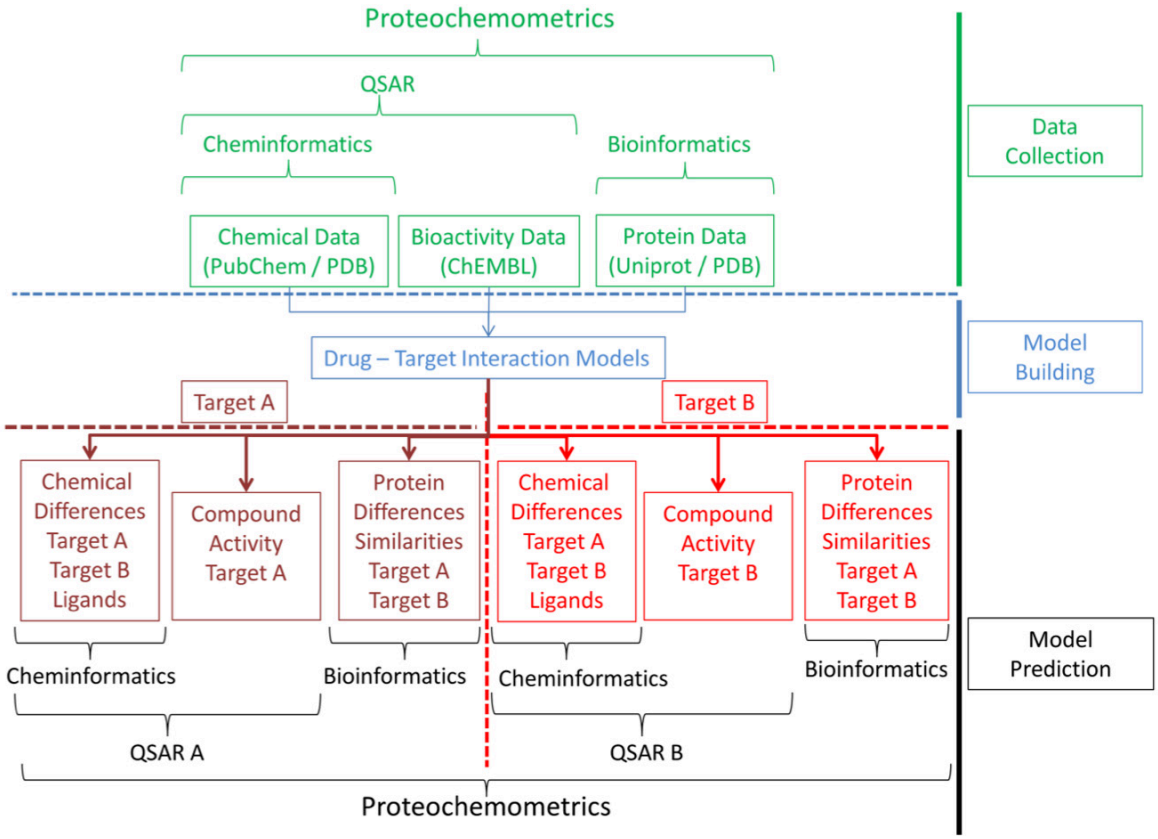
# Déjà Vu - Réjà Vu : on knowledge-based approaches linking ligand and target information to bioactivity

Westen, G.J.P. van

# Chapter 1

# General Introduction



# Contents

1.1 About this thesis. ..... 9
1.2 Chemistry ..... 9
1.2.1 Chemicals and Man. ..... 9
1.2.2 Small Molecules. ..... 9
1.2.3 Chemical Space. ..... 10
1.2.4 Molecular Similarity. ..... 10
1.3 Biology ..... 11
1.3.1 Genomics ..... 11
1.3.2 Proteomics. ..... 12
1.3.3 (Drug) Target Space. ..... 12
1.3.4 Protein Similarity. ..... 12
1.4 Bioactivity ..... 13
1.4.1 Chemistry and Biology ..... 13
1.4.2 Exponential Data Growth. ..... 14
1.5 Exponential Computational Power Growth ..... 15
1.5.1 Smaller and smaller. ..... 15
1.6 Bioinformatics and Cheminformatics ..... 16
1.6.1 Computers in Medicinal Chemistry. ..... 16
1.6.2 Bioinformatics. ..... 16
1.6.3 Cheminformatics. ..... 18
1.7 Current Computational Bioactivity Modeling ..... 20
1.7.1 No standardized tools. ..... 20
1.7.2 Quantitative Structure-Activity Relationships (QSAR). ..... 20
1.7.3 Classification versus Regression. ..... 21
1.7.4 Validation of QSAR models. ..... 21
1.7.5 Classification Validation. ..... 22
1.7.6 ROC Plot. ..... 22
1.7.7 Regression Validation. ..... 23
1.8 Structural Methods ..... 24
1.8.1 X-ray Crystallography. ..... 24
1.9 Combination of Computational Methods With ‘wet’ Experiments ..... 26
1.9.1 Reliability issues. ..... 26
1.10 Informatics Approaches for Bioactivity Data ..... 26
1.10.1 Proteochemometric modeling. ..... 26
1.10.2 Statistical Methods ..... 27
1.10.3 Structural Methods. ..... 28
1.11 Aims of this thesis ..... 28
1.12 References ..... 29

## 1.1 About this thesis.

This thesis focuses on computational approaches able to combine data from different disciplines that are relevant in medicinal chemistry and drug discovery. These different disciplines, or data sources, are: Chemistry, Biology and Bioactivity and will be further explained below. The underlying rationale is that these disciplines are *complementary to* each other rather than *substitutes for* each other. Therefore we expect models created on data from the combination of these disciplines to be more robust than models created from data obtained from a single discipline.

## 1.2 Chemistry

**1.2.1 Chemicals and Man.** Chemicals have long since been recognized to be able to directly affect human beings. For example, at the turn of the 19$^{th}$ and 20$^{th}$ century Hans H. Meyer and Charles E. Overton both published a similar theory stating that the narcotic potency of an anesthetic can be predicted from its solubility in oil.[1, 2] While this is a very crude relationship, it can be considered one of the first attempts to correlate chemical features ("solubility") of a compound with its biological activity ("narcotic potency").

**1.2.2 Small Molecules.** In the 20$^{th}$ century a specific class of chemicals has become dominant as a source of drugs in pharmaceutical research, this class is called 'small molecules'. Small molecules are chemical compounds that are, as the name has it, relatively small. Often this size is expressed as molecular weight, where a molecular weight of approximately 500 Dalton or less is considered to be small while sometimes a limit of 800 Dalton is considered.[3] They are also required to be organic and are often relatively simple to synthesize. It is these properties that make them suitable as drugs and therefore they have been the major focus of pharmaceutical companies and medicinal chemistry research.

**1.2.3 Chemical Space.** With the majority of chemists focusing on the production of either novel small molecules or analogues of existing small molecules, it is not surprising that the total number of known small molecules has risen tremendously over the last years. However, the total amount of known molecules (chemical space) is not even approaching a perceptible fraction of the total number of possible small molecules (estimated at $10^{60}$ compounds).[4, 5] Nonetheless, as structures and properties of compounds can be stored electronically with relative ease; so called virtual libraries (collections of millions of small molecules that are theoretically possible) have appeared and can serve as a source of ideas for small molecules that are actually synthesized.

Simultaneously with virtual libraries, public databases have materialized. Public databases are similar to virtual libraries since both databases and virtual libraries contain electronically stored molecules. However, public databases differ from virtual libraries as databases contain molecules that *have been* synthesized and sometimes tested for biological activity. One of the largest free databases containing small molecules is Pubchem.[6]

**1.2.4 Molecular Similarity.** With the appearance of virtual libraries, a need arose to quantify their similarity (how similar is compound *A* to compound *B*). This principle is demonstrated in **Figure 1.1**. For any pair of compounds the similarity can be quantified between 0 and 1 based on the presence of absence of chemical features. When these compounds are aligned along the edges of a matrix, clusters of similar compounds appear. Given a molecule that exhibits sought properties, molecules that have similar properties can be included in the search for novel drugs. Likewise this similarity measure can be extended to three compounds or an entire virtual library.

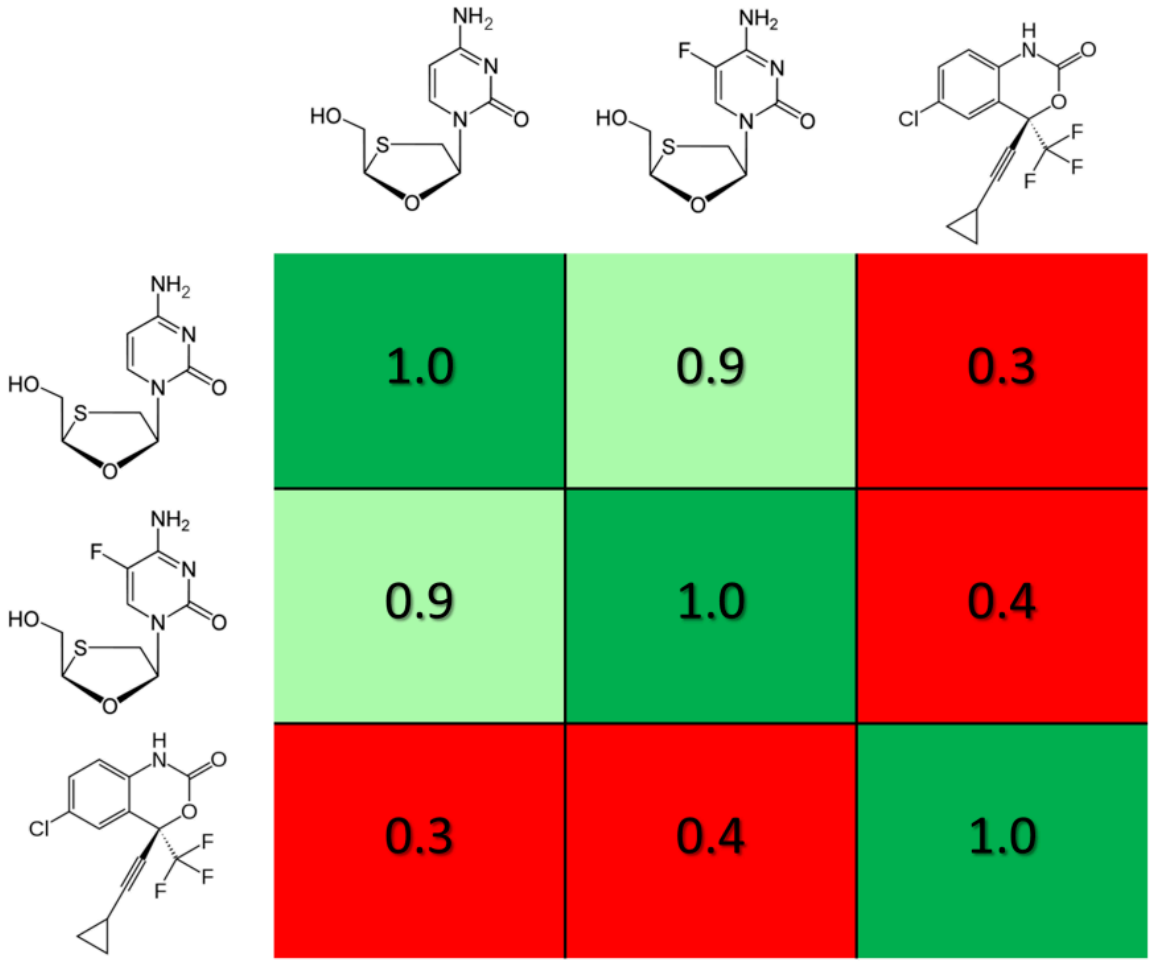



**Figure 1.1: The concept of molecular similarity**. For any pair of compounds the similarity can be quantified between 0 and 1 based on the presence of absence of chemical features. When these compounds are aligned along the edges of a matrix, clusters of similar compounds appear (green squares). Given a molecule that exhibits sought properties, molecules that have similar properties can be included in the search for novel drugs, while compounds that are different (red squares) are avoided.

## 1.3 Biology

**1.3.1 Genomics.** Like the field of chemistry, the field of biology, more specific the field of molecular biology, has flourished over the late 20$^{th}$ and early 21$^{st}$ centuries. Molecular biology has its roots in genetics and biochemistry. The field studies data gathered at a genetic level about gene expression and maps possible functions of these genes onto proteins. The large scale study of genetic data is also known as genomics, the study of an entire genome of an organism. The advent of genomics from molecular biology has introduced numerous techniques for large scale data storage, manipulation and data mining, the process of pattern discovery in large unsorted data sets. Computational genomics approaches use computational analysis methods to obtain these goals.

**1.3.2 Proteomics.** Proteomics is the large scale study of proteins consisting of experimental procedures, large scale data collection and much more. Proteomics links data gathered using genomics to proteins. In the scope of the current thesis proteomics is defined much narrower. Proteomics is of interest as computational approaches have been developed to process the large data sets that are produced on a routine basis, much like in Genomics. One of the largest free databases containing protein sequence information is Uniprot.[7] In addition to sequence information, structural information, elucidating the three dimensional structure of proteins, is also publically available. Structural information is stored in the Protein Data Bank (PDB).[8] It is both sequence information and structural information we are interested in within the scope of this thesis.

**1.3.3 (Drug) Target Space.** Chemistry provides a framework wherein research on small molecule drugs takes place, chemical space. Likewise, the output of genomics and proteomics provides a framework wherein research on novel proteins that can be of interest in medicinal chemistry takes place, the so-called target space. Both spaces are complementary and the advent of large scale public databases has made it possible to mine the data sets that underlie them.

**1.3.4 Protein Similarity.** Similar to the concept of molecular similarity (see **1.2.4**), a concept of protein similarity exists. Protein similarity can be defined on many different levels, from similarity of the three dimensional structure of two or more proteins, to the similarity between a certain region of interest which can be present on two or more proteins. This region of interest can for instance be the location where small molecules bind to the protein ('binding pocket'). The latter definition of protein similarity will be used throughout this thesis. An example is given in **Figure 1.2** using three hypothetical three amino acid peptides (but this approach can easily be up scaled to full proteins). Here, for each pair of peptides the similarity was quantified based on the physicochemical properties of the side chains and the peptides that are most similar cluster together. The rationale is as follows: given a protein that is of interest due to a specific function it performs, similar proteins can be identified which might also be capable of performing a similar function and can hence be also of interest.

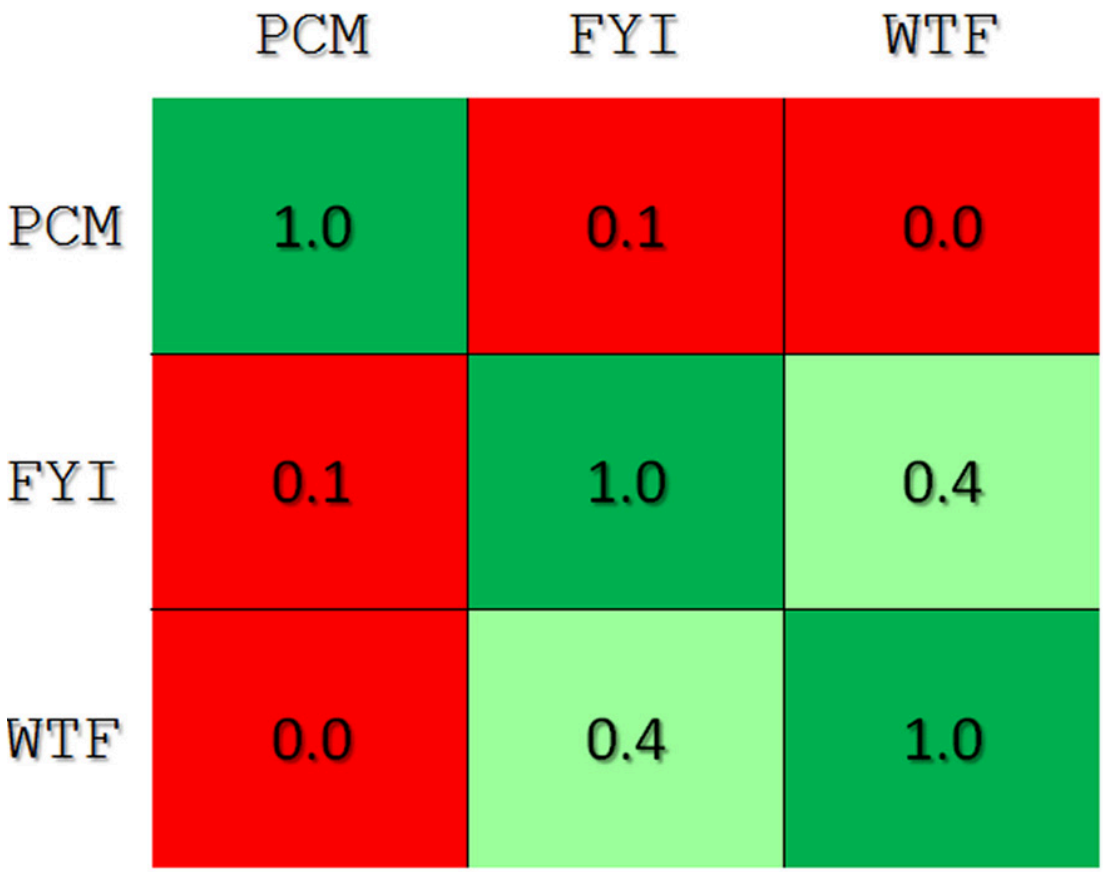



**Figure 1.2: The concept of protein similarity.** For any pair of peptides the similarity can for instance be quantified between 0 and 1 based on the physicochemical characteristics of the amino acid side chains. When these peptides are aligned along the edges of a matrix, clusters of similar peptides appear (green squares). Given a protein that is of interest due to a specific function it performs, similar proteins can be identified which might also be capable of performing a similar function and can hence be also of interest while dissimilar proteins (red squares) are avoided.

## 1.4 Bioactivity

**1.4.1 Chemistry and Biology.** Bioactivity is information about the effects of chemicals on living organisms. In this thesis we will focus on the effect of small molecules on proteins, bioactivity quantifies these effects. The combination of the study of chemical space with the study of target space is what makes the rational development of bioactive compounds possible. Bioactivity data has joined the ranks of molecular biology data and chemical data as publically available information. Recently bioactivity data has been included in Pubchem moreover, there is the advent of ChEMBL,[6, 9] a database completely focused on bioactivity of small molecules.

**1.4.2 Exponential Data Growth.** As public databases have become available, their use has also become commonplace. This has led to a nearly exponential growth in the size of the explored regions of chemical space and target space. This growth will be illustrated using the PDB over the course of its existence as a reference but is equally relevant for other databases (**Figure 1.3A**). More and more data presents scientist with the unique opportunity to start data mining for specific or global bioactivity by linking bioactivity data from sets of related targets and hence possible sets of related molecules that interact with these targets ("ligands"). Effectively, the increase in data enables rational design of desired bioactivity profiles. However, traditional methods are focused on a single target and small analogue series and are ill equipped to mine data on sets this size.

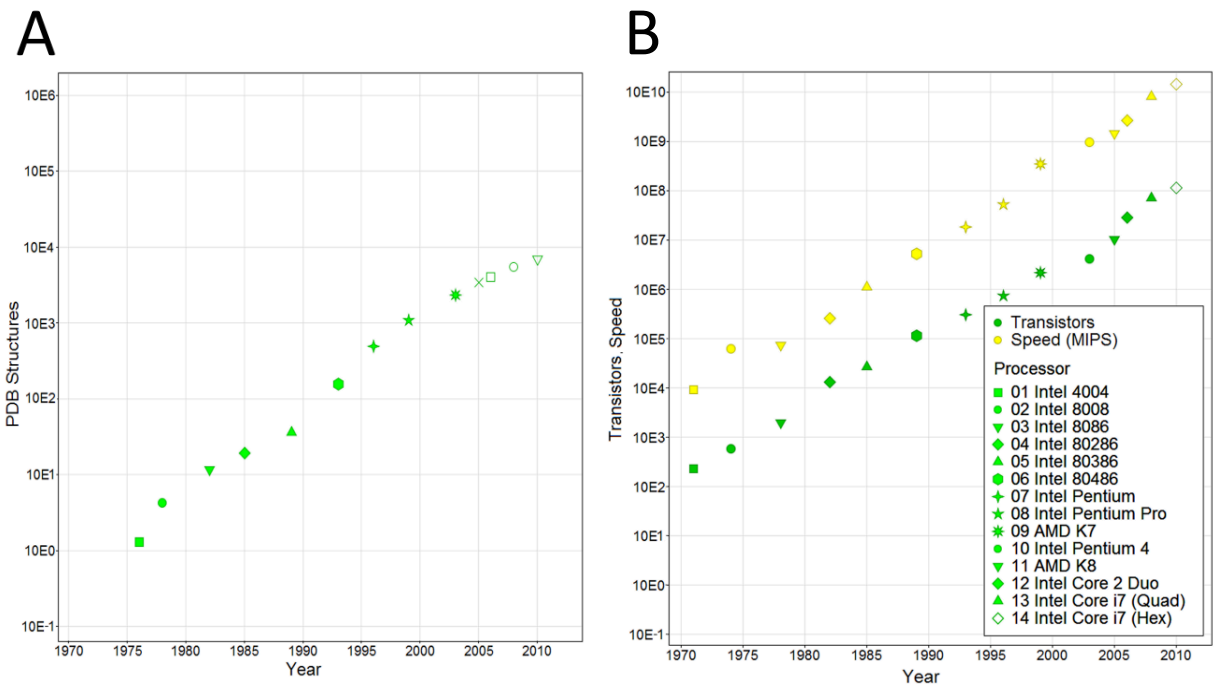



**Figure 1.3: Data growth and processing power growth** (A). The amount of structures available (y-axis) in the Protein Data Bank [8] from 1976 until 2010 (x-axis)(B). The increase in computing power (in Million Instructions Per Second) and transistor amount [10-14] (y-axis) of CPUs in desktop computers. Both Y-axes are drawn on a logarithmic scale.

## 1.5 Exponential Computational Power Growth

**1.5.1 Smaller and smaller.** The exponential growth in data drives a need for data analysis and fuels drug research. However, an exponential growth in the number of scientists cannot be sustained, [15, 16] neither can an exponential growth in research budget.[16, 17] Hence the increase in data analysis capacity needs to come from a different source. It is here that the exponential growth of computing power, available to standard desktop PCs, can prove instrumental (**Figure 1.3B**). One of the main driving forces behind the increase of transistors on a single CPU, hence the increase in speed is the large decrease in minimal feature width. This started as large as 10 µm in 1971 in the Intel 4004, down to 32 nm in the 2010 Intel Core i7 (Hexacore) (**Figure 1.4**).[18] Also shown is the average size of an HIV Virion particle (145 nm),[19] Staphylococcus Aureus bacterium (800 nm) [20] and a red blood cell (90 µm).[21]

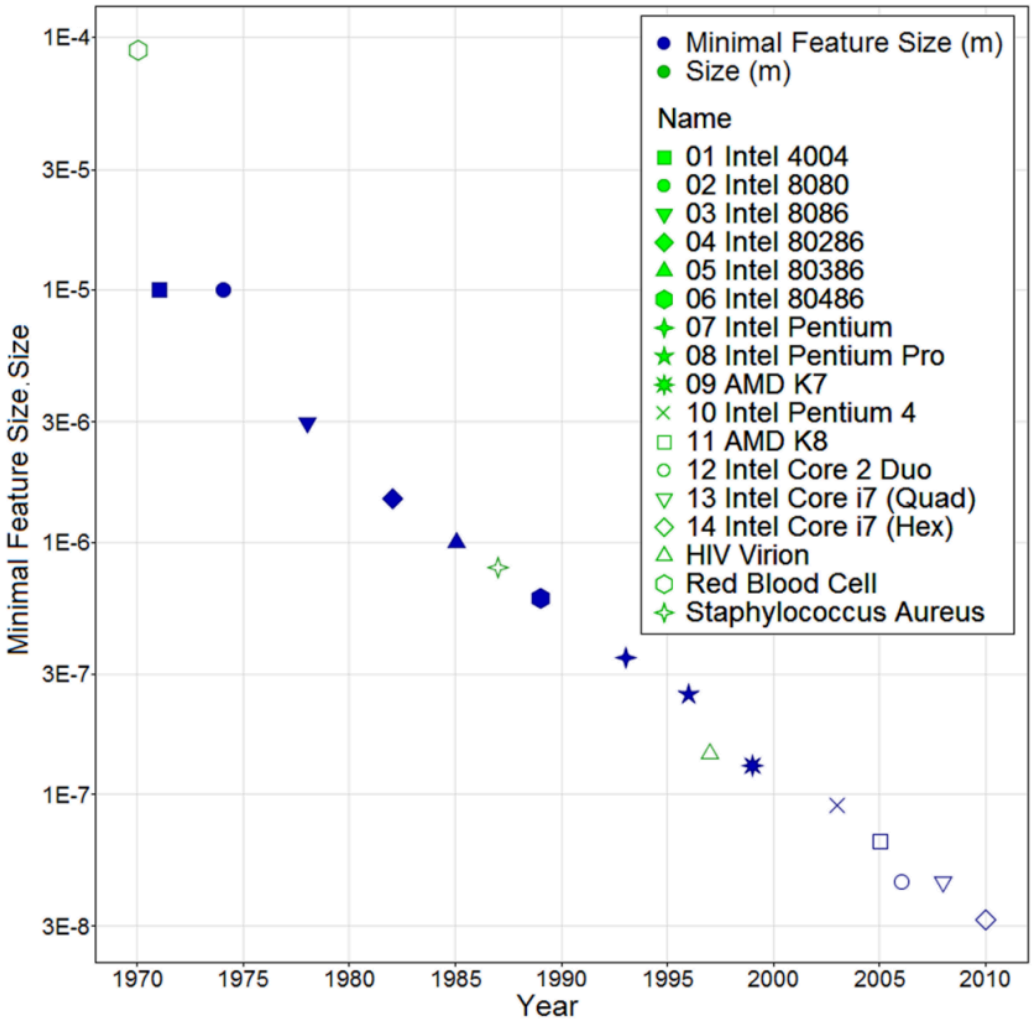



**Figure 1.4: Minimal feature width on integrated circuits since 1971 until 2010.**[10] Also shown are the average sizes of an HIV Virion particle, a human red blood cell, and a Staphylococcus Aureus bacterium.

## 1.6 Bioinformatics and Cheminformatics

**1.6.1 Computers in Medicinal Chemistry.** The use of computers in drug research is not novel. Their role has been invaluable in several parts of the drug design process. Among these is the creation of statistical models that explain the interaction of a small molecule to a target, so called Quantitative Structure-Activity Relationships (QSARs), but computers are also involved in X-ray crystallography. More recently the field of bioinformatics, [7, 22, 23] and cheminformatics have gained solid ground.[24] Both techniques deal with processing large amounts of data (hence the informatics suffix). These concepts will be explained below.

**1.6.2 Bioinformatics.** Bioinformatics combines biological information (nucleotide sequences, amino acid sequences) with computational techniques and here primary applications are storing, retrieving, and evaluating ("mining") data (**Figure 1.5**). Therefore Bioinformatics can be seen as the informatics extension to (molecular) biology and one of the major tools to navigate target space. In order for bioinformatics approaches to function, the data available needs to be transformed into information that is accessible for the computer. In practice this involves structuring the data, standardizing measurements, standardizing descriptive parameters and most important of all removing false information or noise.

After this information has been processed, the output from the computer needs to be interpreted to make it useful information accessible for scientists involved in a research project. This involves creation of plots illustrating possible correlations / inverse correlations or the use of specialized tools. Bioinformatics output can lead to insights about details of cellular modifications that are applied to proteins, or can lead to discovery of similarities between certain sets of proteins.

A simplified example application of a Bioinformatics approach to a dataset consisting of two targets (proteins) is shown in **Figure 1.5**. The scheme distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). Bioinformatics can predict protein differences, protein similarities, and protein properties in general. Data sources in this case can be databases like Uniprot or the PDB.

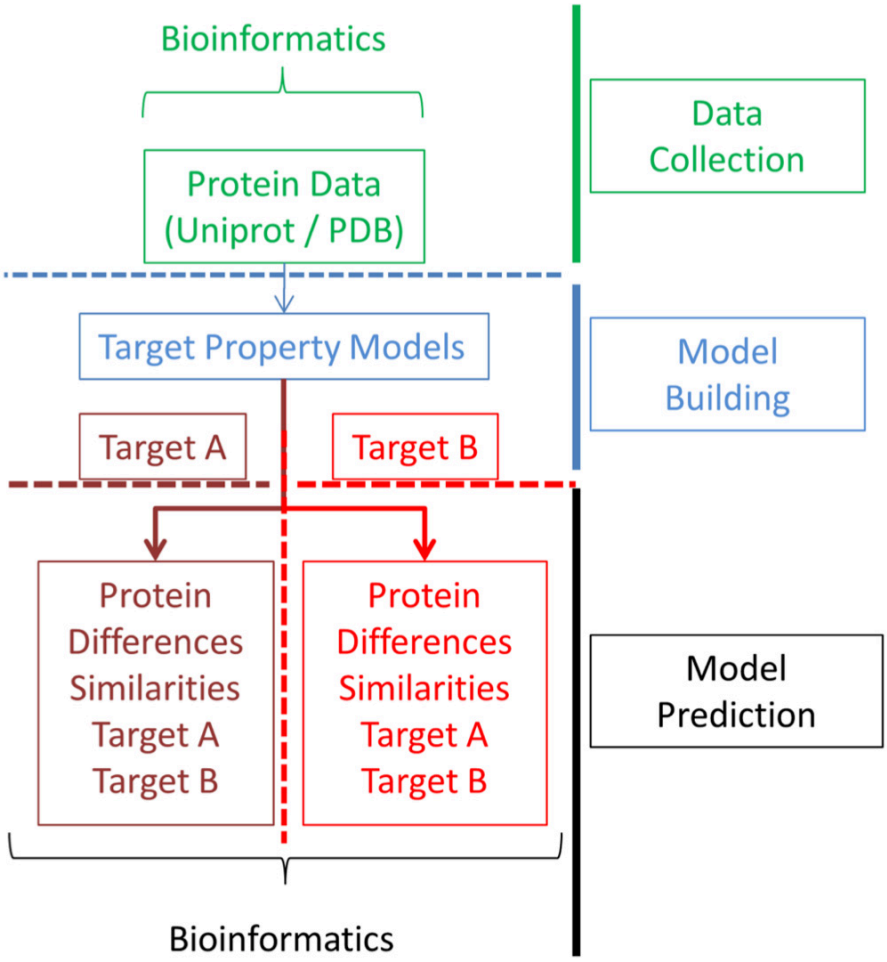



**Figure 1.5: Simplified schematic overview of a bioinformatics project applied to a dataset consisting of two targets (proteins).** The scheme distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). To be able to distinguish between the different targets, they are visualized in different shades of red. See text for further details.

**1.6.3 Cheminformatics.** Cheminformatics combines chemical information with computational techniques and primary applications are storing, retrieving, and data mining of chemical information. Cheminformatics can therefore be seen as the informatics extension to chemistry and one of the major tools to navigate chemical space. Like in bioinformatics, data needs to be structured, standardized and cleared of noise.

Cheminformatics encounters hurdles very similar to bioinformatics, namely data interpretation and presentation of data in an organized fashion to other scientists involved in a research project. Hence specialized tools have been developed to retrieve compounds with desired properties or to visualize a correlation that might not be apparent on first sight.

A simplified example application of a Cheminformatics approach to a dataset consisting of two targets is shown in **Figure 1.6**. This scheme also distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). Cheminformatics can predict chemical differences, chemical similarities, and chemical properties of ligands in general. Data sources in this case can be databases like Pubchem or the PDB. Also shown is the application of statistical QSAR on the same data set, please see below for further details.

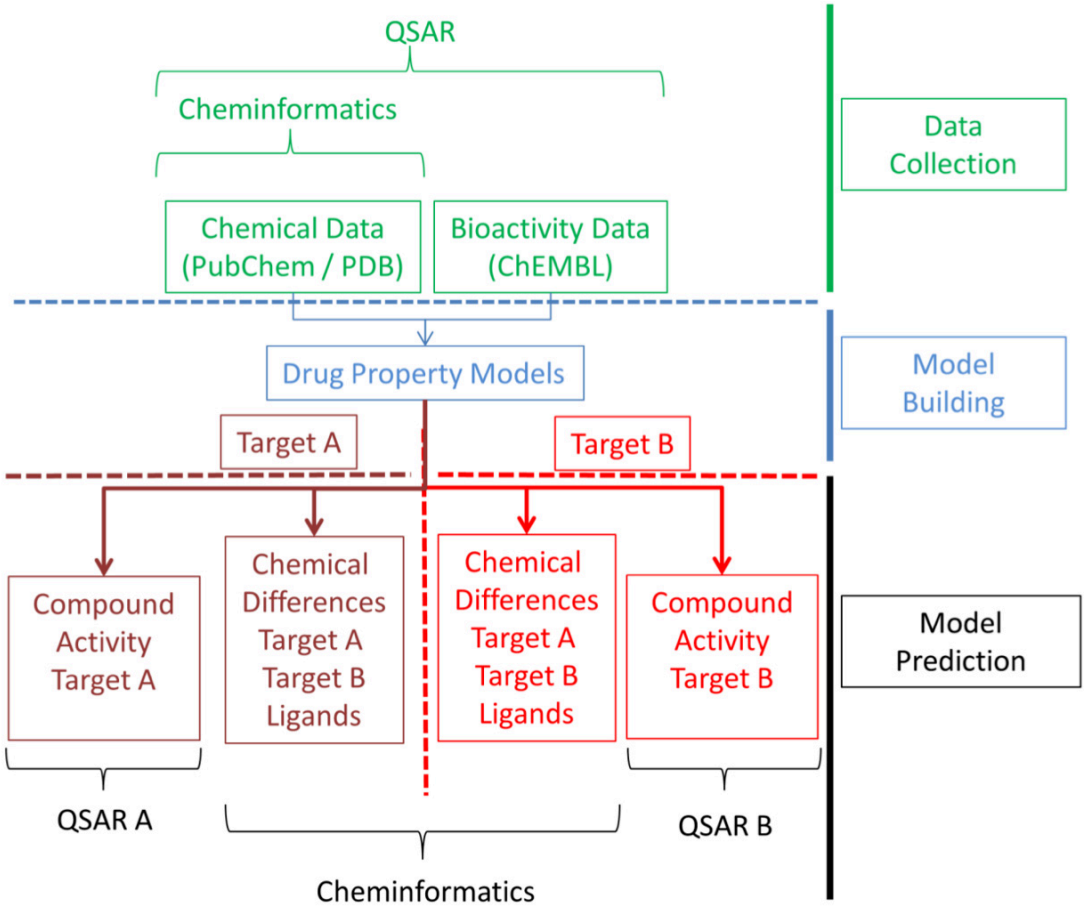



**Figure 1.6: Simplified schematic overview of a cheminformatics project applied to a dataset consisting of two targets.** The scheme also distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). Two statistical QSAR approaches to the same dataset are also shown. To be able to distinguish between the different targets, they are visualized in different shades of red. See text for further details.

## 1.7 Current Computational Bioactivity Modeling

**1.7.1 No standardized tools.** For bioactivity data, no general accepted method to process and mine large amounts of data currently exists. There are several methods to mine bioactivity data available, buts most are focused on small scale datasets. A rough distinction can be made between statistical methods like QSAR and more structural methods (see section **1.8** and **chapter 7**).

**1.7.2 Quantitative Structure-Activity Relationships (QSAR).** The term Quantitative Structure-Activity Relationship is often used to refer to one of two concepts. It can be a quantitative structure-activity relationship focused on a closely related ("congeneric") series of compounds, where the QSAR is driven by relatively slight variations in substitution patterns of the compounds. This is the classical QSAR concept. However, QSARs can also be statistical models that aim to explain the driving forces behind small molecules (ligands) interacting with a given protein (target). In this thesis, QSAR will mean the latter definition. As they are statistical models they do not provide a mechanistic explanation and rely heavily on machine learning. In this light, QSAR models are created ("trained") on a set of compounds for which the activity is known ("training set"). This chemical structure of the compounds in this training set is transformed into a way it can be processed by the computer (usually a numeric description of the compound called 'descriptors',[25] see **chapter 2** for further explanation). After a validation they can then be applied to a set of compounds for which the activity is not known ("test set") to discover novel compounds that display an activity on the target of interest.

Nevertheless, there are some limitations to this approach. Firstly, QSAR models are only able to make valid predictions for compounds that are at least similar to compounds in the training set (the so called applicability domain).[26, 27] This similarity is usually expressed based on the descriptor used to train the model and limited by the chemistry of compounds that have been previously tested on the target of interest. Furthermore, they can only make valid predictions for a single target since they are constructed on the chemical structures of a series of ligands and target information is not present. Finally, the quality of the QSAR (and hence the reliability of the predictions) is defined by the quality of the training set.

A simplified example application of a QSAR approach to a dataset consisting of two targets is shown in **Figure 1.6**. This dataset requires two separate QSAR models to be trained, one for each target. Each QSAR can subsequently predict the activity of ligands on that target. Data sources in this case can be databases like Pubchem or ChEMBL. The PDB can also be used, but needs to be combined with a database like ChEMBL to retrieve the actual bioactivity of known ligands.

**1.7.3 Classification versus Regression.** In machine learning an algorithm can be trained to predict one of two output variable types, each of which will be described below.

The first predicts a class as output variable (classification), in the simplest case the algorithm then decides whether or not the untested ligands belongs to either the 'active' class or the 'inactive' class based on the chemical resemblance to the training set of both the active compounds ('active' class) and the inactive compounds ('inactive' class). However, classification can also be performed using more than two classes and is limited only by computational power and memory.

Secondly, machine learning can predict a numeric output variable for an untested ligand (regression). In regression the predicted value (which can be any numeric value e.g. $pK_i$, $pEC_{50}$, $pIC_{50}$, etc.) is also calculated based on the similarity to the training set and the tested values for the output variable to be predicted. The advantage of regression over classification is that it provides a comparison between untested compound A and untested compound B. Once one of these untested compounds has a higher affinity it is presumed to be more active, whereas in classification both would be 'active'.

**1.7.4 Validation of QSAR models.** As outlined above, QSAR models can predict the activity of untested compounds on certain targets of interest. However, as they rely on statistics, these statistics have to be validated before any meaningful prediction can be made. Due to the different nature of classification and regression models, they require a different form of validation.

**1.7.5 Classification Validation.** In classification-based QSAR validation can be performed by analyzing the fraction of tested compounds that are classified correctly by the model rendering parameters like 'sensitivity', 'specificity', 'positive predictive value', negative predictive value'.[28] Each of these parameters outputs a value between 0 (poor prediction) and 1 (perfect prediction). In addition there is the Matthews correlation coefficient (MCC),[29] which aims to combine all four parameters in a single score between -1 (inverse prediction) and 1 (perfect prediction), in the case of the MCC being '0' also indicates a poor prediction (**Figure 1.7**).

To arrive at these numeric values, the predictions are divided into 4 types of predictions: True Positives (TP, compounds that are tested active and also predicted to be active), True Negatives (TN, compounds that are tested inactive and also predicted to be inactive), False Positives (FP, compounds that are tested inactive but predicted active) and False Negatives (FN, compounds that are tested active but predicted inactive).

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

**Figure 1.7: Definition of some of the validation parameters used to quantify the quality of classification-based QSAR models.** TP stands for True Positives (compounds correctly predicted to be bioactive), TN stands for True Negatives (compounds correctly predicted to not be bioactive), FP stands for False Positives (compounds incorrectly predicted to be bioactive), and FN stands for False Negatives (compounds incorrectly predicted to not be bioactive).

**1.7.6 ROC Plot.** A number of machine learning algorithms, like Support Vector Machines (SVM), are capable of producing a ranking, indicating the likelihood that a predicted compound belongs to the class it is categorized in. This ranking can be used to create a Receiver Operator Characteristic (ROC) plot. In this plot the y-axis denotes the TP rate and the x-axis denotes the False Positive FP rate. To maximize the number of true positives a tradeoff can be made allowing additional false positives. When the ranked predictions are then plotted, the resulting plot shows the tradeoff for every possible threshold.[30] This curve provides a graphical interpretation of model performance (**Figure 1.8A**).

**1.7.7 Regression Validation.** In regression models, validation parameters can be calculated directly from the differences between the measured value and the predicted value for a certain compound as the model provides a numeric value. Usually a standard correlation coefficient ($R^2$) is calculated along with the Root Mean Squared Error (RMSE). However, most models are expected to predict a value for the output variable for a compound which is near identical to the measured value, an ideal model therefore predicts data points on a line that intersects the origin (0,0). Hence it is recommended to also calculate the $R_0^2$, the correlation coefficient while taking into account that the line should intersect the origin.[31]

In addition to calculating quantifiable values, the measured and modeled values for all compounds are also plotted in a scatter plot. This plot provides a direct intuitive overview of the model performance (**Figure 1.8B**); from the plot model bias (biased over- or under-prediction) becomes apparent.

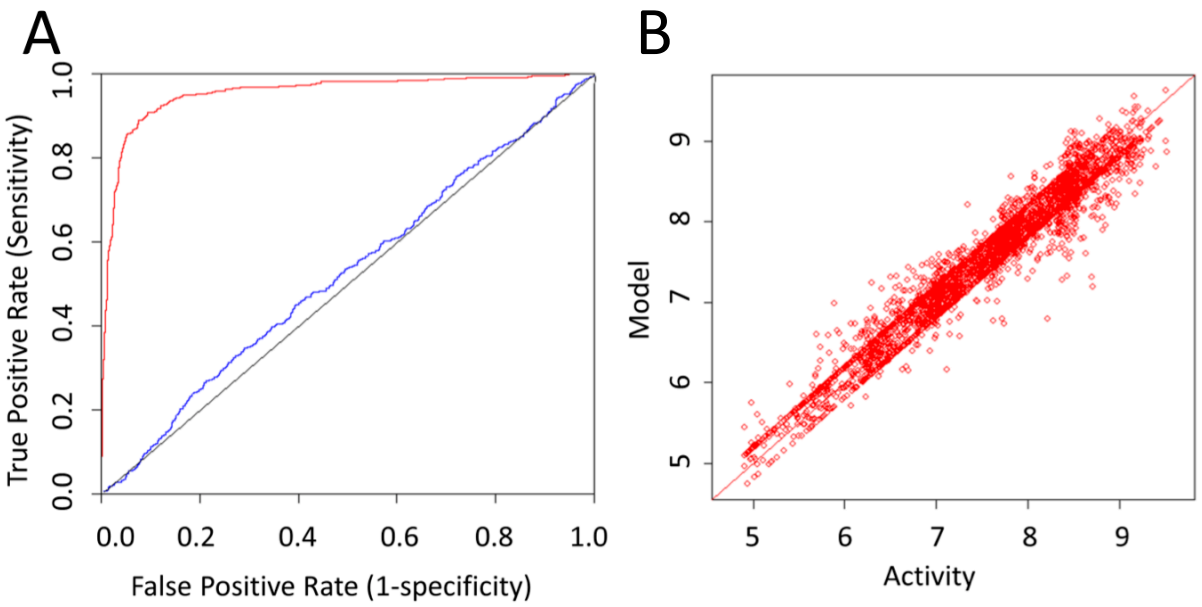



**Figure 1.8: Validation plots in QSAR validation.** (A) Classification validation using an ROC curve. The red line shows a highly predictive model reaching a high TP rate before trading off allowing more FPs. The blue line shows a poor model performing not much better than random showing a roughly equal TP and FP rate. (B) Regression validation using a measured versus predicted scatter plot. The scattering pattern of the plot already intuitively gives an impression of model performance. In addition $R_0^2$, $R^2$ and RMSE provide a quantifiable estimate for model performance.

## 1.8 Structural Methods

**1.8.1 X-ray Crystallography.** While X-ray crystallography is not a computational method itself, it is the most important data source for structure-based approaches and hence will also be introduced here. X-ray crystallography can be applied to both small molecules, entire proteins and proteins with a ligand bound. The technique uses the crystalline form of the analyte. Hence it is important that the analyte can be crystalized, which can be a major hurdle. Subsequently the analyte is subjected to a monochromatic beam of X-rays. These rays scatter as they travel through the analyte as they possess the correct wavelength to be scattered by the electron cloud of an atom (in the order of magnitude of Ångström, $10^{-10}$ m).[32] During this process the analyte is rotated.[32] From the angles and intensities of the scattered rays a crystallographer can produce the electron density map of the analyte. Subsequently, the mean position of atoms in the crystal and the bond orders between them can be determined via mapping onto this electron density map (**Figure 1.9**).

The high resolution information (down to a resolution of 1.4 Å) provides an excellent starting point for structure-based drug design. This is especially true when a known ligand is present in the crystal structure so that the part of the protein involved in the interaction ("binding pocket") is known from its orientation. Not only can this binding pocket be used to define protein similarity (**1.3.4**), complementary information about forces driving the interaction ("pharmacophoric information") can subsequently be obtained from both the binding pocket and the ligand. This information is the starting point for the design of novel ligands. However, a protein is not a static object; it is in fact very dynamic.[33, 34] Therefore a crystal structure can only be seen as a snapshot of one of the states a protein can exist in, but it does not provide any information about the number of other possible states the protein can exist in. However, this flexibility can be of great importance in drug design (See **chapter 7** for further details).[33, 35]

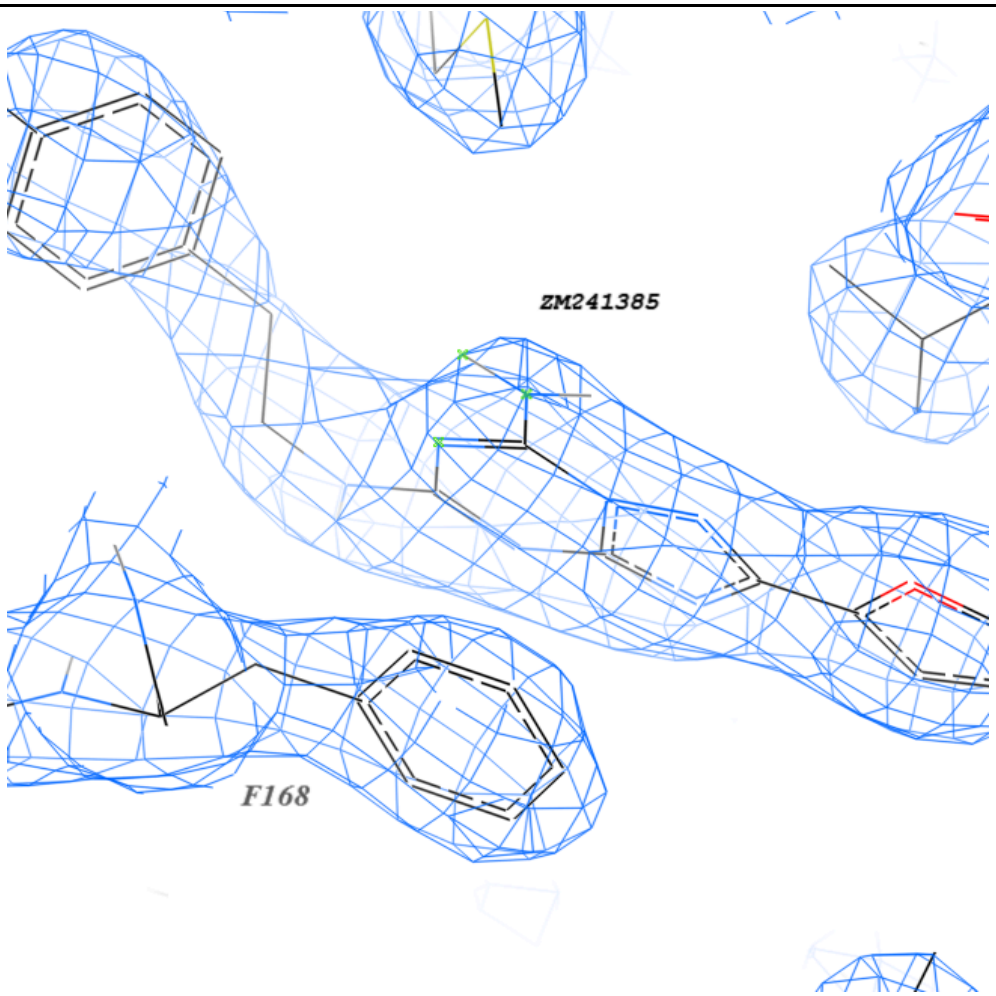



**Figure 1.9: Electron density from PDB structure 3EML visualized as a grid in blue.**[36] Shown within the density are the atoms that were mapped within this density. Both an amino acid from the protein (Phenylalanine, F168) and part of the ligand (ZM-241385) are visualized.

## 1.9 Combination of Computational Methods With 'wet' Experiments

**1.9.1 Reliability issues.** Computational methods have always been met with some reservations. While it is generally acknowledged that computational tools can present an unbiased view of the available data, it has also been shown that blindly following algorithms can lead to expensive experimental failures. However, over the last years the methods have become better accepted in existing research programs. One of the crucial factors to a good integration is reliable predictions, because if one cries wolf too often the computational chemist will lose all credibility. The key to prevent false positives, or rather to minimize the chances of FPs, is to perform a small scale validation of the computational approach that accurately simulates the way it will be integrated in existing research programs. This means that it is absolutely essential to prospectively validate at least some of the predictions made by any model using an actual wet experiment. It also means that is essential to disregard any model unless TPs are identified in such a prospective approach.

## 1.10 Informatics Approaches for Bioactivity Data

**1.10.1 Proteochemometric modeling.** As mentioned before, for bioactivity data, no generally accepted method to process large amounts of data exists. An illustration of the reasons why current methods are ill equipped to deal with large amounts of bioactivity data will be outlined below. An overview of the different computational methods mentioned before is given in **Figure 1.10**, again in the form of a scheme for a medicinal chemistry project involving two targets.

The cheminformatics approaches focus on chemical data; therefore these methods can make predictions about chemical properties relevant for different targets. Bioinformatics focusses on protein data; therefore these methods can make predictions about the differences between targets. QSAR links chemical data to bioactivity data and can therefore make predictions about the activity of compounds on a single target and also rationalize chemically why compounds are active.

Finally there is a relatively young method, proteochemometric (PCM) modeling,[37] which uses all three types of data. Hence it can make predictions about compound activity on multiple targets. Furthermore it can rationalize *why* a compound is active based on chemistry (features of small molecules) or based on biology (features of the proteins).

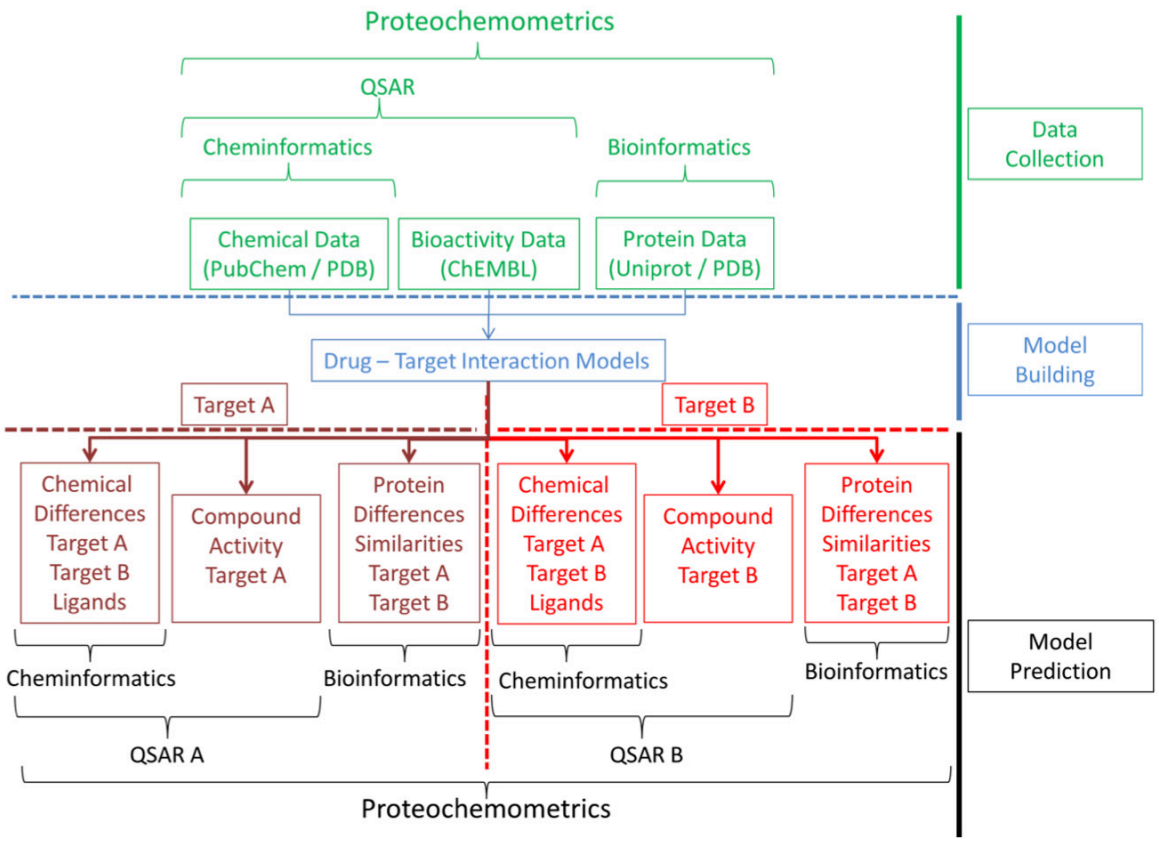



**Figure 1.10: The different computational data analysis methods mentioned in this thesis, the data these methods process and their relationships to one another.** This figure shows a schematic approach in the case of a medicinal chemistry project which involves two targets (deemed A and B). The scheme distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). To be able to distinguish between the different targets, they are visualized in different shades of red. See text for further details.

**1.10.2 Statistical Methods.** Statistical bioactivity modeling traditionally focusses around a single target. While some cheminformatics approaches have been introduced to combine several targets from a chemical point of view,[38] they do not possess the predictive abilities of classical QSAR approaches like affinity prediction and chemical interpretation of the SAR. Moreover, a single QSAR model is unable to explain selectivity e.g. in the case of the adenosine receptor subfamily

To explain selectivity using QSAR, individual QSAR models have to be constructed for each target. The differences between these QSAR models can then explain selectivity. However, this approach is already laborious for small groups of targets. Furthermore, the multiple QSAR approach cannot extrapolate to new (related) targets as it requires the creation of a novel QSAR model for that target. Hence, we cannot use this method to virtually identify small molecules that are active on a novel target as knowledge of compounds active on that target is required.

Therefore improvement of current methods is essential before bioactivity data can be processed on a large scale. An ideal approach would be able to use data gathered via bioinformatics (e.g. protein sequence) and combine that with chemical data (e.g. small molecules known to be active on a related target). PCM modeling might solve these shortcomings of QSAR as major bioactivity modeling approach. PCM is reviewed in **chapter 2** of this thesis.

**1.10.3 Structural Methods.** Structural methods are an advantageous tool to explain in detail what drives the interaction between a ligand and a target. However, due to the high level of detail they require a relatively strenuous interpretation. Furthermore, the interaction relies on both features from the ligand and from the target. Therefore a single structure is not always representative for all ligands that can bind a target. In addition, the aforementioned protein flexibility is difficult to extract from a single structure. Thus a need exists to partially automate the processing of structures, keeping unique features of interactions between a single ligand and target while at the same time also highlighting features that are shared between several ligands.

## 1.11 Aims of this thesis

In this thesis we want to investigate and benchmark new and recent techniques that are equipped to process bioactivity data on a large scale. These techniques should be able to link sets of related targets and ligands and therefore investigate the ligand – target space. As such PCM is a good (statistical modeling) candidate and it will be investigated in **chapters 2, 3, 4, 5**, and **6**.

**Chapter 2** contains a literature review of PCM and other similar approaches carrying a different name. It highlights previous work, application areas and pitfalls. Subsequently **chapter 3** contains an extensive investigation of the physicochemical space of the natural amino acid side chains, and introduces four novel protein descriptors which we have developed and consequently applied in **chapters 4** and **5**. It also investigates co-variation between previously published amino acid descriptors and studies the ability of the descriptors to create bioactivity models.

**Chapter 4** demonstrates a preclinical application of PCM, the goal being the discovery of novel hits (ligands) that are active on one or more targets (the adenosine receptors). **Chapter 5** highlights a later phase in drug discovery, namely the lead optimization stage. This chapter shows how PCM can be used to select the optimal candidate from a group of compounds that best inhibits a group of targets. Lastly, **chapter 6** focusses on a clinical application of PCM.

The technique is used to identify the optimal treatment regimen for individual patient based on the dominant viral genotype they are infected with. Summarizing, **chapters 4 - 6** cover several major phases in drug discovery, and the role PCM can play in that particular phase.

In order to also investigate more structural approaches we introduce in **chapter 7** 'Consensus Structures' and a technique to investigate the ligand – target space using crystal structures. Consensus structures are very similar to PCM as they rely on the combination of ligand information and target information, but differ as they rely on structural information rather than statistics.

Finally **chapter 8** contains general conclusions drawn from the thesis and future perspectives. In this chapter the focus is not so much on PCM but rather on computational methods in drug discovery in general.

Of the here mentioned approaches we will also characterize the limitations and possible pitfalls along with the ability to use these techniques on public data sets as they have become available. In this thesis we have been working both on G Protein-Coupled Receptors (GPCRs) and enzymes. Together these two classes are representative for most drug targets.

## 1.12 References


1. H. Meyer; *Zur theorie der alkoholnarcose.* Arch. Exp. Pathol. Pharmakol.; *1899*. **42**: 109-118.
2. E. Overton; *Studien über die narcose, zugleich ein beitrag zur allgemeinen pharmakologie.* Jena, Gustav Fisher *1901*. **45**: 195.
3. C.A. Lipinski, F. Lombardo, et al.; *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv. Drug Delivery Rev.; *2001*. **46** (1–3): 3-26.
4. P. Kirkpatrick and C. Ellis; *Chemical space.* Nature; *2004*. **432** (7019): 823-823.
5. C. Lipinski and A. Hopkins; *Navigating chemical space for biology and medicine.* Nature; *2004*. **432** (7019): 855-861.
6. E.E. Bolton, Y. Wang, et al.; *PubChem: Integrated Platform of Small Molecules and Biological Activities;* in *Annual Reports in Computational Chemistry*; A.W. Ralph and C.S. David; Editors. *2008*; Elsevier. p. 217-241.
7. E. Jain, A. Bairoch, et al.; *Infrastructure for the life sciences: design and implementation of the UniProt website.* BMC Bioinformatics; *2009*. **10** (1): 136-155.

8. H.M. Berman, J. Westbrook, et al.; *The Protein Data Bank* Nucleic Acids Res.; *2000*. **28**: 235-242.
9. A. Gaulton, L.J. Bellis, et al.; *ChEMBL: a large-scale bioactivity database for drug discovery.* Nucleic Acids Res.; *2011*. **40**: D1100 - D1107.
10. Intel Corporation. *Microprocessor Quick Reference Guide*. 2012 [cited 2012 January 5]; Available from: http://www.intel.com/pressroom/kits/quickreffam.htm.
11. Wikipedia contributors. *Instructions per second*. 2012 10 December 2011 [cited 2012 January 5]; Available from: http://en.wikipedia.org/wiki/Instructions_per_second.
12. J. Culver. *CPU Shack*. 2012 [cited 2012 January 19]; Available from: http://www.cpushack.com.
13. M. White; *Intel Historical Timeline.* Processor; *2003*. **25** (29): 9.
14. G. Shvet. *CPU World*. 2012 [cited 2012 January 18]; Available from: http://www.cpu-world.com.
15. D. Goodstein; *The Big Crunch*; in *NCAR 48 Symposium1994*: Portland.
16. M.F. Perutz; *Will biomedicine outgrow support?* Nature; *1999*. **399**: 299-301.
17. H. Moses, E.R. Dorsey, et al.; *Financial Anatomy of Biomedical Research.* JAMA: The Journal of the American Medical Association; *2005*. **294** (11): 1333-1342.
18. S.M. Sze; *Semiconductor devices: physics and technology*. 2nd ed.*2009*; New York: Wiley.
19. J.A.G. Briggs, T. Wilk, et al.; *Structural organization of authentic, mature HIV-1 virions and cores.* EMBO J.; *2003*. **22** (7): 1707-1715.
20. A. Touhami, M.H. Jericho, and T.J. Beveridge; *Atomic Force Microscopy of Cell Growth and Division in Staphylococcus aureus.* J. Bacteriol.; *2004*. **186** (11): 3286-3295.
21. G. Gulliver; *Observations on the sizes and shapes of the red corpuscles of the blood of vertebrates, with drawings of them to a uniform scale, and extended and revised tables of measurements.* Proceedings of the Zoological Society of London; *1875*: 474-495.
22. The UniProt Consortium; *Ongoing and future developments at the Universal Protein Resource.* Nucleic Acids Res.; *2011*. **39** (suppl 1): D214-D219.
23. W.M. David; *Bioinformatics: sequence and genome analysis*. *2004*; New York: Cold Spring Harbor Laboratory Press.
24. B. Frank K; *Chapter 35. Chemoinformatics: What is it and How does it Impact Drug Discovery;* in *Annu. Rep. Med. Chem.*; A.B. James; Editor *1998*; Academic Press. p. 375-384.
25. N. Nikolova and J. Jaworska; *Approaches to Measure Chemical Similarity – a Review.* QSAR Comb. Sci.; *2003*. **22** (9-10): 1006-1026.

26. H. Dragos, M. Gilles, and V. Alexandre; *Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models.* J. Chem. Inf. Model.; *2009*. **49** (7): 1762-1776.

27. L. Eriksson, J. Jaworska, et al.; *Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs.* Environ. Health Perspect.; *2003*. **111** (10): 1361-1375.

28. P. Baldi, S. Brunak, et al.; *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics; *2000*. **16** (5): 412-424.

29. M. B.W; *Comparison of the predicted and observed secondary structure of T4 phage lysozyme.* Biochimica et Biophysica Acta (BBA) - Protein Structure; *1975*. **405** (2): 442-451.

30. J. Fogarty, R.S. Baker, and S.E. Hudson; *Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction*; in *Proceedings of Graphics Interface2005*; Canadian Human-Computer Communications Society: Victoria, British Columbia. 129-136.

31. A. Tropsha; *Predictive Quantitative Structure-Activity Relationships Modeling;* in *Handbook of Chemoinformatics Algorithms*; J. Faulon and A. Bender; Editors. *2010*.

32. B. Rupp; *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. 1st ed.*2009*; New York: Garland Science. 800.

33. H.A. Carlson; *Protein flexibility and drug design: how to hit a moving target.* Curr. Opin. Chem. Biol.; *2002*. **6** (4): 447-452.

34. H.A. Carlson and J.A. McCammon; *Accommodating Protein Flexibility in Computational Drug Design.* Mol. Pharmacol.; *2000*. **57** (2): 213-218.

35. K. Das, J.D. Bauman, et al.; *High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: Strategic flexibility explains potency against resistance mutations.* Proc. Natl. Acad. Sci. U. S. A.; *2008*. **105** (5): 1466-1471.

36. V.P. Jaakola, M.T. Griffith, et al.; *The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist.* Science; *2008*. **322** (5905): 1211-1217.

37. M. Lapinsh, P. Prusis, et al.; *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochim. Biophys. Acta, Gen. Subj.; *2001*. **1525** (1-2): 180-190.

38. D.E. Gloriam, S.M. Foord, et al.; *Definition of the G Protein-Coupled Receptor Transmembrane Bundle Binding Pocket and Calculation of Receptor Similarities for Drug Design.* J. Med. Chem.; *2009*. **52** (14): 4429-4442.