction of Genetic Variant Effects on Drug Binding Affinity

We developed a protocol to predict the effects of genetic variants in drug targets consisting of two major steps: 1) modeling the effect of a genetic variant on the protein structure using MD, and 2) employing molecular docking to predict the change in binding affinity of the ligand in the altered structure.

We evaluated the pipeline's applicability in a set of clinically relevant protein kinases — a protein family often targeted by precision cancer therapeutics.

Modeled kinases

We evaluated the protocol by modeling mutations in a set of eleven cancer-associated kinases and comparing docking-based prediction performance to experimentally determined binding affinities in the four most commonly affected kinases. We successfully modeled commonly observed mutations in ten different protein kinases (Table 6.1). Simulations in PI3K failed due to problems with the modeling software and were discarded. We obtained trajectories for two different protein structures per kinase and conformation, where possible, to optimally cover the known structural space. For these structures, we simulated the system for 20 ns and extracted snapshots from the trajectory for ligand-docking.

Evaluation of the MD protocol

To test the ability of the protocol to produce stable conformations of the mutated proteins, we compared the modeled mutant structures to known 3D structures containing the same mutation in the PDB[22] (Table 6.2). In *BRAF* the extracted frames of the system with introduced mutation show lower overall RMSDs to the crystallized mutant than frames extracted from the wild-type trajectory. Overall, structural changes in the backbone are small for all tested examples.

Furthermore, when comparing the binding site of a modeled activating variant in an inactive 3D structure to the crystallized active conformation, the RMSDs between modeled structures with activating variants and the crystallized active proteins are consistently smaller than those between active and inactive 3D structures (Table 6.3).

Based on the comparison between known mutant structures and the modeled variants, it appears as if 20 ns are not sufficient to fully cover conformational changes introduced by certain mutations. We therefore compared the trajectories of 20 ns to 100 ns simulations in *BRAF* using the first and second principal component of dihedral space to compare the covered conformational space (Supplementary Figure D.3). In addition to the inactive state structure (DFG-out, αC-out, PDB: 3IDP), we modeled the intermediate conformation

Table 6.2: Comparison between modeled mutated binding site to crystallized mutant protein.

Protein	Mutant Reference Structure	Modeled Structure	Binding Site RMSD	Backbone RMSD
EGFR	2ITN	2GS6	0.95	1.87
	2ITN	2GS6 (G719S)	0.90	1.88
	2ITN	3W32	1.48	3.19
	2ITN	3W32 (G719S)	1.66	3.22
EGFR	4I24	2GS6	1.50	2.86
	4I24	2GS6 (T790M)	2.13	2.84
	4I24	3W32	0.82	1.97
	4I24	3W32 (T790M)	0.98	1.97
EGFR	4LQM	2GS6	0.85	2.11
	4LQM	2GS6 (L858R)	na	2.04
	4LQM	3W32	1.46	2.90
	4LQM	3W32 (L858R)	na	2.89
BRAF	4MNF	3TV6	1.58	2.92
	4MNF	3TV6 (V600E)	1.23	2.55
	4MNF	3IDP	1.42	2.13
	4MNF	3IDP (V600E)	1.38	1.95
KIT	3G0F	1PKG	2.49	3.98
	3G0F	1PKG (D816H)	2.56	3.95
	3G0F	3G0E	2.52	4.24
	3G0F	3G0E (D816H)	2.81	4.29

Table 6.3: RMSD of binding site and all protein backbone residues between modeled activation mutations and active structures

Protein	Reference Structure	Modeled Structure	Activating Mutation	Representative Frame	Binding Site RMSD (Å)
	1PKG (active)	3G0E (inactive)	-		1.84
KIT	1PKG	3G0E	D816H	3001	1.32
	1PKG	3G0E	D816H	57	1.82
	1PKG	3G0E	D816H	685	1.46
	2GS6 (active)	3W32 (inactive)	-		2.62
	2GS6	3W32	L858R	1448	2.23
	2GS6	3W32	L858R	303	2.07
	2GS6	3W32	L858R	1678	1.84
	2GS6	3W32	L858R	1920	2.18
	2GS6	3W32	T790M	685	1.88
EGFR	2GS6	3W32	T790M	1610	1.62
	2GS6	3W32	T790M	2818	1.92
	2GS6	3W32	T790M	3541	1.66
	2GS6	3W32	G719S	439	1.5
	2GS6	3W32	G719S	1369	1.42
	2GS6	3W32	G719S	2447	1.78
	2GS6	3W32	G719S	4004	1.54

(DFG-in, αC-out, PDB: 4E26, 3PRF) and the active conformation (DFG-in, αC-in, PDB:

3TV6, 3C4C). Here, the two trajectories of the intermediate state overlapped in covered conformational space and the 4E26 trajectory further overlaps with the trajectory of the active structure 3TV6.

Evaluation of docking results

We evaluated the ability of the docking protocol to reproduce the known binding pose of a ligand by re-docking the native ligand back into its binding pocket. Overall, RMSDs between co-crystallized and docked ligands are small with an average RMSD of 1.68 Å and below 2 Å for seven out of the eight tested structures, indicating a successful reproduction of the binding pose (Table 6.4). For *FLT3* only an apo-structure existed and we could thus not evaluate the pose reproduction accuracy for this kinase.

Table 6.4: Redocking co-crystallized ligands into native binding pocket.

Protein	Structure	RMSD in Å
EGFR	3W32	1.38
BRAF	3TV6	0.73
	3C4C	0.59
	4E26	6.69
	3PRF	1.41
	3IDP	1.09
	4KSP	1.73
KIT	3G0E	0.71

Since the ligand binding conformation appears to be well reproduced by the proposed docking protocol, we aimed to quantify the extent to which predicted binding affinity changes between wildtype and variants correspond to experimentally determined changes. We calculated the Spearman rank correlation ρ between Glide XP docking scores and experimentally measured changes in the dissociation constant between wild-type and mutant protein[72] (Table 6.5). Given previous reports[14,324,442], we did not expect to see good correlations between Glide scores and experimental binding affinity data. This expectation was met, with overall 15 out of 20 comparisons yielding Spearman correlation coefficients close to random ($-0.3 < \rho < 0.3$).

Based on reports that this problem can be alleviated using a rescoring step[14,132,327], we tried to improve the quantitative correlation between predicted and experimental binding affinities using MM-GBSA. This approach approximates binding free energies (ΔG) and thus predicted changes in ΔG between wild-type and mutant should correlate positively with experimentally determined differences in the dissociation constant K_d. While the correlation between predicted $\Delta\Delta G$ and experimental ΔK_d indeed increased after MM-GBSA rescoring, these results remain statistically non-significant (Table 6.6).

Table 6.5: Correlation of the changes in the docking scores to experimentally determined changes in binding affinities from by Davis *et al.*[72] between wild type and variants

Protein	Mutation	Ligands with data	active		inactive	
			Spearman ρ	p-Value	Spearman ρ	p-Value
EGFR	T790M	27	-0.217	0.278	-0.335	0.087
	L861Q	29	-0.165	0.394	0.000	0.999
	G719C	27	-0.040	0.844	-0.046	0.821
	G719S	27	-0.047	0.817	**0.323**	0.100
	L858R	29	0.118	0.541	-0.224	0.244
KIT	D816V	39	**0.361**	0.024	0.242	0.138
	L576P	41	0.095	0.553	0.049	0.761
	A829P	38	**0.340**	0.037	0.122	0.465
	D816H	36	0.126	0.463	0.140	0.416
BRAF	V600E	19	**0.374**	0.115	-0.200	0.412

BRAF intermediate: Spearman ρ=0.104, p=0.673.

Table 6.6: Correlation of MM-GBSA-predicted to experimentally determined changes[72] in binding affinity between wild type and variants

Protein	Mutation	Ligands with data	active		inactive	
			Spearman ρ	p-Value	Spearman ρ	p-Value
EGFR	T790M	27	-0.342	0.080	-0.495	0.009
	L861Q	29	**0.254**	0.184	0.060	0.759
	G719C	27	**0.292**	0.139	0.070	0.728
	G719S	27	0.031	0.876	**0.423**	0.028
	L858R	29	0.024	0.901	-0.098	0.613

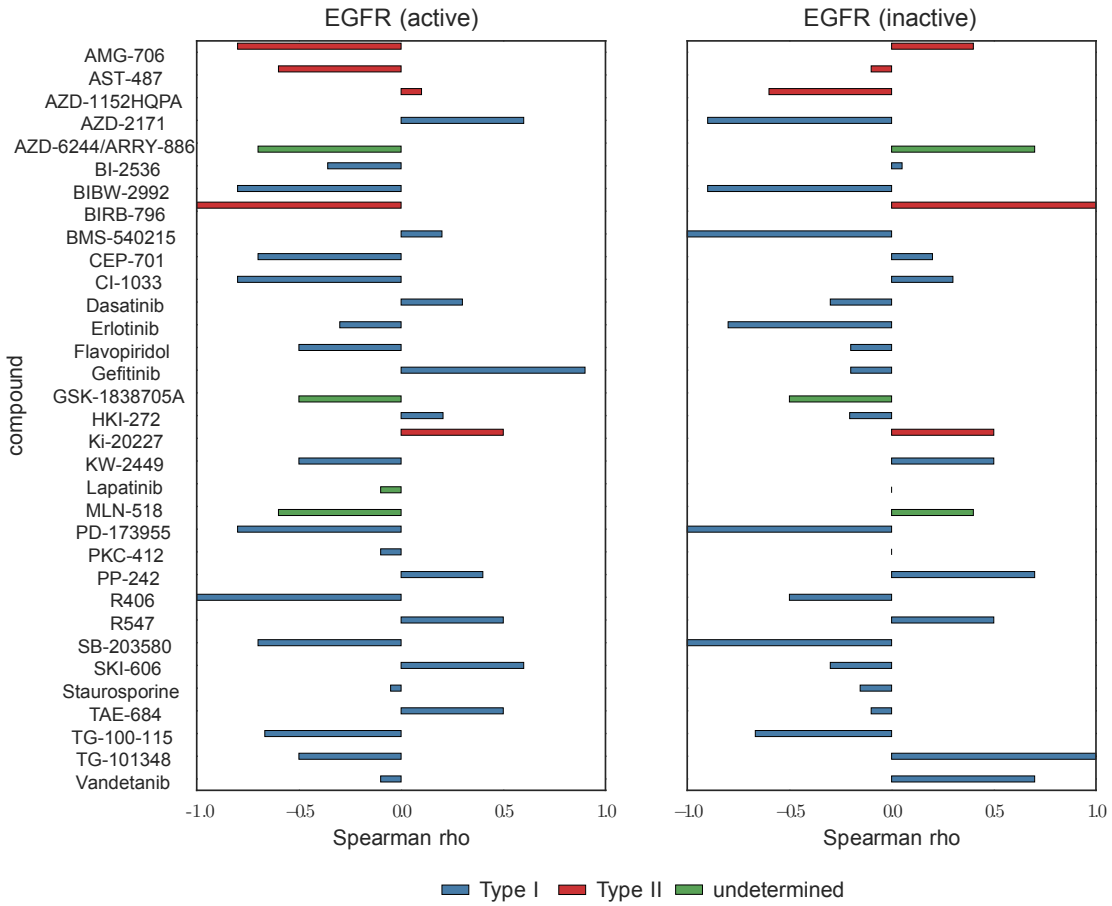


Figure 6.2: Spearman rank correlation between experimental[72] and MM-GBSA predicted binding affinities in active and inactive *EGFR* conformations. With the success of kinase inhibitors in cancer therapy, multiple types of inhibitor classes have emerged that mainly act as ATP mimetics inhibiting the active (type I, blue) or inactive (type II, red) binding site[299]. Inhibitors with unspecified type can be of either type and are colored green.

The correlation coefficients presented above could be influenced by individual outliers due to multiple types of kinase inhibitors in the screening library. We thus also calculated the correlation individually for all five mutations in *EGFR* for the inactive and active conformation: Correlation coefficients for the presented data set varied by inhibitor, but no clear trend was discernible for individual kinase inhibitor types (Figure 6.2).

6.3.2 Identification of Actionable Variants in Patients

Reporting outline

Automatic patient-specific reports could help support medical professionals in their therapy decision process (Figure 6.3). Cancer, for example, is a multi-faceted disease for which it is

now increasingly accepted that classification and treatment decisions should not be made just based on the site of disease, but also its genetic profile[49]. Here, targeted therapies developed and approved for certain genetic variants have been shown to be also effective in other cancer types if the patients tumor also carried that variant. Vemurafenib, a *BRAF* V600E inhibitor, for example, also showed clinical activity in other cancers beyond its primary indication, melanoma[174]. To this end, we supply researchers and physicians with a structured presentation, summarizing genes affected by non-functional variants, including direct drug targets and indirect drug targets, utilizing the evidence found from cross-referencing several public repositories.

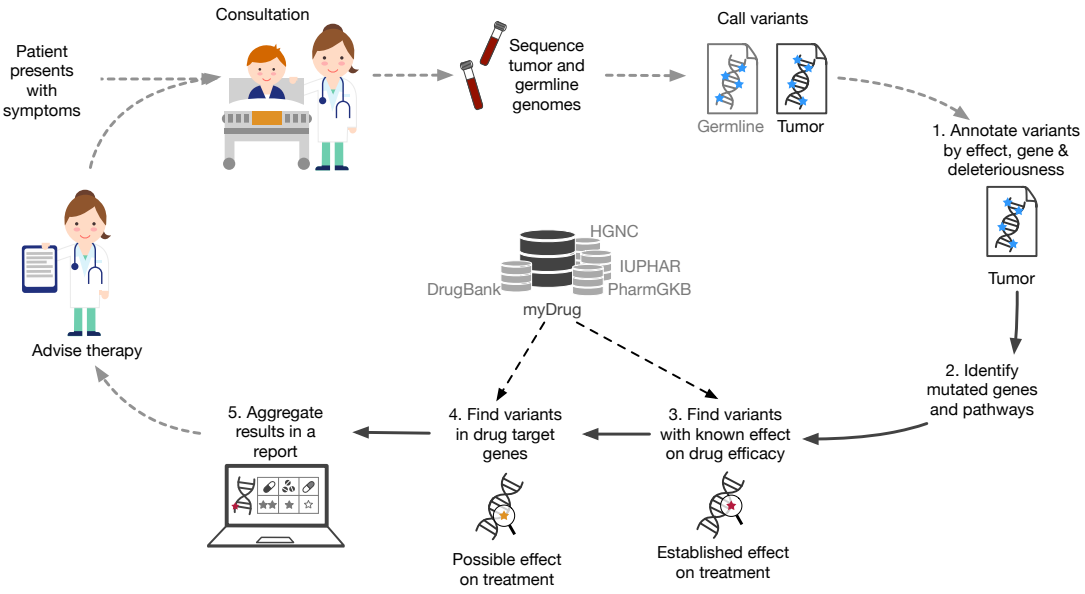


Figure 6.3: Patient-specific reporting of genetic disease profile and actionable variants. In the case of cancer, the patient's tumor and germline genome are sequenced upon diagnosis to identify genetic variants that may affect treatment decisions or treatment success. The variants are automatically annotated and compiled as a web report that can be used by physicians or expert consultants to advise next therapy steps.

To facilitate reporting on clinically interesting variants found in a patient, we developed a webserver that returns information about drug targets, drugs and their primary indication based on either a list of genes or variants (see Methods). Using this server it is possible to create a report of variants found in a patient and further download a structured view of this data (Figure 6.4).

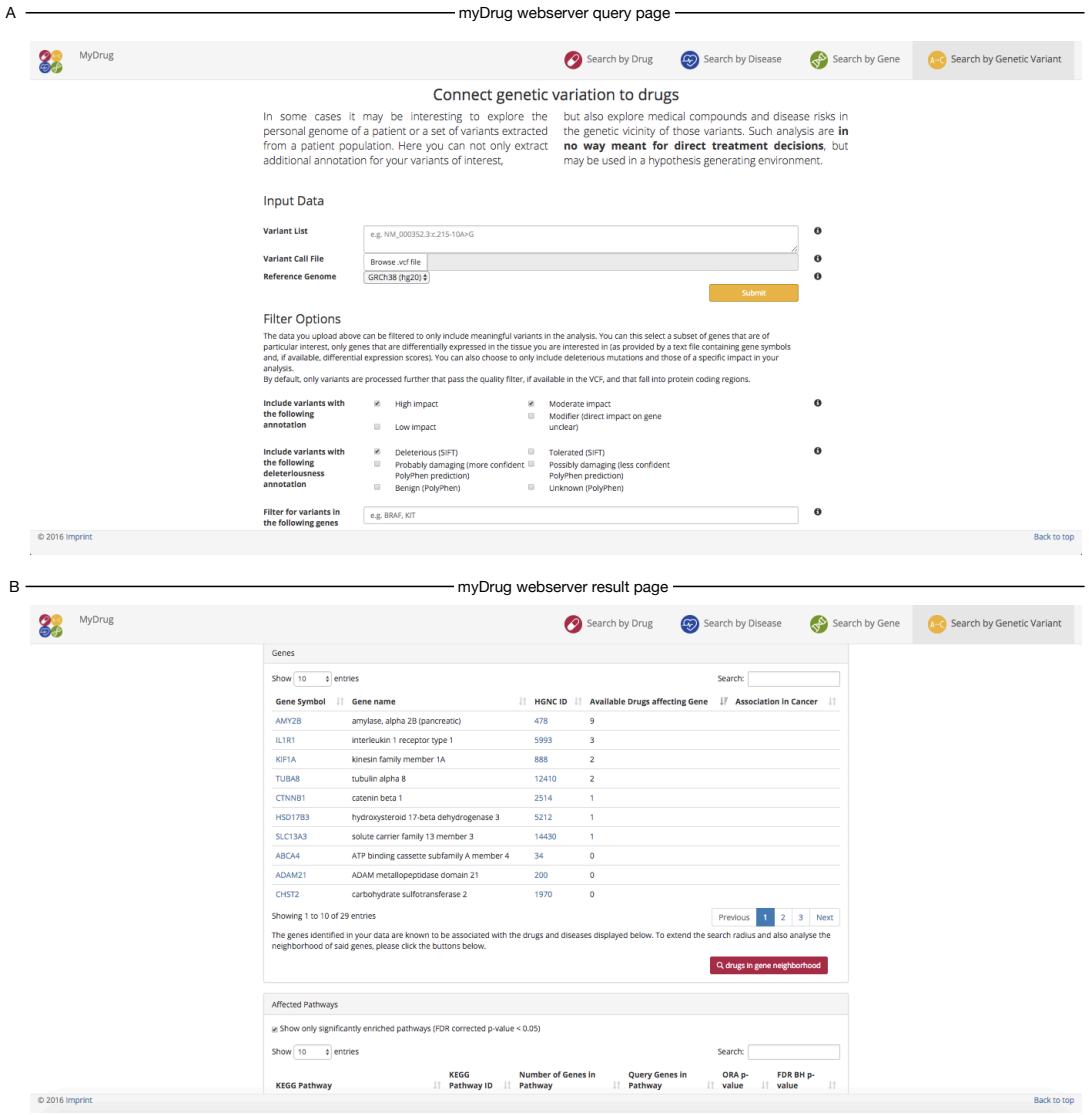


Figure 6.4: Clinical reporting using the `midrug.org` web server. A user can provide variants as a list or a vcf. The variants are then annotated with their genetic impact, affected genes and related drug information.

Case study: Actionable variants in patients with HCC

The added benefit of a personalized reporting of actionable variants in cancer patients can be illustrated on a case study: evaluating the reporting platform on 197 patients with HCC. Using the annotation pipeline, we identified actionable variants for each patient. Of all patients, 194 patients carried at least one non-functional variant in an oncogene or tumor-suppressor gene. Furthermore, 35 carried some variant in one of the genes targeted by the standard of care therapy, sorafenib. For 15 of these patients, the variants were predicted to be non-functional, possibly resulting in the loss of said target in the tumor. However, only two of the patients carried a variant with documented pharmacogenomic effects in a sorafenib target in the external cancer database CiVIC[133].

In addition to the patients with variants in any of the sorafenib targets, an additional 58 patients had non-functional variants in the canonical transcript of another gene targeted by an approved or experimental antineoplastic agent and 22 patients presented with a variant that was associated with altered response to cancer treatment in CiVIC. In 13 of these cases, the variant in the target has documented effects.

6.4 Discussion

In this chapter, we developed a protocol to model protein mutation effects on ligand binding affinity and tested it on clinically relevant protein kinases. We further developed a web server to assess genetic variation in respect to its potential clinical effect and explore alternative treatment options in a patient.

Limitations of Molecular Modeling Protocol

The protocol was capable of reproducing co-crystallized binding poses and, if MD simulations are carried out for 100 ns, also modeled the transition between different protein conformations, but several issues remain: even though 20 ns was shown to distinguish active and inactive designs of cathepsin K variants[208], it appears too short to reach equilibrium in the studied protein kinases. With increasing computational capacity available through GPU clusters and cloud computing, longer MD simulation of mutated protein-ligand complexes for a duration required to achieve equilibrium state[40] will increasingly become feasible.

We expected low correlation between Glide docking scores and experimentally determined binding energy because standard scoring functions have only limited estimates in the gain and loss functions upon binding, omit thermodynamics and build on other simplifications in the empirical function[132,377]. Nevertheless, we anticipated improvements upon rescoring using MM-GBSA[131,327] which were not met. In part this may be explained by positive and negative correlations cancelling out due to inhibitors targeting the active

or inactive conformation. In this case, type I inhibitors would correlate with the active conformation while type II inhibitors correlate with the inactive conformation after MM-GBSA rescoring. Unfortunately, such an effect could not be observed in the data for the five *EGFR* variants (Figure 6.2).

The failure to reproduce experimentally determined changes in ligand binding affinity due to genetic variants in the receptor in a consistent manner may be caused by multiple reasons: 1) structure snapshots obtained from the MD trajectory do not properly represent the protein after mutation, 2) docking does not capture the correct ligand pose, 3) binding affinities estimated by MM-GBSA are not accurate. Additionally, due to the mathematical relationship between binding affinities and binding constants, methods estimating binding affinity need to be extremely accurate to not overestimate changes in binding constants[363]. Given that MM methods lack terms describing charge transfer, many-body effects, and polarization among others, they often do not offer that level of accuracy[363]. It remains to be seen if quantum-mechanical approaches for rescoring could alleviate theses issues[363].

Challenges with Protein Kinases as Evaluation Data Set

Kinase pharmacology further complicates structure-based prediction of variant effects. For instance, the ATP-binding site which serves as primary binding site for type I and II inhibitors is well conserved and common to many kinases, but the secondary binding site and inactive state binding pockets can vary by kinase[299]. It is thus difficult to develop highly specific compounds for the primary binding site and secondary pockets used for the development of selective inhibitors[299]. Nevertheless, many kinases do not properly adopt the DFG-out conformation[143] hampering the computational exploration of that conformational state. Such structural characteristics may also affect docking accuracy when not allowing for additional flexibility during the placement steps.

Secondary variants in the so-called gatekeeper residue (T529 in *BRAF*) can evolve during treatment[467]. Such mutations confer drug resistance by sterically hindering the placement of the inhibitor in the binding pocket or by increasing affinity for the primary ligand ATP (e.g., the T790M mutation in *EGFR*[472])[421]. To observe such effects one would need to model allosteric effects and also include ATP in the screening libary. Given the protocol's overall low performance, we decided against implementing these two extensions in the current study.

Overall, the proposed variant effect prediction method succeeded in reproducing the correct binding pose of kinase ligands, but struggled with accurately predicting the relative effects of mutations on binding affinity. This problem was in part aggravated by the specific properties of kinases as target family. Given this observation, our approach should

be further tested in additional target families and with target-specific (QM-based or fitted) rescoring methods before abandoning it in lieu of more accurate scoring functions.

Limitations of the myDrug reporting service

Despite modeling effects of new mutations, a patient can benefit from reporting existing mutations found in their disease as well as matching potential drugs to the affected genes[425]. We evaluated how many patients could benefit from automated reporting of genetic characteristics in a data set of 197 HCC patients. While only one in ten patients were directly affected with a variant altering drug-binding in the tumor, another two in ten patients had non-functional variants of unknown pharmacogenomic effect in the primary target genes of the standard of care therapy, sorafenib. Furthermore, an additional third of the patients carried variants in genes modulated by antineoplastic agents aside from the standard of care therapy. Thus for more than half of all patients clues for treatment decisions could be found in their tumor genome. This insight is in line with real-world experiences with genotype-directed therapy studies that reported between 40%[425] and up to 95%[20,151,380] of the patients carrying actionable variants. However, these real-world use cases have also shown that, in addition to the identifying relevant variants for (off-)label or experimental treatment options, specific requirements need to be met for a cancer patient to benefit from genotype-based reporting: testing needs to be performed early on in the disease to be of clinical utility[151,425] and/or to match patients to suitable clinical trials[164,185,424].

Given that most variants in drug targets are not yet pharmacologically characterized (see Chapter 4), no readily available approach can be used directly in a clinical context, but interactive tools[164,372] such as the one presented here could be of value to inform expert consortia and molecular tumor boards, by facilitating informed discussion on a patient's treatment plan[155,164,424].

Conclusions

There exists a gap between known annotation of variant effects and those variants observed in patients. The pipeline introduced in this chapter aimed at closing this gap, but similar to other protocols[132], the accurate prediction of binding affinities remained beyond the capabilities of the docking protocol[376]. MM-GBSA rescoring sometimes overcomes shortcomings of conventional scoring functions[131,327], but did not suffice here. It thus remains to be seen if different methods, such as a protein-specific model fitted to experimental binding affinities or the inclusion of quantum mechanics could result in an improvement. The evaluation of such approaches in a recent review remained inconclusive[363].

Our case study of genetic variants found in 197 HCC patients shows that it can be helpful to report actionable variants. But the unknown pharmacological impact of most variants

in drug-related genes observed in the patients underscores the importance of modeling variant effects after overcoming the shortcomings of current scoring and rescoring methods. Until then, an automated annotation workflow collating knowledge from databases such as `myDrug` and ongoing clinical trials, can help guide physician's therapy decisions based on a patient's genotype. Here, the advantage of our approach is that it is not specifically geared towards cancer and can be used in other clinical settings.

Chapter 7

Conclusion and Outlook

In this thesis, we presented contributions to several challenges in the personalized medicine value chain, ranging from early research on drug targets to point-of-care clinical reporting. By utilizing available gene sequences across all kingdoms of life we showed that protein-protein interactions can be modeled and insights into their binding interfaces gained from sequence information alone — these insights can support the drug discovery process in its target identification stage. Connected with additional data about drug and disease mechanism, the networks formed by all PPIs in a cell can further guide the systematic repositioning of drug molecules to additional disease areas, either during the research phase or – hypothetically – also at the point of care if drug resistance to the original treatment is to be expected.

We further demonstrated that the vast majority of genetic variants observed in drug-related genes is still uncharacterized and proposed a method to alleviate this problem using molecular modeling protocols. However, while successful in reproducing known ligand binding poses in mutated proteins, the scoring function could not correctly rank the changes in relative binding affinity of a variant compared to the wild type. Here, future development will be needed to improve the approach. Nevertheless, the automatic reporting of existing knowledge about genetic variants observed in a patient can already be beneficial in the clinical practice.

With more drugs being approved specifically for genetic variants, such as pembrolizumab for all mismatch repair deficient tumors regardless of site of disease, physicians will start to rely on the support of algorithms in their decision processes. The methods presented in this thesis are one step towards bringing such methods to the clinic — also for diseases that are not cancer. The true power of the presented approaches lies, however, in connecting them with other progress in the field to further improve prediction performance. We could already show that ECs can be used to predict protein fitness of gene variants[161] and the effect of variants on protein conformation[383]. Predicting the consequences of genetic variants

on drug binding affinity directly from sequencing data could be valuable as an orthogonal approach to the presented molecular modeling protocol and help overcome some of the weaknesses observed in traditional scoring methods. By filtering the `myDrug` network to variants and gene expression profiles observed in specific patients, we could also tailor drug repurposing predictions to individual patients.

Most of the obstacles in solving each of the presented challenges stem from an incomplete understanding of physiological processes that could possibly be alleviated by the availability of more data. It is, for example, estimated that a complete map of human interactome is at least a decade away[279] and network-based prediction performance will likely increase with a more complete interaction map. Furthermore, a limited number of orthogonal data sources describing different phenotypic and genotypic features can currently be integrated into cellular network models[471], limiting our understanding of the inter-dependencies between genome and phenome.

Combining expert knowledge with external information can already help physicians today find the suitable therapy for their patients[202] or assign them to clinical trials matching their disease (a concept commercialized by companies like FoundationMedicine). By providing earlier and more detailed diagnoses, avoiding obvious pharmacological risks and providing treatment suggestions, personalized medicine algorithms can improve health care for the patient[434] even with the remaining gaps in our understanding of physiological processes[202]. Nevertheless, these gaps have to be accounted for and be openly addressed in the practice. Pre-emptive pharmacogenetic testing with immediate consequences to therapy decisions should, for example, be restricted to validated variants[231], while comprehensive strategies could be used to generate research leads in more complex cases.

Apart from gaps in our understanding of human physiology, imminent hurdles of the healthcare system need to be overcome for the full adoption of personalized medicine in the clinic: this includes financial aspects such as the reimbursement of genetic testing by insurance companies or other payers[434], but even more the availability of suitable and affordable treatment options addressing patient-specific needs. Especially when solving the latter challenge, genetic tests together with personalized treatments will have a tangible positive impact on patient outcome[345].

Bibliography

[1] J. Abbasi. Getting Pharmacogenomics Into the Clinic. *JAMA*, 316(15):1533–1535, 2016. 82

[2] N. S. Abul-Husn, K. Manickam, L. K. Jones, E. A. Wright, D. N. Hartzel, C. Gonzaga-Jauregui, C. O'Dushlaine, J. B. Leader, H. Lester Kirchner, D. M. Lindbuchler, M. L. Barr, M. A. Giovanni, M. D. Ritchie, J. D. Overton, J. G. Reid, R. P. R. Metpally, A. H. Wardeh, I. B. Borecki, G. D. Yancopoulos, A. Baras, A. R. Shuldiner, O. Gottesman, D. H. Ledbetter, D. J. Carey, F. E. Dewey, and M. F. Murray. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science*, 354(6319):aaf7000, Dec. 2016. 13

[3] L. Adnane, P. A. Trail, I. Taylor, and S. M. Wilhelm. Sorafenib (BAY 43-9006, Nexavar (R)), a dual-action inhibitor that targets RAF/MEK/ERK pathway in tumor cells and tyrosine kinases VEGFR/PDGFR in tumor vasculature. *Regulators and Effectors of Small Gtpases: Ras Family*, 407:597–+, 2006. 70

[4] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010. 68, 115

[5] P. Aloy and R. B. Russell. Ten thousand interactions for the molecular biologist. *Nature Biotechnology*, 22(10):1317–1321, Oct. 2004. 19

[6] D. Altschuh, A. M. Lesk, A. C. Bloomer, and A. Klug. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, (193):693–707, 1987. 26

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct. 1990. 38

[8] S. F. Altschul, T. L. Madden, and A. A. Schäffer. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997. 38

[9] J. Andreani and R. Guerois. Evolution of protein interactions: from interactomes to interfaces. *Archives of biochemistry and biophysics*, 554:65–75, July 2014. 36

[10] J. Andreani, G. Faure, and R. Guerois. InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics (Oxford, England)*, 29(14):btt260–1749, May 2013. 36

[11] R. Arnold, F. Goldenberg, H.-W. Mewes, and T. Rattei. SIMAP—the database of all-against-all protein sequence similarities and annotations with new interfaces and increased coverage. *Nucleic Acids Research*, 42(D1):D279–D284, Jan. 2014. 87

[12] X.-c. Bai, G. McMullan, and S. H. W. Scheres. How cryo-EM is revolutionizing structural biology. *Trends in biochemical sciences*, 40(1):49–57, Jan. 2015. 35

[13] D. Baker and A. Sali. Protein Structure Prediction and Structural Genomics. *Science*, 294 (5540):93–96, Oct. 2001. 25

[14] B. Balaji and M. Ramanathan. Prediction of estrogen receptor β ligands potency and selectivity by docking and MM-GBSA scoring methods using three different scaffolds. *Journal of enzyme inhibition and medicinal chemistry*, 27(6):832–844, Dec. 2012. 118

[15] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, and C. J. Langmead. Learning generative models for protein fold families. *Proteins-Structure Function and Bioinformatics*, 79(4):1061–1078, Apr. 2011. 29, 42, 49

[16] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani. Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. *PLoS ONE*, 9(3):e92721, Mar. 2014. 28

[17] J. L. Banks, H. S. Beard, Y. Cao, A. E. Cho, W. Damm, R. Farid, A. K. Felts, T. A. Halgren, D. T. Mainz, J. R. Maple, R. Murphy, D. M. Philipp, M. P. Repasky, L. Y. Zhang, B. J. Berne, R. A. Friesner, E. Gallicchio, and R. M. Levy. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *Journal of Computational Chemistry*, 26(16):1752–1780, Dec. 2005. 14, 109

[18] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, Jan. 2011. 19

[19] M. Bayer. Sqlalchemy. `https://www.sqlalchemy.org`, 2014. Version: 1.1.10. 88, 114

[20] H. Beltran, K. Eng, J. M. Mosquera, A. Sigaras, A. Romanel, H. Rennert, M. Kossai, C. Pauli, B. Faltas, J. Fontugne, K. Park, J. Banfelder, D. Prandi, N. Madhukar, et al. Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA oncology*, 1 (4):466–474, July 2015. 125

[21] T. Berggård, S. Linse, and P. James. Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7(16):2833–2842, Aug. 2007. 17

[22] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, Jan. 2000. 49, 87, 116

[23] M. R. Bertold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel. KNIME – The Konstanz Information Miner. *SIGKDD Explorations*, 11(1): 26–31, Oct. 2009. 84, 109

[24] J. Besnard, G. F. Ruda, V. Setola, K. Abecassis, R. M. Rodriguiz, X.-P. Huang, S. Norval, M. F. Sassano, A. I. Shin, L. A. Webster, F. R. C. Simeons, L. Stojanovski, A. Prat, N. G. Seidah, D. B. Constam, G. R. Bickerton, K. D. Read, W. C. Wetsel, I. H. Gilbert, B. L. Roth, and A. L. Hopkins. Automated design of ligands to polypharmacological profiles. *Nature*, 492(7428):215–220, Dec. 2012. 21

[25] P. J. Beuning, S. M. Simon, V. G. Godoy, D. F. Jarosz, and G. C. Walker. Characterization of Escherichia coli Translesion Synthesis Polymerases and Their Accessory Factors. In *DNA Repair, Part A*, pages 318–340. Elsevier, 2006. 58

[26] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider. An estimation of the number of cells in the human body. *Annals of human biology*, 40(6):463–471, Nov. 2013. 5

[27] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. G. Cassarino, M. Bertoni, L. Bordoli, and T. Schwede. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(W1):W252–W258, 2014. 49

[28] K. Blake and J. Lima. Pharmacogenomics of long-acting β2-agonists. *Expert opinion on drug metabolism & toxicology*, 11(11):1733–1751, 2015. 72

[29] J. M. Blaney and J. S. Dixon. A good ligand is hard to find: Automated docking methods. *Perspectives in Drug Discovery and Design*, 1993. 16

[30] J. M. Blaney, P. K. Weiner, A. Dearing, P. A. Kollman, E. C. Jorgensen, S. J. Oatley, J. M. Burridge, and C. C. F. Blake. Molecular mechanics simulation of protein-ligand interactions: binding of thyroid hormone analogs to prealbumin. *Journal of the American Chemical Society*, 104(23):6424–6434, Nov. 1982. 14

[31] Z. D. Blount, J. E. Barrick, C. J. Davidson, and R. E. Lenski. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*, 489(7417):513–518, Sept. 2012. 8

[32] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–70, Jan. 2004. 86

[33] R. Bonneau and D. Baker. Ab Initio Protein Structure Prediction: Progress and Prospects. *dx.doi.org*, 30(1):173–189, Nov. 2003. 25

[34] L. Bordoli, F. Kiefer, K. Arnold, P. Benkert, J. Battey, and T. Schwede. Protein structure homology modeling using SWISS-MODEL workspace. *Nature protocols*, 4(1):1–13, 2009. 49

[35] K. Bowers, E. Chow, H. Xu, R. Dror, M. Eastwood, B. Gregersen, J. Klepeis, I. Kolossvary, M. Moraes, F. Sacerdoti, J. Salmon, Y. Shan, and D. Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *ACM/IEEE SC 2006 Conference (SC'06)*, pages 43–43. IEEE, 2006. 109

[36] P. Bradley, K. M. S. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, Sept. 2005. 25

[37] K. Brandt, S. Maiwald, B. Herkenhoff-Hesselmann, K. Gnirß, J.-C. Greie, S. D. Dunn, and G. Deckers-Hebestreit. Individual interactions of the b subunits within the stator of the Escherichia coli ATP synthase. *Journal of Biological Chemistry*, 288(34):24465–24479, Aug. 2013. 49

[38] L. Breiman. Random Forests. *Machine learning*, 45(1):5–32, 2001. 94, 104

[39] R. R. Brinkman, M.-P. Dubé, G. A. Rouleau, A. C. Orr, and M. E. Samuels. Human monogenic disorders — a source of novel drug targets. *Nature Reviews Genetics*, 7(4): 249–260, Mar. 2006. 23

[40] S. Buchenberg, F. Sittel, and G. Stock. Time-resolved observation of protein allosteric communication. *Proceedings of the National Academy of Sciences of the United States of America*, 114(33):E6804–E6811, Aug. 2017. 123

[41] M. E. Bunnage. Getting pharmaceutical R&D back on target. *Nature chemical biology*, 2011. 12, 13

[42] L. Burger and E. van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, 6(1):e1000633, Jan. 2010. 27

[43] W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012. 10

[44] W. S. Bush, D. R. Crosslin, A. Owusu Obeng, J. Wallace, B. Almoguera, M. A. Basford, S. J. Bielinski, D. S. Carrell, J. J. Connolly, D. Crawford, K. F. Doheny, C. J. Gallego, A. S. Gordon, B. Keating, J. Kirby, et al. Genetic variation among 82 pharmacogenes: The PGRNseq data from the eMERGE network. *Clinical Pharmacology & Therapeutics*, 100 (2):160–169, Aug. 2016. 14, 63, 186

[45] D. S. Cao, L. X. Zhang, G. S. Tan, Z. Xiang, W. B. Zeng, Q. S. Xu, and A. F. Chen. Computational Prediction of Drug-Target Interactions Using Chemical, Biological, and Network Features. *Molecular Informatics*, pages n/a–n/a, Sept. 2014. 24, 25

[46] J. G. R. Cardoso, M. R. Andersen, M. J. Herrgård, and N. Sonnenschein. Analysis of genetic variation and potential applications in genome-scale metabolic modeling. *Frontiers in bioengineering and biotechnology*, 3:13, 2015. 8, 9

[47] J. Casbon and J. Dougherty. PyVCF - A Variant Call Format reader for Python. `http://pyvcf.readthedocs.io/en/latest`, 2011. Version: 0.6.8. 115

[48] Centers for Disease Control and Prevention. *International classification of diseases, ninth revision, clinical modification (ICD-9-CM)*. URL: http://www. cdc. gov/nchs/about/ ..., 2013. 86

[49] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*, 2(5):401–404, May 2012. 108, 121

[50] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(Database issue):D685–90, Jan. 2011. 19, 87

[51] S. Chakradhar. Tumor sequencing takes off, but insurance reimbursement lags. *Nature Medicine*, 20(11):1220–1221, Nov. 2014. 4, 107

[52] R. Charlab and L. Zhang. Pharmacogenomics: historical perspective and current status. *Methods in molecular biology (Clifton, N.J.)*, 1015(Chapter 1):3–22, 2013. 13

[53] D. I. Chasman, D. Posada, L. Subrahmanyan, N. R. Cook, V. P. Stanton, and P. M. Ridker. Pharmacogenetic study of statin therapy and cholesterol reduction. *JAMA*, 291(23):2821–2827, 2004. 74

[54] S. Chaudhury, M. Berrondo, B. D. Weitzner, P. Muthu, H. Bergman, and J. J. Gray. Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS ONE*, 6(8): e22477, 2011. 36

[55] B. Chen and A. J. Butte. Leveraging Big Data to Transform Target Selection and Drug Discovery. *Clinical Pharmacology & Therapeutics*, 99(3):285–297, Mar. 2016. 13

[56] X. Chen, M.-X. Liu, and G.-Y. Yan. Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. BioSyst.*, 8(7):1970–1978, 2012. 24

[57] R. Cheng, R. K.-K. Leung, Y. Chen, Y. Pan, Y. Tong, Z. Li, L. Ning, X. B. Ling, and J. He. Virtual Pharmacist: A Platform for Pharmacogenomics. *PLoS ONE*, 10(10):e0141105, 2015. 108

[58] Y. Cheng. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell*, 161(3):450–457, Apr. 2015. 35

[59] A. P. Chiang and A. J. Butte. Systematic Evaluation of Drug–Disease Relationships to Identify Leads for Novel Drug Uses. *Clinical Pharmacology & Therapeutics*, 86(5):507–510, Nov. 2009. 23, 93, 103

[60] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, (5):823–826, 1986. 25

[61] S. Claerhout, J. Y. Lim, W. Choi, Y.-Y. Park, K. Kim, S.-B. Kim, J.-S. Lee, G. B. Mills, and J. Y. Cho. Gene Expression Signature Analysis Identifies Vorinostat as a Candidate Therapy for Gastric Cancer. *PLoS ONE*, 6(9), 2011. 23

[62] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71–e71, May 2016. 115

[63] T. . G. P. Consortium. A global reference for human genetic variation. *Nature*, 526(7571): 68–74, Oct. 2015. 3, 8, 13, 23, 63

[64] U. Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 2014. 48, 86

[65] D. Cook, D. Brown, and R. Alexander. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat Rev ...*, 2014. 12, 13

[66] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 118(9):2309–2309, 1996. 14, 15, 16

[67] S. Croset, J. P. Overington, and D. Rebholz-Schuhmann. The functional therapeutic chemical classification system. *Bioinformatics (Oxford, England)*, 30(6):876–883, Mar. 2014. 24

[68] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. 91

[69] K. J. Czogalla, A. Biswas, K. Höning, V. Hornung, K. Liphardt, M. Watzka, and J. Oldenburg. Warfarin and vitamin K compete for binding to Phe55 in human VKOR. *Nature Structural & Molecular Biology*, 24(1):77–85, Jan. 2017. 78

[70] A. K. Daly. Genome-wide association studies in pharmacogenomics. *Nature Reviews Genetics*, 11(4):241–246, Apr. 2010. 3, 64

[71] T. Dandekar. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences*, 23(9):324–328, Sept. 1998. 18

[72] M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046–U124, Nov. 2011. 108, 111, 113, 118, 119, 120

[73] E. De Franchi, C. Schalon, M. Messa, F. Onofri, F. Benfenati, and D. Rognan. Binding of Protein Kinase Inhibitors to Synapsin I Inferred from Pair-Wise Binding Site Similarity Measurements. *PLoS ONE*, 5(8):e12214, Aug. 2010. 21

[74] J. DeLeon-Rangel, D. Zhang, and S. B. Vik. The role of transmembrane span 2 in the structure and function of subunit a of the ATP synthase from Escherichia coli. *Archives of biochemistry and biophysics*, 418(1):55–62, Oct. 2003. 59

[75] J. DeLeon-Rangel, R. R. Ishmukhametov, W. Jiang, R. H. Fillingame, and S. B. Vik. Interactions between subunits aand bin the rotary ATP synthase as determined by cross-linking. *FEBS letters*, 587(7):892–897, Feb. 2013. 49, 59

[76] F. Delsuc, H. Brinkmann, and H. Philippe. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 2005. 1

[77] G. D. Demetri, P. Reichardt, Y.-K. Kang, J.-Y. Blay, P. Rutkowski, H. Gelderblom, P. Hohenberger, M. Leahy, M. von Mehren, H. Joensuu, G. Badalamenti, M. Blackstein, A. Le Cesne, P. Schöffski, R. G. Maki, S. Bauer, B. B. Nguyen, J. Xu, T. Nishida, J. Chung, C. Kappeler, I. Kuss, D. Laurent, and P. G. Casali. The Human Genome Project Completion: Frequently Asked Questions, Oct. 2010. URL `https://www.genome.gov/11006943/`. 5

[78] Z. Dezso, Y. Nikolsky, T. Nikolskaya, J. Miller, D. Cherba, C. Webb, and A. Bugrim. Identifying disease-specific genes based on their topological significance in protein networks. *BMC systems biology*, 3(1):36, 2009. 24

[79] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of health economics*, 47:20–33, May 2016. 12

[80] O. Y. Dmitriev, P. C. Jones, and R. H. Fillingame. Structure of the subunit c oligomer in the F1Fo ATP synthase: model derived from solution structure of the monomer and cross-linking in the native enzyme. *Proceedings of the National Academy of Sciences*, 96(14):7785–7790, July 1999. 59

[81] S. Domcke, R. Sinha, D. A. Levine, C. Sander, and N. Schultz. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature Communications*, 4:2126, 2013. 23

[82] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society*, 125(7):1731–1737, Feb. 2003. 36, 46, 51

[83] R. N. dos Santos, F. Morcos, B. Jana, A. D. Andricopulo, and J. N. Onuchic. Dimeric interactions and complex formation using direct coevolutionary couplings. *Scientific Reports*, 5, 2015. 61

[84] L. Drew. Pharmacogenetics: The right drug for you. *Nature*, 537(7619):S60–2, Sept. 2016. 82

[85] B. I. Droegemoeller, G. E. B. Wright, D. J. H. Niehaus, R. Emsley, and L. Warnich. Next-generation sequencing of pharmacogenes: a critical analysis focusing on schizophrenia treatment. *Pharmacogenetics and genomics*, 23(12):666–674, Dec. 2013. 81

[86] Drugs.com. Top 100 Drugs for 2013 by Units - U.S. Pharmaceutical Statistics, 2013. URL `https://www.drugs.com/stats/top100`. Accessed: January 2018. 68, 72

[87] J. T. Dudley, R. Tibshirani, T. Deshpande, and A. J. Butte. Disease signatures are robust across tissues and experiments. *Molecular Systems Biology*, 5(1):307, 2009. 23

[88] J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha, and A. J. Butte. Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease. *Science Translational Medicine*, 3(96):–96ra76, 2011. 23

[89] S. D. Dunn, L. M. Wahl, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics (Oxford, England)*, 24(3):333–340, Feb. 2008. 30

[90] H. M. Dunnenberger, K. R. Crews, J. M. Hoffman, K. E. Caudle, U. Broeckel, S. C. Howard, R. J. Hunkler, T. E. Klein, W. E. Evans, and M. V. Relling. Preemptive Clinical Pharmacogenetics Implementation: Current programs in five United States medical centers. *Annual review of pharmacology and toxicology*, 55(1):89–106, 2015. 63, 79, 82, 83, 107

[91] M. Duran-Frigola and P. Aloy. Recycling side-effects into clinical markers for drug repositioning. *Genome medicine*, 2012. 22

[92] S. Eddy. *HMMER (Version 2.3. 2) User's Guide*. HHMI/Washington University School of Medicine, 2003. 38

[93] S. R. Eddy. PLOS Computational Biology: Accelerated Profile HMM Searches. *PLoS computational biology*, 2011. 38

[94] A. W. Edwards. The Genetical Theory of Natural Selection. *Genetics*, 154(4):1419–1426, Apr. 2000. 6

[95] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 87(1):012707, Jan. 2013. 27, 28, 29, 42, 49

[96] M. Ekeberg, T. Hartonen, and E. Aurell. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *arXiv.org*, Jan. 2014. 27, 28

[97] D. Emig, A. Ivliev, O. Pustovalova, L. Lancashire, S. Bureeva, Y. Nikolsky, and M. Bessarabova. Drug target prediction and repositioning using an integrated network-based approach. *PLoS ONE*, 8(4):e60618, 2013. 103

[98] M. Y. Eng, S. E. Luczak, and T. L. Wall. ALDH2, ADH1B, and ADH1C genotypes in Asians: a literature review. *Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism*, 30(1):22–27, 2007. 76

[99] A. Ezzat, P. Zhao, M. Wu, X. Li, and C. K. Kwoh. Drug-Target Interaction Prediction with Graph Regularized Matrix Factorization. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, pages 1–1, Feb. 2016. 24

[100] M. A. Fabian, W. H. Biggs, D. K. Treiber, C. E. Atteridge, M. D. Azimioara, M. G. Benedetti, T. A. Carter, P. Ciceri, P. T. Edeen, M. Floyd, J. M. Ford, M. Galvin, J. L. Gerlach, R. M. Grotzfeld, S. Herrgard, D. E. Insko, M. A. Insko, A. G. Lai, J.-M. Lélias, S. A. Mehta, Z. V. Milanov, A. M. Velasco, L. M. Wodicka, H. K. Patel, P. P. Zarrinkar, and D. J. Lockhart. A small molecule–kinase interaction map for clinical kinase inhibitors. *Nature Biotechnology*, 23(3):329–336, Feb. 2005. 108

[101] G. Faure, J. Andreani, and R. Guerois. InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Research*, 40(D1):gkr845–D856, Nov. 2011. 36

[102] M. Ferber, J. Kosinski, A. Ori, U. J. Rashid, M. Moreno-Morcillo, B. Simon, G. Bouvier, P. R. Batista, C. W. Müller, M. Beck, and M. Nilges. Automated structure modeling of large protein assemblies using crosslinks as distance restraints. *Nature methods*, 13(6):515–520, June 2016. 62

[103] L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, Feb. 2006. 8

[104] R. H. Fillingame and P. R. Steed. Half channels mediating H transport and the mechanism of gating in the F-o sector of Escherichia coli F1Fo ATP synthase. *Biochimica Et Biophysica Acta-Bioenergetics*, 1837(7):1063–1068, July 2014. 59

[105] R. D. Finn, J. Clements, and S. R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue):W29–37, July 2011. 39

[106] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222–D230, Jan. 2014. 86, 87

[107] N. M. Fischer. *Modeling Flexibility in Protein-DNA and Protein-Ligand Complexes using Molecular Dynamics*. PhD thesis, Sept. 2013. 110, 111

[108] R. A. Fisher. *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford, 1930. 6

[109] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, R. Stefancsik, B. Harsha, C. Y. Kok, M. Jia, H. Jubb, Z. Sondka, S. Thompson, T. De, and P. J. Campbell. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, Jan. 2017. 107

[110] D. M. Fowler and S. Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, Aug. 2014. 81

[111] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41 (Database issue):D808–15, Jan. 2013. 87

[112] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, Mar. 2004. 111

[113] R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, and D. T. Mainz. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21):6177–6196, Oct. 2006. 111

[114] K. Fujikura, M. Ingelman-Sundberg, and V. M. Lauschke. Genetic variation in the human cytochrome P450 supergene family. *Pharmacogenetics and genomics*, 25(12):584–594, Dec. 2015. 12, 14, 63, 67, 76, 81, 186

[115] B. F. Gage, C. Eby, J. A. Johnson, E. Deych, M. J. Rieder, P. M. Ridker, P. E. Milligan, G. Grice, P. Lenzini, A. E. Rettie, C. L. Aquilante, L. Grosso, S. Marsh, T. Langaee, L. E. Farnett, D. Voora, D. L. Veenstra, R. J. Glynn, A. Barrett, and H. L. McLeod. Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. *Clinical Pharmacology & Therapeutics*, 84(3):326–331, Sept. 2008. 64

[116] M. Y. Galperin and E. V. Koonin. Who's your neighbor? New computational approaches for functional genomics. *Nature Biotechnology*, 2000. 18

[117] L. Gatta, D. Vaira, G. Sorrenti, S. Zucchini, C. Sama, and N. Vakil. Meta-analysis: the efficacy of proton pump inhibitors for laryngeal symptoms attributed to gastro-oesophageal reflux disease. *Alimentary pharmacology & therapeutics*, 25(4):385–392, Feb. 2007. 74

[118] V. Genetics. Veritas Genetics Launches $999 Whole Genome And Sets New Standard For Genetic Testing , Mar. 2016. 1

[119] U. Göbel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins . *Proteins-Structure Function and Bioinformatics*, (18):309–317, 1994. 19, 26, 36

[120] Y. Gofman, C. Schärfe, D. S. Marks, T. Haliloglu, and N. Ben-Tal. Structure, dynamics and implied gating mechanism of a human cyclic nucleotide-gated channel. *PLoS computational biology*, 10(12):e1003976, Dec. 2014. 39, 43, 61

[121] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, May 2007. 20, 86, 94

[122] D. Golan, E. S. Lander, and S. Rosset. Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences of the United States of America*, 111(49):E5272–81, Dec. 2014. 10

[123] Y. Gong, C. W. McDonough, Z. Wang, W. Hou, R. M. Cooper-DeHoff, T. Y. Langaee, A. L. Beitelshees, A. B. Chapman, J. G. Gums, K. R. Bailey, E. Boerwinkle, S. T. Turner, and J. A. Johnson. Hypertension Susceptibility Loci and Blood Pressure Response to Antihypertensives - Results from the Pharmacogenomic Evaluation of Antihypertensive Responses (PEAR) Study. *Circulation: Cardiovascular Genetics*, 5(6):CIRCGENETICS.112.964080–691, Oct. 2012. 64

[124] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos, and N. Lopez-Bigas. IntOGen-mutations identifies cancer drivers across tumor types. *Nature methods*, 10(11):1081–1082, Nov. 2013. 107

[125] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, May 2016. 1

[126] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, May 2016. 1

[127] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7(1): 496–496, Jan. 2011. 24, 103, 105

[128] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Research*, 43(Database issue):D1079–85, Jan. 2015. 86

[129] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, et al. A draft sequence of the Neandertal genome. *Science*, 328 (5979):710–722, May 2010. 1

[130] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, D. I. Chasman, G. A. FitzGerald, K. Dolinski, T. Grosser, and O. G. Troyanskaya. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 47(6):569–576, Apr. 2015. 18, 19

[131] P. A. Greenidge, C. Kramer, J.-C. Mozziconacci, and R. M. Wolf. MM/GBSA binding energy prediction on the PDBbind data set: successes, failures, and directions for further improvement. *Journal of Chemical Information and Modeling*, 53(1):201–209, Jan. 2013. 123, 125

[132] P. A. Greenidge, C. Kramer, J. C. Mozziconacci, and W. Sherman. Improving Docking Results via Reranking of Ensembles of Ligand Poses in Multiple X-ray Protein Conformations with MM-GBSA. *Journal of Chemical Information and Modeling*, 54(10):2697–2717, Oct. 2014. 111, 118, 123, 125

[133] M. Griffith, N. C. Spies, K. Krysiak, J. F. McMichael, A. C. Coffman, A. M. Danos, B. J. Ainscough, C. A. Ramirez, D. T. Rieke, L. Kujan, E. K. Barnell, A. H. Wagner, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature genetics*, 49(2):170–174, Jan. 2017. 107, 108, 123

[134] D. G. Grimm, C.-A. Azencott, F. Aicheler, U. Gieraths, D. G. MacArthur, K. E. Samocha, D. N. Cooper, P. D. Stenson, M. J. Daly, J. W. Smoller, L. E. Duncan, and K. M. Borgwardt. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Human Mutation*, 36(5):513–523, May 2015. 80

[135] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, Oct. 2017. 105

[136] T. Gueudré, C. Baldassi, M. Zamparo, M. Weigt, and A. Pagnani. Simultaneous identification of specifically interacting paralogs and inter-protein contacts by Direct-Coupling Analysis. May 2016. 18, 41, 61

[137] E. Guney, J. Menche, M. Vidal, and A.-L. Barabási. Network-based in silico drug efficacy screening. *Nature Communications*, 7:10331, Feb. 2016. 24

[138] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of Medicinal Chemistry*, 47(7):1750–1759, Mar. 2004. 111

[139] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–55, Jan. 2002. 86

[140] S. M. Han, J. Park, J. H. Lee, S. S. Lee, H. Kim, H. Han, Y. Kim, S. Yi, J. Y. Cho, I. J. Jang, and M. G. Lee. Targeted Next-Generation Sequencing for Comprehensive Genetic Profiling of Pharmacogenes. *Clinical Pharmacology & Therapeutics*, 101(3):396–405, Mar. 2017. 81

[141] C. Hanson, J. Cairns, L. Wang, and S. Sinha. Computational discovery of transcription factors associated with drug response. *Pharmacogenomics Journal*, 16(6):573–582, Nov. 2016. 81

[142] S. B. Hari, E. A. Merritt, and D. J. Maly. Sequence determinants of a specific inactive protein kinase conformation. *Chemistry & Biology*, 20(6):806–815, June 2013. 112

[143] S. B. Hari, E. A. Merritt, and D. J. Maly. Sequence determinants of a specific inactive protein kinase conformation. *Chemistry & Biology*, 20(6):806–815, June 2013. 124

[144] A. R. Harper and E. J. Topol. Pharmacogenomics in clinical practice and drug development. *Nature Biotechnology*, 30(11):1117–1124, Nov. 2012. 13, 14

[145] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009. 94

[146] V. J. Haupt and M. Schroeder. Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Briefings in Bioinformatics*, 12(4):312–326, July 2011. 21

[147] A. Haustant. Flask-RESTPlus. `https://github.com/noirbizarre/flask-restplus`. Version: 0.9.2. 114

[148] P. C. Havugimana, G. T. Hart, T. Nepusz, H. Yang, A. L. Turinsky, Z. Li, P. I. Wang, D. R. Boutz, V. Fong, S. Phanse, M. Babu, S. A. Craig, P. Hu, C. Wan, J. Vlasblom, V.-u.-N. Dar, A. Bezginov, G. W. Clark, G. C. Wu, S. J. Wodak, E. R. M. Tillier, A. Paccanaro, E. M. Marcotte, and A. Emili. A census of human soluble protein complexes. *Cell*, 150(5): 1068–1081, Aug. 2012. 18

[149] S. Hayat, C. Sander, D. S. Marks, and A. Elofsson. All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 112(17):5413–5418, Apr. 2015. 43

[150] B. M. Henn, L. R. Botigué, S. Peischl, I. Dupanloup, M. Lipatov, B. K. Maples, A. R. Martin, S. Musharoff, H. Cann, M. P. Snyder, L. Excoffier, J. M. Kidd, and C. D. Bustamante. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(4):E440–9, Jan. 2016. 74

[151] T. Hilal, M. Nakazawa, J. Hodskins, J. L. Villano, A. Mathew, G. Goel, L. Wagner, S. M. Arnold, P. DeSimone, L. B. Anthony, and P. J. Hosein. Comprehensive genomic profiling in routine clinical practice leads to a low rate of benefit from genotype-directed therapy. *BMC cancer*, 17(1):602, Aug. 2017. 125

[152] A. Hildebrandt, A. K. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N. C. Toussaint, A. Moll, D. Stöckel, S. Nickels, S. C. Mueller, H.-P. Lenhof, and O. Kohlbacher. BALL–biochemical algorithms library 1.3. *BMC Bioinformatics*, 11(1):531, Oct. 2010. 110

[153] D. S. Himmelstein and S. E. Baranzini. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS computational biology*, 11(7):e1004259, July 2015. 19

[154] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *bioRxiv*, page 087619, Nov. 2016. 24, 84, 87, 103, 104, 105

[155] M. Hinderer, M. Boerries, F. Haller, S. Wagner, S. Sollfrank, T. Acker, H.-U. Prokosch, and J. Christoph. Supporting Molecular Tumor Boards in Molecular-Guided Decision-Making -

The Current Status of Five German University Hospitals. *Studies in health technology and informatics*, 236:48–54, 2017. 125

[156] J. Hintzsche, J. Kim, V. Yadav, C. Amato, S. E. Robinson, E. Seelenfreund, Y. Shellman, J. Wisell, A. Applegate, M. McCarter, N. Box, J. Tentler, S. De, W. A. Robinson, and A. C. Tan. IMPACT: a whole-exome sequencing analysis pipeline for integrating molecular profiles with actionable therapeutics in clinical samples. *Journal of the American Medical Informatics Association : JAMIA*, 23(4):721–730, July 2016. 108

[157] T. K. Ho. Random decision forests. In *3rd International Conference on Document Analysis and Recognition*, pages 278–282. IEEE Comput. Soc. Press, 1995. 94

[158] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer*, 13(10):714–726, Oct. 2013. 2

[159] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–1621, June 2012. 2, 36, 39, 42, 43, 49, 59

[160] T. A. Hopf, C. P. I. Schärfe, J. P. G. L. M. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. J. J. Bonvin, and D. S. Marks. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3, 2014. 35, 45, 53, 54, 56, 57, 58

[161] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Schärfe, M. Springer, C. Sander, and D. S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, Feb. 2017. 25, 33, 78, 80, 127

[162] A. L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690, Nov. 2008. 13

[163] A. L. Hopkins and C. R. Groom. The druggable genome. *Nature Reviews Drug Discovery*, 1 (9):727–730, Sept. 2002. 21, 22

[164] P. Horak, B. Klink, C. Heining, S. Gröschel, B. Hutter, M. Fröhlich, S. Uhrig, D. Hübschmann, M. Schlesner, R. Eils, D. Richter, K. Pfütze, C. Geörg, B. Meißburger, S. Wolf, A. Schulz, R. Penzel, E. Herpel, M. Kirchner, A. Lier, V. Endris, S. Singer, P. Schirmacher, W. Weichert, A. Stenzinger, R. F. Schlenk, E. Schröck, B. Brors, C. von Kalle, H. Glimm, and S. Fröhling. Precision oncology based on omics data: The NCT Heidelberg experience. *International journal of cancer. Journal international du cancer*, 141(5):877–886, Sept. 2017. 125

[165] X. Huang, B. Luan, J. Wu, and Y. Shi. An atomic structure of the human 26S proteasome. *Nature Structural & Molecular Biology*, July 2016. 35

[166] Y.-F. Huang, H.-Y. Yeh, and V.-W. Soo. Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. *Bmc Medical Genomics*, 6(Suppl 3), 2013. 23, 24

[167] C. A. Hudis. Trastuzumab — Mechanism of Action and Use in Clinical Practice. *New England Journal of Medicine*, 357(1):39–51, July 2007. 64

[168] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, Mar. 2011. 12

[169] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Y. Dai, Y. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102 (1):109–126, 2000. 23

[170] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–8– 27–8, Feb. 1996. 110

[171] L. D. Hurst. Genetics and the understanding of selection. *Nature Reviews Genetics*, 10(2): 83–93, Feb. 2009. 6, 7

[172] E. L. Huttlin, L. Ting, R. J. Bruckner, F. Gebreab, M. P. Gygi, J. Szpyt, S. Tam, G. Zarraga, G. Colby, K. Baltier, R. Dong, V. Guarani, L. P. Vaites, A. Ordureau, R. Rad, et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, 162(2):425–440, July 2015. 17, 18, 19, 35

[173] R. N. Hvorup, B. A. Goetz, M. Niederer, K. Hollenstein, E. Perozo, and K. P. Locher. Asymmetry in the structure of the ABC transporter-binding protein complex BtuCD-BtuF. *Science*, 317(5843):1387–1390, Sept. 2007. 51

[174] D. M. Hyman, I. Puzanov, V. Subbiah, J. E. Faris, I. Chau, J.-Y. Blay, J. Wolf, N. S. Raje, E. L. Diamond, A. Hollebecque, R. Gervais, M. E. Elez-Fernandez, A. Italiano, R.-D. Hofheinz, M. Hidalgo, E. Chan, M. Schuler, S. F. Lasserre, M. Makrutzki, F. Sirzen, M. L. Veronese, J. Tabernero, and J. Baselga. Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations. *New England Journal of Medicine*, 373(8):726–736, Aug. 2015. 121

[175] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, and D. Di Bernardo. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621–14626, 2010. 23, 83

[176] F. Iorio, A. Isacchi, D. Di Bernardo, and N. Brunetti-Pierri. Identification of small molecules enhancing autophagic function from drug network analysis. *Autophagy*, 6(8):1204–1205, Nov. 2010.

[177] F. Iorio, R. L. Shrestha, N. Levin, V. Boilot, M. J. Garnett, J. Saez-Rodriguez, and V. M. Draviam. A Semi-Supervised Approach for Refining Transcriptional Signatures of Drug Response and Repositioning Predictions. *PLoS ONE*, 10(10):e0139446, Oct. 2015. 23, 83

[178] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754, July 2016. 70

[179] H. Iwata, R. Sawada, S. Mizutani, and Y. Yamanishi. Systematic Drug Repositioning for a Wide Range of Diseases with Integrative Analyses of Phenotypic and Molecular Data. *Journal of Chemical Information and Modeling*, 55(2):446–459, Jan. 2015. 24, 25, 103, 105

[180] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, Apr. 2018. 5

[181] E. T. Jaynes. Information Theory and Statistical Mechanics. II. *Physical review*, 108(2): 171–190, Oct. 1957. 27

[182] L. J. Jensen. STRING and STITCH: known and predicted interactions between proteins and chemicals. *Nature Precedings*, (713), Sept. 2008. 19

[183] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(Database issue):D412–6, Jan. 2009. 19

[184] G. Jin, C. Fu, H. Zhao, K. Cui, J. Chang, and S. T. C. Wong. A Novel Method of Transcriptional Response Analysis to Facilitate Drug Repositioning for Cancer Therapy. *Cancer research*, 72(1):33–44, 2012. 23, 83

[185] D. B. Johnson, K. H. Dahlman, J. Knol, J. Gilbert, I. Puzanov, J. Means-Powell, J. M. Balko, C. M. Lovly, B. A. Murphy, L. W. Goff, V. G. Abramson, M. A. Crispens, I. A. Mayer, J. D. Berlin, L. Horn, V. L. Keedy, N. M. Reddy, C. L. Arteaga, J. A. Sosman, and W. Pao. Enabling a genetically informed approach to cancer medicine: a retrospective evaluation of the impact of comprehensive tumor profiling using a targeted next-generation sequencing panel. *The Oncologist*, 19(6):616–622, June 2014. 125

[186] E. Johnson, P. T. Nguyen, T. O. Yeates, and D. C. Rees. Inward facing conformations of the MetNI methionine ABC transporter: Implications for the mechanism of transinhibition. *Protein science : a publication of the Protein Society*, 21(1):84–96, Jan. 2012. 51

[187] J. A. Johnson, I. Zineh, B. J. Puckett, S. P. McGorray, H. N. Yarandi, and D. F. Pauly. beta(1)-adrenergic receptor polymorphisms and antihypertensive response to metoprolol. *Clinical Pharmacology & Therapeutics*, 74(1):44–52, July 2003. 64

[188] J. A. Johnson, L. Gong, M. Whirl-Carrillo, B. F. Gage, S. A. Scott, C. M. Stein, J. L. Anderson, S. E. Kimmel, M. T. M. Lee, M. Pirmohamed, M. Wadelius, T. E. Klein, and R. B. Altman. Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and

VKORC1 Genotypes and Warfarin Dosing. *Clinical Pharmacology & Therapeutics*, 90(4): 625–629, Oct. 2011. 64, 80

[189] L. S. Johnson, S. R. Eddy, and E. Portugaly. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1):1, Aug. 2010. 38, 48

[190] M. A. Johnson and G. M. Maggiora. *Concepts and applications of molecular similarity*. Wiley, 1990. 21

[191] D. E. Jonas and H. L. McLeod. Genetic and clinical factors relating to warfarin dosing. *Trends in pharmacological sciences*, 30(7):375–386, July 2009. 64

[192] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)*, 28(2):184–190, Jan. 2012. 29, 43, 47

[193] E. Jones, T. Oliphant, and P. Peterson. SciPy: Open source scientific tools for Python. `https://www.scipy.org`, 2001. Version: 0.17.0. 93, 113, 115

[194] W. L. Jorgensen and J. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 1988. 14

[195] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, Aug. 1998. 109

[196] D. Juan, F. Pazos, and A. Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences of the United States of America*, 105(3):934–939, Jan. 2008. 19

[197] H. Kamisetty, S. Ovchinnikov, and D. Baker. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39):15674–15679, Sept. 2013. 29

[198] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(Database Issue): D109–D114, Nov. 2012. 19, 86

[199] E. D. Kantor, C. D. Rehm, J. S. Haas, A. T. Chan, and E. L. Giovannucci. Trends in Prescription Drug Use Among Adults in the United States From 1999-2012. *JAMA*, 314(17): 1818–1830, Nov. 2015. 63, 64, 74

[200] E. Karaca and A. M. J. J. Bonvin. Advances in integrative modeling of biomolecular complexes. *Methods (San Diego, Calif.)*, 59(3):372–381, Mar. 2013. 36

[201] J. Karow. German Health Insurance Allows Reimbursement of NGS Tests Up to Limit; Excludes cfDNA Analysis, PGx, July 2016. URL `https://www.genomeweb.com/molecular-diagnostics/german-health-insurance-allows-reimbursement-ngs-tests-limit-excludes-cfdna`. 107

[202] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ (Clinical research ed.)*, 330(7494):765, Mar. 2005. 4, 128

[203] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, 25(2):197–206, Feb. 2007. 3, 21

[204] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, Nov. 2009. 21

[205] O. Keskin, N. Tuncbag, and A. Gursoy. Predicting Protein–Protein Interactions from the Molecular to the Proteome Level. *Chemical reviews*, 116(8):4884–4909, Apr. 2016. 17, 18, 19, 20

[206] J. Kim, M. Yoo, J. Kang, and A. C. Tan. K-Map: connecting kinases with therapeutics for drug repurposing and development. *Hum Genomics*, 2013. 23, 83

[207] S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie, and P. E. Bourne. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS computational biology*, 5(7):e1000423, July 2009. 21

[208] G. Kiss, D. Röthlisberger, D. Baker, and K. N. Houk. Evaluation and ranking of enzyme designs. *Protein science*, 2010. 123

[209] G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker, and K. N. Houk. Computational enzyme design. *Angewandte Chemie (International ed. in English)*, 52(22):5700–5725, May 2013. 108, 109

[210] H. Kitano. A robustness-based approach to systems-oriented drug design. *Nature Reviews Drug Discovery*, 6(3):202–210, Mar. 2007. 13

[211] S. K. Kjærulff, L. Wich, J. Kringelum, U. P. Jacobsen, I. Kouskoumvekaki, K. Audouze, O. Lund, S. Brunak, T. I. Oprea, and O. Taboureau. ChemProt-2.0: visual navigation in a disease chemical biology database. *Nucleic Acids Research*, 41(D1):D464–D469, Jan. 2013. 84

[212] T. E. Klein, R. B. Altman, N. Eriksson, B. F. Gage, S. E. Kimmel, M. T. M. Lee, N. A. Limdi, D. Page, D. M. Roden, M. J. Wagner, J. A. Johnson, Y. T. Chen, M. S. Wen, Y. Caraco, I. Achache, S. Blotnick, M. Muszkat, J. G. Shin, et al. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. *New England Journal of Medicine*, 360(8):753–764, 2009. 64

[213] T. C. Knepper, G. C. Bell, J. K. Hicks, E. Padron, J. K. Teer, T. T. Vo, N. K. Gillis, N. T. Mason, H. L. McLeod, and C. M. Walko. Key Lessons Learned from Moffitt's Molecular Tumor Board: The Clinical Genomics Action Committee Experience. *The Oncologist*, 22(2): 144–151, Feb. 2017. 4, 107

[214] I. Kola and J. Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–716, Aug. 2004. 12

[215] P. A. Kollman, P. K. Weiner, and A. Dearing. Studies of nucleotide conformations and interactions. The relative stabilities of double-helical B-DNA sequence isomers. *Biopolymers*, 20(12):2583–2621, Dec. 1981. 14

[216] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences*, 99(22):14116–14121, Oct. 2002. 36

[217] T. Kortemme and D. Baker. Computational design of protein–protein interactions. *Current opinion in chemical biology*, 8(1):91–97, Feb. 2004. 36

[218] M. Kozyra, M. Ingelman-Sundberg, and V. M. Lauschke. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. *Genetics in Medicine*, Apr. 2016. 12, 14, 63, 67, 79, 81, 186

[219] K. Kramer, T. Sachsenberg, B. M. Beckmann, S. Qamar, K.-L. Boon, M. W. Hentze, O. Kohlbacher, and H. Urlaub. Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature methods*, 11(10): 1064–1070, Oct. 2014. 36

[220] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins-Structure Function and Bioinformatics*, 77(4): 778–795, Dec. 2009. 49

[221] S. D. Kunkel, M. Suneja, S. M. Ebert, K. S. Bongers, D. K. Fox, S. E. Malmberg, F. Alipour, R. K. Shields, and C. M. Adams. mRNA Expression Signatures of Human Skeletal Muscle Atrophy Identify a Natural Compound that Increases Muscle Mass. *Cell Metabolism*, 13(6): 627–638, June 2011. 23

[222] Z. Kutalik, J. S. Beckmann, and S. Bergmann. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nature Biotechnology*, 26(5):531–539, May 2008. 23

[223] E. Kwong. Advancing Drug Discovery: A Pharmaceutics Perspective. *Journal of Pharmaceutical Sciences*, 104(3):865–871, Mar. 2015. 13

[224] J. Lamb. The Connectivity Map: a new tool for biomedical research. *Nature Reviews Cancer*, 7(1):54–60, Jan. 2007. 23

[225] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, Sept. 2006. 23

[226] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001. 1

[227] G. Landrum. RDKit: Open-source cheminformatics. `http://www.rdkit.org`, 2016. Version: 2016.09.4. 86

[228] M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, and D. R. Maglott. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–8, Jan. 2016. 87

[229] A. Lapedes, B. Giraud, and C. Jarzynski. Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. *arXiv.org*, July 2012. 27, 28

[230] A. Laupacis, D. L. Sackett, and R. S. Roberts. An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine*, 318(26):1728–1733, June 1988. 63

[231] V. M. Lauschke and M. Ingelman-Sundberg. Requirements for comprehensive pharmacogenetic genotyping platforms. *Pharmacogenomics*, 17(8):917–924, June 2016. 128

[232] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 42(Database issue):D1091–7, Jan. 2014. 67, 68, 86, 87, 111, 187

[233] M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, Jan. 2014. 108

[234] C. R. Lee, F. G. Bottone, J. M. Krahn, L. Li, H. W. Mohrenweiser, M. E. Cook, R. M. Petrovich, D. A. Bell, T. E. Eling, and D. C. Zeldin. Identification and functional characterization

of polymorphisms in human cyclooxygenase-1 (PTGS1). *Pharmacogenetics and genomics*, 17 (2):145–160, Feb. 2007. 78

[235] D.-S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, and A. L. Barabasi. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29):9880–9885, July 2008. 20, 86

[236] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616): 285–291, Aug. 2016. 2, 3, 23, 65, 67, 68, 74, 81, 86, 187

[237] R. C. Lewontin. The units of selection. *Annual review of ecology and systematics*, 1970. 6

[238] J. Li and Z. Lu. Pathway-based drug repositioning using causal inference. *BMC Bioinformatics*, 14 Suppl 16(Suppl 16):S3, 2013. 23, 31, 103

[239] Y. Liang, Z. Gao, F. Wang, Y. Zhang, Y. Dong, and Q. Liu. Structural and functional characterization of Escherichia coli toxin-antitoxin complex DinJ-YafQ. *Journal of Biological Chemistry*, 289(30):21191–21202, July 2014. 58

[240] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2012 update. *Nucleic acids research*, 40(Database issue):D857–61, Jan. 2012. 19

[241] S. Lifson and A. Warshel. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *The Journal of Chemical Physics*, 49(11):5116–5129, Dec. 1968. 15

[242] H. Lin, M. F. Sassano, B. L. Roth, and B. K. Shoichet. A pharmacological organization of G protein-coupled receptors. *Nature methods*, 10(2):140–146, Feb. 2013. 21

[243] M. Lin and R. Wu. Theoretical basis for the identification of allelic variants that encode drug efficacy and toxicity. *Genetics*, 170(2):919–928, June 2005. 23

[244] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins-Structure Function and Bioinformatics*, 78(8):1950–1958, June 2010. 14

[245] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, Oct. 2011. 25

[246] K. Lindpaintner. The impact of pharmacogenetics and pharmacogenomics on drug discovery. *Nature Reviews Drug Discovery*, 1(6):463–469, June 2002. 10, 63

[247] A. A. Litonjua, L. Gong, Q. L. Duan, J. Shin, M. J. Moore, S. T. Weiss, J. A. Johnson, T. E. Klein, and R. B. Altman. Very important pharmacogene summary ADRB2. *Pharmacogenetics and genomics*, 20(1):64–69, Jan. 2010. 76, 81

[248] J. Liu, Z. Q. Liu, B. N. Yu, F. H. Xu, W. Mo, G. Zhou, Y. Z. Liu, Q. Li, and H. H. Zhou. β1-Adrenergic receptor polymorphisms influence the response to metoprolol monotherapy in patients with essential hypertension. *Clinical Pharmacology & Therapeutics*, 80(1):23–32, July 2006. 64

[249] R. Liu, N. Singh, G. J. Tawa, A. Wallqvist, and J. Reifman. Exploiting large-scale drug-protein interaction information for computational drug repurposing. *BMC Bioinformatics*, 15(1), 2014. 3, 23

[250] R. Liu, X. Li, W. Zhang, and H. H. Zhou. Comparison of Nine Statistical Model Based Warfarin Pharmacogenetic Dosing Algorithms Using the Racially Diverse International Warfarin Pharmacogenetic Consortium Cohort Database. *PLoS ONE*, 10(8):e0135784, Aug. 2015. 64

[251] G. Livingston and C. Katona. The place of memantine in the treatment of Alzheimer's disease: a number needed to treat analysis. *International Journal of Geriatric Psychiatry*, 19(10):919–925, Oct. 2004. 74

[252] R. Loebstein, I. Dvoskin, H. Halkin, M. Vecsler, A. Lubetsky, G. Rechavi, N. Amariglio, Y. Cohen, G. Ken-Dror, S. Almog, and E. Gak. A coding VKORC1 Asp36Tyr polymorphism predisposes to warfarin resistance. *Blood*, 109(6):2477–2480, Mar. 2007. 78

[253] J. C. Long, J. DeLeon-Rangel, and S. B. Vik. Characterization of the first cytoplasmic loop of subunit a of the Escherichia coli ATP synthase by surface labeling, cross-linking, and mutagenesis. *The Journal of biological chemistry*, 277(30):27288–27293, 2002. 59

[254] V. Lounnas, T. Ritschel, J. Kelder, R. McGuire, R. P. Bywater, and N. Foloppe. Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. *Computational and structural biotechnology journal*, 5(6):e201302011–14, 2013. 22

[255] C. Lu, M. Xie, M. C. Wendl, J. Wang, M. D. McLellan, M. D. M. Leiserson, K.-l. Huang, M. A. Wyczalkowski, R. Jayasinghe, T. Banerjee, J. Ning, P. Tripathi, Q. Zhang, B. Niu, et al. Patterns and functional implications of rare germline variants across 12 cancer types. *Nature Communications*, 6, Dec. 2015. 79

[256] S. Lupoli, E. Salvi, M. Barcella, and C. Barlassina. Pharmacogenomics considerations in the control of hypertension. *dx.doi.org*, 16(17):1951–1964, Nov. 2015. 64

[257] D.-L. Ma, D. S.-H. Chan, and C.-H. Leung. Drug repositioning by structure-based virtual screening. *Chem. Soc. Rev.*, 42(5):2130–2141, Feb. 2013. 21

[258] Q. Ma and A. Y. H. Lu. Pharmacogenetics, pharmacogenomics, and individualized medicine. *Pharmacological Reviews*, 63(2):437–459, June 2011. 2, 11, 63

[259] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, Feb. 2012. 68, 115

[260] A. G. Madian, H. E. Wheeler, R. B. Jones, and M. E. Dolan. Relating human genetic variation to variation in drug responses. *Trends in Genetics*, 28(10):487–495, Jan. 2012. 2, 63

[261] E. Maguire, P. Rocca-Serra, S.-A. Sansone, and M. Chen. Redesigning the Sequence Logo with Glyph-based Approaches to Aid Interpretation. In *Eurographics Conference on Visualization*, pages 1–5, Apr. 2014. 50

[262] M. L. Maitland, A. DiRienzo, and M. J. Ratain. Interpreting Disparate Responses to Cancer Therapy: The Role of Human Population Genetics. *Journal of clinical oncology*, 24(14): 2151–2157, Sept. 2016. 80

[263] T. A. Manolio. Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics*, 14(8):549–558, July 2013. 8, 10

[264] C. Marcotte and E. M. Marcotte. Predicting functional linkages from gene fusions with confidence. *Applied bioinformatics*, 2002. 18

[265] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science*, 285(5428):751–753, July 1999. 18

[266] K. Mardia, J. Kent, and J. M. Bibby. Multivariate analysis. 111

[267] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE*, 6(12): e28766–17, Dec. 2011. 2, 32, 43, 47, 78

[268] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE*, 6(12): e28766, Dec. 2011. 25, 28, 29, 36, 42, 43, 49

[269] D. S. Marks, T. A. Hopf, and C. Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080, Nov. 2012. 27

[270] J. A. Marsh and S. A. Teichmann. Structure, dynamics, assembly, and evolution of protein complexes. *Annual review of biochemistry*, 2015. 2, 17, 35

[271] V. Martınez, C. Navarro, C. Cano, and A. Blanco. Network-based drug-disease relation prioritization using ProphNet. 2013. 23

[272] V. Martínez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. *Artificial Intelligence in Medicine*, 63(1):41–49, Jan. 2015. 3, 23, 103, 105

[273] D. G. McArt and S.-D. Zhang. Identification of Candidate Small-Molecule Therapeutics to Cancer by Gene-Signature Perturbation in Connectivity Mapping. *PLoS ONE*, 6(1):e16382, Jan. 2011. 23, 83

[274] W. McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science ...*, 2010. 115

[275] D. T. McLachlin and S. D. Dunn. Disulfide linkage of the b and delta subunits does not affect the function of the Escherichia coli ATP synthase. *Biochemistry*, 39(12):3486–3490, Mar. 2000. 49

[276] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, June 2016. 115

[277] J. L. Medina-Franco, M. A. Giulianotti, G. S. Welmaker, and R. A. Houghten. Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discovery Today*, 18(9-10): 495–501, May 2013. 13, 19

[278] A. Melnikov, P. Rogov, L. Wang, A. Gnirke, and T. S. Mikkelsen. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Research*, 42(14):–e112, 2014. 81

[279] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347 (6224):1257601–1257601, Feb. 2015. 3, 19, 20, 24, 128

[280] R. Méndez, R. Leplae, L. De Maria, and S. J. Wodak. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins-Structure Function and Bioinformatics*, 52(1):51–67, July 2003. 47

[281] C. Merlin, G. Gardiner, S. Durand, and M. Masters. The Escherichia coli metD locus encodes an ABC transporter which includes Abc (MetN), YaeE (MetI), and YaeC (MetQ). *Journal of Bacteriology*, 184(19):5513–5517, Oct. 2002. 56

[282] L. Mette, K. Mitropoulos, A. Vozikis, and G. P. Patrinos. Pharmacogenomics and public health: implementing 'populationalized' medicine. *Pharmacogenomics*, 13(7):803–813, May 2012. 2, 63

[283] M. Michel, S. Hayat, M. J. Skwark, C. Sander, D. S. Marks, and A. Elofsson. PconsFold: improved contact predictions improve protein models. *Bioinformatics (Oxford, England)*, 30 (17):i482–8, Sept. 2014. 25

[284] S. Mignani, S. Huber, H. Tomás, J. Rodrigues, and J.-P. Majoral. Why and how have drug discovery strategies in pharma changed? What are the new mindsets? *Drug Discovery Today*, 21(2):239–249, Feb. 2016. 12, 13

[285] W. Miller, D. I. Drautz, A. Ratan, B. Pusey, J. Qi, A. M. Lesk, L. P. Tomsho, M. D. Packard, F. Zhao, A. Sher, A. Tikhonov, B. Raney, N. Patterson, K. Lindblad-Toh, E. S. Lander, J. R. Knight, G. P. Irzyk, K. M. Fredrikson, T. T. Harkins, S. Sheridan, T. Pringle, and S. C. Schuster. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, 456(7220): 387–390, Nov. 2008. 1

[286] E. V. Minikel, S. M. Vallabh, M. Lek, K. Estrada, K. E. Samocha, J. F. Sathirapongsasuti, C. Y. McLean, J. Y. Tung, Linda P. C. Yu, P. Gambetti, J. Blevins, S. Zhang, Y. Cohen, et al. Quantifying prion disease penetrance using large population control cohorts. *Science Translational Medicine*, 8(322):322ra9–322ra9, Jan. 2016. 8, 23

[287] C. Mizzi, B. Peters, C. Mitropoulou, K. Mitropoulos, T. Katsila, M. R. Agarwal, R. H. N. van Schaik, R. Drmanac, J. Borg, and G. P. Patrinos. Personalized pharmacogenomics profiling using whole-genome sequencing. *Pharmacogenomics*, 15(9):1223–1234, July 2014. 63, 186

[288] H. Möhler, J. M. Fritschy, and U. Rudolph. A new benzodiazepine pharmacology. *The Journal of pharmacology and experimental therapeutics*, 300(1):2–8, Jan. 2002. 80

[289] S. J. Mojzsis, G. Arrhenius, K. D. McKeegan, T. M. Harrison, A. P. Nutman, and C. R. Friend. Evidence for life on Earth before 3,800 million years ago. *Nature*, 384(6604):55–59, Nov. 1996. 6

[290] A. Moore, S. Collins, D. Carroll, H. McQuay, and J. Edwards. Single dose paracetamol (acetaminophen), with and without codeine, for postoperative pain. *The Cochrane database of systematic reviews*, Sept. 1996. 74

[291] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49):E1293–301, Dec. 2011. 27, 28, 36, 42, 43, 47

[292] C. Morrison. Fresh from the biotech pipeline - 2014. *Nature Biotechnology*, 33(2):125–128, Feb. 2015. 20

[293] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, July 2008. 6

[294] R. Mosca, A. Ceol, and P. Aloy. Interactome3D: adding structural details to protein networks. *Nature methods*, 10(1):47–53, Dec. 2012. 35

[295] D. Mosshammer, E. Schaeffeler, M. Schwab, and K. Moerike. Mechanisms and assessment of statin-related muscular adverse effects. *British Journal of Clinical Pharmacology*, 78(3): 454–466, Sept. 2014. 69

[296] A. A. Motsinger-Reif, E. Jorgenson, M. V. Relling, D. L. Kroetz, R. Weinshilboum, N. J. Cox, and D. M. Roden. Genome-Wide Association Studies in Pharmacogenomics: Successes and Lessons. *Pharmacogenetics and genomics*, 23(8):383–394, July 2013. 3, 64

[297] Y. Mou, P.-S. Huang, L. M. Thomas, and S. L. Mayo. Using Molecular Dynamics Simulations as an Aid in the Prediction of Domain Swapping of Computationally Designed Protein Variants. *Journal of Molecular Biology*, 427(16):2697–2706, Aug. 2015. 108, 109

[298] S. C. Mueller, C. Backes, O. V. Kalinina, B. Meder, D. Stöckel, H.-P. Lenhof, E. Meese, and A. Keller. BALL-SNP: combining genetic and structural information to identify candidate non-synonymous single nucleotide polymorphisms. *Genome medicine*, 7(1):65, 2015. 108

[299] S. Müller, A. Chaikuad, N. S. Gray, and S. Knapp. The ins and outs of selective kinase inhibitor development. *Nature chemical biology*, 11(11):818–821, Nov. 2015. 120, 124

[300] S. Navlakha and C. Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics (Oxford, England)*, 26(8):1057–1063, Apr. 2010. 24

[301] E. Neher. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences*, (91):98–102, 1994. 26

[302] M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zöllner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104, July 2012. 14, 186

[303] P. C. Ng and S. Henikoff. SIFT: predicting amino acid changes that affect protein function. 2003. 68, 115

[304] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261): 272–276, Sept. 2009. 5

[305] D.-T. Nguyen, S. Mathias, C. Bologa, S. Brunak, N. Fernandez, A. Gaulton, A. Hersey, J. Holmes, L. J. Jensen, A. Karlsson, G. Liu, A. Ma'ayan, G. Mandava, S. Mani, S. Mehta, J. Overington, J. Patel, A. D. Rouillard, S. Schürer, T. Sheils, A. Simeonov, L. A. Sklar, N. Southall, O. Ursu, D. Vidovic, A. Waller, J. Yang, A. Jadhav, T. I. Oprea, and R. Guha. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Research*, 45(D1):gkw1072–D1002, Nov. 2016. 84

[306] J. M. Nicoludis, S.-Y. Lau, C. P. I. Schärfe, D. S. Marks, W. A. Weihofen, and R. Gaudet. Structure and Sequence Analyses of Clustered Protocadherins Reveal Antiparallel Interactions that Mediate Homophilic Specificity. *Structure*, 23(11):2087–2098, Nov. 2015. 61

[307] J. M. Nicoludis, B. E. Vogt, A. G. Green, C. P. Schärfe, D. S. Marks, and R. Gaudet. Antiparallel protocadherin homodimers use distinct affinity- and specificity-mediating regions in cadherin repeats 1-4. *eLife*, 5, July 2016. 61

[308] M. Nilges. A calculation strategy for the structure determination of symmetric demers by 1H NMR. *Proteins-Structure Function and Bioinformatics*, 17(3):297–309, Nov. 1993. 62

[309] E. Nogales and S. H. W. Scheres. Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Molecular cell*, 58(4):677–689, May 2015. 18

[310] J. P. Noonan, G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Pääbo, J. K. Pritchard, and E. M. Rubin. Sequencing and analysis of Neanderthal genomic DNA. *Science*, 314(5802):1113–1118, Nov. 2006. 1

[311] P. H. O'Donnell and M. E. Dolan. Cancer Pharmacoethnicity: Ethnic Differences in Susceptibility to the Effects of Chemotherapy. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 15(15):4806–4814, 2009. 2, 63

[312] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, Jan. 1999. 86

[313] K. Oosawa and M. Simon. Analysis of mutations in the transmembrane region of the aspartate chemoreceptor in Escherichia coli. *Proceedings of the National Academy of Sciences*, (83): 6930–6934, 1986. 26

[314] T. I. Oprea, C. G. Bologa, S. Brunak, A. Campbell, G. N. Gan, A. Gaulton, S. M. Gomez, R. Guha, A. Hersey, J. Holmes, A. Jadhav, L. J. Jensen, G. L. Johnson, A. Karlson, A. R. Leach, et al. Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery*, 17(5):317–332, May 2018. 21, 84

[315] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. L. Brinkman, F. Brinkman, G. Cesareni, A. Chatr-aryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. E. W. Hancock, R. Hancock, L. I. Hannick, I. Jurisica, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature methods*, 9(4):345–350, Apr. 2012. 19

[316] V. E. Ortega and D. A. Meyers. Pharmacogenetics: implications of race and ethnicity on defining genetic profiles for personalized medicine. *The Journal of allergy and clinical immunology*, 133(1):16–26, Jan. 2014. 81

[317] S. Ovchinnikov, H. Kamisetty, D. Baker, and B. Roux. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, 3:e02030, May 2014. 46, 59

[318] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins. How many drug targets are there? *Nature Reviews Drug Discovery*, 5(12):993–996, Dec. 2006. 21

[319] R. P. Owen, L. Gong, H. Sagreiya, T. E. Klein, and R. B. Altman. VKORC1 pharmacogenomics summary. *Pharmacogenetics and genomics*, 20(10):642–644, Oct. 2010. 76

[320] N. Pakseresht, B. Alako, C. Amid, A. Cerdeño-Tárraga, I. Cleland, R. Gibson, N. Goodgame, T. Gur, M. Jang, S. Kay, R. Leinonen, W. Li, X. Liu, R. Lopez, et al. Assembly information

services in the European Nucleotide Archive. *Nucleic Acids Research*, 42(Database issue): D38–43, Jan. 2014. 41, 48

[321] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–1415, Dec. 2008. 6

[322] G. V. Paolini, R. H. B. Shapland, W. P. van Hoorn, J. S. Mason, and A. L. Hopkins. Global mapping of pharmacological space. *Nature Biotechnology*, 24(7):805–815, July 2006. 21

[323] J. Park, D.-S. Lee, N. A. Christakis, and A.-L. Barabási. The impact of cellular networks on disease comorbidity. *Molecular Systems Biology*, 5(1):262, 2009. 20, 86

[324] M.-S. Park, A. L. Dessal, A. V. Smrcka, and H. A. Stern. Evaluating docking methods for prediction of binding affinities of small molecules to the G protein betagamma subunits. *Journal of Chemical Information and Modeling*, 49(2):437–443, Feb. 2009. 118

[325] S. M. Paul and F. Lewis-Hall. Drugs in search of diseases. *Science Translational Medicine*, 5 (186):186fs18–186fs18, May 2013. 21

[326] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, Nov. 2010. 12, 13

[327] Paul D Lyne, Michelle L Lamb, and Jamal C Saeh. Accurate Prediction of the Relative Potencies of Members of a Series of Kinase Inhibitors Using Molecular Docking and MM-GBSA Scoring. *Journal of Medicinal Chemistry*, 49(16):4805–4808, July 2006. 118, 123, 125

[328] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, 14(9):609–614, Sept. 2001. 19, 36

[329] F. Pazos and A. Valencia. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins-Structure Function and Bioinformatics*, 47(2):219–227, May 2002.

[330] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, 271(4):511–523, Aug. 1997. 36

[331] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011. 94

[332] E. Pennisi. Genomics. ENCODE project writes eulogy for junk DNA. *Science*, 337(6099): 1159–1161, Sept. 2012. 5

[333] N. Perdigão, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans, and S. I. O'Donoghue. Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences of the United States of America*, 112(52):15898–15903, Dec. 2015. 25

[334] F. Perez and B. E. Granger. IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, 9(3):21–29, 2007. 84, 113, 115

[335] J. R. Perkins, I. Diboun, B. H. Dessailly, J. G. Lees, and C. Orengo. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure*, 18(10):1233–1243, Oct. 2010. 35

[336] E. Perola, W. P. Walters, and P. S. Charifson. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins-Structure Function and Bioinformatics*, 56(2):235–249, Aug. 2004. 21

[337] Peter A Kollman, Irina Massova, Carolina Reyes, Bernd Kuhn, Shuanghong Huo, Lillian Chong, Matthew Lee, Taisung Lee, Yong Duan, Wei Wang, Oreola Donini, Piotr Cieplak, Jaysharee Srinivasan, David A Case, and Thomas E Cheatham, III. Calculating Structures and Free Energies of Complex Molecules:âĂĽCombining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research*, 33(12):889–897, Oct. 2000. 17

[338] M. B. Peters, K. Raha, and K. M. Merz. Quantum mechanics in structure-based drug design. *Current opinion in drug discovery & development*, 9(3):370–379, May 2006. 17

[339] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, Oct. 2004. 49

[340] PharmGKB. Drug Labels. URL `https://www.pharmgkb.org/view/drug-labels.do`. Accessed: Jan 2018. 3

[341] M. Pirmohamed. Personalized Pharmacogenomics: Predicting Efficacy and Adverse Drug Reactions. *Annual Review of Genomics and Human Genetics, Vol 15*, 15(1):349–370, 2014. 2, 63

[342] D. Plewczynski, M. Łaźniewski, R. Augustyniak, and K. Ginalski. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of Computational Chemistry*, 32(4):742–755, Mar. 2011. 21

[343] H. N. Poinar. Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. *Science*, 311(5759):392–394, Jan. 2006. 1

[344] N. Pratanwanich and P. Lio. Pathway-based Bayesian inference of drug-disease interactions. *Mol. BioSyst.*, 10(6):1538–1548, 2014. 23

[345] C. J. Presley, D. Tang, P. R. Soulos, A. C. Chiang, J. A. Longtine, K. B. Adelson, R. S. Herbst, W. Zhu, N. C. Nussbaum, R. A. Sorg, V. Agarwala, A. P. Abernethy, and C. P.

Gross. Association of Broad-Based Genomic Sequencing With Survival Among Patients With Advanced Non-Small Cell Lung Cancer in the Community Oncology Setting. *JAMA*, 320(5): 469–477, Aug. 2018. 128

[346] S. V. Rajagopala, P. Sikorski, A. Kumar, R. Mosca, J. Vlasblom, R. Arnold, J. Franca-Koh, S. B. Pakala, S. Phanse, A. Ceol, R. Häuser, G. Siszler, S. Wuchty, A. Emili, M. Babu, P. Aloy, R. Pieper, and P. Uetz. The binary protein-protein interaction landscape of Escherichia coli. *Nature Biotechnology*, 32(3):285–290, Mar. 2014. 35, 47, 48, 55

[347] I. Rajman. PK/PD modelling and simulations: utility in drug development. *Drug Discovery Today*, 13(7-8):341–346, Apr. 2008. 12

[348] T. N. Raju. *The Nobel chronicles. 1988: James Whyte Black, (b 1924), Gertrude Elion (1918-99), and George H Hitchings (1905-98).*, volume 355. University of Illinois, Chicago, USA., Mar. 2000. 21

[349] M. Rastegar-Mojarad, Z. Ye, J. M. Kolesar, S. J. Hebbring, and S. M. Lin. Opportunities for drug repositioning from phenome-wide association studies. *Nature Biotechnology*, 33(4): 342–345, Apr. 2015. 23, 83, 103

[350] V. K. Rastogi and M. E. Girvin. Structural changes linked to proton translocation by subunit c of the ATP synthase. *Nature*, 402(6759):263–268, Nov. 1999. 58

[351] A. Regev, S. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, et al. The Human Cell Atlas. *bioRxiv*, page 121202, May 2017. 5, 105

[352] M. V. Relling and W. E. Evans. Pharmacogenomics in the clinic. *Nature*, 526(7573):343–350, Oct. 2015. 82

[353] M. Remmert, A. Biegert, A. Hauser, and J. Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 2012. 38

[354] J. P. G. L. M. Rodrigues, A. S. J. Melquiond, E. Karaca, M. Trellet, M. van Dijk, G. C. P. van Zundert, C. Schmitz, S. J. de Vries, A. Bordogna, L. Bonati, P. L. Kastritis, and A. M. J. J. Bonvin. Defining the limits of homology modeling in information-driven protein docking. *Proteins-Structure Function and Bioinformatics*, 81(12):2119–2128, Oct. 2013. 36

[355] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, et al. A Proteome-Scale Map of the Human Interactome Network. *Cell*, 159(5):1212–1226, Nov. 2014. 18, 19

[356] A. Ronacher. Flask. `https://github.com/pallets/flask`, . Version: 0.12.2. 49

[357] A. Ronacher. Jinja2 Template Engine. `https://github.com/pallets/jinja`, . Version: 2.9.6. 50

[358] A. D. Roses. Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nature Reviews Genetics*, 5(9):645–656, Sept. 2004. 14

[359] A. Ruangprasert, T. Maehigashi, S. J. Miles, N. Giridharan, J. X. Liu, and C. M. Dunham. Mechanisms of Toxin Inhibition and Transcriptional Repression by Escherichia coli DinJ-YafQ. *The Journal of biological chemistry*, 289(30):20559–20569, 2014. 58

[360] C. Rubio-Perez, D. Tamborero, M. P. Schroeder, A. A. Antolín, J. Deu-Pons, C. Perez-Llamas, J. Mestres, A. Gonzalez-Perez, and N. Lopez-Bigas. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell*, 27(3):382–396, Mar. 2015. 70, 107, 108

[361] D. G. Rudmann. On-target and Off-target-based Toxicologic Effects. *Toxicologic pathology*, 41(2):0192623312464311–314, Oct. 2012. 13

[362] D. P. Ryan and J. M. Matthews. Protein–protein interactions in human disease. *Current Opinion in Structural Biology*, 15(4):441–446, Aug. 2005. 2

[363] U. Ryde and P. Söderhjelm. Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods. *Chemical reviews*, 116(9):5520–5566, May 2016. 16, 124, 125

[364] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue): D449–51, Jan. 2004. 19

[365] P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon, and V. Mooser. Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, 30(4):317–320, Apr. 2012. 23, 83

[366] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea, and J. P. Overington. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1):19–34, Dec. 2016. 80, 84

[367] M. S. Saporito, C. A. Lipinski, and A. G. Reaume. Phenotypic In VivoScreening to Identify New, Unpredicted Indications for Existing Drugs and Drug Candidates. In *Drug Repositioning*, pages 253–290. John Wiley & Sons, Inc., Hoboken, NJ, USA, Apr. 2012. 22

[368] G. M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, and W. Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3):221–234, Mar. 2013. 49

[369] R. Sawada, H. Iwata, S. Mizutani, and Y. Yamanishi. Target-Based Drug Repositioning Using Large-Scale Chemical-Protein Interactome Data. *Journal of Chemical Information and Modeling*, 55(12):2717–2730, Dec. 2015. 3, 24, 25, 103, 105

[370] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3):191–200, Mar. 2012. 12

[371] C. Schärfe, R. Tremmel, M. Schwab, O. Kohlbacher, and D. S. Marks. Genetic variation in human drug-related genes. *Genome medicine*, 9(1):117, Dec. 2017. 63

[372] L. Schneider, D. Stöckel, T. Kehl, A. Gerasch, N. Ludwig, P. Leidinger, H. Huwer, S. Tenzer, O. Kohlbacher, A. Hildebrandt, M. Kaufmann, M. Gessler, A. Keller, E. Meese, N. Graf, and H.-P. Lenhof. DrugTargetInspector: An assistance tool for patient treatment stratification. *International journal of cancer. Journal international du cancer*, 138(7):1765–1776, Apr. 2016. 125

[373] D. Schneidman-Duhovny, A. Rossi, A. Avila-Sakar, S. J. Kim, J. Velázquez-Muriel, P. Strop, H. Liang, K. A. Krukenberg, M. Liao, H. M. Kim, S. Sobhanifar, V. Dötsch, A. Rajpal, J. Pons, D. A. Agard, Y. Cheng, and A. Sali. A method for integrative structure determination of protein-protein complexes. *Bioinformatics (Oxford, England)*, 28(24):3282–3289, Dec. 2012. 36

[374] N. J. Schork. Time for one-person trials. *Nature*, 520(7549):609–611, 2015. 13, 63

[375] L. M. Schriml, C. Arze, S. Nadendla, Y. W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946, Dec. 2011. 86

[376] Schrödinger, LLC. Schrödinger Knowledgebase: How accurate are the XP GlideScores?, . URL `https://www.schrodinger.com/kb/793`. Accessed: January 2018. 125

[377] Schrödinger, LLC. Schrödinger Knowledgebase: GlideScore/Docking Score does not correlate with my known activities. What is wrong?, . URL `https://www.schrodinger.com/kb/144`. Accessed: January 2018. 123

[378] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. Nov. 2015. xv

[379] B. Schulenberg, R. Aggeler, J. Murray, and R. A. Capaldi. The gammaepsilon-c subunit interface in the ATP synthase of Escherichia coli. cross-linking of the epsilon subunit to the c subunit ring does not impair enzyme function, that of gamma to c subunits leads to uncoupling. *The Journal of biological chemistry*, 274(48):34233–34237, Nov. 1999. 49

[380] M. Schwaederle, B. A. Parker, R. B. Schwab, P. T. Fanta, S. G. Boles, G. A. Daniels, L. A. Bazhenova, R. Subramanian, A. C. Coutinho, H. Ojeda-Fournier, B. Datnow, N. J. Webster, S. M. Lippman, and R. Kurzrock. Molecular tumor board: the University of California-San Diego Moores Cancer Center experience. *The Oncologist*, 19(6):631–636, June 2014. 4, 107, 125

[381] B. E. Schwem and R. H. Fillingame. Cross-linking between helices within subunit a of Escherichia coli ATP synthase defines the transmembrane packing of a four-helix bundle. *The Journal of biological chemistry*, 281(49):37861–37867, Dec. 2006. 58, 59

[382] W. E. C. o. t. Selection and U. o. E. Medicines. WHO Model List of Essential Medicines. Technical report, Nov. 2015. 72

[383] P. Sfriso, M. Duran-Frigola, R. Mosca, A. Emperador, P. Aloy, and M. Orozco. Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins. *Structure*, Dec. 2015. 127

[384] A. Shahandeh, D. M. Johnstone, J. R. Atkins, J.-M. Sontag, M. Heidari, N. Daneshi, E. Freeman-Acquah, and E. A. Milward. Advantages of Array-Based Technologies for Pre-Emptive Pharmacogenomics Testing. *Microarrays (Basel, Switzerland)*, 5(2):12, May 2016. 82

[385] J. L. Sharman, H. E. Benson, A. J. Pawson, V. Lukito, C. P. Mpamhanga, V. Bombail, A. P. Davenport, J. A. Peters, M. Spedding, A. J. Harmar, and NC-IUPHAR. IUPHAR-DB: updated database content and new features. *Nucleic Acids Research*, 41(D1):D1083–D1088, Jan. 2013. 87

[386] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science*, 330(6002):341–346, Oct. 2010. 25

[387] G. Shen, W. Cui, H. Zhang, F. Zhou, W. Huang, Q. Liu, Y. Yang, S. Li, G. R. Bowman, J. E. Sadler, M. L. Gross, and W. Li. Warfarin traps human vitamin K epoxide reductase in an intermediate state during electron transfer. *Nature Structural & Molecular Biology*, 24(1): 69–76, Jan. 2017. 78

[388] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, Jan. 2001. 86

[389] D. Shigemizu, A. Fujimoto, S. Akiyama, T. Abe, K. Nakano, K. A. Boroevich, Y. Yamamoto, M. Furuta, M. Kubo, H. Nakagawa, and T. Tsunoda. A practical method to detect SNVs and indels from whole genome and exome sequencing data. *Scientific Reports*, 3:2161, 2013. 81

[390] I. N. Shindyalov, N. A. Kolchanov, and C. Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering*, (7):349–358, 1994. 26

[391] D. Shivakumar, J. Williams, Y. Wu, W. Damm, J. Shelley, and W. Sherman. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *Journal of Chemical Theory and Computation*, 6(5):1509–1519, Apr. 2010. 109

[392] A. D. Siddiqui and B. Piperdi. KRAS mutation in colon cancer: a marker of resistance to EGFR-I therapy. *Annals of surgical oncology*, 17(4):1168–1176, Apr. 2010. 107

[393] M. S. Singh, P. A. Francis, and M. Michael. Tamoxifen, cytochrome P450 genes and breast cancer clinical outcomes. *Breast (Edinburgh, Scotland)*, 20(2):111–118, Apr. 2011. 14

[394] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte. Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data. *Science Translational Medicine*, 3(96):96ra77–96ra77, Aug. 2011. 23, 83

[395] J. M. Skerker, B. S. Perchuk, A. Siryaporn, E. A. Lubin, O. Ashenberg, M. Goulian, and M. T. Laub. Rewiring the specificity of two-component signal transduction systems. *Cell*, 133(6):1043–1054, June 2008. 41

[396] D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, J. Baselga, and L. Norton. Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. *New England Journal of Medicine*, 344(11):783–792, Mar. 2001. 64

[397] N. Small. Py2neo. `http://py2neo.org/v3`, 2016. Version: 3.1.2. 89

[398] D. Smedley, S. Haider, S. Durinck, L. Pandini, P. Provero, J. Allen, O. Arnaiz, M. H. Awedh, R. Baldock, G. Barbiera, P. Bardou, T. Beck, A. Blake, M. Bonierbale, A. J. Brookes, G. Bucci, I. Buetti, S. Burge, C. Cabau, J. W. Carlson, C. Chelala, C. Chrysostomou, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):W589–98, July 2015. 88

[399] R. D. Smith, J. James B Dunbar, P. M.-U. Ung, E. X. Esposito, C.-Y. Yang, S. Wang, and H. A. Carlson. CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. *Journal of Chemical Information and Modeling*, 51(9):2115–2131, Aug. 2011. 21

[400] R. D. Smith, K. L. Damm-Ganamet, J. James B Dunbar, A. Ahmed, K. Chinnaswamy, J. E. Delproposto, G. M. Kubish, C. E. Tinberg, S. D. Khare, J. Dou, L. Doyle, J. A. Stuckey, D. Baker, and H. A. Carlson. CSAR Benchmark Exercise 2013: Evaluation of Results from a Combined Computational Protein Design, Docking, and Scoring/Ranking Challenge. *Journal of chemical ...*, Oct. 2015. 21

[401] G. G. Sofowora, V. Dishy, M. Muszkat, H. G. Xie, R. B. Kim, P. A. Harris, H. C. Prasad, D. W. Byrne, U. B. Nair, A. J. J. Wood, and C. M. Stein. A common β1-adrenergic receptor polymorphism (Arg389Gly) affects blood pressure response to β-blockade. *Clinical Pharmacology & Therapeutics*, 73(4):366–371, Apr. 2003. 64

[402] R. A. Soo, L. Z. Wang, S. S. Ng, P. Y. Chong, W. P. Yong, S. C. Lee, J. J. Liu, T. B. Choo, L. S. Tham, H. S. Lee, B. C. Goh, and R. Soong. Distribution of gemcitabine pathway genotypes in ethnic Asians and their association with outcome in non-small cell lung cancer patients. *Lung cancer (Amsterdam, Netherlands)*, 63(1):121–127, Jan. 2009. 76

[403] M. E. Sowa, E. J. Bennett, S. P. Gygi, and J. W. Harper. Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138(2):389–403, July 2009. 18

[404] R. R. Stein, D. S. Marks, and C. Sander. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS computational biology*, 11(7):e1004182, July 2015. 42

[405] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson. Big Data: Astronomical or Genomical? *PLoS Biol*, 13(7):e1002195, July 2015. 1

[406] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239): 719–724, Apr. 2009. 70, 79

[407] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, Jan. 2007. 104

[408] M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19):6959–6964, May 2008. 19

[409] H. G. Svensson, W. J. Wedemeyer, J. L. Ekstrom, D. R. Callender, T. Kortemme, D. E. Kim, U. Sjöbring, and D. Baker. Contributions of amino acid side chains to the kinetics and thermodynamics of the bivalent binding of protein L to Ig kappa light chain. *Biochemistry*, 43(9):2445–2457, Mar. 2004. 36

[410] A. T. Swagger. Swagger. `https://swagger.io`. Accessed: January 2018. 114

[411] M. Swain. ChemSpiPy. `https://chemspipy.readthedocs.io`, 2018. Version: 1.0.4. 88

[412] D. C. Swinney and J. Anthony. How were new medicines discovered? *Nature Reviews Drug Discovery*, 10(7):507–519, July 2011. 11, 12, 22

[413] W. R. Taylor and K. Hatrick. Compensating changes in protein multiple sequence alignments. *Protein engineering, design & selection : PEDS*, (7):341–348, 1994. 26

[414] D. J. Thomas, G. Casari, and C. Sander. The prediction of protein contacts from multiple sequence alignments. *Protein Engineering*, (9):941–948, 1996. 26

[415] L. Tolosi and T. Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics (Oxford, England)*, 27(14):1986–1994, July 2011. 104

[416] P. Tompa, N. E. Davey, T. J. Gibson, and M. M. Babu. A million peptide motifs for the molecular biologist. *Molecular cell*, 55(2):161–169, July 2014. 35

[417] A. Torkamani and N. J. Schork. Background gene expression networks significantly enhance drug response prediction by transcriptional profiling. *The Pharmacogenomics Journal*, 12(5): 446–452, Oct. 2012. 23

[418] S. Tornow and H. W. Mewes. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Research*, 31(21):6283–6289, Nov. 2003. 19

[419] A. Toth-Petroczy, P. Palmedo, J. Ingraham, T. A. Hopf, B. Berger, C. Sander, and D. S. Marks. Structured States of Disordered Proteins from Genomic Sequences. *Cell*, 167(1): 158–170.e12, Sept. 2016. 25, 29

[420] A. Tourancheau, G. Margaillan, M. Rouleau, I. Gilbert, L. Villeneuve, E. Lévesque, A. Droit, and C. Guillemette. Unravelling the transcriptomic landscape of the major phase II UDP-glucuronosyltransferase drug metabolizing pathway using targeted RNA sequencing. *The Pharmacogenomics Journal*, 16(1):60–70, Feb. 2016. 81

[421] D. K. Treiber and N. P. Shah. Ins and outs of kinase DFG motifs. *Chemistry & Biology*, 20 (6):745–746, June 2013. 124

[422] S. K. Tripathi, R. Muttineni, and S. K. Singh. Extra precision docking, free energy calculation and molecular dynamics simulation studies of CDK2 inhibitors. *Journal of theoretical biology*, 334:87–100, Oct. 2013. 111

[423] M. d. J. Trovoada, M. Martins, R. Ben Mansour, M. d. R. Sambo, A. B. Fernandes, L. Antunes Gonçalves, A. Borja, R. Moya, P. Almeida, J. Costa, I. Marques, M. P. Macedo, A. Coutinho, D. L. Narum, and C. Penha-Gonçalves. NOS2 variants reveal a dual genetic control of nitric oxide levels, susceptibility to Plasmodium infection, and cerebral malaria. *Infection and Immunity*, 82(3):1287–1295, Mar. 2014. 78

[424] A.-M. Tsimberidou. Initiative for Molecular Profiling and Advanced Cancer Therapy and challenges in the implementation of precision medicine. *Current problems in cancer*, 41(3): 176–181, June 2017. 125

[425] A.-M. Tsimberidou, N. G. Iskander, D. S. Hong, J. J. Wheler, G. S. Falchook, S. Fu, S. Piha-Paul, A. Naing, F. Janku, R. Luthra, Y. Ye, S. Wen, D. Berry, and R. Kurzrock. Personalized medicine in a phase I clinical trials program: the MD Anderson Cancer Center initiative. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 18(22):6373–6383, Nov. 2012. 125

[426] T. Tuntland, B. Ethell, T. Kosaka, F. Blasco, R. X. Zang, M. Jain, T. Gould, and K. Hoffmaster. Implementation of pharmacokinetic and pharmacodynamic strategies in early research phases of drug discovery and development at Novartis Institute of Biomedical Research. *Frontiers in pharmacology*, 5:174, 2014. 12

[427] UK10K Consortium, K. Walter, J. L. Min, J. Huang, L. Crooks, Y. Memari, S. McCarthy, J. R. B. Perry, C. Xu, M. Futema, D. Lawson, V. Iotchkova, S. Schiffels, A. E. Hendricks, et al. The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571): 82–90, Oct. 2015. 3

[428] A. Vangone, J. P. G. L. M. Rodrigues, L. C. Xue, G. C. P. van Zundert, C. Geng, Z. Kurkcuoglu, M. Nellen, S. Narasimhan, E. Karaca, M. van Dijk, A. S. J. Melquiond, K. M. Visscher, M. Trellet, P. L. Kastritis, and A. M. J. J. Bonvin. Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. *Proteins-Structure Function and Bioinformatics*, 85(3):417–423, Mar. 2017. 51

[429] S. Velankar, J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M.-J. Martin, and G. J. Kleywegt. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research*, 41(Database issue):D483–9, Jan. 2013. 87

[430] J. Velázquez-Muriel, K. Lasker, D. Russel, J. Phillips, B. M. Webb, D. Schneidman-Duhovny, and A. Sali. Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images. *Proceedings of the National Academy of Sciences of the United States of America*, 109(46):18821–18826, Nov. 2012. 36

[431] M. Vidal, M. E. Cusick, and A.-L. Barabási. Interactome Networks and Human Disease. *Cell*, 144(6):986–998, Mar. 2011. 19

[432] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American journal of human genetics*, 101(1):5–22, July 2017. 10

[433] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, Mar. 2013. 107

[434] F. R. Vogenberg, C. Isaacson Barash, and M. Pursel. Personalized medicine. *Pharmacy and Therapeutics*, 35(10):560–576, Oct. 2010. 128

[435] J. von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne, and R. Preissner. PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Research*, 39(Database issue):D1060–6, Jan. 2011. 3, 23

[436] J. E. Walker. The ATP synthase: the understood, the uncertain and the unknown. *Biochemical Society Transactions*, 41(1):1–16, Feb. 2013. 58

[437] S. D. Walter. Number needed to treat (NNT): estimation of a measure of clinical benefit. *Statistics in Medicine*, 20(24):3947–3962, Dec. 2001. 74

[438] Y. Wang, S. Chen, N. Deng, Y. Wang, and Y. Wang. Drug Repositioning by Kernel-Based Integration of Molecular Structure, Molecular Activity, and Phenotype Data. *PLoS ONE*, 8 (11):e78518, Nov. 2013. 24, 103, 105

[439] Z.-Y. Wang and H.-Y. Zhang. Rational drug repositioning by medical genetics. *Nature Biotechnology*, 31(12):1080–1082, Dec. 2013. 23, 83

[440] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(Web Server issue):W214–20, July 2010. 19

[441] M. J. Waring, J. Arrowsmith, A. R. Leach, P. D. Leeson, S. Mandrell, R. M. Owen, G. Pairaudeau, W. D. Pennie, S. D. Pickett, J. Wang, O. Wallace, and A. Weir. An analysis of

the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, 14(7):475–486, July 2015. 12, 13

[442] G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head. A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry*, 49(20):5912–5931, Oct. 2006. 118

[443] B. Webb, K. Lasker, J. Velázquez-Muriel, D. Schneidman-Duhovny, R. Pellarin, M. Bonomi, C. Greenberg, B. Raveh, E. Tjioe, D. Russel, and A. Sali. Modeling of proteins and their assemblies with the Integrative Modeling Platform. *Methods in molecular biology (Clifton, N.J.)*, 1091(Chapter 20):277–295, 2014. 36

[444] W.-Q. Wei, R. M. Cronin, H. Xu, T. A. Lasko, L. Bastarache, and J. C. Denny. Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):954–961, Sept. 2013. 87

[445] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1):67–72, Jan. 2009. 41

[446] P. K. Weiner and P. A. Kollman. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry*, 2(3):287–303, 1981. 14, 15

[447] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, Feb. 1984. 14

[448] C. Weinreb, A. J. Riesselman, J. B. Ingraham, T. Gross, C. Sander, and D. S. Marks. 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell*, 165(4):963–975, May 2016. 25, 29

[449] S. Weinreich. Orphanet: A European database for rare diseases. *Nederlands Tijdschrift voor Geneeskunde*, 152(9):518–519, Mar. 2008. 86

[450] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(Database issue):D1001–6, Jan. 2014. 87

[451] C. G. Wermuth. Selective optimization of side activities: another way for drug discovery. *Journal of Medicinal Chemistry*, 47(6):1303–1314, Mar. 2004. 21

[452] C. G. Wermuth. Selective optimization of side activities: the SOSA approach. *Drug Discovery Today*, 11(3-4):160–164, Feb. 2006. 21

[453] M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman, and T. E. Klein. Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, Oct. 2012. 3, 14, 64, 70, 73, 87, 108

[454] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. *Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids*, volume 118. American Chemical Society, Nov. 1996. 14

[455] A. J. Williams. *Chemspider: A Platform for Crowdsourced Collaboration to Curate Data Derived From Public Compound Databases*, volume 30 of *Ekins/Collaborative Computational*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2011. 88

[456] S. J. Wodak, J. Vlasblom, A. L. Turinsky, and S. Pu. Protein-protein interaction networks: the puzzling riches. *Current Opinion in Structural Biology*, 23(6):941–953, Dec. 2013. 18

[457] A. L.-A. Wong, H.-L. Yap, W.-L. Yeo, R. Soong, S. S. Ng, L. Z. Wang, M. T. Cordero, W. P. Yong, B. C. Goh, and S. C. Lee. Gemcitabine and platinum pathway pharmacogenetics in Asian breast cancer patients. *Cancer genomics & proteomics*, 8(5):255–259, Sept. 2011. 76

[458] World Health Organization. ATC - Structure and principles, 2009. URL `http://www.fhi.no/en/hn/drug/who-collaborating-centre-for-drug-statistics-methodology/`. 70

[459] G. E. B. Wright, B. Carleton, M. R. Hayden, and C. J. D. Ross. The global spectrum of protein-coding pharmacogenomic diversity. *The Pharmacogenomics Journal*, Oct. 2016. 14, 63, 79, 81, 186

[460] C. Wu, I. Macleod, and A. I. Su. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Research*, 41(Database issue):D561–5, Jan. 2013. 88

[461] J. Xin, A. Mark, C. Afrasiabi, G. Tsueng, M. Juchler, N. Gopal, G. S. Stupp, T. E. Putman, B. J. Ainscough, O. L. Griffith, A. Torkamani, P. L. Whetzel, C. J. Mungall, S. D. Mooney, A. I. Su, and C. Wu. High-performance web services for querying gene and variant annotation. *Genome Biology*, 17(1):91, May 2016. 87

[462] G. Yachdav, E. Kloppmann, L. Kajan, M. Hecht, T. Goldberg, T. Hamp, P. Hönigschmid, A. Schafferhans, M. Roos, M. Bernhofer, L. Richter, H. Ashkenazy, M. Punta, A. Schlessinger, Y. Bromberg, R. Schneider, G. Vriend, C. Sander, N. Ben-Tal, and B. Rost. PredictProtein–an open resource for online prediction of protein structural and functional features. *Nucleic Acids Research*, 42(Web Server issue):W337–43, July 2014. 25

[463] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics (Oxford, England)*, 24(13):i232–40, July 2008. 24

[464] C.-H. Yang, Y.-H. Cheng, L.-Y. Chuang, and H.-W. Chang. Drug-SNPing: an integrated drug-based, protein interaction-based tagSNP-based pharmacogenomics platform for SNP genotyping. *Bioinformatics (Oxford, England)*, 29(6):758–764, Mar. 2013. 108

[465] H. Yang, C. Qin, Y. H. Li, L. Tao, J. Zhou, C. Y. Yu, F. Xu, Z. Chen, F. Zhu, and Y. Z. Chen. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Research*, 44(D1):D1069–74, Jan. 2016. 87

[466] J. Yang, Z. Li, X. Fan, and Y. Cheng. Drug-Disease Association and Drug-Repositioning Predictions in Complex Diseases Using Causal Inference-Probabilistic Matrix Factorization. *Journal of Chemical Information and Modeling*, 54(9):2562–2569, Sept. 2014. 23

[467] T. A. Yap and P. Workman. Exploiting the cancer genome: strategies for the discovery and clinical development of targeted molecular therapeutics. *Annual review of pharmacology and toxicology*, 52:549–573, 2012. 124

[468] S. U. Yasuda, L. Zhang, and S. M. Huang. The Role of Ethnicity in Variability in Response to Drugs: Focus on Clinical Pharmacology Studies. *Clinical Pharmacology & Therapeutics*, 2008. 2, 63

[469] P. Yeh, H. Chen, J. Andrews, R. Naser, W. Pao, and L. Horn. DNA-Mutation Inventory to Refine and Enhance Cancer Treatment (DIRECT): A Catalog of Clinically Relevant Cancer Mutations to Enable Genome-Directed Anticancer Therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 19(7):1894–1901, Apr. 2013. 108

[470] M. A. Yıldırım, K. I. Goh, M. E. Cusick, and A. L. Barabasi. Drug—target network. *Nature*, 25(10):1119–1126, 2007. 10, 20, 21

[471] K. Yugi, H. Kubota, A. Hatano, and S. Kuroda. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. *Trends in Biotechnology*, 34(4):276–290, Apr. 2016. 128

[472] C.-H. Yun, K. E. Mengwasser, A. V. Toms, M. S. Woo, H. Greulich, K.-K. Wong, M. Meyerson, and M. J. Eck. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):2070–2075, Feb. 2008. 124

[473] S. Zahler, S. Tietze, F. Totzke, M. Kubbutat, L. Meijer, A. M. Vollmar, and J. Apostolakis. Inverse in silico screening for identification of kinase inhibitor targets. *Chemistry & Biology*, 14(11):1207–1214, Nov. 2007. 21

[474] U. M. Zanger and M. Schwab. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*, 138(1):103–141, Apr. 2013. 69

[475] Y. Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, June 2008. 25

[476] J. Zhao, K. Beyrakhova, Y. Liu, C. P. Alvarez, S. A. Bueler, L. Xu, C. Xu, M. T. Boniecki, V. Kanelis, Z.-Q. Luo, M. Cygler, and J. L. Rubinstein. Molecular basis for the binding and modulation of V-ATPase by a bacterial effector protein. *PLoS pathogens*, 13(6):e1006394, June 2017. 59

[477] S. Zhao and S. Li. A co-module approach for elucidating drug-disease associations and revealing their molecular basis. *Bioinformatics (Oxford, England)*, 28(7):955–961, 2012. 24

[478] Y. Zhao, S. Chen, C. Yoshioka, I. Baconguis, and E. Gouaux. Architecture of fully occupied GluA2 AMPA receptor-TARP complex elucidated by cryo-EM. *Nature*, July 2016. 35

[479] W. Zheng, N. Thorne, and J. C. McKew. Phenotypic screens as a renewed approach for drug discovery. *Drug Discovery Today*, 18(21-22):1067–1073, Nov. 2013. 22

[480] A. Zhou, A. Rohou, D. G. Schep, J. V. Bason, M. G. Montgomery, J. E. Walker, N. Grigorieff, and J. L. Rubinstein. Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-EM. *eLife*, 4:e10180, Oct. 2015. 59

[481] I. Zineh and M. A. Pacanowski. Pharmacogenomics in the Assessment of Therapeutic Risks versus Benefits: Inside the United States Food and Drug Administration. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 31(8):729–735, Aug. 2011. 13

[482] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–1198, Jan. 2012. 10

Appendix A

Abbreviations

Amino acids in a general context are not abbreviated. If we refer to a particular amino acid within a peptide or protein, standard abbreviations in 3-letter code are used with the residue position as suffix. Thus, an arbitrary serine at position 256 is abbreviated by Ser256. Within amino acid sequences, standard abbreviations in 1-letter code are used.

2D	*Two-dimensional*
3D	*Three-dimensional*
A	
AC	*Allele count*
ADME(T)	*Absorption, distribution, metabolism, excretion (, toxicity)*
AF	*Allele frequency*
AN	*Allele number*
AP-MS	*Affinity-purification combined with mass spectrometry*
APC	*Average Product Correction*
ATP	*Adenosine triphosphate*
B	
BALL	*Biochemical algorithms library*
BFS	*Breath-first search*
C	
CADD	*Computer-aided drug design*
CAP	*Cumulative Allele Probability*
CDS	*Coding sequence*
Cmap	*Connectivity map*
CNV	*Copy number variant*

cryo-EM	*Cryo-electron microscopy*
D	
DCA	*Direct coupling analysis*
DFG	*Asp-Phe-Gly motif in kinases*
DNA	*Deoxyribonucleic acid*
DO	*Disease Ontology*
DRP	*Drug risk probability*
E	
EC	*Evolutionary coupling*
EFO	*Experimental Factor Ontology*
F	
FDA	*U.S. Food and Drug Administration*
FN	*Frobenius norm*
FP	*False positive*
G	
GB	*Generalized Born*
GBA	*Guilt-by-association*
GEP	*Gene expression profile*
GO	*Gene Ontology*
GoF	*Gain-of-function*
GWAS	*Genome-wide association studies*
H	
HCC	*Hepatocellular carcinoma*
HGNC	*HUGO Gene Nomenclature*
HMM	*Hidden Markov Model*
I	
i.i.d.	*Independently and identically distributed*
ICD	*International Classification of Disease*
indel	*Insertion-deletion*
IP-MS	*Immunoprecipitation combined with mass spectrometry*
L	
LoF	*Loss-of-function*
M	
MD	*Molecular dynamics*

MeSH	*Medical Subject Headings*
MI	*Mutual information*
ML	*Machine learning*
MM	*Molecular mechanics*
MoA	*Mechanism of action*
MS	*Mass spectrometry*
MSA	*Multiple sequence alignment*

N

NMFI	*Naive mean field inversion*
NMR	*Nuclear magnetic resonance*
NNT	*Number needed to treat*
NPT	*Fixed particle number, temperature, and pressure*

O

OMIM	*Online Mendelian Inheritance in Men*
OPLS	*Optimized Potential for Liquid Simulations*

P

PB	*Poisson-Boltzmann*
PCA	*Principal component analysis*
PD	*Pharmacodynamics*
PDB	*Protein Data Bank*
PGx	*Pharmacogenetics/pharmacogenomics*
PharmGKB	*The Pharmacogenomics Knowledgebase*
PK	*Pharmacokinetics*
PLM	*Pseudo-likelihood maximization*
PPI	*Protein-protein interaction*
PPV	*Positive predictive value*

Q

QC	*Quality control*

R

RD	*Risk difference*
RF	*Random Forest*
RL	*Receptor-ligand complex*
RMSD	*Root-mean-square deviation*
RNA	*Ribonucleic acid*
RR	*Risk ratio*

S	
SA	*Surface area*
SASA	*Solvent-accessible surface area*
SE	*Side effect*
SMILES	*Simplified Molecular Input Line Entry System*
SNV	*Single nucleotide variant*
T	
TCGA	*The Cancer Genome Atlas*
TMH	*Transmembrane helix*
TP	*True positive*
Q	
UMLS	*Unified Medical Language System*
V	
VCF	*variant call file*
VEP	*Variant Effect Predictor*
W	
WES	*Whole exome sequencing*
WGS	*Whole genome sequencing*
Y	
Y2H	*Yeast two-hybrid*

Appendix B

Contributions

All ideas, approaches and results presented in this work were developed and discussed with my supervisors Prof. Dr. Oliver Kohlbacher (OK) and Prof. Dr. Debora Marks (DM). The following collaborators also contributed to the different projects:

- Anna Green (AG)
- Bilge Sürün (BS)
- Chris Sander (CS)
- Evelina Pillai (EP)
- Jens Krüger (JK)
- Julian Heinrich (JH)
- Julianus Pfeuffer (JP)
- Mathias Schwab (MT)
- Matthew Divine (MD)
- Philipp Thiel (PT)
- Roman Tremmel (RT)
- Sonja Hägele (SH)
- Thomas Hopf (TH)

Chapter 3: Predicting Protein Interactions using Coevolution

TH, DM and myself: Conception and design of the project, acquisition of data, analysis and interpretation of data, drafting or revising the article; TH and myself: implementation of pipeline and webserver; AG: Comparison of ATP synthase subunit interactions; OK: Acquisition of data, Analysis and interpretation of data; CS: Conception and design, Drafting or revising the article; all other co-authors: Expertise on HADDOCK docking protocols and revising the article.

Chapter 4: Genetic Variation in Drug Targets and Other Pharmacogenes

OK, DM and myself: Conception and design of the project, acquisition of data, analysis and interpretation of data, drafting or revising the article; RT and JS: Expertise on pharmacogenomics. TH: Support in using EVcomplex for predicting variant effects on protein function for selected examples.

Chapter 5: Drug Repurposing using the myDrug network

OK, DM and myself: Conception and design of the project, acquisition of data, analysis and interpretation of data. Implementation of `myDrug` and repurposing methods were performed by myself.

Chapter 6: Personalized Pharmacogene Analysis

OK and myself: Conception and design of the project, acquisition of data, analysis and interpretation of data; JP, SH and EP: Implementation of the protocol as KNIME workflows and partial analysis as part of their Master thesis projects supervised by myself; JK: insights into MM protocols; BS, JH, and MD: Input on conception of clinical reporting pipeline.

Appendix C

Publications

2018

Schubert B., **Schärfe C.P.I.**, Dönnes P., Hopf T., Marks D.S. and Kohlbacher O. (2018) "Population-specific design of de-immunized protein biotherapeutics." *Genome medicine* **9**, 117

2017

Schärfe C.P.I., Tremmel R., Schwab M., Kohlbacher O. and Marks D.S. (2017) "Genetic variation in human drug-related genes." *Genome Medicine* **9**, 117

Hopf T.A., Ingraham J.B., Poelwijk F.J., **Schärfe C.P.I.**, Springer M., Sander C. and Marks D.S. (2017) "Mutation effects predicted from sequence co-variation." *Nature Biotechnology* **35**, 128-135.

2016

Dietsche T., Mebrhatu M.T., Brunner M.J., Abrusci P., Yan J., Franz-Wachtel M., **Schärfe C.P.I.**, Zilkenat S., Grin I., Galán J.E., Kohlbacher O., Lea S., Macek B., Marlovits T.C., Robinson C.V. and Wagner S. (2016) "Structural and functional characterization of the bacterial type III secretion export apparatus." *PLoS Pathogens* **12**, e1006071.

Nicoludis J.M., Vogt B.E., Green A.G., **Schärfe C.P.I.**, Marks D.S. and Gaudet R. (2016) "Antiparallel protocadherin homodimers use distinct affinity-and specificity-mediating regions in cadherin repeats 1-4." *Elife* **5**

2015

Nicoludis J.M., Lau S.Y., **Schärfe C.P.I.**, Marks D.S., Weihofen W.A. and Gaudet R. (2015) "Structure and sequence analyses of clustered protocadherins reveal antiparallel interactions that mediate homophilic specificity" *Structure* **23**, 2087-2098.

2014

Gofman Y., **Schärfe C.P.I.**, Marks D.S., Haliloglu T., Ben-Tal N.(2014) "Structure, Dynamics and Implied Gating Mechanism of a Human Cyclic Nucleotide-Gated Channel." *PLoS Computational Biology* **10**, e1003976.

Hildebrandt A.K., Stöckel D., Fischer N.M., de la Garza Trevino L., Krüger J., Nickels S., Röttig M., **Schärfe C.P.I.**, Schumann M., Thiel P., Lenhof H.-P., Kohlbacher O. and Hildebrandt A. "ballaxy: web services for structural bioinformatics." *Bioinformatics* **31**, 121-122.

Krüger J., Grunzke R., Gesing S., Breuers S., Brinkmann A., de la Garza L., Kohlbacher O., Kruse M., Nagel W.E., Paschies L., Müller-Pfefferkorn R., Schäfer P., **Schärfe C.P.I.**, Steinke T., Schlemmer T., Warzecha K.D., Zink A. and Herres-Pawlies S. (2014) "The MoSGrid science gateway?a complete solution for molecular simulations" *Journal of Chemical Theory and Computation* **10**, 2232-2245.

Hopf T.A.*, **Schärfe C.P.I.***, Rodrigues J.P.*, Green A.G., Kohlbacher O., Sander C., Bonvin A. and Marks D.S.(2014) "Sequence co-evolution gives 3D contacts and structures of protein complexes" *Elife* **3**, e03430.

Avbelj M., Wolz O.-O., Fekonja O., Ben?ina M., Repi? M., Mavri J., Krüger J., **Schärfe C.P.I.**, Delmiro Garcia M., Panter G., Kohlbacher O., Weber A.N.R. and Jerala R. (2014) "Activation of lymphoma-associated MyD88 mutations via allostery-induced TIR domain oligomerization" *Blood* **124**, 3896-3904.

*all authors contributed equally

2013

de la Garza L., Krüger J., **Schärfe C.P.I.**, Röttig M., Aiche S., Reinert K. and Kohlbacher O. (2013) "From the Desktop to the Grid: conversion of KNIME Workflows to gUSE." *IWGS*.

Appendix D

Supporting Material

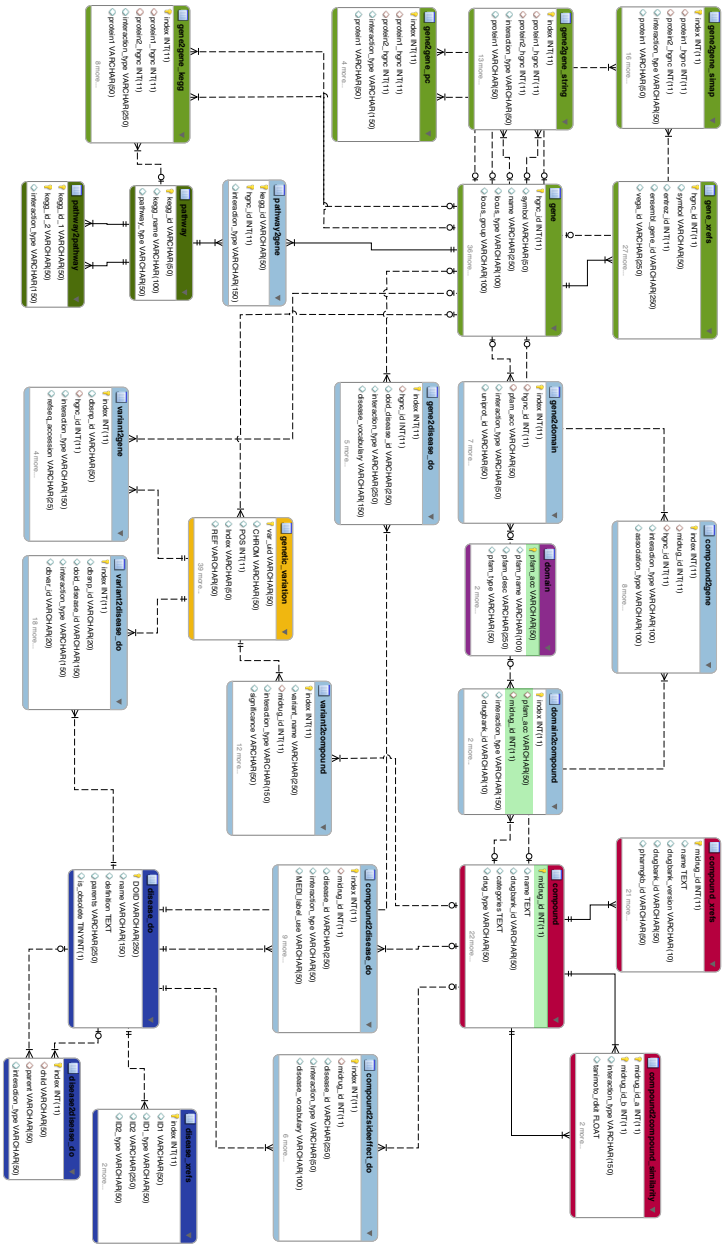


Figure D.1: Entety-Relationshio Diagram of `myDrug` mySQL database schema. Meta data about the node types are stored in respective tables, information about the relations between these nodes are stored in dedicated edge tables. Individual tables exist for disease information from OMIM, ICD9 and Disease Ontology, but for legibility reasons only DO tables are shown.

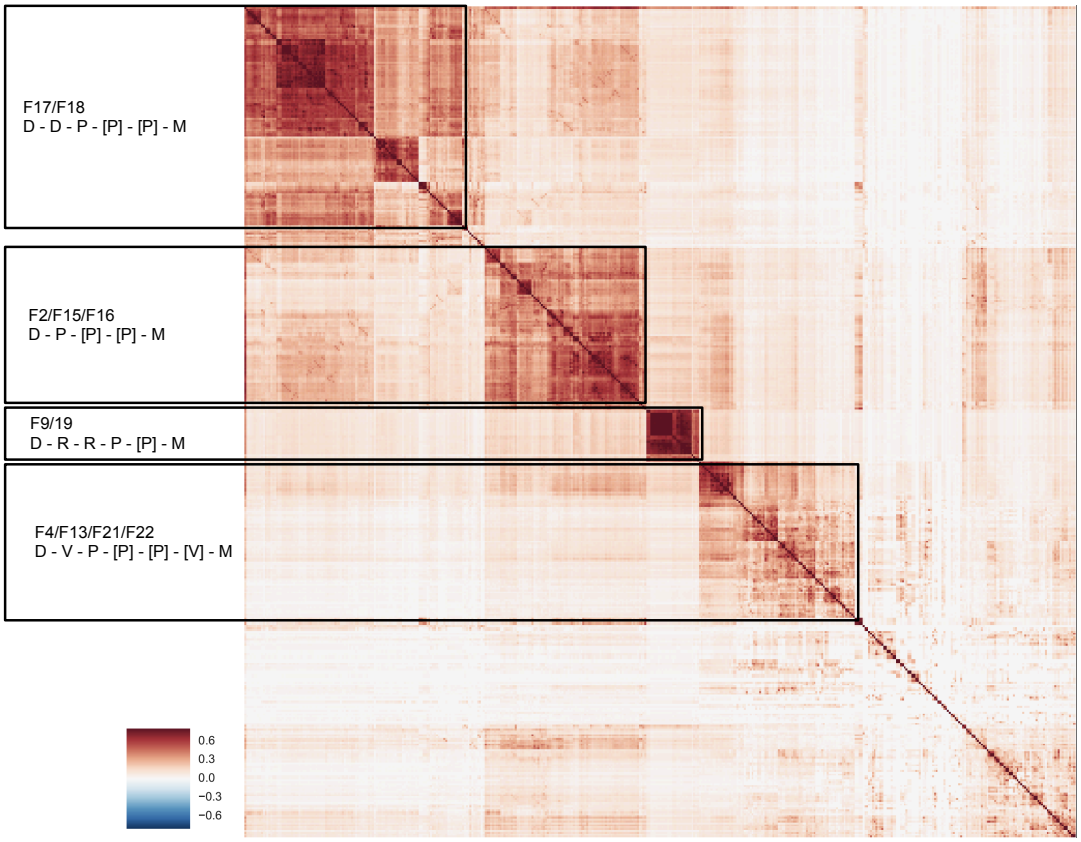


Figure D.2: Correlation matrix of feature instances used in `myDrug` Random Forest classifiers. See 5.3 for full list of features. Clusters of highly correlated features mainly correspond to different instances of similar features, build from different sources and interaction sub-types.

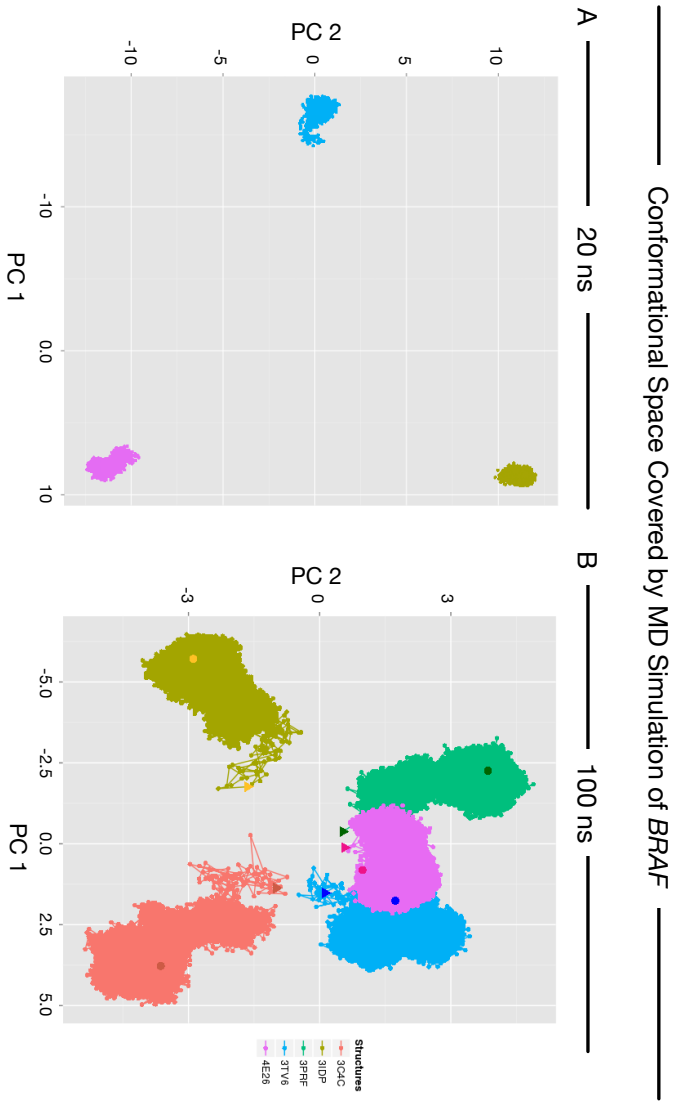


Figure D.3: Conformational space covered by MD simulation trajectories for several *BRAF* structures when simulated for (A) 20 ns and (B) 100 ns. Each frame is represented as a point in 2D, projected on the first and second principal component (PC) of dihedral space, lines connect subsequent frames in the trajectory. Starting conformations were *active* (DFG-in/αC-in) for 3TV6 and 3C4C, *intermediate* (DFG-in/αC-out) for 4E26 and 3PRF, and *inactive* (DFG-out/αC-out) for 3IDP.

Appendix E

Supporting Tables

Table E.1: Existing studies of genetic variants in pharmacogenes and their effect on different human populations.

Study	Genes	PGx Type	Samples	Origin	Findings
Nelson *et al.*, 2012 [302]	202	Targets	14,002	EUR, AFR, SAS	Abundance of rare variants (MAF $< 0.5\%$) across all populations, low level of shared rare variants between populations
Mizzi *et al.*, 2014 [287]	231	ADMET	482	EUR	Majority of variants in PGx genes are rare or singletons, many would not have been found with the most comprehensive existing PGx genotyping platform
Fujikura *et al.*, 2015 [114]	53	CYP450	6503	EUR, AFR, others	CYP genes have different genetic variability, most variants are very rare (MAF $< 0.1\%$), and profiles of genetic variants differ between individuals of European and African origin
Kozyra *et al.*, 2016 [218]	146	ADMET	6503	EUR, AFR	Vast majority of variants is rare (MAF $< 1\%$) and population specific, some genes show large differences in their variation rate between European-American and African-American patients
Wright *et al.*, 2016 [459]	120	ADMET	2504	AFR, AMR, EAS, SAS, EUR	Vast majority of variants is rare (MAF $< 0.5\%$); some of most conserved pharmacogenes are those with somatic mutations in cancer, highly differentiated variants observed between African cohort in relation others, 57% of PGx genes had high-confidence LoF variants, 97% of individuals carried at least one well-established PGx variant
Bush *et al.*, 2016 [44]	82	ADMET	5639	EUR (85%)	96% had at least one actionable PGx variant, between 4% (fluticasone) and 34.6% (simvastatin) of individuals are estimated to have a low-frequency missense variant that may affect its action (from PharmGKB annotation)

Table E.2: Data sources used for construction of myDrug objects.

Entity Type	Source	Origin	Description	Nodes
Compound	DrugBank [232]	curated	Most comprehensive public resource about drugs, their structures and other properties	8,221
Gene	HGNC	curated	Global consortium for gene symbols	40,967
Gene	Uniprot	partially curated	Information about the protein products of genes	19,855
Domain	Pfam	partially curated	Protein domain data created from hidden markov models	16,306
Disease	ICD9	curated	International billing codes	13,296
Disease	OMIM	curated	Compendium of heritable diseases	6,933
Disease	Disease Ontology	curated	Ontology that combines different sources, ranging from MeSH terms to OMIM	9,304
Variant	ExAC [236]	automatic	Gene variants found in 60,000 exomes	
Variant	dbSNP	community contributed	Database with all SNPs found in human population so far (subset only)	45,746
Pathway	KEGG	curated	Compendium of functional pathways in cells	302

Table E.3: Data sources used for construction of relations between `myDrug` objects.

Relation Type	Source	Origin	Description	Links
Compound-Gene	DrugBank	curated (publications)	annotates targets and ADMET interactions	17,992
Compound-Gene	IUPHAR	curated (publications)	drug - target links from literature, also contains activity data	1,703
Compound-Domain	PDB	automatic	binding sites extracted and then region mapped to PFAM domain	41,681
Compound-Disease	MEDI	text mining	disease-drug relations obtained from mining medical data sources	4,195
Compound-Disease	pharmacotherapy DB	curated	discriminates between etiology-centered and symptom-treating indications	1,385
Compound-Side Effect	SIDER	automatic	side effect data from drug labels	35,770
Compound-Compound		automatic	Tanimoto similarity on chemical fingerprints	24,967,714
Disease-Gene	OMIM	curated	morbidmap contains genetic basis for hereditary diseases	1,129
Disease-Gene	CTD	curated	relationships between chemicals, genes and human health	16,031
Disease-Gene	Orphanet	curated	knowledge base dedicated to rare diseases and their genetic basis	32,392
Disease-Gene	TTD	curated (textbooks, publications)	therapeutics centered association of diseases to genes	6,047
Disease-Disease	Disease Ontology	curated	ontology relationships for direct connections in the underlying ontology	8,128
Gene-Domain	Pfam	curated	protein family information based on Hidden Markov Models	41,681
Gene-Gene	KEGG	curated	gene interactions in pathways (directed)	98,000
Gene-Gene	Pathway Commons	automatic	gene interactions from multiple databases (some directed)	582,873
Gene-Gene	STRING	automatic	gene interactions from different levels of evidence (weighted & some directed)	553,069
Gene-Gene	SIMAP2	automatic	Smith-Waterman alignments of all proteins to all others	1,795,884
Pathway-Gene	KEGG	curated	each gene listed in pathway	25,445
Pathway-Pathway	KEGG	curated	pathway crossreferences in pathway map	1,807
Variant-Gene	myvariant.info	atomatic	collection of variant information from several annotation sources	109,081
Variant-Disease	GWAS Catalog	community contributed	collects pubslied GWAS data and the studied phenotypes	2,199
Variant-Disease	ClinVar	community contributed	collects relationships among medically important variants and phenotypes	58,582
Variant-Compound	PharmGKB	curated	haplotypes and variants relevant for PGx	18,880