Bringing bioinformatics into the classroom

Introducing

Computer-Aided Drug Design

A PRACTICAL GUIDE

Version: 30 November 2020

Designing tomorrow's drugs

Overview

This Practical Guide outlines basic computational approaches used in drug discovery. It highlights how bioinformatics can be harnessed to design drug candidates, to predict their affinity for their targets, their fate inside the body, their toxicity and possible side-effects.

Teaching Goals & Learning Outcomes

This Guide introduces bioinformatics tools for designing candidate drug molecules, and for predicting their likely target protein(s) and their drug-like properties. On reading the Guide and completing the exercises, you will be able to:

- design drug-candidate molecules using the structures of known drugs as templates, and dock them to known protein targets;
- compare the protein target-binding strengths of drug candidates with those of known drugs;
- calculate properties of drug candidates and infer whether they need chemical modification to make them more drug-like;
- predict the protein(s) that a drug candidate is likely to target;
- create molecular fingerprints for known drugs, and use these to quantify their similarity.

1 Introduction

Over the past century, the design and production of drugs has had a beneficial impact on life expectancy and quality[1,2]. However, drug-discovery projects are slow and expensive: the design and development process has been estimated to cost >1 billion dollars and to take at least 10 years to complete[3,4]. Despite this labour-intensive process, very few projects successfully lead to the release of new drugs[5,6]. Several technologies have been introduced to reduce the duration, cost and attrition rates of these projects: one of these is Computer-Aided Drug Design (CADD)[7-10]. CADD uses computing resources, algorithms and 3D-visualisation to help create or modify molecules, and rationalise the design process.

Nature has been the most important source of medicinal agents for centuries. Many useful drugs have been developed from plants: consider, for example, morphine derived from the opium poppy (*Papaver somniferum*), used for pain management; quinine from the Cinchona tree's bark, used as an **antimalarial** drug and muscle relaxant; and paclitaxel (also known as taxol) from the Pacific yew tree (*Taxus brevifolia*), used for cancer therapy. Natural molecules are still a major source of inspiration for drug design, but only 5% of small-molecule drugs developed in recent decades are purely natural products, unmodified in structure; the rest are natural-product derivatives containing synthetic modifications (27%), synthetic molecules inspired by natural products (35%), and totally new synthetic compounds (33%)[11]. In other words, 95% of new drugs have, at the least, necessitated chemical modification either to increase their target affinity and selectivity, to correct their Absorption, Distribution, Metabolism or Excretion (**ADME**) and toxicity problems, or to circumvent **Intellectual Property** (IP) issues. Although **serendipity** has had an important role in many therapeutic advances, rational design has become a major factor in developing new agents[12]; consequently, the vast majority of drugs developed in recent years have benefited to various extents from computer-aided approaches[7].

This Guide explores some of the challenges encountered in drug discovery and development, and the role played by CADD. It introduces the basics of drug design, and allows anyone with access to simple computational methodologies to conceive and evaluate molecules for their potential to become drugs[13]. Although macromolecular entities (*e.g.*, like **antibodies**) can act as therapeutic agents, here we consider drugs as small organic molecules (less than ~100 atoms) that activate or inhibit the functions of **proteins**, ultimately with therapeutic, **prophylactic** or diagnostic benefits to patients.

2 About this Guide

This Guide introduces simple computational approaches for drug discovery. It discusses what drugs are, how they work and where they come from, and how to select the best candidates. Exercises are provided both to allow visualisation of candidate molecules docking with known protein targets, and to predict their chemical properties and likely fate (and toxicity) in the body. Throughout the text, key terms – rendered in **bold** type – are defined in boxes. Additional information is provided in figures and supplementary boxes.

KEY TERMS

ADME (Absorption, Distribution, Metabolism, Excretion): properties that are used to investigate how a molecule is processed in the body & hence to determine its suitability as a drug

Amino acid: an organic molecule containing carboxylic acid (COOH) & amine (NH_2) functional moieties. There are 20 common, naturally occurring amino acids that constitute the building-blocks of proteins

Antibodies: large proteins produced by the immune system to fight foreign entities, such as disease-causing bacteria & viruses, or toxins

Antimalarial: an anti-parasitic chemical used to treat or prevent malaria

Intellectual Property (IP): asset or product derived from the human intellect (such as a patent, copyright, *etc.*) that may be protected by law & used for commercial gain

Prophylactic: an action or medicine used to prevent disease

Protein: a macromolecule comprising one or more **amino acid** chains; they are essential for the construction & activity of all biological entities (cells, virus, *etc.*)

Serendipity: a beneficial occurrence that happens by chance, a 'happy accident'

3 What are drugs & drug targets?

In our context, drugs are small molecules that bind to a target – generally a **target protein** – somewhere in the body, and, in so doing, modify its actions. Proteins are large molecules (with typically thousands or tens of thousands of atoms) comprising chains of amino acids, like beads on a necklace; these molecular chains can fold back upon themselves to form specific 3D shapes that depend on their amino acid sequences, as shown in Figure 1.

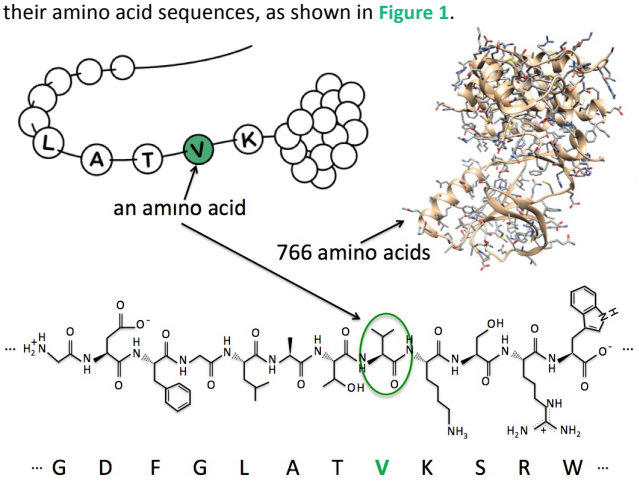

Figure 1 ***Proteins & their component amino acids.*** *A protein is like a string of beads folded over onto itself. Here, the amino acid 'beads' are denoted by a single-letter code: G, D, F… etc. (see box below for details).*

Amino acid residue notation

Figure 1 shows part of a protein sequence depicted using a single-letter code, where each of the letters represents one of the 20 naturally occurring amino acids. This single-letter code is detailed here.

Code	Amino acid	Code	Amino acid	Code	Amino Acid
G	Glycine	C	Cysteine	Q	Glutamine
A	Alanine	F	Phenylalanine	N	Asparagine
V	Valine	Y	Tyrosine	E	Glutamic Acid
L	Leucine	W	Tryptophan	D	Aspartic Acid
I	Isoleucine	H	Histidine	T	Threonine
P	Proline	K	Lysine	S	Serine
M	Methionine	R	Arginine		

Collectively proteins perform thousands of different biological functions in the body: from regulation and repair, transport and immune responses, to digestion, vision and breathing – they are essential building-blocks of life. Cells only make the proteins they need: each cell contains around 10,000 different proteins; and each protein may be present in 100 to 10 million copies (*e.g.*, there are 250 million **haemoglobin** proteins in a single red blood cell).

Protein functions depend on their amino acid sequences and 3D structures. Most proteins don't function in isolation, but work in concert with a range of other proteins and small molecules. Small molecules (*e.g.*, vitamins, sugar, oxygen, water, **lipids**) tend to bind at particular locations of the protein surface – their so-called binding-sites. The shapes and properties of binding-sites determine whether and which molecules can fit inside by 'complementarity'.

3.1 Proteins & disease

People fall ill for many reasons: *e.g.,* if too little of a particular protein is available for the body to use; if too much of a protein is available for the body to be able to use; if a protein's activity is altered in some way; if a protein's structure is altered as a result of **mutation**; and so on. Scenarios like this result in a host of diseases, ranging from diabetes, cancer and Alzheimer's, to arthritis, high blood pressure and heart disease, amongst many others. Knowing this allows the discovery of drugs to treat specific diseases, by targeting particular proteins – indeed, drugs in use today target more than 750 different human proteins[14]. Other drugs can also target bacterial or viral proteins.

4 How are drugs designed?

4.1 Where do drug molecules come from?

Of the drugs approved for therapeutic use today, most (95%) are entirely synthetic or are chemical modifications of natural products.

Potential drugs are often designed to target protein binding-sites. However, there are theoretically 10^{60} small **drug-like molecules** (**virtual compounds**) in 'chemical space'[15]. So where do we start? In practice, only some tens to hundreds of millions of these have been synthesised as real entities – but that's still a lot of molecular structures to consider; to date, only about 2,000 of those have been approved as drugs. This gives an idea of how difficult it is to find molecules with suitable properties to become new and effective drug molecule, like those illustrated in Figure 2.

4.2 How to select the best drug candidates

To help with the challenge of finding the small molecule that best fits a target protein's binding-site from the billions of possibilities available, various computational tools can be used; having discovered such a molecule, computational tools can also be used to optimise its shape and properties to create the best fit with the target. This approach is termed Computer-Aided Drug Design, or CADD.

Whether an optimised candidate molecule will then go on to become a new drug depends on several factors. Many laboratory tests have to be performed (*e.g.*, to investigate ADME and toxicity issues), followed by long clinical trials, in order to determine whether the molecule actually treats the disease efficiently, safely and at what dose (*i.e.*, with few undesirable side-effects). Note, clinical trials have large attrition rates, and only ~10% of trialled compounds are ultimately officially approved as drugs. Overall, from thousands of small molecules actually synthesised, only one is chosen and approved for therapeutic use. From start to finish, this process usually takes around 10 years to complete, at a cost of around a billion US dollars. Despite the time, cost and risk involved, dozens of new molecules hit the market every year.

KEY TERMS

Drug-like molecule: chemical compound showing similar properties (in terms of size, polarity, *etc.*) to existing oral drugs

Haemoglobin: iron-containing, oxygen-carrying protein of the blood

Lipid: a generic name for organic molecules that aren't water-soluble (*e.g.*, fat, oil, steroids, components of cell membranes)

Mutation: a change in genetic material (*e.g.*, a nucleotide change in a gene, or addition/loss of a chromosome or part of a chromosome)

Target protein: a protein that's involved in a disease & has been defined as a target for a drug

Virtual compound: a chemical structure corresponding to a theoretical compound that hasn't yet been synthesised (often compiled in chemical libraries for screening in drug-discovery projects)

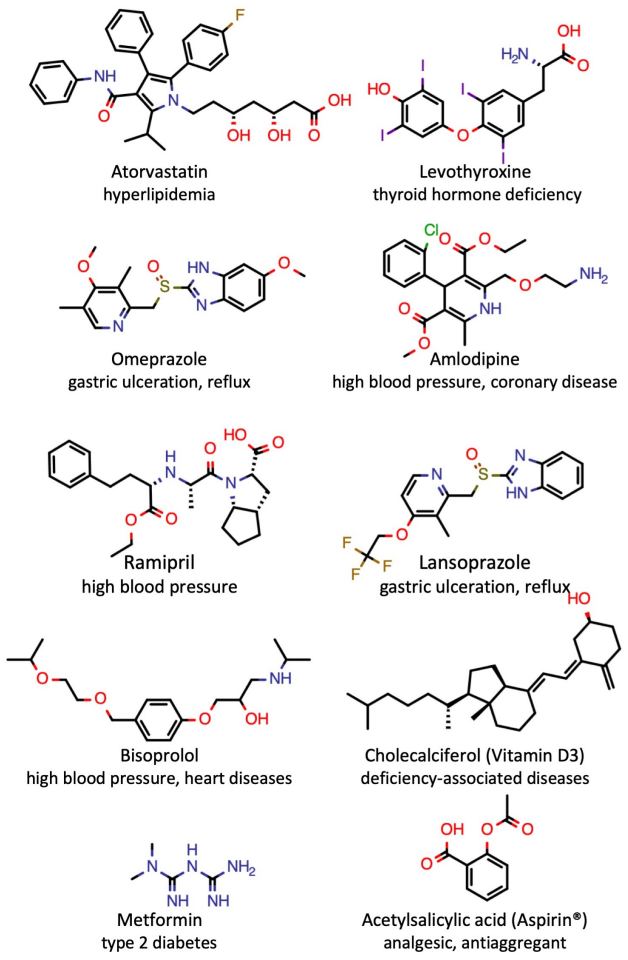

Figure 2 ***Chemical diversity of drug molecules exemplified by the 10 most-prescribed medicines in the UK in 2019.*** *The name and structure of the active ingredient, & the main therapeutic indications are given.*

5 Basic principles of CADD

CADD technologies can be broadly classified into structure-based and **ligand**-based approaches.

5.1 Structure-based approaches

Structure-based CADD approaches use the **3D structures** of protein targets, where these (or reliable models) are available[14]. The cornerstone of structure-based drug design is generally considered to be **molecular docking**. Amongst the most well-known successful applications of this approach are the anti-influenza drugs zanamivir (brand name Relenza) and oseltamivir (brand name, Tamiflu)[15,16].

The main rationale of structure-based drug design relies on how, and how strongly, a given small molecule will bind to a chosen target: the most likely geometry and position of small molecules at a protein surface can be calculated using a **docking program** (*e.g.*, SwissDock.ch[17], Autodock[18], Autodock Vina[19]) to predict the interactions between the molecular partners[8]; the strength of their binding can be evaluated using a score (there are several scoring functions available to estimate the binding free energy). The docking score is a parameter that's optimised when designing a potent drug molecule.

In essence, this approach hinges on optimising molecular recognition (fit, complementarity) and binding affinities (score, free energy), identifying molecules whose shape and properties are complementary to a protein target binding-site. This molecular-docking technique opens the road to *in silico* design and optimisation of virtual compounds, which is the subject of the drug-design-workshop educational website you'll use during the exercises.

5.2 Ligand-based approaches

Ligand-based approaches rely on the information contained in the chemical structures or physical properties (*e.g.*, size, **lipophilicity**, polarity) of other molecules that are known to bind to a chosen protein target. This information can be analysed and used to create predictive models using **machine-learning** methods: these are known as Quantitative Structure–Activity Relationships (QSAR), if the aim is to create new ligands and/or to predict their activity; or Quantitative Structure–Property Relationships (QSPR), if the aim is to predict behaviours related to the fate of the molecule in the body – the so-called **pharmacokinetic** (PK)[21] properties – which are fundamental in drug design. Although high affinity for a protein target is essential, it isn't sufficient for a designed small molecule to become a drug: to achieve a therapeutic effect, molecules must reach their targets in the body, and stay there long enough in a bioactive form to exert their biological effects. Thus, for efficient drug design, it's important to predict PK behaviours with CADD approaches.

Other ligand-based methods follow the principle that small molecules that are similar are more likely to be active on the same target. Such approaches can be used to perform **virtual screening** or **reverse screening**[22–24]. Reverse screening can help to predict both primary targets and potential *secondary* targets – *i.e.*, proteins to which a small molecule may bind, despite having been optimised to target another macromolecule. Secondary targets are often the root of adverse drug side-effects, but they can also open the way to what's known as **drug re-purposing**[25,26].

KEY TERMS

3D structure: the shape a protein or a small molecule adopts in space

Docking program: a software tool designed to predict how and how strong small molecules (*e.g.*, such as drug candidates) bind to a target (usually a protein) of known (or reliably modelled) 3D structure

Drug re-purposing: an ensemble of strategies to evaluate the uses of existing drugs for new therapeutic purposes

Ligand: a molecule that binds to another molecule & serves a biological purpose (*e.g.*, a drug binding to a target protein)

Lipophilicity: the ability of molecule to partition between fats, oils, lipids & water; this property is crucial to cross biological membranes (e.g. gastro-intestinal wall, blood-brain barrier or cell membrane)

Machine learning: a computer algorithm that discovers how to perform tasks (often from a set of training data), without having been explicitly programmed to do so. Here, for Structure-Activity relationships, the algorithm is used to find a mathematical equation linking the activity of the molecules with their properties

Molecular docking: a computational method used to predict how two or more molecules (*e.g.*, such as a drug & its target protein) interact

Pharmacokinetics (PK): study of the fate of a drug molecule, from the point of administration to the point of elimination from the body, to determine how the molecule has been Absorbed, Distributed, Metabolised & Excreted (ADME)

Reverse screening: deducing the potential protein targets of a given molecule by identifying similar compounds whose activity is known experimentally

Virtual screening: a computational method to search for molecules that are similar to known active compounds & are hence potentially also active on the same protein target

A typical CADD workflow is shown in **Figure 3**. Here, molecules can be selected that both have the highest affinity (*i.e.*, best docking score), and are as specific as possible, for their primary target, and have potential to become a drug: *i.e.*, molecules that are sufficiently permeable to be taken orally for the comfort and compliance of patients, have the best toxicity profiles, don't induce adverse drug-drug interactions via the **cytochrome P450** system, and so on.

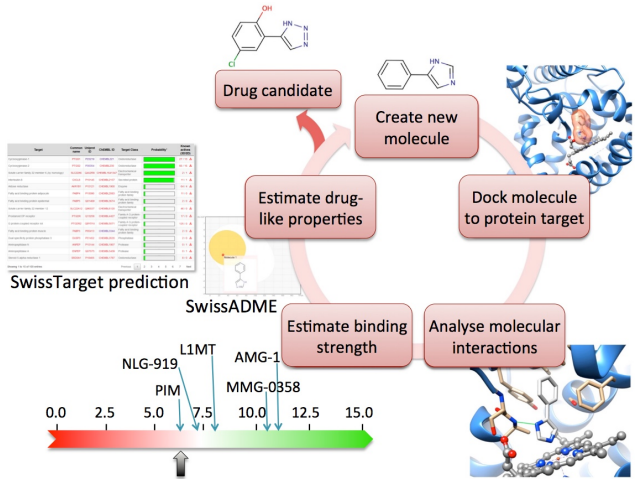

Figure 3 ***Typical CADD cycle.*** *A new molecule is created & docked to its protein target, its interactions are analysed, its binding strength scored, & its selectivity (e.g., calculated by SwissTargetPrediction) & drug-like properties (e.g., predicted by SwissADME) are estimated to assess its viability as a drug candidate.*

6 Hands-on drug design

We're now going to enter the iterative cycle of designing and optimising a molecule to make it a strong ligand for a selected target protein: specifically, cyclooxygenase (also known as COX). In humans, COX exists in two isoforms, which are referred to as COX1 and COX2. These are very similar, but have significantly different biological functions: COX1 is **constitutively expressed** and produces **prostaglandins** to fine-tune physiological processes (*e.g.*, it plays an important role in blood coagulation and in protecting the stomach lining); by contrast, COX2 is expressed locally in the event of inflammation, producing inflammatory prostaglandins that mediate responses to physiological stresses, such as inflammation, and is directly responsible for the sensation of pain.

Today, the most popular drug taken worldwide is Aspirin®, whose active molecule – acetylsalicylic acid – was synthesised at the end of the 19th century from salicylic acid, a natural product extracted from the bark of the willow tree. Willow bark has been known since antiquity for its effects in reducing fever. By the mid-18th century, bark extracts were also recognised for their efficacy in combating pain and inflammation. However, researchers discovered that some extracts caused digestive problems, such as gastric irritation, bleeding and diarrhoea, and even death when ingested in high doses. In 1897, scientists began to investigate acetylsalicylic acid as a less-irritating substitute for common salicylate medicines, and found the way to synthesise it. By 1899, the Bayer company had branded the new drug Aspirin®, and had put it into worldwide circulation.

Acetylsalicylic acid, when taken orally, inhibits both COX1 and COX2, and stops the synthesis of all prostaglandins. In doing so, it has desirable painkiller, anti-inflammatory and anti-fever effects. As a positive consequence of its non-selectivity of targets, Aspirin® is an early example of drug re-purposing, being routinely administered as a platelet anti-aggregant. However, most of its numerous side-effects, such as gastric ulceration, are unwanted.

Seeking to overcome these issues, another class of drugs was discovered: the Non-Steroidal Anti-Inflammatory Drugs (NSAIDs). Like Aspirin, the first NSAIDS inhibited COX1 and COX2, and were hence termed 'non-selective inhibitors'. Later, molecules that specifically inhibit COX2 were designed; these have the desired anti-inflammatory effects without the gastric side-effects – see **Figure 4**.

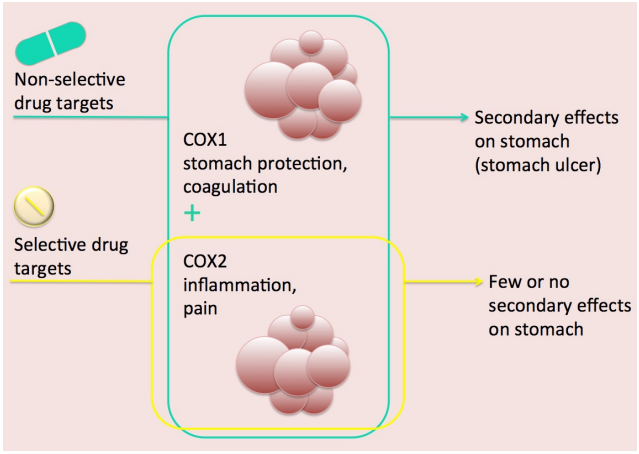

Figure 4 ***Selective & non-selective anti-inflammatory drugs & their gastric side-effects.*** *Non-selective drugs target COX1 & COX2, but have gastrointestinal side-effects; selective drugs target only COX2, with fewer gastric side-effects.*

In the following exercises, we shall use molecular docking to design a drug that targets COX. We'll then compare the effectiveness, safety and specificity predictions of this new molecule with the over-the-counter drug, Ibuprofen. Ibuprofen, an NSAID that inhibits COX1 and COX2, is used routinely for short-term treatments against fever or pain (headaches, toothache, muscle pains, *etc.*). Although proven a safe medication, like Aspirin, Ibuprofen can sometimes have adverse side-effects if taken at high dose and long-term, including potentially stomach or intestinal bleeding. Let's imagine that your aim is to investigate whether it's possible to optimise the Ibuprofen molecule for better selectivity towards COX2 versus COX1.

6.1 Design & dock your own molecule

To begin to answer this question, we'll use an online educational docking tool and CADD websites from the **SIB Swiss Institute of Bioinformatics**. A simplified view of the educational tool's workshop for anti-inflammatory drugs is shown in **Figure 5**.

KEY TERMS

Constitutively expressed: being produced at a constant rate & in constant amount in a given cell, regardless of the cell's metabolic state

Cytochrome P450: an enzyme whose main role is to chemically modify foreign chemicals in order to be able to clear them easily from the body; they are of particular interest in medicine for their role in metabolising exogenous molecules, like drugs

Prostaglandin: physiologically active lipid compounds involved in inflammation & the amplification of pain signals

SIB Swiss Institute of Bioinformatics: a not-for-profit foundation that federates bioinformatics activities across Switzerland, providing bioinformatics resources to the life-science research community

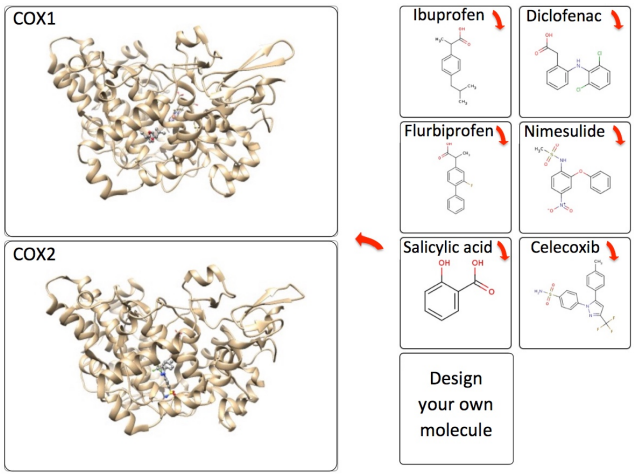

Figure 5 ***Design & docking of drug molecules with target proteins.*** *Drugs (right-hand side) are docked by dragging & dropping them onto the protein targets: COX1, COX2. A molecular 'sketcher' lets you 'Design your own molecule' from scratch or use one of the drugs as template.*

EXERCISES

1 Open a browser (preferably Firefox or Google Chrome) & visit the following website: **www.drug-design-workshop.ch/cox.php** (you might need to allow 'pop-up windows').

2 To design your own molecule, click on the '*Design your own molecule*' box. This opens a '*sketcher box*' in which to start the design process. Let's begin by using Ibuprofen as the template structure.

3 Click on the red '*down arrow*' – top right of Ibuprofen's box. This copies the structure into the '*sketcher*' as the template for modification (for help, see **www.drug-design-workshop.ch/helpsketcher_videos.php**).

4 First, let's add a '*chloro*' substituent in the '*ortho*' position of Ibuprofen's aromatic ring: to do this, refer to **Figure 6**. Add a single bond at the correct position: click 1 then 2, as indicated in **Figure 6**.

5 This adds a methyl group (CH_3). Replace this with a chlorine atom: click 3 then 4 (**Figure 6**). Your new '*o*-chloroibuprofen' is ready.

6 Finally, click on the '*Done*' button beneath the sketcher to transfer the modified structure into '*Your own molecule*' box.

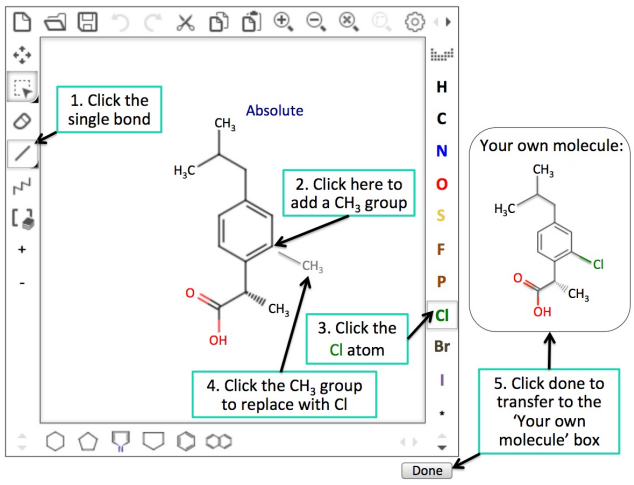

Figure 6 ***Designing a new drug candidate from a known template.*** *The 'sketcher' allows a new compound to be designed or a template structure to be modified, by adding bonds, atoms & chemical groups, or by replacing selected atoms with options from its toolbars. Here, a 'chloro' substituent is added to the aromatic ring of Ibuprofen's molecular template.*

Once you've designed your own molecule, you're ready to investigate how it docks with the target proteins, COX1 or COX2. The docking tool evaluates hundreds of thousands of different geometries and positions of the ligand within the protein; it then reveals the most probable binding mode in an interactive 3D display, alongside an estimation of the binding strength, rendered as a docking score – broadly speaking, the larger the score, the better the ligand. The score of your molecule is set in the context of the relative binding strengths of a range of known drugs (as shown in **Figure 7**); it's thereby possible to compare your molecule's score with those of the reference drugs.

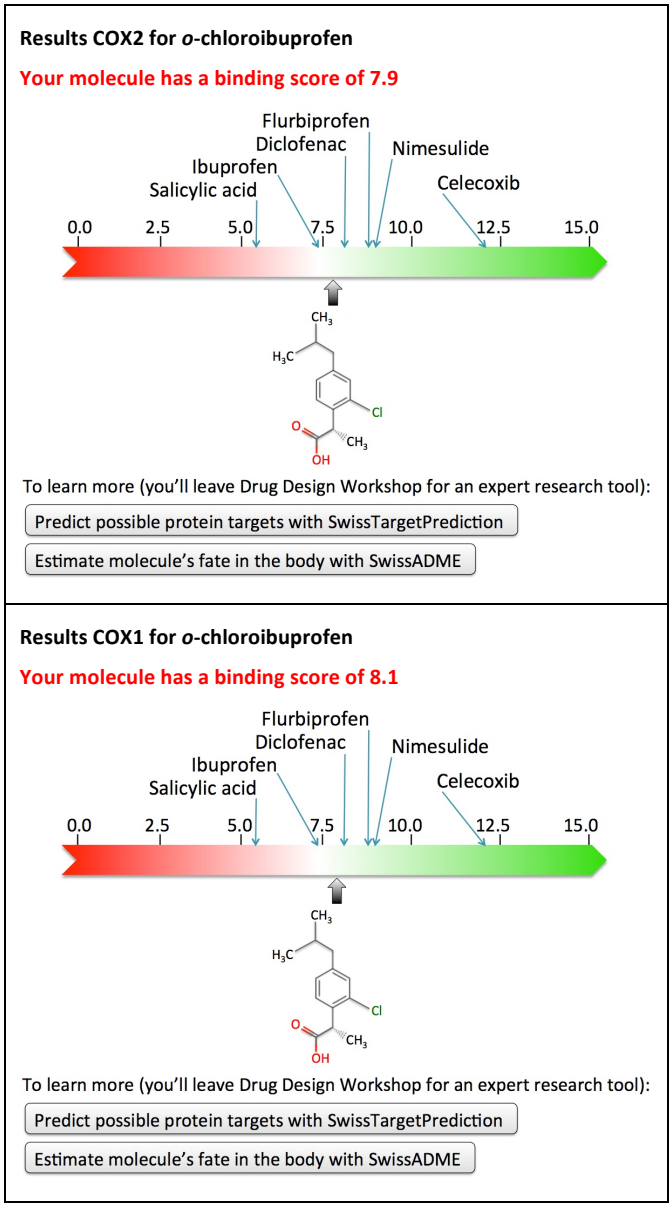

Figure 7 ***Calculated binding score of a drug candidate relative to a set of reference drugs.*** *The drug candidate is shown in the centre of the figure – the black arrow indicates its docking score with target proteins COX1 & COX2 relative to the reference drugs. Here, the drug candidate is o-chloroibuprofen, a modification of Ibuprofen's template molecular structure, & the binding score with COX1 is slightly higher than that for COX2. More detailed analyses (to predict possible protein targets & to estimate the molecule's fate in the body) can be carried out by clicking on the links to the software tools SwissTargetPrediction & SwissADME.*

EXERCISES

1 We will now explore how your molecule binds to COX1 & COX2, & how it compares to Ibuprofen docking in terms of binding strength. Drag & drop your own molecule's image onto the 3D representation of COX1 in the top left-hand panel.

2 This initiates the docking calculations on a remote server, which may take a couple of minutes to run, depending on the load on the server. A new tab opens in your browser to let you monitor the progress of the calculations. When the docking is complete, click on '*here*' to access the results. At the top of the page, you'll see a 3D representation of the computed binding mode between your molecule & COX1.

3 Use your mouse or track-pad to interact with the image. The coloured ribbons represent the **secondary structure** of the protein (see **Figure 8**).

4 Note the docking score. What is the score of your molecule relative to Ibuprofen, whose score is 7.8 on COX1? Is it better or worse?

5 Return to the first tab to re-run the docking calculations, now dragging & dropping your molecule onto the 3D representation of COX2 (bottom left-hand panel).

6 Once completed, note the docking score (refer to **Figure 7**). What is the binding score of your drug candidate molecule relative to Ibuprofen, whose score is 7.6 on COX2? Is it better or worse?

7 Overall, the selectivity of your newly designed molecule for COX2 over COX1 is worse than Ibuprofen. Optimisation of molecules is often very empirical (trail-&-error). Typically, it's necessary to go back to the drawing board (here, the 'sketcher') to generate another hypothesis with a new designed molecule.

Figure 8 ***Representation of the 3D structure of the anti-inflammatory drug, Ibuprofen bound to its target protein, COX2.*** *The Ibuprofen ligand (seen at the centre of the image) is shown in ball-&-stick representation, with carbon atoms coloured brown & oxygen atoms red; the coloured ribbons represent the secondary structure of the protein: α-**helices** are shown in pink, **β-strands** in yellow, & loops in grey.*

Having designed your first drug-like molecule, we're now going to investigate what happens when we add a more bulky purely carbon-based substituent to the same position where we added the 'chloro' substituent to Ibuprofen in the first exercise.

EXERCISES

1 Return to the first tab, & click on '*Your own molecule*' to open the sketcher with the *o*-chloroibuprofen you created previously.

2 First, let's replace the chlorine atom with carbon, & then add three single bonds from that carbon to generate three methyl groups: to do this, refer to **Figure 9** (top panel). Your designed '*o*-tertiobutylibuprofen' is ready.

3 Click on the '*Done*' button beneath the sketcher to transfer the modified structure into '*Your own molecule*' box.

4 Dock the molecule into COX1 & COX2 by repeating steps 1 to 5 of the previous exercise. Note the binding score for each. What conclusions can you draw regarding the selectivity compared to Ibuprofen?

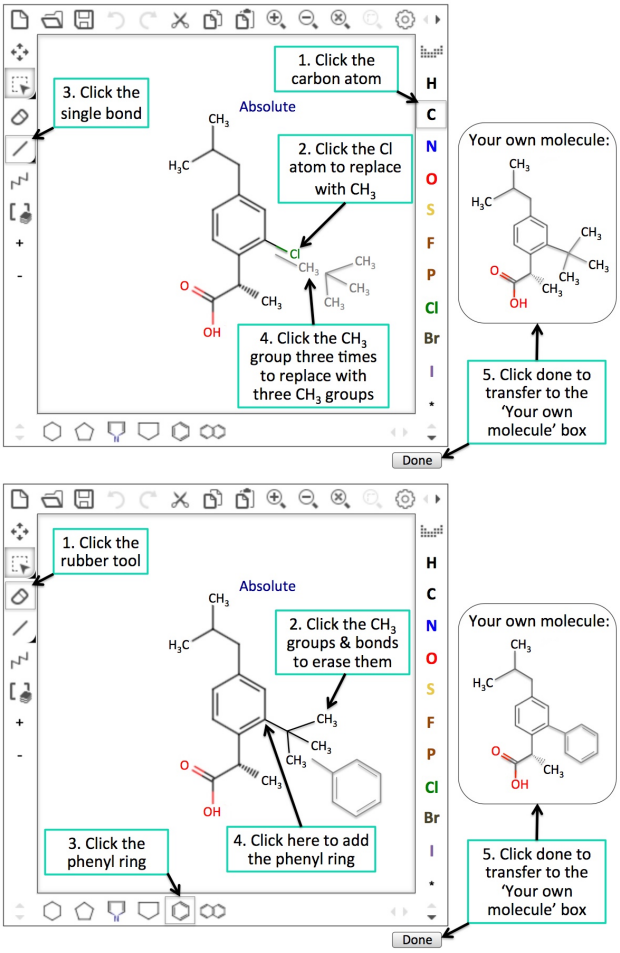

Figure 9 ***Designing a new drug candidate from a known template.*** *The upper panel shows addition of three methyl groups to the 'ortho' position of Ibuprofen's aromatic ring; the lower panel shows replacement of the methyl groups with an even bulkier phenyl group. To see the sketcher in action, visit **www.drug-design-workshop.ch/helpsketcher_videos.php**.*

Overall, the selectivity of your second molecule for COX2 over COX1 is better relative to Ibuprofen. In fact, the binding strength on COX1 is slightly less than Ibuprofen, but that on COX2 is significantly

KEY TERMS

α-helix: region of a protein chain that forms a regular helical structure
β-strand: region of a protein chain that's almost fully extended
Secondary structure: the local structural organisation of a protein, such as helices, beta strands, *etc.*

higher. The iterative optimisation cycle now begins: *i.e.*, we continue with the hypothesis by adding even bulkier substituents at the same position, as this appears to play an important role, favouring binding to COX2 while having a negative impact on binding to COX1.

So, having designed a second molecule with some success, we're now going to investigate further by adding an even bulkier substituent at the same position in the Ibuprofen molecule.

EXERCISES

1 Return to the first tab, & click on '*Your own molecule*' to open the sketcher with the previously created *o*-tertiobutylibuprofen.
2 Let's remove *the 3 methyl groups* using the '*rubber*' tool (left-hand toolbar of sketcher) & then add a phenyl ring at the same position (bottom toolbar of sketcher): refer to **Figure 9** (bottom panel). Your designed '*o*-phenylibuprofen' is ready.
3 Click on the 'Done' button beneath the sketcher to transfer the modified molecule into '*Your own molecule*' box.
4 Dock the molecule into COX1 & COX2 by repeating steps 1 to 5 of the earlier exercise, as before. Note the binding scores. What conclusions can you draw regarding the selectivity compared to Ibuprofen?

Overall, the selectivity of this third molecule for COX2 over COX1 is significantly better: the difference between the COX1 and COX2 docking scores has increased even more. The docking scores for each of the drug-candidate molecules we've investigated so far are summarised in **Table 1**.

Table 1 ***Summary of docking scores for Ibuprofen & three drug-candidate molecules based on Ibuprofen's structural template.***

Docking scores	Ibuprofen	*o*-chloro-ibuprofen	*o*-tertio-butylibuprofen	*o*-phenyl-ibuprofen
COX1	7.8	8.1	7.9	8.6
COX2	7.6	7.9	8.1	9.5

We could continue generating more hypotheses to try to create further molecules with better COX1-/COX2-binding affinities than Ibuprofen. However, we're now going to explore other important aspects of drug design using the software tools *SwissTargetPrediction* and *SwissADME*: the former predicts the most probable protein targets for a drug candidate; the latter calculates its **physicochemical**, pharmacokinetic and drug-like properties, thereby allowing you to determine its likely fate in the body: *e.g.*, is it well absorbed by the gastrointestinal tract? Does it reach the brain? Is it toxic?

6.2 Calculate your molecule's ADME parameters

To explore these questions, let's take a closer look at the *SwissADME* tool. The component of this tool that allows you to compute the lipophilicity and polarity of a small molecule, and thus to predict its likely brain or intestinal permeation, is called *BOILED-Egg* – **Figure 10** shows a typical output. The colour-coded regions (which give the model its name) depict areas where those properties would allow molecules to cross specific biological barriers (*i.e.*, bowel wall or blood-brain barrier). The position of the molecule under investigation is shown as a circle; its colour reflects the likelihood of it being a substrate for the one of the major protein protecting the brain by pumping-out foreign substances, the P-glycoprotein 1, or PGP. The graph is a useful tool, allowing intuitive assessment of the types of chemical modification needed to give a molecule the desired absorption and distribution properties.

As an example, anti-inflammatory drugs must primarily target peripheral proteins. As such, and to avoid unwanted central effects, we can optimise the candidate drug structure to prevent it from accessing the brain. To evade the *BOILED-Egg's* 'yolk' region, we must increase the molecule's polarity. To do this, we can add polar atoms or groups: *e.g.*, by adding OH substituents to the phenyl ring (**Figure 11**). You can apply as many optimisation cycles as necessary; however, the impact of every structural modification must be evaluated for all properties (binding and selectivity towards COX2 and COX1).

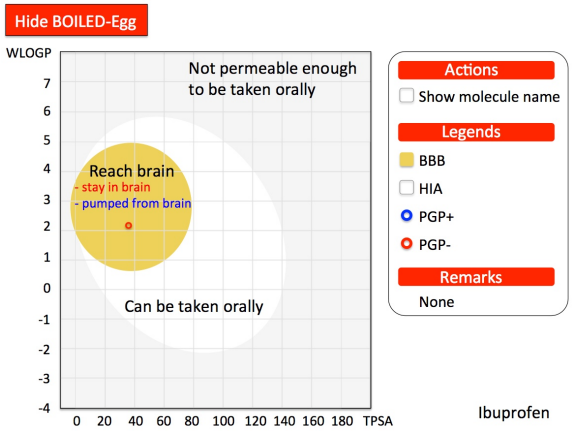

Figure 10 ***Graphical output of the BOILED-Egg model.*** *The y-axis denotes a molecule's lipophilicity (as a predicted partition coefficient between water & octanol, WLOGP), the x-axis its apparent polarity (as a polar surface area, TPSA). The graph predicts its passive Human Intestinal Absorption (HIA, white ellipse) & its permeation through the Blood-Brain Barrier (BBB, yellow 'yoke'). In the grey region, molecules aren't predicted to be well absorbed when taken orally. A circle shows the location of the molecule, its colour denoting whether it's a substrate for P-glycoprotein 1 (PGP), & hence whether it's pumped-out from the brain.*

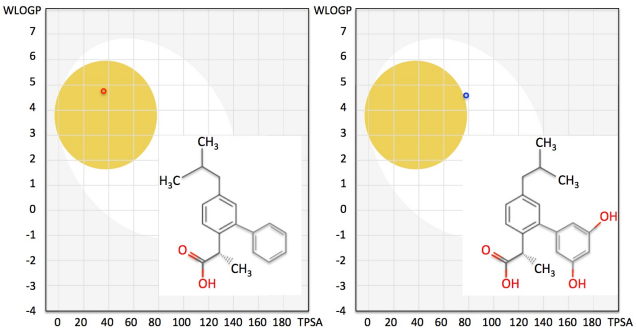

Figure 11 ***Optimising the structure of a drug candidate to prevent it from accessing the brain.*** *Here, the polarity of the molecule is increased by adding OH substituents to the phenyl group.*

EXERCISES

1 To evaluate the likely fate of your *o*-phenylibuprofen drug candidate, click on the '*Estimate molecule's fate in the body with SwissADME*'s box (bottom box, **Figure 7**). The results will appear in a new window.
2 In the results page, click on the red box '*Show BOILED-Egg*'.
3 This reveals a graph (**Figure 10**) where your molecule appears as a coloured circle. In which region does it fall? Could it be taken orally? Does it reach the brain? Is it predicted to be pumped-out by PGP?
4 Your molecule's physicochemical & pharmacokinetic properties are listed below this graph (see **Figure 12**). Inspect the '*Medicinal Properties*' section. Are there any toxicity alerts? If so, what are they?

KEY TERMS

Physicochemical properties: properties relating to the physical chemistry of a molecule: its lipophilicity, polarity, polar surface area, *etc.*

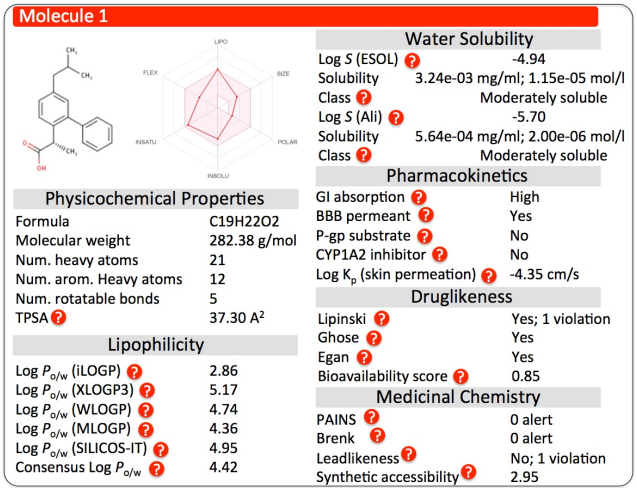

Figure 12 ***Selected physicochemical, pharmacokinetic & drug-like properties of Ibuprofen.*** *The Medicinal Chemistry heading is particularly important: it's here that toxicity alerts are found.*

The *SwissADME* program computes a range of properties that are important in determining whether a molecule has desirable drug-like properties, as illustrated in **Figure 12**. Particularly important here are those properties listed under the *Medicinal Chemistry* heading: notably, the alerts for problematic molecular fragments (those potentially toxic, unstable, reactive or aggregator, *etc.*).

6.3 Predict your molecule's target protein(s)

Let's now consider the specificity of your drug candidate in terms of its likely target protein(s). To do this, we'll use the *SwissTargetPrediction* tool. This returns a table of the 15 most probable targets (by similarity with known active molecules targeting these proteins), and provides links to additional information in the **UniProt** and **ChEMBL** databases – see **Figure 13**; it also provides links to lists of known active molecules with similar 2D or 3D chemical structures to your candidate molecule that could have the same protein target.

Target	Common name	UniProt ID	ChEMBL ID	Target Class	Probability*	Known actives (3D/2D)
Cyclooxygenase-1	PTGS1	P23219	CHEMBL221	Oxidoreductase		36/23
Cyclooxygenase-2	PTGS2	P35354	CHEMBL230	Oxidoreductase		90/60
Solute carrier family 22 member 6 (by homology)	SLC22A6	Q4U2R8	CHEMBL1641347	Electrochemical transporter		2/1
Interleukin-8	CXCL8	P10145	CHEMBL2157	Secreted protein		12/2
Aldose reductase	AKR1B1	P15121	CHEMBL1900	Enzyme		133/7
Fatty acid binding protein epidermal	FABP5	Q01469	CHEMBL3674	Fatty acid binding protein family		4/0
Solute carrier family 22 member 12	SLC22A12	Q96S37	CHEMBL6120	Electrochemical transporter		66/0
Prostanoid DP receptor	PTGDR	Q13258	CHEMBL4427	Family A G protein-coupled receptor		24/0
G protein-coupled receptor 44	PTGDR2	Q9Y5Y4	CHEMBL5071	Family A G protein-coupled receptor		246/0
Aminopeptidase N	ANPEP	P15144	CHEMBL1907	Protease		0/2
Aminopeptidase A	ENPEP	Q07075	CHEMBL3439	Protease		0/2
Steroid 5-alpha-reductase 1	SRD5A1	P18405	CHEMBL1787	Oxidoreductase		9/0
Voltage-gated calcium channel alpha/delta subunit 1 (by homology)	CACNA2D1	P54289	CHEMBL1919	Calcium channel auxiliary subunit alpha2delta family		0/13
Branched-chain-amino-acid aminotransferase, mitochondrial	BCAT2	O15382	CHEMBL3616354	Transferase		2/0
3-phosophoinositide dependent protein kinase -1	PDPK1	O15530	CHEMBL2534	Kinase		2/0

Figure 13 ***Target proteins for Ibuprofen predicted by SwissTargetPrediction.*** *The target name & protein class are indicated, alongside* ***accession numbers*** *for entries in the UniProt & ChEMBL databases. The probability denotes the probability for Ibuprofen to have these proteins as its target.*

EXERCISES

1 To predict the likely target protein for your candidate drug, click on '*Predict possible protein targets with SwissTargetPrediction*' box (upper box, bottom of **Figure 7**). The calculation will take a couple of seconds to run. The results will be displayed in a new window.

2 Your candidate molecule will be shown at the top of the results page, alongside a pie-chart showing the classes of protein it's most likely to target. Beneath this are the 15 most probable protein targets.

3 What is the top target protein? Are Cyclooxygenase-1 (COX 1, also called PTGS1) & Cyclooxygenase-2 (COX 2, PTGS2) in the list?

6.4 Molecular fingerprints

As mentioned in Section 5, ligand-based methods are one of the main forms of CADD. The rational here is grounded in the *similarity principle*: *i.e.*, similar molecules are likely to exhibit similar biological activities[27,28]. This concept underpins the technique of virtual screening, where libraries of molecules (actual or virtual compounds) that are similar to known active molecules are searched to find compounds that should be prioritised for experimental testing.

Nowadays, screenable chemical libraries can contain hundreds of millions to billions of molecules. Dedicated computer methods have been developed to rapidly quantify similarities between them. One such method uses so-called 'molecular fingerprints'. In this approach, the chemical structures of molecules are translated into strings of 0s and 1s – a computer then 'sees' the molecules by means of their digital 'fingerprints'[29,30]. These fingerprints are created by breaking down the molecular structures into their component chemical fragments; the presence (1) or absence (0) of each fragment is then noted, as illustrated in **Figure 14**. Pairs of molecules can then be readily compared via their digital fingerprints.

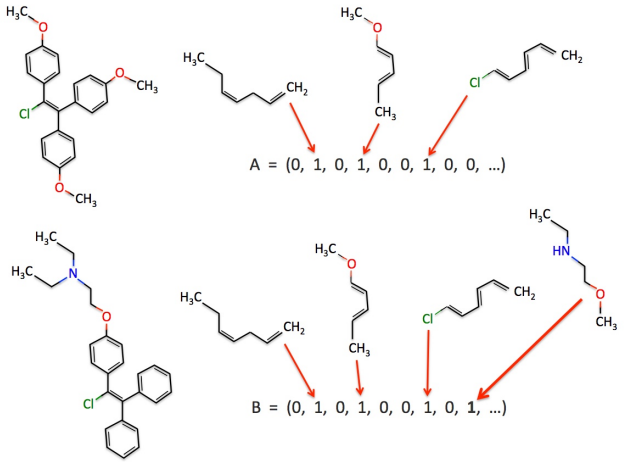

Figure 14 ***Molecular fingerprints.*** *The presence (1) or absence (0) of given chemical fragments create 'fingerprints' for molecules A & B.*

KEY TERMS

Accession number: a unique computer-readable code given to identify a particular entry in a particular database

ChEMBL: a large & highly curated database of molecule bioactivity

UniProt: the world's most comprehensive database of protein sequence & functional information

As shown in Figure 14, the fingerprints of molecules A and B are:

A = (0, 1, 0, 1, 0, 0, 1, 0, 0, …)

B = (0, 1, 0, 1, 0, 0, 1, 0, **1**, …)

The similarity between these molecular fingerprints can be calculated using the Tanimoto coefficient (T)[31]. The value of T is derived by dividing the number of times 1 is found in the same position for both molecules by the number of times 1 is found in at least one of the molecules, as expressed in the equation in Figure 15.

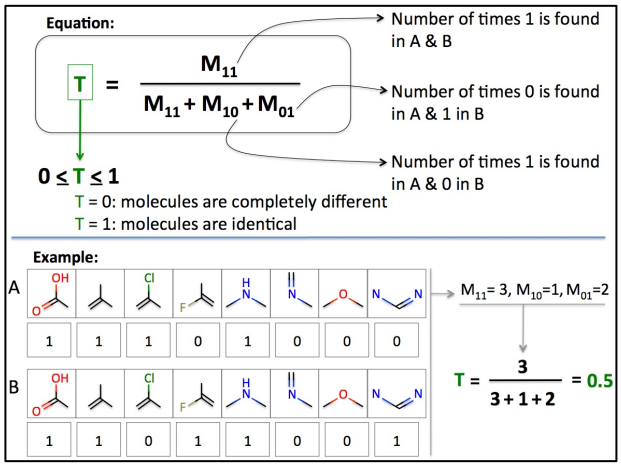

Figure 15 ***The Tanimoto coefficient (T).*** *Comparing the similarity of two molecules, A & B. The value of T ranges from 0 to 1, for molecules that are completely different or identical, respectively.*

To get an idea of how this can be applied in practice, we can calculate the Tanimoto coefficient of short fingerprints with a restricted number of fragments for the set of anti-inflammatory and anti-cancer drugs shown in Figure 16.

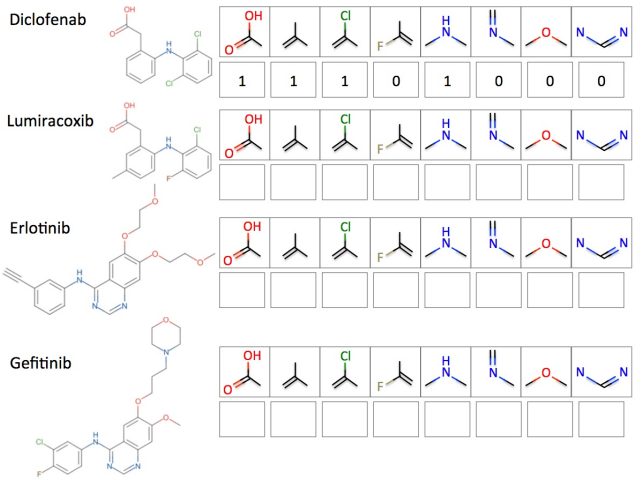

Figure 16 ***Fingerprint-based similarity.*** *The molecular fingerprint of* ***Diclofenac*** *is illustrated at the top of the figure. Noting the presence (1) or absence (0) of the chemical fragments shown allows us to create the corresponding fingerprints for* ***Lumiracoxib***, ***Erlotinib*** & ***Gefitinib****.*

Table 2 ***Tanimoto coefficients of anti-inflammatory & anti-cancer drugs***

	Diclofenac	Lumiracoxib	Erlotinib	Gefitinib
Diclofenac	1			
Lumiracoxib		1		
Erlotinib			1	
Gefitinib				1

EXERCISES

1 To create molecular fingerprints for the anti-inflammatory & anti-cancer drugs shown in **Figure 16**, visit the following website: ***www.drug-design-workshop.ch/pen-and-paper.php***

2 Click on he image shown & print the PDF.

3 Now complete the pen-and-paper exercise, as follows: for each of the drugs shown beneath Diclofenac, record in the empty boxes the presence (1) or absence (0) of the chemical fragments shown in the panel above them.

4 When all fingerprints are completed, calculate the Tanimoto coefficients for each pair of molecules using the equation, and following the example, shown in **Figure 15**. Populate the table (reproduced as **Table 2** opposite) with each value. Correct values can be found online at **www.drug-design-workshop.ch/pen-and-paper.php.**

5 Examine the results in the table. Which molecules are most similar – *i.e.*, likely to target the same protein?

The exercise above gives an idea of how computers translate molecules into fingerprints, and quantify their molecular similarity.

Having calculated the Tanimoto coefficients, two groups of similar molecules emerge in Table 2: Diclofenac and Lumiracoxib on the one hand, and Erlotinib and Gefitinib on the other. This result is consistent with the fact that Diclofenac and Lumiracoxib are COX inhibitors, used to treat pain and inflammatory diseases, while Erlotinib and Gefitinib are kinase inhibitors, used in cancer treatment. This outcome illustrates the similarity principle, frequently used in medicinal chemistry, which claims that similar molecules are prone to share similar biological activities.

TAKE HOMES

Having completed this Practical Guide, you now have a practical sense of how to:

1 Design a drug-candidate molecule, based on a known drug template;

2 Visualise in 3D how a drug candidate binds to a target protein;

3 Optimise the docking score, or binding strength, of a candidate drug with regard to the binding strengths of known drugs;

4 Use the *SwissADME* software tool to predict the fate of a drug or drug candidate in the body, & to calculate the physicochemical & pharmacokinetic properties of such molecules;

5 Infer whether a drug candidate requires chemical modification to optimise some of its ADME properties (such as absorption when taken orally or access to the brain);

6 Use the *SwissTargetPrediction* software tool to predict the specificity of a drug candidate in terms of its likely target protein(s); &

7 Quantify the similarity between molecules using the Tanimoto coefficient.

7 References & further reading

1 Bunker JP. (2001) **The role of medical care in contributing to health improvements within societies.** *Int. J. Epidemiol.*, **30**(6), 1260-1263.

2 Lichtenberg FR. (2005) **The impact of new drug launches on longevity.** *Int. J. Health Care Finance Econ.*, **5**(1), 47–73.

3 Munos B. (2009) **Lessons from 60 years of pharmaceutical innovation.** *Nat. Rev. Drug Discov.*, **8**(12), 959–968.

4 Paul SM *et al.* (2010) **How to Improve R&D productivity: the pharmaceutical industry's grand challenge.** *Nat. Rev. Drug Discov.*, **9**(3), 203–214.

5 Arrowsmith J & Miller P. (2013) **Trial watch: phase II & phase III attrition rates 2011-2012.** *Nat. Rev. Drug Discov.*, **12**, 569.

6 Waring MJ *et al.* (2015) **An analysis of the attrition of drug candidates from four major pharmaceutical companies.** *Nat. Rev. Drug Discov.*, **14**(7), 475-486.

7 Jorgensen WL. (2004) **The many roles of computation in drug discovery.** *Science*, **303**(5665), 1813-1818.

8 Zoete V *et al.* (2009) **Docking, virtual high-throughput screening & *in silico* fragment-based drug design.** *J Cell. Mol. Med.*, **13**(2), 238-248.

9 Schneider G. (2012) **From theory to bench experiment by computer-assisted drug design.** *Chimia,* **66**(3), 120-124.

10 Woltosz WS. (2012) **If we designed airplanes like we design drugs...** *J. Comput.-Aided Mol. Des.,* **26**(1), 159-163.

11 Newman DJ & Cragg GM. (2020) **Natural products as sources of new drugs from 1981 to 2019** *J. Nat. Prod.*, **83**, 770-803.

12 Seddon G *et al.* (2012) **Drug design for ever, from hype to hope.** *J. Comput.-Aided Mol. Des.,* **26**(1), 137-150.

13 **Drug design workshop: www.drug-design-workshop.ch**

14 **The Human Protein Atlas. The druggable proteome: www.proteinatlas.org/humanproteome/tissue/druggable**

15 Bohacek RS *et al.* (1996) **The art and practice of structure-based drug design: a molecular modelling perspective.** *Med. Res. Rev.*, **16**(1) 3-50.

16 Schmidt T *et al.* (2014) **Modelling 3D protein structures for applications in drug design.** *Drug Discov. Today*, **19**(7), 890-897.

17 von Itzstein M *et al.* (1993) **Rational design of potent sialidase-based inhibitors of influenza virus replication.** *Nature*, **363**, 418-423.

18 Kim CU *et al.* (1997) **Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site.** *J. Am. Chem. Soc.*, **119**(4), 681-690.

19 Grosdidier A *et al.* (2011) **SwissDock.** *Nucleic Acids Res.*, **39**, W270-W277.

20 Morris GM *et al.* (2009) **AutoDock4 & AutoDockTools4.** *J. Comput. Chem.*, **30**(16), 2785-2791.

21 Trott O & Olson AJ. (2010) **AutoDock Vina.** *J. Comput. Chem.*, **31**(2), 455-61.

22 Daina A *et al.* (2014) **iLOGP.** *J. Chem. Inf. Model.*, **54**, 3284-3301.

23 Daina A & Zoete V. (2016) **A BOILED-Egg to predict gastrointestinal absorption & brain penetration of small molecules.** *ChemMedChem*, **11**(11), 1117-1121.

24 Gfeller D *et al.* (2013) **Shaping the interaction landscape of bioactive molecules.** *Bioinformatics*, **29**(23), 3073-3079.

25 Gfeller D *et al.* (2013) **SwissTargetPrediction.** *Nucleic Acids Res.*, **42**(W1), W32-W38.

26 Gfeller D & Voete V. (2015) **Protein homology reveals new targets for bioactive small molecules.** *Bioinformatics*, **31**, 2721-7.

27 Oprea TL *et al.* (2011) **Drug repurposing from an academic perspective.** *Drug Discovery Today: Ther. Strategies*, **8**(3-4), 61-69.

28 Bertolini F *et al.* (2015) **Drug repurposing in oncology-patient & health systems opportunities.** *Nat. Rev. Clin. Oncol.*, **12**, 732-42.

29 Martin YC *et al.* (2002) **Do structurally similar molecules have similar biological activity?** *J. Med. Chem.*, **45**(19), 4350-4358.

30 Bender A & Glen RC. (2004) **Molecular similarity: a key technique in molecular informatics.** *Org. Biomol. Chem.*, **2**, 3204-18.

31 Muegge I & Mukherjee P. (2016) **An overview of molecular fingerprint similarity search in virtual screening.** *Expert Opin. Drug Discov.*, **11**(2), 137-148.

32 Cereto-Massague A *et al.* (2015) **Molecular fingerprint similarity search in virtual screening.** *Methods*, **71**, 58-63.

33 Willett P *et al.* (1998) **Chemical similarity searching.** *J. Chem. Inf. Comput. Sci.*, **38**(6), 983-996.

8 Acknowledgements & funding

GOBLET Practical Guides build on GOBLET's Critical Guide concept, using layout ideas from the Higher Apprenticeship specification for college-level students in England. These contents expand on materials made freely available by the SIB Swiss Institute for Bioinformatics (**www.drug-design-workshop.ch/cox.php**).

We thank Studio KO and the Atelier Poisson for their contribution to Figures 1 and 4, respectively.

This Guide was developed with the support of a donation from EMBnet to the GOBLET Foundation. The work was supported by SIB, Swiss Institute of Bioinformatics and by the Swiss National Science Foundation through Agora grant CRAGP3_171653 (to VZ and AD).

Design concepts and cover image contributed by CREACTIVE.

9 Licensing & availability

The Guide is freely available for download via the GOBLET portal (**www.mygoblet.org**), EMBnet website (**www.embnet.org**) and the F1000Research Education and Training Collection (**f1000research.com/collections/bioinformaticsedu?selectedDomain=documents**).

10 Disclaimer

Every effort has been made to ensure the accuracy of this Guide; GOBLET cannot be held responsible for any errors/omissions it may contain, and cannot accept liability arising from reliance placed on the information herein. Examples given here are for educational purposes only; they do not give instruction on drug usage.

About the organisations

GOBLET

GOBLET (Global Organisation for Bioinformatics Learning, Education & Training; **www.mygoblet.org**) was established in 2012 as a not-for profit foundation to unite, inspire and equip bioinformatics trainers worldwide; its mission, to cultivate the global bioinformatics trainer community, set standards and provide high-quality resources to support learning, education and training.

GOBLET's ethos embraces:

- ***inclusivity****: welcoming all relevant organisations & people*
- ***sharing****: expertise, best practices, materials, resources*
- ***openness****: using Creative Commons Licences*
- ***innovation****: welcoming imaginative ideas & approaches*
- ***tolerance****: transcending national, political, cultural, social & disciplinary boundaries*

For general enquiries, contact **info@mygoblet.org**. Further information can be found in the following references:

- Attwood *et al.* (2015) **GOBLET: the Global Organisation for Bioinformatics Learning, Education & Training.** *PLoS Comput. Biol.*, **11**(5), e1004281.
- Corpas *et al.* (2014) **The GOBLET training portal: a global repository of bioinformatics training materials, courses & trainers.** *Bioinformatics*, **31**(1), 140-142.

EMBnet

EMBnet, the Global Bioinformatics Network, is a non-profit organisation, founded in 1988 to establish and maintain bioinformatics services in Europe. Eventually expanding beyond European borders, EMBnet created an international network to support and deliver bioinformatics services across the life sciences: **www.embnet.org**.

Since its foundation, EMBnet has had a keen interest in Education and Training (E&T), and has delivered tutorials and courses worldwide. Perceiving a need to unite and galvanise international E&T activities, EMBnet was a principal founder of GOBLET. For more information and general enquiries, contact **info@embnet.org**.

SIB Swiss Institute of Bioinformatics

The SIB Swiss Institute of Bioinformatics, a non-profit foundation, was created in 1998 as a federation of Swiss bioinformatics research and service groups, and was one of the founders of GOBLET. Its mission is to lead and coordinate the field of bioinformatics in Switzerland. Its data-science experts work to advance biological and medical research, and enhance health: **www.sib.swiss**.

CREACTIVE

CREACTIVE, by Antonio Santovito, specialises in communication and Web marketing, helping its customers to create and manage their online presence: **www.gocreactive.com**.

About the authors

Teresa K Attwood (orcid.org/0000-0003-2409-4235)[†]

Teresa (Terri) Attwood is emeritus Professor of Bioinformatics with more than 25 years' experience teaching bioinformatics in BSc, MSc and PhD programmes, and in courses and summer schools, in the UK and abroad. She wrote the first bioinformatics text-book; her third book was published in 2016:

- Attwood TK & Parry-Smith DJ. (1999) ***Introduction to Bioinformatics.*** Prentice Hall.
- Higgs P & Attwood TK. (2005) ***Bioinformatics & Molecular Evolution.*** Wiley-Blackwell.
- Attwood TK, Pettifer SR & Thorne D. (2016) ***Bioinformatics challenges at the interface of biology and computer science: Mind the Gap.*** Wiley-Blackwell.

Marie-Claude Blatter (orcid.org/0000-0002-7474-1499)[§]

Marie-Claude Blatter leads the SIB Swiss Institute of Bioinformatics' outreach activities. She coordinates and participates in in-house and external teaching, including activities for high-schools and the public, and bioinformatics BSc and MSc programmes. She's also involved in user support of UniProt.

Vivienne Gerritsen (orcid.org/0000-0003-3759-9494)[§]

Vivienne Baillie Gerritsen is a biologist by training, and has been involved in popular science for many years. She is a science writer for SIB, and author of the *Protein Spotlight* articles (**www.proteinspotlight.org**).

Antoine Daina (orcid.org/0000-0002-0074-7601)[‡]

Antoine Daina is in charge of developing and applying novel CADD methods in the SIB's Molecular Modelling Group. His research includes chemoinformatics and machine-learning models for physicochemistry, pharmacokinetics and ADME. He's also involved in developing educational tools for high-school students and the public.

Vincent Zoete (orcid.org/0000-0002-2336-6537)[‡,¥]

Vincent Zoete is one of the Group Leaders of the SIB's Molecular Modelling Group, co-leads SIB's SwissDrug-Design project, and is Assistant Professor in Molecular Modelling at the University of Lausanne, Ludwig Lausanne Branch. He is also active in the field of education, notably in CADD.

- Blatter MC *et al.* (2016) **The Metagenomic Pizza: a simple recipe to introduce bioinformatics to the layman.** *EMBnet J.*, **22**, e864.
- Daina A *et al.* (2017) **Drug Design Workshop: a web-based educational tool to introduce computer-aided drug design to the general public.** *J. Chem. Educ.*, **94**(3), 335–344.
- Daina A *et al.* (2018) **Educational tools to introduce CADD to students & to the public at large.** *Chimia*, **72**(1), 55-61.
- Gerritsen VB *et al.* (2017) **Bioinformatics on a national scale: an example from Switzerland.** *Brief. Bioinform.*, bbx073.

Affiliations

[†]The University of Manchester, Manchester (UK); SIB Swiss Institute of Bioinformatics, [§]Geneva (SwissProt Group), [‡]Lausanne (CH); [¥]Lausanne University, Lausanne (CH).