Molecular Modeling and Computer Aided Drug Design. Examples of their Applications in Medicinal Chemistry

F. Ooms*

Facultés Universitaires Notre-Dame de la Paix. Chimie Moléculaire Structurale, Namur, Belgium

Abstract: The development of new drugs with potential therapeutic applications is one of the most complex and difficult processes in the pharmaceutical industry. Millions of dollars and man-hours are devoted to the discovery of new therapeutic agents. As, the activity of a drug is the result of a multitude of factors such as bioavailability, toxicity and metabolism, rational drug design has been utopian for centuries. Very recently, impressive technological advances in areas such as structural characterization of biomacromolecules, computer sciences and molecular biology have made rational drug design feasible.

The aim of this review is to give an outline of studies in the field of medicinal chemistry in which molecular modeling has helped in the discovery process of new drugs. The emphasis will be on lead generation and optimization.

Introduction

Design, development and commercialization of a drug is a tedious, time-consuming and cost-intensive process [1]. The cost of this process has increased significantly during the past thirty-four years. Industry averages reported to the Pharmaceutical Manufacturer’s Association, have shown that the cost of drug development has increased from $4 million in 1962 to over $350 million in 1996 (Fig. 1). Between 1960 and 1980, the development time of a substance from the first synthesis to its introduction on the market, has almost quadrupled and has remained relatively unchanged since 1980 with a present time period of 9-13 years [2-5]. Moreover, during this process, only a small amount of candidates will be examined in the clinic and few will be marketed. In 1950, it was estimated that 7,000 compounds had to be isolated or synthesized and then tested for therapeutic activity for each one that became a pharmaceutical product. The challenge is becoming more difficult : 10,000

Fig. (1). The cost of drug development from $4 million in 1962 to over $350 million in 1996.

* Address correspondence to this author at the Facultés Universitaires Notre-Dame de la Paix, Laboratoire de Chimie Moléculaire Structurale, Rue de Bruxelles, 61, B-5000 Namur, BELGIUM; Phone: 00-32-81724569; Fax: 00-32-81724530; e-mail: ooms@scf.fundp.ac.be

© 2000 Bentham Science Publishers B.V.
compounds had to be evaluated in 1979, and this number could be as high as 20,000 today (Fig. 2). The reasons for this are several-fold. The market for so-called high value-added compounds is very competitive. The new compound must offer improved characteristics in order to be worthwhile for commercialization. Also there are serious hurdles regarding ease and cost of synthesis, patentability, safety, and social need for the new compound.

Considering both the potential benefits to human health and the enormous costs in time and money of drug discovery, any tool or technique that increases the efficiency of any stage of the drug discovery enterprise will be highly prized. Computer-aided drug design (CADD) is one of these tools which can be used to increase the efficiency of the drug discovery process. CADD cannot, however, maximize its utility in isolation and will not do so. Rather, it can form a valuable partnership with experiment by providing estimates when experiments are difficult, expensive, or impossible, and by coordinating the experimental data available. A close coupling between computational chemists and experimentalists allows information to flow immediately and directly between the two. This helps CADD chemists to better understand the details of the problem and to refine their approach. It also provides valuable information for the experimentalist, it helps to guide further experimental planning and potentially makes this process more efficient. CADD is, however, not a direct route to new drugs, but it provides a somewhat more detailed map to the goal. The hope is that by providing bit and pieces of information, CADD will help to save days and money for drug discovery projects.

The aim of this review is to give an outline of strategies used in CADD and examples of studies in the field of medicinal chemistry in which CADD has helped in the discovery process of new drugs. This review is divided into two major sections: the lead discovery and the lead optimization.

### CADD Strategies in the Drug Discovery Process

Strategies for CADD vary depending on the extent of structural and other information available regarding the target (enzyme/receptor) and the ligands. “Direct” and “indirect” design are the two major modeling strategies currently used in the drug design process [6]. In the indirect approach the design is based on comparative analysis of the structural features of known active and inactive compounds. In the direct design the three-dimensional features of the target (enzyme/receptor) are directly considered (Fig. 3).

**Fig. (3).** Four major cases in CADD also known as "direct" and "indirect design when the structure of the target is respectively known or unknown.
CADD in Lead Generation

3D Structure of the Protein Unknown

In the early stage of a drug discovery process, researchers may be faced with little or no structure-activity relationship (SAR) information. At this point, assay development and screening should be undertaken immediately by the high-throughput screening (HTS) group [7], and chemists should immediately follow up on any screening leads or other sources of initial information. The compounds screened could be commercially available, natural products [8], collections of in-house synthesized compounds or emerge from combinatorial libraries [9]. Computational chemists can, however, help in the choice of the compounds to be selected for HTS. Instead of performing random screening, a set of compounds presenting diversity in their physiochemical properties can be selected to find leads. The aim of these analyses is to select and test fewer compounds, whilst gaining as much information as possible about the dataset [10]. Any reduction of the number of compounds to be tested, while only reducing the amount of redundancy within a database without introducing any voids, should have an important impact on research efficiency and the costs associated [11, 12]. Recently, the use of rational design to maximize the structural diversity of database, for lead findings and refinements, was investigated. Hierarchical clustering and maximum dissimilarity methods were compared to a random approach in order to study their efficiency for the diversity enhancement of three-dimension databases. The investigations were done using two-dimensional fingerprints as a validated molecular descriptor and the performance of rational selection methods vs random approach has been compared [13].

If, however, a lead is known, then more focused approach can be adopted by searching for compounds with similar (two or three-dimensional) structures to the lead candidate or by substructure searching. In substructure searching the query will retrieve those structures from the database that contain groups present in the primary lead. These molecules can then be screened in a biological assay. As an example, substructure searching has been used to identify a potent tyrosine kinase inhibitor (IC_{50}=40nM), starting from an initial lead that was, itself, identified via three-dimensional searching [14].

Once primary leads and their corresponding structural information are or become available the computational chemist can use these data to derive new lead classes and ‘fine tune’ the leads that the chemists have already been pursuing.

Table 1. Main Stereoelectronic Properties used in CADD

<table>
<thead>
<tr>
<th>Steric</th>
<th>L (Substituent length)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BS (Substituent width)</td>
</tr>
<tr>
<td></td>
<td>MR (Molar refractivity)</td>
</tr>
<tr>
<td></td>
<td>Volume</td>
</tr>
<tr>
<td></td>
<td>Surface area</td>
</tr>
</tbody>
</table>

| Electronic | σ (Hammet constant)    |
|           | F, R (Field and resonance parameters) |
|           | pK_a (Ionization constants) |
|           | q (atomic charges)       |
|           | MEP (Molecular Electrostatic Potential) |

| Lipophilic | π (Hansch constant)    |
|           | f (Hydrophobic fragmental constant) |
|           | Log P (partition coefficients) |
|           | Log k_w (capacity factor values from RP-HPLC) |
|           | CLOGP (calculated log P values) |
|           | MLP (Molecular Lipophilic Potential) |

| H-bonding | HA (number of H-bond acceptors) |
|          | HD (number of H-bond donors)    |
|          | Δlop P (oct-hex) (H-bond capability) |

The first step to derive a new lead, also called secondary lead, will be to study the stereoelectronic properties of the selected primary leads (table 1). The primary leads should be (1) selected among a set of compounds showing a large variety in chemical structures, and (2) interact with the same target via the same binding mechanism. By comparison of the stereoelectronic properties of primary leads, a pharmacophore is defined. A pharmacophore model is a spatial arrangement of atoms or functional groups believed to be responsible for biological activity [15]. In this model the rest of the molecule acts as a skeleton to hold the groups in the right place. Typically, the derived pharmacophores consist, generally, of 3-5 features, and the distances between them (angles and other geometric measures are sometimes used) (Fig. 4) [6].

In the course of the pharmacophore identification process, two different steps have to be taken in succession. First a conformational analysis has to be carried out. Indeed, the biological activity of a drug is

\[
\begin{align*}
  d_1 &= 2.3-3.0 \text{ Å} \\
  d_2 &= 3.3-5.7 \text{ Å} \\
  d_3 &= 2.3-2.7 \text{ Å}
\end{align*}
\]

\[
\begin{align*}
  d_{12} &= 1.3-3.6 \text{ Å} \\
  d_{23} &= 2.2-3.7 \text{ Å} \\
  d_{13} &= 2.2-2.7 \text{ Å}
\end{align*}
\]

Fig. (4). Example of a pharmacophore model derived for the MAO-B inhibitors. With X = O, N, S.
supposed to depend on one unique conformation hidden among all the low-energy conformations [16]. Only the so-called bioactive conformation can bind to the specific macromolecular environment at the active site of the target protein. The search for this bioactive conformation among a set of compounds is one of the major tasks in computational chemistry. It has to be noted that the bioactive conformation is not necessarily identical with the lowest-energy conformation but on the other hand, it cannot be a high-energy conformation [17, 18]. Several recent publications and reviews have dealt with the techniques used to derive bioactive conformations [19, 20].

During the second step, the bioactive conformations are used to calculate their stereoelectronic properties. Indeed, the receptor has to recognize whether an approaching molecule possesses the properties necessary for specific and tight binding. This recognition process occurs at rather large distances and precedes the formation of the final interaction complex. At rather large distances, the three-dimensional electrostatic field surrounding each molecule plays a crucial role in recognition while at closer distances between the interacting surfaces other molecular characteristics like hydrogen bonds or hydrophobicity come into play. To perform this analysis, molecular interaction potentials such as (1) the Molecular Electrostatic Potential (MEP) [21], (2) the Molecular Lipophilic Potential (MLP) [22], or the Hydrophobic Interaction Potential (HINT) [23], and the (3) Molecular Interaction Fields Potential (MIP); these fields describe the variation of interaction energy between a target molecule and a chemical probe like water, CH or nitrogen; GRID is one of the most widely used programs for investigation of these fields [24]) are used.

Based on this three-dimensional description of the stereoelectronic properties of the considered compounds, a pharmacophore can be obtained. It is essential that as well as active substances, inactive analogues or at least congeners with low activity can also be described by the pharmacophore. Recently, computer programs which automate this deductive process have been developed.

Once the pharmacophore has been, at least, partially identified, the next step is to find compounds which contain it embedded in their structure by three-dimension database searching [25]. Prerequisites for effective three-dimensional searching are large databases of three-dimensional structures and suitable software to perform this search. The classic database of three-dimensional molecular structures is the Cambridge Structural Database (CSD) [26]. It contains evaluated X-ray and neutron diffraction data for more than 190,000 compounds (small-molecules and polymers). Other commercial three-dimensional databases are available [27]. These commercial databases offer the possibility to use already implemented three-dimensional searching programs.

On the other hand, the user can create his or her own three-dimensional database by converting two-dimensional structures, or “2.5D” structures when the stereoelectrochemical are known, into three-dimension structural information. Research into methods of converting two-dimensional databases into three-dimensional ones has led to the development of a number of software packages (See Sadowski and Gasteiger for a recent review on “automatic” model builders [28]). A comparison of automatic three-dimensional model builders has recently been performed using 639 X-ray structures from the CSD database [29]. Moreover, recent work has been carried out, by using CONCORD [30], on the three-dimensional conversion of more than 400,000 compounds contained in the National Cancer Institute (NCI) two-dimensional database [31]. This three-dimensional database is increasingly, and successfully [32], used in the laboratories’ drug design efforts and can be accessed by internet (http://www.ccc.uni-erlangen.de/services/nci.html). A detailed example of the use of the three-dimensional NCI database will be given in the case study part of this chapter.

More recently, systems taking conformational flexibility into account were developed. The need to perform flexible searches stems from the fact that most molecules in compound collections possess rotatable bonds, and has become possible with hardware advances. At present three main approaches are used:

i) Storing multiple conformations [33-35]
ii) Performing conformational analysis during screen generation and at search time [36]
iii) Performing torsional minimization at search time [37, 38]

Torsional minimization and a genetic algorithm have been found to be efficient for flexible searching, whereas distance geometry and systematic search approaches were found to be too slow for database applications [39].

Conclusion

The use of CADD in lead generation when no structural information about the target protein are available is summarized in figure 5.

Case Studies

Recently, “success stories” have been published in the design and discovery of Human Immunodeficiency
Virus type 1 Integrase (HIV-1 IN) inhibitors [40, 41]. HIV-1 IN is among one of the most important enzyme (with reverse transcriptase and protease) responsible for the HIV-1 replication cycle. HIV-1 IN mediates the integration of HIV-1 DNA into host chromosomal targets and is known to be essential for effective viral replication. Furthermore, no cellular counterpart has been identified. Because of its essential nature in the replicative cycle of HIV-1, HIV-1 IN is an attractive target for the development of anti-AIDS drugs. Starting from a pharmacophore hypothesis derived from a known inhibitor of HIV-1 IN, caffeic acid phenethyl ester (CAPE, 1), a three-dimensional search of the NCI database was performed. From this search, 267 structures were found to match the pharmacophore, 60 of those were tested in an in vitro assay against HIV-1 IN and 19 were found to inhibit both the 3'-processing and strand transfer. The relevance of the proposed pharmacophore was then tested using a small three-dimensional validation database of known HIV-1 IN inhibitors (for a comprehensive review see the review of Neamati, Sunder and Pommier [42]), which had no overlap with the group of compounds found in the initial search. This search strongly supports for the existence of the postulated pharmacophore and in addition, it hinted at the existence of a possible second pharmacophore relevant in the binding to IN [41]. Using the second pharmacophore in a three-dimensional search of the NCI database, 10 novel structurally diverse HIV-1 IN inhibitors were found. Four (2-5) of these 10 inhibitors are particularly potent (IC$_{50}$ < 30 µM) [40]. These studies show that the three-dimensional search technology, coupled with a repository of structures that can be rapidly obtained for testing, is a powerful tool in the development of new leads.

G-protein coupled receptors (GPCR) are another very important drug target and most chemists in a typical medicinal chemistry career will run across a GPCR. To date, no X-ray structure of GPCR is available for structure-based drug design, due to the difficulty to obtain suitable membrane proteins crystals. However, GPCR three-dimensional models can be obtained by homology modeling (For more details on GPCR model construction see references [43-49]) but these are still not enough precise to be used in de novo design. These models are typically used for: (1) to visualize the protein interior and to propose ligand binding modes, (2) to plan mutagenesis experiments, and (3) to support ligand design. Therefore one of the way for designing novel GPCR ligands is still analog-based drug design. Among all the GPCR, the serotoninergic receptors are one of the important targets for central-nervous system (CNS) pathologies (See the review of Gaster and King for the latest developments in serotonin receptor modulation [50]). The multiple
The actions of serotonin (5-HT) are mediated through 14 different receptors and these have been subdivided into 7 distinct groups on the basis of operational pharmacology, sequence analysis and transduction mechanisms. It has to be noted that 5-HT$_3$ receptor is unique in that it is the only one of this receptor family to belong to the superfamily of ion-channel receptors while all the others (5-HT$_1A$, 5-HT$_1B$, 5-HT$_1D$, 5-HT$_1E$, 5-HT$_1F$, 5-HT$_2A$, 5-HT$_2B$, 5-HT$_2C$, 5-HT$_4$, 5-HT$_5A$, 5-HT$_5B$, 5-HT$_6$, 5-HT$_7$) are GPCR.

5-HT$_4$ receptors have during the last years gained interest in medicinal chemistry. The role of 5-HT$_4$ receptors in CNS function has been reviewed and its distinct localization coupled with neurochemical and electrophysiological data suggest an involvement in cognition and anxiety [51, 52]. However, the clinical use of currently available drugs acting at the 5-HT$_4$ receptor has been hampered by their lack of selectivity over 5-HT$_3$ binding site. For this reason, there is considerable interest in the medicinal chemistry of these 5-HT receptor subtypes, and significant effort has been made towards the discovery of potent and selective 5-HT$_4$ ligands.

Recently, CADD has been successfully applied to the design of novel selective 5-HT$_4$ ligands by using a comparative study of the 5-HT$_3$ and 5-HT$_4$ pharmacophores [53]. A conformational analysis performed on active and inactive 5-HT$_3$ and 5-HT$_4$ antagonists has allowed for the proposal of a pharmacophore for both ligands. Regions of steric tolerance and intolerance, based on van der Waals surfaces of active and inactive 5-HT$_3$ and 5-HT$_4$ ligands for both receptors has been proposed. The proposed steric model for the 5-HT$_4$ receptor binding site have received some support from the synthesis of two new active and selective ligands: 6 ($K_i$ (5-HT$_3$) = 3.7 nM; $K_i$ (5-HT$_4$) > 1000 nM) and 7 ($K_i$ (5-HT$_3$) > 10 000 nM; $K_i$ (5-HT$_4$) =13.7 nM).

This study is another example of the use of pharmacophore in the design of novel and selective compounds for a specified target.
3D Structure of the Protein Known

Advances in molecular biology have had a dramatic impact on the drug discovery process [54, 55]. The ability to produce significant quantities of pure protein has led to the structural determination of many biologically relevant targets by X-ray crystallography [56] and nuclear magnetic resonance (NMR) [56, 57]. In most cases, protein crystallography does not determine these structures to atomic resolution, therefore there is some uncertainty in the exact position of each atom. The average errors in the atomic positions depend upon the diffracting quality of the protein crystal. As a general rule of thumb the positional error of atoms amounts to approximately 1/6 of the resolution [58]. Protein structures are available through the Protein Data Bank [59].

Homology modeling can also be used in cases where no experimental proteins structures are available [60]. According to a recent evaluation of homology modeling methods, accurate models, such as those preferred for drug design, can only be obtained in cases in which the two proteins have a high percentage of sequence identity [61].

\[ \Delta G = \Delta H - T \Delta S \]

Fig. (6). Schematic representation showing the components of free-energy changes associated with the formation of the protein-ligand complex.

Given a high resolution structure of proteins, new ligands can be designed *de novo*. The process of forming a complex between a small molecule ligand and a protein is, however, a complex equilibrium (Fig. 6). Ligand binding involves a multiplicity of factors, including changes in intermolecular interactions between the ligand, the solvent water and the target protein, as well as changes in ligand or receptor conformation [62]. However, many promising approaches toward the goal of automated ligand design have been developed (for an overview of the different methods and highlights of their strengths and weaknesses, the reader is referred to references [63, 64]). In general, such programs operate in three broad stages:

**Constraint Definition**

Steric and/or chemical constraints of the design problem are delineated and supplied to the program.

**Structure Generation**

Molecular structures are next assembled, and the designers attempt (1) to find the best position and orientation of a molecule in the binding site of the target, and (2) to meet as many as possible of the imposed constraints.

**Structure Evaluation**

After this step, the binding affinity of the resulting complex is estimated. Binding free energies are extremely difficult to estimate and few of the currently described docking methods attempt this. Docking scoring functions range from extremely computer-intensive theoretical calculations to simple empirical schemes. The field of affinity estimation is too large to be covered in this paper and current approaches to the problem have been recently summarized elsewhere [65, 66].

Because of the need to scan a large number of compounds in an efficient manner, methods that require hours or more, like the free energy perturbation method (FEP) [67], to dock a single ligand, are not practical for searching large chemical database or checking a large number of proposed molecules. Thus, practical scoring functions favor speed over theoretical sophistication in the hope of obtaining a qualitative useful score in a reasonable amount of time. These functions often estimate the free energy of ligand-receptor interactions as a function of hydrophobic contact surface area, number of hydrogen bonds, buried polar surface area, and similar terms. Such methods largely or completely ignore changes in the protein on ligand binding since only the structure of the complex is considered. One of the most widely used scoring function for affinity estimation is LUDI [68, 69], which has inspired several other programs. In this program, the overall free energy of binding being broken down into contributions from hydrogen bonds, ionic interactions, apolar contacts and entropy penalties for fixing rotatable bonds.

A common strategy is to use very simple functions during an early screening step and then use more computer-intensive functions to improve the structures and estimate their affinities.
Case Study

The Agouron group reported the design and evaluation of a novel class of FKBP-12 ligands [70]. While the biological function of FKBP-12 is not completely clear, it has been suggested that it plays a role in protein folding and intracellular transport. The immunosuppressive natural products FK506 8 and rapamycin 9 exert their biological activity indirectly through binding to FKBP-12. In addition to the role of FKBP-12 as an immunosuppressive mediator, it has recently been demonstrated that unnatural ligands to FKBP-12 can be used to regulate gene expression in appropriately genetically engineered cells [71].

Starting from the crystal structure of the FKBP-FK506 complex, from which the ligand (FK506, 8) was removed, the active site of FKBP-12 has been examined using LUDI [68, 69]. Among the many fragments suggested, adamantane appeared to fill the hydrophobic piperolic acid binding site and seemed well suited for further elaboration. From the examination of the docked adamantane, it has been suggested to remove of a methylene bridge with the introduction of a carbonyl group which could allow for a hydrogen bond to the backbone amide NH of Ile-56. A second analyses was then performed using LUDI with the modified ligand in order to fill the FK506 pyran binding site of FKBP-12. LUDI proposed aromatic groups appended through a linker to the bridgehead position of the modified adamantane. Finally, LUDI suggested a hydroxyl group, which could make a favorable contact with Asp-37, in the meta position of an aromatic ring tethered by a methylene linker. Using this template, several compounds were synthesized. Among those compounds the most potent identified (10, Fig. 7), was a 7.9 \( \mu \)M rotamase inhibitor of FKBP-12 [72].

![Fig. (7). 2D structure of the rotamase inhibitor 10 of FKBP-12 derived from the LUDI study.](image)

Another example of the successful use of de novo design is the identification of novel inhibitors of *Pneumocystis carinii* dihydrofolate reductase (DHFR) that are selective versus inhibition of human DHFR [73]. *Pneumocystis carinii* is the principal agent of morbidity and mortality in HIV-infected individuals [74, 75]. Clinically *P. carinii* inhibitors, however, are generally selective for human DHFR [76]. Discovery of novel selective inhibitors of *Pneumocystis carinii* DHFR may possess distinct pharmacological profiles, and avoid sources of toxicity seen with conventional antifolates. The group of Kuntz screened the Fine Chemicals Directory (50,000 compounds) with the DOCK [77] program suite to identify a list of potential selective *P. carinii* DHFR inhibitors. 40 compounds were selected, after a postdocking refinement and manual filtering of DOCK suggestions, their anti-*Pneumocystis* DHFR activity was tested. 13 of the 40 compounds showing IC$_{50}$ values better than 150 µM were further assayed against human DHFR. 10 of the 13 binded preferentially to the fungal enzyme. Among those compounds the most potent identified was a 7 µM inhibitor of *P. carinii* DHFR with 25-fold selectivity. This study represents the first attempts to locate species-selective inhibitors of therapeutic interest. This task was a difficult one because of the great similarity exhibited by the *P. carinii* and human DHFR enzymes.

![Molecule](image)

Recently a pioneering study was published by the group of Fesik [78]. In this work, they elegantly combined the advantage of rational design and combinatorial chemistry by a new procedure called “SAR by NMR”. In the first step of this process, a library of low molecular weight compounds is screened to identify molecules that bind to the protein. Addition of a substrate with sufficient affinity to the $^{15}$N-enriched protein in solution yields a shift of the HSQC NMR signals for all groups near to the binding site. In the next step, once a lead is identified, analogs are screened to optimize binding to this site. Searching for a second binding site is then undertaken in either the original screen or a screen conducted in the presence of the first fragment. The second ligand is then optimized. When the two optimized fragments have been selected, their location and orientation are determined experimentally by NMR or X-ray crystallography. Finally, on the basis of this structural information, suitable linkers for the two ligands are modeled on the computer (Fig. 8).

The advantage of the Fesik approach is that one needs only weak binding for single ligands. Linking such weak binders provides not only the product of binding constants of the single substances but an additional entropic contribution which yields superactive compounds.

A further advantage in this approach, is the possibility of directly investigating allosteric effects. Screening of libraries in the presence of the first ligand allows identification of further ligands that only bind in the presence of the first one.

The “SAR by NMR” procedure has been demonstrated with $^{15}$N-labeled FKBP, which is important for the suppression of the immune response in organ transplantation. Fasik et al. chose two molecules with different but adjacent binding sites. Both compounds bound in the submillimolar range. After an optimization procedure of both ligands, the structure of the ternary complex with FKBP was determined by NMR. Synthesis of only five different

![Diagram](image)

**Fig. (8).** An outline of the SAR by NMR method.
linked compounds yielded nanomolar molecules; the best compound 12 had a $K_d$ of 19 nM.

The same strategy has been used in the rational design of novel, potent bis-phenylamidine carboxylate factor Xa (fXa) inhibitors [79]. In this study, isolated small molecule fragments were docked and minimized in the S1 and S4 aryl binding subsites of the fXa dimer crystal structure. Subsequently, these fragments were connected with a tether, without disturbing the orientation of the fragments in their respective pockets. These modeling studies led to a novel, potent fXa inhibitor 13 ($K_i = 34$ nM).

Combinatorial chemistry and structure-based ligand design have also recently emerged as powerful methods for the discovery and optimization of ligands for a variety of enzymes and receptors [80]. The combination of those two techniques has, recently, led to the discovery of a selective cyclooxygenase-2 (COX-2) inhibitor 14 ($IC_{50} = 1.3 \mu M$), a lead compound for a potentially novel series of anti-inflammatory compounds [81].

**CADD in Lead Optimization**

When leads are available, the next step consists in their optimization. In medicinal chemistry the lead optimization process concerns many aspects such as the optimization of the affinity for the biological target, the toxicity, the oral bioavailability, the cell permeability, the plasma binding, the ease of metabolism. This process requires the synthesis of a series of analogues and testing their biological activities. The principle employed is that any incremental change in the chemical structure produces incremental (positive or negative) changes in bio-activity. A systematic study of such cause and effect relationship is called structure-activity relationship (SAR) study. The process is highly iterative and traditionally based on trial-an-error. Some strategies (Hansch and Leo method [82], Topliss tree [83], Craig plot [84]) have also been advocate to help in focusing on the most informative experiments. CADD
Partial least squares (PLS) can also be used in this process to make it more methodological.

**3D Structure of the Protein Unknown**

When no structural data about the target (receptor/enzyme) is available, the lead optimization process can be made more methodological by using quantitative structure-activity relationship (QSAR) studies. QSAR methods are used to attempt to correlate the biological activities of compounds with their associated stereoelectronic properties (table 1). The aim of such investigations is to produce a suitably robust model capable of reliable predictions for novel chemical species. All QSAR techniques assume that (1) all the compounds being studied bind non covalently to the same biological target; (2) structurally similar compounds are similarly oriented at that common receptor site, and (3) the dynamics of the system can be neglected. Two different approaches can be used in QSAR depending on the available compounds: (1) two-dimensional QSAR (2D-QSAR) and (2) three-dimensional QSAR (3D-QSAR).

**2D-QSAR**

This field was started in the early sixties by the pioneer work of Hansch-Fujita [82] and Free-Wilson [85]. Hansch and Fujita [82] proposed to express the biological activity as a function of molecular or fragmental descriptors:

\[
\text{Biological activity} = f(\text{molecular or fragmental descriptor})
\]

The data used to derive the QSAR equation are assembled into a matrix of numbers in which each row represents the data for a compound and each column a physicochemical property: the descriptor. In 2D QSAR, descriptors are generally substituent constants (π, σ, MR, Es,...) that are assumed to be transferable from one series to another. Large collections of substituent constants have been assembled [86, 87]. A statistical procedure is then used to find a relationship between the biological data and the compounds descriptors.

The first attempts to obtain such a quantitative relationship were done using multilinear regression (MLR). As described in recent reviews, there are, however, a number of pitfalls to this method [88, 89]. To avoid statistically non-significant relationships or chances correlation, one should always apply the following rules: (1) the ratio of compounds to descriptors should be greater than five [90, 91], and (2) the descriptors should not be intercorrelated.

Partial least squares (PLS) can be used to solve these problems. PLS can handle numerous and even collinear variables. PLS is particularly useful when many descriptors are taken into consideration [92]. Another alternative is the use of principal component analysis (PCA) combined with MLR. PCA is a statistical technique which transforms a set of partially cross-correlated data into a smaller set of new orthogonal variables called the principal components (PC's) which still retain much of the descriptive power of the original data. PCA analysis has recently been used to describe the biological activities and molecular diversity of heterocyclic aromatic ring fragments [93]. The aim of this study was to (1) identify principal components which correlate the chemical structures with biological activities, and (2) to enable medicinal chemists to rationally select which heterocyclic rings to synthesize in order to optimize biological activities.

The Free-Wilson model [85] was proposed in 1964 at the same time as the Hansch model, but is far less used. It uses indicator values having a value of unity for the presence of a substructural feature and zero for its absence as sole parameters in a Hansch model-like regression equation. The greatest interest of this method lies in its mixed use with the Hansch analysis. The mixed approach combines the advantages of Hansch and Free Wilson analysis and widens the application of both methods. Physicochemical parameters (Hansch model) describe parts of the molecules with broad structural variation, whereas indicator variables (Free-Wilson method) encode the effects of structural variations that cannot be described otherwise [88].

**3D-QSAR**

When the selected compounds in the dataset present differences in the two-dimensional backbone, classical 2D-QSAR can not be used. 3D-QSAR analyses were developed to overcome this problem. 3D QSAR are quantitative models that relate the biological activity of small molecules with their properties calculated in 3D space. In 1988, an approach was proposed to describe molecular properties by fields (usually steric, electronic, hydrogen bonding, and hydrophobic fields) calculated in a regular grid [94]. This method called Comparative molecular field analysis (CoMFA) is one of the most used in 3D-QSAR as evidenced by the 363 CoMFA models reported recently by Kim et al. [95]. In the past ten years, many successful CoMFA applications proved the value of this method especially in cases where classical 2D-QSAR methods failed. This derives from the fact that molecules are described by properties calculated directly from their (supposed) bioactive conformations. Other methods were also developed to derive a 3D-QSAR model. For a more complete description of these methods, and their philosophies, the reader is referred to the review of Oprea and Waller [96].
One of the most difficult aspects of a 3D QSAR analysis is finding the appropriate alignment rules for the training set i.e. the bioactive conformation. For some datasets, this can present difficulties, for example for compounds with a large number of rotatable bonds, a correct alignment is difficult or even impossible. The investigation of rigid analogs or analogs with different conformational constraints helps to find or confirm the bioactive conformations of the more flexible compounds [97]. This problem limits the applicability of CoMFA. In order to overcome this problem, some new approaches, that do not depend on a common alignment of the molecules, have been recently developed. Comparative molecular moment analysis CoMMA [98], EVA [99] or the WHIM [100] descriptors promise an advantage because they provide three dimensional descriptors that are independent of the orientation of the molecules in space; they do not have to be aligned. The EVA, CoMMA and WHIM descriptors differ from the lattice- or surface-based descriptors, in that they do not consider properties at location in space, but rather 3D properties of the molecules themselves. Hence it is not possible to provide a 3D display of the resulting model.

Moreover, 3D-QSAR studies produce a large number of variables, all of which may not be significant for the final result. Hence, methods, such as GOLPE, were developed for selecting most significant variables [101].

Case Study

One of the early examples of a compound designed using 2D-QSAR is norfloxacin 15 [102]. Structural modifications were guided with the assistance of 2D-QSAR. QSAR equation showed that the antibacterial activities of 6-, 7-, or 8-monosubstituted 1-ethyl-1,4-dihydro-4-oxo-quinoline-3-carboxylic acids 16 were parabolically correlated with steric parameters for R1 and R3 substituents. The Hansch equation also revealed that the activity of the 6,7,8-polysubstituted derivatives of 16 might be more potent than those of monosubstituted compounds. In particular the activities of two disubstituted compounds were expected to be very potent. Norfloxacin 15, one of these compounds has reached market since 1983 under various names including Noroxin.

QSAR analysis played also a role in the discovery of donepizil hydrochloride (E2020, 17), an acetylcholinesterase (AchE) inhibitor [103-105]. Eisai marketed this compound as Aricept for patients with Alzheimer’s disease (AD). A random screening procedure identified a benzylpiperazine compound 18 as a AchE inhibitor. This screening was followed by a second one using a library of compounds of similar structures and led to the discovery of another type of compound which has both benzylpiperazine and benzamide moieties in its structures 19. Over the course of lead optimization it appeared that methyl substitution on the amide nitrogen enhanced activity. Since methyl substitution at the amide nitrogen may change the ratio of cis and trans isomers of the amide, conformational analysis of the N-alkyl-substituted benzamides was performed. QSAR analysis, X-ray diffraction, and molecular modeling were used to provide the structural requirements for lead optimization. The specific structural requirements identified by these analysis were as follows: (1) the cis conformation of the benzamide is the active conformation, (2) the bulky groups at the para position of the benzamide increases activity, and (3) the carbonyl oxygen of the amide is a proton acceptor for an intermolecular hydrogen bond. Using those requirements, analogues were made, one of these was E20202 17. The capacity to perform QSAR, X-ray diffraction, and molecular modeling quickly, relative to synthetic efforts and at critical times in the evolution of the research program, are the reasons these methods have been functional. Recently, the X-ray structure of AchE complexed with E2020 17 has been solved. This study has shown, a posteriori, that the design of E2020 17 took advantage of the results derived by the QSAR, X-ray diffraction, and molecular modeling studies [106].
Other examples of the successful use of QSAR have been given in a review by Fujita [107]. Fujita has given detailed reviews of several QSAR-driven design projects which yielded compounds reaching advanced stages of investigation.

**3D Structure Known**

When the structure of the target protein is available the process of lead optimization can be profoundly influenced and speeded, particularly when the three-dimensional structure of the protein-ligand complex is available. As already discussed, in order to allow the chemist to more fruitfully design modifications of the lead structure, it is helpful to have a three-dimensional structure of the ligand as it binds to the receptor or enzyme. Beyond knowledge of the bioactive conformation of the ligand, the observations of how the ligands bind to their macromolecular target, specific interactions that are important in molecular recognition can be inferred. Protein-ligand complex structures can be either obtained (1) experimentally by X-ray crystallography or NMR [56, 57], or (2) computationally using docking programs [108]. Three-dimensional structures of protein-ligand are available in the PDB. ReLiBase is another very useful novel database containing experimental three-dimensional structures of protein-ligand complexes [109]. ReLiBase, in contrast to the PDB, includes information about bond and atom types of non-protein molecules, and implements query tools (not available in the PDB) for identifying ligands and analysing protein-ligand complexes.

![Diagram](image)

**Fig. (9).** Iterative process used in SBDD.
Those information, obtained from protein-ligand complex structures, help in increasing the effectiveness of the ligand optimization process, analogues or new secondary leads can be designed, synthesized, and assayed. The three-dimensional structures of the protein-ligand complex with the new compounds can be redetermine in order to test experimentally whether the design concept was structurally correct. This process can be itered with further rounds of design, synthesis, testing, and so on to ultimately produce potent and specific compounds (Fig. 9) [72, 110].

Case Study

One of the best known example in structure-based drug design (SBDD) is the discovery of Zanamivir (20), also known as 4-guanidino-Neu5Ac2en and GG167), a potent inhibitor of sialidases from both influenza A and B viruses [111]. Zanamivir displays potent antiviral activity both in vitro and in vivo and is currently undergoing phase III clinical evaluation for the influenza treatment [112]. Influenza virus is a major cause of respiratory disease and produces significant morbidity and mortality (influenza is one of the ten most common causes of death in the USA). Every so often, the virus undergoes a major antigenic transformation resulting in a pandemic strain, such as the strain which killed more than 20 million people in 1918-1919. Although widely recognized for many years, there are currently only few drugs available for influenza treatment. The influenza neuramidase (NA) enzyme is an attractive target for antiviral intervention, its active site is antigenically conserved in all clinically relevant strains and is critical to viral replication. Random screening, however, did not produce potent inhibitors of NA [113]. Transition state analogue design, resulted in a promising lead, 2-deoxy-2,3-dehydro-N-acetylneuramic acid (Neu5Ac2en, 21). Synthesis of a number of Neu5Ac2en analogues led to a halogenated derivative (FANA, 22). FANA 22 had only a micromolar $K_i$, and none of its analogues showed antiviral activity in animals. Further progress was only made when the crystal structure of influenza A NA became available. The structure of the NA-sialic acid complex has since been refined to 1.8 Å and characterised in detail in independant studies [114, 115]. The elucidation of the crystal structure of influenza neuramidase was a key turning point allowing the rational design of more potent and specific inhibitors. von Itzstein et al. probed the active site structure computationally using the GRID program [24]. Results from the GRID analysis revealed a conserved pocket that could be filled by a positively charged nitrogen at the 5 position of the unsaturated sialic derivative (Neu5Ac2en, 20).

More recently, innovative compounds that incorporate a carbocyclic structure into the molecule have been developed, [116, 117] the arrangement offers greater chemical stability than earlier compounds and facilitates modification of the molecule to optimize its properties. The most promising carbocyclic compound is (3R, 4R, 5S)-4-acetamido-5-amino-3-(1-ethylpropoxy)-1-cyclohexane-1-carbocyclic acid, also known as Ro640802 (GS4071, 21) [116].

This compound precisely fits into the three-dimensional structure of the neuramidase active site to interact with antigenically conserved residues and competitively inhibit the enzyme. The incorporation of a lipophilic side chain in this molecule exploits X-ray crystallographic evidence of a hydrophobic pocket in the neuramidase active site, enhancing hte affinity for the target.
Another active area in which structure-based drug design (SBDD) helps to the design and the optimization of new compounds concerns the intensive research effort for safe and effective HIV-1 protease inhibitors [118].

**Conclusion**

CADD approaches aim to increase the speed and efficiency in the drug discovery process. CADD, however, not a direct route to new drugs, but it provides a somewhat more detailed map to the goal. The hope is that providing and pieces of information, and by helping to coordinate the information, CADD will help to make the drug design process more rational. The many success stories of the use of CADD in the discovery of new drugs shows the utility of such analyses used in close coupling with traditional medicinal chemistry techniques.

**Acknowledgements**

F. Ooms is indebted to Professor F. Durant, Dr. J. Wouters and Dr. F. Lebon for their comments and advice.

Finally, recognizing that no review article is ever as thorough as it might be, I apologize to anyone who finds my description of his or her method inadequate, or whose work I have accidentally omitted.

**List of Abbreviations**

- CADD = Computer aided drug design
- HTS = High-throughput screening
- MEP = Molecular electrostatic potential
- MLP = Molecular lipophilic potential
- CSD = Cambridge structural database
- HIV = Human immunodeficiency virus
- GPCR = G-Protein coupled receptors
- 5-HT = Serotonin
- DHFR = Dihydrofolate reductase
- fXa = Carboxylate factor
- COX-2 = Cyclooxygenase-2
- QSAR = Quantitative structure activity relationship
- SAR = Structure activity relationships
- 2D-QSAR = Two-dimensional quantitative structure activity relationship
- 3D-QSAR = Three-dimensional quantitative structure activity relationship
- MLR = Multilinear regression
- PLS = Partial least squares
- PCA = Principal component analysis
- CoMFA = Comparative molecular field analysis
- CoMMA = Comparative molecular moment analysis
- AchE = Acetylcholinesterase
- AD = Alzheimer disease
- SBDD = Structure based drug design
- FEP = Free energy perturbation

**References**

